18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Semiparametric prevalence estimation from a two-phase survey

Denis H. Y. Leung[1] and Jing Qin[2]

[2]*School of Economics, Singapore Management University, 90 Stamford Road, Singapore*
e-mail: denisleung@smu.edu.sg
[2]*Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, NIH, 6700B Rockledge Drive MSC 7609, Bethesda, MD 20892*
e-mail:jingqin@niaid.nih.gov

Abstract

This paper studies a semi-parametric method for estimating the prevalence of a binary outcome using a two-phase survey. The motivation for a two-phase survey is, due to time, money and ethical considerations, it is impossible to carry out comprehensive evaluation on all subjects in a large random sample of the population. Rather, a relatively inexpensive "screening test" is given to all subjects in the random sample and only individuals more likely to have a positive outcome (cases) will be selected for a further "gold standard" test to verify the outcome. Therefore, individuals with verified outcome form a non-random sample from the population and care must be taken when the data are used for estimating the prevalence of the outcome. This paper proposes a semi-parametric method for estimating the outcome prevalence. It requires only an estimate of the probability of selection into the second phase, given the first phase data. This feature is desirable as in most cases, the probability of selection into the second phase is under the control of the researchers, and even when it is not, can be easily estimated given the data. The proposed method uses the empirical likelihood approach (Owen, 1988), which yields consistent prevalence estimates as long as the probability of selection into the second phase is correctly modeled.

Keywords: Empiricial likelihood, missing data, surrogate, two-phase sampling.

1. INTRODUCTION

Two-phase sampling is popular in population surveys where the prevalence of a rare outcome is to be estimated (e.g., Beckett, Scherr and Evans, 1992). In such surveys, data are collected in two phases. In the first phase, a random sample is selected from the population and for each individual in the sample, a surrogate of the outcome is measured. Based on the value of the surrogate, a subset of the first phase sample is then selected, on which the outcome is confirmed using a more extensive test. The data thus consist of two parts. The first part is formed by a random sample of the surrogate in the population while the second part is composed of a biased sample of the outcome. Therefore, the problem of estimating the prevalence using data from a two-phase survey can be seen as that of a biased sampling problem with auxiliary data (Vardi, 1982). Alternatively, it can also be regarded as a case-control study where the individuals with high value of the surrogate are "'cases" and those with low value of the surrogate are "controls".

Since the second phase data is a biased sample selected from the first phase sample, prevalence estimation must be adjusted for selection bias. Previous works in this problem have focused on semi-parametric methods. These works fall primarily into two streams. In the first stream, the probability of selection into phase two is estimated using data from both phases. Each observation in the second phase is then weighted (inversely) by the selection probability estimate, in the spirit of the Horvitz-Thompson estimate (Horvitz and Thompson, 1952). Using this method, Robins and Rotnitzky (1995) proposed a class of estimators for the mean response from a longitudinal study in the presence of missing response data. In the second stream, observations in both phases are directly used for inference. The observations with missing outcome status will have their statuses imputed using estimates. Roberts, Rao and Kumar (1987) used this method to study unemployment rate data collected in a multi-level survey. Pepe, Reilly and Fleming (1994) considered this approach in regression analyses with incomplete covariate information. The imputation method requires modeling of the conditional probability of the outcome given the observed data. It may lead to biased results if the probability of outcome is incorrectly modeled. On the other hand, methods based on weighting require only modeling of the probability of selection into the second phase given the observed data, even though, in small and moderate samples, they are known to be less effecient than imputation methods. Robinson (1988) considered a hybrid of the weighting and the imputation methods. The methods in these works require the estimation of *both* the conditional probability of selection into the second phase and of the outcome, given the observed data. Comparisons of these method can be found in Clayton *et al.* (1998).

In this paper, a method is considered for prevalence estimation in two-phase samples. The method requires an estimate of the probability of selection into the second phase, given the first phase data. But since in most cases, the probability of selection into the second phase is under the control of the researchers, and even when it is not, can be estimated consistently given the data, this criteria can easily be satisfied. The proposed method is based on the empirical likelihood (EL) of Owen (1988).

2. METHOD AND MAIN RESULTS

Assume that N observations are randomly sampled from the population in the first phase. A variable X is measured on each observation. In here, X is assumed to be a scalar. However, the method proposed here generalizes easily to situations where X is a vector. Based on X, a second phase sampling is carried out in which n observations are selected from the N observations. For each of the n observations selected in the second phase, the outcome will be confirmed. Without loss of generality, let the outcome of interest be a binary variable, D (=0 or 1). Then the data are

$$x_1, ..., x_{n_1}, D = 1; \qquad x_{n_1+1}, ..., x_{n_1+n_2}, D = 0; \qquad x_{n_1+n_2+1}, ..., x_N, D \text{ unknown};$$

where $n_1 + n_2 = n$. Let Π_1 and Π_2 denote, respectively, the sub-populations of "case" and "non case" individuals. If the distribution of X in the subpopulations Π_1 and Π_2 are F_1 and F_2, respectively, then the data consist of independent observations $x_i, i = 1, 2, ..., n_1$ from F_1, $x_i, i = n_1 + 1, 2, ..., n_1 + n_2$ from F_2 and $x_i, i = n_1 + n_2 + 1, ..., N$ from the mixture distribution $F = \pi F_1 + (1 - \pi)F_2$, where π represents the proportion in the sub-population, Π_1 and $1 - \pi$ represents the proportion in Π_2. Therefore, the problem can

be treated as a case-control or choice-based sample with contaminated data (the contamination comes from the last $N-n$ observations, Imbens, 1992). The interest is to estimate π. Let S be a binary indicator that denotes whether an observation has been selected into the second phase. Since the second phase selection process depends only on the value of X, we can assume that, given X, the selection process is independent of D. Therefore

$$P(S=1|D,x) = P(S=1|x) = w(x), \tag{1}$$

where w is a known probability function if sampling is controlled by the researchers. Even if w is unknown or the sampling scheme is very complicated, given $x_1, ..., x_N$ and the knowledge of which observations have been selected, w can be estimated. It can easily be shown that (1) is equivalent to the condition

$$P(D=1|S,x) = P(D=1|x) = u(x). \tag{2}$$

The full likelihood based on the observed data is

$$\prod_{i=1}^{n}[w(x_i)\pi f_1(x_i)]^{D_i}[w(x_i)(1-\pi)f_2(x_i)]^{1-D_i}\prod_{j=n+1}^{N}[(1-w(x_j))\{\pi f_1(x_j)+(1-\pi)f_2(x_j)]. \tag{3}$$

When no assumptions are made about f_1 and f_2, it is not clear how to handle the last term of (3). Therefore, instead of working with the full likelihood, the following semi-parametric empirical likelihood (EL) approach will be adopted.

To motivate the method in this paper, consider (1). Note that if random sampling was performed to obtain the second phase data, then, in large sample, one would expect

$$\sum_{i=1}^{n} P(S_i|x_i) \cong \sum_{i=1}^{N} P(S_i|x_i) \cong \frac{n}{N}. \tag{4}$$

However, because sampling into the second phase is non-random, the first term in the above expression will no longer be the approximately the same as the second and third terms, even in large samples. In fact,

$$\sum_{i=1}^{n} P(S_i|x_i)P(x_i|i \in \text{second phase}) \cong \sum_{i=1}^{N} P(S_i|x_i) \cong \frac{n}{N}. \tag{5}$$

Using the same argument, one would expect

$$\sum_{i=1}^{n} P(D_i|x_i)P(x_i|i \in \text{second phase}) \cong \sum_{i=1}^{N} P(D_i|x_i) \cong \pi. \tag{6}$$

Therefore, if estimates of $P(x_i|i \in \text{second phase})$, say p_i, can be obtained, then the p_i's can be substituted into (6) and obtain an estimate of the parameter of interest, π. However, this still requires an estimate of $P(D_i|x_i)$, which needs to be estimated using the (second phase) data by a parametric model such as a

logistic regression model. But since $E(D_i|x_i) = P(D_i|x_i)$, therefore,

$$\sum_{i=1}^{n} D_i P(x_i | i \in \text{second phase}). \tag{7}$$

can be used. The problem now lies in how to obtain $p_i, i = 1, ..., n$. This problem will be approached using the method of EL.

Assume $P(s_i|x_i)$ be estimated by $\hat{w}(x_i) = \hat{w}_i, i = 1, ..., N$ using data from the first and second phase. Let $p_1, \cdots, p_n$ be non-negative weights allocated to the second phase sample $x_i, i = 1, ..., n$. Then the second phase sample has the following EL for the parameter π

$$L(\pi) = \max \prod_{i=1}^{n} p_i \tag{8}$$

subject to

$$0 \le p_i \le 1, i = 1, ..., n, \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i \hat{w}_i = \frac{n}{N}, \quad \sum_{i=1}^{n} p_i D_i = \pi. \tag{9}$$

Note the similarilities in the last constraint in (9) and (7). By introducing Lagrange multipliers and following standard derivations in EL, the optimal p_i's for given π is

$$p_i \quad = \quad \frac{1}{n} \frac{1}{1 + \lambda(\hat{w}_i - \frac{n}{N}) + \gamma(D_i - \pi)}, \quad \text{for} \quad i = 1, \cdots, n, \tag{10}$$

where the Lagrange multipliers λ and γ, satisfy the following equations:

$$\sum_{i=1}^{n} \frac{\hat{w}_i - \frac{n}{N}}{1 + \lambda(\hat{w}_i - \frac{n}{N}) + \gamma(D_i - \pi)} \quad = \quad 0, \tag{11}$$

$$\sum_{i=1}^{n} \frac{D_i - \pi}{1 + \lambda(\hat{w}_i - \frac{n}{N}) + \gamma(D_i - \pi)} \quad = \quad 0. \tag{12}$$

From (10), we have the negative log EL

$$\ell(\pi) = -2 \log\{L(\pi)\} \quad = \quad 2 \sum_{i=1}^{n} \log\{1 + \lambda(\hat{w}_i - \frac{n}{N}) + \gamma(D_i - \pi)\} - 2n \log(n). \tag{13}$$

Differentiating (13) with respect to π and using equations (11) and (12) lead to

$$\sum_{i=1}^{n} \frac{\gamma}{1 + \lambda(\hat{w}_i - \frac{n}{N}) + \gamma(D_i - \pi)} \quad = \quad 0. \tag{14}$$

From (14), it can be deduced that $\gamma = 0$. Therefore, (11) and (12) become

$$\sum_{i=1}^{n} \frac{\hat{w}_i - \frac{n}{N}}{1 + \lambda(\hat{w}_i - \frac{n}{N})} = 0, \quad (15)$$

$$\sum_{i=1}^{n} \frac{D_i - \pi}{1 + \lambda(\hat{w}_i - \frac{n}{N})} = 0. \quad (16)$$

Let $(\hat{\pi}, \hat{\lambda})$ be the solutions to equations (15) and (16). Then, $\hat{\pi}$ is the maximum EL estimate of π. Therefore, $p_i, i = 1, ..., n$ and $\hat{\pi}$ are obtained in one unify framework under EL. It can be shown that $\hat{\pi}$ is asymptotically normal.

The log EL ratio for π evaluated at π_* is $R(\pi_*) = -\ell(\pi_*) + \ell(\hat{\pi})$. It can be shown (proof upon request) that $R(\pi_*) \xrightarrow{d} \chi_1^2$ as $N \to \infty$, $n/N \to \delta > 0$. The large sample EL ratio result can be used to construct confidence limits for π. In particular, let c_α be the upper α-percentile of χ_1^2 for $\alpha \in (0, 1)$. Then an α-level confidence interval is $CR_\alpha = \{\pi | \ell(\pi) \leq c_\alpha\}$.

3. SIMULATION RESULTS

This section reports the results of a simulation study that compares the finite sample properties of the estimator, $\hat{\pi}$ to some existing estimators. The näive complete case estimator (CC) uses only outcome data from the second phase, *i.e.*,

$$\hat{\pi}_{CC} = \frac{1}{N} \sum_{i=1}^{N} S_i D_i.$$

It is well known that $\hat{\pi}_{CC}$ will give biased estimates unless the second phase data form a random sample of the population. Given an estimtor $\hat{w}(x)$ of $P(S = 1|x) = w(x)$, the inverse probability estimator (IPW) (Horvitz and Thompson, 1952) is given by

$$\hat{\pi}_{IPW} = \left(\sum_{i=1}^{n} \hat{w}_i^{-1} S_i \right)^{-1} \sum_{i=1}^{n} \hat{w}_i^{-1} D_i.$$

Similar to $\hat{\pi}$, $\hat{\pi}_{IPW}$ gives unbiased estimates if $w(x)$ is modeled correctly.

Three imputation estimators were considered. The first one imputes the outcome status of all subjects by an estimate, $\hat{u}(x)$, of $P(D = 1|x) = u(x)$ obtained using the second phase data. Hence it is sometimes called the full imputation method (FI). It estimates π by

$$\hat{\pi}_{FI} = \frac{1}{N} \sum_{i=1}^{N} \hat{u}_i.$$

The second imputation estimator solves a score equation of π where the scores for the observations with missing outcome status are imputed by the mean scores over the (unknown) outcome distribution. Pepe *et al.* (1994) and Clayton *et al.* (1998) studied this mean score imputation (MSI) method using a binomial score, which leads to:

$$\hat{\pi}_{MSI} = \frac{1}{N} \sum_{i=1}^{N} \left(S_i D_i + (1 - S_i) \hat{u}_i \right),$$

where $\hat{u}_i$ is an estimate of the probability of outcome, given the observed data. Both $\hat{\pi}_{MSI}$ and $\hat{\pi}_{FI}$ are consistent estimates of π if $u(x)$ is correctly modeled, otherwise both will be biased. The third imputation estimator requires modeling both $u(x)$ and $w(x)$. However, it has the "double robustness" property that unbiased estimates of π can be obtained as long as one of $u(x)$ or $w(x)$ is modeled correctly. The doubly robust (DR) estimator is given by

$$\hat{\pi}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{S_i D_i}{\hat{w}_i} - \left(1 - \frac{S_i}{\hat{w}_i}\right)\hat{u}_i\right].$$

Three values of π were used in the simulations: 0.2, 0.3, 0.4. In all simulations, the first phase sample consisted of $N = 500$ observations. For the i-th individual, the value of x_i was a random variable from a N(0,0.5) distribution if the individual's underlying outcome status was "control" and it was a random variable from a N(1,0.5) distribution if the underlying outcome status was "case". For the i-th observation in the first phase, the probability of selection into the second phase sample was given by a logistic model

$$P(S_i = 1|x_i) = 1/\{1 + \exp(\zeta - \zeta x_i)\}.$$

The i-th observation was selected into the second phase if $P(S_i = 1|x_i) > U_i$, where U_i followed a U(0,1) distribution. In the simulations, the following values of ζ were used: 1.5, 2, 2.5 , 3. These values induced proportions of data with missing outcome (i.e., data not selected into phase two) ranging from 0.648 to 0.822. In the simulations, $u(x)$ and $w(x)$ were modelled by separate logistic models with an intercept and a linear term in X. These were the correct models under the set-up of the simulations. We chose not to study situations where these models are mis-specified because in those situations, the conclusions depend critically on the type and degree of mis-specification and cannot be easily generalized.

Table 1 summarizes the Monte-Carlo means and standard errors of the estimates of π using the different methods. The results clearly show that $\hat{\pi}_{CC}$ is biased in all senarios considered, which is to be expected. The other methods are all approximately unbiased. When the proportion of data with missing outcome is high, there is some slight bias in $\hat{\pi}_{IPW}$, even though in other simulations with a larger N (not shown here) the bias disappeared. It is clear that compared to IPW, EL is more efficient in almost all senarios. The advantage of EL over IPW is higher when the proportion of data with missing outcome is higher and when the unknown π is higher. In fact, EL's efficiency is comparable to the imputation methods, which require a correct model of the probability of outcome. In this simulation study, the imputations are all based on a correct model of the probability of outcome. Therefore, the results for EL is very promising.

4. CONCLUSION

This paper, we introduced a semi-parametric for estimating the prevalence using data from a two-phase survey. Using simulations, we showed that the method gives more accurate estimates than the inverse probability weighted estimator, which uses the same assumptions and set-up as the proposed method. We also demonstrated that the method gives estimates that are competitive to the imputation methods, which depends on correctly modelling the outcome given the observed data. This requirement is often harder to satisfied that the one required by the proposed method (and of the inverse probability weighted method) of a correct model for selection into second phase.

Table 1. Mean (standard error) in estimating π. Based on 1000 simulations.

			$\pi = 0.2$	
proportion missing[1]	0.7360	0.7756	0.8035	0.8220
IPW	0.1997 (0.0264)	0.2014 (0.0312)	0.2024 (0.0407)	0.2120 (0.0551)
EL	0.1989 (0.0268)	0.1993 (0.0289)	0.1980 (0.0318)	0.1991 (0.0398)
CC	0.3762 (0.0438)	0.4466 (0.0475)	0.5083 (0.0509)	0.5654 (0.0536)
FI	0.1997 (0.0258)	0.2009 (0.0277)	0.1997 (0.0291)	0.2020 (0.0322)
MSI	0.1997 (0.0258)	0.2009 (0.0277)	0.1997 (0.0291)	0.2020 (0.0322)
DR	0.1995 (0.0259)	0.2003 (0.0281)	0.1991 (0.0308)	0.2004 (0.0370)
			$\pi = 0.3$	
proportion missing	0.7052	0.7413	0.7656	0.7826
IPW	0.3012 (0.0318)	0.3014 (0.0404)	0.3032 (0.0533)	0.3155 (0.0731)
EL	0.2998 (0.0297)	0.2991 (0.0332)	0.2975 (0.0369)	0.3009 (0.0471)
CC	0.5114 (0.0420)	0.5786 (0.0430)	0.6379 (0.0437)	0.6879 (0.0437)
FI	0.3005 (0.0290)	0.3002 (0.0316)	0.2987 (0.0331)	0.3008 (0.0376)
MSI	0.3005 (0.0290)	0.3002 (0.0316)	0.2987 (0.0331)	0.3008 (0.0376)
DR	0.3005 (0.0293)	0.3000 (0.0328)	0.2982 (0.0353)	0.3001 (0.0405)
			$\pi = 0.4$	
proportion missing	0.6758	0.7064	0.7278	0.7423
IPW	0.4025 (0.0336)	0.4003 (0.0475)	0.4043 (0.0633)	0.4187 (0.0871)
EL	0.4015 (0.0324)	0.3988 (0.0361)	0.3979 (0.0418)	0.4034 (0.0556)
CC	0.6190 (0.0384)	0.6825 (0.0368)	0.7343 (0.0370)	0.7760 (0.0353)
FI	0.4026 (0.0315)	0.3997 (0.0345)	0.3994 (0.0374)	0.4013 (0.0404)
MSI	0.4026 (0.0315)	0.3997 (0.0345)	0.3994 (0.0374)	0.4013 (0.0404)
DR	0.4024 (0.0316)	0.3995 (0.0352)	0.3979 (0.0394)	0.4010 (0.0461)

[1]: average proportion of data with missing outcome

References

Beckett, L., Scherr, P. and Evans, D. (1992) Population prevalence estimates from complex samples. *Journal of Clinical Epidemiology*, **45**, 393–402.

Clayton, D., Dunn, G., Pickles, A. and Spiegelhalter, D. (1998) Analysis of longitudinal binary data from multi-phase sampling. *Journal of the Royal Statistical Society, Series B*, **60**, 71–87.

Horvitz, D. G. and Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663–685.

Imbens, G. (1992) An efficient method of moments estimator for discrete choice models with choice-based sampling. *Econometrica*, **60**, 1187–1214.

Owen, A. (1988) Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–49.

Pepe, M. S., Reilly, M. and Fleming, T. R. (1994) Auxiliary outcome data and the mean-score method. *J. Statist. Planning and Inference*, **42**, 137–160.

Roberts, G., Rao, J. and Kumar, S. (1987) Logistic regression analysis of sample survey data. *Biometrika*, **74**, 1–12.

Robins, J. M. and Rotnitzky, A. (1995) Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, **90**, 122–129.

Robinson, P. M. (1988) Root-n-consistent semiparametric regression. *Econometrica*, **56**, 931–954.

Vardi, Y. (1982) Nonparametric estimation in the presence of length bias. *Annals of Statistics*, **10**, 616–620.