updated on 2015.06.01

Statistical Inference Under Non-ignorable Sampling and Non-response— An Empirical Likelihood Approach

Moshe Feder[1] and Danny Pfeffermann[1,2,3]

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 highly biased inference. Accounting for sample selection bias is relatively simple because the sample selection probabilities are usually known, and several approaches have been proposed in the literature for dealing with this problem. Accounting for nonignorable non-response is much harder since the response probabilities are generally unknown, requiring to assume some structure in the response mechanism. In this article, we develop a new approach for modelling complex survey data, which accounts simultaneously for non-ignorable sampling and non-response. Our approach combines the nonparametric empirical likelihood with a parametric model for the response probabilities, which contains the outcome variable as one of the covariates. The proposed approach is also applicable in principle when the probability of inclusion is unknown, such as data from voluntary Web-based surveys.

We illustrate the robustness of the proposed approach and propose ways of testing the underlying model. Combining the population model with the model for the response probabilities defines the model holding for the missing data and enables imputing the missing sample data from this model. Simulation results illustrate the good performance of the approach in terms of parameter estimation and imputation.

Keywords: Empirical likelihood, Kernel smoothing, Model testing, NMAR non-response, Respondents model, Population model, 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. Familiar examples include the estimation of elasticity of demand from family expenditure surveys, estimation of health risk factors from health surveys and the analysis of labor market dynamics from labor force surveys. 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 model holding for the population values. This will happen, for example, when the selection probabilities are determined by one or more design

[1]Southampton statistical Sciences Research Institute, University of Southampton, UK
[2]Department of Statistics, Hebrew University of Jerusalem, Israel
[3]Central Bureau of Statistics, Israel

variables (stratification variables, size variables used for probability proportional to size sampling, etc.), which are correlated with the model outcome variable, but at least some of them are not included among the model covariates. In an extreme case, the sample selection probabilities are determined directly by the outcome values, as in case-control studies. 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 since 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 corresponding covariate values, 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 (but not necessarily) also the outcome variable. The totals of such calibration variables are often available from administrative or census records. Our proposed approach combines the non-parametric empirical likelihood (EL) under the population model with a parametric model for the response probabilities that contains the outcome variable as one of the covariates. A third component needed for setting the likelihood holding for the responding units is the expectation of the sampling weights given the outcome and the covariates, which we estimate non-parametrically using kernel smoothing. The use of EL for analysing complex survey data has its origins in a landmark paper by Hartley and Rao (1968), and has gained increasing interest in recent years in more general statistical contexts following Owen (1988, 1990, 1991, 2001). Another important paper is Qin and Lawless (1994). The EL combines the robustness of non-parametric 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. As the proposed method is a based on the empirical likelihood, conditional on response, we'll refer to it as 'Conditional Respondents Empirical Likelihood' (CREL).

Chang and Kott (2008) proposed a calibration-based approach where the sampling weights $w_i = 1/\pi_i$ (where π_i is the probability of selecting unit i) are replaced by $w_i^* = w_i g(\tilde{\boldsymbol{\gamma}}' \boldsymbol{z}_i)$, where g is a known and everywhere

monotonic and twice-differentiable function, $\tilde{\gamma}$ is a vector parameter chosen in a manner that minimises the difference $\left[\sum_{i\in\mathcal{U}} \boldsymbol{x}_i - \sum_{i\in\mathcal{R}} w_i^* \boldsymbol{x}_i\right]$, in a weighted least squares sense. Here, $\mathcal{U}$ and $\mathcal{R}$ are the population and the set of respondents, respectively, and $\boldsymbol{x}_i$ is assumed known for each respondent, as well as its population total.

In the next section we define the sample and respondents distribution and in Section 3 we define the empirical likelihood and its components, given the observed data for the responding units. Section 4 provides some details on the maximization of the empirical likelihood and in Section 5 we discuss ways of testing the model. Section 6 reports the results of a simulation study aimed to illustrate the performance of the method. Section 7 contains concluding remarks.

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 $\mathcal{S}$, drawn from a finite population $U = \{1, \ldots, N\}$ with known inclusion probabilities $\pi_i = \Pr(i \in \mathcal{S})$. Let I_i denote the sampling indicator defined as 1 if unit i is sampled, 0 otherwise, and $\boldsymbol{x}_i = (x_{1,i}, \ldots, x_{p,i})'$ denote the values of p auxiliary variables (covariates) associated with unit i. Let $\mathcal{R}$ be the set of respondents and let the response indicator R_i be defined as 1 if unit $i \in \mathcal{S}$ responds, 0 otherwise. We denote by n the size of $\mathcal{S}$ and by r the size of $\mathcal{R}$.

In what follows we assume that the population outcomes are independent realizations from distributions with probability density functions (PDF) $f_u(y_i|\boldsymbol{x}_i)$. Following Pfeffermann *et al.* (1998), the marginal sample PDF, $f_s(y_i|x_i)$ denotes the conditional PDF of y_i given that the unit is sampled, i.e., $f_s(y_i|\boldsymbol{x}_i) = f(y_i|\boldsymbol{x}_i, I_i = 1)$. By Bayes rule,

$$f_s(y_i|\boldsymbol{x}_i) = \frac{\Pr(I_i|\boldsymbol{x}_i, y_i) f_u(y_i|\boldsymbol{x}_i)}{\Pr(I_i|\boldsymbol{x}_i)} 2.0$$

where $\Pr(I_i = 1|\boldsymbol{x}_i) = \int \Pr(I_i = 1|\boldsymbol{x}_i, y_i) f_u(y_i|\boldsymbol{x}_i) dy_i$. Note that $\Pr(I_i = 1|\boldsymbol{x}_i, y_i)$ is generally not the same as the sample selection probability $\pi_i = \Pr(i \in s) = \Pr(I_i = 1|Z_u)$, where Z_u defines a matrix of population values of design variables used for the sample selection. Since $\Pr(I_i = 1|\pi_i, y_i, \boldsymbol{x}_i) = \pi_i$, $\Pr(I_i = 1|y_i, \boldsymbol{x}_i) = E_u(\pi_i|y_i, \boldsymbol{x}_i)$, where E_u is the expectation under the population PDF. The population and sample PDFs differ unless $\Pr(I_i = 1|\boldsymbol{x}_i, y_i) = \Pr(I_i = 1|\boldsymbol{x}_i)$ for all y_i, and when this condition is not met the sampling design is informative and cannot be ignored in the inference process. In particular, it follows from (2.1) that under informative sampling

$$E_s(y_i|\boldsymbol{x}_i) = E_u\left[\frac{\Pr(I_i = 1|\boldsymbol{x}_i, y_i)}{\Pr(I_i = 1|\boldsymbol{x}_i)} y_i \middle| \boldsymbol{x}_i\right] \neq E_u(y_i|\boldsymbol{x}_i) \tag{2.2}$$

where E_s denotes expectation with respect to the sample distribution. Estimating $E_u(y_i|\boldsymbol{x}_i)$ is often the main target of inference. Thus, ignoring an informative sampling scheme, and in fact estimating $E_s(y_i|\boldsymbol{x}_i)$, can severly bias the inference. Next, consider the respondents distribution. The response probabilities may depend on covariates $\boldsymbol{v}$, which may differ from $\boldsymbol{x}$ in one or more components. The marginal PDF for responding unit i, denoted by $f_r(y_i|\boldsymbol{x}_i) = f(y_i|\boldsymbol{x}_i, I_i = 1, R_i = 1)$ is, by Bayes Rule,

$$f_r(y_i|\boldsymbol{x}_i, \boldsymbol{v}_i) = \frac{\Pr(R_i = 1|y_i, \boldsymbol{v}_i, I_i = 1) f_s(y_i|\boldsymbol{x}_i)}{\Pr(R_i = 1|\boldsymbol{v}_i, \boldsymbol{x}_i, I_i = 1)}. \tag{2.3}$$

Note that unless $\Pr(R_i = 1|y_i, \boldsymbol{v}_i, I_i = 1) = \Pr(R_i = 1|\boldsymbol{v}_i, I_i = 1)$ for all i, the respondents PDF differs from the sample PDF.

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

$$L_r(\gamma, \theta) = \prod_{i=1}^{r} \frac{\Pr(R_i = 1|y_i, \boldsymbol{v}_i, I_i = 1; \gamma)\Pr(I_i = 1|y_i, \boldsymbol{x}_i) f_u(y_i|\boldsymbol{x}_i; \theta)}{\Pr(R_i = 1|\boldsymbol{x}_i, \boldsymbol{v}_i, I_i = 1; \theta, \gamma)\Pr(I_i|\boldsymbol{x}_i)} \tag{2.4}$$

In principle, one could maximize the likelihood (2.4) with respect to θ and γ. But our experience shows that this can be very complicated numerically and result in unstable estimates, depending on the population model and the model assumed for the response probabilities. For this reason, we consider in the next section the use of the empirical likelihood paradigm, which enables estimating the parameters $\boldsymbol{\gamma}$ governing the response model without specifying the population model, and is therefore more robust. On the other hand, as discussed in the next section, the parameters underlying an assumed population model can be estimated very easily once the parameters of probabilities underlying the empirical likelihood have been estimated. In this respect the use of the empirical likelihood can be viewed as a convenient way of estimating the parameters of an assumed population model.

REMARK 1. In theory, one also needs to model the probabilities $\Pr(I_i = 1|y_i, \boldsymbol{x}_i)$. However, since $\Pr(I_i = 1|\pi_i, y_i, \boldsymbol{x}_i) = \pi_i$, the probability $\Pr(I_i = 1|y_i, \boldsymbol{x}_i)$ can be estimated outside the likelihood using the relationship $\Pr(I_i = 1|y_i, \boldsymbol{x}_i) = E_u(\pi_i|y_i, \boldsymbol{x}_i) = 1/E_s(w_i|y_i, \boldsymbol{x}_i)$, where $w_i = 1/\pi_i$ is the sampling weight. Thus, the probabilities $\Pr(I_i = 1|y_i, \boldsymbol{x}_i)$ can be estimated by regressing w_i on $(y_i, \boldsymbol{x}_i)$ using the sample data. See Pfeffermann and Sverchkov (2003, 2009) for different approaches to, and examples of, modeling and estimating the expectations $E_s(w_i|y_i, \boldsymbol{x}_i)$. Alternatively, as explained in Section 4 and illustrated in Section 6, the expectations can be estimated nonparametrically using kernel smoothing or other smoothing methods.

REMARK 2. A notable property of the likelihood (2.4) is that it does not require knowledge of the covariates of the nonresponding units. On the other hand, even with good estimates of the probabilities $\Pr(I_i = 1|y_i, \boldsymbol{x}_i)$, the use of this likelihood requires specifying the population model $f_u(y_i|\boldsymbol{x}_i)$, and the response probabilities $\Pr(R_i = 1|y_i, \boldsymbol{v}_i, I_i = 1)$, with no observations obtained directly from either one of the two distributions. Pfeffermann and Landsman (2011) establish conditions under which the likelihood (2.4) is identifiable, but our experience shows that even under these conditions, maximization of the likelihood is often unstable, due to what Lee and Berger (2001) refer to as 'practical non-identifiability.' See Rotnitzky and Robins (1997) for further discussion and theoretical results on the identifiability of likelihoods of the form (2.4).

the remark below added per your comment in the next section

REMARK 3. Whilst we have assumed that the response probabilities are functions of $\boldsymbol{nu}_i$, i.e., $\Pr(R_i = 1|y_i, \boldsymbol{v}_i, I_i = 1)$, we can assume without loss of generality that ν_i is embedded in $\boldsymbol{x}_i$. Hence we'll assume $\rho_i = \Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 1)$.

REMARK 4. Although no observations are available from either the model defining the population PDF or the model assumed for the response probabilities, the respondents model defined by (2.3) can nonetheless be tested using classical test statistics since it relates to the data observed for the responding units. See

Sections 5 and 6 for the test statistic used in our empirical study with illustrations. In the next section we propose an empirical likelihood approach which does not require specifying the population model, thus making the inference more robust.

3 Conditional empirical likelihood for responding units

We assume that for each unit i there corresponds a vector $\boldsymbol{u}_i = (y_i, \boldsymbol{x}_i', \boldsymbol{c}_i', \tau_i, \rho_i)'$ where y_i and $\boldsymbol{x}_i$ are related via a model $f_u(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})$, $\boldsymbol{c}_i$ is a d-dimensional vector of survey values for which the population means $\bar{\boldsymbol{c}}_u$ are known, $\tau_i = E_u(I_i|y_i, \boldsymbol{x}_i)$, and where $\rho_i = \Pr(R_i = 1|y_i, \boldsymbol{x}_i, I_i = 1) = E_u(R_i|y_i, \boldsymbol{x}_i, I_i = 1)$. We employ the load-scale approach of Hartley and Rao (1968) by assuming that the finite population values are generated from a multinomial distribution with a vector of probabilities $\boldsymbol{p} = (p_1, \ldots, p_r)'$. Thus the population distribution has its support in the sample of respondents. Denote by N_i the number of units in the finite population assuming the vector $\boldsymbol{u}_i$ and let $p_i = N_i/N$, where $N = \sum_i N_i$. Under this model, the distribution of the observed data for the responding units (hereafter the respondents' distribution) is also multinomial, with cell probabilities given by $p_i^{(r)} = \Pr(\boldsymbol{u}_i|i \in R) = p_i\tau_i\rho_i/\sum_k p_k\tau_k\rho_k$. Thus, the empirical likelihood is

$$\mathcal{L} = \prod_i p_i^{(r)} = \Pi(\boldsymbol{p}^{(r)}), \tag{3.1}$$

where $\Pi(\boldsymbol{a}) = \prod_i a_i$ denoted the product of the elements of any vector $\boldsymbol{a}$.

Chaudhuri *et al.* (2010) use a similar approach, though restricted to the case of full response (*viz.*, $\rho_i = 1$ for all i). The response probabilities ρ_i in (3.1) are unknown and need to be estimated. In order to account for possible NMAR nonresponse, we model ρ_i as a function of the outcome and covariates. Specifically, we assume $\rho_i(\boldsymbol{\gamma}) = \Pr(R_i = 1|y_i, \boldsymbol{x}_i;\boldsymbol{\gamma}) = \text{logit}^{-1}(\ell(y_i, \boldsymbol{x}_i;\boldsymbol{\gamma})$ where $\text{logit}^{-1}(s) = (1 + e^{-s})^{-1}$ and $\ell(y_i, \boldsymbol{x}_i;\boldsymbol{\gamma})$ is a polynomial in $(y, \boldsymbol{x})$ with coefficients $\boldsymbol{\gamma}$. A function of the form $\text{logit}^{-1}(\ell(y_i, \boldsymbol{x}_i;\boldsymbol{\gamma}))$ can approximate any response function which is a continuous function of $\boldsymbol{x}, y$, arbitrarily close, provided its range is bounded away from 0 and 1 . It thus follows that our EL is a combination of the nonparametric population distribution (but see below), the expectation $\tau_i = E_u(I_i|y_i, \boldsymbol{x}_i)$ and a parametric model for the response probabilities. We mentioned in the introduction that the use of empirical likelihood facilitates the use of calibration constraints for enhancing the efficiency of the estimators. Under our set up and the assumption that the population distribution has its support in the respondents sample, $\sum_{i\in\mathcal{R}} p_i\boldsymbol{c}_i = N^{-1}\sum_{i\in\mathcal{R}} N_i\boldsymbol{c}_i = N^{-1}\sum_{j\in U}\boldsymbol{c_j} = \bar{\boldsymbol{c}}_u$, yielding the $\mathcal{R}$-level constraint

$$\sum_{i\in\mathcal{R}} p_i^{(r)}\tau_i^{-1}\rho_i^{-1}(\boldsymbol{c}_i - \bar{\boldsymbol{c}}_u) = \boldsymbol{0}. \tag{3.2}$$

Denote $\xi_i = \tau_i\rho_i$ and $\sum_{i\in\mathcal{R}} p_i\xi_i = \bar{\xi}_u$. Since $\tau_i = E(I_i|y_i, \boldsymbol{x}_i)$ and $\rho_i = E(R_i|y_i, \boldsymbol{x}_i)$, ξ_i is the probability of a unit i being sampled and subsequently responding, given its outcome and covariate values. Since ρ_i depends on $\boldsymbol{\gamma}$, we'll also use the notation $\rho_i(\boldsymbol{\gamma})$ and $\xi_i(\boldsymbol{\gamma})$. Recall that ${p_i}^{(r)} \propto p_i\tau_i\rho_i = p_i\xi_i$. Thus, ${p_i}^{(r)} = \bar{\xi}_u^{-1}p_i\xi_i$. Denote by E_η the expectation with respect to the combined sampling and response distributions. Then $E_\eta(r) = \sum_{j\in U}\tau_j\rho_j = N\sum_{i\in\mathcal{R}} p_i\tau_i\rho_i = N\sum_{i\in\mathcal{R}} p_i\xi_i = N\bar{\xi}_u$. Thus, $r \approx N\bar{\xi}_u$, leading to the constraint

$r = N\bar{\xi}_u$. Since $\sum_{i\in\mathcal{R}} p_i = 1$, we have $1 = r(N\bar{\xi})^{-1} = r(N\bar{\xi})^{-1}\sum_{i\in\mathcal{R}} p_i = (r/N)\sum_{i\in\mathcal{R}} {p_i}^{(r)}\xi_i^{-1}$, or

$$\sum_{i\in\mathcal{R}} p_i^{(r)}\left(1 - r/(N\tau_i\rho_i)\right) = \sum_{i\in\mathcal{R}} p_i^{(r)}\left(1 - r/(N\xi_i)\right) = 0. \tag{3.3}$$

Note that this constraint is equivalent to $\sum_{i\in\mathcal{R}} p_i^{(r)}\tau_i^{-1}\rho_i^{-1} = N/r$.

Let C be the $r \times d$ matrix, the ith row of which is $\boldsymbol{c}_i - \bar{\boldsymbol{c}}_u$, and let $D(\boldsymbol{\gamma})$ be the $r \times r$ diagonal matrix with $\tau_i\rho_i = \tau_i\rho(y_i, \boldsymbol{x}_i; \boldsymbol{\gamma})$ as its diagonal elements.

Constraints (3.2) and (3.3) can be written in a matrix form as $C'D(\boldsymbol{\gamma})^{-1}\boldsymbol{q} = \mathbf{0}$ and $\mathbf{1}'D^{-1}(\boldsymbol{\gamma})\boldsymbol{q} = N/r$, respectively, where we have denoted $\boldsymbol{q} = \boldsymbol{p}^{(r)}$ for simplicity.

We now have the constrained maximization problem

$$\max_{\boldsymbol{\gamma},\boldsymbol{q}} \Pi(\boldsymbol{q}) \quad \text{s.t.} \quad A(\boldsymbol{\gamma})\boldsymbol{q} = \begin{pmatrix}\mathbf{0}\\ 0\end{pmatrix}, \quad \boldsymbol{q} \in \Omega, \tag{3.4}$$

where $\boldsymbol{\xi}(\boldsymbol{\gamma})^{-1} = (\tau_1^{-1}\rho_1(\boldsymbol{\gamma})^{-1}, \ldots, \tau_r^{-1}\rho(\boldsymbol{\gamma})_r^{-1})'$, $A(\boldsymbol{\gamma}) = \begin{pmatrix} C'D(\boldsymbol{\gamma})^{-1} \\ (rN^{-1}\boldsymbol{\xi}(\boldsymbol{\gamma})^{-1} - 1)' \end{pmatrix}$, and where Ω is the simplex of all non-negative vectors $(a_1, \ldots, a_r)' \in R^r$ with $\sum_i a_i = 1$. The MLE of $\boldsymbol{\gamma}$ and $\boldsymbol{q}$ are the values where (3.4) attains its maximum. In principle, one could maximize (3.4) over all $(\boldsymbol{\gamma}, \boldsymbol{p})$, subject to the constraints. However, this may be impractical, due to the dimension of $\boldsymbol{p}$. Fortunately, there is a much more efficient maximization procedure, which we describe in Section 4.

Existence of a Solution. There are cases where a solution to (3.4) does not exist (and therefore the feasible domain is empty). A simple example is where all observed values of a constraining variable c are all greater (or smaller) than its known population mean. Moreover, multivariate conditions can also preclude a solution. However, if a solution $\boldsymbol{q} \in \Omega$ to (3.4) exists for a given $\boldsymbol{\gamma}$, then there exists a solution in Ω for *any* $\boldsymbol{\gamma}$. In fact, since $D(\boldsymbol{\gamma})$ is a diagonal matrix with a positive diagonal, there is such a solution if and only if there is a positive vector $\boldsymbol{v} \in R^r$ such that $C'\boldsymbol{v} = \mathbf{0}$. The univariate constraint (3.3) can lead to an empty feasible domain in (3.4) for some $\boldsymbol{\gamma}$. Such is the case, for example, if $\rho_i(\boldsymbol{\gamma})$ is very small. By definition, the supremum over an empty set of a real-valued function is $-\infty$. See also Sub-section 4.1.1.

We have assumed so far that the population distribution is multinomial with the unknown probabilities $\boldsymbol{p}$, which are estimated by maximization of the EL. Suppose, however, that the target population distribution is in fact parametric. In particular, consider the general population model

$$y_j = m(x_j; \boldsymbol{\beta}) + \varepsilon_j; \quad E_u(\varepsilon_j|\boldsymbol{x}_j) = 0, \quad E_u(\varepsilon_j^2|\boldsymbol{x}_j) = \sigma_\varepsilon^2, \quad E_u(\varepsilon_j\varepsilon_k|\boldsymbol{x}_j, \boldsymbol{x}_k) = 0, \quad (j \neq k), \tag{3.5}$$

where $m(\boldsymbol{x}_j; \boldsymbol{\beta})$ has a known form and the covariates $\boldsymbol{x}_j$ are random. Under some regularity conditions, the vector parameter $\boldsymbol{\beta}$ and σ_ε^2 are the unique solutions of the equations $E_u\left[(\partial m(\boldsymbol{x}; \boldsymbol{\beta})/\partial\boldsymbol{\beta})(y - m(\boldsymbol{x}; \boldsymbol{\beta}))\right] = \mathbf{0}$ and $\sigma_\varepsilon^2 = E_u[y - m(\boldsymbol{x}; \boldsymbol{\beta})]^2$. Hence $\boldsymbol{\beta}$ and σ_ε^2 satisfy

$$\sum_{i\in\mathcal{R}} p_i \frac{\partial m(\boldsymbol{x}_i; \boldsymbol{\beta})}{\partial\boldsymbol{\beta}}[y_i - m(\boldsymbol{x}_i; \boldsymbol{\beta})] = \mathbf{0} \quad \text{and} \quad \sigma_\varepsilon^2 = \sum_{i\in\mathcal{R}} p_i[y_i - m(\boldsymbol{x}_i; \boldsymbol{\beta})]^2. \tag{3.6}$$

Having estimated the probabilities p_i by maximization of the EL, the estimates of $\boldsymbol{\beta}$ and σ_ε^2 are obtained by solving the estimating equations (3.6) with $p_1, \ldots, p_r$ replaced by their estimates.

The importance of the constraints. The EL maximization problem (3.4) is subject to constraints. The question then arises as to how important these constraints are and how they should be best chosen. Suppose for simplicity that $\tau_i = \Pr(i \in s|y_i, \boldsymbol{x}_i) = \text{const}$ (noninformative sampling) in which case $p_i^{(r)}(\boldsymbol{\gamma}) \propto p_i\rho_i(\boldsymbol{\gamma})$. Then, for a given vector parameter $\boldsymbol{\gamma}$, $\max_{\boldsymbol{p}^{(r)}} \Pi(\boldsymbol{p}^{(r)}) = r^{-r}$. Thus, without any constraints the empirical likelihood is maximized by defining $p_i \propto \rho_i^{-1}$, implying non-identifiability in the estimation of $\boldsymbol{p}$ since this is true for every $\boldsymbol{\gamma}$.

Next is the question of how the survey variables defining the constraints should be chosen. To answer this question, we consider first the constraints underlying the method of Chang and Kott (2008), written in our notation $\bar{\boldsymbol{C}}_u = N^{-1}\sum_{i\in\mathcal{R}} w_i\boldsymbol{c}_i/\rho_i(\boldsymbol{\gamma})$. In this case it is obvious that the variables $\boldsymbol{c}$ should be as highly correlated as possible with y and $\boldsymbol{x}$ because otherwise they provide little or no information on the probabilities $\Pr(R_i = 1|y_i, \boldsymbol{x}_i; \boldsymbol{\gamma})$. Notice also in this respect that while $\hat{\bar{\boldsymbol{C}}}_{HT} = N^{-1}\sum_{i\in\mathcal{R}} w_i\boldsymbol{c}_i/\rho_i(\boldsymbol{\gamma})$ is randomization unbiased for $\bar{\boldsymbol{C}}_u$ over all possible sample selections and nonresponse, the variance of $\hat{\bar{\boldsymbol{C}}}_{HT}$ is minimized when $\boldsymbol{c}_i \propto \rho_i(\boldsymbol{\gamma})$. You've asked why, but I don't know and this came from your text.

In our proposed empirical likelihood approach the constraints (3.2) are of the form $\sum_{i\in\mathcal{R}} p_i\boldsymbol{c}_i = \bar{\boldsymbol{C}}_u$ which are seemingly unrelated to the response probabilities, suggesting that it should not matter which survey variables are used for defining the constraints. This, however, is a false conclusion since the empirical likelihood is defined with respect to the probabilities $p_i^{(r)}(\boldsymbol{\gamma}) = \frac{p_i\tau_i\rho_i(\boldsymbol{\gamma})}{\sum_k p_k\tau_k\rho_k(\boldsymbol{\gamma})}$, such that any constraint on the p_i's effectively defines a constraint on the ρ_i's, implying again that the variables in $\boldsymbol{c}$ should be correlated as highly as possible with y and $\boldsymbol{x}$. I moved the following from the paragraph that you deleted: The number of constraints is of lesser importance than their choice. We illustrate this in the empirical study in Section 6.

4 Estimation of the Model Parameters

4.1 Estimation of the τ's

In the simulation study described in Section 6, we used kernel smoothing to obtain estimates of $\tau_i = E_u(I_i|y_i, \boldsymbol{x}_i)$. Note that $1/\tau_i = E_s(w_i|y_i; x_i) = E_r(w_i|y_i; x_i)$ by applying kernel regression of w_i on (y_i, x_i) and their interaction, using the function `npreg` from the R package `np` at its default setting. See Subsection 6.2 for further details.

4.2 Point Estimation

4.2.1 Estimation of the Response Propensity Model

The profile likelihood of $\boldsymbol{\gamma}$: The maximization problem in (3.4) is equivalent to the maximization $\max_{\boldsymbol{\gamma}} G(\boldsymbol{\gamma})$ where $G(\boldsymbol{\gamma})$ is the *profile likelihood* of $\boldsymbol{\gamma}$, defined as

$$G(\boldsymbol{\gamma}) = \max\{\Pi(\boldsymbol{q}) : A(\boldsymbol{\gamma})\boldsymbol{q} = \boldsymbol{0} \quad \& \quad \boldsymbol{q} \in \Omega\}. \tag{4.1}$$

Danny - note that now $A(\boldsymbol{\gamma})$ is defined right after (3.4)

For a given $\boldsymbol{\gamma}$, the maximization in (4.1) can be done using R function `scel`, written by Art Owen and available from his website `http://statweb.stanford.edu/~owen/empirical`. See Owen (2013) for related

theory and further details.

Estimation of $\boldsymbol{\gamma}$: The MLE of $\boldsymbol{\gamma}$ is $\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} G(\boldsymbol{\gamma})$. In the simulation study described in Section 6, $\arg\max_{\boldsymbol{\gamma}} G(\boldsymbol{\gamma})$ was found by using the R numerical optimization function `optim`. (Since `optim` is a *minimization* routine, we minimized $-G(\boldsymbol{\gamma})$.) The initial point for the optimization was obtained as follows: an initial guess for γ_0 was set as $\text{logit}^{-1}(r/n)$ and the remaining coefficients were initiated by a grid search. NOTE: When the feasible domain in (3.4) for a certain $\boldsymbol{\gamma}$ is empty, $G(\boldsymbol{\gamma}) = -\infty$ by definition.

4.2.2 Estimation of the Population Model

Estimation of the population parameter $\boldsymbol{p} = (p_1, \ldots, p_r)'$: Once we have obtained $\hat{\boldsymbol{\gamma}}$, we can estimate $\hat{\boldsymbol{p}}^{(r)} = \arg\max_{\boldsymbol{q}} \Pi(\boldsymbol{q})$ s.t. $A(\hat{\boldsymbol{\gamma}})\boldsymbol{q} = \boldsymbol{0}, \boldsymbol{q} \in \Omega$. Since $p_i \propto (\tau_i \rho_i)^{-1} p_i^{-1}$, we can now calculate

$$\hat{p}_i = \hat{p}_i^{(r)} [\tau_i \rho_i(\hat{\boldsymbol{\gamma}})]^{-1} \Big/ \sum_{k=1}^{r} \hat{p}_k^{(r)} [\tau_k \rho_k(\hat{\boldsymbol{\gamma}})]^{-1} \tag{4.2}$$

Estimation of the population model's parameter $\boldsymbol{\beta}$: For given $\hat{\boldsymbol{p}}$, $\hat{\boldsymbol{\beta}}$ is the unique solution of the equation $\sum_{i=1}^{r} \hat{p}_i \{(\partial/\partial\boldsymbol{\beta})[y_i - m((\boldsymbol{x}_i; \boldsymbol{\beta})]\} = \boldsymbol{0}$. Any statistical package capable of fitting generalized linear models with survey weights can be used to solve these equations, with $(\hat{p}_1, \ldots, \hat{p}_r)'$ entered as 'weights.' In our simulations, the R package 'survey' was used (Lumley 2004).

Remark: It is not required here to assume the population model holds at all. The parameter $\boldsymbol{\beta}$ can be regarded as merely a population parameter, defined as the solution to the equation $\sum_{i=1}^{r} p_i \{(\partial/\partial\boldsymbol{\beta})[y_i - m((\boldsymbol{x}_i; \boldsymbol{\beta})]\} = \boldsymbol{0}$. For example, in the case of linear regression, $\boldsymbol{\beta} = (X_u' X_u)^{-1} X_u' \boldsymbol{y}_u$ where X_u is the population design matrix, $\boldsymbol{y}_u$ is the population vector of y-values. In this case, $\hat{\boldsymbol{\beta}} = (X' D_{\hat{\boldsymbol{p}}} X)^{-1} X' D_{\hat{\boldsymbol{p}}} \boldsymbol{y}$ where X and $\boldsymbol{y}$ are the corresponding values in the respondents' data and $D_{\hat{\boldsymbol{p}}} = \text{diag}(\hat{p}_1, \ldots, \hat{p}_r)$. In Sub-section 7.4, simulation results are given for this estimate when the true model is non-linear.

Non-parametric estimation of the population model $f_p(y|\boldsymbol{x})$: The proposed approach does not require any specification of a model for $f_p(y|\boldsymbol{x})$. In fact, once estimates $\hat{p}_i$ are obtained (see 4.2 above), and thus an estimate $\hat{F}$ of the population distribution is available, non-parametric estimation of $f_p(y|\boldsymbol{x})$ can be made, for example using smooth polynomial spline. In Sub-section 7.4, results using a smooth cubic spline are given.

4.3 Estimation of Variances

4.3.1 Estimation of $\text{Var}(\boldsymbol{\gamma})$

Whilst a profile likelihood is not, strictly speaking, a likelihood, under general conditions one can still estimate the variance by the inverse profile information $I_{pr}(\boldsymbol{\gamma}) = -(\partial^2/\partial\boldsymbol{\gamma}^2) G(\boldsymbol{\gamma})$ and setting $\widehat{\text{Var}}(\hat{\boldsymbol{\gamma}}) = I_{pr}(\hat{\boldsymbol{\gamma}})^{-1}$. Murphy and van der Vaart (2000) discuss the appropriateness of treating a profile likelihood as proper

likelihood for the purpose of variance estimation. Scott and Wild (2006) show the validity of estimating the covariance of $\hat{\boldsymbol{\gamma}}$ by I_{pr}^{-1} more generally. Below is an outline of their argument.

Let $\ell(\boldsymbol{\gamma}, \boldsymbol{q})$ denote the log-likelihood of the parameters $\boldsymbol{\psi} = (\boldsymbol{\gamma}', \boldsymbol{q}')'$ given the data, then the profile likelihood of $\boldsymbol{\gamma}$ is $\ell_{pr}(\boldsymbol{\gamma}) = \ell(\boldsymbol{\gamma}, \hat{\boldsymbol{q}}(\boldsymbol{\gamma}))$, where $\hat{\boldsymbol{q}}(\boldsymbol{\gamma}) = \arg\max_{\boldsymbol{q}} \ell(\boldsymbol{\gamma}, \boldsymbol{q})$. In principle, the covariance matrix of $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\gamma}}', \hat{\boldsymbol{q}}')'$ could be estimated by I^{-1} where I is the inverse information of $\boldsymbol{\psi}$. Partitioning I into a block matrix with blocks corresponding to $\boldsymbol{\gamma}$ and $\boldsymbol{q}$ and applying the formula for the inverse of a partitioned matrix, the upper left block of I^{-1} equals I_{pr}^{-1}. However, that upper left block is also the estimate of the covariance of $\hat{\boldsymbol{\gamma}}$. (Cf. Scott and Wild 2006 for further details.)

In our simulation study, the use of the inverse profile information gave very good variance estimates of the response model parameter $\boldsymbol{\gamma}$. See more details in Section 6.

4.3.2 Parametric Bootstrap Variance Estimation

In general, the parametric bootstrap approach consists of generating B samples, with each sample consisting of r units independently drawn from the fitted $f_r(\boldsymbol{u})$ distribution, where $\boldsymbol{u}$ stands for all the variables involved. In the case of the empirical likelihood, the fitted f_r distribution is multinomial, with $\boldsymbol{p}^{(r)} = (p_1^{(r)}, \ldots, p_r^{(r)})'$ as its parameter. Therefore, we independently sample r units, with replacement, from $\mathcal{R}$, such that in each of the r draws, the probability of the selected unit being i is $\hat{p}_i{}^{(r)}$. The estimation procedure is applied to the data from each sample. Denote the B estimates of a parameter θ by $\hat{\theta}_1, \ldots, \hat{\theta}_B$. The parametric bootstrap estimate of the variance of $\hat{\theta}$ is $B^{-1} \sum_{b=1}^{B} (\hat{\theta}_b - \bar{\theta})^2$ where $\bar{\theta} = B^{-1} \sum_{b=1}^{B} \hat{\theta}_b$.

NOTE: Whilst the same estimation procedure carried out on the main sample data needs to be applied to each of the B bootstrap samples, two time-saving measures can be used: (a) Starting estimate for $\boldsymbol{\gamma}$ may be taken as the main sample estimate. (b) Additionally, the main sample estimates of τ_i can be used.

Dealing with outlying bootstrap estimates: Bootstrap variance estimates can be too conservative (positively biased) due to outliers (Shao, 1988). This can be remedied by excluding a small portion of the bootstrap samples that appear to be outliers. In our simulation study, we have classified a bootstrap sample as outlier if its estimated $\hat{\boldsymbol{\gamma}}_b$ had any component more than 4SE away from the corresponding main sample estimate, where SE is the estimated standard error calculated as described in Section 4.3.1. We found that on average less than 3% of the bootstrap samples were excluded. The outlying bootstrap samples were excluded from the variance estimation for both $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\beta}}$.

4.4 Dealing With Unknown Probability of Inclusion

There are increasing instances where the probability of inclusion is unknown, such as in voluntary Web-based surveys. The proposed approach is applicable in this case as well, provided the probability of ending up in the sample is positive and provided the population means $\bar{\boldsymbol{c}}$ and respondents' values $\boldsymbol{c}_i$ for some control variables $\boldsymbol{c}$ are known. Assume first that there are no multiple responses coming from the same unit. For a unit i in the population, let ξ_i be the probability that the unit ends up in the sample. Assume a model of the form $\log(\xi_i/(1-\xi_i)) = \boldsymbol{\beta}'\boldsymbol{u}$, where $\boldsymbol{u}$ are survey variables which may include the population model's dependent and independent variables. Note that this situation is mathematically similar to the case where the sampling is non-informative and the response probability of unit i is ξ_i. Therefore, we can take $\tau_i = \text{const}$

(any constant), treat ξ_i as response probability, and proceed as in the case of known sampling probabilities and informative non-response.

It is often the case in Web-based surveys that individuals respond more than once. Moreover, such respondents may do so deliberately, in order to influence the survey results, and without disclosing their multiple responses. This is a challenging problem because those respondents may use different computers, or even the same computer with a different IP address. A possible approach to dealing with this problem is to model the number of times m_i an individual i responds, for example, a log-linear model of the form $\log(m_i) = \boldsymbol{\eta}'\boldsymbol{v}_i$ may be used. It should be noted, however, that this approach is feasible only as long as the population means $\bar{\boldsymbol{c}}$ and respondents' values $\boldsymbol{c}_i$ are available and truthfully reported by the respondents.

5 Model testing

A crucial question regarding any statistical procedure is whether it can be tested. Contrary to the common perception that it is impossible to test whether the nonresponse is ignorable, we contend that under the present approach this is not true. Notice that we have observations from a model fitted to the responding units so that we are basically faced with the classical problem of testing the goodness of fit of observed data to an underlying hypothesized model. The argument in favor of the claim that the model cannot be tested is that it may be the case that there is more than one combination of a population model and a sampling or response mechanism yielding the same respondents models, such that the respondents model is not identifiable or weekly identifiable. Pfeffermann and Landsman (2011) provide conditions under which the respondents model is identifiable, with references to other related studies. On the other hand, notice the role of the constraints that the approach proposed in this paper relies upon. Indeed, as discussed in Section 3 and illustrated in Section 6, if the number of the constraints is deficient, alternative models may better fit the observed data.

Hosmer-Lemeshow-Type Testing

Following Pfeffermann and Landsman (2011), Pfeffermann and Sikov (2011) applied several goodness of fit tests to test the respondents model for the case where the outcome is continuous. Below we describe the application of the Hosmer and Lemeshow (1980, 2000) test statistic for the case of a binary outcome, which performed well in our simulation study. To construct this test statistic, the sample is partitioned into G groups of approximately equal size, based on the predicted probability of 'success.' The test statistic is then defined as

$$\hat{C} = \sum_{k=1}^{G} \frac{(o_k - n_k\bar{\mu}_k)^2}{n_k\bar{\mu}_k(1-\bar{\mu}_k)}, \tag{4.3}$$

where o_k is the number of observed 'successes' in group k, n_k is the size of the group, and $\bar{\mu}_k$ is the mean number of the estimated probabilities of success therein, $\bar{\mu}_k = \sum_{i\in G_k}\hat{\mu}_i/n_k$, where G_k is the kth group, and where $\mu_i = \Pr(y_i = 1, I_i = 1, R_i = 1|x_i)$.

The conditional probability of observing y_i, given that $i \in \mathcal{R}$ and given its covariate values is

$$\Pr(y_i|x_i, i \in \mathcal{R}; \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \Pr(y_i|x_i; \boldsymbol{\beta})\tau(x_i, y_i)\Pr(R_i = 1|x_i, y_i; \boldsymbol{\gamma})$$

and is estimated as follows. First, we estimate $\Pr(y_i|x_i)$ by $\hat{y}_i = \text{logit}^{-1}(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i)$. Next, $\hat{\tau}(x, y = 1)$ is obtained by interpolation of the $\hat{\tau}(x_i, y_i)$ values at the points $\{x_i : y_i = 1\}$. Similarly, $\hat{\tau}(x, y = 0)$ is obtained. Denote $p_{1,i} = \hat{y}_i\hat{\tau}(x, y = 1)\Pr(R = 1|y_i = 1, x_i, \hat{\boldsymbol{\gamma}})$ and $p_{0,i} = (1 - \hat{y}_i)\hat{\tau}(x, y = 0)\Pr(R = 1|y_i = 0, x_i, \hat{\boldsymbol{\gamma}})$. Then $\hat{y}_i^* = p_{1,i}/(p_{1,i} + p_{0,i})$ is the estimate of the probability $\Pr(y_i|x_i, i \in \mathcal{R})$.
Hosmer and Lemeshow (1980) found through an empirical study that their test statistic follows approximately a $\chi^2_{(G-2)}$ distribution under the null hypothesis that the model fits the data. We verified in our empirical study in Section 6 that a similar result holds in the case of our model.

6 Imputation of Non-Respondents' Data

In this section, we propose methods for imputation of the non-respondents data, depending on the whether the auxiliary variables $\boldsymbol{x}$ are available for the non-respondents. The goal is to impute observations for each subject in the non-respondents set $\mathcal{R}^c$ in such a way that the distribution of the variables in the combined data $\mathcal{R} \cup \mathcal{R}^c$ is the same as in that in the original sample, including the unobserved data.
Let us denote the conditional distribution of the response variable given the covariates and given unit i *not* responding by $f_{nr}(y_i|\boldsymbol{x}_i)$, and let $\rho_i = \rho(y_i, \boldsymbol{x}_i; \gamma) = E(R_i = 0|y_i, \boldsymbol{x}_i; \gamma)$. An analogue of of (2.3), and similar to (20) of Pfeffermann-Sikov (2011) is $f_{nr}(y_i|\boldsymbol{x}_i) \propto (1 - \rho_i)f_s(y_i|\boldsymbol{x}_i)$, leading to

$$f_{nr}(y_i|\boldsymbol{x}_i) \propto \frac{1 - \rho_i}{\rho_i} f_r(y_i|\boldsymbol{x}_i). \tag{6.1}$$

Similarly

$$f_{nr}(\boldsymbol{x}_i, y_i) \propto \frac{1 - \rho_i}{\rho_i} f_r(\boldsymbol{x}_i, y_i), \tag{6.2}$$

and

$$f_{nr}(\boldsymbol{x}_i) \propto \frac{1 - \rho_i}{\rho_i} f_r(\boldsymbol{x}_i). \tag{6.3}$$

Let $\hat{p}^{(r)}$ be the EL estimates obtained according to the method described in Section 4, and let $\hat{p}^{(nr)}$ be the corresponding estimates of the multinomial parameter characterising the joint distribution in the non-respondents set of the variables.
Consider now the following two scenarios below.
Scenario 1: The explanatory variables $\boldsymbol{x}_i$ not available for non-responding units.
From (6.2), we get

$$\hat{p}^{(nr)} \propto (1 - \hat{\rho}_i)\hat{\rho}_i{}^{-1}\hat{p}_i^{(r)}. \tag{6.4}$$

Thus, under Scenario 1, data for $\mathcal{R}^c$ can be imputed by drawing $n-r$ independent observations $i(1), \ldots i(n - r)$ from Multinomial$(\hat{p}_1^{(nr)}, \ldots, \hat{p}_r^{(nr)})$. We now have the two alternatives below.

(a) Take $(y_{i(1)}, \boldsymbol{x}_{i(1)}), \ldots, (y_{i(n-r)}, \boldsymbol{x}_{i(n-r)})$ as the imputed data.

(b) Impute the covariates $x_{i(1)}, \ldots, \boldsymbol{x}_{i(n-r)}$ as in (a) and then independently draw $y_{i(j)}$ for each $i(j)$ from the estimated model $f_{nr}(y|\boldsymbol{x})$. Note that informative non-response and sampling alter the model in such a way that even if the population model $f(y|\boldsymbol{x})$ is one of the familiar generalised linear models, such as linear or logistic regression, the distribution $f_r(y|\boldsymbol{x})$ may be very different. One possibility is

to estimate $f_{nr}(y|\boldsymbol{x})$ non-parametrically (e.g., using smooth kernel regression, with weights given by (6.4)) and use this estimated model to draw the imputed y values.

SCENARIO 2: The explanatory variables $\boldsymbol{x}_i$ are available for both responding and non-responding units. In this case, we independently draw y_i for each i from the estimated model $f_{nr}(y|\boldsymbol{x})$, similar to (b) above. Sub-section 7.5 gives simulation results the imputation.

7 Simulation Study

7.1 Simulation set up

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$(2, 2)$. For each value x_j, a binary response y_j was generated with $\Pr(y_j = 1|x_j; \boldsymbol{\beta}) = \text{logit}^{-1}(-0.8+0.8x_j)$. Next, a value of a design variable Z was generated as $z_j = \max[(x_j+1.1)(2y_j+1)+\nu_j, 0.01]$, where $\nu_j \sim \text{Uniform}(-0.2, 0.2)$. Values of calibration variables $\boldsymbol{c}$ were generated as $\boldsymbol{c}_j = (1, x_j, y_j, x_j y_j, x_j^2, x_j^2 y_j)' + \boldsymbol{\varepsilon}_j$, with $\boldsymbol{\varepsilon}_j$ independently drawn from $N(\mathbf{0}_6, I_6)$. (Here, $\mathbf{0}_m$ and I_m respectively are the m-dimensional zero vector and the $m \times m$ identity matrix.) A sample was drawn by Bernoulli sampling ($I_j \overset{\text{indep}}{\sim} \text{Ber}(\pi_j)$), where $\pi_j = \min(3500 z_j^{-1} / \sum_{k=1}^{10000} z_k^{-1}, 0.9999)$. The sampled units were classified as respondents/non-respondents with $\Pr(R_j = 1) = \rho_j = \text{logit}^{-1}(\gamma_0 + \gamma_x x_j + \gamma_y y_j)$; $\gamma_0 = 0.07, \gamma_x = 0.5, \gamma_y = 1.5$. The process of generating the population values and selecting the sample and the subsample of respondents was repeated independently 500 times. (The x-values were generated only once). We used 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 (see next sub-section). For each sample of respondents we estimated the vector coefficients $(\boldsymbol{\gamma}, \boldsymbol{\beta})$ and the variance of the estimators using the procedures described in Section 4.

7.2 Smoothing Weights By Kernel Smoothing

In our simulation study, we used 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. Since in our case, y_i attains only two values, 0 or 1, the smoothing was applied to estimate $E_r(w_i|y_i = 0; x_i)$ and $E_r(w_i|y_i = 1; x_i)$ separately. The kernel regression was performed using the function `npreg` from the R package `np` at its default setting. Specifically, Nadaraya-Watson kernel smoothing was performed, with a bandwidth automatically calculated using the method of Racine and Li (2004) and Li and Racine (2004).

7.3 Estimation of the models' parameters

Tables 1 and 2 below give mean estimates from the simulation study described in subsection 6.1. Table 1 gives the mean estimates of the response probability model parameters ($\boldsymbol{\gamma}$) using the proposed method (CREL), as well as their empirical standard deviation, and the the square root of their mean variance estimates, using

the inverse information and the bootstrap methods. The BS numbers below are being updated and may change a little.

Mean Estimates			Empirical STD			Square Root Mean Variance Estimate			Square Root Mean Bootstrap Variance Estimate		
$\hat{\gamma}_0$	$\hat{\gamma}_x$	$\hat{\gamma}_y$	$\hat{\gamma}_0$	$\hat{\gamma}_x$	$\hat{\gamma}_y$	$\hat{\gamma}_0$	$\hat{\gamma}_x$	$\hat{\gamma}_y$	$\hat{\gamma}_0$	$\hat{\gamma}_x$	$\hat{\gamma}_y$
0.736	0.499	−1.523	0.214	0.212	0.319	0.220	0.212	0.339	0.291	0.223	0.404

Table 1: Estimation of $\boldsymbol{\gamma}$. Mean Estimates, Empirical Standard Deviations (STD), Square Root of Mean Variance Estimates by Inverse Information and by Bootstrap. ($\gamma_0 = 0.7, \gamma_x = -1.5, \gamma_y = -1.5, 3063 \leq n \leq 3954, 2039 \leq r \leq 2636$). Results from 300 main samples.

Table 2 below compares five mean estimates of the population model parameters $\boldsymbol{\beta}$: (1) the unweighted full sample estimate (hypothetically assuming the data for the non-respondents were obtained), (2) the weighted full sample estimate, (3) estimation using the respondents' data, ignoring the weights and the non-ignorability of the non-response, (4) estimation using the respondents' data and their sampling weights, (5) using the proposed method (CREL).

Method	Mean Estimates		Empirical STD		Square Root Mean Variance Estimate	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
FR UW	−1.902	0.802	0.073	0.071	0.073	0.071
FR PW	−0.798	0.799	0.073	0.072	0.075	0.072
MAR UW	−2.665	0.966	0.105	0.093	0.111	0.095
MAR PW	−1.559	0.962	0.106	0.093	0.113	0.097
CREL	−0.797	0.799	0.178	0.104	0.188	0.108
APW	−0.799	0.800	0.180	0.108	0.190	0.112

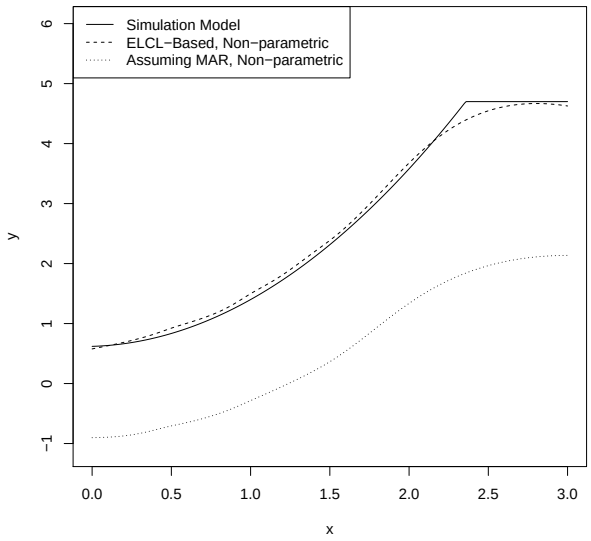
FR= Full response, estimators obtained from all sample data; UW= Unweighted; PW= Probability weighted by sampling weights; MAR= estimators obtained when ignoring response mechanism. CREL= proposed method: constrained respondents' EL. APW = proposed method: adjusted probability weighted (sampling weights divided by response probabilities (estimated by CREL). Variance estimation of the CREL and APW estimates based on parametric bootstrap from 60 main samples.

Table 2: Estimation of $\boldsymbol{\beta}$: Mean Estimates, Empirical Standard Deviations (STD) and Square Root of Mean Variance Estimates. ($\beta_0 = -0.8, \beta_1 = 0.8, 3395 \leq n \leq 3625, 2227 \leq r \leq 2455$).

7.4 Non-parametric Estimation of $f(y|x)$

One hundred samples were generated as follows: $N = 20,000$ x variables were generated once from $\Gamma(2,2)$, and then N y values were generated from the non-linear model $y = g(x) + e$ where $g(x) = -0.3 + \max(0, \min(5, 1 + 0.1x + 0.7x^2))$ and where $e \sim N(0, 3^2)$. A non-informative sample was then drawn using Poisson sampling. Non-response was generated from the model $\text{logit}(\Pr(R = 1|y, x)) = 0.35 + 0.05x - 0.5y$. The generation process was repeated 100 times. The sample size ranged from 4,839 to 5,115. The number of respondents ranged from 1,992 to 2,213. The response rate ranged from 0.40 to 0.44.

Figure 1: Non-parametric Estimation of Population Model.



Estimates of the regression parameter: As discussed in Sub-section 4.2.2, estimation of the regression parameter $\boldsymbol{\beta}$ can be valid even when the assumed model does not hold. In that case, $\boldsymbol{\beta}$ is viewed as only a descriptive parameter as defined by the population estimating equation. Both approaches proposed in this paper: CREL (constrained respondents' EL) and APW (adjusted probability weighted) give good estimates. The table below compares estimates from the 100 samples described above, by these methods to estimates when ignoring the response mechanism (while still using the sampling weights).

MAR-Weighted = estimators obtained when ignoring response mechanism, but using the sampling weights. CREL= proposed method: constrained respondents' EL. APW = proposed method: adjusted probability weighted (sampling weights divided by response probabilities (estimated by CREL).

Estimation of the response probability model:
The 10% and 90% quantiles of the ratio $\hat{\rho}_i/\rho_i$ of the estimated probability of response to the true one were 0.95 and 1.06, respectively.

	β_0	β_1	SE(β_0)	SE(β_1)
Finite Population	0.21	1.50	0.04	0.03
MAR-Weighted	−1.23	1.07	0.09	0.10
CREL	0.21	1.51	0.16	0.13
APW	0.26	1.47	0.31	0.23

Table 3: Mean estimates from 100 samples

	γ_0	γ_x	γ_y	SE(γ_0)	SE(γ_x)	SE(γ_y)
True	0.35	0.05	−0.50			
elcl	0.38	0.04	−0.51	0.18	0.10	0.06

Table 4: Estimates of the response model parameters

Percentile	10%	20%	30%	40%	50%	60%	70%	80%	90%
$\hat{\rho}_i/\rho_i$	0.95	0.97	0.99	1.00	1.00	1.01	1.02	1.04	1.06

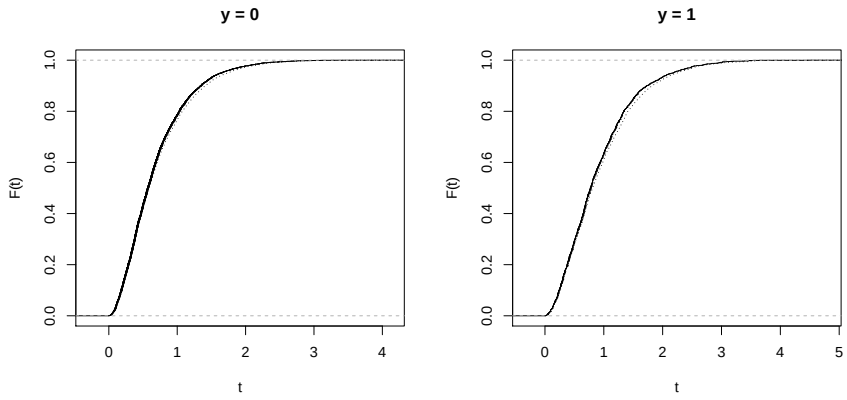
Table 5: The ratio of estimated response probability to its true value

7.5 Imputation Results

In practice, y is not observed for the non-respondents. However, when conducting a simulation, we do know the missing observations and thus we can compare the distributions $f_s(y|x)$ and $f_{imp}(y|x)$.

To illustrate the performance of the imputation procedure when the covariates are not available, we generated 100 samples in the same manner as described in Sub-section 7.1. The average percentage of units with $y = 1$ was 22.1% in the full samples and 22.6% in the imputed data. We have also plotted the empirical CDF (eCDF) of x given y in the full sample, separately for $y = 0$ and $y = 1$, and compared them to the corresponding eCDF of the imputed data. Averaged over 100 samples, the curves were practically identical. Therefore, we also compared the same eCDFs over just three samples. See Figure 2 below.

Figure 2: Comparison of the distribution of x by y in the full sample and in the imputed data, averaged over three samples. (Solid line: Full sample, dotted line: Imputed)



It appears that $f_{imp}(x) \approx f_s(x)$ and $f_{imp}(y|x) \approx f_s(y|x)$, which suggests that $f_{imp}(x, y) \approx f_s(x, y)$.

7.6 Discussion of Results

Parametric Estimation: Both the CREL estimates and the APW estimates were virtually unbiased, whereas ignoring the informativeness of the non-response (*viz.* assuming MAR) showed large biases. (Note that the 'full response' estimates are shown for reference only, and are impossible to obtain in real life.) The performance of the two proposed methods remains good even when the population model is misspecified, or even when no such model is specified, and $f(y|\boldsymbol{x})$ is estimated non-parametrically.
The parametric bootstrap variance estimates of the population model parameter were good as well.

7.7 The Role of the Constraints

An important element of the proposed method is the use of the known population mean of a vector of variables $\boldsymbol{c}_i$. Ideally, the components of $\boldsymbol{c}_i$ should approximate target model's variables. In this section, we study the impact of the choice of the variables $\boldsymbol{c}_i$ and their proximity to the target model's variables. A 6-dimensional $\boldsymbol{c} = (c_0, c_1, c_2, c_3, c_4, c_5)'$ variable was created where $\boldsymbol{c}_i = (1, x_i, y_i, x_i y_i, x_i^2, x_i^2 y_i)' + \boldsymbol{\varepsilon}_i$ and where the $\boldsymbol{\varepsilon}_i$ were independently generated from a multivariate normal $N(\mathbf{0}_6, \sigma_c^2 I_6)$, for a number of values of σ_c. Additionally, we tested the case where there was no correlation between $\boldsymbol{c}$ and the response model covariates. To avoid numerical problems with very large numbers, the $\boldsymbol{c}_i$ were divided by a constant so that their population values were in the $[-1, 1]$ range. The table below demonstrates the dependence of the accuracy of the model parameters estimates on the closeness of the survey calibration variables to the ones in the response model.

		β_0	β_1	γ_0	γ_x	γ_y	$se(\beta_0)$	$se(\beta_1)$	$se(\gamma_0)$	$se(\gamma_x)$	$se(\gamma_y)$
Simulation value ('True')		-0.800	0.800	0.700	0.500	-1.500	—	—	—	—	—
$\sigma_c = 0.5$	$c_0, c_1, c_2, c_3, c_4, c_5$	-0.804	0.802	0.702	0.501	-1.491	0.119	0.086	0.165	0.177	0.198
$\sigma_c = 0.5$	c_1, c_2, c_3, c_4, c_5	-0.804	0.802	0.702	0.501	-1.490	0.119	0.086	0.164	0.177	0.197
$\sigma_c = 0.5$	c_0, c_1, c_4	-1.012	0.754	1.880	0.353	-1.660	0.882	0.215	2.091	0.407	3.67
$\sigma_c = 0.5$	c_2, c_3	-0.796	0.797	0.699	0.516	-1.501	0.135	0.092	0.180	0.226	0.223
$\sigma_c = 0.5$	c_2, c_3, c_5	-0.800	0.800	0.705	0.505	-1.497	0.128	0.091	0.181	0.222	0.214
$\sigma_c = 1.0$	$c_0, c_1, c_2, c_3, c_4, c_5$	-0.802	0.800	0.712	0.507	-1.502	0.177	0.107	0.223	0.214	0.317
$\sigma_c = 1.0$	c_1, c_2, c_3, c_4, c_5	-0.801	0.800	0.712	0.507	-1.503	0.176	0.107	0.222	0.216	0.316
$\sigma_c = 1.0$	c_2, c_3	-0.786	0.784	0.716	0.539	-1.523	0.248	0.143	0.251	0.360	0.418
$\sigma_c = 1.0$	c_0, c_1, c_4	-1.098	0.761	1.759	0.314	-1.256	0.894	0.210	2.076	0.433	3.830
$\sigma_c = 1.0$	c_2, c_3, c_5	-0.790	0.793	0.723	0.517	-1.522	0.201	0.115	0.232	0.265	0.362
$\sigma_c = 9.0$	$c_0, c_1, c_2, c_3, c_4, c_5$	-0.893	0.733	1.358	0.710	-1.852	0.752	0.318	1.799	1.312	2.583
$\sigma_c = 9.0$	c_1, c_2, c_3, c_4, c_5	-0.856	0.727	1.500	0.708	-2.060	0.751	0.315	1.878	1.274	2.632
$\sigma_c = 9.0$	c_0, c_1, c_4	-1.115	0.717	1.445	0.857	-0.922	1.010	0.406	2.167	1.664	4.039
$\sigma_c = 9.0$	c_2, c_3	-0.800	0.671	1.526	1.278	-2.112	0.832	0.467	2.215	2.269	2.746
$\sigma_c = 9.0$	c_2, c_3, c_5	-0.764	0.628	1.656	0.778	-2.023	0.796	0.371	1.864	1.592	2.92
Uncorrelated $\boldsymbol{c} = (c_0, \ldots, c_5)'$		-1.051	0.753	1.143	2.280	-1.310	1.096	0.603	2.826	2.877	4.015

Table 6: Relationship between σ_c and accuracy of estimates. Mean estimates and standard errors (se) from 300 samples in each case.

7.8 Discussion of Results

The results in Table 3 demonstrate the importance of choosing the 'right' constraints. Note that using the pair c_2 and c_3 alone performed rather well, whereas the combination of c_0, c_1 and c_4 resulted in sizeable bias. This is explained by the fact that it is the conditional dependence of y given x that matters and in our case, c_2 and c_3 are proxies for y and xy, respectively. Therefore, to estimate the population model well, constraint variable should be chosen so that the conditional distribution of y given x is accounted for.

As the case c_2, c_3 with $\sigma_c = 1$ shows, even though only two constraints were used, the estimates are good (though less precise than when more constraints are used). Even with large noise ($\sigma_c = 9$) the performance was acceptable.

Finally, note that in our approach, in addition to the respondents data, only population means of the constraints variables need to be known.

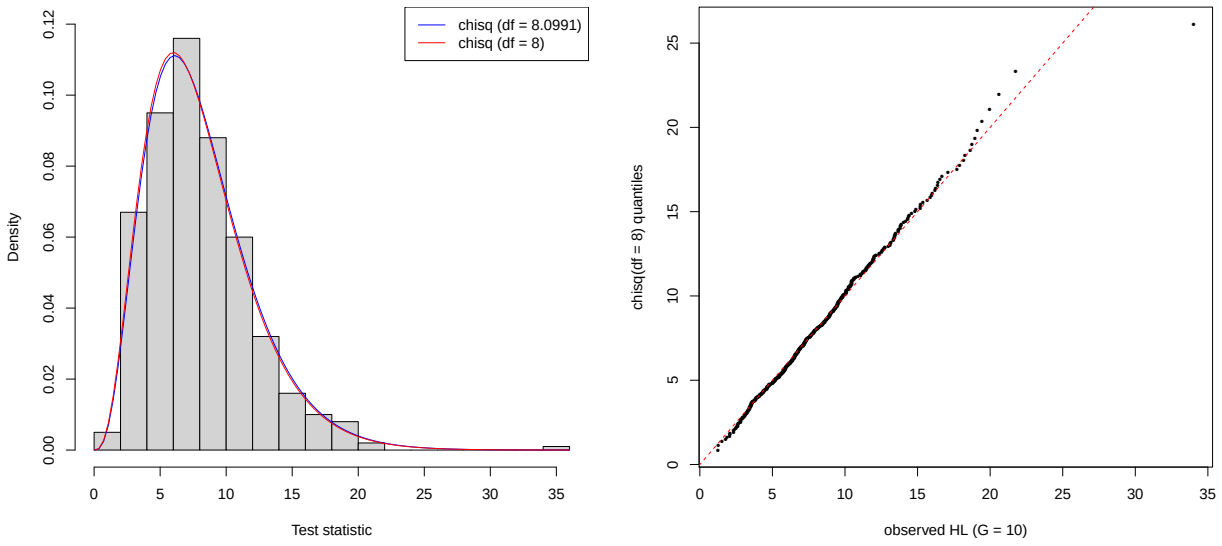
Remarks: The information provided by $c_0 = 1 + \varepsilon$ is essentially the same as that already provided by the r-constraint. Therefore, the c_0, c_2-based estimates (and the c_0, c_1-based estimates) appear to be far less accurate than the c_2, c_3-based estimates.

7.9 Model testing—Numerical Results

7.9.1 Distribution of Test Statistic under H_0

Figure 1 below, shows the empirical distribution of the Hosmer-Lemeshow-type test statistic with $G = 10$ nearly equal groups, calculated from 500 samples. Recall that if $x_1, \ldots, x_n$ are independent draws from a χ^2_d distribution, the log-likelihood of d is $\frac{1}{2}\sum \log x_i - \frac{nd}{2}\log 2 - \log\Gamma(\frac{d}{2}) + H(x_1, \ldots, x_n)$, where Γ is the Gamma function and $H(x)$ is a function of $x_1, \ldots, x_n$ alone. We have estimated the degrees of freedom as approximately 8.0991 using maximum likelihood estimation, and have included a QQ-plot, where the observed quantiles were compared to the expected quantiles from a χ^2_8 distribution.

Figure 3: Distribution of Test Statistics ($G = 10$ equal size groups) under $H_0 : \text{logit}(\rho_i) = 0.7 + 0.5x_i - 1.5y_i$; Comparison to a χ^2_8 and to a $\chi^2_{8.099}$ Distributions. (Based on 500 simulated samples.)



The two figures show a close approximation of the empirical distribution of the test statistic (with G = 10 groups) to the hypothesised χ^2_8 distribution, thus validating the conjecture of Hosmer-Lemeshow (2000).

7.9.2 Power of the test statistic

In order to explore the power of the proposed tests, we tested the null hypothesis that correct response model was of the form $\text{logit}(\rho_i) = 0.7 + 0.5x_i - 1.5y_i$ when in fact the data were generated using a response model of the form $\text{logit}(\rho_i) = 0.7 + 0.5x_i - 1.5y_i + ax_i^2 + bx_iy_i$, for a few combinations of a and b. In these simulations, the population model used to generate the data was $\log[\Pr(y = 1)/\Pr(y = 0)] = \beta_0 + \beta_1 x$, with $\beta_0 = -0.8, \beta_1 = 0.8$. It was assumed the form of the population model was known, but its parameter values were not. Table 4 below shows the rejection rates using the proposed test statistics with $G = 10$ and $G = 20$, at significance levels $\alpha = 0.05$ and $\alpha = 0.10$, for these combinations (labelled as $0, \ldots, 24$).

Outlier Samples. When assumptions made while fitting a model are far from being true, the estimation process may result in estimates of the population model's parameters having extremely high absolute values. In our simulations this was also accompanied with the estimated probability of 'success' of the dependent variable (i.e., $E(y)$) being either 0 or 1. We refer to such samples as 'outliers' and reject the hypothesis that the model fits.

Correct Model			Rejection Rate				Percent
			$G = 10$		$G = 20$		Outlier
Label	a	b	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.10$	Samples
0	0.0	0.0	0.0480	0.1020	0.0340	0.0880	0.0
1	0.0	-1.0	0.0660	0.1240	0.0700	0.1120	0.0
2	0.0	-1.2	0.0900	0.1380	0.0660	0.1260	0.0
3	0.0	-1.3	0.1100	0.1580	0.1140	0.1620	0.0
4	0.0	-1.5	0.2000	0.2620	0.1660	0.2180	0.2
5	-0.5	0.0	0.1400	0.2380	0.0860	0.1520	0.0
6	-0.5	-1.0	0.6700	0.7180	0.6060	0.6540	2.4
7	-0.5	-1.2	0.7420	0.7840	0.7060	0.7380	7.4
8	-0.5	-1.3	0.8000	0.8340	0.7500	0.7840	11.2
9	-0.5	-1.5	0.8600	0.8820	0.8380	0.8600	25.0
10	-0.6	0.0	0.2040	0.2640	0.1440	0.2100	0.0
11	-0.6	-1.0	0.7280	0.7840	0.7040	0.7300	3.6
12	-0.6	-1.2	0.8100	0.8460	0.7880	0.8120	13.0
13	-0.6	-1.3	0.8560	0.8700	0.8340	0.8520	18.6
14	-0.6	-1.5	0.9020	0.9220	0.9020	0.9080	26.2
15	-0.7	0.0	0.2620	0.3500	0.2380	0.2940	0.0
16	-0.7	-1.0	0.8180	0.8380	0.7700	0.8020	7.0
17	-0.7	-1.2	0.8820	0.9020	0.8480	0.8700	17.2
18	-0.7	-1.3	0.8960	0.9080	0.8700	0.8800	21.6
19	-0.7	-1.5	0.9440	0.9540	0.9340	0.9380	30.2
20	-0.8	0.0	0.3900	0.4720	0.3200	0.3900	0.0
21	-0.8	-1.0	0.8500	0.8780	0.8300	0.8500	11.6
22	-0.8	-1.2	0.9080	0.9240	0.8800	0.9000	16.8
23	-0.8	-1.3	0.9360	0.9480	0.9220	0.9340	20.8
24	-0.8	-1.5	0.9600	0.9680	0.9460	0.9560	27.2

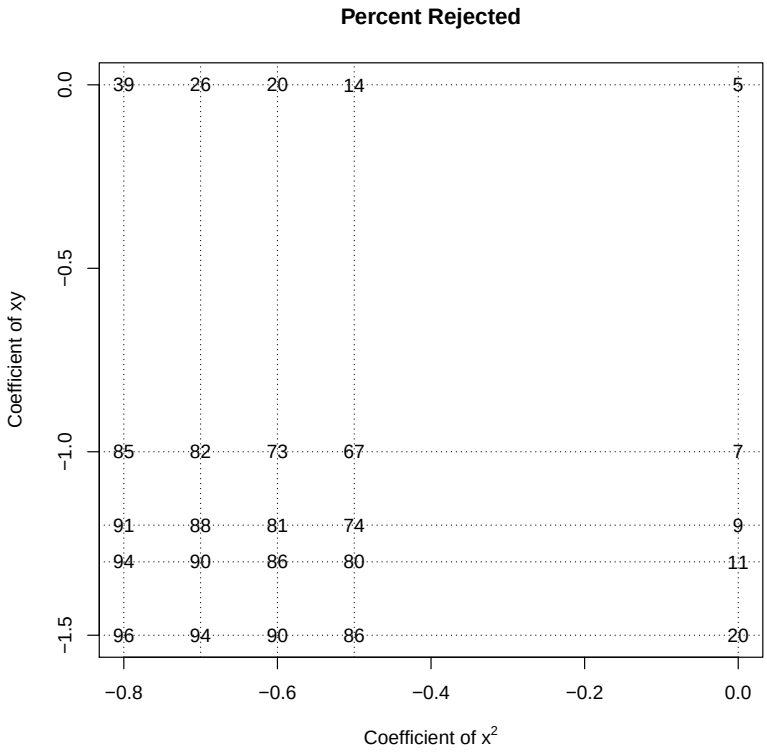
Table 7: Rejection rate and target model estimates when assuming (fitting) the response model $\text{logit}(\rho_i) = 0.7 + 0.5x_i - 1.5y_i$, while the correct response model is $\text{logit}(\rho_i) = 0.7 + 0.5x_i - 1.5y_i + ax_i^2 + bx_iy_i$, assuming the $\chi^2(G-2)$ distribution under the null hypothesis.

Note that with the number of groups $G = 10$, the observed rejection rates when the model was correctly

assumed (Model 0 in the table), were close to the nominal rates (0.05 and 0.10).

REMARK 5. As the dissimilarity between the correct model and the model assumed under the null hypothesis (i.e., as the distance between (a, b) and $(0, 0)$) grows, so do the rejection rates. The figure below illustrates this point.

Figure 4: Rejection rates when the correct response propensity model is $\Pr(R = 1) = (1 + \exp(-\{0.7 + 0.5x - 1.5y + ax^2 + bxy\}))^{-1}$ while assuming $a = b = 0$.



7.10 Discussion of Results

The Hosmer-Lemeshow-type test statistic with $G = 10$ groups showed an empirical distribution very close to χ^2_8. Testing with this test statistic and 10 groups at significance levels 0.05 and 0.1 resulted in rejection rates very close to the nominal rates, when the null hypothesis was true. The test statistic's power to reject the null hypothesis (H_0) is clearly dependent on how far the correct model is from the one under H_0. This was illustrated by adding a quadratic term x^2 and its interaction with y, *viz.* $ax^2 + bxy$, for a range of values. as the distance between a and b and the origin $(0, 0)$ grows, so do the rejection rates. In extreme cases, the fitted model exhibits a behaviour similar to complete separation, with the predicted probabilities at the endpoints of the interval $[0, 1]$.

8 Conclusions

We have developed and illustrated a general procedure for analysing complex survey data subject to informative sampling and NMAR nonresponse, with minimal assumptions. In fact, the only parametric model assumed is the model for the response probabilities as illustrated. Contrary to common misconception, this model can be tested with good power.

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, which is the common situation in large scale surveys used for the production of official statistics.

Acknowledgements

We like to thank Professor Sanjay Chaudhuri for providing us his R code, which helped us develop our own code. This study was funded by a UK ESRC grant No. RES-062-23-2316.

All the numerical calculations were carried out in R (R Development Core Team, 2006).

References

Chang, T. and P.S. Kott(2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, **95**, 555–571.

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, J., Variyath, A.M. and Abraham, B. (2008). Adjusted Empirical Likelihood and its Properties. *Journal of Computational and Graphical Statistics*, **17**, 426–443.

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.

Hosmer, D.W., and Lemeshow, S. (1980). A goodness-of-fit test for the multiple logistic regression model. *Communications in Statistics*, **A10**, 1043–1069.

Hosmer, D.W., Jr. and Lemeshow, S. (2000). Applied Logistic Regression (2nd ed.) New York: Wiley.

Kott, P.S. and Chang T. (2010). Using calibration weighting to adjust for nonignorable unit nonresponse. *Journal of the American Statistical Association*, **105**, 1265–1275.

Lee, J. and Berger, J.O. (2001). Semiparametric Bayesian analysis of selection models. *J. Amer. Statist. Assoc.* **96** 1397–1409.

Lumley, T. (2004). Analysis of Complex Survey Samples. *Journal of Statistical Software*, **9**, 8 (`http://www.jstatsoft.org/v09/i08`).

Murphy, S.A. and van der Vaart, A.W. (2000), On Profile Likelihood. *Journal of the American Statistical Association*, **95**, 449–465.

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.

Owen, A.B. (2013), Self-concordance for empirical likelihood. *Canadian Journal of Statistics*, **41**, 387–397.

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 Landsman, V. (2011). Are Private Schools Really Better Than Public Schools? Assessment by Methods for Observational Studies. *Annals of Applied Statistics*, **5**, 1726–1751.

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

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.

R Development Core Team (2006). R: a Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing. (Available from http://www.R-project.org.)

Li, Q. and J.S. Racine (2004), Cross-validated local linear nonparametric regression, *Statistica Sinica*, **14**, 485–512.

Racine J.S., Li Q (2004), Nonparametric Estimation of Regression Functions with both Categorical and Continuous Data. *Journal of Econometrics*, **119**, 99–130.

Rotnitzky, A. and Robins, J. (1997). Analysis of semi-parametric regression models with non-ignorable non-response. *Stat. Med.* **16** 81–102.

Scott, A.J., and Wild, C. (2006). Calculating efficient semiparametric estimators for a broad class of missing-data problems. Festschrift for Tarmo Pukkila on his 60th Birthday (Eds.: E.P. Liski, J. Isotalo, J. Niemel, S. Puntanen, and G.P.H. Styan.)

Shao, J. (1988). A note on bootstrap variance estimation. Purdue University Technical Report 88-29. (Available on line at `www.dtic.mil/dtic/tr/fulltext/u2/a204266.pdf`.)