Applying a zero-inflated negative binomial model to hierarchical count data.

Nødtvedt A[1]*, Sanchez J[2], Dohoo I[2]. 1 Department of Small Animal Clinical Sciences, Swedish University of Agricultural Science, Uppsala, Sweden. 2 Department of Health Management, Atlantic Veterinary College, University of Prince Edward Island, Charlottetown, PEI, Canada.

Summary

In this paper, fecal egg counts of bovine gastrointestinal nematodes collected from adult cows in 38 different Canadian dairies were analysed. Samples were collected from the same individuals throughout a year, and hence were not independent of each other. The four levels of clustering in the dataset were: region, herd, cow, and test day. A large proportion of samples contained no nematode eggs, and additional extra-Poisson variation existed within the non-zero counts. A zero-inflated negative binomial model and a four level random intercepts negative binomial model were applied to the data and different modelling strategies for over-dispersed and clustered count data are discussed.

Introduction

Poisson models have been regarded the standard way of analysing count data. However, one of the assumptions is that the mean and the variance of the residuals are approximately the same, which is often not the case. Unobserved heterogeneity in the data can yield a variance that is greater than the mean, also known as over-dispersion. Negative binomial models allow for extra-Poisson variation by taking this over-dispersion into account. In special cases of count data, zero counts can be generated by a separate process that prevents some objects from experiencing the event being counted. Zero-inflated negative binomial models allow for additional over-dispersion via a splitting process in which the probability of a zero outcome is modelled by logistic regression and the continuous outcome is modelled using a negative binomial error structure. Alternatively, over-dispersion of negative binomial models can be dealt with by including additional random effects in the model. In this paper, fecal eggs counts of gastrointestinal nematodes from adult dairy cows are used as an example of over-dispersed count data. An important feature of fecal egg count data from adult animals is that there are a high number of zero observations, as no eggs are detected in a high number of individuals. The counts were not independent as repeated samples were collected from the same individuals over time, cows were clustered in herds, and herds were clustered in distinct geographical regions. The lack of independence between clustered observations can be corrected for by using the Huber/White/ sandwich estimator of variance, also known as the robust variance estimator. The dependence of observations within a cluster can also be accounted for by including a cluster-specific random intercept in the regression model. In the case of hierarchical data where the observations are clustered on several levels (region, herd, animal, test day) the intercept can be allowed to have random variation at each of the levels in the hierarchy. The objective of this paper was to explore two different methods for the analysis of over-dispersed, hierarchical count data.

Materials and methods

The dataset consisted of fecal gastrointestinal nematode egg counts (fec) collected from eight adult dairy cows from 38 herds in four Canadian regions throughout a year. All animals had been exposed to pasture to some degree. Animal- and test day variables were recorded by the researchers on site and data on different management factors were obtained using a standardized questionnaire. Two separate regression models were built in order to assess the relationship between fec and the explanatory variables. The first model was a zero-inflated negative binomial (zinb) model, in which herd and region were included as a fixed effects and the Huber/White/ sandwich estimator of variance was applied to correct for the clustering of observations within cow over time. The Vuong statistic was calculated to assess the fit of a zero inflated versus a regular negative binomial model (1). Explanatory variables were tested in both the negative binomial and the logistic part of the model. The statistical software package Stata 7 was used for the zinb model (2). The second model was a negative binomial (nb) model where the intercept was allowed to have random variation at the levels region, herd, animal, and test day. An exchangeable correlation structure was assumed. The same independent variables as in the zinb model were included in the nb model. The second model was built using the software package MLwiN (3).

Results

The dataset consisted of 1840 individual fecal egg counts from 315 animals. The fecal egg counts ranged from 0 to 419 per 5 grams of feces, were heavily skewed to the right and contained a high proportion (46.2%) zero counts. The overall variance was 998.7 and the mean was 9.8, hence a Poisson model was not considered appropriate. The Vuong statistic had a high positive value (8.43; $P<0.001$), indicating that a zero inflated model fit the data better than a regular negative binomial model. Table 1 shows the coefficients and standard errors for the zinb model, separated into the negative binomial (*Log (fec*)) and logistic part (*Logit (probability of zero count*)). The coefficients and standard errors from the random effects model are shown in the same table. Herd was included as a fixed effect in the logistic part of the zinb model but the herd specific coefficients have been excluded from the table. Region was a significant determinant of fec in the negative binomial part of the zinb model but was omitted from the table for ease of comparison with the nb model. Coefficients in the negative binomial part of the zinb model are interpreted as for count models; a negative coefficient means lower expected fec. In the logistic part of the zinb model, a negative coefficient means a lower probability of a zero count. The random effects negative binomial model allows a breakdown of the variance to the four levels of clustering in the data: region (1.6%), herd (3.7%), cow (22.8%), and test day (71.9%).

Table 1. Gastrointestinal nematode eggs per 5 g of feces (fec): Variables, coefficients and robust standard errors from a zero inflated negative binomial model, and coefficients, standard errors and variance components from a random effects negative binomial model.

		Zinb model		*Random effects nb model*	
Variable	Level	Coefficient	Robust S.E.	Coefficient	S.E.
Log(fec)					
Lactation	1st	Baseline		Baseline	
	2nd +	-0.94	0.23	-0.70	0.21
Season	Fall	Baseline		Baseline	
	Winter	-0.70	0.18	-0.82	0.17
	Spring	0.36	0.22	0.28	0.20
	Summer	0.08	0.26	0.15	0.22
Pasture, lactating animals		0.93	0.33	1.49	0.34
Logit (probab. of 0 count)					
Lactation	1st	Baseline		*n.a.*	
	2nd +	1.50	0.53	*n.a.*	

Discussion

As can be seen from table 1, the results from the two models were comparable as to the effect of the independent variables on fecal egg counts. The analysis of crude variance components from the random effects negative binomial model shows that once factors such as age and exposure to pasture were taken into account, very little variation was observed between herds and between regions. The application of the robust variance estimator at the cow level, combined with the negative binomial error distribution, should prevent the clustering of observations within each cow from having a substantial effect on the standard errors in the zinb model. In the zinb model, the clustering of cows within herds was accounted for by including a variable for each of the herds in the logistic part of the model. By including herd as a fixed effect 38 herd-specific coefficients are calculated. However, these herds represent a sample of herds from which general inferences about the effect of the independent variables on fec are to be drawn. Hence, the herd-specific coefficients can be seen as nuisance parameters that are included only to control for unobserved heterogeneity at the herd level and are of little interest in them selves. Including a herd specific random intercept is another way of dealing with the within-herd correlation, but it can be argued this approach is not suitable as the included herds represented a convenience sample rather than a random sample. The zero inflated negative binomial model is theoretically more suitable for dealing with the high proportion of zero counts observed in this dataset. Overall, the zero inflated negative binomial and random effects negative binomial models yielded similar results that were in accordance with previous work on gastrointestinal nematodes in adult dairy cows.

References

1. Long JS. Count outcomes: Regression Models for Counts. Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks: Sage Publications, 1997:217-250.

2. Stata Statistical Software. College Station, Texas: Stata Corporation, 2001.

3. Goldstein H, Rasbash J, PLewis I, et al. A user's guide to MLwiN. 1.1 ed. Institute of Education, University of London, 1998.