Nuanpan Lawson, Chris Skinner

Estimation of a cluster-level regression model under nonresponse within clusters

Article (Published version) (Refereed)

Original citation:
Lawson, Nuanpan and Skinner, Chris (2017) *Estimation of a cluster-level regression model under nonresponse within clusters.* Metron, 75 (3). pp. 319-331. ISSN 0026-1424

DOI: 10.1007/s40300-017-0120-4

METRON (2017) 75:319–331
DOI 10.1007/s40300-017-0120-4

Estimation of a cluster-level regression model under nonresponse within clusters

Nuanpan Lawson[1] · **Chris Skinner**[2]

Received: 28 March 2017 / Accepted: 22 August 2017 / Published online: 9 September 2017

Abstract When sample surveys are clustered and subject to non-response, it is possible to study cluster-level association between response rates and cluster-level quantities derived from survey variables. The existence of association may suggest informative nonresponse with possible biasing effects. In this paper, this problem is studied for the case where the aim is to fit a cluster-level regression model. Two possible underlying models for nonresponse with potential biasing effects are considered. Alternative estimators of regression coefficients under these models are proposed. The properties of these estimators are studied in two simulation studies and with real data from a survey of employees, where the clusters consist of workplaces.

Keywords Cluster specific nonignorable nonresponse · Cluster sample · Informative nonresponse · Regression model · Selection

1 Introduction

A feature of nonresponse in clustered survey data is that it is possible to study cluster-level association between response rates and aggregate statistics, such as means or proportions, for survey variables of interest. Thus, if p_i denotes the response rate among elementary sample units in cluster i and $\bar{y}_{ri}$ denotes the mean of a variable Y among responding units within cluster i then it is possible to study the association between these two quantities across clusters, perhaps conditional on some other cluster-level factors. In contrast, no equivalent association can be observed at the elementary unit level (in unclustered data) since Y is missing for nonresponding units.

✉ Chris Skinner
c.j.skinner@lse.ac.uk

1 Department of Applied Statistics, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

2 Department of Statistics, London School of Economics and Political Science, London, UK

The occurrence of such cluster-level association between p_i and $\bar{y}_{ri}$ may be suggestive of some kind of informative nonresponse. In this paper, we consider this issue when the objective is to fit a regression model at the cluster level, as may be of scientific relevance when the clusters are of analytic interest. As a motivating example in Sect. 3, we consider a survey of employees, where the clusters consist of workplaces and there is analytic interest in how the well-being of employees at a workplace depends upon different kinds of innovations at the workplace. Another example is a survey of hospital patients about the quality of care received, where there may be interest in analysis at the hospital level but nonresponse may arise at the individual patient level [9].

We shall be interested in the case where testing for inclusion of p_i or some function of it as a covariate in the model may be used as some kind of diagnostic for informative nonresponse. We shall introduce two models of nonresponse mechanisms which might provide explanations for such association and which lead to bias in the estimation of the regression coefficients if the nonresponse is ignored by simply running the regression on respondent data. Moreover, we shall consider possible ways of controlling for this bias by including p_i or some function of it as a covariate in the model.

The basic approach of ordinary least squares (OLS) estimation using respondent data will not lead to biased estimation of the regression coefficients if nonresponse is conditionally independent of Y given the explanatory variables included in the model (and if unequal sample inclusion probabilities can be ignored). In standard (unclustered) settings, it can be difficult to detect departures from this conditional independence condition without strong modelling assumptions [8, sect. 1.3] . In our clustered setting, however, we suggest that the use of p_i, or some transformation of it, as an auxiliary explanatory variable can provide a relatively simple way to detect at least some forms of informativeness in the nonresponse. In this paper we consider two possible models which might underly such an effect and, for which, the inclusion of the auxiliary variable in the regression offers some control for nonresponse bias under departures from the conditional independence assumption.

Nonresponse in two-stage surveys can operate at either stage and, in this paper, we focus on the problem when nonresponse occurs at the second stage and suppose that data from all sampled clusters are available, even though data for elementary units are missing from many if not all sampled clusters. For simplicity, we suppose that the explanatory variables in the regression are defined at the cluster level and are not missing.

We present a simulation study in which we consider the performance of estimators based on each model under the assumption that the data are generated from either the assumed model or the alternative model. We also present a real application using data from a survey of workplace employment relations in Great Britain.

The analytic focus of this paper differs from the main literature on clustered survey nonresponse which has considered problems of estimation of finite population parameters, such as means or totals, and associated weighting and imputation questions. Thus, Yuan and Little [12,13] proposed model-based inference approaches for both unit and item nonresponse; Skinner and D'Arrigo [11] and Kim et al. [6] considered weighted estimation; Shao [10], Haziza and Rao [3] and Lago and Clark [7] considered imputation.

The formal framework for the paper is set out in Sect. 2 and the motivating example is given in Sect. 3. Possible models which could account for association between $\bar{y}_{ri}$ and p_i are introduced in Sect. 4 and estimators of regression coefficients based on these models are proposed in Sect. 5. The properties of these estimators are studied in simulation studies in Sect. 6 and with real data from the motivating example in Sect. 7. Some concluding comments are given in Sect. 8.

2 Basic set-up and regression model of interest

We consider a clustered population, containing N clusters with M_i elements in cluster $i = 1, 2, \ldots, N$. We suppose that two-stage sampling is employed, where n clusters are selected and m_i elements are selected in the ith sampled cluster ($i = 1, 2, \ldots, N$). Each stage of sampling may involve simple random sampling, for example, but we discuss the role of the sampling scheme more generally later in this section. Without loss of generality, we write the sampled clusters as $i = 1, \ldots, n$ and the sampled elements in cluster i as $j = 1, \ldots, m_i$. We consider regression analysis for an outcome variable Y and a vector of covariates $\mathbf{x} = (1, x_1, \ldots, x_k)'$. We suppose that Y is defined at the element level with y_{ij} denoting the value of Y for the jth population element ($j = 1, 2, \ldots, M_i$) in the ith cluster ($i = 1, 2, \ldots, N$) and $\bar{Y}_i = M_i^{-1} \sum_{j=1}^{M_i} y_{ij}$ denoting the population mean of Y in the ith cluster. We suppose that interest focusses on the dependence of $\bar{Y}_i$ on a vector of covariates $\mathbf{x}$, defined at the cluster level with $\mathbf{x}_i$ denoting the value of $\mathbf{x}$ for the ith cluster.

We define the regression model of interest by

$$E_m(\bar{Y}_i) = \mathbf{x}_i' \beta, \tag{1}$$

where $E_m(.)$ denotes expectation with respect to a model and β is the vector parameter of interest. The expectation is implicitly taken to be conditional on $\mathbf{x}_i$.

We write $R_{ij} = 1$ if y_{ij} is observed and $R_{ij} = 0$ if y_{ij} is missing as a result of nonresponse by element j in cluster i for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. We suppose $\mathbf{x}_i$ is always observed for $i = 1, \ldots, n$. We denote the number of respondents in cluster i by $r_i = \sum_{j=1}^{m_i} R_{ij}$ and the associated response rate by $p_i = r_i / m_i$.

We shall adopt a purely model-based approach to inference about β in this paper. In particular, as our simplest approach to estimating β, we replace $\bar{Y}_i$ by the respondent mean $\bar{y}_{ri} = \sum_{j=1}^{m_i} R_{ij} y_{ij} / r_i$, where we assume $r_i \geq 1$, and estimate the model in (1) using ordinary least squares (OLS) to regress $\bar{y}_{ri}$ on $\mathbf{x}_i$. We refer to the resulting estimator of β as the OLS estimator. We shall allow for the clustered structure of the population in variance estimation by bootstrapping at the cluster level.

There is a large literature on the role of the sampling scheme in inference about regression models. See e.g. Chambers and Skinner [2]. In this paper, we shall assume that the nature of the sampling scheme is such that it can be ignored for inference about β (beyond the allowance for clustering in bootstrap variance estimation and for nonresponse) and, in particular, sampling weights can be ignored. The implicit assumption here is that the sample inclusion probabilities are unrelated to y_{ij} given $\mathbf{x}_i$. We touch on this point again at the end of the paper. Moreover, we shall not explore standard weighting or imputation methods which might be used to handle the problem of nonresponse here. We suppose that nonresponse will be addressed by appropriate choice of the regression model and associated estimation methods.

3 Example: workplace employment relations survey

To illustrate the set-up, we describe here a regression analysis using data from the 2004 Workplace Employment Relations Survey (WERS) [5], where the elementary units consist of employees and the clusters consist of workplaces. The broad analytic objective is to study how the well-being of employees at a workplace is affected by innovations in working practices at the workplace. Our analysis is based on a much fuller consideration of how innovation affects worker well-being, presented in Bryson et al. [1].

A sample of workplaces in Great Britain with at least 5 employees was selected, with face-to-face interviews conducted with senior managers with responsibility for employee relations and personnel matters. These managers then distributed self-completion questionnaires to 25 randomly selected employees at the workplace (or to every employee in workplaces with 5–25 employees). The response rate of employees to this questionnaire was about 60%. We consider here data on 13,500 employees at 1238 workplaces on a job satisfaction variable y_{ij}, derived from responses by employees to questions about satisfaction with various aspects of their job and an innovations variable x_{1i}, derived from responses by the manager to questions asking about changes initiated by management at the workplace. These variables are described in more detail in Sect. 7. We follow [1] in using only data on private sector workplaces.

Fitting the workplace-level regression model in (1) with $\bar{y}_{ri}$ as dependent variable and just x_{1i} and an intercept as independent variables using OLS gives an estimated coefficient of x_{1i} as -0.29 (with standard error 0.06). This negative association between innovation and job satisfaction was also found by Bryson et al. [1] in their much fuller analyses including other explanatory variables as well as an analysis using instrumental variables to control for the possible endogeneity of innovation. If we also include p_i as an explanatory variable in our linear regression model, we find that it has a least squares coefficient 0.71 (standard error 0.35), differing significantly from 0 at a 0.05 level. The positive coefficient indicates that workplaces with higher response rates to the employee questionnaire tend to have higher levels of job satisfaction, which may be plausible.

The concern here is that if nonresponse is informative, with the probability of response increasing as y_{ij} increases, will this bias the estimation of the regression model of interest and in what way?

4 Models for nonresponse

In this section, we begin by formalising the notion of informative nonresponse in our context and then consider two models of the nonresponse mechanism which might account for association between $\bar{y}_{ri}$ and p_i, conditional on $\mathbf{x}_i$, as observed in the application in the previous section.

We define models for nonresponse in terms of the relationship between the random vectors $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})$ and $\mathbf{R}_i = (R_{i1}, \ldots, R_{im_i})$ for $i = 1, \ldots, n$. Note that noninformative sampling is assumed so that the distribution of y_{ij} is the same whether ij is in the sample or not. Moreover, we assume that the pairs $(\mathbf{y}_i, \mathbf{R}_i)$ are independently distributed for different clusters $i = 1, \ldots, n$.

4.1 Noninformative nonresponse

We say that the nonresponse is noninformative if $\mathbf{y}_i$ and $\mathbf{R}_i$ are conditionally independent given $\mathbf{x}_i$. It follows from (1) that under this condition we have

$$E_m(y_{ij} \mid \mathbf{R}_i) = E_m(\bar{y}_{ri} \mid \mathbf{R}_i) = \mathbf{x}_i'\beta, \tag{2}$$

and that the OLS estimator

$$\hat{\beta}_{OLS} = \left(\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{'} \right)^{-1} \sum_{i=1}^{n} \mathbf{x}_i \bar{y}_{ri}, \tag{3}$$

is unbiased for β.

4.2 Normal selection model

As our first possible source of informative nonresponse, we formulate a two-level selection model following Heckman [4]. See also Little and Rubin [8, sect. 15.4]. This model is related to the parametric cluster-specific nonignorable nonresponse model in Yuan and Little [12]. To model the response outcome R_{ij}, we introduce a variable u_{ij} so that $R_{ij} = 1$ if $u_{ij} > 0$ and $R_{ij} = 0$, otherwise. We then specify a model for both y_{ij} and u_{ij} as:

$$y_{ij} = \mathbf{x}_i'\beta + \epsilon_{ij}, \tag{4}$$

$$u_{ij} = \mathbf{z}_i'\gamma + \delta_{ij}, \tag{5}$$

where $\mathbf{z}_i = (1, z_{i1}, \ldots, z_{il})'$ is a vector of covariates which are assumed to influence nonresponse and may include covariates in $\mathbf{x}_i$. It is assumed that the disturbance terms in (4) and (5) obey

$$\begin{pmatrix} \epsilon_{ij} \\ \delta_{ij} \end{pmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} \begin{bmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon\delta} \\ \sigma_{\epsilon\delta} & \sigma_\delta^2 \end{bmatrix} \right), \tag{6}$$

where the distribution in (6) is taken to be conditional on $\mathbf{x}_i$ and $\mathbf{z}_i$. Nonresponse is noninformative if $\mathbf{y}_i$ and $\mathbf{R}_i$ are conditionally independent given $\mathbf{x}_i$ and $\mathbf{z}_i$, which arises if $\sigma_{\epsilon\delta} = 0$, using (4), (5) and (6). In this case, and assuming (1) holds, the OLS estimator is unbiased for β, as in Sect. 4.1. In general, however, this will not be the case. Thus, we may write

$$\begin{aligned} E(y_{ij}|R_{ij} = 1) &= \mathbf{x}_i'\beta + E(\epsilon_{ij}|u_{ij} > 0) \\ &= \mathbf{x}_i'\beta + E(\epsilon_{ij}|\mathbf{z}_i'\gamma + \delta_{ij} > 0) \\ &= \mathbf{x}_i'\beta + E(\epsilon_{ij}|\delta_{ij} > -\mathbf{z}_i'\gamma). \end{aligned} \tag{7}$$

From (6) we can write, $\epsilon_{ij} = \sigma_{\epsilon\delta}\sigma_\delta^{-2}\delta_{ij} + \xi_{ij}$, where ξ_{ij} is independent of δ_{ij} so that

$$E(y_{ij}|R_{ij} = 1) = \mathbf{x}_i'\beta + E(\sigma_{\epsilon\delta}\sigma_\delta^{-2}\delta_{ij} + \xi_{ij}|\delta_{ij} > -\mathbf{z}_i'\gamma). \tag{8}$$

Following Heckman [4], we have

$$E(y_{ij}|R_{ij} = 1) = \mathbf{x}_i'\beta + c\lambda(\mathbf{z}_i'\psi), \tag{9}$$

where $\psi = \sigma_\delta^{-1}\gamma$, $c = \sigma_{\epsilon\delta}\sigma_\delta^{-1}$, $\lambda(\mathbf{z}_i'\psi) = \phi(\mathbf{z}_i'\psi)/\Phi(\mathbf{z}_i'\psi)$ is the inverse Mills ratio, $\phi(.)$ is the probability density function of the standard normal distribution and $\Phi(.)$ is the cumulative distribution function of this distribution.

We might view the term $\lambda(\mathbf{z}_i'\psi)$ in (9) as a missing auxiliary variable which could induce bias in the OLS estimator $\hat{\beta}_{OLS}$ and could be 'proxied' by p_i if p_i is included as an extra explanatory variable when fitting model (1) to respondent data. To explore this idea, note that we might express p_i approximately as $p_i \approx E(R_{ij})$ and we have

$$E(R_{ij}) = Pr(u_{ij} > 0) = Pr(\delta_{ij} > -\mathbf{z}_i'\gamma) = \Phi(\mathbf{z}_i'\psi). \tag{10}$$

Hence, we might approximate the term $\lambda(\mathbf{z}_i'\psi)$ in (9) by $\lambda(\Phi^{-1}(p_i))$, i.e. a transformation of p_i, and this could provide the basis of one explanation for the significance of p_i when it is added into the regression in Sect. 3.

4.3 A simple informative nonresponse model

We next consider a simple informative nonresponse model, which may be considered as a special case of the previous selection model or as a simple version of a pattern-mixture model considered in Little and Rubin [8, Example 15.10]. We suppose that

$$E(y_{ij}|R_{ij} = 1) - E(y_{ij}|R_{ij} = 0) = \delta, \tag{11}$$

for some constant δ, that is that respondents and nonrespondents differ in their expected value of y_{ij} by a fixed amount, given the values of the covariates. When nonresponse is noninformative we have $\delta = 0$, but nonresponse is informative in general.

It follows that this is a special case of the selection model by noting that, under this model, we may show that $E(y_{ij}|R_{ij} = 0) = \mathbf{x}_i'\beta - c\lambda(-\mathbf{z}_i'\psi)$ and so, from (8), $E(y_{ij}|R_{ij} = 1) - E(y_{ij}|R_{ij} = 0) = c\lambda(\mathbf{z}_i'\psi) - c\lambda(-\mathbf{z}_i'\psi)$, which reduces to a constant if $\mathbf{z}_i$ is fixed.

To explore the consequences of this model, define the nonrespondent mean of y_{ij} by $\bar{y}_{nri} = \sum_{j=1}^{m_i}(1 - R_{ij})y_{ij}/(m_i - r_i)$ and note that we may write the sample mean of y_{ij} as $\bar{y}_i = \sum_{j=1}^{m_i} y_{ij}/m_i = p_i\bar{y}_{ri} + (1 - p_i)\bar{y}_{nri}$. It follows, using (11), that we may write

$$\begin{aligned} E(\bar{y}_i \mid \mathbf{R}_i) &= p_i E(\bar{y}_{ri} \mid \mathbf{R}_i) + (1 - p_i)E[\bar{y}_{nri} \mid \mathbf{R}_i] \\ &= E(\bar{y}_{ri} \mid \mathbf{R}_i) - (1 - p_i)\delta \end{aligned} \tag{12}$$

and hence, using (1), that

$$E(\bar{y}_{ri} \mid \mathbf{R}_i) = \mathbf{x}_i'\beta + (1 - p_i)\delta. \tag{13}$$

Analogous to the way we viewed $\lambda(\mathbf{z}_i'\psi)$ in (9), we may view $(1 - p_i)$ in (13) as a source of bias in the estimation of β arising from the simple informative nonresponse model.

5 Estimation of β and testing of informativeness

It follows from the discussion in the previous section that the OLS estimator $\hat{\beta}_{OLS}$ will, in general, be biased under either of the two informative nonresponse models considered there unless $\sigma_{\epsilon\delta} = 0$ or $\delta = 0$. We now consider how we might seek to remove this bias by constructing estimators based upon each of these models. In both cases, our approach is to include an additional covariate to control for the selection effect.

One approach is to use the simple informative nonresponse model, where it follows from (13) that we just need to include $1 - p_i$ as an additional covariate and regress $\bar{y}_{ri}$ on $\mathbf{x}_i$ and $1 - p_i$. We refer to the resulting estimator of β as the *simple informative* estimator. This is similar to a method discussed in Yuan and Little [13] in which an estimated cluster-level response rate is included as a covariate in a model used to adjust for nonresponse. We next turn to the normal selection model and define a *two-step* estimator of β (c.f. Heckman [4]) as follows

Step 1. Noting that $Pr(R_{ij} = 1) = \Phi(\mathbf{z}_i'\psi)$ from (10), obtain an estimator $\hat{\psi}$ of ψ by probit regression of R_{ij} on $\mathbf{z}_i$.

Step 2. Calculate the estimated inverse Mills ratio,

$$\lambda(\mathbf{z}_i'\hat{\psi}) = \phi(\mathbf{z}_i'\hat{\psi})/\Phi(\mathbf{z}_i'\hat{\psi}), \tag{14}$$

plug this into (9) and regress y_{ij} on $\mathbf{x}_i$ and this estimated inverse Mills ratio to obtain estimators of β and c.

A simpler version of this estimator, which is not dependent on the choice of $\mathbf{z}_i$, is obtained by using the large m_i approximation (based on (10))

$$p_i \approx E(R_{ij}) = \Phi(\mathbf{z}_i'\psi), \tag{15}$$

and replacing $\lambda(\mathbf{z}_i'\hat{\psi})$ by $\lambda(\Phi^{-1}(p_i))$ in the two-step approach. We refer to this as the p_i-*approximate two step* estimator.

An *approximate maximum likelihood* (ML) estimator is obtained as follows. Under the working assumption that the observations are independent, the log likelihood for the observed data and the model in (4)–(6) is

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i}(1-R_{ij})log[1-Pr(u_{ij}>0)]+R_{ij}log[Pr(u_{ij})>0]+R_{ij}log[f(y_{ij}|u_{ij}>0)],$$

where $f(.\mid .)$ denotes the relevant conditional probability distribution. This log likelihood may be maximised numerically using the facts that $f(y_{ij}|u_{ij}>0)=Pr(u_{ij}>0|y_{ij})f(y_{ij})/Pr(u_{ij}>0)$, $y_{ij}\sim N(\mathbf{x}_i'\beta,\sigma_\epsilon^2)$ and $u_{ij}|y_{ij}\sim N[\mathbf{z}_i'\gamma+\sigma_{\epsilon\delta}\sigma_\epsilon^{-2}(y_{ij}-\mathbf{x}_i'\beta),\sigma_\delta^2-\sigma_{\epsilon\delta}^2\sigma_\epsilon^{-2}]$. Evaluation of the log likelihood requires evaluating $Pr(u_{ij}>0)$ for all cases (i,j) and evaluating the probability density function of $N(\mathbf{x}_i'\beta,\sigma_\epsilon^2)$ and $Pr(u_{ij}>0|y_{ij})$ for all cases with $R_{ij}=1$.

We shall estimate the standard errors of each of the above point estimators of β using a bootstrap approach with 1000 replicates in which the sampled clusters $i=1,...,n$ are resampled by simple random sampling with replacement.

The above estimation methods imply approaches to testing the informativeness of the nonresponse. Letting $\hat{\delta}$ denote the estimator of δ in (13) using the simple informative estimator, the 't-statistic' obtained by dividing $\hat{\delta}$ by its bootstrap standard error will have a standard normal distribution under noninformativeness and may be used to test this assumption. Note that the validity of this null distribution does not depend on the model assumption in (11) but only on the general assumption in (2). Similarly, it is possible to construct a test from the t-statistic formed by dividing the two-step estimator of c in (9) by its bootstrap standard error. The validity of the null distribution of this test depends not only on the assumption in (2), but also on the assumption that $\mathbf{z}_i$ does not feature on the right hand side of (4) (other than as a component of $\mathbf{x}_i'\beta$).

6 Simulation studies

We now present two studies of the properties of the estimators introduced in the previous section, one in which values are generated from the normal selection model and one with values generated from the simple informative nonresponse model.

6.1 Study 1 based on normal selection model

Here we vary the correlation $\rho=\sigma_{\epsilon\delta}/(\sigma_\epsilon\sigma_\delta)$, which governs the degree of informative missingness in the model via (6) and fix the scale parameters at $\sigma_\epsilon=3$ and $\sigma_\delta=1$. We set $N=1,000$ and take various combinations of the sample sizes n and m_i, assuming the latter are constant with $m_i=m$. We set $\mathbf{x}_i=(1,x_{1i})'$ so that $k=1$ and similarly $\mathbf{z}_i=(1,z_{1i})'$. The following simulation steps are repeated 10,000 times.

Step 1 Generate ϵ_{ij} and δ_{ij} from a bivariate normal distribution following (6).

Step 2 Generate x_{1i} and z_{1i} from a bivariate normal distribution with means zero, variances one and correlation 0.5.

Step 3 Determine y_{ij} from (4) using values of ϵ_{ij} from step 1 and x_{1i} from step 2 and $\beta_0=0,\beta_1=1$.

Step 4 Similarly, determine u_{ij} from (5) and values of δ_{ij} from Step 1 and z_{1i} from Step 2 and $\gamma_0=0,\gamma_1=1$. These values imply an expected overall response rate of 50 %.

Step 5 Compute values of the alternative estimators defined earlier.

6.2 Study 2 based on simple informative nonresponse model

We again set $N = 1000$ and $\mathbf{x}_i = (1, x_{1i})'$ so that $k = 1$. Now we vary the quantity δ in (11) which governs the degree of informativeness. We repeat the following steps, again 10,000 times.

Step 1 Generate x_{1i} and z_{1i} as in Study 1, and then set

$$\pi_i = \frac{4exp(z_{1i})}{1 + 4exp(z_{1i})}. \tag{16}$$

Step 2 Generate R_{ij} from $P(R_{ij} = 1) = \pi_i$.
Step 3 Generate y_{ij} from

$$y_{ij} = \begin{cases} (\mathbf{x}_i'\beta + \epsilon_{1i} + \epsilon_{2ij}) + (1 - p_i)\delta, & \text{if } R_{ij} = 1 \\ (\mathbf{x}_i'\beta + \epsilon_{1i} + \epsilon_{2ij}) - p_i\delta, & \text{if } R_{ij} = 0, \end{cases} \tag{17}$$

where ϵ_{1i} and ϵ_{2i} are generated from $\epsilon_{1i} \sim N(0, 1)$ and $\epsilon_{2ij} \sim N(0, 9)$, $\beta_0 = 0$, $\beta_1 = 1$, and $\delta = 1, 2$ or 4.
Step 4 Compute values of the alternative estimators defined earlier.

The model in (17) may be viewed as a pattern mixture model in the sense that it generates R before generating $y|R$ in order to determine (R, y) and it ensures that (11) holds. We also undertook a similar study where p_i in (17) is replaced by π_i and obtained very similar results.

6.3 Results of study 1

We present results for two choices of n and m: $n = 20$, $m = 25$ in Table 1 and $n = 100$, $m = 5$ in Table 2. It seems important to consider different values of m since the 'information' in p_i may be expected to decline as m decreases and we would expect this to affect the relative performances of the estimators. The first choice of m corresponds to the maximum value of m_i in the motivating application. Results for the estimators with least absolute bias, variance and mean squared error (MSE) are presented in bold. With 10,000 runs, the simulation error is small. The simulation standard errors for the means of the estimates of β_0 and β_1 in Tables 1 and 2 range from 0.002 to 0.004.

In Table 1 we see that the OLS estimators of β_0 and β_1, as expected, display bias when $\rho \neq 0$. All the other estimators of these parameters tend to have less bias when $\rho \neq 0$ and for all these other estimators (unlike the OLS estimator) the reduced bias tends to be dominated by the standard error when considering MSE. The OLS estimator does, however, have lower variance than the other estimators, at least for smaller values of ρ and it is the preferred estimator when $\rho = 0$. The OLS estimator also has MSE comparable to that of the other estimators when $\rho = 0.2$. For larger values of ρ, however, the OLS estimator tends to have a much higher MSE. The other estimators have broadly similar performance across the range of values of ρ here. Focussing on β_1 which is of primary interest, the approximate ML estimator tends to perform slightly better in terms of MSE when $\rho \neq 0$, followed by the two step estimator.

Table 2 shows some similarity of results to Table 1 but some differences. When $\rho \neq 0$ the OLS estimator is again biased and all the other estimators reduce this bias. However, it is no longer the case that the bias is dominated by the variance for the other estimators. The bias for the simple informative and p_i-approximate two step estimators tends to be worse than that of the two-step and approximate ML estimators and the bias of the former estimators is particularly pronounced when $\rho = 0.5$ or 0.8. The poor performance of the

Table 1 Study 1. Simulation mean, variance and mean square error of estimators of β_0 and β_1 for n = 20, m = 25 and alternative values of ρ under normal selection model

Estimator	Mean		Variance		MSE	
	β_0	β_1	β_0	β_1	β_0	β_1
$\rho = 0$						
1. OLS	**0.009**	**0.973**	**0.047**	**0.048**	**0.047**	**0.049**
2. Simple informative	0.038	0.965	0.173	0.061	0.174	0.062
3. Two-step	0.023	0.968	0.185	0.062	0.186	0.063
4. p_i-approximate two step	0.037	0.965	0.163	0.060	0.164	0.061
5. Approximate ML	0.024	0.968	0.180	0.061	0.180	0.062
$\rho = 0.2$						
1. OLS	0.391	0.897	**0.050**	**0.060**	0.203	0.071
2. Simple informative	**0.001**	**1.007**	0.175	0.071	0.175	0.071
3. Two-step	−0.039	1.019	0.189	0.070	0.191	0.071
4. p_i-approximate two step	0.006	1.008	0.165	0.070	**0.165**	0.070
5. Approximate ML	−0.032	1.017	0.183	0.069	0.184	**0.069**
$\rho = 0.5$						
1. OLS	0.982	0.693	**0.066**	0.069	1.031	0.163
2. Simple informative	0.053	0.963	0.172	0.062	0.175	0.063
3. Two-step	−0.065	0.991	0.183	0.058	0.188	0.058
4. p_i-approximate two step	0.067	0.965	0.163	0.061	0.167	0.062
5. Approximate ML	**−0.035**	0.985	0.163	**0.057**	**0.164**	**0.057**
$\rho = 0.8$						
1. OLS	1.557	0.582	**0.090**	0.092	2.515	0.267
2. Simple informative	0.163	0.984	0.150	0.052	0.177	0.052
3. Two-step	**−0.014**	1.019	0.158	0.042	0.158	0.043
4. p_i-approximate two step	0.178	**0.988**	0.138	0.050	0.169	0.050
5. Approximate ML	0.009	1.023	0.100	**0.039**	**0.101**	**0.040**

The true values are $\beta_0 = 0, \beta_1 = 1$
Figures in bold face identify the estimator with least absolute bias, variance or MSE, depending on the column

simple informative and p_i-approximate two step estimators when m is as small as 5 may be attributed to the dependence of these estimators on p_i, which will be very noisy for such a small value of m. The two-step and approximate ML estimators tend to perform better than the other estimators for larger values of ρ.

6.4 Results of study 2

We consider here the results only for $n = 20, m = 25$, as in Table 1. Table 3 shows that, as in Study 1, the OLS estimators of both β_0 and β_1 are biased for each non-zero value of δ considered here, with bias increasing as δ increases.

Using the simple informative estimator removes the bias but with inflated variance. The p_i-approximate two-step estimator behaves similarly to the simple informative approach. The two-step estimator performs a little worse than these estimators. The approximate ML estimator has more bias than the other estimators, excluding OLS, but the bias is dominated by the variance for these sample sizes and, overall, this estimator performs as well as the simple informative estimator in terms of MSE.

Table 2 Study 1. Simulation mean, variance and mean square error of estimators for n=100, m = 5 and alternative values of ρ under normal selection model.

Estimator	Mean		Variance		MSE	
	β_0	β_1	β_0	β_1	β_0	β_1
$\rho = 0$						
OLS	**−0.063**	0.981	**0.035**	**0.032**	**0.039**	**0.032**
Simple informative	−0.123	**0.998**	0.097	0.037	0.112	0.037
Two-step	−0.108	0.993	0.166	0.040	0.178	0.040
p_i-approximate two step	−0.118	0.997	0.095	0.037	0.109	0.037
Approximate ML	−0.111	0.994	0.188	0.042	0.200	0.042
$\rho = 0.2$						
OLS	0.544	0.883	**0.035**	0.175	0.331	0.049
Simple informative	0.330	0.943	0.095	**0.039**	0.204	**0.042**
Two-step	0.144	0.987	0.171	0.044	**0.191**	0.044
p_i-approximate two step	0.332	0.943	0.094	**0.039**	0.205	**0.042**
Approximate ML	**0.143**	**0.987**	0.171	0.044	**0.191**	0.044
$\rho = 0.5$						
OLS	1.058	0.486	**0.034**	**0.032**	1.153	0.296
Simple informative	0.395	0.685	0.084	0.033	0.240	0.133
Two-step	**−0.155**	0.809	0.143	0.035	0.167	0.071
p_i-approximate two step	0.390	0.689	0.083	0.033	0.236	0.130
Approximate ML	−0.170	**0.818**	0.117	0.034	**0.146**	**0.067**
$\rho = 0.8$						
OLS	1.706	0.442	**0.035**	0.034	2.947	0.345
Simple informative	0.759	0.711	0.077	0.033	0.653	0.117
Two-step	−0.243	**0.969**	0.141	0.032	0.201	**0.033**
p_i-approximate two step	0.751	0.715	0.077	0.033	0.640	0.114
Approximate ML	**0.050**	0.903	0.066	**0.027**	**0.068**	0.037

The true values are $\beta_0 = 0, \beta_1 = 1$
Figures in bold face identify the estimator with least absolute bias, variance or MSE, depending on the column

7 Application to workplace employment relations survey

We now apply the new methods devised in Sect. 5 to the analysis using WERS survey data introduced in Sect. 3.

The outcome variable y_{ij} is a job satisfaction measure, based on responses of employees to the question " How satisfied are you with the following aspects of your job?" for the following eight aspects: achievement you get from your work; the scope for using your own initiative; the amount of influence you have over your job; the training you receive; the amount of pay you receive; your job security; the work itself; the amount of involvement you have in decision-making at this workplace. Responses on each of these eight aspects were recorded on a 5-point Likert scale from very satisfied to very dissatisfied, coded from −2 to 2, and then summed to give an overall measure varying between −16 and 16.

As our primary independent variable of interest x_{1i}, we considered a number of workplace innovation variables, following Bryson et al. [1]. We decided to use a labour innovations variable here, since this was found by Bryson et al. [1] to be most strongly related to job

Table 3 Study 2. Simulation mean, variance and mean square error of estimators for n = 20, m = 25 and alternative value of δ under simple NMAR model

Estimator	Mean		Variance		MSE	
	β_0	β_1	β_0	β_1	β_0	β_1
$\delta = 1$						
1. OLS	0.24	0.92	**0.08**	**0.09**	**0.14**	**0.10**
2. Simple informative	**0.00**	**1.00**	0.28	0.11	0.28	0.11
3. Two-step	−0.02	**1.00**	0.43	0.12	0.43	0.12
4. p_i approximate two-step	−0.01	**1.00**	0.31	0.11	0.31	0.11
5. Approximate ML	0.03	0.99	0.31	0.11	0.31	0.11
$\delta = 2$						
1. OLS	0.48	0.84	**0.09**	**0.10**	0.31	0.12
2. Simple informative	**0.00**	**1.00**	0.28	0.11	**0.28**	**0.11**
3. Two-step	−0.05	**1.00**	0.43	0.12	0.43	0.12
4. p_i approximate two-step	−0.03	**1.00**	0.32	0.11	0.32	**0.11**
5. Approximate ML	0.07	0.97	0.28	0.11	**0.28**	**0.11**
$\delta = 4$						
1. OLS	0.96	0.69	**0.10**	**0.11**	1.02	0.21
2. Simple informative	**0.00**	**1.00**	0.28	**0.11**	0.28	**0.11**
3. Two-step	−0.10	**1.00**	0.44	0.13	0.45	0.13
4. p_i approximate two-step	−0.06	**1.00**	0.32	**0.11**	0.32	**0.11**
5. Approximate ML	0.20	0.93	0.20	**0.11**	**0.24**	**0.11**

The true values are $\beta_0 = 0, \beta_1 = 1$
Figures in bold face identify the estimator with least absolute bias, variance or MSE, depending on the column

satisfaction. This variable is obtained from responses from the manager at each workplace to the question "Over the past two years has management here introduced any of the changes listed on this card?", where the four changes listed are: changes in working time arrangements; changes in the organisation of work; changes in work techniques or procedures; introduction of initiatives to involve employees. The labour innovations variable x_{1i} is then defined as the number of positive responses to these questions and thus ranges from 0 to 4.

In addition to considering regression models with just x_{1i} and an intercept as covariates, we also considered models with a much richer vector $\mathbf{x}_i$, including also a series of control variables, following Bryson et al. [1]. These consist of both workplace level variables, such as industry, union membership, workplace employment and the local unemployment rate, obtained from the manager or from official statistics, and workplace means of individual level variables, such as gender, age, academic qualifications, occupation and disability, obtained from employees.

We considered four of the estimators described in Sect. 5: the OLS estimator $\hat{\beta}_{OLS}$ and the simple informative, two-step and p_i -approximate two step estimators. In the case of the two-step estimator, it is necessary to specify the variable z_i and for this we used the (population) number of employees in the workplace, as a variable with the potential to be related to non-response. We did attempt to use the approximate ML estimator but do not report on results since it frequently failed to converge. The standard error of each estimator of β was estimated from a bootstrap approach which resampled the clusters with replacement 1000 times.

Table 4 Estimates of coefficients in regression model for WERS data with job satisfaction as dependent variable and labour innovation as independent variable

Standard errors in parentheses

Estimator	Intercept, β_0	Innovation, β_1
OLS	4.481 (0.135)	−0.286 (0.057)
Simple informative	4.694 (0.185)	−0.286 (0.035)
Two-step	−0.955 (1.098)	−0.234 (0.059)
p_i-approximate two-step	4.688 (0.180)	−0.286 (0.057)

Table 4 presents estimates of the regression model of job satisfaction on the labour innovation variable. As noted in Sect. 3, higher levels of innovation were associated with lower levels of job satisfaction. The covariate $(1 - p_i)$ was found to be significant (at the 0.05 level) in the simple informative estimator, both when the only covariate consists of labour innovation and when control variables were included in the covariate vector too. However, the estimated coefficient of labour innovation is unchanged by the inclusion of the covariate $(1 - p_i)$ although the standard error is slightly increased, as might be expected.

The coefficient of the estimated inverse mills ratio variable in the two-step estimator was also significant at 0.05 level, although it was not significant when control variables were also included in the model. The estimates of β_0 and β_1 are different for this estimator, especially the intercept which has a greatly increased standard error. The correlation between p_i and z_i here is weak (only -0.05) and the fact that the coefficient of the estimated inverse mills ratio becomes insignificant once control variables are included suggests that the two-step method may be capturing something other than the effect of nonresponse and hence produces somewhat different results. The weak correlation between p_i and z_i may also be a reason behind the non-convergence of the ML estimator.

The p_i—approximate two step estimator makes no use of the variable z_i and performs similarly to the simple informative estimator, although with a larger standard error for β_1. The coefficient of $\lambda(\Phi^{-1}(p_i))$ differs significantly from zero, whether or not the additional control variables are included in the model.

8 Conclusion

In this paper we have studied how the inclusion of a cluster-level non-response rate $1 - p_i$ as a covariate in a cluster-level regression may be used to detect certain kinds of informative nonresponse. In our application, the coefficient of $1 - p_i$ was found to differ significantly from zero. Such a finding might arise because $1 - p_i$ is acting as a proxy for some omitted cluster-level variable, but in our study we found the coefficient still differed significantly from zero, even after including many control variables which might be expected to influence the outcome. We also showed by theory and through our simulation study that the inclusion of $1 - p_i$ as a covariate can reduce the bias of the OLS estimator of the coefficients of interest under certain informative nonresponse models. Whilst it can increase standard errors, we still showed in the simulation study how the mean squared error can be reduced.

For comparison, we also considered a form of normal selection model, as in Heckman [4]. This was used both as a basis for a simulation study to assess the use of $1 - p_i$ as a covariate and also to construct alternative estimators. Inevitably, it was possible to improve on the use of $1 - p_i$ as a covariate, if the model was correct and the covariate vector $\mathbf{z}_i$ used to model nonresponse in this model is known. However, the gain was modest unless the correlation ρ is large and this approach is more complex given the requirement to specify the vector $\mathbf{z}_i$.

We also showed how we could approximate the selection model approach and avoid the need to specify $\mathbf{z}_i$ by incorporating a non-linear transformation $\lambda(\Phi^{-1}(p_i))$ of p_i as a covariate. We found this approach performed similarly to including $1 - p_i$ as a covariate, both in our simulation study and in our application, presumably because the transformation is fairly linear across the range of values of p_i that we used. The simulation study showed a very slight advantage of the non-linear transformation.

We have not considered the impact of weighting in this paper. The procedures we have studied should extend naturally to take account of weighting for unequal sample inclusion probabilities in the definition of p_i, but this does not immediately appear to raise new conceptual issues. Weighting for nonresponse is different. Conventionally, such weighting might be used in surveys when auxiliary variables are available and when a noninformative nonresponse assumption might be plausible when these variables are conditioned upon. However, when the objective is to fit a regression model, it is natural to seek to control for nonresponse by including such auxiliary variables as covariates rather than by employing weights and this is what we have done here. The informative nature of the nonresponse in our application was not controlled for by using auxiliary variables as additional covariates and we do not anticipate that it would be by the use of weights constructed from the same auxiliary variables, but this topic might merit further investigation alongside the role of sampling weights.

Acknowledgements We acknowledge provision of the WERS 2004 Management Questionnaire (MQ) data file by the UK Data Archive and thank Alex Bryson and John Forth for advice on these data and the application.

References

1. Bryson, A., Dale-Olsen, H., Barth, E.: How Does Innovation Affect Worker Well-Being? Centre for Economic Performance, London School of Economics and Political Science, London (2009)
2. Chambers, R.L., Skinner, C.J.: Analysis of Survey Data. Wiley, Chichester (2003)
3. Haziza, D., Rao, J.: Variance estimation in two-stage sampling under imputation for missing survey data. J. Stat. Theory Pract. **4**, 827–848 (2010)
4. Heckman, J.J.: The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. Ann. Econ. Soc. Measur. **5**(4), 475–492 (1976)
5. Kersley, B., Alpin, C., Forth, J., Bryson, A., Bewley, H., Dix, G., Oxenbridge, S.: Inside the Workplace: Findings from the 2004 Workplace Employment Relations Survey. Routledge, Oxfordshire, Abingdon (2006)
6. Kim, J.K., Kwon, Y., Paik, M.C.: Calibrated propensity score method for survey nonresponse in cluster sampling. Biometrika **103**, 461–473 (2016)
7. Lago, L., Clark, R.: Imputation of household survey data using linear mixed models. Aust. N. Z. J. Stat. **57**, 169–187 (2015)
8. Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data. Wiley, New York (2002)
9. O'Malley, A.J., Zaslavsky, A.M.: Domain-Level covariance analysis for multilevel survey data with structured nonresponse. J. Am. Stat. Assoc. **103**, 1405–1418 (2008)
10. Shao, J.: Handling survey nonresponse in cluster sampling. Surv. Methodol. **33**, 81–85 (1970)
11. Skinner, C.J.: D′Arrigo: inverse probability weighting for clustered nonresponse. Biometrika **98**, 953–966 (2011)
12. Yuan, Y., Little, R.J.A.: Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response. J. R. Stat. Soc. Ser. C (Appl. Stat.) **56**, 79–97 (2007)
13. Yuan, Y., Little, R.J.A.: Model-based inference for two-stage cluster samples subject to nonignorable item nonresponse. J. Off. Stat. **24**, 193–211 (2008)