Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?

by Danny Pfeffermann

December 2011

Statistics Canada Statistique Canada

Canada

Survey Methodology, December 2011
Vol. 37, No. 2, pp. 115-136
Statistics Canada, Catalogue No. 12-001-X

Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?

Danny Pfeffermann [1]

Abstract

This article attempts to answer the three questions appearing in the title. It starts by discussing unique features of complex survey data not shared by other data sets, which require special attention but suggest a large variety of diverse inference procedures. Next a large number of different approaches proposed in the literature for handling these features are reviewed with discussion on their merits and limitations. The approaches differ in the conditions underlying their use, additional data required for their application, goodness of fit testing, the inference objectives that they accommodate, statistical efficiency, computational demands, and the skills required from analysts fitting the model. The last part of the paper presents simulation results, which compare the approaches when estimating linear regression coefficients from a stratified sample in terms of bias, variance, and coverage rates. It concludes with a short discussion of pending issues.

Key Words: Informative sampling; NMAR nonresponse; Likelihood-based methods; Probability weighting; Randomization distribution; Sample model.

1. Introduction

Survey data are frequently used for analytic inference on statistical models, which are assumed to hold for the population from which the sample is taken. Familiar examples include the estimation of income elasticities from household surveys, the analysis of labour market dynamics from labour force surveys, comparisons of pupils' achievements from educational surveys and the search for causal relationships between risk factors and disease prevalence from health surveys. An important common feature to all these examples is that interest lies in the structure of the models being estimated and what can be learnt from them. This is different from fitting models merely for prediction purposes, such as when predicting finite population totals or in small area estimation, where the structure and interpretation of the model are of secondary importance. Models are also used implicitly for choosing the sampling design and estimators, such as in stratified sampling, or when defining weighting cells for nonresponse adjustments. However, inference is typically based in these cases on the randomization distribution over all possible sample selections, and not on the model, which is known as 'model assisted inference'.

Survey data typically differ from other data sets in five main aspects.

1. The samples are selected at random with known selection probabilities, which allows using the randomization distribution over all possible sample selections as the basis for inference instead of the hypothetical distribution underlying the population model. As discussed below, a combination of the two distributions is in common use.
2. The sample selection probabilities in at least some stages of the sample selection are often unequal; when these probabilities are related to the model outcome variable, the sampling process becomes informative and the model holding for the sample is then different from the target population model.
3. Survey data are almost inevitably subject to various forms of nonresponse, often of considerable magnitude, which again may distort the population model if the response propensity is associated with the outcome of interest (not missing at random nonresponse).
4. The sample data are often clustered due to the use of multi-stage cluster samples. The clusters are 'natural units' (households, individuals in case of longitudinal surveys…), implying that observations within the same cluster are correlated.
5. The data available to the modeler may be masked ("swapped", "contaminated", suppressed") in order to protect the anonymity of the respondents. When this is the case, the modeler's data differ from the correct data.

Many approaches have been proposed in the literature for estimating population models from complex survey data possessing these features, some of which are more familiar than the others. The approaches differ in the conditions underlying their use, the data required for their application, goodness of fit testing, the inference objectives that they accommodate, statistical efficiency, computational demands, and the skills required from analysts fitting the model. This heterogeneity means that there does not exists any single

1. Danny Pfeffermann, Southampton Statistical Sciences Research Institute, U.K. and Hebrew University of Jerusalem, Israel. E-mail: d.pfeffermann@soton.ac.uk.

approach that can be considered as best in all situations. That being the case, a fundamental question arising is which approach or approaches could or should be used for a given practical application.

The present paper is divided into three parts. In the first part (Section 2) I elaborate on the first four features of complex survey data mentioned above. In the second part (Section 3) I review the various approaches proposed in the literature for dealing with these features, discussing their merits and limitations in light of the properties mentioned above. In the third part (Section 4) I present simulation results which compare the approaches when estimating a linear regression model from a stratified sample in terms of bias, variance, and coverage rates. I conclude with a short discussion of pending issues in Section 5.

2. Why are survey data different from other data?

2.1 The problem of unequal sampling probabilities and nonresponse

Consider a finite population $U = \{1, ..., N\}$ with measurements $\{y_i, \mathrm{x}_i, \mathrm{z}_i\}$ for unit $i = 1, ..., N$, where y represents an outcome variable of interest, x a vector of covariates and z a vector of design variables used for the sample selection. The design variables may include some or all of the covariates, and in special cases also the outcome variable when known for all the population units, such as in case-control studies. The matrix $Z_U = [\mathrm{z}_1, ..., \mathrm{z}_N]$ is known to the sampler drawing the sample, but not necessarily to the analyst fitting the model. Denote by $s = (\mathrm{I}_1, ..., \mathrm{I}_N)$ the selected sample, where I_i is the sampling indicator taking the value 1 if unit $i \in U$ is drawn to the sample and 0 otherwise. In practice, not all the sampled units necessarily respond, and we denote by R_i the response indicator; $R_i = 1(0)$ if unit $i \in S$ responds (does not respond).

The observed data may be viewed as the outcome of three random processes. The first process generates the vectors $\{y_i, \mathrm{x}_i, \mathrm{z}_i\}$ for the N population units. The second process selects a sample s from U at random by a sampling design, $\Pr(s) = \Pr(s \mid Z_U)$. The third process selects the responding units. This process is obviously not part of the original sampling design and is often the result of 'self selection', although nonresponse could be caused by many other reasons. See Brick and Montaquila (2009) for a recent overview.

When the sample selection probabilities and/or the response probabilities are related to the values of the outcome variable even after conditioning on the model covariates, in the sense that $\Pr(\mathrm{I}_i = 1 \mid y_i, \mathrm{x}_i) \neq \Pr(\mathrm{I}_i = 1 \mid \mathrm{x}_i)$ or $\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, \mathrm{I}_i = 1) \neq \Pr(R_i \mid \mathrm{x}_i, \mathrm{I}_i = 1)$, the model holding for the observed outcomes is different from the population model. In symbols, $f_o(y_i|\mathrm{x}_i) \neq f_p(y_i|\mathrm{x}_i)$, where $f_o(y_i|\mathrm{x}_i)$ represents the model holding for a unit selected to the sample and responding, and $f_p(y_i|\mathrm{x}_i)$ is the *population* model (the model holding for the population values). See Equations (2.1) and (2.2) below.

Example 1. Suppose that the population model is the regression model, $f_p(y_i \mid \mathrm{x}_i) = N(\mathrm{x}_i'\beta, \sigma_\varepsilon^2)$, and that the sample is selected with selection probabilities satisfying $\Pr(I_i = 1 \mid y_i, \mathrm{x}_i) = \exp[\gamma_1 y_i + \gamma_2 y_i^2 + g(\mathrm{x}_i)]$, where γ_1 and $\gamma_2 \leq 0$ are constants and $g(\mathrm{x}_i)$ is some nonstochastic function of the covariates. Simple use of Bayes theorem (see below) shows that the model holding for the sample outcomes is in this case, $f_s(y_i \mid \mathrm{x}_i) = N[(\gamma_1\sigma^2 + \mathrm{x}_i'\beta)/C, \sigma_\varepsilon^2/C]$, where $C = (1 - 2\sigma_\varepsilon^2\gamma_2)$. Thus, although the sample residuals have again a normal distribution, the regression coefficients and the residual variance are different from their values under the population model. In the special case $\gamma_2 = 0$, the slope coefficients and the residual variance are the same as under the population model, but not the intercept. If $\gamma_1 = 0$ as well, the sample selection probabilities satisfy $\Pr(\mathrm{I}_i = 1 \mid y_i, \mathrm{x}_i) = \Pr(\mathrm{I}_i = 1 \mid \mathrm{x}_i)$ and the two models are now the same.

Following conventional terminology, when $\Pr(\mathrm{I}_i = 1 \mid y_i, \mathrm{x}_i) \neq \Pr(\mathrm{I}_i = 1 \mid \mathrm{x}_i)$ the sampling design is said to be *informative*. When $\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, \mathrm{I}_i = 1) \neq \Pr(R_i \mid \mathrm{x}_i, \mathrm{I}_i = 1)$, the nonresponse is *not missing at random* (NMAR nonresponse). Notice that whereas the sampling probabilities are typically known to the analyst fitting the model, at least for the sampled units, the response probabilities are generally unknown and need to be modelled under NMAR nonresponse. Ignoring an informative sample or NMAR nonresponse and thus assuming implicitly that the model holding for the observed outcomes is the same as the target population model may yield large biases and erroneous inference. The books edited by Kasprzyk, Duncan, Kalton and Singh (1989), Skinner, Holt and Smith (1989) and Chambers and Skinner (2003) contain many discussions and illustrations of the effect of ignoring informative sampling or NMAR nonresponse. See also Pfeffermann (1993, 1996), Pfeffermann and Sverchkov (2009) and Pfeffermann and Sikov (2011) for further discussions and examples, with many other more recent references.

In what follows, I use the abbreviation "*pdf*" to define the probability density function when the outcome is continuous or the probability function when the outcome is discrete. Suppose first that there is no nonresponse. Following Pfeffermann, Krieger and Rinott (1998a), 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 $(\mathrm{I}_i = 1)$. By Bayes theorem,

$$f_s(y_i \mid \mathrm{x}_i) = f(y_i \mid \mathrm{x}_i, I_i = 1) = \frac{\Pr(I_i = 1 \mid \mathrm{x}_i, y_i) f_p(y_i \mid \mathrm{x}_i)}{\Pr(I_i = 1 \mid \mathrm{x}_i)}, \quad (2.1)$$

where $f_p(y_i \mid \mathrm{x}_i)$ is the corresponding population *pdf*. The probabilities $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i)$ are generally not the same as the sample selection probabilities $\pi_i = \Pr(\mathrm{I}_i = 1)$, which may depend on all the population values Z_U of the design variables. However, the use of the marginal sample *pdf* only requires modelling $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i)$. 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*.

Remark 1. In practice, the covariates featuring in the population model need not be the same as the covariates featuring in the model of the conditional sample inclusion probabilities, $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i)$. In fact, following the results in Pfeffermann and Landsman (2011), identifiability of the sample model often requires that the two sets of covariates are not identical. However, to simplify the presentation in this paper, I assume for convenience that the covariates contained in the population model and the covariates defining the conditional inclusion probabilities are the same, or alternatively, that x_i defines the union of the two sets of covariates.

It follows from (2.1) that unless $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i) = \Pr(I_i = 1 \mid \mathrm{x}_i) \forall y_i$, the sample *pdf* is different from the population *pdf*, in which case the sampling design is informative and cannot be ignored in the inference process. In particular, it follows from (2.1) that under informative sampling,

$$E_s(y_i \mid \mathrm{x}_i) = E_p\left[\frac{\Pr(I_i = 1 \mid \mathrm{x}_i, y_i) y_i}{\Pr(I_i = 1 \mid \mathrm{x}_i)} \,\middle|\, \mathrm{x}_i\right] \neq E_p(y_i \mid \mathrm{x}_i),$$

where $E_s(\cdot)$ is the expectation under the sample *pdf*. Estimating $E_p(y_i \mid \mathrm{x}_i)$ is often the main target of inference, illustrating that ignoring an informative sampling scheme and thus estimating implicitly $E_s(y_i \mid \mathrm{x}_i)$ can bias the inference.

Suppose now the existence of NMAR nonresponse. The marginal sample *pdf* (2.1) can be extended to this case by defining,

$$\begin{aligned} f_o(y_i \mid \mathrm{x}_i) &= f(y_i \mid \mathrm{x}_i, I_i = 1, R_i = 1) \\ &= \frac{\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1) \Pr(I_i = 1 \mid y_i, \mathrm{x}_i) f_p(y_i \mid \mathrm{x}_i)}{\Pr(R_i = 1 \mid \mathrm{x}_i, I_i = 1) \Pr(I_i = 1 \mid \mathrm{x}_i)} \\ &= \frac{\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1) f_s(y_i \mid \mathrm{x}_i)}{\Pr(R_i = 1 \mid \mathrm{x}_i, I_i = 1)}. \end{aligned} \tag{2.2}$$

Notice from (2.2) that unless $\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, \mathrm{I}_i = 1) = \Pr(R_i \mid \mathrm{x}_i, \mathrm{I}_i = 1) \forall y_i$, the *pdf* holding for the observed outcomes is different from the sample *pdf*. Here again I assume for convenience that the response probabilities depend on the same covariates as in the sample model. See Remark 1 above.

The *pdf*s (2.1) and (2.2) define the marginal distributions of the outcome for a given unit. These definitions generalize very naturally to the joint *pdf* of two or more outcomes associated with different units. More generally, define for every plausible sample $s \subset U$ the sample indicator A_s, such that $A_s = 1$ if s is sampled and $A_s = 0$ otherwise, and assume for convenience full response. Denote the data associated with s by $(\mathrm{y}_s, \mathrm{x}_s)$. The joint sample *pdf* of $\mathrm{y}_s \mid \mathrm{x}_s$ is then,

$$\begin{aligned} f_s(\mathrm{y}_s \mid \mathrm{x}_s) &= f(\mathrm{y}_s \mid \mathrm{x}_s, A_s = 1) \\ &= \frac{\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s) f_p(\mathrm{y}_s \mid \mathrm{x}_s)}{\Pr(A_s = 1 \mid \mathrm{x}_s)}. \end{aligned} \tag{2.3}$$

The *pdf* $f_p(\mathrm{y}_s \mid \mathrm{x}_s)$ can be general, allowing in particular for correlated measurements, but modelling the probability $\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s)$ is practically only feasible if the sample can be decomposed into exclusive and exhaustive subsets s_k such that $\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s) \propto \prod_k \Pr(A_{s_k} = 1 \mid \mathrm{y}_{s_k}, \mathrm{x}_{s_k})$ and $\Pr(A_{s_k} = 1 \mid \mathrm{y}_{s_k}, \mathrm{x}_{s_k})$ satisfies the same model for all the subsets (see Example 2). In particular, if the population outcomes are independent given the covariates under the population model and $\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s) \propto \prod_{i \in s} \Pr(\mathrm{I}_i = 1 \mid y_i, \mathrm{x}_i)$, (2.3) takes the form

$$\begin{aligned} f_s(\mathrm{y}_s \mid \mathrm{x}_s) &= \prod\nolimits_{i \in s} \frac{\Pr(\mathrm{I}_i = 1 \mid y_i, \mathrm{x}_i) f_p(y_i \mid \mathrm{x}_i)}{\Pr(\mathrm{I}_i = 1 \mid \mathrm{x}_i)} \\ &= \prod\nolimits_{i \in s} f_s(y_i \mid \mathrm{x}_i), \end{aligned} \tag{2.4}$$

so that the sample outcomes are likewise independent.

Example 2. Consider the case of a clustered population $U = \bigcup_l U_l$, with independent measurements between clusters, such that $f_p(\mathrm{y}_U \mid \mathrm{x}_U) = \prod_l f_p(\mathrm{y}_{U_l} \mid \mathrm{x}_{U_l})$, where $(\mathrm{y}_U, \mathrm{x}_U)$ defines all the population values and $(\mathrm{y}_{U_l}, \mathrm{x}_{U_l})$ the values in cluster l. Let s define the set of sampled clusters, assumed to be drawn independently with probabilities $\Pr(l \in s \mid \mathrm{y}_{U_l}, \mathrm{x}_{U_l}) = r(\mathrm{y}_{U_l}, \mathrm{x}_{U_l})$ for some function $r(\cdot)$, and suppose also that all the units in the sampled clusters are observed (single-stage cluster sampling). Then, $\Pr(A_s = 1 \mid \mathrm{y}_U, \mathrm{x}_U) = \prod_{k \in s} r(\mathrm{y}_{U_k}, \mathrm{x}_{U_k}) \times \prod_{j \notin s} [1 - r(\mathrm{y}_{U_j}, \mathrm{x}_{U_j})]$. Since for $k \in s, (\mathrm{y}_{U_k}, \mathrm{x}_{U_k}) = (\mathrm{y}_{s_k}, \mathrm{x}_{s_k})$, it follows that $\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s) = \prod_{k \in s} r(\mathrm{y}_{s_k}, \mathrm{x}_{s_k}) \times G$, where for given covariates x_{U_j}, $j \notin s_1$, G is a constant satisfying, $G = \prod_{j \notin s} \int [1 - r(\mathrm{y}_{U_j}, \mathrm{x}_{U_j})] f_p(\mathrm{y}_{U_j} \mid \mathrm{x}_{U_j}) d\mathrm{y}_{U_j}$. The case of a non-clustered population with independent measurements and Poisson sampling of individual units is a special case where each cluster consists of a single element, giving rise to (2.4).

Remark 2. The examples considered so far assume independent sampling, which preserves the independence of the outcomes after sampling, but this assumption can

usually be relaxed following a result proved and illustrated in Pfeffermann *et al*. (1998a). By this result, under some general regularity conditions and for many commonly used sampling schemes for selection with unequal probabilities, if the population measurements are independent, the sample measurements are *asymptotically independent* under the sample distribution. The asymptotic framework requires that the population size increases but the sample size is held fixed. As illustrated in section 2.3, the assumption of independent population measurements is often also not restrictive.

So far, we suppressed for convenience from the notation the parameters underlying the population *pdf* and the sampling process. Consider, for example, the sample *pdf* (2.3). With added parameter notation, it can be written as

$$f_s(\mathrm{y}_s \,|\, \mathrm{x}_s; \theta, \gamma) = \frac{\Pr(A_s = 1 \,|\, \mathrm{y}_s, \mathrm{x}_s; \gamma)\, f_p(\mathrm{y}_s \,|\, \mathrm{x}_s; \theta)}{\Pr(A_s = 1 \,|\, \mathrm{x}_s; \theta, \gamma)}. \quad (2.5)$$

Thus, the conditional population and sample *pdf*s are different, unless

$$\Pr(A_s = 1 \,|\, \mathrm{y}_s, \mathrm{x}_s; \gamma) = \Pr(A_s = 1 \,|\, \mathrm{x}_s; \theta, \gamma)\ \forall\, \mathrm{y}_s. \quad (2.6)$$

When (2.6) holds, inference on the target parameter θ can be implemented by fitting the population model to the sample data, ignoring the sample selection. Note that this conclusion refers to the selected sample defined by the event $A_s = 1$.

The condition (2.6) is a strong condition. In a fundamental article on missing values, Rubin (1976) establishes conditions under which the sampling process can be ignored for likelihood, Bayesian or sampling theory (repeated sampling from a model) inference, that is, conditions under which the population model defined by $f_p(\mathrm{y}_s \mid \mathrm{x}_s; \theta)$ can be fitted to the observed data, depending on the inference method used. Little (1982) extends Rubin's results by distinguishing between the sample selection and the response process. Another important distinction is that Little conditions on the population values Z_U of the design variables used for the sample selection. Inference on the target population model $f_p(\mathrm{y}_s \,|\, \mathrm{x}_s; \theta)$ requires therefore integrating the conditional *pdf* of $\mathrm{y}_s \,|\, Z_U, \mathrm{x}_s$ over the distribution of $Z_U \,|\, \mathrm{x}_s$ (see Section 3). Sugden and Smith (1984) establish conditions under which a sampling process that depends on design variables Z is ignorable, given partial information on the design. Let $d_s = D_s(z_U)$ contain all the available design information for a sample s such as strata membership (may only be known for the sampled units), sample selection probabilities *etc*. Using previous notation, a key condition for ignorability of the sampling process given the available design information is that $A_s \perp Z_U \mid d_s$, with "$\perp$" meaning independence, implying $\Pr(A_s = 1 \mid Z_U = z_U) = \Pr(A_s = 1 \mid d_s)$ for any z_U for which $D_s(z_U) = d_s$.

For large scale multi-stage sample surveys with possibly many design variables, it is generally difficult and often impractical to check directly the conditions that permit ignoring the sample selection or nonresponse given the available design information. On the other hand, even when the sample *pdf* is different from the population *pdf*, it does not necessarily imply that inference that ignores the sampling process is wrong. As a simple illustration, consider the special case of Example 1 where $\gamma_2 = 0$. In this case the sample *pdf* is normal with the same slope coefficients and residual variance as under the population *pdf*. Thus, for inference about the slope coefficients one can ignore the sampling process. A similar result holds for logistic models when the sample selection depends on y but not on x. See Pfeffermann *et al*. (1998a) for derivation of this result. Pfeffermann and Sverchkov (2009) review several test statistics proposed in the literature for assessing whether ignoring the sample selection is justified for the intended inference.

2.2 The use of the randomization distribution for inference

A unique feature of sample surveys is that the sample is selected at random by use of a sampling design $[\{s, \Pr(s)\}, s \in S]$. The sampling design induces a (discrete) *randomization distribution* for any statistic T_{ys}, which is the conditional distribution over all possible sample selections, given the finite population values. Thus, the statistic T_{ys} takes the value t_{ys} with probability $\Pr(s)$, $s \in S$. Classical survey sampling inference is based solely on this distribution. For example, the familiar Horvitz-Thompson (HT) estimator T_{ys}^{HT}, which takes the value $t_{ys}^{\mathrm{HT}} = \sum_{i \in s}(y_i / \pi_i)$ if sample s is drawn, is randomization-unbiased for the finite population total $\mathrm{TOT}_y = \sum_{j=1}^{N} y_j$, since $\sum_{s \in S} \Pr(s)\, t_{ys}^{\mathrm{HT}} = T_y$. Its variance is, $\mathrm{Var}(T_{ys}^{\mathrm{HT}}) = \sum_{s \in S} \Pr(s)(t_{ys}^{\mathrm{HT}} - T_y)^2$. Notice that in the case of nonresponse, the use of the randomization distribution requires knowledge of the response probabilities, which in practice can only be estimated. The HT estimator takes in this case the form, $T_{ys}^{\mathrm{HT}} = \sum_{i \in R} y_i / [\pi_i \times \hat{\Pr}(R_i = 1 \mid I_i = 1)]$, where R defines the subsample of respondents. See Fuller (2002) for further discussion.

The randomization distribution conditions on the realized population values. Consequently, it can be used for descriptive inference on known functions of the finite population values, but not for analytic inference on a hypothesized model giving rise to these values. For this, one may consider the joint distribution over all possible sample outcomes for given population values (the *randomization r*-distribution) and all possible realizations of the finite population measurements (the *model p*-distribution). See Binder and Roberts (2009) and the references therein. The combined $r - p$ distribution offers an alternative framework of

inference to the use of the *pdf*s $f_s(y \mid \mathrm{x})$ or $f_o(y \mid \mathrm{x})$ defined before.

Example 3: Suppose that the population model is $y_i \sim \mathrm{Mult}[\{p_k\}, K]$, such that $\Pr_p(y_i = k) = p_k$, $k = 1, ..., K$; $\sum_{k=1}^{K} p_k = 1$. Let $\Pr(i \in s \mid y_i = k) = \pi_k$. Then, by (2.1), $\Pr_s(y_i = k) = \Pr(y_i = k \mid i \in s) = \pi_k p_k / \sum_{j=1}^{K} \pi_j p_j = p_k^*$, or, $y_i \mid i \in s \sim \mathrm{Mult}(\{p_k^*\}, K)$. Assuming independence of the observed outcomes and known selection probabilities, the maximum likelihood estimator (*mle*) of p_k based on the sample distribution is $\tilde{p}_k = (n_k / \pi_k) / \sum_{j=1}^{K} (n_j / \pi_j)$, where n_k is the number of sampled units with outcome $y_i = k$. The use of the $r - p$ distribution suggests estimating p_k by the HT estimator $\hat{p}_k = (1/N) \sum_{i \mid y_i = k} (1/\pi_k) = (n_k / \pi_k)/N$. The estimator $\hat{p}_k$ is randomization-unbiased for $\hat{P}_k = N_k / N$, where N_k is the number of population units with outcome $y_j = k$, and $\hat{P}_k$ is p-unbiased for p_k, such that $\hat{p}_k$ is $r - p$-unbiased for p_k.

The obvious difference between the $r - p$ distribution and the sample distribution, $f_s(y \mid \mathrm{x})$, is that the latter conditions on the observed sample of units (and hence the observed values of the covariates or the selected clusters in a cluster sample), whereas the $r - p$ distribution accounts for all possible sample selections. Consequently, the use of the latter distribution does not lend itself in general to conditional inference. The use of the *pdf*s $f_s(y \mid \mathrm{x})$ or $f_o(y \mid \mathrm{x})$ requires modelling $\Pr(I_i = 1 \mid \mathrm{x}_i, y_i)$ (Equation 2.1) and $\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1)$ in case of nonresponse (Equation 2.2), but it permits the computation (estimation) of the conditional *pdf* of the observed outcomes given the covariates, and hence the use of classical inference tools.

2.3 Data obtained from a cluster sample

Another special feature of survey data mentioned in the introduction is *clustering*, due to the use of multi-stage cluster samples. The clusters are 'natural groups' such as households, residence blocks, schools, or even individuals in the case of longitudinal surveys. Consequently, the outcomes pertaining to the same cluster are generally correlated, known as the *intraclass correlation*. It is important to emphasize that the clusters represent an existing population grouping, such that an intraclass correlation exists also under the population model.

Pfeffermann and Smith (1985) review several classes of plausible regression models for clustered populations, and discuss how they can be estimated from the sample. A population model in common use is the random intercept model,

$$y_{ij} = \mathrm{x}_{ij}'\beta + u_i + \varepsilon_{ij};\ i = 1, ..., M,\ j = 1, ..., N_i;$$
$$u_i \overset{\text{indep.}}{\sim} N(0, \sigma_u^2);\ \varepsilon_{ij} \overset{\text{indep.}}{\sim} N(0, \sigma_\varepsilon^2), \tag{2.7}$$

where M defines the number of clusters and N_i the number of units in cluster i. The model assumes also $E(u_i \varepsilon_{ij}) = 0, \forall i, j$. Notice that under this model $\mathrm{Var}(y_{ij}) = \sigma_u^2 + \sigma_\varepsilon^2$, $E(y_{ij} y_{il}) = \sigma_u^2$ for $j \neq l$ and $E(y_{ij} y_{kl}) = 0$ for $i \neq k$, implying

$$\mathrm{Corr}(y_{ij}, y_{il}) = \sigma_u^2 / (\sigma_u^2 + \sigma_\varepsilon^2) \text{ for } j \neq l;$$
$$\mathrm{Corr}(y_{ij}, y_{kl}) = 0 \text{ for } i \neq k. \tag{2.8}$$

Scott and Holt (1982) show that estimating β in (2.7) by ordinary least squares (OLS) usually results in a small loss of efficiency, compared to the use of the optimal generalized least squares (GLS) estimator. However, ignoring the intracluster correlation when estimating the variance of the OLS estimator may result in considerable variance underestimation and hence wrong size and excessive powers of test statistics and too short confidence intervals.

The results in Scott and Holt (1982) and Pfeffermann and Smith (1985) assume noninformative sampling and full response. When this is not the case, the model holding for the sample data is different from the corresponding population model, although the clustered nature of the model is preserved as we now show. Consider the following two-level population model:

$$\text{Level 1: } u_i \mid \mathrm{t}_i \sim \varphi_p(u_i \mid \mathrm{t}_i; \theta_1),\ i = 1, ..., M$$
$$\text{Level 2: } Y_{ij} \mid (u_i, \mathrm{x}_{ij}) \sim f_p(y_{ij} \mid \mathrm{x}_{ij}, u_i; \theta_2),\ j = 1, ..., N_i, \tag{2.9}$$

where φ_p and f_p denote the first and second-level *pdf*s with known covariates t_i and x_{ij}, governed by the hyperparameters θ_1 and θ_2 respectively. The model (2.7) is a special case of (2.9) by which φ_p and f_p are normal *pdf*s with $\mathrm{t}_i = 0$ (no covariates), $\theta_1 = \sigma_u^2$ and $\theta_2 = (\beta, \sigma_\varepsilon^2)$. Suppose that the sample is drawn by the following two-stage sampling process. In the first stage a sample s_1 of $m < M$ first-level units (clusters; say, schools) is selected with probabilities $\pi_i = \Pr(i \in s_1)$ that may be correlated with the random effects u_i after conditioning on the covariates t_i. In the second stage a sub-sample s_{2i} of $n_i < N_i$ second-level units (ultimate sampling units; say, pupils) is sampled from each selected first-level unit i with probabilities $\pi_{j|i} = \Pr(j \in s_{2i} \mid i \in s_1)$ that may be correlated with the outcomes y_{ij} after conditioning on the covariates x_{ij}. Denote by I_i and $\mathrm{I}_{j|i}$ the first and second-stage sampling indicators. By (2.1), the two-level sample model holding for the observed data, corresponding to the population model (2.9) is,

$$\text{Level 1:}$$
$$f_{s_1}(u_i \mid \mathrm{t}_i; \theta_1, \gamma_1) = \frac{\Pr(\mathrm{I}_i = 1 \mid u_i, \mathrm{t}_i; \gamma_1) \varphi_p(u_i \mid \mathrm{t}_i; \theta_1)}{\Pr(\mathrm{I}_i = 1 \mid \mathrm{t}_i; \theta_1, \gamma_1)}$$
$$\text{Level 2:} \tag{2.10}$$
$$f_{s_{2i}}(y_{ij} \mid \mathrm{x}_{ij}, u_i; \theta_2, \gamma_2) = \frac{\Pr(\mathrm{I}_{j|i} = 1 \mid y_{ij}, \mathrm{x}_{ij}; \gamma_2) f_p(y_{ij} \mid \mathrm{x}_{ij}, u_i; \theta_2)}{\Pr(\mathrm{I}_{j|i} = 1 \mid u_i, \mathrm{x}_{ij}; \theta_2, \gamma_2)},$$

where I assume $\Pr(\mathrm{I}_{j|i} = 1 \mid y_{ij}, u_i, \mathrm{x}_{ij}; \gamma_2) = \Pr(\mathrm{I}_{j|i} = 1 \mid y_{ij}, \mathrm{x}_{ij}; \gamma_2)$.

Remark 3. By the independence result in Remark 2, if $y_{ij} \mid u_i$ are independent under the population model, they are asymptotically independent under the sample model. Similarly, if the random effects u_i are independent under the population model, they are asymptotically independent under the sample model. Thus, the sample model (2.10) is a genuine two-level model, although with different distributions and possibly more parameters. Evidently, the models (2.9) and (2.10) are different, unless $\Pr(\mathrm{I}_{j|i} = 1 \mid y_{ij}, \mathrm{x}_{ij}) = \Pr(\mathrm{I}_{j|i} = 1 \mid u_i, \mathrm{x}_{ij})$ and $\Pr(\mathrm{I}_i = 1 \mid u_i, \mathrm{t}_i) = \Pr(\mathrm{I}_i = 1 \mid \mathrm{t}_i)$.

So far I assumed implicitly full response. Suppose, for example, that in sampled cluster (first level unit) i only a sub-sample $r_{2i} \subset s_{2i}$ respond, and denote by $R_{j|i}$ the response indicator. The second-level model for the observed outcomes is now,

Level l:

$$f_{o2i}(y_{ij} \mid \mathrm{x}_{ij}, u_i; \theta_2, \gamma_2, \gamma_2^*)$$

$$= f(y_{ij} \mid \mathrm{x}_{ij}, u_i, I_{j|i} = 1, R_{j|i} = 1)$$

$$= \frac{\Pr(R_{j|i} = 1 \mid y_{ij}, \mathrm{x}_{ij}, I_{j|i} = 1; \gamma_2^*) f_{s_{2i}}(y_{ij} \mid \mathrm{x}_{ij}, u_i; \theta_2, \gamma_2)}{\Pr(R_{j|i} = 1 \mid \mathrm{x}_{ij}, u_i, I_{j|i} = 1; \theta_2, \gamma_2, \gamma_2^*)}. \quad (2.11)$$

The *pdf* (2.11) coupled with the level 1 *pdf* in (2.10) defines the model holding for the observed data in the case of informative cluster sampling and NMAR nonresponse.

3. How can we estimate population models from complex survey data?

In this section I review the main approaches proposed in the literature to deal with the special features of complex survey data discussed in Section 2, and propose some modifications. In order to simplify the discussion, I consider the following set up used for the simulation study in Section 4.

3.1 Population model and sampling design

Consider a stratified population $U = U_1 \cup \ldots \cup U_H$ of size N. Specifically, define for every unit $j \in U$ a random vector stratification indicator $\mathrm{z}_j = (z_{1j}, \ldots, z_{Hj})'$ such that $\Pr(z_{hj} = 1) = p_h$, $\sum_{h=1}^{H} p_h = 1$ and $j \in U_h$ if $z_{hj} = 1$. The stratification is carried out independently between the units. Values of an outcome variable Y are generated as $y_j = \beta_0 + \beta_1 x_j + \alpha_0 \zeta_j + \alpha_1 \zeta_j x_j + \varepsilon_j$; $\varepsilon_j \sim N(0, \sigma^2)$, where the x_j's are fixed scalar covariates, $(\beta_0, \beta_1, \alpha_0, \alpha_1)$ are fixed coefficients and

$$\zeta_j = \frac{1}{H} \sum_{h=1}^{H} \frac{z_{hj}}{p_h} - 1.$$

Notice that ζ_j is a random variable with mean zero and variance

$$V_\zeta = \left(\frac{1}{H^2} \sum_{h=1}^{H} \frac{1}{p_h} \right) - 1,$$

implying that for given covariates x_j, x_k,

$$E_p(y_j \mid x_j) = \beta_0 + \beta_1 x_j, \ \mathrm{Var}_p(y_j \mid x_j) = (\alpha_0 + \alpha_1 x_j)^2 V_\zeta + \sigma^2, \ \mathrm{Cov}_p(y_j, y_k \mid x_j, x_k) = 0, \ j \neq k. \quad (3.1)$$

However, for unit $j \in U_h$,

$$y_j \mid x_j, \mathrm{z}_{hj} = 1 \sim N[(\beta_0 + \alpha_0 \zeta_h) + (\beta_1 + \alpha_1 \zeta_h) x_j, \sigma^2]; \ \zeta_h = [(1/Hp_h) - 1]. \quad (3.2)$$

Thus, the regression model in each stratum is the classical linear model with constant variance, but the intercepts and slopes change across the strata.

The model defined by (3.1) and (3.2) is a realistic random coefficients regression model, which I think mimics many populations encountered in practice.

We used systematic probability proportional to size (PPS) sampling within the strata for drawing the samples with the size variable defined as $z_j^* = \max\{\min[(|q_j|)^{1.5}, 9], 1\}$; $q_j \sim N(1 + x_j, 1)$. There is nothing novel about the choice of this size variable except that it allows for a clear distinction between the variance of the various estimators. This size z_j^* does not depend on the outcome y_j, and hence the sampling process within each stratum is non-informative. However for disproportionate allocation of the sample between the strata, the sampling scheme is informative because of the different models operating in different strata, such that the observed outcomes carry information on the strata membership and $\Pr(j \in s \mid y_j, x_j) \neq \Pr(j \in s \mid x_j)$. We focus on the estimation of the regression coefficients (β_0, β_1) in (3.1) as the target of inference and assume that the available sample information consists of the observed outcomes and covariates, the strata membership vectors z_{hj} and the strata sizes, $\{N_h\}$.

3.2 Including the design variables among the covariates

As implied by (2.3), the population model (*pdf*), $f_p(\mathrm{y}_s \mid \mathrm{x}_s)$ and the sample model $f_s(\mathrm{y}_s \mid \mathrm{x}_s)$ are the same if $\Pr(A_s = 1 \mid \mathrm{y}_s, \mathrm{x}_s) = \Pr(A_s = 1 \mid \mathrm{x}_s) \forall \mathrm{y}_s$. By (2.2), the response process is ignorable if $\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1) = \Pr(R_i = 1 \mid \mathrm{x}_i, I_i = 1) \forall y_i$. Thus, a possible

way to account for the sampling and response effects is to add to the model covariates all the variables and interactions determining the sample and response probabilities and then integrate them out in order to estimate the model of interest. Denote these variables by $J = Z \cup L$ with population values J_U, where L defines the variables explaining the response probabilities. Assuming $f_p(\mathrm{y}_s|\mathrm{x}_U, j_U) = f_p(\mathrm{y}_s|\mathrm{x}_s, j_U)$, the use of this approach requires to fit first the model

$$f_p(\mathrm{y}_s \mid \mathrm{x}_s, J_U = j_U) = \int f_p(\mathrm{y}_s, \mathrm{y}_{\tilde{s}} \mid \mathrm{x}_U, j_U)\, d\,\mathrm{y}_{\tilde{s}}, \quad (3.3)$$

and then integrate,

$$f_p(\mathrm{y}_s \mid \mathrm{x}_s) = \int f_p(\mathrm{y}_s \mid \mathrm{x}_s, j_U)\, f_p(j_U \mid \mathrm{x}_s)\, dj_U. \quad (3.4)$$

Variants of this approach can be found in DeMets and Halperin (1977), Holt, Smith and Winter (1980), Nathan and Holt (1980), Jewell (1985), Skinner (1994), Chambers and Skinner (2003, Chapter 2) and Gelman (2007).

The use of the approach is appealing, and it has the advantage of allowing classical model based inference procedures once the variables $J_U = Z_U \cup L_U$ are included in the model, but it is often limited in practice for the following reasons:

1. It requires knowledge of the population values of all the variables determining the sample selection and response, and this information is usually unknown to the analyst fitting the model because of confidentiality restrictions or other reasons. Even if known, including in the model all the geographic and operational variables used for the sampling design and the variables explaining the response may be formidable.
2. In practice there may be many covariates and many design variables, and modelling the relationship between the design variables and the covariates in order to integrate out the effect of the former variables can be complicated and may no longer reproduce the original target model.

Feder (2011) proposes the following simple solution to this problem. Suppose first that the design variables and the covariates are known for every element in the population. The proposed solution consists of imputing the missing population outcomes using the model $f_p(\mathrm{y}_s \mid \mathrm{x}_s, J_U = j_U)$ fitted to the sample data, and then fitting the population model $f_p(y_j \mid \mathrm{x}_j)$ using all the population values, with the missing outcomes replaced by their imputed values. When the design variables and the covariates are unknown for the non-sampled units, they need to be imputed as well. The imputation may be carried out by sampling with replacement $(N-n)$ values (x_i, z_i) from the sample values with probabilities $\vec{p}_i = (w_i - 1)/\sum_{k=1}^{n}(w_k - 1)$ on each draw, where the w_i's are the sampling weights. See Pfeffermann and Sikov (2011) for justification of this procedure under the sample model and an extension for the case of NMAR nonresponse.

3. The approach is not operational when the inclusion in the sample depends also on the outcome values, that is, $Z_U = \{Y_U, Z_U^*\}$ and $\Pr(A_s = 1 \mid Y_U, X_U, Z_U^*) \neq \Pr(A_s = 1 \mid X_U, Z_U^*)$. A classical example is *case-control studies* (Scott and Wild 2009), but a similar problem arises when the nonresponse is NMAR.

Remark 4. Including the design variables and the variables explaining the response in the model does not necessarily require integrating them out even if they are not part of the covariates of interest, as the following example shows.

Example 4: Suppose that a sample of size n is selected with probabilities defined by the population values of design variables Z and that all the sampled units respond. Let the population distribution of Y, X, Z be multivariate normal. The data available to the analyst consist of the sample values $[\mathrm{y}_s, \mathrm{x}_s]$ and the population values Z_U. Using properties of the multivariate normal distribution, $E_p(\mathrm{y} \mid x) = \beta_0 + \beta_{yx}x$ for some coefficients (β_0, β_{yx}), but the OLS estimate of β_{yx} is biased because the sampling probabilities depend on Z, which is correlated with Y. The *mle* of β_{yx} for the case of a trivariate normal distribution is (DeMets and Halperin 1977),

$$\hat{\beta}_{yx} = \left\{ s_{xy} + \frac{s_{yz}s_{xz}}{s_{zz}}\left(\frac{\hat{\sigma}_z^2}{s_{zz}} - 1\right)\right\} \Bigg/ \left\{ s_{xx} + \frac{s_{xz}^2}{s_{zz}}\left(\frac{\hat{\sigma}_z^2}{s_{zz}} - 1\right)\right\}, \quad (3.5)$$

where $s_{uv} = n^{-1}\sum_{i=1}^{n}(u_i - \bar{u}_s)(v_i - \bar{v}_s)$ and $\hat{\sigma}_z^2 = N^{-1}\sum_{i=1}^{N}(z_i - \bar{z}_U)^2$, with $\bar{u}_s$, $\bar{v}_s$ and $\bar{z}_U$ defining the corresponding sample and population means. Thus, the population values of Z feature in this case in the optimal estimator of the target parameter β_{yx}. Holt *et al.* (1980) extend this result to the case where Y, X, Z are vector variables. Nathan and Holt (1980) establish conditions under which $\hat{\beta}_{yx}$ is consistent without the multivariate normality assumptions. Pfeffermann and Holmes (1985) study the robustness of the estimator to model misspecification.

3.3 Using the sampling weights as surrogate for the design variables

For situations where there are too many design variables determining the sample selection to include them all in the model, or when some or all of these variables are unknown to the analyst, it is often advocated to include in the model the sampling weights as surrogate of the design variables. Examples of the use of this approach can be found in DuMouchel and Duncan (1983), Särndal and Wright

(1984), Rubin (1985), Chambers, Dorfman and Wang (1998) and Wu and Fuller (2006).

Rubin (1985) defines the vector $a = (a_1, ..., a_N)' = a(Z_U)$ to be an adequate summary of Z_U if $\Pr(A_s = 1 \mid Z_U) = \Pr(A_s = 1 \mid a)$. The author shows that the vector $\pi_U = (\pi_1, ..., \pi_N)$ of the sample inclusion probabilities is the coarsest possible adequate summary of Z_U, though it may be too coarse. It follows therefore that for sampling designs such that $\Pr(A_s = 1 \mid Y_U, Z_U) = \Pr(A_s = 1 \mid Z_U)$, if π_U is an adequate summary, the sample selection can be ignored for inference on the parameters of $f_p(y_s \mid \mathrm{x}_s, \pi_U)$. In order to estimate the target model $f_p(y \mid \mathrm{x})$ in this case, one can follow the same steps as in Section (3.2) with π_U taking the role of Z_U.

The use of this approach reduces the dimension of the added covariates but it requires knowledge of the sample inclusion probabilities (or the sampling weights) for all the population units, which may not be available in the case of a secondary analysis. The case of nonresponse is particularly problematic since the response probabilities are generally unknown and need to be estimated. Another major problem with this approach is that for general sampling designs, the vector π_U may not be an adequate summary of Z. Sugden and Smith (1984) and Smith (1988) establish necessary design information required for sampling ignorability.

Remark 5. Even though the vector π_U is not always an adequate summary of Z_U, for sampling designs such that $\Pr(I_i = 1 \mid y_i, \mathrm{x}_i, \pi_i) = \pi_i$, $f_s(y_i \mid \mathrm{x}_i, \pi_i) = f_p(y_i \mid \mathrm{x}_i, \pi_i)$, so that the marginal population and sample *pdf*s for a given sampled unit are nonetheless the same when adding π_i to the covariates (see Skinner 1994).

Remark 6. In the empirical set up described in Section 3.1 there is a one to one correspondence between the design variables (z'_j, z^*_j) and the sampling weights (w_h, w_j).

3.4 Methods based on probability weighting

So far we considered methods requiring knowledge of the variables J determining the sample selection and response probabilities, or at least an adequate summary of them. The methods considered below only require knowledge of the sampling weights for the responding sampled units. As such, they are restricted to situations of full response, or to cases where the response probabilities can be estimated sufficiently accurately, in which case the sampling weight for a responding unit is the inverse of the product of the unit's selection probability and its estimated response probability. Probability weighting (PW) is discussed in numerous articles; see the recent discussion in Pfeffermann and Sverchkov (2009) and the references therein. As before, we focus here on estimation of population models.

To introduce the idea, consider the case of a *census* with full response. Assuming independent outcomes, the model parameters, θ, are typically estimated in this case by solving *census* estimating equations of the form,

$$\sum_{j=1}^{N} u(y_j, \mathrm{x}_j; \theta) = 0. \qquad (3.6)$$

In the case of *mle*, $u(y_j, \mathrm{x}_j; \theta) = (\partial/\partial\theta)\log f_p(\mathrm{y}_j \mid \mathrm{x}_j; \theta)$, the j^{th} score. In practice, data are available for only a sample $s \subset U$ and the equations (3.6) are replaced by their randomization unbiased Horvitz-Thompson estimator,

$$\sum_{i \in s} w_i\, u(y_i, \mathrm{x}_i; \theta) = 0, \qquad (3.7)$$

where the w_i 's are the sampling weights.

Remark 7. When the census estimating equations (3.6) are the likelihood equations, the estimators obtained by solving (3.7) are known in the sampling literature as 'pseudo *mle*' (*pmle*). See Binder (1983), Skinner *et al*. (1989), Pfeffermann (1993, 1996) and Godambe and Thompson (2009) for discussion with many examples. This approach is implemented in many software packages such as SAS, STATA, SUDAAN, *etc*.

Example 5. In the case of the standard linear regression model, the *pmle* or PW estimator of the vector coefficient β solves the equations $\sum_{i \in s} w_i (y_i - \mathrm{x}'_i \hat{\beta}_{\mathrm{pw}})\mathrm{x}_i = 0$;

$$\hat{\beta}_{\mathrm{pw}} = \Big[\sum_{i \in s} w_i \mathrm{x}_i \mathrm{x}'_i\Big]^{-1} \sum_{i \in s} w_i \mathrm{x}_i y_i. \qquad (3.8)$$

The PW estimator of the residual variance is $\hat{\sigma}^2_{\mathrm{pw}} = \sum_{i \in s} w_i (y_i - \mathrm{x}'_i \hat{\beta}_{\mathrm{pw}})^2 / (\sum_{i \in s} w_i - k)$, where $k = \dim(\beta)$.

For logistic regression, the pseudo likelihood equations (with no explicit solution) are,

$$\sum_{i \in s} w_i [y_i - \tilde{p}_i(\mathrm{x}_i)]\, \mathrm{x}_i = 0;\ \tilde{p}_i(\mathrm{x}_i) = \Pr_p(y_i = 1 \mid \mathrm{x}_i) = \exp(\mathrm{x}'_i\beta)/[1+\exp(\mathrm{x}'_i\beta)]. \qquad (3.9)$$

Example 6. Let $u(y_j; \theta) = [\Delta(\theta - y_j) - F_p(\theta)]$ where $F_p(\theta)$ is the cumulative population distribution at θ and $\Delta(a) = 1(0)$ when $a \geq 0$ ($a < 0$). The PW estimator of $F_p(\theta)$ is $\hat{F}_{p,\,\mathrm{pw}}(\theta) = \sum_{i \in s} w_i \Delta(\theta - y_i) / \sum_{i \in} w_i$, the familiar Hàjek (1971) estimator.

The notable property of PW estimators is that they are generally $r - p$ consistent. (See Section 2.2 for definition of the $r - p$ distribution). This can be seen by decomposing $(\hat{\theta}_{\mathrm{pw}} - \theta) = (\hat{\theta}_{\mathrm{pw}} - \hat{\theta}_{\mathrm{cen}}) + (\hat{\theta}_{\mathrm{cen}} - \theta)$, where $\hat{\theta}_{\mathrm{cen}}$ is the (hypothetical) solution of the census equations (3.6). Under general conditions, $(\hat{\theta}_{\mathrm{pw}} - \hat{\theta}_{\mathrm{cen}}) = O_p(n^{-0.5})$ and $(\hat{\theta}_{\mathrm{cen}} - \theta) = O_p(N^{-0.5})$, thus establishing the $r - p$ consistency of $\hat{\theta}_{\mathrm{cen}}$ under these conditions. The $r - p$ variance of $\hat{\theta}_{\mathrm{pw}}$ can be decomposed as,

$$\mathrm{Var}_{r-p}(\hat{\theta}_{\mathrm{pw}}) = E_p[\mathrm{Var}_r(\hat{\theta}_{\mathrm{pw}})] + \mathrm{Var}_p[E_r(\hat{\theta}_{\mathrm{pw}})]. \quad (3.10)$$

For single stage sampling, if n is much smaller than N as is usually the case, the second term on the right hand side of (3.10) is negligible compared to the first term, and $\mathrm{Var}_{r-p}(\hat{\theta}_{\mathrm{pw}})$ can be estimated by the randomization variance estimator $\hat{\mathrm{Var}}_r(\hat{\theta}_{\mathrm{pw}})$. This result does not necessarily hold for cluster sampling since in this case $\mathrm{Var}_r(\hat{\theta}_{\mathrm{pw}})$ is typically of order $O(1/m)$ where m is the number of sampled clusters, and under a suitable model $\mathrm{Var}_p[E_r(\hat{\theta}_{\mathrm{pw}})]$ is $O(1/M)$ where M is the number of population clusters. For $\hat{\mathrm{Var}}_r(\hat{\theta}_{\mathrm{pw}})$ to be an adequate estimator of $\mathrm{Var}_{r-p}(\hat{\theta}_{\mathrm{pw}})$ in this case, m must be much smaller than M.

Remark 8. The consistency of PW estimators under correct population model specification may also be established under the sample distribution (Equation 2.1). Consider the estimator $\hat{\beta}_{\mathrm{pw}}$ in (3.8) and write $\hat{\beta}_{\mathrm{pw}} = \beta + [\sum_{i\in s} w_i \mathrm{x}_i \mathrm{x}_i']^{-1} \sum_{i\in s} \mathrm{x}_i w_i \varepsilon_i$ where the ε_i's are the population model residuals. The key result leading to the consistency of $\hat{\beta}_{\mathrm{pw}}$ under the sample distribution is that if $\Pr(I_i = 1 | y_i, \mathrm{x}_i, \pi_i) = \pi_i$ then $E_s(w_i\varepsilon_i) = E_s(w_i)E_p(\varepsilon_i) = 0$ (follows from 3.14 below). In fact, by viewing the covariates as random with (y_i, x_i) having some joint distribution,

$$\beta = \arg\min_{\tilde{\beta}} E_p(y_i - \mathrm{x}_i'\tilde{\beta})^2 = \arg\min_{\tilde{\beta}} E_s[w_i(y_i - \mathrm{x}_i'\tilde{\beta})^2],$$

implying that $\hat{\beta}_{\mathrm{pw}}$ is the optimal estimator (in weighted least-squares metric) of β under the sample distribution of (y_i, x_i). See also (3.24) below. Godambe and Thompson (1986, 2009) establish and discuss other optimality properties of estimators solving estimating equations of the form $\sum_{i\in s} w_i u(y_i, \mathrm{x}_i; \theta) = 0$. The following example shows how probability weighting can be used when modelling clustered populations.

Example 7. Consider the population two-level (random intercept) model,

Level 1:

$$u_i \sim N(\mathrm{t}_i'\gamma, \sigma_u^2),\ i = 1,...,M \quad (3.11)$$

Level 2:

$$y_{ij} = \mathrm{x}_{ij}'\beta + u_i + \varepsilon_{ij},\ \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2),\ j = 1...N_i$$

where ε_{ij} and u_i are independent for all i and j. The unknown parameters are the vectors of coefficients $\vartheta = (\beta', \gamma')'$ and the variances $\tau = (\sigma_\varepsilon^2, \sigma_u^2)'$. Assume full response. Under ignorable sampling of first and second-level units, the *mle* of (ϑ, τ) is computed conveniently by iterating between the estimation of ϑ for 'known' τ and the estimation of τ for 'known' ϑ, with the 'known' values defined by the estimators from the previous iteration. The two sets of estimators on the r^{th} iteration are the solutions of linear equations of the form, $P^{(r)}\vartheta = q^{(r)}$, $R^{(r)}\tau = s^{(r)}$, with appropriate definition of the matrices $(P^{(r)}, R^{(r)})$ and the vectors $(q^{(r)}, s^{(r)})$, $r = 1, 2, ...,$ (Goldstein 1986). When applied to all the population values, these equations define the *census* estimating equations.

Suppose, as before, that a sample s_1 of first-level units is sampled with probabilities $\pi_i = \Pr(i \in s_1)$, and that subsamples s_{2i} of size $n_i < N_i$ are sampled from each selected first-level unit i with probabilities $\pi_{j|i} = \Pr(j \in s_{2i} \,|\, i \in s_1)$. The *pmle* for this model can be obtained by first expressing the elements of the matrices $(P^{(r)}, R^{(r)})$ and the vectors $(q^{(r)}, s^{(r)})$ as sums over first and second-level units, and then estimating each population sum of the form $\sum_{i=1}^{M} d_i$ by the H-T estimator $\sum_{i\in s_1}(d_i/\pi_i)$, and each population sum of the form $\sum_{j=1}^{N_i} d_{ij}$ by the H-T estimator $\sum_{j\in s_{2i}}(d_{ij}/\pi_{j|i})$. See Pfeffermann, Skinner, Holmes, Goldstein and Rasbash (1998b). Pfeffermann and Sverchkov (2009) review other methods of probability weighting in two-level models.

Probability weighting is in broad use both for estimation of finite-population quantities, referred to in the literature as descriptive inference, and for 'analytic inference' on population models. The main attraction of this method is its simplicity. It is generally viewed as being 'model free', except when having to estimate the response probabilities, which is often based on models, and hence more robust than other methods, but when used for analytical inference, this view is questionable.

Probability-weighted estimators are randomization consistent for the corresponding descriptive population quantities (CDPQ), defined as the (hypothetical) solutions of the census estimating equations. However, if the population model is misspecified, the target CDPQ are not (model) p-consistent for the true model parameters and the PW estimators are not $r - p$ consistent either. So, probability weighting provides no protection against model misspecification, although the estimated CDPQ may be useful for various kinds of inference. See Pfeffermann (1993) and Binder and Roberts (2009) for discussion and examples.

Estimating the randomization variance of probability-weighted estimators is generally simple, utilizing available techniques in finite population sampling. Binder (1983) developed a general approach for estimating the randomization variance of estimators obtained as the solution of probability-weighted estimating equations; see also Binder and Roberts (2009) and Godambe and Thompson (2009). Fuller (1975), Binder (1983), Chambless and Boyle (1985) and Francisco and Fuller (1991) developed central limit theorems applicable to probability-weighted estimators.

In spite of these desirable properties of probability-weighting, the method has some severe limitations:

1. It is restricted mostly to point estimation. Probabilistic inference like confidence intervals or

hypothesis testing generally requires large sample normality assumptions. In particular, the randomization distribution does not lend itself to the use of classical inference methods such as likelihood-based or Bayesian inference.

2. The variances of probability-weighted estimators are computed with respect to the randomization distribution and the use of this approach does not permit conditioning on the selected sample, for example, conditioning on the observed covariates or the selected clusters in a multi-level model.
3. As often illustrated in the literature, probability-weighted estimators generally have larger variances than model-based estimators, notably for small samples and large variation of the sampling weights.
4. The use of the randomization distribution does not lend itself to prediction problems such as the prediction of the outcome for non-sampled units with known covariates under a regression model, or the prediction of small area means for areas with no samples in a small-area estimation problem.

3.5 Modifications of the sampling weights

When estimating finite population quantities, the sampling weights are often modified by imposing calibration equations, which match the PW estimators of covariates for which the population totals are known with the actual totals. The use of calibration is particularly useful in the case of nonresponse; see Kott (2009) for recent discussion with references. We later discuss the use of *empirical likelihood* for analytical inference on population models, which also attempts to incorporate calibration equations, although in a different manner. Below, I review two modifications of the sampling weights aimed at reducing the variances of the weighted estimators of model parameters under the *sample distribution* (2.1). A combination of the two modifications is also considered.

Magee (1998) considers a linear regression model but the results can be extended to other population models. The author shows that under certain moment assumptions, any estimator $\hat{\beta}_{\rm mg}(a) = [\sum_{i\in s} w_i\, a_i(\alpha)\, \mathrm{x}_i \mathrm{x}_i']^{-1} \sum_{i\in s} w_i\, a_i(\alpha)\, \mathrm{x}_i y_i$ with positive weights $a_i(\alpha) = a(\mathrm{x}_i, \alpha)$ is consistent for β under the sample distribution. The weights $a(\mathrm{x}_i, \alpha)$ belong to a parameterized family of functions with the vector parameter α chosen to minimize a scalar variance criterion such as the determinant or the trace of the asymptotic variance estimator,

$$
\begin{aligned}
&A\hat{\mathrm{var}}[\hat{\beta}_{\rm mg}(a)] \\
&= \Big[\sum_{i\in s} w_i\, a_i(\alpha) \mathrm{x}_i \mathrm{x}_i'\Big]^{-1} \sum_{i\in s} w_i^2 a_i^2(\alpha)\, \hat{\varepsilon}_i^2\, \mathrm{x}_i \mathrm{x}_i' \\
&\quad \Big[\sum_{i\in s} w_i\, a_i(\alpha) \mathrm{x}_i \mathrm{x}_i'\Big]^{-1}, \qquad (3.12)
\end{aligned}
$$

where $\hat{\varepsilon}_i = (y_i - \mathrm{x}_i' \hat{\beta}_{\rm pw})$. The choice of the function $a(\mathrm{x}_i, \alpha)$ is up to the analyst but the obvious idea is to choose a function that is believed to be approximately inversely proportional to the residual variance under the sample model. The resulting 'Quasi-Aitken' estimator is shown to have asymptotically a lower variance under the sample distribution than the probability-weighted estimator $\hat{\beta}_{\rm pw}$. Recall from Remark 8 that $\hat{\beta}_{\rm pw}$ is consistent for β under the sample distribution, justifying comparing the asymptotic variances of the two estimators under this distribution.

Pfeffermann and Sverchkov (1999) propose another modification. Consider the population model,

$$
y_j = m(\mathrm{x}_j; \theta) + \varepsilon_j,\ E_p(\varepsilon_j \mid \mathrm{x}_j) = 0,\ E_p(\varepsilon_j^2 \mid \mathrm{x}_j) = \sigma^2, \qquad (3.13)
$$

where $m(\mathrm{x}_j; \theta)$ has a known form. Let $q_i = w_i / E_s(w_i \mid \mathrm{x}_i)$. The authors show that if $\Pr(I_i = 1 \mid \pi_i, y_i, \mathrm{x}_i) = \pi_i$,

$$
E_p(y_i \mid \mathrm{x}_i) = E_s(w_i y_i \mid \mathrm{x}_i) / E_s(w_i \mid \mathrm{x}_i). \qquad (3.14)
$$

Thus, for vectors $\tilde{\theta}$ in the plausible parameter space Θ,

$$
\begin{aligned}
\theta &= \underset{\tilde{\theta}}{\mathrm{argmin}} \frac{1}{n} \sum_{i\in s} E_p\{[y_i - m(\mathrm{x}_i; \tilde{\theta})]^2 \mid \mathrm{x}_i\} \\
&= \underset{\tilde{\theta}}{\mathrm{argmin}} \frac{1}{n} \sum_{i\in s} E_s\{q_i [y_i - m(\mathrm{x}_i; \tilde{\theta})]^2 \mid \mathrm{x}_i\}.
\end{aligned}
$$

The vector θ can be estimated therefore by solving the minimization problem,

$$
\begin{aligned}
\hat{\theta}_q &= \arg\min_{\tilde{\theta}} \frac{1}{n} \sum_{i=1}^{n} \hat{q}_i [y_i - m(\mathrm{x}_i; \tilde{\theta})]^2; \\
\hat{q}_i &= w_i / \hat{E}_s(w_i \mid \mathrm{x}_i). \qquad (3.15)
\end{aligned}
$$

The use of this estimator requires estimating $E_s(w_i \mid \mathrm{x}_i)$ but under mild regularity conditions $\hat{\theta}_q$ is consistent for θ even when the expectation $E_s(w_i \mid \mathrm{x}_i)$ is misspecified. See Pfeffermann and Sverchkov (2009) and Section 4.1 of this paper for examples of the specification and estimation of $E_s(w_i \mid \mathrm{x}_i)$.

Example 8. Under the linear regression population model with constant variance,

$$\hat{\beta}_q = \left[\sum_{i\in s} \hat{q}_i \mathrm{x}_i \mathrm{x}_i'\right]^{-1} \sum_{i\in s} \hat{q}_i \mathrm{x}_i y_i. \qquad (3.16)$$

As easily verified, $\hat{\beta}_q$ is randomization consistent for the census regression coefficients $\tilde{B} = [\sum_{j=1}^N \mathrm{x}_j \mathrm{x}_j' / E_s(w_j | \mathrm{x}_j)]^{-1} \sum_{j=1}^N \mathrm{x}_j y_j / E_s(w_j | \mathrm{x}_j)$, and hence $p-r$ consistent for β, even when $E_s(w_i | \mathrm{x}_i)$ is misspecified.

The obvious difference between the PW estimator $\hat{\theta}_{\mathrm{pw}}$ and the estimator $\hat{\theta}_q$ is that the latter estimator uses the adjusted weights $q_i = w_i / \hat{E}_s(w_i | \mathrm{x}_i)$. When the sample selection depends only on the covariates, the sampling process is ignorable. Hence, to protect against informative sampling, it is only necessary to account for the net sampling effects on the target conditional *pdf* of $y_i | \mathrm{x}_i$. This is achieved by using the weights q_i. In contrast, the sampling weights w_i account for the sampling effects on the joint distribution of (y_i, x_i). As a result, they tend to be more variable and the estimator $\hat{\theta}_{\mathrm{pw}}$ has a larger variance.

A combination of the last two modifications is also possible and examined in Section 4. The simple idea proposed by Dr. Moshe Feder (private communication) is to apply the modification of Magee (1998) to the estimator $\hat{\beta}_q$ instead of the estimator $\hat{\beta}_{\mathrm{pw}}$, that is, use the estimator,

$$\hat{\beta}_{\mathrm{mg}-q}(a) = \left[\sum_{i\in s} \hat{q}_i \, a_{i,q}(\alpha) \mathrm{x}_i \mathrm{x}_i'\right]^{-1} \sum_{i\in s} \hat{q}_i \, a_{i,q}(\alpha) \mathrm{x}_i y_i, \qquad (3.17)$$

where the vector parameter α is now chosen to minimize a scalar variance criterion of the asymptotic variance estimator, $A\hat{\mathrm{var}}[\hat{\beta}_{\mathrm{mg}-q}(a)]$, computed similarly to (3.12).

3.6 Likelihood based methods

3.6.1 Use of the sample model for maximum likelihood estimation

A natural way of estimating the population model parameters is by maximization of the sample likelihood. Assume first full response and that the sample observations are independent under the sample distribution. The likelihood has then the form,

$$L_s(\theta, \gamma; y_s, \mathrm{x}_s) = \prod_{i\in s} \frac{\Pr(I_i = 1 | \mathrm{x}_i, y_i; \gamma) \, f_p(y_i | \mathrm{x}_i; \theta)}{\Pr(I_i = 1 | \mathrm{x}_i; \gamma, \theta)}. \qquad (3.18)$$

As before, we assume $\Pr(I_i = 1 | \pi_i, y_i, \mathrm{x}_i) = \pi_i$, implying $\Pr(I_i = 1 | \mathrm{x}_i, y_i) = E_p(\pi_i | \mathrm{x}_i, y_i)$. By (3.14), The sample likelihood can be written therefore as,

$$L_s(\theta, \gamma; y_s, \mathrm{x}_s) = \prod_{i\in s} \frac{E_s(w_i | \mathrm{x}_i; \theta, \gamma) f_p(y_i | \mathrm{x}_i; \theta)}{E_s(w_i | y_i, \mathrm{x}_i; \gamma)}. \qquad (3.19)$$

The expectations on the right hand side of (3.19) are with respect to the sample *pdf* of the sampling weights. Thus, when the weights are known for the sampled units as is usually the case under full response, the expectations can be modelled and estimated by regressing w_i against (y_i, x_i), using classical model fitting procedures. Suppose first that the weights are continuous such as in probability proportional to size (PPS) sampling with a continuous size variable. For a given form of the population model, the expectations $E_s(w_i | y_i, \mathrm{x}_i; \gamma)$ and $E_s(w_i | \mathrm{x}_i; \gamma, \theta)$ can be obtained then in two steps:

1. Identify and estimate $\hat{E}_s(w_i | y_i, \mathrm{x}_i; \gamma) = E_s(w_i | y_i, \mathrm{x}_i; \hat{\gamma})$, using the sample data.
2. Integrate $\int [1/E_s(w_i | y, \mathrm{x}_i; \hat{\gamma})] f_p(y | \mathrm{x}_i; \theta) dy$ to obtain $E_p(\pi_i | \mathrm{x}_i; \theta; \hat{\gamma})$. Compute, $\hat{E}_s(w_i | \mathrm{x}_i; \theta, \hat{\gamma}) = 1/E_p(\pi_i | \mathrm{x}_i; \theta, \hat{\gamma})$ (follows from 3.14).

Estimating the vector parameter γ outside the likelihood and then substituting the estimate in (3.19) and maximizing the likelihood as a function of the vector parameter θ only, usually yields more stable results than maximizing the likelihood over (θ, γ) simultaneously.

Estimation of the expectations $E_s(w_i | y_i, \mathrm{x}_i; \gamma)$ and $E_s(w_i | \mathrm{x}_i; \theta, \gamma)$ in the case of discrete inclusion probabilities is similar.

Example 9. Consider the case of multinomial-logistic regression with a discrete covariate x and M possible values of the outcome y. Assuming that $E_s(w_i | y_i = m, x_i = k)$ is not a function of the model parameters, it can be estimated by $\overline{w}_{mk}$, the mean of the weights in cell (m, k), and thence $\hat{\pi}_{mk} = \hat{\Pr}_p(i \in s | y_i = m, x_i = k) = (1 / \overline{w}_{mk})$. We obtain:

$$\Pr_s(y_i = m | x_i = k; \theta) \cong \frac{[\Pr_p(y_i = m | x_i = k; \theta) / \overline{w}_{mk}]}{\sum_{m^*=1}^{M} [\Pr_p(y_i = m^* | x_i = k; \theta) / \overline{w}_{m^*k}]}. \qquad (3.20)$$

The sampling weights feature in the sample model, but this is not an application of classical probability weighting. Notice that with this approximation the parameters in the population and the sample model are the same. In our empirical study we use a similar approximation for the sample distribution by categorizing the values of a continuous outcome. See Pfeffermann and Sverchkov (1999) for other examples.

Next consider the estimation of the vector parameter θ governing the population model. Under mild conditions, θ is the unique solution of the equations,

$$W_U(\theta) = \sum_{j\in U} E_p(\delta_j | \mathrm{x}_j) = 0;$$

$$\delta_j = (\delta_{j,0}, \delta_{j,1}, ..., \delta_{j,k})' = \partial \log f_p(y_j | \mathrm{x}_j; \theta) / \partial\theta. \qquad (3.21)$$

Pfeffermann and Sverchkov (2003) consider three different approaches for estimating θ. The common feature of these

approaches is that the only data used for estimation are the observations $\{(y_i, \mathrm{x}_i, w_i), i \in s\}$, similarly to the PW estimators and their modifications considered in Section 3.5. In Section 3.6.2 we consider the use of the 'full likelihood', which assumes knowledge of the covariates $\{\mathrm{x}_j, j \in U\}$, and possibly also additional design information.

The first approach redefines the parameter equations with respect to the sample model. Assuming that $E_s(w_i \mid \mathrm{x}_i; \theta, \gamma)$ in (3.19) is differentiable with respect to θ, the sample model parameter equations are $W_{1s}(\theta) = \sum_{i\in s} E_s \{[\partial \log f_s(y_i \mid \mathrm{x}_i; \theta, \gamma)/\partial\theta] \mid \mathrm{x}_i\} = \sum_{i\in s} E_s\{[\delta_i + \partial \log E_s(w_i \mid \mathrm{x}_i; \theta, \gamma)/\partial\theta] \mid \mathrm{x}_i\} = 0$. The vector θ is estimated under this approach by solving the equations,

$$W_{1s,e}(\theta) = \sum_{i\in s}[\delta_i + \partial \log E_s(w_i \mid \mathrm{x}_i; \theta, \gamma)/\partial\theta] = 0. \quad (3.22)$$

The second approach applies the relationship (3.14) to the parameter equations (3.21). For a random sample from the sample model, the equations are now $W_{2s}(\theta) = \sum_{i\in s} E_s(q_i \delta_i \mid \mathrm{x}_i) = 0$, where $q_i = w_i / E_s(w_i \mid \mathrm{x}_i)$. The vector θ is estimated under this approach by solving the equations,

$$W_{2s,e}(\theta) = \sum_{i\in s} q_i \delta_i = 0. \quad (3.23)$$

The third approach uses the property that if θ solves (3.21), then it solves also the equations, $\tilde{W}_U(\theta) = \sum_{j\in U} E_p(\delta_j) = E_{\mathrm{x}}[\sum_{j\in U} E_p(\delta_j \mid \mathrm{x}_j)] = 0$, where $E_{\mathrm{x}}(\cdot)$ is the expectation of x (which is viewed as random) with respect to the population distribution. Hence, by (3.14), for a random sample from the sample model, the parameter equations are $W_{3s}(\theta) = \sum_{i\in s} E_s(w_i \delta_i) = 0$, with estimating equations,

$$W_{3s,e}(\theta) = \sum_{i\in s} w_i \delta_i = 0. \quad (3.24)$$

Note that the equations (3.24) are the *pseudo-likelihood* equations (Remark 7).

Remark 9. The use of the weights $q_i = w_i / E_s(w_i \mid \mathrm{x}_i)$ for population model parameter estimation has been justified already in Section 3.5 by reference to least-squares estimation. See the discussion in that section regarding the difference between the use of the weights q_i and the weights w_i. Pfeffermann and Sverchkov (1999, 2003) illustrate that estimating θ by solving the equations (3.23) yields estimators with lower randomization variance than estimating θ by solving the equations (3.24). Notice that under the assumption of a linear regression model operating in the population, the solution of (3.24) yields the PW estimator (3.8), and the solution of (3.23) yields the *q*-weighted estimator (3.16).

Remark 10. The use of the sample model for estimation of multi-level population models is considered in Pfeffermann, Moura and Nascimento-Silva (2006), using the Bayesian approach. Pfeffermann and Sverchkov (2007) fit multi-level models for small area estimation under informative sampling of areas and within the areas, following the frequentist approach.

So far we assumed full response. Next consider the case of NMAR nonresponse. In this case the response process needs to be modelled as well. By (2.2) and with added parameter notation the 'respondents' likelihood takes the form,

$$L_o = \prod_{i=1}^{r} f(y_i \mid \mathrm{x}_i, I_i = 1, R_i = 1; \theta^*, \gamma^*)$$

$$= \prod_{i=1}^{r} \frac{\Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1; \gamma^*) f_s(y_i \mid x_i; \theta^*)}{\Pr(R_i = 1 \mid \mathrm{x}_i, I_i = 1; \gamma^*, \theta^*)}, \quad (3.25)$$

where $\theta^* = (\theta, \gamma)$ represents the parameters of the sample distribution under full response (Equation 3.19), and γ^* represents the parameters of the response process. Notice that unlike the sampling probabilities $\pi_i = \Pr(i \in s)$, which are generally known and can be used for estimating the probabilities $\Pr(I_i = 1 \mid y_i, \mathrm{x}_i; \gamma)$ as explained before, the response probabilities are generally unknown.

Chang and Kott (2008) propose a method of estimating the response probabilities, which uses known totals of calibration variables. The authors assume a parametric model for the response probabilities that may depend on the outcome value, and estimate the unknown parameters of this model by regressing the totals of the calibration variables against their H-T estimators. The weights used for the H-T estimators are the product of the sampling weights and the inverse of the response probabilities under the model. Let c_i define the values of the calibration variables for unit i and denote $p(y_i, \mathrm{x}_i; \gamma^*) = \Pr(R_i = 1 \mid y_i, \mathrm{x}_i, I_i = 1; \gamma^*)$. Chang and Kott (2008) estimate the unknown parameters by setting the nonlinear regression equations,

$$C^U = \sum_{i=1}^{r} w_i \frac{c_i}{p(y_i, \mathrm{x}_i; \gamma^*)} + \varepsilon^*,$$

where $C^U = \sum_{j=1}^{N} c_j$ and ε^* is a vector of errors. The parameters γ^* are estimated by the iterative algorithm

$$\hat{\gamma}^{(j+1)} = \hat{\gamma}^{(j)} + \left\{\hat{H}(\hat{\gamma}^{(j)})^T V^{-1}(\hat{\gamma}^{(j)}) \hat{H}(\hat{\gamma}^{(j)})\right\}^{-1}$$

$$\hat{H}(\hat{\gamma}^{(j)})^T V^{-1}(\hat{\gamma}^{(j)}) \left(C^U - \sum_{i=1}^{r} w_i \frac{c_i}{\pi(y_i, v_i; \hat{\gamma}^{(j)})} \right), \quad (3.26)$$

where

$$\hat{H}(\hat{\gamma}^{(j)}) = \frac{\partial \left[\sum_{i=1}^{r} w_i \frac{c_i}{\pi(y_i, v_i; \gamma)} \right]}{\partial \gamma} \Bigg| \; \gamma = \hat{\gamma}^{(j)} \text{ and } V^{-1}(\hat{\gamma}^{(j)})$$

is the inverse of the estimated quasi-randomization variance of

$$\sum_{i=1}^{r} w_i \frac{c_i}{\pi(y_i, v_i; \gamma)},$$

computed at $\gamma = \hat{\gamma}^{(j)}$.

Chang and Kott (2008) do not assume a model for the outcome and their approach is therefore restricted to estimation of the model for the response probabilities. Pfeffermann and Sikov (2011) use the likelihood (3.25) for estimating population models assuming noninformative sampling. Maximization of the likelihood is carried out by iterating between maximization of the likelihood with respect θ^* for given γ^*, and the solution of calibration equations with respect to γ^* for given θ^*, using known totals of calibration variables, similarly to Chang and Kott (2008). The 'given' parameters are the estimates from the previous iteration. The authors show how to estimate the distribution of the missing covariates and outcome for a nonresponding unit and use this distribution for imputing the missing outcomes and hence predicting the finite population total of the outcome variable.

Estimation of the population model by fitting the sample model has some important advantages not shared by the other approaches considered in this article.

1. Once the sample model is specified, it lends itself to standard model based inference such as likelihood based methods, Bayesian inference or semi-parametric modelling. It is important to emphasize in this regard that the goodness of fit of the postulated population model can be evaluated by testing the goodness of fit of the sample model fitted to the observed outcomes, using classical model diagnostic techniques. See Krieger and Pfeffermann (1997) and Pfeffermann and Sikov (2011) for appropriate test statistics with illustrations.
2. The sample likelihood provides a coherent way of handling NMAR nonresponse when estimating population models. Methods based on probability weighting require knowledge or good estimators of the response probabilities. The use of the full likelihood (see below) requires knowledge of the covariates of nonsampled units.
3. Application of this approach permits the use of conditional inference, given the sample of responding units, for example, conditioning on the observed covariates.
4. The models holding for the observed outcomes and the response probabilities define the model holding for the missing outcomes of the non-sampled units or the nonrespondents, which can be used for imputation of these outcomes. Methods based on probability weighting and variants thereof allow estimating the population model but under informative sampling and NMAR nonresponse, the population model cannot be used for prediction or imputation of the missing outcomes. See Sverchkov and Pfeffermann (2004) and Pfeffermann and Sikov (2011) for illustrations.
5. The use of the sample model enables testing whether the sampling process can be ignored. Pfeffermann and Sverchkov (2009) review several test statistics proposed in the literature for testing the ignorability of the sample selection.

3.6.2 The full likelihood

Theoretically, a more efficient way of estimating the unknown population model parameters is to base the likelihood on the joint distribution of the sample data and the sample membership indicators. Under full response, the *full likelihood* is then,

$$L_f(\theta, \gamma; \mathrm{I}_U, \mathrm{y}_s, \mathrm{x}_s, \mathrm{x}_{\bar{s}}) = \prod_{i \in s} \Pr(I_i = 1 \mid y_i, \mathrm{x}_i; \gamma) f_p(y_i \mid \mathrm{x}_i; \theta) \prod_{j \notin s} [1 - \Pr(I_j = 1 \mid \mathrm{x}_j; \theta, \gamma)], \quad (3.27)$$

where $\mathrm{I}_U = \{I_1, ..., I_N\}$ is the vector of sample inclusion indicators and $\Pr(I_j = 1 \mid \mathrm{x}_j; \theta, \gamma) = \int \Pr(I_j = 1 \mid y_j, \mathrm{x}_j, \gamma) f_p(y_j \mid \mathrm{x}_j, \theta)\, dy_j$ is the *propensity score* of unit j. The likelihood (3.27) assumes $\Pr(\mathrm{I}_U \mid \mathrm{y}_U, \mathrm{x}_U) = \prod_{k \in U} \Pr(I_k \mid y_k, \mathrm{x}_k)$ (Poisson sampling), but it can be generalized to other sampling designs. The full likelihood has the advantage of accounting for the sampling probabilities of units outside the sample, thus utilizing more information, but it requires knowledge of the covariates of all the population units. See, for example, Gelman, Carlin, Stern and Rubin (2003) and Little (2004). Modelling the joint distribution of the covariates for units outside the sample and integrating them out of the likelihood can be very complicated in practice and is formidable when there are many of them. Pfeffermann *et al*. (2006) compare empirically the use of the sample likelihood with the use of the full likelihood for multi-level models in a Bayesian context. The two approaches yield similar results, but this of course may not be the case in other applications.

Another way of defining the full likelihood is by application of the *Missing Information* Principle (MIP, Orchard and Woodbury 1972). The basic idea is to express the sample score function as the conditional expectation of the population score function, given the sample data. Following Chambers and Skinner (2003, Chapter 2), define the *full-sample likelihood* as $L_{fs}(\lambda) = f(\lambda; y_s, \mathrm{x}_s, I_U, z_U)$

where, as before, z_U is a known matrix of population values underlying the sample selection and λ defines the unknown model parameters. The corresponding *full-population* likelihood is $L_{fU}(\lambda) = f(\lambda; y_U, x_U, I_U, z_U)$ where $y_U = (y_s, y_{\bar{s}})$ and $x_U = (x_s, x_{\bar{s}})$. The MIP principle states that,

$$sc_s(\lambda) = (\partial / \partial\lambda)\log[L_{fs}(\lambda)] = E_p[(\partial / \partial\lambda)\log L_{fU}(\lambda) \mid y_s, x_s, I_U, z_U]. \quad (3.28)$$

Another identity defines the relationship between the population likelihood information matrix and the sample likelihood information matrix.

Breckling, Chambers, Dorfman, Tam and Welsh (1994) and Chambers *et al*. (1998) consider applications of the MIP to complex survey data. In particular, Chambers *et al*. (1998) study the use of the MIP when only limited design information is available and not the full information entailed in z_U. The authors show examples where the use of the MIP is more efficient than the use of the sample likelihood $L_s(\theta, \gamma; y_s, x_s)$ defined by (3.19), which only uses the weights $\{w_i,\ i \in s\}$. The likelihood (3.28) can be extended to account for NMAR nonresponse but the application of this approach requires then knowledge of the population values of the variables explaining the response. The computation of the expectation in the right hand side of (3.29) may not be simple either, depending on the population model.

Remark 11. The use of the MIP method in the simulation set up of Section (3.1) requires knowledge of the covariates and stratification membership for units outside the sample. We didn't find a way of applying the method in this case without further assumptions on the joint distribution of the covariates and the design variables.

3.6.3 Empirical likelihood

In recent years there is a growing interest in the use of empirical likelihood (EL) methods for analyzing complex survey data. The EL method as originally proposed by Hartley and Rao (1968) in the survey sample context and by Owen (1988, 2001) combines the robustness of non-parametric methods with the effectiveness of the likelihood approach. Two other important advantages of this method are that it lends itself very naturally to the use of calibration equations and that it enables the construction of confidence intervals without the need for variance estimation.

Consider the model defined by (3.13) where for now we view the covariates as random, and denote $g_i = (y_i, x_i')'$. Under some regularity conditions, the vector parameter θ is the unique solution of the equation

$$E_p\left\{\frac{\partial m(x; \theta)}{\partial\theta}[y - m(x; \theta)]\right\} = 0.$$

Let $p_1, ..., p_n$ be a set of probabilities corresponding to the observations $(g_1, ..., g_n)$ such that p_i is the 'jump' (probability mass) of the population cumulative distribution $F_p(g_i)$ at g_i. It is assumed that F_p has its support on the observed values such that

$$\sum_{i=1}^n p_i \frac{\partial m(x_i; \theta)}{\partial\theta}[y_i - m(x_i; \theta)] = 0. \quad (3.29)$$

Assuming independent observations, the EL of F_p is $L(F_p) = \prod_{i=1}^n p_i$. Notice that if p_i is a known function of some unknown parameters, $L(F_p)$ coincides with the standard parametric likelihood. The (nonparametric) EL estimators of the probabilities p_i are the solution $p_i^{(p)}$ of the maximization problem,

$$\max_{p_1,...,p_n} \prod_{i=1}^n p_i \ \text{ s.t. } p_i \geq 0, \sum_{i=1}^n p_i = 1, \quad (3.30)$$

yielding $p_i^{(p)} = 1/n$, $i = 1, ..., n$. For the linear regression case, $m(x_i; \theta) = x_i'\beta$ and by substituting $p_i^{(p)}$ for p_i in (3.29) and solving the equations we obtain the EL estimator of β as $\hat{\beta}_{el} = \hat{\beta}_{OLS}$. When finite population means $\bar{C}^U$ of variables C measured in the sample are known, they can be added to the maximization problem (3.30) by adding the calibration constraints $\sum_{i=1}^n p_i c_i = \bar{C}^U$. This additional information is expected to enhance the estimation of the p_i's and hence the estimation of the unknown model parameters. See also Remark 12 below.

Suppose now that units are drawn to the sample (or respond) with unequal selection probabilities π_i. In this case it is common to replace the objective empirical likelihood $L(F_p) = \prod_{i=1}^n p_i$ by the pseudo empirical likelihood $L_{pl}(F_p) = \prod_{i=1}^n p_i^{w_i}$, where, as before, $w_i = 1/\pi_i$. Notice that $\log L_{pl}(F_p) = \sum_{i=1}^n w_i \log(p_i)$ is the H-T estimator of $\log L_{pop}(F_p) = \sum_{i=1}^N \log p_i$. The pseudo EL estimators of the p_i's solve the maximization problem,

$$\max_{p_1,...,p_n} = \prod_{i=1}^n p_i^{w_i} \ \text{ s.t. } p_i \geq 0, \ \sum_{i=1}^n p_i = 1. \quad (3.31)$$

See, *e.g.*, Chen and Sitter (1999). It is easy to verify that in the absence of benchmark constraints, the solution of (3.31) is $p_i^{(pel)} = w_i / \sum_{i=1}^n w_i$ and by substituting $p_i^{(pel)}$ for p_i in (3.29), $\hat{\beta}_{pel} = \hat{\beta}_{pw}$, the PW estimator (3.8).

The empirical likelihoods in (3.30) and (3.31) are with respect to the population distribution. Alternatively, one can obtain the EL estimator by defining the likelihood with respect to the sample distribution $f_s(g_i) = \Pr(I_i = 1 \mid g_i)$ $f_p(g_i) / \Pr(I_i = 1)$, where by denoting $\tau_i = \Pr(I_i = 1 \mid g_i)$, $\Pr(I_i = 1) = \sum_{i=1}^n p_i \tau_i$. Following Kim (2009) and Chaudhuri, Handcock and Rendall (2010), the EL estimators of the probabilities p_i are obtained now as the solution of the maximization problem

$$\max_{p_1,\ldots,p_n}\left[\sum_{i=1}^{n}\log(p_i\tau_i) - n\log\sum_{i=1}^{n}p_i\tau_i\right]$$

$$\text{s.t. } p_i \geq 0, \sum_{i=1}^{n}p_i = 1. \quad (3.32)$$

The solution of (3.32) is $p_i^{\text{sel}} = \tau_i^{-1}/\sum_{j=1}^{n}\tau_j^{-1}$ and by substituting in (3.29),

$$\hat{\beta}_{\text{sel}} = \left[\sum_{i=1}^{n}\tau_i^{-1}\mathrm{x}_i\mathrm{x}_i'\right]^{-1}\sum_{i=1}^{n}\tau_i^{-1}\mathrm{x}_i y_i. \quad (3.33)$$

The estimator $\hat{\beta}_{\text{sel}}$ has the same form as the PW estimator $\hat{\beta}_{\text{pw}}$ in (3.8), but with the weights $\tau_i^{-1} = 1/\Pr(i \in s \mid y_i, \mathrm{x}_i)$ instead of the sampling weights w_i. In practice, one has to replace the probabilities τ_i by sample estimates $\hat{\tau}_i$. See Section 4.

Remark 12. The following possible enhancement to the estimation of the probabilities p_i was proposed to me by Dr. Jae Kim in a private communication. Assuming as before that $\Pr(i \in s \mid \pi_i, y_i, \mathrm{x}_i) = \pi_i$, it follows that $\tau_i = \Pr(I_i = 1 \mid y_i, \mathrm{x}_i) = E_p(\pi_i \mid y_i, \mathrm{x}_i)$ and hence that $E_p[(\pi_i - \tau_i) \mid y_i, \mathrm{x}_i] = 0$. This suggests adding calibration constraints of the form

$$\sum_{j=1}^{n}p_j(\pi_j - \hat{\tau}_j)\,k(y_j, \mathrm{x}_j) = 0 \quad (3.34)$$

to enhance the estimation of the probabilities $\{p_i\}$ in (3.31), where $k(y_j, \mathrm{x}_j) = k(g_j)$ is some function of the observed outcome and covariates. Examples for plausible functions for the case of a single covariate x are, $k(g_j) = y_j x_j$, $k(g_j) = y_j/x_j$ *etc*. The notable feature of the constraints (3.34) is that they do not require knowledge of population quantities like means of calibration variables, as is often assumed when advocating the EL approach for sample survey estimation. Clearly, when means $\bar{C}^U$ of calibration variables are known, constraints of the form $\sum_{i=1}^{n}p_i\mathrm{c}_i = \bar{\mathrm{C}}^U$ may be added as well. See also Remark 14.

4. Empirical study

In this section I report the results of a simulation study aimed at assessing and comparing the performance of the methods discussed in Section 3. The simulation set up is described in Section 3.1 and we use $H = 5$ strata. The target parameters are the regression coefficients $\beta' = (\beta_0, \beta_1) = (2,1)$ of the population expectation (3.1). The simulation study consists of generating 2,000 populations and samples (one sample from each population) and computing the estimators, variance estimators and confidence intervals listed below for each sample. The population size is 5,000 with approximate strata sizes $N_h = 363,\ 554,\ 842,\ 1{,}278,\ 1{,}963$. (The strata sizes are random). The sample size is $n = 300$ with $n_h = 60$ sampled units in each stratum. The sampling fractions are therefore highly variable across the strata.

We generated population values of a single discrete covariate x by first generating observations $\tilde{x}_j$ from a *Gamma* distribution with mean 2 and variance 4, and then defining x_j to be the nearest integer to $\tilde{x}_j$ if $\tilde{x}_j < 5$ and $x_j = 5$ otherwise. The covariates are therefore $\mathrm{x}_j = (1, x_j)'$, with $x_j = 0,1, \ldots, 5$. The population covariates were generated once and held fixed for all the populations.

Figure 1 shows the population and sample *pdf*s of the outcome y for $x = 2, 3, 4, 5$.

As can be seen, the population and sample *pdf*s differ, indicating the informativeness of the sampling process. Notice also that the population *pdf* is not normal because the random coefficients ζ_j are not normal.

We study the performance of the various methods in terms of *bias*, *variance*, *variance estimation*, and *confidence interval* coverage. We assume for all the methods that the only available information are the observed outcomes and covariates $(\mathrm{y}_{hs}, \mathrm{x}_{hs})$ for every stratum h, the sample selection probabilities and the true strata sizes $\{N_h\}$. I believe that this is the practice in most real life applications.

4.1 Estimators considered

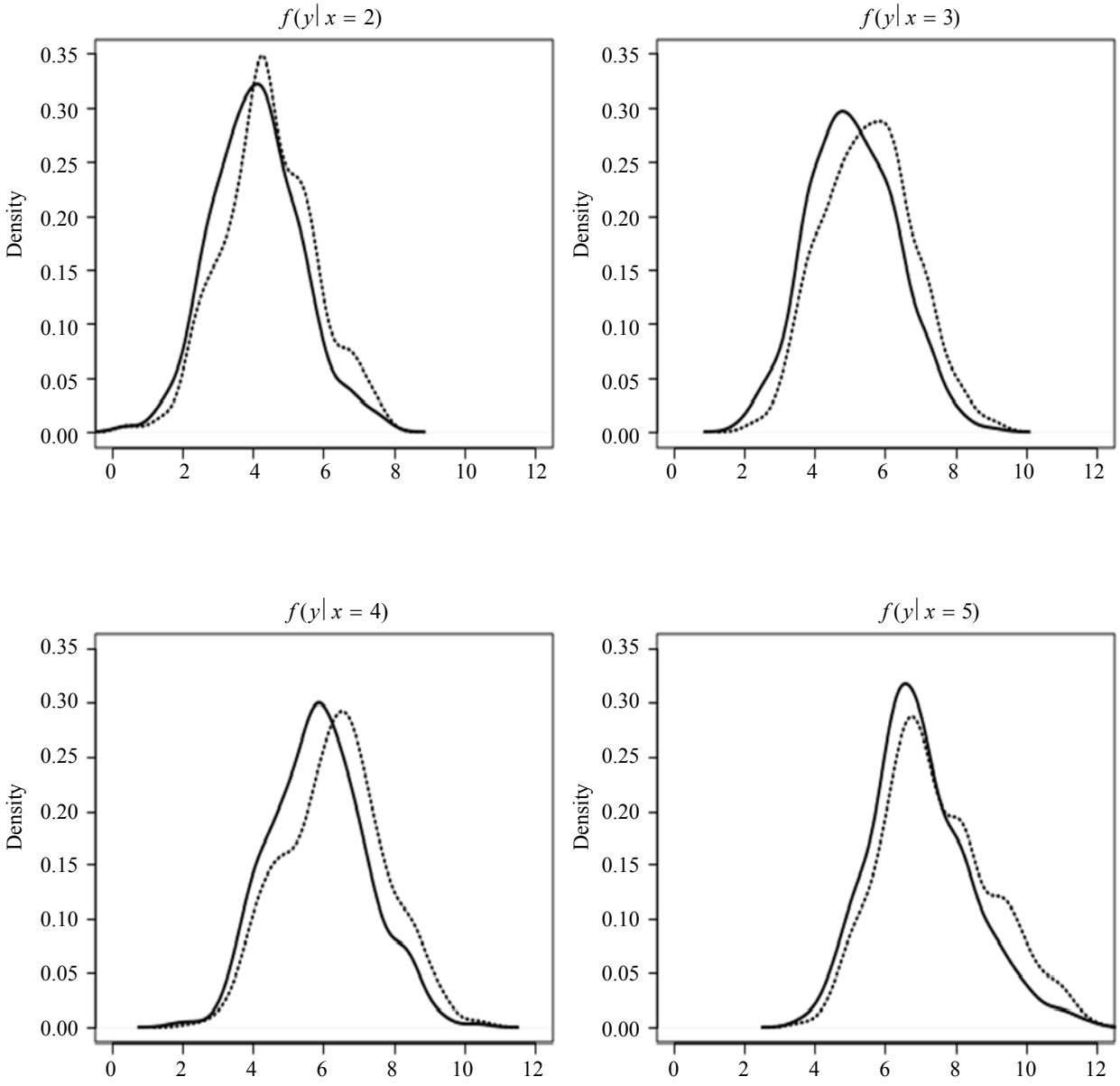
4.1.1 The OLS estimator $\hat{\beta}_{\text{ols}}$. The use of this estimator ignores the sampling process.

4.1.2 The estimator proposed by Feder (2011, see Section 3.2). Application of this approach is in four steps. *i*) fit a linear model with constant residual variance in each stratum, *ii*) impute the missing covariate values for the non-sampled units by sampling with replacement $(N_h - n_h)$ values from the n_h observed values in stratum h with probabilities $\vec{p}_{hi} = (w_{hi} - 1)/\sum_{k=1}^{n_h}(w_{hk} - 1)$ on each draw, where the w_{hi}'s are the sampling weights when sampling from stratum h. *iii*) impute the missing y-values in each stratum by generating observations at random from the model fitted in Step *i*). *iv*) fit the linear regression model of y on x by using all the population data, with the missing values for the non-sampled units replaced by the imputed values. We denote the resulting estimator by $\hat{\beta}_f$.

4.1.3 The PW estimator $\hat{\beta}_{\text{pw}}$ (Equation 3.8).

4.1.4 The estimator $\hat{\beta}_{\text{mg}}$ proposed by Magee (1998, see Section 3.5). In our application we define $a_i(\alpha) = (x_i + 0.1)^{\alpha}$ and search for the optimal power α in the range $[-2, 2]$ minimizing the determinant of the asymptotic variance estimator (3.12).

4.1.5 The estimator $\hat{\beta}_q$ defined by (3.16). For the present study we do not assume any parametric model for the expectation $E_s(w_i \mid \mathrm{x}_i)$ in the denominator of q_i and estimate $\hat{E}_s(w_i \mid \mathrm{x}_i) = \bar{w}_s(\mathrm{x}_i)$, the mean of the observed sampling weights for units with $\mathrm{x} = \mathrm{x}_i$.



Figue 1 Population *pdf* (solid line) and sample *pdf* (dashed line) of *y*|x

4.1.6 The modified q-weighted estimator $\hat{\beta}_{\text{mg}-q}$ defined by (3.17). The weights $\hat{q}_i$ are obtained as in 4.1.5 and the functions $a_{i,q}(\alpha)$ as in 4.1.4.

4.1.7 Estimators derived by maximization of the sample likelihood (3.19). The use of this approach requires specifying the population *pdf* and the expectation $E_s(w_i \mid y_i, \mathrm{x}_i)$. The unknown population model parameters are $\theta' = (\beta', \sigma^2)$ and we assume $f_p(y_i \mid \mathrm{x}_i; \theta) = N(\mathrm{x}_i'\beta, \sigma^2)$, which as noted before and illustrated in Figure 1 is not the correct *pdf* since the random coefficients ζ_j are not normal (see Section 3.1). We estimated $E_s(w_i \mid y_i, \mathrm{x}_i; \gamma)$ nonparametrically and set up the likelihood as follows:

Let s_{x_i} define the sample of units with $\mathrm{x} = \mathrm{x}_i$ of size m_{x_i}. We first divided the sample into $c(\mathrm{x}_i)$ homogeneous clusters based on the ascending values of the outcome y using the R function "hclust". The $c(\mathrm{x}_i)$'s are between 1 and 7, depending on the sample size m_{x_i} (one cluster if $m_{\mathrm{x}_i} \le 10$, 2 clusters if $m_{\mathrm{x}_i} \le 20$, ..., 7 clusters if $m_{\mathrm{x}_i} \ge 70$). Denote by $b_{\mathrm{x}_i, k}$ the midpoint between the highest y-value in cluster k and the lowest y-value in cluster $(k+1)$, $k = 1, ..., c(\mathrm{x}_i) - 1$, and define $b_{\mathrm{x}_i,0} = -\infty$, $b_{\mathrm{x}_i,c(\mathrm{x}_i)} = +\infty$. For $b_{\mathrm{x}_i,k-1} \le y \le b_{\mathrm{x}_i,k}$ we estimated $E_s(w_i \mid y_i, \mathrm{x}_i)$ by the mean $\bar{w}_s(y, \mathrm{x}_i) = \bar{w}_k(\mathrm{x}_i)$ of the sampling weights of units with y-values in the same interval. Substituting $E_s(w_i \mid y_i, \mathrm{x}_i) = \bar{w}_s(y_i, \mathrm{x}_i)$ in (3.19) defines the sample likelihood used for the present simulation study as,

$$L_s(\theta;\, y_s, \mathrm{x}_s) = \prod_{i\in s} \frac{f_p(y_i \mid \mathrm{x}_i;\, \theta) \,/\, \overline{w}_s(y_i, \mathrm{x}_i)}{\sum_{k=1}^{c(\mathrm{x}_i)} [F_p(b_{k,\mathrm{x}_i}) - F_p(b_{k-1,\mathrm{x}_i})] \,/\, \overline{w}_k(\mathrm{x}_i)}, \quad (4.1)$$

where $F_p(b_{k,\,\mathrm{x}_i}) = \int_{-\infty}^{b_k} f_p(y \mid \mathrm{x}_i;\, \theta)\, dy$ (the CDF of the assumed normal *pdf*).

The approximation (4.1) is similar to the approximation (3.20) proposed for the case where both x and y are discrete.

Remark 13. In order to facilitate the numerical optimizations used for the computation of the estimators $\hat{\beta}_{\mathrm{mg}}, \hat{\beta}_{\mathrm{mg}-q}$ and the maximum likelihood estimators in (4.1), we transformed the minimization problem $\min\{f(\theta)\colon \theta \in (a, b)\}$ to $\min\{f[g(\eta)]\colon \eta \in (-\infty, \infty)\}$ with the function $g(\eta)$ defined as $g(\eta) = [(b-a)\tan^{-1}(\eta)] / \pi + 0.5(a+b)$. Notice that every $\theta \in (a, b)$ has an image $\eta \in R$; $g(\eta) = \theta$, and $\arg\min\{f(\theta)\colon \theta \in (a,b)\} = g(\eta_0)$ where $\eta_0 = \arg\min f[g(\eta)]$.

We used the R function *nlm* for the numerical optimization, with the PW estimates as starting values. To prevent numerical overflows of the optimized function by evaluation of exponentials of large numbers, the maximization was limited to the intervals $\{\min[0.5\hat{\beta}_{\mathrm{pw}}, \hat{\beta}_{\mathrm{pw}} - 3\hat{\mathrm{se}}(\hat{\beta}_{\mathrm{pw}})], \max[1.5\hat{\beta}_{\mathrm{pw}}, \hat{\beta}_{\mathrm{pw}} + 3\hat{\mathrm{se}}(\hat{\beta}_{\mathrm{pw}})]\}$ for β, and $[0.5\hat{\sigma}_{\mathrm{pw}}, 1.5\hat{\sigma}_{\mathrm{pw}}]$ for σ.

4.1.8 The empirical likelihood estimator $\hat{\beta}_{\mathrm{sel}}$ defined by (3.33). The computation of this estimator requires estimating the probabilities $\tau_i = \Pr(I_i = 1 \mid \mathrm{y}_i, \mathrm{x}_i) = 1/E_s(w_i \mid \mathrm{y}_i, \mathrm{x}_i)$, and we use the estimator $\hat{E}_s(w_i \mid y_i, \mathrm{x}_i) = \overline{w}_{s,k}(y, \mathrm{x}_i)$ used for defining the likelihood (4.1), such that $\hat{\tau}_i = 1/\overline{w}_k(y, \mathrm{x}_i)$.

4.2 Variance estimation

We applied three approaches for variance estimation. The first approach estimates the randomization variance, the second approach estimates the variance under the sample model, while the third approach uses the nonparametric bootstrap method, which likewise estimates the variance under the sample model.

Consider first the estimators defined by 4.1.1, 4.1.3 – 4.1.6 and 4.1.8 in Section 4.1. All these estimators can be written in the generic form,

$$\hat{\beta}_t = \left[\sum_{i=1}^n w_i t_i \mathrm{x}_i \mathrm{x}_i'\right]^{-1} \sum_{i=1}^n w_i t_i \mathrm{x}_i \mathrm{y}_i = [X_s' W_s T_s X_s]^{-1} \sum_{i=1}^n w_i t_i \mathrm{x}_i \mathrm{y}_i, \quad (4.2)$$

where $X_s' = [\mathrm{x}_1, ..., \mathrm{x}_n]$, $W_s = \mathrm{diag}[w_1, ..., w_n]$ is the diagonal matrix with the sampling weights on the main diagonal and $T_s = \mathrm{diag}[t_1, ..., t_n]$, with the t_i's defined by the estimators. For $\hat{\beta}_{\mathrm{ols}}$ $t_i = 1 / w_i$, for $\hat{\beta}_{\mathrm{sel}}$ $t_i = w_i^{-1}\hat{\tau}_i^{-1}$ and so forth. The randomization variance of these estimators is estimated as,

$$\hat{\mathrm{Var}}_r(\hat{\beta}_t) = [X_s' W_s T_s X_s]^{-1} [\hat{\mathrm{Var}}_r \textstyle\sum_{i=1}^n w_i t_i \mathrm{x}_i e_{it}] [X_s' W_s T_s X_s]^{-1}, \quad (4.3)$$

where $e_{it} = (y_i - \mathrm{x}_i' B)$ and B is the census estimator. Using the double index (hj) to define the j^{th} unit in the sample s_h of size n_h drawn from straum h, we estimated

$$\hat{\mathrm{Var}}_r\left[\sum_{i=1}^n w_i t_i \mathrm{x}_i e_{it}\right] = \sum_{h=1}^5 \hat{\mathrm{Var}}\left(\sum_{j=1}^{n_h} w_{hj} \tilde{e}_{hj,t}\right) = \sum_{h=1}^5 \frac{n_h}{(n_h - 1)} \sum_{j=1}^{n_h} (w_{hj}\tilde{e}_{hj,t} - \overline{e}_{h,t})(w_{hj}\tilde{e}_{hj,t} - \overline{e}_{h,t})', \quad (4.4)$$

where $\tilde{e}_{hj,\,t} = t_{hj}\mathrm{x}_{hj}(y_{hj} - \mathrm{x}_{hj}'\hat{\beta}_t)$ and

$$\overline{e}_{h,t} = \frac{1}{n_h}\sum_{j=1}^{n_h} w_{hj}\tilde{e}_{hj,\,t},$$

assuming with replacement sampling within the strata.

A variance estimator under the sample model which accounts for possible heteroscedasticity is obtained as,

$$\hat{\mathrm{Var}}_{\mathrm{sm}}(\hat{\beta}_t) = [X_s' W_s T_s X_s]^{-1}\left[\sum_{i\in s} w_i^2 t_i^2 \hat{e}_{it}^2 \mathrm{x}_i \mathrm{x}_i'\right][X_s' W_s T_s X_s]^{-1}, \quad (4.5)$$

where $\hat{e}_{it} = (y_i - \mathrm{x}_i'\hat{\beta}_t)$. Randomization and sample model variance estimators for the estimator in 4.1.2 are developed by Feder (2011). For the maximum likelihood estimator under the sample model with the likelihood defined by (4.1) we only estimate the variance under the sample model using the inverse information matrix.

Finally, bootstrap variance estimators for all the estimators are obtained by sampling with replacement n units from the original sample and re-estimating each of the estimators using the same computations as for the original sample. Repeating the same process independently B times, the bootstrap variance estimator is,

$$\hat{\mathrm{Var}}_{BS}(\hat{\beta}) = \frac{1}{B}\sum_{b=1}^B (\hat{\beta}^{(b)} - \overline{\hat{\beta}})(\hat{\beta}^{(b)} - \overline{\hat{\beta}})'; \quad \overline{\hat{\beta}} = \frac{1}{B}\sum_{b=1}^B \hat{\beta}^{(b)}, \quad (4.6)$$

where $\hat{\beta}$ represents any of the estimators defined by 4.1.1 – 4.1.8 and $\hat{\beta}^{(b)}$ is the corresponding estimator computed for bootstrap sample b, $b = 1, ..., B$.

4.3 Computation of confidence intervals

We consider two approaches of $(1-\alpha)$ level confidence interval (C.I.) computation. The first approach is the standard C.I.,

$$\hat{\beta}_k \pm Z_{1-\frac{\alpha}{2}}\, \hat{s}.e(\hat{\beta}_k),\ k = 0, 1,$$

where $\hat{\beta}_k$ stands for any of the estimators considered and $\hat{s}.e(\hat{\beta}_k)$ is the corresponding estimator of the standard error as obtained by one of the methods listed before. The second, "basic bootstrap" approach uses the quantiles $bs(k, \tilde{\alpha})$ of the bootstrap estimators $\hat{\beta}_k^{(b)}$ to compute the C.I.

$$\left[2\hat{\beta}_k - bs\left(k, 1-\frac{\alpha}{2}\right),\ 2\hat{\beta}_k - bs\left(k, \frac{\alpha}{2}\right)\right], k = 1, 2.$$

We tried also the use of the "studentized bootstrap method" but the coverage rates were not better with any of the estimators $\hat{\beta}_k$. See Remark 14 below.

4.4 Simulation results

Table 1 shows the empirical means of the estimates listed in Section 4.1 over the 2,000 populations and samples and the corresponding empirical standard errors (S.E.). Also shown are the square roots of the means of the variance estimates as obtained when estimating the randomization variance ("Ran.") and when estimating the variance under the sample model ("S.M."). Because of computing time limitations, the results for the bootstrap variance estimators ("BS") are based on 300 bootstrap samples drawn from each of 500 original samples. These numbers of original and bootstrap samples were found to produce stable variance estimators.

As expected, given the use of an informative sampling scheme, the OLS estimator has a relatively large bias of 12% (5%) when estimating the intercept (slope). All the other estimators are virtually unbiased, except for $\hat{\beta}_{\text{mle}}$, which has bias of 2% and 1.5%. The almost unbiasedness of the EL estimator $\hat{\beta}_{\text{sel}}$ is particularly encouraging given the somewhat crude nonparametric estimation of the probabilities $\tau_i = \Pr(i \in s \mid y_i, \mathrm{x}_i)$. Notice also that this estimator has similar empirical S.E. to those of the PW estimator. The small (but statistically significant) bias of $\hat{\beta}_{\text{mle}}$ is explained by the fact that we assume a normal distribution under the population model, which as noted and illustrated before is incorrect.

Regarding precision, the OLS estimator has the smallest S.E. but $\hat{\beta}_f$ has almost the same S.E. (and is unbiased). This is explained by the fact that this estimator uses additional stratification information, not used by the other estimators. Note that $\hat{\beta}_{\text{mg}}$, $\hat{\beta}_{\text{mg-}q}$ and particularly $\hat{\beta}_q$ outperform $\hat{\beta}_{\text{pw}}$, but $\hat{\beta}_{\text{mg-}q}$ does not improve over $\hat{\beta}_q$.

Remark 14. Following my presentation of this paper at the 2011 Statistics Canada symposium, Jean-Francois Beaumont suggested to replace the weights $\hat{\tau}_i^{-1}$ used for the computation of $\hat{\beta}_{\text{sel}}$ by the weights $\hat{\tau}_i^{-1} / E_s(\hat{\tau}_i^{-1})$, so as to account for the net sampling effects on the conditional *pdf* $f(y \mid \mathrm{x})$, similarly to the use of the q-weights in $\hat{\beta}_q$. Notice that whereas the sampling weights w_i may depend on y, x and possibly other variables, the weights $\hat{\tau}_i^{-1}$ only depend on y and x. Application of this idea did not affect the bias but the empirical S.E. of the modified estimators are 0.151 and 0.053, smaller than the S.E. of $\hat{\beta}_{\text{sel}}$ and similar to the S.E. of $\hat{\beta}_q$.

Looking at the performance of the variance estimators, the first remarkable outcome is that the randomization and sample model variance estimators (Equations 4.4 and 4.5) are very similar for every estimator of the regression coefficients, even though they are computed very differently. For $\hat{\beta}_{\text{ols}}$, $\hat{\beta}_{\text{pw}}$ and $\hat{\beta}_q$ the variance estimators are almost unbiased but for the other estimators the variance estimators under-estimate the true variance. This is explained by the fact that these variance estimators ignore some of the operations involved in the computation of the estimated regression coefficients. Thus, in the case of the estimators $\hat{\beta}_{\text{mg}}$ and $\hat{\beta}_{\text{mg-}q}$ the variance estimators do not account for the choice of the optimal weights $a_i(\alpha)$, in the case of $\hat{\beta}_f$ the variance estimator does not account for the random imputation of the vectors (y_i, x_i) for $i \in U - s$, and in the case of $\hat{\beta}_{\text{mle}}$ and $\hat{\beta}_{\text{sel}}$ the variance estimators do not account for the estimation of the probabilities $\Pr(i \in s \mid y_i, \mathrm{x}_i)$. This under- estimation of the variance is corrected in almost all cases by use of the bootstrap method, see, in particular, the estimation of the variances of $\hat{\beta}_f$, $\hat{\beta}_{\text{mle}}$ and $\hat{\beta}_{\text{sel}}$.

Figure 2 shows the empirical coverage rates of $(1 - \alpha)$ - level confidence intervals (C.I.) for $\alpha = 0.10, 0.05, 0.01$, as obtained when applying the standard C.I. with the standard errors estimated by the BS method, and when using the basic bootstrap method. The figures in the horizontal axis are the nominal levels

The coverage rates are almost always below the nominal levels but the under- coverage in the case of the standard C.I. is generally less than 4%. The two exceptions are when basing the confidence intervals on the OLS estimators (large under-coverage) and the mle estimator of the slope (under-coverage of 7% at the 90% nominal level), which is explained by the bias of these estimators. The under-coverage percentages when using the basic bootstrap method are generally slightly larger, except for the under-coverage of the C.I. for the intercept based on $\hat{\beta}_{\text{sel}}$, which is more pronounced.

Table 1
Means, standard errors (S.E.) and square roots of means of variance estimates. Population model: $E_p(y_j) = 2 + 1 \times x_j$, $\mathrm{Var}_p(y_j) = (1 + 0.2x_j)^2 \mathrm{V_j} + 1$

Method	Intercept- $\hat{\beta}_0$					Slope- $\hat{\beta}_1$				
	Mean Est.	Emp. S.E.	Ran.	S.M.	BS	Mean Est.	Emp. S.E.	Ran.	S.M.	BS
$\hat{\beta}_{ols}$	2.251	0.133	0.135	0.139	0.140	1.046	0.048	0.048	0.049	0.049
$\hat{\beta}_f$	2.006	0.133	0.126	0.126	0.135	0.999	0.051	0.041	0.041	0.052
$\hat{\beta}_{pw}$	2.008	0.166	0.167	0.169	0.157	0.998	0.059	0.055	0.055	0.056
$\hat{\beta}_{mg}$	2.017	0.158	0.154	0.156	0.154	0.995	0.056	0.050	0.050	0.055
$\hat{\beta}_q$	2.011	0.153	0.157	0.159	0.147	0.999	0.054	0.051	0.051	0.052
$\hat{\beta}_{mg\text{-}q}$	2.020	0.156	0.152	0.154	0.153	0.996	0.055	0.049	0.050	0.054
$\hat{\beta}_{mle}$	1.960	0.159	-----	0.143	0.152	1.026	0.054	-----	0.046	0.053
$\hat{\beta}_{sel}$	2.031	0.164	0.143	0.143	0.159	0.995	0.058	0.049	0.049	0.057

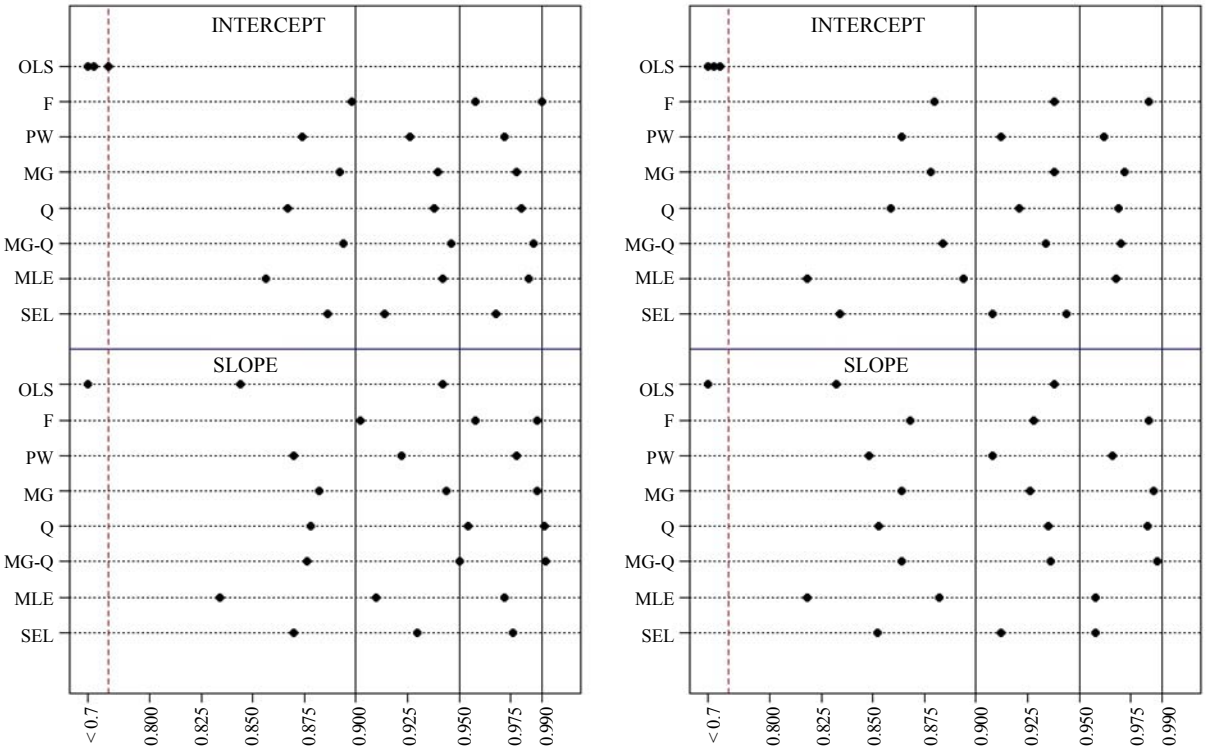


Figure 2 Coverage rates of standard (left) and BS (right) confidence intervals

Remark 15. We computed also the standard C.I. with the S.E. estimated under the randomization distribution (Equation 4.4) and under the sample model (Equation 4.5), but except in the case of the estimators $\hat{\beta}_{pw}$ and $\hat{\beta}_q$, the under-coverage of these intervals was somewhat higher than the coverage rates in Figure 2 because of the under-estimation of the true S.E. by these S.E. estimators discussed before. The same phenomenon was observed when using the "studentized bootstrap method" with these S.E. estimates, which again can be explained by the underestimation of the true S.E.'s. The use of more advanced bootstrap C.I. such as double-bootstrap may correct this under-coverage.

5. Concluding remarks

In this article I discuss alternative procedures proposed in the literature to account for informative sampling and NMAR nonresponse when modeling survey data. The empirical study is restricted so far to the case of linear

regression and single-stage sampling, and an obvious extension would be to consider other models and cluster sampling. The present study illustrates the unbiasedness or approximate unbiasedness of all the point estimators considered, but the standard variance estimators underestimate the true variances in most cases since they fail to account for the extra operations involved in computing the corresponding point estimators. The bootstrap variance estimators produce much better variance estimators in these cases. The confidence intervals applied in the present study yield small under-coverage in most cases, but they should be improved, possibly by use of more advanced bootstrap techniques. Another important extension mentioned in the paper, which we have not investigated empirically so far is to incorporate sample based calibration constraints in the empirical likelihood method when based on the sample distribution.

We plan to apply the various methods to several real data sets. This would require the development of diagnostic procedures that would allow comparing the performance of the methods since unlike in a simulation study, the true distributions and model parameters are seldom known in real applications.

Acknowledgements

I am indebted to Dr. Moshe Feder for carrying out the empirical study and many helpful comments and suggestions. Thanks are due also to Dr. Pedro Silva for his constructive remarks on an earlier draft of the paper and three reviewers for their careful reading and comments in a short time period given to them. This study is funded by a UK ESRC grant No. RES-062-23-2316.

References

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Binder, D., and Roberts, G. (2009). Design and model based inference for model parameters. In *Handbook of Statistics* 29B; *Sample Surveys*: *Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 33-54.

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M. and Welsh, A.H. (1994). Maximum likelihood inference from sample survey data. *International Statistical Review*, 62, 349-363.

Brick, J.M., and Montaquila, J.M. (2009). Nonresponse and weighting. In *Handbook of Statistics* 29A; *Sample Surveys*: *Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 163-185.

Chambers, R.L., Dorfman, A.H. and Wang, S. (1998). Limited information likelihood analysis of survey data. *Journal of the Royal Statistical Society*, Series B, 60, 397-411.

Chambers, R.L., and Skinner, C.J. (2003, Eds.). *Analysis of survey data*. New York: John Wiley & Sons, Inc.

Chambless, L.E., and Boyle, K.E. (1985). Maximum likelihood methods for complex sample data: Logistic, regression and discrete proportional hazards models. *Communication in Statistics-Theory and Methods*, 14, 1377-1392.

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

Chen, J., and Sitter, R.R. (1999). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. *Statistica sinica*, 9, 385-406.

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.

DeMets, D., and Halperin, M. (1977). Estimation of simple regression coefficients in samples arising from sub-sampling procedures. *Biometrics*, 33, 47-56.

DuMouchel, W.H., and Duncan, G.L. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association*, 78, 535-543.

Feder, M. (2011). Fitting Regression Models to Complex Survey Data- Gelman's Estimator Revisited. In Proceedings of the ISI meeting, Ireland, (www.isi2011.ie).

Francisco, C.A., and Fuller, W.A. (1991). Quantile estimation with a complex survey design. *The Annals of Statistics*, 19, 454-469.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, Series C, 37, 117-132.

Fuller, W.A. (2002). Regression estimation for survey samples. *Survey Methodology*, 28, 5-23.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science*, 22, 153-164.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *International Statistical Review*, 54, 127-138.

Godambe, V.P., and Thompson, M.E. (2009). Estimating functions and survey sampling. In *Handbook of Statistics* 29B; *Sample Surveys*: *Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 83-101.

Goldstein, H. (1986). Multi-level mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.

Hàjek, J. (1971). Comments on a paper by D. Basu. In *Foundations of Statistical Inference*, (Eds., V.P. Godambe and D.A. Sprott). Toronto: Holt, Rinehart and Winston.

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

Holt, D., Smith, T.M.F. and Winter, P.D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society*, Series A, 143, 474-487.

Jewell, N.P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika*, 72, 11-21.

Kasprzyk, D., Duncan, G.J., Kalton, G. and Singh, M.P. (1989, Eds.). *Panel Surveys*. New York: John Wiley & Sons, Inc.

Kim, J.K. (2009). Calibration estimation using empirical likelihood in survey sampling. *Statistica Sinica*, 19, 145-157.

Kott, P.S. (2009). Calibration Weighting: Combining Probability Samples and Linear Prediction Models. In *Handbook of Statistics* 29B; *Sample Surveys*: *Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 55-82.

Krieger, A.M., and Pfeffermann D. (1997). Testing of distribution functions from complex sample surveys. *Journal of Official Statistics*, 13, 123-142.

Little, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-249.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99, 546-556.

Magee, L. (1998). Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society*, Series B, 60, 115-126.

Nathan, G., and Holt, D. (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society*, Series B, 42, 377-386.

Orchard, T., and Woodbury, M.A. (1972). A missing information principle: Theory and application. *Proceedings of the 6th Berkeley Symposium on Mathematical Statistics and Probability*, 1, 697-715.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Owen, A.B. (2001). *Empirical likelihood*. New York: Chapman & Hall.

Pfeffermann, D., and Holmes, D. (1985). Robustness consideration in the choice of method of inference for regression analysis of survey data. *Journal of the Royal Statistical Society*, Series A, 148, 268-278.

Pfeffermann, D., and Smith, T.M.F. (1985). Regression models for grouped populations in cross-section surveys. *International Statistical Review*, 53, 37-59.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. *Statistical Methods in Medical Research*, 5, 239-261.

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

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998b). Weighting for unequal selection probabilities in multi-level models (with discussion). *Journal of the Royal Statistical Society*, Series B, 60, 23-76.

Pfeffermann, D., and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā*, 61, 166-186.

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: John Wiley & Sons, Inc., 175-195.

Pfeffermann, D., Moura, F.A.S. and Nascimento-Silva, P.L. (2006). Multilevel modeling under informative sampling. *Biometrika*, 93, 943-959.

Pfeffermann., D., and Sverchkov, M. (2007). Small area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, 102, 1427-1439.

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.

Pfeffermann, D., and Landsman, V. (2011). Are private schools better than public schools? Appraisal for Ireland by methods for observational studies. *The Annals of Applied Statistics*, 5, 1726-1751.

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.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 605-614.

Rubin, D.B. (1985). The use of propensity scores in applied Bayesian inference. In *Bayesian Statistics* 2, (Eds., J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith), Elsevier Science Publishers B.V., 463-472.

Särndal, C.-E., and Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.

Scott, A.J., and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares. *Journal of the American Statistical Association*, 77, 848-854.

Scott, A.J., and Wild, C.J. (2009). Population-based case-control studies. In *Handbook of Statistics* 29B; *Sample Surveys*: *Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Amsterdam: North Holland, 431-453.

Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds.) (1989). *Analysis of complex surveys*. New York: John Wiley & Sons, Inc.

Skinner, C.J. (1994). Sample models and weights. *Proceedings of the Section on Survey Research Methods*, 133-142.

Smith, T.M.F. (1988). To weight or not to weight, that is the question. In *Bayesian Statistics* 3, (Eds., J.M. Bernardo, M.H. Degroot, D.V. Lindley and A.F.M. Smith), Oxford University Press, 437-451.

Sugden, R.A., and Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71, 495-506.

Sverchkov, M., and Pfeffermann, D. (2004). Prediction of finite population totals based on the sample distribution. *Survey Methodology*, 30, 79-92.

Wu, Y.Y., and Fuller, W.A. (2006). Estimation of regression coefficients with unequal probability samples. *Proceedings of the American Statistical Association*, *Section on Survey Research Methods*, 3892-3899.