Malaysian Journal of Mathematical Sciences 5(1): 27- 44 (2011)

# Analyzing Data with Missing Continuous Covariates by Multiple Imputation Using Proper Imputation

**M. Ganjali and H. Zahed**
*Department of Statistics,*
*Shahid Beheshti University,*
*Evin, Tehran, Iran.*
*E-mail: m-ganjali@sbu.ac.ir*

## ABSTRACT

Missing covariate data occur inevitably in various scientific researches. The response variable of interest in these studies may be continuous or categorical and the covariates may have a continuous or discrete nature. Multiple Imputation (MI) procedures may be used to properly or improperly impute the missing data several times and to find parameter estimates by combining the pseudo-complete-case analyses of the imputed data-sets. Although many efforts in the literature have been placed on analyzing continuous response data with missing covariates using MI, models for ordinal response data with missing covariates have received less attention. In this paper four different models for imputation of a missing continuous covariate, of which three are proper and one improper, are compared in models for ordinal responses. All models can be easily implemented in existing software. Data from a Steatosis study is used to illustrate the use of these models. The importance of using a fuller model for imputation compared to that of the analysis model is finally underlined.

Keywords: Missing Data, Categorical Response Data, Generalized Linear Models (GLMs), Multiple Imputation, Predictive Distribution, Proper Imputation.

## INTRODUCTION

Incomplete data, in which parts of the information relating to some subjects are not available, is a common problem in both experimental and observational studies. Standard statistical modeling methods require a complete dataset in which all responses $(y_i's)$ and covariates $(x_i's)$ have been fully recorded. When parts of the data are not available, a typical method usually referred to as "Complete-Case analysis" (CC), is to disregard all cases with missing values and estimate model parameters using the remaining complete dataset. Though simple, this could result in serious bias and inefficiency (Little and Rubin (2002)). Hence much effort in the past has been put into methods

concerning data-analysis and statistical modeling in the presence of missing data. Little and Rubin (2002) and Schafer (1997) give thorough reviews of available methods in the literature for handling missing data. Most of these methods concern themselves primarily with incomplete data on behalf of the response variable. Little and Schluchter (1985) and Little (1992) discuss missing covariates in models for normal data. Ibrahim (1990) and Horton and Laird (1998) focus on missing categorical covariates in Generalized Linear models (GLMs); and Ibrahim *et al.* (1999) extend this approach to the case of missing continuous covariates. In a review article, Ibrahim *et al.* (2005) cover the literature for missing covariates in GLMs. For a thorough review of Generalized Linear Models the reader can see McCullagh and Nelder (1989). Fahrmeir and Tutz (1994) and Agresti (2002) pay special attention to the modeling of categorical (nominal and ordinal) responses.

However, models for ordinal data with missing covariate have received less attention. A variety of ordinal regression models can be used to model ordinal data (Snell (1964); Agresti (2002); McCullagh (1980)) when there is no missing covariate. These include cumulative models (with logit, probit or complementary log-log link functions) and sequential models (Fahrmeir and Tutz (1994); Berridge (1995); Berridge and Dos Santos (1996); Tutz (2005)) for which different link functions can be used. In this paper we shall use the logit link since it is the canonical link (Nelder and Wedderburn (1972)) and leads to a likelihood that is simpler to optimize compared to the likelihood derived using the other links. This paper will concern itself with modeling ordinal data when the information on behalf of some continuous covariate is partly not available due to missingness.

## MODELS AND CONCEPTS

In this section some primary concepts regarding GLMs and the modeling of ordinal data are presented. Also issues concerning incomplete datasets and approaches to the analysis of such data are briefly reviewed. For more detailed discussions the reader can refer to the mentioned references.

## GLMs and Modeling Categorical Responses

A Generalized Linear Model, as a generalization of the familiar linear model, is characterized by:

$$\mu_i = \mathrm{E}\left(y_i \middle| x_i\right) = h\left(z_i'\boldsymbol{\beta}\right);$$

in which $h$ is a known one-to-one link function, $\boldsymbol{\beta}$ is a vector of parameters and $z_i$, a function of $x_i$, is referred to as the design vector. The model errors here do not necessarily follow the Normal distribution, but rather a distribution from the simple exponential family with the general form:

$$f\left(y_i;\theta_i,\phi\right) = \exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi} + c\left(y_i,\phi\right)\right\};$$

where $\theta_i$ is the natural parameter (possibly a vector), $\phi$ is an additional dispersion parameter and $b$ and $c$ are known functions relating to the specific distribution in the family (Normal, Poisson, Gamma, etc.).

Estimation of model parameters $\boldsymbol{\beta}$ and possibly $\phi$ are carried out by maximizing the model likelihood using iterative methods such as Newton-Raphson. Different choices of the link function $h$ and error distributions result in different models. If the errors follow a Poisson distribution and $\eta = h^{-1}\left(\mu\right) = \log\left(\mu\right)$, the resulting model is a log-linear model suitable for count data. For binary data a choice of $\eta = \log\left(\pi/\left(1-\pi\right)\right)$, where $\pi = P\left(y_i = 1 \middle| x_i\right)$ prompts the familiar logistic regression model. For normal data one may use the identity function as a link.

For nominal responses however, the same idea is extended to form a multivariate version of a GLM. If $y_i$ is a nominal variable with $J$ categories, $y_i$ can be assumed to follow a Multinomial distribution:

$$y_i \sim \mathit{Multinomial}\left(n_i, p_{i1}, p_{i2}, ..., p_{iJ}\right).$$

Now each category when compared to another can resemble a similarity to the case of binary responses. With the last category treated as reference, for each category $j$ we can write:

$$\log \frac{p_{ij}}{p_{iJ}} = z_i' \boldsymbol{\beta}_j \qquad j = 1, ..., J-1;$$

which is sometimes referred to as a base-line category logit model.

Finally for ordinal responses with $J$ ordered categories, the same idea leads to a cumulative logit model with the general form:

$$\log \left( \frac{P\left(y_i \le r \middle| x_i\right)}{P\left(y_i > r \middle| x_i\right)} \right) = \theta_r + z_i' \boldsymbol{\beta}; \tag{1}$$

where $r = 1, ..., J-1$ and $\theta_r's$ are the so-called threshold parameters, corresponding to each level of the ordinal variable. The cumulative logit model can be better interpreted using the idea of a latent variable. In this approach, the ordinal variable $y$ is thought of as an observable version of the continuous, yet unobservable variable $y^*$, which is in turn related to the design vector $z$ in a linear fashion:

$$y_i^* = -z_i' \boldsymbol{\beta} + \varepsilon$$

Here, $\theta_r's$ are thought of as cut points or thresholds which define the relationship between $y$ and $y^*$:

$$y_i = r \quad \Leftrightarrow \quad \theta_{r-1} < y_i^* < \theta_r, \qquad r = 1, 2, ..., J.$$

These thresholds are additional parameters themselves, which will need to be estimated. Although usually they are not of interest and only used to compute response probabilities, they can be interpreted as partition-specific intercepts (cut-points) indicating the logarithms of odds of selecting lower, rather than higher, categories when all explanatory variables are set to zero. It can be easily seen that y is related to the linear predictor $z'\boldsymbol{\beta}$ in the form:

$$P\left(y_i \le r \middle| z_i\right) = F\left(\theta_r + z'\boldsymbol{\beta}_i\right),$$

where $F$ is the distribution function of the random variable $\varepsilon$. When a logistic distribution is assumed for $F$, (1) is simply obtained.

**Patterns and Mechanisms of Missing Data**

A distinction is usually made between the placement of unavailable (missing) items in the dataset, i.e. missing-data pattern, and the mechanism leading to this unavailability, referring to the relationship between missingness and the values of variables. Let $R=\{r_{ij}\}$ be an indicator matrix denoting whether the item in the $i$'th row and $j$'th column of the dataset $Y=\{Y_{ij}\}$ is observed. The missing-data pattern is said to be monotone if it can be arranged in such a way that if $r_{ij}=0$ then $r_{ik}=0$ for all $k>j$. This means that if there is some order on the variables (e.g. longitudinal response data) and the $j$-th response is missing, all responses after the $j$-th response are also missing. If not, the pattern is said to be general or intermittent i.e. the individual with some missing values may be returned to the study after a period of time. The $Y$ here represents the whole data matrix including the covariates.

Using the terminology of Little and Rubin (2002) and Diggle and Kenward (1994), the missing-data mechanism is said to be Missing Completely at Random (MCAR) if:

$$f\left(R|Y,\boldsymbol{\phi}\right)=f\left(R|\boldsymbol{\phi}\right) \qquad \forall Y,\boldsymbol{\phi};$$

that is missingness does not depend on any variable in the data-set. Here $\boldsymbol{\phi}$ is a vector of parameters.

If $Y=\left(Y_{obs},Y_{mis}\right)$, where $Y_{obs}$ is the observed part of $Y$ and $Y_{mis}$ is the missing part, the missing-data mechanism is said to be Missing at Random (MAR) if:

$$f\left(R|Y,\boldsymbol{\phi}\right)=f\left(R|Y_{obs},\boldsymbol{\phi}\right) \qquad \forall Y_{mis},\boldsymbol{\phi};$$

which means that missingness only depends on the observed values of the dataset and not on the missing ones.

MCAR and MAR mechanisms are usually referred to as ignorable. In contrast to these ignorable mechanisms (so called because in fully parametric analysis, based on likelihood function the model for the missing data mechanism may be ignored with the assumption that parameters associated with missing mechanism and the responses of interest are disjoint), if the distribution of $R$ also depends on the missing values $Y_{mis}$, the mechanism is said to be Not Missing at Random (NMAR). In this paper we shall assume that missing covariate data are MAR and so missing mechanism can be ignored.

## Approaches to the Analysis of Incomplete Data

Four general approaches in handling missing data can be distinguished:

1. Complete-Case Analysis is the simplest method for the analysis of incomplete data which involves the analysis of the set of observation with no missing values. Although quite simple and tractable, this approach is only advised when the missing-data mechanism is known to be MCAR and the rate of missingness is fairly low.

2. Weighting Procedures which are similar to randomization inferences in sample survey data where the complete cases are re-weighted to adjust for non-response (Robins *et al.* (1995)). In these methods a model for the probability of missingness is fitted, and the inverse of these probabilities are used as weights for the complete cases.

3. Likelihood Based Approaches involve method where one needs to model the distribution of the missing covariates (vide Lipsitz and Ibrahim, (1996)) and to include it in the model likelihood. Monte-Carlo and iterative methods are then used to make inferences about the model using the obtained likelihood.

4. Imputation Based Methods where missing values are "filled-in" once or several times (resulting in Single Imputation and Multiple Imputation) using a specific method such as mean imputation, hot-deck imputation, regression imputation, etc. and the resulting imputed dataset is treated as a complete dataset in which standard statistical methods are applicable.

Issues concerning incomplete datasets and general approaches to modeling data in the presence of missingness are given a standard review in Little and Rubin (2002). Schafer and Graham (2002) also provide a concise and explicit introduction to the subject.

## MULTIPLE IMPUTATION

Single Imputation, though practically convenient, treats the resulting dataset as if complete, failing to take into account the uncertainty regarding the imputed values. This usually results in biased estimates of population quantities. Thus Rubin (1978) proposed Multiple Imputation (MI), in which the advantages of imputation are retained whilst the drawbacks are also addressed.

MI consists of three successive steps:

1. Imputing each missing value $M$ times to obtain $M$ pseudo-complete datasets,

2. Analyzing each data-set using standard statistical methods for the estimation of a desired population quantity $Q$, and obtaining $\hat{Q}_1,...,\hat{Q}_M$ with respective variances $\hat{V}_1,...,\hat{V}_M$,

3. Combining the results using Rubin's rules:

$$\hat{Q} = \frac{1}{M}\sum_{m=1}^{M}\hat{Q}_m$$

$$V = \bar{V} + \left(1+\frac{1}{M}\right)\hat{B}$$

in which $\bar{V} = \frac{1}{M}\sum_{m=1}^{M}\hat{V}_m$ is the average within imputation variance and $\hat{B} = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{Q}_m - \hat{Q}\right)$ is the between imputation variability.

The important part of MI is the process by which the $M$ imputations are generated. When the missingness has occurred in one or more of the covariates $(x's)$, the general idea is to make draws from $P(x_{mis}|y, x_{obs}; \gamma)$ in which $x_{mis}$ indicates the missing part of $x$ and $x_{obs}$ the observed part. If the missing data mechanism is assumed to be MAR, in order to generate $x_{mis,i}$, the imputation distribution:

$$P(x_{mis,i}|y_i, x_{obs,i}; \gamma) \propto P(y_i|x_i; \beta) P(x_{mis,i}|x_{obs,i}; \alpha) \tag{2}$$

may be used. Here $i$ indicates the $i$'th subject in the study and $\gamma = (\alpha, \beta)$. The first term on the right hand side of (2) indicates the analysis model (the original model of interest), relating the response variable $y$ to the covariates. If $y$ has an ordinal nature this model will have the form of a cumulative logit model. The $\beta$ is a vector of parameters for this model which needs to be estimated. The second term represents the distribution of the missing covariate(s) with nuisance parameters $\alpha$. According to the nature of the missing covariate(s) this term could take different forms. In this paper we will be concerning ourselves with continuous missing covariates which we will assume to follow a Normal distribution. For continuous $x_{mis,i}$ we have:

$$P(x_{mis,i}|y_i, x_{obs,i}; \gamma) = \frac{P(y_i|x_i; \beta) P(x_{mis,i}|x_{obs,i}; \alpha)}{\int P(y_j|x_j; \beta) P(x_{mis}|x_{obs,j}; \alpha) dx_{mis}}$$

Rubin (1987) distinguishes proper from improper imputation and suggests a Bayesian approach to the process. In improper imputation some estimate $\tilde{\gamma}$ is substituted for $\gamma$ and draws are subsequently made from the imputation distribution. He shows that this approach doesn't in general provide valid inferences and so calls it "improper". In "proper" imputation however, a prior distribution is first chosen for $\gamma$ and then draws are made from the joint posterior distribution of $(\gamma, x_{mis}|x_{obs}, y)$ using MCMC methods. In other words, in order to fill-in the missing values, draws can be made from the posterior predictive distribution, via the Gibbs Sampler (Ibrahim *et al.* (2005)):

$$P(x_{mis,i}|y_i, x_{obs,i}) \propto \iint P(x_{mis,i}|y_i, x_{obs,i}, \gamma) \pi(\alpha, \beta|y, x_{obs}) d\alpha d\beta \tag{3}$$

In the Equation (3), $\pi(\alpha,\beta|y,x_{obs})$ is the joint posterior distribution for $(\alpha,\beta)$ based on the observed data.

The imputation distribution could very well differ from the analysis model. When the imputation model uses more information compared to the analysis model, the analysis tends to be more efficient than expected and the confidence intervals would for example have greater than nominal convergence rates. Rubin (1996) calls this "super efficiency". For a more detailed discussion see Rubin and Schenker (1986), Fay (1992, 1996) and Little and Rubin (2002).

The posterior distribution of model parameters $\gamma$ under the MAR assumption, can be written as

$$P(\gamma|y,x_{obs},R)=P(\gamma|y,x_{obs})=c\times\pi(\gamma)\times f(y,x_{obs}|\gamma),$$

in which $\pi(\gamma)$ is the prior distribution of parameters, $f$ is the density of the observed data, $c$ is a constant and similar to before, $x_{obs}$ represents the observed part of the covariates. Now by simple probability theory, we have

$$\begin{aligned} P(\gamma|y,x_{obs}) &= \int P(\gamma,x_{mis}|y,x_{obs})\,dx_{mis} \\ &= \int P(\gamma|y,x_{obs},x_{mis})P(x_{mis}|y,x_{obs})\,dx_{mis}. \end{aligned} \tag{4}$$

The equation (4) suggests that we can first generate $x_{mis}$ from the posterior distribution $P(x_{mis}|y,x_{obs})$ and then use the values to generate $\gamma$ from $P(\gamma|y,x_{obs},x_{mis})$.

What proper MI does is that it consistently estimates from Equation (4) by averaging over the missing values

$$P(\gamma|y,x_{obs}) \approx \frac{1}{D}\sum_{d=1}^{D}P(\gamma|y,x_{obs},x_{mis}^{(d)}),$$

where $x_{mis}^{(d)}$'s are draws from the posterior predictive distribution $P(x_{mis}|y,x_{obs})$.

Other distribution properties, like expectations and variances can in turn be estimated in a similar fashion for vector of parameter $\gamma$ and sufficiently large positive integer $D$ as follows

$$\mathrm{E}\left(\gamma \middle| y, x_{obs}\right) \approx \int \gamma \frac{1}{D}\sum_{d=1}^{D} P\left(\gamma \middle| y, x_{mis}, x_{mis}^{(d)}\right) d\gamma = \bar{\gamma},$$

$$Var\left(\gamma \middle| y, x_{obs}\right) \approx \frac{1}{D}\sum_{d=1}^{D} V_d + \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\gamma} - \bar{\gamma}\right)^2 = \bar{V} + B,$$

in which $\bar{\gamma} = \sum_{d=1}^{D} \hat{\gamma}_d / D$, $\hat{\gamma}_d = \mathrm{E}\left(\gamma \middle| y, x_{obs}, x_{mis}^{D}\right)$; $V_d = Var\left(\gamma \middle| y, x_{obs}, x_{mis}^{d}\right)$ the posterior variance from the $d$ 'th imputed data-set, $\bar{V} = \sum_{d=1}^{D} V_d / D$ and $B = \sum_{d=1}^{D}\left(\hat{\gamma}_d - \bar{\gamma}\right)/\left(D-1\right)$.

## APPLICATION: STEATOSIS STUDY

In this section first a dataset with ordinal responses is introduced and then missingness is generated using MAR mechanism, on a continuous covariate. Four approaches to MI for this data are then discussed and implemented.

### The Data

Steatosis is the infiltration of liver cells with fat, usually associated with disturbance of the metabolism. It is measured by means of sonography and on an ordinal scale with three levels:

1. The individual has no sign of Steatosis (None),
2. The individual has signs of mild Steatosis (Mild),
3. The individual has signs of severe Steatosis (Severe).

The data used here is recorded at Taleghani hospital, Tehran, and from 60 diabetic patients as part of an observational study on Steatosis and overweight. Explanatory variables (covariates) used in the model include: age, sex, duration (indicating the duration in years that the individual has been diabetic), blood pressure (in mmHG) and Body Mass Index (BMI,

calculated by dividing an individual's weight in kilograms by his/her height in meters squared).

TABLE 1: Sample mean and standard error of continuous variables in the study

| Variable | Mean | SE |
|---|---|---|
| **Age** $(z_1)$ | 60.317 | 9.177 |
| **Duration** $(z_2)$ | 9.717 | 5.969 |
| **Blood Preasure** $(Z_3)$ | 136.800 | 23.585 |
| **BMI** $(x)$ | 28.044 | 3.923 |

TABLE 2: An overview of discrete variables in the study

| Variable | Levels | Frequency | Percentage |
|---|---|---|---|
| **Steatosis** $(y)$ | None | 12 | 20% |
| | Mild | 18 | 30% |
| | Severe | 30 | 50% |
| **Sex** $(z_4)$ | Female | 24 | 40% |
| | Male | 36 | 60% |

Tables 1 and 2 show an overview of the dataset. Results of Table 1 shows that in average sample includes old people with high BMI and blood pressure. Table 2 shows that half of the people in the sample have severe Steatosis.

In order to examine the performance of different MI procedures, the following missing-data mechanism is implemented on the continuous covariate BMI:

$$P\left(R_i = 0 \middle| y_i, \phi_0, \phi_1, \phi_2\right) = \Phi\left(\phi_0 + \phi_1 I_{\{y_i=1\}} + \phi_2 I_{\{y_i=2\}}\right). \tag{5}$$

As 20% of respondents had no sign of Steatosis $\left(y_i = 1\right)$ and 30% showed signs of mild Steatosis $\left(y_i = 2\right)$, thus a selection of $\phi_0 = -0.9, \phi_1 = -0.2$ and $\phi_2 = -0.2$, would result in an approximate missing rate of 15% $\left[\Phi\left(-0.9 - 0.2\left(0.2\right) - 0.2\left(0.3\right)\right) = \Phi\left(-1\right) = 0.1587\right]$ which is convenient for our purpose.

The BMI values for all the respondents and also for male and female respondents separately, were checked by using the Kolmogorov-Smirnov (KS) test and a Normal distribution was found appropriate in all three cases (respective p-values were 0.9534, 0.9121 and 0.9067). Alternatively, the Anderson-Darling Test may be used to test for normality. This test gives more weight to the tails compared to the more common KS test. In our data, normality is accepted for three cases by this method.

TABLE 3: Parameter estimates and standard errors for the original data and complete-case analysis.

| Effect | Original Data | | Complete-Case Analysis | |
|---|---|---|---|---|
| | Estimate | SE | Estimate | SE |
| **Threshold 1** | 0.8468 | 2.7514 | 0.2086 | 2.8640 |
| **Threshold 2** | 2.5353 | 2.7727 | 2.2254 | 2.8862 |
| **Age** | -0.0472 | 0.0291 | -0.0429 | 0.0298 |
| **Duration** | 0.0532 | 0.0466 | 0.0203 | 0.0506 |
| **Blood Preassure** | -0.0127 | 0.0117 | -0.0148 | 0.0125 |
| **Sex (Male)** | -1.2798 | 0.5987 | -1.0989 | 0.6343 |
| **BMI** | 0.2619 | 0.0835 | 0.2556 | 0.0856 |

Table 3 shows parameter estimates and standard errors for the original data (with no missing data) and for the complete-case analysis (were all cases with missing values were omitted from the study). The original-data results show that blood pressure, duration and age are not significant (p-values = 0.1389, 0.1268 and 0.0524 respectively), whilst BMI and Sex have some significant effects (p-values = 0.0009 and 0.0163, respectively). The odds of lower levels of Steatosis in men are less than that of in women and increase in the BMI index increases the odds of higher levels of Steatosis.

## Four approaches to MI

In this subsection four approaches (methods A, B, C and D) to Multiple Imputation will be discussed and applied to the introduced dataset. The first three of these approaches (methods A, B and C) use proper imputation method and the last one (method D) uses the improper approach. These approaches differ in the amount of information they use for imputation

- Method A: In the first method the missing variable *(x)* is modeled strictly on the other covariates $(z_1,...,z_4)$,

$$\left(x_i \middle| z_{1i}, z_{2i}, z_{3i}, z_{4i}\right) \sim N\left(\mu_i, \sigma^2\right); \tag{6}$$

in which $\mu_i = \sum_{j=0}^{4} \alpha_j z_{ji}$, and it is defined that: $z_{0i} = 1$ for all $i$. Now the joint distribution of $(y, x)$ given $z$ can be expressed as:

$$P\left(y, x \middle| z\right) = P\left(y \middle| x, z\right) P\left(x \middle| z\right).$$

TABLE 4: Parameter estimates and standard errors for four methods of MI.

| Effect | Method A | | Method B | | Method C | | Method D | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | SE | Estimate | SE | Estimate | SE | Estimate | SE |
| **Threshold 1** | 0.2007 | 2.9456 | 1.3062 | 2.8611 | 1.2513 | 2.7703 | 0.4867 | 2.7969 |
| **Threshold 2** | 1.8413 | 1.2243 | 3.0780 | 0.9784 | 3.0404 | 0.6725 | 2.1875 | 0.9086 |
| **Age** | 0.0459 | 0.0289 | -0.0477 | 0.0293 | -0.0484 | 0.0294 | -0.0464 | 0.0293 |
| **Duration** | 0.0508 | 0.0489 | 0.0496 | 0.0474 | 0.0485 | 0.0470 | 0.0513 | 0.0471 |
| **Blood Pressure** | -0.0101 | 0.0119 | -0.0123 | 0.0122 | -0.0127 | 0.0122 | -0.0125 | 0.0118 |
| **Sex (Male)** | -1.2087 | 0.6246 | -1.3180 | 0.6134 | -1.2843 | 0.6038 | -1.2563 | 0.6072 |
| **BMI** | 0.2208 | 0.0954 | 0.2808 | 0.0872 | 0.2823 | 0.0886 | 0.2473 | 0.0833 |

- Method B: In the second method the response variable $y$ (with three ordinal categories) is also included in the model, using two dummy variables that is

$$\left(x_i \middle| y_i, z_{1i}, z_{2i}, z_{3i}, z_{4i}\right) \sim N\left(\mu_i, \sigma^2\right) \tag{7}$$

where $\mu_i = \sum_{j=0}^{4} \alpha_j z_{ji} + \phi_1 y_{2i}^{*}$, and again $z_{0i} = 1$. It is obvious that this method is using more information for imputing the missing values compared to the one before.

- Method C: In the third method, separate models are fitted for each level of the ordinal response variable that is

$$\left(x_i \middle| y_i, z_{1i}, z_{2i}, z_{3i}, z_{4i}\right) \sim N\left(\mu_{yi}, \sigma_{yi}^2\right) \quad for \quad y_i = 1, 2, 3; \tag{8}$$

where $\mu_{yi} = \sum_{j=0}^{4} \alpha_{jy_i} z_{ji}$, and $z_{0i} = 1$. Equation (8) can also be written as a full model which includes the interactions between the response variable $y$ and the covariates $z_1, ..., z_4$. This model is using all the available information in the data for imputation.

- Method D: In this method we use an improper imputation approach where we generate missing values by

$$\left(x_i \,|\, y_i, z_{1i}, z_{2i}, z_{3i}, z_{4i}, \gamma\right) \sim N\left(\mu_{y_i}, \sigma^2_{y_i}\right) \quad \text{for } y_i = 1,2,3 \tag{9}$$

where $\mu_i = \sum_{j=0}^{4} \alpha_j z_{ji} + \phi_1 y^*_{1i} + \phi_2 y^*_{2i}$, $\gamma = (\alpha, \phi)$ and $z_{0i} = 1$. For computational convenience, we will use a complete-case estimate of $\gamma$ for generating values from the distribution in Equation (9).

TABLE 5: Parameter estimates and standard errors for the original data and complete-case analysis with insignificant covariates removed.

| Effect | Original Data | | Complete-Case Analysis | |
|---|---|---|---|---|
| | Estimate | SE | Estimate | SE |
| **Threshold 1** | 3.6013 | 1.9029 | 3.7141 | 1.9759 |
| **Threshold 2** | 5.1767 | 1.9665 | 5.6069 | 2.0705 |
| **Sex (Male)** | -1.1983 | 0.5828 | -1.1126 | 0.6200 |
| **BMI** | 0.2096 | 0.0741 | 0.2187 | 0.0768 |

TABLE 6: Parameter estimates and standard errors for four methods of MI with insignificant covariates removed.

| Effect | Method A | | Method B | | Method C | | Method D | |
|---|---|---|---|---|---|---|---|---|
| | Estimate | SE | Estimate | SE | Estimate | SE | Estimate | SE |
| **Threshold 1** | 2.7200 | 1.9629 | 4.5235 | 1.9610 | 3.9097 | 1.8700 | 4.1505 | 1.8811 |
| **Threshold 2** | 4.2529 | 0.8261 | 6.2006 | 0.6413 | 5.5315 | 0.4422 | 5.7840 | 0.3330 |
| **Sex (Male)** | -1.0659 | 0.5795 | -1.3159 | 0.5923 | -1.243 | 0.5873 | -1.2507 | 0.5927 |
| **BMI** | 0.1736 | 0.0760 | 0.2484 | 0.0770 | 0.2213 | 0.0728 | 0.2315 | 0.0737 |

The "mi" package (Yu-Sung *et al*. (2009)) in the statistical software R, was used to create $M = 5$ draws from each distribution. A cumulative logit model was also implemented in R for the analysis model using the function "polr (MASS)". Table 4 shows parameter estimates and respective standard errors obtained using each method.

Due to the rather small sample size of 60 individuals, and in order to further see the performance of the four MI methods, insignificant covariates in the model for the original data were omitted from the analysis and all six models were fitted again using only the covariates "Sex" and "BMI". The results can be seen in Tables 5 and 6.

For the covariate effects, the Complete-Case results show some bias and inefficiency regarding coefficient estimates. The bias is more or less corrected in all imputation methods, although for this dataset, method B shows a slightly better performance for the effects of age, duration and BMI compared to that of the method C on the other hand is less biased for the effects of blood pressure and sex. The gain of efficiency is again addressed in all four methods, compared to Complete-Case (CC) analysis, although again methods B and C perform better than methods A and D.

The superiority of the methods B and C can be better seen when insignificant covariates are removed from the model. Here, CC analysis and imputation method A fail to show the significance of the covariate sex. This is corrected in both imputation methods B and C. The method C, compared to the method B, shows less bias and more efficiency for all model parameters. Method D, due to the use of an improper approach, gives biased estimates of cut points parameters, comparing with the results of the original data.

## DISCUSSION

It can be concluded that use of proper methods of imputation gives better results. It can also be concluded that as more information is used in the imputation procedure, estimates of model parameters tend to be less biased and more efficient.

## ACKNOWLEDGEMENTS

This work is guided by the statistical research group of “Bayesian Inferential Statistics” in Shahid Beheshti University.

## REFERENCES

Agresti, A. 2002. *Categorical Data Analysis.* New Jersey: Wiley.

Berridge, D. M. 1995. Modeling Ordinal Recurrent Events. *Journal of Statistical Planning and Inference*, **47**: 71-78

Berridge, D. M. and Dos Santos, D. M. 1996. Fitting a random effects model to ordinal recurrent events using existing software. *J. Stat. Comput. Simu.,* **55**(1-2): 73-86.

Diggle, P. J. and Kenward, M. G. 1994. Informative Drop-out in Longitudinal Data Analysis. *Appl. Stat.,* **43**: 49-93.

Fahrmeir, L. and Tutz, G. 1994. *Multivariate Statistical Modeling Based on Generalized Linear Models*. New-York: Springer-Verlag.

Fay, R. E. 1992. When are inferences from multiple imputation valid? *Proc. Survey Research Methodology Sec., American Statistical Association,* 227-232.

Fay, R. E. 1996. Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association,* **91**: 490-498.

Horton, N. J. and Laird, N. M. 1998. Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research,* **8**: 37-50.

Ibrahim, J.G. 1990. Incomplete data in generalized linear models. *Journal of the American Statistical Association,* **85**: 765-769.

Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. 1999. Monte Carlo EM for missing covariates in parametric regression models. *Biometrics,* **55**: 591-596.

Ibrahim, J. G., Chen, M.H., Lipsitz, S. R. and Herring, A. H. 2005. Missing-data methods for generalized linear models: a comparative review. *Journal of the American Statistical Association,* **100**: 332-346.

Lipsitz, S. R and Ibrahim, J. G. 1996. A conditional model for incomplete covariates in parametric regression models. *Biometrika*, **83**(4): 916-922.

Little, R. J. A. 1992. Regression with Missing X's: a review. *Journal of the American Statistical Association,* **87**: 1227-1237.

Little, R. J. A. and Rubin, D. B. 2002. *Statistical Analysis with Missing Data.* New Jersey: Wiley.

Little, R. J. A. and Schluchter, M. D. 1985. Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika,* **72**: 497-512.

McCullagh, P. and Nelder, J. 1989. *Generalized Linear Models, 2nd Edition.* New-York: CRC Press.

McCullagh, P. 1980. Regression models for ordinal data (with discussion). J. Roy. *Statist. Soc.*, **B42**: 109-142.

Nelder, J. and Wedderburn, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society, Series A,* **135**: 370-384.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *J. Amer. Statist. Assoc.,* **90**: 106-121.

Rubin, D. B. 1978. Multiple Imputation in sample surveys *Proc. of Survey Research Methodology Sec., American Statistical Association 1978,* 20-34.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New Jersey: Wiley.

Rubin, D. B. 1996. Multiple Imputation after 18+ years (with discussion) *Journal of the American Statistical Society,* **91**: 473-489.

Rubin, D. B. and Schenker, N. 1986. Multiple Imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association,* **81**: 366-374.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data.* New-York: CRC Press.

Schafer, J. L. and Graham, J. W. 2002. Missing Data: our view of the state of the art. *Psychological Methods,* **7**(2): 147-177.

Snell, E. J. 1964. A scaling procedure for ordered categorical data. *Biometrics*, **20**(3): 592-60

Tutz, G. 2005. Modelling of repeated ordered measurements by isotonic sequential regression. *Stat. Mod.,* **5**: 269-287.

Yu-Sung, Su, Gelman, A, Jennifer, H and Masanao Yajima. 2009. Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software,* (forthcoming). Available on: http://www.stat.ucla.edu/ yajima/Publication/mipaper.rev04.pdf.