UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

REGRESIÓN LOGÍSTICA MULTINOMIAL

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

A C T U A R I O

P R E S E N T A:

JORGE ARTURO MUÑOZ ARISTIZABAL

DIRECTORA DE TESIS:
Mat. MARGARITA ELVIRA CHÁVEZ CANO
Junio 2017

Ciudad Universitaria, CD. MX.

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

1.Datos del alumno
Muñoz
Aristizabal
Jorge Arturo
5527023150
Universidad Nacional Autónoma de México
Facultad de Ciencias
Actuaría
307189846

2. Datos del tutor
Mat.
Margarita Elvira
Chávez
Cano

3. Datos del sinodal 1
Dra.
Ruth Selene
Fuentes
García

4. Datos del sinodal 2
Dra.
Lizbeth
Naranjo
Albarrán

5. Datos del sinodal 3
Act.
Jaime
Vázquez
Alamilla

6. Datos del sinodal 4
Act.
Francisco
Sánchez
Villarreal

7.Datos del trabajo escrito.
Regresión Logística Multinomial
104 p
2017

Agradecimientos

A la Facultad de Ciencias por todo el aprendizaje recibido, las amistades que me ha dado, y por ser mi segundo hogar por tanto tiempo. Siempre llevaré un grato recuerdo de todo lo vivido en ella.

A los profesores de mi carrera universitaria que me han transmitido su valioso conocimiento.

A mis sinodales por su tiempo y comentarios para este trabajo.

A la profesora Margarita por su paciencia, tiempo y dedicación en este trabajo. Gracias por sus enseñanzas y consejos en lo académico y personal.

A Luis Rodrigo Gallardo, Mauricio Estrada, Rodrigo Sánchez, Néstor Issel Hernandez, León Felipe Gómez, María Fernanda Olmedo, Axel Moreno, Mónica Aguirre, y Jorge Aurelio Carreño por su amistad de tantos años, y que a pesar de la distancia sé que estarán siempre a mi ahí.

A Tania por toda la paciencia y amor que me has brindado en este viaje juntos.
Tus consejos siempre me ayudan a ser una mejor persona.
Esta aventura solo está empezando.

A mi familia por su apoyo y cariño incondicional en todas las etapas de mi vida.
Por apoyarme y aconsejarme cuando más lo necesito, sin ellos no hubiera podido llegar tan lejos. Siempre estarán presentes en mi corazón.

Índice

Introducción

El análisis de regresión es una de las herramientas estadísticas más utilizadas para analizar datos donde se desee encontrar alguna relación entre diferentes variables. Dependiendo del tipo de datos que se tenga para realizar dicho análisis será el tipo de regresión que se utilizará.

En este trabajo se presenta uno de los modelos más utilizados en el análisis de regresión, la regresión logística, donde se da mayor importancia al desarrollo de la metodología de la regresión logística multinomial.

La regresión logística multinomial es el tema principal en este trabajo, y se presenta el desarrollo intuitivo sobre el modelo, así como la estimación de los parámetros, y la bondad de ajuste del mismo. Presentando ejemplos para ilustrar el tema.

Los datos que se presentan son extraídos de las bases de datos del estudio *Adversidad psicosocial, psicopatología y funcionamiento en hermanos adolescentes en alto riesgo (HAR) con y sin* ***trastorno por déficit de atención con hiperactividad*** *(TDAH)*[17], con las variables de Folio, Expediente, Peso, Talla, IMC, Edad, Hermano probando, Sexo, Hospital, TDAHINATDXS, y TDAHMIXTDXS.

En el primer capítulo se desarrolla el modelo de regresión logística simple para el caso binario, donde se da una explicación sobre el modelo desde su deducción hasta sus pruebas de bondad de ajuste. (En ocasiones, teniendo como referencia la regresión lineal para hacer una comparación del mismo).

En el segundo capítulo se describe el modelo de regresión logística múltiple, con más de K variables independientes.

En el último capítulo se desarrolla el modelo de regresión logística multinomial. Dicho modelo no es una generalización del modelo de regresión logística para el caso binario, pero el caso binario se puede obtener como un caso particular de este modelo. Para el modelo multinomial, la variable dependiente puede tomar J valores diferentes.

El propósito de este trabajo es ejemplificar y explicar de forma clara y concreta el modelo de regresión logística multinomial, el cual es ampliamente utilizado. Cabe mencionar que encontrar ejemplos y literatura de ese modelo es difícil ya que la información es escasa y solo está hecha para casos particulares.

Capítulo 1

1 Regresión logística binaria

1.1 Introducción

En estadística los métodos de regresión son un componente integral dentro de cualquier análisis de datos enfocado a describir la relación entre una variable dependiente y una o más variables independientes. Sin dejar de lado el poder hacer estimaciones a partir de la información previa que proporcione una muestra aleatoria. Un ejemplo seria la regresión lineal en donde se busca una relación entre una variable dependiente continua y una o varias variables dependientes.

En la práctica el uso de la regresión lineal y la regresión logística tienen mucha semejanza, aunque sus enfoques matemáticos son muy diferentes se diferencian principalmente por la variable dependiente.

En el modelo de regresión logística, la variable dependiente es dicotómica, y la variable independiente puede ser categórica o de razón. La ecuación del modelo no es una función lineal sino exponencial, ya que, debido a una simple transformación logarítmica, puede presentarse como una función lineal.

De esta forma el modelo será útil para situaciones de investigación en donde la respuesta pueda tomar únicamente dos valores: *ausencia o presencia de un determinado evento*. Cabe mencionar que el modelo permite que la variable independiente pueda tomar valores categóricos o de razón.

Ejemplo 1.1

A continuación se muestra una tabla de frecuencias con una muestra de 121 pacientes sobre la variable *Edad*, en referencia la base de datos del estudio

"*Adversidad psicosocial, psicopatología y funcionamiento en hermanos adolescentes en alto riesgo" (HAR) con y sin trastorno por déficit de atención con hiperactividad (TDAH)*[16], en donde se usan las variables de Edad, TDAH, y TDAHMIXTDXS. En este ejemplo se utilizan solo los datos de pacientes diagnosticados con algún tipo de TDAH, tomando en cuenta que su edad es menor a 21 años. Si un paciente no fue diagnosticado con TDAH mixto será codificado con 0 en la variable de TDAHMIXTDXS y 1 si presenta dicho diagnóstico.

Edad	Frecuencia	Porcentaje	Porcentaje acumulado
12	4	3%	3%
13	26	21%	25%
14	17	14%	39%
15	22	18%	57%
16	20	17%	74%
17	16	13%	87%
18	4	3%	90%
19	6	5%	95%
20	6	5%	100%
Total	121	100.0	

Tabla 1.1 de Frecuencias por Edad

En este caso nos interesa saber la relación entre la edad y la presencia o ausencia de TDAH mixto. La dificultad que se observa con los datos presentados, es la naturaleza de la variable dependiente, es decir, la variable de presencia y ausencia de TDAH mixto, ya que al graficarla queda como se muestra el Figura 1.1. Esta figura no muestra una relación directa con la variable independiente, por lo que se buscará una transformación en los datos para pasar a una medida de proporción como se muestra en la tabla 1.2.

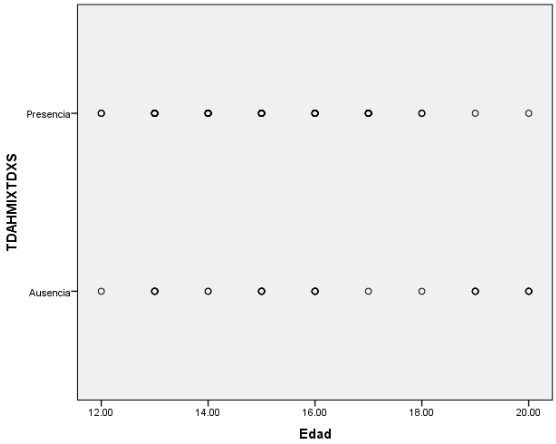

Figura 1.1 Diagrama de dispersión de Tabla 1.1.

La tabla 1.2 contiene la frecuencia de cada evento por cada año de edad, así como la media (o proporción con presencia de TDAH mixto) de cada grupo.

Edad	0 Ausencia	1 Presencia	Total	Proporción
12	1	3	4	0.75
13	7	19	26	0.73
14	2	15	17	0.88
15	8	14	22	0.63
16	8	12	20	0.6
17	1	15	16	0.94
18	1	3	4	0.75
19	5	1	6	0.16
20	5	1	6	0.16
Total	38	83	121	0.686

Tabla 1.2 de Frecuencias por Edad por TDAH

En la tabla 1.2 se observa que entre mayor edad hay menor presencia de TDAH mixto. La Figura 1.2 presenta la gráfica sobre la proporción de pacientes con TDAH mixto contra la edad de estos.

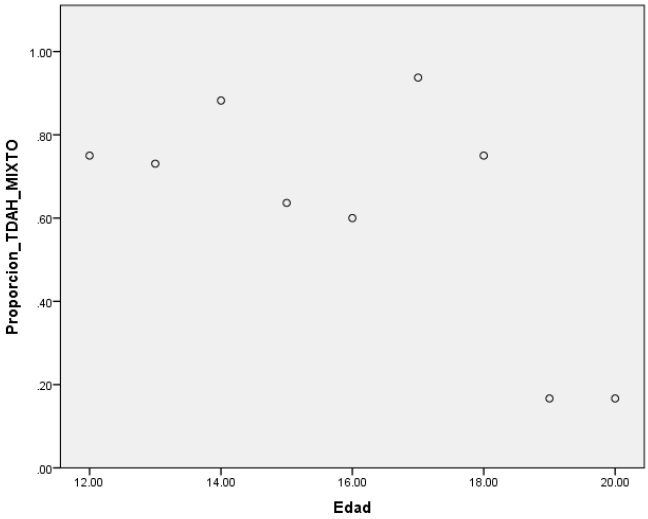

Figura 1.2 Diagrama de dispersión de Tabla 1.2.

Por lo que, es importante encontrar una distribución que describa el estudio. La gráfica que se muestra en la figura 1.2 presenta datos con cierta tendencia, lo que nos lleva a buscar un modelo de regresión.

1.2 El modelo logístico binario

En un modelo de regresión se expresa la esperanza de la variable dependiente dado el valor de la variable independiente. Como una ecuación lineal:

$$E(Y_i|X = x_i) = \beta_0 + \beta_1 x_i$$

Pero la expresión anterior implica que es posible que el valor de $E(Y_i)$ pueda tomar cualquier valor entre $-\infty$ y ∞.

La columna de "Proporción" de la tabla 1.2 da muestra un valor estimado de $E(Y_i|X = x_i)$. Se supone que el valor estimado en la figura 1.2 está suficientemente

cercano al valor real de $E(Y_i)$ para proveer una relación razonable entre el TDAH y la Edad. Con la información que se tiene en forma dicotómica, la esperanza condicional tiene que ser mayor o igual a cero y menor o igual a uno. El cambio por unidad en X tiende a ser más pequeño cuando la esperanza condicional se acerca a cero o a uno.

Gran parte de las bondades del modelo logístico se heredan de la función logística, que describe la base del modelo. Dicha función es la siguiente:

$$f(z) = \frac{1}{1 + e^{-z}} = \frac{e^{z}}{1 + e^{z}}$$

con z que toma valores entre $(-\infty, \infty)$.

En esta clase de estudio se utilizar la distribución logística, la cual se denota de la siguiente forma $\pi(x_i) = E(Y|X = x_i) = P(Y = 1|X = x_i)$ para la esperanza condicional de Y dado un valor en X usando la distribución logística. La forma específica del modelo de regresión logística está dada de la siguiente forma:

$$\pi(x_i) = P(Y = 1|X = x_i)$$

$$= \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

La transformación de esta función a una forma lineal es:

$$g(x_i) = \ln\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \beta_1 x_i$$

ya que si:

$$\alpha = \beta_0 + \beta_1 x_i$$

y

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

entonces:

$$g(x_i) = ln\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = ln\left(\frac{\frac{e^{\alpha}}{1+e^{\alpha}}}{1-\frac{e^{\alpha}}{1+e^{\alpha}}}\right)$$

$$= ln\left(\frac{\frac{e^{\alpha}}{1+e^{\alpha}}}{\frac{1+e^{\alpha}-e^{\alpha}}{1+e^{\alpha}}}\right)$$

$$= ln(e^{\alpha}) = \alpha = \beta_0 + \beta_1 x_i$$

La importancia de esta transformación es que se obtiene un modelo lineal, que hereda propiedades para poder hacer un análisis de regresión lineal. La función $g(X_i)$ es lineal en sus parámetros, es continua, y tiene valores dentro del rango de $(-\infty, \infty)$ dependiendo del valor de x_i.

1.2.1 Objetivos de la regresión logística

El objetivo primordial de la regresión logística es modelar la influencia de las variables independientes en la probabilidad de ocurrencia de un suceso particular. También tiene como objetivo investigar su influencia en la probabilidad de ocurrencia de un suceso, la presencia o no de diversos factores y el valor o nivel de los mismos. Y, como último objetivo poder determinar el modelo más adecuado

según el ajuste, que, describa mejor la relación entre la variable dependiente y el conjunto de variables independientes.

La regresión lineal está basada en una distribución normal con varianza constante de los errores, pero no es el caso con una variable dependiente dicotómica. En este caso se tiene que, el valor la variable dependiente dada la variable independiente *x*, se puede escribir como $Y = \pi(x) + \varepsilon$. Aquí la cantidad ε, se puede suponer que toma dos posibles valores, si $Y = 1$ entonces $\varepsilon = 1 - \pi(x)$ con una probabilidad de $\pi(x)$, y si $Y = 0$ entonces $\varepsilon = -\pi(x)$ con probabilidad de $1 - \pi(x)$. ε tiene una distribución con media cero y varianza $\pi(x)[1 - \pi(x)]$. La distribución condicional de Y, denotada por $f(Y|x)$, se distribuye como una variable aleatoria $Binomial\big(1, \pi(x)\big)$ [11].

Los *umbrales* son propiedades importantes de la función logística. Si se empieza a evaluar la función desde $z = -\infty$ se tiene que los valores de $f(z)$ son cercanos a 0, luego $f(z)$ incrementa drásticamente hasta llegar a valores cercanos a 1 donde su aumento se estabiliza hasta $z = \infty$. A estos puntos de inflexión se les llama umbrales.

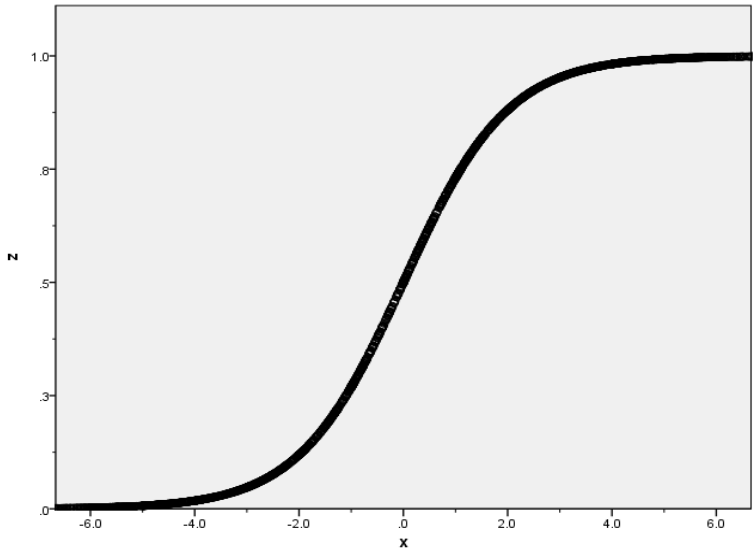

Figura 1.3 Simulación función logística

El modelo logístico con una variable independiente resulta de aplicar la función a la combinación lineal de x, $z = \beta_0 + \beta_1 x$, en donde β_0 y β_1 son términos constantes que representan parámetros desconocidos.

Por lo Tanto:

$$f(z) = f(\beta_0 + \beta_1 x) = \frac{1}{1 + exp\big(-(\beta_0 + \beta_1 x)\big)} = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}$$

1.3 Ajuste del modelo de regresión logística

1.3.1 Modelo lineal de probabilidad

Sea una muestra de n observaciones independientes de la pareja (x_i, y_i) $i = 1,2, \dots, n,$ donde y_i es el valor de la variable dependiente dicotómica y x_i es el valor de la variable independiente para el i −ésimo entrevistado.

Para una respuesta binomial, el modelo de regresión:

$$E(Y|X = x) = \pi(x) = \beta_0 + \beta_1 x$$

el cual se le conoce como un modelo de probabilidad lineal. Cuando las observaciones en Y son independientes, se dice que este modelo es un Modelo Lineal Generalizado (MLG) con componentes aleatorios binomiales y función de liga identidad. Este modelo tiene los siguientes inconvenientes:

- La probabilidad debe estar entre 0 y 1, mientras que las funciones lineales toman valores sobre la recta real completa.

- El modelo toma $\pi(x) < 0$ y $\pi(x) > 1$ para valores suficientemente grandes o pequeños de x.

- La relación entre x y $\pi(x)$ es no lineal.

El modelo puede ser válido sobre un rango finito de valores de x, sin embargo, puede tener problemas si se usa el método de mínimos cuadrados, pues las condiciones que hace que los estimadores sean óptimos no son satisfactorias. Por esta razón se estudia una función que tiene la forma de una $\boldsymbol{S}$, y por lo tanto se usa $\pi(x)$ de la siguiente forma [23].

La probabilidad de que Y tome el valor de la categoría 1 es:

$$P(Y = 1|X = x) = \pi(x) = \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)} \quad (1.3)$$

La probabilidad de que Y tome el valor de la categoría 0 es:

$$P(Y = 0|X = x) = 1 - \pi(x) = \frac{1}{1 + exp(\beta_0 + \beta_1 x)}$$

Cuando $\beta_1 < 0$,

$$\lim_{x\to\infty} \pi(x) = \lim_{x\to\infty} \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)} = 0$$

y cuando $\beta_1 > 0$

$$\lim_{x\to\infty} \pi(x) = \lim_{x\to\infty} \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)} = 1$$

La función de regresión logística tiene como primera derivada:

$$\frac{\partial \pi(x)}{\partial x} =$$

$$= \frac{\big(1 + exp(\beta_0 + \beta_1 x)\big)\beta_0 exp(\beta_0 + \beta_1 x) - ex\,p(\beta_0 + \beta_1 x)\beta_0 ex\,p(\beta_0 + \beta_1 x)}{(1 + exp(\beta_0 + \beta_1 x))^2}$$

$$= \frac{\big(1 + exp(\beta_0 + \beta_1 x) - exp(\beta_0 + \beta_1 x)\big)\beta_0 exp(\beta_0 + \beta_1 x)}{(1 + exp(\beta_0 + \beta_1 x))^2}$$

$$= \frac{\beta_0 exp(\beta_0 + \beta_1 x)}{(1 + exp(\beta_0 + \beta_1 x))^2} = \frac{\beta_0 \pi(x)}{1 + exp(\beta_0 + \beta_1 x)}$$

$$= \beta_0 \pi(x) \left(1 + \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}\right) = \beta_0 \pi(x)[1 - \pi(x)]$$

De la igualdad anterior se obtiene la segunda derivada de la función de regresión logística:

$$\frac{\partial^2 \pi(x)}{\partial x^2} = \beta_0 \pi(x)\big(-\pi'(x)\big) + \beta_0 \pi'(x)\big(1 - \pi(x)\big)$$

$$= \pi'(x)\left(-\beta_0 \pi(x) + \beta_0\big(1 - \pi(x)\big)\right)$$

$$= \pi'(x)\{\beta_0 - 2\beta_0 \pi(x)\}$$

A continuación, se igualará a cero la segunda derivada para obtener el punto donde se maximiza la función de regresión logística:

$$\frac{\beta_0 exp(\beta_0 + \beta_1 x)}{(1 + exp(\beta_0 + \beta_1 x))^2} \cdot \left(\beta_0 - 2\beta_0 \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}\right) =$$

$$\frac{\beta_0^2 exp(\beta_0 + \beta_1 x)}{(1 + exp(\beta_0 + \beta_1 x))^2} = 2\beta_0^2 \frac{(exp(\beta_0 + \beta_1 x))^2}{\big(1 + exp(\beta_0 + \beta_1 x)\big)^3}$$

$$\frac{2 \cdot exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)} = 1$$

Por lo tanto:

$$\frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)} = \frac{1}{2}$$

Del resultado anterior se obtiene que $exp(-\beta_0 - \beta_1 X) = 1$, y $x = -\frac{\beta_0}{\beta_1}$.

Por lo tanto, la función se maximiza cuando $\pi(x) = \frac{1}{2}$, y $x = -\frac{\beta_0}{\beta_1}$.

La ecuación en (1.3) no es lineal en los parámetros, por lo tanto, al utilizar la función *logit*, $g(x)$, es posible estimar los parámetros en el modelo de regresión lineal.

$$g(x) = ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x \qquad (1.4)$$

El método para estimar los parámetros en la regresión logística es diferente al utilizado en la regresión lineal, en el que se usa el método de mínimos cuadrados. En la regresión logística se estiman los parámetros por máxima verosimilitud.

1.3.2 Método de máxima verosimilitud

Sea una muestra de n observaciones independientes de la pareja (x_i, y_i) $i = 1,2, \dots, n$, donde Y_i es el valor de la variable dependiente dicotómica y x_i es el valor de la variable independiente para el i −ésimo entrevistado. Si $\boldsymbol{\beta} = (\beta_0, \beta_1)'$, y además se tiene que para la pareja observada (x_i, y_i) la probabilidad se puede expresar de la siguiente manera de acuerdo a la función de densidad de probabilidad $Binomial(1, \pi(x))$:

$$P[Y = y_i] = \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i}$$

Entonces la función de verosimilitud de esta muestra está dada por:

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} \tag{1.5}$$

Aplicando el logaritmo natural en la ecuación (1.5) da como resultado la función de log-verosimilitud:

$$\mathcal{L}(\boldsymbol{\beta}) = ln\big(\ell(\boldsymbol{\beta})\big) =$$

$$= \sum_{i=1}^{n} y_i \, ln\big(\pi(x_i)\big) + \sum_{i=1}^{n} (1 - y_i) ln(1 - \pi(x_i)) \tag{1.6}$$

$$= \sum_{i=1}^{n} y_i \, ln\big(\pi(x_i)\big) - \sum_{i=1}^{n} y_i \, ln\big(1 - \pi(x_i)\big) + \sum_{i=1}^{n} ln(1 - \pi(x_i))$$

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

$$= \sum_{i=1}^{n} y_i \, (\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} ln(1 + exp(\beta_0 + \beta_1 x_i))$$

Se obtiene la derivada con respecto a β_0 y β_1 y se iguala a cero

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}$$

$$= \sum_{i=1}^{n} [y_i - \pi(x_i)] \tag{1.7}$$

y con respecto a β_1 la derivada queda de la siguiente forma:

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_1} = \sum_{i=1}^{n} y_i \, x_i - \sum_{i=1}^{n} x_i \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}$$

$$= \sum_{i=1}^{n} x_i [y_i - \pi(x_i)] \tag{1.8}$$

Los estimadores de máxima verosimilitud para β_0 y β_1 se obtienen igualando a cero las ecuaciones de (1.7) y (1.8) y resolviendo simultáneamente.

En el modelo de regresión lineal, las ecuaciones de verosimilitud obtenidas mediante la derivada de la suma de cuadrados con respecto a $\boldsymbol{\beta}$ son lineales en los parámetros desconocidos y por ello son fáciles de obtener. Para la regresión logística las expresiones en (1.7) y (1.8) son no lineales en β_0 y β_1, por lo tanto, requieren un método especial para su solución. Los métodos utilizados son de naturaleza iterativa que han sido programados en diferentes softwares estadísticos, el método más utilizado es el de Newton-Raphson, que se planteará más adelante.

El valor de $\boldsymbol{\beta}$ dada la solución en (1.7) y (1.8) y se denota como $\widehat{\boldsymbol{\beta}}$. En general el uso del símbolo "^" denota al estimador por máxima verosimilitud de un parámetro. Por ejemplo, $\hat{\pi}(x_i)$ es estimador por máxima verosimilitud de $\pi(x_i)$.

$$\hat{\pi}(x_i) = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}$$

Esta cantidad da un valor estimado de la probabilidad condicional de Y igual a 1, dado que $X = x_i$.

Como tal, representa el valor ajustado o estimado por el modelo de regresión logística. Una consecuencia interesante de la ecuación en (1.7) es que:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\pi}(x_i) \tag{1.9}$$

Esto es, la suma de los datos observados $\boldsymbol{y}$ que es igual a la suma de los valores estimados [11].

Ejemplo 1.2

Del conjunto de datos del ejemplo 1.1 el modelo de regresión logística con el software estadístico SPSS, se toma como variable independiente la *Edad*, y la ausencia o presencia de TDAH mixto como variable dependiente.

Variables	B	Exp(B)
Edad	-.249	.780
Constante	4.633	102.862

Tabla 1.3 Coeficientes estimados de la regresión logística

Los valores de los estimadores por máxima verosimilitud de los coeficientes β_0 y β_1 están dadas por $\hat{\beta}_0 = 4.633$ y $\hat{\beta}_1 = -0.249$. La ecuación para el valor estimado de la probabilidad condicional de $Y = 1$ dado que $X = x_i$ para el paciente i, es:

$$\hat{\pi}(x_i) = \frac{exp(4.633 - 0.249 \times Edad_i)}{1 + exp(4.633 - 0.249 \times Edad_i)}$$

Lo que puede apreciarse del ejemplo anterior es que se confirma la hipótesis planteada en el ejemplo 1.1, en donde a mayor edad menor presencia de TDAH mixto, aspecto que se refleja por el signo negativo en $\hat{\beta}_1$.

1.4 Pruebas de hipótesis

Existen varias formas de calcular la prueba de bondad de ajuste de un modelo de regresión logística. Entre las más usadas son la estadística de log-verosimilitud, y la estadística de Wald, usualmente para evaluar el ajuste del modelo y por lo que son consideradas como medidas de bondad de ajuste.

El principio usado en la bondad de ajuste de la regresión logística, es el mismo que se utiliza en la regresión lineal (comparar los valores observados de la variable de respuesta con los valores ajustados obtenidos de los modelos con la variable en cuestión). En la regresión logística la comparación de los valores observados contra los valores ajustados está basada en la función *log-verosimilitud* definida en la ecuación (1.7). Para entender mejor dicha comparación se debe tomar en cuenta que un valor observado de la variable de respuesta es también un valor ajustado.

1.4.1 Prueba del cociente de verosimilitudes

En la prueba del cociente de verosimilitudes la hipótesis que se desea probar es:

$$H_0: \beta_1 = 0 \quad vs \quad H_a: \beta_1 \neq 0$$

El principio en la regresión logística es: *comparar los valores observados de la variable dependiente para estimar los valores obtenidos a partir del modelo con y*

sin la variable en cuestión. En donde la comparación de los valores observados contra los estimados se basa en la función log-verosimilitud definida en (1.6).

Un modelo saturado es aquel que incluye el mismo número de observaciones que parámetros en el modelo. El cociente de verosimilitudes queda de la siguiente manera:

$$LR = \frac{verosimilitud\ del\ modelo\ ajustado}{verosimilitud\ del\ modelo\ saturado} \tag{1.10}$$

En un modelo saturado, por definición del modelo mismo, se tiene que $E(Y|X) = y$ donde los valores observados son iguales a los valores estimados del modelo consecuentemente, en el modelo logístico se tiene que en un modelo saturado $\widehat{\pi(x_i)} = y_i$, por lo que la verosimilitud será

$$l(modelo\ saturado) = \prod_{i=1}^{n} {y_i}^{y_i}(1 - y_i)^{1-y_i} = 1 \tag{1.11}$$

la comparación de los valores observados con los valores esperados utilizando la función de verosimilitud está basada en la estadística D llamada devianza, definida como:

$$D = -2\ ln(LR) = -2l\,n\left(\frac{verosimilitud\ del\ modelo\ ajustado}{verosimilitud\ del\ modelo\ saturado}\right) \tag{1.12}$$

Es necesario utilizar menos dos veces el logaritmo natural de la ecuación (1.10) para obtener una estadística cuya distribución sea conocida y por lo tanto se pueda usar para probar la hipótesis. Tal prueba es conocida como distribución asintótica del cociente de verosimilitudes. De la ecuación (1.6), y la ecuación (1.12) queda de la siguiente forma:

$$D = -2\sum_{i=1}^{n}\left(y_i\, ln\left(\frac{\hat{\pi}(x_i)}{y_i}\right) + (1-y_i) ln\left(\frac{(1-\hat{\pi}(x_i))}{1-y_i}\right)\right)$$

La devianza es equivalente a la suma de los cuadrados de residuales en la regresión lineal. De este modo, la devianza, como se expresa en la ecuación (1.12), y al calcularse para la regresión lineal, es idéntica a la *Suma de Cuadrados del Error*. Para probar la significancia de las variables independientes, se compara el valor de D obtenido al incluir la variable en el modelo y el valor obtenido al excluirla [11].

$$G = D(modelo\ excluyendo\ la\ variable) - D(modelo\ incluyendo\ la\ variable)$$

$$= -2ln\left(\frac{verosimilitud\ excluyendo\ la\ variable}{verosimilitud\ incluyendo\ la\ variable}\right)$$

El papel que juega esta estadística en la regresión logística podría verse como el numerador en la prueba F parcial en la regresión lineal, lo cual se debe a que la verosimilitud del modelo saturado es común en los dos valores de la estadística D de donde se obtiene la diferencia para la estadística G. La estadística G se distribuye aproximadamente como una ji-cuadrada con un grado de libertad por la transformación en $-2ln(x)$.

Para calcular la verosimilitud del modelo excluyendo la variable independiente, se tiene que, al suponer $\beta_1 = 0$, la probabilidad de éxito de la variable de respuesta, denotada con $\hat{\pi}(x)'$, estará dada por:

$$\hat{\pi}(x)' = \frac{exp(\hat{\beta}_0)}{1 + exp(\hat{\beta}_0)}$$

Por lo que se debe encontrar el estimador máximo verosímil de β_0 bajo este modelo. La log-verosimilitud del modelo excluyendo las variables independientes es:

$$\mathcal{L}(\beta_0) = \sum_{i=1}^{n} \left(y_i \, ln\left(\frac{exp(\beta_0)}{1+exp(\beta_0)}\right) + ((1+y_i) ln\left(\frac{1}{1+exp(\beta_0)}\right)\right)$$

$$= \sum_{i=1}^{n} (y_i \, \beta_0 - ln(1+exp(\beta_0)))$$

entonces, la derivada parcial de β_0 es:

$$\frac{\partial \mathcal{L}(\beta_0)}{\partial \beta_0} = \sum_{i=1}^{n} \left(y_i - \frac{exp(\beta_0)}{1+exp(\beta_0)}\right) = 0$$

por lo que el estimador máximo verosímil de $\widehat{\pi(x)}'$ es:

$$\hat{\pi}(x)' = \frac{\sum_{i=1}^{n} y_i}{n}$$

de donde:

$$\hat{\pi}(x)' = \frac{n_1}{n} \qquad y \qquad (1-\hat{\pi}(x)') = \frac{n_0}{n}$$

donde:

$$n_1 = \sum_{i=1}^{n} y_i \qquad y \qquad n_0 = \sum_{i=1}^{n} (1-y_i)$$

Aplicando la función *logit* como en la ecuación (1.4) y por propiedades de los estimadores máximo verosímiles se tiene:

$$ln\left(\frac{\hat{\pi}(x)'}{1-\hat{\pi}(x)'}\right) = \hat{\beta}_0$$

Por lo que el estimador máximo verosímil de β_0, bajo el modelo excluyendo la variable independiente es:

$$\hat{\beta}_0 = ln\left(\frac{n_1}{n_0}\right)$$

De lo anterior, se tiene que la verosimilitud del modelo excluyendo la variable independiente es:

$$verosimilitud\ excluyendo\ la\ variable = \prod_{i=1}^{n}\left(\left(\frac{n_1}{n}\right)^{y_i}\left(\frac{n_0}{n}\right)^{(1-y_i)}\right)$$

$$= \left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}$$

por lo tanto, para probar que:

$$H_0: \beta_1 = 0 \quad vs \quad H_a: \beta_1 \neq 0$$

Se utiliza la estadística:

$$G = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}(x_i)^{y_i}(1-\hat{\pi}(x_i))^{1-y_i}}\right]$$

$$= 2\left(\sum_{i=1}^{N}\left(y_i ln(\hat{\pi}(x_i)) + (1-y_i)ln(1-\hat{\pi}(x_i))\right)\right.$$

$$\left. - \left(n_1 ln(n_1) + n_0 ln(n_0) - nln(n)\right)\right) \qquad (1.13)$$

que tiene una distribución ji-cuadrada con un grado de libertad, que denota $G \sim \chi_1^2$. Para un nivel de significancia α, se rechazará la hipótesis nula si $G > \chi_1^{2^{(1-\alpha)}}$.

Ejemplo 1.3

Se tiene en cuenta los datos del ejemplo 1.1, se sabe que $n_1 = 83$ y $n_0 = 38$.Al obtener los parámetros estimados del modelo con el software de SPSS se obtiene también el resumen del modelo donde $-2\log(verosimilitud\ incluyendo\ las\ variables)$ que tiene un valor de -71.925. Se evalúa G en la ecuación en (1.13), por lo que se obtiene:

$$G = 2\left(-71.925 - \left(83ln(83) + 38ln(38) - 121ln(121)\right)\right) = 6.7472$$

$$P[G > 6.7472] = .009 = p - value$$

Con lo anterior se evidencia que la variable independiente (*Edad*) es una variable significativa para predecir la variable TDAH mixto.

La estadística G tiene una relación cercana con la usada en la regresión lineal, es así que, la suma de cuadrados de residuales, el cual es un análogo a la medida de

devianza en la regresión lineal, puede ser vista como un análogo a G, la cual es la medida de la devianza para la función *logit*. De la misma forma, la estadística F usada en la regresión lineal puede ser vista como el análogo a la estadística ji-Cuadrada en la regresión logística.

1.4.2 Prueba de Wald

Al igual que la prueba del cociente de verosimilitudes, la hipótesis que se desea probar en la prueba de Wald es:

$$H_0: \beta_1 = 0 \quad vs \quad H_a: \beta_1 \neq 0$$

Esta prueba se obtiene al hacer el cociente del parámetro estimado por máxima verosimilitud con su error estándar estimado, donde la razón resultante sigue una distribución aproximada de una normal estándar.

Para probar la significancia de un coeficiente de regresión individual, β_1 utiliza la prueba de Wald univariada. La estadística de prueba se expresa como:

$$W = \frac{\hat{\beta}_1}{\sqrt{\widehat{Var(\hat{\beta}_1)}}}$$

Por lo que, a un nivel de significancia α se rechazará H_0 si

$$W > \left| z_{\left(1-\frac{\alpha}{2}\right)} \right|$$

donde $z_{\left(1-\frac{\alpha}{2}\right)}$ es el cuantil $\left(1-\frac{\alpha}{2}\right)$ de una distribución normal con media igual a 0 y desviación estándar igual a 1.

Una prueba de Wald significativa sugiere que la variable independiente tiene un efecto en la variable dependiente, la cual es fácil de calcular e interpretar, tomando en cuenta que debería ser usada con precaución debido a que tiende a sobrestimar la significancia de la variable independiente cuando su coeficiente tiene una magnitud grande y puede también ser de poca confianza cuando la muestra es pequeña.

1.5 Intervalos de Confianza

Se presentarán los intervalos de confianza de los parámetros del modelo. Los cálculos de los intervalos están basados en una aproximación al método de máxima verosimilitud, que por ende tienen una distribución normal (para muestras grandes). En general el estimador por máxima verosimilitud para un parámetro $\theta \sim Normal(\hat{\theta}, \widehat{SE_{\hat{\theta}}})$ donde $\widehat{SE_{\hat{\theta}}}$ es el error estándar se denota como $\hat{\theta}$ [20].

1.5.1 Intervalo de confianza para β_0 y β_1

Un intervalo de confianza para β_0 es un intervalo de confianza para el cambio en el logaritmo de las proporciones.

El intervalo para β_0 es obtenido como:

$$\hat{\beta}_0 \pm z_{1-\alpha/2}\widehat{SE_{\hat{\beta}_0}}$$

Un intervalo de confianza para β_1 es un intervalo de confianza para el cambio en el logaritmo de las proporciones.

El intervalo para β_1 es obtenido como:

$$\hat{\beta}_1 \pm z_{1-\alpha/2}\widehat{SE_{\hat{\beta}_1}}$$

Donde $z_{1-\alpha/2}$ es el quantíl $1-\alpha/2$ de una normal estándar y $\widehat{SE_{\hat{\beta}_1}}$ es el estimador del error estándar del parámetro $\hat{\beta}_1$. En este caso se utiliza z en lugar de t que es usada para intervalos de confianza en una regresión lineal. Lo cual se debe a que no hay normalidad en la regresión logística.

Un intervalo para β_1 debe ser usado cuidadosamente, especialmente si el tamaño de la muestra no es grande.

Ejemplo 1.4

Con los datos del ejemplo 1.1, se utiliza el software SPSS para obtener los parámetros estimados del modelo, así mismo se obtienen los errores estándar de cada uno, y con ellos se pueden construir los intervalos de confianza. Además, se obtienen las estadísticas de Wald, como se muestra en la siguiente tabla

Variables en la Equación	B	SE	Wald	Grados	Sig.
Edad	-.249	.098	6.438	1	.011
Constant	4.633	1.546	8.983	1	.003

Tabla 1.4 Estimación de los coeficientes de la regresión logística, junto con S.E. y la estadística de Wald

Con el 95% de confianza, los intervalos de confianza para los parámetros quedan de la siguiente forma

$$\hat{\beta}_0 \pm z_{.975}\widehat{SE_{\hat{\beta}_0}} = 4.633 \pm (1.96\times 1.546) = 4.633 \pm 3.03016$$

$$\hat{\beta}_1 \pm z_{.975}\widehat{SE_{\hat{\beta}_1}} = -0.249 \pm (1.96\times 0.098) = -0.249 \pm 0.19208$$

Por lo que el intervalo de confianza para β_0 es el siguiente $(1.60284, 7.66316)$, y para β_1 $(-0.44108, -0.05692)$. Los intervalos de confianza de los dos coeficientes

no contienen al cero, aspecto que confirma la significancia para los dos coeficientes con la prueba de Wald.

1.6 Momios

La proporción de que un evento suceda esta descrita como:

$$\frac{\pi(x)}{1-\pi(x)}$$

Al contar con un modelo que tenga un buen ajuste, la interpretación de los coeficientes es lo que permite la comprensión de la presencia de cada una de las variables.

Hoy en día las estimaciones existentes más comunes y útiles para este tipo de regresión son los cocientes de *Momios*, que son valores mayores a cero y sin cotas superiores, vistos como medidas de disparidad sin importar el tipo de estudio que sea o el método de selección de variable. Los momios tienen una interpretación directa sobre la contribución de cada variable al modelo, aspecto que no se presenta en todos los métodos para identificación de los niveles de riesgo.

Así mismo el momio es una medida que intuitivamente ha sido ubicada para referir expectativas en apuestas. Un ejemplo de esto se puede ver en las carreras de caballos, donde el caballo A tiene probabilidad del 0.6 de ganar la carrera, por lo tanto, tiene 0.4 de probabilidad de no ganarla. En este caso, los momios de que el caballo gane la carrera son:

$$\frac{P[Probabilidad\ de\ ganar]}{1-P[Probabilidad\ de\ ganar]} = \frac{0.6}{0.4} = 1.5$$

Dichos cocientes cuentan el número de veces que será más probable que ocurra el éxito de un determinado evento.

El concepto de momio se puede aplicar a la regresión logística reemplazando la probabilidad de ocurrencia del evento a estudiar por $\pi(x)$ cuando $X = x$, obteniendo así una medida para los momios de presentar una respuesta positiva para un individuo con una especificación particular de x.

$$Momio = \frac{\pi(x)}{1 - \pi(x)} = \frac{\frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}}{1 - \frac{exp(\beta_0 + \beta_1 x)}{1 + exp(\beta_0 + \beta_1 x)}} = exp(\beta_0 + \beta_1 x)$$

Con lo anterior se puede ver que:

$$\pi(x) = \frac{Momio}{1 + Momio}$$

Capítulo 2

2. Regresión logística múltiple

2.1 Introducción

En el capítulo anterior se explicó el modelo de regresión logística para el caso de una sola variable independiente. Así como en el caso de la regresión lineal, la mayor ventaja de este modelo es que incluye varias variables, en donde algunas pueden estar en diferentes escalas.

En este capítulo se generaliza el modelo de regresión logística al caso de más de una variable independiente, en donde, una consideración importante será estimar los coeficientes del modelo y probar su significancia, así como también considerar el uso de variables independientes continuas.

El reto que enfrenta el modelo de regresión logística múltiple es explicar la presencia o ausencia de un evento cuando una sola variable dependiente no es suficiente para explicar dicho evento, por tal razón, se busca el uso del modelo de regresión logística múltiple, donde se relaciona la variable dependiente con K variables independientes.

2.2 Ajuste del modelo de regresión logística múltiple

Se supone que se tienen n observaciones, en donde y_i es la i-ésima respuesta observada, x_{ij} la i-ésima observación de la $j-$ésima variable, y siguiendo el esquema de la regresión logística con una sola variable independiente, se tiene para cada observación i con

$$i = 1,2,\dots,n \quad ; \quad j = 1,\dots,K$$

las siguientes ecuaciones:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_K X_{1K} \\ Y_2 &= \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_K X_{2K} \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_K X_{nK} \end{aligned}$$

Estas ecuaciones son el resultado de aplicar la función logit a $\frac{\pi(\underline{x_i})}{1-\pi(\underline{x_i})}$,

esto es:

$$y_i = ln\left(\frac{\pi\left(\underline{x_i}\right)}{1-\pi\left(\underline{x_i}\right)}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK}$$

donde

$$\pi\left(\underline{x_i}\right) = \frac{exp\left(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}\right)}{1 + exp\left(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}\right)}$$

y

$$\underline{x'_i} = (x_{i1}, x_{i2}, \ldots, x_{iK})$$

Para expresar en forma matricial, al modelo se define:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1K} \\ 1 & X_{21} & X_{22} & \cdots & X_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nK} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$$

Por lo tanto el modelo de regresión lineal múltiple expresado en notación matricial queda denotado de la siguiente manera:

$$\begin{bmatrix} Y_1 \\ Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1K} \\ 1 & X_{21} & X_{22} & \cdots & X_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nK} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$$

o

$$\boldsymbol{Y} = \boldsymbol{X\beta}$$

Con esta notación, el modelo de regresión logística múltiple puede escribirse como:

$$ln\left(\frac{\pi\left(\underline{x_i}\right)}{1-\pi\left(\underline{x_i}\right)}\right) = \boldsymbol{X\beta}$$

entonces la función de densidad conjunta de esta muestra está dada por:

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\underline{x_i})^{y_i} \left(1-\pi(\underline{x_i})\right)^{1-y_i} \qquad (2.1)$$

Se aplica el logaritmo natural en la ecuación (2.1) y da como resultado la función de log-verosimilitud siguiente:

$$\mathcal{L}(\boldsymbol{\beta}) = ln\big(\ell(\boldsymbol{\beta})\big) = \sum_{i=1}^{n} y_i \, ln\left(\pi\left(\underline{x_i}\right)\right) + \sum_{i=1}^{n} (1-y_i) ln(1-\pi(\underline{x_i}))$$

$$= \sum_{i=1}^{n} y_i \, ln\left(\pi\left(\underline{x_i}\right)\right) - \sum_{i=1}^{n} y_i \, ln\left(1-\pi\left(\underline{x_i}\right)\right) + \sum_{i=1}^{n} ln(1-\pi(\underline{x_i}))$$

$$= \sum_{i=1}^{n} y_i \, ln\left(\frac{\pi\left(\underline{x_i}\right)}{1-\pi\left(\underline{x_i}\right)}\right) + \sum_{i=1}^{n} ln(1-\pi(\underline{x_i}))$$

$$= \sum_{i=1}^{n} y_i \left(\beta_0 + \sum\nolimits_{j=1}^{K} \beta_j x_{ij}\right) - \sum_{i=1}^{n} ln\left(1 + exp\left(\beta_0 + \sum\nolimits_{j=1}^{K} \beta_j x_{ij}\right)\right)$$

la derivada con respecto a β_0 es:

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n}\left[\frac{exp\left(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}\right)}{1 + exp\left(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}\right)}\right] =$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi(\underline{x_i})$$

la derivada con respecto a β_l

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_l} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} x_{ij}\left[\frac{exp\left(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}\right)}{1 + exp\left(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}\right)}\right]$$

$$= \sum_{i=1}^{n} x_{il}\left(y_i - \pi\left(\underline{x_i}\right)\right) \qquad (2.2)$$

Para encontrar el estimador máximo verosímil de $\boldsymbol{\beta}$ se deben resolver las $K+1$ derivadas de (2.2) igualándolas a cero.

Para obtener cada una de las soluciones, si es que existen, se tiene que encontrar el punto crítico sea este un máximo o un mínimo. El punto crítico va a ser máximo si la matriz de las segundas derivadas parciales es definida negativa. Una propiedad importante de la matriz de segundas derivadas parciales, es que esta forma la matriz de varianzas-covarianzas de los parámetros estimados. La forma general de los elementos de la matriz de segundas derivadas es:

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_l \partial \beta_{l'}} = \frac{\partial}{\partial \beta_{l'}} \sum_{i=1}^{n} x_{il} \left(y_i - \pi \left(\underline{x_i} \right) \right)$$

$$= -\frac{\partial}{\partial \beta_{l'}} \sum_{i=1}^{n} x_{il} \pi \left(\underline{x_i} \right)$$

$$= -\sum_{i=1}^{n} x_{il} \frac{\partial}{\partial \beta_{l'}} \left[\frac{exp\left(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}\right)}{1 + exp\left(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}\right)} \right] \tag{2.3}$$

2.2.1 El método de Newton-Raphson

Teniendo en cuenta el hecho de tener que resolver un sistema de $K + 1$ ecuaciones no lineales con $K + 1$ variables desconocidas, la solución a este sistema de ecuaciones es un vector con elementos de la forma β_l, después de verificar que la matriz de las segundas derivadas parciales es definida negativa, y la solución es un máximo global en vez de un máximo local, es entonces cuando se puede concluir que este vector contiene las estimaciones de los parámetros donde los datos observados tendrían la más alta probabilidad de ocurrencia. De cualquier forma, resolver un sistema de ecuaciones no lineal no es una tarea sencilla, la solución podría ser que no se derive fácilmente como en el caso de las ecuaciones lineales, y deba ser estimada numéricamente usando un proceso iterativo. Quizá el método más

popular para el sistema de ecuaciones no lineales es el método de Newton Raphson [7].

El método de Newton Raphson comienza con una aproximación inicial para la solución, posteriormente utiliza los dos primeros términos del polinomio de Taylor evaluado en la estimación inicial para llegar a otra estimación que está más cerca de la solución. Este proceso continuará hasta que converja, (en el mejor de los casos) a una solución. Tomando en cuenta que el polinomio de Taylor de grado n para una función f en el punto $x = x_0$ se define como los primeros n términos de Taylor de la serie de f:

$$\sum_{i=0}^{n} \frac{f^{(i)}(x_0)}{i!}(x - x_0)^i \tag{2.4}$$

siempre y cuando las primeras n derivadas de f existan en x_0. El polinomio de Taylor de primer grado es también la ecuación de la recta tangente de f en el punto $\big(x_0, f(x_0)\big)$. El punto en donde la línea tangente cruce el eje x, $(x_1, 0)$, es usada para la siguiente iteración en la aproximación de la raíz para encontrar donde $f(x) = 0$.

El primer paso en el método de Newton Raphson es tomar el primer grado del polinomio de Taylor como una aproximación para f, en donde se iguala a cero:

$$f(x_0) + f'(x_0)(x - x_0)^1 = 0$$

Se resuelve para x, se tiene que:

$$x = x_0 - \frac{f(x_0)}{f'(x_0)} \tag{2.5}$$

Este nuevo valor de x es la nueva aproximación para la raíz, de tal forma que se deja $x_1 = x$, y se continua con el mismo proceso para obtener $x_2, x_3, ...$, hasta que las aproximaciones converjan.

Generalizar el método de Newton Raphson a un sistema de ecuaciones no es difícil. En este caso, las ecuaciones cuyas raíces se quiere obtener son las presentadas en la ecuación (2.2), es decir, la primera derivada de la función de log-verosimilitud. La igualdad en (2.2) es en realidad un sistema de $K + 1$ ecuaciones cuyas raíces se quieren encontrar al mismo tiempo, de manera que, es más conveniente utilizar la notación matricial para expresar cada paso del método de Newton Raphson. La ecuación (2.2) puede ser escrita como $\mathcal{L}'(\boldsymbol{\beta})$. $\boldsymbol{\beta}^{(0)}$ es el vector inicial con aporximaciónes de cada β_l, así el primer paso del método de Newton Raphson se puede expresar como:

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + \left[-\mathcal{L}''\left(\boldsymbol{\beta}^{(0)}\right)\right]^{-1}\mathcal{L}'^{(\boldsymbol{\beta}^{(0)})} \tag{2.6}$$

Sea $\boldsymbol{\mu}$ el vector columna de longitud N con elementos $\mu_i = \pi_i$. Se puede notar que cada elemento de $\boldsymbol{\mu}$ puede ser escrito también como $\mu_i = E(Y_i)$, el valor esperado de y_i. Usando la multiplicación de matrices, se puede mostrar que:

$$\mathcal{L}'(\boldsymbol{\beta}) = \boldsymbol{X^T}(\boldsymbol{Y} - \boldsymbol{\mu})$$

quien es un vector columna de orden $K + 1$ cuyos elementos son $\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_l}$, como se deriva en (2.2). Sea $\boldsymbol{V}$ una matriz cuadrada de orden N, con elementos de la forma $\pi_i(1 - \pi_i)$ en la diagonal y ceros en las demás entradas. Usando la multiplicación de matrices se puede verificar que:

$$l''(\boldsymbol{\beta}) = -\boldsymbol{X^T V X} \tag{2.7}$$

como resultado queda una matriz de $(K+1)\times(K+1)$ con elementos de la forma $\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_l \beta_{l'}}$, en donde la igualdad en (2.6) puede ser reescrita como:

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + [-\boldsymbol{X^T V X}]^{-1}\boldsymbol{X^T}(\boldsymbol{Y} - \boldsymbol{\mu})$$

Se debe continuar aplicando la última igualdad hasta que no haya ningún cambio significativo en los valores de $\boldsymbol{\beta}$ de una iteración a otra. En la ecuación (2.7) se tiene la información de varianzas y covarianzas de la matriz de los estimadores.

Los estimadores de las varianzas y covarianzas se obtienen de la matriz de segundas derivadas parciales de la función de log verosimilitud. Estas derivadas parciales tienen la forma de (2.3)

$$\frac{\partial^2 \mathcal{L}(\beta)}{\partial {\beta_l}^2} =$$

$$= -\sum_{i=1}^{n} x_{il}\left[\frac{x_{il}exp\left(\beta_0 + \sum_{j=1}^{k}\beta_j x_{ij}\right)\left(1 + exp\left(\beta_0 + \sum_{j=1}^{K}\beta_j x_{ij}\right)\right)}{\left(1 + exp\left(\beta_1 + \sum_{j=2}^{K}\beta_j x_{ij}\right)\right)^2} - \frac{-x_{il}\left(exp\left(\beta_0 + \sum_{j=1}^{K}\beta_j x_{ij}\right)\right)^2}{\left(1 + exp\left(\beta_1 + \sum_{j=2}^{K}\beta_j x_{ij}\right)\right)^2}\right]$$

$$= -\sum_{i=1}^{n} {x_{il}}^2\left[\frac{exp\left(\beta_0 + \sum_{j=1}^{K}\beta_j x_{ij}\right)}{\left(1 + exp\left(\beta_0 + \sum_{j=1}^{K}\beta_j x_{ij}\right)\right)^2}\right]$$

$$= -\sum_{i=1}^{n} {x_{il}}^2 \pi(x_i)\big(1 - \pi(x_i)\big) \tag{2.8}$$

y

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_l \beta_{l'}} = -\sum_{i=1}^{n} x_{il} x_{il'} \left[\frac{exp\big(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}\big)}{\Big(1 + exp\big(\beta_0 + \sum_{j=1}^{K} \beta_j x_{ij}\big)\Big)^2} \right]$$

$$= -\sum_{i=1}^{n} x_{il} x_{il'} \pi\left(\underline{x_i}\right)\left(1 - \pi\left(\underline{x_i}\right)\right) \tag{2.9}$$

$$l, l' = 1, \dots, K$$

Se denota como $I(\boldsymbol{\beta})$ a la matriz de $(K+1)\times(K+1)$ que contiene los negativos de los términos dados en las igualdades de (2.8) y (2.9). Esta matriz es conocida como la matriz de información observada[7].

Para obtener las varianzas y covarianzas de los coeficientes estimados se obtiene la inversa de la matriz de información la cual queda denotada como $\Sigma\big(\widehat{\boldsymbol{\beta}}\big) = [l''(\boldsymbol{\beta})]^{-1}$, teniendo como resultado lo siguiente:

$$\widehat{\boldsymbol{Var}(\widehat{\boldsymbol{\beta}})} = Diag([l''(\boldsymbol{\beta})^{-1}])$$

Ejemplo 2.1

Para esta sección del modelo logístico múltiple se toman en cuenta las variables de TDAHMIXTDXS, *Edad*, y *Talla*, donde TDAHMIXTDXS es la variable dependiente y Edad y Talla son las variables independientes, recalcando que se

tomarán en cuenta los pacientes que hayan proporcionado la información de Talla, que sean menores de 21 años, y hayan sido diagnosticados con algún tipo de TDAH. La variable TDAHMIXTDXS muestra si algún paciente fue diagnosticado con TDAH tipo Mixto o no (si presenta TDAH tipo Mixto la variable TDAHMIXTDXS será igual a 1 y 0 para cualquier otro caso).

Variables	B	Exp(B)
Edad (x_1)	-.389	.678
Talla (x_2)	5.055	156.854
Constante	-1.262	.283

Tabla 2.1 Coeficientes Estimados de la Regresión Logística Múltiple.

Las estimaciones por máxima verosimilitud de los coeficientes β_0, β_1 y β_2, están dadas por $\hat{\beta}_0 = 1.262$, $\hat{\beta}_1 = -0.389$ y $\hat{\beta}_2 = -5.055$. La ecuación para el valor estimado de la probabilidad condicional donde $Y = 1$ dado que $X = x_i$ para el paciente i, es:

$$\hat{\pi}\left(\underline{x_i}\right) = \frac{exp(-1.262 + (-0.389 \times Edad_i) + (5.055 \times Talla_i))}{1 + exp(-1.262 + (-0.389 \times Edad_i) + (5.055 \times Talla_i))}$$

2.3 Intervalos de confianza

2.3.1 Intervalo de confianza para $\boldsymbol{\beta_i}$

Así como en el modelo simple, el intervalo de confianza para un coeficiente está centrado en el estimador del mismo más menos el cuantil $\left(1 - {}^{\alpha}/_{2}\right)$ de una distribución normal estándar por la estimación del error estándar sobre este parámetro, de tal forma el intervalo de confianza de $(1-\alpha)\times 100\%$ para el parámetro β_i con $i = 1,2,\ldots,K$ será:

$$\hat{\beta}_i \pm z_{(1-{}^{\alpha}/_{2})} \widehat{SE}_{\hat{\beta}_i}$$

2.3.2 Intervalos de confianza para una *Y* fija

El cálculo de los intervalos de confianza para una *Y* fija cuando se tienen múltiples variables independientes es la misma que cuando se tiene una sola variable independiente. Este intervalo es calculado primero para $ln\big(\pi(x)\big)$ y después se transformará para $\pi(x)$.

Así el intervalo de confianza para $ln\big(\pi(x)\big)$ queda de la siguiente forma:

$$ln\big(\widehat{\pi(x_i)}\big) \pm z_{1-\alpha/2}\widehat{SE}_{\hat{\beta}_i}$$

la dificultad con el cálculo de este intervalo está en el determinar$\widehat{SE}_{\hat{\pi}(x_i)}$, cuando los intervalos de confianza no pueden ser obtenidos directamente a través de un software. Para el cálculo de $\widehat{SE}_{\hat{\pi}(x_i)}$ en este trabajo, solo se revisará un método.

Si se puede obtener la matriz de varianzas y covarianzas para cualquier pareja de $\hat{\beta}_j, \hat{\beta}_l$, se obtiene $\widehat{SE}_{\hat{\pi}(x_i)}$ de la siguiente manera:

$$\widehat{SE}_{\hat{\pi}(x_i)} = \sqrt{\sum_{j=1}^{K} x_j^2 \widehat{Var(\hat{\beta}_J)} + 2\sum_{j=1}^{K}\sum_{l=j+1}^{K} x_j x_l \widehat{Cov\big(\hat{\beta}_j, \hat{\beta}_l\big)}}$$

2.4 Pruebas de hipótesis

En el modelo de regresión logística simple mencionados anteriormente, se revisaron algunas pruebas de significancia, sin embargo, se tiene el caso en donde se agregan más variables al modelo de las cuales algunas varían y otras cambian, por lo tanto, se van a presentar las pruebas de significancia más conocidas.

2.4.1 Devianza

La hipótesis que se desea probar será:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$$
$$vs$$
$$H_a: \beta_j \neq 0 \qquad para\ al\ menos\ algún\ j$$

Como se señaló en el capítulo anterior, derivado de la función de verosimilitud, se define la estadística *D* como:

$$D = -2ln\left(\frac{versomilitud\ del\ modelo\ ajustado}{versomilitud\ del\ modelo\ saturado}\right)$$

que evalúa qué tan drásticos son los cambios que se producen al agregar o quitar variables independientes al modelo.

Se puede destacar que la característica de la estadística *D* en relación al cociente de verosimilitudes es que dicho, cociente al tomar un valor cercano a uno la devianza tendrá un valor muy cercano a cero.

El modelo saturado para el modelo de regresión logística múltiple, al igual que en el modelo logístico simple, los valores observados son iguales a los valores estimados.

Por otro lado, en el modelo logístico ajustado o aquel en el que están consideradas cierto número de variables, con parámetros estimados $\widehat{\boldsymbol{\beta}}$, al aplicarle la función de log-verosimilitud, se aprecia que el valor máximo lo alcanza en:

$$ln\left(L(\widehat{\boldsymbol{\beta}})\right) = \sum_{i=1}^{n}\left(y_i \boldsymbol{x'}_i \widehat{\boldsymbol{\beta}} - ln\left(1 - e^{\boldsymbol{x'}_i \widehat{\boldsymbol{\beta}}}\right)\right)$$

La devianza del modelo que establece la comparación entre los logaritmos de verosimilitudes de los modelos saturado y ajustados queda definida como:

$$D = -2\sum_{i=1}^{n}\left(y_i\, ln\left(\frac{\hat{\pi}\left(\underline{x_i}\right)}{y_i}\right) + (1 - y_i) ln\left(\frac{(1 - \hat{\pi}\left(\underline{x_i}\right))}{1 - y_i}\right)\right)$$

$$= -2ln\left(\frac{verosimilitud\ del\ modelo\ ajusatdo}{verosimilitud\ del\ modelo\ saturado}\right) \qquad (2.10)$$

Esta estadística deberá ser usada para determinar si una variable debe ser agregada a un modelo, y no como una medida absoluta de la calidad de bondad de ajuste.

La prueba de razón de verosimilitudes para el modelo múltiple es análoga al caso univariado, sin embargo la versión múltiple de la prueba tiene la finalidad de verificar la significancia de al menos una de las K variables independientes $(X_1, X_2, \dots, X_K)$, de manera simultánea. Debido a que la función de verosimilitud de un modelo saturado siempre es igual a 1, el cálculo de la estadística G se basará en la misma expresión que el caso univariado, sin embargo, se comparará el modelo con K variables independientes y un modelo con ninguna variable independiente y un parámetro β_0. La verosimilitud del modelo sin variables explicativas se conserva y G resulta de la siguiente manera:

$$G = -2ln\left[\frac{\left(\frac{n_1}{n}\right)^{n_1}\left(\frac{n_0}{n}\right)^{n_0}}{\prod_{i=1}^{n}\hat{\pi}\left(\underline{x_i}\right)^{y_i}\left(1 - \hat{\pi}\left(\underline{x_i}\right)\right)^{1-y_i}}\right]$$

$$= 2\left(\sum_{i=1}^{n}\left[y_i \ln\left(\hat{\pi}\left(\underline{x_i}\right)\right) + (1-y_i) ln\left(1-\hat{\pi}\left(\underline{x_i}\right)\right)\right]\right.$$
$$\left. - \left[n_1 ln(n_1) + n_0 ln(n_0) - n\, ln(n)\right]\right) \qquad (2.11)$$

Bajo el supuesto de que todos los parámetros son igual a cero, la estadística G se distribuye ji-cuadrada con K grados de libertad $G \sim {\chi^2}_{(K)}$, es decir, para un nivel de significancia α se rechazará la hipótesis nula si $G > {\chi^2}_{(K)}^{(1-\alpha)}$,en donde ${\chi^2}_{(K)}^{(1-\alpha)}$ es el cuantil de $1-\alpha$ de una ji-cuadrada con K grados de libertad.

Si la condición se cumple, se dice que al menos una variable es significativamente distinta de cero [11].

Usando esta misma notación, el *p-value* asociado a esta prueba es

$$P\left[{\chi^2}_{(k)} > G\right] = p - value$$

Ejemplo 2.2

Para los datos del ejemplo 2.1 se tiene que $n_1 = 51$ y $n_0 = 21$. Al obtener los parámetros estimados del modelo dentro del software de SPSS también se obtiene el resumen del modelo donde $-2\log(verosimilitud\ incluyendo\ las\ variables)$ tiene un valor de 80.321. Evaluando G en la ecuación en (2.11) se tiene:

$$G = -2\left(40.16 + \left(51ln(51) + 21ln(21) - 72ln(72)\right)\right) = 6.6022$$

Por un lado:

$$\chi^2{}_{(2)}^{(.95)} = 5.99 < 6.6022 = G$$

por otro

$$P[G > 6.6022] = 0.0368 = p - value$$

Con lo anterior se evidencia que al menos una variable independiente es significativa para el modelo donde la variable dependiente es TDAH mixto a un nivel significancia del 5%.

2.4.2 Prueba de Wald multivariada

La prueba de Wald en su forma univariada es útil en el contexto de un modelo múltiple para comprobar la significancia de variables individualmente.

De tal forma que, para probar la significancia de un coeficiente de regresión individual, β_j con $j = 1, \dots, K$, es decir:

$$H_0: \beta_j = 0 \quad vs \quad H_a: \beta_j \neq 0$$

La estadística de prueba es:

$$W_j = \frac{\hat{\beta}_j}{\left(\widehat{S_{\hat{\beta}_j}}\right)^{\frac{1}{2}}}$$

Bajo la hipótesis nula esta estadística se distribuye normal estándar, por lo que se rechaza H_0 si $W_j > z_{\left(1-\frac{\alpha}{2}\right)}$ o $W_j < z_{\left(\frac{\alpha}{2}\right)}$, donde $z_{(\alpha)}$ es el α cuantil de una distribución normal estándar.

W_j denota la estadística de Wald en su versión univariada, prueba que puede ser adoptada en los modelos logísticos múltiples cuando se desea evaluar la significancia individual de una cierta variable j.

Existe una prueba análoga a la estadística de Wald en caso en que se requiera probar la significancia de las variables explicativas, es decir:

$$H_0: \beta_1 = \cdots = \beta_K \quad vs \quad H_a: \beta_j \neq 0 \quad para\ algún\ j = 1, \dots, K$$

El análogo de la estadística de Wald para el caso múltiple se puede calcular de la siguiente forma:

$$\boldsymbol{W} = \widehat{\boldsymbol{\beta}}^T \left[\widehat{Var(\widehat{\boldsymbol{\beta}})}\right]^{-1} \widehat{\boldsymbol{\beta}}$$

$$= \widehat{\boldsymbol{\beta}}^T \left(\boldsymbol{X}^T \widehat{\boldsymbol{V}} \boldsymbol{X}\right) \widehat{\boldsymbol{\beta}}$$

Bajo la hipótesis nula la estadística $\boldsymbol{W}$ en su versión múltiple se distribuye aproximadamente ${\chi^2}_{(K+1)}$.

La prueba del cociente de verosimilitudes encuentra un equivalente en la prueba de Wald múltiple si la matriz $\boldsymbol{X}$ y el vector de parámetros $\widehat{\boldsymbol{\beta}}$ se modifican para contener únicamente la información de los K parámetros restantes en el modelo, es decir, al eliminar $\hat{\beta}_0$ de $\widehat{\boldsymbol{\beta}}$ y la primera fila y columna de la matriz $\boldsymbol{X}^T \widehat{\boldsymbol{V}} \boldsymbol{X}$ relacionada con la información en $\boldsymbol{X}$ de $\hat{\beta}_0$. En este caso la estadística $\boldsymbol{W}$ se distribuye ji-cuadrada con K grados de libertad.

De esta forma, para probar la significancia del modelo se utiliza la siguiente comparación:

$$W > \chi^{2}{}_{(K)}^{(1-\alpha)}$$

esto significa que si se cumple, se rechazará la hipótesis nula si se cumple lo anterior.

2.4.3 Estadística de Hosmer-Lemeshow

Hosmer y Lemeshow propusieron una estadística de bondad de ajuste para el modelo de regresión logística que se basa en la agrupación de los sujetos por probabilidades estimadas $\hat{\pi}\left(\underline{x_i}\right)$.

La estadística de Hosmer-Lemeshow ayuda a comparar y evaluar los modelos ajustados con los datos reales. Es sencillo pensar que se pueden ir revisando pareja por pareja, sin embargo, al tener una muestra de tamaño significativo, dicha idea resulta un tanto complicada.

De esta forma, partiendo de ideas sencillas, se sabe que la prueba de Hosmer-Lemeshow es una técnica para evaluar los modelos de regresión. La estadística de Hosmer-Lemeshow, se diseñó para evitar el uso de la estadística devianza en casos donde su aproximación a una distribución conocida (ji-cuadrada) resulta cuestionable.

La idea general es agrupar las probabilidades estimadas por el modelo logístico, $\left(\hat{\pi}\left(\underline{x_1}\right), \hat{\pi}\left(\underline{x_2}\right), \ldots, \hat{\pi}\left(\underline{x_n}\right)\right)$ según su valor. Luego se forman G grupos, cada uno con n/G observaciones. Para cada grupo, se calculan las sumas de las frecuencias observadas y estimadas de cada respuesta de la variable dependiente. El número de respuestas positivas y negativas esperadas, en el grupo percentil g son:

$$E_{1g} = \sum_{i=1}^{n_g} \hat{\pi}\left(\underline{x_{ig}}\right)$$

y

$$E_{0g} = \sum_{i=1}^{n_g} \left(1 - \hat{\pi}\left(\underline{x_{ig}}\right)\right)$$

El número de respuestas positivas (categoría 1) y negativas (categoría 0) observadas, en el grupo percentil g son:

$$O_{1g} = \sum_{i=1}^{n_g} Y_i$$

y

$$O_{0g} = \sum_{i=1}^{n_g} (1 - Y_i)$$

donde, tanto para el valor esperado como para el valor observado, n_g es la cantidad de individuos en el grupo g y x_{ig} son los valores para las covariables del i −ésimo sujeto en el mismo grupo.

La estadística $\hat{C}$ se calcula comparando los valores observados y esperado por los grupos definidos de manera que:

$$\hat{C} = \sum_{g=1}^{G} \frac{\left(O_{1g} - E_{1g}\right)^2}{E_{1g}} + \sum_{g=1}^{G} \frac{\left(O_{0g} - E_{0g}\right)^2}{E_{0g}}$$

Es conveniente remarcar que una de las restricciones de este modelo es que los valores de E_{ig} deben ser diferentes de cero [14].

La estadística $\hat{C}$ se distribuye ji-cuadrada con $G-2$ grados de libertad $G \sim {\chi^2}_{(G-2)}$. Esto es, para un nivel de significancia α, se rechazará la hipótesis nula si $G > {\chi^2}_{(G-2)}^{(1-\alpha)}$. Donde ${\chi^2}_{(G-2)}^{(1-\alpha)}$ es el cuantil de $1-\alpha$ de una ji-cuadrada con $G-2$ grados de libertad.

Ejemplo 2.3

El software estadístico SPSS tiene un módulo en donde se puede hacer la prueba de bondad de Ajuste de Hosmer-Lemeshow. Normalmente este paquete utiliza grupos de 10 ($G = 10$) para esta prueba.

Con los datos del ejemplo 2.1, se aplica la prueba de Hosmer-Lemeshow con los siguientes resultados:

Tabla de Contingencia Prueba Hosmer-Lemeshow					
	TDAHMIXTDXS = 0 Ausencia		TDAHMIXTDXS = 1 Presencia		
Grupos	Observado	Esperado	Observado	Esperado	Total
1	5	4.501	2	2.499	7
2	1	2.947	6	4.053	7
3	3	2.38	4	4.62	7
4	2	2.01	5	4.99	7
5	2	1.874	5	5.126	7
6	0	1.673	7	5.327	7
7	2	1.727	6	6.273	8
8	3	1.395	4	5.605	7
9	2	1.325	5	5.675	7
10	1	1.168	7	6.832	8

Tabla 2.2 Tabla de Contingencia Prueba Hosmer Lemeshow

Prueba Hosme-Lemeshow		
Ji-Cuadrada	Grados de libertad	Sig.
7.644	8	.469

Tabla 2.3 Prueba Hosmer-Lemeshow

$$\chi^2{}_{(8)}^{(.95)} = 15.51$$

por lo tanto

$$\hat{C} = 7.644 < 15.51 = \chi^2{}_{(8)}^{(.95)}$$

Lo que hace esta prueba es, comprobar si el modelo propuesto puede explicar lo que se observa. Es una prueba donde se evalúa la distancia entre lo observado en los datos que se tienen de la realidad y lo esperado bajo el modelo.

Se puede observar que el $p - value$ es grande. Aquí la Hipótesis nula es que el modelo se ajusta a la realidad. En una prueba de bondad de ajuste siempre en la hipótesis nula se afirma que el modelo propuesto se ajusta a lo observado. Por lo tanto, un $p - value$ superior a 0.05 implica que lo que se observa se ajusta suficientemente a lo que esperado bajo el modelo.

2.5 Razón de Momios

Los momios de un determinado evento se definen como la razón de la probabilidad de que ocurra este evento entre la probabilidad de que el evento no ocurra, esta cantidad obtenida es que tanto es más probable que ocurra el evento a que no ocurra.

$$O\left(\underline{x_i}\right) = \frac{\pi\left(\underline{x_i}\right)}{1 - \pi\left(\underline{x_i}\right)} = exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_K x_{Ki})$$

La razón de momios, como su nombre lo dice, es la razón de diferentes momios. Esta medida está asociada a una tabla de contingencia de 2×2, donde se van a comparar las probabilidades de los dos diferentes valores que tiene la variable dependiente condicionado al incremento en una unidad de una variable independiente.

$$OR_k\left(\underline{x_i}\right) = \frac{O(x_1, x_2, \dots, x_k + 1, \dots, x_K)}{O(x_1, x_2, \dots, x_k, \dots, x_K)} = exp(\beta_k)$$

2.5.1 Intervalos de confianza para la razón de momios

En la sección 2.5, se establecio que la razón de momios en cambio de una unidad $OR_k\left(\underline{x_i}\right)$ esta dada por:

$$OR_k\left(\underline{x_i}\right) = exp(\beta_k)$$

Si se conoce $\widehat{SE_{\hat{\beta}_k}}$, es posible establecer el intervalo de confianza para la razón de momios.

Por lo tanto, el intervalo de confianza del $(1-\alpha)\times100\%$ de $OR_k\left(\underline{x_i}\right)$ es:

$$\left[exp\left(\hat{\beta}_k - z_{\alpha/2}\widehat{SE_{\hat{\beta}_k}}\right), exp\left(\hat{\beta}_k + z_{\alpha/2}\widehat{SE_{\hat{\beta}_k}}\right)\right]$$

donde $z_{\left(\frac{\alpha}{2}\right)}$ es el cuantil $\left(\frac{\alpha}{2}\right)$ de una distribución normal con media igual a 0 y desviación estándar igual a 1.

Capítulo 3

3. Regresión Logística Multinomial

3.1 Introducción

En los capítulos anteriores se presentaron modelos donde la variable dependiente es de forma binaria y supone un componente aleatorio binario. Para el caso en el que la variable de respuesta es multicategórica, se supone un componente aleatorio multinomial. En este capítulo se presenta una generalización de la regresión logística cuando la variable dependiente sea multinomial.

Se presenta un modelo para variables con respuesta nominal. Este usa ecuaciones binarias separadas para cada par de categorías de respuesta. Un tipo importante de análisis es el efecto de las variables independientes en la elección de una persona en un conjunto de opciones a elegir, ya sea el elegir una determinada marca para comprar o la elección de votar por un determinado partido político. Asimismo, se presentarán ejemplos prácticos.

En este capítulo se expresa el modelo en términos de datos desagrupados. Como en el caso binario, es mejor agrupar las N observaciones de acuerdo al tipo de categoría en el que caen, antes de calcular la devianza y otras pruebas estadísticas de bondad de ajusta y residuales.

Tradicionalmente, las variables dependientes multicategóricas han sido modeladas mediante análisis discriminante, pero gracias al creciente desarrollo de las técnicas de cálculo, cada vez es más habitual el uso de modelos de regresión ya implementados en paquetes estadísticos como SAS o SPSS, debido a la mejor interpretación de los resultados que proporciona.

Lo que se busca en este capítulo es presentar las bases teóricas de este modelo, ya sea para su formulación o para el ajuste del mismo.

3.2 Modelo Logístico Multinomial

En el Anexo se presenta la distribución multinomial y su deducción.

Suponiendo que se tiene una muestra aleatoria de tamaño n, se considera una variable aleatoria Y_i que puede tomar uno de los valores $1, 2, \dots, J$. Sea $\pi_{ij} = P(Y_i = j)$ que denota la probabilidad que la variable aleatoria Y_i tome el valor j.

Suponiendo que los diferentes tipos de respuesta de la variable aleatoria son mutuamente excluyentes, se tiene que $\sum_{j=1}^{J} \pi_{ij} = 1$. Al ser la escala de medida nominal; el orden entre las categorías es irrelevante.

Recodificando $Y_i = j$, sea Y_{ij} el evento que el individuo i responda a la categoría $j-$ésima, con un valor observado de y_{ij} donde y_{ij} solo puede tener los valores de 0 o 1 para cada individuo.

$$Y_{ij} = \begin{cases} 1 & si \quad Y_i = j \\ 0 & Cualquier\ otro\ caso. \end{cases}$$

Se puede notar que $\sum_j y_{ij} = 1$.

La distribución de los conteos de Y_{ij} queda denotado de la siguiente forma:

$$P\left(Y_{i1} = y_{i1}, \dots, Y_{iJ} = y_{iJ}\right) = {\pi_{i1}}^{y_{i1}} \dots {\pi_{ij}}^{y_{iJ}}$$

Con el modelo anterior se considerarán modelos de probabilidades para π_{ij}. En particular se desean modelos donde las probabilidades dependan de un vector $\underline{x_i}$ de

variables independientes asociadas al i −ésimo individuo o grupo. Se toma una categoría como la respuesta base, y se le denomina como categoría de referencia, para el cual existirán $J-1$ categorías restantes con las que se contrastará, por lo tanto, se define un modelo logístico con respecto a ella y se obtiene una función lineal [21].

Con ello se define la razón de las probabilidades de la siguiente forma:

$$\frac{P(Y_i = j)}{P(Y_i = J)} = \frac{\pi_{ij}}{\pi_{iJ}}$$

Suponiendo además que para y existen K variables independientes $x_1, \dots, x_K$ la probabilidad de que y_i tenga el valor de J es descrita como función de las $J-1$ variables independientes como:

$$ln\left(\frac{P(Y_i = j)}{P(Y_i = J)}\right) = ln\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \sum_{k=0}^{K} \beta_{kj} x_{ik} \qquad (3.1)$$

Donde para cada $j \neq J$ se tiene un conjunto de parámetros $\boldsymbol{\beta}_j = (\beta_{1j}, \beta_{2j}, \dots, \beta_{Kj})$ relacionados a la categoría j. El uso de la categoría J como referencia es arbitrario, pero cualquier otra categoría puede ser la de referencia.

Este modelo es análogo al modelo logístico binomial, con excepción de que la distribución de probabilidad de la variable dependiente es multinomial en lugar de binomial y se tienen $J-1$ ecuaciones en lugar de solo una. Las $J-1$ ecuaciones logísticas multinomiales contrastan cada una de las categorías contra la categoría J, mientras que en el modelo logístico binomial la ecuación es un contraste entre éxito contra fracaso.

Se puede observar que solo se necesitan $J-1$ ecuaciones para describir una variable con J posibles respuestas y con ello no hay diferencia alguna en qué categoría es la que se escoge como referencia, ya que siempre se puede convertir de una formulación a otra. Suponiendo que se tienen 3 categorías se contrasta la categoría a contra la categoría J y la categoría b contra la J. Con lo anterior solo haría falta contrastar la categoría a contra la b, pero esta puede ser fácil de obtener en término de las otras dos:

$$ln\left(\frac{\pi_{ia}}{\pi_{ib}}\right) = ln\left(\frac{\pi_{ia}}{\pi_{iJ}}\right) - ln\left(\frac{\pi_{ib}}{\pi_{iJ}}\right)$$

El modelo logístico multinomial se puede escribir en términos de π_{ij}, de acuerdo a la ecuación (3.1), se puede escribir π_{ij} como:

$$\pi_{ij} = \frac{exp\{\sum_{k=0}^{K}\beta_{kj}x_{ik}\}}{\sum_{j=1}^{J}exp\{\sum_{k=0}^{K}\beta_{kj}x_{ik}\}} \tag{3.2}$$

Para verificar el resultado anterior, solo se necesita aplicar la exponencial en la ecuación 3.1 para así obtener:

$$\pi_{ij} = \pi_{iJ}exp\left\{\sum_{k=0}^{K}\beta_{kj}x_{ik}\right\} \tag{3.3}$$

Como suposición adicional se tiene que:

$$\sum_{k=0}^{K}\beta_{kJ}x_{ik} = ln\left(\frac{\pi_{iJ}}{\pi_{iJ}}\right) = \ln(1) = 0$$

Además, se sabe que $\sum_{j=1}^{J}\pi_{ij} = 1$ de donde se obtiene:

$$\pi_{iJ} = \frac{1}{1 + \sum_{j=1}^{J-1} exp\{\sum_{k=0}^{K} \beta_{kJ} x_{ik}\}} \tag{3.4}$$

Finalmente se sustituye la ecuación (3.4) en la ecuación (3.3) para obtener la igualdad de (3.2) [7].

Un dato particular que se puede encontrar en la probabilidad de la $j-$ésima categoría es que esta puede ser escrita en términos de la razón de probabilidades.

Ejemplo 3.1

Las variables del estudio *Adversidad psicosocial, psicopatología y funcionamiento en hermanos adolescentes en alto riesgo (HAR) con y sin trastorno por déficit de atención con hiperactividad (TDAH)*[16], que se usan en capítulo son: $Edad$, TDAHINATDXS, y TDAHMIXTDXS. Las variables de TDAHINATDXS, TDAHHIPERIMPDXS, y TDAHMIXTDXS, son variables dicotómicas de presencia o ausencia de diferentes tipos de TDAH, con ellas se define una variable llamada Tipo_TDAH, en donde sus posibles respuestas son: Mixto, Inatento, o Ausente.

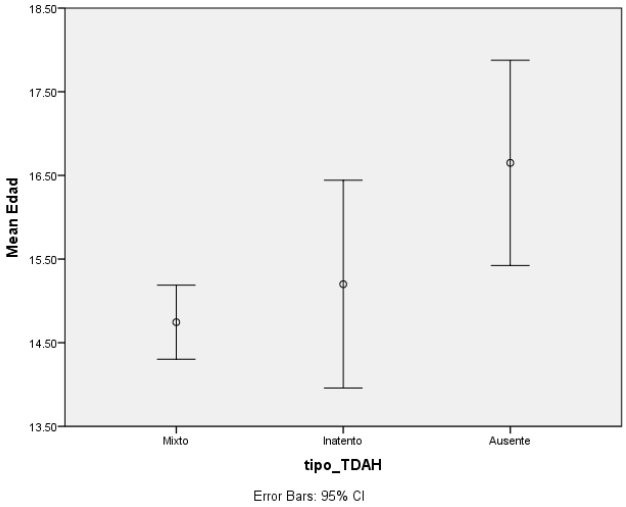

Figura 3.1 Diagrama de cajas Tipo_TDAH vs Edad

La figura 3.1 muestra donde se acumulan las medias de edad dependiendo del tipo de TDAH, lo que sugiere es que dependiendo de la edad hay una determinada probabilidad de que la variable Tipo_TDAH tome un determinado tipo de TDAH.

Al igual que en modelo binario, se desea encontrar una relación entre la variable dependiente con la variable independiente. En el caso binario a mayor edad cambia la proporción de pacientes que padecen algún tipo de TDAH. En la figura 3.1, dependiendo del tipo de TDAH cambia el promedio de edad en los pacientes.

3.3 Estimación de los Parámetros

Para cada población dentro de cada categoría, la variable dependiente tiene una distribución multinomial con J categorías, y por ello su función de densidad de probabilidad conjunta es:

$$f(y|\boldsymbol{\beta}) = \prod_{j=1}^{J} \pi_{ij}{}^{y_{ij}} \qquad (3.5)$$

Cuando $J = 2$ se reduce a la función de densidad de probabilidad conjunta de una binomial. La función de verosimilitud es algebraicamente igual a la ecuación 3.5, donde la única diferencia se encuentra en que la función de verosimilitud expresa los valores del parámetro desconocido $\boldsymbol{\beta}$ en términos de los valores de y. Dado que se desea maximizar la ecuación en 3.5 con respecto a $\boldsymbol{\beta}$, los términos factoriales que no contienen ninguno de los términos de π_{ij} se pueden tratar como constantes. De esta forma, la función de verosimilitud para una regresión logística multinomial es:

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{j=1}^{J} {\pi_{ij}}^{y_{ij}}$$

Reemplazando el término J −ésimo, la igualdad anterior queda como:

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{j=1}^{J-1} {\pi_{ij}}^{y_{ij}} {\pi_{iJ}}^{1-\sum_{j=1}^{J-1} y_{ij}}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{J-1} {\pi_{ij}}^{y_{ij}} \frac{{\pi_{iJ}}^{1}}{{\pi_{iJ}}^{\sum_{j=1}^{J-1} y_{ij}}}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{J-1} {\pi_{ij}}^{y_{ij}} \frac{\pi_{iJ}}{\prod_{j=1}^{J-1} {\pi_{iJ}}^{y_{ij}}}$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{J-1} \left(\frac{\pi_{ij}}{\pi_{iJ}}\right)^{y_{ij}} \pi_{iJ}$$

Como $ln\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \sum_{k=0}^{K} \beta_{kj} x_{ik}$ y el resultado de la ecuación (3.4), entonces se tiene que la ecuación anterior se puede escribir como:

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{j=1}^{J-1} exp\left(\sum_{k=0}^{K} \beta_{kj} x_{ik}\right)^{y_{ij}} \left(\frac{1}{1 + \sum_{j=1}^{J-1} exp\left\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\right\}}\right)$$

$$= \prod_{i=1}^{n} \prod_{j=1}^{J-1} \exp\left(y_{ij}\left(\sum_{k=0}^{K} \beta_{kj} x_{ik}\right)\right)\left(1 + \sum_{j=1}^{J-1} exp\left\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\right\}\right)^{-1}$$

Aplicando el logaritmo natural a la igualdad anterior, nos da la función de log-verosimilitud para el modelo de regresión logística multinomial:

$$\mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{J-1} \left(y_{ij}\left(\sum_{k=0}^{K} \beta_{kj} x_{ik}\right)\right) - ln\left(1 + \sum_{j=1}^{J-1} exp\left\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\right\}\right)$$

Como en el caso binomial, se quieren encontrar los valores para $\boldsymbol{\beta}$ donde se maximice la función anterior. Para ello se tiene que encontrar la primera y segunda derivada de la función de log-verosimilitud.

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_{kj}} = \sum_{i=1}^{n} y_{ij} x_{ik} -$$

$$\frac{1}{1 + \sum_{j=1}^{J-1} exp\left\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\right\}} \frac{\partial}{\partial \beta_{kj}} \left(1 + \sum_{j=1}^{J-1} exp\left\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\right\}\right)$$

$$= \sum_{i=1}^{n} y_{ij} x_{ik} - \frac{exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}}{1 + \sum_{j=1}^{J-1} exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}} \frac{\partial}{\partial \beta_{kj}} \left(\sum_{k=0}^{K} \beta_{kj} x_{ik} \right)$$

$$= \sum_{i=1}^{n} y_{ij} x_{ik} - \frac{exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}}{1 + \sum_{j=1}^{J-1} exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}} x_{ik}$$

$$= \sum_{i=1}^{n} y_{ij} x_{ik} - \pi_{ij} x_{ik} \tag{3.6}$$

El valor que maximiza a $\mathcal{L}(\boldsymbol{\beta})$ se obtiene a partir de la ecuación (3.6) igualándola a cero y resolviendo la ecuación para $\boldsymbol{\beta}$. Pero para esto, se necesita obtener la segunda derivada para cada β_{kj}.

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\beta})}{\partial \beta_{kj} \beta_{k'j'}} = \frac{\partial}{\partial \beta_{k'j'}} \sum_{i=1}^{n} y_{ij} x_{ik} - \pi_{ij} x_{ik}$$

$$= \frac{\partial}{\partial \beta_{k'j'}} \sum_{i=1}^{n} -\pi_{ij} x_{ik}$$

$$= \sum_{i=1}^{n} -x_{ik} \frac{\partial}{\partial \beta_{k'j'}} \frac{exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}}{\sum_{j=1}^{J-1} exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}}$$

De aquí la derivada depende completamente si $j' = j$, ya que la derivada del numerador cambia dependiendo del valor de j'.

$$\left(\frac{f}{g}\right)'(a) = \frac{g(a)f'(a) - f(a)g'(a)}{[g(a)]^2}$$

$$f'(a) = g'(a) = exp\left\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\right\} x_{ik\prime} \qquad j' = j$$

$$f'(a) = 0 \qquad g'(a) = exp\left\{\sum_{k=0}^{K} \beta_{kj\prime} x_{ik}\right\} x_{ik\prime} \qquad j' \neq j$$

Entonces, cuando $j' = j$, la segunda derivada parcial quedaría de la siguiente forma:

$$\frac{\left(1 + \sum_{j=1}^{J-1} exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}\right) exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\} x_{ik\prime} - \left(exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}\right)^2 x_{ik\prime}}{\left(\sum_{j=1}^{J-1} exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}\right)^2}$$

$$= \frac{exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\} x_{ik\prime} \left(1 + \sum_{j=1}^{J-1} exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\} - exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}\right)}{\left(1 + \sum_{j=1}^{J-1} exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}\right)^2}$$

$$= \pi_{ij} x_{ik\prime} (1 - \pi_{ij})$$

Ahora, para el caso de $j' \neq j$

$$\frac{0 - exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\} exp\{\sum_{k=0}^{K} \beta_{kj\prime} x_{ik}\} x_{ik\prime}}{\left(1 + \sum_{j=1}^{J-1} exp\{\sum_{k=0}^{K} \beta_{kj} x_{ik}\}\right)^2} = -\pi_{ij} x_{ik\prime} \pi_{ij\prime}$$

La forma general de los elementos de la matriz de segundas derivadas es el que nos ayudara a obtener el estimador de matriz de covarianzas del estimador de máxima verosimilitud. Donde estos elementos de la matriz de las segundas derivadas quedaran de la siguiente forma:

$$\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_{kj} \beta_{k'j'}} = -\sum_{i=1}^{n} x_{ik}\, \pi_{ij} x_{ik'} \left(1 - \pi_{ij}\right) \tag{3.7}$$

y

$$\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta_{kj} \beta_{k'j'}} = \sum_{i=1}^{n} x_{ik}\, \pi_{ij} x_{ik'} \pi_{ij'} \tag{3.8}$$

3.3.1 Método de Newton- Raphson

Para ilustrar el procedimiento iterativo de Newton-Raphson en el modelo de regresión logística multinomial, se toma la siguiente ecuación:

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + \left[-l''\left(\boldsymbol{\beta}^{(0)}\right)\right]^{-\mathbf{1}} l'^{\left(\boldsymbol{\beta}^{(0)}\right)} \tag{3.9}$$

Sea $\boldsymbol{\mu}$ una matriz con n renglones y $J-1$ columnas con elementos de la forma $\mu_{ij} = \pi_{ij}$. Se puede notar que cada elemento de $\boldsymbol{\mu}$ puede ser escrito también como $\mu_{ij} = E\left(y_{ij}\right)$, el valor esperado de y_{ij}. Usando la multiplicación de matrices, se puede mostrar que:

$$l'(\boldsymbol{\beta}) = \boldsymbol{X^T}(\boldsymbol{y} - \boldsymbol{\mu}) \tag{3.10}$$

La expresión en (3.10) es una matriz con $K+1$ renglones y $J-1$ cuyos elementos son $\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_{kj}}$, como se deriva de (3.6).

El caso de la segunda derivada, es diferente del caso binomial, dado que (3.7) y (3.8) depende de sí $j' = j$ o no.

Para los elementos de la diagonal de la matriz de segundas derivadas donde $j' = j$, se tiene $\boldsymbol{V}$ una matriz cuadrada de orden n con elementos de la forma $\pi_{ij}\left(1 - \pi_{ij}\right)$

en la diagonal y ceros en cualquier otro lado. La matriz de segundas derivadas queda como:

$$l''(\boldsymbol{\beta}) = -\boldsymbol{X^T V X} \tag{3.11}$$

(3.11) es una matriz de orden $(K+1)\times(K+1)$. Solo se puede utilizar la igualdad en (3.11) para los elementos de la diagonal. Para los elementos fuera de la diagonal en donde $j' \neq j$, se define $\boldsymbol{V}$ como una matriz con elementos sobre la diagonal $\pi_{ij}\pi_{ik}$.

Utilizando los supuestos de $\boldsymbol{V}$, en cada paso para el método de Newton Raphson se procede como en el caso del modelo de regresión logística binaria con la siguiente ecuación:

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + [-\boldsymbol{X^T V X}]^{-1}\boldsymbol{X^T}(\boldsymbol{y} - \boldsymbol{\mu})$$

Se tiene que continuar aplicando la última igualdad hasta que no haya ningún cambio significativo en los valores de $\boldsymbol{\beta}$ de una iteración a otra [7].

Ejemplo 3.2

El software estadístico SPSS tiene un módulo para aplicar la regresión logística multinomial. Se usa como variable dependiente Tipo_TDAH y como variable independiente $Edad$. Tomando como categoría de referencia la $Ausencia$, sus parámetros estimados quedan de la siguiente forma:

Estimación de Parámetros			
Tipo_TDAH	Parámetro	B	Exp(B)
1 Mixto	Constante	7.773	
	Edad	-.438	1.137
2 Inatento	Constante	4.631	
	Edad	-.309	1.549

Tabla 3.1 Estimación de Parámetros regresión logística multinomial Tipo_TDAH vs Edad

Las ecuaciones para el valor estimado de la probabilidad condicional donde la variable dependiente Y puede tener como respuesta Mixto o Inatento (1 o 2 respectivamente), dado que $X = x_i$ para el paciente i, son:

$$\hat{\pi}(x_{i1}) = \frac{exp(7.773 - 0.438 \times Edad_i)}{1 + exp(7.773 - 0.438 \times Edad_i) + exp(4.631 - 0.309 \times Edad_i)}$$

$$\hat{\pi}(x_{i2}) = \frac{exp(4.631 - 0.309 \times Edad_i)}{1 + exp(7.773 - 0.438 \times Edad_i) + exp(4.631 - 0.309 \times Edad_i)}$$

Y para calcular la categoría de ausencia:

$$\hat{\pi}(x_{i3}) = 1 - \widehat{\pi(x_{i1})} - \widehat{\pi(x_{i2})}$$

Al final para asignar una determinada categoría, se calculan las probabilidades de cada categoría y se escoge la que tenga mayor probabilidad. Si $Edad = 15$, entonces $\hat{\pi}(x_{i1}) = 0.69$, $\hat{\pi}(x_{i2}) = 0.18$, y $\hat{\pi}(x_{i3}) = 0.13$, como $\hat{\pi}(x_{i1})$ es la probabilidad más alta, la categoría que se le asigna a un paciente con 15 años de edad es la categoría 1.

3.4 Momios y razón de momios

Los momios y la razón de momios son de suma importancia para el modelo multinomial de igual forma como lo es en el modelo binario descrito en el capítulo anterior. En el modelo multinomial se obtienen los momios entre la categoría base (J) contra cualquiera (j) para cualquier i.

$$\frac{P(Y_i = j)}{P(Y_i = J)} = \frac{\hat{\pi}_{ij}}{\hat{\pi}_{iJ}} = \frac{\frac{exp\left(\sum_{k=0}^{K} \hat{\beta}_{kj} x_{ik}\right)}{1 + \sum_{j=1}^{J-1} exp\left(\sum_{k=0}^{K} \hat{\beta}_{kj} x_{ik}\right)}}{\frac{1}{1 + \sum_{j=1}^{J-1} exp\left(\sum_{k=0}^{K} \hat{\beta}_{kj} x_{ik}\right)}} =$$

$$= exp\left(\sum_{k=0}^{K} \hat{\beta}_{kj} x_{ik}\right) \qquad (3.12)$$

Aplicando el logaritmo natural en (3.12) se obtiene una función lineal como:

$$ln\left(\frac{P(Y_i = j)}{P(Y_i = J)}\right) = ln\left(\frac{\hat{\pi}_{ij}}{\hat{\pi}_{iJ}}\right) = \sum_{k=o}^{K} \hat{\beta}_{kj} x_{ik} \qquad (3.13)$$

El momio de la categoría j respecto a la categoría J se representa como $O_j(x_1, x_2, \dots, x_k, \dots, x_K) = O_j$. De este modo se puede observar que la razón de cambio de O_j cuando x_k tiene un incremento en una unidad manteniéndose constantes las demás variables independientes es:

$$OR_j(x_k) = \frac{O_j(x_1, x_2, \dots, x_k + 1, \dots, x_K)}{O_j(x_1, x_2, \dots, x_k, \dots, x_K)} = exp\left(\hat{\beta}_{kj}\right) \qquad (3.14)$$

La ecuación (3.14) recibe el nombre de *razón de momios* de la categoría j respecto a la variable independiente x_k y se representa como $OR_j(x_k)$.

Es interesante observar como la razón de momios depende de las unidades en que vengan medidas las variables independientes. Por lo tanto, la importancia de cada variable independiente en el modelo debería medirse por el valor en la *razón de momios* suponiendo estandarizada esta variable. Ese es el motivo por el que se necesita la *razón de momios estandarizada* en las variables independientes. Esto es

$$OR_j(x_k') = exp\left(\hat{\beta}_{kj} S_{x_k}\right)$$

Donde S_{x_k} es la desviación estándar de la variable x_k. Por lo tanto, mientras más grande sea la *razón de momios estandarizada*, más importante será la variable dentro del modelo [9].

Ejemplo 3.3

Tomando como referencia los datos del Ejemplo 3.2 el momio de la categoría 2 (Tipo TDAH Inatento), contra la categoría de referencia es:

$$\frac{P(Y_i = 2)}{P(Y_i = 3)} = \frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}} = \exp\left(4.631 + (-.309 \times \text{Edad}_i)\right)$$

Suponiendo que se tiene una edad de 12 años, el momio quedaría de la siguiente forma:

$$\frac{P(Y_i = 2)}{P(Y_i = 3)} = \frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}} = \exp\left(4.631 + (-.309 \times 12)\right) = 2.5$$

Lo cual nos dice que un paciente que tiene 12 años es 2.5 veces más probable que tenga TDAH tipo Inatento a que no presente algún tipo de TDAH.

Suponiendo ahora que un paciente tiene la edad de 18 años, el momio resultante sería:

$$\frac{P(Y_i = 2)}{P(Y_i = 3)} = \frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}} = \exp\left(4.631 + (-.309 \times 18)\right) = 0.3941$$

Lo cual nos dice que un paciente que tiene 18 años es 2.54 (esto es igual a $^{1}/_{0.3941}$ comparado a la categoría de referencia.) veces más probable que tenga no presente algún tipo de a que TDAH tipo Inatento.

3.5 Intervalos de confianza para los parámetros

Basándose en la normalidad asintótica de los estimadores por máxima verosimilitud se pueden construir los intervalos de confianza para cada uno de los parámetros del modelo y mediante las transformaciones correspondientes, así como los intervalos de confianza para la razón de momios [9].

3.5.1 Intervalo de confianza para $\boldsymbol{\beta_{kj}}$

De igual forma que el modelo binomial, el intervalo de confianza de un coeficiente es conformado por la estimación del parámetro más o menos el punto porcentual de una distribución *normal* multiplicado por la estimación del error estándar sobre este parámetro. Así el intervalo de confianza de $(1-\alpha)\times 100\%$ para el parámetro β_{kj}

$$\hat{\beta}_{kj} \pm z_{(1-\alpha/2)}\widehat{SE}_{\hat{\beta}_{kj}}$$

3.5.2 Intervalos de confianza para $\boldsymbol{\pi_{ij}}$

Así el intervalo de $(1-\alpha)\times 100\%$ confianza para $\pi(x_{ij})$ queda de la siguiente forma:

$$\hat{\pi}(x_{ij}) \pm z_{1-\alpha/2}\widehat{SE}_{\hat{\pi}(x_{ij})}$$

3.5.3 Intervalo de confianza para $OR_j(x_k)$

Así el intervalo de $(1-\alpha)\times 100\%$ confianza para $OR_j(x_k)$ queda de la siguiente forma:

$$\left(exp\left(\hat{\beta}_{kj} - z_{(1-\alpha/2)}\widehat{SE}_{\hat{\beta}_{kj}}\right), exp\left(\hat{\beta}_{kj} + z_{(1-\alpha/2)}\widehat{SE}_{\hat{\beta}_{kj}}\right)\right)$$

3.5.4 Intervalo de confianza para $OR_j(x'_k)$

Así el intervalo de $(1-\alpha)\times 100\%$ confianza para $OR_j(x'_k)$ queda de la siguiente forma:

$$\left(exp\left(S_{x_k}\left(\hat{\beta}_{kj} - z_{(1-\alpha/2)}\widehat{SE}_{\hat{\beta}_{kj}}\right)\right), exp\left(S_{x_k}\left(\hat{\beta}_{kj} + z_{(1-\alpha/2)}\widehat{SE}_{\hat{\beta}_{kj}}\right)\right)\right)$$

3.6 Pruebas de Hipótesis

3.6.1 Estadística Devianza

La estadística devianza, es una prueba sobre la estimación de máxima verosimilitud, en donde se compara el modelo estudiado con un modelo de referencia que provee un ajuste perfecto, en este caso, un modelo saturado.

Se sabe que el estimador por máxima verosimilitud para la función de la regresión logística multinomial está dada por:

$$\ell(\beta) = \prod_{i=1}^{n}\prod_{j=1}^{J}{\pi_{ij}}^{y_{ij}}$$

La estadística dada por el cociente de verosimilitudes prueba la hipótesis nula de un modelo saturado contra la hipótesis alternativa de un modelo más general. Un modelo saturado es aquel que incluye el mismo número de observaciones que parámetros en el modelo.

En un modelo saturado, por definición del modelo mismo, se tiene que $E(Y|X) = y$, los valores observados son iguales a los valores estimados del modelo, entonces, en el modelo logístico se tiene que en un modelo saturado $\hat{\pi}(x_i) = y_i$, por lo que la verosimilitud será:

$$l(modelo\ saturado) = \prod_{i=1}^{n} \prod_{j=1}^{J} {y_{ij}}^{y_{ij}} = 1$$

La devianza del modelo que realiza la comparación entre los logaritmos de verosimilitudes de los modelos saturado y ajustado, queda definida como:

$$D = -2\big[\big(ln(verosimilitud\ del\ modelo\ saturado)\big) - \big(ln(verosimilitud\ del\ modelo\ ajustado)\big)\big]$$

Donde el modelo ajustado es una estimación del modelo restringiendo la variable independiente x_k

En la práctica el problema con esta prueba radica en el tener que estimar el modelo completo junto con los K modelos restringidos por las variables independientes [19].

La prueba de razón de verosimilitudes para el modelo múltiple es análoga al caso binario. Como la función de verosimilitud de un modelo saturado siempre es igual a 1, el cálculo de la estadística G se basa en la misma expresión que el caso binomial, sin embargo, se compara el modelo con K variables independientes con J categorías.

Las hipótesis son:

$$H_0: \beta_{kj} = 0 \quad vs \quad H_a: \beta_{kj} \neq 0$$

$$G = (-2ln(Modelo\ sin\ las\ variables\ independientes)) \\ - (-2ln(Modelo\ con\ las\ variables\ independientes))$$

El modelo con las variables independientes es igual a la verosimilitud del modelo ajustado definido anteriormente.

Para calcular la verosimilitud del modelo excluyendo las variables independientes, se tiene que, al suponer $\hat{\beta}_{kj} = 0$ para toda j y k, la probabilidad de éxito de la variable de respuesta, denotada con $\hat{\pi}_{ij}{}'$, estará dada por:

$$\hat{\pi}_{ij}{}' = \frac{exp(\hat{\beta}_{0j})}{1 + \sum_{j=1}^{J-1} exp(\hat{\beta}_{0j})}$$

Por lo que, hay que encontrar el estimador máximo verosímil de $\pi_{ij}{}'$ bajo este modelo. Por la ecuación 3.6 se tiene que la log-verosimilitud del modelo excluyendo las variables independientes es:

$$\mathcal{L}(\boldsymbol{\beta_0}) = \sum_{i=1}^{n} \sum_{j=1}^{J-1} \left[y_{ij}\, ln\big(exp(\beta_{0j})\big) - ln(1 + \sum_{j=1}^{J-1} exp(\beta_{0j})) \right]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{J-1} \left[y_{ij}\, \beta_{0j} - ln(1 + \sum_{j=1}^{J-1} exp(\beta_{0j})) \right]$$

Entonces, la derivada parcial de β_{0j} es:

$$\frac{\partial \mathcal{L}(\boldsymbol{\beta_0})}{\partial \beta_{0j}} = \sum_{i=1}^{n} \left[y_{ij} - \frac{exp(\beta_{0j})}{1 + \sum_{j=1}^{J-1} exp(\beta_{0j})} \right] = 0$$

Por lo que el estimador máximo verosímil de $\hat{\pi}(x)'$ es:

$$\hat{\pi}_{ij}{}' = \frac{\sum_{i=1}^{n} y_{ij}}{n}$$

donde se supondrá que:

$$n_j = \sum_{i=1}^{n} y_{ij}$$

entonces

$$\hat{\pi}_{ij}{}' = \frac{n_j}{n}$$

De lo anterior, se tiene que la verosimilitud del modelo excluyendo las variables independientes es:

$$verosimilitud\ excluyendo\ las\ variables = \prod_{i=1}^{n} \prod_{j=1}^{J} \left[\left(\frac{n_j}{n} \right)^{y_{ij}} \right]$$

$$G = -2ln\left[\frac{\prod_{i=1}^{n}\prod_{j=1}^{J}\left[\left(\frac{n_j}{n}\right)^{y_{ij}}\right]}{\prod_{i=1}^{n}\prod_{j=1}^{J}\hat{\pi}_{ij}{}^{y_{ij}}}\right]$$

Bajo el supuesto de que todos los parámetros son igual a cero, la estadística G se distribuye ji-cuadrada con K grados de libertad $G \sim \chi^2_{((K)\times(J-1))}$. Esto es, para un nivel de significancia α, se rechazará la hipótesis nula si $G > \chi^{2\,(1-\alpha)}_{((K)\times(J-1))}$. Donde $\chi^{2\,(1-\alpha)}_{((K)\times(J-1))}$ es el cuantil de $1-\alpha$ de una ji-cuadrada con $((K)\times(J-1))$ grados de libertad.

Al nivel de significancia α. Si la condición se cumple, se dice que al menos una variable es significativamente distinta de cero [11].

Usando esta misma notación, el *p-value* asociado a esta prueba es:

$$P\left[\chi^2_{((K)\times(J-1))} > G\right] = p - value$$

Ejemplo 3.4

Con los datos del ejemplo 3.2, se tiene $n_1 = 51$, $n_2 = 15$, y $n_3 = 20$. También se obtiene $-2\log(verosimilitud\ incluyendo\ las\ variables)$ tiene un valor de -151.869. Evaluando G se tiene:

$$G = 2\{-75.93 + [51ln(51) + 15\,ln(15) + 20\,ln(20) - 86ln(86)]\}$$
$$= 12.1609$$

donde

$$\chi^2{}_{(2)}^{(.95)} = 5.99$$

por lo tanto

$$G = 12.1609 > 5.99 = \chi^2{}_{(2)}^{(.95)}$$

$$P[G > 12.1609] = .0023$$

Con lo anterior se evidencia que la variable independiente (Edad) es una variable significativa para predecir la variable Tipo_TDAH a un nivel confianza del 5%.

3.6.2 Prueba de Wald

La prueba de Wald en su forma univariada es útil para comprobar la significancia de variables individualmente.

La hipótesis que se desea probar en la prueba de Wald es:

$$H_0: \beta_{kj} = 0 \quad vs \quad H_a: \beta_{kj} \neq 0$$

Esta prueba se obtiene al hacer el cociente del parámetro estimado bajo máxima verosimilitud con su error estándar estimado, la razón resultante sigue una distribución aproximada de una normal estándar.

La expresión

$$W_{kj} = \frac{\hat{\beta}_{kj}}{\left(\widehat{SE}_{\hat{\beta}_{kj}}\right)^{\frac{1}{2}}}$$

denota la estadística de Wald en su versión univariada. Esta prueba puede ser adoptada en los modelos logísticos múltiples cuando se desea evaluar la significancia individual de una cierta variable k para cierta categoría j.

Por lo que a un nivel de significancia α, se rechazará H_0 si

$$W_j > z_{\left(1-\frac{\alpha}{2}\right)} \quad o \quad W_j < z_{\left(\frac{\alpha}{2}\right)}$$

donde $z_{(\alpha)}$ es el cuantil α de una distribución Normal con media igual a 0 y desviación estándar igual a 1.

El análogo de la estadística de Wald para el caso múltiple se puede calcular de la siguiente forma:

$$\boldsymbol{W} = \widehat{\boldsymbol{\beta}}^T \left[\widehat{Var(\boldsymbol{\beta})}\right]^{-1} \widehat{\boldsymbol{\beta}}$$

$$= \widehat{\boldsymbol{\beta}}^T \left(\boldsymbol{X}^T \widehat{\boldsymbol{V}} \boldsymbol{X}\right) \widehat{\boldsymbol{\beta}}$$

Bajo la hipótesis nula, la estadística $\boldsymbol{W}$ en su versión múltiple se distribuye aproximadamente ${\chi^2}_{(K-1)\times(J-1)}$ [15].

De esta forma, para comprobar la significancia del modelo se utiliza la comparación

$$\boldsymbol{W} > {\chi^2}_{(K-1)\times(J-1)}^{(1-\alpha)}$$

esto es que, se rechazara la hipótesis nula si se cumple lo anterior.

3.6.3 Prueba de Hosmer-Lemeshow

La regresión logística es actualmente un modelo estándar para describir la relación entre variables dependientes con naturaleza nominal y una o más variables independientes. Existen varias pruebas de bondad de ajuste cuando la variable independiente es binaria. Una de las pruebas más utilizadas es la prueba de Hosmer-Lemeshow.

Para la prueba de Hosmer-Lemeshow se recodifica la variable dependiente y_i como un indicador binario $\tilde{y}_{ij}$, donde $\tilde{y}_{ij} = 1$ cuando $y_i = j$ y $\tilde{y}_{ij} = 0$ en cualquier otro caso. Se tiene además las probabilidades estimadas $\hat{\pi}_{ij}$ en cada observación de cada posible resultado de la variable dependiente.

Esta prueba se basa en ordenar las observaciones de acuerdo al complemento de las probabilidades estimadas de la categoría de referencia $\left(\sum_{j=1}^{J-1} \hat{\pi}_{ij}\right)$. Luego se forman G grupos, cada uno con n/G observaciones. Para cada grupo, se calculan las sumas de las frecuencias observadas y estimadas de cada categoría obtenida

$$O_{gj} = \sum_{i \in \Omega_g} \tilde{y}_{ij}$$

$$E_{gj} = \sum_{i \in \Omega_g} \hat{\pi}_{ij}$$

Donde $g = 1, \ldots, G$; y Ω_g denota los índices de cada observación en cada grupo g. Una forma muy útil de resumir el modelo es mediante una tabla de contingencia con los valores O_{gj} y E_{gj}, como se muestra a continuación.

	Y=1	*Y=2*	⋯	*Y=J*
Grupos	*Obs Est*	*Obs Est*		*Obs Est*
1	O_{11} E_{11}	O_{12} E_{12}	...	$O_{1,J}$ $E_{1,J}$
2	O_{21} E_{21}	O_{22} E_{22}	...	$O_{2,J}$ $E_{2,J}$
⋮	⋮	⋮		⋮
G	O_{G1} E_{G1}	O_{G2} E_{G2}	...	$O_{G,J}$ $E_{G,J}$

Tabla 3.2 Frecuencias observadas(O_{gj}) y estimadas(E_{gj}) ordenadas en G grupos.

La estadística para prueba de bondad de ajuste del modelo multinomial es una ji-cuadrada, que se construye a partir de las frecuencias de los datos observados y estimados.

$$C_G = \sum_{g=1}^{G}\sum_{j=1}^{J}\frac{\left(O_{gj}-E_{gj}\right)^2}{E_{gj}}$$

Bajo la hipótesis nula de que el modelo ajustado es el modelo correcto y la muestra es suficientemente grande, la estadística C_G se distribuye como una ji-cuadrada con $(G-2)\times(J-1)$ grados de libertad [8].

La estadística $\hat{C}$ se distribuye ji-cuadrada con $(G-2)\times(J-1)$ grados de libertad $G\sim\chi^2_{(G-2)\times(J-1)}$. Esto es, para un nivel de significancia α, se rechazará la hipótesis nula si $G > \chi^{2\ (1-\alpha)}_{(G-2)\times(J-1)}$. Donde $\chi^{2\ (1-\alpha)}_{(G-2)\times(J-1)}$ es el cuantil de $1-\alpha$ de una ji-cuadrada con $(G-2)\times(J-1)$ grados de libertad.

Ejemplo 3.5

Tomando en cuenta los datos del ejemplo 3.2 junto con las probabilidades estimadas para cada una de las categorías, y ordenando los datos de mayor a menor probabilidad estimada de la categoría $Ausencia$, se obtiene la siguiente tabla:

Tabla de Contingencia Prueba Hosmer-Lemeshow							
	Tipo_TDAH=1 Mixto		Tipo_TDAH=2 Inatento		Tipo_TDAH=3 Ausente		
Grupos	Observado	Esperado	Observado	Esperado	Observad	Esperado	Total
1	0	2.37	2	1.23	7	5.4	9
2	5	3.66	0	1.42	3	2.92	8
3	7	4.56	1	1.65	1	2.82	9
4	5	4.48	2	1.52	1	2.08	8
5	4	5.6	3	1.71	2	1.78	9
6	7	5.97	2	1.66	0	1.41	9
7	7	5.52	1	1.44	0	1.04	8
8	5	6.41	0	1.58	4	1.01	9
9	5	5.92	2	1.36	1	0.72	8
10	6	6.7	2	1.52	1	0.78	9

Tabla 3.3 Tabla de Contingencia Prueba Hosmer Lemeshow

Prueba Hosme-Lemeshow		
ji-Cuadrada	Grados de libertad	Sig.
25.015	16	0.069

Tabla 3.4 Prueba Hosmer-Lemeshow

$$\chi^2{}_{(16)}^{(.95)} = 26.3$$

por lo tanto

$$\hat{C} = 25.015 < 26.3 = \chi^2{}_{(16)}^{(.95)}$$

La prueba muestra que es no significativa a un nivel de confianza del 95%. Eso quiere decir que no se rechaza la hipótesis nula, en donde las variables observadas son iguales a los valores esperados.

Como se mostró en el capítulo anterior, la prueba de bondad de ajuste de Hosmer-Lemeshow consiste en una prueba de comparación de datos observados y estimados. Si se llegaran a escoger diferentes categorías de referencia, esto

provocaría diferentes resultados. La sensibilidad de la prueba al momento de escoger la categoría de referencia generalmente es pequeña, es preferible evitar el uso de una categoría de referencia que tenga pocas observaciones.

Las pruebas de bondad de ajuste están hechas para detectar un mal ajuste en el modelo. Sin embargo, estas pruebas no pueden evaluar completamente el ajuste del modelo. Las pruebas de bondad de ajuste deben ser consideradas como una de varias herramientas para evaluar la bondad de ajuste de un modelo. En concreto no se puede concluir que el modelo ajusta o no con base a un resultado de una prueba de bondad de ajuste. Muchas veces es necesario hacer análisis sobre observaciones en particular y con ello es mejor utilizar otra clase de procedimientos, por ejemplo, el diagnóstico de regresión o ciertas técnicas gráficas.

Las pruebas de bondad de ajuste no son herramientas para la construcción de modelos, ni tampoco para su comparación o su selección, son para ajustar el modelo que se tiene pensado usar.

El problema que tienen los modelos de regresión logística es la poca potencia que tienen en las pruebas de bondad de ajuste. Esto quiere decir que, se necesita tener un tamaño de muestra grande para detectar pequeñas y medianas diferencias en el modelo.

Comentarios Finales

En este trabajo se trató de presentar una introducción sobre la regresión logística multinomial, y con ello presentar algunos de los modelos que dan una imagen de cómo va surgiendo la idea de este modelo. Con la finalidad de proporcionar material a futuras generaciones que deseen información introductoria a este tipo de regresión, ya que el material con el que normalmente algún estudiante tiene a la mano, sean libro o internet, en muchas ocasiones es escasa, pobremente desarrollada, y poco intuitiva. Y así poder aplicar el modelo a la investigación que así lo requiera.

Los modelos logísticos presentados en este trabajo son herramientas esenciales en los campos de investigación. Esto se debe a que dentro de la naturaleza de los datos es común encontrar datos categóricos, a diferencia de datos de tipo continuo, debido a que, es más sencillo que en las encuestas un entrevistado responda a una pregunta categórica que a una continua.

El modelo multinomial no es una generalización del modelo binomial sino el modelo binomial es un caso particular del modelo multinomial. Esto lleva como resultado la generalización del modelo para la estimación de los parámetros, los momios y la razón de momios, los intervalos de confianza, y sobre todo las pruebas de bondad de ajuste. Sobre este último, es interesante la prueba de Hosmer-Lemeshow, que con la generalización del modelo se puede apreciar que es una tabla de contingencia.

De entre los problemas que se pueden encontrar al querer implementar dicho modelo, está el decidir qué categoría es la que se tomará como referencia, ya que, con base en esto todos los resultados pueden llegar a cambiar, desde los parámetros hasta las pruebas de bondad de ajuste, ya que la sensibilidad de muchas pruebas se basa en el número de casos que caigan en la categoría que se esté tomando como

base, por ejemplo, la prueba de Hosmer-Lemeshow. Es por eso que tiene mucha importancia la bondad de ajuste, ya que en la práctica se decide qué categoría será la base dependiendo en qué tanto mejore su bondad de ajuste.

Una prueba interesante en el modelo multinomial es la prueba de independencia de alternativas irrelevantes, la cual implica que cada posible categoría no posee alguna otra categoría que abarque alguna otra categoría cercana como sustituto entre las posibles categorías a elegir. Este tema no se menciona en este trabajo debido a su complejidad, sin embargo, si se desea buscar más información, se puede consultar el articulo *Testing for IIA with the Hausman-McFadden Test*e de Wim Vijverberg [22].

Apéndice

A1. Datos de Capítulo 1

Tabla 1

Folio	Edad	TDAHMIXTDXS	Pi_gorro_1	1-Pi_gorro_1	LRE_1	DEV_1
1	16	0	0.65743	0.34257	-2.91915	-1.46376
2	14	1	0.75943	0.24057	1.31677	0.74187
3	13	1	0.80193	0.19807	1.24699	0.66443
4	15	0	0.7111	0.2889	-3.46139	-1.57586
5	13	1	0.80193	0.19807	1.24699	0.66443
5	15	1	0.7111	0.2889	1.40627	0.82577
6	14	1	0.75943	0.24057	1.31677	0.74187
7	14	1	0.75943	0.24057	1.31677	0.74187
8	14	0	0.75943	0.24057	-4.15684	-1.68805
8	16	0	0.65743	0.34257	-2.91915	-1.46376
9	14	1	0.75943	0.24057	1.31677	0.74187
10	16	1	0.65743	0.34257	1.52106	0.91587
11	14	1	0.75943	0.24057	1.31677	0.74187
11	18	1	0.53847	0.46153	1.85711	1.11267
12	17	1	0.59942	0.40058	1.66829	1.01173
12	19	0	0.47636	0.52364	-1.90969	-1.13749
13	13	1	0.80193	0.19807	1.24699	0.66443
14	14	1	0.75943	0.24057	1.31677	0.74187
14	12	0	0.83852	0.16148	-6.19273	-1.90965
15	16	1	0.65743	0.34257	1.52106	0.91587
16	16	0	0.65743	0.34257	-2.91915	-1.46376
16	15	0	0.7111	0.2889	-3.46139	-1.57586

17	13	0	0.80193	0.19807	-5.04878	-1.79953
18	17	1	0.59942	0.40058	1.66829	1.01173
19	13	1	0.80193	0.19807	1.24699	0.66443
19	20	0	0.41496	0.58504	-1.70929	-1.03545
20	19	0	0.47636	0.52364	-1.90969	-1.13749
20	15	1	0.7111	0.2889	1.40627	0.82577
21	15	0	0.7111	0.2889	-3.46139	-1.57586
22	13	1	0.80193	0.19807	1.24699	0.66443
22	17	1	0.59942	0.40058	1.66829	1.01173
23	16	1	0.65743	0.34257	1.52106	0.91587
24	16	0	0.65743	0.34257	-2.91915	-1.46376
24	14	1	0.75943	0.24057	1.31677	0.74187
25	17	1	0.59942	0.40058	1.66829	1.01173
25	13	1	0.80193	0.19807	1.24699	0.66443
26	17	0	0.59942	0.40058	-2.49636	-1.35265
27	12	1	0.83852	0.16148	1.19258	0.59349
27	17	1	0.59942	0.40058	1.66829	1.01173
28	20	1	0.41496	0.58504	2.40987	1.32633
28	16	0	0.65743	0.34257	-2.91915	-1.46376
29	15	1	0.7111	0.2889	1.40627	0.82577
29	12	1	0.83852	0.16148	1.19258	0.59349
30	16	1	0.65743	0.34257	1.52106	0.91587
30	19	1	0.47636	0.52364	2.09927	1.21786
31	15	1	0.7111	0.2889	1.40627	0.82577
33	16	1	0.65743	0.34257	1.52106	0.91587
34	13	1	0.80193	0.19807	1.24699	0.66443
35	17	1	0.59942	0.40058	1.66829	1.01173
36	13	1	0.80193	0.19807	1.24699	0.66443
36	19	0	0.47636	0.52364	-1.90969	-1.13749

37	14	1	0.75943	0.24057	1.31677	0.74187
37	18	1	0.53847	0.46153	1.85711	1.11267
38	13	0	0.80193	0.19807	-5.04878	-1.79953
39	15	1	0.7111	0.2889	1.40627	0.82577
40	16	0	0.65743	0.34257	-2.91915	-1.46376
40	15	1	0.7111	0.2889	1.40627	0.82577
41	17	1	0.59942	0.40058	1.66829	1.01173
42	16	0	0.65743	0.34257	-2.91915	-1.46376
43	16	1	0.65743	0.34257	1.52106	0.91587
44	14	1	0.75943	0.24057	1.31677	0.74187
44	19	0	0.47636	0.52364	-1.90969	-1.13749
45	15	0	0.7111	0.2889	-3.46139	-1.57586
45	16	1	0.65743	0.34257	1.52106	0.91587
46	15	1	0.7111	0.2889	1.40627	0.82577
46	13	1	0.80193	0.19807	1.24699	0.66443
48	18	0	0.53847	0.46153	-2.16672	-1.24355
49	17	1	0.59942	0.40058	1.66829	1.01173
52	16	1	0.65743	0.34257	1.52106	0.91587
52	20	0	0.41496	0.58504	-1.70929	-1.03545
53	16	1	0.65743	0.34257	1.52106	0.91587
54	12	1	0.83852	0.16148	1.19258	0.59349
55	15	1	0.7111	0.2889	1.40627	0.82577
56	16	0	0.65743	0.34257	-2.91915	-1.46376
56	13	1	0.80193	0.19807	1.24699	0.66443
57	17	1	0.59942	0.40058	1.66829	1.01173
58	17	1	0.59942	0.40058	1.66829	1.01173
59	15	1	0.7111	0.2889	1.40627	0.82577
59	17	1	0.59942	0.40058	1.66829	1.01173
60	13	1	0.80193	0.19807	1.24699	0.66443

60	14	1	0.75943	0.24057	1.31677	0.74187
61	15	1	0.7111	0.2889	1.40627	0.82577
61	13	0	0.80193	0.19807	-5.04878	-1.79953
62	14	1	0.75943	0.24057	1.31677	0.74187
62	15	0	0.7111	0.2889	-3.46139	-1.57586
63	14	1	0.75943	0.24057	1.31677	0.74187
63	15	0	0.7111	0.2889	-3.46139	-1.57586
64	14	0	0.75943	0.24057	-4.15684	-1.68805
64	19	0	0.47636	0.52364	-1.90969	-1.13749
65	17	1	0.59942	0.40058	1.66829	1.01173
65	13	1	0.80193	0.19807	1.24699	0.66443
66	16	1	0.65743	0.34257	1.52106	0.91587
67	16	1	0.65743	0.34257	1.52106	0.91587
67	13	0	0.80193	0.19807	-5.04878	-1.79953
68	17	1	0.59942	0.40058	1.66829	1.01173
68	20	0	0.41496	0.58504	-1.70929	-1.03545
69	13	1	0.80193	0.19807	1.24699	0.66443
69	20	0	0.41496	0.58504	-1.70929	-1.03545
70	13	1	0.80193	0.19807	1.24699	0.66443
70	14	1	0.75943	0.24057	1.31677	0.74187
71	14	1	0.75943	0.24057	1.31677	0.74187
71	13	1	0.80193	0.19807	1.24699	0.66443
72	17	1	0.59942	0.40058	1.66829	1.01173
72	20	0	0.41496	0.58504	-1.70929	-1.03545
73	14	1	0.75943	0.24057	1.31677	0.74187
73	13	0	0.80193	0.19807	-5.04878	-1.79953
74	13	0	0.80193	0.19807	-5.04878	-1.79953
75	15	1	0.7111	0.2889	1.40627	0.82577
76	16	1	0.65743	0.34257	1.52106	0.91587

77	15	1	0.7111	0.2889	1.40627	0.82577
77	18	1	0.53847	0.46153	1.85711	1.11267
78	15	1	0.7111	0.2889	1.40627	0.82577
78	15	0	0.7111	0.2889	-3.46139	-1.57586
79	17	1	0.59942	0.40058	1.66829	1.01173
80	13	1	0.80193	0.19807	1.24699	0.66443
81	13	1	0.80193	0.19807	1.24699	0.66443
82	13	1	0.80193	0.19807	1.24699	0.66443
84	15	1	0.7111	0.2889	1.40627	0.82577
84	13	0	0.80193	0.19807	-5.04878	-1.79953
85	13	1	0.80193	0.19807	1.24699	0.66443
93	15	0	0.7111	0.2889	-3.46139	-1.57586

Syntax Ejemplo 1.1

Se obtienen los datos dentro de una base de datos con format “sav”.

```
get file 'F:\Tesis\Tesis\Base Regresión Logística Simple.sav'.
```

Syntax Tabla 1.1

```
FREQUENCIES   Edad.
```

Syntax Figura 1.1

```
* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=Edad TDAHMIXTDXS
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: Edad=col(source(s), name("Edad"))
```

```
  DATA: TDAHMIXTDXS=col(source(s), name("TDAHMIXTDXS"),
unit.category())
  GUIDE: axis(dim(1), label("Edad"))
  GUIDE: axis(dim(2), label("TDAHMIXTDXS"))
  SCALE: cat(dim(2), include(".00", "1.00"))
  ELEMENT: point(position(Edad*TDAHMIXTDXS))
END GPL.
```

Syntax Tabla 1.2

```
VALUE LABELS   TDAHMIXTDXS
0'Ausencia'
1'Presencia'.

* Custom Tables.
CTABLES
  /VLABELS VARIABLES=Edad TDAHMIXTDXS DISPLAY=DEFAULT
  /TABLE Edad [COUNT F40.0] BY TDAHMIXTDXS
  /CATEGORIES VARIABLES=Edad TDAHMIXTDXS ORDER=A KEY=VALUE
EMPTY=EXCLUDE.
```

Syntax Figura 1.2

```
Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=Edad
Proporcion_TDAH_MIXTO MISSING=LISTWISE
    REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: Edad2=col(source(s), name("Edad"))
  DATA: Proporcion_TDAH_MIXTO=col(source(s),
name("Proporcion_TDAH_MIXTO"))
  GUIDE: axis(dim(1), label("Edad"))
  GUIDE: axis(dim(2), label("Proporcion_TDAH_MIXTO"))
  ELEMENT: point(position(Edad2*Proporcion_TDAH_MIXTO))
```

```
END GPL.
```

Syntax Ejemplo 1.2 Tabla 1.3, Ejemplo 1.3, y Ejemplo 1.4

```
LOGISTIC REGRESSION VARIABLES TDAHMIXTDXS
  /METHOD=ENTER Edad
  /SAVE=PRED LRESID DEV
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

A2. Datos de Capítulo 2

Tabla 2

Folio	TDAHMIXTDXS	Talla	Edad	TDAHMIXTDXS	Pi_gorro_1	1-Pi_gorro_1	LRE_1	DEV_1	pi_gorro_yi
52	0	1.58	20	0	0.259	0.741	-1.34952	-0.774	-0.299754654
68	0	1.58	20	0	0.259	0.741	-1.34952	-0.774	-0.299754654
72	0	1.65	20	0	0.33241	0.66759	-1.49792	-0.899	-0.404081066
64	0	1.6	19	0	0.36325	0.63675	-1.57048	-0.950	-0.451378165
69	0	1.69	20	0	0.37869	0.62131	-1.60951	-0.976	-0.475925127
57	1	1.49	17	1	0.41587	0.58413	2.40458	1.325	-0.877382568
59	1	1.55	17	1	0.4909	0.5091	2.03709	1.193	-0.711514838
56	0	1.5	16	0	0.52489	0.47511	-2.10476	-1.220	-0.744208923
25	1	1.59	17	1	0.54135	0.45865	1.84722	1.108	-0.613689259
30	1	1.54	16	1	0.57489	0.42511	1.73946	1.052	-0.553576561
18	1	1.63	17	1	0.59098	0.40902	1.69211	1.026	-0.525973103
72	1	1.63	17	1	0.59098	0.40902	1.69211	1.026	-0.525973103
77	1	1.72	18	1	0.60687	0.39313	1.6478	0.999	-0.499440679
67	1	1.58	16	1	0.62341	0.37659	1.60408	0.972	-0.472550871
44	0	1.82	19	0	0.63434	0.36566	-2.73481	-1.418	-1.006051339
1	0	1.59	16	0	0.6352	0.3648	-2.74125	-1.420	-1.008406021
58	1	1.68	17	1	0.65039	0.34961	1.53753	0.928	-0.430183096

68	1	1.69	17	1	0.6618	0.3382	1.51103	0.909	-0.412791883
65	1	1.7	17	1	0.67302	0.32698	1.48584	0.890	-0.395980232
78	1	1.55	15	1	0.67726	0.32274	1.47654	0.883	-0.389700033
78	0	1.56	15	0	0.68821	0.31179	-3.20726	-1.527	-1.165425395
23	1	1.64	16	1	0.69155	0.30845	1.44603	0.859	-0.368819824
4	0	1.58	15	0	0.70948	0.29052	-3.44211	-1.572	-1.236082858
10	1	1.66	16	1	0.71269	0.28731	1.40314	0.823	-0.338708736
19	1	1.43	13	1	0.71345	0.28655	1.40165	0.822	-0.337642922
61	1	1.59	15	1	0.71979	0.28021	1.3893	0.811	-0.328795776
62	0	1.59	15	0	0.71979	0.28021	-3.56874	-1.595	-1.272215957
52	1	1.67	16	1	0.72293	0.27707	1.38326	0.806	-0.32444288
79	1	1.75	17	1	0.72604	0.27396	1.37733	0.800	-0.320150169
73	1	1.52	14	1	0.72678	0.27322	1.37593	0.799	-0.319131461
24	0	1.68	16	0	0.73294	0.26706	-3.74445	-1.625	-1.320281927
76	1	1.68	16	1	0.73294	0.26706	1.36437	0.788	-0.310691436
5	1	1.45	13	1	0.73366	0.26634	1.36302	0.787	-0.309709573
8	0	1.53	14	0	0.7367	0.2633	-3.798	-1.634	-1.334461212
70	1	1.53	14	1	0.7367	0.2633	1.3574	0.782	-0.305574525
53	1	1.69	16	1	0.74272	0.25728	1.34641	0.771	-0.297436156
55	1	1.63	15	1	0.75871	0.24129	1.31802	0.743	-0.276135656
70	1	1.48	13	1	0.76223	0.23777	1.31194	0.737	-0.271506932
11	1	1.56	14	1	0.76505	0.23495	1.30711	0.732	-0.267814088
60	1	1.56	14	1	0.76505	0.23495	1.30711	0.732	-0.267814088
71	1	1.56	14	1	0.76505	0.23495	1.30711	0.732	-0.267814088
77	1	1.64	15	1	0.76784	0.23216	1.30235	0.727	-0.264173901
73	0	1.49	13	0	0.77127	0.22873	-4.37199	-1.718	-1.47521301
59	1	1.65	15	1	0.77673	0.22327	1.28744	0.711	-0.252662479
60	1	1.5	13	1	0.78007	0.21993	1.28194	0.705	-0.24837162
6	1	1.58	14	1	0.78273	0.21727	1.27757	0.700	-0.24496747

81	1	1.51	13	1	0.78862	0.21138	1.26804	0.689	-0.237470696
2	1	1.59	14	1	0.79121	0.20879	1.26389	0.684	-0.23419186
63	1	1.59	14	1	0.79121	0.20879	1.26389	0.684	-0.23419186
64	0	1.59	14	0	0.79121	0.20879	-4.78946	-1.770	-1.566426317
17	0	1.52	13	0	0.79692	0.20308	-4.9242	-1.786	-1.594155289
71	1	1.52	13	1	0.79692	0.20308	1.25483	0.674	-0.227000982
80	1	1.52	13	1	0.79692	0.20308	1.25483	0.674	-0.227000982
24	1	1.6	14	1	0.79944	0.20056	1.25088	0.669	-0.223843796
66	1	1.76	16	1	0.8044	0.1956	1.24317	0.660	-0.217658621
61	0	1.53	13	0	0.80498	0.19502	-5.12768	-1.808	-1.634653162
67	0	1.53	13	0	0.80498	0.19502	-5.12768	-1.808	-1.634653162
14	1	1.61	14	1	0.80742	0.19258	1.23851	0.654	-0.2139113
62	1	1.61	14	1	0.80742	0.19258	1.23851	0.654	-0.2139113
21	0	1.69	15	0	0.80984	0.19016	-5.25862	-1.822	-1.659889456
63	0	1.69	15	0	0.80984	0.19016	-5.25862	-1.822	-1.659889456
54	1	1.46	12	1	0.81041	0.18959	1.23395	0.648	-0.210214987
56	1	1.54	13	1	0.81279	0.18721	1.23032	0.644	-0.207282505
75	1	1.7	15	1	0.8175	0.1825	1.22324	0.635	-0.201504376
9	1	1.63	14	1	0.82265	0.17735	1.21558	0.625	-0.195224442
3	1	1.56	13	1	0.82769	0.17231	1.20818	0.615	-0.189116591
13	1	1.57	13	1	0.83479	0.16521	1.19791	0.601	-0.180575083
34	1	1.58	13	1	0.84164	0.15836	1.18816	0.587	-0.17240291
74	0	1.58	13	0	0.84164	0.15836	-6.31472	-1.920	-1.842884357
65	1	1.59	13	1	0.84826	0.15174	1.17888	0.574	-0.164568086
69	1	1.66	13	1	0.88844	0.11156	1.12557	0.486	-0.118288163
7	1	1.83	14	1	0.92727	0.07273	1.07843	0.389	-0.075510494

Se obtienen los datos dentro de una base de datos con formato “sav”.

```
get file 'F:\Tesis\Tesis\Base Regresión Logística Multiple.sav'.
```

Syntax Ejemplo 2.1 Tabla 2.1, Ejemplo 2.2, y Ejemplo 2.3

```
LOGISTIC REGRESSION VARIABLES TDAHMIXTDXS
  /METHOD=ENTER Edad  Talla
  /SAVE=PRED LRESID DEV
  /PRINT=GOODFIT
  /CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

A3. Datos de Capítulo 3

Tabla 3

Folio	Edad	tipo_TDAH	Pi_gorro_1	Pi_gorro_2	Pi_gorro_3	PRE_1	PCP_1	nj/n	pi_gorro_yij	Grupo
52	20	2	0.24	0.13	0.63	3	0.63	0.174	0.13	1
53	20	3	0.24	0.13	0.63	3	0.63	0.233	0.63	1
66	20	3	0.24	0.13	0.63	3	0.63	0.233	0.63	1
68	20	2	0.24	0.13	0.63	3	0.63	0.174	0.13	1
69	20	3	0.24	0.13	0.63	3	0.63	0.233	0.63	1
72	20	3	0.24	0.13	0.63	3	0.63	0.233	0.63	1
21	19	3	0.31	0.15	0.54	3	0.54	0.233	0.54	1
44	19	3	0.31	0.15	0.54	3	0.54	0.233	0.54	1
64	19	3	0.31	0.15	0.54	3	0.54	0.233	0.54	1
54	18	3	0.39	0.17	0.44	3	0.44	0.233	0.44	2
77	18	1	0.39	0.17	0.44	3	0.44	0.593	0.39	2
17	17	3	0.48	0.18	0.34	1	0.48	0.233	0.34	2
18	17	1	0.48	0.18	0.34	1	0.48	0.593	0.48	2
25	17	1	0.48	0.18	0.34	1	0.48	0.593	0.48	2
55	17	3	0.48	0.18	0.34	1	0.48	0.233	0.34	2
57	17	1	0.48	0.18	0.34	1	0.48	0.593	0.48	2
58	17	1	0.48	0.18	0.34	1	0.48	0.593	0.48	2
59	17	1	0.48	0.18	0.34	1	0.48	0.593	0.48	3

65	17	1	0.48	0.18	0.34	1	0.48	0.593	0.48	3
68	17	1	0.48	0.18	0.34	1	0.48	0.593	0.48	3
72	17	1	0.48	0.18	0.34	1	0.48	0.593	0.48	3
74	17	3	0.48	0.18	0.34	1	0.48	0.233	0.34	3
79	17	1	0.48	0.18	0.34	1	0.48	0.593	0.48	3
1	16	2	0.56	0.19	0.26	1	0.56	0.174	0.19	3
10	16	1	0.56	0.19	0.26	1	0.56	0.593	0.56	3
23	16	1	0.56	0.19	0.26	1	0.56	0.593	0.56	3
24	16	2	0.56	0.19	0.26	1	0.56	0.174	0.19	4
30	16	1	0.56	0.19	0.26	1	0.56	0.593	0.56	4
52	16	1	0.56	0.19	0.26	1	0.56	0.593	0.56	4
53	16	1	0.56	0.19	0.26	1	0.56	0.593	0.56	4
56	16	2	0.56	0.19	0.26	1	0.56	0.174	0.19	4
57	16	3	0.56	0.19	0.26	1	0.56	0.233	0.26	4
66	16	1	0.56	0.19	0.26	1	0.56	0.593	0.56	4
67	16	1	0.56	0.19	0.26	1	0.56	0.593	0.56	4
76	16	1	0.56	0.19	0.26	1	0.56	0.593	0.56	5
4	15	2	0.63	0.19	0.19	1	0.63	0.174	0.19	5
21	15	2	0.63	0.19	0.19	1	0.63	0.174	0.19	5
55	15	1	0.63	0.19	0.19	1	0.63	0.593	0.63	5
58	15	3	0.63	0.19	0.19	1	0.63	0.233	0.19	5
59	15	1	0.63	0.19	0.19	1	0.63	0.593	0.63	5
61	15	1	0.63	0.19	0.19	1	0.63	0.593	0.63	5
62	15	2	0.63	0.19	0.19	1	0.63	0.174	0.19	5
63	15	3	0.63	0.19	0.19	1	0.63	0.233	0.19	5
75	15	1	0.63	0.19	0.19	1	0.63	0.593	0.63	6
77	15	1	0.63	0.19	0.19	1	0.63	0.593	0.63	6
78	15	1	0.63	0.19	0.19	1	0.63	0.593	0.63	6
78	15	2	0.63	0.19	0.19	1	0.63	0.174	0.19	6

2	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	6
6	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	6
7	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	6
8	14	2	0.69	0.18	0.13	1	0.69	0.174	0.18	6
9	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	6
11	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	7
14	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	7
24	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	7
60	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	7
62	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	7
63	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	7
64	14	2	0.69	0.18	0.13	1	0.69	0.174	0.18	7
70	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	7
71	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	8
73	14	1	0.69	0.18	0.13	1	0.69	0.593	0.69	8
79	14	3	0.69	0.18	0.13	1	0.69	0.233	0.13	8
80	14	3	0.69	0.18	0.13	1	0.69	0.233	0.13	8
81	14	3	0.69	0.18	0.13	1	0.69	0.233	0.13	8
3	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	8
5	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	8
13	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	8
15	13	3	0.74	0.17	0.09	1	0.74	0.233	0.09	8
17	13	2	0.74	0.17	0.09	1	0.74	0.174	0.17	9
19	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	9
34	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	9
56	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	9
60	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	9
61	13	3	0.74	0.17	0.09	1	0.74	0.233	0.09	9
65	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	9

67	13	2	0.74	0.17	0.09	1	0.74	0.174	0.17	9
69	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	10
70	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	10
71	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	10
73	13	2	0.74	0.17	0.09	1	0.74	0.174	0.17	10
74	13	2	0.74	0.17	0.09	1	0.74	0.174	0.17	10
75	13	3	0.74	0.17	0.09	1	0.74	0.233	0.09	10
80	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	10
81	13	1	0.74	0.17	0.09	1	0.74	0.593	0.74	10
54	12	1	0.78	0.16	0.06	1	0.78	0.593	0.78	10

Se obtienen los datos dentro de una base de datos con formato “sav”.

```
get file 'F:\Tesis\Tesis\Base Regresión Logística Multinomial.sav'.
```

Syntax Ejemplo 3.1 Figura 3.1

```
VALUE LABELS   tipo_TDAH
1 'Inatento'
2 'Mixto'
3 'Ausente'.

* Chart Builder.
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=tipo_TDAH
MEANCI(Edad, 95)[name="MEAN_Edad"
    LOW="MEAN_Edad_LOW" HIGH="MEAN_Edad_HIGH"] MISSING=LISTWISE
REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: tipo_TDAH=col(source(s), name("tipo_TDAH"),
unit.category())
  DATA: MEAN_Edad=col(source(s), name("MEAN_Edad"))
```

```
  DATA: LOW=col(source(s), name("MEAN_Edad_LOW"))
  DATA: HIGH=col(source(s), name("MEAN_Edad_HIGH"))
  GUIDE: axis(dim(1), label("tipo_TDAH"))
  GUIDE: axis(dim(2), label("Mean Edad"))
  GUIDE: text.footnote(label("Error Bars: 95% CI"))
  SCALE: cat(dim(1), include("1.00", "2.00", "3.00"))
  SCALE: linear(dim(2), include(0))
  ELEMENT: point(position(tipo_TDAH*MEAN_Edad))
  ELEMENT:
interval(position(region.spread.range(tipo_TDAH*(LOW+HIGH))),
    shape.interior(shape.ibeam))
END GPL.
```

Syntax Ejemplo 3.2, Ejemplo 3.3, Ejemplo 3.4, y Ejempo 3.5

```
NOMREG tipo_TDAH (BASE=3 ORDER=ASCENDING) WITH edad
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5) CHKSEP(20)
LCONVERGE(0) PCONVERGE(0.000001)
    SINGULAR(0.00000001)
  /MODEL
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0) RULE(SINGLE)
ENTRYMETHOD(LR) REMOVALMETHOD(LR)
  /INTERCEPT=INCLUDE
  /PRINT=CELLPROB CLASSTABLE FIT PARAMETER SUMMARY LRT CPS STEP
MFI
  /SAVE ESTPROB PREDCAT PCPROB.
```

Anexos

Distribución Multinomial

Se considera una muestra aleatoria de tamaño n, se desea obtener la probabilidad de un cierto evento del cual hay J posibles resultados. Lo anterior se puede modelar con una variable aleatoria X, la cual indica el resultado de dicho evento. Sean $y_1, \dots, y_J$ los distintos tipos de resultados dentro del evento. Por lo tanto, X toma valores en un conjunto de $\{y_1, \dots, y_J\}$, y se definen las probabilidades como $p_j = P(X = y_j)$ donde $\sum_{i=1}^{J} p_i = 1$

Teniendo en cuenta la muestra aleatoria se define el vector aleatorio $\boldsymbol{N} = (N_1, \dots, N_J)$ que indica en cada entrada la frecuencia de la $j-$esima ocurrencia del tipo y_i en la muestra. Entonces la distribución de $\boldsymbol{N}$ es una multinomial con parámetros n y $\boldsymbol{p} = (p_1, \dots, p_J)$.

Entonces

$$P(N_1 = n_1, \dots, N_J = n_J) = \binom{n}{n_1, \dots, n_J} {p_1}^{n_1} \dots {p_J}^{n_J}$$

Donde

$$\binom{n}{n_1, \dots, n_J} = \frac{n!}{n_1!\, n_2! \dots n_J!}$$

La deducción de esta distribución viene en dos partes, la parte de las probabilidades, y el coeficiente que viene asociado a $(n_1, \dots, n_J)$, que se denotará como $\alpha_{(n_1,\dots,n_J)}$.

La parte de las probabilidades viene del hecho de obtener n_1 resultados del tipo y_1, obtener n_2 resultados del tipo y_2,…, obtener $n_J$resultados del tipo y_J (sin importar el orden en que salen con respecto al total), esto es ${p_1}^{n_1} \dots {p_J}^{n_J}$.

Sea $\alpha_1 = \begin{Bmatrix} número\ de\ formas\ de\ obtener\ n_1\ resultados\ del\ tipo\ y_1 \\ de\ entre\ n\ disponibles \end{Bmatrix}$

$$= \frac{n!}{(n-n_1)!\,n_1!}$$

Sea $\alpha_2 = \begin{Bmatrix} número\ de\ formas\ de\ obtener\ n_2\ resultados\ del\ tipo\ y_2 \\ de\ entre\ n-n_1\ disponibles \end{Bmatrix}$

$$= \frac{(n-n_1)!}{(n-n_1-n_2)!\,n_2!}$$

$$\vdots$$

Sea $\alpha_{J-1} = \begin{Bmatrix} número\ de\ formas\ de\ obtener\ n_{J-1}\ resultados \\ del\ tipo\ y_{J-1} \\ de\ entre\ n-n_1-\cdots-n_{J-2}\ disponibles \end{Bmatrix}$

$$= \frac{(n-n_1-\cdots-n_{J-2})!}{(n-n_1-\cdots-n_{J-2}-n_{J-1})!\,n_{J-1}!}$$

Sea $\alpha_J = \begin{Bmatrix} número\ de\ formas\ de\ obtener\ n_k\ resultados\ del \\ tipo\ y_J \\ de\ entre\ n-n_1-\cdots-n_{J-1}=n_J\ disponibles \end{Bmatrix}$

$$= \frac{(n-n_1-\cdots-n_{J-1})!}{(n-n_1-\cdots-n_{J-1}-n_J)!\,n_J!}$$

De aquí se puede mostrar fácilmente que $\alpha_{(n_1,\ldots,n_J)} = \alpha_1\alpha_2 \ldots \alpha_J$, y desarrollando algo de álgebra la expresión queda como el coeficiente de la distribución multinomial.

Para el caso en el que $J = 2$, se puede notar de inmediato que la distribución multinomial coincide con la binomial. Como $p_1 + p_2 = 1$ además de que $p_2 = 1 - p_1$, y se define $p = p_1$ y $q = p_2$. De igual forma $n_2 = n - n_1$. Reemplazando en la distribución multinomial los valores anteriores, se obtiene que $P(N_1 = n_1, N_2 = n_2) = P(N_1 = n_1)$, donde N_1 se distribuye como una binomial de parámetros n y $p = p_1$ [3].

Bibliografía

[1] Agresti, Alan. (2007). *An Introduction to Categorical Data Analysis*. New Jersey: John Wiley & Sons.

[2] Agresti, Alan. (2007). *Categorical Data Analysis*. New Jersey: John Wiley & Sons.

[3] Assar, Rodrigo. Iturriaga, Andrés. y Riquelme, Victor. (2012) *Distribución Multinomial*. Chile.

[4] Chatterjee, Samprit. y Hadi, Alis. (2006). *Regression Analysis by Example*. New Jersey: Wiley Series in Probability and Statistics.

[5] Cirillo, Marcelo. y de Siqueria, Patrícia. (2014) *Goodness-of-fit Test for Modified Multinomial Logit Models*. Chilean Journal Statistics. Vol. 5, No. 1, April, 73–85.

[6] Cramer J.S., (2003). *Logit Models from Economics and Other Fields*. New York: Cambridge Press

[7] Czepiel, Scott. (2015). *Maximun Likehood Estimation of Logistic Regression Models*: Theory and Implementation.

[8] Fagerland, Morten. y Hosmer, David. (2012). *A Genralized Hosmer-Lemeshow Goodness-of-fit Test for Multinomial Logistic Regression Models*. The Stata Journal. 12, Number 3, pp. 447–453.

[9] Fernández, Pando. y San Martín, R. (2004). *Regresión Logística Multinomial*. Valladolid: Actas de la Reunión de Modernización Forestal.

[10] Freese, Jeremy. y Long, Scott. (2000). *Test for the Multinomial Logit Model*.

[11] Hosmer, D. y Lemeshow, S. (2000). *Applied Logistic Regression*. New York. Wiley Interscience.

[12] Hsiao, Cheng. y Samll, Kenneth. (1983). *Multinomial Logit Specification Tests*. Econometric Research Program Research Memorandum No. 305.

[13] Jiménez, Ezequiel. (2013). *Introducción a la Econometría*. Valencia.

[14] Kunter, Michael. Nachtsheim, Christopher. Neterm, John. y Li, Wiliam. (2005). *Applied Linear Statistical Models*. New York: McGraw Hill Irwin.

[15] Long, Scott. (1997). *Regression Models for Categorical and Limited Dependent Variables*. California: Sage Publications.

[16] Montgomery, Douglas. Peck, Elizabeth. y Vining, Geoffrey. (2006). *Introducción al Análisis de Regresión Lineal*. CECSA. México.

[17] Palacios, Lino. y Arias, Adriana. (2014). *Adversidad psicosocial, psicopatología y funcionamiento en hermanos adolescentes en alto riesgo (HAR) con y sin trastorno por déficit de atención con hiperactividad (TDAH)*. México. Salud Mente vol.37 no.6.

[18] Powers, Daniel. y Xie, Yu. (1999). *Statistical Methods for Categorical Data Analysis*. Academic Press.

[19] Ryoo, Jihoon. (2008). *Proof of MLE for Multinomial Distribution*.

[20] Simonoff, Jeffrey. (2003). *Analyzing Categorical Data*. New York: Springer Texts in Statistics.

[21] Rodríguez, G. (2007). *Lecture Notes on Generalized Linear Models*. URL: http://data.princeton.edu/wws509/notes/

[22] Vijverberg, Wim. (2011). *Testing for IIA with the Hausman-McFadden Test*e Germany: IZA DP No. 5826.

[23] Ying, Chao. Lee, Kuk, Lida. y Ingersoll, Gary, M. (2002). *An Introduction to Logistic Regression Analysis and Reporting*. Indiana University-Boomington