

Why limit yourself to one logistic regression model?

Menzies FD, Gordon AW[~], Courcier EA, Robinson PA, Abernethy DA

Veterinary Epidemiology Unit, Department of Agriculture and Rural Development (DARD),
Belfast, Northern Ireland. Email: fraser.menzies@dardni.gov.uk

[~] Biometrics Branch, Agri-Food and Bioscience Institute (AFBI), Belfast, Northern Ireland

Introduction

Choice of a regression model usually relies upon use of a forward selection or backward elimination process. This may lead to equivalent models being overlooked particularly in observational studies.

A preferable approach is to fit all possible models and evaluate them. This can be a useful approach where sample size is limited although care must be taken in selection of the terms to be incorporated in the procedure.

Various programs directly offer this procedure (e.g. R's `bestglm`; Genstat's RSEARCH) but it can be carried out in other programs (Stata, SAS, SPSS).

The following example illustrates a study where more than one model is relevant (otherwise information would be lost).

Example

Methods outline

A regional case-control study on the herd risk factors for the disclosure of tuberculin skin test reactors was carried out on a random sample of cattle herds where badger activity had been surveyed (cases = 77 herds ; controls = 108 herds). Main risk factors are listed in the table below.

Cases = Herds that did have one or more TB test reactors during a 3 year period.

Controls = Herds that did not have any TB test reactor during a 3 year period.

GenStat® used for statistical analyses:

Logistic regression ($P \leq 0.15$ cut off for including in multivariate model) and RSEARCH procedure used to select best fit models.

The eight best subsets for inclusion of a different number of terms (2-5) were examined

Results/Discussion

Three models were deemed to be biologically plausible (see table).

Best subsets with 3 terms

Akaike Information Criterion	Schwarz Information Criterion	DF	Previous TB history	Badger activity	Herd Size	Market to Farm Movements	Cattle Tracing
212.02	231.34	6	0.001	0.032	0.001	-	-
219.89	232.77	4	0.000	0.006	-	0.016	-
219.92	232.80	4	0.000	0.002	-	-	0.017

The three models consistently indicated that a previous history of TB in a herd and the observation of badger activity on the farm were risk factors for the presence of TB in a herd. The third term was dependent upon which model was used, with all three considered as valid risk factors (increased herd size, cattle movements onto the farm from markets and epidemiological tracing of cattle).

Conclusions

- This approach allows you to look at all possible model permutations and highlights other logistic regression models that are worthy of consideration.
- In the example, two valid risk factors would be ignored if more traditional methods were used.