Nonresponse Adjustment using Auxiliary Variables Subject themselves to Missing Data

CHRIS SKINNER[1], NUANPAN LAWSON[2,*]
[1]Department of Statistics,
London School of Economics and Political Science,
WC2A 2AE,
UNITED KINGDOM

[2]Department of Applied Statistics, Faculty of Applied Science,
King Mongkut's University of Technology North Bangkok,
1518 Pracharat 1 Road, Wongsawang, Bangsue, Bangkok 10800,
THAILAND

Abstract: - Nonresponse is a significant matter that cannot be denied in a sample survey. Declining response rates lead to increasing nonresponse bias which affects the estimated bias. Nonresponse adjustment can be used to deal with unit nonresponse by using nonresponse weight. Two possible models in which missingness in an ancillary database may be correlated with missingness in a survey are considered in this study for estimating the population mean when nonresponse occurs on both the study and auxiliary variables. Two auxiliary variables where one auxiliary variable is fully observed and some part of the other is missing are considered in the possible models. Simulation studies are carried on to see how the nonresponse adjustment using auxiliary variables that subject themselves to nonresponse work under the possible models. The simulation results show that the weighted mean performed the best in removing the bias and gave the minimum mean square error compared to the unweighted mean which was affected by nonresponse.

Key-Words: - Nonresponse adjustment, Missing data, Propensity score weights, Logistic regression, Auxiliary variables, Bias, Mean square error.

Received: November 16, 2024. Revised: December 19, 2024. Accepted: February 6, 2025. Published: April 1, 2025.

1 Introduction

Sample survey inevitably faces the problem of non-response despite how intricate the sample survey design is as it can seldom be controlled. Non-response can occur in many ways. For example, the survey participant may refuse to answer some questions such as privacy-related or sensitive issues, or not answer due to a language barrier or sickness. On the other hand, the survey taker might be unable to reach some respondents. To prevent this, the survey must be designed to be easily understandable and able to engage the respondent. However, in the end, that cannot always ensure a complete dataset and non-response does not occur due to a flaw in the design, so to decrease bias and variance, standard statistical techniques to adjust for non-response before analysis are utilized. Weighting methods can assist in dealing with unit non-response in a post-survey; this has the added benefit of reducing non-response bias. It is imperative to deal with non-response to prevent errors leading to inconclusive results. A strong relationship between the response propensity and the variable of interest in the sample survey can be utilized for non-response adjustment. The auxiliary variables have been used as predictors in propensity models, [1], [2], [3], [4].

A cluster-level regression model under non-response was studied to solve the problem of biasing effects caused by cluster-level association between response rates and cluster-level quantities obtained from survey variables, [5]. They considered the case where testing for inclusion of a non-response rate or some function of it as a covariate in the model may indicate nonresponse. Two models of nonresponse mechanisms with potential biasing effects were introduced along with methods to control the bias by including a non-response rate or some function of it as a covariate in the model. The results found that biases and mean square errors decreased as the non-response rate was included in the model, [6], [7].

Many researchers studied the useful information on auxiliary variables for survey adjustment. For example, [8] investigated the bias and variance of the adjusted response means by using multiple auxiliary variables that correlated to the response indicator and the survey outcome variable in different directions. They found that the differences in the direction of the relationship between the predictors and either propensity or the survey variables gave different bias and mean square error (MSE) for the adjusted respondent mean, [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19].

In this paper, we consider two possible models in which missingness in an ancillary database may be correlated with missingness in the survey. We focus on the problem of estimating the population mean of the response variable when nonresponse occurs on both the study and the auxiliary variables. For simplicity, we consider two auxiliary variables in the possible models where one auxiliary variable is fully observed and some part of the other is missing. Simulation studies are presented in which we consider the properties of these estimators based on two possible models under the assumption that the data are generated from the assumed models in order to see how they are going to account for nonresponse bias.

The formal framework for the paper is set out in Section 2 and the possible models that could account for the correlation between missingness in the ancillary database and missingness in the survey are given in Section 3. The simulation studies are used to see the performance of these estimators in Section 4. Conclusions are given in Section 5.

2 Framework

Assume survey sample s. Respondent set $r \subset s$. Let

$$R_i = \begin{cases} 1 & \text{if } i \text{ in } r \\ 0 & \text{if } i \text{ not in } r. \end{cases}$$

Measure survey variables y_i for i in $r,\ i = 1,2,...,n$ and consider weighted estimator using weights w_i for i in r. If we observe x_i for i in s then might determine w_i by propensity score weights (based on logistic regression of R_i on x_i).

3 Possible Models

In this study, we suppose that $x_i = (x_{1i}, x_{2i})$, where x_{1i} is observed for all i in s and x_{2i} is observed for i in r_2, where r_2 is some subset of s, which will typically include some units from r and some units from s/r. Let

$$R_{2i} = \begin{cases} 1 & \text{if } i \text{ in } r_2 \\ 0 & \text{if } i \text{ not in } r_2. \end{cases}$$

This suggests that missingness in the ancillary database may be correlated with missingness in the survey. So may expect R_i and R_{2i} to be correlated. However, we may not find R_{2i} to be very related to y_i conditional on $R_i = 1$.

Simple Model A:

Suppose A1: R_i is conditionally independent of y_i given $x_i = (x_{1i}, x_{2i})$ and $R_{2i} = 1$.

Suppose A2: R_i is conditionally independent of y_i given x_{1i} and $R_{2i} = 0$.

Under these assumptions, we can estimate $P(R_i = 1)$ by logistic regression of R_i on x_i for cases with $R_{2i} = 1$ and by logistic regression of R_i on x_{1i} for cases with $R_{2i} = 0$. We can then set nonresponse weight to be $P(R_i = 1)^{-1}$. We evaluate properties of weighting following [8]. We could also consider cases where x_{2i} which is strongly related to y_i and different amounts of missingness on R_{2i}.

4 Simulation Studies

In this section we follow [8] to generate y_i and response propensity using R program, [20]. We consider cases where x_{2i} is strongly related to y_i and different amounts of missingness on R_{2i}. The simulation steps are as follows.

Simulation steps:

Step 1 Generate x_{1i} and x_{2i} from bivariate normal distribution with a mutual correlation of -0.2, 0 and 0.2 and mean is equal to zero and variance equal to one with a population of size $N = 100{,}000$.

Step 2 Generate u_i from a normal distribution with mean equal to zero and variance equal to one. Then generate $y_i = 10 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$, where β_1 and β_2 are varied in order to generate the different levels for the correlation between y_i and x_{2i}.

Step 3 Select simple random samples of sizes $n =$ 1,000 and 2,500 and repeat M = 1,000 times.

Step 4 Generate a response probability π_{2i},

$$\pi_{2i} = \frac{e^{1+\gamma_2 x_{2i}}}{1+e^{1+\gamma_2 x_{2i}}},$$

$\gamma_2 = 2$ and then generate a binary response indicator R_{2i} from a binomial distribution with probability $p_{2i}, R_2 \sim B(1, p_{2i})$.

Step 5 Generate a response probability π_i,

$$\pi_i = \begin{cases} \dfrac{e^{1+\gamma_1 x_{1i}+\gamma_2 x_{2i}}}{1+e^{1+\gamma_1 x_{1i}+\gamma_2 x_{2i}}} & \text{if } R_{2i} = 1 \\ \dfrac{e^{1+\gamma_1 x_{1i}}}{1+e^{1+\gamma_1 x_{1i}}} & \text{if } R_{2i} = 0 \end{cases}$$

, $\gamma_1 = 0.1,\ 1,\ 2,\ \gamma_2 = 2$ and then generate a binary response indicator R_i from a binomial distribution with probability $p_i, R_i \sim B(1, p_i)$.

Step 6 Assume A1 and A2 hold, we can estimate $P(R_i = 1)$ by logistic regression of R_i on x_i for cases with $R_{2i} = 1$ and by logistic regression of R_i on x_i for cases with $R_{2i} = 0$ as follows.

Assume A1 holds;

$$logit\left(P\left(R_i = 1\right)\middle|R_{2i} = 1\right) = \hat{\gamma}_{01} + \hat{\gamma}_{11}x_{1i} + \hat{\gamma}_{01}x_{2i}$$

Assume A2 holds;

$$logit\left(P\left(R_i = 1\right)\middle|R_{2i} = 0\right) = \hat{\gamma}_{02} + \hat{\gamma}_{12}x_{1i}$$

Step 7 Calculate the weight w_i by,

$$w_i = \frac{1}{\pi_i p\left(x_i, \hat{\beta}\right)}$$

, where

$$p\left(x_i, \hat{\beta}\right) = \begin{cases} \dfrac{e^{1+\hat{\gamma}_1 x_{1i}+\hat{\gamma}_2 x_{2i}}}{1+e^{1+\hat{\gamma}_1 x_{1i}+\hat{\gamma}_2 x_{2i}}} & \text{if } R_{2i} = 1 \\ \dfrac{e^{1+\hat{\gamma}_1 x_{1i}}}{1+e^{1+\hat{\gamma}_1 x_{1i}}} & \text{if } R_{2i} = 0 \end{cases}$$

Step 8 Compute the unweighted mean and the weighted mean from

$$\bar{y} = \frac{\sum_{i=1}^{n} R_i y_i}{\sum_{i=1}^{n} R_i} \quad (1)$$

$$\bar{y}_{weighted} = \frac{\sum_{i=1}^{n_r} w_i y_i}{\sum_{i=1}^{n_r} w_i} \quad (2)$$

where w_i is the estimated weights from Step 7 and n_r is the number of respondents.

Step 9 Compare each estimator using bias and MSE. The bias and MSE formulas are

$$Bias(\bar{y}) = \frac{1}{1000}\sum_{i=1}^{1000}\left|\bar{y}_i - \bar{Y}\right| \quad (3)$$

$$MSE(\bar{y}) = \frac{1}{1000}\sum_{i=1}^{1000}\left(\bar{y}_i - \bar{Y}\right)^2 \quad (4)$$

The results are shown in Table 1, Table 2, Table 3, Table 4, Table 5 and Table 6.

Table 1, Table 2, Table 3, Table 4, Table 5 and Table 6 showed the bias and mean square error for the weighted mean using x_{1i} and x_{2i} and the weighted mean using x_{1i} compared to the unweighted mean when response rates (r) are varied between 0.68 and 0.76 and the response rate is 0.65 as a result the nonresponse rate is 35% in this study. The correlation between y and x_2 and y and x_1 are varied between 0.28 and 0.9 and the sample of sizes n are equal to 1,000 and 2,500 respectively.

Table 1. Simulation results for n = 1,000 and $\rho_{x1x2} = -0.2$.

r	ρ_{yx_2}	ρ_{yx_1}	β_2	β_1	γ_2	γ_1	Estimator	Bias	MSE
0.76	0.6	0.6	2	2	2	0.1	1.Unweighted mean	0.278	0.087
							2.Weighted mean using x_{1i} and x_{2i}	0.216	0.057
							3.Weighted mean using x_{1i}	0.236	0.067
0.74						1	1.Unweighted mean	0.571	0.336
							2.Weighted mean using x_{1i} and x_{2i}	0.075	0.014
							3.Weighted mean using x_{1i}	0.133	0.034
0.70						2	1.Unweighted mean	0.831	0.701
							2.Weighted mean using x_{1i} and x_{2i}	0.009	0.057
							3.Weighted mean using x_{1i}	0.045	0.079
0.76	0.3	0.85		4	2	0.1	1.Unweighted mean	0.276	0.098
							2.Weighted mean using x_{1i} and x_{2i}	0.170	0.048
							3.Weighted mean using x_{1i}	0.179	0.059
0.74						1	1.Unweighted mean	0.901	0.834
							2.Weighted mean using x_{1i} and x_{2i}	0.020	0.022
							3.Weighted mean using x_{1i}	0.195	0.116
0.70						2	1.Unweighted mean	1.470	2.182
							2.Weighted mean using x_{1i} and x_{2i}	0.036	0.161
							3.Weighted mean using x_{1i}	0.910	1.401
0.77	0.86	0.29	4	2	2	0.1	1.Unweighted mean	0.563	0.341
							2.Weighted mean using x_{1i} and x_{2i}	0.484	0.260
							3.Weighted mean using x_{1i}	0.533	0.308
0.75						1	1.Unweighted mean	0.818	0.693
							2.Weighted mean using x_{1i} and x_{2i}	0.209	0.066
							3.Weighted mean using x_{1i}	0.598	0.383
0.71						2	1.Unweighted mean	1.028	1.083
							2.Weighted mean using x_{1i} and x_{2i}	0.003	0.100
							3.Weighted mean using x_{1i}	1.052	1.246
0.77	0.63	0.62		4	2	0.1	1.Unweighted mean	0.561	0.349
							2.Weighted mean using x_{1i} and x_{2i}	0.438	0.225
							3.Weighted mean using x_{1i}	0.476	0.264
0.75						1	1.Unweighted mean	1.147	1.350
							2.Weighted mean using x_{1i} and x_{2i}	0.155	0.056
							3.Weighted mean using x_{1i}	0.270	0.132
0.70						2	1.Unweighted mean	1.667	2.813
							2.Weighted mean using x_{1i} and x_{2i}	0.024	0.214
							3.Weighted mean using x_{1i}	0.095	0.283

For n = 1,000, Table 1 showed that the weighted mean using x_{1i} and x_{2i} performed the best in terms of both minimum bias and mean square error which gave a lot better results than the unweighted mean in all situations. The weighted mean using x_{1i} performed the second best and the unweighted mean

DOI: 10.37394/23202.2025.24.12

performed the worst. Unweighted mean is biased due to nonresponse and therefore gave the highest bias and mean square errors.

Table 2 and Table 3 show the results for n = 1,000 when the correlation between x_1 and x_2 are equal to 0 and 0.2 respectively, which also gave similar results to Table 1. The weighted mean using x_{1i} and x_{2i} performed the best in all situations.

Table 2. Simulation results for n = 1,000 and $\rho_{x1x2} = 0$.

r	ρ_{yx_2}	ρ_{yx_1}	β_2	β_1	γ_2	γ_1	Estimator	Bias	MSE
0.77	0.67	0.67	2	2	2	0.1	1.Unweighted mean	0.323	0.116
							2.Weighted mean using x_{1i} and x_{2i}	0.244	0.072
							3.Weighted mean using x_{1i}	0.267	0.085
0.75						1	1.Unweighted mean	0.676	0.468
							2.Weighted mean using x_{1i} and x_{2i}	0.074	0.017
							3.Weighted mean using x_{1i}	0.162	0.043
0.70						2	1.Unweighted mean	0.980	0.973
							2.Weighted mean using x_{1i} and x_{2i}	0.010	0.084
							3.Weighted mean using x_{1i}	0.082	0.093
0.77	0.44	0.87		4	2	0.1	1.Unweighted mean	0.367	0.163
							2.Weighted mean using x_{1i} and x_{2i}	0.237	0.082
							3.Weighted mean using x_{1i}	0.255	0.098
0.75						1	1.Unweighted mean	1.059	1.149
							2.Weighted mean using x_{1i} and x_{2i}	0.039	0.030
							3.Weighted mean using x_{1i}	0.047	0.065
0.71						2	1.Unweighted mean	1.667	2.804
							2.Weighted mean using x_{1i} and x_{2i}	0.000	0.245
							3.Weighted mean using x_{1i}	0.529	0.719
0.77	0.87	0.44	4	2	2	0.1	1.Unweighted mean	0.607	0.397
							2.Weighted mean using x_{1i} and x_{2i}	0.500	0.283
							3.Weighted mean using x_{1i}	0.553	0.336
0.75						1	1.Unweighted mean	0.973	0.975
							2.Weighted mean using x_{1i} and x_{2i}	0.188	0.063
							3.Weighted mean using x_{1i}	0.540	0.322
0.70						2	1.Unweighted mean	1.278	1.665
							2.Weighted mean using x_{1i} and x_{2i}	0.029	0.150
							3.Weighted mean using x_{1i}	0.782	0.732
0.77	0.7	0.7		4	2	0.1	1.Unweighted mean	0.651	0.469
							2.Weighted mean using x_{1i} and x_{2i}	0.493	0.290
							3.Weighted mean using x_{1i}	0.540	0.341
0.75						1	1.Unweighted mean	1.356	1.883
							2.Weighted mean using x_{1i} and x_{2i}	0.153	0.067
							3.Weighted mean using x_{1i}	0.330	0.174
0.71						2	1.Unweighted mean	1.965	3.905
							2.Weighted mean using x_{1i} and x_{2i}	0.018	0.331
							3.Weighted mean using x_{1i}	0.170	0.355

Table 3. Simulation results for n = 1,000 and $\rho_{x1x2} = 0.2$.

r	ρ_{yx_2}	ρ_{yx_1}	β_2	β_1	γ_2	γ_1	Estimator	Bias	MSE
0.77	0.74	0.74	2	2	2	0.1	1.Unweighted mean	0.371	0.152
							2.Weighted mean using x_{1i} and x_{2i}	0.284	0.097
							3.Weighted mean using x_{1i}	0.309	0.112
0.74						1	1.Unweighted mean	0.787	0.633
							2.Weighted mean using x_{1i} and x_{2i}	0.082	0.021
							3.Weighted mean using x_{1i}	0.193	0.056
0.70						2	1.Unweighted mean	1.194	1.439
							2.Weighted mean using x_{1i} and x_{2i}	0.042	0.080
							3.Weighted mean using x_{1i}	0.184	0.108
0.77	0.6	0.9		4	2	0.1	1.Unweighted mean	0.454	0.239
							2.Weighted mean using x_{1i} and x_{2i}	0.314	0.133
							3.Weighted mean using x_{1i}	0.340	0.154
0.74						1	1.Unweighted mean	1.215	1.507
							2.Weighted mean using x_{1i} and x_{2i}	0.061	0.037
							3.Weighted mean using x_{1i}	0.076	0.060
0.70						2	1.Unweighted mean	1.867	3.516
							2.Weighted mean using x_{1i} and x_{2i}	0.028	0.254
							3.Weighted mean using x_{1i}	0.226	0.342
0.77	0.9	0.6	4	2	2	0.1	1.Unweighted mean	0.665	0.476
							2.Weighted mean using x_{1i} and x_{2i}	0.544	0.335
							3.Weighted mean using x_{1i}	0.594	0.388
0.74						1	1.Unweighted mean	1.152	1.359
							2.Weighted mean using x_{1i} and x_{2i}	0.189	0.070
							3.Weighted mean using x_{1i}	0.510	0.296
0.70						2	1.Unweighted mean	1.558	2.460
							2.Weighted mean using x_{1i} and x_{2i}	0.044	0.186
							3.Weighted mean using x_{1i}	0.619	0.492
0.77	0.76	0.77		4	2	0.1	1.Unweighted mean	0.748	0.614
							2.Weighted mean using x_{1i} and x_{2i}	0.574	0.391
							3.Weighted mean using x_{1i}	0.624	0.451
0.74						1	1.Unweighted mean	1.580	2.547
							2.Weighted mean using x_{1i} and x_{2i}	0.168	0.084
							3.Weighted mean using x_{1i}	0.392	0.224
0.70						2	1.Unweighted mean	2.387	5.749
							2.Weighted mean using x_{1i} and x_{2i}	0.084	0.308
							3.Weighted mean using x_{1i}	0.365	0.422

Table 3 showed that the positive higher correlation between x_1 and x_2 ($\rho_{x1x2} = 0.2$) gave higher biases and mean square errors compared to the results for $\rho_{x1x2} = -0.2$ and $\rho_{x1x2} = 0$. When the nonresponse rate increases, nonresponse adjustment using the weighted mean using both x_{1i} and x_{2i} works very well and lead to declining nonresponse bias.

Table 4. Simulation results for n = 2,500 and $\rho_{x1x2} = -0.2$.

r	ρ_{yx_2}	ρ_{yx_1}	β_2	β_1	γ_2	γ_1	Estimator	Bias	MSE
0.76	0.6	0.6	2	2	2	0.1	1.Unweighted mean	0.286	0.085
							2.Weighted mean using x_{1i} and x_{2i}	0.223	0.053
							3.Weighted mean using x_{1i}	0.243	0.063
0.74						1	1.Unweighted mean	0.578	0.338
							2.Weighted mean using x_{1i} and x_{2i}	0.080	0.010
							3.Weighted mean using x_{1i}	0.147	0.027
0.70						2	1.Unweighted mean	0.835	0.700
							2.Weighted mean using x_{1i} and x_{2i}	0.008	0.027
							3.Weighted mean using x_{1i}	0.049	0.038
0.76	0.28	0.85		4	2	0.1	1.Unweighted mean	0.273	0.084
							2.Weighted mean using x_{1i} and x_{2i}	0.165	0.035
							3.Weighted mean using x_{1i}	0.175	0.042
0.74						1	1.Unweighted mean	0.897	0.815
							2.Weighted mean using x_{1i} and x_{2i}	0.013	0.009
							3.Weighted mean using x_{1i}	0.179	0.059
0.70						2	1.Unweighted mean	1.461	2.144
							2.Weighted mean using x_{1i} and x_{2i}	0.019	0.083
							3.Weighted mean using x_{1i}	0.894	1.075
0.76	0.85	0.28	4	2	2	0.1	1.Unweighted mean	0.583	0.349
							2.Weighted mean using x_{1i} and x_{2i}	0.502	0.262
							3.Weighted mean using x_{1i}	0.552	0.314
0.74						1	1.Unweighted mean	0.835	0.707
							2.Weighted mean using x_{1i} and x_{2i}	0.224	0.058
							3.Weighted mean using x_{1i}	0.619	0.393
0.70						2	1.Unweighted mean	1.040	1.093
							2.Weighted mean using x_{1i} and x_{2i}	0.001	0.047
							3.Weighted mean using x_{1i}	1.037	1.123
0.76	0.62	0.62		4	2	0.1	1.Unweighted mean	0.570	0.339
							2.Weighted mean using x_{1i} and x_{2i}	0.444	0.217
							3.Weighted mean using x_{1i}	0.484	0.250
0.74						1	1.Unweighted mean	1.154	1.346
							2.Weighted mean using x_{1i} and x_{2i}	0.157	0.037
							3.Weighted mean using x_{1i}	0.292	0.107
0.70						2	1.Unweighted mean	1.667	2.792
							2.Weighted mean using x_{1i} and x_{2i}	0.012	0.107
							3.Weighted mean using x_{1i}	0.093	0.144

Similar results were found for n = 2500 in Table 4, Table 5 and Table 6. We can see that when x_{1i} and x_{2i} or only x_{1i} are included in the model, the biases and mean square errors are reduced using the weighted mean. The unweighted mean had more biases and mean square errors than the other estimators. Increasing nonresponse rates and declining bias and mean square error by using the weighted mean using x_{1i} and x_{2i} in adjusting for nonresponse for estimating the response variable outperformed the unweighted mean that was affected by nonresponse for all levels of correlations between y and x_2 and y and x_1.

Table 5. Simulation results for $n = 2{,}500$ and $\rho_{x1x2} = 0$.

r	ρ_{yx_2}	ρ_{yx_1}	β_2	β_1	γ_2	γ_1	Estimator	Bias	MSE
0.76	0.67	0.67	2	2	2	0.1	1.Unweighted mean	0.357	0.132
							2.Weighted mean using x_{1i} and x_{2i}	0.280	0.083
							3.Weighted mean using x_{1i}	0.307	0.099
0.73						1	1.Unweighted mean	0.716	0.518
							2.Weighted mean using x_{1i} and x_{2i}	0.112	0.017
							3.Weighted mean using x_{1i}	0.203	0.047
0.69						2	1.Unweighted mean	1.021	1.048
							2.Weighted mean using x_{1i} and x_{2i}	0.019	0.028
							3.Weighted mean using x_{1i}	0.123	0.047
0.76	0.44	0.87		4	2	0.1	1.Unweighted mean	0.400	0.171
							2.Weighted mean using x_{1i} and x_{2i}	0.273	0.085
							3.Weighted mean using x_{1i}	0.299	0.102
0.73						1	1.Unweighted mean	1.059	1.133
							2.Weighted mean using x_{1i} and x_{2i}	0.039	0.012
							3.Weighted mean using x_{1i}	0.039	0.024
0.69						2	1.Unweighted mean	1.709	2.934
							2.Weighted mean using x_{1i} and x_{2i}	0.026	0.082
							3.Weighted mean using x_{1i}	0.494	0.425
0.76	0.87	0.44	4	2	2	0.1	1.Unweighted mean	0.675	0.467
							2.Weighted mean using x_{1i} and x_{2i}	0.570	0.339
							3.Weighted mean using x_{1i}	0.626	0.403
0.73						1	1.Unweighted mean	1.050	1.114
							2.Weighted mean using x_{1i} and x_{2i}	0.262	0.079
							3.Weighted mean using x_{1i}	0.616	0.392
0.69						2	1.Unweighted mean	1.356	1.852
							2.Weighted mean using x_{1i} and x_{2i}	0.032	0.054
							3.Weighted mean using x_{1i}	0.863	0.789
0.76	0.7	0.7		4	2	0.1	1.Unweighted mean	0.717	0.532
							2.Weighted mean using x_{1i} and x_{2i}	0.563	0.336
							3.Weighted mean using x_{1i}	0.618	0.401
0.73						1	1.Unweighted mean	1.434	2.076
							2.Weighted mean using x_{1i} and x_{2i}	0.227	0.068
							3.Weighted mean using x_{1i}	0.408	0.190
0.69						2	1.Unweighted mean	2.044	4.199
							2.Weighted mean using x_{1i} and x_{2i}	0.039	0.115
							3.Weighted mean using x_{1i}	0.246	0.181

Table 6. Simulation results for $n = 2{,}500$ and $\rho_{x1x2} = 0.2$.

r	ρ_{yx_2}	ρ_{yx_1}	β_2	β_1	γ_2	γ_1	Estimator	Bias	MSE
0.76	0.74	0.74	2	2	2	0.1	1.Unweighted mean	0.428	0.189
							2.Weighted mean using x_{1i} and x_{2i}	0.335	0.117
							3.Weighted mean using x_{1i}	0.366	0.140
0.72						1	1.Unweighted mean	0.847	0.723
							2.Weighted mean using x_{1i} and x_{2i}	0.139	0.025
							3.Weighted mean using x_{1i}	0.258	0.073
0.68						2	1.Unweighted mean	1.201	1.450
							2.Weighted mean using x_{1i} and x_{2i}	0.027	0.039
							3.Weighted mean using x_{1i}	0.192	0.064
0.76	0.6	0.9		4	2	0.1	1.Unweighted mean	0.540	0.304
							2.Weighted mean using x_{1i} and x_{2i}	0.389	0.164
							3.Weighted mean using x_{1i}	0.426	0.197
0.72						1	1.Unweighted mean	1.304	1.715
							2.Weighted mean using x_{1i} and x_{2i}	0.147	0.034
							3.Weighted mean using x_{1i}	0.176	0.050
0.68						2	1.Unweighted mean	1.958	3.850
							2.Weighted mean using x_{1i} and x_{2i}	0.046	0.111
							3.Weighted mean using x_{1i}	0.132	0.133
0.76	0.9	0.6	4	2	2	0.1	1.Unweighted mean	0.747	0.572
							2.Weighted mean using x_{1i} and x_{2i}	0.614	0.392
							3.Weighted mean using x_{1i}	0.675	0.471
0.72						1	1.Unweighted mean	1.238	1.546
							2.Weighted mean using x_{1i} and x_{2i}	0.273	0.087
							3.Weighted mean using x_{1i}	0.599	0.373
0.68						2	1.Unweighted mean	1.648	2.732
							2.Weighted mean using x_{1i} and x_{2i}	0.038	0.072
							3.Weighted mean using x_{1i}	0.711	0.543
0.76	0.76	0.77		4	2	0.1	1.Unweighted mean	0.859	0.759
							2.Weighted mean using x_{1i} and x_{2i}	0.669	0.471
							3.Weighted mean using x_{1i}	0.735	0.565
0.72						1	1.Unweighted mean	1.695	2.898
							2.Weighted mean using x_{1i} and x_{2i}	0.280	0.099
							3.Weighted mean using x_{1i}	0.518	0.293
0.68						2	1.Unweighted mean	2.405	5.810
							2.Weighted mean using x_{1i} and x_{2i}	0.056	0.157
							3.Weighted mean using x_{1i}	0.386	0.256

5 Conclusions

Dealing with nonresponse is imperative in sample survey analysis as fewer responses allow space for increasing nonresponse bias which affects the estimated bias. When revision of the survey design cannot yield full responses, adjustment of nonresponse can tackle the issue using nonresponse weight to deter increasing bias. Nonresponse adjustment using the weighting method is considered in this study. We consider two possible models in which missingness in the ancillary database may be correlated with missingness in the survey when nonresponse occurs on both the study and the auxiliary variables focusing on two auxiliary variables in the possible models where one auxiliary variable is fully observed, and some part of the other is missing. These models were studied as potential effects on reducing bias after receiving survey results were of interest. The results showed that the weighted mean using nonresponse adjustment by propensity score weights based on logistic regression of R_i on x_i performed the best in terms of removing the bias and also minimum mean square error when compared to the unweighted mean. The unweighted mean gave poorly biased estimates due to nonresponse especially when the nonresponse rate is high.

We can see that considering the connection between missingness in the auxiliary variable and the missingness in the survey in this study can benefit in reducing nonresponse bias and mean square error for estimating population mean using the weight. In future work, other propensity score weights may be considered use in creating the weighted in order to adjust for nonresponse.

Acknowledgement:
We would like to thank all the unknown referees for their valuable comments on the manuscript.

References:

[1] T. Chang, P. S. Kott, Using calibration weighting to adjust for nonresponse under a plausible model, *Biometrika*, Vol. 95, No. 3, 2008, pp. 555-571. https://doi.org/10.1093/biomet/asn022.

[2] P.S. Kott, Using calibration weighting to adjust for nonresponse and coverage errors, *Survey Methodology,* Vol. 32, No. 2, 2006, pp.133-42.

[3] N. Shlomo, T. Krenzke, J. Li, Comparison of three post-tabular confidentiality approaches for survey weighted frequency tables, *Transactions on Data Privacy,* Vol. 12, No. 3, 2019, pp. 145-168, [Online]. http://www.tdp.cat/issues16/abs.a329a18.php (Accessed Date: November 20, 2024).

[4] S. Lundstrom, C. E. Sarndal, Calibration as a standard method for the treatment of nonresponse, *Journal of Official Statistics*, Vol. 15, No. 2, 1999, pp. 305-327.

[5] N. Lawson, C. J. Skinner, Estimation of a cluster-level regression model under nonresponse within clusters, *Metron*, Vol. 75, 2017, pp. 319-331. https://doi.org/10.1007/s40300-017-0120-4.

[6] N. Nangsue, *Adjusting for nonresponse in the analysis and estimation of sample survey data for cluster designs*, Thesis, University of Southampton, UK, 2014.

[7] C. Ponkaew, N. Lawson, New product estimators for population mean under unequal probability sampling with missing data: a case study on the number of new COVID-19 patients. *Thailand Statistician*, Vol. 22, No. 3, 2024, pp. 634–656.

[8] F. Kreuter, K. Olson, Multiple auxiliary variables in nonresponse adjustment, *Sociological Methods & Research*, Vol. 40, No. 2, 2011, pp. 311-332. https://doi.org/10.1177/0049124111400042.

[9] R. J. Little, Survey nonresponse adjustments for estimates of means, *International Statistical Review,* Vol. 54, No. 2, 1986, pp. 139-57. https://doi.org/10.2307/1403140.

[10] R. J. Little, S. Vartivarian, On weighting the rates in non-response weights, *Statistics in Medicine,* Vol. 22, No. 9, 2003, pp. 1589-599. https://doi.org/10.1002/sim.1513.

[11] R. J. Little, S. Vartivarian, Does weighting for nonresponse increase the variance of survey means, *Survey Methodology,* Vol. 31, No. 2, 2005, pp. 161-68.

[12] N. Nangsue, Y. G. Berger, Optimal regression estimator for stratified two-stage sampling, In: F., Mecatti, P., Conti, M., Ranalli, (eds) Contributions to Sampling Statistics. *Contributions to Statistics,* Springer, Cham, 2014. https://doi.org/10.1007/978-3-319-05320-2_11.

[13] R. Groves, M. Couper, *Nonresponse in Household Interview Surveys*. New York: John Wiley, 1998.

[14] R. Groves, Nonresponse rates and nonresponse bias in household surveys, *Public Opinion Quarterly*, Vol. 70, No. 5, 2006, pp. 646-675. https://doi.org/10.1093/poq/nfl033.

[15] G. Kalton, I. Flores-Cervantes, Weighting methods, *Journal of Official Statistics,* Vol. 19, No. 2, 2003, pp. 81-97.

[16] G. Kalton, D. Maligalig, A comparison of methods of weighting adjustment for nonresponse, 1991, pp. 409-28 in *Proceedings of the 1991 Annual Re- search Conference*. Washington, DC: U.S. Bureau of the Census.

[17] B. T. West, R. J. A. Little, Non-response adjustment of survey estimates based on auxiliary variables subject to error, *Journal of the Royal Statistical Society Series C: Applied Statistics*, Vol. 62, No. 2, 2013, pp. 213–231. https://doi.org/10.1111/j.1467-9876.2012.01058.x.

[18] J. M. Brick, Unit nonresponse and weighting adjustments: a critical review, *Journal of Official Statistics*, Vol. 29, No. 3, 2013, pp. 329–353. https://doi.org/10.2478/jos-2013-0026.

[19] D. Haziza, E. Lesage, A discussion of weighting procedures for unit nonresponse, *Journal of Official Statistics*, Vol. 32, No. 1, 2016, pp. 129–145. https://doi.org/10.1515/jos-2016-0006.

[20] R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2021. https://www.R-project.org/ (Accessed Date: November 5, 2024).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

- Chris Skinner was responsible for the research planning and execution, and writing the manuscript.
- Nuanpan Lawson carried out the simulation studies, was responsible for the statistics, writing, review and editing of the manuscript.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

This research was funded by the National Science, Research and Innovation Fund (NSRF), and King Mongkut's University of Technology North Bangkok Contract no. KMUTNB-FF-68-B-23.

Conflict of Interest

The authors have no conflicts of interest to declare.