Simulation Studies on the Effects of Clustering

Dohoo, I.R. and Stryhn, H.
Department of Health Management, University of PEI

Abstract

A series of simulation studies were carried out to investigate the effects of clustering (hierarchical data) on analyses of both simple and confounded associations between a predictor (X) and an outcome (Y). Data sets consisting of cows within 100 herds were simulated 1000 times and analysed. For the simple association model the bias in the standard error of the estimate of the coefficient for X (•) depended on the intra-class correlation (ICC) of X. Although asymptotically unbiased, individual estimates of • were much more variable than expected and this additional variance was also related to the ICC of X. For the confounding model, the ability of a random effects model to control bias resulting from unmeasured confounders (Z) depended on the ICC of Z.

Introduction

It has long been recognized that lack of independence among observations results in incorrect results from statistical analyses, if the lack of independence is not correctly dealt with (McDermott and Schukken, 1994). The most serious concern has been with the underestimation of the standard errors (SE) of parameters. Under certain assumptions, the SE of the estimate of the effect of a group-level predictor is known to be underestimated by a factor equal to the square root of the variance inflation factor (VIF – also known as the design effect) (Dohoo et al, 2003). Less attention has been paid to the situation in which the predictor is an animal-level (eg cow-level) predictor which might be independent within herds or may itself be clustered within herds.

In many analyses evaluating the effects of various factors on either risk of disease or some measure of productivity, "herds" are often included in analyses as either fixed or random effects. This is done partially to deal with the statistical issue of clustering, but also an attempt to control for unmeasured confounders that may be related to the management of the herd. It is not clear how well this approach works in removing this potential confounding effect.

To address the issues raised above, a series of simulation studies were carried out using hypothetical two-level (cows within herds) data with the specific questions to be answered as follows.

1. How does clustering bias estimates of • and its SE when the predictor is a herd-level or animal-level predictor?
2. How well does inclusion of "herd" as a random effect control for the confounding effect of unmeasured herd- or animal-level confounders which may be related to herd management?

This manuscript deals with the issue of bias in • when both the predictor and outcome are continuous (linear regression model). A more complete manuscript will cover dichotomous predictors, logistic regression models and impacts on the intercept.

Methods

Simulations were carried out to determine the effects of clustering on associations between a confounder (Z), a predictor (X or XZ) and an outcome (Y) in relationships such as those shown in Figure 1 (**simple model** in left panel, **confounding model** in the right panel). A dataset of 100 virtual herds was created with 50 "small" herds (number of cows distributed N(50,10), and

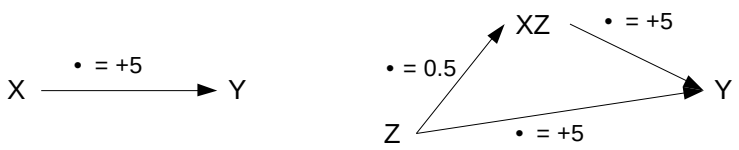

Figure 1 Causal diagram of the "simple association model" (left) and "confounded association model" (right)

50 “large” herds (number of cows distributed N(250,50). All results are presented in the framework of a two level hierarchy consisting of cows within herds.

The outcome (Y = milk production in kg/day) had a between herd-variance of 25 and within herd variance of 49. A one unit increase in either X and XZ increased milk production by 5 kg/day. The predictor X was Normally distributed (X~N(0,1)) and was generated to have one of 17 levels of within-herd clustering, ranging from 0 (complete independence) to 1 (a herd-level predictor). The confounder Z was similarly constructed. XZ was then computed to also be N(0,1) and to have a correlation with Z of 0.5.

A total of 1000 datasets were generated of each of the 17 levels of ICC. For the **simple model** (X-->Y) the following analyses were carried out. Bias in estimates of • and SE(•) for X were determined by dividing the means of their respective estimates derived from an ordinary linear regression by the means of the estimates from a random effects model (restricted maximum likelihood estimate). The effect of clustering on the variability of the individual estimates was evaluated by determining the standard deviation (SD) of the individual estimates of • from both the ordinary and random effects models across the 1000 datasets.

For the **confounding model** (involving Z), bias in estimates of • and SE(•) for XZ were computed by dividing their respective means of estimates from a random effects model which did not include Z by those from a model which did include Z. All simulations and analyses were carried out using Stata (Version 9).

Results and Discussion

Figure 2 (left panel) presents the mean bias observed for • and SE(•) of X for various levels of ICC of X in the **simple model**. There is no systematic bias in the estimate of • and this is in agreement with Liang and Zeger (1993). When ICC=1 (X is a herd-level predictor), the SE(•) is grossly underestimated (bias = 0.138). This corresponds to what would be predicted based on standard formulae for the VIF associated with a herd-level predictor and reflects the serious problem of potentially reporting spurious significance for herd-level predictors if clustering is not taken into account. When ICC=0 (X is a cow-level predictor that doesn't cluster at all within herd), the SE(•) is overestimated because the variance used to compute the SE is the total variance of Y in the ordinary model, but the within-herd variance in the random effects model. This has important implications for sample size computations. Sample size calculations for studies in which cow-level variables have no within-herd clustering (eg. treatment assigned to 50% of animals in a randomized controlled trial) will be conservative estimates (ie overestimate the required sample size) if the hierarchical nature is not taken into consideration. This is the opposite of what would happen for a herd-level factor. For the particular scenario investigated in this study, the bias in the SE(•) switches from overestimation to

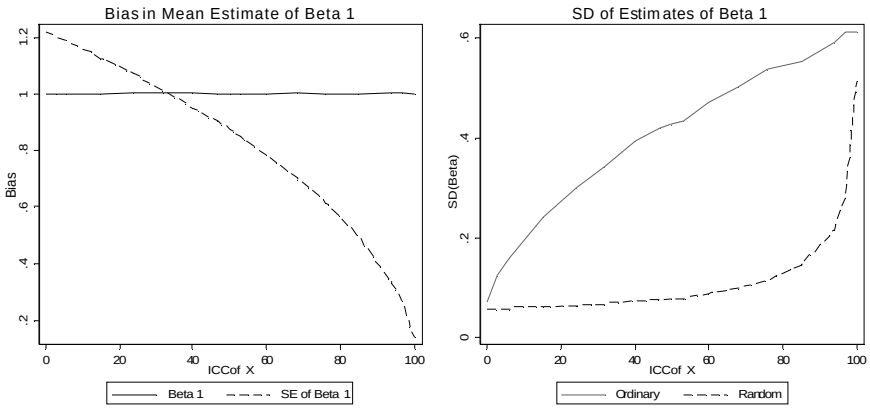

Figure 2 Results from analyses of simple model. Bias (see text for definition) in estimates of • and SE(•) in left panel. Standard deviation of individual estimates of • from ordinary and random effects analyses in right panel.

underestimation at about ICC = 0.35, although this transition point varies with the degree of clustering in Y (data not shown)

Although there is no systematic bias in the estimate of •, individual estimates derived from an ordinary linear regression are much more variable than those from the random effects model (Figure 2 – right panel). This is particularly true at moderate values for ICC. When ICC~=0.5, the SD of the individual estimates of • was approximately 5 times that of what it should be, based on the random effects model. As a consequence, individual estimates of • derived from an ordinary linear model may be badly biased, but it would be impossible to predict the direction of the bias. At extreme values of ICC (0 or 1), the variability of the ordinary regression estimates are much closer to what would be expected.

Figure 3 shows the mean bias observed for • and SE(•) of XZ for various levels of ICC of Z in the **confounding model**. If the confounder is a true herd-level factor (ie ICC=1) inclusion of herd as a random effect in the model completely eliminates any confounding effect of Z. On the other had, if there is some variation in Z among cows within a herd (ie. ICC<1), then there is residual confounding due to Z and not including Z in the model will produce a biased estimate – in this case an estimate of • that is up to 1.5 times larger than it should be. The actual magnitude and direction of the confounding bias will depend on the strength and direction of the relationships between Z and Y and between Z and XZ. While the estimate of the • is biased if ICC<1, the estimate of the SE(•) is virtually unbiased regardless of the ICC of Z, because both models had a correct specification of the variance structure.

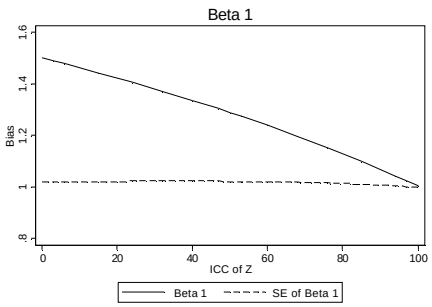

Figure 3 Bias in • and SE(•) from the confounding model.

Conclusions

There are a number of important conclusions which can be drawn from the above simulations. Based on the **simple model** , failure to adequately account for the hierarchical nature of a dataset in the analysis of an association between X and Y will have the following effects.

1. There will be no systematic bias in the estimate of •, but individual estimates of • may be badly biased
2. The bias in the estimation of the SE (•) will depend on the nature of the predictor (X).
   (a) If X is a herd level variable, the SE (•) will be substantially underestimated.
   (b) If X is a cow-level variable with no within-herd clustering, the SE (•) is overestimated.
   (c) If X is a cow-level variable with some within-herd clustering, the degree of over- or under-estimation will depend on the ICC of X and Y.

The analyses based on the **confounding model** showed that inclusion of herd as a random effect in a model has the following effects in terms of controlling confounding.

1. If the confounder is a herd-level factor, the estimate of • will be unbiased.
3. If the confounder is a cow-level confounder (even if it is clustered within herds), inclusion of herd random effects will not remove all of the confounding effects on •.

References

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

Liang, K. and Zeger, S. (1993). Regression Analysis for Correlated Data. *Annual Review of Public Health* 14, 43-68.

McDermott, J.J. and Schukken, Y.H. (1994). A review of methods used to adjust for cluster effects in explanatory epidemiological studies of animal populations. *Preventive Veterinary Medicine* 18, 155-176.