Rezaee *et al.*
Swiss Journal of Economics and Statistics (2022) 158:8
https://doi.org/10.1186/s41937-022-00089-1

Swiss Journal of
Economics and Statistics

ORIGINAL ARTICLE

Sample selection bias with multiple dependent selection rules: an application to survey data analysis with multilevel nonresponse

Alireza Rezaee, Mojtaba Ganjali* and Ehsan Bahrami Samani

Abstract

The microdata of surveys are valuable resources for analyzing and modeling relationships between variables of interest. These microdata are often incomplete because of nonresponses in surveys and, if not considered, may lead to model misspecification and biased results. Nonresponse variable is usually assumed as a binary variable, and it is used to construct a sample selection model in many researches. However, this variable is a multilevel variable related to its reasons of occurring. Missing mechanism may differ among the levels of nonresponse, and merging the levels of nonresponse may cause bias in the results of the analysis. In this paper, a method is proposed for analyzing survey data with respect to reasons for the nonresponse based on sample selection model. Each nonresponse level is considered as a selection rule, and classical Heckman model is extended. Simulation studies and an analysis of a real data set from an establishment survey are presented to demonstrate the performance and practical usefulness of the proposed method.

Keywords: Establishment survey, Heckman model, Multivariate sample selection model, Nonresponse mechanism, Probit model, Truncated normal distribution

1 Introduction

Modeling relationships between variables based on survey microdata is an essential part of many researches and analyses. For example, in survey methodology, determining a proper approach for some problems, such as appropriate strategies for following up nonresponse units in the phase of data collection, assessment of measurement errors of main variables among individuals that responded and imputation of nonresponse, is usually based on the modeling. Another example in economics is to model productivity or efficiency in a sector in the form of secondary analysis of survey data or to examine the relationship between turnover and value added of an establishment with some factors related to production such as the number of employees based on microdata of an establishment survey.

Most surveys suffer from nonresponse and their microdata are often incomplete. Nonresponse can increase errors of estimates or lead to model misspecification and biased results, especially in the case of nonignorable nonresponse. Heckman (1976, 1979) presented a method to adjust bias due to the nonresponse in modeling of a dependent variable. He considered a model for nonresponse called sample selection model and presented the estimates of the parameters and the variances of the estimators by assuming nonresponse variable as binary and normal distribution for the errors components of two models (nonresponse and variable of interest models). Hanoch (1976) extended the Heckman approach for multivariate dependent variables with one equation for nonresponse mechanism and investigated main factors

*Correspondence: m_ganjali43@yahoo.com
Department of Statistics, Shahid Beheshti University, Tehran, Iran

on labor force. Catsiapis and Robinson (1978, 1982) developed the Heckman model by two and then multi-equations for nonresponse mechanisms and obtained estimators for model parameters with independent assumptions between the random effects in the equations of selection mechanism. Since then, in recent years, some developments have been performed on Heckman model. Jolani (2014) worked on longitudinal data in the presence of nonresponse by presenting an extension of Heckman model. He modeled dependent variable with some explanatory variables at time t, and for each time before t, considered a model as selection model. He obtained the estimates of the parameters by assuming nonresponse variable as binary and multivariate normal distribution for error components in the models. Kim and Kim (2016) presented a method to analyze data with multivariate sample selection model. They assumed elliptically contoured (EC) distribution for the errors in the models to obtain robustness against departures from normality.

However, nonresponse can be caused by different reasons, and therefore it is in fact a multilevel variable. Merging of the levels may lead to model misspecifications and biased results, especially in cases where the mechanism of nonresponse is not the same at different levels of nonresponse. In other words, different covariates may be related to different reasons of nonresponse or the effects of covariates are of different strengths, or go in opposite directions. So, it makes sense to consider a different selection model for each level of nonresponse in such cases.

Most of the researches about analysis of survey data are based on using only one binary variable for nonresponse. Also, there are a few works on nonresponse in establishment surveys in recent years. Earp et al. (2014, 2018), Kirchner and Signorino (2018), Phipps and Toth (2012), Seiler (2010) and Rezaee et al. (2021) used logistic regression, classification tree and support vector machine methods to investigate nonresponse in establishment surveys. Paiva and Reiter (2017) provided a way to follow nonresponse samples in an establishment survey using a mixture pattern model and the assumption of a nonrandom nonresponse mechanism. Refusal and noncontact are two levels of nonresponse variable and were studied in some of the researches about household surveys. Heerwegh et al. (2007) examined the effect of nonresponse error due to refusal and noncontact in a household survey and concluded that the error due to noncontact nonresponse is 2.56 times greater than the error due to refusal. Durrant and Steele (2009) examined the factors influencing the nonresponse by distinguishing refusal from noncontact for a set of UK household surveys using a multivariate logistic regression model. Steele and Durrant (2011) examined alternative approaches to multilevel modeling of survey noncontact and refusal. They reviewed multinomial and sequential models and compare them with a sample selection model that allows for residual correlation between a sample unit's noncontact and refusal propensities. Vassallo et al. (2015) also examined interviewer's experience effects on nonresponse in a panel survey in the case of multilevel nonresponse.

In this paper, we provided a method for analyzing incomplete survey data with considering nonresponse as a dependent multilevel variable. We extended the classical Heckman model via increasing the number of selection models, caused by the number of nonresponse reasons, considering the dependency between nonresponse levels, then we evaluated the performance of the proposed method using a simulation study and implemented it on an establishment survey with two reasons, refusal and noncontact for nonresponse. We compared the results of the proposed method with those of the univariate selection model and investigated the influence of nonrandom nonresponse by a sensitivity study.

This paper is organized as follows. In Sect. 2, Heckman model is reviewed, then in Sect. 3, sample selection model with multiple selection rules is presented and discussed. In Sect. 4, simulation studies are given, and in Sect. 5, the proposed method is implemented on an establishment survey microdata and the results are compared with those of using univariate selection model. Also, the influence of nonrandom nonresponse is investigated using likelihood displacement. In Sect. 6, conclusion and discussion are given.

2 Univariate selection model

Heckman (1976, 1979) proposed a method for bias correction due to nonresponse samples in an ordinary regression model. He wanted to estimate the parameters in the model

$$y_i = \mathbf{x_i}\beta + e_i, i = 1, 2, \ldots, n \tag{1}$$

in the presence of nonresponse on some y_is. He considered the model

$$y_i^* = \mathbf{w_i}\alpha + u_i, i = 1, 2, \ldots, n \tag{2}$$

for nonresponse as sample selection model, where y_i^* is a latent variable such that if $y_i^* < 0$ then nonresponse occurs for y_i and if $y_i^* \geq 0$, then y_i is observed. He assumed bivariate normal distribution for the errors (e_i, u_i) with parameters $(0, 0, \sigma_1^2, \sigma_2^2, \rho)$ and found estimators of parameters in models (1) and (2). In order to calculate sample selection bias, Heckman first obtained $E[y_i|\mathbf{x_i}, y_i^* \geq 0] = \mathbf{x_i}\beta + \frac{\sigma_{12}}{\sigma_2}\lambda_i$ where $\lambda_i = \frac{f(z_i)}{1-F(z_i)}$ is known as inverse Mills ratio, $z_i = -\mathbf{w_i}\alpha/\sigma_2$, $\sigma_{12} = \rho\sigma_1\sigma_2$ and f and F are density and distribution functions of the standard normal distribution. Therefore, sample selection

bias is equal to $\frac{\sigma_{12}}{\sigma_2}\lambda_i$. Then, he rewrote the regression model as

$$y_i = \mathbf{x_i}\beta + \frac{\sigma_{12}}{\sigma_2}\lambda_i + \nu_i \tag{3}$$

where ν_i has mean 0 and variance $\sigma_{11}[(1-\rho^2)+\rho^2(1+z_i\lambda_i-\lambda_i^2)]$ such that $0 \le 1+z_i\lambda_i-\lambda_i^2 \le 1$. It is possible to construct likelihood function and estimate the unknown parameters but since this task involved complex calculations, especially at that time, he estimated the unknown parameters in two steps. Firstly, he estimated α by maximum likelihood estimator using likelihood function:

$$L = \prod_{i=1}^{n} F(z_i)^{m_i}[1-F(z_i)]^{(1-m_i)} \tag{4}$$

where $m_i = 1$ if $y_i^* < 0$ and $m_i = 0$ if $y_i^* \ge 0$, then inverse Mills ratio λ_i was estimated by $\frac{f(\hat{z}_i)}{1-F(\hat{z}_i)}$ and secondly, using $\hat{\lambda}_i$ instead of λ_i in model (3), β, σ_{12} and σ_1^2 were estimated by ordinary least squares regression (OLS). He adjusted the estimator of σ_1^2 in form of $\hat{\sigma_1}^2 = \sum_{i\in S_0}(\hat{\nu}_i^2 - \hat{\alpha}(\hat{\alpha}\hat{z}_i - \hat{z}_i^2))/n_0$, where S_0 is the set of individuals who responded y_i and $\hat{\nu}_i$ is the estimation of residuals that can be obtained from OLS. For identification of probit model in (4), one has to assume $\sigma_2^2 = 1$ (Long 1997, p. 47). Heckman (1979) presented a method for estimation of variance of estimators of parameters based on asymptotic distribution of them. Heckman two-step method is a convenient method for bias correction but it has some weakness including assumption of bivariate normal distribution for errors and not using the exact likelihood of the observations.

3 Sample selection with multiple selection rules

Nonresponse occurs for a variety of reasons (here, levels) in many surveys. Combining these levels into a single category and using a selection model to show their relationship with the main variable of interest may lead to an increasing error or model misspecification. Kim and Kim (2016) presented a method for multivariate selection regression model assuming the errors come from a family of elliptical distributions. They used exact likelihood to drive estimates using an extended version of the EM algorithm and a hierarchical model. In their method, it is not possible to generalize the Heckman's two-step procedure because of non-normality of the errors and finite boundary values of the latent variable for determining nonresponse.

3.1 Model structure

In this section, we increase the number of sample selection models to be equal to the number of nonresponse reasons. We consider the problem to be the study of the relationship between Y and X based on a survey data, in which a percent of samples are nonresponse for some recorded reasons. Let the set of observations in survey be $S = \{(y_1, M_1, \mathbf{x_1}, \mathbf{w_{11}}, \ldots, \mathbf{w_{1K}}), \ldots, (y_n, M_n, \mathbf{x_n}, \mathbf{w_{n1}}, \ldots, \mathbf{w_{nK}})\}$ where y_i is the variable of interest, M_i is the nonresponse indicator of y_i , i.e., 0 for response and j in the case that y_i is nonresponse due to reason j, $j = 1, \ldots, K$ and for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, K$, $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $\mathbf{w_{ij}} = (w_{ij_1}, w_{ij_2}, \ldots, w_{ij_{q_j}})$ are the vectors of known explanatory variables. We use following models to show the relationship between variable of interest (main model), explanatory variables and levels of nonresponses (selection models):

$$y_i = \mathbf{x_i}\beta + e_i, i = 1, 2, \ldots n, \tag{5}$$

with $K(K > 1)$ sample selection models,

$$y_{ij}^* = \mathbf{w_{ij}}\alpha_j + u_{ij}, i = 1, 2, \ldots n, \quad \text{and} \quad j = 1, 2, \ldots, K, \tag{6}$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)', \alpha_j = (\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jq_j})', j = 1, \ldots, K$. We set $y_{ij}^* \ge 0 \Longleftrightarrow M_{ij} = 0, y_{ij}^* < 0 \Longleftrightarrow M_{ij} = j, i = 1, \ldots, n, j = 1, 2, \ldots, K$ and e_i and $u_i = (u_{i1}, u_{i2}, \ldots, u_{iK})$ have multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma} = \begin{bmatrix} \sigma_{00} & \mathbf{\Sigma_{eu}} \\ \mathbf{\Sigma_{ue}} & \mathbf{\Sigma_{uu}} \end{bmatrix}$ where $\mathbf{\Sigma_{eu}} = [\sigma_{01}, \sigma_{02}, \ldots, \sigma_{0K}]$, $\mathbf{\Sigma_{ue}} = \mathbf{\Sigma}'_{\mathbf{eu}}$ and $\mathbf{\Sigma_{uu}}$ is the $K \times K$ covariance matrix of u_i with diagonal elements of 1 due to identifiability and off-diagonal elements of $\sigma_{kl} = \rho_{kl}$. Our main goal is to find estimates of parameters in model (5) and (6) such that bias of sample selections be corrected.

Usually in surveys, reasons of nonresponse have priority of observing over each other, i.e., it is not possible to observe nonresponse reasons all at the same time and only one reason is observable and the others may or may not. For example, if the nonresponse in a survey is due to noncontact and refusal, then for some individuals, only noncontact is observable, i.e., refusal will be observable if contact with respondent can be done. Therefore, by prioritizing the nonresponse reasons, observing $M_i = k$ means that nonresponse reasons were not related to 1 to $k-1$ first items of reasons and also we cannot have any judgment for reasons $k+1$ to K. In this paper, we assume that the reasons for the nonresponse are prioritized, therefore $M_i = 0$ if we have $M_{ij} = 0$ for all of $j = 1, \ldots, K$ and $M_i = j$ if $M_{ij} = 1$ and $M_{it} = 0$ for all of $t = 1, \ldots, j-1$. In this case, the value of M_{it} are unobservable for all $t = j+1, \ldots, K$.

Reviewing literature, there are other models such as sequential, nested, Tobit or double-hurdle models which are commonly used. However, these models cannot cover the issue of priority well. Sequential or nested model can

be used to analyze multilevel nonresponse with priorities, but using this model, it is not possible to consider correlation between the reasons of the nonresponse. Failure to account for this correlation may lead to biased parameter estimates. For example, if the reasons of nonresponse are noncontact and refusal, the sequential model cannot take into account the dependency between the noncontact and refusal processes and this dependency will be unexplained by the covariates in the model (Steele & Durrant, 2011). In the extended Tobit model, there are one selection model for each of dependent variable of interest. The double-hurdle model can be considered as a special case of the method presented in this paper when the number of selection models is 2, the correlation coefficient between the nonresponse reasons is zero and the nonresponse reasons have priority over each other. See, Bruno (2013) and Engel and Moffatt (2014) for more details.

We set $S_j = \{i|M_i = j\}, j = 0, 1, 2, \ldots K$ and use respondent samples to estimate the parameters in models (5) and (6). Since samples are respondent if the corresponding latent variables are nonnegative, we should obtain $E(y_i|x_i, y_{i1}^* \geq 0, \ldots, y_{iK}^* \geq 0)$ and $\mathrm{Var}(y_i|x_i, y_{i1}^* \geq 0, \ldots, y_{iK}^* \geq 0)$. McGill (1992) investigated the moment generating function of truncated normal distribution. Based on his work and Jolani (2014), we have:

$$E(y_i|x_i, y_{i1}^* \geq 0, \ldots, y_{iK}^* \geq 0) = \mathbf{x_i}\beta + \sum_{j=1}^{K} \sigma_{0j}\lambda_{ij} \quad (7)$$

$$\mathrm{Var}(y_i|x_i, y_{i1}^* \geq 0, \ldots, y_{iK}^* \geq 0) = \sigma_{00} + \mathbf{\Sigma_{eu}} H_i \mathbf{\Sigma_{ue}}. \quad (8)$$

In Eq. (7),

$$\lambda_{ij} = \frac{\phi_1(\mathbf{w_{ij}}\alpha_j)\Phi_{K-1}^*(\mathbf{w_{ij}}\alpha_j)}{\Phi_K(\mathbf{w_{i1}}\alpha_1, \ldots, \mathbf{w_{iK}}\alpha_K)}$$

where $\phi_1(\mathbf{w_{ij}}\alpha_j)$ is the density function of the univariate standard normal distribution evaluated at $\mathbf{w_{ij}}\alpha_j$, $\Phi_K(\mathbf{w_{i1}}\alpha_1, \ldots, \mathbf{w_{iK}}\alpha_K)$ is the cumulative distribution function (cdf) of a K-variates normal distribution with mean zero and covariance matrix $\mathbf{\Sigma_{uu}}$ in the form of:

$$\Phi_K(\mathbf{w_{i1}}\alpha_1, \ldots, \mathbf{w_{iK}}\alpha_K) = \int_{-\infty}^{\mathbf{w_{i1}}\alpha_1} \ldots \int_{-\infty}^{\mathbf{w_{iK}}\alpha_K} \phi_K(u_{i1}, \ldots u_{iK}; \mathbf{\Sigma_{uu}}) \mathrm{d}u_{i1} \ldots \mathrm{d}u_{iK},$$

$\Phi_{K-1}^*(\mathbf{w_{ij}}\alpha_j) \equiv 1$ for $K = 1$ and

$$\Phi_{K-1}^*(\mathbf{w_{ij}}\alpha_j) = \int_{-\infty}^{\mathbf{w_{i(j)}}\alpha_{(j)}} \phi_{k-1}(u_{i(j)}|u_{ij} = \mathbf{w_{ij}}\alpha_j; \mathbf{\Sigma_{uu.j}}) \mathrm{d}u_{i(j)}, \quad K > 1$$

where $\phi_{k-1}(u_{i(j)}|u_{ij} = \mathbf{w_{ij}}\alpha_j; \mathbf{\Sigma_{uu.j}})$ is the conditional density function of a $K-1$ variates normal distribution evaluated at the $u_{i(j)}$ (all u_is without the j-th variable) given u_{ij} and $\int_{-\infty}^{\mathbf{w_{i(j)}}\alpha_{(j)}} \phi_{K-1}(u_{i(j)}|u_{ij} = \mathbf{w_{ij}}\alpha_j; \mathbf{\Sigma_{uu.j}}) \mathrm{d}u_{i(j)}$ is the $K-1$ integral of $\phi_{K-1}(u_{i(j)}|u_{ij} = \mathbf{w_{ij}}\alpha_j; \mathbf{\Sigma_{uu.j}})$ on all of $u_{it}, t = 1, \ldots, K, t \neq j$.

In equation (8), H_i is the $K \times K$ matrix with diagonal elements of $h_{jj} = -\mathbf{w_{ij}}\alpha_j\lambda_{ij} - \lambda_{ij}^2$ and off-diagonal elements of $h_{kl} = \lambda_{i,kl}^* - \lambda_{ik}\lambda_{il}$ where

$$\lambda_{i,kl}^* = \frac{\phi_2(\mathbf{w_{ik}}\alpha_k, \mathbf{w_{il}}\alpha_l; \mathbf{\Sigma_{uu}^{kl}})\Phi_{K-2}^*(\mathbf{w_{ik}}\alpha_k, \mathbf{w_{il}}\alpha_l)}{\Phi_K(\mathbf{w_{i1}}\alpha_1, \ldots, \mathbf{w_{iK}}\alpha_K)}$$

and $\phi_2(\mathbf{w_{ik}}\alpha_k, \mathbf{w_{il}}\alpha_l; \mathbf{\Sigma_{uu}^{kl}})$ is the density function of the standard bivariate normal distribution evaluated at $\mathbf{w_{ik}}\alpha_k$ and $\mathbf{w_{il}}\alpha_l$, $\mathbf{\Sigma_{uu}^{kl}}$ is the covariance matrix of u_k and u_l, $\Phi_{K-2}^*(\mathbf{w_{ik}}\alpha_k, \mathbf{w_{il}}\alpha_l) \equiv 1$, $\mathbf{\Sigma_{uu}^{kl}} = \mathbf{\Sigma_{uu}}$ for $K = 2$ and

$$\Phi_{K-2}^*(\mathbf{w_{ik}}\alpha_k, \mathbf{w_{il}}\alpha_l) = \int_{-\infty}^{\mathbf{w_{i(kl)}}\alpha_{(kl)}} \phi_{K-2}(u_{i(kl)}|u_{ik} = \mathbf{w_{ik}}\alpha_k, u_{il} = \mathbf{w_{il}}\alpha_l; \mathbf{\Sigma_{uu.kl}}) \mathrm{d}u_{i(kl)}, \quad K > 2$$

where $\phi_{K-2}(u_{i(k)}, u_{i(l)}|u_{ik} = \mathbf{w_{ik}}\alpha_k, u_{il} = \mathbf{w_{il}}\alpha_l; \mathbf{\Sigma_{uu.kl}})$ is the conditional density function of a $K-2$ variates normal distribution evaluated at the $u_{i(kl)}$ (all u_is without the k-th and the l-th variable) given u_{ik} and u_{il} and $\int_{-\infty}^{\mathbf{w_{i(kl)}}\alpha_{(kl)}} \phi_{K-2}(u_{i(kl)}|u_{ik} = \mathbf{w_{ik}}\alpha_k, u_{il} = \mathbf{w_{il}}\alpha_l; \mathbf{\Sigma_{uu.kl}}) \mathrm{d}u_{i(kl)}$ is the $K-2$ integral of $\phi_{K-2}(u_{i(k)}, u_{i(l)}|u_{ik} = \mathbf{w_{ik}}\alpha_k, u_{il} = \mathbf{w_{il}}\alpha_l; \mathbf{\Sigma_{uu.kl}})$ on all of $u_{it}, t = 1, \ldots, K, t \neq k, l$. Also in equation 8, $\sigma_{00} + \mathbf{\Sigma_{eu}} H_i \mathbf{\Sigma_{ue}}$ should be positive.

The model 5 can be rewritten as follows:

$$y_i = \mathbf{x_i}\beta + \sum_{j=1}^{K} \lambda_{ij}\sigma_{0j} + \nu_i \quad (9)$$

where ν_i is the random error with $E(\nu_i) = 0$ and $\mathrm{Var}(\nu_i) = \mathrm{Var}(y_i)$ for all $i = 1, 2, \ldots, n$.

3.2 Two-step estimation

Now, it is possible to apply Heckman's two-step method. The likelihood function in the firs step with priority of nonresponse reasons is in the form of:

$$
\begin{aligned}
&L(\alpha_1, \alpha_2, \ldots, \alpha_K, \mathbf{\Sigma_{uu}} | M_1, M_2, \ldots, M_K) \\
&= \prod_{i \in S_0} \Phi_K(\mathbf{w_{i1}}\alpha_1, \ldots \mathbf{w_{iK}}\alpha_K) \prod_{i \in S_1} \Phi_1(-\mathbf{w_{i1}}\alpha_1) \prod_{i \in S_2} P(u_{i1} \geq -\mathbf{w_{i1}}\alpha_1, u_{i2} < -\mathbf{w_{i2}}\alpha_1) \\
&\ldots \\
&\prod_{i \in S_{K-1}} P(u_{i1} \geq -\mathbf{w_{i1}}\alpha_1, \ldots, u_{i(K-2)} \geq -\mathbf{w_{i(K-2)}}\alpha_{K-2}, u_{i(K-1)} < -\mathbf{w_{i(K-1)}}\alpha_{(K-1)}) \\
&\prod_{i \in S_K} P(u_{i1} \geq -\mathbf{w_{i1}}\alpha_1, \ldots, u_{i(K-1)} \geq -\mathbf{w_{i(K-1)}}\alpha_{K-1}, u_{iK} < -\mathbf{w_{iK}}\alpha_K)
\end{aligned}
$$

$\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{iK}$ and $\mathbf{\Sigma_{uu}}$ can be estimated by maximum likelihood estimation method. With substitution of the value of $\hat{\lambda_{ij}}$ in equation (9), β and σ_{0j} can be estimated by OLS. $\mathrm{Var}(Y_i | x_i, Y_{i1}^* > 0, \ldots, Y_{iK}^* > 0)$ has heteroscedasticity and the estimate of σ_{00} should be adjusted. Since $\mathrm{Var}(v_i^2) = E(v_i^2)$, it is expected that $\Sigma_{i=1}^{n_0} \hat{v_i}^2 / n_0$ be equal to $\sigma_{00} + \mathbf{\Sigma_{eu}} H_i \mathbf{\Sigma_{ue}}$ and therefore we can adjust the estimator of σ_{00} in form of:

$$
n_0 \hat{\sigma_{00}} = \sum_{i \in n_0} \hat{v}_i^2 - \sum_{i \in n_0} \hat{\mathbf{\Sigma_{eu}}} \hat{H}_i \hat{\mathbf{\Sigma}}_{\mathbf{ue}}.
$$

3.2.1 Estimation of standard errors of the estimators

We can consider model (9) as follows to find the standard errors of the estimators of parameters:

$$
Y = \mathbf{G}\beta^* + V \tag{10}
$$

where $\mathbf{G} = [\mathbf{X}, \hat{\Lambda}]$, $\mathbf{X} = [\mathbf{x_1'}, \ldots, \mathbf{x_n'}]'$, $\hat{\Lambda} = [\hat{\lambda}_1', \hat{\lambda}_2', \ldots, \hat{\lambda}_n']$, $\hat{\boldsymbol{\lambda}}_\mathbf{i} = (\hat{\lambda}_{i1}, \ldots, \hat{\lambda}_{iK}), i = 1, 2, \ldots, n, \quad \beta^* = [\beta', \mathbf{\Sigma_{ue}}]'$ and $\hat{\beta}^* = (\mathbf{G'G})^{-1}\mathbf{G'}Y$. Based on the work done by Lee et al. (1980), the appropriate forms of the standard errors of the parameters in the multivariate sample selection can be expressed by

$$
\begin{aligned}
\mathrm{Cov}(\hat{\beta}^*) =& \sigma_{00}(\mathbf{G'G})^{-1} - (\mathbf{G'G})^{-1}\mathbf{G'\Sigma_{eu}} \\
&[\Delta - \Delta \mathbf{W} \mathbf{\Sigma}^* \mathbf{W}' \Delta] \mathbf{\Sigma_{ue}} \mathbf{G} (\mathbf{G'G})^{-1}
\end{aligned}
$$

$\mathbf{\Sigma_{eu}}$ is a $n \times nK$ dimension matrix with diagonal elements Σ_{eu} and off-diagonal elements 0, Δ is a $nK \times nK$ dimension matrix with diagonal elements $w_{ij}\lambda_{ij} - \lambda_{ij}^2, i = 1, \ldots, n; \quad j = 1, \ldots, K$ and off-diagonal elements 0, $\mathbf{W}$ is a $nK \times (\sum_{j=1}^{K} q_j)$ diagonal block matrix with elements of $[w_{i1}', \ldots, w_{iK}']'$and $\mathbf{\Sigma}^*$ is the asymptotic covariance matrix for the parameters of the first step. The standard errors of the parameters in vector β^* are given by the squared root of the diagonal elements of $\mathrm{Cov}(\hat{\beta}^*)$.

3.3 One-step estimation

With the development of methods to compute the multiple integrals and to optimize multivariate functions in major statistical software, it is possible to obtain estimators by maximizing the exact likelihood. The exact likelihood, with considering priority of nonresponse reasons, is in the form of:

$$
\begin{aligned}
&L(\beta, \alpha_1, \alpha_2, \alpha_K, \mathbf{\Sigma} | \mathbf{Y_{obs}}, M_1, \ldots, M_n) \\
&= \prod_{i \in S_0} \phi_1\left(\frac{y_i - \mathbf{x_i}\beta}{\sigma_{00}^{1/2}}\right) \int_{-\mathbf{w_{i1}}\alpha_1}^{+\infty} \cdots \int_{-\mathbf{w_{iK}}\alpha_K}^{+\infty} \phi_K(u_{i1}, \ldots, u_{iK} | y_i, \mathbf{\Sigma_{uu.0}}) \mathrm{d}u_{iK} \ldots \mathrm{d}u_{i1} \\
&\prod_{i \in S_1} \Phi_1(\mathbf{w_{i1}}\alpha_1) \prod_{i \in S_2} \int_{-\mathbf{w_{i1}}\alpha_1}^{+\infty} \int_{-\infty}^{-\mathbf{w_{i2}}\alpha_2} \phi_2(u_{i1}, u_{i2}; \mathbf{\Sigma_{uu}}) \mathrm{d}u_{i2} \mathrm{d}u_{i1} \\
&\ldots \\
&\prod_{i \in S_{K-1}} \int_{-\mathbf{w_{i1}}\alpha_1}^{+\infty} \cdots \int_{-\mathbf{w_{i[K-2]}}\alpha_{K-2}}^{+\infty} \int_{-\infty}^{-\mathbf{w_{i[K-1]}}\alpha_{K-1}} \phi_{K-1}(u_{i1}, \ldots, u_{i[K-1]}; \mathbf{\Sigma_{uu}}) \mathrm{d}u_{i[K-1]} \ldots \mathrm{d}u_{i1} \\
&\prod_{i \in S_K} \int_{-\mathbf{w_{i1}}\alpha_1}^{+\infty} \cdots \int_{-\mathbf{w_{i[K-1]}}\alpha_{K-1}}^{+\infty} \int_{-\infty}^{-\mathbf{w_{iK}}\alpha_K} \phi_K(u_{i1}, \ldots, u_{iK}; \mathbf{\Sigma_{uu}}) \mathrm{d}u_{iK} \ldots \mathrm{d}u_{i1}
\end{aligned} \tag{11}
$$

where $\mathbf{Y_{obs}}$ is the vector containing the observed y_is. Likelihood function (11) can effectively evaluated by many statistical software such as R.

3.3.1 Estimation of standard errors of the estimators

In this method, since the estimators are obtained from the maximum likelihood estimation method, the variance of the estimators can be obtained approximately from the inverse of the diagonal components of Fisher information. Also, the bootstrap and Jacknife methods may be used.

3.3.2 Test of significancy of the model parameters

In the one-step method, we have exact likelihood, and we can test significancy of the model parameters using likelihood ratio test as follows: With the exact likelihood in (11), it is possible to obtain the estimators with K sample selection models, and

$$X^2 = -2\log\frac{L(\beta^0, \alpha_1^0, \ldots, \alpha_K^0, \Sigma^0|\mathbf{Y_{obs}}, M_1, \ldots, M_n)}{L(\hat{\beta}, \hat{\alpha_1}, \ldots, \hat{\alpha_K}, \hat{\Sigma}|\mathbf{Y_{obs}}, M_1, \ldots, M_n)} \sim \chi^2(df) \tag{12}$$

where $\beta^0, \alpha_1^0, \ldots, \alpha_K^0, \Sigma^0$ are the values of the parameters under the null hypothesis (H_0), $\hat{\beta}, \hat{\alpha_1}, \ldots, \hat{\alpha_K}, \hat{\Sigma}$ are maximum likelihood estimators that are obtained from (11) and df is the number of parameters which are not assumed to be known in H_0.

3.3.3 Sensitivity analysis

By the specification of the exact likelihood in (11), it is possible to use likelihood displacement approach to study the influence of sample selection on estimates of the parameters. The method of local influence was introduced by Cook (1986) and developed by others as a general tool for assessing the influence of local departures from the assumptions underlying the models. These assumptions, since we desire to study the departure of random nonresponse to nonrandom nonresponse, are about the elements of $\mathbf{\Sigma}$, for example $\rho_{01} = 0$, $\rho_{02} = 0$ or $\rho_{01} = \cdots = \rho_{0K} = 0$ may be considered to see the influence of nonresponse on the results. The likelihood displacement LD (w) is defined as:

$$\mathrm{LD}(w) = 2[l(\hat{\theta}) - l(\hat{\theta}|w)] \tag{13}$$

where $\hat{\theta} = (\hat{\alpha_1}, \hat{\alpha_2}, \hat{\beta}, \hat{\mathbf{\Sigma}})$ and w is the $q \times 1$ perturbation vector which shows the departures from the assumptions. In the cases where we desire to study the influence of each reason of nonresponse, e.g., $k - th$ reason, $k = 1, \ldots, K$, $\rho_{0k} = 0$, and w is a scalar around 0 and for influence study of a subset of reasons, simultaneously, w is a multi-dimensional vector. $l(\hat{\theta})$ $(l(\theta|\hat{w}))$ is the maximum log-likelihood with no perturbation (perturbation). When w is univariate, influence graph LD (w) around zero is a convenient tool for studying the local behavior of w. If the graph is strongly curved at zero, it means that sample selection is nonrandom and the parameters are estimated with high precision, and otherwise, sample selection is random.

In the cases when w is multi-dimensional, there are several curvatures. Cook (1986) suggests investigating the direction in which this influence measure changes most rapidly locally. The maximum curvature $C_{\max}$ of the LD (w) surface is given by:

$$C = 2\left|l'\Delta'Q^{-1}\Delta l\right| \tag{14}$$

where Δ is the $P \times q$ matrix with elements of $\frac{\partial^2 l(\theta|w)}{\partial\theta_i \partial w_j}|_{\theta=\hat{\theta}, w=0}, i = 1, \ldots, P; j = 1, \ldots, q$, P is the dimension of the θ with respect to not having perturbation, $\hat{\theta}$ is the estimation of θ under no perturbation, $w = 0$ denotes no perturbation, Q is the $P \times P$ matrix with the elements of $\frac{\partial^2 l(\theta|w)}{\partial\theta_i \partial\theta_j}|_{\theta=\hat{\theta}, w=0}, i = 1, \ldots, P; j = 1, \ldots, P$ and l is the eigen vector corresponding to the maximum absolute eigen value of the matrix $\Delta'Q^{-1}\Delta$. It is straightforward to apply this approach to multivariate selection model. For more details see Billor and Loynes (1993), Cook (1986), Ganjali and Rezaei (2005) and Razie et al. (2013).

4 Simulation studies

The multivariate sample selection model was examined in this section by comparing its performance with those of the univariate selection model (USM) and the regression model with removing nonresponse observations (complete cases, CC). It is possible to run simulations for any number of selection models, but due to the number of different combinations of the nonresponse mechanisms at the nonresponse levels, there will be many cases and reporting them is out of the aim of this paper. For this reason, bivariate selection model (BSM) was considered. To compare the results with the USM and its relationship with the variable of interest, the same explanatory variable was chosen for the selection models and the main model. This variable extracted from a uniform distribution between 1 and 10. In order to investigate the importance of normality assumption for errors, we run simulation in two parts, at first assuming normality and secondly considering non-normality assumption. Moreover, to study the behavior of the proposed method

in different states of the nonresponse mechanisms at the different levels of nonresponse, we consider the following three cases.

- Case 1: The mechanisms of nonresponse at one level is random and at another level is nonrandom $\alpha_{01} = 4.5, \alpha_{11} = -0.6, \alpha_{02} = 1, \alpha_{12} = 0, \sigma_{01} = -0.5, \sigma_{02} = 0, \rho_{12} = 0$
- Case2: Missing not at random (MNAR) mechanisms in the same direction with the variable of interest for both levels of nonresponse $\alpha_{01} = 2$, $\alpha_{11} = -0.2$, $\alpha_{02} = 5$, $\alpha_{12} = -0.7$, $\sigma_{01} = -0.5$, $\sigma_{02} = -0.5$, $\rho_{12} = 0.5$
- Case3: MNAR mechanisms for both levels of nonresponse with different directional with the variable of interest $\alpha_{01} = 4.5$, $\alpha_{11} = -0.5$, $\alpha_{02} = -3$, $\alpha_{12} = 1$, $\sigma_{01} = -0.5, \sigma_{02} = 0.5, \rho_{12} = -0.5$.

We set $\beta_0 = -1$, $\beta_1 = 1.5$, $\sigma_{00} = 1$, the number of iterations of 500 and a sample number of 1000 were used to generate the data. Other sample sizes such as 200 and 500 were used, and the results were consistent with the results of sample size of 1000. All analysis were done in

Table 1 Estimates of the parameters with mean squared errors (in parentheses) with the normality assumption for errors

Case	Parameter	TRUE	BSM (two-step)	BSM (one-step)	USM (one-step)	CC
1	α_{11}	4.5	4.54 (0.20)	4.54 (0.20)	1.95 (2.55)	4.54 (0.19)
	α_{12}	− 0.6	− 0.60 (0.03)	− 0.60 (0.03)	− 0.30 (0.31)	− 0.60 (0.03)
	α_{21}	1	1.05 (0.25)	1.05 (0.25)	–	0.69 (0.34)
	α_{22}	0	− 0.02 (0.07)	− 0.02 (0.07)		0.10 (0.10)
	β_1	− 1	− 0.80 (0.22)	− 0.86 (0.23)	− 0.85 (0.19)	− 0.80 (0.23)
	β_2	1.5	1.44 (0.07)	1.48 (0.05)	1.49 (0.06)	1.43 (0.07)
	ρ_{01}	− 0.5	− 0.01 (0.39)	− 0.31 (0.28)	− 0.35 (0.37)	–
	ρ_{02}	0	0.00 (0.00)	− 0.22 (0.38)	–	–
	ρ_{12}	0	0.05 (0.34)	0.05 (0.37)	–	–
	σ_{00}	1	0.96 (0.06)	1.05 (0.11)	1.10 (0.11)	0.97 (0.06)
2	α_{11}	2	2.07 (0.17)	2.07 (0.17)	2.96 (0.98)	2.07 (0.17)
	α_{12}	− 0.2	− 0.21 (0.02)	− 0.21 (0.02)	− 0.45 (0.25)	− 0.21 (0.02)
	α_{21}	5	4.96 (0.72)	5.02 (0.64)	–	3.92 (1.11)
	α_{22}	− 0.7	− 0.69 (0.08)	− 0.69 (0.09)	–	− 0.43 (0.27)
	β_1	− 1	− 0.96 (0.19)	− 1.01 (0.11)	− 1.04 (0.11)	− 0.83 (0.19)
	β_2	1.5	1.48 (0.12)	1.49 (0.03)	1.52 (0.03)	1.41 (0.09)
	ρ_{01}	− 0.5	− 0.32 (0.49)	− 0.26 (0.42)	− 0.60 (0.21)	–
	ρ_{02}	− 0.5	− 0.10 (0.56)	− 0.47 (0.30)	–	–
	ρ_{12}	0.5	0.63 (0.29)	0.47 (0.55)	–	–
	σ_{00}	1	1.67 (2.73)	1.01 (0.10)	1.03 (0.09)	0.91 (0.11)
3	$\alpha 11$	4.5	4.68 (0.45)	4.68 (0.45)	3.13 (2.61)	4.68 (0.45)
	$\alpha 12$	− 0.5	− 0.52 (0.05)	− 0.52 (0.05)	− 0.32 (0.33)	− 0.52 (0.05)
	$\alpha 21$	− 3	− 3.06 (0.26)	− 3.04 (0.27)	–	− 3.06 (0.26)
	$\alpha 22$	1	1.03 (0.09)	1.02 (0.09)	–	1.03 (0.09)
	$\beta 0$	− 1	− 0.65 (0.60)	− 1.00 (0.25)	− 0.38 (0.69)	− 0.20 (0.81)
	β_1	1.5	1.43 (0.11)	1.50 (0.06)	1.37 (0.13)	1.36 (0.14)
	ρ_{01}	− 0.5	0.12 (0.81)	− 0.42 (0.36)	0.63 (1.13)	–
	ρ_{02}	0.5	0.12 (0.56)	0.63 (0.10)	–	–
	ρ_{12}	− 0.5	− 0.50 (0.10)	− 0.51 (0.21)	–	–
	σ_{00}	1	1.01 (0.17)	1.05 (0.08)	0.99 (0.14)	0.91 (0.10)

Estimates are based on the bivariate selection model using two-step and one-step methods, BSM (two-step) and BSM (one-step), univariate selection model in one-step method, USM (one-step), and complete case analysis (CC), i.e., data after deleting nonresponse cases.

The sample size is $N = 1000$.

Case1: Random nonresponse at one level and MNAR at another level of nonresponse,

Case2: MNAR in the same direction with the variable of interest at both levels of nonresponse,

Case3: MNAR at both levels of nonresponse and with different direction with the variable of interest

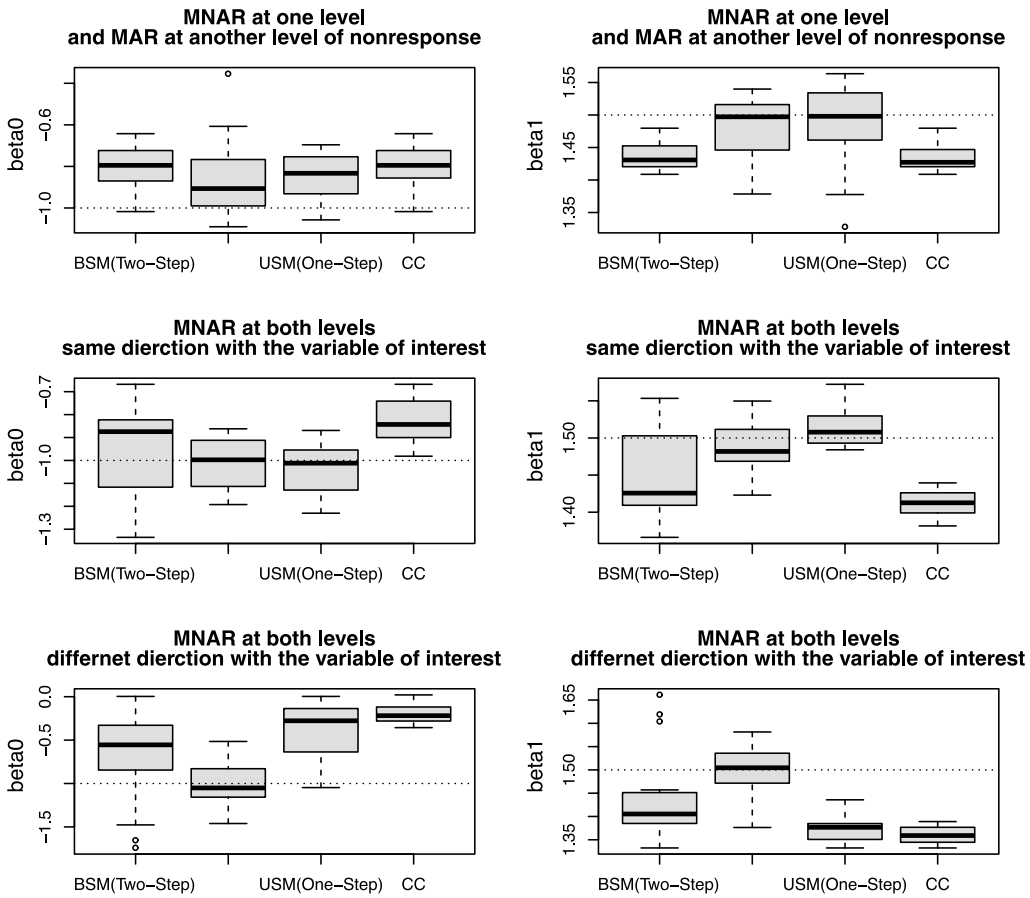

Fig. 1 Boxplots of the estimated regression coefficients of β_0 and β_1 with normality assumption from 500 simulation runs. Methods are the bivariate selection model using two-step and one-step methods, BSM (two-step) and BSM (one-step), respectively, univariate selection model using one-step method, USM (one-step), and complete case analysis (CC), i.e., data after deleting nonresponse cases

Table 2 Root of the mean squared error (RMSE) for three cases of simulation with the normality assumption

Method	BSM (two-step)	BSM (one-step)	USM (one-step)	CC
Case1:	0.056	0.035	0.099	0.058
Case2:	0.428	0.028	0.017	0.146
Case3:	0.094	0.027	0.160	0.138

The same as that of Table 1

R. We applied some packages in R such as "maxLik" and "mvtnorm" to do calculations. The corresponding source code is available on request.

4.1 Normality assumption

We assume the stochastic errors come from a three-variate normal distribution with mean zero and covariance structure as:

$$\Sigma = \begin{bmatrix} \sigma_{00} & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & 1 & \rho_{12} \\ \sigma_{02} & \rho_{12} & 1 \end{bmatrix}.$$

The main model and the selection models are as follows, respectively:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + e_i, \\ y_{i1}^* &= \alpha_{01} + \alpha_{11} w_i + u_{i1}, \\ y_{i2}^* &= \alpha_{02} + \alpha_{12} w_i + u_{i2}. \end{aligned}$$

The average response rate is about 60% in case 1 and about 63% in cases 2 and 3. The average nonresponse rates at level 1 in cases 1 to 3 are about 29, 21 and 15 percent, respectively, and at level 2 are around 11, 16 and 23 percent, respectively. Table 1 shows the results of the simulation.

Figure 1 shows a boxplot of the main model's coefficients estimates to evaluate the performance of the methods. It is observed that in all three cases, the CC

Table 3 Estimates of the parameters with mean squared errors (in parentheses) using the t distribution (3 degrees of freedom)

Case	Parameter	TRUE	BSM (two-step)	BSM (one-step)	USM (one-step)	CC
1	α_{11}	4.50	3.23 (1.30)	3.23 (1.30)	2.88 (1.76)	3.23 (1.30)
	α_{12}	− 0.60	− 0.43 (0.17)	− 0.43 (0.17)	− 0.39 (0.22)	− 0.43 (0.17)
	α_{21}	1.00	0.99 (0.27)	0.98 (0.27)	–	0.63 (0.39)
	α_{22}	0.00	− 0.04 (0.09)	− 0.05 (0.10)	–	0.09 (0.09)
	β_1	− 1.00	− 0.78 (0.24)	− 0.81 (0.25)	− 0.80 (0.22)	− 0.78 (0.24)
	β_2	1.50	1.42 (0.09)	1.45 (0.06)	1.45 (0.08)	1.42 (0.09)
	ρ_{01}	− 0.50	− 0.01 (0.40)	− 0.17 (0.28)	− 0.80 (0.40)	–
	ρ_{02}	0.00	0.00 (0.00)	− 0.08 (0.42)	–	–
	ρ_{12}	0.00	0.14 (0.43)	0.17 (0.47)	–	–
	σ_{00}	1.00	1.92 (1.00)	2.12 (1.18)	2.09 (1.18)	1.93 (1.01)
2	α_{11}	2.00	1.58 (0.43)	1.52 (0.49)	2.18 (0.18)	1.58 (0.43)
	α_{12}	− 0.20	− 0.15 (0.05)	− 0.14 (0.06)	− 0.34 (0.14)	− 0.15 (0.05)
	α_{21}	5.00	4.64 (0.49)	4.12 (1.01)	–	3.69 (1.31)
	α_{22}	− 0.70	− 0.58 (0.13)	− 0.54 (0.16)	–	− 0.40 (0.30)
	β_1	− 1.00	1.60 (3.50)	− 0.88 (0.16)	− 1.09 (0.18)	− 0.83 (0.22)
	β_2	1.50	2.81 (1.85)	1.47 (0.03)	1.56 (0.07)	1.38 (0.12)
	ρ_{01}	− 0.50	− 0.60 (0.53)	− 0.61 (0.36)	− 0.67 (0.28)	–
	ρ_{02}	− 0.50	− 0.15 (0.78)	− 0.25 (0.44)	–	–
	ρ_{12}	0.50	− 0.25 (1.02)	0.14 (0.92)	–	–
	σ_{00}	> 5.00	1.67 (> 5.00)	2.43 (1.45)	2.49 (1.50)	2.02 (1.03)
3	$\alpha 11$	4.5	3.01 (1.50)	3.08 (1.42)	1.06 (3.74)	3.00 (1.50)
	$\alpha 12$	− 0.5	− 0.33 (0.17)	− 0.34 (0.16)	− 0.08 (0.46)	− 0.33 (0.17)
	$\alpha 21$	− 3	− 2.08 (0.95)	− 2.17 (0.89)	–	− 2.00 (1.01)
	$\alpha 22$	1	0.71 (0.30)	0.73 (0.28)	–	0.69 (0.31)
	$\beta 0$	− 1	− 2.12 (2.88)	− 0.38 (0.73)	− 0.43 (0.70)	0.06 (1.07)
	β_1	1.5	1.79 (0.64)	1.39 (0.13)	1.34 (0.16)	1.32 (0.18)
	ρ_{01}	− 0.5	− 0.40 (0.50)	− 0.15 (0.46)	0.63 (1.14)	–
	ρ_{02}	0.5	0.28 (0.47)	0.31 (0.38)	–	–
	ρ_{12}	− 0.5	− 0.69 (0.36)	− 0.80 (0.45)	–	–
	σ_{00}	> 5.00	1.01 (> 5.00)	1.73 (0.76)	2.01 (1.09)	1.69 (0.71)

The same as that of Table 1

method has biases in estimating β_0 and β_1. The method of USM in case 1 and 3 has biases in estimating the intercept, but in case 2, there is almost no bias. It can also be seen in the estimating of β_1 that, this method, in case 3, gives biased estimate. The BSM has no bias in estimating the intercept in cases 2 and 3, and in case 1, the bias of this method is much less than that of the biases of other methods. To estimate β_1, this method gives no bias in cases 1 and 3, and in case 2, its bias is slightly higher than that of the USM. The two-step BSM gives almost bias in all three cases. Considering the above, it can be stated that the efficiency of the one-step BSM is more than that of the univariate, and according to the advances it made in maximization algorithms and computer programs, the method is acceptable. Comparison of the performance of this method with that of the USM in cases 1 and 3, especially 3, has a significant advantage, but in case 2, the USM has more advantages.

Table 2 compares the performances of the methods based on the root of the mean square error (RMSE) criterion using the regression model $y_i = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n,$ where $\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n}}$. Given that all the y_i values of this criterion are observed, it seems to be a suitable criterion for comparing methods. It is observed that the value of this criterion for the one-step BSM method in cases 1 and 3 is less than those of the other methods. The method of USM in case 2 has less RMSE than those of the other methods.

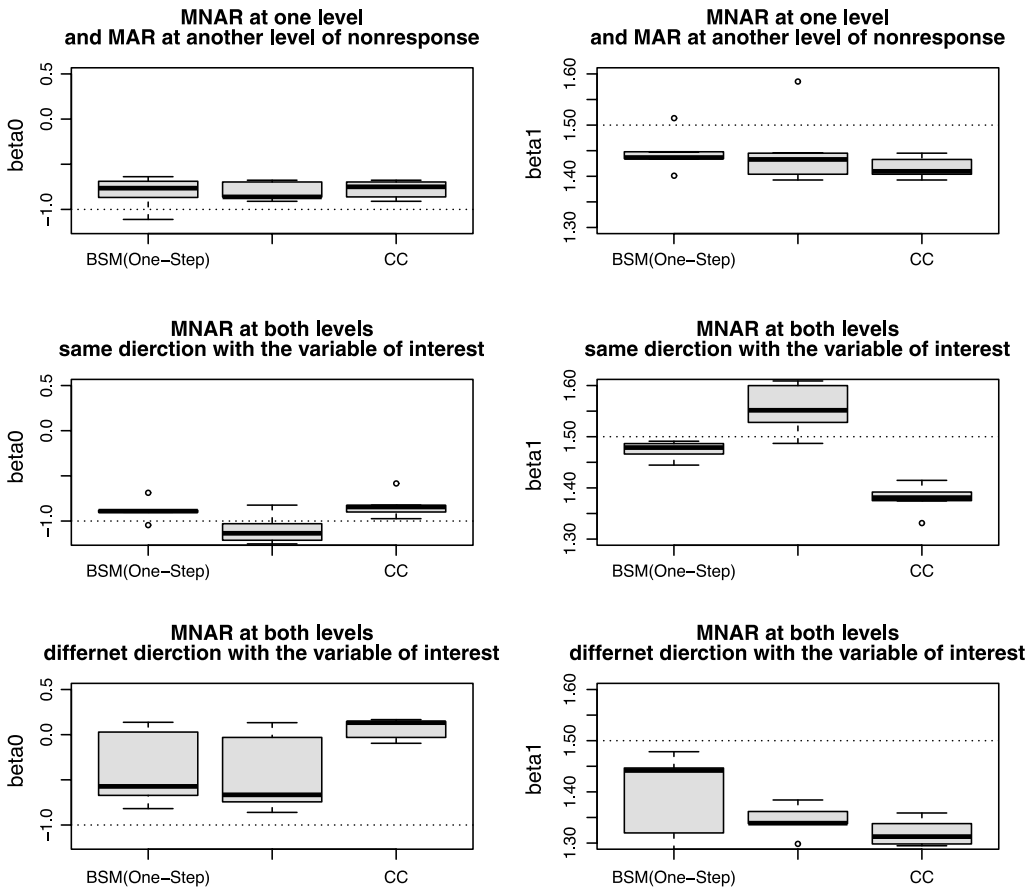

Fig. 2 Boxplots of the estimated regression coefficients of β_0 and β_1 with t distribution assumption from 500 simulation runs. Methods are the bivariate selection model using two-step and one-step methods, BSM (two-step), BSM (one-step), respectively, univariate selection model using one-step method, USM (one-step), and complete case analysis (CC), i.e., data after deleting nonresponse cases

4.2 Non-normality assumption

In considering selection models, errors are mostly assumed to have a multivariate normal distribution due to flexibility in computation and mathematical formulation. However, sensitivity to such an assumption should be considered. For this purpose, the multivariate t distribution can be used, because of its heavier tail than that of the multivariate normal distribution. A simulation study is done in this section to investigate the change of results due to using the distribution of t with 3 or more degrees of freedom. The simulation results are given in Table 3, which show almost close results to those of the normal distribution model, except the use of BSM (two-step) method in the cases 2 and 3.

Figure 2 shows a boxplot of the main model coefficient estimates to evaluate the performance of the methods. Although it is observed that all three cases have a bias in estimating β_0 and β_1, the bias of one-step BSM is less than those of others. Because of low efficiency of two-step BSM, boxplot of using this method was removed from Fig. 2.

Table 4 compares the methods by the RMSE criterion. As in the normality assumption, the efficiency of the BSM (one-step) method is higher than those of other methods. The two-step method is not as efficient as USM (one-step) and CC methods.

Table 4 Root of mean squared error (RMSE) in three cases of simulation, t distribution with 3 degrees of freedom

Method	BSM (two-step)	BSM (one-step)	USM (one-step)	CC
Case1:	> 5.00	0.123	0.167	0.107
Case2:	> 5.00	0.036	0.089	0.351
Case3:	3.079	0.121	0.345	0.227

The same as Table 1

Table 5 Explanatory variables used in models for response and nonresponses

Dependent variable	Explanatory variables	Values
Logarithm of output	Logarithm of the number of employees	Real number
	Status of registration in Bourse	2 categories: 1 for registered and 2 for not registered
	Economic activity code (ISIC)	2 categories: 1 for Manufacture of rubber tyres and tubes plastics products and 2 for Manufacture of other rubber products (Isic4_2219)
Noncontact	Organization of conducting survey	3 categories: Org. 1, Org. 7 and Org. 11
	Status of participation in previous survey	2 categories: 1 for refusal and 2 for noncontact
	Having ancillary unit	2 categories: 1 for having and 2 for not
	Location of establishment	2 categories: 1 for in industrial zone and 2 for out of industrial zone
Refusal	Organization of conducting survey	6 categories: Org. 3, Org. 10, Org. 12, Org. 20, Org. 23 and Org. 25
	Status of participation in previous survey	3 categories: 0 for respondent, 1 for refusal and 2 for noncontact
	Status of being in sample in previous survey	2 categories: 1 for in sample and 2 for not in sample
	Having ancillary unit	2 categories: 1 for having and 2 for not having

Table 6 Estimates of parameters, standard errors (s.e.) and *P* values in the bivariate selection model using two-step and one-step methods

Reason	Parameter	Two-step			One-step		
		Estimate	s.e.	*P* value	Estimate	s.e.	*P* value
Noncontact	Intercept	1.605	0.127	0.000	1.604	0.127	0.000
	Org. 1	− 0.971	0.297	0.000	− 0.970	0.297	0.001
	Org. 7	− 1.019	0.199	0.000	− 1.012	0.199	0.000
	Org. 11	− 1.892	0.504	0.000	− 1.943	0.509	0.000
	Previous status in survey	− 0.378	0.186	0.042	− 0.377	0.185	0.041
	Not having ancillary unit	0.379	0.176	0.031	0.378	0.179	0.035
	Position in industrial zone	0.356	0.140	0.011	0.358	0.141	0.011
Refusal	Intercept	1.155	0.082	0.000	1.184	0.080	0.000
	Org. 3	0.343	0.207	0.096	0.471	0.208	0.023
	Org. 10	0.285	0.162	0.079	0.153	0.166	0.358
	Org. 12	0.694	0.289	0.016	0.675	0.284	0.018
	Org. 20	− 1.083	0.186	0.000	− 1.076	0.181	0.000
	Org. 23	− 0.615	0.112	0.000	− 0.681	0.110	0.000
	Org. 25	− 1.072	0.201	0.000	− 1.059	0.195	0.000
	Previous status in survey	− 1.076	0.152	0.000	− 1.024	0.162	0.000
	Previous status in sample	− 0.595	0.093	0.000	− 0.567	0.093	0.000
	Not having ancillary unit	0.240	0.099	0.015	0.146	0.102	0.151
Correlation between noncontact and refusal		− 0.440	0.535	0.411	− 0.657	0.785	0.402

The sample size is 1236 establishments with 865 respondents, 55 noncontacts and 316 refusals

Table 7 Estimates of correlations and *P* values in the bivariate selection model using two-step and one-step methods

Method	Correlation	Two-step	One-step	*P* value
Bivariate selection models	Output value and noncontact	0.069	0.052	0.852
	Output value and refusal	− 0.576	− 0.411	0.000
Univariate selection model	Output value and nonresponse	− 0.393	− 0.287	0.000

The sample size is 1236 establishments with 865 respondents, 55 noncontacts and 316 refusals. *P* values are calculated for one-step method

5 Application: analysis of an establishment survey

In this section, we apply the presented method on the data of manufacturing with ten employees or more which is one of the most important surveys implemented in the statistical center of Iran. Its results are used for calculation of value added in the manufacturing sector in whole country and provinces. We use this survey on industry of "manufacture of rubber and plastics products" to investigate the effect of covariates on output variable, i.e., the value of all sales of goods and services for each of establishment. Noncontact and refusal are two reasons of nonresponse in this survey, and so we consider two selection models.

5.1 Estimation

Table 5 shows the explanatory variables used in main model and selection models. The values of the explanatory variables are known for all samples before conducting the survey. Initially, we use separate probit models for refusal and noncontact to determine the explanatory variables and exclude variables which are not significant. The number of samples is 1236 establishments of which 865 establishments are respondent, 55 establishments are noncontact and 316 establishments are refusal.

In order to obtain the parameters estimates in models (5) and (6), we applied BSM and USM using both two-step and one-step approaches and also CC, i.e., only using those with observed values. In finding explanatory

Table 8 Estimates of parameters with standard errors (in parentheses) in the main model using the bivariate selection model, the univariate selection model and the complete cases analysis

Parameter	Bivariate selection model		Univariate selection model		Complete cases
	Two-step	One-step	Two-step	One-step	
Intercept	21.249 (0.138)	21.156 (0.119)	21.161 (0.129)	21.098 (0.120)	20.940 (0.106)
Logarithm of the number of employees	1.126 (0.031)	1.132 (0.031)	1.135 (0.031)	1.139 (0.031)	1.149 (0.030)
Registered in Bourse	0.607 (0.252)	0.611 (0.270)	0.649 (0.268)	0.640 (0.269)	0.632 (0.270)
Isic4_2219	− 0.687 (0.112)	− 0.678 (0.112)	− 0.691 (0.115)	− 0.685 (0.114)	− 0.669 (0.113)
Inv. Mills ratio 1	0.074 (0.297)	0.054 (0.086)	− 0.407 (0.137)	− 0.294 (0.096)	
Inv. Mills ratio 2	− 0.619 (0.193)	− 0.427 (0.441)			
SSE	858.368	855.872	866.945	866.338	876.035
MSE	0.999	0.996	1.008	1.007	1.017

The sample size is 1236 establishments with 865 respondents, 55 noncontacts and 316 refusals

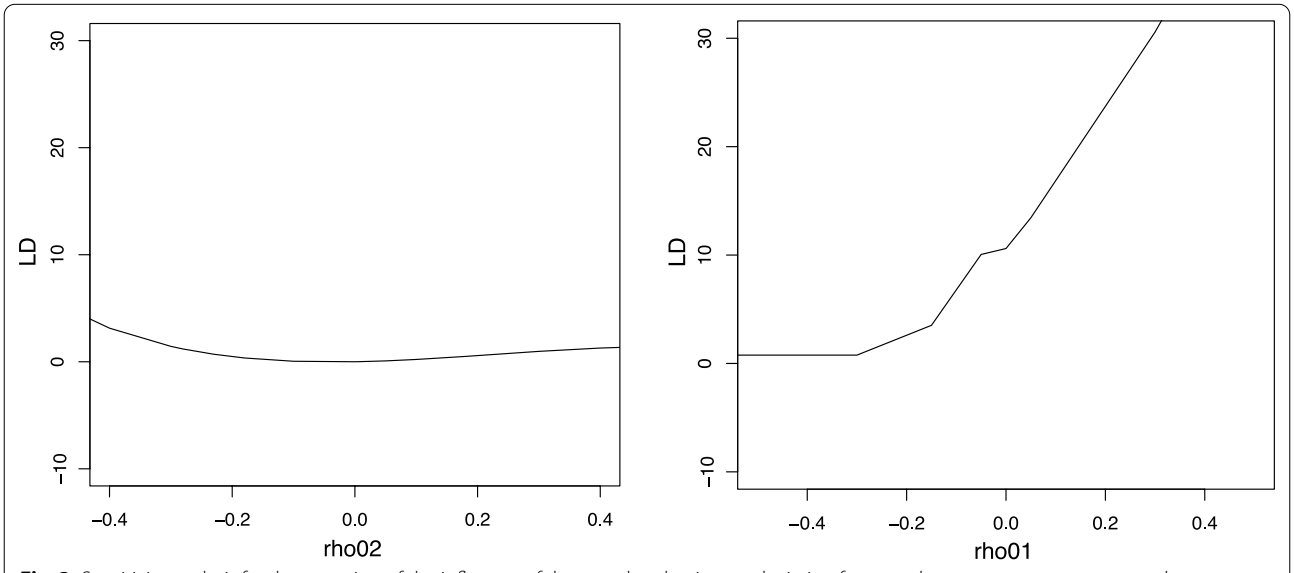

Fig. 3 Sensitivity analysis for the assessing of the influence of the sample selection on deviation from random nonresponse to nonrandom nonresponse, noncontact in the left panel and refusal in the right panel

variables for the USM, variables that did not have significant coefficient were excluded from the model. Therefore, the set of explanatory variables for the USM and the BSM are not the same.

Table 6 shows the estimates in using the selection models provided with standard errors and P values. We use likelihood ratio test to obtain P values. The correlation coefficient between refusal and noncontact is estimated to be negative, and it is equal to -0.440 and -0.657 with respect to using two-step and one-step methods, but it insignificant.

Table 7 shows estimates for the correlations between output value and the reasons of nonresponse. It is seen that correlation between output value and refusal is -0.576 and -0.411 in BSM using two-step and one-step methods, respectively, which is significant, i.e., the higher is the value of the output, the higher is the probability of refusal. In addition, correlation between output value and noncontact is 0.069 and 0.052 in using two- and one-step methods, respectively, which is not significant, that is, nonresponse due to noncontact has no association with the value of output. This shows that the nonresponse mechanism is different among the levels of nonresponse, so with considering just one level as USM, the estimates will be biased as it is shown in the case 1 of the simulation study. Moreover, in USM, it is seen that the correlation between output value and nonresponse is -0.393 and -0.287 in using two- and one-step methods, respectively, which is significant, but by this model, it is not known which levels of nonresponse causes this noningnorable nonresponse.

Based on the results of Table 8, the main model coefficient estimates in the BSM method have differences with those of the USM and CC methods, but their standard errors are almost the same. These differences are due to the distinction between reasons of nonresponse, consideration of the BSM, the unbiased property of the estimates in the BSM method (as shown in Sect. 3) and consideration of different nonresponse mechanisms among different nonresponse levels. The causes of having biases of the parameters estimates of using the USM and CC methods are the lack of the above-mentioned reasons. Moreover, BSM method in two forms (two-step and one-step) has lower MSE than other methods. Also, the MSEs using USM method have less value than that of the complete case method.

5.2 Sensitivity analysis

In order to assess the influence of the sample selections on the estimates, we consider three cases of deviation $\rho_{02} = 0$, $\rho_{01} = 0$ and $\rho_{01} = \rho_{02} = 0$. Figure 3 shows the graph of likelihood-displacement for the first two cases. These graphs are obtained by the equation given in (13). It can be seen from the left panel of Fig. 3 that the value of $\mathrm{LD}(\rho)$ is not large and $l(.|\rho = 0)$ is not curved around zero, and it can be concluded that the estimates will not be affected by noncontact nonresponse. But for refusal nonresponse, it can be seen in Fig. 3, the right panel, that the value of $\mathrm{LD}(\rho)$ is large and so the refusal nonresponse has large effect on the estimates of the parameters. In order to assess the influence of the third case, we apply the equation given in (14). The maximum curvature ($C_{\max}$) in this case is larger than 3, and it can be concluded that estimates are sensitive to the kind of missing mechanism. These results are also consistent with the assessment of the model by using the Akaike information criterion (AIC). The values of AIC are 4075.748, 4089.000 and 4087.746, respectively, for case 1 to case 3. AIC for the model with three correlation coefficient is 4077.728 which is slightly more than that of the model with no ρ_{02}.

6 Conclusion and discussion

In this paper, we presented a method for analysis of survey data with modeling of dependent multilevel nonresponse. In this method, the number of selection models is equal to the number of reasons of nonresponse. We assumed a multivariate normal distribution for the error terms of these models and the response model. The parameters can be estimated using a two-step method or the one-step (full likelihood) method. In this approach, we assumed that there is only one variable of interest as response for modeling. However, this approach can be extended to cases where there are more than one variable of interest.

In a set of simulation studies, performance of the proposed method in the case of BSM, and that of Heckman model were compared. It turns out that the proposed method (in two forms of one-step and two-step) has better performance than that of USM in the cases with different signs of correlation for dependent variable of interest and the nonresponse levels. Moreover, it performs at least as well as the USM when the sign of correlation is the same and one-step method is used. In other words, well-performance of the BSM using two-step method is less than that of the USM using one-step method.

The normality assumption for errors is mostly assumed due to flexibility in computation and mathematical formulation. However, sensitivity to such an assumption was considered by using multivariate t distribution with degrees of freedom of 3 or more, because of its heavier tail than that of the multivariate normal distribution. The results of simulation show that the estimates are biased in both bivariate and univariate selection models and CC analysis, but the bias of BSM

using one-step method is less than those of others. Of course, for higher degrees of freedom than 3, the bias will be small because of convergence of the t distribution to normal distribution. The results show that BSM using two-step method without normality assumption is not very effective.

The results of using this method on the data of an establishment survey show that the MSE obtained using the proposed model is less than that of the USM. This is consistent with the simulation studies where nonresponse mechanism at the noncontact is random and at the refusal is nonrandom.

The AIC value of this method is less than that of the method without consideration of correlation between output value and refusal and noncontact. Noncontact reason is not associated with output value significantly and therefore, the AIC value of the model without correlation between output value and noncontact is less than that of the model. Although it is not possible to compare the AIC values between this method and the USM because of having different likelihoods in two methods, in the univariate case, merging of the reasons of nonresponse causes nonsignificancy of the correlation between noncontact and output value, and this may lead to loss of information in our inference about variable of interest.

We applied the proposed method on an establishment survey, but this method can also be used for household surveys. In using this method, there must be a sufficient number of nonresponse samples at different levels of nonresponse so that the estimation of the parameters in the selection models can reach acceptable accuracy.

Abbreviations

cdf: Cumulative distribution function; ISIC4: International Standard Industrial Classification of All Economic Activities (ISIC), Revision 4; Org.: Organization of conducting survey; BSM: Bivariate Selection Model; USM: Univariate Selection Model; CC: Complete Case analysis; AIC: Akaike information criterion.

Acknowledgements

We would like to thank the editor, referees and Adeniyi Francis Fagbamigbe for reading and giving many improving comments.

Authors' contributions

MG was supervisor of the paper and proposed the idea of considering nonresponse according to its reasons in modeling the survey results in the form of a multivariate selection model. He also proposed sensitivity analysis for the proposed model. AR obtained detailed solution for idea and applied it in an establishment survey and analyzed and interpreted the results. EBS was a major contributor in writing the manuscript and improve the sensitivity analysis. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The data that support the findings of this study have been provided by the Statistical Center of Iran. Such data are not publicly available.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 17 July 2021 Accepted: 7 March 2022
Published online: 02 April 2022

References

Billor, N., & Loynes, R. (1993). Local influence: A new approach. *Communications in Statistics-Theory and Methods, 22,* 1595–1611. https://doi.org/10.1080/03610929308831105.

Bruno, G. (2013). Implementation of a double-hurdle model. *The Stata Journal, 13* (4), 776–794. https://doi.org/10.1177/1536867X1301300406.

Catsiapis, G. & Robinson, C. (1978). *Sample selection bias with two selection rules: An application to student aid grants*. UWO Department of Economics Working Papers 7833, University of Western Ontario, Department of Economics.

Catsiapis, G., & Robinson, C. (1982). Sample selection bias with multiple selection rules: An application to student aid grants. *Journal of Econometrics, 18,* 351–368. https://doi.org/10.1016/0304-4076(82)90088-4.

Cook, R. D. (1986). Assessment of local influence (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological), 48* (2), 133–169. https://doi.org/10.1111/j.2517-6161.1986.tb01398.x.

Durrant, G. B., & Steele, F. (2009). Multilevel modelling of refusal and noncontact in household surveys: Evidence from six UK Government surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 172* (2), 361–381. https://doi.org/10.1111/j.1467-985X.2008.00565.x.

Earp, M., Mitchell, M., McCarthy, J., & Kreuter, F. (2014). Modeling nonresponse in establishment surveys: Using an ensemble tree model to create nonresponse propensity scores and detect potential bias in an agricultural survey. *Journal of Official Statistics, 30* (4), 701–719. https://doi.org/10.2478/JOS-2014-0044.

Earp, M., Toth, D., Phipps, P., & Oslund, C. (2018). Assessing nonresponse in a longitudinal establishment survey using regression trees. *Journal of Official Statistics, 34* (2), 463–481. https://doi.org/10.2478/jos-2018-0021.

Engel, C., & Moffatt, P. G. (2014). dhreg, xtdhreg, and bootdhreg: Commands to implement double-hurdle regression. *The Stata Journal, 14* (4), 778–797. https://doi.org/10.1177/1536867X1401400405.

Ganjali, M., & Rezaei, M. (2005). An influence approach for sensitivity analysis of non-random dropout based on the covariance structure. *Iranian Journal of Science & Technology, Transaction A, 29* (A2), 287–294.

Hanoch, G. (1976). *A multivariate model of labor supply: Methodology for estimation*. RAND Corporation. https://www.rand.org/pubs/reports/R1869.html. (**Also available in print form**).

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement, 5* (4), 475–492.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47* (1), 153–161. https://doi.org/10.2307/1912352.

Heerwegh, D., Abts, K., & Loosveldt, G. (2007). Minimizing survey refusal and noncontact rates: Do our efforts pay off? *Survey Research Methods, 1* (1), 3–10. https://doi.org/10.18148/srm/2007.v1i1.46.

Jolani, S. (2014). An analysis of longitudinal data with nonignorable dropout using the truncated multivariate normal distribution. *Journal of Multivariate Analysis, 131,* 163–173. https://doi.org/10.1016/j.jmva.2014.06.016.

Kim, H. J., & Kim, H. M. (2016). Elliptical regression models for multivariate sample-selection bias correction. *Journal of the Korean Statistical Society, 45* (3), 422–438. https://doi.org/10.1016/j.jkss.2016.01.003.

Kirchner, A., & Signorino, C. S. (2018). Using support vector machines for survey research. *Survey Practice, 11* (1), 1–11. https://doi.org/10.29115/SP-2018-0001.

Lee, L. F., Maddala, G. S., & Trost, R. P. (1980). Asymptotic covariance matrices of two-stage probit and two-stage tobit methods for simultaneous equations models with selectivity. *Econometrica, 48* (2), 491–503. https://doi.org/10.2307/1911112.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. SAGE Publications Inc.
McGill, J. I. (1992). The multivariate hazard gradient and moments of the truncated multinormal distribution. *Communications in Statistics-Theory and Methods, 21* (11), 3053–3060. https://doi.org/10.1080/03610929208830962.
Paiva, T., & Reiter, J. P. (2017). Stop or continue data collection: A nonignorable missing data approach for continuous variables. *Journal of Official Statistics, 33* (3), 579–599. https://doi.org/10.1515/JOS-2017-0028.
Phipps, P., & Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *The Annals of Applied Statistics, 6* (2), 772–794. https://doi.org/10.1214/11-AOAS521.
Razie, F., Bahrami, E., & Ganjali, M. (2013). Analysis of mixed correlated bivariate negative binomial and continuous responses. *Application and Allied Mathematics, 8* (2), 404–415.
Rezaee, A., Ganjali, M., & Bahrami, E. (2021). Nonresponse prediction in an establishment survey using combination of machine learning methods. *Andishe, 25* (1), 101–109.
Seiler, C. (2010). *Dynamic modelling of nonresponse in business surveys*. Ifo Working Paper 93, ifo Institute - Leibniz Institute for Economic Research at the University of Munich.
Steele, F., & Durrant, G. B. (2011). Alternative approaches to multilevel modelling of survey noncontact and refusal. *International Statistical Review, 79* (1), 70–91. https://doi.org/10.1111/j.1751-5823.2011.00133.x.
Vassallo, R., Durrant, G. B., & Smith, P. F. (2015). Interviewer effects on nonresponse propensity in longitudinal surveys: A multilevel modelling approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 178* (1), 83–99. https://doi.org/10.1111/rssa.12049.

Publisher's Note