Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression

Michael P. Jones

Journal of the American Statistical Association, Vol. 91, No. 433 (Mar., 1996), 222-230.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28199603%2991%3A433%3C222%3AIASMFM%3E2.0.CO%3B2-C

Journal of the American Statistical Association is currently published by American Statistical Association.

Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression

Michael P. JONES

The statistical literature and folklore contain many methods for handling missing explanatory variable data in multiple linear regression. One such approach is to incorporate into the regression model an indicator variable for whether an explanatory variable is observed. Another approach is to stratify the model based on the range of values for an explanatory variable, with a separate stratum for those individuals in which the explanatory variable is missing. For a least squares regression analysis using either of these two missing-data approaches, the exact biases of the estimators for the regression coefficients and the residual variance are derived and reported. The complete-case analysis, in which individuals with any missing data are omitted, is also investigated theoretically and is found to be free of bias in many situations, though often wasteful of information. A numerical evaluation of the bias of two missing-indicator methods and the complete-case analysis is reported. The missing-indicator methods show unacceptably large biases in practical situations and are not advisable in general.

KEY WORDS: Epidemiology; Incomplete data; Missing data; Psychology.

1. INTRODUCTION

It is quite common in practice that a statistician planning on performing a regression analysis finds that the explanatory variable information is incomplete on some subjects. There can be several mechanisms at work that produce the incompleteness. The statistician, of course, wishes to apply a strategy that comes as close as possible to the true regression had data not been missing. For the purpose of this article, it is sufficient to assume that the true regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \qquad i = 1, \ldots, n, \quad (1)$$

where the ε_i are independent error terms with mean zero and common variance σ^2. We also assume throughout that Y and X_1 are always measured but X_2 may be missing. The most commonly used method for handling such data is the complete-case analysis, so called because any subject with missing data is removed from the analysis. This approach is valid when X_2 is not missing as a function of either Y or ε and is useful when the vast majority of the cases are complete. But when a fair proportion of the data are missing, the complete-case method is very wasteful of information. A multitude of alternative missing-data techniques have been devised to reclaim as much of the available data as possible. Review articles by Afifi and Elashoff (1967), Anderson, Basilevsky, and Hum (1983), and Little (1992) have summarized many of the missing–explanatory variable regression methods. There are, however, a couple of classes of missing-data methods in common use in various disciplines that have not yet been investigated for their validity.

The two classes of missing-data methods to be studied here fall under the headings of missing-indicator methods and stratification methods. These methods have been proposed for use in the areas of behavioral sciences, epidemiology, sample survey research, and business and economics. In the *missing-indicator* method, an indicator of whether an explanatory variable is missing is worked into the regression model (1). In particular, if Q_2 is a binary indicator of whether X_2 is observed, then (1) is modified by replacing X_2 by X_2Q_2 and by adding $1 - Q_2$ as another predictor. This procedure has been suggested by Anderson et al. (1983), Chow (1979), Cohen and Cohen (1975), and Miettinen (1985). Cohen and Cohen (1975, p. 274) argued that such methodology uses all the available information, including the presence or absence of values on the explanatory variable. Thus one avoids both "the risk of nonrepresentativeness in dropping subjects if data are missing nonrandomly" and as well lower power from reduced sample size "even if data are missing randomly." In the context of logistic regression, Chow (1979) proposed a variation on this method in which the interaction term X_1Q_2 is also added as a predictor to the model. A special case of the missing-indicator method is of interest. In a one-way analysis-of-variance problem, a common method of handling observations with unknown group identification is to create a separate group for them. This "missing" group is created by adding a $1 - Q$ term to the regression model and replacing the true dummy group indicators X_j by the observed group indicators X_jQ.

A close relative to the missing-indicator methods is the class of stratification methods. Suppose that X_2 in model (1) can assume only k distinct values. In this class of procedures, the analysis is stratified into $k + 1$ submodels, one for each group of subjects with a distinct value of X_2 and then a $(k + 1)$th stratum for those with unknown X_2. This is commonly done when X_2 is a confounder and X_1 is the predictor of interest.

These classes of missing-data procedures for regression have been proposed for use regardless of the missing-data mechanism. These methods' appeal is that they incorporate the observed "missingness" into the model. A basic check

Michael P. Jones is Associate Professor, Department of Preventive Medicine and Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242. This work was supported in part by National Cancer Institute Grant CA55212. The author thanks Jon Lemke for helpful discussion and the referees and associate editor for suggestions that improved the recommendations for what to do in practice.

on the validity of such modeling is whether the methods produce biased estimation if no information about the regression of Y on (X_1, X_2) is contained within the "missingness" aspect of the data. The goal of this article is to investigate whether these procedures are biased when the true regression model is given in (1) and X_2 can be missing as a function of X_1 or X_2 but not as a function of either Y or ε. These missing–explanatory variable methodologies are quite general in that they can be implemented for any type of regression: least squares, generalized linear models, Cox proportional hazards regression, and others. Here we treat least squares, because we can then derive exact biases of the estimators for the regression coefficients and variances. It is reasonable to conjecture that if parameter estimators are biased for least squares, then they are probably biased for the other regressions as well.

The complete-case method is reviewed in Section 2, which will help to set the stage for the other methods. The missing-indicator methods are investigated in Section 3, with a special section looking at the missing-group method. In Section 4 the stratification methods are considered. An evaluation of the magnitude of bias is carried out in Section 5 for the purpose of recommending procedures to use in practice. Some concluding comments are contained in Section 6.

2. REVIEW OF THE COMPLETE-CASE METHOD

Before studying the missing-indicator and stratification methods for handling missing data, it will be helpful to review a standard missing-data technique—the complete-case analysis, so called because any individuals with missing data are excluded from the analysis. The true model is given by (1). Let Q_i be 1 if the ith individual has complete data and be zero otherwise. Define $\mathbf{Q} = \text{diag}(Q_1, \ldots, Q_n)$. Then the complete-case model is

$$Y_i Q_i = \theta_{c0} Q_i + \theta_{c1} X_{1i} Q_i + \theta_{c2} X_{2i} Q_i + \varepsilon_i Q_i, \quad i = 1, \ldots, n \tag{2}$$

or, equivalently, as $\mathbf{QY} = \mathbf{QX}\boldsymbol{\theta}_c + \mathbf{Q}\boldsymbol{\varepsilon}$. The following theorem summarizes the properties of the complete-case least squares analysis.

Theorem 2.1. The least squares estimator for model (2) is

$$\hat{\boldsymbol{\theta}}_c = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{QX})^{-1}\mathbf{X}'\mathbf{Q}\boldsymbol{\varepsilon}. \tag{3}$$

If $\mathbf{Q}$ is independent of ε, then, conditional on $\mathbf{Q}$,

$$\begin{aligned} E(\text{RSS}) &= E(\mathbf{QY} - \mathbf{QX}\hat{\boldsymbol{\theta}}_c)'(\mathbf{QY} - \mathbf{QX}\hat{\boldsymbol{\theta}}_c) \\ &= \sigma^2\left(\sum Q_i - 3\right). \end{aligned}$$

Proof. Let $\mathbf{X}_c = \mathbf{QX}$ and $\mathbf{Y}_c = \mathbf{QY}$. Then $\hat{\boldsymbol{\theta}}_c = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{Y}_c = (\mathbf{X}'\mathbf{QX})^{-1}\mathbf{X}'\mathbf{QY} = (\mathbf{X}'\mathbf{QX})^{-1}\mathbf{X}'\mathbf{Q}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{QX})^{-1}\mathbf{X}'\mathbf{Q}\boldsymbol{\varepsilon}$, giving the first result. Next, recall the standard result

$$\begin{aligned} \text{RSS} &= (\mathbf{Y}_c - \mathbf{X}_c\hat{\boldsymbol{\theta}}_c)'(\mathbf{Y}_c - \mathbf{X}_c\hat{\boldsymbol{\theta}}_c) \\ &= (\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\beta})'(\mathbf{Y}_c - \mathbf{X}_c\boldsymbol{\beta}) \\ &\quad - (\hat{\boldsymbol{\theta}}_c - \boldsymbol{\beta})'\mathbf{X}_c'\mathbf{X}_c(\hat{\boldsymbol{\theta}}_c - \boldsymbol{\beta}). \end{aligned}$$

The first term on the right is $\boldsymbol{\varepsilon}'\mathbf{Q}\boldsymbol{\varepsilon}$, and the second term is easily seen to be $\boldsymbol{\varepsilon}'\mathbf{H}_c\boldsymbol{\varepsilon}$, where $\mathbf{H}_c = \mathbf{X}_c(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'$. Hence $\text{RSS} = \boldsymbol{\varepsilon}'(\mathbf{Q} - \mathbf{H}_c)\boldsymbol{\varepsilon}$. By theorem 1.7 of Seber (1977) and the assumption that $\mathbf{Q}$ is independent of $\boldsymbol{\varepsilon}$,

$$\begin{aligned} E(\text{RSS}) &= E[\boldsymbol{\varepsilon}'(\mathbf{Q} - \mathbf{H}_c)\boldsymbol{\varepsilon}] = \text{tr}[(\mathbf{Q} - \mathbf{H}_c)\text{Var}(\boldsymbol{\varepsilon})] \\ &\quad + [(E\boldsymbol{\varepsilon})'(\mathbf{Q} - \mathbf{H}_c)(E\boldsymbol{\varepsilon})] = \sigma^2[\text{tr}\,\mathbf{Q} - \text{tr}\,\mathbf{H}_c] \\ &= \sigma^2\left(\sum Q_i - 3\right) \end{aligned}$$

since $\text{tr}[\mathbf{X}_c(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'] = \text{tr}[(\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{X}_c] = 3$.

According to Theorem 2.1, for the method of complete-case analysis to yield unbiased estimates of the regression coefficients and error variance, the missing-data mechanism can depend on the values of the covariates (i.e., Q_i can be a function of X_{1i} and X_{2i}) but not on the error term. Two common ways in which the missing-data mechanism may depend on the error term are through the existence of an omitted covariate and through mismeasured covariates. Suppose in (1) that $\varepsilon_i = \gamma X_{3i} + \varepsilon_i^*$, where ε_i^* has the usual properties of an error term and X_3 is the omitted covariate. For example, in a forward stepwise regression procedure, the second stage of analysis may only include X_1 and X_2. If X_3 is orthogonal to $(1, X_1, X_2)$ and there is no missing data ($\mathbf{Q} = \mathbf{I}_n$), then by (3), estimation is unbiased. But if data are missing as a function of either X_1, X_2, or X_3, then the bias is

$$(\mathbf{X}'\mathbf{QX})^{-1}\gamma \begin{pmatrix} \sum X_{3i}Q_i \\ \sum X_{1i}X_{3i}Q_i \\ \sum X_{2i}X_{3i}Q_i \end{pmatrix}, \tag{4}$$

which could be quite different from the zero vector.

The other case involves mismeasured covariates. Suppose that X_{1i} is a mismeasurement of the true covariate S_{1i} so that in (1), $\varepsilon_i = \beta_1(S_{1i} - X_{1i}) + e_i^*$, where e_i^* has the usual properties of an error term. Then, regardless of the type of missing-data pattern (including no missing data at all), the estimation of $\boldsymbol{\beta}$ can be biased. The resultant bias is given by (4) with β_1 replacing γ and $S_{1i} - X_{1i}$ replacing X_{3i}.

3. MISSING-INDICATOR METHODS

As defined in Section 1, these methods modify the original model by adding a missingness indicator and possibly interactions between this indicator and the covariates. In the first two methods discussed, the true model is of the form (1), and only X_2 is liable to be missing. The third method involves adding a missing indicator to simple linear regression. The two-sample problem with missing-group information is considered as a special case.

3.1 Missing-Indicator Methods I and II

The true model is assumed to be given by (1); however, X_2 is not always observed. Let Q_{2i} equal 1 when X_{2i} is observed and zero when it is missing. The first missing-indicator method, as described by Anderson et al. (1983, p. 456), Chow (1979), Cohen and Cohen (1975, secs. 7.4.3 and 9.3.5), and Mietinnen (1985, sec. 18.2.5), is

$$Y_i = \gamma_0 + \gamma_1 X_{1i} + \gamma_2 X_{2i} Q_{2i} + \gamma_3(1 - Q_{2i}) + e_i. \tag{5}$$

The original model (1) is modified by replacing X_{2i} by $X_{2i}Q_{2i}$ and adding the missing indicator $1 - Q_{2i}$. This model is commonly used to test whether the data are missing at random by checking if γ_3 is significantly different from zero. Furthermore, regardless of the test result, the estimates of γ_0, γ_1, and γ_2 are used. A slight modification of (5) that is mathematically easier to work with is

$$Y_i = \theta_0 Q_{2i} + \theta_1 X_{1i} + \theta_2 X_{2i} Q_{2i} + \theta_3(1 - Q_{2i}) + e_i. \quad (6)$$

The least squares estimators of $(\gamma_0, \gamma_1, \gamma_2)$ are identical to those of $(\theta_0, \theta_1, \theta_2)$. The least squares estimators are biased for $(\beta_0, \beta_1, \beta_2)$, as stated in the following theorem.

Theorem 3.1. If ε is independent of (X_1, X_2, Q_2), then, conditional on (X_1, X_2, Q_2), the expected least squares estimators for model (6) are

$$E(\hat{\theta}_0) = \beta_0 + \beta_2 P_m S_{12}^m F_0,$$

$$E(\hat{\theta}_1) = \beta_1 + \beta_2 P_m S_{12}^m F_1,$$

and

$$E(\hat{\theta}_2) = \beta_2(1 - P_m S_{12}^m F_2),$$

where $P_m = 1 - \bar{Q}_2$, the proportion missing X_2; S_{12}^m is the sample covariance of X_1 and X_2 for those missing X_2; and (F_0, F_1, F_2) are functions of the means, variances, and covariances of X_1 and X_2 and are defined in the proof.

Note that the estimators are unbiased if either $P_m = 0$ or $S_{12}^m = 0$. The proof is given in the Appendix.

By writing down model (6) for the subsets with and without X_2 information, one can gain an intuitive understanding why this model produces biased estimators. For those with X_2 information $(Q_{2i} = 1)$, model (6) is

$$Y_i = \theta_0 + \theta_1 X_{1i} + \theta_2 X_{2i} = e_i, \quad (7)$$

whereas for those missing X_2 $(Q_{2i} = 0)$, it is

$$Y_i = \theta_3 + \theta_1 X_{1i} + e_i. \quad (8)$$

The intercept terms are different, but the X_1 coefficients are required to be the same, regardless of the adjustment value of X_2. This causes the bias. A simple way around this constraint is to add another term to the model, which allows the X_1 coefficients for the subsets to differ, so that for those subjects missing X_2,

$$Y_i = \theta_3 + \theta_4 X_{1i} + e_i. \quad (9)$$

This new model is described next.

The second-missing indicator method is to use the model

$$Y_i = \theta_0 Q_{2i} + \theta_1 X_{1i} Q_{2i} + \theta_2 X_{2i} Q_{2i} + \theta_3(1 - Q_{2i}) + \theta_4 X_{1i}(1 - Q_{2i}) + e_i. \quad (10)$$

This generalizes the first missing-indicator method by splitting the X_{1i} term of model (6) into two predictor variables with corresponding coefficients θ_1 and θ_4. The subset models are given by (7) and (9). Model (10) is only a slight variation on that proposed by Chow (1979). The modification was made to facilitate the proof of the following theorem.

Theorem 3.2. Assume that ε is independent of (X_1, X_2, Q_2). Then, conditional on (X_1, X_2, Q_2), the expected least squares estimators for model (10) are $E(\hat{\theta}_0) = \beta_0, E(\hat{\theta}_1) = \beta_1, E(\hat{\theta}_2) = \beta_2, E(\hat{\theta}_3) = \beta_0 + \beta_2 \hat{\beta}_{X_2|1}^m$, and $E(\hat{\theta}_4) = \beta_1 + \beta_2 \hat{\beta}_{X_2|X_1}^m$, where $\hat{\beta}_{X_2|1}^m$ and $\hat{\beta}_{X_2|X_1}^m$ are the least squares intercept and slope estimators from the regression of X_2 on X_1 for those missing X_2. Furthermore, conditional on (X_1, X_2, Q_2), the expected mean squared error for model (10) is

$$\sigma^2 + \frac{1}{n-5} [\mathrm{RSS}^m(X_2|X_1)]\beta_2^2,$$

where $\mathrm{RSS}^m(X_2|X_1)$ is the residual sum of squares from regressing X_2 on X_1 for the subset of individuals missing X_2.

Note that $(\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2)$ are unbiased estimators of $(\beta_0, \beta_1, \beta_2)$, whereas $(\hat{\theta}_3, \hat{\theta}_4)$ are biased for (β_0, β_1); if X_1 and X_2 are uncorrelated among those missing X_2, then $\hat{\beta}_{X_2|X_1}^m = 0$, in which case $\hat{\theta}_4$ is unbiased for β_1. Rewriting (10) for those with X_2 information and those without gives the subset models (7) and (9), which contain no common regression coefficients. As such, model (7) is the complete-case analysis, which, by Theorem 2.1, allows unbiased estimation, whereas model (9) is incorrectly specified if Y is truly modeled by (1). Overestimation of σ^2 is also not surprising because of the assumption that $\mathrm{var}(e_i) = \sigma^2$ in (10) regardless of whether Q_{2i} is zero or 1. In particular, as seen by comparing (7) and (9), the contribution to the residual sum of squares for those missing X_2 will be larger than for those with X_2, unless $\beta_2 = 0$. The residual sum of squares (and hence the bias for σ^2) is largest when X_1 and X_2 are uncorrelated for those subjects missing X_2.

Because model (9) excludes the predictor X_2, it will be helpful to review the effect of omitting a predictor variable before proving Theorem 3.2. That proof is found in the Appendix. The following lemma is from section 6.1.1 of Seber (1977).

Lemma 3.1. Suppose that the true model is given by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, where $E(\boldsymbol{\varepsilon}) = \mathbf{0}, E(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors. Furthermore, suppose that $\mathbf{Z}$ is omitted from the model fit. Then

$$E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\boldsymbol{\gamma}.$$

Given $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathrm{MSE} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{H})\mathbf{Y}/(n-p)$, where $p = \dim(\mathbf{X})$, then

$$E(\mathrm{MSE}) = \sigma^2 + \frac{\boldsymbol{\gamma}'\mathbf{Z}'(\mathbf{I}_n - \mathbf{H})\mathbf{Z}\boldsymbol{\gamma}}{n-p} > \sigma^2.$$

The estimated variance of $\hat{\boldsymbol{\beta}}$ is $\mathrm{MSE}(\mathbf{X}'\mathbf{X})^{-1}$, which is larger on average than $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

3.2 Missing-Indicator Method III and the Missing-Group Method

Cohen and Cohen (1975, chap. 7) described a missing-indicator method for simple linear regression similar to

those described earlier. Assuming that the true model is

$$Y_i = \beta_0 + \beta_1 Z_i + \varepsilon_i,$$

they recommended the data analyst use the model

$$Y_i = \gamma_0 + \gamma_1[Z_i Q_i + c(1 - Q_i)] + \gamma_2(1 - Q_i) + e_i, \quad (11)$$

where c is an arbitrary constant and Q_i takes the value 1 when Z_i is measured and zero when Z_i is missing.

Cohen and Cohen (1975) correctly noted that the estimation of γ_0 and γ_1 is independent of the choice of the constant c.

Theorem 3.3. Assume that ε is independent of (Z, Q). Then, conditional on (Z, Q), the expected least squares estimators for model (11) are $E(\hat{\gamma}_0) = \beta_0, E(\hat{\gamma}_1) = \beta_1$ and $E(\hat{\gamma}_2) = \beta_1(\bar{Z}^m - c)$, where $\bar{Z}^m$ is the average of the Z's among those missing Z. Moreover, $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are exactly the least squares estimators from the complete-case analysis. Conditional on (Z, Q), the expected mean squared error (MSE) from fitting (11) is $\sigma^2 + \beta_1^2 \text{MSE}^*$, where MSE^* is the mean squared error from fitting the model

$$(1 - Q_i)(Z_i - c) = \eta_0 + \eta_1\{Z_i Q_i + c(1 - Q_i)\} + \eta_2(1 - Q_i) + e_i.$$

The proof of Theorem 3.3 is given in the Appendix.

This missing-indicator methods can be further illustrated by a special case, called the missing-group method. Often in analysis of variance, group membership is unknown for some individuals. One method from the statistical folklore for dealing with this type of missing data is to form another group and then to perform the analysis of variance on the augmented set of groups. Cohen and Cohen (1975, chap. 7) recommended this approach. To investigate possible bias in this procedure, the two-sample problem is considered here for simplicity. The true model is assumed to be

$$Y_i = \mu_0 + \alpha_0 Z_i + \varepsilon_i,$$

where Z_i assumes the value zero for group 1 and 1 for group 2 and ε_i has mean zero and variance σ^2. Again let Q_i be 1 if Z_i is observed and zero if not. The third missing-indicator method, described previously, uses the model

$$Y_i = \mu + \alpha Z_i Q_i + \theta(1 - Q_i) + e_i, \quad (12)$$

where the arbitrary constant c of the third-missing indicator model is chosen here to be zero. An artificial group 3 is thereby created to consist of those subjects with missing-group identification. The three group means are $\mu, \mu + \alpha$, and $\mu + \theta$.

Corollary 3.1. Assume that ε is independent of (Z, Q). Then, conditional on (Z, Q), the expected least squares estimators for model (12) are $E(\hat{\mu}) = \mu_0, E(\hat{\alpha}) = \alpha_0$, and $E(\hat{\theta}) = \alpha_0$ times the proportion of group 2 subjects among those with missing-group identification. But conditional on (Z, Q), the expected MSE from this analysis is

$$\sigma^2 + \frac{m_1 m_2}{(m_1 + m_2)(n - 3)} \alpha_0^2,$$

where m_j is the number missing from group $j, j = 1, 2$.

The proof is given in the Appendix. This MSE bias would affect the standard t test for the difference between the two nonmissing groups

$$t = (\bar{Y}_1 - \bar{Y}_2) \Big/ \sqrt{\text{MSE}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

The complete-case analysis, which does not include the missing group, uses a smaller, unbiased estimator of σ^2, but its t test is based on fewer degrees of freedom. Naturally, any comparison of t tests based on these two methods depends on the values of m_1, m_2, n, σ^2, and α_0. Two examples illustrate this. Suppose that the two groups are of equal size, each with half their observations missing, so that $m_1 = m_2 = n/4$ and $\sigma^2 = 1, \alpha_0 = 2$. The MSE bias of the missing-group method is roughly 0.5, which would produce a less powerful test than the complete-case analysis for the usual α levels at moderate degrees of freedom. On the other hand, suppose that one group has no missing observations, perhaps $m_1 = 0$. Then the MSE bias is zero, and the missing-group method produces the larger critical region. In practice, though, m_1 and m_2 are unknown, and caution would suggest using the complete-case method.

4. STRATIFICATION METHODS

Another class of methods for handling missing data fall under the category of stratification methods because they involve stratifying the data set into smaller pieces for analysis. Once again, suppose that the true model for Y is linear in the predictor variables X_1 and X_2, given by (1), and that X_1 is always observed but X_2 is sometimes missing. Furthermore, suppose that X_2 can only take on values in $\{c_1, \ldots, c_k\}$, with at least two observations per category. There are several possible scenarios for X_1 and X_2. Consider here the case in which X_1 is the predictor of interest and X_2 is an important confounder but β_2 is considered a nuisance parameter. For example, X_1 might be a treatment indicator or dosage level, and X_2 the age category or sex of a subject. In general a stratification method would control for the confounder X_2 by creating k strata corresponding to the k subsets $\{\{i | X_{2i} = c_j\}, j = 1, \ldots, k\}$ and then by modeling Y in terms of X_1 in each stratum assuming a constant β_1 across strata. In such a stratified setting, the true model (1) for subjects in stratum j is $Y_i = (\beta_0 + \beta_2 c_j) + \beta_1 X_{1i} + \varepsilon_i$. Data analysts sometimes modify this method for missing-data problems by adding a stratum to include all those with unobserved X_2 information. In particular, define the stratum indicator I_{ji} to be 1 when X_{2i} is observed to be c_j and zero otherwise for $j = 1, \ldots, k$, and let $I_{k+1,i}$ be 1 if X_{2i} is missing and zero otherwise. The resulting model is

$$Y_i = \sum_{j=1}^{k+1} \gamma_{0j} I_{ji} + \gamma_1 X_{1i} + e_i. \quad (13)$$

Theorem 4.1. The least squares estimators for model (13) are

$$\hat{\gamma}_{0j} = \beta_0 + (c_j - \bar{X}_{1(j)} P_m S_{12}^m / S_{11}^o)\beta_2 + \sum I_{ji}\varepsilon_i \quad (j = 1, \ldots, k),$$

$$\hat{\gamma}_{0,k+1} = \beta_0 + (\bar{X}_{2(k+1)} - \bar{X}_{1(k+1)} P_m S_{12}^m / S_{11}^o)\beta_2 + \sum I_{k+1,i}\varepsilon_i,$$

and

$$\hat{\gamma}_1 = \beta_1 + (P_m S_{12}^m / S_{11}^o)\beta_2 + \sum X_{1i}\varepsilon_i,$$

where $\bar{X}_{1(j)}$ is the average X_1 within the jth stratum, P_m is the proportion of individuals in stratum $k+1$ (unknown X_2), S_{12}^m is the sample covariance of X_1 and X_2 for those missing X_2, $S_{11}^o = n^{-1}\sum\sum I_{ji}(X_{1i} - \bar{X}_{1(j)})^2$, and ε_i is the error term from the true model (1).

The proof is somewhat tedious but straightforward and is omitted here. Of particular interest is that even if ε_i is independent of X_{1i} and the I_{ji}'s, $\hat{\gamma}_1$ is a biased estimator of β_1 unless $P_m = 0$ or $S_{12}^m = 0$. The flavor of this result is reminiscent of the first missing-indicator method of Section 3, because the β_1 coefficient is assumed to be the same in the "missing-data" stratum (in which X_2 may be heterogeneous and thus S_{12}^m nonzero) as in the other strata (in which X_2 is homogeneous). Adding a different β_1 parameter for the "missing-data" stratum gets around this constraint. Two alternative stratification methods that allow more flexibility through additional stratum-specific parameters are

$$Y_i = \sum_{j=1}^{k+1} (\gamma_{0j} + \gamma_{1j} X_{1i}) I_{ji} + e_i \tag{14}$$

and

$$Y_i = \sum_{j=1}^{k} (\gamma_{0j} + \gamma_1 X_{1i}) I_{ji} + (\gamma_{0,k+1} + \gamma_{1,k+1} X_{1i}) I_{k+1,i} + e_i. \tag{15}$$

Theorem 4.2. Assume that ε_i is independent of X_{1i} and the I_{ji}'s. The least squares estimators $(\hat{\gamma}_{0j}, \hat{\gamma}_{1j})$ for model (14) and $(\hat{\gamma}_{0j}, \hat{\gamma}_1)$ for model (15) are unbiased for $(\beta_0 + \beta_2 c_j, \beta_1)$ for $j = 1, \ldots, k$. In stratum $k+1$, conditional on (X_1, X_2, Q), $(\hat{\gamma}_{0,k+1}, \hat{\gamma}_{1,k+1})$ are unbiased for $(\beta_0 + \beta_2(\bar{X}_{2(k+1)} - \hat{\beta}_{X_2|X_1}^m \bar{X}_{1(k+1)}), \beta_1 + \beta_2 \hat{\beta}_{X_2|X_1}^m)$, where $\hat{\beta}_{X_2|X_1}^m$ is the least squares slope estimator from regression of X_2 on X_1 in stratum $k+1$. Conditional on (X_1, X_2, Q), the expected MSE's for these models are

$$E[\text{MSE}(14)] = \sigma^2 + \beta_2^2[\text{RSS}^m(X_2|X_1)]/(n - 2(k+1))$$

and

$$E[\text{MSE}(15)] = \sigma^2 + \beta_2^2[\text{RSS}^m(X_2|X_1)]/(n - k - 3),$$

where $\text{RSS}^m(X_2|X_1)$ is the residual sum of squares after regressing X_2 on X_1 for those missing X_2 (stratum $k+1$).

The proof is given in the Appendix.

Table 1. Asymptotic Biases of β_1, β_2, and σ^2

σ_2	ρ_{12}	MIM I bias(β_1)	MIM I bias(β_2)	MIM II bias(σ^2)
1	0	0	0	$.50\beta_2^2$
	.5	$.29\beta_2$	$-.14\beta_2$	$.37\beta_2^2$
	.9	$.76\beta_2$	$-.68\beta_2$	$.10\beta_2^2$
2	0	0	0	$2.00\beta_2^2$
	.5	$.57\beta_2$	$-.14\beta_2$	$1.50\beta_2^2$
	.9	$1.51\beta_2$	$-.68\beta_2$	$.38\beta_2^2$
5	0	0	0	$12.50\beta_2^2$
	.5	$1.43\beta_2$	$-.14\beta_2$	$9.37\beta_2^2$
	.9	$3.78\beta_2$	$-.68\beta_2$	$2.37\beta_2^2$

5. EVALUATION OF THE MAGNITUDE OF BIAS

Proper assessment of whether the missing-covariate methods should be used in practice requires an evaluation of the magnitude of these methods' biases for $\boldsymbol{\beta}$ and σ^2. The focus here is on missing-indicator methods I and II, modeled in (6) and (10). The ultimate question of interest is whether either of these methods should be preferred over the complete-case analysis, based on model (2).

Once again, the true model is assumed to be (1), where X_2 may be missing for some individuals. The bias in estimating $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ by missing-indicator method I (MIM I) is given in Theorem 3.1. The biases in estimating $\boldsymbol{\beta}$ and σ^2 by missing-indicator method II (MIM II) are given in Theorem 3.2. Although the bias in estimating σ^2 by MIM I is not derived herein, it is at least as large as that for MIM II, because the MIM I model (6) is restricted relative to the MIM II model (10). The evaluation study consists of two parts. In Section 5.1 no particular distribution for (X_1, X_2) is assumed, but the measurement indicator Q_2 is assumed to be independent of X_1, X_2, and ε. That is, X_2 is missing completely at random. In Section 5.2 (X_1, X_2) are assumed to be bivariate Bernoulli and Q_2 may be missing as a function of X_1 and/or X_2. This allows a more complete study by pattern of missingness. Bias is evaluated by asymptotic theory and computer simulation.

5.1 General Covariate Distribution

In this part of the evaluation study, the covariates X_1 and X_2 are assumed to have variances σ_1^2 and σ_2^2 and correlation ρ_{12}. Furthermore, Q_2 is assumed to be independent of X_1, X_2, and ε. Let $p_m = P(Q_2 = 0)$ be the proportion missing. Replacing terms in the bias expressions in Theorem 3.1 by their almost sure limits, the asymptotic biases in MIM I for estimating β_1 and β_2 become

$$\text{bias}_n(\beta_1) \xrightarrow{\text{as}} \beta_2 \frac{p_m \sigma_2 \rho_{12}}{\sigma_1[1 - \rho_{12}^2(1 - p_m)]}, \tag{16}$$

and

$$\text{bias}_n(\beta_2) \xrightarrow{\text{as}} -\beta_2 \frac{p_m \rho_{12}^2}{[1 - \rho_{12}^2(1 - \rho_m)]}. \tag{17}$$

As expected, these biases are zero when either $\beta_2 = 0$ or $p_m = 0$ or $\rho_{12} = 0$. When $p_m \neq 0$, bias(β_1) ranges from $-\beta_2(\sigma_2/\sigma_1)$ when $\rho_{12} = -1$ to $\beta_2(\sigma_2/\sigma_1)$ when $\rho_{12} = 1$. Obviously, the larger the ratio σ_2/σ_1, the larger this bias

Table 2. Averages and Standard Deviations Over Simulated Data Sets

Analysis	Ave($\hat{\beta}_1$)	SD($\hat{\beta}_1$)	Ave(se($\hat{\beta}_1$))	Ave($\hat{\beta}_2$)	SD($\hat{\beta}_2$)	Ave(se($\hat{\beta}_2$))	Ave($\hat{\sigma}^2$)
			$\rho_{12} = 0$				
All data	1.002	.072	.071	2.003	.035	.035	.991
Complete-case	1.006	.100	.101	1.999	.050	.051	.988
MIM I	1.001	.212	.215	1.999	.051	.154	9.115
MIM II	1.006	.100	.307	1.999	.050	.154	9.115
			$\rho_{12} = .5$				
All data	.996	.078	.082	2.001	.041	.041	1.002
Complete-case	.994	.115	.118	2.002	.060	.059	.995
MIM I	2.128	.247	.214	1.717	.095	.152	7.858
MIM II	.994	.115	.313	2.002	.060	.156	7.038

becomes. When $p_m \neq 0$, the range of bias(β_2) is from zero when $\rho_{12} = 0$ to $-\beta_2$ when $\rho_{12}^2 = 1$.

The regression-coefficient estimators for MIM II are the same as those of the complete-case analysis and thus are unbiased. The MIM II estimator of σ^2 is not unbiased. In fact, from Theorem 3.2, the MSE from MIM II can be shown to converge almost surely to $\sigma^2 + \beta_2^2 p_m \sigma_2^2 (1 - \rho_{12}^2)$. This bias is very large when σ_2^2 is large and ρ_{12} is zero. Table 1 summarizes the MIM I asymptotic bias in (β_1, β_2) and the MIM II asymptotic bias in σ^2 when $p_m = .5$ and $\sigma_1^2 = 1$. As already mentioned, the MIM I estimate of σ^2 is greater than or equal to that of MIM II. Settings that produce the largest (smallest) bias in the regression coefficient estimation produce the smallest (largest) bias in residual variance estimation.

It is also of interest to consider the standard errors of the MIM regression coefficient estimators and how they compare to those of the complete-case analysis. Using the GAUSS matrix language (Aptech Systems, Inc. 1991), simulation was used to answer this question. Two bivariate normal covariates were generated with zero means; $\sigma_1^2 = 1$ and $\sigma_2^2 = 4$; $\rho_{12} = 0$ in the first simulation and .5 in the second. In the assumed true model (1), $\beta_0 = 1, \beta_1 = 1, \beta_2 = 2, \sigma^2 = 1$. Also, $p_m = .5$. The results in Table 2 are based on 500 simulated data sets, each with $n = 200$. The standard deviations of the simulated estimators are not comparable to the averages of the simulated standard errors for the missing-indicator methods. The MSE's of $\hat{\beta}_1$ and $\hat{\beta}_2$ for MIM I and MIM II far exceed those of the complete-case analysis.

5.2 Various Patterns of Missingness

In the previous section X_2 are missing completely at random. Now the various possible patterns of X_2 missing and their effect on bias are investigated for bivariate Bernoulli covariates. Let $p_{jk} = P(X_1 = j, X_2 = k)$ for $j, k = 1, 2$ be the joint frequency function of (X_1, X_2). Also let $q_{jk} = P(Q_2 = 1 | X_1 = j, X_2 = k)$ be the conditional frequency function of Q_2. As before, Q_2 is independent of ε. There are four possible patterns of missing X_2 data:

P1. X_2 missing completely at random: $P(Q_2 = 1 | X_1, X_2) = P(Q_2 = 1)$

P2. X_2 missing as a function of X_1: $P(Q_2 = 1 | X_1, X_2) = P(Q_2 = 1 | X_1)$

P3. X_2 missing as a function of X_2: $P(Q_2 = 1 | X_1, X_2) = P(Q_2 = 1 | X_2)$

P4. X_2 missing as a function of X_1 and X_2.

First, the bias of the MIM I regression coefficient estimators is considered. Using Theorem 3.1, the asymptotic biases are zero if the limiting covariance between X_1 and X_2 for subjects with X_2 missing is zero; that is, if $\sigma_{12}^m = 0$. It can be shown that $\sigma_{12}^m = 0$ if and only if

$$\frac{p_{00}p_{11}}{p_{10}p_{01}} \frac{(1 - q_{00})(1 - q_{11})}{(1 - q_{10})(1 - q_{01})} = 1.$$

Let A represent the first term and B the second. Then $A = 1$ if and only if X_1 and X_2 are independent. In missing-data patterns P1–P3, $B = 1$, but in P4, $B \neq 1$. Hence even if the covariates are independent, MIM I will give biased estimates if X_2 is missing as a function of both X_1 and X_2. Two pattern P4 examples illustrate this bias. First, suppose that X_2 is missing whenever $X_1 \neq X_2$; that is, $q_{00} = q_{11} = 1, q_{01} = q_{10} = 0$. Then $\text{bias}_n(\beta_1) \xrightarrow{\text{as}} -\beta_2$ and $\text{bias}_n(\beta_2) \xrightarrow{\text{as}} \beta_2$. Second, suppose that X_2 is missing whenever $X_1 = X_2$; that is, $q_{00} = q_{11} = 0, q_{01} = q_{10} = 1$. Then $\text{bias}_n(\beta_1) \xrightarrow{\text{as}} \beta_2$ and $\text{bias}_n(\beta_2) \xrightarrow{\text{as}} \beta_2$.

Finally, the properties of the missing-indicator methods' estimators can be compared to those of the complete-case analysis through computer simulation. The analysis of all data is given as the baseline standard of comparison. As before, 500 simulated data sets of size $n = 200$ each were

Table 3. Simulation Results When X_2 is Missing Completely at Random

Analysis	Ave($\hat{\beta}_1$)	SD($\hat{\beta}_1$)	Ave(se($\hat{\beta}_1$))	Ave($\hat{\beta}_2$)	SD($\hat{\beta}_2$)	Ave(se($\hat{\beta}_2$))	Ave($\hat{\sigma}^2$)
All data	.997	.139	.142	2.001	.145	.142	.994
Complete-case	1.004	.198	.202	1.994	.201	.202	.998
MIM I	.996	.174	.174	1.995	.201	.248	1.503
MIM II	1.004	.198	.249	1.994	.201	.249	1.503

Table 4. Simulation Results When X_2 is Missing as a Function of X_1 and X_2

Analysis	Ave($\hat{\beta}_1$)	SD($\hat{\beta}_1$)	Ave(se($\hat{\beta}_1$))	Ave($\hat{\beta}_2$)	SD($\hat{\beta}_2$)	Ave(se($\hat{\beta}_2$))	Ave($\hat{\sigma}^2$)
All data	.990	.135	.142	2.006	.131	.142	.996
Complete-case	.989	.167	.174	2.008	.169	.174	.996
MIM I	.443	.159	.162	2.195	.178	.186	1.177
MIM II	.989	.167	.175	2.008	.169	.175	.996

generated for model (1) with $\beta_0 = \beta_1 = 1, \beta_2 = 2, \sigma^2 = 1$, and $p_{jk} \equiv .25$. Note that X_1 and X_2 are independent. Table 3 summarizes the results when X_2 is missing completely at random and $P(Q_2 = 1) = .5$. Table 4 summarizes the results when X_2 is missing as a function of X_1 and X_2; in particular, X_2 is present whenever $X_1 = X_2$ but is missing with probability .5 whenever $X_1 \neq X_2$.

In the setting of Table 3, MIM I is unbiased for β_1 and has a lower estimated standard error of $\hat{\beta}_1$ on average than does the complete-case analysis. This advantage disappears when X_1 and X_2 are correlated or when X_2 is missing as a function of both X_1 and X_2, as shown in Table 4 where the bias in estimating β_1 is unacceptable. Knowledge of why X_2 is missing and of the correlation among covariates when X_2 is missing is essential when using MIM I for $\hat{\boldsymbol{\beta}}$. MIM I overestimates the standard error of $\hat{\beta}_2$. Use of MIM I is generally ill-advised. But some special cases in which a covariate is missing by design due to cost considerations should be investigated. MIM II shows no advantage over the complete-case analysis.

6. DISCUSSION

This article has investigated the possible bias in the estimators of the regression coefficients and residual variance derived from the missing-indicator and stratification methods of handling missing data. In particular, the true regression relationship between Y and (X_1, X_2) is assumed to be given in (1), in which X_2 can be missing as a function of X_1 and/or X_2 but not as a function of ε. Hence, given X_1 and X_2, missingness of X_2 is conditionally independent of Y. The missing-data methods studied in this article include (a) complete-case analysis, modeled in (2), with biases in Theorem 2.1; (b) missing-indicator method (MIM) I (6) with biases in Theorem 3.1; (c) MIM II (10) with biases in Theorem 3.1; (d) stratification method I (13), with biases in Theorem 4.1; and (e) stratification methods II and III, modeled in (14) and (15), with biases given in Theorem 4.2. In summary, the complete-case analysis is valid so long as the covariates are independent of ε; they need not be missing completely at random. MIM I and stratification method I produce biased regression parameter estimators. MIM's II and III and stratification methods II and III give the same regression estimators as the complete-case analysis and thus are unbiased. But they each overestimate the residual variance. The magnitude of these biases was studied further in Section 5 via asymptotic theory and computer simulation. In Section 5.1, where X_2 is missing completely at random, the biases in estimating β_1, β_2, and σ^2 by MIM I were seen to be arbitrarily large, depending on the percentage of missing data, the value of β_2, the ratio of X variances, and the correlation between X_1 and X_2. Various patterns of missing data were investigated in Section 5.2 for the case of binary covariates. As discussed there, MIM I is not advised as a general method. MIM II produces the same regression parameter estimates as the complete-case analysis, but often with considerably larger standard errors.

In Section 3.2 the true model was assumed to contain only a single explanatory variable. MIM III, modeled in (11), was found to have biases given in Theorem 3.3, and for the special case of the so-called missing-group method (12), biases were given in Corollary 3.1. The missing-group method overestimates the residual variance. Hence the addition of a missing group can weaken the power of the t test that compares two of the nonmissing groups. On the other hand, the complete-case analysis uses an unbiased estimator of the residual variance, but its t test is based on fewer degrees of freedom. As discussed at the end of Section 3, the complete-case analysis is generally preferable.

Of those researchers suggesting the missing-data methods studied in this article, Cohen and Cohen (1975) recognized that the residual variance may be overestimated, as they stated that the power of the regression analysis may be weakened if one truly knows that the data are missing at random. By "missing at random" they mean that X_2 is not missing as a function of $(Y, X_1, X_2, \varepsilon)$. This corresponds to the "missing completely at random" definition given by Little and Rubin (1987). In general, Cohen and Cohen (1975) recommended against such an assumption; however, they did not recognize that MIM I will produce biased regression coefficient estimators, even when the assumption is correct.

APPENDIX: PROOFS OF THEOREMS

With the exception of Theorem 2.1, the proofs of the theorems are given in this Appendix.

Proof of Theorem 3.1

This proof is very lengthy and tedious, so only a sketch of it is given. Define $\mathbf{Y}' = (Y_1, \ldots, Y_n), \boldsymbol{\varepsilon}' = (\varepsilon_1, \ldots, \varepsilon_n)$,

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{21} \\ \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} \end{pmatrix},$$

and

$$\mathbf{X}_I = \begin{pmatrix} Q_{21} & X_{11} & X_{21}Q_{21} & 1 - Q_{21} \\ \vdots & \vdots & \vdots & \vdots \\ Q_{2n} & X_{1n} & X_{2n}Q_{2n} & 1 - Q_{2n} \end{pmatrix}.$$

Next, some definitions are necessary. Define $\bar{Q}_2 = \sum Q_{2i}/n$. For $j, k \in \{1, 2\}$, let

$$\bar{X}_j^c = \sum Q_{2i} X_{ji}/n\bar{Q}_2,$$

$$\bar{X}_j^m = \sum (1 - Q_{2i}) X_{ji}/n(1 - \bar{Q}_2),$$

$$S_{jk}^c = \sum Q_{2i}(X_{ji} - \bar{X}_j^c)(X_{ki} - \bar{X}_k^c)/n\bar{Q}_2,$$

$$S_{jk}^m = \sum (1 - Q_{2i})(X_{ji} - \bar{X}_j^m)(X_{ki} - \bar{X}_k^m)/n(1 - \bar{Q}_2),$$

and

$$r_{12}^c = S_{12}^c/\sqrt{S_{11}^c S_{22}^c}.$$

These terms represent the sample means, variances, and covariances for those with and without X_2 information. Because

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I'\mathbf{Y} = (\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}),$$

and because, by assumption,

$$E[(\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I'\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{Q}] = (\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I'E(\boldsymbol{\varepsilon}) = \mathbf{0},$$

the proof consists of showing that $(\mathbf{X}_I'\mathbf{X}_I)^{-1}\mathbf{X}_I'\mathbf{X}\boldsymbol{\beta}$ is given as in Theorem 3.1 with

$$F_0 = (S_{12}^c\bar{X}_2^c - S_{22}^c\bar{X}_1^c)/D,$$

$$F_1 = S_{22}^c/D,$$

and

$$F_2 = S_{12}^c/D,$$

where $D = (1 - \bar{Q}_2)S_{11}^m S_{22}^c + \bar{Q}_2 S_{11}^c S_{22}^c[1 - (r_{12}^c)^2]$. The expectation of $\hat{\theta}_3$ is $\beta_0 + \beta_2$ times a very messy expression, which is omitted here.

Proof of Theorem 3.2

As already stated, MIM II, modeled by (10), can be rewritten as submodels (7) and (9) for the subsets of individuals with and without X_2 information. For the purpose of finding the bias in the predictor coefficients, it is sufficient to consider the submodels separately, because the coefficients are different between the two submodels. Model (7) is in fact a complete-case analysis for that subset, and thus by Theorem 2.1, $E(\hat{\theta}_0) = \beta_0, E(\hat{\theta}_1) = \beta_1$, and $E(\hat{\theta}_2) = \beta_2$. Model (9) underfits the true model by omitting X_2. By Lemma 3.1,

$$\begin{pmatrix} \hat{\theta}_3 \\ \hat{\theta}_4 \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{X}_2\beta_2,$$

where $\mathbf{X}^*$ is the $n \times 2$ design matrix $(1, X_1)$. The latter term is β_2 times the least squares intercept and slope estimators from the regression of X_2 on X_1.

Model (10) assumes the $\text{var}(e_i) = \sigma^2$, regardless of whether an individual possesses or is missing X_2 information. The RSS for model (10) can be split into two parts, one for each submodel. Because submodel (7) is a complete-case analysis for that subset, then by Theorem 2.1, the RSS for (7) in the subset with X_2 data is unbiased for $[(\sum Q_{2i}) - 3]\sigma^2$. Because submodel (9) omits the predictor X_2, then by Lemma 3.1, the RSS for (9) in the subset missing X_2 is $[(n - \sum Q_{2i}) - 2]\sigma^2 + \beta_2^2\mathbf{X}_2'(\mathbf{I}_n - \mathbf{H}^*)\mathbf{X}_2$, where $\mathbf{H}^* = \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}$. Note that $\mathbf{X}_2'(\mathbf{I}_n - \mathbf{H}^*)\mathbf{X}_2$ is the RSS from regressing X_2 on X_1 in the subset missing X_2. Hence the RSS for the entire data set is unbiased for

$$(n - 5)\sigma^2 + \beta_2^2\text{RSS}^m(X_2|X_1).$$

Proof of Theorem 3.3

The data analyst–assumed models for the subsets of the data with and without measured Z_i values are

$$Y_i = \gamma_0 + \gamma_1 Z_i + e_i \qquad (Q_i = 1)$$

and

$$Y_i = \gamma_0 + \gamma_1 c + \gamma_2 + e_i \qquad (Q_i = 0).$$

The least squares estimators $\hat{\gamma}_0$ and $\hat{\gamma}_1$ are derived completely from the complete-case subset of the data, while

$$\hat{\gamma}_2 = \bar{Y}^m - \hat{\gamma}_0 - \hat{\gamma}_1 c,$$

where $\bar{Y}^m$ is the average Y value in the missing-data subset. Conditional on $\mathbf{Z}$ and $\mathbf{Q}, E\bar{Y}^m = \beta_0 + \beta_1\bar{Z}^m$, and because $(\hat{\gamma}_0, \hat{\gamma}_1)$ are unbiased complete-case estimators, $E\hat{\gamma}_2 = \beta_0 + \beta_1\bar{Z}^m - \beta_0 - \beta_1 c = \beta_1(\bar{Z}^m - c)$. These results can also be derived from the usual straightforward but tedious solution of the normal equations.

Let $\mathbf{X}$ be the $n \times 3$ matrix whose ith row is $(1, Z_iQ_i + c(1 - Q_i), 1 - Q_i)$ and let $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then

$$\begin{aligned} E(\text{RSS}) &= E[\mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}] \\ &= E[\boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}] + E[(\beta_0\mathbf{1} + \beta_1\mathbf{Z})' \\ &\quad \times (\mathbf{I} - \mathbf{H})(\beta_0\mathbf{1} + \beta_1\mathbf{Z})]. \end{aligned}$$

By theorem 1.7 of Seber (1977), $E[\boldsymbol{\varepsilon}'(\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}] = \sigma^2(n - 3)$. Because $\mathbf{I} - \mathbf{H}$ is a projection matrix, the last term of $E(\text{RSS})$ is the RSS from regressing $\beta_0\mathbf{1} + \beta_1\mathbf{Z}$ onto the columns of $\mathbf{X}$. To see the rest of the result, let $\boldsymbol{\beta}^* = (\beta_0, \beta_1, 0)'$ and let $\mathbf{Z}^*$ be an n vector with ith element $(1 - Q_i)(Z_i - c)$. Then one can easily show that $\beta_0\mathbf{1} + \beta_1\mathbf{Z} = \mathbf{X}\boldsymbol{\beta}^* + \beta_1\mathbf{Z}^*$. Because $\mathbf{X}$ is orthogonal to $\mathbf{I} - \mathbf{H}$,

$$(\beta_0\mathbf{1} + \beta_1\mathbf{Z})'(\mathbf{I} - \mathbf{H})(\beta_0\mathbf{1} + \beta_1\mathbf{Z}) = \beta_1^2\mathbf{Z}^{*\prime}(\mathbf{I} - \mathbf{H})\mathbf{Z}^*,$$

which is β_1^2 times the RSS from regressing $\mathbf{Z}^*$ on the columns of $\mathbf{X}$.

Proof of Corollary 3.1

The least squares estimators follow directly from Theorem 3.3. Here $c = 0$. With regards to the $E(\text{MSE})$, defined in Theorem 3.3, $\hat{\eta}_0 = \hat{\eta}_1 = 0$ and $\hat{\eta}_2 = \bar{Z}^m = m_2/(m_1 + m_2)$. The rest follows after noting that the MSE* of Theorem 3.3 is $\sum Q_i(Z_i - \bar{Z}^m)^2/(n - 3)$.

Proof of Theorem 4.2

The true model for subjects in stratum j is

$$Y_i = (\beta_0 + \beta_2 c_j) + \beta_1 X_{1i} + \varepsilon_i, \qquad j = 1, \ldots, k.$$

Model (14) is considered first. This model allows separate $(\gamma_{0j}, \gamma_{1j})$ for each stratum and hence the least squares estimators of them are unbiased for $j = 1, \ldots, k$. Estimation of $(\gamma_{0,k+1}, \gamma_{1,k+1})$ is the omitted covariate problem, and the result follows directly from Lemma 3.1. The expected RSS for the complete-case strata is $(n - \sum I_{k+1,i} - 2k)\sigma^2$, whereas that of the missing-data stratum, according to Lemma 3.1, is

$$\left(\sum I_{k+1,i} - 2\right)\sigma^2 + \beta_2^2\text{RSS}^m(X_2|X_1).$$

Adding these sums of squares and dividing by $n - 2(k + 1)$ yields the expected MSE for (14).

Model (15) is considered next. Estimation of $\{\gamma_{01}, \ldots, \gamma_{0k}, \gamma_1\}$ is equivalent to estimation in a complete-case analysis that, by Theorem 2.1, produces unbiased estimators. Estimation of $(\gamma_{0,k+1}, \gamma_{1,k+1})$ is the same as under model (14). The verification of the expected MSE parallels that of model (14), except that the number of parameters is $k + 3$.

REFERENCES

Afifi, A. A., and Elashoff, R. M. (1967), "Missing Observations in Multivariate Statistics II: Point Estimation in Simple Linear Regression," *Journal of the American Statistical Association*, 62, 10–29.

Anderson, A. B., Basilevsky, A., and Hum, D. P. J. (1983), "Missing Data: A Review of the Literature," in *Handbook of Survey Research*, eds. P. H. Rossi, J. D. Wright, and A. Anderson, New York: Academic Press, pp. 415–492.

Aptech Systems, Inc. (1991), *GAUSS* (Version 2.2), Maple Valley, WA: Author.

Chow, W. K. (1979), "A Look at Various Estimators in Logistic Models in the Presence of Missing Values," in *Proceedings of Business and Economics Section, American Statistical Association*, pp. 417–420.

Cohen, J., and Cohen, P. (1975), *Applied Multiple Regression Correlation Analysis for the Behavioral Sciences*, New York: John Wiley.

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. (1987), *Statistical Analysis With Missing Data*, New York: John Wiley.

Miettinen, O. S. (1985), *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*, New York: John Wiley.

Seber, G. A. F. (1977), *Linear Regression Analysis*, New York: John Wiley.