The Effect of Different Imputation Methods on Analytical Statistics of Simple Linear Regression

Dr. Jann-Huei Jinn
Department of Mathematics and Statistics
Grand Valley State University
Allendale, Michigan 49401
e-mail: jinnj@gvsu.edu

1. Introduction

Most surveys face problems of both unit and item nonresponse. Unit nonresponse occurs when no information is collected from a sample unit, and item nonresponse occurs when most of the questions for a unit are answered, but for certain questions either no answer is given or the answer is judged to be inconsistent with other answers and is deleted during editing. Compensation for unit nonresponse is usually carried out by some form of weighting adjustment, while compensation for item nonresponse is commonly made by imputation; i.e., by assigning one or more values for each missing response. This is, a priori, an appealing general purpose strategy given the vary large size and complexity of many sample surveys. In their excellent review paper Kalton and Kasprzyk (1982) describe the desirable features of imputation: " First… it aims to reduce biases in survey estimates from missing data…Second, by assigning values at the micro level and thus allowing analyses to be conducted as if the data set were complete, imputation makes analyses easier to conduct and results easier to present. Complex algorithms to estimate population parameters in the presence of missing data (e.g., the EM algorithm of Dempester, Laird and Rubin, 1977) are not required. Third, the results obtained from different analyses are bound to be consistent, a feature which need not apply with an incomplete data set."

On the other hand, imputation has its dangers. Kalton and Kasprzyk also point out that imputation "does not necessarily lead to estimates that are less biased than those obtained from the incomplete data set; indeed the biases could be much greater, depending on the imputation procedure and the form of estimate. There is also the risk that analysts may treat the completed data set as if all the data were actual responses, thereby overstating the precision of the survey estimates." Even if the biases of univariate statistics are reduced, the relationship between variables may be distorted (see Santos, 1981a.)

Undoubtedly, when used, imputation should be applied cautiously and analysts of the completed data set should be fully warned of the potential dangers created by the imputation. At least, the imputed values should be flagged, so that the careful analyst can assess the effect that imputations may have on the analysis. The flagging concept is important because, among other things, it also allows for the data analyst to "second guess" the survey statistician and apply whatever missing data procedure he wishes.

It is our experience that the overwhelming majority of secondary data analysts proceed as if the completed data set contains only observed responses, and it is our belief that they will continue to do so. The objective of our research is to try to discern the effect on the properties of standard statistical techniques of proceeding in this way. That is, we view imputation methodology from the perspective of the secondary data analyst who does not take cognizance of the presence of imputed values in the data set.

Let $\hat{\theta}_c$ denote a random quantity of interest which contains both observed and imputed values (e.g., $\hat{\theta}_c = \hat{\beta}_{1c}$ in (2.2.8).) Moments of $\hat{\theta}_c$ are obtained conditionally on the observed values of X, and are evaluated by averaging using (a) the model specification (e.g. (2.1.1)) and (b) any random features of the imputation method. For these evaluations we assume that the model in (a) is correct, because this is the assumption that the secondary data analyst will make. Regarding (b), we average over the assignments of respondents' values to nonrespondents for RI, RC and RRS (see Section 2.2.2 for definitions.)

Although finding properties of standard statistical techniques when imputed values are used is a very important problem, the research is difficult. Once a statistic, $\hat{\theta}(x_1,...,x_n)$, is perturbed by including imputed values, any degree of symmetry in $\hat{\theta}(x_1,...,x_n)$ is lost. Thus, derivations of moments of statistics are tedious and the resulting expressions are cumbersome. Thus, comparisons of properties of $\hat{\theta}_c$ when alternative imputation methods are employed are difficult to make. For these reasons we have considered simple statistical models and simple analytical objectives, and have made a few simplifying assumptions. Even so, it has been difficult to obtain the results presented in this paper (see, for example, Appendix B) and only a limited number of analytical comparisons are possible. Specifically, we have considered the simple linear regression model

$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i \qquad (1.1)$$

where the errors are normally and independently distributed with $E(\varepsilon_i | X_i)=0$ and $Var(\varepsilon_i | X_i)=\sigma^2$. Let $\hat{\beta}_{oc}$, $\hat{\beta}_{1c}$ and $\hat{\sigma}_c^2$ denote the estimates of β_o, β_1, and σ^2 where "c" indicates that these are estimates from a completed data set. We present the results of a theoretical investigation of the effect that commonly used imputation methods have on the properties of confidence for β_o and β_1. Specifically, we will try to identify "good" imputation procedures and conditions where specific imputation procedures work well. For an arbitrary imputation method we evaluate: (a) properties of regression residuals, $e_i = Y_i - \hat{Y}_i$, i=1,…,n, (b) point estimation of β_o and β_1, (c) point estimation of σ^2, (d) separate confidence intervals for β_o and β_1. By examining properties of the residuals from the regression we hope to be able to discriminate among the imputation methods; i.e., for "good" imputation methods the properties of the residuals would more closely agree with those from a random sample of the same size. Evaluating the biases of the point estimators and the properties of the associated confidence intervals will determine which imputation methods are the better ones.

Most of the research concerning incomplete data makes the assumption that the cause of values being unobserved is unrelated to the relationships under study. Rubin (1976) has made a precise distinction between two types of incomplete data, i.e., missing at random (MAR) and missing not at random. In this paper we assume that any missing data is missing at random (MAR), and that (2.1.1) adequately describes the relationship between Y and X for the entire population under study.

While the specification just describe is a simple one, there are practical situations where (2.1.1) is a model of interest and the assumptions associated with (2.1.1) are not unrealistic. Our assumptions that any missing data are MAR at random is for convenience. Considering nonignorable missing data mechanism is essential, but doing so will add enormous complexity to the model and to the resulting derivations and comparisons.

While imputation has been used for a long period of time, systematic research on properties of imputation methods is recent. Early published papers whose objectives were to determine analytical properties of estimators containing both observed and imputed data include Bailar and Bailar (1978), Bailar and Corby (1978), Ernest (1978) and Platek, Singh and Tremblay (1978).

Almost all of these researchers investigate properties of univariate descriptive statistics such as means and totals. Analytical uses of survey data have only been considered in the very simplest situation: Herzog and Rubin (1983) study the effects of several imputation methods on the usual confidence interval for a population mean. Bivariate statistics such as sample covariance, correlation and regression coefficients have been studied by Santos (1981a,b); these results are summarized in Kalton and Kasprzyk (1982). However, only biases of point estimators are considered.

The paper is organized as follows. The notation is defined and the imputation methods are described in section 2. In Section 3 we consider the simple linear regression model (1.1) . Our analytical procedure is described first and the properties of a selected set of statistics are presented for each of the imputation methods. The results are in Sections 3.1 and 3.2. Comparisons of the alternative imputation methods are made in Section 3.3. while Section 3.4 summarizes the results in Section 3.3. Section 4 is a discussion of the difficulties associated with uncritical use of data sets containing imputed values. Since the algebra needed to derive the results is complicated and tedious we illustrated by outlining some of the derivations in the Appendix.

2. Notation and Imputation Schemes

2.1 Notation

Suppose we have the simple linear regression model

$$Y = \beta_o + \beta_1 X + \varepsilon \qquad (2.1.1)$$

with $E(\varepsilon|X) = 0$ and $Var(\varepsilon|X) = \sigma^2$ where σ^2 is unknown. Given a random sample of size n with X observed for all sampled units, let $\{x_{ri} : i = 1,...,r\}$ and $\{x_{mj} : j = r+1,...,n\}$ denote the observed X values which correspond to the r observed Y values and m missing Y values, respectively. When $\{x_{mj} : j = r+1,...,n\}$ are also missing we use $\{x^*_{mj} : j = r+1,...,n\}$ to denote the imputed values. When X has no missing values, define

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\left\{\sum_{i=1}^{r} x_{ri} + \sum_{j=r+1}^{n} x_{mj}\right\}}{n}$$

as the overall sample mean, $\bar{x}_r = \frac{\sum_{i=1}^{r} x_{ri}}{r}$ as the mean corresponding to the r respondents on Y, and $\bar{x}_m = \frac{\sum_{j=r+1}^{n} x_{mj}}{m}$ as the mean corresponding to the m nonrespondents on Y. Define

$s^2_{rx} = \frac{\sum_{i=1}^{r} (x_{ri} - \bar{x}_r)^2}{r}$ as the sample variance of X corresponding to the r respondents on Y.

If X has imputed values, define $\bar{x}_c = \frac{\sum_{i=1}^{r} x_{ri} + \sum_{j=r+1}^{n} x^*_{mj}}{n}$ to be the (completed) sample mean of the r observed and m imputed values of X.

Suppose that an auxiliary variable Z is used to create L imputation cells. The variable Z may be a composite variable formed from several basic auxiliary variables; for example, Z may represent the cells in a cross classification of, say, age, sex, and race.

In cell h, let (n_h, r_h, m_h) denote, respectively, the numbers of sampled units, potential donors (units with responses to the item,) and recipients (units with missing responses to the item.) In the h^{th} imputation cell, let x_{rhi}, x_{mhj}, and x^*_{mhj} correspond to x_{ri}, x_{mj}, and x^*_{mj} as defined earlier. When X has no missing data, define

$$\bar{x}_h = \frac{\sum_{i=1}^{n_h} x_{hi}}{n_h} = \frac{\sum_{i=1}^{r_h} x_{rhi} + \sum_{j=r_h+1}^{n_h} x_{mhj}}{n_h}$$

as the cell mean of size n_h. Also, $\bar{x}_{rh}$, $\bar{x}_{mh}$ and s^2_{rxh} just like $\bar{x}_r$, $\bar{x}_m$, and s^2_{rx} but for cell h.

Define $\bar{x}_r = \frac{\sum_{h=1}^{L} r_h \bar{x}_{rh}}{r}$ and $\bar{x}_m = \frac{\sum_{h=1}^{L} m_h \bar{x}_{mh}}{m}$ as the means corresponding to r respondents on Y and corresponding to m nonrespondents on Y, respectively.

If X has imputed values, define $\bar{x}_c = \frac{r\bar{x}_r + \sum_{h=1}^{L} \sum_{j=r_h+1}^{n_h} x^*_{mhj}}{n}$ as the sample mean of the completed data set.

Assuming the linear regression model (2.1.1), a random sample of size n and no missing values, the usual unbiased estimators of β_o, β_1, and σ^2 are $\hat{\beta}_o$, $\hat{\beta}_1$, and $\hat{\sigma}^2$, where

$$\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}, \tag{2.1.2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{2.1.3}$$

and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} \tag{2.1.4}$$

where $\hat{y}_i$ is the predicted value of y_i corresponding to x_i. In this chapter, we assume that the dependent variable Y always has missing values, and define

$$y_i^+ = \begin{cases} y_{ri} & if\ Y\ is\ observed; \quad i = 1,...,r \\ y^*_{mj} & if\ Y\ is\ mis\sin g; \quad j = r+1,...,n \end{cases} \tag{2.1.5}$$

When there is missing data on X, x_i^+ is defined analogously.

Let $\hat{\beta}_{oc}$, $\hat{\beta}_{1c}$, and $\hat{\sigma}_c^2$ denote the estimators of β_o, β_1, and σ^2 using the completed data set. When there are some missing values for X, define the predicted value of y_i^+, $\hat{y}_{ic}^+$, for a given x_i^+ as:

$$\hat{y}_{ic}^+ = \begin{cases} \hat{y}_{ric} = \hat{\beta}_{oc} + \hat{\beta}_{1c} x_{ri} & if\ X\ is\ observed \\ \hat{y}^*_{mjc} = \hat{\beta}_{oc} + \hat{\beta}_{1c} x^*_{mj} & if\ X\ is\ mis\sin g \end{cases} \tag{2.1.6.a}$$

When X has no missing values, define the predicted value of y_i^+, $\hat{y}_{ic}^+$, for a given x_i as:

$$\hat{y}_{ic}^+ = \begin{cases} \hat{y}_{ric} = \hat{\beta}_{oc} + \hat{\beta}_{1c} x_{ri} & if \quad Y \quad is \quad observed \\ \hat{y}_{mjc} = \hat{\beta}_{oc} + \hat{\beta}_{1c} x_{mj} & if \quad Y \quad is \quad mis\sin g. \end{cases} \qquad (2.1.6.b)$$

Then, we define

$$\hat{\beta}_{oc} = \begin{cases} \bar{y}_c - \hat{\beta}_{1c}\bar{x}_c & if \quad X \quad has \quad imputed \quad data \\ \bar{y}_c - \hat{\beta}_{1c}\bar{x} & Otherwise \end{cases}, \qquad (2.1.7)$$

$$\hat{\beta}_{1c} = \begin{cases} \dfrac{\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)(y_i^+ - \bar{y}_c)}{\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)^2} & if \quad both \quad X \quad and \quad Y \quad have \quad mis\sin g \quad values \\ \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i^+ - \bar{y}_c)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} & if \quad only \quad Y \quad has \quad mis\sin g \quad values \end{cases}, \qquad (2.1.8)$$

and

$$\hat{\sigma}_c^2 = \frac{\sum_{i=1}^{n}(y_i^+ - \hat{y}_{ic}^+)^2}{n-2} =$$

$$\begin{cases} \dfrac{\sum_{i=1}^{r}(y_{ri} - \hat{y}_{ric})^2 + \sum_{j=r+1}^{n}(y_{mj}^* - \hat{y}_{mjc}^*)^2}{n-2} & if \quad both \quad X \quad and \quad Y \quad have \quad mis\sin g \quad values \\ \dfrac{\sum_{i=1}^{r}(y_{ri} - \hat{y}_{ric})^2 + \sum_{j=r+1}^{n}(y_{mj}^* - \hat{y}_{mjc})^2}{n-2} & if \quad only \quad Y \quad has \quad mis\sin g \quad values \end{cases}. \qquad (2.1.9)$$

Let $\hat{\beta}_{or}$, $\hat{\beta}_{1r}$ and $\hat{\sigma}_r^2$ denote the unbiased estimators of β_o, β_1, and σ^2 using only the observed data set $\{(x_{ri}, y_{ri}) : i = 1,...,r\}$:

$$\hat{\beta}_{or} = \bar{y}_r - \hat{\beta}_{1r}\bar{x}_r, \qquad (2.1.10)$$

$$\hat{\beta}_{1r} = \frac{\sum_{i=1}^{r}(x_{ri} - \bar{x}_r)(y_{ri} - \bar{y}_r)}{\sum_{i=1}^{r}(x_{ri} - \bar{x}_r)^2}, \qquad (2.1.11)$$

and

$$\hat{\sigma}_r^2 = \frac{\sum_{i=1}^{r}(y_{ri} - \hat{y}_{ri})^2}{r-2}. \qquad (2.1.12)$$

When there are imputed values, define the "observed" residuals for a completed data set as follows:

$$e_{ic} = y_i^+ - \hat{y}_{ic}^+ = \begin{cases} e_{ric} = y_{ri} - \hat{y}_{ric} & r \text{ pairs of } (X,Y) \text{ observed data} \\ e_{mjc} = y_{mj}^* - \hat{y}_{mjc} & X \text{ all observed with } m \text{ imputed } Y \text{ values} \\ e_{mjc}^* = y_{mj}^* - \hat{y}_{mjc}^* & m \text{ pairs of } (X,Y) \text{ imputed data} \end{cases} \qquad (2.1.13)$$

where the terms in (2.1.13) are defined in (2.1.5) and (2.1.6).

2.2 Imputation Schemes

The following imputation methods are considered in this paper:

(a) Mean Overall Imputation (MO). This method imputes a constant, the overall mean of the respondents, to all missing values.

(b) Random Imputation (RI). Given a sample of size n with m=n-r missing values. A random sample of size m is taken with replacement from the r observed values. The selected respondents act as "donors" and their values are randomly assigned to the nonrespondents.

(c) Mean Imputation Within Cells (MC). This method assigns each sampled unit to one of L mutually exclusive and exhaustive imputation cells. The cells are defined by the values of the auxiliary variables, assumed to be known for each sample member. Within each cell the observed cell mean is assigned to each of the nonrespondents in the cell.

(d) Random Imputation Within Cells (RC). This method is a simple generation of RI; it is applied within imputation cells. Randomly selected respondents within each cell are used to assign values to the nonrespondents in the same cell.

(e) Simple Regression Prediction Imputation (RG). This method uses the respondent data $\{(x_{ri}, y_{ri}) : i = 1,...,r\}$ to estimate regression coefficients. When X values are all present, the missing Y values are imputed by the predicted values from the regression equation (for example, $y_{mj}^* = \hat{\beta}_{or} + \hat{\beta}_{1r} x_{mj}$ where $\hat{\beta}_{or}$, $\hat{\beta}_{1r}$ are as given by (2.1.10) and (2.1.11). If X and Y both have missing values, we impute for missing X values by using the overall mean of the respondents, i.e., $x_{mj}^* = \bar{x}_r$, and then the missing Y values are imputed by $y_{mj}^* = \hat{\beta}_{or} + \hat{\beta}_{1r}\ \bar{x}_r$.

(f) Random Regression Imputation (RRS, RRN). The RG method imputes values directly from the estimated regression line. Random residual errors can be added to the regression prediction to provide dispersion about the regression line. The residual can be obtained in various ways, including: (1) Draw a random sample of size m with replacement from the r observed residuals, $\left\{e_{ri} = y_{ri} - \hat{\beta}_{or} - \hat{\beta}_{ir} x_{ri}\right\}$(2) A residual can be chosen at random from a distribution with mean zero and variance $\hat{\sigma}_r^2$ where $\hat{\sigma}_r^2$ is the residual variance of the regression using the respondents' data. Thus, the RRS method imputes the missing Y values by $y_{mj}^* = \tilde{y}_{mj} + \tilde{e}_{mj}$, where $\tilde{y}_{mj}$ is the regression prediction for unit j and $\tilde{e}_{mj}$ is a randomly selected respondent residual. The RRN method imputes the missing Y values by $y_{mj}^* = \tilde{y}_{mj} + e_{mj}$, where e_{mj} is randomly drawn from a distribution with mean zero and variance $\hat{\sigma}_r^2$.

When both X and Y have missing values, we impute the missing X values by $x_{mj}^* = \bar{x}_r$, then use RRS or RRN to impute the missing Y values. While (a) and (b) are simplistic methods included for illustration, (d) approximates a fixed replicate of the "Statistical Matching Procedure" used for the CPS (Current Population Survey) March Income Supplement and RRS and RRN in (f) are prototypes for sensible procedures when there are good covariates available.

2.3 Simple Linear Regression

Suppose the relationship between an independent variable X and a dependent variable Y is given by (2.1.1). Assume that m Y values are missing at random (MAR). We consider two cases: (i) X has no missing values (ii) X has missing values. When there is no missing data for a random sample of size n, the usual unbiased estimators of β_o, β_1, and σ^2 are given by (2.1.2), (2.1.3), and (2.1.4). If we assume that the variations of the observations about the line are normal, the 100(1-α)% confidence intervals for β_o and β_1 are given by, respectively,

$$\hat{\beta}_o \pm t(n-2,1-\frac{\alpha}{2})\left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]^{1/2}\hat{\sigma} \qquad (2.3.1)$$

and

$$\hat{\beta}_1 \pm \frac{t(n-2,1-\frac{\alpha}{2})\hat{\sigma}}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2}} \qquad (2.3.2)$$

where $t(n-2,1-\frac{\alpha}{2})$ is the $100\,(1-\frac{\alpha}{2})$ percentage point of a t-distribution with n-2 degrees of freedom.

Let $\hat{\beta}_{oc}$, $\hat{\beta}_{1c}$, and $\hat{\sigma}_c^2$ (see (2.1.7), (2.1.8) and (2.1.9)) denote the estimators of β_o, β_1, and σ^2 using the completed data. The secondary data analyst will use

$$\hat{\beta}_{oc} \pm t(n-2,1-\frac{\alpha}{2})\left\{\frac{\sum_{i=1}^{n}(x_i^+)^2}{n\sum_{i=1}^{n}(x_i^+-\bar{x}_c)^2}\right\}^{1/2}\hat{\sigma}_c \qquad (2.3.3)$$

and

$$\hat{\beta}_{1c} \pm \frac{t(n-2,1-\frac{\alpha}{2})\hat{\sigma}_c}{\left[\sum_{i=1}^{n}(x_i^+-\bar{x}_c)^2\right]^{1/2}} \qquad (2.3.4)$$

as the confidence interval for β_o and β_1, respectively.

The properties of statistics associated with (2.1.1) that have been investigated are : (a) properties of the observed residuals (see (2.1.13)) (b) the biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ (c) the bias of $\hat{\sigma}_c^2$ (d) properties of the secondary data analyst's confidence intervals for β_o and β_1, i.e. (2.3.3) and (2.3.4), where x_i and $\bar{x}$ replace x_i^+ and $\bar{x}_c$, respectively, in (2.3.3) and (2.3.4) if there are no missing values of X.
The most desirable way to ascertain the properties of (2.3.3) and (2.3.4) would be to determine whether

$$t_a = \frac{\hat{\beta}_{oc}-\beta_o}{\hat{\sigma}_c\left\{\frac{\sum_{i=1}^{n}(x_i^+)^2}{n\sum_{i=1}^{n}(x_i^+-\bar{x}_c)^2}\right\}^{1/2}} \qquad (2.3.5)$$

and

$$t_b = \frac{(\hat{\beta}_{1c}-\beta_1)\left[\sum_{i=1}^{n}(x_i^+-\bar{x}_c)^2\right]^{1/2}}{\hat{\sigma}_c} \qquad (2.3.6)$$

are well-approximated by t-distribution with (n-2) degrees of freedom.

Because the algebraic expressions for Var($\hat{\beta}_{oc}$) are very complicated (see, for example, Table 2.3.1 for the case where X has no missing values) we only present the results for E($\hat{\beta}_{oc}$) and Var($\hat{\beta}_{oc}$). We mainly discuss the properties of (2.3.4) via t_b.

Since it is difficult to consider (2.3.6) directly we proceed in stages: (a) If the bias of $\hat{\beta}_{1c}$ is large then the approximation will not be satisfactory; (b) Since Var($\hat{\beta}_{1c}$) is a constant multiple of σ^2 (perhaps under some assumptions), a large bias for $\hat{\sigma}_c^2$ provides strong evidence of a poor approximation. If the biases in $\hat{\beta}_{1c}$ and $\hat{\sigma}_c^2$ are not large, then one should investigate

$$Q^2 = \frac{E\left[\dfrac{\hat{\sigma}_c^2}{\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)^2}\right]}{Var(\hat{\beta}_{1c})}. \qquad (2.3.7)$$

However, we have sometimes obtained more useful results by considering

$$R^2 = \frac{E\left[\dfrac{\hat{\sigma}_c^2}{\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)^2}\right]}{Var(\hat{\beta}_{1r})} \qquad (2.3.8)$$

where $Var(\hat{\beta}_{1r}) = \dfrac{\sigma^2}{\sum_{i=1}^{r}(x_{ri} - \bar{x}_r)^2}$ (see (2.1.11)). Since for all analytical situations of interest we can show that Var($\hat{\beta}_{1c}$) $\geq$ Var($\hat{\beta}_{1r}$), if R^2 is small then Q^2 will be small and the approximation will be unsatisfactory.

It should be noted that even if the biases of $\hat{\beta}_{1c}$ and $\hat{\sigma}_c^2$ are small and $Q^2 \cong 1$, there is no guarantee that t_b will be well approximated by a t distribution. Unfortunately, even in simple cases it is very difficult to find the exact distribution of $\hat{\sigma}_c^2$ and to investigate the dependence between $\hat{\beta}_{1c}$ and $\hat{\sigma}_c^2$.

The first evaluation of the alternative imputation methods by considering the properties of the observed residuals, $\{e_{ic}\}$, and investigating $\text{cov}(e_{ric}, e_{rlc})$, $\text{cov}(e_{ric}, e_{mjc})$, $\text{cov}(e_{mlc}, e_{mkc})$, etc., were unsuccessful.

In the remainder of section 2.3, we mainly discuss the results for the biases of $\hat{\beta}_{1c}$ and $\hat{\sigma}_c^2$, the values of the variance of $\hat{\beta}_{1c}$, Q^2 and R^2. To permit investigators to compare the properties of the various imputation methods for specific populations, the formulas for the biases and variances are given in their most general forms. Section 2.3.1 considers the case where X has no missing values. Section 2.3.2 considers the case where both X and y have missing values. Comparisons of the different imputation methods are made in Section 2.3.3. The results in Section 2.3 are summarized in Section 2.3.4.

2.3.1 X has no Missing Values

Considering each of the seven imputation methods listed in Section 2.2, we present the biases of $\hat{\beta}_{oc}$, $\hat{\beta}_{1c}$, and $\hat{\sigma}_c^2$, and the variances of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$. Expressions for Q^2 and R^2 are also given. See formulas (2.1.7), (2.1.8), (2.1.9), (2.3.7) and (2.3.8) for definitions. The expectations of $\hat{\beta}_{oc}$, $\hat{\beta}_{1c}$, $\hat{\sigma}_c^2$, $\dfrac{\hat{\sigma}_c^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$, Var($\hat{\beta}_{oc}$) and Var($\hat{\beta}_{1c}$) are taken over the model (1.1) but conditional on the observed X values, $\{x_i : i=1,2,...,n\}$. Since

$$\sum_{i=1}^{n}(y_i^+ - \hat{y}_{ic}^+)\hat{y}_{ic}^+ = 0,\ E\left(\hat{\sigma}_c^2 \middle| \{x_{i=1}^n\}\right) = E\left(\frac{\sum_{i=1}^{n}(y_i^+ - \hat{y}_{ic}^+)y_i^+}{n-2} \middle| \{x_{i=1}^n\} \right).$$

2.3.1-1 Mean Overall Imputation Methods (MO)

The missing Y values are imputed by the overall respondent mean, i.e. $y_{mj}^* = \bar{y}_r$. The conditional expectations of $\hat{\beta}_{oc}$, $\hat{\beta}_{1c}$, and $\hat{\sigma}_c^2$ are given by, respectively,

$$\mathrm{E}(\hat{\beta}_{oc} \mid \{x_{i=1}^n\})_{MO} = \beta_o + \beta_1 \left\{ \bar{x}_r - \frac{\bar{x}\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\}, \tag{2.3.9}$$

$$\mathrm{E}(\hat{\beta}_{1c} \mid \{x_{i=1}^n\})_{MO} = \beta_1 \left\{ \frac{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\}, \tag{2.3.10}$$

$$\begin{aligned} \mathrm{E}(\hat{\sigma}_c^2 \mid \{x_{i=1}^n\})_{MO} = \frac{\sigma^2}{n-2}\{(\mathrm{r}\text{-}1) - \left\{ \bar{x}_r - \frac{\bar{x}\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\}\} \\ + \frac{\beta_1^2}{n-2}\{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2\ [1 - \left\{ \frac{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\}]\}. \end{aligned} \tag{2.3.11}$$

$\mathrm{Var}(\hat{\beta}_{1c}\,|\,\{x_{i=1}^{n}\})_{MO} = \sigma^2 \left\{ \dfrac{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2} \right\}$. The expression for $\mathrm{Var}(\hat{\beta}_{oc})_{MO}$ is given in Table 2.3.1.

To obtain further insight we make the simplifying assumption that $x_{mj} = \bar{x}_m = \bar{x}_r$ for j=r+1,…,n. Then,

$$\mathrm{E}(\hat{\beta}_{oc}\,|\,\{x_{i=1}^{n}\})_{MO} = \beta_o, \tag{2.3.12}$$

$$\mathrm{E}(\hat{\beta}_{1c}\,|\,\{x_{i=1}^{n}\})_{MO} = \beta_1, \tag{2.3.13}$$

and $$\mathrm{E}(\hat{\sigma}_c^2\,|\,\{x_{i=1}^{n}\})_{MO} = (\text{r-2})\ \sigma^2/(\text{n-2}). \tag{2.3.14}$$

Also, $\mathrm{Var}(\hat{\beta}_{1c}\,|\,\{x_{i=1}^{n}\})_{MO} = \sigma^2 / \sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2 = \mathrm{Var}((\hat{\beta}_{1r})$

and $\mathrm{Var}(\hat{\beta}_{oc}\,|\,\{x_{i=1}^{n}\})_{MO} = \sigma^2 \left\{ \dfrac{1}{r} + \dfrac{\bar{x}_r^2}{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2} \right\} = \mathrm{Var}(\hat{\beta}_{or})$.

We obtain unbiased estimators for β_o and β_1, the bias of $\hat{\sigma}_c^2$ is $-\dfrac{m\sigma^2}{n-2}$. Also, $Q^2 = R^2 \cong \dfrac{r}{n}$ because $\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2$. Although under this special assumption E($\hat{\beta}_{1c}$)= β_1, the large bias of $\hat{\sigma}_c^2$ and the result that $Q^2 \cong \dfrac{r}{n}$ imply that t_b in (2.3.6) will not be well-approximated by a t-distribution. Recall that we assume a nonnegligible rate of nonresponse.

2.3.1-2 Random Imputation (RI)

For this method the m imputed Y values, $\{y_{mj}^{*}: \text{j=r+1,…,n}\}$, will vary from imputation to imputation. Thus, to obtain the required expected values one must first take an expectation over repeated imputations and conditional on **all** the observed data, $(\{(x_{ri}, y_{ri}): i = 1,2,...,r\},\ \{x_{mj}: j = r+1,...,n\})$, and then taken an expectation over the model. After considerable algebraic manipulation it can be shown that the conditional expectations of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are exactly the same as (2.3.9) and (2.3.10), respectively, and the conditional expectation of $\hat{\sigma}_c^2$ is given by

$$E(\hat{\sigma}_c^2 \left| \{x_{i=1}^n\}\right)_{RI} = \frac{\sigma^2}{n-2}\left\{(n-2) - \frac{m(n+r-1)}{nr} - \frac{m(\bar{x}_m - \bar{x})(\bar{x}_r - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}$$

$$+\frac{\beta_1^2}{n-2}\left\{\sum_{i=1}^{r}(x_{ri} - \bar{x}_r)^2\left[\frac{n^2-m}{nr} - \frac{1}{r}\left(1 - \frac{\sum_{i=1}^{r}(x_{ri}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\right] - \frac{\left[\sum_{i=1}^{r}(x_{ri}-\bar{x})^2\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right\}. \qquad (2.3.15)$$

Also,

$$Var(\hat{\beta}_{1c} \left| \{x_{i=1}^n\}\right)_{RI} = \frac{\sigma^2\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2 + 2m(\bar{x}_m-\bar{x})(\bar{x}_r-\bar{x}) + \frac{\beta_1^2 s_{rx}^2}{\sigma^2}\sum_{j=r+1}^{n}(x_{mj}-\bar{x})^2\right\}}{\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}^2}.$$

The expectation for $Var(\hat{\beta}_{oc} \left| \{x_{i=1}^n\}\right)_{RI}$ is given in Table 2.3.1.

Assuming that $x_{mj} = \bar{x}_m = \bar{x}_r$ for j=r+1,…,n, we obtain $\mathrm{E}(\hat{\beta}_{oc} \left| \{x_{i=1}^n\}\right)_{RI} = \beta_o$,

$\mathrm{E}(\hat{\beta}_{1c} \left| \{x_{i=1}^n\}\right)_{RI} = \beta_1$, and

$$E(\hat{\sigma}_c^2 \left| \{x_{i=1}^n\}\right)_{RI} = \frac{\sigma^2\left\{(n-2) - \frac{m(n+r-1)}{nr} + \frac{\beta_1^2 m(n-1)s_{rx}^2}{n\sigma^2}\right\}}{n-2}, \qquad (2.3.16)$$

$$Var(\hat{\beta}_{1c} \left| \{x_{i=1}^n\}\right)_{RI} = Var(\hat{\beta}_{1r}) = \frac{\sigma^2}{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}.$$

$$Var(\hat{\beta}_{oc} \left| \{x_{i=1}^n\}\right)_{RI} = \sigma^2\left\{\left(\frac{1}{n} + \frac{2m}{n^2}\right) + \frac{\bar{x}^2}{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2} + \frac{m}{n^2}\frac{\beta_1^2 s_{rx}^2}{\sigma^2}\right\} \geq Var(\hat{\beta}_{1r}).$$

If r is large, $\frac{m(n+r-1)}{nr(n-2)} \cong \left(1 - \left(\frac{r}{n}\right)^2\right)\frac{1}{r}$ is negligible, and

$$E(\hat{\sigma}_c^2 \left| \{x_{i=1}^n\}\right)_{RI} \cong \sigma^2\left\{1 + \frac{\beta_1^2}{\sigma^2}s_{rx}^2\left(1 - \frac{r}{n}\right)\right\}. \qquad (2.3.17)$$

Therefore, the bias of $\hat{\sigma}_c^2$ is $\frac{\beta_1^2}{\sigma^2}s_{rx}^2\left(1-\frac{r}{n}\right)$ which will be small if the r observed X values,

$\{x_{ri} : i = 1,...,r\}$, are close to each other. Also, $Q^2 = R^2 \cong \{1+\frac{\beta_1^2}{\sigma^2} s_{rx}^2 (1-\frac{r}{n})\}$.

Under the special assumption E($\hat{\beta}_{1c}$)= β_1, the bias of $\hat{\sigma}_c^2$ and the result that $Q^2 \cong \{1+\frac{\beta_1^2}{\sigma^2} s_{rx}^2 (1-\frac{r}{n})\}$ imply that t_b in (2.3.6) may not well-approximated by a t-distribution.

2.3.1-3 Mean Imputation Within Cells (MC)

For this method $y_{mhj}^* = \bar{y}_{rh}$, the sample mean of the respondents in imputation cell h. The expectations and variances of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ over the model (2.1.1), but conditional on the observed values of X, can be shown to be given by

$$\mathrm{E}(\hat{\beta}_{oc} \left| \{x_{i=1}^n\}\right)_{MC} = \beta_o + \beta_1 \left\{ \sum_{h=1}^{L} \frac{n_h}{n} \bar{x}_{rh} - \frac{\bar{x}\left[\sum_{i=1}^{r}(x_{ri}-\bar{x})x_{ri} + \sum_{h=1}^{L} m_h \bar{x}_{rh}(\bar{x}_{mh}-\bar{x})\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\}, \quad (2.3.18)$$

$$\mathrm{E}(\hat{\beta}_{1c} \left| \{x_{i=1}^n\}\right)_{MC} = \frac{\beta_1 \left\{ \sum_{i=1}^{r}(x_{ri}-\bar{x})x_{ri} + \sum_{h=1}^{L} m_h \bar{x}_{rh}(\bar{x}_{mh}-\bar{x}) \right\}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \quad (2.3.19)$$

Also,

$$\mathrm{Var}(\hat{\beta}_{1c} \left| \{x_{i=1}^n\}\right)_{MC} = \frac{\sigma^2 \left\{ \sum_{h=1}^{L} \sum_{i=1}^{r_h} \left[(x_{rhi}-\bar{x}) + \frac{m_h}{r_h}(\bar{x}_{mh}-\bar{x}) \right]^2 \right\}}{\left\{ \sum_{i=1}^{n}(x_i-\bar{x})^2 \right\}^2}. \quad (2.3.20)$$

The expression for $\mathrm{Var}(\hat{\beta}_{oc})_{MC}$ is given in Table 2.3.1. and

$$\mathrm{E}(\hat{\sigma}_c^2 \left| \{x_{i=1}^n\}\right)_{MC} = \frac{\sigma^2}{n-2} \left\{ (r-1) + \sum_{h=1}^{L}(1-\frac{n_h}{n})\frac{m_h}{r_h} - \frac{\sum_{h=1}^{L}\sum_{i=1}^{r_h}\left[(x_{rhi}-\bar{x}) + \frac{m_h}{r_h}(\bar{x}_{mh}-\bar{x}) \right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\}$$

$$+ \frac{\beta_1^2}{n-2} \left\{ \sum_{i=1}^{r} x_{ri}^2 + \sum_{h=1}^{L} m_h \bar{x}_{rh}^2 - \frac{(\sum_{h=1}^{L} n_h \bar{x}_{rh})^2}{n} - \frac{\left[\sum_{i=1}^{r} x_{ri}(x_{ri}-\bar{x}) + \sum_{h=1}^{L} m_h \bar{x}_{rh}(\bar{x}_{mh}-\bar{x}) \right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\}. \quad (2.3.21)$$

To obtain further insight we make the simplifying assumption that all of the X values within the same cell are equal (i.e., $x_{hi} = x_h$). This approximates the (realistic) situation where there is little variation in X within each of the imputation cells. Then

$$\mathrm{E}(\hat{\beta}_{oc} \left| \{x_{i=1}^n\}\right)_{MC} = \beta_o\,, \tag{2.3.22}$$

$$\mathrm{E}(\hat{\beta}_{1c} \left| \{x_{i=1}^n\}\right)_{MC} = \beta_1\,, \tag{2.3.23}$$

and

$$\mathrm{E}(\hat{\sigma}_c^2 \left| \{x_{i=1}^n\}\right)_{MC} = \frac{\sigma^2}{n-2}\left\{(r-2) + \sum_{h=1}^{L}(1-\frac{n_h}{n})\frac{m_h}{r_h} - \frac{\sum_{h=1}^{L} n_h \frac{m_h}{r_h}(x_h-\bar{x})^2}{\sum_{h=1}^{L} n_h (x_h-\bar{x})^2}\right\}. \tag{2.3.24}$$

$$\mathrm{Var}(\hat{\beta}_{1c} \left| \{x_{i=1}^n\}\right)_{MC} = \frac{\sigma^2\left\{\sum_{h=1}^{L}\frac{n_h^2}{r_h}(x_h-\bar{x})^2\right\}}{\left\{\sum_{h=1}^{L} n_h (x_h-\bar{x})^2\right\}^2} \geq Var(\hat{\beta}_{1r}) \quad \text{if } \frac{r}{n} = \frac{r_h}{n_h} \text{ for h=1,2,…,L}$$

$$\mathrm{Var}(\hat{\beta}_{oc} \left| \{x_{i=1}^n\}\right)_{MC} = \sigma^2\left\{\sum_{h=1}^{L}\frac{n_h^2}{r_h}\left[\frac{1}{n} - \frac{\bar{x}(x_h-\bar{x})}{\sum_{h=1}^{L} n_h (x_h-\bar{x})^2}\right]^2\right\}.$$

There is no easy comparison between $Var(\hat{\beta}_{oc})$ and $\mathrm{var}(\hat{\beta}_{or})$. Assuming $r_h \geq \frac{n_h}{2}$, n is large and L is small relative to n, the last two terms in (2.3.24) will be negligible. Then

$$\mathrm{E}(\hat{\sigma}_c^2 \left| \{x_{i=1}^n\}\right)_{MC} \cong \frac{r\sigma^2}{n}\,. \tag{2.3.25}$$

Using (2.3.25), it can be shown that

$$Q^2 \cong \frac{r}{n\left\{1+\left[\frac{\sum_{h=1}^{L} n_h \frac{m_h}{r_h}(x_h-\bar{x})^2}{\sum_{h=1}^{L} n_h (x_h-\bar{x})^2}\right]\right\}} \geq \frac{r}{2n}\text{, a and } R^2 \cong \frac{r\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

If we make the additional assumption that $\{x_{ri} : i = 1,...,r\}$ is a random sample from $\{x_i : i = 1,2,...,n\}$ then $E(\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2 \left| \{x_{i=1}^n\}\right) \cong \frac{r}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$, and $R^2 \cong \left(\frac{r}{n}\right)^2$ (2.3.26)

Where the approximation in (2.3.26) comes from using (2.3.25), and replacing R^2 by its expected value. Given the results in (2.3.25), (2.3.26), and $\frac{r}{2n} \le Q^2 \le \left(\frac{r}{n}\right)^2$, it is to be expected that t_b in (2.3.6) will not be well-approximated by a t-distribution.

2.3.1-4 Random Imputation Within Cells (RC)

This method is a generation of RI; i.e., RC is RI applied independently within each of the imputation cells. After a considerable amount of algebraic manipulation it can be shown that the conditional expectations of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are the same as (2.3.18) and (2.3.19), respectively. The conditional expectation of $\hat{\sigma}_c^2$ is

$$E(\hat{\sigma}_c^2 \left| \{x_{i=1}^n\}\right)_{RC} =$$

$$\frac{\sigma^2}{n-2}\left\{(n-1)-\frac{m}{n}-\sum_{h=1}^{L}(\frac{n_h-1}{n})\frac{m_h}{r_h}-\frac{\sum_{h=1}^{L}\sum_{J=r_h+1}^{n_h}(1-\frac{1}{r_h})(x_{mhj}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}-\frac{\sum_{h=1}^{L}\sum_{i=1}^{r_h}\left[(x_{rhi}-\bar{x})+\frac{m_h}{r_h}(\bar{x}_{mh}-\bar{x})\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right\}$$

$$+\frac{\beta_1^2}{n-2}\left\{\sum_{i=1}^{r}x_{ri}^2+\sum_{h=1}^{L}m_h\bar{x}_{rh}^2-\frac{(\sum_{h=1}^{L}n_h\bar{x}_{rh})^2}{n}+\sum_{h=1}^{L}\sum_{j=r_h+1}^{n_h}\left[1-\frac{1}{n}-\frac{(x_{mhj}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]s_{rxh}^2\right.$$

$$\left.-\frac{\left[\sum_{i=1}^{r}(x_{ri}-\bar{x})x_{ri}+\sum_{h=1}^{L}m_h\bar{x}_{rh}(\bar{x}_{mh}-\bar{x})\right]^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right\} \qquad (2.3.27)$$

$$Var(\hat{\beta}_{1c} \left| \{x_{i=1}^n\}\right)_{RC} =$$

$$\sigma^2\frac{\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2+2\sum_{h=1}^{L}m_h(\bar{x}_{mh}-\bar{x})(\bar{x}_{rh}-\bar{x})+\frac{\beta_1^2\sum_{h=1}^{L}\sum_{j=r_h+1}^{n_h}(x_{mhj}-\bar{x})^2s_{rxh}^2}{\sigma^2}\right\}}{\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}^2} \qquad (2.3.28)$$

The expression for $Var(\hat{\beta}_{oc})_{RC}$ is given in Table 2.3.1. Under the assumption that $x_{hi} = x_h$, we obtain $E(\hat{\beta}_{oc} \left| \{x_{i=1}^n\}\right)_{RC} = \beta_o$, $E(\hat{\beta}_{1c} \left| \{x_{i=1}^n\}\right)_{RC} = \beta_1$, and

$$E(\hat{\sigma}_c^2 \mid \{x_{i=1}^n\})_{RC} = \sigma^2 \frac{\left\{(n-2) - \frac{m}{n} - \sum_{h=1}^{L}\left(\frac{n_h - 1}{n}\right)\frac{m_h}{r_h} - \frac{\sum_{h=1}^{L}(n_h + r_h - 1)\frac{m_h}{r_h}(x_h - \bar{x})^2}{\sum_{h=1}^{L} n_h (x_h - \bar{x})^2}\right\}}{n-2} \tag{2.3.29}$$

$$Var(\hat{\beta}_{1c} \mid \{x_{i=1}^n\})_{RC} = \sigma^2 \frac{\left\{\sum_{h=1}^{L} n_h (x_h - \bar{x})^2 + 2\sum_{h=1}^{L} m_h (x_h - \bar{x})^2\right\}}{\left\{\sum_{h=1}^{L} n_h (x_h - \bar{x})^2\right\}^2} \geq Var(\hat{\beta}_{1r}) \text{ if } \frac{r_h}{n_h} = \frac{r}{n} \geq \frac{1}{2} \text{ for}$$

h=1,2,…,L.

$Var(\hat{\beta}_{oc} \mid \{x_{i=1}^n\})_{RC} =$

$$\sigma^2 \left\{\frac{n+2m}{n^2} + \frac{\bar{x}^2}{\sum_{h=1}^{L} n_h (x_h - \bar{x})^2}\left[1 + \frac{2\sum_{h=1}^{L} m_h (x_h - \bar{x})^2}{\sum_{h=1}^{L} n_h (x_h - \bar{x})^2}\right] - \frac{4\bar{x}}{n}\frac{\sum_{h=1}^{L} m_h (x_h - \bar{x})}{\sum_{h=1}^{L} n_h (x_h - \bar{x})^2}\right\}.$$ There is no

easy comparison between $Var(\hat{\beta}_{oc})$ and $Var(\hat{\beta}_{or})$. Assuming $r_h \geq \frac{n_h}{2}$, n is large and L is small relative to n, then

$$E(\hat{\sigma}_c^2 \mid \{x_{i=1}^n\})_{RC} \cong \sigma^2 \tag{2.3.30}$$

Using (2.3.30), it can be shown that

$$Q^2 \cong \frac{1}{\left\{1 + \frac{2\sum_{h=1}^{L} m_h (x_h - \bar{x})^2}{\sum_{h=1}^{L} n_h (x_h - \bar{x})^2}\right\}} \geq \frac{1}{2} \text{ and } R^2 \cong \frac{\sum_{i=1}^{r} (x_{ri} - \bar{x}_r)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

If we assume that that $\{x_{ri} : i = 1,...,r\}$ is a random sample from $\{x_i : i = 1,2,...,n\}$ then $E(\sum_{i=1}^{r} (x_{ri} - \bar{x}_r)^2 \mid \{x_{i=1}^n\}) \cong \frac{r}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$, and $R^2 \cong \frac{r}{n}$. The result that $\frac{1}{2} \leq Q^2 \leq \frac{r}{n}$ is not reassuring about the quality of the approximation of t_b in (2.3.6) by a t-distribution.

2.3.1-5 Simple Regression Prediction Imputation (RG)

In the RG method, the simple regression of Y on X is used to impute the missing Y values, i.e., $y_{mj}^* = \hat{\beta}_{or} + \hat{\beta}_{1r}\, x_{mj}$. The resultant conditional expectations of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are given by, respectively,

$$E(\hat{\beta}_{oc}\left|\{x_{i=1}^{n}\}\right)_{RG} = \beta_o \text{ and } \mathrm{E}(\hat{\beta}_{1c}\left|\{x_{i=1}^{n}\}\right)_{RG} = \beta_1. \quad (2.3.31)$$

The conditional expectation of $\hat{\sigma}_c^2$ is given by

$$E(\hat{\sigma}_c^2\left|\{x_{i=1}^{n}\}\right)_{RG} = \frac{r-2}{n-2}\sigma^2 \quad (2.3.32)$$

The bias for the estimator of σ^2 is $-\frac{m}{n-2}\sigma^2$. These are the same results obtained by the MO method when we assume that $x_{mj} = \bar{x}_m = \bar{x}_r$ (see Section 2.3.1-1).

$\mathrm{Var}(\hat{\beta}_{1c}\left|\{x_{i=1}^{n}\}\right)_{RG}$ =

$$\frac{\sigma^2}{\left\{\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}^2}\left\{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2 + 2m(\bar{x}_m-\bar{x})(\bar{x}_r-\bar{x}) + \frac{\sum_{j=r+1}^{n}(x_{mj}-\bar{x})^2}{r} + \frac{\sum_{j=r+1}^{n}(x_{mj}-\bar{x})^2(x_{mj}-\bar{x}_r)^2}{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}\right\}$$

The expression for $Var(\hat{\beta}_{oc})_{RG}$ is given in table 2.3.1. If we make the simplifyinf assumption that $x_{mj} = \bar{x}_m = \bar{x}_r$ for j=r=1,…,n, $\mathrm{Var}(\hat{\beta}_{1c}\left|\{x_{i=1}^{n}\}\right)_{RG}$ =Var($\hat{\beta}_{1r}$) and

$Var(\hat{\beta}_{oc})_{RG} = \sigma^2\left\{\frac{1}{r} + \frac{\bar{x}_r^2}{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}\right\}$ =Var($\hat{\beta}_{or}$). Under the assumption and using (2.3.32), we obtain $R^2 = Q^2 \cong \frac{r}{n}\frac{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$. If in addition, it is assumed that $\{x_{ri} : i = 1,...,r\}$ is a random sample from $\{x_i : i = 1,2,...,n\}$ then $R^2 = Q^2 \cong \left(\frac{r}{n}\right)^2$. From (2.3.32) and the value of Q^2, t_b in (2.3.6) will not be well-approximated by a t-distribution (under our assumption of a nonnegligible rate of nonresponse.)

2.3.1-6 Random Regression Imputation (RRS)

For this method the respondents' residuals, $\{e_{ri} = y_{ri} - \hat{\beta}_{or} - \hat{\beta}_{1r}\, x_{ri} : \mathrm{i=1,\ldots,r}\}$, are randomly allocated to the nonrespondents and added to $\{y_{mj}^{*} : \mathrm{j=r+1,\ldots,n}\}$ as defined in Section 2.3.1-5. Then, after considerable algebraic manipulation, it can be shown that

$$\mathrm{E}(\hat{\beta}_{oc}\left|\{x_{i=1}^{n}\}\right)_{RRS} = \beta_o \text{ and } \mathrm{E}(\hat{\beta}_{1c}\left|\{x_{i=1}^{n}\}\right)_{RRS} = \beta_1 \quad (2.3.33)$$

The conditional expectation of $\hat{\sigma}_c^2$ is

$$\mathrm{E}(\hat{\sigma}_c^2 \mid \{x_{i=1}^n\})_{RRS} =$$

$$\frac{\sigma^2}{n-2}\left\{(n-2)-\frac{\left[m(m+1)+\frac{m(n-1)}{r}\right]}{n}-\frac{(r-1)m^2(\bar{x}_m-\bar{x}_r)^2}{r\sum_{i=1}^{n}(x_i-\bar{x})^2}+\frac{\sum_{j=r+1}^{n}(x_{mj}-\bar{x})^2}{r\sum_{i=1}^{n}(x_i-\bar{x})^2}\right\} \quad (2.3.34)$$

while $\mathrm{Var}(\hat{\beta}_{1c} \mid \{x_{i=1}^n\})_{RRS} = \sigma^2\left\{\frac{1}{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}+\frac{(r-2)\sum_{j=r+1}^{n}(x_{mj}-\bar{x})^2}{r\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}\right\} \geq \mathrm{Var}(\hat{\beta}_{1r})$.

The expression for $\mathrm{Var}(\hat{\beta}_{oc})_{RRS}$ is given in Table 2.3.1. Assuming $x_{mj} = \bar{x}_m = \bar{x}_r$ for j=r=1,…,n,

$$\mathrm{E}(\hat{\sigma}_c^2 \mid \{x_{i=1}^n\})_{RRS} = \frac{\sigma^2}{n-2}\left\{(n-2)-\frac{m(m+1)+\frac{m(n-1)}{r}}{n}\right\},$$

$$\mathrm{Var}(\hat{\beta}_{1c} \mid \{x_{i=1}^n\})_{RRS} = \frac{\sigma^2}{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2} = \mathrm{Var}(\hat{\beta}_{1r}), \text{ and}$$

$$\mathrm{Var}(\hat{\beta}_{oc})_{RRS} = \sigma^2\left\{\frac{n+2m-\frac{m}{r}}{n^2}+\frac{\bar{x}_r^2}{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2}\right\} \geq \mathrm{Var}(\hat{\beta}_{or}) \text{ if } r > \frac{n}{2}.$$ If r is large,

$$\mathrm{E}(\hat{\sigma}_c^2 \mid \{x_{i=1}^n\})_{RRS} \cong \sigma^2\left\{1-(1-\frac{r}{n})^2\right\} \quad (2.3.35)$$

and then $R^2 = Q^2 \cong \left\{1-(1-\frac{r}{n})^2\right\}$. From (2.3.35) and the value of Q^2, t_b in (2.3.6) may be well-approximated by a t-distribution.

2.3.1-7 <u>Random Regression Imputation (RRN)</u>

For the RRN scheme we impute the missing Y values by using $y_{mj}^* = \tilde{y}_{mj} + e_{mj}$ where $\tilde{y}_{mj} = \hat{\beta}_{or} + \hat{\beta}_{1r}\, x_{mj}$ is the simple regression prediction and e_{mj} is chosen at random from a distribution with zero mean and variance equal to the residual variance of the regression using the respondents' data. The conditional expectations of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ both are unbiased, $\mathrm{E}(\hat{\beta}_{oc} \mid \{x_{i=1}^n\})_{RRN} = \beta_o$ and $\mathrm{E}(\hat{\beta}_{1c} \mid \{x_{i=1}^n\})_{RRN} = \beta_1$. The conditional expectation of $\hat{\sigma}_c^2$ is given by

$$\mathrm{E}(\hat{\sigma}_c^2 \left| \{x_{i=1}^n\}\right)_{RRN} = \sigma^2 - \frac{\sigma^2}{n-2}\left\{\frac{m}{n} + \frac{\sum_{j=r+1}^{n}(x_{mj}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right\} \qquad (2.3.36)$$

while

$$\mathrm{Var}(\hat{\beta}_{1c} \left| \{x_{i=1}^n\}\right)_{RRN} = \sigma^2 \left\{\frac{1}{\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2} + \frac{\sum_{j=r+1}^{n}(x_{mj}-\bar{x})^2}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2}\right\} \geq \mathrm{Var}(\hat{\beta}_{1r}).$$

The expression for $\mathrm{Var}(\hat{\beta}_{oc})_{RRN}$ is given in Table 2.3.1. Since $0 < \frac{\sum_{j=r+1}^{n}(x_{mj}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} < 1$, if n is large then $\mathrm{E}(\hat{\sigma}_c^2 \left| \{x_{i=1}^n\}\right)_{RRN} \cong \sigma^2$. If, in addition, it is assumed that $\{x_{ri} : i = 1,...,r\}$ is a random sample from $\{x_i : i = 1,2,...,n\}$ then $E(\sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2 \left| \{x_{i=1}^n\}\right) \cong \frac{r}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$, $Q^2 \cong \frac{r}{n}\frac{1}{\left\{1+\frac{r}{n}\left(1-\frac{r}{n}\right)\right\}}$, and $R^2 \cong \frac{r}{n}$.

The result that $Q^2 \leq R^2 \cong \frac{r}{n}$ is not reassuring about the quality of the approximation of t_b in (2.3.6) by a t-distribution. However, if we assume that $x_{mj} = \bar{x}_m = \bar{x}_r$ and n is large then $\mathrm{E}(\hat{\sigma}_c^2 \left| \{x_{i=1}^n\}\right)_{RRN} \cong \sigma^2$ and $\mathrm{Var}(\hat{\beta}_{1c} \left| \{x_{i=1}^n\}\right)_{RRN} = \mathrm{Var}(\hat{\beta}_{1r})$. Since $\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{r}(x_{ri}-\bar{x}_r)^2$, $Q^2 = R^2 \cong 1$. Therefore, t_b in (2.3.6) may be well approximated by a t-distribution.

2.3.2 X Has Missing values

In this case, both X and Y have missing values. Considering each of the seven imputation methods listed in Section 2.2, we present the biases of $\hat{\beta}_{oc}$, $\hat{\beta}_{1c}$ and $\hat{\sigma}_c^2$, and the variances of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$. Expressions for Q^2 and R^2 are also given. See formulas (2.1.7), (2.1.8), (2.1.9), (2.3.7) and (2.3.8) for definitions, but note that for some methods Q^2 and R^2 are approximated by

$$Q^2 \cong \frac{E\left(\hat{\sigma}_c^2 \left| \{x_{r_i=1}^r\}\right.\right)}{E\left[\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)^2 \left| \{x_{r_i=1}^r\}\right.\right] Var(\hat{\beta}_{1c})}\,, \tag{2.3.37}$$

$$R^2 \cong \frac{\sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2 E\left(\hat{\sigma}_c^2 \left| \{x_{r_i=1}^r\}\right.\right)}{\sigma^2 E\left[\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)^2 \left| \{x_{r_i=1}^r\}\right.\right]}, \tag{2.3.38}$$

respectively. The expectations of $\hat{\beta}_{oc}$, $\hat{\beta}_{1c}$, $\hat{\sigma}_c^2$, $\frac{\hat{\sigma}_c^2}{\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)^2}$, $Var(\hat{\beta}_{oc})$, and $Var(\hat{\beta}_{1c})$ are taken over the model (2.1.1) but conditional on the observed X values, $\{x_{r_i} : i = 1,...,r\}$.

2.3.2-1 Mean Overall Imputation Method (MO)

For this method $x_{mj}^* = \bar{x}_r$ and $y_{mj}^* = \bar{y}_r$. The conditional expectations of $\hat{\beta}_{oc}$, $\hat{\beta}_{1c}$, and $\hat{\sigma}_c^2$ can be shown to be given by

$$\mathrm{E}(\hat{\beta}_{oc} \left| \{x_{r_i=1}^r\}\right.)_{MO} = \beta_o\,,\ \mathrm{E}(\hat{\beta}_{1c} \left| \{x_{r_i=1}^r\}\right.)_{MO} = \beta_1, \tag{2.3.39}$$

and

$$\mathrm{E}(\hat{\sigma}_c^2 \left| \{x_{r_i=1}^r\}\right.)_{MO} = \frac{(r-2)\sigma^2}{(n-2)}\,. \tag{2.3.40}$$

The biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are zero and the bias of $\hat{\sigma}_c^2$ is $-\frac{m\sigma^2}{n-2}$. The variances of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are $\mathrm{Var}(\hat{\beta}_{oc} \left| \{x_{r_i=1}^r\}\right.)_{MO} = \sigma^2 \left\{\frac{1}{r} + \frac{\bar{x}_r^2}{\sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2}\right\} = \mathrm{Var}(\hat{\beta}_{or})$, and

$$\mathrm{Var}(\hat{\beta}_{1c} \left| \{x_{r_i=1}^r\}\right.)_{MO} = \sigma^2 / \sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2 = \mathrm{Var}(\hat{\beta}_{1r}).$$

Since $\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)^2 = \sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2$, if r is large then $Q^2 = R^2 \cong$ r/n. Although for MO, $\mathrm{E}(\hat{\beta}_{1c}) = \beta_1$ the large bias of $\hat{\sigma}_c^2$ and the value of Q^2 indicate that t_b in (2.3.6) will not be well approximated by a t-distribution.

2.3.2-2 Random Imputation (RI)

The values imputed for $\{(x_{mj}, y_{mj})$: j=r+1,…,n$\}$, $\{(x^*_{mj}, y^*_{mj})$: j=r+1,…,n$\}$, are obtained by selecting a random sample (with replacement) from $\{(x_{ri}, y_{ri}) : i = 1,...,r\}$.
To obtain the required expected values, one first takes an expectation over the model (2.1.1), but
Conditional on the r observed X values, $\{x_{ri} : i = 1,...,r\}$, and $\{(x^*_{mj}, y^*_{mj})$: j=r+1,…,n$\}$.
Then the expectation is taken over the repeated imputations. Doing so, it can be shown that

$$\mathrm{E}(\hat{\beta}_{oc} \left| \{x^r_{r_i=1}\}\right)_{RI} = \beta_o, \ \mathrm{E}(\hat{\beta}_{1c} \left| \{x^r_{r_i=1}\}\right)_{RI} = \beta_1. \tag{2.3.41}$$

Using a first order Taylor series approximations it can be shown, after considerable algebraic manipulation, that

$$\mathrm{E}(\hat{\sigma}^2_c \left| \{x^r_{r_i=1}\}\right)_{RI} \cong \sigma^2 \tag{2.3.42}$$

Similarly, it can be shown that

$$\mathrm{Var}(\hat{\beta}_{1c} \left| \{x^r_{r_i=1}\}\right)_{RI} \cong \sigma^2 / \sum_{i=1}^{r} (x_{r_i} - \bar{x}_r)^2 = \mathrm{Var}(\hat{\beta}_{1r}),$$

$$\mathrm{Var}(\hat{\beta}_{oc} \left| \{x^r_{r_i=1}\}\right)_{RI} \cong \sigma^2 \left\{ \frac{nr + 2mr + m}{n^2 r} + \frac{\bar{x}_r^2}{\sum_{i=1}^{r} (x_{r_i} - \bar{x}_r)^2} \right\} \geq \mathrm{Var}(\hat{\beta}_{or}) \text{ if } \mathrm{r} \geq \frac{n}{2}.$$

Since

$$E\left(\sum_{i=1}^{n} (x_i^+ - \bar{x}_c)^2 \left| \{x^r_{r_{i=1}}\} \right. \right) = \left(n - \frac{m}{n} \right) \sum_{i=1}^{r} (x_{r_i} - \bar{x}_r)^2 \text{ /r}, \ Q^2 = R^2 \cong \mathrm{r} / \left(n - \frac{m}{n} \right).$$

From (2.3.42) and the value of $Q^2 \geq$ r/n, t_b in (2.3.6) may be well approximated by a t-distribution.

2.3.2-3 Mean Imputation Within cells (MC)

For this method $(x^*_{mhj}, y^*_{mhj}) = (\bar{x}_{rh}, \bar{y}_{rh})$ for j= $r_h + 1,...,n_h$. The expectations of $\hat{\beta}_{oc}$, $\hat{\beta}_{1c}$, and $\hat{\sigma}^2_c$ over the model (2.1.1) but conditional on the r observed X values are given by

$$\mathrm{E}(\hat{\beta}_{oc} \left| \{x^r_{r_{i=1}}\} \right)_{MC} = \beta_o, \quad \mathrm{E}(\hat{\beta}_{1c} \left| \{x^r_{r_{i=1}}\} \right)_{MC} = \beta_1, \tag{2.3.43}$$

and

$$\mathrm{E}(\hat{\sigma}^2_c \left| \{x^r_{r_{i=1}}\} \right)_{MC} = \frac{\sigma^2}{n-2} \left\{ (r-2) + \sum_{h=1}^{L} (1 - \frac{n_h}{n}) \frac{m_h}{r_h} - \frac{\sum_{h=1}^{L} n_h \frac{m_h}{r_h} (\bar{x}_{rh} - \bar{x}_c)^2}{\sum_{i=1}^{r} (x_{r_i} - \bar{x}_c)^2 + \sum_{h=1}^{L} m_h (\bar{x}_{rh} - \bar{x}_c)^2} \right\}. \tag{2.3.44}$$

The variances of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are

$$\mathrm{Var}(\hat{\beta}_{1c}\,|\,\{x^{r}_{r_{i=1}}\})_{MC}=\frac{\sigma^2\left\{\sum_{h=1}^{L}\sum_{i=1}^{r_h}\left[(x_{r_{hi}}-\bar{x}_c)+\frac{m_h}{r_h}(\bar{x}_{rh}-\bar{x}_c)\right]^2\right\}}{\left\{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2+\sum_{h=1}^{L}m_h(\bar{x}_{rh}-\bar{x}_c)^2\right\}^2},$$

$$\mathrm{Var}(\hat{\beta}_{oc}\,|\,\{x^{r}_{r_{i=1}}\})_{MC}=\sigma^2\times$$

$$\left\{\sum_{h=1}^{L}(\frac{n_h}{n})^2\frac{1}{r_h}+\frac{\bar{x}_c^2\sum_{h=1}^{L}\sum_{i=1}^{r_h}[(x_{rh_i}-\bar{x}_c)+\frac{m_h}{r_h}(\bar{x}_{rh}-\bar{x}_c)]^2}{[\sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2+\sum_{h=1}^{L}m_h(\bar{x}_{rh}-\bar{x}_c)^2]^2}-\frac{2\bar{x}_c\sum_{h=1}^{L}\frac{n_h^2}{nr_h}(\bar{x}_{rh}-\bar{x}_c)}{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2+\sum_{h=1}^{L}m_h(\bar{x}_{rh}-\bar{x}_c)^2}\right\}$$

Now, assume that $x_{rh_i}=x_{rh}\,(=\bar{x}_{rh})$ for $i=1,...,r_h$. Then

$$\mathrm{Var}(\hat{\beta}_{1c}\,|\,\{x^{r}_{r_{i=1}}\})_{MC}=\frac{\sigma^2\sum_{h=1}^{L}\frac{n_h^2}{r_h}(\bar{x}_{rh}-\bar{x}_c)^2}{\left[\sum_{h=1}^{L}n_h(\bar{x}_{rh}-\bar{x}_c)^2\right]^2}\geq\mathrm{Var}(\hat{\beta}_{1r})\text{ if }\frac{r_h}{n_h}=\frac{r}{n}\text{ for }h=1,...,L\text{ and}$$

$$\mathrm{Var}(\hat{\beta}_{oc}\,|\,\{x^{r}_{r_{i=1}}\})_{MC}=\sigma^2\left\{\sum_{h=1}^{L}\frac{n_h^2}{r_h}\left[\frac{1}{n}-\frac{\bar{x}_c(\bar{x}_{rh}-\bar{x}_c)}{\sum_{h=1}^{L}n_h(\bar{x}_{rh}-\bar{x}_c)^2}\right]^2\right\}.$$ There is no easy comparison

Between $\mathrm{Var}(\hat{\beta}_{oc})$ and $\mathrm{Var}(\hat{\beta}_{or})$. If , in addition, it is assumed that n is large and L is small relative to n and $r_h\geq\frac{n_h}{2}$, the last two terms in (2.3.44) are negligible. Then,

$$\mathrm{E}(\hat{\sigma}_c^2\,|\,\{x^{r}_{r_{i=1}}\})_{MC}\cong\sigma^2\frac{r}{n}. \qquad (2.3.45)$$

Under the same assumption $x_{rh_i}=x_{rh}$, $\sum_{i=1}^{n}(x_i^{+}-\bar{x}_c)^2=\sum_{h=1}^{L}n_h(\bar{x}_{rh}-\bar{x}_c)^2$. Using (2.3.45) and same assumptions, it can be shown that

$$Q^2\cong(\frac{r}{n})\frac{1}{1+\left[\frac{\sum_{h=1}^{L}n_h\frac{m_h}{r_h}(\bar{x}_{rh}-\bar{x}_c)^2}{\sum_{h=1}^{L}n_h(\bar{x}_{rh}-\bar{x}_c)^2}\right]}\geq\frac{r}{2n},\text{ and }R^2\cong\frac{r\sum_{h=1}^{L}r_h(\bar{x}_{rh}-\bar{x}_r)^2}{n\sum_{h=1}^{L}n_h(\bar{x}_{rh}-\bar{x}_c)^2}. \qquad (2.3.46)$$

Hence, $\frac{r}{n} \geq R^2 \geq Q^2 \geq \frac{r}{2n}$.

Although for MC, E($\hat{\beta}_{1c}$)= β_1 the large bias of $\hat{\sigma}_c^2$ and the upper bound on Q^2 indicate that t_b in (2.3.6) will not be well approximated by a t-distribution.

2.3.2-4 Random Imputation Within Cells (RC)

This method is a generalization of RI; i.e., RC is RI applied independently within each of the imputation cells. As noted in Section 2.3.2-2, two levels of expectation and first order of Taylor series approximations are needed to obtain most of the expected values presented below. After considerable algebraic manipulation, it can be shown that

$$\mathrm{E}(\hat{\beta}_{oc} \left| \{x_{r_{i=1}}^r\} \right.)_{RC} = \beta_o \,, \mathrm{E}(\hat{\beta}_{1c} \left| \{x_{r_{i=1}}^r\} \right.)_{RC} = \beta_1, \tag{2.3.47}$$

and

$$\mathrm{E}(\hat{\sigma}_c^2 \left| \{x_{r_{i=1}}^r\} \right.)_{RC} \cong \sigma^2 . \tag{2.3.48}$$

$$\mathrm{Var}(\hat{\beta}_{1c} \left| \{x_{r_{i=1}}^r\} \right.)_{RC} \cong \sigma^2 \left\{ \frac{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2 + \sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2 + 2\sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2}{\left[\sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2 + \sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2\right]^2} \right\},$$

$$\mathrm{Var}(\hat{\beta}_{oc} \left| \{x_{r_{i=1}}^r\} \right.)_{RC} \cong \sigma^2 \left(\frac{n+2m}{n^2}\right) +$$

$$\sigma^2 \left\{ \frac{\sum_{h=1}^{L} m_h s_{rxh}^2}{n^2} + \bar{x}_c^2 \right\} \left\{ \frac{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2 + \sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2 + 2\sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2}{\left[\sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2 + \sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2\right]^2} \right\}.$$ Also,

$$E\left(\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)^2 \left| \{x_{r_{i=1}}^r\} \right.\right) = \sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2 + \sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2 + \left(1-\frac{1}{n}\right)\sum_{h=1}^{L} m_h s_{rxh}^2 .$$ Thus,

$$Q^2 \cong$$

$$\left\{ 1 + \frac{2\sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2 + \left(1-\frac{1}{n}\right)\sum_{h=1}^{L} m_h s_{rxh}^2}{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2 + \sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2} + \frac{\left[\left(1-\frac{1}{n}\right)\sum_{h=1}^{L} m_h s_{rxh}^2\right]\left[2\sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2\right]}{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2 + \sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2} \right\}^{-1}$$

and $R^2 \cong \dfrac{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_r)^2}{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2 + \sum_{h=1}^{L} m_h(\bar{x}_{rh}-\bar{x}_c)^2 + \left(1-\frac{1}{n}\right)\sum_{h=1}^{L} m_h s_{rxh}^2}$.

Note that when L=1 these results are essentially the same as the comparable results for RI. If we assume that $x_{rh_i} = x_{rh} (= \bar{x}_{rh})$ for $i = 1,...,r_h$ then

$$\text{Var}(\hat{\beta}_{1c} \left| \{x^r_{r_{i=1}}\} \right.)_{RC} \cong \sigma^2 \left\{ \frac{\sum_{h=1}^{L} n_h (\bar{x}_{rh} - \bar{x}_c)^2 + 2\sum_{h=1}^{L} m_h (\bar{x}_{rh} - \bar{x}_c)^2}{\left[\sum_{h=1}^{L} n_h (\bar{x}_{rh} - \bar{x}_c)^2 \right]^2} \right\} \geq \text{Var}(\hat{\beta}_{1r}) \text{ if}$$

$\frac{r_h}{n_h} = \frac{r}{n} \geq \frac{1}{2}$ for $h = 1,...,L$ and

$$\text{Var}(\hat{\beta}_{oc} \left| \{x^r_{r_{i=1}}\} \right.)_{RC} \cong \sigma^2 \left\{ (\frac{n+2m}{n^2}) + \bar{x}_c^2 \left[\frac{\sum_{h=1}^{L} n_h (\bar{x}_{rh} - \bar{x}_c)^2 + 2\sum_{h=1}^{L} m_h (\bar{x}_{rh} - \bar{x}_c)^2}{\left[\sum_{h=1}^{L} n_h (\bar{x}_{rh} - \bar{x}_c)^2 \right]^2} \right] \right\}.$$

There is no easy comparison between $Var(\hat{\beta}_{oc})$ and $\text{Var}(\hat{\beta}_{or})$. Also,

$E(\sum_{i=1}^{n} (x_i^+ - \bar{x}_c)^2 \left| \{x^r_{r_{i=1}}\}) = \sum_{h=1}^{L} n_h (\bar{x}_{rh} - \bar{x}_c)^2 \right.$. Under the above assumptions and results it can be shown that

$$Q^2 \cong \frac{1}{\left[1 + \frac{2\sum_{h=1}^{L} m_h (\bar{x}_{rh} - \bar{x}_c)^2}{\sum_{h=1}^{L} n_h (\bar{x}_{rh} - \bar{x}_c)^2} \right]} \geq \frac{1}{2}, \text{ and } R^2 \cong \frac{\sum_{h=1}^{L} r_h (\bar{x}_{rh} - \bar{x}_r)^2}{\sum_{h=1}^{L} n_h (\bar{x}_{rh} - \bar{x}_c)^2}. \quad (2.3.49)$$

Note that $\bar{x}_c = \frac{\sum_{h=1}^{L} n_h \bar{x}_{rh}}{n}$ and $\bar{x}_r = \frac{\sum_{h=1}^{L} r_h \bar{x}_{rh}}{r}$. The value of Q^2 indicates that t_b in (2.3.6) may not be well approximated by a t-distribution.

2.3.2-5 Simple Regression Prediction Imputation (RG)

Here, $x^*_{mj} = \bar{x}_r$ for j=r+1,…,n, and $y^*_{mj} = \hat{\beta}_{or} + \hat{\beta}_{1r} \bar{x}_r$ where $\hat{\beta}_{or}$ and $\hat{\beta}_{1r}$ are defined in (2.1.10) and (2.1.11). It can be shown that

$$\text{E}(\hat{\beta}_{oc} \left| \{x^r_{r_{i=1}}\} \right.)_{RG} = \beta_o, \qquad \text{E}(\hat{\beta}_{1c} \left| \{x^r_{r_{i=1}}\} \right.)_{RG} = \beta_1, \quad (2.3.50)$$

and

$$\text{E}(\hat{\sigma}_c^2 \left| \{x^r_{r_{i=1}}\} \right.)_{RG} = \frac{(r-2)\sigma^2}{n-2}. \quad (2.3.51)$$

The bias of $\hat{\sigma}_c^2$ is $\frac{-m\sigma^2}{n-2}$. The variances of $\hat{\beta}_{1c}$ and $\hat{\beta}_{oc}$ are

$$\mathrm{Var}(\hat{\beta}_{1c}\,|\{x_{r_{i=1}}^r\})_{RG} = \frac{\sigma^2}{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_r)^2} = \mathrm{Var}(\hat{\beta}_{1r}),$$

$$\mathrm{Var}(\hat{\beta}_{oc}\,|\{x_{r_{i=1}}^r\})_{RG} = \sigma^2\left\{\frac{1}{r}+\frac{\bar{x}_r^2}{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_r)^2}\right\} = \mathrm{Var}(\hat{\beta}_{or}).$$

Since $\sum_{i=1}^{n}(x_i^+-\bar{x}_c)^2 = \sum_{i=1}^{r}(x_{r_i}-\bar{x}_r)^2$, if r is large then $Q^2 = R^2 \cong \frac{r}{n}$. Note that these results are essentially the same as the comparable results for the MO methods.

2.3.2-6 Random Regression Imputation (RRS)

For this method $x_{mj}^* = \bar{x}_r$ for j=r+1,…,n, and $y_{mj}^* = \hat{\beta}_{or} + \hat{\beta}_{1r}\,\bar{x}_r + \tilde{e}_{mj}$ where $\tilde{e}_{mj}$ is a residual randomly selected from the respondents' residuals. Then after considerable algebraic manipulation it can be shown that

$$\mathrm{E}(\hat{\beta}_{oc}\,|\{x_{r_{i=1}}^r\})_{RRS} = \beta_o, \qquad \mathrm{E}(\hat{\beta}_{1c}\,|\{x_{r_{i=1}}^r\})_{RRS} = \beta_1, \tag{2.3.52}$$

and

$$\mathrm{E}(\hat{\sigma}_c^2\,|\{x_{r_{i=1}}^r\})_{RRS} = \sigma^2\left\{1-\frac{m(m+1)}{n(n-2)}-\frac{m(n-1)}{nr(n-2)}\right\}. \tag{2.3.53}$$

$$\mathrm{Var}(\hat{\beta}_{1c}\,|\{x_{r_{i=1}}^r\})_{RRS} = \frac{\sigma^2}{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_r)^2} = \mathrm{Var}(\hat{\beta}_{1r}),$$

$$\mathrm{Var}(\hat{\beta}_{oc}\,|\{x_{r_{i=1}}^r\})_{RRS} = \sigma^2\left\{\frac{1}{r}+\frac{\bar{x}_r^2}{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_r)^2}\right\}+\frac{m(r-2)\sigma^2}{n^2 r} \geq \mathrm{Var}(\hat{\beta}_{or}).$$

If r is large, the third term in (2.3.53) will be negligible and

$$\mathrm{E}(\hat{\sigma}_c^2\,|\{x_{r_{i=1}}^r\})_{RRS} \cong \sigma^2\left\{1-\left(1-\frac{r}{n}\right)^2\right\}. \tag{2.3.54}$$

Since $\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)^2 = \sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2$, $Q^2 = R^2 \cong \left\{1 - \left(1 - \frac{r}{n}\right)^2\right\}$.

From (2.3.54) and the value of Q^2, t_b in (2.3.6) may be well approximated by a t-distribution.

2.3.2-7 Random Regression Imputation (RRN)

For this method $x_{mj}^* = \bar{x}_r$ and $y_{mj}^* = \hat{\beta}_{or} + \hat{\beta}_{1r}\,\bar{x}_r + e_{mj}$ for j=r+1,…,n, where e_{mj} is drawn From a distribution with mean zero and variance $\hat{\sigma}_r^2$ (see (2.1.12)). Then it can be shown that $\mathrm{E}(\hat{\beta}_{oc} \mid \{x_{r_{i=1}}^r\})_{RRN} = \beta_o$, $\mathrm{E}(\hat{\beta}_{1c} \mid \{x_{r_{i=1}}^r\})_{RRN} = \beta_1$, (2.3.55)

and for large n

$$\mathrm{E}(\hat{\sigma}_c^2 \mid \{x_{r_{i=1}}^r\})_{RRN} = \sigma^2\left\{1 - \frac{m}{n(n-2)}\right\} \cong \sigma^2. \quad (2.3.56)$$

$$\mathrm{Var}(\hat{\beta}_{1c} \mid \{x_{r_{i=1}}^r\})_{RRN} = \frac{\sigma^2}{\sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2} = \mathrm{Var}(\hat{\beta}_{1r}),$$

$$\mathrm{Var}(\hat{\beta}_{oc} \mid \{x_{r_{i=1}}^r\})_{RRN} = \sigma^2\left\{\frac{1}{r} + \frac{\bar{x}_r^2}{\sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2}\right\} + \frac{m}{n^2}\sigma^2 \geq \mathrm{Var}(\hat{\beta}_{or}).$$

Since $\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)^2 = \sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2$, $Q^2 = R^2 \cong 1$. Therefore, t_b in (2.3.6) may be well approximated by a t-distribution.

2.3.3 Comparisons

The best way to choose among the imputation methods is to assess their properties considering data analytical objectives and populations of interest. Thus, one might evaluate $\mathrm{E}(\hat{\beta}_{1c})$, $\mathrm{E}(\hat{\sigma}_c^2)$ and Q^2 for each of the imputation methods using values of n, $\{r_h\}$, $\{n_h\}$, $\{x_{rh_i}\}$ and $\{x_{mh_j}\}$ corresponding to applications of interest.

In this section we compare the alternative imputation methods by evaluating $\mathrm{E}(\hat{\beta}_{1c})$, $\mathrm{E}(\hat{\sigma}_c^2)$ and Q^2 (or R^2). The case where X has no missing values is considered first.

2.3.3-1 MO vs RI

In general, the biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ for MO and RI are equal (see (2.3.9), (2.3.10).) Because the expressions for E($\hat{\sigma}_c^2$) are complicated (see (2.3.11) and (2.3.15)), we made a simplifying assumption that $x_{mj} = \bar{x}_m = \bar{x}_r$. Then, the biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are zero for both MO and RI (see (2.3.12) and (2.3.13)), the bias of $\hat{\sigma}_c^2$ for RI will be small if the observed X values are close to each other (see Section 2.3.1-2) and the bias of $\hat{\sigma}_c^2$ for MO is equal to $-\frac{m\sigma^2}{n-2}$. Similarly, by assuming small variability of the X values we obtain $Q_{RI}^2 > Q_{MO}^2$ when r is large. These results suggest that RI is preferable to MO.

2.3.3-2 MC vs RC

Making no assumptions the biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ for these two methods are equal (see (2.3.18) and (2.3.19).) Under the simplifying assumption that $x_{hi} = x_h$ for i=1,…, n_h, the biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are zero for MC and RC (see (2.3.22) and (2.3.23)). If r is large, $\frac{r_h}{n_h} \geq \frac{n_h}{2}$ and L is small relative to n then $E(\hat{\sigma}_c^2)_{RC} \cong \sigma^2 > E(\hat{\sigma}_c^2)_{MC} \cong \frac{r\sigma^2}{n}$ (see(2.3.25) and (2.3.30)). If, in addition, it is assumed that $\{x_{ri} : i = 1,...,r\}$is a random sample from $\{x_i : i = 1,2,...,n\}$ then $R_{MC}^2 \cong \left(\frac{r}{n}\right)^2 < \left(\frac{r}{n}\right) \cong R_{RC}^2$. These results suggest that RC is preferable to MC (no simple comparison can be made between Q_{MC}^2 and Q_{RC}^2.)

2.3.3-3 MO vs MC and RI vs RC

Under the simplifying assumption, $x_{hi} = x_h$ for $i = 1,...,n_h$, $E(\hat{\beta}_{oc})_{MC} = \beta_o$ and $E(\hat{\beta}_{1c})_{MC} = \beta_1$ but the biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ for MO are equal to, respectively,

$$\beta_1 \left\{ \sum_{h=1}^{L} \frac{r_h x_h}{r} - \left(\sum_{h=1}^{L} \frac{n_h x_h}{n} \right) \frac{\sum_{h=1}^{L} r_h (x_h - \sum_{h=1}^{L} \frac{r_h x_h}{r})^2}{\sum_{h=1}^{L} n_h (x_h - \sum_{h=1}^{L} \frac{n_h x_h}{n})^2} \right\} \text{ and } \beta_1 \left\{ \frac{\sum_{h=1}^{L} r_h (x_h - \sum_{h=1}^{L} \frac{r_h x_h}{r})^2}{\sum_{h=1}^{L} n_h (x_h - \sum_{h=1}^{L} \frac{n_h x_h}{n})^2} - 1 \right\}.$$

Since MC passes the first test corresponding to the t_b-statistic (see (2.3.6)) while MO doesn't. Thus, MC is better than MO. Same result apply to the comparison that RC is better than RI.

2.3.3-4 RG vs RRS

Under the assumptions that $x_{mj} = \bar{x}_m = \bar{x}_r$ and r is large, $Q^2_{RRS} \cong 1\text{-}(1-\frac{r}{n})^2$ and

$Q^2_{RG} \cong (\frac{r}{n}) \frac{\sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} < \frac{r}{n} < Q^2_{RRS}$. Thus, we prefer RRS to RG.

2.3.3-5 RRS vs RRN

If r is large, $E(\hat{\sigma}^2_c)_{RRN} \cong \sigma^2 > E(\hat{\sigma}^2_c)_{RRS}$ (see (2.3.34)) and (2.3.36). In addition, if $x_{mj} = \bar{x}_m = \bar{x}_r$ then $Var(\hat{\beta}_{1c})_{RRN} = Var(\hat{\beta}_{1c})_{RRS} = Var(\hat{\beta}_{1r})$. Therefore, $Q^2_{RRN} > Q^2_{RRS}$ and RRN is preferable to RRS.

Considering the case where both X and Y have missing values the biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are zero for each of the methods. Hence, we only need to evaluate E($\hat{\sigma}^2_c$) and Q^2 (or R^2) for each of the imputation methods.

2.3.3-6 MO vs RI

Since $E(\hat{\sigma}^2_c)_{MO} = \frac{r-2}{n-2}\sigma^2 < E(\hat{\sigma}^2_c)_{RI} \cong \sigma^2$ (see (2.3.40) and (2.3.42)) and

$Q^2_{MO} \cong \frac{r}{n} < \frac{r}{n-\frac{m}{n}} \cong Q^2_{RI}$, RI is preferable to MO.

2.3.3-7 MC vs RC

Using first order Taylor series approximations, $E(\hat{\sigma}^2_c)_{RC} \cong \sigma^2$. If we assume that $r_{h\geq} \frac{n_h}{2}$, r is large and L is small relative to n, $E(\hat{\sigma}^2_c)_{MC} \cong \frac{r\sigma^2}{n}$. If, in addition, it is assumed that $x_{rh_i} = x_{rh}$ for $i = 1,...,r_h$, then $R^2_{RC} \geq R^2_{MC}$ (see (2.3.46) and (2.3.49)). Hence, RC is preferable to MC. No simple comparison can be made between Q^2_{MC} and Q^2_{RC}.

Without numerical studies there are no clear comparisons of MO vs MC or RI vs RC.

2.3.3-8 RRS vs RI

If r is large, $Q^2_{RRS} \cong 1\text{-}(1-\frac{r}{n})^2 > \frac{r}{n-\frac{m}{n}} \cong Q^2_{RI}$. Hence, RRS is preferable to RI.

2.3.3-9 RRN vs RRS

If r is large, $E(\hat{\sigma}^2_c)_{RRN} \cong \sigma^2 > E(\hat{\sigma}^2_c)_{RRS} \cong \sigma^2 \{1\text{-}(1-\frac{r}{n})^2\}$ and $Q^2_{RRN} \cong 1 > Q^2_{RRS} \cong 1\text{-}(1-\frac{r}{n})^2$. Thus, RRN is preferable to RRS.

2.3.3-10 RRN and RRS vs RG

Since $Q^2_{RG} = Q^2_{MO} \cong \frac{r}{n}$ and RI is preferable to MO, RRN and RRS are preferable to RG.

2.3.4 Summary Tables and Conclusions

In section 2.3 we investigated the effect of different imputation methods on the properties of the completed interval (2.3.4). Assuming that X has no missing values, Table 2.3.1 presents the general forms of the variances of $\hat{\beta}_{oc}$ for seven imputation methods. Table 2.3.2 summarizes the results for the biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$. Since the expressions for E($\hat{\sigma}^2_c$) are rather complicated for most of the imputation methods, Table 2.3.3 summarizes the biases of $\hat{\sigma}^2_c$ (under special assumptions) for seven imputation methods when X has no missing values. Note that for large n the bias of $\hat{\sigma}^2_c$ is approximated by zero for RRN (without having to make any special assumptions.) Table 2.3.4 summarizes the biases of $\hat{\sigma}^2_c$ when both X and Y have missing values. Table 2.3.5 and 2.3.6 summarize the approximate values of Q^2 and the conditions under which these approximations hold. Note that RRN is the only method which can yield $Q^2 \cong 1$.

From the results presented in Section 2.3 only a very few general conclusions about the relative merits of the methods can be made. Additional conclusions can be drawn by making assumptions which approximate conditions typical in applications; e.g., n large or the values of X in an imputation cell are equal.

The case where X has no missing values is considered first. While the biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are zero for RG, RRS and RRN it is, in general, nonzero for MO, RI, MC and RC.

Making the assumption that all values of X within an imputation cell are equal, the biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are zero for MC and RC but not for MO and RI (see Section 2.3.3-3.) This lead us to exclude MO and RI from further consideration.

We next consider the bias of $\hat{\sigma}_c^2$. Assuming that all values of X within an imputation cell are equal, $r_{h\geq} \frac{n_h}{2}$, n is large and L is small relative to n, then the bias of $\hat{\sigma}_c^2$ is approximated by zero for RC but $-\frac{m}{n}\sigma^2$ for MC (see (2.3.25) and (2.3.30).) Thus, we prefer RC to MC. If we assume that $x_{mj} = \bar{x}_m = \bar{x}_r$ and r is large, $Q^2_{RRN} > Q^2_{RRS} > Q^2_{RG}$. Thus, we prefer RRN to RRS and RG.

As a consequence of the evaluation of the biases of $\hat{\beta}_{oc}$, $\hat{\beta}_{1c}$, $\hat{\sigma}_c^2$ and the values of Q^2, we regard RRN and RC as preferable to the remainder. The choice between them clearly depends on the characteristics of the imputation cells, the fit of the model (2.1.1) and the response rate.

Considering the case where both X and Y have missing values the biases of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ are zero for each of the methods. However, if r is large the bias of $\hat{\sigma}_c^2$ is substantially larger for MO than RI and for RG than for RRS or RRN. Similarly, if $r_{h\geq} \frac{n_h}{2}$, n is large and L is small relative to n, then the bias of $\hat{\sigma}_c^2$ is substantially larger for MC than for RC. This leads us to exclude MO, RG and MC from further consideration.

If r is large, $Q^2_{RRN} > Q^2_{RRS} > Q^2_{RG}$ which implies a preference for RRN over RRS and RI. As a consequence of the evaluation of the biases of $\hat{\sigma}_c^2$ and values of Q^2, we regard RRN and RC as preferable to the remainder. Again, the choice between them will depend on the characteristics of the imputation cells, the fit of the model (2.1.1) and the response rate.

2.4 Discussion

In Section 2.3 evaluations of the formulas for populations of interest are needed to obtain more definite conclusions about the properties of the alternative imputation methods. However, in view of the results of our investigations, the conclusions in Section 2.3.4

seems reasonable: Under the model (2.1.1) RC and RRN are preferable to the alternatives. The choice between RC and RRN will depend on the characteristics of the imputation cells, the fit of the model and the response rate. (Note that when L=1 RC reduces to RI.)

Although the model (2.1.1) is simple and the analytical objectives that we have considered are modest, our results are important because simple linear regression is a viable model and providing appropriate confidence intervals for β_o and β_1 is a reasonable objective. For the types of imputation method studied in this article we believe that the results will be replicated in situations where there are more complicated models and/or more sophisticated data analytic procedures are employed: The variability of statistics will tend to be underestimated by use of the completed data sets as if they only contained observed data. Specifically, the upper bounds for Q^2 do not provide adequate that t_b in (2.3.6) may be well approximated by t-distributions. The value of Q^2 may be too small for two reasons: (a) $\hat{\sigma}_c^2$ is an underestimate of σ^2 or (b) $\sum_{i=1}^{n}(x_i^+ - \bar{x}_c)^2$ is too large. As is evident from the results in Section 2.3 there are imputation methods that will provide completed data sets leading to reasonable estimators of σ^2. The most important concern is, then, (b).

Assume the model (2.1.1), both X and Y have missing values, a confidence interval for β_1 is desired (see (2.3.4)) and the imputation method is RRN (see Section 2.3.2-7). Then, if n is large, $Q^2 \cong 1$.

Thus, we believe that there are situations (i.e., models and analytical objectives) where use of a specific imputation method will yield a completed data set which a secondary data analyst can properly treat as if it contained only observed values. However, if the model and/or analytical objectives are changed, treating the same data set as if it had only observed responses could lead to poor inferences. We are not confident about finding a universally (minimally) acceptable imputation method. The simple model considered in this article clearly indicate the difficulties. Appropriate treatment of missing values in survey data may thus require special computer software as, for example, is need to implement the multiple imputation methodology. We do not find this to be a felicitous prospect since there are a multitude of secondary data analysts with minimal statistical competence and with even less knowledge of the mechanism generating the missing data.

It seems unlikely that comparing the seven imputation methods using additional models and data analytical objectives will prove to be especially useful. This article probably displays most of the difficulties associated with the uncritical use of imputed data. Rather, researchers should consider situations where the missing data cannot be regarded as missing at random. While many of the difficulties associated with using a completed data set as if it contained only observed values will persist, alternative ways of handling very large but incomplete data sets may be far worse.

Table 2.3.1

The variances of $\hat{\beta}_{oc}$ for seven imputation methods when X has no missing values

$$\mathrm{Var}(\hat{\beta}_{oc})_{MO} = \sigma^2 \left\{ \frac{1}{r} + \frac{\bar{x}^2 \sum_{i=1}^{r} (x_{r_i} - \bar{x}_r)^2}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2} \right\}$$

$$\mathrm{Var}(\hat{\beta}_{oc})_{MC} = \sigma^2 \left\{ \sum_{h=1}^{L} \left(\frac{n_h}{n} \right)^2 \frac{1}{r_h} + \frac{\bar{x}^2}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2} \sum_{h=1}^{L} \sum_{i=1}^{r_h} \left[(x_{r_{h_i}} - \bar{x}) + \frac{m_h (\bar{x}_{m_h} - \bar{x})}{r_h} \right]^2 - \right.$$

$$\left. \frac{2\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \left[\sum_{h=1}^{L} \frac{n_h}{n} (\bar{x}_{r_h} - \bar{x}) + \sum_{h=1}^{L} \frac{n_h m_h}{n r_h} (\bar{x}_{m_h} - \bar{x}) \right] \right\}$$

$$\mathrm{Var}(\hat{\beta}_{oc})_{RI} = \beta_1^2 s_{r_x}^2 \sum_{j=r+1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_{m_j} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)^2 + \sigma^2 \left\{ \left(\frac{1}{n} + \frac{2m}{n^2} \right) + \right.$$

$$\left. \frac{\bar{x}}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \left[n\bar{x} + \frac{2m(r-m)}{n} (\bar{x}_r - \bar{x}_m) \right] + \frac{2\bar{x}^2 m (\bar{x}_m - \bar{x})(\bar{x}_r - \bar{x})}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2} \right\}$$

$$\mathrm{Var}(\hat{\beta}_{oc})_{RC} = \beta_1^2 s_{rx_h}^2 \left\{ \sum_{h=1}^{L} \sum_{j=r_h+1}^{n_h} \left[\frac{1}{n} - \frac{\bar{x}(x_{m_{h_j}} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^2 \right\} + \sigma^2 \left\{ \left(\frac{1}{n} + \frac{2m}{n^2} \right) + \right.$$

$$\left. \frac{\bar{x}}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \left[n\bar{x} - 2 \sum_{h=1}^{L} m_h (\bar{x}_{m_h} + \bar{x}_{r_h} - 2\bar{x}) \right] + \frac{2\bar{x}^2 \sum_{h=1}^{L} m_h (\bar{x}_{m_h} - \bar{x})(\bar{x}_{r_h} - \bar{x})}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2 \right]^2} \right\}$$

Table 2.3.1 (continued)

$$\mathrm{Var}(\hat{\beta}_{oc})_{RG} = \sigma^2 \left\{ \frac{1}{r} + \frac{(\bar{x}_r - \bar{x}_m)^2}{\sum_{i=1}^{r} (x_{r_i} - \bar{x}_r)^2} \right\} -$$

$$\frac{2m\bar{x}\sigma^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \left\{ (\bar{x}_r - \bar{x}) + m(\bar{x}_m - \bar{x}_r) \left[\frac{1}{r} + \frac{\bar{x}_r(\bar{x}_r - \bar{x}_m)}{\sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2} \right] + \frac{\sum_{j=r+1}^{n}(x_{m_j} - \bar{x})x_{m_j}}{\sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2} (\bar{x}_m - \bar{x}_r) \right\} +$$

$$\frac{\bar{x}^2\sigma^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} \left\{ \sum_{i=1}^{r}(x_{r_i} - \bar{x})^2 + \frac{\sum_{j=r+1}^{n}(x_{m_j} - \bar{x})^2}{r} + \frac{\sum_{j=r+1}^{n}(x_{m_j} - \bar{x})^2 (x_{m_j} - \bar{x}_r)^2}{\sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2} \right.$$

$$\left. - \frac{2rm^2(\bar{x}_r - \bar{x}_m)^2}{n^2} \right\}$$

$$\mathrm{Var}(\hat{\beta}_{oc})_{RRS} = \frac{\sigma^2(n + 2m + \frac{m}{r})}{n^2} + \frac{\sigma^2 \sum_{j=r+1}^{n}(x_{m_j} - \bar{x}_r)^2}{n^2 \sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2} + \frac{\bar{x}\sigma^2 \left[\bar{x}_r + \frac{m(\bar{x}_r - \bar{x}_m)}{n} \right]}{\sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2}$$

$$+ \frac{\bar{x}\sigma^2(r-2)}{r\sum_{i=1}^{n}(x_i - \bar{x})^2} \left\{ \frac{\sum_{j=r+1}^{n}(x_{m_j} - \bar{x})^2 \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \frac{2m(\bar{x}_m - \bar{x})}{n} \right\}$$

$$\mathrm{Var}(\hat{\beta}_{oc})_{RRN} = \frac{\sigma^2(n + 2m + \frac{m}{r})}{n^2} + \frac{\sigma^2 \sum_{j=r+1}^{n}(x_{m_j} - \bar{x}_r)^2}{n^2 \sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2} + \frac{\bar{x}\sigma^2 \left[\bar{x}_r + \frac{m(\bar{x}_r - \bar{x}_m)}{n} \right]}{\sum_{i=1}^{r}(x_{r_i} - \bar{x}_r)^2}$$

$$+ \frac{\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \left\{ \frac{\sum_{j=r+1}^{n}(x_{m_j} - \bar{x})^2 \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} - \frac{2m(\bar{x}_m - \bar{x})}{n} \right\}$$

Table 2.3.2 Biases of the conditional expectations of $\hat{\beta}_{oc}$ and $\hat{\beta}_{1c}$ for seven imputation Methods when X has no missing values.

Imputation method	Bias ($\hat{\beta}_{oc}$)	Bias($\hat{\beta}_{1c}$)
MO and RI	$b_1 = \beta_1 \left\{ \bar{x}_r - \frac{\bar{x}\sum_{i=1}^{r}(x_{r_i}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\}$	$-\beta_1 \left\{ \frac{r(\bar{x}_r-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} + \frac{\sum_{j=r+1}^{n}(x_{m_j}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\} = b_3$
MC and RC	$b_2 = \beta_1 \left\{ \sum_{h=1}^{L}\frac{n_h\bar{x}_{r_h}}{n} - \frac{\bar{x}\left[\sum_{i=1}^{r}(x_{r_i}-\bar{x})x_{r_i}\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2} - \frac{\bar{x}\left[\sum_{h=1}^{L}m_h\bar{x}_{r_h}(\bar{x}_{m_h}-\bar{x})\right]}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\}$	$\beta_1 \left\{ \frac{r\bar{x}(\bar{x}_r-\bar{x}) + \sum_{h=1}^{L}m_h\bar{x}_{r_h}(\bar{x}_{m_h}-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} - \frac{\sum_{j=r+1}^{n}(x_{m_j}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right\} = b_4$
RG	0	0
RRS	0	0
RRN	0	0

Note: If we assume that $x_{m_j} = \bar{x}_m = \bar{x}_r$ then $b_1 = b_3 = 0$. If we assume that $x_{h_i} = x_h$ then $b_2 = b_4 = 0$. When X and Y both have missing values, all the seven imputation methods provide unbiased estimators for β_o and β_1.

Table 2.3.3 Biases* of $\hat{\sigma}_c^2$ for seven imputation Methods when X has no missing values

Imputation method	Bias $(E(\hat{\sigma}_c^2 \mid \{x_{i=1}^n\}) - \sigma^2)$	Approximate Bias
MO	$-\frac{m}{n-2}\sigma^2$	$-\frac{m}{n}\sigma^2$ (if n is large)
RI	$-\frac{m(n+r-1)\sigma^2}{nr(n-2)} + \frac{\beta_1^2 s_{rx}^2 m(n-1)}{n(n-2)}$	$\sigma^2\left\{1+\frac{\beta_1^2 s_{rx}^2}{\sigma^2}(1-\frac{r}{n})\right\}$ (if r is large)
MC	$\frac{\sigma^2}{n-2}\left\{-m+\sum_{h=1}^{L}(1-\frac{n_h}{n})\frac{m_h}{r_h} - \frac{\sum_{h=1}^{L} n_h \frac{m_h}{r_h}(x_h-\bar{x})^2}{\sum_{h=1}^{L} n_h (x_h-\bar{x})^2}\right\}$	$-\frac{m}{n}\sigma^2$ (if $r_h \geq \frac{n_h}{2}$, n is large and L is small relative to n)
RC	$-\frac{\sigma^2}{n-2}\left\{\frac{m}{n}+\sum_{h=1}^{L}\frac{(n_h-1)m_h}{nr_h} + \frac{\sum_{h=1}^{L}\frac{m_h(n_h+r_h-1)}{r_h}(x_h-\bar{x})^2}{\sum_{h=1}^{L} n_h (x_h-\bar{x})^2}\right\}$	0 (if $r_h \geq \frac{n_h}{2}$, n is large and L is small relative to n)
RG	$-\frac{m}{n-2}\sigma^2$	$-\frac{m}{n}\sigma^2$ (if n is large)
RRS	$-\frac{\sigma^2}{n-2}\left\{\frac{m(m+1)+\frac{m(n-1)}{r}}{n}\right\}$	$-(\frac{m}{n})^2\sigma^2$ (if r is large)
RRN	$-\frac{\sigma^2}{n-2}\left\{\frac{m}{n}+\frac{\sum_{j=r+1}^{n}(x_{m_j}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right\}$	0 (if n is large)

*The biases under MO, RI and RRS assume $x_{m_j} = \bar{x}_m = \bar{x}_r$, the biases under MC and RC assume $x_{h_i} = x_h$ while no assumptions are made for the biases under RG and RRN.

Table 2.3.4 Biases of $\hat{\sigma}_c^2$ for seven imputation Methods when X has missing values

Imputation method	Bias $(E(\hat{\sigma}_c^2 \mid \{x_{r_i=1}^r\}) - \sigma^2)$.	Approximate Bias	
MO	$-\frac{m}{n-2}\sigma^2$	$-\frac{m}{n}\sigma^2$	(if r is large)
RI	0	0	
MC	$\frac{\sigma^2}{n-2}\left\{-m+\sum_{h=1}^{L}(1-\frac{n_h}{n})\frac{m_h}{r_h} - \frac{\sum_{h=1}^{L} n_h \frac{m_h}{r_h}(\bar{x}_{r_h}-\bar{x}_c)^2}{\sum_{i=1}^{r}(x_{r_i}-\bar{x}_c)^2+\sum_{h=1}^{L} m_h(\bar{x}_{r_h}-\bar{x}_c)^2}\right\}$	$-\frac{m}{n}\sigma^2$	(if $r_h \geq \frac{n_h}{2}$, n is large and L is small relative to n)
RC	0	0	
RG	$-\frac{m}{n-2}\sigma^2$	$-\frac{m}{n}\sigma^2$	(if r is large)
RRS	$-\sigma^2\left\{\frac{m(m+1)}{n(n-2)}+\frac{m(n-1)}{rn(n-2)}\right\}$	$-(\frac{m}{n})^2\sigma^2$	(if r is large)
RRN	$-\sigma^2\left\{\frac{m}{n(n-2)}\right\}$	0	(if n is large)

Note: The unbiasedness of $\hat{\sigma}_c^2$ for RI and RC are obtained by using first order Tayler Series approximation.

Table 2.3.5 The approximate values of Q^2 for seven imputation Methods when X has no Missing values

Imputation Method	Approximate value of Q^2	
MO	$\frac{r}{n}$	(if $x_{m_j} = \bar{x}_m = \bar{x}_r$ and r is large)
RI	$\left\{1+\frac{\beta_1^2 s_{rx}^2}{\sigma^2}(1-\frac{r}{n})\right\}$	(if $x_{m_j} = \bar{x}_m = \bar{x}_r$ and r is large)
MC	$\frac{\frac{r}{n}}{1+\frac{\sum_{h=1}^{L} n_h \frac{m_h}{r_h}(x_h-\bar{x})}{\sum_{h=1}^{L} n_h (x_h-\bar{x})^2}}$	(if $x_{h_i} = x_h$, $r_h \geq \frac{n_h}{2}$, n is large and L is small relative to n)
RC	$\frac{1}{1+\frac{2\sum_{h=1}^{L} m_h (x_h-\bar{x})^2}{\sum_{h=1}^{L} n_h (x_h-\bar{x})^2}}$	(if $x_{h_i} = x_h$, $r_h \geq \frac{n_h}{2}$, n is large and L is small relative to n)
RG	$\left(\frac{r}{n}\right)^2$	(if r is large, $x_{m_j} = \bar{x}_m = \bar{x}_r$ and $\{x_{r_i} : i = 1,...,r\}$ is a random sample from $\{x_i : i = 1,...,n\}$)
RRS	$1-(1-\frac{r}{n})^2$	(if r is large and $x_{m_j} = \bar{x}_m = \bar{x}_r$)
RRN	$\frac{\frac{r}{n}}{1+(\frac{r}{n})(1-\frac{r}{n})}$	(if n is large and $\{x_{r_i} : i = 1,...,r\}$ is a random sample from $\{x_i : i = 1,...,n\}$)

Note: $Q^2_{RRN} \cong 1$ if we assume $x_{m_j} = \bar{x}_m = \bar{x}_r$ and n is large

Table 2.3.6 The approximate values of Q^2 for seven imputation Methods when X has Missing values

Imputation Method	Approximate value of Q^2	
MO	$\frac{r}{n}$	(if r is large)
RI	$\frac{r}{(n-\frac{m}{n})}$	(by using first order Taylor Series approximation)
MC	$\frac{\frac{r}{n}}{1+\frac{\sum_{h=1}^{L} n_h \frac{m_h}{r_h}(\bar{x}_{r_h}-\bar{x}_c)^2}{\sum_{h=1}^{L} n_h(\bar{x}_{r_h}-\bar{x}_c)^2}}$	(if $x_{r_{h_i}} = x_{r_h}$, $r_h \geq \frac{n_h}{2}$, n is large and L is small relative to n)
RC	$\frac{1}{1+\frac{2\sum_{h=1}^{L} m_h(\bar{x}_{r_h}-\bar{x}_c)^2}{\sum_{h=1}^{L} n_h(\bar{x}_{r_h}-\bar{x}_c)^2}}$	(if $x_{r_{h_i}} = x_{r_h}$, $r_h \geq \frac{n_h}{2}$ and by using first order Taylor Series approximation)
RG	$\frac{r}{n}$	(if r is large)
RRS	$1-(1-\frac{r}{n})^2$	(if r is large)
RRN	1	(if n is large)

REFERENCE

Bailar, B. A., Bailey, L. and Corby, C. A. (1978), " A Comparison of Some Adjustment and Weighting Procedures for Survey Data," *Survey Sampling and Measurement* (Namboodiri, N. K. ed.), 175-198, New York: Academic Press.

Bailar, B. A., Bailey, L. (1978), " Comparison of Two Procedures for Imputing Missing Survey Values," *Proceedings Section of Survey Research Method, American Statistical Association*, 462-467.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), " Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of Royal Statistical Society*, B, 39, 1-38.

Ernst, L. R. (1978), " Weighting to Adjust for Partial Nonresponses," *Proceedings Section of Survey Research Method, American Statistical Association*, 468-473

Kalton, G. and Kasprzyk, D. (1982), " Imputing for Missing Survey Responses," *Proceedings Section of Survey Research Method, American Statistical Association*, 22-33.

Little, J. A. and Rubin, D. B. (1987), " *Statistical Analysis With Missing Data*," New York: Wiley.

Platek, R., Singh, M. P. and Tremblay, V. (1978), " Adjustment for Nonresponse in Surveys," *Survey Sampling and Measurement* (Namboodiri, N. K. ed.), 157-174, New York: Academic Press.

Santos, R. L. (1981b), " Effects of Imputation on Regression Coefficients," *Proceedings Section of Survey Research Method, American Statistical Association*, 140-145.