UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

“ANALISIS DE LOS FACTORES QUE INFLUYEN EN EL RENDIMIENTO DE LA CAÑA DE AZUCAR”

TRABAJO DE GRADUACIÓN PRESENTADO POR:

JOSÉ DAVID ESCOBAR MUÑOZ

PARA OPTAR AL GRADO DE:

MAESTRO EN ESTADÍSTICA

CIUDAD UNIVERSITARIA, ENERO DE 2016

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

“ANALISIS DE LOS FACTORES QUE INFLUYEN EN EL RENDIMIENTO DE LA CAÑA DE AZUCAR”

TRABAJO DE GRADUACIÓN PRESENTADO POR:

JOSÉ DAVID ESCOBAR MUÑOZ

ASESORES:

DR. JOSÉ NERYS FUNES TORRES

DR. DOMINGO MORALES

CIUDAD UNIVERSITARIA, ENERO DE 2016

AUTORIDADES

RECTOR INTERINO:
LIC. JOSÉ LUIS ARGUETA ANTILLÓN

SECRETARIA GENERAL:
DRA. ANA LETICIA ZAVALETA DE AMAYA

FISCAL GENERAL INTERINO:
LICDA. NORA BEATRÍZ MELÉNDEZ

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

DECANO:
LIC. MAURICIO HERNÁN LOVO CÓRDOVA

SECRETARIO:
LIC. CARLOS ANTONIO QUINTANILLA APARICIO

ESCUELA DE MATEMÁTICA

DIRECTOR:
DR. JOSÉ NERYS FUNES TORRES

SECRETARIA:
MSC. ALBA IDALIA CÓRDOVA CUÉLLAR

CIUDAD UNIVERSITARIA, ENERO DE 2016

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

ASESOR INTERNO

DR. JOSÉ NERYS FUNES TORRES

UNIVERSIDAD DE EL SALVADOR

ASESOR EXTERNO

DR. DOMINGO MORALES

UNIVERSIDAD MIGUEL HERNÁNDEZ DE ELCHE

CIUDAD UNIVERSITARIA, ENERO DE 2016

A mis padres, mi hermana,

familiares, buenos docentes y amigos.

Índice general

Índice de tablas

Índice de figuras

Agradecimientos

Agradezco a Dios, a mis padres, a mis asesores (Dr. Nerys Funes y Dr. Domingo Morales), a los buenos docentes de la Facultad de Ciencias Naturales y Matemática, especialmente a los de la Esculea de Matemática, por su valiosa ayuda en mi formación profesional y especialmente a mis amigos: Dra. Begoña Vitoriano, Dr. Francisco Javier Martín Campo, MSc. Ricardo Ríos, y a mis compañeros y amigos de trabajo por todo su apoyo.

Gracias!!

Introducción

La agroindustria azucarera es actualmente una de las principales actividades económicas del país, ya que, es una fuente de empleo para muchas familias salvadoreñas y el azúcar es uno de los principales productos de exportación. La producción del azúcar o también conocida como sacarosa, se lleva a cabo en ingenios o centrales azucareras, en los cuales la materia prima es la caña de azúcar. La caña de azúcar constituye el cultivo sacarífero más importante del mundo, responsable del 70 % de la producción total de azúcar. En este estudio se presenta la descripción de algunos factores agrícolas y ambientales que afectan al rendimiento de la caña de azúcar, los cuales serán utilizados para poder encontrar relaciones que permitan mejorar el rendimiento.

En este trabajo se analizará el rendimiento de la caña de azúcar en función de los factores con información indispensable que influyen en él, para poder plantear un modelo con su respectiva interpretación, que describa este problema y obtener una explicación estadística, que ayude a la toma de decisiones que contribuyan de manera significativa a mejorar la producción.

En el Capítulo 1, se presentan detalles técnicos de la caña de azúcar y sus componentes básicos, para poder familiarizarnos con ésta y tener una mejor idea de su importancia en la actualidad, y así entender sus necesidades para su desarrollo adecuado y poder aprovechar al máximo su rendimiento. Además, se describen algunos factores agrícolas y ambientales que influyen en el rendimiento de la caña, tanto a nivel de campo como industrial, para tener una

mejor comprensión de cada uno de estos factores, así como algunos componentes básicos de la caña y su composición física. En el capítulo 2 se desarrolla la teoría de los métodos estadísticos usados para modelar la relación entre una variable de interés Y, llamada variable respuesta o variable dependiente, y una o más variables $X_1, \ldots, X_p$, llamadas variables explicativas o independientes para predecir el comportamiento de la variable Y. En este caso la variable Y es el rendimiento y las variables explicativas son los factores agrícolas y ambientales. Se trata de modelos de regresión lineal, que son adecuados cuando todas las variables explicativas son numéricas y de los modelos ANCOVA, que son utilizados cuando se tienen variables numéricas y variables categóricas o cualitativas, siendo estos últimos los indicados para el modelado del rendimiento de la caña, (ya que se tienen variables numéricas como la edad en meses de la caña y variables categóricas o cualitativas como la variedad de caña cosechada). En el Capítulo 3 se desarrolla el modelado para los rendimientos de la caña, se describen las variables para su respectiva comprensión, se identifican, estiman y evalúan posibles modelos. Se seleccionarán los mejores modelos en base a criterios de selección para poder obtener resultados confiables e interpretables que ayuden a una buena toma de decisiones que contribuyan significativamente a obtener mejores rendimientos.

Objetivos

General

- Analizar y modelar el rendimiento de caña en campo y en la producción de azúcar e identificar los factores agrícolas y ambientales que tienen una mayor influencia en la producción.

Específicos

- Analizar el efecto de las variables agrícolas y ambientales en el rendimiento de la caña de azúcar usando modelos ANCOVA.
- Modelar los rendimientos de la caña de azúcar en función de las variables agrícolas y ambientales.
- Estimar intervalos con un 95 % de confianza para los rendimientos medios de la caña de azúcar.
- Recomendar acciones que permitan obtener un incremento significativo en el rendimiento y en el rendimiento industrial.

Antecedentes y justificación

Antecedentes

La caña de azúcar es originaria de Nueva Guinea, de donde se distribuyó a toda Asia. Los árabes la trasladaron a Siria, Palestina, Arabia y Egipto, de donde se extendió por África. Cristóbal Colón la llevó a las islas del Caribe y de ahí pasó a América. Actualmente, la caña de azúcar es el cultivo más importante para la producción de azúcar. Además de la producción de azúcar provee subproductos como el etanol para uso energético, bagazo utilizado para la generación de energía eléctrica y otros.

El cultivo de la caña de azúcar está muy difundido en el continente americano debido a las condiciones climáticas, las cuales propician su producción. La mayoría de países en Latinoamérica cultiva la caña de azúcar para la producción de azúcar. Los principales productores son Brasil, México y Colombia.

Actualmente se han realizado muchos estudios sobre el manejo agronómico, la fertilización y el rendimiento de este cultivo. Algunos ejemplos son:

- Luna González, C. A; Cock, J. H; Palma, A. E, Díaz, L. V y Moreno, C. A. Análisis de la Productividad en la Agroindustria Azucarera de Colombia y Perspectivas para aumentarla (1995).
- Quintero Durán, Rafael. Fertilización y Nutrición de la caña de azúcar (1995).
- Suárez García, Luis Fernando. Manejo agronómico del cultivo

de la caña de azúcar (2012).

- Zossi, Silvia; Cárdenas, Gerónimo; Sorol, Natalia y Sastre Marcos , Influencia de compuestos azúcares y no azúcares en la calidad industrial de caña de azúcar en Tucumán, R. Argentina: caña verde y quemada, Parte 1 y 2 (2010 - 2011).

En estos trabajos se han analizado las necesidades de la caña en base a diferentes componentes y la influencia de estos, así como el manejo o manipulación adecuados para la caña, para no dañar su calidad y garantizar que se está trabajando con materia prima adecuada, que permita obtener un producto final que cumpla con los requisitos y necesidades de los clientes y poder incursionar en nuevos mercados, y obtener mejores ingresos económicos.

Justificación

La importancia del cultivo de la caña de azúcar se refleja en su presencia mundial. Actualmente para el área centroamericana es el rubro agroindustrial más estable debido al colapso de la producción de café. Igualmente para el resto de América es un cultivo de suma importancia, siendo reflejado en la generación de empleos directos e indirectos en la industria.

A pesar de la importancia que tiene la caña de azúcar, a nivel centroamericano no existen estudios que permitan estimar o diagnosticar objetivamente el rendimiento cañero, mucho menos modelos probados y validados que se puedan replicar en sectores con similares características agronómicas o climáticas.

Los diferentes factores climáticos y agrícolas que actúan sobre un lugar determinado condicionan en gran medida las fases del ciclo productivo de la caña y los resultados finales de ésta. A cada lugar corresponde un rendimiento máximo dependiente de las condiciones agrícolas y climáticas del año. A la media de esas condiciones climáticas corresponde una media de rendimiento máximo, o rendimiento potencial específico.

Debido a esto, nace la necesidad de realizar un estudio que involucre tanto factores agrícolas como climáticos para analizar el impacto que estos tienen en el rendimiento de la caña de azúcar, ya que la composición de la caña de azúcar depende de un gran número de factores, entre ellos están su edad, su tolerancia a enfermedades, condiciones de cultivo y el uso de madurantes.

Planteamiento del problema

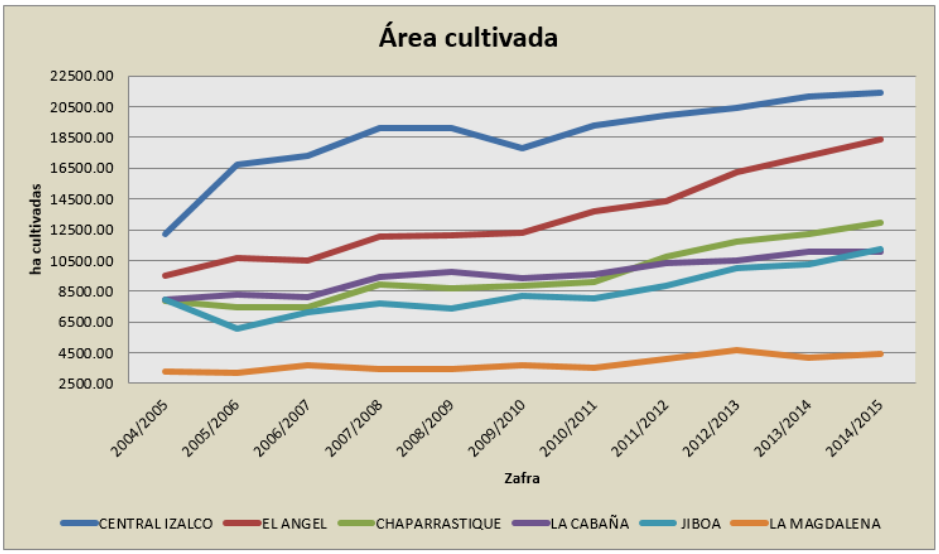
La producción de la caña de azúcar se ha vuelto una necesidad que ha venido acompañada del rápido crecimiento de la población humana que trae consigo una mayor demanda de recursos. Esta producción varía significativamente de un área a otra, dependiendo de la variedad, factores climáticos, disponibilidad de agua, prácticas agrícolas y la duración del periodo de crecimiento.

La cosecha de caña y producción de azúcar, se realiza en un periodo de producción específico al que se le llama zafra. En El Salvador está comprendido entre los meses de noviembre a abril, periodo en el cual la caña se encuentra en condiciones adecuadas para su cosecha (maduración). El principal interés al cultivar caña de azúcar se centra en obtener un buen rendimiento de campo y un buen rendimiento industrial, los cuales se definen a continuación.

- Rendimiento de Campo o de caña: Toneladas de caña producidas por hectárea cultivada (t/ha).
- Rendimiento Industrial o de azúcar: Kilogramos de azúcar producidos por tonelada de caña cosechada (kg/t).

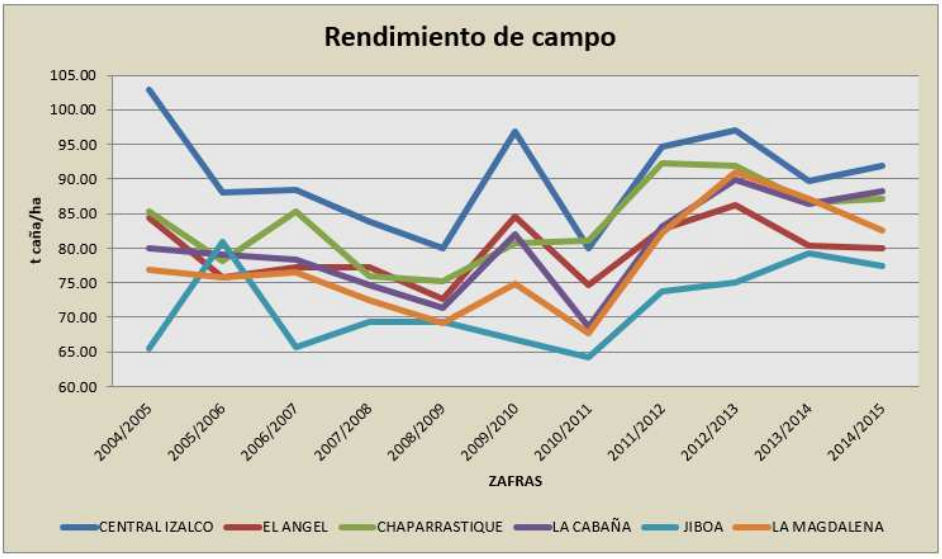
En El Salvador existen 6 ingenios azucareros, en los cuales la superficie utilizada para la siembra de caña de azúcar se ha incrementado de manera significativa en las últimas diez zafras, mostrando una clara tendencia a incrementarse año tras año, (ver Figura1).

En la Figura 2 se muestra el rendimiento agrícola de la caña de azúcar en las últimas 11 zafras



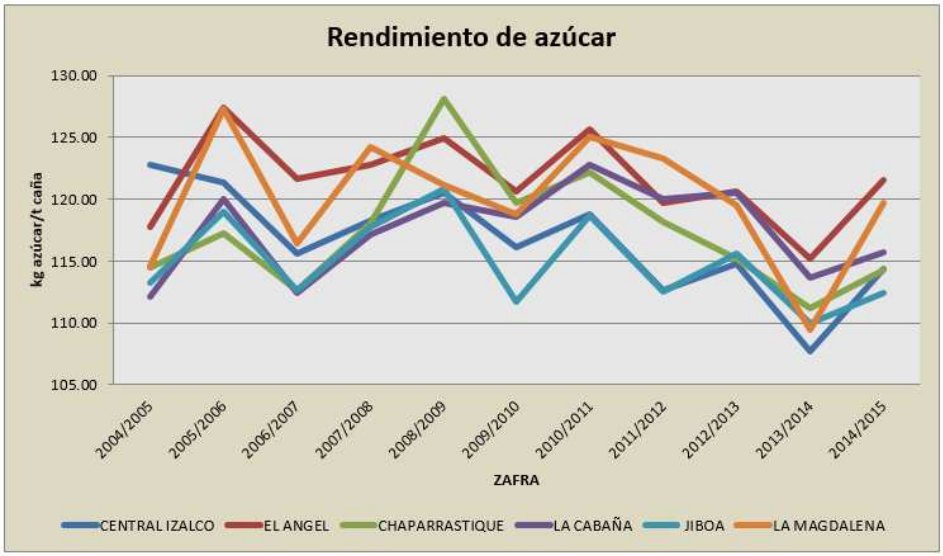
Fuente: Consejo Salvadoreño de la Agroindustria Azucarera. CONSAA

Figura 1: Superficie cultivada por ingenio



Fuente: Consejo Salvadoreño de la Agroindustria Azucarera. CONSAA

Figura 2: Rendimiento de campo por ingenio



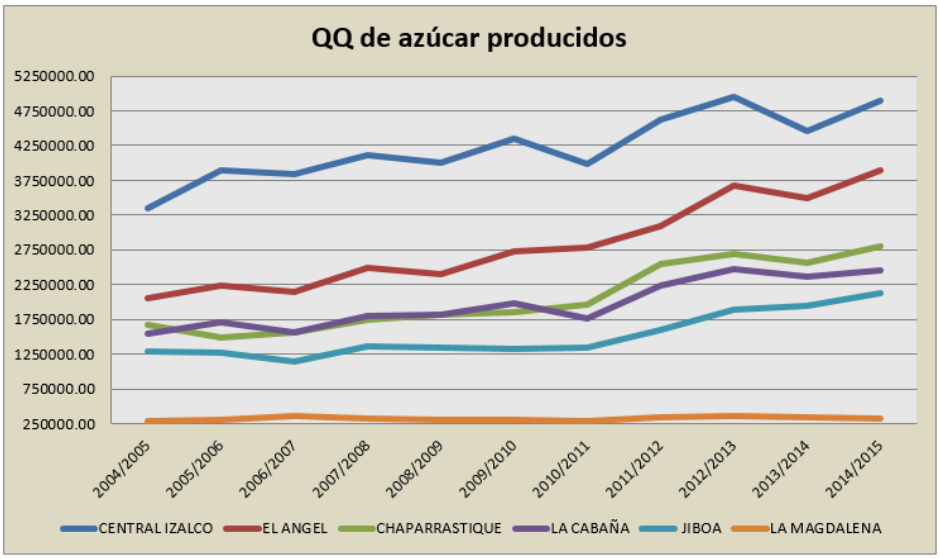
Fuente: Consejo Salvadoreño de la Agroindustria Azucarera. CONSAA

Figura 3: Rendimiento industrial por ingenio

En la zafra 2014-2015 se obtuvo un promedio general de 116.33 kilogramos de azúcar por tonelada de caña, siendo un rendimiento un poco bajo con respecto a otras zafras con apenas 5.12 kg más que la zafra 2013-2014.

Según informes presentados por cada uno de los ingenios, los mayores rendimientos nacionales se alcanzaron en la zafra 2008-2009, cuando el ingenio Chaparrastique tuvo un rendimiento de 128.57 kg/t, superando el promedio nacional que fue de 122.53 kg/t (ver Figura 3).

La Figura 4 muestra la producción de azúcar de las últimas 11 Zafras para cada ingenio, donde se puede notar que año con año, la producción ha aumentado en cada uno de los ingenios. Esto se debe a que se ha estado cultivando más.



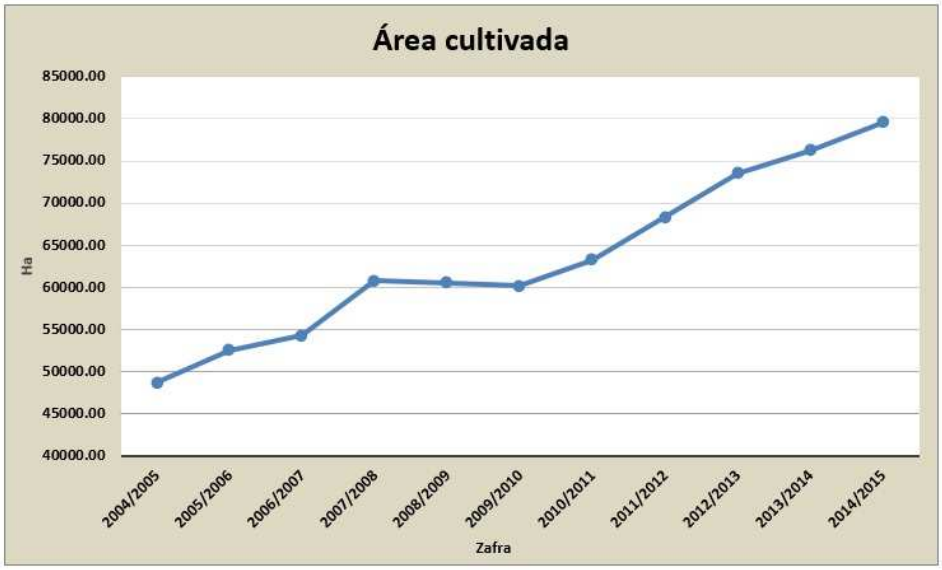
Fuente: Consejo Salvadoreño de la Agroindustria Azucarera. CONSAA

Figura 4: Producción de azúcar por ingenio

A diferencia del comportamiento de la cantidad de la superficie de caña de azúcar cultivada, los rendimientos no muestran un incremento significativo. Por esta razón se hace necesario identificar los factores que contribuyan a un incremento de ambos rendimientos.

En las Figuras 5, 6, 7, 8 se muestra los resultados obtenidos a nivel nacional para las últimas 11 Zafras.

Necesidad: Identificar las variables que maximicen los rendimientos tanto de campo (t/ha), como el industrial (kg/t) y de esta forma establecer un plan de acción para encontrar la manera de obtener mejores rendimientos y así contribuir con el desarrollo económico del país.



Fuente: Consejo Salvadoreño de la Agroindustria Azucarera. CONSAA

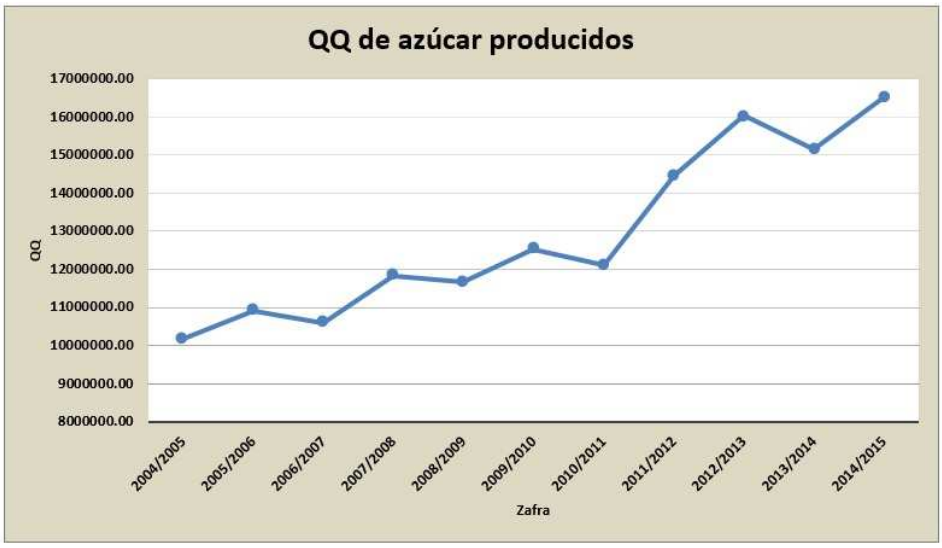
Figura 5: Área total cultivada

Fuente: Consejo Salvadoreño de la Agroindustria Azucarera. CONSAA

Figura 6: Rendimiento total de caña

Fuente: Consejo Salvadoreño de la Agroindustria Azucarera. CONSAA

Figura 7: Rendimiento total de azúcar



Fuente: Consejo Salvadoreño de la Agroindustria Azucarera. CONSAA

Figura 8: Producción total de azúcar

Capítulo 1

La caña de azúcar

1.1. El cultivo de la caña de azúcar.

Nombre científico: Saccharum officinarum.
Nombres comunes: Caña de azúcar, caña miel, caña dulce (en español), sugar cane, noble cane, white salt (en inglés).

La caña de azúcar es una planta herbácea de gran tamaño que se cultiva en países tropicales y subtropicales. Es un híbrido complejo de varias especies, derivadas principalmente del *Saccharum officinarum* y otras especies de Saccharum. La caña se propaga vegetativamente sembrando trozos de sus tallos. La nueva planta de retoño crece a partir de las yemas contenidas en los nudos del tallo, asegurando así una descendencia uniforme. En el proceso de preparación de la caña se desarrollan y ensayan nuevas variedades en búsqueda de nuevas y mejores plantas. Este procedimiento se ha constituido en un factor fundamental para la mejora de la productividad en la industria de la caña de azúcar.

La duración del cultivo varía entre ocho meses en Luisiana y

cerca de dos años en Hawaii. La caña producida puede estar entre 50 t/ha bajo condiciones desfavorables y cifras próximas a 200 t/ha bajo condiciones excepcionales con largos periodos de crecimiento. La producción del azúcar varía de 5 a 25 t/ha [11].

Generalmente no se requiere volver a sembrar caña luego de cada cosecha, sino que se deja crecer de nuevo para producir la siguiente cosecha, denominada soca o rebrote. La producción de caña se reduce después de varias socas, llegando a un punto en que se debe arar la tierra y sembrar caña nuevamente. Este proceso se conoce como renovación. En El Salvador la caña se cosecha en el periodo de noviembre a abril, principalmente por condiciones meteorológicas como la lluvia.

El principal objetivo al procesar la caña es recobrar el azúcar, que en su estado puro se conoce con el nombre químico de sacarosa. Esta se forma en la planta a través de un proceso complejo que esencialmente consiste en la combinación de dos azúcares: fructosa y glucosa.

La caña de azúcar es esencialmente una combinación de jugo y fibra. El jugo es una solución acuosa de sacarosa y otras sustancias orgánicas e inorgánicas. La fibra se define como todo material insoluble en la caña y, por tanto, incluye cualquier suciedad, suelo o cualquier tipo de materia extraña.

El análisis más básico de la caña considera que ésta consiste en agua, sólidos disueltos o sustancia seca refractométrica (RSD) y fibra. El RSD se mide generalmente empleando un refractómetro y a menudo se designa simplemente como Brix. La agroindustria

azucarera es de suma importancia para la economía salvadoreña. A pesar de la crisis económica presentada en los últimos años en nuestro país, la caña de azúcar ha sido una fuente importante de empleo directo e indirecto en las diferentes regiones cañeras del país, ya que se necesita mano de obra para la ejecución de la cosecha, transporte y siembra. Influye a su vez en las actividades propias del sector terciario (servicios), ya que proporciona ingresos a la población que toma parte en la economía de esas regiones agroindustriales durante la duración de la zafra.

La caña de azúcar, más que un cultivo y una actividad empresarial, ha representado toda una cultura para los más de 130 países productores. En virtud de que su presencia ha sido muy amplia e intensa desde el siglo XVI y ha acompañado los procesos de colonización y desarrollo de numerosos países, y son muchas las formas y manifestaciones a través de las cuales esa planta y sus subproductos han intervenido en el quehacer de los pueblos.

Cada una de las regiones cañeras posee características y condiciones productivas singulares que hacen que el potencial productivo, la expectativa de rendimientos agroindustriales y los costos de producción involucrados varíen significativamente.

Se ha comprobado que los rendimientos máximos de caña de azúcar alcanzan aproximadamente un 65 % del rendimiento esperado, por lo que existe un alto potencial para incrementar la acumulación de sacarosa si los limites bioquímicos y fisiológicos pueden ser identificados y modificados [1].

El fruto agrícola de esta planta o agroindustrialmente útil para

múltiples producciones es el tallo, en el cual se acumula sacarosa en el período de maduración, y que tiene una gran importancia para la producción de azucares. Los dos componentes del rendimiento de caña son la cantidad de sacarosa y la producción de biomasa; incrementar uno o ambos eleva el rendimiento.

La composición de la caña de azúcar depende de un gran número de factores, incluyendo su edad, su tolerancia a enfermedades, las condiciones de cultivo y el uso o no de madurantes [11].

1.2. Principales componentes de la caña de azúcar

1.2.1. Raíces.

Las raíces tienen la función de absorber las sustancias nutritivas del suelo y al mismo tiempo sirven de sostén a la planta. Las raíces poseen un sistema radicular fasciculado o fibroso. Al plantar una estaca de caña se desarrollan dos clases de raíces: las transitorias y las permanentes.

Las raíces transitorias, primarias o temporales son aquellas que nacen de la estaca o trozo plantado desarrollándose a partir de la semilla agrícola en los puntos del anillo radicular del nudo, son finas y muy ramificadas de color blancuzco o amarillento. Su función consiste en tomar las sustancias nutritivas del suelo necesarias para la primera etapa de la vida del retoño. Eventualmente estas raíces desaparecen para ser sustituidas por las permanentes.

Las raíces permanentes, secundarias o definitivas son aquellas que nacen de la cepa que se forma desde los primeros momentos en la base del nuevo retoño. Son gruesas, largas y de color blanco y tienen la función de absorber agua y demás nutrientes del suelo necesarios para el desarrollo de la planta, sin olvidar que le sirven de anclaje.

1.2.2. Tallo.

El tallo es la parte más importante de la caña, al constituir el fruto agrícola de la misma: en él se almacena el azúcar. Está formado por unidades conocidas como canutos (entrenudos), que varían en longitud, grosor, forma y color según su variedad. Los canutos están unidos por los nudos, en donde se insertan las hojas. En los nudos encontramos entre otros: el anillo de crecimiento, la banda o anillo de raíces, la cicatriz foliar y la yema, en el cual hay un pequeño hundimiento llamado canal o surco de la yema (véase Figura 1.1). El anillo de crecimiento es donde se produce el alargamiento de los canutos, es decir, donde tiene lugar el crecimiento del tallo. Por lo delicado de los tejidos que lo forman es donde se produce la ruptura o quiebra del tallo por acción de los vientos u otras causas.

El anillo de raíces es la zona donde pueden observarse varias filas circulares de puntos redondos y blanquecinos, de donde brotan las raíces transitorias al tener la caña la humedad necesaria para el brote y desarrollo. Su forma, color y tamaño de los puntos y número de filas es característico de cada variedad.

La cicatriz foliar es el punto que une la hoja al tallo. Después de la caída de la hoja una parte de la base permanece fija al tallo.

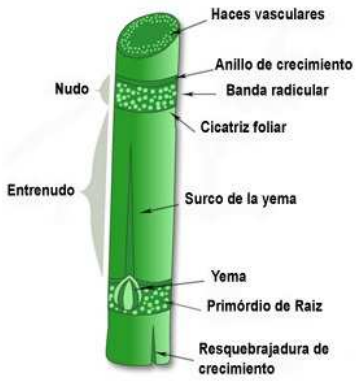


Figura 1.1: Tallo de la caña

La yema es una planta en miniatura cuyo tallo no se ha prolongado. Las hojas son diminutas y muy unidas y las exteriores adquieren forma de escama. En la caña las yemas se encuentran situadas en un nudo en posición opuesta a la anterior (intercalada). Puede haber una o más yemas en el nudo.

Al presentarse las condiciones de humedad y temperatura óptimas, la yema se desarrolla originando una nueva planta. Ésta puede adoptar diversas formas tales como: triangular, ovalada, pentagonal, romboide, redonda, oval, rectangular o picuela, puede tener o no pelos en la base y su valor germinativo disminuye del cogollo hacia la base, elemento a considerar en el momento de seleccionar la semilla para la siembra.

El canal de la yema es una depresión o hundimiento en el canuto, que comienza en la yema y termina en la mitad o más arriba del mismo, según la variedad.

El color y la forma del canuto depende de la variedad y de las condiciones del medio donde se desarrolla la planta de caña. Por

ejemplo, los rayos del sol pueden cambiar el color de la planta.

Generalmente el color es completo, pero algunas variedades presentan colores segmentados, los que reciben el nombre de cañas cintas, siendo los colores más comunes: rojo, verde morado y amarillo y las formas más comunes: cilíndrica, tonel, carrete, cónico y curvada o zigzag.

El grosor y el largo del canuto son característicos y de gran importancia en la selección de las variedades a utilizar en la siembra, por lo que representarán en términos de rendimientos.

Las características internas del tallo se pueden observar si damos un corte transversal al mismo, caracterizando los tallos de las plantas monocotiledóneas con la médula al centro, formada por un tejido esponjoso, que contiene jugos ricos en azúcar, atravesado por vasos cribosos que van unidos mientras más se acercan al exterior.

1.2.3. Hoja.

La hoja brota del nudo del tallo, son lanceoladas, lineales, largas y agudas, tienen un nervio o vena central fuerte, dispuesta en el tallo de forma alterna, de color verde y cambia su tonalidad de acuerdo a la variedad y al medio en que se desarrolla. Su borde es dentado y se distinguen en la hoja tres partes fundamentales: la vaina, la lígula y el limbo.

La vaina es la parte de la hoja que abraza el tallo, cubriendo por entero el canuto del cual nace. A medida que la planta va madurando la vaina se separa del canuto y termina por desprenderse. En algunas

Figura 1.2: Hojas de la caña

variedades la vaina se encuentra cubierta de pelos caedizos.

La lígula es una delgada membrana en forma de lengüeta, que se observa en el punto de unión entre el limbo y la vaina, que tiene como función principal evitar que el agua penetre entre el tallo y la vaina. De suceder lo contrario, la humedad provoca el desarrollo de raicillas en el nudo que provocan un desarrollo anormal de la yema.

El limbo está formado por la lámina de la hoja, encontrándose en el la vena o nervio central y las paralelas a ésta. Generalmente el limbo mide de 1.0 a 1.8 m de longitud por 0.5 a 0.7 m. de ancho, de color verde más o menos intenso dependiendo del medio en que se desarrollan (ver Figura 1.2).

1.2.4. Flores.

La inflorescencia de la caña aparece en forma de panícula, que se desarrolla a partir del último canuto. Puede ser de diversas formas: ancha, estrecha, corta, larga, cónica, cilíndrica, dependiendo de cada variedad. Está constituida por un eje principal al cual se insertan los ejes laterales primarios que a su vez conforman unos ejes secundarios y a su vez terciarios. Esta ramificación está más desarrollada en la base que en el vértice. Las espiguillas están dispuestas por pares en cada articulación, una es sesil y la otra es pedunculada. Está rodeada de largos pelos que hacen ver a la inflorescencia con un aspecto sedoso o afelpado. La flor es bisexual, la semilla de caña es extremadamente pequeña siendo un fruto cariópside. Las condiciones que favorecen la floración son: la duración del día, próxima a 12 horas, temperatura mínima superior a 18°C, humedad suficiente de la planta y perfecto estado vegetativo foliar.

1.3. La variedad de la caña.

Generalmente la selección y multiplicación de las variedades de caña busca una producción elevada de ésta por hectárea y un alto contenido de sacarosa. Sin embargo, es igualmente importante su resistencia o susceptibilidad a enfermedades. Existen algunas diferencias significativas entre variedades, las cuales son generalmente elegidas para satisfacer las condiciones agronómicas por ejemplo: lluvia, maduración temprana, tipo de suelo, duración del periodo de crecimiento y el sistema de cosecha.

Figura 1.3: Flores de la caña

Con el tiempo, las variedades más populares que se cultivan en un área particular sufren cambios. Las diferencias en la composición del jugo entre diferentes variedades no son suficientemente grandes o previsibles como para afectar ajustes en el procesamiento según la variedad. Sin embargo, la dureza de la caña y la disposición de la fibra varían de una variedad de caña a otra, lo que puede afectar su comportamiento en la planta de extracción.

Las diferencias en componentes no-sacarosas son normalmente influenciadas en gran parte por la variedad de la caña.

Hay cientos de variedades en todo el mundo. En España por ejemplo más del 80 % de la superficie plantada es de la variedad NC0310, que procede de África del Sur. El cultivo de cada variedad depende de las condiciones de cada región. Algunas de las variedades cultivadas en El Salvador se presentan a continuación:

B-34-104, C-116-67, CP-64-388, CP-72-1210, CP-72-1312, CP-72-2086, CP-73-1547, CP-80-1557, CP-81-1384, CP-83-1499, CP-84-1198, CP-88-1165, CP-88-1508, CP-89-2143, MEX-69-290, MEX-79-431, MY-5465, PGM-89-118, PGM-89-968, PINDAR, PR-75-2002, PR-83-1172, PR-87-2080, SP-79-1011, VARIAS. Más del 40 % del área cosechada es de la variedad CP-72-2086, y cerca del 22 % es de la variedad MEX-79-431, siendo estas las dos variedades más usadas.

1.4. Etapas del cultivo.

1.4.1. Establecimiento (germinación 30 - 50 días).

La germinación se refiere a la iniciación del crecimiento a partir de las yemas presentes en los tallos plantados o en los que quedan en pie después de la cosecha del cultivo anterior. Durante esta fase es necesaria la disponibilidad adecuada de agua y el control de malezas. El déficit hídrico tiene un impacto significativo sobre el rendimiento de azúcar ya que propicia la reducción de la densidad de población de adultos debido al nuevo e insuficiente sistema de raíces pequeñas y poco profundas [1].

La germinación de las yemas es influenciada por factores externos e internos. Los factores externos son la humedad, la temperatura y la aireación del suelo. Los factores internos son la sanidad de la yema, la humedad y el contenido de azúcar reductor del esqueje y su estado nutricional. La germinación produce una mayor respiración y, por ello, es importante tener una buena aireación del

suelo. Por esta razón, los suelos abiertos, bien estructurados y porosos permiten una mejor germinación.

1.4.2. Crecimiento vegetativo, amacollamiento o ahijamiento, elongación del tallo y cierre de la plantación (50 -70 días).

El crecimiento y el rendimiento son muy sensibles a cualquier déficit de agua en esta etapa exigente. Además la planta amacolla y desarrolla mayor cantidad de follaje y la plantación comienza a cerrar. Es necesario aplicar fertilizante, para que las plantas puedan desarrollarse satisfactoriamente en la siguiente fase. La elongación del tallo es inicialmente rápida y, durante esta fase, el contenido de fibra del tallo es elevado, mientras que los niveles de sacarosa son todavía bastante bajos. Una temperatura cercana a 30^{o}C es considerada óptima para el ahijamiento [1].

El ahijamiento es el proceso fisiológico de ramificación subterránea múltiple, que se origina a partir de las articulaciones nodales compactas del tallo primario. El ahijamiento le da al cultivo un número adecuado de hojas activas y tallos, que permiten obtener un buen rendimiento. Diversos factores, tales como la variedad, la luz, la temperatura, el riego (humedad del suelo) y las prácticas de fertilización afectan al ahijamiento. La incidencia de una iluminación adecuada en la base de la planta de caña durante el período de ahijamiento es de vital importancia. Los hijuelos o retoños que se forman primero dan origen a tallos más gruesos y pesados. Los retoños formados más tarde mueren o se quedan cortos o inmaduros. Manejos culturales co-

mo el espaciamiento, la fertilización, la disponibilidad de agua y el control de las arvenses afectan al ahijamiento.

1.4.3. Crecimiento rápido e incremento del rendimiento (180 - 220 días).

Comprende desde el cierre del dosel hasta el inicio del periodo de madurez de los tallos. Se caracteriza por el aumento de biomasa y del número de tallos por área. La humedad es fundamental para que el sistema radical se desarrolle y pueda absorber los nutrientes. Cualquier déficit de agua comenzaría el proceso de maduración y detendría la acumulación de sacarosa antes de su etapa óptima.

Durante la primera etapa de esta fase ocurre la estabilización de los retoños. De todos los retoños formados sólo el 40 - 50 % sobrevive y llega a formar cañas triturables. Esta es la fase más importante del cultivo, en la que se determinan la formación y elongación real de la caña y su rendimiento. En esta fase ocurre un crecimiento rápido de los tallos con la formación de 4-5 nudos por mes. Así como una foliación frecuente y rápida hasta alcanzar un Índice de Área Foliar (IAF) de 6-7.

1.4.4. Maduración y sazonado (60 - 140 días).

Se inicia alrededor de dos a tres meses antes de la cosecha para cultivos con ciclo de 12 meses, y de los 12 a los 16 meses de edad para los que completan el ciclo en 18 a 24 meses. En esta fase se requiere un bajo contenido de humedad del suelo, por lo que el riego debe

ser reducido y luego detenerse para llevar la caña a la madurez. Así se detiene el crecimiento y se propicia la acumulación de carbohidratos y la conversión de azúcares reductores (glucosa y fructosa) a sacarosa. La maduración de la caña ocurre desde la base hacia el ápice y por esta razón la parte basal contiene más azúcares que la parte superior de la planta [1].

Condiciones de abundante luminosidad, cielos claros, noches frescas y días calurosos (es decir, con mayor variación diaria de temperatura) y climas secos son altamente estimulantes para la maduración. La consecuencia práctica del conocimiento de estas etapas permite al productor una mejor comprensión de lo que ocurre con la planta y ayuda a un manejo eficiente del agua y los nutrientes. El control parcial del crecimiento vegetativo y la manipulación de la producción de azúcar es factible. El conocimiento de las fases fenológicas de la planta es esencial para maximizar los rendimientos de caña y la recuperación del azúcar.

1.4.5. Cosecha.

Los factores que afectan el sazonado de la planta de caña de azúcar son la edad, el contenido de nitrógeno del suelo y la humedad. Los factores ambientales pueden influir en la acumulación de sacarosa, incluido el estrés hídrico, los nutrientes y la temperatura. Por regla general, la caña de azúcar es cosechada mediante un corte en la base del tallo, el cual se hace de forma manual o mecánica. La paja se elimina manualmente o es quemada previamente a la cosecha. Ésta ocurre antes de la floración (12 a 18 meses después de la

siembra), debido a que la antesis o floración, conduce a la reducción en el contenido de azúcar en los tallos. Estas etapas se traslapan cíclicamente entre los ciclos planta, soca y resocas y determinan el calendario de los periodos de zafra y no zafra azucarera y las actividades de campo. Se esperan mayores producciones de la caña plantada y un decrecimiento a medida que la edad aumenta [1].

1.5. Necesidades climáticas y agrícolas de la caña de azúcar.

La caña es frecuente en los climas tropicales y pueden producirse hasta los 35° de latitud norte y sur. Se desempeña mejor en altitudes que van desde 0 a 1,000 metros sobre el nivel del mar, aunque los requerimientos obtenibles hasta 1,500 metros son económicamente aceptables. Se desempeña bien con una temperatura media de 24 °C además de una precipitación anual de 1,500 mm bien distribuidos durante su ciclo de crecimiento. Cuando las temperaturas de la noche y del día son uniformes, una caña no cesa de crecer y en sus tejidos siempre habrá un alto porcentaje de azúcares reductores. Las variaciones de temperatura superiores a 8 °C son muy importantes en la fase de maduración, porque ayudan a formar y a retener la sacarosa. Este cultivo se desempeña bien en suelos profundos y fértiles. Si se cuenta con un sistema de riego podremos lograr mejores rendimientos que en suelos sin regar. Puede producirse también en suelos marginales como arenosos y suelos arcillosos con un buen drenaje. No se recomienda para suelos franco-limosos y limosos.

En áreas donde la caña es cultivada utilizando riego y el abastecimiento de agua está asegurado, la composición de caña generalmente varía poco de una temporada a la siguiente. Sin embargo, en áreas dependientes de la lluvia, la producción y la composición pueden ser fuertemente afectadas por variaciones en la precipitación. El efecto sobre la composición de la caña puede apreciarse con las cifras comparativas de temporadas normales y secas o la longitud del entre nudo que es generalmente función de la velocidad de crecimiento. En periodos de sequía, la longitud de los entre nudos se reduce y como resultado el contenido de fibra se incrementa. En casos extremos los cogollos se dejan con la caña para facilitar la formación de pilas donde este método de manejo de caña es utilizado. Esto resulta en un menor contenido de sacarosa, menor pureza del jugo y aumento de color en el jugo. Por el contrario, en áreas donde se presentan fuertes lluvias durante la temporada de zafra, generalmente se observa un incremento en la materia extraña, principalmente en términos del contenido de suelo. Dependiendo de los factores climáticos, puede presentarse un efecto pronunciado del momento de la zafra sobre la composición de la caña. En caso de que la caña esté inmadura al inicio de la zafra, se puede esperar que el contenido de sacarosa y la pureza del jugo sean bajos y que el contenido de azúcares invertidos sea alto. A medida que la zafra transcurre y la caña madura, el contenido de sacarosa y la pureza se incrementan. En este momento se alcanza la mejor recuperación de sacarosa dado que el contenido de no-sacarosa es bajo. Hacia el final de la zafra, a veces se observa un aumento del contenido de fibra. Debido a que la temporada de lluvias es normalmente el factor que determina la duración de la zafra, la lluvia al final de la zafra causa

caídas en la calidad y un incremento en la materia extraña.

Capítulo 2

Métodos estadísticos

"No hay conocimiento que pueda contribuir tanto a mejorar la calidad, productividad y competitividad de las empresas como el de los métodos estadísticos" (Domingo Morales).

2.1. El Modelo de Regresión Lineal.

Son frecuentes en la práctica, situaciones en las que se cuenta con observaciones de diversas variables, y es razonable pensar en una relación entre ellas. El poder determinar si existe esta relación y, en su caso, una forma funcional para la misma es de sumo interés. Por una parte, ello permitiría, conocidos los valores de algunas variables, efectuar predicciones sobre los valores previsibles de otra. Podríamos también responder con criterio estadístico a cuestiones acerca de la relación de una variable sobre otra. Las variables que influyen en una variable dependiente Y, pueden dividirse en dos grupos: El primero contiene a un conjunto de variables X, que son llamadas variables explicativas, exógenas o independientes, el segundo grupo incluye los demás factores que influyen en la variable respuesta y son llamadas

perturbaciones aleatorias.

Es de interés señalar que el ajuste de un modelo de regresión no se limita a analizar la relación entre dos variables; en general, buscaremos relaciones del tipo:

$$Y = f(X_0, X_1, X_2, \cdots, X_k) + U \tag{2.1}$$

Suponemos que en el rango de valores de interés, la función (2.1) admite una aproximación lineal, con lo que resulta el modelo:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + U \tag{2.2}$$

Siendo:

- $\beta_0, \cdots, \beta_k$, parámetros fijos desconocidos.
- $X_0, \cdots, X_k$, variables explicativas no aleatorias e independientes entre sí, llamadas regresores, cuyos valores son fijados por el investigador. Frecuentemente X_0, toma el valor constante "uno".
- U, una variable aleatoria inobservable, llamada perturbaciones. Además suponemos que las perturbaciones deben cumplir las siguientes propiedades:

  a) Su esperanza es cero.

  b) Su varianza es constante, σ^2.

  c) Las perturbaciones son independientes entre sí.

  d) Su distribución es normal.

La ecuación (2.2), indica que la variable aleatoria Y se genera como combinación lineal de las variables explicativas, salvo en una perturbación aleatoria U.

En estos problemas, contamos con una muestra de n observaciones de la variable aleatoria Y , y de los correspondientes valores de las variables explicativas X. Como se ha dicho, U es inobservable. La muestra nos permitirá escribir n igualdades similares a (2.2):

$$\begin{aligned} y_1 &= \beta_0 x_{1,0} + \beta_1 x_{1,1} + \cdots + \beta_k x_{1,k} + u_1 \\ y_2 &= \beta_0 x_{2,0} + \beta_1 x_{2,1} + \cdots + \beta_k x_{2,k} + u_2 \\ &\vdots \\ y_n &= \beta_0 x_{n,0} + \beta_1 x_{n,1} + \cdots + \beta_k x_{n,k} + u_n \end{aligned}$$

En forma matricial, escribiremos las n igualdades así:

$$Y = X\beta + e \tag{2.3}$$

Siendo:

- Y, el vector $n \times 1$ de observaciones de la variable aleatoria Y.
- X, la matriz $n \times k$ de valores de las variables explicativas, x_{ij} denota el valor que la j-ésima variable explicativa toma en la i-ésima observación.
- β, el vector $k \times 1$de parámetros $(\beta_0, \cdots, \beta_k)'$.
- U, el vector $n \times 1$ de valores de la perturbación aleatoria U.

De lo anterior podemos representar cada elemento como sigue:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},$$

$$U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Denotaremos mediante $\hat{\beta}$ al vector de estimadores de los parámetros, y por $\hat{e}$ al vector $n \times 1$ de residuos, definido matricialmente por:

$$\hat{e} = Y - X\hat{\beta} \tag{2.4}$$

Esto significa que los residuos recogen la diferencia entre los valores muestrales observados y los ajustados de la variable aleatoria Y.

Los supuestos respecto a las perturbaciones o errores, pueden escribirse en términos de la variables respuesta como sigue:

a) Para cada conjunto fijo de las X, la distribución de Y tiene media:

$$E[Y] = \beta_0 + \beta_1 X_1 + \cdots , \beta_k X_k$$

b) Conocidos los valores de las X, la varianza de Y es constante, y no depende de los valores de las X:

$$Var[Y] = \sigma^2$$

c) Las y_i son independientes entre sí;

d) Y tiene distribución normal.

2.1.1. Estimación de los parámetros

El problema que abordamos es el de estimar los parámetros desconocidos $\beta_0, \cdots , \beta_k$. Se puede aplicar el método de Mínimos Cuadrados Ordinarios (MCO), para estimar los coeficientes de regresión de la ecuación (2.2).

Se desea determinar el vector $\vec{\beta}$ de estimadores de MCO que minimice la suma de los cuadrados de los residuos.

$$S(\beta) = \sum_{i=1}^{n} e_i^2 = e'e = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \qquad (2.5)$$

Al desarrollar (2.5), obtenemos el siguiente resultado:

$$\begin{aligned} S(\beta) &= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

ya que $\hat{\beta}'X'Y$ es una matriz de 1×1, es decir, un escalar, y que su transpuesta $(\hat{\beta}'X'Y)' = Y'X\hat{\beta}$ es el mismo escalar (ver Apéndice A.1). Los estimadores de mínimos cuadrados deben satisfacer:

$$\frac{\partial S}{\partial \beta} = -2X'Y + 2X'X\hat{\beta} = O \tag{2.6}$$

que se simplifica a

$$X'X\hat{\beta} = X'Y \tag{2.7}$$

La ecuación (2.7) representa las ecuaciones normales de MCO. Para resolver las ecuaciones normales se multiplican ambos lados de (2.7) por la inversa de $X'X$. Así, el estimador de β por MCO es

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2.8}$$

siempre y cuando exista la matriz inversa $(X'X)^{-1}$. La matriz $(X'X)^{-1}$ siempre existe si los regresores son linealmente independientes, esto es, si ninguna columna de la matriz X es una combinación lineal de las demás columnas.

Por tanto los valores estimados por el modelo $\hat{Y}$ que corresponden a los valores observados Y son:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$$

en forma matricial:

$$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y$$

Luego de estimar los parámetros del modelo se hace necesario estimar el peso relativo de cada variable en el modelo, sin importar la unidad

de medida en que se encuentren expresadas. Esta información se obtiene con los coeficientes estandarizados y se calculan multiplicando el coeficiente por la desviación estándar de la variable de interés y dividiendo entre la desviación estándar de la variable dependiente.

2.1.2. Interpretación de los coeficientes

La interpretación de los coeficientes de la ecuación de regresión múltiple, debe realizarse con mucha cautela. Un coeficiente pequeño no es reflejo de una baja correlación entre la variable que lo acompaña y el criterio sino que puede significar que la información compartida entre dicha variable y otra (u otras) predictoras del modelo es muy alta. La recomendación es, teniendo en cuenta que el modelo se elabora con estos condicionantes, que las variables predictoras que explican Y sean lo más independientes entre sí, es decir, no mantengan relación o colinealidad. Sin embargo, la mayoría de las veces la ausencia total de colinealidad entre ellas es imposible de satisfacer y, por tanto hay, que jugar con dicha información compartida entre predictores a la hora de aportar la interpretación última de la ecuación de regresión.

Así el coeficiente de una variable X_i en una regresión multiple, $\hat{\beta}_i$, representa el efecto sobre la respuesta cuando la variable X_i aumenta en una unidad y las demás variables permanecen constantes. Puede interpretarse como el efecto *diferencial* de esta variable cuando eliminamos o controlamos el efecto de las otras variables explicativas. Por tanto, para interpretar el coeficiente de una variable es imprescindible conocer cuáles son el resto de variables ecplicativas

incluidas en la regresión.

El coeficiente de regresión de una variable X_i en regresión múltiple cuando todas las variables explicativas son incorreladas, es el cociente entre la covarianza de la respuesta y la variable, dividido por su varianza.

2.1.3. Propiedades de los estimadores

Los estimadores de los parámetros son, según (2.8), funciones lineales de Y y, recordemo que $Y \sim N(\beta_0 + \beta_1 X_1 + \cdots, \beta_k X_k,\ \sigma^2)$, por tanto los estimadores también se distribuirán normalmente. Vamos a calcular su media y su varianza.

Media de los estimadores

Partiendo de (2.8) y llamando C a la matriz $(X'X)^{-1}X'$, tendremos $\hat{\beta} = CY = C(X\beta + U)$, y como: $CX = (X'X)^{-1}X'X = I$, tenemos que:

$$\hat{\beta} = \beta + CU \tag{2.9}$$

Tomando esperanzas, y recordando que las esperanzas de un vector aleatorio es el vector obtenido tomando esperanzas a los componentes,

$$E[\hat{\beta}] = \beta + CE[U]$$

como, por hipótesis, la perturbación tiene esperanza nula,

$$E[\hat{\beta}] = \beta \tag{2.10}$$

De este resultado se concluye que los estimadores son insesgados o centrados.

Varianza de los estimadores

La matriz de varianzas y covarianzas de los estimadores, $Var(\hat{\beta})$, es una matriz cuadrada de orden $k+1$ y tendrá en la diagonal las varianzas y fuera de la diagonal las covarianzas de las variables. Se define como:

$$Var(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']$$

utilizando (2.9)

$$Var(\hat{\beta}) = E[CUU'C'] = CE[UU']C' = \sigma^2 CC'$$

donde se ha usado que $E[UU'] = \sigma^2 I$, donde I es la matriz unidad. Sustituyendo $C = (X'X)^{-1}X'$, se tiene que (Ver Apéndice A.1):

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

y obtenemos que:

$$Var(\hat{\beta}_i) = \frac{\sigma^2}{VNE_{i,R}}$$

siendo $VNE_{i,R} = \sum_{j=1}^{n}(e_{i.R}^2) = \sum_{j=1}^{n}(x_{ij} - \hat{x}_{ij.R})^2$, la variabilidad diferencial de x_i, que se calcula como la suma de cuadrados de los residuos de una regresión de x_i respecto a las demás variables

explicativas. Por ejemplo $\sum_{j=1}^{n}(e_{2.R}^2)$, son los residuos obtenidos de la regresión $X_2 = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \cdots + \beta_k X_k$. De los resultados anteriores tenemos que:

$$\hat{\beta}_i \sim N(\beta_i, \sigma/\sqrt{VNE_{i.R}}) \tag{2.11}$$

Estimación de la varianza

El modelo quedará especificado al estimar β y la varianza σ^2 de la perturación. Vamos a construir un estimador de σ^2 y estudiar sus propiedades.

Se ha visto que los residuos se calculan usando (2.4), así:

$$\hat{\sigma^2} = \frac{1}{n} ee' = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \tag{2.12}$$

Este es un estimador sesgado para la varianza. Se obtiene un estimador insesgado dividiendo (2.12) por el número de grados de libertad o número de residuos independientes. Se tienen $n - 1 - k$ grados de libertad, ya que existen los $k + 1$ estimadores obtenidos de (2.8). En consecuencia, el estimador centrado o insesgado de la variabilidad será la varianza residual, definida por:

$$\hat{S}_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - k - 1}$$

Así:

$$\frac{(n-k-1)\hat{S}_R^2}{\sigma^2} = \sum_{i=1}^{n}(\frac{e_i}{\sigma})^2 \ \sim \chi^2_{n-k-1}$$

Como la esperanza de una distribición χ^2 es igual al número de grados de libertad $1/\sigma E[ee'] = n - k - 1$, se tiene:

$$E[\hat{S}_R^2] = E\left[\frac{ee'}{n-k-1}\right] = \sigma^2$$

2.1.4. Intervalos de confianza de los coeficientes de regresión

Para la construcción de intervalos de confianza de los coeficientes de regresión $\hat{\beta}$, se continuará suponiendo que los errores e_i son independientes y están distribuidos normalmente, con promedio 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. Así por (2.11), podemos construir un intervalo de confianza para un coeficiente mediante:

$$t_{n-k-1} = \frac{\hat{\beta}_i - \beta_i}{\hat{S}_R\sqrt{VNE_{i.R}}};\ i = 1, \cdots, k \tag{2.13}$$

de donde se contruye el intervalo de confianza al $(1-\alpha)\,\%$:

$$(\hat{\beta}_i \pm t_{n-k-1,\alpha/2}\hat{S}_R\sqrt{VNE_{i.R}}) \tag{2.14}$$

donde $t_{n-k-1,\alpha/2}$ es el valor correspondiente de la distribución $t - Student$ con $n - k - 1$ grados de libertad.

2.1.5. Contrastes de hipótesis

En muchas aplicaciones de la regresión existe una teoría que establece que el efecto de una variable determinada X_i, medido por el valor de un parámetro β_i, debe ser igual a un valor β_i^* prefijado. Esto equivale a contrastar que la variable aleatoria $\hat{\beta}_i$ tiene media β_i^*. El test se realiza calculando el estadístico (2.13) con $\beta_i = \beta_i^*$, que sigue una distribución $t - Student$ con $n - k - 1$ grados de libertad.

Un contraste importante para probar la significancia de cualquier coeficiente individual de regresión es el siguiente:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

Si esto es cierto, es estadístico:

$$t_{n-k-1} = \frac{\hat{\beta}_i}{\hat{S}_R\sqrt{VNE_{i.R}}}$$

seguirá una distribució $t - Student$ con $n - k - 1$ grados de libertad y rechazaremos H_0 si:

$$|t_{n-k-1}| > t_{n-k-1,\alpha/2}$$

2.1.6. Intervalos de confianza para la varianza

El intervalo de confianza para σ^2, se contruye calculando:

$$\frac{(n-k-1)\hat{S_R^2}}{\chi^2_{\alpha/2}} \leqslant \sigma^2 \leqslant \frac{(n-k-1)\hat{S_R^2}}{\chi^2_{1-\alpha/2}}$$

2.1.7. La significancia en la regresión

La prueba de la significancia de la regresión se realiza para determinar si hay una relación lineal entre la respuesta Y y cualquiera de las variables regresoras $X_1, X_2, \cdots, X_k$. Este procedimiento suele considerarse como una prueba general o global de la adecuación del modelo. Las hipótesis pertinentes son:

$$\begin{cases} H_0 & : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 & : \beta_j \neq 0,\ al\ menos\ para\ un\ j = 1,\ \cdots,\ k. \end{cases}$$

El rechazo de la hipótesis nula implica que al menos uno de los regresores $X_1, X_2, \cdots, X_k$, contribuye al modelo en forma significativa. El procedimiento de prueba consiste en una descomposición de la varianza que se describe a continuación.

La variabilidad de la variable respuesta puede descomponerse utilizando:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2 \tag{2.15}$$

que expresa la variabilidad total (VT) como suma de la variabilidad

explicada (VE) y la variabilidad no explicada o residual (VNE).

El contraste de la regresión establece que la VE es significativamente mayor que VNE. Los grados de libertad de VE son k, ya que el vector $\hat{Y}$ se mueve en el muestreo en un espacio de dimensión $k+1$.

El procedimiento se basa en el análisis de varianza (ADEVA o ANOVA), que se presenta en la Tabla 2.1

Fuente	Suma de cuadrados	Grados de libertad	Varianza	Contraste
VE	$\sum(\hat{y}_i - \bar{y})^2$	k	$\hat{S}_e^2$	$F = \frac{\hat{S}_e^2}{\hat{S}_R^2}$
VNE	$\sum(y_i - \hat{y}_i)^2$	$n-k-1$	$\hat{S}_R^2$	
VT	$\sum(y_i - \bar{y})^2$	$n-1$	$\hat{S}_y^2$	

Tabla 2.1: La Tabla ADEVA en regresión

Para probar la hipótesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = O$, se calcula el estadístico de prueba $F - Snedecor$ de la Tabla 2.1 y se rechaza H_0 si $F > F_{k,n-k-1,\alpha}$.

2.1.8. El coeficiente de determinación

Para construir una medida descriptiva del ajuste global del modelo se utiliza el cociente entre la variabilidad explicada por la regresión y la variabilidad total. Esta medida se llama *coeficiente de determinación*.

$$R^2 = \frac{VE}{VT} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Al valor R se le denomina también *coeficiente de correlación múltiple*.

EL coeficiente de correlación múltiple tiene las propiedades siguientes:

a) $|R| \leqslant 1$. Cuando $R = 1$ existe una relaión funcional exacta entre la variable respuesta y las variables explicativas.

b) $(1 - R^2) * 100$, representa el porcentaje de variablidad no explicada por la relación, y se espera que éste sea bajo.

El coeficiente de determinación presenta el problema de que aumenta siempre que se introducen nuevas variables en el modelo, aunque su efecto no sea significativo. Para evitar que R^2 aumente siempre al introducir nuevas variables, se define el *coeficiente de determinación corregido por grados de libertad*, $\bar{R^2}$, como:

$$\bar{R^2} = \frac{\sum e_i^2/(n-k-1)}{\sum (y_i - \bar{y_i})^2/(n-1)} \tag{2.16}$$

2.1.9. Selección de modelos

El estadístico AIC, criterio de información de Akaike, está basado en la función de verosimilitud e incluye una penalización que aumenta con el número de parámetros estimados en el modelo. Premia pues, los modelos que dan un buen ajuste en términos de verosimilitud y a la vez son parsimoniosos (tienen pocos parámetros).

Si $\hat{\beta}$ es el estimador máximo-verosimil del modelo (2.3), de

dimensión p, el estadístico AIC se define por:

$$AIC = -2l(\hat{\beta}) + 2p \tag{2.17}$$

Valores pequeños de (2.17), identifican mejores modelos.

2.1.10. Multicolinealidad

En el modelo de regresión la estimación del efecto de una variable depende del efecto diferencial, es decir, la parte de la variable que no está relacionada linealmente con las demás variables incluidas en el modelo. Cuando tenemos modelos con un gran número de variables explicativas puede ocurrir que dichas variables sean redundantes o, lo que es lo mismo, que muchas de estas variables estén correlacionadas entre sí. Al introducir variables correlacionadas en un modelo, el modelo se vuelve inestable. Por un lado, las estimaciones de los parámetros del modelo se vuelven imprecisas y los signos de los coeficientes pueden llegar incluso a ser opuestos a lo que la intuición nos sugiere. Por otro, se inflan los errores estándar de dichos coeficientes por lo que los test estadísticos pueden fallar a la hora de revelar la significación de estas variables. Por tanto, siempre que tengamos varias variables explicativas (sobretodo cuando tenemos un gran núumero de ellas), es importante explorar la relación entre ellas previamente al ajuste del modelo estadístico.

La estimación de los parámetros del modelo de regresión requiere la inversión de la matriz $X'X$. Si una de las variables explicativas es combinación lineal exacta de las demás (colinealidad con el resto), la matriz $X'X$ será singular y el sistema de ecuaciones

que determina los parámetros no tendrá solución única. Si esto pasa tendremos una situación de alta multicolinealidad y además:

a) Los estimadores $\hat{\beta}$ tendrán varianzas muy altas

b) las estimaciones $\hat{\beta}_i$ serán muy dependientes entre sí.

2.1.11. Tratamiento de la multicolinealidad

Cuando la recogida de datos se diseñe a priori, la multicolinealidad puede evitarse tomando las observaciones de manera que $X'X$ sea diagonal, lo que aumenta la precisión de la estimación (los estimadores tienen menor varianza). La multicolinealidad es un problema de la muestra y, por tanto, no tiene solución simple, ya que estamos pidiendo a los datos más información de la que contienen. Las dos únicas soluciones son:

1. Eliminar los regresores, reduciendo el número de parámetros a estimar
2. Incluir información externa a los datos.

La primera conduce a eliminar o bien ciertas variables muy correladas con las incluidas, o bien ciertas combinaciones lineales de ellas mediante componentes principales. La segunda a estimadores contraídos y bayesianos. Matemáticamente, estas dos soluciones suponen sustituir los estimadores minimocuadráticos por estimadores sesgados, pero de menor error cuadrático medio.

2.1.12. Diagnosis del modelo de regresión

Análisis de los residuos

La herramienta básica para el diagnóstico del modelo es el análisis de los residuos, tanto a través de gráficos, como de test que verifican la validez de las hipótesis asumidas en el ajuste del modelo de regresión lineal, y que son:

a) $E(e_i) = 0; \forall i = 1, \cdots, n$. Su esperanza es cero.

b) $Cov(e_i, e_j) = 0; \forall i \neq j$. Independientes entre sí.

c) $Var(e_i) = \sigma^2; \forall i$. Su varianza es constante.

d) $e \sim N(0; \sigma I)$. Su distribución es normal.

Tipos de Residuos

Residuos comunes: Los residuos comunes del modelo lineal $Y = \beta X + e$ consisten simplemente en las desviaciones entre los datos observados y_i y los predichos $\hat{y}_i$, obtenidos de:

$$e = Y - \hat{Y} = Y - X\hat{\beta} = (I - X(X'X)^{-1}X')Y$$

cuando $X'X$ es no singular.

Surge así una matriz básica en la definición de los residuos, denominada matriz gorro y definida por:

$$H = X(X'X)^{-1}X',$$

que tiene su importancia en la interpretacóon y redefinición de nuevos

tipos de residuos, como veremos. A sus elementos nos referiremos como h_{ij}. Esta matriz H es simétrica $(H' = H)$ e idempotente $(HH = H)$, de dimensión $n \times n$ y de rango $p = rang(X)$.

En términos de H, los residuos e se pueden escribir como:

$$e = Y - \hat{Y} = (I - H)Y,$$

esto es,

$$e_i = (1 - \sum_{j=1}^{n} h_{ij})y_i = y_i - \hat{y}_i,\ i = 1, \cdots,\ n.$$

Así, los residuos e_i son estimadores sesgados de los errores aleatorios u_i, cuyo sesgo y varianza depende de la matriz H gorro, y por lo tanto de la matriz de diseño X:

$$\begin{aligned} e - E[e] &= (I - H)(Y - X\beta) = (I - H)U \\ Var(e) &= (I - H)\sigma^2, \end{aligned}$$

de donde la varianza de cada residuo es:

$$Var(e_i) = (1 - h_{ii})\sigma^2; i = 1, \cdots,\ n,$$

y la correlación entre los residuos e_i y e_j:

$$Cor(e_i, e_j) = \frac{-h_{ij}}{\sqrt{(1 - h_{ii})(1 - h_{jj})}}$$

Residuos estandarizados: Son residuos de media cero y

varianza aproximadamente unidad, definidos por:

$$d_i = \frac{e_i}{\sqrt{s^2}},\ i = 1, \cdots,\ n,$$

donde s^2 es la estimación habitual de σ^2.

Residuos (internamente) estudentizados: Son unos residuos estandarizados (con varianza aproximadamente igual a 1) que tratan de librar a éstos de la variabilidad que introduce la matriz de diseño:

$$r_i = \frac{ei}{\sqrt{s^2(1-h_{ii})}}.$$

La distribución marginal de $r_i/(n-p)$ es una $Be(1/2,(n-p-1)/2)$. Así, cuando los grados de libertad del error, $n-p$, son pequeños, ningún residuo $|r_i|$ será demasiado grande. Se suelen utilizar para la detección de outliers u observaciones influyentes.

Residuos externamente estudentizados: Se trata de otro tipo de residuos estandarizados calculados según:

$$rt_i = \frac{ei}{\sqrt{s^2_{(i)}(1-h_{ii})}},$$

donde $s^2_{(i)}$ es la estimación de la varianza en el ajuste sin la observación i-ésima,

$$s^2_{(i)} = \frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}.$$

Estos residuos siguen una distribución t_{n-p-1} cuando los errores U son normales. Proporcionan pues, un procedimiento más formal

para la detección de outliers vía contraste de hipótesis. De hecho, se puede utilizar una corrección de Bonferroni para comparar todos los n valores $|rt_i|$ con el cuantil $t_{n-p-1,\alpha/2n}$, e identificar así los más "raros". Una ventaja de estos residuos externamente estudentizados es que si e_i es grande, la observación correspondiente aún destaca más a través del residuo rt_i.

Residuos parciales: Para una covariable x_j, estos residuos se obtienen cuando se prescinde de ella en el ajuste, y se calculan según:

$$e^*_{ij} = y_i - \sum_{k \neq j} \hat{\beta} X_{ik} = e_i + \hat{\beta}_j x_{ij}, i = 1, \cdots, n,$$

Se utilizan, como veremos, para valorar la linealidad entre una covariable y la variable respuesta, en presencia de los restantes predictores.

Residuos de predicción PRESS: Estos residuos se utilizan para cuantificar el error de predicción, y se calculan a partir de la diferencia entre la respuesta observada y_i y la predicción que se obtiene ajustando el modelo propuesto sólo con las restantes $n-1$ observaciones, $\hat{y}_i^{(i)}$. Están relacionados con los residuos habituales según:

$$e(i) = y_i - \hat{y}_i^{(i)} = \frac{e_i}{1 - h_{ii}}$$

La varianza del residuo $e_{(i)}$ es:

$$Var(e_{(i)}) = \sigma^2/(1 - h_{ii}),$$

con lo que, la versión estandarizada de los residuos de predicción utilizando como estimación de σ^2, s^2, coincide con el residuo

estudentizado r_i. Estos residuos se utilizan en el análisis de influencia y en la validación del modelo.

Linealidad

Una especificación deficiente del modelo de predicción, esto es, que no incluya como predictoras algunas variables que son útiles para explicar la respuesta, provoca estimaciones sesgadas. Si hay alguna variable explicativa que no ha sido incluida en el ajuste del modelo, representarla versus los residuos ayuda a identificar algún tipo de tendencia que dicha variable pueda explicar. Si se trata de una covariable, utilizaremos un gráfico de dispersión. Si se trata de un factor, diagramas de cajas, una por cada nivel de respuesta del factor. Si no se detecta ninguna tendencia en el gráfico de dispersión, o diferencia en las cajas, en principio no tenemos ninguna evidencia que nos sugiera incorporar dicha variable al modelo para predecir mejor la respuesta.

Los gráficos de residuos parciales sirven para valorar la linealidad entre una covariable y la variable respuesta, en presencia de los restantes predictores.

Homocedasticidad

La heterocedasticidad, que es como se denomina el problema de varianza no constante, aparece generalmente cuando el modelo está mal especificado, bien en la relación de la respuesta con los predictores, bien en la distribución de la respuesta o bien en ambas

cuestiones.

La violación de la hipótesis de varianza constante, $Var(U) = \sigma^2 I$, se detecta usualmente a través del análisis gráfico de los residuos:

- Gráficos de residuos ver sus valores ajustados $\hat{y}_i$. Cuando aparece alguna tendencia como una forma de embudo o un abombamiento, etc., entonces decimos que podemos tener algún problema con la violación de la hipótesis de varianza constante para los errores.
- Gráficos de residuos versus predictores x_j . Básicamente se interpretan como los gráficos de residuos versus valores ajustados y_i. Es deseable que los residuos aparezcan representados en una banda horizontal sin tendencias alrededor del cero.

Normalidad

La hipótesis de normalidad de los errores e_i en el modelo lineal, justifica la utilización de los test F y t para realizar los contrastes habituales y obtener conclusiones confiables a cierto nivel de confianza $(1 - \alpha)$ dado. En muestras pequeñas, la no normalidad de los errores es muy difícil de diagnosticar a través del análisis de los residuos, pues éstos pueden diferir notablemente de los errores aleatorios u_i. De hecho, la relación entre los residuos e_i y los errores aleatorios u_i, viene dada por:

$$e = (I - H)Y = (I - H)(X\beta + U) = (I - H)U$$

es decir,

$$e_i = u_i - \left(\sum_{j=1}^{n} h_{ij} u_j \right) \tag{2.18}$$

En muestras grandes no se esperan demasiadas diferencias entre residuos y errores, y por lo tanto hacer un diagnóstico de normalidad sobre los residuos equivale prácticamente a hacerlo sobre los errores mismos. Esto es debido a que por el teorema central del límite, el término $e_i - u_i$, al ser una suma converge a una distribución normal, incluso aunque los u_i no sigan tal distribución. Los términos h_{ii} tenderán a cero, y en consecuencia el término u_i en 2.18, tenderá a dominar en la expresión para los residuos e_i.

La forma habitual de diagnosticar no normalidad es a través de los gráficos de normalidad y de test como el de Shapiro-Wilks, específico para normalidad, o el de bondad de ajuste de Kolmogorov-Smirnov.

Un método muy sencillo de comprobar la suposición de normalidad es trazar una gráfica de probabilidad normal de los residuales. Es una gráfica diseñada para que al graficarse la distribución normal acumulada parezca una línea recta. Sean $e_{[1]} < e_{[2]} < \cdots < e_{[n]}$ los residuales ordenados en orden creciente. Si se grafican $e_{[i]}$ en función de la probabilidad acumulada $P_i = (i - 1/2)/n,\ i = 1,\ 2, \cdots,\ n$, en papel de probabilidad normal, los puntos que resulten deberían estar aproximadamente sobre una línea recta. Esa recta se suele determinar en forma visual, con atención en los valores centrales (por ejemplo, los puntos de probabilidad acumulada 0.33 y 0.67), y no en los extremos. Las diferencias apreciables respecto a la recta indican

que la distribución no es normal. A veces, las gráficas de probabilidad normal se trazan graficando el residual clasificado $e_{[i]}$ en función del "valor normal esperado", $\Phi^{-1}[(i-1/2)/n]$, donde Φ representa la distribución acumulada normal estándar. Esto es consecuencia de que $E(e_{[i]}) \backsimeq \Phi^{-1}[(i-1/2)/n]$, (ver [5]).

Incorrelación

Para el modelo lineal general asumimos que los errores observacionales están incorrelados dos a dos. Si esta hipótesis no es cierta, cabe esperar que un gráfico secuencial de los residuos manifieste alguna tendencia. Sin embargo, hay muchas formas en que los errores pueden estar correlados. De hecho, la independencia entre observaciones es una cuestión justificada básicamente por el muestreo realizado.

Un gráfico de los residuos en función de la secuencia temporal en que se observaron los datos puede ayudar a apreciar un problema de correlación de los residuos (llamada autocorrelación), o de inestabilidad de la varianza a lo largo del tiempo. Detectar autocorrelación ha de conducir a considerar otro tipo de modelos distintos.

Un tipo de correlación bastante habitual, es la correlación temporal, consistente en que las correlaciones entre los errores que distan s posiciones, correlaciones que denotamos en adelante por ρ_s, son siempre las mismas.

$$\rho_s = Cov(r_i, r_{i+s}).$$

Si hay correlación temporal positiva, los residuos tienden a ser consecutivos en la secuencia temporal. Si la correlación temporal es negativa, un residuo positivo suele ser seguido de uno positivo y viceversa. Los **gráficos lag** ayudan a detectar este tipo de correlación temporal. Dichos gráficos consisten en representar cada residuo (excepto el primero) versus el residuo anterior en la secuencia temporal sospechosa de inducir la correlación.

Un test habitual para detectar cierto tipo de correlación temporal es el test de **Durbin-Watson**. Asumiendo normalidad, todas las correlaciones temporales entre los residuos han de ser cero, $\rho_s = 0$. El test de Durbin-Watson nos permite contrastar esto:

$$H_0 : \rho_s = 0; \text{ versus } H_1 : \rho_s \neq 0.$$

Para resolver el test se utiliza el estadístico de Durbin-Watson, definido por:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

La distribución de d depende de los datos X, está definida entre 0 y 4 y es simétrica alrededor de 2. Los valores críticos por tanto, han de calcularse para cada problema concreto.

En el caso de detectar correlación temporal, cabe estudiar la naturaleza de la correlación de los errores y buscar algún modelo válido para modelizar la dependencia temporal (análisis de series temporales). El planteamiento entonces es utilizar modelos de tendencias temporales para predecir la respuesta, puesto que ésta depende de la secuencia temporal en que se han observado los datos.

2.1.13. Validación del modelo de regresión

Validación cruzada

Una vez ajustado un modelo cabe comprobar su validez predictiva, esto es, si es o no, especialmente sensible a la aleatoriedad de la muestra con la que se realiza el ajuste. Esto es lo que se denomina validación cruzada del modelo. La validación cruzada consiste en realizar el ajuste con una parte de los datos y compararla con la que se obtiene con otra parte de los datos. Si se obtienen diferencias severas, concluiremos con que el ajuste no es robusto y por lo tanto su validez predictiva queda en cuestión. Para realizar la validación cruzada disponemos de dos procedimientos básicos [2].

1. Un procedimiento riguroso de validación del modelo es verificar sus resultados con una muestra independiente de la utilizada para construirlo. Cuando no es posible muestrear más, se puede considerar el ajuste con una parte de los datos y dejar los restantes para la validación del mismo. Por supuesto, este procedimiento resta fiabilidad a las estimaciones, ya que al obtenerse con menos datos producen errores estándar mayores. Si por ejemplo se ha ajustado el modelo con una muestra de n_1 observaciones denotada por $M_1,\ y_1 = X_1\beta_1 + e_1$, cabe plantear el ajuste con una muestra independiente M_2 con n_2 observaciones, $y_2 = X_2\beta_2 + e_2$ y a continuación resolver el contraste:

$$H_0 : \beta_1 = \beta_2$$

$$H_1 : \beta_1 \neq \beta_2$$

Dicho contraste se resuelve, cuando $n_2 > p$, con p el núumero de coeficientes del modelo, con el estadístico F basado en las sumas de cuadrados residuales del modelo completo, SSE_T , y de los parciales para el conjunto de observaciones y_1, SSE_1, y para y_2, SSE_2, definido por:

$$\frac{(SSE_T - SSE_1 - SSE_2)/p}{(SSE_1 + SSE_2)/(n_1 + n_2 - 2p)} \sim F_{(p,n_1+n_2-2p)} \qquad (2.19)$$

y cuando $n_2 < p$ (en cuyo caso $SSE_2 = 0$), con:

$$\frac{(SSE_T - SSE_1)/n2}{SSE_1/(n_1 - p)} \sim F_{(n_2,n_1-p)} \qquad (2.20)$$

2. Otro procedimiento interesante para juzgar la robustez de un modelo es la validación cruzada una a una, y consiste en que, para cada observación i, ajustar el modelo con las $n - 1$ observaciones restantes y con él predecir la respuesta i, $\hat{y}_i^{(i)}$.

   Se define el error cuadrático de validación como:

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

   Algunos autores propusieron elegir los estimadores de β que minimizaran la expresión 2.21, procedimiento que requiere de los mínimos cuadrados ponderados y que da lugar a estimadores poco robustos, pues las observaciones más influyentes (alejadas del resto), son las que más peso tienen en la estimación de los parámetros. Otra medida de robustez la proporciona el

coeficiente de robustez:

$$B^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i^{(i)})^2} \tag{2.21}$$

que da un valor entre 0 y 1 y cuantifica la robustez del modelo. Cuando las predicciones $\hat{y}_i^{(i)}$, sean próximas a las del ajuste con todos los datos, $\hat{y}_i$, el valor de B^2 será próximo a 1, y si hay mucha diferencia, será casi cero.

2.2. Modelos ANCOVA.

2.2.1. Variables ficticias

Estos son los modelos más usados, debido a que en los modelos de regresión se encuentran situaciones en las cuales las variables explicativas cambian bruscamente su impacto en la variable respuesta y que la naturaleza de este cambio, no se puede atribuir a una variable que sea medible, ya que estas variables son de naturaleza cualitativa y pueden modelizarse, aplicando dos valores: cero y uno.

El efecto de una variable cualitativa indica la presencia o la ausencia de una cualidad o atributo, tales como el sexo, educación, religión, etc.

El propósito de estos modelos consiste en incluir la variable atributo dentro del modelo. Por ejemplo, supongamos que se tiene una variable cualitativa z con dos categorías y una variable cuantitativa x,

entonces escribiremos el modelo como:

$$y = \beta_0 + \beta_1 x + \beta_2 z + e$$

Esto implica que el valor esperado de la respuesta donde $z = 0$, es:

$$E[y/z = 0] = \beta_0 + \beta_1 x$$

mientras que, para $z = 1$:

$$E[y/z = 1] = \beta_0 + \beta_1 x + \beta_2$$

Por tanto el coeficiente β_2 de la variable binaria medirá el efecto *incremental* que produce $z = 1$ sobre $z = 0$. Llamaremos entonces *variables ficticias* a las variables binarias que, como z, representan atributos o formas de clasificar los datos muestrales.

Esta idea puede generalizarse a cualquier número de variables. Supongamos que queremos separar las observaciones en D grupos distintos, en función de su respuesta media, a igualdad de valores de las variables cuantitativas.

Para este tipo de problemas existe dos posibles soluciones. La más utilizada es introducir $D - 1$ variables ficticias:

$$z_i = \begin{cases} 0 \; si \; la \; observación \; no \; pertence \; al \; grupo \; i \\ 1 \; si \; la \; observación \; pertence \; al \; grupo \; i \end{cases} \quad ; i = 1, \cdots, D-1$$

y establecer la regresión de forma habitual, tratando a estas $D - 1$ nuevas variables z como nuevos regresores. La razón de introducir

$D-1$ variables, en lugar de D, es evitar que la matriz $\tilde{X}'\tilde{X}$ sea singular. En efecto, si definimos una variable para cada grupo, la matriz $\tilde{X}$ sería:

$$\tilde{X} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & \tilde{x}_{11} & \cdots & \tilde{x}_{k1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & 1 & 0 & \cdots & 0 & \vdots & & \vdots \\ \vdots & & & \vdots & & \vdots & & \vdots \\ \vdots & 0 & 1 & \cdots & 0 & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & 1 & \cdots & 0 & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ \vdots & \vdots & 0 & \cdots & 1 & \vdots & & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 & \tilde{x}_{1n} & \cdots & \tilde{x}_{kn} \end{bmatrix}$$

la primera columna corresponde al término constante, β_0, las siguientes a las D variables ficticias y las últimas a las variables explicativas cuantitativas. Se observa que la suma de las columnas de las variables ficticias es uno, igual a la columna de β_0, con lo que la matriz $\tilde{X}'\tilde{X}$ no será invertible. Así si $z_i = 0,\ i = 1, \cdots, D-1$, β_0 representa el valor medio de la respuesta para el grupo definido por valores cero de todas las variables ficticias. Este es el grupo que actúa como grupo de referencia. Para cualquier otro grupo el coeficiente de $z_i = 1$ representa el valor incremental de la respuesta media en el grupo i respecto al grupo de referencia.

La segunda solución es eliminar el término β_0, lo que permite incluir D variables ficticias. Entonces cada coeficiente de z_i estima la

respuesta media de su grupo.

Estas ideas se generalizan para cualquier número A de atributos distintos o variables cualitativas, cada una a n_a niveles distintos.

2.2.2. Formulación del modelo ANCOVA

Suponemos que para explicar una variable respuesta disponemos de una variable cualitativa, que clasificamos en p grupos o clases y k variables explicativas cuantitativas $(x_1, \cdots, x_k)$. Se toman muestras, mediante muestreo aleatorio simple, de la variable respuesta en los distintos grupos, y se desea estudiar cómo estos grupos afectan a la respuesta. En concreto, sea X la matriz de datos de las variables cuantitativas y Z la matriz de las $p-1$ variables ficticias. Se trata de elegir entre los modelos siguientes:

1. Los grupos no influyen en absoluto y todas las observaciones se generan con el mismo modelo:

$$Y = X\beta_1 + U_1$$

2. Los grupos difieren en la respuesta media, pero el efecto de las x es idéntico entre ellos:

$$Y = X\beta_2 + Z\alpha + U_2$$

2.2.3. Estimación del modelo

Si los coeficientes de regresión son idénticos en todas las regresiones, el efecto del grupo se recogerá con un término independiente distinto en cada grupo. Esto equivale a definir $p-1$ variables ficticias $z_2, \cdots, z_p$ por:

$$z_i = \begin{cases} 1 \; si \; la \; observación \; es \; de \; la \; clase \; j \\ 0 \; si \; la \; observación \;\; no \; es \; de \; la \; clase \; j \end{cases}$$

resultando el modelo:

$$Y = Z\alpha + X\beta + U_2$$

donde la matriz Z tiene n filas y $p-1$ columnas; el vector α, $p-1$ coeficientes; X es, como siempre, $n(k+1)$ y β es un vector $k+1$. Supondremos que las n_1 primeras observaciones corresponden al grupo 1, las n_2 siguientes al grupo 2, etc. Entonces:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ \vdots \\ y_{n_1+n_2+1} \\ \vdots \\ \vdots \\ y_{n_1+n_2+n_3+1} \\ \vdots \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \vdots & \cdots & \vdots \\ 1 & 0 & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & \cdots & \vdots \\ 0 & 0 & 1 & \cdots & \vdots \\ \vdots & \vdots & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \alpha_2 \\ \vdots \\ \vdots \\ \alpha_p \end{bmatrix} + \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} U_1 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ U_n \end{bmatrix}$$

uniendo las matrices Z y X en una nueva matriz de datos X^* y los vectores α y β en un vector β^*, se obtiene un modelo general de regresión que se estimará por el procedimiento descrito en la sección 2.1.1. Llamaremos e_2 al vector de residuos estimados:

$$e_2 = Y - \hat{Y} = Y - Z\hat{\alpha} - X\hat{\beta}_2$$

que tendrá $n - (k+1) - (p-1)$ grados de libertad.

Al estimar un modelo ANCOVA y comprobar que una o más variables cualitativas son significativas, es necesario realizar comparaciones múltiples entre los diferentes niveles o categorías que éstas tengan. Esto se hace usando la prueba de Diferencia Significativa Ho-

nesta (DSH) de Tukey. El estadístico de prueba es:

$$DSH = F_{1-\alpha/2,k-1,n-k}\sqrt{\frac{SCM_D}{n}}$$

donde SCM_D, es la suma de cuadrados medios dentro de cada categoría de la variable. Se calculan los valores absolutos de las diferencias de medias entre las categorías y se observa si hay diferencias cuyo valor es mayor que el DSH, si esto sucede se concluye que hay diferencias entre esas categorías. Inicialmente suponemos que cada categoría tiene n elementos. En caso de no verificarse esta condición, es posible aplicar el método tomando

$$n = \frac{1}{\sum_{i=1}^{a} \frac{1}{n_i}}$$

donde a es el número de categorías de cada variable.

2.3. Ejemplo ilustrativo

En esta sección, se ilustra mediante un ejemplo sencillo, el manejo y aplicación de los modelos ANCOVA de regresores cualitativos y cuantitativos.

Supongamos que se desea estimar un modelo para explicar el salario de 13 personas en función de las siguientes variables:

- **Salarios:** Variable dependiente (en dólares).
- **Estudio:** Nivel de estudio de la persona (Primaria, secundaria, técnica, pregrado y postgrado).

- **Sexo:** Sexo de la persona (M, F).
- **Gasto:** Gasto personal (en dólares).
- **Consumo:** Consumo familiar (en dólares).

La información se presenta en la Tabla A.1. De acuerdo a lo estudiado en las secciones 2.1 y 2.2 (tomando como referencia la categoría Postgrado de la variable Estudio), podría plantearse el siguiente modelo:

$$\begin{aligned} Y =& \alpha_0 + \alpha_1 Primaria + \alpha_2 Secundaria + \alpha_3 Técnica + \\ & \alpha_4 Pregrado + \beta_1 Gasto + \beta_2 Consumo + u_t \end{aligned}$$

En este modelo, cada α_i reflejará correspondientemente la incidencia de la Educación Secundaria, Técnica, de Pregrado y de Postgrado y los β_1 y β_2 miden la incidencia del Gasto Familiar y del Consumo Personal sobre el salario nominal respectivamente.

En base a los resultados obtenidos mediante el código del Apéndice A.3, podemos observar que resultados muestran que existen diferencias significativas entre los niveles de la variable Estudio.

donde podemos observar que todas las categorías de la variable Estudio son significativas y, las variables Gasto y Consumo no lo son. Así el modelo obtenido es:

$$\begin{aligned} Y =& 352.42 - 293.64 Primaria - 333.59 Secundaria - \\ & 297.69 Técnica - 247.39 Pregrado. \end{aligned}$$

Esto significa por ejemplo, que aquellas personas que poseen un postgrado tienen un ingreso medio de $293.64 más, con respecto

a aquellas personas que tienen Primaria; $333.59 más, que aquellas personas que poseen Secundaria.

Capítulo 3

Modelos para el rendimiento

3.1. Análisis descriptivos de los datos

En un principio se cuenta con un total de 2070 observaciones, de las cuales, cada una corresponde a un lote cosechado en la Zafra 2014-2015. A continuación se presenta un análisis descriptivo de los datos para poder tener una mejor visión de los mismos.

Se cuenta con las siguientes variables:

Variables dependientes

- **Rendimiento de Campo o rendimiento de caña:** Toneladas de caña producidas por hectárea cultivada (t/ha).
- **Rendimiento Industrial o rendimiento de azúcar:** Kilogramos de azúcar producidos por tonelada de caña molida (kg/t).

Variables independientes

Variables agrícolas.

Son las prácticas y condiciones agrícolas a las que ha sido sometido

cada lote por parte de los encargados con el fin de obtener buenos rendiminetos.

- **Altura:** Altura con respecto al nivel del mar a la que se encuentra el cultivo. Es expresada en metros (m.s.n.m).
- **Tipo de corte:** Forma en la que se ha cosechado la caña (Manual o Mecanizada).
- **Edad:** Período de tiempo desde el último corte. Por lo general se expresa en meses.
- **Etapa de maduración o madurez:** Tiempo que tarda la caña en llegar a su madurez, categorizada en:
  - Intermedia
  - Intermedia tardía
  - Temprana
  - Temprana intermedia
  - Otra
- **Humedad:** Variable de calidad que indica el contenido de agua que posee la caña.
- **Madurante:** Indica el tipo de producto que se aplica para modificar el desarrollo y crecimiento de la caña.
  - MODDUS
  - NO
  - OTRO

    - ROUNDUP SL
    - SELECT
    - TOUCHDOWNS

- **Número de cortes:** Número de veces en las que ha sido cosechado el cañal, sin haber sido renovado.
- **Textura del suelo:** Característica del suelo donde se cultiva la caña de azúcar.

    - Arcilloso
    - Franco
    - Franco - Franco Arenoso
    - Franco Arcilloso
    - Franco Limoso
    - Otro

- **Variedad:** Tipo de variedad a la que pertenece la caña.

    - CP-72-1210
    - CP-72-2086
    - CP-73-1547
    - CP-88-1165
    - CP-89-2143
    - MEX-79-431
    - OTRAS
    - PR-87-2080

- VARIAS

Variables ambientales.
Condiciones climáticas a las que ha estado expuesto cada lote.

- **Lluvia:** Milímetros de lluvia que ha recibido la caña desde su corte anterior, hasta su fecha de corte actual.
- **Amplitud térmica:** Promedio de la diferencia entre la temperatura más alta del día y la menor temperatura durante la noche y madrugada.
- **Temperatura:** Temperatura media a la que ha estado expuesta la caña, reportada en grados centígrados (°C).

La Altura, Edad, Humedad, Número de cortes, Lluvia acumulada, Amplitud térmica, y la Temperatura son variables continuas y, por tanto consideradas como covariables. Las demás variables son cualitativas y consideradas como factores. Se pretende modelar los rendimientos en función de las variables independientes, y determinar la influencia de éstas para tomar decisiones que contribuyan a mejorar los rendimientos.

En la Tabla 3.1 se presentan los principales estadísticos para los rendimientos de la caña

Se ha tenido un rendimiento medio de azúcar de $104.69\ Kg/t$ con una desviación típica de $8.76\ Kg/t$. Además se ha obtenido un rendimiento máximo de $132\ kg/t$ y un rendimiento mínimo de $71.71\ kg/t$, y puede notarse que sólo el $25\,\%$ de los lotes ha tenido un rendimiento de más de $110\ kg/t$. Para el rendimiento de caña, se

Estadísticos		**kg az/t**	**t caña/ha**
Número de casos		2070	2070
Media		104.69	90.74
Mediana		105.22	90.18
Desv. típ.		8.76	32.29
Varianza		76.67	1042.82
Asimetría		-.38	0.52
Curtosis		.37	3.76
Mínimo		71.71	2.82
Máximo		132.97	379.57
Percentiles	25	99.15	71.08
	50	105.22	90.18
	75	110.78	110.37

Tabla 3.1: Estadísticos para los rendimientos

tiene un rendimiento medio de $90.74\ t/ha$ con una desviación típica de $32.29\ t/ha$. Además un rendimiento máximo de $379\ t/ha$ y un rendimiento mínimo de $2.82\ t/ha$, puede notarse que sólo el $25\,\%$ de los lotes ha tenido un rendimiento mayor que $110.37\ t/ha$.

La Figura 3.1 muestran los gráficos Boxplot para ambos rendimientos.

Un intervalo de confianza del $95\,\%$ para el rendimiento de azúcar es: **[104.31, 105.06]**, significa que de cada 100 lotes se espera que 95 tengan rendimientos medios entre $104.31\ kg/t$ y $105.06\ kg/t$.

Un intervalo de confianza del $95\,\%$ para el rendimiento de caña es: **[89.34, 92.13]**, significa que de cada 100 lotes se espera que 95 tengan rendimientos medios entre $89.34\ t/ha$ y $92.13\ t/ha$.

Algunos de los datos presentan incongruencias, esto se debe a que en ocasiones las entregas de un mismo lote pueden ser registradas en otro diferente, obteniéndose de esta forma rendimientos

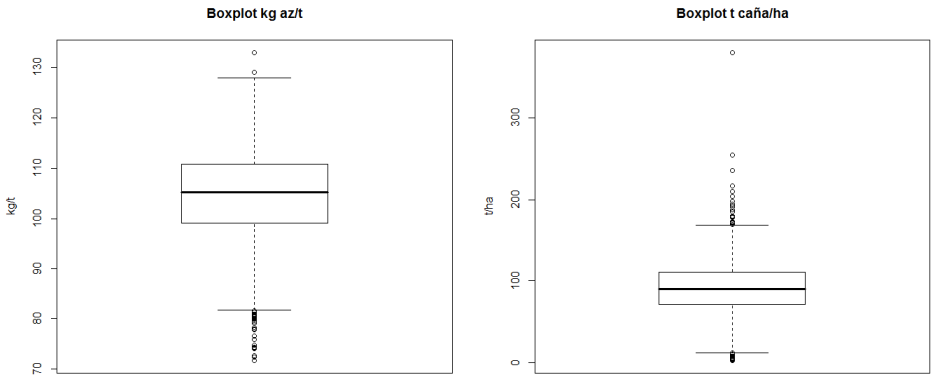


Figura 3.1: Boxplot para los rendimientos

que rondan los extremos, altos y bajos tal como se observa en la Figura 3.1. De acuerdo a los encargados agrícolas el rendimiento debe estar entre 15 y 180 t/ha. Por lo que se descartan del análisis de rendimiento de campo aquellos lotes que presentan datos extremos.

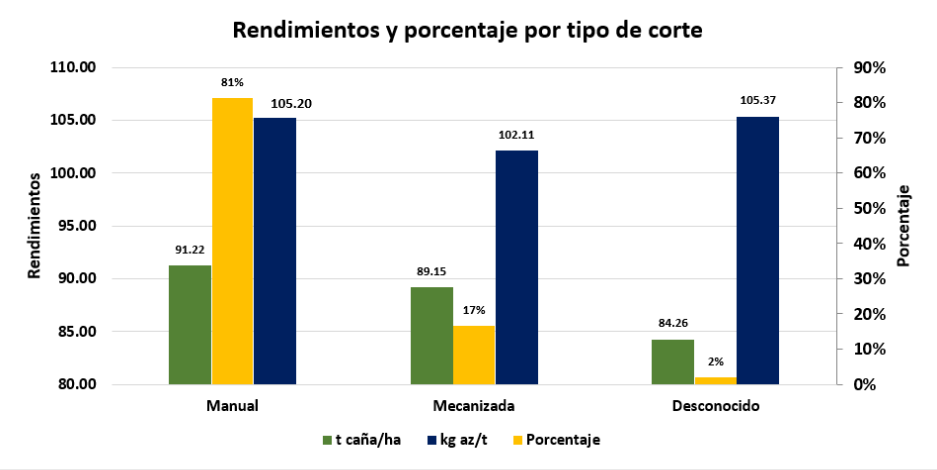
Los descriptivos de las demás variables se presentan en el Apéndice B.1.

3.1.1. Descriptivos en base a las variables cualitativas

- **Tipo de corte**

  En la Figura 3.2, se muestran los redimientos y el porcentaje por tipo de corte usado en la cosecha de la caña. Se observa que el $81\,\%$ de la caña ha sido cosechada de manera manual, un $17\,\%$ de manera mecanizada y sólo para un $2\,\%$ de los lotes se desconoce la manera de corte, (estos lotes no se usarán en el análisis).

  En base al tipo de corte, el mayor rendimiento medio de caña se obtiene cosechándola de manera manual. Este rendimiento es



Fuente: Elaboración propia

Figura 3.2: Rendimientos y porcentaje por tipo de corte

de $91.22\ t/ha$, y se obtiene un rendimiento medio de azúcar de $105.27\ kg/t$.

- **Etapa de maduración o madurez**

  En la Tabla 3.2, se presenta el tipo de madurez para cada variedad.

  Se observa que la mayoría de caña es de madurez intermedia. Esta comprende el 42 % del total cosechado. La minoría es de madurez otra con un 6 % del total cosechado. Este tipo de madurez se presenta en lotes en los que no se ha cosechado una misma variedad de caña.

  En la Figura 3.3, se muestran los rendimientos y el porcentaje por tipo de madurez de la caña cosechada.

  En base al tipo de madurez, el mayor rendimiento medio de caña se obtiene del tipo Intermedia. Este rendimiento es de $95.62\ t/ha$, y se obtiene un rendimiento medio de azúcar de

Madurez/ Variedad	Frecuencia
Intermedia	
CP-72-2086	813
OTRAS	56
Total	**869**
Intermedia tardía	
MEX-79-431	482
OTRAS	23
PR-87-2080	42
Total	**547**
Otra	
OTRAS	5
VARIAS	113
Total	**118**
Temprana	
CP-72-1210	120
CP-73-1547	185
OTRAS	30
Total	**335**
Temprana intermedia	
CP-88-1165	63
CP-89-2143	138
Total	**201**
Total general	**2070**

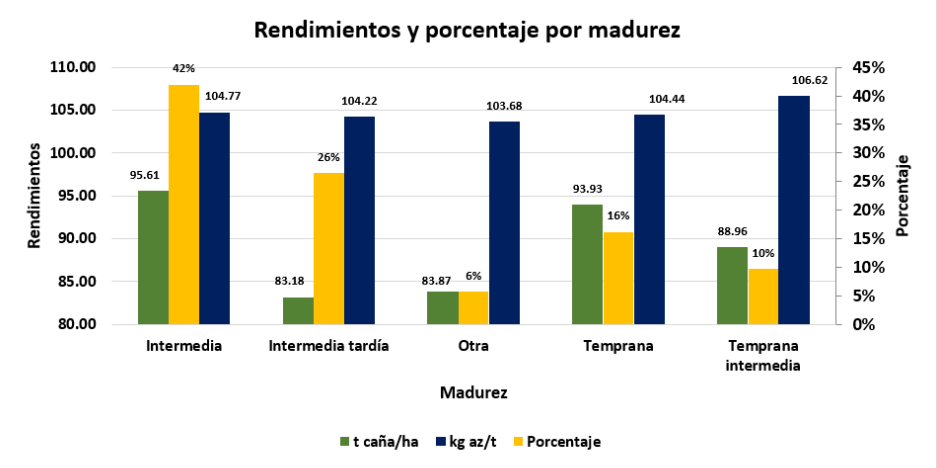
Tabla 3.2: Frecuencia de madurez

$104.77\ kg/t$.

- **Tipo de madurante**

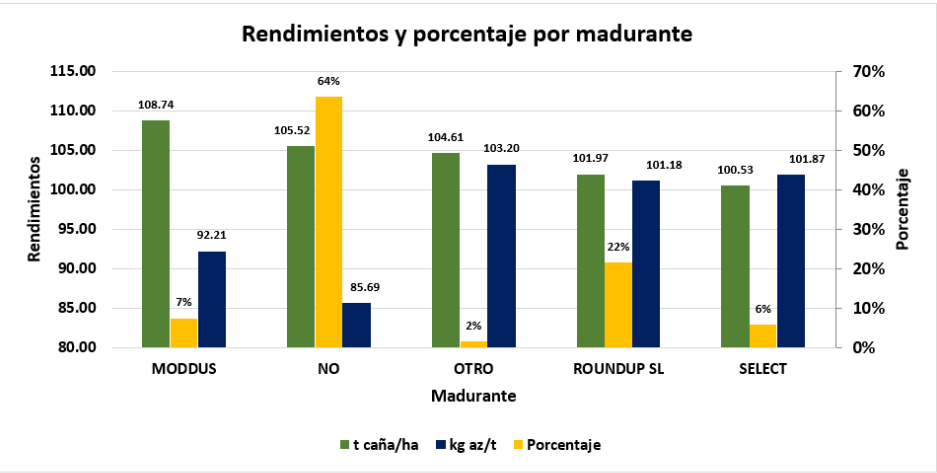
  En la Figura3.4 se muestran los rendimientos y el porcentaje por tipo de madurante usado en la caña cosechada.

  El madurante **ROUNDUP SL** es el más usado y este se ha utilizado en el $22\,\%$ de los lotes cosechados. El siguiente madurante más usado es el **MODDUS**, aplicado al $7\,\%$ de los lotes cosechados. Puede notarse que se ha usado madurante en



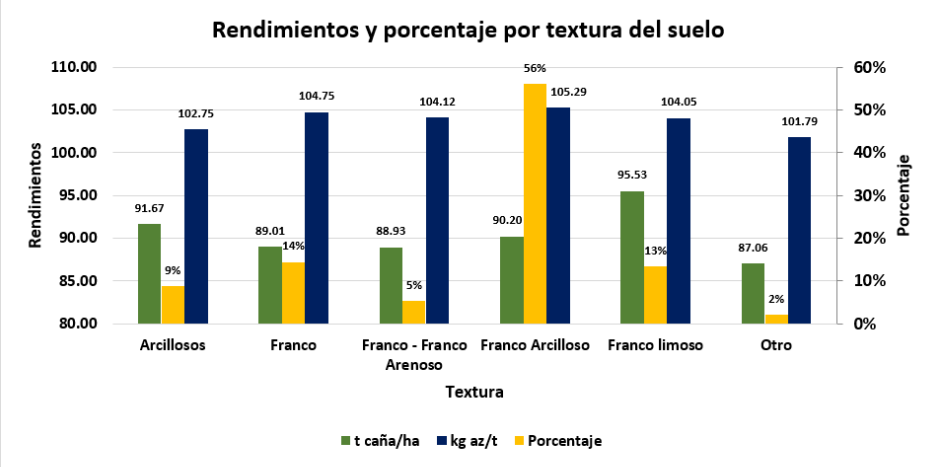
Fuente: Elaboración propia

Figura 3.3: Rendimientos y porcentaje por tipo de madurez



Fuente: Elaboración propia

Figura 3.4: Rendimientos y porcentaje por madurante aplicado



Fuente: Elaboración propia

Figura 3.5: Rendimientos y porcentaje por textura del suelo

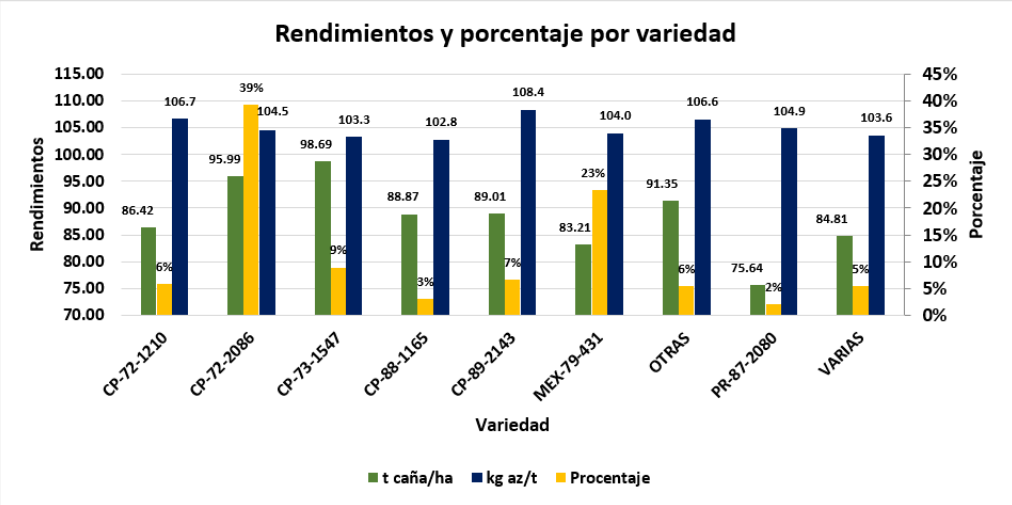
el $36\,\%$ de los lotes, mientras que al $64\,\%$ de los lotes cosechados no se les ha aplicado madurante.

El mayor rendimiento medio de caña ha sido obtenido de los lotes en los que se ha aplicado el tipo de madurante **MODDUS**. Su rendimiento es de $108.74\ t/ha$ y con un rendimiento medio de azúcar de $92.21\ kg/t$, este madurante se ha aplicado al $7\,\%$ de los lotes. El mayor rendimiento medio de azúcar se ha obtenido de los lotes en los que se ha aplicado **OTRO**. Su rendimiento medio de $103.2\ kg/t$ y un rendimiento de caña de $104.61\ t/ha$. Este tipo de madurante sólo ha sido aplicado en el $2\,\%$ de los lotes.

- **Textura del suelo**

  En la figura 3.5, se muestran los rendimientos y el porcentaje por textura del suelo donde ha sido cultivada la caña.

  La textura Franco Arcilloso es la que se presenta en la mayoría



Fuente: Elaboración propia

Figura 3.6: Rendimientos y porcentaje por variedad

de lotes, concretamente en el $56\,\%$ de ellos, Franco es la segunda textura más común, presente en el $14\,\%$ de los lotes.

El mayor rendimiento medio de caña ha sido obtenido en el suelo con textura Franco limoso, siendo de $95.53\,t/ha$. Este suelo tiene un rendimiento medio de azúcar de $95.53\ kg/t$, y esta textura la posee el $13\,\%$ de los lotes. El mayor rendimiento medio de azúcar se ha obtenido en el suelo con textura Franco Arcilloso, éste rendimiento es de $105.29\,kg/t$ y esta textura la posee el $56\,\%$ de los lotes.

- **Variedad cosechada**

  La figura 3.6, muestran los rendimientos y el porcentaje para cada variedad cosechada. Puede notarse que la variedad **CP-72-2086** es la que más se cosecha, con el $39\,\%$ de los lotes cultivados. La siguiente variedad más cosechada es la **MEX-79-431**, con un $23\,\%$, y la variedad con el menor porcentaje

cosechado es la **PR-87-2080**.

El mayor rendimiento medio de caña ha sido obtenido por la variedad **CP-73-1547**, siendo de $98.69\ t/ha$. Esta variedad tiene un rendimiento medio de azúcar de $103.3\ kg/t$, y en la Figura 3.6 podemos ver que esta variedad representa solo el $9\,\%$ de la cosecha.

El mayor rendimiento medio de azúcar se ha obtenido de la variedad **CP-89-2143**, este rendimiento es de $108.4\ kg/t$ y esta variedad representa solo el $7\,\%$ de la cosecha.

3.2. Modelo para el rendimiento de fábrica o de azúcar (kg azúcar/t de caña)

Luego de verificar la información se encontraron datos incongruentes o faltantes. Debido a esto se realizó la corrección en conjunto con los especialistas y en base a revisión bibliográfica para tener información confiable y verdadera. Esto llevó a quedarnos con 1631 observaciones de las 2070 que se tenían al principio.

3.2.1. Estimación y selección del modelo

Primero se realiza un análisis de correlación entre las variables continuas para identificar los posibles tipos de relación y la presencia de colinealidad entre las variables.

En la Tabla 3.3 se observa que se tiene una fuerte correlación negativa entre el rendimiento de azúcar y la humedad de la caña,

	kg az/t caña	Humedad	Ncortes	Edad	Altura	Lluvia	Amp térmica	Temp
kg az/t caña	1	-0.8773	-0.0253	0.2901	0.2684	0.2301	0.0891	-0.2255
Humeda	-0.8773	1	0.0734	-0.3082	-0.1796	-0.1683	0.0216	0.1881
Ncortes	-0.0253	0.0734	1	-0.0279	0.0216	0.0432	0.0097	-0.0377
Edad	0.2901	-0.3082	-0.0279	1	0.0974	0.1064	-0.0504	-0.3147
Altura	0.2684	-0.1796	0.0216	0.0974	1	0.3996	0.3106	-0.3559
Lluvia	0.2301	-0.1683	0.0432	0.1064	0.3996	1	0.3156	-0.4663
Amp térmica	0.0891	0.0216	0.0097	-0.0504	0.3106	0.3156	1	0.2610
Temp	-0.2255	0.1881	-0.0377	-0.3147	-0.3559	-0.4663	0.2610	1

Tabla 3.3: Correlaciones para el rendimiento de azúcar

por lo que puede esperarse que la humedad tenga un coeficiente negativo con un fuerte aporte en el modelo. Con las demás variables, el rendimiento de azúcar presenta una relación baja, pero debido a los conocimientos teóricos, éstas tienen influencia en el rendimiento y, por lo tanto, no serán descartadas del análisis. La relación entre las variables explicativas es baja, por lo que puede descartarse la existencia de colinealidad.

3.2.2. Gráficos de medias para el rendimiento de azúcar

Se presentan los gráficos para el rendimiento medio con un intervalo de confianza del $95\,\%$, por categoría para cada factor para así poder identificar posibles diferencias entre estos.

- **Rendimiento medio por tipo de corte**

  La Figura 3.7 muestra para cada tipo de corte el rendimiento medio de azúcar y su correspondiente intervalo de confianza.

  Se observa una clara diferencia en el rendimiento medio obtenido de manera manual con el obtenido de manera mecanizada. La caña que ha sido cosechada de manera manual presenta un mayor rendimiento con una menor variabilidad.

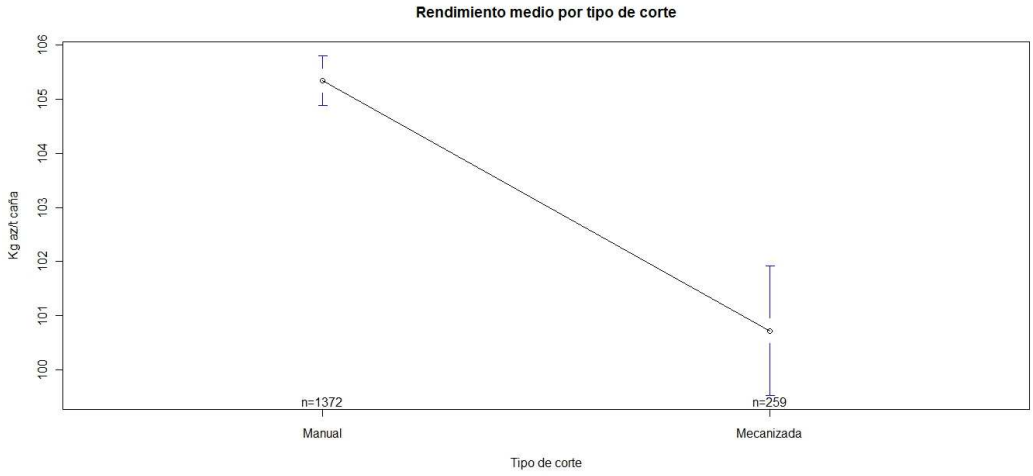


Figura 3.7: Rendimiento medio por tipo de corte

- **Rendimiento medio por tipo de madurez**
  La Figura 3.8, muestra para cada tipo de madurez, el rendimiento medio de azúcar y su correspondiente intervalo de confianza. Se observan diferencia entre los tipos de madurez, el mayor rendimiento se tiene en los lotes de madurez Temprana intermedia, se tienen 156 lotes con este tipo de caña y el menor rendimiento es obtenido en los lotes con madurez Otra.

- **Rendimiento medio por tipo de madurante**
  La Figura 3.9 muestra para cada tipo de madurante, el rendimiento medio de azúcar y su correspondiente intervalo de confianza. Puede notarse que existe diferencia en los rendimientos medios por madurante, el madurante con los mayores rendimientos es el MODDUS, los lotes a los que se les ha aplicado ROUNDUP SL presentan baja variabilidad, pero también bajos rendimientos. Además, a la mayoría de lotes no se las ha aplicado madurante y estos presentan menor variabilidad en el rendimiento, es decir,

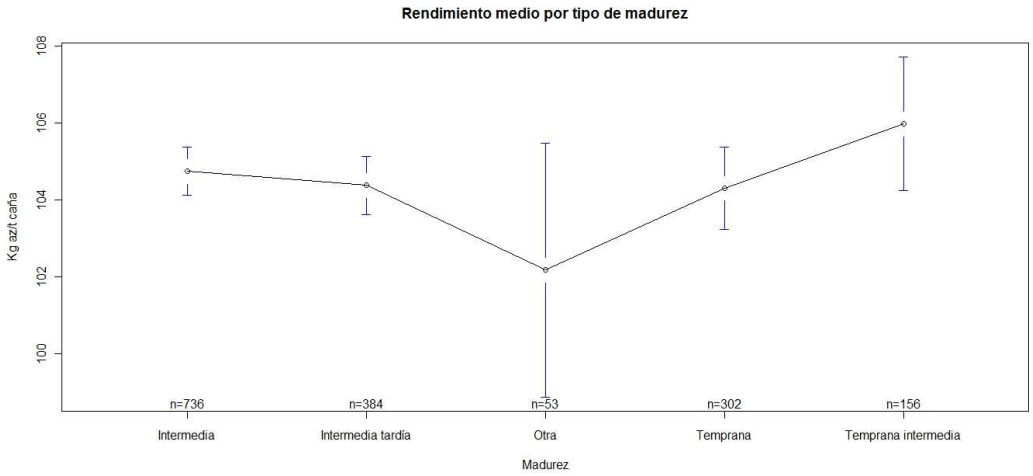


Figura 3.8: Rendimiento medio por tipo de madurez

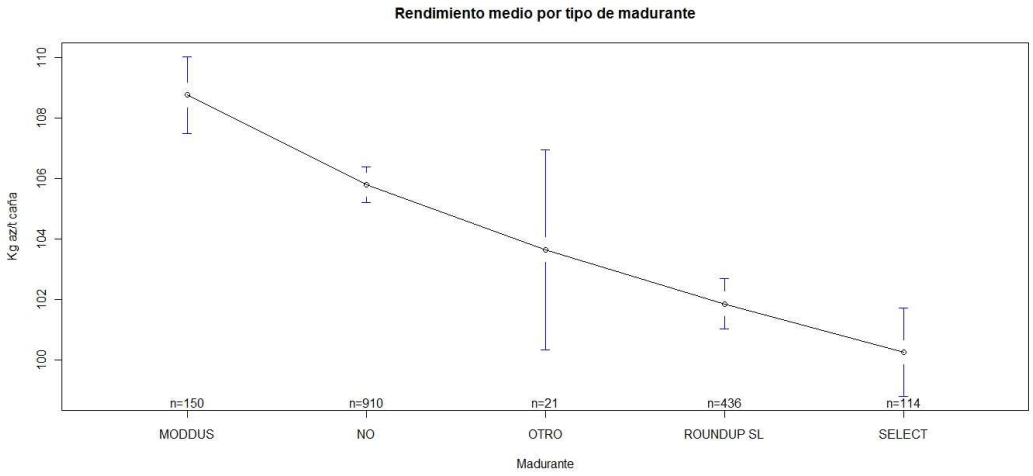


Figura 3.9: Rendimiento medio por tipo de madurante

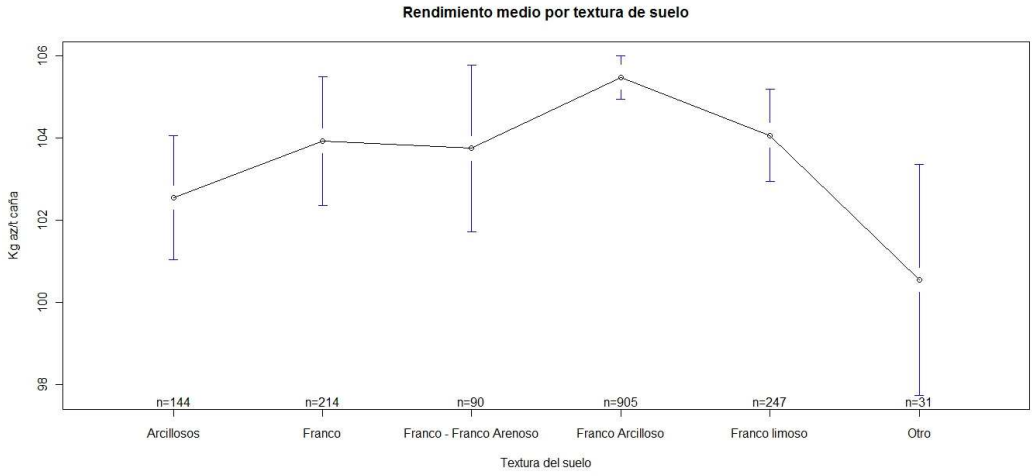


Figura 3.10: Rendimiento medio por textura del suelo

en estos lotes se puede esperar tener rendimientos medios muy parecidos.

- **Rendimiento medio por textura del suelo**
  La Figura 3.10 muestra para cada tipo de textura del suelo, el rendimiento medio de azúcar y su correspondiente intervalo de confianza. Al observar los rendimientos medios puede notarse que se presentan diferencias de acuerdo a la textura del suelo. El mayor rendimiento se presenta en los lotes con textura Franco Arcilloso, que es a la vez la textura con mayor frecuencia y donde se presenta la menor variabilidad.

- **Rendimiento medio por variedad**
  La Figura 3.11 muestra para cada tipo de variedad, el rendimiento medio de azúcar y su correspondiente intervalo de confianza.

  En el gráfico puede notarse que existen diferencias entre

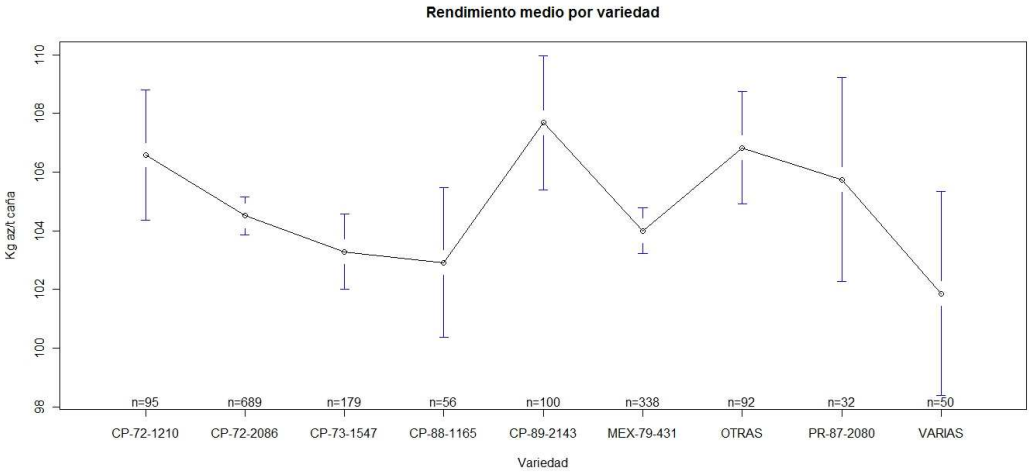


Figura 3.11: Rendimiento medio por variedad

los rendimientos medios obtenidos para cada variedad. Las variedades con una mayor variabilidad son la PR-2080 y VARIAS, y las variedades con una menor variabilidad son la CP-72-2086 y la MEX-79-431.

Para hacer la estimación del modelo se eligen las categorías de referencia de cada variable cualitativa, estas son las categorías que presentan una mayor frecuencia. Son:

Corte: Manual
Madurante: ROUNDUP SL
Textura del suelo: Franco Arcilloso
Madurez: Intermedia
Variedad: CP-20-86

De acuerdo a la información disponible se plantea el modelo con todas

las covariables y factores.

$$\begin{aligned} Y_1 =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \\ & \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \alpha_1 z_1 + \alpha_2 z_2 + \\ & \alpha_3 z_3 + \alpha_4 z_4 + \alpha_5 z_5 \end{aligned} \tag{3.1}$$

Donde: Y_1: kg de azúcar/t de caña.
x_1: Edad.
x_2: Número de cortes.
x_3: Humedad.
x_4: Altura.
x_5: Lluvia.
x_6: Amplitud térmica.
x_7: Temperatura.
z_1: Textura del suelo.
z_2: Tipo de madurante.
z_3: Tipo de corte.
z_4: Variedad del lote.
z_5: Madurez.

Notese que las variables x_i son cuantitativas y las variables z_i, son variables cualitativas como las presentadas en la sección 2.2.1, cada una con sus respectivas categorías, indicando la presencia o ausencia de una cualidad. Finalmente α_i es el efecto una determinada característica de z_i sobre la variable respuesta.

Primero se estima el modelo (3.1), con un nivel de significancia del $5\,\%$. La Tabla 3.4, es la tabla ANOVA para el modelo 3.1, donde se comprueba si las variables son influyentes y si diferencias observadas en los gráficos anteriores son significativas, resultando no

significativas la Edad, Número de cortes, Lluvia, y Etapa de madurez.

Respuesta: kg de azúcar/ t de caña						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Edad meses	1	11143	11143	769.6424	< 2.2e-16	***
Ncortes	1	39	39	2.7174	0.09946	.
Humedad	1	91018	91018	6286.5955	< 2.2e-16	***
Altura	1	1618	1618	111.787	< 2.2e-16	***
Lluvia	1	232	232	16.0116	0.901059	
Amplitud térmica	1	638	638	44.0884	4.29e-11	***
Temperatura	1	363	363	25.08	6.11e-07	***
Textura del suelo	5	435	87	6.0071	1.63e-05	***
Tipo de madurante	4	1042	261	17.9803	1.87e-14	***
Tipo de corte	1	545	545	37.6226	1.08e-09	***
Variedad del lote	8	2074	259	17.8856	< 2.2e-16	***
Madurez	3	79	26	1.8083	0.14366	
Residuales	1602	23218	14			
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

Tabla 3.4: ANOVA para el modelo propuesto

Se estima el nuevo modelo usando solo las variables significativas y usando el AIC definido en la sección 2.1.9. La ecuación (3.2) representa el modelo estimado con las variables significativas. La salida de la estimación de este modelo se presenta en la Tabla B.8.

$$\begin{aligned} Y_1 \;=\; & 520.4 - 5.93x_3 + 0.0026x_4 + 0.424x_6 - 0.182x_7 - \\ & 0.811z_1Fco - 0.979z_1FcoFcoArenoso - \\ & 1.24z_1FcoLimoso - 2.11z_1Otro - 1.53z_2No - \\ & 1.69z_3Mecanizada - 2.32z_4(CP-88-1165) + \\ & 2.86z_4(CP-89-2143) - 1.95z_4(MEX-79-431) - \\ & 1.63z_4(PR-87-2080) \end{aligned} \tag{3.2}$$

Entre paréntesis están los nombres de las variedades.

La Tabla 3.5, es la tabla ANOVA para el modelo (3.2), que contiene solo variables significativas.

Respuesta: kg de azúcar/ t de caña						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Humedad	1	101939	101939	7009.69	< 2.2e-16	***
Altura	1	1682	1682	115.656	< 2.2e-16	***
Amplitud térmica	1	765	765	52.6952	6.04e-13	***
Temperatura	1	516	516	35.4621	3.19e-09	***
Textura de suelo	5	442	88	6.0778	1.39e-05	***
Tipo de madurante	4	1099	275	18.8996	3.37e-15	***
Tipo de corte	1	533	533	36.6235	1.78e-09	***
Variedad del lote	8	2085	261	17.9183	< 2.2e-16	***
Residuales	1608	23385	15			
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 ' '	1

Tabla 3.5: ANOVA para el modelo estimado

La Tabla 3.6, muestra los AIC estimados, observandose que el mejor modelo es el que incluye todas las variables significativas.

	Df	Sum of Sq	RSS	AIC
<none>			23385	4389.2
-Altura	1	255	23639	4404.8
-Testura de suelo	5	373	23758	4405
-Temperatura	1	306	23691	4410
-Tipo de corte	1	466	23850	4420
-Tipo de madurante	4	924	24308	4440
-Amplitud térmica	1	971	24355	4453.5
-Variedad del lote	8	2090	25469	4512.4
-Humedad	1	78770	102111	6791.2

Tabla 3.6: AIC para el modelo estimado

Este modelo explica el $82.34\,\%$ de la variabilidad total de los datos con un coeficiente de determinación ajustado del $82.1\,\%$, por lo que puede deducirse que es un modelo bastante bueno y será sometido a prueba en las siguientes secciones para comprobar si es adecuado.

3.2.3. Interpretación de los resultados

Variables Cuantitativas

La Humedad presenta un coeficiente negativo e indica que un incremento unitario de ésta, manteniendo las demás variables constantes, el rendimiento de azúcar disminuye en $5.93\ kg$ por tonelada de caña. Si la Altura incrementa en una unidad, el rendimiento de azúcar incrementa en promedio $0.0026\ kg$ por tonelada de caña. Por cada incremento unitario en la Amplitud térmica, puede esperarse un incremento de $0.42\ kg$ en el rendimiento de azúcar. Si la temperatura media a la que se expone la caña durante su ciclo de vida incrementa una unidad, se espera una reducción de $0.18\ kg$ de azúcar por tonelada de caña.

Variables cualitativas

- **Textura del suelo**

  La categoría que se ha tomado como referencia es la textura Franco Arcilloso y es contra esta que se hacen las comparaciones. En el modelo solo se incluyen las texturas que presentan diferencias significativas con la textura de referencia. Todos los coeficientes de las texturas son negativos por lo que puede deducirse que una caña cosechada en la textura de referencia tiene un rendimiento medio mayor que la cosechada en otra textura. Por ejemplo: La caña cosechada de un suelo con textura Franco Franco Arenoso se espera que tenga un rendimiento de $0.979\ kg/t$ menos que la cosechada en un suelo con textura Franco Arcilloso.

- **Tipo de madurante**

  Para el madurante se tomó como referencia el ROUNDUP SL. Aquellos lotes a los que se les ha aplicado madurante OTRO, presentan un rendmiento medio de $2.11\ kg/t$ menos que a loa que se les ha aplicado ROUNDUO SL. Otra diferencia significativa se da con aquellos lotes a los que no se les ha aplicado madurante, es decir, los lotes en los que no se aplica madurante presentan un rendimiento medio de $1.53\ kg/t$ menos que aquellos a los que se les ha aplicado RPUNDUP SL.

- **Tipo de corte**

  Esta variable solo tiene dos categorías (Manual y mecanizada), y se ha tomado como referencia aquellos lotes que se han cosechado de manera manual. Un lote que se cosecha de manera mecanizada tiene un rendimiento medio de $1.63\ kg/t$ menos que uno cosechado de manera manual.

- **Variedad del lote**

  La variedad que se ha tomado como referencia es la CP-72-20-86. En el modelo se incluyen solo aquellas variedades que presentan diferencias significativas con esta. Puede notarse que los lotes con la variedad de referencia presentan rendimientos medios mayores que los lotes en los que se cosecha otra variedad, excepto los lotes donde se cosecha la variedad CP-89-2143, ya que estos lotes tienen un rendimiento medio de $2.86\ kg/t$ más con respecto a la variedad de referencia. Los lotes con variedades MEX-79-431 y PR-87-2080, presentan un rendimiento de $1.95\ kg/t$ y $1.63\ kg/t$ respectivamente, menor que los lotes con variedad CP-2086.

Una vez estimado el modelo es necesario identificar las variables más influyentes. Para ello se usan los estimadores estandarizados, calculados de la forma descrita en la sección 2.1.1, los cuales indican la influencia de cada variable: A mayor valor del estimador, mayor influencia de la variable sobre la variable respuesta, estos estimadores se presentan en la Tabla 3.7. También puede determinarse la influencia de la variable usando el p-valor del contraste de significatividad. Cuanto más pequeño sea el p-valor más significativa es la variable. Para el modelo (3.2), la variable más influyente es la Humedad, por lo que debe tenerse mucho cuidado con esta variable y buscar la manera de controlarla para reducir su efecto negativo, la segunda variable más influyente es la Amplitud térmica y esta depende de las condiciones climáticas que se presenten.

Estimadores estandarizados	
Variable	**Valor del estimador**
Humedad	-0.8672
Amplitud térmica	0.117
Altura	0.061
Temperatura	-0.064

Tabla 3.7: Estimadores estandarizados

3.2.4. Diagnosis del modelo

Normalidad de los residuos

Se analizan los residuos estandarizados. En la Figura 3.12 se presenta el histograma con la línea de densidad normal para tener una visión de la normalidad de los residuos obtenidos con el modelo seleccionado.

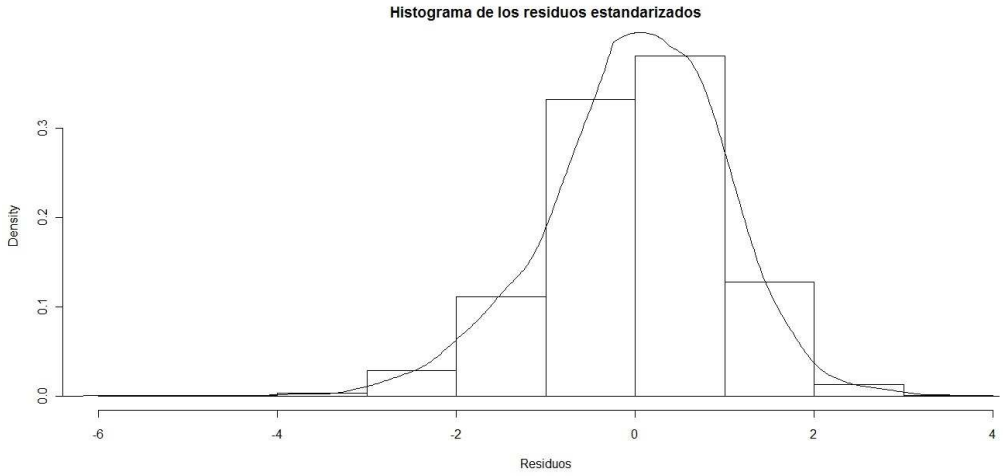


Figura 3.12: Histograma de los residuos estandarizados

El histograma indica una posible normalidad de los residuos, lo que se refuerza con el gráfico “qqplot” presentado en la Figura 3.13.

Puede notarse que los residuos tienen un comportamiento aproximadamente normal. Para reforzar estos resultados se realiza el test de tets de Kolmogorov Smirnov obteniendose:

D = 0.0293, p-value = 0.1215

con lo que puede concluirse que los residuos son normales.

Homocedasticidad de los residuos

Se presenta el diagrama de dispersión de residuos versus observaciones predichas, este gráfico no debe presentar ningún tipo de patrón o tendencia.

En la Figura 3.14 puede notarse que la variabilidad de los residuos permanece constante, por lo que puede decirse que los

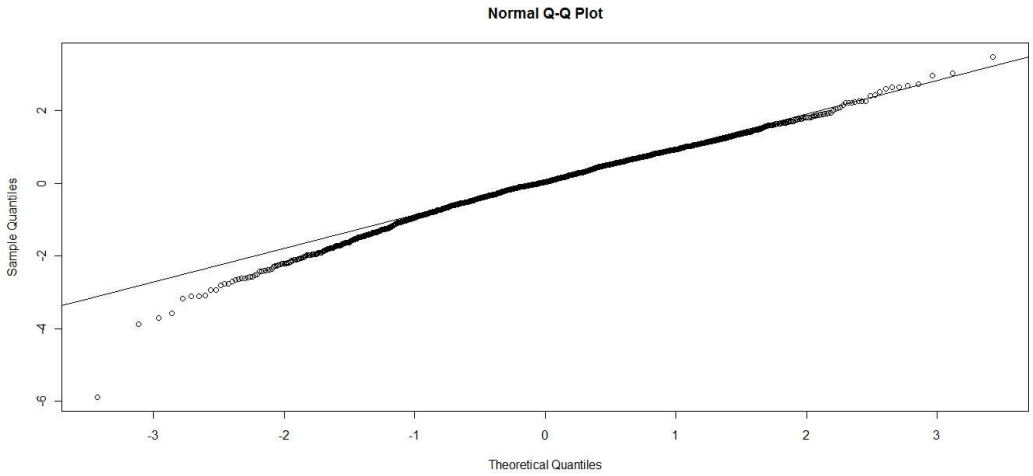


Figura 3.13: QQPLOT de los residuos estandarizados

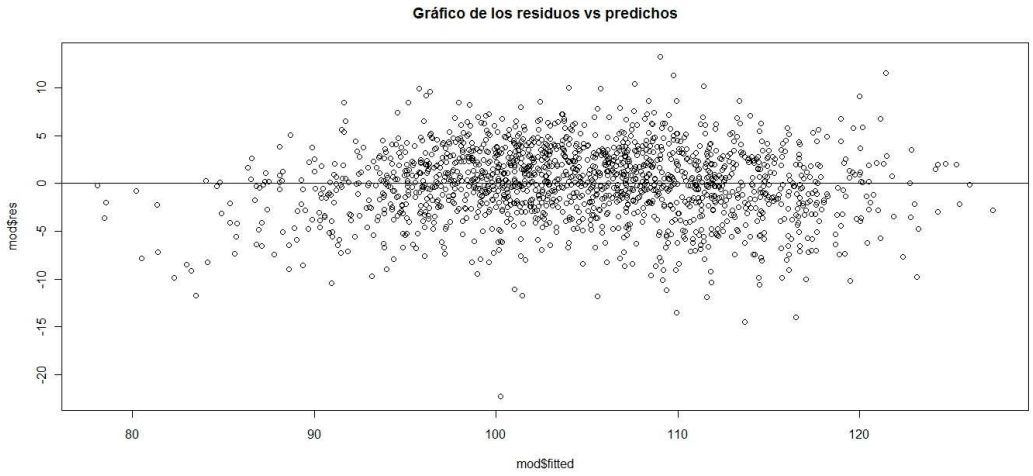


Figura 3.14: Gráfico de los residuos versus predichos

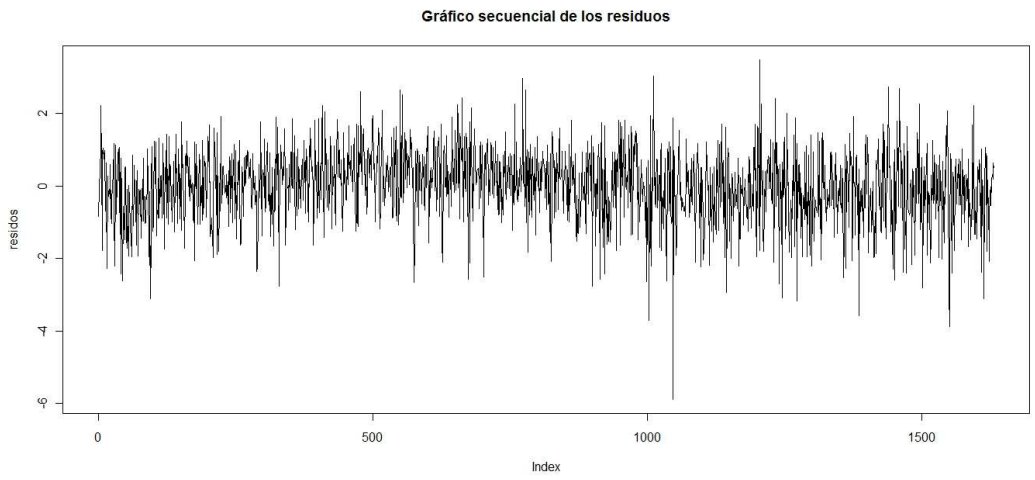


Figura 3.15: Gráfico secuencial de los residuos

residuos son homocedásticos o de varianza constante.

Independencia de los residuos

Para observar si existe relación entre los residuos se hace el gráfico secuencial de los residuos y así detectar alguna tendencia.

En el gráfico 3.15, puede observarse que no existe tendencia por lo que podemos decir que los residuos son independientes. Esto se refuerza con el test de Durbin-Watson con el que se obteienen los siguientes resultados:

DW = 2.0514, p-value = 0.2986

con los resultados anteriores podemos concluir que los residuos son independientes.

3.2.5. Validación del modelo: Validación cruzada

En esta fase se procede según lo descrito en la sección 2.1.13, primero seleccionando dos muestras aleatorias de los datos, estimando un modelo para cada una y usando el estadístico (2.19) y luego usando el (2.21), para comprobar la validez predictiva del modelo (3.2).

- **Método 1**

  $$\frac{(SSE_T - SSE_1 - SSE_2)/p}{(SSE_1 + SSE_2)/(n_1 + n_2 - 2p)} \sim F_{(p,n_1+n_2-2p)} = 1.006534$$

  donde $n = 1631$, $p = 26$, y el p-valor correspondiente es 0.4548, el cual es mayor que 0.05, por lo que no se rechaza la hipótesis de que los modelos son iguales.

- **Método 2**

  $$B^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i^{(i)})^2} = 0.9662$$

  El valor de B^2 es aproximadamente de 0.97 que es muy cercano a 1 indicando que se tienen modelos muy parecidos.

De acuerdo a los resultados obtenidos con ambos métodos, el ajuste lo podemos catalogar como robusto debido a que no se evidencian variaciones relevantes entre los ajustes obtenidos con todos los datos y otros conseguidos con sólo una parte de ellos.

Se ha estimado un modelo sin interacciones y se ha comprobado que es un modelo bastante bueno y el incluir interacciones no mejora el modelo, ya que se logra explicar únicamente un $1\,\%$

más de la variabilidad de los datos, es decir, se obtiene un R^2 de aproximadamente $83\,\%$, obteniendo muchos más coeficientes y por tanto un modelo complicado.

En las tabalas B.9, B.10, B.11 y B.12, se muestran las comparaciones múltiples para cada factor respectivamante, donde se obtienen resultados parecidos a los obtenidos en el modelo (3.2), esto significa que el modelo representa adecuadamente las diferencias que existen en el rendimiento de azúcar entre las categorías de cada factor.

3.3. Modelo para el rendimiento de campo o rendimiento de caña (t de caña/ha)

3.3.1. Estimación y selección del modelo

Primero se realiza un análisis de correlación entre las variables continuas para identificar los posibles tipos de relación y la presencia de colinealidad entre las variables.

	t caña/ha	Humedad	Ncortes	Edad	Altura	Lluvia	Amp térmica	Temp
t caña/ha	1	0.3874	0.1273	0.1318	-0.0463	-0.0499	-0.0652	-0.0489
Humedad	0.3874	1	0.0744	-0.3134	-0.1771	-0.1678	0.0243	0.1895
Ncortes	0.1273	0.0744	1	-0.0303	0.0208	0.0395	0.0102	-0.0335
Edad	0.1318	-0.3134	-0.0303	1	0.0953	0.1011	-0.0543	-0.3098
Altura	-0.0463	-0.1771	0.0208	0.0953	1	0.3998	0.3092	-0.3577
Lluvia	-0.0499	-0.1678	0.0395	0.1011	0.3998	1	0.3152	-0.4629
Amp térmica	-0.0652	0.0243	0.0102	-0.0543	0.3092	0.3152	1	0.2626
Temp	-0.0489	0.1895	-0.0335	-0.3098	-0.3577	-0.4629	0.2626	1

Tabla 3.8: Correlaciones para el rendimiento de caña

En la Tabla 3.8, se observa una baja correlación entre las variables. La correlación mayor se tiene entre las Toneladas de caña por hectárea (t/ha) y la Humedad. Esta relación implica que puede

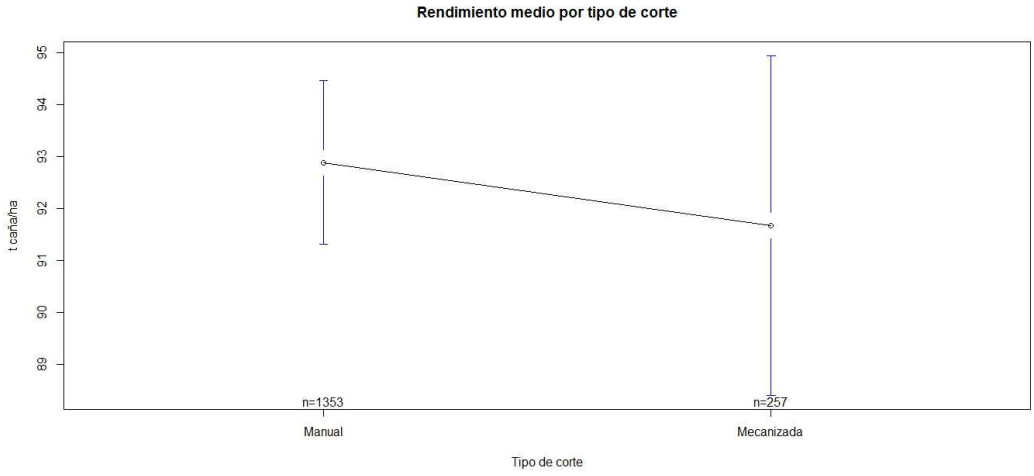


Figura 3.16: Rendimiento medio por tipo de corte

esperarse que la Humedad tenga un coeficiente positivo. Con las demás variables, el rendimiento de caña presenta una relación baja, pero los conocimientos teóricos indican que éstas tienen influencia en el rendimiento y por tanto serán incluidas en el análisis. La relación entre las variables explicativas es baja, por lo que puede descartarse la existencia de colinealidad.

3.3.2. Gráficos de medias para el rendimiento de caña

Se presentan los gráficos para el rendimiento medio con un intervalo de confianza del $95\,\%$, por categoría para cada factor para poder identificar posibles diferencias entre estos.

- **Rendimiento medio por tipo de corte**
  La Figura 3.16 muestra para cada tipo de corte el rendimiento medio de caña y su correspondiente intervalo de confianza.

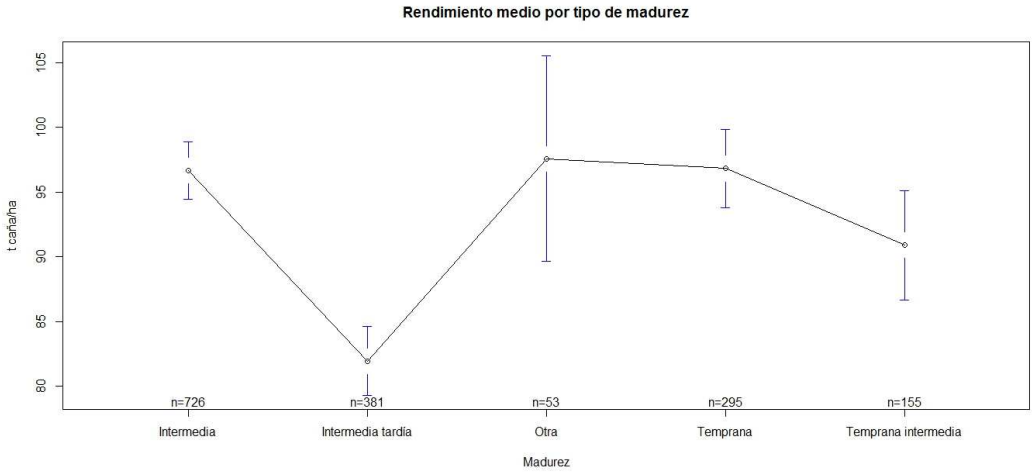


Figura 3.17: Rendimiento medio por tipo de madurez

Se observa una clara diferencia en el rendimiento medio obtenido de manera manual con el obtenido de manera mecanizada, donde la caña que ha sido cosechada de manera manual presenta un mayor rendimiento con una menor variabilidad.

- **Rendimiento medio por etapa de madurez**

  La Figura 3.17 muestra para cada tipo de madurez de la caña, el rendimiento medio de caña y su correspondiente intervalo de confianza. Se observa que los lotes que tienen caña con madurez intermedia tardía es la que presenta el menor rendimiento medio, entre los demás tipos de madurez se observa poca diferencia en el rendimiento medio.

- **Rendimiento medio por madurante**

  La Figura 3.18 muestra para cada tipo de madurante el rendimiento medio de caña y su correspondiente intervalo de confianza.

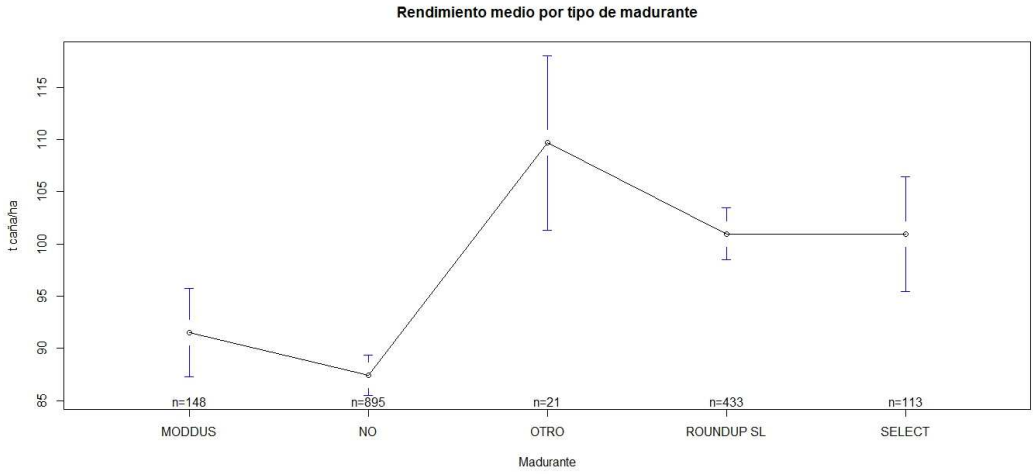


Figura 3.18: Rendimiento medio por tipo de madurante

Puede notarse que existe diferencia en los rendimientos medios por madurante, el madurante con los mayores rendimientos es OTRO. Sólo ha sido aplicado en 21 lotes de la muestra y presentan una alta variabilidad. Los lotes a los que se les ha aplicado ROUNDUP SL presentan rendimientos medios altos y baja variabilidad, es decir, en estos lotes se puede esperar tener rendimientos medios muy parecidos.

- **Rendimiento medio por textura del suelo**

  La Figura 3.19 muestra para cada textura de suelo el rendimiento medio de caña y su correspondiente intervalo de confianza. Al observar los rendimientos medios puede notarse que se presentan diferencias de acuerdo a la textura del suelo. El mayor rendimiento se presenta en los lotes con textura Franco limoso, que es la segunda textura con mayor frecuencia y donde se presenta poca variabilidad. La textura más frecuente es la Franco arcilloso y esta presenta la menor variabilidad.

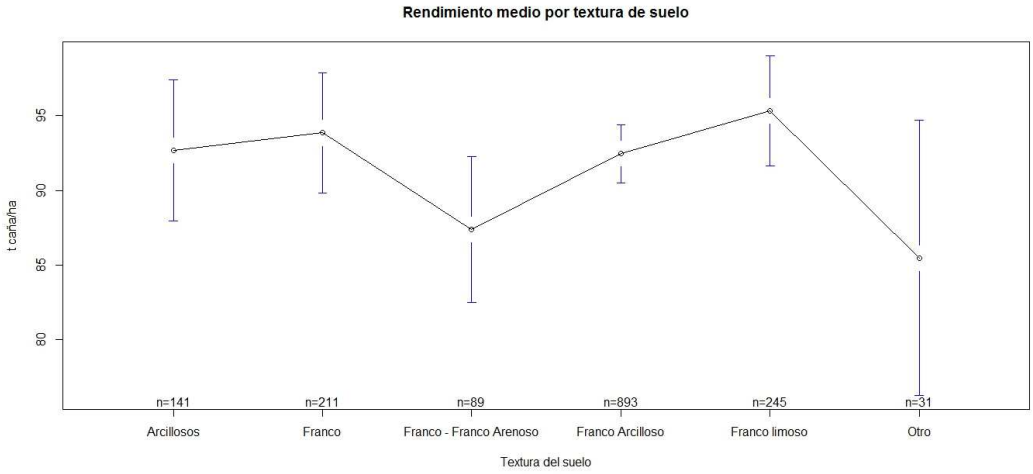


Figura 3.19: Rendimiento medio por textura de suelo

- **Rendimiento medio por variedad**

  La Figura 3.20 muestra para cada tipo de corte el rendimiento medio de caña y su correspondiente intervalo de confianza.

  En el gráfico puede notarse que existen diferencias entre los rendimientos medios obtenidos para cada variedad. La variedad con un mayor rendimiento es la CP-73-1547 y la segunda con mayores rendimientos es la CP-72-2086. En ambas variedades se tiene poca variabilidad mientras que la variedad con mayor variabilidad y menor rendimiento es la PR-87-2080.

Para hacer la estimación del modelo se eligen las categorías de referencia de cada variable cualitativa, estas son las categorías que presentan una mayor frecuencia. Son:

Corte: Manual
Madurante: ROUNDUP SL
Textura del suelo: Franco Arcilloso

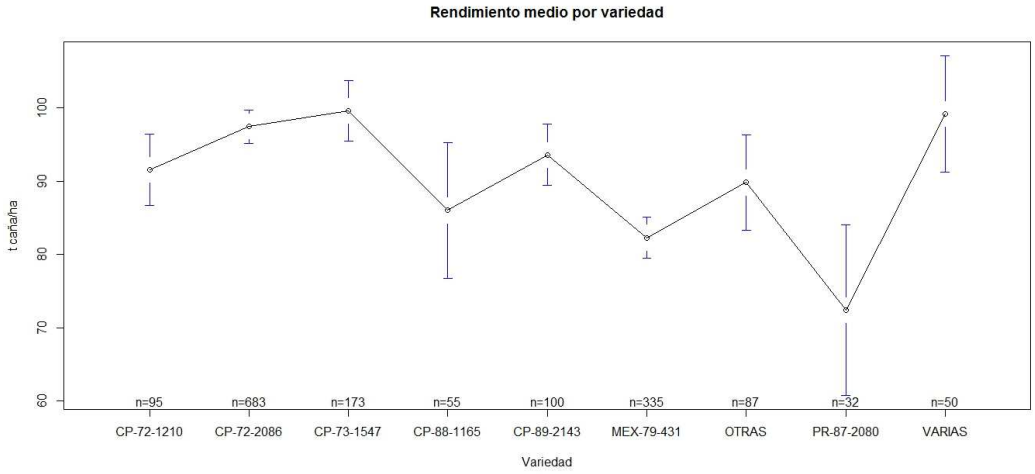


Figura 3.20: Rendimiento medio por tipo de corte

Madurez: Intermedia

Variedad: CP-20-86

De acuerdo a la información disponible se plantea el modelo con todas las covariables y factores.

$$\begin{aligned} Y_2 =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \\ & \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \alpha_1 z_1 + \alpha_2 z_2 + \\ & \alpha_3 z_3 + \alpha_4 z_4 + \alpha_5 z_5 \end{aligned} \tag{3.3}$$

Donde:

Y_2: Toneladas de caña por hectárea (t/ha).

x_1: Edad.

x_2: Número de cortes.

x_3: Humedad.

x_4: Altura.

x_5: Lluvia.

x_6: Amplitud térmica.
x_7: Temperatura.
z_1: Textura del suelo.
z_2: Tipo de madurante.
z_3: Tipo de corte.
z_4: Variedad del lote.
z_5: Madurez.

Notese que las variables x_i son variables continuas y las variables z_i, son variables cualitativas como las presentadas en la sección 2.2.1, cada una con sus respectivas categorías, las cuales indican la presencia o ausencia de una cualidad y α_i es el efecto una determinada característica de z_i sobre la variable respuesta.

Primero se estima el modelo (3.3), con un nivel de significancia del $5\,\%$, resultando no significativas la Altura, Lluvia, Temperatura, Textura del suelo y Etapa de madurez, observando los resultados en la Tabla 3.9, donde se comprueba si las variables son influyentes y si diferencias observadas en los gráficos anteriores son significativas.

Se estima el nuevo modelo usando solo las variables significativas y usando el AIC definido en la sección 2.1.9. La salida de la estimación de este modelo se presenta en la Tabla B.13.

$$\begin{aligned} Y_2 \;=\; & -625.114 + 6.4852x_1 + 0.621x_2 + 9.4011x_3 - \\ & 0.868x_4 - 7.3831z_2No - 4.848z_3Mecanizada - \\ & 10.05z_4(MEX - 79 - 431) - 15.378z_4(PR - 87 - 2080) \end{aligned} \tag{3.4}$$

Respuesta: toneladas de caña por hectárea						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Edad meses	1	23686	23686	38.2541	7.89e-10	***
Ncortes	1	23515	23515	37.979	9.06e-10	***
Humedad	1	268003	268003	432.8417	< 2.2e-16	***
Altura	1	84	84	0.1354	0.712908	
Lluvia	1	74	74	0.1197	0.729372	
Amplitud térmica	1	6449	6449	10.4151	0.001276	**
Temperatura	1	682	682	1.1009	0.294221	
Textura del suelo	5	6713	1343	2.1683	0.075182	.
Tipo madurante	4	19690	4923	7.9503	2.41e-06	***
Tipo de corte	1	3170	3170	5.1193	0.023796	*
Variedad del lote	8	27874	3484	5.6272	4.63e-07	***
Madurez	3	3641	1214	1.9601	0.118041	
Residuales	1578	977440	619			
Signif. codes:	0 ’***’	0.001 ’**’	0.01 ’*’	0.05 ’.’	0.1 ’ ’	1

Tabla 3.9: ANOVA para el modelo propuesto

Entre paréntesis están los nombres de las variedades. El modelo (3.4) contiene únicamente covariables y categorías que son significativas.

Se presenta la tabla ANOVA del modelo para observar la significancia de las variables en la Tabla 3.10, donde sólo aparecen variables significativas.

Respuesta: toneladas de caña por hectárea						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Edad meses	1	23686	23686	38.044	8.75e-10	***
Ncortes	1	23515	23515	37.7705	1.00e-09	***
Humedad	1	268003	268003	430.4651	< 2.2e-16	***
Amplitud térmica	1	5314	5314	8.5347	0.003533	**
Tipo madurante	4	14888	3722	5.9782	8.96e-05	***
Tipo de corte	1	6354	6354	10.2055	0.001428	**
Variedad del lote	8	29567	3696	5.9363	1.61e-07	***
Residuales	1592	991162	623			
Signif. codes:	0 ’***’	0.001 ’**’	0.01 ’*’	0.05 ’.’	0.1 ’ ’	1

Tabla 3.10: ANOVA para el modelo estimado

La Tabla 3.11, muestra los AIC estimados y se observa que el mejor modelo es el que incluye todas las variables significativas.

	Df	Sum of Sq	RSS	AIC
<none>			991162	10376
-Número de cortes	1	4068	995485	10381
-Tipo de corte	1	4323	995485	10382
-Amplitud térmica	1	6611	997773	10385
-Tipo de madurante	4	16332	1007494	10395
-Variedad del lote	8	29567	1020729	10408
-Edad	1	76567	1067729	10494
-Humedad	1	193607	1184769	10662

Tabla 3.11: AIC para el modelo estimado

Este modelo explica el $27.25\,\%$ de la variabilidad total de los datos con un coeficiente de determinación ajustado del $26.48\,\%$, indicando que las variables consideradas en el análisis no logran predecir de forma adecuada las toneladas de caña por hectárea (t/ha), por lo que puede decirse que este es un modelo explicativo y no predictivo. Lo anterior parece indicar que existen otras variables importantes que no han sido consideradas en el análisis.

3.3.3. Interpretación de los resultados

Variables Cuantitativas

La Edad presenta un coeficiente positivo e indica que un incremento unitario de ésta, manteniendo las demás variables constantes, el rendimiento de caña incrementa en promedio $6.48\ t/ha$, sin que la edad exceda los límites permitidos ya que una caña muy vieja puede perder peso. Si el número de cortes incrementa una unidad, el rendimiento de caña incrementa en promedio $0.62\ t/ha$, sabiendo que no

se deben realizar más de 10 cortes. Por cada incremento unitario en la Humedad, el rendimiento de caña aumenta un promedio de 9.4 t/ha. Sin embargo debe tenerse mucho cuidado con esta variable, ya que su incremento impacta de forma negativa al rendimiento de azúcar. Si la Amplitud térmica tiene un incremento unitario, el rendimiento de caña tiene una reducción promedio de $0.87\ t/ha$.

Variables cualitativas

- **Tipo de madurante**
  Para el madurante se tomó como referencia el ROUNDUP SL. La única diferencia significativa se da con aquellos lotes a los que no se les ha aplicado madurante, es decir, los lotes en los que no se aplica madurante presentan un rendimiento medio de $7.4\ t/ha$ menos que aquellos a los que se les ha aplicado ROUNDUP SL.

- **Tipo de corte**
  Se han tomado como referencia los lotes que se han cosechado de manera manual. Un lote que se cosecha de manera mecanizada tiene un rendimiento medio de $4.85\ t/ha$ menos que uno cosechado de manera manual.

- **Variedad del lote**
  La variedad que se ha tomado como referencia es la CP-72-20-86, y puede notarse que los lotes con la variedad MEX-79-431 presentan un rendimiento medio de $10.05\ t/ha$ menos que los lotes con la variedad de referencia, y los lotes con PR-87-2080 presentan un rendimiento medio de $15.4\ t/ha$ menos que los

lotes con la variedad de referencia.

Una vez estimado el modelo es necesario identificar las variables más influyentes y para ello se usan los estimadores estandarizados, calculados de la forma descrita en la sección 2.1.1, los cuales indican la influencia de cada variable: A mayor valor del estimador, mayor influencia de la variable sobre la variable respuesta. Estos estimadores se presentan en la Tabla 3.12. También puede determinarse la influencia de la variable usando el p-valor del contraste de significatividad. Cuanto más pequeño sea el p-valor más significativa es la variable. Para el modelo (3.4), las variables más influyentes son Humedad y Edad por lo que debe tenerse mucho cuidado con estas variables para tener un buen rendimiento.

Estimadores estandarizados	
Variable	**Valor del estimador**
Humedad	0.4266
Edad	0.2547
Número de cortes	0.0584
Amplitud térmica	-0.0744

Tabla 3.12: Estimadores estandarizados

3.3.4. Diagnosis del modelo

Normalidad de los residuos

En la Figura 3.21, se presenta el histograma con la línea de densidad normal para tener una visión de la normalidad de los residuos obtenidos con el modelo seleccionado. El histograma indica una posible normalidad de los residuos, lo que se refuerza con el

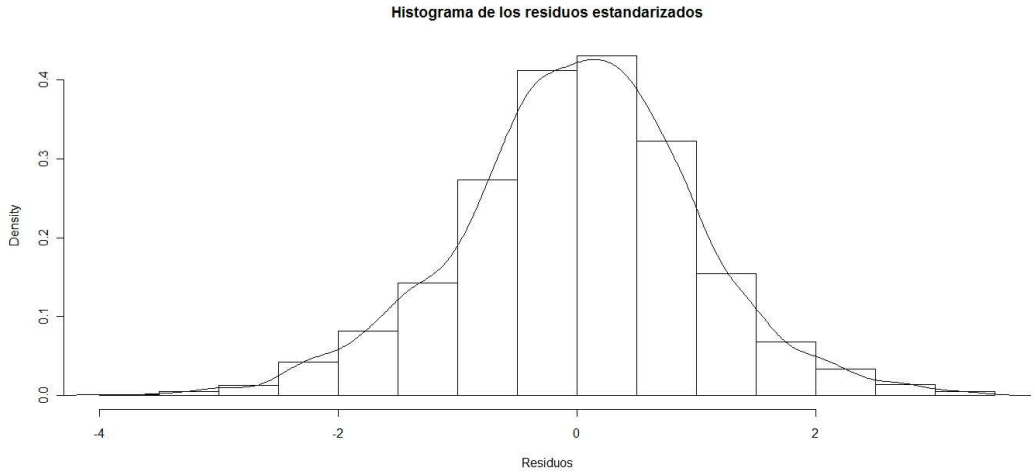


Figura 3.21: Histograma de los residuos estandarizados

gráfico “qqplot” presentado en la Figura 3.22. De acuerdo a los gráficos, puede notarse que los residuos tienen un comportamiento aproximadamente normal. Estos resultados se refuerzan usando el tets de Kolmogorov Smirnov:

D = 0.026, p-value = 0.2189

Los residuos presentan una distribución normal.

Homocedasticidad de los residuos

Se presenta en la Figura 3.23, el diagrama de dispersión de residuos versus observaciones predichas. Este gráfico no debe presentar ningún tipo de patrón o tendencia para aceptar varianza constante en los residuos.

En la Figura 3.23 puede notarse que la variabilidad de los residuos permanece constante, por lo que puede decirse que los residuos son homocedásticos o de varianza constante. En el gráfico puede no-

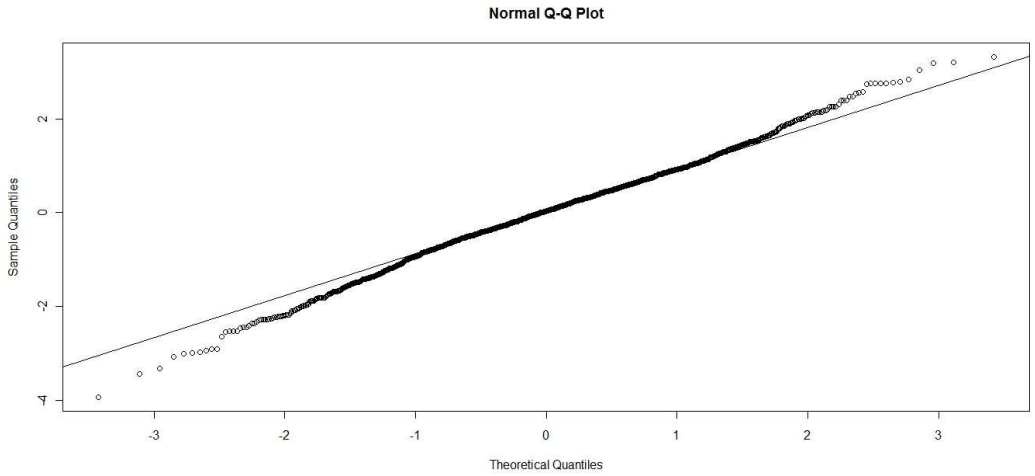


Figura 3.22: QQPLOT de los residuos estandarizados

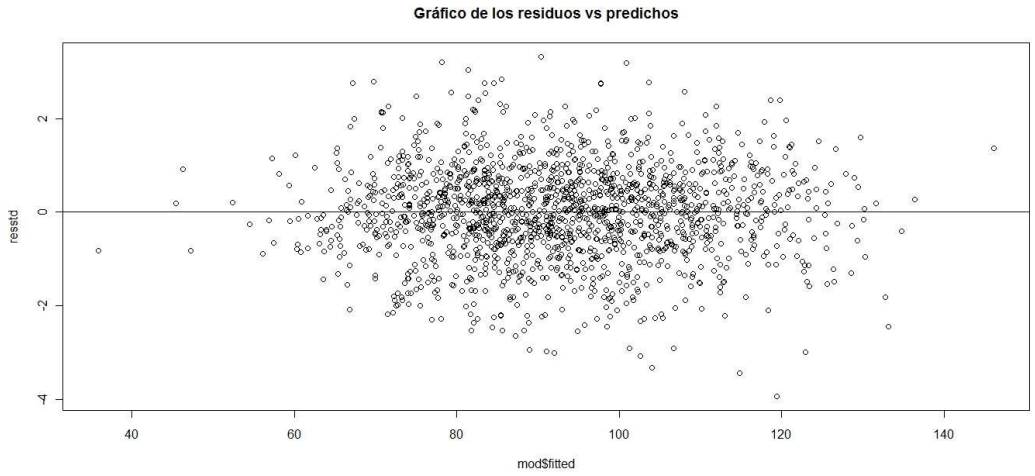


Figura 3.23: Gráfico de los residuos versus predichos

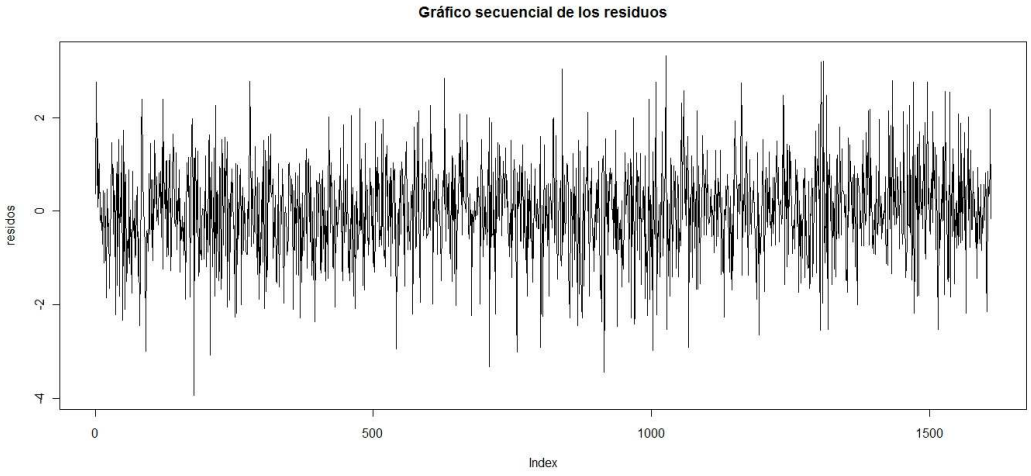


Figura 3.24: Gráfico seciencial de los residuos

tarse que la variabilidad de los residuos permanece constante, por lo que los residuos son homocedásticos o de varianza constante.

Independencia de los residuos

Para observar si existe relación entre los residuos se hace el gráfico secuencial de los residuos presentado en la Figura 3.24.

En la Figura 3.24, puede observarse que no existe tendencia por lo que podemos decir que los residuos son independientes. Esta independencia se comprueba con el test de Durbin-Watson:

DW = 2.0453, p-value = 0.3608

De acuerdo a los resultados puede concluirse que losresiduos son independientes.

3.3.5. Validación del modelo: Validación cruzada

De acuerdo a los resultados obtenidos, no se ha encontrado un modelo para predecir el tonelaje de caña, sino solo para explicarlo, por lo tanto no se hace la validación cruzada.

Se ha estimado un modelo sin interacciones y se ha comprobado que es un modelo explicatico y el incluir interacciones no mejora el modelo, ya que se logra explicar únicamente un $2\,\%$ más de la variabilidad de los datos, es decir, se obtiene un R^2 de aproximadamente $29\,\%$, obteniendo muchos más coeficientes y por tanto un modelo complicado.

En las tabalas B.14, B.15 y B.16 se muestran las comparaciones múltiples para cada factor respectivamante, donde se obtienen resultados parecidos a los obtenidos en el modelo (3.4), esto significa que el modelo representa adecuadamente las diferencias que existen en el rendimiento de azúcar entre las categorías de cada factor.

Conclusiones y recomendaciones

Conclusiones

- El modelo estimado para el rendimiento de azúcar explica el $82.3\,\%$ de la variabilidad de los datos y se ha comprobado que es coherente y puede ser usado para predecir el rendimiento de azúcar.
- El modelo estimado para el rendimiento de caña explica el $27.2\,\%$ de la variabilidad de los datos este modelo puede ser

usado para explicar el rendimiento de caña.

- La variable que más influye en los rendimientos es la Humedad, por lo que deben buscarse los medios adecuados para poder controlarla.
- La variedad con el mayor rendimiento de azúcar es la CP-89-2143, y la variedad con el mayor rendmiento de caña es la CP-73-1547, por lo que deben analizarse las necesidades y características de estas variedades para poder aprovecharlas al máximo.
- Si se aplica madurante, se recomienda aplicar MODUS para un buen rendimiento de azúcar y ROUNDUP para un buen rendimiento de caña.

Recomendaciones

- Para tener mejores modelos para ambos rendimientos, debe tenerse mayor control sobre los registros de las variables y factores que influyen en los rendimientos, para que esto facilite la ejecución de investigaciones de interés.
- Para tener un buen rendimiento de azúcar y de caña, debe implementarse y mantenerse un balance entre el tipo de corte de la caña.
- Deben hacerse análisis previos a la cosecha de la caña que permitan determinar la humedad que esta tiene ya que la humedad a pesar de tener un impacto positivo en el rendimiento

de caña, tiene un impacto negativo muy fuerte en el rendimiento de azúcar.

- Antes de cultivar una variedad, debe conocerse con exactitud su periodo de madurez para determinar la edad adecuada de cosecha.
- Para el cultivo de nuevos lotes es necesario determinar la textura del suelo y de preferencia esta debe ser Franco Arcilloso.
- Este estudio puede ser tomado como base para estudios posteriores en los que se cuente con mayor información.
- Las variables ambientales no son controlables, pero sí se pueden monitorear. Es posible controlar variables y factores agrícolas, por lo que el obtener mejores rendimientos es una tarea factible.
- Los ingenios y cañeros deben trabajar de la mano y buscar mejores rendimientos para obterner un beneficio común.

Apéndice A

Capítulo 2

A.1. Desarrollo de las ecuaciones.

$$\hat{\beta}' X' Y = [\hat{\beta}_0 \ \hat{\beta}_1 \ \cdots \ \hat{\beta}_k] \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$= \left[[\hat{\beta}_0 + \sum \hat{\beta}_j x_{1j}] y_1 \quad [\hat{\beta}_0 + \sum \hat{\beta}_j x_{2j}] y_2 \quad \cdots \quad [\hat{\beta}_0 + \sum \hat{\beta}_j x_{nj}] y_n \right]$$

$$= \left[\hat{\beta}_0 \sum y_i + \sum \sum y_i \hat{\beta}_j x_{ij} \right]$$

también se tiene que:

$$(\hat{\beta}' X' Y)' = Y' X \hat{\beta}$$

por tanto:

$$Y' X \hat{\beta} = \hat{\beta}' X' Y = \left[\hat{\beta}_0 \sum y_i + \sum \sum y_i \hat{\beta}_j x_{ij} \right]$$

Desarrollo de la matriz C.

$$
\begin{aligned}
C &= (X'X)^{-1}X' \\
C' &= [(X'X)^{-1}X']' \\
&= X[(X'X)^{-1}]' \\
&= X[(X'X)']^{-1} \\
&= X[X'X]^{-1} \\
CC' &= (X'X)^{-1}X'X[X'X]^{-1} \\
&= (X'X)^{-1}
\end{aligned}
$$

A.2. Datos para el ejemplo ilustrativo.

Salarios	Estudio	Sexo	Procedencia	Gasto	Consumo
200	Primaria	M	Putumayo	125	50
205	Primaria	M	Nariño	130	65
220	Secundaria	M	Cauca	140	80
228	Secundaria	F	Nariño	142	86
252	Técnica	F	Nariño	162	90
264	Técnica	M	Putumayo	172	90
272	Técnica	M	Putumayo	200	70
315	Pregrado	F	Cauca	215	80
324	Pregrado	F	Putumayo	225	80
340	Pregrado	M	Putumayo	400	100
618	Postgrado	F	Nariño	325	130
720	Postgrado	M	Nariño	360	140
800	Postgrado	F	Cauca	380	160

Tabla A.1: Datos para el ejemplo ilustrativo

A.3. Script para el ejemplo ilustrativo.

```
##################################
#LEEMOS LOS DATOS Ingresos←read.table(“Ingresos1.csv”,header=TRUE,sep
= “,”)
Ingresos

#NOMBRES DE LAS VARIABLES
names(Ingresos)

#ACCEDIENDO A LOS DATOS PARA USARLOS DIRECTA-
MENTE
attach(Ingresos)

#VERIFICANDO QUE SE ESTÁ TRABAJANDO CON FACTO-
RES
is.factor(Sexo)
is.factor(Estudio)
is.factor(Procedencia)


##################################
#MODELO
#SELECCIONANDO LA CATEGORÍA DE EFERENCIA
Estudio1 ← relevel(Estudio, ”Postgrado”)
Estudio1

#ESTIMACIÓN DEL MODELO
modelo←lm(Salarios Estudio1+Gasto+Consumo)
summary(modelo)
```

```
#SIGNIFICANCIA DEL MODELO
anova(modelo)

#INTERVALOS DE CONFIANZA PARA LOS ESTIMADORES
confint(modelo)
```

Apéndice B

Capítulo 3

B.1. Estadísticos descriptivos, estimación de los modelos y comparaciones múltiples

Altura		
Número de casos		2070
Media		174.66
Mediana		75.99
Desv. típ.		206.44
Varianza		42617.88
Asimetría		1.06
Curtosis		-0.31
Mínimo		1.00
Máximo		836.97
Percentiles	25	10
	50	75.99
	75	250.00

Tabla B.1: Estadísticos para la Altura

Edad		
Número de casos		2070
Media		11.88
Mediana		11.96
Desv. típ.		1.18
Varianza		1.39
Asimetría		-0.16
Curtosis		1.88
Mínimo		6.44
Máximo		16.79
Percentiles	25	11.27
	50	11.96
	75	12.45

Tabla B.2: Estadísticos para la Edad

Humedad		
Número de casos		2070
Media		69.66
Mediana		69.58
Desv. típ.		1.31
Varianza		1.72
Asimetría		0.23
Curtosis		-0.19
Mínimo		64.60
Máximo		73.85
Percentiles	25	68.70
	50	69.58
	75	70.59

Tabla B.3: Estadísticos para la Humedad

Número de cortes		
Número de casos		2070
Media		3.55
Mediana		3.00
Desv. típ.		2.73
Varianza		7.43
Asimetría		0.82
Curtosis		0.10
Mínimo		1.00
Máximo		11.00
Percentiles	25	2.00
	50	3.00
	75	5.00

Tabla B.4: Estadísticos para el Número de cortes

Lluvia		
Número de casos		2070
Media		1687.43
Mediana		1495.50
Desv. típ.		334.19
Varianza		111685.78
Asimetría		0.47
Curtosis		1.12
Mínimo		0.00
Máximo		2400.40
Percentiles	25	1446.10
	50	1495.50
	75	1851.50

Tabla B.5: Estadísticos para la Lluvia acumulada

Amplitud térmica		
Número de casos		2070
Media		9.8
Mediana		9.28
Desv. típ.		2.45
Varianza		6.01
Asimetría		0.60
Curtosis		0.07
Mínimo		0.00
Máximo		16.05
Percentiles	25	7.97
	50	9.28
	75	10.97

Tabla B.6: Estadísticos para la Amplitud térmica

Temperatura		
Número de casos		2070
Media		23.53
Mediana		23.92
Desv. típ.		3.18
Varianza		10.11
Asimetría		-1.32
Curtosis		7.56
Mínimo		0.00
Máximo		29.07
Percentiles	25	21.75
	50	23.92
	75	25.41

Tabla B.7: Estadísticos para la Temperatura

Estimación del modelo para el rendimiento de azúcar					
Coefficients:					
	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	520	5.56	93.595	< 2e-16	***
Humedad	-5.926	0.081	-73.577	< 2e-16	***
Altura	0.003	0.001	4.186	2.99e-05	***
Amplitud.térmica	0.424	0.052	8.17	6.18e-16	***
Temperatura	-0.182	0.04	-4.589	4.80e-06	***
Textura.suelo1Arcilloso	-0.445	0.355	-1.254	2.10e-01	
Textura.suelo1Fr - franco arenoso	-0.979	0.481	-2.038	4.17e-02	*
Textura.suelo1Franco	-0.811	0.323	-2.509	1.22e-02	**
Textura.suelo1Franco limoso	-1.235	0.291	-4.244	2.32e-05	***
Textura.suelo1Otro	-2.112	0.723	-2.923	3.52e-03	
Tipo.mad1MODDUS	0.671	0.384	1.747	8.08e-02	.
Tipo.mad1NO	-1.531	0.253	-6.049	1.81e-09	***
Tipo.mad1OTRO	-0.446	0.861	-0.518	6.04e-01	
Tipo.mad1SELECT	-0.599	0.422	-1.419	1.56e-0.1	
Tipo.corteMecanizada	-1.688	0.298	-5.660	1.79e-08	***
Variedad.del.lote1CP-72-1210	-0.758	0.424	-1.788	7.40e-02	
Variedad.del.lote1CP-73-1547	-0.168	0.332	-0.508	6.12e-01	
Variedad.del.lote1CP-88-1165	-2.315	0.538	-4.301	1.80e-05	***
Variedad.del.lote1CP-89-2143	2.859	0.417	6.86	1.01e-11	***
Variedad.del.lote1MEX-79-431	-1.947	0.267	-7.292	4.78e-13	***
Variedad.del.lote1Otras	-0.668	0.433	-1.542	1.23e-01	
Variedad.del.lote1PR-87-2080	-1.628	0.706	-2.304	2.13e-02	*
Variedad.del.lote1Varias	-0.914	0.576	-1.589	1.12e-01	
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1 Residual standard error: 3.813 on 1608 degrees of freedom Multiple R-squared: 0.8234, Adjusted R-squared: 0.821 F-statistic: 340.9 on 22 and 1608 DF, p-value: < 2.2e-16					

Tabla B.8: Estimación del modelo para el rendimiento de azúcar

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
Arcillosos - Franco Arcilloso == 0	-0.4452	0.3549	-1.254	0.7921	
Franco - Franco Arcilloso == 0	-0.8106	0.3231	-2.509	0.1094	
Franco - Franco Arenoso - Franco Arcilloso == 0	-0.9794	0.4805	-2.038	0.2968	
Franco limoso - Franco Arcilloso == 0	-1.2347	0.2909	-4.244	<0.001	***
Otro - Franco Arcilloso == 0	-2.1125	0.7228	-2.923	0.0361	*
Franco - Arcillosos == 0	-0.3655	0.4253	-0.859	0.9506	
Franco - Franco Arenoso - Arcillosos == 0	-0.5343	0.5502	-0.971	0.9189	
Franco limoso - Arcillosos == 0	-0.7895	0.4083	-1.934	0.3561	
Otro - Arcillosos == 0	-1.6673	0.7763	-2.148	0.2413	
Franco - Franco Arenoso - Franco == 0	-0.1688	0.5201	-0.325	0.9995	
Franco limoso - Franco == 0	-0.4240	0.3702	-1.146	0.8483	
Otro - Franco == 0	-1.3018	0.7539	-1.727	0.4871	
Franco limoso - Franco - Franco Arenoso == 0	-0.2552	0.5001	-0.510	0.9952	
Otro - Franco - Franco Arenoso == 0	-1.1330	0.8180	-1.385	0.7145	
Otro - Franco limoso == 0	-0.8778	0.7446	-1.179	0.8321	
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1					
(Adjusted p values reported – single-step method)					

Tabla B.9: Comparaciones múltiples entre textura de suelo para kg az/t de caña

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
MODDUS - ROUNDUP SL == 0	0.6709	0.3840	1.747	0.3724	
NO - ROUNDUP SL == 0	-1.5308	0.2531	-6.049	<0.001	***
OTRO - ROUNDUP SL == 0	-0.4460	0.8607	-0.518	0.9833	
SELECT - ROUNDUP SL == 0	-0.5992	0.4222	-1.419	0.5833	
NO - MODDUS == 0	-2.2016	0.3477	-6.332	<0.001	***
OTRO - MODDUS == 0	-1.1169	0.9039	-1.236	0.7036	
SELECT - MODDUS == 0	-1.2701	0.4936	-2.573	0.0651	.
OTRO - NO == 0	1.0848	0.8547	1.269	0.6823	
SELECT - NO == 0	0.9315	0.4026	2.314	0.1231	
SELECT - OTRO == 0	-0.1532	0.9268	-0.165	0.9998	
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1					
(Adjusted p values reported – single-step method)					

Tabla B.10: Comparaciones múltiples entre tipo de madurante para kg az/t de caña

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
Mecanizada - Manual == 0	-1.6876	0.2981	-5.66	1.79e-08	***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Adjusted p values reported – single-step method)					

Tabla B.11: Comparaciones múltiples entre tipo de corte para kg az/t de caña

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
CP-72-1210 - CP-72-2086 == 0	-0.7583	0.4242	-1.788	0.447561	
CP-73-1547 - CP-72-2086 == 0	-0.1684	0.3316	-0.508	0.999391	
CP-88-1165 - CP-72-2086 == 0	-2.3154	0.5383	-4.301	0.000144	***
CP-89-2143 - CP-72-2086 == 0	2.8595	0.4171	6.856	$<$ 1e-04	***
MEX-79-431 - CP-72-2086 == 0	-1.9467	0.2670	-7.292	$<$ 1e-04	***
OTRAS - CP-72-2086 == 0	-0.6677	0.4330	-1.542	0.636726	
PR-87-2080 - CP-72-2086 == 0	-1.6278	0.7064	-2.304	0.154802	
VARIAS - CP-72-2086 == 0	-0.9144	0.5755	-1.589	0.600326	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Adjusted p values reported – single-step method)					

Tabla B.12: Comparaciones múltiples entre variedades para kg az/t de caña

Estimación del modelo para el rendimiento de caña					
Coefficients:					
	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	-625.1144	40.2717	-15.522	< 2e-16	***
Edad.meses	6.485	0.585	11.090	< 2e-16	***
Ncortes	0.6211	0.243	2.556	0.010678	*
Humedad	9.401	0.533	17.634	< 2e-16	***
Amplitud.térmica	-0.868	0.266	-3.259	0.001143	**
Tipo.mad1MUDDUS	-2.142	2.492	-0.860	0.390151	
Tipo.mad1NO	-7.383	1.600	-4.613	4.28e-06	***
Tipo.mad1OTRO	5.232	5.619	0.931	0.351997	
Tipo.mad1SELECT	-3.823	2.716	-1.408	0.159468	
Tipo.corteMecanizada	-4.848	1.840	-2.635	0.008493	**
Variedad.del.lote1CP-72-1210	-2.953	2.774	-1.064	0.287268	
Variedad.del.lote1CP-73-1547	0.5096	2.1626	0.236	0.813675	
Variedad.del.lote1CP-88-1165	-6.688	3.553	-1.883	0.059933	.
Variedad.del.lote1CP-89-2143	-2.776	2.752	-1.009	0.313293	
Variedad.del.lote1MEX-79-431	-10.055	1.726	-5.826	6.85e-09	***
Variedad.del.lote1Otras	-0.2537	2.8799	-0.091	0.927851	
Variedad.del.lote1PR-87-2080	-15.379	4.613	-3.334	0.000876	***
Variedad.del.lote1Varias	2.586	3.822	0.677	0.498815	
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1					
Residual standard error: 24.95 on 1592 degrees of freedom					
Multiple R-squared: 0.2725, Adjusted R-squared: 0.2648					
F-statistic: 35.08 on 17 and 1591 DF, p-value: < 2.2e-16					

Tabla B.13: Estimación del modelo para el rendimiento de caña

	Estimate	Std. Error	t value	Pr(> |t|)	
MODDUS - ROUNDUP SL == 0	-2.142	2.492	-0.860	0.900	
NO - ROUNDUP SL == 0	-7.383	1.600	-4.613	<0.001	***
OTRO - ROUNDUP SL == 0	5.232	5.619	0.931	0.870	
SELECT - ROUNDUP SL == 0	-3.823	2.716	-1.408	0.591	
NO - MODDUS == 0	-5.241	2.260	-2.319	0.121	
OTRO - MODDUS == 0	7.374	5.892	1.252	0.693	
SELECT - MODDUS == 0	-1.681	3.215	-0.523	0.983	
OTRO - NO == 0	12.614	5.568	2.265	0.137	
SELECT - NO == 0	3.559	2.626	1.356	0.625	
SELECT - OTRO == 0	-9.055	6.030	-1.502	0.528	
Signif. codes: 0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1					
(Adjusted p values reported – single-step method)					

Tabla B.14: Comparaciones múltiples entre tipo de madurante para t de caña/ha

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
Mecanizada - Manual == 0	-4.848	1.840	-2.635	0.00849	**
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Adjusted p values reported – single-step method)					

Tabla B.15: Comparaciones múltiples entre tipo de corte para t de caña/ha

	Estimate	Std. Error	t value	Pr($>\|t\|$)	
CP-72-1210 - CP-72-2086 == 0	-2.9525	2.7736	-1.064	0.92678	
CP-73-1547 - CP-72-2086 == 0	0.5096	2.1620	0.236	1.00000	
CP-88-1165 - CP-72-2086 == 0	-6.6884	3.5527	-1.883	0.37954	
CP-89-2143 - CP-72-2086 == 0	-2.7760	2.7521	-1.009	0.94522	
MEX-79-431 - CP-72-2086 == 0	-10.0546	1.7258	-5.826	$<$ 1e-04	***
OTRAS - CP-72-2086 == 0	-0.2537	2.8789	-0.088	1.00000	
PR-87-2080 - CP-72-2086 == 0	-15.3787	4.6130	-3.334	0.00695	**
VARIAS - CP-72-2086 == 0	2.5857	3.8222	0.677	0.99539	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Adjusted p values reported – single-step method)					

Tabla B.16: Comparaciones múltiples entre variedades para kg az/t de caña

Bibliografía

Bibliografía

[1] Aguilar Rivera, Noé. *Ficha técnica del cultivo de caña de azúcar (2013)*, Universidad Veracruzana, Facultad de Ciencias Biológicas y Agropecuarias.

[2] Aparicio, Juan; Martínez Mayoral, Ma Asunción y Morales, Javier. *Modelos Lineales Aplicados en R*, Depto. Estadística, Matemáticas e Informática. Centro de Investigación Operativa, Universidad Miguel Hernández.

[3] Arias Flores, Enrique Sebastián. *Diagnóstico de rendimientos de caña de azúcar utilizando factores climatológicos múltiples (2008)*), Zamorano, Honduras.

[4] C. Montgomery, Douglas. *Diseño y análisis de experimentos (2004)*, Editorial Limusa, Limusa Wiley, Universidad Estatal de Arizona.

[5] C. Montgomery, Douglas. *Introducción al análisis de regresión lineal (2006)*, tercera edición, Universidad Estatal de Arizona.

[6] Díaz Montejo, Lucas Lizandro; Portocarrero Rivera, Eduardo Tomás. *Manual de Producción de Caña de Azúcar (2002)*, Honduras.

[7] Durán, Rafael Quintero. *Fertilización y nutrición de la caña de azúcar (1995)*, Cali, CENICAÑA.

[8] Morales González, Domingo. *Diseño de Experimentos (2014)*, Departamento de Estadística y Matemática Aplicada, Universidad Miguel Hernández de Elche.

[9] Peña, Daniel. *Regresión y diseño de experimentos (2002)*, Alianza Editorial.

[10] R. Romero, Eduardo; Scandaliaris Jorge. *La caña de azúcar Características y ecofisiología (2009)*

[11] Rein, Peter. *Ingeniería de la caña de azúcar (2012)*, Berlin, Alemania.

[12] Suarez García, Luis Fernando. *Manejo Agronómico del cultivo de la caña de azúcar (2012)*, Universidad Veracruzana.