Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 7 (12): 1762-1767, 2007
ISSN 1812-5654

The Application of Overdispersion and Generalized Estimating Equations in Repeated Categorical Data Related to the Sexual Behaviour Traits of Farm Animals

Abdullah Yeşilova and Ayhan Yılmaz
Department of Animal Science, Faculty of Agriculture, Yüzüncü Yil University, Turkey

Abstract: In this study, the Poison regression, negative binomial regression and generalized estimating equations were applied to the repeated measurements based on count data obtained from the sexual behaviors of ram lambs. Negative binomial regression was more effective to handle the over dispersion that causes bias in parameter estimations in Poison regression. The generalized estimating equations were used for analyzing repeated categorical data. GEE estimates were obtained by using the exchangeable working correlation. As a result of GEE analyses, it was concluded that flehmen lip curl response, tail raising, mount duration, vocalization and weight of the ram lamb were statistically important ($p<0.05$) for mount frequent. However, the anogenital sniff found be not significant.

Key words: Over dispersion, Poison regression, negative binominal regression, generalized estimating equations

INTRODUCTION

When researchers gather data from individual unit on multiple occasions (i.e., repeated measurements), such data are routinely analyzed with classical univariate ANOVA methods (e.g., linear regression, repeated measures design ANOVA). These procedures assume normality, independence of observations and equality of variance (or sphericity). Data obtained in many experimental setting (e.g., longitudinal) rarely satisfy these assumptions (Frome *et al*., 1973; Agresti, 1997; Gardner *et al*., 1995). The response variables used in this study were collected within subjects across time. Poison or negative binomial regression using Generalized Linear Models (GLM) has been accepted as the best means of estimating probabilities in cases in which the dependent variable consists of counted data (Gardner *et al*., 1995; Cameron and Trivedi, 1998).

The most significant property of the Poison Regression (PR) is the equality of the mean and variance. But in the practice this equality does not realize always. When the variances derived from the data are higher or lower than mean in the model, the data may be over-or underdispersed (Cox, 1983; Breslow, 1990; Stokes *et al*., 2000). In such cases ordinary PR produces biased estimates for the response variable modeled. Instead, quasi likelihood methods (Mccullagh and Nelder, 1989; Breslow, 1990) and random effect models and mixture models are advised (Lawles, 1987; Wang *et al*., 1996, 1998; Dalrymple *et al*., 2003)

When analyzing counted data, a Negative Binomial Regression (NBR), one of the models of random effects (Wang *et al*., 1996) should be specified in cases in which the dispersion is high. In NBR uses log link function, which links the linear structure of the explanatory variables to the expected value of the dependent variable, to model the data (Frome, 1983; Mccullagh and Nelder, 1989; Dobson, 1990; Agresti, 1997; Cameron and Trivedi, 1998). Negative binomial allows for extra-Poison variation du to other variables not included in the model.

Generalized Estimating Equations (GEE) approach for counted data was developed by Liang and Zeger (1986) to produce more efficient and unbiased regression estimates for use in analyzing longitudinal or repeated measures research designs with non-normal response variables. GEE provides a semi-parametric approach on individuals for observances obtained in longitudinal data (Stokes *et al*., 2000). The GEE for estimating parameter is an extension of the independent estimating equation to correlated data. In contrast to the method of estimating equation requires only assumptions of relevant moments such as means, variances and correlation. GEEs permit specification of distributions from the exponential family of distributions, which includes normal, inverse normal,

Corresponding Author: Abdullah Yeşilova, Department of Animal Science, Faculty of Agriculture, Yüzüncü Yil University, Turkey Tel: +90 0432 225 10024 Fax: +90 0432 2251104

binomial, Poison, negative binomial and Gamma distributions. As in GLMs, the variance needs to be expressed as a function of the mean; this is then incorporated in the calculation of the covariance matrix by multiplying the components against an N×N matrix (W) with a value W_i (i = 1,2,…,N) on the diagonal that is determined by the variance function (Davis, 2002; Stokes *et al*., 2000). GEE assume the relationship between the mean and variance of the responses is known, but remaining characteristics of the distribution need not be specified. Experimental units are assumed to be independent, but observations over time from a given unit are allowed to be correlated. This correlation is considered a nuisance to be adjusted for but not of interest in and of itself (Okut *et al*., 1999).

MATERIALS AND METHODS

Materials: The animal material of this study was carried out Norduz Ram lambs rearing in the Investigation and Practicing Farm of the Agriculture Faculty of Yüzüncü Yil University in Van province of Turkey in 2003. The studies of determination sexual behaviors of the ram lambs were started when the ram lambs were six months old. Sexual behaviors tests were done once a fortnight until they were 12 months old and between 12 and 13 just one test was carried out. In order to value the sexual behaviors of the each lamb at a 5.40×5.00 m^2 with 1-3 stimulus sheep for 15 min was individually tested. The stimulus sheep used in the test were fixed during the mating season in the evening by using the searching rams, in the same evening were taken to the testing department and were got accustomed to there. Apart from mating season, inter vaginal sponges were applied to 6-9 sheep for 12-14 days. The sponges were taken out in the early morning and after that 500 IU PMSG was injected into each sheep intramuscularly. Between 24-72 h following the PMSG injection in the morning and evening hours a rake of estrus was done and sheep giving estrus symptoms were distinguished from the others and immediately sexual behaviors of male lambs were analyzed. In order to remove the effect of the test time the ram lambs were tested after being selected randomly. During the testing the same test department was used. The determination of sexual behaviors of the male lambs was done as described Price (1993).

Mount duration: The time that passed until the ram lamb determined the oestrus sheep and mounted on it.

Mount frequent: The number of mountings that were done by Ram lamb in order to determine the oestrus sheep and display the act of climbing over. The intention of mounting and mounting itself was appraised together (Katz *et al*., 1988).

Flehmen lip curl: The act of curling its upper lip upward performed by ram lamb after sniffing its urine and its genital zone. At the same time, the male lambs were recorded to anogenital sniffing, the ability of tail raising and a frequency of vocalization.

Methods

Poison regression: In the PR the y_i (i = 1,…,n) dependent variable which is the number of the interested event is supposed to have the Poison distribution when the x_i independent variables are given. In this case the logarithm of μ, the average of the Poison, is supposed to be a linear function of independent variables (SAS, 2005). The function, using GLMs, is given as,

$$\log(\mu) = b_0+b_1x_1+...+b_nx_n$$

or

$$\mu = \exp(b_0+b_1x_1+...+b_nx_n) \quad (1)$$

Describing the dispersion parameter that is resulted from the inequality of the means and the variance, it possible to explain the dispersion occurring in the PR. When we take the dispersion parameter ø into account, the relation between the means and the variance is described as,

$$Var(Y_i) = \phi\mu$$

When ϕ>1, it is assumed that there is over dispersion (SAS, 2005).

Negative binomial regression: In GLM, Poison and negative binomial regression models share similar procedures to estimate the parameters in the model. The expected value of y_i denoted by μ_i for given x_i is

$$\mu(x) = g(x;\beta)$$

Where, $g(x;\beta)$ is a positive valued function of the x, β is the vector of regression parameters.

In the log linear form, it is described as,

$$g(x;\beta) = \exp(x_i'\beta)$$

NBR model is given as,

$$\Pr(Y = y \mid x) = \frac{\Gamma(y+a^{-1})}{y!\Gamma(a^{-1})}\left(\frac{\mu(x)}{1+a\mu(x)}\right)^y\left(\frac{1}{1+a\mu(x)}\right)^{a^{-1}}$$

$$, a>0,y=0,1,... \quad (2)$$

In Eq. 2, a is called dispersion parameter for the model. In the NBR model, the means and the variance can be given as,

$$E(Y_i) = \mu(x)$$

and,

$$Var(y_i) = \mu(x)+a\mu(x^2)$$

(Lawles, 1987; Yeşilova, 2003). The distribution of response variables in NBR model is

$$Y \approx Poison(\mu(x),a)$$

If a = 0, then the model reduced to Poison model.

When we consider the ϕ as distribution parameter then the variance of NBR would be rewritten as,

$$Var(Y_i) = \phi(\mu+a\mu^2)$$

The negative binomial distribution contains a parameter a, called the negative binomial dispersion parameter. This is not the same as the GLM dispersion ϕ, but it is an additional distribution parameter that must be estimated or set to a fixed value (SAS, 2005).

Generalized estimating equations: The GEE approach of Zeger and Liang facilitates analysis of data collected in longitudinal, nested, or repeated measures designs. GEEs use the GLM to estimate more efficient and unbiased regression parameters relative to ordinary least squares regression in part because they permit specification of a working correlation matrix that accounts for the form of within-subject correlation of responses on dependent variables of many different distributions, including normal, binomial and Poison (Davis, 2002; Stokes *et al*., 2000). Correlated data are modeled using the same link function and linear predictor setup (systematic component) as in the independence case. The random component is described by the same variance functions as in the independence case, but the covariance structure of the correlated measurements must also be modeled (Littell *et al*., 1996; SAS, 2005).

GEE uses quasi-likelihood methods for estimating parameters. To define a quasi-likelihood function only the first two moments of the dependent variable need to be specified. The motivation for using quasi-likelihood methodology is clear: A possible consequence of violating the distributional assumption is that the estimates of the parameters or their variances estimated by maximum likelihood may be biased, or may not have an asymptotic normal distribution (Okut *et al*., 1999). The estimate from quasi-likelihood is no longer a likelihood estimate, since the quasi-likelihood may not be a likelihood function at all. Nevertheless, the quasi-likelihood estimate possesses all of the properties of likelihood estimate. In the quasi-likelihood regression method, both the marginal mean and the marginal variance are assumed, although the marginal variances or their parameters are often not of interest in most regression problems (Breslow, 1990).

The GEE for parameters vector is an extension of the independence estimating equation to correlated data and is given by (Littell *et al*., 1996; SAS, 2005). Therefore, in order to estimate β, GEE, similar to GLM, can be written as,

$$\sum_{i=1}^{n} \frac{\partial \mu'}{\partial \beta} V_i^{-1}\left(Y_i - \mu_i(\beta)\right) = 0 \quad (3)$$

In the Eq. 3, μ_i, $\mu_i = (\mu_{i1},...,\mu_{it_i})'$ is the vector of the means; $Y_i = (Y_{i1},...,Y_{it_i})$ is the observation vector and V_i is an estimation of the covariance matrix. All these described equations are similar the GLM equations. Typically, covariance V_i is written as;

$$V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2} \quad (4)$$

In the Eq. 4, ϕ is scale parameter, A_i is a $t_i \times t_i$ diagonal matrix, A_i = diag $[V(\mu_{ij})]$ and $R(\alpha)$ is $n_i \times n_i$ working correlation matrix that is specified by the vector parameters and estimated as

$$r_{ij} = \frac{y_{ij} - \mu_{ij}}{\sqrt{V(\mu_{ij})}} \quad ,$$

using the current value of the parameter vector β to compute appropriate functions of the Pearson residual (Stokes *et al*., 2000; Davis, 2002).

For each $y_i = (y_{i1},...,y_{it_i})$, $R_i(\alpha)$ working correlation matrix for repeated measurements for each individual is calculated. When $t_i = 1$ then GEE is equal to GLM. The following step require for obtain the estimates with GEE (Stokes *et al*., 2000).

Compute an initial estimate of β by GLMs with specify an appropriate link function, $g(E(y_{ij})) = X'_{ij}\beta$ where g is the link function (logit, probit, identity etc.,). $E(y_{ij}) = \mu_{ij}$ which is the marginal proportion.

- Specify the variance of y_{ij}, $Var(y_{ij}) = V(\mu_{ij})\phi$, for binomial data, for example, Var(y)-pq, so $V(\mu_{ij}) = p(1-p) = \mu_{ij}(1-\mu_{ij})$. ϕ is the scale parameter and $\phi = 1$ for logit data.
- Choose a working correlation, $R_{ij}(\alpha)$.
- Compute an initial estimate of β. This can be estimated with Ordinal Least Square (OLS) assuming independence.

- Compute the working correlation $R_i(\alpha)$ based on the standardized residuals. Standardized residuals for normal case, for example, $r_i = (y_i - \mu_i)/\sigma_i$. There are structure of working correlation i.e., fixed, m-dependent, exchangeable, unstructured and Autoregressive AR(1).
- Compute an estimate of the covariance matrix, V_i
- Estimate regression parameters β, $cov(\beta_t, \beta_j)$ and update. Iterative quasi-likelihood methods for estimating β,

$$\beta_t = \beta_{t-1} - \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta}\right]^{-1} \left[\sum_{i=1}^{K} \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} (Y_i - \mu_i)\right]$$

where the working covariance matrix for y_i is V_i and diagonal elements of A are $V(\mu_{ij})$

- Update β_{t+1} and iterate until convergence.

The GEE parameter estimates are consistent as the number of clusters becomes large, even if the working correlation matrix have mis-specified, as long as the mean model is correct. In this study exchangeable correlation matrix was used. The exchangeable correlation matrix presumes the correlation of the measurements between any two observations when time is constant. The exchangeable correlation matrix is described as,

$$Corr\left(y_{ij}, y_{i,j'} = \begin{cases} 1 & j = j' \\ \alpha & j \neq j' \end{cases}\right)$$

(Stokes *et al*., 2000; Davis, 2001).

RESULTS

The descriptive statistics about the variables used in the model are presented in Table 1

The criteria for assessing goodness of fit are displayed in Table 2, contains statistics that summarize the fit of the specified model. These statistics are used in judging the adequacy of a model and in comparing it with other models under consideration. The value/DF of deviance for PR and NBR (4.5006 vs 1.0446) indicate that NBR fits the data reasonably well for the specified model. When the value/DF of deviance approach to 1, there is no over dispersion.

Table 3 displays the analysis of initial parameter estimates. Entries in the chi-square column are likelihood ratio statistics for testing the significance of the effect added to the model containing all the preceding effects. On the Table 3, analysis of initial parameter estimates about the variables modeled is given. It was seen that over flehmen lip curl, the weight of the ram, the tail raising, vocalization and the mounting duration on the number mounting are important ($p<0.01$) but the behaviour of anogenital sniffing is unimportant ($p>0.01$).

Forteen separate inspections were applied to 32 animals of the data set. Because of that for repeated measurement values GEE estimations were obtained. On the Table 4 GEE parameter values are given. When the values of the parameters are evaluated it was seen that over flehmen lip curl, the weight of the ram, the tail raising, vocalization and the mounting duration on the number mounting are important ($p<0.01$) but the behaviour of anogenital sniffing is unimportant ($p>0.01$). It was determined that flehmen lip curl, the weight of the ram and the mounting duration effected the number of mounting negatively. However it was observed that vocalization and the tail raising have a positive effect on the number of mounting.

Table 1: Basic statistics for the data used

Variables	Mean	Variance	Min.	Max.
Mount frequent	10.7916	117.5128	0.00	68.00
Flehmen lip curl	1.1499	2.2076	0.00	11.00
Mount duration	109.5399	11810.13	10.00	515.00
Anogenital sniffing	4.4132	11.8177	0.00	20.00
Tail raising	0.7606	2.0696	0.00	13.00
Vocalization	8.0771	71.7786	0.00	49.00
Weight of ram lamb	49.413	114.2452	26.10	75.70

Table 2: Goodness of fit criteria for Poison regression and negative binomial regression

	Poison regression			Negative Binomial regression		
Criterion	df	Value	Value/df	df	Value	Value/df
Deviance	313	1408.683	4.5006	313	333.229	1.0446
Pearson Chi-square	313	1527.285	4.8795	313	251.067	1.1216

Bold numbers is the value/df of deviance of the selected model.

Table 3: Initial parameter estimates

Parameter	df	Estimate (standard error)	Wald %95 C.I. limits	Chi-square
Intercept	1	3.2872(0.1844)	(2.9260, 3.6485)	317.87**
Mount duration	1	-0.0798(0.0273)	(-0.0263, -0.1333)	8.55**
Flehmen lip curl	1	-0.0180(0.0004)	(-0.0251, -0.0110)	25.16**
Anogenital sniffing	1	-0.0170(0.0121)	(-0.0406, 0.0066)	1.99
Tail raising	1	0.1736(0.0250)	(0.1245, 0.2227)	48.05**
Vocalization	1	0.0273(0.0043)	(0.0189, 0.0357)	40.21**
Weight of ram lamb	1	-0.0019(0.0004)	(-0.0026, -0.0011)	25.54**

**p<0.01

Table 4: GEE parameter estimates

Parameter	df	Estimate (standard error)	Wald %95 C.I. limits	Z
Intercept	1	2.8973(0.1067)	(2.6881, 3.1065)	27.14**
Mount duration	1	-0.0665 (0.0240)	(-0.1135, -0.0195)	-2.77**
Flehmen lip curl	1	-0.0095 (0.0029)	(-0.0152, -0.0038)	-3.28**
Anogenital sniffing	1	-0.0194(0.0152)	(-0.0493, 0.0105)	-1.27
Tail raising	1	0.1642(0.0223)	(0.1205, 0.2078)	7.38**
Vocalization	1	0.0248(0.0061)	(0.0128, 0.0368)	4.06**
Weight of lam lamb	1	-0.0021(0.0004)	(-0.0028, -0.0014)	-5.81**

**p<0.01

It is clear that the parameter estimates observed on the Table 4 are in accordance with initial parameter estimates. Besides, exchangeable correlation matrix, used for working correlation, are obtained as:

$$\text{Corr}\left(y_{ij}, y_{i,j'} = \begin{cases} 1 & j = j' \\ 0.112 & j \neq j' \end{cases}\right)$$

DISCUSSION

The results about Flehmen lip curl and mount duration and their effects on the number of the mount are expected behaviors for sexual behaviors. In fact, when rams do not display the mating behaviors they show the behaviors of courting (Katz *et al*., 1988). This state both increase the mount frequent and negatively affects the number of the mating behaviors performed in the unit time. But vocalization and tail raising behaviors affected positively the number of mount.

Negative Binomial Regression is widely used for investigating the overdispersion happening in the Poisson regression (Wang *et al*., 1998; Dalrymple *et al*., 2003; Yeşilova, 2003). Analysis without considering the overdispersion causes incorrect parameter estimations. In this study, it is revealed that NBR is very effective for explaining the overdispersion. Besides NBR is a quite effective method for ceasing the later heterogeneity of the population (Dalrymple *et al*., 2003; Yeşilova, 2003; Wang *et al*., 1996, 1998). Using GEEs, parameter estimations were obtained. Especially in the studies related with cattle species, repeated measurements are widely used. Some reproduction characteristics of the domesticated animals are categorically obtained. In such data normal distribution assumption are not obtained. (Tempelman and Gianola, 1993; 1996; Tempelman, 1998). GEE is widely used in the analysis of repeated data that the interested variable is categorical. GEE uses a working estimation about the forms of correlations. When the correlations are of medium degree, GEE gives similar results for all working correlations (Davis, 2002; Stokes *et al*., 2000). In this study the exchangeable was used as the working correlation. Also analyses were made for the other working correlations. Especially in the form of unstructured correlation for the mounting frequency, vocalization and the tail raising it was determined that the algorithms of GEE were not converged.

REFERENCES

Agresti, A., 1997. Categorical Data Analysis. John and Wiley and Sons, Inc.

Breslow, N., 1990. Tests of hypotheses in overdispersed poisson regression and other quasi-likelihood models. J. Am. Stat. Assoc., 85: 565-571.

Cameron, A.C. and P.K. Trivedi, 1998. Regression Analysis of Count Data. Cambridge University Press.

Cox, R., 1983. Some remarks on overdispersion. Biometrica, 70: 269-274.

Dalrymple, M.L., I.L. Hudson and R.P.K. Ford, 2003. Finite mixture, zero-inflated poisson and hurdle models with application to SIDS. Comput. Stat. Data Anal., 41: 491-504.

Davis, C.S., 2002. Statistical Methods for the Analysis of Repeteated Measurements. Heidelberg: Springer Verlag.

Dobson, J.A., 1990. An Introduction to Generalized Linear Models. Chapman and Hall, New York, pp: 174.

Frome, E.L., M.H. Kutner and J.J. Beachamp, 1973. Regression analysis of poisson-distributed data. J. Am. Stat. Associ., 68: 935-940.

Frome, E.L., 1983. The analysis of rates using poisson regression models. Biometrics, 39: 665-674.

Gardner, W., E.P. Mulvey and E.C. Shaw, 1995. Regression analyses of counts and rates: Poisson overdispersed poisson and negative binomial models. Psychol. Bull., 118: 392-404.

Katz, L.S., E.O. Price, S.J.R. Wallach and J.J. Zenchak, 1988. Sexual performance of rams reared with or without females after weaning. J. Anim. Sci., 66: 1166-1173.

Lawles, J., 1987. Negative binomial and mixed poisson regression. Can. J. Stat., 15: 209-225.

Liang, K.Y. and S.L. Zeger, 1986. Logitudional data analysis using generalized linear models. Biometrica, 73: 13-22.

Littell, C.R., A.G. Milliken, W.W. Stroup and D.R. Wolfinger, 1996. SAS system for mixed models, SAS Institute Inc., Cary, NC.

Mccullagh, P. and J.A. Nelder, 1989. Generalized Linear Models. 2nd Edn., Chapmann and Hall, London.

Okut, H., S.A. Gökdere and A. Yeşilova, 1999. Application Generalized Linear Mixed Models. III. National Conference of the Italian Biometric Society, Roma, 1999.

Price, E.O., 1993. Practical considerations in the measurement of sexual behavior. Meth. Neurosci., 14: 16-31.

SAS: SAS/STAT Software, 2005. Hangen and Enhanced. SAS, Inst. Inc., USA.

Stokes, M.E., C.S. Davis and G.G. Koch, 2000. Categorical Data Analysis Using the SAS System. John and Wiley and Sons, Inc, pp: 626.

Tempelman, R.J. and D. Gianola, 1993. Genetic analysis of fertility in dairy cattle using negative binomial mixed models. J. Dairy Sci., 72: 1557-1568

Tempelman, R.J. and D. Gianola, 1996. A mixed effects model for overdispersed count data in animal breeding. Biometrics, 52: 265-279.

Tempelman, R.J., 1998. Generalized linear mixed models in dairy cattle breeding. J. Dairy Sci., 81: 1428-1444.

Wang, P., M.L. Putterman, I.M. Cockburn and N. Le, 1996. Mixed poisson regression models with covariate dependent rates. Biometrics, 52: 381-400.

Wang, P., I.M. Cockburn and M.L. Putterman, 1998. Analysis of patent data- mixed poisson regression model approach. J. Busi. Econ. Stat., 16: 27-41.

Yeşilova, A., 2003. Using of poisson mixture regression models for categorical data in biology. Unpublished Ph.D. Thesis, University of Yüzüncü Yil, Van, Turkey.