Fitting Models to Complex Survey Data Accounting for Nonignorable Sampling and Nonresponse

Danny Pfeffermann[1,2,3], and Moshe Feder[1]
[1] Southampton statistical Sciences Research Institute, University of Southampton, UK
[2] Department of Statistics, Hebrew University of Jerusalem, Israel
[3] Corresponding author: Danny Pfeffermann, e-mail: msdanny@soton.ac.uk

Abstract

When the sample selection probabilities and/or the response probabilities are related to the model dependent variable even after conditioning on the model covariates, the model holding for the sample data is different from the model holding in the population from which the sample is taken. Ignoring the sample selection or response mechanism in this case may result in biased inference. Accounting for sample selection bias is relatively simple because the sample selection probabilities are usually known. In this paper we consider the much harder problem where in addition to sample selection bias, the response mechanism is also not ignorable with unknown response probabilities. Our approach is based on empirical likelihood, which is defined with respect to the model holding for the data observed for the responding units. Simulation results with binary dependent outcomes illustrate the good performance of the proposed approach.

Keywords: Empirical likelihood, NMAR nonresponse, Sample model.

1. Introduction

Survey data are often used for analytic inference on statistical models assumed to hold for the population from which the sample is taken. It is often the case, however, that the sampling design used to select the sample is informative for the population model in the sense that the sample selection probabilities are correlated with the target outcome variables even after conditioning on the model covariates, in which case the model holding for the sample data is different from the population model.

Inevitably, sample data are subject to non-response, which is informative for the population model if the response probabilities are correlated with the outcome values after conditioning on the model covariates, known as `not missing at random' (NMAR) non-response. Here again, the model holding for the data observed for the responding units is different from the sample model under complete response, which as noted above is different from the population model under informative sampling. Clearly, ignoring an informative sampling design and/or response mechanism may yield highly biased estimators and distort the inference.

Pfeffermann (2011) reviews several approaches proposed in the literature to deal with informative sampling, ranging from weighting each sample observation by the corresponding sampling weight, to maximization of the sample likelihood as defined by the model holding for the sample data. A common feature of these approaches is that they utilize the sampling weights in the inference process, although in different ways. On the other hand, accounting for NMAR non-response is much more complicated as the response probabilities are practically never known, requiring some assumptions on them. Pfeffermann and Sikov (2011) review approaches proposed in the literature to deal with NMAR non-response, but these approaches are quite limited. In particular, most of the approaches assume that the model covariates are known also for the non-respondents, which is often not the case.

Evidently, accounting for both informative sampling and NMAR non-response in a single analysis is a major undertaking, and the present article attempts to tackle this problem. We assume that not only the outcome values are missing for the non-responding units but also the covariates, known as unit non-response. The only additional information beyond the data observed for the responding units assumed to

be known is the population totals of calibration variables, which may include some of the model covariates, and possibly also the outcome variable. The totals of such calibration variables are often available from administrative or census records. Our proposed approach uses the empirical likelihood (EL) as the basis for inference on the target population model. The use of EL for analyzing complex survey data has its origins in a landmark paper by Hartley and Rao (1968), and has gained increasing interest in recent years in a general context following Owen (1988, 1990, 1991, 2001). Another important reference is Qin and Lawless (1994). The EL combines the robustness of nonparametric methods with the effectiveness of the likelihood approach. Another important advantage of this method is that it lends itself very naturally to the use of calibration constraints, thus enhancing the precision of the estimators. See, *e.g.*, Chen and Keilegom (2009) for a recent review. The use of this approach has also computational advantages over fully parametric approaches.

2. The sample and respondents distributions

Let y_i denote the value of an outcome variable Y associated with unit i belonging to a sample S drawn from a finite population $U = \{1,...,N\}$ with known inclusion probabilities $\pi_i = Pr(i \in s)$. Let I_i define the sampling indicator; $I_i = 1(0)$ if unit i is sampled (not sampled), and $\mathrm{x}_i = (x_{1i},...,x_{pi})'$ denote the values of p auxiliary variables (covariates) associated with unit i. In what follows we assume that the population outcomes are independent realizations from distributions with probability density functions (*pdf*) $f_P(y_i|x_i)$, where for now we omit for convenience the underlying parameters from the notation. Following Pfeffermann *et al.* (1998), the *marginal sample pdf*, $f_S(y_i \mid \mathrm{x}_i)$, defines the conditional *pdf* of y_i given that unit i is in the sample ($I_i = 1$). By Bayes rule, $f_S(y_i \mid \mathrm{x}_i) = f(y_i \mid \mathrm{x}_i, I_i = 1)$ $= \Pr(I_i = 1 \mid \mathrm{x}_i, y_i) f_p(y_i \mid \mathrm{x}_i) / \Pr(I_i = 1 \mid \mathrm{x}_i)$. Note that $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i)$ is generally not the same as the sample selection probability $\pi_i = \Pr(I_i = 1 \mid Z_U)$, where Z_U defines a matrix of population values of design variables used for the sample selection. Typically, $\Pr(I_i = 1 \mid \pi_i, y_i, \mathrm{x}_i) = \pi_i$, in which case $\Pr(I_i = 1 \mid y_i, \mathrm{x}_i)$ $= E_P(\pi_i \mid y_i, \mathrm{x}_i)$, where $E_P(\cdot)$ is the expectation under the population *pdf*. The population and sample *pdf*s differ unless $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i) = \Pr(I_i = 1 \mid \mathrm{x}_i)$ for all y_i, and when this condition is not met, the sampling design is informative and cannot be ignored in the inference process.

Next consider the respondents distribution. Let $R = \{\mathrm{i} \mid R_i = 1\}$ define the sample of respondents of size r with observed outcomes and covariates where $R_i = 1(0)$ if sample unit i responds (does not respond). The response probabilities may depend on covariates $\mathbf{v}$, which may differ from x in one or more components. The marginal *pdf* for responding unit i is then $f_R(y_i \mid \mathrm{x}_i) = f(y_i \mid \mathrm{x}_i, I_i = 1, R_i = 1)$ $= \Pr(R_i = 1 \mid y_i, \mathrm{v}_i, I_i = 1) f_S(y_i \mid \mathrm{x}_i) / \Pr(R_i = 1 \mid \mathrm{v}_i, \mathrm{x}_i, I_i = 1)$. Note that unless $\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1) = \Pr(R_i \mid \mathrm{x}_i, I_i = 1)$ for all y_i, the respondents *pdf* differs from the sample *pdf*.

So far we excluded for convenience from the notation the parameters governing the various distributions. If the outcome and the response are independent between the units, the respondents' likelihood has the form,

$$L_{\mathrm{Resp}}(\gamma, \theta) = \prod_{i=1}^{r} \frac{\Pr(R_i = 1 \mid y_i, \mathrm{v}_i, I_i = 1; \gamma) \Pr(I_i = 1 \mid y_i, \mathrm{x}_i) f_P(y_i \mid \mathrm{x}_i; \theta)}{\Pr(R_i = 1 \mid \mathrm{x}_i, \mathrm{v}_i, I_i = 1; \theta, \gamma) \Pr(I_i = 1 \mid \mathrm{x}_i)}. \qquad (1)$$

Remark 1. In theory, one needs to model also the probabilities $\Pr(I_i = 1 | y_i, \mathrm{x}_i)$ but under the mild assumption that $\Pr(I_i = 1 | \pi_i, y_i, \mathrm{x}_i) = \pi_i$, the probability $\Pr(I_i = 1 | y_i, \mathrm{x}_i)$ can be estimated outside the likelihood using the relationship $\Pr(I_i = 1 | y_i, \mathrm{x}_i) = E_P(\pi_i | y_i, \mathrm{x}_i) = 1 / E_S(w_i | y_i, \mathrm{x}_i)$, where $w_i = 1/\pi_i$ is the sampling weight. Thus, the probabilities $\Pr(I_i = 1 | y_i, \mathrm{x}_i)$ can be estimated by regressing w_i against (y_i, x_i) using the sample data, and then plugging the estimates into (1). See Pfeffermann and Sverchkov (2003, 2009) for different approaches and examples of modeling and estimating the expectations $E_S(w_i | y_i, \mathrm{x}_i)$. Alternatively, and as illustrated in Pfeffermann (2011), the expectations $E_S(w_i | y_i, \mathrm{x}_i)$ can be estimated nonparametrically, and this is the approach adopted in the present paper.

3. Conditional empirical likelihood, given the Respondents' data

We assume that for each unit i corresponds a vector $u_i = (y_i, \mathrm{x}_i', c_i', \tau_i, \rho_i)'$ where y_i and x_i are related via a model $f_p(y_i | \mathrm{x}_i; \beta)$, c_i is a vector of survey values for which the population means are known, $\tau_i = E_p(I_i | y_i, \mathrm{x}_i)$ and $\rho_i = E_p(R_i | y_i, \mathrm{x}_i, I_i = 1)$. We employ the load-scale approach of Hartley and Rao (1968), assuming that the finite population values are generated from a multinomial distribution. Suppose that the only values attained in the finite population U are those observed for the respondents. Denote by N_i the number of units $j \in U$ assuming the vector u_i and let $p_i = N_i / N$, where $N = \sum_i N_i$. The vector parameter of interest is $p = (p_1, ..., p_r), \sum_{i=1}^{r} p_i = 1$.

<u>*R-level empirical likelihood*</u>: The distribution of the observed data is the respondents' distribution (hereafter the R-level distribution), and the likelihood is, $EL = \prod_{i \in R} p_i^{(r)}$, where $p_i^{(r)} = \Pr(u_i | i \in R) = p_i \tau_i \rho_i / \sum_{k \in R} p_k \tau_k \rho_k$. Chaudhuri *et al.* (2010) use the same likelihood for the case of full response.

<u>*R-level constraints*</u>: Under the assumptions above, $\sum_{i \in R} p_i c_i = N^{-1} \sum_{i \in R} N_i c_i$ $= N^{-1} \sum_{j \in U} c_j = \bar{c}_U$, yielding the R-level constraints $\sum_{i \in R} p_i^{(r)} \tau_i^{-1} \rho_i^{-1} (c_i - \bar{c}_U) = 0$. In particular, denoting $\zeta_i = \tau_i \rho_i$ we have, $\sum_{i \in R} p_i \zeta_i = \bar{\zeta}_U$, which is equivalent to $\sum_{i \in R} p_i^{(r)} [1 - (\bar{\zeta} / \zeta_i)] = 0$ and by approximating $\bar{\zeta} \cong r / N$ we have $\sum_{i \in R} p_i^{(r)} [1 - r / (N \tau_i \rho_i)] = 0$. We refer to the last equation as the *r*-constraint. We found in our empirical study that imposing this constraint "as is" causes numerical difficulties, possibly due to the fact that the optimization solution is not necessarily an internal point of the feasible domain (a familiar problem in EL maximization under constraints). However, adding a random noise solves the problem. This was done by rewriting the *r*-constraint as $\sum_{i \in R} p_i [\tau_i \rho_i - (r / N) + \delta_i] = 0$, $\delta_i \sim \mathrm{Unifom}(-a, a)$ with $a > 0$ a small number. Since $E(\delta_i) = 0$, this modification is valid.

<u>*Response model*</u>: The response probabilities, ρ_i, are unknown and need to be estimated. In order to account for possible NMAR nonresponse, we model them as a function of the outcome and the covariates. Specifically, we assume $\rho_i(\gamma) = \Pr(R_i = 1 | y_i, \mathrm{v}_i; \gamma) = \mathrm{logit}^{-1}[l(y_i, \mathrm{v}_i; \gamma)]$ where $\mathrm{logit}^{-1}(s) = (1 + e^{-s})^{-1}$

and $l(y_i, \mathrm{v}_i; \gamma)$ is a polynomial in (y, x) with coefficients γ. A function of the form $\mathrm{logit}^{-1}[l(y_i, \mathrm{v}_i; \gamma)]$ can approximate the true response function arbitrarily close.

Population model: We assume that the target population model has the general form $E(y_j \mid \mathrm{x}_j; \beta) = m(\mathrm{x}_j; \beta)$, where $m(\mathrm{x}_j; \beta)$ has a known form and the covariates are taken as random. Under some regularity conditions, the estimate of the vector parameter β is the unique solution of the equations $E_p\{[\partial m(\mathrm{x}; \beta) / \partial \beta][y - m(\mathrm{x}; \beta)]\} = 0$.

Estimation: Incorporating the response model into the R-level EL, results in the following maximization problem where we denote $p^{(r)}(\gamma) = [p_1^{(r)}(\gamma), ..., p_r^{(r)}(\gamma)]$:

$$\max_{\gamma, p^{(r)}(\gamma)} \prod_{i=1}^{r} p_i^{(r)}(\gamma), \text{ s.t. } A(\gamma) p^{(r)}(\gamma) = 0, \qquad (2)$$

where the rows of the matrix $A(\gamma)$ consist of $\tau_i^{-1} \rho_i^{-1} (c_i - \bar{c}_U)$ and $1 - r(N \tau_i \rho_i)^{-1}$.

The profile likelihood: The maximization in (2) is equivalent to $\max_\gamma G(\gamma)$ with the same constraints where $G(\gamma) = \max_{p^{(r)}(\gamma)} \prod_{i=1}^{r} p_i^{(r)}(\gamma)$. Therefore, for given γ, the computation of $G(\gamma)$ reduces to maximization of the type $\max\{\prod_{i=1}^{r} h_i : h_i > 0, \sum_{i=1}^{r} h_i = 1, A(h_1, ..., h_r)' = 0\}$. We solve this maximization problem by adapting the *S* function **elm**, written by A. Owen to the *R* software, with some modifications. Note that $G(\gamma)$ is the *profile likelihood* of γ.

Maximum likelihood estimation of γ: Finding $\arg\max_\gamma G(\gamma)$ is done by using the numerical optimization routine **optim** in R. The initial point for the maximization can be found by a grid search.

Parametric estimation of $Var(\hat{\gamma})$: Whilst a profile likelihood is not, properly speaking a likelihood, under general conditions one can still estimate the variance using the Hessian $H(\gamma) = (\partial^2 / \partial \gamma^2) G(\gamma)$ and estimate $Var(\hat{\gamma}) = -H^{-1}(\gamma)$.

Estimation of the population parameter $p = (p_1, ..., p_r)$: Once we obtain $\hat{\gamma}$, $\hat{p}^{(r)} = \arg\max_{p^{(r)}(\hat{\gamma})} \prod_{i=1}^{r} p_i^{(r)}(\hat{\gamma})$, s.t. $A(\hat{\gamma}) p^{(r)}(\hat{\gamma}) = 0$. Since $p_i \propto \tau_i \rho_i p_i^{(r)}$,

$\hat{p}_i = \hat{p}_i^{(r)} [\tau_i \rho_i(\hat{\gamma})]^{-1} / \sum_{j=1}^{r} \hat{p}_j^{(r)} [\tau_j \rho_j(\hat{\gamma})]^{-1}$.

Estimation of target parameter β: for given $\hat{p}$, $\hat{\beta}$ is the unique solution of the equations $\sum_{i=1}^{r} \hat{p}_i \{[\partial m(\mathrm{x}_i; \beta) / \partial \beta][y_i - m(\mathrm{x}_i; \beta)]\} = 0$.

Parametric estimation of $Var(\hat{\beta})$: $Var(\hat{\beta}) = Var[E(\hat{\beta} \mid \hat{\gamma})] + E[Var(\hat{\beta} \mid \hat{\gamma})]$. The 1st term can be estimated by drawing at random K vectors $\gamma_1, ..., \gamma_K$ from $N[\hat{\gamma}, Var(\hat{\gamma})]$ and computing, $\hat{Var}[E(\hat{\beta} \mid \hat{\gamma})]$

$= \frac{1}{K-1} \sum_{k=1}^{K} [E(\hat{\beta} \mid \gamma_k) - \bar{E}(\hat{\beta} \mid \gamma)][E(\hat{\beta} \mid \gamma_k) - \bar{E}(\hat{\beta} \mid \gamma)]'$; $\bar{E}(\hat{\beta} \mid \gamma)] = K^{-1} \sum_{k=1}^{K} E(\hat{\beta} \mid \gamma_k)$.

The 2nd term can be estimated by the "sandwich estimator" (Owen, 2001, pp.55-56).

4. Simulation results

In order to test the performance of our proposed approach we conducted a small simulation study as follows: A population of values $x_j, j = 1,...,10000$ was generated from $gamma(2,2)$ and truncated at 3. For each x_j, a binary response y_j was generated as $\Pr(y_j = 1 | \mathrm{x}_j; \beta) = \mathrm{logit}^{-1}(-0.8 + 0.8x_j)$. Next, a value of a design variable Z was generated as $z_j = \max[(x_j + 1.1)(2y_j + 1) + v_j, 0.01]$; $v_j \sim \mathrm{Uniform}(-0.2, 0.2)$. Values of calibration variables c were generated as $c_j = (1, x_j, y_j, x_j y_j, x_j^2, x_j^2 y_j)' + \varepsilon_j$; $\varepsilon_j \sim MN(0, \mathrm{I})$. A sample was selected via Bernoulli sampling, $I_j \sim Ber(\pi_j)$, where $\pi_j = \min(3500 z_j^{-1} / \sum_{k=1}^{10000} z_k^{-1}, 0.9999)$. The sampled units were classified as respondents with probabilities $R_i \sim (\rho_i)$, where $\rho_i = \mathrm{logit}^{-1}(\gamma_0 + \gamma_x x_i + \gamma_y y_i)$, with $\gamma_0 = 0.7, \gamma_x = 0.5, \gamma_y = 1.5$. The process of generating the population values and selecting the sample of respondents was repeated independently 100 times. (The x-values were generated only once). We use kernel smoothing to obtain estimates of $E_S(w_i | y_i; x_i) = E_R(w_i | y_i; x_i)$ by applying kernel regression of w_i on (y_i, x_i) and their interaction using the R function **npreg** from the **np** package at its default setting. For each sample of respondents we estimate (γ, β) and the variance of the estimators using the procedures described above. We also applied the Bootstrap (BS) method for estimating the variances of the estimators of the $\beta-$coefficients. This was done by sampling $B = 40$ simple random samples with replacement from each of the 100 primary samples and estimating the coefficients from the BS samples. We are running more BS samples at present.

Table 1. Mean Estimates, Standard Errors and SQRT of Mean of Variance estimates, β-coefficients. $\beta_0 = -0.8, \beta_1 = 0.8$. Mean sample size 3622.12; mean number of respondents, 2438.4.

	Mean Est.		**Empirical SE**		**Par. SE Est.**		**BS SE est.**	
Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
FR UW	-1.91	0.81	0.07	0.07			0.07	0.07
FR PW	-0.80	0.80	0.07	0.07			0.07	0.07
FR EL	-0.80	0.80	0.07	0.07			0.07	0.07
MAR UW	-2.66	0.97	0.12	0.10			0.11	0.09
MAR PW	-1.55	0.96	0.12	0.10			0.11	0.10
CCREL	-0.77	0.79	0.17	0.10	0.18	0.09	0.19	0.11

FR= Full response, estimators obtained from all sample data; **UW**= Unweighted; **PW**= Probability weighted; **MAR**= estimators obtained when ignoring response mechanism. **CCREL**= proposed method: Constrained conditional respondents' EL.

Table 2. Mean Estimates, Standard Errors and SQRT of Mean of Variance estimates, γ-coefficients. $\gamma_0 = 0.7, \gamma_x = 0.5, \gamma_y = 1.5$.

Mean Est.			**Empirical SE**			**Par. SE Est.**		
$\hat{\gamma}_0$	$\hat{\gamma}_x$	$\hat{\gamma}_y$	$\hat{\gamma}_0$	$\hat{\gamma}_x$	$\hat{\gamma}_y$	$\hat{\gamma}_0$	$\hat{\gamma}_x$	$\hat{\gamma}_y$
0.74	0.53	-1.55	0.21	0.20	0.32	0.21	0.19	0.34

5. Conclusions

The proposed approach is numerically simpler and much more stable than fully parametric alternatives. The results from our simulation study demonstrate the good properties of the method for sufficiently large number of respondents

References

Chaudhuri, S., Handcock, M.S. and Rendall, M.S. (2010). A conditional empirical likelihood approach to combine sampling design and population level information. Technical report No. 3/2010, National University of Singapore, Singapore, 117546.

Chen, S.X. and Van Keilegom, I. (2009). A review on empirical likelihood methods for regression. *Test*, **18**, 415-447.

Hartley, H.O. and Rao, J.N.K. (1968). A new estimation theory for sample surveys . *Biometrika*, **55**, 547-557.

Pfeffermann, D. (2011). Modeling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, 37, 115-136.

Pfeffermann, D., Krieger, A. M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8, 1087-1114.

Pfeffermann, D. and Sikov N. (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *Journal of Official Statistics*, 27, 181–209.

Pfeffermann, D. and Sverchkov, M. (2003). Fitting generalized linear models under informative probability sampling. In *Analysis of Survey Data* (Eds., R. L. Chambers and C. J. Skinner), New York: Wiley, 175-195.

Pfeffermann, D. and Sverchkov, M. (2009). Inference under Informative Sampling. In *Handbook of Statistics* 29B; *Sample Surveys: Inference and Analysis* (Eds., D. Pfeffermann and C.R. Rao), Amsterdam: North Holland, 455-487.

Qin, J., Leung, D., and Shao, J. (2002). Estimation with Survey data under nonignorable nonresponse or informative sampling. *Journal of the American Statistical Association*, 97, 193-200.

Qin, J., Shao, J., and Zhang, B. (2008). Efficient and doubly robust imputation for covariate-dependent missing response. *Journal of the American Statistical Association*, 103, 797-810.

Owen, A. (1988), Empirical Likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika*, 75, 237-249.

Owen, A. (1990). Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, **18**, 90-120.

Owen, A. (1991), Empirical Likelihood for Linear Models. *The Annals of Statistics*, 19, 1725–1747.

Owen, A. (2001), Empirical Likelihood. Boca Raton:Chapman & Hall/CRC.

Qin, J., and Lawless, J. (1994), Empirical Likelihood and General Estimating Equations. *The Annals of Statistics*, 22, 300–325.