Dealing with Data Deficiencies:

A Comparison of Methods for Modelling Sparse Data

A.E. Mather[1], J.D. Holt[2], D.J. Mellor[3], S.A. McEwen[1], R. Reid-Smith[1,4], S.W.J. Reid[3]

[1]Department of Population Medicine, Ontario Veterinary College, University of Guelph, Canada [2]Department of Mathematics and Statistics, University of Guelph, Canada [3]Faculty of Veterinary Medicine, University of Glasgow, UK [4]Laboratory for Foodborne Zoonoses, Public Health Agency of Canada, Guelph, Canada

Introduction

Data deficiencies, specifically data sparsity and zero cells, present a recurring problem that is found in all areas of research, including veterinary epidemiology. This takes the form of too few counts in the cells of multidimensional contingency tables, and can result from either a small sample size, or a large sample size but with many variables. A special problem arises when, by chance, zero cells occur, and consequently some of the parameter estimates are infinite. This may be recognised by very large standard errors in the output of the logistic regression model, as there is a lack of convergence in the iterative fitting process (Agresti 2002). This may lead one to believe that the variable is not significant, when in fact it may be.

This poster investigates several different methods of dealing with data sparsity, and compares their relative merit with respect to a veterinary epidemiology data set.

The Data Set

Data sparsity was encountered during the analysis of a data set examining *Escherichia coli* O157 hide contamination of 222 Scottish cattle at abattoir. One variable, *Feed in lairage*, examined whether or not feed was supplied to the cattle waiting in lairage, and if so, what type of feed. There were three levels to this variable: no feed provided, hay or straw. In both the hay and straw levels, there were complete separation of the data: all animals provided hay did not have contaminated hides, and all animals provided straw had contaminated hides.

Table 1. Contingency table of hide contamination status (0=negative, 1=positive) and feed in lairage (0=no feed, 1=hay, 2=straw)

		Feed in Lairage		
Hide		**0**	**1**	**2**
Contamination	**0**	81	19	0
Status	**1**	102	0	20

Methods of Dealing with Data Deficiencies

A lot of attention has been focussed on how best to deal with data sparsities. Many methods exist; here, several ad hoc methods as well as more formal methods were applied to the *E. coli* O157 hide contamination data set.

A) Switching select outcomes to eliminate zero cells

Method: switch outcome for one record in each of the *Hay* and *Straw* levels from either a success (1) to a failure (0), or from failure to success

Advantage: easy to implement

Disadvantage: creates artificial data; altered records may become influential; internal validity may be questioned

B) Downweighting select outcomes to eliminate zero cells

Method: downweight the outcome of one record in each of the levels containing a zero cell (from 0 to 0.05, or from 1 to 0.95)

Advantage: easy to implement; alteration not as dramatic as completing switching outcome; allows convergence of the parameter estimate

Disadvantage: still creates artificial data

C) Exact methods

Method: calculate median unbiased estimates (the average of the endpoints of a 50% confidence interval estimator (Hosmer & Lemeshow 2000)), as conditional MLEs do not exist for zero cells

Advantage: probability of overestimation equals that of underestimation; effects of a certain parameter can be determined in cases where lack of convergence causes the unconditional and conditional maximum likelihood procedures to fail (Collett 2003)

Disadvantage: computationally intense, so more suited to small data sets; exact methods more conservative (tends to over-estimate p-values); less powerful than large sample unconditional maximum likelihood estimation (Agresti 2002); some evidence that median unbiased estimates may be unreliable (Collett 2003)

D) Profile likelihood

Method: for each possible pair of parameter values for *Hay* and *Straw*, the profile log-likelihood is obtained by maximising the log-likelihood over the other parameters in the model. Then plot contours of constant profile log-likelihood

Advantage: no artificial data are introduced; obtain joint interval estimates for *Hay* and *Straw*

Disadvantage: the regions shown in Figure 1 correspond to approximate 95% and 99% confidence regions for the 2 parameters; however, the actual confidence coverage of regions based on the profile likelihood is unknown (Aitken *et al.* 1994)

Results

As shown in Table 2, using the switched, downweighted and exact methods, *Hay* remains statistically significant and in a negative (protective) direction; even using the downweighted and exact methods, the upper limit for the odds ratio is well below 1.0 (the lower limit is bounded by zero). The results for *Straw* appear to be more method dependent, with odds ratios ranging from 0.87 to 4.4e+6. However, the majority of the methods (excluding the downweighted method) indicate that the association is not significant. A problem was identified when the switching method (e.g. success to failure) was used, as the switched records became influential.

Table 2. Comparison of the different sparse data methods on the *E. coli* O157 data set for ***Hay***

Method	Coefficient	S.E.	OR	95% CI
Original	-18.3	1410.6	1.20E-08	(0.00, 1.4e+33)
Switched	-2.7	1.1	0.07	(0.004, 0.39)
Downweighted	-5.8	4.5	0.003	(NA, 0.16)
Exact	-3.0	NA	0.05	(NA, 0.34)

Table 3. Comparison of the different sparse data methods on the *E. coli* O157 data set for ***Straw***

Method	Coefficient	S.E.	OR	95% CI
Original	15.3	1433.8	4.40E+06	(1.0e-40, NA)
Switched	-0.09	1.2	0.91	(0.11, 20.23)
Downweighted	2.8	4.5	15.98	(2.22, NA)
Exact	-0.13	NA	0.87	(0.09, NA)

Examination of Figure 1 shows that the upper confidence limit of the effect of *Hay* is -1.65, implying that the odds of hide contamination when eating hay are less than one-fifth of the odds when feed is not provided. Conversely, the odds of hide contamination with *Straw* may range from two-fifths to infinitely larger than that when feed is not provided, as a consequence of its statistical nonsignificance. Of some importance, when the model was reparameterised to compare *Hay* and *Straw* directly, a profile-likelihood based confidence interval demonstrated that the odds of hide contamination was at least 30 times higher in animals provided straw. This method requires more advanced statistical methods than the first two (ad hoc) methods.

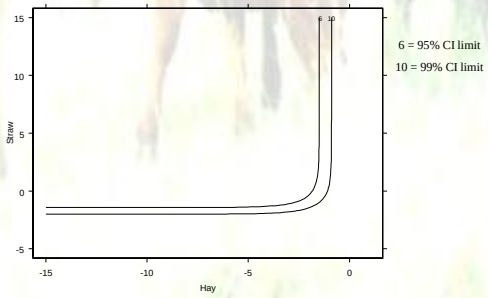

Figure 1. Profile likelihood contour plot for *Hay* (x-axis) and *Straw* (y-axis).

Conclusions

- While the point estimates may vary, virtually all examined methods of dealing with the zero cells in this data set are in agreement with respect to direction of effect and significance
- While this is true for this particular data set, such agreement may not occur with all data sets
- When deciding what approach to take when modelling sparse data, validity of the approach and ease of implementation should be considered

Acknowledgements

This study was funded by an International Partnership Research Award in Veterinary Epidemiology provided by the Wellcome Trust.

References

Agresti A. 2002. Categorical Data Analysis. 2nd Ed. New York: Wiley Interscience.

Aitken A, Anderson D, Francis B, Hinde J. 1994. New York: Oxford University Press Inc.

Collett D. 2003. Modelling Binary Data. Boca Raton: Chapman & Hall/CRC

Dohoo I, Martin W, Stryhn H. 2003. Veterinary Epidemiologic Research. Charlottetown: AVC Inc.

Hosmer DW, Lemeshow S. 2000. Applied Logistic Regression. New York: Wiley.