UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

TESIS:

MODELACIÓN LOGÍSTICA MULTINOMIAL PARA CLASIFICAR LOS HOGARES DE EL SALVADOR POR NIVEL DE POBREZA

PRESENTADO POR:

SANDRA ELIZABETH GÓMEZ HERNÁNDEZ.

DARWIN ERNESTO PALACIOS ARIAS.

PARA OPTAR AL GRADO DE:

LICENCIADO(A) EN ESTADÍSTICA.

Ciudad Universitaria, Febrero de 2013.

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

TESIS:

MODELACIÓN LOGÍSTICA MULTINOMIAL PARA CLASIFICAR LOS HOGARES DE EL SALVADOR POR NIVEL DE POBREZA

PRESENTADO POR:

SANDRA ELIZABETH GÓMEZ HERNÁNDEZ.

DARWIN ERNESTO PALACIOS ARIAS.

ASESOR:

DR. JOSÉ NERYS FUNES TORRES

Ciudad Universitaria, Febrero de 2013.

AUTORIDADES

RECTOR UNIVERSITARIO:
ING.MARIO ROBERTO NIETO LOVO

SECRETARIA GENERAL:
DRA. ANA LETICIA ZAVALETA DE AMAYA

FISCAL GENERAL:
LIC. FRANCISCO CRUZ LETONA

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

DECANO:
MSC. MARTÍN ENRIQUE GUERRA CÁCERES

SECRETARIO:
CARLOS QUINTANILLA

ESCUELA DE MATEMÁTICA

DIRECTOR
DR. JOSÉ NERYS FUNES TORRES

SECRETARIA
ALBA IDALIA CÓRDOVA CUÉLLAR

Ciudad Universitaria, Febrero de 2013.

UNIVERSIDAD DE EL SALVADOR

FACULTAD DE CIENCIAS NATURALES Y MATEMÁTICA

ESCUELA DE MATEMÁTICA

ASESOR:

DR. JOSÉ NERYS FUNES TORRES

_____

Ciudad Universitaria, Febrero de 2013.

ÍNDICE DE CONTENIDO

Introducción.

En la presente investigación se realiza una aplicación de los modelos de regresión logística multinomial para determinar las variables de mayor incidencia en la clasificación de los hogares de El Salvador por nivel de pobreza, utilizando como información principal la base de datos obtenida a través de las Encuestas de Hogares de Propósitos Múltiples del año 2010.

El objetivo principal es encontrar un modelo de regresión logística multinomial para la clasificación de los hogares salvadoreños en tres niveles: pobre extremo, pobre relativo y no pobre, para clasificar la situación que presenta un hogar, se buscó perfilarlos de acuerdo a un conjunto de características referidas al Jefe de Hogar y que tengan relación con los niveles de pobreza, considerando que el tipo de pobreza, está determinado por un conjunto de características estructurales del Jefe de hogar, vinculadas a las siguientes dimensiones: Geográficas, Demográficas, Mercado Laboral, Educación, Vivienda, Ingresos, Patrimonio, por lo que puede utilizarse dicho modelo como referencia para la toma de decisiones sobre programas sociales para los hogares más pobres de El Salvador.

Para el desarrollo de nuestra investigación fue necesaria la siguiente organización:
En el Capítulo I, se describen los fundamentos teóricos necesarios sobre la construcción de los modelos de regresión logística multinomial, es decir, una exposición de las principales herramientas estadísticas necesarias, y el contraste de hipótesis que se plantean en la investigación. En el Capítulo II, se presenta las características principales de la Encuestas de Hogares de Propósitos Múltiples del año 2010 de la cual obtenemos la información primaria a utilizar, también se presentan las variables consideradas en el estudio, además de exponer la aplicación del modelo de regresión logística multinomial donde muestra la estimación de los parámetros, interpretación del modelo, su bondad y calidad de ajuste y la validación del modelo encontrado. Y por último, se presentan las conclusiones de los principales resultados obtenidos.

Para poder realizar el análisis que se presenta en esta investigación se ha utilizado el paquete estadístico SPSS 18 (PASWS Statistics 18)

Antecedentes y Justificación.

En El Salvador en los últimos años, se ha generado una serie de estudios acerca de la pobreza, entre ellos se puede citar al desarrollado por FLACSO-MINEC-PNUD (2010), titulado: "Mapa de pobreza urbana y exclusión social". Este estudio constituye una valiosa herramienta para localizar geográficamente y dimensionar la pobreza en las zonas urbanas de El Salvador. Se creó con el objetivo de apoyar con criterios técnicos la toma de decisiones de focalización de los programas sociales, para así contribuir a mejorar la calidad de vida de las familias urbanas que viven en condiciones de pobreza y exclusión social. La elaboración de este Mapa de pobreza urbana y exclusión social involucra una serie de cálculos complejos, a partir de definiciones conceptuales y operativas, encaminada a situar los asentamientos urbanos precarios o AUP en la cartografía de las áreas urbanas de El Salvador. Asimismo, permite comparar los asentamientos entre ellos, determinando aquellos que merecen mayor atención por su nivel de pobreza urbana y exclusión social.

A diferencia de esta investigación nuestro modelo se basará en identificar los niveles de pobreza por hogar y no por áreas o conglomerados de viviendas como lo hace el mapa de pobreza urbana y exclusión social. Y para ello se recurrirá a un modelo lineal generalizado, específicamente un caso particular de estos, el cual es un modelo de regresión logístico multinomial.

El estudio y análisis de los modelos lineales generalizados es extenso, sin embargo las investigaciones estadísticas en el país utilizan muy poco la aplicación de estos modelos en el análisis de datos. Este hecho se debe a que dichos métodos, a pesar de ser importantes, son de escaso dominio a nivel estudiantil y profesional. Son temáticas que se desarrollan en estudios de postgrado en el campo de la Estadística, por lo que se conoce muy poco en nuestro país. En la Maestría en Estadística, en el área de Modelos Lineales Generalizados, se han realizado los siguientes trabajos de investigación:

1- "Construcción de un Modelo de Regresión Logístico Sobre la Oferta Laboral a Jefes(as) de Hogares en El Salvador", la que consiste en determinar las probabilidades de que un(a) jefe(a) de hogar se encuentre desocupado(a) a partir del conocimiento de variables socio-económicas y demográficas.

2- “Comparación entre el análisis discriminante y la regresión logística en la clasificación de una colonia de cangrejos herradura (Limulus Polyphemus)”, la que consiste en aplicar las técnicas (Regresión logística y análisis discriminante) a un conjunto de variables que representan características cualitativas y cuantitativas de una muestra de 173 cangrejos de herradura.

3- “Métodos robustos aplicados a la clasificación del estado nutricional de la niñez salvadoreña (FESAL 2008)”, la que consiste en una aplicación de los modelos de regresión logística y del análisis discriminante con el fin de determinar el conjunto de variables más importantes en la estimación del estado nutricional de la niñez salvadoreña. Dichas técnicas, se utilizan en la base de datos de la Encuesta de Salud Familiar (FESAL) realizada en el año 2008 por la Asociación Demográfica Salvadoreña.

En nuestra investigación se realizó una aplicación del método de regresión logística multinomial para la clasificación de un hogar en el nivel de pobreza adecuado en virtud de sus características demográficas y socioeconómicas del hogar y su jefatura.

El análisis de los determinantes de la pobreza que se ha realizado permite establecer los factores que determinan la pobreza, por medio del estudio de características del hogar y del jefe(a) de hogar que pueden tener efectos sobre la pobreza de los hogares. Por lo que se ajusta un modelo estadístico de regresiones categóricas que permite seleccionar las mejores variables para la identificación de pobres relativos, pobres extremos y no pobres. El modelo probabilístico encontrado permite una medición alternativa de la pobreza en El Salvador, el cual es validado con los datos obtenidos por medio de “La Encuesta de Hogares de Propósitos Múltiples del año 2010” (EHPM-2010) mostrando el grado de predicción.

En base al resultado obtenido, es posible calcular la probabilidad de ser pobre para cada una de las familias entrevistadas por la EHPM realizada en el año 2010 y así obtener, por medio de estas variables, una medición alternativa de pobreza.

Planteamiento del problema.

En El Salvador, el tema de la pobreza, es reconocido como uno de los problemas que exigen rápidamente solución y que enfrenta actualmente el país. Tanto a nivel nacional como municipal, la discusión se centra, por una parte en la identificación de indicadores que caracterizan la pobreza, por otra en la formulación de políticas que disminuyan las condiciones desfavorables de los grupos más vulnerables de la población.

En El Salvador, la mayoría de los programas sociales (educación, salud, subsidios, etc.) se asignan o se operan a nivel comunal y la asignación de recursos a éstos se efectúa de acuerdo a los niveles de pobreza que muestran las comunas. Uno de los principales objetivos de la política social en El Salvador es la focalización de los recursos en los sectores más pobres de la población.

Se tiene como antecedente que la medición de la pobreza en El Salvador se realiza utilizando la metodología de la Línea del ingreso, el cual cuantifica y clasifica a los hogares según su ingreso, considerándose pobres a los que no alcanzan un umbral predeterminado de ingreso. El umbral para medir la pobreza es la Canasta Básica Alimentaria (CBA). La CBA es un conjunto de alimentos básicos que conforman la dieta usual de una población en cantidades suficientes para cubrir adecuadamente, por lo menos, las necesidades energéticas de todo individuo.

En nuestro trabajo se desarrolló un modelo probabilístico utilizando "Regresión Logística Multinomial" el cual permite calcular la probabilidad de que el hogar sea pobre relativo, pobre extremo o no pobre para cada una de las familias entrevistadas por la Encuesta de Hogares de Propósitos Múltiples del año 2010 obteniendo, en base a variables referidas al jefe de hogar como las Geográficas, Demográficas, de Mercado Laboral, Educación, Vivienda, Ingresos, Patrimonio, una medición alternativa de pobreza.

Objetivos.

Objetivos Generales

- Obtener un modelo que permita identificar las variables que son determinantes de la pobreza en los hogares.
- Tener un modelo de estimación de pobreza, que sea usado por encuestas de muestras representativas de la población.

Objetivos Específicos

- Describir la teoría del análisis de regresión logística multinomial y aplicarla a datos reales como lo es la Encuesta de Hogares de Propósitos Múltiples del año 2010.
- Cuantificar e interpretar los efectos de cada variable sobre la probabilidad de encontrarse un hogar en algún tipo de pobreza, a través de las estimaciones obtenidas en el modelo.

Hipótesis.

Hipótesis de trabajo

La pobreza corresponde a múltiples factores, que se ven reflejados por variables vinculadas al hogar y su jefatura. El ser pobre o no, está determinado por un conjunto de características estructurales del hogar, vinculadas a las siguientes dimensiones:

- Geográficas
- Demográficas
- Mercado Laboral
- Educación
- Vivienda
- Ingresos
- Patrimonio

Hipótesis Específicas

- La inserción de los Jefes(as) de Hogar en el mercado laboral tiene una fuerte incidencia en la probabilidad de que el hogar sea pobre o no.
- La edad del Jefe(a) de Hogar tiene incidencia en la probabilidad de que el hogar sea pobre o no, a menor edad mayor probabilidad de ser pobre.
- La escolaridad del Jefe(a) de Hogar tiene incidencia en la probabilidad de que el hogar sea pobre o no.
- A mayor número de personas en el hogar mayor es la posibilidad de que el hogar sea pobre.
- El género del Jefe(a) de Hogar tiene incidencia en la probabilidad de que el hogar sea pobre.
- El acceso a bienes y servicios básicos incide en menores niveles de pobreza.

CAPÍTULO I: MODELOS DE REGRESIÓN LOGÍSTICA MULTINOMIAL.

Introducción.

Los modelos de regresión logística son modelos estadísticos en los que se pretende conocer la relación entre una variable dependiente cualitativa, dicotómica (regresión logística binaria o binomial) y entre variables explicativas independientes, que pueden ser cualitativas o cuantitativas. Las covariables cualitativas que sean dicotómicas, es aconsejable que se codifiquen tomando valores 0 para una de las categorías o para su ausencia, y 1 para la otra categoría o para su presencia (esta codificación es importante ya que cualquier otra codificación podría provocar modificaciones en la interpretación del modelo). Pero si la covariable cualitativa tuviera más de dos categorías, se realiza una transformación, para poderla incluir en el modelo. Esta transformación consiste en crear varias variables cualitativas dicotómicas ficticias o de diseño, llamadas variables dummies, de forma que una de las variables se tomaría como referencia y cada una de las variables creadas entraría en el modelo de forma individual. En general, si la covariable cualitativa posee c categorías, habrá que realizar c-1 covariables ficticias.

La regresión logística multinomial es utilizada en modelos con variable dependiente de tipo nominal con más de dos categorías (politómica) y es una extensión multivariante de la regresión logística binaria clásica. Las variables independientes pueden ser tanto continuas (covariables) como categóricas (factores).

Tradicionalmente las variables dependientes politómicas han sido modeladas mediante análisis discriminante pero, gracias al creciente desarrollo de las técnicas de cálculo, cada vez es más habitual el uso de los modelos de regresión logística multinomial, implementados en paquetes estadísticos, debido a la mejor interpretatividad de los resultados que proporcionan.

Estos modelos se analizan eligiendo una categoría como referencia de la variable respuesta y se modelan varias ecuaciones simultáneamente, una para cada una de las categorías respecto a la de referencia.

En general, los requisitos y etapas de la regresión logística son los que se muestran a continuación, posteriormente detallaremos esas etapas.

- Recodificar las variables independientes categóricas u ordinales en variables ficticias o simuladas y así como también la variable dependiente.
- Evaluar efectos de confusión y de interacción del modelo explicativo.
- Evaluar la bondad de ajuste de los modelos.
- Analizar la fuerza, sentido y significación de los coeficientes, sus exponenciales y estadísticos de prueba (por ejemplo, el estadístico de Wald).

A continuación, mostraremos la formulación de los modelos de regresión logística multinomial, así como los contrastes aplicados sobre este modelo y la inferencia, para ampliar esta teoría se puede consultar el libro: Aguilera del Pino, A. M. (2002), Modelos de Respuestas Discreta.

1.1 Formulación e interpretación del modelo.

1.1.1 Formulación.

En el modelo de regresión logística se codifican los valores de la variable dependiente como 0 y 1, lo que da como resultado que la media de la variable represente la proporción de casos que ocurren en una de sus dos categorías (en el caso binomial) o en una de sus múltiples categorías (en el caso multinomial). El valor predicho de la probabilidad por el modelo según la categoría puede ser interpretado como la probabilidad de que un caso caiga en esa categoría (según libro de Menard, 2002). Un modelo lineal no se ajusta apropiadamente a variables binomiales, dado que los valores predichos de la variable dependiente con este modelo (ajustados mediante la ecuación de una recta), pueden tomar valores imposibles de probabilidad mayores que 1 o menores que 0, a pesar de que los valores observados estén entre 0 y 1. La misma situación se extiende a variables multinomiales.

El mejor modelo que linealiza la relación entre variable dependiente e independiente es el modelo logit, construido a través de regresión logística (según libro de Debella-Gilo et al., 2007). En una variable dependiente binomial (Y=0; Y=1), si se conoce la probabilidad de pertenecer a una clase (Y=0), se puede conocer la probabilidad de pertenecer a la otra clase (Y=1), es decir:

$$P(Y = 1) = [1 - P(Y = 0)] \tag{1.1}$$

Se puede tratar de aplicar el modelo lineal de probabilidad, expresado como:

$$P(Y = 1) = \beta_0 + \beta_1 X \tag{1.2}$$

Donde:

$P(Y = 1)$, es la probabilidad asociada a la variable predicha (Y),

X, es la variable independiente o predictora,

β, son los parámetros de la población a ser estimados.

Este modelo de probabilidad presenta el problema de no linealidad, con valores predichos que pueden ser menores que cero o mayores que uno.

Un paso a la solución del problema es reemplazar la probabilidad $[P(Y=1)]$ con el Odds (Y=1), el cual se expresa como:

$$Odds(Y=1) = \frac{P(Y=1)}{[1-P(Y=1)]} \tag{1.3}$$

Como se deduce de la Ecuación 1.3, el Odds puede variar entre 0 y $+\infty$, para valores de P entre 0 y 1, por lo que valores de Odss menores que 0 generen valores imposibles de P (el valor de una variable puede ser cualquier cifra entre $-\infty$ y $+\infty$, sin embargo, en este modelo está restringida a valores entre 0 y $+\infty$). Para evitar lo anterior se requiere otra transformación, conocida como logit(Y), que se expresa:

$$logit(Y) = Ln\left(\frac{P(Y=1)}{[1-P(Y=1)]}\right) \tag{1.4}$$

El logit varía entre $-\infty$ y $+\infty$, cuando el Odds varía entre 0 y $+\infty$, para valores de P entre 0 y 1, por lo que se elimina el problema de que la probabilidad estimada puede exceder el máximo o el mínimo posible.

Así, la ecuación de la relación entre la variable dependiente y las variables independientes es la siguiente:

$$logit(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_q X_q \tag{1.5}$$

El logit de la variable dependiente se puede convertir a Odds mediante la ecuación:

$$Odds(Y=1) = e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_q X_q)} \tag{1.6}$$

Se puede encontrar la probabilidad $P(Y=1)$ igualando y despejando dicha probabilidad, de las ecuaciones 1.3 y 1.6, llegando a la expresión:

$$P(Y=1) = \frac{e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_q X_q)}}{1+e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots . + \beta_q X_q)}} \tag{1.7}$$

Cuando la variable dependiente tiene más de dos categorías, se aplica la versión multinomial de la regresión logística. Para una variable dependiente con k categorías, la

regresión requiere k-1 ecuaciones logísticas, una para cada categoría, en relación a otra categoría tomada como referencia. La relación se representa por las siguientes funciones (según libro de Eastman, 2006):

$$g_1(x) = Ln\left[\frac{P(Y=1|X)}{P(Y=k|X)}\right] = \beta_{10} + \beta_{11}x_1 + \beta_{12}x_2 + \cdots . + \beta_{1q}x_q = X'B_1' \quad (1.8)$$

$$g_2(x) = Ln\left[\frac{P(Y=2|X)}{P(Y=k|X)}\right] = \beta_{20} + \beta_{21}x_1 + \beta_{22}x_2 + \cdots . + \beta_{2q}x_q = X'B_2' \quad (1.9)$$

.

.

.

$$g_{k-1}(x) = Ln\left[\frac{P(Y=k-1|X)}{P(Y=k|X)}\right] = \beta_{(k-1)0} + \beta_{(k-1)1}x_1 + \beta_{(k-1)2}x_2 + \cdots . + \beta_{(k-1)q}x_q = X'B_{k-1}' \quad (1.10)$$

Donde:

$g_i(x)$, es la función logit de la categoría i contra la categoría de referencia.

X, es el vector de variables independientes, $X = (x_0, x_1, \ldots, x_q)'$ con $x_0 = 1$

β_{ij}, es el vector de coeficientes, estimado para la categoría i y la variable j.

Veamos la formulación de estos modelos de forma general.

Consideremos una variable de respuesta politómica Y con más de dos categorías de respuestas que denotaremos por $Y_1, Y_2, \ldots, Y_k$.

Se pretende explicar la probabilidad de cada categoría de respuesta en función de un conjunto de covariables $X = \{x_1, x_2, \ldots, x_q\}$ observadas. Es decir, ajustar un modelo de

la forma $\pi_j = P\left(Y = Y_j \middle| X = x\right), \quad \forall\, j = 1, \dots, k$. Para cada vector x de valores observados de las variables explicativas X.

En el caso de una variable de respuesta binaria, la distribución condicionada a cada combinación de valores observados de las covariables sigue una distribución de Bernoulli. Cuando la variable de respuesta es politómica, la distribución de Bernoulli se convierte en una distribución multinomial de parámetros igual a las probabilidades de cada una de las categorías de respuesta. Es decir, $(Y|X = x) \rightarrow M(n;\, \pi_1, \dots \pi_k)$, siendo $\sum_{j=1}^{k} \pi_j = 1$.

Así que para obtener un modelo lineal, obtendremos $\binom{k}{2}$ transformaciones logit para comparar cada par de categorías de la variable respuesta, que sería de este tipo:

$$Ln\left[\frac{\frac{\pi_i}{\pi_i+\pi_j}}{\frac{\pi_j}{\pi_i+\pi_j}}\right] = Ln\left[\frac{\pi_i}{\pi_j}\right], \quad \forall i, j = 1, \dots k(i \neq j) \tag{1.11}$$

Que representa el logaritmo de la ventaja de respuesta Y_i frente a Y_j condicionado a las observaciones de las variables independientes que caen en uno de ambos niveles. Pero para construir el modelo logit de respuesta multinomial bastaría con considerar (k-1) transformaciones logit básicas, definidas con respecto a una categoría de referencia. Tomando como categoría de referencia la última Y_k. Así las transformaciones logit generalizadas se definen como $logit_j(x) = ln\left[\frac{\pi_j}{\pi_k}\right], \quad \forall\, j = 1, \dots, k-1$, siendo $logit_j(x)$ el logaritmo de la ventaja de respuesta Y_j dado que las observaciones de las variables independientes caen en la categoría Y_j o en la Y_k.

El modelo lineal para cada una de las transformaciones logit generalizado, para q variables explicativas, es de la siguiente forma:

$$logit_j(x) = \sum_{i=0}^{q} \beta_{ij} X_i = x'\beta_j\,, \quad \forall\, j = 1, \dots, k-1 \tag{1.12}$$

Para cada vector de valores observados de las variables explicativas $x = \left(x_0, x_1, x_2, \dots, x_q\right)'$ con $x_0 = 1$ y $\beta_j = (\beta_{0j}, \beta_{1j}, \dots \beta_{qj})'$ el vector de parámetros asociado a la categoría Y_j.

Para las probabilidades de respuesta, podemos escribir el modelo de la siguiente forma:

$$\pi_j = \frac{e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}}{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}}, \quad \forall\, j = 1, \dots, k-1. \tag{1.13}$$

Entonces para encontrar la probabilidad de respuesta para la categoría tomada como referencia π_k, se ocupa la propiedad que $\sum_{j=1}^{k}\pi_j = 1$.

Por lo tanto:

$$\pi_k = 1 - \sum_{j=1}^{k-1}\pi_j$$

$$\pi_k = 1 - \left(\frac{e^{\left(\sum_{i=0}^{q}\beta_{i1}X_i\right)}}{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}} + \frac{e^{\left(\sum_{i=0}^{q}\beta_{i2}X_i\right)}}{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}} + \cdots + \frac{e^{\left(\sum_{i=0}^{q}\beta_{i(k-1)}X_i\right)}}{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}}\right)$$

$$\pi_k = \frac{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}}{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}} - \left(\frac{e^{\left(\sum_{i=0}^{q}\beta_{i1}X_i\right)}}{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}} + \frac{e^{\left(\sum_{i=0}^{q}\beta_{i2}X_i\right)}}{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}} + \cdots + \frac{e^{\left(\sum_{i=0}^{q}\beta_{i(k-1)}X_i\right)}}{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}}\right)$$

$$\pi_k = \frac{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)} - e^{\left(\sum_{i=0}^{q}\beta_{i1}X_i\right)} - e^{\left(\sum_{i=0}^{q}\beta_{i2}X_i\right)} - \cdots - e^{\left(\sum_{i=0}^{q}\beta_{i(k-1)}X_i\right)}}{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}}$$

$$\pi_k = \frac{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)} - \sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}}{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}}$$

Eliminando términos semejantes en la diferencia, se tiene que:

$$\pi_k = \frac{1}{1+\sum_{j=1}^{k-1}e^{\left(\sum_{i=0}^{q}\beta_{ij}X_i\right)}} \tag{1.14}$$

1.1.2 Interpretación del modelo.

A continuación mostramos la interpretación de los parámetros del modelo, pero distinguiendo los casos según de qué tipo son las variables explicativas, cuantitativas o cualitativas.

- **Una variable predictora cuantitativa X.**

Si en el modelo tenemos solo una única covariable cuantitativa X, el modelo para cada valor observado x de la variable X viene dada por:

$$logit_j(x) = \beta_{0j} + \beta_{1j}x\,, \quad \forall\, j = 1, \dots, k-1 \tag{1.15}$$

A continuación mostramos la exponencial de los parámetros β_{1j} asociados a cada categoría de la variable dependiente, que se interpreta en términos de cocientes de ventajas (odds_ratio):

$$Odds\_ratio_j(\Delta X = 1) = \frac{\frac{\pi_j(x+1)}{\pi_k(x+1)}}{\frac{\pi_j}{\pi_k}} = \frac{e^{\beta_{0j}+\beta_{1j}(x+1)}}{e^{\beta_{0j}+\beta_{1j}x}} = e^{(\beta_{1j})}, \quad \forall\, j = 1, \dots, k-1 \tag{1.16}$$

$Odds\_ratio_j(\Delta X = 1)$, es el cociente de ventajas de respuesta Y_j frente a la última categoría Y_k cuando aumenta en una unidad la variable X.

- **Más de una variable predictora cuantitativa**.

Para el modelo logit generalizado múltiple, los cocientes de ventajas se definen incrementando una de las variables y manteniendo fijas las demás.

$$Odds\_ratio_j(\Delta X_r = 1 | X_s = x_s, s \neq r) = \frac{\frac{p\left(Y = Y_j \middle| X_r = x_r + 1, X_s = x_s, s \neq r\right)}{p(Y = Y_k | X_r = x_r + 1, X_s = x_s, s \neq r)}}{\frac{p\left(Y = Y_j \middle| X_r = x_r, X_s = x_s, s \neq r\right)}{p(Y = Y_k | X_r = x_r, X_s = x_s, s \neq r)}}$$

$$= e^{(\beta_{rj})}, \quad \forall\, j = 1, \dots, k-1, \qquad s = 1, \dots, q, r = 1, \dots, q, \qquad s \neq r$$

Siendo $Odds\_ratio_j(\Delta X_r = 1 | X_s = x_s, s \neq r)$ el cociente de ventajas de respuesta Y_j frente a la última categoría Y_k cuando aumenta en una unidad la variable X_r y las demás se mantiene fijas.

- **Variables predictoras categóricas.**

Si se incluyen en el modelo variables independientes categóricas, se introducen mediante sus variables del diseño asociados (variables dummies).

Supongamos que tenemos la variable categórica A con categorías $A_1, A_2, \dots, A_p$. si de esta variable realizamos la transformación a variables de diseño mediante el método

parcial que asigna un uno a la variable asociada a cada categoría y un cero al resto, y tomando como categoría de referencia la primera, obtenemos p-1 variables que las denotaremos como $X_m^A \quad (m = 2, \dots p)$.

Así el modelo de regresión logística multinomial generalizado que obtenemos sigue siendo un modelo lineal, como en los casos anteriores, para cada logit generalizado en función de esas variables de diseño procedentes de la variable X y viene dado por:

$$logit_{lj}(x) = \ln\left[\frac{\pi_{lj}}{\pi_{lk}}\right] = \beta_{0j} + \sum_{m=2}^{p} \tau_{mj}^{A} X_{lm}^{A} \; ; \quad l = 1, .., p; \;\; j = 1, \dots, k-1 \qquad (1.17)$$

Siendo $\pi_{lj} = p\left(Y = Y_j \middle| A = A_l\right)$, la probabilidad de respuesta Y_j en la categoría A_l.
Dado una observación para un individuo en específico y sabiendo que solo puede pertenecer a una de las categorías de las variables de diseño $X_m^A \quad (m = 2, \dots p)$. se puede definir el modelo como:

$$logit_{lj}(x) = \beta_{0j} + \tau_{lj}, \; para\ cualquier \;\; l = 1, \dots, p; \;\; y \;\; j = 1, \dots, k-1. \qquad (1.18)$$

Siendo $\tau_{1j} = 0, \forall\, j = 1, \dots, k-1$, ya que es tomada como referencia.
Este modelo en términos de cocientes de ventajas viene dado por:

$$Odds\_ratio_{(l1)j} = \frac{\frac{\pi_{lj}}{\pi_{lk}}}{\frac{\pi_{1j}}{\pi_{1k}}} = \frac{e^{(\beta_{0j}+\tau_{lj})}}{e^{(\beta_{0j})}} = e^{(\tau_{lj})}, \quad \forall\, j = 1, \dots, k-1, \;\; \forall\, l = 2, \dots, p$$, que es el cociente de respuesta Y_j frente a la última categoría Y_k para la categoría A_l de A respecto a la primera categoría A_1.

1.1.3 Otros aspectos a tener en cuenta sobre las variables.

Para seleccionar el conjunto de variables predictoras que se incluyen en el modelo, los criterios a seguir son:

Incluir todas aquellas variables que se consideran importante para el modelo, independientemente de si se ha demostrado o no significación estadística en un análisis univariado previo, ya que puede conducir a dejar de incluir en el modelo covariables

con una débil asociación a la variable dependiente en solitario, pero que podrían demostrar ser fuerte predictores de la misma al tomarlas en conjunto con el resto de covariables. Aunque se aconseja incluir todas las variables que aparentemente tienen relación con la variable dependiente. También es de tener presente que debemos de conseguir un modelo que sea lo más reducido posible que explique los datos (principio de parsimonia), y que además sea congruente e interpretable.

Para obtener el mejor modelo, se puede recurrir a métodos de selección paso a paso, mediante inclusión hacia adelante, o por eliminación hacia atrás, o a la selección de variables por mejores subconjuntos de covariables. Estos métodos se encuentran implementados en la mayoría de los paquetes estadísticos y se describen posteriormente.

Otro aspecto a tener en cuenta para elegir el número de covariables a incluir en un modelo de regresión logística es el tamaño muestral. Ya que modelos excesivamente grandes para muestras con tamaños muestrales relativamente pequeños podrían provocar errores estándar grandes o coeficientes estimados falsamente muy elevados (sobreajuste). Por lo que se suele recomendar, que por cada covariable se cuenta con un número mínimo de 10 individuos por cada categoría de la variable dependiente con menor representación.

También otra cuestión a tener en cuenta de los modelos de regresión logística, es la inclusión de factores de interacción, para estudiar como la asociación de dos o más covariables puede influir en la variable dependiente. Estas interacciones pueden ser de primer orden (tomadas las covariables dos a dos o de mayor orden, pero estas últimas suelen ser de difícil interpretación). Las interacciones se incluyen siempre que sean interpretables y tengan significado desde el punto de vista de la investigación. Si en un modelo se incluye una interacción de dos o más covariables, estas deben de estar incluidas también en el modelo de forma aislada.

1.2. Método de estimación. Estimación por máxima verosimilitud.

Para la estimación de los coeficientes del modelo y de sus errores estándar se utiliza la estimación por máxima verosimilitud, es decir, estimaciones que hagan máxima la

probabilidad de obtener los valores de la variable dependiente y proporcionada por los datos de la muestra. Al contrario de lo que ocurre con la estimación de los coeficientes de regresión lineal múltiple que se utiliza el método de los mínimos cuadrados, los cálculos para las estimaciones de los coeficientes de la regresión logística multinomial no son directos, hay que llevar a cabo métodos iterativos, como por ejemplo el método de Newton-Raphson, que se explicará posteriormente.

Al aplicar estos métodos además de obtener las estimaciones de los coeficientes de regresión, se obtienen sus errores estándar y las covarianzas entre las covariables del modelo.

A continuación describimos el método de estimación de máxima verosimilitud para el cálculo de los coeficientes de nuestro modelo de regresión logística multinomial. Primero veamos un ejemplo de cómo es la estructura de datos de un modelo logístico multinomial.

Ejemplo 1.1. En la **Tabla 1.1** se presentan datos enumerados por el número de orden del individuo. Obsérvese que hay q=2 variables explicativas, N=14 datos y la variable respuesta Y posee 3 categorías.

Tabla 1.1 *Ejemplo 1.1*

N. del individuo	Variables explicativas		Variable respuesta
J	X_1	X_2	Y
1	1	1	1
2	1	2	2
3	1	2	1
4	2	1	1
5	2	2	2
6	1	2	2
7	1	1	2
8	2	2	3
9	1	2	3
10	1	1	3
11	1	1	2
12	1	2	1
13	1	2	2
14	2	1	2

Frecuentemente ocurre que $X_j^T, j = 1, \dots, N$ toma solamente n valores distintos (clases covariantes o combinaciones diferentes de valores de las variables explicativas). En el **Ejemplo 1.1** hay cuatro clases covariantes: (1.1), (1,2), (2,1), (2,2).

Como ejemplo se puede observar en la tabla 1.1 que para la primer combinación (1,1), se repite cuatro veces: en la fila 1, 7, 10 y 11, por lo que se denomina m_1=4, al número total de observaciones de los datos para la combinación 1.

En la **Tabla 1.2** se presentan los datos del ejemplo con estructura de modelo multinomial. Los datos se ordenan por número de clases covariantes (o combinaciones). Además se observa que en la variable respuesta Y se toma solo 2 de sus 3 categorías, y Y_{i1}, Y_{i2} son el número de observaciones que caen en la categoría 1 y 2 respectivamente de la clase o combinación i. Cada m_i es el número total de observaciones de los datos anteriores para la clase covariante i (ó combinación i). Además p=2, N=14 y n=4.

Tabla 1.2. *Datos con estructura de modelo multinomial*

Unidad	Categoría de respuesta Y						
	1			**2**			
i	Y_1	X_{11}	X_{12}	Y_2	X_{21}	X_{22}	m_i
1	1	1	1	2	1	1	**4**
2	2	1	2	2	1	2	**6**
3	1	2	1	1	2	1	**2**
4	0	2	2	1	2	2	**2**

De lo anterior se puede observar que de manera general si disponemos de una muestra aleatoria de tamaño N con n combinaciones diferentes de valores de las variables explicativas $X_1, X_2, \dots, X_q$. Denotemos a cada combinación de valores de las variables explicativas por $x_i = (x_{i0}, x_{i1}, \dots, x_{iq})' \; con \; x_{i0} = 1 \quad \forall \, i = 1, \dots n$. En cada una de estas combinaciones se tiene una muestra aleatoria de m_i observaciones independientes de la variable de respuesta politómica Y, de entre las cuales denotamos por y_{ij} al número de observaciones que caen en la categoría de respuesta $Y_j \quad \forall \, j = 1, \dots, k$.

Así, se verifica que $\sum_{j=1}^{k} y_{ij} = m_i \quad y \quad \sum_{i=1}^{n} m_i = N$.

Además que: $y_{ik} = m_i - \sum_{j=1}^{k-1} y_{ij}$ y $\pi_{ik} = 1 - \sum_{j=1}^{k-1} \pi_{ij}$

Tal como se puede observar en la **Tabla 1.3** siguiente:

Tabla 1.3. *Representación multinomial*

	CATEGORÍA													
	1					**J**				**K-1**				
Unidad	**Y_1**	**X_{11}**	**...**	**X_{1q}**	**...**	**Y_j**	**X_{j1}**	**...**	**X_{jq}**	**...**	**Y_{k-1}**	**$X_{(k-1)1}$**	**...**	**$X_{(k-1)q}$**	m_i
1	y_{11}	X_{111}	...	X_{11q}	...	y_{1j}	X_{1j1}	...	X_{1jq}	...	$y_{1(k-1)}$	$X_{1(k-1)1}$	...	$X_{1(k-1)q}$	$\sum_{j=1}^{k} y_{1j} = m_1$
.			.....		.....			...		.....			.....		
i	y_{i1}	X_{i11}	...	X_{i1q}	...	y_{ij}	X_{ij1}	...	X_{ijq}	...	$y_{i(k-1)}$	$X_{i(k-1)1}$	...	$X_{i(k-1)q}$	$\sum_{j=1}^{k} y_{ij} = m_i$
.			.....		.....			...		.....			.....		
n	y_{n1}	X_{n11}	...	X_{n1q}	...	y_{nj}	X_{nj1}	...	X_{njq}	...	$y_{n(k-1)}$	$X_{n(k-1)1}$	...	$X_{n(k-1)q}$	$\sum_{j=1}^{k} y_{nj} = m_n$

Los vectores $\left(y_{i1}, \dots, y_{i(k-1)}\right)' \ \forall\, i = 1, \dots, n$, siguen una distribución de probabilidad multinomial independiente $M\left(m_i\,;\, \pi_{i1}, \dots, \pi_{i(k-1)}\right)$, siendo $\pi_{ij} = p\left(Y = Y_j \middle| X = x_i\right)$ y verificando que $\sum_{j=1}^{k} \pi_{ij} = 1$.

Su función de densidad viene dada por:

$$f(y;\pi) = \frac{m_i!}{y_{i1}! \dots y_{ik}!}\ {\pi_{i1}}^{y_{i1}} \dots {\pi_{ik}}^{y_{ik}} \tag{1.19}$$

Esta misma función se puede reescribir aplicando logaritmo y exponencial, quedando de la siguiente forma.

$$f(y;\pi) = exp\left\{\ln\left(\frac{m_i!}{y_{i1}! \dots y_{ik}!}\ {\pi_{i1}}^{y_{i1}} \dots {\pi_{ik}}^{y_{ik}}\right)\right\}$$

Aplicando propiedad de logaritmos y sabiendo que:

$y_{ik} = m_i - \sum_{j=1}^{k-1} y_{ij}$

Se tiene:

$$f(y;\pi) = exp\left\{\sum_{j=1}^{k-1} y_{ij} \ln\left(\pi_{ij}\right) + \left(m_i - \sum_{j=1}^{k-1} y_{ij}\right) \ln\left(\pi_{ik}\right) + \ln\left(\frac{m_i!}{y_{i1}! \dots y_{ik}!}\right)\right\}$$

$$f(y;\pi) = exp\left\{\sum_{j=1}^{k-1} y_{ij} \ln\left(\pi_{ij}\right) + m_i \ln\left(\pi_{ik}\right) - \sum_{j=1}^{k-1} y_{ij} \ln\left(\pi_{ik}\right) + \ln\left(\frac{m_i!}{y_{i1}! \dots y_{ik}!}\right)\right\}$$

$$f(y;\pi) = exp\left\{\sum_{j=1}^{k-1} y_{ij} \ln\left(\frac{\pi_{ij}}{\pi_{ik}}\right) + m_i \ln\left(\pi_{ik}\right) + \ln\left(\frac{m_i!}{y_{i1}! \dots y_{ik}!}\right)\right\} \tag{1.20}$$

Donde:

$\pi_{ik} = 1 - \sum_{j=1}^{k-1} \pi_{ij}$ e $y_{ik} = m_i - \sum_{j=1}^{k-1} y_{ij}$ y $E[y_{ij}] = \mu_{ij} = m_i \pi_{ij}$

$C(y_1, \dots, y_{k-1}) = \ln\left(\frac{m_i!}{y_{i1}! \dots y_{ik}!}\right)$

Con $\theta_{ij} = \ln\left(\frac{\pi_{ij}}{\pi_{ik}}\right) = \beta_{0j} + \sum_{p=1}^{q} \beta_{pj} x_{ipj}$. Para j=1,…k-1. (1.21)

Además aplicando exponencial a θ_{ij} se tiene que:

$e^{\theta_{ij}} = \frac{\pi_{ij}}{\pi_{ik}}$

$\Rightarrow \quad \pi_{ij} = \pi_{ik} e^{\theta_{ij}}$ (1.22)

También, dado que $\sum_{j=1}^{k} \pi_{ij} = 1$, se tiene:

$1 - \pi_{ik} = \sum_{j=1}^{k-1} \pi_{ij}$ (1.23)

Por (1.22)

$\sum_{j=1}^{k-1} \pi_{ij} = \pi_{ik} \sum_{j=1}^{k-1} e^{\theta_{ij}}$ (1.24)

Por transitividad dado (1.23) y (1.24)

$1 - \pi_{ik} = \pi_{ik} \sum_{j=1}^{k-1} e^{\theta_{ij}}$

$\Leftrightarrow \left(1 + \sum_{j=1}^{k-1} e^{\theta_{ij}}\right) \pi_{ik} = 1$

$\Leftrightarrow \pi_{ik} = \frac{1}{(1+\sum_{j=1}^{k-1} e^{\theta_{ij}})}$ (1.25)

Por tanto de (1.22) y (1.25):

$\pi_{ij} = \frac{1}{(1+\sum_{j=1}^{k-1} e^{\theta_{ij}})} e^{\theta_{ij}}$ (1.26)

Siendo θ_{ij} función de los parámetros $\beta's$ definido como en (1.21).

y $b(\theta_{i1}, \dots, \theta_{ik-1}) = -m_i \ln(\pi_{ik}) = -m_i \ln\left(\frac{1}{(1+\sum_{j=1}^{k-1} e^{\theta_{ij}})}\right)$

$b(\theta_{i1}, \dots, \theta_{ik-1}) = -m_i \ln(1) + m_i \ln\left(1 + \sum_{j=1}^{k-1} e^{\theta_{ij}}\right) = m_i \ln\left(1 + \sum_{j=1}^{k-1} e^{\theta_{ij}}\right)$

Como los θ_{ij} están en función de los $\beta's$ pero los π_{ij} están en función de los θ_{ij}.

Se tiene:

$\pi_{ij} = \pi_{ij}(\beta), \quad i = 1, \dots, n, \; j = 1, \dots, k-1.$

Y la función de verosimilitud para (1.20) viene dada por:

$L(\beta; y_1, \dots, y_n) = \prod_{i=1}^{n} exp\left\{\sum_{j=1}^{k-1} y_{ij} \ln\left(\frac{\pi_{ij}(\beta)}{\pi_{ik}(\beta)}\right) + m_i \ln(\pi_{ik}(\beta)) + \ln\left(\frac{m_i!}{y_{i1}! \dots y_{ik}!}\right)\right\}$ (1.27)

La función de log-verosimilitud es:

$$l = l(\beta; y_1, \dots, y_n) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{k-1} \mathrm{y_{ij}} \ln\left(\frac{\pi_{ij}(\beta)}{\pi_{ik}(\beta)}\right) + m_i \ln(\pi_{ik}(\beta)) + \ln\left(\frac{m_i!}{y_{i1}! \dots y_{ik}!}\right) \right\} \quad (1.28)$$

$$l = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{k-1} \mathrm{y_{ij}} \left(\mathbf{x_{ij}\beta_j}\right) - m_i \ln\left(1 + \sum_{r=1}^{k-1} \exp\{\mathbf{x_{ir}\beta_r}\}\right) + \ln\left(\frac{m_i!}{y_{i1}! \dots y_{ik}!}\right) \right\} \quad (1.29)$$

Para obtener las ecuaciones de verosimilitud se tiene que derivar (1.28) ó (1.29) con respecto al parámetro de interés e igualar a cero, teniendo en cuenta que:

$\theta_{ij} = \ln\left(\frac{\pi_{ij}}{\pi_{ik}}\right) = \beta_{0j} + \sum_{p=1}^{q} \beta_{pj} x_{ipj}$

Entonces se deducen las ecuaciones de verosimilitud para los parámetros respectivos y definido π_{ij} como en (1.26):

$$0 = U_{j0} = \frac{\partial l}{\partial \beta_{0j}} \sum_{i=1}^{n} \left(\mathrm{y_{ij}} - m_i \frac{\exp\{\mathbf{x_{ij}\beta_j}\}}{1 + \sum_{r=1}^{k-1} \exp\{\mathbf{x_{ir}\beta_r}\}} \right) = \sum_{i=1}^{n} \left(\mathrm{y_{ij}} - m_i \pi_{ij} \right) \quad (1.30)$$

$j = 1, \dots, k - 1$.

$$0 = U_{jp} = \frac{\partial l}{\partial \beta_{pj}} \sum_{i=1}^{n} \mathrm{x_{ijp}} \left(\mathrm{y_{ij}} - m_i \frac{\exp\{\mathbf{x_{ij}\beta_j}\}}{1 + \sum_{r=1}^{k-1} \exp\{\mathbf{x_{ir}\beta_r}\}} \right) = \sum_{i=1}^{n} \mathrm{x_{ijp}} \left(\mathrm{y_{ij}} - m_i \pi_{ij} \right) \quad (1.31)$$

$j = 1, \dots, k - 1, \qquad p = 1, \dots, q.$

Para obtener los estimadores de máxima verosimilitud hay que resolver k-1 sistemas de q+1 ecuaciones no lineales, que no se pueden resolver explícitamente. Es necesario acudir a métodos numéricos para obtener aproximaciones de $\beta_j = (\beta_{0j}, \dots, \beta_{qj})$.

1.3. Métodos numéricos para la obtención de estimadores máximo-verosímiles.

Las ecuaciones de verosimilitud no siempre proporcionan soluciones explícitas para β_j , $j = 1, \dots k - 1$. De ahí la necesidad de disponer de métodos numéricos para obtener estimaciones máximo verosímiles, $\hat{\beta}_j$, $j = 1, \dots, k - 1$, y en consecuencia obtener el ajuste $\hat{\mu}_i = g^{-1}\left(x_i{}^t \, \hat{\beta} \right)$.

1.3.1 El método de Newton-Raphson.

El algoritmo de Newton-Raphson se apoya en el desarrollo en serie de Taylor. Si β^* es solución de las ecuaciones de verosimilitud; es decir;

$$U(\beta^*) = 0$$

Y $\beta^{(0)}$ es un valor arbitrario de β, entonces el desarrollo de Taylor de primer orden garantiza la siguiente aproximación $0 = U(\beta^*) \cong U(\beta^{(0)}) + H(\beta^{(0)})(\beta^* - \beta^{(0)})$

Donde $H = \frac{\partial U}{\partial \beta} = \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right)$ es la matriz Hessiana. Despejando β^* de la aproximación, se obtiene $\beta^* \cong \beta^{(0)} - H^{-1}(\beta^{(0)})U(\beta^{(0)})$, que sirve de base para plantear la ecuación recurrente:

$$\hat{\beta}^{(r)} \cong \hat{\beta}^{(r-1)} - H^{-1}\left(\hat{\beta}^{(r-1)}\right)U(\hat{\beta}^{(r-1)}) \tag{1.32}$$

Donde $U = \left(\frac{\partial l}{\partial \beta_1}, \dots, \frac{\partial l}{\partial \beta_q}\right)^t$, $\hat{\beta}^{(r)} = \left(\hat{\beta}_1^{(r)}, \dots, \hat{\beta}_q^{(r)}\right)^t$, $H = \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right)_{j,k=1,\dots q}$, H es conocida como la matriz Hessiana de orden qxq y está formada por las segundas derivadas parciales respecto a los parámetros betas.

Además, $\hat{\beta}^{(r)}$ es el valor estimado de $\hat{\beta}$ en la r-ésima iteración del algoritmo y $H^{-1}\left(\hat{\beta}^{(r-1)}\right), U\left(\hat{\beta}^{(r-1)}\right)$ son H^{-1} y U evaluadas en $\hat{\beta}^{(r-1)}$.

1.3.2 El método de puntuaciones de Fisher.

El método de puntuaciones de Fisher utiliza el mismo algoritmo de Newton-Raphson, pero sustituye la matriz Hessiana H, por su esperanza; es decir, por la matriz información de Fisher cambiada de signo $I = -E[H]$. La ecuación recurrente sería

$$\hat{\beta}^{(r)} = \hat{\beta}^{(r-1)} + I^{-1}\left(\hat{\beta}^{(r-1)}\right) U\left(\hat{\beta}^{(r-1)}\right) \tag{1.33}$$

Donde $I = -E[H] = \left(-E\left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right]\right)_{j,k=1,\dots q}$

Observación. La estimación inicial de β para los métodos de Newton-Raphson y de puntuaciones de Fisher se puede obtener tomando $\hat{\mu}^{(0)} = y$. En tal caso $g(y) = X\hat{\beta}^{(0)} \Longrightarrow X^t g(y) = X^t X \hat{\beta}^{(0)} \implies \hat{\beta}^{(0)} = (X^t X)^{-1} X^t g(y)$.
Donde $g(y)$, sería estimada evaluando la función nexo $g(.)$ para el modelo en estudio con el vector de observaciones y.

1.4. Contrastes sobre los parámetros del modelo.

Una vez estimado el modelo, el siguiente paso será comprobar la significación estadística de cada uno de los coeficientes de regresión en el modelo. Para ello, se pueden emplear básicamente dos métodos para los modelos de regresión logística multinomial: el Estadístico de Wald y el Estadístico condicional de razón de verosimilitudes.

Así que nos planteamos contrastar si un subconjunto de los parámetros del modelo de regresión logística multinomial, que denotaremos por $\beta = (\beta_1, \dots, \beta_r)'$, es nulo, es decir, se realiza el siguiente contraste de hipótesis:

$$H_0: \beta = 0$$
$$H_1: \beta \neq 0$$

Veamos los dos tipos de contrastes mencionados anteriormente que se utiliza para decidir si rechazamos o no la hipótesis nula.

1.4.1 Contraste de Wald.

Se basa en la normalidad asintótica de los estimadores de máxima verosimilitud. El estimador de máxima verosimilitud de β, $\hat{\beta}$, tiene distribución normal asintótica de media β y matriz de covarianzas estimada $\widehat{cov}(\hat{\beta})$ obtenida a partir de la matriz de covarianza $cov(\hat{\beta})$. Así que el estadístico de Wald presenta la forma cuadrática: $\hat{\beta}'\left[\widehat{cov}(\hat{\beta})\right]^{-1}\hat{\beta}$, que tiene distribución chi-cuadrado asintótica con r grados de libertad (r número de parámetros nulos bajo la hipótesis nula).

Así que se rechaza la hipótesis nula al nivel de significación α cuando el valor observado de este estadístico sea mayor o igual que el cuantil de orden (1- α) de la distribución χ_r^2.

Su valor para un coeficiente concreto viene dado por el cociente entre el valor del coeficiente y su correspondiente error estándar. Es decir si se quiere contrastar:

$$H_0: \beta_{pj} = 0$$
$$H_1: \beta_{pj} \neq 0$$

El estadístico será: $W = \frac{\hat{\beta}_{pj}^2}{\hat{\sigma}^2(\hat{\beta}_{pj})}$, que tiene distribución chi-cuadrado asintótica con un grado de libertad. Así que se rechaza la hipótesis nula con un nivel de confianza 1- α si

$W_{Obs} \geq \chi^2_{1;1-\alpha}$. En caso de rechazar la hipótesis nula indica que dicho coeficiente es diferente de 0 y merece la pena su conservación en el modelo.

En modelos con errores estándar grandes, el estadístico de Wald puede proporcionar falsas ausencias de significación. Tampoco es recomendable su uso si se está empleando variables de diseño. En estos casos se recomienda el uso del test de razón de verosimilitudes.

1.4.2 Contrastes condicionales de razón de verosimilitud.

Se trata de ir contrastando cada modelo que surge de eliminar de forma aislada cada una de las covariables frente al modelo completo. La ausencia de significación implica que el modelo sin la covariable no empeora respecto al modelo completo (Es decir, da igual su presencia que su ausencia), por lo que según la estrategia de obtención del modelo más reducido (principio de parsimonia), dicha covariable debe ser eliminada del modelo ya que no aporta nada al mismo.

Supongamos que tenemos un modelo de regresión logística multinomial M_G que se ajusta bien y se desea contrastar si un subconjunto de parámetros, $\beta = (\beta_1, \dots, \beta_r)'$, son nulos. Sea M_p el modelo con ese subconjunto de parámetros ceros. Así que M_p está anidado en el modelo general M_G. Así que planteamos el contraste:

$$H_0: \beta = 0 \ (M_P \ se\ verifica\ )$$
$$H_1: \beta \neq 0 \ (\text{asumiendo cierto } M_G\ )$$

Si asumimos que M_G se verifica, el estadístico del test de razón de verosimilitud para contrastar si M_p se verifica es: $G^2(M_P|M_G) = -2(L_P - L_G) = G^2(M_P) - G^2(M_G)$, siendo L_p y L_G los máximos de la log-verosimilitud bajo la suposición de que se verifican los modelos saturados, M_p y M_G, respectivamente. Es decir, el test de razón de verosimilitudes para contrastar dos modelos anidados es la diferencia de los contrastes de razón de verosimilitudes de bondad de ajuste para cada modelo.

El estadístico $G^2(M_P|M_G)$ tiene distribución chi-cuadrado con grados de libertad la deferencia entre los grados de libertad de las distribuciones chi-cuadrado asintóticas de $G^2(M_P)$ y $G^2(M_G)$, es decir, el número de parámetros que se anulan para H_0, estos es r.

Así que se rechaza la hipótesis nula al nivel de significancia α cuando $G^2_{Obs}(M_P|M_G) \geq \chi^2_{r;\alpha}$.

1.5 Inferencia en regresión logística multinomial.

Lo principal que se pretende cuando se realiza un modelo estadístico a través de los datos procedentes de una muestra, es extrapolar los resultados muestrales a la población general, es por ello que, para nuestro caso particular de haber estimado los parámetros del modelo de regresión logística multinomial pretendemos hacer inferencia.

1.5.1 Intervalos de confianza.

Basándonos en la normalidad asintótica de los estimadores de máxima verosimilitud se pueden construir intervalos de confianza asintóticos para cada uno de los parámetros del modelo, utilizando la distribución normal, y mediante las transformaciones correspondientes, intervalos de confianza para los odds_ratio.

- **Intervalos de confianza para los parámetros.**

Construimos un intervalo de confianza 1- α para cada parámetro del modelo de regresión logística multinomial, β_{sj} con j=1,…, k-1, s=0,…, q. La distribución asintótica de $\hat{\beta}_{sj}$ es $N\left(\beta_{sj}, \hat{\sigma}^2(\hat{\beta}_{sj})\right)$, donde $\hat{\sigma}^2(\hat{\beta}_{sj})$ es el valor correspondiente al error estándar del estimador del parámetro β_{sj}.

Así que tenemos que: $P\left[-z_{\alpha/2} \leq \frac{\hat{\beta}_{sj}-\beta_{sj}}{\hat{\sigma}(\hat{\beta}_{sj})} \leq z_{\alpha/2}\right] = 1 -$ α. Por lo que obtenemos así el intervalo de confianza aproximado para β_{sj} al nivel $1 -$ α:

$$IC(\beta_{sj}) = \left(\hat{\beta}_{sj} \pm z_{\alpha/2}\hat{\sigma}(\hat{\beta}_{sj})\right) \tag{1.34}$$

- **Intervalos de confianza para los odds_ratio.**

Sabemos que los cocientes de ventajas vienen dados por:

$Odds\_ratio_j\left((\Delta X_r = 1)\middle|X_s = x_s, s \neq r\right) = \exp(\beta_{rj}) \quad \forall\, r = 1, \dots q;\ j = 1, \dots, k-1.$

Por lo tanto, el intervalo de confianza para los cocientes de ventajas se calcula tomando exponenciales en el intervalo de confianza obtenido anteriormente para cada uno de los (β_{pj}) al nivel de confianza $1 - \alpha$, viene dado por:

$$IC(\exp(\beta_{pj})) = exp\left(\hat{\beta}_{pj} \pm z_{\alpha/2}\hat{\sigma}(\hat{\beta}_{pj})\right) \quad (1.35)$$

1.6 Bondad de Ajuste del Modelo.

1.6.1 Contrastes de bondad de ajuste del modelo.

Uno de los primeros indicadores de importancia para apreciar el ajuste del modelo logístico multinomial es el doble logaritmo del estadístico de verosimilitud (likelihood), que veremos posteriormente. Se trata de un estadístico que sigue una distribución similar a una χ^2.

Sea y_{ij} el número de observaciones que caen en la categoría de respuesta Y_j $\forall$ j=1,2,…, k. y sean las m_i observaciones correspondientes a la i-ésima combinación de valores de las variables explicativas.

Denotaremos por $\hat{d}_{ij}$ la frecuencia esperada de respuesta Y_j en la combinación de x_i de valores observados de las variables predictoras, estimada bajo el modelo y definida como $\hat{d}_{ij} = m_i\,\hat{\pi}_{ij}$, siendo $\hat{\pi}_{ij}$ el estimador por máxima verosimilitud de π_{ij}.
Así que para contrastar la bondad del ajuste global del modelo cuando el número de observaciones en cada combinación de valores de las variables explicativas es grande se utiliza el estadístico chi-cuadrado de Pearson y el estadístico de Wilks de razón de verosimilitudes.
El test global de bondad de ajuste del modelo de regresión logística multinomial múltiple contrasta la siguiente hipótesis:

$$H_0: \pi_{ij} = \frac{e^{(\Sigma_{p=0}^{q}\beta_{pj}x_{ip})}}{1+\sum_{j=1}^{k-1}e^{(\Sigma_{p=0}^{q}\beta_{pj}x_{ip})}} \;\forall\, i = 1, \ldots..n; \quad \forall\, j = 1, \ldots., k$$

$$H_1: \pi_{ij} \neq \frac{e^{(\Sigma_{p=0}^{q}\beta_{pj}x_{ip})}}{1+\sum_{j=1}^{k-1}e^{(\Sigma_{p=0}^{q}\beta_{pj}x_{ip})}} \text{ Para algún i, j.}$$

Con lo cual H_0 trata de comprobar que los datos multinomiales se ajustan a un modelo de regresión logística multinomial.

Para realizar el contraste anterior se utilizan los siguientes test.

1.6.1.1. Test chi-cuadrado de Pearson.

El estadístico chi-cuadrado de Pearson de bondad de ajuste a un modelo de regresión logística multinomial M, de la forma anterior viene dado por:

$$\chi^2(M) = \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(y_{ij}-m_i\hat{\pi}_{ij})^2}{m_i\hat{\pi}_{ij}} \tag{1.36}$$

Este estadístico tiene distribución asintótica chi-cuadrado con grados de libertad obtenidos como la diferencia entre el número de parámetros π_{ij} y el número de parámetros independientes en el modelo, (n-q-1)(k-1). Es decir, $\chi^2(M) \xrightarrow{d} \chi^2_{(n-q-1)(k-1),1-\alpha}$ si $m_i \to \infty$.

Así que se rechaza la hipótesis nula con un nivel de significación de α cuando $\chi^2(M)_{Obs} \geq \chi^2_{(n-q-1)(k-1),1-\alpha}$ o equivalentemente podemos definir el p_valor del contraste como la probabilidad acumulada a la derecha del valor observado: $p\_valor = P\left[\ \chi^2(M)_{Obs} \geq \chi^2_{(n-q-1)(k-1),1-\alpha}\ \right]$, se rechaza la hipótesis nula cuando el p_valor ≤ α.

1.6.1.2 Test chi-cuadrado de razón de verosimilitudes. Estadístico de Wilks. Devianza.

El estadístico de Wilks de razón de verosimilitudes para el contraste de bondad de ajuste del modelo de regresión logística multinomial M se obtiene como menos dos veces el logaritmo del cociente entre el supremo de la verosimilitud bajo la hipótesis nula y el supremo de la verosimilitud en la población.

Vamos a deducir la región crítica y la expresión del estadístico desviación (test de la razón de verosimilitudes) para el contraste H_0.

Sea $\pi = (\pi_{11}, \dots, \pi_{1(k-1)}, \dots, \pi_{n1}, \dots, \pi_{n(k-1)})$. El espacio paramétrico es:

$$\Theta^{(1)} = \left\{ \pi \in \mathcal{R}^{n(k-1)} / \pi_{ij} \in (0,1), \sum_{r=1}^{k-1} \pi_{ir} \leq 1, \qquad i = 1, \dots, n, \qquad j = 1, \dots, k-1 \right\}$$

El subespacio paramétrico que define la hipótesis nula es

$$\Theta_0^{(1)} = \left\{ \pi \in \mathcal{R}^{n(k-1)} / \pi_{ij} = \frac{\exp\{\beta_{0j} + \sum_{p=1}^{q} \beta_{pj} x_{ipj}\}}{1 + \sum_{j=1}^{k-1} \exp\{\beta_{0j} + \sum_{p=1}^{q} \beta_{pj} x_{ipj}\}}, i = 1, \dots, n, \right.$$

$$\left. j = 1, \dots, k-1 \right\}$$

Sean $Y_1 = y_1, \dots, Y_n = y_n$, la función de logverosimilitud no restringida es

$$\ell^{(1)} = \ell^{(1)}(\pi; y_1, \dots, y_n) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{k-1} y_{ij} ln \frac{\pi_{ij}}{\pi_{ik}} + m_i ln \pi_{ik} + ln \frac{m_i!}{y_{i1}!, \dots, y_{ik}!} \right\} \quad (1.37)$$

Con $\pi_{ik} = 1 - \sum_{j=1}^{k-1} \pi_{ij}$. Derivando parcialmente (1.37) respecto de π_{ij} e igualando a cero, se obtiene:

$$0 = \frac{\partial \ell^{(1)}}{\partial \pi_{ij}} = y_{ij} \left(\frac{1}{\pi_{ij}} + \frac{1}{\pi_{ik}} \right) + \sum_{r=1, r \neq j}^{k-1} \frac{y_{ir}}{\pi_{ik}} - \frac{m_i}{\pi_{ik}} = \frac{y_{ij}}{\pi_{ij}} + \frac{m_i - y_{ik}}{\pi_{ik}} - \frac{m_i}{\pi_{ik}} = \frac{y_{ij}}{\pi_{ij}} - \frac{y_{ik}}{\pi_{ik}} \quad (1.38)$$

Con lo cual $y_{ij} \pi_{ik} = \pi_{ij} y_{ik}$, y sumando en j=1,…,k se obtiene

$$\pi_{ik} = \frac{y_{ik}}{m_i}, \qquad \pi_{ij} = \frac{y_{ij}}{m_i}, \qquad j = 1, \dots, k-1$$

Así pues $\hat{\pi}_{ij} = \frac{y_{ij}}{m_i}$, es el EMV de π_{ij} en el modelo saturado.

Sea $\beta = (\beta_1, \dots, \beta_{k-1}) = (\beta_{10}, \beta_{11}, \dots, \beta_{1q}, \dots, \beta_{k-1,0}, \beta_{k-1,1}, \dots, \beta_{k-1,q})$, la función de logverosimilitud restringida a $\Theta_0^{(1)}$ es:

$$\ell_0^{(1)} = \ell_0^{(1)}(\beta; y_1, \dots, y_n) = \sum_{i=1}^{n} \left\{ \sum_{j=1}^{k-1} y_{ij} \left(\beta_{0j} + \sum_{p=1}^{q} \beta_{pj} x_{ipj} \right) - m_i \ln\left(1 + \right.\right.$$

$$\left.\left. \sum_{j=1}^{k-1} \{ \beta_{0j} + \sum_{p=1}^{q} \beta_{pj} x_{ipj} \} \right) + ln \frac{m_i!}{y_{i1}!, \dots, y_{iq}!} \right\} \quad (1.39)$$

Aplicando el algoritmo de las puntuaciones de Fisher se obtienen los estimadores de máxima verosimilitud $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_{k-1})$. El test de la razón de verosimilitudes para contrastar $H_0^{(1)}$ usa el estadístico desviación:

$$S_n^{(1)} = 2 \left[\ell^{(1)}(\hat{\pi}; y_1, \dots, y_n) - \ell_0^{(1)}(\hat{\beta}; y_1, \dots, y_n) \right]$$

$$= 2 \sum_{i=1}^{n} \left\{ \sum_{j=1}^{k-1} y_{ij} \left(ln \frac{y_{ij}}{y_{ik}} - \hat{\beta}_{j0} - \sum_{p=1}^{q} \hat{\beta}_{jp} x_{ijp} \right) + m_i \left[ln \frac{y_{ik}}{m_i} + ln\left(1 + \sum_{j=1}^{k-1} \{ \hat{\beta}_{j0} + \right.\right.\right.$$

$$\left.\left.\left. \sum_{p=1}^{q} \hat{\beta}_{jp} x_{ijp} \} \right) \right] \right\} \quad (1.40)$$

Se rechaza la hipótesis nula si $S_n^{(1)} > \chi^2_{(n-q-1)(k-1),1-\alpha}$ o equivalente cuando el $p\_valor = P\left[S_n^{(1)} \geq \chi^2{}_{(n-q-1))(k-1),1-\alpha}\right] \leq \alpha$.

1.6.2 Calidad del Ajuste.

Además de los contrastes que hemos visto anteriormente, podemos calcular otras medidas que nos dan información sobre la calidad del modelo. Por ejemplo en los modelos de regresión logística binaria, la calidad del ajuste se mide mediante coeficientes de determinación conocidos como Pseudo- R^2, para la regresión logística multinomial también se utilizan estos coeficientes. De entre todos los que existen, los más usados son el de Mc-Fadden, el de Cox-Snell y el de Nagelkerke. Veamos cómo se calculan cada uno de ellos.

1.6.2.1 Coeficiente pseudo- R^2 de Mc-Fadden.

Si tenemos $\Lambda = -2\ln(L(\beta; y_1, \dots, y_n))$, identificamos por Λ_0 el valor inicial de esta función, es decir el mínimo Λ bajo el modelo nulo dado solo por un término constante de la siguiente forma:

$$\Lambda_0 = -2\ln(L_0(\beta; y_1, \dots, y_n)) = -2\sum_{i=1}^{n}\left\{\sum_{j=1}^{k-1} y_{ij} \ln\left(\frac{\pi_{ij}(\beta)}{\pi_{ik}(\beta)}\right) + m_i \ln(\pi_{ik}(\beta)) + \ln\left(\frac{m_i!}{y_{i1}!\dots y_{ik}!}\right)\right\}$$

Con $\theta_{ij} = \ln\left(\frac{\pi_{ij}}{\pi_{ik}}\right) = \beta_{0j}$.

Y por Λ_f el mínimo de Λ bajo el modelo ajustado con todos los parámetros, y se define de la siguiente forma:

$$\Lambda_f = -2\ln(L(\beta; y_1, \dots, y_n)) = -2\sum_{i=1}^{n}\left\{\sum_{j=1}^{k-1} y_{ij} \ln\left(\frac{\pi_{ij}(\beta)}{\pi_{ik}(\beta)}\right) + m_i \ln(\pi_{ik}(\beta)) + \ln\left(\frac{m_i!}{y_{i1}!\dots y_{ik}!}\right)\right\}$$

Con $\theta_{ij} = \ln\left(\frac{\pi_{ij}}{\pi_{ik}}\right) = \beta_{0j} + \sum_{p=1}^{q}\beta_{pj}x_{ipj}$

Dado esto se obtiene la siguiente expresión del pseudo-R^2 de Mc-Fadden:

$$R^2_{MF} = 1 - \frac{\Lambda_f}{\Lambda_0}. \tag{1.41}$$

Siendo su rango teórico de valores $0 \leq R^2_{MF} \leq 1$, pero muy raramente su valor se aproxima a 1. Suele considerarse una buena calidad del ajuste cuando $0.2 \leq R^2_{MF} \leq 0.4$ y excelente para valores superiores, siendo malo para valores menores a 0.2.

1.6.2.2 Coeficiente pseudo-R^2 de Cox-Snell.

En este caso se utiliza directamente la función de verosimilitud $L(\beta; y_1, \dots, y_n)$, y no la función auxiliar Λ. Por lo que si denotamos por $V_0 = \exp\left(-\frac{\Lambda_0}{2}\right)$ el máximo de verosimilitud bajo el modelo nulo dado solo por un término constante y por $\mathrm{V}_f = \exp\left(-\Lambda_f/2\right)$ el máximo de verosimilitud bajo el modelo ajustado con todos los parámetros, definimos el coeficiente pseudo-R^2 de Cox-Snell como:

$$R^2_{CS} = 1 - \left(\frac{\mathrm{V}_0}{\mathrm{V}_f}\right)^{\frac{2}{N}} = 1 - \exp\left(\frac{\Lambda_f - \Lambda_0}{\mathrm{N}}\right) \tag{1.42}$$

El rango teórico de valores para el coeficiente es $0 \leq R^2_{CS} \leq 1 - \mathrm{V}_0^{\frac{2}{N}}$, lo que le hace poco interpretable al depender de V_0. Ya que puede ser próximo a cero cuando hay pocos datos. Por ello es preferible utilizar el siguiente coeficiente como medida de bondad de ajuste.

1.6.2.3 Coeficiente pseudo-R^2 de Nagelkerke.

Viene dado por la siguiente expresión:

$$R^2_N = \frac{R^2_{CS}}{1 - \mathrm{V}_0^{\frac{2}{N}}} = \frac{1 - \exp\left(\frac{\Lambda_f - \Lambda_0}{\mathrm{N}}\right)}{1 - exp\left(\frac{-\Lambda_0}{N}\right)} \tag{1.43}$$

Y en este caso, su rango de valores es de $0 \leq R^2_N \leq 1$, por lo que puede interpretarse del mismo modo que el coeficiente de determinación de la regresión lineal clásica, aunque es más difícil que alcance valores cercanos a 1.

1.6.3. Tasa de clasificaciones correctas.

Para cuantificar la bondad del ajuste global del modelo se dispone también de otra medida como es la tasa de clasificaciones correctas. Es decir, a partir del modelo ajustado, se clasifica cada observación en la categoría más probable, construyendo así una matriz de clasificación: ***observados-predichos*** y se utiliza el porcentaje de clasificaciones correctas como una medida de calidad de predicción, del mismo modo que se hace en el análisis discriminante. Se define como la proporción de individuos

clasificados correctamente por el modelo y se calcula como el cociente entre el número de observaciones clasificadas correctamente y el tamaño muestral N.
Un individuo es clasificado correctamente por el modelo cuando su valor observado de la variable respuesta Y (Y_1, Y_2,..,Y_k) coincide con su valor estimado por el modelo.

1.7. Métodos de selección del modelo.

Una vez conocido el procedimiento de ajuste de modelos de regresión logística multinomial, el siguiente paso es el desarrollo de estrategias para seleccionar las variables que mejor explican a la variable de respuesta. Para ello se adoptará el principio de parsimonia que consiste en seleccionar el modelo que con menor número de parámetros se ajuste bien a los datos y lleve a una interpretación sencilla en términos de cocientes de ventajas.
Hay que tener especial atención a las covariables cualitativas que se transforman en varias variables dummies. Siempre que se incluya o excluya una de estas variables, todas las demás categorías deben ser incluidas o excluidas en bloque.

Si no se tiene en cuenta esta consideración, implicaría que se habría recodificado la variable, y por tanto la interpretación de la misma no sería la correcta. Además, hay que tener en cuenta la significación que pudiera tener cada variable dummy. No siempre todas las categorías de una covariable son significativas, o todas no significativas. Por lo que, cuando ocurra esta situación es recomendable contrastar el modelo completo frente al modelo sin la covariable mediante la prueba de razón de verosimilitud, decidiendo incluir o excluir la covariable dependiendo del resultado de la prueba. Si se obtiene significación en este contraste, la variable permanecería en el modelo.

1.7.1. Selección Hacia adelante.

1. Se inicia con un modelo vacio (sólo constante).
2. Se ajusta un modelo y se calcula el p-valor del contraste de razón de verosimilitud que resulta de incluir cada variable por separado.
3. Se selecciona el modelo con el p-valor más significativo.
4. Se ajusta de nuevo un modelo con la(s) variable(s) seleccionada(s) y se calcula el p-valor de añadir cada variable no seleccionada anteriormente por separado.
5. Se selecciona el modelo con el más p_valor significativo.

6. Se repite 4 – 5 hasta que no queden variables significativas para incluir.

1.7.2. Selección Hacia atrás.

1. Se inicia con un modelo con todas las variables candidatas.
2. Se eliminan, una a una, cada variable y se calcula la pérdida de ajuste al eliminar.
3. Se selecciona para eliminar la menos significativa.
4. Se repite 2 – 3 hasta que todas las variables incluidas sean significativas y no pueda eliminarse ninguna sin que se pierda ajuste.

1.7.3. Selección Stepwise (Paso a Paso).

En este modelo se combinan los métodos adelante y atrás. Puede empezarse por el modelo vacío o por el completo, pero en cada paso se exploran las variables incluidas, por si deben salir y las no seleccionadas, por si deben entrar. Pero no todos los métodos llegan a la misma solución necesariamente. El modelo de stepwise, está basado en contrastes condicionales de razón de verosimilitudes.

Si partimos del modelo vacío, sólo con la constante, este método consiste en partir de ese modelo inicial, y en cada paso se ajustarán todos aquellos modelos que resultan de incluir cada una de las variables explicativas que no están en el modelo seleccionado en el paso anterior. Entonces se llevan a cabo contrastes condicionales de razón de verosimilitudes que tiene en la hipótesis nula el modelo seleccionado en el paso anterior y en la hipótesis alternativa el modelo resultante de la inclusión de cada variable. De este modo se seleccionarán las variables para las que el contraste sea significativo, y se incluiría en el modelo aquella variable asociada al mínimo p-valor de entre todos los menores o iguales que α_1. La inclusión de variable mediante este método continua hasta que ninguno de estos contrastes condicionales sea significativo.

Por otra parte, a la misma vez, se considera en cada paso la posibilidad de eliminar alguno de los parámetros del modelo seleccionado en el paso anterior (método hacia atrás). Pero no se puede eliminar en un paso la variable que acaba de entrar en el paso anterior, por lo que se fijará para la eliminación de variables un nivel de significación α_2 mayor que α_1. Al igual que antes, para la eliminación de variables se realizarán contrastes condicionales de razón de verosimilitudes que tienen en la hipótesis nula el

modelo que resulta de la eliminación de cada variable y en la hipótesis alternativa el modelo seleccionado en el paso anterior. Así, las variables candidatas a eliminar serán aquellas cuyo p-valor sea mayor de α_2 y se eliminará la variable con el mayor p-valor de éstos. La eliminación de variables continúa hasta que todos estos contrastes condicionales resulten significativos.

Así finalmente, se llegará a un paso en el que ninguno de los contrastes condicionales de introducción de variables sean significativos y todos los de eliminación de variables sean significativos.

CAPÍTULO II. REGRESIÓN LOGÍSTICA MULTINOMIAL PARA CLASIFICAR LOS HOGARES DE EL SALVADOR POR NIVEL DE POBREZA.

Introducción.

En este Capítulo II nos dedicamos a llevar a la práctica la teoría explicada en el Capítulo I, es decir, la aplicación del método de regresión logística multinomial, para la clasificación de un hogar en el nivel de pobreza adecuado en virtud de las características demográficas y socioeconómicas del hogar y su jefatura, para ello se requiere tener en cuenta varios aspectos previos para la preparación de la base de datos, entre ellos se incluyen la selección tanto de la variable dependiente como las variables independientes, el tamaño de la muestra para la estimación del modelo, la partición de la muestra con fines de validación, además de explicar la información primaria que se utilizará para la realización de ésta aplicación, donde la base de datos a utilizar es la proveniente de la encuesta de hogares de propósitos múltiples del año 2010.

2.1. Características de la Encuesta de Hogares de Propósitos Múltiples del año 2010.

La principal fuente de información empleada para la realización de este trabajo es la Encuesta de Hogares de Propósito Múltiples (EHPM), captada y procesada por la Dirección General de Estadísticas y Censos (DIGESTYC). Desde 1975 ésta encuesta constituye la fuente de información oficial sobre los principales indicadores socioeconómicos de la población salvadoreña. Desde el año 2000 dicha encuesta se realiza anualmente.

La Encuesta de Hogares de Propósitos Múltiples es un instrumento estadístico con que cuenta el país, para obtener diagnósticos de su situación, para implementar acciones apropiadas a favor de su desarrollo y por otro lado, facilitar el seguimiento de los efectos que producen las medidas de políticas adoptadas.

La EHPM del año 2010 se desarrolló entre los meses de enero a diciembre, con una muestra de 19,968 viviendas obteniendo representatividad a nivel de país: por zona urbana y rural, área metropolitana de San Salvador (AMSS), departamental y de los 50 municipios más grandes del país.

El objetivo general de la EHPM es generar información estadística actualizada, tanto cualitativa como cuantitativa, relacionada con las condiciones socioeconómicas y demográficas de la población salvadoreña, para facilitar el diseño o rediseño de políticas, planes, programas y proyectos que desarrollan las instituciones públicas, que contribuyan a elevar el bienestar de la población y que a la vez sea de utilidad a otros organismos nacionales e internacionales para los mismos propósitos.

La Encuesta de Hogares de Propósitos Múltiples (EHPM) se diseñó como un sistema continuo de encuestas, basada en submuestras mensuales representativas del país, con aplicación de ocho módulos permanentes, y dos módulos especiales: Uso del Tiempo y Monitoreo de educación.

Se investigaron durante todo el período de la encuesta, rubros básicos sobre población (características personales de los miembros de los hogares, edad, sexo, parentesco, etc.), educación, empleo e ingreso, salud, remesas familiares y gastos del hogar.

Se incluyeron los siguientes módulos de la encuesta:

Cuadro 2.1: *Módulos de la Encuesta.*

SECCIÓN	TEMÁTICAS
1	Características Sociodemográficas
2	Características de Educación
2A	Tecnología de Información y Comunicación
3	Características generales de la Vivienda
4	Empleo e Ingreso
5	Actividad del Productor Agropecuario
6	Salud
7	Remesas Familiares
8	Gasto del hogar

2.2. Variables consideradas en la clasificación de hogares.

2.2.1. Variable Dependiente.

Condición de pobreza: La condición de pobreza calculada a partir de la EHPM 2010, clasifica a los hogares de El Salvador en tres grupos diferenciados, pobres extremos, pobres relativos y no pobres. Esta clasificación se realiza mediante la construcción de una Canasta Básica Alimentaria (línea de pobreza) la cual constituye un punto referencial cuyo "estatus de pobreza o no" se encuentran por encima y por debajo de él; a su vez esta línea de pobreza tiene dos componentes que clasifica a los hogares pobres en "pobres relativos" y "pobres extremos".

2.2.2. Variables Independientes.

- **Características Individuales.**

Sexo del(la) jefe(a) de hogar: Un factor importante a considerar es el sexo del jefe del hogar, cuando el jefe de hogar es hombre la inserción laboral suele darse con mayor facilidad que cuando es mujer; además se encuentran en condiciones de acceder a mejores puestos de trabajo en comparación a las jefas de hogar ya que en su mayoría se ven libres de tener que asumir labores domésticos porque encuentran en su pareja el apoyo necesario para cumplir con dicha actividad. Para las jefas de hogar, la inserción laboral es más difícil, más aún si son madres solteras o divorciadas porque además de trabajar tienen que asumir la responsabilidad de atender a su familia, que las coloca en una situación de desventaja para acceder a mejores oportunidades de trabajo; en otros casos las mujeres son víctimas de discriminación laboral, y se ven limitadas en adquirir activos que les dotaría de nuevas y mejores posibilidades de empleo. Se incluye esta variable en el estudio, a fin de conocer las diferencias relevantes en la condición de pobreza cuando un hogar se encuentra jefaturado por un hombre o una mujer, y si es causa importante del estatus de pobreza del hogar en presencia de otros activos.

Edad del(la) jefe(a) de hogar: Es una característica importante a considerar en este estudio, para determinar si la condición de pobreza depende de la edad del(la) jefe(a) del hogar además de su inserción laboral en el mercado, pues algunos estudios han puesto en evidencia que los adultos mayores son miembros que contribuyen de manera

efectiva a la economía del hogar. Mientras que otros afirman que el estatus de pobreza suele ser precaria cuando los jefes de hogar son muy jóvenes o muy adultos.

Estado Familiar: El estado familiar es una característica muy importante, ya que contar con una pareja o no puede influir en la superación económicamente del hogar.

Nivel de estudio aprobado del(la) jefe(a) de hogar: La educación como capital cultural es fundamental para el desarrollo de las personas pues es un activo que permite generar o acceder a nuevas oportunidades laborales que incidirán en un mejor estatus de vida para el hogar, más aún cuando se es jefe de hogar.

Ocupado: Los bajos ingresos y las condiciones desfavorables en las que se insertan los jefes de hogar al mundo laboral, los apremios económicos por los que tienen que pasar y la responsabilidad de dotar de recursos al hogar impulsan a los jefes de hogar a buscar fuentes de ingresos tal que les permita hacer frente a la pobreza. Por tal motivo se ha incluido en el estudio, la variable ocupación del jefe de hogar a fin de evaluar si el hecho de que el jefe de hogar se encuentre con un empleo actúa como un factor reductor de la condición de pobreza de los hogares salvadoreños.

Es asegurado(a) el(la) jefe(a) de hogar: Un elevado porcentaje de la población mundial ha sufrido padecimientos de salud imprevistos, que generan grandes costos, muchas veces difíciles de afrontar. El seguro médico es un complemento valioso a la salud de cada persona individual o perteneciente a una familia u otro grupo social, porque suple económicamente parte de los auxilios o servicios.

- **Características del hogar.**

Total de Miembros en el hogar: En distintas investigaciones sociales realizadas en América Latina se ha determinado que el tamaño del hogar es un componente que incide en la condición de pobreza de los hogares, pues, se ha detectado que a mayor número de miembros, mayor es la posibilidad de ser pobre, por eso se considera importante la inclusión de esta variable en el estudio.

Ingreso Familiar: Esta variable recoge información esencial para la clasificación de los hogares en pobreza o no.

- **Servicios públicos.**

Disponibilidad de energía eléctrica: Proveer al hogar de energía eléctrica favorece a sus integrantes, especialmente a los que están en edad escolar, poder realizar sus actividades en condiciones más adecuadas (ambientes más iluminados) así como de acceder a nuevas tecnologías de comunicación e información, asimismo su uso favorece a la generación de nuevas fuentes de trabajo apoyados en este bien. En este caso la variable capta la información de la disponibilidad o no de alumbrado eléctrico en la vivienda, la cual conjuntamente con las demás características ayudará a determinar el estatus de pobreza del hogar

Disponibilidad de servicios higiénicos (Alcantarillado): la dotación de este activo a la vivienda se traduce en mejoras en su infraestructura, lo cual favorece al bienestar del hogar, cuando los servicios higiénicos están conectados a un sistema de eliminación de excretas se reduce el riesgo que la familia contraiga enfermedades infecciosas ya que las fuentes naturales de abastecimiento de agua como ríos y manantiales se ven protegidos ante la posibilidad de que los desechos humanos sean arrojados en sus cauces y se conviertan en focos de transmisión de enfermedades.

Disponibilidad de agua: Disponer de algunos servicios públicos básicos como el acceso a agua potable coloca a los hogares en situación de menor vulnerabilidad frente a la pobreza; esto debido a que al disponer de agua en condiciones salubres se reduce el riesgo de que los miembros del hogar, especialmente los más pequeños, sufran de enfermedades infecciosas; así mismo significa ahorro de tiempo para las madres, pues ya no tendrán que emplear parte de su tiempo en abastecerse de este líquido y tendrán la posibilidad de emplear su tiempo de manera más productiva.

Tenencia de teléfono: El teléfono constituye un activo de gran utilidad cuando se trata de propiciar cambios en los hogares pues al igual que la energía eléctrica es un medio de integración poderoso y una herramienta útil y complementaria en el desempeño de labores de carácter productivo y de servicios.

- **Capital institucional**

Propiedad de la vivienda: Poseer derechos sobre la vivienda resulta ser ventajoso para los integrantes de un hogar pues se convierte en un recurso valioso que le permitirá acceder a créditos o crear microempresas en su domicilio.

- **Ubicación Geográfica**

Región: para una mejor cobertura geográfica, el país se divide en cinco regiones, estando constituida por la región Occidental, Central I, Centra II, Oriental, AMSS. De acuerdo a esto podríamos observar la(s) Región(es) que podrían incidir en los diferentes niveles de pobreza.

Área de ubicación del hogar: debido a que la pobreza se mide de acuerdo a la canasta básica alimentaria, y ésta no es igual para el área urbana y rural, se desearía tener el peso que influye esta variable para la clasificación de la pobreza.

- **Patrimonio**

Se puede pensar que un hogar que posea más bienes como: **lavadora, refrigeradora, computadora y microonda,** reduciría la probabilidad de ser pobre. En tal sentido han sido tomadas estas variables en la investigación.

Cuadro 2.2: *Variables y sus Códigos Consideradas en la estimación del modelo.*

VARIABLES		CÓDIGOS Y CATEGORÍAS
Variable Dependiente		
Condición de Pobreza	**POBREZA**	*1- Pobre Extremo*
		2- Pobre Relativo
		3- No Pobre
Características Individuales		
Sexo del Jefe(a) de Hogar	**SEXO**	*1- Mujer*
		2- Hombre
Edad del Jefe(a) de Hogar	**EDAD**	*1- Si 15<=Edad<=20*
		2- Si 21<=Edad<=30
		3- Si 31<=Edad<=40
		4- Si 41<=Edad<=50
		5- Si 51<=Edad<=65
		6- Si Edad>=65
Estado Familiar del Jefe(a) de Hogar	**ESTAFAMILIAR**	*1- Acompañado (a)*
		2- Casado (a)
		3- Viudo (a)
		4- Divorciado (a)
		5- Separado (a)
		6- Soltero (a)
Educación del Jefe(a) de Hogar	**EDUCACION**	*1- Básica(1º a 9º)*
		2- Media(10º a 13º)
		3- universitario(1º a 15º) y no universitario(1º a 3º)
		4- Parvularia o Ninguno
Es Ocupado el Jefe(a) de Hogar	**OCUPACION**	*1- SI*
		2- NO
Es asegurado el Jefe(a) de Hogar	**SEGURO**	*1- SI*
		2- NO
Características del Hogar		
Ingreso Familiar	**INGREFA**	*Valor Real Positivo*
Número de Miembros del Hogar	**MIENH**	*Valor Entero Positivo*
Servicios Públicos		
Cuenta el Hogar con Energía Eléctrica	**ELECTRICIDAD**	*1- SI*
		2- NO
Cuenta el Hogar con Agua Potable	**AGUA**	*1- SI*
		2- NO
Cuenta el Hogar con Alcantarillado	**ALCANTARILLADO**	*1- SI*
		2- NO
Cuenta el Hogar con Teléfono	**TELEFONO**	*1- SI*
		2- NO
Ubicación Geográfica		
Región de Ubicación	**REGION**	*1- Occidental*
		2- Central 1
		3- Central 2
		4- Oriental
		5- AMSS
Área de Ubicación	**AREA**	*1- Urbana*
		2- Rural
Capital Institucional		
Propiedad de la Vivienda	**TENENCIA**	*1- SI*
		2- NO
Patrimonio		
Cuenta el Hogar con Refrigeradora	**REFRIGERADORA**	*1- SI*
		2- NO
Cuenta el Hogar con Lavadora	**LAVADORA**	*1- SI*
		2- NO
Cuenta el Hogar con Computadora	**COMPUTADORA**	*1- SI*
		2- NO
Cuenta el Hogar con Microonda	**MICROONDA**	*1- SI*
		2- NO

2.3. Tamaño Muestral.

La base de datos de la Encuesta de Hogares de Propósitos Múltiples (EHPM) del 2010, contiene información para 21,166 hogares encuestados, de los cuales se hizo una selección de las variables de interés; luego se procedió a seleccionar de dicha base sólo aquellos que son jefes(as) de hogar, eliminando los casos perdidos de las variables Nivel de estudios aprobados (EDUCACIÓN) y Condición de Actividad (OCUPACIÓN) que eran las únicas que presentaban esos casos, teniendo como resultado 16,387 Jefes (as) de Hogar.

2.4. Partición de la Muestra.

La muestra se dividió en dos submuestras; una utilizada para la estimación del modelo de regresión logística multinomial, y otra con fines de validación. Para ello se consideró que las submuestras tuvieran el tamaño adecuado para apoyar las conclusiones de los resultados.

El procedimiento utilizado para la partición de la muestra se denomina: División de la muestra o enfoque de validación cruzada, que consiste en desarrollar el modelo de regresión logística multinomial en un grupo y luego probarla con un segundo grupo. Este procedimiento consistió en dividir aleatoriamente la muestra total de los 16,387 jefes(as) de hogar en dos grupos. Uno de estos grupos, la muestra de análisis con 12,297 jefes(as) de hogar, que equivale aproximadamente al 75%, que se utiliza para construir el modelo. En la **Tabla 2.1** se puede observar como quedó distribuida la variable dependiente Pobreza en la muestra para el análisis.

Tabla 2.1*: Totales y Porcentajes de los niveles de la variable dependiente en la muestra de análisis.*

Pobreza

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Pobre Extremo	1347	11,0	11,0	11,0
	Pobre Relativo	3230	26,3	26,3	37,2
	No Pobre	7720	62,8	62,8	100,0
	Total	12297	100,0	100,0	

El segundo grupo, la ampliación de la muestra con 4,090 jefes(as) de hogar equivalente al aproximado del 25% que permite validar el modelo. En la **Tabla 2.2** se puede observar como quedó distribuida la variable dependiente Pobreza en la muestra para la validación.

Tabla 2.2: *Totales y Porcentajes de los niveles de la variable dependiente en la muestra de validación.*

Tipo de Pobreza de la muestra para la validación

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Pobre Extremo	460	11,2	11,2	11,2
	Pobre Relativo	1081	26,4	26,4	37,7
	No Pobre	2549	62,3	62,3	100,0
	Total	4090	100,0	100,0	

2.5. Modelo de regresión logística multinomial. Factores asociados a la Pobreza de los Hogares.

Como resultado final y para alcanzar el objetivo principal de esta aplicación, se realiza un ajuste de un modelo de regresión logística multinomial, aplicando lo explicado en el Capítulo I. Se empieza estimando el modelo con las variables explicativas anteriormente mencionadas, y teniendo en cuenta que para cada variable independiente categórica, se generan las variables indicadoras correspondientes.

2.5.1. Estimación del modelo

Al efectuar el análisis en *SPSS* (siguiendo la ruta *Analizar>Regresión>Logística-Multinomial*) con el método de "paso a paso hacia adelante" y con el criterio de contrastar la entrada basándose en la significación del Estadístico de puntuación y contrasta la eliminación basándose en la probabilidad del Estadístico de la razón de verosimilitud, que se basa en estimaciones de la máxima verosimilitud parcial.

En la **Tabla 2.3** se puede ver un resumen de los pasos de la entrada de variables por este método.

Tabla 2.3: *Resumen de pasos: Entrada hacia adelante.*

Resumen de los pasos

Modelo	Acción	Efecto(s)	Criterio de ajuste del modelo	Contrastes de selección de efectos		
			-2 log verosimilitud	Chi-cuadrado[a]	gl	Sig.
0	Introducido	Intersección	21781,868	.		
1	Introducido	INGREFA	15162,920	6618,948	2	,000
2	Introducido	MIEMH	8214,903	6948,017	2	,000
3	Introducido	AREA	3947,766	4267,136	2	,000
4	Introducido	EDAD	3883,506	64,260	10	,000
5	Introducido	ESTAFAMILIAR	3845,666	37,841	10	,000
6	Introducido	SEXO	3830,149	15,516	2	,000
7	Introducido	ELECTRICIDAD	3816,936	13,213	2	,001
8	Introducido	LAVADORA	3808,475	8,461	2	,015
9	Introducido	REGION	3790,764	17,711	8	,024

Método por pasos: Entrada hacia adelante

a. El valor de chi-cuadrado para su inclusión se basa en la prueba de la razón de verosimilitudes.

En la **Tabla 2.4** se presenta los resultados de las estimaciones de los parámetros del modelo, desviación estándar(E.T.), Estadístico de Wald, grados de libertad(gl), Nivel de significancia(Sig.), valor exponente de la base e (Exp(B)) y los límites del intervalo de confianza al 95% para el valor exponente base e (I.C. 95% para Exp(B)).

Tabla 2.4: *Método de Razón de Verosimilitud: Variables en la Ecuación.*

Estimaciones de los parámetros

Pobreza[a]		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
								Límite inferior	Límite superior
Pobre Extremo	Intersección	-10,424	,614	288,692	1	,000			
	MIEMH	8,253	,216	1460,606	1	,000	3839,015	2514,241	5861,824
	[EDAD=1]	2,106	,770	7,473	1	,006	8,213	1,815	37,168
	[EDAD=2]	1,241	,344	12,977	1	,000	3,458	1,761	6,793
	[EDAD=3]	1,016	,331	9,444	1	,002	2,762	1,445	5,279
	[EDAD=4]	,155	,356	,188	1	,664	1,167	,580	2,347
	[EDAD=5]	-,025	,331	,006	1	,939	,975	,510	1,864
	[EDAD=6]	0[b]	.	.	0	.	.	.	.
	[ELECTRICIDAD=1]	-,344	,205	2,821	1	,093	,709	,474	1,059
	[ELECTRICIDAD=2]	0[b]	.	.	0	.	.	.	.
	[SEXO=1]	,883	,289	9,315	1	,002	2,419	1,372	4,265
	[SEXO=2]	0[b]	.	.	0	.	.	.	.
	[REGION=1]	,152	,322	,223	1	,637	1,164	,620	2,186
	[REGION=2]	,525	,323	2,645	1	,104	1,691	,898	3,184
	[REGION=3]	,323	,339	,906	1	,341	1,381	,710	2,684
	[REGION=4]	,683	,331	4,265	1	,039	1,981	1,035	3,789
	[REGION=5]	0[b]	.	.	0	.	.	.	.
	[AREA=1]	15,208	,440	1193,841	1	,000	4023356,311	1698020,501	9533098,099
	[AREA=2]	0[b]	.	.	0	.	.	.	.
	[ESTAFAMILIAR=1]	2,167	,443	23,936	1	,000	8,734	3,666	20,810
	[ESTAFAMILIAR=2]	2,109	,446	22,341	1	,000	8,239	3,436	19,752
	[ESTAFAMILIAR=3]	,943	,470	4,023	1	,045	2,567	1,022	6,451
	[ESTAFAMILIAR=4]	,001	1,118	,000	1	,999	1,001	,112	8,960
	[ESTAFAMILIAR=5]	,561	,411	1,869	1	,172	1,753	,784	3,919
	[ESTAFAMILIAR=6]	0[b]	.	.	0	.	.	.	.
	[LAVADORA=1]	-6,499	8,409	,597	1	,440	,002	1,047E-10	21615,312
	[LAVADORA=2]	0[b]	.	.	0	.	.	.	.
	INGREFA	-,172	,005	1098,248	1	,000	,842	,834	,851
Pobre Relativo	Intersección	-4,788	,316	230,151	1	,000			
	MIEMH	4,279	,126	1150,680	1	,000	72,169	56,361	92,412
	[EDAD=1]	,778	,384	4,099	1	,043	2,177	1,025	4,625
	[EDAD=2]	,371	,189	3,858	1	,049	1,450	1,001	2,100
	[EDAD=3]	,221	,182	1,475	1	,225	1,247	,873	1,781
	[EDAD=4]	,092	,188	,242	1	,623	1,097	,759	1,585
	[EDAD=5]	,103	,178	,334	1	,563	1,109	,782	1,572
	[EDAD=6]	0[b]	.	.	0	.	.	.	.
	[ELECTRICIDAD=1]	-,450	,122	13,610	1	,000	,638	,502	,810
	[ELECTRICIDAD=2]	0[b]	.	.	0	.	.	.	.
	[SEXO=1]	,558	,152	13,457	1	,000	1,747	1,297	2,353
	[SEXO=2]	0[b]	.	.	0	.	.	.	.
	[REGION=1]	-,343	,168	4,176	1	,041	,710	,511	,986
	[REGION=2]	,055	,165	,112	1	,738	1,057	,765	1,460
	[REGION=3]	-,207	,180	1,315	1	,251	,813	,571	1,158
	[REGION=4]	,133	,176	,576	1	,448	1,142	,810	1,612
	[REGION=5]	0[b]	.	.	0	.	.	.	.
	[AREA=1]	7,311	,236	962,714	1	,000	1496,230	942,842	2374,422
	[AREA=2]	0[b]	.	.	0	.	.	.	.
	[ESTAFAMILIAR=1]	1,184	,229	26,834	1	,000	3,267	2,088	5,113
	[ESTAFAMILIAR=2]	1,128	,230	24,166	1	,000	3,090	1,971	4,846
	[ESTAFAMILIAR=3]	,779	,247	9,963	1	,002	2,179	1,344	3,535
	[ESTAFAMILIAR=4]	1,006	,611	2,707	1	,100	2,733	,825	9,056
	[ESTAFAMILIAR=5]	,383	,211	3,299	1	,069	1,466	,970	2,216
	[ESTAFAMILIAR=6]	0[b]	.	.	0	.	.	.	.
	[LAVADORA=1]	-,674	,302	4,992	1	,025	,510	,282	,921
	[LAVADORA=2]	0[b]	.	.	0	.	.	.	.
	INGREFA	-,059	,002	1178,895	1	,000	,943	,940	,946

a. La categoría de referencia es: No Pobre.
b. Este parámetro se ha establecido a cero porque es redundante.

Se puede notar que el SPSS permite indicarle en la opción *Analizar>Regresión>Logística-Multinomial*, al momento del análisis, cuales son las variables cualitativas a la que habrá que generarles las variables indicadoras; señalando cual de las alternativas es la de referencia (a la que le asigna código de 0) en este caso se ha señalado la última categoría de referencia para cada una de las variables consideradas como cualitativas, esto permite establecer el contraste de significación de los coeficientes en cada una de las variables indicadoras, bajo el supuesto que:

$$H_0\colon \beta_{ij} = 0$$

Que significa probar la hipótesis que el coeficiente correspondiente en el modelo logístico lineal para la categoría j y la variable i es cero. Esta prueba como ya se mencionó en el Capítulo I, se puede realizar con el Estadístico de Wald. Rechazando si el valor de este Estadístico es mayor que el valor en tablas para una chi-cuadrado con un grado de libertad a un nivel de significancia de α.

Revisando este criterio se puede observar que en la **Tabla 2.4** existen variables no significativas en el modelo para "*pobre extremo"* y significativas para el modelo *"pobre relativo"* y viceversa, pero estas son variables de diseño (indicadoras) en las cuales como se mencionó en el Capítulo I, no es muy aconsejable el uso del Estadístico de Wald. En ese caso se recomienda el uso del test de razón de verosimilitudes, que sirve para problemas como el que se está dando, en el cual hay variables cualitativas transformadas en variables de diseño (indicadoras) y resultan no ser todas significativas, y como el criterio es que: siempre que se incluya o se excluya una de estas variables, todas las demás deben ser incluidas o excluidas en el bloque. Dado esto, la **Tabla 2.5** muestra un resumen de las pruebas de inclusión de las variables por el método de razón de verosimilitudes, decidiendo incluir o excluir la variable dependiendo del resultado de la prueba, y observando si se obtiene significación en ese contraste, las variables permanecerían en el modelo, dado que en esa tabla todas resultan ser significativas, entonces se deciden conservar en el modelo a pesar de la prueba del Estadístico de Wald.

Tabla 2.5: *Contraste de significación de las variables.*

Contrastes de la razón de verosimilitud

Efecto	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
	-2 log verosimilitud del modelo reducido	Chi-cuadrado	gl	Sig.
Intersección	3790,764[a]	,000	0	.
MIEMH	13122,312	9331,548	2	,000
EDAD	3825,899	35,135	10	,000
ELECTRICIDAD	3804,849	14,085	2	,001
SEXO	3806,221	15,456	2	,000
REGION	3808,475	17,711	8	,024
AREA	7894,034	4103,270	2	,000
ESTAFAMILIAR	3841,675	50,911	10	,000
LAVADORA	3799,373	8,608	2	,014
INGREFA	19422,239	15631,474	2	,000

El estadístico de chi-cuadrado es la diferencia en las -2 log verosimilitudes entre el modelo final y el modelo reducido. El modelo reducido se forma omitiendo un efecto del modelo final. La hipótesis nula es que todos los parámetros de ese efecto son 0.
a. Este modelo reducido es equivalente al modelo final ya que la omisión del efecto no incrementa los grados de libertad.

2.5.2. Interpretación del Modelo.

Cuando interpretamos los Odds_ratios de cada variable, se asume que el resto de variables independientes se mantienen fijas. Interpretamos cada una de las variables independientes entre los distintos tipo de pobreza del hogar tomando como referencia "No Pobre".

El odds_ratio cambia cuando la i-ésima variable explicativa regresora se incrementa en una unidad, si:

$\beta_i > 0$ significa que el odds_ratio se incrementa.

$\beta_i < 0$ significa que el odds_ratio decrece.

$\beta_i = 0$ significa que el factor es igual a uno, lo cual hace que el odds_ratio no varía.

En la columna exp(β) de la **Tabla 2.4**, se observa que:

Para la categoría "Pobre Extremo" β_1, β_2, β_3, β_4, β_5, β_8, β_9, β_{10}, β_{11}, β_{12}, β_{13}, β_{14}, β_{15}, β_{16}, β_{17}, y $\beta_{18} > 0$, por lo tanto, por cada unidad que aumente el número de miembros del hogar , es aproximadamente 3,839 veces más probable que el hogar sea Pobre Extremo a que el hogar sea No Pobre. De la misma manera se puede decir que, si la Edad de el/la jefe(a) de hogar está entre 15 y 20 años (Primera categoría de la variable EDAD) es 8.213 veces más probable que el hogar sea Pobre Extremo a que el hogar sea No Pobre con respecto a que la edad de el/la jefe(a) de hogar sea mayor a 65 años, en cambio, si la edad de el/la jefe(a) de hogar se encuentra entre los 21 y los 30 años es solo 3.458 veces más probable que el hogar sea Pobre Extremo a que el hogar sea No Pobre con

respecto a que la edad de el/la jefe(a) de hogar sea mayor a 65 años, y se observa que esta ventaja va decreciendo conforme la Edad de el/la jefe(a) de hogar va aumentando.

Se puede observar también que si el SEXO de el/la jefe(a) de hogar es femenino es 2.419 veces más probable que el hogar sea Pobre Extremo a que el hogar sea No Pobre con respecto a que el sexo del jefe de hogar sea masculino. También, la probabilidad de que el hogar sea Pobre Extremo aumenta un 16.4% en la Región Occidental(REGION=1) con respecto a que el hogar sea No Pobre, a que si el hogar perteneciera en la región del área metropolitana de San Salvador, en cambio en la región Central 1(REGION=2) la probabilidad de ser el hogar Pobre Extremo aumenta el 69.1%, en la región Central 2(REGION=3) ésta aumenta el 38.1% y en la región Oriental(REGION=4) aumenta un 98.1% la probabilidad de que el hogar sea Pobre Extremo con respecto a que el hogar sea No Pobre a que si el hogar perteneciera en la región del área metropolitana de San salvador. También el hecho que el Estado Familiar de el/la jefe(a) de hogar sea acompañado(a) es 8.73 veces más probable que el hogar sea Pobre Extremo a que el hogar sea No Pobre con respecto a que el Estado Familiar de el/la jefe(a) de hogar sea Soltero, en cambio, si el Estado Familiar de el/la jefe(a) de hogar es casado(a) es 8.239 veces más probable que el hogar sea Pobre Extremo a que el hogar sea No Pobre con respecto a que el Estado Familiar de el/la jefe(a) de hogar sea Soltero, en cambio, si el Estado Familiar de el/la jefe(a) de hogar es Viudo(a), es 2.567 veces más probable, y si el Estado Familiar de el/la jefe(a) de hogar es Divorciado(a), se mantiene intacta esta probabilidad con respecto a que si es Soltero, y si es Separado(a) aumenta un 75.3% la probabilidad de que el hogar sea Pobre Extremo a que sea No Pobre con respecto a que si es Soltero.

Para la categoría "Pobre Extremo" β_6, β_7, β_{19} y $\beta_{20} < 0$, por lo tanto, las variables EDAD de el/la jefe(a) de hogar en la categoría de 51 a 65 años, poseer energía Eléctrica en el hogar, poseer Lavadora e incrementar el Ingreso Familiar, son factores que disminuyen la probabilidad de que el hogar sea Pobre Extremo en comparación a que el hogar sea No Pobre. Así el hogar que posea lavadora disminuye la probabilidad de ser Pobre Extremo en 99.8% con respecto a ser el hogar No Pobre, si se mantienen constante el resto de las variables, también poseer el servicio de energía Eléctrica en el hogar disminuye la probabilidad de ser Pobre Extremo en 29.1% con respecto a ser No Pobre, y aumentar el Ingreso Familiar en una unidad disminuye la probabilidad de ser Pobre

Extremo en 15.8% con respecto a ser No Pobre (nuevamente asumiendo que mantenemos constante al resto de las variables).

Para la categoría "Pobre Relativo" la interpretación es similar, se observa que β_1, β_2, β_3, β_4, β_5, β_6, β_8, β_{10}, β_{12}, β_{13}, β_{14}, β_{15}, β_{16}, β_{17}, y $\beta_{18} > 0$, por lo tanto, por cada unidad que aumente el número de miembros del hogar, es aproximadamente 72.17 veces más probable que el hogar sea Pobre Relativo a que el hogar sea No Pobre. También, si la Edad de el/la jefe(a) de hogar está entre 15 y 20 años (Primera categoría de la variable EDAD) es 2.18 veces más probable que el hogar sea Pobre Relativo a que el hogar sea No Pobre con respecto a que la edad de el/la jefe(a) de hogar sea mayor a 65 años. Se puede observar también que si el SEXO de el/la jefe(a) de hogar es femenino es 74.7% más probable que el hogar sea Pobre Relativo a que el hogar sea No Pobre con respecto a que el sexo del jefe de hogar sea masculino. También el hecho que el Estado Familiar de el/la jefe(a) de hogar sea acompañado(a) es 3.267 veces más probable que el hogar sea Pobre Relativo a que el hogar sea No Pobre con respecto a que el Estado Familiar de el/la jefe(a) de hogar sea Soltero, en cambio, si el Estado Familiar de el/la jefe(a) de hogar es casado(a) es 3.09 veces más probable que el hogar sea Pobre Relativo a que el hogar sea No Pobre con respecto a que el Estado Familiar de el/la jefe(a) de hogar sea Soltero, en cambio, si el Estado Familiar de el/la jefe(a) de hogar es Viudo(a), es 2.17 veces más probable, y si el Estado Familiar de el/la jefe(a) de hogar es Divorciado(a), es 2.74 veces más probable que el hogar sea Pobre Relativo a que el hogar sea No Pobre con respecto a que el Estado Familiar de el/la jefe(a) de hogar sea Soltero, y si es Separado(a) aumenta un 46.6% la probabilidad de que el hogar sea Pobre Relativo a que sea No Pobre con respecto a que si es Soltero.

Para la categoría "Pobre Relativo" β_7 , β_9 , β_{11} , β_{19} y $\beta_{20} < 0$, por lo tanto, las variables poseer Energía Eléctrica, encontrarse en la Región Occidental o en la Región Central 2, poseer lavadora y el aumento del Ingreso Familiar, son factores que disminuyen la probabilidad de que el hogar se encuentre en Pobreza Relativa, en comparación a que el hogar sea No Pobre.

Luego, con la información de la **Tabla 2.4** y definiendo:

- $\hat{\pi}_{i1}$, siendo la probabilidad estimada de que un hogar "i" dado la información de las variables significativas sea clasificado como "Pobre Extremo" en la variable cualitativa "Pobreza del Hogar".
- $\hat{\pi}_{i2}$, siendo la probabilidad estimada de que un hogar "i" dada la información de las variables significativas sea clasificado como "Pobre Relativo" en la variable cualitativa "Pobreza del Hogar".
- $\hat{\pi}_{i3}$, siendo la probabilidad estimada de que un hogar "i" dada la información de las variables significativas sea clasificado como "No Pobre" en la variable cualitativa "Pobreza del Hogar".

Se tienen los modelos logit para las categorías de la variable dependiente "Pobreza del Hogar", ya que tiene 3 categorías, se construyen modelos logit sólo para dos de sus categorías tomando una como referencia, en este caso la categoría de referencia es "No Pobre", entonces los modelos para las dos categorías restantes "Pobre Extremo" y "Pobre Relativo" son:

$$g_1(x) = ln\left(\frac{\hat{\pi}_{i1}}{\hat{\pi}_{i3}}\right) = -10.424 + 8.253 * MIEMH + 2.106 * EDAD(1) + 1.241 * EDAD(2) + 1.016 * EDAD(3) + 0.155 * EDAD(4) - 0.025 * EDAD(5) - 0.344 * ELECTRICIDAD + 0.883 * SEXO + 0.152 * REGION(1) + 0.525 * REGION(2) + 0.323 * REGION(3) + 0.683 * REGION(4) + 15.208 * AREA + 2.167 * ESTAFAMILIAR(1) + 2.109 * ESTAFAMILIAR(2) + 0.943 * ESTAFAMILIAR(3) + 0.001 * ESTAFAMILIAR(4) + 0.561 * ESTAFAMILIAR(5) - 6.499 * LAVADORA - 0.172 * INGREFA$$

$$g_2(x) = ln\left(\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}}\right) = -4.788 + 4.279 * MIEMH + 0.778 * EDAD(1) + 0.371 * EDAD(2) + 0.221 * EDAD(3) + 0.092 * EDAD(4) + 0.103 * EDAD(5) - 0.450 * ELECTRICIDAD + 0.558 * SEXO - 0.343 * REGION(1) + 0.055 * REGION(2) - 0.207 * REGION(3) + 0.133 * REGION(4) + 7.311 * AREA + 1.184 * ESTAFAMILIAR(1) + 1.128 * ESTAFAMILIAR(2) + 0.779 * ESTAFAMILIAR(3) + 1.006 * ESTAFAMILIAR(4) + 0.383 * ESTAFAMILIAR(5) - 0.674 * LAVADORA - 0.059 * INGREFA$$

Por lo visto en el Capítulo I, se obtienen las probabilidades estimadas $\hat{\pi}_{i1}$, $\hat{\pi}_{i2}$ y $\hat{\pi}_{i3}$, resultando las siguientes expresiones:

$$\hat{\pi}_{i1} = \frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$\hat{\pi}_{i2} = \frac{e^{g_2(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

$$\hat{\pi}_{i3} = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}}$$

Un ejemplo de cómo se utilizaría este modelo, tendríamos el siguiente caso (hogar) tomado de la muestra de análisis para clasificarlo de acuerdo a la información brindada por el modelo.

POBREZA	REGION(1)	REGION(2)	REGION(3)	REGION(4)	AREA	SEXO
1	0	1	0	0	1	1

ESTAFAMILIAR(1)	ESTAFAMILIAR(2)	ESTAFAMILIAR(3)	ESTAFAMILIAR(4)	ESTAFAMILIAR(5)
0	1	0	0	0

INGREFA	MIEMH	ELECTRICIDAD	LAVADORA	EDAD(1)	EDAD(2)	EDAD(3)	EDAD(4)	EDAD(5)
65	2	0	0	0	0	0	0	0

Siguiendo este perfil, sustituiríamos los valores para las variables en los predictores lineales estimados, quedando de la siguiente forma.

$$g_1(x) = ln\left(\frac{\hat{\pi}_{i1}}{\hat{\pi}_{i3}}\right) = -10.424 + 8.253 * (2) + 2.106 * (0) + 1.241 * (0) + 1.016 * (0) + 0.155 * (0) - 0.025 * (0) - 0.344 * (0) + 0.883 * (1) + 0.152 * (0) + 0.525 * (1) + 0.323 * (0) + 0.683 * (0) + 15.208 * (1) + 2.167 * (0) + 2.109 * (1) + 0.943 * (0) + 0.001 * (0) + 0.561 * (0) - 6.499 * (0) - 0.172 * (65)$$

$$g_1(x) = ln\left(\frac{\hat{\pi}_{i1}}{\hat{\pi}_{i3}}\right) = 13.627$$

$$g_2(x) = ln\left(\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}}\right) = -4.788 + 4.279 * (2) + 0.778 * (0) + 0.371 * (0) + 0.221 * (0) + 0.092 * (0) + 0.103 * (0) - 0.450 * (0) + 0.558 * (1) - 0.343 * (0) + 0.055 * (1) - 0.207 * (0) + 0.133 * (0) + 7.311 * (1) + 1.184 * (0) + 1.128 * (1) + 0.779 * (0) + 1.006 * (0) + 0.383 * (0) - 0.674 * (0) - 0.059 * (65)$$

$$g_1(x) = ln\left(\frac{\hat{\pi}_{i1}}{\hat{\pi}_{i3}}\right) = 8.987$$

Ahora que se tienen los predictores calculados, se pasaría a estimar las probabilidades.

$$\hat{\pi}_{i1} = \frac{e^{g_1(x)}}{1+e^{g_1(x)}+e^{g_2(x)}} = \frac{e^{13.627}}{1+e^{13.627}+e^{8.987}} = \frac{828191.7594}{1+828191.7594+7998.45559} = 0.9904$$

$$\hat{\pi}_{i2} = \frac{e^{g_2(x)}}{1+e^{g_1(x)}+e^{g_2(x)}} = \frac{e^{8.987}}{1+e^{13.627}+e^{8.987}} = \frac{7998.45559}{1+828191.7594+7998.45559} = 0.0096$$

$$\hat{\pi}_{i3} = \frac{1}{1+e^{g_1(x)}+e^{g_2(x)}} = \frac{1}{1+e^{13.627}+e^{8.987}} = \frac{1}{1+828191.7594+7998.45559} \cong 0$$

Por lo que se observa que la mayor probabilidad es $\hat{\pi}_{i1} = 0.9904$, que representa la probabilidad que este hogar sea Pobre Extremo, por lo que a este hogar se le asigna a la

variable POBREZA (predicha) el valor de 1, que coincide con el valor de la variable POBREZA que se observa para este perfil.

Antes de crear conclusiones acerca de este modelo, primero se tiene que evaluar su bondad de ajuste, su calidad de ajuste y su validación, que es lo que se pasará a realizar.

2.5.3. Bondad del ajuste.

La bondad y ajuste de este modelo logístico multinomial se puede contrastar con la información de la siguiente tabla generada con SPSS>Analizar>Regresión>Logística-Multinomial:

Tabla 2.6: *Estadísticos sobre la bondad de ajuste del modelo.*

Bondad de ajuste

	Chi-cuadrado	GI	Sig.
Pearson	14252,474	24534	1,000
Desviación	3790,764	24534	1,000

De la tabla anterior se pueden probar las siguientes hipótesis de bondad de ajuste:

$$H_0: \pi_{ij} = \frac{e^{(\sum_{p=0}^{q} \beta_{pj} x_{ip})}}{1+\sum_{j=1}^{k-1} e^{\left(\sum_{p=0}^{q} \beta_{pj} X_{ip}\right)}} \quad \forall\, i = 1, \ldots . n; \quad \forall\, j = 1, \ldots , k$$

$$H_1: \pi_{ij} \neq \frac{e^{(\sum_{p=0}^{q} \beta_{pj} x_{ip})}}{1+\sum_{j=1}^{k-1} e^{\left(\sum_{p=0}^{q} \beta_{pj} X_{ip}\right)}} \text{ Para algún i, j.}$$

Los estadísticos de prueba para contrastar esta hipótesis nula tal y como se vio en el Capítulo I son: el Estadístico de Pearson y el Estadístico de Desviación, los cuales se definen de la siguiente forma:

2.5.3.1. Estadístico de Pearson.

$$\chi^2(M) = \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{(y_{ij} - m_i \hat{\pi}_{ij})^2}{m_i \hat{\pi}_{ij}}$$

Sean y_{ij} el número de observaciones que caen en la categoría de respuesta Y_j $\forall$ j=1,2,…, k. y sean las m_i observaciones correspondientes a la i-ésima

combinación de valores de las variables explicativas. Siendo $\hat{\pi}_{ij}$ el estimador por máxima verosimilitud de π_{ij}.

2.5.3.2. Estadístico de Wilks. Desvianza.

El estadístico de Wilks de razón de verosimilitudes para el contraste de bondad de ajuste el modelo de regresión logística multinomial M se obtiene como menos dos veces el logaritmo del cociente entre el supremo de la verosimilitud bajo la hipótesis nula y el supremo de la verosimilitud en la población.

$$S_n^{(1)} = 2\left[\ell^{(1)}(\hat{\pi}; y_1, \dots, y_n) - \ell_0^{(1)}(\hat{\beta}; y_1, \dots, y_n)\right]$$

De la **Tabla 2.6** se observa que el valor del **Estadístico de Pearson** es de: 14,252.474, o sea:

$\chi^2(M) = 14{,}252.474$, que tiene una chi-cuadrado con 24,534 grados de libertad $\chi^2_{24,534}$(de los 12,297 casos hay 12,288 clases covariantes y 21 parámetros estimados) y el valor en tablas para su valor chi-cuadrado respectivo es: $\chi^2_{24,534,0.95} = 24{,}899.49$. Por tanto no puede rechazarse H_0 y el modelo estimado es aceptable para estimar las probabilidades de pertenencia de las categorías de la variable dependiente "Tipo de Pobreza del Hogar".

Lo mismo pasa con el **Estadístico de Desviación** $\mathbf{S_n^{(1)}}$ que es:

$3{,}790.764 < \chi^2_{24,534,0.95} = 24{,}899.49$, llevando al no rechazo de H_0.

Por lo que no se puede rechazar la hipótesis de que el ajuste del modelo sea válido, es decir el modelo es adecuado para el ajuste de los datos.

2.5.3.3. Tasa de Clasificaciones correctas

También para cuantificar la bondad del ajuste global del modelo utilizamos la tasa de clasificaciones correctas, que nos permite clasificar cada observación en la categoría más probable, construyendo así una matriz de clasificación *Observados-Predichos*, como se observa en la siguiente tabla.

Tabla 2.7: *Matriz de clasificación Observados-Predichos para el modelo estimado.*

Clasificación

Observado	Pronosticado			
	Pobre Extremo	Pobre Relativo	No Pobre	Porcentaje correcto
Pobre Extremo	1210	129	8	89,8%
Pobre Relativo	118	2827	285	87,5%
No Pobre	0	281	7439	96,4%
Porcentaje global	10,8%	26,3%	62,9%	93,3%

Se observa que la tasa de clasificaciones correctas es de 93.3%, A través de la misma se puede concluir acerca de la eficacia predictiva del modelo, que al ser de un 93.3%, se puede decir que el modelo posee una muy buena predicción, es decir que un 93.3% de los casos analizados logran ser correctamente clasificados, al coincidir el tipo de Pobreza del Hogar observado con el pronosticado por el modelo.

2.5.4. Calidad del Ajuste del Modelo.

Para medir la calidad del ajuste del modelo se utiliza como se vió en el Capítulo I, los coeficientes **Pseudo-R^2 de Mc-Fadden, de Cox-Snell** y de **Nagelkerke.** El cálculo de éstos depende del valor de las funciones de log-verosimilitud del modelo final y del modelo inicial con solo la constante, por lo que se calculan de la siguiente manera:

Si se tiene el modelo inicial con solo la constante y el modelo final, y sean respectivamente, Λ_0 y Λ_f sus funciones de log-verosimilitud, obtenemos los siguientes coeficientes de la **Tabla 2.8**:

Tabla 2.8: *Coeficientes de Medición de Ajuste*

Pseudo R-cuadrado

Cox y Snell	,768
Nagelkerke	,926
McFadden	,826

2.5.4.1. Coeficiente pseudo- R^2 de Mc-Fadden.

$$R^2_{MF} = 1 - \frac{\Lambda_f}{\Lambda_0} = 0.826$$

2.5.4.2. Coeficiente pseudo-R^2 de Cox-Snell.

$$R_{CS}^2 = 1 - \left(\frac{V_0}{V_f}\right)^{\frac{2}{N}} = 1 - \exp\left(\frac{\Lambda_f - \Lambda_0}{N}\right) = 0.768$$

2.5.4.3. Coeficiente pseudo-R^2 de Nagelkerke.

$$R_N^2 = \frac{R_{CS}^2}{1 - V_0^{\frac{2}{N}}} = \frac{1 - \exp\left(\frac{\Lambda_f - \Lambda_0}{N}\right)}{1 - exp\left(\frac{-\Lambda_0}{N}\right)} = 0.926$$

Se puede observar que los coeficientes pseudo-R^2, alcanzan valores altos cercanos a 1, por lo que se puede concluir que el modelo presenta muy buena calidad de ajuste.

2.5.5. Validación del modelo.

Una vez encontrado el mejor modelo, en esta fase tenemos que validarlo. En un modelo predictivo se trata de ver si el modelo predice bien la variable dependiente en un nuevo hogar encuestado.

Como se había mencionado anteriormente en este capítulo, la muestra es N=16,337 casos válidos observados para este análisis fue dividida en dos submuestras, una para el análisis y la otra con fines de validación, la primera con un tamaño de 12,297 casos y la segunda con 4,090 casos, y dadas las probabilidades de clasificación del Hogar por condición de pobreza estimadas $\hat{\pi}_{i1}$, $\hat{\pi}_{i2}$ y $\hat{\pi}_{i3}$. Se define los siguientes criterios de clasificación:

1- Si se cumple la condición: $\hat{\pi}_{i1} >= \hat{\pi}_{i2}$ y $\hat{\pi}_{i1} > \hat{\pi}_{i3}$, entonces el Hogar se clasifica como: “Hogar en Pobreza Extrema”.

2- Si se cumple la condición: $\hat{\pi}_{i2} > \hat{\pi}_{i1}$ y $\hat{\pi}_{i2} >= \hat{\pi}_{i3}$, entonces el Hogar se clasifica como: “Hogar en Pobreza Relativa”.

3- Si se cumple la condición: $\hat{\pi}_{i3} > \hat{\pi}_{i1}$ y $\hat{\pi}_{i3} > \hat{\pi}_{i2}$, entonces el Hogar se clasifica como: “Hogar No Pobre”.

Entonces, se pueden usar estos criterios de clasificación para un hogar nuevo encuestado, tomando los 4,090 casos de la muestra de validación, para observar el nivel de predicción que tiene el modelo. Veamos cómo podríamos ir construyendo una tabla

de clasificación con el modelo, teniendo en cuenta que para los 4,090 casos ya se tienen los niveles de los hogares, y el objetivo sería observar que tanto acierto tiene el modelo en su predicción al clasificarlos comparando con lo que ya se tiene observado.

Como un segundo ejemplo a manera de ilustrar lo que se quiere realizar, se toma un caso de los 4,090 casos de la muestra de validación, y se observa cómo se clasifica de acuerdo a los criterios anteriores.

Se tiene el siguiente caso de ejemplo para clasificarlo en su nivel de pobreza usando el modelo, con su información para las variables respectivas:

POBREZA	REGION(1)	REGION(2)	REGION(3)	REGION(4)	AREA	SEXO
2	1	0	0	0	0	0

ESTAFAMILIAR(1)	ESTAFAMILIAR(2)	ESTAFAMILIAR(3)	ESTAFAMILIAR(4)	ESTAFAMILIAR(5)
1	0	0	0	0

INGREFA	MIEMH	ELECTRICIDAD	LAVADORA	EDAD(1)	EDAD(2)	EDAD(3)	EDAD(4)	EDAD(5)
129,46	3	0	0	0	1	0	0	0

De acuerdo a esta información, podemos evaluar estos datos para este caso (Hogar i) en las funciones $g_1(x)$ y $g_2(x)$, respectivamente, y elevando esos resultados en exponenciales, dando lo siguientes:

$e^{g_1(x)} = 0{,}012624448$.

$e^{g_2(x)} = 5{,}067562841$.

Ahora, con estos resultados podemos encontrar las probabilidades estimadas para observar en qué tipo de pobreza es clasificado este caso (Hogar i).

$$\hat{\pi}_{i1} = \frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}} = \frac{0{,}012624448}{1 + 0{,}012624448 + 5{,}067562841} = 0{,}002076326$$

$$\hat{\pi}_{i2} = \frac{e^{g_2(x)}}{1 + e^{g_1(x)} + e^{g_2(x)}} = \frac{5{,}067562841}{1 + 0{,}012624448 + 5{,}067562841} = 0{,}833455057$$

$$\hat{\pi}_{i3} = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)}} = \frac{1}{1 + 0{,}012624448 + 5{,}067562841} = 0{,}164468618$$

Se observa que según los criterios mencionados anteriormente, este hogar "i" cumple el criterio de clasificación 2. Por lo tanto este hogar es clasificado como "Pobre Relativo"

(o consecuentemente para este hogar la variable POBREZA toma el valor 2, el cual coincide con el valor actual de esa variable).

Si hacemos este mismo procedimiento para el total de casos de la muestra de validación (los 4,090 casos que se dejaron fuera de la estimación), se puede observar el porcentaje de aciertos que tiene el modelo, y esto se ve reflejado en la **Tabla 2.9** que se muestra a continuación:

Tabla 2.9: *Predicción del modelo para los 4,090 casos que quedaron fuera.*

Tipo de Pobreza del Hogar Pronosticado Vs Tipo de Pobreza del Hogar Observado

		Tipo de Pobreza del Hogar Observado			
		Pobre Extremo	Pobre Relativo	No Pobre	Total
Tipo de Pobreza del Hogar Pronosticado	Pobre Extremo	414	31	0	445
	Pobre Relativo	43	948	77	1068
	No Pobre	3	102	2472	2577
Total		460	1081	2549	4090

De esta tabla se puede encontrar el porcentaje de aciertos que tuvo el modelo con la muestra de validación, y se obtiene de la siguiente manera:
(414+948+2472) / 4090 = 0,937.

Por lo tanto se puede decir que el modelo acertó en clasificar según su tipo de pobreza a los hogares el 93,7% ó 3,834 casos de los 4,090 casos que se dejaron para la validación.

2.6 Resumen Ilustrativo de la Aplicación del Modelo de Regresión Multinomial.

A continuación se muestran los cuadros del número de hogares clasificados como "Pobre Extremo", "Pobre Relativo" y "No Pobre", en comparación con la clasificación realizada con el modelo logístico multinomial estimado, en cada una de las regiones y áreas del país identificadas por la Dirección General de Estadísticas y Censos (DIGESTYC), se utilizan todos los datos (datos usados para la validación y para la estimación del modelo de regresión logístico multinomial).

Tabla 2.10: *Aplicación del modelo de regresión logística multinomial para la zona Occidental.*

Tabla de contingencia Categoría de respuesta Observada * Categoría de respuesta pronosticada

Región				Categoría de respuesta pronosticada			
				Pobre Extremo	Pobre Relativo	No Pobre	Total
Occidental	Categoría de respuesta Observada	Pobre Extremo	Recuento	456	39	3	498
			% del total	12,6%	1,1%	,1%	13,7%
		Pobre Relativo	Recuento	43	865	102	1010
			% del total	1,2%	23,9%	2,8%	27,9%
		No Pobre	Recuento	0	80	2038	2118
			% del total	,0%	2,2%	56,2%	58,4%
	Total		Recuento	499	984	2143	3626
			% del total	13,8%	27,1%	59,1%	100,0%

Tabla 2.11: *Aplicación del modelo de regresión logística multinomial para la zona Central I.*

Tabla de contingencia Categoría de respuesta Observada * Categoría de respuesta pronosticada

Región				Categoría de respuesta pronosticada			
				Pobre Extremo	Pobre Relativo	No Pobre	Total
Central 1	Categoría de respuesta Observada	Pobre Extremo	Recuento	325	36	6	367
			% del total	8,9%	1,0%	,2%	10,0%
		Pobre Relativo	Recuento	31	875	84	990
			% del total	,8%	23,9%	2,3%	27,1%
		No Pobre	Recuento	0	80	2217	2297
			% del total	,0%	2,2%	60,7%	62,9%
	Total		Recuento	356	991	2307	3654
			% del total	9,7%	27,1%	63,1%	100,0%

Tabla 2.12: *Aplicación del modelo de regresión logística multinomial para la zona Central II.*

Tabla de contingencia Categoría de respuesta Observada * Categoría de respuesta pronosticada

Región				Categoría de respuesta pronosticada			
				Pobre Extremo	Pobre Relativo	No Pobre	Total
Central 2	Categoría de respuesta Observada	Pobre Extremo	Recuento	307	37	2	346
			% del total	11,5%	1,4%	,1%	13,0%
		Pobre Relativo	Recuento	32	673	64	769
			% del total	1,2%	25,3%	2,4%	28,9%
		No Pobre	Recuento	0	59	1488	1547
			% del total	,0%	2,2%	55,9%	58,1%
	Total		Recuento	339	769	1554	2662
			% del total	12,7%	28,9%	58,4%	100,0%

Tabla 2.13: *Aplicación del modelo de regresión logística multinomial para la zona Oriental.*

Tabla de contingencia Categoría de respuesta Observada * Categoría de respuesta pronosticada

Región				Categoría de respuesta pronosticada			
				Pobre Extremo	Pobre Relativo	No Pobre	Total
Oriental	Categoría de respuesta Observada	Pobre Extremo	Recuento	405	38	0	443
			% del total	12,1%	1,1%	,0%	13,2%
		Pobre Relativo	Recuento	30	786	82	898
			% del total	,9%	23,4%	2,4%	26,7%
		No Pobre	Recuento	0	83	1934	2017
			% del total	,0%	2,5%	57,6%	60,1%
	Total		Recuento	435	907	2016	3358
			% del total	13,0%	27,0%	60,0%	100,0%

Tabla 2.14: *Aplicación del modelo de regresión logística multinomial para la zona del Área Metropolitana de San Salvador (AMSS).*

Tabla de contingencia Categoría de respuesta Observada * Categoría de respuesta pronosticada

Región				Categoría de respuesta pronosticada			
				Pobre Extremo	Pobre Relativo	No Pobre	Total
AMSS	Categoría de respuesta Observada	Pobre Extremo	Recuento	131	22	0	153
			% del total	4,2%	,7%	,0%	5,0%
		Pobre Relativo	Recuento	13	576	55	644
			% del total	,4%	18,7%	1,8%	20,9%
		No Pobre	Recuento	0	56	2234	2290
			% del total	,0%	1,8%	72,4%	74,2%
	Total		Recuento	144	654	2289	3087
			% del total	4,7%	21,2%	74,1%	100,0%

Tabla 2.15: *Aplicación del modelo de regresión logística multinomial por área Urbana.*

Tabla de contingencia Categoría de respuesta Observada * Categoría de respuesta pronosticada

Área				Categoría de respuesta pronosticada			
				Pobre Extremo	Pobre Relativo	No Pobre	Total
Urbana	Categoría de respuesta Observada	Pobre Extremo	Recuento	837	95	0	932
			% del total	8,6%	1,0%	,0%	9,5%
		Pobre Relativo	Recuento	78	2208	192	2478
			% del total	,8%	22,6%	2,0%	25,4%
		No Pobre	Recuento	0	187	6170	6357
			% del total	,0%	1,9%	63,2%	65,1%
	Total		Recuento	915	2490	6362	9767
			% del total	9,4%	25,5%	65,1%	100,0%

Tabla 2.16: *Aplicación del modelo de regresión logística multinomial por área Rural.*

Tabla de contingencia Categoría de respuesta Observada * Categoría de respuesta pronosticada

Área				Categoría de respuesta pronosticada			
				Pobre Extremo	Pobre Relativo	No Pobre	Total
Rural	Categoría de respuesta Observada	Pobre Extremo	Recuento	787	77	11	875
			% del total	11,9%	1,2%	,2%	13,2%
		Pobre Relativo	Recuento	71	1567	195	1833
			% del total	1,1%	23,7%	2,9%	27,7%
		No Pobre	Recuento	0	171	3741	3912
			% del total	,0%	2,6%	56,5%	59,1%
	Total		Recuento	858	1815	3947	6620
			% del total	13,0%	27,4%	59,6%	100,0%

Conclusiones.

El modelo de regresión logística multinomial permite relacionar la variable dependiente Tipo de Pobreza del Hogar con las variables referidas al hogar y su jefatura como lo son: la Región, el Área, el Ingreso Familiar, el Número de Miembros del hogar, si posee Electricidad el hogar, si posee lavadora, el Sexo del Jefe(a) de Hogar, el estado Familiar del Jefe(a) de Hogar y la Edad del Jefe(a) de Hogar en la Encuesta de Hogares de Propósitos Múltiples del año 2010.

Las variables que en mayor medida permiten clasificar a un determinado hogar como "Pobre Relativo" o como "Pobre Extremo" son el Número de Miembros del Hogar y el Área Geográfica, se observa que a mayor número de miembros del hogar, mayor es la probabilidad de que el hogar sea Pobre Extremo o Pobre Relativo, también, si el hogar es del Área urbana aumenta la probabilidad de que el hogar sea Pobre Extremo o Pobre Relativo.

También se puede observar que a medida el jefe(a) de hogar tiende a ser más joven, mayor es la probabilidad de que el hogar sea Pobre Extremo o Pobre Relativo, se observa que cuando el Sexo del jefe(a) de hogar es femenino es 2.419 veces más probable que el hogar sea Pobre Extremo a que el hogar sea No Pobre y es 74.7% más probable que el hogar sea Pobre Relativo a que el hogar sea No Pobre con respecto a que el sexo del jefe de hogar sea masculino.

Las variables que reducen tanto la probabilidad de que el hogar sea clasificado como "Pobre Relativo" y la probabilidad de que el hogar sea clasificado como "Pobre Extremo" son: el Ingreso Familiar (un mayor ingreso familiar aumenta la probabilidad de que el hogar sea "No Pobre"), también si el hogar posee Energía Eléctrica y si el hogar posee Lavadora.

La Regresión Logística Multinomial es una de las herramientas estadísticas con mejor capacidad para el análisis, cuando la variable dependiente es categórica, sirve para determinar los factores de riesgo y factores de prevención frente a la situación de pobreza que enfrentan los hogares. En esta investigación se obtiene un modelo que cumple con los supuestos requeridos y que cumple con todos los test estadísticos requeridos.

El modelo obtenido hace un buen ajuste a los datos y ha superado la etapa de validación, además presenta buena calidad predictiva, se observa que predice de forma similar en la muestra de validación como lo hace en la muestra de análisis, por lo que puede ser extendido a toda la población de hogares con las variables referidas a ellos y su jefatura esperando un porcentaje global de aciertos del 93,3%.

Además el modelo estimado permite predecir con una muestra más pequeña y con un margen de error pequeño los porcentajes de cada nivel de pobreza: "pobre extremo", "pobre relativo" y "no pobre" por Zona (Rural y Urbana), Región, Departamentos y municipios.

Referencias Bibliográficas.

[1] Abarca, Oscar (2010): Desarrollo de un modelo de geoprocesamiento para la valoración productiva y tributaria de tierras agrícolas en Venezuela. Tesis doctoral, Universidad politécnica de Madrid. Madrid, España.

[2] Álvarez Fernández Antonio, 2007, Memoria presentada como requisito para obtener el Diploma de Estudios Avanzados, Universidad de Almería, España.

[3] Agresti A. (2002), Categorical Data Analysis. Second Edition ed. New York: Wiley.

[4] Aguilera del Pino, A. M. (2002). Modelos de respuesta Discreta. Granada: Copias Coca, Dep. Legal GR-11554-02.

[5] Andersen E. (1990). The Statistical Analysis of Categorical Data. New York: Springer-Verlag.

[6] Barrios Sosa, Yolanda (2006): "Determinantes de la pobreza en los hogares con adultos mayores. Costa Rica, 2005". Trabajo final de graduación presentado a la Escuela de Estadística, para optar al título de Máster en Población y Salud. Universidad de Costa Rica. Costa Rica.

[7] Blázquez Zaballos Antonio, 2006, Modelos con respuesta binaria: La regresión logística, Salamanca España.

[8] Canizales Rivera, Carlos Ernesto, 2012, Métodos robustos aplicados a la clasificación del estado nutricional de la niñez salvadoreña (FESAL 2008), Universidad de El Salvador, El Salvador.

[9] Clark, W. y Hosking, P. (1986). Statistical Methods for geographers. New York: John Wiley & Sons. 518 pp.

[10] Debella-Gilo, M.; Etzelmuller, B; Klakegg, O. (2007). Digital soil mapping using Digital Terrain analysis and statistical modelling integrated into GIS: Examples from Vestfold County of Norway. ScanGIS´2007 – Proceedings of the 11th Scandinavian

Research Conference on Geographical Information Sciences (pp. 237-253). As-Noruega: University of Life Sciences.

[11] Dueñas Rodríguez, María Ángeles, 2006, Modelos de respuesta discreta en R y aplicación con datos reales, Universidad de Granada, España.

[12] Eastman, R. Idrisi Andes (2006). Guide to GIS and Image Processing. Worcester, Ma: Clark University. 327 pp.

[13] Encuesta de Hogares de Propósitos Múltiples 2010, Dirección General de Estadística y Censos, Ministerio de Economía, República de El Salvador.

[14] Espinoza Silicia, Núñez Jairo (2005), Determinantes de la pobreza y la vulnerabilidad. Misión para el Diseño de una Estrategia para la Reducción de la Pobreza y la Desigualdad.

[15] Facultad Latinoamericana de Ciencias Sociales (FLACSO) Programa El Salvador, Ministerio de Economía (MINEC), Programa de las Naciones Unidas para el Desarrollo (PNUD).2010: Mapa de pobreza urbana y exclusión social El Salvador. Volumen 1. Conceptos y metodología. San Salvador, El Salvador.

[16] Fagerland MW, Hosmer DW, Bofin AM, (2008). Multinomial goodness-of-fit tests for logistic regression models. Stat Med 2008 Sep 20;27(21):4238-4253.

[17] García Mendoza Saúl, Misari Atanancio Julissa, Villacorta Olazabal Mirlena (2011): Perú: Determinantes de la pobreza, 2009. Instituto Nacional de Estadística e Información. Centro de Investigación y Desarrollo (CIDE). Lima, Perú.

[18] Gontero Sonia, Ojeda Silvia, Pereyra Liliana E.(2005): La pobreza en los hogares del Gran Córdoba: aplicación del modelo de regresión logística. Universidad Nacional de Córdoba. Instituto de Economía y Finanzas Facultad de Ciencias Económicas y Facultad de Matemática Astronomía y Física. Argentina.

[19] Gonzales Fernández María de Lourdes, Pérez Izquierdo Victoria. Determinantes de la pobreza y la vulnerabilidad económica en Cuba. Un estudio empírico. Instituto Nacional de Investigaciones Económicas (INIE). Cuba.

[20] Lemus Gómez Oscar Hernán, Lemus Gómez Rolando (2005). "Construcción de un modelo de regresión logístico sobre la oferta laboral a jefes(as) de hogares en El Salvador". Trabajo de graduación para optar al grado de: maestro en estadística. Universidad de El Salvador, San Salvador.

[21] McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour. En. P. Zarembka (Ed.), Frontiers in econometrics, New York, Academic Press:105-142.

[22] Menard, S. (2002). Applied Logistic Regression Analysis. Series: Quantitative Applications in the Social Sciences, N° 7-106. Thousand Oaks, California-London-New Delhi: Sage Publications. 111pp.

[23] Ministerio de Economía de El salvador (2009). Midiendo la pobreza en El Salvador: Valoraciones conceptuales y desafíos metodológicos. México.

[24] Morales González, Domingo, 2004, Modelos Lineales Generalizados, Departamento de Estadística y Matemática Aplicada, Universidad Miguel Hernández de Elche, España.

[25] Pando Fernández V, San Martín Fernández R. Regresión logística multinomial. Cuad Soc Esp Cien For 2004;18.

[26] Peña Aguilar René Armando, 2009, Comparación entre el análisis discriminante y la regresión logística en la clasificación de una colonia de cangrejos herradura (Limulus polyphemus). Anteproyecto de Tesis para optar al grado de Maestro en Estadística, Universidad de El Salvador. El Salvador.

[27] Silva Ayçaguer LC. Excursión a la regresión logística en ciencias de la salud. Madrid: Díaz de Santos; 1995.

[28] Teitelboim, Berta, 2006, Factores concluyentes de la pobreza en base a un modelo logístico. Tesis para optar al Grado de Magíster en Bioestadística. Universidad de Chile. Santiago de Chile.