Semiparametric-efficient estimation for multi-stage case-control studies

Alan Lee[1], Alastair Scott[1] and Chris Wild[1]

[1]Department of Statistics, University of Auckland, New Zealand

In this paper, we describe the analysis of multi-stage case-control studies. In a simple stratified case-control study, a finite population, or prospective cohort selected from the population, is stratified according to some variables known for the whole cohort. Separate random samples of cases (units with some characteristic of interest) and controls (units without the characteristic) are then drawn from each stratum and values of covariates are obtained for each of the sampled units. In a two-stage study, some of the more expensive or difficult covariates are not measured on all the sampled units, but only on a subsample drawn from them. This process can be continued indefinitely: In a three-stage study, for example, some of the extra covariates are measured on all individuals sampled at the second stage sample while others are only obtained for a further subsample of second stage individuals, and so on.

We present a semi-parametric maximum likelihood approach that extends earlier work, including the seminal case-control paper by Prentice and Pike (1979), Scott and Wild (1997, 2001) Breslow and Holubkov (1997) and others. We derive a set of estimating equations for the semi-parametric maximum likelihood estimator, and demonstrate the efficiency of the solutions to these equations. These take the form of penalized pseudo-likelihood equations, with a term corresponding to the ordinary prospective likelihood, plus penalty terms corresponding to the different stages of sampling. Our methods apply to arbitrary binary regression models, but take on a particularly simple form for logistic regression.

References

[1] Breslow, N.E. and Holubkov, R. (1997). Maximum likelihood estimation of logistic regression parameters for two-phase outcome-dependent sampling. J. Roy. Statist. Soc, B, 59, 447-461.

[2] Prentice, R.L., and Pyke, R. (1979). Logistic disease incidence models and case-control studies. Biometrika, 66, 403-11.

[3] Scott, A.J. and Wild, C.J. (1997). Fitting regression models to case-control data by maximum likelihood. Biometrika, 84, 57-71.

[4] Scott, A.J. and Wild, C.J. (2001). Maximum likelihood for generalised case-control studies. Journal of Statistical Planning and Inference, 96 , 3-27.