Missing Data in Willingness-to-pay Measures and Its Influences in Analyses

Jane Lu Hsu

Department of Marketing, National Chung Hsing University, 250 Kuo Kuang Road, Taichung, 40227, Taiwan

jlu@dragon.nchu.edu.tw

Keywords: Missing data, Willingness-to-pay measures, Quantitative analysis.

Abstract. Missing data in willingness-to-pay measures appears in nearly every dataset utilizing consumer surveys. This study utilizes consumer survey data to examine the differences in distributional properties using various missing data replacement methods, including replacing by mean or median values of the complete observations, replacing by predicted values of multiple regressions, replacing by predicted values of simultaneous system equations, replacing by cluster means, and utilizing Bayesian approach to estimate the posterior distributions for missing observations. Results of this study indicate that for datasets of consumer surveys with relatively large portions of missing data, replacing the incomplete observations with the mean or median values of complete observations can distort the distributional properties of the variables. Clustering or Bayesian approach is recommended to maintain the distributional properties of variables to the third and the fourth moment (skewness and kurtosis). Missing data in willingness-to-pay measures needs to be dealt with thorough considerations in quantitative analyses. The implication of this study is that missing data should not be ignored in willingness-to pay measures and researchers need to be aware of different approaches in handling missing data in quantitative analyses.

1. Introduction

Consumer surveys are administered frequently to gather information of opinions, preferences, purchasing patterns, and reactions to certain marketing information. Missing data in consumer surveys are a common phenomenon, caused by ignorance of surveyors, unwillingness to answer, incompleteness or incapability in response to certain questions of respondents. Missing data can cause serious biasedness in estimation and interpretations of analytical results. Kao and Liu [1] mention that the decision-making individuals sometimes are unable to provide certain data, and then the possibility distributions of imprecise data need to be reconstructed.

In cases that complete observations are used in analyses while the incomplete observations are salvaged to extract limited information, Griliches [2] refers this as the 'ignorable case.' Afifi and Elashoff [3] state that the missing data can be considered as functions of additional parameters. Further research regards missing data as random variables and can be replaced by real values under certain assumptions [4,5]. Greene [6] suggests two schemes to replace missing data using the zero-order or the first-order regressions. The zero-order regression refers to replacing the missing values by the means of the complete observations. The first-order regression is to use predicted values of the variables resulting from a regression model to restore the missing observations.

Sargan and Drettakis [7] use simultaneous equation models and treat missing data as parameters. Using an iterative procedure, the predicted values of missing data are re-estimated with the

ISBN: 978-605-64453-0-9

parameters of the whole sample period. Greene [6] mentions that properties of predicted values of variables used in replacing missing observations are not thoroughly discussed in literature.

Vriens and Melton [8] suggest several methods to replace missing data, including replacing the missing data by the mean values, using the grouping techniques and replacing by the nearest complete record (random hot-deck imputation), replacing by the predicted values based on regressions using other variables with complete observations, and utilizing model-based multiple imputation for the missing values. Little and Rubin [9] describe the method of the Bayesian approach to restore nonignorable missing data. The advantage of utilizing the Bayesian approach is that uncertainty about the value of an unknown parameter can be expressed using a probability distribution [10]. The influence of the missing data can be assessed by a probability of the predictive distribution of the complete-data statistics [9].

This paper utilizes consumer survey data to examine general statistical properties of missing data replacements using six different approaches to provide further insights into the features of incomplete observation restorations. The missing data replacement approaches utilized in this study include methods of replacing by the mean values of the complete observations, replacing by the median values of the compete observations, using predicted values of multiple regressions to replace missing observations, using predicted values of a system of simultaneous equations to replace missing observations, using cluster means to replace missing observations within clusters, and utilizing the Bayesian approach to estimate the posterior distributions.

2. Methodology

2.1 Missing data replacement

2.1.1 Replacing missing data by the mean values

The dataset can be partitioned into subsets of complete and incomplete observations for the respective variables. The mean value for each variable based on the observations available is calculated to substitute the missing observations.

2.1.2 Replacing missing data by the median values

The method also uses the complete observations to generate the distributions of the variables. The median values are used to substitute the missing observations of variables.

2.1.3 Replacing missing data by the predicted values of multiple regressions

Vriens and Melton [8] refer this method to ‘model-based mean imputations’ and state that this method makes better use of the data structure. This method includes two steps. First, a regression is performed on the variable with incomplete observations with other variables as predictors. Second, the predicted value of each missing observation is imputed as the replacement. This procedure is repeated for all the variables with incomplete observations.

2.1.4 Replacing missing data by the predicted values of a system of simultaneous equations

In cases the variables with missing observations are in a category that respondents consider using the same or similar criterion, the replacements of missing observations need to be calculated using a system of simultaneous equations instead of individual regressions. The replacements of missing observations are generated simultaneously.

2.1.5 Replacing missing data by the cluster means

Complete observations are used in the clustering procedure to form groups with similar attributes. The most commonly used technique for grouping individuals or households with similar characteristics is the cluster analysis [11]. The essential criterion needed for cluster analysis is to classify experimental units into classes or groups, so that the units within a class or group are similar to one another while units in distinct classes or groups are different [12]. Two basic methods,

ISBN: 978-605-64453-0-9

hierarchical and the nonhierarchical methods, can be used to search and define clusters [13]. The hierarchical method separates units into different groups in a nested sequence of clustering. The major defect of hierarchical method is that subsequent steps can never repair mistakes made in earlier steps [14]. The nonhierarchical method, referred to as the *K*-means clustering method, is a popular practical technique in the clustering analysis [15] and is applied in this study. Observations with incomplete data are grouped using the variables of complete observations or weaved among observations with complete data. The mean values of clusters are used to substitute the missing values of observations within clusters.

2.1.6 Replacing missing data using the Bayesian approach

Assuming the prior probabilities of variable x are known, the posterior probabilities of observations with missing data can be calculated using the Bayes' theorem:

$$p\langle \varphi | x \rangle = \frac{q_{\varphi} f_{\varphi}(x)}{f(x)} \tag{1}$$

where f(x) is the unconditional density distribution of x, q_{φ} is the prior probability of observations having the values represented by parameter φ, $f_{\varphi}(x)$ is the value-specific density distribution of x, and $p(\varphi|x)$ is the posterior probability of the observation having specific value.

2.1.7 Comparisons of distributional properties of missing data replacements

General distributional properties of variables before and after missing data replacements are presented in this study for comparisons. Missing data restorations are not to reshape the distributions of the variables to the normal distributions. The principles of missing data replacements are to retain the original distributional properties of the variables so the subsequent analyses are not severely affected by data restorations.

Four fundamental distributional properties are examined in this study, including mean, variance, skewness, and kurtosis. For each variable with missing data that are replaced by certain values, the original distributional properties and the properties of restored data up to the fourth moment are compared to examine the distortions of data replacements.

The skewness of a variable describes the asymmetry of the probability distribution around the sample mean. Negative skewness indicates that the data are spread out more to the left of the mean, and positive skewness specifies that the data are spread out more to the right of the mean. The skewness of a distribution is the following.

$$\text{skewness}(x) = \frac{E\left[(x - \mu)^3\right]}{\sigma^3} \tag{2}$$

The kurtosis of a variable measures the peakedness of the probability distribution. Normal distributions have kurtosis of three. More peaked distributions have kurtosis larger than three, and less peaked distributions have kurtosis less than three. The kurtosis of a distribution is the following.

$$\text{kurtosis}(x) = \frac{E\left[(x - \mu)^4\right]}{\sigma^4} \tag{3}$$

2.2 Data

A dataset generated from a consumer survey administered in 2013 in the three most populated metropolitan areas in Taiwan (Taipei, Taichung, and Kaohsiung) is utilized in this study to examine

ISBN: 978-605-64453-0-9

the differences of general distributional properties of missing data replacements. Stratified sampling was applied following the gender and age distribution of the population between the ages of 18 to 49. Questionnaires were designed based on findings in the literature and discussions with professionals and practitioners. A pilot survey of six valid samples was administered prior to the formal survey. The dataset used in this study are real consumer survey data to reduce the limitations of applicability of the results. The considerations are to have the datasets with missing observations replaced in a way that the general statistical properties are not seriously distorted. The closer the properties of the dataset with restorations are to the properties of the originals, the better the replacements are.

A total of 407 valid samples were obtained using personal interviews. The formal survey was conducted in memorial parks, at central stations, on university campuses, at traditional markets, hypermarts, business districts, and science museums in three metropolitan areas to ensure the diversity of respondents. Qualified respondents needed to be residence of the city where survey was administered and had personal computers (desktop or notebook) with self-installed software. A gift worth approximately three US dollars was provided to each respondent who was willing to participate. Respondents needed 20 to 30 minutes to answer all of the questions in the questionnaire. Trained surveyors provided necessary assistance in explaining the questions but would not interfere using personal judgments. Refusal rate was close to 10 percent, mainly due to time constrains of potential respondents.

Variables used in missing data replacements are willingness-to-pay (WTP) measures of computer software, *Windows 8* and *Office 2013*. WTP measures are commonly applied in research to generate data for quantitative analyses. In this study, WTPs are measured using the payment card method, and *Microsoft Windows 8* and *Office 2013* were objects used for respondents to assess WTP values. Ranges of WTP were provided for each software product, covering the recommended market price of the software and equally divided into 14 smaller ranges for respondents to choose from. An additional question was added to ask for respondents to write down the exact dollar amount of WTP for the software within the chosen price range.

In demographics, 49.63% of the respondents were females and 45.70% of the respondents were married. The average age of respondents was 33.73 years old, living in households of 4.04 persons on average. About 90% of respondents had educational levels of college or above, reflecting relatively high educational levels of people living in metropolitan areas in general. Slightly less than one-third of respondents worked in the business sector, and additional 20.64% were students. Less than 5% of respondents were housewives. Average monthly household income was USD 3132.55 approximately.

3. Results

This study intends to examine the general statistical properties of missing data replacements for consumer surveys. In order to demonstrate distortions in general statistical properties in missing data replacements, one-third of observations were removed from the dataset. Observations were coded with sequential numbers and the data deleting process was randomized using SAS software to remove one-third of the samples, 135 observations out of 407 samples, leaving 272 observations as remaining samples.

The first and second methods of missing data replacements use mean and median values, respectively, and results are listed in Table 1. Mean and median values are commonly used for replacing missing data in statistical software. The variances of restored samples are reduced due to increased numbers of observations after replacement. For skewness, replacing the missing data by median values skews the data distributions more than replacing by the mean values. Both methods for missing data replacements cannot maintain the distributional properties at a higher moment. The kurtosis of the distribution becomes much larger with mean or median value replacement. The peakedness of distributions is hardly stable when missing data are replaced by mean or median values of complete observations. In sum, both mean and median values distorted distributional properties of

ISBN: 978-605-64453-0-9

observations once used in missing data replacement. These two commonly applied methods in commercial software need to be used with caution. With more advanced methods available, researchers are recommended to check for missing data handling of commercial software before using it.

Table 1. Distributional properties of missing data replaced by mean and median values

WTP	Distributional properties	Original data	Missing data with replacements	
			Replaced by mean values	Replaced by median values
Windows 8	Mean	58.82	58.82	56.05
Office 2013		60.45	60.45	57.14
Windows 8	Variance	1381.61	922.21	937.73
Office 2013		1544.51	1030.94	1053.13
Windows 8	Skewness	0.92	1.12	1.36
Office 2013		0.84	1.03	1.29
Windows 8	Kurtosis	3.62	5.42	5.68
Office 2013		3.13	4.69	4.96

The third method is to use the expected values of linear regression models to replace missing observations as suggested by Gourieroux and Monfort [5]. The explanatory variables included in the regression models are considering factors such as price and convenience, whether authorized software has been installed for personal usage, whether respondents would recommend authorized software to others, possibilities to purchase authorized software in the future, whether respondents believe using pirated software is just a behavior like everyone else does (i.e., crowd behavior), whether there is a not-in-my-back-yard phenomenon regarding huge loss in profit of the software industry from piracy, and demographic variables. Regression models are individually estimated for *Windows 8* and *Office 2013*. Expected values of WTP for *Windows 8* and *Office 2013* are used for missing data replacement. The fourth method uses system equations to estimate expected values of WTP for *Windows 8* and *Office 2013* simultaneously. Explanatory variables are the same as individual regression models. Expected values of WTP for *Windows 8* and *Office 2013* from simultaneous are used for missing data replacements. Results are listed in Table 2. Variances are largely shrunk using regression or simultaneous equations after replacing missing data. Peakedness of data distribution (kurtosis) is reduced, and data is more likely to be negatively skewed. In sum, predicted values of regression models or simultaneous equations can maintain unbiasedness of mean values, but not other distributional properties in missing data replacement.

Table 2. Distributional properties of missing data replaced by predicted values of regressions and simultaneous equations

WTP	Distributional properties	Original data	Missing data with replacements	
			Replaced by predicted value of regression	Replaced by predicted values of simultaneous equations
Windows 8	Mean	58.82	59.12	59.05
Office 2013		60.45	60.19	60.28
Windows 8	Variance	1381.61	294.85	219.79
Office 2013		1544.51	358.28	285.07
Windows 8	Skewness	0.92	-0.08	-0.13
Office 2013		0.84	0.04	-0.05
Windows 8	Kurtosis	3.62	2.31	2.46
Office 2013		3.13	2.12	2.20

ISBN: 978-605-64453-0-9

The fifth method is to use a nonhierarchical clustering method to replace missing data that have been randomly removed. The mean values of clusters were utilized to replace missing values. The last method is to use the Bayesian approach in order to generate the values of missing data with the highest posterior possibilities. A nonparametric method, kernel density estimates with normal kernels, with unequal bandwidths in smoothing parameters is used. The general statistical properties of the original data and the data with restorations are listed in Table 3. The Bayesian approach provides the mean values close to the original means. Variances are not largely shrunk, and distributional properties of higher moments (i.e., skewness and kurtosis) are close to those of the original distributions. Replacing missing data with cluster means provides distributional properties close to those using Bayesian approach, but are slightly more positive skewed in skewness. Both of these two methods are considered appropriate to replace missing data values when large proportions of observations are missing.

Table 3 Distributional properties of missing data replaced by cluster means and using Bayesian approach

WTP	Distributional properties	Original data	Missing data with replacements	
			Replaced by cluster means	Replaced missing data using Bayesian approach
Windows 8	Mean	58.82	58.38	58.90
Office 2013		60.45	59.87	60.36
Windows 8	Variance	1381.61	990.83	1014.87
Office 2013		1544.51	1140.41	1152.67
Windows 8	Skewness	0.92	1.02	0.97
Office 2013		0.84	0.94	0.89
Windows 8	Kurtosis	3.62	4.77	4.51
Office 2013		3.13	3.95	3.81

4. Conclusion

This study utilized real consumer survey data to examine the differences in distributional properties using various missing data replacement methods. A dataset from consumer surveys conducted in 2013 in the three most populated metropolitan areas in Taiwan (Taipei, Taichung, and Kaohsiung) are utilized in the study. A total of six methods are used to generate replacements for missing data to fit the types of the datasets. The considerations are to have the datasets with missing observations replaced while the general statistical properties can be maintained.

The results of this study indicate that when the missing observations are accounted for large proportions of the dataset, using the mean or median values of complete observations to replace the missing values can seriously affect the distributional properties of the variables. Advanced methods like clustering or Bayesian approach are recommended to maintain the distributional properties of variables with replacements to the third and fourth moment statistics. Furthermore, researchers who need to replace missing data in consumer surveys may need to use different approaches and compare the results of distributional properties to determine the suitable method for missing data replacements.

Limitations of this study are only one dataset is utilized in analysis. Further studies may use different types of survey data and may consider other advanced method like neutral network-based analysis for missing data replacement as suggested in Gheyas and Smith [16].

References

[1] C. Kao and S. T. Liu, Data development analysis with missing data: an application to University libraries in Taiwan, *Journal of the Operational Research Society*, vol. 51, pp. 897-905, 2000.

ISBN: 978-605-64453-0-9

[2] Z. Griliches, Economic data issues. In: Griliches Z and Intriligator M (eds). *Handbook of Econometrics*. North Holland: Amsterdam, vol. 3, 1986.

[3] A. A. Afifi and R. M. Elashoff, Missing observation in multivariate statistics II: point estimation in simple linear regression, *Journal of the American Statistical Association,* vol. 62, pp. 10-29, 1967.

[4] M. G. Dagenais, The use of incomplete observations in multiple regression analysis, *Journal of Econometrics,* vol. 1, pp. 317-328, 1973.

[5] C. Gourieroux and A. Monfort, On the problem of missing data in linear models, *Review of Economics Studies,* vol. 48, pp. 579-586, 1981.

[6] W. H. Greene, *Econometric Analysis*, 4th ed. Prentice-Hall: Upper Saddle River, NJ, 2000.

[7] J. D. Sargan and E. G. Drettakis, Missing data in an autoregressive model, *International Economic Review,* vol. 15, pp. 39-58, 1974.

[8] M. Vriens and E. Melton, Managing missing data, *Marketing Research,* vol. 14, pp. 12-17, 2002.

[9] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data,* John Wiley & Sons: New York, 1987.

[10] G. G. Judge, W. E. Griffiths, R. C. Hill, H. Lütkepohl, and T. C. Lee, *The Theory and Practice of Econometrics*, 2nd ed. John Wiley & Sons: New York, 1985.

[11] J. F. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black, *Multivariate Data Analysis with Readings,* 3rd ed. Macmillam Publishing Company: New York, 1992.

[12] D. E. Johnson, *Applied Multivariate Methods for Data Analysis*, Brooks/Cole Publishing: California, 1998.

[13] D. J. Hand, *Discrimination and Classification*, John Wiley & Sons: Chichester, U.K., 1981.

[14] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data - An Introduction to Cluster Analysis*, John Wiley & Sons: New York, 1990.

[15] A. A. Afifi and V. Clark, *Computer-Aided Multivariate Analysis*, 2nd ed. Van Nostrand Reinhold: New York, 1990.

[16] I. A. Gheyas and L. S. Smith, A neural network-based framework for the reconstruction of incomplete data sets, *Neurocomputing*, vol. 73, pp. 3039-3065, 2010.

ISBN: 978-605-64453-0-9