Menemui Matematik (Discovering Mathematics)
Vol. 34, No. 1: 7– 15 (2012)

Analysis of Repeated Measures via Simulation (Pengukuran Berulang Analisis melalui Simulasi)

Chin Wan Yoke[1], Zarina Mohd Khalid[2] & Ho Ming Kang[3]
[1,2,3] Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia
[1]wanyoke86@yahoo.com, [2]zarinamkhalid@utm.my, [3]ryanho1025@gmail.com

ABSTRACT

Longitudinal data are measures collected repeatedly at different time-points; with each time of the collection has the tendency to be affected by a covariate. Furthermore, such data can be subjected to missing values. In this study, a multilevel longitudinal model based on a regression concept, with assumptions of varying degree of correlation between the response and covariate, is suggested. The performance of the estimators in the model parameters is then investigated by using a simulated longitudinal data. The strength of the correlation is responsible to manipulate the distribution of the response variables throughout the simulation process. It is found that as the correlation increases, the performance of the estimators becomes better. In addition to the simulated data, a logistic regression model with some degree of missingness is further implemented to generate the missing values. The model parameters are re-estimated using a reduced data set which ignores the missing values. As expected, the performance of the estimators gets worse as the degree of missing values increases.

Keywords: Correlation, Logistic Regression, Missing Data, Simulation, Longitudinal Model

ABSTRAK

Data pengukuran berulang sering dikumpul berulang-kali pada suatu masa yang berlainan; dengan setiap kali pengumpulan mempunyai kecenderungan untuk dipengaruhi oleh covariat. Selain itu, data tersebut boleh tertakluk kepada nilai-nilai yang hilang. Dalam kajian ini, model pengukuran berulang pelbagai-peringkat yang berdasarkan konsep regresi, bersama dengan andaian bahawa tahap korelasi antara pembolehubah sambutan dan covariat yang berbeza-beza, adalah disyorkan. Prestasi penganggar parameter model kemudian diselidik dengan menggunakan data simulasi pengukuran berulang. Kekuatan kolerasi bertanggungjawab untuk memanipulasi taburan pembolehubah sambutan disepanjang proses simulasi ini. Kajian ini mendapati bahawa korelasi yang lebih kukuh akan menghasilkan prestasi penganggar yang lebih baik. Daripada data yang dijana, model regresi logistik yang mengawal peratusan kehilangan digunakan untuk menjana kehilangan data. Parameter model dianggar semula dengan menggunakan set data yang kurang dengan mengabaikan nilai-nilai yang hilang. Seperti yang dijangka, prestasi penganggar menjadi kurang tepat apabila tahap kehilangan nilai bertambah.

Katakunci: Korelasi, Regresi Logistik, Kehilangan Data, Simulasi, Model Pengukuran Berulang

INTRODUCTION

Repeated measurements are defined as a set of data which is collected continuously at a sequence of time-points. A popular mathematical model for this type of data was presented and investigated by Laird and Ware (1982) through a class of random effect models. Different from the two-stage models in Laird and Ware (1982), this study will incorporate a multilevel model introduced by Goldstein (1986) to analyze the repeated measurements. Simplifying the longitudinal model into a multilevel model, it involves analysis of linear regression. With the proof of Goldstein (1986), the iterative generalized least squares estimation is equivalent to the maximum likelihood estimation (MLE). Since the researchers had produced a general model on the multilevel data in Goldstein and McDonald (1988), it can also be applied to some other special cases such as repeated measures designs. Using the same approach, Hedeker and Gibbons (1997) applied the multilevel model ideas to model the random-effects pattern-mixture model. Gibbons and Hedeker (1997) further improved the two-level hierarchical model into three-level probit and logistic model and then the work was advanced to four-level hierarchical mixed-effects regression models (Gibbons *et. al*., 2010).

As mentioned earlier, longitudinal data are measures collected repeatedly at different time-points. Therefore Diggle (1988) suggested that correlation should be constructed within each time sequence of measurements. In addition, correlation should be introduced in a simple linear regression model due to a close relationship between time-independent or time-dependent covariates (predictors) and the response variable (Ismail Mohamad, 2003). Therefore, it is reasonable to implement this approach into the longitudinal model by using a simplified multilevel path to generate the repeated measurements.

Missing data is always a common problem occurs in longitudinal studies. Therefore, some proportions of covariate might contingently be missing. From the previous review, Little (1992) produced a review paper on the regression analysis with missing *X*'s. The author summarized some of the available methods such as complete case analysis, available case analysis, least squares on imputed data, maximum likelihood, Bayesian and multiple imputation to estimate the missing estimators. Next, according to Diggle and Kenward (1994), the partition of missing data mechanism can be restricted to the assumption of missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR). The authors thus proposed a useful methodology to analyze the longitudinal dropout process. For more information, the most common and alternative approach in analyzing the longitudinal model with missing data was the likelihood-based methods such as selection models (Wu and Carroll, 1988; Diggle and Kenward, 1994; Follmann and Wu, 1995), pattern-mixture models (Wu and Bailey, 1989; Little, 1993, 1994, 1996; Hedeker and Gibbons, 1997) and mixed-effects hybrid models (Little, 2008 and Yuan and Little, 2009).

In particular, the recent works have presented a multilevel longitudinal model based on a linear regression concept, with the assumptions of varying degree of correlation between the response and predictor variables. The performance of estimators of the model parameters is then investigated by using a simulated longitudinal data. With the presence of certain proportions of missing data in the covariate, the model parameters are re-estimated using a reduced data set which ignores the missing values.

LONGITUDINAL MODEL WITH MISSING *X* VALUES

Generally, for *i* subject of interests, the longitudinal model is represented by

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i \quad \text{for ,} \quad i= 1, 2, ..., n, \tag{1}$$

with the error term ε_i follows the normal distribution of mean 0 and variance α_ε^2 and the random effects b_i follow the bivariate normal with mean 0 and variance-covariance matrix Σ.

In this study, it is assumed that the full longitudinal model with a single covariate varies across the k number of time-points, so the model is as follows:

$$y_{ik} = \beta_0 + \beta_1 t_{ik} + \beta_2 x_i + \beta_3 (x_i t_{ik}) + b_{0i} + b_{1i} t_{ik} + \varepsilon_{ik}. \tag{2}$$

or in matrix form, it is represented by

$$\underset{\mathrm{y}_i}{\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ \vdots \\ y_{ik} \end{bmatrix}} = \underset{\mathrm{X}_i}{\begin{bmatrix} 1 & t_{i1} & x_i & x_i \times t_{i1} \\ 1 & t_{i2} & x_i & x_i \times t_{i2} \\ 1 & t_{i3} & x_i & x_i \times t_{i3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{ik} & x_i & x_i \times t_{ik} \end{bmatrix}} \underset{\beta}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}} + \underset{\mathrm{Z}_i}{\begin{bmatrix} 1 & t_{i1} \\ 1 & t_{i2} \\ 1 & t_{i3} \\ \vdots & \vdots \\ 1 & t_{ik} \end{bmatrix}} \underset{\mathrm{b}_i}{\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix}} + \underset{\varepsilon_i}{\begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \vdots \\ \varepsilon_{ik} \end{bmatrix}}. \tag{3}$$

Describing Eq. (2) in the form of multilevel model, it is further split into a within-subjects model, written as

$$y_{ik} = v_{0i} + v_{1i} t_{ik} + \varepsilon_{ik}, \tag{4}$$

and between-subjects model where

$$v_{0i} = \beta_0 + \beta_2 x_i + b_{0i} \quad \text{and} \quad v_{1i} = \beta_1 + \beta_3 x_i + b_{1i}. \tag{5}$$

For the simulation study, correlation is suggested to be introduced for both the within-subjects model (4) and between-subjects models (5). The same covariate x_i is correlated to v_{0i} and v_{1i}, with the b_i being the bivariate normal, so v_i does also follow the bivariate normal. The simulated v_{0i} and v_{1i} are then substituted into (4) in order to generate the repeated measurement y_{ik}. Similar to the between-subjects models, correlation is incorporated in the within-subjects model in which the response variables are correlated to the independent variable such as time t_{ik}.

Following Ismail Mohamad (2003) idea, the simulation algorithm for the simple linear regression model with correlation is developed as follows:

1. Set a value for the correlation coefficients r_{xy}, and the estimators β_0 and β_1.
2. Simulate a set of data $x_{sim} \sim N(\mu_x, \sigma_x^2)$.
3. Find the variance σ_{xs}^2 and σ_{xs} standard deviation for the simulated data x.
4. Find the standard deviation $\sigma_{ys} = \beta_1 \frac{\sigma_{xs}}{r_{xy}}$ and variance σ_{ys}^2 for the simulated data y.
5. Calculate $\sigma_{\varepsilon s}^2 = \sigma_{ys}^2 - \beta_1^2 \sigma_{xs}^2$.
6. Simulate a set of $\varepsilon_{ys} \sim N(0, \sigma_{\varepsilon s}^2)$.
7. Calculate $y_{sim} = \beta_0 + \beta_1 x_{sim} + \varepsilon_{ys}$.

Next, the percentage of missing X values is controlled by the logistic regression model adjusted from Ismail Mohamad (2003):

$$P(X \text{ is missing} \mid \text{data}) = \frac{\exp(a + bY + cX + dif)}{1 + \exp(a + bY + cX + dif)}, \tag{6}$$

where

a is the intercept point; with $a = \ln\left(\frac{q}{1-q}\right) - b\bar{Y} - c\bar{X}$, it is a shifted logistic regression line which is responsible to control for the q proportions of missing X values,

b and c control the type of missing mechanism; for instance, if $b = c = 0$, it will contribute as a MCAR missing mechanism and if $b = 1$, $c = 0$, the production will be a MAR model, and

an additional of *dif* to the logistic regression model can generate the proportions of missing data exactly; let the shifted $dif = \ln\left(\frac{tu}{1-tu}\right) - \frac{\exp(a+bY+cX)}{1+\exp(a+bY+cX)}$ and *tu* followed a pseudo random number 0.

From Eq. (6), the *q* proportions (percentage) of highest probability generated will be defined as the missing data. In this study the missing *X* values will be ignored in the complete case analysis which will then result in deleting the corresponding *Y* values. Hence the simulated data used will be a reduced set of data which ignores *q* proportion of data that contain missing values.

ANALYSIS OF SIMULATION STUDIES

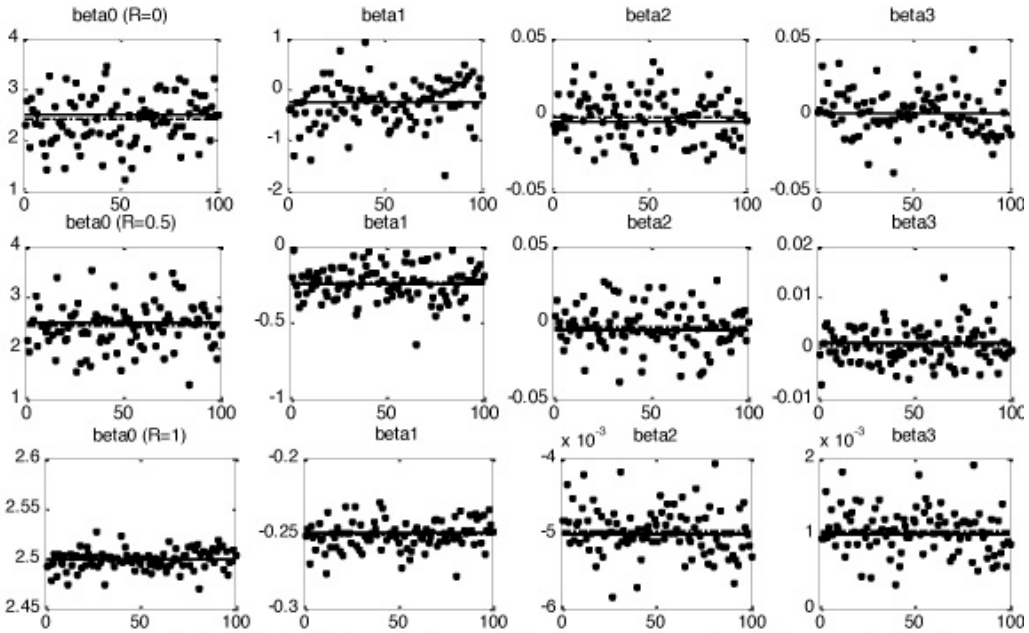
First, the parameters are chosen arbitrarily as β_0 = 2.50, β_1= -0.25, β_2= -0.005 and β_3= 0.001. Followed with the pre-setting time sequences 0, 3 and 6 for all subjects; the *X* values are generated from a pseudorandom numbers which are normally distributed with mean 30 and standard deviation 8. The longitudinal model is then being simulated and the parameters are estimated for 100 times by varying the correlation from 0 to 1. Table 1 shows the results using longitudinal model with and without the presence of missing values.

Referring to Table 1, it suggests that as the strength of correlation increases, the estimators tend to go closer to the true values of the assumed model parameters; these statements is true for three of the complete observation, MCAR and MAR cases. Additionally, Fig. 1, Fig. 2 and Fig. 3 show the different plots of estimates from 100 simulations which can be seen to vary differently around the true assumed parameter values. Bias and standard error (s.e.) significantly decrease when the correlation coefficient increases, as shown in Table 1. In short, the strength of the correlation influences the longitudinal simulation, and thus influences the accuracy of the attained estimators.

This paper only presented the graphical views for the three cases: (1) complete observation, (2) MCAR and (3) MAR, both (2) and (3) are with 60% missing *X* values. From Fig. 1, the distributions of estimates for the complete observation cases were scattered very close around the assumed parameter values, as expected. On the other hand, it is evident from Fig. 2 and Fig. 3 that the distributions of the estimates for the case with 60% missing *X* values were largely dispersed from the assumed parameter values.

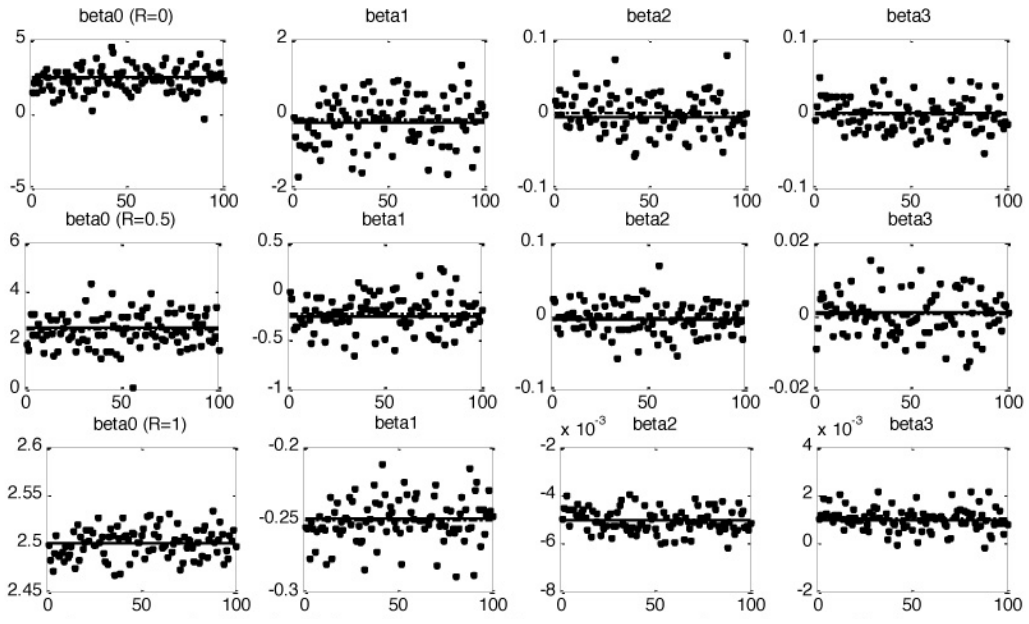
Table 1: Comparisons of the performance of the assumed estimators by varying the percentage of x missingness and correlation.

Correlation			0			0.5			1		
		j	$\hat{\beta}_j$	Bias$_j$	s.e.$_j$	$\hat{\beta}_j$	Bias$_j$	s.e.$_j$	$\hat{\beta}_j$	Bias$_j$ (1.0e-003)	s.e.$_j$
True β_j		0 1 2 3	2.5000 -0.2500 -0.0050 0.0010								
Complete Observation		0 1 2 3	2.4161 -0.2657 -0.0014 0.0013	-0.0839 -0.0157 0.0036 0.0003	0.4818 0.4283 0.0152 0.0134	2.4426 -0.2365 -0.0027 0.0006	-0.0574 0.0135 0.0023 -0.0004	0.4413 0.1113 0.0140 0.0036	2.4990 -0.2508 -0.0050 0.0010	-0.9596 -0.8232 0.0304 0.0344	0.0100 0.0095 0.0003 0.0003
MCAR	30% x missing	0 1 2 3	2.3970 -0.2321 -0.0009 0.0001	-0.1030 0.0179 0.0041 -0.0009	0.6097 0.4981 0.0190 0.0155	2.4699 -0.2399 -0.0037 0.0008	-0.0301 0.0101 0.0013 -0.0002	0.5469 0.1372 0.0166 0.0043	2.4999 -0.2506 -0.0050 0.0010	-0.1096 -0.6136 -0.0029 0.0261	0.0109 0.0120 0.0003 0.0004
	60% x missing	0 1 2 3	2.3387 -0.1839 0.0010 -0.0010	-0.1613 0.0661 0.0060 -0.0020	0.8598 0.6543 0.0264 0.0214	2.4372 -0.2223 -0.0034 0.0004	-0.0628 0.0277 0.0016 -0.0006	0.6946 0.1825 0.0214 0.0057	2.5008 -0.2506 -0.0050 0.0010	0.8466 -0.6032 -0.0284 0.0348	0.0148 0.0150 0.0005 0.0005
MAR	30% x missing	0 1 2 3	2.1506 -0.3508 -0.0007 0.0005	-0.3494 -0.1008 0.0043 -0.0005	0.6509 0.4797 0.0173 0.0145	2.1833 -0.1749 -0.0019 0.0002	-0.3167 0.0751 0.0031 -0.0008	0.5773 0.1497 0.0153 0.0042	2.4999 -0.2506 -0.0050 0.0010	-0.0941 -0.6157 -0.0070 0.0231	0.0113 0.0123 0.0003 0.0004
	60% x missing	0 1 2 3	1.9174 -0.4069 0.0001 -0.0001	-0.5826 -0.1569 0.0051 -0.0011	0.9814 0.7085 0.0243 0.0225	1.9613 -0.1300 -0.0017 0.0001	-0.5387 0.1200 0.0033 -0.0009	0.9469 0.2231 0.0238 0.0060	2.5005 -0.2510 -0.0050 0.0010	0.4637 -1.0000 -0.0247 0.0410	0.0146 0.0150 0.0005 0.0005



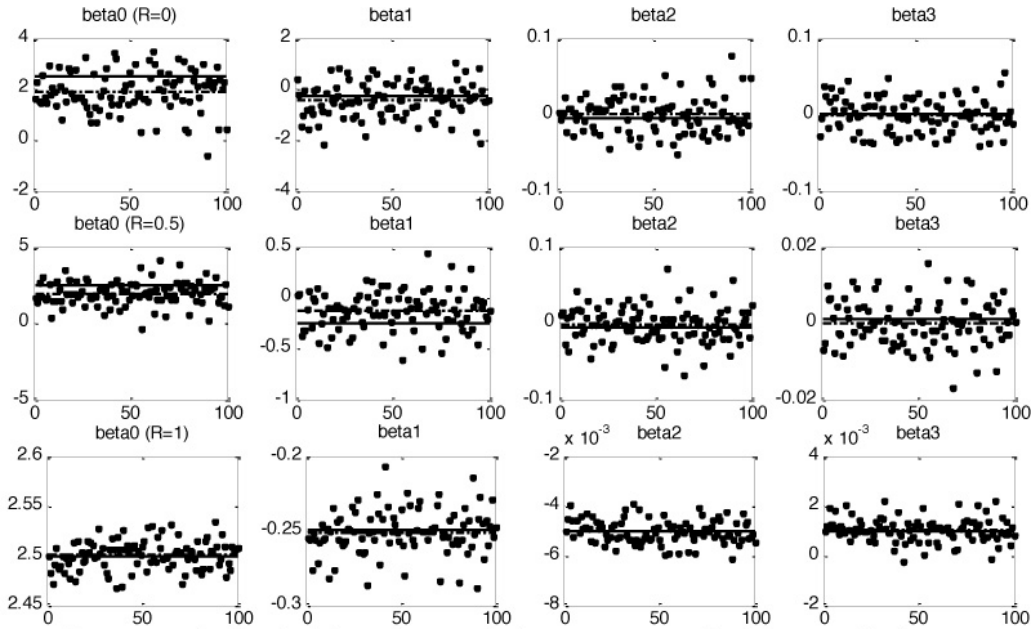
*represent each of the simulation estimators, ─ is the true parameters and –. is the assumed estimators

Figure 1: Comparisons of the complete observation estimators and the true parameters for complete data by varying the correlation coefficient.



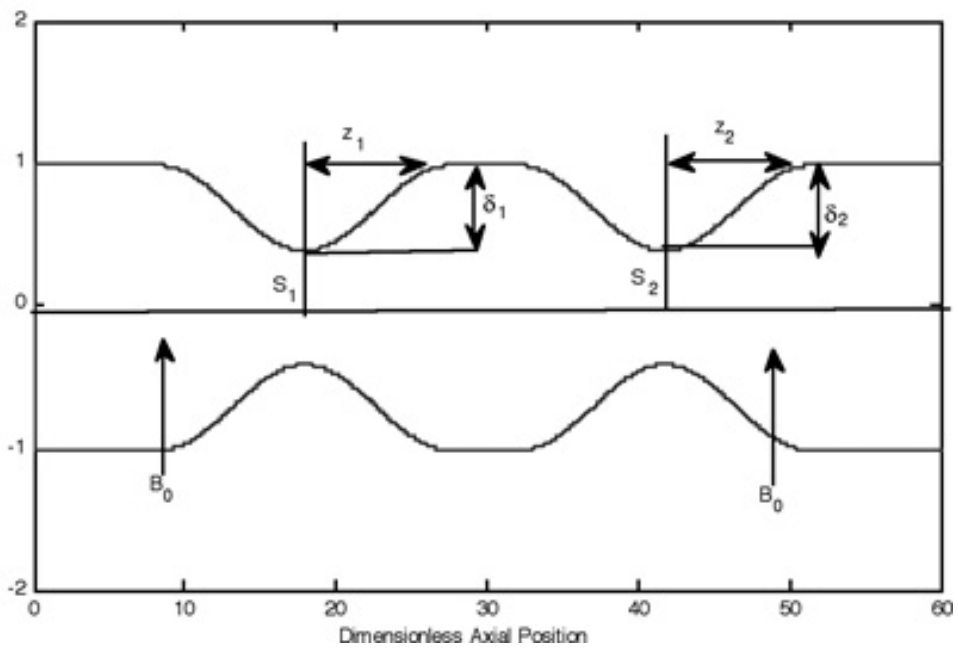
*represent each of the simulation estimators, ─ is the true parameters and –. is the assumed estimators

Figure 2: Comparisons of the MCAR estimators and the true parameters for data with 60% missing X values by varying the correlation.



*represent each of the simulation estimators, ─ is the true parameters and –. is the assumed estimators

Figure 3: Comparisons of the MAR estimators and the true parameters for data with 60% missing *X* values by varying the correlation.



CONCLUSION

In conclusion, the strength of correlation is responsible to manipulate the distribution of the response variable throughout the simulation process. It will directly influence the performance of the estimators. When the correlation coefficient approaches to 1, there exists a substantially strong relationship between the response and the predictor variables, therefore the estimators are very close to the assumed true parameter values. On the other hand, the relationship between response and predictor variables is getting weak when the correlation coefficient moves towards zero, consequently the estimators will diverge from the assumed true parameter values.

The same condition is applied to the model with a certain proportions of missing X data. However, the estimators for the case with missing values, either MCAR or MAR, have performed less efficiently than the estimators with complete observations. The accuracy also reduces as the degree of missing values increases. Lastly, a further consideration of the missing data on the response variable is suggested to be implemented in the model so that the simulation study will fit better in the real environment.

ACKNOWLEDGEMENT

This paper was supported in part by Fundamental Research Grant Scheme (FRGS) grant 4F004.

REFERENCES

Diggle, Peter J. "An Approach to the Analysis of Repeated Measurements." *Biometrics* 44 (1988): 959–971.

Diggle, Peter and Kenward, Mike G. "Informative Dropout in Longitudinal Data Analysis (with discussion)." *Applied Statistics* 43 (1994): 49–94.

Follmann, Dean and Wu, Margaret. "An Approximate Generalized Linear Model with Random Effects for Informative Missing Data." *Biometrics* 51 (1995): 151–168.

Gibbons, Robert D. and Hedeker, Donald. "Random Effects Probit and Logistic Regression Models for Three-Level Data." *Biometrics* 53 (1997): 1527–1537.

Gibbons, Robert D., Hedeker, Donald and Du'Toit, Stephen. "Advances in Analysis of Longitudinal Data." *Annual Review of Clinical Psychology* 6 (2010): 79–107.

Goldstein, Harvey. "Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Square." *Biometrika* 73 (1986): 43–56.

Goldstein, Harvey and McDonald, Roderick P. "A General Model for the Analysis of Multilevel Data." *Psychometrika* 53 (1988): 455–467.

Hedeker, Donald and Gibbons, Robert D. "Application of Random-Effects Pattern-Mixture Models for Missing Data in Longitudinal Studies." *Psychological Methods* 2 (1997) : 64–78.

Ismail Mohamad. "Data Analysis in the Presence of Missing Data." PhD Thesis, Lanchester University, 2003.

Laird, Nan M. and Ware, James H. "Random-Effects for Longitudinal Data." *Biometrics* 38 (1982): 963–974.

Little, Roderick J. A. "Regression With Missing X's: A Review." *Journal of the American Statistical Association* 87 (1992), 1227–1237.

Little, Roderick J. A. "Pattern-Mixture Models for Multivariate Incomplete Data." *Journal of the American Statistical Association* 88 (1993): 125–134.

Little, Roderick J. A. "A Class of Pattern-Mixture Models for Normal Incomplete Data." *Biometrika* 81 (1994): 471–483.

Little, Roderick J. A. "Selection and Pattern-Mixture Models." In *Advances in Longitudinal Data Analysis*, edited by Garrett Fitzmaurice, Marie Davidian, Geert Verbeke and Geert Molenberghs, 409–431. London: CRC Press, 2008.

Little, Roderick J. A. and Wang, Yongxiao "Pattern-Mixture Models for Multivariate Incomplete Data with Covariates." *Biometrics* 52 (1996): 98–111.

Wu, Margaret C. and Bailey, Kent R. "Estimation and Comparison of Changes In the Presence of Informative Right Censoring: Conditional Linear Model." *Biometrics* 45 (1989): 939–955.

Wu, Margaret C. and Carroll, Raymond J. "Estimation and Comparison of Changes In the Presence of Informative Right Censoring by Modeling the Censoring Process." *Biometrics* 44 (1988): 175–188.

Yuan, Ying and Little, Roderick J. A. "Mixed-Effect Hybrid Model for Longitudinal Data with Nonignorable Dropout." *Biometrics* 65 (2009): 478–486.