PESQUIMAT Revista de la Fac.CC.MM.de la
Universidad Nacional Mayor de San Marcos
Vol.III, $N°2$, Noviembre 2000, pag.23-33. Lima-Perú

ALGUNAS APLICACIONES DEL MODELO LOG LINEAL Y LA REGRESIÓN LOGíSTICA

Rosa Maria Inga Santivañez [1]

ABSTRACT. *En este artículo se presentan dos métodos para el análisis de datos categorizados: la regresión logística, la cual nos permite estudiar una variable respuesta cualitativa con respecto a variables explicativas cualitativas o cuantitativas; y los modelos log lineal los cuales nos permiten analizar posibles relaciones entre las variables. Estos métodos son aplicados en el estudio del riesgo nutricional, obteniéndose que variables determinan un riesgo nutricional.*

1. INTRODUCCIÓN

En los estudios de datos categorizados es de interés estudiar una variable respuesta cualitativa en función de variables explicativas cuantitativas o cualitativas, un método para analizar este tipo de información es la regresión logística, el cual es un caso especial de la regresión múltiple. En el análisis de datos categorizados también es interesante examinar los tipos de relación que existe entre las variables categóricas (o factores de una tabla de contingencia) para realizar este análisis se puede asociar modelos log lineal a hipótesis estadística y luego analizar el ajuste de estos modelos.

2. METODOLOGÍA

A continuación presentamos los métodos para el análisis de datos categorizados, la regresión logística y los modelos log lineal.

[1]Univ. Peruana Cayetano Heredia; Univ. Nacional Mayor de San Marcos.
e-mail: rmis@upch.edu.pe

2.1. Modelo Log Lineal

Un método para analizar los tipos de relación que se pueden establecer entre factores de una tabla de contingencia es mediante los modelos log lineal. Mediante el modelo log lineal asociado a cierta hipótesis H_0 hallamos los valores esperados (E) de la tabla de contingencia.

MODELO LOG LINEAL (ASOCIADO H_0) $\rightarrow$ E (VALOR ESPERADO)

Contrastamos la bondad de ajuste mediante el estadístico

$$\chi_0^2 = \sum_{\substack{todas \\ celdas}} \frac{(O-E)^2}{E} \qquad \text{ó} \qquad G = \sum_{\substack{todas \\ celdas}} O^* \ln\left(\frac{O}{E}\right) \sim \chi^2$$

donde: E: valor esperado; O:valor observado.

MODELOS LOG LINEAL PARA UNA TABLA DE TRES FACTORES DE CLASIFICACIÓN

$$\ln m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} + \mu_{123(ijk)} \quad (1)$$

donde:

m_{ijk}	: frecuencia esperada de la celda (i,j,k)
μ	: la media de los logaritmos de las frecuencias
$\mu_{1(i)}$	: efecto de la categoria i del factor Fila
$\mu_{2(j)}$	: efecto de la categoria j del factor Columna
$\mu_{3(k)}$	: efecto de la categoria k del factor Profundidad
$\mu_{12(ij)}$	: efecto de interacción de la categoria i del factor Fila y la categoria j del factor Columna
$\mu_{13(ik)}$	: efecto de interacción de la categoria i del factor Fila y la categoria k del factor Profundidad
$\mu_{23(jk)}$	: efecto de interacción de la categoria j del factor Columna y la categoria k del factor Profundidad
$\mu_{123(ijk)}$	: efecto de la categoria i del factor Fila, la categoria j del factor Columna y la categoria k del factor Profundidad

Los modelos asociados con varias hipótesis relacionadas a tablas de contingencia se obtienen cuando se hacen cero ciertos términos de (1)

(estamos tratando con modelos jerárquicos, cuando efecto de interacción de mayor orden se incluyen en el modelo efecto de menor orden también se incluyen en el modelo).

EJEMPLO: H_0: $\mu_{123(ijk)} = 0 \quad i = 1,\ldots,r \quad j = 1,\ldots,c \quad k = 1,\ldots,d$

El modelo asociado a H_0 es.

$$\ln m_{ijk} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{3(k)} + \mu_{12(ij)} + \mu_{13(ik)} + \mu_{23(jk)} \qquad (2)$$

Multiplicando (2) por x_{ijk} y luego sumando sobre i, j, k y si además sumamos y restamos $\left[\sum_i \mu_{1(i)} x_{i..} + \sum_j \mu_{2(j)} x_{.j.} + \sum_k \mu_{3(k)} x_{..k}\right]$, obtenemos,

$$\sum_i \sum_j \sum_k x_{ijk} \ln m_{ijk} = N\mu + \left[\sum_i \sum_j x_{ij.} U_{12} + \sum_i \sum_k x_{i.k} U_{13} + \sum_j \sum_k x_{.jk} U_{23}\right] - \left[\sum_i x_{i..} \mu_1 + \sum_j x_{.j.} \mu_2 + \sum_k x_{..k} \mu_3\right]$$

donde: $U_{12} = \mu_1 + \mu_2 + \mu_{12}$, $\quad U_{13} = \mu_1 + \mu_3 + \mu_{13}$, $\quad U_{23} = \mu_2 + \mu_3 + \mu_{23}$.

Observamos que los datos siguen una distribución Multinomial y bajo H_0 tenemos que

$$p(x) = e^{\ln p(x)} = exp(\sum_i \sum_j \sum_k x_{ijk} \ln m_{ijk} + \ln[n!/N^N \prod_{i,j,k} x_{ijk}!])$$

Luego las configuraciones suficientes de U_{12}, U_{13}, U_{23} son C_{12}, C_{13}, C_{23}.

MODELOS EN TABLAS DE TRES DIMENSIONES

Tipo de Modelo	Términos Ausentes	Configuraciones suficientes	Grados de Libertad	H_0
1	μ_{123}	C_{12}, C_{13}, C_{23}	$(r-1)(c-1)(d-1)$	NO EXISTE INTERACCIÓN ENTRE LOS FACTORES
2	μ_{12}, μ_{123}	C_{13}, C_{23}	$(r-1)(c-1)d$	$(F^*C)/D$
3	$\mu_{12}, \mu_{13}, \mu_{123}$	C_{23}, C_1	$(r-1)(cd-1)$	$F^*(C,D)$
4	$\mu_{12}, \mu_{13}, \mu_{23}, \mu_{123}$	C_1, C_2, C_3	$rcd-(r+c+d)+2$	F^*C^*D

Examinando las configuraciones suficientes se determina si los estimadores de "m" se van a obtener por

MÉTODO DIRECTO → POR LOS ESTIMADORES DE MÁXIMA VEROSIMILITUD.

MÉTODO ITERATIVO → POR AJUSTE ITERATIVO DE LAS CONFIGURACIONES.

REGLA PARA DETECTAR LA EXISTENCIA DE ESTIMADORES DIRECTOS

Mediante un examen de las configuraciones suficientes nos permiten determinar si un modelo puede ser ajustado directamente o no.
PASOS

1. Agrupar cualquier grupo de variables que siempre aparecen juntas en una sola variable.
2. Suprimir alguna variable que aparezca en toda configuración.
3. Suprimir alguna variable que sólo aparezca en una configuración.
4. Cambiar alguna configuración redundante.
5. Repetir los pasos del 1 al 4 hasta que:
   (a) No haya más de dos configuraciones, esto es un indicador de la existencia de estimadores directos.
   (b) Si no se pueden obtener menos de tres configuraciones, nos indicaría que debemos hacer interacciones para obtener los estimadores.

Tipo de Modelo	Términos Ausentes	Configuraciones suficientes	Método de Estimación
1	μ_{123}	C_{12}, C_{13}, C_{23}	NO SE PUEDE APLICAR NINGÚN PASO, APLICAR MÉTODO ITERATIVO
2	μ_{12}, μ_{123}	C_{13}, C_{23}	EL 3 ES COMÚN POR PASO 2, LUEGO C_1,C_2, ASÍ QUE SE PUEDE APLICAR UN MÉTODO DIRECTO.
3	$\mu_{12}, \mu_{13}, \mu_{123}$	C_{23}, C_1	POR PASO 5, ENTONCES SE APLICA MÉTODO DIRECTO.
4	$\mu_{12}, \mu_{13}, \mu_{23}, \mu_{123}$	C_1, C_2, C_3	POR PASO 5, USAR MÉTODO ITERATIVO.

PROCEDIMIENTO DE AJUSTE ITERATIVO

Mostraremos el procedimiento iterativo mediante un ejemplo; consideremos que se desea realizar el siguiente contraste $H_0 : \mu_{123} = 0$. Las configuraciones suficientes asociadas son C_{12}, C_{13}, C_{23}.

PROCESO

- Considerar estimadores preliminares $\hat{m}^{(0)}_{ijk} = 1$
- Ciclo (ajustamos los estimadores preliminares para ajustar sucesivamente C_{12}, C_{13}, C_{23}).

PASO 1.- Ajustando C_{12} tenemos $\hat{m}^{(1)}_{ijk} - \dfrac{\hat{m}^{(0)}_{ijk}\, x_{ij.}}{\hat{m}^{(0)}_{ij.}}$

PASO 2.- Ajustando C_{13} tenemos $\hat{m}^{(2)}_{ijk} - \dfrac{\hat{m}^{(1)}_{ijk}\, x_{i.k}}{\hat{m}^{(1)}_{i.k}}$

PASO 3.- Ajustando C_{23} tenemos $\hat{m}^{(3)}_{ijk} - \dfrac{\hat{m}^{(2)}_{ijk}\, x_{.jk}}{\hat{m}^{(2)}_{.jk}}$

Repetimos el ciclo hasta que la convergencia con la aproximación deseada se alcance

Regla de parada $\left| \hat{m}^{(3r)}_{ijk} - \hat{m}^{(3r-3)}_{ijk} \right| < \delta$

donde δ: Precisión r : Número de ciclo.

- Si existen estimadores directos el procedimiento iterativo daría los estimadores exactos en el primer ciclo.

CONVERGENCIA DEL PROCESO

Darroch y Ratclif (1972) demostraron la convergencia del proceso iterativo para hallar los estimadores de máxima verosimilitud $\{\hat{m}_{ijk}\}$ de $\{m_{ijk}\}$ Fienberg (1970), Gockhale (1971), Ireland y Kullback (1968) nos presentaron otras demostraciones de la convergencia del proceso iterativo.

2.2. La Regresión Logística

El objetivo de la regresión logística es obtener un modelo especial de regresión múltiple, con las siguientes características diferenciales: a) la variable dependiente o respuesta no es continua, sino discreta (generalmente 1, 0); b) las variables explicativas pueden ser cuantitativas o

cualitativas; c) la ecuación del modelo no es una función lineal de partida, sino exponencial; si bien por sencilla transformación logarítmica (logit), puede finalmente presentarse como función lineal.

MODELO LOGISTICO

Sea Y una variable dicotómica.

$$Y - \begin{cases} 1 & \text{enfermo} \\ 0 & \text{sano} \end{cases}$$

La ecuación de curva signoide, matemáticamente sencilla y flexible, y biológicamente interpretable es en el caso de una sola variable predictiva x,

$$P = \frac{e^{b_0+b_1x}}{1-e^{b_0+b_1x}} \tag{3}$$

y se denomina distribución logística.

Se necesita calcular b_0 y b_1 para poder cuantificar la probabilidad de la "enfermedad" $P = P(Y = 1)$ en función de los distintos valores que este puede presentar en su factor de riesgo X.

$$\frac{P}{1-P} = e^{b_0 \mid b_1 x} \tag{4}$$

Tomando logaritmo neperiano a la expresión dada en (4) se obtiene el logit de P,

$$log\, it\,(P) = log\left(\frac{P}{1-P}\right) = b_0 + b_1 x \tag{5}$$

X es una variable cuantitativa o cualitativa $(1, 0)$, pero si X es una variable cualitativa con varias categorías se deberá usar variables ficticias o DUMMY.

El modelo dado en (5) se puede generalizar de la siguiente manera:

$$log\left(\frac{P}{1-P}\right) = b_0 + b_1\, x_1 + \ldots + b_p\, x_p \tag{6}$$

que proporciona el *logit* de la probabilidad de enfermar de un individuo que presenta perfil $x_1, \ldots, x_p$ de factores de riesgo, cuando se haya estimado los coeficientes de regresión $b_0, \ldots, b_p$.

La expresión dada en (6) se puede interpretar como el logaritmo del número de veces que es más probable que un individuo de perfil $x_1, \ldots, x_p$ sea un enfermo a que sea un sano. Su antilogaritmo es por lo tanto el número de veces que es más probable que sea un enfermo.

En la regresión logística el error ϵ sigue una distribución Binomial con probabilidad P. Y los valores de los estimadores $b_0, b_1, \ldots, b_p$ se obtienen mediante un procedimiento iterativo a través de programas estadísticos automáticos.

Algunos métodos para analizar la validez del modelo son: la prueba G y el Indice de Wald.

3. APLICACIÓN

Objetivo: Determinación de los factores que influyen en el riesgo nutricional.

Unidad de Observación: La familia.

Número de encuestas: 1360.

VARIABLE	DESCRIPCIÓN	UNIDADES
PDALTITUD	ALTITUD SOBRE EL NIVEL DEL MAR	METROS
NMEN24	NÚMERO DE NIÑOS MENORES DE 2 AÑOS	
NMEF	NÚMERO DE MUJERES EN EDAD FÉRTIL	
TOTRAB	TOTAL DE MIEMBROS DE LA FAMILIA QUE TRABAJAN	
NUSUJS	NÚMERO DE MIEMBROS DE LA FAMILIA	
		CATEGORIAS
NPARED	MATERIAL PREDOMINANTE EN PAREDES EXTERIORES	MATERIAL LIGERO(0)
		MATERIAL NOBLE (1)
NTECHO	MATERIAL PREDOMINANTE EN EL TECHO	MATERIAL LIGERO (0)
		MATERIAL NOBLE (1)
NSHIG	SERVICIOS HIGIÉNICOS	MATERIAL LIGERO (0)
		MATERIAL NOBLE (1)
NPISO	MATERIAL EN EL PISO	MATERIAL LIGERO (0)
		MATERIAL NOBLE (1)
NAGUA	ABASTECIMIENTO DE AGUA PARA BEBER	NO RED PÚBLICA (0)
		RED PÚBLICA (1)
RECUER	CONOCE SOBRE LA AYUDA EN ALIMENTOS	SI (0)
		NO (1)
AMBITO	AMBITO	COSTA (1)
		SIERRA NORTE (2)

		SIERRA CENTRO (3)
		SELVA (4)
		LIMA (5)
		SIERRA SUR (6)
RIESGO	RIESGO DE DESNUTRICIÓN	BAJO RIESGO (0)
		ALTO RIESGO (1)

Los métodos expuestos fueron utilizados para el estudio del riesgo nutricional. A continuación presentamos algunos resultados de la aplicación de la regresión logística.

Variable independiente: RIESGO.
Variables dependientes: Todas las otras variables.

MODELO:

$$\begin{aligned} log\,(P/(1-P)) \;=\; & 1.0705 - 0.0013\,[\text{AMBITO}\,(1)] - 0.943\,[\text{AMBITO}\,(2)] \\ & - 0.6716\,[\text{AMBITO}\,(3)] - 0.1560\,[\text{AMBITO(4)}] \\ & - 0.3848\,[\text{AMBITO}\,(5)] - 0.3770\,[\text{BENEFI}] \\ & - 0.2292\,[\text{NMEN24}] + 0.1857\,[\text{NMEF}] \\ & - 0.0686\,[\text{NUSUJS}] - 0.3718\,[\text{NTECHO}] \end{aligned} \tag{7}$$

Se han obtenido los ODDS RATIOS (EXP(B)), los cuales presentamos a continuación.

TABLA 1

Variable	Exp(B)
AMBITO	
AMBITO (1)	.9987
AMBITO (2)	.9100
AMBITO (3)	.5109
AMBITO (4)	.8555
AMBITO (5)	.6806
BENEFI	.6859
NMEN24	.7952
NMEF	1.2040
NUSUJS	.9337
NTECHO	.6895

TABLA 2.

TABLA DE CLASIFICACIÓN PARA RIESGO

Observada	Predicción		Porcentaje Correcto
	0	1	
0	70	445	13.59 %
1	51	794	93.96 %
	Total		63.53 %

TABLA 3.

Nivel de Probabilidad (P)	Sensibilidad $1-\beta$		Especificidad $1-\alpha$		Falso Positivo α	Falso Negativo β
0.10	100		62.2		37.8	0
0.20	100		62.2		37.8	0
0.30	80		62.3		37.7	20
0.40	72.2		62.6		37.4	27.8
0.50	57.9	*	64.1	*	35.9	42.1
0.60	47.5		68		32	52.5
0.70	40		72		28	60
0.80	37.9		100		0	62.1

A continuación presentamos algunos resultados obtenidos mediante la aplicación de los modelos *log* lineal en el análisis de un estudio de riesgo nutricional, todas las decisiones se tomaron considerando un nivel de significación del 5%.

1. Hipótesis nula: El riesgo de desnutrición es independiente del ámbito.
   Goodness-of-fit chi square $= 17.28246$ $DF = 5$ $P = .004$
   Pearson chi square $= 17.34830$ $DF = 5$ $P = .004$
2. Hipótesis nula: El riesgo de desnutrición es independiente de que si recuerda el programa de asistencia alimentaria.
   Goodness-of-fit chi square $= 0.03919$ $DF = 1$ $P = .843$
   Pearson chi square $= 0.03916$ $DF = 1$ $P = .843$
3. Hipótesis nula: Existe independencia condicional entre el riesgo de desnutrición y el recuerdo dado el ámbito.
   Goodness-of-fit chi square $= 2.88238$ $DF = 6$ $P = .823$
   Pearson chi square $= 2.89980$ $DF = 6$ $P = .821$
4. Hipótesis nula: Existe independencia entre el riesgo nutricional y el hecho de que la familia ha sido beneficiada dado el ámbito.

Goodness-of-fit chi square = 8.19730 $DF = 6$ $P = .224$
Pearson chi square = 8.33373 $DF = 6$ $P = .215$

5. Hipótesis nula: El que si recuerda el programa de ayuda es independiente del ámbito de la familia.
   Goodness-of-fit chi square = 82.71712 $DF = 5$ $P = .000$
   Pearson chi square = 76.92225 $DF = 5$ $P = .000$
6. Hipótesis nula: El riesgo de desnutrición, sistema de agua para beber y el sistema de alumbrado son independientes.
   Goodness-of-fit chi square = 683.91191 $DF = 4$ $P = .000$
   Pearson chi square = 645.95320 $DF = 4$ $P = .000$
7. Hipótesis nula: El riesgo de desnutrición, el material del techo y el material de la pared son independientes.
   Goodness-of-fit chi square = 13.36658 $DF = 4$ $P = .010$
   Pearson chi square = 13.69371 $DF = 4$ $P = .008$
8. Hipótesis nula: El riesgo de desnutrición es independiente de que la familia ha sido beneficiada.
   Goodness-of-fit chi square = 5.38449 $DF = 1$ $P = .020$
   Pearson chi square = 5.53547 $DF = 1$ $P = .019$
9. Hipótesis nula: El riesgo nutricional, el material del techo, el material de la pared, agua, alumbrado, servicios higiénicos son independientes.
   Goodness-of-fit chi square = 2124.18339 $DF = 120$ $P = .000$
   Pearson chi square = 3490.76946 $DF = 120$ $P = .000$

4. CONCLUSIÓN

Analizando los resultados de aplicar la regresión logística observamos lo siguiente:
El modelo (7) nos permite pronosticar el riesgo dado un perfil de familia.
Mediante la Tabla 1 observamos lo siguiente:

- En cuanto al AMBITO más riesgo de desnutrición en el AMBITO (1), es decir la Costa, y la zona de menor riesgo nutricional es en el AMBITO (3), es decir en la Sierra Central.
- Si la familia es beneficiada con ayuda alimentaria el riesgo de desnutrición disminuye.
- Para cada incremento de una mujer en edad fértil en la familia el riesgo de desnutrición se incrementa.
- En cuanto al NTECHO, si el material del techo de la casa es noble el riesgo de desnutrición disminuye.

En cuanto a la tabla 2 de clasificación para riesgo se observa:

- Si la familia esta en riesgo de desnutrición es clasificada el 93.96% correctamente.
- Si la familia no está en riesgo de desnutrición la clasifica correctamente el 13.59 %.
- En general el porcentaje de clasificación es de 63.53 %.

En cuanto a la tabla 3 se observa que la situación más equilibrada se encuentra en $P = 0.50$, la sensibilidad es de 57.9 % y la especificidad es de 64.1 %. Por lo tanto $P = 0.50$ puede ser el punto de corte:

- Si $P < 0.50$, entonces la familia no esta en riesgo de desnutrición.
- Si $P \geq 0.50$, entonces la familia esta en riesgo de desnutrición.

Analizando los resultados de aplicar los modelos *log* lineal observamos lo siguiente:

- El riesgo depende del ámbito.
- El riesgo de desnutrición es independiente de que si recuerda el programa de asistencia alimentaria.
- Existe independencia condicional entre el riesgo de desnutrición y el recuerdo dado el ámbito.
- Existe independencia entre el riesgo nutricional y el hecho de que la familia ha sido beneficiada dado el ámbito.
- El que si recuerda el programa de ayuda depende del ámbito de la familia.
- Existe relación entre el riesgo de desnutrición, sistema de agua para beber y el sistema de alumbrado.
- Existe relación entre el riesgo de desnutrición, el material del techo y el material de la pared.
- El riesgo de desnutrición depende de que la familia ha sido beneficiada.
- Existe relación entre el riesgo nutricional, el material del techo, el material de la pared, agua, alumbrado, servicios higiénicos.

5. BIBLIOGRAFIA

[1] Agresti Alan, *An Introduction to Categorical Data Analysis.*, Willey InterScience,(1986).

[2] Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland, *Discrete Multivariate Analysis.* MA: MIT Press. Cambridge, (1975).

[3] Christensen Roland, *Log Linear Moldels.* Springer-Verlag, (1990).