Logistic Regression with Missing Data: A Comparison of Handling Methods, and Effects of Percent Missing Values

Sutthipong Meeyai
School of Transportation Engineering, Suranaree University of Technology, 111 University Ave., Muang, Nakhon Ratchasima, 30000, Thailand
Email: sutthi@sut.ac.th

***Abstract*—The aim of this article is to compare five popular missing data handling methods: listwise deletion, mean substitution, regression imputation, stochastic imputation, and multiple imputation. Three missing data mechanisms are investigated: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). A Monte Carlo simulation is applied to simulate data and then logistic regression parameters are estimated. Our findings show that, among the five missing data handling methods, multiple imputation performs well on both MCAR and MAR. There is no evidence indicating that listwise deletion and multiple imputation produce biased parameters for MCAR. None of these techniques can handle MNAR. Finally, this article suggests maximum percent missing data and a sample size for listwise deletion and multiple imputation techniques.**

***Index Terms*—logistic regression, multiple imputation, listwise deletion, missing at random, incomplete data**

I. Introduction

Missing data has been pervasive in several research areas e.g. psychology [1], public health [2], and education [3]. In transportation planning, however, missing data is often ignored and is not reported how it is remedied. Missing data handling techniques for a logistic regression, which is often used to model a choice among alternatives, need more investigation.

Most statistical analyses and software packages assume that all variables in the model are measured. The default procedure normally deletes cases with missing data on the variables of interest, which known as listwise deletion. The major disadvantage of this method is that it regularly removes a large proportion of the sample, leading to a severe loss of statistical power [4]. Wilkinson [5] stated that deletion methods are among the worst methods available for practical applications.

Ignoring missing data is typically resulted from lack of awareness or failure to recognize the significance of the problem [6]. Saunders, Morrow-Howell [7] reported that only 15 percent of literature reviews, approximately 100 articles between 2001 and 2003, provided information about the amount of missing data or showing how missing data were handled in their studies. Davey and Savla [8] also noted that the implications of missing data for social research had not received well-known treatment to date.

Although the importance of missing data have long been realized, the comparison of missing data handling methods for a logistic regression is rarely discussed. The aim of this research is to provide researchers guidelines to choose appropriate missing data handling techniques for logistic regression analysis and to suggest maximum percent missing data and a sample size.

II. Theoretical Background

For any data set, one can define indicator variables R as missingness, that is identify what are known and what are missing. In modern missing-data procedures, missingness has been regarded as a probabilistic phenomenon [9]. Rubin treated R as a set of random variables having a joint probability distribution. He developed a typology of missing data which became widely practiced by researchers since then. The mechanisms consist of missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). These mechanisms define relationships between interesting variables and the likelihood of missing data. Each type of missing data dictates the performance of imputation techniques.

Let X_{com} denote the complete set of data. Its components may be expressed as $X_{\text{com}} = (X_{\text{obs}}, X_{\text{mis}})$, where X_{obs} and X_{mis} are the observed and missing parts, respectively. Rubin [9] defines missing data as missing at random (MAR) if the distribution of missingness does not depend on X_{mis},

$$P(R \mid X_{com}) = P(R \mid X_{obs}) \tag{1}$$

MAR assumes that the probability of missingness depends on observed data but not on missing data. A special case of MAR is missing completely at random (MCAR) which occurs when the distribution does not depend on either,

$$P(R \mid X_{com}) = P(R) \tag{2}$$

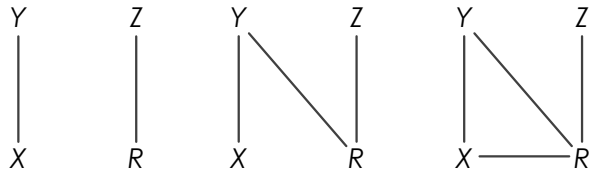
Manuscript received February 14, 2016; revised June 5, 2016.

doi: 10.18178/jtle.4.2.128-134

In other words, the missing data are MCAR if the probability of missing data on a variable X is not associated with other measured variables and with the values of X itself. When the previous assumption are violated, the distribution depends on X_{mis},

$$P(R \mid X_{com}) = P(R \mid X_{miss}) \quad (3)$$

Such missing data are considered Missing Not At Random (MNAR). Schafer and Graham [10] provide useful graphical relationships. They let Z denote the components that is not related to X and Y, where variables Y are known for all participants and related to partially missing variables X. MCAR, MAR, and MNAR can be illustrated in Fig. 1. In MCAR the missingness only relates to Z. Under MAR not only does the missingness relate to Z, but also links to Y. In MNAR the missingness is associated with all X, Y, and Z.



(a) MCAR *(b) MAR* *(c) MNAR*

Source: adapted from Schafer & Graham [10]

(a) missing completely at random (MCAR), (b) missing at random (MAR), and (c) missing not at random (MNAR) in a univariate missing-data pattern. *Y* represent a variable that is completely observed, *X* represent a variable that is partly missing, *Z* represents the component of the causes of missingness unrelated to *X* and *Y*, and *R* represents the missingness.

Figure 1. Graphical representation of missing data mechanisms.

Of three mechanisms, MCAR is the sole mechanism that can be tested empirically [1]. Unfortunately, MAR and MNAR mechanisms cannot be verified because they depend on unobserved data. Table I illustrates three mechanisms in the context of a binary logistic regression for a transportation mode choice model. Y represents a choice between two travel alternatives (e.g. $Y = 1$, choose auto; and $Y = 0$, choose transit). X_1 is the difference between the first variable (e.g. travel time) and X_2 is the difference between the second variable (e.g. travel cost) of the two alternatives. Y and X_1 are complete data, whereas X_2 is missing. MCAR does not depend on Y or X; it happens at complete random. MAR depends on Y only; it is missing when Y is not chosen ($Y = 0$). MNAR depends on X_2 itself; when X_2 is sorted, the values are missing for the first five rows.

Schafer and Graham [10], and Baraldi and Enders [1] point out that the most misunderstandings of MCAR, MAR, and MNAR are from general concepts of the meaning of random. In fact the MAR mechanism is not random at all, it presents systematic missingness related to other variables in analysis. To a statistician, a random phenomenon implies a probabilistic rather than deterministic property. On the other hand, to a psychologist, a random phenomenon suggests a predictable process that is extraneous to variables in a present study [10].

The goal of a statistical procedure is to make valid and efficient inferences on population of interest. Allison [4] suggested three general agreements of good missing data handling methods: minimize bias, maximum the use of available information, and yield good estimates of uncertainty.

TABLE I. ILLUSTRATION OF COMPLETE DATA AND THREE MISSING DATA MECHANISMS: MCAR, MAR, AND MNAR

Y	X_1 (Complete data)	X_2 (Complete data)	X_2		
			MCAR	MAR	MNAR
$Y_1 = 0$	X_{11}	X_{21}	X_{21}	-	-
$Y_2 = 1$	X_{12}	X_{22}	-	X_{22}	-
$Y_3 = 1$	X_{13}	X_{23}	X_{23}	X_{23}	-
$Y_4 = 0$	X_{14}	X_{24}	-	-	-
$Y_5 = 0$	X_{15}	X_{25}	X_{25}	-	-
$Y_6 = 0$	X_{16}	X_{26}	X_{26}	-	X_{26}
$Y_7 = 1$	X_{17}	X_{27}	-	X_{27}	X_{27}
$Y_8 = 0$	X_{18}	X_{28}	-	-	X_{28}
$Y_9 = 1$	X_{19}	X_{29}	X_{29}	X_{29}	X_{29}
…	...	...	...	...	...
Y_n	X_{1n}	X_{2n}	X_{2n}	X_{2n}	X_{2n}

Rubin [9] developed a framework of inference from available data. One of the most common traditional missing data techniques include deletion and single imputation approaches [3]. The modern approaches, maximum likelihood estimation and multiple imputation, are considered "*state of the art*" missing data techniques [10]. Most techniques assume missing data are MCAR, while some also assume MAR. However, typical missing data techniques do not assume MNAR. Every such techniques perform poorly for MNAR, although maximum likelihood estimation and multiple imputation tend to perform better than most traditional techniques [1].

III. MISSING DATA TECHNIQUES

The goal of a statistical procedure is to make valid and efficient inferences on population of interest. Allison [4] suggested three general agreements of good missing data handling methods: minimize bias, maximum the use of available information, and yield good estimates of uncertainty.

Researchers had employed a wide variety of techniques to deal with missing data; until Rubin [9] developed a framework of inference from available data. One of the most common traditional missing data techniques include deletion and single imputation approaches [3]. Schafer and Graham [10] defined imputation as the practice of filling in missing items. The modern approaches, maximum likelihood estimation and multiple imputation, are considered "state of the art" missing data techniques [10]. Most techniques assume

missing data are MCAR, while some also assume MAR. However, typical missing data techniques do not assume MNAR. Every such techniques perform poorly for MNAR, although maximum likelihood estimation and multiple imputation tend to perform better than most traditional techniques [1].

A. Deletion Methods

Among traditional missing data handling methods, one of the most widely used techniques is the deletion technique. It discards cases with one or more missing values. It is commonly known as listwise deletion (also called case deletion or complete-case analysis). The main advantage of this technique is its straightforwardness. It produces a simple complete data set, which in turn allows for the use of standard analysis techniques [1]. However, in most situations disadvantages are over advantages [11]. The total sample size can be reduced considerably, in particular for the high percentage of missing data and the large number of missing variables. Consequently, significant tests lead to a loss of power [4]. Additionally, listwise deletion assumes the missing data are MCAR; if this assumption is violated, the analyses will produce biased estimates [1].

B. Single Imputation Methods

Single imputation refers to traditional techniques that fills in missing data with plausible values [1]. There are a number of single imputation techniques. Imputation is potentially more efficient than deletion methods because no units are thrown away. As a result, the remaining units help prevent loss of statistical power. As the observed values hold useful information for predicting the missing values, the imputation process yield better accuracy [10]. As well as deletion methods, single imputation methods produce a complete data set, which can be analyzed by standard techniques. However, single imputation methods are difficult to implement, particularly in multivariate and some techniques can distort the distribution of the data. This paper focuses on common techniques including mean substitution, regression imputation, and stochastic regression imputation.

First, mean substitution or mean imputation substitutes missing values with an arithmetic mean of available values. This method is simple and commonly-used among imputation approaches, but it causes biased estimates [11]. Second, regression imputation, also called conditional mean imputation, estimates missing values from a regression equation of available data. Regression imputation performs much better than the first method, but still gives biased estimation of some parameters. In particular, the variance tends to be underestimated, which leads to the bias of any parameters that based on variances e.g. regression coefficients [4]. In addition, it is not recommended for an analysis of covariance or correlation because of their high correlation between imputed and imputing variables [10]. Finally, stochastic regression imputation is similar to linear regression but the estimated values are included a random error term drawn from a normal distribution with a zero mean and a variance estimated by a residual mean square. Stochastic regression imputation produces unbiased estimated parameters under both MCAR and MAR assumptions. Nevertheless, stochastic regression imputation is to guess about true values and is performed with no mechanism for correcting standard errors [1].

C. Maximum Likelihood Estimation

Maximum likelihood utilizes all available data, both complete and incomplete, to determine parameters that maximize probability of generating sample data. The estimation process uses a log likelihood function to quantify a standardized distance between observed data points and the parameters of interest. The main objective is to identify parameters that minimize such distance. Thorough descriptions of maximum likelihood estimation for missing data are in several sources [e.g. 4, 11, 12].

Maximum likelihood can handle missing data in a variety of situations [4]. Maximum likelihood produces estimates that have the following properties: consistency, asymptotic efficiency and asymptotic normality, when assumptions are met. Consistency means that estimated parameters are unbiased in large samples. Asymptotic efficiency implies that the estimates are close to efficient, i.e. minimal standard errors. Asymptotic normality means that researchers can use a normal approximation to predict confidence intervals and *p*-values. Importantly unlike single imputation, maximum likelihood can estimate accurately standard errors that totally account for the missing data [4].

Allison [13] claimed that maximum likelihood had benefit over multiple imputation in several ways. First, maximum likelihood is more efficient than multiple imputation. Second, maximum likelihood provides the same result, while multiple imputation produces a different result every time when researchers use it. Third, the implementation of multiple imputation requires several different decisions, each of which involves uncertainty, whereas maximum likelihood involves far fewer decisions. Forth, with multiple imputation, there is always a potential conflict between the imputation model and the analysis model, while there is no potential conflict in maximum likelihood because everything is done under one model.

However, maximum likelihood is based on a parametric model for a joint distribution of all variables with missing data. The drawback of maximum likelihood is calculation complexity. A specialized software is normally required to obtain the result. Additionally, large samples are needed to meet the property of maximum likelihood estimation. If sample size is small, the approximation may produce poor results [12].

D. Multiple Imputation Methods

Proposed by Rubin [14], multiple imputation generates a number of copies of data sets which consist of different imputed values, and analyzes each data set separately. The data set produces multiple sets of estimated parameters and standard errors which are combined into a single set of results. Reviews of multiple imputation have been published in various sources [e.g. 4, 11, 12].

Multiple imputation is an excellent alternative to impute missing data, which has properties that are approximately as good as maximum likelihood [4]. Similar to maximum likelihood, multiple imputation estimates are consistent, asymptotically normal and efficient. Researchers can produce the estimates close to asymptotical efficiency by increasing the number of imputations.

Multiple imputation has benefits over maximum likelihood in twofold [4]. First, multiple imputation can be applied to any type of data of a model. Second, it can be applied by conventional software packages rather than a special package. However, two key disadvantages of multiple imputation are that it generates different outcomes every time due to a random draw, and there are several ways to utilize multiple imputation leading to uncertainty and confusion.

IV. Methodology

Five popular missing data handling methods are compared including: listwise deletion, mean substitution, regression imputation, stochastic regression imputation and multiple imputation. Due to the requirement of a specific software to tackle missing data for a logistic regression using maximum likelihood estimation, this study employs multiple imputation. The estimates by multiple imputation are as good as maximum likelihood [4]. These five methods combining with three missing data mechanisms: MCAR, MAR, and MNAR are investigated.

To compare the results from different techniques, this study uses a simulation approach. A Monte Carlo simulation generates a number of samples from a population with a specified set of parameters. The simulation program generates 10, 20, 30, 40, 50, 100, 250, 500, 1000, 2500, and 5000 samples with 100 runs for each scenario. The main reason for using simulated data is to ensure that a distribution assumption is not violated.

To simplify the analysis and focus on the comparison of missing data handling techniques and missing data mechanisms, the model is a binary logistic regression (choosing between an auto and a transit) with two continuous independent variables: travel time and cost. The missing data are merely costs, whereas travel times are complete data. The simulated data are drawn from a bivariate normal population with mean times of auto and transit $\mu_{auto} = 80$ and $\mu_{transit} = 100$, standard deviations $\sigma_{auto} = 20$ and $\sigma_{transit} = 20$, and correlation $\rho = 0.3$. Cohen [15] suggests a small, medium, and strong correlation as a rule of the thumb are 0.1, 0.3, and 0.5, respectively. Although a coefficient of correlation of 0.7 may be considered weak in a physical experiment or an engineering context, the same value is considered very strong in the context of social science [16].

Other variables are auto cost and transit cost that are generated from a relationship with times plus a random term. A binary choice of an alternative is assigned according to the different of time and cost of alternatives. Then, the MCAR, MAR, and MNAR conditions are created. Finally, five missing data handling methods are performed. Because there are two independent variables in the model, the imputation equation consists of only two explanatory variables: a dependent variable (choose) and an independent variable (time). The number of imputation is set to 20 which was suggested as an effective number by Schafer and Graham [10] in several practical applications.

Saunders, Morrow-Howell [7] note that a large sample with a small percentage of missing values is not influenced to the same degree by data imputation methods as are smaller data sets. Although it is not clear to what extent missing data can be imputed, the literature suggests that 20 percent or less is acceptable [12]. To compare the outcome of different situations, the percentage of missing data is set to 10, 20, 30, 40, 50, 60, 70, and 80 percent, respectively. It could be argued that greater than 60 percent missing data is high. However, Schafer and Graham [10] claim that it is not difficult to find published papers with higher missing data. King, Honaker [17] report high percentage of deletion with serious implications for parameter bias and inefficiency.

TABLE II. Estimated Parameters from a Simulation with 20% Missing Data and 1,000 Samples

Estimates	Population parameters	Missing data techniques				
		LD	MS	RI	SI	MI
		MCAR simulation				
Mean diff. cost	6.265	6.262	6.262	6.264	6.264	6.269
		(0.722)	(0.722)	(0.902)	(0.928)	(0.677)
SD diff. cost	1.555	1.554	**1.390**	**1.514**	1.553	1.553
		(0.791)	**(0.000)**	**(0.000)**	(0.664)	(0.715)
Coefficient diff. time	-0.016	-0.016	**-0.032**	**-0.007**	**-0.023**	-0.016
		(0.801)	**(0.000)**	**(0.000)**	**(0.000)**	(0.900)
SD diff. time	0.006	0.007	0.005	0.007	0.006	0.007
Coefficient diff. cost	-0.637	-0.638	**-0.441**	**-0.808**	**-0.501**	-0.642
		(0.913)	**(0.000)**	**(0.000)**	**(0.000)**	(0.708)
SE diff. cost	0.098	0.110	0.090	0.113	0.096	0.110
Constant	4.350	4.359	**2.791**	**5.611**	**3.346**	4.386
		(0.930)	**(0.000)**	**(0.000)**	**(0.000)**	(0.731)
SE constant	0.729	0.816	0.643	0.836	0.713	0.814
Log Likelihood	-523	-419	-533	-517	-531	*n/a*
Rho square	0.21	0.21	0.19	0.22	0.20	*n/a*

		MAR simulation				
Mean diff. cost	6.265	**6.414**	**6.414**	6.266	**6.293**	**5.795**
		(0.000)	**(0.000)**	(0.910)	**(0.000)**	**(0.000)**
SD diff. cost	1.555	**1.575**	**1.408**	**1.524**	**1.572**	**1.434**
		(0.000)	**(0.000)**	**(0.000)**	**(0.001)**	**(0.000)**
Coefficient diff. time	-0.016	-0.017	**-0.027**	**-0.009**	**-0.022**	-0.017
		(0.614)	**(0.000)**	**(0.000)**	**(0.000)**	(0.277)
SD diff. time	0.006	0.007	0.005	0.007	0.006	0.007
Coefficient diff. cost	-0.637	-0.619	**-0.549**	**-0.753**	**-0.508**	-0.612
		(0.175)	**(0.000)**	**(0.000)**	**(0.000)**	(0.060)
SE diff. cost	0.098	0.107	0.093	0.110	0.096	0.106
Constant	4.350	**3.845**	**3.654**	**5.210**	**3.426**	4.174
		(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	(0.077)
SE constant	0.729	0.792	0.682	0.812	0.714	0.785
Log Likelihood	-523	-437	-527	-519	-531	*n/a*
Rho square	0.21	0.21	0.20	0.21	0.20	*n/a*
		MNAR simulation				
Mean diff. cost	6.265	**5.707**	**5.707**	**6.043**	**6.033**	**7.110**
		(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
SD diff. cost	1.555	**1.156**	**1.034**	**1.269**	**1.315**	**1.183**
		(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
Coefficient diff. time	-0.016	**-0.029**	**-0.048**	**-0.028**	**-0.038**	**-0.033**
		(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
SD diff. time	0.006	**0.007**	**0.004**	**0.007**	**0.006**	**0.007**
Coefficient diff. cost	-0.637	**-0.419**	**-0.117**	**-0.519**	**-0.276**	**-0.398**
		(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
SE diff. cost	0.098	**0.127**	**0.094**	**0.126**	**0.104**	**0.127**
Constant	4.350	**2.794**	**0.382**	**3.269**	**1.565**	**2.419**
		(0.000)	**(0.000)**	**(0.000)**	**(0.000)**	**(0.000)**
SE constant	0.729	0.891	0.596	0.881	0.724	0.890
Log Likelihood	-523	-412	-545	-536	-542	*n/a*
Rho square	0.21	0.12	0.18	0.19	0.18	*n/a*

Remark: LD = listwise deletion; MS = mean substitution; RI = regression imputation; SI = stochastic imputation; MI = multiple imputation; values in parenthesis is p-value, parameters in **bold** mean p-value < 0.001, H_0: an estimated parameter equals to the population, H_a: an estimated parameter is different from the population

V. Results and Discussion

A. An Explanatory Variable with Imputation of Missing Values

This section compares means and standard deviations of missing values after an imputation, i.e. impute missing some values of a cost. A case of 20 percent missing data, and 1,000 samples is presented. A large sample size is selected to ensure the efficiency of estimated parameters [18] whereas 20 percent missing data is picked for the reason that the literature suggests that 20 percent or less is acceptable [12].

As shown in Table II, a population mean of a cost (to make it short, different cost is referred as *cost*) is 6.265, whereas estimated means are 6.262-6.269 for MCAR missing data. Estimated means are 5.795-6.414 and 5.707-7.110 for MAR and MNAR, respectively. No evidence shows that five handling methods providing biased means (i.e. difference between the population mean and the estimated means) for MCAR, but it indicates that these values are different for MAR and MNAR apart from regression imputation for MAR. Population standard deviation of the cost is 1.555, while estimated standard deviations decrease to 1.390 and 1.514 for mean substitution and regression imputation, respectively, when the missing data is MCAR. As expected, the estimated standard deviation from the mean substitution is least among the five methods, following by that from the regression imputation. Listwise deletion, stochastic imputation, and multiple imputation provide no evidence that the estimated standard deviations are biased for MCAR. The estimated standard deviations are biased when the missing data is MAR and MNAR.

B. Estimated Parameters

Population coefficients for travel time (to make it short, different travel time is referred as *travel time* or *time*), cost and constant are -0.016, -0.637, and 4.350, respectively; estimated coefficient from listwise deletion and multiple imputation are unbiased when missing data is MCAR. On the contrary, estimates from mean substitution, regression imputation, and stochastic imputation are biased; they vary from -0.032 to -0.007, -0.808 to -0.441, and 2.791 to 5.611 for time, cost and constant, respectively. Thus, for MCAR, listwise deletion and multiple imputation provide no evidence that the coefficients are biased, while mean substitution, regression imputation and stochastic imputation produce biased coefficients.

When missing data is MAR, interestingly, there is no evidence proving that multiple imputation provides biased coefficients, although its mean and standard deviation differ from the population parameters. The population mean is 6.265, whereas the estimates are 5.795. The reason is the imputation equation has only two explanatory variables. The estimates should be improved if the model have more explanatory variables. Besides the multiple imputation, the rest four methods provide biased

coefficients. The estimates from listwise deletion, mean substitution, regression imputation, and stochastic imputation vary from -0.027 to -0.009, -0.549 to -0.753, and 3.426 to 5.210 for time, cost and constant, respectively. As we expect, the five methods cannot handling MNAR; the results shows strongly evidence that estimates are biased.

Finally, other statistical indicators are described. Log likelihood increases extensively by listwise deletion (from -523 to -419) because missing cases are removed from an analysis; as a result, the number of cases to be calculate log likelihood function are considerably decreased. Thus, one cannot carelessly conclude that the listwise deletion improves the model because of increasing in log likelihood index. In terms of MCAR and MAR, all methods provide estimated rho squares between 0.19 and 0.22, while a population value is 0.21. The estimated rho squares vary from 0.12 to 0.19, much less than the population value when the missing data is MNAR.

TABLE III. BIASED PARAMETERS WITH THE VARIATION OF SAMPLE SIZES AND PERCENT MISSING VALUES

Sample size	Percent missing values							
	10	**20**	**30**	**40**	**50**	**60**	**70**	**80**
	Listwise deletion							
10	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
20	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
30	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
40	*aab*	*aab*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
50	*aab*	*aab*	*bab*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
100	*aab*	*aad*	*abd*	*acd*	*acc*	*n/a*	*n/a*	*n/a*
250	*aac*	*abd*	*acd*	***add***	***add***	*n/a*	*n/a*	*n/a*
500	*aad*	***add***	***add***	***add***	***add***	*n/a*	*n/a*	*n/a*
1000	*abd*	***add***	***add***	***add***	***add***	*n/a*	*n/a*	*n/a*
2500	***add***	***add***	***add***	***add***	***add***	***add***	*n/a*	*n/a*
5000	***add***	***abd***	***add***	***add***	***add***	***add***	*n/a*	*n/a*
	Multiple imputation							
10	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
20	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
30	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
40	*aab*	*aac*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
50	*aac*	*aad*	*aad*	*n/a*	*n/a*	*n/a*	*n/a*	*n/a*
100	*aac*	*aad*	*aad*	*abd*	*aad*	*n/a*	*n/a*	*n/a*
250	*aad*	*aad*	*aad*	*abd*	*aad*	*n/a*	*n/a*	*n/a*
500	*aad*	*aad*	*abd*	*abd*	*abd*	*n/a*	*n/a*	*n/a*
1000	*Aad*	*abd*	*acd*	***bdd***	***add***	*n/a*	*n/a*	*n/a*
2500	*Aad*	*abd*	***add***	***add***	***cdd***	***bdd***	*n/a*	*n/a*
5000	*Aad*	*acd*	***add***	***add***	***add***	***bdd***	*n/a*	*n/a*

Remark: a = p-value ≥ 0.1, b = p-value < 0.1 to 0.01, c = p-value < 0.01 to 0.001, d = p-value < 0.001; the value in *italic* e.g. *aaa* means that the first is MCAR, the second is MAR, and the third is MNAR; a parameter in **bold** means the first, or second value has p-value < 0.001, H_0: an estimated parameter equals to the population, H_a: an estimated parameter is different from the population

C. Effects of Sample Sizes and Percent Missing Data

This section focuses on the sample size and percent missing data for listwise deletion and multiple imputation because they can cope with MCAR and multiple imputation is only technique that is able to handle MAR. Percent missing data and sample size and are closely relationship. As shown in Table III, although high percentage of missing data up to 60%, the simulation shows that estimated parameters are unbiased when the sample sizes are greater than 1000 samples for MCAR using listwise deletion and multiple imputation.

Table III illustrates biased parameters by the number of sample sizes, and percent missing values for listwise deletion and multiple imputation. The values above the diagonal line show an unavailable sample size and percent missing values, while the bold values indicate that at least one parameter for MCAR and MAR has p-value < 0.001 (indicated by *d* in the first or second value, respectively).

VI. CONCLUSION

Allison [4] states that "Somewhat surprisingly, listwise deletion is very robust to violations of MCAR (or even MAR) for predictor variables in a regression analysis. Specifically, so long as missingness on the predictors does not depend on the dependent variable, listwise deletion will yield approximately unbiased estimates of regression coefficients [12]. And this holds for virtually any kind of regression – linear, logistic, Poisson, Cox, etc." Nevertheless, our study has found evidence that listwise deletion is not as robust as we first expected. For MAR we have found that listwise deletion produces biased coefficients when missing data greater than 20 percent and sample size more than 500 samples; however,

there is no evidence to show that on listwise deletion provide a biased coefficient on MCAR.

There are a number of limitations on this study. First, the model is a binary logistic regression with two continuous independent variables. There is no a categorical variable to be considered in this analysis. Second, missing values occur only to one variable, but in a real-world situation the missing values may occur to multiple variables. Third, we consider five well-known techniques; however there are other methods e.g. new developed techniques to tackle MNAR.

Supporting the mention by Saunders, Morrow-Howell [7] in section 3, to maintain the unbiased parameters, the sample size rises when the percentage of missing data increase. Enhancing a practice suggested by Little and Rubin [12] that 20 percent missing data or less is acceptable. We would recommend using multiple imputation with the maximum 20 percent missing data for sample size at least 40 samples in MAR condition.

The final verdict on multiple imputation cannot be made in this article because of the limitations. We are able to make a conclusion of some fact findings in this study. First, for a logistic regression estimation, of five missing data handling methods, multiple imputation performs well on both MCAR and MAR. Second, there is no evidence to indicate that listwise deletion and multiple imputation produce biased coefficients for a logistic regression on MCAR. Finally, all these techniques cannot handle MNAR.

The simulated data in this study can be classified as reveal preference data according to Louviere, Hensher [19]. It cannot be simulated as stated preference data by missing some independent variables because such variables are predefined by researchers. The key question becomes whether one should impute the dependent variable itself. Allison [4] points out that if the data are MAR and there are no auxiliary variables; the answer is no. Imputation of dependent variable itself will increase sampling variability [20].

For future research, other missing data handling techniques, particular one dealing with MNAR, could be examined as well as a categorical variable and a mixed continuous and categorical variable. One can determine an effect of missing data when the data are conducted by stated preference survey. It means that the missing data is a dependent variable itself.

Finally, we totally agree with Saunders, Morrow-Howell [7] who point out that the trade-off between an accurate result and time-consuming to learn and use the multiple imputation is well worth the investment.

REFERENCES

[1] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *Journal of School Psychology*, vol. 48, no. 1, pp. 5-37, 2010.

[2] T. E. Raghunathan, "What do we do with missing data? Some options for analysis of incomplete data," *Annu. Rev. Public Health*, vol. 25, pp. 99-117, 2004.

[3] J. L. Peugh and C. K. Enders, "Missing data in educational research: A review of reporting practices and suggestions for improvement," *Review of educational research*, vol. 74, no. 4, pp. 525-556, 2004.

[4] P. D. Allison, *Missing Data*, Sage Publications, vol. 136. 2001.

[5] L. Wilkinson, "Statistical methods in psychology journals: Guidelines and explanations," *American Psychologist*, vol. 54, no. 8, p. 594, 1999.

[6] A. J. Figueredo, *et al*., "Multivariate modeling of missing data within and across assessment waves," *Addiction*, vol. 95, pp. 361-380, 2000.

[7] J. A. Saunders, *et al*., "Imputing missing data: A comparison of methods for social work researchers," *Social Work Research*, vol. 30, no. 1, pp. 19-31, 2006.

[8] A. Davey and J. Savla, *Statistical Power Analysis with Missing Data: A Structural Equation Modeling Approach*, Routledge, 2010.

[9] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581-592, 1976.

[10] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art," *Psychological Methods*, vol. 7, no. 2, p. 147, 2002.

[11] C. K. Enders, *Applied Missing Data Analysis*, Guilford Publications, 2010.

[12] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed., New York: John Wiley & Sons, 2002.

[13] P. D. Allison, "Handling missing data by maximum likelihood," in *Proc. SAS Global Forum*, 2012.

[14] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley & Sons, 1987.

[15] J. Cohen, "Quantitative methods in psychology: A power primer," *Psychological Bulletin*, vol. 112, no. 1, p. 155, 1992.

[16] J. R. Cheema, "Some general guidelines for choosing missing data handling methods in educational research," *Journal of Modern Applied Statistical Methods*, vol. 13, no. 2, p. 3, 2014.

[17] G. King, *et al*., "Analyzing incomplete political science data: An alternative algorithm for multiple imputation," in *American Political Science Association*, Cambridge University Press, 2001.

[18] C. Dougherty, *Introduction to Econometrics*, Oxford University Press, 2007.

[19] J. J. Louviere, D. A. Hensher, and J. D. Swait, *Stated Choice Methods: Analysis and Applications*, Cambridge University Press, 2000.

[20] R. J. Little, "Regression with missing X's: A review," *Journal of the American Statistical Association*, vol. 87, no. 420, pp. 1227-1237, 1992.

Sutthipong Meeyai was born in Thailand, in 1966. He obtained his BSc in Civil Engineering in 1989 from Chulachomklao Royal Military Academy, MEng in Transportation Engineering in 1993 from Chiang Mai University, MSc in Logistics Management in 2007 from Chulalongkorn University, Thailand.

He has worked as a transportation planning and traffic engineering consultant. He also has experiences in transportation planning using strategic transport model and discrete choice models application for various topics such as marketing, transportation etc. He has published papers in a field of transportation e.g. Transportmetrica A, and marketing e.g. Journal of International Conference on Marketing and Business Development. Currently, he is a lecturer in the School of Transportation Engineering, Suranaree University of Technology, Thailand. Mr. Sutthipong Meeyai is a memberships in Professional Engineering Board of Thailand.