UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2018, Vol. 11, No. 2, pp. 153–170

Regresión Logística Bivariable para Tablas de Contingencia Usando Metodología GSK

Bivariable Logistic Regression for Contingency Tables Via GSK

Kelly Johana Henao Zuluaga[a]
kjhenao@unal.edu.co

Juan Carlos Correa Morales[b]
jccorrea@unal.edu.co

Resumen

La regresión logística bivariable considera una respuesta que contiene dos variables dicótomas que son explicadas por un conjunto de variables y que permite tener en cuenta el nivel de asociación que existe entre las variables dicótomas a diferencia de los modelos marginales usualmente utilizados. Se desarrolla la metodología para la estimación de modelos logísticos bivariables para datos en tablas de contingencia utilizando metodología GSK. Esto incluye tanto la estimación del modelo como la evaluación de la calidad del modelo la cual incluye pruebas de hipótesis.

***Palabras clave*:** Regresión logística, Variable categórica, respuesta binaria, tablas de contingencia, metodología GSK.

Abstract

The bivariable logistic regression takes into account an answer based on two dichotomous variables, that are explained by a set of variables and that also allows to bear in mind the existing level of association between the dichotomous variables, in contrast to the marginal models commonly used. This paper will be developed under the estimation of bivariable logistic models for data in contingency tables using the GSK methodology. This covers both the model estimation and the quality evaluation model which includes hypothesis testing.

***Keywords*:** Logistic regression, categorical variables, binary outcome, contingency tables, GSK methodology.

1. Introducción

La regresión logística bivariable considera una respuesta bivariable que está conformada por dos variables dicótomas y que pueden ser explicadas por un conjunto

[a]Magíster en Ciencias-Estadística, Universidad Nacional de Colombia, sede Medellín
[b]Profesor asociado, Escuela de Estadística, Universidad Nacional de Colombia, sede Medellín

de covariables. Los investigadores usualmente ajustan dos modelos marginales los cuales no tienen en cuenta la asociación que existe entre las variables dicótomas dependientes como es el caso de McDonald (McDonald, 1993). Para el caso donde se modelan individuos vía regresión logística, existe un extenso desarrollo (Agresti, 2013). Para el análisis de estos modelos se utiliza principalmente la máxima verosimilitud para la realización de inferencias . Existen otros métodos con propiedades asintóticas similares, uno de estos métodos es el de mínimos cuadrados ponderados. En el caso de la regresión logóstica bivariable cuando las variables explicativas son continuas y los datos vienen en forma individual, el método de estimación basado en máxima verosimilitud ha sido desarrollado por varios autores (Schmidt and Strauss, 1975), pero esto exige que el investigador establezca la forma de asociación entre ellas, por ejemplo una correlación lineal que se estima mediante un coeficiente tetracórico, lo cual exige ordinalidad en ambas variables (Le Cessie y Van Houwelingen, 1994). Una posible solución aún no explorada estaría basada en cópulas, pero tiene la desventaja de la no unicidad en la cópula para una estructura específica (Genest y Neslehova, 2007).

Para la modelación de tablas de contingencia existe una metodología de carácter general conocida como GSK. Esta es una metodología para el análisis de tablas de conteo, la cual puede ser mucho más fácil de aplicar que los métodos de máxima verosimilitud. El método GSK permite respuestas correlacionadas y no requiere varianza constante, además los cálculos tienen una forma estándar fácil de aplicar para gran cantidad de modelos (Stokes et al,2012).

En este trabajo se propone el desarrollo de la metodología para la estimación de los parámetros de los modelos logísticos bivariables para tablas de contingencia utilizando metodología GSK, incluyendola parte inferencial del modelo (pruebas de hipótesis e intervalos de confianza).

2. Regresión Logística Bivariable para Tablas de Contingencia usando Metodología GSK

Sean Y_1 y Y_2 dos variables respuesta dicótomas, es decir, asumen solo dos posiles valores que corresponden a asignaciones arbitrarias de una respuesta cualitativa. Se puede obtener así la Tabla 1, una tabla 2×2 que relacionan a estas variables respusta con dos posibles respuestas una correspondiente al fracaso, 0 y otra correspondiente a éxito, 1. Cada celda de la tabla representa la probabilidad de ocurrencia de dos de las categorías, una categoría de Y_1 y una categoría de Y_2 simultáneamente que se denota como π_{ql}, donde q se refiere a la clasificación en la tabla con respecto a la variable Y_2 y l se refiere se refiere a la clasificación con respecto a la variable Y_1. Si q y l toman un valor de 2 se refiere a un éxito y si toman valor de 1 se refiere a fracaso.

Un analista de datos puede estar interesado en modelar a Y_1 y Y_2, y determinar la relación que existe entre estas variables bajo unas condiciones específicas, que son

		Y_1		
		0	1	
Y_2	0	$\pi_{11}^{(i)}$	$\pi_{12}^{(i)}$	$\pi_{1+}^{(i)}$
	1	$\pi_{21}^{(i)}$	$\pi_{22}^{(i)}$	$\pi_{2+}^{(i)}$
		$\pi_{+1}^{(i)}$	$\pi_{+2}^{(i)}$	1

Tabla 1: Tabla de contingencia para Y_1 y Y_2

Subpoblación	1	2	3	4	Total
1	$\pi_{11}^{(1)}$	$\pi_{12}^{(1)}$	$\pi_{21}^{(1)}$	$\pi_{22}^{(1)}$	1
2	$\pi_{11}^{(2)}$	$\pi_{12}^{(2)}$	$\pi_{21}^{(2)}$	$\pi_{22}^{(2)}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$\pi_{11}^{(i)}$	$\pi_{12}^{(i)}$	$\pi_{21}^{(i)}$	$\pi_{22}^{(i)}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	$\pi_{11}^{(I)}$	$\pi_{12}^{(I)}$	$\pi_{21}^{(I)}$	$\pi_{22}^{(I)}$	1

Tabla 2: Distribución de probabilidades para I subpoblaciones generadas.

las que determinan cierta subpoblación. Y_1 y Y_2 pueden surgir como funciones de una tabla de variables categóricas multidimensional que se puede acomodar a un esquema de muestreo multinomial mediante modelos tipo logístico, en este caso llamados logísticos bivariables, que dependen de un conjunto de covariables a su vez categóricas. Las variables respuestas Y_1 y Y_2 al ser relacionadas en una tabla de contingencia con k variables regresoras, cada una de ellas con un determinado número de categorías siendo $I = P_1 \times P_2 \times \cdots \times P_k$ las I combinaciones que corresponden a las diferentes subpoblaciones, donde P corresponde al número de categorías que tiene cada covariable y el subíndice corresponde a la covariable a la cual pertenece determinado número de categorías.

Así se tiene que la distribución de la población es conceptualizada como el producto de I subconjuntos multinomiales, donde cada subpoblaciíon tiene el mismo número de categorías y cada celda representa una combinación determinada de atributos. Se tienen así 4 categorías respuesta e I subpoblaciones. Para este caso define la probabilidad de ocurrencia de dos de las categorías en la i-ésima subpoblación como $\pi_{ql}^{(i)}$ y la distribución de probabilidades para las I subpoblaciones quedaria como muestra la Tabla 2.

La muestra tomada de la población puede ser representada como la Tabla 3 la cual tiene la misma estructura del modelo poblacional. El modelo asume que de cada subpoblación se toma una muestra aleatoria e independiente. Para el modelo poblacional se asume que cada subpoblación es independiente de las demás subpoblaciones.

A esta Tabla 3 se le ajusta un modelo multinomial para la i-ésima subpoblación

Subpoblación	1	2	3	4	Total
1	$n_{11}^{(1)}$	$n_{12}^{(1)}$	$n_{21}^{(1)}$	$n_{22}^{(1)}$	n_1
2	$n_{11}^{(2)}$	$n_{12}^{(2)}$	$n_{21}^{(2)}$	$n_{22}^{(2)}$	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$n_{11}^{(i)}$	$n_{12}^{(i)}$	$n_{21}^{(i)}$	$n_{22}^{(i)}$	n_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
I	$n_{11}^{(I)}$	$n_{12}^{(I)}$	$n_{21}^{(I)}$	$n_{22}^{(I)}$	n_I

Tabla 3: Tabla de muestras

con la función de masa de probabilidad

$$P\left(n_{11}^{(i)}, n_{12}^{(i)}, n_{21}^{(i)}, n_{22}^{(i)} | \pi_{11}^{(i)}, \pi_{12}^{(i)}, \pi_{21}^{(i)}, \pi_{22}^{(i)}\right) = \frac{n_i}{n_{11}^{(i)}! n_{12}^{(i)}! n_{21}^{(i)}! n_{22}^{(i)}!} {\pi_{11}^{(i)}}^{n_{11}^{(i)}} {\pi_{12}^{(i)}}^{n_{12}^{(i)}} {\pi_{21}^{(i)}}^{n_{21}^{(i)}} {\pi_{22}^{(i)}}^{n_{22}^{(i)}}$$

Las n_i observaciones fijas de Y, que es la variable respuesta, en la subpoblación i de X, que es el conjunto de variables explicativas, tienen distribución de probabilidad $\left[\pi_{11}^{(i)}, \pi_{12}^{(i)}, \pi_{21}^{(i)}, \pi_{22}^{(i)}\right] = \left[\pi_{00}^{(i)}, \pi_{01}^{(i)}, \pi_{10}^{(i)}, \pi_{11}^{(i)}\right]$, los conteos $n_{ql}^{(i)}$ con $q = 1, 2$ y $l = 1, 2$ satisfacen $n_{11}^{(i)} + n_{12}^{(i)} + n_{21}^{(i)} + n_{22}^{(i)} = n_i$. En el proceso de estimación se requieren muestras independientes de cada subpoblació de tamaño suficientemente grande para garantizar aproximaciones asintóticas.

3. Definición del modelo

El modelo lineal paramétrico que relaciona las variables de interés con un conjunto de covariables para cada subpoblación es $\hat{f} = X\beta + \epsilon$, donde $\hat{f}$ es la función respuesta muestral. Como se tiene un modelo con dos respuestas, así f esta dado por:

$$f' = \left[f_1^{(1)}, f_2^{(1)}, f_1^{(2)}, f_2^{(2)}, \ldots, f_1^{(i)}, f_2^{(i)}, \ldots, f_1^{(I)}, f_2^{(I)}, \right]$$

donde $f_1^{(i)}$ y $f_2^{(i)}$ son las dos funciones respuestas generadas para la i-ísima subpoblación que van a ser modeladas linealmente. Así para el modelo logistico tenemos:

$$f_1^{(i)} = log\left(\frac{\pi_1^{(i)}}{1 - \pi_1^{(i)}}\right) = logit\left(\pi_1^{(i)}\right)$$

$$f_2^{(i)} = log\left(\frac{\pi_2^{(i)}}{1 - \pi_2^{(i)}}\right) = logit\left(\pi_2^{(i)}\right)$$

donde π_1 y π_2 son funciones que tienen una distribución Bernoulli con probabilidades de éxito distintas, entonces:

$$\pi_1(X) = P(W_1 = 1|X) \qquad \pi_2(X) = P(W_2 = 1|X)$$

Así vemos que W_1 y W_2 son dos nuevas variables condicionadas a X que surgen permitiendo modelar la probabilidad de éxito de un evento especifico relacionado con mi tabla de contingencia. No son las probabilidades directas de la tabla de contingencia.

3.1. Matriz de Diseño para el Modelo General

Se debe considerar el modelo:

$$f = [O \circ X]\,\beta + \epsilon$$

donde O es una matriz con la siguiente estructura para cada una de sus columnas, el cual hace referencia a la primera o a la segunda respuesta:

$$[0, 1, 0, 1, \ldots, 0, 1,]'$$

Entonces la matriz O tendrá la forma:

$$O = \begin{bmatrix} 0 & 0 & \ldots & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 1 & \ldots & 1 \\ 0 & 0 & \ldots & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 1 & \ldots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & \ldots & 0 \\ 1 & 1 & \ldots & 1 & \ldots & 1 \end{bmatrix}$$

La operación $O \circ X$ corresponde al producto Hadamard entre la matriz O y la matriz X. La matriz X es una matriz de diseño tiene en cuenta los parametros tanto para f_1 como para f_2, esta matriz tiene dimensiones $_{2I \times \left(1 + P_1^{(1)} - 1 + \cdots + P_k^{(2,k)} - 1\right)}$, β es un vector de parámetros desconocidos de dimensión $_{\left(1 + P_1^{(1)} - 1 + \cdots + P_k^{(2,k)} - 1\right) \times 1}$ y ϵ se distribuye aproximadamente normal.
Para cada nivel de la variable X definimos una variable indicadora, es decir una variable que toma el valor 1 para denotar la presencia de un atributo cualitativo y usa el valor 0 para denotar la ausencia de este atributo, es decir si el nivel de la categoría considerada es o no observada en una subpoblación (Dutta, 1975). Sea la variable $z_r^{(k)}$ donde r es la r-ésima categoría de una variable X con $r = 1, \ldots, P_k - 1$ y k se refiere a la k-ésima varible X tal que:

$$z_r^{(k)} = \begin{cases} 0, & \text{si el nivel } r \text{ para la covariable } k \text{ no es observado} \\ 1, & \text{si el nivel } r \text{ para la covariable } k \text{ es observado} \end{cases}$$

Los coeficientes del modelo para la variable indicadora $z_r^{(k)}$ están dados por $\beta_r^{(k)}$, donde k hace referencia a la covariable X_k y r a la r-ésima categoría de esta covariable y el primer componente del vetor es β_0. Se obtiene el vector de parámetros:

$$\beta' = \left[\beta_0, \beta_1^{(1)}, \ldots, \beta_{P_1-1}^{(1)}, \ldots, \beta_1^{(k-1)}, \beta_2^{(k-1)}, \ldots \beta_{P_{k-1}-1}^{(k-1)}, \ldots, \beta_1^{(k)}, \beta_2^{(k)}, \ldots, \beta_{P_k-1}^{(k)}\right]$$

El vector de error ϵ es

$$\epsilon' = \left[\epsilon_1^{(1)}, \epsilon_2^{(1)}, \ldots, \epsilon_1^{(i)}, \epsilon_2^{(i)}, \ldots, \epsilon_1^{(I)}, \epsilon_2^{(I)}\right]$$

4. Estimación del Modelo vía GSK

La distribución de probabilidades en la tabla de contingencia puede ser representada como un vector donde, los primeros 4 componentes corresponden a la primera subpoblación y así hasta los últimos 4 componentes que corresponden a la última subpoblación I. Este vector permite bajo ciertas operaciones generar la función respuesta que modela las variables categóricas bajo el enfoque GSK, así se tiene que:

$$\pi' = \left[\pi_{11}^{(1)}, \pi_{12}^{(1)}, \pi_{21}^{(1)}, \pi_{22}^{(1)}, \ldots, \pi_{11}^{(i)}, \pi_{12}^{(i)}, \pi_{21}^{(i)}, \pi_{22}^{(i)}, \ldots, \pi_{11}^{(I)}, \pi_{12}^{(I)}, \pi_{21}^{(I)}, \pi_{22}^{(I)}\right]$$

Para definir las probabilidades para cada subpoblación en términos de $\pi_1(x)$ y $\pi_2(x)$ el vector π se multiplica por una matriz A de dimensiones $4I \times 4I$ que es una matriz en bloque y cuya estructura dependerá del problema que se está abordando. El resultado es un vector $4I \times 1$ en el cual cada 4 componentes aparecen las probabilidades de éxito y fracaso en términos de $\pi_1(x)$ y $\pi_2(x)$ para cada estrato o subpoblación:

$$(A\pi)' = \left[\pi_1^{(1)}, 1-\pi_1^{(1)}, \ldots, \pi_1^{(i)}, 1-\pi_1^{(i)}, \ldots, \pi_1^{(I)}, 1-\pi_1^{(I)}, \pi_2^{(I)}, 1-\pi_2^{(I)}\right]$$

Tomando el logaritmo se genera el vector columna Q de dimensión $4I$:

$$Q' = \left[\log\left(\pi_1^{(1)}\right), \log\left(1-\pi_1^{(1)}\right), \ldots, \log\left(\pi_2^{(I)}\right), \log\left(1-\pi_2^{(I)}\right)\right]$$

Sea la matriz K una matriz de dimensiones $2I \times 4I$:

$$K = \begin{bmatrix} 1 & -1 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 1 & -1 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 & -1 & \ldots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & 1 & -1 \end{bmatrix}$$

La matriz K premultipicando a Q genera el vector respuesta f:

$$KQ = \begin{bmatrix} \log\left(\frac{\pi_1^{(1)}}{1-\pi_1^{(1)}}\right) \\ \log\left(\frac{\pi_2^{(1)}}{1-\pi_2^{(1)}}\right) \\ \cdots \\ \log\left(\frac{\pi_1^{(i)}}{1-\pi_1^{(i)}}\right) \\ \log\left(\frac{\pi_2^{(i)}}{1-\pi_2^{(i)}}\right) \\ \cdots \\ \log\left(\frac{\pi_1^{(I)}}{1-\pi_1^{(I)}}\right) \\ \log\left(\frac{\pi_2^{(I)}}{1-\pi_2^{(I)}}\right) \end{bmatrix} = \begin{bmatrix} f_1^{(1)} \\ f_2^{(1)} \\ \vdots \\ f_1^{(i)} \\ f_2^{(i)} \\ \vdots \\ f_1^{(I)} \\ f_2^{(I)} \end{bmatrix}$$

Un estimador de máxima verosimilitud para π es $\hat{\pi}$, donde:

$$\hat{\pi}_{11}^{(i)} = \frac{n_{11}^{(i)}}{n_i}, \quad \hat{\pi}_{12}^{(i)} = \frac{n_{12}^{(i)}}{n_i}, \quad \hat{\pi}_{21}^{(i)} = \frac{n_{21}^{(i)}}{n_i} \quad y \quad \hat{\pi}_{22}^{(i)} = \frac{n_{22}^{(i)}}{n_i}$$

Con $n_{11}^{(i)}, n_{12}^{(i)}, n_{21}^{(i)} \quad y \quad n_{22}^{(i)}$ frecuencias observadas en la i-ésima subpoblación y n_i el total de observaciones de la i-ésima subpoblacón (Grizzle et al.,1969).

$$\hat{\pi}' = \left[\hat{\pi}_{11}^{(1)}, \hat{\pi}_{12}^{(1)}, \hat{\pi}_{21}^{(1)}, \hat{\pi}_{22}^{(1)}, \ldots, \hat{\pi}_{11}^{(i)}, \hat{\pi}_{12}^{(i)}, \hat{\pi}_{21}^{(i)}, \hat{\pi}_{22}^{(i)}, \ldots, \hat{\pi}_{11}^{(I)}, \hat{\pi}_{12}^{(I)}, \hat{\pi}_{21}^{(I)}, \hat{\pi}_{22}^{(I)}\right]$$

$$\hat{\pi}' = \left[\frac{n_{11}^{(1)}}{n_1}, \frac{n_{12}^{(1)}}{n_1}, \frac{n_{21}^{(1)}}{n_1}, \frac{n_{22}^{(1)}}{n_1}, \ldots, \frac{n_{11}^{(i)}}{n_i}, \frac{n_{12}^{(i)}}{n_i}, \frac{n_{21}^{(i)}}{n_i}, \frac{n_{22}^{(i)}}{n_i}, \ldots, \frac{n_{11}^{(I)}}{n_I}, \frac{n_{12}^{(I)}}{n_I}, \frac{n_{21}^{(I)}}{n_I}, \frac{n_{22}^{(I)}}{n_I}\right]$$

Los valores esperados de los estimadores para la i-ésima subpoblacién son:

$$E\left(\hat{\pi}_{11}^{(i)}\right) = \pi_{11}^{(i)}, \quad E\left(\hat{\pi}_{12}^{(i)}\right) = \pi_{12}^{(i)}, \quad E\left(\hat{\pi}_{21}^{(i)}\right) = \pi_{21}^{(i)} \quad y \quad E\left(\hat{\pi}_{22}^{(i)}\right) = \pi_{22}^{(i)}$$

El análisis vía GSK requiere estimar las varianzas y covarianzas de $\hat{\pi}$. Como cada subpoblacion corresponde una variable multinomial, para la i-ésima subpoblación de una muestra aleatoria de tamaño n_i se tiene que las varianzas de los estimadores en la i-ésima subpoblacién son:

$$var\left(\hat{\pi}_{11}^{(i)}\right) = \frac{1}{n_i}\pi_{11}^{(i)}\left(1-\pi_{11}^{(i)}\right), \quad var\left(\hat{\pi}_{12}^{(i)}\right) = \frac{1}{n_i}\pi_{12}^{(i)}\left(1-\pi_{12}^{(i)}\right),$$

$$var\left(\hat{\pi}_{21}^{(i)}\right) = \frac{1}{n_i}\pi_{21}^{(i)}\left(1-\pi_{21}^{(i)}\right) \quad \text{y} \quad var\left(\hat{\pi}_{22}^{(i)}\right) = \frac{1}{n_i}\pi_{22}^{(i)}\left(1-\pi_{22}^{(i)}\right)$$

En el caso de la distribución multinomial, la matriz de covarianzas no es de rango completo y frecuentemente requiere ser trabajada con una inversa generalizada (Tanabe,1992).Las covarianzas estan dadas por:

$$cov\left(\hat{\pi}_{11}^{(i)},\hat{\pi}_{12}^{(i)}\right)=\frac{1}{n_i}\left(-\pi_{11}^{(i)}\pi_{12}^{(i)}\right),\quad cov\left(\hat{\pi}_{11}^{(i)},\hat{\pi}_{21}^{(i)}\right)=\frac{1}{n_i}\left(-\pi_{11}^{(i)}\pi_{21}^{(i)}\right),$$

$$cov\left(\hat{\pi}_{11}^{(i)},\hat{\pi}_{22}^{(i)}\right)=\frac{1}{n_i}\left(-\pi_{11}^{(i)}\pi_{22}^{(i)}\right)$$

Sea

$$\hat{\pi}^{(i)\prime}=\left[\frac{n_{11}^{(i)}}{n_i},\frac{n_{12}^{(i)}}{n_i},\frac{n_{21}^{(i)}}{n_i},\frac{n_{22}^{(i)}}{n_i}\right]$$

La matriz de covarianza estimada en la i-ésima subpoblación, es decir, para $\hat{\pi}^{(i)}$:

$$Cov\left(\hat{\pi}^{(i)}\right)=\Sigma_i=\frac{1}{n_i}\begin{bmatrix}\pi_{11}^{(i)}\left(1-\pi_{11}^{(i)}\right) & -\pi_{11}^{(i)}\pi_{12}^{(i)} & -\pi_{11}^{(i)}\pi_{21}^{(i)} & -\pi_{11}^{(i)}\pi_{22}^{(i)}\\ -\pi_{11}^{(i)}\pi_{12}^{(i)} & \pi_{12}^{(i)}\left(1-\pi_{12}^{(i)}\right) & -\pi_{12}^{(i)}\pi_{21}^{(i)} & -\pi_{12}^{(i)}\pi_{22}^{(i)}\\ -\pi_{11}^{(i)}\pi_{21}^{(i)} & -\pi_{12}^{(i)}\pi_{21}^{(i)} & \pi_{21}^{(i)}\left(1-\pi_{21}^{(i)}\right) & -\pi_{21}^{(i)}\pi_{22}^{(i)}\\ -\pi_{11}^{(i)}\pi_{22}^{(i)} & -\pi_{12}^{(i)}\pi_{22}^{(i)} & -\pi_{21}^{(i)}\pi_{22}^{(i)} & \pi_{22}^{(i)}\left(1-\pi_{22}^{(i)}\right)\end{bmatrix}$$

Luego $\Sigma_{\hat{\pi}}$ es una matriz en bloque diagonal de dimensión $4I\times 4I$ con elementos cero fuera de la diagonal principal:

$$\Sigma_{\hat{\pi}}=\begin{bmatrix}\Sigma_1 & 0 & \dots & 0 & \dots & 0\\ 0 & \Sigma_2 & \dots & 0 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots\\ 0 & 0 & \dots & \Sigma_i & \dots & 0\\ \vdots & \vdots & \dots & 0 & \ddots & 0\\ 0 & 0 & \dots & 0 & \dots & \Sigma_I\end{bmatrix}$$

Se define $\hat{f}$ como el estimador del vector f

$$\hat{f}'=\left[\hat{f}_1^{(1)},\hat{f}_2^{(1)},\hat{f}_1^{(2)},\hat{f}_2^{(2)},\vdots,\hat{f}_1^{(i)},\hat{f}_2^{(i)},\vdots,\hat{f}_1^{(I)},\hat{f}_2^{(I)},\right]$$

Entonces la matriz $\Sigma_{\hat{f}}$ de covarianza de $\hat{f}$ es (Serfling, 2002)(Grizzle, 1969):

$$\Sigma_{\hat{f}}=KD_{lineal}^{-1}A\Sigma_{\hat{\pi}}A'D_{lineal}^{-1}K'$$

Para funciones logaritmicas se tiene una matriz diagonal D_{lineal} $4I\times 4I$, con su diagonal principal compuesta por a'_i,que es la i-ésima fila de A:

$$D_{lineal} = \begin{bmatrix} a'_1\pi & 0 & \dots & 0 \\ 0 & a'_2\pi & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a'_{4I}\pi \end{bmatrix}$$

La matriz de covarianza estimada del vector de probabilidades estimadas para la i-esima población es:

$$\hat{\Sigma}_i = \frac{1}{n_i} \begin{bmatrix} \hat{\pi}_{11}^{(i)}\left(1-\hat{\pi}_{11}^{(i)}\right) & -\hat{\pi}_{11}^{(i)}\hat{\pi}_{12}^{(i)} & -\hat{\pi}_{11}^{(i)}\hat{\pi}_{21}^{(i)} & -\hat{\pi}_{11}^{(i)}\hat{\pi}_{22}^{(i)} \\ -\hat{\pi}_{11}^{(i)}\hat{\pi}_{12}^{(i)} & \hat{\pi}_{12}^{(i)}\left(1-\hat{\pi}_{12}^{(i)}\right) & -\hat{\pi}_{12}^{(i)}\hat{\pi}_{21}^{(i)} & -\hat{\pi}_{12}^{(i)}\hat{\pi}_{22}^{(i)} \\ -\hat{\pi}_{11}^{(i)}\hat{\pi}_{21}^{(i)} & -\hat{\pi}_{12}^{(i)}\hat{\pi}_{21}^{(i)} & \hat{\pi}_{21}^{(i)}\left(1-\hat{\pi}_{21}^{(i)}\right) & -\hat{\pi}_{21}^{(i)}\hat{\pi}_{22}^{(i)} \\ -\hat{\pi}_{11}^{(i)}\hat{\pi}_{22}^{(i)} & -\hat{\pi}_{12}^{(i)}\hat{\pi}_{22}^{(i)} & -\hat{\pi}_{21}^{(i)}\hat{\pi}_{22}^{(i)} & \hat{\pi}_{22}^{(i)}\left(1-\hat{\pi}_{22}^{(i)}\right) \end{bmatrix}$$

La matriz de covarianzas estimada para las I subpoblaciones será:

$$\hat{\Sigma}_{\hat{\pi}} = \begin{bmatrix} \hat{\Sigma}_1 & 0 & \dots & 0 & \dots & 0 \\ 0 & \hat{\Sigma}_2 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & \hat{\Sigma}_i & \dots & 0 \\ \vdots & \vdots & \dots & 0 & \ddots & 0 \\ 0 & 0 & \dots & 0 & \dots & \hat{\Sigma}_I \end{bmatrix}$$

La matriz diagonal estimada es:

$$\hat{D}_{lineal} = \begin{bmatrix} a'_1\hat{\pi} & 0 & \dots & 0 \\ 0 & a'_2\hat{\pi} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a'_{4I}\hat{\pi} \end{bmatrix}$$

Entonces la matriz estimada $\hat{\Sigma}_{\hat{f}}$ de covarianza estimada de $\hat{f}$ es:

$$\hat{\Sigma}_{\hat{f}} = K\hat{D}_{lineal}^{-1}A\hat{\Sigma}_{\hat{\pi}}A'\hat{D}_{lineal}^{-1}K'$$

4.1. Estimación de parámetros del modelo

Para el modelo $\hat{f} = X\beta + \epsilon$, donde $Var(\epsilon) = \Sigma_{\hat{\epsilon}}$ y $\epsilon \sim AN(0, \Sigma_{\hat{\epsilon}})$ [1] el estimador vía mínimos cuadrados ponderados de β, el cual es un estimador BAN (best

[1] AN asintóticamente normal

asymptotic normal) (Rao, 2008) es,

$$\hat{\beta} = \left(X'\hat{\Sigma}_{\hat{f}}^{-1}X\right)^{-1}X'\hat{\Sigma}_{\hat{f}}^{-1}\hat{f}$$

$\hat{\beta}$ es un estimador asintóticamente insesgado de β y es el valor que minimiza la ecuación:

$$S(\beta) = \left(\hat{f} - X\beta\right)' \hat{\Sigma}_{\hat{f}}^{-1}\left(\hat{f} - X\beta\right)$$

La matriz de covarianzas de $\hat{\beta}$ es $\Sigma_{\hat{\beta}}$, y su estimación está dada por:

$$\hat{\Sigma}_{\hat{\beta}} = \left(X'\hat{\Sigma}_{\hat{f}}^{-1}X\right)^{-1}$$

La estimacion de $\hat{f}$ denotada $\hat{f}*$ esta dada por:

$$\begin{aligned}\hat{f}* &= X\hat{\beta} \\ &= X\left(X'\hat{\Sigma}_{\hat{f}}^{-1}X\right)^{-1}X'\hat{\Sigma}_{\hat{f}}^{-1}\hat{f}\end{aligned}$$

4.2. Inferencia sobre el modelo

Un intervalo de confianza para $\left(\beta_r^{(k)} + \beta_r^{(k,2)}\right)$ está dado por:

$$C\hat{\beta} \pm Z_{\frac{\alpha}{2}}\sqrt{C\hat{\Sigma}_{\hat{\beta}}C'}$$

donde $C\Sigma_{\hat{\beta}}C'$ es la varianza estimada de $C\hat{\beta}$ (Grizzle et al,1969).
Las pruebas globales para el modelo, pruebas para los coeficientes individuales, pruebas para las variables y pruebas de contraste tienen en forma general el esadistico de prueba dado por:

$$\chi^2 = \left(C\hat{\beta}\right)'\left(C\hat{\hat{\Sigma}}_{\hat{\beta}}C'\right)^{-1}C\hat{\beta} \sim \chi^2_{(g)}$$

donde C es una matriz que define la prueba de hipótesis que se quiere probar de acuerdo al caso y g corresponde al rango de la matriz C definida para la prueba de hipótesis.

4.3. Ilustración

Teniendo en cuenta la información recolectada en la encuesta de calidad de vida del municipio de Medellín del año 2011 (Departamento administrativo de Planeación Municipio de Medellin, 2012), donde se tiene información de la posesión de vehículos motorizados, es decir la posesión de moto y carro , por parte de los hogares del municipio y se quiere relacionar estas dos variables con la disponibilidad

Garaje	Estrato	$\bar{m}\bar{c}$	$m\bar{c}$	$\bar{m}c$	mc
No	1	62534	3667	491	75
	2	92104	8964	2282	514
	3	51027	6852	5192	824
	4	5516	751	2179	277
	5	1245	118	1596	170
	6	174	16	913	8
Sí	1	340	45	19	5
	2	983	153	214	41
	3	2687	499	1309	255
	4	1834	256	2213	296
	5	1283	136	3200	244
	6	235	15	1553	92

Tabla 4: Tabla de contingencia para posesión de vehículos motorizados en el municipio de Medellín según la encuesta de calidad de vida de 2011 (Departamento administrativo de plnaneación Municipio de Medellín, 2012).

de garaje y un determinado estrato socioeconómico de los hogares, tenemos así que mc corresponde a tener moto y carro, $m\bar{c}$ tener moto y no tener carro, $\bar{m}c$ no tener moto y tener carro y $\bar{m}\bar{c}$ no tener moto ni tener carro. Una subpoblación o estrato definido para este ejemplo serian los hogares de estrato 1 y que no tienen garaje y se relacionarán las variables categóricas como se muestra en la Tabla 5.

		Moto	
		No	Sí
Carro	No	62534	3667
	Sí	491	75

Tabla 5: Tabla de contingencia para posesión de vehículos motorizados en el estrato 1 y sin disponibilidad de garaje en el hogar

En la Tabla 4 se ilustra la situación para el caso del ejemplo de la posesión de vehículos motorizados en la ciudad de Medellín, donde se obtiene una tabla de contingencia $6 \times 2 \times 4$, se generan para este ejemplo 12 subpoblaciones o estratos, que corresponden a las posibles combinaciones de los diferentes niveles para diferentes categorías de las covariables seleccionadas, para este ejemplo la disponibilidad de garaje con 2 niveles y el estrato socioeconómico con 6 niveles. A partir de esta tabla se puede obtener simultáneamente información de todos los estratos socioeconómicos y de aquellos hogares que disponen o no disponen garaje, relacionado con la posesión de vehículos motorizados.

Considere que se puede estar interesado en la probabilidad de que los hogares tengan moto y la probabilidad de que los hogares tengan carro, sabiendo que Y_1 es la variable respuesta tener o no tener moto y Y_2 es la variable respuesta tener o no tener carro .

$$\pi_1(X) = P(W_1 = 1|X) = P(\text{Tener únicamente moto}|X)$$

$$\pi_2(X) = P(W_2 = 1|X) = P(\text{Tener únicamente carro}|X)$$

$$\{W_1\} \Leftrightarrow \{Y_1 = 1, Y_2 = 1\} \qquad \{W_2\} \Leftrightarrow \{Y_1 = 0, Y_2 = 1\}$$

En este caso se tiene que para la i-ésima población:

$$\pi_1^{(i)}(X) = \pi_{22}^{(i)} \qquad \pi_2^{(i)}(X) = \pi_{21}^{(i)}$$

Así vemos que W_1 y W_2 son dos nuevas variables condicionadas a X que surge permitiendo modelar la probabilidad de éxito de un evento específico relacionado con la tabla de contingencia, ya que dicho evento puede involucrar combinaciones de las categorías del par de variables Y_1 y Y_2.

Para la posesión de vehículos motorizados por parte de los hogares según la encuesta de calidad de vida del municipio de Medellín del año 2011 aparecen las covariables X_1 = estrato socioeconómico del hogar definido en los niveles 1,2, 3, 4, 5 y 6 y X_2 = disponibilidad de garaje con dos niveles: sí tiene garaje y no tiene garaje. Para cada nivel de cada covariable se define una varible indicadora, por ejemplo para la covariable estrato socioeconómico en el nivel estrato 1 se tiene:

$$z_1^{(1)} = \begin{cases} 0, & \text{Si no pertenece al estrato 1 socioeconómico.} \\ 1, & \text{Si pertenece al estrato 1 socioeconómico} \end{cases}$$

Así, cada nivel de cada covariable se convierte en una variable indicadora y se tiene como nivel de referencia a el último nivel para cada covariable, es decir para el caso de estrato socioeconómico sería el estrato 6 y para el caso de la disponibilidad de garaje en el hogar sería, el sí tiene garaje. Se genera la siguiente matriz de diseño ilustrada en la Tabla 6.

El vector de parametros β estaria dado por :

$$\beta' = \left[\beta_0\, \beta_1^{(1)}, \beta_1^{(2)}, \beta_2^{(2)}, \beta_3^{(2)}, \beta_4^{(2)}, \beta_5^{(2)}, \beta_0^{(2)}, \beta_1^{(1,2)}, \beta_1^{(2,2)}, \beta_2^{(2,2)}, \beta_3^{(2,2)}, \beta_4^{(2,2)}, \beta_5^{(2,2)}\right]$$

La matriz A es:

β_0	$z_1^{(1)}$	$z_1^{(2)}$	$z_2^{(2)}$	$z_3^{(2)}$	$z_4^{(2)}$	$z_5^{(2)}$	$\beta_0^{(,2)}$	$z_1^{(1,2)}$	$z_1^{(2,2)}$	$z_2^{(2,2)}$	$z_3^{(2,2)}$	$z_4^{(2,2)}$	$z_5^{(2,2)}$
1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1	0	0	0	0	0	0
1	0	1	0	0	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	0	1	0	1	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	1	0	0	1	0	0	0
1	0	0	0	1	0	0	0	0	0	0	0	0	0
1	0	0	0	1	0	0	1	0	0	0	1	0	0
1	0	0	0	0	1	0	0	0	0	0	0	0	0
1	0	0	0	0	1	0	1	0	0	0	0	1	0
1	0	0	0	0	0	1	0	0	0	0	0	0	0
1	0	0	0	0	0	1	1	0	0	0	0	0	1
1	1	0	0	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	1	1	0	0	0	0	0
1	1	1	0	0	0	0	0	0	0	0	0	0	0
1	1	1	0	0	0	0	1	1	1	0	0	0	0
1	1	0	1	0	0	0	0	0	0	0	0	0	0
1	1	0	1	0	0	0	1	1	0	1	0	0	0
1	1	0	0	1	0	0	0	0	0	0	0	0	0
1	1	0	0	1	0	0	1	1	0	0	1	0	0
1	1	0	0	0	1	0	0	0	0	0	0	0	0
1	1	0	0	0	1	0	1	1	0	0	0	1	0
1	1	0	0	0	0	1	0	0	0	0	0	0	0
1	1	0	0	0	0	1	1	1	0	0	0	0	1

Tabla 6: Elementos de la Matriz de diseño X para un Modelo General en el Ejemplo vehículos motorizados

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 1 & 0 & 0 \end{bmatrix}$$

$$A\pi = \begin{bmatrix} \pi_{12}^{(1)} + \pi_{22}^{(1)} \\ \pi_{11}^{(1)} + \pi_{21}^{(1)} \\ \pi_{21}^{(1)} + \pi_{22}^{(1)} \\ \pi_{11}^{(1)} + \pi_{12}^{(1)} \\ \pi_{12}^{(2)} + \pi_{22}^{(2)} \\ \pi_{11}^{(2)} + \pi_{21}^{(2)} \\ \pi_{21}^{(2)} + \pi_{22}^{(2)} \\ \pi_{11}^{(2)} + \pi_{12}^{(2)} \\ \vdots \\ \pi_{12}^{(12)} + \pi_{22}^{(12)} \\ \pi_{11}^{(12)} + \pi_{21}^{(12)} \\ \pi_{21}^{(12)} + \pi_{22}^{(12)} \\ \pi_{11}^{(12)} + \pi_{12}^{(12)} \end{bmatrix} = \begin{bmatrix} \pi_1^{(1)} \\ 1 - \pi_1^{(1)} \\ \pi_2^{(1)} \\ 1 - \pi_2^{(1)} \\ \pi_1^{(2)} \\ 1 - \pi_1^{(2)} \\ \pi_2^{(2)} \\ 1 - \pi_2^{(2)} \\ \vdots \\ \pi_1^{(12)} \\ 1 - \pi_1^{(12)} \\ \pi_2^{(12)} \\ 1 - \pi_2^{(12)} \end{bmatrix}$$

Así se tiene que para f_1 y f_2 se define una indicadora para cada nivel de cada covariable como se muestra en la Tabla 7.

Para el modelo las respuestas f_1 y f_2 en el que se determinan las probabilidades de manera independiente para cada una de las respuestas planteadas, están dadas para la i-esima subpoblación por:

$$\hat{f}_1^{(i)} = \beta_0 + \beta_1^{(1)} z_1^{(1)} + \beta_1^{(2)} z_1^{(2)} + \beta_2^{(2)} z_2^{(2)} + \beta_3^{(2)} z_3^{(2)} + \beta_4^{(2)} z_4^{(2)} + \beta_5^{(2)} z_5^{(2)} + \epsilon_1^{(i)}$$

f_1	f_2	
$Z_1^{(1)}=1$	$Z_1^{(1,2)}=1$	No tiene garaje
$Z_1^{(1)}=0$	$Z_1^{(1,2)}=0$	Otro
$Z_2^{(1)}=1$	$Z_2^{(1,2)}=1$	Sí tiene garaje
$Z_2^{(1)}=0$	$Z_2^{(1,2)}=0$	Otro
$Z_1^{(2)}=1$	$Z_1^{(2,2)}=1$	Estrato
$Z_1^{(2)}=0$	$Z_1^{(2,2)}=0$	Otro
$Z_2^{(2)}=1$	$Z_2^{(2,2)}=1$	Estrato 2
$Z_2^{(2)}=0$	$Z_2^{(2,2)}=0$	Otro
$Z_3^{(2)}=1$	$Z_3^{(2,2)}=1$	Estrato 3
$Z_3^{(2)}=0$	$Z_3^{(2,2)}=0$	Otro
$Z_4^{(2)}=1$	$Z_4^{(2,2)}=1$	Estrato 4
$Z_4^{(2)}=0$	$Z_4^{(2,2)}=0$	Otro
$Z_5^{(2)}=1$	$Z_5^{(2,2)}=1$	Estrato 5
$Z_5^{(2)}=0$	$Z_5^{(2,2)}=0$	Otro
$Z_6^{(2)}=1$	$Z_6^{(2,2)}=1$	Estrato 6
$Z_6^{(2)}=0$	$Z_6^{(2,2)}=0$	Otro

Tabla 7: Nivel de cada covariable definida como indicadora.

$$\begin{aligned}\hat{f}_2^{(i)} = &\left(\beta_0+\beta_0^{(,2)}\right)+\left(\beta_1^{(1)}+\beta_1^{(1,2)}\right)z_1^{(1,2)}+\left(\beta_1^{(2)}+\beta_1^{(2,2)}\right)z_1^{(2,2)}\\ &+\left(\beta_2^{(2)}+\beta_2^{(2,2)}\right)z_2^{(2,2)}+\left(\beta_3^{(2)}+\beta_3^{(2,2)}\right)z_3^{(2,2)}+\left(\beta_4^{(2)}+\beta_4^{(2,2)}\right)z_4^{(2,2)}\\ &+\left(\beta_5^{(2)}+\beta_5^{(2,2)}\right)z_5^{(2,2)}+\epsilon_2^{(i)}\end{aligned}$$

Las funciones se puede interpretar como a continuación, nótese que este análisis supone la aditividad lineal de los coeficientes de las variables indicadoras. Para f_1 se tiene que para el caso de no tener garaje y ser estrato cinco es:

$$E\left[f\left(\pi_1|\,X\right)\right]=\beta_0+\beta_1^{(1)}+\beta_5^{(2)} \tag{1}$$

Para f_2 no tener garaje y ser estrato cinco es:

$$E\left[f\left(\pi_2|\,X\right)\right]=\left(\beta_0+\beta_0^{(2)}\right)+\left(\beta_1^{(1)}+\beta_1^{(1,2)}\right)+\left(\beta_5^{(2)}+\beta_5^{(2,2)}\right) \tag{2}$$

El signo de los parámetros representa una influencia negativa o positiva en la variable dependiente, en este caso es la influencia sobre los logit de π_1 y π_2. Así se tiene que para f_1 según la tabla 5.13 el $\beta_4^{(2)}$ el cúal esta asociado con el estrato socioeconómico 5 no es significativo, los demás parámetros si son significativos y

Parámetro	Estimación	Intervalo de Confianza	Valor chi-cuadrado	Valor p
β_0	-2.678	-2.706 ,-2.649	32931.877	0.000
$\beta_1^{(1)}$	0.647	0.613, 0.681	1364.034	0.000
$\beta_1^{(2)}$	0.3491	0.230, 0.469	32.844	0.000
$\beta_2^{(2)}$	0.575	0.439, 0.711	68.438	0.000
$\beta_3^{(2)}$	0.730	0.668, 0.791	534.569	0.000
$\beta_4^{(2)}$	0.0615	-0.0337, 0.158	1.603	2.1e-01
$\beta_5^{(2)}$	0.319	0.287, 0.351	378.240	0.000
$\beta_0^{(,2)}$	-2.051	-2.105, -1.996	5442.006	0.000
$\beta_1^{(1,2)}$	1.865	1.808, 1.922	4004.241	0.000
$\beta_1^{(2,2)}$	4.478	4.335, 4.622	3738.7410	0.000
$\beta_2^{(2,2)}$	1.514	1.328, 1.701	250.840	0.000
$\beta_3^{(2,2)}$	3.163	3.080, 3.245	5618.144	0.000
$\beta_4^{(2,2)}$	5.533	5.413, 5.652	8221.5548	0.000
$\beta_5^{(2,2)}$	0.811	0.761, 0.861	1018.510	0.000

Tabla 8: Parámetros estimados

tienen una influencia positiva sobre la respuesta.
Para f_2 según la tabla 5.15 todos los parámetros son significativos y tienen una influencia positiva sobre la respuesta. En conclusión el tener garaje en el hogar y el estrato socioeconómico tienen una influencia positiva sobre la variable dependiente del modelo.
Con respecto a las probabilidades, se tiene que $\hat{f}_1$ y $\hat{f}_2$ son los logit de $\hat{\pi}_1$ y $\hat{\pi}_2$, la probabilidad de exito π_1 y π_2 se toman de el vector $\hat{f}$ donde:

$$E\left[f\left(\pi_1|\, X\right)\right] = \beta_0 + \beta_1^{(2)}$$

$$\hat{\pi}_2 = \frac{exp(\hat{f}_2)}{1 + exp(\hat{f}_2)}$$

Por ejemplo para la probabilidad de que el estrato 1 con garaje tenga moto sería:

$$\text{Estrato 1 + Sí garaje} \quad E\left[f\left(\pi_1|\, X\right)\right] = \beta_0 + \beta_1^{(2)}$$

La probabilidad estimada es:

$$\hat{\pi*}_1 = \frac{exp(-2.67752495)}{1 + exp(-2.67752495)} = 0.062$$

La probabilidad observada es:

$$\hat{\pi}_1 = \frac{exp(-2.77)}{1 + exp(-2.77)} = 0.066$$

5. Conclusiones

- La metodología GSK se ha utilizado para desarrollar el modelo log"istico bivariable en el caso donde se tengan variables explicativas categóricas, o sea tablas de conteos. Una ventaja de esta aproximación es que no requiere especificar la estructura de asociación entre las dos variables dependientes
- La metodología en la parte inferencial se puede desarrollar de una forma directa y simple aún en situaciones complejas tales como pruebas simultáneas.
- El método propuesto permite modelar simultáneamente parámetros que son de relevancia en ciertas areas como lo es la epidemiología, por ejemplo en el análisis de sensibilidad y especificidad permitiendo calcularlas de manera directa y no de manera marginal como si fueran independientes.

Recibido: 2017-08-02
Aceptado: 2018-12-11

Referencias

Agresti, A. (2013), *Categorical data analysis*, John Wiley & Sons, New Jersey.

Cengiz, M. (2005), ' Bayesian inference for bivariate generalized linear models in diagnosing renal arterial obstruction', *Statistical Methodology* **2**(3), 168–174.

Genest, C. y Neslehova, J. (n.d.), 'A primer on copulas for count data'.

Glonek, G.F., M. P. (1995), ' Multivariate logistic models ', *Journal of the royal statistical society. Series B (Metho dological)* **25**(3), 553–546.

Grizzle, J.E., S. C. y. K. G. (1969), 'Analysis of categorical data by linear models ', *Biometrics* **57**, 489–540.

Le Cessie, S. y Van Houwelingen, J. C. (1994), 'Logistic regression for correlated binary data ', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **43**(1), 95–108.

McDonald, B. (1993), ' Estimating logistic regression parameters for bivariate binary data ', *Journal of the Royal Statistical Society* **55**(2), 391–397.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Rao, C. R., T. H. S. H. C. y. S. M. (2008), *Linear models and generalizations. Least Squares and Alternatives*, Springer, Berlin.

Schmidt, P. y Strauss, R. P. (1975), ' Estimation of models with jointly dependent qualitative variables: a simultaneous logit approach ', *Estimation of models with jointly dependent qualitative variables: a simultaneous logit approach* **43**(4), 745–755.

Serfling, R. (2002), *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New Jersey.

Stokes, M. E., D. C. S. y. K. G. G. (2012), *Categorical data analysis using SAS*, SAS institute, North Carolina.