UNIVERSIDAD DE EL SALVADOR

FACULTAD MUTIDISCIPLINARIA ORIENTAL

DEPARTAMENTO DE CIENCIAS NATURALES Y MATEMÁTICA

SECCIÓN DE MATEMÁTICA.

Universidad de El Salvador

Hacia la libertad por la cultura

TESIS:

"MODELOS LINEALES Y ALGUNAS APLICACIONES"

PRESENTADO POR:

VELÁSQUEZ LÓPEZ, ONEYDA YASMÍN

VELÁSQUEZ BONILLA, MARÍA ELVIRENA

PARA OPTAR AL TÍTULO DE:

LICENCIATURA EN ESTADÍSTICA

DICIEMBRE DE 2008

SAN MIGUEL, EL SALVADOR, CENTROAMÉRICA.

UNIVERSIDAD DE EL SALVADOR

FACULTAD MUTIDISCIPLINARIA ORIENTAL

DEPARTAMENTO DE CIENCIAS NATURALES Y MATEMÁTICA

SECCIÓN DE MATEMÁTICA.

Universidad de El Salvador

Hacia la libertad por la cultura

TESIS:

"MODELOS LINEALES Y ALGUNAS APLICACIONES"

PRESENTADO POR:

VELÁSQUEZ LÓPEZ, ONEYDA YASMÍN

VELÁSQUEZ BONILLA, MARÍA ELVIRENA

PARA OPTAR AL TÍTULO DE:

LICENCIATURA EN ESTADÍSTICA

DOCENTE DIRECTOR:

Msc. Est. MARÍA DEL TRANSITO GUTIERREZ REYES

ASESOR METODOLÓGICO:

Msc. Est. JOSÉ ENRY GARCÍA

DICIEMBRE DE 2008

SAN MIGUEL, EL SALVADOR, CENTROAMÉRICA.

UNIVERSIDAD DE EL SALVADOR

FACULTAD MUTIDISCIPLINARIA ORIENTAL

AUTORIDADES UNIVERSITARIAS

RECTOR: Msc. RUFINO ANTONIO QUEZADA SANCHEZ

SECRETARIO GENERAL: Lic. DOUGLAS VLADIMIR ALFARO CHAVEZ

FISCAL GENERAL: Dr. RENE MADECADEL PERLA JIMENEZ

DECANO: Ing. DAVID ARNOLDO CHAVEZ SARAVIA

VICEDECANA: Dra. ANA JUDITH GUATEMALA DE CASTRO

DEPARTAMENTO DE CIENCIAS NATURALES Y MATEMÁTICA

JEFE DEL DEPARTAMENTO: Lic. ABEL MARTÍNEZ LÓPEZ

SECCIÓN DE MATEMÁTICA

COORDINADORA: Licda. MARÍA OLGA QUINTANILLA DE LOVO

AGRADECIMIENTOS

A DIOS TODO PODEROSO:

En este momento en cual he culminado mis estudios, quiero darle gracias a Dios, por haberme permitido lograr mi sueño, además de brindarme la Sabiduría y Bendición en este proceso.

A MIS PADRES:

Florida Arjen López. Por brindarme su amor, dedicación, entrega y por toda la ayuda que me ha brindado siempre, y porque es un ejemplo de que cuando se quiere algo en la vida se puede lograr.

Vidal Velásquez Paz. Por su ayuda.

A MIS HERMANOS:

José Mauricio, Darwin Antonio y **Alma Graciela**, por el apoyo que me dieron cuando más lo necesitaba.

A MIS ABUELOS:

Antonio Velásquez, Catalina Paz, Francisco López y Virginia López. Por sus palabras, y por toda la ayuda que me brindaron.

A MIS AMIGOS:

A todos mis amigos, especialmente a **María Elvirena Velásquez**, por haberme ayudado en los momentos más difíciles de mi vida.

Oneyda Yasmín Velásquez López

AGRADECIMIENTOS

A DIOS TODO PODEROSO:

A Dios Padre por darme su amor en abundancia, a Dios Hijo por darme su gracia, a Dios Espíritu Santo por darme sabiduría y a la Virgen Maria por interceder a su hijo amado por mí. Gracias Santísima Trinidad por darme todo lo necesario para lograr este éxito.

A MIS AMADOS PADRES:

José Serapio Velásquez y María Dora Bonilla de Velásquez por su apoyo incondicional y por la educación moral y religiosa que me dieron e hicieron de mí una persona de bien.

A MIS ABUELOS:

Josefina Zavala de Bonilla por todo el apoyo que siempre me ha dado y por las muchas oraciones que hace en intersección por mí.

Catalino Velásquez por enseñarme que todo lo que uno se propone lo puede lograr y por todo su apoyo incondicional.

A MIS HERMANOS:

Gracias por apoyarme moral y económicamente en todos los momentos de mi carrera.

A MIS TIOS:

Por ayudarme económicamente en especial a mi tío **Carlos Salvador** y demás familiares y amigos que de alguna forma me ayudaron.

María Elvirena Velásquez Bonilla.

ÍNDICE

Contenidos **Pág.**

Introducción……xiv

Antecedentes…….....xvi

Justificación…….xxi

Objetivos generales y específicos……..xxii

Capítulo 1: Modelo de Regresión Lineal Simple…….23

1.1 Introducción al Modelo de Regresión Lineal Simple……...23

1.2 Aplicaciones del Modelo de Regresión Lineal Simple……25

1.3 Definición de Términos Básicos……..26

1.4 Estadística Descriptiva Bidimensional…….28

1.4.1 Distribuciones Marginales y Distribución Condicional……...……30

1.4.2 Diagramas de Dispersión…….…….....31

1.4.3 Covarianza…….……...……36

1.4.4 Coeficiente de Correlación…...…….…….....38

1.5 Construcción de un Modelo Estadístico……...41

1.5.1 Concepto de la Función de Regresión Poblacional (FRP)……..47

1.5.2 Especificación Estocástica de la Función de Regresión Poblacional……..49

1.5.3 Naturaleza Estocástica del Error o Término de Perturbación…….51

1.5.4 Función de Regresión Muestral (FRM)…….…..53

1.6 Asunciones del Modelo de Regresión Lineal Simple……...58

1.6.1 Comentarios a las Asunciones Anteriores……...62

Ejercicios 1……..……...63

Apéndice 1: Deducción de Ecuaciones y Propiedades……..….66

1.1 Deducción de Ecuaciones Utilizadas en el Capítulo 1…….…...…….....66

1.2 Solución de Ejemplos Haciendo Uso del Software Estadístico SPSS v15.0...71

Capítulo 2: Estimación y Prueba de Hipótesis ……...…….76

2.1 Introducción a la Estimación y Prueba de Hipótesis…...…….....76

2.2 Definición de Términos Básicos……..…77

2.3 Estimación de los Parámetros por el Método de Mínimos Cuadrados Ordinarios (MCO)……...…….……..79

2.3.1 Estimación de β_0 y de β_1 ……82

2.3.2 Propiedades de los Estimadores de Mínimos Cuadrados y el Modelo de Regresión Ajustado……..86

2.4 Estimación de σ^2…...…….…...91

2.5 Coeficiente de Determinación r^2: Medida de la Bondad del Ajuste ……..92

2.6 Prueba de Hipótesis de la Pendiente $\hat{\beta}_1$ y del Intercepto $\hat{\beta}_0$ ……108

2.6.1 Uso de las Pruebas t……....…….....108

2.6.2 Prueba de Significancia de la Regresión……....….110

2.6.3 Análisis de Varianza……...114

2.6.4 Prueba de Hipótesis de la Correlación……120

2.7 Estimación de Intervalo en la Regresión Lineal Simple…….....122

2.7.1 Intervalos de Confianza de β_0, β_1, σ^2…….122

2.8 Estimación por Máxima Verosimilitud…….……..…126

Ejercicios 2……..…….……129

Apéndice 2: Deducción de Ecuaciones…….……..….133

2.1 Deducción de Ecuaciones Utilizadas en el Capítulo 2…….…….……..133

2.2 Solución de Ejemplos Haciendo uso del Software Estadístico SPSS v15.0...156

Capítulo 3: Validación del Modelo y Predicción……...…….....161

3.1 Introducción a la Validación del Modelo y Predicción.……161

3.2 Análisis de los residuos……..162

3.3 Validación del Modelo Mediante los Residuos……..165

3.3.1 Linealidad…..……165

3.3.2 Homoscedasticidad……166

3.3.3 Normalidad……166

3.3.4 Independencia…...…….167

3.4 Predicción Usando el Modelo……173

3.4.1 Predicción Media…….……..173

3.4.2 Predicción Individual……..…177

Ejercicios 3……..…….180

3.5 Análisis de los Residuos Haciendo uso del SPSS v15.0……181

Capítulo 4: Modelo de Regresión Lineal Múltiple.……188

4.1 Introducción al Modelo de Regresión Lineal Múltiple.……188

4.2 Definición de Términos Básicos……189

4.3 Asunciones del Modelo de tres Variables……..190

4.4 Interpretación de la Ecuación de Regresión Lineal Múltiple…….191

4.5 Significado de los Coeficientes de Regresión Parcial……191

4.6 Estimación de los Coeficientes de Regresión Parciales por Mínimos Cuadrados Ordinarios (MCO)……..192

4.6.1 Estimadores de MCO……192

4.6.2 Varianza y Errores Estándar de los Estimadores de MCO……194

4.6.3 Propiedades de los Estimadores de MCO……196

4.7 Coeficiente de Determinación Múltiple R^2 y el Coeficiente de Correlación Múltiple R…….…..199

4.7.1 Comparación de Dos o Más Valores de R^2: El R^2 Ajustado……..…201

4.7.2 Coeficientes de Correlación Parcial…….….204

4.8 Supuesto de Normalidad…….……220

4.8.1 Pruebas de Hipótesis sobre Coeficientes Individuales de Regresión Parcial……..….222

4.8.2 Pruebas de la Significación Global de la Regresión Muestral……..226

4.8.3 Análisis de Varianza en las Pruebas de Significancia Global de una Regresión Múltiple……227

4.8.4 Importancia de la Relación entre R^2 y F……231

4.8.5 Intervalos de Confianza en Regresión Múltiple……233

4.8.5.1 Intervalos de Confianza de los Coeficientes de Regresión…….....233

Ejercicios 4……..…….236

Apéndice 4: Deducción de Ecuaciones…….……..…….243

4.1 Deducción de Ecuaciones Utilizadas en el Capítulo 4…….……..……..…243

4.2 Solución de Ejemplos Haciendo uso del Software Estadístico SPSS v15.0...251

Capítulo 5: Modelo de Regresión Lineal Múltiple Haciendo Uso del Algebra Matricial…….259

5.1 Introducción al Modelo de Regresión Lineal Múltiple…….…….259

5.2 Definición de Términos Básicos……260

5.3 Modelos de Regresión Lineal con k Variables……..261

5.4 Asunciones del Modelo Regresión Lineal con k Variables en Notación Matricial…….264

5.5 Estimación de los Coeficientes de Regresión por Mínimos Cuadrados Ordinarios (MCO)……...267

5.5.1 Matriz de Varianza- Covarianza de $\hat{\beta}$ …….274

5.5.2 Propiedades del Vector $\hat{\beta}$ de Mínimos Cuadrados Ordinarios……277

5.6 Coeficiente de Determinación R^2 en Notación Matricial……..278

5.7 Pruebas de Hipótesis con Notación Matricial……279

5.7.1 Pruebas de la Significación de la Regresión…….281

5.7.2 Análisis de Varianza en Notación Matricial…….282

5.7.3 Intervalos de Confianza en Regresión Múltiple……284

5.7.3.1 Intervalos de Confianza de los Coeficientes de Regresión…….284

5.7.3.2 Estimación del Intervalo de Confianza de la Predicción Media……285

5.7.3.3 Intervalo de Confianza para la Predicción Individual……286

5.8 Matriz de Correlación……287

Ejercicios 5…….298

Apéndice 5: Deducción de Ecuaciones……302

5.1 Deducción de Ecuaciones Utilizadas en el Capítulo 5…….. …….302

Capítulo 6: Modelo de Regresión Lineal con Variable Independiente Cualitativa……306

6.1 Introducción al Modelo de Regresión con Variable Cualitativa……306

6.2 Definición de Términos Básicos……307

6.3 Naturaleza de las Variables Cualitativas……..…….308

6.4 Regresión de una Variable Cuantitativa y una Cualitativa con dos Categorías.310

6.5 Regresión de una Variable Cuantitativa y una Cualitativa con más de dos Categorías……...315

6.6 Regresión de una Variable Cuantitativa y dos Variables Cualitativas…..…….317

6.7 Interacción entre Variables Cualitativas y Cuantitativas……329

6.8 Comparación de Modelos de Regresión…….343

6.9 Uso de las Variables Dicótomas en el Análisis Estacional……345

6.10 Regresión Lineal por Tramos……..350

Ejercicios 6……...352

Capítulo 7: Extensiones del Modelo de Regresión y Violación de Supuestos…..……355

7.1 Introducción……..…….355

7.2 Definición de Términos Básicos……356

7.3 Modelos de Regresión Lineal…….357

7.3.1 Modelos Polinomiales en una Variable…….....358

7.4 Modelos no Lineales y Transformaciones……..……..369

7.5 Regresión con Variable Dependiente Cualitativa…….375

7.5.1 Estimación de Modelos Lineales de Probabilidad……..…….…….....377

7.6 Multicolinealidad……381

7.6.1 Estimación en el caso de la Multicolinealidad Perfecta…….…….....385

7.6.2 Estimación en caso de Multicolinealidad Alta pero Imperfecta…….387

7.6.2 Consecuencias de la Multicolinealidad…….…….…389

7.6.4 Como Detectar la Multicolinealidad…….……..396

7.6.5 Multicolinealidad y Predicción…….……..398

7.6.6 Medidas Remediales…….……..398

7.7 Heteroscedasticidad……404

7.7.1 Consecuencias de la Heteroscedasticidad……..409

7.7.2 Como Detectar la Heteroscedasticidad……..415

7.7.3 Medidas Remediales……..427

7.7.3.1 Cuando se conoce σ_i^2 : Método de Mínimos Cuadrados Ponderados.427

7.7.3.2 Cuando no se conoce σ_i^2 ……430

7.8 Autocorrelación……..435

7.8.1 Consecuencias de la Autocorrelación……445

7.8.2 Como Detectar la Autocorrelación……450

7.8.2.1 Prueba de Durbin-Watson……...454

7.8.3 Medidas Remediales…….460

7.8.3.1 Cuando se conoce la Estructura de la Autocorrelación……...460

Ejercicios 7……...471

Apéndice 7.1: Solución del Ejemplo 1 Haciendo uso del Software Estadístico SPSS v15.0……...……..478
Capítulo 8: Método de Selección de Variables……..482
8.1 Introducción……482
8.2 Construcción de Modelos de Regresión…….483
8.3 Métodos de Selección de Variables……483
8.3.1 Selección Hacia Adelante…….……484
8.3.2 Eliminación Hacia Atrás……485
8.3.3 Regresión Paso a Paso……...485
8.4 Métodos de Selección de Variables Haciendo Uso del SPSS v15.0……..496
Ejercicios 8……...509
Apéndice A: Elementos del Álgebra Matricial……...…….510
Apéndice B: Tablas Estadísticas……..531
Respuesta a los ejercicios planteados……...538
Bibliografía……...552

INTRODUCCIÓN

Los Modelos Lineales han sido usados durante décadas tanto intensiva como extensivamente en aplicaciones Estadísticas.

Llamamos Modelos Lineales a aquellas situaciones que después de haber sido analizadas Matemáticamente, se representan por medio de una función lineal, los cuales son lineales en los parámetros desconocidos e incluyen un componente de error. El componente de error es el que los convierte en Modelos Estadísticos. Estos modelos son la base de la metodología que usualmente llamamos Regresión Múltiple. Por esta razón el manejo de los Modelos Lineales es indispensable para comprender y aplicar correctamente los Métodos Estadísticos.

En algunos casos el modelo coincide precisamente con una recta; en otros casos, a pesar de que las variables que interesan no pertenecen todas a la misma línea, es posible encontrar una función lineal que mejor se aproxime al problema, ayudando a obtener información valiosa.

Un Modelo Lineal se puede determinar de manera gráfica o bien, por medio de una ecuación. Existen ocasiones en que en una de las variables se quiere que cumpla varias condiciones a la vez, entonces surge un conjunto de ecuaciones donde el punto de intersección de dichas ecuaciones representa la solución del problema.

El presente trabajo pretende contribuir al desarrollo de esta rama de la Estadística por medio de la aplicación de la teoría a un problema real y que a su vez pueda ser utilizado como una guía de estudio para los estudiantes de la Licenciatura en Estadística como también por los docentes para el desarrollo del curso de Modelos Lineales, ya que no se encuentra bibliografía completa para el desarrollo del curso.

Se desarrollará la teoría de los Modelos de Regresión Lineal Simple, Estimación y Prueba de Hipótesis, Validación del Modelo y Predicción, Modelos de Regresión Lineal Múltiple, Pruebas de los Parámetros y Validación del Modelo de Regresión Lineal Múltiple, Modelos de Regresión con Variables Cualitativas, otros Modelos y Problemas, y Métodos de Selección de Variables.

Para el desarrollo de los ejemplos o aplicaciones que se realizaran se hará uso del paquete estadístico SPSS v15.0

En cada uno de los capítulos se presenta una pequeña introducción así como también una definición de términos básicos.

Y por último se presentan los apéndices y las referencias bibliográficas que se han utilizado durante la investigación.

ANTECEDENTES

Los primeros intentos de modelar la relación estadística entre dos variables se hicieron en Astronomía en el siglo XVIII con el objeto de contrastar la teoría de Newton.

Adrien M. Legendre (1752-1833) y Carl F. Gauss (1777-1855) resuelven de manera general el problema de explicar la posición de un planeta, variable respuesta, como función de las posiciones de otros cuerpos. Aunque según la teoría de Newton la relación es Matemática o Determinista, los errores de observación de los instrumentos existentes requerían un procedimiento Estadístico para modelar la relación entre las variables observadas. Legendre resolvió este problema inventando el Método de Estimación de Mínimos Cuadrados, que es aún la herramienta más utilizada para la Estimación de Modelos Estadísticos. Gauss, independientemente, obtuvo tambien este resultado y demostró su optimalidad cuando los errores de medida siguen una Distribución Normal.

Francis Galton (1822-1911) fue un hombre de profunda curiosidad intelectual que le llevo a viajar por todo el mundo, a realizar actividades tan diversas como redactar leyes para los hotentotes[*] que gobernaban en el sur de África, realizar investigaciones productivas en Meteorología (a él le debemos el termino anticiclón) o descubrir la

[*] Los khoikhoi ("hombres de los hombres"), a veces llamados hotentotes o simplemente khoi, son una raza nómada del sudoeste de África.

singularidad de las huellas digitales en el cuerpo humano. Galton se interesó en estudiar la transmisión de características entre generaciones, con el objetivo de contrastar las teorías de su primo Darwin, y comparó las estaturas de padres e hijos. Encontró que los padres altos tenían, en promedio, hijos altos, pero en promedio mas bajos que sus padres, mientras que los padres bajos tenían hijos bajos, pero, en promedio, mas altos que sus padres. Este fenómeno, que él denominó de regresión a la media, se ha encontrado en muchas características hereditarias, de manera que los descendientes de personas extremas en alguna característica estarán, en promedio, mas cerca de la media de la población que sus progenitores. El trabajo de Galton condujo ha denominar Métodos de Regresión a los desarrollados para medir la relación Estadística entre dos variables, y estimuló a Karl Pearson (1857-1936), Matemático y Filósofo inglés para inventar el Coeficiente de Correlación Lineal.

Francis Y.Edgeworth (1845-1926), Economista inglés influido por la obra de Galton, estudia la conexión entre los Modelos de Regresión y las distribuciones condicionadas en la Normal Multivariante. Edgeworth encontró procedimientos para calcular la esperanza y la varianza condicionada de la Normal Multivariante sin ninguna referencia al Método de Mínimos Cuadrados.

George U.Yule (1871-1951) introdujo el Coeficiente de Correlación Múltiple y Parcial.

Cualquiera que sea el origen de la Modelación Estadística, hay que reconocer que es hasta la década de los años treinta del siglo XX cuando Ronald A. Fisher desarrolló de forma integral una familia de Modelos para resolver un tipo genérico de problemas, inventando el Análisis de la Varianza (ANOVA) y los correspondientes Modelos, hoy conocidos como Modelos ANOVA. Siguiendo esta perspectiva, Bartlett en 1935 publicó un trabajo para modelar tablas de contingencia donde ya se percibe el germen de un modelo equivalente a los modelos ANOVA para datos discretos. Sin embargo, no es hasta los años cincuenta cuando Lancaster, Roy y Kastenbaun desarrollan los Modelos Log-Lineales y Bhapkar, Koch, Grizzle y Starmer, los Modelos Lineales Generales para datos en tablas de contingencia. Después de las propuestas de estos modelos, una gran cantidad de autores han contribuido a su desarrollo (para una literatura hasta 1944, ver Killion and Zahn, 1976), destacándose Goodman, Mosteller y Cox, entre los más importantes. Hay que resaltar aquí la contribución de Birch (1963), quien expresó el Modelo Log-Lineal en la forma actual, equivalente a los Modelos ANOVA. Sin temor a equívoco, es posible asegurar que el detonante de la Modelación Estadística en datos discretos lo constituyen el trabajo de Nelder y Wedderburn (1972), que presenta, a partir de los Modelos Lineales Generalizados, un marco teorico general para el estudio de los Modelos Estadísticos, incluyendo los Modelos de Regresión Lineal para respuestas continuas, dicótomas (logística), de conteos (Poisson) y los Modelos de medias (ANOVA).

La Modelación requiere necesariamente de supuestos, pues de otra manera no podríamos representar a escala y con sencillez una realidad compleja.

Un buen modelo puede ser aquel que se enfoque principalmente en describir la realidad, pero también aquel que tenga capacidad de hacernos ver mas allá de lo que a primera vista parece ofrecer. Un modelo "malo" es aquel altamente realista, pero tan complicado que se vuelve inmanejable; en este caso no hay razón para construirlo.

A menudo se usan o se hacen pronósticos de una forma u otra. Pocos reconocen sin embargo, que alguna clase de estructura lógica o modelo, está implícita en cada pronóstico. Por tanto, incluso un pronosticador intuitivo construye algún tipo de modelo, quizá sin percatarse de que lo hace. Construir modelos obliga al individuo a pensar con claridad y explicar todas las interrelaciones importantes implicadas en un problema. Fiarse de la intuición puede ser peligroso a veces debido a la posibilidad de que se ignoren o se usen de manera inapropiada relaciones importantes.

Además, es importante que las relaciones individuales sean validadas de alguna manera. Pero, generalmente no se hace esto cuando se realizan pronósticos intuitivos. Sin embargo, en el proceso de construir un modelo, una persona debe validar no sólo el modelo en conjunto sino también las relaciones individuales que forman el modelo.

Al hacer un pronóstico, también es importante proporcionar una medida de la precisión que esperamos del pronóstico. El uso de métodos intuitivos, por lo general, impide cualquier medida cuantitativa de confianza en el pronóstico resultante. El Análisis

Estadístico de las relaciones individuales que forman un modelo, y del modelo como un conjunto, hace posible adjuntar una medida de confianza a los pronósticos del modelo.

Una vez que se ha construido un modelo y se ha adecuado a los datos, puede usarse un análisis de sensibilidad para estudiar muchas de sus propiedades. En particular, pueden evaluarse los efectos de cambios pequeños en variables individuales en el modelo. Por ejemplo, en el caso de un modelo que describe y predice tasas de interés, uno podría medir el efecto en una tasa de interés particular de un cambio en el índice de inflación. Este tipo de estudio de sensibilidad sólo puede realizarse si el modelo está en forma explícita.

JUSTIFICACIÓN

Los Modelos Lineales constituyen una de las Metodologías Estadísticas más ampliamente utilizadas en la Modelización y el análisis de datos de todo tipo, estos se encuentran además en la base de técnicas tan populares como la Regresión y Análisis de Varianza, también el estudio de los Modelos Lineales requiere de conocimientos teóricos en un nivel avanzado sobre Álgebra Lineal y Estadística.

Es por ello que se desea conocer mas a fondo la teoría de los Modelos Lineales y conocer las áreas de aplicación de los modelos, además de la necesidad que tienen los estudiantes de la carrera de Licenciatura en Estadística a tener acceso a un documento que se adecue a las exigencias que tendrán al someterse a un curso de Modelos Lineales, y es una de las áreas que corresponde al plan de estudios, la cual tiene un soporte bibliográfico limitado en el sentido de que los textos existentes no enfocan problemas de nuestra realidad, además la mayoría esta escrito en el idioma inglés.

Otra razón es que con la facilitación de este material vamos a poder colaborar con la enseñanza de Los Modelos Lineales, para que se obtenga una mejor profesionalización en el área de la Estadística.

OBJETIVOS

OBJETIVOS GENERALES

- Adquirir dominio de la teoría Matemática y aplicaciones de los Modelos Estadísticos Lineales, para ajustar Modelos de Regresión Lineal Simple o Múltiple a un conjunto de datos.

- Ilustrar como construir Modelos que expliquen el comportamiento de una variable de interés, la variable respuesta, como resultado del efecto de un conjunto de variables explicativas y mostrar la utilización de estos Modelos para hacer predicciones o tomar decisiones.

OBJETIVOS ESPECÍFICOS

- Evaluar la bondad de ajuste en los Modelos estimados.

- Proporcionar las herramientas de cómo construir un Modelo a partir de un conjunto de datos.

- Estudiar la Multicolinealidad en un conjunto de datos, la Heteroscedasticidad y la Autocorrelación en los residuos.

- Utilizar el software SPSS v15.0 como una herramienta en la aplicación de los Modelos a estudiar.

Capítulo 1

Modelo de Regresión Lineal Simple.

1.1 Introducción al Modelo de Regresión Lineal Simple.

El modelo de regresión lineal simple permite explicar la relación entre dos variables.

El objetivo es explicar el comportamiento de una variable "y", que denominaremos variable explicada (dependiente, endógena o respuesta), a partir de otra variable "x", que llamaremos variable explicativa (independiente o exógena).

Este modelo es muy utilizado y su estudio conforma un área de Investigación Clásica dentro de la Ciencia Estadística desde hace muchos años.

Mediante la Regresión Lineal Simple, se busca hallar la línea recta que mejor explica la relación entre una variable independiente y una variable dependiente. Se trata de cuantificar cuánto varía la variable respuesta con cada cambio en la variable independiente. Cuando sólo se incluye en el modelo una variable independiente se habla de Regresión Lineal Simple. En los modelos de Regresión Lineal Simple la variable dependiente será siempre cuantitativa.

Son numerosas las aplicaciones de la regresión, y, las hay en diversos campos como:

Ingeniería, Ciencias Físicas, Ciencias Químicas, Economía, Administración, Ciencias Biológicas y Ciencias Sociales, entre otras.

Como ejemplo de un problema real aplicado a la Economía, se puede estudiar la relación que existe entre los ingresos y gastos de un grupo de estudiantes.

Si "y" representa los gastos semanales de los estudiantes y "x" representa los ingresos semanales, la ecuación de una recta que relaciona estas dos variables es:

$$y = \beta_0 + \beta_1 x \qquad \textbf{(1.1)}$$

Donde:

β_0: Es la ordenada al origen.

β_1: Es la pendiente.

Ahora bien, los datos no caen exactamente sobre una recta, por lo que se debe modificar la ecuación (1.1), para tomar en cuenta esto; sea ε la diferencia entre el valor observado de "y" y el de la línea recta ($\beta_0 + \beta_1 x$) un **error**. Conviene imaginar que ε es un error estadístico, esto es, que es una variable aleatoria que explica por qué el modelo no ajusta exactamente los datos.

Este error puede estar formado por los efectos de otras variables sobre los gastos de los estudiantes, por errores de medición, etc. Así, un modelo más adecuado para los datos de los gastos de los estudiantes es:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \textbf{(1.2)}$$

La ecuación (1.2) se llama Modelo de Regresión Lineal.

Por costumbre se dice que “x” es la variable independiente y “y” la variable dependiente. Como la ecuación (1.2) sólo tiene una variable independiente, se le llama Modelo de Regresión Lineal Simple.

1.2 Aplicaciones del Modelo de Regresión Lineal Simple.

Son muchas las ciencias en las cuales se pueden observar las diferentes aplicaciones del modelo de Regresión Lineal Simple, entre las cuales podemos mencionar:

1. Economía:
   - Se puede estudiar si la demanda de un determinado producto está relacionado con el precio de éste.
   - Si el salario de una persona está relacionado con la experiencia laboral.

2. Medicina:
   - Efecto de la quimioterapia en los enfermos de cáncer.
   - Analizar la relación entre presión sanguínea y edad.
   - Estudiar la relación entre la estatura y el peso.
   - Investigar si el peso está relacionado con el colesterol.
   - Se puede estudiar la relación entre la concentración de un medicamento inyectable y la frecuencia cardiaca.

3. Agronomía:
    - Determinar si la cantidad de abono está relacionado con el crecimiento del maíz.
    - Analizar la relación de determinada vitamina en la producción de leche.

4. Ingeniería:
    - Estudiar si la construcción de un edificio está relacionado con el tiempo.

5. En la Industria:
    - Se puede saber si el contenido de alquitrán en el producto de salida de un proceso químico está relacionado con la temperatura con la que se lleva a cabo.

6. Educación:
    - Determinar si el rendimiento académico de un estudiante está relacionado con el tiempo que dedique a estudiar.

1.3 Definición de Términos Básicos.

Bidimensional: Son dos variables aleatorias definidas sobre el mismo espacio de probabilidad.

Coeficiente de Correlación: Raíz cuadrada del coeficiente de determinación. Su signo indica la dirección de la relación entre dos variables, directa o inversa.

Diagrama de Dispersión: Gráfica de puntos en una red rectangular; las coordenadas "x" y "y" de cada punto corresponden a las dos mediciones hechas sobre un elemento particular de muestra, y el patrón de puntos ilustra la relación entre las dos variables. El diagrama de dispersión también se conoce como nube de puntos.

Error ε: Error que surge de diferencias o cambios aleatorios en los entrevistados o las situaciones de medición.

Heteroscedasticidad: Es una característica del modelo por la que las varianzas del error no son constantes.

Homoscedasticidad: Es una característica del modelo por la que las varianzas del error son constantes.

Linealidad en las Variables: Una función y = f(x) se dice que es lineal en "x", si "x" aparece con una potencia de 1 y no está multiplicada ni dividida por otra variable.

Linealidad en los Parámetros: Una función es lineal en los parámetros digamos β_1, si β_1 aparece con una potencia de 1 y no está multiplicado ni dividido por otro parámetro.

L.q.q.d: Se utilizará al final de cada deducción de fórmula y significa Lo que se quería deducir.

Regresión: Proceso general que consiste en predecir una variable a partir de otra mediante medios estadísticos, utilizando datos anteriores.

Tabla de Contingencia: Tabla que contiene R renglones y C columnas. Cada renglón corresponde a un nivel de una variable; cada columna, a un nivel de otra variable. Las entradas del cuerpo de las tablas son las frecuencias con que cada combinación de variables se presenta.

Valor Atípico: Es un valor inusualmente muy pequeño o muy grande para un conjunto de datos. Gráficamente es un valor que "está lejos" de la mayoría de valores.

Variable Aleatoria: Variable que toma diferentes valores como resultado de un experimento aleatorio.

1.4 Estadística Descriptiva Bidimensional.

Definición: Se denomina variable aleatoria bidimensional al conjunto de dos variables aleatorias unidimensionales X e Y, definidas sobre el mismo espacio de probabilidad.

Más rigurosamente, una variable aleatoria bidimensional (X, Y) es una función que asigna a cada resultado posible de un experimento aleatorio un par de números reales.

Si el número de datos bidimensionales es pequeño, los datos se disponen en dos columnas o en dos filas sobre las que se emparejan los correspondientes valores unidimensionales de una misma realización de la variable bidimensional, como se expresa en la tabla siguiente:

Tabla 1.1 Tabulación de los datos en dos columnas.

Variable X	Variable Y
X_1	y_1
x_2	y_2
.	.
.	.
.	.
x_n	y_n

Es posible estudiar las variables aleatorias bidimensionales, con las dos componentes de naturaleza cualitativa, con las tablas de frecuencias cruzadas o tablas de contingencia.

Si el número de observaciones bidimensionales es grande, se clasifican los n individuos de la muestra en r clases ($A_1,\ldots, A_r$) respecto de la variable X, y en k clases ($B_1,\ldots,B_k$) respecto de la variable Y, entonces los datos suelen organizarse en una tabla como la siguiente:

Tabla 1.2 Doble entrada o contingencia.

Y / X	B_1	B_2	…	B_j	…	B_k	Suma
A_1	f_{11}	f_{12}	…	f_{1j}	…	f_{1k}	f_{1*}
A_2	f_{21}	f_{22}	…	f_{2j}	…	f_{2k}	f_{2*}
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
A_i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{ik}	f_{i*}
.	.	.		.		.	.
.	.	.		.		.	.
.	.	.		.		.	.
A_r	f_{r1}	f_{r2}	…	f_{rj}	…	f_{rk}	f_{r*}
Suma	f_{*1}	f_{*2}	…	f_{*j}	…	f_{*k}	N

En donde f_{ij} es el número de individuos que pertenecen a la clase A_i de la variable X y la clase B_j de la variable Y, y se llama frecuencia absoluta conjunta de la clase A_i x B_j de la variable bidimensional (X, Y).

La frecuencia relativa conjunta de la clase bidimensional A_i x B_j es igual a:

$$h_{ij} = \frac{f_{ij}}{n} \qquad \textbf{(1.3)}$$

1.4.1 Distribuciones Marginales y Distribución Condicional.

Cuando sobre cada individuo de la población se observan dos características aleatorias expresables numéricamente, se tiene una variable aleatoria bidimensional.

Ejemplo 1: Se tiene la población de 40 estudiantes del curso de Estadística Aplicada a la Educación II del ciclo I 2008 de la UES-FMO, en la que se analizan las variables ingresos y gastos semanales de dichos estudiantes.

Ejemplo 2: En la población constituida por 40 estudiantes de Estadística Aplicada a la Educación II del ciclo I 2008 de la UES-FMO, se observa la estatura en cm., y el peso en kg. de cada estudiante.

Mediante una tabla de contingencia se podría describir la relación entre las dos componentes de una variable bidimensional.

En el caso de que ambas variables sean de tipo discreto, como es especialmente el caso cuando las variables son de naturaleza básicamente cualitativa.

Cuando las dos variables sean de tipo cuantitativo, y especialmente cuando se trate de variables continuas como se muestra en los ejemplos anteriores es posible utilizar técnicas más adecuadas para describir y analizar la relación existente entre ambas.

Por supuesto es posible, en primer lugar, construir una tabla de frecuencias cruzadas entre las dos variables, aunque será necesario previamente agruparlas en intervalos.

1.4.2 Diagramas de Dispersión.

Una forma sencilla de describir gráficamente las relaciones constatadas entre dos variables, consiste en representar cada observación por un punto en el plano cuya abscisa sea el valor de la primera variable y cuya ordenada sea el de la segunda. A este tipo de gráfico se le denomina Diagrama de Dispersión.

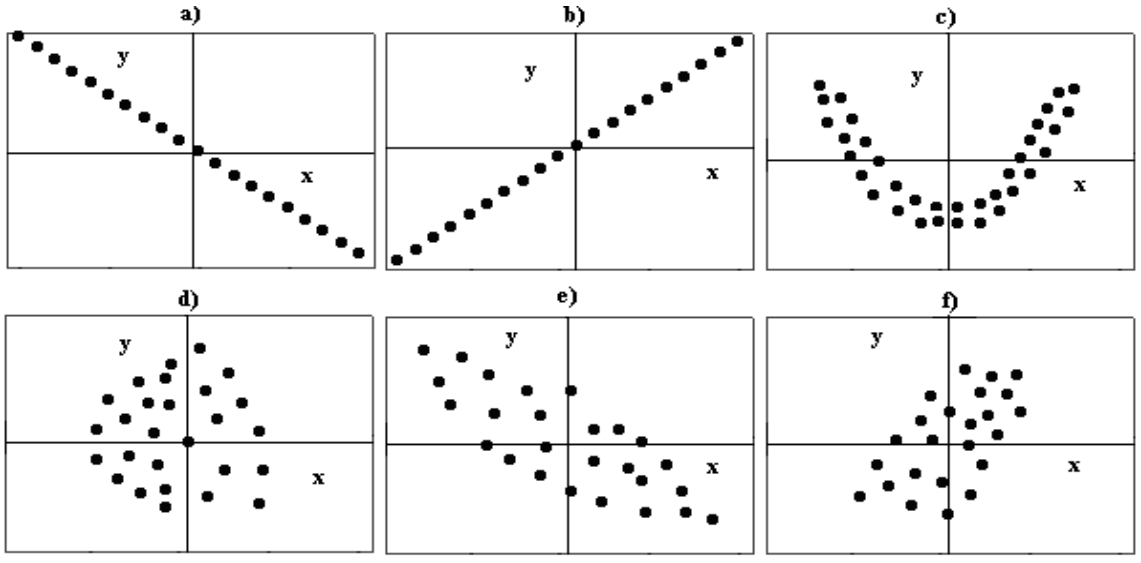
A partir de un conjunto de observaciones de dos variables X e Y sobre una muestra de individuos, el primer paso en un análisis de regresión es representar estos datos sobre los ejes coordenados x, y; esto puede ayudar mucho en la búsqueda de un modelo que describa la relación entre las dos variables.

El diagrama de dispersión se obtiene representando cada observación (x_i, y_i) como un punto en el plano cartesiano xy.

Ejemplo de diagramas de dispersión.

El diagrama de dispersión puede presentar formas diversas:

Figura 1.1 Diagramas de dispersión.



- En los casos a) y b) se tiene que las observaciones se encuentran sobre una recta. En el primer caso, con pendiente negativa, esto indica que a medida que "x" aumenta, la "y" es cada vez menor y en el segundo caso la pendiente es positiva, indicando esto que a medida que la variable "x" aumenta también la variable "y". En estos dos casos los puntos se ajustan perfectamente sobre una recta, de manera que tenemos una relación funcional entre las dos variables dadas por la ecuación de la recta.

- En el caso c) los puntos se encuentran situados en una franja bastante estrecha que tiene una forma bien determinada, se puede observar que no se trata de una relación lineal ya que la nube de puntos tiene forma cuadrática.

- En el caso d) no se tiene ningún tipo de relación entre las variables. La nube de puntos no presenta una forma "tabular" bien determinada; los puntos se encuentran absolutamente dispersos.

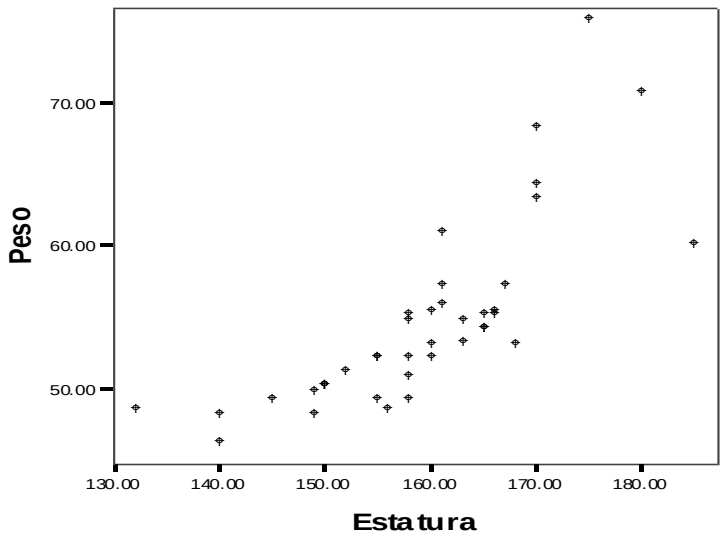
- En los casos e) y f) se puede observar que sí existe algún tipo de relación entre las dos variables. En el caso e) se puede ver un tipo de dependencia lineal con pendiente negativa, ya que a medida que el valor de "x" aumenta, el valor de "y" disminuye. Los puntos no están sobre una línea recta, pero se acercan bastante, de manera que se puede pensar en una fuerte relación lineal. En el caso f) se observa una relación lineal con pendiente positiva, pero no tan fuerte como la anterior.

Ejemplo 3: Si los datos de la población de 40 estudiantes de Estadística Aplicada a la Educación II del ciclo I 2008 de la UES-FMO, de la estatura en cm., y el peso en kg. de cada estudiante, no están agrupados en intervalos (como en la tabla 1.3), entonces el gráfico de dispersión se hace como se muestra en la figura 1.2.

Tabla 1.3 Datos de los 40 estudiantes de Estadística Aplicada a la Educación II.

Individuo	Estatura cm. X	Peso kg. Y	Individuo	Estatura cm. X	Peso kg. Y
1	132	48.3	21	160	52.9
2	140	46	22	160	55.2
3	140	48	23	161	55.66
4	145	49	24	161	57
5	149	48	25	161	60.72
6	149	49.5	26	163	53
7	150	50	27	163	54.5
8	150	50	28	165	54
9	150	50	29	165	54
10	152	51	30	165	55
11	155	49	31	166	55
12	155	52	32	166	55.2
13	155	52	33	167	57
14	156	48.3	34	168	52.9
15	158	49	35	170	63
16	158	50.6	36	170	64
17	158	52	37	170	68
18	158	54.5	38	175	75.5
19	158	55	39	180	70.5
20	160	52	40	185	59.8

Figura 1.2 Diagrama de dispersión de Peso vs. Estatura.



En el diagrama de dispersión figura 1.2 se puede ver claramente la relación positiva entre las dos variables estudiadas, que se refleja en una nube de puntos cuyo eje principal tiene un sentido creciente, como consecuencia del hecho de que, en términos generales, los individuos más altos pesan más que los más bajos.

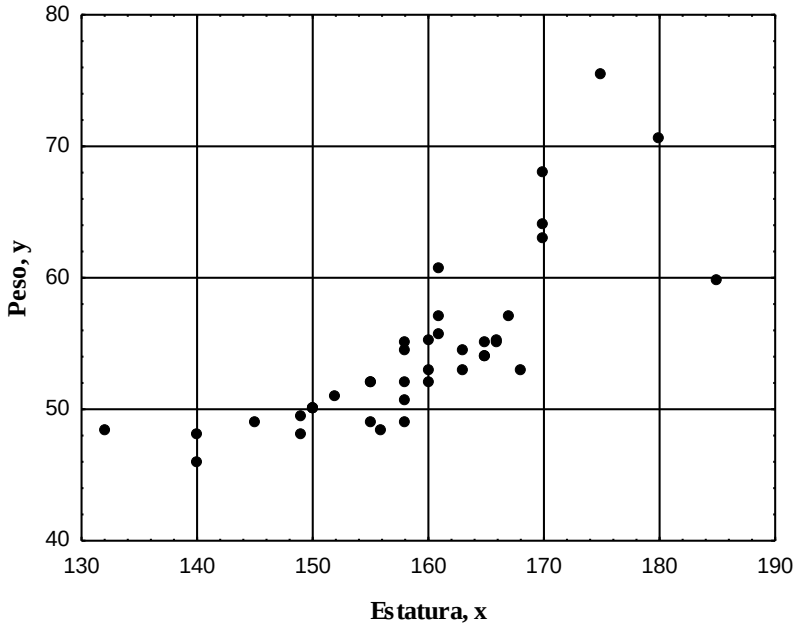
En general cuanto más estrechamente se agrupen los puntos del diagrama de dispersión alrededor de una recta, más fuerte es el grado de relación lineal existente entre las dos variables consideradas. El diagrama de dispersión también puede ayudar a encontrar algún valor atípico, entre los datos de la muestra que pueda tener su origen en una mala observación o en el hecho de ser una observación correspondiente a un individuo excepcional dentro de la muestra. Cuando tenemos un valor atípico, debemos controlar las influencias que pueda tener en el análisis.

Si los datos están agrupados en intervalos como en la tabla 1.4, entonces el diagrama de dispersión se hace como se muestra en la figura 1.3.

Tabla 1.4 Tabla de contingencia.

y / x	40 a < 50	50 a < 60	60 a < 70	70 a < 80	Total
130 a < 140	1	0	0	0	1
140 a < 150	5	0	0	0	5
150 a < 160	3	10	0	0	13
160 a < 170	0	14	1	0	15
170 a < 180	0	0	3	1	4
180 a < 190	0	1	0	1	2
Total	9	25	4	2	40

Figura 1.3 Diagrama de dispersión para datos agrupados en intervalos.



En la figura 1.2 y 1.3 se puede observar que ambos gráficos tienen el mismo comportamiento independientemente de la forma en que se presenten los datos, la ventaja de agrupar es que se reduce el tamaño de la tabla 1.3.

1.4.3 Covarianza.

Con el fin de cuantificar con un índice numérico el grado de relación lineal existente entre dos variables, se utilizan en Estadística dos parámetros: la Covarianza y el Coeficiente de Correlación.

Por definición la Covarianza entre dos variables no es más que el promedio de los productos de las desviaciones de ambas variables respecto a sus medias.

Entre las medidas descriptivas bidimensionales, más utilizadas se tiene la Covarianza entre "x" y "y", que se calcula de la siguiente forma:

1) Si los datos se tabulan en dos columnas (o dos filas), la Covarianza entre "x" y "y" es:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} \qquad \textbf{(1.4)}$$

La deducción de la ecuación (1.4) puede verse en el apéndice **1.1a)**.

2) Si los datos se organizan en una tabla de doble entrada como la 1.2, la Covarianza entre "x" e "y" es:

$$S_{xy} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{k}(x_i - \bar{x})(y_j - \bar{y})f_{ij}}{n} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{k} x_i y_j f_{ij}}{n} - \bar{x}\bar{y} \qquad \textbf{(1.5)}$$

Donde:

x_i : Es la marca de la clase A_i.

y_j : Es la marca de la clase B_j.

f_{ij} : Es la frecuencia absoluta conjunta de la clase bidimensional $A_i * B_j$.

Si en lugar de dividir por n se divide por (n-1) se tiene la Cuasicovarianza o Covarianza modificada o corregida entre “x” y “y”; cuya definición es la siguiente:

1) Si los datos se tabulan en dos columnas (o dos filas), la Cuasicovarianza entre “x” y “y” es:

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad \textbf{(1.6)}$$

2) Si los datos se organizan en una tabla de doble entrada como la 1.2, la Cuasicovarianza entre “x” y “y” es:

$$S_{xy} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{k}(x_i - \bar{x})(y_j - \bar{y})f_{ij}}{n-1} \qquad \textbf{(1.7)}$$

En consecuencia, la Covarianza y la Cuasicovarianza están relacionadas de la siguiente forma:

$$(n-1)S_{xy} = nS_{xy} \qquad \textbf{(1.8)}$$

Por tanto se puede calcular una de ellas a partir de la otra.

La Covarianza (y, por tanto la Cuasicovarianza) es capaz de discriminar entre los dos tipos de relación lineal pues:

1. Si $S_{xy} > 0$, entonces hay relación lineal directa entre “x” y “y”.

2. Si $S_{xy} < 0$, entonces hay relación lineal inversa entre “x” y “y”.

3. Si $S_{xy} = 0$, entonces no hay relación lineal entre “x” y “y”.

1.4.4 Coeficiente de Correlación.

La Covarianza presenta el inconveniente de que depende de las dimensiones en que se expresan las variables. Es decir que la Covarianza entre estatura y peso será 100 veces mayor si la variable estatura se mide en centímetros que si se mide en metros.

Para obviar este problema se utiliza universalmente en Estadística el Coeficiente de Correlación Lineal, como medida del grado de relación lineal existente entre dos variables, que no es más que la covarianza dividida por el producto de las desviaciones típicas de las dos variables, se denota por la letra **r** y se define como:

$$r = \frac{S_{xy}}{S_x S_y} \qquad \textbf{(1.9)}$$

Donde:

S_x: Es la desviación típica de la variable "x".

S_y: Es la desviación típica de la variable "y".

Si la tabulación de datos se hace en dos columnas, entonces una fórmula alternativa equivalente a la ecuación (1.9) es la siguiente:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}} \qquad \textbf{(1.10)}$$

La deducción de la ecuación 1.10 se puede ver en el apéndice **1.1b).**

El coeficiente de correlación lineal está comprendido entre $-1 \leq r \leq 1$.

Los valores extremos de -1 y +1 sólo los toma en el caso de que los puntos del diagrama de dispersión están alineados exactamente en una línea recta.

La interpretación descriptiva de r es la siguiente:

a. Si r = 1, entonces existe una dependencia lineal directa exacta entre las variables "x" y "y". Los puntos del diagrama de dispersión están sobre una línea recta de pendiente positiva figura 1.1 b).
b. Si r = −1, entonces existe dependencia lineal inversa exacta entre "x" y "y". Los puntos del diagrama de dispersión están sobre una línea recta de pendiente negativa figura 1.1 a).
c. Si r = 0, entonces no existe dependencia lineal entre "x" y "y" figura 1.1 d).
d. Cuanto más se aproxime r a −1 ó a 1, más dependencia lineal existe entre "x" y "y". Cuando esto ocurra, el diagrama de dispersión se aproxima a una línea recta.
e. Cuanto más se aproxime r a 0, más independencia lineal existe entre "x" y "y", es decir la variable "y" no depende de "x". Cuando esto ocurra, el diagrama de dispersión no se aproxima a una recta figura 1.1 d).
f. Si r es positivo, entonces al aumentar el valor de la variable "x", aumenta el valor de la variable "y", es decir es directamente proporcional.
g. Si r es negativo, entonces al aumentar el valor de la variable "x", disminuye el valor de la variable "y", en este caso es inversamente proporcional.

Ejemplo 4: Calcular el coeficiente de correlación entre Estatura "x" y el Peso "y" haciendo uso de los datos de la tabla 1.3 y de la ecuación (1.10).

$$\sum_{i=1}^{n} x_i = 132 + 140 + 140 + ... + 185 = 6369 \qquad \sum_{i=1}^{n} y_i = 48.3 + 46 + 48 + ... + 59.8 = 2177.08$$

$$\sum_{i=1}^{n} x_i^2 = (32)^2 + (40)^2 (40)^2 + ... + (85)^2 = 1018437$$

$$\sum_{i=1}^{n} y_i^2 = (48.3)^2 + (46)^2 (48)^2 + ... + (59.8)^2 = 120086.2840$$

$$\sum_{i=1}^{n} x_i y_i = (132)(48.3) + (140)(46) + (140)(48) + ... + (185)(59.8) = 348686.28$$

Sustituyendo estos resultados en la ecuación

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r = \frac{40(348686.28) - (6369)(2177.08)}{\sqrt{40(1018437) - (6369)^2}\sqrt{40(120086.2840) - (2177.08)^2}}$$

$$r = \frac{13947451.2 - 13865822.52}{(416.3159858)(252.5352126)} = \frac{81628.68}{105134.446} = 0.776$$

El coeficiente de correlación lineal obtenido para el ejemplo de Estaturas y Pesos de los estudiantes es 0.776, dado que este valor es cercano a 1 se puede ver que existe relación entre las dos variables así como de que, a medida que la Estatura aumenta, el Peso también lo hace, ya que el valor calculado para **r** es positivo.

En el apéndice **1.2** pueden verse los pasos a seguir para el cálculo del coeficiente de correlación mediante el software estadístico SPSS v15.0

1.5 Construcción de un Modelo Estadístico.

Un modelo estadístico es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión, para indicar los diferentes factores que modifican la variable de respuesta. Si las mediciones se refieren a dos variables, el análisis estadístico puede producir una asociación estadística en las variables.

El análisis de regresión se propone estimar o predecir el valor medio o promedio (poblacional) de la variable dependiente con base en los valores fijos o conocidos de la variable explicatoria, para entender como se lleva a cabo este análisis, examinamos el siguiente ejemplo en el cual la población con la que se trabaja son 40 estudiantes de Estadística Aplicada a la Educación II del ciclo I 2008 de la UES-FMO.

Se tienen los ingresos y los gastos de dichos estudiantes. Se cree que los gastos semanales de un estudiante se relacionan con los ingresos. Las 40 observaciones se presentan en la tabla 1.5

Donde:

x : Ingreso de los estudiantes por semana, en dólares.

y : Gasto de los estudiantes por semana, en dólares.

Tabla 1.5 Ingreso de estudiantes por semana.

x / y	10 a < 20	20 a < 30	30 a < 40	40 a < 50	50 a < 60	60 a < 70	Total
	15	20	30	40	50	55	210
	15	20	30	40			105
	15	20	35	40			110
	15	20					35
	15	20					35
	15	20					35
	16	20					36
	17	20					37
	18	20					38
		23					23
		24					24
		24					24
		25					25
		25					25
		25					25
		25					25
		25					25
		25					25
		25					25
		25					25
		25					25
		25					25
		28					28
		28					28
Total	141	532	95	120	50	55	993

Como en la tabla 1.5 la variable “x” está en intervalos de clase, en la tabla 1.6 los valores de la variable “x” corresponden al valor promedio de cada intervalo con el fin de tener un sólo valor en la variable “x”, por ejemplo para el intervalo de 10-20 el valor promedio o punto medio es $\frac{10+20}{2}=15,$ y así sucesivamente.

Tabla 1.6 Ingreso de estudiantes por semana.

y \ x	15	25	35	45	55	65	Total
	15	20	30	40	50	55	210
	15	20	30	40			105
	15	20	35	40			110
	15	20					35
	15	20					35
	15	20					35
	16	20					36
	17	20					37
	18	20					38
		23					23
		24					24
		24					24
		25					25
		25					25
		25					25
		25					25
		25					25
		25					25
		25					25
		25					25
		25					25
		25					25
		28					28
		28					28
Total	141	532	95	120	50	55	993

La tabla 1.6 debe interpretarse de la siguiente manera: Para un ingreso promedio semanal de $15 hay 9 estudiantes cuyos gastos de consumo semanales oscilan entre $15 y $18. Similarmente, para x = $55 hay un estudiante cuyo gasto de consumo semanal es $50. En otras palabras cada columna de la tabla 1.6 muestra la distribución de los gastos de consumo “y” correspondiente a un nivel fijo de ingreso “x”; esto es, muestra la distribución condicional de “y” condicionada por los valores dados de “x”.

Dado que la tabla 1.6 representa la población, se pueden calcular fácilmente las probabilidades condicionales de “y” $p(y \mid x)$, o probabilidad de “y” dado “x”, de la

manera siguiente. Para x = $25 por ejemplo, hay 23 valores de y: 20, 20, 20, 20, 20, 20, 20, 20, 20, 23, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 25, 28, 28, es decir, dado x = $25, la probabilidad de obtener un gasto cualquiera de estos es 1/23. Simbólicamente $p(y = 28 \mid x = 25) = \frac{1}{23}$ ó para otro valor $p(y = 40 \mid x = 45) = \frac{1}{3}$ y así sucesivamente. Las probabilidades condicionales para los datos de la tabla 1.6 se presentan en la tabla 1.7

Tabla 1.7 Probabilidades condicionales $p(y \mid x_i)$ para los datos de la tabla 1.6.

$p(y \mid x_i)$ \ x	15	25	35	45	55	65
Probabilidades condicionales	1/9	1/23	1/3	1/3	1/1	1/1
	1/9	1/23	1/3	1/3		
	1/9	1/23	1/3	1/3		
	1/9	1/23				
	1/9	1/23				
	1/9	1/23				
	1/9	1/23				
	1/9	1/23				
	1/9	1/23				
	1/9	1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
		1/23				
Media condicional de y	47/3	532/23	95/3	40	50	55

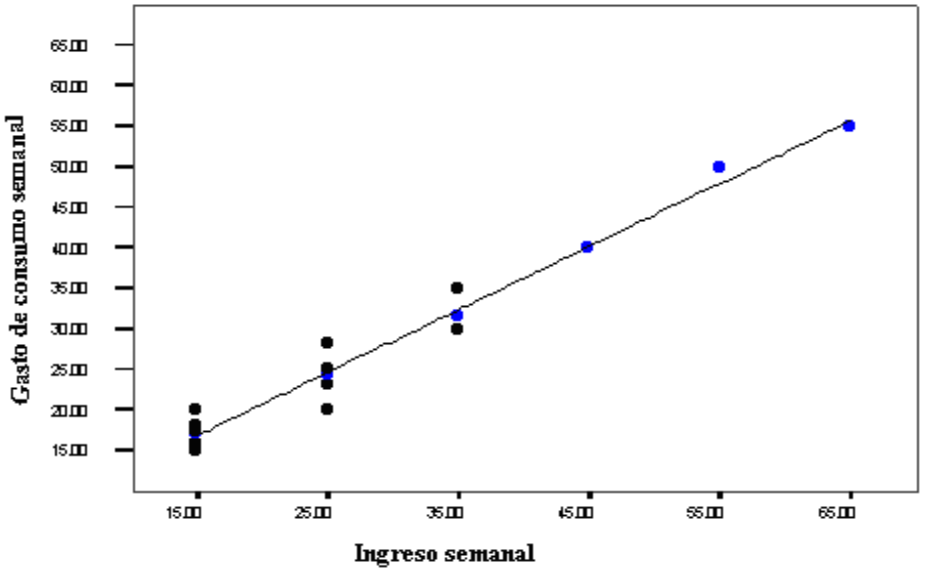
Ahora bien, para cada una de las distribuciones de probabilidad condicionales de "y" se puede calcular su valor medio o promedio, término conocido también como la media condicional o expectativa condicional, que se denota por $E(y \mid x)$ y se lee "el valor esperado de "y" dado x".

Para los datos de la tabla 1.6 las expectativas condicionales pueden ser calculadas fácilmente multiplicando los valores relevantes de "y", dados en la tabla 1.6 por sus probabilidades condicionales dadas en la tabla 1.7 y luego obteniendo la sumatoria de estos productos. Para ilustrar lo anterior se tiene la media condicional o expectativa de "y" dado x = $15 que es igual a:

$$15\left(\frac{1}{9}\right)+15\left(\frac{1}{9}\right)+15\left(\frac{1}{9}\right)+15\left(\frac{1}{9}\right)+15\left(\frac{1}{9}\right)+15\left(\frac{1}{9}\right)+16\left(\frac{1}{9}\right)+17\left(\frac{1}{9}\right)+18\left(\frac{1}{9}\right) \quad = \frac{47}{3}$$

De este modo las medias condicionales aparecen en la última fila de la tabla 1.7

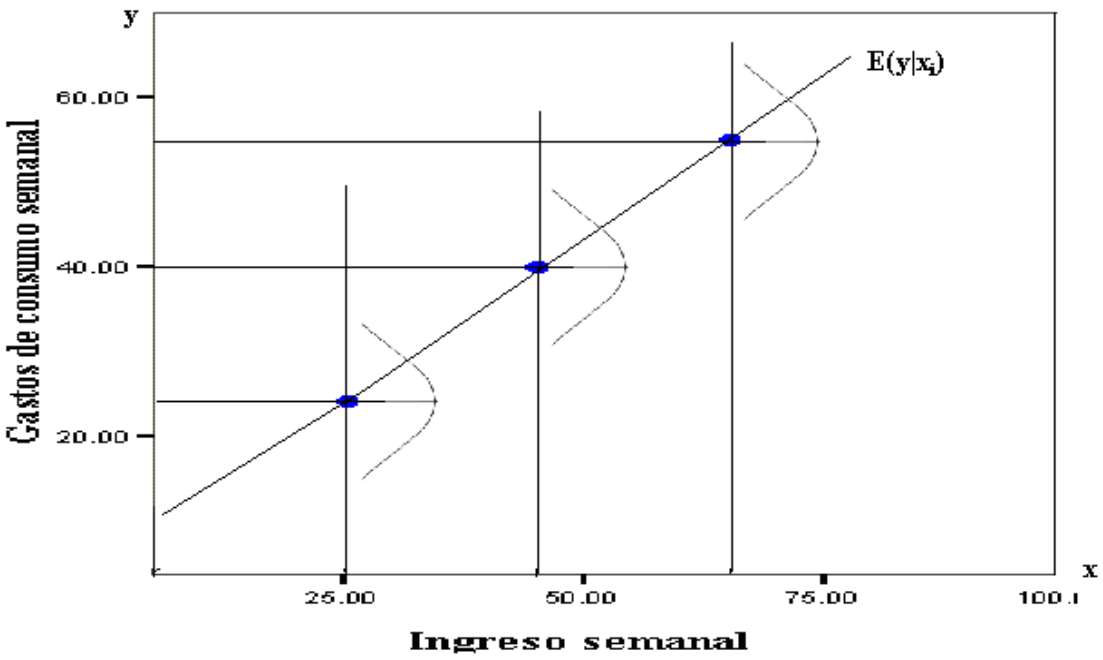
Figura 1.4 Distribución condicional del gasto para varios niveles de ingreso dados en la tabla 1.6



En la Figura 1.4 se presentan los valores de la tabla 1.6 dispuestos en forma de gráfico, además se muestra la distribución condicional de "y" (puntos azules) correspondiente a los valores promedios de "x". A pesar de que ocurren variaciones en los gastos de consumo de los estudiantes, la figura muestra claramente que en promedio los gastos de consumo aumentan al aumentar el ingreso. Dicho de otra manera la figura sugiere que los valores (condicionales) promedios de "y" aumentan al aumentar "x". La afirmación anterior resulta más objetiva si se presta atención en los puntos azules que representan los valores condicionales medios de "y". Estos puntos aparecen sobre una línea recta con pendiente positiva. Esta línea se denomina línea de regresión o más generalmente, curva de regresión o más precisamente, curva de regresión de "y" sobre "x".

Además las medias condicionales no siempre estarán sobre una línea recta pueden perfectamente estar sobre una línea curva, en la figura 1.4 se puede observar que solamente una media condicional está fuera de la curva de regresión de "y" sobre "x" que es la $p(y = 50 \mid x = 55)$ desde el punto de vista de la geometría, una curva de regresión es simplemente el lugar geométrico de las medias condicionales o expectativas de la variable dependiente para los valores fijos de las variables explicatorias. En la figura 1.5 se puede observar que para cada x_i existen ciertos valores poblacionales de "y" y una media (condicional) correspondiente. La línea o curva de regresión atraviesa estas medias condicionales.

Figura 1.5 Línea de regresión.



1.5.1 Concepto de la Función de Regresión Poblacional (FRP).

De las figuras 1.4 y 1.5, se deduce claramente que cada media condicional $E(y \mid x_i)$ es una función de x_i. Simbólicamente, se tiene:

$$E(y \mid x_i) = f(x_i) \qquad \textbf{(1.11)}$$

En donde $f(x_i)$ denota una función de la variable explicatoria x_i. En el ejemplo de construcción del modelo sección 1.5 la $E(y \mid x_i)$ es una función lineal de x_i. La ecuación (1.11) se conoce como la función (de dos variables) de regresión poblacional (FRP) o simplemente regresión poblacional (RP) y denota únicamente que la media (poblacional)

de la distribución de "y" dado x_i está funcionalmente relacionada con x_i. En otras palabras, muestra como el valor promedio (poblacional) de "y" varía con las x_i.

¿Qué forma tiene la función f (x_i)?. Esta pregunta es bastante importante ya que hay situaciones en las que no se dispone de toda la población para el análisis. La forma funcional de FRP es, por lo tanto, un hecho empírico aunque en ocasiones, es necesario recurrir a la teoría. Como se observó en el ejemplo el gasto de consumo de los estudiantes está linealmente relacionado con el ingreso. En consecuencia, como una primera aproximación o hipótesis de trabajo se puede suponer que la FRP: $E(Y \mid x_i)$ es una función lineal de x_i, del siguiente tipo:

$$E(Y \mid x_i) = \beta_0 + \beta_1 x_i \qquad \textbf{(1.12)}$$

En la cual β_0 y β_1 son parámetros desconocidos pero fijos que se conocen con el nombre de coeficiente de regresión, donde β_0 es la ordenada al origen y β_1 es la pendiente.

Se puede interpretar que β_0 representa el valor medio de "y" cuando "x" es cero y la pendiente β_1 es el cambio de la media de "y" para un cambio unitario de "x".

La ecuación (1.12) se conoce como la función de regresión lineal poblacional o simplemente como la regresión lineal poblacional. En el análisis de regresión, nos interesa estimar una FRP como la de la ecuación (1.12), esto es, estimar los valores de las incógnitas β_0 y β_1 con base en las observaciones de "y" y "x" (esto se estudiará en el Capítulo 2).

1.5.2 Especificación Estocástica de la Función de Regresión Poblacional (FRP).

Como claramente se observa en la figura 1.4, al aumentar el ingreso de los estudiantes el gasto de consumo en promedio también aumenta.

¿Qué puede entonces decirse acerca de la relación entre el gasto de consumo de un estudiante y un nivel de ingreso dado?. Observando la figura 1.4 se ve que para un nivel de ingreso dado x_i, el gasto de consumo de un estudiante está concentrado alrededor del consumo promedio de todos los estudiantes para ese mismo x_i, esto es, alrededor de su expectativa condicional. Por consiguiente se puede expresar la desviación de un y_i individual alrededor de su valor esperado de la siguiente manera:

$$\varepsilon_i = y_i - E(y \mid x_i)$$

ó

$$y_i = E(y \mid x_i) + \varepsilon_i \qquad \textbf{(1.13)}$$

En donde la desviación ε_i es una variable aleatoria, no observable, que puede tomar valores positivos o negativos. Técnicamente se conoce a ε_i como la perturbación estocástica o término de error estadístico.

La ecuación (1.13) postula que el gasto de consumo semanal de un estudiante dado su nivel de ingreso, es igual al promedio del gasto de consumo de todos los estudiantes con ese nivel de ingreso, más una cantidad positiva o negativa que es aleatoria. Se supone que el término de error que se agrega al modelo es una variable sustitutiva de todas las variables omitidas que pueden afectar a “y”, pero que por una razón u otra no pueden incluirse en el modelo de regresión.

Si $E(y \mid x_i) = \beta_0 + \beta_1 x_i$ se supone lineal en x_i, como en (1.12), la ecuación (1.13) puede escribirse:

$$\begin{aligned} y_i &= E(y \mid x_i) + \varepsilon_i \\ y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \end{aligned} \qquad \textbf{(1.14)}$$

La ecuación (1.14) plantea el hecho de que el gasto de consumo condicional de un estudiante está relacionado linealmente con su ingreso más un término de perturbación; así, los gastos de consumo dado x = $35 (ver tabla 1.6) pueden expresarse como:

$$\begin{aligned} y_1 &= 30 = \beta_0 + \beta_1(35) + \varepsilon_1 \\ y_2 &= 30 = \beta_0 + \beta_1(35) + \varepsilon_2 \\ y_3 &= 35 = \beta_0 + \beta_1(35) + \varepsilon_3 \end{aligned} \qquad \textbf{(1.15)}$$

Ahora bien, si se toma el valor esperado de (1.13) en ambos lados, se obtendrá:

$$\begin{aligned} E(y \mid x_i) &= E[E(y \mid x_i)] + E(\varepsilon \mid x_i) \\ E(y \mid x_i) &= E(y \mid x_i) + E(\varepsilon \mid x_i) \end{aligned} \qquad \textbf{(1.16)}$$

Habiendo hecho uso de la propiedad que dice que el valor esperado de una constante es igual a la misma constante[1]. Puede verse que en la ecuación (1.16) se ha tomado la expectativa condicional, siendo las x_i la condicionante.

La ecuación (1.16) indica que:

$$E(\varepsilon \mid x_i) = 0 \qquad \textbf{(1.17)}$$

En otras palabras el supuesto de que la línea de regresión pase por las medias condicionales de "y" (ver figura 1.5) implica que los valores medios condicionales de ε_i

[1] Ver apéndice **1.1c)** para una breve discusión de las propiedades del operador E. Nótese que el E $(y|x_i)$, es una constante.

(condicionales a los x_i dados) son cero, dicho de otra manera la media de los errores es cero.

De lo anterior se deduce que (1.12) y (1.14) son formas equivalentes si E ($\varepsilon|x_i$) = 0[2]. Sin embargo, la especificación estocástica (1.14) ofrece la ventaja de mostrar claramente que además del ingreso hay otras variables que afectan el gasto de consumo, y que el gasto de consumo de un estudiante no puede ser totalmente explicado sólo por la o las variables incluidas en el modelo de regresión.

1.5.3 Naturaleza Estocástica del Error o Término de Perturbación.

Como pudo verse en la sección 1.5.2, el término de perturbación ε_i, sustituye a todas aquellas variables que han sido excluidas del modelo, pero que conjuntamente afectan a "y". La pregunta obvia es ¿Por qué no se introducen explícitamente en el modelo todas estas variables? o dicho de otro modo, ¿Por qué no desarrollar un modelo de regresión múltiple con tantas variables como sea posible? Esta interrogante tiene varias respuestas a saber:

1. La teoría, si existe alguna, que determina el comportamiento de "y", suele ser incompleta. Se puede estar seguro de que el ingreso semanal "x" afecta el gasto de consumo "y", pero por otra parte, se puede no estar seguro o desconocer otras variables que afectan a "y". Por lo tanto ε_i puede ser usada como un sustituto de todas las variables excluidas en el modelo.

[2] En efecto, en el método de Mínimos Cuadrados Ordinarios que se desarrollará en el Capítulo 2 se supone explícitamente que E ($\varepsilon|x$) = 0.

2. Aunque se sepa qué variables entre las omitidas son relevantes y se incluyan en una regresión múltiple, es posible que no existan cifras sobre ellas. Es muy común en el análisis empírico que los datos que se desean tener no se encuentren a la disposición. Por ejemplo, se puede en principio introducir la riqueza de los estudiantes, como una variable explicatoria, además del ingreso, para explicar el consumo de los estudiantes. Desafortunadamente, ocurre a menudo que no se encuentra información sobre esta variable, lo cual nos obliga a excluir del modelo la variable riqueza, a pesar de su relevancia teórica en la explicación del gasto de consumo de los estudiantes.
3. Supongamos que además del ingreso x_1, también afecta el gasto de consumo el número de hermanos que estén estudiando x_2, el sexo x_3, la religión x_4 y la región geográfica x_5. Es muy posible que la influencia conjunta de todas o algunas de estas variables sea insignificante o a lo mejor aleatoria o no sistemática y que desde el punto de vista práctico y por razones de costo, no justifique su introducción explícita en el modelo. Cuando así ocurre el efecto combinado de todas las variables, puede ser tratado como una variable aleatoria ε_i [3].
4. Aunque se tenga éxito en la inclusión de todas las variables en el modelo, no deja de existir cierta aleatoriedad "intrínseca" en "y", que a pesar de muchos esfuerzos no puede ser explicada. En tal forma las ε_i pueden reflejar la mencionada aleatoriedad intrínseca.

[3] Las variables sexo y religión son cualitativas y pueden ser de difícil cuantificación.

5. Finalmente siguiendo el principio que dice "las descripciones deben ser tan simples como sea posible a menos que resulten inadecuadas", lo ideal sería tener un modelo de regresión lo más simple posible. Si se puede explicar "sustancialmente" el comportamiento de "y" (vía el r^2 o coeficiente de determinación que se considera en el Capítulo 2) con dos o tres variables, y si además, la teoría no es lo suficientemente sólida como para abarcar otras variables, para qué incluir más variables. Más bien representamos con ε_i todas las demás variables. Sobra decir, que no se deben excluir las variables importantes si se quiere mantener un modelo de regresión sencillo.

Por todas las razones mencionadas anteriormente, la perturbación estocástica ε_i, tiene un papel crítico en el análisis de regresión, que se estudiarán en Capítulos posteriores.

1.5.4 Función de Regresión Muestral (FRM).

Hasta aquí se han limitado los planteamientos a los valores poblacionales de "y" correspondientes a unos x_i fijos. Se ha hecho de manera deliberada, pues no se deseaba hacer consideraciones de muestreo. Obsérvese que las cifras de la tabla 1.6 representan la población de los estudiantes de Estadística Aplicada a la Educación II del ciclo I 2008 de la UES-FMO y no la muestra. Se quiere ahora referirse a la muestra porque en la práctica lo que está a nuestro alcance es una muestra de valores de "y" correspondientes a x_i fijos. Por consiguiente, la tarea actual es la estimación de la FRP con base en la información muestral.

Por ejemplo si se supone que no se conoce la población de la tabla 1.6 y que todo lo que se tiene es una muestra de "y" seleccionada aleatoriamente para los valores fijos de "x" (tabla 1.8). Ahora, no conociendo la tabla 1.6 se tiene un solo valor de "y" para cada "x" dado; cada "y" (dado un x_i) de la tabla 1.8 ha sido escogido aleatoriamente entre sus equivalentes de la tabla 1.6 para cada x_i.

De este modo, se puede formular la siguiente pregunta: ¿De la muestra de la tabla 1.8 es posible predecir el promedio del gasto de consumo de los estudiantes de la población como un todo para las x_i escogidas? En otras palabras. ¿Es posible estimar la FRP con base en los datos muéstrales?. No es factible estimar "con precisión" la FRP debido a las fluctuaciones muéstrales. Para examinar este punto supongamos otra muestra de la población de la tabla 1.6, tal como se presenta en la tabla 1.9.

Tabla 1.8 Primera Muestra Aleatoria de la Población de la Tabla 1.6.

x	15	25	35	45	55	65
y	16	20	35	40	50	55

Tabla 1.9 Segunda Muestra Aleatoria de la Población de la Tabla 1.6.

x	15	25	35	45	55	65
y	20	25	30	40	50	55

Al hacer un diagrama con los datos de las tablas 1.8 y 1.9 se obtiene la figura 1.6, en la cual se dibujan dos líneas de regresión que tratan de "ajustar" los puntos dispersos.

FRM_1 y FRM_2 representan la primera y segunda muestra respectivamente. Sin embargo, la pregunta inicial es: ¿Cuál de las dos líneas de regresión es la "verdadera" línea de regresión de la población? No existe modo alguno de afirmar con certeza, cual de las dos líneas que aparecen en la figura 1.6, representa la verdadera línea de regresión poblacional, aparentemente ambas representan la línea de regresión poblacional pero en razón de fluctuaciones muéstrales, en el mejor de los casos, son una aproximación de la verdadera regresión poblacional.

De manera análoga a la FRP que subraya la regresión lineal poblacional, es posible desarrollar el concepto de Función de Regresión Muestral (FRM) que representa la línea de regresión muestral. La contraparte muestral de la ecuación (1.12) puede escribirse como:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad \textbf{(1.18)}$$

Donde:

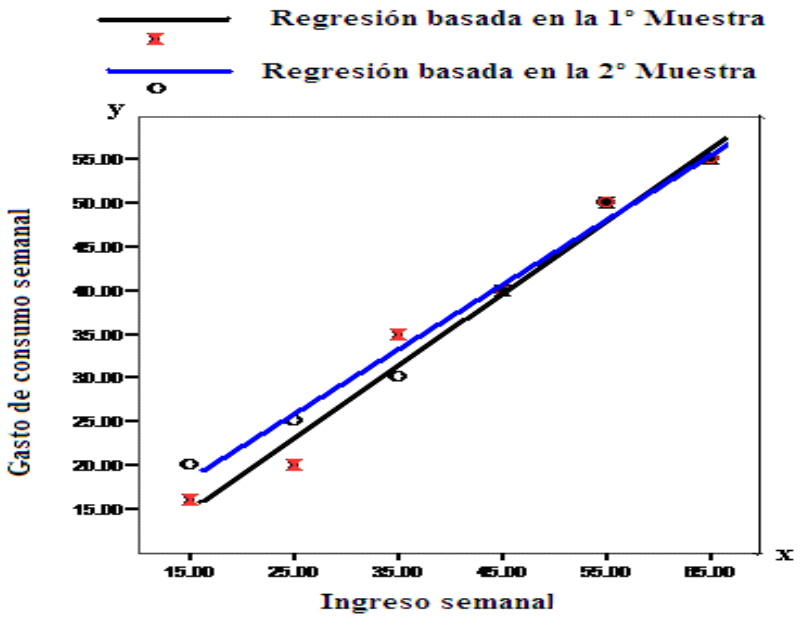
^ : Se lee como sombrero o gorro.

$\hat{y}_i$: Estimador de la E ($y|x_i$).

$\hat{\beta}_0$: Estimador de β_0.

$\hat{\beta}_1$: Estimador de β_1.

Figura 1.6 Líneas de regresión basadas en dos muestras diferentes.



Nótese que un estimador también conocido como un estadístico (muestral), es simplemente una fórmula, que nos dice como estimar el parámetro poblacional a partir de la información proporcionada por la muestra. El valor particular obtenido por el estimador después de una aplicación se conoce con el nombre de estimado[4].

Así como se expresaba la FRP en dos formas equivalentes como las ecuaciones (1.12) y (1.14), se puede también expresar la FRM ecuación (1.18) en su forma estocástica de la siguiente manera:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \textbf{(1.19)}$$

Donde además de los símbolos definidos anteriormente, e_i denota el término residual (muestral). Conceptualmente es análogo a ε_i, y puede ser considerado como un

[4] De aquí en adelante el ^ sobre una variable significará un estimador o estimado del valor poblacional relevante.

estimador de ε_i. Se introduce en la FRM por las mismas razones por las que ε_i fue introducido en la FRP. Resumiendo, el objetivo principal al hacer análisis de regresión, es estimar la FRP

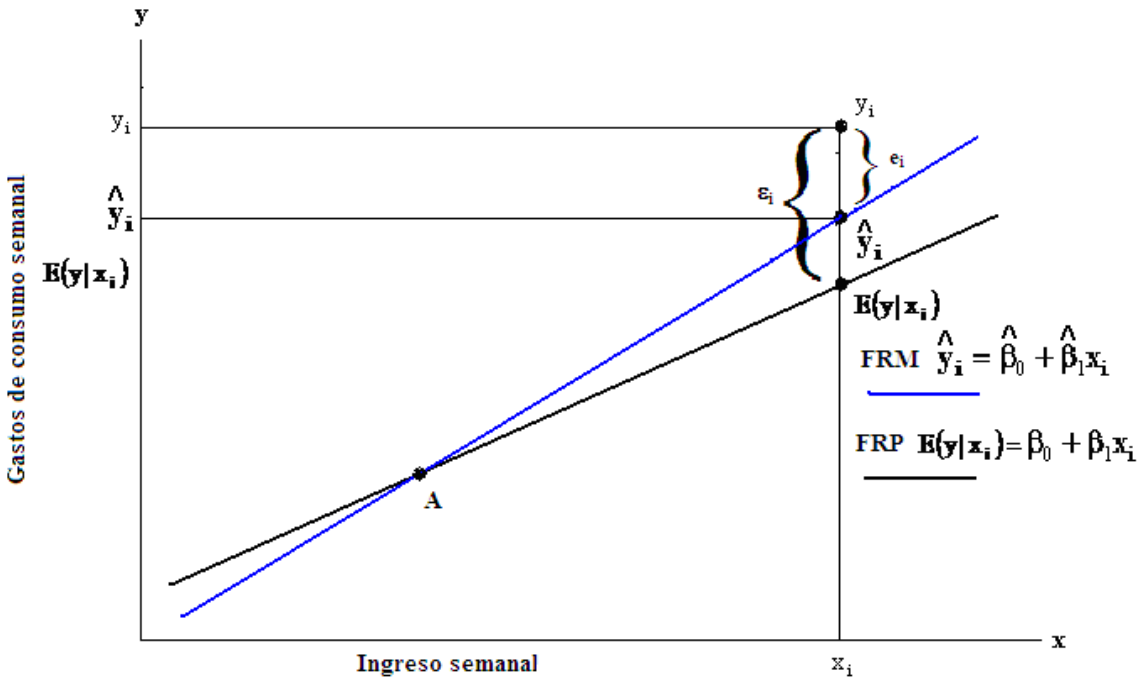
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \textbf{(1.20)}$$

Con base en la FRM

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \textbf{(1.21)}$$

En razón de que en la mayoría de las veces, el análisis se debe llevar a cabo con base en una muestra tomada de una población. Como ya se ha dicho, por fluctuaciones entre una muestra y otra, la estimación de la FRP con base en la FRM es en el mejor de los casos "aproximada". Esta aproximación se representa en forma de diagrama en la figura 1.7.

Figura 1.7 Líneas de regresión poblacional y muestral.



Surge ahora la siguiente pregunta crítica: puesto que se sabe que la FRM es una aproximación a la FRP, ¿Es posible encontrar un método que "acerque" esta aproximación cuanto sea posible? En otros términos, ¿Cómo se debe construir la FRM, para que $\hat{\beta}_0$ y $\hat{\beta}_1$ estén tan cerca como sea posible a β_0 y β_1 respectivamente? Se tratará de dar respuesta a esta pregunta en el Capítulo 2.

1.6 Asunciones del Modelo de Regresión Lineal Simple.

Se admite que todos los factores o causas que influyen en una variable respuesta, pueden dividirse en dos grupos: el primero contiene una variable "x" que se le llamará variable explicativa, que se supone no aleatoria y conocida al observar "y"; el segundo incluye el resto de los factores, cada uno de los cuales influye en la variable respuesta sólo en pequeña magnitud, que se le llama comúnmente perturbación aleatoria. La hipótesis estructural básica del modelo es:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \textbf{(1.22)}$$

Donde:

y_i y ε_i: Son variables aleatorias.

x_i: Es una variable predeterminada con valores conocidos.

β_0 y β_1: Son parámetros desconocidos.

Se establecen las siguientes asunciones:

a. La perturbación tiene esperanza nula, es decir:

$$E(\varepsilon_i) = 0 \qquad \textbf{(1.23)}$$

b. La varianza de la perturbación es siempre constante, y no depende de "x"; lo expresaremos diciendo que la perturbación es homoscedástica:

$$\text{Var}(\varepsilon_i) = \sigma^2 \qquad \textbf{(1.24)}$$

La ecuación (1.24) expresa que la varianza de ε_i es un número positivo constante e igual a σ^2, prácticamente (1.24) representa el supuesto de homoscedasticidad o igual (homos) dispersión (cedasticidad) o igual varianza. Dicho de otra manera, (1.24) quiere decir que las "y" poblacionales que corresponden a varios valores de "x" tienen la misma varianza.

Para examinar el caso opuesto obsérvese la figura 1.9 en la que la varianza condicional de la población "y" aumenta a medida que "x" aumenta igualmente. Esta situación se conoce propiamente con el nombre de heteroscedasticidad o dispersión desigual o varianza desigual, simbólicamente esta situación puede escribirse como:

$$\text{Var}(\varepsilon_i) = \sigma_i^2 \qquad \textbf{(1.25)}$$

Como se ve en la ecuación (1.25) aparece un subíndice, lo cual quiere decir que la varianza de la población ya no es constante.

c. La perturbación ε_i tiene una distribución normal. Esta asunción es consecuencia del Teorema Central de Limite.

d. Las perturbaciones ε_i son independientes entre sí, es decir:

$$E(\varepsilon_i \varepsilon_j) = 0 \qquad i \neq j \qquad \textbf{(1.26)}$$

Estas cuatro ecuaciones pueden expresarse igualmente respecto a la variable respuesta, como sigue:

a. La esperanza de la respuesta depende linealmente de "x". Tomando esperanzas en la ecuación (1.22), como las x_i se suponen no aleatorias:

$$E(y_i) = \beta_0 + \beta_1 x_i \qquad \textbf{(1.27)}$$

El parámetro β_0 representa el valor medio de "y" cuando "x" es cero, β_1 representa el incremento que experimenta la media de "y" cuando "x" aumenta en una unidad.

b. La varianza de la distribución de y_i es constante.

$$\text{Var}(y_i) = \sigma^2 \qquad \textbf{(1.28)}$$

c. La distribución de "y" para cada "x" es normal.

d. Las observaciones y_i son independientes entre si.

Gráficamente, las hipótesis anteriores (excepto la ecuación (1.25) que se muestra en la figura 1.9) indican que, para "x" fija, la distribución de probabilidad de "y" es normal, con varianza constante σ^2 y media que varía linealmente con "x", como indica la figura 1.8.

Figura 1.8 Asunciones del modelo de regresión simple para varianzas iguales.

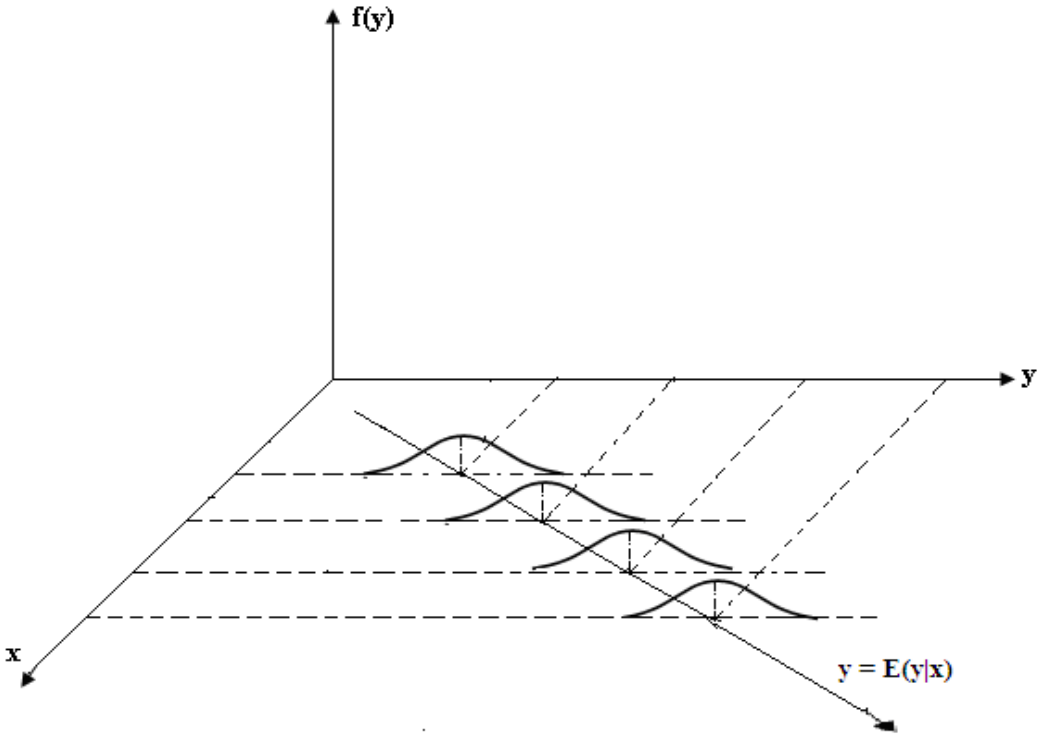
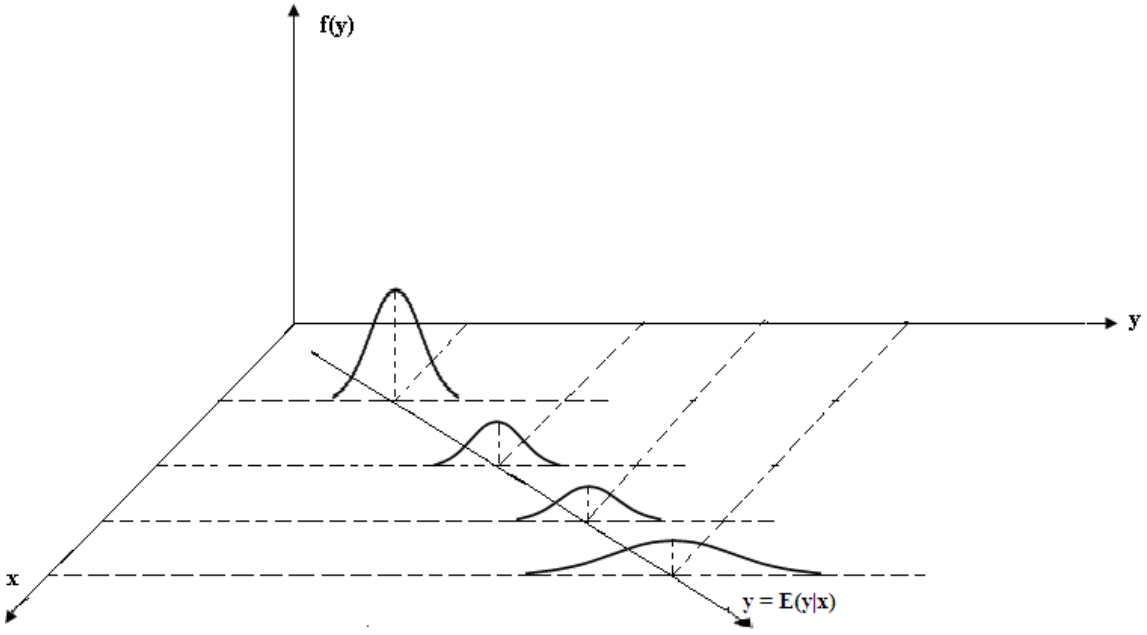


Figura 1.9 Asunciones del modelo de regresión simple para varianzas desiguales.



1.6.1 Comentarios a las Ecuaciones Anteriores.

La suposición principal del modelo es que la media de la distribución de "y", para "x" fija, varía linealmente con "x". Como veremos estas hipótesis deben comprobarse siempre, ya que condicionan toda la construcción del modelo.

La utilidad del modelo lineal $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ radica en que muchas relaciones no lineales pueden convertirse en lineales transformando las variables adecuadamente.

En cualquier caso, conviene tener en cuenta que una relación lineal debe en general considerarse como una aproximación simple, en un rango de valores limitados a una relación más compleja. En consecuencia es necesario tener presente:

a. El rango de los valores dentro del cual vamos a trabajar.

b. El peligro de extrapolar fuera de ese rango.

Las suposiciones de que las perturbaciones tienen media cero, no serán ciertas cuando existan observaciones tomadas en condiciones heterogéneas con el resto. Este hecho puede a veces detectarse mediante un análisis de los residuos del modelo y es importante porque una única observación atípica puede tener gran influencia en la estimación.

La hipótesis de homoscedasticidad no se cumplirá si la variabilidad de cada distribución condicionada depende de la media de dicha distribución: como se observó en el ejemplo de ingresos "x" y gastos "y" que cuando los ingresos son pocos, el gasto es para todos ellos muy pequeño, es decir si se tiene un ingreso promedio de $15 sus gastos son menores o iguales a $18 y existe muy poca variabilidad entre los estudiantes. Sin embargo para ingresos altos hay más variabilidad porque los gastos aumentan.

Ejercicios 1.

1. El departamento de informática de Estadísticos y Censos de El Salvador dedicado a la introducción de datos ha llevado a cabo un programa de formación inicial del personal. La tabla siguiente indica el progreso en pulsaciones por minuto (p.p.m) obtenido en mecanografía de ocho estudiantes que siguieron el programa y el número de semanas que hace que lo siguen:

Individuo	1	2	3	4	5	6	7	8
Nº Semanas x	3	5	2	8	6	9	3	4
Ganancia de velocidad y	87	119	47	195	162	234	72	110

   a) Representar el diagrama de dispersión.

   b) Calcular el Coeficiente de Correlación.

   c) Interpretar si existe relación o no de acuerdo al diagrama y el valor del Coeficiente de Correlación.

2. Se toma una muestra aleatoria de 19 alumnos de la Universidad de El Salvador y se estudian las variables x = número medio de hijos entre sus abuelos maternos y paternos; y = número de hijos de sus padres. Los resultados obtenidos son:

x	6	4	3	4	6.5	2	4.5	3	5	1	2.5	2.5	4.5	3	2.5	5.5	3	2	4
y	4	3	4	4	8	1	4	5	4	2	7	3	4	3	5	8	2	2	6

   a) Construir el diagrama de dispersión.

   b) Calcular el Coeficiente de Correlación.

   c) Interpretar los resultados obtenidos en a) y b).

3. Un comerciante al menudeo de la ciudad de San Miguel llevó a cabo un estudio para determinar la relación que existe entre los gastos "x" ($) de publicidad semanal y las ventas "y" ($). Se obtuvieron los datos siguientes:

x	40	20	25	20	30	50	40	20	50	40	25	50
y	385	400	395	365	475	440	490	420	560	525	480	510

a) Dibujar el diagrama de dispersión.
b) Calcular el Coeficiente de Correlación.
c) Concluir de a cuerdo a los resultados obtenidos en el diagrama y el valor del Coeficiente de Correlación. Es decir si existe o no relación entre las variables gasto en publicidad y ventas.

4. Un psicólogo afirma en base a los datos obtenidos, que a medida que un niño crece, menor es el número de respuestas inadecuadas que da, "x" representa la edad en años, y "y" representa el número de respuestas inadecuadas. Los datos son:

x	2	3	4	4	5	5	6	7	7	9	9	10	11	11	12
y	11	12	10	13	11	9	10	7	12	8	7	3	6	5	5

a) Elaborar el diagrama de dispersión.
b) Determinar la validez de esta conclusión por medio del valor del Coeficiente de Correlación entre las variables "x" y "y".

5. En la tabla siguiente se presenta la información sobre el número de horas de estudio "x" para preparar un examen de Estadística, y la calificación obtenida en dicho examen "y".

x	1	2	2	3	3	3.5	4	4	4.5	4.5	5	5.5	5.5	6
y	4	5	6	6	8	7	8	6	7	8	9	8	9	10

   a) Haga la gráfica (diagrama de dispersión).
   b) Calcule el Coeficiente de Correlación.
   c) Concluya de acuerdo a lo obtenido en a) y b).

6. La Escuela de Biología de la Universidad de El Salvador realizó un estudio biológico de unos peces denominados nariz-negra. Se registraron la longitud "y", en milímetros y la edad "x". Los datos se muestran en la tabla siguiente:

x	0	3	2	2	1	3	2	4	1	1
y	25	80	45	40	36	75	50	95	30	15

   a) Elaborar el diagrama de dispersión para estos datos.
   b) Calcular el Coeficiente de Correlación.
   c) Explicar el significado de las respuestas anteriores.

Apéndice 1: Deducción de Ecuaciones y Propiedades.

1.1 Deducción de Ecuaciones Utilizadas en el Capítulo 1.

a) Deducción de la ecuación (1.4) de la covarianza entre "x" y "y".

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i y_i - \bar{x}y_i - x_i\bar{y} + \bar{x}\bar{y})$$

$$S_{xy} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}\bar{x}y_i - \sum_{i=1}^{n} x_i\bar{y} + \sum_{i=1}^{n}\bar{x}\bar{y}\right)$$

$$S_{xy} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y}\right)$$

Multiplicando y dividiendo por n los dos términos del centro se tiene:

$$S_{xy} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\frac{1}{n}\sum_{i=1}^{n} y_i - n\bar{y}\frac{1}{n}\sum_{i=1}^{n} x_i + n\bar{x}\bar{y}\right)$$

pero $\left[\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \quad y \quad \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}\right]$ y sustituyendo se llega a

$$S_{xy} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{y}\bar{x} + n\bar{x}\bar{y}\right) \qquad \left| -n\bar{y}\bar{x} + n\bar{x}\bar{y} = 0\right.$$

$$S_{xy} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}(n\bar{x}\bar{y})$$

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}$$

L.q.q.d

b) Deducción de la ecuación (1.10) Coeficiente de Correlación r.

$$r = \frac{S_{xy}}{S_x S_y} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$S_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y} \quad , \quad S_x = \sqrt{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \quad \text{y} \quad S_y = \sqrt{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}$$

Sustituyendo las ecuaciones anteriores en **r** y tomando en cuenta que $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Se tiene:

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\quad)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - (\quad)^2}}$$

$$r = \frac{\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2}}$$

$$r = \frac{\frac{\sum_{i=1}^{n} x_i y_i - n\left(\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) * \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\right)}{n}}{\sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)}\sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\left(\frac{1}{n^2}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{\sqrt{\frac{1}{n}}\sqrt{\frac{1}{n}}\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)}\sqrt{\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

$$r = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - n\left(\frac{1}{n^2}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)\right)}{\frac{1}{n}\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)}\sqrt{\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

$$r = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \left(\frac{n}{n^2}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)\right)}{\frac{1}{n}\sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n}}\sqrt{\frac{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}{n}}}$$

$$r = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)\right)}{\frac{1}{n}\frac{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}{\sqrt{n}}\frac{\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}{\sqrt{n}}}$$

$$r = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)\right)}{\frac{1}{n} * \frac{1}{\sqrt{n}\sqrt{n}}\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)\right)}{\frac{1}{n} * \frac{1}{n}\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r = \frac{\frac{1}{n}\left(\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)\right)}{\frac{1}{n^2}\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r = \frac{\frac{\left(\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)\right)}{n}}{\frac{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}{n^2}}$$

$$r = \frac{n^2\left(\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)\right)}{n\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - n\left(\frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\frac{n}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Por lo tanto $$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

L.q.q.d

c) Propiedades de la Esperanza (E).

- El valor esperado de una constante es igual a la constante. Si b es una constante

$$E(b) = b$$

- Si a y b son constantes,

$$E(ax + b) = aE(x) + b$$

Lo cual puede generalizarse así: Si $x_1, x_2, ..., x_N$ son N variables aleatorias y $a_1, a_2, ..., a_N$ y b son constantes, entonces

$$E(a_1x_1 + a_2x_2 + ... + a_Nx_N + b) = a_1E(x_1) + a_2E(x_2) + ... + a_NE(x_N) + b$$

- Si "x" y "y" son dos variables aleatorias independientes, entonces

$$E(xy) = E(x)E(y)$$

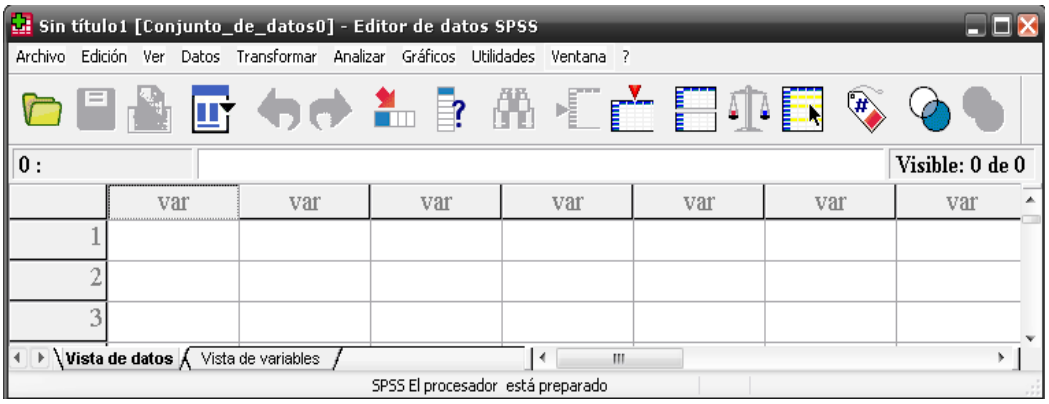
Es decir, la esperanza del producto de xy es igual al producto de las esperanzas individuales de "x" y "y".

Apéndice 1.2: Solución de Ejemplos Haciendo uso del Software Estadístico SPSS v15.0.

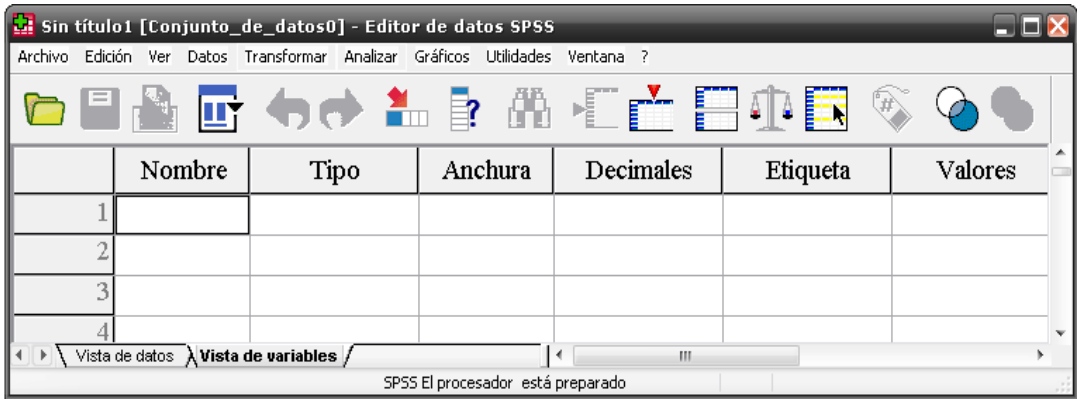
Ejemplo 4: Calcular el coeficiente de correlación entre Estatura "x" y Peso "y" haciendo uso de los datos de la tabla 1.3.

Pasos para la solución de ejemplos con SPSS.

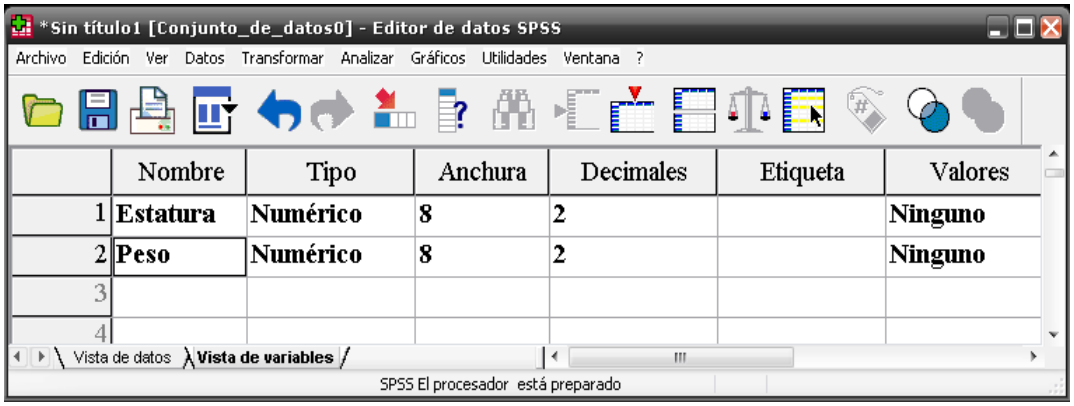
1. Inicie SPSS para Windows. Se presentará el editor de datos como se muestra a continuación:



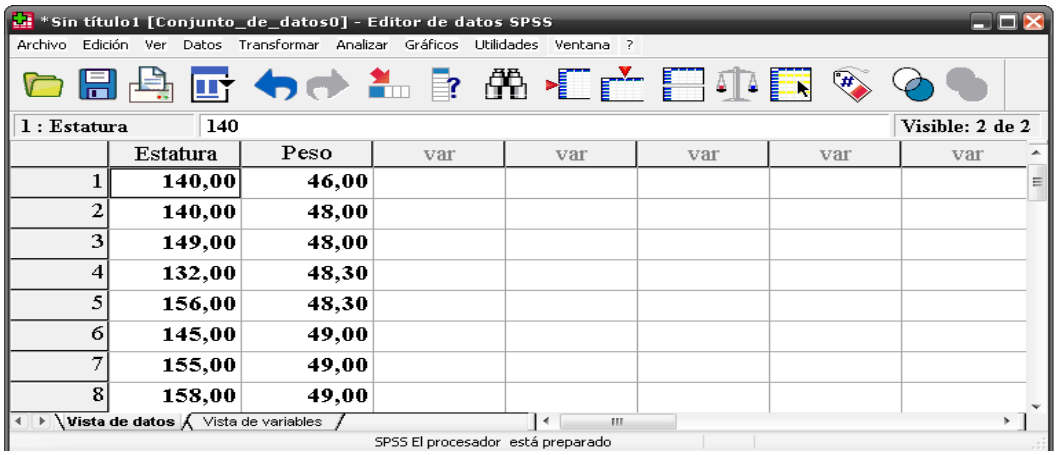
2. Haciendo un click en la pestaña **vista de variable** se obtiene la siguiente ventana.



3. En esta ventana se declaran las variables, es decir, se les coloca un nombre a las variables, para el ejemplo queda de la forma siguiente, en donde Estatura es la variable independiente “x” y Peso es la variable dependiente “y” :

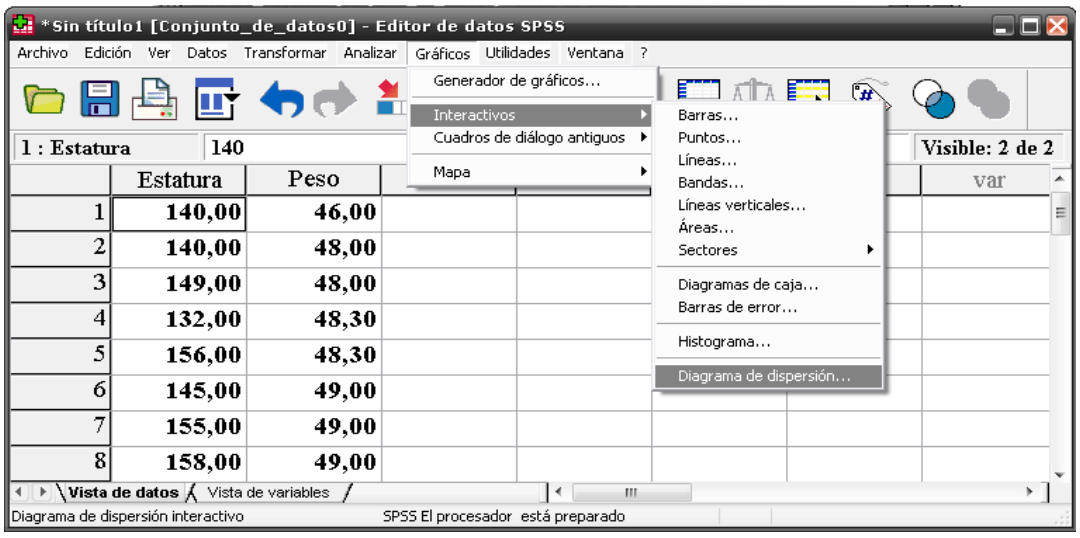


4. Haciendo click en la pestaña **vista de datos** e introduciendo los datos para cada variable se obtiene la ventana siguiente:

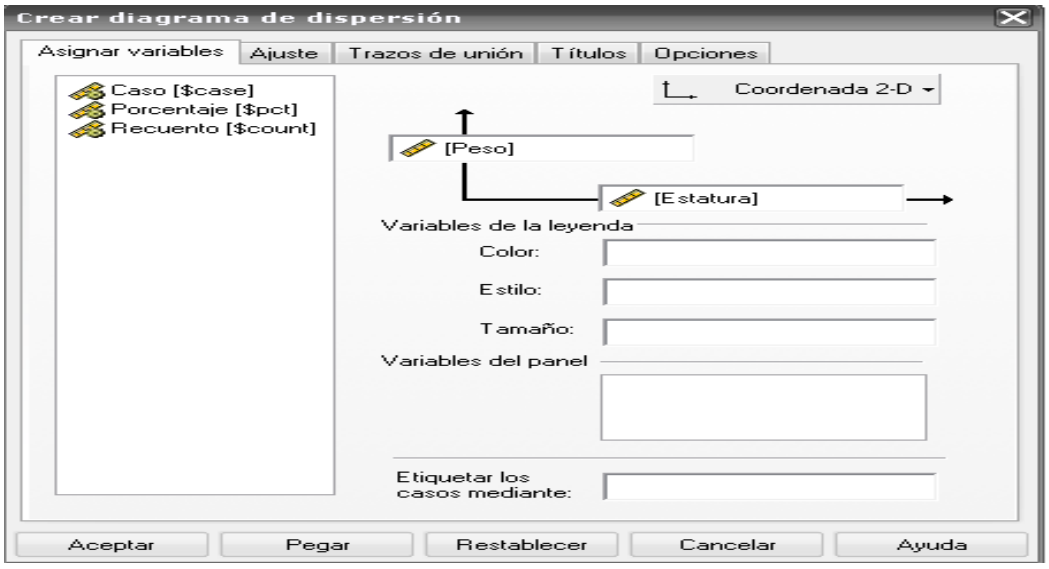


En la que se muestran solamente 8 datos de un total de 40 observaciones.

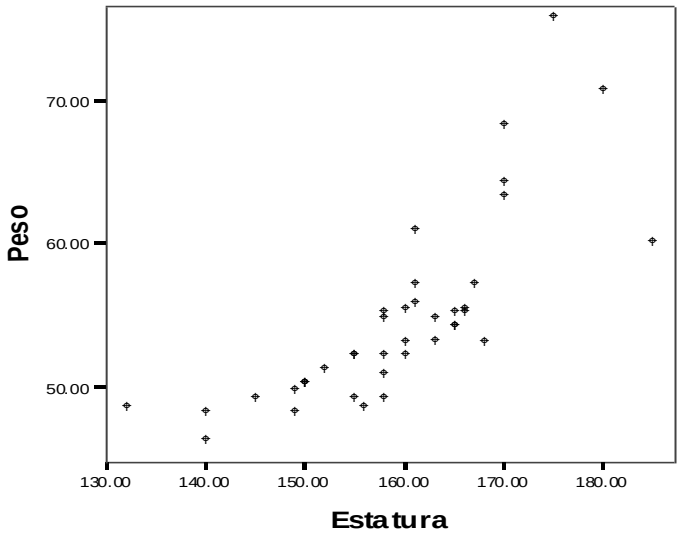
5. Teniendo todos los datos en el editor de datos se pide el diagrama de dispersión el cual se obtiene siguiendo la ruta: Gráficos →Interactivos→ Diagrama de dispersión, como se muestra a continuación.



6. Haciendo click en diagrama de dispersión y colocando cada una de las variables en el eje correspondiente se obtiene lo siguiente: (en donde Estatura va en el eje de las "x" y Peso en el eje de las "y").

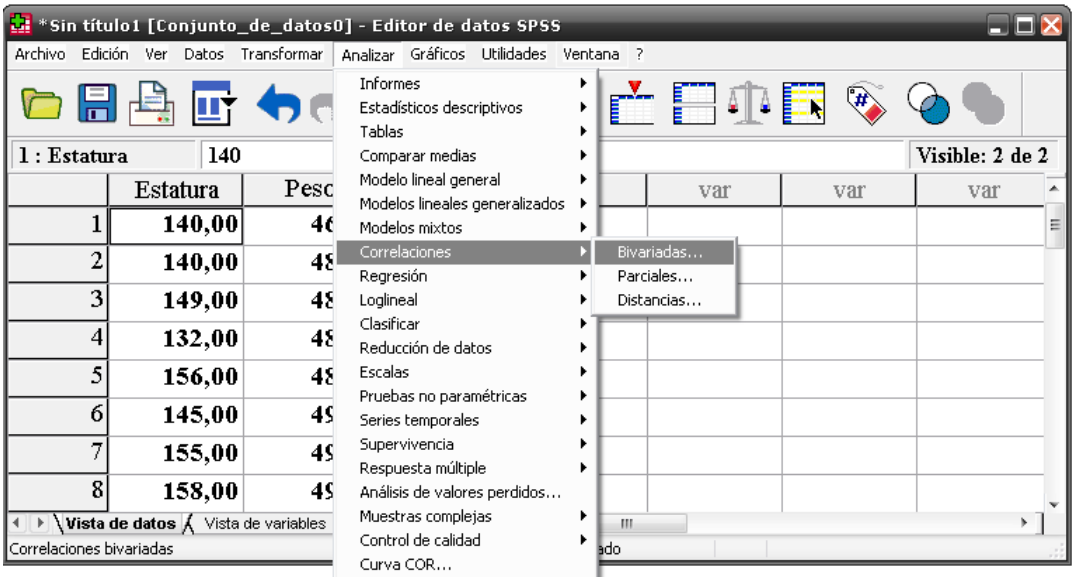


7. Dando click en aceptar se obtiene el diagrama de dispersión siguiente:



De acuerdo con el diagrama de dispersión mostrado en la gráfica se puede decir que existe relación lineal entre las variables Estatura-Peso.

8. Ahora se calculará el coeficiente de correlación **r** o coeficiente de Pearson como sigue: Analizar → Correlaciones→ Bivariadas.



9. Haciendo click en la opción bivariadas se obtiene el cuadro siguiente en el que se trasladan las variables de la izquierda a la derecha haciendo uso de la flecha y

seleccionando el tipo de coeficiente que se desea calcular, en este caso se ha seleccionado Pearson que es la opción que tiene por defecto y la prueba de significación es bilateral:

Y dando click en aceptar se obtiene el resultado siguiente:

Correlaciones

		Estatura	Peso
Estatura	Correlación de Pearson	1	.776
	n	40	40
Peso	Correlación de Pearson	.776	1
	n	40	40

Las correlaciones que se muestran son para la variable “x” con ella misma, y para la variable “y”, por esto el primer valor es 1, n que es igual 40, y el coeficiente de Correlación de Pearson que es 0.776, que es el mismo que se obtuvo en el ejemplo 4 desarrollado en este Capítulo.

Se puede concluir entonces con el diagrama de dispersión y el valor del coeficiente de correlación de Pearson que, existe relación alta entre la variable Peso y Estatura.

La utilización del software reduce el trabajo ya que los diagramas de dispersión y el cálculo del coeficiente se realizan de una forma muy rápida.

Capítulo 2

Estimación y Prueba de Hipótesis.

2.1 Introducción a la Estimación y Prueba de Hipótesis.

La Estimación y la Prueba de Hipótesis constituyen las dos principales ramas de la estadística clásica. La teoría de la estimación consta de dos partes: Estimación puntual y Estimación por intervalo.

En la estimación el principal interés radica en poder estimar la Función de Regresión Poblacional (FRP) con base en la Función de Regresión Muestral (FRM), de la manera más precisa posible. Como se vio en el Capítulo 1 en el Modelo de Regresión Lineal Simple hay tres parámetros que se deben estimar: Los coeficientes de la recta de regresión, β_0 y β_1; y la varianza de la distribución normal, σ^2.

En la actualidad el cálculo de los estimadores de los parámetros para construir la FRM se realiza por los siguientes métodos:

- $\rightarrow$ Mínimos Cuadrados Ordinarios (MCO).
- $\rightarrow$ Máxima Verosimilitud (MV).

Pero en lo concerniente al análisis de regresión, el método más usado es el de los Mínimos Cuadrados Ordinarios. En el presente Capítulo se tratan los dos métodos en términos del modelo de regresión con dos variables, pero se hace más énfasis en el MCO. Además se trata la estimación por intervalo la cual está relacionada con la prueba de hipótesis.

2.2 Definición de Términos Básicos.

Análisis de Varianza (ANOVA): Técnica estadística utilizada para probar la igualdad de tres o más medias de muestra y, de este modo, hacer inferencias sobre si las muestras provienen de poblaciones que tienen la misma media.

Coeficiente de Determinación: Medida de la proporción de la variable dependiente, que es explicada por la línea de regresión, esto es, por la relación de "y" con la variable independiente "x".

Estimación: Valor específico observado de un estimador.

Estimación por Intervalo: Estimación del parámetro de la población indicando un valor máximo y un valor mínimo dentro del cual se encuentra el parámetro poblacional.

Estimación Puntual: Estimación del parámetro de la población calculado con la información de la muestra.

Estimador Insesgado: Estimador cuyo valor esperado es el parámetro de la población.

Estimador Eficiente: Estimador con un menor error estándar que algún otro estimador del parámetro de la población, esto es, cuando más pequeño sea el error estándar de un estimador, más eficiente será ese estimador.

Estimador Consistente: Estadístico que se aproxima al parámetro de la población a medida que aumenta el tamaño de la muestra.

es: Error estándar o desviación típica.

Hipótesis: Enunciado o proposición no probados acerca de un factor o fenómeno de interés para el investigador. Una hipótesis estadística es un enunciado respecto a una

población y usualmente es un enunciado respecto a uno o más parámetros de la población.

Hipótesis Alternativa: Afirmación de que se espera alguna diferencia o efecto. La aceptación de la hipótesis alternativa dará lugar a cambios en las opiniones o acciones.

Hipótesis Nula: Afirmación en la cual no se espera ninguna diferencia ni efecto. Si la hipótesis nula no se rechaza, no se hará ningún cambio.

Intervalo de Confianza: Intervalo de valores que tiene designada una probabilidad de que incluya el valor real del parámetro de la población.

Prueba de Hipótesis: Procedimiento a través del cual se rechaza o no la hipótesis nula.

SS_{Res}: Suma de cuadrados residuales, o suma de cuadrados de error.

SS_T: Suma total de cuadrados, o suma corregida de cuadrados de las observaciones.

SS_R: Suma de cuadrados de regresión, o suma de cuadrados del modelo.

Varianza: Desviación cuadrada media de todos los valores de la media.

2.3 Estimación de los Parámetros por el Método de Mínimos Cuadrados Ordinarios (MCO).

La función de regresión poblacional no es observable directamente; es preciso estimarla a partir de la FRM, motivo por el cual se explica a continuación como se determina la FRM. Recordando la FRM lineal con dos variables, se puede escribir:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad \textbf{(2.1)}$$

$$y_i = \hat{y}_i + e_i \qquad \textbf{(2.2)}$$

Donde $\hat{y}_i$ es el valor estimado (media condicional) de y_i. También la ecuación (2.2) puede expresarse como:

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ e_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \end{aligned} \qquad \textbf{(2.3)}$$

Lo que muestra que los e_i (los residuos) son simplemente las diferencias entre los valores verdaderos y los estimados de "y".

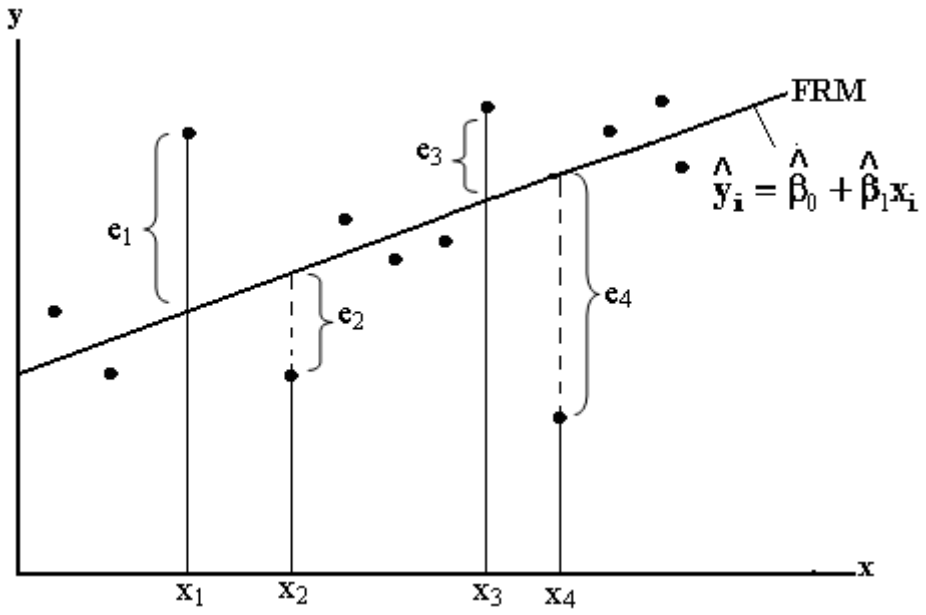
Los parámetros β_0 y β_1 son desconocidos, y se deben estimar con los datos de la muestra. Supongamos que hay n pares de datos: (y_1, x_1), (y_2, x_2),…, (y_n, x_n) estos datos pueden obtenerse en un experimento controlado, diseñado en forma específica para recolectarlos, en un estudio mediante la observación, o a partir de registros históricos existentes.

Estamos interesados en determinar la FRM de forma tal que esté tan cerca como sea posible al "y" real. Con este fin se puede adoptar el siguiente criterio:

Elegir la FRM de manera tal que la suma de los residuos $\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)$ sea tan pequeña como sea posible.

Aunque intuitivamente este criterio parece atractivo, no es necesariamente un buen criterio como se muestra en la figura 2.1.

Figura 2.1 Criterio de Mínimos Cuadrados Ordinarios.



Si se adopta el criterio de minimizar $\sum_{i=1}^{n} e_i$, se observa en la figura 2.1 cómo los residuos e_2 y e_3 así como los residuos e_1 y e_4 reciben la misma ponderación en la suma ($e_1 + e_2 + e_3 + e_4$) aunque los dos primeros estén mucho más cerca de la FRM que los dos últimos. En otras palabras, todos los residuos tienen igual relevancia sin que importe qué tan cerca o qué tan dispersas estén las observaciones originales de la FRM. Como consecuencia, la suma algebraica de los e_i puede ser pequeña (a un cero o igual a cero)

aunque los e_i estén muy dispersos alrededor de la FRM. Para verificarlo supongamos que e_1, e_2, e_3 y e_4 tienen valores 10, -2, 2 y -10 respectivamente; la suma algebraica de estos residuos es cero, aunque e_1 y e_4 estén más dispersos alrededor de la FRM que e_2 y e_3. Este problema puede evitarse adoptando el criterio de Mínimos Cuadrados según el cual la FRM puede establecerse en forma tal que:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \qquad \textbf{(2.4)}$$

Sea tan pequeña como sea posible y donde e_i^2 representa los residuos al cuadrado. Elevando al cuadrado los residuos e_i, este método destaca mejor los residuos e_1 y e_4 que los residuos e_2 y e_3. Como ya se vio, bajo el criterio de minimizar $\sum_{i=1}^{n} e_i$ la suma puede ser pequeña con los e_i bien dispersos alrededor de la FRM, situación que no puede representarse con el Método de los Mínimos Cuadrados, por cuanto entre más grandes sean los e_i (en valor absoluto) más grande será la $\sum_{i=1}^{n} e_i^2$. Una justificación adicional para el Método de los Mínimos Cuadrados es la de que los estimadores obtenidos por este método tienen propiedades muy deseables desde el punto de vista estadístico.

De la ecuación (2.4) se puede deducir que:

$$\sum_{i=1}^{n} e_i^2 = f(\hat{\beta}_0, \hat{\beta}_1) \qquad \textbf{(2.5)}$$

O sea que la suma de los residuos al cuadrado es una función de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$. Para un conjunto dado de datos con diferentes valores $\hat{\beta}_0$ y $\hat{\beta}_1$ se obtendrán diferentes e_i y por lo tanto diferentes valores de $\sum_{i=1}^{n} e_i^2$. El principio de Mínimos Cuadrados escoge $\hat{\beta}_0$ y $\hat{\beta}_1$ en forma tal que para una muestra dada la $\sum_{i=1}^{n} e_i^2$ resulte tan pequeña como sea posible.

2.3.1 Estimación de β_0 y β_1.

Para estimar β_0 y β_1 se usa el Método de Mínimos Cuadrados. Esto es, se estiman β_0 y β_1 tales que la suma de los cuadrados de la diferencia entre las observaciones y_i y la línea recta sea mínima según la ecuación:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \textbf{(2.6)}$$

La ecuación (2.6) se puede escribir como:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i \qquad i = 1, 2, \ldots, n \qquad \textbf{(2.7)}$$

Se puede considerar que la ecuación (2.6) es un Modelo de Regresión Poblacional, mientras que la ecuación (2.7) es un Modelo de Regresión Muestral, escrito en términos de los n pares de datos (y_i, x_i) (i = 1, 2,…, n). Así, el criterio de Mínimos Cuadrados es:

$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \qquad \textbf{(2.8)}$$

Los estimadores por Mínimos Cuadrados de $\hat{\beta}_0$ y $\hat{\beta}_1$ deben satisfacer:

$$\frac{\partial S}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \qquad \textbf{(2.9)}$$

y

$$\frac{\partial S}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0 \qquad \textbf{(2.10)}$$

Simplificando estas dos ecuaciones se obtiene:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad \textbf{(2.11)}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \qquad \textbf{(2.12)}$$

Las ecuaciones anteriores se conocen como ecuaciones normales de Mínimos Cuadrados y al resolverlas se obtiene:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \textbf{(2.13)}$$

En donde $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ y $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, son las medias muéstrales de “x” y “y”.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad \textbf{(2.14)}$$

La deducción de las ecuaciones (2.13) y (2.14) se muestra en el apéndice **2.1 a) y 2.1 b).** Una forma alternativa de calcular $\hat{\beta}_1$ es:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Por consiguiente, $\hat{\beta}_0$ y $\hat{\beta}_1$ en las ecuaciones (2.13) y (2.14) son los estimadores por Mínimos Cuadrados de la ordenada al origen y la pendiente, respectivamente. El modelo ajustado de Regresión Lineal Simple es:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad \textbf{(2.15)}$$

La ecuación (2.15) produce un estimador puntual, de la media de "y", para una determinada "x". Como el denominador de la ecuación (2.14) es la suma corregida de cuadrados de la x_i, y, el numerador es la suma corregida de los productos cruzados (covarianza) de x_i y y_i, estas ecuaciones se pueden escribir en una forma más compacta como sigue:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \qquad \textbf{(2.16)}$$
$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

y

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \qquad \textbf{(2.17)}$$
$$S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

Entonces, una forma cómoda de escribir la ecuación (2.14) es:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \textbf{(2.18)}$$

La diferencia entre el valor observado y_i y el valor ajustado correspondiente $\hat{y}_i$ se llama residuo o residual. Matemáticamente el i-ésimo residual es:

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, 2, \ldots, n \qquad \textbf{(2.19)}$$

Los residuales tienen un papel importante para investigar la adecuación del modelo de regresión ajustado, y para detectar diferencias respecto a los supuestos básicos.

Los estimadores previamente obtenidos, se conocen como estimadores de Mínimos Cuadrados, por derivarse del principio de los Mínimos Cuadrados. Obsérvese a continuación las características de estos estimadores.

1. Están expresados únicamente en términos de cantidades observables (de “y” y “x”).
2. Son estimadores puntuales; es decir, que dada la muestra, cada estimador proporcionará un solo (punto) valor del parámetro poblacional relevante.

Una vez obtenidos los estimadores de los Mínimos Cuadrados a partir de los datos que se tengan es muy fácil ajustar la Línea de Regresión Muestral (figura 2.2).

2.3.2 Propiedades de los Estimadores de Mínimos Cuadrados y el Modelo de Regresión Ajustado.

Los estimadores por Mínimos Cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen algunas propiedades importantes.

Primero, obsérvese que, según las ecuaciones (2.13) y (2.14), $\hat{\beta}_0$ y $\hat{\beta}_1$ son combinaciones lineales de las observaciones y_i. Por ejemplo $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i$

Donde $c_i = \frac{x_i - \bar{x}}{S_{xx}}$, para i = 1, 2,…, n.

Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ por Mínimos Cuadrados Ordinarios son estimadores insesgados de los parámetros β_0 y β_1 del modelo. Para demostrarlo con $\hat{\beta}_1$, considérese

La esperanza o valor medio de $\hat{\beta}_1$.

$$E(\hat{\beta}_1) = E(\sum_{i=1}^{n} c_i y_i) = \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

Ya que se supuso que, $E(\varepsilon_i) = 0$. Ahora se puede deducir en forma directa que $\sum_{i=1}^{n} c_i = 0$

y que $\sum_{i=1}^{n} c_i x_i = 1$, y entonces

$$E(\hat{\beta}_1) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_0 (0) + \beta_1 (1) = \beta_1$$

$$E(\hat{\beta}_1) = \beta_1 \qquad \textbf{(2.20)}$$

La deducción completa de este resultado se muestra en el apéndice **2.1 c)**.

Esto es, si se supone que el modelo es correcto [que $E(y_i) = \beta_0 + \beta_1 x_i$], entonces $\hat{\beta}_1$ es un estimador insesgado de β_1 de igual forma se puede deducir que $\hat{\beta}_0$ es un estimador insesgado de β_0, es decir,

La esperanza o valor medio de $\hat{\beta}_0$.

$$E(\hat{\beta}_0) = \beta_0 \qquad \textbf{(2.21)}$$

La deducción de este resultado se muestra en el apéndice **2.1 d)**.

La varianza de $\hat{\beta}_1$.

$$\begin{aligned} \text{var}(\hat{\beta}_1) &= \text{var}\left(\sum_{i=1}^{n} c_i y_i \right) \\ \text{var}(\hat{\beta}_1) &= \sum_{i=1}^{n} c_i^2 \, \text{var}(y_i) \end{aligned} \qquad \textbf{(2.22)}$$

Ya que las observaciones y_i son no correlacionadas, por lo que la varianza de la suma es igual a la suma de las varianzas. La varianza de cada término en la suma es $c_i^2 \, \text{var}(y_i)$ y en la ecuación (1.28) Capítulo 1 se hizo el supuesto que $\text{Var}(y_i) = \sigma^2$; en consecuencia,

$$\text{var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} c_i^2$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{S_{xx}^2}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2 S_{xx}}{S_{xx}^2}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \qquad \textbf{(2.23)}$$

y el error estándar de $\hat{\beta}_1$ está dado por: $es(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} = \sqrt{\frac{\sigma^2}{S_{xx}}} = \frac{\sigma}{\sqrt{S_{xx}}}$

La varianza de $\hat{\beta}_0$ es

$$\begin{aligned} var(\hat{\beta}_0) &= var(\bar{y} - \hat{\beta}_1 \bar{x}) \\ var(\hat{\beta}_0) &= var(\bar{y}) + var((-\bar{x})\hat{\beta}_1) \\ var(\hat{\beta}_0) &= var(\bar{y}) + (-\bar{x})^2 \, var(\hat{\beta}_1) \\ var(\hat{\beta}_0) &= \frac{\sigma^2}{n} + \bar{x}^2 \, var(\hat{\beta}_1) \\ var(\hat{\beta}_0) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) \end{aligned} \qquad \textbf{(2.24)}$$

Y el error estándar $\hat{\beta}_0$ está dado por:

$$es(\hat{\beta}_0) = \sqrt{var(\hat{\beta}_0)} = \sqrt{\sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)} = \sigma \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

var = varianza y σ^2 es la constante o varianza homoscedástica (ecuación 1.24 Capítulo 1) y se puede estimar como se muestra en la sección 2.4.

La deducción de las ecuaciones (2.23) y (2.24) se muestra en el apéndice **2.1 e) y f).**

Otro resultado importante a cerca de la calidad de los estimadores por Mínimos Cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ es el **Teorema de Gauss –Markov**, que establece que para el modelo de regresión (ecuación (1.2) del Capítulo 1) con las hipótesis $E(\varepsilon) = 0$, $var(\varepsilon) = \sigma^2$ y con errores no correlacionados, los estimadores por Mínimos Cuadrados Ordinarios son insesgados y tienen varianza mínima en comparación con todos los demás estimadores insesgados que sean combinaciones lineales de las y_i. Con frecuencia

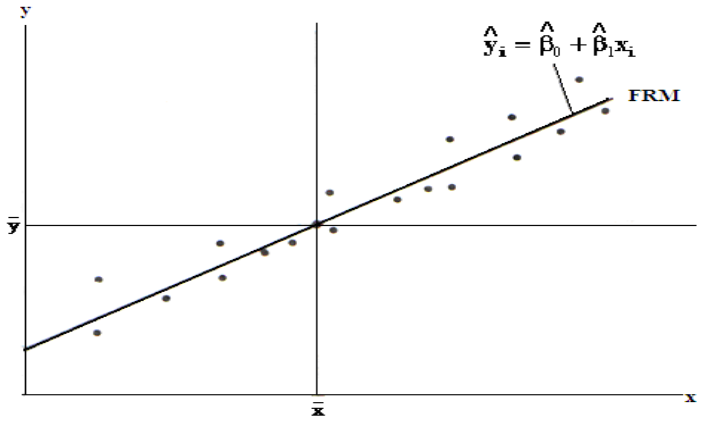
se dice que los estimadores por Mínimos Cuadrados son los **Estimadores Lineales Insesgados Óptimos,** donde "óptimos" implica que son de varianza mínima.

En el apéndice **2.1 g)** se demuestra el teorema de Gauss-Markov.

Hay otras propiedades útiles del ajuste por Mínimos Cuadrados que se muestran a continuación:

1. La línea de Regresión Muestral (figura 2.2) pasa a través de la media muestral de "x" y "y". Esto se puede ver partiendo de (2.13) puesto que ésta puede reescribirse como $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$, como se observa en la figura 2.2.

Figura 2.2 Diagrama que muestra como la línea de regresión muestral pasa a través de los valores de las medias muéstrales de "y" y "x".



2. El valor medio de "y" estimado ($\hat{y}_i$) es igual al valor medio del "y" observado debido a que:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) \quad \textbf{(2.25)}$$

Sumando a ambos lados, en la última igualdad, sobre los valores muéstrales y dividiendo por el tamaño de la muestra n se obtiene:

$$\frac{\sum_{i=1}^{n}\hat{y}_i}{n}=\frac{\sum_{i=1}^{n}\bar{y}_i}{n}+\frac{\sum_{i=1}^{n}\hat{\beta}_1(x_i-\bar{x})}{n}$$

$$\bar{\hat{y}}=\frac{n\bar{y}}{n}+\frac{\sum_{i=1}^{n}\hat{\beta}_1 x_i}{n}+\frac{\sum_{i=1}^{n}\hat{\beta}_1\bar{x}}{n}$$

$$\bar{\hat{y}}=\bar{y}+\hat{\beta}_1\frac{\sum_{i=1}^{n}x_i}{n}+\frac{\hat{\beta}_1 n\bar{x}}{n}$$

$$\bar{\hat{y}}=\bar{y}+\hat{\beta}\bar{x}-\hat{\beta}\bar{x}$$

$$\bar{\hat{y}}=\bar{y} \qquad \textbf{(2.26)}$$

L.q.q.d

3. El valor medio de los residuos e_i es cero del apéndice 2.1 a) la primer ecuación es: $-2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)=0$ pero dado que $e_i=y_i-\hat{\beta}_0-\hat{\beta}_1x_i$, la anterior ecuación se reduce a $-2\sum_{i=1}^{n}e_i=0$ donde $\bar{e}=0$ como resultado de la propiedad anterior, la Regresión Muestral es:

$$y_i=\hat{\beta}_0+\hat{\beta}_1x_i+e_i \qquad \textbf{(2.27)}$$

4. Los residuos e_i no están correlacionados con el valor predicho de y_i, lo cual se puede verificar como sigue:

$$\sum_{i=1}^{n}\hat{y}_i e_i=0 \qquad \textbf{(2.28)}$$

5. Los residuos e_i no están correlacionados con x_i esto es $\sum_{i=1}^{n}x_i e_i=0$.

2.4 Estimación de σ^2.

Además de estimar $\hat{\beta}_0$ y $\hat{\beta}_1$, se requiere un estimador de σ^2 para probar hipótesis y formar estimados de intervalos pertinentes al modelo de regresión. En el caso ideal este estimado no debería depender de la adecuación del modelo ajustado, eso sólo es posible cuando hay varias observaciones de "y", para al menos un valor de "x" o cuando se dispone de información anterior acerca de σ^2. Cuando no se puede usar este método, el estimador de σ^2 se obtiene de la suma de cuadrados residuales, o suma de cuadrados de error:

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \textbf{(2.29)}$$

Se puede deducir una fórmula cómoda para calcular SS_{Res} sustituyendo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ en la ecuación (2.29), y simplificando se llega a:

$$SS_{Res} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy} \qquad \textbf{(2.30)}$$

Pero

$$\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 = S_{yy} = SS_T$$

Es justo la suma de cuadrados corregida, de las observaciones de la respuesta, por lo que

$$SS_{Res} = S_{yy} - \hat{\beta}_1 S_{xy} = SS_T - \hat{\beta}_1 S_{xy} \qquad \textbf{(2.31)}$$

La deducción de (2.31) se presenta en el apéndice **2.1 h).**

La suma de cuadrados residuales tiene n-2 grados de libertad, porque dos grados de libertad se asocian con los estimados $\hat{\beta}_0$ y $\hat{\beta}_1$ que se usan para obtener $\hat{y}_i$.

En el apéndice **2.1 i)** se demuestra que el valor esperado de SS_{Res} es $E(SS_{Res}) = (n-2)\sigma^2$

Por lo que un estimador insesgado de σ^2 es:

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = MS_{Res} \qquad \textbf{(2.32)}$$

La cantidad MS_{Res} se llama cuadrado medio residual.

La raíz cuadrada de $\hat{\sigma}^2$ ($es(\hat{\sigma}^2) = \sqrt{\hat{\sigma}^2} = \sqrt{MS_{Res}}$) se llama, error estándar de la regresión y tiene las mismas unidades que la variable de respuesta "y".

Ya que $\hat{\sigma}^2$ depende de la suma de cuadrados residuales, cualquier violación de los supuestos sobre los errores del modelo, o cualquier especificación equivocada de la forma del modelo pueden dañar gravemente la utilidad de $\hat{\sigma}^2$ como estimador de σ^2.

Como $\hat{\sigma}^2$ se calcula con los residuales del modelo de regresión, se dice que es un estimador de σ^2 dependiente del modelo.

2.5 Coeficiente de Determinación r^2: Medida de la Bondad del Ajuste.

Hasta el momento, nos hemos referido al problema de la estimación de los coeficientes de regresión, a sus errores estándar y algunas de sus propiedades.

Consideraremos ahora la bondad del ajuste de la línea de regresión ajustada al conjunto de datos, es decir, se trata de encontrar en qué medida se ajusta la línea de regresión muestral a los datos. De la figura 2.1 se desprende claramente que si todas las observaciones coincidieran con la línea de regresión, obtendríamos un ajuste "perfecto", lo que raras veces ocurre. Generalmente tienden a haber algunos e_i positivos y otros negativos, con la esperanza de que los residuos localizados alrededor de la línea de regresión sean lo más pequeños posible. Ahora bien, el coeficiente de determinación r^2 (caso de dos variables) o R^2 (regresión múltiple) es una medida de resumen que nos dice qué tan exactamente la línea de regresión muestral se ajusta a los datos, y se denota de la forma siguiente:

$$r^2 = \frac{S_{xy}^2}{S_x^2 S_y^2} = \frac{S_{xy}^2}{S_{xx} S_{yy}} \qquad \textbf{(2.33)}$$

Donde:

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

La cantidad definida como r^2 se conoce como el coeficiente de determinación (muestral) y es ampliamente utilizado como una medida de la bondad del ajuste de una

línea de regresión. Es decir, el r^2 mide la proporción o porcentaje de la variación total en "y" explicada por el modelo de regresión. Sus propiedades más importantes son:

1. Es una cantidad no negativa.
2. Sus limites son $0 \leq r^2 \leq 1$. Un r^2 de 1 quiere decir ajuste perfecto, mientras que un r^2 de 0 quiere decir que no hay relación entre la variable dependiente y las variables explicatorias.

Aunque el r^2 puede calcularse directamente a partir de la ecuación (2.33) se puede obtener más rápidamente haciendo uso de la siguiente ecuación:

$$r^2 = \hat{\beta}_1^2 \left(\frac{S_x^2}{S_y^2} \right) = \hat{\beta}_1^2 \left(\frac{S_{xx}}{S_{yy}} \right) \qquad \textbf{(2.34)}$$

Donde S_{xx} y S_{yy} son las varianzas muéstrales de "x" y "y" respectivamente.

Una cantidad muy relacionada con el r^2 pero conceptualmente diferente, es el coeficiente de correlación, que como se vio en el Capítulo 1 es una medida del grado de asociación entre dos variables. Puede calcularse bien como:

$$r = \pm\sqrt{r^2} \qquad \textbf{(2.35)}$$

O a partir de su definición dada en la ecuación (1.10) del Capítulo 1. El r puede tomar dos valores un positivo y un negativo, se tomará el positivo cuando la pendiente de la ecuación de regresión sea positiva y el negativo en el caso contrario.

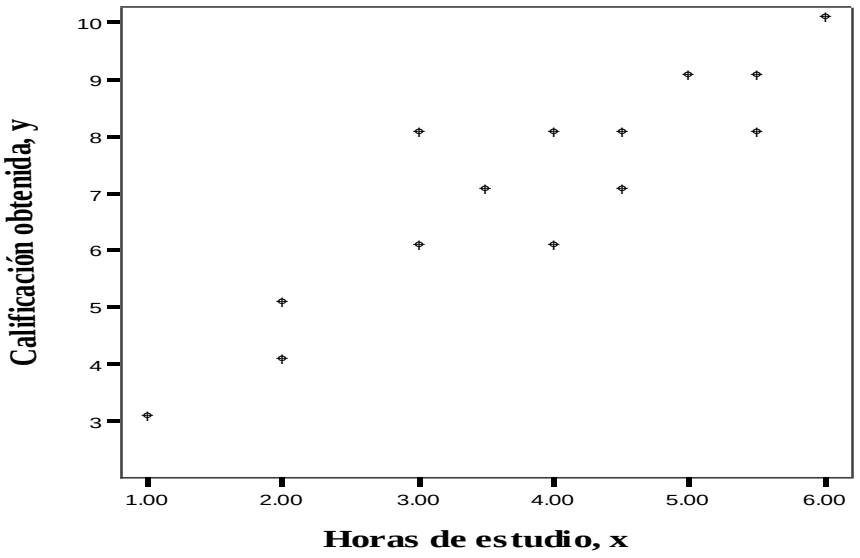
Ejemplo 1: A continuación se presenta información de 14 estudiantes sobre el número de Horas de estudio "x" para preparar un examen de Estadística, y la Calificación obtenida en dicho examen "y".

Tabla 2.1 Observaciones de 14 estudiantes.

x	1	2	2	3	3	3.5	4	4	4.5	4.5	5	5.5	5.5	6
y	3	4	5	6	8	7	8	6	7	8	9	8	9	10

Solución:

Figura 2.3 Diagrama de dispersión para las Horas de estudio vs. Calificación.



El diagrama de dispersión figura 2.3 nos muestra que la relación entre las dos variables (Horas de estudio y Calificación obtenida) es lineal con pendiente positiva, de manera que cuantas más horas dedique a estudiar mayor es la calificación obtenida en el examen. Por tanto, tiene sentido buscar la recta de regresión.

Se calculará la recta de regresión haciendo uso de ecuaciones y propiedades expuestas anteriormente.

Tabla 2.2 Resultados basados en la tabla 2.1

n	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	3	1	9	3
2	2	4	4	16	8
3	2	5	4	25	10
4	3	6	9	36	18
5	3	8	9	64	24
6	3.5	7	12.25	49	24.5
7	4	8	16	64	32
8	4	6	16	36	24
9	4.5	7	20.25	49	31.5
10	4.5	8	20.25	64	36
11	5	9	25	81	45
12	5.5	8	30.25	64	44
13	5.5	9	30.25	81	49.5
14	6	10	36	100	60
sumas	$\sum_{i=1}^{n} x_i = 53.5$	$\sum_{i=1}^{n} y_i = 98$	$\sum_{i=1}^{n} x_i^2 = 233.25$	$\sum_{i=1}^{n} y_i^2 = 738$	$\sum_{i=1}^{n} x_i y_i = 409.5$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{14}(98) = 7 \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{14}(53.5) = 3.8214$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$\hat{\beta}_1 = \frac{409.5 - \frac{(53.5)(98)}{14}}{233.25 - \frac{(53.5)^2}{14}}$$

$$\hat{\beta}_1 = \frac{35}{28.803}$$

$$\hat{\beta}_1 = 1.215$$

y

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_0 = 7 - 1.215(3.8214)$$
$$\hat{\beta}_0 = 2.356$$

Por tanto la ecuación de regresión muestral es:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 2.356 + 1.215 x_i \qquad \textbf{(2.36)}$$

Calificación $= 2.356 + 1.215$ Horas de estudio

Tabla 2.3 Resultados basados en la tabla 2.2

$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	e_i^2
3.571	-0.571	0.326
4.786	-0.786	0.618
4.786	0.214	0.046
6.001	-0.001	0.000
6.001	1.999	3.996
6.6085	0.392	0.153
7.216	0.784	0.615
7.216	-1.216	1.479
7.8235	-0.824	0.678
7.8235	0.177	0.031
8.431	0.569	0.324
9.0385	-1.039	1.078
9.0385	-0.038	0.001
9.646	0.354	0.125
$\sum_{i=1}^{n} \hat{y}_i = 97.9865$	$\sum_{i=1}^{n} e_i = 0.013$	$\sum_{i=1}^{n} e_i^2 = 9.471$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$
$$\hat{\sigma}^2 = \frac{9.471}{14-2}$$
$$\hat{\sigma}^2 = \frac{9.471}{12}$$
$$\hat{\sigma}^2 = 0.789$$

$\hat{\sigma}^2 = 0.789$ es un estimador de σ^2 y $S_{xx} = 28.803$ denominador de $\hat{\beta}_1$, con estos datos calculamos:

La varianza de $\hat{\beta}_1$ es:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$
$$\text{var}(\hat{\beta}_1) = \frac{0.789}{28.803}$$
$$\text{var}(\hat{\beta}_1) = 0.027393$$

El error estándar de $\hat{\beta}_1$ es:

$$\text{es}(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)}$$
$$\text{es}(\hat{\beta}_1) = \sqrt{0.027393}$$
$$\text{es}(\hat{\beta}_1) = 0.1655$$
$$\text{es}(\hat{\beta}_1) \approx 0.166$$

La varianza de $\hat{\beta}_0$ es:

$$\text{var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$
$$\text{var}(\hat{\beta}_0) = 0.789\left(\frac{1}{14} + \frac{(3.8214)^2}{28.803}\right)$$
$$\text{var}(\hat{\beta}_0) = 0.789(0.5784)$$
$$\text{var}(\hat{\beta}_0) = 0.456$$

El error estándar de $\hat{\beta}_0$ es:

$$\text{es}(\hat{\beta}_0) = \sqrt{\text{var}(\hat{\beta}_0)}$$
$$\text{es}(\hat{\beta}_0) = \sqrt{0.456}$$
$$\text{es}(\hat{\beta}_0) = 0.675$$

Con los datos obtenidos anteriormente calculamos el valor de r^2 y el valor de r así.

$$\hat{\beta}_1 = 1.215\ ,\ S_{xx} = 28.803$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$S_{yy} = \sum_{i=}^{n} y_i^2 - n\bar{y}^2$$

$$S_{yy} = 738 - 14(7)^2$$

$$S_{yy} = 52$$

$$r^2 = \hat{\beta}_1^2\left(\frac{S_{xx}}{S_{yy}}\right)$$

$$r^2 = (1.215)^2\left(\frac{28.803}{52}\right)$$

$$r^2 = (1.476)(0.554)$$

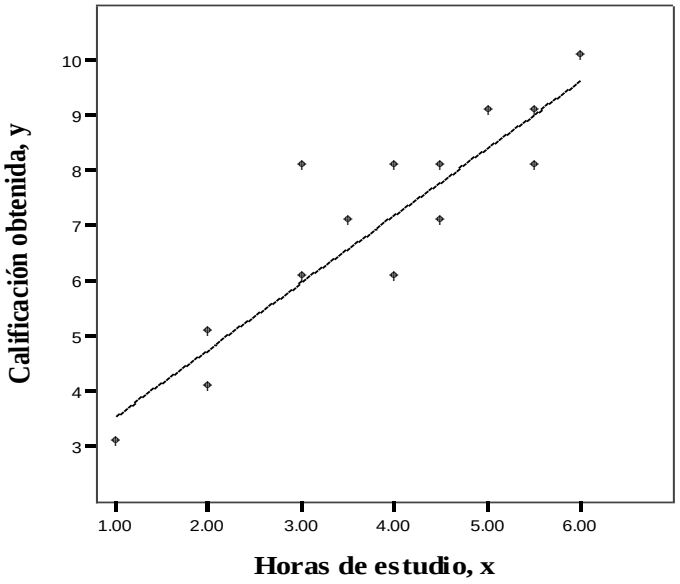
$$r^2 = 0.818$$

y

$$r = \pm\sqrt{r^2} = \pm\sqrt{0.818} = \pm 0.904$$

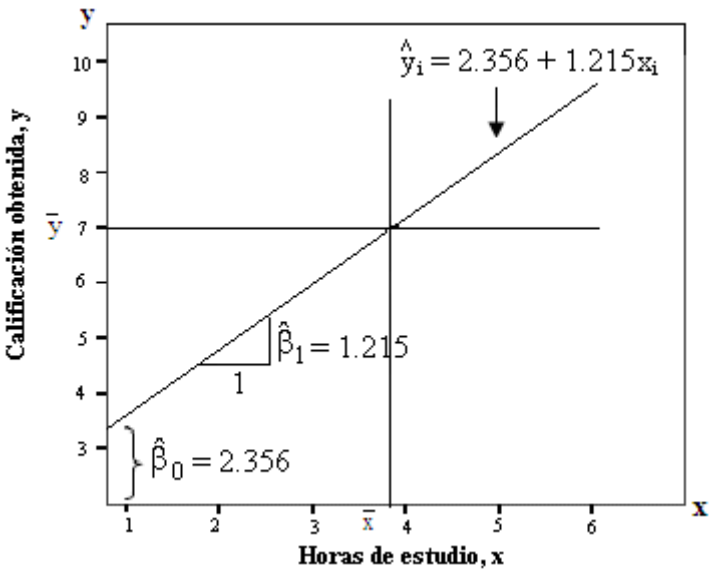
Se puede observar que existen dos valores para r, + 0.904 y - 0.904 para este ejemplo tomaremos el valor positivo r = 0.904, debido a que la relación que existe entre las variables es directamente proporcional, es decir, que a medida que crece la variable "x" también lo hace la variable "y", en la siguiente figura se puede observar que la pendiente es positiva.

Figura 2.4 Recta de regresión para los datos de la tabla 2.1.



La figura 2.4 muestra que la relación que existe entre las dos variables es positiva o tiene pendiente positiva, es decir, que por cada hora más que dedique a estudiar mayor será su calificación.

Figura 2.5 Línea de regresión muestral basadas en las cifras de la tabla 2.1



La FRM ecuación (2.36) y la línea de regresión asociada se interpretan de la siguiente manera: cada punto de la línea de regresión proporciona una estimación del valor esperado o valor promedio de "y" correspondiente al valor escogido de "x" es decir $\hat{y}_i$ es una estimación del $E(y|x_i)$. El valor de $\hat{\beta}_1 = 1.215$ que mide la pendiente de la recta e indica que para los valores de x =1, 2, 3, 3.5, 4, 4.5, 5,5.5, 6 Horas de estudio, a medida que "x" aumenta digamos en 1 hora, el aumento estimado en el valor medio o promedio de la Calificación obtenida en el examen es aproximadamente 1.215.

El valor de $\hat{\beta}_0 = 2.356$ o intercepto de la línea indica el nivel promedio de la calificación obtenida en el examen cuando ha estudiado cero horas.

El valor de $r^2 = 0.818$ significa que aproximadamente el 81.8% de la variación de las Calificaciones obtenidas en el examen está explicada por el número de Horas dedicadas a estudiar.

El coeficiente de correlación de r = 0.904 muestra que las dos variables, Calificación obtenida y Horas dedicadas a estudiar están positivamente asociadas.

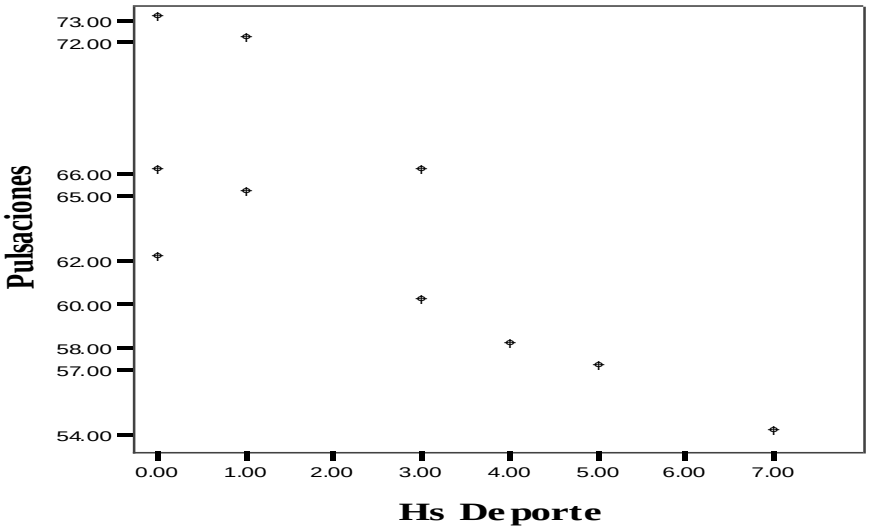
Ejemplo 2: La siguiente tabla recoge los datos de 10 personas, donde "x" es el número de horas semanales que éstas dedican a hacer Deporte (Hs Deporte), y "y" el número de pulsaciones por minuto que las personas tienen cuando están en reposo, estimar los parámetros β_0 y β_1.

Tabla 2.4 Observaciones de 10 personas que practican deporte.

Hs Deporte, x	0	0	0	1	1	3	3	4	5	7
Pulsaciones, y	66	62	73	72	65	60	66	58	57	54

Solución:

Figura 2.6 Diagrama de dispersión de las Pulsaciones vs. Hs Deporte.



En el diagrama de dispersión (figura 2.6) se puede observar que para valores pequeños de "x" los valores de "y" son altos, y para valores altos de "x" los valores de "y" son pequeños lo que indica que cuando una persona dedica pocas horas a hacer deporte sus pulsaciones cuando esta descansando son mayores, y cuando dedican varias horas a hacer deporte sus pulsaciones son mucho menor.

Tabla 2.5 Resultados basados en la tabla 2.4

n	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	0	66	0	4356	0
2	0	62	0	3844	0
3	0	73	0	5329	0
4	1	72	1	5184	72
5	1	65	1	4225	65
6	3	60	9	3600	180
7	3	66	9	4356	198
8	4	58	16	3364	232
9	5	57	25	3249	285
10	7	54	49	2916	378
sumas	$\sum_{i=1}^{n} x_i = 24$	$\sum_{i=1}^{n} y_i = 633$	$\sum_{i=1}^{n} x_i^2 = 110$	$\sum_{i=1}^{n} y_i^2 = 40423$	$\sum_{i=1}^{n} x_i y_i = 1410$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10}(633) = 63.3 \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{10}(24) = 2.4$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$\hat{\beta}_1 = \frac{1410 - \frac{24(633)}{10}}{110 - \frac{(24)^2}{10}}$$

$$\hat{\beta}_1 = \frac{1410 - 1519.2}{110 - 57.6}$$

$$\hat{\beta}_1 = \frac{-109.2}{52.4}$$

$$\hat{\beta}_1 = -2.084$$

Como la pendiente es -2.084 esto confirma lo que se observa en la figura 2.6, es decir datos con pendiente negativa.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_0 = 63.3 - (-2.084)(2.4)$$
$$\hat{\beta}_0 = 63.3 + 5.001$$
$$\hat{\beta}_0 = 68.3016$$
$$\hat{\beta}_0 \approx 68.302$$

Por tanto la ecuación de regresión muestral es:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
$$\hat{y}_i = 68.302 - 2.084 x_i \qquad \textbf{(2.37)}$$

$$\text{Pulsaciones} = 68.302 - 2.084\,(\text{Hs Deporte})$$

Se puede interpretar que la pendiente de -2.084 es la disminución semanal promedio de pulsaciones debido al número de horas dedicadas a hacer deporte, la ordenada al origen de 68.302 representa el número de pulsaciones antes de hacer ejercicio.

Tabla 2.6 Resultados basados en la tabla 2.5

$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	e_i^2
68.302	-2.302	5.299
68.302	-6.302	39.715
68.302	4.698	22.071
66.218	5.782	33.432
66.218	-1.218	1.484
62.05	-2.05	4.203
62.05	3.95	15.603
59.966	-1.966	3.865
57.882	-0.882	0.778
53.714	0.286	0.082
$\sum_{i=1}^{n} \hat{y}_i = 633.004$	$\sum_{i=1}^{n} e_i = -0.004$	$\sum_{i=1}^{n} e_i^2 = 126.531$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

$$\hat{\sigma}^2 = \frac{126.531}{10-2}$$

$$\hat{\sigma}^2 = \frac{126.531}{8}$$

$$\hat{\sigma}^2 = 15.816$$

$\hat{\sigma}^2 = 15.816$, es un estimador de σ^2 y $S_{xx} = 52.4$ denominador de $\hat{\beta}_1$, con estos datos calculamos:

La varianza de $\hat{\beta}_1$:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$
$$\text{var}(\hat{\beta}_1) = \frac{15.816}{52.4}$$
$$\text{var}(\hat{\beta}_1) = 0.30183$$

El error estándar de $\hat{\beta}_1$:

$$\text{es}(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)}$$
$$\text{es}(\hat{\beta}_1) = \sqrt{0.30183}$$
$$\text{es}(\hat{\beta}_1) = 0.5494$$

La varianza de $\hat{\beta}_0$:

$$\text{var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$
$$\text{var}(\hat{\beta}_0) = 15.816\left(\frac{1}{10} + \frac{(2.4)^2}{52.4}\right)$$
$$\text{var}(\hat{\beta}_0) = 15.816(0.2099)$$
$$\text{var}(\hat{\beta}_0) = 3.320$$

El error estándar de $\hat{\beta}_0$:

$$\text{es}(\hat{\beta}_0) = \sqrt{\text{var}(\hat{\beta}_0)}$$
$$\text{es}(\hat{\beta}_0) = \sqrt{3.320}$$
$$\text{es}(\hat{\beta}_0) = 1.822$$

Con los datos obtenidos anteriormente calculamos el valor de r^2 y el valor de r como sigue:

$$\hat{\beta}_1 = -2.083\,, \quad S_{xx} = 52.4$$

$$S_{yy} = \sum_{i=}^{n} y_i^2 - n\bar{y}^2$$

$$S_{yy} = 40423 - 10(63.3)^2$$

$$S_{yy} = 354.1$$

$$r^2 = \hat{\beta}_1^2\left(\frac{S_{xx}}{S_{yy}}\right)$$

$$r^2 = (-2.083)^2\left(\frac{52.4}{354.1}\right)$$

$$r^2 = (4.3388)(0.1479)$$

$$r^2 = 0.643$$

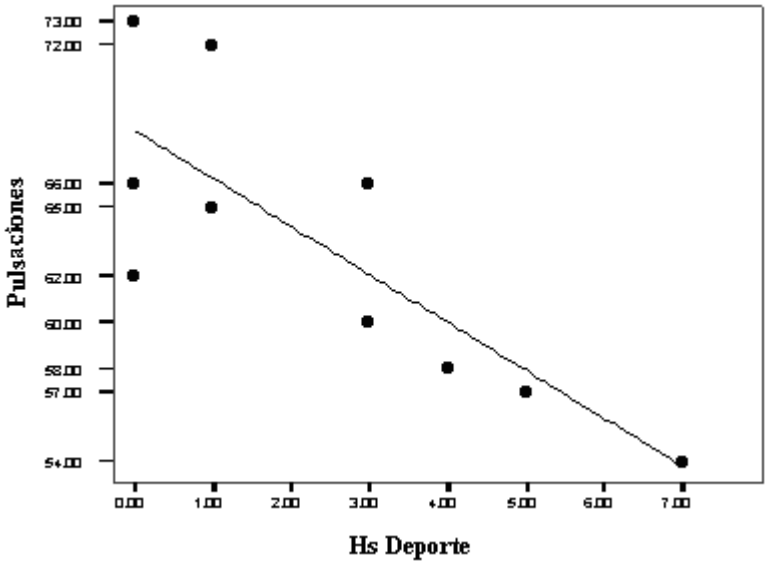
y

$$r = \pm\sqrt{r^2}$$

$$r = \pm\sqrt{0.643}$$

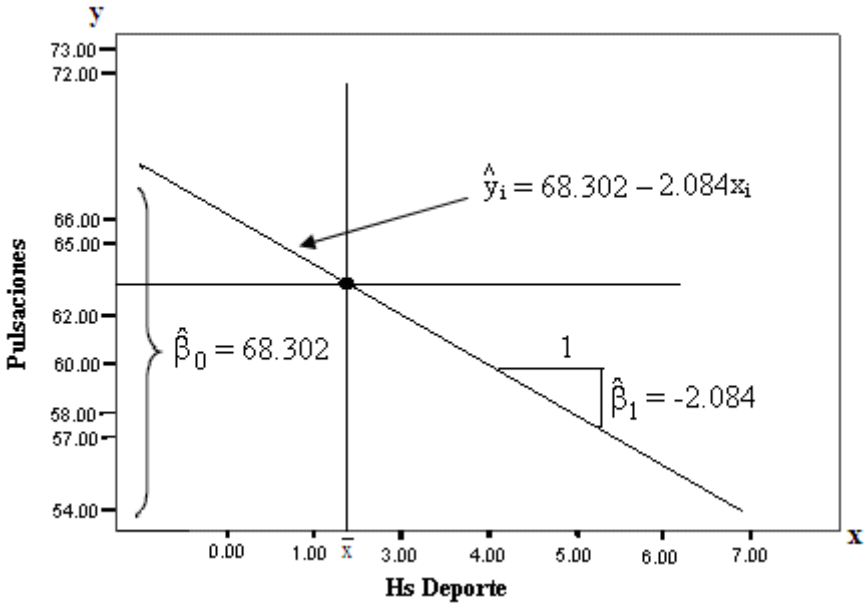
$$r = \pm 0.801$$

Figura 2.7 Diagrama de dispersión con ajuste.



En la figura 2.7 se puede observar que la relación que existe entre las dos variables es negativa o tiene pendiente negativa, es decir que por cada hora más que dedique a hacer deporte una persona sus pulsaciones disminuirán.

Figura 2.8 Línea de regresión muestral basadas en las cifras de la tabla 2.4



La FRM ecuación (2.37) y la línea de regresión asociada se interpretan de la siguiente manera: cada punto de la línea de regresión proporciona una estimación del valor esperado o valor promedio de "y" correspondiente al valor escogido de "x" es decir $\hat{y}_i$ es una estimación del $E(y|x_i)$. El valor de $\hat{\beta}_1 = -2.084$ que mide la pendiente de la recta e indica que para los valores de x = 0, 1, 3, 4, 5, 7 horas semanales, a medida que "x" aumenta digamos en 1 hora, la disminución estimada en el valor medio o promedio de las pulsaciones es aproximadamente -2.084. El valor de $\hat{\beta}_0 = 68.302$ o intercepto de la línea indica el nivel promedio de las pulsaciones cuando no se ha hecho ningún deporte.

El valor de $r^2 = 0.643$ significa que aproximadamente el 64.3% de la variación de las pulsaciones está explicada por el número de horas semanales dedicadas a hacer deporte.

El coeficiente de correlación de r = - 0.801 muestra que las dos variables, Pulsaciones y las Horas dedicadas a hacer deporte están negativamente asociadas, es decir que, a medida que aumentan las Horas dedicadas a hacer deporte las Pulsaciones disminuyen.

2.6 Prueba de Hipótesis de la Pendiente $\hat{\beta}_1$ y del Intercepto $\hat{\beta}_0$.

Con frecuencia interesa probar hipótesis y establecer intervalos de confianza de los parámetros del modelo; estos procedimientos requieren hacer el supuesto adicional de que los errores ε_i del modelo estén distribuidos normalmente. Así, los supuestos son: que los errores estén distribuidos en forma normal e independiente, con media cero y varianza σ^2, lo cual se abrevia "NID (0, σ^2)". NID viene de Normalmente e Independientemente Distribuido.

2.6.1 Uso de las Pruebas t.

Supongamos que se desea probar la hipótesis que la pendiente es igual a una constante por ejemplo a β_{10}. Las hipótesis correspondientes son:

$$\begin{aligned} H_0 &: \beta_1 = \beta_{10} \\ H_1 &: \beta_1 \neq \beta_{10} \end{aligned} \qquad \textbf{(2.38)}$$

En donde se ha especificado una hipótesis alternativa bilateral. Como los errores ε_i son NID (0, σ^2), las observaciones y_i son NID ($\beta_0 + \beta_1 x_i$, σ^2). Ahora, $\hat{\beta}_1$ es una combinación lineal de las observaciones, de modo que $\hat{\beta}_1$ está distribuido normalmente

con promedio β_1 y varianza $\frac{\sigma^2}{S_{xx}}$, usando la media y la varianza de $\hat{\beta}_1$ que se determinó en la sección 2.3.2. Por consiguiente, el estadístico

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2 / S_{xx}}}$$

Está distribuido N(0, 1). Si se conoce σ^2, se podría usar Z_0 para probar la hipótesis (2.38). Comúnmente se desconoce σ^2. Ya se ha visto que MS_{Res} es un estimador insesgado de σ^2. En el apéndice 2.1 (propiedad 6 de los estimadores) se establece que $(n-2)MS_{Res}/\sigma^2$ tiene una distribución ji-cuadrada (χ^2_{n-2}) con n – 2 grados de libertad y que MS_{Res} y $\hat{\beta}_1$ son independientes. De acuerdo con la definición del estadístico t que se presenta en el apéndice 2.1 j) se tiene que:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{10}}{es(\hat{\beta}_1)} \qquad \textbf{(2.39)}$$

Sigue una distribución t_{n-2} si es cierta la hipótesis nula $H_0 : \beta_1 = \beta_{10}$. La cantidad de grados de libertad asociados con t_0 es igual a la cantidad de grados de libertad asociados con MS_{Res}. Así, la razón t_0 es el estadístico con que se prueba $H_0 : \beta_1 = \beta_{10}$. El procedimiento de prueba calcula t_0 y compara su valor observado de acuerdo con la ecuación (2.39) con el punto porcentual $\alpha/2$ superior de t_{n-2} la distribución $t_{(\alpha/2,\ n-2)}$. Este procedimiento rechaza la hipótesis nula si:

$$|t_0| > t_{(\alpha/2,\ n-2)} \qquad \textbf{(2.40)}$$

También se podría usar el método del valor **p** para tomar la decisión.

El denominador del estadístico t_0 en la ecuación (2.39) se llama con frecuencia el error estándar estimado, o más sencillamente el error estándar de la pendiente. Esto es,

$$\text{es}(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}} \qquad \textbf{(2.41)}$$

Por lo anterior, se ve con frecuencia a t_0 escrito en la forma:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\text{es}(\hat{\beta}_1)} \qquad \textbf{(2.42)}$$

Se puede usar un procedimiento parecido para probar hipótesis a cerca de la ordenada al origen. Para probar

$$\begin{aligned} H_0 &: \beta_0 = \beta_{00} \\ H_1 &: \beta_0 \neq \beta_{00} \end{aligned} \qquad \textbf{(2.43)}$$

Se podría usar el estadístico de prueba

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{\beta}_0 - \beta_{00}}{\text{es}(\hat{\beta}_0)} \qquad \textbf{(2.44)}$$

En donde $\text{es}(\hat{\beta}_0) = \sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$ es el error estándar de la ordenada al origen.

La hipótesis nula $H_0 : \beta_0 = \beta_{00}$ se rechaza si $|t_0| > t_{(\alpha/2,\ n-2)}$.

2.6.2 Prueba de Significancia de la Regresión.

Un caso especial muy importante de la hipótesis en la ecuación (2.38) es el siguiente:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0 \qquad \textbf{(2.45)}$$

Estas hipótesis se relacionan con la significancia de la regresión. El no rechazar $H_0 : \beta_1 = 0$ implica que no hay relación lineal entre "x" y "y". Este caso se ilustra en la figura 2.9. Nótese que eso puede implicar que "x" tiene muy poco valor para explicar la variación de "y" y que el mejor estimador para cualquier "x" es $\hat{y} = \bar{y}$ figura 2.9a), o que la verdadera relación entre "x" y "y" no es lineal figura 2.9b). Por consiguiente, si no se rechaza $H_0 : \beta_1 = 0$, equivale a decir que no hay relación lineal entre "x" y "y".

Figura 2.9. Casos en los que no se rechaza la hipótesis H_0.

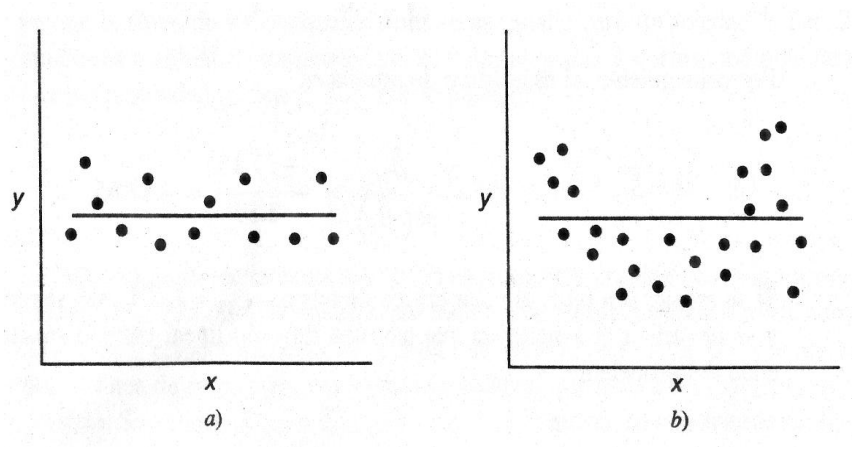
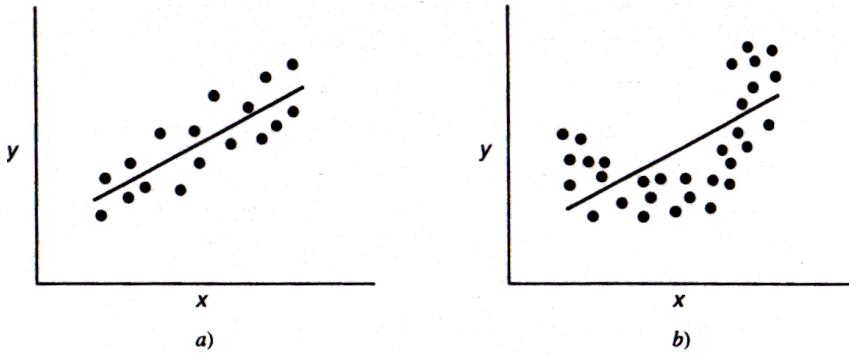


Figura 2.10. Casos en los que se rechaza la hipótesis H_0.



También, si se rechaza H_0, eso implica que "x" sí tiene valor para explicar la variabilidad de "y". Esto se ilustra en la figura 2.10. Sin embargo rechazar H_0 podría equivaler a que el modelo de línea recta es adecuado figura 2.10a), o que aunque hay un efecto lineal de "x", se podrían obtener mejores resultados agregando términos polinomiales en "x" figura 2.10b).

El procedimiento de prueba para H_0 se puede establecer con dos métodos. El primero usa el estadístico t dado en la ecuación (2.41), con $\beta_{10} = 0$, es decir,

$$t_0 = \frac{\hat{\beta}_1 - 0}{es(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{es(\hat{\beta}_1)}$$

La hipótesis de la significancia de la regresión se rechazaría sí $|t_0| > t_{(\alpha/2,\ n-2)}$, y el segundo es el método de análisis de varianza.

Ejemplo 3. Se probará la significancia de la regresión en el modelo de las horas dedicadas a estudiar del ejemplo 1 es decir, $H_0 : \beta_1 = 0$ y $H_1 : \beta_1 \neq 0$.

Datos:

El estimado de la pendiente es $\hat{\beta}_1 = 1.215$.

El estimado de σ^2 que resultó $MS_{Res} = \hat{\sigma}^2 = 0.789$.

El error estándar de la pendiente es $es(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} = \sqrt{0.027393} = 0.1655 \approx 0.166$.

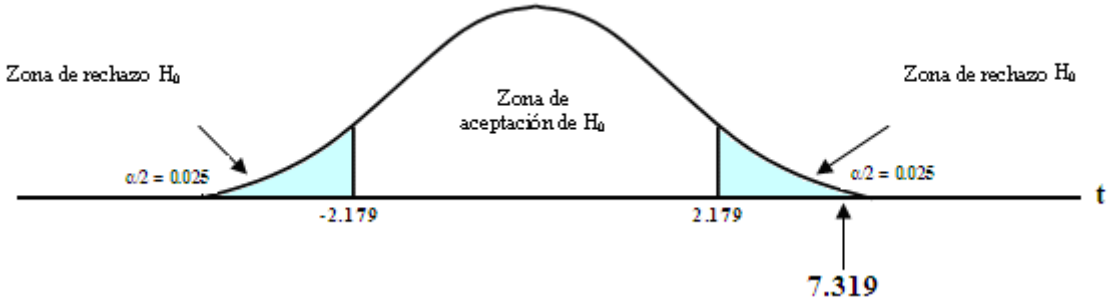
Solución:

1. $H_0 : \beta_1 = 0$
2. $H_1 : \beta_1 \neq 0$

3. Se selecciona un nivel de significancia de $\alpha = 0.05$ y como la prueba es de dos colas $\alpha/2 = 0.05/2 = 0.025$ y se tiene que el valor de la tabla de t es $t_{(0.05/2,\ 14-2)} = t_{(0.025,\ 12)} = 2.179$
4. Región critica: si $t < -2.179$ ó $t > 2.179$, entonces rechazamos H_0.
5. Cálculos:

$$t_0 = \frac{\hat{\beta}_1}{es(\hat{\beta}_1)} = \frac{1.215}{0.166} = 7.319$$

Figura 2.11 de la Distribución t.



6. Decisión Estadística: se rechaza H_0 porque el valor calculado para t_0 cae en la zona de rechazo de H_0, es decir que β_1 es estadísticamente significativa, esto es, significativamente diferente de cero.
7. Conclusión: dado que el valor calculado para t_0 (7.319) es mayor que el de la tabla (2.179) se concluye que hay una relación lineal entre la calificación obtenida en el examen y las horas dedicadas a estudiar.

2.6.3 Análisis de Varianza.

También se puede utilizar un método de análisis de varianza para probar el significado de la regresión. Este análisis se basa en una partición de la variabilidad total de la variable "y" de respuesta. Para obtener esta partición se comienza con la identidad:

$$y_i - \bar{y} = (y_i - \bar{y}) + (y_i - \hat{y}_i) \qquad \textbf{(2.46)}$$

Se elevan al cuadrado ambos miembros de la ecuación (2.46) y aplicando sumatorias, se obtiene:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} [(y_i - \bar{y}) + (y_i - \hat{y}_i)]^2$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n} (y_i - \bar{y})(y_i - \hat{y}_i)$$

Nótese que el tercer término del lado derecho de esta ecuación se puede escribir de la siguiente forma:

$$2\sum_{i=1}^{n} (y_i - \bar{y})(y_i - \hat{y}_i) = 2\sum_{i=1}^{n} \hat{y}_i (y_i - \hat{y}_i) - 2\bar{y}\sum_{i=1}^{n} (y_i - \hat{y}_i) = 2\sum_{i=1}^{n} \hat{y}_i e_i - 2\bar{y}\sum_{i=1}^{n} e_i = 0$$

Ya que la suma de los residuales es siempre cero y la suma de los residuales ponderados por el valor ajustado de $\hat{y}_i$ correspondiente, también es igual a cero por lo anterior,

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \textbf{(2.47)}$$

El lado izquierdo de la ecuación (2.47) es la suma corregida de cuadrados de las observaciones, SS_T (S_{yy}), que mide la variabilidad total en las observaciones. Los dos componentes de SS_T miden, respectivamente, la cantidad de variabilidad en las

observaciones y_i explicada por la línea de regresión, y la variación residual que queda sin explicar por la línea de regresión. Se ve que $SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y})^2$ es la suma de cuadrados de los residuos o la suma de cuadrados de error de la ecuación (2.29). Se acostumbra llamar a $SS_R = \sum_{i=1}^{n} (y_i - \bar{y})^2$ la suma de cuadrados de regresión, o del modelo. La ecuación (2.47) es la identidad fundamental del análisis de varianza para un modelo de regresión. En forma simbólica, se acostumbra a escribir:

$$SS_T = SS_R + SS_{Res} \qquad \textbf{(2.48)}$$

Si se comparan las ecuaciones (2.48) y (2.31), se ve que la suma de cuadrados de regresión se puede calcular como sigue:

$$SS_R = \hat{\beta}_1 S_{xy} \qquad \textbf{(2.49)}$$

La cantidad de grados de libertad se determina como sigue. La suma total de cuadrados, SS_T, tiene df_T = n-1 grados de libertad, porque se perdió un grado de libertad como resultado de la restricción $\sum_{i=1}^{n} (y_i - \bar{y}) = 0$ para las desviaciones $y_i - \bar{y}$. La suma de cuadrados del modelo, o de la regresión es SS_R y tiene df_R = 1 grado de libertad, porque SS_R queda completamente determinado por un parámetro, que es $\hat{\beta}_1$. Por último, antes se dijo que SS_{Res} tiene df_{Res} = n-2 grados de libertad, porque se imponen dos restricciones a las desviaciones $y_i - \hat{y}$ como resultado de estimar $\hat{\beta}_0$ y $\hat{\beta}_1$. Obsérvese que los grados de libertad tienen una propiedad aditiva

$$df_T = df_R + df_{Res}$$

$$n - 1 = 1 + (n - 2) \qquad \textbf{(2.50)}$$

Se puede aplicar la prueba F normal del análisis de varianza para probar la hipótesis $H_0 : \beta_1 = 0$. En el apéndice 2.1 propiedad 6 de los estimadores se puede ver que:

1. $SS_{Res} = (n-2)\, MS_{Res}$ sigue una distribución χ^2_{n-2}.
2. Si es cierta la hipótesis nula $H_0 : \beta_1 = 0$, entonces SS_R tiene una distribución χ^2_{n-2}.
3. SS_{Res} y SS_R son independientes. De acuerdo con la definición del estadístico F.

$$F_0 = \frac{SS_R / df_R}{SS_{Res} / df_{Res}} = \frac{SS_R / 1}{SS_{Res} / (n-2)} = \frac{MS_R}{MS_{Res}} \qquad \textbf{(2.51)}$$

Sigue la distribución $F_{(1,\, n-2)}$. Y los valores esperados de estos cuadrados medios son:

$$E(MS_{Res}) = \sigma^2$$
$$E(MS_R) = \sigma^2 + \beta_1^2 S_{xx}$$

Estos cuadrados medios esperados indican que si es grande el valor esperado de F_0, es probable que la pendiente $\beta_1 \neq 0$. Para probar la hipótesis $H_0 : \beta_1 = 0$, se calcula el estadístico F_0 de prueba y se rechaza H_0 si $F_0 > F_{(\alpha,\, 1,\, n-2)}$.

El procedimiento de prueba se resume en la tabla 2.7.

Tabla 2.7 Análisis de varianza para probar el significado de la regresión.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F_0
Regresión	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_{Res}
Residual	$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}$	n-2	MS_{Res}	
Total	SS_T	n-1		

Ejemplo 4. Se probará el significado de la regresión para el modelo desarrollado en el ejemplo 1, es decir si $H_0 : \beta_1 = 0$ ó $H_1 : \beta_1 \neq 0$ de los datos de las horas dedicadas a estudiar y la calificación obtenida en el examen.

Datos:

El modelo ajustado es $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 2.356 + 1.215 x_i$

El valor para $SS_T = S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=}^{n} y_i^2 - n\bar{y}^2 = 738 - 14(7)^2 = 52$

y $S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = 409.5 - \frac{(53.5)(98)}{14} = 35$

La suma de cuadrados de regresión se calcula con la ecuación $SS_R = \hat{\beta}_1 S_{xy}$, como sigue:

$SS_R = \hat{\beta}_1 S_{xy} = 1.215(35) = 42.525.$

Y la suma de los errores al cuadrado $SS_{Res} = \sum_{i=1}^{n} e_i^2 = 9.471$

El análisis de varianza se resume en la tabla 2.8.

Solución:

1. $H_0 : \beta_1 = 0$
2. $H_1 : \beta_1 \neq 0$
3. Se selecciona un nivel de significancia de $\alpha = 0.05$ y se tiene el valor de la tabla F es $F_{(0.05, 1, 12)} = 4.75$
4. Cálculos:

$$F_0 = \frac{SS_R / 1}{SS_{Res} / (n-2)} = \frac{MS_R}{MS_{Res}} = \frac{\hat{\beta}_1 S_{xy}}{\hat{\sigma}^2} = \frac{42.525}{0.789} = 53.897$$

El valor calculado de $F_0 = 53.897$ y el de la tabla $F_{(0.05, 1, 12)} = 4.75$

Tabla 2.8 Análisis de varianza para el modelo de regresión horas de estudio.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F_0
Regresión	42.525	1	42.525	42.525/0.789 = 53.897
Residual	9.471	12	0.789	
Total	52	13		

5. Decisión Estadística: se rechaza H_0 porque el valor calculado para F_0 (53.897) es mayor que el de la tabla (4.75).
6. Conclusión: Se concluye que la variación en la calificación obtenida puede atribuirse a las horas dedicadas a estudiar.

Más a cerca de la prueba t.

Se dijo, en la sección 2.6.1, que el estadístico de prueba

$$t_0 = \frac{\hat{\beta}_1}{es(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{MS_{Res}/S_{xx}}} \qquad \textbf{(2.52)}$$

Se podría usar para probar la significancia de la regresión. Sin embargo, nótese que al elevar al cuadrado ambos miembros de la ecuación (2.52) se obtiene

$$t_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_{Res}} = \frac{MS_R}{MS_{Res}} \qquad \textbf{(2.53)}$$

Así, t_0^2 en la ecuación (2.53) es idéntica a F_0 del método de análisis de varianza en la ecuación (2.51). Una muestra es el ejemplo 3 de las horas dedicadas a estudiar, $t_0 = 7.319$, así que $t_0^2 = (7.319)^2 = 53.567 \approx F_0 = 53.897$. En general, el cuadrado de una variable aleatoria t con f grados de libertad es una variable aleatoria F con 1 y f grados de libertad en el numerador y el denominador respectivamente. Aunque la prueba t para $H_0 : \beta_1 = 0$ equivale a la prueba F en la regresión lineal simple, la prueba t es algo más adaptable, porque se podría usar para probar hipótesis alternativas unilaterales (sea H_1: $\beta_1 < 0$ o H_1: $\beta_1 > 0$), mientras que la prueba F sólo considera la alternativa bilateral.

Por último, recuérdese que decidir que $\beta_1 = 0$ es una conclusión muy importante que sólo es apoyada por la prueba t o la prueba F. La incapacidad de demostrar que la pendiente no es estadísticamente distinta de cero no necesariamente quiere decir que “x” y “y” no están relacionadas. Puede indicar que la capacidad de detectar esta relación se ha confundido por la varianza del proceso de medición, o que el intervalo de valores de

"x" es inadecuado. Se requiere una gran cantidad de evidencia no estadística y conocimiento del problema, para llegar a la conclusión que $\beta_1 = 0$.

2.6.4 Prueba de Hipótesis de la Correlación.

Como se vio en el Capítulo 1, el Análisis de Correlación intenta medir la fuerza de tales relaciones entre dos variables por medio de un simple número que recibe el nombre de coeficiente de correlación.

La constante ρ (rho) recibe el nombre de coeficiente de correlación poblacional y juega un papel importante en muchos problemas de análisis de datos de dos variables.
El valor de ρ es 0 cuando $\beta_1 = 0$, lo cual resulta cuando esencialmente no hay regresión lineal; esto es, la línea de regresión es horizontal y cualquier conocimiento de "x" no es de utilidad para predecir "y".

Ejemplo 5. Para los datos de la tabla 2.1 Horas dedicadas a estudiar y la Calificación obtenida se encuentra que $r = \pm\sqrt{r^2} = \pm\sqrt{0.818} = \pm 0.904$.

Un coeficiente de correlación de 0.904 indica una buena relación lineal positiva entre "x" y "y". Dado que $r^2 = 0.818$, se puede afirmar que aproximadamente el 81.8% de la variación de los valores de "y" se deben a una relación lineal con "x".

Una prueba de la hipótesis especial $\rho = 0$ contra una alternativa apropiada es equivalente a probar $\beta_1 = 0$, para el modelo de regresión lineal simple y, por lo tanto, son

aplicables los procedimientos de la sección 2.6.1 en los que se utiliza la distribución t con n-2 grados de libertad o la distribución F con 1 y n-2 grados de libertad.

El valor de t_0 está dado por:

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} \qquad \textbf{(2.54)}$$

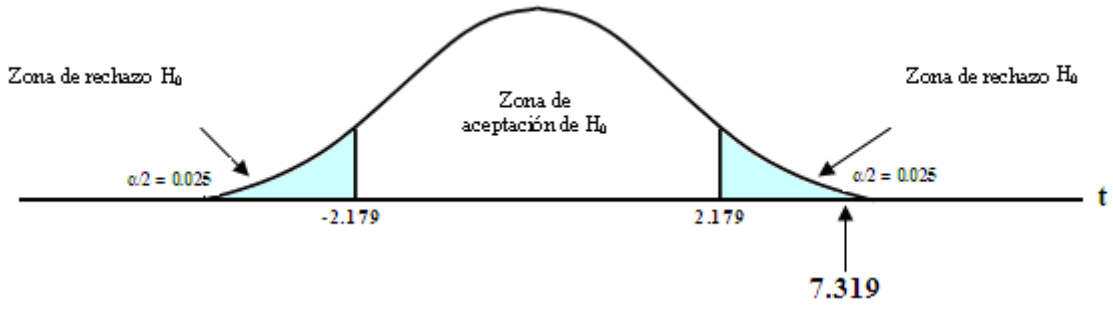
La hipótesis nula se rechazaría sí $|t_0| > t_{(\alpha/2,\ n-2)}$

Solución:

1. $H_0 : \rho = 0$
2. $H_1 : \rho \neq 0$
3. Se selecciona un nivel de significancia de $\alpha = 0.05$ y como la prueba es de dos colas $\alpha/2 = 0.05/2 = 0.025$ y se tiene que el valor de la tabla de t es $t_{(0.05/2,\ 14-2)} = t_{(0.\ 025,\ 12)} = 2.179$.
4. Región critica: si t < - 2.179 ó t > 2.179, entonces rechazamos H_0.
5. Cálculos:

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{1.215}{\sqrt{0.789/28.803}} = \frac{1.215}{0.166} = 7.319$$

Figura 2.12 de la Distribución t.



6. Decisión Estadística: se rechaza la hipótesis de no asociación lineal.
7. Conclusión: dado que el valor calculado para t_0 (7.319) es mayor que el de la tabla (2.179) se concluye que hay una relación lineal entre la calificación obtenida en el examen y las horas dedicadas a estudiar.

2.7 Estimación de Intervalo en la Regresión Lineal Simple.

En esta sección se describirá la estimación del intervalo de confianza de los parámetros del modelo de regresión.

2.7.1 Intervalos de confianza de β_0, β_1 y σ^2.

Además de los estimadores puntuales de β_0, β_1 y σ^2, también se pueden obtener estimados de intervalos de confianza para esos parámetros. El ancho de dichos intervalos es una medida de la calidad general de la recta de regresión. Si los errores se distribuyen en forma normal e independiente, entonces la distribución de muestreo tanto de

$$t = \frac{\hat{\beta}_1 - \beta_1}{es(\hat{\beta}_1)} \quad y \quad t = \frac{\hat{\beta}_0 - \beta_0}{es(\hat{\beta}_0)}$$

Es la distribución t con n-2 grados de libertad. Así, un intervalo de confianza de 100(1-α) por ciento para la pendiente β_1 se determina como:

$$\hat{\beta}_1 - t_{(\alpha/2,n-2)} es(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{(\alpha/2,n-2)} es(\hat{\beta}_1) \qquad \textbf{(2.55)}$$

Y un intervalo de confianza de 100(1-α) por ciento para la ordenada al origen β_0 es:

$$\hat{\beta}_0 - t_{(\alpha/2,n-2)} es(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{(\alpha/2,n-2)} es(\hat{\beta}_0) \qquad \textbf{(2.56)}$$

Estos intervalos de confianza tienen la interpretación usual, por lo tanto, si hubiese que tomar muestras repetidas del mismo tamaño a los mismos valores de "x", y formar, por ejemplo, intervalos de confianza de 95% de la pendiente para cada muestra entonces el 95% de esos intervalos contendrán el verdadero valor de β_1.

Si los errores están distribuidos en forma normal e independiente, el apéndice 2.1 propiedad 6 de los estimadores detalla que la distribución de muestreo de $(n-2)\hat{\sigma}^2/\sigma^2 = (n-2)MS_{Res}/\sigma^2$ es ji-cuadrada, con n-2 grados de libertad. Así:

$$P\left\{\chi^2_{(1-\alpha/2,\ n-2)} \le \frac{(n-2)MS_{Res}}{\sigma^2} \le \chi^2_{(\alpha/2,\ n-2)}\right\} = 1-\alpha$$

Y en consecuencia, un intervalo de confianza de 100(1-α) por ciento para σ^2 es:

$$\frac{(n-2)MS_{Res}}{\chi^2_{(\alpha/2,\ n-2)}} \le \sigma^2 \le \frac{(n-2)MS_{Res}}{\chi^2_{(1-\alpha/2,\ n-2)}} \qquad \textbf{(2.57)}$$

Ejemplo 6: Se establecerán los intervalos de confianza del 95% para β_0, β_1 y σ^2 con los datos de las horas dedicadas a estudiar del ejemplo 1.

Intervalo de confianza para β_0.

Datos:

El valor estimado del intercepto es: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 7 - 1.215(3.8214) = 2.356$

El error estándar de $\hat{\beta}_0$ es$(\hat{\beta}_0) = \sqrt{\text{var}(\hat{\beta}_0)} = \sqrt{0.456} = 0.675$

El valor de la tabla de t es $t_{(0.05/2,\ 14-2)} = t_{(0.025,\ 12)} = 2.179$

Sustituyen los datos anteriores en la ecuación (2.56) se tiene:

$$\hat{\beta}_0 - t_{(\alpha/2,\ n-2)}es(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{(\alpha/2,\ n-2)}es(\hat{\beta}_0)$$
$$2.356 - 2.179(0.675) \le \beta_0 \le 2.356 + 2.179(0.675)$$
$$2.356 - 1.47 \le \beta_0 \le 2.356 + 1.47$$
$$0.88 \le \beta_0 \le 3.826$$

El 95% de esos intervalos incluirán el verdadero valor del intercepto.

Si se escoge un valor distinto de α y se utilizan los mismos datos, el ancho del intervalo de confianza resultante será distinto.

Intervalo de confianza para β_1.

Datos:

El valor estimado de la pendiente es $\hat{\beta}_1 = 1.215$

El error estándar de $\hat{\beta}_1$ $es(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} = \sqrt{0.027393} = 0.1655 \approx 0.166$

El valor de la tabla de t es $t_{(0.05/2,\ 14-2)} = t_{(0.025,\ 12)} = 2.179$

Solución:

$$\hat{\beta}_1 - t_{(\alpha/2,\ n-2)}es(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{(\alpha/2,\ n-2)}es(\hat{\beta}_1)$$
$$1.215 - 2.179(0.166) \le \beta_1 \le 1.215 + 2.179(0.166)$$
$$1.215 - 0.361714 \le \beta_1 \le 1.215 + 0.361714$$
$$0.85 \le \beta_1 \le 1.57$$

En otras palabras, el 95% de esos intervalos incluirán el verdadero valor de la pendiente.

En general, cuando más grande sea el coeficiente de confianza 1-α, el intervalo de confianza será mayor.

Intervalo de confianza para σ^2.

El intervalo de confianza de 95%, para σ^2 se determina a partir de la ecuación (2.57).

Utilizando los datos:

Cuadrado medio: $\hat{\sigma}^2 = MS_{Res} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{9.471}{14-2} = \frac{9.471}{12} = 0.789$

Grados de libertad: n-2 = 14-2 = 12

Limite inferior: $\chi^2_{(\alpha/2,\ n-2)} = \chi^2_{(0.05/2,\ 12)} = \chi^2_{(0.025,\ 12)} = 23.337$

Limite superior: $\chi^2_{(1-\alpha/2,\ n-2)} = \chi^2_{(1-0.05/2,\ 12)} = \chi^2_{(0.975,\ 12)} = 4.404$

El intervalo de confianza queda de la siguiente manera:

$$\frac{(12)(0.789)}{23.337} \leq \sigma^2 \leq \frac{(12)(0.789)}{4.404}$$

$$\frac{9.468}{23.337} \leq \sigma^2 \leq \frac{9.468}{4.404}$$

$$0.40 \leq \sigma^2 \leq 2.15$$

Se puede interpretar de la siguiente forma: dado un coeficiente de confianza del 95%, en 95 de cada 100 intervalos tales como (0.40, 2.15) deberán contener el verdadero valor de σ^2.

2.8 Estimación por Máxima Verosimilitud.

Un método general de estimación puntual con algunas propiedades teóricas más definidas que las del método de los MCO es el método de Máxima Verosimilitud (MV), cuya idea fundamental consiste en estimar parámetros de modo tal que la probabilidad de observar "y" sea lo máximo posible maximizar la FMV.

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

$$\begin{aligned} FMV(y_i, x_i, \beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left[\frac{-1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right] \\ FMV(y_i, x_i, \beta_0, \beta_1, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right] \end{aligned} \qquad \textbf{(2.58)}$$

Y por la independencia de las observaciones, tomando logaritmo natural a ambos lados resulta que la función es:

$$\ln FMV(y_i, x_i, \beta_0, \beta_1, \sigma^2) = \ln\left[(2\pi\sigma^2)^{-n/2} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]\right]$$

$$\ln FMV(y_i, x_i, \beta_0, \beta_1, \sigma^2) = \ln(2\pi\sigma^2)^{-n/2} + \ln\exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right]$$

Por propiedades de logaritmo

$$\ln FMVo(y_i, x_i, \beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) + \left(\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)$$

$$\ln FMV(y_i, x_i, \beta_0, \beta_1, \sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \qquad \textbf{(2.59)}$$

Para obtener los estimadores de β_0 y β_1 derivaremos esta función respecto a cada uno de los parámetros.

$$\frac{\partial \ln FMV}{\partial \tilde{\beta}_0} = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$$

$$\frac{\partial \ln FMV}{\partial \tilde{\beta}_0} = 0 - 0 - \frac{1}{2\tilde{\sigma}^2}(2)\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)(-1)$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x} \qquad \textbf{(2.60)}$$

$$\frac{\partial \ln FMV}{\partial \tilde{\beta}_1} = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$$

$$\frac{\partial \ln FMV}{\partial \tilde{\beta}_1} = 0 - 0 - \frac{1}{2\tilde{\sigma}^2}(2)\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)(-x_i)$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad \textbf{(2.61)}$$

La deducción de las ecuaciones (2.60) y (2.61) se encuentra en el apéndice **2.1 k)**

$$\frac{\partial \ln FMV}{\partial \tilde{\sigma}^2} = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$$

$$0 = -\frac{n}{\hat{\sigma}} + \frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$$

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n} \qquad \textbf{(2.62)}$$

Obsérvese que los estimadores de Máxima Verosimilitud de la ordenada al origen y de la pendiente, $\tilde{\beta}_0$ y $\tilde{\beta}_1$, son idénticos a los obtenidos con los Mínimos Cuadrados. También $\tilde{\sigma}^2$ es un estimador sesgado de σ^2. El estimador sesgado se

relaciona con el estimador insesgado $\hat{\sigma}^2$ ecuación (2.32) mediante $\tilde{\sigma}^2 = \frac{n-1}{n}\hat{\sigma}^2$. El sesgo es pequeño cuando n es moderadamente grande, por lo general se usa el estimador insesgado $\hat{\sigma}^2$.

En este Capítulo se hizo más énfasis en el método de Mínimos Cuadrados Ordinarios por lo siguiente:

Se minimiza la suma de cuadrados de los residuos por varias razones:

- $\rightarrow$ Es fácil obtener la fórmula de los estimadores.
- $\rightarrow$ Sin técnicas de optimización numérica.
- $\rightarrow$ Teoría estadística es sencilla: insesgadez, consistencia, etc.
- $\rightarrow$ Solución coincide con las propiedades deducidas de la esperanza condicional.

Ejercicios 2.

1. En la siguiente tabla se muestran 8 observaciones donde "x" es el ingreso de los padres en miles de dólares y "y" promedio de calificaciones de un grupo de estudiantes.

x	21	15	15	9	12	18	6	12
y	4	3	3.5	2	3	3.5	2.5	2.5

   a) Calcular los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ para la curva de regresión y formar la ecuación.

   b) Calcular la varianza $\hat{\sigma}^2$.

   c) Calcular la varianza de los parámetros $\text{var}(\hat{\beta}_1)$, $\text{var}(\hat{\beta}_0)$ y sus errores estándar.

   d) Calcular el coeficiente de determinación r^2.

   e) Realizar la prueba de hipótesis para la pendiente y para la ordenada al origen.

   f) Establecer los intervalos de confianza del 95% para β_0, β_1 y σ^2.

2. Se cree que la pureza del oxigeno producido con un proceso de fraccionamiento está relacionada con el porcentaje de hidrocarburos en el condensador principal de la unidad de procesamiento. A continuación se muestran los datos.

Pureza (%)	Hidrocarburos (%)	Pureza (%)	Hidrocarburos (%)
86.91	1.02	96.73	1.46
89.85	1.11	99.42	1.55
90.28	1.43	98.66	1.55
86.34	1.11	96.07	1.55
92.58	1.01	93.65	1.40
87.33	0.95	87.31	1.15
86.29	1.11	95.00	1.01
91.86	0.87	96.85	0.99
95.61	1.43	85.20	0.95
89.86	1.02	90.56	0.98

a) Ajustar un modelo de regresión lineal simple a los datos.

b) Probar la hipótesis H_0: $\beta_1 = 0$.

c) Calcular r^2.

d) Determinar un intervalo de confianza de 95% para la pendiente.

e) Concluir de acuerdo a lo obtenido en los literales anteriores.

3. En la tabla siguiente aparecen los datos sobre el desempeño de los 26 equipos de la liga nacional de fútbol en 1976. Se cree que la cantidad de yardas ganadas por tierra por los equipos contrarios "x" tiene un efecto sobre la cantidad de juegos que gana un equipo "y".

Cantidad de juegos	Yardas por tierra del contrario	Cantidad de juegos	Yardas por tierra del contrario
10	2205	6	1901
11	2096	5	2288
11	1847	5	2072
13	1903	5	2861
10	1457	6	2411
11	1848	4	2289
10	1564	3	2203
11	1821	3	2592
4	2577	4	2053
2	2476	10	1979
7	1984	6	2048
10	1917	8	1786
9	1761	2	287
9	1709	0	2560

a) Formar la tabla del análisis de la varianza y probar el significado de la regresión.

b) Determinar un intervalo de confianza de 95% para la pendiente.

c) Concluir en base a los resultados.

4. Construir la recta de regresión y formar los intervalos de 90% de confianza para los parámetros de regresión de los datos siguientes, donde x = nº de revoluciones por minuto, y = potencia en Kw. de una maquina diesel.

x	400	500	600	700	750
y	580	1030	1420	1880	2100

5. La estatura de un bebe al nacer (en cm.) y el periodo de embarazo (en días) son:

x	277.1	279.3	281.4	283.2	284.8
y	48	49	50	51	52

Ajustar una recta de regresión y construir intervalos de confianza para sus coeficientes.

¿Es lineal la relación entre las variables “x” y “y”?.

6. Calcular la varianza residual y el coeficiente de correlación para los datos siguientes:

Presión	Temperatura
20.79	194.5
22.40	197.9
23.15	199.4
23.89	200.9
24.02	201.4
25.14	203.6
28.49	209.5
29.04	210.7
29.88	211.9
30.06	212.2

7. Para los datos del ejercicio 2 del Capítulo 1 realizar lo siguiente:

   a) Calcular los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$, para la curva de regresión y formar la ecuación.

   b) Calcular la varianza $\hat{\sigma}^2$.

   c) Calcular la varianza de los parámetros $\text{var}(\hat{\beta}_1)$ y $\text{var}(\hat{\beta}_0)$ y el error estándar de $\sqrt{\text{var}(\hat{\beta}_0)}$ y $\sqrt{\text{var}(\hat{\beta}_1)}$.

   d) Calcular el coeficiente de determinación r^2.

   e) Realizar la prueba de hipótesis para la pendiente y para la ordenada al origen.

   f) Establecer los intervalos de confianza del 95% para β_0, β_1 y σ^2.

8. Ha salido al mercado un nuevo modelo de grabadora de DVD, un poco más caro que los anteriores, pero con unas prestaciones muy superiores, de manera que la labor de los técnicos de los grandes centros comerciales es muy importante a la hora de presentar este producto al cliente. Con el objetivo de saber si el "número de técnicos comerciales presentes en una tienda" (x) puede tener alguna incidencia en el "número de aparatos vendidos durante una semana" (y), se observaron quince centros comerciales con los resultados que se muestran a continuación:

$$\sum_{i=1}^{15} x_i = 215;\ \sum_{i=1}^{15} x_1^2 = 3567;\ \sum_{i=1}^{15} y_i = 1700;\ \sum_{i=1}^{15} x_i y_i = 28300$$

   a) Encontrar la recta de regresión

   b) Cual es el número de aparatos que se puede estimar que se venderán en un centro con 17 comerciales.

Apéndice 2: Deducción de Ecuaciones.

2.1 Deducción de ecuaciones utilizadas en el Capítulo 2.

a) Deducción de $\hat{\beta}_0$ ecuación (2.13).

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$
$$y_i = \hat{y}_i + e_i$$
$$e_i = y_i - \hat{y}_i$$

y $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ entonces, $\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} e_i^2$

Derivando ambos lados de la ecuación anterior con respecto ha $\hat{\beta}_0$, es decir que los demás términos se toman como constantes y la derivada de una constante es cero.

$$\frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2\right) = \frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=1}^{n} e_i^2\right)$$

$$2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(0 - 1 - 0) = 0$$

$$2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-1) = 0$$

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \frac{0}{-2}$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i = n\hat{\beta}_0$$

$$\frac{\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i}{n} = \hat{\beta}_0$$

$$\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n} = \hat{\beta}_0$$

$$\bar{y} - \hat{\beta}_1 \bar{x} = \hat{\beta}_0$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

L.q.q.d

b) Deducción de $\hat{\beta}_1$ ecuación (2.14) derivando ahora con respecto a $\hat{\beta}_1$ se tiene:

$$\frac{\partial}{\partial \hat{\beta}_1}\left(\sum_{i=}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right) = \frac{\partial}{\partial \hat{\beta}_1}\left(\sum_{i=1}^{n} e_i^2\right)$$

$$2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(0 - 0 - x_i) = 0$$

$$2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(-x_i) = 0$$

$$-2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i) = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i) = \frac{0}{-2}$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)(x_i) = 0$$

$$\sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i x_i = 0$$

$$\sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\sum_{i=1}^{n} y_i x_i - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i = \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \hat{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i \right)$$

$$\frac{\sum_{i=1}^{n} y_i x_i - \bar{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i} = \hat{\beta}_1$$

$$\frac{\sum_{i=1}^{n} y_i x_i - \dfrac{\sum_{i=1}^{n} y_i}{n} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \dfrac{\sum_{i=1}^{n} x_i}{n} \sum_{i=1}^{n} x_i} = \hat{\beta}_1$$

$$\frac{\sum_{i=1}^{n} y_i x_i - \dfrac{\left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i \right)^2}{n}} = \hat{\beta}_1$$

L.q.q.d

c) Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ por Mínimos Cuadrados son estimadores insesgados de los parámetros β_0 y β_1 del modelo, es decir que $E(\hat{\beta}_0) = \beta_0$ y $E(\hat{\beta}_1) = \beta_1$, Para demostrarlo con $\hat{\beta}_1$, primero se tiene que:

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ Donde $S_{xy} = \sum_{i=1}^{n} y_i (x_i - \bar{x})$ y $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ entonces $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$

Si se hace $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$ donde

$$c_i = \frac{(x_i - \bar{x})}{S_{xx}}$$

$$c_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Se facilitan los cálculos.

La esperanza o valor medio de $\hat{\beta}_1$ ecuación (2.20).

$$E(\hat{\beta}_1) = E(\sum_{i=1}^{n} c_i y_i) = \sum_{i=1}^{n} c_i E(y_i) = \sum_{i=1}^{n} c_i (\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

Pero $\sum_{i=1}^{n} c_i = 0$ porque.

$$\sum_{i=1}^{n} c_i = \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$\sum_{i=1}^{n} c_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} c_i = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} c_i = \frac{\sum_{i=1}^{n} x_i - n\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} c_i = \frac{\sum_{i=1}^{n} x_i - n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} c_i = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} c_i = \frac{0}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} c_i = 0$$

y $\sum_{i=1}^{n} c_i x_i = 1$ Porque

$$\sum_{i=1}^{n} c_i x_i = \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) x_i$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n} x_i x_i - \sum_{i=1}^{n} \bar{x} x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2)}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n} x_i^2 - \left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - 2\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)\sum_{i=1}^{n} x_i + n\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)^2}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{\sum_{i=1}^{n} x_i^2 - 2\frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} + \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\dfrac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n}}{\sum_{i=1}^{n} x_i^2 - 2\dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} + \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{\dfrac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n}}{\dfrac{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$\sum_{i=1}^{n} c_i x_i = \frac{n\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)}{n\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)}$$

$$\sum_{i=1}^{n} c_i x_i = 1$$

Entonces,

$$E(\hat{\beta}_1) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

$$E(\hat{\beta}_1) = \beta_0 (0) + \beta_1 (1)$$

$$E(\hat{\beta}_1) = \beta_1$$

L.q.q.d

d) Para probar que $E(\hat{\beta}_0) = \beta_0$ se parte de que: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \dfrac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \bar{x}$ pero en el desarrollo de $E(\hat{\beta}_1)$ se consideró $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$ entonces,

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum_{i=1}^{n} y_i}{n} - \bar{x}\sum_{i=1}^{n} c_i y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}c_i\right)y_i$ y luego haciendo r igual a lo que está dentro del paréntesis para facilitar el desarrollo de la deducción de la ecuación $r = \frac{1}{n} - \bar{x}c_i$ entonces, $\hat{\beta}_0 = \sum_{i=1}^{n} r_i y_i$

La esperanza o valor medio de $\hat{\beta}_0$ ecuación (2.21).

$$E(\hat{\beta}_0) = E\left(\sum_{i=1}^{n} r_i y_i\right)$$

$$E(\hat{\beta}_0) = \sum_{i=1}^{n} r_i E(y_i)$$

$$E(\hat{\beta}_0) = \sum_{i=1}^{n} r_i (\beta_0 + \beta_1 x_i)$$

$$E(\hat{\beta}_0) = \sum_{i=1}^{n} r_i \beta_0 + \sum_{i=1}^{n} r_i \beta_1 x_i$$

$$E(\hat{\beta}_0) = \beta_0 \sum_{i=1}^{n} r_i + \beta_1 \sum_{i=1}^{n} r_i x_i$$

Pero $\sum_{i=1}^{n} r_i = 1$ porque

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}c_i\right)$$

$$\sum_{i=1}^{n} r_i = \frac{\sum_{i=1}^{n} 1}{n} - \bar{x}\sum_{i=1}^{n} c_i \quad \text{pero} \sum_{i=1}^{n} c_i = 0$$

Así:

$$\sum_{i=1}^{n} r_i = \frac{n}{n} - \bar{x}(0)$$

$$\sum_{i=1}^{n} r_i = 1$$

y

$$\sum_{i=1}^{n} r_i x_i = 0$$

$$\sum_{i=1}^{n} r_i x_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \bar{x}c_i\right)x_i$$

$$\sum_{i=1}^{n} r_i x_i = \frac{\sum_{i=1}^{n} x_i}{n} - \bar{x}\sum_{i=1}^{n} c_i x_i \quad \text{pero} \sum_{i=1}^{n} c_i x_i = 1 \text{ así}$$

$$\sum_{i=1}^{n} r_i x_i = \bar{x} - \bar{x}(1)$$

$$\sum_{i=1}^{n} r_i x_i = 0$$

Por lo que

$$E(\hat{\beta}_0) = \beta_0 \sum_{i=1}^{n} r_i + \beta_1 \sum_{i=1}^{n} r_i x_i$$

$$E(\hat{\beta}_0) = \beta_0 (1) + \beta_1 (0)$$

$$E(\hat{\beta}_0) = \beta_0$$

Por lo tanto $\hat{\beta}_0$ es un estimador insesgado de β_0.

L.q.q.d

e) Deducción de la varianza de $\hat{\beta}_1$ ecuación (2.23).

$$\text{var}(\hat{\beta}_1) = \text{var}\left(\sum_{i=1}^{n} c_i y_i\right)$$

$$\text{var}(\hat{\beta}_1) = \sum_{i=1}^{n} c_i^2 \text{ var}(y_i)$$

En las asunciones del modelo sección 1.6 del Capítulo 1 se vio que $\text{Var}(\varepsilon_i) = \sigma^2$.

Tomando en cuenta esto se tiene,

$$\text{var}(\hat{\beta}_1) = \sum_{i=1}^{n} c_i^2 \, \text{var}(y_i)$$

$$\text{var}(\hat{\beta}_1) = \sum_{i=1}^{n} c_i^2 \sigma^2$$

$$\text{var}(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} c_i^2$$

$$\text{var}(\hat{\beta}_1) = \sigma^2 \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2$$

$$\text{var}(\hat{\beta}_1) = \sigma^2 \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^4} \right)$$

$$\text{var}(\hat{\beta}_1) = \sigma^2 \left(\frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Por lo tanto $\text{var}(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$

L.q.q.d

f) Deducción de la varianza de $\hat{\beta}_0$ ecuación (2.24), antes se tomó $\hat{\beta}_0 = \sum_{i=1}^{n} r_i y_i$ con el propósito de facilitar el desarrollo, entonces,

$$\operatorname{var}(\hat{\beta}_0) = \operatorname{var}\left(\sum_{i=1}^{n} r_i y_i\right)$$

$$\operatorname{var}(\hat{\beta}_0) = \sum_{i=1}^{n} r_i^2 \operatorname{var}(y_i)$$

$$\operatorname{var}(\hat{\beta}_0) = \sum_{i=1}^{n} r_i^2 \sigma^2$$

$$\operatorname{var}(\hat{\beta}_0) = \sigma^2 \sum_{i=1}^{n} r_i^2$$

$$\operatorname{var}(\hat{\beta}_0) = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} c_i\right)^2$$

$$\operatorname{var}(\hat{\beta}_0) = \sigma^2 \sum_{i=1}^{n} \left(\left(\frac{1}{n}\right)^2 - 2\frac{1}{n}\bar{x} c_i + \bar{x}^2 c_i^{\,2}\right)$$

$$\operatorname{var}(\hat{\beta}_0) = \sigma^2 \left(\sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 - 2\frac{1}{n}\bar{x} \sum_{i=1}^{n} c_i + \bar{x}^2 \sum_{i=1}^{n} c_i^{\,2}\right)$$

$$\text{pero} \sum_{i=1}^{n} c_i = 0 \text{ entonces,}$$

$$\operatorname{var}(\hat{\beta}_0) = \sigma^2 \left(n\left(\frac{1}{n}\right)^2 - 2\frac{1}{n}\bar{x}(0) + \bar{x}^2 \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \right)$$

$$\operatorname{var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \bar{x}^2 \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \right)$$

$$\operatorname{var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \bar{x}^2 \left(\frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \right)$$

$$\operatorname{var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \bar{x}^2 \frac{1}{S_{xx}} \right) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

Por lo tanto $\text{var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$

L.q.q.d

g) Teorema de Gauss-Markov.

Los estimadores de Mínimos Cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ son lineales e insesgados para mostrar que estos estimadores tienen varianza mínima dentro de la clase de todos los estimadores lineales e insesgados consideremos el estimador de $\hat{\beta}_1$

$$\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i \quad \text{Donde } c_i = k_i = \frac{(x_i - \bar{x})}{S_{xx}} = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Que muestra que $\hat{\beta}_1$ es promedio ponderado de los "y" con c_i sirviendo como ponderaciones.

Definiendo un estimador alterno de β_1 así:

$$\beta_1^* = \sum_{i=1}^{n} k_i y_i$$

Donde k_i son también ponderaciones iguales a c_i. Ahora bien

$$E(\beta_1^*) = E(\sum_{i=1}^{n} k_i y_i)$$

$$E(\beta_1^*) = \sum_{i=1}^{n} k_i E(y_i)$$

$$E(\beta_1^*) = \sum_{i=1}^{n} k_i (\beta_0 + \beta_1 x_i)$$

$$E(\beta_1^*) = \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} k_i x_i$$

Por lo tanto para que β_1^* sea insesgado se requiere que:

$$\sum_{i=1}^{n} k_i = 0 \text{ y } \sum_{i=1}^{n} k_i x_i = 1$$

$$\text{Así } E(\beta_1^*) = 0 + \beta_1(1) = \beta_1$$

Ahora

$$\text{var}(\beta_1^*) = \text{var}\left(\sum_{i=1}^{n} c_i y_i\right)$$

$$\text{var}(\beta_1^*) = \text{var}\left(\sum_{i=1}^{n} k_i y_i\right)$$

$$\text{var}(\beta_1^*) = \sum_{i=1}^{n} k_i^2 \text{ var}(y_i) \qquad \text{Pero Var}(y_i) = \sigma^2, \text{entonces}$$

$$\text{var}(\beta_1^*) = \sigma^2 \sum_{i=1}^{n} k_i^2$$

$$\text{var}(\beta_1^*) = \sigma^2 \sum_{i=1}^{n} \left(k_i - \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2$$

Como se puede observar se ha sumado un cero adecuado, ahora agrupando términos y desarrollando el cuadrado se tiene:

$$\text{var}(\beta_1^*) = \sigma^2 \sum_{i=1}^{n} \left(\left(k_i - \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) + \left(\frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \right)^2$$

$$\text{var}(\beta_1^*) = \sigma^2 \sum_{i=1}^{n} \left(k_i - \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2 + 2\sigma^2 \sum_{i=1}^{n} \left(k_i - \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \left(\frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) + \sigma^2 \sum_{i=1}^{n} \left(\frac{(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2$$

$$\text{var}(\beta_1^*) = \sigma^2 \sum_{i=1}^{n} \left(k_i - \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2 + 2\sigma^2 \sum_{i=1}^{n} \left(k_i - \frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \left(\frac{x_i - \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) + \sigma^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2}$$

Sustituyendo k_i los dos primeros términos se hacen cero y solamente queda:

$$\text{var}(\beta_1^*) = \sigma^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2}$$

$$\text{var}(\beta_1^*) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\text{var}(\beta_1^*) = \frac{\sigma^2}{S_{xx}}$$

Por lo tanto $\text{var}(\beta_1^*) = \text{var}(\hat{\beta}_1)$

Puede entonces decirse que con ponderaciones $k_i = c_i$, que son las ponderaciones de Mínimos Cuadrados Ordinarios, la varianza del estimador lineal β_1^* es igual a la varianza del estimador de Mínimos Cuadrados $\hat{\beta}_1$; o si no $\text{var}(\beta_1^*) > \text{var}(\hat{\beta}_1)$. Dicho de otra forma, si hay un estimador lineal insesgado de β_1 con varianza mínima, debe ser el estimador de Mínimos Cuadrados Ordinarios. Igualmente puede mostrarse que $\hat{\beta}_0$ es un estimador lineal insesgado de β_0 con varianza mínima.

h) Deducción de SS_{Res} ecuación (2.31).

$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ Entonces, $SS_{Res} = S_{yy} - \hat{\beta}_1 S_{xy} = SS_T - \hat{\beta}_1 S_{xy}$

Primeramente se tiene:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{xx} = \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^2$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2$$

Multiplicando y dividiendo por n el término del centro se tiene

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{n}{n}\sum_{i=1}^{n} x_i + n\bar{x}^2$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\frac{\sum_{i=1}^{n} x_i}{n} + n\bar{x}^2$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - 2\bar{x}n\bar{x} + n\bar{x}^2$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 n + n\bar{x}^2$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Así se tiene:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} y_i (x_i - \bar{x})$$

Entonces,

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$SS_{Res} = \sum_{i=1}^{n} (y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i)^2$$

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2$$

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})^2 \text{ agrupandose tiene}$$

$$SS_{Res} = \sum_{i=1}^{n} ((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})^2$$

$$SS_{Res} = \sum_{i=1}^{n} \left((y_i - \bar{y})^2 - 2(y_i - \bar{y})\hat{\beta}_1 (x_i - \bar{x}) + \hat{\beta}_1 (x_i - \bar{x})^2 \right)$$

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \bar{y})^2 - 2\hat{\beta}_1 \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$SS_{Res} = S_{yy} - 2\hat{\beta}_1 S_{xy} + \hat{\beta}_1^2 S_{xx} \qquad \text{pero } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$SS_{Res} = S_{yy} - 2\left(\frac{S_{xy}}{S_{xx}}\right) S_{xy} + \left(\frac{S_{xy}}{S_{xx}}\right)^2 S_{xx}$$

$$SS_{Res} = S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}^2} S_{xx}$$

$$SS_{Res} = S_{yy} - 2\frac{S_{xy}^2}{S_{xx}} + \frac{S_{xy}^2}{S_{xx}}$$

$$SS_{Res} = S_{yy} - \frac{S_{xy}}{S_{xx}} S_{xy}$$

$$SS_{Res} = S_{yy} - \hat{\beta}_1 S_{xy}$$

Por lo tanto $SS_{Res} = S_{yy} - \hat{\beta}_1 S_{xy}$

L.q.q.d

i) Un estimador insesgado de σ^2 es $\hat{\sigma}^2$.

$$\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = MS_{Res} \text{ Cuadrado Medio Residual}$$

$$SS_{Res} = S_{yy} - \hat{\beta}_1 S_{xy} \text{ y como } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \Rightarrow \hat{\beta}_1 S_{xx} = S_{xy} \text{ , entonces}$$

$$SS_{Res} = S_{yy} - \hat{\beta}_1 S_{xy}$$
$$SS_{Res} = S_{yy} - \hat{\beta}_1 (\hat{\beta}_1 S_{xx})$$
$$SS_{Res} = S_{yy} - \hat{\beta}_1^2 (S_{xx})$$
$$SS_{Res} = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$SS_{Res} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \hat{\beta}_1^2 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right)$$
$$SS_{Res} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 - \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) \hat{\beta}_1^2$$

Ahora al tomar los valores esperados se tiene:

$$E(SS_{Res}) = \sum_{i=1}^{n} E(y_i^2) - nE(\bar{y}^2) - \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right) E(\hat{\beta}_1^2)$$

Por teorema se sabe que una forma de calcular la varianza de una variable aleatoria es:

$$\sigma^2 = E(x^2) - \mu^2 \text{ despejando la } E(x^2) \text{ se tiene } \sigma^2 + \mu^2 = E(x^2)$$

Se pueden sustituir las cantidades

$$E(y_i^2) = \sigma_{y_i}^2 + \mu_{y_i}^2$$

$$E(\bar{y}^2) = \sigma_{\bar{y}}^2 + \mu_{\bar{y}}^2$$

$$E(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 + \mu_{\hat{\beta}_1}^2$$

La ecuación $E(S_{Res}) = \sum_{i=1}^{n} E(y_i^2) - nE(\bar{y}^2) - \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) E(\hat{\beta}_1^2)$ queda de la forma siguiente:

$$E(S_{Res}) = \sum_{i=1}^{n} (\sigma_{y_i}^2 + \mu_{y_i}^2) - n(\sigma_{\bar{y}}^2 + \mu_{\bar{y}}^2) - \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)(\sigma_{\hat{\beta}_1}^2 + \mu_{\hat{\beta}_1}^2) \text{ Pero}$$

$$\mu_y = \beta_0 + \beta_1 x$$
$$\mu_{\bar{y}} = \beta_0 + \beta_1 \bar{x}$$
$$\mu_{\hat{\beta}_1} = \beta_1$$
$$\sigma_{\bar{y}}^2 = \sigma^2/n$$
$$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{S_{xx}}$$

Entonces distribuyendo el símbolo de sumatoria y sustituyendo se tiene:

$$E(S_{Res}) = \sum_{i=1}^{n} (\sigma_{y_i}^2 + \mu_{y_i}^2) - n(\sigma_{\bar{y}}^2 + \mu_{\bar{y}}^2) - \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)(\sigma_{\hat{\beta}_1}^2 + \mu_{\hat{\beta}_1}^2)$$

$$E(S_{Res}) = \sum_{i=1}^{n} \sigma_{y_i}^2 + \sum_{i=1}^{n} \mu_{y_i}^2 - n(\sigma_{\bar{y}}^2 + \mu_{\bar{y}}^2) - \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)(\sigma_{\hat{\beta}_1}^2 + \mu_{\hat{\beta}_1}^2)$$

$$E(S_{Res}) = n\sigma^2 + \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i)^2 - n\left(\frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x})^2\right) - \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right)$$

$$E(S_{Res}) = n\sigma^2 + \sum_{i=1}^{n} (\beta_0^2 + 2\beta_0\beta_1 x_i + \beta_1^2 x_i^2) - \frac{n\sigma^2}{n} - n(\beta_0^2 + 2\beta_0\beta_1\bar{x} + \beta_1^2\bar{x}^2) - \left(\frac{\sum_{i=1}^{n} x_i^2\sigma^2}{S_{xx}} + \beta_1^2\sum_{i=1}^{n} x_i^2 - \frac{n\bar{x}^2\sigma^2}{S_{xx}} - n\bar{x}^2\beta_1^2\right)$$

$$E(S_{Res}) = n\sigma^2 + \sum_{i=1}^{n} \beta_0^2 + 2\beta_0\beta_1\sum_{i=1}^{n} x_i + \beta_1^2\sum_{i=1}^{n} x_i^2 - \sigma^2 - n\beta_0^2 - 2\beta_0\beta_1 n\bar{x} - \beta_1^2 n\bar{x}^2 - \left(\frac{\sigma^2\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} + \beta_1^2\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\beta_1^2\right)$$

$$E(SS_{Res}) = n\sigma^2 + n\beta_0^2 + 2\beta_0\beta_1\sum_{i=1}^{n} x_i + \beta_1^2\sum_{i=1}^{n} x_i^2 - \sigma^2 - n\beta_0^2 - 2\beta_0\beta_1 n\frac{\sum_{i=1}^{n} x_i}{n} - \beta_1^2 n\bar{x}^2 - \left(\sigma^2 + \beta_1^2\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\beta_1^2\right)$$

Reduciendo términos semejantes se obtiene:

$$E(SS_{Res}) = n\sigma^2 + 2\beta_0\beta_1\sum_{i=1}^{n} x_i + \beta_1^2\sum_{i=1}^{n} x_i^2 - \sigma^2 - 2\beta_0\beta_1\sum_{i=1}^{n} x_i - \beta_1^2 n\bar{x}^2 - \sigma^2 - \beta_1^2\sum_{i=1}^{n} x_i^2 + n\bar{x}^2\beta_1^2$$

$$E(SS_{Res}) = n\sigma^2 - \sigma^2 - \sigma^2$$

$$E(SS_{Res}) = n\sigma^2 - 2\sigma^2$$

$$E(SS_{Res}) = (n-2)\sigma^2$$

Por lo tanto $E(SS_{Res}) = (n-2)\sigma^2$ ahora tomando esperanza de

$$E(\hat{\sigma}^2) = \frac{E(SS_{Res})}{n-2}$$

$$E(\hat{\sigma}^2) = \frac{E\left(\sum_{i=1}^{n} e_i^2\right)}{n-2}$$

$$E(\hat{\sigma}^2) = \frac{(n-2)\sigma^2}{n-2}$$

$$E(\hat{\sigma}^2) = \sigma^2$$

Se concluye entonces que $\hat{\sigma}^2$ es un estimador insesgado de σ^2.

L.q.q.d

En resumen y de acuerdo con el supuesto de normalidad, los estimadores por Mínimos Cuadrados $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\sigma}^2$ poseen las siguientes propiedades estadísticas.

1. Son insesgados.
2. Tienen varianza mínima, tomando en cuenta la propiedad anterior esto quiere decir que son insesgados con varianza mínima, es decir estimadores eficientes.

3. Consistentes, esto es, que a medida que el tamaño de la muestra aumenta indefinidamente, los estimadores convergen al valor poblacional verdadero.

4. $\hat{\beta}_0$ está normalmente distribuida con

$$\text{Media: } E(\hat{\beta}_0) = \beta_0$$

$$\text{Varianza: } \text{var}(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$$

Se puede escribir como $\hat{\beta}_0 \sim N(\beta_0, \sigma^2_{\hat{\beta}_0})$

5. $\hat{\beta}_1$ está normalmente distribuida con

$$\text{Media: } E(\hat{\beta}_1) = \beta_1$$

$$\text{Varianza: } \text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

Y puede escribirse también como $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$

6. $(n-2)\hat{\sigma}^2/\sigma^2 = (n-2)MS_{Res}/\sigma^2$ está distribuida como la distribución χ^2 (ji-cuadrado) con n-2 grados de libertad, porque $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$ y sustituyendo este valor en $(n-2)\hat{\sigma}^2/\sigma^2$ en vez de $\hat{\sigma}^2$ se tiene: $\frac{(n-2)\left(\frac{\sum_{i=1}^{n} e_i^2}{n-2}\right)}{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2}$

7. $(\hat{\beta}_0, \hat{\beta}_1)$ están distribuidas independientemente de $\hat{\sigma}^2$.

j) Fórmula de la Distribución Normal y la Distribución t.

La fórmula de la distribución normal es: $Z = \frac{x - \mu}{\sigma}$ pero comúnmente se desconoce la varianza poblacional (σ^2) entonces se utiliza la distribución t, la fórmula es: $t = \frac{x - \mu}{s}$ partiendo de esta definición se tienen los valores de t para los parámetros.

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} = \frac{\hat{\beta}_1 - \beta_{10}}{es(\hat{\beta}_1)} \quad y \quad t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{\beta}_0 - \beta_{00}}{es(\hat{\beta}_0)}$$

k) Deducción de los parámetros de regresión por el método de Máxima Verosimilitud.

Derivando primero con respecto a $\tilde{\beta}_0$.

$$\frac{\partial \ln FMV}{\partial \tilde{\beta}_0} = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2$$

$$\frac{\partial \ln FMV}{\partial \tilde{\beta}_0} = 0 - 0 - \frac{1}{2\tilde{\sigma}^2}(2)\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)(-1)$$

$$\frac{\partial \ln FMV}{\partial \tilde{\beta}_0} = \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)$$

Ahora igualando a cero la derivada parcial y despejando $\tilde{\beta}_0$ se tiene:

$$0 = \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)$$

$$0 = \sum_{i=1}^{n}(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)$$

$$0 = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\tilde{\beta}_0 - \sum_{i=1}^{n}\tilde{\beta}_1 x_i$$

$$0 = \sum_{i=1}^{n} y_i - n\tilde{\beta}_0 - \tilde{\beta}_1 \sum_{i=1}^{n} x_i$$

$$n\tilde{\beta}_0 = \sum_{i=1}^{n} y_i - \tilde{\beta}_1 \sum_{i=1}^{n} x_i$$

$$\tilde{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{\tilde{\beta}_1 \sum_{i=1}^{n} x_i}{n}$$

$$\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1 \bar{x}$$

L.q.q.d

Derivando ahora con respecto a $\tilde{\beta}_1$

$$\frac{\partial \ln FMV}{\partial \tilde{\beta}_1} = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right)^2$$

$$\frac{\partial \ln FMV}{\partial \tilde{\beta}_1} = 0 - 0 - \frac{1}{2\tilde{\sigma}^2}(2)\sum_{i=1}^{n}\left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right)(-x_i)$$

$$\frac{\partial \ln FMV}{\partial \tilde{\beta}_1} = \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}\left(y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\right)x_i$$

Igualado la derivada parcial a cero y despejando $\tilde{\beta}_1$ se tiene:

$$0 = \frac{1}{\tilde{\sigma}^2}\sum_{i=1}^{n}\left(x_i y_i - x_i\tilde{\beta}_0 - \tilde{\beta}_1 x_i^2\right)$$

$$0 = \sum_{i=1}^{n}\left(x_i y_i - x_i\tilde{\beta}_0 - \tilde{\beta}_1 x_i^2\right)$$

$$0 = \sum_{i=1}^{n} x_i y_i - \tilde{\beta}_0\sum_{i=1}^{n} x_i - \tilde{\beta}_1\sum_{i=1}^{n} x_i^2$$

$$0 = \sum_{i=1}^{n} x_i y_i - (\bar{y} - \tilde{\beta}_1\bar{x})\sum_{i=1}^{n} x_i - \tilde{\beta}_1\sum_{i=1}^{n} x_i^2$$

$$0 = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i + \tilde{\beta}_1\bar{x}\sum_{i=1}^{n} x_i - \tilde{\beta}_1\sum_{i=1}^{n} x_i^2$$

$$-\tilde{\beta}_1 \bar{x} \sum_{i=1}^{n} x_i + \tilde{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i$$

$$\tilde{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i$$

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i}$$

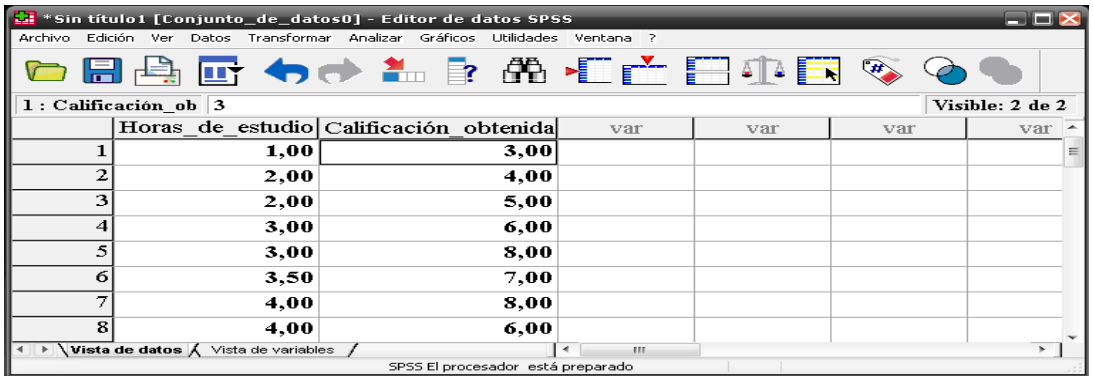
$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i \right)^2}{n}}$$

L.q.q.d

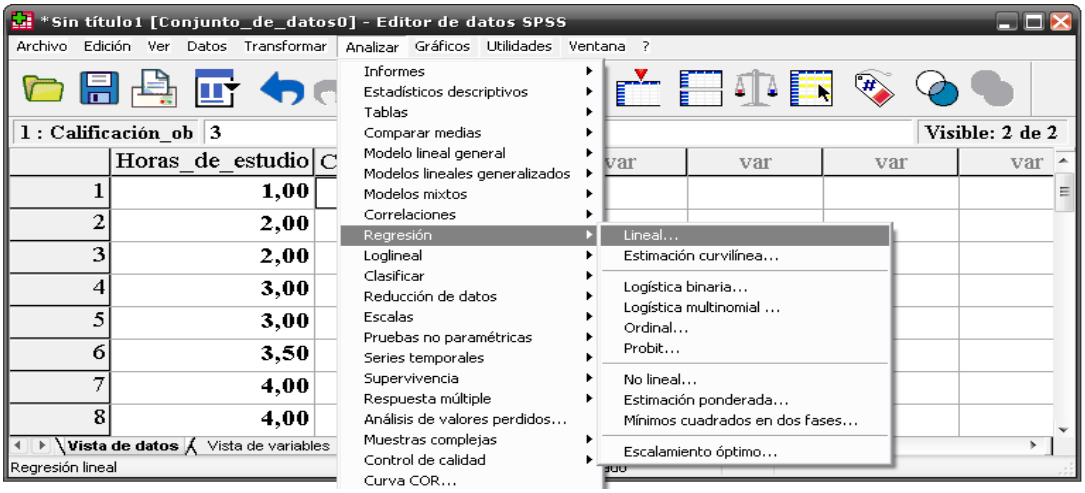
Apéndice 2.2: Solución de Ejemplos Haciendo uso del Software Estadístico SPSS v15.0.

Haciendo uso del software se pueden obtener los resultados de los ejemplos 1, 3, 4, 6, en una sola ejecución siguiendo los pasos que se muestran a continuación:

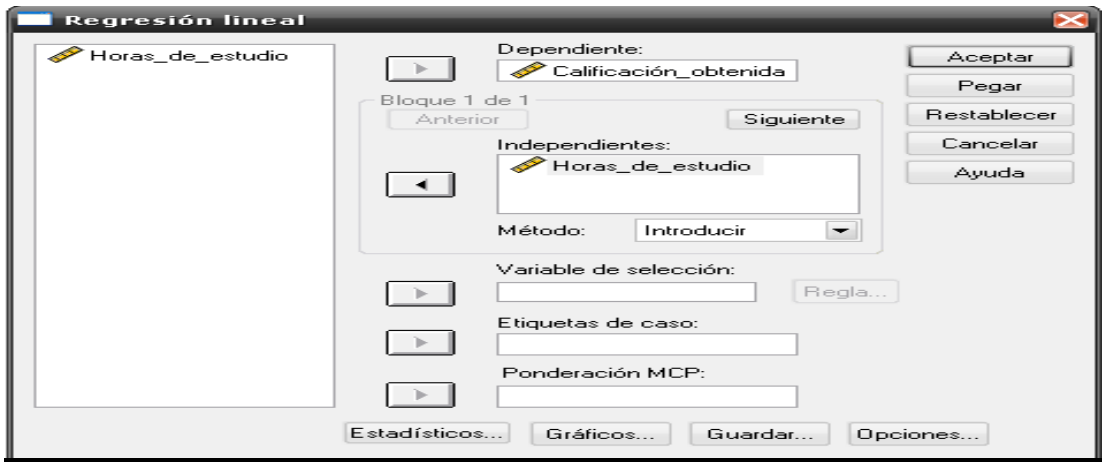
1. Se les da un nombre a las dos variables en estudio, se digitan los datos para cada variable y se obtiene la ventana siguiente en la cual solamente se muestran 8 observaciones del total (14).



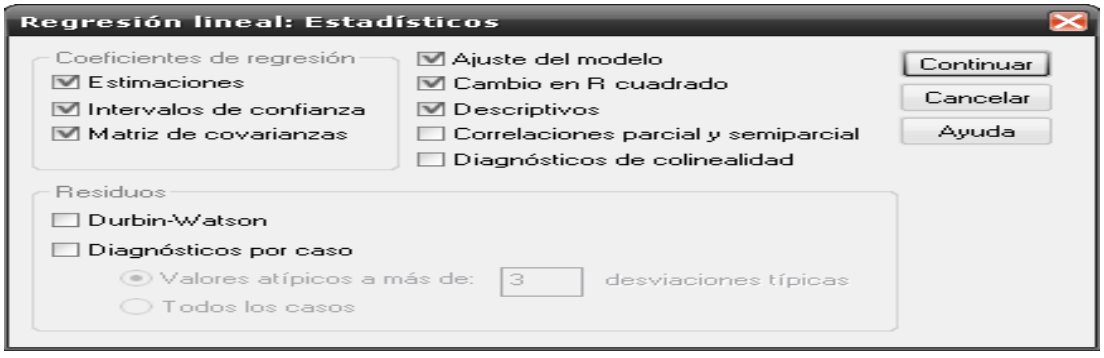
2. En la barra de menú se selecciona la opción Analizar → Regresión → Lineal como se muestra a continuación:

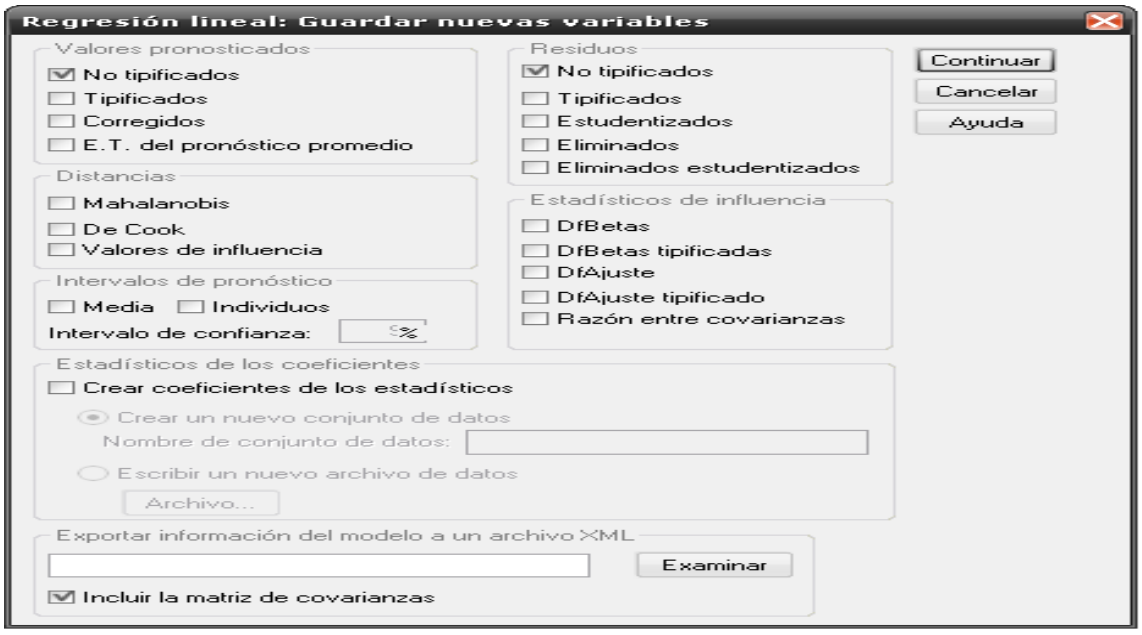


3. Al hacer click en la opción lineal aparece la siguiente ventana en la cual se colocan las variables cada una en su lugar.



Al pulsar en los botones **Estadístico** y **Guardar** aparecen los cuadros siguientes:





Dando un click en el botón aceptar aparecen los siguientes resultados

Variables introducidas/eliminadas[b]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Horas_de_estudio[a]	.	Introducir

a. Todas las variables solicitadas introducidas

b. Variable dependiente: Calificación_obtenida

En la tabla de variables introducidas se observa que no se ha eliminado ninguna variable

Estadísticos descriptivos

	Media	Desviación típ.	N
Calificación_obtenida	7.0000	2.00000	14
Horas_de_estudio	3.8214	1.48851	14

La tabla de estadísticos descriptivos muestra la media que son exactamente las obtenidas en el ejemplo y la desviación típica para cada una de las variables, también puede observarse que aparece el número de observaciones n = 14.

Correlaciones

		Horas_de_estudio	Calificación_obtenida
Horas_de_estudio	Correlación de Pearson	1	.904**
	Sig. (bilateral)		.000
	Suma de cuadrados y productos cruzados	28.804	35.000
	N	14	14
Calificación_obtenida	Correlación de Pearson	.904**	1
	Sig. (bilateral)	.000	
	Suma de cuadrados y productos cruzados	35.000	52.000
	N	14	14

**. La correlación es significativa al nivel 0,01 (bilateral).

En la tabla correlaciones se presenta la correlación de cada variable que es el 1 que aparece, eso quiere decir que la correlación de una variable con ella misma es 1 ó correlación perfecta, el valor de 0.904 es el coeficiente de correlación, para las dos variables ó r que como se vio en el Capítulo 1 es una medida del grado de relación lineal existente entre dos variables, el valor de 28.804 es la varianza de la variable "x" o S_{xx}, se tiene también la varianza de la variable "y" esto es $S_{yy} = 52.000$, además se muestran los productos cruzados es decir $S_{xy} = 35$ o sea la covarianza de las variables "x" y "y".

Coeficientes[a]

Modelo		Coeficientes no estandarizados		t	Sig.	Intervalo de confianza para B al 95%	
		B	Error típ.			Límite inferior	Límite superior
1	(Constante)	2.356	.676	3.488	.004	.884	3.829
	Horas_de_estudio	1.215	.166	7.341	.000	.854	1.576

a. Variable dependiente: Calificación_obtenida

En la tabla coeficientes se muestran los coeficientes $\hat{\beta}_0 = 2.356$ y $\hat{\beta}_1 = 1.215$, que son los mismos valores obtenidos en el desarrollo del ejemplo 1. El error estándar de la pendiente y de la ordenada al origen es: $es(\hat{\beta}_1) = 0.166$ y $es(\hat{\beta}_0) = 0.675$, se tienen los

valores de t para la pendiente y la ordenada 3.488 y 7.341 el valor de t para la pendiente es el mismo que se obtuvo en el ejemplo 3, se puede observar también que los intervalos de confianza para los parámetros son $0.884 \leq \beta_0 \leq 3.829$ y $0.854 \leq \beta_1 \leq 1.576$ es casi igual al obtenido en el ejemplo 6 sólo que varía un poco por algunas aproximaciones internas del software.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	42.529	1	42.529	53.888	.000[a]
	Residual	9.471	12	.789		
	Total	52.000	13			

a. Variables predictoras: (Constante), Horas_de_estudio

b. Variable dependiente: Calificación_obtenida

La tabla ANOVA, es la misma del análisis de la varianza se puede observar que los valores obtenidos en esta son iguales a los obtenidos en el desarrollo del ejemplo 4.

Resumen del modelo[b]

Modelo	R	R cuadrado	Error típ. de la estimación	Estadísticos de cambio			
				Cambio en F	gl1	gl2	Sig. del cambio en F
1	.904[a]	.818	.88838	53.888	1	12	.000

a. Variables predictoras: (Constante), Horas_de_estudio

b. Variable dependiente: Calificación_obtenida

En la tabla resumen del modelo se observa el valor del coeficiente de correlación, pero, también muestra el coeficiente de determinación r^2 o bondad de ajuste de la línea de regresión al conjunto de datos, con el cual se puede decir que hay un buen ajuste ya que este valor es 0.818 cerca de 1 como se mostró en el ejemplo 1.

Capítulo 3

Validación del Modelo y Predicción.

3.1 Introducción a Validación del Modelo y Predicción.

Una vez estimado el modelo de regresión y obtenidos los residuos, hay que comprobar si los supuestos que se han utilizado para construirlo no están en contradicción con los datos; a este proceso se le denomina validación del modelo. Si los supuestos son adecuados, se puede utilizar el modelo de regresión lineal para generar predicciones y/o tomar decisiones. Los supuestos de un modelo estadístico se refieren a una serie de condiciones que deben darse para garantizar la validez del modelo. Al efectuar aplicaciones prácticas del modelo de regresión, nos veremos en la necesidad de examinar muchos de estos supuestos. En este Capítulo se estudian los cuatro supuestos del modelo:

- → Linealidad: La relación entre las dos variables es lineal.
- → Homoscedasticidad: La variabilidad de los residuos es constante.
- → Normalidad: Los residuos siguen una distribución normal.
- → Independencia: Los residuos son independientes entre sí.

Los dos primeros supuestos pueden generalmente comprobarse antes de construir el modelo, observando el gráfico de dispersión entre las dos variables. Los supuestos de normalidad e independencia conviene comprobarlos, analizando los residuos después de ajustar el modelo. Los residuos también dan información respecto a la linealidad y a la

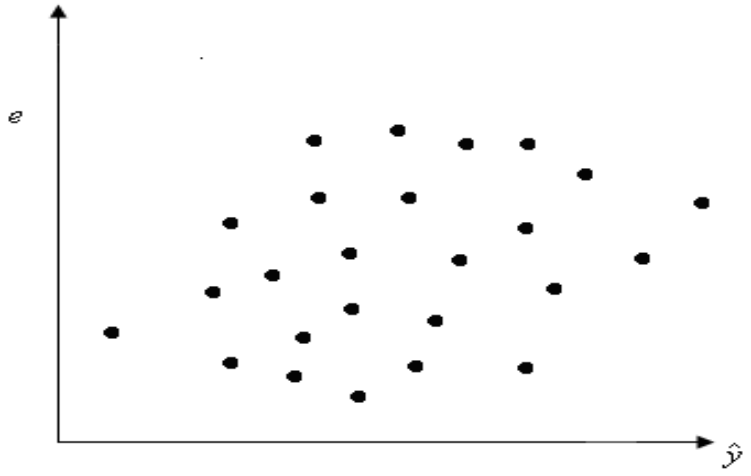
moscedasticidad. En este Capítulo también se utiliza el modelo para hacer predicciones.

3.2 Análisis de los Residuos.

El análisis de los residuos consiste en ver la distribución de los residuos; esto se realiza gráficamente representando en un diagrama de dispersión los puntos $(\hat{y}_i, e_i)$; es decir, sobre el eje de las abscisas se representa el valor estimado $\hat{y}_i$ y sobre el eje de las ordenadas, el valor correspondiente del residuo, es decir $e_i = y_i - \hat{y}_i$.

Veamos un ejemplo.

Figura 3.1. Diagrama de dispersión de los valores estimados y los residuos.



Si el modelo lineal obtenido se ajusta bien a los datos entonces la nube de puntos $(\hat{y}_i, e_i)$ no debe mostrar ningún tipo de estructura.

Para ilustrar la utilidad del análisis de los residuos del modelo estimado, la tabla 3.1 presenta cuatro conjuntos de datos distintos en los que los valores de "x" para las tres regresiones primeras son los mismos, este ejemplo es debido a Anscombe (1973)[1].

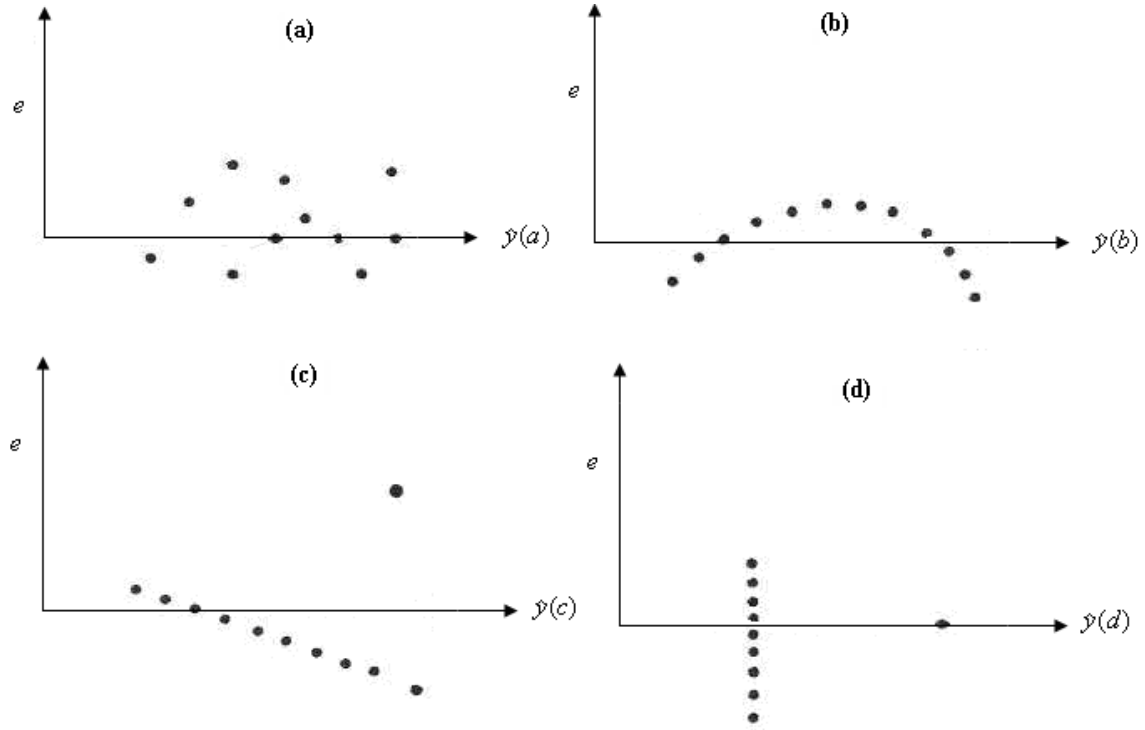
Tabla 3.1 Datos de Anscombe (1973).

Caso (a)		**Caso (b)**		**Caso (c)**		**Caso (d)**	
x(a)	y(a)	x(b)	y(b)	x(c)	y(c)	x(d)	y(d)
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

Al hacer la regresión de "y" sobre "x" en los cuatro casos, se obtiene exactamente la misma recta $\hat{y} = 3 + 0.5x$, la variación explicada, la no explicada y la varianza residual son iguales en las cuatro regresiones ($\hat{\sigma}^2 = 1.52$), así como el valor del estadístico t para el contraste $H_0 : \beta_1 = 0$. El coeficiente de correlación **r** es también igual en los cuatro modelos (0.82). Por lo tanto, las cuatro regresiones parecen ser formalmente idénticas. Sin embargo, si se estudian sus residuos, la situación se modifica radicalmente: la figura 3.2 presenta gráficos de los residuos e_i frente a los valores estimados $\hat{y}_i$ para los cuatro conjuntos de datos.

[1] El ejemplo de Anscombe se puede encontrar en el artículo siguiente: TW. Anscombe (1973) "grapas in Statistical analysis". The American Statistician (núm. 27, Pág. 17-21).

Figura 3.2 Gráficos de los residuos para el ejemplo de Anscombe.



De acuerdo con los gráficos de los residuos, el modelo (a) no ofrece ninguna evidencia de error de especificación, el modelo (b) no verifica el supuesto de linealidad ya que los residuos muestran claramente una estructura curvilínea, el modelo (c) no verifica el supuesto de normalidad de las perturbaciones ya que tiene un valor atípico incompatible con una distribución normal y que afecta mucho a la estimación de la regresión.

Finalmente, en el modelo (d) no podemos comprobar si los supuestos son ciertos o no, ya que la pendiente de la recta viene determinada únicamente por un valor, y tendríamos que ser extraordinariamente cautelosos a cerca de las posibles utilizaciones de este modelo.

Este ejemplo ilustra la importancia de analizar cuidadosamente los residuos del modelo estimado.

3.3 Validación del Modelo Mediante los Residuos.

Es frecuente que la muestra disponible contenga únicamente un valor de "y" para cada "x", y por lo tanto, los contrastes básicos de linealidad, homoscedasticidad y normalidad de las distribuciones condicionadas no pueden realizarse a priori; entonces la validación del modelo hay que hacerla sobre los residuos. Se verá a continuación el efecto del incumplimiento de cada supuesto sobre el modelo, y la forma de contrastarlos.

3.3.1 Linealidad.

El supuesto de linealidad establece el rango de valores observados para las variables: es decir que la media de la variable dependiente crece linealmente con la variable independiente. Es importante tener en cuenta que sólo se puede contrastar la linealidad en el rango de valores observados de las variables y que esto no implica que la linealidad se mantenga para otros posibles valores no incluidos en la muestra. Para comprobar la linealidad, además del gráfico de dispersión de las variables se debe hacer un gráfico de los residuos frente a los valores estimados. Cuando se detecta falta de linealidad[2] el modelo es inadecuado y conducirá a malas predicciones.

[2] No linealidad: La relación entre las variables independientes y la dependiente no es lineal.

El incumplimiento del supuesto de linealidad suele denominarse error de especificación, algunos ejemplos son: omisión de variables independientes importantes, inclusión de variables independientes irrelevantes.

3.3.2 Homoscedasticidad.

Para cada valor de la variable independiente o combinación de valores de las variables independientes, la varianza de los residuos es constante.

Si la varianza de los errores es muy diferente para unos valores de la variable explicativa que para otros, se tiene heteroscedasticidad, y las varianzas calculadas para los estimadores son erróneas. Además, los estimadores por Mínimos Cuadrados o Máxima Verosimilitud no son buenos estimadores, porque no tienen en cuenta la distinta precisión de los datos. Si la varianza de los errores varía aleatoriamente de unas partes a otras, el efecto de este tipo de heteroscedasticidad puede ser pequeño. Sin embargo, cuando hay pautas sistemáticas de variación en la variabilidad, se deben tener en cuenta para mejorar el modelo.

3.3.3 Normalidad.

El supuesto de normalidad es necesario para justificar el método de estimación y las distribuciones de los estimadores. Los efectos de la falta de normalidad dependen crucialmente de si la distribución que generan las perturbaciones tiene alta kurtosis[3] (colas pesadas) o no. Las distribuciones con alta kurtosis o colas pesadas pueden generar

[3] Kurtosis es una medida de la presencia de los valores extremos de la distribución.

con la probabilidad apreciables datos que se apartan más de 4 ó 5 desviaciones típicas de la media de la distribución. Si la distribución es aproximadamente simétrica y con colas similares o menos pesadas que la normal, el efecto de la falta de normalidad sobre el modelo de regresión es muy pequeño y los resultados obtenidos bajo normalidad son aproximadamente correctos. Sin embargo, cuando la distribución tiene colas pesadas, el efecto de la estimación de los parámetros de los valores extremos o atípicos puede ser muy grande. Entonces el Método de Mínimos Cuadrados o Máxima Verosimilitud (suponiendo normalidad) es un mal procedimiento de estimación: es decir los estimadores tienen varianza mucho mayor que la calculada bajo Mínimos Cuadrados y los intervalos y contrastes serán invalidados.

La normalidad de los residuos puede contrastarse gráficamente representando su distribución acumulada en papel probabilístico normal, el gráfico resultante se denomina gráfico probabilístico normal de los residuos, y, si la distribución de los residuos es normal, el gráfico tiene que mostrar aproximadamente una línea recta.

Existe normalidad en los residuos si su media es cero y la varianza es constante.

3.3.4 Independencia.

La dependencia temporal del error aleatorio es esperable cuando los datos de las variables correspondan a una serie temporal. Por ejemplo, si relacionamos las ventas de helados cada mes con la temperatura del mes, la secuencia temporal de los datos es

importante y no se tiene en cuenta en el modelo de regresión, que es invariante ante permutaciones de los datos.

Cuando los datos corresponden al mismo momento temporal (se dice entonces que se tiene una muestra de corte transversal) es esperable que las perturbaciones sean independientes.

Los residuos son independientes entre sí, es decir, los residuos constituyen una variable aleatoria, recuérdese que los residuos son las diferencias entre los valores observados y los pronosticados. Es frecuente encontrarse con residuos autocorrelacionados cuando se trabaja con series temporales.

Ejemplo 1: Con la información de los 14 estudiantes del ejemplo 1 (número de Horas de estudio "x" y la Calificación obtenida en dicho examen "y") del Capítulo 2, se realiza el análisis de los residuos para el cual se obtuvo la siguiente recta de regresión.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad \textbf{(3.1)}$$
$$\hat{y}_i = 2.356 + 1.215 x_i$$

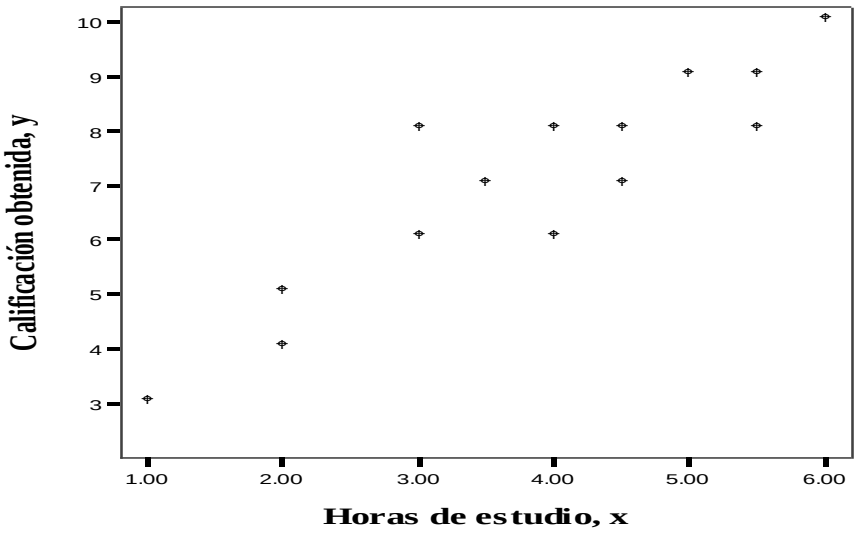
Calificación = 2.356 + 1.215 (Horas de estudio)

Tabla 3.2 Datos de Horas de estudio, Calificación obtenida, estimación y residuos.

n	x_i	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	e_i^2
1	1	3	3.571	-0.571	0.326
2	2	4	4.786	-0.786	0.618
3	2	5	4.786	0.214	0.046
4	3	6	6.001	-0.001	0.000
5	3	8	6.001	1.999	3.996
6	3.5	7	6.6085	0.392	0.153
7	4	8	7.216	0.784	0.615
8	4	6	7.216	-1.216	1.479
9	4.5	7	7.8235	-0.824	0.678
10	4.5	8	7.8235	0.177	0.031
11	5	9	8.431	0.569	0.324
12	5.5	8	9.0385	-1.039	1.078
13	5.5	9	9.0385	-0.038	0.001
14	6	10	9.646	0.354	0.125
Sumas	$\sum_{i=1}^{n} x_i = 53.5$	$\sum_{i=1}^{n} y_i = 98$	$\sum_{i=1}^{n} \hat{y}_i = 97.9865$	$\sum_{i=1}^{n} e_i = 0.013$	$\sum_{i=1}^{n} e_i^2 = 9.47$

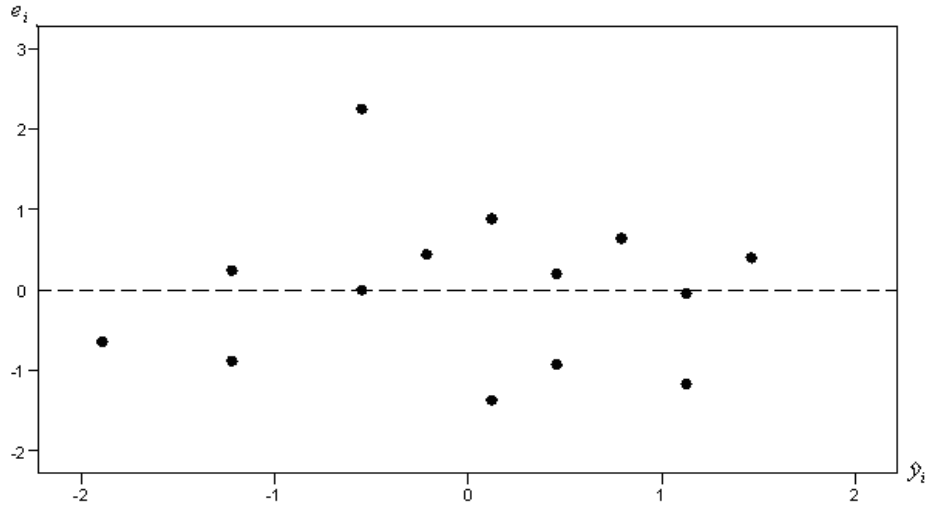
El diagrama de dispersión obtenido para estos datos es el siguiente:

Figura 3.3 Diagrama de dispersión para las 14 observaciones.



De acuerdo con la forma que tiene la figura 3.3 se puede ver que se cumple el supuesto de linealidad de los datos, a medida que aumentan los valores de la variable "x" también lo hace la variable "y". Para comprobar el supuesto de linealidad además del diagrama de dispersión, se muestra el gráfico de los residuos frente a los valores estimados.

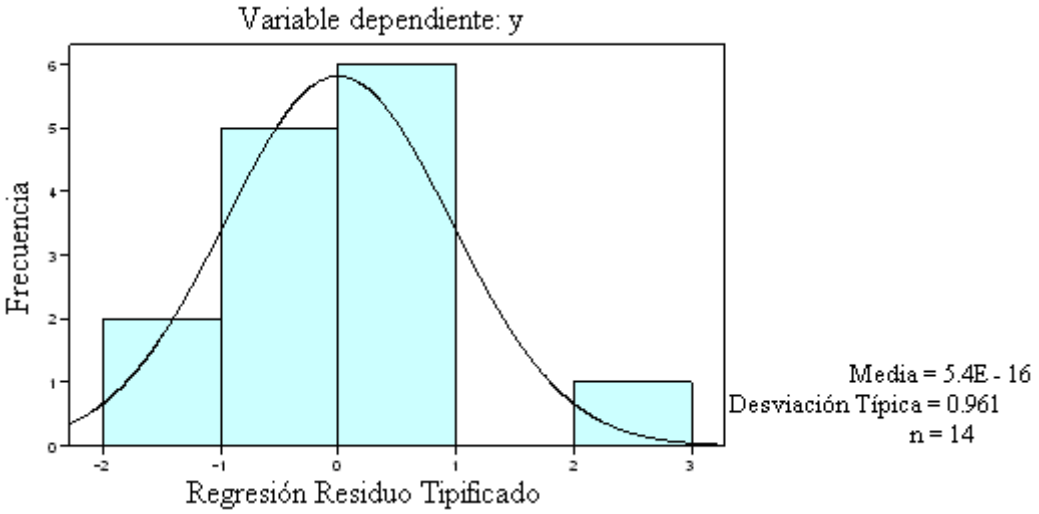
Figura 3.4 Gráfico de los residuos frente a los valores estimados.



En el gráfico de los residuos se puede observar que la nube de puntos no sigue ningún tipo de estructura, de manera que se puede decir que tiene sentido la regresión hecha sobre la muestra. En las figuras 3.3 y 3.4 se comprobaron los supuestos de linealidad y homoscedasticidad.

Para comprobar el supuesto de normalidad se hace el histograma de los residuos con una curva normal superpuesta; como se muestra en la figura 3.5.

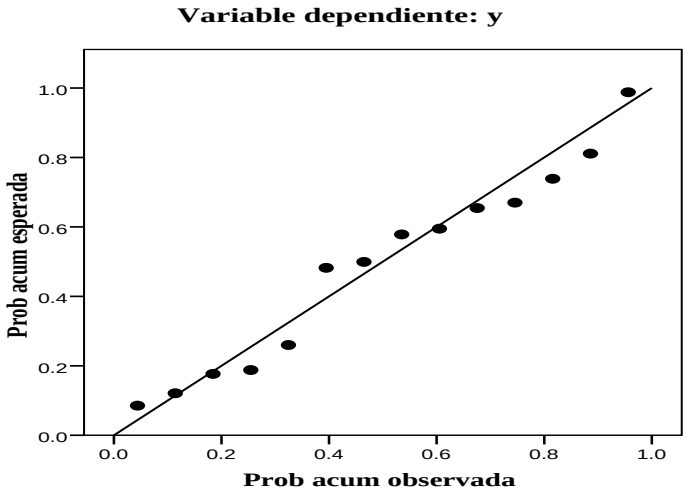
Figura 3.5 Histograma de los residuos.



La curva se construye tomando una media de 0 y una desviación típica de aproximadamente 1, como se ve en el gráfico; es decir la misma media y la misma desviación típica que los residuos tipificados.

Para comprobar el supuesto de normalidad también se muestra el gráfico probabilístico normal de los residuos.

Figura 3.6 Gráfico de probabilidad normal de los residuos.



El gráfico de probabilidad normal de la figura 3.6 muestra información similar a la obtenida en el histograma. Como se tiene que la distribución de los residuos es aproximadamente normal, se puede observar que los puntos se aproximan a la recta.

El supuesto de independencia entre los residuos se cumple dado que los datos corresponden al mismo momento temporal, pero también se puede comprobar el grado de independencia, con el estadístico d de Durbin-Watson (1951) que se define como sigue:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad \textbf{(3.2)}$$

El estadístico d oscila entre 0 y 4, y toma el valor de 2 cuando los residuos son independientes. Los valores menores que 2 indican autocorrelación positiva y los valores mayores que 2 autocorrelación negativa. Podemos asumir independencia entre los residuos cuando el estadístico d toma valores entre 1.5 y 2.5.

Para nuestro ejemplo el estadístico d es el siguiente:

$$d = \frac{(0.786 - (-0.571))^2 + (0.214 - (-0.786))^2 + \ldots + (0 - (0.354))^2}{0.571^2 + 0.786^2 + \ldots + 0.354^2} = \frac{17.005}{9.47} = 1.79$$

Dado que el valor de d = 1.79 se encuentra entre 1.5 y 2.5 podemos asumir que los residuos son independientes.

Nota: El análisis de los residuos se puede realizar haciendo uso del paquete estadístico SPSS v15.0 como se muestra al final de este Capítulo.

3.4 Predicción Usando el Modelo.

Con base en los datos muestrales de la tabla 2.1 (observaciones de 14 estudiantes) se obtiene la siguiente regresión muestral:

$$\hat{y}_i = 2.356 + 1.215\, x_i \qquad \textbf{(3.3)}$$

Donde:

$\hat{y}_i$: Es el estimador del verdadero $E(y_i)$ correspondiente a un "x" dado.

Una aplicación de la regresión muestral consiste en predecir sobre el futuro de "y" para algún valor dado de "x".

Existen dos clases de predicciones:

1. Predicción del valor medio condicional de "y" correspondiente a un determinado "x".
2. Predicción de un valor individual de "y" correspondiente a x_0.

A estas dos predicciones se les llama predicción media y predicción individual.

3.4.1 Predicción Media.

Para concretar los conceptos, supongamos que $x_0 = 3.5$ y que se quiere predecir $E(y_0 \mid x_0 = 3.5)$. Ahora es posible mostrar que la regresión muestral ecuación (3.3) proporciona la estimación puntual de esta predicción media de la siguiente manera:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$
$$\hat{y}_0 = 2.356 + 1.215(3.5) \qquad \textbf{(3.4)}$$
$$\hat{y}_0 = 6.6085$$

Donde:

$\hat{y}_0$: Es el estimador de $E(y_0 \mid x_0)$.

Como $\hat{y}_0$ es un estimador; no es extraño que sea diferente de su verdadero valor, la diferencia entre los dos valores $(y_0 - \hat{y}_0)$ nos da una idea de la fiabilidad de la predicción.

Para estimar este error se necesita encontrar la distribución muestral de $\hat{y}_0$. También es posible mostrar que $\hat{y}_0$ es una variable aleatoria que está normalmente distribuida, con media $(\beta_0 + \beta_1 x_0)$, y varianza dada por la siguiente ecuación:

$$\text{var}(\hat{y}_0) = \text{var}(\hat{\beta}_0 + \hat{\beta}_1 x_0)$$
$$\text{var}(\hat{y}_0) = \text{var}(\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_0)$$
$$\text{var}(\hat{y}_0) = \text{var}[\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})]$$
$$\text{var}(\hat{y}_0) = \text{var}(\bar{y}) + \text{var}[(x_0 - \bar{x})\hat{\beta}_1]$$
$$\text{var}(\hat{y}_0) = \text{var}(\bar{y}) + (x_0 - \bar{x})^2 \, \text{var}(\hat{\beta}_1)$$
$$\text{var}(\hat{y}_0) = \frac{\sigma^2}{n} + (x_0 - \bar{x})^2 \frac{\sigma^2}{S_{xx}}$$

$$\text{var}(\hat{y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \qquad \textbf{(3.5)}$$

Reemplazando σ^2 por su estimador insesgado $\hat{\sigma}^2$ y haciendo uso de la ecuación dada en el apéndice 2.1 j) del Capitulo 2 se tiene que la variable

$$t = \frac{\hat{y}_0 - E(y \mid x_0)}{\sqrt{var(\hat{y}_0)}} = \frac{\hat{y}_0 - (\beta_0 + \beta_1 x_0)}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} = \frac{\hat{y}_0 - (\beta_0 + \beta_1 x_0)}{es(\hat{y}_0)} \qquad \textbf{(3.6)}$$

Sigue la distribución t con n-2 grados de libertad. La distribución t puede por lo tanto, emplearse para encontrar intervalos de confianza del verdadero $E(y_0 \mid x_0)$, y para hacer pruebas de hipótesis a cerca del mencionado valor en la forma usual.

Un intervalo de confianza de 100(1-α) por ciento para la respuesta media en el punto $x = x_0$ es:

$$\begin{gathered} \hat{y}_0 - t_{(\alpha/2,\ n-2)} es(\hat{y}_0) \le E(y_0 \mid x_0) \le \hat{y}_0 + t_{(\alpha/2,\ n-2)} es(\hat{y}_0) \\ \hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{(\alpha/2,\ n-2)} es(\hat{y}_0) \le \beta_0 + \beta_1 x_0 \le \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{(\alpha/2,\ n-2)} es(\hat{y}_0) \end{gathered} \qquad \textbf{(3.7)}$$

Ejemplo 2: Haciendo uso de los datos obtenidos en el ejemplo 1 del Capítulo 2 se tiene:

Datos:

$x_0 = 3.5$, $\hat{\sigma}^2 = 0.789$, $n = 14$, $S_{xx} = 28.803$, $\bar{x} = 3.8214$, $t_{(0.05/2,\ 14-2)} = t_{(0.025,\ 12)} = 2.179$,

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$\hat{y}_0 = 2.356 + 1.215(3.5)$$

$$\hat{y}_0 = 6.6085$$

$$es(\hat{y}_0) = \sqrt{(0.789)\left(\frac{1}{14} + \frac{(3.5 - 3.8214)^2}{28.803}\right)}$$

$$es(\hat{y}_0) = 0.243$$

Por lo tanto, el intervalo de confianza del 95%, para el verdadero valor $E(y_0 \mid x_0) = \beta_0 + \beta_1 x_0$ está dado por:

$$6.6085 - 2.179(0.243) \le E(y_0 \mid x_0 = 3.5) \le 6.6085 + 2.179(0.243) \qquad \textbf{(3.8)}$$

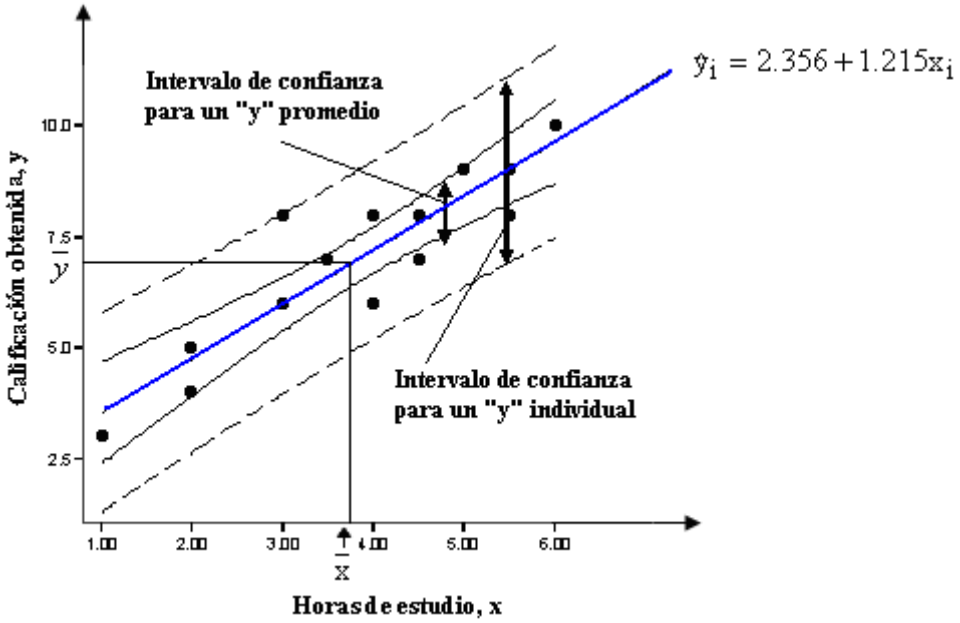
Es decir,

$$6.079 \leq E(y_0 \mid x_0 = 3.5) \leq 7.139$$

De este modo dado un x_0 = 3.5, con muestras repetidas en 95 de cada 100 intervalos como el de la ecuación (3.8) estará incluido el verdadero valor medio; la mejor estimación de este valor medio verdadero es obviamente la estimación puntual 6.6085.

Si obtenemos intervalos de confianza del 95%, como el de ecuación (3.8), para cada uno de los "x" dados en la tabla 2.1, hallaremos lo que se conoce como intervalo de confianza o banda de confianza para la función de regresión poblacional que se muestra en la figura 3.7.

Figura 3.7 Intervalos de confianza para el promedio "y" y para un "y" individual.



3.4.2 Predicción Individual.

Si nos proponemos predecir un valor individual de "y" como y_0, que corresponde a un valor dado de "x" como x_0, es posible probar que el mejor estimador lineal insesgado de y_0 está también dado por la ecuación (3.4), pero su varianza será:

$$\text{var}(y_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \qquad \textbf{(3.9)}$$

Reemplazando σ^2 por su estimador insesgado $\hat{\sigma}^2$ y haciendo uso de la ecuación dada en el apéndice 2.1 j) del Capitulo 2 se tiene que la variable

$$t = \frac{y_0 - \hat{y}_0}{\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}} = \frac{y_0 - \hat{y}_0}{es(y_0)} \qquad \textbf{(3.10)}$$

También sigue la distribución t. Por consiguiente la distribución t puede utilizarse para hacer inferencias a cerca del verdadero valor y_0.

Así, el intervalo de confianza de 100(1-α) por ciento para y_0 en el punto $x = x_0$ es:

$$\begin{aligned} &\hat{y}_0 - t_{(\alpha/2,\ n-2)}es(y_0) \le y_0 \mid x_0 \le \hat{y}_0 + t_{(\alpha/2,\ n-2)}es(y_0) \\ &\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{(\alpha/2,\ n-2)}es(y_0) \le y_0 \mid x_0 \le \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{(\alpha/2,\ n-2)}es(y_0) \end{aligned} \qquad \textbf{(3.11)}$$

Ejemplo 3: Haciendo uso de los datos obtenidos en el ejemplo 1 del Capitulo 2 se tiene:

Datos:

$x_0 = 3.5$, $\hat{\sigma}^2 = 0.789$, $n = 14$, $S_{xx} = 28.803$, $\bar{x} = 3.8214$, $t_{(0.05/2,\ 14-2)} = t_{(0.025,\ 12)} = 2.179$,

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$\hat{y}_0 = 2.356 + 1.215(3.5)\ y$$

$$\hat{y}_0 = 6.6085$$

$$es(y_0) = \sqrt{(0.789)\left(1 + \frac{1}{14} + \frac{(3.5 - 3.8214)^2}{28.803}\right)}$$

$$es(y_0) = 0.920$$

Sustituyendo en la ecuación (3.11) se tiene el intervalo de confianza del 95% para y_0 correspondiente a $x_0 = 3.5$ será:

$$\begin{gathered} 6.6085 - 2.179(0.920) \le y_0 \mid x_0 = 3.5 \le 6.6085 + 2.179(0.920) \\ 4.6038 \le y_0 \mid x_0 = 3.5 \le 8.6132 \end{gathered} \qquad \textbf{(3.12)}$$

Comparando este intervalo con el de la ecuación (3.8), se puede ver que el intervalo de confianza para y_0 individual es más ancho que el intervalo de confianza para el valor medio de y_0.

Calculando intervalos de confianza como el de la ecuación (3.12) condicionales a los valores de "x" de la tabla 2.1, obtenemos una banda de confianza del 95% para los valores individuales de "y" que corresponden a los valores mencionados de "x". La banda de confianza para nuestros x_i individuales al igual que la banda para y_0 se representa en la figura 3.7.

Nótese que una característica importante de las bandas de confianza de la figura 3.7 es la amplitud (anchura) de las bandas es menor cuando $x_0 = \bar{x}$. Esto podría sugerir que la habilidad predictiva de la línea de regresión muestral decrece a medida que x_0 se separa progresivamente de $\bar{x}$. En conclusión, hay que ser muy cautelosos al "extrapolar" la línea de regresión cuando se trata de predecir $\hat{y}_0$ o un y_0 asociado con un x_0 dado, que esté más o menos lejos de la media muestral $\bar{x}$.

Ejercicios 3.

1. Para los datos del ejercicio 1 del Capítulo 2 hacer:

   a) La gráfica de los residuos.

   b) Análisis de los residuos.

2. Consideremos las observaciones de los Pesos y Alturas de un conjunto de 10 personas: el individuo 1 tiene 161 cm. de altura y 63 kg. de peso, el individuo 2 tiene 152 cm de altura y 56 kg de peso, etc., tal como se ve en la tabla siguiente:

Individuo	1	2	3	4	5	6	7	8	9	10
Altura cm. x	161	152	167	153	161	168	167	153	159	173
Peso kg. y	63	56	77	49	72	62	68	48	57	67

   c) Estimar la ecuación de regresión.

   d) Hacer el análisis de los residuos.

   e) Determinar el intervalo de confianza del 95% para la predicción media y para la predicción individual dado $x_0 = 162$.

3. Para los datos del ejercicio 4 del Capítulo 2 realizar:

   f) El análisis de los residuos.

   g) Determinar el intervalo de confianza del 95% para la predicción media y para la predicción individual dado $x_0 = 650$.

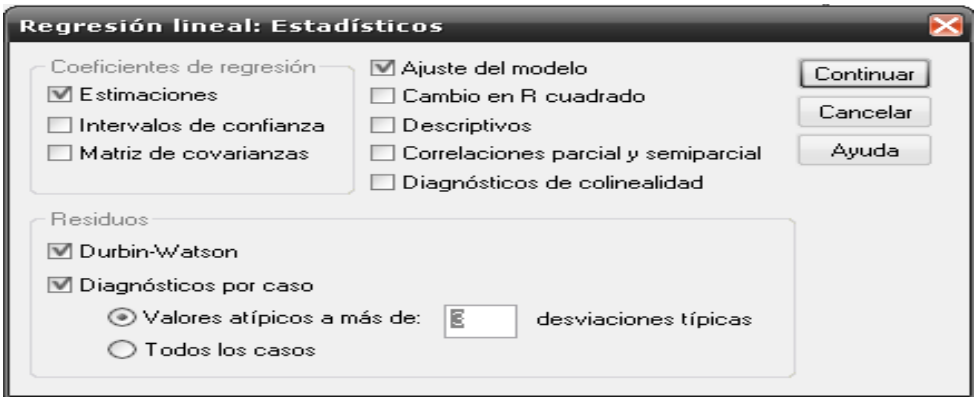
3.5 Análisis de los Residuos Haciendo uso del SPSS V15.0.

Llamamos residuos a las diferencias entre los valores observados y los pronosticados: $(y_i - \hat{y}_i)$.

Después de haber digitado los datos en el editor, se realiza el análisis siguiendo los pasos que se muestran a continuación:

Analizar ⟶ Regresión ⟶ Lineal, luego aparece el siguiente cuadro:

En el que se colocan las variables, haciendo click en el botón **Estadísticos** se obtiene el cuadro **Regresión lineal: Estadísticos**, como se muestra a continuación:



Por defecto, el SPSS lista los residuos que se alejan de cero a más de 3 desviaciones típicas, pero el usuario puede cambiar este valor introduciendo el valor deseado. Para obtener un listado de los residuos que se alejan de cero de por lo menos más de tres desviaciones típicas.

Haciendo click en la opción **Guardar** de la ventana **Regresión lineal** se obtiene la ventana siguiente:

En la cual se marca la opción **No tipificados** del recuadro **Residuos** y aceptando esas opciones se obtiene la tabla resumen que se presenta a continuación:

Estadísticos sobre los residuos[a]

	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	3.5716	9.6472	7.0000	1.80873	14
Residuo bruto	-1.21699	1.99814	.00000	.85352	14
Valor pronosticado tip.	-1.895	1.464	.000	1.000	14
Residuo tip.	-1.370	2.249	.000	.961	14

a. Variable dependiente: Calificación obtenida, y

Con información sobre el valor máximo y mínimo, la media y la desviación típica de los pronósticos, de los residuos, de los pronósticos tipificados y de los residuos tipificados.

Es especialmente importante señalar que la media de los residuos vale cero y la desviación típica de los residuos está acercándose a uno.

- **Independencia**

  Uno de los supuestos básicos del modelo de regresión lineal simple es el de independencia entre los residuos. El estadístico **Durbin-Watson** proporciona información sobre el grado de independencia existente entre ellos.

  En el cuadro **Regresión lineal: Estadísticos** se seleccionó la opción **Durbin-Watson** esta elección permite obtener la tabla que se muestra a continuación:

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	.904[a]	.818	.803	.88838	1.782

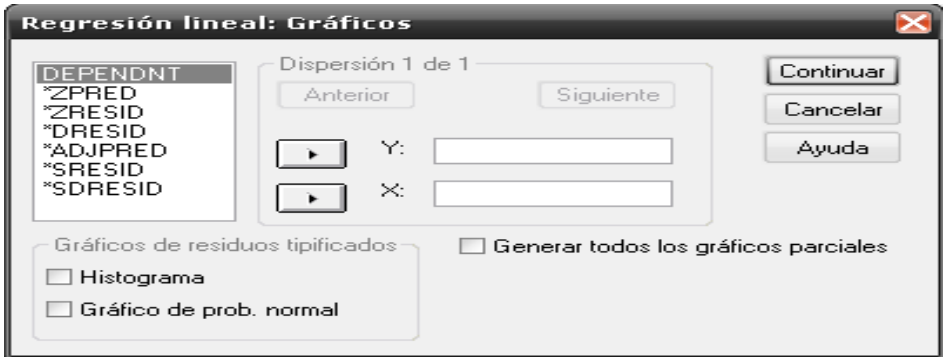
a. Variables predictoras: (Constante), Horas de estudio, x

b. Variable dependiente: Calificación obtenida, y

Como se dijo antes podemos asumir independencia entre los residuos cuando **Durbin-Watson** toma valores entre 1.5 y 2.5, en la tabla **Resumen del modelo** se observa que el valor es de 1.782 ≈ 1.79 que es el que se obtuvo al hacerlo a mano utilizando los residuos, por lo cual se puede decir que existe independencia entre los residuos.

- **Homoscedasticidad**

El procedimiento **Regresión lineal** dispone de una serie de gráficos que permiten, entre otras cosas, obtener información sobre el grado de cumplimiento de los supuestos de homoscedasticidad y normalidad de los residuos. Para utilizar estos gráficos en el cuadro **Regresión lineal** pulsamos el botón **gráficos** y se obtiene la ventana siguiente:



Las variables listadas permiten obtener diferentes gráficos de dispersión. Las variables precedidas por asterisco son las variables creadas por el SPSS.

ZRESID: (residuos eliminados o corregidos): residuos obtenidos al efectuar los pronósticos eliminando de la ecuación de regresión el caso sobre el que se efectúa el pronóstico.

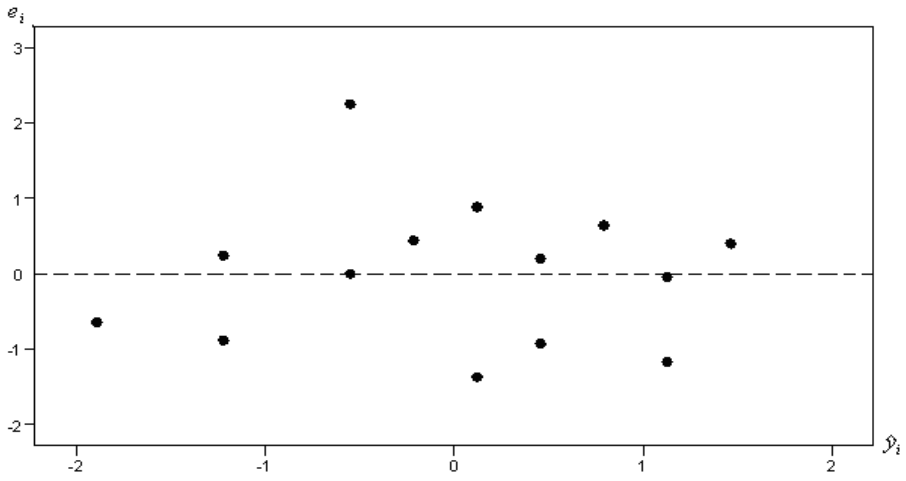
ZPRED (pronósticos tipificados): pronósticos divididos por su desviación típica.

Son pronósticos transformados en puntuaciones ≈ (media 0 y desviación típica 1).

Trasladar la variable **ZRESID** al cuadro **Y**: del recuadro Dispersión 1 de 1.

Trasladar la variable **ZPRED** al cuadro **X**: del recuadro Dispersión 1 de 1.

Aceptando estas elecciones el visor ofrece el diagrama de dispersión que se muestra en la figura siguiente:



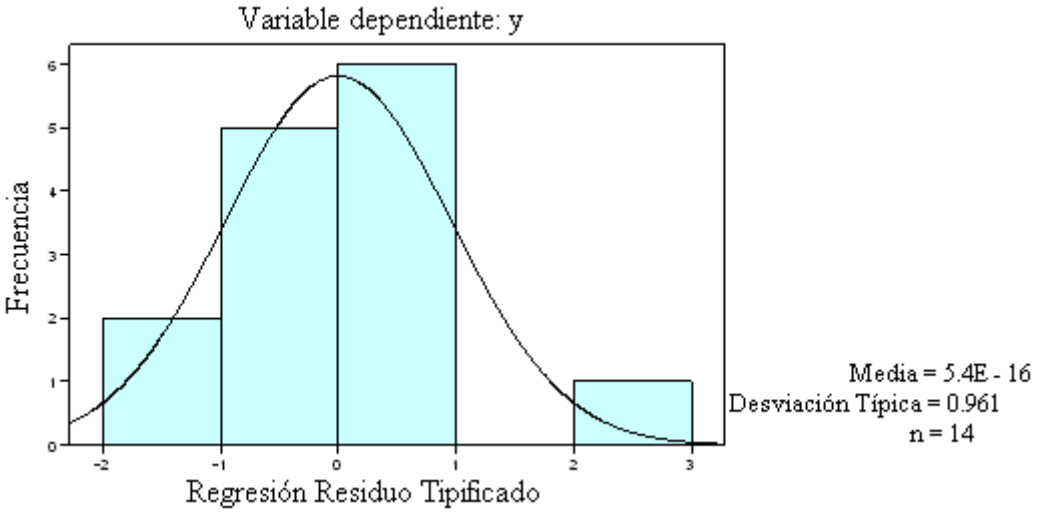
Observando el diagrama de dispersión podemos ver que no sigue ningún tipo de estructura, entonces, se puede decir que tiene sentido la regresión hecha sobre la muestra.

El diagrama de dispersión de las variables **ZPRED** y **ZRESID** posee la utilidad adicional de permitir detectar relaciones de tipo no lineal entre las variables. Si la relación es, de hecho, no lineal, el diagrama puede contener indicios sobre otro tipo de función de ajuste: por ejemplo, los residuos estandarizados podrían, en lugar de estar homogéneamente dispersos, seguir un trazado curvilíneo.

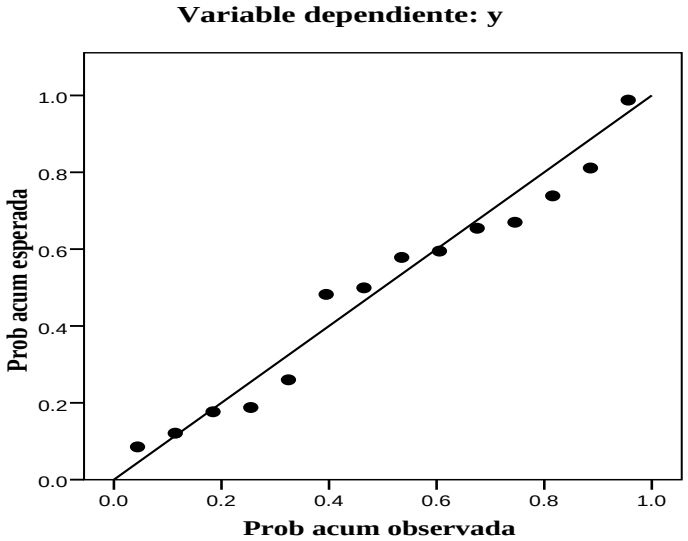
- **Normalidad**

El recuadro **Regresión lineal: Gráficos** contiene dos opciones gráficas que informan sobre el grado en el que los residuos tipificados se aproximan a una distribución normal.

Histograma: Ofrece un histograma de los residuos tipificados con una curva normal superpuesta como se muestra en la figura siguiente:



Según este gráfico se puede ver que los residuos son aproximadamente normales, pero además del histograma, se tiene el gráfico de probabilidad normal que se muestra a continuación:



En el que se puede observar que los puntos se aproximan a la diagonal, si la relación entre las variables fuera perfecta todos los puntos estarían sobre la línea, pero esos son casos remotos que cuando se trabaja con datos reales casi nunca se cumple.

- **Linealidad**

  Por último se tiene la linealidad, que se puede observar en el diagrama de dispersión como se mostró en la figura 3.3.

Capítulo 4

Modelo de Regresión Lineal Múltiple.

4.1 Introducción al Modelo de Regresión Lineal Múltiple.

El modelo de dos variables, que estudiamos en el Capítulo 2, es más bien inadecuado en la práctica. Por esta razón, necesitamos extender nuestro modelo simple con dos variables a un modelo que contenga más de dos variables. Esto nos conduce al estudio de los modelos de regresión múltiple, es decir, a los modelos en que la variable dependiente "y" depende de dos o más variables explicatorias.

El modelo de regresión múltiple más simple es el de la regresión de tres variables, una dependiente y dos explicatorias, en este Capítulo se estudiará este modelo y lo generalizaremos para más de tres variables en el Capítulo 5.

El procedimiento de Regresión Lineal permite utilizar más de una variable independiente, y, por tanto, permite llevar a cabo análisis de regresión múltiple, la ecuación de regresión ya no define una recta en el plano, si no un hiperplano en un espacio multidimensional.

Para el modelo de regresión múltiple se describen los supuestos que subyacen al modelo.

Además, la estimación de los parámetros se realiza por el método de Mínimos Cuadrados Ordinarios, y haciendo uso del algebra matricial para el caso de k variables.

En este Capítulo nos ocuparemos también, de la prueba de hipótesis y luego de la estimación por intervalo para modelos que incorporen tres variables.

4.2 Definición de Términos Básicos.

Coeficiente de Determinación Múltiple (R^2): Representa el porcentaje de variabilidad de "y" debida a la recta de regresión.

Coeficiente de Correlación Múltiple (R): Representa el porcentaje de variabilidad de "y" que explica el modelo de regresión.

Coeficiente de Correlación Parcial: Mide la asociación entre dos variables después de controlar los efectos de una o más variables adicionales.

Colinealidad: Es un problema del análisis de regresión y se da cuando las variables explicativas del modelo están relacionadas constituyendo una combinación lineal.

Diagrama de Dispersión Múltiple: También llamado hiperplano de regresión que pasa necesariamente por el punto $(\overline{y}, \overline{x}_1, \overline{x}_2)$.

Hipótesis Estadísticas: Es un enunciado acerca de la distribución de probabilidad de una variable aleatoria. Las hipótesis estadísticas a menudo involucran una o más características de la distribución, como por ejemplo forma o independencia de la variable aleatoria.

Multicolinealidad: Problema estadístico que se presenta en el análisis de regresión múltiple, en el que la confiabilidad de los coeficientes de regresión se ve reducida debido a un alto nivel de correlación entre las variables independientes.

No Multicolinealidad: Ocurre cuando las variables explicativas del modelo no están correlacionadas.

4.3 Asunciones del Modelo de tres Variables.

Al generalizar la función de regresión poblacional con dos variables (FRP) ecuación (1.20), podemos escribir la FRP para tres variables como sigue:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \qquad \textbf{(4.1)}$$

Donde:

y: Es la variable dependiente.

x_1 y x_2: Las variables explicatorias.

ε: El término del error.

i: La i-ésima observación.

Dentro del esquema del modelo de regresión lineal presentado en el Capítulo 2, específicamente suponemos que:

$$E(\varepsilon_i) = 0 \text{ Para cada i} \qquad \textbf{(4.2)}$$

$$E(\varepsilon_i \varepsilon_j) = 0 \qquad i \neq j \qquad \textbf{(4.3)}$$

$$\text{Var}(\varepsilon_i) = \sigma^2 \text{ Para cada i} \qquad \textbf{(4.4)}$$

$$E(\varepsilon_i, x_1) = E(\varepsilon_i, x_2) = 0 \qquad \textbf{(4.5)}$$

A esta lista añadimos ahora otro supuesto que denominamos el supuesto de no multicolinealidad, que significa que no existe una relación lineal exacta entre las variables explicatorias. Formalmente, no multicolinealidad significa que no existe un conjunto de números λ_1 y λ_2, distintos de cero, tales que:

$$\lambda_1 x_1 + \lambda_2 x_2 = 0 \qquad \textbf{(4.6)}$$

Si tal relación lineal existe, entonces, se dice que x_1 y x_2 son colineales o linealmente dependientes. De otra forma, si la ecuación (4.6) se cumple sólo cuando $\lambda_1 = \lambda_2 = 0$, entonces, se dice que x_1 y x_2 son linealmente independientes.

El supuesto de no multicolinealidad requiere que en la función de regresión poblacional teórica se incluyan únicamente aquellas variables que no sean funciones lineales de algunas de las variables del modelo.

4.4 Interpretación de la Ecuación de Regresión Lineal Múltiple.

De los supuestos del modelo de regresión clásico, se deduce que, tomando el valor condicional esperado de "y" en ambos lados de la ecuación (4.1), obtendremos:

$$E(y_i \mid x_1, x_2) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \qquad \textbf{(4.7)}$$

O sea que, la ecuación (4.7) nos da la media condicional o valor esperado de "y" condicionado por los valores fijos o dados de x_1 y x_2. Por lo tanto, como en el caso de dos variables, el análisis de regresión múltiple es un análisis de regresión condicional; condicional en los valores fijos de las variables explicatorias, y lo que obtenemos es el promedio o valor medio de "y" para los valores fijos de las variables x_i.

4.5 Significado de los Coeficientes de Regresión Parcial.

El significado de los coeficientes de regresión parcial es el siguiente:

β_0: Se puede interpretar como el valor medio de "y" cuando las x_i son cero.

β_1: Mide el cambio en el valor medio de "y", $E(y_i \mid x_1, x_2)$ por cambio de una unidad en x_1, manteniéndose x_2 constante. En otras palabras nos da la pendiente de $E(y_i \mid x_1, x_2)$ con respecto a x_1, manteniéndose x_2 constante.

β_2: Mide el cambio en el valor medio de "y" por unidad de cambio en x_2, manteniéndose x_1 constante.

4.6 Estimación de los Coeficientes de Regresión Parciales por Mínimos Cuadrados Ordinarios (MCO).

Para estimar los parámetros del modelo de regresión con tres variables, ecuación (4.1), usamos el método de Mínimos Cuadrados Ordinarios visto en el Capítulo 2.

4.6.1 Estimadores de MCO.

Para encontrar los estimadores de MCO, escribimos primero la función de regresión muestral (FRM) correspondiente a la FRP de la ecuación (4.1) como sigue:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i \qquad \textbf{(4.8)}$$

$$y_i = \hat{y}_i + e_i \qquad \textbf{(4.9)}$$

$$y_i - \hat{y}_i = e_i \qquad \textbf{(4.10)}$$

Donde:

e_i : Es el término residual.

El procedimiento MCO consiste en buscar los valores de los parámetros desconocidos, de tal forma que la suma residual de cuadrados sea tan pequeña como sea posible. Simbólicamente, lo que se quiere es:

$$\min \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 \qquad \textbf{(4.11)}$$

Donde SS_{Res} se obtiene por manipulación algebraica de la ecuación (4.8), derivando con respecto a las variables desconocidas, igualando las expresiones resultantes a cero y resolviéndolas simultáneamente se obtiene:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 \qquad \textbf{(4.12)}$$

$$\sum_{i=1}^{n} y_i x_{1i} = \hat{\beta}_0 \sum_{i=1}^{n} x_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{1i} x_{2i} \qquad \textbf{(4.13)}$$

$$\sum_{i=1}^{n} y_i x_{2i} = \hat{\beta}_0 \sum_{i=1}^{n} x_{2i} + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} x_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i}^2 \qquad \textbf{(4.14)}$$

De la ecuación (4.12) se ve claramente que:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 \qquad \textbf{(4.15)}$$

Que es el estimador de MCO del intercepto poblacional β_0.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2} \qquad \textbf{(4.16)}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)(y_i-\bar{y})\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2 - \sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(y_i-\bar{y})\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2 - \left(\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)\right)^2} \quad \textbf{(4.17)}$$

La deducción de las ecuaciones (4.15), (4.16) y (4.17) se muestran en el apéndice **4.1 a), b) y c).**

Las ecuaciones (4.16) y (4.17) nos dan los estimadores MCO de los coeficientes de regresión parcial poblacional, β_1 y β_2 respectivamente.

Los valores de los coeficientes se pueden obtener también encontrando las sumatorias, sustituyéndolas en las ecuaciones (4.12), (4.13) y (4.14) y luego simultaneando las ecuaciones para despejar los coeficientes.

Recapitulando se tiene que:

1. Las ecuaciones (4.16) y (4.17) son de naturaleza simétrica pues una puede obtenerse a partir de la otra intercambiando los papeles de x_1 y de x_2.
2. Los denominadores de estas ecuaciones son idénticos.
3. El caso de tres variables es una extensión natural del de dos variables.

4.6.2 Varianza y Errores Estándar de los Estimadores de MCO.

Una vez obtenidos los estimadores de los coeficientes de regresión parciales, se pueden encontrar las varianzas y los errores estándar de estos estimadores en la forma indicada en el Capítulo 2 apéndice 2.1 e). Como en el caso de dos variables, necesitamos los errores estándar para dos propósitos:

- Para probar hipótesis estadísticas.
- Para establecer los intervalos de confianza.

Las ecuaciones son como sigue:

$$\text{var}(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{\bar{x}_1^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 + \bar{x}_2^2 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 - 2\bar{x}_1 \bar{x}_2 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2} \right] * \sigma^2 \quad \textbf{(4.18)}$$

$$\text{es}(\hat{\beta}_0) = \sqrt{\text{var}(\hat{\beta}_0)} \quad \textbf{(4.19)}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2} * \sigma^2 \quad \textbf{(4.20)}$$

$$\text{es}(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)} \quad \textbf{(4.21)}$$

$$\text{var}(\hat{\beta}_2) = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2} * \sigma^2 \quad \textbf{(4.22)}$$

$$\text{es}(\hat{\beta}_2) = \sqrt{\text{var}(\hat{\beta}_2)} \quad \textbf{(4.23)}$$

Donde σ^2 es la varianza (homoscedástica) de los errores poblacionales ε_i.

Un estimador insesgado de σ^2 está dado por:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-3} \quad \textbf{(4.24)}$$

Observe la similitud de este estimador de σ^2 con el de dos variables $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$;

Para el caso de la regresión múltiple con tres variables los grados de libertad son n-3, pues al estimar $\sum_{i=1}^{n} e_i^2$ debemos estimar primero β_0, β_1 y β_2 lo cual consume tres grados de libertad.

El estimador $\hat{\sigma}^2$ puede calcularse a partir de la ecuación (4.24), una vez que los residuos e_i estén disponibles, pero puede también obtenerse más rápidamente usando la siguiente relación:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) - \hat{\beta}_2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \qquad \textbf{(4.25)}$$

La deducción de este resultado se muestra en el apéndice **4.1 d).**

4.6.3 Propiedades de los Estimadores de MCO.

Los estimadores de los coeficientes de regresión parcial de MCO satisfacen el teorema de Gauss-Markov, el cual establece que de todos los estimadores lineales insesgados, los de MCO tienen la mínima varianza.

A propósito, vale la pena anotar los siguientes aspectos de la función de regresión muestral ecuación (4.8).

1. Como en el caso de dos variables, la línea (superficie) de regresión de tres variables pasa por las medias $\bar{y}$, $\bar{x}_1$ y $\bar{x}_2$. Esto se deduce fácilmente de la ecuación (4.12).

2. El valor medio de $\hat{y}_i$ es igual al valor medio de los valores observados y_i, lo cual se puede ver fácilmente:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$
$$\hat{y}_i = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$
$$\frac{\sum_{i=1}^{n} \hat{y}_i}{n} = \frac{\sum_{i=1}^{n} \bar{y}}{n} - \frac{\hat{\beta}_1 \sum_{i=1}^{n} \bar{x}_1}{n} - \frac{\hat{\beta}_2 \sum_{i=1}^{n} \bar{x}_2}{n} + \frac{\hat{\beta}_1 \sum_{i=1}^{n} x_{1i}}{n} + \frac{\hat{\beta}_2 \sum_{i=1}^{n} x_{2i}}{n}$$
$$\frac{\sum_{i=1}^{n} \hat{y}_i}{n} = \frac{n\bar{y}}{n} - \frac{\hat{\beta}_1 n\bar{x}_1}{n} - \frac{\hat{\beta}_2 n\bar{x}_2}{n} + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2$$
$$\bar{\hat{y}}_i = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2$$

$$\bar{\hat{y}}_i = \bar{y} \qquad \textbf{(4.26)}$$

3. $\sum_{i=1}^{n} e_i = \bar{e} = 0$ (la sumatoria de los errores es aproximadamente cero, entonces, la media es cero).

4. Los residuos e_i no están correlacionados con $\hat{y}_i$, es decir $\sum_{i=1}^{n} e_i \hat{y}_i = 0$

5. Los residuos e_i no están correlacionados con x_1 ni con x_2, es decir, $\sum_{i=1}^{n} e_i x_1 = \sum_{i=1}^{n} e_i x_2 = 0$

6. Como se vio en el Capítulo 2, para las pruebas de hipótesis suponemos que los errores ε_i están distribuidos normalmente con media cero y varianza σ^2 con este supuesto los estimadores $\hat{\beta}_0, \hat{\beta}_1$ y $\hat{\beta}_2$ están también distribuidos normalmente con medias iguales a β_0, β_1 y β_2 respectivamente y con las varianzas dadas anteriormente.

7. Siguiendo la lógica del modelo de dos variables dado en el Capítulo 2, bajo los supuestos de normalidad puede demostrarse que $(n-3)\hat{\sigma}^2/\sigma^2$ sigue la distribución ji-cuadrada (χ^2) con n-3 grados de libertad, esto nos permite hacer pruebas de hipótesis a cerca del verdadero valor de σ^2.

En el Capítulo 2, se anotó que, bajo los supuestos de normalidad, los estimadores de MCO y MV de los coeficientes de regresión del modelo de dos variables son idénticos. Esta igualdad se extiende a otros modelos que contenga cualquier número de variables. Las pruebas de esta afirmación se encuentran en el Capítulo 2 apéndice 2.1. No obstante, esto no se cumple para el estimador de σ^2. Se puede demostrar que el estimador MV de σ^2 es: $\frac{\sum_{i=1}^{n} e_i^2}{n}$ independiente del número de variables del modelo, mientras que el estimador de MCO de σ^2 es: $\frac{\sum_{i=1}^{n} e_i^2}{n-2}$ en el caso de dos variables, $\frac{\sum_{i=1}^{n} e_i^2}{n-3}$ en el caso de tres variables y $\frac{\sum_{i=1}^{n} e_i^2}{n-k}$ en el caso del modelo con k variables. En otras palabras, el estimador de MCO de σ^2 tiene en cuenta el número de grados de libertad, mientras que el estimador de σ^2 de MV no lo hace. Naturalmente, si **n** es muy grande los estimadores de σ^2 de MCO y de MV tienden a ser iguales.

4.7 Coeficiente de Determinación Múltiple R^2 y el Coeficiente de Correlación Múltiple R.

En el caso de dos variables vimos que r^2 definido como la ecuación (2.33) mide la bondad de ajuste de la ecuación de regresión; es decir, nos da la proporción o porcentaje de variación total en la variable dependiente "y" explicada por la variable "x". Esta definición de r^2 puede fácilmente extenderse a modelos de regresión de más de dos variables. Por consiguiente, en el modelo de tres variables estamos interesados en conocer la proporción de la variación en "y" explicada conjuntamente por las variables x_1 y x_2. El valor que nos da esta información se conoce como el coeficiente de determinación múltiple y se denota con R^2; conceptualmente es igual a r^2.

Para encontrar el R^2 se puede seguir el procedimiento siguiente:

Para cada observación, podemos descomponer la diferencia entre y_i, y su media $\bar{y}$ como sigue:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Elevando al cuadrado ambos lados y aplicando sumatorias obtenemos:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad \textbf{(4.27)}$$

Variación en y = Variación residual + Variación explicada

La deducción del resultado anterior se muestra en el apéndice **4.1 e)**

Usando la terminología introducida en el Capítulo 2:

$$SS_T = SS_{Res} + SS_R$$

Dividiendo ambos lados de la ecuación por SS_T se tiene:

$$\frac{SS_T}{SS_T} = \frac{SS_R}{SS_T} + \frac{SS_{Res}}{SS_T}$$

$$1 = \frac{SS_R}{SS_T} + \frac{SS_{Res}}{SS_T}$$

$$1 - \frac{SS_{Res}}{SS_T} = \frac{SS_R}{SS_T}$$

El R^2 mide la proporción de la variación en "y", que es explicada por la ecuación de regresión múltiple, se define como el cociente de la suma de cuadrados debida a la regresión entre la suma de cuadrados totales y se denota de la siguiente forma:

$$R^2 = \frac{SS_R}{SS_T}$$

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad \textbf{(4.28)}$$

Dado que los valores en la ecuación (4.28) son generalmente calculados en forma rutinaria, R^2 puede calcularse fácilmente. Note que R^2 al igual que r^2 está comprendido entre 0 y 1. Si es 1, significa que la línea de regresión ajustada explica el ciento por ciento de la variación en "y". De otra forma si es cero, el modelo no explica nada de las variaciones en "y".

Se dice que el ajuste del modelo es "mejor" mientras más cerca de 1 esté el R^2.

Recuerde que en el caso de dos variables definimos el valor r como el coeficiente de correlación e indicamos que medía el grado de asociación (lineal) entre dos variables.

El análogo de r en el caso de tres o más variables es el coeficiente de correlación múltiple, denotado por R, y es una medida del grado de asociación entre "y" y todas las variables explicatorias conjuntamente. Aunque r puede ser positivo o negativo, R siempre es positivo. En la práctica, R tiene poca importancia. El más significativo es R^2.

4.7.1 Comparación de Dos o Más Valores de R^2: El R^2 Ajustado.

Una propiedad importante del R^2 es el hecho de ser una función no dependiente del número de variables explicatorias del modelo; a medida que aumenta el número de variables explicatorias, R^2 casi invariablemente crece y nunca decrece, en otras palabras, una variable "x" adicional no disminuirá el R^2. Para ver eso, recordemos la definición del coeficiente de determinación:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad \textbf{(4.29)}$$

Ahora, $\sum_{i=1}^{n} (y_i - \bar{y})^2$ es independiente del número de variables "x" del modelo, sin embargo la suma de cuadrados residuales $\left(\sum_{i=1}^{n} e_i^2 \right)$ depende del número de variables explicatorias (incluyendo el intercepto). Por intuición, resulta claro que a medida que el

número de variables "x" aumenta, $\sum_{i=1}^{n} e_i^2$ debe decrecer o mantenerse; por lo tanto, el R^2 como se definió en la ecuación (4.29) crecerá. En vista de lo anterior al comparar dos modelos de regresión con la misma variable dependiente pero con distinto número de variables "x", es necesario tener cuidado de escoger el modelo que tenga el mayor R^2.

Para comparar dos R^2, hay que tener en cuenta el número de variables "x" del modelo, lo cual puede hacerse rápidamente mediante un coeficiente de determinación alterno, como sigue:

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2 \Big/ (n-k)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 \Big/ (n-1)} \qquad \textbf{(4.30)}$$

Donde:

k: Número de parámetros en el modelo incluyendo el término de intercepto.

(En el modelo de 3 variables k = 3, porque se estima β_0, β_1 y β_2). El R^2 definido de esta forma se conoce como el R^2 ajustado ($\bar{R}^2$). El término ajustado significa ajustado por los grados de libertad asociados con las sumas de cuadrados que aparecen en la ecuación (4.29): $\sum_{i=1}^{n} e_i^2$ tiene n-k grados de libertad en un modelo con k parámetros, que incluyen el intercepto, y $\sum_{i=1}^{n} (y_i - \bar{y})^2$ tiene n-1 grados de libertad. Para el caso de tres variables sabemos que $\sum_{i=1}^{n} e_i^2$ tiene n-3 grados de libertad.

La ecuación (4.30) puede escribirse como:

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{S_{yy}} \qquad \textbf{(4.31)}$$

Donde:

$\hat{\sigma}^2$: Es la varianza residual, un estimador insesgado del verdadero σ^2.

S_{yy}: Es la varianza muestral de "y".

Es fácil ver que $\bar{R}^2$ y R^2 están relacionados, sustituyendo la ecuación (4.29) en (4.30) obtenemos:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k} \qquad \textbf{(4.32)}$$

La deducción de este resultado se presenta en el apéndice **4.1 f).**

De la ecuación (4.32) se deduce inmediatamente que:

- Para k >1, $\bar{R}^2 < R^2$, lo que implica que a medida que el número de variables "x" aumenta, el R^2 ajustado es cada vez menor que el R^2 no ajustado.
- $\bar{R}^2$ puede ser negativo, aunque R^2 es necesariamente no negativo. En el caso de que $\bar{R}^2$ resulte negativo se debe tomar como cero.

Es importante notar que al comparar dos modelos por medio de los coeficientes de determinación, ya sea ajustado o no, la variable dependiente debe ser la misma, mientras que las variables explicatorias pueden tomar cualquier forma.

4.7.2 Coeficientes de Correlación Parcial.

Hasta ahora, nuestra consideración del análisis de regresión múltiple ha sido básicamente una extensión del caso de regresión simple. Introduciremos ahora un nuevo concepto llamado coeficiente de correlación parcial, que se da cuando tres o más variables son consideradas en el análisis de correlación (la correlación entre la variable dependiente, y solamente una de las variables independientes; la influencia de las otras variables independientes se mantiene constante en el análisis de correlación parcial). Por ejemplo, el coeficiente de correlación parcial para medir la correlación entre y_i y x_1, manteniendo constante x_2, es denotado con el símbolo $r_{y x_1 \bullet x_2}$.

Los subíndices primarios representan las variables para las cuales la correlación parcial está siendo medida, mientras que el subíndice secundario representa la variable que se mantiene constante.

Las correlaciones parciales pueden variar entre -1 y +1, al igual que en el caso de la correlación simple.

Utilizando la ecuación (1.10) del Capítulo 1:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Definimos ahora los coeficientes de correlación simple para el caso de tres variables.

Coeficiente de correlación simple entre "y" y x_1.

$$r_{yx_1} = \frac{n\sum_{i=1}^{n} x_{1i}y_i - \left(\sum_{i=1}^{n} x_{1i}\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_{1i}^2 - \left(\sum_{i=1}^{n} x_{1i}\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Coeficiente de correlación simple entre "y" y x_2.

$$r_{yx_2} = \frac{n\sum_{i=1}^{n} x_{2i}y_i - \left(\sum_{i=1}^{n} x_{2i}\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_{2i}^2 - \left(\sum_{i=1}^{n} x_{2i}\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

Coeficiente de correlación simple entre x_1 y x_2.

$$r_{x_1x_2} = \frac{n\sum_{i=1}^{n} x_{1i}x_{2i} - \left(\sum_{i=1}^{n} x_{1i}\right)\left(\sum_{i=1}^{n} x_{2i}\right)}{\sqrt{n\sum_{i=1}^{n} x_{2i}^2 - \left(\sum_{i=1}^{n} x_{2i}\right)^2}\sqrt{n\sum_{i=1}^{n} x_{1i}^2 - \left(\sum_{i=1}^{n} x_{1i}\right)^2}}$$

Con los valores de los coeficientes de correlación simple determinados, se pueden definir los coeficientes de correlación parcial para el caso de tres variables, en términos de estos valores de la siguiente manera.

Coeficiente de correlación parcial entre "y" y x_1, manteniéndose constante x_2:

$$r_{yx_1 \bullet x_2} = \frac{r_{yx_1} - r_{yx_2}r_{x_1x_2}}{\sqrt{1-(r_{yx_2})^2}\sqrt{1-(r_{x_1x_2})^2}} \qquad \textbf{(4.33)}$$

Coeficiente de correlación parcial entre "y" y x_2, manteniéndose constante x_1:

$$r_{yx_2 \bullet x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{\left[1-(r_{yx1})^2\right]\left[1-(r_{x_1x_2})^2\right]}} \qquad \textbf{(4.34)}$$

Coeficiente de correlación parcial entre x_1 y x_2, manteniéndose constante "y":

$$r_{x_2x_1 \bullet y} = \frac{r_{x_1x_2} - r_{yx_1} r_{yx_2}}{\sqrt{\left[1-(r_{yx_1})^2\right]\left[1-(r_{yx_2})^2\right]}} \qquad \textbf{(4.35)}$$

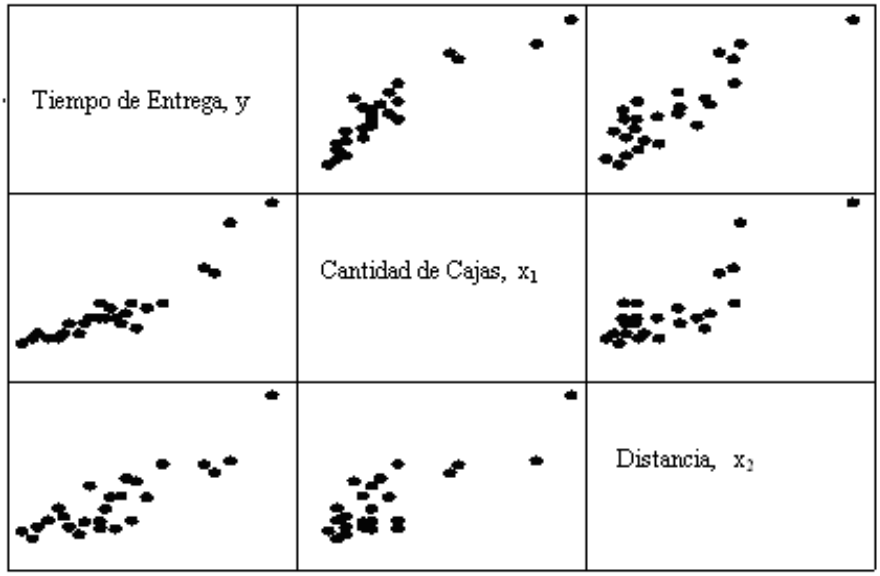
Las correlaciones parciales dadas en las ecuaciones (4.33) a (4.35) se llaman coeficientes de correlación parcial de primer orden; por orden se entiende el número de subíndices secundarios. Así, $r_{yx_1 \bullet x_2x_3}$ será el coeficiente de correlación de orden dos, $r_{yx_1 \bullet x_2x_3x_4}$ sería de orden tres y así sucesivamente. r_{yx_1} y los sucesivos se llaman correlaciones simples o de orden cero.

Ejemplo 1: Un Ingeniero Industrial empleado por la Compañía de la Coca-Cola, analiza las operaciones de entrega y servicio de producto en máquinas tragamonedas. Cree que el tiempo utilizado por un repartidor, en cargar y dar servicio a una máquina, se relaciona con la cantidad de cajas de productos entregadas y la distancia recorrida por el repartidor. El Ingeniero visita 25 tiendas de menudeo, escogidas al azar, con máquinas tragamonedas, y anota el tiempo de entrega en la tienda (en minutos), el volumen del producto entregado (en cajas) y la distancia recorrida (en pies), para cada una. Con los datos que se muestran en la tabla 4.1 ajustar un modelo de regresión lineal múltiple.

Tabla 4.1 Datos de tiempo de entrega.

Observaciones	Tiempo de Entrega, y (min.)	Cantidad de Cajas, x_1	Distancia, x_2 (pies)
1	26	7	330
2	10	2	110
3	25	7	210
4	15	3	220
5	17	3	340
6	21	4	80
7	19	6	150
8	58	30	1460
9	32	5	605
10	47	16	688
11	31	10	215
12	18	4	255
13	29	6	462
14	34	9	448
15	37	10	776
16	22	6	200
17	28	7	132
18	12	3	36
19	45	17	770
20	25	10	140
21	50	26	810
22	27	9	450
23	30	8	635
24	13	4	150
25	23	7	560

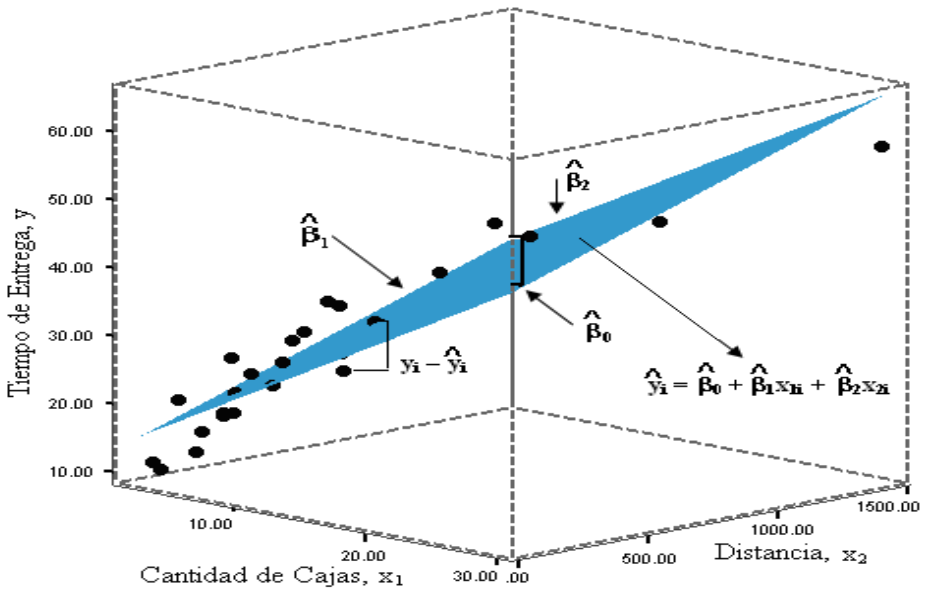
Figura 4.1 Matriz de diagramas de dispersión para datos de la tabla 4.1.



La figura 4.1 es una matriz de dispersión de los datos de tiempo de entrega. Es un arreglo bidimensional de graficas bidimensionales, en las que, a excepción de los de la diagonal, cada cuadro contiene un diagrama de dispersión. Así, cada cuadro nos muestra la relación entre un par de variables. Con frecuencia esto es un mejor resumen de las relaciones; que una presentación numérica, como por ejemplo mostrar los coeficientes de correlación entre cada par de variables, porque muestra un sentido de linealidad o de no linealidad en la relación, y cierta percepción de cómo se arreglan los datos individuales en la región.

Cuando sólo hay dos variables independientes, a veces un diagrama tridimensional de dispersión es útil para visualizar la regresión entre la variable dependiente y las independientes. La figura 4.2 muestra esta gráfica para los datos de Tiempo de Entrega.

Figura 4.2 Diagrama de dispersión con ajuste para los datos de la tabla 4.1.



La figura 4.2 muestra la relación que existe entre las tres variables, se puede observar que los puntos están cerca de la región sombreada lo que indica un buen ajuste, si la relación entre las variables fuera perfecta todos los puntos estarían en la región sombreada.

Con más de una variable independiente, la representación gráfica de las relaciones presentes en un modelo de regresión resulta poco intuitiva, muy complicada y nada útil. Es más fácil y práctico partir de la ecuación del modelo de regresión lineal estimado: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$ para la cual es necesario estimar los parámetros de regresión, en la tabla 4.2 se muestra como sigue:

Tabla 4.2 Resultados basados en los datos de la tabla 4.1

n	y	x_1	x_2	$(y_i - \bar{y})$	$(x_{1i} - \bar{x}_1)$	$(x_{2i} - \bar{x}_2)$	$(x_{1i} - \bar{x}_1)^2$	$(x_{2i} - \bar{x}_2)^2$	$(x_{1i} - \bar{x}_1)(y_i - \bar{y})$	$(x_{2i} - \bar{x}_2)(y_i - \bar{y})$	$(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$
1	26	7	330	-1. 76	-1. 76	-79. 28	3. 098	6285. 318	3.098	139.533	139.533
2	10	2	110	-17. 76	-6. 76	-299. 28	45. 698	89568. 518	120.058	5315.213	2023.133
3	25	7	210	-2. 76	-1. 76	-199. 28	3. 098	39712. 518	4.858	550.013	350.733
4	15	3	220	-12. 76	-5. 76	-189. 28	33. 178	35826. 918	73.498	2415.213	1090.253
5	17	3	340	-10. 76	-5. 76	-69. 28	33. 178	4799. 718	61.978	745.453	399.053
6	21	4	80	-6. 76	-4. 76	-329. 28	22. 658	108425. 318	32.178	2225.933	1567.373
7	19	6	150	-8. 76	-2. 76	-259. 28	7. 618	67226. 118	24.178	2271.293	715.613
8	58	30	1460	30. 24	21. 24	1050. 72	451. 138	1104012. 518	642.298	31773.773	22317.293
9	32	5	605	4. 24	-3. 76	195. 72	14. 138	38306. 318	-15.942	829.853	-735.907
10	47	16	688	19. 24	7. 24	278. 72	52. 418	77684. 838	139.298	5362.573	2017.933
11	31	10	215	3. 24	1. 24	-194. 28	1. 538	37744. 718	4.018	-629.467	-240.907
12	18	4	255	-9. 76	-4. 76	-154. 28	22. 658	23802. 318	46.458	1505.773	734.373
13	29	6	462	1. 24	-2. 76	52. 72	7. 618	2779. 398	-3.422	65.373	-145.507
14	34	9	448	6. 24	0. 24	38. 72	0. 058	1499. 238	1.498	241.613	9.293
15	37	10	776	9. 24	1. 24	366. 72	1. 538	134483. 558	11.458	3388.493	454.733
16	22	6	200	-5. 76	-2. 76	-209. 28	7. 618	43798. 118	15.898	1205.453	577.613
17	28	7	132	0. 24	-1. 76	-277. 28	3. 098	76884. 198	-0.422	-66.547	488.013
18	12	3	36	-15. 76	-5. 76	-373. 28	33. 178	139337. 958	90.778	5882.893	2150.093
19	45	17	770	17. 24	8. 24	360. 72	67. 898	130118. 918	142.058	6218.813	2972.333
20	25	10	140	-2. 76	1. 24	-269. 28	1. 538	72511. 718	-3.422	743.213	-333.907
21	50	26	810	22. 24	17. 24	400. 72	297. 218	160576. 518	383.418	8912.013	6908.413
22	27	9	450	-0. 76	0. 24	40. 72	0. 58	1658. 118	-0.182	-30.947	9.773
23	30	8	635	2. 24	-0. 76	225.72	0. 578	50949. 518	-1.702	505.613	-171.547
24	13	4	150	-14. 76	-4. 76	-259. 28	22. 658	67226. 118	70.258	3826.973	1234.173
25	23	7	560	-4. 76	-1. 76	150. 72	3. 098	22716. 518	8.378	-717.427	-265.267

Sumando los valores de las columnas de la tabla 4.2 se obtienen:

$$\sum_{i=1}^{25} y_i = 694, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{25}(694) = 27.76$$

$$\sum_{i=1}^{25} x_{1i} = 219, \quad \bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i} = \frac{1}{25}(219) = 8.76$$

$$\sum_{i=1}^{25} x_{2i} = 10232, \quad \bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{2i} = \frac{1}{25}(10232) = 409.28$$

$$\sum_{i=1}^{25} (x_{1i} - \bar{x}_1)^2 = 1136.56, \quad \sum_{i=1}^{25} (x_{2i} - \bar{x}_2)^2 = 2537935.04$$

$$\sum_{i=1}^{25} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) = 1850.56, \quad \sum_{i=1}^{25} (x_{2i} - \bar{x}_2)(y_i - \bar{y}) = 82680.68$$

$$\sum_{i=1}^{25} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = 44266.680, \quad \sum_{i=1}^{n} (y_i - \bar{y})^2 = 3588.56$$

Sustituyendo los datos anteriores en la ecuación siguiente se tiene:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2}$$

$$\hat{\beta}_1 = \frac{(1850.56)(2537935.04) - (82680.68)(44266.680)}{(1136.56)(2537935.04) - (44266.680)^2}$$

$$\hat{\beta}_1 = \frac{4696601068 - 3659999204}{2884515449 - 1959538958}$$

$$\hat{\beta}_1 = \frac{1036601864}{924976491}$$

$$\hat{\beta}_1 = 1.120679146$$

$$\hat{\beta}_1 \approx 1.121$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)(y_i - \bar{y})\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 - \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y})\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right)^2}$$

$$\hat{\beta}_2 = \frac{(82680.68)(1136.56) - (1850.56)(44266.680)}{(1136.56)(2537935.04) - (44266.680)^2}$$

$$\hat{\beta}_2 = \frac{93971553.66 - 81918147.34}{2884515449 - 1959538958}$$

$$\hat{\beta}_2 = \frac{12053406.32}{924976491}$$

$$\hat{\beta}_2 = 0.013$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2$$

$$\hat{\beta}_0 = 27.76 - (1.121)(876) - (0.013)(409.28)$$

$$\hat{\beta}_0 = 12.6194$$

$$\hat{\beta}_0 \approx 12.610$$

Sustituyendo en la ecuación de regresión lineal estimada los valores de los parámetros, se tiene:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

$$\hat{y}_i = 12.610 + 1.121\, x_{1i} + 0.013\, x_{2i} \qquad \textbf{(4.36)}$$

$$\text{Tiempo de Entrega} = 12.610 + 1.121\,\text{Cantidad de Cajas} + 0.013\,\text{Distancia recorrida}$$

La interpretación de la ecuación (4.36) es la siguiente: si las variables Cantidad de Cajas de producto y Distancia recorrida por el repartidor se fijan o se igualan a cero, el promedio o valor medio del Tiempo de Entrega (que refleja la influencia de todas las variables omitidas) es aproximadamente 12.610.

El coeficiente de regresión parcial $\hat{\beta}_1 = 1.121$ mide la cantidad promedio en que se espera, que un cambio en una unidad en la variable Cantidad de Cajas afecte al Tiempo de Entrega cuando la variable Distancia recorrida se mantiene constante.

El coeficiente de regresión parcial $\hat{\beta}_2 = 0.013$ mide la cantidad promedio de cambio en el Tiempo de Entrega por unidad de cambio en la Distancia recorrida cuando Cantidad de Cajas se mantiene constante.

Tabla 4.3 Datos originales, valores estimados usando la ecuación (4.36) y residuos.

n	y	x_1	x_2	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$	e_i^2
1	26	7	330	24.747	1.253	1.570
2	10	2	110	16.282	-6.282	39.464
3	25	7	210	23.187	1.813	3.287
4	15	3	220	18.833	-3.833	14.692
5	17	3	340	20.393	-3.393	11.512
6	21	4	80	18.134	2.866	8.214
7	19	6	150	21.286	-2.286	5.226
8	58	30	1460	65.22	-7.22	52.128
9	32	5	605	26.08	5.92	35.046
10	47	16	688	39.49	7.51	56.400
11	31	10	215	26.615	4.385	19.228
12	18	4	255	20.409	-2.409	5.803
13	29	6	462	25.342	3.658	13.381
14	34	9	448	28.523	5.477	29.998
15	37	10	776	33.908	3.092	9.560
16	22	6	200	21.936	0.064	0.004
17	28	7	132	22.173	5.827	33.954
18	12	3	36	16.441	-4.441	19.722
19	45	17	770	41.677	3.323	11.042
20	25	10	140	25.64	-0.64	0.410
21	50	26	810	52.286	-2.286	5.226
22	27	9	450	28.549	-1.549	2.399
23	30	8	635	29.833	0.167	0.028
24	13	4	150	19.044	-6.044	36.530
25	23	7	560	27.737	-4.737	22.439
						$\sum_{i=1}^{25} e_i^2 = 437.265$

Para el cálculo de la varianza y el error estándar de los parámetros de regresión lineal se necesita la varianza de los errores, y se calculan de la forma siguiente:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-3}$$
$$\hat{\sigma}^2 = \frac{437.265}{25-3}$$
$$\hat{\sigma}^2 = \frac{437.265}{22}$$
$$\hat{\sigma}^2 = 19.8756$$

La varianza de $\hat{\beta}_0$:

$$var(\hat{\beta}_0) = \left[\frac{1}{n} + \frac{\bar{x}_1^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 + \bar{x}_2^2 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 - 2\bar{x}_1 \bar{x}_2 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2} \right] * \sigma^2$$

$$var(\hat{\beta}_0) = \left[\frac{1}{25} + \frac{(8.76)^2 (2537935.04) + (409.28)^2 (1136.56) - 2(8.76)(409.28)(44266.680)}{(1136.56)(2537935.04) - (44266.680)^2} \right] * (19.8756)$$

$$var(\hat{\beta}_0) = \left[\frac{1}{25} + \frac{67722325.93}{924976490.8} \right] * (19.8756)$$

$$var(\hat{\beta}_0) = 2.2502$$

El error estándar de $\hat{\beta}_0$:

$$es(\hat{\beta}_0) = \sqrt{var(\hat{\beta}_0)}$$
$$es(\hat{\beta}_0) = \sqrt{2.2502}$$
$$es(\hat{\beta}_0) = 1.500$$

La varianza de $\hat{\beta}_1$:

$$\text{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right)^2} * \sigma^2$$

$$\text{var}(\hat{\beta}_1) = \frac{2537935.04}{(1136.56)(2537935.04) - (44266.680)^2} * (19.8756)$$

$$\text{var}(\hat{\beta}_1) = \frac{2537935.04}{924976490.8} * (19.8756)$$

$$\text{var}(\hat{\beta}_1) = (0.002743783)(19.8756)$$

$$\text{var}(\hat{\beta}_1) \approx 0.05453$$

El error estándar de $\hat{\beta}_1$:

$$\text{es}(\hat{\beta}_1) = \sqrt{\text{var}(\hat{\beta}_1)}$$

$$\text{es}(\hat{\beta}_1) = \sqrt{0.05453}$$

$$\text{es}(\hat{\beta}_1) = 0.234$$

La varianza de $\hat{\beta}_2$ es:

$$\text{var}(\hat{\beta}_2) = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)\right)^2} * \sigma^2$$

$$\text{var}(\hat{\beta}_2) = \frac{1136.56}{(1136.56)(2537935.04) - (44266.680)^2} * (19.8756)$$

$$\text{var}(\hat{\beta}_2) = \frac{1136.56}{924976490.8} * (19.8756)$$

$$\text{var}(\hat{\beta}_2) = 0.000024422$$

El error estándar de $\hat{\beta}_2$ es:

$$es(\hat{\beta}_2) = \sqrt{var(\hat{\beta}_2)}$$
$$es(\hat{\beta}_2) = \sqrt{0.000024422}$$
$$es(\hat{\beta}_2) = 0.005$$

Haciendo uso de los datos anteriores se calcula el valor de R^2, R y $\overline{R}^2$.

El R^2 se obtiene de la siguiente forma:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$
$$R^2 = 1 - \frac{437.265}{3588.560}$$
$$R^2 = 1 - 0.1218$$
$$R^2 = 0.8781$$

$$R = \sqrt{R^2}$$
$$R = \sqrt{0.8781}$$
$$R = 0.937$$

El valor de $R^2 = 0.8781$, el cual muestra que las dos variables Cantidad de Cajas y Distancia recorrida explican alrededor del 87.81% de la variación en el Tiempo de Entrega.

El valor de $\overline{R}^2$ se obtiene:

$$\overline{R}^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2 \Big/ (n-k)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 \Big/ (n-1)}$$

$$\overline{R}^2 = 1 - \frac{437.265/(25-3)}{3588.560/(25-1)}$$

$$\overline{R}^2 = 1 - \frac{19.8756}{149.5233}$$

$$\overline{R}^2 = 1 - 0.1329$$

$$\overline{R}^2 = 0.867$$

El valor de $\overline{R}^2 = 0.867$ nos indica que después de tener en cuenta los grados de libertad, las variables Cantidad de Cajas y Distancia recorrida aún explican también el 86.7% de la variación en el Tiempo de Entrega.

Calculamos ahora los coeficientes de correlación parcial para los datos del Tiempo de Entrega, para ello se necesita encontrar el coeficiente de correlación simple entre cada par de variables.

Coeficiente de correlación simple entre Tiempo de Entrega y Cantidad de Cajas.

$$r_{yx_1} = \frac{n\sum_{i=1}^{n} x_{1i}y_i - \left(\sum_{i=1}^{n} x_{1i}\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_{1i}^2 - \left(\sum_{i=1}^{n} x_{1i}\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r_{yx_1} = \frac{(25)(7930) - (219)(694)}{\sqrt{(25)(3055) - (219)^2}\sqrt{(25)(22854) - (694)^2}}$$

$$r_{yx_1} = \frac{46264}{50488.94528}$$

$$r_{yx_1} = 0.916$$

Coeficiente de correlación simple entre Tiempo de Entrega y Distancia recorrida.

$$r_{yx_2} = \frac{n\sum_{i=1}^{n} x_{2i}y_i - \left(\sum_{i=1}^{n} x_{2i}\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_{2i}^2 - \left(\sum_{i=1}^{n} x_{2i}\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}$$

$$r_{yx_2} = \frac{(25)(366721) - (10232)(694)}{\sqrt{(25)(6725688) - (10232)^2}\sqrt{(25)(22854) - (694)^2}}$$

$$r_{yx_2} = \frac{2067017}{2385834.782}$$

$$r_{yx_2} = 0.866$$

Coeficiente de correlación simple entre Cantidad de Cajas y Distancia recorrida.

$$r_{x_1x_2} = \frac{n\sum_{i=1}^{n} x_{1i}x_{2i} - \left(\sum_{i=1}^{n} x_{1i}\right)\left(\sum_{i=1}^{n} x_{2i}\right)}{\sqrt{n\sum_{i=1}^{n} x_{2i}^2 - \left(\sum_{i=1}^{n} x_{2i}\right)^2}\sqrt{n\sum_{i=1}^{n} x_{1i}^2 - \left(\sum_{i=1}^{n} x_{1i}\right)^2}}$$

$$r_{x_1x_2} = \frac{(25)(133899) - (219)(10232)}{\sqrt{(25)(6725688) - (10232)^2}\sqrt{(25)(3055) - (219)^2}}$$

$$r_{x_1x_2} = \frac{1106667}{1342692.13}$$

$$r_{x_1x_2} = 0.824$$

Se puede observar que los valores de los coeficientes de correlación simple obtenidos, están cerca de 1, que indica que existe una buena asociación lineal entre cada par de variables.

Con los valores de los coeficientes de correlación simple determinados anteriormente, podemos definir los coeficientes de correlación parcial para el caso de tres variables, en términos de estos valores de la siguiente manera:

Coeficiente de correlación parcial entre el Tiempo de Entrega y Cantidad de Cajas, manteniéndose constante Distancia recorrida:

$$r_{yx_1 \bullet x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left[1-(r_{yx_2})^2\right]\left[1-(r_{x_1x_2})^2\right]}}$$

$$r_{yx_1 \bullet x_2} = \frac{0.916 - (0.866)(0.824)}{\sqrt{\left[1-(0.866)^2\right]\left[1-(0.824)^2\right]}}$$

$$r_{yx_1 \bullet x_2} = \frac{0.202416}{0.283319828}$$

$$r_{yx_1 \bullet x_2} \approx 0.715$$

El valor de $r_{yx_1 \bullet x_2} \approx 0.715$ indica que existe una buena asociación entre las variables Tiempo de Entrega y Cantidad de Cajas cuando no interviene la variable Distancia recorrida.

Coeficiente de correlación parcial entre Tiempo de Entrega y Distancia recorrida Manteniéndose constante Cantidad de Cajas:

$$r_{yx_2 \bullet x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1x_2}}{\sqrt{\left[1-(r_{yx1})^2\right]\left[1-(r_{x_1x_2})^2\right]}}$$

$$r_{yx_2 \bullet x_1} = \frac{0.866 - (0.916)(0.824)}{\sqrt{\left[1-(0.916)^2\right]\left[1-(0.824)^2\right]}}$$

$$r_{yx_2 \bullet x_1} = \frac{0.111216}{0.227303512}$$

$$r_{yx_2 \bullet x_1} = 0.489$$

$$r_{yx_2 \bullet x_1} \approx 0.49$$

El valor de $r_{yx_2 \bullet x_1} \approx 0.49$ indica que existe poca relación entre las variables Tiempo de entrega y Distancia recorrida cuando se mantiene constante la variable Cantidad de Cajas.

Coeficiente de correlación parcial entre Cantidad de Cajas y Distancia recorrida, manteniéndose constante Tiempo de Entrega:

$$r_{x_2x_1 \bullet y} = \frac{r_{x_1x_2} - r_{yx_1} r_{yx_2}}{\sqrt{\left[1 - (r_{yx_1})^2\right]\left[1 - (r_{yx_2})^2\right]}}$$

$$r_{x_2x_1 \bullet y} = \frac{0.824 - (0.916)(0.866)}{\sqrt{\left[1 - (0.916)^2\right]\left[1 - (0.866)^2\right]}}$$

$$r_{x_2x_1 \bullet y} = \frac{0.030744}{0.200606783}$$

$$r_{x_2x_1 \bullet y} = 0.153$$

La cantidad de $r_{x_2x_1 \bullet y} = 0.153$ muestra que no existe relación entre las variables Cantidad de Cajas y Distancia recorrida, cuando se mantiene constante la variable Tiempo de Entrega, porque el valor está más cerca de 0 que de 1.

4.8 Supuesto de Normalidad.

Sabemos que si nuestro único objetivo es la estimación puntual de los parámetros de los modelos de regresión, el método de Mínimos Cuadrados Ordinarios (MCO), que no hace ningún supuesto respecto a la distribución de probabilidad de las perturbaciones ε_i, será más que suficiente. Pero si además nuestro objetivo es tanto la estimación como

la inferencia, entonces, es necesario suponer que ε_i sigue alguna distribución de probabilidad.

Hemos supuesto que ε_i sigue la distribución normal con media cero y varianza constante σ^2, supuesto que mantendremos para el modelo de regresión múltiple. Con el supuesto de normalidad y siguiendo lo expuesto en el Capítulo 2, los estimadores de MCO de los coeficientes de regresión parcial, que además son idénticos a los estimadores de Máxima Verosimilitud (MV), son los mejores estimadores lineales insesgados. Y aun más, los estimadores $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$ están normalmente distribuidos con medias iguales a β_0, β_1 y β_2, y con varianzas dadas en las ecuaciones (4.18), (4.20) y (4.22). Igualmente $(n-3)\hat{\sigma}^2/\sigma^2$ sigue la distribución ji-cuadrada (χ^2) con n-3 grados de libertad, y los tres estimadores de MCO están distribuidos independientemente de $\hat{\sigma}^2$.

Reemplazando σ^2 por su estimador insesgado $\hat{\sigma}^2$ en el cálculo de los errores estándar, cada una de las variables sigue la distribución t con n-3 grados de libertad.

$$t = \frac{x - \mu}{s}$$

Sustituyendo los parámetros estimados, poblacionales y los errores estándar en la ecuación anterior se obtiene:

$$t_0 = \frac{\hat{\beta}_0 - \beta_0}{es(\hat{\beta}_0)} \qquad \textbf{(4.37)}$$

$$t_0 = \frac{\hat{\beta}_1 - \beta_1}{es(\hat{\beta}_1)} \qquad \textbf{(4.38)}$$

$$t_0 = \frac{\hat{\beta}_2 - \beta_2}{es(\hat{\beta}_2)} \qquad \textbf{(4.39)}$$

Nótese que ahora los grados de libertad son n-3, debido a que en el cálculo de la $\sum_{i=1}^{n} e_i^2$ o de $\hat{\sigma}^2$ se estimaron primero tres coeficiente de regresión parcial, que obviamente impusieron tres restricciones en la suma de cuadrados residuales. Por lo tanto, la distribución t puede utilizarse no sólo para establecer intervalos de confianza si no para probar hipótesis estadísticas respecto a los coeficientes de regresión parcial de la verdadera población.

Así mismo, la distribución χ^2 puede emplearse para hacer prueba de hipótesis respecto a σ^2.

4.8.1 Pruebas de Hipótesis sobre Coeficientes Individuales de Regresión Parcial.

Teniendo en cuenta el supuesto de que $\varepsilon_i \sim$ NID(0, σ^2), podemos utilizar la prueba t para hacer pruebas de hipótesis a cerca de cualquier coeficiente individual de regresión parcial.

Tomando la información del ejemplo 1.

Supongamos que se desean probar las hipótesis:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0 \qquad \textbf{(4.40)}$$

La hipótesis nula establece que manteniendo x_2 constante, la Distancia recorrida por el repartidor no tiene influencia (lineal) sobre el Tiempo de Entrega. Para verificar la hipótesis nula se hace uso de la prueba t dada en la ecuación (4.38). Si el valor del t calculado excede el t crítico para el nivel de significancia escogido, podemos rechazar la hipótesis nula; de lo contrario podemos aceptarla.

Ejemplo 2: Se probará la significancia de la regresión $(\hat{\beta}_1)$ para el modelo de Tiempo de Entrega, ejemplo 1 es decir, $H_0 : \beta_1 = 0$ y $H_1 : \beta_1 \neq 0$.

Datos:

El valor estimado de $\hat{\beta}_1 = 1.121$

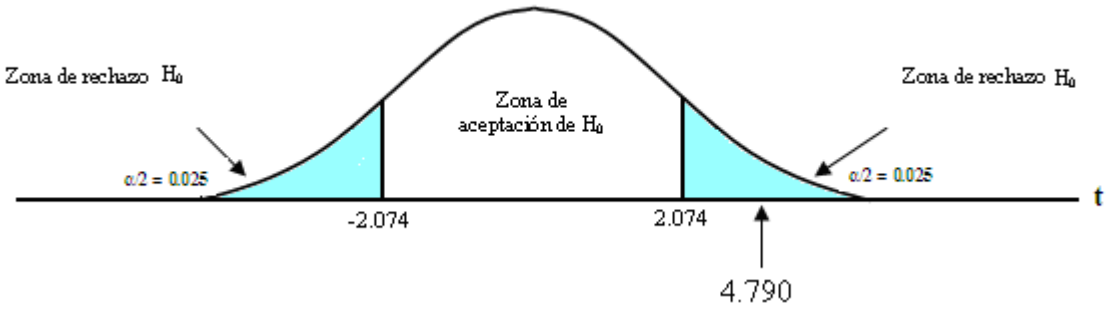
El error estándar $es(\hat{\beta}_1) = 0.234$

Solución:

1. $H_0 : \beta_1 = 0$
2. $H_1 : \beta_1 \neq 0$
3. Se selecciona un nivel de significancia de α = 0.05 y como la prueba es de dos colas $\alpha/2$ = 0.05/2 = 0.025 y se tiene que el valor de la tabla de t es $t_{(0.05/2,\ 25-3)} = t_{(0.025,\ 22)} = 2.074$
4. Región crítica: si t < - 2.074 ó t > 2.074, entonces rechazamos H_0.
5. Cálculos:

$$t_0 = \frac{\hat{\beta}_1}{es(\hat{\beta}_1)} = \frac{1.121}{0.234} = 4.790$$

Figura 4.3 de la Distribución t.



6. Decisión Estadística: se rechaza H_0, porque el valor calculado para t_0 cae en la zona de rechazo de H_0, es decir que β_1 es estadísticamente significativa, esto es, significativamente diferente de cero.
7. Conclusión: se concluye que hay una relación lineal entre Tiempo de Entrega y Cantidad de Cajas.

Ejemplo 3: Se probará la significancia de la regresión $(\hat{\beta}_2)$ para el modelo de Tiempo de Entrega, ejemplo 1, es decir, $H_0 : \beta_2 = 0$ y $H_1 : \beta_2 \neq 0$.

Datos:

El valor estimado de $\hat{\beta}_2 = 0.013$

El error estándar $es(\hat{\beta}_2) = 0.005$

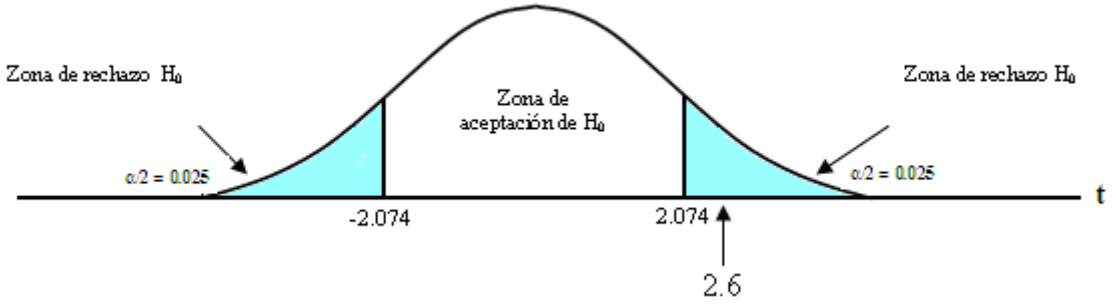
Solución:

1. $H_0 : \beta_2 = 0$
2. $H_1 : \beta_2 \neq 0$

3. Se selecciona un nivel de significancia de $\alpha = 0.05$ y como la prueba es de dos colas $\alpha/2 = 0.05/2 = 0.025$ y se tiene que el valor de la tabla de t es $t_{(0.05/2,\ 25-3)} = t_{(0.025,\ 22)} = 2.074$
4. Región crítica: si $t < -2.074$ ó $t > 2.074$, entonces rechazamos H_0.
5. Cálculos:

$$t_0 = \frac{\hat{\beta}_2}{es(\hat{\beta}_2)} = \frac{0.013}{0.005} = 2.6$$

Figura 4.4 de la Distribución t.



6. Decisión Estadística: se rechaza H_0, porque el valor calculado cae en la zona de rechazo de H_0, es decir que β_2 es estadísticamente significativa, esto es, significativamente diferente de cero.
7. Conclusión: se concluye que hay una relación lineal entre Tiempo de Entrega y Distancia recorrida.

4.8.2 Pruebas de la Significación Global de la Regresión Muestral.

En la sección anterior nos limitamos a verificar individualmente la significancia de los coeficientes de regresión parcial estimados, es decir, bajo la hipótesis separada de que cada coeficiente de regresión parcial correspondiente a la población verdadera es igual a cero.

Ahora se considera la siguiente hipótesis

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \beta_j \neq 0, \quad \text{al menos para un j.} \qquad \textbf{(4.41)}$$

Esta hipótesis nula es una hipótesis conjunta según la que β_1 y β_2 son simultáneamente iguales a cero. La prueba o verificación de una hipótesis como esta se denomina prueba de la significancia global de la línea de regresión observada o estimada, es decir, si es cierto que "y" está linealmente relacionada tanto con x_1 como con x_2.

¿Puede verificarse la hipótesis conjunta dada en la ecuación (4.41) probando la significancia de $\hat{\beta}_1$ y $\hat{\beta}_2$ individualmente como en la sección 4.8.1?. La respuesta es negativa por lo siguiente:

Al verificar en la sección 4.8.1 la significancia individual de un coeficiente de regresión parcial observado, se supuso implícitamente que cada prueba de significancia estaba basada en una muestra diferente. De tal manera que cuando verificábamos la significancia de $\hat{\beta}_1$ bajo la hipótesis de que $\hat{\beta}_1 = 0$, se suponía implícitamente que la verificación o prueba se basaba en una muestra distinta a la que se utilizó para verificar

la significancia de $\hat{\beta}_2$ bajo la hipótesis nula de que $\hat{\beta}_2 = 0$. Pero si en el proceso de verificar conjuntamente la hipótesis dada en la ecuación (4.41) utilizamos las mismas cifras muéstrales (tabla 4.1), se violaría el supuesto anterior del método de verificación.

El resultado del planteamiento anterior es el de que para un ejemplo dado, tan solo puede encontrarse una prueba de hipótesis. La pregunta obvia será entonces, ¿Cómo verificar simultáneamente la hipótesis nula $H_0 : \beta_1 = \beta_2 = 0$?. La respuesta a esta pregunta se da en la sección siguiente.

4.8.3 Análisis de Varianza en las Pruebas de Significancia Global de una Regresión Múltiple.

Por la razón expuesta en la sección anterior no se puede utilizar la prueba t para verificar la hipótesis conjunta según la cual las pendientes de las distintas variables son simultáneamente cero. Sin embargo, esta hipótesis conjunta puede verificarse mediante la técnica de Análisis de Varianza y puede demostrarse del modo siguiente:

Recordando la identidad

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \hat{\beta}_1 \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y}) + \hat{\beta}_2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)(y_i - \bar{y}) + \sum_{i=1}^{n} e_i^2 \qquad \textbf{(4.42)}$$

Donde:

$\sum_{i=1}^{n}(y_i - \bar{y})^2 =$ Suma Total de Cuadrados (SS_T).

$\hat{\beta}_1 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) + \hat{\beta}_2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(y_i - \bar{y}) =$ Suma de Cuadrados de Regresión (SS_R).

$\sum_{i=1}^{n} e_i^2 =$ Suma de Cuadrados de Error (SS_{Res}).

SS_T tiene n-1 grados de libertad, SS_R tiene 2 grados de libertad en razón de que es una función de $\hat{\beta}_1$ y $\hat{\beta}_2$, y SS_{Res} tiene n-3 por lo que se dijo antes. Por lo tanto, siguiendo el procedimiento de análisis de varianza comentado en el Capítulo 2, sección 2.6.3 se puede elaborar la tabla 4.4.

Ahora bajo el supuesto de que los ε_i están distribuidos normalmente y de que la hipótesis nula $H_0 : \beta_1 = \beta_2 = 0$, la variable

$$F_0 = \frac{\left. \hat{\beta}_1 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) + \hat{\beta}_2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \right/ 2}{\left. \sum_{i=1}^{n} e_i^2 \right/ (n-3)} = \frac{SS_R / 2}{SS_{Res} / (n-3)} = \frac{MS_R}{MS_{Res}} \qquad \textbf{(4.43)}$$

Está distribuida como la distribución F con 2 y n-3 grados de libertad.

Tabla 4.4 Análisis de varianza para una regresión de tres variables.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F_0
Regresión	$SS_R = SS_T - SS_{Res}$	2	MS_R	MS_R/MS_{Res}
Residual	$SS_{Res} = SS_T - SS_R$	n-3	MS_{Res}	
Total	$SS_T = SS_R + SS_{Res}$	n-1		

El valor de F_0 dado en la ecuación (4.43) proporciona una prueba de la hipótesis nula, o sea que los coeficientes verdaderos correspondientes a las pendientes son simultáneamente iguales a cero. Si el valor de F_0 calculado es mayor que el valor tomado de la tabla F para un nivel de significancia α, rechazamos H_0; de lo contrario la aceptamos.

Ejemplo 4. Se probará la significancia de la regresión para el modelo del ejemplo 1, de los datos Tiempo de Entrega, Cantidad de Cajas y Distancia recorrida, es decir $H_0 : \beta_1 = \beta_2 = 0$ y $H_1 : \beta_j \neq 0$, al menos para un j.

Datos:

El modelo ajustado es: $\hat{y}_i = 12.610 + 1.121\, x_{1i} + 0.013\, x_{2i}$

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 3588.56$$

$$\hat{\beta}_1 = 1.121\,,\ \hat{\beta}_2 = 0.013\,,\ \sum_{i=1}^{25} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) = 1850.56$$

$$\sum_{i=1}^{25} (x_{2i} - \bar{x}_2)(y_i - \bar{y}) = 82680.68$$

$$SS_R = \hat{\beta}_1 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) + \hat{\beta}_2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(y_i - \bar{y})$$

$$SS_R = 1.121(1850.56) + 0.013(82680.68)$$

$$SS_R = 3149.3266$$

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = 437.265$$

Solución:

1. $H_0 : \beta_1 = \beta_2 = 0$
2. $H_1 : \beta_j \neq 0$, al menos para un j.
3. Se selecciona un nivel de significancia de α = 0.05 y se tiene que el valor de la tabla F es $F_{(0.05,\ 2,\ 22)}$ = 3.44.
4. Cálculos:

$$F_0 = \frac{SS_R/2}{SS_{Res}/(n-3)} = \frac{3149.3266/2}{437.265/(25-3)} = \frac{3149.3266/2}{437.265/(25-3)} = \frac{1574.6633}{19.8756} = 79.225$$

Tabla 4.5 Análisis de varianza para las variables del ejemplo 1.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F_0
Regresión	3149.3266	2	1574.6633	79.225
Residual	437.265	22	19.8756	
Total	3588.56	24		

5. Decisión Estadística: Se rechaza H_0, porque el valor calculado para F_0 (79.225) es mayor que el de la tabla (3.44).
6. Conclusión: Se concluye que el Tiempo de Entrega se relaciona con la Cantidad de Cajas y con la Distancia recorrida.

4.8.4 Importancia de la Relación entre R^2 y F.

Existe una relación íntima entre el coeficiente de determinación R^2 y la prueba F utilizada en el análisis de varianza. Suponiendo que los ε_i están distribuidos normalmente y que $\beta_1 = \beta_2 = 0$, es decir la hipótesis nula, hemos visto que:

$$F_0 = \frac{SS_R/2}{SS_{Res}/(n-3)} \qquad \textbf{(4.44)}$$

Está distribuido como la distribución F con 2 y n-3 grados de libertad.

En general en el caso de k variables (incluido el intercepto), si suponemos que los errores están distribuidos normalmente y que la hipótesis nula es:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \qquad \textbf{(4.45)}$$

Se deduce que:

$$F_0 = \frac{SS_R/(k-1)}{SS_{Res}/(n-k)} \qquad \textbf{(4.46)}$$

Tiene la distribución F con k-1 y n-k grados de libertad.

Nota: El número de parámetros a estimar es k, de los cuales uno corresponde al intercepto.

Manipulando la ecuación (4.46) se tiene:

$$F_0 = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \qquad \textbf{(4.47)}$$

Habiendo hecho uso de la definición de $R^2 = \frac{SS_R}{SS_T}$, la ecuación (4.47) muestra cómo están relacionados R^2 y F. Estos dos estadísticos varían directamente cuando $R^2 = 0$, F es cero inmediatamente. Mientras mayor sea el R^2 mayor será el F. En el limite cuando $R^2 = 1$, F es infinito. De este modo la prueba F, que es una medida de la significancia global de la regresión estimada, es también una prueba para el R^2. En otros términos verificar la hipótesis nula dada en la ecuación (4.45) es equivalente a verificar la hipótesis nula de que el R^2 (de la población) es cero.

Para el caso de tres variables la ecuación (4.47) se convierte en

$$F_0 = \frac{R^2/2}{(1-R^2)/(n-3)} \qquad \textbf{(4.48)}$$

Por la conexión que hay entre R^2 y F, la tabla 4.6 de análisis de varianza puede rotularse del mismo modo que la tabla 4.4.

Tabla 4.6 Análisis de varianza en términos de R^2.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio
Regresión	$R^2\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)$	2	$R^2\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)\Big/2$
Residual	$(1-R^2)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)$	n-3	$(1-R^2)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)\Big/(n-3)$
Total	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	n-1	

Ejemplo 5. Encontrar el valor de F_0 para los datos del ejemplo 1, haciendo uso de la ecuación (4.47).

Datos:

$R^2 = 0.8781$

$k = 3$

$n = 25$

Solución:

$$F_0 = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} = \frac{0.8781/(3-1)}{(1-0.8781)/(25-3)} = \frac{0.43905}{0.005540909} = 79.237$$

El valor de 79.237, es aproximadamente igual a 79.225 obtenido con la ecuación (4.43); la diferencia se debe a errores de redondeo. El valor de $F_0 = 79.237$ es mayor que el de la tabla (3.44) lo que nos permite rechazar la hipótesis nula.

4.8.5 Intervalos de Confianza en Regresión Múltiple.

Los intervalos de confianza de los coeficientes de regresión individuales, juegan el mismo papel importante que en la regresión lineal simple.

4.8.5.1 Intervalos de Confianza de los Coeficientes de Regresión.

Para construir estimados de intervalos de confianza de los coeficientes de regresión β_j, se continuará suponiendo que los errores ε_i están distribuidos normalmente, con media cero y varianza σ^2. En consecuencia las observaciones y_i están distribuidas en

forma normal e independiente, con media $\sum_{j=1}^{k}\beta_j x_{ij}$, y varianza σ^2. Como el estimador $\hat{\beta}$ por Mínimos Cuadrados es una combinación lineal de las observaciones, también está distribuido normalmente.

Entonces la distribución de muestreo para el caso de tres variables:

$$t_0 = \frac{\hat{\beta}_0 - \beta_0}{es(\hat{\beta}_0)}, \quad t_0 = \frac{\hat{\beta}_1 - \beta_1}{es(\hat{\beta}_1)} \quad y \quad t_0 = \frac{\hat{\beta}_2 - \beta_2}{es(\hat{\beta}_2)}$$

Tiene n-3 grados de libertad. Así se puede definir un intervalo de confianza de 100(1-α) por ciento para la ordenada al origen β_0 como sigue:

$$\hat{\beta}_0 - t_{(\alpha/2,\, n-3)}es(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t_{(\alpha/2,\, n-3)}es(\hat{\beta}_0) \qquad \textbf{(4.49)}$$

Un intervalo de confianza de 100(1-α) por ciento para la pendiente β_1 es:

$$\hat{\beta}_1 - t_{(\alpha/2,\, n-3)}es(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t_{(\alpha/2,\, n-3)}es(\hat{\beta}_1) \qquad \textbf{(4.50)}$$

Y un intervalo de confianza de 100(1-α) por ciento para la pendiente β_2 es:

$$\hat{\beta}_2 - t_{(\alpha/2,\, n-3)}es(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{(\alpha/2,\, n-3)}es(\hat{\beta}_2) \qquad \textbf{(4.51)}$$

Ejemplo 6: Calcular el intervalo de confianza del 95% para el parámetro β_1, para los datos del ejemplo 1.

Datos:

El estimador puntual de β_1 es $\hat{\beta}_1 = 1.121$

El valor para $t_{(\alpha/2,\, n-3)}$ es: $t_{(.05/2,\ 25-3)} = t_{(0.025,\ 22)} = 2.074$

El error estándar de $\hat{\beta}_1$: $es(\hat{\beta}_1) = 0.234$

Solución:

Sustituyendo estos datos en la ecuación (4.50) se tiene:

$$\hat{\beta}_1 - t_{(\alpha/2,\, n-3)} es(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{(\alpha/2,\, n-3)} es(\hat{\beta}_1)$$
$$1.121 - 2.074(0.234) \leq \beta_1 \leq 1.121 + 2.074(0.234)$$

$$0.636 \leq \beta_1 \leq 1.606 \qquad \textbf{(4.52)}$$

Esto es β_1 cae entre 0.636 y 1.606 con un coeficiente de confianza del 95%, lo cual quiere decir que si se seleccionan 100 muestras de tamaño 25, y se construyen 100 intervalos de confianza como $\hat{\beta}_1 \pm t_{(\alpha/2,\, n-3)} es(\hat{\beta}_1)$, podemos esperar que 95 de ellos contengan el verdadero parámetro poblacional β_1.

Como se puede observar el valor hipotético nulo de cero no cae dentro del intervalo dado en la ecuación (4.52), podemos rechazar la hipótesis nula según la cual $\beta_1 = 0$ con un coeficiente de confianza del 95 por ciento. Así pues, usando la prueba de significancia o la estimación del intervalo de confianza, llegamos a la misma conclusión, cosa que no debe sorprendernos en razón del vinculo entre la estimación de intervalos de confianza y las pruebas de hipótesis.

Ejercicios 4

1. Se lleva a cabo un experimento para determinar si el peso de un animal se puede predecir después de un tiempo dado, sobre la base del peso inicial del animal y la cantidad de alimento que consume. Se registran los datos siguientes en kilogramos:

Peso final (kg.)	Peso inicial (kg.)	Alimento consumido (kg.)
95	42	272
77	33	226
80	33	259
100	45	292
97	39	311
70	36	183
50	32	173
80	41	236
92	40	230
84	38	235

h) Determinar la ecuación de regresión múltiple.

i) Calcular los coeficientes de correlación parcial e interpretarlos.

j) Calcular el coeficiente de determinación e interpretarlo.

k) Calcular el peso final de un animal cuando el peso inicial es 45 kg. y 250 kg. de alimento consumido.

l) Realizar la prueba de hipótesis individual y global de los coeficientes de regresión.

m) Encontrar los intervalos de confianza para los coeficientes de regresión.

2. La tensión de la pierna es un ingrediente necesario para un pateador exitoso en el fútbol americano. Una medida de la calidad de una buena patada es la distancia a la que se lanza el ovoide (pelota en forma de huevo). Para determinar si la tensión de las piernas influye en la distancia de pateo, se eligieron 13 pateadores para el experimento y cada uno pateó 10 veces un ovoide. La distancia promedio en pies, junto con la tensión en libras, se registraron como sigue:

Distancia (pies)	Tensión pierna izq. (lbs.)	Tensión pierna der. (lbs.)
162.50	170	170
144.00	130	140
105.67	110	120
147.50	170	180
117.59	120	130
163.50	160	160
140.25	140	120
192.50	150	170
150.17	130	140
171.75	150	150
165.16	150	160
162.00	180	170
104.93	110	110

a) Determinar la ecuación de regresión múltiple.

b) Calcular el coeficiente de determinación e interpretarlo.

c) Calcular el coeficiente de determinación ajustado.

d) Calcule la distancia de pateo de un jugador con tensión en ambas piernas de 145 lbs.

e) Realizar la prueba de hipótesis individual y global de los coeficientes de regresión.

f) Encontrar los intervalos de confianza para los coeficientes de regresión.

3. Los datos de la siguiente tabla corresponden a un estudio sobre la contaminación acústica realizado en distintas zonas de la misma ciudad. La variable "y" mide la contaminación acústica en decibelios, la variable x_1 la hora del día y x_2 el tráfico de vehículos por minuto.

Decibelios	0.9	1.6	4.7	2.8	5.6	2.4	1.0	1.5
Hora	14	15	16	13	17	18	19	20
Trafico de Vehículos (min.)	1	2	5	2	6	4	3	4

a) Determinar la ecuación de regresión múltiple.

b) Calcular el coeficiente de determinación e interpretarlo.

c) Calcular el coeficiente de determinación ajustado.

d) Realizar la prueba de hipótesis individual y global de los coeficientes de regresión.

4. Se lleva a cabo un conjunto de eventos experimentales para determinar una forma de pronosticar los tiempos de cocimiento "y", a varios niveles de ancho de horno x_1 y la temperatura de los conductos interiores x_2. Los datos codificados se registran como se muestra a continuación.

y (º)	**x_1 (cm.)**	**x_2 (º)**
6.40	1.32	1.15
15.05	2.69	3.40
18.75	3.56	4.10
30.25	4.41	8.75
44.85	5.35	14.82
48.94	6.20	15.15
51.55	7.12	15.32
61.50	8.87	18.18
100.44	9.80	35.19
111.42	10.65	40.40

a) Determinar la ecuación de regresión múltiple.

b) Calcular el coeficiente de determinación e interpretarlo.

c) Calcular el coeficiente de determinación ajustado.

d) Realizar la prueba de hipótesis individual y global de los coeficientes de regresión.

e) Realizar la estimación por intervalo para un $\alpha = 0.05$

5. En el diseño de un modelo de simulación necesitamos disponer de una función de consumo de bienes de origen industrial; para lograrlo tenemos los siguientes datos:

Años	**y ($)**	**x_1 ($)**	**x_2 ($)**
1970	45	52	10
1971	42	58	13
1972	48	58	10
1973	55	60	14
1974	53	65	16
1975	65	70	18

Donde:

y: Consumo de bienes industriales (medido en unidades monetarias constantes).

x_1: Ingreso disponible (monetarias constantes).

x_2: Importaciones de bienes de consumo.

a) Determinar la ecuación de regresión múltiple.

b) Calcular el coeficiente de determinación e interpretarlo.

c) Estimar el consumo de bienes industriales para 1976; si asumimos que para dicho año el ingreso disponible fue de 72 y las importaciones de bienes de consumo 17.

6. Se quiere disponer de estimaciones de las variaciones en los precios de bienes agrícolas de consumo esencial. Para lograrlo, después de algunos estudios, se concluyó que una metodología posible podría ser el ajuste de una ecuación de regresión a los siguientes datos:

Período	**y (% precios)**	**x_1 (% costo unitario)**	**x_2 (%)**
1	7	6	11
2	9	7	14
3	11	12	7
4	12	13	12
5	14	15	21
6	22	23	21
7	25	24	14

Donde:

y: Porcentaje de los precios de bienes agrícolas.

x_1: Porcentaje del costo unitario de producción.

x_2: Tasa de inflación (%).

a) Calcular la ecuación de regresión múltiple.

b) Calcular el coeficiente de determinación e interpretarlo.

7. Los datos de la tabla que se muestran a continuación son mediciones realizadas a 9 niños con el propósito de llegar a una ecuación de estimación que se relacione con su estatura al nacer y con su edad en número de días.

Estatura del niño (cm.)	Edad (días)	Estatura al nacer (cm.)
57.5	78	48.2
52.8	69	45.5
61.3	77	46.3
67.0	88	49.0
53.5	67	43.0
62.7	80	48.0
56.2	74	48.0
68.5	94	53.0
69.2	102	58.0

a) Determinar la ecuación de regresión múltiple.

b) Calcular el coeficiente de determinación e interpretarlo.

c) Calcular el coeficiente de determinación ajustado.

d) Realizar la prueba de hipótesis individual y global de los coeficientes de regresión.

e) Realizar la estimación por intervalo para un $\alpha = 0.05$

8. Se quiere determinar si la demanda de café, depende del precio del café y del precio del cacao. Para ello se presentan los datos en la siguiente tabla.

Demanda de café, y ($)	**Precio de café, x_1 ($)**	**Precio de cacao, x_2 ($)**
10	3	5
8	5	4
5	4	3
6	8	2
2	10	2

a) Hacer un diagrama de dispersión tridimensional.

b) Determinar la ecuación de regresión múltiple.

c) Calcular el coeficiente de determinación e interpretarlo.

d) Calcular el coeficiente de determinación ajustado.

e) Realizar la prueba de hipótesis individual y global de los coeficientes de regresión.

f) Realizar la estimación por intervalo para un $\alpha = 0.05$.

g) Realizar la predicción para nuevos valores del café y el cacao, donde: $x_1 = 12$ y $x_2 = 2$.

Apéndice 4: Deducción de Ecuaciones.

4.1 Deducción de ecuaciones utilizadas en el Capítulo 4.

a) Deducción de $\hat{\beta}_o$ ecuación (4.15).

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i$$
$$y_i = \hat{y}_i + e_i$$
$$y_i - \hat{y}_i = e_i$$

y $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ entonces, $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2$

Derivando parcialmente ambos lados de la ecuación anterior con respecto a $\hat{\beta}_0$ se tiene:

$$\frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2\right) = \frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=1}^{n} e_i^2\right)$$
$$2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})(-1) = 0$$
$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) = 0$$
$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_{1i} - \sum_{i=1}^{n} \hat{\beta}_2 x_{2i} = 0$$
$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} = 0$$
$$\sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} = n\hat{\beta}_0$$
$$\frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_{1i}}{n} - \hat{\beta}_2 \frac{\sum_{i=1}^{n} x_{2i}}{n} = \hat{\beta}_0$$
$$\bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2 = \hat{\beta}_0$$

L.q.q.d

Como ya conocemos $\hat{\beta}_0$, sustituimos en la ecuación $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2$ para expresarla en función de $\hat{\beta}_1$ y $\hat{\beta}_2$, derivando con respecto a $\hat{\beta}_1$ se tiene la ecuación **(1)** así:

$$\sum_{i=1}^{n}(y_i - (\bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2) - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 = \sum_{i=1}^{n} e_i^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y} + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 = \sum_{i=1}^{n} e_i^2$$

$$\sum_{i=1}^{n}[(y_i - \bar{y}) - \hat{\beta}_1(x_{1i} - \bar{x}_1) - \hat{\beta}_2(x_{2i} - \bar{x}_2)]^2 = \sum_{i=1}^{n} e_i^2$$

$$\frac{\partial}{\partial\hat{\beta}_1}\left(\sum_{i=1}^{n}[(y_i - \bar{y}) - \hat{\beta}_1(x_{1i} - \bar{x}_1) - \hat{\beta}_2(x_{2i} - \bar{x}_2)]^2\right) = \frac{\partial}{\partial\hat{\beta}_1}\left(\sum_{i=1}^{n} e_i^2\right)$$

$$2\sum_{i=1}^{n}[(y_i - \bar{y}) - \hat{\beta}_1(x_{1i} - \bar{x}_1) - \hat{\beta}_2(x_{2i} - \bar{x}_2)](x_{1i} - \bar{x}_1)(-1) = 0$$

$$\sum_{i=1}^{n}[(y_i - \bar{y}) - \hat{\beta}_1(x_{1i} - \bar{x}_1) - \hat{\beta}_2(x_{2i} - \bar{x}_2)](x_{1i} - \bar{x}_1) = 0$$

$$\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y}) - \hat{\beta}_1\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 - \hat{\beta}_2\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = 0$$

$$\hat{\beta}_1\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 + \hat{\beta}_2\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) = \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y}) \qquad \textbf{(1)}$$

Derivando con respecto a $\hat{\beta}_2$ se tiene la ecuación **(2)** así:

$$\sum_{i=1}^{n}(y_i - (\bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2) - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 = \sum_{i=1}^{n} e_i^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y} + \hat{\beta}_1\bar{x}_1 + \hat{\beta}_2\bar{x}_2 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})^2 = \sum_{i=1}^{n} e_i^2$$

$$\sum_{i=1}^{n}[(y_i - \bar{y}) - \hat{\beta}_1(x_{1i} - \bar{x}_1) - \hat{\beta}_2(x_{2i} - \bar{x}_2)]^2 = \sum_{i=1}^{n} e_i^2$$

$$\frac{\partial}{\partial\hat{\beta}_2}\left(\sum_{i=1}^{n}\left[(y_i-\bar{y})-\hat{\beta}_1(x_{1i}-\bar{x}_1)-\hat{\beta}_2(x_{2i}-\bar{x}_2)\right]^2\right)=\frac{\partial}{\partial\hat{\beta}_2}\left(\sum_{i=1}^{n}e_i^2\right)$$

$$2\sum_{i=1}^{n}\left[(y_i-\bar{y})-\hat{\beta}_1(x_{1i}-\bar{x}_1)-\hat{\beta}_2(x_{2i}-\bar{x}_2)\right](x_{2i}-\bar{x}_2)(-1)=0$$

$$\sum_{i=1}^{n}\left[(y_i-\bar{y})-\hat{\beta}_1(x_{1i}-\bar{x}_1)-\hat{\beta}_2(x_{2i}-\bar{x}_2)\right](x_{2i}-\bar{x}_2)=0$$

$$\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)(y_i-\bar{y})-\hat{\beta}_1\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)-\hat{\beta}_2\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2=0$$

$$\hat{\beta}_1\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)+\hat{\beta}_2\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2=\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)(y_i-\bar{y}) \quad \textbf{(2)}$$

b) Deducción de $\hat{\beta}_1$ ecuación (4.16).

Tomando las ecuaciones **(1)** y **(2)** y para facilitar el proceso hacemos:

$$\sum_{i=1}^{n}X_{12}=\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)$$

$$\sum_{i=1}^{n}X_1=\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)\ \text{ y }\ \sum_{i=1}^{n}X_1^2=\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2$$

$$\sum_{i=1}^{n}X_2=\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)\ \text{ y }\ \sum_{i=1}^{n}X_2^2=\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2$$

$$\sum_{i=1}^{n}Y_i=\sum_{i=1}^{n}(y_i-\bar{y})$$

$$\sum_{i=1}^{n}X_1Y_i=\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(y_i-\bar{y})$$

$$\sum_{i=1}^{n}X_2Y_i=\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)(y_i-\bar{y})$$

y multiplicando **(1)** por $\sum_{i=1}^{n} X_2^2$ y **(2)** por $-\sum_{i=1}^{n} X_{12}$ se tiene:

$$\hat{\beta}_1 \sum_{i=1}^{n} X_1^2 \sum_{i=1}^{n} X_2^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{12} \sum_{i=1}^{n} X_2^2 = \sum_{i=1}^{n} X_1 Y_i \sum_{i=1}^{n} X_2^2 \qquad (1)$$

$$-\hat{\beta}_1 \sum_{i=1}^{n} X_{12} \sum_{i=1}^{n} X_{12} - \hat{\beta}_2 \sum_{i=1}^{n} X_2^2 \sum_{i=1}^{n} X_{12} = -\sum_{i=1}^{n} X_2 Y_i \sum_{i=1}^{n} X_{12} \qquad (2)$$

Sumando las ecuaciones (1) y (2) tenemos:

$$\hat{\beta}_1 \sum_{i=1}^{n} X_1^2 \sum_{i=1}^{n} X_2^2 - \hat{\beta}_1 \sum_{i=1}^{n} X_{12} \sum_{i=1}^{n} X_{12} = \sum_{i=1}^{n} X_1 Y_i \sum_{i=1}^{n} X_2^2 - \sum_{i=1}^{n} X_2 Y_i \sum_{i=1}^{n} X_{12}$$

$$\hat{\beta}_1 \left[\sum_{i=1}^{n} X_1^2 \sum_{i=1}^{n} X_2^2 - \left(\sum_{i=1}^{n} X_{12} \right)^2 \right] = \sum_{i=1}^{n} X_1 Y_i \sum_{i=1}^{n} X_2^2 - \sum_{i=1}^{n} X_2 Y_i \sum_{i=1}^{n} X_{12}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_1 Y_i \sum_{i=1}^{n} X_2^2 - \sum_{i=1}^{n} X_2 Y_i \sum_{i=1}^{n} X_{12}}{\sum_{i=1}^{n} X_1^2 \sum_{i=1}^{n} X_2^2 - \left(\sum_{i=1}^{n} X_{12} \right)^2}$$

Sustituyendo en $\hat{\beta}_1$ las ecuaciones originales se tiene:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 - \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2}$$

L.q.q.d

c) Deducción de $\hat{\beta}_2$ ecuación (4.17).

Para despejar $\hat{\beta}_2$ multiplicamos **(1)** por $-\sum_{i=1}^{n} X_{12}$ y **(2)** por $\sum_{i=1}^{n} X_1^2$ se tiene:

$$-\hat{\beta}_1 \sum_{i=1}^{n} X_1^2 \sum_{i=1}^{n} X_{12} - \hat{\beta}_2 \sum_{i=1}^{n} X_{12} \sum_{i=1}^{n} X_{12} = -\sum_{i=1}^{n} X_1 Y_i \sum_{i=1}^{n} X_{12} \qquad (1)$$

$$\hat{\beta}_1 \sum_{i=1}^{n} X_{12} \sum_{i=1}^{n} X_1^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_2^2 \sum_{i=1}^{n} X_1^2 = \sum_{i=1}^{n} X_2 Y_i \sum_{i=1}^{n} X_1^2 \qquad (2)$$

Sumando las ecuaciones (1) y (2) obtenemos:

$$\hat{\beta}_2 \sum_{i=1}^{n} X_2^2 \sum_{i=1}^{n} X_1^2 - \hat{\beta}_2 \sum_{i=1}^{n} X_{12} \sum_{i=1}^{n} X_{12} = -\sum_{i=1}^{n} X_1 Y_i \sum_{i=1}^{n} X_{12} + \sum_{i=1}^{n} X_2 Y_i \sum_{i=1}^{n} X_1^2$$

$$\hat{\beta}_2 \left[\sum_{i=1}^{n} X_2^2 \sum_{i=1}^{n} X_1^2 - \left(\sum_{i=1}^{n} X_{12} \right)^2 \right] = \sum_{i=1}^{n} X_2 Y_i \sum_{i=1}^{n} X_1^2 - \sum_{i=1}^{n} X_1 Y_i \sum_{i=1}^{n} X_{12}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} X_2 Y_i \sum_{i=1}^{n} X_1^2 - \sum_{i=1}^{n} X_1 Y_i \sum_{i=1}^{n} X_{12}}{\sum_{i=1}^{n} X_1^2 \sum_{i=1}^{n} X_2^2 - \left(\sum_{i=1}^{n} X_{12} \right)^2}$$

Sustituyendo en $\hat{\beta}_2$ las ecuaciones originales se tiene:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 - \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2 - \left(\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) \right)^2}$$

L.q.q.d

d) Deducción de la ecuación (4.25).

La ecuación de regresión estimada está dada por: $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i$ a partir de la cual se puede despejar el e_i como se muestra a continuación:

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}$$
$$e_i = y_i - (\bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2) - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}$$
$$e_i = y_i - \bar{y} + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}$$
$$e_i = (y_i - \bar{y}) - \hat{\beta}_1 (x_{1i} - \bar{x}_1) - \hat{\beta}_2 (x_{2i} - \bar{x}_2)$$

Entonces:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (e_i e_i)$$
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} e_i \left[(y_i - \bar{y}) - \hat{\beta}_1 (x_{1i} - \bar{x}_1) - \hat{\beta}_2 (x_{2i} - \bar{x}_2) \right]$$
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} e_i (y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n} e_i (x_{1i} - \bar{x}_1) - \hat{\beta}_2 \sum_{i=1}^{n} e_i (x_{2i} - \bar{x}_2)$$

Donde:

$$\hat{\beta}_1 \sum_{i=1}^{n} e_i (x_{1i} - \bar{x}_1) = 0 \quad \text{y} \quad \hat{\beta}_2 \sum_{i=1}^{n} e_i (x_{2i} - \bar{x}_2) = 0$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} e_i (y_i - \bar{y})$$
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y}) e_i$$
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y}) \left[(y_i - \bar{y}) - \hat{\beta}_1 (x_{1i} - \bar{x}_1) - \hat{\beta}_2 (x_{2i} - \bar{x}_2) \right]$$
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) - \hat{\beta}_2 (x_{2i} - \bar{x}_2)(y_i - \bar{y})$$

L.q.q.d

e) Deducción de la ecuación (4.27).

Como R^2 es una medida de bondad del ajuste en el modelo de regresión múltiple, para cada observación podemos descomponer la diferencia entre y_i y su media $\bar{y}$ como sigue:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Elevando al cuadrado ambos lados de la ecuación anterior y aplicando sumatoria tenemos:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \left[\sum_{i=1}^{n}(y_i - \hat{y}_i) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})\right]^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)\sum_{i=1}^{n}(\hat{y}_i - \bar{y})$$

Pero el último término es idénticamente cero ya que:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}e_i(\hat{y}_i - \bar{y})$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}e_i\hat{y}_i - \sum_{i=1}^{n}e_i\bar{y}$$

Donde $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_k x_{ki}$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = \sum_{i=1}^{n}e_i(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_k x_{ki}) - \bar{y}\sum_{i=1}^{n}e_i$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = \hat{\beta}_0\sum_{i=1}^{n}e_i + \hat{\beta}_1\sum_{i=1}^{n}e_i x_{1i} + ... + \hat{\beta}_k\sum_{i=1}^{n}e_i x_{ki} - \bar{y}\sum_{i=1}^{n}e_i$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)\sum_{i=1}^{n}(\hat{y}_i - \bar{y}) = 0$$

Dado que $\sum_{i=1}^{n}e_i = 0$ y $\sum_{i=1}^{n}e_i x_{ji} = 0$ para j = 1, 2,…, k

Así:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$$

L.q.q.d

f) Deducción de la ecuación (4.32).

Para ver la relación entre $\bar{R}^2$ y R^2 sustituimos

$R^2=1-\dfrac{\sum_{i=1}^{n}e_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$ en $\bar{R}^2=1-\dfrac{\sum_{i=1}^{n}e_i^2\,/(n-k)}{\sum_{i=1}^{n}(y_i-\bar{y})^2\,/(n-1)}$ y se tiene entonces:

$$R^2=1-\frac{\sum_{i=1}^{n}e_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$R^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}e_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$R^2\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\bar{y})^2-\sum_{i=1}^{n}e_i^2$$

$$\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\bar{y})^2-R^2\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\bar{y})^2(1-R^2)$$

Sustituyendo este resultado en $\bar{R}^2$ tenemos que:

$$\bar{R}^2=1-\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2(1-R^2)}{\sum_{i=1}^{n}(y_i-\bar{y})^2}*\frac{(n-1)}{(n-k)}$$

$$\bar{R}^2=1-(1-R^2)*\frac{(n-1)}{(n-k)}$$

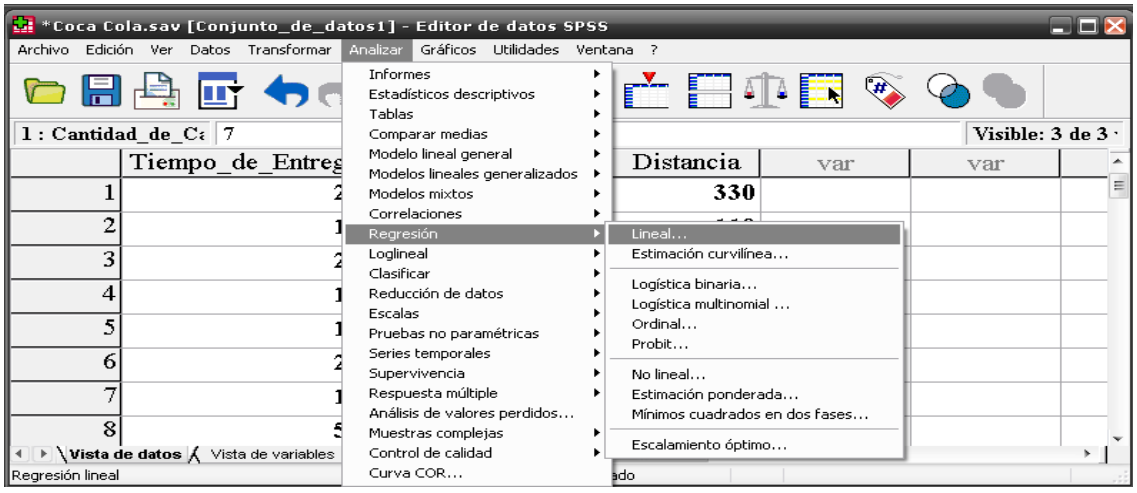
L.q.q.d

Apéndice 4.2: Solución de Ejemplos Haciendo uso del Software Estadístico SPSS v15.0

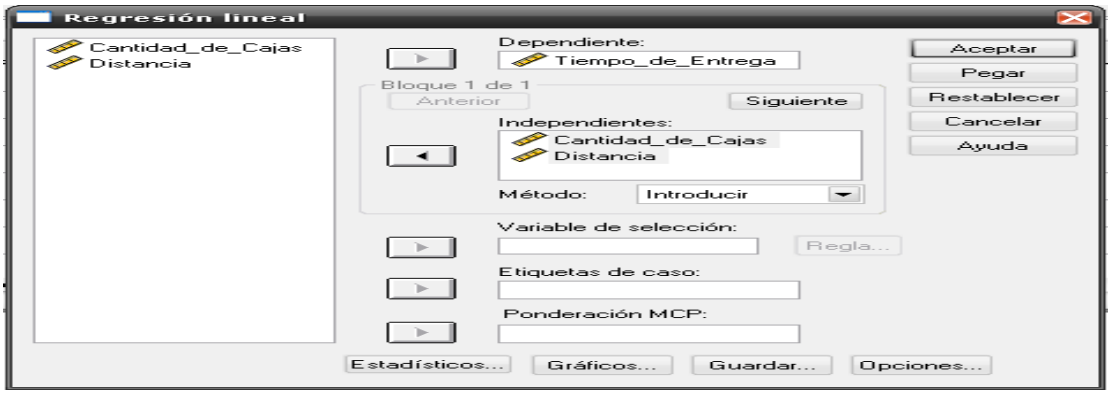
Haciendo uso del software se pueden obtener los resultados de los ejemplos 1, 2, 3, 4, 5, 6 en una sola ejecución siguiendo los siguientes pasos.

1. Se les da un nombre a las tres variables en estudio, se digitan los datos para cada variable y se obtiene la ventana siguiente en la cual solamente se muestran 8 observaciones del total (25).

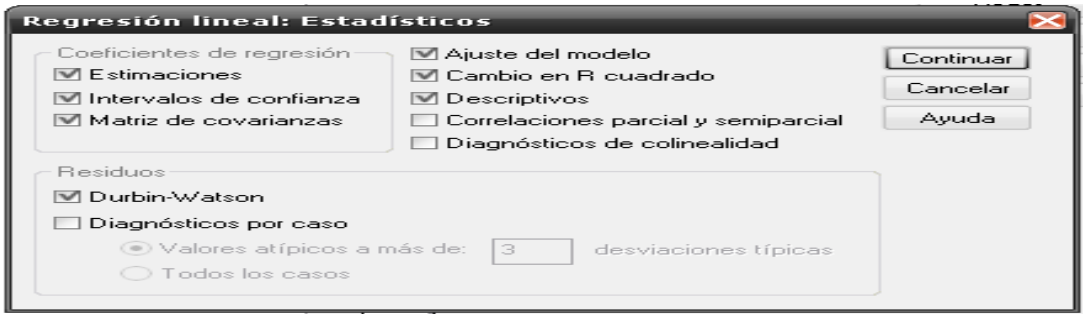
2. En la barra de menú se selecciona la opción Analizar → Regresión → Lineal como se muestra a continuación.

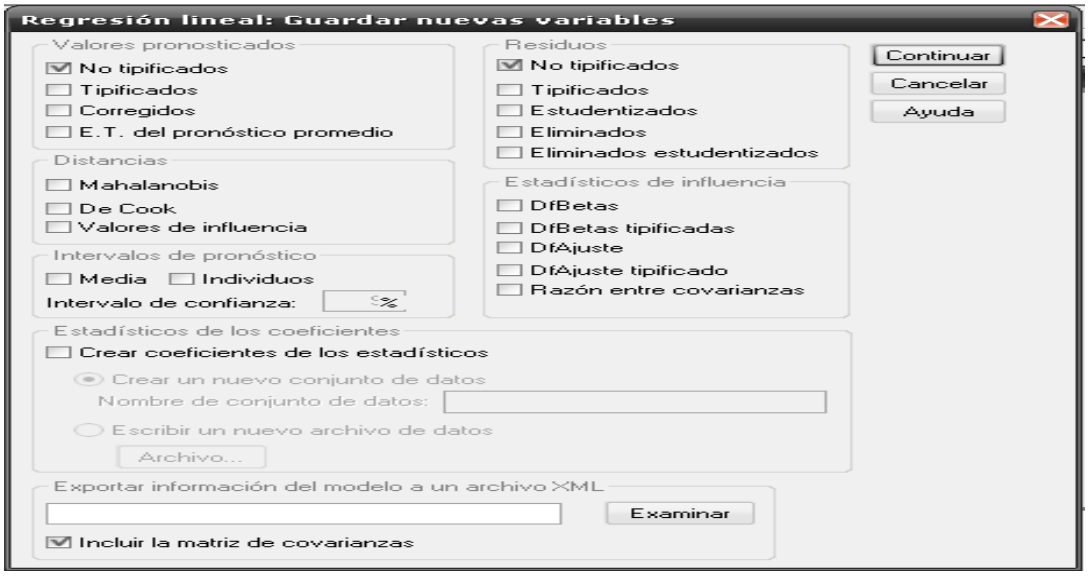


3. Al hacer click en la opción lineal aparece la siguiente ventana en la cual se colocan las variables cada una en su lugar en este caso hay dos variables independientes.



Al pulsar en los botones **Estadístico** y **Guardar** aparecen los cuadros siguientes:





Dando un click en el botón aceptar aparecen los siguientes resultados:

Variables introducidas/eliminadas[b]

Modelo	Variables introducidas	Variables eliminadas	Método
1	Distancia, Cantidad_de_Cajas[a]	.	Introducir

a. Todas las variables solicitadas introducidas

b. Variable dependiente: Tiempo_de_Entrega

En la tabla de variables introducidas se observa que no se ha eliminado ninguna variable

Estadísticos descriptivos

	Media	Desviación típ.	N
Tiempo_de_Entrega	27.76	12.228	25
Cantidad_de_Cajas	8.76	6.882	25
Distancia	409.28	325.188	25

La tabla de estadísticos descriptivos muestra la media, que son exactamente las obtenidas en el ejemplo y la desviación típica para cada una de las variables, también puede observarse que aparece el número de observaciones.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		t	Sig.	Intervalo de confianza para B al 95%	
		B	Error típ.			Límite inferior	Límite superior
1	(Constante)	12.610	1.500	8.406	.000	9.499	15.720
	Cantidad_de_Cajas	1.121	.234	4.799	.000	.636	1.605
	Distancia	.013	.005	2.637	.015	.003	.023

a. Variable dependiente: Tiempo_de_Entrega

Los valores obtenidos en la tabla son iguales a los obtenidos en el ejemplo haciendo uso de las ecuaciones mostradas en este capítulo, y algunas diferencias que se dan son debido a las aproximaciones que se hacen.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	3151.299	2	1575.650	79.276	.000[a]
	Residual	437.261	22	19.875		
	Total	3588.560	24			

a. Variables predictoras: (Constante), Distancia, Cantidad_de_Cajas

b. Variable dependiente: Tiempo_de_Entrega

La tabla ANOVA es la del análisis de varianza en la cual se presenta un resumen de los valores que se necesitan para realizar la prueba de hipótesis global de los parámetros de regresión, se puede ver que los valores son casi iguales salvo por algunas aproximaciones.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				Durbin-Watson
					Cambio en F	gl1	gl2	Sig. del cambio en F	
1	.937[a]	.878	.867	4.458	79.276	2	22	.000	1.811

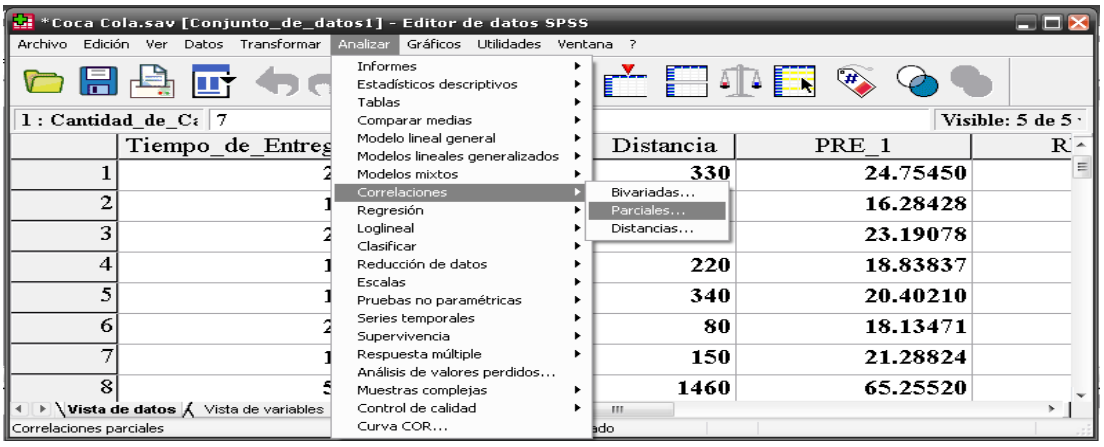
a. Variables predictoras: (Constante), Distancia, Cantidad_de_Cajas

b. Variable dependiente: Tiempo_de_Entrega

En la tabla resumen del modelo se muestran datos interesantes que se necesitan para ver si el modelo que hemos ajustado es bueno, se muestran los valores de los R los cuales son iguales a los obtenidos haciendo uso de las ecuaciones, el valor del estadístico F, los grados de libertad y el valor del estadístico d = 1.811 que como se dijo en el Capítulo 3, cuando el valor de Durbin-Watson se encuentra entre 1.5 y 2.5 podemos asumir independencia entre los residuos, en este caso los valores de los R y el valor de Durbin-Watson muestran que el modelo ajustado es adecuado.

Ahora calculamos la correlación simple y parcial de las variables siguiendo los pasos que se muestran a continuación:

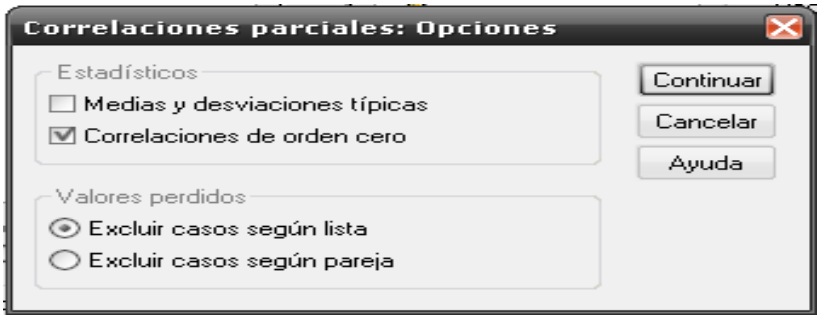
4. Después haber hecho la regresión de las variables hacemos la correlación simple o de orden cero y la correlación parcial o de orden uno, con los datos que ya se tienen, recuérdese que con la correlación lo que se quiere ver es la asociación que existe entre las variables. En la barra de menú se selecciona la opción Analizar → Correlación → Parciales como se muestra a continuación.



5. Dando click en la opción parciales se obtiene la ventana siguiente:

En la que se han trasladado las variables Tiempo de Entrega y Cantidad de Cajas a la primera casilla, y en la segunda casilla se ha trasladado la variable Distancia que se mantendrá constante en este caso.

Luego dando un click en opciones se muestra la ventana siguiente:



En la que se ha seleccionado la opción correlaciones de orden cero, es decir que los resultados que se obtendrán serán de las correlaciones de orden cero y de orden 1 como se muestra, dando click en aceptar se tiene:

Correlaciones

Variables de control			Tiempo_de_ Entrega	Cantidad_de_ Cajas	Distancia
-ninguno-[a]	Tiempo_de_Entrega	Correlación	1.000	.916	.866
		Significación (bilateral)	.	.000	.000
		gl	0	23	23
	Cantidad_de_Cajas	Correlación	.916	1.000	.824
		Significación (bilateral)	.000	.	.000
		gl	23	0	23
	Distancia	Correlación	.866	.824	1.000
		Significación (bilateral)	.000	.000	.
		gl	23	23	0
Distancia	Tiempo_de_Entrega	Correlación	1.000	.715	
		Significación (bilateral)	.	.000	
		gl	0	22	
	Cantidad_de_Cajas	Correlación	.715	1.000	
		Significación (bilateral)	.000	.	
		gl	22	0	

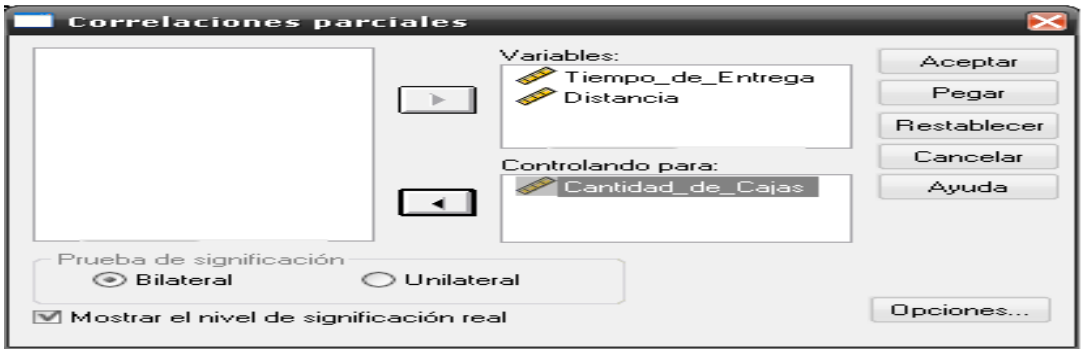
a. Las casillas contienen correlaciones de orden cero (de Pearson).

La tabla correlaciones está dividida en tres filas en la primer fila está el nombre de las variables, en la segunda fila se tienen los valores obtenidos para la correlación simple entre cada par de variables estos son: la correlación entre Tiempo de Entrega y Cantidad de Cajas es r = 0.916, entre Tiempo de Entrega y Distancia recorrida es

r = 0.866 y entre Cantidad de Cajas y Distancia recorrida es r = 0.824 igual a los obtenidos en el desarrollo del ejemplo 1 en este Capítulo.

En la tercer fila de la tabla se tienen la correlación parcial entre las variables Tiempo de Entrega y Cantidad de Cajas manteniendo constante o controlando la variable Distancia recorrida, y se puede ver que la correlación en este caso es de 0.715 igual que antes.

Hacemos la correlación entre la variable Tiempo de Entrega y Distancia recorrida y mantenemos constante la variable Cantidad de Cajas así:



Correlaciones

Variables de control			Tiempo_de_Entrega	Distancia	Cantidad_de_Cajas
-ninguno-[a]	Tiempo_de_Entrega	Correlación	1.000	.866	.916
		Significación (bilateral)	.	.000	.000
		gl	0	23	23
	Distancia	Correlación	.866	1.000	.824
		Significación (bilateral)	.000	.	.000
		gl	23	0	23
	Cantidad_de_Cajas	Correlación	.916	.824	1.000
		Significación (bilateral)	.000	.000	.
		gl	23	23	0
Cantidad_de_Cajas	Tiempo_de_Entrega	Correlación	1.000	.490	
		Significación (bilateral)	.	.015	
		gl	0	22	
	Distancia	Correlación	.490	1.000	
		Significación (bilateral)	.015	.	
		gl	22	0	

a. Las casillas contienen correlaciones de orden cero (de Pearson).

Se puede observar que los valores de la segunda fila, solamente han cambiado de posición y son iguales a los obtenidos en la tabla correlaciones que se mostró anteriormente. Pero los valores de la tercer fila son distintos porque es la correlación de las variables Tiempo de Entrega y Distancia recorrida manteniendo constante la variable Cantidad de Cajas, el valor en este caso es r = 0.490.

De la misma forma se puede obtener la correlación entre las variables Cantidad de Cajas y Distancia recorrida manteniendo constante la variable Tiempo de Entrega.

Para la elaboración de los gráficos solamente se sigue la ruta

Gráficos → Interactivos → Diagramas de dispersión → Coordenadas → Traslado de variables→Ajuste→ Regresión.

Capítulo 5

Modelo de Regresión Lineal Múltiple Haciendo Uso del Algebra Matricial.

5.1 Introducción al Modelo de Regresión Lineal Múltiple.

Este Capítulo presenta el modelo de regresión lineal con k variables ("y" y x_1, x_2,…, x_k) en notación de álgebra matricial. Conceptualmente, el modelo de k variables es una extensión lógica de los modelos de dos y tres variables que se han visto hasta el momento. Por esta razón, el presente Capítulo muestra muy pocos conceptos nuevos salvo la notación matricial.

Una gran ventaja del álgebra matricial sobre el álgebra escalar (algebra elemental que trata con escalares o números reales) consiste en que proporciona un método resumido para manejar los modelos de regresión con cualquier número de variables independientes; una vez formulado el modelo de k variables y resuelto en notación matricial, la solución se puede aplicar a una, dos, tres o cualquier número de variables.

5.2 Definición de Términos Básicos.

Correlación Serial: Existe cuando las observaciones sucesivas a través del tiempo se relacionan entre sí.

Escalar: El escalar es un solo número real. Dicho de otra forma, un escalar es una matriz de 1*1.

Matriz: Es una disposición de números u otros elementos en M filas y N columnas.

Matriz Transpuesta: La transpuesta de una matriz **A** de orden (M * N) es una matriz $\mathbf{A}'$ (N * M), obtenida mediante el intercambio de filas y columnas.

Matriz Cuadrada: Una matriz es cuadrada si el número de filas es igual al número de columnas.

Matriz Simétrica: Una matriz cuadrada es simétrica si se verifica que la transpuesta es igual a ella misma.

Matriz Identidad: Es una matriz cuyos elementos de la diagonal son todos iguales a 1 y se simboliza con **I**.

Vector Columna: Un vector columna es una ordenación de elementos dispuestos en M filas y 1 columna.

Vector Fila: Un vector fila es una ordenación de elementos dispuestos en 1 fila y N columnas. La transpuesta de un vector fila es un vector columna.

Vector Nulo: Es el vector fila o columna cuyos elementos son todos cero.

5.3 Modelo de Regresión Lineal con k Variables.

Generalizando el modelo de regresión lineal de dos y tres variables, el modelo de regresión poblacional (FRP) de k variables que tiene la variable dependiente "y" y k-1 variables explicatorias $x_1, x_2,..., x_k$, puede escribirse de la siguiente forma:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i \qquad i = 1, 2, 3,\ldots, n \qquad \textbf{(5.1)}$$

Donde:

β_0: Es el intercepto.

β_1 a β_k: Coeficientes (pendientes) parciales.

ε_i: Error aleatorio.

i : Es la i-ésima observación.

La ecuación (5.1) se debe interpretar como se vio en el Capítulo 4: nos proporciona la media o valor esperado de "y" condicional a los valores fijos (en muestras repetidas) de $x_1, x_2,..., x_k$, es decir $E(y \mid x_{1i}, x_{2i},..., x_{ki})$.

La ecuación (5.1) es una expresión abreviada del siguiente conjunto de n ecuaciones simultáneas:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \ldots + \beta_k x_{k1} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \ldots + \beta_k x_{k2} + \varepsilon_2 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \ldots + \beta_k x_{kn} + \varepsilon_n \end{aligned} \qquad \textbf{(5.2)}$$

Escribiendo el sistema de ecuaciones (5.2) en forma alterna pero más ilustrativa se tiene:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \qquad \textbf{(5.3)}$$

$$\begin{array}{ccccccc} \mathbf{y} & = & & \mathbf{x} & & \boldsymbol{\beta} & + \quad \boldsymbol{\varepsilon} \\ n * 1 & & & n * k & & k * 1 & \quad n * 1 \end{array}$$

Donde[1]:

y: Es un vector columna n * 1 de observaciones de la variable dependiente "y".

x: Es una matriz n * k que nos da n observaciones de las k-1 variables de x_1 a x_k. La primera columna de 1's representa el intercepto. (Esta matriz se conoce también como la matriz de observaciones).

β: Es un vector columna k * 1 de los parámetros desconocidos $\beta_1, \beta_2, \ldots, \beta_k$.

ε: Es un vector columna n * 1 de las n perturbaciones ε_i.

El sistema de ecuaciones dado en (5.3) se conoce como la representación matricial del modelo de regresión lineal general (de k variables). Se puede escribir de forma resumida como:

$$\begin{array}{ccccccc} \mathbf{y} & = & \mathbf{x} & & \boldsymbol{\beta} & + & \boldsymbol{\varepsilon} \\ n * 1 & & n * k & & k * 1 & & n * 1 \end{array} \qquad \textbf{(5.4)}$$

Donde no hay confusión a cerca de las dimensiones u orden de la matriz **x** y de los vectores **y**, β y ε. La ecuación (5.4) puede escribirse simplemente como:

$$\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \textbf{(5.5)}$$

A manera de ilustración de la representación matricial, se considera el modelo de dos

[1] Los vectores y las matrices se denotaran por letras minúsculas en negritas.

variables de las horas dedicadas hacer deporte y el número de pulsaciones, visto en el ejemplo 2 del Capítulo 2, o sea, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Donde:

y: Pulsaciones.

x: Hs Deporte.

Usando los datos de la tabla 2.4, la expresión matricial es:

$$\begin{bmatrix} 66 \\ 62 \\ 73 \\ 72 \\ 65 \\ 60 \\ 66 \\ 58 \\ 57 \\ 54 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 3 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 7 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \end{bmatrix} \qquad \textbf{(5.6)}$$

$$\underset{10*1}{\mathbf{y}} = \underset{10*2}{\mathbf{x}} \quad \underset{2*1}{\boldsymbol{\beta}} + \underset{10*1}{\boldsymbol{\varepsilon}}$$

Como en los casos de dos y tres variables nuestro objetivo es el de estimar los parámetros de la regresión múltiple ecuación (5.1) y hacer inferencias a cerca de ellos con la información disponible. En notación matricial esto equivale a estimar β y hacer inferencias a cerca de este β. Para la estimación de los parámetros se puede utilizar el método de Mínimos Cuadrados Ordinarios (MCO) o el de Máxima Verosimilitud (MV).

Pero como se mostró anteriormente, estos dos métodos nos proporcionan estimadores idénticos de los coeficientes de regresión. Por lo tanto utilizamos el método de Mínimos Cuadrados Ordinarios para la estimación de los parámetros.

5.4 Asunciones del Modelo de Regresión Lineal con k Variables en Notación Matricial.

Supondremos que se desea explicar los valores de una variable aleatoria "y" por un conjunto de k variables matemáticas (x_1, x_2,…,x_k), que toman en los elementos estudiados valores predeterminados conocidos. La relación entre estas variables es como se presentó en la ecuación (5.1) donde y_i es el valor de la variable dependiente en el elemento i, x_{1i},.., x_{ki} los valores de las variables independientes, cada coeficiente β_i mide el efecto marginal sobre la variable dependiente de un aumento unitario en la variable independiente x_i cuando el resto de las variables independientes permanecen constantes y el término ε_i es, como en modelos anteriores, el efecto de todas las variables que afectan a la dependiente y no están incluidas en el modelo (5.1). Para el término de error aleatorio ε_i y las variables independientes x_i se detallan los supuestos en notación matricial como se muestra a continuación:

Supuesto 1.

$$E\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix}=\begin{bmatrix}E(\varepsilon_1)\\ E(\varepsilon_2)\\ \vdots\\ E(\varepsilon_n)\end{bmatrix}=\begin{bmatrix}0\\ 0\\ \vdots\\ 0\end{bmatrix} \qquad \textbf{(5.7)}$$

El supuesto 1 dado en la ecuación (5.7) significa que el valor esperado del vector de perturbaciones **ε,** o sea, de cada elemento es cero.

Supuesto 2.

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = E\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{bmatrix}$$

Donde:

$\boldsymbol{\varepsilon}'$: Es la transposición del vector columna **ε**, es decir, el vector fila.

Haciendo la multiplicación se obtiene:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = E\begin{bmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \cdots & \varepsilon_2\varepsilon_n \\ \cdots & \cdots & \cdots & \cdots \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \cdots & \varepsilon_n^2 \end{bmatrix}$$

Aplicando el operador del valor esperado E a cada elemento de la matriz anterior, se obtiene:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \cdots & E(\varepsilon_2\varepsilon_n) \\ \cdots & \cdots & \cdots & \cdots \\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \cdots & E(\varepsilon_n^2) \end{bmatrix} \qquad \textbf{(5.8)}$$

Debido al supuesto de homoscedasticidad y no correlación serial, la matriz de ecuaciones dada en (5.8) se reduce a:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \mathbf{I} \qquad \textbf{(5.9)}$$

Donde: **I:** Es una matriz identidad de n * n.

La matriz de ecuaciones (5.8) y su representación dada en (5.9) se llama matriz de Varianza – Covarianza de las perturbaciones ε_i; los elementos de la diagonal principal de esta matriz (que van de la esquina superior izquierda a la esquina inferior derecha), nos dan las varianzas y los elementos localizados fuera de la diagonal, las covarianzas[2].

Nótese que la matriz de Varianza – Covarianza es simétrica: los elementos localizados a la derecha de la diagonal principal son el reflejo de los de la izquierda.

Supuesto 3.

El supuesto 3 afirma que la matriz $\mathbf{x} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$ **(5.10)**

[2] Por definición la varianza de $\varepsilon_i = E[\varepsilon_i - E(\varepsilon_i)]^2$ y la covarianza entre ε_i y $\varepsilon_j = E[\varepsilon_i - E(\varepsilon_i)][\varepsilon_j - E(\varepsilon_j)]$. Pero dado el supuesto $E(\varepsilon_i) = 0$ para cada i, tenemos la matriz de varianza – covarianza dada en la ecuación (5.8).

De orden n * k es no estocástica, o sea que consiste en números fijos. Como se mencionó anteriormente, nuestro análisis de regresión, es análisis de regresión condicional, condicional a los valores fijos de las variables x_i.

Supuesto 4.

El supuesto 4 dice que la matriz $\mathbf{x} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}$ **(5.11)**

Tiene rango (columna) completo igual a k, que es el número de columnas de la matriz. Esto significa que las columnas de la matriz son linealmente independientes, es decir, que no existe una relación lineal exacta entre las variables x_i. En otras palabras no hay multicolinealidad, en notación matricial esto es:

$$\boldsymbol{\lambda}'\mathbf{x} = 0 \qquad \textbf{(5.12)}$$

Donde:

$\boldsymbol{\lambda}'$: Es un vector fila de 1 * k.

$\mathbf{x}$: Es un vector columna k * 1.

5.5 Estimación de los Coeficientes de Regresión por Mínimos Cuadrados Ordinarios (MCO).

Para encontrar el estimador de β, por MCO escribamos primero la Función de Regresión Muestral (FRM):

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki} + e_i \qquad \textbf{(5.13)}$$

La cual puede escribirse de manera resumida en notación matricial de la siguiente forma:

$$\mathbf{y} = \mathbf{x}\hat{\boldsymbol{\beta}} + \mathbf{e} \tag{5.14}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{5.15}$$

$$\begin{array}{ccccccc} \mathbf{y} & = & & \mathbf{x} & & \hat{\boldsymbol{\beta}} & + \quad \mathbf{e} \\ n * 1 & & & n * k & & k * 1 & \quad n * 1 \end{array}$$

Donde:

$\hat{\boldsymbol{\beta}}$: Es un vector columna de k elementos que son los estimadores de MCO de los coeficientes de regresión.

e: Es un vector columna n * 1 de los residuos.

De la misma forma que en los modelos de dos y tres variables, en el caso de k variables los estimadores MCO se obtienen minimizando:

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_k x_{ki})^2 \tag{5.16}$$

En notación matricial esto equivale a minimizar $\mathbf{e}'\mathbf{e}$ dado que:

$$\mathbf{e}'\mathbf{e} = \begin{bmatrix} e_1 & e_2 & \cdots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

$$\mathbf{e}'\mathbf{e} = e_1^2 + e_2^2 + \ldots + e_n^2 = \sum_{i=1}^{n} e_i^2 \tag{5.17}$$

A partir de la ecuación (5.14) se tiene que:

$$\mathbf{e = y - x\hat{\beta}} \qquad \textbf{(5.18)}$$

Por lo tanto

$$\mathbf{e'e = (y - x\hat{\beta})'(y - x\hat{\beta})}$$

Haciendo uso de las propiedades de la transposición de matrices dadas en apéndice **A**, explícitamente $\mathbf{(x\hat{\beta})' = \hat{\beta}'x'}$; y dado que $\mathbf{\hat{\beta}'x'y}$ es un escalar (un número real), igual a su transposición $\mathbf{y'x\hat{\beta}}$.

$$\mathbf{e'e = y'y - 2\hat{\beta}'x'y + \hat{\beta}'x'x\hat{\beta}} \qquad \textbf{(5.19)}$$

La ecuación (5.19) es la representación matricial de la ecuación (5.16). En la notación escalar, el método de MCO consiste en estimar β_0, $\beta_{1,\ldots}$, β_k de tal manera que $\left(\sum_{i=1}^{n} e_i^2\right)$ sea lo más pequeña posible. Esto se logra derivando la ecuación (5.16) parcialmente con respecto a $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ e igualando los resultados a cero. Este procedimiento nos resulta en k ecuaciones simultáneas para k incógnitas, las ecuaciones normales de la teoría de MCO. Como se muestra en el apéndice **5.1 a)**, estas ecuaciones son como siguen:

$$
\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}x_{1i} + \hat{\beta}_2\sum_{i=1}^{n}x_{2i} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{ki} &= \sum_{i=1}^{n}y_i \\
\hat{\beta}_0\sum_{i=1}^{n}x_{1i} + \hat{\beta}_1\sum_{i=1}^{n}x_{1i}^2 + \hat{\beta}_2\sum_{i=1}^{n}x_{1i}x_{2i} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{1i}x_{ki} &= \sum_{i=1}^{n}x_{1i}y_i \\
\hat{\beta}_0\sum_{i=1}^{n}x_{2i} + \hat{\beta}_1\sum_{i=1}^{n}x_{1i}x_{2i} + \hat{\beta}_2\sum_{i=1}^{n}x_{2i}^2 + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{2i}x_{ki} &= \sum_{i=1}^{n}x_{2i}y_i \qquad \textbf{(5.20)} \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \cdots\cdots \\
\hat{\beta}_0\sum_{i=1}^{n}x_{ki} + \hat{\beta}_1\sum_{i=1}^{n}x_{ki}x_{1i} + \hat{\beta}_2\sum_{i=1}^{n}x_{ki}x_{2i} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{ki}^2 &= \sum_{i=1}^{n}x_{ki}y_i
\end{aligned}
$$

En forma de matrices las ecuaciones dadas en (5.20) pueden representarse como:

$$
\underbrace{\begin{bmatrix}
n & \sum_{i=1}^{n}x_{1i} & \sum_{i=1}^{n}x_{2i} & \cdots & \sum_{i=1}^{n}x_{ki} \\
\sum_{i=1}^{n}x_{1i} & \sum_{i=1}^{n}x_{1i}^2 & \sum_{i=1}^{n}x_{1i}x_{2i} & \cdots & \sum_{i=1}^{n}x_{1i}x_{ki} \\
\sum_{i=1}^{n}x_{2i} & \sum_{i=1}^{n}x_{1i}x_{2i} & \sum_{i=1}^{n}x_{2i}^2 & \cdots & \sum_{i=1}^{n}x_{2i}x_{ki} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\sum_{i=1}^{n}x_{ki} & \sum_{i=1}^{n}x_{ki}x_{1i} & \sum_{i=1}^{n}x_{ki}x_{2i} & \cdots & \sum_{i=1}^{n}x_{ki}^2
\end{bmatrix}}_{(\mathbf{x}'\mathbf{x})}
\underbrace{\begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k\end{bmatrix}}_{\hat{\boldsymbol{\beta}}}
=
\underbrace{\begin{bmatrix}
1 & 1 & \cdots & 1 \\
x_{11} & x_{12} & \cdots & x_{1n} \\
x_{21} & x_{22} & \cdots & x_{2n} \\
\cdots & \cdots & \cdots & \cdots \\
x_{k1} & x_{k2} & \cdots & x_{kn}
\end{bmatrix}}_{\mathbf{x}'}
\underbrace{\begin{bmatrix}y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n\end{bmatrix}}_{\mathbf{y}}
\qquad \textbf{(5.21)}
$$

O de manera resumida como:

$$
(\mathbf{x}'\mathbf{x})\hat{\boldsymbol{\beta}} = \mathbf{x}'\mathbf{y} \qquad \textbf{(5.22)}
$$

Observe las siguientes características de la matriz $(\mathbf{x}'\mathbf{x})$:

1. Nos da las sumas brutas de cuadrados y los productos cruzados de las variables x_i, uno de los cuales es el intercepto que toma el valor de uno para cada observación. Los elementos de la diagonal principal dan las sumas brutas de los cuadrados y los demás dan las sumas brutas de los productos cruzados (por sumas brutas entendemos la suma de las unidades originales de medida).

2. Es simétrica dado que el producto cruzado entre x_{1i} y x_{2i} es el mismo que entre x_{2i} y x_{1i}.
3. Es de orden k * k, esto es, que el número de filas es igual al número de columnas.

En la ecuación (5.22) los valores conocidos son $\mathbf{X'x}$ y $\mathbf{X'y}$ (el producto cruzado, entre las variables "x" y "y") la incógnita es $\hat{\boldsymbol{\beta}}$. Usando ahora el álgebra matricial, si la inversa $\mathbf{X'x}$ existe, digamos $\mathbf{(X'x)^{-1}}$, multiplicando ambos lados de la ecuación (5.22) por esta inversa, se obtiene:

$$\mathbf{(X'x)^{-1}(X'x)\hat{\beta} = (X'x)^{-1}X'y}$$

Pero dado $\mathbf{(X'x)^{-1}(X'x) = I}$, una matriz identidad de orden k * k, se tiene:

$$\begin{aligned} \mathbf{I\hat{\beta}} &= \mathbf{(X'x)^{-1}X'y} \\ \mathbf{\hat{\beta}} &= \mathbf{(X'x)^{-1}X'y} \\ k * 1 \quad k * k \quad & (k * n)(n * 1) \end{aligned} \qquad \textbf{(5.23)}$$

La ecuación (5.23) es un resultado fundamental de la teoría de Mínimos Cuadrados Ordinarios en notación matricial, que nos muestra como el vector $\hat{\boldsymbol{\beta}}$ puede estimarse a partir de la información dada. Aunque la ecuación (5.23) se obtuvo de la ecuación (5.21), se puede obtener directamente de la ecuación (5.19) diferenciando $\mathbf{e'e}$ con respecto a $\hat{\boldsymbol{\beta}}$ como se muestra en el apéndice **5.1 b)**.

Ejemplo Ilustrativo.

Haciendo uso de los datos del ejemplo 2 del Capítulo 2 ilustramos el método matricial desarrollado hasta el momento, para el caso de dos variables se tiene:

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & x_{13} & \cdots & x_{110} \end{bmatrix} \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ 1 & x_{13} \\ \vdots & \vdots \\ 1 & x_{110} \end{bmatrix} = \begin{bmatrix} 10 & \sum_{i=1}^{10} x_i \\ \sum_{i=1}^{10} x_i & \sum_{i=1}^{10} x_i^2 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & x_{13} & \cdots & x_{110} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{10} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{10} y_i \\ \sum_{i=1}^{10} x_i y_i \end{bmatrix}$$

Empleando la información dada en la ecuación (5.6), se obtiene:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 3 & 3 & 4 & 5 & 7 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 3 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 7 \end{bmatrix} = \begin{bmatrix} 10 & 24 \\ 24 & 110 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 3 & 3 & 4 & 5 & 7 \end{bmatrix} \begin{bmatrix} 66 \\ 62 \\ 73 \\ 72 \\ 65 \\ 60 \\ 66 \\ 58 \\ 57 \\ 54 \end{bmatrix} = \begin{bmatrix} 633 \\ 1410 \end{bmatrix}$$

Usando las reglas de la inversión de matrices dadas en el apéndice **A**, se puede ver que la inversa de la matriz $\mathbf{X}'\mathbf{X}$ es:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \frac{5}{262} & \frac{-6}{131} \\ \frac{-6}{131} & \frac{5}{262} \end{bmatrix}$$

Por lo tanto

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \frac{5}{262} & \frac{-6}{131} \\ \frac{-6}{131} & \frac{5}{262} \end{bmatrix} \begin{bmatrix} 633 \\ 1410 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 68.301 \\ -2.084 \end{bmatrix}$$

En el Capítulo 2 se obtuvo $\hat{\beta}_0 = 68.302$ y $\hat{\beta}_1 = -2.084$. La diferencia entre las dos estimaciones se debe a los errores de redondeo.

5.5.1 Matriz de Varianza – Covarianza de $\hat{\boldsymbol{\beta}}$.

El método matricial nos permite desarrollar fórmulas, no sólo para la varianza de $\hat{\beta}_i$, cualquier elemento del vector $\hat{\boldsymbol{\beta}}$, si no además para las covarianzas entre los dos elementos de $\hat{\boldsymbol{\beta}}$, digamos, $\hat{\beta}_i$ y $\hat{\beta}_j$. Estas varianzas y covarianzas se necesitan para la inferencia estadística.

Por definición la matriz de varianza covarianza de $\hat{\boldsymbol{\beta}}$ es:

$$\text{var} - \text{cov}\,(\hat{\boldsymbol{\beta}}) = E\{[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})]\,[\hat{\boldsymbol{\beta}} - E(\hat{\boldsymbol{\beta}})]'\}$$

Lo cual se puede escribir explícitamente como:

$$\text{var} - \text{cov}\,(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \cdots & \text{cov}(\hat{\beta}_0, \hat{\beta}_k) \\ \text{cov}(\hat{\beta}_1, \hat{\beta}_0) & \text{var}(\hat{\beta}_1) & \cdots & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ \text{cov}(\hat{\beta}_k, \hat{\beta}_0) & \text{cov}(\hat{\beta}_k, \hat{\beta}_1) & \cdots & \text{var}(\hat{\beta}_k) \end{bmatrix} \qquad \textbf{(5.24)}$$

En el apéndice **5.1 c)** se muestra que la matriz de varianzas y covarianzas puede obtenerse de la siguiente forma:

$$\text{var} - \text{cov}\,(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{x}'\mathbf{x})^{-1} \qquad \textbf{(5.25)}$$

Donde:

σ^2: Es la varianza homoscedástica de ε_i.

$(\mathbf{x}'\mathbf{x})^{-1}$: Es la matriz inversa dada en la ecuación (5.23) que nos da el estimador $\hat{\boldsymbol{\beta}}$ de MCO.

En el modelo de regresión lineal de dos y tres variables, un estimador insesgado de σ^2 estaba dado por:

$\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-2}$ y $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-3}$, respectivamente. En el caso de k variables la fórmula correspondiente es:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k}$$

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} \qquad \textbf{(5.26)}$$

Donde hay n – k grados de libertad.

Aunque en principio $\mathbf{e}'\mathbf{e}$ puede calcularse a partir de los residuos estimados, en la práctica puede obtenerse directamente de la siguiente manera. Recordando que

$SS_{Res} = \sum_{i=1}^{n} e_i^2 = SS_T - SS_R$, en el caso de dos variables,

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum_{i=1}^{n} (x_{1i} - \bar{x})^2 \qquad \textbf{(5.27)}$$

En el caso de tres variables

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) - \hat{\beta}_2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(y_i - \bar{y}) \qquad \textbf{(5.28)}$$

Extendiendo este principio al modelo de k variables se puede ver que:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \hat{\beta}_1 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(y_i - \bar{y}) - \cdots - \hat{\beta}_k \sum_{i=1}^{n} (x_{ki} - \bar{x}_k)(y_i - \bar{y}) \qquad \textbf{(5.29)}$$

En notación matricial:

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{y'y} - n\bar{y}^2 \qquad \textbf{(5.30)}$$

$$SS_R = \hat{\beta}_1\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y}) + \cdots + \hat{\beta}_k\sum_{i=1}^{n}(x_{ki} - \bar{x}_k)(y_i - \bar{y}) = \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2 \qquad \textbf{(5.31)}$$

Donde el término $n\bar{y}^2$ se conoce como la corrección de la media. Entonces,

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} \qquad \textbf{(5.32)}$$

Una vez estimado $\mathbf{e}'\mathbf{e}$, el valor de $\hat{\sigma}^2$ puede calcularse fácilmente como en la ecuación (5.26) lo que a su vez nos permitirá estimar la matriz de varianza – covarianza como en la ecuación (5.25).

Para el caso del ejemplo ilustrativo,

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}$$

$$\mathbf{e}'\mathbf{e} = \begin{bmatrix}66 & 62 & 73 & 72 & 65 & 60 & 66 & 58 & 57 & 54\end{bmatrix}\begin{bmatrix}66\\62\\73\\72\\65\\60\\66\\58\\57\\54\end{bmatrix} - [68.301 \quad -2.084]\begin{bmatrix}633\\1410\end{bmatrix}$$

$$\mathbf{e}'\mathbf{e} = 40423 - [68.301 \quad -2.084]\begin{bmatrix}633\\1410\end{bmatrix}$$

$\mathbf{e}'\mathbf{e}$ = 40423 – 40296.093

$\mathbf{e}'\mathbf{e}$ = 126.907

En consecuencia $\hat{\sigma}^2 = \frac{126.907}{10-2} = 15.863$, valor que se aproxima al que se obtuvo en el Capítulo 2.

5.5.2 Propiedades del Vector $\hat{\beta}$ de Mínimos Cuadrados Ordinarios.

En el caso de dos y tres variables se sabe que los estimadores MCO son lineales, insesgados y entre todos los estimadores insesgados, tienen varianza mínima (Teorema de Gauss-Markov). En resumen, los estimadores MCO son los mejores estimadores lineales insesgados.

Esta propiedad es extensiva a todo el vector $\hat{\beta}$; esto es, $\hat{\beta}$ es lineal (cada uno de los elementos es una función lineal de "y"). E($\hat{\beta}$) = β, o sea, el valor esperado de cada elemento del vector $\hat{\beta}$ es igual al elemento correspondiente del verdadero β, y de todos los estimadores lineales insesgados de β, el estimador por MCO de $\hat{\beta}$ tiene varianza mínima.

Como se afirmó en la introducción, el caso de k variables es generalmente una extensión directa de los casos de dos y tres variables.

5.6 Coeficiente de Determinación R^2 en Notación Matricial.

El coeficiente de determinación R^2 se ha definido como:

$$R^2 = \frac{SS_R}{SS_T}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

En el caso de dos variables:

$$r^2 = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \hat{\beta}_1^2 \left(\frac{S_{xx}}{S_{yy}} \right) \qquad \textbf{(5.33)}$$

En el caso de tres variables:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\hat{\beta}_1 \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y}) + \hat{\beta}_2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \textbf{(5.34)}$$

Generalizando, para el caso de k variables tendremos:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2 = \frac{\hat{\beta}_1 \sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y}) + \hat{\beta}_2 \sum_{i=1}^{n}(x_{2i} - \bar{x}_2)(y_i - \bar{y}) + \cdots + \hat{\beta}_k \sum_{i=1}^{n}(x_{ki} - \bar{x}_k)(y_i - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad \textbf{(5.35)}$$

Usando las ecuaciones (5.30) y (5.31), la ecuación (5.35) puede escribirse como:

$$R^2 = \frac{\hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2}{\mathbf{y}'\mathbf{y} - n\bar{y}^2} \qquad \textbf{(5.36)}$$

Que nos muestra la representación matricial de R^2.

Para nuestro ejemplo ilustrativo.

$$\hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} = [68.301 \quad -2.084]\begin{bmatrix} 633 \\ 1410 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} = 40296.093$$

$$\mathbf{y}'\mathbf{y} = 40423$$

$$n\bar{y}^2 = 10(63.3)^2 = 40068.9$$

Reemplazando estos valores en la ecuación (5.36) se puede ver que:

$$R^2 = \frac{40296.093 - 40068.9}{40423 - 40068.9}$$

$$R^2 = 0.641$$

Que es aproximadamente igual al valor que se obtuvo en el Capítulo 2, salvo por los errores de redondeo.

5.7 Pruebas de Hipótesis con Notación Matricial.

Por las razones dadas en capítulos anteriores, si nuestro objetivo es la inferencia además de la estimación, debemos suponer que las perturbaciones ε_i siguen alguna distribución de probabilidad. En el análisis de regresión usualmente suponemos que cada

ε_i sigue la distribución normal con media $E(\varepsilon_i) = 0$ y varianza $var(\varepsilon_i) = \sigma^2$. En notación matricial, se tiene que:

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \qquad \textbf{(5.37)}$$

Donde:

$\boldsymbol{\varepsilon}$ y $\mathbf{0}$: Son vectores columna de n * 1.

I: Es una matriz identidad de n * n.

0 : Es el vector nulo.

Según el supuesto de normalidad, sabemos que en los casos de dos y tres variables:

1. Los estimadores $\hat{\beta}_i$ de MCO y los estimadores $\tilde{\beta}_i$ de MV son idénticos, pero el estimador $\tilde{\sigma}^2$ de MV es sesgado, por esta razón al calcular el estimador de σ^2 se utiliza el método de MCO.
2. Los estimadores $\hat{\beta}_i$ están también normalmente distribuidos.

Generalizando, en el caso de k variables se puede mostrar que:

$$\hat{\boldsymbol{\beta}} \sim N [\beta, \sigma^2 (\mathbf{x}'\mathbf{x})^{-1}] \qquad \textbf{(5.38)}$$

Esto es, cada elemento de $\hat{\boldsymbol{\beta}}$ está distribuido normalmente con media igual al verdadero β y la varianza dada por σ^2 multiplicada por el correspondiente elemento de la diagonal de la matriz inversa $(\mathbf{x}'\mathbf{x})^{-1}$.

Debido a que en la práctica σ^2 es desconocida, se estima por $\hat{\sigma}^2$. Luego por el cambio común a la distribución t, se sigue que cada elemento de $\hat{\boldsymbol{\beta}}$ sigue la distribución t con n-k grados de libertad.

Simbólicamente esto es:

$$t = \frac{\hat{\beta}_i - \beta_i}{es(\hat{\beta}_i)} \qquad \textbf{(5.39)}$$

Con n-k grados de libertad

Donde:

$\hat{\beta}_i$: Es cualquier elemento del vector $\hat{\boldsymbol{\beta}}$.

La distribución t puede, por lo tanto, usarse para pruebas de hipótesis acerca del verdadero valor β_i así como para establecer intervalos de confianza acerca de dicho valor.

5.7.1 Prueba de la Significancia de la Regresión.

La prueba de la significancia de la regresión es para determinar si hay una relación lineal entre la variable respuesta "y" y cualquiera de las variables explicativas $x_1, x_2, ..., x_k$. Este procedimiento suele considerarse como una prueba general o global de la adecuación del modelo.

Las hipótesis correspondientes son:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1 : \beta_j \neq 0, \quad \text{al menos para un j.}$$

El rechazo de la hipótesis nula implica que al menos una de las variables explicatorias $x_1, x_2, ..., x_k$ contribuye en el modelo de forma significativa.

El procedimiento de prueba es una generalización del análisis de varianza que se usó en la regresión lineal simple dada en el Capítulo 2.

5.7.2 Análisis de Varianza en Notación Matricial.

La técnica de análisis de varianza se utiliza:

1. Para probar la significancia de la regresión estimada, es decir, para probar la hipótesis nula según la cual los verdaderos coeficientes parciales (pendientes) son simultáneamente iguales a cero.
2. Para estimar la contribución incremental de una variable explicatoria.

La técnica del análisis de varianza se puede hacer extensiva al caso de k variables. Recuerde que la técnica de análisis de varianza consiste en descomponer la suma total de cuadrados (SS_T) en dos componentes: la suma de cuadrados de regresión (SS_R), y la suma de cuadrados residuales (SS_{Res}). Así:

$$SS_T = SS_R + SS_{Res}$$

Las expresiones matriciales para estas tres sumas ya se mostraron en las ecuaciones (5.30), (5.31) y (5.32), respectivamente. Los grados de libertad asociados con estas sumas de cuadrados son n-1, k-1 y n-k, respectivamente.

De acuerdo con la definición del estadístico F se tiene que:

$$F_0 = \frac{\dfrac{\hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2}{k-1}}{\dfrac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}}{n-k}} \qquad \textbf{(5.40)}$$

Tiene distribución F con k-1 y n-k grados de libertad.

En esta forma y de acuerdo con el Capítulo 4, tabla 4.4, podemos construir la tabla 5.1.

Tabla 5.1 Formulación matricial del cuadro de análisis de varianza para el modelo de regresión lineal de k variables.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F_0
Regresión	$\hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2$	k-1	$(\hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2)/(k-1)$	$\dfrac{\hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2}{k-1}$
Residual	$\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}$	n-k	$(\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y})/(n-k)$	$\dfrac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}}{n-k}$
Total	$\mathbf{y}'\mathbf{y} - n\bar{y}^2$	n-1		

En el Capítulo 4 se vio que bajo los supuestos formulados, existe una relación muy cercana entre F y R^2; explícitamente:

$$F_0 = \frac{\dfrac{R^2(\mathbf{y}'\mathbf{y} - n\bar{y}^2)}{k-1}}{\dfrac{(1-R^2)(\mathbf{y}'\mathbf{y} - n\bar{y}^2)}{n-k}} \qquad \textbf{(5.41)}$$

Por lo tanto, la tabla 5.1 de análisis de varianza se muestra en una forma alterna en la tabla 5.2. Una ventaja de la tabla 5.2 respecto a la 5.1 es que todo el análisis puede hacerse en términos del R^2; no es necesario tener en cuenta $(\mathbf{y}'\mathbf{y} - n\bar{y}^2)$ en razón de que este desaparece en la relación F.

Tabla 5.2 Análisis de varianza para k variables forma matricial en términos de R^2.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F_0
Regresión	$R^2(\mathbf{y}'\mathbf{y} - n\bar{y}^2)$	k-1	$R^2(\mathbf{y}'\mathbf{y} - n\bar{y}^2)/k-1$	$\dfrac{R^2(\mathbf{y}'\mathbf{y} - n\bar{y}^2)}{k-1}$
Residual	$(1-R^2)(\mathbf{y}'\mathbf{y} - n\bar{y}^2)$	n-k	$(1-R^2)(\mathbf{y}'\mathbf{y} - n\bar{y}^2)/n-k$	$\dfrac{(1-R^2)(\mathbf{y}'\mathbf{y} - n\bar{y}^2)}{n-k}$
Total	$\mathbf{y}'\mathbf{y} - n\bar{y}^2$	n-1		

5.7.3 Intervalos de Confianza en Regresión Múltiple.

Los intervalos de confianza de los coeficientes de regresión individual y los intervalos de confianza para la predicción media, para niveles específicos de las variables explicativas, juegan un papel importante igual que en la regresión lineal simple. En esta sección se desarrollan los intervalos de confianza, uno por uno, para estos casos. También se presentará en forma breve los intervalos simultáneos de confianza para los coeficientes de regresión.

5.7.3.1 Intervalos de Confianza de los Coeficientes de Regresión.

Para construir estimados de intervalo de confianza de los coeficientes de regresión β_j, se continuará suponiendo que los errores ε_i están distribuidos normal e independientemente, con media cero y varianza σ^2. En consecuencia, las observaciones y_i están distribuidas en forma normal e independiente, con media $\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$, y varianza σ^2. Como el estimador $\hat{\boldsymbol{\beta}}$ obtenido por Mínimos Cuadrados es una combinación lineal de las observaciones, también está distribuida normalmente, con vector medio $\boldsymbol{\beta}$ y matriz de varianza-covarianza $\sigma^2 (\mathbf{x}'\mathbf{x})$. Esto implica que la distribución marginal de cualquier coeficiente de regresión $\hat{\beta}_j$ es normal, con media β_j y varianza $\sigma^2 C_{jj}$, donde C_{jj} es el j- ésimo elemento de la diagonal de la matriz $(\mathbf{x}'\mathbf{x})$. Debido a que en la práctica σ^2 es desconocido, se estima por $\hat{\sigma}^2$. Luego, por el cambio

común a la distribución t, se sigue que cada elemento de $\hat{\boldsymbol{\beta}}$ sigue la distribución t con n – k grados de libertad. Simbólicamente es:

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}},\ j = 0, 1,\ldots, k \qquad \textbf{(5.42)}$$

De acuerdo con el resultado de la ecuación (5.42), se puede definir un intervalo de confianza de 100(1 - α) por ciento para el coeficiente de regresión β_j, j = 0, 1,…, k, como sigue:

$$\hat{\beta}_j - t_{(\alpha/2,\ n-k)}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{(\alpha/2,\ n-k)}\sqrt{\hat{\sigma}^2 C_{jj}} \qquad \textbf{(5.43)}$$

Recuérdese que la cantidad:

$$es(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}} \qquad \textbf{(5.44)}$$

Es el error estándar del coeficiente de regresión $\hat{\beta}_j$.

5.7.3.2 Estimación del Intervalo de Confianza de la Predicción Media.

Se puede establecer un intervalo de confianza para la predicción media en determinado punto, como x_{01}, x_{02},…, x_{0k}. Defínase el vector $\mathbf{x}_0$ como sigue:

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

El valor ajustado en este punto es:

$$\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} \qquad \textbf{(5.45)}$$

Es un estimador insesgado de E(y|$\mathbf{x}_0$), porque la $E(\hat{y}_0) = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$ = E(y|$\mathbf{x}_0$), y la varianza de $\hat{y}_0$ es:

$$\text{var}(\hat{y}_0) = \sigma^2\, \mathbf{x}_0'(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_0 \qquad \textbf{(5.46)}$$

Por consiguiente, un intervalo de confianza de 100(1 - α) por ciento de la predicción media en el punto $x_{01}, x_{02},\ldots, x_{0k}$ es:

$$\hat{y}_0 - t_{(\alpha/2,n-k)}\sqrt{\sigma^2\, \mathbf{x}_0'(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_0} \le E(y|\mathbf{x}_0) \le \hat{y}_0 + t_{(\alpha/2,n-k)}\sqrt{\sigma^2\, \mathbf{x}_0'(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_0} \qquad \textbf{(5.47)}$$

Es la generalización del caso de regresión simple.

5.7.3.3 Intervalo de Confianza para la Predicción Individual.

Con el modelo de regresión se pueden predecir observaciones futuras de "y" que correspondan a determinados valores de las variables explicativas, por ejemplo $x_{01}, x_{02},\ldots, x_{0k}$. Si $\mathbf{x}_0'$ = [1, x_{01}, $x_{02},\ldots, x_{0k}$], entonces un estimador puntual de la observación futura $\mathbf{y}_0$ en el punto $x_{01}, x_{02},\ldots, x_{0k}$ es:

$$\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} \qquad \textbf{(5.48)}$$

Un intervalo de predicción de 100(1 - α) por ciento para esta futura observación es:

$$\hat{y}_0 - t_{(\alpha/2,\ n\text{-}k)}\sqrt{\sigma^2(1+\mathbf{x}_0'(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_0\ )} \le y_0 \le \hat{y}_0 + t_{(\alpha/2,\ n\text{-}k)}\sqrt{\sigma^2(1+\mathbf{x}_0'(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_0\ )}$$

Es una generalización del intervalo de predicción para una futura observación en la regresión lineal simple.

5.8 Matriz de Correlación.

En los Capítulos anteriores, vimos los coeficientes de correlación simple o de orden cero r_{12}, r_{13}, r_{23} y las correlaciones parciales o de primer orden $r_{12.3}, r_{13.2}, r_{23.1}$ y sus interrelaciones. En el caso de k variables tendremos k(k - 1)/2 coeficientes de correlación de orden cero. Estas k(k-1)/2 correlaciones pueden escribirse en una matriz llamada matriz de correlación **R**, de la forma siguiente:

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22} & \cdots & r_{2k} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ r_{k1} & a_{k2} & \cdots & r_{kk} \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots\cdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix} \qquad \textbf{(5.49)}$$

Donde el subíndice 1, denota la variable dependiente (r_{12} significa coeficiente de correlación entre "y" y x_2) y donde el coeficiente de correlación de una variable con respecto a ella misma es siempre 1 ($r_{11} = r_{22} = \ldots = r_{kk} = 1$).

A partir de la matriz de correlación **R**, podemos obtener los coeficientes de correlación de primer orden y de órdenes más altos.

Ejemplo 1: Para resumir el uso de matrices del análisis de regresión, se presenta este ejemplo numérico de tres variables.

De los datos de la población de 40 estudiantes de Estadística Aplicada a la Educación II del ciclo I 2008 de la UES-FMO, tomamos una muestra de 10 estudiantes, estamos interesados en estudiar, si existe relación entre el peso de un estudiante, la estatura y los años de edad que este tenga. En donde la variable dependiente es Peso en kilogramos (y), las variables independientes son Estatura en centímetros (x_{1i}) y Años (x_{2i}), los datos se muestran en la tabla siguiente:

y(kg.)	54.5	50	49.5	52	54	50	63	48	49	54
x_{1i}	163	150	149	155	165	150	170	140	145	165
x_{2i}	21	23	24	23	19	24	18	19	30	25

La ecuación de regresión es: $\mathbf{y = x\hat{\beta} + e}$

En notación matricial, este problema puede escribirse como:

$$\begin{bmatrix} 54.5 \\ 50 \\ 49.5 \\ 52 \\ 54 \\ 50 \\ 63 \\ 48 \\ 49 \\ 54 \end{bmatrix} = \begin{bmatrix} 1 & 163 & 21 \\ 1 & 150 & 23 \\ 1 & 149 & 24 \\ 1 & 155 & 23 \\ 1 & 165 & 19 \\ 1 & 150 & 24 \\ 1 & 170 & 18 \\ 1 & 140 & 19 \\ 1 & 145 & 30 \\ 1 & 165 & 25 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \\ e_8 \\ e_9 \\ e_{10} \end{bmatrix}$$

$$\underset{10*1}{\mathbf{y}} \quad = \quad \underset{10*3}{\mathbf{x}} \quad \underset{3*1}{\hat{\boldsymbol{\beta}}} \quad + \quad \underset{10*1}{\mathbf{e}}$$

Con la información anterior se obtienen los valores siguientes:

$$\bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = \frac{524}{10} = 52.4, \quad \bar{x}_1 = \frac{\sum_{i=1}^{10} x_{1i}}{10} = \frac{1552}{10} = 155.20, \quad \bar{x}_2 = \frac{\sum_{i=1}^{10} x_{2i}}{10} = \frac{226}{10} = 22.6$$

$$\sum_{i=1}^{10} x_{1i} = 1552, \quad \sum_{i=1}^{10} x_{2i} = 226, \quad \sum_{i=1}^{10} x_{1i}^2 = 241770, \quad \sum_{i=1}^{10} x_{2i}^2 = 5222, \quad \sum_{i=1}^{10} x_{1i}x_{2i} = 34944$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & x_{16} & x_{17} & x_{18} & x_{19} & x_{110} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & x_{26} & x_{27} & x_{28} & x_{29} & x_{210} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ 1 & x_{13} & x_{23} \\ 1 & x_{14} & x_{24} \\ 1 & x_{15} & x_{25} \\ 1 & x_{16} & x_{26} \\ 1 & x_{17} & x_{27} \\ 1 & x_{18} & x_{28} \\ 1 & x_{19} & x_{29} \\ 1 & x_{110} & x_{210} \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 163 & 150 & 149 & 155 & 165 & 150 & 170 & 140 & 145 & 165 \\ 21 & 23 & 24 & 23 & 19 & 24 & 18 & 19 & 30 & 25 \end{bmatrix} \begin{bmatrix} 1 & 163 & 21 \\ 1 & 150 & 23 \\ 1 & 149 & 24 \\ 1 & 155 & 23 \\ 1 & 165 & 19 \\ 1 & 150 & 24 \\ 1 & 170 & 18 \\ 1 & 140 & 19 \\ 1 & 145 & 30 \\ 1 & 165 & 25 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^{10} x_{1i} & \sum_{i=1}^{10} x_{2i} \\ \sum_{i=1}^{10} x_{1i} & \sum_{i=1}^{10} x_{1i}^2 & \sum_{i=1}^{10} x_{1i}x_{2i} \\ \sum_{i=1}^{10} x_{2i} & \sum_{i=1}^{10} x_{1i}x_{2i} & \sum_{i=1}^{10} x_{2i}^2 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 10 & 1552 & 226 \\ 1552 & 241770 & 34944 \\ 226 & 34944 & 5222 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ x_{11} & x_{12} & x_{13} & x_{14} & x_{15} & x_{16} & x_{17} & x_{18} & x_{19} & x_{110} \\ x_{21} & x_{22} & x_{23} & x_{24} & x_{25} & x_{26} & x_{27} & x_{28} & x_{29} & x_{210} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{10} y_i \\ \sum_{i=1}^{10} x_{1i}y_i \\ \sum_{i=1}^{10} x_{2i}y_i \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_{i=1}^{10} y_i \\ \sum_{i=1}^{10} x_{1i}y_i \\ \sum_{i=1}^{10} x_{2i}y_i \end{bmatrix} = \begin{bmatrix} 524 \\ 81674 \\ 11770.5 \end{bmatrix}$$

Para encontrar el valor de los coeficientes de regresión, necesitamos calcular la inversa de la matriz $\mathbf{X}'\mathbf{X}$, para ello hacemos uso de las reglas de inversión de matrices dadas en el apéndice **A**.

Calculamos el determinante de la matriz $\mathbf{x'x}$ como se muestra:

$$|\mathbf{x'x}| = \begin{vmatrix} 10 & 1552 & 226 \\ 1552 & 241770 & 34944 \\ 226 & 34944 & 5222 \end{vmatrix}$$

$$|\mathbf{x'x}| = 10\begin{vmatrix} 241770 & 34944 \\ 34944 & 5222 \end{vmatrix} - 1552\begin{vmatrix} 1552 & 226 \\ 34944 & 5222 \end{vmatrix} + 226\begin{vmatrix} 1552 & 226 \\ 241770 & 34944 \end{vmatrix}$$

$$|\mathbf{x'x}| = 414398040 - 321574400 - 91966632$$

$$|\mathbf{x'x}| = 857008$$

Obtenemos ahora la matriz de cofactores, o sea **C.**

$$\mathbf{C} = \begin{bmatrix} \begin{vmatrix} 241770 & 34944 \\ 34944 & 5222 \end{vmatrix} & - \begin{vmatrix} 1552 & 226 \\ 34944 & 5222 \end{vmatrix} & \begin{vmatrix} 1552 & 226 \\ 241770 & 34944 \end{vmatrix} \\ - \begin{vmatrix} 1552 & 34944 \\ 226 & 5222 \end{vmatrix} & \begin{vmatrix} 10 & 226 \\ 226 & 5222 \end{vmatrix} & - \begin{vmatrix} 10 & 226 \\ 1552 & 34944 \end{vmatrix} \\ \begin{vmatrix} 1552 & 241770 \\ 226 & 34944 \end{vmatrix} & - \begin{vmatrix} 10 & 1552 \\ 226 & 34944 \end{vmatrix} & \begin{vmatrix} 10 & 1552 \\ 1552 & 241770 \end{vmatrix} \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 41439804 & -207200 & -406932 \\ -207200 & 1144 & 1312 \\ -406932 & 1312 & 8996 \end{bmatrix}$$

Transponiendo la matriz de cofactores anterior se obtiene la matriz adjunta:

$$(\text{adj}\,\mathbf{x'x}) = \begin{bmatrix} 41439804 & -207200 & -406932 \\ -207200 & 1144 & 1312 \\ -406932 & 1312 & 8996 \end{bmatrix}$$

La matriz es la misma, dado que los elementos por encima de la diagonal son iguales a los que están debajo de la diagonal.

Dividimos los elementos de la (adj $\mathbf{x}'\mathbf{x}$) por el valor del determinante $|\mathbf{x}'\mathbf{x}|$ = 857008 y obtenemos:

$$(\mathbf{x}'\mathbf{x})^{-1} = \frac{1}{|\mathbf{x}'\mathbf{x}|}\ (\text{adj}\,\mathbf{x}'\mathbf{x}) = \begin{bmatrix} \frac{41439804}{857008} & -\frac{207200}{857008} & -\frac{406932}{857008} \\ -\frac{207200}{857008} & \frac{1144}{857008} & \frac{1312}{857008} \\ -\frac{406932}{857008} & \frac{1312}{857008} & \frac{8996}{857008} \end{bmatrix}$$

Ahora obtenemos los valores de los coeficientes de la forma siguiente:

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$$

$$\hat{\beta} = \begin{bmatrix} \frac{41439804}{857008} & -\frac{207200}{857008} & -\frac{406932}{857008} \\ -\frac{207200}{857008} & \frac{1144}{857008} & \frac{1312}{857008} \\ -\frac{406932}{857008} & \frac{1312}{857008} & \frac{8996}{857008} \end{bmatrix} \begin{bmatrix} 524 \\ 81674 \\ 11770\ .5 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 2.11362092 \\ 0.356066688 \\ -0.220284472 \end{bmatrix}$$

La suma de los errores al cuadrado puede calcularse como:

$$\sum_{i=1}^{10} e_i^2 = \mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}$$

$$\mathbf{y}'\mathbf{y} = \sum_{i=1}^{10} y_i^2 = 27630.5$$

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}$$

$$\mathbf{e}'\mathbf{e} = 27630.5 - \begin{bmatrix} 2.11362092 & 0.356066688 & -0.220284472 \end{bmatrix} \begin{bmatrix} 524 \\ 81674 \\ 11770.5 \end{bmatrix}$$

$$\mathbf{e'e} = 27630.5 - 27596.06966$$

$$\mathbf{e'e} = 34.430$$

De donde obtenemos:

$$\hat{\sigma}^2 = \frac{\mathbf{e'e}}{n-k} = \frac{\mathbf{y'y} - \hat{\boldsymbol{\beta}}'\mathbf{x'y}}{n-k}$$

$$\hat{\sigma}^2 = \frac{34.430}{10-3} = \frac{34.430}{7} = 4.92$$

La matriz de varianza-covarianza para $\hat{\boldsymbol{\beta}}$ puede escribirse como:

$$\text{var} - \text{cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{x'x})^{-1} = \begin{bmatrix} \text{var}(\hat{\beta}_0) & \text{cov}(\hat{\beta}_0, \hat{\beta}_1) & \cdots & \text{cov}(\hat{\beta}_0, \hat{\beta}_k) \\ \text{cov}(\hat{\beta}_1, \hat{\beta}_0) & \text{var}(\hat{\beta}_1) & \cdots & \text{cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \cdots & \cdots & \cdots & \cdots \\ \text{cov}(\hat{\beta}_k, \hat{\beta}_0) & \text{cov}(\hat{\beta}_k, \hat{\beta}_1) & \cdots & \text{var}(\hat{\beta}_k) \end{bmatrix}$$

$$\text{var} - \text{cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{x'x})^{-1} = 4.92 \begin{bmatrix} \frac{41439804}{857008} & -\frac{207200}{857008} & -\frac{406932}{857008} \\ -\frac{207200}{857008} & \frac{1144}{857008} & \frac{1312}{857008} \\ -\frac{406932}{857008} & \frac{1312}{857008} & \frac{8996}{857008} \end{bmatrix}$$

$$\text{var} - \text{cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{x'x})^{-1} = \begin{bmatrix} 237.9019 & -1.1895 & -2.3362 \\ -1.1895 & 0.0066 & 0.0075 \\ -2.3362 & 0.0075 & 0.0516 \end{bmatrix}$$

Los elementos de la diagonal de esta matriz nos dan las varianzas de $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$, respectivamente, y sus raíces cuadradas positivas nos dan los correspondientes errores estándar.

Con la información anterior encontramos ahora el valor de R^2.

$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{x'y} - n\bar{y}^2 = 27596.06966 - 10\,(52.4)^2 = 138.46966$

$SS_T = \mathbf{y'y} - n\bar{y}^2 = 27630.5 - 10\,(52.4)^2 = 172.9$

$$R^2 = \frac{\hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2}{\mathbf{y}'\mathbf{y} - n\bar{y}^2}$$

$$R^2 = \frac{138.46966}{172.9} = 0.8009 \approx 0.801$$

Con la información obtenida hasta el momento escribimos la ecuación de regresión estimada así:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$
$$\hat{y}_i = 2.11362092 + 0.356066688 x_{1i} - 0.220284472 x_{2i}$$

La interpretación de la ecuación anterior es: si ambos x_1 y x_2 están fijos en cero, el valor promedio de la variable dependiente Peso se estima en $\hat{\beta}_0 = 2.11362092$ kg., el coeficiente de regresión parcial $\hat{\beta}_1 = 0.356066688$, significa que manteniendo todas las demás variables constantes, un aumento en el Peso de, por ejemplo 1 kg. va acompañado de un aumento en la Estatura de los estudiantes alrededor de 0.35cm., de forma similar se puede interpretar $\hat{\beta}_2 = -0.220284472$, manteniendo todas las demás variables constantes el Peso promedio disminuye.

El valor de $R^2 = 0.801$ muestra que las dos variables independientes explican el 80.1% de la variación en el Peso de los estudiantes.

Prueba de hipótesis para los coeficientes individuales de regresión.

Con los datos obtenidos anteriormente realizamos la prueba de hipótesis individual para $\hat{\beta}_1$ es decir, $H_0 : \beta_1 = 0$ y $H_1 : \beta_1 \neq 0$.

Solución:

1. $H_0 : \beta_1 = 0$
2. $H_1 : \beta_1 \neq 0$
3. Se selecciona un nivel de significancia de $\alpha = 0.05$ y como la prueba es de dos colas $\alpha/2 = 0.05/2 = 0.025$ y se tiene que el valor de la tabla de t es:

   $t_{(0.05/2,\ 10-3)} = t_{(0.025,\ 7)} = 2.365$
4. Región critica: si $t < -2.365$ ó $t > 2.365$, entonces rechazamos H_0.
5. Cálculos:

$$t_0 = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 C_{11}}} = \frac{0.356066688 - 0}{\sqrt{0.0066}} = 4.383$$

6. Decisión Estadística: se rechaza H_0 porque el valor calculado $t_0 = 4.383$ es mayor de la tabla 2.365.
7. Conclusión: se concluye que hay una relación lineal entre el Peso y la Estatura.

De igual forma se realiza la prueba de hipótesis parcial para los demás coeficientes de regresión.

Como se mencionó en el Capítulo 4, no es posible aplicar la prueba t para verificar la hipótesis global según la cual $H_0 : \beta_1 = \beta_2 = 0$.

Sin embargo, recuérdese que una hipótesis nula $H_0 : \beta_1 = \beta_2 = 0$, puede ser verificada mediante la técnica de análisis de varianza y la prueba F dadas anteriormente.

Se probará la significancia global de la regresión para los datos Peso, Estatura y Edad de la muestra de 10 estudiantes, es decir, $H_0 : \beta_1 = \beta_2 = 0$ y $H_1 : \beta_j \neq 0$, al menos para un j.

Datos:

El modelo ajustado es: $\hat{y}_i = 2.11362092 + 0.356066688\ x_{1i} - 0.220284472\ x_{2i}$

$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2 = 27596.06966 - 10\ (52.4)^2 = 138.46966$

$SS_{Res} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} = 27630.5 - 27596.06966 = 34.430$

$SS_T = \mathbf{y}'\mathbf{y} - n\bar{y}^2 = 27630.5 - 10\ (52.4)^2 = 172.9$

Solución:

1. $H_0 : \beta_1 = \beta_2 = 0$

2. $H_1 : \beta_j \neq 0$, al menos para un j.

3. Se selecciona un nivel de significancia de $\alpha = 0.05$ y se tiene que el valor de la tabla F es $F_{(0.05,\ 2,\ 7)} = 4.74$

4. Cálculos:

$$F_0 = \frac{\dfrac{\hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2}{k - 1}}{\dfrac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}}{n - k}}$$

$$F_0 = \frac{\dfrac{138.46966}{3-1}}{\dfrac{34.430}{10-3}} = 14.076$$

Tabla 5.3 Análisis de varianza para las variables del ejemplo 1.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F_0
Regresión	138.46966	2	69.23483	14.076
Residual	34.430	7	4.91857	
Total	172.9	9		

5. Decisión Estadística: se rechaza H_0, porque el valor calculado para F_0 (14.076) es mayor que el de la tabla (4.74).
6. Conclusión: Se concluye que el Peso se relaciona con la Estatura y con la Edad para la muestra de 10 estudiantes.

Como se pudo observar la notación matricial proporciona un método resumido para tratar los modelos de regresión lineal que contienen cualquier número de variables.

Al igual que en los Capítulos anteriores se puede utilizar el Software estadístico SPSS para realizar la regresión lineal con cualquier número de variables.

Ejercicios 5

1. Los datos de la siguiente tabla corresponden a un estudio sobre la contaminación acústica realizado en distintas zonas de la misma ciudad. La variable "y" mide la contaminación acústica en decibelios, la variable x_1 la hora del día y x_2 el tráfico de vehículos por minuto.

Decibelios	0.9	1.6	4.7	2.8	5.6	2.4	1.0	1.5
Hora	14	15	16	13	17	18	19	20
Vehículos (min.)	1	2	5	2	6	4	3	4

Haciendo uso del algebra matricial:

e) Determinar la ecuación de regresión múltiple.

f) Calcular el coeficiente de determinación e interpretarlo.

g) Realizar la prueba de hipótesis individual y global de los coeficientes de regresión.

h) Realizar la estimación por intervalo para un $\alpha = 0.05$.

2. Para el ejercicio 7 del Capítulo 4 realizar los siguientes cálculos, haciendo uso del algebra matricial.

a) Determinar la ecuación de regresión múltiple.

b) Calcular el coeficiente de determinación e interpretarlo.

c) Realizar la prueba de hipótesis individual y global de los coeficientes de regresión.

d) Realizar la estimación por intervalo para un $\alpha = 0.05$.

3. Se quiere probar si la cobertura de la canopia (parte verde de un árbol) "y" en m^2, es una función del diámetro de los árboles por encima de 1mt. x_1; altura de la primera rama principal x_2; distancia al árbol más cercano x_3.

y (m^2)	x_1(cm.)	x_2(m)	x_3(m)
630	1112	5	22
960	810	6	19
930	1996	6	28
150	420	5	14
740	1580	3	20
180	515	3	20
690	1404	8	13
880	1720	4	26
320	620	9	18
440	880	4	22

a) Determinar la ecuación de regresión múltiple.

b) Realizar la prueba de hipótesis para los parámetros individuales y globales.

c) Determinar intervalos de confianza del 95% para los parámetros.

d) Definir el vector $\mathbf{x}_0$ con $x_{01} = 800$, $x_{02} = 7$ y $x_{03} = 17$ y realizar la predicción media.

4. Se tomaron medidas de 9 regiones geográficas sobre nivel de urbanización relativa x_1, nivel educativo x_2 e ingreso relativo x_3, para determinar su influencia sobre la demanda de un producto "y". Los datos se muestran a continuación:

Nivel de urbanización	42.2	48.6	42.6	39.0	34.7	44.5	39.1	40.1	45.9
Nivel educativo	11.2	10.6	10.6	10.4	9.3	10.8	10.7	10.0	12.0
Ingreso relativo	31.9	13.2	28.7	26.1	30.1	8.5	24.3	18.6	20.4
Consumo	167.1	174.4	160.8	162.0	140.8	174.6	163.7	174.5	185.7

a) Determinar la ecuación de regresión múltiple.

b) Calcular el valor de R^2.

c) Realizar prueba de hipótesis para los parámetros individuales y globales.

d) Determinar intervalos de confianza del 99% para los parámetros.

5. Se quiere ajustar un modelo de regresión lineal múltiple, que relacione los precios en miles de dólares de viviendas (y) con impuestos (x_1), cantidad de baños (x_2), tamaño del terreno en pies cuadrados (x_3), superficie construida (x_4), cantidad de cajones en cochera (x_5), cantidad de habitaciones (x_6), cantidad de recamaras (x_7), edad de la casa en años (x_8) y cantidad de chimeneas (x_9).

y	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
25.9	4.9176	1.0	3.4720	0.9980	1.0	7	4	42	0
29.5	5.0208	1.0	3.5310	1.5000	2.0	7	4	62	0
27.9	4.5429	1.0	2.2750	1.1750	1.0	6	3	40	0
25.9	4.5573	1.0	4.0500	1.2320	1.0	6	3	54	0
29.9	5.0597	1.0	4.4550	1.1210	1.0	6	3	42	0
29.9	3.8910	1.0	4.4550	0.9880	1.0	6	3	56	0
30.9	5.8980	1.0	5.8500	1.2400	1.0	7	3	51	1
28.9	5.6039	1.0	9.5200	1.5010	0.0	6	3	32	0
35.9	5.8282	1.0	6.4350	1.2250	2.0	6	3	32	0
31.5	5.3003	1.0	4.9883	1.5520	1.0	6	3	30	0
31.0	6.2712	1.0	5.5200	0.9750	1.0	5	2	30	0
30.9	5.9592	1.0	6.6660	1.1210	2.0	6	4	32	0
30.0	5.0500	1.0	5.0000	1.0200	0.0	5	3	46	1
36.9	8.2464	1.5	5.1500	1.6640	2.0	8	3	50	0
41.9	6.6969	1.5	6.9020	1.4880	1.5	7	3	22	1
40.5	7.7841	1.5	7.1020	1.3760	1.0	6	3	17	0
43.9	9.0384	1.0	7.8000	1.5000	1.5	7	3	23	0
37.5	5.9894	1.0	5.5200	1.2560	2.0	6	3	40	1
37.9	7.5452	1.5	5.0000	1.6900	1.0	6	3	22	0
44.5	8.7951	1.5	9.8900	1.8200	2.0	8	4	50	1
37.9	6.0831	1.5	6.7265	1.6520	1.0	6	3	44	0
38.9	8.3607	1.5	9.1500	1.7770	2.0	8	4	48	1
36.9	8.1400	1.0	8.0000	1.5040	2.0	7	3	3	0
45.8	9.1416	1.5	7.3262	1.8310	1.5	8	4	31	0

a) Determinar la ecuación de regresión múltiple.

b) Calcular el valor de R^2.

c) Realizar el análisis de los residuos.

d) Realizar prueba de hipótesis para los parámetros individuales y globales.

e) Determinar intervalos de confianza del 95% para los parámetros.

f) Concluir de acuerdo a los resultados obtenidos en los literales anteriores.

6. Para los datos del ejemplo 1 desarrollado en este Capítulo:

a) Determinar intervalos de confianza del 95% para los parámetros.

b) Determinar intervalos de confianza del 99% para los parámetros.

c) Realizar el análisis de los residuos.

d) Interpretar los resultados obtenidos en a), b) y c).

Apéndice 5: Deducción de Ecuaciones.

5.1 Deducción de ecuaciones utilizadas en el Capítulo 5.

a) Deducción de ecuación (5.20).

Partiendo de:

$$SS_{Res} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_k x_{ki})^2$$

Derivando parcialmente con respecto a $\hat{\beta}_0$ obtenemos:

$$\frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_k x_{ki})^2\right) = \frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=1}^{n} e_i^2\right)$$

$$2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_k x_{ki})(-1) = 0$$

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_k x_{ki}) = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} - \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} - \ldots - \hat{\beta}_k \sum_{i=1}^{n} x_{ki} = 0$$

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} x_{2i} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ki} = \sum_{i=1}^{n} y_i$$

Con respecto a $\hat{\beta}_1$.

$$\frac{\partial}{\partial \hat{\beta}_1}\left(\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_k x_{ki})^2\right) = \frac{\partial}{\partial \hat{\beta}_1}\left(\sum_{i=1}^{n} e_i^2\right)$$

$$2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \ldots - \hat{\beta}_k x_{ki})(-x_{1i}) = 0$$

$$\sum_{i=1}^{n} x_{1i} y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_{1i} - \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 - \hat{\beta}_2 \sum_{i=1}^{n} x_{1i} x_{2i} - \ldots - \hat{\beta}_k \sum_{i=1}^{n} x_{1i} x_{ki} = 0$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} x_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{1i} x_{2i} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{1i} x_{ki} = \sum_{i=1}^{n} x_{1i} y_i$$

Con respecto a $\hat{\beta}_2$.

$$\frac{\partial}{\partial\hat{\beta}_2}\left(\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{1i}-\hat{\beta}_2x_{2i}-\ldots-\hat{\beta}_kx_{ki})^2\right)=\frac{\partial}{\partial\hat{\beta}_2}\left(\sum_{i=1}^{n}e_i^2\right)$$

$$2\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_{1i}-\hat{\beta}_2x_{2i}-\ldots-\hat{\beta}_kx_{ki})(-x_{2i})=0$$

$$\sum_{i=1}^{n}x_{2i}y_i-\hat{\beta}_0\sum_{i=1}^{n}x_{2i}-\hat{\beta}_1\sum_{i=1}^{n}x_{1i}x_{2i}-\hat{\beta}_2\sum_{i=1}^{n}x_{2i}^2-\ldots-\hat{\beta}_k\sum_{i=1}^{n}x_{2i}x_{ki}=0$$

$$\hat{\beta}_0\sum_{i=1}^{n}x_{2i}+\hat{\beta}_1\sum_{i=1}^{n}x_{1i}x_{2i}+\hat{\beta}_2\sum_{i=1}^{n}x_{2i}^2+\ldots+\hat{\beta}_k\sum_{i=1}^{n}x_{2i}x_{ki}=\sum_{i=1}^{n}x_{2i}y_i$$

Y así sucesivamente, obtenemos así la ecuación (5.20):

$$\begin{aligned}
n\hat{\beta}_0 &+ \hat{\beta}_1\sum_{i=1}^{n}x_{1i} + \hat{\beta}_2\sum_{i=1}^{n}x_{2i} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{ki} = \sum_{i=1}^{n}y_i\\
\hat{\beta}_0\sum_{i=1}^{n}x_{1i} &+ \hat{\beta}_1\sum_{i=1}^{n}x_{1i}^2 + \hat{\beta}_2\sum_{i=1}^{n}x_{1i}x_{2i} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{1i}x_{ki} = \sum_{i=1}^{n}x_{1i}y_i\\
\hat{\beta}_0\sum_{i=1}^{n}x_{2i} &+ \hat{\beta}_1\sum_{i=1}^{n}x_{1i}x_{2i} + \hat{\beta}_2\sum_{i=1}^{n}x_{2i}^2 + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{2i}x_{ki} = \sum_{i=1}^{n}x_{2i}y_i\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
\hat{\beta}_0\sum_{i=1}^{n}x_{ki} &+ \hat{\beta}_1\sum_{i=1}^{n}x_{ki}x_{1i} + \hat{\beta}_1\sum_{i=1}^{n}x_{ki}x_{2i} + \cdots + \hat{\beta}_k\sum_{i=1}^{n}x_{ki}^2 = \sum_{i=1}^{n}x_{ki}y_i
\end{aligned}$$

L.q.q.d

b) Deducción de ecuación (5.23).

Se sabe que:

$$\mathbf{y}=\mathbf{x}\hat{\boldsymbol{\beta}}+\mathbf{e}$$

$$\mathbf{e}=\mathbf{y}-\mathbf{x}\hat{\boldsymbol{\beta}}$$

Entonces:

$$\mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}})$$
$$\mathbf{e}'\mathbf{e} = (\mathbf{y}' - \mathbf{x}'\hat{\boldsymbol{\beta}}')(\mathbf{y} - \mathbf{x}\hat{\boldsymbol{\beta}})$$
$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{x}\hat{\boldsymbol{\beta}} - \mathbf{x}'\hat{\boldsymbol{\beta}}'\mathbf{y} + \mathbf{x}'\mathbf{x}\hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\beta}}$$

Y dado que $\mathbf{x}'\hat{\boldsymbol{\beta}}'\mathbf{y}$ es un número real, igual a su transposición, entonces $\mathbf{x}'\hat{\boldsymbol{\beta}}'\mathbf{y} = \mathbf{y}'\mathbf{x}\hat{\boldsymbol{\beta}}$ así:

$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \mathbf{x}'\hat{\boldsymbol{\beta}}'\mathbf{y} - \mathbf{x}'\hat{\boldsymbol{\beta}}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{x}\hat{\boldsymbol{\beta}}$$
$$\mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - 2\mathbf{x}'\hat{\boldsymbol{\beta}}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{x}\hat{\boldsymbol{\beta}}$$

Derivado parcialmente la ecuación anterior con respecto a $\hat{\boldsymbol{\beta}}$, haciendo uso de las reglas de derivación matricial dadas en el apéndice **A**.

$$\frac{\partial}{\partial\hat{\boldsymbol{\beta}}}(\mathbf{e}'\mathbf{e}) = \frac{\partial}{\partial\hat{\boldsymbol{\beta}}}(\mathbf{y}'\mathbf{y} - 2\mathbf{x}'\hat{\boldsymbol{\beta}}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{x}\hat{\boldsymbol{\beta}})$$
$$0 = 0 - 2\mathbf{x}'\mathbf{y} + 2\mathbf{x}'\mathbf{x}\hat{\boldsymbol{\beta}}$$
$$2\mathbf{x}'\mathbf{y} = 2\mathbf{x}'\mathbf{x}\hat{\boldsymbol{\beta}}$$
$$\mathbf{x}'\mathbf{y} = (\mathbf{x}'\mathbf{x})\hat{\boldsymbol{\beta}}$$
$$\frac{\mathbf{x}'\mathbf{y}}{(\mathbf{x}'\mathbf{x})} = \hat{\boldsymbol{\beta}}$$
$$(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \hat{\boldsymbol{\beta}}$$

L.q.q.d

c) Deducción de ecuación (5.25) var-cov de $\hat{\boldsymbol{\beta}}$.

Tenemos que

$$(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y} = \hat{\boldsymbol{\beta}}$$

Entonces sustituyendo $\mathbf{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ en la ecuación anterior:

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'(\mathbf{x}\beta + \varepsilon)$$
$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{x}\hat{\beta} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon$$
$$\hat{\beta} = \mathbf{I}\hat{\beta} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon$$
$$\hat{\beta} = \hat{\beta} + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon$$
$$\hat{\beta} - \beta = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon$$

Por definición:

$$\text{var-cov}(\hat{\beta}) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right]$$

$$\text{var-cov}(\hat{\beta}) = E\left[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon\,(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon\right]$$

$$\text{var-cov}(\hat{\beta}) = E\left[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon\,(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}\varepsilon'\right]$$

$$\text{var-cov}(\hat{\beta}) = E\left[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\varepsilon\varepsilon'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\right]$$

$$\text{var-cov}(\hat{\beta}) = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\,E(\varepsilon\varepsilon')\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}$$

Recordando que: las $\mathbf{x_i}$ son valores dados y $E(\varepsilon\varepsilon') = \sigma^2\mathbf{I}$ se tiene entonces que:

$$\text{var-cov}(\hat{\beta}) = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\sigma^2\mathbf{I}\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}$$
$$\text{var-cov}(\hat{\beta}) = (\mathbf{x}'\mathbf{x})^{-1}\sigma^2\mathbf{I}\mathbf{x}'\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}$$
$$\text{var-cov}(\hat{\beta}) = (\mathbf{x}'\mathbf{x})^{-1}\sigma^2\mathbf{I}\frac{\mathbf{x}'\mathbf{x}}{\mathbf{x}'\mathbf{x}}$$
$$\text{var-cov}(\hat{\beta}) = \sigma^2\mathbf{I}(\mathbf{x}'\mathbf{x})^{-1}$$
$$\text{var-cov}(\hat{\beta}) = \sigma^2(\mathbf{x}'\mathbf{x})^{-1}$$

L.q.q.d

Capítulo 6

Modelo de Regresión Lineal con Variable Independiente Cualitativa.

6.1 Introducción al Modelo de Regresión con Variable Cualitativa.

Las variables usadas en las ecuaciones de regresión, se suelen llamar variables cuantitativas, lo que significa que las variables tienen una escala bien definida de medición. Las variables como temperatura, distancia, presión e ingreso son cuantitativas, sin embargo, esto no siempre tiene que ser así y a veces es necesario usar variables cualitativas o categóricas como variables independientes en el modelo de regresión.

Las variables cualitativas son las variables que expresan distintas cualidades, características o modalidad. Cada modalidad que se presenta se denomina atributo o categoría y la medición consiste en una clasificación de dichos atributos.

El propósito del presente Capítulo es el estudio de las variables independientes de tipo cualitativo en el análisis de regresión. Veremos como la introducción de variables cualitativas, llamadas también dicótomas, convierte el análisis de regresión en un instrumento muy flexible, capaz de resolver muchos problemas.

6.2 Definición de Términos Básicos.

Análisis de Covarianza: Representa una extensión del análisis de varianza, y, es particularmente útil cuando no ha sido posible comparar muestras seleccionadas al azar.

Desestacionalización: Proceso estadístico utilizado para eliminar los efectos de la estacionalidad de una serie temporal.

Dicotomía: Es el proceso de categorización de una variable en sus modalidades posibles.

Estacionalidad: Período de tiempo asociado a determinadas actividades productivas, que se repite cíclicamente todos los años.

Interacción: Se presenta cuando la relación entre una variable independiente y una dependiente es diferente para diferentes categorías de otra variable independiente.

Variable Cualitativa: Aquellas que no aparecen en forma numérica, sino como categorías o atributos (sexo, profesión, color de ojos) y sólo pueden ser nominales u ordinales.

Variables Dicótomas: Son aquellas que, por su propia naturaleza sólo permiten 2 opciones es decir, que manifiestan o traducen una modalidad llamada atributo o categoría. Ejemplo: blanco o negro.

Se les agrupa en nominales cuando no pueden ser agrupadas numéricamente o variables ordinales como seria establecer un orden progresivo entre malo o poco, mediano o mucho.

6.3 Naturaleza de las Variables Cualitativas.

En el análisis de regresión sucede con frecuencia que la variable dependiente está influenciada no sólo por las variables fácilmente cuantificables, si no también por variables que son de naturaleza cualitativa, por ejemplo sexo, raza, color, religión guerras, huelgas, entre otras.

Como estas variables cualitativas nos indican la presencia o ausencia de una "cualidad" o "atributo", como femenino o masculino, blanco o negro, católico o no católico, una manera de cuantificar tales atributos consiste en construir variables artificiales que tomen los valores de 1 ó 0; 0 para indicar ausencia y 1 para indicar la presencia del atributo. Por ejemplo, 1 puede indicar que la persona es hombre y 0 que es mujer; 1 puede indicar que la persona es estudiante universitario graduado y 0 que no lo es, etc. Estas variables que asumen valores de 0 ó 1 se denominan variables dicótomas.

Las variables dicótomas se pueden usar en los modelos de regresión con la misma facilidad que las variables cuantitativas. De igual forma, un modelo de regresión puede contener exclusivamente variables dicótomas o de naturaleza cualitativa. Tales modelos se denominan modelos de análisis de varianza.

A manera de ejemplo, supóngase que un ingeniero mecánico desea relacionar la vida útil "y" de una cuchilla en un torno, con la clase de cuchilla que se usa para hacer las piezas, se tiene el siguiente modelo:

$$y_i = \beta_0 + \beta_1 D_i + \varepsilon_i \qquad \textbf{(6.1)}$$

Donde:

y_i: Es la vida útil de una cuchilla en un torno.

$$D_i = \begin{cases} 0 \text{ si la pieza procede de la cuchilla tipo A} \\ 1 \text{ si la pieza procede de la cuchilla tipo B} \end{cases}$$

Nótese que la ecuación (6.1) es como el modelo de regresión de dos variables visto anteriormente, con la única diferencia de que en lugar de la variable cuantitativa x_i, tenemos una variable dicótoma D_i (en adelante todas las variables dicótomas se denotarán con la letra D).

El modelo (6.1) nos permitirá saber si la clase de herramienta que se usa para hacer las piezas influye en la vida útil de estas, suponiendo, naturalmente, que todas las demás variables, se mantienen constantes. Para interpretar los parámetros en el modelo (6.1) y Bajo los supuestos del modelo de regresión lineal, se examinará el primer tipo de cuchilla el A, para el cual D = 0. El modelo de regresión se transforma en:

$$E(y_i \mid D_i = 0) = E(\beta_0) + E(\beta_1(0)) + E(\varepsilon_i)$$
$$E(y_i \mid D_i = 0) = \beta_0$$

Así, el intercepto β_0 nos da la vida útil de una herramienta para la cuchilla tipo A.

Para el tipo de cuchilla B, para el cual D = 1. El modelo es:

$$E(y_i \mid D_i = 1) = E(\beta_0) + E(\beta_1(1)) + E(\varepsilon_i)$$
$$E(y_i \mid D_i = 1) = \beta_0 + \beta_1$$

El coeficiente β_1 nos dice en cuanto difiere la vida útil de una herramienta si se hace con el tipo de cuchilla B.

La hipótesis nula de que no hay discriminación (H_0: $\beta_1 = 0$) puede verificarse fácilmente corriendo la regresión (6.1) en la forma usual y observando, por medio de la prueba t, si el $\hat{\beta}_1$ es estadísticamente significativo.

Los modelos de análisis de varianza del tipo (6.1), aunque muy comunes en Sociología, Psicología, Educación e Investigación de Mercadeos, no son tan comunes en Economía. Típicamente en la mayoría de los modelos de regresión en investigaciones económicas se encuentran tanto variables cualitativas como cuantitativas. Los modelos que contienen los dos tipos de variables se denominan modelos de análisis de covarianza.

Nos ocuparemos de ellos en este Capítulo.

6.4 Regresión de una Variable Cuantitativa y una Cualitativa con dos Categorías.

Como ejemplo de los modelos de análisis de covarianza, modifiquemos la ecuación (6.1) de la siguiente forma:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \varepsilon_i \qquad \textbf{(6.2)}$$

Donde:

y_i: Es la vida útil de una herramienta en un torno.

x_i: Velocidad del torno en revoluciones por minuto.

$$D_i = \begin{cases} 0 & \text{si la pieza procede de la cuchilla tipo A} \\ 1 & \text{si la pieza procede de la cuchilla tipo B} \end{cases}$$

El modelo dado en la ecuación (6.2) contiene dos variables independientes de las cuales una es cuantitativa (revoluciones por minuto) y la otra es cualitativa (el tipo de cuchilla) que tiene dos categorías o sea tipo A y tipo B.

Entonces el significado de la ecuación (6.2) suponiendo, como siempre que $E(\varepsilon_i) = 0$, es:

Vida útil promedio de una herramienta procedente del tipo de cuchilla A.

$$E(y_i \mid x_i, D_i = 0) = \beta_0 + \beta_1 x_i \qquad \textbf{(6.3)}$$

Así, la relación entre la vida útil promedio y la velocidad del torno para la herramienta procedente del tipo de cuchilla A es una recta con ordenada al origen β_0 y pendiente β_1.

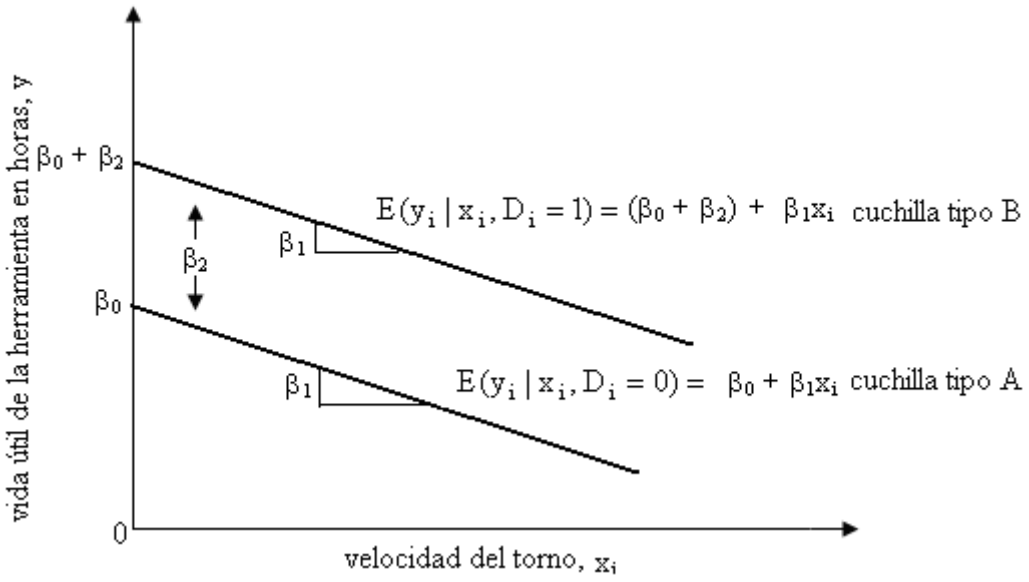
Vida útil promedio de una herramienta procedente del tipo de cuchilla B.

$$E(y_i \mid x_i, D_i = 1) = (\beta_0 + \beta_2) + \beta_1 x_i \qquad \textbf{(6.4)}$$

Esto es, para la cuchilla de tipo B la relación entre la vida útil promedio de la herramienta y la velocidad del torno también es una recta con pendiente β_1, pero con ordenada al origen $(\beta_0 + \beta_2)$.

Las dos funciones de respuesta se ven en la figura 6.1. Los modelos (6.3) y (6.4) describen dos líneas de regresión paralelas, esto es, dos rectas con una pendiente común β_1 y con distintas ordenadas al origen. También, se supone que la varianza de los errores ε_i es igual para ambos tipos de herramientas, A y B. El parámetro β_2 expresa la diferencia de alturas entre las dos líneas de regresión, ya que, β_2 es una medida de la diferencia de vida media de la herramienta que resulta de cambiar del tipo A al tipo B.

Figura 6.1 Funciones de respuesta para la vida útil de una herramienta.



Antes de continuar, es necesario anotar los siguientes puntos del modelo de regresión lineal con una variable independiente cualitativa como el que acabamos de ver:

1. Para distinguir las dos categorías, tipo A y tipo B, se introdujo una variable dicótoma D_i, dado que $D_i = 0$ denota que la herramienta procede del tipo A y $D_i = 1$ denota que la herramienta procede del tipo B, ya que sólo existen 2 posibles resultados. De este modo, una sola variable D_i es suficiente para distinguir dos categorías. Suponiendo que el modelo de regresión tiene un intercepto, si escribiéramos el modelo (6.2) como:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \varepsilon_i \qquad \textbf{(6.5)}$$

Donde: y_i y x_i son como ya se definieron,

$$D_{1i} = \begin{cases} 0 \text{ si la pieza procede de la cuchilla tipo A} \\ 1 \text{ si la pieza procede de la cuchilla tipo B} \end{cases}$$

$$D_{2i} = \begin{cases} 1 \text{ si la pieza procede de la cuchilla tipo A} \\ 0 \text{ si la pieza procede de la cuchilla tipo B} \end{cases}$$

Entonces el modelo (6.5) no podría estimarse tal como se presenta, pues hay perfecta colinealidad entre D_1 y D_2. Para verificarlo supongamos que se tiene una muestra de dos observaciones procedentes de la cuchilla tipo A y tres de la cuchilla tipo B. La matriz de datos será como se muestra a continuación:

$$\begin{array}{lc} & \begin{array}{cccc} & D_1 & D_2 & x \end{array} \\ \begin{array}{ll} \text{tipo B} & y_1 \\ \text{tipo B} & y_2 \\ \text{tipo A} & y_3 \\ \text{tipo B} & y_4 \\ \text{tipo A} & y_5 \end{array} & = \begin{array}{cccc} 1 & 1 & 0 & x_1 \\ 1 & 1 & 0 & x_2 \\ 1 & 0 & 1 & x_3 \\ 1 & 1 & 0 & x_4 \\ 1 & 0 & 1 & x_5 \end{array} \end{array}$$

La primera columna de la derecha de la matriz representa el intercepto. Se puede ver fácilmente que $D_1 = 1 - D_2$ ó $D_2 = 1 - D_1$; es decir, D_1 y D_2 son perfectamente colineales y como se verá más adelante en casos de perfecta colinealidad no es posible la estimación de Mínimos Cuadrados Ordinarios. Existen varias formas de resolver el problema, pero la más simple consiste en introducir la variable dicótoma como lo hicimos en el modelo (6.2), esto es usar únicamente una variable dicótoma si solamente hay dos categorías para la variable independiente cualitativa, en este caso la matriz anterior no tendrá la columna D_2, lo que evita el problema de la multicolinealidad.

La regla general es: *Si una variable cualitativa tiene m categorías, se deben introducir m – 1 variables dicótomas.* En nuestro ejemplo, hay dos tipos de cuchillas A y B, y, por lo tanto introdujimos sólo una variable dicótoma. Si esta regla no se sigue, caeremos en lo que se llama la trampa de la variable dicótoma, esto es, en una situación de perfecta multicolinealidad.

2. La asignación de los valores 0 y 1 a las categorías es arbitraria, en el sentido de que hubiéramos podido asignar D = 1 al tipo de cuchilla A y D = 0 al tipo de cuchilla B. Por lo tanto, para interpretar los resultados de un modelo de variables dicótomas es indispensable saber cómo se asignan los valores 0 y 1.
3. El grupo, categoría al que se le asigna el valor de cero recibe el nombre de categoría base, o de control. Es la base en el sentido de que todas las demás comparaciones se hacen con esa categoría. En el modelo (6.2) la cuchilla tipo A es la categoría base, pues si corremos la regresión con D = 0, esto es, sólo con las piezas que proceden de la cuchilla tipo A, el intercepto será β_0. Nótese además que elegir qué categoría sirve de base es un asunto de preferencias, basado algunas veces en consideraciones dadas.
4. El coeficiente β_2 correspondiente a la variable dicótoma D puede llamarse coeficiente diferencial de intercepto, pues nos dice en cuanto difiere el intercepto de la categoría que recibe el valor de 1, del coeficiente de la categoría base.

6.5 Regresión de una Variable Cuantitativa y una Cualitativa con más de dos Categorías.

Supongamos que basados en información de corte transversal queremos ver si el gasto anual de un individuo depende del ingreso y la educación que este tenga. Dado que la variable educación es de naturaleza cualitativa y considerando, tres categorías de educación mutuamente excluyentes: menos que bachiller, bachiller y nivel universitario. A diferencia del caso anterior, tenemos más de 2 categorías de la variable cualitativa educación. Siguiendo la regla de que el número de variables dicótomas debe ser uno menor que el número de categorías, debemos introducir dos variables dicótomas que tengan en cuenta las tres categorías de educación. Suponiendo que los tres grupos de educación tienen la misma pendiente pero distinto intercepto en la regresión del gasto anual en salud contra el ingreso, podemos usar el siguiente modelo:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \varepsilon_i \qquad \textbf{(6.6)}$$

Donde:

y_i: Gasto anual en salud.

x_i: Ingreso anual.

$$D_{1i} = \begin{cases} 1 \text{ si bachiller} \\ 0 \text{ si no lo es} \end{cases}$$

$$D_{2i} = \begin{cases} 1 \text{ si tiene educación universitaria} \\ 0 \text{ si no la tiene} \end{cases}$$

Nótese que en la asignación anterior de variables dicótomas estamos tratando, arbitrariamente, la categoría “menos de bachiller” como la categoría base. Por lo tanto,

el intercepto β_0 reflejará el intercepto de esta categoría. Los interceptos diferenciales β_2 y β_3 nos dicen en cuanto difieren los interceptos de las otras dos categorías, del intercepto de la categoría base. Esto puede comprobarse fácilmente de la forma siguiente:

Suponiendo $E(\varepsilon_i) = 0$ de la ecuación (6.6) se tiene:

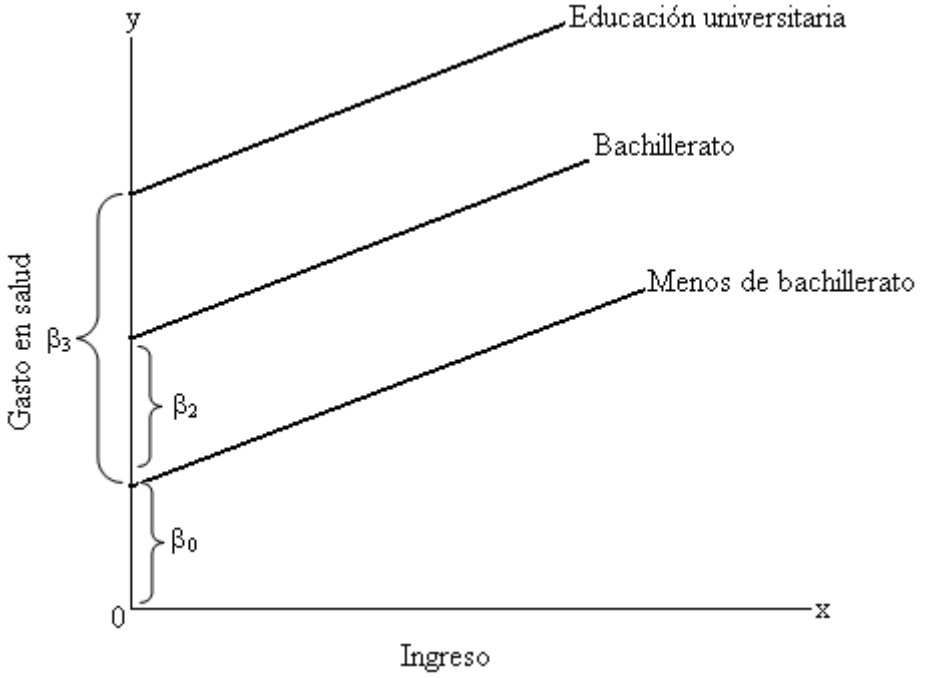
$$E(y_i \mid D_1 = 0, D_2 = 0, x_i) = \beta_0 + \beta_1 x_i \qquad \textbf{(6.7)}$$

$$E(y_i \mid D_1 = 1, D_2 = 0, x_i) = (\beta_0 + \beta_2) + \beta_1 x_i \qquad \textbf{(6.8)}$$

$$E(y_i \mid D_1 = 0, D_2 = 1, x_i) = (\beta_0 + \beta_3) + \beta_1 x_i \qquad \textbf{(6.9)}$$

Que son, respectivamente, las funciones para los tres niveles de educación: menor que el bachillerato, bachillerato y educación universitaria. Las ecuaciones anteriores se muestran en la figura 6.2 (para fines ilustrativos se supone que $\beta_3 > \beta_2$).

Figura 6.2 Gasto en salud con relación al ingreso, para tres niveles de educación.



Después de realizar la regresión (6.6), se puede averiguar si los interceptos diferenciales β_2 y β_3 son de manera individual estadísticamente significativos, es decir, diferentes del de base. Una verificación de la hipótesis nula H_0: $\beta_2 = \beta_3 = 0$ puede hacerse simultáneamente mediante la técnica análisis de varianza y la prueba F, como se vio en el Capítulo 4.

Obsérvese que la interpretación de la ecuación (6.6) cambiará si adoptamos un esquema diferente para la asignación de los valores de las variables dicótomas. Por ejemplo, si designamos $D_1 = 1$ a la categoría menor que el bachillerato y $D_2 = 1$ a bachillerato, la categoría base será la educación universitaria y todas las comparaciones se harán con relación a esa categoría.

6.6 Regresión de una Variable Cuantitativa y dos Variables Cualitativas.

La técnica de variables dicótomas puede extenderse fácilmente a más de una variable cualitativa. Para ilustrarlo, supóngase que en el ejemplo de vida útil de una herramienta ecuación (6.2), se debe considerar un segundo factor cualitativo, el tipo de lubricante de corte que se usa, suponiendo que este factor tiene dos categorías, se puede definir una segunda variable indicadora, D_{2i}, entonces un modelo de regresión que relacione la vida útil de una herramienta (y) con la velocidad de corte (x_1), el tipo de cuchilla (D_{1i}) y el tipo de lubricante de corte (D_{2i}) es:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_{1i} + \beta_3 D_{2i} + \varepsilon_i \qquad \textbf{(6.10)}$$

Donde:

y_i: Es la vida útil de una herramienta en un torno.

x_i: Velocidad del torno en revoluciones por minuto.

$$D_{1i} = \begin{cases} 0 & \text{si la pieza procede de la cuchilla tipo A} \\ 1 & \text{si la pieza procede de la cuchilla tipo B} \end{cases}$$

$$D_{2i} = \begin{cases} 0 & \text{si se usa aceite de baja viscosidad} \\ 1 & \text{si se usa aceite de viscosidad intermedia} \end{cases}$$

Se puede ver que cada una de las variables cualitativas, tiene dos categorías y por lo tanto sólo se necesita una variable dicótoma para cada una.

Se puede observar de la ecuación (6.10) que la pendiente β_1, del modelo de regresión que relaciona la vida útil de la herramienta con la velocidad de corte no depende ni del tipo de cuchilla ni del tipo de lubricante de corte. La ordenada al origen de la recta de regresión sí depende de esos factores de una forma aditiva.

Ahora suponiendo que $E(\varepsilon_i) = 0$, a partir de la ecuación (6.10) podemos obtener:

Vida promedio de la herramienta procedente del tipo A, usando aceite de baja viscosidad:

$$E(y_i \mid D_1 = 0, D_2 = 0, x_i) = \beta_0 + \beta_1 x_i \qquad \textbf{(6.11)}$$

Vida promedio de la herramienta procedente del tipo B, usando aceite de baja viscosidad:

$$E(y_i \mid D_1 = 1, D_2 = 0, x_i) = (\beta_0 + \beta_2) + \beta_1 x_i \qquad \textbf{(6.12)}$$

Vida promedio de la herramienta procedente del tipo A, usando aceite de viscosidad intermedia:

$$E(y_i \mid D_1 = 0, D_2 = 1, x_i) = (\beta_0 + \beta_3) + \beta_1 x_i \qquad \textbf{(6.13)}$$

Vida promedio de la herramienta procedente del tipo B usando aceite de viscosidad intermedia:

$$E(y_i \mid D_1 = 1, D_2 = 1, x_i) = (\beta_0 + \beta_2 + \beta_3) + \beta_1 x_i \qquad \textbf{(6.14)}$$

Una vez, más suponemos que las regresiones anteriores difieren solamente en el intercepto y no en la pendiente.

Una estimación por Mínimos Cuadrados Ordinarios de la ecuación (6.10) nos permitirá verificar una variedad de hipótesis. De este modo, si β_3 es estadísticamente significativa, esto nos dará a entender que el tipo de lubricante que se usa en el corte de la herramienta sí afecta la vida útil de esta. De igual forma, si β_2 es significativa, esto significará que el tipo de cuchilla que se utiliza también influye en la vida útil de la herramienta. Si ambos interceptos diferenciales son estadísticamente significativos, querrá decir que tanto el tipo de cuchilla como el tipo de lubricante, son importantes en la determinación de la vida útil de la herramienta.

En general y siguiendo la exposición anterior, podemos extender nuestro modelo a más de una variable cuantitativa y dos cualitativas. La única precaución que debemos tener es que el número de variables dicótomas para cada variable cualitativa sea uno menos que el número de categorías de esa variable.

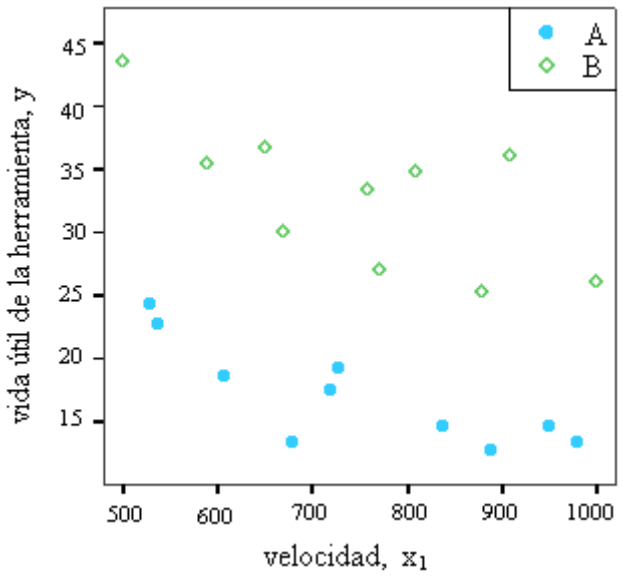
Ejemplo 1: Datos de vida de herramienta.

En la tabla 6.1 se presentan 20 observaciones de duración de la herramienta "y" y velocidad del torno (rpm) x_{1i}, el diagrama de dispersión se ve en la figura 6.3.

Tabla 6.1 Datos de vida de la herramienta.

y (horas)	x(rpm)	tipo de herramienta	y (horas)	x(rpm)	tipo de herramienta
18.73	610	A	30.16	670	B
14.52	950	A	27.09	770	B
17.43	720	A	25.4	880	B
14.54	840	A	26.05	1000	B
13.44	980	A	33.49	760	B
24.39	530	A	35.62	590	B
13.34	680	A	26.07	910	B
22.71	540	A	36.78	650	B
12.68	890	A	34.95	810	B
19.32	730	A	43.67	500	B

Figura 6.3 Vida útil de la herramienta "y" en función de la velocidad del torno x_{1i}, para los tipos de cuchillas A y B.



Se ajustará el siguiente modelo:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \varepsilon_i$$

En donde la variable indicadora $D_i = 0$ si la observación procede de la cuchilla tipo A, y $D_i = 1$ si procede de la cuchilla tipo B. La matriz **x** y el vector **y** para ajustar este modelo son:

$$\mathbf{x} = \begin{bmatrix} 1 & 610 & 0 \\ 1 & 950 & 0 \\ 1 & 720 & 0 \\ 1 & 840 & 0 \\ 1 & 980 & 0 \\ 1 & 530 & 0 \\ 1 & 680 & 0 \\ 1 & 540 & 0 \\ 1 & 890 & 0 \\ 1 & 730 & 0 \\ 1 & 670 & 1 \\ 1 & 770 & 1 \\ 1 & 880 & 1 \\ 1 & 1000 & 1 \\ 1 & 760 & 1 \\ 1 & 590 & 1 \\ 1 & 910 & 1 \\ 1 & 650 & 1 \\ 1 & 810 & 1 \\ 1 & 500 & 1 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 18.73 \\ 14.52 \\ 17.43 \\ 14.54 \\ 13.44 \\ 24.39 \\ 13.34 \\ 22.71 \\ 13.68 \\ 19.32 \\ 30.16 \\ 27.09 \\ 25.40 \\ 2.05 \\ 33.49 \\ 35.62 \\ 26.07 \\ 36.78 \\ 34.95 \\ 43.67 \end{bmatrix}$$

Haciendo uso del algebra matricial y siguiendo los pasos dados en el ejemplo 1 del Capítulo 5 se obtiene:

$$\mathbf{x'x} = \begin{bmatrix} n & \sum_{i=1}^{20} x_{1i} & \sum_{i=1}^{20} D_i \\ \sum_{i=1}^{20} x_{1i} & \sum_{i=1}^{20} x_{1i}^2 & \sum_{i=1}^{20} x_{1i} D_i \\ \sum_{i=1}^{20} D_i & \sum_{i=1}^{20} x_{1i} D_i & \sum_{i=1}^{20} D_i^2 \end{bmatrix} = \begin{bmatrix} 20 & 15010 & 10 \\ 15010 & 11717500 & 7540 \\ 10 & 7540 & 10 \end{bmatrix}$$

$$\mathbf{x'y} = \begin{bmatrix} \sum_{i=1}^{20} y_i \\ \sum_{i=1}^{20} x_{1i} y_i \\ \sum_{i=1}^{20} D_i y_i \end{bmatrix} = \begin{bmatrix} 490.38 \\ 356515.7 \\ 319.28 \end{bmatrix}$$

Para encontrar el valor de los coeficientes de regresión, necesitamos calcular la inversa de la matriz $\mathbf{x'x}$, para ello hacemos uso de las reglas de inversión de matrices dadas en el apéndice **A**.

Calculamos el determinante de la matriz $\mathbf{x'x}$ como se muestra:

$$|\mathbf{x'x}| = \begin{vmatrix} 20 & 15010 & 10 \\ 15010 & 11717500 & 7540 \\ 10 & 7540 & 10 \end{vmatrix}$$

$$|\mathbf{x'x}| = 20 \begin{vmatrix} 11717500 & 7540 \\ 7540 & 10 \end{vmatrix} - 15010 \begin{vmatrix} 15010 & 10 \\ 7540 & 10 \end{vmatrix} + 10 \begin{vmatrix} 15010 & 10 \\ 11717500 & 7540 \end{vmatrix}$$

$$|\mathbf{x'x}| = 45225000$$

La matriz de cofactores es la que se muestra a continuación

$$\mathbf{C} = \begin{bmatrix} 60323400 & -74700 & -3999600 \\ -74700 & 100 & -700 \\ -3999600 & -700 & 9049900 \end{bmatrix}$$

Transponiendo la matriz de cofactores anterior se obtiene la matriz adjunta:

$$(\text{adj}\,\mathbf{x}'\mathbf{x}) = \begin{bmatrix} 60323400 & -74700 & -3999600 \\ -74700 & 100 & -700 \\ -3999600 & -700 & 9049900 \end{bmatrix}$$

Dividimos los elementos de la $(\text{adj}\,\mathbf{x}'\mathbf{x})$ por el valor del determinante $|\mathbf{x}'\mathbf{x}|$ = 45225000 y obtenemos:

$$(\mathbf{x}'\mathbf{x})^{-1} = \frac{1}{|\mathbf{x}'\mathbf{x}|}\,(\text{adj}\,\mathbf{x}'\mathbf{x}) = \begin{bmatrix} \frac{60323400}{45225000} & -\frac{74700}{45225000} & -\frac{3999600}{45225000} \\ -\frac{74700}{45225000} & \frac{100}{45225000} & -\frac{700}{45225000} \\ -\frac{3999600}{45225000} & -\frac{700}{45225000} & \frac{9049900}{45225000} \end{bmatrix}$$

Ahora obtenemos los valores de los coeficientes de la forma siguiente:

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \frac{60323400}{45225000} & -\frac{74700}{45225000} & -\frac{3999600}{45225000} \\ -\frac{74700}{45225000} & \frac{100}{45225000} & -\frac{700}{45225000} \\ -\frac{3999600}{45225000} & -\frac{700}{45225000} & \frac{9049900}{45225000} \end{bmatrix} \begin{bmatrix} 490\ .\ 38 \\ 356515\ .7 \\ 319\ .\ 28 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 36.986012 \\ -0.02660723 \\ 15.00425061 \end{bmatrix}$$

El ajuste del modelo por Mínimos Cuadrados Ordinarios es:

$$\hat{y} = 36.986012 - 0.02660723\ x_1 + 15.00425061\ D_1 \qquad \textbf{(6.15)}$$

La suma de los errores al cuadrado puede calcularse como:

$$\sum_{i=1}^{20} e_i^2 = \mathbf{e'e} = \mathbf{y'y} - \hat{\boldsymbol{\beta}}'\mathbf{x'y}$$

$$\mathbf{y'y} = \sum_{i=1}^{20} y_i^2 = 13598.7154$$

$$\hat{\boldsymbol{\beta}}'\mathbf{x'y} = \begin{bmatrix} 6.986012 & -0.02660723 & 15.00425061 \end{bmatrix} \begin{bmatrix} 490.38 \\ 356515.7 \\ 319.28 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}}'\mathbf{x'y} = 13441.86247$$

Por lo tanto la suma de los errores al cuadrado es:

$$\sum_{i=1}^{20} e_i^2 = \mathbf{e'e} = \mathbf{y'y} - \hat{\boldsymbol{\beta}}'\mathbf{x'y}$$

$$\sum_{i=1}^{20} e_i^2 = 13598.7154 - 13441.86247$$

$$\sum_{i=1}^{20} e_i^2 = 156.85293$$

De donde obtenemos:

$$\hat{\sigma}^2 = \frac{\mathbf{e'e}}{n-k} = \frac{\mathbf{y'y} - \hat{\boldsymbol{\beta}}'\mathbf{x'y}}{n-k}$$

$$\hat{\sigma}^2 = \frac{156.85293}{20-3} = 9.226$$

La matriz de varianza-covarianza para $\hat{\boldsymbol{\beta}}$ puede escribirse como:

$$\text{var} - \text{cov}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{x'x})^{-1} = 9.226 \begin{bmatrix} \frac{60323400}{45225000} & -\frac{74700}{45225000} & -\frac{3999600}{45225000} \\ -\frac{74700}{45225000} & \frac{100}{45225000} & -\frac{700}{45225000} \\ -\frac{3999600}{45225000} & -\frac{700}{45225000} & \frac{9049900}{45225000} \end{bmatrix}$$

$$\text{var} - \text{cov}(\hat{\beta}) = \hat{\sigma}^2 (\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} 12.3061 & -0.0152 & -0.8159 \\ -0.0152 & 0.0000204 & -0.0001 \\ -0.8159 & -0.0001 & 1.8462 \end{bmatrix}$$

Los elementos de la diagonal de esta matriz nos dan las varianzas de $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$, respectivamente, y sus raíces cuadradas positivas nos dan los correspondientes errores estándar.

Con la información anterior encontramos ahora el valor de R^2 así.

$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2$

$SS_R = 13441.86247 - 20(24.519)^2$

$SS_R = 1418.235$

$SS_T = \mathbf{y}'\mathbf{y} - n\bar{y}^2$

$SS_T = 13598.7154 - 20(24.519)^2$

$SS_T = 1575.088$

$$R^2 = \frac{\hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2}{\mathbf{y}'\mathbf{y} - n\bar{y}^2}$$

$$R^2 = \frac{SS_R}{SS_T} = \frac{1418.235}{1575.088} = 0.9004$$

La interpretación de la ecuación (6.15) es: si ambos D_1 y x_1 están fijos en cero, el valor promedio de la variable dependiente (Vida útil) se estima en $\hat{\beta}_0 = 36.986012$. El valor de la pendiente $\hat{\beta}_1 = -0.02660723$ es la disminución promedio en la vida útil de la herramienta, debido a la velocidad del torno en revoluciones por minuto. El coeficiente

de regresión parcial $\hat{\beta}_2 = 15.00425061$ significa que manteniendo todas las demás variables constantes, un aumento en la vida promedio de la herramienta de, por ejemplo 1 hora depende del tipo de cuchilla que se utiliza.

El valor de $R^2 = 0.9004$ muestra que las dos variables independientes (tipos de cuchillas y velocidad del torno) explican el 90.04% de la variación en la vida útil promedio de la herramienta.

Prueba de hipótesis para los coeficientes individuales de regresión.

Con los datos obtenidos anteriormente realizamos la prueba de hipótesis individual para $\hat{\beta}_2$ es decir, $H_0 : \beta_2 = 0$ y $H_1 : \beta_2 \neq 0$.

Solución:

1. $H_0 : \beta_2 = 0$
2. $H_1 : \beta_2 \neq 0$
3. Se selecciona un nivel de significancia de $\alpha = 0.05$ y como la prueba es de dos colas $\alpha/2 = 0.05/2 = 0.025$ y se tiene que el valor de la tabla de t es:

   $t_{(0.05/2,\ 20-3)} = t_{(0.025,\ 17)} = 2.110$
4. Región critica: si $t < -2.110$ ó $t > 2.110$, entonces rechazamos H_0.
5. Cálculos:

$$t_0 = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{33}}} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\hat{\sigma}^2 C_{33}}} = \frac{15.00425061}{\sqrt{1.8462}} = 11.042$$

6. Decisión Estadística: se rechaza H_0 porque el valor calculado t_0 = 11.042 es mayor que el de la tabla (2.110).
7. Conclusión: se concluye que hay una relación lineal entre el tipo de cuchilla y la vida útil de la herramienta.

De igual forma se realiza la prueba de hipótesis parcial para los demás coeficientes de regresión.

Como se mencionó en el Capítulo 4, no es posible aplicar la prueba t para verificar la hipótesis global según la cual $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$.

Sin embargo, recuérdese que una hipótesis nula $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$ puede ser verificada mediante la técnica de análisis de varianza y la prueba F dadas anteriormente.

Se probará la significancia global de la regresión para los datos de los tipos de herramientas, es decir, $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$ y $H_1 : \beta_j \neq 0$, al menos para un j.

Datos:

El modelo ajustado es: $\hat{y} = 36.986012 - 0.02660723\ x_1 + 15.00425061\ D_1$

SS_R = 1418.235

SS_{Res} = 156.85293

SS_T = 1575.088

Solución:

1. $H_0 : \beta_0 = \beta_1 = \beta_2 = 0$
2. $H_1 : \beta_j \neq 0$, al menos para un j.

3. Se selecciona un nivel de significancia de $\alpha = 0.05$ y se tiene que el valor de la tabla F es $F_{(0.05,\ 2,\ 17)} = 3.59$.

4. Cálculos:

$$F_0 = \frac{\dfrac{\hat{\beta}'x'y - n\bar{y}^2}{k - 1}}{\dfrac{y'y - \hat{\beta}'x'y}{n - k}} = \frac{\dfrac{SS_R}{k-1}}{\dfrac{SS_{Res}}{n-k}} = \frac{\dfrac{1418.235}{3-1}}{\dfrac{156.85293}{20-3}} = 76.855$$

Tabla 6.2 Análisis de varianza para las variables del ejemplo de herramientas.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F_0
Regresión	1418.235	2	709.1175	76.855
Residual	156.85293	17	9.2266	
Total	1575.088	19		

5. Decisión Estadística: se rechaza H_0, porque el valor calculado para F_0 (76.855) es mayor que el de la tabla (3.59).

6. Conclusión: Se concluye que la vida útil de la herramienta se relaciona con el tipo de cuchilla que se usa y con la velocidad del torno, en revoluciones por minuto, para la muestra dada.

El intervalo de confianza del 95% para β_2 es:

$$\hat{\beta}_2 - t_{(\alpha/2,\ n-k)}es(\hat{\beta}_2) \le \beta_2 \le \hat{\beta}_2 + t_{(\alpha/2,\ n-k)}es(\hat{\beta}_2)$$
$$15.00425061 - 2.110(1.360) \le \beta_2 \le 15.00425061 + 2.110(1.360)$$
$$12.135 \le \beta_2 \le 17.873$$

Se tiene el 95% de confianza de que el verdadero parámetro β_2 se encuentra entre 12.135 y 17.873.

Nota:

Al igual que en los Capítulos anteriores se puede utilizar el Software estadístico SPSS para realizar la regresión lineal con variables cuantitativas y cualitativas, la única diferencia es que los datos de la variable cualitativa son ceros y unos.

6.7 Interacción entre Variables Cualitativas y Cuantitativas.

Al revisar el diagrama de dispersión figura 6.3 se ve que se requieren dos líneas de regresión para modelar bien los datos, y que la ordenada al origen depende del tipo de cuchilla que se usa.

En vista de que se requieren dos líneas de regresión distintas para modelar la relación entre la "vida útil de la herramienta" y la "velocidad del torno", se podrían ajustar dos modelos separados rectilíneos, en lugar de uno solo con una variable indicadora.

Sin embargo, se prefiere el método con un solo modelo, porque sólo se tiene una ecuación final con la que se trabaja, y no dos, es un resultado práctico mucho más simple; además, como se supone que las dos rectas tienen la misma pendiente, tiene sentido combinar los datos de ambos tipos para producir un solo estimado de este parámetro común; este método también proporciona una estimación de la varianza común del error σ^2, y se tienen más grados de libertad que los que resultarían de ajustar dos líneas separadas de regresión.

Supongamos que se espera que las rectas de regresión que relacionan la vida útil de la herramienta con la velocidad del torno difieren tanto en la ordenada al origen como en la pendiente. Es posible modelar este caso con una sola ecuación de regresión, usando variables indicadoras, el modelo es:

$$y = \beta_0 + \beta_1 x_i + \beta_2 D_1 + \beta_3 D_1 x_i + \varepsilon \qquad \textbf{(6.16)}$$

Al comparar las ecuaciones (6.16) con la (6.2) se observa que se agregó al modelo un producto cruzado entre x_1, la velocidad del torno y la variable indicadora que representa el tipo de cuchilla D_1. Para interpretar los parámetros en este modelo, se examinará primero la cuchilla tipo A, para la que $D_1 = 0$. El modelo (6.16) se transforma en:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_i + \beta_2 (0) + \beta_3 (0) x_i + \varepsilon \\ y &= \beta_0 + \beta_1 x_i + \varepsilon \end{aligned} \qquad \textbf{(6.17)}$$

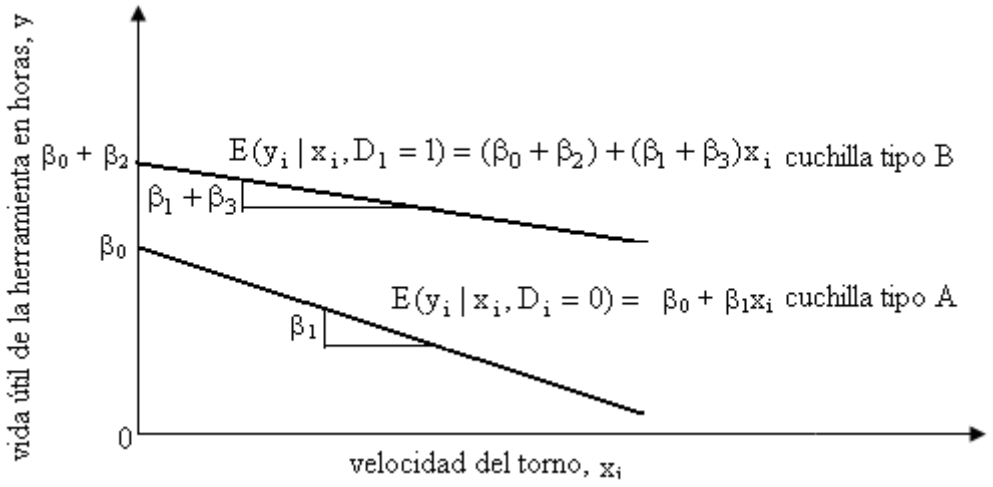
Que es una recta con ordenada al origen β_0 y pendiente β_1.

Para la cuchilla tipo B, $D_1 = 1$ es:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_i + \beta_2 (1) + \beta_3 (1) x_i + \varepsilon \\ y &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_i + \varepsilon \end{aligned} \qquad \textbf{(6.18)}$$

Es un modelo rectilíneo con ordenada al origen $\beta_0 + \beta_2$ y pendiente $\beta_1 + \beta_3$. Las dos funciones de regresión se grafican en la figura 6.4. Se puede ver que la ecuación (6.16) define dos rectas de regresión con distintas pendientes y ordenadas al origen. En consecuencia, el parámetro β_2 refleja el cambio de la ordenada al origen asociado con el cambio de cuchilla tipo A, a cuchilla tipo B (las clases 0 y 1 de la variable indicadora D_1), y β_3 indica el cambio de pendiente asociado con el cambio de tipos de cuchillas, de A a B.

Figura 6.4 Funciones de respuesta para la ecuación (6.16).



Una ventaja del uso de variables indicadoras es que las pruebas de hipótesis se pueden hacer en forma directa, con el método de la suma extra de cuadrados (o prueba F parcial).

Para el caso de una variable se vio anteriormente que la contribución de cada variable independiente se puede probar utilizando pruebas individuales sobre los parámetros por medio de la distribución t – de Student.

El método estadístico suma extra de cuadrados permite conocer no solamente la contribución de una variable sino la de cualquier subconjunto de variables.

Para ilustrar la utilidad de este procedimiento, se considera el siguiente modelo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Las sumas de cuadrados

$$SS_{RE}(\beta_1 \mid \beta_0, \beta_2, \beta_3)$$
$$SS_{RE}(\beta_2 \mid \beta_0, \beta_1, \beta_3)$$

y

$$SS_{RE}(\beta_3 \mid \beta_0, \beta_1, \beta_2)$$

Donde:

SS_{RE}: Suma de Cuadrados de Regresión del modelo reducido.

Son las sumas de cuadrados de regresión de un grado de libertad que miden la contribución de cada variable x_j, j = 1, 2, 3, al modelo, dado que todas las demás variables ya estaban en él. Esto es, se evalúa la ventaja de agregar x_j a un modelo que no incluía a esta variable. En general se puede determinar:

$$SS_{RE}(\beta_j \mid \beta_0, \beta_1, ..., \beta_{j-1}, \beta_{j+1}, ..., \beta_k), \qquad 1 \le j \le k$$

Que es el aumento en la suma de cuadrados de regresión, debido a agregar x_j a un modelo que ya contiene $x_1, \ldots, x_{j-1}, \ldots, x_k$.

Por ejemplo, para ver la contribución de x_1, se obtiene de la diferencia entre la suma de cuadrados de los coeficientes de regresión del modelo completo (SS_R) y la suma de cuadrados de los coeficientes de regresión del modelo reducido (SS_{RE}) así:

$$SS_{RE}(\beta_1 \mid \beta_0, \beta_2, \beta_3, ..., \beta_k) = SS_R(\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_k) - SS_{RE}(\beta_0, \beta_2, \beta_3, ..., \beta_k)$$

Donde $SS_R(\beta_0,\beta_1,\beta_2,\beta_3,...,\beta_k)$ es la Suma de Cuadrados de Regresión del modelo completo, y $SS_{RE}(\beta_0,\beta_2,\beta_3,...,\beta_k)$ es la Suma de Cuadrados de Regresión del modelo reducido, es decir, eliminada $\beta_1 x_1$ del modelo.

Para probar la hipótesis

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Se calcula:

$$F_0 = \frac{SS_{RE}(\beta_1 \mid \beta_0,\beta_2,\beta_3,...,\beta_k)}{\hat{\sigma}^2} = \frac{[SS_R(\beta_0,\beta_1,\beta_2,\beta_3,...,\beta_k) - SS_{RE}(\beta_0,\beta_2,\beta_3,...,\beta_k)]/1}{MS_{Res}}$$

Si el valor calculado de F_0 es mayor que el de la tabla $F_{\alpha\,(1,\,n\,-\,k)}$ (con un grado de libertad en el numerador debido a que sólo se está probando la contribución de x_1) y n-k en el denominador se rechaza la hipótesis nula.

De manera similar, se puede probar la significancia de un subconjunto de las variables. Por ejemplo, para investigar simultáneamente la importancia de incluir x_1 y x_2 en el modelo, se prueba la hipótesis

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \beta_j \neq 0, \quad \text{al menos para un j.}$$

Se calcula:

$$F_0 = \frac{[SS_{RE}(\beta_1,\beta_2 \mid \beta_0,\beta_3,\beta_4,...,\beta_k)]/2}{\hat{\sigma}^2}$$
$$F_0 = \frac{[SS_R(\beta_0,\beta_1,\beta_2,\beta_3,...,\beta_k) - SS_{RE}(\beta_0,\beta_3,\beta_4,...,\beta_k)]/2}{MS_{Res}}$$

Y se compara con el de la tabla, si el valor calculado F_0 es mayor que el de la tabla $F_{\alpha\,(2,\,n\,-\,k)}$, se rechaza la hipótesis nula.

El número de grados de libertad asociados con el numerador, es igual al número de variables en el subconjunto, en el caso anterior tenemos las variables x_1 y x_2 en el subconjunto, por lo que los grados de libertad del numerador es igual a 2. Los grados de libertad del denominador se calculan igual que antes n – k (n – número de parámetros estimados en el modelo completo).

Por ejemplo, para probar si los dos modelos de regresión (ejemplo1) son idénticos, las hipótesis serían:

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \beta_j \neq 0, \quad \text{al menos para un j.}$$

Si no se rechaza $H_0 : \beta_2 = \beta_3 = 0$, entonces un solo modelo de regresión puede explicar la relación entre la vida útil de la herramienta y la velocidad del torno. Para probar si las dos rectas de regresión tienen la misma pendiente pero quizá distintas ordenadas al origen, las hipótesis son:

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

Si se usa el modelo (6.16), las dos rectas de regresión se pueden ajustar, y se pueden hacer esas pruebas calculando la suma de cuadrados SS_{RE} (β_1, β_0) que es el modelo de regresión lineal simple, SS_{RE} (β_0, β_1, β_2) es un modelo de regresión lineal múltiple con dos variables independientes y SS_{RE} ($\beta_3|\beta_0$, β_1, β_2) es un modelo de regresión lineal

múltiple con tres variables independientes, donde se quiere ver la contribución de la variable x_3 al modelo.

Ejemplo 2: Datos de duración de herramienta.

Se ajustará el modelo de regresión:

$$y = \beta_0 + \beta_1 x_i + \beta_2 D_1 + \beta_3 D_1 x_i + \varepsilon$$

A los datos de vida útil de herramienta de la tabla 6.1. La matriz **x** y el vector **y** para este modelo son:

$$\mathbf{x} = \begin{array}{c} \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array} \begin{array}{c} x_1 \\ 610 \\ 950 \\ 720 \\ 840 \\ 980 \\ 530 \\ 680 \\ 540 \\ 890 \\ 730 \\ 670 \\ 770 \\ 880 \\ 1000 \\ 760 \\ 590 \\ 910 \\ 650 \\ 810 \\ 500 \end{array} \begin{array}{c} D_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array} \begin{array}{c} x_1 D_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 670 \\ 770 \\ 880 \\ 1000 \\ 760 \\ 590 \\ 910 \\ 650 \\ 810 \\ 500 \end{array}$$

$$\mathbf{y} = \begin{array}{c} \\ 18.73 \\ 14.52 \\ 17.43 \\ 14.54 \\ 13.44 \\ 24.39 \\ 13.34 \\ 22.71 \\ 13.68 \\ 19.32 \\ 30.16 \\ 27.09 \\ 25.40 \\ 2.05 \\ 33.49 \\ 35.62 \\ 26.07 \\ 36.78 \\ 34.95 \\ 43.67 \end{array}$$

Para estimar los parámetros del modelo se sigue el procedimiento mostrado anteriormente.

El modelo de regresión estimado es:

$$\hat{y} = 32.775 - 0.021 x_1 + 23.971 D_1 - 0.012 x_1 D_1 \qquad \textbf{(6.19)}$$

Para probar la hipótesis que los dos modelos de regresión son idénticos, se usa la estadística,

$$F_0 = \frac{SS_{RE}(\beta_2,\ \beta_3 \mid \beta_0, \beta_1)/2}{\hat{\sigma}^2} = \frac{[SS_R(\beta_0, \beta_1, \beta_2, \beta_3) - SS_{RE}(\beta_0, \beta_1)]/2}{MS_{Res}}$$

Si el F calculado excede el de la tabla, rechazar la hipótesis de que los dos modelos de regresión son iguales.

Para calcular el valor de F_0 se necesitan las sumas de cuadrados debida a la regresión del modelo completo (SS_R) y del modelo reducido (SS_{RE}). Llamamos modelo completo a la regresión hecha con las dos variables independientes más el término de interacción, ecuación (6.19), es decir que, para obtener SS_R, se debe ejecutar un análisis de regresión múltiple entre "y" y las variables x_1, D_1, x_1D_1.

Un modelo reducido se hace eliminando una de las variables cualitativas, en nuestro ejemplo al eliminar la variable cualitativa se elimina también el término de interacción, quedando así un modelo de regresión simple como modelo reducido, así para obtener SS_{RE} se debe ejecutar un análisis de regresión simple entre "y" y la variable x_1.

Datos:

La suma de cuadrados debida a la regresión (SS_R) del modelo completo y el modelo reducido (SS_{RE}) es:

$$SS_R(\beta_0, \beta_1, \beta_2, \beta_3) = \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2 = 1434.112$$

$$SS_{RE}(\beta_0, \beta_1) = \hat{\beta}_1 S_{xy} = 293.005$$

$$SS_{RE}(\beta_2, \beta_3 \mid \beta_0, \beta_1) = SS_R(\beta_0, \beta_1, \beta_2, \beta_3) - SS_{RE}(\beta_0, \beta_1)$$
$$SS_{RE}(\beta_2, \beta_3 \mid \beta_0, \beta_1) = 1434.112 - 293.005$$
$$SS_{RE}(\beta_2, \beta_3 \mid \beta_0, \beta_1) = 1141.107$$

La varianza de los residuos es la siguiente:

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}}{n-k}$$

$$\hat{\sigma}^2 = \frac{140.976}{20-4} = 8.811$$

Solución:

1. $H_0 : \beta_2 = \beta_3 = 0$
2. $H_1 : \beta_j \neq 0$, al menos para un j.
3. Se selecciona un nivel de significancia de $\alpha = 0.05$, se tiene que el valor de la tabla F es $F_{(0.05, 2, 16)} = 3.63$
4. Cálculos:

$$F_0 = \frac{SS_{RE}(\beta_2, \beta_3 \mid \beta_0, \beta_1)/2}{\hat{\sigma}^2} = \frac{1141.107/2}{8.811} = 64.75$$

5. Decisión Estadística: Se rechaza $H_0 : \beta_2 = \beta_3 = 0$ porque el valor calculado para F_0 (64.75) es mayor que el de la tabla (3.63).
6. Conclusión: Se concluye que los dos modelos de regresión no son idénticos.

Para probar la hipótesis que las dos rectas tienen distintas ordenadas al origen y una pendiente común ($H_0 : \beta_3 = 0$) se usa el estadístico:

$$F_0 = \frac{SS_{RE}(\beta_3 \mid \beta_0, \beta_1, \beta_2)/1}{\hat{\sigma}^2} = \frac{[SS_R(\beta_0, \beta_1, \beta_2, \beta_3) - SS_{RE}(\beta_0, \beta_1, \beta_2)]/1}{MS_{Res}}$$

Si el F calculado excede el de la tabla, rechazar la hipótesis de que los dos modelos de regresión tienen la misma pendiente. Para obtener $SS_{RE}(\beta_0, \beta_1, \beta_2)$ se debe ejecutar un análisis de regresión múltiple entre "y" y las variables x_1 y D_1.

Datos:

La suma de cuadrados debida a la regresión (SS_R) y la del modelo reducido (SS_{RE}) es:

$$SS_R(\beta_0, \beta_1, \beta_2, \beta_3) = \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2 = 1434.112$$

$$SS_{RE}(\beta_0, \beta_1, \beta_2) = \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2 = 1418.034$$

$$SS_{RE}(\beta_3 \mid \beta_0, \beta_1, \beta_2) = SS_R(\beta_0, \beta_1, \beta_2, \beta_3) - SS_{RE}(\beta_0, \beta_1, \beta_2)$$
$$SS_{RE}(\beta_3 \mid \beta_0, \beta_1, \beta_2) = 1434.112 - 1418.034$$
$$SS_{RE}(\beta_3 \mid \beta_0, \beta_1, \beta_2) = 16.078$$

La varianza de los residuos es la siguiente:

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}}{n-k}$$

$$\hat{\sigma}^2 = \frac{140.976}{20-4} = 8.811$$

Solución:

1. $H_0 : \beta_3 = 0$
2. $H_1 : \beta_3 \neq 0$
3. Se selecciona un nivel de significancia de $\alpha = 0.05$, se tiene que el valor de la tabla F es $F_{(0.05,\ 1,\ 16)} = 4.49$
4. Cálculos:

$$F_0 = \frac{SS_{RE}(\beta_3 \mid \beta_0, \beta_1, \beta_2)/1}{\hat{\sigma}^2} = \frac{16.078}{8.811} = 1.82$$

5. Decisión Estadística: no se rechaza $H_0 : \beta_3 = 0$ porque el valor calculado para F_0 (1.82) es menor que el de la tabla (4.49).
6. Conclusión: Se concluye que las pendientes de las dos rectas son iguales.

Las variables cualitativas son útiles en diversos casos de regresión, el ejemplo siguiente es una de muchas aplicaciones de estas.

Ejemplo 3:

Una empresa eléctrica esta investigando el efecto que tiene el tamaño de una vivienda familiar y el tipo de acondicionamiento de aire que se usa en ella, sobre el consumo total de electricidad durante los meses calurosos. Sea "y" el consumo eléctrico total (en kilowatts-horas), durante el periodo de febrero a mayo, y x_1 el tamaño de la casa (pies cuadrados de construcción). Hay cuatro tipos de sistemas de acondicionamiento de aire:

1) Sin acondicionamiento.
2) Unidades de ventanas.
3) Bomba térmica.
4) Acondicionamiento central.

Los cuatro niveles de ese factor se pueden modelar con tres variables indicadoras, D_1, D_2 y D_3, que se definen como sigue:

Tipo de acondicionamiento de aire	D_1	D_2	D_3
Sin acondicionamiento de aire	0	0	0
Unidades de ventanas	1	0	0
Bomba térmica	0	1	0
Acondicionamiento central de aire	0	0	1

El modelo de regresión es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + \varepsilon \qquad \textbf{(6.20)}$$

Si la casa no tiene acondicionamiento de aire, la ecuación (6.20) se transforma en:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Si la casa tiene unidades de ventanas, entonces:

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$$

Si la casa tiene bomba térmica, el modelo de regresión es:

$$y = (\beta_0 + \beta_3) + \beta_1 x_1 + \varepsilon$$

Y si la casa tiene acondicionamiento central, entonces:

$$y = (\beta_0 + \beta_4) + \beta_1 x_1 + \varepsilon$$

Así, en el modelo (6.20) se supone que la relación entre el consumo eléctrico en tiempo caluroso, y el tamaño de la casa es lineal, y que la pendiente no depende del tipo de sistema de acondicionamiento de aire que se emplea. Los parámetros β_2, β_3 y β_4 modifican la altura (u ordenada al origen) del modelo de regresión para los distintos sistemas de acondicionamiento de aire. Esto es, β_2, β_3 y β_4 miden el efecto de las unidades de ventanas, de bomba térmica y de acondicionamiento central, respectivamente, en comparación con la falta de acondicionamiento de aire. Además se pueden determinar otros efectos comparando en forma directa los coeficientes adecuados de regresión. Por ejemplo, $\beta_3 - \beta_4$ refleja la eficiencia relativa de una bomba térmica respecto al acondicionamiento central de aire, también nótese la hipótesis que la varianza del consumo de energía no depende del tipo de sistema de acondicionamiento usado; esta hipótesis puede ser inadecuada.

En este problema parece irreal suponer que la pendiente de la función de regresión que relaciona el consumo eléctrico medio con el tamaño de la vivienda no depende del tipo de sistema de acondicionamiento de aire. Por ejemplo, se puede esperar que el consumo eléctrico medio aumente al aumentar el tamaño de la casa, pero la tasa de aumento debería de ser distinta para un sistema de acondicionamiento de aire que para las unidades de ventanas, porque el primero debería ser más eficiente que las unidades de ventanas para las casas más grandes.

Esto es, debería haber una interacción entre el tamaño de la casa y la clase de sistema de acondicionamiento. Esto se puede incorporar al modelo ampliando la ecuación (6.20) para incluir términos de interacción.

El modelo resultante es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D_1 + \beta_3 D_2 + \beta_4 D_3 + \beta_5 x_1 D_1 + \beta_6 x_1 D_2 + \beta_7 x_1 D_3 + \varepsilon \qquad \textbf{(6.21)}$$

Los cuatro modelos de regresión, que corresponden a las cuatro clases de sistema de acondicionamiento de aire son:

$y = \beta_0 + \beta_1 x_1 + \varepsilon$ (Sin acondicionamiento de aire)

$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_5) x_1 + \varepsilon$ (Unidades de ventanas)

$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_6) x_1 + \varepsilon$ (Bomba térmica)

$y = (\beta_0 + \beta_4) + (\beta_1 + \beta_7) x_1 + \varepsilon$ (Acondicionamiento central de aire)

Nótese que el modelo (6.21) implica que cada clase de sistema de acondicionamiento de aire puede tener una recta separada de regresión, con su pendiente y ordenada al origen correspondiente.

6.8 Comparación de Modelos de Regresión.

Se examinará el caso de la regresión lineal simple, en el que las n observaciones se pueden dividir en M grupos, y el m-ésimo grupo tiene n_m observaciones. El modelo más general consiste en M ecuaciones separadas, como por ejemplo:

Modelo Nº.	Modelo
1	$y = \beta_{01} + \beta_{11}x + \varepsilon$
2	$y = \beta_{01} + \beta_{11}x + \varepsilon$
⋮	⋮
M	$y = \beta_{0M} + \beta_{1M}x + \varepsilon$

O se puede escribir como:

$$y = \beta_{0m} + \beta_{1m}x + \varepsilon, \qquad m = 1, 2,. . ., M \qquad \textbf{(6.22)}$$

Con frecuencia interesa comparar este modelo general con uno más restrictivo; las variables cualitativas son útiles en este aspecto. Se consideran los siguientes casos:

a) Líneas Paralelas: En este caso todas las M pendientes son idénticas, $\beta_{11} = \beta_{12} = ... = \beta_{1M}$, pero las ordenadas al origen pueden ser distintas, nótese que esta es la clase de problema que se vio en el ejemplo 1 (en donde M = 2); condujo al uso de una variable indicadora. En forma más general se puede aplicar el método de la suma extra de cuadrados para probar la hipótesis $H_0 : \beta_{11} = \beta_{12} = ... = \beta_{1M}$. Recuérdese que este procedimiento implica ajustar un modelo completo y un modelo reducido restringido a la hipótesis nula, y calcular el estadístico F:

$$F_0 = \frac{[SS_{Res(MR)} - SS_{Res(MC)}]/(gl_{(MR)} - gl_{(MC)})}{SS_{Res(MC)}/gl_{(MC)}} \qquad \textbf{(6.23)}$$

Si el modelo reducido es tan satisfactorio como el modelo completo, entonces F_0 será pequeña en comparación con $F_{(\alpha,\ gl(MR)\ -\ gl(MC),\ gl\ (MC))}$. Los valores grandes de F_0 implican que el modelo reducido es inadecuado. Para ajustar el modelo completo (6.22) sólo se ajustan M ecuaciones separadas de regresión, a continuación se calcula $SS_{Res(MC)}$ sumando las sumas de cuadrados residuales obtenidas en cada regresión separada. Los grados de libertad $SS_{Res(MC)}$ son $gl_{MC} = \sum_{m=1}^{M}(n_m - 2) = n - 2M$. Para ajustar el modelo reducido se definen $M - 1$ variables indicadoras, D_1, D_2,…, D_{M-1} que corresponden a los M grupos, y entonces se ajusta:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 D_1 + \beta_3 D_2 + ... + \beta_M D_{M-1} + \varepsilon$$

La suma de cuadrados residuales de este modelo es $SS_{Res(MR)}$ con $gl_{(MR)}$ = n- k = n- (M + 1) grados de libertad donde k es el número de parámetros del modelo anterior.

Si la prueba F, ecuación (6.23) indica que los M modelos de regresión tienen una pendiente común, entonces $\hat{\beta}_1$ obtenida en el modelo reducido es un estimado de este parámetro, que se determina agrupando o combinando todos los datos, esto se mostró en el ejemplo 2.

En forma más general, el análisis de covarianza se usa para agrupar los datos, para estimar la pendiente común. En consecuencia, el análisis de covarianza es un tipo especial de modelo lineal, que es una combinación de un modelo de

regresión (con factores cuantitativos) con un modelo de análisis de varianza (con factores cualitativos).

b) Líneas Concurrentes: Las M ordenadas al origen son iguales $\beta_{01} = \beta_{02} = ... = \beta_{0M}$, pero las pendientes pueden ser distintas. El modelo reducido es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 Z_1 + \beta_3 Z_2 + ... + \beta_M Z_{M-1} + \varepsilon$$

En donde $Z_k = xD_k$, k = 1, 2,…, M – 1. La suma de cuadrados residuales de este modelo es $SS_{Res(MR)}$ y $gl_{(MR)}$ = n- (M + 1) grados de libertad, nótese que se está suponiendo la concurrencia en el origen.

c) Líneas Coincidentes: En este caso las M pendientes y las M ordenadas al origen son iguales, es decir $\beta_{01} = \beta_{02} = ... = \beta_{0M}$ y $\beta_{11} = \beta_{12} = ... = \beta_{1M}$. El modelo reducido es sólo:

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Y la suma de cuadrados residuales $SS_{Res(MR)}$ tiene $gl_{(MR)}$ = n - 2 grados de libertad. No son necesarias variables indicadoras en la prueba de coincidencia, pero se incluye este caso para completar la explicación.

6.9 Uso de las variables Dicótomas en el Análisis Estacional.

Muchas series de tiempos de las variables económicas basadas en información mensual o trimestral presentan patrones estacionales (movimiento oscilatorio regular). Algunos ejemplos de estas variables son: ventas de los almacenes en época de navidad,

demanda de dinero (saldos monetarios) de las familias en épocas de vacaciones, demanda por helados y bebidas durante el verano y precios de la cosecha cuando apenas termina la estación de la recolección. En ocasiones es conveniente eliminar el factor o "componente" estacional de las series de tiempo para poder prestar toda la atención a los demás factores, como por ejemplo, la tendencia[1]. El proceso de eliminación del componente estacional de una serie se conoce como la "desestacionalización" o el "ajuste estacional" y la serie resultante se denomina desestacionalizada o estacionalmente ajustada. Series económicas importantes tales como el índice de precios al consumidor, el índice de precios al por mayor, el índice de producción industrial, se publican en general ajustadas estacionalmente.

Existen varios métodos de desestacionalizar una serie, pero sólo nos ocuparemos de uno de ellos el llamado método de las variables dicótomas.

Ejemplo 4:

Si se desea ver como se usan las variables dicótomas para desestacionalizar una serie de tiempo podemos suponer que hacemos la regresión de las utilidades de empresas manufactureras de Estados Unidos contra las ventas en los periodos trimestrales de 1995 – 2000. La información pertinente, sin ajustes estacionales, se muestra en la tabla 6.3, la que también nos muestra como preparamos la matriz de información para incluir las variables dicótomas. Si observamos dicha información descubriremos un patrón

[1] La serie de tiempo puede tener cuatro componentes: estacional, cíclico, de tendencia y estrictamente aleatorio.

interesante. Tanto las utilidades como las ventas, son más altas en el segundo trimestre que en el primero o el tercero de cada año. Quizá el segundo trimestre presenta un efecto estacional. Para investigarlo hacemos lo siguiente:

$$\text{Utilidades}_t = \beta_0 + \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta x_t + \varepsilon_t \qquad \textbf{(6.24)}$$

Donde:

$$D_1 = \begin{cases} 1 \text{ para el segundo trimestre} \\ 0 \text{ para otro trimestre} \end{cases}$$

$$D_2 = \begin{cases} 1 \text{ para el tercer trimestre} \\ 0 \text{ para otro trimestre} \end{cases}$$

$$D_3 = \begin{cases} 1 \text{ para el cuarto trimestre} \\ 0 \text{ para otro trimestre} \end{cases}$$

Hay que indicar que suponemos que la variable "estación" tiene cuatro categorías, los cuatro trimestres del año, lo que requiere el uso de tres variables dicótomas. En estas condiciones si existe un patrón estacional en varios trimestres, los interceptos diferenciales, si son estadísticamente significativos, lo reflejará. Es posible que sólo algunos de estos interceptos diferenciales sean significativos estadísticamente lo que indica que sólo algunos trimestres reflejan la estacionalidad. El modelo (6.24) es un modelo general que se ajusta a todos los casos (recordemos, que se toma el primer trimestre del año como el de base).

Tabla 6.3 Matriz de datos para la regresión (6.24).

Año y trimestre	Ganancias (millones de $)	Ventas (millones de $)	D_1	D_2	D_3
1995 I	10503	114862	0	0	0
II	12092	123968	1	0	0
III	10834	121454	0	1	0
IV	12201	131917	0	0	1
1996 I	12245	129911	0	0	0
II	14001	140976	1	0	0
III	12213	137828	0	1	0
IV	12820	145465	0	0	1
1997 I	11349	136989	0	0	0
II	12615	145126	1	0	0
III	11014	141536	0	1	0
IV	12730	151776	0	0	1
1998 I	12539	148862	0	0	0
II	14849	158913	1	0	0
III	13203	155727	0	1	0
IV	14947	168409	0	0	1
1999 I	14151	162781	0	0	0
II	15949	176057	1	0	0
III	14024	172419	0	1	1
IV	14315	183327	0	0	1
2000 I	12381	170415	0	0	0
II	13991	181313	1	0	0
III	12174	176712	0	1	0
IV	10985	180370	0	0	1

Utilizando la información de la tabla 6.3, se obtienen los siguientes resultados:

$$\text{Utilidades}_t = 6899.346 + 1453.342\ D_{1t} - 167.405\ D_{2t} + 434.576\ D_{3t} + 0.036\ \text{Ventas}_t \quad \textbf{(6.25)}$$

Errores estándar de los coeficientes y los valores t son los siguientes:

$$es(\hat{\beta}_1) = 617.214 \qquad es(\hat{\beta}_3) = 588.337$$

$$es(\hat{\beta}_2) = 569.817 \qquad es(\hat{\beta}_4) = 0.012$$

$t_{\hat{\beta}_1} = 2.355$ $\qquad$ $t_{\hat{\beta}_3} = 0.739$

$t_{\hat{\beta}_2} = -0.294$ $\qquad$ $t_{\hat{\beta}_4} = 3.088$

$R^2 = 0.537$

Los resultados nos muestran que sólo el coeficiente de las ventas y el intercepto diferencial del segundo trimestre son significativos al nivel del 95% de confianza. Se puede entonces concluir que hay algún factor estacional en el segundo trimestre del año. El coeficiente de las ventas de 0.036 nos indica que después de tomar en cuenta el factor estacional, si las ventas aumentan en un dólar la utilidad promedio aumentará en aproximadamente 4 centavos.

En la formulación del modelo (6.24) se supuso que los trimestres se diferenciaban sólo en el intercepto siendo el coeficiente de las ventas el mismo para todos los trimestres.

De la regresión estimada (6.25) se pueden deducir las siguientes regresiones individuales:

Trimestres primero, tercero y cuarto:

$$\begin{aligned} E(y_t \mid x_t, D_1 = 0, D_2 = D_3 = 0) &= 6899.346 + 1453.342(0) + 0.036x_t \\ E(y_t \mid x_t, D_1 = 0, D_2 = D_3 = 0) &= 6899.346 + 0.036x_t \end{aligned} \qquad \textbf{(6.26)}$$

Segundo trimestre:

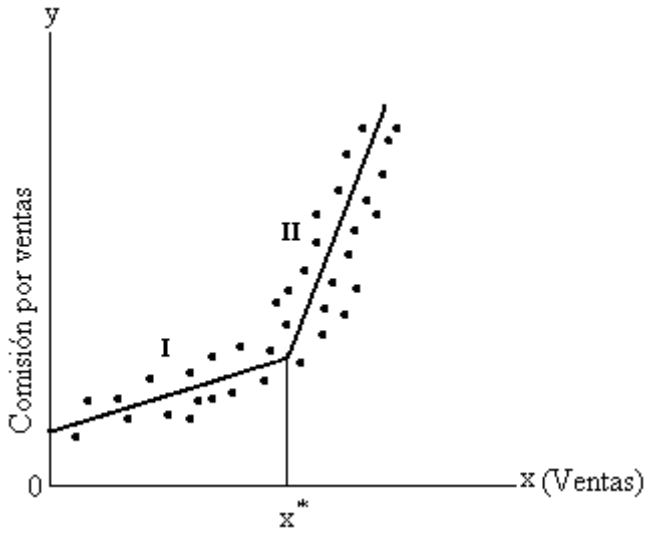
$$\begin{aligned} E(y_t \mid x_t, D_1 = 1, D_2 = D_3 = 0) &= 6899.346 + 1453.342(1) + 0.036x_t \\ E(y_t \mid x_t, D_1 = 1, D_2 = D_3 = 0) &= 8352.688 + 0.036x_t \end{aligned} \qquad \textbf{(6.27)}$$

De las ecuaciones (6.26) y (6.27) se puede observar que la utilidad promedio es mayor en el segundo trimestre que en el primero.

6.10 Regresión Lineal por Tramos.

Para ilustrar otro uso de las variables dicótomas, consideremos la figura 6.5 que nos muestra las remuneraciones percibidas por los representantes de ventas de una empresa hipotética. Dicha empresa paga comisiones por ventas de modo que hasta cierto nivel, denominado el objetivo o la meta, x^*, hay una estructura de comisiones, y por debajo de este nivel hay otra. Más específicamente, se supone que las comisiones aumentan linealmente con las ventas hasta el nivel objetivo x^*, después del cual aumentan también linealmente pero a una tasa más rápida. Se tiene entonces una regresión lineal por tramos que consiste en dos pedazos o segmentos que hemos denominado **I** y **II** en la figura 6.5. La función de comisiones por ventas cambia de pendiente en el valor del nivel objetivo x^*. El intercepto en el eje "y" denota la comisión mínima base.

Figura 6.5 Relación hipotética entre comisiones y volumen de ventas.



Con la información sobre las comisiones, ventas y el valor del nivel objetivo o meta x^*, la técnica de las variables dicótomas puede servir para estimar las diferentes pendientes de los segmentos de la regresión lineal por tramos presentada en la figura 6.5. El procedimiento es el siguiente:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - x^*) D_i + \varepsilon_i \qquad \textbf{(6.28)}$$

Donde:

y_i: Comisión por ventas.

x_i: Volumen de ventas realizado por el vendedor.

x^*: Valor objetivo de ventas (conocido de antemano).

$$D_i = \begin{cases} 1 \text{ si } x_i > x^* \\ 0 \text{ si } x_i < x^* \end{cases}$$

Suponiendo que $E(\varepsilon_i) = 0$, vemos enseguida que:

$$E(y_i \mid D_i = 0, x_i, x^*) = \beta_0 + \beta_1 x_i \qquad \textbf{(6.29)}$$

Que nos da las comisiones por ventas promedio hasta el nivel x^*, y

$$E(y_i \mid D_i = 1, x_i, x^*) = \beta_0 - \beta_2 x^* + (\beta_1 + \beta_2) x_i \qquad \textbf{(6.30)}$$

Que nos da las comisiones por ventas promedio, mas allá del nivel x^*.

De este modo, β_1 representa la pendiente de la línea de regresión en el segmento **I** y $\beta_1 + \beta_2$ representa la pendiente de la línea de regresión del segmento **II** de la regresión lineal por tramos de la figura 6.5. La hipótesis H_0 de que no hay "inflexión" en la regresión al nivel x^* puede llevarse a cabo examinando la significación estadística del coeficiente diferencial de la pendiente estimada $\hat{\beta}_2$.

Ejercicios 6

1. En la tabla siguiente se presenta una muestra de 20 estudiantes del curso de Estadística Aplicada a la Educación II del ciclo I 2008 de la UES-FMO, con la que se estudian las variables Peso, Estatura y Sexo. La variable sexo toma el valor de 1 si el estudiante es hombre y 0 si es mujer.

Peso (kg.)	Estatura (cm.)	Sexo	Peso (kg.)	Estatura (cm.)	Sexo
54.5	163	0	75.5	175	1
50	150	0	50	150	0
49.5	149	0	52	160	1
52	155	0	70.5	180	0
54	165	0	51	152	0
50	150	1	55	158	0
63	170	0	54.5	158	0
48	140	0	48	149	0
49	145	0	52	158	0
54	165	0	57	161	0

a) Estimar un modelo de regresión lineal que relacione el Peso “y” con la Estatura y el sexo del estudiante.

b) Realizar la prueba de hipótesis para el coeficiente de la variable sexo.

c) Construir un intervalo de confianza de 95% para el coeficiente de la variable Sexo.

d) Modificar el modelo desarrollado en la parte a), para incluir una interacción entre la variable Estatura y la variable Sexo.

e) Interpretar los parámetros de los modelos estimados en a) y d).

2. En la tabla siguiente se muestran los datos de rendimiento de gasolina en 32 automóviles, en la que "y" es el rendimiento de gasolina (millas/galón), "x" la cilíndrica del motor (pulgadas cúbicas), y D el tipo de transmisión (1 = automática, 0 = manual).

y (m/g)	x (p^3)	D	y (m/g)	x (p^3)	D
18.90	350	1	14.39	500.0	1
17.00	350	1	14.89	440.0	1
20.00	250	1	17.80	350.0	1
18.25	351	1	16.41	318.0	1
20.07	225	0	23.54	231.0	1
11.20	440	1	21.47	360.0	1
22.12	231	1	16.59	400.0	1
21.47	262	1	31.90	96.9	0
34.70	89.7	0	29.40	140.0	0
30.40	96.9	0	13.27	460.0	1
16.50	350	1	23.90	133.6	0
36.50	85.3	0	19.73	318.0	1
21.50	171	0	13.90	351.0	1
19.70	258	1	13.27	351.0	1
20.30	140	0	13.77	360.0	1
17.80	302	1	16.50	350.0	1

a) Formar un modelo de regresión lineal que relacione el rendimiento de la gasolina con la cilíndrica del motor y el tipo de transmisión. ¿afecta en forma importante el tipo de transmisión al rendimiento de la gasolina?.

b) Modificar el modelo desarrollado en la parte a), para incluir una interacción entre la cilíndrica del motor y el tipo de transmisión.

c) Realizar la prueba de hipótesis individual y global de los coeficientes de regresión. Estimar los intervalos de confianza del 95% de los parámetros.

3. La desestacionalización de cifras. El ejemplo 4 de la sección 6.9 señaló cómo las variables dicótomas pueden usarse para tomar en cuenta los efectos estacionales. Después de estimar la regresión (6.25) se encontró que solamente la variable dicótoma asociada al segundo trimestre del año era estadísticamente significativa, indicando que sólo este trimestre presentaba un patrón estacional. Por este motivo, un método de desestacionalizar la serie consiste en sustraer de los datos de utilidades y ventas, el segundo trimestre de cada año, la suma 1453.342 (millones de dólares), valor del coeficiente de la variable dicótoma para ese trimestre, y hacer la regresión de utilidades contra ventas mediante el empleo de la información transformada.

   a) Con la información dada en la tabla 6.3 hacer la regresión. No introducir ninguna variable dicótoma en esta regresión.

   b) Comparar el coeficiente de la variable ventas, en la regresión estimada en a) con el de la regresión (6.25). ¿Se espera que estos dos coeficientes sean estadísticamente iguales?.

4. Con los datos que se muestran en la tabla siguiente ajustar una regresión lineal por tramos, haciendo la regresión del costo total en dólares (y) de producción contra el producto (x) y la variable cualitativa D, que toma valores de 0, si $x_i > x^*$ y 1 si $x_i < x^*$ sabiendo además que la función de costo total cambia su pendiente para un nivel de producto de 5500 (x^*) unidades.

y ($)	256	414	634	778	1003	1839	2081	2423	2734	2814
x (u)	1000	2000	3000	4000	5000	6000	7000	8000	9000	10000
D	0	0	0	0	0	1	1	1	1	1

Capítulo 7

Extensiones del Modelo de Regresión y Violación de Supuestos.

7.1 Introducción.

Este Capítulo trata otros modelos de regresión como: modelos de regresión polinómicos, modelos de regresión no lineales y los modelos de regresión con variable cualitativa dependiente. El modelo de regresión polinomial permite aproximar relaciones no lineales de las variables, con lo que se amplia el modelo de regresión como herramienta muy poderosa para la investigación científica. Aunque los modelos polinómicos pueden verse como casos particulares del modelo de regresión múltiple, presenta ciertas peculiaridades que justifican su estudio independiente.

Hay muchos problemas donde es necesario utilizar algunas transformaciones para linealizar los datos. Además nos ocuparemos de los modelos de regresión en los cuales la variable dependiente es de naturaleza dicótoma, tomando los valores de 1 ó 0; y señalaremos algunos de los problemas de estimación que presenta.

También estudiaremos la violación de los supuestos básicos de la regresión; la Multicolinealidad que es la relación exacta entre las variables independientes, la Heteroscedasticidad que se da cuando la varianza de los residuos no es constante y la Autocorrelación que es cuando existe dependencia entre los residuos.

7.2 Definición de Términos Básicos.

Ad hoc: Es una expresión latina que significa literalmente "para esto". Generalmente se refiere a una solución elaborada específicamente para un problema o fin preciso y, por tanto, no es generalizable ni utilizable para otros propósitos. Se usa pues para referirse a algo que es adecuado sólo para un determinado fin. En sentido amplio, ad hoc puede traducirse como "específico" o "específicamente".

Autocorrelación: Es el hecho de que existen indicios de una fuerte relación (dependencia) lineal entre el término de error $\varepsilon_{t;}$ para un periodo de tiempo t y sus retardos (ε_{t-1}, ε_{t-2}) o adelantos (ε_{t+1}, ε_{t+2}).

Espuria: En estadística, una relación espuria (o, a veces, correlación espuria) es una relación matemática en la cual dos acontecimientos no tienen conexión lógica, aunque se puede implicar que la tienen debido a un tercer factor no considerado aún (llamado "factor de confusión" o "variable escondida").

Multicolinealidad Perfecta: Es cuando los coeficientes de regresión son indeterminados y sus desviaciones estándar infinitas, por lo tanto el modelo de regresión no puede ser estimado.

Multicolinealidad Menos Perfecta: Cuando los coeficientes de regresión aunque determinados o finitos, poseen errores estándar demasiado grandes, lo cual implica que los coeficientes no se pueden estimar con gran precisión o exactitud.

Regresión Curvilínea: Asociación entre dos variables que no es descrito por una línea por ejemplo la función exponencial, la función potencia, entre otras.

Regresión Polinómica: Es un tipo especial de regresión múltiple, donde aparecen como variables independientes una única variable y potencias de ésta (función cuadrática, función cúbica).

Transformaciones: Manipulación matemática para convertir una variable a una forma diferente, de modo que podamos ajustar curvas así como líneas rectas mediante regresión.

Transformación Lineal: Es un conjunto de operaciones que se realizan sobre un elemento de un sub-espacio, para transformarlo en un elemento de otro sub-espacio.

7.3 Modelos de Regresión Polinomial.

Los modelos de regresión polinomial más utilizados en la práctica son los de primer orden y los de segundo orden, en los capítulos anteriores se ha trabajado con el modelo de regresión polinomial de primer orden, es decir, con el modelo de regresión lineal como el siguiente:

Polinomio de primer orden o caso lineal:

$$\mathbf{y} = \mathbf{x}\beta + \boldsymbol{\varepsilon}$$

Que es un modelo general de ajuste de toda relación lineal en los parámetros desconocidos β y en las variables.

En esta sección estudiaremos los modelos de regresión polinomial de orden mayor que uno, como el siguiente:

Polinomio de segundo orden en una variable:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Y el polinomio de segundo orden de dos variables independientes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

Son modelos de regresión polinomial.

Los polinomios de orden mayor que 1 se usan mucho en casos en los que la respuesta es curvilínea (esto se puede observar a partir del diagrama de dispersión de los datos) y aun las relaciones no lineales complejas (por ejemplo: polinomios de orden mayor que 2) se pueden modelar en forma adecuada con polinomios dentro de límites razonablemente pequeños de las x_i.

7.3.1 Modelos Polinomiales en una Variable.

Como ejemplo de un modelo de regresión polinomial se considera el siguiente:

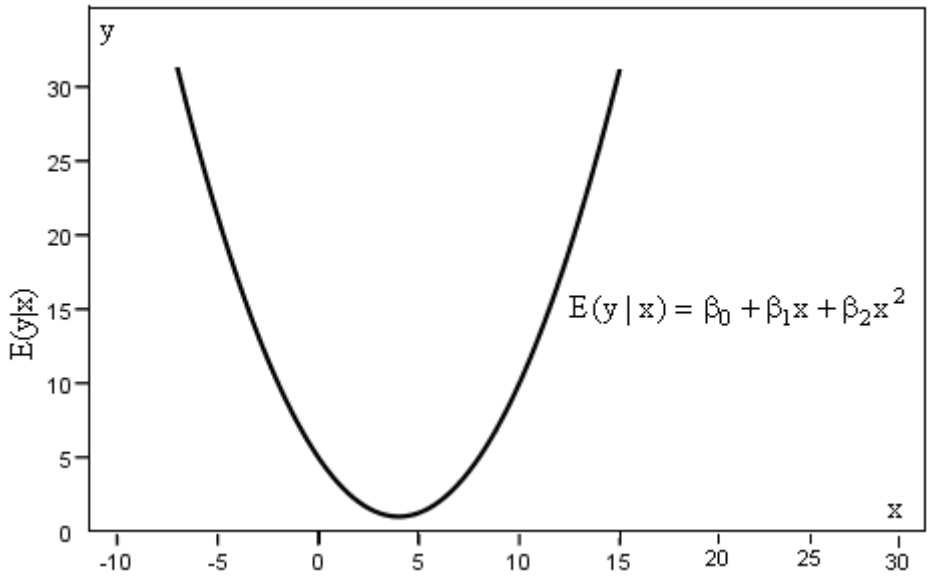
$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \qquad \textbf{(7.1)}$$

Este modelo se llama modelo de segundo orden en una variable. También a veces se llama modelo cuadrático, por que el valor esperado de “y” es:

$$E(y \mid x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Lo cual describe una función cuadrática. Un ejemplo típico se ve en la figura 7.1. Con frecuencia, a β_1 se le llama parámetro de efecto lineal y a β_2 parámetro de efecto cuadrático. El parámetro β_0 es el promedio de “y” cuando $x = 0$.

Figura 7.1 Ejemplo de polinomio cuadrático.



En general, el modelo polinomial de k-ésimo orden en una variable es:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon \qquad \textbf{(7.2)}$$

Si se define $x_j = x^j$, desde j = 1,…, k, la ecuación (7.2) se transforma en un modelo de regresión lineal múltiple con las k variables independientes x_1, x_2,…, x_k. Así, un modelo polinomial de orden k se puede ajustar con las técnicas que ya se estudiaron (MCO).

Los modelos polinomiales son útiles en casos cuando el investigador sabe (a través del diagrama de dispersión) que hay efectos curvilíneos presentes en la función verdadera de respuesta. También son útiles como funciones de aproximación a relaciones no lineales desconocidas y posiblemente muy complejas.

Hay varias consideraciones importantes que se presentan cuando se ajusta un polinomio de una variable. Algunas de ellas se describen a continuación:

1. **Orden del modelo:** Es importante mantener tan bajo como sea posible el orden del modelo. Cuando la función de respuesta parezca ser curvilínea se deben intentar transformaciones para mantener el modelo como de primer orden si fallan las transformaciones se debe intentar un polinomio de segundo orden. Como regla general, se debe evitar el uso de polinomios de orden superior ($k > 2$), a menos que se puedan justificar por razones ajenas a los datos. Un modelo de orden menor en una variable transformada casi siempre es preferible a un modelo de orden superior en la métrica original. El ajuste arbitrario (ilegal) de polinomios de orden superior es un grave abuso del análisis de regresión. Siempre se debe mantener un sentido de parsimonia, esto es, se debe usar el modelo más simple posible que sea consistente con los datos y el conocimiento del ambiente del problema. Recuérdese que en un caso extremo siempre es posible hacer pasar un polinomio de orden $n - 1$ por n puntos, por lo que siempre se puede encontrar un polinomio con grado suficientemente alto que produzca un ajuste ”bueno” con los datos. Ese modelo no contribuirá a mejorar el conocimiento de la función desconocida, ni es probable que sea un buen predictor.

2. **Estrategia para la construcción del modelo:** Se han sugerido diversas estrategias para elegir el orden de un polinomio de aproximación. Un método es ajustar en forma sucesiva modelos de orden creciente hasta que la prueba t para el término de orden máximo sea no significativa. Un procedimiento alterno es ajustar el modelo de orden máximo adecuado, y a continuación eliminar términos, uno por uno, comenzando por el de orden máximo hasta que el término que quede de orden máximo tenga una estadística t significativa. Esos dos procedimientos se llaman selección en avance y eliminación en reversa, respectivamente, no necesariamente conducen al mismo modelo. En vista del comentario del punto 1, se deben usar con cuidado esos procedimientos. En la mayor parte de los casos se debería restringir la atención a polinomios de primero y segundo orden.

3. **Extrapolación:** La extrapolación con modelos polinomiales puede ser peligrosa en extremo. En general, los modelos polinomiales pueden dirigirse hacia direcciones imprevistas e inadecuadas, tanto en la interpolación como en la extrapolación.

4. **Mal acondicionamiento I:** A medida que aumenta el orden del polinomio, la matriz $\mathbf{x'x}$ se vuelve mal acondicionada. Esto quiere decir que los cálculos de inversión de matrices serán inexactos y se puede introducir error considerable en los estimados de los parámetros. El mal acondicionamiento no esencial

causado por la elección arbitraria del origen, se puede eliminar centrando primero las variables independientes, es decir corregir "x" por su promedio $\bar{x}$.

5. **Mal acondicionamiento II:** Si los valores de "x" se limitan a un rango estrecho, puede haber mal acondicionamiento o multicolinealidad apreciables en las columnas de la matriz **x**. Por ejemplo si "x" varía entre 1 y 2, entonces x^2 varía entre 1 y 4, lo cual podría crear una fuerte multicolinealidad entre "x" y x^2

6. **Jerarquía:** El modelo de regresión:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

Se llama jerárquico por que contiene todos los términos de orden tres y menores. En cambio, el modelo:

$$y = \beta_0 + \beta_1 x + \beta_3 x^3 + \varepsilon$$

No es jerárquico porque no tiene el término $\beta_2 x^2$.

Lo mejor que se debe hacer es ajustar un modelo que contenga todos los términos significativos y usar el conocimiento de la disciplina más que una regla arbitraria, como guía adicional para formular el modelo.

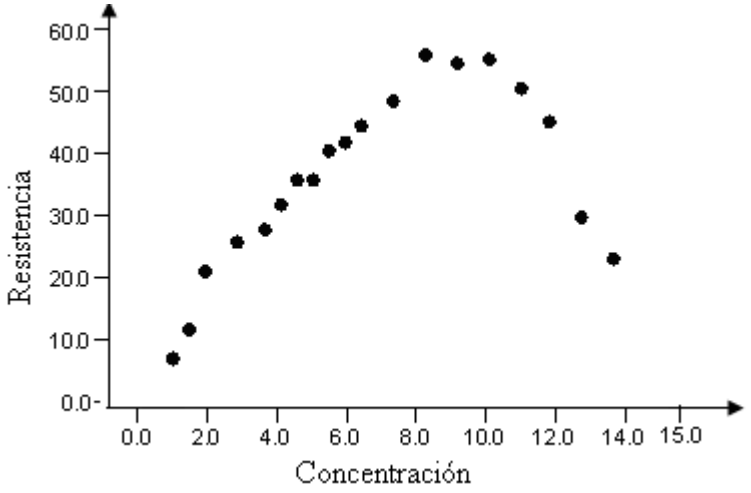
Ejemplo 1: Datos de madera dura.

La tabla 7.1 presenta datos a cerca de la resistencia del papel kraft y el porcentaje de madera dura en el lote de pulpa con el que se fabricó.

Tabla 7.1 Concentración de madera dura en la pulpa, y resistencia del papel kraft a la tensión.

Resistencia a la tensión (psi)	Concentración de madera dura (%)
6.3	1.0
11.1	1.5
20.0	2.0
24.0	3.0
26.1	4.0
30.0	4.5
33.8	5.0
34.0	5.5
38.1	6.0
39.9	6.5
42.0	7.0
46.1	8.0
53.1	9.0
52.0	10.0
52.5	11.0
48.0	12.0
42.8	13.0
27.8	14.0
21.9	15.0

Figura 7.2 Diagrama de dispersión del ejemplo 1.



En la figura 7.2 se ve el diagrama de dispersión para los datos del ejemplo 1. Esta presentación y el conocimiento del proceso de producción parecen indicar que un modelo cuadrático puede describir en forma adecuada la relación entre la resistencia a la tensión y la concentración de fibra corta (es decir, de madera dura). Si se adopta la recomendación de que al centrar los datos se puede eliminar el mal acondicionamiento no esencial, se ajustará el modelo:

$$y = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_2(x_i - \bar{x})^2 + \varepsilon$$

Para ello estimamos los parámetros de regresión haciendo uso de las ecuaciones siguientes, donde se puede observar que se ha sustituido la $\sum_{i=1}^{n} x_i$ por $\sum_{i=1}^{n}(x_i - \bar{x})$ con el propósito de eliminar el mal acondicionamiento no esencial:

$$\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x}) + \hat{\beta}_2\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0\sum_{i=1}^{n}(x_i - \bar{x}) + \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x})^2 + \hat{\beta}_2\sum_{i=1}^{n}(x_i - \bar{x})^3 &= \sum_{i=1}^{n}(x_i - \bar{x})y_i \\
\hat{\beta}_0\sum_{i=1}^{n}(x_i - \bar{x})^2 + \hat{\beta}_1\sum_{i=1}^{n}(x_i - \bar{x})^3 + \hat{\beta}_2\sum_{i=1}^{n}(x_i - \bar{x})^4 &= \sum_{i=1}^{n}(x_i - \bar{x})^2 y_i
\end{aligned}$$

$$\begin{aligned}
19\hat{\beta}_0 + 0\hat{\beta}_1 + 332.68\,\hat{\beta}_2 &= 649.5 \\
0\hat{\beta}_0 + 332.68\,\hat{\beta}_1 + 406.67\,\hat{\beta}_2 &= 589.15 \\
332.68\,\hat{\beta}_0 + 406.67\,\hat{\beta}_1 + 11439.95\,\hat{\beta}_2 &= 8844.73
\end{aligned}$$

Resolviendo el sistema de ecuaciones se encuentran los siguientes valores de los coeficientes de regresión:

$$\hat{\beta}_0 = 45.296$$

$$\hat{\beta}_1 = 2.546$$

$$\hat{\beta}_2 = -0.635$$

Así el modelo ajustado es:

$$\hat{y} = 45.296 + 2.546(x_i - 7.2632) - 0.635(x_i - 7.2632)^2$$

Prueba de hipótesis para los coeficientes de regresión.

Se probará la significancia global de la regresión polinomial para los datos de la Resistencia a la tensión y la Concentración de madera dura, es decir, $H_0 : \beta_1 = \beta_2 = 0$ y $H_1 : \beta_j \neq 0$, al menos para un j.

Datos:

$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2 = 3104.247$

$SS_{Res} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} = 312.638$

$SS_T = \mathbf{y}'\mathbf{y} - n\bar{y}^2 = 3416.885$

Solución:

1. $H_0 : \beta_1 = \beta_2 = 0$
2. $H_1 : \beta_j \neq 0$, al menos para un j.
3. Se selecciona un nivel de significancia de $\alpha = 0.05$ y se tiene que el valor de la tabla F es $F_{(0.05, 2, 16)} = 3.63$
4. Cálculos:

$$F_0 = \frac{\dfrac{\hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2}{k - 1}}{\dfrac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y}}{n - k}}$$

$$F_0 = \frac{\frac{3104.247}{3-1}}{\frac{312.638}{19-3}} = 79.434$$

En la tabla 7.2 se presenta el análisis de varianza para este modelo.

Tabla 7.2 Análisis de varianza para el modelo cuadrático del ejemplo 1.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F_0
Regresión	3104.247	2	1552.123	79.434
Residual	312.638	16	19.540	
Total	3416.885	18		

5. Decisión Estadística: se rechaza H_0, porque el valor calculado para F_0 (79.434) es mayor que el de la tabla (3.63).
6. Conclusión: Se concluye que el término lineal o el cuadrático (o ambos) contribuyen al modelo en forma significativa.

Las demás estadísticas de resumen para este modelo son: $R^2 = 0.9085$, el error estándar $es(\hat{\beta}_1) = 0.254$ y $es(\hat{\beta}_2) = 0.062$.

En la figura 7.3 se ve la gráfica de los residuos en función de $\hat{y}_i$. En ella no se ve inadecuación grave del modelo. En la figura 7.4 se muestra la gráfica de probabilidad normal de los residuos, en la que se puede observar que los puntos se aproximan a una recta; si la distribución de los residuos fuera normal todos los puntos estarían alineados formando una diagonal. Sin embargo, aún no se cuestiona seriamente la suposición de normalidad.

Figura 7.3 Gráfica de los residuos en función de los valores ajustados.

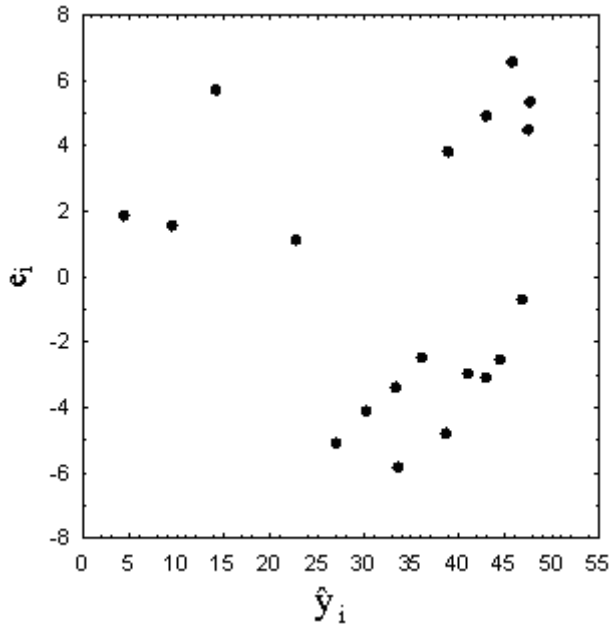
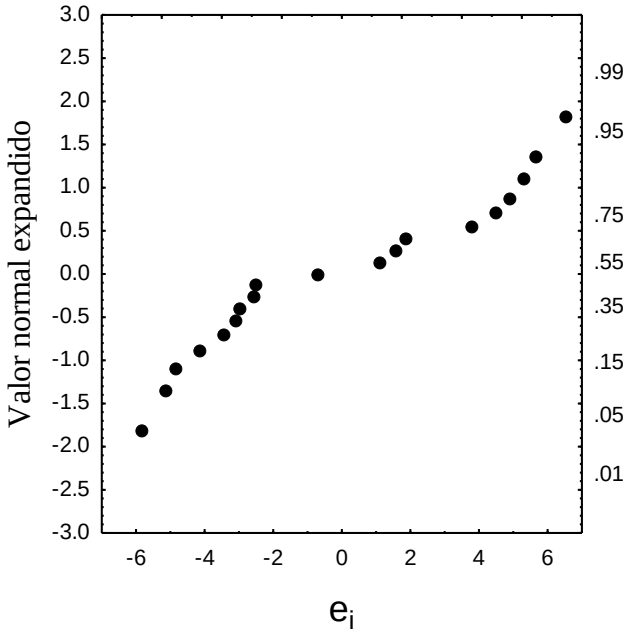


Figura 7.4 Gráfica de probabilidad normal de los residuos.



Ahora supóngase que se desea investigar la contribución del término cuadrático al modelo, esto es, se quiere probar:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Se probará esta hipótesis con el método de la suma extra de cuadrados dada en el Capítulo 6. Si $\beta_2 = 0$, el modelo reducido es la recta $y = \beta_0 + \beta_1(x_i - \bar{x}) + \varepsilon$. El ajuste por mínimos cuadrados es:

$$\hat{y} = 34.184 + 1.771(x_i - 7.2632)$$

Las estadísticas de resumen para este modelo son MS_{Res} = 139.615, R^2 = 0.3054, $es(\hat{\beta}_1) = 0.648$ y $SS_{RE}(\beta_1 \mid \beta_2) = 1043.427$. Se ve que al eliminar el término cuadrático se afectó R^2 drásticamente, así como el cuadrado medio residual (MS_{Res}) y $es(\hat{\beta}_1)$. Estas estadísticas de resumen son muy inferiores que las del modelo cuadrático. La suma extra de cuadrados para probar $H_0 : \beta_2 = 0$ es:

$SS_R(\beta_0, \beta_1, \beta_2) = \hat{\boldsymbol{\beta}}'\mathbf{x}'\mathbf{y} - n\bar{y}^2 = 3104.247$

$SS_{RE}(\beta_0, \beta_1) = \hat{\beta}_1 S_{xy} = 1043.427$

$$SS_{RE}(\beta_2 \mid \beta_0, \beta_1) = SS_R(\beta_0, \beta_1, \beta_2) - SS_{RE}(\beta_0, \beta_1)$$
$$SS_{RE}(\beta_2 \mid \beta_0, \beta_1) = 3104.247 - 1043.427$$
$$SS_{RE}(\beta_2 \mid \beta_0, \beta_1) = 2060.820$$

Con un grado de libertad la estadística F es:

$$F_0 = \frac{SS_R(\beta_2 \mid \beta_1, \beta_0)/1}{MS_{Res}} = \frac{2060.820/1}{19.540} = 105.47$$

Y como $F_{(0.05,\ 1,\ 16)} = 4.49$, se llega a la conclusión que $\beta_2 \neq 0$. Por lo anterior, el término cuadrático contribuye al modelo en forma significativa.

7.4 Modelos no Lineales y Transformaciones.

Los Capítulos anteriores han tratado de la creación de modelos de regresión en los cuales hay una ó más variables independientes. Además, se asume, a lo largo de la formulación del modelo, que tanto "x" como "y" entran al modelo en una forma lineal. Con frecuencia es aconsejable trabajar con un modelo alterno en el cual "x" o "y" (o ambas) entren en una forma no lineal. Puede indicarse una transformación de los datos debido a las consideraciones teóricas esenciales en el estudio científico, o una gráfica simple de los datos puede sugerir la necesidad de transformar las variables en el modelo. La necesidad de realizar una transformación es bastante simple de diagnosticar en el caso de regresión lineal simple debido a que las gráficas en dos dimensiones dan una imagen real de cómo entra cada variable en el modelo.

Un modelo en el cual "x" o "y" se ha transformado no debe considerarse como un modelo de regresión no lineal. Por lo general un modelo de regresión se considera como lineal cuando es lineal en los parámetros. En otras palabras. Supóngase que la naturaleza de los datos u otra información científica sugiere que se debe realizar la regresión y^* contra x^*, donde cada una es una transformación de las variables naturales "x" y "y". Entonces el modelo de la forma:

$$y_i^* = \alpha + \beta x_i^* + \varepsilon_i$$

Es un modelo lineal dado que es lineal en los parámetros α y β y el método de Mínimos Cuadrados Ordinarios permanece válido con y^* y x^* reemplazando a y_i y x_i.

Un ejemplo es el modelo log-log dado por:

$$\log y_i = \alpha + \beta \log x_i + \varepsilon_i$$

No obstante que este modelo no es lineal en "x" y "y", es lineal en los parámetros y es entonces considerado como un modelo lineal. Por otro lado, un ejemplo de un modelo no lineal verdadero está dado por:

$$y_i = \beta_0 + \beta_1 x^{\beta_2} + \varepsilon_i$$

Donde el parámetro β_2 (así como β_0 y β_1) debe estimarse. El modelo no es lineal en β_2.

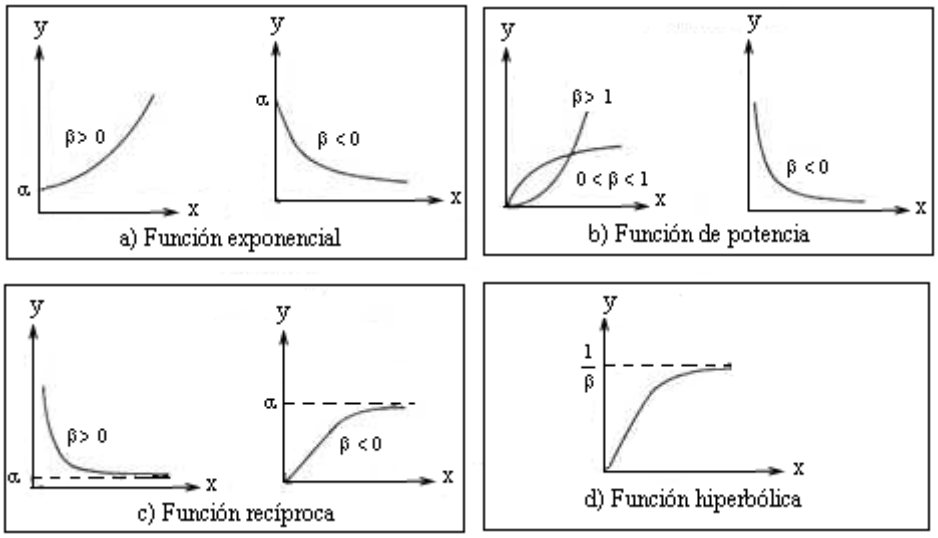
Las transformaciones que pueden mejorar el ajuste y el pronóstico son muchas. Aquí se tratan algunas de ellas y se presenta la gráfica que sirve como diagnóstico. En la tabla 7.3, se presentan algunas transformaciones. Las diferentes funciones que se dan representan las relaciones entre "x" y "y" que pueden producir una regresión lineal a lo largo de la transformación indicada. Además, se dan las variables dependientes e independientes para utilizarse en la regresión lineal simple resultante.

Tabla 7.3 Algunas transformaciones útiles para linealizar.

Forma funcional que relaciona "y" con "x".	Transformación apropiada	Forma de regresión lineal simple
Exponencial: $y = \alpha \ell^{\beta x}$	$y^* = \ln y$	Regresión y^* contra "x"
Potencia: $y = \alpha x^{\beta}$	$y^* = \log y;\ x^* = \log x$	Regresión y^* contra x^*
Recíproca: $y = \alpha + \beta\left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Regresión "y" contra x^*
Función hiperbólica: $y = \frac{x}{\alpha + \beta x}$	$y^* = \frac{1}{y};\ x^* = \frac{1}{x}$	Regresión y^* contra x^*

La figura 7.5 presenta las diferentes gráficas de las situaciones descritas en la tabla 7.3. Éstas sirven como una guía para que el investigador seleccione una gráfica de transformación de la observación de la curva de "y" contra "x".

Figura 7.5 Diagramas que muestran las funciones descritas en la tabla 7.3.



Lo anterior pretende ser una ayuda para el investigador cuando es aparente que una transformación proporcionará una mejora. Sin embargo, se deben considerar dos puntos importantes. El primero de ellos gira alrededor de la escritura formal del modelo una vez que los datos se transforman. Con bastante frecuencia el investigador no piensa nada al respecto, solamente realiza la transformación sin interesarse en la forma del modelo antes y después de la misma. El modelo exponencial sirve como un buen ejemplo. El modelo con las variables naturales (no transformadas) y que produce un modelo de error aditivo de las variables transformadas. Está dado por:

$$y_i = \alpha \ell^{\beta x_i} \cdot \varepsilon_i$$

El cual es un modelo de error multiplicativo. Resulta evidente que al tomar los logaritmos se produce:

$$\ln y_i = \ln \alpha + \beta x_i + \ln \varepsilon_i$$

Como resultado, las suposiciones básicas se realizan sobre $\ln \varepsilon_i$. El propósito de esta presentación es recordar, que no se debe considerar una transformación como solamente una manipulación algebraica con un error agregado. Con frecuencia un modelo de las variables transformadas que tiene una estructura de error aditivo es resultado de un modelo de las variables naturales con un tipo diferente de estructura de error.

El segundo punto importante es en relación con la noción de mediciones de mejora. Las mediciones obvias de comparación son, por supuesto, R^2 y el cuadrado medio residual, $\hat{\sigma}^2$. Ahora, si la respuesta "y" no se transforma, entonces es evidente que se pueden utilizar R^2 y $\hat{\sigma}^2$ para medir la utilidad de la transformación. Los residuos estarán en las

mismas unidades para ambos modelos transformados y no transformados. Pero cuando "y" se transforma, el criterio de comportamiento para el modelo transformado deberá basarse en los valores de los residuales en la métrica de la respuesta no transformada. De esta manera, las comparaciones que se realizan son adecuadas. El ejemplo que sigue proporciona una demostración clara de esto.

Ejemplo 2:

La presión P de un gas correspondiente a varios volúmenes V se registró de la siguiente manera:

Tabla 7.4 Datos de presión y volumen.

V (cm^3)	50	60	70	90	100
P (kg/cm^2)	64.7	51.3	40.5	25.9	7.8

La ley de los gases ideales está dada por la forma funcional $PV^{\gamma} = C$, donde γ y C son constantes. Estimar las constantes anteriores.

Solución: Se toman logaritmos naturales a ambos lados del modelo:

$$P_i V_i^{\gamma} = C \cdot \varepsilon_i, \quad i = 1, 2, 3, 4, 5.$$

Como resultado de aplicar logaritmo natural a la ecuación anterior, puede escribirse un modelo de regresión lineal:

$$\ln P_i = \ln C - \gamma \ln V_i + \varepsilon_i^*, \qquad i = 1, 2, 3, 4, 5.$$

Donde $\varepsilon_i^* = \ln \varepsilon_i$. Así se obtienen los resultados de una regresión lineal simple:

Intercepto: $\widehat{\ln C} = 14.7589739$, $\hat{C} = \exp(\widehat{\ln C}) = 2568862.88$

Pendiente: $\hat{\gamma} = -2.65347221$

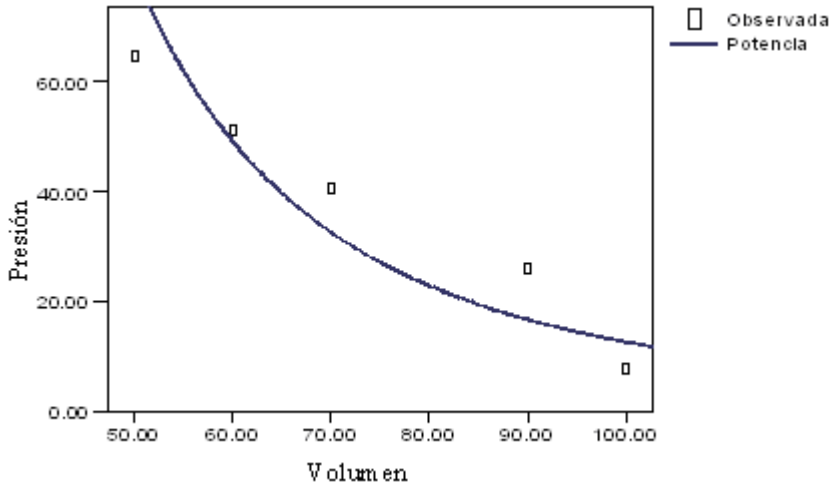
El modelo de regresión estimado es:

$$\widehat{\ln P_i} = 14.7589739 - 2.65347221 \ln V_i$$

Haciendo uso de la ecuación anterior se obtienen los siguientes resultados.

P_i	V_i	$\ln P_i$	$\ln V_i$	$\widehat{\ln P_i}$	$\hat{P}_i = \exp(\widehat{\ln P_i})$	$e_i = P_i - \hat{P}_i$
64.7	50	4.16976	3.91202	4.37853	79.721	-15.021
51.3	60	3.93769	4.09434	3.89474	49.143	2.157
40.5	70	3.70130	4.24850	3.48571	32.646	7.854
25.9	90	3.25424	4.49981	2.81885	16.758	9.142
7.8	100	2.05412	4.60517	2.53928	12.671	-4.871

Figura 7.6 Datos de presión y volumen, y regresión ajustada.



En la figura anterior se muestran los datos de la presión y el volumen no transformados, y la curva que representa la ecuación de regresión.

7.5 Regresión con Variable Dependiente Cualitativa.

Cuando una o más de las variables independientes en un modelo de regresión son dicótomas, podemos representarlas como variables indicadoras y proceder como se hizo en el Capítulo 6. Sin embargo es más compleja la aplicación del modelo de regresión lineal cuando la variable dependiente es dicótoma. Los modelos de elección binaria asumen que los individuos se enfrentan con una elección entre dos alternativas y que la elección depende de características identificables.

Supóngase, que vamos a estudiar la participación de los hombres adultos en la fuerza laboral como función de la tasa de desempleo, la tasa promedio de salarios, el ingreso familiar, la educación etc. En un momento determinado, una persona hace parte de la fuerza de trabajo o no lo hace. Por lo tanto, la variable dependiente puede tomar sólo dos valores: 1, si la persona hace parte de la fuerza de trabajo, y, 0 si no lo hace. Existen muchos ejemplos de este tipo, con variables dependientes dicótomas. Una familia, por ejemplo, tiene casa propia o no la tiene, ambos cónyuges están en el trabajo o sólo uno de ellos, etc. Lo único que tienen en común estos ejemplos es que la variable dependiente requiere una respuesta afirmativa o negativa: es decir, es dicótoma por naturaleza.

Para ver como se manejan estos modelos que tienen una variable dependiente dicótoma, consideremos el siguiente modelo:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \textbf{(7.3)}$$

Donde:

x_i: Ingreso familiar.

$$y_i = \begin{cases} 1; \text{si la familia posee casa propia} \\ 0; \text{si no la posee} \end{cases}$$

Modelos como el (7.3), que representa la variable dicótoma y_i como una función lineal de las variables explicatorias x_i, se denominan modelos lineales de probabilidad dado que $E(y_i \mid x_i)$, valor esperado condicional de y_i dado x_i, puede interpretarse como la probabilidad condicional de que el hecho ocurra, dado x_i. Es decir, $Pr(y_i = 1 \mid x_i)$. Por esto, en el caso anterior, $E(y_i \mid x_i)$ nos da la probabilidad de que una familia, cuyo ingreso es x_i, tenga casa. La justificación del nombre de modelo lineal de probabilidad para estos modelos (7.3) se puede explicar de la siguiente manera:

$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i \qquad \textbf{(7.4)}$$

Dado que y_i sólo puede tomar dos valores 1 y 0, podemos escribir la distribución de probabilidad de "y" suponiendo que: $P_i = Pr(y_i = 1 \mid x_i)$ es decir, de que el evento ocurra dada x_i y $1 - P_i = Pr(y_i = 0 \mid x_i)$ es decir, de que el evento no ocurra dada x_i, la variable y_i tiene la siguiente distribución:

y_i	Probabilidad
0	$1 - P_i$
1	P_i
	1

Entonces, por la definición de esperanza matemática obtenemos:

$$\begin{aligned} E(y_i \mid x_i) &= 0 * Pr(y_i = 0 \mid x_i) + 1 * Pr(y_i = 1 \mid x_i) \\ E(y_i \mid x_i) &= 1 * Pr(y_i = 1 \mid x_i) \\ E(y_i \mid x_i) &= P_i \end{aligned} \qquad \textbf{(7.5)}$$

Comparando la ecuación (7.4) con la (7.5), podemos igualar

$$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i = P_i \qquad \textbf{(7.6)}$$

Esto es, la esperanza condicional del modelo (7.3) puede efectivamente interpretarse como la probabilidad condicional de y_i.

Dado que P_i debe estar entre 0 y 1 inclusive, podemos dar la restricción siguiente:

$$0 \leq E(y_i \mid x_i) \leq 1$$

Es decir, la expectativa condicional, o probabilidad condicional, debe estar entre 0 y 1.

7.5.1 Estimación de Modelos Lineales de Probabilidad.

A primera vista parece que el modelo (7.3) es como cualquier otro modelo de regresión ya que sus parámetros pueden estimarse por el método de MCO. No obstante, examinaremos a continuación algunos problemas que se presentan:

1. Normalidad del error ε_i. Aunque el método de MCO no requiere que los errores estén normalmente distribuidos, hemos supuesto que lo están con fines de inferencia estadística, es decir, para la prueba de hipótesis. Sin embargo, el supuesto de normalidad de los ε_i no es válido para los modelos lineales de probabilidad pues como ocurre en los y_i, ε_i toma sólo dos valores. Para aclarar este punto escribiremos (7.3) como:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i \qquad \textbf{(7.7)}$$

Ahora, cuando $y_i = 1$ $\varepsilon_i = 1 - \beta_0 - \beta_1 x_i$

Y cuando

$$y_i = 0 \qquad \varepsilon_i = -\beta_0 - \beta_1 x_i \qquad \textbf{(7.8)}$$

No podemos suponer que los ε_i están normalmente distribuidos.

No obstante el hecho de no cumplir con el supuesto de normalidad no es tan crítico como parece pues, como sabemos, las estimaciones puntuales de MCO siguen siendo insesgadas (recuerde que si el objetivo es la estimación puntual, el supuesto de normalidad no tiene importancia). Además, a medida que aumenta el tamaño de la muestra, se puede demostrar que los estimadores de MCO tienden por lo general a estar normalmente distribuidos. Por lo tanto, en muestras grandes, la inferencia estadística de los modelos lineales de probabilidad seguirá el procedimiento usual de MCO bajo las condiciones de normalidad.

2. Varianzas heteroscedásticas de los errores. Aunque $E(\varepsilon_i) = 0$ y $E(\varepsilon_i \varepsilon_j) = 0$, para $i \neq j$, no se cumple el hecho de que los errores sean homoscedásticos. Para verlo más claramente, los ε_i dados en la ecuación (7.8) tienen la siguiente distribución de probabilidad:

ε_i	Probabilidad
$-\beta_0 - \beta_1 x_i$	$1-P_i$
$1-\beta_0 - \beta_1 x_i$	P_i
	1

Esta distribución de probabilidad se desprende de la distribución de probabilidad para y_i dada previamente.

Por definición,

$$\text{var}(\varepsilon_i) = E[\varepsilon_i - E(\varepsilon_i)]^2$$
$$\text{var}(\varepsilon_i) = E(\varepsilon_i)^2, \quad \text{para} \quad E(\varepsilon_i) = 0 \ \text{por presunción}$$

Por lo tanto, usando esta distribución de probabilidad de ε_i, obtenemos:

$$\text{var}(\varepsilon_i) = E(\varepsilon_i)^2$$
$$\text{var}(\varepsilon_i) = (-\beta_0 - \beta_1 x_i)^2 (1 - P_i) + (1 - \beta_0 - \beta_1 x_i)^2 (P_i)$$
$$\text{var}(\varepsilon_i) = (-\beta_0 - \beta_1 x_i)^2 (1 - \beta_0 - \beta_1 x_i) + (1 - \beta_0 - \beta_1 x_i)^2 (\beta_0 + \beta_1 x_i)$$
$$\text{var}(\varepsilon_i) = (\beta_0 + \beta_1 x_i)^2 (1 - \beta_0 - \beta_1 x_i) + (1 - \beta_0 - \beta_1 x_i)^2 (\beta_0 + \beta_1 x_i)$$

$$\text{var}(\varepsilon_i) = (\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i) \qquad \textbf{(7.9)}$$

o

$$\text{var}(\varepsilon_i) = E(y_i \mid x_i)[1 - E(y_i \mid x_i)]$$
$$\text{var}(\varepsilon_i) = P_i(1 - P_i) \qquad \textbf{(7.10)}$$

De donde se tiene en cuenta el hecho de que $E(y_i \mid x_i) = \beta_0 + \beta_1 x_i - P_i$. La ecuación (7.10) muestra que la varianza de ε_i es heteroscedástica porque depende de la esperanza condicional de "y", que depende naturalmente, del valor que tome "x". En último término, la varianza de ε_i depende de "x" y por lo tanto no es homoscedástica.

Sabemos que en presencia de heteroscedásticidad, los estimadores de MCO aunque sean insesgados no son eficientes; es decir, no tienen varianza mínima. Pero tampoco en este caso el problema de la heteroscedásticidad es grave, en

la sección 7.7 se discutirán varios métodos para manejar la heteroscedasticidad. Dado que la varianza de ε_i depende del valor esperado de "y" condicional en "x", como se vio en la ecuación (7.9), una forma de resolver el problema de la heteroscedasticidad consiste en transformar la información dividiendo ambos lados del modelo (7.3) por:

$$\sqrt{E(y_i \mid x_i)[1-E(y_i \mid x_i)]} = \sqrt{P_i(1-P_i)} = \sqrt{w_i}$$

$$\frac{y_i}{\sqrt{w_i}} = \frac{\beta_0}{\sqrt{w_i}} + \frac{\beta_1 x_i}{\sqrt{w_i}} + \frac{\varepsilon_i}{\sqrt{w_i}} \qquad \textbf{(7.11)}$$

Podemos seguir entonces con la estimación por MCO de (7.11).

Naturalmente, el verdadero $E(y_i \mid x_i)$ no se conoce por lo tanto $\sqrt{w_i}$ tampoco se conoce.

Para estimar $\sqrt{w_i}$ podemos usar el siguiente procedimiento en dos etapas:

Etapa I: Correr la regresión (7.3) por MCO a pesar del problema de la heteroscedásticidad y obtener $\hat{y}_i$ = estimación del verdadero $E(y_i \mid x_i)$. Luego, obtenga $\hat{w}_i = \hat{y}_i(1-\hat{y}_i)$, la estimación de w_i.

Etapa II: Utilice el estimador w_i para transformar la información como en la ecuación (7.11), y corra la regresión con los datos transformados por MCO.

3. Si no se cumple $0 \le E(y_i \mid x_i) \le 1$. Dado que $E(y_i \mid x_i)$ en los modelos lineales de probabilidad mide la probabilidad condicional de que ocurra el evento "y" dado "x", necesariamente estará comprendido entre 0 y 1. Aunque esto es verdad, no se puede garantizar que $\hat{y}_i$, los estimadores de $E(y_i \mid x_i)$,

cumplan necesariamente esta restricción, lo que constituye el mayor problema de la estimación de MCO de los modelos lineales de probabilidad. Existen dos métodos para saber si los estimadores $\hat{y}_i$ están efectivamente entre 0 y 1.

a) El primero consiste en estimar el modelo lineal de probabilidad por el método de MCO y ver si los $\hat{y}_i$ estimados se encuentran entre 0 y 1. si algunos son menores que cero (es decir negativos), se supone que para estos casos el $\hat{y}_i$ es cero; si son mayores que 1, se suponen iguales a 1.

b) El segundo procedimiento es el de diseñar una técnica de estimación que nos garantice que las probabilidades condicionales estimadas $\hat{y}_i$ estén entre 0 y 1.

7.6 Multicolinealidad.

Uno de los supuestos del modelo de regresión lineal clásico es el de que no existe multicolinealidad entre las variables independientes incluidas en el. En esta sección se tratará de examinar más detenidamente este supuesto.

En el modelo de regresión múltiple la estimación del efecto de una variable depende de su efecto diferencial, es decir, la parte de la variable que no está relacionada linealmente con las demás variables incluidas en el modelo. Si una variable independiente está relacionada exactamente con las restantes, entonces no disponemos de información libre sobre ella y, por tanto, no es posible estimar sus efectos. Este es el problema de multicolinealidad.

Se dice que existe una relación lineal exacta si se satisface la siguiente condición:

$$\lambda_1 x_1 + \lambda_2 x_2 + ... + \lambda_k x_k = 0 \qquad \textbf{(7.12)}$$

Donde λ_1, λ_2,..., λ_k son constantes, sin que todas ellas sean simultáneamente 0.

Sin embargo ahora, el término multicolinealidad se utiliza en un sentido más amplio con el fin de incluir el caso de la multicolinealidad perfecta, como se muestra en la ecuación (7.12) así como en el caso donde las variables "x" están intercorrelacionadas pero no perfectamente si no en la forma[1]:

$$\lambda_1 x_1 + \lambda_2 x_2 + ... + \lambda_k x_k + v_i = 0 \qquad \textbf{(7.13)}$$

Donde v_i es un término estocástico de error.

Para ver la diferencia entre la multicolinealidad perfecta y la menos perfecta, supongamos, por ejemplo, que $\lambda_2 \neq 0$. Entonces, (7.12) puede escribirse como:

$$x_{2i} = \frac{\lambda_1}{\lambda_2} x_{1i} - \frac{\lambda_3}{\lambda_2} x_{3i} - ... - \frac{\lambda_k}{\lambda_2} x_{ki} \qquad \textbf{(7.14)}$$

Que muestra como x_2 está exactamente relacionada de manera lineal con las otras variables o como puede ser derivada de una combinación lineal de las otras variables "x". En situaciones como esta, el coeficiente de correlación entre la variable x_2 y la combinación lineal del lado derecho de la ecuación (7.14) debe ser igual a la unidad.

Igualmente, si $\lambda_2 \neq 0$, la ecuación (7.13) puede reescribirse como:

$$x_{2i} = \frac{\lambda_1}{\lambda_2} x_{1i} - \frac{\lambda_3}{\lambda_2} x_{3i} - ... - \frac{\lambda_k}{\lambda_2} x_{ki} - \frac{1}{\lambda_2} v_i \qquad \textbf{(7.15)}$$

[1] Si hay sólo dos variables explicatorias, la intercorrelación puede medirse por el coeficiente de correlación de orden cero o por el simple. Pero si hay más de dos variables "x", la intercorrelación puede medirse por el coeficiente de correlación parcial o por el coeficiente de correlación múltiple R de una variable "x" contra todas las otras "x" variables agrupadas.

Que muestra como x_2 no es una combinación lineal exacta de las otras "x" sino que está también determinada por el error estocástico v_i.

Como ejemplo, se consideran las siguientes cifras:

x_2	x_3	x_3^*
2	10	12
4	20	20
6	30	37
8	40	49
10	50	52

Se puede notar que $x_{3i} = 5x_{2i}$; por lo tanto, hay perfecta colinealidad entre x_2 y x_3 puesto que el coeficiente de correlación $r_{23} = 1$. La variable x_3^* fue creada a partir de x_3 simplemente agregándole a esta última las siguientes cifras tomadas de una tabla de números aleatorios: 2, 0, 7, 9, 2; en esta forma, no hay ya perfecta colinealidad entre x_2 y x_3^*. Sin embargo, las dos variables están altamente correlacionadas como lo muestra el cálculo del coeficiente de correlación entre ellas que es de 0.992.

Obsérvese que la multicolinealidad, como la acabamos de definir, hace referencia sólo a la relación lineal entre las variables "x", dejando por fuera las relaciones no lineales; por ejemplo, si consideramos el siguiente modelo de regresión:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i \qquad \textbf{(7.16)}$$

Donde:

y_i: Costo total de producción.

x_i: Producción.

x_i^2: Producción al cuadrado.

x_i^3: Producción al cubo.

Las variables x_i^2 y x_i^3 están funcionalmente relacionadas con x_i, aunque es claro que la relación no es lineal. Por consiguiente, los modelos del tipo (7.16) no violan el supuesto de la no multicolinealidad; en efecto, para describir las curvas de costos medios y marginales en forma de U el modelo (7.16) es muy apropiado.

¿Por qué en el modelo de regresión lineal clásico se supone que no hay multicolinealidad entre las "x"? la razón es que: Si la multicolinealidad es perfecta en el sentido de (7.12) los coeficientes de regresión de las variables "x" son indeterminados y sus errores estándar infinitos. Si la multicolinealidad es menos perfecta como en (7.13), los coeficientes de regresión aunque determinados poseen grandes errores estándar (en relación a los propios coeficientes) lo que significa que los coeficientes no se pueden estimar con gran precisión.

Debe enfatizarse, que, si las "x" se suponen fijas o no estocásticas, la multicolinealidad es esencialmente un fenómeno muestral (de regresión)[2] . Cuando postulamos la función de regresión poblacional o teórica (FRP), dijimos que todas las

[2] Si hay razón para pensar que las variables "x" son estocásticas y que en la población están relacionadas linealmente, debemos desarrollar nuestra FRP teniendo esto en cuenta. Lo que afirmamos es que aunque las "x" no estén relacionadas en la población, pueden estarlo en la muestra. En este sentido, la multicolinealidad es un fenómeno muestral.

variables “x” incluidas en el modelo tienen una influencia separada o independiente sobre “y”. Pero puede suceder que en una muestra utilizada para verificar la FRP, algunas o todas las variables “x” sean tan altamente colineales que no podamos aislar su influencia sobre “y”. Por así decirlo, nuestra muestra nos falla aunque la teoría nos diga que todas las “x” son importantes. En resumen, nuestra muestra puede no ser lo suficientemente significativa como para acomodar todas las variables “x” en el análisis.

Tomando el ejemplo de gastos e ingresos del Capitulo 1 podemos suponer que fuera del ingreso, la riqueza es otro determinante valioso en los gastos de consumo, lo cual nos permite escribir:

$$\text{Gastos}_i = \beta_0 + \beta_1 \text{ Ingreso}_i + \beta_2 \text{ Riqueza}_i + \varepsilon_i$$

Puede suceder ahora que al obtener cifras de ingreso y riqueza, las dos variables pueden ser altamente, o incluso perfectamente, correlacionadas, pues las personas ricas tienden a tener ingresos más altos. De este modo aunque en teoría el ingreso y la riqueza son razones lógicas para explicar el comportamiento de los gastos de consumo, en la práctica (por ejemplo), en la muestra puede ser difícil separar la influencia del ingreso y la riqueza sobre el consumo.

7.6.1 Estimación en el caso de la Multicolinealidad Perfecta.

Ya se estableció que en el caso de multicolinealidad perfecta los coeficientes de regresión son indeterminados y que sus errores estándar son infinitos. Por ejemplo para el modelo de tres variables tenemos:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + e_i \qquad \textbf{(7.17)}$$

Teniendo en cuenta las ecuaciones (4.15), (4.16) y (4.17) del Capítulo 4 y suponiendo que $(x_{2i} - \bar{x}_2) = \lambda(x_{1i} - \bar{x}_1)$, donde $\lambda \neq 0$. Reemplazando esto en la ecuación (4.16) obtendremos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y})\sum_{i=1}^{n}(\lambda(x_{1i} - \bar{x}_1))^2 - \sum_{i=1}^{n}(\lambda(x_{1i} - \bar{x}_1))(y_i - \bar{y})\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(\lambda(x_{1i} - \bar{x}_1))}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2\sum_{i=1}^{n}(\lambda(x_{1i} - \bar{x}_1))^2 - \left(\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(\lambda(x_{1i} - \bar{x}_1))\right)^2}$$

$$\hat{\beta}_1 = \frac{\lambda^2\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y})\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 - \lambda^2\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y})\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}{\lambda^2\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 - \lambda^2\left(\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1)\right)^2}$$

$$\hat{\beta}_1 = \frac{\lambda^2\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y})\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 - \lambda^2\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(y_i - \bar{y})\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}{\lambda^2\left(\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2\right)^2 - \lambda^2\left(\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2\right)^2}$$

$$\hat{\beta}_1 = \frac{0}{0} \qquad \textbf{(7.18)}$$

Que es una expresión indeterminada, de forma similar se puede verificar que $\hat{\beta}_2$ es también indeterminada.

¿Por qué se obtiene el resultado que se muestra en la ecuación (7.18)? Recordemos el significado de $\hat{\beta}_1$: que nos da la tasa de variación promedio de "y" cuando x_1 cambia en una unidad, manteniendo x_2 constante. Sin embargo, si x_1 y x_2 son perfectamente colineales, no hay manera de que se mantenga x_2 constante: a medida que x_1 cambia también x_2 cambia en el factor λ, lo anterior significa que no existe un medio de extraer las influencias separadas de x_1 y x_2 a partir de la muestra dada.

Volviendo a las varianzas dadas en las ecuaciones (4.20) y (4.22) del Capítulo 4 y reemplazando $(x_{2i} - \bar{x}_2) = \lambda(x_{1i} - \bar{x}_1)$ en la ecuación (4.20) obtenemos:

$$\text{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(\lambda(x_{1i} - \bar{x}_1))^2}{\left(\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2\right)\left(\sum_{i=1}^{n}(\lambda(x_{1i} - \bar{x}_1))^2\right) - \left(\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)(\lambda(x_{1i} - \bar{x}_1))\right)^2} * \sigma^2$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{0}$$

$$\text{var}(\hat{\beta}_1) = \infty \qquad \textbf{(7.19)}$$

La $\text{var}(\hat{\beta}_2) = \infty$. De este modo, las varianzas tanto de $\hat{\beta}_1$ y $\hat{\beta}_2$ son indefinidas y por lo tanto "infinitas", y sus errores estándar son también indefinidos e infinitos.

7.6.2 Estimación en caso de Multicolinealidad Alta pero Imperfecta.

La situación de perfecta multicolinealidad es bastante extrema, generalmente no existen relaciones lineales exactas entre las variables "x", en especial para cifras de series de tiempo.

En esta forma, en cuanto al modelo de tres variables (7.16), en lugar de multicolinealidad exacta podemos tener más bien:

$$(x_{2i} - \bar{x}_2) = \lambda(x_{1i} - \bar{x}_1) + v_i \qquad \textbf{(7.20)}$$

Donde $\lambda \neq 0$, y donde v_i es un término que capta el error estocástico de modo que

$\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)v_i=0$. En este caso puede ser posible la estimación de los coeficientes de regresión β_1 y β_2. Por ejemplo reemplazando (7.20) en (4.16) tendremos:

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(y_i-\bar{y})\sum_{i=1}^{n}(\lambda(x_{1i}-\bar{x}_1)+v_i)^2-\sum_{i=1}^{n}(\lambda(x_{1i}-\bar{x}_1)+v_i)(y_i-\bar{y})\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(\lambda(x_{1i}-\bar{x}_1)+v_i)}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2\sum_{i=1}^{n}(\lambda(x_{1i}-\bar{x}_1)+v_i)^2-\left(\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(\lambda(x_{1i}-\bar{x}_1)+v_i)\right)^2}$$

$$\hat{\beta}_1=\frac{\left(\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(y_i-\bar{y})\right)\left(\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2+\sum_{i=1}^{n}v_i^2\right)-\left(\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(y_i-\bar{y})+\sum_{i=1}^{n}(y_i-\bar{y})v_i\right)+\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2\left(\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2+\sum_{i=1}^{n}v_i^2\right)-\left(\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2\right)^2} \quad \textbf{(7.21)}$$

Donde hemos aprovechado que $\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)v_i=0$. Una expresión similar puede derivarse para $\hat{\beta}_2$. Ahora no existen razones a priori para creer que (7.21) no puede estimarse. Desde luego, si v_i es lo suficientemente pequeña, digamos muy cercano a cero, (7.20) indicará casi perfecta colinealidad volviendo al caso indeterminado (7.18).

Si las varianzas de $\hat{\beta}_1$ y $\hat{\beta}_2$ se definen de la forma siguiente:

$$\text{var}(\hat{\beta}_1)=\frac{\sigma^2}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2(1-r_{12}^2)} \quad \textbf{(7.22)}$$

$$\text{var}(\hat{\beta}_2)=\frac{\sigma^2}{\sum_{i=1}^{n}(x_{2i}-\bar{x}_2)^2(1-r_{12}^2)} \quad \textbf{(7.23)}$$

Donde r_{12} es el coeficiente de correlación entre x_1 y x_2 de las ecuaciones (7.22) y (7.23) se puede ver que si r_{12} tiende a 1, es decir, a medida que la colinealidad aumenta, las

varianzas de los estimadores aumentan y en el limite, cuando $r_{12} = 1$, se vuelven infinitas.

7.6.3 Consecuencias de la Multicolinealidad.

Propiedades de los estimadores de MCO.

Tengamos en cuenta que si los supuestos de los modelos de regresión lineal clásico se cumplen, los estimadores de MCO de los coeficientes de regresión lineal serán lineales insesgados y con varianzas mínimas; en pocas palabras, son los mejores estimadores lineales insesgados. Ahora bien, si la multicolinealidad es alta los estimadores de MCO siguen siendo los mejores estimadores lineales insesgados aunque es necesario considerar lo siguiente:

Ser insesgado es una propiedad multimuestral o de muestras repetidas que dice que manteniendo fijos los valores de la variable "x", si se toman muestras repetidas y se calculan los estimadores de MCO, para cada una de estas muestras, el promedio de los valores muestrales, convergerá al verdadero valor poblacional de los estimadores, a medida que el número de muestras aumenta. Sin embargo, esto no se refiere a las propiedades de los estimadores en una muestra dada.

Es verdad que la colinealidad no destruye la propiedad de varianza mínima; en efecto, dentro de la clase de estimadores lineales insesgados, los estimadores de MCO tienen varianza mínima, es decir, son eficientes. Aunque esto no quiere decir que la varianza de un estimador de MCO sea necesariamente pequeña (con relación al valor del estimador) en una muestra dada.

La multicolinealidad es un fenómeno esencialmente muestral. Por consiguiente, el hecho de que los estimadores de MCO sean los mejores estimadores lineales insesgados es, en la práctica, de poco valor. Veremos entonces que pasa o que puede pasar en una muestra cualquiera.

Consecuencias prácticas de la multicolinealidad.

Como se mostró en la sección 7.6.1, en el caso de perfecta multicolinealidad los estimadores de MCO son indeterminados y sus varianzas y errores estándar son indefinidos. Si por el contrario hay colinealidad severa, aunque no perfecta, las consecuencias serán las siguientes:

1. Aunque los estimadores de MCO son obtenibles, sus errores estándar tienden a ser mayores a medida que aumenta el grado de colinealidad entre las variables. Esto se mostró en la sección 7.6.2 para el caso de tres variables.
2. Debido al gran tamaño de los errores estándar, los intervalos de confianza para los parámetros poblacionales relevantes (β_j, para j = 1, 2,…, k) tienden a ser grandes. Así, en el caso de tres variables si no hay colinealidad ($r_{12} = 0$) y suponiendo que conocemos a σ^2 el intervalo de confianza del 95%, para β_1 puede obtenerse como[3]:

$$es(\hat{\beta}_1) = \sqrt{var(\hat{\beta}_1)} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2(1-0)}} = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}}$$

[3] Nótese que estamos utilizando la distribución normal en razón de que por conveniencia suponemos conocer a σ^2.

$$P(\hat{\beta}_1 - 1.96\,es(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + 1.96\,es(\hat{\beta}_1)) = 0.95 \qquad \textbf{(7.24)}$$

3. En virtud del punto dos para casos con alta multicolinealidad, las cifras muestrales pueden ser compatibles con un conjunto de diversas hipótesis, por lo que la probabilidad de aceptar una hipótesis falsa (de error tipo II) aumenta.
4. Si la multicolinealidad no es perfecta, es posible la estimación de los coeficientes de regresión pero los estimadores y sus errores estándar se vuelven muy sensibles incluso con mínimos cambios en las cifras. Para ver esto consideremos la tabla 7.5.

Tabla 7.5 Cifras hipotéticas de "y", x_1 y x_2.

y	x_1	x_2
1	2	4
2	0	2
3	4	12
4	6	0
5	8	16

Los cálculos muestran lo siguiente:

$$\begin{aligned} \hat{y}_i &= 1.1939 + 0.446x_{1i} + 0.0030x_{2i} \\ es(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) &= (0.7737)\ (0.1848)\quad (0.0841) \\ t(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) &= (1.5431)\ \ (2.4151)\ \ (0.0358) \\ cov(\hat{\beta}_1, \hat{\beta}_2) &= -0.00088, \quad g\,de\,l = 2 \\ R^2 &= 0.8101 \quad r_{12} = 0.5523 \end{aligned} \qquad \textbf{(7.25)}$$

La regresión (7.25) muestra que ninguno de los coeficientes de regresión es significativo individualmente al nivel convencional del 1% o 5% de significancia aunque $\hat{\beta}_1$ es significativo al nivel del 10% con base en una prueba t de una cola.

Ahora veamos la tabla 7.6. La única diferencia entre las tabla 7.5 y 7.6 es que los terceros y cuartos valores de x_2 están intercambiados.

Tabla 7.6 Cifras hipotéticas de "y", x_1 y x_2.

y	x_1	x_2
1	2	4
2	0	2
3	4	0
4	6	12
5	8	16

Los cálculos muestran lo siguiente:

$$\begin{aligned} \hat{y}_i &= 1.2108 + 0.4014x_{1i} + 0.0270x_{2i} \\ (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) &= (0.7480)\ (0.2721)\quad (0.1252) \\ t(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) &= (1.6187)\ \ (1.4752)\ \ (0.2158) \\ \text{cov}(\hat{\beta}_1, \hat{\beta}_2) &= -0.0282, \quad \text{g de l} = 2 \\ R^2 &= 0.8143 \quad r_{12} = 0.8285 \end{aligned} \qquad \textbf{(7.26)}$$

Como resultado de la pequeña diferencia en las cifras vemos que $\hat{\beta}_1$, que antes era estadísticamente significativa al nivel del 10%, deja de serlo. También se ve en la ecuación (7.25) la $\text{cov}(\hat{\beta}_1, \hat{\beta}_2) = -0.00088$ mientras que en la ecuación (7.26) es -0.0282, es decir, mas de 30 veces diferente. Todos estos cambios pueden atribuirse a un aumento en la multicolinealidad: en (7.25) el $r_{12} = 0.5523$ mientras que en (7.26) es 0.8285. Igualmente, los errores estándar de $\hat{\beta}_1$ y $\hat{\beta}_2$ aumentan de una regresión a otra, síntoma común de colinealidad.

5. Si la multicolinealidad es alta, se puede obtener un R^2 alto aunque con pocos o casi ningún coeficiente estimado estadísticamente significativo. De este modo, en la regresión (7.26) el $R^2 = 0.8143$ que quiere decir que alrededor de 81.43% de la variación de "y" se explica por x_1 y x_2, y ninguno de los coeficientes individuales es estadísticamente significativos al nivel del 10%. En conclusión, la alta multicolinealidad puede hacer imposible separar los efectos individuales de las variables independientes.

Ejemplo 3:

Para ilustrar los puntos antes mencionados, consideramos los datos de consumo, ingreso y riqueza del consumidor que se muestran en la tabla siguiente:

Tabla 7.7 Cifras hipotéticas de gastos de consumo "y", ingreso x_1 y riqueza x_2.

y ($)	x_1 ($)	x_2 ($)
70.00	80.00	810.00
65.00	100.00	1009.00
90.00	120.00	1273.00
95.00	140.00	1425.00
110.00	160.00	1633.00
115.00	180.00	1876.00
120.00	200.00	2052.00
140.00	220.00	2201.00
155.00	240.00	2435.00
150.00	260.00	2686.00

Si se supone que el gasto de consumo está linealmente relacionado con el ingreso y la riqueza, con los datos de la tabla 7.7 obtenemos la siguiente regresión:

$$\hat{y}_i = 24.7747 + 0.9415\, x_{1i} - 0.0424\, x_{2i} \qquad \textbf{(7.27)}$$

Tabla 7.8 Estadísticos de resumen.

Parámetro	Estimado	Error estándar	Estadístico t.
β_0	24.7747	6.7525	3.6690
β_1	0.9415	0.8229	1.1442
β_2	-0.0424	0.0807	-0.5261
n = 10	$R^2 = 0.9635$	$\overline{R}^2 = 0.9531$	g de l = 7

La regresión (7.27) muestra que el ingreso y la riqueza conjuntamente explican alrededor del 96.35% de la variación en los ingresos de consumo y ninguno de los coeficientes de la pendiente es individualmente significativo. Además, no solamente la riqueza no es significativa sino que tiene un signo contrario; pues a priori uno esperaría una relación positiva entre consumo y riqueza. Aunque $\hat{\beta}_1$ y $\hat{\beta}_2$ no son significativos individualmente (al nivel del 5%) desde el punto de vista estadístico, si verificamos la hipótesis simultánea de que $\beta_1 = \beta_2 = 0$, puede rechazarse, como se muestra en la tabla 7.9.

Bajo los supuestos convencionales obtenemos:

$$F_0 = \frac{4282.7770}{46.3494} = 92.4019 \qquad \textbf{(7.28)}$$

Donde el valor F_0 es significativo en alto grado.

Tabla 7.9 Cuadro del análisis de varianza para ejemplo 3.

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios
Debido a la regresión	8565.5541	2	4282.7770
Debido a los residuos	324.4459	7	46.3494

Este ejemplo muestra con características dramáticas los efectos de la multicolinealidad. El hecho de que la prueba F sea significativa al nivel del 5% pero que los valores de t de β_1 y β_2 no sean individualmente significativos quiere decir que las dos variables están tan correlacionadas que se hace imposible aislar el efecto individual de la riqueza y del ingreso. En efecto, si corremos la regresión de x_2 contra x_1 tendremos:

$$\hat{x}_{2i} = 7.5454 + 10.1909\, x_{1i} \qquad \textbf{(7.29)}$$

Tabla 7.10 Estadísticos de resumen.

Parámetro	Estimado	Error estándar	Estadístico t.
β_0	7.5454	29.4758	0.2560
β_1	10.1909	0.1643	62.0405
n = 10		$R^2 = 0.9979$	

Que claramente muestra que hay casi perfecta colinealidad entre x_2 y x_1.

Ahora veamos qué ocurre si corremos la regresión de "y" contra x_1:

$$\hat{y}_i = 24.4545 + 0.5091\, x_{1i} \qquad \textbf{(7.30)}$$

Tabla 7.11 Estadísticos de resumen.

Parámetro	Estimado	Error estándar	Estadístico t.
β_0	24.4545	6.4138	3.8128
β_1	0.5091	0.0357	14.2432
n = 10		$R^2 = 0.9621$	

En la ecuación (7.27) la variable ingreso no era estadísticamente significativa para $\alpha = 0.05$, mientras que ahora lo es altamente.

Si en vez de correr la regresión de "y" contra x_1 la corremos contra x_2 obtendremos:

$$\hat{y}_i = 24.411 + 0.050\, x_{2i} \qquad \textbf{(7.31)}$$

Tabla 7.12 Estadísticos de resumen.

Parámetro	Estimado	Error estándar	Estadístico t.
β_0	24.411	6.874	3.551
β_1	0.050	0.004	13.292
n = 10		$R^2 = 0.957$	

Vemos ahora que la riqueza tiene un impacto significativo sobre los gastos de consumo mientras que en (7.27) no tenía tales efectos.

Las regresiones (7.30) y (7.31) muestran claramente que en situaciones de extrema multicolinealidad al descartar la variable altamente colineal se vuelve a la otra variable "x" estadísticamente significativa. Esto siguiere que una salida a la extrema colinealidad implicaría descartar la variable colineal.

7.6.4 Como Detectar la Multicolinealidad.

Una vez estudiadas la naturaleza y las consecuencias de la multicolinealidad, debemos formularnos la siguiente pregunta: ¿Cómo saber que la multicolinealidad está presente en una situación dada, especialmente en los modelos en que se involucran más de dos variables independientes? Existen varios métodos para detectarla, algunos de los cuales se comentan a continuación:

1. Se sospecha que la colinealidad está presente en situaciones en que el R^2 es alto (por ejemplo, entre 0.7 y 1) y cuando las correlaciones de orden cero son altas y a la vez ninguno o pocos de los coeficientes de regresión parcial son individualmente significativos, con base en la prueba t convencional. Si el R^2 es alto quiere decir que la prueba F del análisis de varianza, en la mayoría de los casos, rechazará la hipótesis nula de que el valor verdadero de todos los coeficientes parciales de la pendiente sean simultáneamente cero, independientemente de la prueba t.
2. Las correlaciones simples relativamente altas entre uno o más pares de variables independientes puede indicar multicolinealidad. Sin embargo, las conclusiones sobre la presencia o ausencia de multicolinealidad que sólo se basan en estas correlaciones deben hacerse con cuidado. Es posible que con algunos conjuntos de datos, en especial aquellos que implican series de tiempo, las correlaciones entre muchos pares de variables serán altas, pero los datos le permitirán al investigador separar los efectos de las variables explicativas individuales sobre la variable dependiente. Una limitación adicional es que un examen de las correlaciones simples entre pares de variables no permitirán detectar la multicolinealidad que surge debido a que tres o cuatro variables están relacionadas entre sí.
3. Se han propuesto varias pruebas formales para detectar la multicolinealidad a lo largo de los años, pero ninguna ha encontrado una aceptación amplia. Cuáles de

las pruebas nos permitirán detectar la multicolinealidad dependerá de la naturaleza específica del problema.

7.6.5 Multicolinealidad y Predicción.

Si la predicción es el único propósito del análisis de regresión, el problema de la multicolinealidad no es serio porque mientras mayor sea el R^2, mejor será la predicción. Nótese que esto es válido en la medida en que la colinealidad existente entre las variables "x" en una muestra dada, se mantenga en el futuro.

Sin embargo, si la relación lineal aproximada entre las variables "x" de la muestra no se presentan en muestras futuras, la predicción será sin duda incierta. Pero si el objetivo del análisis no es la predicción si no la estimación confiable de los parámetros, la multicolinealidad es todo un problema por que conlleva a grandes errores estándar de los estimadores.

7.6.6 Medidas Remediales.

¿Qué puede hacerse si la multicolinealidad es seria? Al igual que en el caso de la detección, no hay guías seguras porque justamente la multicolinealidad es un problema muestral. Sin embargo, se pueden ensayar las siguientes reglas generales, sin olvidar que el éxito dependerá de la severidad del problema de la multicolinealidad.

1. Información a priori. Consideremos el siguiente modelo

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Donde:

y_i: Consumo.

x_1: Ingreso.

x_2: Riqueza.

Como se dijo antes las variables ingreso y riqueza tienden a ser altamente colineales. Pero supongamos a priori que $\beta_2 = 0.10\beta_1$; es decir, que la tasa de variación del consumo con respecto a la riqueza es un décimo de la correspondiente tasa con respecto al ingreso. Podemos entonces correr la siguiente regresión:

$$y_i = \beta_0 + \beta_1 x_{1i} + 0.10\beta_2 x_{2i} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Donde $x_i = x_{1i} + 0.10\, x_{2i}$. Una vez obtenido $\hat{\beta}_1$ podemos seguir a estimar $\hat{\beta}_2$ a partir de la relación postulada entre β_1 y β_2.

¿Cómo se obtiene la información a priori? Puede provenir de la teoría Económica o de trabajos empíricos en los cuales el problema de la colinealidad es menos serio.

2. Combinación de cifras de corte transversal y de series de tiempo. Una variante de la técnica de la información a priori es la combinación de las cifras de corte transversal y de series de tiempo, conocida como mezcla de datos. Supongamos que queremos estudiar la demanda de automóviles en El Salvador y que se dispone de series de tiempo del número de carros vendidos, precio promedio del carro, e ingreso del consumidor. Supongamos también que:

$$\ln y_i = \beta_0 + \beta_1 \ln P_t + \beta_2 \ln I_t + \varepsilon_t$$

Donde:

y: Número de carros vendidos.

P: Precio promedio.

I: Ingreso.

t: Tiempo.

Nuestro objetivo es estimar la elasticidad precio β_1 y la elasticidad ingreso β_2.

Ahora bien, tratándose de series de tiempo las variables precio e ingreso tienden a ser altamente colineales. Por consiguiente, si corremos la regresión anterior nos enfrentaremos al problema usual de la multicolinealidad. Una salida del problema ha sido sugerida por Tobin[4] quien sugiere que si tenemos datos de corte transversal (como los que se generan por paneles de consumidores o por estudios presupuestales de los que llevan a cabo agencias privadas y gubernamentales), podemos obtener estimación relativamente precisa de la elasticidad y el ingreso β_2 porque con estos datos, que son un punto en el tiempo, los precios no varían mucho. Sea $\hat{\beta}_2$, la elasticidad ingreso, estimada a partir de los datos de corte transversal. Utilizando esta estimación, la anterior regresión con series de tiempo puede escribirse como:

$$y_t^* = \beta_0 + \beta_1 \ln P_t + \varepsilon_t$$

[4] J. Tobin, "a Statistical Demand Fuction for Food in the U.S.A.", Journal of the Royal Statistical Society, ser, A, pp. 113-141, 1950.

Donde $y_t^* = \ln y - \beta_2 \ln I_t$, que representa el valor de "y" después de suprimirle el efecto del ingreso. Es claro que ahora se puede obtener una estimación de la elasticidad precio β_1 a partir de la regresión anterior.

Aunque la técnica parece atractiva, "mezclar" las cifras de corte transversal con las de series de tiempo puede crear problemas de interpretación, porque en este caso suponemos implícitamente que la elasticidad del ingreso, estimada a partir de cifras de corte transversal, es igual a la que se hubiera obtenido a partir del análisis de series de tiempo.

Sin embargo, la técnica ha tenido muchas aplicaciones y es particularmente valiosa en situaciones en las cuales los estimadores de corte transversal no varían sustancialmente de una muestra a otra.

3. Eliminación de variables y sesgo de especificación. Cuando enfrentamos el problema de la multicolinealidad severa, una de las soluciones más "simples" es omitir una de las variables colineales. Lo problemático al descartar una variable puede ser que estemos incurriendo en sesgo de especificación o error de especificación, que generalmente aparece como consecuencia de una especificación incorrecta del modelo analizado.
4. Transformaciones de variables. Supongamos que poseemos cifras en forma de series de tiempo para los gastos de consumo, ingreso y riqueza. Una razón que explica la alta multicolinealidad entre ingreso y riqueza en estos datos, es la de que en el tiempo ambas variables tienden a moverse en la misma dirección. Una manera de minimizar esta dependencia es la siguiente:

Si la relación

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t \qquad \textbf{(7.32)}$$

Se cumple en el tiempo t, también debe cumplirse en t – 1 en razón de que el origen del tiempo es arbitrario. Por lo tanto, tenemos que:

$$y_{t-1} = \beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{2,t-1} + \varepsilon_{t-1} \qquad \textbf{(7.33)}$$

Si restamos la ecuación (7.32) de (7.33) obtendremos:

$$y_t - y_{t-1} = \beta_1 (x_{1t} - x_{1,t-1}) + \beta_2 (x_{2t} - x_{2,t-1}) + v_t \qquad \textbf{(7.34)}$$

Donde $v_t = u_t - u_{t-1}$.

La ecuación (7.34) se conoce como forma de primeras diferencias en razón de que se corre la regresión no sobre las variables originales sino sobre las diferencias de sus valores sucesivos.

El modelo de primeras diferencias reduce a menudo la severidad de la multicolinealidad porque aunque los niveles de x_1 y x_2 estén altamente correlacionados no existe razón a priori para pensar que sus diferencias estén correlacionadas también en alto grado.

La transformación de primeras diferencias crea sin embargo, otros problemas. El término de error v_t que aparece en (7.34) puede no satisfacer uno de los supuestos del modelo de regresión lineal clásico según el cual las perturbaciones no están correlacionadas serialmente. Como se verá más adelante, si el u_t original no está autocorrelacionado, o lo que es igual, es serialmente independiente, el término de error v_t previamente obtenido será el

mayor número de veces correlacionado serialmente. En este caso, el remedio vuelve a ser peor que la enfermedad, y se pierde además una observación al sacar las diferencias, reduciendo así, en uno los grados de libertad. En una muestra pequeña, este factor puede ser considerable. Más aún, el procedimiento de las primeras diferencias puede no ser apropiado para cifras de corte transversal en que no hay un ordenamiento lógico de las observaciones.

5. Datos nuevos o adicionales. Como la multicolinealidad es un problema muestral, es posible que en otras muestras con las mismas variables, la colinealidad no sea tan seria como en la primera muestra. En algunas ocasiones, con aumentar tan solo el tamaño de la muestra (de ser posible) se atenúa el problema: por ejemplo, en el modelo de tres variables vimos que:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2(1 - r_{12}^2)}$$

Ahora, a medida que el tamaño de la muestra aumenta, $\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2$ también aumentará. En consecuencia, para cualquier r_{12} dado, la varianza de $\hat{\beta}_1$ disminuirá reduciéndose por ende el error estándar lo cual nos permite estimar más precisamente a $\hat{\beta}_1$.

Debemos tener en cuenta en el análisis de regresión que cuando se obtiene un valor para t no significativo para los coeficientes de regresión, existe la

tendencia de culpar de la falta de significancia a la multicolinealidad, pudiendo ser más bien la culpa de un sesgo de especificación. Talvez el modelo usado en el análisis está mal especificado, o el soporte teórico para el modelo es muy débil, con lo cual podemos afirmar que antes de que el investigador le atribuya la culpa de sus problemas de t insignificantes a la multicolinealidad, debe revisar el modelo desde el punto de vista teórico, siendo probable que la misma bibliografía sugiera una especificación alterna del mismo.

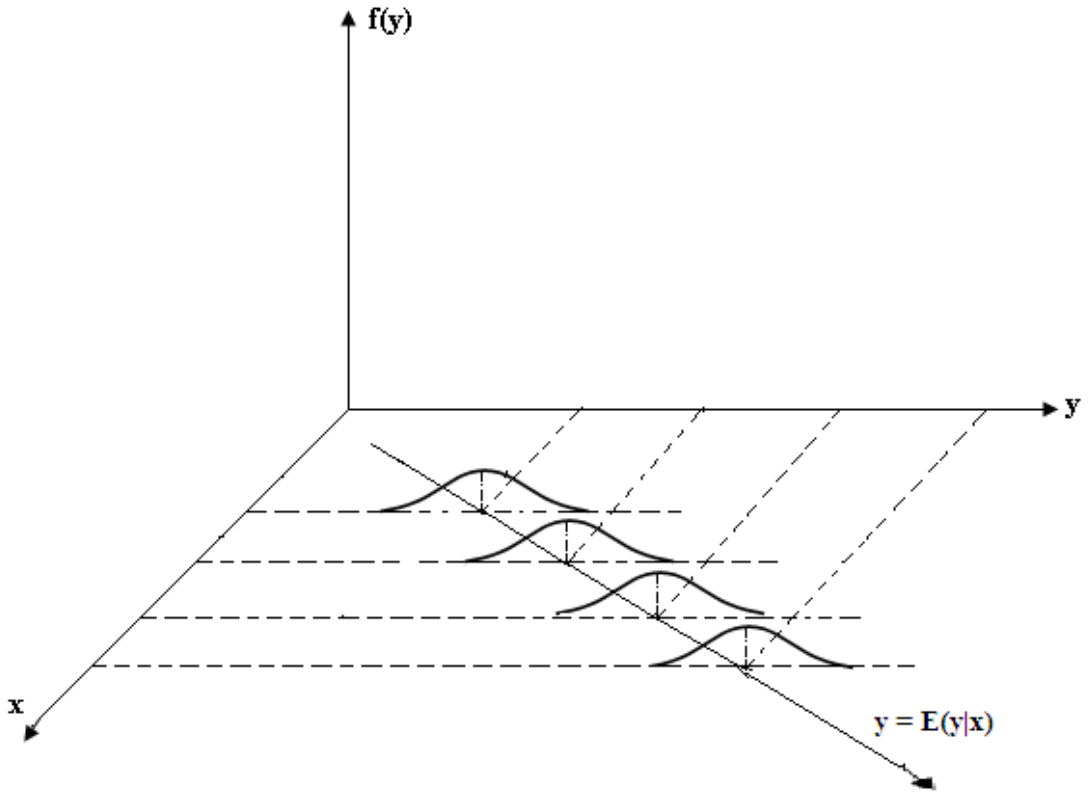
7.7 Heteroscedasticidad.

Uno de los supuestos importantes del modelo de regresión lineal clásico consiste en que la varianza de cada perturbación ε_i, condicional a los valores escogidos de las variables independientes, es una constante igual a σ^2. Este es el supuesto de homoscedasticidad, que viene de (homo) igual y (cedasticidad) dispersión, es decir, igual varianza. Simbólicamente:

$$E(\varepsilon_i^2) = \sigma^2 \qquad i = 1, 2, \ldots, n \qquad \textbf{(7.35)}$$

Gráficamente, en el modelo de regresión lineal de dos variables, la homoscedasticidad puede representarse como en la figura 1.8 que por conveniencia, se reproduce como la figura 7.7.

Figura 7.7 Perturbaciones homoscedásticas.



Como la figura lo muestra, la varianza condicional de y_i (que es igual a la de ε_i), condicional a los valores dados de x_i, permanece constante independientemente de los valores que tome "x".

En contraste con esta figura, considere la figura 7.8 que muestra como la varianza condicional de y_i aumenta a medida que "x" aumenta. En este caso, las varianzas de y_i no son iguales, por lo cual se presenta la heteroscedasticidad.

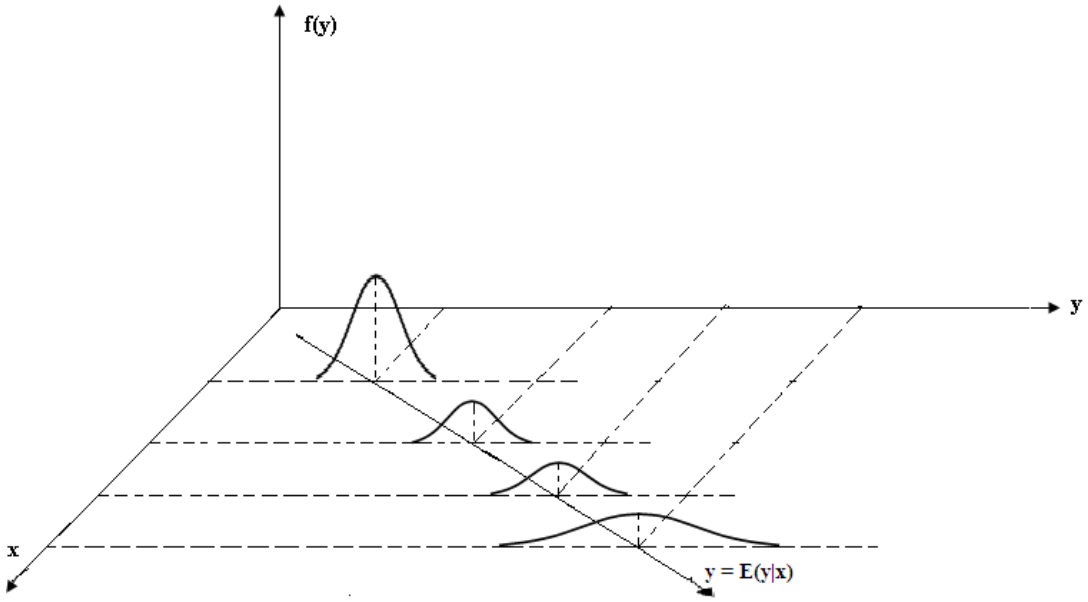
Simbólicamente:

$$E(\varepsilon_i^2) = \sigma_i^2 \qquad \textbf{(7.36)}$$

Nótese el subíndice en σ^2 que nos recuerda que las varianzas condicionales de ε_i (varianza condicional de y_i) ya no son constantes.

Para establecer claramente la diferencia entre homoscedasticidad y heteroscedasticidad suponga que en el modelo de dos variables $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, "y" representa el ahorro y "x" representa el ingreso. Las figuras 7.7 y 7.8 muestran que a medida que aumenta el ingreso, el ahorro también aumenta, en promedio. Sin embargo, en la figura 7.7 la varianza del ahorro permanece constante en todos los niveles de ingreso, mientras que en la figura 7.8 la varianza aumenta con el ingreso. Parece, según la figura 7.8 que las familias de más altos ingresos, en promedio, ahorran más que las familias de bajos ingresos, pero también que hay más variabilidad en sus ahorros.

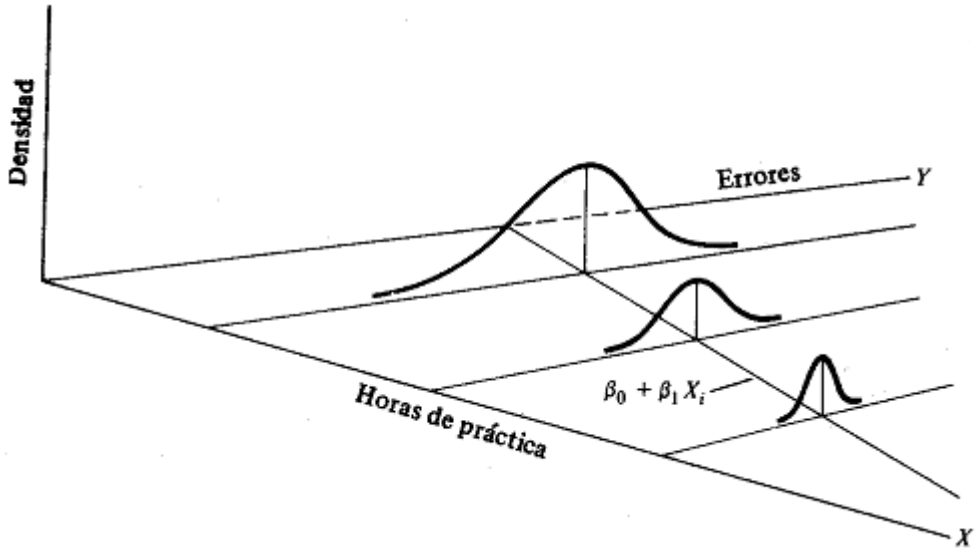
Figura 7.8 Perturbaciones heteroscedásticas.



Existen varias razones para que las varianzas de ε_i sean variables, entre las cuales se destacan las siguientes:

1. Siguiendo los modelos de aprendizaje por errores, a medida que la gente aprende, sus errores en el comportamiento van disminuyendo en el tiempo. En este caso, se espera que σ_i^2 disminuya. A manera de ejemplo, considere la figura 7.9 que nos presenta el número de errores de mecanografía cometidos en un período determinado, en una prueba, contra el número de horas de práctica.

Figura 7.9 Ilustración de la heteroscedasticidad.



Como se ve en la figura, a medida que el número de horas de práctica aumenta, el número promedio de errores disminuye y su varianza también disminuye.

2. A medida que los ingresos aumentan, la gente tiene más ingreso discrecional y por lo tanto más oportunidad para elegir cómo disponer de sus ingresos. De este modo σ_i^2 tiende a aumentar con el ingreso por lo cual, en la regresión del ahorro

contra el ingreso es muy factible encontrar que σ_i^2 aumente con el ingreso (como en la figura 7.8) ya que la gente tiene más oportunidades para colocar sus ahorros. De igual forma, las compañías que obtienen grandes utilidades tienden a presentar más variabilidad en cuanto a sus políticas de dividendos que las de menores ganancias. Las empresas orientadas hacia la expansión por lo general presentan más variabilidad en sus tasas de dividendos pagados que las compañías ya establecidas.

3. A medida que las técnicas de recolección mejoran, σ_i^2 tiende a disminuir. Los bancos que disponen de equipos sofisticados de procesamiento de datos tienen menos posibilidad de cometer errores en sus informes mensuales o trimestrales que los que no disponen de tales facilidades.

Debe señalarse que el problema de heteroscedasticidad tiende a ser más común en las informaciones de corte transversal que en las series de tiempo. En la información de corte transversal por lo general se trabaja con miembros de una población, en un momento determinado, tales como consumidores individuales o sus familias, firmas, industrias, subdivisiones geográficas como países, estados o ciudades, etc. Además, estos miembros pueden ser de diferentes tamaños como firmas grandes, pequeñas o medianas o de ingresos altos, bajos o medianos. En la información de series de tiempo, por otra parte, la variable tiende a ser de órdenes de magnitud similares porque generalmente se recoge información para la misma entidad durante un período de tiempo. Como ejemplos podemos citar el PNB (Producto Nacional Bruto), el consumo, el ahorro o el empleo en El Salvador en el período de 1950-1975.

7.7.1 Consecuencias de la Heteroscedasticidad.

Tenga en cuenta que si todos los supuestos del modelo clásico se cumplen, los estimadores de MCO son Mejores Estimadores Lineales Insesgados, es decir, entre todos los estimadores insesgados, tienen la mínima varianza. En síntesis, son eficientes. Si mantenemos ahora todos los supuestos excepto el de homoscedasticidad, podemos probar que los estimadores de MCO siguen siendo insesgados y consistentes pero ya no son eficientes para ningún tipo de muestras, grandes o pequeñas. En otras palabras, en muestras repetidas los estimadores MCO son iguales, en promedio, a los verdaderos valores poblacionales (la propiedad de ser insesgados), y a medida que el tamaño de la muestra crece indefinidamente, convergen a su verdadero valor (la propiedad de consistencia), pero sus varianzas ya no son mínimas inclusive cuando el tamaño de la muestra crece indefinidamente (la propiedad de eficiencia asintótica).

Para concretar mejor la idea, volvamos al caso de dos variables:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Ahora dejando $E(\varepsilon_i^2) = \sigma_i^2$ pero manteniendo todos los demás supuestos de MCO se puede demostrar que el método de mínimos cuadrados ponderados (se estudiará más adelante) nos da el mejor estimador lineal insesgado de β_1, digamos β_1^*, que es como sigue:

$$\beta_1^* = \frac{\left(\sum w_i\right)\left(\sum w_i x_i y_i\right) - \left(\sum w_i x_i\right)\left(\sum w_i y_i\right)}{\left(\sum w_i\right)\left(\sum w_i x_i^2\right) - \left(\sum w_i x_i\right)^2} \qquad \textbf{(7.37)}$$

Y su varianza está dada por:

$$\text{var}(\beta_1^*) = \frac{\sum w_i}{\left(\sum w_i\right)\left(\sum w_i x_i^2\right) - \left(\sum w_i x_i\right)^2} \qquad \textbf{(7.38)}$$

Donde

$$w_i = \frac{1}{\sigma_i^2} \qquad \textbf{(7.39)}$$

El estimador β_1^* se conoce como estimador de mínimos cuadrados ponderados por las razones que se explicaran más adelante.

De otra forma, el estimador común de β_1 de MCO es:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad \textbf{(7.40)}$$

Y si ocurre heteroscedasticidad su varianza será:

$$\text{var}(\hat{\beta}_1) = \frac{S_{xx}\sigma_i^2}{S_{xx}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_i^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \qquad \textbf{(7.41)}$$

Del apéndice 2 del Capítulo 2, se puede deducir que $\hat{\beta}_1$ sigue siendo insesgado, de hecho, la propiedad de ser insesgado no requiere que las perturbaciones ε_i sean homoscedásticas. Sin embargo, la varianza de $\hat{\beta}_1$ dada en (7.40) es diferente (efectivamente mayor que) de la varianza de β_1^* dada en (7.37) y ya habíamos

establecido que β_1^* es mejor estimador lineal insesgado. La conclusión de nuestro análisis es, entonces, que $\hat{\beta}_1$ aunque insesgado es ineficiente, su varianza es mayor, es decir mayor que la de β_1^*.

En la práctica lo que puede suceder es que no sepamos que en una determinada situación existe heteroscedasticidad y, por lo tanto, resultemos usando equivocadamente las fórmulas comunes de MCO derivadas para la homoscedasticidad. ¿Cuáles serán las consecuencias de este hecho? Para responder, continuemos con el modelo de dos variables. Como antes, el estimador de $\hat{\beta}_1$ es dado por (7.40) y debido al supuesto de homoscedasticidad su varianza es la fórmula común:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \qquad \textbf{(7.42)}$$

Si existe además heteroscedasticidad, debemos utilizar (7.41) aun cuando la varianza obtenida sea ineficiente. Para ver las consecuencias de la utilización de (7.42) en lugar de (7.41), digamos que:

$$\sigma_i^2 = \sigma^2 c_i \qquad \textbf{(7.43)}$$

Donde c_i son algunas ponderaciones constantes (no estocásticas) no necesariamente todas iguales. La ecuación (7.43) nos dice que las varianzas heteroscedásticas son proporcionales a c_i, siendo σ^2 el factor de proporcionalidad. (Nota: A diferencia de σ_i^2, σ^2 es una constante.)

Sustituyendo (7.43) en (7.41) obtenemos:

$$\text{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma^2 c_i}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 c_i}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)} \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 c_i\right)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)}$$

$$\text{var}(\hat{\beta}_1) = \text{var}(\hat{\beta}_1^{MCO}) \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 c_i\right)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)} \qquad \textbf{(7.44)}$$

Donde $\text{var}(\hat{\beta}_1^{MCO})$ es la varianza de $\hat{\beta}_1$ bajo el supuesto de homoscedasticidad, como se mostró en (7.42).

Se ve claramente en (7.44) que si $(x_i - \bar{x})^2$ y c_i, están correlacionadas positivamente, como puede asegurarse en la mayoría de datos económicos, y si $\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 c_i\right) \Big/ \left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$ es mayor que 1, la varianza de $\hat{\beta}_1$ cuando existe heteroscedasticidad será mayor que su varianza en caso de homoscedasticidad. En estas condiciones, la fórmula común de MCO (7.42) subestimará la verdadera varianza de $\hat{\beta}_1$ dada en (7.41) por ser ineficiente. Por consiguiente, subestimaríamos el verdadero error

estándar de $\hat{\beta}_1$ y por lo tanto sobreestimaremos el valor de t asociado con $\hat{\beta}_1$ [recuerde que bajo la hipótesis nula $\beta_1 = 0$, $t = \hat{\beta}_1 / es(\hat{\beta}_1)$], lo que nos puede llevar a la conclusión de que, en el caso específico que analizamos, $\hat{\beta}_1$ es estadísticamente significativo. Naturalmente, si la verdadera varianza dada en (7.41) fuera conocida, el "correcto" valor de t podría mostrar que $\hat{\beta}_1$ es, de hecho, estadísticamente insignificante. Todo esto nos permite pensar que la heteroscedasticidad es potencialmente un problema complicado.

Por consiguiente el resultado final de la discusión anterior se puede concretar así:

1. Cuando existe heteroscedasticidad o se sospecha que existe, teóricamente el mejor estimador lineal insesgado de $\hat{\beta}_1$ es el estimador de mínimos cuadrados ponderados β_1^* , no el estimador convencional $\hat{\beta}_1$, aunque éste sea insesgado.
2. La varianza de $\hat{\beta}_1$ obtenida bajo el supuesto de heteroscedasticidad y dada por (7.41) ya no es la mínima. La mínima es la varianza de β_1^* dada en (7.38).
3. Respecto de 2, si usamos la fórmula de la varianza dada en (7.41) en lugar de (7.38), el intervalo de confianza para $\hat{\beta}_1$ es innecesariamente ancho y las pruebas de significación tienen menos fuerza.
4. El problema se complica más si en condiciones de heteroscedasticidad, en lugar de usar (7.41), que es ineficiente como ya vimos, usamos la fórmula común de MCO (7.42). Para estimar la varianza de $\hat{\beta}_1$. Como se anotó anteriormente, (7.42) es un estimador sesgado de (7.41), resultando el sesgo del hecho de que el estimador convencional de σ^2, $\hat{\sigma}^2$, no es insesgado. La naturaleza del sesgo

depende de la relación entre σ_i^2 y los valores que toman las variables explicatorias.

5. Como consecuencia de 4, si en las condiciones de heteroscedasticidad continuamos aplicando equivocadamente las fórmulas tradicionales de MCO (obtenidas bajo los supuestos de homoscedasticidad), las conclusiones serán falsas pues las pruebas t y F tienden a exagerar la significancia estadística de los parámetros estimados convencionalmente. Por lo tanto, en casos de heteroscedasticidad el estimador convencional de (7.42) es inapropiado. Debemos utilizar al menos (7.41) aun cuando la varianza obtenida con esta fórmula no sea la mínima. Lo ideal es, naturalmente, utilizar (7.38), reemplazando $\hat{\beta}_1$ por β_1^*.

Aunque en algunos casos y bajo hipótesis específicas acerca de la forma de σ_i^2 se puede saber la naturaleza del sesgo de las varianzas y los errores estándar, hallados equivocadamente con las fórmulas corrientes de MCO, para el caso de homoscedasticidad, en general no es posible detectarlo tan rápidamente. Esto se debe a que el sesgo de las varianzas estimadas depende de la naturaleza de la heteroscedasticidad misma (es decir, de la forma de σ_i^2), así como también de la naturaleza de los valores de "x" que aparecen en la muestra. En la práctica muy rara vez se sabe cuál es la verdadera σ_i^2. Por consiguiente, a pesar de su superioridad teórica, el estimador de mínimos cuadrados ponderados β_1^* no se puede obtener fácilmente. Lo

usual para tratar el problema de la heteroscedasticidad es hacer algunos supuestos ad hoc acerca de σ_i^2. La ecuación (7.43) representa uno de tales supuestos.

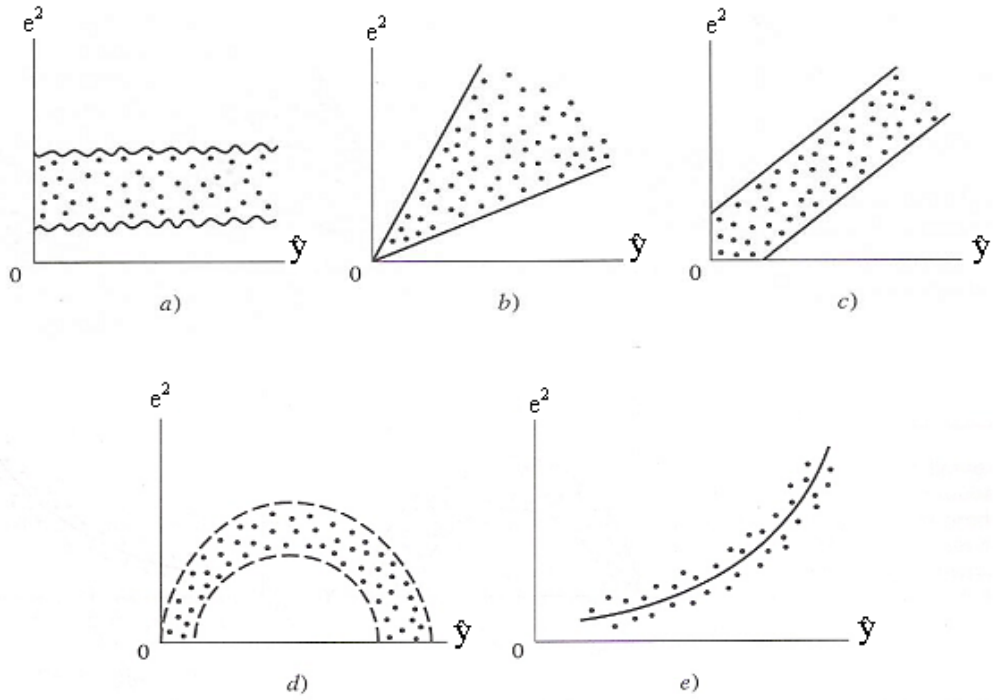
7.7.2 Como Detectar la Heteroscedasticidad.

Como lo hicimos con la multicolinealidad, es preciso preguntarse: ¿cómo sabemos que en una situación específica se presenta la heteroscedasticidad? Otra vez, como en el caso de multicolinealidad, no existen reglas fijas y seguras para detectarla sino solamente unas cuantas normas muy generales. Esto es inevitable ya que σ_i^2 se puede conocer solamente cuando tenemos toda la población "y" correspondiente a las "x" escogidas. No obstante, sólo se cuenta con esta información excepcionalmente en la mayoría de las investigaciones económicas. En esto difieren los econometristas de los científicos de otros campos como la agricultura y la biología, en donde se puede tener el suficiente control sobre los objetos de la investigación. Lo más corriente en estudios económicos es tener sólo un valor muestral de "y" para cada valor particular de "x" y por esto no hay manera de conocer σ_i^2 a partir de una sola observación de "y". Es así como en la mayoría de las investigaciones econométricas, la heteroscedasticidad puede ser motivo de "especulación" o de "soluciones ad hoc".

Teniendo en cuenta la advertencia anterior, examinemos algunos de los métodos formales e informales para detectar la heteroscedasticidad.

1. **Naturaleza del problema**. A menudo la naturaleza del problema sugiere cuándo existe la heteroscedasticidad. Por ejemplo, siguiendo el trabajo de Prais y Houthakker sobre los presupuestos familiares, en los que encontraron que la varianza residual de la regresión del consumo contra el ingreso aumentaba con el ingreso, se supone generalmente ahora, que en estudios similares se pueden esperar diferentes varianzas en las perturbaciones. Efectivamente, en la información de corte transversal que contiene unidades heterogéneas, lo más común es que exista heteroscedasticidad. Por lo tanto, en un análisis de corte transversal que incluya los gastos de inversión con relación a las ventas, a la tasa de interés, etc. es muy probable que haya heteroscedasticidad si se han tomado conjuntamente como muestra empresas pequeñas, medianas y grandes.
2. **Método gráfico**. En la práctica, cuando no existe información a priori o empírica acerca de la naturaleza de la heteroscedasticidad, se puede hacer el análisis de regresión sobre el supuesto de que no existe heteroscedasticidad y luego hacer un examen posterior de los residuos estimados al cuadrado e_i^2, para ver si presentan algún patrón sistemático. Aunque e_i^2 y ε_i^2 no son la misma cosa, pueden usarse los unos como aproximaciones de los otros especialmente cuando la muestra es lo suficientemente grande. Al examinar la SS_{Res} podemos encontrar patrones como los que aparecen en la figura 7.10.

Figura 7.10 Patrones hipotéticos de los residuos estimados al cuadrado.



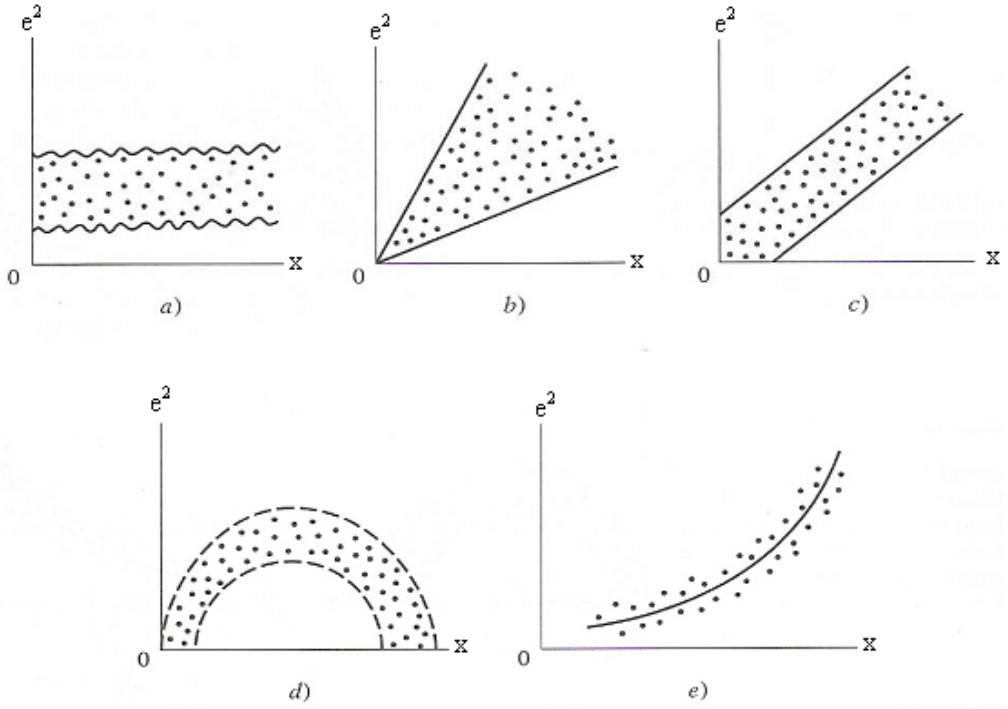
En la figura 7.10, e_i^2 están dibujados contra $\hat{y}_i$, los y_i estimados, a partir de la línea de regresión, con la idea de ver si el valor medio estimado de "y" está relacionado sistemáticamente con el residuo al cuadrado. En la figura 7.10a) se advierte que no hay un patrón sistemático entre las dos variables, lo que sugiere la inexistencia de heteroscedasticidad en la información. Las figuras de la 7.10b) a la 7.10e), muestran patrones definidos. Por ejemplo, la figura 7.10c) sugiere una relación lineal mientras que las figuras 7.10d) a 7.10e) muestran una relación cuadrática entre e_i^2 y $\hat{y}_i$. Utilizando esta información, aunque informal,

podemos transformar los datos de modo que una vez transformados, no presenten heteroscedasticidad.

En lugar de dibujar e_i^2 contra $\hat{y}_i$ podemos trazarlos contra una de las variables independientes, especialmente si al dibujarlos contra $\hat{y}_i$ nos resulta un patrón como el que muestra en la figura 7.10a). El dibujo resultante, que aparece en la figura 7.11 puede dar patrones semejantes a los de la figura 7.10. (En el modelo de dos variables, dibujar e_i^2 contra $\hat{y}_i$ es equivalente a dibujarlos contra x_i, y por lo tanto la figura 7.11 es similar a la 7.10. Sin embargo, no es esta la situación al considerar un modelo de más de dos variables, porque en este caso e_i^2 puede dibujarse contra cualquiera de las variables "x" del modelo.)

Un patrón como el de la figura 7.11c), por ejemplo, sugiere que la varianza del término de error está relacionada linealmente con la variable "x". De este modo, si en la regresión de ahorro contra ingreso encontramos un patrón de este tipo, esto nos sugiere que la varianza heteroscedástica puede ser proporcional al valor de la variable ingreso. Esta información puede ayudar a transformar nuestros datos de modo que en la regresión que contiene los datos transformados la varianza de la perturbación sea homoscedástica.

Figura 7.11 Diagrama de los residuos estimados al cuadrado, contra x.



3. **Prueba de Park**. Park formaliza el método gráfico sugiriendo que σ_i^2 es una función de la variable explicatoria x_i. La forma funcional propuesta por Park es:

$$\sigma_i^2 = \sigma^2 x_i^{\beta} \ell^{v_i}$$

o

$$\ln \sigma_i^2 = \ln \sigma^2 + \beta \ln x_i + v_i \qquad \textbf{(7.45)}$$

Donde v_i es el término estocástico de perturbación.

Dado que σ_i^2 es por lo general desconocida, Park propone que se use e_i^2 como aproximación y que se realice la siguiente regresión:

$$\begin{aligned} \ln e_i^2 &= \ln \sigma^2 + \beta \ln x_i + v_i \\ \ln e_i^2 &= \alpha + \beta \ln x_i + v_i \end{aligned} \qquad \textbf{(7.46)}$$

Si β resulta estadísticamente significativa eso nos sugiere que existe heteroscedasticidad. Si resulta no significativa, podemos aceptar la hipótesis de homoscedasticidad. La prueba de Park es, por lo tanto, un procedimiento en dos etapas. En la primera etapa se realiza la regresión de MCO sin tener en cuenta el problema de la heteroscedasticidad. De esta regresión obtenemos e_i y luego, en la segunda etapa, llevamos a cabo la regresión (7.46).

A pesar de todo, la prueba de Park presenta algunos problemas. Goldfeld y Quandt afirman que el término de error v_i en (7.46) es posible que no cumpla los supuestos de MCO y puede ser el mismo heteroscedástico. Sin embargo, se puede usar como método estrictamente indicativo.

Con el fin de ilustrar el enfoque de Park, empleamos los datos de las últimas filas de la tabla 7.20 que se muestra al final del apéndice 7.1, de ahí se obtiene la tabla 7.13.

Ejemplo 4:

Tabla 7.13 Remuneración media y productividad media según la escala de empleo del establecimiento.

Remuneración media	Productividad media
3396	9355
3787	8584
4013	7962
4104	8275
4146	8389
4241	9418
4387	9795
4538	10281
4843	11750

Para hacer la siguiente regresión:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Donde:

y_i: Remuneración media en miles de dólares.

x_i: Productividad media en miles de dólares.

i: La i-ésima escala de empleo del establecimiento.

Los resultados de la regresión son los siguientes:

$$\hat{y}_i = 1992.345 + 0.233 x_i \qquad \textbf{(7.47)}$$

$$es(\hat{\beta}_1) = 0.1000 \qquad t(\hat{\beta}_1) = (2.333) \qquad R^2 = 0.438$$

Los resultados nos indican que el coeficiente de la pendiente estimado es significativo al nivel del 5%, con base en una prueba t de una cola. La ecuación (7.47) muestra que a medida que la productividad aumenta en un dólar, por ejemplo, la remuneración media del trabajo aumenta en cerca de 23 centavos.

Ahora haciendo uso de la ecuación (7.47) obtenemos los residuos, los elevamos al cuadrado, calculamos el logaritmo natural a los residuos al cuadrado y a la variable “x”. Estos cálculos se muestran en la tabla siguiente:

Tabla 7.14 Resultados obtenidos haciendo uso de la ecuación (7.47).

e_i	e_i^2	$\ln (e_i^2)$	$\ln (x_i)$
-775.6579	601645.23	13.31	9.14
-205.0481	42044.72	10.65	9.06
165.85117	27506.61	10.22	8.98
183.93563	33832.32	10.43	9.02
199.37853	39751.8	10.59	9.03
54.66578	2988.35	8.00	9.15
112.84099	12733.09	9.45	9.19
150.62388	22687.55	10.03	9.24
113.41004	12861.84	9.46	9.37

Con la información de la tabla anterior estimamos los valores de los coeficientes tomando como variable dependiente $\ln (e_i^2)$ y como variable independiente $\ln (x_i)$ así se obtiene la siguiente ecuación de regresión:

$$\begin{aligned} \ln e_i^2 &= \ln \sigma^2 + \beta \ln x_i + v_i \\ \ln e_i^2 &= \alpha + \beta \ln x_i + v_i \qquad \textbf{(7.48)} \\ \ln e_i^2 &= 35.817 - 2.801 \ln x_i \end{aligned}$$

$es(\hat{\beta}) = 4.196 \qquad t(\hat{\beta}) = (-0.668) \qquad R^2 = 0.060$

Se puede ver que no hay una relación estadísticamente significativa entre las dos variables. Siguiendo la prueba de Park, podemos concluir que no hay heteroscedasticidad en la varianza del error[5].

4. **Prueba de Glejser**. La prueba de Glejser es esencialmente similar a la prueba de Park. Después de obtener los residuos e_i de la regresión de MCO, Glejser sugiere

[5] La forma funcional escogida por Park es tan sólo una sugerencia. Una forma funcional diferente puede revelar una relación significativa. Por ejemplo, podemos usar e_i^2 en lugar de $\ln e_i^2$ como variable dependiente.

que se calcule la regresión de los valores absolutos de e_i, $|e_i|$, contra la variable "x" que se supone asociada íntimamente con σ_i^2. En este experimento Glejser usó las siguientes fórmulas funcionales:

$$|e_i| = \beta_1 x_i + v_i$$
$$|e_i| = \beta_1 \sqrt{x_i} + v_i$$
$$|e_i| = \beta_1 \frac{1}{x_i} + v_i$$
$$|e_i| = \beta_1 \frac{1}{\sqrt{x_i}} + v_i$$

$$|e_i| = \beta_0 + \beta_1 x_i + v_i$$
$$|e_i| = \sqrt{\beta_0 + \beta_1 x_i} + v_i$$
$$|e_i| = \sqrt{\beta_0 + \beta_1 x_i^2} + v_i$$

Donde v_i es el término de error.

El método de Glejser puede utilizarse también como solución empírica; pero Golfeld y Quandt afirman que el término de error v_i tiene algunos problemas por cuanto su valor esperado no es cero, está serialmente correlacionado, e irónicamente es heteroscedástico. Otra dificultad del método de Glejser es que los modelos como:

$$|e_i| = \sqrt{\beta_0 + \beta_1 x_i} + v_i \quad \text{y} \quad |e_i| = \sqrt{\beta_0 + \beta_1 x_i^2} + v_i$$

Son no lineales en los parámetros y por lo tanto no pueden estimarse con el procedimiento corriente de MCO.

Glejser encontró que para muestras grandes los cuatro primeros modelos, entre los anteriores, dan generalmente, resultados satisfactorios para detectar la

heteroscedasticidad. En la práctica, la técnica de Glejser puede usarse entonces para muestras grandes, y para muestras pequeñas puede tomarse como un recurso cualitativo para iniciarse en los problemas de la heteroscedasticidad.

5. **Prueba de correlación de rango de Spearman**. El coeficiente de correlación de rango de Spearman se define como:

$$r_s = 1 - 6\left[\frac{\sum_{i=1}^{n} d_i^2}{n(n^2-1)}\right] \qquad \textbf{(7.49)}$$

Donde:

d_i: Diferencia en los rangos atribuida a dos características diferentes del i-ésimo individuo o fenómeno.

n: Número de individuos o fenómenos clasificados.

Puede emplearse este coeficiente de correlación de rango para detectar la heteroscedasticidad de la siguiente manera: suponga que $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Etapa I: Con la información ajuste la regresión de "y" contra "x" y obtenga los residuos e_i.

Etapa II: Ignorando el signo de e_i, es decir, tomando su valor absoluto, ordene tanto $|e_i|$ como x_i en forma ascendente o descendente y calcule el coeficiente de correlación de rango de Spearman dado anteriormente.

Etapa III: Suponiendo que el coeficiente de correlación de rango de la población ρ_s es cero, y n > 8, la significancia del coeficiente de correlación de rango muestral r_s puede verificarse con la prueba t de la manera siguiente:

$$t = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}} \qquad \textbf{(7.50)}$$

Con n - 2 grados de libertad.

Si el valor calculado de t es mayor que el valor critico de t, podemos aceptar la hipótesis de heteroscedasticidad; si no, debemos rechazarla. Si el modelo de regresión contiene más de una variable "x", r_s puede calcularse entre $|e_i|$ y cada una de las "x" por separado y puede verificarse, en cada caso, para ver su significancia estadística por medio de la prueba t.

Ejemplo 5:

Se requiere la estimación de la línea del mercado de capitales de la teoría del portafolio. Dado que la información se relaciona con 10 fondos mutuos de diferentes tamaños y objetivos de inversión, a priori se puede esperar que hay heteroscedasticidad. En la tabla 7.15 se muestran los valores para la variable "y" (rendimiento anual promedio %), "x" (desviación estándar del rendimiento anual %), el valor absoluto de los residuos, el rango de la variable "x", rango del valor absoluto de los residuos, las diferencias y las diferencias al cuadrado.

Tabla 7.15 Datos para el ejemplo 5.

y (%)	x (%)	$\|e_i\|$	Rango de x	Rango de $\|e_i\|$	d	d^2
12.4	12.1	1.017	4	9	-5	25
14.4	21.4	1.260	9	10	-1	1
14.6	18.7	0.181	7	4	3	9
16.0	21.7	0.202	10	5	5	25
11.3	12.4	0.221	5	6	-1	1
10.0	10.4	0.602	2	7	-5	25
16.2	20.6	0.908	8	8	0	0
10.4	10.2	0.110	1	3	-2	4
13.1	16.0	0.077	6	2	4	16
11.3	12	0.037	3	1	2	4
				Suma	0	110

Con los datos de la tabla anterior calculamos el coeficiente de correlación de Spearman:

$$r_s = 1 - 6\left[\frac{\sum_{i=1}^{n} d_i^2}{n(n^2-1)}\right]$$

$$r_s = 1 - 6\left[\frac{110}{10(100-1)}\right]$$

$$r_s = 1 - 6(0.11111111)$$

$$r_s = 0.33333$$

La significancia del coeficiente de correlación de rango muestral r_s puede verificarse con la prueba t de la manera siguiente:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

$$t = \frac{0.2333\sqrt{10-2}}{\sqrt{1-(0.2333)^2}}$$

$$t = \frac{(0.2333)*(2.8284)}{\sqrt{1-0.0544}}$$

$$t = 0.99998$$

Con 10 – 2 = 8 grados de libertad este valor de t no es significativo inclusive a un nivel de significancia del 10%. De esta forma, no hay evidencia de una relación sistemática entre la variable independiente y los valores absolutos de los residuos, lo que puede sugerir que no hay heteroscedasticidad.

7.7.3 Medidas Remediales.

Como hemos visto, la heteroscedasticidad no destruye las propiedades de insesgamiento y de consistencia de los estimadores de MCO, pero ya no son eficientes, ni siquiera asintóticamente (es decir, en muestras grandes). Esta falta de eficiencia le resta credibilidad al procedimiento de la prueba de hipótesis. Por esto son necesarias las medidas remediales. Existen dos enfoques para remediar la heteroscedasticidad:

- Cuando se conoce σ_i^2.
- Cuando no se conoce σ_i^2.

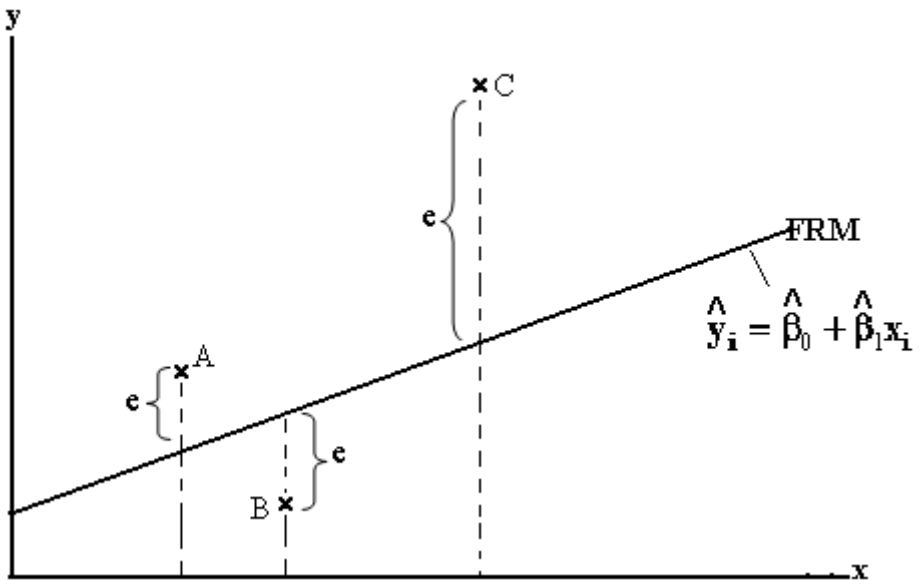
7.7.3.1 Cuando se conoce σ_i^2: Método de Mínimos Cuadrados Ponderados.

Cuando se conoce σ_i^2 o se puede estimar, el método más sencillo de tratar la heteroscedasticidad es el de mínimos cuadrados ponderados. Para ilustrar este método consideramos el modelo de dos variables:

$$\text{FRP}: y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$
$$\text{FRM}: y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$$

El método usual, no ponderado, consiste en minimizar SS_{Res}: $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$ con respecto a las incógnitas. Al minimizar esta SS_{Res}, el método MCO da implícitamente la misma ponderación a cada e_i^2. Por esto, en el diagrama hipotético de la figura 7.12 los puntos A, B y C tienen el mismo peso en el cálculo de $\sum_{i=1}^{n} e_i^2$. Se puede ver, que en este caso los e_i^2 asociados con el punto C dominaran la SS_{Res}.

Figura 7.12 Diagrama hipotético.



El método de los mínimos cuadrados ponderados toma en cuenta puntos extremos, como por ejemplo C en la figura 7.12, por minimización, no el ponderado usual SS_{Res}, si no el siguiente SS_{Res}:

$$\min : \sum_{i=1}^{n} w_i e_i^2 = \sum_{i=1}^{n} w_i (y_i - \beta_0^* - \beta_1^* x_i)^2 \qquad \textbf{(7.51)}$$

Donde:

w_i: Las ponderaciones, son ciertos números constantes (no estocásticos).

β_0^* y β_1^*: Son los estimadores de mínimos cuadrados ponderados.

Los w_i se escogen de tal manera que las observaciones extremas (por ejemplo C en la figura 7.12) reciban menor ponderación. Si σ_i^2 se conoce podemos tener:

$$w_i = \frac{1}{\sigma_i^2} \qquad \textbf{(7.52)}$$

Es decir, ponderar cada observación de manera inversamente proporcional a σ_i^2. Este sistema de ponderación "descuenta" observaciones muy pesadas provenientes de poblaciones con varianzas muy grandes, tales como el punto C de la figura 7.12.

La mecánica de minimizar (7.51) sigue los métodos usuales del cálculo, las ecuaciones son las siguientes:

$$\beta_0^* = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} - \beta_1^* \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad \textbf{(7.53)}$$

$$\beta_0^* = \bar{y}^* - \beta_1^* \bar{x}^*$$

Donde $\bar{y}^*$ y $\bar{x}^*$ son medias muestrales ponderadas con ε_i como ponderación y

$$\beta_1^* = \frac{\sum_{i=1}^{n} w_i (x_i - \bar{x}^*)(y_i - \bar{y}^*)}{\sum_{i=1}^{n} w_i (x_i - \bar{x}^*)^2} \qquad \textbf{(7.54)}$$

Se puede observar que si $w_1 = w_2 = \ldots = w_n$, es decir, si cada observación tiene el mismo peso, los estimadores de mínimos cuadrados ponderados, dados anteriormente, coinciden con los estimadores de MCO.

7.7.3.2 Cuando no se conoce σ_i^2.

En los estudios econométricos, el conocimiento previo de σ_i^2 es muy poco común, por lo que el método de mínimos cuadrados ponderados visto anteriormente no puede usarse tan sencillamente. En la práctica, por lo tanto, debemos recurrir a algunos supuestos ad hoc, aunque razonablemente plausibles, sobre σ_i^2 y transformar el modelo de regresión original de tal manera que satisfaga el supuesto de homoscedasticidad. Sin una transformación de este tipo el problema de heteroscedasticidad se torna prácticamente insoluble. A continuación presentamos algunas de esas transformaciones, con la ayuda del modelo de dos variables:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Consideramos algunos supuestos posibles sobre el patrón de heteroscedasticidad.

Supuesto 1.

$$E(\varepsilon_i^2) = \sigma^2 x_i^2 \qquad \textbf{(7.55)}$$

Si como producto de la "especulación", de los métodos gráficos o de los enfoques de Park y Glejser se cree que la varianza de ε_i es proporcional al cuadrado de la variable independiente "x", podemos transformar el modelo original de la siguiente manera. Dividiendo todo el modelo original por x_i:

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 \frac{x_i}{x_i} + \frac{\varepsilon_i}{x_i}$$
$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + v_i \qquad \textbf{(7.56)}$$

Donde v_i es el término de perturbación transformado y es igual a ε_i / x_i. Ahora es fácil verificar que:

$$E(\varepsilon_i^2) = E\left(\frac{\varepsilon_i}{x_i}\right)^2$$
$$E(\varepsilon_i^2) = \frac{1}{x_i^2} E(\varepsilon_i^2)$$
$$E(\varepsilon_i^2) = \frac{1}{x_i^2} (\sigma^2 x_i^2)$$
$$E(\varepsilon_i^2) = \sigma^2$$

Por lo tato, la varianza de v_i es homoscedástica y podemos proceder a aplicar MCO a la ecuación transformada (7.56), estimando la regresión de y_i / x_i contra $1/x_i$.

En la regresión transformada el intercepto β_1 es la pendiente de la ecuación original y la pendiente β_0 es el intercepto del modelo original. Por lo que para volver al modelo original hay que multiplicar (7.56) por x_i.

Supuesto 2.

$$E(\varepsilon_i^2) = \sigma^2 x_i \qquad \textbf{(7.57)}$$

Si se cree que la varianza de ε_i en lugar de ser proporcional al cuadrado de x_i es proporcional a x_i el modelo original puede transformarse en:

$$\frac{y_i}{\sqrt{x_i}} = \frac{\beta_0}{\sqrt{x_i}} + \beta_1\sqrt{x_i} + \frac{\varepsilon_i}{\sqrt{x_i}}$$
$$\frac{y_i}{\sqrt{x_i}} = \beta_0\frac{1}{\sqrt{x_i}} + \beta_1\sqrt{x_i} + v_i \qquad \textbf{(7.58)}$$

Donde $x_i > 0$.

Dado el supuesto 2, se puede verificar que $E(v_i^2) = \sigma^2$, situación homoscedástica y, por consiguiente podemos proceder a aplicar MCO a (7.58) haciendo la regresión de $y_i/\sqrt{x_i}$ contra $1/\sqrt{x_i}$ y $\sqrt{x_i}$.

Supuesto 3.

$$E(\varepsilon_i^2) = \sigma^2[E(y_i)]^2 \qquad \textbf{(7.59)}$$

La ecuación (7.53) postula que la varianza de ε_i es proporcional al cuadrado del valor esperado de "y" (ver figura 7.10e). Ahora,

$$E(y_i) = \beta_0 + \beta_1 x_i$$

Por consiguiente, si transformamos la ecuación original de la siguiente manera:

$$\frac{y_i}{E(y_i)} = \frac{\beta_0}{E(y_i)} + \frac{\beta_1 x_i}{E(y_i)} + \frac{\varepsilon_i}{E(y_i)}$$
$$\frac{y_i}{E(y_i)} = \frac{\beta_0}{E(y_i)} + \frac{\beta_1 x_i}{E(y_i)} + v_i \qquad \textbf{(7.60)}$$

Donde $v_i = \varepsilon_i / E(y_i)$, se podrá mostrar que $E(v_i^2) = \sigma^2$, es decir, las perturbaciones v_i son homoscedásticas y, por lo tanto la regresión de (7.60) satisface el supuesto de homoscedasticidad del modelo de regresión lineal clásico.

La transformación de (7.60) es, sin embargo, inoperante pues la $E(y_i)$ depende de β_0 y β_1 que son desconocidas. Lógicamente conocemos $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ que es el valor estimado de $E(y_i)$ y podemos proceder en dos etapas:

Primero hacemos la regresión normal MCO sin tener en cuenta el problema de heteroscedasticidad y obtenemos $\hat{y}_i$. Luego, usando $\hat{y}_i$ transformamos el modelo de la siguiente manera:

$$\frac{y_i}{\hat{y}_i} = \frac{\beta_0}{\hat{y}_i} + \frac{\beta_1 x_i}{\hat{y}_i} + \frac{\varepsilon_i}{\hat{y}_i}$$
$$\frac{y_i}{\hat{y}_i} = \frac{\beta_0}{\hat{y}_i} + \frac{\beta_1 x_i}{\hat{y}_i} + v_i \qquad \textbf{(7.61)}$$

Donde $v_i = \varepsilon_i / \hat{y}_i$.

En la segunda etapa hacemos la regresión (7.61). Aunque $\hat{y}_i$ no son exactamente $E(y_i)$, son estimadores consistentes, es decir, a medida que el tamaño de la muestra aumenta indefinidamente, convergen al verdadero valor $E(y_i)$. Por esto la transformación (7.61) funcionará en la práctica si el tamaño de la muestra es razonablemente grande.

Supuesto 4. Transformación Logarítmica.

Si en lugar de correr la regresión $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ corremos:

$$\ln y_i = \beta_0 + \beta_1 \ln x_i + \varepsilon_i \qquad \textbf{(7.62)}$$

Se reduce frecuentemente la homoscedasticidad. Esto se debe a que la transformación logarítmica comprime las escalas en que están medidas las variables, reduciendo una

diferencia de 10 veces en una de 2 veces. El número 80 es diez veces el número 8, pero ln 80 = 4.3820 es sólo dos veces más grande que ln 8 = 2.0794.

Una ventaja más de la transformación logarítmica es que el coeficiente de la pendiente β_1 mide la elasticidad de "y" con respecto a "x", es decir, el cambio porcentual en "y" debido a un cambio porcentual en "x". Por ejemplo, si "y" es consumo y "x" ingreso, β_1 en la ecuación (7.62) medirá la elasticidad de ingreso, mientras que en el modelo original β_1 mide sólo la tasa de cambio del consumo medio por una unidad de cambio en el ingreso. Por esta razón los modelos logarítmicos son tan populares en la econometría empírica.

Para concluir la discusión sobre las medidas remediales se debe enfatizar el hecho de que todas las transformaciones vistas anteriormente son ad hoc. Se está especulando esencialmente sobre la naturaleza de σ_i^2. ¿Cuál de las transformaciones expuestas dependerá de la naturaleza del problema y de la severidad de la heteroscedasticidad? Existen algunos problemas adicionales en relación con las transformaciones vistas. Por ejemplo, cuando vamos más allá del modelo de dos variables, no sabemos a priori cual de las variables "x" debe transformarse[6]. Surge entonces un problema de correlación espuria.

Esta expresión, debida a Park, se refiere a una situación en la que existe correlación entre las razones de variables (x_1/x_2), aunque las variables originales no estén

[6] No obstante, en el caso práctico, podemos dibujar e_i^2 contra cada variable y decidir que variable "x" puede usarse para transformar los datos (ver figura 7.11).

correlacionadas, o sean aleatorias[7]. En el modelo $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, "y" y "x" pueden no estar correlacionadas, pero en el modelo transformado $y_i / x_i = \beta_0 / x_i + \beta_1 + v_i$, y_i / x_i y $1/x_i$ si lo están, por lo general.

7.8 Autocorrelación.

Otro de los supuestos importantes del modelo de regresión lineal es el de que no existe autocorrelación o correlación serial entre las perturbaciones ε_i que entran en la función de regresión poblacional.

La dependencia entre las perturbaciones del modelo de regresión es un problema frecuente cuando las variables que estudiamos se observan a lo largo del tiempo como una serie temporal. Entonces, es esperable que todas las variables que influyen sobre la variable respuesta tengan estructura temporal y, por lo tanto las perturbaciones (que recogen el efecto de las variables omitidas) tendrán dependencia temporal. Por ejemplo, si estudiamos las ventas anuales de un producto en función del precio y de los gastos en publicidad, la perturbación sintetizará los efectos de los gustos de los consumidores, de las decisiones de la competencia, de la evolución del consumo, etc. Todas estas variables se modifican a lo largo del tiempo y, por tanto, las perturbaciones de años consecutivos serán probablemente, dependientes.

[7] Por ejemplo si x_1, x_2 y x_3 no están mutuamente correlacionadas $r_{12} = r_{13} = r_{23} = 0$ y encontramos que (los valores de) las razones x_1/x_3 y x_2/x_3 están correlacionadas, entonces hay correlación espuria. "de manera más general, la correlación se denomina espuria si es inducida al manipular los datos y no existe en la información original".

El término autocorrelación puede definirse como la "correlación existente entre los miembros de una serie de observaciones ordenadas en el tiempo (como las cifras de series de tiempo) o en el espacio (como las cifras de corte transversal)".

En el contexto de la regresión, el modelo de regresión lineal clásico supone que dicha autocorrelación no existe en las perturbaciones ε_i. Simbólicamente:

$$E(\varepsilon_i \varepsilon_j) = 0 \qquad i \neq j$$

Sencillamente, el modelo clásico supone que el término de perturbación perteneciente a una observación no está influenciado por el término de perturbación perteneciente a otra Por ejemplo, si tratamos con series de tiempo trimestrales sobre la regresión de la producción contra los insumos de capital y trabajo y de pronto se presenta una huelga o paro laboral que afecta la producción en un trimestre, no existen razones para pensar que esta interrupción se extienda al siguiente trimestre. Es decir, si la producción es baja este trimestre no hay razón para pensar que sea más baja en el siguiente. Igualmente si se trata de cifras de corte transversal sobre la regresión de los gastos de consumo de una familia contra su ingreso, el efecto de un aumento en el ingreso de una familia sobre su consumo no tiene por qué verse afectado por el gasto de consumo de otra familia.

Sin embargo, si existe dicha dependencia, tendríamos autocorrelación. Simbólicamente:

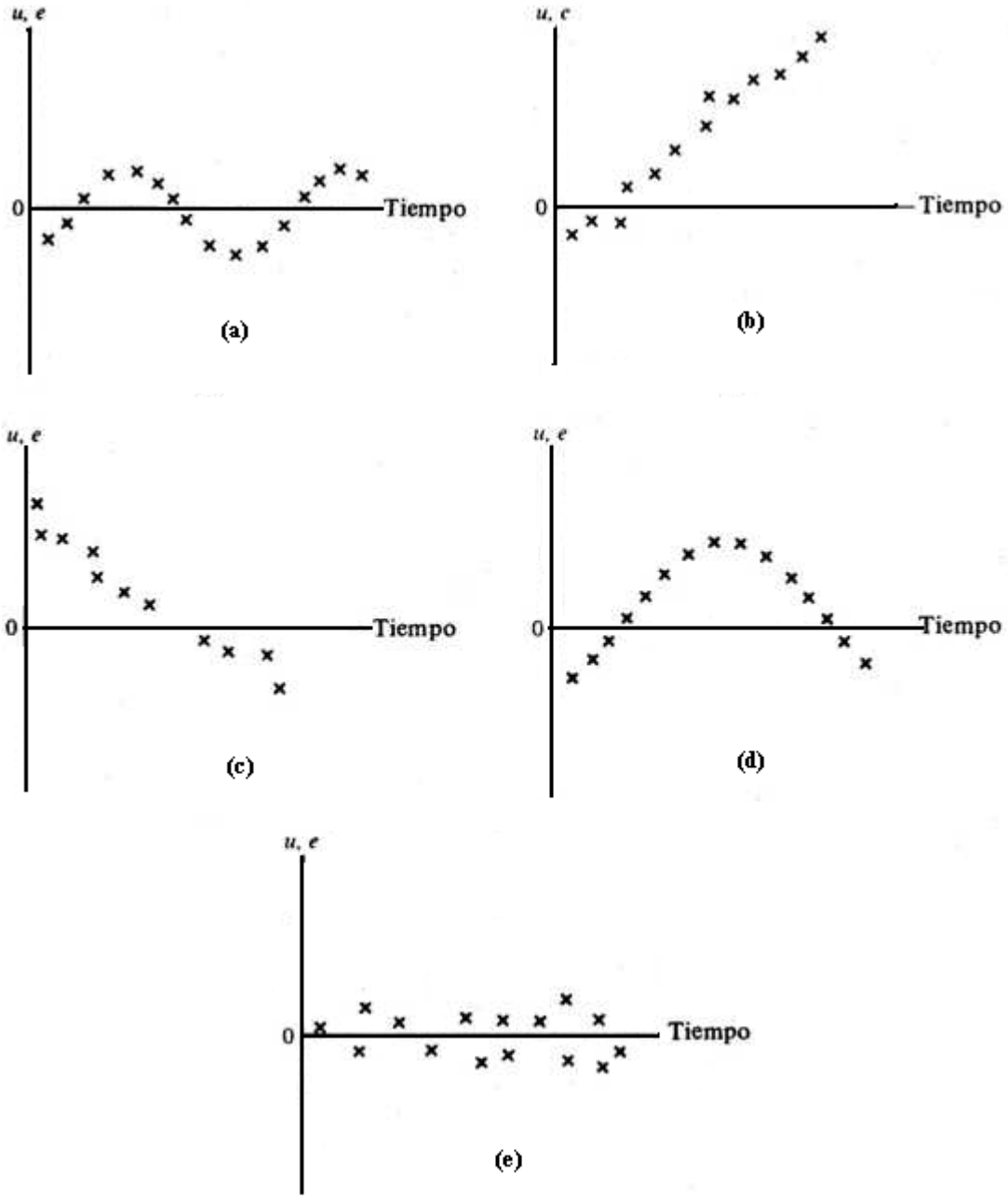
$$E(\varepsilon_i \varepsilon_j) \neq 0 \qquad i \neq j \qquad \textbf{(7.63)}$$

En tal situación, la interrupción causada por la huelga en un trimestre puede afectar la producción del siguiente trimestre, o los aumentos en los gastos de consumo de una familia pueden motivar a otra familia a aumentar los suyos, por el deseo de no quedarse atrás.

Antes de averiguar por qué existe la autocorrelación es indispensable aclarar el aspecto relativo a la terminología. Aunque hoy en día es común el empleo de los términos autocorrelación y correlación serial como sinónimos, algunos autores prefieren hacer distinción entre los dos términos. Tintner, por ejemplo, define la autocorrelación como una "correlación de una serie con rezago consigo misma, rezagada un cierto número de unidades de tiempo" mientras que reserva el término correlación serial para una "correlación rezagada entre dos series diferentes". Por lo tanto, la correlación entre dos series de tiempo como $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{10}$ y $\varepsilon_2, \varepsilon_3, \ldots, \varepsilon_{11}$ donde la primera es igual a la segunda retrasada un período es autocorrelación mientras que la correlación entre las series de tiempo tales como $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{10}$ y $v_2, v_3, \ldots, v_{11}$, donde ε y v son dos series de tiempo diferentes se llama correlación serial. Pero aunque en algún contexto pueda ser útil la distinción entre los términos, en este documento se utilizarán como sinónimos.

Puede resultar interesante ver gráficamente algunos de los posibles patrones de autocorrelación y de no autocorrelación que se muestran en la figura 7.13, que en su parte (a) muestra un patrón cíclico mientras que la (b) y la (c) sugieren una tendencia lineal en las perturbaciones hacia arriba y hacia abajo, y la parte (d) indica que tanto la tendencia lineal como cuadrática está presente en las perturbaciones. Sólo la figura 7.13(e) indica un patrón no sistemático, respaldando el supuesto de no autocorrelación del modelo de regresión lineal clásico.

Figura 7.13 Patrones de autocorrelación.



Obviamente debemos preguntarnos ahora: ¿por qué ocurre la correlación serial?

Existen varias razones; veamos algunas:

1. Inercia. Una de las características más importantes de la mayoría de las series estadísticas de tiempo es la inercia o "inactividad". Como es bien sabido, las series de tiempo como el PNB (Producto Nacional Bruto), los índices de precios, la producción, el empleo y el desempleo presentan ciclos (económicos). Partiendo del fondo de la recesión, cuando comienza la recuperación económica, la mayoría de estas series empieza a moverse hacia arriba; en este ciclo ascendente, el valor de la serie en un punto del tiempo es mayor que su valor previo; entonces, hay un "impulso" en la serie que continúa hasta que sucede algo (por ejemplo, un aumento en la tasa de interés, en los impuestos o en ambas cosas) que los hace descender lentamente. Finalmente, en las regresiones de cifras sobre series de tiempo es muy probable que las observaciones sucesivas sean interdependientes.
2. Sesgo de especificación: el caso de las variables excluidas. En el análisis empírico es común que el investigador comience con un modelo de regresión que puede ser aceptable pero no "perfecto". Después de analizar la regresión, el investigador realiza el examen posterior para ver si los resultados están de acuerdo con lo que se espera, si no, para recurrir a una solución extrema. Por ejemplo, el investigador puede expresar gráficamente los residuos e_i obtenidos a partir de la regresión ajustada y observar si se presentan patrones como los que se muestran en las figuras 7.13(a) a (d). Estos residuos (que son aproximaciones de ε_i) pueden sugerir que algunas de las variables que originalmente pretendían incluirse en el modelo, pero que fueran excluidas, deben ahora excluirse. Este es

el caso del sesgo de especificación con variables excluidas. Frecuentemente ocurre que al incluir estas variables, desaparece el patrón de correlación observado entre los residuos. Por ejemplo, supongamos el siguiente modelo de demanda:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t \qquad \textbf{(7.64)}$$

Donde:

y: Cantidad demandada de carne de res.

x_1: Precio de la carne de res.

x_2: Ingreso del consumidor

x_3: Precio de la carne de cerdo.

t: Tiempo[8].

Sin embargo, por alguna razón hemos corrido la siguiente regresión:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + v_t \qquad \textbf{(7.65)}$$

Ahora, si la ecuación (7.64) es el modelo "correcto" o verdadera relación, correr (7.65) equivale a decir que $v_t = \beta_3 x_{3t} + \varepsilon_t$, y en la medida en que el precio de la carne de cerdo afecte el consumo de carne de res, el término de error o perturbación v_t reflejará un patrón sistemático, creando por consiguiente (una falsa) autocorrelación. Una prueba sencilla de lo anterior sería correr tanto (7.64)

[8] Por convención, se utiliza el subíndice t para series de tiempo e i para cifras de corte transversal.

como (7.65) y ver si en caso de autocorrelación en (7.65), ésta desaparece cuando se corre (7.64)[9].

3. Sesgo de especificación: Forma funcional incorrecta. Suponga que el modelo verdadero o "correcto" en un estudio sobre costos y producción es como sigue:

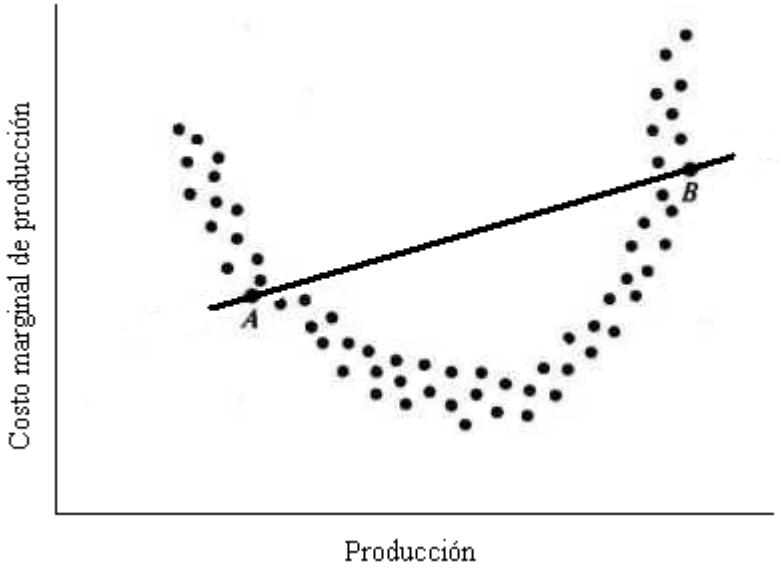
$$\text{Costo marginal}_i = \beta_0 + \beta_1\,(\text{producción}_i) + \beta_2\,(\text{producción}_i)^2 + \varepsilon_i \qquad \textbf{(7.66)}$$

Pero nosotros ajustamos el siguiente modelo:

$$\text{Costo marginal}_i = \alpha_0 + \alpha_1\,(\text{producción}_i) + v_i \qquad \textbf{(7.67)}$$

La curva de costos marginales que corresponde al "verdadero" modelo se muestra en la figura 7.14, así como la curva lineal "incorrecta".

Figura 7.14 Sesgo de especificación, forma funcional incorrecta.



Como se observa en la figura 7.14, entre los puntos A y B la curva lineal de costo marginal sobreestimará consistentemente el verdadero costo marginal, mientras

[9] Si se encuentra que el verdadero problema es el de un sesgo de especificación y no de autocorrelación, los estimadores de MCO de los parámetros (7.65) pueden ser sesgados e inconsistentes.

que por detrás de estos puntos subestimará consistentemente el costo marginal. Esto es de esperarse en razón de que el término de perturbación v_i es realmente igual a la (producción)^2 + ε_i, y por lo tanto capta el efecto sistemático del término (producción)^2 sobre el costo marginal. En este caso v_i, reflejará la autocorrelación por haber utilizado una forma funcional incorrecta.

4. El fenómeno de la telaraña. La oferta de muchos bienes agrícolas refleja el llamado "fenómeno de la telaraña", que consiste en que la oferta reacciona ante el precio con un rezago de un período de tiempo porque se requiere cierto tiempo para implementar las decisiones de la oferta (el periodo de gestación). De tal manera que al comienzo de la cosecha de un año, los granjeros están influenciados por el precio prevaleciente el año anterior de suerte que su función de oferta será:

$$\text{Oferta}_t = \beta_0 + \beta_1 P_{t-1} + \varepsilon_t \qquad \textbf{(7.68)}$$

Suponga que al final del período t, el precio P_t resulta ser más bajo que P_{t-1}. Con lo cual en el período t + 1, los granjeros pueden decidirse a producir menos de lo que produjeron en el período t. Obviamente en esta situación no se espera que las perturbaciones ε_i sean aleatorias porque si los granjeros sobreproducen en el año t, es muy probable que reduzcan su producción en t + 1, y así sucesivamente, creando así un patrón de tipo telaraña.

5. Rezagos. No es extraño encontrar, en una regresión de gastos de consumo contra el ingreso, que los gastos de consumo en determinado periodo dependen entre otras cosas de los gastos de consumo en el periodo anterior. Es decir:

$$\text{Consumo}_t = \beta_0 + \beta_1 \text{ingreso}_t + \beta_2 \text{consumo}_{t-1} + \varepsilon_t \qquad \textbf{(7.69)}$$

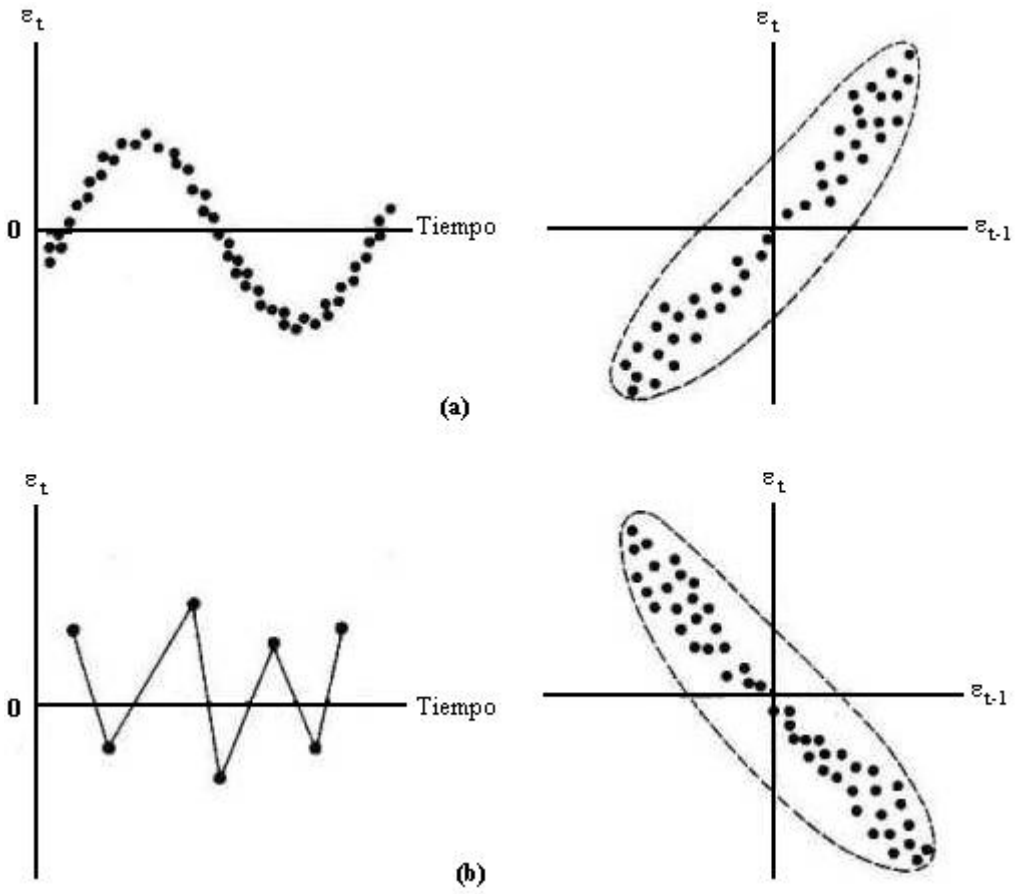
La regresión como la que se da en la ecuación (7.69) se conoce con el nombre de autorregresión, justamente porque una de las variables explicatorias es el valor retrasado o sesgado de la variable dependiente. La justificación teórica de un modelo como el de (7.69) resulta simple, ya que los consumidores no cambian muy a menudo sus hábitos de consumo por razones psicológicas, tecnológicas e institucionales. Ahora, si dejamos de lado el término rezagado en (7.69), el error resultante reflejará un patrón sistemático, debido a la influencia del consumo rezagado sobre el consumo corriente.

6. "Manipulación" de datos. En el análisis empírico comúnmente se manipulan los datos básicos; por ejemplo, en las regresiones de series temporales trimestrales, estas se derivan a partir de los datos mensuales, mediante la simple adición de las cifras de 3 meses y luego dividiendo por 3. Este procedimiento de promediar las cifras permite uniformarlas, eliminando las fluctuaciones mensuales que ofrezcan. Por lo tanto, un gráfico que contenga cifras trimestrales debe ser más uniforme que uno que contenga cifras mensuales, uniformidad que puede llevar a un patrón sistemático en las perturbaciones, introduciendo de este modo la autocorrelación. Otra forma de manipulación es la interpolación y extrapolación de cifras; por ejemplo, el censo de población se lleva a cabo cada 10 años (en EEUU); el último se hizo en 2000 y el anterior en 1990; entonces si hay necesidad de obtener datos de un año comprendido en el período intercensal - 1990-2000, se recurre comúnmente a la interpolación con base en algunos

supuestos ad hoc. En general todas estas técnicas que emparejan las cifras suelen introducir patrones sistemáticos que normalmente no existen en los datos originales.

Debe tenerse en cuenta además que la autocorrelación puede ser positiva o negativa; se presenta con más frecuencia la positiva debido a que la mayoría de las series económicas se mueven hacia arriba o hacia abajo todo el tiempo y no con movimientos ascendentes-descendentes como los que se muestran en la figura 7.15(b).

Figura 7.15 Autocorrelación (a) positiva y (b) negativa.



7.8.1 Consecuencias de la Autocorrelación.

Recordemos que si todos los supuestos del modelo de regresión clásico se cumplen, el teorema de Gauss-Markov afirma que entre todos los estimadores lineales los estimadores de MCO son los mejores, es decir tienen la mínima varianza; en resumen, son eficientes. Si mantenemos ahora todos los supuestos del modelo clásico, excepto el de no autocorrelación, los estimadores de MCO tendrán entonces las siguientes propiedades:

1. Son insesgados, es decir, en muestras repetidas (condicionales a los valores fijos de "x") sus valores medios son iguales a los verdaderos valores poblacionales.
2. Son consistentes, o sea que a medida que el tamaño de la muestra crece indefinidamente, se aproximan a los verdaderos valores.
3. Como en el caso de heteroscedasticidad, ya no son eficientes (mínima varianza) ni para muestras pequeñas ni para muestras grandes.

Por consiguiente, si persistimos en aplicar MCO en situaciones de autocorrelación tendremos las siguientes consecuencias:

1. Aunque tengamos en cuenta la correlación serial en los estimadores comunes de MCO y sus varianzas, los estimadores serán aun ineficientes (comparados con los mejores estimadores lineales insesgados). Por lo tanto, los intervalos de confianza serán más anchos de lo necesario y la prueba de significancia menos fuerte.
2. Si olvidamos por completo el problema de la autocorrelación y seguimos aplicando las fórmulas clásicas de MCO (derivadas bajo el supuesto de no

autocorrelación), las consecuencias serán todavía más serias, por las siguientes razones:

a) La varianza residual $\hat{\sigma}^2$ tiende a subestimar la verdadera σ^2.

b) Incluso si σ^2 no está subestimada, las varianzas y los errores estándar de los estimadores MCO tienden a subestimar las verdaderas varianzas y errores estándar.

c) Las pruebas usuales de significación t y F ya no son válidas y si se aplican tienden a dar conclusiones erróneas acerca de la significación estadística de los coeficientes de regresión estimados.

3. Aunque los estimadores de MCO sean insesgados, lo cual es una propiedad de muestras repetidas, para una muestra en particular tienden a dar una visión distorsionada de los verdaderos valores poblacionales. En otras palabras, los estimadores de MCO se vuelven sensibles a las fluctuaciones muestrales.

Para concretar algunas de las proposiciones anteriores, volvamos al modelo con dos variables:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \qquad \textbf{(7.70)}$$

Donde t denota la observación en el tiempo t. Ahora, para poder continuar, debemos suponer algún mecanismo que genere los ε_t, lo cual es inevitable dado que ε_t no es observable. Como punto de partida, podemos suponer que las perturbaciones se generan de la siguiente forma:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t \qquad -1 \leq \rho \leq +1 \qquad \textbf{(7.71)}$$

Donde ρ se conoce como el coeficiente de autocovarianza y donde ν_t es una perturbación estocástica de tal forma que satisface todos los supuestos de MCO, siendo éstos:

$$\begin{aligned} E(\nu_t) &= 0 \\ \text{var}(\nu_t) &= \sigma^2 \\ \text{cov}(\nu_t, \nu_{t+s}) &= 0 \qquad s \neq 0 \end{aligned} \tag{7.72}$$

El esquema (7.71) se conoce como el esquema autorregresivo de primer orden de Markov, o simplemente un esquema autorregresivo de primer orden. El término autorregresivo resulta apropiado porque (7.71) puede interpretarse como la regresión de ε_t contra si mismo, retrasado un período. Es de primer orden pues sólo entran en el modelo ε_t y un valor inmediatamente anterior. Si el modelo fuera $\varepsilon_t = \rho\varepsilon_{t-2} + \nu_t$, seria un esquema autorregresivo de segundo orden, y así sucesivamente. Debe anotarse que el coeficiente de autocovarianza puede también interpretarse como el coeficiente de autocorrelación de primer orden o, más precisamente, el coeficiente de autocorrelación de 1 rezago. Este nombre se explica de la siguiente manera:

Por definición el coeficiente (poblacional) de correlación entre ε_t y ε_{t-1} es:

$$\rho = \frac{E\{[\varepsilon_t - E(\varepsilon_t)][\varepsilon_{t-1} - E(\varepsilon_{t-1})]\}}{\sqrt{\text{var}(\varepsilon_t)}\sqrt{\text{var}(\varepsilon_{t-1})}}$$

$$\rho = \frac{E(\varepsilon_t \varepsilon_{t-1})}{\text{var}(\varepsilon_{t-1})}$$

Dado que $\varepsilon_t = 0$ para cada t y $\text{var}(\varepsilon_t) = \text{var}(\varepsilon_{t-1})$ ya que mantenemos el supuesto de homoscedasticidad.

Lo que la ecuación (7.71) plantea es que el movimiento o cambio en ε_t, se compone de dos partes: una parte $\rho\varepsilon_{t-1}$ que capta un cambio sistemático y otra que es puramente aleatoria.

Con el esquema autorregresivo de primer orden se tiene que[10]:

$$\text{var}(\beta_1^*) = \frac{\sigma^2}{\sum_{t=1}^{n}(x_t - \bar{x}_t)^2}\left(1 + \rho\frac{\sum_{t=1}^{n-1}(x_t - \bar{x}_t)(x_{t+1} - \bar{x}_{t+1})}{\sum_{t=1}^{n}(x_t - \bar{x}_t)^2} + \ldots + 2\rho^{n-1}\frac{(x_1 - \bar{x}_1)(x_n - \bar{x}_n)}{\sum_{t=1}^{n}(x_t - \bar{x}_t)^2}\right) \quad \textbf{(7.73)}$$

Donde $\text{var}(\beta_1^*)$ es la varianza del estimador usual de MCO bajo correlación serial (de primer orden). Es importante anotar que $\text{var}(\beta_1^*)$ no es aún la mínima pues:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

Ya no es el mejor estimador lineal insesgado. Suponiendo que un esquema autorregresivo de primer orden, el mejor estimador lineal insesgado de β_1 llamémoslo b_1, está dado por:

$$b_1 = \frac{\sum_{t=1}^{n}\left((x_t - \bar{x}_t) - \rho(x_{t-1} - \bar{x}_{t-1})(y_t - \bar{y}_t) - \rho(y_{t-1} - \bar{y}_{t-1})\right)}{\sum_{t=1}^{n}\left((x_t - \bar{x}_t) - \rho(x_{t-1} - \bar{x}_{t-1})\right)^2} + C$$

[10] No se presentan detalles de esta ecuación dado que se trata de series temporales.

$$\text{var}(b_1) = \frac{\sigma^2}{\sum_{t=1}^{n} \left[(x_t - \bar{x}_t) - \rho(x_{t-1} - \bar{x}_{t-1})\right]^2} + D$$

Donde C y D son factores de corrección que pueden descartarse en la práctica.

En contraste, la fórmula usual (homoscedástica) para la varianza del estimador MCO es:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} = \frac{\sigma^2}{\sum_{t=1}^{n} (x_t - \bar{x}_t)^2} \qquad \textbf{(7.74)}$$

Comparando (7.73) con (7.74) vemos claramente que la primera excluye todo, menos el primer término localizado antes del paréntesis de (7.73). Ahora, si ρ es positivo (lo que ocurre en la mayoría de series económicas) y las "x" están positivamente correlacionadas (también cierto en la mayoría de series), entonces es evidente que:

$$\text{var}(\hat{\beta}_1) < \text{var}(\beta_1^*) \qquad \textbf{(7.75)}$$

es decir, la varianza usual de MCO de $\hat{\beta}_1$ subestimará su verdadera varianza (bajo correlación serial de primer orden). Por lo tanto bajo las condiciones supuestas debemos utilizar $\text{var}(\beta_1^*)$ y no $\text{var}(\hat{\beta}_1)$.

Si utilizamos $\text{var}(\hat{\beta}_1)$, estaremos inflando la precisión (es decir, subestimando el error estándar) del estimador $\hat{\beta}_1$ y por consiguiente al calcular la razón t como $t = \hat{\beta}_1 / \text{es}(\hat{\beta}_1)$ (bajo la hipótesis nula de que $\beta_1 = 0$) estaríamos sobreestimando el valor de t y por ende la significancia estadística del β_1 estimado. Como en el caso de la heteroscedasticidad, el mismo σ^2 puede estar subestimado. Recordemos que para el modelo de regresión lineal clásico de dos variables:

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{n} e_t^2}{n-2} \qquad \textbf{(7.76)}$$

Proporciona un estimador insesgado de σ^2; es decir, $E(\hat{\sigma}^2) = \sigma^2$. Si hay autocorrelación generada según el esquema autorregresivo de primer orden se puede mostrar que:

$$E(\hat{\sigma}^2) = \frac{\sigma^2 [n - [2/(1-\rho)] - 2\rho r]}{n-2} \qquad \textbf{(7.77)}$$

Donde $r = \sum_{t=1}^{n-1}(x_t - \bar{x}_t)(x_{t-1} - \bar{x}_{t-1}) \Big/ \sum_{t=1}^{n}(x_t - \bar{x}_t)^2$ que puede interpretarse como el coeficiente de correlación (muestral) entre los valores sucesivos de "x". Si ρ y r son positivos (supuesto aceptable en la mayoría de series económicas) es obvio que a partir de (7.77) la $E(\hat{\sigma}^2) < \sigma^2$, es decir, que la fórmula convencional de la varianza residual en promedio subestimará el verdadero σ^2. En otras palabras, $\hat{\sigma}^2$ será sesgado hacia abajo. No es necesario decir que el sesgo en $\hat{\sigma}^2$ se transmite a la $\text{var}(\hat{\beta}_1)$ porque en la práctica estimamos esta última con la ecuación $\text{var}(\hat{\beta}_1) = \hat{\sigma}^2 / \sum_{t=1}^{n}(x_t - \bar{x}_t)^2$.

7.8.2 Como Detectar la Autocorrelación.

Como se señaló en la sección 7.8.1, la autocorrelación es un problema relativamente serio que requiere el concurso de medidas remediales. Desde luego, antes de hacer algo, es necesario saber si la autocorrelación está presente en determinada situación; presentamos en esta sección algunas pruebas de correlación serial.

Método gráfico.

Recordemos que los supuestos del modelo clásico de la no autocorrelación hacen referencia a las perturbaciones poblacionales que no son directamente observables. Disponemos solamente de sus aproximaciones de los residuos, que se obtienen mediante el método de MCO. Aunque los e_i y los ε_i no son lo mismo, están relacionados, como puede verse a continuación:

Para el modelo de dos variables

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

o en forma de desviaciones

$$y_i - \bar{y} = \beta_1 (x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \qquad \textbf{(7.78)}$$

Nótese que $\bar{\varepsilon}$ y $E(\varepsilon_i)$ no son lo mismo.

Sabemos ya que

$$\begin{aligned} e_i &= (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \\ e_i &= [\beta_1 (x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})] - \hat{\beta}_1 (x_i - \bar{x}) \\ e_i &= (\beta_1 - \hat{\beta}_1)(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \end{aligned} \qquad \textbf{(7.79)}$$

Ahora

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \textbf{(7.80)}$$

Por lo tanto, reemplazando (7.80) en (7.79) obtenemos:

$$e_i = (\varepsilon_i - \bar{\varepsilon}) - (x_i - \bar{x}) \left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \qquad \textbf{(7.81)}$$

Como consecuencia, si existe algún grado de autocorrelación entre los ε_i se reflejará, en virtud de (7.81), en las e_i. Por lo tanto, podrán examinarse las e_i en busca de posibles pistas de correlación serial en las ε_i. Respecto a las series de tiempo, los e_t pueden dibujarse contra el tiempo como se muestra en la figura 7.13; y si se presentaran patrones como los de la figura 7.13(a) a (d), se podría sospechar la existencia de autocorrelación, en tanto que si se dan patrones como los de 7.13(e) de la misma figura, es posible que no la haya.

Un examen de los residuos, como el que acabamos de exponer, puede por sí solo sugerir varias formas de enfrentar el problema de la correlación serial. Por ejemplo, si los residuos presentan un patrón como el de la figura 7.13(d) se puede pensar en incluir una variable de tendencia o variable-tiempo en el modelo. Si en cambio, el patrón de residuos es como el de la figura 7.13(d) puede pensarse en incluir tanto una variable de segundo como de primer grado.

Ejemplo 6:

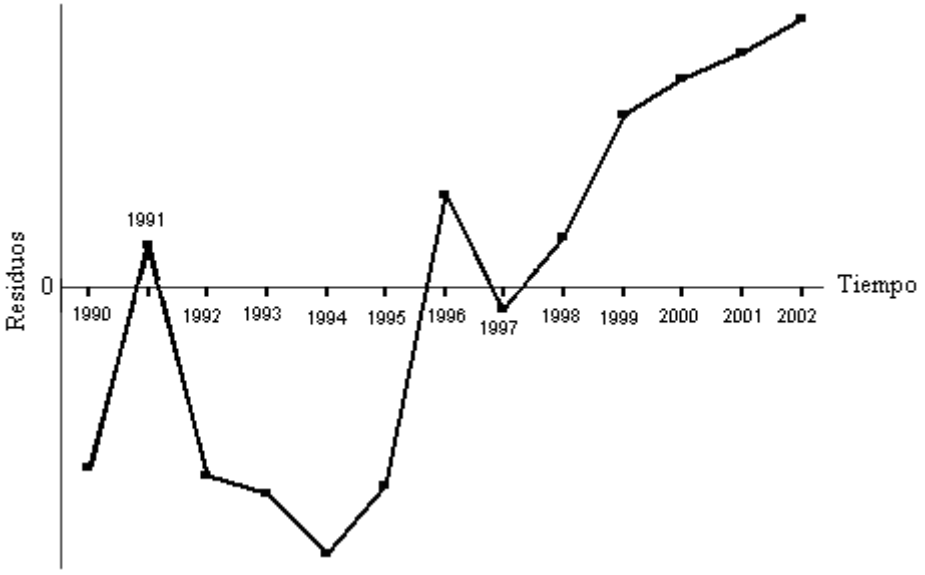
Para ilustrar el método gráfico, la tabla 7.16 nos presenta los datos donde se corre la regresión de la tasa de retiro contra la tasa de desempleo, se presentan los residuos. Dibujando los residuos contra el tiempo, en la tabla 7.16 se observa que no son aleatorios. Hasta 1994 (con excepción de 1991) los residuos son cada vez más negativos,

mientras que a partir de 1996 (con excepción de 1997) son cada vez más positivos. Tenemos pues, autocorrelación positiva en los residuos.

Tabla 7.16 Tasa de retiro y desempleo en la industria manufacturera de los EE.UU, 1990-2002 valores estimados y residuos.

Año	Tasa de retiro por cada 100 empleados, y	Tasa de desempleo (%), x	"y" estimado	Residuos, e_i
1990	1.3	6.2	1.592	-0.292
1991	1.2	7.8	1.134	0.066
1992	1.4	5.8	1.706	-0.306
1993	1.4	5.7	1.735	-0.335
1994	1.5	5.0	1.935	-0.435
1995	1.9	4.0	2.221	-0.321
1996	2.6	3.2	2.450	0.150
1997	2.3	3.6	2.336	-0.036
1998	2.5	3.3	2.422	0.078
1999	2.7	3.3	2.422	0.278
2000	2.1	5.6	1.763	0.337
2001	1.8	6.8	1.420	0.380
2002	2.2	5.6	1.763	0.437

Figura 7.16 Residuos de la regresión de la tasa de retiro contra la tasa de desempleo.



La figura 7.16 que muestra un patrón casi cíclico para los e_t, sugiere que puede introducirse en el modelo otra variable que se mueva cíclicamente con la tasa de retiro; por ejemplo, la tasa de acceso (número de nuevos alistamientos para 100 empleados), que es un indicador de la demanda de trabajo, puede tenerse en cuenta en razón de que, manteniendo constante lo demás, a mayor tasa de acceso mayor tasa de retiro.

La mayor virtud del método gráfico es su simplicidad; los residuos se pueden dibujar contra el tiempo independientemente de que el modelo tenga una o diez variables explicatorias. Existen muchos programas estadísticos (SPSS, STATISTICA, etc.) que, calculan automáticamente los residuos, incluyendo el respectivo gráfico, lo que constituye una gran ayuda visual para determinar la presencia de la autocorrelación.

Los métodos analíticos pueden sustituir al método gráfico, proporcionando una prueba estadística para establecer si el patrón no aleatorio de los e_i es estadísticamente significativo. El más reconocido de estos métodos es el de la prueba estadística Durbin-Watson.

7.8.2.1 Prueba Durbin - Watson.

El estadístico Durbin – Watson que se representa con la letra d y se define como:

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad \textbf{(7.82)}$$

o simplemente la razón de la suma de las diferencias al cuadrado de residuos sucesivos,

a la SS_{Res}. Obsérvese que en el numerador del estadístico d el número de observaciones es n - 1 por haberse perdido una de ellas al tomar las diferencias consecutivas.

Una gran ventaja del estadístico d consiste en estar basado en los residuos estimados que se calculan automáticamente en el análisis de la regresión.

1. El modelo de regresión incluye el intercepto; si no está presente como en la regresión que pasa por el origen, es indispensable volver a correr la regresión incluyendo el intercepto antes de obtener la SS_{Res}.
2. Las variables explicatorias "x", no son estocásticas, o son fijas en muestras repetidas.
3. Las perturbaciones ε_t, se generan mediante un esquema autorregresivo de primer orden: $\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t$.
4. El modelo de regresión no incluye valores rezagados de la variable dependiente como una de las variables explicatorias, por lo cual la prueba no es aplicable a modelos cómo:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + ... + \beta_k x_{kt} + \gamma y_{t-1} + \varepsilon_t \qquad \textbf{(7.83)}$$

Donde:

y_{t-1}: Es la variable "y" rezagada un período, estos modelos se llaman autorregresivos.

La distribución de probabilidad exacta del estadístico d (7.82) es difícil de encontrar, ya que, como lo han demostrado Durbin y Watson, depende en forma complicada de los valores de "x" de una muestra dada. Esto es comprensible, puesto que d es calculado con base en e_i, que a su vez depende de los "x" dados. Por consiguiente, a diferencia de las

pruebas t, F, ó χ^2 no hay un valor crítico único que nos lleve a rechazar o a aceptar la hipótesis nula de que no hay correlación serial de primer orden en las perturbaciones ε_i. No obstante, Durbin y Watson tuvieron éxito al poder encontrar un límite inferior d_L y un límite superior d_U tales que si el d calculado en (7.82) cae por fuera de estos valores críticos, puede tomarse una decisión sobre la posible presencia de correlación serial positiva o negativa. Además, estos límites dependen únicamente del número de observaciones n y del número de variables independientes y no de los valores que tomen esas variables independientes. Dichos límites para n, entre 15 y 100 y hasta para 5 variables independientes han sido tabulados por Durbin y Watson.

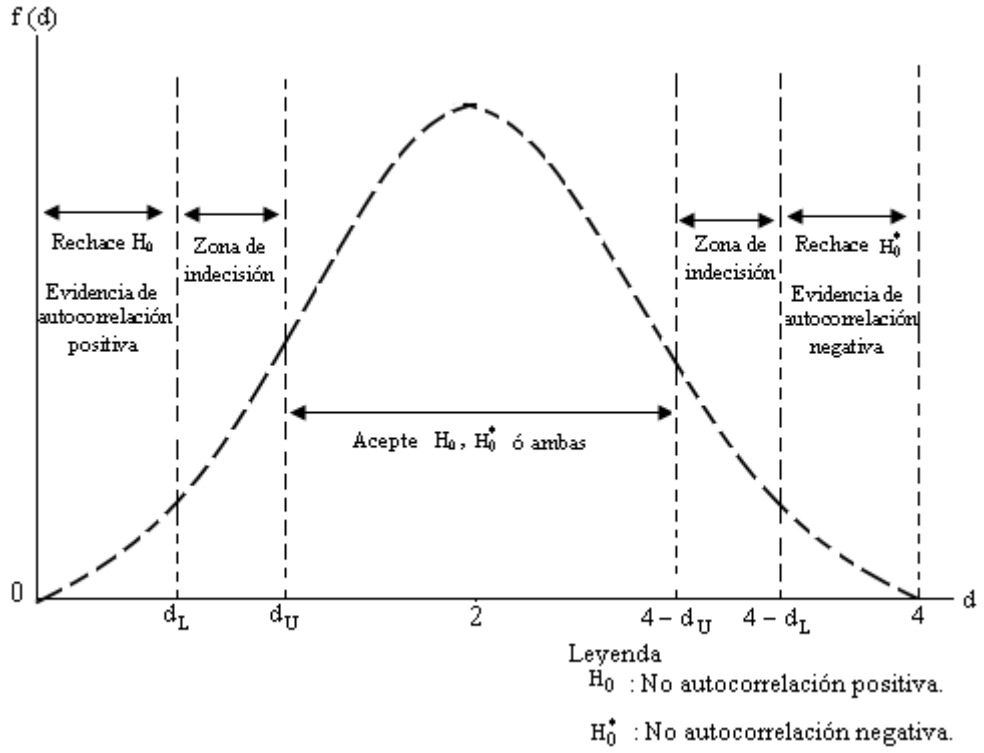
El procedimiento para llevar a cabo la prueba se explica mejor con la ayuda de la figura 7.17, que muestra que los límites de d están entre 0 y 4, lo que se establece expandiendo (7.82), para obtener:

$$d = \frac{\sum_{t=1}^{n} e_t^2 + \sum_{t=1}^{n} e_{t-1}^2 - 2\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \qquad \textbf{(7.84)}$$

Como $\sum_{t=1}^{n} e_t^2$ y $\sum_{t=1}^{n} e_{t-1}^2$ difieren entre sí en una sola observación, se consideran aproximadamente iguales. Entonces haciendo $\sum_{t=1}^{n} e_{t-1}^2 = \sum_{t=1}^{n} e_t^2$ la ecuación (7.84) puede escribirse como:

$$d \approx 2\left(1 - \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2}\right) \qquad \textbf{(7.85)}$$

Figura 7.17 Estadístico Durbin – Watson.



Ahora definamos

$$\hat{\rho} = \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} e_t^2} \qquad \textbf{(7.86)}$$

Como el coeficiente de autocorrelación muestral de primer orden, un estimador de ρ. Utilizando (7.86), (7.85) puede expresarse como:

$$d \approx 2(1-\hat{\rho}) \qquad \textbf{(7.87)}$$

Resulta evidente por la ecuación (7.87) que si $\hat{\rho} = 0$, entonces d = 2; es decir, si no hay correlación serial (de primer orden), se espera que d sea igual a 2. Por lo tanto, como regla general, si se encuentra d igual a 2 en una aplicación, se puede suponer que no hay autocorrelación de primer orden, ni positiva ni negativa. Si $\hat{\rho} = 1$, es decir, si hay autocorrelación, $d \approx 0$; en otras palabras mientras más cerca esté d de 0, mayor será la evidencia de correlación serial positiva, lo que sería evidente con base a la ecuación (7.82) puesto que si existe autocorrelación positiva, los e_t estarán todos juntos y sus diferencias tenderán a ser pequeñas, y por lo tanto el numerador (suma de cuadrados) será menor en comparación con el denominador (suma de cuadrados) que es un valor que permanece fijo para una regresión dada.

Si $\hat{\rho} = -1$, es decir, hay perfecta correlación negativa entre los valores consecutivos de los residuos, entonces $d \approx 4$. Esto es, entre más cerca esté d de 4, mayor será la evidencia de correlación serial negativa, un e_t positivo será seguido por un e_t negativo y viceversa, de tal manera que $|e_t - e_{t-1}|$ será mayor que $|e_t|$, por consiguiente el numerador de d será comparativamente mayor que el denominador.

La mecánica de la prueba de Durbin – Watson es la siguiente, si se cuenta con que los supuestos subyacentes se satisfacen:

1. Corra la regresión de MCO y obtenga los residuos e_i.
2. Calcule el estadístico d usando la ecuación (7.82). (Con el paquete estadístico SPSS se obtiene más fácilmente).

3. Encuentre los valores críticos d_L y d_U para el tamaño de la muestra y el número de variables independientes dadas.

4. Si la hipótesis nula H_0 es la de que no hay correlación serial positiva, entonces si:

$$d < d_L : \text{rechace } H_0$$
$$d > d_U : \text{no rechace } H_0$$
$$d_L \le d \le d_U : \text{la prueba no es concluyente}$$ [11]

5. Si la hipótesis nula H_0 es la de que no hay correlación serial negativa, entonces si:

$$d > 4 - d_L : \text{rechace } H_0$$
$$d < 4 - d_U : \text{no rechace } H_0$$
$$4 - d_U \le d \le 4 - d_L : \text{la prueba no es concluyente}$$

6. Si H_0 es de dos colas, es decir, que no hay autocorrelación serial positiva o negativa, entonces:

$$d < d_L : \text{rechace } H_0$$
$$d > 4 - d_L : \text{rechace } H_0$$
$$d_U < d < 4 - d_U : \text{no rechace } H_0$$
$$\left.\begin{array}{c} d_L \le d \le d_U \\ 4 - d_U \le d \le 4 - d_L \end{array}\right\} \text{la prueba no es concluyente}$$

[11] Theil y Nagar han mostrado, sin embargo, que el límite superior d_U es "aproximadamente igual al verdadero límite de significancia en todos aquellos casos en que el comportamiento de las variables independientes es uniforme, en el sentido de que las primeras y segundas diferencias son pequeñas en comparación con el rango de la variable correspondiente". Ver Henri Theil, Principles of Econometrics, John wiley & Sons, Inc., New York, 1971, p.201.

Como los pasos anteriores lo indican, es una gran desventaja para la prueba d si se cae en la zona de indecisión o región de ignorancia, puesto que no es posible concluir si existe o no la autocorrelación.

Al emplear la prueba Durbin-Watson, es conveniente tener en cuenta que no puede aplicarse en situaciones donde se violen los supuestos. En particular, no puede usarse con modelos autorregresivos, es decir modelos que contienen valores rezagados de la variable dependiente como variables explicatorias. Si se aplica equivocadamente en esté tipo de situaciones, el valor de d estará alrededor de 2, que es el valor de d esperado en ausencia de autocorrelación [ver (7.87)]. Por lo tanto, hay un sesgo "incorporado" en contra del descubrimiento de la correlación serial en tales modelos, lo cual no significa que los modelos autorregresivos no sufran del problema de la autocorrelación.

7.8.3 Medidas Remediales.

Dado que en presencia de correlación serial los estimadores MCO son ineficientes, es necesario buscar medidas remediales. El remedio, sin embargo, depende del conocimiento que se tenga sobre la naturaleza de la interdependencia entre las perturbaciones. A este respecto, se distinguen dos situaciones: cuando se conoce la estructura de la autocorrelación y cuando no se conoce.

7.8.3.1 Cuando se Conoce la Estructura de la Autocorrelación.

Debido a que las perturbaciones ε_t no son observables, la naturaleza de la correlación serial es un asunto de especulación o exigencias prácticas. En la práctica, se

supone frecuentemente que ε_t sigue un esquema autorregresivo de primer orden, como el siguiente:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t \qquad \textbf{(7.88)}$$

donde $|\rho| < 1$ y donde el ν_t sigue los supuestos de MCO de valor esperado cero, varianza constante y no autocorrelación, como se muestra en (7.72).

Si (7.88) es válida, el problema de correlación serial puede resolverse satisfactoriamente si ρ, el coeficiente de correlación se conoce. Para verlo volvamos al modelo de dos variables:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t \qquad \textbf{(7.89)}$$

Si (7.89) se cumple en t, se cumple también en t – 1. Luego,

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \varepsilon_{t-1} \qquad \textbf{(7.90)}$$

Multiplicando (7.90) a ambos lados por ρ, obtenemos:

$$\rho y_{t-1} = \rho\beta_0 + \rho\beta_1 x_{t-1} + \rho\varepsilon_{t-1} \qquad \textbf{(7.91)}$$

Restando (7.91) de (7.89) tendremos:

$$\begin{aligned}(y_t - \rho y_{t-1}) &= \beta_0(1-\rho) + \beta_1 x_t - \rho\beta_1 x_{t-1} + (\varepsilon_t - \rho\varepsilon_{t-1}) \\ (y_t - \rho y_{t-1}) &= \beta_0(1-\rho) + \beta_1(x_t - \rho x_{t-1}) + \nu_t\end{aligned} \qquad \textbf{(7.92)}$$

Donde se utilizó la ecuación (7.88) en el último paso.

Como ν_t, satisface todos los supuestos de MCO, se puede proceder a aplicar el método de MCO a (7.92) y obtener estimadores con todas las propiedades óptimas (insesgados, varianza mínima, etc.). La regresión (7.92) se conoce como la ecuación de diferencias generalizadas; contempla a “y” contra “x” no en la forma original sino en forma de

diferencias, que se obtienen restando una proporción (= ρ) del valor de la variable en el período anterior, del valor de la variable en el período corriente. Obteniendo estas diferencias se pierde una observación porque la primera no tiene un antecesor; para evitar esto, la primera observación se transforma de la siguiente manera[12]:

$$y_1\sqrt{(1-\rho^2)} \quad \text{y} \quad x_1\sqrt{(1-\rho^2)}$$

Cuando no se conoce ρ.

Siendo más o menos directo el método anterior, la regresión con diferencias generalizadas suele ser difícil de correr porque ρ rara vez se conoce en la práctica. Algunos métodos alternos se comentan a continuación:

1. El método de primera diferencia. Como ρ cae entre 0 y ±1, se puede comenzar por dos posiciones extremas. Si suponemos que $\rho = 0$, no existe autocorrelación si $\rho = \pm 1$, entonces existe autocorrelación positiva o negativa perfecta. En la práctica cuando se corre una regresión se suele suponer que no existe autocorrelación, dejando que la prueba de Durbin-Watson u otras pruebas nos digan si el supuesto es justificado. Si $\rho = +1$, entonces la ecuación de diferencia generalizada (7.92) se reduce a la ecuación de primera diferencia.

$$y_t - \rho y_{t-1} = \beta_1(x_t - \rho x_{t-1}) + (\varepsilon_t - \rho\varepsilon_{t-1})$$
$$y_t - \rho y_{t-1} = \beta_1(x_t - \rho x_{t-1}) + \nu_t$$

o

$$\Delta y_t = \beta_1 \Delta x_t + \nu_t \qquad \textbf{(7.93)}$$

[12] Es importante que se transformen las primeras observaciones de "x" y "y"; de no ser así, el método de primeras diferencias puede no ser mejor que el MCO común y corriente.

Donde Δ, la letra griega delta, es el operador primera diferencia y se utiliza como símbolo u operador (como el operador valor esperado E) para diferencias entre dos valores consecutivos. (Nota: generalmente un operador es un símbolo que expresa una operación matemática.) Al correr (7.93) todo lo que hay que hacer es formar las primeras diferencias tanto de la variable dependiente como de las variables independientes, y utilizar como insumos en la regresión de las nuevas cifras.

Obsérvese que una de las características importantes del modelo de primera diferencia es que el intercepto es cero, por lo que al correr (7.93) debe utilizarse una regresión que pase por el origen. Supongamos sin embargo, que el modelo original fuera

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 t + \varepsilon_t \qquad \textbf{(7.94)}$$

Donde t es la variable tendencia y donde ε_t sigue el esquema autorregresivo de primer orden. Así se tiene que la transformación de primera diferencia de (7.94) es:

$$\Delta y_t = \beta_1 \Delta x_t + \beta_2 + \nu_t \qquad \textbf{(7.95)}$$

Donde: $\Delta y_t = y_t - y_{t-1}$ y $\Delta x_t = x_t - x_{t-1}$. La ecuación (7.95) muestra un intercepto en la forma de primera diferencia que contrasta con (7.93) y donde desde luego , β_2 es el coeficiente de la variable de tendencia en el modelo original. En conclusión, si existe un intercepto en la forma con primera diferencia es porque hay en el modelo original, un término de tendencia lineal,

siendo el intercepto el coeficiente de la mencionada variable de tendencia. Si β_2 es, por ejemplo, positiva en (7.95), quiere decir que hay una tendencia hacia arriba en "y", una vez considerada la influencia de las otras variables.

Si en el lugar de suponer $\rho = +1$, suponemos que $\rho = -1$ es decir perfecta correlación serial negativa (lo que no es precisamente típico en las series económicas), la ecuación de diferencia generalizada (7.92) se convierte en:

$$y_t + y_{t-1} = 2\beta_0 + \beta_1(x_t + x_{t-1}) + \nu_t$$

o

$$\frac{y_t + y_{t-1}}{2} = \beta_0 + \frac{\beta_1(x_t + x_{t-1})}{2} + \frac{\nu_t}{2} \qquad \textbf{(7.96)}$$

El modelo anterior se conoce como el modelo de regresión de promedios móviles (en dos períodos) porque se trata de una regresión de un promedio móvil contra otro promedio móvil[13].

La transformación anterior de primera diferencia es muy popular en la econometría aplicada por ser muy fácil de interpretar. Pero observe que esta transformación se apoya en el supuesto de que $\rho = +1$, es decir, las perturbaciones están perfectamente correlacionadas positivamente. Si no es éste el caso, el remedio puede ser peor que la enfermedad. Nos resta comentar cómo saber si el supuesto de que $\rho = +1$ es justificable en una situación dada. La respuesta se da a continuación:

[13] Como $(y_t + y_{t-1})/2$ y $(x_t + x_{t-1})/2$ son los promedios de dos valores adyacentes (vecinos), son llamados promedios de dos períodos. Son móviles porque al calcular en periodos sucesivos estos promedios se prescinde de una observación y se añade otra. Así $(y_{t+1} + y_t)/2$ será el siguiente promedio de dos períodos, etc.

2. ρ basado en el estadístico Durbin-Watson d. Recordemos que anteriormente establecimos la siguiente relación:

$$d \approx 2(1-\hat{\rho}) \qquad \textbf{(7.97)}$$

ó

$$\hat{\rho} \approx 1-\frac{d}{2} \qquad \textbf{(7.98)}$$

Que sugiere una manera sencilla de obtener una estimación de ρ a partir del estadístico estimado d. A partir de (7.98) resulta claro que el supuesto de primera diferencia ρ = + 1 es válido sólo si d = 0, o aproximadamente igual a cero. También es claro que cuando d = 2, $\hat{\rho} = 0$ y cuando d = 4, $\hat{\rho} = -1$. Entonces, el estadístico d nos proporciona un método "listo" para obtener una estimación de ρ. Nótese sin embargo, que la relación (7.98) es aproximada y es posible que no se cumpla en muestras pequeñas. Theil y Nagar han sugerido la siguiente relación[14]:

$$\hat{\rho} = \frac{n^2(1-d/2)+k^2}{n^2-k^2} \qquad \textbf{(7.99)}$$

Donde:

n: Número total de observaciones.

d: Durbin-Watson.

k: Número de coeficientes (incluyendo el intercepto) que van a ser estimados.

[14] Estos autores suponen que las variables independientes se mueven suavemente; especialmente las primeras y segundas diferencias de estas variables son pequeñas en valor absoluto en relación al rango de las mismas variables.

Es fácil verificar que para n grande la formulación de Theil-Nagar coincide con la relación (7.98). Una vez que se ha estimado ρ a partir de (7.98) y (7.99) se pueden transformar los datos utilizando la ecuación de diferencia generalizada (7.92) y a continuación proceder con la estimación usual de MCO. Recuérdese que las primeras observaciones de "x" y "y" tienen que ser multiplicadas por $\sqrt{1-\hat{\rho}^2}$ evitando así la pérdida de la primera observación.

Ejemplo 7: Ventas de concentrado para bebidas gaseosas.

Una empresa fabricante de bebidas gaseosas desea pronosticar las ventas anuales regionales del concentrado de uno de sus productos, en función de los gastos de promoción regional de ese producto. En las columnas 1 y 2 de la tabla 7.17 se ven los datos de 20 años. Suponiendo que sea adecuada una relación lineal, se ajustó un modelo lineal de regresión con los Mínimos Cuadrados Ordinarios. En la columna 3 de la tabla 7.17 se ven los residuos de este modelo rectilíneo, y en la tabla 7.18 se presentan otros estadísticos de resumen para el modelo. Como las variables independientes de las dependientes son de serie temporal, se cree que puede haber autocorrelación. En la figura 7.18 se muestra la gráfica de los residuos en función del tiempo, en la que se puede observar que hay un desplazamiento definido, primero hacia arriba y después hacia abajo, en los residuos. La autocorrelación podría ser la responsable de ese comportamiento.

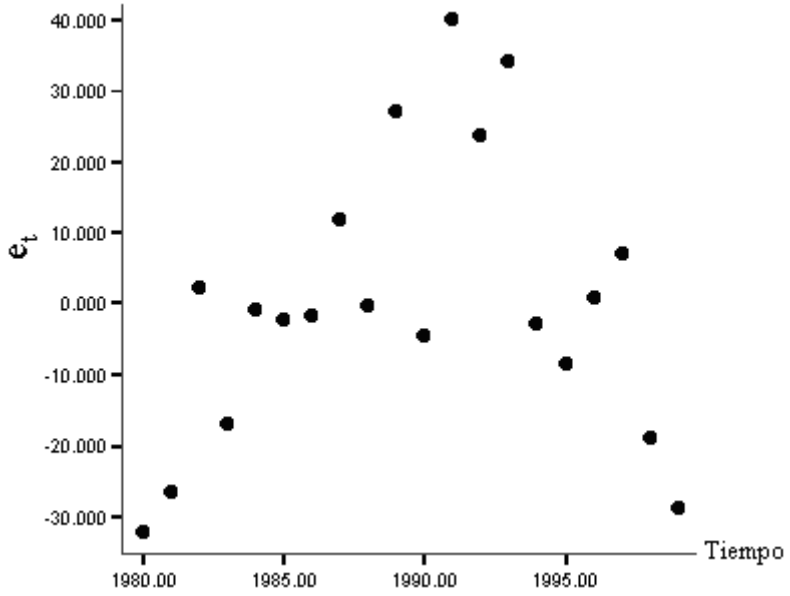
Tabla 7.17 Datos del ejemplo de ventas de concentrado de bebida gaseosa.

Año	t	(1) Ventas anuales regionales de concentrado (unidades) y_t	(2) Gastos anuales de publicidad ($*1000) x_t	(3) Residuos de mínimos cuadrados e_t	(4) e_t^2	(5) $(e_t - e_{t-1})^2$	(6) Población regional anual z_t
1980	1	3083	75	-32.330	1045.2289		825000
1981	2	3149	78	-26.603	707.7196	32.7985	830445
1982	3	3218	80	2.215	4.9062	830.4771	838750
1983	4	3239	82	-16.967	287.8791	367.9491	842940
1984	5	3295	84	-1.148	1.3179	250.2408	846315
1985	6	3374	88	-2.512	6.3101	1.8605	852240
1986	7	3475	93	-1.967	3.8691	0.2970	860760
1987	8	3569	97	11.669	136.1656	185.9405	865925
1988	9	3597	99	-0.513	0.2632	148.4011	871640
1989	10	3725	104	27.032	730.7290	758.7270	877745
1990	11	3794	109	-4.422	19.5541	989.3541	886520
1991	12	3959	115	40.032	1602.5610	1976.1581	894500
1992	13	4043	120	23.577	555.8749	270.7670	900400
1993	14	4194	127	33.940	1151.9236	107.3918	904005
1994	15	4318	135	-2.787	7.7674	1348.8725	908525
1995	16	4493	144	-8.606	74.0632	33.8608	912160
1966	17	4683	153	0.575	0.3306	84.2908	917630
1997	18	4850	161	6.848	46.8951	39.3505	922220
1998	19	5005	170	-18.971	359.8988	666.6208	925910
1999	20	5236	182	-29.063	844.6580	101.8485	929610
		$\sum_{t=1}^{20} e_t^2 = 7587.9154$		$\sum_{t=2}^{20} (e_t - e_{t-1})^2 = 8195.2065$			

Tabla 7.18 Estadísticos de resumen para el modelo de mínimos cuadrados del ejemplo 7.

Parámetro	Estimado	Error estándar	Estadístico t.
β_0	1608.508	17.0223	94.49
β_1	20.091	0.1428	140.71
n = 20	$R^2 = 0.9991$	$MS_{Res} = 421.5485$	

Figura 7.18 Residuos, e_t, en función del tiempo, ejemplo 7.



También se utiliza la prueba de Durbin-Watson como sigue:

Solución:

1. $H_0 : \rho = 0$

2. $H_1 : \rho > 0$

3.Se selecciona un nivel de significancia de α = 0.05 y los valores críticos (de la tabla) correspondientes para n = 20 y una variable independiente, son d_L = 1.20 y d_U = 1.41.

4.Cálculos:

$$d = \frac{\sum_{t=1}^{20}(e_t - e_{t-1})^2}{\sum_{t=1}^{20} e_t^2} = \frac{8195.2065}{7587.9154} = 1.08$$

5.Decisión Estadística: Se rechaza la hipótesis nula.

6. Conclusión: Dado que el valor d = 1.08 es menor que d_L = 1.20 se concluye que los errores tienen autocorrelación positiva.

Un valor significativo en el estadístico de Durbin – Watson, o una gráfica dudosa de residuales, indica que hay un error de especificación del modelo. Esta mala especificación podría ser una dependencia real de los errores respecto al tiempo, o una dependencia "artificial", causada por la omisión de una variable independiente importante. Si la autocorrelación aparente se debe a variables independientes faltantes, y si se pueden identificar e incorporar al modelo esas variables faltantes, se podrá eliminar la autocorrelación aparente. Esto se ilustra en el siguiente ejemplo.

Ejemplo 8:

Se tienen los datos de las ventas de concentrado para bebidas gaseosas que se presentaron en el ejemplo 7. La prueba de Durbin – Watson ha indicado que los errores en el modelo de regresión lineal, que relaciona las ventas de concentrado con los gastos de promoción, tienen autocorrelación positiva. En este ejemplo es relativamente fácil imaginar otros regresores probables que puedan estar positivamente correlacionados con las ventas. Por ejemplo, es muy probable que la población de la región afecte las ventas de concentrado.

En la columna 6 de la tabla 7.17 se muestran datos sobre la población de la región durante los años 1980 a 1999. Si se agrega esta variable al modelo, la ecuación tentativa será:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t$$

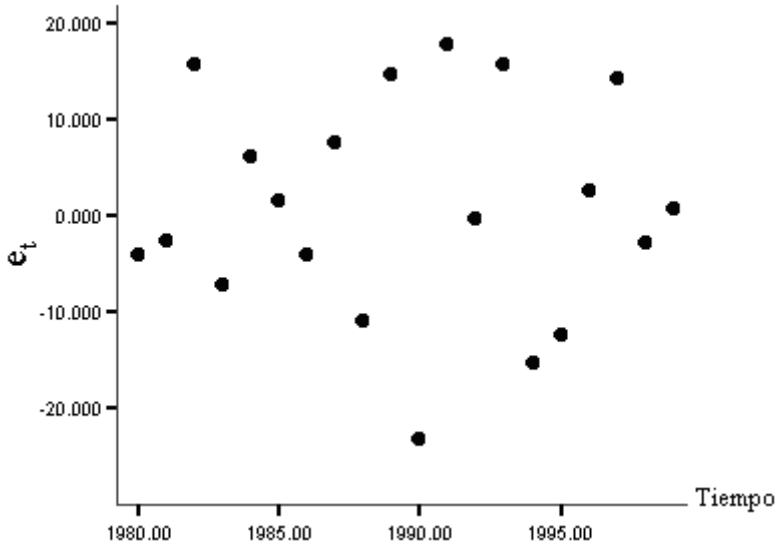
La tabla 7.19 contiene los estadísticos de resumen para el análisis de esos datos por Mínimos Cuadrados.

Tabla 7.19 Estadísticos de resumen para el modelo del ejemplo 8.

Parámetro	Estimado	Error estándar	Estadístico t.
β_0	320.340	217.3278	1.47
β_1	18.434	0.2915	63.23
β_2	0.002	0.0003	5.93
n = 20	$R^2 = 0.9997$	d = 3.06	$MS_{Res} = 145.3408$

Se ve en la tabla que el estadístico Durbin – Watson es d = 3.06, porque el 5% de los valores críticos ahora con dos variables independientes, son $d_L = 1.10$ y $d_U = 1.54$, harían llegar a la conclusión de que no hay autocorrelación positiva en los errores.

Figura 7.19 Residuos, e_t, en función del tiempo, ejemplo 8.



La gráfica de los residuos en función del tiempo se ve en la figura 7.19, y mejoró mucho, en comparación con la figura 7.18; por consiguiente, al agregar el tamaño de la población al modelo se ha eliminado el problema aparente de la autocorrelación.

Ejercicios 7.

1. Un combustible sólido para cohetes pierde peso después de haber sido producido. Se disponen de los siguientes datos:

Meses después de producido, x	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
Pérdida de peso, y (kg)	1.42	1.39	1.55	1.89	2.43	3.15	4.05	5.15	6.43	7.89

   a) Ajustar un polinomio de segundo orden que exprese la pérdida de peso en función de la cantidad de meses después de haber sido producido.

   b) Probar la significancia de la regresión con $\alpha = 0.05$.

   c) Probar la hipótesis H_0: $\beta_2 = 0$. Comente la necesidad del término cuadrático en este modelo.

2. Para el ejercicio 1, calcular los residuos del modelo de segundo orden. Analizar los residuos y comentar la adecuación del modelo.

3. Se llevó a cabo un experimento con el objetivo de determinar si el flujo sanguíneo cerebral en los seres humanos podía pronosticarse a partir de la presión del oxígeno arterial (milímetros de mercurio). Se utilizaron quince pacientes en este estudio y los datos observados fueron los que se muestran en la tabla siguiente:

   a) Estimar la ecuación de regresión cuadrática.

   b) Probar la significancia de la regresión con $\alpha = 0.05$.

   c) Probar la hipótesis H_0: $\beta_2 = 0$. comente la necesidad del término cuadrático en esta ecuación.

Flujo sanguíneo, y	Presión del oxígeno arterial, x
84.33	603.40
87.80	582.50
82.20	556.20
78.21	594.60
78.44	558.90
80.01	575.20
83.53	580.10
79.46	451.20
75.22	404.00
76.58	484.00
77.90	452.40
78.80	448.40
80.67	334.80
86.60	320.30
78.20	350.30

4. Usando los siguientes seis puntos de datos, estime un modelo lineal de probabilidad haciendo uso de Mínimos Cuadrados Ordinarios:

x	-1	-2	0	1	1	1
y	0	0	0	1	1	1

Calcule R^2 para el modelo. Luego use el modelo estimado para clasificar a los individuos en dos categorías. Calcule el número de clasificaciones correctas usando la siguiente regla de clasificación:

$$\text{Clasificar} = \begin{cases} \text{primer grupo}(y=1) & \text{si } \hat{y} > 1/2 \\ \text{segundo grupo}(y=0) & \text{si } \hat{y} \leq 1/2 \end{cases}$$

Discuta las ventajas y desventajas de usar R^2 o el porcentaje de clasificaciones correctas como una medida de la bondad del ajuste en el modelo lineal de probabilidad.

5. La siguiente tabla presenta cifras hipotéticas para 40 familias respecto de tener casa propia “y” (1 = tiene casa propia, 0 = no tiene casa propia) y al ingreso familiar “x” (en miles de dólares).

Familia	y	x	Familia	y	x
1	0	8	21	1	22
2	1	16	22	1	16
3	1	18	23	0	12
4	0	11	24	0	11
5	0	12	25	1	16
6	1	19	26	0	11
7	1	20	27	1	20
8	0	13	28	1	18
9	0	9	29	0	11
10	0	10	30	0	10
11	1	17	31	1	17
12	1	18	32	0	13
13	0	14	33	1	21
14	1	20	34	1	20
15	0	6	35	0	11
16	1	19	36	0	8
17	1	16	37	1	17
18	0	10	38	1	16
19	0	8	39	0	7
20	1	18	40	1	17

a) Ajuste a los datos un modelo lineal de probabilidad e interprete la ecuación resultante.

b) Para cada familia obtenga el “y” estimado. ¿Cómo trataría el “y” estimado que sea negativo o mayor que 1?

6. Se quiere analizar la relación existente entre el grado de estrés de los trabajadores "y", medido a partir del tamaño de la empresa en que trabajan, x_1, el número de años que llevan en el puesto de trabajo, x_2, el salario anual percibido, x_3 y la edad del trabajador, x_4. Se pide:

   a) Estimar la ecuación de regresión.

   b) Calcular el valor de R^2 y el valor de R^2 ajustado.

   c) Realizar la prueba de hipótesis individual y global de los coeficientes.

   d) Analizar el problema y ver si es posible descartar alguna de las variables independientes que resulte colineal.

Para ello se dispone de las observaciones siguientes:

y	x_1	x_2	x_3	x_4
101	812	15	30	38
60	334	8.0	20	52
10	377	5.0	20	27
27	303	10	54	36
89	505	13	52	34
60	401	4	27	45
16	177	6	26	50
184	598	9	52	60
34	412	16	34	44
17	127	2	28	39
78	601	8	42	41
141	297	11	84	58
11	205	4	31	51
104	603	5	38	63
76	484	8	41	30

7. Considere el siguiente conjunto de datos hipotéticos:

y	x_1	x_2
-10	1	1
-8	2	3
-6	3	5
-4	4	7
-2	5	9
0	6	11
2	7	13
4	8	15
6	9	17
8	10	19
10	11	21

Si se quiere ajustar el modelo $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$, ¿pueden estimarse los coeficientes de regresión? ¿Por qué si o por qué no?

8. Consideramos un estudio de corte transversal de los gastos de vivienda anuales e ingresos anuales de cuatro grupos de familias donde y_i son los gastos de vivienda y x_i es el ingreso.

Grupo	Gastos de vivienda, (miles de $)					Ingreso (miles de $)
1	1.8	2.0	2.0	2.0	2.1	5.0
2	3.0	3.2	3.5	3.5	3.6	10.0
3	4.2	4.2	4.5	4.8	5.0	15.0
4	4.8	5.0	5.7	6.0	6.2	20.0

a) Estimar la ecuación de regresión.

b) Calcular el valor de R^2, t y F.

c) Realizar un examen gráfico de los residuos para determinar si está presente la heteroscedasticidad en el modelo.

9. Los datos de la tabla siguiente muestran las ventas mensuales de un fabricante de cosméticos (y_t) y las ventas mensuales correspondientes de toda la industria (x_t). Las unidades de x_t y y_t son millones de dólares.

t	x_t	y_t
1	5.00	0.318
2	5.06	0.330
3	5.12	0.356
4	5.10	0.334
5	5.35	0.386
6	5.57	0.455
7	5.61	0.460
8	5.80	0.527
9	6.04	0.598
10	6.16	0.650
11	6.22	0.685
12	6.31	0.713
13	6.38	0.724
14	6.54	0.775
15	6.68	0.782
16	6.73	0.796
17	6.89	0.859
18	6.97	0.883

a) Ajustar un modelo de regresión lineal simple a los datos.

b) Graficar los residuos en función del tiempo. ¿hay algún indicio de autocorrelación?.

c) Calcular el valor d (Durbin - Watson).

d) Aplicar la prueba de Durbin – Watson para determinar si hay autocorrelación positiva de los errores.

e) Estimar ρ por el método de Theil – Nagar.

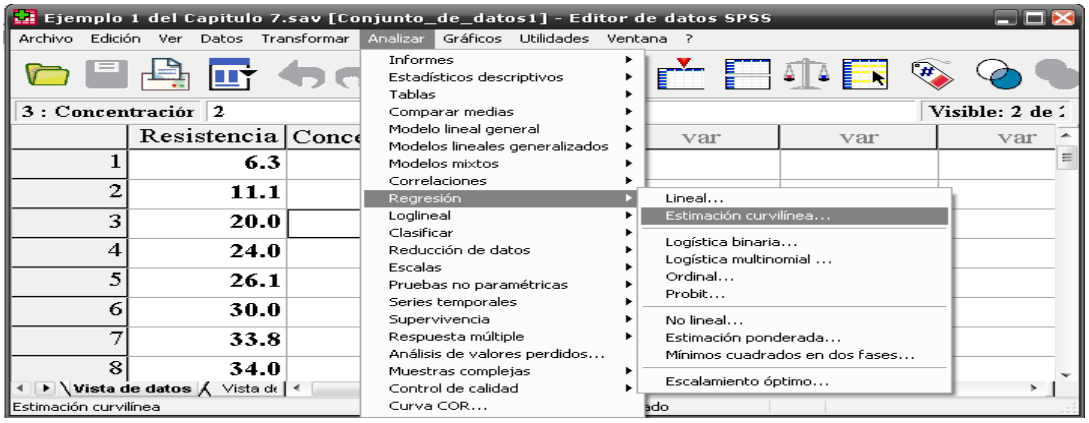
10. Dada una muestra de 50 observaciones y 4 variables independientes, ¿Qué se puede decir acerca de la autocorrelación si: a) d = 1.05? b) d = 1.40? c) d = 2.50? d) d = 3.97?.
11. ¿Por qué es improbable que los errores en los estudios de corte transversal estén correlacionados serialmente? ¿puede dar un ejemplo en el que esté presente la correlación serial?.

Apéndice 7.1: Solución del Ejemplo 1 Haciendo uso del Software Estadístico SPSS v15.0.

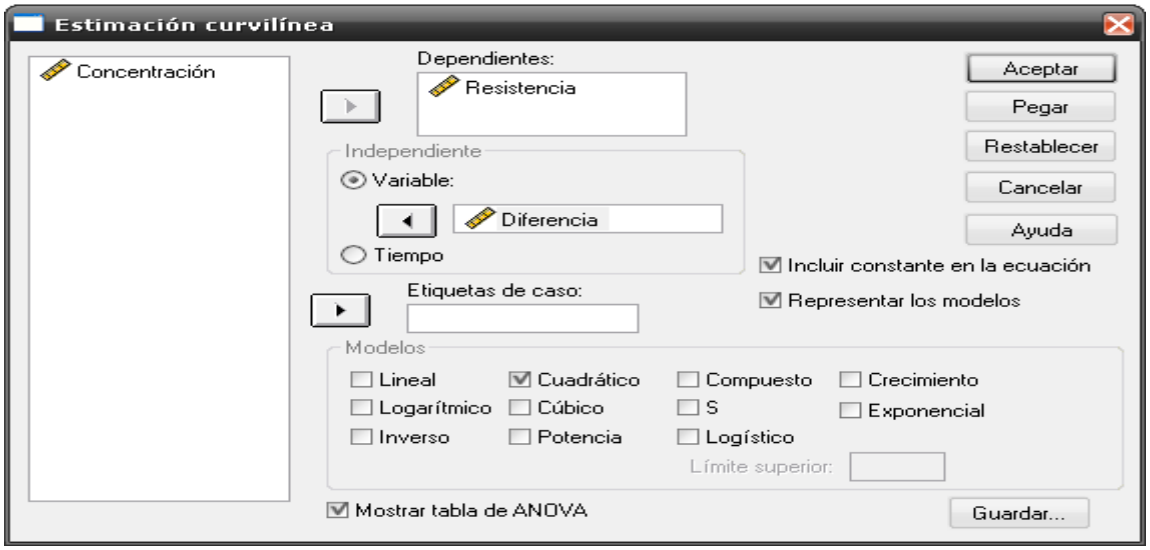
Haciendo uso del software se puede obtener los resultados del ejemplo 1, en una sola ejecución siguiendo los siguientes pasos:

1. Se les da un nombre a las dos variables en estudio se digitan los datos para cada variable y se obtiene la ventana siguiente en la cual solamente se muestran 5 observaciones del total (19) nuestra variable independiente será diferencia que se obtuvo de $(x_i - \bar{x}) = (x_i - 7.2632)$.

2. En la barra de menú se selecciona la opción Analizar → Regresión → Estimación curvilínea como se muestra a continuación:



3. Al hacer click en la opción Estimación curvilínea aparece la siguiente ventana en la cual se colocan las variables cada una en su lugar, en este cuadro aparecen los distintos tipos de modelos que se pueden ajustar, en nuestro caso hemos seleccionado el modelo cuadrático, se puede obtener la tabla de análisis de varianza seleccionando mostrar la tabla de ANOVA, al seleccionar la opción guardar se pueden obtener los valores estimados y los residuos.



4. Haciendo un click en aceptar del cuadro anterior se obtienen los resultados siguientes:

Resumen del modelo

R	R cuadrado	R cuadrado corregida
.953	.909	.897

La variable independiente esConcentració

ANOVA

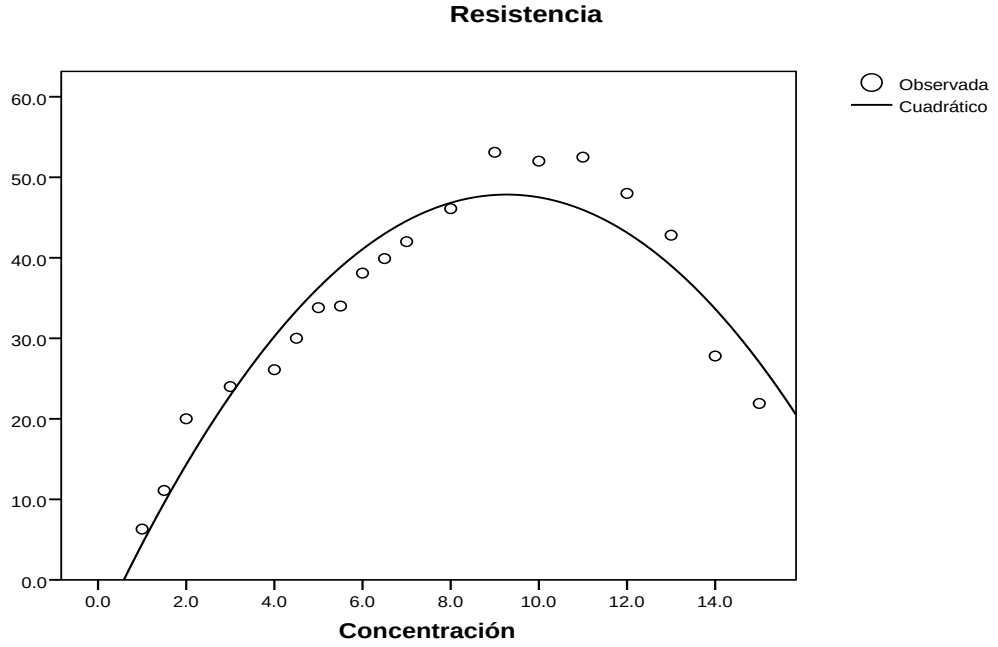
	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Regresión	3104.247	2	1552.123	79.434	.000
Residual	312.638	16	19.540		
Total	3416.885	18			

La variable independiente esConcentración.

Coeficientes

	Coeficientes no estandarizados		
	B	Error típico	t
Diferencia	2.546	.254	10.031
Diferencia ** 2	-.635	.062	-10.270
(Constante)	45.295	1.483	30.545

Se puede observar que los resultados obtenidos con SPSS son los mismos que se obtuvieron anteriormente, también se obtiene el diagrama de dispersión con ajuste como se muestra a continuación:



Se puede observar que los valores observados se aproximan a la curva de regresión ajustada lo que indica que el modelo cuadrático es el adecuado para estos datos.

5. En el paso 3 al hacer un click en la opción guardar se pueden obtener los valores estimados y los residuos con los cuales se puede hacer el diagrama de dispersión de los residuos frente a los valores estimados además la gráfica de probabilidad normal y se obtienen las figuras 7.3 y 7.4 dadas anteriormente.

Tabla 7.20 Remuneración por empleados ($) en industrias manufactureras de bienes perecederos según la escala de empleo del establecimiento.

Industria	Escala de empleo (Nº. promedio de empleados)								
	1-4	5-9	10-19	20-49	50-99	100-249	250-499	500-999	1000-2499
Alimentos y productos afines	2994	3295	3565	3907	4189	4486	4676	4968	5342
Productos del tabaco	1721	2057	3336	3320	2980	2848	3072	2969	3822
Textiles confecciones	3600	3657	3674	3437	3340	3334	3225	3163	3168
y productos relacionados	3494	3787	3533	3215	3030	2834	2750	2967	3453
Papel y productos afines	3498	3847	3913	4135	4445	4885	5132	5342	5326
Editorial y artes gráficas	3611	4206	4695	5083	5301	5269	5182	5395	5552
Productos químicos y derivados	3875	4660	4930	5005	5114	5248	5630	5870	5876
Productos del petróleo y del carbón	4616	5181	5317	5337	5421	5710	6316	6455	6347
Caucho y productos plásticos	3538	3984	4014	4287	4221	4539	4721	4905	5481
Cuero y productos de cuero	3016	3196	3149	3317	3414	3254	3177	3346	4067
Remuneración media	3396	3787	4013	4104	4146	4241	4387	4538	4843
Desviación estándar	743.7	851.4	727.8	746.3	929.9	1080.6	1243.2	1307.7	1112.5
Productividad media	9355	8584	7962	8275	8389	9418	9795	10281	11750

Capítulo 8

Métodos de Selección de Variables.

8.1 Introducción.

Un problema importante cuando se dispone de un amplio conjunto de variables independientes, es seleccionar un subconjunto de ellas que proporciona el mejor modelo de regresión. Cuando el número de variables es grande (mayor de 10), es frecuente que un modelo con un subconjunto de variables proporcione predicciones mucho mejores que el modelo con todas las variables.

En este Capítulo se presentan tres métodos de selección de variables (método de selección hacia adelante, método de eliminación hacia atrás y regresión paso a paso), donde la función de cada método es la de exponer las variables a una metodología sistemática diseñada para asegurar la inclusión de las mejores combinaciones de variables, que se van a utilizar en la ecuación final.

Los modernos paquetes de computadora realizan los cálculos y elaboran el resumen de información cuantitativa de todos los modelos para cada posible subconjunto de variables, en nuestro caso utilizamos el software estadístico SPSS para desarrollar cada uno de los tres métodos.

8.2 Construcción de Modelos de Regresión.

Cuando se dispone de un conjunto amplio de variables independientes, existen varias estrategias de regresión para seleccionar las variables que tienen un aporte significativo al modelo, aquí se muestran tres métodos de selección de variables. La varianza promedio de predicción en los puntos observados es $\sigma^2(k + 1)/n$, y aumenta en σ^2/n por cada variable innecesaria introducida. Estas estrategias tratan de evitar seleccionar modelos que incluyan variables innecesarias, lo que mejorará su comportamiento predictivo. En especial, cuando tengamos variables muy correlacionadas entre sí, hemos visto que incluyendo variables muy correlacionadas en el modelo de regresión, inflamos las varianzas de los coeficientes estimados y, por lo tanto, del modelo ajustado y de sus predicciones.

8.3 Métodos de Selección de Variables en Regresión.

Como la evaluación de todas las variables independientes posibles puede ser difícil, se han desarrollado varios métodos para evaluar sólo una pequeña cantidad de modelos de regresión con un subconjunto, agregando o eliminando variables una por una. Esos métodos pueden clasificarse en tres categorías principales:

a) Selección hacia adelante.

b) Eliminación hacia atrás.

c) Regresión paso a paso.

8.3.1 Selección Hacia Adelante.

En este procedimiento comenzamos con una única variable y vamos incluyendo el resto, una a una, hasta obtener la ecuación definitiva. El procedimiento puede resumirse así: escogemos como variable de entrada la más correlacionada con "y" o de manera equivalente, la que da el valor más grande de R^2, sea esta x_1; calculamos la regresión simple entre ambas (x_1, y) y los coeficientes de correlación parcial entre el resto de las variables (x_2,..., x_k) y la variable "y" eliminando el efecto de la variable x_1. Introducimos entonces como segunda variable aquella que presente un coeficiente de correlación parcial con la variable independiente más alto. Supongamos que es x_2. Calculamos la ecuación de regresión con las variables (y, x_1, x_2) y comprobamos si el estadístico t para el coeficiente de regresión $\hat{\beta}_2$ de x_2, es significativo. Si no lo es, terminamos el proceso; si lo es, introducimos como nueva variable la más correlacionada con la respuesta eliminando el efecto de x_1 y x_2. El proceso continúa hasta obtener un valor de t no significativo.

El método de selección hacia adelante tiene la ventaja de requerir una menor capacidad de cálculo. Sin embargo, es peor respecto al error de especificación, ya que no es capaz de eliminar variables cuando la introducción de otras nuevas hacen innecesaria su presencia. Por ejemplo, es posible que la primera variable introducida pierda su eficacia al introducir nuevas variables y deba eliminarse en una etapa posterior de la regresión, lo que no es posible con este procedimiento. Además, es posible que alguna variable aparezca como no significativa cuando realmente lo es pero tiene interacción con alguna variable no incluida. Por esta razón este método se utiliza poco en la práctica.

8.3.2 Eliminación Hacia Atrás.

Este método comienza con una regresión que incorpora todas las variables independientes potencialmente influyentes. A continuación, se calculan los estadísticos t para cada coeficiente, y si alguno de estos valores no es significativo para un nivel de significancia dado, se elimina esta variable. Se calcula la regresión con las k – 1 variables restantes, y se repite el procedimiento de eliminación de variables no significativas.

La estrategia de eliminación hacia atrás tiene el inconveniente de utilizar mucha capacidad de cálculo, es posible que únicamente un subconjunto pequeño de las k variables sea significativo, y este procedimiento obliga a efectuar regresiones muy extensas. Además, conduce fácilmente al problema de multicolinealidad si hay variables muy relacionadas o el número de variables es muy elevado. En contrapartida, es excelente para evitar la exclusión de alguna variable significativa, por lo que se utiliza con frecuencia cuando el número de variables es pequeño. Para problemas grandes, esta estrategia es lenta y poco utilizada.

8.3.3 Regresión Paso a Paso.

El procedimiento de regresión paso a paso, que ha adquirido gran popularidad, trata de evitar los inconvenientes de la selección hacia adelante de variables, manteniendo su relativa economía de cálculo. Se diferencia de éste (método de selección hacia adelante) en que, en cada paso, al incluir una nueva variable, el papel de todas las

ya presentes es reevaluado mediante un contraste t (o F, que es equivalente), pudiendo rechazarse alguna de las ya incluidas.

1. Una regla de entrada de nuevas variables: introducimos una variable cuado:
    a) Produce el máximo incremento de la variabilidad explicada por el modelo al incluirla.
    b) La variabilidad explicada por ella es significativa a un nivel prefijado.

    Estas condiciones suponen introducir la variable cuyo coeficiente de regresión tiene el máximo valor del estadístico t de Student.
2. Una regla de salida: excluimos una variable introducida en una etapa anterior, cuando su estadístico t no sea significativo.

Esta estrategia de regresión es muy utilizada. Sin embargo, es peligroso confiar en la selección automática que realiza el ordenador, especialmente en problemas con muchas variables donde desconocemos el nivel de significación que estamos utilizando en los contrastes por los problemas de contrastes múltiples. En general, recomendamos trabajar con un nivel de significación muy bajo, de manera que el ordenador incluya en el proceso todas las variables que puedan tener efectos. Esto nos permite observar si la introducción de alguna variable altera profundamente los coeficientes anteriores a pesar de tener un bajo poder explicativo, señal, en muchos casos, de alta multicolinealidad.

Ejemplo 1: Método de selección hacia adelante.

Se consideran los datos de la tabla 8.1 en la cual se tomaron mediciones de nueve recién nacidos. El propósito es llegar a una ecuación de estimación apropiada que relacione la

talla del recién nacido (y) en centímetros con todas o con un subconjunto de las variables independientes (x_i).

Tabla 8.1 Datos relacionados a la talla de recién nacidos.

Talla del recién nacido, y (cm.)	Edad, x_1(días)	Talla al nacer, x_2 (cm.)	Peso al nacer, x_3 (kg.)	Tamaño del tórax al nacer, x_4 (cm.)
57.5	78	48.2	2.75	29.5
52.8	69	45.5	2.15	26.3
61.3	77	46.3	4.41	32.2
67.0	88	49.0	5.52	36.5
53.5	67	43.0	3.21	27.2
62.7	80	48.0	4.32	27.7
56.2	74	48.0	2.31	28.3
68.5	94	53.0	4.30	30.3
69.2	102	58.0	3.71	28.7

Antes de mostrar los resultados que se obtienen con el software estadístico SPSS, se detalla el proceso que se hace en cada método de selección de variables.

Haciendo uso de los datos de la tabla 8.1, primeramente se detalla el procedimiento de selección hacia adelante:

Paso 1: Se halla la regresión simple con la variable independiente más altamente correlacionada con la variable dependiente, para poder ver esto se necesita hacer la regresión de la variable dependiente con todas las independientes. En la tabla 8.2 se muestra la matriz de correlación de las variables independientes y la dependiente:

Tabla 8.2 Correlaciones de Pearson.

Variables	y	x_1	x_2	x_3	x_4
y	1.000	0.947	0.819	0.761	0.560
x_1	0.947	1.000	0.952	0.534	0.390
x_2	0.819	0.952	1.000	0.263	0.155
x_3	0.761	0.534	0.263	1.000	0.784
x_4	0.560	0.390	0.155	0.784	1.000

En este caso se puede observar que la variable independiente más altamente correlacionada con la dependiente es Edad (x_1) que tiene una correlación de 0.947 con Talla del recién nacido (y), así la primera variable en el modelo es x_1, entonces calculamos la regresión lineal simple entre x_1 y "y", obteniendo la ecuación siguiente:

$$\hat{y} = 19.011 + 0.518\,x_1$$

Paso 2: Se introduce la segunda variable, aquella que presente un coeficiente de correlación más alto eliminando el efecto de x_1 y se obtiene el siguiente resultado:

Tabla 8.3 Correlaciones de Pearson manteniendo constante x_1.

Variables de control	Variables	y	x_1	x_2	x_3
x_1	y	1.000	-0.849	0.941	0.646
	x_2	-0.849	1.000	-0.953	-0.770
	x_3	0.941	-0.953	1.000	0.740
	x_4	0.646	-0.770	0.740	1.000

La tabla 8.3 muestra que la siguiente variable que debemos introducir es Peso al nacer (x_3) que tiene una correlación de 0.941 con Talla del recién nacido (y). Ahora calculamos la ecuación de regresión con las dos variables (x_1 y x_3) y comprobamos si el estadístico t para el coeficiente de regresión $\hat{\beta}_3$ de x_3 es significativo.

Tabla 8.4 Estadísticos de resumen.

Parámetro	Estimado	Error estándar	Estadístico t.
β_0	20.108	1.987	10.119
β_1	0.414	0.029	14.431
β_3	2.025	0.297	6.817
n = 9	$R^2 = 0.988$	$\bar{R}^2 = 0.984$	g de l = 6

Así obtenemos la ecuación de regresión siguiente:

$$\hat{y} = 20.108 + 0.414\,x_1 + 2.025\,x_3$$

Se puede observar de la tabla 8.4 que el estadístico t para el coeficiente de regresión $\hat{\beta}_3$ es significativo ya que el valor calculado (6.817) es mayor que el de la tabla ($t_{(0.05/2,\ 6)} = 2.447$) por lo tanto seguimos con el proceso para ver si hay otras variables que se deben introducir en el modelo.

Paso 3: Para introducir la tercer variable, necesitamos saber cual es la que presenta un coeficiente de correlación más alto eliminando el efecto de x_1 y x_3, así:

Tabla 8.5 Correlaciones de Pearson manteniendo constante x_1 y x_3.

Variables de control	Variables	y	x2	x4
x1 y x3	y	1.000	0.458	-0.221
	x2	0.458	1.000	-0.318
	x4	-0.221	-0.318	1.000

En la tabla 8.5 se observa que la variable que se debe introducir es Talla al nacer (x_2) entonces con los estadísticos de la tabla 8.6 se escribe el siguiente modelo:

$$\hat{y} = 5.630 + 0.081\,x_1 + 0.771\,x_2 + 3.069\,x_3$$

Tabla 8.6 Estadísticos de resumen.

Parámetro	Estimado	Error estándar	Estadístico t.
β_0	5.630	12.707	0.443
β_1	0.081	0.290	0.279
β_2	0.771	0.669	1.153
β_3	3.069	0.951	3.229
n = 9	$R^2 = 0.991$	$\overline{R}^2 = 0.985$	g de l = 5

En la tabla 8.6 se observa que el estadístico t (1.153) para el coeficiente de regresión $\hat{\beta}_2$ es menor que el de la tabla ($t_{(0.05/2,\ 5)} = 2.571$), es decir que no es significativo al nivel de 5%, por lo que se termina el procedimiento de selección hacia adelante con:

$$\hat{y} = 20.108 + 0.414\,x_1 + 2.025\,x_3 \qquad \textbf{(8.1)}$$

Por tanto en el modelo final, no se incluye x_2 porque el estadístico t de $\hat{\beta}_2$ no es significativo.

Ejemplo 2: Método de eliminación hacia atrás.

Se ilustrará el método de eliminación hacia atrás haciendo uso de los datos mostrados en la tabla 8.1. Este método involucra los mismos conceptos de la selección hacia delante excepto que se inicia con todas las variables en el modelo.

Paso 1: Se ajusta una ecuación con las cuatro variables independientes, se calculan los estadísticos t para cada coeficiente como se muestra en la tabla 8.7 la ecuación de regresión es la siguiente:

$$\hat{y} = 7.148 + 0.100\,x_1 + 0.726\,x_2 + 3.076\,x_3 - 0.030\,x_4$$

Tabla 8.7 Estadísticos de resumen para el modelo con todas las variables.

Parámetro	Estimado	Error estándar	Estadístico t.
β_0	7.148	16.460	0.434
β_1	0.100	0.340	0.295
β_2	0.726	0.786	0.924
β_3	3.076	1.059	2.904
β_4	-0.030	0.166	-0.180
n = 9	$R^2 = 0.991$	$\overline{R}^2 = 0.982$	g de l = 4

En la tabla 8.7 se puede observar que el estadístico t para el coeficiente $\hat{\beta}_4$ es el más pequeño, por lo que se elimina la variable x_4 del modelo.

Paso 2: Se corre la regresión eliminando la variable x_4, al eliminar esta variable obtenemos los resultados que se muestran en la tabla 8.8 con los que se puede escribir la ecuación de regresión siguiente:

$$\hat{y} = 5.630 + 0.081\,x_1 + 0.771\,x_2 + 3.069\,x_3$$

Tabla 8.8 Estadísticos de resumen para el modelo sin x_4.

Parámetro	Estimado	Error estándar	Estadístico t.
β_0	5.630	12.707	0.443
β_1	0.081	0.290	0.279
β_2	0.771	0.669	1.153
β_3	3.069	0.951	3.229
n = 9	$R^2 = 0.991$	$\overline{R}^2 = 0.985$	g de l = 5

Ahora el estadístico t más pequeño que se tiene es el del coeficiente $\hat{\beta}_1$, por lo que eliminamos la variable x_1 del modelo.

Paso 3: Se corre la regresión eliminando las variables x_1 y x_4 obteniendo los estadísticos de la tabla 8.9 a partir de los cuales se escribe la ecuación siguiente:

$$\hat{y} = 2.183 + 0.958\,x_2 + 3.325\,x_3 \qquad \textbf{(8.2)}$$

Tabla 8.9 Estadísticos de resumen para el modelo sin x_1 y x_4.

Parámetro	Estimado	Error estándar	Estadístico t.
β_0	2.183	2.801	0.779
β_2	0.958	0.059	16.156
β_3	3.325	0.233	14.260
n = 9	$R^2 = 0.991$	$\overline{R}^2 = 0.987$	g de l = 6

El proceso termina porque los estadísticos t para las variables x_2 y x_3 son significativos, es decir, que los valores calculados para t son mayores que el de la tabla ($t_{(0.05/2,\ 6)} = 2.447$) por lo que el modelo que se inicio con cuatro variables independientes solamente queda con dos, este resultado es el que se muestra en la ecuación anterior (8.2).

Ejemplo 3: Método de regresión paso a paso.

Se utilizarán los datos de la tabla 8.1, para ejemplificar la regresión paso a paso. La regresión paso a paso se lleva a cabo con una ligera pero importante modificación del procedimiento de selección hacia adelante, los pasos son:

Paso 1: Calcular el coeficiente de correlación entre “y” y todas las variables “x”.

Tabla 8.10 Correlaciones de Pearson.

Variables	y	x_1	x_2	x_3	x_4
y	1.000	0.947	0.819	0.761	0.560
x_1	0.947	1.000	0.952	0.534	0.390
x_2	0.819	0.952	1.000	0.263	0.155
x_3	0.761	0.534	0.263	1.000	0.784
x_4	0.560	0.390	0.155	0.784	1.000

La variable con mayor coeficiente de correlación es x_1, así calculamos la regresión con x_1 como variable independiente y obtenemos.

$$\hat{y} = 19.011 + 0.518x_1$$
$$es(\hat{\beta}_1) = 0.066$$
$$t(\hat{\beta}_1) = 7.807$$

El valor de t calculado para el coeficiente $\hat{\beta}_1$ es significativo al nivel del 5% de significancia, es decir, que es mayor que el valor de la tabla ($t_{(0.05/2,\ 7)} = 2.365$), entonces la primera variable que entra en el modelo es x_1.

Paso 2: En esta etapa se ajustan tres regresiones, conteniendo todas a x_1. Los resultados importantes para las combinaciones (x_1, x_2), (x_1, x_3) y (x_1, x_4) son:

- Regresión de "y" contra x_1 y x_2.

$$\hat{y} = 44.100 + 0.983x_1 - 1.287x_2$$
$$es(\beta_1) = 0.124 \quad es(\beta_2) = 0.326$$
$$t(\hat{\beta}_1) = 7.932 \quad t(\hat{\beta}_2) = -3.941$$

- Regresión de “y” contra x_1 y x_3.

$$\hat{y} = 20.108 + 0.414x_1 + 2.025x_3$$
$$es(\beta_1) = 0.029 \quad es(\beta_3) = 0.297$$
$$t(\hat{\beta}_1) = 14.431 \quad t(\hat{\beta}_3) = 6.817$$

- Regresión de “y” contra x_1 y x_4.

$$\hat{y} = 9.324 + 0.470x_1 + 0.458x_4$$
$$es(\beta_1) = 0.059 \quad es(\beta_4) = 0.221$$
$$t(\hat{\beta}_1) = 7.912 \quad t(\hat{\beta}_4) = 2.074$$

De las tres regresiones anteriores se puede observar que solamente el estadístico t para el coeficiente de la variable x_3 es significativo al nivel de significancia del 5%, es decir, que el valor calculado es mayor que el de la tabla, por lo que la siguiente variable que se introduce en el modelo es x_3 junto con x_1.

Paso 3: Con x_1 y x_3 ya en el modelo, se ajustan dos regresiones conteniendo a x_1 y x_3 los resultados para las combinaciones (x_1, x_3, x_2) y (x_1, x_3, x_4) son:

- Regresión de “y” contra x_1, x_2 y x_3.

$$\hat{y} = 5.630 + 0.081x_1 + 0.771x_2 + 3.069x_3$$
$$es(\beta_1) = 0.290 \quad es(\beta_2) = 0.669 \quad es(\beta_3) = 0.951$$
$$t(\hat{\beta}_1) = 0.279 \quad t(\hat{\beta}_2) = 1.153 \quad t(\hat{\beta}_3) = 3.229$$

- Regresión de “y” contra x_1, x_3 y x_4.

$$\hat{y} = 21.874 + 0.413x_1 + 2.203x_3 - 0.079x_4$$
$$es(\beta_1) = 0.031 \quad es(\beta_3) = 0.0.472 \quad es(\beta_4) = 0.156$$
$$t(\hat{\beta}_1) = 13.460 \quad t(\hat{\beta}_3) = 4.667 \quad t(\hat{\beta}_4) = -0.508$$

Se puede observar en las ecuaciones anteriores que, ningún estadístico t de los coeficientes para las variables que se agregaron al modelo es significativo al nivel del 5%, por lo que el modelo final incluye únicamente las variables x_1 y x_3. Se encuentra que la ecuación de estimación es:

$$\hat{y} = 20.108 + 0.414\,x_1 + 2.025\,x_3 \qquad \textbf{(8.3)}$$

Y el coeficiente de determinación para este modelo es $R^2 = 0.988$.

No obstante que (x_1, x_3) es la combinación que selecciona la regresión paso a paso y la selección hacia adelante, no necesariamente es la combinación de dos variables que da el valor más grande de R^2.

Se puede observar en los métodos de selección de variables que, el orden en el que entran o salen las variables del modelo no necesariamente implica un orden de importancia de las variables. No es raro ver que una variable que entró al modelo al principio se vuelve sin importancia en un paso posterior; esto de hecho es un problema general con el procedimiento de selección hacia adelante, porque una vez agregada una variable no se puede eliminar en un paso posterior.

Nótese que la selección hacia adelante, la eliminación hacia atrás y la regresión paso a paso no necesariamente conducen a la misma elección del modelo final. La intercorrelación entre las variables afecta el orden de entrada y la eliminación, por ejemplo al usar los datos de la tabla 8.1 se vio que las variables seleccionadas por cada procedimiento fueron las siguientes:

Selección hacia adelante	(x_1, x_3)	ecuación	**(8.1)**
Eliminación hacia atrás	(x_2, x_3)	ecuación	**(8.2)**
Regresión paso a paso	(x_1, x_3)	ecuación	**(8.3)**

Se recomienda que se apliquen todos los procedimientos, para aprender algo acerca de la estructura de los datos, que pudiera haberse escapado si solamente se usa un procedimiento de selección de variables.

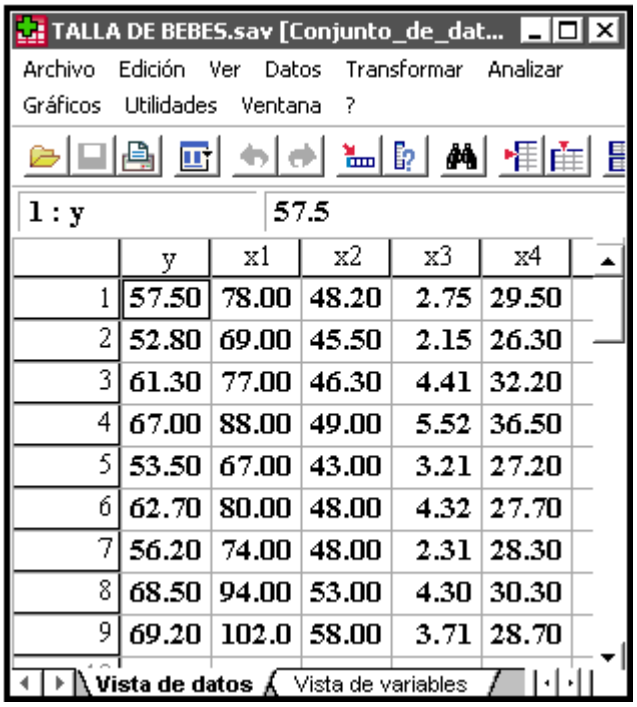
Los procedimientos de selección de variables se deben usar con precaución, la forma más recomendable de utilizarlos es: primeramente la regresión paso a paso seguido de la eliminación hacia atrás ya que frecuentemente la eliminación hacia atrás; se afecta menos por la estructura correlativa de las variables que la selección hacia adelante.

8.4 Métodos de Selección de Variables Haciendo Uso del SPSS V15.0.

Método de selección hacia adelante.

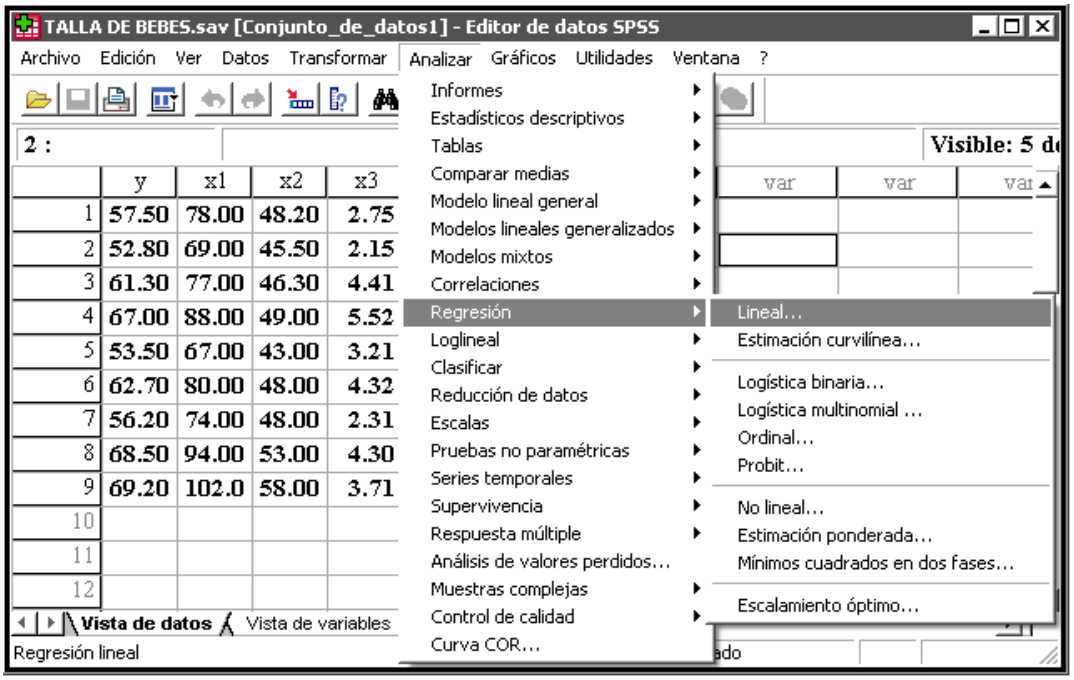
Haciendo uso de los datos de la tabla 8.1 se desarrolla el método de selección hacia adelante, siguiendo los pasos siguientes:

1. Se les da un nombre a las variables en estudio, en este caso las hemos representado como: y, x_1, x_2, x_3, x_4, en la tabla 8.1 se escribió que significa cada una de las variables, se digitan los datos para cada variable y se obtiene la ventana siguiente:



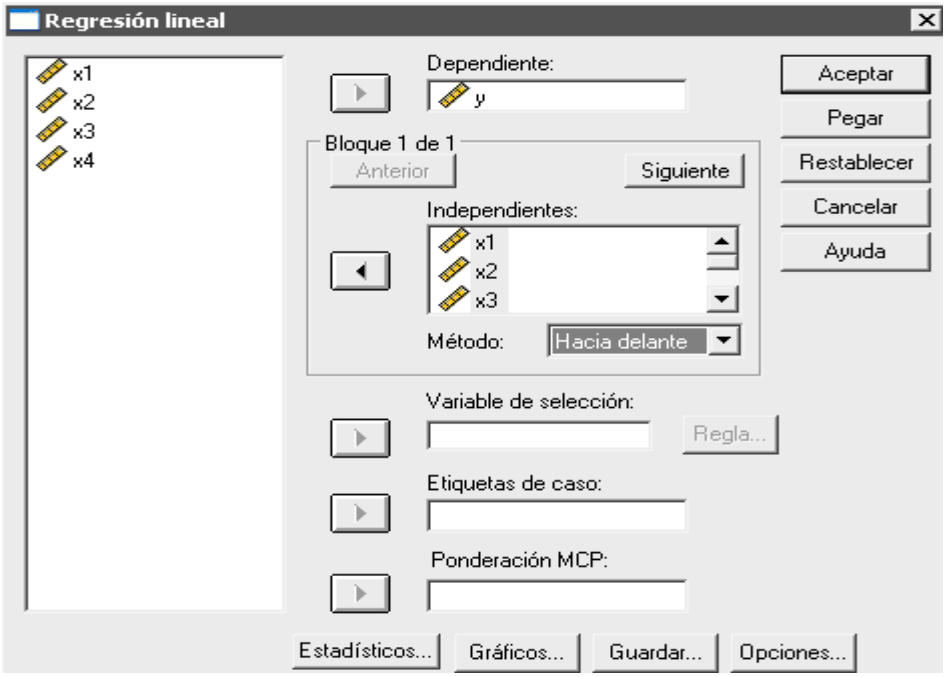
	y	x1	x2	x3	x4
1	57.50	78.00	48.20	2.75	29.50
2	52.80	69.00	45.50	2.15	26.30
3	61.30	77.00	46.30	4.41	32.20
4	67.00	88.00	49.00	5.52	36.50
5	53.50	67.00	43.00	3.21	27.20
6	62.70	80.00	48.00	4.32	27.70
7	56.20	74.00	48.00	2.31	28.30
8	68.50	94.00	53.00	4.30	30.30
9	69.20	102.0	58.00	3.71	28.70

2. En la barra de menú se selecciona la opción Analizar → Regresión → Lineal, como se muestra a continuación:

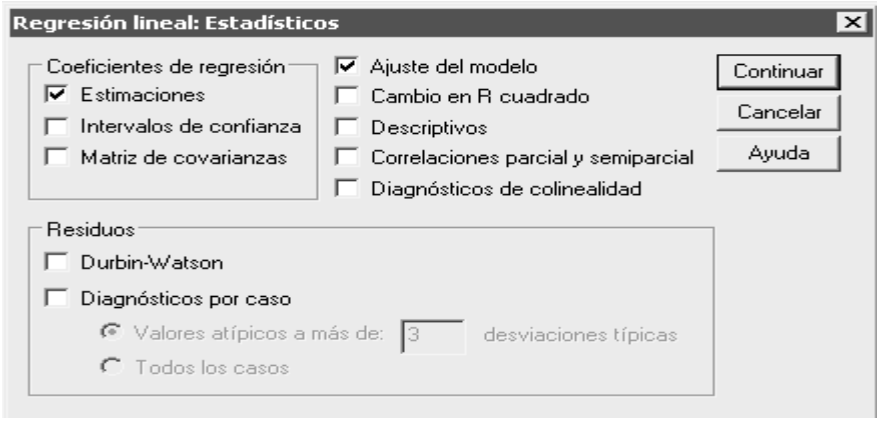


	y	x1	x2	x3
1	57.50	78.00	48.20	2.75
2	52.80	69.00	45.50	2.15
3	61.30	77.00	46.30	4.41
4	67.00	88.00	49.00	5.52
5	53.50	67.00	43.00	3.21
6	62.70	80.00	48.00	4.32
7	56.20	74.00	48.00	2.31
8	68.50	94.00	53.00	4.30
9	69.20	102.0	58.00	3.71
10				
11				
12				

3. Al hacer click en la opción lineal aparece la siguiente ventana en la cual se colocan las variables cada una en su lugar y se elige (en el recuadro de método) el método que se va ha utilizar en este caso se ha elegido el método hacia adelante.



4. Haciendo click en el botón Estadísticos en la parte inferior de la ventana anterior, se abrirá la ventana siguiente:



Donde las opciones Estimaciones y Ajuste del modelo están seleccionadas por determinación, pero hay muchas otras opciones disponibles, en este caso solamente necesitamos esos estadísticos dando click en continuar volvemos al cuadro Regresión lineal presentado en el paso 3.

5. Dando un click en aceptar del cuadro Regresión lineal dado en el paso 3 se obtienen los resultados siguientes:

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	x1	.	Hacia adelante
2	x3	.	Hacia adelante

a. Variable dependiente: y

En la tabla variable introducidas/eliminadas se muestra cuantos modelos se han formado, en este caso son 2 modelo 1 y modelo 2, y las variables introducidas son x_1 y x_3, se puede ver que no se muestran las variables eliminadas en esta tabla; se muestran más adelante; también se presenta el método que se ha utilizado.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida
1	.947[a]	.897	.882
2	.994[b]	.988	.984

a. Variables predictoras: (Constante), x1

b. Variables predictoras: (Constante), x1, x3

En la tabla resumen del modelo se muestra el valor de R, R^2 y $\bar{R}^2$ para el modelo 1 y para el modelo 2, el valor de R muestra que hay buena relación lineal entre la

variable independiente y la dependiente en los dos modelos de regresión. El valor de R^2 representa un buen ajuste en los dos modelos de regresión que se han formado.

A continuación se muestra la tabla coeficientes:

Coeficientes[a]

Modelo		Coeficientes no estandarizados		
		B	Error típ.	t
1	(Constante)	19.011	5.423	3.506
	x1	.518	.066	7.807
2	(Constante)	20.108	1.987	10.119
	x1	.414	.029	14.431
	x3	2.025	.297	6.817

a. Variable dependiente: y

Donde visualizamos los valores de los coeficientes que se han utilizado para formar los dos modelos de regresión, así como también los errores estándar de los coeficientes y los valores t significativos de cada uno de los coeficientes. Se puede observar que la primera variable que se utilizó para formar el modelo 1 es x_1, y para formar el modelo dos se han utilizado dos variables x_1 y x_3 como se mostró anteriormente, las demás variables no aparecen debido a que el estadístico t de los coeficientes no es significativo al nivel del 5%.

La tabla siguiente es la de análisis de varianza (ANOVA):

ANOVA[c]

Modelo		Suma de cuadrados	gl	Media cuadrática	F
1	Regresión	288.147	1	288.147	60.950
	Residual	33.093	7	4.728	
	Total	321.240	8		
2	Regresión	317.456	2	158.728	251.650
	Residual	3.784	6	.631	
	Total	321.240	8		

c. Variable dependiente: y

La cual muestra los resultados de sumas de cuadrados de las tres fuentes de variación (Regresión, Residual y Total), grados de libertad, media cuadrática y los valores de F calculados tanto para el modelo 1 como para el modelo 2. Los valores de F son significativos al nivel del 5%, es decir, que son más grandes que los de la tabla ($F_{(0.05,\ 1,\ 7)} = 5.59$ para el modelo 1 y $F_{(0.05,\ 2,\ 6)} = 5.14$ para el modelo 2).

Se presenta también la tabla de variables excluidas:

Variables excluidas[c]

Modelo		Beta dentro	t	Sig.	Correlación parcial
1	x2	-.893[a]	-3.941	.008	-.849
	x3	.357[a]	6.817	.000	.941
	x4	.225[a]	2.074	.083	.646
2	x2	.535[b]	1.153	.301	.458
	x4	-.039[b]	-.508	.633	-.221

a. Variables predictoras en el modelo: (Constante), x1

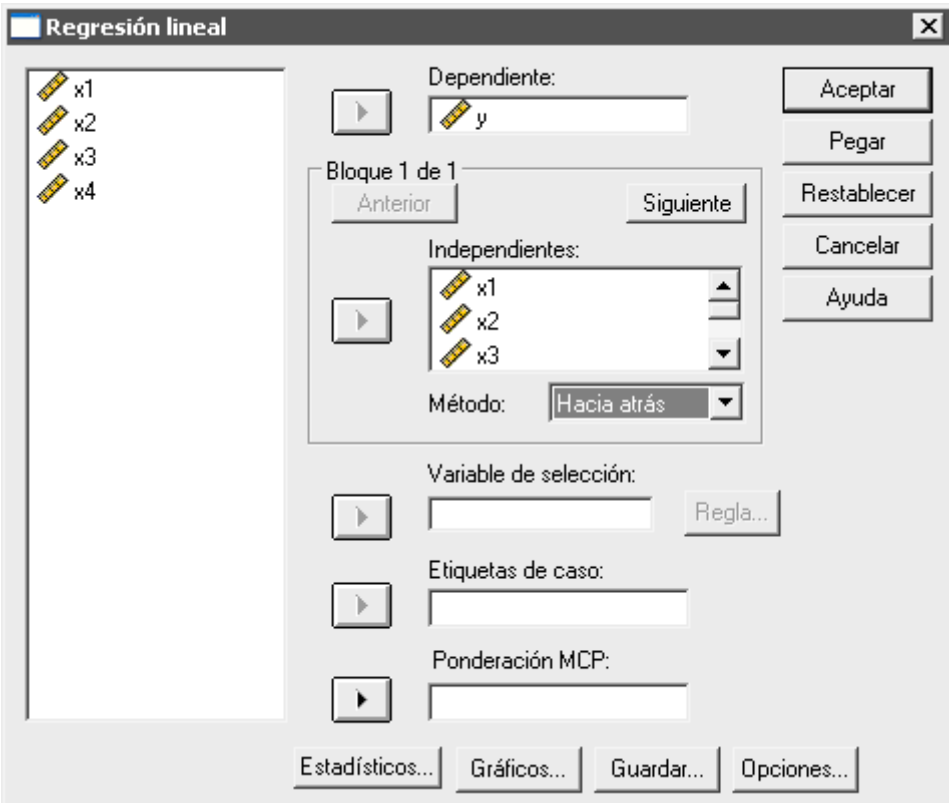
b. Variables predictoras en el modelo: (Constante), x1, x3

c. Variable dependiente: y

En la cual se muestra la información a cerca de las variables que no se agregan a la ecuación de regresión en cada paso o modelo. Esta información incluye el valor que tendría el coeficiente beta si se añadiera la variable a la ecuación. Obsérvese que en el modelo 1 se excluyeron tres variables y solamente se incluyó x_1, debido a que tiene un coeficiente de correlación igual a 0.947 como se mostró anteriormente en la tabla resumen del modelo, además se puede ver que las variables que se excluyen del modelo 2 son x_2 y x_4, la variable que se incluye en el modelo 2 es x_3 porque el coeficiente de correlación es 0.941 que es mayor que el de las demás variables, además de que el estadístico t de $\hat{\beta}_3$ es significativo al nivel del 5%.

Método de eliminación hacia atrás.

Para obtener la regresión con el método de eliminación hacia atrás, se realiza el paso 1, 2 y 4 como se hizo con el método de selección hacia adelante, pero en el paso 3 hay un cambio y es el que se muestra en el siguiente cuadro, donde el método elegido es el de eliminación hacia atrás.



Dando click en aceptar en el cuadro Regresión lineal se obtienen los resultados siguientes:

Variables introducidas/eliminadas[b]

Modelo	Variables introducidas	Variables eliminadas	Método
1	x4, x2, x3, x1[a]	.	Introducir
2	.	x4	Hacia atrás
3	.	x1	Hacia atrás

a. Todas las variables solicitadas introducidas

b. Variable dependiente: y

En la tabla variables introducidas/eliminadas se muestran tres modelos, donde en el modelo 1 se han introducido todas las variables y el método utilizado es introducir que es el que utiliza el SPSS por determinación, así teniendo un modelo con todas las variables se comienzan a eliminar las variables con el método de eliminación hacia atrás. Se puede observar en la tabla que en el modelo 2 se elimina la variable x_4 y en el modelo 3 se elimina la variable x_1.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida
1	.995[a]	.991	.982
2	.995[b]	.991	.985
3	.995[c]	.991	.987

a. Variables predictoras: (Constante), x4, x2, x3, x1

b. Variables predictoras: (Constante), x2, x3, x1

c. Variables predictoras: (Constante), x2, x3

Los resultados mostrados en la tabla Resumen del modelo incluyen dos conjuntos de datos, uno concerniente a la correlación múltiple, y otro, a la regresión múltiple. Estos resultados indican que la correlación múltiple de "y" con las variables independientes en el modelo 1 es de 0.995, pero resulta que en el modelo 2 y 3 este resultado no cambia, esto se da porque las variables que se han eliminado x_1 y x_4 aportan nada al modelo, es decir que, en este ejemplo el valor de R y el valor de R^2 = 0.991 no sufren ningún cambio cuando se eliminan las dos variables. Se puede observar que el $\overline{R}^2$ sí muestra cambios, esto es porque este coeficiente es ajustado por los grados de libertad y como se dijo antes los grados de libertad se obtienen de la diferencia de la muestra y el número

de parámetros a estimar en el modelo (n – número de parámetros estimados), debido a esto el valor de $\overline{R}^2$ es distinto para los tres modelos.

La tabla que se muestra a continuación es la de coeficientes:

Coeficientes[a]

Modelo		Coeficientes no estandarizados		t
		B	Error típ.	
1	(Constante)	7.148	16.460	.434
	x1	.100	.340	.295
	x2	.726	.786	.924
	x3	3.076	1.059	2.904
	x4	-.030	.166	-.180
2	(Constante)	5.630	12.707	.443
	x1	.081	.290	.279
	x2	.771	.669	1.153
	x3	3.069	.951	3.229
3	(Constante)	2.183	2.801	.779
	x2	.958	.059	16.156
	x3	3.325	.233	14.260

a. Variable dependiente: y

Y se utiliza para ir formando las ecuaciones y para decidir que variable es la que entra primero, se tiene el modelo 1 con todas las variables y se elimina la variable que tiene el estadístico t del parámetro más pequeño; quedando el modelo 3 o modelo final solamente con las dos variables (x_2 y x_3) que contribuyen de forma significativa a la predicción.

Se tiene también la tabla ANOVA siguiente:

ANOVA[d]

Modelo		Suma de cuadrados	gl	Media cuadrática	F
1	Regresión	318.274	4	79.569	107.323
	Residual	2.966	4	.741	
	Total	321.240	8		
2	Regresión	318.250	3	106.083	177.413
	Residual	2.990	5	.598	
	Total	321.240	8		
3	Regresión	318.204	2	159.102	314.389
	Residual	3.036	6	.506	
	Total	321.240	8		

d. Variable dependiente: y

Que muestra los resultados de sumas de cuadrados de las tres fuentes de variación (Regresión, Residual y Total), grados de libertad, media cuadrática y los valores de F calculados para los tres modelos que se han formado. Los valores de F son significativos al nivel del 5%, es decir, que son más grandes que los de la tabla.

Finalmente se tiene la tabla de variables excluidas:

Variables excluidas[c]

Modelo		Beta dentro	t	Sig.	Correlación parcial
2	x4	-.015[a]	-.180	.866	-.090
3	x4	-.007[b]	-.103	.922	-.046
	x1	.148[b]	.279	.791	.124

a. Variables predictoras en el modelo: (Constante), x2, x3, x1

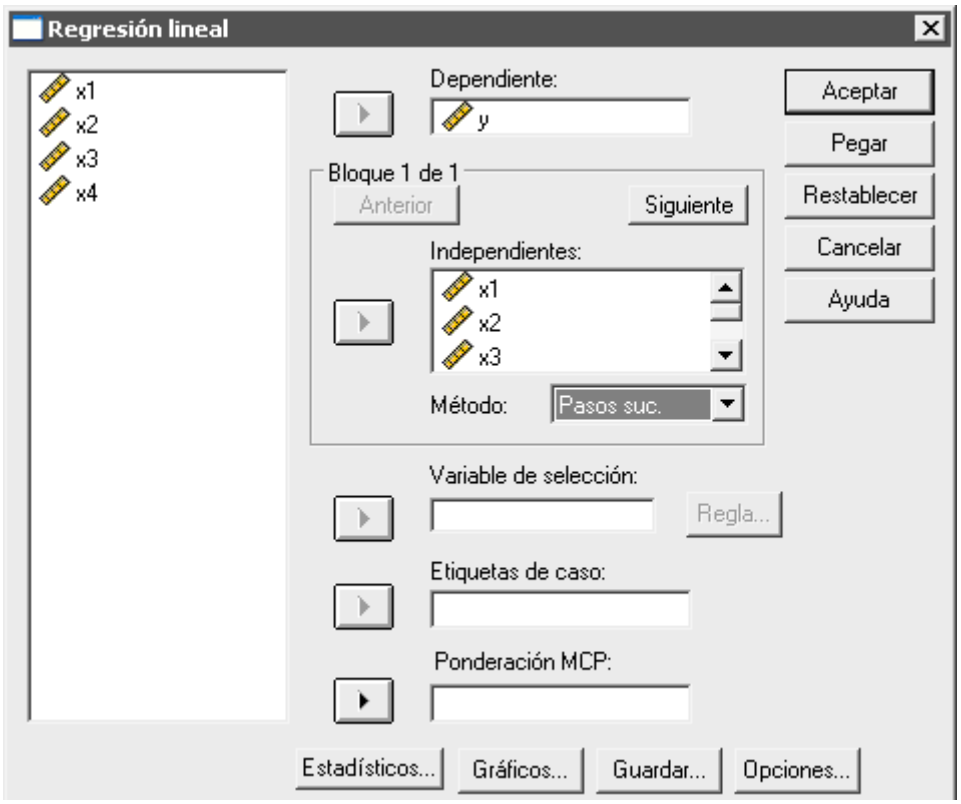
b. Variables predictoras en el modelo: (Constante), x2, x3

c. Variable dependiente: y

En la que los resultados muestran que las variables se excluyeron del modelo porque no son significativas, es decir, que ni la variable x_1 ni x_4 contribuyen de forma significativa a la predicción. Los estadísticos t de los parámetros son muy pequeños también las correlaciones parciales, por lo tanto el proceso termina.

Método de regresión paso a paso.

Para obtener la regresión con el método de regresión paso a paso, se realiza el paso 1, 2 y 4 como se hizo con el método de selección hacia adelante, ya el paso 3 tiene un cambio y es el que se muestra en el siguiente cuadro, donde el método elegido es el de pasos sucesivos o paso a paso.



Para obtener los resultados que se muestra a continuación se ha dado un click en la opción aceptar del cuadro Regresión lineal mostrado en el cuadro anterior.

Variables introducidas/eliminadas[a]

Modelo	Variables introducidas	Variables eliminadas	Método
1	x1	.	Por pasos
2	x3	.	Por pasos

a. Variable dependiente: y

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida
1	.947[a]	.897	.882
2	.994[b]	.988	.984

a. Variables predictoras: (Constante), x1

b. Variables predictoras: (Constante), x1, x3

Coeficientes[a]

Modelo		Coeficientes no estandarizados		
		B	Error típ.	t
1	(Constante)	19.011	5.423	3.506
	x1	.518	.066	7.807
2	(Constante)	20.108	1.987	10.119
	x1	.414	.029	14.431
	x3	2.025	.297	6.817

a. Variable dependiente: y

ANOVA[c]

Modelo		Suma de cuadrados	gl	Media cuadrática	F
1	Regresión	288.147	1	288.147	60.950
	Residual	33.093	7	4.728	
	Total	321.240	8		
2	Regresión	317.456	2	158.728	251.650
	Residual	3.784	6	.631	
	Total	321.240	8		

c. Variable dependiente: y

Variables excluidas[c]

Modelo		Beta dentro	t	Sig.	Correlación parcial
1	x2	-.893[a]	-3.941	.008	-.849
	x3	.357[a]	6.817	.000	.941
	x4	.225[a]	2.074	.083	.646
2	x2	.535[b]	1.153	.301	.458
	x4	-.039[b]	-.508	.633	-.221

a. Variables predictoras en el modelo: (Constante), x1

b. Variables predictoras en el modelo: (Constante), x1, x3

c. Variable dependiente: y

Los resultados obtenidos con el método de regresión paso a paso, son los mismos que se obtuvieron con el método de selección hacia adelante para este ejemplo en particular, esto significa que, para otros ejemplos puede variar. En este caso es igual porque las dos variables independientes que los dos métodos eligen para formar el modelo final son x_1 y x_3.

Ejercicios 8

1. El departamento de personal de una empresa utilizó a doce individuos en un estudio para determinar la relación entre su comportamiento hacia el trabajo (y) y las calificaciones de cuatro pruebas (x_1, x_2, x_3 y x_4). Los datos son los siguientes:

y	x_1	x_2	x_3	x_4
11.2	56.5	71.0	38.5	43.0
14.5	59.5	72.5	38.2	44.8
17.2	69.2	76.0	42.5	49.0
17.8	74.5	79.5	43.4	56.3
19.3	81.2	84.0	47.5	60.2
24.5	88.0	86.2	47.4	62.0
21.2	78.2	80.5	44.5	58.1
16.9	69.0	72.0	41.8	48.1
14.8	58.1	68.0	42.1	46.0
20.0	80.5	85.0	48.1	60.3
13.2	58.3	71.0	37.5	47.1
22.5	84.0	87.2	51.0	65.2

   Realizar el análisis de regresión haciendo uso de los tres métodos de selección de variables mostrados en este Capítulo.

2. Con los datos mostrados en el ejercicio 5 del Capítulo 5 realizar el análisis de regresión con los tres métodos de selección de variables mostrados en este Capítulo.

Apéndice A: Elementos de Álgebra Matricial.

Este apéndice nos ofrece los principales elementos del álgebra matricial, necesarios para comprender de una forma más fácil el Capítulo 5. La exposición no es compleja ni rigurosa.

A.1 Definiciones.

Matriz: La matriz es una disposición rectangular de números u otros elementos en filas y columnas. Es decir una matriz de orden o dimensión M por N (escrita M * N) es un conjunto de M * N elementos distribuidos en M filas y N columnas. De este modo, simbolizando las matrices con negritas, una matriz **A** (de orden M * N) puede expresarse así:

$$\mathbf{A} = [a_{ij}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1N} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{M1} & a_{M2} & a_{M3} & \cdots & a_{MN} \end{bmatrix}$$

Donde a_{ij} es el elemento que aparece en la fila i-ésima y en la columna j-ésima de **A** y donde $[a_{ij}]$ es una expresión abreviada de la matriz **A** cuyo elemento típico es a_{ij}. El orden o dimensión de una matriz, es decir, el número de filas y columnas, se escribe a menudo debajo de ella con el fin de facilitar la referencia.

Ejemplos

$$\underset{2*3}{\mathbf{A}} = \begin{bmatrix} 5 & 2 & 3 \\ 3 & 1 & 6 \end{bmatrix} \qquad \underset{3*3}{\mathbf{B}} = \begin{bmatrix} 1 & 2 & 4 \\ -1 & 0 & 7 \\ 9 & 8 & 11 \end{bmatrix}$$

Vector columna: La matriz que consta de M filas y solo de una columna se denomina vector columna. Denotando los vectores con negritas minúsculas, veamos el siguiente

Ejemplo

$$\underset{4*1}{\mathbf{x}} = \begin{bmatrix} 4 \\ 5 \\ 9 \\ 3 \end{bmatrix}$$

Vector fila: La matriz que consta de una sola fila y N columnas se denomina vector fila.

Ejemplos

$$\underset{1*4}{\mathbf{x}} = \begin{bmatrix} 2 & 5 & -4 \end{bmatrix} \qquad \underset{1*5}{\mathbf{y}} = \begin{bmatrix} 5 & -9 & 6 & 10 \end{bmatrix}$$

Transposición: La transposición de una matriz **A** de orden M * N, se denota $\mathbf{A}'$ (se lee **A** prima o **A** transpuesta), y es una matriz de N * M que se obtiene intercambiando las filas y las columnas de **A**; es decir la i-ésima fila de **A** se convierte en la j-ésima columna de $\mathbf{A}'$.

Ejemplos

$$\underset{3*2}{\mathbf{A}} = \begin{bmatrix} 4 & 5 \\ 3 & 1 \\ 5 & 0 \end{bmatrix} \qquad \underset{2*3}{\mathbf{A}'} = \begin{bmatrix} 4 & 3 & 5 \\ 5 & 1 & 0 \end{bmatrix}$$

Como los vectores son un tipo especial de matriz, la transpuesta de un vector fila es un vector columna y la transpuesta de un vector columna es un vector fila. Así:

$$\mathbf{x} = \begin{bmatrix} 4 \\ 5 \\ 9 \end{bmatrix} \quad \text{y} \quad \mathbf{x}' = \begin{bmatrix} 4 & 5 & 9 \end{bmatrix}$$

De aquí en adelante denotaremos los vectores filas con letras primas.

Submatriz: Dada una matriz **A** de orden M * N, si descartamos todas las filas menos r y todas las columnas menos s, la matriz resultante r*s se llamará submatriz de **A**. De este modo, si

$$\underset{3*3}{\mathbf{A}} = \begin{bmatrix} 3 & 5 & 7 \\ 5 & 2 & 1 \\ 8 & 2 & 1 \end{bmatrix}$$

Y descartamos la tercera fila y la tercera columna de esta matriz obtendremos:

$$\underset{2*2}{\mathbf{B}} = \begin{bmatrix} 3 & 5 \\ 5 & 2 \end{bmatrix}$$

Que es una submatriz de **A** y cuyo orden es 2*2.

A.2 Tipos de Matrices.

Matriz cuadrada: Es la matriz que tiene el mismo número de filas y columnas.

Ejemplos

$$\underset{2*2}{\mathbf{A}} = \begin{bmatrix} 3 & 4 \\ 5 & 6 \end{bmatrix} \qquad \underset{3*3}{\mathbf{B}} = \begin{bmatrix} 3 & 5 & 8 \\ 7 & 3 & 1 \\ 4 & 5 & 0 \end{bmatrix}$$

Matriz diagonal: La matriz cuadrada que presenta por lo menos un elemento diferente de cero en la diagonal principal (que va de la esquina superior izquierda a la esquina inferior derecha) y ceros en las demás posiciones, recibe el nombre de diagonal.

Ejemplos

$$\underset{2*2}{\mathbf{A}} = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix} \qquad \underset{3*3}{\mathbf{B}} = \begin{bmatrix} -2 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Matriz escalar: La matriz diagonal cuyos elementos de la diagonal son todos iguales se le llama matriz escalar. Un ejemplo es la matriz de varianza – covarianza de las perturbaciones poblacionales del modelo de regresión lineal clásico; o sea,

$$\text{var}-\text{cov}(\boldsymbol{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix}$$

Matriz simétrica: A la matriz cuadrada en que los elementos que van por encima de la diagonal principal son imágenes reflejas de los elementos que van por debajo de ella, se le denomina matriz simétrica. Alternativamente, una matriz simétrica es aquella cuya transpuesta es igual a ella misma; o sea que $\mathbf{A} = \mathbf{A}'$. Es decir, que los elementos a_{ij} de **A** son iguales a los elementos a_{ji} de $\mathbf{A}'$.

Ejemplo: la matriz de varianza – covarianza dada anteriormente y la matriz de correlaciones dada en (5.49).

Matriz nula: La matriz cuyos elementos son todos cero se denomina matriz nula y se simboliza con **0**.

Vector nulo: El vector fila o columna cuyos elementos son todos cero se denomina vector nulo y también se designa con **0.**

Matrices iguales: Dos matrices son iguales si son del mismo orden y sus elementos correspondientes son iguales; es decir, que $a_{ij} = b_{ij}$ para todo i y todo j.

Ejemplo

Si

$$\underset{3*3}{\mathbf{A}} = \begin{bmatrix} 3 & 4 & 5 \\ 0 & -1 & 2 \\ 5 & 1 & 3 \end{bmatrix} \qquad \text{y} \qquad \underset{3*3}{\mathbf{B}} = \begin{bmatrix} 3 & 4 & 5 \\ 0 & -1 & 2 \\ 5 & 1 & 3 \end{bmatrix}$$

Entonces, **A = B.**

A.3 Operaciones Matriciales.

Suma de matrices.

Sea **A** = $[a_{ij}]$ y **B** = $[b_{ij}]$. Si **A** y **B** son del mismo orden, la suma de matrices se define como:

$$\mathbf{C = A + B}$$

Donde **C** es del mismo orden que **A** y **B** y se obtiene como $c_{ij} = a_{ij} + b_{ij}$ para todo i y para todo j; es decir, que **C** se obtiene sumando los elementos correspondientes de cada matriz.

Si se puede hacer dicha suma, se dice que **A** y **B** son conformables para la suma.

Ejemplo

Si

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 \end{bmatrix} \quad \text{y} \quad \mathbf{B} = \begin{bmatrix} 1 & 0 & -1 & 3 \\ -2 & 0 & 1 & 5 \end{bmatrix}$$

Entonces

$$\mathbf{C} = \begin{bmatrix} 3 & 3 & 3 & 8 \\ 4 & 7 & 9 & 14 \end{bmatrix}$$

Resta de matrices.

La resta de matrices sigue el mismo principio que la suma excepto que **C = A – B**; es decir, restamos los elementos de la matriz **B** de los elementos correspondientes de la matriz **A**, siempre que **A** y **B** sean del mismo orden.

Multiplicación escalar.

Para multiplicar una matriz **A** por un escalar λ (número real), multiplicamos cada elemento de la matriz por λ:

$$\lambda\mathbf{A} = [\lambda a_{ij}]$$

Ejemplo:

Si $\lambda = 2$ y $\mathbf{A} = \begin{bmatrix} -3 & 5 \\ 8 & 7 \end{bmatrix}$, entonces $2\mathbf{A} = \begin{bmatrix} -6 & 10 \\ 16 & 14 \end{bmatrix}$

Multiplicación de matrices.

Sea **A** una matriz de M * N y **B** otra de N * P. El producto **AB** (en este orden) se define como la matriz **C** de orden M * P tal que:

$$c_{ij} = \sum_{k=1}^{N} a_{ik} b_{kj} \quad i = 1, 2,\ldots, M \quad \text{y} \quad j = 1, 2,\ldots, P$$

Es decir, el elemento en la i-ésima fila y en la j-ésima columna de **C** se obtiene multiplicando los elementos de la i-ésima fila de **A** por los elementos correspondientes de la j-ésima columna de **B** y sumando todos los términos; esta operación se conoce como la regla de multiplicación de fila por columna. Entonces, para obtener c_{11}, el elemento de la primera fila y la primera columna de **C**, multiplicamos los elementos de la primera fila de **A** por los elementos correspondientes de la primera columna de **B** y sumamos todos los elementos. De igual manera, para obtener c_{12}, multiplicamos los elementos de la primera fila de **A** por los elementos correspondientes a la segunda columna de **B** y sumamos todos los términos, y así sucesivamente.

Nótese que para que la multiplicación exista, las matrices **A** y **B** deben ser conformables con respecto a la multiplicación; es decir, el número de columnas de **A** debe ser igual al número de filas de **B**.

Ejemplos

Si

$$\underset{2*3}{\mathbf{A}} = \begin{bmatrix} 3 & 4 & 7 \\ 5 & 6 & 1 \end{bmatrix} \qquad \text{y} \qquad \underset{3*2}{\mathrm{B}} = \begin{bmatrix} 2 & 1 \\ 3 & 5 \\ 6 & 2 \end{bmatrix}$$

Entonces,

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} (3*2) + (4*3) + (7*6) & (3*1) + (4*5) + (7*2) \\ (5*2) + (6*3) + (1*6) & (5*1) + (6*5) + (1*2) \end{bmatrix}$$

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 60 & 37 \\ 34 & 37 \end{bmatrix}$$

Pero si,

$$\underset{2*3}{\mathbf{A}} = \begin{bmatrix} 3 & 4 & 7 \\ 5 & 6 & 1 \end{bmatrix} \quad \text{y} \quad \underset{2*2}{\mathbf{B}} = \begin{bmatrix} 2 & 3 \\ 5 & 6 \end{bmatrix}$$

El producto **AB** no está definido dado que **A** y **B** no son conformables con respecto a la multiplicación.

Propiedades de la multiplicación de matrices.

1. La multiplicación de matrices no es conmutativa. Es decir, en general **AB ≠ BA**. Por lo tanto, el orden en el cual se multiplican las matices es muy importante. **AB** significa que **A** es postmultiplicada por **B** o que **B** es premultiplicada por **A**.
2. Si **AB** y **BA** existen, las matrices resultantes no son del mismo orden.
3. Si **A** y **B** son matrices cuadradas tales que **AB** y **BA** están ambas definidas, las matrices resultantes no son necesariamente iguales.

   Ejemplo

   Si

   $$\mathbf{A} = \begin{bmatrix} 4 & 7 \\ 3 & 2 \end{bmatrix} \qquad \text{y} \qquad \mathbf{B} = \begin{bmatrix} 1 & 5 \\ 6 & 8 \end{bmatrix}$$

   Entonces,

   $$\mathbf{AB} = \begin{bmatrix} 46 & 76 \\ 15 & 31 \end{bmatrix} \qquad \text{y} \qquad \mathbf{BA} = \begin{bmatrix} 19 & 17 \\ 48 & 58 \end{bmatrix}$$

   Por lo tanto **AB ≠ BA**. Un ejemplo de **AB = BA** se da cuando ambas matrices **A** y **B** son matrices cuadradas y unitarias.

4. Un vector fila postmultiplicado por un vector columna es un escalar. Considere por ejemplo los Mínimos Cuadrados Ordinarios e_1, e_2,…, e_n. Siendo **e** un vector fila y $\mathbf{e}'$ un vector columna, se tiene:

$$\mathbf{e}'\mathbf{e} = \begin{bmatrix} e_1 & e_2 & e_3 & \cdots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

$$\mathbf{e}'\mathbf{e} = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2$$

$$\mathbf{e}'\mathbf{e} = \sum_{i=1}^{n} e_i^2 \text{ un escalar.}$$

5. Un vector columna postmultiplicado por un vector fila es una matriz. Considere por ejemplo las perturbaciones poblacionales del modelo de regresión lineal, ε_1, ε_2,…, ε_n. Siendo $\boldsymbol{\varepsilon}$ un vector columna y $\boldsymbol{\varepsilon}'$ un vector fila, se tiene:

$$\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_N \end{bmatrix}$$

$$\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' = \begin{bmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_N \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \cdots & \varepsilon_2\varepsilon_N \\ \cdots & \cdots & \cdots & \cdots \\ \varepsilon_N\varepsilon_1 & \varepsilon_N\varepsilon_2 & \cdots & \varepsilon_N^2 \end{bmatrix}$$

Que es una matriz de orden N * N, observemos que la matriz anterior es simétrica.

6. Una matriz postmultiplicada por un vector columna es un vector columna.

7. Un vector fila postmultiplicado por una matriz es un vector fila.
8. La multiplicación de matrices es asociativa; es decir, (**AB**)**C** = **A**(**BC**), donde **A** es de M * N, **B** de N * P y **C** de P * K.
9. La multiplicación de matrices es distributiva con respecto a la suma; es decir, **A**(**B** + **C**) = **AB** + **AC** y (**B** + **C**)**A** = **BA** + **CA**.

Transposición de matrices.

Ya hemos definido el proceso de transformación de matrices como el intercambio de las filas y las columnas de una matriz o vector. Ahora, enunciaremos algunas de las propiedades de la transposición de matrices.

1. La transpuesta de una matriz transpuesta es la matriz original misma: $(\mathbf{A}')' = \mathbf{A}$.
2. Si **A** y **B** son conformables para la suma, entonces $\mathbf{C} = \mathbf{A} + \mathbf{B}$ y $\mathbf{C}' = (\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$. O sea, que la transpuesta de la suma de dos matrices es la suma de las matrices transpuestas.
3. Si **AB** está definida, entonces $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$. Es decir, la transpuesta del producto de dos matrices es el producto de sus transpuestas en orden inverso. Esto puede generalizarse así: $(\mathbf{ABCD})' = \mathbf{D}'\mathbf{C}'\mathbf{B}'\mathbf{A}'$.
4. La transpuesta de una matriz identidad **I** es la misma matriz identidad; esto es, $\mathbf{I}' = \mathbf{I}$.
5. La transpuesta de un escalar es el mismo escalar. Si λ es un escalar $\lambda' = \lambda$.

6. La transpuesta de $(\lambda\mathbf{A})'$ es $\lambda\mathbf{A}'$ donde λ es un escalar.

7. Si $\mathbf{A}$ es una matriz cuadrada tal que $\mathbf{A} = \mathbf{A}'$, entonces $\mathbf{A}$ es una matriz simétrica.

Inversión de matrices.

La inversa de una matriz **A**, que se marca con $\mathbf{A}^{-1}$ (se lee inversa de **A**), si existe, es una matriz única tal que:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Donde **I** es una matriz identidad cuyo orden es el mismo de **A**.

Ejemplo

Si $\mathbf{A} = \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix}$ Entonces, $\mathbf{A}^{-1} = \begin{bmatrix} -1 & 1/2 \\ 6/8 & -1/4 \end{bmatrix}$, así $\mathbf{A}\mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathbf{I}$

Después de estudiar el tema de los determinantes, veremos cómo se calcula la matriz inversa. Por lo pronto anotaremos las propiedades siguientes.

1. $(\mathbf{A}\mathbf{B})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$; o sea, la inversa del producto de dos matrices es igual al producto de sus inversas en el orden contrario.

2. $(\mathbf{A}^{-1})' = (\mathbf{A}')^{-1}$; es decir, la transpuesta de **A** inversa es igual a la inversa de **A** transpuesta.

A.4 Determinantes.

A cualquier matriz cuadrada **A**, corresponde un número escalar conocido como el determinante de la matriz que se designa det **A** o por medio del símbolo $|\mathbf{A}|$, donde $|\ |$

significa “el determinante de”. Note que una matriz no tiene un valor numérico por si misma, pero el determinante de ella sí es un número.

Ejemplo

Si

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & -7 \\ 2 & 5 & 0 \\ 3 & 8 & 6 \end{bmatrix} \text{ entonces, } |\mathbf{A}| = \begin{vmatrix} 1 & 3 & -7 \\ 2 & 5 & 0 \\ 3 & 8 & 6 \end{vmatrix}$$

El $|\mathbf{A}|$ en el ejemplo se llama determinante de orden 3 puesto que está asociado con una matriz de orden 3 * 3.

Evaluación de un determinante.

El proceso de encontrar el valor numérico de un determinante recibe el nombre de evaluación, expansión o reducción del determinante. Esto se hace manipulando los datos de la matriz de manera muy bien definida.

Evaluación de un determinante de 2 * 2:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

Su determinante se evalúa como sigue:

$$|\mathbf{A}| = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

Que se obtiene multiplicando en cruz los elementos de la diagonal principal y restando de ellos el producto de los elementos de la otra diagonal de la matriz **A**.

Evaluación de un determinante de 3 * 3:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Su determinante se calcula como se muestra a continuación:

$$|\mathbf{A}| = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}$$

Un examen cuidadoso de la evaluación de un determinante de 3 * 3 nos muestra que:

1. Cada término en la expansión del determinante contiene un solo elemento de cada fila y de cada columna.
2. El número de elementos de cada término es el mismo que el número de filas o columnas de la matriz. De modo que un determinante de 2 * 2 tiene dos elementos en cada término de su expansión, uno de 3 * 3 tiene tres elementos en cada término de su expansión y así sucesivamente.
3. Los términos de la expansión tienen los signos + y – alternados.
4. Un determinante de 2 * 2 tiene dos términos en su expansión, y uno de 3 * 3 tiene 6. La regla general es:

El determinante de N * N tiene N! = N(N - 1)(N - 2)…3 * 2 * 1 términos en su expansión, donde N! significa “N factorial”. Siguiendo esta regla, un determinante de orden 5 * 5 tendrá 5* 4*3*2*1 = 120 términos en su expansión.

Propiedades de los determinantes.

1. La matriz cuyo determinante es igual a cero se denomina matriz singular, mientras que la matriz con un determinante distinto de cero se llama matriz no singular. La inversa de una matriz no existe cuando su determinante es cero, es decir, cuando se trata de una matriz singular.
2. Si todos los elementos de una fila de **A** son cero, su determinante es cero. Entonces,

$$|\mathbf{A}| = \begin{vmatrix} 0 & 0 & 0 \\ 3 & 4 & 5 \\ 6 & 7 & 8 \end{vmatrix} = 0$$

3. $|\mathbf{A}'| = |\mathbf{A}|$; es decir, el determinante de **A** es igual al determinante de **A** transpuesta.
4. Si intercambiamos dos filas o dos columnas de una matriz, el signo de su determinante cambia.

Ejemplo

$$\mathbf{A} = \begin{bmatrix} 6 & 9 \\ -1 & 4 \end{bmatrix} \qquad \text{y} \qquad \mathbf{B} = \begin{bmatrix} -1 & 4 \\ 6 & 9 \end{bmatrix}$$

Donde **B** se obtiene al intercambiar las filas de **A**, luego

$$|\mathbf{A}| = 24 - (-9) = 33 \qquad \text{y} \qquad |\mathbf{B}| = -9 - (24) = -33$$

5. Si cada elemento de una fila o de una columna se multiplica por un escalar λ, esto equivale a multiplicar $|\mathbf{A}|$ por λ.

   Ejemplo

   $$\text{Si } \lambda = 5 \text{ y } \mathbf{A} = \begin{bmatrix} 5 & -8 \\ 2 & 4 \end{bmatrix}$$

   Multiplicando la primera fila de **A** por 5 se obtiene:

   $$\mathbf{B} = \begin{bmatrix} 25 & -40 \\ 2 & 4 \end{bmatrix}$$

   Se puede ver que $|\mathbf{A}| = 36$ y $|\mathbf{B}| = 180$ que es igual a $5|\mathbf{A}|$.

6. Si dos filas o dos columnas de una matriz son idénticas, su determinante es cero.
7. Si una fila o una columna de una matriz es múltiplo de la otra fila o columna, respectivamente, su determinante es cero. Entonces, si:

   $$\mathbf{A} = \begin{bmatrix} 4 & 8 \\ 2 & 4 \end{bmatrix}$$

   Donde la primera fila de **A** es dos veces la segunda, $|\mathbf{A}| = 0$. De forma más general, si cualquier fila o columna de una matriz es una combinación lineal de las otras filas o columnas, su determinante es cero.

8. $|\mathbf{AB}| = |\mathbf{A}||\mathbf{B}|$; es decir, el determinante del producto de dos matrices es igual al producto de sus determinantes.

Rango de una matriz.

El rango de una matriz es el orden de la submatriz cuadrada más grande cuyo determinante es diferente de cero.

Ejemplo

$$\mathbf{A} = \begin{bmatrix} 3 & 6 & 6 \\ 0 & 4 & 5 \\ 3 & 2 & 1 \end{bmatrix}$$

Se puede ver que $|\mathbf{A}| = 0$. En otras palabras, **A** es una matriz singular. Entonces, aunque su orden es 3 * 3, su rango es menor que 3. En efecto, su rango es 2, por cuanto podemos encontrar una submatriz de 2 * 2 cuyo determinante es diferente de cero. Por ejemplo, si borramos la primera fila y la primera columna de **A** obtenemos:

$$\mathbf{B} = \begin{bmatrix} 4 & 5 \\ 2 & 1 \end{bmatrix}$$

Cuyo determinante es -6, que es diferente de cero. Por lo tanto, el rango de **A** es 2. Como se anotó anteriormente, la inversa de una matriz singular no existe; por lo tanto para que la inversa de una matriz **A** de N * N exista, su rango debe ser **A**. Si es inferior a N, **A** es singular.

Menor.

Si la i-ésima fila y la j-ésima columna de una matriz de N * N se borran, o no se tienen en cuenta, el determinante de la matriz resultante se denomina el menor del elemento a_{ij} (el elemento situado en la intersección de i-ésima fila con la j-ésima columna) y se marca como $|\mathbf{M}_{ij}|$.

Ejemplo

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

El menor de a_{11} es:

$$|\mathbf{M}_{11}| = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} = a_{22}a_{33} - a_{23}a_{32}$$

De igual manera, el menor de a_{21} es:

$$|\mathbf{M}_{21}| = \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} = a_{12}a_{33} - a_{13}a_{32}$$

Los menores de otros elementos de **A** se hallan de modo semejante.

Cofactor.

El cofactor del elemento a_{ij} de una matriz **A** de N * N, denominado c_{ij}, se define como:

$$c_{ij} = (-1)^{i+j} |\mathbf{M}_{ij}|$$

En otras palabras, el cofactor es un menor con el signo correspondiente. El signo es positivo si i + j es par y negativo si i + j es impar. De este modo, el cofactor del elemento a_{11} de la matriz **A** de 3 * 3, dada anteriormente es $a_{22}a_{33} - a_{23}a_{32}$, mientras que el

elemento a_{21} es $-(a_{22}a_{33} - a_{23}a_{32})$, ya que la suma de los subíndices 2 y 1 es 3 que es un número impar.

Matriz de cofactores.

Reemplazando los elementos a_{ij} de la matriz **A,** por sus cofactores obtenemos la matriz que se conoce como matriz de cofactores, que se denota como (cof **A**).

Matriz adjunta.

La matriz adjunta, que se marca como (adj **A**), es la transpuesta de la matriz de cofactores; es decir (adj **A**) = (cof **A**)′.

A.5 Cálculo de la inversa de una matriz cuadrada.

Si **A** es una matriz cuadrada no singular (es decir, $|\mathbf{A}| \neq 0$), su inversa $\mathbf{A}^{-1}$ se puede hallar de la siguiente manera:

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}(\text{adj } \mathbf{A})$$

Las etapas que se requieren para calcularla son las siguientes:

1. Hallar el determinante de **A**. Si es diferente de cero, siga con la etapa 2.
2. Reemplazar cada elemento a_{ij} de **A** por su cofactor para obtener la matriz de cofactores.
3. Transponer la matriz de cofactores y obtener la matriz adjunta.
4. Dividir cada elemento de la matriz adjunta por $|\mathbf{A}|$.

   Ejemplo: Supongamos que queremos hallar la inversa de la matriz siguiente:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 7 & 4 \\ 2 & 1 & 3 \end{bmatrix}$$

Etapa 1: Primero hallamos el determinante de la matriz. Aplicando las reglas para expandir un determinante de 3 * 3 como se vio antes, así obtenemos que:

$$|\mathbf{A}| = -24$$

Etapa 2: Obtenemos ahora la matriz de cofactores, o sea, **C**.

$$\mathbf{C} = \begin{bmatrix} \begin{vmatrix} 7 & 4 \\ 1 & 3 \end{vmatrix} & -\begin{vmatrix} 5 & 4 \\ 2 & 3 \end{vmatrix} & \begin{vmatrix} 5 & 7 \\ 2 & 1 \end{vmatrix} \\ -\begin{vmatrix} 2 & 3 \\ 1 & 3 \end{vmatrix} & \begin{vmatrix} 1 & 3 \\ 2 & 3 \end{vmatrix} & -\begin{vmatrix} 1 & 2 \\ 2 & 1 \end{vmatrix} \\ \begin{vmatrix} 2 & 3 \\ 7 & 4 \end{vmatrix} & -\begin{vmatrix} 1 & 3 \\ 5 & 4 \end{vmatrix} & \begin{vmatrix} 1 & 2 \\ 5 & 7 \end{vmatrix} \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 17 & -7 & -9 \\ -3 & -3 & 3 \\ -13 & 11 & -3 \end{bmatrix}$$

Etapa 3: Transponiendo la matriz de cofactores obtenemos la matriz adjunta:

$$(\text{adj } \mathbf{A}) = \begin{bmatrix} 17 & -3 & -13 \\ -7 & -3 & 11 \\ -9 & 3 & -3 \end{bmatrix}$$

Etapa 4: Dividimos los elementos de la (adj **A**) por el valor del determinante -24, y se obtiene:

$$\mathbf{A}^{-1} = -\frac{1}{24}\begin{bmatrix} 17 & -3 & -13 \\ -7 & -3 & 11 \\ -9 & 3 & -3 \end{bmatrix}$$

$$\mathbf{A}^{-1} = \begin{bmatrix} -\frac{17}{24} & \frac{3}{24} & \frac{13}{24} \\ \frac{7}{24} & \frac{3}{24} & -\frac{11}{24} \\ \frac{9}{24} & -\frac{3}{24} & \frac{3}{24} \end{bmatrix}$$

Se puede verificar que:

$$\mathbf{A}\mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Que es la matriz identidad.

A.6 Derivación de matrices.

Para seguir el material del apéndice 5, es necesario conocer algunas reglas de la derivación de matrices.

Regla 1: Si $\mathbf{a}' = [a_1, a_2, \ldots, a_n]$ es un vector fila de números, y

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Es un vector columna de variables $x_1, x_2, \ldots, x_n$, entonces,

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{a}'\mathbf{x}) = \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Considere la matriz $\mathbf{x}'\mathbf{A}\mathbf{x}$ tal que:

$$\mathbf{x}'\mathbf{A}\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Entonces,

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}'\mathbf{x}) = 2\mathbf{A}\mathbf{x}$$

Que es un vector columna de n elementos, o

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}'\mathbf{x}) = 2\mathbf{x}'\mathbf{A}$$

Que es un vector fila de n elementos.

Apéndice B: Tablas Estadísticas.

Tabla B.1 Distribución normal estándar acumulada.

Tabla B.2 Puntos porcentuales de la distribución t.

Tabla B.3 Puntos porcentuales de la distribución F.

Tabla B.4 Puntos porcentuales de la distribución χ^2.

Tabla B.5 Estadístico de Durbin-Watson d: Puntos de significancia de d_L y d_U para el nivel de significancia $\alpha = 0.05$.

Tabla B.1 Distribución normal estándar acumulada N(0, 1).

$$f(Z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \ell^{-u^2/2} du$$

Z	0.0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.504	0.508	0.512	0.516	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.591	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.648	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.67	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.695	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.719	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.758	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.791	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.834	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.877	0.879	0.881	0.883
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.898	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.937	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.975	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.983	0.9834	0.9838	0.9842	0.9846	0.985	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.989
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.992	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.994	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.996	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.997	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.998	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.999	0.999

Tabla B.2 Puntos porcentuales de la distribución t.

	α									
g de l (v)	**0.4**	**0.3**	**0.2**	**0.15**	**0.1**	**0.075**	**0.05**	**0.025**	**0.01**	**0.005**
1	0.325	0.727	1.376	1.963	3.078	4.165	6.314	12.706	31.821	63.656
2	0.289	0.617	1.061	1.386	1.886	2.282	2.92	4.303	6.965	9.925
3	0.277	0.584	0.978	1.25	1.638	1.924	2.353	3.182	4.541	5.841
4	0.271	0.569	0.941	1.19	1.533	1.778	2.132	2.776	3.747	4.604
5	0.267	0.559	0.92	1.156	1.476	1.699	2.015	2.571	3.365	4.032
6	0.265	0.553	0.906	1.134	1.44	1.65	1.943	2.447	3.143	3.707
7	0.263	0.549	0.896	1.119	1.415	1.617	1.895	2.365	2.998	3.499
8	0.262	0.546	0.889	1.108	1.397	1.592	1.86	2.306	2.896	3.355
9	0.261	0.543	0.883	1.1	1.383	1.574	1.833	2.262	2.821	3.25
10	0.26	0.542	0.879	1.093	1.372	1.559	1.812	2.228	2.764	3.169
11	0.26	0.54	0.876	1.088	1.363	1.548	1.796	2.201	2.718	3.106
12	0.259	0.539	0.873	1.083	1.356	1.538	1.782	2.179	2.681	3.055
13	0.259	0.538	0.87	1.079	1.35	1.53	1.771	2.16	2.65	3.012
14	0.258	0.537	0.868	1.076	1.345	1.523	1.761	2.145	2.624	2.977
15	0.258	0.536	0.866	1.074	1.341	1.517	1.753	2.131	2.602	2.947
16	0.258	0.535	0.865	1.071	1.337	1.512	1.746	2.12	2.583	2.921
17	0.257	0.534	0.863	1.069	1.333	1.508	1.74	2.11	2.567	2.898
18	0.257	0.534	0.862	1.067	1.33	1.504	1.734	2.101	2.552	2.878
19	0.257	0.533	0.861	1.066	1.328	1.5	1.729	2.093	2.539	2.861
20	0.257	0.533	0.86	1.064	1.325	1.497	1.725	2.086	2.528	2.845
21	0.257	0.532	0.859	1.063	1.323	1.494	1.721	2.08	2.518	2.831
22	0.256	0.532	0.858	1.061	1.321	1.492	1.717	2.074	2.508	2.819
23	0.256	0.532	0.858	1.06	1.319	1.489	1.714	2.069	2.5	2.807
24	0.256	0.531	0.857	1.059	1.318	1.487	1.711	2.064	2.492	2.797
25	0.256	0.531	0.856	1.058	1.316	1.485	1.708	2.06	2.485	2.787
26	0.256	0.531	0.856	1.058	1.315	1.483	1.706	2.056	2.479	2.779
27	0.256	0.531	0.855	1.057	1.314	1.482	1.703	2.052	2.473	2.771
28	0.256	0.53	0.855	1.056	1.313	1.48	1.701	2.048	2.467	2.763
29	0.256	0.53	0.854	1.055	1.311	1.479	1.699	2.045	2.462	2.756
30	0.256	0.53	0.854	1.055	1.31	1.477	1.697	2.042	2.457	2.75
> 30	0.253	0.524	0.842	1.036	1.282	1.44	1.645	1.96	2.326	2.576

Tabla B.3 Puntos porcentuales de la distribución F.

$F_{(0.05,\ v_1,\ v_2)}$

Grados de libertad para el denominador (v_2)	Grados de libertad para el numerador (v_1)									
	1	2	3	4	5	6	7	8	9	10
1	161.4	199.5	215.7	224.6	230.2	234	236.8	238.9	240.5	241.9
2	18.51	19	19.16	19.25	19.3	19.33	19.35	19.37	19.38	19.4
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.1	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.5	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.1	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.48	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.8	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.9	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.1	2.87	2.71	2.6	2.51	2.45	2.39	2.35
25	4.24	3.39	2.99	2.76	2.6	2.49	2.4	2.34	2.28	2.24
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
50	4.03	3.18	2.79	2.56	2.4	2.29	2.2	2.13	2.07	2.03
60	4	3.15	2.76	2.53	2.37	2.25	2.17	2.1	2.04	1.99
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.0	1.95
100	3.94	3.09	2.7	2.46	2.31	2.19	2.1	2.03	1.97	1.93
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Tabla B.3 (Continuación)

$$F_{(0.05,\ v_1,\ v_2)}$$

Grados de libertad para el denominador (v_2)	Grados de libertad para el numerador (v_1)						
	12	15	20	24	30	40	50
1	243.9	245.9	248	249.1	250.1	251.1	252
2	19.41	19.43	19.45	19.46	19.46	19.47	19.48
3	8.74	8.7	8.66	8.63	8.62	8.59	8.58
4	5.91	5.86	5.8	5.77	5.75	5.72	5.7
5	4.68	4.62	4.56	4.52	4.5	4.46	4.44
6	4.0	3.94	3.87	3.83	3.81	3.77	3.75
7	3.57	3.51	3.44	3.4	3.38	3.34	3.32
8	3.28	3.22	3.15	3.11	3.08	3.04	3.02
9	3.07	3.01	2.94	2.89	2.86	2.83	2.8
10	2.91	2.85	2.77	2.73	2.7	2.66	2.64
11	2.49	2.72	2.65	2.61	2.57	2.53	2.51
12	2.69	2.62	2.54	2.5	2.47	2.43	2.4
13	2.60	2.53	2.46	2.42	2.38	2.34	2.31
14	2.53	2.46	2.39	2.35	2.31	2.27	2.24
15	2.48	2.4	2.33	2.28	2.25	2.2	2.18
16	2.42	2.35	2.28	2.24	2.19	2.15	2.12
17	2.38	2.31	2.23	2.19	2.15	2.10	2.08
18	2.34	2.27	2.19	2.15	2.11	2.06	2.04
19	2.31	2.23	2.16	2.11	2.07	2.03	2.00
20	2.28	2.2	2.12	2.07	2.04	1.99	1.97
25	2.16	2.09	2.01	1.96	1.92	1.87	1.84
30	2.09	2.01	1.93	1.88	1.84	1.79	1.76
40	2.0	1.92	1.84	1.78	1.74	1.69	1.66
50	1.95	1.87	1.78	1.73	1.69	1.63	1.6
60	1.92	1.84	1.75	1.69	1.65	1.59	1.56
80	1.88	1.79	1.7	1.64	1.6	1.54	1.51
100	1.85	1.77	1.68	1.62	1.57	1.52	1.48
120	1.83	1.75	1.66	1.6	1.55	1.5	1.46

Tabla B.4 Puntos porcentuales de la distribución χ^2.

g de l \ α	0.995	0.990	0.975	0.950	0.900	0.100	0.050	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.1	0.21	4.61	5.99	7.38	9.21	10.6
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.2	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.6	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.3
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.6	5.23	6.26	7.26	8.55	22.31	25.0	27.49	30.58	32.8
16	5.14	5.81	6.91	7.96	9.31	23.54	26.3	28.85	32.0	34.27
17	5.7	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.2	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.0
21	8.03	8.9	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.4
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.8
23	9.26	10.2	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.4	13.85	15.66	33.2	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
30	13.79	14.95	16.79	18.49	20.6	40.26	43.77	46.98	50.89	53.67
35	17.19	18.51	20.57	22.47	24.8	46.06	49.8	53.2	57.34	60.27
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
45	24.31	25.9	28.37	30.61	33.35	57.51	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	37.69	63.17	67.5	71.42	76.15	79.49
55	31.73	33.57	36.4	38.96	42.06	68.8	73.31	77.38	82.29	85.75
60	35.53	37.48	40.48	43.19	46.46	74.4	79.08	83.3	88.38	91.95
65	39.38	41.44	44.6	47.45	50.88	79.97	84.82	89.18	94.42	98.1
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
75	47.21	49.48	52.94	56.05	59.79	91.06	96.22	100.8	106.4	110.3
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
85	55.17	57.63	61.39	64.75	68.78	102.1	107.5	112.4	118.2	122.3

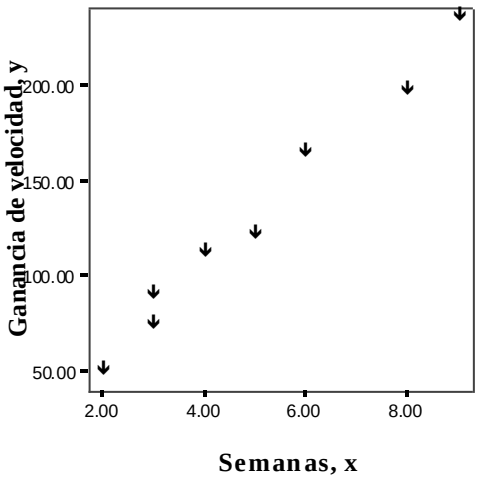
Tabla B.4 Estadísticos de Durbin-Watson d: puntos de significancia de d_L y d_U con $\alpha = 0.05$.

n	k = 1		k = 2		k = 3		k = 4		k = 5	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77

Respuestas a los ejercicios planteados.

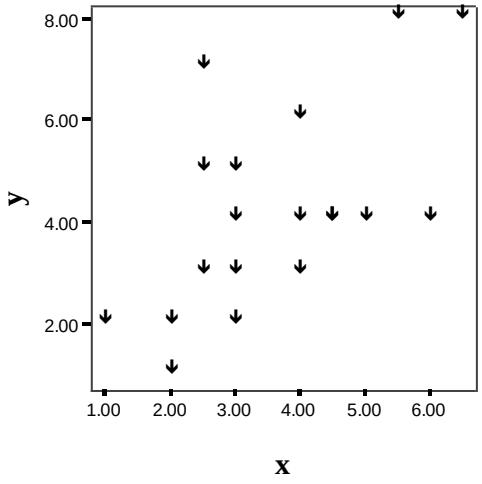
Capítulo 1.

1. a).



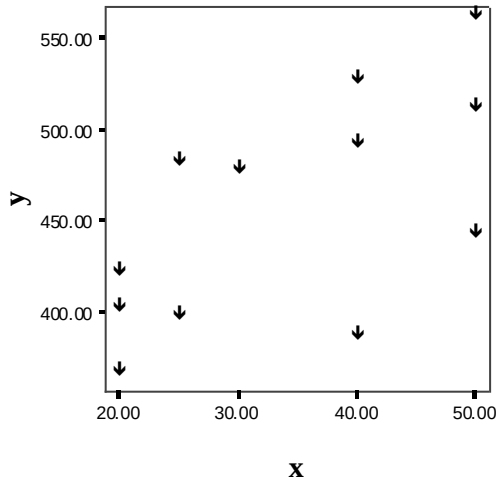
b). $r = 0.992$

2. a).



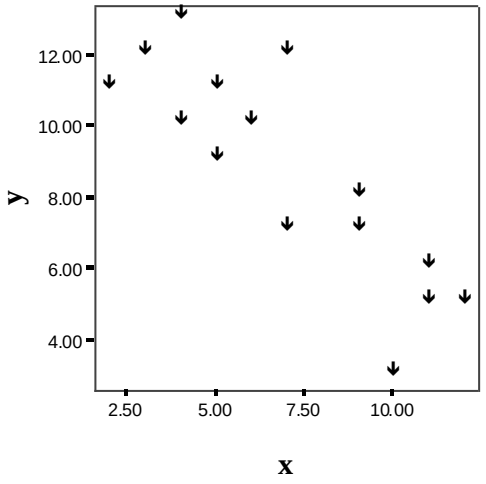
b). r = 0.587.

3. a).



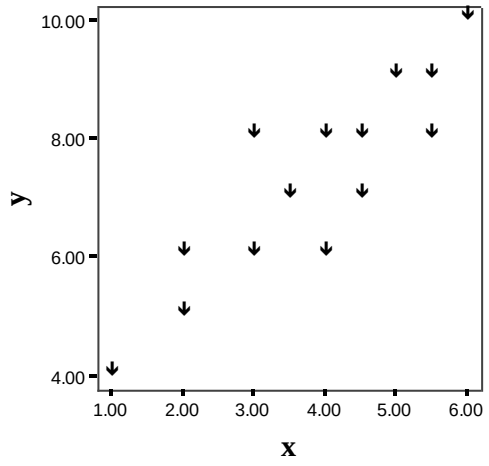
b). r = 0.635.

4. a).



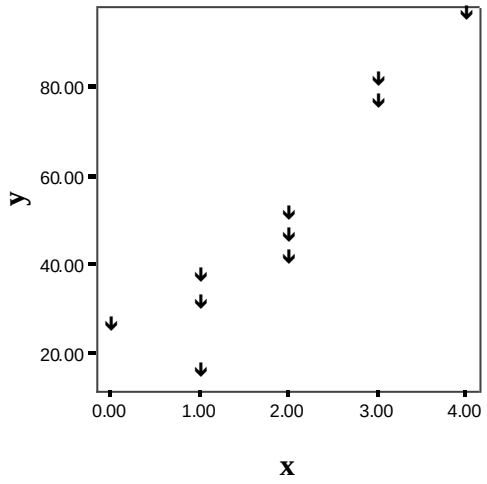
b). r = -0.847

5. a).



b). r = 0.882

6. a).



b). r = 0.973

Capítulo 2.

1.

a) $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i = 1.375 + 0.120\, x_i$.

b) $\hat{\sigma}^2 = 0.109$.

c) $\text{var}(\hat{\beta}_0) = 0.136161$, $\text{var}(\hat{\beta}_1) = 0.000676$, $\text{es}(\hat{\beta}_0) = 0.369$ y $\text{es}(\hat{\beta}_1) = 0.026$.

d) $r^2 = 0.782$.

e) Se rechaza la hipótesis nula $H_0 : \beta_1 = 0$ para la pendiente.

f) $0.473 \le \beta_0 \le 2.277$, $0.057 \le \beta_1 \le 0.184$ y $0.045 \le \sigma^2 \le 0.527$.

2.

a) $\hat{y} = 77.863 + 11.801 x_i$.

b) Se rechaza la hipótesis nula $H_0 : \beta_1 = 0$ para la pendiente.

c) $r^2 = 0.389$.

d) $4.479 \le \beta_1 \le 19.123$

3.

a)

Fuente de Variación	**Suma de Cuadrados**	**Grados de Libertad**	**Cuadrado Medio**	**F_0**
Regresión	25.580	1	25.58	2.207
Residual	301.384	26	11.592	
Total	326.964	27		

No se rechaza la hipótesis nula $H_0 : \beta_1 = 0$

b) $-0.005 \le \beta_1 \le 0.001$

4.

a) $\hat{y} = -1145.793 + 4.318\,x_i$.

b) $4.1697 \le \beta_1 \le 4.466$

5.

a) $\hat{y} = -95.044 + 0.516\,x_i$.

b) $-112.674 \le \beta_0 \le -77.414$ y $0.453 \le \beta_1 \le 0.579$.

6.

a) $\hat{\sigma}^2 = 0.188$.

b) $r = 0.999$.

7.

a) $\hat{y} = 1.306 + 0.791\,x_i$.

b) $\hat{\sigma}^2 = 2.721$.

c) $\text{var}(\hat{\beta}_0) = 1.054729$, $\text{var}(\hat{\beta}_1) = 0.070225$, $\text{es}(\hat{\beta}_0) = 1.027$ y $\text{es}(\hat{\beta}_1) = 0.265$.

d) $r^2 = 0.344$

e) Se rechaza la hipótesis nula $H_0 : \beta_1 = 0$ para la pendiente.

f) $-0.861 \le \beta_0 \le 3.473$, $0.232 \le \beta_1 \le 1.350$ y $1.532 \le \sigma^2 \le 6.12$.

8.

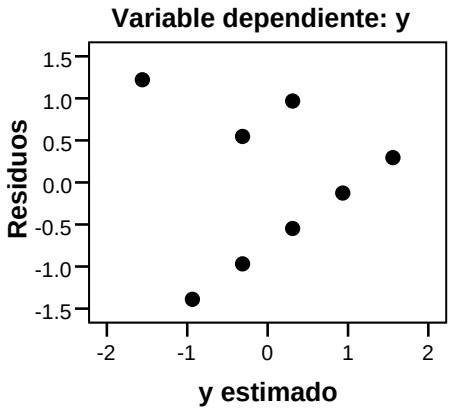
a) $\hat{y} = -2.821 + 8.104\,x_i$.

b) $\hat{y}_{(x=17)} = -2.821 + 8.104(17) = 134.947 \approx 135$ se venderán aproximadamente 135 aparatos.

Capítulo 3.

1.

a) Gráfico de los residuos



2.

a) $\hat{y} = -96.112 + 0.979\, x_i$.

c) $57.468 \le E(y_0 \mid x_0 = 162) \le 67.503$, $46.330 \le y_0 \mid x_0 = 162 \le 78.64$

Capítulo 4.

1.

a) $\hat{y} = -22.993 + 1.396\, x_{1i} + 0.218\, x_{2i}$.

b) $r_{yx_1 \bullet x_2} = 0.671$, $r_{yx_2 \bullet x_1} = 0.818$, $r_{x_2 x_1 \bullet y} = -0.293$.

c) $R^2 = 0.873$.

d) $\hat{y}_{(x_1=45,\ x_2=250)} = -22.993 + 1.396(45) + 0.218(250) = 94.327$

e) Se rechaza la hipótesis nula para las pruebas individual y global de los coeficientes de regresión β_1 y β_2.

f) $-64.995 \le \beta_0 \le 19.009$, $0.018 \le \beta_1 \le 2.773$, $0.081 \le \beta_2 \le 0.354$.

2.

a) $\hat{y} = 12.685 + 0.196\,x_{1i} + 0.728\,x_{2i}$.

b) $R^2 = 0.631$.

c) $\overline{R}^2 = 0.558$.

d) $\hat{y}_{(x_1=145,\ x_2=145)} = 12.685 + 0.196(145) + 0.728(145) = 146.665$.

e) Se acepta la hipótesis nula para la prueba individual y se rechaza la hipótesis nula para la prueba global de los coeficientes de regresión para $\alpha = 0.05$.

f) $-61.245 \le \beta_0 \le 86.615$, $-0.897 \le \beta_1 \le 1.288$, $-0.372 \le \beta_2 \le 1.827$.

3.

a) $\hat{y} = 6.900 - 0.511\,x_{1i} + 1.214\,x_{2i}$.

b) $R^2 = 0.996$.

c) $\overline{R}^2 = 0.994$

d) Se rechaza la hipótesis nula para la prueba individual y global de los coeficientes (β_1 y β_2) de regresión para $\alpha = 0.05$.

4.

a) $\hat{y} = 0.580 + 2.712\,x_{1i} + 2.050\,x_{2i}$.

b) $R^2 = 1.00$.

c) $\overline{R}^2 = 1.00$.

d) Se rechaza la hipótesis nula individual y global de los coeficientes de regresión β_1 y β_2 para $\alpha = 0.05$.

e) $-0.855 \leq \beta_0 \leq 2.015$, $2.234 \leq \beta_1 \leq 3.190$, $1.936 \leq \beta_2 \leq 2.163$.

5.

a) $\hat{y} = -16.365 + 1.109\, x_{1i} + 0.045\, x_{2i}$.

b) $R^2 = 0.732$.

c) $\hat{y}_{(x_1=72,\ x_2=17)} = -16.365 + 1.109\,(72) + 0.045\,(17) = 64.248$

6.

a) $\hat{y} = 0.987 + 0.940\, x_{1i} - 0.009\, x_{2i}$.

b) $R^2 = 0.968$.

7.

a) $\hat{y} = 44.100 + 0.983\, x_{1i} - 1.287\, x_{2i}$.

b) $R^2 = 0.971$.

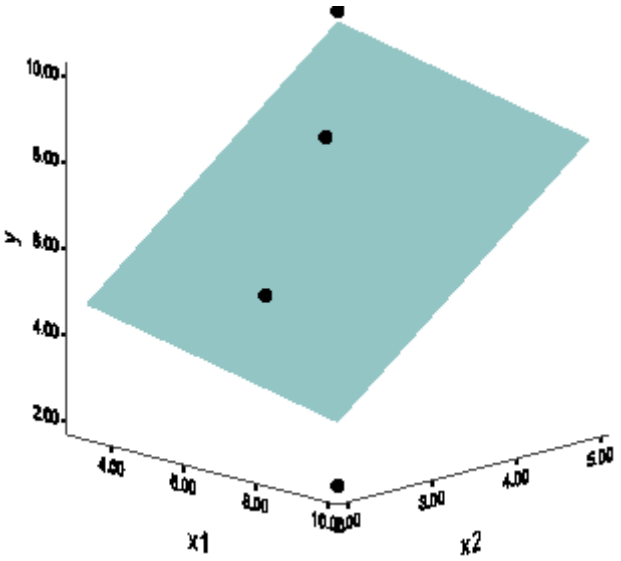
c) $\overline{R}^2 = 0.962$.

d) Se rechaza la hipótesis nula individual y global de los coeficientes de regresión β_1 y β_2 para $\alpha = 0.05$.

e) $26.781 \leq \beta_0 \leq 61.419$, $0.680 \leq \beta_1 \leq 1.286$, $-2.086 \leq \beta_2 \leq -0.488$.

8.

a)



b) $\hat{y} = 1.849 - 0.177\,x_{1i} + 1.691\,x_{2i}$.

c) $R^2 = 0.769$.

d) $\overline{R}^2 = 0.538$.

e) Se acepta la hipótesis nula individual y global de los coeficientes de regresión β_1 y β_2 para $\alpha = 0.05$.

f) $-35.549 \le \beta_0 \le 39.246$, $-3.111 \le \beta_1 \le 2.757$, $-4.870 \le \beta_2 \le 8.252$.

g) $\hat{y}_{(x_1=12,\ x_2=2)} = 1.849 - 0.177\,(12) + 1.691\,(2) = 3.107$.

Capítulo 5.

1.

a) $\hat{y} = 6.900 - 0.511\,x_{1i} + 1.214\,x_{2i}$.

b) $R^2 = 0.996$.

c) Se rechaza la hipótesis nula para la prueba individual y global de los coeficientes (β_1 y β_2) de regresión para $\alpha = 0.05$.

d) $5.957 \leq \beta_0 \leq 7.842$, $-0.576 \leq \beta_1 \leq -0.447$, $1.121 \leq \beta_2 \leq 1.308$.

2.

a) $\hat{y} = 44.100 + 0.983\,x_{1i} - 1.287\,x_{2i}$.

b) $R^2 = 0.971$.

c) Se rechaza la hipótesis nula individual y global de los coeficientes de regresión β_1 y β_2 para $\alpha = 0.05$.

d) $26.781 \leq \beta_0 \leq 61.419$, $0.680 \leq \beta_1 \leq 1.286$, $-2.086 \leq \beta_2 \leq -0.488$.

3.

a) $\hat{y} = -117.121 + 0.410\,x_{1i} + 21.325\,x_{2i} + 7.060\,x_{3i}$.

b) No se rechaza la hipótesis nula individual y global de los coeficientes de regresión β_1, β_2 y β_3 para $\alpha = 0.05$.

c) $-1289.826 \leq \beta_0 \leq 1055.584$, $-0.009 \leq \beta_1 \leq 0.829$, $-77.124 \leq \beta_2 \leq 119.774$, $-45.147 \leq \beta_3 \leq 59.267$.

4.

a) $\hat{y} = 60.014 + 0.240\,x_{1i} + 10.718\,x_{2i} - 0.751\,x_{3i}$.

b) $R^2 = 0.845$.

c) No se rechaza la hipótesis nula individual y se rechaza la hipótesis global de los coeficientes de regresión β_1, β_2 y β_3 para $\alpha = 0.05$.

Capítulo 6.

1.

a) $\hat{y} = -43.593 + 0.620\,x_i + 2.592\,D_i$.

b) No se rechaza la hipótesis nula para el coeficiente de regresión β_2.

c) $-2.078 \leq \beta_2 \leq 7.262$.

d) $\hat{y} = -29.543 + 0.530\,x_i - 83.168\,D_i + 0.533\,x_i D_i$

2.

a) $\hat{y} = 33.619 - 0.046\,x_i - 0.517\,D_i$.

b) $\hat{y} = 42.920 - 0.117\,x_i - 13.483\,D_i + 0.082\,x_i D_i$

e) Se rechaza la hipótesis nula individual y global de los coeficientes de regresión estimados en el modelo del literal b) para un $\alpha = 0.05$.

f) $37.312 \leq \beta_0 \leq 48.527$, $-0.157 \leq \beta_1 \leq -0.076$, $-21.363 \leq \beta_2 \leq -5.603$, $0.038 \leq \beta_3 \leq 0.125$.

3.

a) $\hat{y} = 6597.974 + 0.041\,x_i$.

b) El coeficiente de ventas de la regresión estimada en a) es 0.041 y el de la regresión estimada en la ecuación 6.25 es de 0.036 se puede decir que son estadísticamente iguales.

4. $\hat{y} = -52.150 + 0.223\,x_i + 645.950\,(x_i - x^*)D_i$

Capítulo 7.

1.

a) $\hat{y} = 1.633 - 1.232\,x_i + 1.495\,x_i^2$.

b) Se rechaza la hipótesis nula de la prueba de significancia global de los coeficientes de regresión.

c) Se rechaza la hipótesis nula H_0: $\beta_2 = 0$ con la suma extra de cuadrados y por medio de la prueba t.

2.

a) $\hat{y} = 141.612 - 0.282\,x_i + 0x_i^2$.

b) Se rechaza la hipótesis nula de la prueba de significancia global de los coeficientes de regresión.

c) Se rechaza la hipótesis nula H_0: $\beta_2 = 0$ con la suma extra de cuadrados y por medio de la prueba t.

3. $\hat{y} = 0.5 + 0.375\,x_i$, $R^2 = 0.75$, el número de clasificaciones correctas es cinco.

4.

a) $\hat{y} = -0.946 + 0.102\,x_i$. Los coeficientes de regresión son significativos con un nivel de significancia del 5%.

5.

a) $\hat{y} = -126.505 + 0.176\,x_{1i} - 1.563\,x_{2i} + 1.575\,x_{3i} + 1.629\,x_{4i}$.

b) $R^2 = 0.842$, $\bar{R}^2 = 0.779$.

c) Se rechaza la hipótesis nula de los coeficientes de regresión individual β_1, β_3 y β_4. y se acepta para β_2.

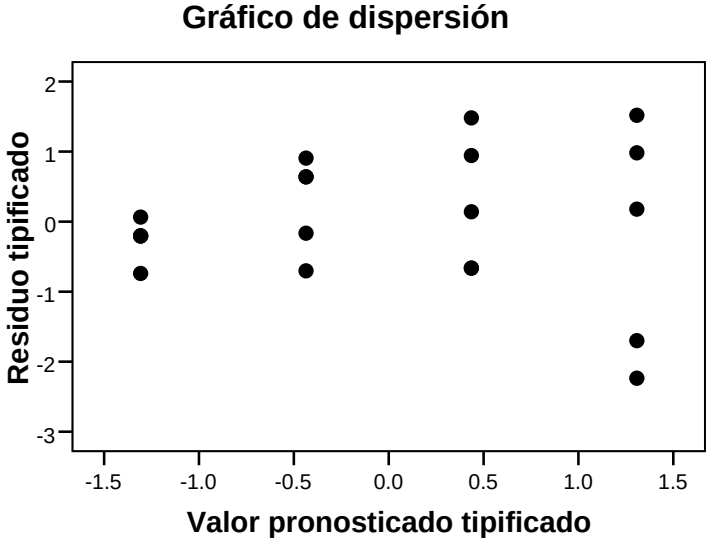
6. Este caso no se pueden estimar los coeficientes de regresión porque existe perfecta colinealidad ya que la variable x_3 puede ser formada como una combinación lineal de la variable x_2 en la forma $x_3 = 2x_2 - 1$.

7.

a) $\hat{y} = 0.89 + 0.237\, x_i$.

b) $R^2 = 0.0.93$, $t(\hat{\beta}_0) = 4.356$, $t(\hat{\beta}_1) = 15.897$ y F = 252.722.

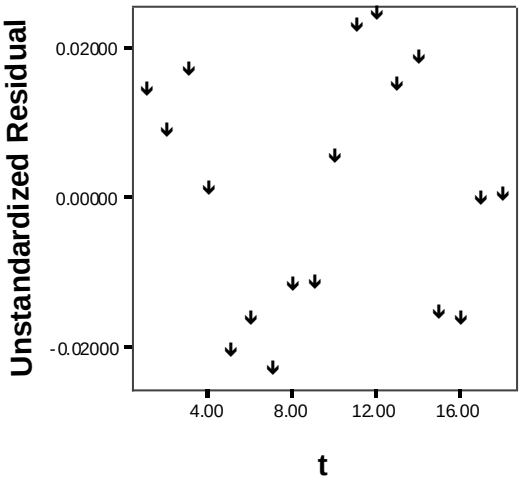
c) En la figura siguiente se presenta el diagrama de dispersión de los residuos en el que se puede observar que existe heteroscedasticidad entre los residuos.



8.

a) $\hat{y} = -1.165 + 0.294\, x_i$.

b) El siguiente diagrama de dispersión de los residuos contra el tiempo muestra que hay autocorrelación entre los residuos pero en menor grado.



c) $d = 0.736$.

d) Dado que d = 0.736 es menor que d_L = 1.16. Se rechaza la hipótesis nula.

e) $\hat{\rho} = \dfrac{n^2(1-d/2)+k^2}{n^2-k^2} = \dfrac{(18)^2(1-0.736/2)+(4)^2}{(18^2)-(2)^2} = \dfrac{220.768}{320} = 0.6899$

Capítulo 8.

1. Método de selección hacia adelante: $\hat{y} = -6.336 + 0.337\,x_1$.

Método de eliminación hacia atrás: $\hat{y} = -6.336 + 0.337\,x_1$.

Método de regresión paso a paso: $\hat{y} = -6.336 + 0.337\,x_1$.

2. Método de selección hacia adelante: $\hat{y} = 13.321 + 3.324\,x_1$.

Método de eliminación hacia atrás: $\hat{y} = 10.044 + 2.713\,x_1 + 6.163\,x_2$.

Método de regresión paso a paso: $\hat{y} = 13.321 + 3.324\,x_1$.

BIBLIOGRAFIA.

1. PEÑA SANCHEZ, DANIEL. ESTADISTICA: *Modelos y Métodos*. Tomo 2. 1987. Alianza Editorial.
2. DRAPER, N.R. Y SMITH, H. *APPLIEDREGRESION ANALISIS*. 1966. John Wiley & Sons.
3. GUJARATI, DAMODAR. *ECONOMETRIA*. 1992. MC GRAWHILL.
4. MONTGOMERY, D.C, PECH_E. Y G.G. Vining. *Introducción al Análisis de Regresión*. 2002. CECSA.
5. GALLASTEGUI FERNÁNDEZ, ALONSO. *ECONOMETRIA*. Madrid 2005. PEARSON prentice hall.
6. LORIA DIAZ DE GUZMAN, EDUARDO G. *ECONOMETRIA CON APLICACIONES*. MEXICO 2007. PEARSON prentice hall.
7. GARDNER ROBERT C. *Estadística para psicología usando SPSS para Windows*. Primera edición, 2003. . PEARSON prentice hall.
8. PINDYCK, ROBERT S., Rubinfeld Daniel L. *Econometria Modelos y Pronósticos*. Cuarta edición MC GRAWHILL.
9. BONILLA, GILDABERTO. ESTADÍSTICA II: *Métodos Prácticos de Inferencia Estadística*. Segunda Edición, San Salvador El Salvador, 1992. Editorial UCA Editores.
10. MYERS. WALPOLE. *Probabilidad y Estadística*. Cuarta Edición, México 1992. Editorial MC GRAWHILL.

11. http://tarwi.lamolina.edu.pe/~arrubio/pag06.htm.
12. http://supervisadaextraccionrecuperacioninformacion.iespana.es/modelos-lineales.html.
13. http://www.virtual.unal.edu.co/cursos/ciencias/2007315/lecciones_html/capitulo_8/leccion0/introduccion.html