Turkish Journal of Computer and Mathematics Education *Vol.12 No.3(2021), 4077-4088*

Research Article

Statistical Evaluation of Item Nonresponse Methods Using the World Bank's 2015 Philippines Enterprise Survey

Madeline D. Cabauatan[a], Laurence P. Usona[b]

[a] Consultant, Asian Development Bank
[b]Professor, Polytechnic University of the Philippines
Email:[a]dumaua@gmail.com,[b]laurence_usona@yahoo.com

Article History: Received: 10 November 2020; Revised 12 January 2021 Accepted: 27 January 2021; Published online: 5 April 2021

Abstract: The main objective of the study was to evaluate item nonresponse procedures through a simulation study of different nonresponse levels or missing rates. A simulation study was used to explore how each of the response rates performs under a variety of circumstances. It also investigated the performance of procedures suggested for item nonresponse under various conditions and variable trends. The imputation methods considered were the cell mean imputation, random hotdeck, nearest neighbor, and simple regression. These variables are some of the major indicators for measuring productive labor and decent work in the country. For the purpose of this study, the researcher is interested in evaluating methods for imputing missing data for the number of workers and total cost of labor per establishment from the World Bank's 2015 Enterprise Survey for the Philippines.

The performances of the imputation techniques for item nonresponse were evaluated in terms of bias and coefficient of variation for accuracy and precision. Based on the results, the cell-mean imputation was seen to be most appropriate for imputing missing values for the total number of workers and total cost of labor per establishment. Since the study was limited to the variables cited, it is recommended to explore other labor indicators. Moreover, exploring choice of other clustering groups is highly recommended as clustering groups have great effect in the resulting estimates of imputation estimation. It is also recommended to explore other imputation techniques like multiple regression and other parametric models for nonresponse such as the Bayes estimation method. For regression based imputation, since the study is limited only in using the cluster groupings estimation, it is highly recommended to use other possible variables that might be related to the variable of interest to verify the results of this study.

Keywords: cell mean imputation; imputation; random hotdeck imputation; item nonresponse; missingness; nonresponse rates; nearest-neighbor; single regression imputation

1. Introduction

One major challenge of conducting surveys is that of having nonresponse. It has been proven repeatedly that nonresponse can have large effects on the results of survey. Nonresponse, interchangeably termed as missing or incomplete data, is a common occurrence in surveys, even if great care is taken before and during the data collection. Missing data, either unit or item, creates potential for bias in estimates derived from survey data (Lohr, 2010).

This study aimed at evaluating item nonresponse procedures through a simulation study of different nonresponse levels or missing rates using the World Bank's 2015 Philippines Enterprise Survey. A simulation study was conducted to explore how each of the response rates perform under a variety of circumstances. Also, the performance of procedures suggested for item nonresponse has been investigated under various conditions and variable trends from the survey.

2.Methodology:

2.1.Sources of Data

The study was conducted to compare imputation methods that would best conform for both discrete and continuous type of variables. For the purpose of this study, the survey data of the World Bank's 2015 Philippines Enterprise Survey was used. The data are not publicly available, therefore, one has to apply for access to the World Bank.

2.2.Statistical Treatment of Data

The data were examined and analyzed using the statistical software R. The researcher employed the following statistical processes and procedures to attain the objectives of the study:

1. Created a database file using R and MS Excel;
2. Computed for the characteristics of the selected discrete and continuous variables such as means and variances;
3. Performed simulation under different levels of nonresponse using the Bootsrap resampling method;
4. Evaluated and compared the characteristics of estimates for the different nonresponse rates from the pseudo-population estimates; and
5. Imputed missing values using selected procedures for item nonresponse;
6. Evaluated the procedures by comparing estimates using Bias and Variances.

2.2.1.Selection of Variables

For trend of variables, the following important indicators from the Enterprise Survey both for discrete and continuous type were used in the study:

Discrete: total number of workers per establishment; and

Continuous: average cost of labor per establishment.

2.2.2.Characteristics of the Pseudo-Population

The full sample data for the variables on total number of workers and average cost of labor per establishment was treated as the pseudo-population. Hence, evaluation and description of the characteristics of the population were done in terms of means and variances.

2.2.3.Simulation Using Different Levels of Nonresponse

The simulation experiments were done to evaluate the procedures across the different percentage of nonresponses: 5%, 10%, and 20%:

1. Given the database of all responding sampling units of the Enterprise Survey for selected variables, a sample without replacement 1,000 times was drawn. Bootstrap, one of the popular resampling methods discussed in the book of Lohr (2010), was used by simply drawing the sample using a simple random sampling without replacement of size n, which will reproduce properties of the whole population.
2. To simulate nonresponses, the values of the selected variables from the database equal to the level of nonresponses: 5%, 10%, and 20% were dropped at random.
3. Using the database with values of the variables dropped in some portions of the database in step 2, the statistics of interest of the variables for each of the samples using the different methods for item nonresponse were calculated.
4. The characteristics of the estimates for the different nonresponse rates with that of the pseudo-population estimates were finally compared.

The following options were explored for both labor cost and number of workers during simulation:

- Sample sizes – 100 and 200
- Nonresponse rates – 5%, 10%, and 20%
- Imputation methods – cell mean, nearest neighbor, random hotdeck and regression
- Classes – size, region, sector, size-region, size-sector, region sector, and size-region-sector

3.Evaluation of Methods

The imputation methods used are the cell mean imputation, random hotdeck imputation, nearest neighbor imputation, and regression imputation. For the purpose of this study, the following methods for imputing missing data for the total number of workers and average cost of labor per establishment were evaluated.

3.1Cell Mean Imputation

This method assumes that missing values within the cells are missing completely at random. First, classes or cells are used to group the respondents according to known variables. The average of all responding establishments in a class or a cell is used to replace for each missing data.

3.2.Random HotDeck Imputation

A donor is randomly chosen from the establishment in the cell with information on all missing items. To preserve multivariate relationships, usually values from the same donor are used for all missing items of an establishment.

3.3.Nearest-Neighbor HotDeck Imputation

This method works by defining a distance measure based on one or more clustering variables then by imputing the missing value of a unit using the non-missing value of a unit nearest to it (the nearest neighbor) based on the distance measure.

3.4.Regression Imputation

Regression imputation predicts the missing value by using a regression of the item of interest on variables observed for all cases. A variation is stochastic regression imputation, in which the missing value is replaced by the predicted value from the regression model, plus a randomly generated error term.

The following are regression models used for regression imputation:

1. Ordinary Least Squares regression model for Labor.Cost

$$\log(Labor.Cost) = \beta_0 + \underline{\beta}' \underline{X} + \epsilon$$

2. Negative Binomial Generalized Linear Model for Number.of.Workers

$$\log(Number.of.Workers) = \beta_0 + \underline{\beta}' \underline{X}$$

$$E(Number.of.Workers) = \mu$$

$$Var(Number.of.Workers) = \mu + \frac{\mu^2}{dispersion\ parameter}$$

Where:

$\underline{X}$ = vector of 1's and 0's as indicator of the clustering variables

$\underline{\beta}$ = vector of corresponding coefficients

β_0 = intercept term

$\epsilon \sim Normal(0, \sigma^2)$ = normal error term for OLS model

The natural logarithm of labor cost was used since it is highly skewed and do not scale linearly. The Negative Binomial generalized linear model was used for number of workers since the variable is of discrete type and exhibits overdispersion (variance = 74483.18 is so much larger than the mean = 111.2349) which violates the characteristic of the Poisson distribution where the mean and variance are equal.

3.5.Comparison of the Estimates/Assessment of the Performance of the Techniques

The estimates to be obtained from the methods will be compared using a set of criteria for selecting a better procedure to compensate missing data for the variables on total number of workers and average cost of labor per establishment. The criteria to be used in assessing the estimates include measures of accuracy and precision.

To mitigate the effects of sampling error, 1000 simulated simple random sampling without replacement of size samples were used to obtain the expected accuracy (average percent bias and average absolute percent bias) and precision (average CV of sample mean) of the sample mean per scenario. Furthermore, comparability was ensured by using the same set of 1000 simulated samples per scenario. A value for the bias that is near zero indicates better estimator. Estimates are said to be precise if it has a coefficient of variation below 10%.

Estimators for a Single Stage SRSWOR

$$sample\ mean: \bar{y} = \frac{1}{n_{res}} \sum_{i=1}^{n_{res}} y_i$$

$$finite\ population\ correction: fpc = 1 - \frac{n_{res}}{N}$$

$$sample\ variance: s^2 = \frac{1}{n_{res} - 1} \sum_{i=1}^{n_{res}} (y_i - \bar{y})^2$$

$$unbiased\ estimator\ of\ the\ variance\ of\ \bar{y}: \hat{V}(\bar{y}) = fpc \frac{s^2}{n_{res}}$$

$$estimated\ CV\ of\ \bar{y}\text{: } \widehat{CV}(\bar{y}) = \frac{\sqrt{\hat{V}(\bar{y})}}{\bar{y}}$$

$$average\ \widehat{CV}(\bar{y})\ of\ B\ simulated\ samples\text{: } AVE\ \widehat{CV}(\bar{y}) = \frac{1}{B}\sum_{b=1}^{B} \widehat{CV}(\bar{y}_b)$$

$$percent\ bias\text{: } PBIAS = 100\% \times \frac{(\bar{y} - \bar{y}_U)}{\bar{y}_U}$$

$$absolute\ percent\ bias\text{: } ABS\_PBIAS = 100\% \times \frac{|\bar{y} - \bar{y}_U|}{\bar{y}_U}$$

$$average\ PBIAS\ of\ B\ simulated\ samples\text{: } \overline{PBIAS} = \frac{1}{B}\sum_{b=1}^{B} PBIAS_b$$

$$average\ ABS\_PBIAS\ of\ B\ simulated\ samples\text{: } \overline{ABS\_PBIAS} = \frac{1}{B}\sum_{b=1}^{B} ABS\_PBIAS_b$$

Where
n_{res} = number of responding units
$N = 1141$ = number of units in the pseudopopulation
$B = 1000$ = number of simulated samples
y_i = value of the i[th] sampled units (Labor.Cost or Number.of.workers)
$\bar{y}_U$ = true mean

4.Result

4.1.Simulation of Samples and Clusters

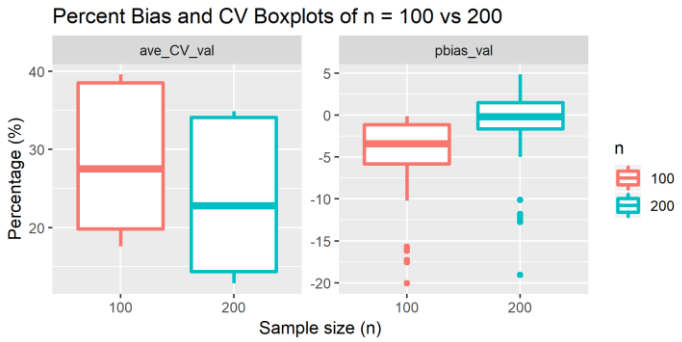
For both number of workers and labor cost, sample sizes of n=100 and n=200 were generated during simulation using the Bootstrap method of resampling. Expected accuracy (average percent bias and average absolute percent bias) and precision (average Coefficient of Variation or CV of sample mean) of the sample mean per scenario were obtained. For the number of workers and labor cost, sample size at n=200 shows more accurate and precise estimates than at sample size of n=100 (Table 1). Accuracy and precision of the simulated samples can also be visualize in the boxplot presented in Figure 1.

Table 1. Characteristics of Simulated Samples (n=100 and n=200)

Variable	n	$\bar{y}$	$\bar{y}_U$	$\overline{ABS\_PBIAS}$	$AVE\ \widehat{CV}(\bar{y})$
Labor.Cost (Php '000)	100	69,785.3	71,615.3	2.6	39.6
Labor.Cost (Php '000)	200	73,463.9	71,615.3	2.6	34.9
Number.of.workers	100	110.4	111.2	0.8	20.0
Number.of.workers	200	111.1	111.2	0.1	14.5

Source: Author's calculations

Figures 1. Boxplot of the Simulated Samples (Accuracy and Precision)



To decide which clustering group to use in the analysis, a comparison of the different combinations of clusters was performed in terms of accuracy and precision. There are 7 possible clusters: (i) size, (ii) region, (iii) sector, (iv) size-region, (v) size-sector, (vi) region sector, and (vii) size-region-sector.

Figure 2. Boxplots of CVs for Clusters of the Number of Workers

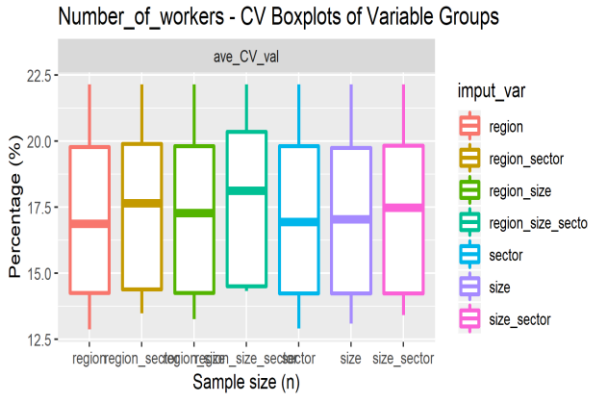
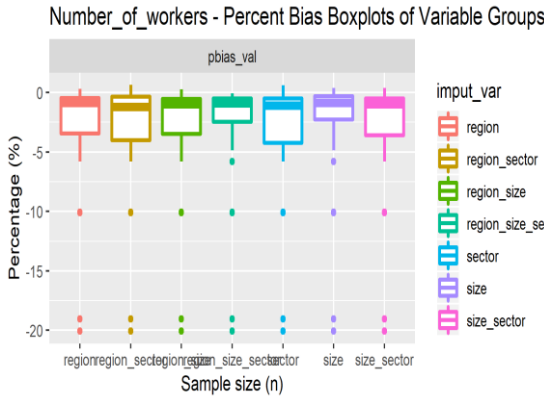


Figure 3. Boxplots of Percent Bias for Clusters of the Number of Workers



Figures 2 and 3 show that single clusters such as size, region, and sector have performed better in terms of accuracy and precision exhibiting lower CVs and Bias closer to zero for the number of workers than the pairwise combination of the stratification variables. In terms of precision, the combination of three cluster group size-region-sector also showed a value of bias that is near zero. However, having a narrowed down groupings can affect the source of donor values from clusters during imputation. Missing values may not be filled if its cluster did not match any donor values of non-missing values in the sample. Hence, for a given sample, clustering based on all grouping variables will least likely impute all missing values. The same results were generated for labor cost (Figure 4 and Figure 5).

Figure 4. Boxplots of CVs for Clusters of the Labor Cost

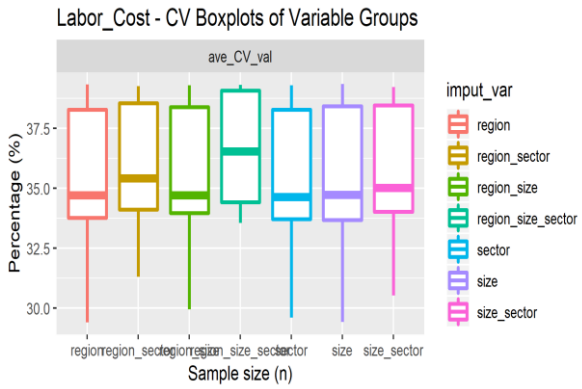
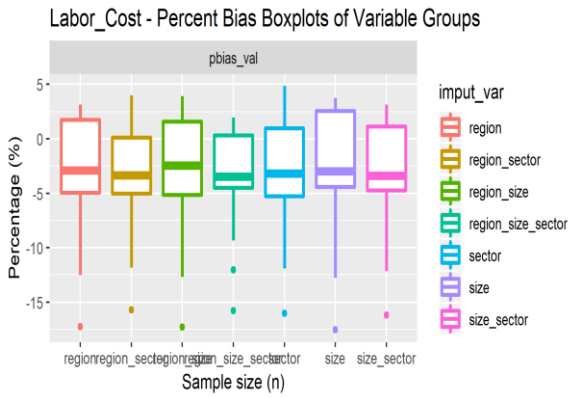


Figure 5. Boxplots of Percent Bias for Clusters of the Labor Cost



4.2.Levels of Nonresponses

The following options for nonresponse rates were explored for both number of workers and labor cost during simulation: 5%, 10%, and 20%. To artificially create nonresponses from the sample of size n=200, dropped at random the values of the selected variables equal to the nonresponse levels.

Table 2. Characteristics of the Number of Workers at Different Nonresponse Rates

Nonresponse Rate	n	$\bar{y}$	$\bar{y}_U$	$\overline{PBIAS}$	fpc	$\hat{V}(\bar{y})$	$\hat{V}(\bar{y}_U)$
0%	200	111.2	111.2	0.04%	0.8	302.5	307.1
5%	190	110.9	111.2	0.33%	0.8	320.2	326.7
10%	180	112.2	111.2	0.84%	0.8	355.6	348.5
20%	160	111.8	111.2	0.55%	0.9	406.9	400.2

Source: Author's calculations

Table 3. Characteristics of Labor Cost (Php '000) at Different Nonresponse Rates

Nonresponse Rate	n	$\bar{y}$	$\bar{y}_U$	$\overline{PBIAS}$	fpc	$\hat{V}(\bar{y})$	$\hat{V}(\bar{y}_U)$
0%	200	72,806.5	71,615.3	1.66%	0.8	1,199,875,054,268.6	1,170,256,532,299.9
5%	190	71,097.1	71,615.3	1.72%	0.8	1,210,585,012,732.2	1,244,939,831,329.7
10%	180	73,034.2	71,615.3	1.98%	0.8	1,435,057,926,518.3	1,327,921,274,696.2
20%	160	69,987.8	71,615.3	2.27%	0.9	1,434,915,602,020.8	1,525,002,202,691.6

Source: Author's calculations

A simple analysis of the newly created databases for the number of workers and labor cost with missing values at 5%, 10%, and 20% reveals that when the level of missing items increases, the estimates become less accurate and less precise (Tables 2 and 3).

4.3.Evaluation of Imputation Methods

Results on the evaluation of imputation methods for the "number of workers" using the simulated samples of "n=200" and cluster group "size" showed that cell mean and regression imputation methods have the same performance in terms of precision at 5%, 10%, and 20% nonresponse. The three methods of imputation - random hotdeck, cell mean, and regression techniques have the same performances in terms of the accuracy of estimates at 5% level of nonresponses. While cell mean imputation outperformed the other methods in terms of accuracy at 10% nonreponse rate, the cell mean and regression gave accurate estimates at 20% missingness (Table 4).

Table 4. Evaluation of Imputation Methods for the Number of Workers Using the Simulated Samples (n=200) and Cluster Group (Size) Under Different Nonresponse Rates

Nonresponse Rate	Number of Workers, Sample n=200, Cluster: Size				
	Method	$\bar{y}$	$\bar{y}_U$	$\overline{ABS\_PBIAS}$	$AVE\ \widehat{CV}(\bar{y})$
5%	random.hot.deck	111.3	111.2	0.0	14.4
5%	cell.mean	111.2	111.2	0.0	14.1
5%	nearest.neighbor	111.2	111.2	0.1	14.4
5%	regression	111.2	111.2	0.0	14.1
10%	random.hot.deck	110.6	111.2	0.6	14.2
10%	cell.mean	110.8	111.2	0.4	13.7
10%	nearest.neighbor	110.6	111.2	0.6	14.2
10%	regression	110.8	111.2	0.4	13.7
20%	random.hot.deck	111.5	111.2	0.2	14.2
20%	cell.mean	111.6	111.2	0.4	13.1
20%	nearest.neighbor	111.7	111.2	0.4	14.2
20%	regression	111.6	111.2	0.4	13.1

Source: Author's calculations

Table 5. Evaluation of Imputation Methods for the Number of Workers Using the Simulated Samples (n=200) and Cluster Group (Region) Under Different Nonresponse Rates

Nonresponse Rate	Number of Workers, Sample n=200, Cluster: Region				
	Method	$\bar{y}$	$\bar{y}_U$	$\overline{ABS\_PBIAS}$	$AVE\ \widehat{CV}(\bar{y})$
5%	random.hot.deck	110.9	111.2	0.3	14.4
5%	cell.mean	111.2	111.2	0.1	14.1
5%	nearest.neighbor	110.9	111.2	0.3	14.4
5%	regression	111.2	111.2	0.1	14.1
10%	random.hot.deck	110.7	111.2	0.5	14.2
10%	cell.mean	110.7	111.2	0.5	13.6
10%	nearest.neighbor	110.4	111.2	0.8	14.3
10%	regression	110.7	111.2	0.5	13.6
20%	random.hot.deck	111.6	111.2	0.3	14.2
20%	cell.mean	111.6	111.2	0.3	12.9
20%	nearest.neighbor	110.7	111.2	0.5	14.2
20%	regression	111.6	111.2	0.3	12.9

Source: Author's calculations

Table 5 shows an evaluation of the imputation methods for the "number of workers" using the simulated samples of "n=200" and cluster group "region". As displayed in the table above, the cell mean and regression imputation methods have the same performance in terms of accuracy and precision of estimates at 5%, 10%, and 20% nonresponse.

Table 6 shows an evaluation of the imputation methods for the "number of workers" using the simulated samples of "n=200" and cluster group "sector". As displayed in Table 6, the cell mean and regression imputation methods have the same performance in terms of precision at 5%, 10%, and 20% nonresponses. In terms of the accuracy of the estimates, cell mean and regression performed best among all methods at 5 and 10% nonresponses while random hotdeck performed best at 20% missing rate.

Table 6. Evaluation of Imputation Methods for the Number of Workers Using the Simulated Samples (n=200) and Cluster Group (Sector) Under Different Nonresponse Rates

Nonresponse Rate	Number of Workers, Sample n=200, Cluster: Sector				
	Method	$\bar{y}$	$\bar{y}_U$	$\overline{ABS\_PBIAS}$	$AVE\ \widehat{CV}(\bar{y})$
5%	random.hot.deck	111.0	111.2	0.2	14.4
5%	cell.mean	111.2	111.2	0.0	14.1
5%	nearest.neighbor	110.5	111.2	0.7	14.4
5%	regression	111.2	111.2	0.0	14.1
10%	random.hot.deck	110.7	1 11.2	0.5	14.2

10%	cell.mean	110.8	111.2	0.4	13.6
10%	nearest.neighbor	109.7	111.2	1.4	14.2
10%	regression	110.8	111.2	0.4	13.6
20%	random.hot.deck	111.6	111.2	0.3	14.2
20%	cell.mean	111.6	111.2	0.4	12.9
20%	nearest.neighbor	108.8	111.2	2.2	14.2
20%	regression	111.6	111.2	0.4	12.9

Source: Author's calculations

On the other hand, it is shown that random hotdeck imputation method is the best performer in terms of accuracy of estimates only at 10% and 20% nonresponses, while regression perfomed best at 5% missing rate for "labor cost (Php '000)" using the simulated samples of "n=200" and cluster group "size". In terms of the precision of estimates at 5% level of nonresponses, the cell mean imputation outperformed the other methods all levels of nonreponse rates (Table 7).

Table 7. Evaluation of Imputation Methods for the Labor Cost (Php '000) Using the Simulated Samples (n=200) and Cluster Group (Size) Under Different Nonresponse Rates

	Labor Cost (Php '000), Sample n=200, Cluster: Size				
Nonresponse Rate	Method	$\bar{y}$	$\bar{y}_U$	$\overline{ABS\_PBIAS}$	AVE $\widehat{CV}(\bar{y})$
5%	random.hot.deck	73,427.3	71,615.3	2.5	34.1
5%	cell.mean	73,240.2	71,615.3	2.3	33.3
5%	nearest.neighbor	73,445.8	71,615.3	2.6	34.1
5%	regression	70,133.8	71,615.3	2.1	34.8
10%	random.hot.deck	73,571.3	71,615.3	2.7	33.8
10%	cell.mean	73,695.7	71,615.3	2.9	32.2
10%	nearest.neighbor	73,952.9	71,615.3	3.3	33.9
10%	regression	67,387.6	71,615.3	5.9	35.1
20%	random.hot.deck	73,362.5	71,615.3	2.4	32.7
20%	cell.mean	73,915.7	71,615.3	3.2	29.4
20%	nearest.neighbor	73,777.2	71,615.3	3.0	32.7
20%	regression	61,320.7	71,615.3	14.4	35.3

Source: Author's calculations

Table 8 shows that cell mean imputation method outperformed all other techniques in terms of precision at all levels of nonresponses for the "labor cost (Php '000)" using the simulated samples of "n=200" and cluster group "region". In terms of the accuracy of the estimates, nearest neighbor imputation technique performed best among all methods at 5%, 10%, 20% levels of nonresponses.

Table 8. Evaluation of Imputation Methods for the Labor Cost (Php '000) Using the Simulated Samples (n=200) and Cluster Group (Region) Under Different Nonresponse Rates

	Labor Cost (Php '000), Sample n=200, Cluster: Region				
Nonresponse Rate	Method	$\bar{y}$	$\bar{y}_U$	$\overline{ABS\_PBIAS}$	AVE $\widehat{CV}(\bar{y})$
5%	random.hot.deck	73,132.3	71,615.3	2.1	34.1
5%	cell.mean	73,167.7	71,615.3	2.2	33.3
5%	nearest.neighbor	72,812.4	71,615.3	1.7	34.1
5%	regression	69,868.0	71,615.3	2.4	34.9
10%	random.hot.deck	74,576.4	71,615.3	4.1	33.7
10%	cell.mean	73,693.9	71,615.3	2.9	32.2
10%	nearest.neighbor	72,991.8	71,615.3	1.9	33.9
10%	regression	66,850.8	71,615.3	6.7	35.4
20%	random.hot.deck	73,657.4	71,615.3	2.9	32.6
20%	cell.mean	73,875.5	71,615.3	3.2	29.4
20%	nearest.neighbor	70,718.1	71,615.3	1.3	33.0
20%	regression	60,272.7	71,615.3	15.8	35.9

Source: Author's calculations

Table 9 shows that the cell mean imputation method outperformed all other techniques in terms of precision at 5%, 10%, 20% levels of nonresponses for the "labor cost (Php '000)" using the simulated samples of "n=200" and cluster group "region". In terms of the accuracy of the estimates, nearest neighbor imputation technique performed best among all methods at all levels of nonresponses.

Table 9. Evaluation of Imputation Methods for the Labor Cost (Php '000) Using the Simulated Samples (n=200) and Cluster Group (Sector) Under Different Nonresponse Rates

	Labor Cost (Php '000), Sample n=200, Cluster: Sector				
Nonresponse Rate	Method	$\bar{y}$	$\bar{y}_U$	$\overline{ABS\_PBIAS}$	AVE $\widehat{CV}(\bar{y})$
5%	random.hot.deck	73,247.7	71,615.3	2.3	34.0
5%	cell.mean	73,159.6	71,615.3	2.2	33.4
5%	nearest.neighbor	72,099.3	71,615.3	0.7	34.2
5%	regression	69,892.3	71,615.3	2.4	34.9
10%	random.hot.deck	73,726.4	71,615.3	2.9	33.8
10%	cell.mean	73,756.5	71,615.3	3.0	32.2
10%	nearest.neighbor	71,913.4	71,615.3	0.4	34.1
10%	regression	66,906.4	71,615.3	6.6	35.4
20%	random.hot.deck	73,815.1	71,615.3	3.1	32.7
20%	cell.mean	73,653.1	71,615.3	2.8	29.6
20%	nearest.neighbor	69,961.0	71,615.3	2.3	33.1
20%	regression	60,391.3	71,615.3	15.7	35.8

Source: Author's calculations

Boxplots of the performance of the four evaluation techniques were compared at sample size of "n=200" and cluster group "size" regardless of the level of nonresponses for the number of workers and labor cost.

For the number of workers, cell mean imputation outperformed all other techniques in terms of producing precise estimates, exhibiting the lowest CV among all methods. This was followed by random hotdeck and nearest neighbor imputation methods. The regression imputation have yielded the highest CV (Figure 6).

Moreover, cell mean and random hotdeck imputation methods have almost the same performance in terms of producing accurate estimates with biases almost close to zero for the number of workers. This was followed by nearest neighbor imputation imputation method. The regression imputation have yielded a more bias estimates (Figure 7).

Figure 6. Boxplots of CVs of Imputation Methods for Number of Workers

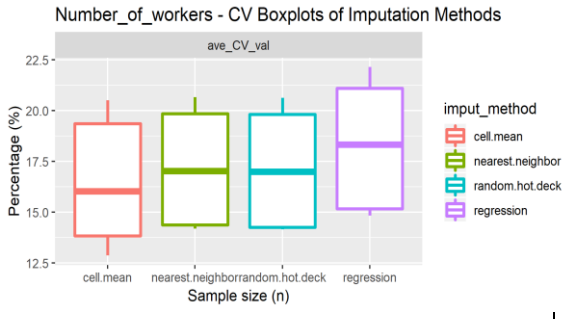
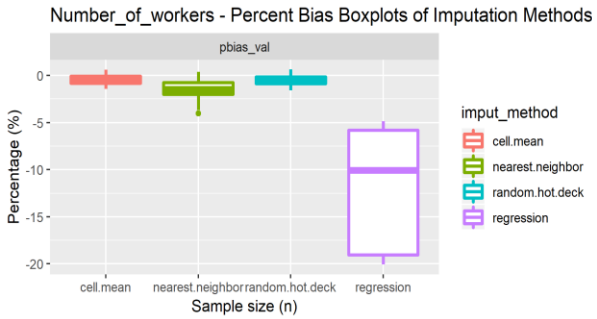


Figure 7. Boxplots of Bias of Imputation Methods for Number of Workers



For labor cost, cell mean imputation outperformed all other techniques in terms of producing precise estimates, exhibiting the lowest CV among all methods. This was followed by random hotdeck and nearest neighbor imputation methods. The regression imputation have yielded the highest CV (Figure 8).

Figure 8. Boxplots of CVs of Imputation Methods for Labor Cost

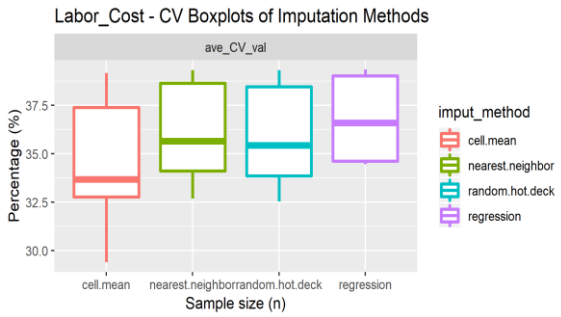
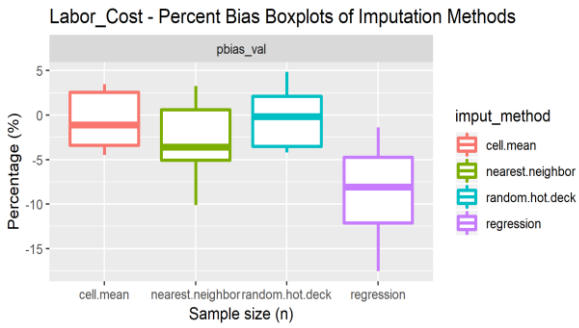


Figure 9. Boxplots of Percent Bias of Imputation Methods for Labor Cost



Moreover, cell mean and random hotdeck imputation methods have almost the same performance in terms of producing accurate estimates with biases almost close to zero for labor cost. This was followed by nearest neighbor imputation imputation method. The regression imputation have yielded a more bias estimates (Figure 9).

5.Summary of Findings, Conclusions, and Recommendations:

5.1.Summary of Findings

The findings of the study are summarized as follows:

1. Newly created databases for the number of workers and labor cost across different levels of nonresponses or missingness of 5%, 10%, and 20% reveal that means of the database with 5% missingness are closest to the true population mean (number of workers - 111 and labor cost - Php 71.615 million). The lowest percentage bias can be found at 5% missingness and the highest bias is seen at 20% missingness both for the number of workers and labor cost. The same result can be generated in terms of variances. Moreover, for the number of workers and labor cost, sample size at n=200 shows more accurate and precise estimates than at sample size of n=100. Single clusters or classes such as size, region, and sector have performed better in terms of accuracy and precision exhibiting lower CVs and Bias closer to zero for the number of workers and labor cost as compared to the pairwise combinations of the stratification variables.

2. Results on the evaluation of imputation methods for the “number of workers” using the simulated samples of “n=200” and cluster group “size” showed that cell mean and regression imputation methods have the same performance in terms of precision at 5%, 10%, and 20% nonresponse. The three methods of imputation - random hotdeck, cell mean, and regression techniques have the same performances in terms of the accuracy of estimates at 5% level of nonresponses. While cell mean imputation outperformed the other methods in terms of accuracy at 10% nonreponse rate, the cell mean and regression gave accurate estimates at 20% missingness. Evaluation of the imputation methods for the “number of workers” using the simulated samples of “n=200” and cluster group “region” showed that cell mean and regression imputation methods have the same performance in terms of accuracy and precision of estimates at 5%, 10%, and 20% nonresponse. Evaluation of the imputation methods for the “number of workers” using the simulated samples of “n=200” and cluster group “sector” showed that cell mean and regression imputation methods have the same performance in terms of precision at 5%, 10%, and 20% nonresponses. In terms of the accuracy of the estimates, cell mean and regression performed best among all methods at 5 and 10% nonresponses while random hotdeck performed best at 20% missing rate.

3. On the other hand, evaluation of imputation methods for “labor cost (Php ‘000)” using the simulated samples of “n=200” and cluster group “size” showed that random hotdeck imputation method is the best performer in terms of accuracy of estimates only at 10% and 20% nonresponses, while regression perfomed best at 5% missing rate. In terms of the precision of estimates at 5% level of nonresponses, the cell mean imputation outperformed the other methods all levels of nonreponse rates. Evaluation of the imputation methods for the “labor cost (Php ‘000)” using the simulated samples of “n=200” and cluster group “region” showed that the cell mean imputation method outperformed all other techniques in terms of precision at all levels of nonresponses. In terms of the accuracy of the estimates, nearest neighbor imputation technique performed best among all methods at 5%, 10%, 20% levels of nonresponses. Evaluation of the imputation methods for the “labor cost (Php ‘000)” using the simulated samples of “n=200” and cluster group “region” showed that the cell mean imputation method outperformed all other techniques in terms of precision at 5%, 10%, 20% levels of nonresponses. In terms of the

accuracy of the estimates, nearest neighbor imputation technique performed best among all methods at all levels of nonresponses.

4. The most appropriate imputation technique for estimating the item nonreponse for the number of workers are cell mean imputation and regression imputation at all levels of missingness (5%, 10%, 20%) and for all cluster groups (size, region, sector). For the labor cost using the clustering group size, cell mean imputation and regression imputation showed as the superior techniques for estimating item nonresponse at 5% missingness while cell mean imputation and random hotdeck imputation showed superiority at 10% and 20% missingness. For the clustering group region, the best method for estimating item nonresponse for labor cost is the nearest neighbor imputation for all levels of nonresponses in terms of the accuracy of estimates (bias). In terms of the precision of estimates (CVs), cell mean imputation is the most appropriate technique to impute for missing items at all levels of missingness. On the other hand, for the clustering group sector, the best method for estimating item nonresponse for labor cost is the nearest neighbor imputation for all levels of nonresponses in terms of the accuracy of estimates (bias). In terms of the precision of estimates (CVs), cell mean imputation is the most appropriate technique to impute for missing items at all levels of missingness.

6.Conclusions

The following conclusions were drawn based on the findings of the study:

1. The newly created databases for the number of workers and labor cost with missing values at 5%, 10%, and 20% reveals that when the level of missing items increases, the estimates become less accurate and less precise. Therefore, it would be best to treat our data using appropriate techniques when missingness occurs. Moreover, higher sample sizes provide better estimates in terms of accuracy and precision. Simple clusters or classes can be used to select the donor value for a missing item. Clustering based on all grouping variables will least likely impute all missing values.

2. Imputed estimates for the number of workers using the clustering groups of size, region, sector showed accurate and precise estimates at all levels of missingness for cell mean and regression imputation techniques.

3. For labor cost, imputed estimates using the clustering groups of region and sector showed that the cell mean imputation and nearest neighbor provided more accurate and precise estimates at all levels of nonresponses. For cluster group size, cell mean imputation and random hotdeck provided better estimates at 10% and 20% missingness while cell mean imputation and regression imputation gave more accurate and precise estimates at 5% missingness.

4. Overall, cell mean imputation method has provided the best estimates for both discrete and continuous variables (number of workers and labor cost) at different levels of nonresponses (5%, 10%, 20%) in terms of providing accurate and precise estimates for item nonresponses.

7.Recommendations

The following recommendations are offered based on the derived conclusions:

1. For regression based imputation, since study is limited only in using the cluster groupings estimation, it is highly recommended to use other possible variables that might be related to the variable of interest to verify the results of this study.

2. Explore choice of other clustering groups. Clustering groups greatly affects the resulting estimates of imputation estimation.

3. Explore multiple imputation method with different models for nonresponse, where each missing value is imputed m (>=2) different times.

Also, explore the use of other parametric models for nonresponse by fitting a superpopulation model such as the Bayes estimation method..

References

Capatoy, Reynario A. (2004). Use of Regression Analysis in Imputing Item Nonreponse in the Annual Survey of Establishments in the Real Estate Renting and Miscellaneous Business Activities Sector. Unpublished Master's Thesis, Polytechnic University of the Philippines, Sta. Mesa, Manila.

De Leeuw, Hox, Dillman (2008). International Handbook of Survey Methodology.

Del Prado, Divina Grace L. (1999). "Imputation of the Philippine Quarterly Survey of Establishments for Mining and Quarrying Sector" Unpublished Master's Thesis, Katholieke Universiteit Leuven, Belgium.

Divino, A.M. (2005). An Evaluation on Imputation Techniques of the Quarterly Survey of Philippine Business and Industry for Wholesale and Retail Trade Sector. Unpublished Master's Thesis, Polytechnic University of the Philippines, Sta. Mesa, Manila.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. Journal of the American Statistical Association, 91, 490–498.

Gay, L.R., Mills, G.E. and Airasian, P. (2009). Educational Research Competencies for Analysis and Applications. Pearson, Columbus.

Little, Roderick J.A. and Rubin Donald (1987). Statistical Analysis with Missing Data. New York: Wiley.

Little, R. J. A., and Rubin, D. B. (2002). Statistical analysis with missing data, 2nd ed. New York: Wiley.

Leeuw,E.D. de, Hox, J.J., and Huisman, M. (2003). Prevention and Treatment of Item Nonresponse. Journal of Official Statistics, Vol. 19, No.2, 2003, pp. 153-176.

Lohr, Sharon S.(1999 and 2010). Sampling Design and Analysis, 1st and 2nd Edition. New York: Wiley.

OECD Glossary of Statistical Terms (2002). Retrieved 18 June 2013, from http://stats.oecd.org/glossary/detail.asp?ID=3764.

Pauwels, L. and Svensson, R. (2008). How Serious Is The Problem Of Item Nonresponse in Scale Constructs of Delinquency and Ley Social Mechanisms? A Cross-National Inquire of two Calssroom PAPI-Self Report Studies in Antwerp and Halmstad, The European Journal of Criminology: 289-309.

Platek, R. (1977). Some factors affecting non-response. Survey Methodology, 3, 191–214.

Rao, J. N. K. (1996). On variance estimation with imputed survey data. Journal of the American Statistical Association, 91, 499–506.

Rubin, Donald B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, Inc.

Shao, J. (2003). Impact of the bootstrap on sample surveys. Statistical Science, 18, 191–198.

Stef van Buuren. Flexible Imputation of Missing Data. Chapman and Hall/CRC Interdisciplinary Statistics Series. CRC Press, Taylor and Francis Group, Boca Raton, London New York. 2012.

World Bank Enterprise Survey (2017). https://www.enterprisesurveys.org/.