Comparación de cuatro modelos de regresión para el estudio de factores de riesgo de hato para datos binomiales correlacionados

Comparison of four regression models for the study of herd risk factors for binomial correlated data

José C. Segura Correa* Antonio Salinas-Meléndez** José Jesús Solís Calderón***
José Luis Solorio Rivera†

Abstract

The objectives of this study were to compare four linear or logistic regression models and to determine their effects on the level of significance and parameter estimates, using the data from a study on seroprevalence of brucellosis in goats. Information on 5 114 does tested during 2002-2003 from 79 herds in the Bajio region in Michoacan, Mexico was used. The models were: the prevalence of seropositive animals per herd (V1), analyzed by a general linear model (GLM), herds with at least one seropositive animal, analyzed by standard logistic regression (SLRH); V1 analyzed by standard logistic regression (SLR), assuming independence among results within a same herd (SLRA); and V1 analyzed by mixed LR, considering the herd as random effect (MLR). The risk factors included in the four models were: the presence of abortions the year previous to the study, cleanness of the corral (hygiene) and length of lactation. The V1 variable transformed to arcsine-square root did not show a normal distribution. SLRH model (SLR assuming the herd as the unit of interest) and MLR were not compared because they were not nested models. MLR model adjusted the data better than the SLRA model. The deviance (–2LL) from model SLRH (70.6) was similar to their degrees of freedom (75), suggesting that the model adjusted the data very well. Levels of significance for the risk factors were different, depending of the model used. GLM and SLRH models showed significant effects ($P < 0.02$) only for the presence of abortions; SLRA model showed significant effect ($P < 0.05$) for the three risk factors, and MLR, effects of the presence of abortions and lactation length, but not for hygiene. The values for the odd ratios (OR) for the SLRA and MLR models were different; the narrowest confidence intervals corresponded to the SLRA model, and the widest to the SLRH model.

Key words: LOGISTIC REGRESSION, FIXED EFFECTS, RANDOM EFFECTS, CLUSTERS.

Resumen

Los objetivos de este estudio fueron comparar cuatro modelos de regresión lineal o logística y determinar sus efectos sobre los niveles de significancia y parámetros, utilizando los datos de un estudio de seroprevalencia de brucelosis en cabras. Se utilizó la información de 5 114 cabras en 79 hatos de la región del Bajío, en Michoacán, México, durante 2002-2003. Los modelos fueron: la prevalencia de animales seropositivos por hato (V1), analizados mediante un modelo lineal general (MLG); hatos con al menos un animal seropositivo, analizados mediante regresión logística estándar (RLEH), V1 analizada mediante RLE, suponiendo independencia entre resultados dentro de un mismo hato (RLEA) y V1 analizada mediante RL mixta, considerando al hato como efecto aleatorio (RLM). Los factores de riesgo incluidos en los cuatro modelos fueron: presencia de abortos el año anterior al estudio, limpieza del corral (higiene) y duración de la lactancia. La variable V1 transformada a arcoseno-raíz cuadrada no mostró distribución normal. El modelo RLEH (RL estándar considerando al hato como la unidad de interés) y RLM no se compararon por no ser modelos anidados. El modelo RLM ajustó mejor los datos que el modelo RLEA. La deviance (–2LL) del modelo RLEH (70.6) fue similar a sus grados de libertad (75), ello sugiere que este modelo ajustó estadísticamente bien los datos. Se encontraron niveles de significancia diferentes para los factores de riesgo, según el modelo estadístico utilizado. Los modelos MLG y RLEH mostraron efectos significativos ($P < 0.02$) sólo de presencia de abortos; el modelo RLEA mostró efecto significativo ($P < 0.05$) para los tres factores de riesgo, y el RLM, efectos de presencia de abortos y duración de la lactancia, pero no de higiene. Los valores de la razón de momios (OR) para los modelos RLEA y RLM fueron diferentes; los intervalos de confianza más estrechos correspondieron al modelo RLEA, y los más amplios, al RLEH.

Palabras clave: REGRESIÓN LOGÍSTICA, EFECTOS FIJOS, EFECTOS ALEATORIOS, CONGLOMERADOS.

Recibido el 30 de noviembre de 2006 y aceptado el 12 de enero de 2008.

*Facultad de Medicina Veterinaria y Zootecnia, Universidad Autónoma de Yucatán, Km 15.5, Carretera Mérida-Xmatkuil, Mérida, Yucatán, México, Tel. 01(999) 9423200, correo electrónico: scorrea@tunku.uady.mx
**Departamento de Microbiología y Unidad de Biotecnología, Facultad de Medicina Veterinaria y Zootecnia, Universidad Autónoma de Nuevo León, Av. Lázaro Cárdenas 4600, Unidad Mederos, Monterrey, Nuevo León, 69930, México.
***Centro de Investigación Regional del Sureste, Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias, Km 25, Carretera Mérida-Motul, 97454, Mocochá, Yucatán, México.
†Facultad de Medicina Veterinaria y Zootecnia, Universidad Michoacana de San Nicolás de Hidalgo, Av. Acueducto y Tzintzuntzan, 58000, Morelia, Michoacán, México.

Introduction

Epidemiology researchers study animal diseases in populations, where the unit of study is generally the herd, even though observations are done in the animals. The study of herds as the sampling unit, limit the use of standard logistic regression methodology (statistical tool for the study of risk factors), because it supposes independence among animals sampled within the same herd; that is, the presence of disease in an animal is independent of the presence or absence of disease in other animal. However, because of the nature of diseases, the response of animals in a same herd usually is correlated (they are not independent).

The lack of independence generally leads to a subestimation of the variability of the data, increasing the probability of rejecting the null hypothesis stated on the risk factors.[1] There are some papers in the literature that suggest how to adjust for the within herd correlation;[2] the general trend is the use of mixed (fixed and random effects) logistic regression models, which adjust using within herd correlation.[1,3] However, even today is common to find papers where models of mixed effects are not used.

To avoid the random effects in the model or by ignorance of the mixed models, epidemiologists use as response variable the seroprevalence of the herd, classifying them as positives or negatives, if they have or have not at least one positive animal or a determined level of seroprevalence.[4-6] A disadvantage of this approach is that information is not efficiently used. The existence or not of the correlation among individuals of the same herd is by itself information of interest in the design of prevention programs or control of infections.[1] Also, the criteria to classify a herd as positive or negative are chosen generally by convenience. In animal production, the dichotomic variables are generally studied as proportions, dividing the number of seropositive animals between the total number of animals sampled, and the resulting arcsine value is used in analysis of variance test,[7] which, as the case of dichotomized data, it does not use the information efficiently.

The objectives of this study were to compare four linear or logistic regression models and to determine their effect on the level of significance and parameter estimates, using data from a study of seroprevalence in goats.

Material and methods

The information on 5 114 goats from 79 herds of the state of Michoacan, Mexico, sampled from July 2002 to December 2003 was used. The study popula-

Introducción

Los epidemiólogos estudian las enfermedades de los animales en el ámbito poblacional y la unidad de estudio generalmente es el hato, aunque las observaciones se hacen en los animales. El estudio de hatos como unidad de muestreo limita el uso de la metodología de regresión logística estándar (herramienta estadística para el estudio de factores de riesgo), ya que ésta supone independencia entre cada uno de los animales muestreados dentro de cada hato; es decir, la presencia de enfermedad en un animal es independiente de la presencia o ausencia de enfermedad en otro animal. Sin embargo, debido a la naturaleza de las enfermedades, las respuestas de los animales en un mismo hato a menudo están correlacionadas (no son independientes).

El hecho de no cumplir con la independencia generalmente conduce a la subestimación de la variabilidad de los datos, incrementando la probabilidad de rechazar las hipótesis nulas planteadas sobre los factores de riesgo.[1] Existen algunos artículos en la literatura que indican cómo ajustarlos mediante la correlación dentro de hatos;[2] la tendencia general ha sido el uso de modelos de regresión logística de efectos mixtos (fijos y aleatorios), que corrigen por medio de la correlación de resultados dentro de hatos.[1,3] Sin embargo, aun hoy en día es común encontrar estudios donde no se hace uso de los modelos de efectos mixtos.

Para evitar los efectos aleatorios o por desconocimiento de los modelos mixtos, los epidemiólogos utilizan como variable de respuesta la seroprevalencia del hato, clasificándolo como positivo o negativo, según tenga o no al menos un animal positivo o determinado nivel de seroprevalencia.[4-6] Una desventaja de esta aproximación es que no se hace uso eficiente de toda la información. La existencia o no de correlación entre individuos de un mismo hato es en sí información de interés en la elaboración de programas de prevención o control de las infecciones.[1] Además, los criterios para clasificar a un hato como positivo o negativo son generalmente escogidos por conveniencia. En producción animal, las variables dicotómicas generalmente son estudiadas en forma de proporciones, dividiendo el número de animales seropositivos entre el total de animales muestreados, y el arcoseno del valor resultante se utiliza en un análisis de varianza,[7] que así como al dicotomizar grupos de datos, no hace uso eficiente de la información.

Los objetivos de este estudio fueron: comparar cuatro modelos de regresión lineal o logística y determinar su efecto sobre los niveles de significancia y parámetros, utilizando los datos de un estudio de seroprevalencia de brucelosis en cabras.

tion, the data collection, the processing methods and the epidemiological description of brucellosis seroprevalence has been described in other study.[8] The responses of interest were: the proportion of animals in the sample studied (p), proportion of seropositive animals per herd (p_l) and proportion of herds with at least one seropositive animal (p*). Information on herd-risk factors was collected by personal interview, using a questionnaire with questions about the characteristics, management and biosecurity of the herds. The response variable and the herd-risk factors evaluated in this study were described in detail in another study.[8]

The information obtained was analyzed using the following models:

Linear regression model with angular transformation of the data:

$$ar\cos ine\sqrt{p_l} = \alpha + \beta_1 A_i + \beta_2 H_j + B_3 L_k + e_{ijkl}$$

Standard logistic regression model considering the herds as positives or negatives:

$$\log it\,(p^*) = \mathrm{In}\left(\frac{p^*}{1-p^*}\right) = \alpha + \beta_1 A_i + \beta_2 H_j + B_3 L_k$$

Standard logistic regression model considering each observation as independent:

$$\log it\,(p) = \mathrm{In}\left(\frac{p}{1-p}\right) = \alpha + \beta_1 A_i + \beta_2 H_j + B_3 L_k$$

Logistic regression model with herds as random effects:

$$\log it\,(p_l) = \mathrm{In}\left(\frac{p_l}{1-p_l}\right) = \alpha + \beta_1 A_i + \beta_2 H_j + B_3 L_k$$

where:
α = intercept;
βs = regression coefficients for the risk factors considered in the model;
A_i = presence (1) or not (0) of abortions in the herd the year previous to the study;
Hj = hygiene of the herd (clean (0) or not (1) the manure of the corrals);
Lk = lactation length of the does in the herd (0 = lactation length ≤ 90, 1= lactation length > 90 days);
σ = standard deviation of the distribution of the random effects in each case (models the amount of overdispersion for each group of data given);
ul = random sample from a standardized distribution;[9]
e_{ijkl} = residual error NID $(0,\sigma^2_e)$.

Material y métodos

Se utilizó la información de 5 114 cabras en 79 hatos, en Michoacán, México, muestreadas de julio de 2002 a diciembre de 2003. La población de estudio, la colección de los datos, los métodos de procesamiento y la descripción epidemiológica de la seroprevalencia de brucelosis ha sido descrita en otro estudio. Las respuestas de interés fueron: proporción de animales en la muestra estudiada (p), proporción de animales seropositivos por hato (p_l) y proporción de hatos con al menos un animal seropositivo (p*). Se recabó información sobre los factores de riesgo de manejo del hato mediante entrevista personal, utilizando un cuestionario que contenía preguntas sobre característica, manejo y bioseguridad de los hatos. La variable de respuesta y los factores de riesgo del hato examinados en este trabajo se describieron detalladamente en otro estudio.[8]

La información obtenida se analizó utilizando los siguientes modelos:

Modelo de regresión lineal con transformación angular de los datos:

$$ar\cos eno\sqrt{p_l} = \alpha + \beta_1 A_i + \beta_2 H_j + B_3 L_k + e_{ijkl}$$

Modelo de regresión logística estándar considerando al hato como positivo o negativo:

$$\log it\,(p^*) = \mathrm{In}\left(\frac{p^*}{1-p^*}\right) = \alpha + \beta_1 A_i + \beta_2 H_j + B_3 L_k$$

Modelo de regresión logística estándar, considerando cada observación como independiente:

$$\log it\,(p) = \mathrm{In}\left(\frac{p}{1-p}\right) = \alpha + \beta_1 A_i + \beta_2 H_j + B_3 L_k$$

Modelo de regresión logística con efecto aleatorio de hato:

$$\log it\,(p_l) = \mathrm{In}\left(\frac{p_l}{1-p_l}\right) = \alpha + \beta_1 A_i + \beta_2 H_j + B_3 L_k$$

donde
α = intercepto;
βs = coeficientes de regresión para los factores de riesgo considerados en el modelo;
A_i = presencia (1) o no (0) de abortos en el hato el año anterior al estudio;
H_j = higiene del hato (limpia (o) o no (1) el corral de estiércol);
Lk = duración de la lactancia en el hato (0 = lactancia ≤ 90, 1= lactancia > 90 días);

General linear and standard logistic regression (LR) models were used to fit the prevalence of each herd, as a binomial proportion or as positive or negative herds. After, mixed LR was used to model the same relationship, but taking into account the expected extra binomial variation due to the correlation of results within herds. The risk factors included in the models were those that were significant ($P < 0.10$) in the univariable mixed LR tests; the four models included the same risk factors.

The data on the proportion of seropositive animals were transformed to angle function before analyzed using the GLM procedure of the SAS package.[10] The fixed and mixed LR models were carried out by the statistical package EGRET.[11] This last package produces parameter estimates of maximum likelihood (probability that the observed values of the dependent variable are predicted from the observed values of the independent variables) through methods of iterative optimization using the New-Raphson (standard LR) and quasi-Newton-Raphson (mixed LR) algorithms.

The choice of the best model and the null hypothesis of no herd effect was based on the likelihood ratio test = $(-1)(2LL_1-2LL_2)$, where LL_1 and LL_2 are the logarithms of the likelihood of the standard LR and mixed LR models. The term –2LL is also known as deviance. In this study, the likelihood ratio test, also tests the null hypothesis of no herd effect. According to Breslow and Day,[12] if the deviance and degree of freedom of a model are similar, then this indicates that the model adjust the data well. The statistical significance of the regression coefficients (β) was checked using the Wald's chi-square test: $\chi^2 = \beta^2/Var(\beta)$. The normality of the data for the general linear model was checked using the Wilk-Shapiro test.[11]

Results

Goodness of fit

The data on the proportion of seropositive animals, transform to arcsine-root square, did not show a normal distribution according to the Wilk-Shapiro test ($P < 0.001$). The linear regression models for the data transform to angles and the LR models were not compared, because they uses two different methodologies. The linear model uses least square procedures and the LR model uses maximum likelihood procedures. Base on the likelihood ratio test (lower value is better), the mixed LR model adjusted the data better than the standard LR model, assuming independence between the results of seropositivity of each animal. The deviance (–2LL) of the standard LR model for the positive or negative herd variable (70.6) was similar to their degree of freedom (75), which suggests that

σ = desviación estándar de la distribución de los efectos aleatorios en cada caso (modela la cantidad de sobredispersión para un grupo de datos dado);
ul = muestra aleatoria de una distribución estandarizada;[9]
e_{ijkl} = error residual NID $(0,\sigma^2_e)$.

Se usaron modelos lineales generales y de regresión logística (RL) estándar para modelar la prevalencia de cada hato, como proporción binomial o como hato positivo o negativo. Luego se usó RL binomial mixta para modelar la misma relación, pero considerando la variación binomial extra esperada por la correlación de resultados dentro de los hatos. Los factores de riesgo incluidos en los modelos fueron los que resultaron significativos ($P < 0.10$) en pruebas univariadas de RL mixta; los cuatro modelos incluyeron los mismos factores de riesgo.

La proporción de animales seropositivos transformados a ángulos fue analizada usando el procedimiento MLG del paquete SAS.[10] Las RL de efectos fijos y mixtos se realizaron mediante el paquete estadístico EGRET.[11] Este último produce estimadores de máxima verosimilitud (probabilidad de que los valores observados de la variable dependiente sean predichos a partir de los valores observados de las variables independientes) mediante métodos de optimización iterativos a través de los algoritmos de Newton-Raphson (RL estándar) y quasi-Newton-Raphson (RL mixta).

La selección del mejor modelo y la hipótesis nula de no efecto de hato se basaron en la prueba de razón de verosimilitudes = $(-1)(2LL_1-2LL_2)$ donde LL_1 y LL_2 son los logaritmos de verosimilitud de los modelos de RL estándar y RL mixta. El término –2LL es también conocido como lejanía o *deviance*. En este estudio, la prueba de razón de verosimilitud también prueba la hipótesis nula de no efecto de hato. Según Breslow y Day,[12] si la *deviance* y los grados de libertad de un modelo son relativamente iguales, entonces esto indica que el modelo ajusta bien los datos. La significancia estadística de los coeficientes de regresión (β) fue corroborada usando la prueba de Ji-cuadrada de Wald: $\chi^2 = \beta^2/Var(\beta)$. La normalidad de los datos del modelo de regresión lineal se comprobó utilizando la prueba de Wilk-Shapiro.[11]

Resultados

Bondad de ajuste

Los datos de proporción de animales seropositivos transformados a arcoseno-raíz cuadrada no mostraron distribución normal de acuerdo con la prueba de Wilk-Shapiro ($P < 0.001$). Los modelos de regresión lineal para los datos transformados a ángulos y

the model adjusted the data well. The deviance of the standard LR model without adjustment for herd effect was 5.2 times their degree of freedom (75), which suggests a greater variability than that explained by the binomial distribution and a bad adjustment of the data. Also, the mixed LR model did not adjust the data well, because its deviance was 2.6 times the degrees of freedom of the model (74). However, the difference of the deviance of the standard LR without adjusting for the herd effect and the mixed LR (difference = 206.1) models suggest that the inclusion in the model of the herds as random effect was appropriated.

Levels of significance and parameter estimates

Different levels of significance were found for the risk factors, depending on he model used. The linear regression and LR models for the herd-dichotomized variable showed significant effects ($P < 0.02$ and $P < 0.002$, respectively) only on the presence of abortions the year previous to the study. The standard LR using each data as independent, showed significant effect ($P < 0.05$) of the three risk factors, and the mixed LR only showed effect of the variables: presence of abortions and lactation length, but not of hygiene (Table 1).

The OR values for the standard LR model under the assumption of independence and mixed LR were different; the narrowest confidence intervals were for the standard LR, which considered each observation as independent, and the widest for the standard LR that considered the herds as positives or negatives (Table 2). The value of the variance component estimate, $\sigma^2_e = 0.972$, in relation to its standard error (0.084), reflect the heterogeneity of the seroprevalences among herds.

Discussion

Goodness of fit

The transformation of the data to proportion of seropositive animals using the arcsine-square root function did not approximate the data to the normal distribution ($P = 0.001$), which violates the assumption of the data imposed by the least squares methodology that use the GLM procedure of the SAS package. The consequence of this is that the significant levels of the risk factors here studied were biased, as indicated by the results of the mixed LR, taken as the golden test. When the linear regression is used to fit binomial data (expressed as percentage), three problems occur: the error variance is not constant, the error term is not

los modelos de RL no se compararon, ya que utilizan dos metodologías diferentes. El modelo lineal utiliza procedimientos de cuadros mínimos, y los modelos de RL utilizan procedimientos de máxima verosimilitud. Basado en la prueba de razón de verosimilitudes (menor valor es mejor), el modelo de RL mixta ajustó mejor los datos que el modelo de RL estándar, suponiendo independencia entre resultados de seropositividad de cada animal. La *deviance* (–2LL) del modelo de RL estándar para la variable hato positivo o negativo (70.6) fue similar a sus grados de libertad (75), lo que sugiere que este modelo ajustó estadísticamente bien los datos. La *deviance* del modelo de RL estándar sin ajustar por el efecto de hato fue 5.2 veces sus grados de libertad (75), ello sugiere una variabilidad mayor que la explicada por la distribución binomial y un mal ajuste de los datos. Asimismo, el modelo de RL mixta no ajustó bien los datos, ya que su *deviance* fue 2.6 veces los grados de libertad del modelo (74). Sin embargo, la diferencia de las *deviance* de los modelos de RL estándar sin ajustar por el efecto de hato y la RL mixta (diferencia = 206.1), sugiere que la inclusión del efecto aleatorio de hato en el modelo fue apropiada.

Niveles de significancia y parámetros

Se encontraron niveles de significancia diferentes para los factores de riesgo, según el modelo estadístico utilizado. Los modelos de regresión lineal y RL estándar para la variable hato dicotomizado, mostraron efectos significativos ($P < 0.02$ y $P < 0.002$, respectivamente) sólo de presencia de abortos el año anterior al estudio; la RL estándar, considerando cada dato como independiente, mostró efecto significativo ($P < 0.05$) para los tres factores de riesgo, y la RL mixta, efectos de presencia de abortos y duración de la lactancia, pero no de higiene (Cuadro 1).

Los valores de OR para los modelos de RL estándar bajo la suposición de independencia y RL mixta fueron diferentes; los intervalos de confianza más estrechos correspondieron a la RL estándar, que considera cada observación como independiente, y los más amplios a la RL estándar, que consideró al hato como positivo o negativo (Cuadro 2). El valor del estimador de la componente de varianza, $\sigma^2_e = 0.972$ en relación con su error estándar (0.084), refleja la heterogeneidad de las seroprevalencias entre hatos.

Discusión

Bondad de ajuste

La transformación de los datos de proporción de animales seropositivos, mediante las funciones arcoseno-raíz cuadrada, no aproximó los datos a la distribución

Cuadro 1

VALORES DE PROBABILIDAD DE LA PRUEBA DE JI-CUADRADA DE WALD, PARA ALGUNOS FACTORES DE RIESGO DEL HATO, USANDO MODELOS DE REGRESIÓN FIJOS Y ALEATORIOS EN CABRAS

PROBABILITY VALUES FOR THE WALD'S CHI-SQUARE TEST FOR SOME HERD-RISK FACTORS IN GOATS, USING FIXED AND RANDOM LOGISTIC REGRESSION MODELS.

	Risk factors			
Model	*Abortion*	*Hygiene*	*Length of lactation*	*–2 likelihood logarithm*
GLM	0.0221	0.9108	0.6525	-
Standard LR[a]	0.0018	0.1012	0.2555	70.60 (75)
Standard LR[b]	0.0010	0.0029	0.0308	386.9 (75)
Random LR	0.0150	0.9935	0.0139	189.9 (74)

GLM = General linear models; LR = logistic regression; [a]Herd dichotomized; [b]Each observation was considered independent; between parenthesis the degrees of freedom.

Cuadro 2

RAZÓN DE MOMIOS E INTERVALOS DE CONFIANZA A 95% PARA ALGUNOS FACTORES DE RIESGO DEL HATO, PARA SEROPREVALENCIA EN CABRAS, USANDO MODELOS DE REGRESIÓN LOGÍSTICA (RL) DE EFECTOS FIJOS Y MIXTOS

ODD RATIOS AND 95% CONFEDENCE INTERVALS FOR SOME HERD-RISK FACTORS, FOR SEROPREVALENCE IN GOATS, USING FIXED AND RANDOM LOGISTIC REGRESSION MODELS

	Risk factors		
Model	*Abortion*	*Hygiene*	*Length of lactation*
Standard LR[a]	7.57 (2.12, 27.06)	2.86 (0.81, 10.02)	0.49 (0.14, 1.68)
Standard LR[b]	2.05 (1.41, 2.98)	0.63 (0.47, 0.85)	1.24 (1.02, 1.50)
Random LR	2.12 (1.57, 3.88)	1.00 (0.64, 1.58)	1.49 (1.08, 2.04)

RL = Logistic regression; [a]Herd dichotomized; [b]Each observation considered independent.

normally distributed (as observed in this study) and may predict percentage out of range 0-1.[13]

With respect to the results of the LR, based on the similarity of the deviance value and degrees of freedom,[11,14] the best model was LR for herds classified as positive or negative. The advantage of this model in comparison with the mixed LR model might be explained by the presence of extra binomial variation, even though the herd effect was included in the model. This result disagrees to that observed in a paper on pig mortality,[15] in which, three of the LR models here used were compared. In that study, it was found that the mixed LR fitted better the data on mortality at birth. The extra binomial variation might be explained partially due to management differences between herds and to the relatively large size of the herds. The extra binomial variation depends on herd size and the intracluster correlation; greater extra binomial variation is observed in clusters of

normal (P = 0.001), lo cual viola la suposición de normalidad de los datos impuesta por la metodología de cuadrados mínimos que utiliza el procedimiento MLG del paquete SAS. La consecuencia de esto último es que los niveles de significancia de los factores de riesgo estudiados estuvieron sesgados, como lo indican los resultados de la RL de efectos mixtos, considerada como la prueba de oro. Cuando se usa la regresión lineal para ajustar datos binomiales (expresados como porcentajes), surgen tres problemas: la varianza del error no es constante, el error no se distribuye normalmente (como se observó en este estudio) y predice porcentajes fuera del rango de 0 a 1.[13]

Con respecto a los resultados de la RL, basados en la similitud de los valores de *deviance* y grados de libertad,[11,14] el mejor modelo de ajuste fue el de RL para hatos clasificados como positivos o negativos. La ventaja de este modelo, en comparación con el modelo de RL mixta, puede explicarse debido a la gran variación

large size.[2] Noordhulzen *et al.*[16] mention that the use of random effect models is the best way of dealing with the effect of clusters, and that they increase the standard error of the coefficients and could change their values.

Levels of significance and parameter estimates

The cluster of animal data in herds classified as positives or negatives prevent errors in the statistical inferences, because the sampling unit is the herd.[2] However, using this approach, much information is lost, because herds with 1% prevalence are classified with herds with 100% prevalence, which conduce to a lost of power of the statistical test.[2] Also, the dichotomization of herds to justify the use of standard LR models does not take into account the heterogeneity among herds and the intraclass correlation of the data, neither allow for the study of animal-risk factors. The differences in the significant levels obtained with this model (compared with the mixed LR model) could conduce to establish prevention and control programs directed to the wrong risk factors.[17]

To use standard LR without considering the effect of herd assumes that the animals sampled within each herd are independent; in consequence, the P values associated with the statistical tests, normally are smaller and produce bias towards the alternative hypothesis.[1] In this study, the lowest Wald's significant values and narrowest confidence intervals were for the standard LR model, with the assumption of independence. Some authors[18] suggest to reduce the level of significance at 1%, as a measure to reduce the bias in such levels. However, this is not an adequate solution, because depending on the model used (marginal, random etc) the point and dispersion estimates are also affected.[19]

Mixed LR allow to model the prevalence within herds without the need of dichotomize the data; therefore, the information on the seropositivity of the animals within herds is not lost. Also, the mixed models take into account the heterogeneity of the risk of disease among herds, and they are considered the model of choice for the study of correlated binomial data. The inclusion of herd in the model changed the interpretation of the contribution of some of the risk factors and its association with the seropositivity of exposition to the causal agent of brucellosis. Changes in the magnitude of the levels of significance of risk factors have been notified by other authors.[3,5,17]

In a study of risk factors for Bovine Herpes Virus type 1 (BHV1), Schukken *et al.*[1] found that the herd-risk factors (herd size, grazing, and control programs) showed important differences in the size of the param-

extrabinomial existente, a pesar de que ésta incluyó el efecto de hato en el modelo. Este resultado es contrario al observado en un trabajo sobre mortalidad en cerdos,[15] en el que se compararon los tres modelos de RL evaluados en este estudio. En aquél se encontró que el modelo de RL mixta ajustó mejor los datos de mortalidad al nacer. La variación binomial extra puede explicarse, en parte, por la diferencia en el manejo y tamaño relativamente grande de los hatos. La variación binomial extra depende del tamaño del hato y la correlación intraconglomerados; se observa mayor variación extrabinomial en los conglomerados de mayor tamaño.[2] Noordhulzen *et al.*[16] mencionan que el uso de modelos de efectos aleatorios es la mejor forma de tratar con el efecto de conglomerados, y que éstos aumentan el error estándar de los coeficientes y pueden cambiar sus valores.

Niveles de significancia y parámetros

La agrupación de los datos de cada individuo en hatos clasificados como positivos o negativos previene de cometer inferencias estadísticas, ya que la unidad de muestreo es el hato.[2] Sin embargo, utilizando esta aproximación se pierde mucha información, ya que los hatos con 1% de prevalencia son clasificados con hatos con 100% de prevalencia, lo que conduce a pérdida de poder de la prueba.[2] Además, dicotomizar los hatos para justificar el uso de modelos de RL estándar no considera la heterogeneidad entre hatos, y la correlación de resultados dentro de éstos no permite estudiar factores de riesgo del animal. Las diferencias en los niveles de significancia obtenidos con este modelo (en comparación con el modelo de RL mixta) podría conducir a establecer programas de prevención y control dirigidos a los factores de riesgo equivocados.[17]

Utilizar RL estándar sin considerar el efecto de hato es suponer que los animales muestreados dentro del hato son independientes; en consecuencia, los valores de P asociados con las pruebas estadísticas normalmente son más pequeños y producen sesgo hacia la hipótesis alterna.[1] En este estudio, los valores de significancia de Wald y los intervalos de confianza más estrechos correspondieron al modelo de RL estándar, con la suposición de independencia. Algunos autores[18] sugieren reducir el nivel de significancia a 1%, como medida para disminuir el sesgo en dicho nivel. Sin embargo, ésta no es una solución adecuada, ya que dependiendo del modelo utilizado (marginal, aleatorio), los estimadores de punto y dispersión también son afectados.[19]

Las RL mixtas permiten modelar la prevalencia de hato sin tener que dicotomizar; por lo tanto, la información sobre la seropositividad de los animales dentro del hato no se pierde. Además, los modelos

eter estimates and their standard errors. For example, the estimate for herd size changed from 1.33 in the fixed effect model to 0.46 for the random effect model, whereas, the standard error changed from 0.19 to 0.46. Therefore, when risk factors are evaluated for a given disease, the within herd correlation must be considered to reach correct conclusions about the impact of risk factors. The values of the LR parameter estimates are usually greater in the models adjusted for the random effects and increase with the variation of this effects.[17] Similar results were observed in this study, where the OR values increased when the herd effect was included in the model (Table 2). According to Curtis *et al.*,[3] the parameter estimates obtained from mixed LR models are more trustable than those obtained from standard LR models, under the assumption of independence, because this model commonly provides smaller standard errors. Other authors16 mention an increase in the standard error of the coefficients with the use of LR models with random effects and changes in the regression coefficients with regard to the standard LR results.

In this study, the variation due to herd was different from zero, as indicated by the significant value (206.1) of the ratio likelihood test for the standard LR with assumption of independence and the mixed LR model, and by the herd variance (0.972 ± 0.084), which indicates that the inclusion of the herd effect in the model was adequate, because it affected the levels of significance and the OR values of the risk factors. The extra binomial variation could be due to differences in herd management, susceptibility in the herds to the diseases of interest, herd size, environmental effects, etc. Also, overdispersion may occur when important risk factors are not included in the model or when within-herd correlation exist.[19]

In conclusion, the results of this study show that the use of different linear or logistic regression models change the levels of significance and the magnitude of the regression coefficients, which might conduce to different results of the studied risk factors. The mixed LR model allowed that the within herd prevalence could be modeled without having to dichotomize the information; therefore, it did a better use of the available information. Also, because the sampling design used in this study was a two-step-cluster design, mixed LR models are recommended, because they take into account the random effect of herds.

Referencias

1. Schukken YH, Grohn YT, McDermott B, McDermott JJ. Analysis of correlated discrete observations: background, examples and solutions. Prev Vet Med 2003;59: 223-240.

mixtos consideran la heterogeneidad del riesgo de enfermedad entre hatos, por lo que son considerados como modelo de elección para el estudio de datos binomiales correlacionados. La inclusión de hato en el modelo cambió la interpretación de la contribución de algunos factores de riesgo y su asociación con la seropositividad a la exposición del agente causal de la brucelosis. Los cambios en la magnitud de los niveles de significancia de los factores de riesgo han sido notificados por otros autores.[3,15,17]

En un estudio de factores de riesgo para prevalencia de herpes virus bovino tipo 1 (BHV1), Schukken *et al.*[1] encontraron que los factores de riesgo medidos en el hato (tamaño de hato, pastoreo y programas de control) mostraron importantes diferencias en el tamaño de los parámetros y sus errores estándares. Por ejemplo, el parámetro para tamaño de hato cambió de 1.33 en el modelo de efectos fijos a 0.46 para el modelo de efectos aleatorios, mientras que el error estándar cambió de 0.19 a 0.46. Por lo tanto, cuando se evalúan factores de riesgo para una enfermedad, la correlación entre individuos dentro del hato debe ser considerada para alcanzar conclusiones correctas acerca del impacto de los factores de riesgo. Los valores de los estimadores de regresión de la RL son usualmente mayores en los modelos ajustados por los efectos aleatorios y aumentan con la variabilidad de estos efectos.[17] Resultados similares se observaron en este estudio, en donde los valores de OR aumentaron al incluir el efecto de hato en el modelo (Cuadro 2). Según Curtis *et al.*,[3] los estimadores de los parámetros obtenidos de los modelos de RL mixta son más confiables que los obtenidos de la RL estándar con suposición de independencia, ya que ésta comúnmente presenta errores estándar más pequeños. Otros autores[16] mencionan un aumento en los errores estándar de los coeficientes con el uso de modelos de RL con efectos aleatorios y cambios en los coeficientes de regresión con respecto a los resultados de la RL.

La variación debida al hato en este estudio fue diferente a cero, como lo indican el valor (206.1) significativo de la prueba de razón de similitudes para los modelos de RL estándar con suposición de independencia, y de RL mixta y la varianza de hato (0.972 ± 0.084), ello sugiere que la inclusión de efecto de hato en el modelo fue apropiada, ya que afecta los niveles de significancia y los valores de OR de los factores de riesgo. La variación extrabinomial pudo deberse a diferencias en el manejo de los hatos, susceptibilidad de los hatos a la enfermedad en cuestión, tamaño del hato, microclima, etc. Asimismo, la sobredispersión puede ocurrir cuando los factores de riesgo importantes no se incluyen en el modelo o cuando existe correlación dentro de los hatos.[19]

En conclusión, los resultados de este estudio mues-

tran que el uso de diferentes modelos de regresión lineal o logística modifican los niveles de significancia y magnitud de los coeficientes de regresión, lo que podría conducir a diferentes resultados sobre los factores de riesgo estudiados. El modelo de RL mixta permitió que la prevalencia entre hatos se modelara sin tener que recurrir a la dicotomización; es decir, hizo mejor uso de la información disponible. Asimismo, puesto que el diseño de este estudio consistió en un muestreo por conglomerados, deberían usarse modelos estadísticos como los de RL mixta, que contemplan los efectos aleatorios de hato.

2. McDermott JJ, Schukken YH. A review of methods used to adjust for cluster effects in explanatory epidemiological studies of animal populations. Prev Vet Med 1994;18: 155-173.
3. McDermott JJ, Schukken YH, Shoukri MM. Study design and analytic methods for data collected from clusters of animals. Prev Vet Med 1994;18: 175-191.
4. Curtis CR, Mauritsen RH, Kass PH, Salman MD, Erb HN. Ordinary *versus* random-effects logistic regression for analyzing herd-level calf morbidity and mortality data. Prev Vet Med 1993;16:207-22.
5. Solorio-Rivera JL, Rodriguez-Vivas RI, Perez-Gutierrez E, Wagner G. Management factors associated with *Babesia bovis* seroprevalence in cattle from eastern Yucatan, Mexico. Prev Vet Med 1999;40: 261-269.
6. Riveriego FJ, Moreno MA, Dominguez L. Risk factors for brucellosis seroprevalence of sheep and goat flocks in Spain. Prev Vet Med 2000;44:167-173.
7. Al-Talafhah AH, Lafi SQ, Al-Tarazi Y. Epidemiology of ovine brucellosis in Awassi sheep in Northern Jordan. Prev Vet Med 2003;60:297-306.
8. Steel RGD, Torrie JH. Principles and Procedures of Statistics. A Biometrical Approach. 2nd ed. New York: McGraw-Hill Book Company. 1980.
9. Solorio-Rivera JL, Segura Correa JC, Sanchez-Gil LG. Seroprevalence of antibodies and risk factors for brucellosis of goats in the Bajio region of Michoacan, Mexico. Prev Vet Med 2007, 82 (In press).
10. Cochran Ch, Coull B, Patel A. EGRET Users Manual for Windows (Version 2.0.3) Seattle WA: Cytel Software Corporation. 1999.
11. SAS. SAS/STAT User's Guide (Version 8.1) Cary NC, USA: SAS Inst. Inc. 2000.
12. EGRET for Windows (Version 2.0.3) Seattle WA: Cytel Software Corporation. 1999.
13. Breslow NE, Day NE. Statistical Methods in Cancer Research. Vol. I. The analysis of Case-Control Studies. International Agency for Research on Cancer. Lyon, France: Scientific Publications No. 32, 1980.
14. Zhao L, Chen Y, Schaffner DW. Comparison of logistic regression and lineal regression in modeling percentage data. Appl Environ Microbiol 2001; 67: 2129-2135.
15. Hosmer DW, Lemeshow S. Applied Logistic Regression. New York: Wiley, 1989.
16. Segura-Correa JC, Alzina-López A, Solorio-Rivera JL. Evaluación de tres modelos y factores de riesgo asociados a la mortalidad de lechones al nacimiento en el trópico de México. Téc Pecu Méx 2007; 45:227-236.
17. Noordhulzen JPTM, Frankena K, Van der Hoofd CM, Graat EAM. Application of Quantitative Methods in Veterinary Epidemiology. Wageningen: Wageningen Pers, 1997.
18. McDermott JJ, Kadohira M, O´Callaghan CJ, Shoukri MM. A comparison of different models for assessing variation in the sero-prevalence of infectious bovine rhinotracheitis by farm, area and district in Kenya. Prev Vet Med 1997;32: 219-234.
19. Bendixen PH, Vilson B, Ekesbo I, Astrand DB. Disease frequencies of tied zero-grazed dairy cows and of dairy cows on pasture during summer and tied during winter. Prev Vet Med 1986; 4: 291-306.
20. Condon J, Kelly G, Bradshow B, Leonard N. Estimation of infection prevalence from correlated binomial samples. Prev Vet Med 2004;64: 1-14.