INTRODUCCIÓN AL ESTUDIO DE MODELOS CATEGÓRICOS CON APLICACIONES

ELSA EDITH RIVERA ROSALES

T E S I S

Presentada como requisito parcial
para obtener el grado de
Maestro Profesional en
Estadística Aplicada

Universidad Autónoma Agraria
Antonio Narro
PROGRAMA DE GRADUADOS
Buenavista, Saltillo, Coahuila, México
Diciembre de 2008

Universidad Autónoma Agraria Antonio Narro

Dirección Postgrado

INTRODUCCIÓN AL ESTUDIO DE MODELOS CATEGÓRICOS

CON APLICACIONES

TESIS

Por:

ELSA EDITH RIVERA ROSALES

Elaborada bajo la supervisión del comité particular de asesoría y aprobada como

requisito parcial, para optar al grado de

MAESTRO PROFESIONAL

EN ESTADÍSTICA APLICADA

Comité Particular

Asesor principal: _____

M.C. Félix de Jesús Sánchez Pérez

Asesor: _____

Dr. Rolando Cavazos Cadena

Asesor: _____

Dr. Fernando Esquivel Bocanegra

_____

Dr. Jerónimo Landeros Flores
Director de Postgrado

Buenavista, Saltillo, Coahuila. Diciembre de 2008

AGRADECIMIENTOS

• A Dios.

• A mi familia, por el apoyo que me han dado siempre.

• Al M.C. Félix de Jesús Sánchez Pérez, por ser pieza fundamental en esta etapa de mi formación académica, por contar con Usted cuando lo necesitaba, por todo el apoyo que me brindó en la realización de esta tesis y por todo lo que ha hecho por mí: GRACIAS.

• Al Dr. José Antonio Díaz García, por estar al pendiente de mi desempeño, por todos sus consejos y por confiar en mí.

• Al Dr. Rolando Cavazos Cadena, por su dedicación, por su cariño y por su trato tan amable.

• Al Dr. Fernando Esquivel Bocanegra, por su ayuda y su intervención en este trabajo.

• A Laura Leticia Aguirre Padilla, por ser mi verdadera amiga, y por todos los momentos inolvidables que pasamos.

• A todos los maestros que a lo largo de mi camino compartieron conmigo sus conocimientos, que me son de gran utilidad, no solo en cuestiones estadísticas sino también en mi vida.

• A todos mis amigas y amigos, por estar a mi lado en todo momento, que aunque no mencione sus nombres saben que pueden contar conmigo siempre, así como sé que cuento con ustedes.

• A *Juan Patishtan Pérez*, gracias por TODO.

DEDICATORIA

A mi Gordo,
a mi Mamá
y a Claudia

COMPENDIO

Introducción al Estudio de Modelos Categóricos con Aplicaciones

POR

ELSA EDITH RIVERA ROSALES

MAESTRÍA PROFESIONAL
EN ESTADÍSTICA APLICADA

UNIVERSIDAD AUTÓNOMA AGRARIA ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA. DICIEMBRE DE 2008

M.C. Félix de Jesús Sánchez Pérez -Asesor-

Palabras clave: Datos Categóricos, Verosimilitud Máxima, Razón de Oportunidades, Modelo de Regresión Logística, Modelo Logit.

En el presente trabajo, se describen algunas herramientas básicas del análisis de datos categóricos y se mencionan ciertas aplicaciones. Lo anterior se justifica por el hecho de que actualmente muchos experimentos en las ciencias sociales, biológicas, económicas, etc., arrojan este tipo de datos por lo que enseguida se concentran algunos fundamentos de la teoría; y además, el análisis estadístico de este trabajo será apoyado con el programa R creando un paquete que permita realizar los cálculos y estimación de los parámetros del modelo, con mayor facilidad.

ABSTRACT

Introduction to the Study of Categorical Models with Applications

BY

ELSA EDITH RIVERA ROSALES

MASTER

APPLIED STATISTICS

UNIVERSIDAD AUTÓNOMA AGRARIA ANTONIO NARRO

BUENAVISTA, SALTILLO, COAHUILA. DECEMBER, 2008

M.C. Félix de Jesús Sánchez Pérez -Advisor-

Key Words: Categorical Data, Maximum Likelihood, Odds Ratio, Logistic Regression Model, Logit Model.

In this work, it describe some basic tools of the analysis of categorical data and certain applications are mentioned. The above-mentioned is justified for the fact that at the moment many experiments in the social, biological, economic sciences, among others, throw this type of data for that at once concentrate some foundations of the theory; and also, the statistical analysis of this work will be leaning with the program R creating a package that allows to carry out the calculations and estimate of the parameters of the models, with more facility.

Índice de Contenido

Índice de Tablas

Índice de Figuras

INTRODUCCIÓN

Generalmente, los procedimientos de prueba de hipótesis e intervalos de confianza se basan en muestras aleatorias tomadas de poblaciones normales. A menudo, éstos no funcionan bien para un conjunto de datos que disponen de: muy pocas observaciones, demasiadas variables, o bien, debido a la naturaleza que pueden tomar las variables de estudio cuyo comportamiento no es Gaussiano. Para solucionar tal problema, los métodos no parámetricos ofrecen una gran utilidad, ya que proporcionan de manera habitual mejoras considerables en comparación con los métodos parámetricos.

En particular, los métodos no parámetricos para datos categóricos han incrementado su uso especialmente en aplicaciones a ciencias de la salud, educación, sociales, biológicas, entre otras. La modelación de las variables consideradas en estas aplicaciones se conoce comúnmente con el nombre de modelos de elección discreta, dentro de la cual existe una amplia variedad de modelos. En concreto, según el número de alternativas incluidas en la variable respuesta, se distinguen los modelos

de respuesta simple frente a los denominados modelos de elección múltiple.
Según la función utilizada para la estimación de la probabilidad, existe el modelo de probabilidad lineal, el modelo Logit y el modelo Probit (Medina, 2003).

La importancia que tiene el estudio de variables categóricas, obliga a contar con literatura actualizada que concentre la teoría y aplicación con apoyo de paquetes estadísticos computacionales; con los cuales, y para el uso adecuado de éstos, implica saber elegir la técnica idónea para cada situación, aplicarla e interpretar correctamente los resultados. Por lo anterior, la presente tesis tiene como objetivo general el presentar algunos fundamentos y aplicaciones de los modelos categóricos, teniendo como apoyo el programa estadístico R en la estimación de los parámetros de los modelos.

La justificación a lo planteado se da a que existe poca literatura técnica que aborde los tópicos anteriormente señalados; y las referencias que se tienen, por lo general, están en otro idioma. Además, actualmente las encuestas socioeconómicas, industriales, ecológicas y biológicas necesitan herramientas computacionales que faciliten el análisis de datos categóricos. Otro punto esencial es que, la adquisición de paquetes estadísticos comerciales implica grandes costos. Afortunadamente, existe un programa estadístico con acceso gratuito denominado R; el cual es un conjunto integrado de programas para la manipulación de datos, cálculo y gráficas. Entonces, desarrollando programas en R enfocados a la modelación categórica incrementará la

facilidad del análisis estadístico.

La presentación de la tesis queda articulada en cuatro capítulos, el primero presenta las ideas fundamentales para el análisis de datos categóricos, en el segundo capítulo se introducen algunos conceptos de los modelos generalizados. Por su parte, el análisis de la regresión logística se incluye en el tercer capítulo; y por último, en el cuarto capítulo se establece la metodología para una generalización del modelo de regresión logística. Además, como una información adicional, se explica en el apéndice cómo se lleva a cabo la creación de un paquete en R.

CAPÍTULO 1

CONCEPTOS BÁSICOS EN EL ANÁLISIS DE DATOS CATEGÓRICOS

Sabiendo que las variables discretas pueden estar formadas por un número contable de alternativas que miden cualidades, ésta característica exige la codificación como paso previo a la modelación, proceso por el cual las alternativas de las variables se transforman en datos categóricos, por lo que éste capítulo introduce los conceptos básicos esenciales para su análisis.

1.1. Datos de Respuesta Categórica

Una variable cualitativa es tal que la escala de medida consiste en un conjunto de categorías, tomando dicha variable sólo un conjunto finito de valores, es decir sus valores son discretos. Por otra parte, existen modelos estadísticos que distinguen entre variable respuesta y explicatoria; por ejemplo, los modelos de regresión describen cómo la media de la variable respuesta, tal como el precio de venta de una casa, cambia de acuerdo a los valores de las variables explicatorias, como puede ser

la ubicación o la medida de la superficie del terreno.

La variable respuesta es algunas veces llamada variable dependiente denotada por Y, la variable explicatoria es llamada variable independiente, representada por X. Además, las variables explicatorias pueden ser categóricas o continuas, dentro del primer grupo están las variables nominales y ordinales, entendiendo por variables nominales a las que tienen desordenadas sus categorías, es decir no presentan un orden natural entre ellas, por ejemplo, el modo de transporte al trabajo (automóvil, bicicleta, autobús, metro o caminando), o la preferencia de café (caliente, frío o tibio), para éste tipo de variables el orden en que se listan las categorías es irrelevante y el análisis estadístico no depende de este orden. A diferencia de las variables nominales, las ordinales sí tienen un orden natural, por ejemplo, la respuesta a un tratamiento médico (excelente, bueno, regular o malo) o el desempeño del trabajo de un gobernador (malo, regular o bueno).

Por otra parte, a continuación se introduce la siguiente idea de variable categórica asociada a una partición $\mathcal{P}$ que se analiza en Rojas (2007).

Definición 1.1. Considere una variable aleatoria arbitraria T que toma valores en el espacio muestral Ω, y sea $\mathcal{P} = \{A_1, A_2, \ldots, A_N\}$ una partición de Ω en categorías $A_1, A_2, \ldots, A_N$. La variable categórica X introducida por T y la partición $\mathcal{P}$ toma valores en el conjunto de categorías $A_1, A_2, \ldots, A_N$, y se define como

$$X = A_i \Longleftrightarrow T \in A_i \, .$$

Además, el análisis de datos categóricos, requiere suposiciones sobre el mecanismo aleatorio por el cual se generan los datos, como lo es su distribución, es por tal motivo que enseguida se presenta la distribución binomial y multinomial que tienen un papel clave en el estudio de datos categóricos.

1.2. Distribución Binomial

Ciertas aplicaciones se refieren a un número fijo de n observaciones binarias, es decir con dos posibles resultados. Sea $y_1, y_2, \ldots, y_n$ que denota la respuesta para n ensayos independientes e idénticos tal que $P(Y_i = 1) = \pi$ y $P(Y_i = 0) = 1 - \pi$, donde se utilizará la notación de 1 y 0 para los resultados "éxito" y "fracaso", respectivamente. La media de los ensayos idénticos que tienen probabilidad de éxito π es la misma para cada ensayo. Así, el número total de éxitos, $\mathbf{Y} = \sum_{i=1}^{n} Y_i$, tiene *distribución binomial* con notación n y parámetro π, denotado por $\text{bin}(n, \pi)$. La función de probabilidad para los y posibles resultados de Y es

$$p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \qquad y = 0, 1, 2, \ldots, n$$

donde el coeficiente binomial $\binom{n}{y} = n!/[y!(n-y)!]$. Puesto que $E(Y_i) = E(Y_i^2) = 1 \times \pi + 0 \times (1 - \pi) = \pi$, entonces la esperanza y varianza son

$$E(Y_i) = \pi \quad \text{y} \quad \text{var}(Y_i) = \pi(1-\pi).$$

La distribución binomial para $\mathbf{Y} = \sum_i Y_i$ tiene media y varianza

$$\mu = E(Y) = n\pi \quad \text{y} \quad \sigma^2 = \text{var}(Y) = n\pi(1-\pi)$$

Para la demostración de la media, se presenta la siguiente definición.

Definición 1.2. Sea Y una variable aleatoria discreta con función de probabilidad $p(y)$. Entonces el valor esperado de Y, $E(Y)$ se define como

$$E(Y) = \sum_y yp(y). \tag{1.1}$$

De acuerdo con (1.1), se tiene que

$$E(Y) = \sum_y yp(y) = \sum_{y=0}^{n} y\binom{n}{y}\pi^y(1-\pi)^{n-y}$$

y puesto que el primer término de la suma anterior es cero, resulta

$$\begin{aligned} E(Y) &= \sum_{y=1}^{n} y\frac{n!}{(n-y)!y!}\pi^y(1-\pi)^{n-y} \\ &= \sum_{y=1}^{n} \frac{n!}{(n-y)!(y-1)!}\pi^y(1-\pi)^{n-y} \end{aligned}$$

Factorice $n\pi$ de cada término de la suma, y con $z = y - 1$

$$\begin{aligned} E(Y) &= n\pi\sum_{y=1}^{n} \frac{(n-1)!}{(n-y)!(y-1)!}\pi^{y-1}(1-\pi)^{n-y} \\ &= n\pi\sum_{z=0}^{n-1} \frac{(n-1)!}{(n-1-z)!z!}\pi^{z}(1-\pi)^{n-1-z} \\ &= n\pi\sum_{z=0}^{n-1} \binom{n-1}{z}\pi^{z}(1-\pi)^{n-1-z} \end{aligned}$$

Observe que $\sum_{z=0}^{n-1}\binom{n-1}{z}\pi^z(1-\pi)^{n-1-z}$ es la función de probabilidad binomial basada en $(n-1)$ ensayos, entonces $\sum_z p(z) = 1$, se deduce que

$$E(Y) = n\pi.$$

Para la demostración de la varianza es de gran ayuda la siguiente definición.

Definición 1.3. Si Y es una variable aleatoria discreta con función de probabilidad $p(y)$, entonces

$$V(Y) = \sigma^2 = E[(Y-\mu)^2] = E(Y^2) - \mu^2. \tag{1.2}$$

De (1.2) se sabe que $\sigma^2 = V(Y) = E(Y^2) - \mu^2$. Por lo tanto σ^2 se puede calcular determinando $E(Y^2)$. Así

$$\begin{aligned} E(Y^2) &= \sum_{y=0}^{n} y^2 p(y) \\ &= \sum_{y=0}^{n} y^2 \binom{n}{y} \pi^y (1-\pi)^{n-y} \\ &= \sum_{y=0}^{n} y^2 \frac{n!}{y!(n-y)!} \pi^y (1-\pi)^{n-y} \end{aligned}$$

Ahora bien, observe que

$$E[Y(Y-1)] = E(Y^2 - Y) = E(Y^2) - E(Y)$$

por lo tanto,

$$E(Y^2) = E[Y(Y-1)] + E(Y) = E[Y(Y-1)] + \mu.$$

En este caso

$$E[Y(Y-1)] = \sum_{y=0}^{n} y(y-1) \frac{n!}{y!(n-y)!} \pi^y (1-\pi)^{n-y}$$

El primer y segundo término de la suma anterior son cero (cuando $y = 0$ y $y = 1$). Por consiguiente,

$$E[Y(Y-1)] = \sum_{y=2}^{n} \frac{n!}{(y-2)!(n-y)!} \pi^y (1-\pi)^{n-y}$$

Los téminos de la última expresión se asemejan mucho a las probabilidades binomiales. Escribiendo el factor constante $n(n-1)\pi^2$, fuera de la suma y si $z = y - 2$,

se obtiene

$$\begin{aligned} E[Y(Y-1)] &= n(n-1)\pi^2 \sum_{y=2}^{n} \frac{(n-2)!}{(y-2)!(n-y)!}\pi^{y-2}(1-\pi)^{n-y} \\ &= n(n-1)\pi^2 \sum_{z=0}^{n-2} \frac{(n-2)!}{z!(n-2-z)!}\pi^{z}(1-\pi)^{n-2-z} \\ &= n(n-1)\pi^2 \sum_{z=0}^{n-2} \binom{n-2}{z}\pi^{z}(1-\pi)^{n-2-z} \end{aligned}$$

Observe que de nuevo $p(z) = \binom{n-2}{z}\pi^z(1-\pi)^{n-2-z}$ es la función de probabilidad binomial basada en (n - 2) ensayos. Así $\sum_{z=0}^{n-2} p(z) = 1$ y

$$E[Y(Y-1)] = n(n-1)\pi^2.$$

Resultando

$$\begin{aligned} E(Y^2) &= E[Y(Y-1)] + \mu \\ &= n(n-1)\pi^2 + n\pi \end{aligned}$$

y

$$\begin{aligned} \sigma^2 &= E(Y^2) - \mu^2 \\ &= n(n-1)\pi^2 + n\pi - n^2\pi^2 \\ &= n\pi[(n-1)\pi + 1 - n\pi] \\ &= n\pi(1-\pi). \end{aligned}$$

Para tener una idea más clara de ésta distribución, se analiza el siguiente ejemplo.

Ejemplo 1.1. Suponga que un lote de 5000 fusibles contiene un 5 % de defectuosos. Si se toma una muestra de cinco fusibles, encuentre la probabilidad de encontrar por lo menos uno defectuoso.

Solución. Suponga que N es la variable aleatoria que denota el número de fusibles defectuosos observados y puesto que el lote tiene una distribución binomial, es decir bin(5, 0.05). Así

$$\begin{aligned} P(N \geq 1) &= 1 - P(N = 0) \\ &= 1 - \binom{5}{0}\pi^0(1-\pi)^{5-0} \\ &= 1 - 1(0.05)^0(0.95)^5 \\ &= 1 - 0.773780937 \\ &= 0.226219062 \end{aligned}$$

Entonces, la probabilidad de encontrar por lo menos un fusible defectuoso es 0.2262.

De aquí en adelante se trabajará con el paquete "Categoricos"creado en R. Primero se tendrá que cargar dicho paquete, para hacer esto se tiene que abrir la consola de R, ir a la opción "Paquetes", enseguida seleccionar "Cargar paquete" y elegir el paquete "Categoricos", esto permitirá que las funciones creadas para analizar los datos estén listas para usar.

Luego, para activar el paquete se tendrá que escribir en la consola de R la instrucción: library(Categoricos) que indica que el paquete nombrado estará disponible para utilizar.

Para los cálculos en R, de la distribución binomial, se tendrá que proporcionar: y (valor de la variable aleatoria), n (número de ensayos) y π (probabilidad de éxito en cada ensayo), en la función "bynomial"del paquete "Categoricos".

1.3. Distribución Multinomial

Enseguida se considera una variable aleatoria discreta importante de mayor dimensión, que puede ser considerada como una generalización de la distribución binomial.

Suponga un experimento en el que cada uno de los n ensayos independientes e idénticos, puede tener un resultado en cualquiera de c categorías mutuamente excluyentes. Sea $y_{ij} = 1$ si el ensayo i tiene un resultado en la categoría j y $y_{ij} = 0$ de otra manera, en donde $i = 1, 2, \ldots, n$ y $j = 1, 2, \ldots, c$. Entonces $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ic})$ representa un ensayo multinomial, con $\sum_j y_{ij} = 1$ para todo $i = 1, 2, \ldots, n$; por ejemplo, $(0, 0, 1, 0)$ denota el resultado en la categoría tres de cuatro posibles, note que y_{ic} es redundante, porque es linealmente dependiente de los otros. Sea $n_j = \sum_i y_{ij}$ que denota el número de ensayos que tienen resultado en la categoría j. El cálculo de $n_1, n_2, \ldots, n_c$ tiene distribución multinomial y sea $\pi_j = P(Y_{ij} = 1)$ que denota la probabilidad de un resultado en la categoría j para cada ensayo. La función de probabilidad multinomial es

$$p(n_1, n_2, \ldots, n_c) = \left(\frac{n!}{n_1! n_2! \ldots n_c!} \right) \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$$

en donde $\sum_j n_j = n$; observe que $n_c = n - (n_1 + \ldots + n_{c-1})$.

Para la distribución multinomial, enseguida se tienen la esperanza, varianza y covariaza, respectivamente.

$$E(n_j) = n\pi_j, \qquad \text{var}(n_j) = n\pi_j(1 - \pi_j), \qquad \text{cov}(n_j, n_k) = -n\pi_j\pi_k$$

En la demostración de la esperanza y varianza multinomiales, se puede utilizar la

distribución marginal de n_j para calcular la media y la varianza. Recuerde que n_j puede interpretarse como el número de ensayos que caen en la categoría j. De modo que n_j tiene una distribución de probabilidad marginal binomial. Por consiguiente la media y varianza multinomiales son,

$$E(n_j) = n\pi_j \quad \text{y} \quad V(n_j) = n\pi_j(1-\pi_j).$$

Ahora la covarianza se demuestra con ayuda de la siguiente definición

Definición 1.4. Sean $Y_1, Y_2, \ldots, Y_n$ y $X_1, X_2, \ldots, X_m$ variables aleatorias con $E(Y_i) = \mu_i$ y $E(X_j) = \xi_j$. Se define ahora

$$U_1 = \sum_{i=1}^{n} a_i Y_i \quad \text{y} \quad U_2 = \sum_{j=1}^{m} b_j X_j$$

para las constantes $a_1, a_2, \ldots, a_n$ y $b_1, b_2, \ldots, b_m$. Así, las siguientes proposiciones se cumplen:

a)

$$E(U_1) = \sum_{i=1}^{n} a_i \mu_i$$

b)

$$V(U_1) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2 \sum \sum_{i<j} a_i a_j Cov(Y_i, Y_j)$$

donde la doble suma está sobre todos sus pares (i, j) con $i < j$

c)

$$Cov(U_1, U_2) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j Cov(Y_i, X_j).$$

Demostración de las proposiciones

a)

$$\begin{aligned} E(U_1) &= E\left(\sum_{i=1}^{n} a_i Y_i\right) \\ &= \sum_{i=1}^{n} E(a_i Y_i) \\ &= \sum_{i=1}^{n} a_i E(Y_i) \\ &= \sum_{i=1}^{n} a_i \mu_i \end{aligned}$$

b)

$$\begin{aligned} V(U_1) &= V\left(\sum_{i=1}^{n} a_i Y_i\right) \\ &= E[U_1 - E(U_1)]^2 = E\left[\sum_{i=1}^{n} a_i Y_i - \sum_{i=1}^{n} a_i \mu_i\right]^2 \\ &= E\left[\sum_{i=1}^{n} a_i (Y_i - \mu_i)\right]^2 \\ &= E\left[\sum_{i=1}^{n} a_i^2 (Y_i - \mu_i)^2 + \sum\sum_{i \neq j} a_i a_j (Y_i - \mu_i)(Y_j - \mu_j)\right] \\ &= \sum_{i=1}^{n} a_i^2 E(Y_i - \mu_i)^2 + \sum\sum_{i \neq j} a_i a_j E[(Y_i - \mu_i)(Y_j - \mu_j)] \end{aligned}$$

Mediante la definición de la varianza y covarianza se obtiene

$$V(U_1) = \sum_{i=1}^{n} a_i^2 V(Y_i) + \sum\sum_{i \neq j} a_i a_j cov(Y_i, Y_j)$$

como $cov(Y_i, Y_j) = cov(Y_j, Y_i)$, se puede escribir

$$V(U_1) = \sum_{i=1}^{n} a_i^2 V(Y_i) + 2\sum\sum_{i < j} a_i a_j cov(Y_i, Y_j)$$

c) Se tiene

$$\begin{aligned}
cov(U_1, U_2) &= E\{[U_1 - E(U_1)][U_2 - E(U_2)]\} \\
&= E\left[\left(\sum_{i=1}^{n} a_i Y_i - \sum_{i=1}^{n} a_i \mu_i\right)\left(\sum_{j=1}^{m} b_j X_j - \sum_{j=1}^{m} b_j \xi_j\right)\right] \\
&= E\left\{\left[\sum_{i=1}^{n} a_i (Y_i - \mu_i)\right]\left[\sum_{j=1}^{m} b_j (X_j - \xi_j)\right]\right\} \\
&= E\left[\sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j (Y_i - \mu_i)(X_j - \xi_j)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j E[(Y_i - \mu_i)(X_j - \xi_j)] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j cov(Y_i, X_j).
\end{aligned}$$

Ahora, considere el experimento multinomial como una serie de n ensayos independientes y defina, para $s \neq t$,

$$U_i = \begin{cases} 1, & \text{si el resultado del ensayo } i \text{ cae en la clase (o celda) } s \\ 0, & \text{en caso contrario} \end{cases}$$

y

$$W_i = \begin{cases} 1, & \text{si el resultado del ensayo } i \text{ cae en la clase } t \\ 0, & \text{en caso contrario} \end{cases}$$

Así, $Y_s = \sum_{i=1}^{n} U_i$ y $Y_t = \sum_{j=1}^{n} W_j$.

(Ya que $U_i = 1$ ó 0 en caso de que el i-ésimo ensayo haya dado o no como resultado la clase s, respectivamente; Y_s es sencillamente la suma de una serie de ceros y unos. Una interpretación semejante se puede hacer con Y_t). Advierta que U_i y W_i no pueden ser simultáneamente iguales a 1 (el resultado en el ensayo i no puede pertenecer simultáneamente a las clases s y t). Por lo tanto, el producto $U_i W_i$ siempre es igual

a cero y $E(U_iW_i) = 0$. Los siguientes resultados permiten evaluar $Cov(Y_s, Y_t)$:

$$E(U_i) = \pi_s \quad E(W_j) = \pi_t$$

$Cov(U_i, W_j) = 0$, si $i \neq j$, puesto que los ensayos son independientes.

$Cov(U_i, W_i) = E(U_iW_i) - E(U_i)E(W_i) = 0 - \pi_s\pi_t$.

Así, de acuerdo con la definición (1.4), se tiene

$$\begin{aligned} Cov(Y_s, Y_t) &= \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(U_i, W_j) \\ &= \sum_{i=1}^{n} Cov(U_i, W_i) + \sum\sum_{i\neq j} Cov(U_i, W_j) \\ &= \sum_{i=1}^{n}(-\pi_s\pi_t) \\ &= -n\pi_s\pi_t \end{aligned}$$

La covarianza es negativa, puesto que un gran número de resultados en la celda s hará que el número de resultados en la celda t sea menor.

Se aborda la distribución multinomial con el ejemplo siguiente.

Ejemplo 1.2. De acuerdo con las cifras obtenidas en un censo reciente, las fracciones de población adulta (individuos mayores de 18 años de edad) de Estados Unidos relacionadas con cinco categorías de edad que aparecen en la tabla (1.1).
Si se eligen cinco adultos aleatoriamente, calcule la probabilidad de que la muestra incluya a un individuo entre 18 y 24 años, dos entre 25 y 34 años y dos entre 45 y 64 años.

Solución. Se numeran las cinco categorías de edad como 1, 2, 3, 4 y 5, de arriba hacia abajo y suponga que las proporciones que aparecen en la tabla son

Edad	Proporción
18 - 24	0.18
25 - 34	0.23
35 - 44	0.16
45 - 64	0.27
65 $\uparrow$	0.16

Tabla 1.1: Categorías de Edad del Estudio de Estados Unidos

las probabilidades relacionadas con cada categoría. Así se determina la distribución multinomial siguiente

$$p(n_1, n_2, n_3, n_4, n_5) = \frac{n!}{n_1!n_2!n_3!n_4!n_5!}\pi_1^{n_1}\pi_2^{n_2}\pi_3^{n_3}\pi_4^{n_4}\pi_5^{n_5}$$

con n = 5 (que es la suma de los resultados para las diferentes categorías) y $n_1 = 1$, $n_2 = 2$, $n_3 = 0$, $n_4 = 2$ y $n_5 = 0$. Ahora sustituyendo dichos valores en la fórmula de la función de probabilidad se obtiene

$$\begin{aligned} p(1,2,0,2,0) &= \frac{5!}{1!2!0!2!0!}(0.18)^1(0.23)^2(0.16)^0(0.27)^2(0.16)^0 \\ &= 30(0.18)(0.23)^2(0.27)^2 \\ &= 0.020824614 \end{aligned}$$

Por lo cual la muestra puede incluir el número de las personas deseadas en las distintas categorías con una probabilidad del 2.082 por ciento.

Para llevar a cabo los cálculos de esta distribución se usará la función "multinomial"del paquete "Categoricos"donde se proporcionan: $n_i's$ (vector que indica el

número de ensayos que caen en las c categorías), n (número total de ensayos) y π (vector que especifica la probabilidad para las c categorías).

A continuación se plantea como se realiza la estimación de parámetros en una variable categórica usando verosimilitud máxima.

1.4. Estimación de la Verosimilitud Máxima

El método de verosimilitud máxima, elige como estimaciones los valores de los parámetros que maximizan la función de probabilidad conjunta de la muestra observada.

La función de verosimilitud máxima de n variables aleatorias $X_1, X_2, \ldots, X_n$ se define como

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

Definición 1.5. Método de Verosimilitud Máxima: Suponga que la función de verosimilitud depende de k parámetros $\theta_1, \theta_2, \ldots, \theta_k$. Elija como estimaciones aquellos valores de los parámetros que maximizan la verosimilitud $L(\theta_1, \theta_2, \ldots, \theta_k | y_1, y_2, \ldots, y_n)$.

Hay que destacar el hecho de que la función de verosimilitud es una función de los parámetros $\theta_1, \theta_2, \ldots, \theta_k$ que a menudo se expresa como $L(\theta_1, \theta_2, \ldots, \theta_k)$; usualmente, para abreviar, se hace referencia a la verosimilitud máxima como ML, por sus siglas en ingles.

Para ilustrar la estimación de la verosimilitud máxima analizada anteriormente se tiene el siguiente ejemplo.

Ejemplo 1.3. Un experimento binomial que consta de n ensayos genera los resultados $y_1, y_2, \ldots, y_n$, donde $y_i = 1$ si el i-ésimo ensayo es un éxito y $y_i = 0$ si es un fracaso. Encuentre el estimador de verosimilitud máxima de π, la probabilidad de éxito.

Solución. La verosimilitud de la muestra observada es la probabilidad de observar $y_1, y_2, \ldots, y_n$. Por consiguiente,

$$\begin{aligned} L(\pi) &= L(\pi|y_1, y_2, \ldots, y_n) \\ &= \pi^y(1-\pi)^{n-y} \qquad \text{donde } y = \sum_{i=1}^{n} y_i \end{aligned}$$

Se quiere determinar el valor de π que maximiza $L(\pi)$. Si $y = 0$, $L(\pi) = (1-\pi)^n$, y $L(\pi)$ se maximiza cuando $\pi = 0$. De la misma manera, si $y = n$, $L(\pi) = \pi^n$, y $L(\pi)$ se maximiza cuando $\pi = 1$. Si $y = 1, 2, \ldots, n-1$, entonces $L(\pi) = \pi^y(1-\pi)^{n-y}$ es cero en $\pi = 0$ y $\pi = 1$, y es continua para valores de π entre 0 y 1. Por consiguiente, para $1, 2, \ldots, n-1$, se puede determinar el valor de π que maximiza $L(\pi)$ al igualar la derivada $d\,L(\pi)/d\,\pi$ a cero y despejar π. Para ver esto, note que $ln[L(\pi)]$ es una función monótona creciente de $L(\pi)$. Entonces, tanto $ln[L(\pi)]$ como $L(\pi)$ se maximizan en el mismo valor de π. Como $L(\pi)$ es el producto de funciones de π, y obtener la derivada de productos puede resultar un tanto laborioso, es más fácil encontrar el valor de π que maximiza $ln[L(\pi)]$. Así, se tiene que

$$ln[L(\pi)] = ln[\pi^y(1-\pi)^{n-y}] = y\,ln\pi + (n-y)\,ln(1-\pi)$$

Si $y = 1, 2, \ldots, n-1$, la derivada de $ln[L(\pi)]$ respecto a π, es

$$\begin{aligned} \frac{d\, ln[L(\pi)]}{d\pi} &= y\left(\frac{1}{\pi}\right) + (n-y)\left(\frac{-1}{1-\pi}\right) \\ &= \frac{y}{\pi} - \frac{n-y}{1-\pi} \end{aligned}$$

Entonces la ecuación de verosimilitud es

$$\begin{aligned} \frac{y}{\pi} - \frac{n-y}{1-\pi} &= 0 \\ \frac{y(1-\pi) - \pi(n-y)}{\pi(1-\pi)} &= 0 \\ \frac{y - n\pi}{\pi(1-\pi)} &= 0 \\ y - n\pi &= 0 \end{aligned}$$

cuya solución es $\pi = y/n$.

Puesto que $L(\pi)$ se maximiza en $\pi = 0$ cuando $y = 0$, en $\pi = 1$ cuando $y = n$ y en $\pi = y/n$ cuando $y = 1, 2, \ldots, n-1$, cualquiera que sea el valor observado de y, $L(\pi)$ se maximiza cuando $\pi = y/n$. El estimador de verosimilitud máxima que se obtiene al sustituir y por la variable aleatoria Y para generalizar la fórmula de $\hat{\pi}$ es, $\hat{\pi} = Y/n$, que es la fracción de éxitos del número total de pruebas de n.

1.5. Pruebas de Independencia

Enseguida se muestra cómo se prueba la hipótesis nula H_0 de que las probabilidades de las celdas de la tabla sea igual a un valor fijo $\{\pi_{ij}\}$, es decir $H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$. Para una muestra de tamaño n con datos $\{n_{ij}\}$, los valores $\{\mu_{ij} = n\pi_{\cdot j}\}$ son llamados frecuencias esperadas y representan las esperanzas $\{E(n_{ij})\}$ cuando H_0 es verdadera.

1.5.1. Estadístico de Pearson

Un procedimiento para probar $H_0 : \pi_{ij} = \pi_{i.}\pi_{.j}$ se puede llevar a cabo usando el estadístico de Pearson, que se determina mediante

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}, \qquad i = 1, \ldots, r; j = 1, \ldots, c$$

Para un tamaño de muestra fijo, observe que si la frecuencia observada n_{ij} es similar a la correspondiente frecuencia esperada μ_{ij}, entonces $(n_{ij} - \mu_{ij})^2$ no será grande, y por lo tanto X^2 tomará valores moderados, de esta manera valores grandes de X^2 son evidencia contra H_0. El estadístico X^2 tiene aproximadamente una distribución ji-cuadrada para tamaños de muestra grandes, ésto es dificil al especificar que las medias sean "grandes"pero con $\{\mu_{ij} \geq 5\}$ es suficiente. Bajo H_0, el estadístico X^2 tiene una distribución ji-cuadrada con (r - 1)(c - 1) grados de libertad, donde r denota el número de filas en la tabla y la letra c se refiere al número de columnas. La hipótesis H_0 se rechaza al nivel de significancia α si y sólo si $X^2 > \chi^2_{(r-1)(c-1),\alpha}$. Más adelante se analiza este estadístico con un ejemplo.

1.5.2. Razón de Verosimilitud

Un procedimiento alternativo para probar H_0 es el método de la razón de verosimilitud. La prueba determina el valor del parámetro que maximiza la función de verosimilitud bajo la suposición de que H_0 sea verdadera, además establece el valor que maximiza dicha función bajo la condición más general que H_0 pueda o no ser cierta.

La fórmula del estadístico es

$$G^2 = 2\sum_{i,j} n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right),$$

donde G^2 es llamado estadístico ji-cuadrado de la razón de verosimilitud.

Cuando H_0 es verdadera y la muestra es grande, los dos estadísticos tienen la misma distribución ji-cuadrada y sus valores numéricos son similares. Como en el caso anterior, bajo H_0 este estadístico tiene una distribución ji-cuadrada con (r - 1)(c - 1) grados de libertad, y la hipótesis H_0 se rechaza al nivel de significancia α si y sólo si $G^2 > \chi^2_{(r-1)(c-1),\alpha}$.

Los dos estadísticos para probar el supuesto de independencia tienen un nivel de significancia aproximadamente igual a α, y la aproximación es mejor conforme el número de datos crece (Rojas, 2007). En el ejemplo (1.4) se discute el estadístico G^2, así como el estadístico X^2.

1.6. Medidas de Asociación

El interés en estas medidas es el determinar si existe un cambio en la variable respuesta Y para los distintos valores que puede tomar la variable X. A continuación se presentan tres de ellas.

1.6.1. Diferencia de Proporciones

La diferencia de proporciones $\pi_1 - \pi_2$ compara la probabilidad de éxito en las dos filas de la tabla de datos. Sea p_1 y p_2 que denotan las proporciones muestrales de éxito para las dos filas, respectivamente. La diferencia muestral $p_1 - p_2$ es una estimación de $\pi_1 - \pi_2$.

Entonces la diferencia denotada por δ es

$$\delta = \pi_1 - \pi_2$$

Su estimador de verosimilitud máxima es

$$\hat{\delta} = p_1 - p_2$$

y su error estándar asintótico denotado por ASE es

$$ASE(\hat{\delta}) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Entonces el intervalo de confianza para $\pi_1 - \pi_2$ es

$$\hat{\delta} \pm z_{\alpha/2} ASE(\hat{\delta})$$

donde $z_{\alpha/2}$ es el percentil de orden $\alpha/2$ en la distribución normal estándar.

1.6.2. Riesgo Relativo

El riesgo relativo es definido como

$$\rho = \frac{\pi_1}{\pi_2}$$

Teniendo como estimador de verosimilitud a

$$\hat{\rho} = \frac{p_1}{p_2}$$

Como la razón de oportunidades converge más rápido a la normalidad en la escala logarítmica, se tiene que el error estándar asintótico de $\log(\hat{\rho})$ es

$$ASE(\log \hat{\rho}) = \sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}}$$

El intervalo de confianza es

$$\log \hat{\rho} \pm z_{\alpha/2} ASE(\log \hat{\rho})$$

donde $z_{\alpha/2}$ es el percentil de orden $\alpha/2$ en la distribución normal estándar.

1.6.3. Razón de Oportunidades

La razón de oportunidades denotada por θ en las dos filas (tabla 2×2) es

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

Su estimador de verosimilitud máxima es

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

Como el logarítmo de $\hat{\theta}$ converge más rápido a la normalidad, se tiene que el error estándar asintótico para $\log \hat{\theta}$ es

$$ASE(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

El intervalo de confianza es

$$\log \hat{\theta} \pm z_{\alpha/2} ASE(\log \hat{\theta})$$

donde $z_{\alpha/2}$ es el percentil de orden $\alpha/2$ en la distribución normal estándar.

Para aplicar la asociación entre dos variables categóricas, enseguida se analiza un ejemplo propuesto en Agresti, (2002).

Ejemplo 1.4. En un estudio realizado por doctores británicos se notó que las muertes anuales por cáncer pulmonar eran 140 para los fumadores y 10 para los no fumadores. Las personas que fallecieron por enfermedad del corazón son 669 para los fumadores y 413 para los no fumadores.

Describa la asociación de fumadores con cáncer pulmonar y enfermedad del corazón, usando a) la diferencia de proporciones, b) riesgo relativo y c) la razón de oportunidades, e interprete, además pruebe la independencia entre la causa de muerte y el grupo al que pertenece.

	Muertes por	
Grupo	Cáncer Pulmonar	Enfermedad del Corazón
Fumadores	140	669
No Fumadores	10	413

Solución. Las proporciones de muerte debido al cáncer pulmonar y enfermedad del corazón en la muestra del grupo de fumadores, denotadas por p_1 y p_2, respectivamente son $p_1 = 140/150 = 0.9333$ y $p_2 = 669/1082 = 0.6182$.

a) La diferencia de proporciones estimada es

$$\begin{aligned} \hat{\delta} &= p_1 - p_2 = 0.9333 - 0.6182 \\ &= 0.3151 \end{aligned}$$

Interpretación: Este valor indica que, en la muestra analizada, la proporción de muerte por cáncer pulmonar es mayor que una muerte por enfermedad del corazón. El intervalo de confianza correspondiente es

$$\hat{\delta} \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

donde n_1 y n_2 son los tamaños de muestra en los grupos de cáncer pulmonar y enfermedad del corazón, respectivamente. Efectuando los cálculos se obtiene que

$$\begin{aligned} 0.3151 - 1.96(0.025162443) \leq \ & \delta \leq 0.3151 + 1.96(0.025162443) \\ 0.265781611 \leq \ & \delta \leq 0.364418388 \end{aligned}$$

es el intervalo para δ con un 95 % de confianza, puesto que el intervalo contiene valores positivos, se puede establecer que las personas son más propensas a morir debido al cáncer pulmonar.

b) El riesgo relativo estimado es

$$\begin{aligned} \hat{\rho} &= \frac{p_1}{p_2} = \frac{0.9333}{0.6182} \\ &= 1.509705597 \end{aligned}$$

Interpretación: $\hat{\rho}$ indica que el riesgo de muertes por cáncer pulmonar es casi 1.5 veces más que morir por alguna enfermedad del corazón.

El intervalo de confianza correspondiente, con $\log(\rho) = \log(1.5097) = 0.411910955$, es

$$\log(\hat{\rho}) \pm z_{\alpha/2}\sqrt{\frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}}$$

efectuando las operaciones se tiene que

$$\log(1.5097) - 1.96(0.032361086) \leq \log(\rho) \leq \log(1.5097) + 1.96(0.032361086)$$

$$0.348483227 \leq \log(\rho) \leq 1.573127729$$

y por lo tanto escribiendo en forma exponencial cada extremo, se tiene que

$$e^{0.3484} \leq \rho \leq e^{0.4753}$$

$$1.4167 \leq \rho \leq 1.6084$$

es el intervalo para ρ con un 95 % de confianza, debido a que el intervalo contiene valores mayores que uno, se puede establecer que la causa principal de muerte es el cáncer pulmonar.

c) La razón de oportunidades estimada es

$$\begin{aligned} \hat{\theta} &= \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{13.99250375}{1.619172342} \\ &= 8.641763072 \end{aligned}$$

Interpretación: La oportunidad de muerte debido al cáncer pulmonar será 8.64 veces mayor que morir por una enfermedad del corazón, para las personas fumadoras.

El intervalo de confianza para la razón de oportunidades es

$$\log\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

donde n_{ij} son las frecuencias en la tabla de datos y sabiendo que $\log(\hat{\theta}) = \log(8.6417) = 2.156599323$, se utiliza un intervalo de confianza de 95 %, de forma que $z_{\alpha/2} = 1.96$, los extremos del intervalo para $\log(\theta)$ son

$$2.1565 - 1.96(0.333255056) \leq \log(\theta) \leq 2.1565 + 1.96(0.333255056)$$

$$1.50332009 \leq \log(\theta) \leq 2.80967991$$

y por lo tanto expresando en forma exponencial cada término

$$e^{1.5033} \leq \quad \theta \quad \leq e^{2.8096}$$

$$4.4965 \leq \quad \theta \quad \leq 16.6032$$

que es el intervalo para θ con un 95 % de confianza, note que este intervalo contiene puntos mayores que 1, de manera que no puede concluirse que la razón de oportunidades de morir por cáncer pulmonar sea menor que morir por una enfermedad del corazón.

Ahora para probar la independencia entre la causa de muerte y el grupo al que se pertenece se calculan los estadísticos X^2 y G^2. Para calcular tales estadísticos es necesario sumar los valores de las filas y columnas y obtener el gran total.

	Cáncer Pulmonar	Enfermedad del Corazón	Total fila
	$n_{11} = 140$	$n_{12} = 669$	$809 = n_{1\cdot}$
	$n_{21} = 10$	$n_{22} = 413$	$423 = n_{2\cdot}$
Total columna	$n_{\cdot 1} = 150$	$n_{\cdot 2} = 1082$	1232 : Gran total $= n$

Se obtienen las frecuencias esperadas

$$\mu_{11} = \frac{809(150)}{1232} = 98.49837662 \qquad \mu_{12} = \frac{809(1082)}{1232} = 710.5016234$$

$$\mu_{21} = \frac{423(150)}{1232} = 51.50162338 \qquad \mu_{22} = \frac{423(1082)}{1232} = 371.4983766$$

Cálculo del estadístico de Pearson X^2

$$\begin{aligned} X_{11} &= \frac{(140-98.4983)^2}{98.4983} = 17.48650589 \\ X_{12} &= \frac{(669-710.5016)^2}{710.5016} = 2.42417864 \\ X_{21} &= \frac{(10-51.5016)^2}{51.5016} = 33.44328725 \\ X_{22} &= \frac{(413-371.4983)^2}{371.4983} = 4.636336432 \end{aligned}$$

Sumando cada uno de los términos anteriores se obtiene el estadístico X^2

$$\begin{aligned} X^2 &= 17.48650589 + 2.42417864 + 33.44328725 + 4.636336432 \\ &= 57.99030821 \end{aligned}$$

En la presente tabla con dos filas y dos columnas, X^2 tiene aproximadamente una distribución ji-cuadrada con un grado de libertad, y el valor crítico $\chi^2_{1,0.95} = 0.003$ es prácticamente cero. Por lo tanto, el supuesto de independencia $H_0 : \pi_{ij} = \pi_i\pi_j$ entre la causa de muerte y el grupo al que pertenece la persona, se rechaza ($X^2 > \chi^2_{(1-r)(1-c),\alpha}$).

Cálculo del estadístico G^2

$$\begin{aligned} G_{11} &= 2(140)\log\left(\frac{140}{98.4983}\right) = 98.44887737 \\ G_{12} &= 2(669)\log\left(\frac{669}{710.5016}\right) = -80.53039244 \\ G_{21} &= 2(10)\log\left(\frac{10}{51.5016}\right) = -32.78055564 \\ G_{22} &= 2(413)\log\left(\frac{413}{371.4983}\right) = 87.47612981 \end{aligned}$$

Sumando los términos anteriores se obtiene el estadístico

$$\begin{aligned} G^2 &= 98.44887737 - 80.53039244 - 32.78055564 + 87.47612981 \\ &= 72.6140591 \end{aligned}$$

De manera similar, el estadístico G^2 tiene aproximadamente una distribución ji-cuadrada con un grado de libertad, y el valor crítico $\chi^2_{1,0.95} = 0.003$ es casi cero. Por lo que, se rechaza el supuesto de independencia entre las variables causa de muerte y grupo al que pertenece la persona.

Para obtener las medidas de asociación junto con sus intervalos de confianza y los estadísticos de prueba X^2 y G^2, en R se proporciona la matriz de los datos de la tabla, es decir las frecuencias de los datos en la función "medasoc"del paquete "Categoricos".

CAPÍTULO 2

INTRODUCCIÓN A LOS MODELOS LINEALES GENERALIZADOS

En el capítulo anterior se presentaron métodos para analizar la asociación entre dos variables, sin embargo muchos estudios tienen varias variables explicatorias, la meta usualmente es describir los efectos en las variables respuesta (Fahrmeir and Tutz, 2001). En este capítulo se estudia la manera de usar los modelos lineales generalizados como la base del análisis al modelar variables respuesta categóricas, se estudia éste tema ya que el modelo logístico que se analiza más adelante es un caso particular en los modelos lineales generalizados.

Dado que el modelo lineal general es un buen punto de partida para el estudio de los modelos lineales generalizados, se empieza este capítulo con una somera revisión de los principales aspectos del modelo lineal general.

2.1. Modelo Lineal General

El modelo lineal general surge por la necesidad de expresar en forma cuantitativa relaciones entre un conjunto de variables, en la que una de ellas se denomina variable respuesta o dependiente y las restantes son llamadas variables explicativas o independientes.

Sea Y una variable aleatoria cuya función de distribución de probabilidad pertenece a una familia de distribuciones de probabilidades **H**, y es explicada por un conjunto de variables $X_1, X_2, \ldots, X_k$ las cuales son fijadas antes de conocer a Y. Así la esperanza condicional de Y dado X se expresa como

$$E(Y/X_1, X_2, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k = \mu$$

Luego, si se extrae una muestra aleatoria de tamaño n $\{(y_i, x_{i1}, \ldots, x_{ik}) : i = 1, 2, \ldots, n\}$, de una población en la cual la variable respuesta Y, y las variables independientes se relacionan linealmente. Cada observación de la muestra puede ser expresada como

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i, \qquad i = 1, \ldots, n \tag{2.1}$$

en la ecuación (2.1), el término ε_i es una perturbación aleatoria no observable denominada error aleatorio, el cual tiene el supuesto de una esperanza cero y varianza σ^2. Utilizando notación matricial, se puede expresar (2.1) como

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{2.2}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

donde $\mathbf{Y}$ es un vector de variables aleatorias observables, denominado vector respuesta de orden n, $\mathbf{X}$ es la matriz de variables independientes de orden $n \times (k+1)$ y β el vector de parámetros desconocidos de orden $(k+1)$. El vector de respuestas $\mathbf{Y}$ de la expresión (2.2) está formado por dos componentes, uno sistemático y otro aleatorio. El primer componente constituido por la combinación lineal $X\beta$, llamado predictor lineal, el cual es representado como

$$\eta = \mathbf{X}\beta.$$

El segundo componente, formado por el vector aleatorio ε, con elementos independientes entre sí, es caracterizado por una distribución $h \in \mathbf{H}$ con vector de esperanzas $\mathbf{0}$ y matriz de covarianza $\sigma^2 I$. Por otro lado, calculando la esperanza de Y en (2.2) se tiene que

$$E(\mathbf{Y}) = \mathbf{X}\beta = \mu.$$

Ahora se presentan los modelos lineales generalizados extendiendo la teoría de los modelos lineales, algunos modelos que forman parte de esta familia son los loglineales, logit, probit y logístico.

2.2. Modelo Lineal Generalizado

Tres componentes especifican a un modelo lineal generalizado: 1) El *componente aleatorio* identifica la variable respuesta Y y su distribución de probabilidad, 2) El *componente sistemático* especifica las variables explicatorias usadas como predictores en el modelo, y 3) La *función enlace* describe la relación funcional entre el componente sistemático y el valor esperado $E(Y)$ del componente aleatorio. El modelo lineal generalizado relaciona una función de la media para las variables explicatorias a través de una ecuación que tiene forma lineal.

2.2.1. Componentes del Modelo Lineal Generalizado

El *componente aleatorio* consite en seleccionar una distribución de probabilidad para una muestra de tamaño N. Sea Y una variable respuesta con observaciones independientes $(y_1, \ldots, y_N)$ de una distribución de la familia exponencial. Esta familia tiene funciones de la forma

$$f(y_i; \theta_i) = a(\theta_i)b(y_i)\exp[c(y_i)Q(\theta_i)], \tag{2.3}$$

$a(\cdot)$, $b(\cdot)$, $c(\cdot)$ y $Q(\cdot)$ son funciones dadas, y donde el valor del parámetro θ_i puede variar para $i = 1, \ldots, N$ dependiendo de los valores de la variable explicatoria, el término $Q(\theta)$ es llamado parámetro natural.

El *componente sistemático* relaciona un vector $(\eta_1, \ldots, \eta_N)$ con las variables explicatorias a través de un modelo lineal. Sea x_{ij} que denota el valor del predictor

j (con $j = 1, 2, \ldots, p$) para el objeto i. Entonces

$$\eta_i = \sum_j \beta_j x_{ij}, \qquad i = 1, \ldots, N$$

esta combinación lineal de variables explicatorias es llamado predictor lineal, donde una de las x_{ij} es igual a 1, para el coeficiente del intercepto.

El tercer componente de un modelo lineal generalizado es la *función enlace* que conecta los componentes aleatorio y sistemático. Sea $\mu_i = E(Y_i)$, $i = 1, \ldots, N$. Los enlaces del modelo μ_i para η_i son $\eta_i = g(\mu_i)$, donde la función enlace g es una función monótona y diferenciable. Así la función enlace tiene la fórmula

$$g(\mu_i) = \sum_j \beta_j x_{ij}, \qquad i = 1, \ldots, N$$

La función enlace $g(\mu) = \mu$ es llamada enlace identidad, con $\eta_i = \mu_i$. Por su parte, la función enlace que transforma la media al parámetro natural es llamada enlace canónico.

En resumen, un modelo lineal generalizado es un modelo lineal para una media transformada de una variable respuesta que tiene distribución en la familia exponencial.

2.2.2. Modelos Logit Binomial

Muchas variables respuesta son binarias, y representan el éxito y fracaso por cero y uno, respectivamente. La distribución Bernoulli especifica las probabilidades $P(Y = 1) = \pi$ y $P(Y = 0) = 1 - \pi$ en donde $E(Y) = \pi$. La función de probabilidad

es

$$\begin{aligned} f(y;\pi) &= \pi^y(1-\pi)^{1-y} = (1-\pi)[\pi/(1-\pi)]^y \\ &= (1-\pi)\exp\left(y\log\frac{\pi}{1-\pi}\right) \end{aligned}$$

para y = 0 y 1. Esto es en la familia exponencial (2.3), identificando a θ como π, $a(\pi) = 1-\pi$, $b(y) = 1$, $c(y) = y$ y $Q(\pi) = \log[\pi/(1-\pi)]$, el parámetro natural $\log[\pi/(1-\pi)]$ es el logarítmo de la oportunidad de respuesta 1, es decir el *logit* de π. Los modelos lineales generalizados que utilizan el enlace logit son frecuentemente llamados modelos logit, denotado por $\text{logit}[\pi(x)] = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$. Este modelo será analizado con más detalle en el cuarto capítulo.

2.3. Estimación del Modelo Lineal Generalizado

Los dos métodos clásicos para estimar los parámetros desconocidos de un modelo lineal generalizado son el de verosimilitud máxima y mínimos cuadrados generalizados, a continuación se exponen los aspectos más importantes de ambos métodos de estimación.

2.3.1. Estimación por Mínimos Cuadrados Generalizados

Bajo las suposiciones establecidas para los errores al formular el modelo (2.2), los parámetros desconocidos pueden ser estimados por el método de mínimos cuadrados ordinarios. Estos parámetros se obtienen minimizando la suma de cuadrados de los errores, $S(\beta)$ (la suma de cuadrados de las desviaciones de los

valores observados y los esperados), es decir se trata de

$$Min_{\beta}S(\beta) = Min(Y - X\beta)'(Y - X\beta)$$

derivando con respecto a β e igualando a cero $S(\beta)$ se obtienen las ecuaciones normales

$$X'X\beta = X'Y \tag{2.4}$$

Resolviendo el sistema de ecuaciones (2.4) se tiene que

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

siempre que $(X'X)^{-1}$ exista. Por otro lado, cuando la suposición de varianza constante no se verifica, es decir $V(\varepsilon) = \sigma^2 V$, donde V es una matriz no singular y definida positiva, el método de mínimos cuadrados ordinarios no funciona, por lo que se debe considerar una reparametrización del modelo para que se cumplan las suposiciones establecidas al formular el modelo (2.1). Para estimar el vector de parámetros será necesario entonces

$$Min_{\beta}S(\beta) = Min(Y - X\beta)'V^{-1}(Y - X\beta) \tag{2.5}$$

la función objetivo $S(\beta)$ en (2.5) es una función continua y derivable, por lo que el mínimo se obtiene

$$\frac{\partial S(\beta)}{\beta} = \sum_{i=1}^{n} V(y_i)^{-1}[y_i - x_i\beta]x_{ij} = 0 \qquad j = 0, 1, \ldots, k.$$

El sistema de ecuaciones normales es

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

Finalmente, el estimador de mínimos cuadrados generalizados de β es

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

este estimador es insesgado, con matriz de covarianza dada por

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

2.3.2. Estimación por Verosimilitud Máxima

Si la función de distribución de Y pertenece a una familia de distribuciones **H** conocida, un método alternativo para estimar el vector de parámetros desconocidos β es el método de verosimilitud máxima.

Dado un vector de observaciones $\mathbf{y'} = (\mathbf{y_1}, \ldots, \mathbf{y_n})$, la función de verosimilitud cuantifica la posibilidad de que un vector β haya generado el vector de respuestas observado. La función de verosimilitud está dada por la función de densidad conjunta cuando se trate de una variable aleatoria continua (o por la función de probabilidad cuando sea el caso de una variable aleatoria discreta) de las variables aleatorias independientes $Y_1, \ldots, Y_n$ reduciendose la expresión a

$$L(\beta) = h(y_1, \ldots, y_n) = h(y_1, \beta)h(y_2, \beta)\ldots h(y_k, \beta) = \prod_{i=1}^{n} h(y_i; \beta) \qquad (2.6)$$

El estimador de verosimilitud máxima de β es el vector que maximiza $L(\beta)$ y para obtener el estimador máximo verosimil se necesita resolver el problema de maximizar $L(\beta)$ para β y dado que la función logarítmo es monótona, se aplica a la expresión (2.6) y se tiene

$$l(\beta) = \log L(\beta) = \sum_{i=1}^{n} \log h(y_i, \beta)$$

En consecuencia, si la función $l(\beta)$ es continua y derivable, maximizar $L(\beta)$ o $l(\beta)$ son procesos equivalentes. Para ilustrar, esto se obtiene el estimador de verosimilitud máxima para el caso del modelo lineal general, bajo la suposición que los errores se distribuyen normalmente con vector de medias cero y matriz de covarianza $\mathbf{V}$. La función de verosimilitud es

$$L = (2\pi)^{-n/2}|\mathbf{V}|^{-1/2}\exp\{-1/2(Y - X\beta)'\mathbf{V}^{-1}(Y - X\beta)\}$$

y su función transformada por el logarítmo es

$$l(\beta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}(Y - X\beta)'\mathbf{V}^{-1}(Y - X\beta)$$

$l(\beta)$ es una función continua y derivable, por lo tanto, derivando respecto a β e igualando a cero la expresión anterior, se obtiene

$$\mathbf{X'V^{-1}X}\beta = \mathbf{X'V^{-1}Y}$$

La solución de este sistema de ecuaciones conduce al estimador de verosimilitud máxima para β.

$$\hat{\beta} = (\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}Y}. \tag{2.7}$$

De las ecuaciones (2.7) y (2.5) se puede concluir que, bajo la suposición de normalidad los métodos de mínimos cuadrados y de verosimilitud máxima producen los mismos estimadores.

2.4. Modelos Lineales Generalizados para Datos Binarios

2.4.1. Modelo de Probabilidad Lineal

Para una respuesta binaria, el modelo de regresión

$$\pi(x) = \alpha + \beta x \tag{2.8}$$

es llamado modelo de probabilidad lineal. Con observaciones independientes es un modelo lineal generalizado con componente aleatoria binomial y función enlace identidad. El parámetro β representa el cambio en $\pi(x)$ para una unidad de incremento en x.

Se establece el modelo de probabilidad lineal con el siguiente ejemplo.

Ejemplo 2.1. La tabla (2.1) basada en un estudio epidemiológico de 2484 individuos para investigar si el ronquido es un posible factor de riesgo para una enfermedad del corazón; estos estudios fueron clasificados de acuerdo a lo reportado por sus esposas de cuánto roncan. El modelo manifiesta que la probabilidad de una enfermedad del corazón, denotada por $\pi(x)$ es relacionada linealmente con el nivel de ronquido x.

Se trata a las filas de la tabla como muestras binomiales independientes con esa probabilidad como parámetro, y se utiliza la notación (0, 2, 4, 5) correspondientes a las categorías de ronquido, tratando a los dos últimos niveles como más cercanos que las otras categorías adyacentes.

Solución. Según lo reportado por la función "generalizado"del paquete

	Enfermedad del corazón	
Ronquido	Si	No
Nunca	24	1355
Ocasional	35	603
Casi cada noche	21	192
Cada noche	30	224

Tabla 2.1: Relación entre ronquido y enfermedad del corazón

"Categoricos"en R, donde se proporciona la matriz de los datos indicando el número de filas y columnas, se tiene que el ajuste del modelo está dado por

$$\hat{\pi}(x) = \alpha + \beta x = 0.0172 + 0.0198x.$$

Por ejemplo, para los individuos que no roncan ($x = 0$), la proporción estimada de sujetos que tiene enfermedad del corazón es $\hat{\pi}(0) = 0.0172 + 0.0198(0) = 0.0172$, haciendo referencia a los valores estimados de $E(Y)$ para un modelo lineal generalizado como *valores ajustados*. La tabla (2.2) exhibe el ajuste para el modelo de probabilidad lineal, teniendo como la interpretación del modelo que la probabilidad estimada de enfermedad del corazón es por ejemplo, alrededor de 0.02 para individuos que no roncan.

Suponga ahora que se eligen para los niveles de ronquido la notación diferente al $\{0, 2, 4, 5\}$, por ejemplo $\{0, 1, 2, 3\}$ entonces los valores ajustados para las cuatro categorías de ronquido pueden cambiar un poco, teniendo que para cualquier

	Ajuste
Ronquido	Lineal
Nunca	0.017
Ocasional	0.057
Casi cada noche	0.096
Cada noche	0.116

Tabla 2.2: Ajuste para los datos de ronquido

incremento en las notaciones produce la conclusión que la oportunidad de una enfermedad del corazón incrementa conforme aumenta el nivel de ronquido.

2.4.2. Modelo Probit

En muchos análisis se ha utilizado la distribución normal, dando lugar al *modelo probit*, con lo que el modelo queda especificado a través de la siguiente expresión

$$\begin{aligned} \text{probit}[\pi(x)] &= \int_{-\infty}^{\alpha+\beta x} \phi(t)dt \\ &= \Phi(\alpha + \beta x) \end{aligned}$$

la función de distribución normal estándar acumulada se representa habitualmente como $\Phi^{-1}[\pi(x)] = \alpha + \beta x$ (Greene, 1999). El enlace probit aplicado a una probabilidad $\pi(x)$ transforma la normal estándar z en la cual la probabilidad de la cola izquierda es igual a $\pi(x)$, por ejemplo, probit(0.05) = -1.645, probit(0.50) = 0, probit(0.95) = 1.645 y probit(0.975) = 1.96. El modelo probit es un modelo li-

neal generalizado con componente aleatorio binomial y enlace probit. Enseguida se expone el modelo probit usando los datos de ronquido y enfermedad del corazón.

Ejemplo 2.2. El ajuste de la verosimilitud máxima del modelo probit, usando las categorías $\{0, 2, 4, 5\}$ para el nivel de ronquido, es

$$\text{probit}[\hat{\pi}(x)] = -2.061 + 0.188x$$

El ajuste se obtuvo mediante la función "probit"del paquete "Categoricos"en R, proporcionando la matriz con los datos de la tabla de referencia.
Con nivel de ronquido $x = 0$, el probit ajustado es igual a $-2.061 + 0.188(0) = -2.061$, y la probabilidad ajustada $\hat{\pi}(0)$ es la probabilidad de la cola izquierda para la distribución normal estándar en -2.061, igual a 0.020, con el nivel de ronquido $x = 5$, el probit ajustado es -2.061 + 0.188(5) = -1.12, con su correspondiente probabilidad ajustada de 0.131.

La transformación probit traza a $\pi(x)$ como la curva de regresión para $\pi(x)$ (o $1 - \pi(x)$, cuando $\beta < 0$) con apariencia de la función de distribución acumulada normal con media $\mu = -\alpha/\beta$ y desviación estándar $\sigma = 1/|\beta|$.

Para los datos del ejemplo (2.1) referente al ronquido y enfermedad del corazón, el ajuste probit corresponde a una función de distribución acumulada normal con media $-\hat{\alpha}/\hat{\beta} = 2.061/0.188 = 11$ y desviación estándar $1/|\hat{\beta}| = 1/0.188 = 5.3191$. La probabilidad predicha de enfermedad del corazón es 1/2 con nivel de ronquido $x = 11.0$; es decir, $x = 11$ tiene un probit ajustado de $-2.061 + 0.188(11) = 0$

que es el correspondiente valor de z para la probabilidad de la cola izquierda en 1/2. El valor de -2.06 en el probit ajustado con $x = 0$, implica que el cero es 2.06 desviaciones estándar bajo la media de una distribución normal con media 11.0 y desviación estándar 5.3.

2.5. Modelo Loglineal Poisson

Algunas variables respuesta tienen como resultados eventos poco comunes que se presentan en el espacio, tiempo, volúmen o cualquier otra dimensión, por ejemplo, para una muestra de obleas de silicón usadas en la manufactura de un placa de computadora, cada observación puede ser el número de imperfecciones en la oblea. La distribución más simple para estos datos es la Poisson, puesto que los resultados toman cualquier valor entero no negativo. Entonces sea Y que denota un resultado y con $\mu = E(Y)$, la función de probabilidad Poisson es

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!} = \exp(-\mu)\left(\frac{1}{y!}\right)\exp(y\log\mu) \qquad y = 0, 1, 2, \ldots$$

como en la familia exponencial (2.3), con $\theta = \mu$, $a(\mu) = \exp(-\mu)$, $b(y) = 1/y!$, $c(y) = y$ y $Q(\mu) = \log\mu$, el parámetro natural es el logarítmo de μ, así la función enlace es $\eta = \log\mu$. El modelo es

$$\log\mu = \alpha + \beta x$$

este modelo es llamado modelo loglineal Poisson con variable explicatoria X.

Además, la media satisface la siguiente relación exponencial

$$\mu = \exp(\alpha + \beta x) = e^{\alpha}(e^{\beta})^{x}$$

Para ilustrar el modelo loglineal Poisson se presenta el siguiente ejemplo con un estudio del apareamiento del cangrejo.

Ejemplo 2.3. Se analiza la tabla (2.3) que muestra el estudio de anidamiento de los cangrejos. Cada cangrejo hembra tiene un cangrejo macho residente en su nido. El estudio investiga los factores que afectan que el cangrejo hembra tenga otros machos, llamados *satélites*, residiendo cerca del nido. Las variables explicatorias son el color del cangrejo, condición de las espinas, peso y anchura del caparazón. El resultado o variable respuesta para cada cangrejo hembra es su número de satélites, utilizando la anchura como predictor, que se encuentra en la tabla (2.3) expresada en centímetros.

Solución. La media muestral del ancho es 26.3 y su desviación estándar es 2.1. Para un cangrejo hembra, sea μ el número esperado de satélites y sea x su ancho, se tiene que el ajuste de la verosimilitud máxima del modelo loglineal de Poisson ($\log \mu = \alpha + \beta x$) es

$$\log \hat{\mu} = \hat{\alpha} + \hat{\beta} x = -3.305 + 0.164x$$

puesto que $\hat{\beta} > 0$, el ancho tiene un efecto estimado positivo sobre el número de satélites. El ajuste del modelo produce la media estimada del número de satélites $\hat{\mu}$, un valor ajustado, en cualquier valor del ancho. Por ejemplo, el valor ajustado al

Tabla 2.3: Número de satélites por las características del cangrejo hembra

C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa	C	S	W	Wt	Sa
2	3	28.3	3.05	8	3	3	22.5	1.55	0	1	1	26.0	2.30	9	3	3	24.8	2.10	0
3	3	26.0	2.60	4	2	3	23.8	2.10	0	3	2	24.7	1.90	0	2	1	23.7	1.95	0
3	3	25.6	2.15	0	3	3	24.3	2.15	0	2	3	25.8	2.65	0	2	3	28.2	3.05	11
4	2	21.0	1.85	0	2	1	26.0	2.30	14	1	1	27.1	2.95	8	2	3	25.2	2.00	1
2	3	29.0	3.00	1	4	3	24.7	2.20	0	2	3	27.4	2.70	5	2	2	23.2	1.95	4
1	2	25.0	2.30	3	2	1	22.5	1.60	1	3	3	26.7	2.60	2	4	3	25.8	2.00	3
4	3	26.2	1.30	0	2	3	28.7	3.15	3	2	1	26.8	2.70	5	4	3	27.5	2.60	0
2	3	24.9	2.10	0	1	1	29.3	3.20	4	1	3	25.8	2.60	0	2	2	25.7	2.00	0
2	1	25.7	2.00	8	2	1	26.7	2.70	5	4	3	23.7	1.85	0	2	3	26.8	2.65	0
2	3	27.5	3.15	6	4	3	23.4	1.90	0	2	3	27.9	2.80	6	3	3	27.5	3.10	3
1	1	26.1	2.80	5	1	1	27.7	2.50	6	2	1	30.0	3.30	5	3	1	28.5	3.25	9
3	3	28.9	2.80	4	2	3	28.2	2.60	6	2	3	25.0	2.10	4	2	3	28.5	3.00	3
2	1	30.3	3.60	3	4	3	24.7	2.10	5	2	3	27.7	2.90	5	1	1	27.4	2.70	6
2	3	22.9	1.60	4	2	1	25.7	2.00	5	2	3	28.3	3.00	15	2	3	27.2	2.70	3
3	3	26.2	2.30	3	2	1	27.8	2.75	0	4	3	25.5	2.25	0	3	3	27.1	2.55	0
3	3	24.5	2.05	5	3	1	27.0	2.45	3	2	3	26.0	2.15	5	2	3	28.0	2.80	1
2	3	30.0	3.05	8	2	3	29.0	3.20	10	2	3	26.2	2.40	0	2	1	26.5	1.30	0
2	3	26.2	2.40	3	3	3	25.6	2.80	7	3	3	23.0	1.65	1	3	3	23.0	1.80	0
2	3	25.4	2.25	6	3	3	24.2	1.90	0	2	2	22.9	1.60	0	3	2	26.0	2.20	3
2	3	25.4	2.25	4	3	3	25.7	1.20	0	2	3	25.1	2.10	5	3	2	24.5	2.25	0
4	3	27.5	2.90	0	3	3	23.1	1.65	0	3	1	25.9	2.55	4	2	3	25.8	2.30	0
4	3	27.0	2.25	3	2	3	28.5	3.05	0	4	1	25.5	2.75	0	4	3	23.5	1.90	0
2	2	24.0	1.70	0	2	1	29.7	3.85	5	2	1	26.8	2.55	0	4	3	26.7	2.45	0
2	1	28.7	3.20	0	3	3	23.1	1.55	0	2	1	29.0	2.80	1	3	3	25.5	2.25	0
3	3	26.5	1.97	1	3	3	24.5	2.20	1	3	3	28.5	3.00	1	2	3	28.2	2.87	1
2	3	24.5	1.60	1	2	3	27.5	2.55	1	2	2	24.7	2.55	4	2	1	25.2	2.00	1
3	3	27.3	2.90	1	2	3	26.3	2.40	1	2	3	29.0	3.10	1	2	3	25.3	1.90	2
2	3	26.5	2.30	4	2	3	27.8	3.25	3	2	3	27.0	2.50	6	3	3	25.7	2.10	0
2	3	25.0	2.10	2	2	3	31.9	3.33	2	4	3	23.7	1.80	0	4	3	29.3	3.23	12
3	3	22.0	1.40	0	2	3	25.0	2.40	5	3	3	27.0	2.50	6	3	3	23.8	1.80	6
1	1	30.2	3.28	2	3	3	26.2	2.22	0	2	3	24.2	1.65	2	2	3	27.4	2.90	3
2	2	25.4	2.30	0	3	3	28.4	3.20	3	4	3	22.5	1.47	4	2	3	26.2	2.02	2
2	1	24.9	2.30	6	1	2	24.5	1.95	6	2	3	25.1	1.80	0	2	1	28.0	2.90	4
4	3	25.8	2.25	10	2	3	27.9	3.05	7	2	3	24.9	2.20	0	2	1	28.4	3.10	5
3	3	27.2	2.40	5	2	2	25.0	2.25	6	2	3	27.5	2.63	6	2	1	33.5	5.20	7
2	3	30.5	3.32	3	3	3	29.0	2.92	3	2	1	24.3	2.00	0	2	3	25.8	2.40	0
4	3	25.0	2.10	8	2	1	31.7	3.73	4	2	3	29.5	3.02	4	3	3	24.0	1.90	10
2	3	30.0	3.00	9	2	3	27.6	2.85	4	2	3	26.2	2.30	0	2	1	23.1	2.00	0
2	1	22.9	1.60	0	4	3	24.5	1.90	0	2	3	24.7	1.95	4	2	3	28.3	3.20	0
2	3	23.9	1.85	2	3	3	23.8	1.80	0	3	2	29.8	3.50	4	2	3	26.5	2.35	4
2	3	26.0	2.28	3	2	3	28.2	3.05	8	4	3	25.7	2.15	0	2	3	26.5	2.75	7
2	3	25.8	2.20	0	3	3	24.1	1.80	0	3	3	26.2	2.17	2	3	3	26.1	2.75	3
3	3	29.0	3.28	4	1	1	28.0	2.62	0	4	3	27.0	2.63	0	2	2	24.5	2.00	0
1	1	26.5	2.35	0															

C: color(1, claro medio; 2, medio; 3, oscuro medio; 4, oscuro), S: condición de la espina (1, ambas buenas; 2, una gastada o rota; 3, ambas gastadas o rotas), W: ancho del caparazón (cm); Wt, peso (kg); Sa, número de satélites.

ancho de la media de $x = 26.3$ es

$$\hat{\mu} = \exp(\hat{\alpha} + \hat{\beta}x) = \exp[-3.305 + 0.164(26.3)] = 2.74$$

es el valor del número promedio de satélites que tiene un cangrejo con ancho de 26.3 centímetros.

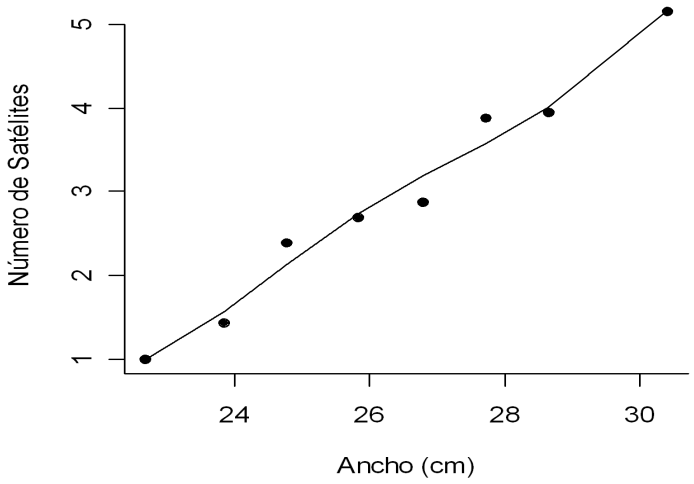

Figura 2.1: Número de satélites por anchura del cangrejo hembra

La figura (2.1) muestra que el número de satélites puede tener un crecimiento aproximadamente lineal con el ancho, ésto sugiere un modelo Poisson con enlace idéntico y se tiene el ajuste de la verosimilitud máxima

$$\hat{\mu} = \hat{\alpha} + \hat{\beta}x = -11.53 + 0.55x$$

El efecto de X en μ en este modelo es aditivo, más que multiplicativo, por un centímetro de incremento en x se tiene un incremento estimado de $\hat{\beta} = 0.55$ en $\hat{\mu}$. Por ejemplo el valor ajustado en la media del ancho de x $= 26.3$ es $\hat{\mu} = -11.53 + 0.55(26.3) = 2.93$; en x $= 27.3 = 26.3 + 1$, es $\hat{\mu} = 2.93 + 0.55 = 3.48$. Los valores ajustados son positivos para toda la muestra x, y el modelo describe el efecto de que en promedio se incrementa 2 centímetros aproximadamente de ancho está asociado con un extra satélite.

En la tabla (2.4) se muestran los resultados de la media muestral para el número de satélites, tomando en cuenta el ancho del cangrejo hembra, donde la media es el

Ancho (cm)	Número de Casos	Número de Satélites	Media Muestral
< 23.25	14	14	1.00
23.25 - 24.25	14	20	1.43
24.25 - 25.25	28	67	2.39
25.25 - 26.25	39	105	2.69
26.25 - 27.25	22	63	2.86
27.25 - 28.25	24	93	3.87
28.25 - 29.25	18	71	3.94
> 29.25	14	72	5.14

Tabla 2.4: Media y varianza muestral para el número de satélites

número de satélites entre el número de casos.

Para obtener los ajustes para los datos se utilizará la función "loglineal"del paquete "Categoricos", se proporcionará la tabla analizada, indicando la variable respuesta y predictor así como el número de columnas y de filas.

2.6. Ventajas de los Modelos Lineales Generalizados

Hace dos décadas, el desarrollo de la teoría de los modelos lineales generalizados unificó importantes modelos para variables respuesta continuas y categóricas. Por razones teóricas, el componente aleatorio en la definición de un modelo lineal generalizado debe tener una distribución de la familia exponencial, esta restricción no es grave puesto que esta familia contiene distribuciones más importantes, incluyendo Poisson, binomial, multinomial y normal, un rasgo bueno para los modelos tratados en el capítulo es que el algoritmo de ajuste del modelo es el mismo para cualquier

modelo lineal generalizado, ésto se sostiene sin tomar en cuenta la elección de la distribución para el componente aleatorio o la elección de la función enlace.

CAPÍTULO 3

REGRESIÓN LOGÍSTICA

La regresión logística es el modelo más popular para modelar datos de respuesta categórica y ha incrementado su uso en una extensa variedad de aplicaciones (Agresti, 1996, Kleinbaum et al. 1998). Es por tal motivo que en este capítulo se estudia en detalle dicha regresión.

3.1. Interpretación del modelo de Regresión Logística

Sea Y una variable de respuesta binaria y X una variable explicatoria, además sea $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$ que denota la probabilidad de "éxito"cuando X toma x valores. El modelo de regresión logística es

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \tag{3.1}$$

donde α es el parámetro de intercepto y el parámetro β representa el cambio en la probabilidad por unidad de cambio en x, la notación exp se refiere a la función

exponencial. Equivalentemente, el logarítmo de la oportunidad, llamada logit, tiene la relación lineal

$$\text{logit}[\pi(x)] = \log\frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x \tag{3.2}$$

La ecuación anterior se obtiene como

$$\begin{aligned}
\pi(x) &= \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} \\
\pi(x)[1+\exp(\alpha+\beta x)] &= \exp(\alpha+\beta x) \\
\pi(x)+\pi(x)\exp(\alpha+\beta x) &= \exp(\alpha+\beta x) \\
\pi(x) &= [1-\pi(x)]\exp(\alpha+\beta x) \\
\frac{\pi(x)}{1-\pi(x)} &= \exp(\alpha+\beta x)
\end{aligned}$$

y aplicando logarítmos en ambos lados

$$\begin{aligned}
\log\left(\frac{\pi(x)}{1-\pi(x)}\right) &= \log\left[\exp(\alpha+\beta x)\right] \\
&= \alpha+\beta x
\end{aligned}$$

se tiene la expresión deseada.

Esto implica que $\pi(x)$ incrementa o decrementa como una función de x en forma de S, con x que está en el rango (0, 1).

3.1.1. Interpretación de β

El signo de β indica si $\pi(x)$ aumenta o disminuye conforme x crece. El coeficiente de ascenso o descenso incrementa conforme el valor de $|\beta|$ aumenta, cuando β tiende a cero la curva se achata a una línea horizontal. Cuando $\beta = 0$, Y es independiente a X.

Para x cuantitativa, con $\beta > 0$, la curva $\pi(x)$ tiene la forma de la función de distribución acumulada de la distribución logística y puesto que la densidad logística es simétrica, $\pi(x)$ se aproxima a uno con la misma proporción que se aproxima a cero. Expresando en forma exponencial ambos lados de (3.2) se muestra que las oportunidades son una función exponencial de x, esto proporciona una interpretación básica para la magnitud de β: La oportunidad incrementa multiplicativamente en e^{β} por cada unidad que se incrementa en x. Por otra parte, la inclinación de la curva ocurre cuando $x = -\alpha/\beta$, este valor frecuentemente es llamado *nivel medio eficaz* y se denota por EL_{50} y representa el nivel en el cual el resultado tiene un 50 por ciento de acierto.

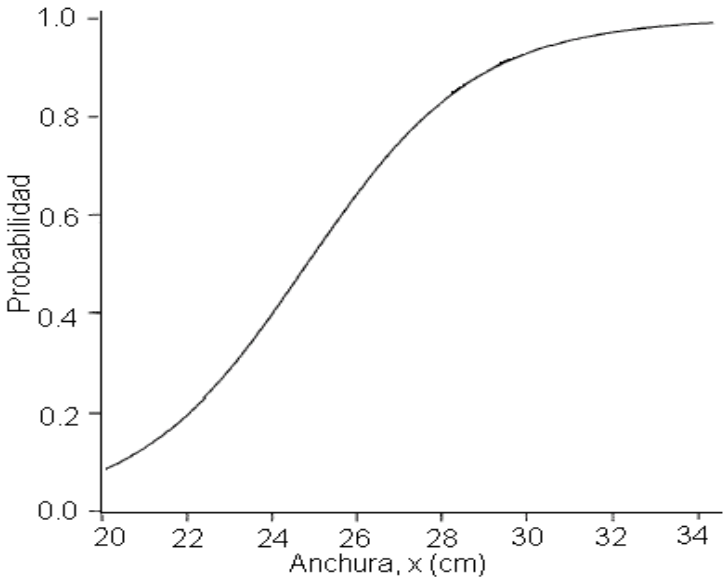

Figura 3.1: Aproximación lineal a la curva de regresión logística

La figura (3.1) muestra el modelo de regresión logística para $\pi(x)$ con apa-

riencia en forma de S, la función (3.1) implica que el coeficiente de cambio en $\pi(x)$ varía por una unidad de cambio en x.

Ejemplo 3.1. Para ilustrar el modelo, se retoma el ejemplo (2.3) introducido en el capítulo anterior de los datos del apareamiento del cangrejo.

Se usa respuesta binaria cuando un cangrejo hembra presenta satélites, esto es, $Y = 1$ si un cangrejo hembra tiene por lo menos un satélite, y $Y = 0$ si no tiene satélite. Puesto que Y toma sólo valores 0 y 1, la dificultad para determinar la regresión logística del modelo es razonable al graficar Y contra x. Para tener resultados con una mejor información se agrupan los valores del ancho y se calcula una proporción de la muestra de cangrejos llamados satélites para cada categoría.

Sea $\pi(x)$ que denota la probabilidad de que un cangrejo hembra con un ancho x tenga un satélite, el modelo a interpretar es un modelo de probabilidad lineal $\pi(x) = \alpha + \beta x$.

El parámetro de la verosimilitud máxima estimada para el modelo de regresión logística es $\hat{\alpha} = -12.351$ y $\hat{\beta} = 0.497$. La probabilidad predicha de un satélite análoga a (3.1) es

$$\hat{\pi}(x) \quad = \quad \frac{\exp(-12.351 + 0.497x)}{1 + \exp(-12.351 + 0.497x)}$$

El ancho mínimo en la muestra es 21.0 centímetros, así que la probabilidad esperada es $\hat{\pi}(x) = \exp(-12.351 + 0.497(21))/[1 + \exp(-12.351 + 0.497(21))] = 0.129$, con el ancho máximo de 33.5 centímetros, la probabilidad esperada es $\exp(-12.351 + 0.497(33.5))/[1 + \exp(-12.351 + 0.497(33.5))] = 0.987$, y el nivel medio eficaz del ancho con la probabilidad esperada igual a 0.5, es $x = EL_{50} = \hat{\alpha}/\hat{\beta} = 12.351/0.497 =$

24.8. Teniendo que el modelo de regresión logística permite variar la razón de cambio conforme x (el ancho) varía.

Para concluir el ejemplo se puede ver que la oportunidad estimada de un satélite multiplicada por $\exp(\hat{\beta}) = \exp(0.497)$ es 1.64 por cada centímetro de incremento en el ancho, es decir, un 64 % de incremento. Para la media del ancho $x = 26.3$, la probabilidad esperada $\hat{\pi}(x)$ es

$$\begin{aligned} \hat{\pi}(x) &= \frac{\exp(-12.351 + 0.497(26.3))}{1 + \exp(-12.351 + 0.497(26.3))} \\ &= 0.674 \end{aligned}$$

la razón de incremento en la probabilidad ajustada en un punto es $\hat{\beta}\hat{\pi}(x)[1-\hat{\pi}(x)] = 0.497(0.674)(0.326) = 0.11$ que describe la proporción de cambio. Para los cangrejos hembra que tienen su ancho cerca de la media, la probabilidad estimada de un satélite incrementa 0.11 por centímetro de aumento en el ancho.

Los cálculos efectuados para el ejemplo se verifican con la función "regresion"del paquete "Categoricos"introduciendo la tabla en forma de matriz donde se indicará la variable respuesta y predictor así como el número de filas y de columnas.

3.2. Inferencia en la Regresión Logística

Los estudios que se tienen del ajuste del modelo de regresión logística ayudan a describir los efectos de los predictores en una variable de respuesta binaria. Enseguida se presenta la inferencia estadística para los parámetros del modelo, para ayudar a juzgar la significancia y el tamaño de los efectos.

3.2.1. Intervalo de Confianza para los Efectos

El intervalo de confianza para el parámetro β en el modelo de regresión logística, $\text{logit}[\pi(x)] = \alpha + \beta x$ es

$$\hat{\beta} \pm z_{\alpha/2}(ASE)$$

donde $z_{\alpha/2}$ es el valor de z con área $z/2$ en la cola derecha de una distribución normal estándar, y ASE es el error estándar asintótico de $\hat{\beta}$ con expresión igual a $\sqrt{\hat{\sigma}^2(\hat{\beta}_j)}$ donde la notación $\hat{\sigma}_j^2(\hat{\beta}_j)$ denota la varianza estimada de $\hat{\beta}_j$.
Expresando en forma exponencial los extremos del intervalo se produce, $e^{\hat{\beta}}$, que es el efecto multiplicativo en la oportunidad por una unidad de incremento en X.

Para ilustrar lo precedente, se expone el ejemplo de los cangrejos que ha sido revisado con anterioridad, teniendo que el efecto estimado del ancho en la ecuación ajustada para la probabilidad de un sátelite, es $\hat{\beta} = 0.497$, con $ASE = 0.102$. Un intervalo de confianza al 95 % para β es $0.497 \pm 1.96(0.102)$, es decir $(0.298, 0.697)$.

3.2.2. Prueba de Significancia

Para el modelo de regresión logística, la hipótesis nula $H_0 : \beta = 0$ es que la probabilidad de éxito sea independiente de X. Para muestras grandes, el estadístico de prueba es

$$z = \frac{\hat{\beta}}{ASE}$$

que tiene una distribución normal cuando $\beta = 0$, donde $ASE = \sqrt{\hat{\sigma}^2(\hat{\beta}_j)}$ y $\hat{\sigma}^2(\hat{\beta}_j)$ denota la varianza estimada de $\hat{\beta}_j$; donde $z_{\alpha/2}$ es el valor de z con área $z/2$ en la cola

derecha de una distribución normal estándar. La prueba de la razón de verosimilitud es potente y fiable para las muestras en la práctica, la prueba estadística compara el máximo L_0 de la función de log-verosimilitud cuando $\beta = 0$ (es decir, cuando $\pi(x)$ es forzado a ser idéntico para todos los valores de x) con el máximo L_1 de la función de log-verosimilitud para β.

Algunos paquetes computacionales estadísticos, para la regresión logística reportan el aumento máximo de la log-verosimilitud L_0 y L_1 y el estadístico de la razón de verosimilitud derivado de éste máximo.

Por ejemplo, para los datos de los cangrejos analizados previamente, el estadístico $z = \hat{\beta}/ASE$ es igual a $0.497/0.102 = 4.9$ que muestra fuerte evidencia de un efecto positivo en el ancho con la presencia de un satélite ($p < 0.0001$).

3.3. Modelo Logit con Predictores Categóricos

La regresión logística, tal como la regresión ordinaria se extiende a modelos que incorporan variables explicatorias múltiples. En esta sección se mostrará el uso de variables *dummy* que incluyen predictores cualitativos, frecuentemente llamados *factores*.

3.3.1. Variables Dummy en el Modelo Logit

Una expresión equivalente para el modelo

$$\log\frac{\pi_i}{1-\pi_i} = \alpha + \beta_i x \tag{3.3}$$

es utilizando variables dummy. Sea $x_i = 1$ para observaciones en la fila i y $x_i = 0$ en otra parte, $i = 1, \ldots, I-1$, donde I son las categorías existentes. El modelo es

$$\text{logit}(\pi_i) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{I-1} x_{I-1}$$

Estos cálculos para parámetros redundantes no forman una variable dummy para la categoría I, el contraste $\beta_I = 0$ en (3.3) corresponde a esta forma de variable dummy, y la selección de la categoría para excluir dicha variable es arbitraria. Otra manera de imponer un conjunto de contrastes es que $\sum_i \beta_i = 0$. Suponga que X tiene $I = 2$ categorías, de esta manera $\beta_1 = -\beta_2$, este resultado tiene efecto codificado para variables dummy, $x = 1$ en la categoría 1 y $x = -1$ en la categoría 2, así el mismo resultado ocurre para cualquier esquema de código. Las diferencias $\hat{\beta}_a - \hat{\beta}_b$ para pares (a, b) de categorías de X son idénticas y representan la proporción estimada del logarítmo de la razón de oportunidades. Así, $\exp(\hat{\beta}_a - \hat{\beta}_b)$ es la oportunidad estimada de éxito en la categoría a de X dividida por la oportunidad de éxito en la categoría b de X. Este tipo de variables se emplean en el ejemplo (3.2) que se expone más adelante.

3.4. Regresión Logística Múltiple

Tal como la regresión ordinaria, la regresión logística se generaliza a modelos con variables explicatorias múltiples. Los predictores pueden ser cuantitativos, cualitativos, o de ambos tipos. Por ejemplo, el modelo para $\pi(\mathbf{x}) = P(Y = 1)$ con

valores $\mathbf{x} = (x_1, \ldots, x_p)$ para p predictores es

$$\text{logit}[\pi(x)] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{3.4}$$

La fórmula alternativa, especificando directamente a $\pi(\mathbf{x})$, es

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)} \tag{3.5}$$

donde el parámetro β_i se refiere al efecto de x_i en el logarítmo de la oportunidad cuando $Y = 1$, y el otro x_j controlado, por ejemplo, $\exp(\beta_i)$ es el efecto multiplicativo en la oportunidad por unidad de incremento en x_i. Se tiene que una variable explicatoria puede ser cualitativa, usando variables dummy para sus categorías. Enseguida se explica la regresión logística múltiple con un ejemplo.

Ejemplo 3.2. Continuando con el ejemplo (2.3) analizado durante el desarrollo de éste trabajo, en la regresión logística se puede tener una mezcla de predictores cuantitativos y cualitativos, esto se expresa con los datos de los cangrejos mostrados en la tabla (2.3), usando el ancho del cangrejo hembra y el color como predictores. En general, el color tiene cinco categorías: claro, claro medio, medio, oscuro medio y oscuro, y el color es un sustituto para la edad, ya que cangejos más viejos tienden a ser más oscuros. La muestra no contiene cangrejos de color claro, así que el modelo sólo tendrá cuatro categorías del color.

Primero, se trata el color como variable cualitativa, de manera que se usan tres variables dummy para representar las cuatro categorías. El modelo es

$$\text{logit}(\pi) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x \tag{3.6}$$

donde $\pi = P(Y = 1)$ y x denota el ancho en centímetros y

$$\begin{aligned} c_1 &= 1 \text{ para color claro medio, y 0 de otra forma,} \\ c_2 &= 1 \text{ para color medio, y 0 de otra forma,} \\ c_3 &= 1 \text{ para color oscuro medio, y 0 de otra forma.} \end{aligned}$$

El color del cangrejo es oscuro (categoría 4) cuando $c_1 = c_2 = c_3 = 0$. La verosimilitud máxima estimada de los parámetros son

$$\begin{aligned} Intercepto: \quad \hat{\alpha} = & \quad -12.715, \quad ASE = 2.762 \\ c_1: \quad \hat{\beta}_1 = & \quad 1.330, \quad ASE = 0.852 \\ c_2: \quad \hat{\beta}_2 = & \quad 1.402, \quad ASE = 0.548 \\ c_3: \quad \hat{\beta}_3 = & \quad 1.106, \quad ASE = 0.592 \\ Ancho: \quad \hat{\beta}_4 = & \quad 0.468, \quad ASE = 0.106 \end{aligned}$$

por ejemplo, para un cangrejo oscuro, con $c_1 = c_2 = c_3 = 0$, la ecuación de predicción es $\text{logit}(\hat{\pi}) = -12.715 + 0.468x$, contrastando, para un cangrejo de color claro medio, $c_1 = 1$, la ecuación predictora es $\text{logit}(\hat{\pi}) = (-12.715 + 1.330) + 0.468x = -11.385 + 0.468x$. Para ilustrar las probabilidades predichas de un satélite, se utiliza un cangrejo de color claro medio con un promedio de ancho de 26.3 cm, la probabilidad predicha es

$$\begin{aligned} \hat{\pi}(x) &= \frac{\exp[-11.385 + 0.468(26.3)]}{1 + \exp[-11.385 + 0.468(26.3)]} \\ &= 0.715 \end{aligned}$$

En comparación, un cangrejo oscuro con el mismo ancho tiene probabilidad predicha

de

$$\begin{aligned}\hat{\pi}(x) &= \frac{\exp[-12.715 + 0.468(26.3)]}{1 + \exp[-12.715 + 0.468(26.3)]} \\ &= 0.399\end{aligned}$$

La estimación del color indica que, en la muestra, los cangrejos oscuros tienen menos probabilidad de tener satélites que los cangrejos de otros colores. El modelo asume una pérdida de interacción entre color y ancho para los efectos en la respuesta.

Los valores se obtuvieron utilizando la función "dummy"del paquete "Categoricos"en R donde se introducirá la tabla de referencia en forma de matriz con las columnas de la variable respuesta y el predictor.

3.5. Tamaño de Muestra y Potencia para Regresión Logística

El objetivo de muchos estudios es determinar si una variable en particular X tiene efectos en la respuesta binaria. El análisis debe tomar en cuenta el tamaño de muestra N que se necesita para estipular una buena posibilidad de detectar el efecto de un tamaño dado.

3.5.1. Tamaño de Muestra con Predictores Cuantitativos

Para modelos de la forma

$$\text{logit}[\pi(x)] = \alpha + \beta x_i$$

en el cual x es un indicador cuantitativo, el tamaño de muestra necesario para probar $H_0 : \beta = 0$ depende de la distribución de los valores de x, además suponga la probabilidad de éxito $\bar{\pi}$ como la media de x. El tamaño del efecto es la razón de oportunidades θ comparando la probabilidad de éxito en ese punto, donde dicha probabilidad y la desviación estándar están por encima de la media de x y, sea $\lambda = log(\theta)$. F. Y. Hsieh (1989) proporcionó la aproximación al tamaño de muestra para una prueba de una cola, sea

$$N = \frac{[z_\alpha z_\beta \exp(-\lambda^2/4)]^2(1 + 2\bar{\pi}\delta)}{\bar{\pi}\lambda^2}$$

donde

$$\delta = \frac{1 + (1 + \lambda^2)\exp(5\lambda^2/4)}{1 + \exp(-\lambda^2/4)}$$

Ejemplo 3.3. Para modelar la probabilidad de la dependencia de varias enfermedades del corazón con x = nivel de colesterol para una población de edad mediana, considere la prueba de indepencia $\beta = 0$ contra la alternativa $\beta > 0$ de que incremente el riesgo así como incrementa el colesterol. Suponga que estudios previos han sugerido que la probabilidad de varias enfermedades del corazón en un nivel promedio de colesterol es 0.08, además se quiere probar que puede ser susceptible a un 50 % de incremento en esta probabilidad, es decir a 0.12, para una desviación estándar aumentando el colesterol.

Solución. La oportunidad de varias enfermedades del corazón con un nivel medio de colesterol es igual a $0.08/0.092 = 0.087$, y la oportunidad con desviación por encima de la media es $0.12/0.88 = 0.136$. Así, la razón de oportunidades es igual a $\theta = 0.136/0.087 = 1.57$ y $\lambda = \log(1.57) = 0.450$, $\lambda^2 = 0.202$. Para $\beta = 0.10$

de cambio en un error tipo II y $\alpha = 0.05$, donde α y β son las probabilidades de cometer un error tipo I y error tipo II, respectivamente, $z_\alpha = z_{0.05} = 1.645$ y $z_\beta = z_{0.10} = 1.28$, Así

$$\begin{aligned} \delta &= \frac{1 + (1.202)\exp(5(0.202)/4)}{1 + \exp(-0.202/4)} \\ &= \frac{2.547261874}{1.950753929} \\ &= 1.306 \end{aligned}$$

$$\begin{aligned} N &= \frac{[1.645 + 1.28\exp(-0.202/4)]^2(1 + 2(0.08)(1.306))}{(0.08)(0.202)} \\ &= 612 \end{aligned}$$

El valor de N disminuye conforme $\bar{\pi}$ está cercano a 0.5 y conforme $|\lambda|$ toma valores más alejados del valor nulo. Esta derivación supone que la variable explicatoria tiene aproximadamente una distribución normal.

3.5.2. Tamaño de Muestra en la Regresión Logística Múltiple

Un modelo de regresión logística múltiple requiere tamaños de muestra grandes para detectar efectos parciales. Sea R que denota la correlación múltiple entre el predictor de interés X y otros en el modelo. El tamaño de muestra se da al dividir N entre $(1 - R^2)$, en la fórmula, $\bar{\pi}$ denota la probabilidad de la media de todas las variables explicatorias, y la razón de oportunidades se refiere al efecto del predictor de interés al nivel medio de los otros. A manera de ejemplificar lo anterior, se retoma el ejemplo previo, considere una prueba para el efecto del colesterol en

varias enfermedades del corazón, mientras se controla el nivel de presión sanguínea. Si la correlación entre colesterol y el nivel de presión sanguínea es 0.40, el tamaño de muestra aproximado es

$$N \approx \frac{N}{(1 - R^2)} = \frac{612}{(1 - 0.40^2)} = 728.5714286 \approx 729.$$

CAPÍTULO 4

MODELO LOGIT PARA DATOS MULTINOMIALES

Para cada variable explicatoria, el modelo asume que la respuesta para las categorías de Y tienen distribución multinomial. Este capítulo presenta una generalización del modelo de regresión logística que maneja respuestas multinomiales.

4.1. Modelo Logit para respuestas nominales

Suponga Y es una variable nominal con J categorías, el orden en que se listan las categorías es irrelevante. Sea $\{\pi_1(x), \ldots, \pi_J(x)\}$ que denotan la probabilidad de respuesta, satisfaciendo que $\sum_j \pi_j(x) = 1$. Cuando se toman n observaciones independientes basadas en esas probabilidades, la distribución de probabilidad para el número de resultados que ocurran de cada una de las J categorías, es multinomial, especificando la probabilidad para cada posible manera de asignar las n observaciones a las J categorías. Los modelos logit multicategóricos se refieren a todos los pares de categorías, y describen la oportunidad de respuesta en cada categoría.

4.2. Categoría Básica del Logit

Sea $\pi_j(\mathbf{x}) = P(Y = j|\mathbf{x})$ en un conjunto fijo $\mathbf{x}$ de variables explicatorias, con $\sum_j \pi_j(\mathbf{x}) = 1$, y para las observaciones, en ese conjunto los cálculos en las J categorías de Y son tratados como multinomiales con probabilidades $\{\pi_1(\mathbf{x}), \ldots, \pi_j(\mathbf{x})\}$. En el modelo logit cada par de categorías respuesta y base, es arbitrario. Cuando la última categoría (J) es la base, el modelo logit es

$$\log \frac{\pi_j(\mathbf{x})}{\pi_J(\mathbf{x})} = \alpha_j + \beta_j'\mathbf{x} \qquad j = 1, \ldots, J-1 \tag{4.1}$$

simultáneamente describe los efectos de $\mathbf{x}$ en esos $J-1$ logits, los efectos varían de acuerdo al par de respuesta con la base. Éstas $J-1$ ecuaciones determinan los parámetros para los logit's con otros pares de categorías de respuesta, por ejemplo, para un par arbitrario de categorías a y b,

$$\begin{aligned}
\log \frac{\pi_a(\mathbf{x})}{\pi_b(\mathbf{x})} &= \log \left(\frac{\pi_a(\mathbf{x})/\pi_J(\mathbf{x})}{\pi_b(\mathbf{x})/\pi_J(\mathbf{x})} \right) \\
&= \log \frac{\pi_a(\mathbf{x})}{\pi_J(\mathbf{x})} - \log \frac{\pi_b(\mathbf{x})}{\pi_J(\mathbf{x})} \\
&= (\alpha_a + \beta_a\mathbf{x}) - (\alpha_b + \beta_b\mathbf{x}) \\
&= (\alpha_a - \alpha_b) + (\beta_a - \beta_b)\mathbf{x}
\end{aligned}$$

así, la ecuación del logit para las categorías a y b tiene parámetro de intercepto $(\alpha_a - \alpha_b)$ y parámetro de inclinación $(\beta_a - \beta_b)$.

Con predictores categóricos, X^2 y G^2 los estadísticos de bondad de ajuste proporcionan un control del modelo cuando los datos no están dispersos. Cuando una variable explicatoria es continua o los datos son dispersos, tales estadísticos son válidos para comparar modelos anidados difiriendo por los relativos pocos términos.

Una generalización del modelo (4.1) permite que las variables explicatorias tomen diferentes valores para distintas categorías de respuesta, por ejemplo, la elección de una marca de automóvil probablemente dependa del precio, de la variación entre las opciones de la marca, entre otras. Éste modelo generalizado, frecuentemente es llamado *modelo logit multinomial* $\log(\pi_j(\mathbf{x})/\pi_J(\mathbf{x})) = \alpha_j + \beta_j'\mathbf{x}$.

Éste concepto se ilustra con el siguiente ejemplo, que muestra el análisis en la selección de comida de un caimán.

Ejemplo 4.1. En la tabla (4.1) se muestra un estudio de los factores que influyen en la elección de la comida primaria de los caimánes.

		Tamaño	Elección de comida primaria				
Lago	Género	(metros)	Pez	Invertebrado	Reptil	Ave	Otro
Hancock	Macho	≤ 2.3	7	1	0	0	5
		> 2.3	4	0	0	1	2
	Hembra	≤ 2.3	16	3	2	2	3
		> 2.3	3	0	1	2	3
Oklawaha	Macho	≤ 2.3	2	2	0	0	1
		> 2.3	13	7	6	0	0
	Hembra	≤ 2.3	3	9	1	0	2
		> 2.3	0	1	0	1	0
Trafford	Macho	≤ 2.3	3	7	1	0	1
		> 2.3	8	6	6	3	5
	Hembra	≤ 2.3	2	4	1	1	4
		> 2.3	0	1	0	0	0
George	Macho	≤ 2.3	13	10	0	2	2
		> 2.3	9	0	0	1	2
	Hembra	≤ 2.3	3	9	1	0	1
		> 2.3	8	1	0	0	1

Tabla 4.1: Elección de comida en los caimánes

Se usan 219 caimánes capturados en cuatro lagos de Florida. La variable respuesta nominal es el tipo de comida primaria, en volumen, encontrada en el

estómago del caimán, esto tiene cinco categorías: pez, invertebrado, reptil, ave y otro. En la tabla (4.1) se clasifican los caimánes de acuerdo a L: lago de captura (Hancock, Oklawaha, Trafford, George), G: género (macho, hembra) y S: tamaño (≤ 2.3 metros de largo, > 2.3 metros de largo).

Solución. La tabla (4.2) muestra los valores ajustados para el modelo $(L + S)$

	Tamaño	Elección de comida primaria				
Lago	(metros)	Pez	Invertebrado	Reptil	Ave	Otro
Hancock	≤ 2.3	23	4	2	2	8
		(20.9)	(3.6)	(1.9)	(2.7)	(9.9)
	> 2.3	7	0	1	3	5
		(9.1)	(0.4)	(1.1)	(2.3)	(3.1)
Oklawaha	≤ 2.3	5	11	1	0	3
		(5.2)	(12.0)	(1.5)	(0.2)	(1.1)
	> 2.3	13	8	6	1	0
		(12.8)	(7.0)	(5.5)	(0.8)	(1.9)
Trafford	≤ 2.3	5	11	2	1	5
		(4.4)	(12.4)	(2.1)	(0.9)	(4.2)
	> 2.3	89	7	6	3	5
		(8.6)	(5.6)	(5.9)	(3.1)	(5.8)
George	≤ 2.3	16	19	1	2	3
		(18.5)	(16.9)	(0.5)	(1.2)	(3.8)
	> 2.3	17	1	0	1	3
		(14.5)	(3.1)	(0.5)	(1.8)	(2.2)

Tabla 4.2: Valores observados y ajustados

Utilizando al pez la categoría base, la tabla (4.3) contiene los estimadores de verosimilitud máxima de los parámetros.

La ecuación que expresa al modelo logit multinomial directamente en térmi-

			Lago		
Logit	Intercepto	Tamaño ≤ 2.3	Hancock	Oklawaha	Trafford
$\log(\pi_I/\pi_P)$	-1.55	1.46	-1.66	0.94	1.12
$\log(\pi_R/\pi_P)$	-3.31	-0.35	-1.24	2.46	2.94
$\log(\pi_A/\pi_P)$	-2.09	0.63	0.70	-0.65	1.09
$\log(\pi_O/\pi_P)$	-1.90	0.33	0.83	0.01	1.52

I: invertebrado, R: reptil, A: ave, O: otro, P: pez

Tabla 4.3: Parámetros estimados en el modelo Logit para la elección de la comida

nos de probabilidades de respuesta $\{\pi_j(\mathbf{x})\}$ es

$$\pi_j(\mathbf{x}) = \frac{\exp(\alpha_j + \beta_j'\mathbf{x})}{1 + \sum_{h=1}^{J-1} \exp(\alpha_h + \beta_h'\mathbf{x})} \tag{4.2}$$

con $\alpha_j = 0$ y $\beta_J = \mathbf{0}$ y el denominador de (4.2) es el mismo para cada j.

Para la tabla (4.3), la probabilidad estimada de que un caimán en el lago Hancock sea invertebrado es

$$\begin{aligned}
\hat{\pi}_I &= \frac{e^{-1.55-1.66}}{1 + e^{-1.55-1.66} + e^{-3.31+1.24} + e^{-2.09+0.70} + e^{-1.90+0.83}} \\
&= 0.023 \\
\hat{\pi}_R &= \frac{e^{-3.31+1.24}}{1 + e^{-1.55-1.66} + e^{-3.31+1.24} + e^{-2.09+0.70} + e^{-1.90+0.83}} \\
&= 0.072 \\
\hat{\pi}_A &= \frac{e^{-2.09+0.70}}{1 + e^{-1.55-1.66} + e^{-3.31+1.24} + e^{-2.09+0.70} + e^{-1.90+0.83}} \\
&= 0.141 \\
\hat{\pi}_O &= \frac{e^{-1.90+0.83}}{1 + e^{-1.55-1.66} + e^{-3.31+1.24} + e^{-2.09+0.70} + e^{-1.90+0.83}} \\
&= 0.194 \\
\hat{\pi}_P &= \frac{e^{0}}{1 + e^{-1.55-1.66} + e^{-3.31+1.24} + e^{-2.09+0.70} + e^{-1.90+0.83}} \\
&= 0.570
\end{aligned}$$

Así las probabilidades estimadas para reptil, ave, otro y pez son 0.072, 0.141, 0.194 y 0.570, respectivamente. Recuerde que los cálculos se encuentran en el script de R con el nombre referente al ejemplo; en la carpeta 'Programas en R'.

Los modelos logit multinomial también pueden contener predictores cuantitativos; en este caso se usa el tamaño como variable dummy o ficticia para distinguir entre un caimán adulto y un adulto sub-alterno. Sin embargo la longitud actual de los caimánes fue medida y es variable cuantitativa.

4.3. Ajuste del Modelo Logit

Ajustando la verosimilitud máxima a un modelo logit multinomial se maximiza la verosimilitud sujeta a que $\{\pi_j(x)\}$ satisfaga simultáneamente las $J-1$ ecuaciones que especifica el modelo. Sea $\mathbf{y}_i = (y_{i1}, \ldots, y_{iJ})'$ con $i = 1, \ldots, n$ que representa el ensayo multinomial para el dominio i, donde $y_{ij} = 1$ cuando la respuesta es en la categoría j y $y_{ij} = 0$ en otro caso. Así, $\sum_j y_{ij} = 1$, sea $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ que denota los valores de la variable explicatoria para i y sea $\beta_j = (\beta_{j1}, \ldots, \beta_{jp})'$ que denota los parámetros para el j-ésimo logit. Puesto que $\pi_J = 1 - (\pi_1 + \ldots + \pi_{J-1})$ y $y_{iJ} = 1 - (y_{i1} + \ldots + y_{i,J-1})$, la contribución al logarítmo de la verosimilitud para i es

$$\begin{aligned}
\log\left[\prod_{j=1}^{J} \pi_j(\mathbf{x}_i)^{y_{ij}}\right] &= \sum_{j=1}^{J-1} y_{ij} \log \pi_j(\mathbf{x}_i) + \left(1 - \sum_{j=1}^{J-1} y_{ij}\right) \log\left[1 - \sum_{j=1}^{J-1} \pi_j(\mathbf{x}_i)\right] \\
&= \sum_{j=1}^{J-1} y_{ij} \log \frac{\pi_j(\mathbf{x}_i)}{1 - \sum_{j=1}^{J-1} \pi_j(\mathbf{x}_i)} + \log\left[1 - \sum_{j=1}^{J-1} \pi_j(\mathbf{x}_i)\right]
\end{aligned}$$

Así, la categoría base de los logits son los parámetros naturales para la

distribución multinomial. Ahora, suponga n observaciones independientes. En la última expresión anterior, sustituyendo $\alpha_j + \beta_j' \mathbf{x}_i$ para el logit en el primer término y $\pi_J(\mathbf{x}_i) = 1/[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j' \mathbf{x}_i)]$ en el segundo término, el logarítmo de la verosimilitud es

$$\begin{aligned}
\log \prod_{i=1}^{n} \left[\prod_{j=1}^{J} \pi_j(\mathbf{x}_i)^{y_{ij}} \right] &= \sum_{i=1}^{n} \left\{ \sum_{j=1}^{J-1} y_{ij}(\alpha_j + \beta_j' \mathbf{x}_i) - \log \left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j' \mathbf{x}_i) \right] \right\} \\
&= \sum_{j=1}^{J-1} \left[\alpha_j \left(\sum_{i=1}^{n} y_{ij} \right) + \sum_{k=1}^{p} \beta_{jk} \left(\sum_{i=1}^{n} x_{ik} y_{ij} \right) \right] \\
&\quad - \sum_{i=1}^{n} \log \left[1 + \sum_{j=1}^{J-1} \exp(\alpha_j + \beta_j' \mathbf{x}_i) \right].
\end{aligned}$$

El estadístico suficiente para β_{jk} es $\sum_i x_{ik} y_{ij}$, con $j = 1, \ldots, J-1$, $k = 1, \ldots, p$. El estadístico suficiente para α_j es $\sum_i y_{ij} = \sum_i x_{i0} y_{ij}$, con $x_{i0} = 1$; esto es el número total de resultados en la categoría j. Además, los estimadores de muestras grandes tienen distribución normal y sus errores estándar asintóticos son las raíces de los elementos de la diagonal en la inversa de la matriz de información.

4.4. Modelo Logit para respuestas ordinales

Cuando las categorías de respuesta son ordenadas, los logits pueden incorporar directamente un orden, este resultado en el modelo tiene una interpretación simple y posiblemente más potente que los modelos logits multinomiales. Así, los modelos con términos que reflejan características ordinales tales como tendencia monótona tienen modelos mejorados parsimoniosos (o con pocos parámetros) y fuertes.

4.4.1. Logit Acumulado

Una forma de usar el orden de las categorías en el logit con probabilidad acumulada es

$$P(Y \leq j|\mathbf{x}) = \pi_1(\mathbf{x}) + \ldots + \pi_j(\mathbf{x}), \qquad j = 1, \ldots, J$$

Las probabilidades acumuladas reflejan el orden, con $P(Y \leq 1) \leq P(Y \leq 2) \leq \ldots \leq P(Y \leq J) = 1$, el modelo de probabilidades acumuladas no utiliza $P(Y \leq J)$, puesto que es necesariamente igual a 1. Entonces, el logit acumulado para las primeras $J-1$ probabilidades acumuladas es definido como

$$\begin{aligned} \text{logit}[P(Y \leq j|\mathbf{x})] &= \log \frac{P(Y \leq j|\mathbf{x})}{1 - P(Y \leq j|\mathbf{x})} \\ &= \log \frac{\pi_1(\mathbf{x}) + \ldots + \pi_j(\mathbf{x})}{\pi_{j+1}(\mathbf{x}) + \ldots + \pi_J(\mathbf{x})} \qquad j = 1, \ldots, J-1 \end{aligned} \quad (4.3)$$

donde cada logit acumulado utiliza todas las J categorías. Un modelo logit$[P(Y \leq j)]$ solo es un modelo logit ordinario para respuestas binarias en el cual las categorías 1 a la j forman un resultado y las categorías $j+1$ a la J forman el segundo. Teniendo que en los modelos se pueden usar todos los $J-1$ logit acumulados en un modelo simple que tenga pocos parámetros para mayor facilidad y manejo.

4.4.2. Modelo de Oportunidades Proporcional

Un modelo que simultáneamente utiliza todos los logit acumulados es

$$\text{logit}[P(Y \leq j|\mathbf{x})] = \alpha_j + \beta'\mathbf{x}, \qquad j = 1, \ldots, J-1 \quad (4.4)$$

Donde cada logit acumulado tiene un intercepto. Los $\{\alpha_j\}$ son incrementos en j, dado que $P(Y \leq j|\mathbf{x})$ incrementa en j para un $\mathbf{x}$ fijo y el logit tiene un incremento

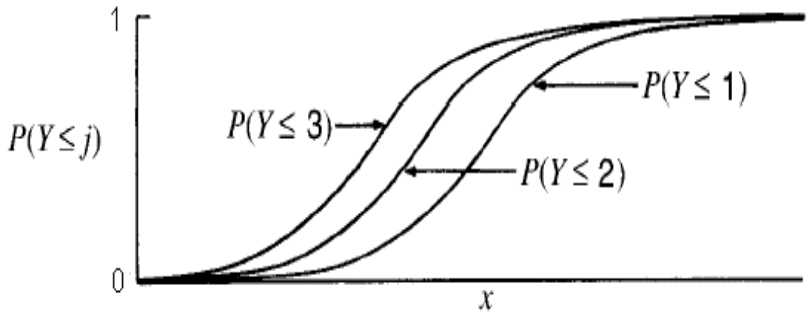

Figura 4.1: Representación de las probabilidades acumuladas en el modelo de oportunidades proporcional.

en la función de esta probabilidad.

El modelo tiene el mismo efecto β para cada logit. Para un predictor continuo x, la figura (4.1) describe el modelo cuando $J = 4$. Para un j fijo la curva de respuesta es una curva de regresión logística para respuesta binaria con resultado $Y \leq j$ y $Y > j$. La curva de respuesta para $j = 1, 2 \text{ y } 3$ tienen la misma forma, comparten exactamente la misma razón de incremento o decremento pero cambian de sitio horizontalmente. Para $j < k$, la curva para $P(Y \leq k)$ es la curva para $P(Y \leq j)$ trasladada $(\alpha_k - \alpha_j)/\beta$ unidades en el eje de las $x's$, esto es

$$P(Y \leq k|X = x) \quad = \quad P(Y \leq j|X = x + (\alpha_k - \alpha_j)/\beta)$$

El modelo logit acumulado (4.4) satisface

$$\begin{aligned} \text{logit}\left[P(Y \leq j|\mathbf{x}_1)\right] - \text{logit}\left[P(Y \leq j|\mathbf{x}_2)\right] &= \log \frac{P(Y \leq j|\mathbf{x}_1)/P(Y > j|\mathbf{x}_1)}{P(Y \leq j|\mathbf{x}_2)/P(Y > j|\mathbf{x}_2)} \\ &= \beta'(\mathbf{x}_1 - \mathbf{x}_2) \end{aligned}$$

La probabilidad acumulada de la razón de oportunidades es llamada *razón de oportunidades acumulada*. El logarítmo de la razón de oportunidades acumulada es pro-

porcional a la distancia entre $\mathbf{x}_1$ y $\mathbf{x}_2$, es decir $\beta'(\mathbf{x}_1 - \mathbf{x}_2)$, y ésta misma constante de proporcionalidad β' se aplica a cada logit. A ésta propiedad, McCullagh (1980) llamó a (4.4) modelo de oportunidades proporcional.

El modelo (4.4) restringe las $J-1$ curvas de respuesta que tienen la misma forma. Así, el ajuste no es de la misma forma como se tiene, ajustando por separado cada modelo logit para cada j. De nuevo sea $(y_{i1}, \ldots, y_{iJ})$ un indicador de respuesta binaria para cada i, la función de verosimilitud es

$$\begin{aligned}\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\pi_j(\mathbf{x}_i)^{y_{ij}}\right] &= \prod_{i=1}^{n}\left[\prod_{j=1}^{J}(P(Y\leq j|\mathbf{x}_i)-P(Y\leq j-1|\mathbf{x}_i))^{y_{ij}}\right] \\ &= \prod_{i=1}^{n}\left[\prod_{j=1}^{J}\left(\frac{\exp(\alpha_j+\beta'\mathbf{x}_i)}{1+\exp(\alpha_j+\beta'\mathbf{x}_i)}-\frac{\exp(\alpha_{j-1}+\beta'\mathbf{x}_i)}{1+\exp(\alpha_{j-1}+\beta'\mathbf{x}_i)}\right)^{y_{ij}}\right]\end{aligned} \tag{4.5}$$

vista como una función de $(\{\alpha_j\}, \beta)$, McCullagh (1989).

Se estudian los dos conceptos anteriores con un ejemplo.

Ejemplo 4.2. En la tabla (4.4) se tiene un estudio de salud mental para una muestra aleatoria de adultos residentes de Alachua Country, Florida, que relaciona el deterioro mental con dos variables explicativas. El deterioro mental es una respuesta ordinal con categorías (bueno, formación del síntoma leve, formación del síntoma moderado, deteriorado).

Los eventos de vida con índice x_1 son una medida del número y la gravedad de eventos importantes en la vida, como lo puede ser cumpleaños de un hijo, nuevo trabajo, divorcio o muertes familiares que le ocurrieron al sujeto en los 3 años pasados. La condición socioeconómica ($x_2 = SES$) es medida como respuesta binaria

Tabla 4.4: Deterioro mental

Sujeto	Deterioro Mental	SES^1 x_2	Eventos de Vida x_1	Sujeto	Deterioro Mental	SES^1 x_2	Eventos de Vida x_1
1	Bueno	1	1	21	Leve	1	9
2	Bueno	1	9	22	Leve	0	3
3	Bueno	1	4	23	Leve	1	3
4	Bueno	1	3	24	Leve	1	1
5	Bueno	0	2	25	Moderado	0	0
6	Bueno	1	0	26	Moderado	1	4
7	Bueno	0	1	27	Moderado	0	3
8	Bueno	1	3	28	Moderado	0	9
9	Bueno	1	3	29	Moderado	1	6
10	Bueno	1	7	30	Moderado	0	4
11	Bueno	0	1	31	Moderado	0	3
12	Bueno	0	2	32	Deteriorado	1	8
13	Leve	1	5	33	Deteriorado	1	2
14	Leve	0	6	34	Deteriorado	1	7
15	Leve	1	3	35	Deteriorado	0	5
16	Leve	0	1	36	Deteriorado	0	4
17	Leve	1	8	37	Deteriorado	0	4
18	Leve	1	2	38	Deteriorado	1	8
19	Leve	0	5	39	Deteriorado	0	8
20	Leve	1	5	40	Deteriorado	0	9

[1] 0: bajo, 1: alto

(1 = alto, 0 = bajo).

Solución. El efecto principal del modelo de la forma (4.4) es

$$\text{logit}[P(Y \leq j|\mathbf{x})] = \alpha_j + \beta_1 x_1 + \beta_2 x_2$$

En la tabla (4.5) se muestra el ajuste del modelo logit acumulado para los datos analizados. Con $J = 4$ categorías de respuesta, el modelo tiene tres interceptos $\{\alpha_j\}$, estos parámetros producen logits estimados y por lo tanto la estimación de $P(Y \leq j)$, $P(Y > j)$ o $P(Y = j)$. Por ejemplo, para la media de los eventos de vida

x_1 y condición socioeconómica (SES) baja $x_2 = 0$, y puesto que $\hat{\alpha}_1 = -0.282$, la probabilidad estimada de una respuesta binaria es

$$\begin{aligned}
\hat{P}(Y = 1) &= \hat{P}(Y \leq 1) \\
&= \frac{\exp(-0.282 - 0.319(4.275))}{1 + \exp(-0.282 - 0.319(4.275))} \\
&= \frac{0.192872679}{1.192872679} \\
&= 0.16
\end{aligned}$$

	Coeficientes de Regresión	
Parámetro	Estimado	Error Std.
Intercepto 0	-0.2819	0.6423
Intercepto 1	1.2128	0.6607
Intercepto 2	2.2094	0.7210
ses	1.1112	0.6109
vida	-0.3189	0.1210

Tabla 4.5: Ajuste del modelo logit de la tabla (4.4)

Para el ajuste en R se introduce la matriz de los datos analizados en la función 'logitacu'.

Los efectos estimados $\hat{\beta}_1 = -0.319$ y $\hat{\beta}_2 = 1.111$ sugieren que la probabilidad acumulada disminuye conforme los eventos de vida incrementan y aumenta con el nivel más alto de la condición socioeconómica (SES). Dada la notación de los eventos de vida, en el nivel más alto de SES las oportunidades estimadas de deterioro mental bajo cualquier nivel fijo son $e^{1.111} = 3.03$ veces las oportunidades estimadas en el nivel bajo de SES. La descripción de los efectos puede comparar probabilidades acumuladas más que razones de oportunidades.

La salida de la tabla (7.6) tomada del script generado en R, donde se introduce la matriz con las columnas deterioro, ses y vida en la función 'logitacu', presenta los cálculos para la prueba de oportunidades proporcional. En ésta prueba los efectos son los mismos para cada logit acumulado contra la alternativa de efectos distintos, comparando el modelo con un parámetro para x_1 y uno para x_2 o modelos más complejos con tres parámetros para cada uno, permitiendo efectos diferentes para logit$[P(Y \leq 1), P(Y \leq 2)]$, y logit$[P(Y \leq 3)]$. El estadístico es igual a 2.33, con 4 grados de libertad, puesto que el modelo más complejo tiene cuatro parámetros adicionales, pero el modelo más complejo no tiene un ajuste significativamente mejor (P = 0.68).

Otras Distribuciones de Estudio

En este trabajo de tesis se han presentado sólo ciertos modelos categóricos más aún existen otros modelos empleados para analizar datos del tipo categóricos como lo es el modelo logit anidado y modelo probit condicional, además de que en el capítulo de los modelos lineales generalizados exiten otras distribuciones de la familia exponencial que no se discutieron en este proyecto como lo es la Gamma y la distribución Gaussiana Inversa.

LITERATURA CITADA

[1] Agresti, A., 1996. An Introduction to Categorical Data Analysis. New York: Wiley.

[2] Agresti, A., 2002. Categorical Data Analysis. New York: Wiley.

[3] Bishop, Y. M. M., S. E. Fienberg, y P.W. Holland. 1975. Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: MIT Press.

[4] Rojas, M. E., Cavazos, R. 2007. Ejemplos sobre el Análisis de Datos Categóricos. Tesis de Maestría en Estadística Experimental. UAAAN.

[5] Fahrmeir, L., G. Tutz. 2001. Multivariate Statistical Modelling Based on Generalized Linear Models. New York: Springer.

[6] Greene, W. H. 1999. Análisis Econométrico. Madrid: Prentice-Hall.

[7] Kleinbaum, D. G., L. L. Kupper, K. E. Muller, A. Nizam. 1998. Applied Regression Analysis and Other Multivariable Methods. 3rd ed. Duxbury Press.

[8] McCullagh, P., J. A. Nelder. 1989. Generalized Linear Models, 2nd ed. London: Chapman & Hall.

[9] Medina, E. 2003. Modelos de elección discreta. (www.eva.medinaam.es).

[10] Neter, J., M. H. Kutner, C. J. Nachtsheim, W. Wasserman. 1996. Applied Linear Statistical Models, 4nd ed. McGraw-Hill.

[11] Wackerly, D. D., W. Mendenhall III, R. L. Scheaffer. 2002. Estadística Matemática con Aplicaciones. Thomson.

APÉNDICE A

PAQUETES EN R

A.1. ¿Qué es un Paquete en R?

Un paquete es una colección de funciones que realizan ciertas tareas específicas y que han sido construidas por diversos usuarios de R. La razón por la cual se crea un paquete en R es para tener un código eficiente que asegure los cálculos requeridos para ciertos trabajos o ejemplos.

A.2. Instrucciones para Construir un Paquete en R

1.- Abrir una nueva sesión de R con el propósito de que no haya elementos en la memoria.

2.- Abrir el archivo "nombre_del_paquete.R"de la siguiente manera:

Desde el menú: Archivo / Interpretar código fuente R... / Explorar la ubicación del

archivo señalado y abrirlo. Con esto, todas las funciones y conjuntos de datos están cargados en la memoria de R.

3.- Una vez hecho lo anterior, crear una carpeta llamada "Rpackages"en la unidad "C", de modo que la carpeta tenga la ubicación: "C:\Rpackages"(en esta carpeta quedará almacenado el esqueleto del paquete).

4.- El esqueleto del paquete se crea mediante la instrucción:

package.skeleton(name = "nombre_del_paquete", path = "C://Rpackages")

De esta manera, se ha creado la estructura del paquete, para su posterior compilación, la cual se encuentra almacenada en la carpeta "C:\Rpackages\nombre_del_paquete".

A continuación se deben retocar algunos archivos que están contenidos en la carpeta "C:\Rpackages\nombre_del_paquete", como por ejemplo, los archivos de ayuda. Los archivos de ayuda se encuentran en la carpeta "C:\Rpackages\nombre_del_paquete\man", los cuales tienen extensión *.Rd.

Existe un archivo .Rd por cada función considerada en el paquete, es decir, cada función debe tener un archivo de ayuda.

Observación: Para obtener una versión compilada del paquete y lista para su utilización, el paso de creación del manual de ayuda puede omitirse (ya que la creación del manual de ayuda puede tomar mucho tiempo).

5.- Para compilar el paquete, el sistema operativo Windows XP debe contar con una serie de programas que interactúan con un entorno del sistema operativo Unix (necesarios para generar un paquete para Microsoft Windows XP). Estos programas son:

1.- cygwin

2.- MinGW

3.- ActivePerl

4.- HTML Help Workshop

5.- Inno Setup 5

Los cuales se pueden obtener fácilmente desde Internet. Una vez descargados e instalados todos estos programas, se debe configurar el "path"del sistema operativo, esto de hace de la siguiente manera:

1.- Ir a "Panel de control"

2.- Ir a "Sistema", en la pestaña "Opciones avanzadas"

2.1.- Seleccionar: "Variables de entorno"

2.2.- Seleccionar: "Variables de usuario"

2.3.- Seleccionar "Nueva"y completar lo siguiente:

Nombre de variable: TMPDIR

Valor de variable: C:\WINDOWS\Temp

Luego, en "Variables de Sistema", seleccionar la variable "Path"y luego "Modificar", ahora debe incorporar lo siguiente:

C:\Perl\bin;C:\cygwin\bin;C:\MinGW\bin;C:\HTML Help Workshop;C:\Archivos de programa\R\R-2.6.1\bin;C:\Archivos de programa\R;

Note que todos los elementos del "path"deben estar separados por un punto y coma (;). Además las rutas de los archivos deben ser adecuadas a las de cada usuario.
Luego presionar "aceptar" hasta que se cierren todas las ventanas, luego se debe de reiniciar el equipo.
Una vez instalados todos lo programas y hecha la configuración anterior, se tiene el PC listo para compilar paquetes de R.

A.3. Instrucciones para Compilar un Paquete en R

Una vez instalados todos los programas, habiendo hecho el esqueleto del paquete y estando el PC configurado para compilar un paquete R, se debe hacer lo siguiente:

1.- Abrir una ventana de DOS de la siguiente manera:
Ir a "Inicio", "Ejecutar"y en abrir, escribir: cmd y luego aceptar. Con esto aparecerá la pantalla negra de DOS.

2.- Luego escribir: cd.. (cd punto punto, sin espacios y en minúsculas) y presionar

"enter". Repetir lo anterior hasta que aparezca la ubicación: C:\>

3.- Una vez que aparezca sólo el símbolo: C:\> escribir: cd Rpackages y presionar "enter". Aparecerá escrito: C:\Rpackages

Luego para compilar el paquete, se debe teclear lo siguiente:

RCMD build –binary nombre_del_paquete

Y presionar "enter". Así, en la carpeta Rpaquetes habrá un archivo zip listo para instalarlo en R.

Adicionalmente, se debe escribir el comando:

RCMD build nombre_del_paquete

Con el cual se crea otro archivo compilado (.tar) que se exige también al someter este tipo de paquetes.