ANÁLISIS DE REGRESIÓN LOGÍSTICA PARA DATOS CORRELACIONADOS UTILIZANDO TRES PROCEDIMIENTOS DEL SISTEMA ESTADÍSTICO SAS

Logistic Regression Analysis For Correlated Data Using Three Procedures of the SAS Statystical System

José Candelario Segura Correa[1], José de Jesús Solis Calderón[2] y Víctor Manuel Segura Correa[2]

Facultad de Medicina Veterinaria y Zootecnia, Universidad Autónoma de Yucatán. Km 15,5 carrera Mérida-Xmatkuil. Mérida, Yucatán, México. E-mail segura52@hotmail.com. Fax: +52 9999 423205.

[2]Centro de Investigación Regional del Sureste, INIFAP. Km 25 carretera Mérida-Motul, C.P. 97454. Mocochá, Yucatán, México

RESUMEN

El objetivo de este estudio fue comparar los resultados de regresiones logísticas para datos correlacionados obtenidos mediante tres procedimientos del sistema de análisis estadístico (SAS). Se utilizó la información de un estudio transversal sobre rinotraqueitis infecciosa bovina en el estado de Yucatán, México, donde la unidad de muestreo fue el hato y la unidad de interés el animal. Los datos se analizaron mediante los procedimientos LOGISTIC, GENMOD y NLMIXED del SAS. El modelo logístico utilizado incluyó los factores de riesgo: tamaño de hato (≤75, 76-147, 148-261, 262-480 animales) y etapa de vida del animal (crecimiento, desarrollo y producción). Se obtuvieron las razones de probabilidad (OR) e intervalos de confianza al 95% (IC95). También se calcularon las razones de probabilidad (OR) e IC95 utilizando regresiones logísticas estándar como referencia. Los ORs para tamaño de hato: ≤75, 76-147, 148-261 y 262-480 animales, utilizando los procedimientos LOGISTIC, GENMOD y NLMIXED fueron: 0,34; 0,68; 0,62 y 1; 0,34; 0,63; 0,63 y 1; y 0,29; 0,62; 0,57 y 1, respectivamente; y para los animales en crecimiento, desarrollo y producción fueron: 0,12; 0,15 y 1; 0,15; 0,17 y 1; y 0,12; 0,14 y 1, respectivamente. Los menores IC95 fueron para los ORs de la regresión logística estándar y los mayores para los OR obtenidos mediante el procedimiento NLMIXED. En conclusión, los tres procedimientos ajustaron por el efecto de hato siendo el más estricto NLMIXED.

Palabras clave: Regresión logística, modelo marginal, modelo aleatorio.

ABSTRACT

The objective of this study was to compare the results of logistic regressions for correlated data obtained by using three procedures of the Statistical Analysis System package (SAS). Information from a cross sectional study on infectious bovine rhinotracheitis in Yucatan state, Mexico, was used; where the sampling unit was the herd and the unit of interest was the animal. Data were analyzed by using the LOGISTIC, GENMOD and NLMIXED procedures of the SAS. The logistic model used included the risk factors: herd size (≤75, 76-147, 148-261, 262-480 animals) and stage of life (growth, development and production). The odd ratios (OR) and 95% confidence intervals (CI95) were obtained. Also the OR and CI95 for the risk factors were estimated using ordinary logistic regression as a reference. The ORs for herd size ≤75, 76-147, 148-261 y 262-480 animals, using LOGISTIC, GENMOD and NLMIXED procedures were: 0.34, 0.68, 0.62 and 1; 0.34, 0.63, 0.63 and 1; and 0.29, 0.62, 0.57 and 1, respectively; and for the animals in the stage of growth, development and production were: 0.12, 0.15 and 1; 0.15, 0.17 and 1; and 0.12, 0.14 and 1, respectively. The narrower CI95 were for the ORs from the ordinary logistic regression and the widest for the OR obtained using the NLMIXED procedure. In conclusion, the three procedures adjusted for herd effect, being the most precise those of the NLMIXED procedure.

Key words: Logistic regression, marginal model, random model.

Recibido: 06 / 10 / 2004. Aceptado: 17 / 03 / 2006.

INTRODUCCIÓN

Respuestas binarias o dicotómicas surgen en muchos campos de estudios. En las ciencias de la salud las respuestas dicotómicas más comunes son la presencia o ausencia de enfermedad o infección en un individuo. El análisis de regresión logística es usado para investigar la relación entre este tipo de respuestas y un conjunto de variables explicatorias o factores de riesgo. En las poblaciones de animales domésticos, éstos son explotados en grupos, comúnmente en hatos, parvadas u otro tipo de agrupaciones; por lo tanto, los hatos son usualmente la unidad de muestreo y en consecuencia las respuestas de los individuos dentro de hato no son independientes de sus compañeros de grupo. Si la respuesta dentro de hato está correlacionada, la pruebas estadísticas basadas en la suposición de independencia serían incorrectas y la varianza estimada sería más pequeña. El grado con que la varianza disminuye esta en relación con la correlación intra-conglomerado y el tamaño del hato [8]. Según Dargatz y Hill [4] no considerar el efecto de diseño de muestreo puede resultar en sesgo en la estimación del parámetro y de su varianza.

McDermott y Schukken [8] en un estudio para describir los métodos estadísticos utilizados para ajustar por el efecto de conglomerados (hatos) señalan que el 54% de los 67 trabajos revisados por ellos utilizaron algún tipo de ajuste por conglomerado. En 27 artículos se utilizó el efecto de hato como efecto fijo, 4 artículos utilizaron diseños pareados, dos utilizaron análisis ponderados y tres incluyeron hato como efecto aleatorio. También señalan que en 20 de 31 trabajos las inferencias hechas fueron incorrectas.

Se puede usar una gran variedad de métodos para ajustar por los efectos de hato. Para aquellos datos que no se distribuyen normalmente (binomial o poisson), los efectos de hato (ya sea aleatorio o fijo) han sido estimados utilizando distribuciones de verosimilitud combinadas o parámetros sobredispersos [8, 9]. El procedimiento LOGISTIC es la herramienta estándar del paquete estadístico SAS [12] para la descripción de modelos de regresión logística, pero soluciones con los procedimientos GENMOD, PROBIT, CATMOD o NLMIXED son también posibles; los cuales proporciona similares resultados para el caso de modelos de efectos fijos (modelos de regresión logística estándar). Los procedimientos LOGISTIC, GENMOD y NLMIXED permiten ajustar por el efecto de hato.

El objetivo de este estudio fue comparar los resultados de regresiones logísticas, para datos correlacionados, mediante tres procedimientos de SAS, que permiten ajustar por el efecto aleatorio de hato.

MATERIALES Y MÉTODOS

Origen de los datos y factores de riesgo

La información utilizada para explicar los diferentes procedimientos computacionales corresponde a un estudio transversal (sección cruzada) para estimar la prevalencia y los factores de riesgo asociados a la rinotraqueítis infecciosa bovina [13]. Se obtuvo información del estado serológico de 564 animales procedentes de 35 hatos. El tipo de muestreo utilizado fue un muestreo en dos etapas con un número constante de animales por hato (n = 17). Los factores de riesgo analizados fueron obtenidos utilizando una entrevista directa a los productores al momento de la toma de la muestra de sangre. Los factores de riesgo estudiados fueron: tamaño del hato (TH), densidad animal, introducción de animales al hato, grupo genético, etapa productiva del animal (EV). Debido a que el propósito de esta investigación fue comparar diferentes procedimientos sólo se utilizaron los efectos fijos de TH (≤75, 76-147, 148-261 y 262-480 animales) y EV (crecimiento, desarrollo y producción) los cuales fueron los factores de riesgo encontrados significativos ($P < 0,05$) en el estudio de Solis-Calderon y col. [13]. Asimismo se incluyó el efecto de hato para corregir por la correlación de los datos dentro de hato.

Procedimientos del SAS

Los procedimientos del sistema estadístico SAS [12] que se compararon fueron: el procedimiento LOGISTIC, el cual ajusta modelos de regresión logística lineal para datos binarios u ordinales por el método de máxima verosimilitud. Con este procedimiento se corrieron regresiones logísticas estándar y marginales. El procedimiento LOGISTIC corrige por el aumento o disminución de la dispersión de los datos, multiplicando la matriz de covarianzas por el parámetro de dispersión, que se obtiene dividiendo el valor de Ji-cuadrado por los grados de libertad. La regresión logística estándar se corrió como punto de referencia. El procedimiento GENMOD ajusta modelos lineales generales permitiendo el modelaje de datos correlacionados a través de la sentencia REPEATED, donde el método de estimación implementado es el GEE (Generalized Estimation Equation) de Liang y Zeger [5]. El procedimiento NLMIXED fija modelos mixtos no lineales**.** NLMIXED maximiza la verosimilitud directamente por métodos de integración numérica (Adaptative Gaussian quadrature), el cual permite el modelaje explícito de los efectos aleatorios. En este, la heterogeneidad entre hatos cuenta por la correlación entre animales de un hato [7].

Modelos estadísticos

El modelo de regresión logística que describió los datos mediante los procedimientos LOGISTIC y GENMOD fue:

$$\text{Logit}(p_{ijk}) = b_o + b_1X_1 + b_2X_2$$

con varianza = $p_{ijk}(1 - p_{ijk})$ y correlación $(y_{ijk}, y_{ijk}') = \alpha$.

Mientras que el modelo de regresión logística mixta que describió los datos mediante el procedimiento NLMIXED fue:

$$\text{Logit}(p_{ijk}/u_i) = b_o + b_1X_1 + b_2X_2 + u_i$$

donde y_{ijk} denota si el ijk-ésimo animal perteneciente al i-ésimo hato, j-ésimo TH y k-ésima EV fue seropositivo (yijk = 1) o seronegativo (yijk = 0) a rinotraqueitis infecciosa bovina; b_o = es el intercepto de la ecuación de regresión logística, b_1 = es el logaritmo natural (ln) de las razones de probabilidad (OR) para seropositividad de un animal perteneciente al j-ésimo TH dividido por el ln de los odds para un animal perteneciente a hatos con 262 a 480 animales; b_2 es el logaritmo natural de los odds para seropositividad para un animal perteneciente a la k-ésima EV dividido por el ln de los odds para un animal en producción; p_{ijk} es la probabilidad de un animal seropositivo y u_i es el efecto aletorio de hato con $u_i \sim N(0, \sigma_u^2)$. Los valores de OR y sus IC95 fueron obtenidos mediante la opción "estimate" del procedimiento GENMOD y como $\exp^b$ y $\exp^{(b \pm t*EEb)}$ para el procedimiento NLMIXED; donde exp se refiere a la base de los logaritmos naturales 2,178; t es el valor de tablas de t de Student y EEb denota el error estándar de b.

La bondad de ajuste del modelo para cada procedimiento, se basó en el logaritmo de la verosimilitud. El criterio de convergencia fue el mismo en los tres procedimientos (1×10^{-8}). Los estimadores de los parámetros entre procedimientos sólo se compararon subjetivamente dado que los métodos estadísticos que utilizan son diferentes. No existiendo actualmente pruebas estadísticas formales que indiquen si dos procedimientos distintos, para los mismos datos, son diferentes estadísticamente [2]. Para comparar el modelo de regresión logística estándar y el ajustado por el efecto de hato se utilizó la prueba de razón de verosimilitudes [6]. Este método se basa en que la diferencia $-2(\log L_1 - \log L_2)$ tiene una distribución Ji-cuadrada, donde L_1 y L_2 son los valores de las funciones de verosimilitud. El número de grados de libertad esta dado por la diferencia en los grados de libertad de ambos modelos.

RESULTADOS Y DISCUSIÓN

Los tres procedimientos utilizados mostraron efecto significativo de tamaño de hato y etapa de vida del animal. Las razones de probabilidad (OR) e intervalos de confianza del 95% de confianza (IC95) proporcionados por los procedimientos LOGISTIC y GENMOD sin ajustar por el efecto de hato fueron iguales (TABLA I). Los ORs del procedimiento LOGISTIC para los datos ajustados o no por el efecto de hato fueron los mismos aunque los IC95 para el procedimiento ajustado, por sobredispersión de los datos fueron mayores (TABLA II). El ajuste por efecto de hato en los análisis de regresión logística, mejoró el modelo, como lo indica la diferencia significativa (P < 0,05) de la razón de verosimilitud del modelo de regresión logística estándar y el ajustado por el efecto de hato. Los OR estimados mediante el procedimiento GENMOD fueron ligeramente diferentes y los IC95 fueron mayores cuando se consideró el efecto de hato en el modelo (TABLAS III y IV). Con respecto al procedimiento NLMIXED los OR e IC95 obtenidos fueron diferentes de los proporcionados por los modelos marginales (LOGISTIC y GENMOD), teniendo asimismo IC95 mayores que los otros dos procedimientos (TABLA IV). La varianza estimada para el efecto de hato por el procedimiento NLMIXED fue 0,44.

El problema serio que surge con el uso de las regresiones logísticas estándar, cuando los datos contienen el efecto del hato, es que no toda la variación en los hatos es tomada en cuenta; de aquí que tales datos exhiban variación binomial extra [3]. Como resultado de esto los valores de probabilidad asociados con pruebas estadísticas son menores que lo esperado bajo el nivel de significancia deseado y por lo tanto produce un sesgo hacia el rechazo de la hipótesis nula. Esta aseveración coincide con los resultados de este estudio, en donde los efectos de los factores de riesgo fueron más significativos

TABLA I

VALORES DE BETA (B), RAZÓN DE PROBABILIDAD (OR) E INTERVALOS DE CONFIANZA AL 95% (IC95) OBTENIDOS MEDIANTE EL PROCEDIMIENTO LOGISTIC O GENMOD SIN CONSIDERAR EL EFECTO DE HATO / BETA (B) VALUES, ODD RATIOS AND 95% CONFIDENCE INTERVALS (IC95) OBTAINED BY THE LOGISTIC OR GENMOD PROCEDURES WITHOUT CONSIDERING THE EFFECT OF HERD

	b	OR	IC95
Tamaño de hato			
≤75	–1,086	0,34	0,20; 0.58
76-147	–0,393	0,68	0,41; 1,12
148-261	–0,477	0,62	0,37; 1,04
262-480	0	1	
Etapa de vida			
Crecimiento	–2,107	0,12	0,06; 0,26
Desarrollo	–1,910	0,15	0,08; 0,27
Producción	0	1	

–2 Log verosimilitud = 658,3. Prueba de bondad de ajuste de Hosmer y Lemeshow (P = 0,98).

TABLA II

VALORES DE BETA (B), RAZÓN DE PROBABILIDAD (OR) E INTERVALOS DE CONFIANZA AL 95% (IC95) OBTENIDOS MEDIANTE EL PROCEDIMIENTO LOGISTIC CONSIDERANDO EL EFECTO DE HATO / BETA (B) VALUES, ODD RATIOS AND 95% CONFIDENCE INTERVALS (IC95) OBTAINED BY THE LOGISTIC OR GENMOD PROCEDURES CONSIDERING THE EFFECT OF HERD

	b	OR	IC95
Tamaño de hato			
≤75	–1,086	0,34	0,18; 0,64
76-147	–0,393	0,68	0,37; 1,24
148-261	–0,477	0,62	0,34; 1,15
262-480	0	1	
Etapa de vida			
Crecimiento	–2,107	0,12	0,05; 0,30
Desarrollo	–1,910	0,15	0,07; 0,31
Producción	0	1	

–2 Log verosimilitud = 462,4. Factor de heterogeneidad (Pearson Ji-cuadrada/DF) = 1,42. Prueba de bondad de ajuste de Hosmer y Lemeshow (P = 0,98).

TABLA III

VALORES DE BETA (B), RAZÓN DE PROBABILIDAD (OR) E INTERVALOS DE CONFIANZA AL 95% (IC95) OBTENIDOS MEDIANTE EL PROCEDIMIENTO GENMOD CONSIDERANDO EL EFECTO DE HATO / BETA (B) VALUES, ODD RATIOS AND 95% CONFIDENCE INTERVALS (IC95) OBTAINED BY THE GENMOD PROCEDURE CONSIDERING THE EFFECT OF HERD

	b	OR	IC95
Tamaño de hato			
≤75	–1,086	0,34	0,14; 0,71
76-147	–0,466	0,63	0,30; 1,31
148-261	–0,470	0,63	0,28; 1,38
262-480	0	1	
Etapa de vida			
Crecimiento	–1,908	0,15	0,07; 0,30
Desarrollo	–1,785	0,17	0,09; 0,32
Producción	0	1	

–2 Log verosimilitud = 658,3.

TABLA IV

VALORES DE BETA (B), RAZÓN DE PROBABILIDAD (OR) E INTERVALOS DE CONFIANZA AL 95% (IC95) OBTENIDOS MEDIANTE EL PROCEDIMIENTO NLMIXED CONSIDERANDO EL EFECTO DE HATO / BETA (B) VALUES, ODD RATIOS AND 95% CONFIDENCE INTERVALS (IC95) OBTAINED BY THE NLMIXED PROCEDURE CONSIDERING THE EFFECT OF HERD

	b	OR	IC95
Tamaño de hato			
≤75	–1,226	0,29	0,12; 0,71
76-147	–0,483	0,62	0,26; 1,44
148-261	–0,570	0,57	0,24; 1,35
262-480	0	1	
Etapa de vida			
Crecimiento	–2,099	0,12	0,05; 0,29
Desarrollo	–1,965	0,14	0,07; 0,30
Producción	0	1	

–2 Log verosimilitud = 643,5. Varianza de hato = 0,44.

(menor IC95) en el análisis de regresión estándar que en aquellos que ajustaron por el efecto del hato (TABLAS I-IV). Por lo tanto, el efecto de hato debería ser considerado en el diseño y análisis estadístico de los estudios por conglomerados. McDermott y col. [10] señalan que de los cinco factores de riesgo encontrados significativos en la regresión logística estándar sólo uno (pastoreo) fue consistentemente significativo en los modelos, que ellos utilizaron, para ajustar por efecto de hato. Dichos autores observaron que al incluir los efectos aleatorios, la magnitud de los efectos de la regresión disminuyó y el error estándar estimado aumento. Esto se puede apreciar en este estudio como una menor amplitud de los IC95. El aumento del error estándar de la seroprevalencia de una enfermedad en un muestreo por conglomerados en relación con el error estándar de un muestreo simple aleatorio es conocido como el efecto de diseño [1, 11], el cual varía de una enfermedad y de una población a otra.

La diferencia entre los resultados de los procedimientos LOGISTIC o GENMOD y NLMIXED se debe a que pertenecen a dos familias de modelos estadísticos, que ajustan por la dependencia o correlación de los datos de una manera diferente. Los primeros dos procedimientos, corresponden a la familia de los modelos marginales, donde el efecto de los factores de riesgo es modelado separadamente a través de la correlación intra-conglomerado. En estos modelos se ajusta por la correlación entre animales del mismo hato y se supone que esta correlación es la misma para cada par de animales del mismo hato. La varianza para una variable binomial bajo la suposición de independencia (varianza = p(1 – q)) es modificada para incorporar el factor de inflación de la varianza, α, para cada hato (varianza = α_i * ni * p_i * (1 – p_i).

El modelo de efectos aleatorios (NLMIXED) modela la heterogeneidad natural entre hatos mediante una distribución de probabilidad (la distribución normal en este estudio) ocasionando que los coeficientes de regresión varíen de un hato a otro. McDermott y col. [10] mencionan que los valores absolutos de las estimaciones de la regresión son normalmente más grandes en los modelos que incluyen efectos aleatorios y que estos valores aumentan con la variabilidad de los efectos aleatorios. Diferencias en los resultados de los modelos pueden deberse también al tipo de factores de riesgo estudiados.

La varianza de hato (0,44) proporcionada por el procedimiento NLMIXED mide el grado de heterogeneidad en la probabilidad de seroprevalencia a RIB que no puede ser explicada por los efectos fijos (factores de riesgo). Por lo tanto, el efecto de hato en este estudio es muy importante, como ya ha sido manifestado en otros casos [3, 10]. Según Curtis y col. [3] la variación binomial extra ocasionada por el efecto de hato puede ser debida a diferentes factores tales como: diferencias en la genética del hato, clima, diferencias en manejo, morbilidad y susceptibilidad a la enfermedad en cuestión, forma de llevar los registros, etc.

Aunque no es posible valorar con métodos cuantitativos cual de los tres procedimientos aquí utilizados es mejor; se recomienda el uso de los procedimientos GENMOD y NLMIXED ya que el primero proporciona opciones que permiten utilizar diferentes estructuras de la matriz de covarianzas y el segundo tiene la propiedad de incluir directamente el efecto aleatorio en el modelo. Condon y col. [2] compararon cuatro modelos comúnmente utilizados para datos binomiales correlacionados (GENMOD ajustando o no por efectos aleatorios, GLIMMIX y NLMIXED) y concluyeron que NLMIXED era el más plausible.

CONCLUSIONES E IMPLICACIONES

Los tres procedimientos mostraron efectos significativos de los factores de riesgo estudiados aunque con diferente precisión. El análisis de regresión logística estándar fue el menos estricto y el análisis del procedimiento NLMIXED el más estricto. El uso de análisis de regresión estándar para estudios de muestreo por conglomerados o multi-etapas conduce a inferencias estadísticas invalidas, rechazando con mayor facilidad las hipótesis nulas planteadas, por lo que es recomendable utilizar el modelo adecuado al tipo estudio realizado (muestreo simple o por conglomerados). Los procedimientos del SAS que se recomiendan para el análisis de estudios transversales donde las unidades de muestreo son conglomerados y se consideran factores de riesgo del hato y del huésped es el procedimiento NLMIXED cuya limitante (al menos hasta la versión 8 del SAS) es que sólo puede incluir un efecto aleatorio y el procedimiento GENMOD permite modelar diferentes estructuras de la matriz de covarianzas.

REFERENCIAS BIBLIOGRÁFICAS

[1] BENNETT, S.; WOODS, T.; LIYANAGE, W.M.; SMITH, D.L. A simplified general method for cluster-sampling survey of health in developing countries. **World Health Stat. Quarterly**. 44: 98-106. 1991.

[2] CONDON, J.; KELLY, G.; BRADSHAW, B; LEONARD, N. Estimation of infection prevalence from correlated binomial samples. **Prev. Vet. Med.** 64: 1-14. 2004.

[3] CURTIS, C.R.; MAURITSEN, R.H.; KASS, P.H.; SALMAN, M.D.; ERB, H.N. Ordinary versus random-effects logistic regression for analyzing herd-level calf morbidity and mortality data. **Prev. Vet. Med.** 16: 207-222. 1993.

[4] DARGATZ, D.A.; HILL, G.W. Analysis of survey data. **Prev. Vet. Med**. 28:225-237. 1996.

[5] LIANG, K.Y.; ZEGER, Z. Longitudinal data analysis using generalized linear models. **Biometrica** 73: 13-22. 1986.

[6] LYNCH, M.; WALSH, B. **Genetics and analysis of quantitative traits**. Massachusetts: Sinauer Associates. 1980 pp 1998.

[7] MCDERMOTT, J.J. Progress in analytical methods – more sophistication or back to basics? **Prev. Vet. Med**. 25: 121-133. 1995.

[8] MCDERMOTT, J.J.; SCHUKKEN, Y.H. A review of methods used to adjust for cluster effects in explanatory epidemiological studies of animal populations. **Prev. Vet. Med.** 18: 155-173. 1994.

[9] MCDERMOTT, J.J.; SCHUKKEN, Y.H.; SHOUKRI, M.M. Study design and analytic methods for data collected from clusters of animals. **Prev. Vet. Med**. 18: 175-191. 1994.

[10] MCDERMOTT, J.J.; KADOHIRA, M.; O'CALLAGHAN, C.J.; SHOUKRI, M.M. A comparison of different models for assessing variation in the seroprevalence of infectious bovine rhinotracheitis by farm, area and district in Kenya. **Prev. Vet. Med**. 32: 219-234. 1997.

[11] OTTE, M.J.; GUMM, I.D. Intra-cluster correlation coefficients of 20 infections calculated from the results of cluster-sample surveys. **Prev. Vet. Med**. 31: 147-150. 1997.

[12] STATISTICAL ANALYSIS SYSTEMS INSTITUTE (SAS). **User's Guide Statistic**, SAS Institute, Cary, North Carolina. 646 pp. (Version 8,1). 2000.

[13] SOLIS-CALDERON, J.J.; SEGURA-CORREA, V.M.; SEGURA-CORREA, J.C.; ALVARADO-ISLAS, A. Seroprevalence of and risk factors for infectious bovine rhinotracheitis in beef cattle herds of Yucatan, Mexico. **Prev. Vet. Med**. 57: 199-208. 2003.