ACTA UNIVERSITARIA
Guanajuato, Gto., México

Comparación de Tres Métodos de Análisis de Variables Binarias

Hugo H. Montaldo*

RESUMEN / ABSTRACT

El presente trabajo compara los modelos lineales (ML), las tablas de contingencia con Chi-cuadrado (TC) y la regresión logística (RL) en cuanto a su eficiencia para probar hipótesis de diferencias entre porcentajes de dos tratamientos. Se definió eficiencia como la probabilidad de no cometer error tipo I. Se analizó una serie de casos simulados con 4 y 25 repeticiones. En conclusión, considerando la menor eficiencia de los ML, particularmente con pocas repeticiones y las limitaciones de las TC para el análisis de variables binarias, los resultados de este estudio preliminar ilustran las expectativas teóricas de que el método de RL, actualmente disponible en varios programas evaluados de software estadístico, debería ser preferido para el análisis de experimentos con variables dependientes binarias.

This study compares lineal models (ML), contingency tables using Chi-square (TC) and logistic regression (RL) on their efficiency to test hypothesis about differences of percentages between two treatments. Efficiency was defined as the probability of not commit error type I. A series of simulated cases with 4 and 25 repetitions were analyzed. Considering the lower efficiency of ML, particularly with few replicates and the limitations of TC for the analysis of the binary variables, the results of this preliminary study show that in agreement with the theoretical expectations, the RL method, readily available in several tested programs of statistical software, should be prefered for the analysis of experiments with binary dependent variables.

Recibido: 20 de Febrero de 2001

Aceptado: 26 de Agosto de 2001

* Instituto de Ciencias Agrícolas, Universidad de Guanajuato, Irapuato, 36500, México.
cuauhtli@veterin.unam.mx

INTRODUCCIÓN

El uso de variables binarias está muy extendido en muchas áreas del conocimiento tales como las ciencias agropecuarias, biomédicas y en general en la biología. Ejemplos de este tipo de variables "cero, uno" (0,1), son: sano, enfermo, vivo, muerto, etc. Este tipo de variables son también de uso común en ingeniería (fallo, no fallo).

Las propiedades estadísticas de estas variables son diferentes de aquéllas con una distribución continua, por lo que requieren métodos específicos para su análisis en experimentos y otros estudios comparativos (Agresti, 1990). En términos prácticos, sin embargo, en muchas investigaciones se analiza este tipo de variables usando métodos que no son en sentido estricto idóneos en cuanto a satisfacer las suposiciones del modelo. Este es el caso de los modelos lineales (ML) cuando se usan para hacer análisis de varianza con el fin de probar hipótesis de igualdad entre las medias de los tratamientos usando datos binarios. Los ML tienen suposiciones como normalidad, independencia de los errores y homogeneidad de varianzas entre grupos, que no pueden ser satisfechas en todas las situaciones al analizar variables binarias (Gill, 1978).

Probablemente la razón principal para usar los ML en el análisis de variables binarias es la mayor familiaridad de los investigadores con ellos y la posibilidad de incluir factores de variación adicionales, tales como variables de confusión categóricas o continuas y usar diseños estándar en los experimentos, tales como bloques al azar, cuadros latinos, parcelas divididas, factoriales etc. (Gill, 1978). Este tipo de variables y diseños son difíciles de tratar con algunos de los métodos más comunes de análisis no paramétrico de variables binarias, tales como las tablas de contingencia usando la distribución de Chi-cuadrado (TC).

Como alternativa a estos métodos, se han desarrollado metodologías de análisis de variables binarias en base a la generalización de los ML de acuerdo a la distribución normal, a la familia más amplia de distribuciones estadísticas exponenciales, de modo que sea posible analizar variables binarias con las ventajas que

PALABRAS CLAVE: Regresión logística, Modelos lineales, Tablas de contingencia, Eficiencia.

KEYWORDS: Logistic regression, Linear models, Contingency tables, Efficiency.

usualmente tienen los ML, pero sin tener que establecer suposiciones inapropiadas. Entre estos desarrollos se encuentran los métodos de regresión logística (RL), fundamentados en el desarrollo de modelos que analizan datos binarios transformados a logits, los que permiten un análisis similar a los de los ML, pero sin las limitaciones de las TC (Agresti, 1990; Dobson, 1990).

Esta nota tiene como propósito comparar en forma práctica y desde la perspectiva de un usuario, los métodos de ML, TC y RL para el análisis de experimentos donde la variable dependiente es binaria. Con este fin, se analizaron algunos ejemplos numéricos (casos) desarrollados para ilustrar los resultados de los diferentes métodos. En este estudio se usó como criterio de comparación la eficiencia de cada método, la que se definió como la probabilidad de no cometer error tipo I, es decir, la probabilidad de aceptar la hipótesis nula (Ho) cuando esta es cierta.

MATERIAL Y MÉTODOS

Para cada caso se obtuvo la probabilidad observada de cometer error tipo I, denotada como P y a partir de esta, la eficiencia (1-P) que es la probabilidad de no cometer error tipo I, es decir, la probabilidad de aceptar la hipótesis nula (Ho) cuando ésta es cierta. Esto implica que para los mismos datos, el método con menores valores de P será el más eficiente. En la tabla 1, se muestran los casos estudiados para dos tratamientos con 25 observaciones o repeticiones (N) en cada uno. En la tabla 2, se muestran los casos estudiados cuando se tienen 4 observaciones en cada uno. Los análisis fueron realizados con el Software Statgraphics (Statistical Graphics Corp., 1996). Los valores de P para las TC fueron obtenidos sin la corrección de Yates (Gill, 1978).

RESULTADOS Y DISCUSIÓN

Los resultados de eficiencia mostrados en las tablas 1 y 2, ilustran la posibilidad de analizar los datos utilizando los tres métodos en la mayor parte de los casos, sin embargo, para situaciones donde la diferencia entre los tratamientos fue de 100% (casos 1 en tabla 1 y 2), los ML no pudieron ser usados, ya que no fue posible estimar la varianza del error para realizar la prueba de F. Esto es una limitación de gran importancia práctica en experimentos donde en algún tratamiento todos son unos o ceros. Los valores de P para los ML pueden estar subestimados o sobreestimados por desviaciones de normalidad de la variable dependiente y la heterogeneidad de varianzas que se producirán en una variable binaria, particularmente para diferencias extremas entre los valores de las proporciones en cada muestra (Gill, 1978). También es necesario considerar que los valores calculados de P en TC, están subestimados, particularmente cuando los valores esperados para una celda en la tabla de contingencia son menores a 5, y en tablas 2x2 (Gill, 1978).

En términos generales, la RL resultó más eficiente que las TC y los ML, tanto para 25 observaciones (tabla 1), como para 4 observaciones (tabla 2). Generalmente los ML fueron el método menos eficiente, aunque las diferencias con las TC fueron relativamente pequeñas. Las diferencias a favor de la RL fueron generalmente mayores para 4 observaciones (tabla 1) que para 25 observaciones (tabla 2). Con 25 repeticiones, el rango de superioridad en eficiencia de la RL sobre ML fue de 0 a 13.24% y de RL sobre TC de 0 a 11.59% (tabla 1). Con 4 repeticiones, el rango de superioridad en eficiencia de RL sobre ML fue de 1.02 a 21.74% y de RL sobre TC de 1.07 a 9.64% (tabla 2). Los casos en los cuales la superioridad de la RL es mayor para N=25, los 4 y 8 (tabla 1), en ambos, las diferencias fueron las menores posibles (4%). En forma similar, para N=4 los casos 12 y 14, con las menores diferencias posibles (25%) fueron los que mostraron una mayor superioridad de la RL sobre los ML y las TC (tabla 2).

En conclusión, considerando la menor eficiencia de los ML, particularmente con pocas repeticiones y para diferencias menores, además

Tabla 1. Resultados de métodos de análisis de variables binarias, N=25.

CASO	Tratamientos				Eficiencia (1-P)			Incremento de la Eficiencia de RL vs. (%)	
	n_{1*}		Diferencia 1 vs. 2		Modelo Lineal (ML)	Tabla de Contingencia (TC)	Regresión Logística (RL)	Modelo Lineal (ML)	Tabla de Contingencia (TC)
	1	2	n_1	%					
1	0	25	25	100	NE**	1.000	1.000	--	0.00
2	0	5	5	20	0.982	0.981	0.994	1.22	1.33
3	1	5	4	16	0.915	0.918	0.930	1.64	1.31
4	0	1	1	4	0.680	0.690	0.770	13.24	11.59
5	10	20	10	40	0.997	0.996	0.997	0.00	0.10
6	10	15	5	20	0.836	0.843	0.844	0.96	0.12
7	11	15	4	16	0.733	0.742	0.743	1.36	0.13
8	10	11	1	4	0.220	0.225	0.226	2.73	0.44

* n_1; número de "unos" en cada muestra
**NE; no estimable porque la varianza del error fué = 0

Tabla 2. Resultados de métodos de análisis de variables binarias, N=4.

CASO	Tratamientos				Eficiencia (1-P)			Incremento de la Eficiencia de RL vs. (%)	
	n_{1*}		Diferencia 1 vs. 2		Modelo Lineal (ML)	Tabla de Contingencia (TC)	Regresión Logística (RL)	Modelo Lineal (ML)	Tabla de Contingencia (TC)
	1	2	n_1	%					
9	0	4	4	100	NE**	0.995	0.999	--	0.40
10	0	3	3	75	0.976	0.971	0.986	1.02	1.54
11	0	2	2	50	0.866	0.897	0.937	8.20	4.46
12	0	1	1	25	0.644	0.715	0.784	21.74	9.65
13	1	3	2	50	0.793	0.843	0.852	7.44	1.07
14	2	3	1	25	0.463	0.535	0.538	16.20	0.56

* n_1; número de "unos" en cada muestra
**NE; no estimable porque la varianza del error fué = 0

de las limitaciones de las TC para el análisis de variables binarias, los resultados parecen indicar que en concordancia con las expectativas teóricas (Dobson, 1990), el método de RL, actualmente disponible en varios programas de software estadístico evaluados, como los mencionados por Agresti (1990), Dobson (1990) y Wegman y Hayes (1988), debería ser preferido para el análisis de experimentos con variables dependientes binarias a los métodos de ML o TC.

Estos resultados ilustran lo que un investigador puede encontrar en la práctica al usar los diferentes métodos en datos similares, sin embargo, estos resultados preliminares son válidos sólo para un número limitado de casos estudiados, para dos tratamientos y en ausencia de otros factores de confusión categóricos o covariables. Considerando que las suposiciones del método ML no están totalmente satisfechas en cuanto a normalidad y homogeneidad de varianzas, se espera que ocurra una sobreestimación o subestimación de los valores de P, dependiente del caso estudiado. Si los valores de P para ML están subestimados, su eficiencia real comparada con la de la RL, será aun menor que la obtenida en este estudio.

REFERENCIAS

Agresti, A. 1990. *Categorical data analysis.* Wiley, New York.

Dobson, A. J. 1990. *An introduction to generalized linear models*. Chapman and Hall, London.

Gill, J. L. 1978. *Design and analysis of experiments in the animal and medical sciences.* The Iowa State University Press, Ames, Iowa.

Statistical Graphics Corp. 1996. *Statgraphics Plus for Windows 2.1.*

Wegman, E. J. and Hayes, A. R. 1988. Statistical software. *Encyclopedia of Statistical Sciences*, 8, 667-674.