Efficient analysis with biased samples

Gustavo G. C. Amorim, Chris J. Wild and Alastair J. Scott

Department of Statistics, University of Auckland, Auckland, New Zealand

Introduction

Biased sampling can be found in most data collection processes and is quite common in medical studies. It may arise by design for example, oversampling specific subgroups in order to gain efficiency or inadvertently if someone refuses to respond.

The best-known example occurs in a simple case-control study. The disease status is obtained from a random population, and potential covariates are observed for independent samples taken from each group (case or control).

Note that the sampling is stratified on the response variable. The likelihood depends on $g(x)$, the distribution of the covariate x, which is of no interest and usually too complicated for modelling. On the other hand, if the sampling were performed unconditionally or conditionally on x, the likelihood would be given by

$$L(\theta) = \prod_{i=1}^{N} f(y|x;\theta)g(x) \Rightarrow \frac{\partial}{\partial\theta}\log L(\theta) = \sum_{i=1}^{N}\frac{\partial}{\partial\theta}\log f(y|x;\theta) \tag{1}$$

where, $f(y|x;\theta)$ is the regression model. Note that it is independent of $g(x)$ and so maximization is straightforward.

Anderson [1] and Prentice and Pike [4] have shown that for the binary logistic regression model with an intercept, maximum likelihood estimates from all regression coefficients except for the constant term can be obtained by ignoring the case-control scheme; i.e, the case-control problem can be treated as a prospective one.

Scott and Wild [6] have shown how to adjust the intercept. They have also extended that result for a broader class of models called "multiplicative intercept models" [7]. However, since this property does not hold for arbitrary categorical regression models, methods that can provide efficient estimation without modelling $g(x)$ are of interest.

Missing data

A more general class of problems where biased samples can be found is related to missing data.

Sampling scheme

Let y be the response variable and x a covariate vector. Suppose that complete information is available on y for all units and x is observed for a sample of N individuals generated from $f(y|x;\theta)g(x)$. Let R_i be an indicator variable denoting a complete ($R_i = 1$) or incomplete ($R_i = 0$) observation, for $i = 1, \dots, N$, and $\pi(y_i, x_i) = \mathbb{P}(R_i = 1|y_i, x_i)$. The likelihood is given by

$$\prod_{i=1}^{N}\left[\pi(y_i,x_i)f(y|x;\theta)g(x)\right]^{R_i}\left[(1-\pi(y_i,x_i))f(y)\right]^{1-R_i}, \tag{2}$$

where

$$f(y) = \int f(y|x;\theta)dG(x).$$

Suppose now that y is partitioned into K strata and n_i units are selected from each stratum for full observation. Note that the set $\{i : R_i = 1\}$ corresponds to those units sampled on the second phase. The likelihood can be written as

$$\prod_{i=1}^{I}\left\{\prod_{j=1}^{n_i} f(y_{ij}|x_{ij};\theta)g(x_{ij})\right\} f(y_i)^{N_i-n_i} \tag{3}$$

which is the likelihood used by Lawless et.at [2] for response-selective problems.

Methods

The most basic approach is to ignore the sampling scheme and consider only the complete observed units. In general, this will not lead to an efficient analysis of the data. With the purpose of improving it, several methods have been proposed.

Weighted likelihood

The weighted approach considers only the observed data, ignoring the incomplete ones. Here, each unit is weighted by the inverse of its probability of being selected for full observation. The likelihood is

$$L(\theta) = \prod_{i=1}^{I}\prod_{j=1}^{n_i} w_i \log f(y_i|x_{ij};\theta), \quad \text{where } w_i^{-1} = \mathbb{P}(R=1|y_i,x_i) = \pi(y_i,x_i) \tag{4}$$

It is known to be robust, but inefficient.

Conditional likelihood

Conditional likelihood is an alternative approach used to increase efficiency. Lawless et. al. [2] have shown that this approach is more efficient than the weighted method for different situations. The likelihood is given by

$$L(\theta, G) = \prod_{i:R_i=1}\mathbb{P}(y_i, x_i|R_i = 1) = \prod_{i:R_i=1}\frac{\mathbb{P}(R_i = 1|y_i,x_i)f(y_i|x_i\theta)g(x_i)}{\mathbb{P}(R_i = 1)} \tag{5}$$

It leads to a dependency on G, because

$$\mathbb{P}(R_i = 1) = \int \mathbb{P}(R_i = 1|y,x)f(y|x)g(x)dxdy$$

In order to avoid this dependence we can condition the likelihood on x as well

$$L(\theta) = \prod_{i:R_i=1}\mathbb{P}(y_i|x_i, R_i = 1) = \prod_{i:R_i=1}\frac{\pi(y_i,x_i)f(y_i|x_i)}{\int \pi(y,x_i)f(y|x_i)dy} \tag{6}$$

and maximization is now straightforward. Note that both methods depend on π. Lee et. al. [3] have shown that by estimating that probability, there will be a gain in efficiency even if π is known.

Maximum likelihood method

The most efficient methods can be found here. They use complete likelihood to make inferences on y, achieving full efficiency in special cases. The most basic methods are the estimated likelihood and the EM algorithm.

- ***EM algorithm***: Reilly and Pepe [5] applied the EM algorithm to maximize the likelihood (2). Although it uses the complete data set, the mean score, as it is called, can be shown to be related to the weighted method.
- ***Estimated likelihood***: Since $G(x) = P(X \leq x) = \sum_{i=1}^{K}\mathbb{P}(y = i)\mathbb{P}(X \leq x|y = i)$. we can maximize (2) by replacing $G(x)$ with $\tilde{G}(x)$, where

$$\tilde{G}(x) = \sum_{i=1}^{I}\frac{N_i}{N}\hat{G}_i(x), \quad \text{where } \hat{G}_i(x) \text{ is the ECDF.} \tag{7}$$

Weaver and Zhou have also used $\hat{G}_i(x)$, but for continuous y with no stratum information for the missing data.

Fully efficient estimators can be obtained by profiling the full likelihood L_F, i.e, $L_{FP}(\theta) = \sup_G L_F(\theta, G)$. Maximization should be made over all possible values of θ and distributions of G. However, Zhang and Rockette [9] suggested working with a simpler maximization that is asymptotically equivalent. Instead of using the global MLE, full likelihood can be maximized under the restriction that G is supported by the observed values of x.

Fully efficient methods have been developed for the following special cases

- The multivariate case-control problem for discrete y and x: an algorithm for obtaining the MLE was developed by Scott and Wild [7]. algorithm for obtaining the MLE for the multivariate case-control problem for discrete y and x.
- Song et.al. [8] obtained fully efficient estimator with a continuous response, but discrete covariates.
- Zhao et.al. [10] worked with three distinct problems: x MAR, y MAR and both variables MAR, when both variables are discrete.

Objectives and Perspectives

We are interested in the following objectives:

- Unify the work that has been done on conditional likelihood;
- To develop asymptotic expressions for the efficiency of weighted, conditional and full likelihood methods;
- Describe situations where the conditional approach is substantially better than weighting, when both methods achieve full efficiency and when the loss is neglible.

References

[1] Anderson, J. A. Separate sample logistic discrimination. *Biometrika 59* (1972), 19–35.

[2] Lawless, J. F., kalbfleisch, J. D., and Wild, C. J. Semiparametric methods for response-selective and missing data problemas in regression. *Journal of the Royal Statistical Society B 61*, 2 (2006), 413–438.

[3] Lee, A. J., Scott, A. J., and Wild, C. J. Efficient estimation in muti-phase case-control studies. *Biometrika 97* (2010), 361–374.

[4] Prentice, R. L., and Pike, R. Logistic disease incidence models with case-control studies. *Biometrika 66* (1979), 403–411.

[5] Reilly, M., and Pepe, M. S. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika 82* (1995), 299–314.

[6] Scott, A. J., and Wild, C. J. Fitting logistic model under case-control or choice based sampling. *Journal of the Royal Statistical Society B 48*, 2 (1986), 170–182.

[7] Scott, A. J., and Wild, C. J. Fitting regression models to case-control data by maximum likelihood. *Biometrika 84*, 1 (2006), 57–71.

[8] Song, R., Zhou, H., and Kosorok, M. R. A note on semiparametric efficient inference for two-stage outcome-dependent sampling with a continuous outcome. *Biometrika 96*, 1 (2009), 221–228.

[9] Zhang, Z., and Rockette, H. E. On maximum likelihood estimation in parametric regression with missing covariates. *Journal of Statistical Planning and Inference*, 134 (2005), 206–223.

[10] Zhao, Y., Lawlles, J. F., and McLeish, D. L. Likelihood methods for regression models with expensive variables missing by design. *Biometrical Journal 51*, 1 (2009), 123–136.