Identification and Analysis of Factors Affecting Missing Values in Survey Data

Ranasinghe DAN[1]
Napagoda NADN[2]

ABSTRACT

Survey data comprises missing values which create serious problems for all the parties who involve with surveys. Incomplete data occurs if there is no observation for particular item in survey data. Therefore, it makes difficulties when analyzing data. Finally, existing of missing data misleads conclusions of survey results. In fact, inappropriate handling of the missing data in the analysis may introduce bias and can also limit the general purpose of the research findings. Therefore, reducing occurrence of missing values is most significant solution for this problem. Hence the purpose of the study becomes to identify the factors affecting missing values. It may help to collect complete data for a survey with the awareness of affecting factors. Survey data of Annual Surveys of Industries (ASI) conducted by the Department of Census and Statistics (DCS) were used for the study. Diffusion of missing values among each identified variables were comprehend and the factors affecting missing values were employed using binary logistic regression model. The results suggest that there are four factors affecting missing values; such as Establishment type, Area, Legality type and Number of total employee. This study is mostly important to handle missing values in ASI. Since ASI results provide the annual summary of the circumstance of local industries, it affect for the social and economical state of the country. Hence accuracy of the ASI results will be very significant to the development of the country.

KEYWORDS: Binary Logistic Regression, Department of Census and Statistics, Missing Values, Survey Data

INTRODUCTION

The Department of Census and Statistics (DCS) is the central government agency which is responsible for data collection, process, analysis and dissemination of statistical information and computation of statistical indicators related to the socio economic condition of the country. The department provides data required for national planning and policy formulation of the government as well as that needed for monitoring the progress in implementing such policies. Department also provides data to the needs of local and international agencies regularly on a variety of subject areas drawn from different sources, including its own census and surveys. The data collected covers subjects such as population, agriculture, trade, industry, prices and national income.

DCS has 16 head office divisions and Industry, Trade and Services Division (ITSD) is one of the divisions out of them. ITSD conducts several surveys annually and quarterly. Annual Surveys of Industries (ASI) is a survey program mainly conducted by ITSD which is covered the economic activities in the areas of mining and quarrying, manufacturing, production and distribution of electricity, gas and water. Establishments with 25 or more persons engaged were fully canvassed and a sample of establishment was selected from the other establishments with 5 or more persons engaged for the survey. ITSD frequently faces the problem of occurring missing values in ASI data. It is a major barrier to obtain accurate survey results.

Data of most surveys comprise missing values. Missing values create serious problems for all the parties involved

[1]Graduate, Department of Mathematical Sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka

[2]Senior Lecturer, Department of Mathematical Sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka

in surveys. Incomplete data occur when there are no answers for a particular question in the survey questionnaire and it makes analysis of data more challenging. Furthermore, missing values can lead to incorrect decisions. Therefore, identifying the factors affecting to occur missing values are very significant to interpret about data clearly.

RESEARCH OBJECTIVE

The main objective of the study is to identify the factors affecting missing values in ASI.

LITERATURE REVIEW

Missing data arise in almost all statistical surveys. There are various reasons for their existence, such as non-responsive, manual data entry procedures, equipment errors and incorrect measurements (Pigott, 2001). Non-responsiveness is the major issue in missing data (Wagner, 2000). Survey questionnaires frequently contain missing values because of the users refuse to answer some sensitive questions such as income level and age or they simply have no opinions about them and so on. Missing values create complex environment and make difficulties to understand about the data when analyzing. The occurrence of missing values can also create serious problems for researchers.

Identification of the patterns of missing data helps the researcher to determine whether the missing values are Missing Completely at Random (MCAR), Missing at Random (MAR) or Missing Not at Random (MNAR). Missing Completely at Random means that the probability of missingness of a variable is not related to any of the study variables. That is, the data are missing due to some totally unrelated event. This type of event occurs rarely that it is usually best to categorize the missing data as MCAR. If data are MAR, omitted data may be related to at least one other variable in the study but not to the outcome being measured. It is often difficult, however, for researchers to be certain of the relationship between missing data and these variables. Consequently, they may be unable to distinguish data that are truly MAR from data MCAR. Most frequently, data are MNAR. This means that the reason for the missingness is related to one or more of the outcome variables or that the missingness has a systematic pattern (Schafer et al., 2002).

Three general methods have been used for handling missing values in statistical analysis. One is the so-called "complete case analysis", which ignores the observations with missing values and base of the analysis on the complete case data. This approach is the default of many statistical software packages. The disadvantages of this approach are the loss of efficiency due to discarding the incomplete observations and biases in estimates when data are missing in a systematic way. The second approach for handling missing values is the imputation method, which imputes values for the missing covariates and carries out the analysis as if the imputed values were observed data. This approach may reduce the bias of the complete case analysis but may lead to additional bias in multivariate analysis if the imputation fails to control for all multivariate relationships. The third approach is to assume some models for the covariates with missing values and then use a maximum likelihood approach to obtain estimates for the models (Sujuan, 1997).

RESEARCH PROBLEM

Surveys conducted on industrial sector containing missing values. In statistics, missing data occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Mainly, non-responsiveness is the major reason for their existence. Manual data entry procedures, equipment errors and incorrect measurements are some other

reasons of missing value existence. The presence of such failures usually requires a preprocessing stage in which the data preparation. Mainly missing data can occur because of nonresponse. That means no information is provided for several items or no information is provided for a whole unit. Some items are more sensitive for nonresponse than others, for example items about private subjects such as income.

METHODOLOGY

Since the main objective of the study is to identify the factors affecting missing values in ASI, the literatures were reviewed to acquaint about the behavior of missing values in survey data. Secondary data were collected from the survey data of ASI in 2012 conducted by the ITSD of DCS. Few expertise officers were interviewed to attain basic conception of data. Variables were identified by referring collected data. The binary logistic regression model was developed to identify the factors affecting missing values in ASI.

DATA COLLECTION AND ANALYSIS

ASI data in 2012 were collected which had 3121 cases. Those collected data were used for the analysis of the research.

Mostly missing values are occurring in the income variable of ASI. Therefore the dependent variable of the binary logistic regression model becomes as whether a value for income has or not. Six independent variables were selected after studying the data set. They are employment size, sector, area, industry type, establishment type and legality type.

Hypothesis

H_0: There is no relationship between dependent variable (income) and independent variables

H_1: There is a relationship between dependent variable (income) and independent variables

Table 1: Values of Pearson Correlation Test

Relationship between	P value	Sig: value	Result
Income and Establishment type	0.000	< 0.01	**Reject H_0**
Income and Area	0.000	< 0.01	**Reject H_0**
Income and Sector	0.000	< 0.01	**Reject H_0**
Income and Industry type	0.025	< 0.05	**Reject H_0**
Income and legality type	0.000	< 0.01	**Reject H_0**
Income and Total Employee	0.004	<0.01	**Reject H_0**

According to the correlation test, it has been identified that there is a relationship between dependent variable (A value for income has or not) and independent variables at 5% level of significance. Henceforth binary logistic regression was conducted to identify factors affecting missing values.

Table 2: Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	4198.497[a]	.040	.053
2	4142.581[a]	.057	.076
3	4112.956[a]	.066	.088
4	**4089.759[a]**	**.073**	**.097**

The model summary illustrates that 7.3% of the variation in the dependent variable is explained by the logistic model and there is a 9.7% of relationship between predictor variables. Though the model summary table represents a poor model, practically that was the best fitted model for the data and it was considered to construct the final conclusion.

Table 3: Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	2	1.000
2	.972	5	.965
3	3.278	6	.773
4	**7.295**	**6**	**.294**

Hypothesis

H_0: Model is significant
H_1: Model is not significant.

The Hosmer and lemeshow test result shows that significant value is greater than 0.05. Hence the selected model is a significant to explain the dependent variable.

In pursue of Table 4, four variables were remained out of the six variables after the fourth step which gives a significant contribution to the dependent variable at 5% significant level.

Table 4 – Variables in the Equation

Step	Variable	Wald statistic	Df	p-value
Step 1	Legality Type	112.493	6	.000
Step 2	Area	54.578	1	.000
	Legality Type	115.336	6	.000
Step 3	Area	54.578	1	.000
	Legality Type	115.336	6	.000
	Total Employee	29.133	3	.000
Step 4	**Establishment Type**	22.969	3	**.000**
	Area	50.173	1	**.000**
	Legality Type	93.817	6	**.000**
	Total Employee	35.434	3	**.000**

Affected categories of each identified variables and their coefficients were used to develop the binary logistic regression model. Therefore, the logistic model can be formulated as;

$\text{Logit}\left(\frac{P}{1-P}\right)$ = -0.815 + 0.514 (Establishment Type 1) + 0.410 (Establishment Type 2) +0.636 (Area 1) + 0.937 (Legality Type 1) + 0.488 (Legality Type 2) +0.813 (Legality Type 3) + 1.514 (Legality Type 4) + 1.819 (Legality Type 5) + 1.107 (Legality Type 6) – 0.505 (Total Employee 1) – 0.943(Total Employee 2)

Where;
P = the probability of occurring missing values of income rather than not occurring missing values of income given the values of independent variables.

Establishment Type 1 = Head office and factory (No other office)
Establishment Type 2 = Head office and factory (other factory somewhere)
Area 1 = Urban
Legality Type 1 = Individual Owner
Legality Type 2 = Partnership
Legality Type 3 = Private limited liability Co-operation
Legality Type 4 = Public limited liability Co-operation
Legality Type 5 = Co-operative Society
Legality Type 6 = State Corporation
Total Employee 1 = Less than 100
Total Employee 2 = Between 100-500

Then logistic model became as;

$$P = \frac{exp\{\gamma\}}{1 + exp\{\gamma\}}$$

Where;

γ = – 0.815 + 0.514 (Establishment Type 1) + 0.410 (Establishment Type 2) + 0.636 (Area 1) + 0.937 (Legality Type 1) + 0.488 (Legality Type 2) + 0.813 (Legality Type 3) + 1.514 (Legality Type 4) + 1.819 (Legality Type 5) +1.107 (Legality Type 6) – 0.505 (Total Employee 1) – 0.943(Total Employee2)

RESULTS AND DISCUSSION

This study mainly considers exposing the factors affecting missing values. There had been some discussions with few expertise officers about occurring missing values in their survey data to obtain the basic knowledge. Data set was clearly studied to identify the distribution of missing values among categories of each variable. Analysis results revealed that four variables were contributed for the model out of selected six variables such as Establishment Type, Legality Type, Area and Number of Total Employee.

CONCLUSION

Analysis results demonstrate that main four factors affect for missing values in ASI as Establishment Type, Legality Type, Area and Number of Total Employee. Those factors can be further expanded with their categories. Establishment Type categorizes according to the place where the data were collected for the ASI. According to that, two establishment types are affected for occurring missing values. Those are establishments which have only head office and factory (no other factory) and establishments which have head office and factory (other factory somewhere). There are six legality type categories are affected for missing values. Those are individual owner, partnership, private limited liability co-operation, public limited liability co-operation, co-operative society and state co-operation. Establishments which have above legality type categories are affected for occurring missing values. Establishments which are in urban areas also contribute to occurring missing values. Establishments which have total employees less than 100 and total employees between 100-500 are affected for missing values. It can be concluded that small establishments are refused to attend to the survey than large establishments. With those all of the categories, eleven factors are affected for occurring missing values in ASI. The best approach to reduce missing values in ASI is being attentive about establishments which have above identified factors during data collection stage.

REFERENCES

- Pigott,T. D. (2001). A review of methods for missing data, *Educational Research and Evaluation*, 7, 353-383.
- Schafer, J. L., & Graham, J.W. (2002). Missing data: Overview of the state of the art, *Psychological Methods, 7*, 147-177.
- Sujuan gao & Siu L.H.(1997). Logistic regression models with missing covariate values for complex survey data, *Statistics in Medicine*, 16, 2419-2428.
- Wagner A.K., & Michel W. (2000). Factor Analysis and Missing data, *Marketing Research*, 37,490-498.