A Method to Identify Missing Data Mechanism in Incomplete Dataset

Sonam TSHERING[†], Takeo OKAZAKI[††] AND Satoshi ENDO[†††]

Faculty of Information Engineering, University of the Ryukyus, Okinawa 903-0213, Japan

Summary
This paper proposes a sequential method of identifying missing data mechanism in incomplete dataset. A simulation study in the context of GNH dataset is designed to validate the proposed method. The dataset is explored for latent factors and a model is constructed from the discovery. Using this model, pseudorandom data are generated for a simulation study. Specific missingness probability is invoked in pseudorandom data to create respective MCAR, MAR and MNAR data. The proposed method is validated and tested for robustness by employing a specific statistical test for each of MCAR, MAR and MNAR data.

Key words:
Simulation, Probability, Exploratory, Ignorability, Test.

1. Introduction

The issue of missing data is common in many applied fields, particularly in medical and social studies involving survey questionnaires [1]. The reasons for missing data are manifold. They may be missing because of malfunctioning equipment, unfavorable weather conditions, and data entry errors [2]. Survey questionnaires also suffer from missing data when respondents refuse, or do not know the answer to or accidentally skip an item. This ubiquitous problem complicates statistical data analysis [3].

Traditionally, a simple solution that is usually a default for many statistical packages is employed to compensate for missing data. The solution is if a case has any missing data for any of the variables in the analysis, then the case was simply excluded from the analysis. This leaves a data set without any missing data and can be analyzed by any conventional methods. This strategy is commonly known as list-wise deletion or case-wise deletion, and also goes by the name of complete case analysis [4]. This strategy has many attractive properties but the main apparent drawback is it excludes a large fraction of original sample and thus regarded as "not generally recommended" [5].

In light of this drawback many alternative methods, namely pairwise deletion, single imputation, arithmetic mean imputation, regression imputation, stochastic regression imputation, last observation carried forward, etc. have been proposed but unfortunately most of them have little value or are inferior to the method of list-wise deletion [6, 7, 8, 9, 10]. The good news however is that statisticians have developed two modern and novel approaches. They are maximum likelihood method and multiple imputation method. They offer substantial improvement over list-wise deletion and widely regarded as state-of-the-art methodologies [11].

Although these two methods of missing data analysis have reasonably good statistical properties, like any other methods the reliability and accuracy largely depend on certain assumptions. Both the methods assume that the missing data are missing completely at random (MCAR) or at least missing at random (MAR). MCAR and MAR are considered "ignorable" meaning modeling the missing data mechanism as part of the estimation process is unnecessary. Missing data other than MCAR and MAR is missing not at random (MNAR). MNAR is considered "non-ignorable" [12]. This non-ignorability results in inconsistent parameters and inflated standard errors. As a consequence it could have a detrimental effect on the results or estimates. This calls for modeling the missing data mechanism as part of the estimation process to get good estimates of the parameters of interest.

Without the knowledge of missing data mechanism it is impractical to identify missing data as "ignorable" or "non-ignorable." And without this categorization, an appropriate analysis method cannot be recommended for missing data. This makes the missing data mechanism identification method all the more important [13].

Gross National Happiness (GNH) is the Bhutan government's development concept coined in an attempt to invent an indicator that assesses the quality of life in more holistic and psychological terms than Gross Domestic Product (GDP). To this end the government has taken initiatives of developing GNH indicators by designing questionnaires and conducting surveys. The survey questionnaire covers a wide range of data on multiple domains of well-being, namely psychological well-being, health, education, culture, time use, good governance, community vitality, ecological diversity and resilience, and living standards. A pilot survey was first conducted in 2006 with 350 respondents. A full-fledged survey was later conducted in 2007 and 2010 [14]. As with any survey, GNH survey is also riddled with missing data problem for various reasons.

Manuscript received March 5, 2013
Manuscript revised March 20, 2013

2. Missing Data Mechanism

Rubin [15] presents the standard definition of missing data mechanism that are classified into three categories, namely MCAR, MAR, and MNAR. Suppose (Y, R) is a data matrix with complete data; R being the missing data indicator matrix is defined as:

$$R = \begin{cases} 1, \text{ if } Y \text{ is observed} \\ 0, \text{ otherwise} \end{cases}$$

Y^o and Y^m are the respective observed and missing parts of Y.

An observation is said to be MCAR if the missingness is independent of all observed and unobserved values.

$$P(R \mid Y) = P(R) \text{ for all } Y$$
$$R \text{ is independent of both } Y^{(o)} \text{ and } Y^{(m)} \quad (1)$$

An observation is said to be MAR if the missingness is independent of unobserved values but dependent on the observed values.

$$P(R \mid Y) = P(R \mid Y^{(o)}) \text{ for all } Y$$
$$R \text{ is independent of } Y^{(m)} \quad (2)$$

MNAR is a missingness mechanism that is neither MCAR nor MAR. It occurs when the missingness depends not only on observed values but also on the unobserved values.

$$P(R \mid Y) \text{ depends on } Y^{(m)} \quad (3)$$

2.1 Traditional Method of Identifying Missing Data Mechanism

Missing or incomplete data analysis methods are being developed for specific missing data mechanism in mind. As a result these methods produce biased results when they are not applied according to the missing data mechanism. For instance, complete-case analysis will be unbiased only if the data are MCAR. Likewise only mixed-effects models are identified suitable for MAR data. And MNAR data require more sophisticated analysis methods [16]. There are a number of hypothesis tests that can be employed to test the assumption of MCAR.

Dixon's [17] MCAR test uses a series of independent $t-test$ to compare the means of complete and incomplete cases. An insignificant $t-statistic$ shows evidence in favor of the null hypothesis–that missing data are indeed MCAR. On the other hand, a significant $t-statistic$ shows evidence against the null hypothesis–that the missing data are not MCAR.

Alternatively, Little [18] developed a test based on means under different missing data patterns. The test statistic is a weighted sum of the standardized differences between the subgroup means and the grand means:

$$d^2 = \sum_{j=1}^{j} n_j \left(\bar{\mu}_j - \mu_j^{(ML)} \right)^T \hat{\Sigma}_j - 1 \left(\bar{\mu}_j - \mu_j^{(ML)} \right),$$

Where n_j is the number of cases in missing data pattern j; $\hat{\mu}_j$ contains the variable means for the cases in missing data pattern j; $\hat{\mu}_j^{(ML)}$ contains maximum likelihood estimates of the grand means; and $\hat{\Sigma}_j$ is the maximum likelihood estimate of the covariance matrix. The drawback of $t-test$ are twofold: Generating the test statistic is very cumbersome for multivariate data and the test doesn't take correlation between the variables into account. Although the Little's global test is the widely used MCAR test, it is not without flaws. First, it doesn't identify the specific variable that violates MCAR assumption. Second, the test assumes that the missing data patterns share a common covariance matrix. And finally, past studies suggest that the test suffer from low power and consequently it has a propensity to produce Type II errors that can lead to a false sense of security about the missing data mechanism [19].

A hypothesis test for MAR was proposed by Diggle [20]. He performed a test for each time point t_j that the patients who drop out at t_{j+1} are a random sample from all patients who continued the study up till t_j. As for whether *p-values* are uniformly distributed under the null hypothesis, he recommended Kolmogorov test to decide if the resulting *p-values*, p_j behave like a random sample from a uniform distribution. The better way of combining *p-values* of independent statistical tests was proposed by Fisher [21]. He proposed the use of

$$C = -2\sum_{j=1}^{k} \ln p_j \text{ as overall test criterion.}$$

The drawbacks of these tests are: First, their null distributions depend on the independence of the tests to be combined; Second, the Kolmogorov test lacks power since the number of *p-values* are small; And third, the approaches to combine *p-values* do not take into account the detailed characteristics of the data [22]. A non-parametric test for random dropouts that don't suffer from these pitfalls was later proposed by the same authors [23].

Fairclough outlined an approach to identify MNAR mechanism using logistic regression. In it the dataset is restricted to responders only and regarded the reminder-response as missing. If the current score is significant in the logistic model after having adjusted for covariates and previously observed score, then there is evidence of MNAR data. However the prerequisite for this approach is that all responder data must be used and that the true value of the data regarded as missing must be known, otherwise making an inference whether MNAR or not is impossible.

3. Proposed Method of Identifying Missing Data Mechanism

Identifying MCAR data: From equation (1), for an observation to be MCAR, the homogeneity of means and covariances between the observed and unobserved is desired. Hinging on this condition, let Y be data matrix with complete data and R be the missing data indicator, such that:

$$R = \begin{cases} 1, \text{ if } Y \text{ is observed} \\ 0, \text{ otherwise} \end{cases} \quad (4)$$

Based on this binary missing data indicator we assess mean, μ and covariance, Σ differences between the observed and unobserved values. We then test if:

$$\begin{cases} H_0 : \mu_{R=1} = \mu_{R=0} \\ H_a : \text{ otherwise} \end{cases}, \text{ and}$$

$$\begin{cases} H_0 : \Sigma_{R=1} = \Sigma_{R=0} \\ H_a : \text{ otherwise} \end{cases}$$

Acceptance of H_0 would mean no significant difference between the groups both in terms of means and covariances, thereby confirming that an observation is MCAR. To perform this task the method of analysis of variance with Wilks test is proposed. Inference as to whether MCAR or not is drawn from Wilks test and $F-value.$ Wilks test demonstrates the amount of variance accounted for in the response variable by the explanatory variables.

$$\text{Wilk's lambda} = \Lambda = \frac{|\mathrm{E}|}{|\mathrm{H+E}|} = \prod_{i=1}^{q} \frac{1}{1+\lambda_i} \text{ and}$$

$F-value$ takes into account the covariance of the variables apart from its similarity to the univariate $F-value.$

$$F = \frac{MS_h}{MS_e}.$$

Identifying MAR data: According to equation (2), an observation is MAR if the probability of missingness depends on observed values but not on unobserved values. Premised on this definition, let x_i be linear predictors, we have missing data indicator, R from equation (4) that takes the value either 0 or 1. So we fit logistic regression model as:

$$P(R=1) = \log it^{-1}(X_i\beta).$$

A statistically significant association between missingness indicator and observed data reveals evidence for MAR data.

Identifying MNAR data: From equation (3), an observation is MNAR if the probability of missingness depends not only on observed values but also on unobserved values. Mathematically:

$$P(R \mid Y) \text{ depends on } \mathrm{Y}^{(\mathrm{m})}$$

The joint distribution of the data and the probability of missingness is:

$$P(Y, R)$$

Joint distribution can be factored into the product of two component distributions:

$$P(Y, R) = P(R \mid Y)P(Y)$$

This is a two-part model that combines the substantive regression model with an additional regression equation that predicts response probabilities. First, substantive regression model:

$$P(R=1) = \log it^{-1}(x_i\beta) \quad (5)$$

Then, the propensity for missing data on the outcome variable as a normally distributed latent variable:

$$R^* = \gamma_0 + \gamma_i x_i + +\zeta \quad (6)$$

With independent error, ζ that has the logistic probability distribution which is defined so that:

$$\Pr(\zeta < x) = \log it^{-1}(x) \text{ for all } x.$$

Thus, equation (5) and (6) are equivalent. Latent R^* values are themselves completely missing. So, we use the binary missing data indicator R as the dependent variable. In effect, R serves as a manifest indicator for R^*, such that the cases that score above some threshold on R^* have complete data $(R=1)$, and cases that fall below the threshold have missing values $(R=0)$:

$$R = \begin{cases} 1 \text{ if } R^* > 0 \\ 0 \text{ if } R^* < 0 \end{cases}$$

Hence the model expresses the predicted probability of a complete response as:

$$P(R = 1 \mid x^*) = \Phi[\gamma_0 + \gamma_i x_i]$$

Where Φ is the cumulative normal distribution function.

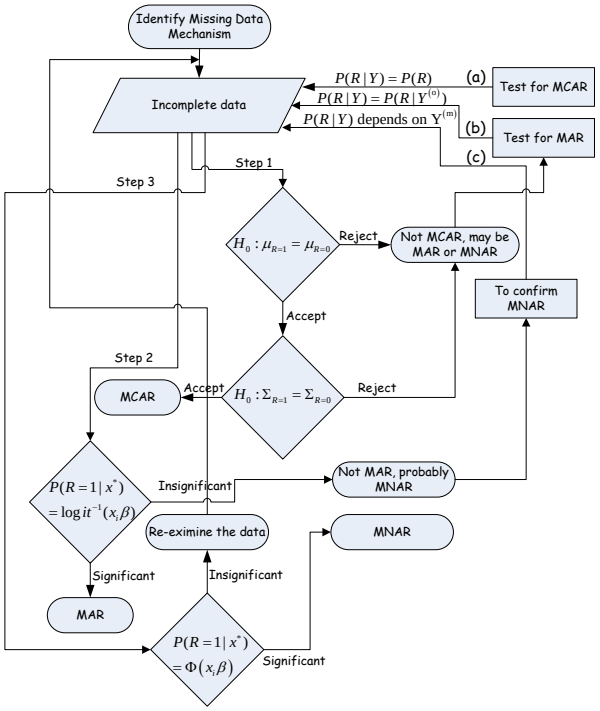


Fig. 1 A flow chart of missing data mechanism identification method

The flowchart consists of three steps: One, in which an incomplete data will be tested for MCAR by employing a hypothesis test to see if the means and covariances of the observed and unobserved groups are homogeneous. Acceptance of the means' null hypothesis will lead to further testing of covariances' null hypothesis while the rejection will terminate MCAR testing, and would recommend performing MAR test. Acceptance of the null hypothesis of the covariance test would confirm that the data is MCAR.

Second, following the MCAR test the incomplete data will be tested for MAR. Significant coefficients of logistic regression would provide evidence in favor of MAR data while insignificant coefficients are by default MNAR data. This is so because MNAR is a missingness mechanism that is neither MCAR nor MAR. Furthermore, a latent variable model would reconfirm it.

4. Simulation Study

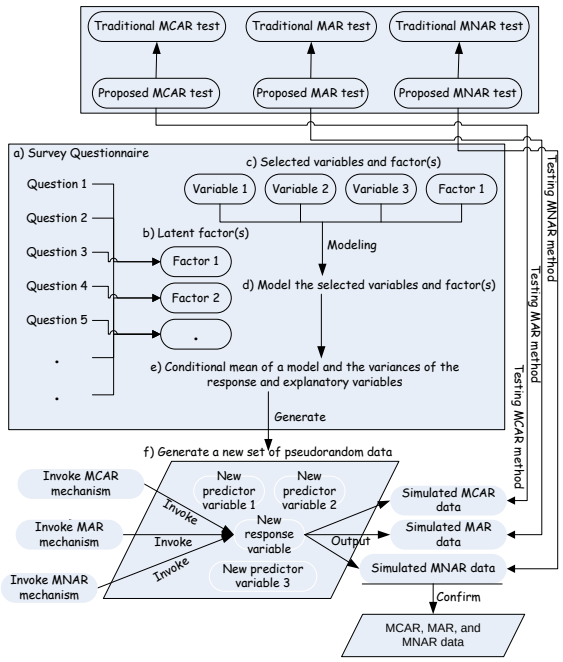


Fig. 2 Outline sketch of a simulation design

Simulation study begins with the exploration of GNH survey data set (a). Then extract hypothetical latent factors (b) and select a pool of variables closely associated with latent factor(s) (c). With the extracted latent factor as a response variable we model the selected variables (d). Then obtain the conditional mean of the model and the variances of the response and explanatory variables (e). Using the obtained conditional mean and the variances we generate a new set of pseudorandom data (f). Next invoke MCAR, MAR, and MNAR missingness probability in the response variable and as a result, output simulated MCAR, MAR, and MNAR data. Apply the proposed methods to respective simulated missing data to test the validity, efficiency and robustness of the method.

4.1 Dataset Exploration

Simulation study explored only a part of GNH survey that pertains to psychological domain [14]. That too, of the four hypothetical indicators and 19 manifest variables, only one indicator and the corresponding five manifest variables are considered. The indicator variable are, namely 'life satisfaction' and the variables are: 'Satisfaction with health,' 'satisfaction with standard of living,' 'satisfaction with occupation,' 'satisfaction with family relationship,' 'satisfactions with work life balance.' All the variables are measured on a five-point satisfaction scale, 1 being the lowest and 5 the highest.

Let x_1, x_2, x_3, x_4, x_5 represent five manifest variables assumed to be linked to f_1, a latent indicator. Performing factor analysis tests the assumption if single factor is adequate or not and that if manifest variables are closely associated with it or not.

Table 1: Test for number of factors

Test of the hypothesis that 1 factor is sufficient.
The chi square statistics is 18.24 on 5 degrees of freedom
The p-value is 0.00266

Since a single factor is adequate for the data, a factor analysis model is built thus:

$$x_1 = \lambda_1 f_1 + u_1$$
$$x_2 = \lambda_2 f_1 + u_2$$
$$x_3 = \lambda_3 f_1 + u_3$$
$$x_4 = \lambda_4 f_1 + u_4$$
$$x_5 = \lambda_5 f_1 + u_5$$

λ is a weight or factor loading on the common factor. The factor loadings are calculated as:

Table 2: Factor loadings

Variable	$\lambda_i f_1$
x_1	0.859
x_2	0.690
x_3	-0.225
x_4	0.958
x_5	-0.521

Only the significant variables $x_1, x_2,$ and x_4 that have high loadings on f_1 are included in the simulation study.

4.2 Modeling

A regression model is fitted as the response variable (f_1) is a continuous variable:

$$f_1 = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4$$

Table 3: Parameter estimates of the fitted model

Variable	**Parameter est.**	**Std. err.**	**p-value**
Intercept	-2.274	0.003	< 2.2e-16 ***
x_1	0.006	0.001	3.425e-10 ***
x_2	0.012	0.001	9.858e-15 ***
x_4	0.638	0.001	< 2.2e-16 ***

The estimates of the model are used to obtain the conditional mean $E[f_1]$ of the response variable which, in turn, is used to generate pseudorandom data.

4.3 Data Generation and Invoking Missingness Probability

Using the conditional mean, $E[f_1]$ and the variance, σ_r^2 of the response variable, a new response variable f_1^* is generated from a normal distribution, $f_1^* \sim N(E[f_1], \sigma_r^2)$. In a similar fashion three more variables, x_1^*, x_2^*, x_4^* are generated using the $E[f_1]$ and the variances of the three select manifest variables. As a result we have a set of new data $\left[x_1^*, x_2^*, x_4^*, f_1^*\right]$. f_1^* is used to create a binary missing data indicator:

$$R_{ij} = \begin{cases} 1 \text{ if } f_{1ij}^* \text{ is observed} \\ 0 \text{ otherwise} \end{cases} \quad (7)$$

A missing data probability π is then invoked in f_1^* according to the standard definition of missing data mechanism as presented in equation (1), (2), and (3):

$$\pi(x) = \exp it(\varphi_0) \text{ for MCAR} \quad (8)$$

$$\pi(\text{x}) = \text{expit}(\varphi_0 + \varphi_1 x_1^* + \varphi_2 x_2^* + \varphi_3 x_4^*) \text{ for MAR} \quad (9)$$

$$\pi(\text{x}) = \text{expit}(\varphi_0 + \varphi_1 x_1^* + \varphi_2 x_2^* + \varphi_3 x_4^* + \varphi_4 f_1^*) \text{ for MNAR} \quad (10)$$

$$\text{Where } \exp it(x) = \frac{\exp(x)}{1+\exp(x)}$$

φ_i is used to produce a specific level of missing data. We set it at 30% in all simulations. Three new response variables with different missing data pattern are generated: f_{1MCAR}^* generated by invoking equation (8), f_{1MAR}^* from equation (9), and f_{1MNAR}^* from equation (10). The final data set $\left[x_1^*, x_2^*, x_4^*, f_1^*, f_{1MCAR}^*, f_{1MAR}^*, f_{1MNAR}^*\right]$ is used to test the validity, efficiency and the robustness of the proposed method.

Table 4: Values of coefficient used in the missing data mechanism models to set 30% missingness probability

	Missingness level 30%				
Mechanism	φ_0	φ_1	φ_2	φ_3	φ_4
MCAR	0.89	-	-	-	-
MAR	6.09	4.2	5.3	1.002	-
MNAR	0.01	3.04	4.03	0.1	5.91

4.4 Results of Simulation Study

These simulations are carried out at $\pi = 30\%$, $\alpha = 0.05$, $N = 350$, and the number of runs =1000. Simulation parameters are also the same for MCAR, MAR and MNAR test.

4.4.1 MCAR Testing

Table 5: MCAR multivariate test with R_{MCAR} as a factor

	Wilks	Approx. F	P-value	Inference
as.factor (R_{MCAR})	0.992	0.931	0.426	MCAR
Residuals				

The critical value of $F_{crit}(3,346) = 2.631$ at $\alpha = 0.05$. Since $F = 0.931 < 2.631$ and $\Lambda = 0.992 < 2.631$, it is insignificant at 5% significance level. So, we accept the H_0, deducing that there's no significant variation between the groups. The $p-value$ confirms it. Hence the model is MCAR.

Table 6: MCAR multivariate test with R_{MAR} as a factor

	Wilks	Approx. F	P-value	Inference
as.factor (R_{MAR})	0.978	2.607	2.2e-16 ***	Not MCAR
Residuals				

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

As expected when MCAR test is wrongly applied to a simulated MAR data, it infers that the data is not MCAR. This also means the MCAR test is not prone to Type II errors.

Table 7: MCAR multivariate test with R_{MNAR} as a factor

	Wilks	Approx. F	P-value	Inference
as.factor (R_{MAR})	0.917	10.482	1.288e-06 ***	Not MCAR
Residuals				

Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Here too when MCAR test is wrongly applied to a simulated MNAR data, it infers that the data is not MCAR.

Table 8: MCAR simulation result

Simulation I		Simulation II		Simulation III	
Correct ratio (p>a)	Incorrect ratio (p<a)	Incorrect ratio (p<a)	Correct ratio (p>a)	Incorrect ratio (p<a)	Correct ratio (p>a)
0.948	0.052	1	0	1	0

Simulation I: MCAR test to MCAR data; simulation II: MCAR test to MAR data; and simulation III: MCAR test to MNAR data.

4.4.2 MAR Testing

Table 9: MAR test with R_{MAR} as responsible variable

	Estimate	Std. Error	P-value	Inference
x_1^*	-4.738	1.136	5.03e-05 ***	MAR
x_2^*	-6.335	1.480	2.84e-05 ***	MAR
x_4^*	-1.056	0.282	0.000238 ***	MAR

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The probability of missingness significantly depends on the observed variables when MAR test is applied to MAR data.

Table 10: MAR test with R_{MCAR} as responsible variable

	Estimate	Std. Error	P-value	Inference
x_1^*	-0.072	0.060	0.236	Not MAR
x_2^*	0.165	0.105	0.111	Not MAR
x_4^*	-0.056	0.050	0.270	Not MAR

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

There's no significant association between the probability of missingness and the observed variables when MAR test is applied to MCAR data.

Table 11: MAR test with R_{MNAR} as responsible variable

	Estimate	Std. Error	P-value	Inference
x_1^*	-2.753	0.555	7.17e-07 ***	MAR or MNAR
x_2^*	-3.717	0.847	1.14e-05 ***	MAR or MNAR
x_4^*	-0.141	0.144	0.325	Not MAR

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

There's also no significant association between the missingness probability and the observed variables when MAR test is applied to MNAR variables.

Table 12: MAR simulation result

Var	Simulation IV		Simulation V		Simulation VI	
	Correct ratio (p<a)	Incorrect ratio (p>a)	Inorrect ratio (p>a)	Correct ratio (p<a)	Incorrect ratio (p>a)	Correct ratio (p<a)
x_1^*	0.992	0.008	0.944	0.056	0	1
x_2^*	0.992	0.008	0.958	0.042	0	1
x_4^*	0.991	0.009	0.947	0.053	0.754	0.246

Simulation VI: MAR test to MAR data; simulation V: MAR test to MCAR data; and simulation VI: MAR test to MNAR data.

4.4.3 MNAR Testing

Table 13: MNAR test with R_{MNAR} as responsible variable

	Estimate	Std. Error	P-value	Inference
x_1^*	-1.464	0.265	3.18e-08 ***	MNAR
x_2^*	-1.983	0.423	2.68e-06 ***	MNAR
x_4^*	-0.064	0.078	0.409	Not MNAR

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Only two variables are found significantly associated with a latent variable.

Table 14: MNAR test with R_{MAR} as responsible variable

	Estimate	Std. Error	P-value	Inference
x_1^*	-2.723	0.615	9.69e-06 ***	Not MNAR
x_2^*	-3.610	0.787	4.44e-06 ***	Not MNAR
x_4^*	-0.614	0.155	7.07e-05 ***	Not MNAR

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

MNAR data are those whose missingness probability depends not only on observed data but also on unobserved data. So, it is no surprise that there is strong association between the two when MNAR test is applied to MAR data.

Table 15: MNAR test with R_{MCAR} as responsible variable

	Estimate	Std. Error	P-value	Inference
x_1^*	-0.043	0.036	0.236	Not MNAR
x_2^*	0.102	0.063	0.103	Not MNAR
x_4^*	-0.034	0.030	0.262	Not MNAR

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

None of the variables have significant association with the missingness probability when MNAR test is applied to MCAR data.

Table 16: MNAR simulation result

Var	Simulation VII		Simulation VIII		Simulation XI	
	Correct ratio (p<a)	Incorrect ratio (p>a)	Incorrect ratio (p>a)	Correct ratio (p<a)	Incorrect ratio (p<a)	Correct ratio (p>a)
x_1^*	1	0	0.945	0.055	0.995	0.005
x_2^*	1	0	0.954	0.046	0.993	0.007
x_4^*	0.246	0.754	0.945	0.055	0.993	0.007

Simulation VII: MNAR test to MNAR data; simulation VIII: MNAR test to MCAR data; and simulation IX: MNAR test to MAR data.

4.5 Conclusion

MNAR test clearly identifies MNAR data from MCAR data but it doesn't crisply distinguishes MNAR data from MAR data. There is an overlap between the two and some

theoretical background knowledge is warranted to make the distinction between the two. Although the simulation results authenticate the adequacy of the each proposed method in terms of both power and efficiency, there is still scope for further exploration. For an instance, for MCAR testing, one may further look into how introducing correlation between dependent variables and changing the covariance structure would affect the empirical type I error rate. Likewise, for MNAR testing in longitudinal studies we recommend collecting responder/reminder data to reconfirm the inference drawn.

5. Application of the Proposed Method to GNH Dataset

The proposed method of identifying missing data mechanism in incomplete dataset is applied to GNH survey dataset that is riddled with missing values problem.

Table 17: Result of application of the proposed method to the variables of 'life satisfaction' indicator of 'psychological well-being' domain of GNH

Variable	Inference
Satisfaction with health	MCAR
Satisfaction with standard of living	MNAR
Satisfaction with occupation	MAR
Satisfaction with family relationship	MCAR
Satisfaction with work life balance	MCAR

Table 18: Data description of 'life satisfaction' indicator

Variable	Question using a five-point Likert item (strongly disagree; disagree; neither agree nor disagree; agree; strongly agree)	Missing ratio
Satisfaction with health	How satisfied are you with your health?	113/350
Satisfaction with standard living	How satisfied are you with your standard of living?	98/350
Satisfaction with occupation	How satisfied are tour with your major occupation?	98/350
Satisfaction family relationship	How satisfied are you with your family relationship?	103/350
Satisfaction with work balance	How satisfied are you with your work life balance?	56/350

The proposed method can be applied to incomplete dataset at two levels. One, to individual variables; and two, to models. Both are applicable to GNH and longitudinal studies.

References

[1] Finch W. H. Imputation methods for missing categorical questionnaire data: a comparison of approaches. *Journal of Data Sciences*, 8 (2010), 361-378.

[2] Fitzmaurice G. Missing data: implications for analysis. *Nutrition*, 24 (2008), 200-202.

[3] Howell D. C. "Treatment of missing data." The University of Vermont. 2009. http://www.uvm.edu/~dhowell/StatPages/MoreStuff/Missing_Data/Missing.html. Accessed 12 January 2012.

[4] Myers W. R. Handling missing data in clinical trials: an overview. *Drug Information Journal*, 34 (2000), 525-533.

[5] Diggle P. J., Heagerty P., Liang, K. Y. et al. *Analysis of Longitudinal Data.* 2nd ed.. Oxford University Press Inc., (2002).

[6] Schafer J. L. *Analysis of Incomplete Multivariate Data*. Chapman and Hall, (1997).

[7] Beale E. M. L., Little R. J. A. Missing values in multivariate analysis. *Journal of the Royal Statistical Society*, 37 (1975), 129-145.

[8] Kim, Jae-On, James C. The treatment of missing data in multivariate analysis. *Sociological Methods and Research*, 6 (1977), 215:240.

[9] Little R. J. A. Missing data in large surveys. *Journal of Business and Economic Statistics*, 6 (1988), 287-201.

[10] Little R. J. A., Rubin D. B. *Statistical Analysis with Missing Data.* New York: Wiley, (1987).

[11] Schafer J. L., Graham J. W. Missing data: over view of the state of the art. *Psychological Methods*, 7 (2002), 147:177.

[12] Little R.J.A. Rubin DB. Statistical analysis with missing data. Wiley, (2002).

[13] Fielding S., Fayers P. M., Ramsay C. Health and quality of life outcomes. *BioMed Central.* 7 (2009), 57.

[14] "Gross National Happiness." The Centre for Bhutan Studies. 2012. http://www.grossnationalhappiness.com/. Accessed 19 February 2012.

[15] Rubin D.B. Inference and missing data. *Biometrika*, 72 (1976), 359-364.

[16] Fairclough D.L. *Design and analysis of quality of life studies in clinical trials.* Chapman and Hall, (2002).

[17] Dixon W. J. *BMDP statistical software.* Los Angeles. University of California Press, (1988).

[18] Little R.J.A. A test of missing completely at random for multivariate data with missing values. *Journal of American Statistical Association*, 83 (1988), 1198-1202.

[19] Thoemmes F., Enders C. K. A structural equation model for testing whether data are missing completely at random, (2007).

[20] Diggle P.J. Testing for random dropouts in repeated measurements data. *Biometrics*, 45 (1989), 1255-1258.

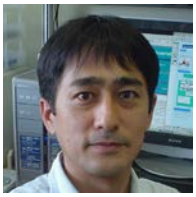
[21] Fisher R. A. *Statistical methods for research workers*, 12th edition. Oliver and Boyd, London, (1954).

[22] Listing J., Schlittgen R. Tests if dropouts are missed at random. *Biometrical Journal*, 40 (8) (1998), 929-935.

[23] Listing J., Schlittgen R. A non-parametric test for random dropouts. *Biometrical Journal*, 45 (1) (2003), 113-127.

[24] Kim K. H. Bentler P. M. Tests of homogeneity of means and covariance matrices for multivariate incomplete data. *Psychometrika*, 67 (2002), 609-624.

Sonam TSHERING is a PhD candidate at the University of the Ryukyus, Japan. He received his Master's Degree in Information Engineering, Bachelor's Degree in Computer Application, and Diploma in Information Management Systems from the Ryukyus University, Japan; Vinayaka Missions University, India; and Royal Institute of Management, Bhutan respectively. He was working as a system administrator at Ministry of Labour and Human Resources, Bhutan prior to his coming to Japan for studies. His research interests include handling missing data, design and analysis of QoL studies, constructing socio-economic indicators, causal modeling, web services and predictive analytics.



Takeo OKAZAKI took B.Sc., M.Sc. from Kyushu University in 1987 and1989, respectively. He had been a re-search assistant at Kyushu University from 1989 to 1995. He has been a lecturer at University of the Ryukyus since1995. His research interests are statistical data normalization for analysis, statistical causal relationship analysis.

Satoshi ENDO took M.E., D.E. from Hokkaido University in 1990 and 1995,respectively. He had been a research assistant at Hokkaido University from1990 to 1995. He has been a professor at University of the Ryukyus since 2005.His research interests are intelligent informatics and sensitivity informatics/soft computing.