C.J. Skinner and J.D'Arrigo

Inverse probability weighting for clustered nonresponse

Article (Accepted version) (Refereed)

Original citation:
Skinner, Chris J. and D'Arrigo, Julia (2011) *Inverse probability weighting for clustered nonresponse.* *Biometrika*, 98 (4). pp. 953-966. ISSN 0006-3444

DOI: 10.1093/biomet/asr058

Inverse Probability Weighting for Clustered Nonresponse

C.J. Skinner* J.D'Arrigo†

August 2011

Abstract

Correlated nonresponse within clusters arises in certain survey settings. It is represented by a random effects model and assumed to be cluster-specific nonignorable, in the sense that survey and nonresponse outcomes are conditionally independent given cluster-level random effects. Two basic forms of inverse probability weights are considered: response propensity weights based on a marginal model, and weights based on predicted random effects. It is shown that both approaches can lead to biased estimation under cluster-specific nonignorable nonresponse, when the cluster sample sizes are small. We propose a new form of weighted estimator based upon conditional logistic regression, which can avoid this bias. An associated estimator of variance and an extension to observational studies with clustered treatment assignment are also described. Properties of the alternative estimators are illustrated in a small simulation study.

Keywords: Conditional logistic regression; Nonresponse; Response propensity; Survey weight.

*Department of Statistics, London School of Economics and Political Science, London WC2A 2AE, U.K.

†Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, SO17 1BJ, U.K.

1 Introduction

Survey weighting is widely used to correct for the potential biasing impact of nonresponse [12, 13, 18]. An important tool in the construction of survey weights is inverse probability weighting, defined here as weighting by the reciprocal of a response probability, estimated under a model. Such weights may be combined with sampling weights for an integrated treatment of nonresponse and sampling. They may also be combined with model-based predictors of a survey variable to improve efficiency [4, 9]. Such combined estimators may be doubly robust in the sense that consistent estimation can be achieved in a modelling framework if either the response model or the model for the survey variable is correct [2, 9].

Most discussions of inverse probability weighting (e.g. 3) assume that responses for different units are independent. It is not uncommon in surveys, however, for nonresponse to be correlated within clusters. Access of interviewers to respondents in some surveys is dependent on authorities at a cluster level, for example in surveys of employees within a firm, and response for individuals within such a cluster may be influenced by the extent to which the authorities encourage participation, inducing correlation. In other surveys, nonresponse may display intra-cluster correlation simply because of heterogeneity between clusters used for multistage sampling.

This paper investigates how to construct inverse probability weights, when response is clustered and cluster membership is observed for both responding and nonresponding units, as is the case when the clusters define a stage in a multi-stage sampling design. One established approach is to use such design clusters or homogeneous sets of clusters as weighting adjustment

cells [12], where the implicit model is that response probabilities vary just by cell and may be estimated by cell-level response rates. We consider the more general setting when auxiliary information at the sample level may include other variables in addition to cluster membership. A natural model for nonresponse, given such auxiliary information, is a multilevel model (e.g., 6), where clustered nonresponse is captured via random effect terms. Our interest is in how to construct inverse probability weights based on such models.

Some methods to correct for nonresponse bias in a clustered survey were proposed by **(author?)** [22]. These methods were based on a random effects model for the survey variable and thus fall outside the class of weighting methods we consider. We refer briefly to the relation between these different approaches in §7. We make use of the concept of cluster-specific non-ignorable nonresponse introduced by **(author?)** [22] to describe the case when nonresponse may depend on unobserved cluster random effects which may be correlated with the survey variables. This condition is weaker than the usual missing at random condition, which is conventionally assumed if inverse probability weighting is to correct for bias [20, p.146]. A key aim of this paper is to construct weights which exploit the auxiliary information on cluster membership to correct for bias when nonresponse is cluster-specific non-ignorable, not just missing at random. The cluster-specific noninorable condition has also been discussed, at least implicitly, by **(author?)** [14, Example 6.24], **(author?)** [19] and **(author?)** [23].

2 Estimation and Modelling Framework

Let $U = \{(i,j) : i = 1, \ldots, N, j = 1, \ldots, M_i\}$ denote a finite population, with the jth unit in the ith cluster labelled (i,j). The population size is denoted $N_0 = \sum_1^N M_i$. Suppose the objective is to estimate $T_y = \sum_{(i,j)\in U} y_{ij}$, where y_{ij} is a generic survey variable of interest. Many other parameters may be expressed as a function of such totals and estimated by this function of the corresponding estimated totals.

Let $s = \{(i,j) : i = 1, \ldots, n, j = 1, \ldots, m_i\} \subset U$ denote a sample selected by a probability sampling design from U, where the sample labels are ordered in this way without loss of generality. Suppose that π_{ij}, the probability of selection of (i,j) under the sampling design, is known and non-zero for each $(i,j) \in s$.

Let R_{ij} denote the response indicator variable, which is defined for all units $(i,j) \in U$, irrespective of which sample s is selected, such that a sample unit responds when $R_{ij} = 1$ and not if $R_{ij} = 0$. Thus, nonresponse is stable in the terminology of **(author?)** [17]. Suppose that R_{ij}, a $1 \times k$ vector of auxiliary variables x_{ij} and the cluster membership indicator i are observed for all units in s, but that y_{ij} is only observed for sample units if $R_{ij} = 1$.

Our primary focus will be on the inverse probability weighted estimator of T_y given by

$$\hat{T}_y = \sum_{(i,j)\in s} d_{ij}\hat{q}_{ij}R_{ij}y_{ij}, \tag{1}$$

where $d_{ij} = \pi_{ij}^{-1}$ is the design weight and $\hat{q}_{ij}$ is a non-response weight, representing an inverse estimated response probability, to be discussed in §3. The estimator in (1) is called the two-phase nonresponse adjusted estimator in

(author?) [18, equation (6.3)].

We also consider

$$\hat{T}_{y,\text{reg}} = \hat{T}_y + (\hat{T}_{xs} - \hat{T}_x)\hat{\lambda}, \tag{2}$$

where

$$\hat{T}_{xs} = \sum_{(i,j)\in s} d_{ij}x_{ij}, \quad \hat{T}_x = \sum_{(i,j)\in s} d_{ij}\hat{q}_{ij}R_{ij}x_{ij},$$

and

$$\hat{\lambda} = \Big(\sum_{(i,j)\in s} d_{ij}\hat{q}_{ij}R_{ij}x_{ij}^T x_{ij} \Big)^{-1} \sum_{(i,j)\in s} d_{ij}\hat{q}_{ij}R_{ij}x_{ij}^T y_{ij},$$

introduced by **(author?)** [4] and called the two-phase generalized regression estimator by **(author?)** [18, equation (6.4)] and an augmented inverse probability weighted complete case estimator by **(author?)** [20, p.148].

In order to construct the nonresponse weights $\hat{q}_{ij}$ and to assess the properties of the estimators of T_y, we introduce a model framework ξ for the generation of the R_{ij} and y_{ij}. We assume that the distribution of (R_{ij}, y_{ij}) implied by ξ does not depend on the sample outcome s. Sampling and nonresponse may thus be said to be unconfounded and sampling is noninformative with respect to y_{ij}.

The basic parametric model we consider for R_{ij}, unconditional on y_{ij}, is

$$\text{pr}(R_{ij} = 1 \mid u_i) = h(x_{ij}\beta + u_i), \quad u_i \sim N(0, \tau^2), \tag{3}$$

where the u_i are independent random effects, $h(\cdot)$ is a specified inverse link function, such as the inverse logit function, the x_{ij} are treated as fixed, and the $k \times 1$ vector β and τ^2 are unknown parameters. The R_{ij} are assumed mutually independent conditional on the u_i.

We shall only consider estimation in the case when the number of respondents in each cluster, $R_{i+} = \sum_{j=1}^{m_i} R_{ij}$, is non-zero. **(author?)** [22] comment

on ways in which biased estimation can arise when this is not the case. The event that $R_{i+} = 0$ could arise for two main reasons. First, in some surveys, nonresponse may occur as a cluster-level process, for example at the school level for a survey of children clustered in schools. We do not consider this possibility further here. Secondly, this event could arise when nonresponse is only an individual-level process, as in model (3), but m_i is small, as for example in a survey of adults clustered in households. The practical applicability of this paper will be to surveys where the m_i may not be large, e.g., values of 5 and 10 are considered in §6, but they are large enough for this event to occur under model (3) with negligible probability.

In addition to the random effects model (3), we also consider the implied marginal model:

$$\text{pr}(R_{ij} = 1) = g(x_{ij}\beta), \tag{4}$$

where $g(x_{ij}\beta) = E\{h(x_{ij}\beta + u_i)\}$ and the expectation is taken across the distribution of u_i. This random effect will induce a correlation between R_{ij} and R_{ik} for $j \neq k$.

We consider two principal assumptions regarding the relation between R_{ij} and y_{ij}. Nonresponse is said to be missing at random if the R_{ij} and y_{ij} are pairwise independent. The mechanism is said to be cluster-specific nonignorable nonresponse, following **(author?)** [22], if model (3) holds and the R_{ij} and y_{ij} are independent conditional on the u_i, that is $\text{pr}(R_{ij} = 1|y_{ij}, u_i) = \text{pr}(R_{ij} = 1|u_i)$.

To illustrate and motivate the cluster-specific nonignorable nonresponse assumption, suppose that y_{ij} obeys a linear mixed effects model

$$y_{ij} = x_{ij}\lambda + \nu_i + \epsilon_{ij}, \tag{5}$$

where ν_i and ϵ_{ij} are random effects with zero means, such that the R_{ij} are conditionally independent of the ν_i and ϵ_{ij} given the u_i and, furthermore, u_i is conditionally independent of the ϵ_{ij} given the ν_i. Then, when both models (3) and (5) hold, nonresponse is missing at random when u_i and ν_i are independent and cluster-specific nonignorable otherwise. The principal relevance of this paper is to cases when nonresponse is cluster-specific nonignorable but not missing at random. The key motivating application arises when both nonresponse and the survey variable exhibit clustering, which may be represented by the kind of joint cluster effect model for (R_{ij}, y_{ij}) in (3) and (5), where the cluster effects u_i and ν_i display correlation after controlling for observable x_{ij}. For example, when clustering is by geography, correlation between area-level response rates and area means of the survey variable may be induced by a common correlation with average area-level income which is not available as an x_{ij} variable.

3 Construction of Nonresponse Weight

We now consider the construction of the weight $\hat{q}_{ij}$ used in the estimators in (1) and (2), when model (3) holds. We first consider three basic options to serve as benchmarks for assessing the proposed option.

(i) Response propensity weights (13): the inverse link function $g(\cdot)$ in the marginal probability $\text{pr}(R_{ij} = 1)$ in (4) is assumed known and the weights are set to be $\hat{q}_{ij}^{\text{M}} = g(x_{ij}\hat{\beta}^{\text{M}})^{-1}$, where $\hat{\beta}^{\text{M}}$ is obtained, for example, by maximum likelihood estimation under the working model of independent observations.

(ii)Weights based on predicted random effects: set $\hat{q}_{ij}^{\text{RE}} = h(x_{ij}\hat{\beta}^{\text{RE}} + \hat{u}_i^{\text{RE}})^{-1}$, based on the random effects model in (3), where $\hat{\beta}^{\text{RE}}$ and the $\hat{u}_i^{\text{RE}}$

and implicitly $\hat{\tau}^{2\mathrm{RE}}$ might be obtained using an approximate maximum likelihood method, such as in **(author?)** [5, p.174], where $\hat{u}_i^{\mathrm{RE}}$ is the mode of an approximate predictive distribution for u_i given the observed R_{ij}.

(iii)Weights based on estimated fixed effects: set $\hat{q}_{ij}^{\mathrm{FE}} = h(x_{ij}\hat{\beta}^{\mathrm{FE}}+\hat{u}_i^{\mathrm{FE}})^{-1}$ as in (ii), but where the u_i in (3) are now treated as unknown parameters (fixed effects) and $\hat{\beta}^{\mathrm{FE}}$ and the $\hat{u}_i^{\mathrm{FE}}$ are maximum likelihood estimators. One advantage of this approach compared to (ii) when $h(\cdot)$ is the inverse logit function is that it avoids numerical integration in the computation.

We shall present theoretical reasons in the next section why each of the above options may not correct adequately for bias from cluster-specific non-ignorable nonresponse when the m_i may be small. We now propose an alternative conditional logistic regression approach for this case, designed to remove the dependence of the weighting method on the random effects. The basic idea is to construct the weight as $\mathrm{pr}(R_{ij} = 1 \mid R_{i+})^{-1}$. It may be shown (e.g., 1, p.251) that when model (3) holds and $h(\cdot)$ is the inverse logit function, we have

$$\mathrm{pr}(R_{ij} = 1 \mid R_{i+}) = \frac{\sum_{r_i \in B_{1ij}} \exp(\sum_{j=1}^{m_i} r_{ij}x_{ij}\beta)}{\sum_{r_i \in B_{2i}} \exp(\sum_{j=1}^{m_i} r_{ij}x_{ij}\beta)}, \tag{6}$$

where r_{ij} denotes a possible value taken by R_{ij}, $r_i = (r_{i1}, \ldots, r_{im_i})$, $r_{i+} = \sum_j r_{ij}$, $B_{1ij} = \{r_i : r_{ij} = 1, r_{i+} = R_{i+}\}$ and $B_{2i} = \{r_i : r_{i+} = R_{i+}\}$. The absence of the u_i in (6) arises from the sufficiency of R_{i+} for u_i. In practice, β is unknown and we propose to estimate it by conditional maximum likelihood. We suppose that the first element of x_{ij} is the intercept and write $x_{ij} = (1\ x_{1ij})$, $\beta = (\beta_0\ \beta_1^T)^T$ and $x_{ij}\beta = \beta_0 + x_{1ij}\beta_1$. The parameter β_0 cancels from (6) and we express the weight as $\hat{q}_{ij}^{\mathrm{CML}} = \mathrm{pr}(R_{ij} = 1 \mid R_{i+}; \beta_1 = \hat{\beta}_1^{\mathrm{CML}})^{-1}$,

where $\hat{\beta}_1^{\mathrm{CML}}$ is obtained by solving the conditional score equation given by

$$U(\beta_1) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} U_{ij}(\beta_1) = 0, \qquad U_{ij}(\beta_1) = R_{ij} x_{1ij} - \sum_{r_i \in B_{2i}} a_i(\beta_1) x_{1i+r} / \{ m_i \sum_{r_i \in B_{2i}} a_i(\beta_1) \},$$

$$a_i(\beta_1) = \exp\Big(\sum_{j=1}^{m_i} r_{ij} x_{1ij} \beta_1 \Big), \qquad x_{1i+r} = \sum_{j=1}^{m_i} r_{ij} x_{1ij}.$$

The conditional logistic approach is closer to the fixed effects than the random effects approach in the sense that, given β, the weights in cluster i depend only on the R_{ij} in cluster i and they are not shrunk to a cluster average using outcomes from other clusters. In the special case when $x_{ij} = x_i$ and we replace $x_{ij}\beta + u_i$ by u_i, since x_i is effectively confounded with u_i, both the conditional logistic and fixed effects weights reduce to m_i/R_{i+}, the inverse response rate in cluster i, a traditional choice of weight with clustered survey data [22]. In the general case, $\hat{u}_i^{\mathrm{FE}}$ is the solution of $R_{i+} = \sum_j h(x_{ij}\beta + \hat{u}_i^{\mathrm{FE}})$, for given β, and thus both sets of weights in cluster i may be viewed as functions of the cluster level response rate R_{i+}/m_i, with the functions depending, in slightly different ways, on $h(\cdot)$ and β. Compared to the random effects approach, the conditional logistic approach has the advantage that it does not depend on assumptions about the distribution of u_i nor about the relation of u_i to x_{ij}. On the other hand, it does depend on the assumption that $h(\cdot)$ is the inverse logit function in order that (6) holds and is free of u_i. Since we have assumed that sampling and nonresponse are unconfounded, we have not incorporated design weights in either the expression in (6) or the construction of $\hat{\beta}_1^{\mathrm{CML}}$.

4 Bias Properties of Weighted Estimators

We now consider how well the four weighting approaches described in the previous section correct for the bias arising from nonresponse which is either missing at random or cluster-specific nonignorable. This bias is approximated in an asymptotic framework, where increasing values of n, the number of sampled clusters, index a sequence of finite populations and samples (8, sect. 1.3), such that the population size N_0 also increases but the cluster sample sizes, m_i, may remain small. We ignore stratification for simplicity but note that, in practice, the sampling of clusters is usually stratified and it may be more appropriate to assume that the number of strata increases, with the numbers of clusters within each stratum remaining small and fixed.

We suppose that under the asymptotic framework, $\hat{\beta}^{\mathrm{M}} = \beta^{\mathrm{M}} + o_p(1)$, $\hat{\beta}^{\mathrm{RE}} = \beta^{\mathrm{RE}} + o_p(1)$, $\hat{\tau}^{2\mathrm{RE}} = \tau^{2\mathrm{RE}} + o_p(1)$, $\hat{\beta}^{\mathrm{CML}} = \beta^{\mathrm{RE}} + o_p(1)$, where $g(x_{ij}\beta^{\mathrm{M}})$ is the true value of $\mathrm{pr}(R_{ij} = 1)$ in (4) and $(\beta^{\mathrm{RE}}, \tau^{2\mathrm{RE}})$ define the true model when (3) holds. See, e.g., **(author?)** [11] for the consistency of $\hat{\beta}^{\mathrm{M}}$. The consistency of $\hat{\beta}^{\mathrm{CML}}$ depends on $h(\cdot)$ being the inverse logit function. It is well-known, however, that the fixed effects estimator $\hat{\beta}^{\mathrm{FE}}$ is not consistent under model (3), where the u_i are treated as unknown parameters and the m_i may be small (1, p.496). We therefore do not attempt here to approximate the bias of the corresponding estimator of T_y, although we shall consider this estimator in the simulation study in §6.

Let q_{ij}^{M} and q_{ij}^{CML} denote the values of $\hat{q}_{ij}^{\mathrm{M}}$ and $\hat{q}_{ij}^{\mathrm{CML}}$ obtained when $\hat{\beta}^{\mathrm{M}}$ and $\hat{\beta}^{\mathrm{CML}}$ are replaced by β^{M} and β^{RE} respectively and let $q_{ij}^{\mathrm{RE}} = h(x_{ij}\beta^{\mathrm{RE}} + \tilde{u}_i^{\mathrm{RE}})^{-1}$, where $\tilde{u}_i^{\mathrm{RE}}$ is the limiting value of $\hat{u}_i^{\mathrm{RE}}$, with $(\hat{\beta}^{\mathrm{RE}}, \hat{\tau}^{2\mathrm{RE}})$ replaced by $(\beta^{\mathrm{RE}}, \tau^{2\mathrm{RE}})$.

Let $\hat{T}_y$ in (1) be denoted $\hat{T}_y^{\mathrm{M}}$, $\hat{T}_y^{\mathrm{RE}}$ or $\hat{T}_y^{\mathrm{CML}}$ when $\hat{q}_{ij} = \hat{q}_{ij}^{\mathrm{M}}$, $\hat{q}_{ij}^{\mathrm{RE}}$ or $\hat{q}_{ij}^{\mathrm{CML}}$, respectively, and let $\tilde{T}_y = \sum d_{ij} q_{ij} R_{ij} y_{ij}$ be denoted $\tilde{T}_y^{\mathrm{M}}$, $\tilde{T}_y^{\mathrm{RE}}$ or $\tilde{T}_y^{\mathrm{CML}}$ when $q_{ij} = q_{ij}^{\mathrm{M}}$, q_{ij}^{RE} or q_{ij}^{CML}, respectively. We are interested in the biases of $\hat{T}_y^{\mathrm{M}}$, $\hat{T}_y^{\mathrm{RE}}$ and $\hat{T}_y^{\mathrm{CML}}$ when nonresponse is either missing at random or cluster-specific nonignorable. We shall approximate these by the biases of $\tilde{T}_y^{\mathrm{M}}$, $\tilde{T}_y^{\mathrm{RE}}$ and $\tilde{T}_y^{\mathrm{CML}}$, for which expressions are given in the following result, together with a form (in (7)) of asymptotic equivalence between ($\hat{T}_y^{\mathrm{M}}$, $\hat{T}_y^{\mathrm{RE}}$, $\hat{T}_y^{\mathrm{CML}}$) and ($\tilde{T}_y^{\mathrm{M}}$, $\tilde{T}_y^{\mathrm{RE}}$, $\tilde{T}_y^{\mathrm{CML}}$). Since $\tilde{T}_y^{\mathrm{RE}}$ is biased in general when nonresponse is missing at random, it will also be when it is cluster-specific nonignorable and thus we do not present a bias expression for that case.

Theorem 1 *Under conditions given below expression (11),*

$$N_0^{-1}\hat{T}_y^{\mathrm{M}} = N_0^{-1}\tilde{T}_y^{\mathrm{M}} + o_p(1), \quad N_0^{-1}\hat{T}_y^{\mathrm{RE}} = N_0^{-1}\tilde{T}_y^{\mathrm{RE}} + o_p(1),$$

$$N_0^{-1}\hat{T}_y^{\mathrm{CML}} = N_0^{-1}\tilde{T}_y^{\mathrm{CML}} + o_p(1). \tag{7}$$

When nonresponse is missing at random,

$$E_p\{E_\xi(\tilde{T}_y^{\mathrm{M}} - T_y)\} = 0, \tag{8}$$

where E_p and E_ξ denote expectation under the sampling design and the model, respectively.

When nonresponse is cluster-specific nonignorable and model (5) holds,

$$E_p\{E_\xi(\tilde{T}_y^{\mathrm{M}} - T_y)\} = \sum_U \frac{E_\xi\{h(x_{ij}\beta + u_i)E_\xi(\nu_i \mid u_i)\}}{E_\xi\{h(x_{ij}\beta + u_i)\}}. \tag{9}$$

When nonresponse is missing at random and model (5) holds,

$$E_p\{E_\xi(\tilde{T}_y^{\mathrm{RE}} - T_y)\} = E_p[\sum_s d_{ij}\{E_\xi(q_{ij}^{\mathrm{RE}} R_{ij}) - 1\}x_{ij}\lambda]. \tag{10}$$

When nonreponse is either missing at random or cluster-specific nonignorable and when model (5) need not necessarily hold,

$$E_p\{E_\xi(\tilde{T}_y^{\mathrm{CML}} - T_y)\} = 0. \tag{11}$$

The expressions in (7) may be proved through Taylor expansion of the $\hat{q}_{ij}$ as functions of $\hat{\beta}$ around β, and $\hat{\tau}^2$ around τ^2 in the case of $\hat{q}_{ij}^{\mathrm{RE}}$, under the model ξ for the R_{ij}. The proof uses the consistency of the $\hat{\beta}$ and $\hat{\tau}^2$. It also assumes that the functions of $\hat{\beta}$ and $\hat{\tau}^2$ have continuous first derivatives and that $N_0^{-1}\sum_s d_{ij}R_{ij}\delta_{ij}y_{ij} = O_p(1)$, for the derivatives δ_{ij} of each of these functions with respect to $\hat{\beta}$, and $\hat{\tau}^2$ in the case of $\hat{q}_{ij}^{\mathrm{RE}}$, evaluated at their true values.

Expressions (8)–(11) follow by direct evaluation. For illustration, the key result (11) is obtained by noting that when nonresponse is cluster-specific nonignorable we have

$$E_p\{E_\xi(\tilde{T}_y^{\mathrm{CML}} - T_y)\} = E_pE_\xi[\sum_{(i,j)\in s} d_{ij}\{E_\xi(q_{ij}^{\mathrm{CML}}R_{ij} \mid u_i) - 1\}E_\xi(y_{ij} \mid u_i)]$$

and

$$E_\xi(q_{ij}^{\mathrm{CML}}R_{ij} \mid u_i) = E_\xi\{E_\xi(q_{ij}^{\mathrm{CML}}R_{ij} \mid R_{i+}, u_i) \mid u_i\} = E_\xi\{q_{ij}^{\mathrm{CML}}E_\xi(R_{ij} \mid R_{i+}) \mid u_i\} = 1.$$

Hence, the proposed weighting approach results in removal of bias for $\tilde{T}_y^{\mathrm{CML}}$ when nonresponse is cluster-specific nonignorable or is missing at random.

By comparison, we see from expressions (8) and (9) that $\tilde{T}_y^{\mathrm{M}}$ is unbiased only when nonresponse is missing at random. When nonresponse is cluster-specific nonignorable, response propensity weighting may lead to bias. This is not surprising since this weighting approach makes no attempt to control

for clustering. The bias expression in (9) will generally be non-zero when u_i and ν_i are correlated. If $h(\cdot)$ is an increasing function, as for the logit function, and u_i and ν_i are positively correlated then we may expect the bias to be positive.

Turning to weighting based on predicted random effects, we observe that $\tilde{T}_y^{\mathrm{RE}}$ may be biased, even when nonreponse is missing at random. This may occur when the $E_\xi(q_{ij}^{\mathrm{RE}} R_{ij}) = E_\xi\{h(x_{ij}\beta^{\mathrm{RE}} + \tilde{u}_i^{\mathrm{RE}})^{-1} R_{ij}\}$ differ from unity. As the m_i increase, the $\tilde{u}_i^{\mathrm{RE}}$ will approach u_i and $E_\xi\{h(x_{ij}\beta^{\mathrm{RE}} + \tilde{u}_i^{\mathrm{RE}})\}$ will approach 1. But for small m_i this will not generally be the case. The problem is that, when m_i is small, there may be correlation between R_{ij} and $\tilde{u}_i^{\mathrm{RE}}$, which depends on $R_{i1}, \ldots, R_{im_i}$, conditional on u_i. Assuming again that $h(\cdot)$ is increasing, we may expect that $h(x_{ij}\beta + \tilde{u}_{\mathrm{I}}^{\mathrm{RE}})$ and R_{ij} are negatively correlated, conditional on u_i, suggesting that the biasing effect when nonresponse is missing at random will be to shrink $\tilde{T}_y^{\mathrm{RE}}$ towards zero. Even if nonresponse does not depend on x_{ij}, so that $\beta^{\mathrm{RE}} = 0$, we may still have bias, unless τ^2 is also zero.

Our discussion in this section has so far related to the basic weighted estimator $\tilde{T}_y$. We now consider parallel results to those in Theorem 1 for the regression estimator $\tilde{T}_{y,\mathrm{reg}}$ in (2). It follows, by analogy to (8) and (9) that

$$E_p\{E_\xi(\tilde{T}_x^{\mathrm{M}} - \hat{T}_{xs})\} = 0,$$

whether nonresponse is missing at random or cluster-specific nonignorable. Hence the bias properties of $\hat{T}_{y,\mathrm{reg}}^{\mathrm{M}}$ follow those of $\hat{T}_y^{\mathrm{M}}$, that is these estimators are both approximately unbiased when nonresponse is missing at random but are subject to potential bias (as in (9)) when it is cluster-specific nonignor-

able. Turning to $\hat{T}_{y,\text{reg}}^{\text{RE}}$, we note that, analogous to (10) we have:

$$E_p\{E_\xi(\tilde{T}_x^{\text{RE}} - \hat{T}_{xs})\} = E_p[\sum_s d_{ij}\{E_\xi(q_{ij}^{\text{RE}} R_{ij}) - 1\}x_{ij}].$$

Under the assumption that sampling is noninformative, we have $\hat{\lambda} = \lambda + o_p(1)$ under model (5) and it follows that, in the limit, $\hat{T}_{y,\text{reg}}^{\text{RE}}$ is approximately unbiased when nonresponse is missing at random, unlike $\hat{T}_y^{\text{RE}}$. This depends on the truth of (5), unlike the approximate unbiasedness of $\hat{T}_y^{\text{CML}}$. When nonresponse is cluster-specific nonignorable, $\hat{T}_{y,\text{reg}}^{\text{RE}}$ will be subject to potential bias like $\hat{T}_y^{\text{RE}}$. Its approximate form, analogous to (10), is:

$$E_p\{E_\xi(\tilde{T}_{y,\text{reg}}^{\text{RE}} - T_y)\} = E_p \sum_s d_{ij} E_\xi[\{E_\xi(q_{ij}^{\text{RE}} R_{ij}) - 1\}\nu_i].$$

Finally, we note that the large sample bias of $\hat{T}_{y,\text{reg}}^{\text{CML}}$ follows that of $\hat{T}_y^{\text{CML}}$, with both being zero when nonresponse is cluster-specific nonignorable or missing at random.

5 Variance Estimation

In this section we outline an approach to estimating the variance of the proposed estimator $\hat{T}_y^{\text{CML}}$. We adopt a linearization approach in which a linearized variable z_{ij} is determined, such that the variance of $\hat{T}_y^{\text{CML}}$ may be approximated by the variance of $\sum_s z_{ij}$ (21). The variance estimator may then be constructed following a standard survey sampling approach for linear statistics. In order to construct z_{ij} we first recall from §3 that we may write the conditional logistic weight as a function of $\hat{\beta}_1^{\text{CML}}$. As a first order Taylor expansion we have

$$\hat{T}_y^{\mathrm{CML}} = \tilde{T}_y^{\mathrm{CML}} + \sum_s d_{ij} R_{ij} y_{ij} \delta_{ij}(\beta_1^{\mathrm{RE}})(\hat{\beta}_1^{\mathrm{CML}} - \beta_1^{\mathrm{RE}}), \tag{12}$$

where $\delta_{ij}(\beta_1) = \partial q_{ij}^{\mathrm{CML}}(\beta_1)/\partial\beta_1$.

A Taylor expansion of $\hat{\beta}_1^{\mathrm{CML}}$ is

$$\hat{\beta}_1^{\mathrm{CML}} = \beta_1^{\mathrm{RE}} + I(\beta_1^{\mathrm{RE}})^{-1} U(\beta_1^{\mathrm{RE}}), \tag{13}$$

where $I(\beta_1)$ is the information matrix (cf. 7):

$$\begin{aligned} I(\beta_1) &= -\frac{\partial U(\beta_1)}{\partial \beta_1} \\ &= \sum_{i=1}^{n} \left\{ \frac{\sum_{r_i \in B_{2i}} a_i(\beta_1) x_{1i+r}^T x_{1i+r}}{\sum_{r_i \in B_{2i}} a_i(\beta_1)} - \frac{\sum_{r_i \in B_{2i}} a_i(\beta_1) x_{1i+r}^T}{\sum_{r_i \in B_{2i}} a_i(\beta_1)} \frac{\sum_{r_i \in B_{2i}} a_i(\beta_1) x_{1i+r}}{\sum_{r_i \in B_{2i}} a_i(\beta_1)} \right\}. \end{aligned}$$

Substituting (13) in (12) , we obtain the linearized variable as

$$z_{ij} = d_{ij} R_{ij} q_{ij}^{\mathrm{CML}} y_{ij} + \{ \sum_{(i,j) \in s} d_{ij} R_{ij} y_{ij} \delta_{ij}(\beta_1^{\mathrm{RE}}) \} I(\beta_1^{\mathrm{RE}})^{-1} U_{ij}(\beta_1^{\mathrm{RE}}). \tag{14}$$

In order to construct a variance estimator it is necessary to replace β_1^{RE} in (14) by $\hat{\beta}_1^{\mathrm{CML}}$ and q_{ij}^{CML} by $\hat{q}_{ij}^{\mathrm{CML}}$. The first term in (14) would be the linearized variable, were the weight $\hat{q}_{ij}^{\mathrm{CML}}$ to be treated as fixed. The remaining term provides an adjustment for the fact that $\hat{q}_{ij}^{\mathrm{CML}}$ is an estimator. An analogous expression to (14) is provided in **(author?)** [10, Theorem 1] for the case when the weight $\hat{q}_{ij}^{\mathrm{M}}$ is used and there is no clustering.

One commonly used estimator of the variance of a linear statistic, in the case of stratified selection of clusters, is given by

$$\nu = \sum_{h=1}^{H} \frac{n_h}{(n_h - 1)} \sum_{i \in s_h} (z_{i+} - \bar{z}_h)^2, \tag{15}$$

where $z_{i+} = \sum_{j=1}^{m_i} z_{ij}$, $\bar{z}_h = n_h^{-1} \sum_{i \in s_h} z_{i+}$, and s_h denotes the set of n_h clusters drawn in stratum h for $h = 1, \ldots, H$ and it is assumed that $n_h \geq 2$ for each h. This effectively assumes that the z_{i+} may be treated as independent and identically distributed within strata, which may be a reasonable approximation for many sampling schemes where clusters are selected as primary sampling units and the fraction of primary sampling units selected in each stratum is small and when nonresponse is independent between clusters. A practical advantage of this approach is that it allows for clustered nonresponse as well as complex forms of sampling within clusters.

6 Simulation Study

A small simulation study is now undertaken to illustrate the properties of the four weighted point estimators in §3 and the variance estimator derived in the previous section. We created a finite population with $N = 200$ and $M_i = M = 10$, where the values of x_{ij}, R_{ij} and y_{ij} were generated, respectively, from: $x_{ij} = (1, x_{1ij})$, $x_{1ij} \sim N(2, 1)$, truncated below by 0 and above by 4; $R_{ij} \sim$ model (3) with $h(\cdot)$ the inverse logit function, $\beta = (\beta_0, \beta_1)^T$, τ^2=1; $y_{ij} \sim$ model (5) with $\lambda = 5$, $\epsilon_{ij} \sim N(0, 1)$ and $\nu_i = \alpha_i + \delta u_i$, where $\alpha_i \sim N(0, 1)$.

Since α_i, u_i and ϵ_{ij} are generated independently, nonresponse is missing at random if $\delta = 0$ and cluster-specific nonignorable otherwise. We consider four possible sets of values for the parameters $\beta = (\beta_0, \beta_1)^T$ and δ, representing different missing data mechanisms: (i) MCAR: $(\beta_0, \beta_1) = (1, 0)$, $\delta = 0$, (ii) MAR: $(\beta_0, \beta_1) = (0, 0.5)$, $\delta = 0$, (iii) CSNI1: $(\beta_0, \beta_1) = (1, 0)$, $\delta = 5$ and (iv) CSNI2: $(\beta_0, \beta_1) = (0, 0.5)$, $\delta = 5$.

Both choices of (β_0, β_1) generate an overall response rate of about 70%. We drew 1000 samples using (a) simple random cluster sampling with $n = 50$, $m_i = M = 10$ and (b) two stage sampling, with simple random sampling at each stage with $n = 50$, $m_i = 5$. For each of the 1000 replications, new values of the R_{ij} were generated along with the new samples. Other finite population values were kept fixed. Any samples for which $R_{i+} = 0$ for some i were rejected.

Simulation results are presented in Tables 1 and 2 for these four missing data mechanisms, for the four weighting approaches in §3 and for the two choices of (n, m_i) above. The relative bias reported in the tables is the mean of the estimated total across the 1000 replications less the true finite population total, divided by this population total. The relative standard error reported is the standard deviation of the estimated total across the 1000 replications divided by the true finite population total. To help understand the impact of estimating β^{RE} by $\hat{\beta}^{\mathrm{CML}}$, we also include results for $\tilde{T}_y^{\mathrm{CML}}$, i.e., $\hat{T}_y^{\mathrm{CML}}$ with $\hat{\beta}^{\mathrm{CML}}$ replaced by β^{RE}.

We comment first on the bias properties. There is no evidence of bias in $\tilde{T}_y^{\mathrm{CML}}$, as should be the case from (11), nor is there evidence of bias in $\hat{T}_y^{\mathrm{CML}}$. The asymptotic equivalence of $\tilde{T}_y^{\mathrm{CML}}$ and $\hat{T}_y^{\mathrm{CML}}$ in (7) holds here to a suitable approximation. We observe evidence of bias in $\hat{T}_y^{\mathrm{M}}$ under missing data mechanisms (iii) and (iv), where $\delta = 5$, but not for mechanisms (i) and (ii), as expected, in approximation, from (8) and (9). Since the value, 0.5, of the intra-cluster correlation implied by $\delta = 5$ is fairly large, we repeated the simulations with $\delta = 1$, implying an intra-cluster correlation of 0.07, and found the bias of $\hat{T}_y^{\mathrm{M}}$ to be reduced but still clearly the worst of all

Table 1: Simulation estimates, based on 1000 replicates, of relative bias, standard errors and root mean squared errors of weighted estimates of totals for alternative weighting methods and missing data mechanisms. Cluster sampling with $n = 50$, $m_i = 10$

Missing data mechanism	Weighting method	Relative Bias (%)	Relative SE (%)	Relative RMSE (%)
MCAR	Response propensity	(−0.1)	2.3	2.3
	Fixed effects	(0.0)	2.5	2.5
	CML, estimated	(0.0)	2.5	2.5
	CML, true parameter	(0.0)	2.9	2.9
	Random effects	−2.8	2.4	3.6
MAR	Response propensity	(−0.1)	2.3	2.3
	Fixed effects	(0.1)	2.3	2.3
	CML, estimated	(0.1)	2.3	2.3
	CML, true parameter	(0.1)	2.5	2.5
	Random effects	−2.4	2.3	3.3
CNI1	Response propensity	11.1	6.2	12.7
	Fixed effects	(−0.1)	6.3	6.3
	CML, estimated	(−0.1)	6.3	6.3
	CML, true parameter	(−0.2)	6.4	6.4
	Random effects	2.2	6.1	6.4
CNI2	Response propensity	11.4	6.2	12.9
	Fixed effects	(−0.1)	6.3	6.3
	CML, estimated	(−0.1)	6.3	6.3
	CML, true parameter	(−0.1)	6.4	6.4
	Random effects	2.7	6.0	6.6

Parentheses surround estimates which are within two simulation standard errors of 0. SE, standard error; RMSE, root mean squared error; MCAR, missing completely at random; MAR, missing at random; CSNI, cluster-specific nonignorable; CML, conditional maximum likelihood.

Table 2: Simulation estimates, based on 1000 replicates, of relative bias, standard errors and root mean squared errors of weighted estimates of totals for alternative weighting methods and missing data mechanisms. Two stage sampling with $n = 50$, $m_i = 5$

Missing data mechanism	Weighting method	Relative Bias (%)	Relative SE (%)	Relative RMSE (%)
MCAR	Response propensity	(−0.1)	3.2	3.2
	Fixed effects	(0.2)	3.7	3.7
	CML, estimated	(0.2)	3.6	3.6
	CML, true parameter	(0.1)	4.0	4.0
	Random effects	−3.1	3.3	4.5
MAR	Response propensity	(0.0)	3.1	3.1
	Fixed effects	(0.2)	3.2	3.2
	CML, estimated	(0.2)	3.2	3.2
	CML, true parameter	(0.1)	3.4	3.4
	Random effects	−2.6	3.2	4.1
CNI1	Response propensity	10.4	6.6	12.3
	Fixed effects	(0.0)	6.7	6.7
	CML, estimated	(0.0)	6.6	6.6
	CML, true parameter	(−0.1)	7.0	7.0
	Random effects	3.6	6.4	6.4
CNI2	Response propensity	10.8	6.6	12.6
	Fixed effects	(−0.1)	6.8	6.8
	CML, estimated	(0.0)	6.7	6.7
	CML, true parameter	(−0.1)	7.0	7.0
	Random effects	4.2	6.4	7.6

Parentheses surround estimates which are within two simulation standard errors of 0. SE, standard error; RMSE, root mean squared error; MCAR, missing completely at random; MAR, missing at random; CSNI, cluster-specific nonignorable; CML, conditional maximum likelihood.

estimators. As anticipated in §4, there is evidence of negative bias in $\hat{T}_y^{\mathrm{RE}}$ under mechanisms (i) and (ii), when $\delta = 0$. As δ increases we found the bias of $\hat{T}_y^{\mathrm{RE}}$ to shift in the direction towards that of $\hat{T}_y^{\mathrm{M}}$. For $\delta = 1$ we found it still negative. For missing data mechanisms (iii) and (iv) with $\delta = 5$ we see in Tables 1 and 2 that the bias is positive, as for $\hat{T}_y^{\mathrm{M}}$. The bias of $\hat{T}_y^{\mathrm{RE}}$ under mechanism (ii) does decline as m_i increases but, repeating the simulation for $m_i = 20$, we still find a relative bias of -1.7% so the decline is not rapid. We presented no theory for $\hat{T}_y^{\mathrm{FE}}$ in §4. We observe in Tables 1 and 2 that it seems to share a similar absence of bias to $\hat{T}_y^{\mathrm{CML}}$.

Turning to the standard errors, we first compare $\tilde{T}_y^{\mathrm{CML}}$ and $\hat{T}_y^{\mathrm{CML}}$. Some results in the literature (e.g. 15, 10) suggest that the use of an unweighted estimate of the response propensity rather than its true value can, paradoxically, reduce variance and this is indeed observed in Tables 1 and 2 in all cases. There is some evidence in these tables that the variance of $\hat{T}_y^{\mathrm{CML}}$ can be a little larger than those of $\hat{T}_y^{\mathrm{M}}$ and $\hat{T}_y^{\mathrm{RE}}$, but the smaller bias of $\hat{T}_y^{\mathrm{CML}}$ offsets this effect. The root mean squared error of $\hat{T}_y^{\mathrm{CML}}$ is always smaller than that of $\hat{T}_y^{\mathrm{RE}}$, sometimes substantially so, and it is also considerably smaller than that of $\hat{T}_y^{\mathrm{M}}$ for the cluster-specific nonignorable cases. Of course, the relative root mean squared error and the extent of the bias-variance trade-off will depend on sample size.

The somewhat larger variances of $\hat{T}_y^{\mathrm{CML}}$ and $\hat{T}_y^{\mathrm{FE}}$ observed in Tables 1 and 2 seem to be associated with greater variability in the weights $\hat{q}_{ij}^{\mathrm{CML}}$ than the $\hat{q}_{ij}^{\mathrm{M}}$ or $\hat{q}_{ij}^{\mathrm{RE}}$. These weights are truncated below by unity and it is the very large weights that are of potential concern. Comparison of the weights resulting from the use of $\hat{\beta}^{\mathrm{CML}}$ versus its true value suggests that the estimation

of β is not a major source of the weight variability in the simulation study. Large weights $\hat{q}_{ij}^{\mathrm{CML}}$ arise primarily when one of the conditional probabilities of response in (6) is small. This may be partly because the response rate in the cluster is low, perhaps by chance, which will also lead to a larger value of $\hat{q}_{ij}^{\mathrm{FE}}$, and partly because an outlying value of x_{ij}, with $x_{ij}\beta$ unusually low, is included in the sample of m_i units and that unit responds.

We now turn to results on the regression estimator $\hat{T}_{y,\mathrm{reg}}$ in Table 3. Results for $\hat{T}_{y,\mathrm{reg}}^{\mathrm{M}}$ were almost identical to those for $\hat{T}_{y}^{\mathrm{M}}$, as anticipated for bias in §4, and are thus not included in the table. Results for $\hat{T}_{y,\mathrm{reg}}^{\mathrm{FE}}$ and $\tilde{T}_{y,\mathrm{reg}}^{\mathrm{CML}}$ were almost identical to those for $\hat{T}_{y,\mathrm{reg}}^{\mathrm{CML}}$ and are also thus not included, although it is of interest to note that the reduction in variance of $\hat{T}_{y,\mathrm{reg}}^{\mathrm{CML}}$ vs. $\tilde{T}_{y,\mathrm{reg}}^{\mathrm{CML}}$ observed in Tables 1 and 2 seems to disappear once regression estimation is used. Table 3 shows how the bias of $\hat{T}^{\mathrm{RE}}$ under the first two missing data mechanisms is removed by regression estimation, as anticipated in §4. However, $\hat{T}_{y,\mathrm{reg}}^{\mathrm{RE}}$ remains biased under the cluster-specific nonignorable mechanisms. As expected, regression estimation does lead to some reduction in variance. As in Tables 1 and 2, $\hat{T}_{y,\mathrm{reg}}^{\mathrm{RE}}$ does show some slight variance gains relative to $\hat{T}_{y,\mathrm{reg}}^{\mathrm{CML}}$ but this is more than offset by bias and the root mean squared error of $\hat{T}_{y,\mathrm{reg}}^{\mathrm{CML}}$ is in no cases greater than that of $\hat{T}_{y,\mathrm{reg}}^{\mathrm{RE}}$.

Finally we present in Table 4 some results on the estimation of the variance of $\hat{T}_{y}^{\mathrm{CML}}$ for the case of cluster sampling. We consider two versions of the variance estimator in (15). Both include a finite population correction $(1-n/N)$. Estimator (i) includes only the first term from (14) and so treats the weight $\hat{q}_{ij}^{\mathrm{CML}}$ as fixed. Estimator (ii) includes both terms in (14) and so allows for variation in $\hat{\beta}_{1}^{\mathrm{CML}}$. Allowing for uncertainty in estimation of β_1

Table 3: Simulation estimates, based on 1000 replicates, of relative bias, standard errors and root mean squared errors of regression weighted estimates of totals for alternative weighting methods and missing data mechanisms

Missing data mechanism	Weighting method	Relative Bias (%)	Relative SE (%)	Relative RMSE (%)
n=50, m_i=10				
MCAR	CML, estimated	(0.1)	2.3	2.3
	Random effects	(0.0)	2.3	2.3
MAR	CML, estimated	(0.1)	2.3	2.3
	Random effects	(0.0)	2.3	2.3
CSNI1	CML, estimated	(−0.1)	6.2	6.2
	Random effects	5.0	5.9	7.8
CSNI2	CML, estimated	(−0.1)	6.3	6.7
	Random effects	4.2	6.4	7.6
n=50, m_i=5				
MCAR	CML, estimated	(0.1)	3.2	3.2
	Random effects	(0.0)	3.2	3.2
MAR	CML, estimated	(0.2)	3.1	3.1
	Random effects	(0.1)	3.1	3.1
CSNI1	CML, estimated	(−0.1)	6.5	6.5
	Random effects	6.8	6.3	9.3
CSNI2	CML, estimated	(0.0)	6.7	6.7
	Random effects	7.0	6.3	9.4

Parentheses surround estimates which are within two simulation standard errors of 0. SE, standard error; RMSE, root mean squared error; MCAR, missing completely at random; MAR, missing at random; CSNI, cluster-specific nonignorable; CML, conditional maximum likelihood.

Table 4: Simulation estimates, based on 1000 replicates, of relative bias, standard errors and root mean squared errors of standard error estimators for conditional maximum likelihood estimation of totals for alternative missing data mechanisms. Cluster sampling with $n = 50$, $m_i = 10$

Missing data mechanism	Standard Error Estimator	Relative Bias (%)	Relative SE (%)	Relative RMSE (%)
MCAR	Treating weight as fixed	10.7	13.7	17.4
	Allowing for variation in $\hat{\beta}$	−3.2	10.3	10.8
MAR	Treating weight as fixed	3.9	10.3	11.0
	Allowing for variation in $\hat{\beta}$	−1.0	9.3	9.4
CSNI1	Treating weight as fixed	3.1	10.3	10.8
	Allowing for variation in $\hat{\beta}$	1.0	9.9	9.9
CSNI2	Treating weight as fixed	1.4	13.1	13.1
	Allowing for variation in $\hat{\beta}$	(0.2)	12.0	12.0

Parentheses surround estimates which are within two simulation standard errors of 0. SE, standard error; RMSE, root mean squared error; MCAR, missing completely at random; MAR, missing at random; CSNI, cluster-specific nonignorable.

reduces the variance estimates, as is appropriate given that the variance of $\hat{T}_y^{\mathrm{CML}}$ is smaller than that of $\tilde{T}_y^{\mathrm{CML}}$ in Tables 1 and 2. Estimator (ii) does display significant if relatively modest bias in three out of four cases. This may be attributed to the small between-cluster degrees of freedom. Estimator (i) has larger root mean squared error than estimator (ii) in each case, but is always conservative and this may be attractive in some applications, especially since this estimator is simpler to compute.

7 Extension to Observational Studies with Clustered Treatment Assignment

The use of the conditional maximum likelihood estimator to correct for large-sample bias may be extended to treatment effect estimation in observational studies with clustered treatment assignment . Suppose that a_{ij} denotes a 0-1 treatment assignment variable which is subject to clustering and obeys

model (3) with R_{ij} replaced by a_{ij}, where u_i is the random effect term. Let $y_{ij} = (y_{0ij}, y_{1ij})$ denote the potential outcomes under either treatment (16). Write $y_i = (y_{i1}^T, \ldots, y_{im_i}^T)^T$, $a_i = (a_{i1}, \ldots, a_{im_i})$ and $x_i = (x_{i1}^T, \ldots, x_{im_i}^T)^T$ and define the conditional propensity score as $e_{ij} = \text{pr}(a_{ij} = 1 \mid a_{i+}, x_i)$, where $a_{i+} = \sum_j a_{ij}$. Just as in (6), e_{ij} is free of u_i and observable, subject to parameter estimation. Let $e_i = (e_{i1}, \ldots, e_{im_i})^T$. The treatment assignment assumption corresponding to cluster-specific nonignorable nonresponse is that a_i and y_i are conditionally independent given u_i and x_i. We might refer to this as cluster specific nonignorable treatment assignment. Then we may show, corresponding to Theorem 3 of **(author?)** [16], that a_i and y_i are conditionally independent given e_i. This enables treatment effects to be estimated consistently under cluster-specific nonignorable treatment assignment using the conditional propensity score in an analogous way to the use of standard propensity scores. This will be of most interest when the potential outcomes also display clustering and have associated random effects which are correlated with u_i conditional on x_i.

8 Discussion

We have shown, theoretically and with simulation evidence, that an attempt to allow for clustered response via the introduction of predicted random effects into the estimated probability of response can induce negative relative bias in the inverse probability weighted estimator when nonresponse is missing at random and the cluster sizes are not large. In our simulation study we found a negative relative bias of about 2% for cluster sizes of between 5 and 20, declining to about 1% as these sizes increased to 50. In such cir-

cumstances, if the missing at random assumption is plausible, it seems safer to employ simple response propensity weights based upon a marginal model for response. If nonresponse is cluster-specific nonignorable but not missing at random then the latter approach may be subject to bias. We found the relative bias could be as high as 10% with high intra-cluster correlations in both the survey variable and the nonresponse process. With a more modest intra-cluster correlation of about 0.01 in the survey variable, we found this bias reduced to about 2%. The proposed conditional maximum likelihood approach removes this bias, when the number of sampled clusters is large even if the cluster sizes are small. We have also shown in §7 how this conditional maximum likelihood approach might be extended to the estimation of treatment effects in observational studies.

In addition to its bias correction advantage, the conditional maximum likelihood approach is not dependent on the assumption that the u_i term in (3) is Gaussian, nor that it is independent of x_{ij}. There are, however, potential disadvantages to the conditional maximum likelihood approach. It depends on the logistic form of the model in (3). It becomes increasingly computationally intensive as the sizes of the sets B_{1ij} and B_{2i} grow. And, as we observed in the simulation study, it can lead to more variable weights and lower efficiency.

Efficiency considerations need not be overriding. There is considerable interest among survey researchers in methods which may help detect or correct for bias when sample sizes are large. Moreover, the efficiency of the simple estimator in (1) may be expected to be improved by the use of the regression estimator in (2). The improvement will depend on how well x_{ij} predicts y_{ij}.

The regression estimator also has the double robustness benefit, mentioned in the introduction, that consistency may be achievable when nonresponse is missing at random even if the nonresponse model is misspecified, provided the model for the survey variable in (5) holds. Furthermore, like the simple estimator, it may be expressed as a weighted estimator with weights which do not depend on y_{ij}. This has various practical advantages in multipurpose surveys. For an estimator outside this class of weighted estimators, which is efficient even under cluster-specific nonignorable nonresponse, see **(author?)** [22]. A simpler modification of the conditional maximum likelihood approach would be to use what **(author?)** [12] calls response propensity stratification, forming classes by grouping values of $\hat{q}_{ij}^{\mathrm{CML}}$ and then replacing this weight by the inverse observed response rate in the group. This approach may be less sensitive to the logistic link function assumption and may help smooth large values of $\hat{q}_{ij}^{\mathrm{CML}}$.

In the simulation study we observed that the fixed effects estimator performed similarly to the conditional maximum likelihood estimator and it may be that in practice it will often provide a reasonable proxy to this estimator, while not requiring such strong model assumptions nor so much computation.

Acknowledgement

Research was supported by the Economic and Social Research Council. We are grateful to two referees for constructive comments.

References

[1] A. Agresti. *Categorical Data Analysis*. Hoboken: Wiley, second edition, 2002.

[2] H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–72, 2005.

[3] W. Cao, A. A. Tsiatis, and M. Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96:723–34, 2009.

[4] C.-M. Cassel, C.-E. Särndal, and J. H. Wretman. Some uses of statistical models in connection with the nonresponse problem. In W. G. Madow and I. Olkin, editors, *Incomplete Data in Sample Surveys III. Symposium on Incomplete Data, Proceedings*, pages 143–60. New York: Academic Press, 1983.

[5] P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford: Oxford University Press, second edition, 2002.

[6] G. B. Durrant and F. Steele. Multilevel modelling of refusal and non-contact nonresponse in household surveys: evidence from six UK government surveys. *J. Roy. Statist. Soc., Series A*, 172:361–81, 2009.

[7] M. P. Fay, B. I. Graubard, L. S. Freedman, and D. N. Midthune. Conditional logistic regression with sandwich estimators: application to meta analysis. *Biometrics*, 54:195–208, 1998.

[8] W. A. Fuller. *Sampling Statistics*. Hoboken: Wiley, 2009.

[9] D. Y. J. Kang and J. L. Schafer. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion and rejoinder). *Statist. Sci.*, 22:523–80, 2007.

[10] J. K. Kim and J. J. Kim. Nonresponse weighting adjustment using estimated probability. *Can. J. Statist.*, 35:501–14, 2007.

[11] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73:13–22, 1986.

[12] R. J. A. Little. Survey nonresponse adjustments for estimates of means. *Internat. Statist. Rev.*, 54:139–57, 1986.

[13] R. J. A. Little. Missing data adjustments in large surveys (with discussion). *J. Bus. Econ. Statist.*, 6:267–301, 1988.

[14] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Hoboken: Wiley, second edition, 2002.

[15] P. R. Rosenbaum. Model-based direct adjustment. *J. Amer. Statist. Ass.*, 82:387–94, 1987.

[16] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[17] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.

[18] C.-E. Särndal and S. Lundström. *Estimation in Surveys with Nonresponse*. Chichester: Wiley, 2005.

[19] J. Shao. Handling survey nonresponse in cluster sampling. *Survey Methodology*, 33:81–85, 2007.

[20] A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

[21] R. S. Woodruff. A simple method for approximating the variance of a complicated estimate. *J. Amer. Statist. Ass.*, 66:411–4, 1971.

[22] Y. Yuan and R. J. A. Little. Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response. *Appl. Statist.*, 56:79–97, 2007.

[23] Y. Yuan and R. J. A. Little. Model-based inference for two-stage cluster samples subject to nonignorable item nonresponse. *J. Off. Statist.*, 24:193–211, 2008.