Semiparametric Maximum Likelihood for Response-selective and Missing-data problems

Chris Wild
Alastair Scott
University of Auckland, Department of Statistics
Private Bag 92019
Auckland, New Zealand
c.wild@auckland.ac.nz
scott@stat.auckland.ac.nz

The Auckland Birthweight Cooperative (ABC) Study was designed as a hospital-based case-control study for babies being born "small for gestational age" (SGA) defined approximately as being in the lowest 10% of birthweights for gestational age and sex. Approximately 900 cases (SGA) and 900 controls (non-SGA) were sampled and a large number of covariates measured. There were approximately 20,000 births in the contributing hospitals over the same time period, however, and for "all" of them, birthweight, gestational age, sex, mother's "dimensions", smoking habits and a number of other variables were recorded as part of routine hospital procedures. In studies of this sort, only data on case-control study subjects would normally be analysed. Use of data on the other babies is problematic because, for them, there is no data on many of the study covariates of interest. Intuitively, however, it would seem that standard analyses that ignore this additional information may be seriously inefficient.

Millar (1991) gives data from a study of the catching properties of fishing nets in which a fine meshed net and a test net are dragged behind a fishing trawler. Fish sizes were measured for the fish caught in each net. No fish escape the fine net giving us an unconditional sample from the size distribution of fish $g(\boldsymbol{z})$. Many of the smaller fish escape from the test net leaving us with a sample from $g(\boldsymbol{z} \mid y = \text{"caught"})$. For this we wish to model the probability that the test net would retain a fish of a given size, $\text{pr}(y = \text{"caught"} \mid \boldsymbol{x}; \boldsymbol{\theta})$, using a logistic regression model, for example. We have response-selective sampling with the test net, whereas for the fine-net data we have unconditional sampling, but y is missing.

Whittemore (1995) presents data from a case-control family study in which ovarian cancer subjects and controls were sampled and then data obtained about them and their mothers. The data were analysed as a bivariate binary-response (daughter's status, mother's status) regression. Sampling is response selective depending only on the response of the first member of each pair.

What these, and many other, problems have in common is that in each case the likelihood is a special case of the class of likelihoods given below. We need some notation as follows. Let $\boldsymbol{y}$ be a response variable, $\boldsymbol{z}$ contain the covariates that are only available for a subset of subjects, and $\boldsymbol{v}$ contains covariates and correlated variables that are always available. Interest centres

on a parametric regression model $f(\boldsymbol{y} \mid \boldsymbol{x}; \boldsymbol{\theta})$ where $\boldsymbol{x}$ contains $\boldsymbol{z}$ and $\boldsymbol{v}_1$, a subset of $\boldsymbol{v}$. We impose no further parametric assumptions. Our class of likelihood functions is

$$\prod_{s=1}^{S} \prod_{i:\boldsymbol{v}_i=\boldsymbol{v}_s} f(\boldsymbol{y}_i \mid \boldsymbol{z}_i, \boldsymbol{v}_{1s}; \boldsymbol{\theta})^{\Delta_{1i}} g_1(\boldsymbol{z}_i \mid \boldsymbol{v}_s)^{\Delta_{2i}} \mathrm{pr}\{\boldsymbol{h}_{\boldsymbol{v}_s}(\boldsymbol{y}, \boldsymbol{z}) = \boldsymbol{h}^{[i]} \mid \boldsymbol{v}_s; \boldsymbol{\theta}\}^{\Delta_{3i}}, \tag{1}$$

where Δ_{1i} and Δ_{2i} are binary indicator variables, Δ_{3i} takes values $0, \pm 1$, and $\boldsymbol{h}(\cdot)$ is a known function. This latter construction allows sampling conditional on a pattern in multivariate response.

The authors have developed semiparametric maximum likelihood estimation for this class of problems building on the ideas in Scott and Wild (1997, 2001) and Lawless et al. (1999). In our approach, $g(\boldsymbol{z}|\boldsymbol{v}_s)$ is replaced by a discrete distribution with probability mass concentrated on the $\boldsymbol{z}$ values observed in the $\boldsymbol{v} = \boldsymbol{v}_s$ stratum. This gives us a likelihood of the form $L(\boldsymbol{\theta}, \boldsymbol{\delta})$. We profile out $\boldsymbol{\delta}$ to get profile likelihood $L_P(\boldsymbol{\theta}) = L[\boldsymbol{\theta}, \widehat{\boldsymbol{\delta}}(\boldsymbol{\theta})]$. Inference is based upon this profile likelihood. Advantages of the profile likelihood approach include conceptual simplicity and the fact that one can use the same program-engine for a variety of sampling schemes. Estimating equation based methods require a new method for estimating variances for every new missing-data pattern. No modelling of missing data patterns is necessary which makes life easier for the data analyst who needs only build the model of primary interest.

We have written freely available R functions to implement the methodology. The program engine allows for arbitrary models and a variety of restrospective sampling/missing-data patterns. In addition, we have written easy-to-use "front-end" functions for popular models such as logistic regression and linear regression catering for specific retrospective sampling and/or missing-data patterns. Functions for multivariate binary responses, including random intercept models to allow for clustering, are also available. We will show analyses of the data sets above and discuss the strengths and weaknesses of the semiparametric maximum likelihood methodology.

REFERENCES

Lawless, J.L., Kalbfleisch, J.D. and Wild, C.J. (1999). Estimation for Response-selective and Missing Data Problems in Regression, *J. R. Statist. Soc.*, B, **61**, 413-438.

Millar, R.B. (1992). Estimating the size-selectivity of fishing gear by conditioning on the total catch. *J. Am. Statist. Assoc.*, **87**, 962-968.

Scott, A.J. and Wild (1997). Maximum Likelihood Estimation for Case-Control Data. *Biometrika*, **84**, 57-71.

Scott, A.J. and Wild, C.J. (2001a). Maximum Likelihood For Generalised Case-Control Studies, *J. of Statist. Plan. Inf.*, **96**, 327.

Whittemore, A.S. (1995). Logistic regression of family data from case-control studies. *Biometrika* **82**, 57–67. (Correction: **84**, 989–90.)