Interpretation of variance parameters in multilevel Poisson regression models

Stryhn, H.[1], Sanchez, J.[1] , Morley, P.[2] , Booker, C.[2] and Dohoo, I.R.[1]
[1] Atlantic Veterinary College, University of Prince Edward Island, Canada
[2] Feedlot Health Management Services, Alberta, Canada

Introduction

Multilevel models have become a standard statistical tool for veterinary epidemiology. Goldstein et al (2002) review interpretation of variance parameters in 2-level logistic regression models in terms of intra-class correlations (ICCs) and variance partition coefficients (VPCs). Briefly, ICCs are correlations between two observations that share the same unit(s) of the hierarchical structure, e.g. animals in the same herd in a two-level structure. VPCs are the proportions of variation at different levels of the hierarchy. In discrete models such as logistic and Poisson regression, the ICC and VPC are not constant across the data but depend on the fixed part of the model. For a 2-level model, the VPC coincides with the ICC between two observations with the same predictors; here, we restrict attention to such ICCs. Goldstein et al (2002) describe four procedures to compute ICCs (VPCs), and note their methods to be applicable to models with multiple hierarchical levels and other discrete distributions than binomial. Vigre et al (2004) and Browne et al (2005) describe extensions to 3-level logistic regression models. The objective of this paper is to describe the application of these methods to Poisson regression models with 2 or more hierarchical levels, and to provide formulae for computation of ICCs in such models.

Materials and Methods

Consider a 2-level Poisson regression model (say, for counts of events in groups of animals within herds, as in Example 1 below) with log (natural logarithm) link, fixed part $X\beta$ (short for $\beta_0 + \beta_1 x_1 + ... + \beta_k x_k$) and herd random effects (intercepts) $u \sim \mathrm{N}(0,\sigma^2)$. The equation for the linear predictor takes the form: $\log(\lambda) = X\beta + u$, and given the random effects, the observed counts are independent and Poisson-distributed with mean λ. The fixed part may include a term with a regression coefficient fixed at a value of 1 (often termed an offset), corresponding to the (logarithmic) population at risk.

Model linearisation method

By the linearisation method, the non-linear inverse link function (here, the exponential function) is approximated by a first order Taylor expansion which yields a single equation for the observed count Y containing both the random effect and a standardised error term e:

$$Y \approx \text{(fixed terms)} + u \exp(X\beta) + \lambda^{½} e,$$

where λ is evaluated at the mean of the random effects, i.e. $\lambda = \exp(X\beta)$. From this equation it follows that $\mathrm{var}(Y) \approx \sigma^2\exp(2X\beta) + \lambda \equiv \sigma^2(2) + \sigma^2(1)$, say, where $\sigma^2(1) = \lambda$ and $\sigma^2(2) = \sigma^2\exp(2X\beta)$ may be interpreted as variance components at the lower (animal group) and upper (herd) levels, respectively. Then, ICC $\approx \sigma^2(2) / [\sigma^2(2) + \sigma^2(1)]$.

Simulation method

The Appendix of Vigre et al (2002) gives a detailed account of the simulation method for 2-level and 3-level logistic regressions. We describe here only the necessary modifications for a 2-level Poisson regression model. Simulate a large number (say N=100,000) of standard normal variables $z_1,...,z_N$, replace both $p_i(x)$ and $\tau_i(x)$ by $\lambda_i(x)$ calculated as $\lambda_\mathrm{i}(x) = \exp(X\beta + \sigma z_i)$, and compute the ICC by the same formula: $\mathrm{ICC}(x) = \mathrm{var}\ \{\lambda_\mathrm{i}(x)\}\ /\ (\mathrm{var}\ \{\lambda_\mathrm{i}(x)\} + \mathrm{mean}\ \{\lambda_\mathrm{i}(x)\})$. As above, the equation may be expressed as: $\mathrm{ICC}(x) = \sigma^2(2) / [\sigma^2(2) + \sigma^2(1)]$, with similar interpretations of $\sigma^2(1)$ and $\sigma^2(2)$.

Normal distribution model

In order to make a linear mixed (normal distribution) model at least somewhat comparable to a Poisson regression with an offset, one should normalise (divide) the observed counts with the population at risk. If the latter measures risk time, the model is effectively for the incidence rate.

Exact calculation

Using integration formulae for exponential functions it can be shown (Stryhn, 2006) that the ICC for two observations with common $X\beta$ within the same hierarchical unit (herd), is given by

$$\text{ICC} = [\exp(2X\beta + 2\sigma^2) - \exp(2X\beta + \sigma^2)] / [\exp(2X\beta + 2\sigma^2) - \exp(2X\beta + \sigma^2) + \exp(X\beta + \sigma^2/2)],$$

which is again of the form $\sigma^2(2) / [\sigma^2(2) + \sigma^2(1)]$, with $\text{var}(Y) = \sigma^2(1) + \sigma^2(2)$.

Example 1: Tuberculosis cases

The textbook Dohoo et al (2003) gives data on outbreaks of tuberculosis in domestic animals (tb_real dataset). In each of the 30 farms where tuberculosis was observed, the animals were grouped by age, sex and animal type, and the number of animal days at risk within each group was recorded together with the observed number of (new) tuberculosis infections. The farms comprise 1-13 (median=4) animal groups for a total of 134 groups. A random effects Poisson regression model was fitted to the data with fixed effects of age group (0-12, 12-24, >24 months), sex and animal type and normally distributed farm random effects. Among the fixed effects, only age group showed a substantial effect, and we consider here a simplified model with a dichotomous age grouping only (young (0-12 months) and older (>12 months) animals). The resulting equation for the linear predictor on log scale, including also the log(animal days at risk) as an offset, was:

$$\text{linear predictor} = (-11.49)*\text{young} + (-8.87)*\text{older} + 1.267*u_{\text{farm}} + \log(\text{animal days at risk}),$$

with $u_{\text{farm}} \sim N(0,1)$; that is, $\sigma^2 = 1.267^2 = 1.605$. To illustrate the implications of this model, we consider six scenarios (1-6) with age group being either young (scenarios 1-3) or older (4-6), and the animal days at risk equal to 700 (scenarios 1,4), 2100 (2,5) or 8000 (3,6). The three values for animal days at risk are close to the 25%, 50% and 75% percentiles in the data, respectively.

Example 2: Infectious disease mortality in feedlots

Data covering a six-year period in beef cattle feedlots in Western Canada were collected with the purpose of determining factors affecting mortality (caused by bovine respiratory disease complex, *Histophilus somni* or infectious arthritis) in groups of animals purchased and raised together (Animal Health Database, Feedlot Health Management Services, Okotoks, Alberta). The data considered here comprises 44 feedlots, 6678 lots and a total of 42722 animal groups. As a descriptive tool to assess the clustering in the data, a ”null” Poisson model with no fixed effects except the offset determined from the population at risk was fitted; 1st order MQL estimation in the MLwiN software package produced the equation:

$$\text{linear predictor} = (-4.568) + 1.092*u_{\text{lot}} + 0.3975*v_{\text{feedlot}} + \log(\text{population at risk}).$$

A 3-level Poisson regression model permits similar exact calculations of ICCs as described above (see Stryhn (2006) for details), except that there are two ICCs of interest: between two animal groups within the same lot (and feedlot), and within the same feedlot only (i.e., in different lots).

Results

Example 1: Tuberculosis cases

The variance components and resulting ICC, computed by three different methods for each of the six scenarios, are shown in Table 1 (next page). A two-level normal distribution model had a constant ICC=0.185; the variance components for the number of cases per 1000 animal days at risk were 0.529 and 0.120 at the lower (animal group) and upper (herd) levels, respectively.

The results show strong variation of the ICCs across the six scenarios; higher values of the linear predictor imply stronger correlations. Therefore the constant ICC provided by the normal distribution approximation is quite useless. Also the model linearisation method is seen to perform poorly. Although the approximation may be improved by higher order Taylor expansion, for Poisson regression models the effort does not seem worthwhile. The simulation method seems to give an acceptable accuracy with N=100,000 simulations, but does not offer any advantages over an exact calculation.

Table 1 Lower (animal group) and upper (herd) level variance components (σ^2(1) and σ^2(2), respectively) and corresponding intra-class correlation coefficient (ICC) of a random effects Poisson regression model for tuberculosis data, computed in six different scenarios and by three different procedures.

Scen-ario	Linear	Model linearisation			Simulation-based			Exact formulae		
	pred.	σ^2(1)	σ^2(2)	ICC	σ^2(1)	σ^2(2)	ICC	σ^2(1)	σ^2(2)	ICC
1	-4.94	0.007	0.000	0.011	0.016	0.001	0.055	0.016	0.001	0.060
2	-3.84	0.021	0.001	0.033	0.048	0.008	0.147	0.048	0.009	0.160
3	-2.50	0.082	0.011	0.116	0.181	0.137	0.430	0.183	0.133	0.421
4	-2.32	0.098	0.016	0.136	0.221	0.228	0.508	0.220	0.192	0.466
5	-1.22	0.295	0.140	0.321	0.658	1.618	0.711	0.659	1.726	0.724
6	0.12	1.124	2.029	0.643	2.527	26.01	0.911	2.509	25.05	0.909

Example 2: Infectious disease mortality in feedlots

Exact ICCs for two animal groups of the same size were computed for the range of animal group sizes represented in the data (Figure 1). The correlation within feedlot only is seen to be low irrespective of the group size, whereas the correlation within lot (and feedlot) ranges from moderate to very high.

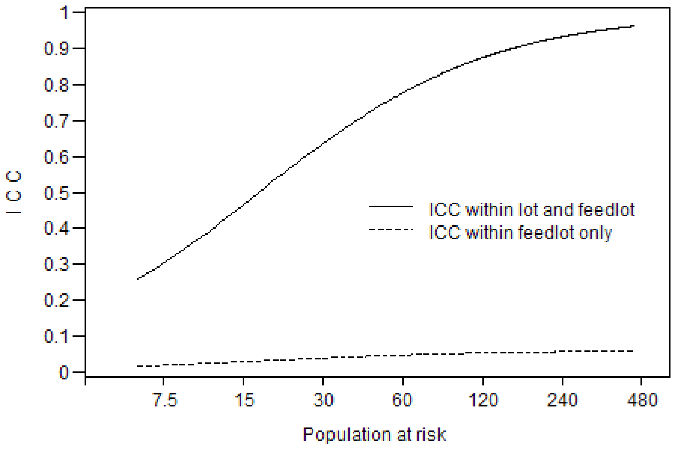

Figure 1 Intra-class correlation coefficients (ICCs) in a 3-level Poisson regression model for a range of values of the population at risk (6-450 animals).

References

Browne, W.J., Subramanian, S.V., Jones, K. and Goldstein, H. (2005). Variance partitioning in multilevel models that exhibit overdispersion. *Journal of the Royal Statistical Society A,* 168, 599-614.

Dohoo, I.R., Martin, S.W. and Stryhn, H. (2003). *Veterinary Epidemiologic Research*. AVC-Inc., Charlottetown, Canada.

Goldstein, H., Browne, W.J. and Rasbach, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, 1, 223-232.

Stryhn, H. (2006). A note on variance partition and intra-class correlation coefficients in multilevel Poisson and negative binomial regression models. Submitted manuscript.

Vigre, H., Dohoo, I.R., Stryhn, H. and Busch, M.E. (2004). Intra-unit correlations in seroconversion to *Actinobacillus pleuropneumoniae* and *Mycoplasma hyopneumoniae* at different levels in Danish multi-site pig production facilities. *Preventive Veterinary Medicine*, 63, 9-28.