Missing covariates in multiple linear regression when the data is missing at random

G.P. Nthoiwa[1] and J.O Owino[2]
[1]Botswana College of Agriculture, Private Bag 0027, Gaborone. Email gnthoiwa@bca.bw
[2]University of Nairobi, School of Mathematics, P.O. Box 30197, Nairobi 0100, Kenya.

ABSTRACT
The occurrence of missing values is frequent in data collected for different uses such as in surveys, censuses, balanced experiments. On the other end most statistical analysis methods have been developed for complete rectangular data. This paper uses a simulated data set to examine the performance of recently available methods for treating data with missing values. Multiple imputations (MI) and maximum likelihood (ML) methods through expectation maximization (EM) were compared with the complete case (CC) analysis which is the default method in statistical computer packages. The effects of the data treatment methods were examined on the regression coefficients. The results indicate that ML through EM and MI methods are both superior than the commonly available complete case analysis (CC).
Key words: Multiple regression; complete case; maximum likelihood; multiple imputations; missing data

INTRODUCTION
It is a common place experience of any practicing statistician to be faced with the task of analyzing data with missing values. This is despite the fact that most standard methods have been developed to analyze rectangular sets. Missing values are most frequently encountered with data routinely collected for data information systems such as those providing official statistics based on census and surveys or in clinical records systems of hospitals. Missing values also arise by default in data recording activities which are designed to give complete records such as in balanced experiments. The occurrence of missing values is so frequent that it has become one of the most important problems in statistical analysis (Hartly and Hocking, 1971). When encountered, missing values are usually taken as nuisance as they are not the main focus of inquiry. The widespread occurrence of missing values may be regarded as a necessary evil realistically associated with data collection. However, literature provides plenty of evidence to reveal that statisticians are already blessed with complex methodologies for treating missing data situations in which alternative methods of analysis are often put forward and compared.

The traditional theory for estimation in regression models gives no hint to deal with missing values in the covariates. Intuitively, when the subjects with missing covariate values differ systematically from those with complete data with respect to the outcome of interest, results from a traditional data analysis omitting the missing cases may no longer be valid. Because standard techniques for regression models require full covariate information, one simple way to avoid the problem of missing data is to use the complete case (CC) analysis where only those subjects who are completely observed are used in the analysis. Complete case analysis is the technique most commonly used with missing values in the covariates and/or response, and it is still the default method in most software packages, despite the development of statistical methods that handle missing data more appropriately. The objective of this paper is to compare different methods of treating missing

continuous covariate data in multiple regression when the data is missing at random. The paper aims to evaluate two methods in comparison to the complete case analysis – that is a likelihood based method which uses Expectation Maximization (EM) algorithm and a Multiple Imputation (MI) method. The paper describes general notation for general linear models (GLM's) and explains the likelihood based method and the multiple imputation method.

General Linear Models

Suppose that $(x_1, y_1), \ldots (x_n, y_n)$ are independent observation where each y_i is the response variable and each x_i is a $P \times 1$ random vector of covariates. The joint distribution of (x_i, y_i) is specified by a conditional distribution of y_i given x_i and a marginal distribution of x_i. Suppose $(y_i \mid x_i)$ has a density in the exponential class with the form:

$$p(y_i \mid x_i, \theta_i, \tau) = \exp\left\{a_i^{-1}(\tau)[y_i\theta_i - b(\theta_i)] + c(y_i, \tau)\right\}, i = 1, \ldots n \quad (1)$$

Indexed by the canonical parameter θ_i and the scale parameter τ. The functions b and c determine a particular family in the class, such as the Binomial, Normal or Poisson. These functions $a_i(\tau)$ are commonly of the form $a_i(\tau) = \tau^{-1}k_i^{-1}$ where k_i are the known weights. Further suppose that the θ_i's satisfy the equations $\theta_i = \theta(\eta_i)$, $i = 1, \cdots, n$ and $\eta = X\beta$ where $\eta_i = x_i'\beta$ are the components of η, X is an $n \times p$ full rank matrix of covariates, $\beta = (\beta_1, \cdots, \beta_p)'$ is a $p \times 1$ vector of regression coefficients, and θ is a monotone differentiable function. The model described in equation (1), is called a general linear model and has proven very useful in many applications. When $\theta_i = \eta_i$, the link is said to be a canonical link. The GLMs include a large class of regression models, such as normal linear regression, logistic and probit regression, Poisson regression, gamma regression, and some proportional hazards models (McCullagh and Nelder, 1989). The complete-data likelihood for GLM based on all observations is given by

$$L(\beta \mid x, y) = \prod_{i=1}^{n} p(y_i \mid x_i, \beta) \quad (2)$$

where $y = (y_1, \cdots, y_n)$ and $x = (x_1, \cdots, x_n)$. The observed data likelihood is obtained by integrating equation (2) over the missing values, with respect to the distribution of the missing values (Little and Rubin, 2002).

Methods of Treating Data with Missing Values

Likelihood Based Methods

A large class of model based procedure arises from defining a model for variables with missing values and making statistical inference based on a maximum likelihood method. An important issue is the specification of the covariate distribution. In missing data problems one should consider strategies for reducing the number of nuisance parameters in the covariate distribution (Lipsitz and Ibrahim, 1996; Ibrahim *et al.*, 1999). In this case if the missing data mechanism is non ignorable, it is typical to specify a parametric model for the missing data mechanism and to incorporate it into the complete data log-likelihood. Let x_{ik} denote the k^{th} component of x_i. The missing data mechanism is defined as the distribution of the $p \times 1$ random vector r_i, whose k^{th} component r_{ik} equals 1 if x_{ik} is observed for the subject i, and 0 if x_{ik} is missing. The conditional distribution of r_i given (y_i, x_i), denoted $[r_i \mid y_i, x_i, \phi]$, is indexed by the parameter vector ϕ and is a multinomial

distribution with 2^p cell probabilities. The marginal density of x_i is denoted by $p(x_i \mid \alpha)$, where α is a vector of indexing parameters. The complete data density of (y_i, x_i, r_i) for subject i is then given by

$$p(y_i, x_i, r_i \mid \beta, \alpha, \phi) = p(y_i \mid x_i, \beta)p(x_i \mid \alpha)p(r_i \mid y_i, x_i, \phi) \quad (3)$$

This leads to the complete data log-likelihood

$$l(\gamma) = \sum_{i=1}^{n} l(\gamma; x_i, y_i, r_i) \equiv \sum_{i=1}^{n} \log\{p(y_i \mid x_i, \beta)\} + \log\{p(r_i \mid y_i, x_i, \phi)\} \quad (4)$$

where $\gamma = (\beta, \alpha, \phi)$ and $l(\gamma; x_i, y_i, r_i)$ is the distribution to the contribution on the complete data log-likelihood for the *ith* observation. The main interest here is in the estimation of β, with α and ϕ being viewed as nuisance parameters. Log-linear is chosen for specifying multinomial model $p(r_i \mid y_i, x_i, \phi)$.

Estimating Maximum likelihood using EM Algorithm

The observed data likelihood is generally difficult to obtain in closed form for most missing data problems, including GLM. In this regard, the Expectation Maximization (EM) algorithm has been a popular technique for obtaining maximum likelihood estimates (MLE) in GLM's with missing covariates. In this paper the EM algorithm for estimating missing continuous covariates is used. The E-step for the EM algorithm used for continuous covariates consists of an integral, which typically does not have a closed form for GLMs. The E-step for the ith observation would therefore be

$$Q(\gamma \mid \gamma^{(r)}) = \int \log\{p(y_i \mid x_i, \beta)\} p(x_{mis,i} \mid x_{obs,i}, y_i, r_i, \gamma^{(r)}) dx + \int \log\{p(x_i \mid \alpha)\} p(x_{mis,i} \mid x_{obs,i}, y_i, r_i, \gamma) dx_{mis,i} + \int \log\{p(r_i \mid y_i, x_i, \phi)\} p(x_{mis,i} \mid x_{obs,i}, y_i, r_i, \gamma) dx_{mis,i} \quad (5)$$

To evaluate equation (5) at the $(r+1)^{th}$ iteration of the EM algorithm, a Monte Carlo version of the EM algorithm as given by Wei and Turner (1990) is used. To do this, one must first generate a sample from $p(x_{mis,i} \mid x_{obs,i}, y_i, \gamma^{(r)})$. A straightforward derivation yields

$$p(x_{mis,i} \mid x_{obs,i}, y_i, \gamma^{(r)}) \propto p(y_i \mid x_i, \beta^{(r)}) p(x_i \mid \alpha^{(r)}) \quad (6)$$

The product on the right hand side of equation (6) has an elegant form for efficient sampling as discussed by Ibrahim et al. (1999).

Imputation Methods

Dempster and Rubin (1983) stated that the idea of imputation tempts analysts to believe that data are complete after all. This authors further state that this idea is dangerous because it results in a situation where the problem is assumed to be sufficiently minor such that standard estimators are applied to the real and imputed data, thus creating substantial biases. Multiple Imputation (MI) appears to be one of the most attractive methods for general purpose handling of missing data in multivariate analysis, and this is the method of interest in this paper. The basic idea, first proposed by Rubin (1977) and elaborated in his book (Rubin, 1987), is as follows:

(1) Construct M "complete data sets;

(2) Obtain $\hat{\gamma}^{(m)}$ for the *mth* imputed dataset, $m = 1, 2, ...M$; and

(3) The parameter estimate is

$$\gamma = \frac{1}{M}\sum_{i=1}^{m}\gamma^{(m)}$$

The variance estimate from the m^{th} imputed dataset is obtained by assuming that the imputed data set is the complete data set and calculating the usual variance estimate. MI works similar to EM algorithm in that MI replaces the conditional mean imputation in the E-step of the EM algorithm by a single draw from the imputation distribution, defined as

$$p(x_{miss,i} \mid x_{obs,i}, y_i, \gamma) \propto p(y_i \mid x_i, \beta)p(x_i \mid \alpha) \quad (7)$$

METHODOLOGY

To achieve the objective, a simulated da set of a sample size n (=45) was creat using SAS (9.1). Then a dataset with 30 missing values (missing at random) w created from the complete data set. Th complete data set was compared to th missing data set using complete ca analysis (default in SAS), EM algorithm ar MI methods to observe how much of bias created. All the comparisons were don using the standard error of regressio coefficient (β).

Table 1. Effect of missing values on parameter estimation

Variable	Full Data Set			Complete Case Analysis		
	Parameter Estimate	Standard error	t-value	Parameter Estimate	Standard error	t-value
X_1	0.71564	0.13486	5.31	0.65082	0.27925	2.33
X_2	1.29529	0.36802	3.52	0.78319	0.60445	1.30
X_3	-0.15212	0.15629	-0.97	-0.07819	0.30134	-0.26

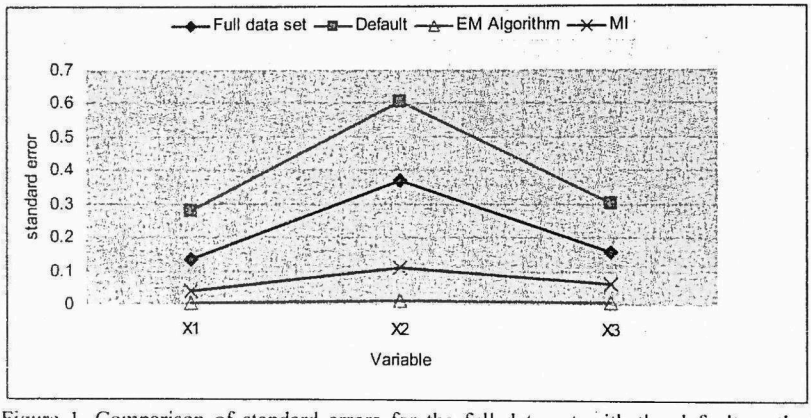

Figure 1. Comparison of standard errors for the full data set with the default method, E algorithm and MI methods

RESULTS AND DISCUSSION

Table 1 shows the effects of missing values on the estimation of regression coefficients and standard error for a full data set and a data set with 30% of the values missing

The results show that the regression coefficients estimated for the full data set are higher than for the missing values data set, thus the complete case analysis method underestimates the coefficients. On the other hand, the standard errors for the full data set are found to be lower than those for the data set with missing values. These two scenarios have an effect on hypothesis testing since the test statistic t is equal $t = \frac{b_i}{s_{b_i}}$, where b_i is the sample regression coefficient and s_{b_i} is the sample standard error. This can lead to a wrong interpretation of the multiple regressions because it is difficult to tell which of the variable have an effect on the dependent variable. The above comparison shows that the default method in statistical packages has bias in estimating the multiple regression coefficients.

In addition to the default method, applied statisticians have more choices for analyzing multiple regression data sets with missing values, such the EM algorithm in finding the maximum likelihood and multiple imputation method. As one of the objectives, this paper compares the two methods against the default method to find out which one gives the best estimate. The best estimate method is the one that shows the lowest standard error. Figure 1 shows the results of the comparison of the standard errors estimated for the full data set, default method, EM algorithm and multiple imputation method.

The EM algorithm method has the lowest standard error, up to 10 times lower than the MI method for all variables. The default method (complete case analysis) showed the largest standard error. From these results, the EM algorithm is the best method. The low standard error estimated from the EM algorithm indicates that the regression coefficient is close to that of the full data set. The MI method also gives good results and even though standard errors that are above those of the EM algorithm. Both these methods are generally better than the complete case analysis which gives inflated standard errors and mean square errors. Ibrahim et al. (2005) also reported similar results when they compared maximum likelihood (ML) using EM, MI, weight estimating equation (WEE), Fully Bayesian (FB), and the CC methods on a set of simulated data. They also found that CC was in general outperformed by ML, MI FB and WEE methods regardless of whether the covariates were correctly specified. However, their analysis was based on the data that has categorical missing covariates that were assumed to be missing at random. The superiority of maximum likelihood and multiple imputation analyses have also been reported by Graham et al. (1996), Graham et al. (1997) and Wothke (1998) for non random incomplete data.

CONCLUSION

This paper demonstrates that misleading conclusions can be drawn if complete case analysis is conducted on data with missing values. The best method to use for estimating regression coefficients is the EM algorithm, followed by MI, when analyzing data by multiple regressions if the covariates are continuous and missing at random.

REFERENCE

Dempster, A.P and Rubin B.B. (1983). Introduction. In: Incomplete Data in Sample Surveys (volume 2): Theory and Bibliography. pp 3-10. New York, Academic press.

Graham, J.W., Hofer, S.M. and Mackinnon, D.P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: an application of maximum likelihood procedures. *Multivariate Behavioral Research*. 3: 197-218.

Graham, J.W., Hofer, S.M., Donaldson, S.I., Mackinnon, D.P. and Schafer, J.L. (1997). Analysis with missing data in prevention research. In: The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research. Brynant, K. Windle, M. and West, S. (Eds). National Institute on Drug Abuse. Washington DC.

Hartley, H.O., and Hocking, R.R. (1971). The analysis of incomplete data. *Biometrics* 27: 783-808.

Ibrahim, J.G., Lipsitz, S.R. and Chen, M.H. and Herring, H. (2005). Missing data methods for generalized linear models: a comparative review. *Journal of American Statistical Association* 469: 332-346.

Ibrahim, J.G., Lipsitz, S.R. and Chen, M.H. (1999). Missing covariate in generalized linear models when the data mechanism is Nonignorable. Journal of Royal Statistical Society Series B 61: 173-190.

Lipsitz, S.R. and Ibrahim, J.G. (1996). A conditional model for incomplete covariates in parametric regression models. *Biometrika*, 83: 916-922.

Little and Rubin, (2002). Statistical analysis with missing data 2nd e.d New York: Wiley

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models" 2nd ed. New York, CRC Press.

Rubin, B.D. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* 7: 34-58.

Rubin, B.D. (1987). Multiple imputation for nonresponse in surveys. New York, Wiley.

Wei, G.C. and Turner, M.A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association* 85:699-704.

Wothke, W.W. (1998). Longitudinal and multi-group modelling with missing data. In: Modeling Longitudinal and Multiple Group Data: Practical Issues, Applied Approaches and Specific Examples. Little, T.D, Schnabel, K.U. and Baumert, J. (eds). Mawah, NJ. Lawrence Erlbaum Associates