PLANTEAMIENTO DEL MODELO LOGÍSTICO MULTINOMIAL A TRAVÉS DE LA FUNCIÓN CANÓNICA DE ENLACE DE LA FAMILIA EXPONENCIAL

Daniel F. Osorio

Universidad del Valle, Escuela de Ingeniería Industrial y Estadística, Cali
daosorio@univalle.edu.co,

Johann A. Ospina

Universidad del Valle, Escuela de Ingeniería Industrial y Estadística, Cali
joalexos@univalle.edu.co,

Danny A. Lenis

Universidad del Valle, Escuela de Ingeniería Industrial y Estadística, Cali
danyleni@univalle.edu.co

Resumen

Este articulo pretende mostrar formalmente el planteamiento del modelo logístico multinomial a través de la función de enlace canónico de la distribución multinomial a la familia exponencial, así como también las expresiones con las cuales se estiman las probabilidades de los individuos en dicho modelo; Debido que al realizar una revisión de la literatura de algunos autores como McCullagh, Nelder, Jobson, Silva, Hosmer y Lemeshow se encuentra poca información referente a este antecedente teórico. No se abordaran los aspectos referentes a la inferencia de los parámetros del modelo, pero se realizara una ilustración de cómo se aplica este modelo utilizando una base de datos del programa estadístico SPSS 15.0®.

Palabras Claves: Modelo logístico multinomial, Función de enlace canónico, Familia exponencial.

Abstract

This article tries to show in a formal way the posing of the multinomial logistic model by means of the function canonical link of the multinomial distribution to the exponential family, as well as the expressions which stimate the probabilities of individuals in this model; owing to the fact that at the moment of making a revision of the litterature of some authors like McCullagh, Nelder, Jobson, Silva, Hosmer and Lemeshow very little information is found related to this theoric antecedent. The aspects related to the inference of the parameters of the model are not raised, but an illustration about how this model is applied will be done using a data base of the statistical software SSPSS 15.0®.

Key words: Multinomial logistic model, Function canonical link, Exponential family.

1. Introducción

El modelo de regresión logística multinomial o también conocido como modelo con respuesta politómica, es una generalización del modelo de regresión logístico binomial (Mc Cullagh, 1989) en el que se desea estimar la probabilidad de que el individuo presente o no un evento especifico, dado un conjunto de variables que explican características particulares de los individuos. En el caso del modelo multinomial, la variable endógena tiene más de dos alternativas a considerar como posibles respuestas, por lo cual la distribución de probabilidad adecuada para modelar este fenómeno es la distribución multinomial. Se debe tener en cuenta que la regresión logística multinomial difiere de la regresión logística condicional y ordinal, pues en la regresión condicional las variables explicativas hacen referencia a atributos de las alternativas, variando sus valores para cada una de ellas, mientras que pueden o no variar para cada individuo, además solo se estima un vector de parámetros, a diferencia de el caso multinomial en el que hay tantos vectores como categorías menos una. Por otra parte en la regresión ordinal, las estimaciones no se pueden realizar con los modelos mencionados anteriormente, ya que esta tiene como particularidad el uso de la información adicional suministrado por el orden de las categorías de la variable respuesta.

Ahora bien, la variable respuesta del modelo de regresión logística multinomial es una variable aleatoria con distribución multinomial, que se puede considerar como el numero de éxitos en cada una de las *g* categorías que se presentan en *n* ensayos independientes, y su función de distribución viene dada por (Feller, 1967).

$$P(Y_1 = n_1, \dots, Y_{\mathrm{g}} = n_{\mathrm{g}}; n, p_1, p_2, \dots, p_{\mathrm{g}} =$$

$$P\binom{n}{n_1 \;\; n_2 \dots \;\; n_{\mathrm{g}}} p_1^{n_1} \, p_2^{n_2} \dots p_{\mathrm{g}}^{n_{\mathrm{g}}} \qquad (1)$$

Donde $\sum_{j=1}^{\mathrm{g}} n_j = n$ y $\sum_{j=1}^{\mathrm{g}} p_j = 1$. En esencia, la regresión logística multinomial estima las probabilidades p_j de esta distribución para cada individuo *i*, teniendo en cuenta un conjunto de variables explicativas (Mc Fadden, 1974). Por ejemplo considérese un estudio realizado en un hospital en el área de quemados en el que a cada individuo al ingresar se le miden cuatro variables que se piensa influyen en su evolución, las cuales son: la edad, porcentaje de área corporal afectados por quemaduras hipodérmicas, dérmicas de tipo B y dérmicas de tipo A. se consideran tres posibles respuestas al momento de la salida del paciente: muere, sobrevive con secuelas, sobrevive sin secuelas y se desea estimar la probabilidad de que un individuo con cierta edad y porcentajes de quemaduras desarrolle alguno de los tres eventos mencionados (Silva, 1995). Otras aplicaciones de la regresión logística multinomial se presentan en la investigación de mercados, en la cual se investiga sobre la probabilidad de que un individuo seleccione un determinado producto de un conjunto de estos, dadas algunas características de sus condiciones económicas, sociales y culturales.

Para realizar las estimaciones de los parámetros en el modelo logístico multinomial, se utiliza la función canónica de enlace de la distribución multinomial a la familia exponencial, la cual es llamada transformación logit:

$$\mathrm{l}\;\; E\;\; Y_i \;\; \overset{def}{=} \log\left(\frac{p_{ij}}{p_{ig}}\right) \qquad (2)$$

Donde p_{ij} es la probabilidad del individuo *i* pertenezca a la categoría *j*, p_{ig} corresponde a la probabilidad del mismo individuo en la categoría *g*, la cual es definida como la categoría de referencia, de la variable con distribución multinomial Y. Al realizar una revisión de la literatura sobre el enlace canónico multinomial, se encuentra poca información de los antecedentes

teóricos de esta función, debido a que el modelo multinomial puede verse como una extensión del modelo de regresión logístico binomial, del cual muchos autores realizan el desarrollo formal y luego generalizan al caso politomico, mostrando directamente los resultados, como Hosmer y Lemeshow (2000), McCullagh y Nelder (1989), Jobson (1991) , Silva (1995) entre otros. Dado esta situación este trabajo pretende mostrar formalmente el planteamiento del modelo de regresión logístico multinomial partiendo de dicha función, y las expresiones a través de las cuales se estiman las probabilidades de los individuos.

2. Aspectos Teóricos de la Función Canónica de Enlace

Como se mencionó anteriormente para la estimación de los parámetros del modelo se utiliza el hecho de que la distribución multinomial pertenece a la familia exponencial, lo cual se puede demostrar de forma sencilla. Supóngase que $F_Y\ y;\theta$ es una función de distribución dependiendo de un único parámetro θ . Se dice que pertenece a la familia exponencial si su función de densidad (o de masa, en el caso discreto) puede expresarse de la forma:

$$f_Y\ y;\theta\ = \exp\ a(\theta)b(y)+c(\theta)+d(y) \qquad (3)$$

En el caso de distribuciones que dependen de k parámetros, por ejemplo $\theta = \theta_1,\theta_2,...,\theta_k$, la definición anterior se generaliza de la siguiente manera (Tusell, 2003):

$$f_Y\ y;\theta\ = \exp\left\{\sum_{i=1}^{k} a_i(\theta)b_i(y)+c(\theta)+d(y)\right\} \quad (4)$$

Por lo tanto si el vector aleatorio Y de dimensión g tiene distribución multinomial con parámetros n $\vec{p} = p_1, p_2,..., p_g$, entonces la función de probabilidad correspondiente bajo esta distribución es:

$$P(Y_{i1} = n_1, \dots, Y_{ig} = n_{\mathrm{g}}; n, p_{i1}, p_{i2}, \dots, p_{ig} =$$

$$P\binom{n}{n_1\ \ n_2 \dots\ n_{\mathrm{g}}} p_{i1}^{n_1}\ p_{i2}^{n_2} \dots p_{ig}^{n_{\mathrm{g}}} \qquad (5)$$

Se denota $\binom{n}{n_1\ \ n_2\ \ \dots\ \ n_g} = \frac{n!}{n_1!n_2!\dots n_g!} = A$ para simplificar las expresiones, entonces:

$$P = (A)p_{i1}^{n_1} \dots p_{i(\mathrm{g}-1)}^{n_{(\mathrm{g}-1)}} p_{ig}^{n_{\mathrm{g}}}$$

Como $n_1 + n_2 + \dots + n_g = n$, por lo tanto

$$P = (A)p_{i1}^{n_1} \dots p_{i(\mathrm{g}-1)}^{n_{(\mathrm{g}-1)}} p_{ig}^{n-n_1-\dots-n_{\mathrm{g}-1}}$$

$$P = (A)\frac{p_{i1}^{n_1} \dots p_{i(\mathrm{g}-1)}^{n_{(\mathrm{g}-1)}} p_{ig}^{n}}{p_{ig}^{n_1+n_2+\dots+n_{\mathrm{g}-1}}}$$

$$P = (A)\frac{p_{i1}^{n_1} \dots p_{i(\mathrm{g}-1)}^{n_{(\mathrm{g}-1)}} p_{ig}^{n}}{p_{ig}^{n_1} \dots p_{ig}^{n_{\mathrm{g}-1}}}$$

$$P = (A)\frac{p_{i1}^{n_1}}{p_{ig}^{n_1}}\frac{p_{i2}^{n_2}}{p_{ig}^{n_2}}\dots\frac{p_{i(\mathrm{g}-1)}^{n_{(\mathrm{g}-1)}}}{p_{ig}^{n_{\mathrm{g}-1}}}\ p_{ig}^{n}$$

$$P = (A)\left(\frac{p_{i1}}{p_{ig}}\right)^{n_1}\left(\frac{p_{i2}}{p_{ig}}\right)^{n}\dots\left(\frac{p_{i(\mathrm{g}-1)}}{p_{ig}}\right)^{n_{\mathrm{g}-1}}\ p_{ig}^{n}$$

Sin alterar la función, se aplica exp log P = $e^{\log P} = P$, es así como:

$$P = \exp\left\{ \log\left[A \left(\frac{p_{i1}}{p_{ig}}\right)^{n_1} \left(\frac{p_{i2}}{p_{ig}}\right)^{n_2} .... \left(\frac{p_{i(g-1)}}{p_{ig}}\right)^{n_{g-1}} p_{ig}^{n} \right] \right\}$$

$$P = \exp\left\{ \log\ A + n_1 \log\left(\frac{p_{i1}}{p_{ig}}\right) + ... + n_{g-1} \log\left(\frac{p_{i(g-1)}}{p_{ig}}\right) + n\log(p_{ig}) \right\}$$

Manipulando algebraicamente y utilizando algunas propiedades de las funciones logarítmicas y exponenciales se obtiene finalmente:

$$P = \exp\left\{ \sum_{j=1}^{g-1} n_j \log\left(\frac{p_{ij}}{p_{ig}}\right) + n\log(p_{ig}) + \log\ A \right\} \quad (6)$$

Lo cual puede ser expresado como una función de la familia exponencial, mostrada en (4) donde:

$$b_i(y) = n_j$$

$$c(\theta) = n\log\ p_{ig}$$

$$d(y) = \log(A) = \log\left(\frac{n!}{n_1!n_2!...n_g!}\right)$$

$a_i(\theta) = \log\left(\frac{p_{ij}}{p_{ig}}\right) = \mathrm{l}\ E\ Y_i$ Que es el parámetro natural de la familia y también llamada la función de enlace canónico (2). Así queda demostrado que la distribución Multinomial pertenece a la familia exponencial.

3. Planteamiento del Modelo Logístico Multinomial

Dado lo preliminar, note que la función (2) toma valores de modo continuo entre $-\infty$ y ∞, entonces es razonable pensar en hacer que $\mathrm{l}\ E\ Y_i$ dependa linealmente de los regresores[1] :

$$\mathrm{l}\ E\ Y_i = \log\left(\frac{p_{ij}}{p_{ig}}\right) = \beta_{0j} + \beta_{1j}x_{i1} + ... + \beta_{pj}x_{ip} \quad (7)$$

Ahora aplicando la exponencial a ambos lados de la ecuación lineal (7), se tiene:

$$\frac{p_{ij}}{p_{ig}} = e^{\beta_{oj} + \beta_{1j}x_{i1} + ... + \beta_{pj}\ x_{ip}}$$

$$\Rightarrow p_{ij} = p_{ig} e^{\beta_{oj} + \beta_{1j}x_{i1} + ... + \beta_{pj}\ x_{ip}} \quad (8)$$

[1] Transformar la media $E(Y_i)$ es un enfoque alternativo al de transformar Y_i, y en muchos aspectos un refinamiento. Una transformación de la respuesta como, por ejemplo, las de la familia de Box-Cox, tiene que cumplir varios objetivos. Por un lado, se desea que la variable respuesta se acerque a la normalidad. Por otro, que la varianza sea homogénea, y que la dependencia de los regresores sea lineal. Este enfoque de hacer depender linealmente de los regresores función de la media de la variable respuesta es mucho más flexible. Se puede escoger entonces la función de la media que sea más aproximada, y especificar separadamente la distribución de la variable respuesta en torno a su media. El enfoque goza así de una enorme flexibilidad. (Núñez y Tusell, 2005)

Como

$$\sum_{j=1}^{g} p_{ij} = p_{i1} + ... + p_{i(g-1)} + p_{ig} = 1$$

$$e\ p_{ig} = 1 - p_{i1} - ... - p_{i(g-1)} = 1 - \sum_{j=1}^{g-1} p_{ij}$$

Entonces

$$p_{ig} = 1 - \sum_{j=1}^{g-1} p_{ig} e^{\beta_{0j} + \beta_{1j} x_{i1} + ... + \beta_{pj} x_{ip}}$$

$$= 1 - p_{ig} \sum_{j=1}^{g-1} e^{\beta_{0j} + \beta_{1j} x_{i1} + ... + \beta_{pj} x_{ip}}$$

$$p_{ig} \left(1 + \sum_{j=1}^{g-1} e^{\beta_{0i} + \beta_{1i} x_{i1} + ... + \beta_{pi} x_{ip}} \right) = 1$$

Y a partir de la probabilidad de que el individuo i pertenezca a la categoría g, la cual de ahora en adelante se denotara como la categoría de referencia:

$$p_{ig} = \frac{1}{\left(1 + \sum_{j=1}^{g-1} e^{\beta_{0i} + \beta_{1i} x_{i1} + ... + \beta_{pi} x_{ip}} \right)} \qquad (9)$$

Se obtienen finalmente las siguientes probabilidades condicionales al reemplazar (9) en (8)

$$p_{ij} = \frac{e^{\beta_{0j} + \beta_{1j} x_{i1} + ... + \beta_{pj} x_{ip}}}{\left(1 + \sum_{j=1}^{g-1} e^{\beta_{0j} + \beta_{1j} x_{i1} + ... + \beta_{pj} x_{ip}} \right)} \qquad (10)$$

Para $i = 1,\ldots,n$ y $j = 1,\ldots,g\text{-}1$

Entonces de esta forma quedan definidas las expresiones utilizadas para el cálculo de las probabilidades:

$$p_{ij} = P\left[Y_i = j \mid x_1, x_{2,\ \ldots,} x_p \right] ; j = 1,2,\ldots,g-1$$

Que en esencia se interpretan como la probabilidad de que el individuo *i* pertenezca a la categoría j, dado un conjunto de covariables[2].

Ahora bien, empleando el método de estimación por máxima verosimilitud que no establece restricción alguna respecto de las características de las variables predictoras, se estiman los parámetros β_{ij} de la regresión lineal, la función de verosimilitud se define como sigue:

$$L = \prod_{i=1}^{n} \ p_{i1}^{y_{i1}} \ p_{i2}^{y_{i2}} \ \ldots \ p_{ig}^{y_{ig}} \qquad (11)$$

Que al aplicarle el logaritmo y remplazando los resultados anteriores se obtiene (12). Al maximiza la función de verosimilitud se obtiene un sistema de ecuaciones que requiere de los métodos numéricos para encontrar la mejor solución.

[2] Note que si el número de categorías es dos, se estará trabajando un modelo logístico binario, siendo cero la categoría de referencia g.

$$\ln L = \sum_{i=1}^{n}\sum_{j=1}^{g} y_{ij} \ln p_{ij} = \sum_{i=1}^{n}\left[\sum_{j=1}^{g-1} y_{ij} \ln\left(\frac{e^{x_i'\beta_j}}{1+\sum_{j=1}^{g-1} e^{x_i'\beta_j}}\right) + y_{ig} \ln\left(\frac{1}{1+\sum_{j=1}^{g-1} e^{x_i'\beta_j}}\right)\right] \quad (12)$$

$$\ln L = \sum_{i=1}^{n}\left[\sum_{j=1}^{g-1} y_{ij} \ln e^{x_i'\beta_j} - \sum_{j=1}^{g} y_{ij} \ln\left(1+\sum_{j=1}^{g-1} e^{x_i'\beta_j}\right)\right]$$

Los algoritmos iterativos, como el método de Newton-Raphson, son una buena alternativa. Este trabajo no pretende abordar estos problemas de inferencia.

4. Ejemplo Práctico

Para ilustrar la parte practica de la regresión logística multinomial, se tomara la base de datos **cereal.sav** que se encuentra en los ejemplos ilustrativos del paquete estadístico SPSS 15.0®. El objetivo del ejemplo es mejorar el marketing de opciones de desayuno de la compañía Consumer Packages Goods, para esto encuestó a 880 personas, registrando su edad, género y si tiene o no un estilo de vida activo. Cada participante degustó tres tipos de desayunos y se le preguntó sobre su preferencia. A continuación se presentan las definiciones de las variables y las frecuencias correspondientes.

Tabla 1. *Definición y categorización de las variables.*

Variable	Definición	Categoría
Edad	Representa la edad de la persona encuestada. Variable de tipo cualitativo con escala ordinal.	1. Menor de 30 años 2. 31 a 45 años 3. 46 a 60 años. 4. Más de 60 años
Género	Se refiere al género de la persona encuestada. Variable de tipo cualitativo con escala nominal.	0. Hombre. 1. Mujer
Estilo de Vida	Representa el estilo de la persona encuestada en cuanto a si realiza (Activo) o no (No activo) actividad física (Se considera estilo de vida activo si la persona en las últimas dos semanas ha realizado ejercicio).Variable de tipo cualitativo con escala ordinal.	0. No activo 1. Activo
Desayuno Preferido	Representa el tipo de desayuno que la persona selecciono como el mejor de los 3 tipos. Variable de tipo cualitativo con escala ordinal.	1. Barra Energética. 2. Avena. 3. Cereales

Tabla 2. *Frecuencias absolutas y relativas.*

Variable	Categoría	Frecuencia	%
Edad	Menor de 30 años	181	20.6
	31 a 45 años	206	23.4
	46 a 60 años.	231	26.3
	Más de 60 años	262	29.8
	Total	880	100
Género	Hombre.	424	48.2
	Mujer	456	51.8
	Total	880	100
Estilo de Vida	No activo	474	53.9
	Activo	406	46.1
	Total	880	100
Desayuno Preferido	Barra Energética.	231	26.3
	Avena.	310	35.2
	Cereales	339	38.5
	Total	880	100

4.1. . Ajuste del modelo

La estimación del modelo se realiza con base en las 880 observaciones, con el objetivo de explicar el tipo de desayuno en función de la edad, el género y el estilo de vida. Para esto se deben crear variables ficticias o variables dummy que representan las categorías de cada variable. Estas variables son dicotómicas, 1 si pertenece a la categoría y 0 cuando no pertenece:

$$x_i = \begin{cases} 0;\ Si\ pertenece\ a\ la\ categoría \\ 1;\ Si\ no\ pertenece\ a\ la\ categoría \end{cases};\ i = 1,2,\dots,8$$

x_1 = Edad 1 x_5 = *Hombre*

x_2 = Edad 2 x_6 = *Mujer*

x_3 = Edad 3 x_7 = *Estilo de vida activo*

x_4 = Edad 4 x_8 = *Estilo de vida no activo*

Especificación del modelo tomando desayuno con *cereales* como la categoría de referencia (*g*=3). Ecuación (13). La estimación de los parámetros de los modelos son mostrados en la Tabla 3.

$$ln\left(\frac{p_j}{p_3}\right) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{3j}x_3 + \beta_{4j}x_4 + \beta_{5j}x_5 + \beta_{6j}x_6 + \beta_{7j}x_7 + \beta_{8j}x_8\ ;\ j = 1,2 \qquad (13)$$

Tabla 3. *Estimación de los parámetros.*

Desayuno preferido		B	Error típ.	Wald	gl	Sig.	Exp(β)	Intervalo de confianza al 95% para Exp(B)	
								Límite inferior	Límite superior
Barra Energetica	Intersección	-0.68	0.30	5.22	1	0.02			
	Edadcat=1	0.94	0.31	9.07	1	0.00	2.57	1.39	4.74
	Edadcat=2	1.05	0.31	11.41	1	0.00	2.86	1.55	5.27
	Edadcat=3	0.27	0.33	0.66	1	0.42	1.31	0.68	2.51
	Edadcat=4	0[b]	.	.	0	.	.	.	.
	Hombre	-0.15	0.18	0.68	1	0.41	0.86	0.61	1.22
	Mujer	0[b]	.	.	0	.	.	.	.
	No Activo	-0.79	0.18	18.94	1	0.00	0.46	0.32	0.65
	Activo	0[b]	.	.	0	.	.	.	.
Avena	Intersección	1.00	0.21	22.37	1	0.00			
	Edadcat=1	-4.26	0.53	63.77	1	0.00	0.01	0.00	0.04
	Edadcat=2	-2.46	0.27	80.19	1	0.00	0.09	0.05	0.15
	Edadcat=3	-1.12	0.21	28.72	1	0.00	0.33	0.22	0.49
	Edadcat=4	0[b]	.	.	0	.	.	.	.
	Hombre	0.04	0.18	0.05	1	0.82	1.04	0.73	1.48
	Mujer	0[b]	.	.	0	.	.	.	.
	No Activo	0.17	0.19	0.87	1	0.35	1.19	0.83	1.72
	Activo	0[b]	.	.	0	.	.	.	.

b: Este parámetro se ha establecido a cero porque es redundante.

Tabla 4. *Información sobre el ajuste del modelo.*

Modelo	Criterio de ajuste del modelo	Contrastes de la razón de verosimilitud		
	-2 log verosimilitud	Chi-cuadrado	g. l.	Sig.
Inicial	511.64			
Final	135.00	376.64	10	8.7674E-75

Como se observa en la Tabla 4, el contraste de la razón de verosimilitud fue significativo (α=0.05) por lo cual se rechaza la hipótesis nula de que todo los parámetros incluidos en el modelo son iguales a cero. Con respecto a las estimaciones de los parámetros en la Tabla 3 para el modelo correspondiente a la categoría *barra energética* las variables que resultaron significativas fuerón x_1, x_2 y x_7. Mientras que para el modelo correspondiente a la categoría *avena* las variables que resultaron significativas fuerón x_1, x_2 y x_3. La interpretación de estos coeficientes se debe realizar en términos de la expresión $\exp(\beta)$ llamada *OR* (*odds ratio*):

$$\exp(\beta_{ij}) = \frac{odds\ (x_i = k-esima\ categoría)}{odds\ (x_i = última\ categoría)}; \quad i = 1,\ldots,8;\ \ j = 1,2 \tag{14}$$

Donde los *odds* representan la oportunidad de un evento que tiene probabilidad *p* de ocurrencia respecto de su no ocurrencia o en tal caso de la ocurrencia de otro evento con probabilidad *q*. En este sentido, la interpretación de los coeficientes se realiza de acuerdo a cada modelo. Para el primer modelo $\exp(\beta_{71}) = 0.46$ (correspondiente a la variable x_7) significa que la preferencia[3] por la *barra energética* como desayuno frente a los *cereales* para los que tienen un *estilo de vida no activo* es 0.46 veces la preferencia de los que tienen un *estilo de vida activo*, indicando que una persona con *estilo de vida activo* tiene mayor preferencia por la *barra energética* frente a los *cereales*, sin tener en cuenta las demás variables. Así mismo $\exp(\beta_{21}) = 2.86$ significa que la predilección por la *barra energética* como desayuno preferido frente a los *cereales* para los que tienen una *edad* entre 31 y 45 años (x_2) es 2.86 veces la de los que tienen una *edad* superior a 60 años. De la misma forma se interpretan los coeficientes para el segundo modelo teniendo en cuenta que ahora la elección es entre la *avena* y los *cereales*. Además, para los coeficientes del *género* es indiferente la escogencia de un tipo de desayuno en ambos modelos. Esto sucede cuando los *odds* son estadísticamente iguales, es decir que la opción de elegir un producto u otro es indiferente a la categoría de la variable.

De acuerdo a los resultados obtenidos los modelos finales se pueden expresar como Ecuación (15) y (16). Ahora bien, la probabilidad de que un individuo prefiera como desayuno la *barra energética* dada unas características es:

$$P\ \ y = 1\ = \frac{e_1}{1 + e_1 + e_2}$$

$$e_1 = \left(\frac{p_1}{p_3}\right) = \exp\ \ \beta_{01} + \beta_{11}x_1 + \beta_{21}x_2 + \beta_{31}x_3 + \beta_{51}x_5 + \beta_{71}x_7 \tag{15}$$

$$e_2 = \left(\frac{p_2}{p_3}\right) = \exp\ \ \beta_{02} + \beta_{12}x_1 + \beta_{22}x_2 + \beta_{32}x_3 + \beta_{52}x_5 + \beta_{72}x_7 \tag{16}$$

[3] En salud se habla del riesgo relativo, debido al contexto del estudio se trata de una opción, oportunidad, lección o alternativa

Y la probabilidad de que su desayuno preferido sea *avena*:

$$P\ \ y=2\ \ =\frac{e_2}{1+e_1+e_2}$$

Por lo tanto la probabilidad de que prefiera los *cereales* como desayuno, es el complemento:

$$P(y=3)=1-\ \ P(y=1)+P(y=2)$$

Como se observa las probabilidades relacionan ambos modelos. Para interpretar estas estimaciones considérese una mujer que realice actividad física y sea menor a 30 años, las probabilidades de su desayuno preferido son:

- $P(\ y = 1\ ) = 0.77$
- $P(\ y = 2\ ) = 0.23$
- $P(\ y = 3\ ) = 0.00$

Ahora si es un hombre con una edad superior a 60 años y además no realiza actividad física, las probabilidades serian:

- $P(\ y = 1\ ) = 0.10$
- $P(\ y = 2\ ) = 0.66$
- $P(\ y = 3\ ) = 0.23$

Apreciando los valores de los $\exp(\beta_{ij})$ y de las probabilidades, cuando la *edad* es mayor a los 60 años el individuo prefiere la *avena*, mientras si su edad es menor, preferirá la *barra energética*. Esto explica porque los $\exp(\beta_{i2})$ son menores a 1 y los $\exp(\beta_{i1})$ son mayores a 1.

5. Conclusiones

El modelo de regresión logística es de gran importancia en la parte aplicativa debido a que facilita la modelación de una variable de naturaleza cualitativa en términos de un conjunto tienen más de dos categorías de carácter nominal, la distribución multinomial es la más adecuada, pues debido a su pertenencia a la familia exponencial, permite realizar una transformación para expresar el logaritmo del valor esperado de la variable respuesta en términos de la regresión lineal, esta transformación es conocida como la función de enlace canónico.

Al igual que en el caso dicotómico la estimación de las probabilidades correspondientes a cada evento o alternativa considerada son de gran utilidad en áreas como la economía y la salud, pues en muchos casos se desea conocer cómo cambian estas probabilidades con relación a características inherentes de los individuos. En los estudios de mercadeo esta técnica es muy utilizada debido a la subjetividad de las variables, similarmente en el análisis de riesgos la regresión logística binomial y multinomial juegan un papel importante, debido a que se busca caracterizar perfiles de individuos que maximizan las probabilidades a estimar, por ende hay un interés relevante en el conjunto de variables explicativas, pues se busca en lo posible que este conjunto o un subconjunto de ellas discriminen en términos significativos a los individuos con referencia a la variable respuesta.

Cuando se utiliza el modelo de regresión multinomial, se deben estimar tantos modelos como categorías menos una tenga la variable respuesta, digamos *g*-1, y en consecuencia se deben estimar *g*-1 vectores de regresores. Como el sistema de ecuaciones que se tiene es indeterminado, no existen soluciones matemáticas explicitas a las ecuaciones y debe recurrirse a algoritmos iterativos de los métodos numéricos para encontrar una solución.

6. Referencias Bibliográficas

1. Feller, W. (1967). *An introduction to probability theory and its application*, tercera edición, John Wiley & Sons, USA.

2. Hosmer, D. & Lemeshow, S. (2000). *Applied Logistic Regresion*, segunda edición, John Wiley & Sons, USA.

3. Jobson, J. D. (1991). *Categorical & multivariate methods*, vol 2, Springer verlag, USA.

4. McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Model*, segunda edición, Chapman & Hall, London.

5. McFadden, D. (1974). *Conditional logit analysis of qualitative choice behavior*, Frontiers of Econometrics. Academic Press, New York.

6. Núñez, A. & Tusell, F. (2007). Regresión y análisis de varianza. Notas de clase. http://www.et.bs.ehu.es/~etptupaf/nuevo/ficheros/estad3/reg.pdf

7. Silva, L. C. (1995). *Excursión a la regresión logística en ciencias de la salud (1994*), primera edición, Díaz de santos S.A.

8. Tussel, F. (2003). *Estadistica matemática*, Notas de clase Economía aplicada III, universidad del país vasco: http://www.et.bs.ehu.es/~etptupaf/nuevo/ficheros/estad5/estad5.pdf