The Development of Hot-deck Corrected Item Mean (HDD-CIM) for Estimating Missing Data

Paitoon Muliwan[1*], Nipaporn Chutiman[2], Prapas Pue-on[3]

Received: 15 March 2013 Accepted: 15 May 2013

Abstract

The purpose of this study was to develop the hot-deck corrected item mean (HDD-CIM) for missing data estimation and compare its under missing complete at random (MCAR) and simple random sampling with two methods, namely; Corrected item mean (CIM) and hot-deck (HDD). The secondary data from a survey of public information about the prevalence of drugs, 2011, a survey by the Bureau of Statistics Mahasarakham Province were used and the comparisons were made with three sample sizes (100, 200 and 500) and four levels of percentage of missing data (5%, 10%, 15% and 20%). It appears that the HDD-CIM method has most efficiency in estimating missing data.

Keywords: Missing data, MCAR, Simple random sampling, Corrected item mean (CIM) , hot-deck (HDD)

Introduction

Missing data a problem in many field[1] of research, and the researcher must consider to appropriate process for management of missing data in every case. Sometimes it may not bea serious problem that may be regarded as a trivial matter, nevertheless the experiments showed that, if each variable with random missing data, only 10% the unit of analysis will cut off 59% (Roth, 1995). Missing completely at random (MCAR) is a process in which the missingness of the data is completely independent of both the observed and the missing values, The study found that if the missing data mechanism is MCAR, then the results from many missing data procedures would be valid. On the other hand, if data are not MCAR, care must be exercised in employing routine missing data procedures. Thus, statistical tests of MCAR are important and of interest (Jamshidian and Jalal, 2010; Little, 1988). The study found that CIM is the best technique and easy to compute and yields good estimates of scale score of person, although one should bare in mind the overestimation of the scale quality. In general the performance of CIM is best and it shows that there is much to gain when measurement models are used for the imputation of missing values to test-items (Huisman, 2000). Hot-deck (HDD) as a replacement for the loss information by donor from same research / explore, so the survey are similar to the units with missing data. It was found that the HDD is the smallest bias and highest precision (Montree Piriyakul, 2005), the disadvantage of this method is a practical way and less theory to support. Although flexible and widely used by practitioners to handle item non-response. But may have a theoretical objection (Montree Piriyakul, 2005). And the hot deck is widely used by practitioners to handle item non-response. Its strengths are that it imputes real (and hence realistic) values, it avoids strong parametric assumptions, it can incorporate covariate information, and it can provide good inferences for linear and non-linear statistics if appropriate attention is paid to propagating imputation uncertainty. A weakness is that it requires good matches of donors to recipients that reflect available covariate information; finding good matches is more likely in large than in small samples (Andridge, R. R. and Little, R. J. A., 2010). Therefore, the objective was to take advantage of these

[1,2,3] *Mathematics Department Faculty of Science Mahasarakham University Mahasarakham, Thailand E-mail: paitoon.m@msu.ac.th*

[*] *Coressponding author: Email: paitoon.m@msu.ac.th*

two methods to develop a method for estimating the lost data with the HDD-CIM. This study using secondary data from a survey of public information about the prevalence of drugs by the Bureau of Statistics Mahasarakham Province 2011, there were 2,970 records.

Methods

The sample was selected from the population of 2,970 records by simple random sampling with 100, 200 and 500, generated missing data using MCAR at 5%, 10%, 15% and 20%. Calculated mean square error (MSE) from Eq(1).

$$MSE = \frac{\sum_{i=1}^{n}(\theta_i - \hat{\theta}_i)^2}{n} \quad (1)$$

where θ_i is the old value and $\hat{\theta}_i$ is the new value and *n* is number of missing value. And to compare the results of the estimation of missing data between CIM , hot-deck (HDD) and HDD-CIM.

CIM replaces missing values by the item mean which is corrected for the ability of the respondent, i.e., the score on the observed items of the respondent compared with the mean score on these items.

The operation of CIM as shown from Eq(2).

$$CIM_{vi} = \frac{\bar{x}_{.1}^{(i)} \times \sum_{i \in obs(v)} x_{vi}}{\sum_{i \in obs(v)} \bar{x}_{.i}^{(i)}} \quad (2)$$
$$; v = 1,2,...,m_i$$

where $\bar{x}_{.i}^{(i)}$ is the mean score on item i for non-missing data and *obs (v)* is the collection of observed items.

Hot-deck deterministic method (HDD) uses the complete case for which the distance function is minimized. When several complete cases are at the same minimal distance of the currently considered incomplete case, the complete case which is nearest to the incomplete case with respect to its place in the data matrix is used as a donor case.

The operation of HDD as shown from Eq(3).

$$d_{vv'}^2 = \sum_{i \in obs}(x_{vi} - x_{v'i})^2 \quad (3)$$

where v is incomplete and v' a complete case.

HDD-CIM is a mix methods by using CIM to replace missing value, calculate the distance between the survey data as a way to decide by HDD.

The operation of the HDD-CIM, as shown below.

Step 1: replace missing data with the mean of the variables according to CIM for the temporary file following Eq(4).

$$CIM_{vi} = \frac{\bar{x}_{.1}^{(i)} \times \sum_{i \in obs(v)} x_{vi}}{\sum_{i \in obs(v)} \bar{x}_{.i}^{(i)}} \quad (4)$$
$$; v = 1,2,...,m_i$$

where $\bar{x}_{.i}^{(i)}$ is the mean score on item i for non-missing data and *obs (v)* is the collection of observed items.

Step 2: Calculate the distance between the survey and the HDD can be calculated using the following Eq(5).

$$d_{vv'}^2 = \sum_{i \in obs}(x_{vi} - x_{v'i})^2 \quad (5)$$

where v is incomplete and v' a complete case.

Step 3: use the complete case for which the distance function is minimized. When several complete cases are at the same minimal distance of the currently considered incomplete case, the complete case which is nearest to the incomplete case with respect to its place in the data matrix is used as a donor case.

Results

This study by experiment performed with real data. The secondary data from a survey of public information about the prevalence of drugs, 2011, a survey by the Bureau of Statistics Mahasarakham Province were used. The sample was selected from the population by simple random sampling and to compare the results of the estimation of missing data between the HDD-CIM, CIM and hot-deck (HDD) by using the mean square error (MSE). In the ex-

periments, the percentage of missing data are 5%, 10%, 15% and 20%, and three samples sizes are 100, 200 and 500. Table 1 present the average values of the MSE of the imputation techniques for all factors of the design.

Table 1 Mean Square Error of CIM, HDD and HDD-CIM classified by sample sizes and percentage of missing data.

Sample sizes	Percentage of missing data	Methods		
		CIM	***HDD***	***HDD-CIM***
100	5	3.0798	3.1357	1.9130
	10	2.8944	2.9868	1.9595
	15	2.0867	2.8821	1.8491
	20	2.8628	2.5492	1.7024
200	5	3.1830	3.3952	2.2013
	10	3.0852	3.4571	2.0366
	15	2.9263	3.3973	2.0613
	20	3.1714	3.4623	2.1883
500	5	3.0589	3.3672	1.9357
	10	3.0766	3.4057	2.0683
	15	3.0132	3.2320	2.0695
	20	2.8856	3.0806	1.8270

Bold numbers mean having the most efficiency

From Table 1 it follows that across all factors HDD-CIM is the best technique. For each independent variable separately, HDD-CIM also performs the best, closely follows by CIM and HDD in varying order. This means that the HDD-CIM has most efficiency for all factors.

Table 2 Mean Square Error of CIM, HDD and HDD-CIM classified by percentage of missing data.

Percentage of missing data	Methods		
	CIM	***HDD***	***HDD-CIM***
5	3.1072	3.2993	2.0166
10	3.0187	3.2832	2.0214
15	2.9154	3.1705	1.9933
20	2.9732	3.0307	1.9059

From Table 2 When classify by percentage of missing data. 20% of missing data also performs best, closely follows by 15%, 10% and 5% in varying order.

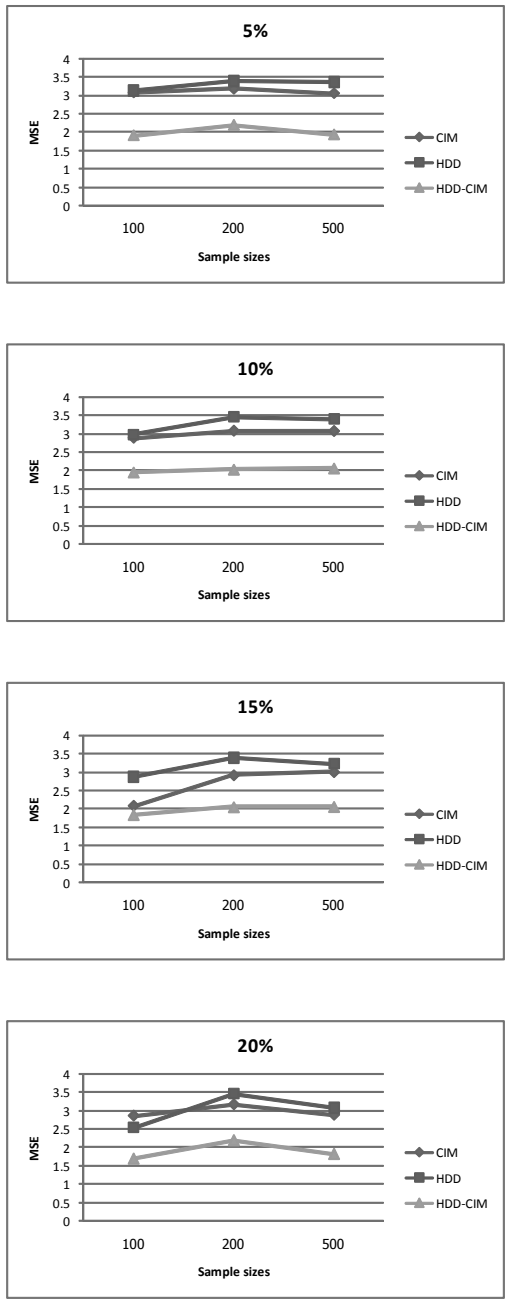

Figure 1 Mean Square Error of CIM, HDD and HDD-CIM classified by percentage of missing data.

Summary and Concluding

We demonstrated the effectiveness of the proposed HDD-CIM method using sample sizes and percentage of missing data. In table 1-2 HDD-CIM has the lowest mean square error. This means that the HDD-CIM has most efficiency for each sample size and each percentage of missing data.

References

[1] Andridge, R. R. and Little, R. J. A., "A Review of Hot Deck Imputation for Survey Non-response", International Statistical Review. 78(1): 2010, 40-64.

[2] Jamshidian, M., and Jalal, S., "Test of homoscedasticity, normality and missing completely at random for incomplete multivariate data", Psychometrika. 75(4): 2010, 649 – 674.

[3] Little, R. J. A., "A test of Missing completely at random for multivariate data with missing values", Journal of the American Statistical Association. 1988 83(404).

[4] Mark Huisman."Imputation of Missing Item Responses: Some Simple Techniques", Quality & Quantity. 2000, 331–351.

[5] Montree Piriyakul. Missing Data Replacement Models in Social Science Research : Simulation Study of Simple Models. Ramkhamhaeng University, 2005.

[6] Roth, P.L. and others, "The Impact of Four Missing Data Techniques on Validity Estimates in Human Resource Management", Journal of Business and Psychology. 11(1) :1996, 101 - 112.