Regression With Missing X's: A Review

RODERICK J. A. LITTLE*

The literature of regression analysis with missing values of the independent variables is reviewed. Six classes of procedures are distinguished: complete case analysis, available case methods, least squares on imputed data, maximum likelihood, Bayesian methods, and multiple imputation. Methods are compared and illustrated when missing data are confined to one independent variable, and extensions to more general patterns are indicated. Attention is paid to the performance of methods when the missing data are not missing completely at random. Least squares methods that fill in missing X's using only data on the X's are contrasted with likelihood-based methods that use data on the X's and Y. The latter approach is preferred and provides methods for elaboration of the basic normal linear regression model. It is suggested that more widely distributed software is needed that advances beyond complete-case analysis, available-case analysis, and naive imputation methods. Bayesian simulation methods and multiple imputation are reviewed; these provide fruitful avenues for future research.

KEY WORDS: Bayesian inference; Imputation; Incomplete data; Multiple imputation.

1. INTRODUCTION

1.1 Statement of the Problem

Statistical inference with missing data is an important applied problem, because missing values (planned or unplanned) are commonly encountered in practice. Recent advances in computing power and theory have made the topic an active area of statistics research in the last 20 years, and the fruits of this research are becoming available to applied workers in statistics packages. In this article I update earlier reviews (Afifi and Elashoff 1966; Anderson, Basilevsky, and Hum 1983; Hartley and Hocking 1971) for a particular missing-data problem, namely, inference for the regression, of $Y \equiv X_{p+1}$ on p variables $X_1, \ldots, X_p$ based on a random sample of n cases, when some of the X values are missing.

Research has primarily focused on homoscedastic linear regression, where

$$E(Y \mid X_1, \ldots, X_p) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j; \qquad \operatorname{var}(Y \mid X_1, \ldots, X_p) = \sigma^2. \quad (1)$$

Write $\beta = (\beta_1, \ldots, \beta_p)$. If $X_1, \ldots, X_p$, Y have a joint distribution with mean $\mu = (\mu_1, \ldots, \mu_p, \mu_y)$ and covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \sigma_{yy} \end{pmatrix},$$

then standard regression theory gives

$$\beta = \Sigma_{yx}\Sigma_{xx}^{-1}; \qquad \beta_0 = \mu_y - \sum_{j=1}^{p} \beta_j \mu_j;$$

$$\sigma^2 = \sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \quad (2)$$

With complete data, least squares (LS) estimates are obtained by replacing μ and Σ by sample first and second moments; the primary problem considered is to develop estimates of parameters and associated precision when some data are missing. In my review I shall focus on this problem but also mention work on other problems, including incomplete categorical X's, logistic regression, nonlinear regression, and survival analysis with missing X's.

The focus on regression needs little justification in view of its importance as a statistical tool. The related problem of missing values in the outcome Y was prominent in the early history of missing-data methods, but is less interesting in the following sense: If the X's are complete and the missing values of Y are missing at random, then the incomplete cases contribute no information to the regression of Y on $X_1, \ldots, X_p$. The only advantage of methods that use the incomplete cases is computational—for example, to retain a balanced design in ANOVA (see, for example, Little and Rubin 1987, chap. 2). Computational issues have less importance in the era of high speed computers, although they do still arise when p is very large. If values of X are missing as well as Y, then cases with Y missing can provide a minor amount of information for the regression of interest, by improving predictions of missing X's for cases with Y present. The likelihood-based methods described here can be extended to handle this case. Difficult modeling issues not addressed here arise when missing Y's are not missing at random, as when missingness is monotonically related to the value of Y (see, for example, Amemiya 1984; Little and Rubin 1987, chap. 11).

Example 1. Missing Data in a Mental Health Survey. To illustrate missing data problems and methods, I use a subset of the data from a survey of depression in Los Angeles, analyzed in Afifi and Clark (1984), and used to illustrate methods in BMDP (Dixon 1988, app. D9). The regression model (1) was applied to $n = 294$ respondents, where $p = 4$, Y = square root total depression score, X_1 = log(income), X_2 = Age, X_3 = Healthy (a four-point scale indicating general health), and X_4 = Bed-Days (a binary variable indicating whether the respondent had an entire day in bed in the last two months). LS regression yields the coefficient estimates and standard errors in Row 0 of Table 1. Standard diagnostics are satisfactory. The R-squared is

* Roderick J. A. Little is Professor, Department of Biomathematics, UCLA School of Medicine, Los Angeles, CA 90024. This research was supported by JSA 89-17 between the Bureau of the Census and the Regents of the University of California and by USPHS Grant MH37188 from the National Institute of Mental Health. The author thanks Nat Schenker and two referees for useful comments.

Journal of the American Statistical Association
December 1992, Vol. 87, No. 420, Review Paper

Table 1. Regression Coefficients and Standard Errors for Depression Data

		Regressor variable							
		X_1 = Log(Income)		X_2 = Age		X_3 = Healthy		X_4 = Bed-Days	
	Intercept	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error	Coefficient	Standard error
OLS, Data Before Deletion									
0.	3.571	−.744	.246	−.021	.0049	.427	.106	.673	.206
Complete Case Analysis									
1. MCAR	2.202	−.949	.373	−.0224	.0075	.565	.167	.695	.321
2. Selection on X	4.139	−.860	.486	−.0232	.0064	.153	.151	.890	.313
3. Selection on Y	1.557	−.110	.277	−.0145	.0055	.463	.135	.294	.253
Ave. vs. row 0	−.94	.104	54%	.014	32%	−.033	42%	−.046	44%
Available Case Analysis									
4. MCAR	4.066	−1.075	.278	−.0234	.0053	.391	.115	.720	.223
5. Selection on X	3.094	−.590	.364	−.0199	.0053	.435	.119	.658	.227
6. Selection on Y	2.870	−.262	.292	−.0209	.0056	.459	.117	.654	.227
Ave. vs. row 0	−.23	.102	26%	.000	10%	.001	10%	.004	10%
OLS, Conditional Mean Imputed									
7. MCAR	3.914	−.949	.385	−.0230	.0051	.388	.112	.757	.212
8. Selection on X	3.378	−.860	.483	−.0199	.0049	.422	.111	.653	.208
9. Selection on Y	2.627	−.110	.389	−.0200	.0054	.467	.114	.654	.210
Ave. vs. row 0	−.26	.104	71%	.004	5%	.001	6%	.015	2%
Maximum Likelihood, Normal Model									
10. MCAR		−1.00	.356	−.0233	.0051	.392	.112	.759	.214
11. Selection on X		−.857	.471	−.0199	.0048	.416	.110	.658	.208
12. Selection on Y		−.220	.272	−.0206	.0051	.456	.109	.658	.208
Ave. vs. row 0		.052	49%	.001	2%	−.006	4%	.019	2%
Bayesian Simulation, Normal Model									
13. MCAR		−.955	.367	−.0233	.0052	.397	.110	.753	.216
14. Selection on X		−.822	.459	−.0199	.0049	.417	.111	.658	.211
15. Selection on Y		−.200	.517	−.0200	.0058	.458	.118	.650	.210
Ave. vs. row 0		.085	82%	.003	8%	−.003	7%	.014	3%

quite low (.18), but the included variables are clearly statistically significant. Correlations between the regressors are moderate. The model is selected for illustrative purposes and is by no means the best substantive analysis of the data. These results will be compared with those from incomplete data methods after values of X_1 have been deleted in various ways. The deletion of values from a complete data set is rather artificial, but it allows the mechanism of deletion to be varied and provides comparisons with the regression results before deletion.

1.2 Patterns of Missing Data

Some methods apply only to special patterns of missing data, whereas others apply to any pattern. Consider the four examples of missing-data patterns among the X's in Figures 1–4. For *univariate missing data* (see Fig. 1), missing values are confined to a single X, say X_1. This is a special case of *monotone* or *nested* missing data (see Fig. 2), where the columns can be arranged so that X_{j+1} is observed for every case where X_j is observed, for $j = 1, \ldots, p$. Figure 3 displays a pattern where two X's (X_1 and X_2) are never observed together. Such data arise when two samples containing data on X_1 and Y and X_2 and Y are merged into a single data base. Estimates of the regression from this pattern require an assumption (explicit or implicit) about the conditional association of X_1 and X_2 given X_3 and Y. Finally, Figure 4 represents a general pattern with no special structure.

1.3 Missing-Data Mechanisms

Methods differ in assumptions made about the mechanisms leading to missing values. The key issue is whether missingness is related to the data values. For example, given data as in Figure 1, the probability that X_1 is missing for a

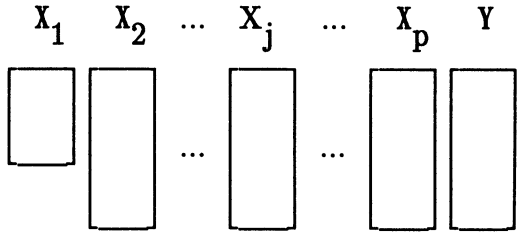

Figure 1. Pattern of Univariate Missing Data.

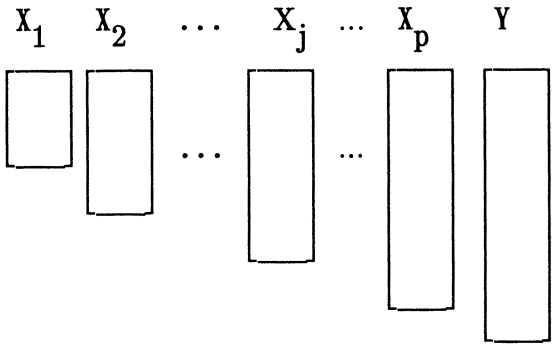

Figure 2. Pattern of Monotone Missing Data.

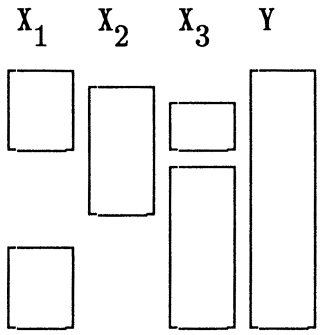

Figure 4. General Pattern of Missing Data.

case may (a) be independent of data values, (b) depend on the value of X_1 for that case, (c) depend on the value of X_2, ..., X_p for that case, or (d) depend on the values of $X_2, \ldots, X_p$ and Y for that case, to name some of the possibilities. If missing values of X_1 are created by subsampling according to a controlled sampling strategy, then case (a) may be justified. If X_1 is income and individuals with high incomes are less likely than others to respond, then case (b) may apply.

Formally, let $\mathbf{Z}$ denote the $n \times (p + 1)$ data matrix, including observed and missing values, let $\mathbf{Z}_{obs}$ denote the set of observed values of $\mathbf{Z}$, and let $\mathbf{Z}_{mis}$ denote the set of missing values. Rubin (1976a) introduced a *missing-data indicator matrix* $\mathbf{R}$, with (i, j)th element $\mathbf{R}_{ij} = 1$ if X_{ij} is observed and $\mathbf{R}_{ij} = 0$ if X_{ij} is missing, and then formalized the notion of a *missing-data mechanism* in terms of a model for the conditional distribution $p(\mathbf{R}|\mathbf{Z}, \varphi)$ of $\mathbf{R}$ given $\mathbf{Z}$, indexed by unknown parameters φ. Data are missing at random (MAR) if this distribution depends on the data $\mathbf{Z}$ only through the observed values $\mathbf{Z}_{obs}$; that is,

$$p(\mathbf{R}|\mathbf{Z}, \varphi) = p(\mathbf{R}|\mathbf{Z}_{obs}, \varphi) \quad \text{for all } \mathbf{Z}_{mis}.$$

Data are missing completely at random (MCAR) if the distribution of R does not depend on the observed or missing values of $\mathbf{Z}$; that is,

$$p(\mathbf{R}|\mathbf{Z}, \varphi) = p(\mathbf{R}|\varphi) \quad \text{for all } \mathbf{Z}.$$

Thus mechanism (a) is MCAR, mechanisms (a), (c), and (d) are MAR because $X_2, \ldots, X_p$ and Y are fully observed, and mechanism (b) is not MAR because X_1 is not fully observed.

A different issue is whether the *pattern* of missing data is random, in the sense that missing-data indicators for the different variables are independent (see Anderson et al. 1983; Glasser 1964). In survey nonresponse settings, missing-indicator variables are often highly correlated for blocks of variables with similar content, because such variables tend to be observed or missing together. The observed pattern clearly affects the information content of the data, but the question of whether missingness is related to the data values (i.e., whether or not the data are MAR or MCAR) is the key to nonresponse bias. Much previous work on missing data in regression compares methods under the (often unrealistic) assumption that the data are MCAR. Here attention will be paid to performance under alternative assumptions about the missing-data mechanism.

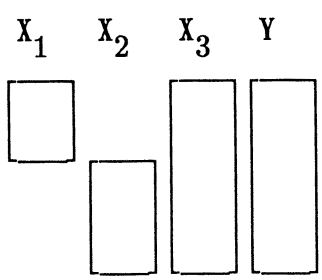

Figure 3. Special Pattern of Missing Data with Unidentified Parameters.

1.4 Taxonomy of Methods

Most proposed methods can be classified into one of the following classes:

1. complete-case (CC) analysis
2. available-case (AC) methods
3. least squares (LS) on imputed data
4. maximum likelihood (ML)
5. Bayesian methods
6. multiple imputation (MI)

A common theme of the latter three methods is that they are based on *models* for the data and missing-data mechanism (although the models may be implicit rather than explicitly formulated). For missing data in regression it is interesting to compare model-based methods with LS methods, which have received considerable attention in the econometrics literature. Refined LS methods can be quite competitive in some settings; however, in my view LS is inferior to methods 4–6, for reasons that I hope will become apparent. I now review each of these strategies for the linear regression problem.

2. COMPLETE-CASE ANALYSIS

The standard treatment of missing data in statistical packages is *complete-case analysis* (CC), where cases with any missing values are simply discarded. This method is also known as *listwise deletion.* Advantages are ease of implementation and the fact that valid inference is obtained when missingness depends on the regressors, as in case (b) (Glynn and Laird 1986). This is a useful property that is not shared by other, more sophisticated approaches. On the other hand, the rejection of incomplete cases seems an unnecessary waste

of information. If the number of X's is large, then even a sparse pattern of missing X's can result in a substantial number of incomplete cases. It seems reasonable to seek ways to incorporate the incomplete cases into the analysis. One approach is to drop regressor variables with high levels of nonresponse; in this context Rubin (1976b) described measures of a covariate's predictive value that take into account degree of incompleteness. Other ways of incorporating incomplete cases are discussed in Sections 3–8.

CC serves as a useful baseline method for comparisons. It is a least squares method in the sense that it results from minimizing the sum of squares of residuals with respect to the parameters *and* the missing values (Afifi and Elashoff 1966; Yates 1933). To see this, note that any incomplete case with missing values can be assigned a zero residual by suitable choices of the missing values, thus effectively removing that case for estimation of the regression parameters. Treating missing values like parameters in this way is a poor general strategy (Little and Rubin 1983). A more useful approach is to treat the missing values as random variables, as in likelihood-based approaches 4–6.

3. AVAILABLE-CASE ANALYSIS

Available-case analysis (AC) methods use the largest sets of available cases for estimating individual parameters (Little and Rubin 1987). In particular, for regression Glasser (1964) substituted AC estimates of the first two moments of $(X_1, \ldots, X_p, Y)$ in (1): μ_j and σ_{jj} were estimated using the $n^{(j)}$ cases with X_j observed, and $\sigma_{jk}(j \neq k)$ was estimated using the $n^{(jk)}$ cases with X_j and X_k observed. Other versions of AC analysis can be developed, depending on the choice of parameterization.

Although AC appears to use information in the incomplete cases in a plausible way, a defect is that the estimated covariance matrix of the X's is not necessarily positive definite, yielding indeterminate slopes when it is not. This problem is severe when X variables are highly correlated; Haitovsky's (1968) simulations on highly correlated data found the method to be markedly inferior to CC. On the other hand, Kim and Curry (1977) found AC to be superior to CC in simulations based on weakly correlated data. Simulation studies comparing AC regression estimates with maximum likelihood ML under normality (see Sec. 5) suggest superiority for ML even when underlying normality assumptions are violated (Azen, Van Guilder, and Hill 1989; Little 1988a; Muthen, Kaplan, and Hollis 1987).

Example 2. CC and AC Analysis of Depression Data with Missing Values of Income. Using the data in Example 1, an artificial variable $U = \alpha_1 I^* + \alpha_2 D^* + \mathbf{Z}$ was created, where I^* is $X_1 = \log(\text{Income})$, D^* is $Y = \sqrt{\text{Depression}}$ standardized to mean 0 and variance 1 and $\mathbf{Z}$ is an independent standard normal random deviate. Values of $X_1 = \log(\text{Income})$ were deleted when U was positive, yielding a data set with about half the values of Income missing. The large fraction of missing values is chosen to exaggerate differences between methods. Three mechanisms were simulated by the following choices of α_1 and α_2, using in each case the same random number seed for $\mathbf{Z}$:

1. MCAR Selection: $\alpha_1 = \alpha_2 = 0$; this yielded 157 missing values and 137 complete cases.
2. Selection on X_1: $\alpha_1 = 1$, $\alpha_2 = 0$; this yielded 155 missing values and 139 complete cases
3. Selection on Y: $\alpha_1 = 0$, $\alpha_2 = 1$; this yielded 150 missing values and 144 complete cases

The results from CC analysis of the three data sets obtained in this way are given in rows 1–3 of Table 1. The row "Ave. vs. row 0" provides a summary comparison with the estimates before deletion: for estimated coefficients it gives the average deviation of the estimates in rows 1–3 from the estimate in row 0; for estimated standard errors it gives the average percent increase of the standard errors in rows 1–3 over the standard error in row 0. Similar summaries are presented for AC and the other methods, to be discussed later. Results in Table 1 are, of course, illustrative and are not intended to be a basis for generalization.

Under assumptions of the model, estimates from MCAR (row 1) and selection on X (row 2) remain valid; the standard errors increase by 30–50%, reflecting the halving of sample size. The CC estimates and standard errors after selection on Y are biased, as reflected in the marked change in the coefficient of X_1 ($-.110$, compared with $-.744$ from the original data).

Rows 4–6 of Table 1 present results from an AC analysis. Coefficients were obtained by computing the covariance matrix of $(X_1, \ldots, X_4, Y)$ using the "all value" option in the BMDPAM program (Dixon 1988), a method that differs slightly from Glasser's (1964) method. The covariance matrix is then inputted into a regression program. Estimates $\tilde{\beta}$ are consistent under MCAR but generally are not consistent under selection on X or Y; estimated coefficients for X_2–X_4 are closer to row 0 than are those from CC analysis, and are probably acceptable for this data set because correlations between the X's are moderate. AC estimates are less successful for data sets with high correlations.

Standard errors for the AC estimates are taken from BMDPAM and are based on the expression $\text{var}(\tilde{\beta}) \simeq \tilde{\boldsymbol{\Sigma}}_{xx}^{-1}\tilde{\sigma}^2/\tilde{n}$, where $\tilde{\boldsymbol{\Sigma}}_{xx}$ and $\tilde{\sigma}^2$ are AC estimates of $\boldsymbol{\Sigma}_{xx}$ and σ^2 and $\tilde{n}$ is the harmonic mean of the sample sizes of the individual variables (Dixon 1988, p. 684). This method seems to have no theoretical basis; in particular, the standard errors for X_1 appear too small, because additional information in the incomplete cases is modest in this example. Asymptotically consistent estimates of standard errors require more complex formulas (Van Praag, Dijkstra, and Van Velzen 1985).

4. LEAST SQUARES ON IMPUTED DATA

LS methods fill in (or *impute*) the missing X's. The regression of Y on the X's is then computed on the filled-in data by ordinary least squares (OLS) or by a weighted least squares (WLS) scheme that downweights incomplete cases.

4.1 Unconditional Mean Imputation

A simple approach imputes missing X's by their unconditional sample means. Discussions of the method in bivar-

iate settings were given in Wilks (1932) and Afifi and Elashoff (1967). The method yields an inconsistent estimate of Σ (Haitovsky 1968): Assuming MCAR, the sample variance of X_j is biased by a factor $(n^{(J)} - 1)/(n - 1)$, and the sample covariance of X_j and X_k is biased by a factor $(n^{(jk)} - 1)/(n - 1)$. Biases in the resulting estimated slopes can be compensated for by reductions in variance relative to CC if the fraction of complete cases is small (Afifi and Elashoff 1967). But inferences (tests and confidence intervals) are seriously distorted by bias and overstated precision. Correcting the sample covariance matrix for bias leads to the AC method discussed in the previous section (Little and Rubin 1987, chap. 3). Unconditional mean imputation cannot be generally recommended.

4.2 Conditional Mean Imputation Based on X's

A worthwhile improvement is to use information in the observed X's in a case to impute the missing X's. Some authors base imputations on a principal component analysis (Dear 1959; Timm 1970). But a more obvious (and in my view, more promising) approach is to impute for a missing X by linear regression on the observed X's in that case, estimated from the complete cases (Dagenais 1973). WLS regression of the imputed data is recommended.

For univariate missing data as in Figure 1, suppose that X_1 is observed for m cases (say $i = 1, \ldots, m$) and missing for $n - m$ cases (say $i = m + 1, \ldots, n$). Because $E(Y_i|X_i) = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$,

$$E(Y_i|X_{i2}, \ldots, X_{ip}) = \beta_0 + \beta_1 X_{i1}^* + \sum_{j=2}^{p} \beta_j X_{ij},$$

where $X_{i1}^* = E(X_{i1}|X_{i2}, \ldots, X_{ip})$. Thus, if conditional means X_{i1}^* are substituted for missing values of X_{i1}, then LS on the filled-in data produces consistent estimates of the regression coefficients, assuming MCAR. Also, write s for the set of subscripts $(2, \ldots, p)$ and let $\sigma_{yy \cdot s}$ and $\sigma_{yy \cdot 1s}$ denote the residual variances of Y given $X_2, \ldots, X_p$ and Y given $X_1, X_2, \ldots, X_p$. Then, to compensate for the increased residual variance when X_1 is missing, incomplete cases should be assigned the reduced weight

$$w^* = \sigma_{yy \cdot 1s}/\sigma_{yy \cdot s} = 1 - \rho_{1y \cdot s}^2, \tag{3}$$

where $\rho_{1y \cdot s}$ is the partial correlation of X_1 and Y given $X_2, \ldots, X_p$. Replacing the parameters in (3) by sample estimates yields weights proposed by Dagenais (1973) and Beale and Little (1975).

The imputations X_{i1}^* depend on the unknown regression parameters, which in practice must be estimated from the data. Gourieroux and Montfort (1981) and Conniffe (1983a) noted that estimation error in regression coefficients inflates the residual variance and also introduces a correlation between the incomplete observations. This does not affect the consistency of the WLS estimates, but does affect the best choice of weight and consistency of estimates of standard error. Arguing by a rather loose analogy with generalized least squares (GLS), these authors proposed the improved weight

$$w = \frac{(1 - \rho_{1y \cdot s}^2)m/n}{\rho_{1y \cdot s}^2 + (1 - \rho_{1y \cdot s}^2)m/n}, \tag{4}$$

which approximates (3) when the fraction of complete cases is large but gives less weight to the incomplete cases.

For general patterns of missing data, the appropriate generalization of the weight (3) is the ratio of the residual variance of Y given all the X's to the residual variance of Y given the observed X's for that case. Dagenais (1973) proposed this weighting with imputations and weights based on the complete cases; Beale and Little (1975) studied a similar method but with imputations based on an estimate of the covariance matrix Σ_{xx} that used all the data. Analogs of the weight (4) for a general pattern of missing data have not been developed.

Example 3. Imputation of Depression Data Sets. Rows 7–9 of Table 1 show the results of imputing the conditional mean of X_1 given $X_2 \cdots X_p$. Because X_1 is weakly related to $X_2 \cdots X_p$ here, results from imputing the unconditional mean are similar and are not presented; OLS results are presented, WLS being similar in this case. Note that the coefficients for X_1 are the same as for complete cases, a result that holds in general for this pattern. Coefficients for X_2–X_4 are like those from AC. Standard errors are based on the filled-in data. Because they do not account for uncertainty in the imputed values, they are underestimated on the average; however, the standard errors for coefficients of X_1 are in fact higher than for complete cases in this realization.

4.3 Conditional Mean Imputation Based on X's and Y

If the partial correlation of Y and a missing X given the observed X's is high, then better imputations can be obtained by using Y as well as the observed X's for imputation. It seems like cheating to use Y to fill in missing X's when the objective is to regress Y on the X's. Indeed, if the regression of Y on the X's is computed by LS on the filled-in data, then biased regression estimates result. Afifi and Elashoff (1969a,b) studied bias-corrected versions of this approach for the case of univariate X.

For a general pattern of missing data, μ and Σ can be estimated by Buck's (1960) method and substituted in (2). Buck imputed missing X's by a regression of the missing X's on the observed X's and Y, with coefficients based on the complete cases. Easily computed corrections to the estimated covariance matrix from the filled-in data correct for bias. Specifically, if X_j is missing for a case, then the residual variance of X_j given y and the observed X's in that case (computed from the regression on complete cases) is added to the sum of squares of X_j; and if both X_j and X_k are missing, the residual covariance of X_j and X_k given y and the observed X's is added to the sum of cross products of X_j and X_k. Buck (1960) gave the corrections for the variances but omitted the corrections for the covariances; see Beale and Little (1975). The corrected version of Buck's method is in fact closely related to normal ML, as discussed in the next section.

Whether the missing X's are imputed using only observed

X's or using observed X's and Y, estimated standard errors of the regression coefficients from OLS or WLS on the filled-in data will tend to be too small, because imputation error is not taken into account. Formulas for standard errors can be developed for special patterns such as Figure 1, but appear harder to derive for general patterns. One possibility is to compute the sample variance of the slopes over a set of bootstrap samples; properties of this approach do not appear to have been explored. Another approach is multiple imputation, as discussed in Section 7.

5. MAXIMUM LIKELIHOOD

5.1 ML Estimation for the Multivariate Normal Model: Factored Likelihood Methods

Another approach computes ML estimates for a model for the joint distribution of Y and X. A basic choice is the multivariate normal distribution with mean μ and covariance matrix Σ. ML estimates of μ and Σ are substituted into (1), yielding ML estimates of the regression parameters.

Afifi and Elashoff (1966) reviewed early contributions to ML estimation of μ and Σ by Wilks (1932), Lord (1955), Edgett (1956), Rao (1956), Nicholson (1957), and Anderson (1957). Anderson introduced the important idea of factoring the likelihood to obtain explicit ML solutions for special patterns of missing data. Gourieroux and Montfort (1981) applied Anderson's method to regression with missing X's. For the data pattern of Figure 1, the distribution of X_1 and Y given the other X's can be factored as

$$p(X_1, Y \mid X_2, \ldots, X_p; \theta) = p(X_1 \mid X_2, \ldots, X_p, Y; \varphi_1)p(Y \mid X_2, \ldots, X_p; \varphi_2).$$

The corresponding likelihood of φ_1 and φ_2 factors as

$$L(\varphi_1, \varphi_2) = L_1(\varphi_1)L_2(\varphi_2), \tag{5}$$

where L_1 is a product of the normal density of X_1 given $X_2, \ldots, X_p$ and Y over the m complete observations, and L_2 is a product of the normal density of Y given $X_2, \ldots, X_p$ over all n observations. Because φ_1 and φ_2 are distinct sets of parameters, their ML estimates are obtained by maximizing L_1 and L_2 separately, two standard complete-data problems. ML estimates of parameters of the regression of interest are then obtained by expressing them as functions of φ_1 and φ_2 and substituting their ML estimates. In particular, ML estimates of the coefficients of X_1 and X_j ($j > 1$) on Y take the form

$$\hat{\beta}_{y1\cdot 1s} = \frac{\tilde{\beta}_{1y\cdot sy}\hat{\sigma}_{yy\cdot s}}{\tilde{\sigma}_{11\cdot sy} + \tilde{\beta}^2_{1y\cdot sy}\hat{\sigma}_{yy\cdot s}};$$

$$\hat{\beta}_{yj\cdot 1s} = \frac{\hat{\beta}_{yj\cdot s}\tilde{\sigma}_{11\cdot sy} - \tilde{\beta}_{1y\cdot sy}\tilde{\beta}_{1j\cdot sy}\hat{\sigma}_{yy\cdot s}}{\tilde{\sigma}_{11\cdot sy} + \tilde{\beta}^2_{1y\cdot sy}\hat{\sigma}_{yy\cdot s}}. \tag{6}$$

Here s again represents the set of subscripts $\{2, \ldots, p\}$, and $\beta_{1y\cdot sy}$ and $\sigma_{11\cdot sy}$ denote the slope of Y and residual variance for the regression of X_1 on $X_2, \ldots, X_p$ and Y, etc. Parameters with tildes on the right sides of (6) belong to φ_1 and are estimated from the m complete cases; parameters with hats belong to φ_2 and are estimated from all n cases. Equivalent expressions for $p = 2$ are given in the ML row of Table 2, in a form that shows how they relate to WLS estimates.

Asymptotic standard errors under MAR can be obtained from the inverse of the information matrix, obtained by twice differentiating $\log L(\varphi_1, \varphi_2)$. In particular, let $\hat{\theta} = \theta(\tilde{\varphi}_1, \hat{\varphi}_2)$ be the ML estimate of a parameter $\theta(\varphi_1, \varphi_2)$. Because $\tilde{\varphi}_1$ and $\hat{\varphi}_2$ are asymptotically uncorrelated,

$$\text{var}(\hat{\theta}) = \hat{\theta}_1^T\mathbf{I}_m^{-1}(\tilde{\varphi}_1)\hat{\theta}_1 + \hat{\theta}_2^T\mathbf{I}_n^{-1}(\hat{\varphi}_2)\hat{\theta}_2, \tag{7}$$

where $\hat{\theta}_j$ is the vector of partial derivatives of θ with respect to φ_j evaluated at $\hat{\varphi}$ and $\mathbf{I}_m$ and $\mathbf{I}_n$ are the information matrices of φ_1 and φ_2. The variance estimate of the CC estimator $\tilde{\theta}$ is

$$\text{var}(\tilde{\theta}) = \tilde{\theta}_1^T\mathbf{I}_m^{-1}(\tilde{\varphi}_1)\tilde{\theta}_1 + \tilde{\theta}_2^T\mathbf{I}_m^{-1}(\tilde{\varphi}_2)\tilde{\theta}_2,$$

where the derivatives $\tilde{\theta}_j$ are evaluated at $\tilde{\varphi} = (\tilde{\varphi}_1, \tilde{\varphi}_2)$. If the data are MCAR, then $\tilde{\varphi} \simeq \hat{\varphi}$, $\tilde{\theta}_j \simeq \hat{\theta}_j$ and

$$\text{var}(\hat{\theta}) \simeq \text{var}(\tilde{\theta}) - \hat{\theta}_2^T\{(\mathbf{I}_m^{-1}(\hat{\varphi}_1) - \mathbf{I}_n^{-1}(\hat{\varphi}_2)\}\hat{\theta}_2. \tag{8}$$

This expression is simpler than (7) in that it does not involve partial derivatives with respect to φ_1. The last term, which is positive, represents the reduction in variance from including the incomplete cases and is used to provide the proportional variance reductions in Table 3.

Factored likelihood methods can also be applied to more complex missing-data patterns, such as Figures 2 and 3 (Little and Rubin 1987, chap. 6; Rubin 1974).

Example 4. ML for Depression Data Sets. Rows 10–12 of Table 1 present results from applying the normal ML method to the depression data. Estimates can be found by computing the ML estimate of Σ in BMDPAM (Dixon 1988) and then inputting the result into a regression routine. Standard errors are not currently implemented in BMDP—those in the table were derived using (8) and hence assume MCAR. Results are quite similar to previous analyses. One would hope that the coefficient of X_1 for row 12 is improved by ML, because this method is consistent when missingness depends on Y, under normal assumptions. The estimate is still too large (−.22), although it is closer to the estimate before deletion (−.74) than are estimates from the previously discussed methods. Standard errors of the other coefficients reflect gains in efficiency over CC analysis.

Table 2. Regression of Y on X_1 and X_2, X_1 Observed for m Cases and Missing for n − m Cases: Estimators From Four Methods

	Parameter	
Method	$\beta_{y1\cdot 12}$	$\beta_{y2\cdot 12}$
CC	$\tilde{\beta}_{y1\cdot 12}$	$\tilde{\beta}_{y2\cdot 12}$
OLS	$\tilde{\beta}_{y1\cdot 12}$	$\hat{\beta}_{y2\cdot 2} - \tilde{\beta}_{y2\cdot 12}\tilde{\beta}_{12\cdot 2}$
WLS	$\tilde{\beta}_{y1\cdot 12}$	$(1 - \tilde{\rho}^2_{1y\cdot 2})\hat{\beta}_{y2\cdot 2} + \tilde{\rho}^2_{1y\cdot 2}\tilde{\beta}_{y2\cdot 2} - \tilde{\beta}_{y1\cdot 12}\tilde{\beta}_{12\cdot 2}$
ML	$\frac{\hat{\rho}^2_{1y\cdot 2}}{\tilde{\rho}^2_{1y\cdot 2}}\tilde{\beta}_{y1\cdot 12}$	$(1 - \hat{\rho}^2_{1y\cdot 2})\hat{\beta}_{y2\cdot 2} + \hat{\rho}^2_{1y\cdot 2}\tilde{\beta}_{y2\cdot 2} - \frac{\hat{\rho}^2_{1y\cdot 2}}{\tilde{\rho}^2_{1y\cdot 2}}\tilde{\beta}_{y1\cdot 12}\tilde{\beta}_{12\cdot 2}$

NOTE: Quantities with a tilde (˜) are standard complete-data ML estimates based on the m complete cases; $\hat{\beta}_{y2\cdot 2}$ and $\hat{\sigma}_{yy\cdot 2}$ are standard complete-data ML estimates based on all n cases; and $\hat{\rho}^2_{1y\cdot 2}$ is the ML estimate of $\rho^2_{1y\cdot 2}$, namely,

$$\hat{\rho}^2_{1y\cdot 2} = \frac{\hat{\beta}^2_{1y\cdot 2y}\hat{\sigma}_{yy\cdot 2}}{\tilde{\sigma}_{11\cdot 2y} + \hat{\beta}^2_{1y\cdot 2y}\hat{\sigma}_{yy\cdot 2}}.$$

5.2 Theoretical Comparisons of ML, LS, and CC

Efficiency comparisons of ML with CC and/or LS were given by Conniffe (1983a,b), Donner and Rosner (1982), Gourieroux and Montfort (1981), Hill and Ziemer (1983), Hocking and Smith (1972), and Nijman and Palm (1988). I compare here LS, ML, and CC for Figure 1 with $p = 2$ regressors, assuming MCAR. Table 2 gives expressions for estimates of the regression coefficients of Y on X_1 and X_2 ($\beta_{y1\cdot 12}$ and $\beta_{y2\cdot 12}$), based on OLS, WLS with weights given by (3), and ML under normality. In these expressions $\hat{\beta}_{y2\cdot 2}$ is the LS regression coefficient of Y on X_2 based on all n cases, and estimates with a tilde ($\tilde{}$) are based on LS on the m complete cases. The large sample variances of these estimates are presented in Table 3, expressed as a proportional decrease in variance relative to CC. Note the following:

1. OLS and WLS yield the same estimate of $\beta_{y1\cdot 12}$ as does CC; ML is more efficient, with incomplete cases contributing a fraction $2\rho^2_{1y\cdot 2}(1 - \rho^2_{1y\cdot 2})$ of a CC, which is bounded by $\frac{1}{2}$ when $\rho^2_{1y\cdot 2} = \frac{1}{2}$.
2. As might be expected, all three alternatives to CC achieve maximum efficiency for estimating $\beta_{y2\cdot 12}$ when both ρ^2_{12} and $\rho^2_{1y\cdot 2}$ are small; that is, the incomplete variable X_1 has no effect on the regression of Y on X_2. Because the regression of X_1 on X_2 yields good predictions of the missing values when ρ^2_{12} is large, one might expect LS methods to do well in that case; however, this is not true for the regression that conditions on both X_1 and X_2. The efficiency result helps to explain the performance of LS in the simulations of Hill and Ziemer (1983).
3. OLS is worse than CC when $\rho^2_{1y\cdot 2} > \frac{1}{2}$. WLS corrects the deficiencies of OLS, and ML is slightly more efficient than WLS.

These comparisons suggest that OLS is not a reliable way to recover information from the incomplete cases, and WLS can come close to ML in terms of efficiency, provided that the weights are well chosen.

ML, unlike LS, yields consistent estimates of other regression parameters. For example, suppose that there are three variables, the regression of Y on X_1 is under study, X_1 is hard or expensive to measure, and X_2 is a proxy for X_1 that is easy or inexpensive to measure. Then an efficient design might collect a large sample on X_2 and Y and a small subsample on X_1, X_2, and Y, yielding the pattern of Figure 1 with $p = 2$. But for the regression of Y on X_1, it usually does not make sense to condition on the proxy variable X_2; for this reason, $\beta_{y1\cdot 1}$ rather than $\beta_{y1\cdot 12}$ is the parameter of interest.

LS methods (unweighted or weighted) do not yield consistent estimates of $\beta_{y1\cdot 1}$. To see this, note that $\beta_{y1\cdot 1} = \beta_{y1\cdot 12} + \beta_{y2\cdot 12}\beta_{21\cdot 1}$ for both parameters and LS estimates, and LS yields consistent estimates of $\beta_{y1\cdot 12}$ and $\beta_{y2\cdot 12}$ but does *not* yield a consistent estimate of $\beta_{21\cdot 1}$, because the denominator of the LS estimate from the filled-in data underestimates the quantity $\{n\sigma_{11}\}$.

The ML estimate of $\beta_{y1\cdot 1}$ is

$$\hat{\beta}_{y1\cdot 1} = \frac{\hat{\beta}_{1y\cdot y}\hat{\sigma}_{yy}}{\hat{\sigma}_{11\cdot y} + \hat{\beta}^2_{1y\cdot y}\hat{\sigma}_{yy}},$$

where $\hat{\beta}_{1y\cdot y} = \tilde{\beta}_{1y\cdot 2y} + \tilde{\beta}_{12\cdot 2y}\hat{\beta}_{2y\cdot y}$ and $\hat{\sigma}_{11\cdot y} = \tilde{\sigma}_{11\cdot 2y} + \tilde{\beta}^2_{12\cdot 2y}\hat{\sigma}_{22\cdot y}$. Here $\hat{\sigma}_{yy}$, $\hat{\beta}_{2y\cdot y}$, and $\hat{\sigma}_{22\cdot y}$ are computed using all n cases, and $\tilde{\beta}_{1y\cdot 2y}$ and $\beta_{12\cdot 2y}$ are computed using the m complete cases. This estimate is consistent. Under MCAR, the proportional reduction in asymptotic variance over CC from (8) is

$$\begin{aligned}\operatorname{var}(\tilde{\beta}_{y1\cdot 1})/\operatorname{var}(\hat{\beta}_{y1\cdot 1}) - 1 &= \left(1 - \frac{m}{n}\right)\{(2\rho^2_{1y}(1 - \rho^2_{1y}) + (1 - 2\rho^2_{1y})\rho^2_{21\cdot y} \\ &\quad + 2\rho^2_{1y}(1 - \rho^2_{1y})\rho^4_{21\cdot y}\}.\end{aligned}$$

Hence the value of an incomplete case ranges from $2\rho^2_{1y}(1 - \rho^2_{1y})$ when $\rho^2_{21\cdot y} = 0$ and data on the proxy supply no information to 1 when $\rho^2_{21\cdot y} = 1$ and data on the proxy fully recover the missing information on X_1.

Table 3. Regression of Y on X_1 and X_2, X_1 Observed for m Cases and Missing for n − m Cases: Proportional Decrease in Variance of Estimators from OLS, WLS, and ML Relative to CC Estimate

	Parameter	
Method	$\beta_{y1\cdot 12}$	$\beta_{y2\cdot 12}$
OLS	0	$\left(1 - \frac{m}{n}\right)\frac{(1 - \rho^2_{12})(1 - 2\rho^2_{1y\cdot 2})}{1 - \rho^2_{1y\cdot 2}}$
WLS	0	$\left(1 - \frac{m}{n}\right)(1 - \rho^2_{1y\cdot 2})(1 - \rho^2_{12})$
ML	$\left(1 - \frac{m}{n}\right)2\rho^2_{1y\cdot 2}(1 - \rho^2_{1y\cdot 2})$	$\left(1 - \frac{m}{n}\right)(1 - \rho^2_{1y\cdot 2})[1 - \rho^2_{12}(1 - 2\rho^2_{1y\cdot 2})]$

5.3 ML for General Patterns

ML for a general pattern of missing data requires iterative methods. Scoring algorithms for the normal model were developed by Trawinski and Bargmann (1964) and Hartley and Hocking (1971). Orchard and Woodbury (1972) formulated an alternative approach (a "missing information principle"), later called the EM algorithm by Dempster et al. (1977). The E step of EM computes the expected value of the complete-data log-likelihood, given the observed data and current parameter estimates, and the M step of EM maximizes the resulting function to provide new parameter estimates. The appeal of EM lies in the fact that for many problems the M step is a complete-data problem with an easy or existing solution. History, theory, and examples of EM were given in Dempster et al. (1977), Little and Rubin (1987), and Orchard and Woodbury (1972).

The E step of EM for ML estimation of (μ, Σ) under multivariate normality effectively imputes the missing X's in a case by regression on the observed X's and Y, as in Buck's method. In fact, EM for this problem is simply an iterated version of Buck's method, with the correction to the covariances noted earlier (Beale and Little 1975). For certain patterns (such as Figure 1), EM starting from CC estimates converges in one iteration, in which case Buck's method is ML under the normal model. Little (1993) showed that

Buck's method is also ML for a general pattern of missing data, under a "pattern mixture" model with different normal distributions for each pattern of missing values and certain restrictions to identify inestimable parameters.

Asymptotic standard errors of ML estimates of slopes can be computed from the inverse of the observed or expected information matrix. This calculation forms part of scoring or Newton algorithms, but is not an output of EM. Other approaches to standard errors are (1) to bootstrap the sample (Little 1988a; Su 1988), (2) to use EM computations to construct numerical approximations to the information matrix (Meng and Rubin 1991), or (3) to use an approximate formula for standard errors, as in Beale and Little (1975). The latter approach performed well in simulations in Little (1979). Su's (1988) simulations compared standard errors computed by the observed or expected information matrix and by the bootstrap, finding the observed information superior to the expected and bootstrap methods better when the model was misspecified.

Some have claimed computational advantages of LS over ML methods, but these advantages seem minor. Both methods yield explicit estimates for special patterns. For general patterns ML requires iteration, but extensions of LS to such cases are not necessarily simpler, because filling in the missing X's is a more complex problem. WLS methods for general patterns in Beale and Little (1975) were dominated by ML in simulations on normal data.

5.4 Assumptions About the Missing-Data Mechanism

Simonoff (1988) stated that CC, LS, and ML require an MCAR assumption, but in fact weaker assumptions suffice. As noted earlier, CC estimates are unbiased if missingness depends on the values of the X's (Glynn and Laird 1986). LS remains valid when missingness depends on the fully observed covariates, provided the missing X's have a linear regression on the observed X's. Under model assumptions, ML remains valid when the data are MAR and in particular when missingness depends on the fully observed covariates and Y, because these variables are fully observed (Rubin 1976a).

In particular, for data in Figure 1, LS is valid if missingness of X_1 depends only on $X_2, \ldots, X_p$; CC is valid if missingness of X_1 depends on $X_1, X_2, \ldots, X_p$; and ML is valid when missingness of X_1 depends on $X_2, \ldots, X_p$ and Y. ML can be extended to handle non-MAR mechanisms by adding terms for the missing-data mechanism $p(\mathbf{R}|\mathbf{Z}, \varphi)$ in the likelihood (Little and Rubin 1987, chap. 11).

6. BAYESIAN METHODS

ML is essentially a large sample tool and has limitations in small samples. Conniffe (1983b) suggested advantages for LS in small samples, based on a small simulation comparison of point estimates. In fact neither LS nor ML methods are satisfactory for small sample inference, and more work is needed on methods that do not rely on asymptotic results. One approach is to add a prior to the likelihood and base inference on the posterior distribution. Work on the related problem of inference about a mean from incomplete data suggests that this approach can yield inferences with good frequentist properties (Little 1988b).

The Bayesian approach has been applied to multivariate problems with incomplete dependent variables (Chen 1986; Guttman and Menzefricke 1983; Little 1988b; Press and Scott 1976; Rubin 1977, 1978, 1987a, in press; Swamy and Mehta 1975), but applications to regression with missing X's seem more limited. The complexity of the likelihood function does not allow explicit expressions for marginal posterior distributions of parameters; these distributions need to be approximated by numerical integration or simulation.

For monotone patterns where the likelihood factorizes into complete-data components, parameters of the components can be drawn from their complete-data posterior distributions and then transformed to yield draws of the regression parameters of interest.

Example 5. Bayesian Simulation for Depression Data. In particular, for draws from the posterior distributions of $\beta_{yj\cdot 1s}$ given the data pattern of Figure 1, $(\beta_{1y\cdot sy}, \beta_{1j\cdot sy}(2 \le j \le p), \sigma_{11\cdot sy})$ are drawn from their complete-data posterior distribution based on the m complete cases and $(\beta_{yj\cdot s}(2 \le j \le p), \sigma_{yy\cdot s})$ are drawn from their complete-data posterior distribution based on all n cases. For conjugate normal priors, these draws are readily constructed from chi-squared and normal random numbers (see, for example, Little and Rubin 1987, sec. 6.3.2). These draws are then transformed into draws for $\beta_{y1\cdot 1s}$ and $\beta_{yj\cdot 1s}$ using the formulas (6) used for transforming ML estimates.

Rows 13–15 in Table 1 present the posterior mean and standard deviation of the regression coefficients for the depression data sets, based on 2,000 draws. For this quite sizable data set, histograms of the posterior draws look normal and the posterior means are close to the ML estimates in rows 10–12. Standard errors are also similar to the standard errors from ML, with the exception of the standard error of X_1 in row 12, which is significantly larger than that for ML in row 9. Note, however, that the latter is computed using (8), which assumes MCAR; this assumption is violated because selection was based on values of Y. ML standard errors based on (7) would be more appropriate and similar to the Bayesian posterior standard errors. For smaller samples the Bayesian and ML results would show more disparity.

For general patterns of missing data, more complex simulation techniques (Tanner 1991), such as data augmentation (Tanner and Wong 1987), the Gibbs sampler (Gelfand and Smith 1990), and importance sampling (Rubin 1987b), are needed to simulate the posterior distribution. Rubin and Schafer (1990) applied these ideas to the normal model. Further developments along these lines can be expected in the future.

7. MULTIPLE IMPUTATION

The imputation methods of Section 4 require special formulas for standard errors; if regression is applied to the filled-in data, then the complete-data standard errors will be too small, because errors in the imputations are not taken into

account. Rubin (1978, 1987a) proposed *multiple imputation* (MI) as a solution to this problem. Instead of imputing a single mean for each missing value, $I \geq 2$ values are drawn from the predictive distribution and then complete-data analyses are repeated I times, once with each imputation substituted. Let $\hat{\theta}_m$ be the estimate of a particular regression parameter θ from the mth analysis, and let $\hat{v}_m$ be the estimated variance. The final estimate of θ is $\hat{\theta} = \Sigma\hat{\theta}_m/I$, with estimated variance

$$\hat{v}^2 = s_w^2 + (1 + I^{-1})s_b^2,$$

where $s_w^2 = \Sigma\hat{v}_m/I$ is the average variance within imputed data sets and $s_b^2 = \Sigma(\hat{\theta}_m - \hat{\theta})^2/(I - 1)$ is the between-imputation variance and reflects uncertainty in the imputation process. Large sample inference for θ is based on treating $(\hat{\theta} - \theta)/\hat{v}$ as t distributed with $\nu = (I - 1)[1 + \{I/(I + 1)\}s_w^2/s_b^2]$ degrees of freedom. For theory underlying the method and practical examples, see Rubin and Schenker (1986) and Rubin (1987a).

Note that MI draws are from the predictive distribution of the missing values and as such condition on the observed data, including the Y's. Thus for data in Figure 1, imputations of X_1 are drawn from the conditional distribution of X_1 given X_2 *and* Y. The draws are thus closer in spirit to the imputations in Section 4.3 or Section 5 than to the LS imputations in Section 4.2, which do not condition on Y. When *means* are imputed, conditioning on X's alone can yield consistent regression estimates, but conditioning of X's and Y requires bias adjustments such as those in Buck (1960) and in the E step of the normal EM algorithm. On the other hand, when *draws* are imputed, conditioning on X's and Y does not lead to bias, whereas conditioning only on the X's does! Specifically, for data of the form of Figure 1, if imputations are draws from the conditional distribution of X_1 given $X_2, \ldots, X_p$, then the regression coefficient of X_1 is attenuated, because the noise added to the conditional means does not account for partial correlation of X_1 and Y given $X_2, \ldots, X_p$. The imputations proposed in Simonoff (1988) have this characteristic and hence yielded biased regression coefficients, although the bias was not a crucial issue in the diagnostic setting of that article.

MI is particularly useful for data base construction, because once the imputations are created, analysis by the user requires only complete-data methods. If imputations are predictions based on an explicit model, then MI is closely related to ML inference. For example, as the sample size and I increase, $\hat{\theta}$ converges to the ML estimate of θ and v converges to the variance of the ML estimate based on the information matrix. Multiple imputations can also be constructed based on an implicit model for the missing values. For example, hot deck imputations match incomplete cases to complete cases using covariate information and then impute values from the complete case. Multiple imputation versions match an incomplete case to a set of complete cases similar with respect to some metric, and then impute more than once by drawing from the set (see, for example, Heitjan and Little 1991). Rubin and Schafer (1990) applied multiple imputation to normal model regression problems.

8. MODELS FOR NONNORMAL DATA

Although this approach lacks strict chronological accuracy, I like to view the literature as evolving from relatively simple but limited methods such as CC and AC analysis, through the imputation methods described in Section 4, to methods based on models—namely ML methods that work well in large samples and multiple imputation and Bayesian methods that may be preferable in small samples.

Under the assumed model, the usual optimal large sample properties of consistency and efficiency of ML apply to incomplete data problems. Sensitivity to model assumptions is an important issue. Rubin (1974) noted that the regression parameter ML estimates based on the multivariate normal model remain ML for a model that fixes the fully observed X's and hence allows for dummy variables, polynomials, and interactions among X's that are fully observed. In particular, for the data in Figure 1, $X_2, \ldots, X_p$ are fully observed, so we can avoid distributional assumptions about those variables and restrict multivariate normal assumptions to the conditional distribution of Y and X_1 given $X_2, \ldots, X_p$.

The normal ML method does not require normality to yield consistent estimates under MCAR, but it is not necessarily efficient when the data are nonnormal. Nijman and Palm (1987) provided some evidence that WLS methods can have a slight edge over ML when the data are elliptically symmetric with very long tails. An alternative to normal ML estimation in such settings is ML for distributions with longer-than-normal tails. In particular, Little (1988a) showed that the normal EM algorithm is easily modified to provide ML estimates for multivariate t and contaminated multivariate normal models, and Lange, Little, and Taylor (1989) discussed adaptive robust inference for the t model.

Neither normal ML nor LS methods seem appropriate for incomplete categorical X's, represented in the regression by binary dummy variables; in particular, both methods can yield linear predictions outside the allowable range. Little and Schluchter (1985) presented an ML method for mixed continuous and categorical variables with missing data based on Olkin and Tate's (1961) extension of the model for discriminant analysis. This method yields a tractable EM algorithm for regression with missing categorical X's and for logistic regression with missing X's. Schafer (1991) provided algorithms for simulating Bayesian posterior distributions under this model. Ibrahim (1990) and Schluchter and Jackson (1989) discussed EM algorithms for respectively generalized linear models and survival analysis with missing categorical X's. Pepe and Fleming (1991) gave an approximate likelihood method for nonlinear regression with missing X's.

Model analyses should be accompanied by diagnostics to detect model sensitivity. This seems to be an area where more work would be useful. Mahalanobis-type distance plots for multivariate normality with missing data were considered by Little (1988a) and Lange et al. (1989). More specifically, for regression, Shih and Weisberg (1986) discussed influence measures with incomplete data using an ML estimation framework and Simon and Simonoff (1986) and Simonoff (1988) considered diagnostic plots and tests for consonance of the data with MCAR.

The models discussed here all assume MAR; that is, they do not model the missing-data mechanism. Some work has been done on non-MAR missing-data mechanisms, although here sensitivity to misspecification is a serious issue (Brown 1990). Two areas where more work appears to be needed are small sample inference methods (Sec. 7) and methods for nonnormal data where likelihood methods are hard to implement. Examples of the latter include nonlinear or generalized linear models with continuous missing X's and survival analysis with interval-censored covariates (Little 1992). Approximate approaches based on multiple imputation may yield adequate answers for such problems (Dorey, Little, and Schenker in press).

For practitioners, CC analysis remains the most common method in the absence of readily available alternatives in software packages. We have seen that naive alternatives to CC are not necessarily improvements and that if the amount of missing data is minor, then CC may be a tolerable option. But this method becomes markedly less attractive as the fraction of incomplete cases increases. Aside from efficiency concerns, dropping variables or incomplete cases is philosophically unappealing, because I think that the statistician's role is to analyze all available data in the best way possible and resist restriction to subsets of data unless the reasons for doing so are truly compelling. AC analysis and imputation methods both attempt to avoid discarding data but have deficiencies as general methods. Model-based estimation methods (i.e., ML estimation, multiple imputation, or Bayes) seem preferable, because they use all the data and are grounded in established principles of statistical inference.

In contrast to incomplete repeated-measures analysis, where software development is currently active, most widely distributed software for regression with incomplete covariates is still restricted to AC and CC analysis. Algorithms for ML estimation for the multivariate normal model are available in BMDP (Dixon 1988) and Gauss (Aptech Systems 1988), but some postprocessing is needed to yield ML estimates of the regression coefficients and associated standard errors. Other methods described in Sections 5–7 are not currently available in the major packages, although programs may be obtainable from individual researchers. For example, software to carry out the Bayesian methods in Rubin and Schafer (1990) and Schafer (1991) is available from Schafer. It is hoped that the year 2000 review of this topic will find these methods more widely available to users.

[*Received September 1990. Revised March 1992.*]

REFERENCES

Afifi, A. A., and Clark, V. (1984), *Computer-Aided Multivariate Analysis: A Computer Oriented Approach,* 2nd ed., New York: Academic Press.

Afifi, A. A., and Elashoff, R. M. (1966), "Missing Observations in Multivariate Statistics: I: Review of the Literature," *Journal of the American Statistical Association,* 61, 595–605.

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

——— (1969a), "Missing Observations in Multivariate Statistics: III: Large Sample Analysis of Simple Linear Regression," *Journal of the American Statistical Association,* 64, 337–358.

——— (1969b), "Missing Observations in Multivariate Statistics: IV: A Note on Simple Linear Regression," *Journal of the American Statistical Association,* 64, 358–365.

Amemiya, T. (1984), "Tobit Models: A Survey," *Journal of Econometrics,* 24, 3–61.

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.

Anderson, T. W. (1957), "Maximum Likelihood Estimation for the Multivariate Normal Distribution When Some Observations are Missing," *Journal of the American Statistical Association,* 52, 200–203.

Aptech Systems (1988), *GAUSS Programming Language,* Kent, WA: Author.

Azen, S. P., Van Guilder, M., and Hill, M. A. (1989), "Estimation of Parameters and Missing Values Under a Regression Model With Non-Normally Distributed and Non-Randomly Incomplete Data," *Statistics in Medicine,* 8, 217–228.

Beale, E. M. L., and Little, R. J. A. (1975), "Missing Values in Multivariate Analysis," *Journal of Royal Statistical Society,* Ser. B, 37, 129–145.

Brown, C. H. (1990), "Protecting Against Nonrandomly Missing Data in Longitudinal Studies," *Biometrics,* 33, pp. 143–156.

Buck, S. F. (1960), "A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer," *Journal of the Royal Statistical Society,* Ser. B, 22, 302–306.

Chen, C.-F. (1986), "A Bayesian Approach to Nested Missing-Data Problems," in *Bayesian Inference and Decision Techniques,* eds. P. Goel and A. Zellner, New York: Elsevier Press, pp. 355–361.

Conniffe, D. (1983a), "Comments on the Weighted Regression Approach to Missing Values," *Economic and Social Review,* 14, 259–272.

——— (1983b), "Small-Sample Properties of Estimators of Regression Coefficients Given a Common Pattern of Missing Data," *Review of Economic Studies,* L, 111–120.

Dagenais, M. G. (1973), "The Use of Incomplete Observations in Multiple Regression Analysis: A Generalized Least Squares Approach," *Journal of Econometrics,* 1, 317–328.

Dear, R. E. (1959), "A Principal Components Missing Data Method for Multiple Regression Models," SP-86, Santa Monica, CA: Systems Development Corp.

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

Dixon, W. J. (ed.) (1988), *BMDP Statistical Software Manual,* (Vol. 2), Berkeley, CA: University of California Press.

Donner, A., and Rosner, B. (1982), "Missing Value Problems in Multiple Linear Regression with Two Independent Variables," *Communications in Statistics, Part A—Theory and Methods,* 11, 127–140.

Dorey, F., Little, R. J. A., and Schenker, N. (in press), "Multiple Imputation of Interval-Censored Data with a Threshold Response," *Statistical Medicine,* 12.

Edgett, G. L. (1956), "Multiple Regression with Missing Observations Among the Independent Variables," *Journal of the American Statistical Association,* 51, 122–131.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association,* 85, 398–409.

Glasser, M. (1964), "Linear Regression Analysis with Missing Observations Among the Independent Variables," *Journal of the American Statistical Association,* 59, 834–44.

Glynn, R. J., and Laird, N. M., (1986), "Regression Estimates and Missing Data: Complete-Case Analysis," Technical Report, Harvard School of Public Health, Dept. of Biostatistics.

Gourieroux, C., and Montfort, A. (1981), "On the Problem of Missing Data in Linear Models," *Review of Economic Studies,* XLVIII, 579–586.

Guttman, I., and Menzefricke, U. (1983), "Bayesian Inference in Multivariate Regression with Missing Observations on the Response Variables," *Journal of Business and Economic Statistics,* 1, 239–248.

Haitovsky, Y. (1968), "Missing Data in Regression Analysis," *Journal of the Royal Statistical Society,* Ser. B, 30, 67–82.

Hartley, H. O., and Hocking, R. R. (1971), "The Analysis of Incomplete Data," *Biometrics,* 14, 174–194.

Heitjan, D. F., and Little, R. J. A. (1991), "Multiple Imputation in the Fatal Accident Reporting System," *Applied Statistics,* 40, 13–29.

Hill, R. C., and Ziemer, R. F. (1983), "Missing Regressor Values Under Conditions of Multicollinearity," *Communications in Statistics, Part A—Theory and Methods,* 12, 2557–2573.

Hocking, R. R., and Smith, W. B. (1972), "Optimum Incomplete Multinormal Samples," *Technometrics,* 14, 299–307.

Ibrahim, J. G. (1990), "Incomplete Data in Generalized Linear Models," *Journal of the American Statistical Association,* 82, 765–769.
Kim, J.-O., and Curry, J. (1977), "Treatment of Missing Data in Multivariate Analysis," *Sociological Methods and Research,* 6, 215–240.
Lange, K., Little, R. J. A., and Taylor, J. M. G. (1989), "Robust Statistical Inference Using the T Distribution," *Journal of the American Statistical Association,* 84, 881–896.
Little, R. J. A. (1979), "Maximum Likelihood Inference for Multiple Regression with Missing Values: A Simulation Study," *Journal of the Royal Statistical Society,* Ser. B, 41, 76–87.
——— (1988a), "Robust Estimation of the Mean and Covariance Matrix From Data with Missing Values," *Applied Statistics,* 37, 23–29.
——— (1988b), "Small Sample Inference About Means from Bivariate Normal Data with Missing Values," *Computational Statistics and Data Analysis,* 7, 161–178.
——— (1992), "Incomplete Data in Event History Analysis," in *Demographic Applications of Event History Analysis,* eds. J. Trussell, R. Hankinson and J. Tilton, Oxford, U.K.: Clarendon Press, pp. 209–230.
——— (1993), "Pattern-Mixture Models for Multivariate Incomplete Data," To appear in *Journal of the American Statistical Association,* 88.
Little, R. J. A., and Rubin, D. B. (1983), "On Jointly Estimating Parameters and Missing Data by Maximizing the Complete-Data Log-likelihood," *The American Statistician,* 37, 218–220.
——— (1987), *Statistical Analysis with Missing Data,* New York: John Wiley.
——— (1989), "Missing Data in Social Science Data Sets," *Sociological Methods and Research,* 18, 292–326; reprinted in *Modern Methods of Data Analysis,* eds. J. S. Long and J. Fox, Newbury Park, CA: Sage Press, pp. 374–409.
Little, R. J. A., and Schluchter, M. D. (1985), "Maximum Likelihood Estimation With Mixed Continuous and Categorical Data With Missing Values," *Biometrika,* 72, 497–512.
Lord, F. M. (1955), "Estimation of Parameters From Incomplete Data," *Journal of the American Statistical Association,* 50, 870–876.
Meng, X. L., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association,* 86, 899–909.
Muthen, B., Kaplan, D., and Hollis, M. (1987), "On Structural Equation Modeling with Data that Are Not Missing Completely at Random," *Psychometrika,* 52, 431–462.
Nicholson, G. E. (1957), "Estimation of Parameters From Incomplete Multivariate Samples," *Journal of the American Statistical Association,* 52, 523–526.
Nijman, T., and Palm, F. (1988), "Efficiency Gains Due to Using Missing Data Procedures in Regression Models," *Statistical Papers,* 29, 249–256.
Olkin, I., and Tate, R. F. (1961), "Multivariate Correlation Models With Mixed Discrete and Continuous Variables," *Annals of Mathematical Statistics,* 32, 448–465.
Orchard, T., and Woodbury, M. A. (1972), "A Missing Information Principle: Theory and Applications," *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability,* 1, 697–715.
Pepe, M. S., and Fleming, T. R. (1991), "A Nonparametric Method for Dealing With Mismeasured Covariate Data," *Journal of the American Statistical Association,* 86, 108–113.
Press, S. J., and Scott, A. J. (1976), "Missing Variables in Bayesian Regression, II," *Journal of the American Statistical Association,* 71, 366–369.
Rao, C. R. (1956), "Analysis of Dispersion With Incomplete Observations on One of the Characters," *Journal of the Royal Statistical Society,* Ser. B, 18, 259–264.
Rubin, D. B. (1974), "Characterizing the Estimation of Parameters in Incomplete Data Problems," *Journal of the American Statistical Association,* 69, 467–474.
——— (1976a), "Inference and Missing Data," *Biometrika,* 63, 581–592.
——— (1976b), "Comparing Regressions When Some Predictor Variables Are Missing," *Technometrics,* 18, 201–205.
——— (1977), "Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association,* 72, 538–543.
——— (1978), "Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach," in *Proceedings of the Survey Research Methods Section, American Statistical Association,* pp. 20–34.
——— (1987a), *Multiple Imputation for Nonresponse in Surveys,* New York: John Wiley.
——— (1987b), "Comment on "The Calculation of Posterior Distributions by Data Augmentation" by M. A. Tanner and W. H. Wong," *Journal of the American Statistical Association,* 82, 543–546.
——— (in press), "Computational Aspects of Analyzing Random Effects/Longitudinal Models," *Statistics in Medicine,* 11.
Rubin, D. B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse," *Journal of the American Statistical Association,* 81, 366–374.
Rubin, D. B., and Schafer, J. L. (1990), "Efficiently Creating Multiple Imputations for Incomplete Multivariate Normal Data," *Proceedings of the Statistical Computing Section, American Statistical Association.*
Schafer, J. L. (1991), "Algorithms for Multiple Imputation and Posterior Simulation from Incomplete Multivariate Data with Ignorable Nonresponse," unpublished Ph.D. dissertation, Harvard University, Dept. of Statistics.
Schluchter, M. D., and Jackson, K. L. (1989), "Log-Linear Analysis of Censored Survival Data with Partially Observed Covariates," *Journal of the American Statistical Association,* 84, 42–52.
Shih, W. J., and Weisberg, S. (1986), "Assessing Influence in Multiple Linear Regression with Incomplete Data," *Technometrics,* 28, 231–239.
Simon, G. A., and Simonoff, J. S. (1986), "Diagnostic Plots for Missing Data in Least Squares Regression," *Journal of the American Statistical Association,* 81, 501–509.
Simonoff, J. S. (1988), "Regression Diagnostics to Detect Nonrandom Missingness in Linear Regression," *Technometrics,* 30, 205–214.
Su, H.-L. (1988), "Estimation of Standard Errors in Some Multivariate Models with Missing Data," unpublished Ph.D., dissertation, UCLA School of Public Health, Dept. of Biostatistics.
Swamy, P. A. V. B., and Mehta, J. S. (1975), "On Bayesian Estimation of Seemingly Unrelated Regressions When Some Observations are Missing," *Journal of Econometrics,* 3, 157–169.
Tanner, M. A. (1991), *Tools for Statistical Inference: Observed Data and Data Augmentation Methods,* New York: Springer-Verlag.
Tanner, M., and Wong, W. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association,* 82, 528–550.
Timm, N. H. (1970), "The Estimation of Variance-Covariance and Correlation Matrices from Incomplete Data," *Psychometrika,* 35, 417–437.
Trawinski, I. M., and Bargmann, R. W. (1964), "Maximum Likelihood with Incomplete Normal Data," *Annals of Mathematical Statistics,* 35, 647–657.
Van Praag, B. M. S., Dijkstra, T. K., and Van Velzen, J. (1985), "Least Squares Theory Based on Distributional Assumptions With an Application to the Incomplete Observations Problem," *Psychometrika,* 50, 25–36.
Wilks, S. S. (1932), "Moments and Distributions of Estimates of Population Parameters From Fragmentary Samples," *Annals of Mathematical Statistics,* 3, 163–195.
Yates, F. (1933), "The Analysis of Replicated Experiments When the Field Results Are Incomplete," *The Empire Journal of Experimental Agriculture,* 1, 129–142.