TECHNOMETRICS, AUGUST 1986, VOL. 28, NO. 3

Assessing Influence in Multiple Linear Regression With Incomplete Data

Weichung J. Shih

Clinical Biostatistics
Merck Sharp and Dohme
P.O. Box 2000, WB-D216
Rahway, NJ 07065-0914

Sanford Weisberg

Department of Applied Statistics
University of Minnesota
St. Paul, MN 55108

The problem of assessing influence and detecting influential cases in multiple linear regression with incomplete data is considered. A case is said to be influential if appreciable changes in fitted regression coefficients occur when it is removed from the data. A one-step influence measure is derived, based on the EM algorithm for detecting cases that are influential in the maximum likelihood estimation of the regression coefficients. Results are compared with the (complete data) Cook's distance measure. Techniques are demonstrated by examples.

KEY WORDS: Diagnostics; Cook's distance.

1. INTRODUCTION

A case may be judged influential if appreciable changes in important features of the fitted model occur when the case is deleted. Finding such cases has become a standard part of statistical modeling. One may view the search for influential cases as part of the concern of the analyst over robustness of a fitted model. If conclusions can change when a case is deleted, then the usefulness of the fitted model may be in doubt.

Consider the model

$$y = \beta_0 + \mathbf{x}^T\boldsymbol{\beta} + e,$$

where y is a response variable, $\mathbf{x}^T = (x_1, \ldots, x_p)$ is a vector of predictor variables, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression coefficients, and e is an error with mean 0 and variance σ^2. For fitting with complete data, the influence of the ith case can be assessed by a distance measure (Cook 1977, 1979; Cook and Weisberg 1980, 1982),

$$D_i(\mathbf{A}) = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T\mathbf{A}^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}), \qquad (1.1)$$

where $\hat{\boldsymbol{\beta}}$ denotes an estimate of $\boldsymbol{\beta}$ [or of $(\beta_0, \boldsymbol{\beta}^T)^T$] based on the full data set, $\hat{\boldsymbol{\beta}}_{(i)}$ is the analogous estimate without the ith case, and $\mathbf{A}$ is a positive definite matrix representing the metric chosen. When data are incomplete, in the sense that some values of the $n \times (p + 1)$ data matrix $\mathbf{Z} = (\mathbf{Y}, \mathbf{X}) = (z_{ij})$ are "missing at random" (Rubin 1976), one approach to fitting the regression that uses all of the observed data is to maximize the likelihood of an approximating joint distribution for $\mathbf{z} = (y, \mathbf{x}^T)$. A common choice would be the multinormal distribution, $\mathbf{z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\mathbf{z}_i = \begin{pmatrix} y_i \\ \mathbf{x}_i \end{pmatrix} \sim N_{p+1}\left[\begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}\right]. \qquad (1.2)$$

Given the maximum likelihood estimates (MLE's) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, usual methods for obtaining estimates for the conditional distribution of $y \mid \mathbf{x}$ can then be used.

In this article we obtain influence measures for incomplete data problems in which partially observed predictors can be assumed jointly normally distributed. Fully observed though nonnormal predictors may also be included, although these are not specifically discussed here; Little (1979) gave the details. In Section 2 we outline the EM algorithm (Dempster, Laird, and Rubin 1977) for finding MLE's. In Section 3 we discuss the problem of deleting cases, and we elaborate on the results in Section 4. In Section 5 we present two influence norms, and we give some numerical results in Section 6. Section 7 contains conclusions.

2. MAXIMUM LIKELIHOOD ESTIMATION

Iterative procedures for computing the MLE of $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, using the EM algorithm can be summarized as follows (Beale and Little 1975). Let $\boldsymbol{\theta}_S$ be a starting value or an intermediate value of $\boldsymbol{\theta}$, $\mathbf{R}_i$ be the vector of observed variables in the ith case, and σ_{jk} be the (j, k)th element of $\boldsymbol{\Sigma}$. Each iterative cycle of the EM algorithm consists of two steps: expectation (E step) and maximization (M step). The E step fills in the data matrix and estimates conditional covariances of the unobserved given the observed. The fill-in values are merely conditional expectations: for

each $i = 1, \ldots, n$ and $j = 1, \ldots, p+1$,

$$\hat{z}_{ij,S} = E(z_{ij} \mid \mathbf{R}_i, \boldsymbol{\theta}_S), \qquad z_{ij} \text{ not observed}$$
$$= z_{ij}, \qquad z_{ij} \text{ observed.} \tag{2.1}$$

Similarly, the conditional covariance matrix, $\hat{\mathbf{C}}_{i,S}$ for case i, given $\mathbf{R}_i$ and $\boldsymbol{\theta}_S$ has (j, k)th element

$$\hat{c}_{jk,S|R_i} = \operatorname{cov}(z_{ij}, z_{ik} \mid \mathbf{R}_i, \boldsymbol{\theta}_S), \quad z_{ij}, z_{ik} \text{ not observed}$$
$$= 0, \quad \text{at least one of } z_{ij}, z_{ik} \text{ observed.} \tag{2.2}$$

These values are easily computed from $\boldsymbol{\theta}_S$ and $\mathbf{R}_i$ by use of a "sweep" or Gaussian elimination routine.

The M step obtains the estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by

$$\hat{\mu}_{j,M} = \frac{1}{n} \sum_{i=1}^{n} \hat{z}_{ij,S}, \qquad j = 1, \ldots, p+1, \tag{2.3}$$

and, for each $j, k = 1, 2, \ldots, p+1$,

$$\hat{\sigma}_{jk,M} = \frac{1}{n} \sum_{i=1}^{n} \{(\hat{z}_{ij,S} - \hat{\mu}_{j,M})(\hat{z}_{ik,S} - \hat{\mu}_{k,M}) + \hat{c}_{jk,S|R_i}\}. \tag{2.4}$$

For the iterative procedure, substitute $\hat{\boldsymbol{\theta}}_M = (\hat{\boldsymbol{\mu}}_M, \hat{\boldsymbol{\Sigma}}_M)$ for $\boldsymbol{\theta}_S$, where

$$\hat{\mu}_M = (\hat{\mu}_{j,M}), \qquad j = 1, \ldots, p+1,$$

and

$$\hat{\Sigma}_M = (\hat{\sigma}_{jk,M}), \qquad j, k = 1, \ldots, p+1,$$

and cycle through (2.1)–(2.4) until a convergence criterion is met. At convergence, we denote the fitted matrix by $\hat{\mathbf{Z}} = (\hat{\mathbf{Y}}, \hat{\mathbf{X}}) = (\hat{z}_{ij})$, the conditional covariance matrix for the ith case by $\hat{\mathbf{C}}_i$, and the MLE's of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$, respectively. The MLE's of the regression parameters $\boldsymbol{\beta}$ and σ^2 can be obtained by the usual transformations:

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{xx}^{-1} \hat{\boldsymbol{\Sigma}}_{xy}$$
$$\hat{\beta}_0 = \hat{\boldsymbol{\mu}}_y - \hat{\boldsymbol{\beta}}^T \hat{\boldsymbol{\mu}}_x$$
$$\hat{\sigma}^2 = \hat{\sigma}_y^2 - \hat{\boldsymbol{\Sigma}}_{yx} \hat{\boldsymbol{\Sigma}}_{xx}^{-1} \hat{\boldsymbol{\Sigma}}_{xy}. \tag{2.5}$$

As has been pointed out by many authors (Hocking and Smith 1972; Press and Scott 1974; Rubin 1974; Shih 1981, chap. 5), cases with the response variable y missing convey no information on the estimation of regression parameters. We therefore exclude such cases from consideration in the following discussion. From (2.4) and (2.5) one can show that the convergent form of the MLE of $\boldsymbol{\beta}$ can then be written as

$$\hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}^T \hat{\mathbf{X}} + \hat{\mathbf{C}})^{-1} \hat{\mathbf{X}}^T \mathbf{Y}, \tag{2.6}$$

where $\hat{\mathbf{C}} = \sum_{i=1}^{n} \hat{\mathbf{C}}_i$.

3. DELETING ONE CASE AT A TIME

To measure influence, we need to compute the MLE, or an approximation to it, with one case deleted. We shall develop an approximation to $\hat{\boldsymbol{\beta}}_{(i)}$ that does not include the intercept. If inclusion of the intercept is of interest, then in Equation (2.6) it is necessary to add only an initial column of 1s to $\hat{\mathbf{X}}$ and adjust accordingly.

To obtain $\hat{\boldsymbol{\beta}}_{(i)}$, one can delete the ith case and follow the EM algorithm (described in Section 2) on the remaining $n-1$ cases. That is, choose initial estimates (such as $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ from all n cases) for the mean and covariance matrix and go through the EM iterations until convergence is reached. We denote $\hat{\boldsymbol{\beta}}_{(i)}$, obtained by this "regular" method, by

$$\hat{\boldsymbol{\beta}}_{(i)} = (\hat{\mathbf{X}}_{(i)}^T \hat{\mathbf{X}}_{(i)} + \hat{\mathbf{C}}_{(i)})^{-1} \hat{\mathbf{X}}_{(i)}^T \mathbf{Y}_{(i)}, \tag{3.1}$$

where the notation indicates that case i has been deleted. The concept for this "regular" method is simple, as it is the same as that for obtaining $\hat{\boldsymbol{\beta}}$. The iterations involved can be quite expensive, however, since $\hat{\boldsymbol{\beta}}_{(i)}$ is calculated for every case.

To avoid iteration, we can use a single EM step to approximate $\hat{\boldsymbol{\beta}}_{(i)}$. Let $\mathbf{U}_i$ be the vector of observed x variables in the ith case and $\mathbf{R}_i = (y_i, \mathbf{U}_i)$, since the y is always observed (see the discussion in Section 2). At convergence of the EM algorithm with all of the data, from (2.1) and (2.2) we obtain

$$\hat{x}_{ij} = E(x_{ij} \mid y_i, \mathbf{U}_i, \hat{\boldsymbol{\theta}}),$$
$$\hat{\mathbf{C}}_i = [\operatorname{cov}(x_{ij}, x_{ik} \mid y_i, \mathbf{U}_i, \hat{\boldsymbol{\theta}})].$$

Using $\hat{\boldsymbol{\theta}}$ (hence $\hat{\boldsymbol{\beta}}$) as the initial estimate, when the ith row is deleted from $\hat{\mathbf{X}}$ the first E step in obtaining $\hat{\boldsymbol{\beta}}_{(i)}$ will not change any of the fill-in values $\hat{x}_{ij}$ or the correction terms $\hat{\mathbf{C}}_i$. Hence

$$\hat{\mathbf{X}}_{(i)}^{1\,T} \hat{\mathbf{X}}_{(i)}^{1} = \hat{\mathbf{X}}^T \hat{\mathbf{X}} - \hat{\mathbf{x}}_i \hat{\mathbf{x}}_i^T$$

is a one-step approximation of $\hat{\mathbf{X}}_{(i)}^T \hat{\mathbf{X}}_{(i)}$ and $\hat{\mathbf{C}}_{(i)}^1 = \hat{\mathbf{C}} - \hat{\mathbf{C}}_i$ is a one-step approximation of $\hat{\mathbf{C}}_{(i)}$, where the superscript 1 denotes *one step*. The following M step will be carried out as usual. Thus the one-step approximation to $\hat{\boldsymbol{\beta}}_{(i)}$ is

$$\hat{\boldsymbol{\beta}}_{(i)}^1 = (\hat{\mathbf{X}}_{(i)}^{1\,T} \hat{\mathbf{X}}_{(i)}^1 + \hat{\mathbf{C}}_{(i)}^1)^{-1} \hat{\mathbf{X}}_{(i)}^{1\,T} \mathbf{Y}_{(i)}. \tag{3.2}$$

A simple updating relationship between $\hat{\boldsymbol{\beta}}_{(i)}^1$ and $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}}_{(i)}^1 = \hat{\boldsymbol{\beta}} - [\hat{\mathbf{X}}_{(i)}^{1\,T} \hat{\mathbf{X}}_{(i)}^1 + \hat{\mathbf{C}}_{(i)}^1]^{-1} [\hat{\mathbf{x}}_i \hat{r}_i - \hat{\mathbf{C}}_i \hat{\boldsymbol{\beta}}], \tag{3.3}$$

where $\hat{r}_i = y_i - \hat{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}}$ is the ith *estimated residual* [see the Appendix for a derivation of (3.3)]. If the ith case is complete, then $\hat{\mathbf{x}}_i = \mathbf{x}_i$, $\hat{\mathbf{C}}_i = \mathbf{0}$, $\hat{r}_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$, and (3.3) becomes

$$\hat{\boldsymbol{\beta}}_{(i)}^1 = \hat{\boldsymbol{\beta}} - [\hat{\mathbf{X}}_{(i)}^{1\,T} \hat{\mathbf{X}}_{(i)}^1 + \hat{\mathbf{C}}]^{-1} \mathbf{x}_i \hat{r}_i. \tag{3.4}$$

When the data are complete, Equation (3.4) reduces to the least squares result given in Cook (1977, eq. 5).

The trade-off for saving the extra iterations is in the precision of the approximation. Figure 1 is a graph of log-likelihoods $L(\boldsymbol{\beta})$ and $L_{(i)}(\boldsymbol{\beta})$; of course

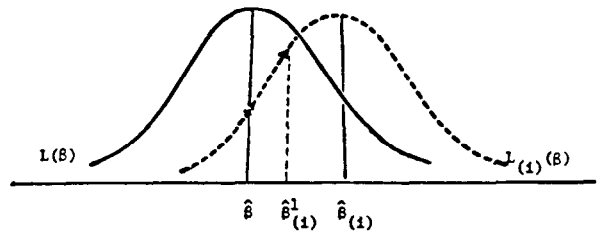

Figure 1. One-Step Approximation to $\hat{\beta}_{(i)}$, Using $\hat{\beta}$ as the Initial Estimate. $L(\beta)$ ——, log-likelihood function of β based on all of the cases; $L_{(i)}(\beta)$ ---, log-likelihood function of β based on all of the cases but case i.

these are multidimensional, so the figure is merely a device to suggest the behavior of these functions. In Figure 1 the one-step estimate is always moving toward the fully iterated estimate, $\hat{\boldsymbol{\beta}}_{(i)}$; that is, it is always moving in the right direction, since the increase of the likelihood in each iteration is guaranteed (Dempster et al. 1977, theorem 1). The one-step approximation may underestimate $\hat{\boldsymbol{\beta}}_{(i)}$ and might be improved by further iterative steps. For diagnostics where the purpose is to point out big individual effects, however, the one-step approximation might be an adequate tool for influence even if the one-step estimator is poor, as we shall see in Section 6. It should be mentioned that this type of one-step procedure occurs in other problems, such as nonlinear and generalized linear models, as given in several examples in Cook and Weisberg (1982, chap. 5).

4. FURTHER DISCUSSION AND COMPUTATIONAL CONSIDERATIONS

Equation (3.3) can be rewritten as

$$\hat{\boldsymbol{\beta}}^1_{(i)} = [I + (\hat{\mathbf{X}}^{1\,T}_{(i)}\hat{\mathbf{X}}^1_{(i)} + \hat{\mathbf{C}}^1_{(i)})^{-1}\hat{\mathbf{C}}_i]\hat{\boldsymbol{\beta}} - [\hat{\mathbf{X}}^{1\,T}_{(i)}\hat{\mathbf{X}}^1_{(i)} + \hat{\mathbf{C}}^1_{(i)}]^{-1}\hat{\mathbf{x}}_i\hat{r}_i,$$

where I is the $p \times p$ identity matrix. The complete and incomplete data results can be written in one form as

$$\boldsymbol{\beta}^{*1}_{(i)} = \mathbf{A}_i\boldsymbol{\beta}^* - \mathbf{d}_i, \tag{4.1}$$

where $\boldsymbol{\beta}^*$, $\mathbf{A}_i$, and $\mathbf{d}_i$ are as follows. When data are complete, $\boldsymbol{\beta}^*$ is the least squares estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Then

$$\boldsymbol{\beta}^{*1}_{(i)} = \tilde{\boldsymbol{\beta}}_{(i)}, \quad \mathbf{A}_i = \mathbf{I}, \quad \mathbf{d}_i = (\mathbf{X}^T_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{x}_i r_i.$$

When the data are incomplete, the MLE $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is used. For incomplete cases,

$$\boldsymbol{\beta}^{*1}_{(i)} = \hat{\boldsymbol{\beta}}^1_{(i)}$$
$$\mathbf{A}_i = \mathbf{I} + (\hat{\mathbf{X}}^{1\,T}_{(i)}\hat{\mathbf{X}}^1_{(i)} + \hat{\mathbf{C}}^1_{(i)})^{-1}\hat{\mathbf{C}}_i$$
$$\mathbf{d}_i = (\hat{\mathbf{X}}^{1\,T}_{(i)}\hat{\mathbf{X}}^1_{(i)} + \hat{\mathbf{C}}^1_{(i)})^{-1}\hat{\mathbf{x}}_i\hat{r}_i,$$

whereas for complete cases,

$$\mathbf{A}_i = \mathbf{I}$$
$$\mathbf{d}_i = (\hat{\mathbf{X}}^{1\,T}_{(i)}\hat{\mathbf{X}}^1_{(i)} + \hat{\mathbf{C}})^{-1}\hat{\mathbf{x}}_i\hat{r}_i.$$

Equation (4.1) states that the change of $\boldsymbol{\beta}^*$ to $\boldsymbol{\beta}^{*1}_{(i)}$ in general consists of rotation, stretch or shrinkage of some components of $\boldsymbol{\beta}^*$ by $\mathbf{A}_i$, and translation by $\mathbf{d}_i$. When data are complete or at a fully observed case in an incomplete data set, the only action is translation. The translation is proportional to the least squares residual or estimated residual. When not null, the rotation, stretch, and shrinkage seem to relate to the proportion of the correction $\hat{\mathbf{C}}_i$ made by removing the ith case. The definition of $\hat{\mathbf{C}}_i$ explains that this proportion depends on the covariances of the x's that are missing in the ith case.

A useful computation form for (3.3) is the following:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^1_{(i)} &= \hat{\boldsymbol{\beta}} - [\hat{\mathbf{X}}^{1\,T}_{(i)}\hat{\mathbf{X}}^1_{(i)} + \hat{\mathbf{C}}^1_{(i)}]^{-1}[\hat{\mathbf{x}}_i\hat{r}_i - \hat{\mathbf{C}}_i\hat{\boldsymbol{\beta}}] \\ &= \hat{\boldsymbol{\beta}} - [\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}}^1_{(i)} - \hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^T]^{-1}(\hat{\mathbf{x}}_i\hat{r}_i - \hat{\mathbf{C}}_i\hat{\boldsymbol{\beta}}) \\ &= \hat{\boldsymbol{\beta}} - \frac{[\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}}^1_{(i)}]^{-1}\hat{\mathbf{x}}_i\hat{r}_i}{1 - \hat{\mathbf{x}}_i^T[\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}}^1_{(i)}]^{-1}\hat{\mathbf{x}}_i} \\ &\quad + \left[\mathbf{I} + \frac{(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}}^1_{(i)})^{-1}\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^T}{1 - \hat{\mathbf{x}}_i^T[\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}}^1_{(i)}]^{-1}\hat{\mathbf{x}}_i}\right] \\ &\quad \times [\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}}^1_{(i)}]^{-1}\hat{\mathbf{C}}_i\hat{\boldsymbol{\beta}}. \end{aligned} \tag{4.2}$$

Since cases with the same missing variables will have the same correction matrix $\hat{\mathbf{C}}_i$ and the same matrix $\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}}^1_{(i)}$, matrix inversion can be done once for each pattern of incomplete data. Expression (4.2) corresponds to the complete data result in Cook (1977, between eqs. 6 and 7) with an extra correction term.

5. INFLUENCE NORMS

One goal of influence analysis is to identify those cases that give the largest change in a specific aspect of an analysis when a case is removed. The identified cases can then be studied individually. To rank cases on influence, we must define a norm of the vector $\hat{\boldsymbol{\beta}}_{(i)}$ or its one-step approximation $\hat{\boldsymbol{\beta}}^1_{(i)}$. Cook and Weisberg (1982, secs. 3.5 and 5.2) considered several methods for defining a norm. Because of the complexity of the incomplete data problem, there is no obvious way to do this. We consider two possibilities. They are both elliptical norms, defined for $\hat{\boldsymbol{\beta}}_{(i)}$ by

$$D_i(\mathbf{A}) = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T\mathbf{A}^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})$$

and for one-step estimators by

$$D^1_i(\mathbf{A}) = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^1_{(i)})^T\mathbf{A}^{-1}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^1_{(i)}).$$

The character of the norm is determined by the choice of $\mathbf{A}$.

In general, choices of $\mathbf{A}$ can refer to either an external reference scale or an internal scatter of the changes of $\hat{\boldsymbol{\beta}}$ to $\hat{\boldsymbol{\beta}}_{(i)}$ or $\hat{\boldsymbol{\beta}}^1_{(i)}$ for one-step estimators. For external norms, $\mathbf{A}$ can be chosen to be proportional

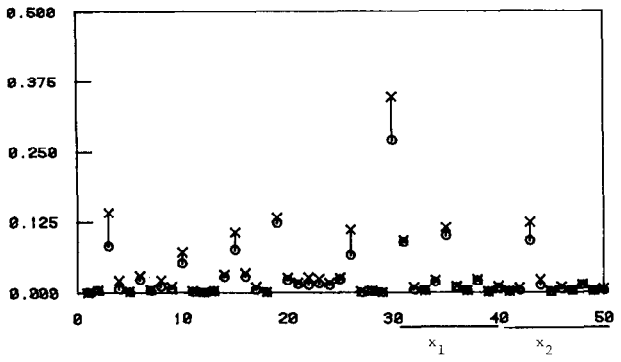

Figure 2. × = $D_i(\mathbf{A}_W)$ *and* ○ = $D_i^1(\mathbf{A}_W)$ *for the first simulated example. The underlining pattern gives the variables missing for the indicated case. Cases 31–40 are missing* x_1 *and 41–50 are missing* x_2.

to an estimate of the variance of $\hat{\boldsymbol{\beta}}$. For internal norms, the n values $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$ (or $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}^1$) are treated as an unstructured sample for which the Wilks (1963) multivariate outlier technique may be used to order the values. For the complete data problem, Cook's (external) measure is preferable because it is computationally simple and is made up of a few fundamental quantities, whereas computation of the internal norm is more complex. For the incomplete data problem, both measures are of about equal complexity, so the choice of one over the other is less clear.

Beale and Little (1975) and Little (1979) proposed the use of

$$\mathbf{A}_w = \hat{\sigma}^2 \mathbf{S}_W^{-1} = \hat{\sigma}^2 (\hat{\mathbf{X}}^T \hat{\mathbf{W}} \hat{\mathbf{X}})^{-1} \tag{5.1}$$

as an estimate of var($\hat{\boldsymbol{\beta}}$) in the incomplete data problem, where $\hat{\mathbf{W}}$ is a diagonal matrix with entries equal to the ratio of the estimated residual variance of y given $\mathbf{x}$ to the estimated residual variance of y given the observed part of $\mathbf{x}$ for the ith case, $w_i = \sigma / \sigma_{y \cdot U_i}$. Here w_i is bounded between 0 and 1 and can be thought of as a completeness index. The first norm for one-step estimators, for $i = 1, 2, \ldots, n$, is

$$D_i^1(\mathbf{A}_W) = \frac{1}{p\hat{\sigma}^2} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}^1)^T (\hat{\mathbf{X}}^T \hat{\mathbf{W}} \hat{\mathbf{X}}) (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}^1), \tag{5.2}$$

where p = number of elements in $\boldsymbol{\beta}$. If we define $\hat{\mathbf{Y}}_{(i)}^1 = \hat{\mathbf{X}} \hat{\boldsymbol{\beta}}_{(i)}^1$, with elements $\hat{y}_{(i)j}^1, j = 1, \ldots, n$, this measure can be written as

$$D_i^1(\mathbf{A}_W) = \sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{(i)j}^1)^2 / p\hat{\sigma}_{y \cdot U_j}^2, \tag{5.3}$$

a weighted sum of the squared changes in the fitted values. Unlike Cook's distance, $D_i^1(\mathbf{A}_W)$ does not factor into a fixed part (proportional to the diagonals of a projection matrix) and a random part. As Cook and Weisberg (1982, p. 137) remarked, the factorization may be unique to complete-data, one-case-at-a-time deletion. Hence computation of $D_i^1(\mathbf{A}_W)$ is more intensive than is the comparable statistic for complete data.

Both the one-step and the fully iterated external norms use $p\mathbf{A}_W^{-1}$ to define the inner product. Since $\boldsymbol{\beta}_{(i)}^1$ is moving from $\hat{\boldsymbol{\beta}}$ toward $\hat{\boldsymbol{\beta}}_{(i)}$, we can expect that usually $D_i^1(\mathbf{A}_W) \leq D_i(\mathbf{A}_W)$, and $D_i^1(\mathbf{A}_W)$ may be much smaller. Using this norm the one-step measure provides a lower bound for the fully iterated measure.

The second norm is defined by the internal scatter of $\boldsymbol{\Delta}_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}$ or the $\boldsymbol{\Delta}_i^1 = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}^1$. A norm can be defined using Wilks's (1963) statistic for a multivariate outlier,

$$(\boldsymbol{\Delta}_i - \bar{\boldsymbol{\Delta}})^T \mathbf{A}_I^{-1} (\boldsymbol{\Delta}_i - \bar{\boldsymbol{\Delta}}), \tag{5.4}$$

where

$$\bar{\boldsymbol{\Delta}} = (1/n) \Sigma \boldsymbol{\Delta}_i$$

$$\mathbf{A}_I = \Sigma (\boldsymbol{\Delta}_i - \bar{\boldsymbol{\Delta}})(\boldsymbol{\Delta}_i - \bar{\boldsymbol{\Delta}})^T.$$

For the complete data problem, $\mathbf{A}_I^{-1}$ is closely related to the weighted jackknife estimate (Hinkley 1977) of var($\hat{\boldsymbol{\beta}}$). For the incomplete data problem, this estimate may be poor if the one-step estimates are poor.

Substituting for $\boldsymbol{\Delta}_i$ in (5.4) and simplifying will give the measures. Generally, the correction for centering at $\bar{\boldsymbol{\Delta}}$ rather than $\mathbf{0}$ is negligible and may be dropped for simplicity. Then

$$D_i(\mathbf{A}_I) = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T \times \left[\sum_{i=1}^{n} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})^T \right]^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}) \tag{5.5}$$

and

$$D_i^1(\mathbf{A}_I^1) = (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}^1)^T \times \left[\sum_{i=1}^{n} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}^1)(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}^1)^T \right]^{-1} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}^1). \tag{5.6}$$

As defined, $D_i(\mathbf{A}_I)$ and $D_i^1(\mathbf{A}_I^1)$ are bounded between 0 and 1, with large values corresponding to unusual rows.

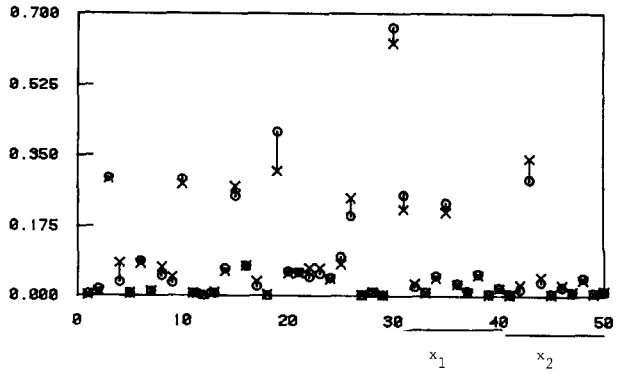

Figure 3. × = $D_i(\mathbf{A}_I)$ *and* ○ = $D_i^1(\mathbf{A}_I^1)$ *for the first simulated example.*

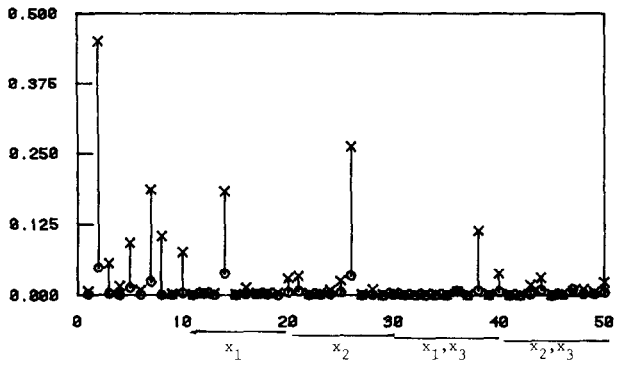

Figure 4. $\times = D_i(\mathbf{A}_W)$ *and* $\circ = D_i(\mathbf{A}_W)$ *for the second simulated example.*

Unlike the externally scaled measures, (5.5) and (5.6) each use a different inner product; hence $D_i^1(\mathbf{A}_I^1)$ need not be smaller than $D_i(\mathbf{A}_I)$. Depending on circumstances, this could prove either an advantage or a disadvantage.

Other choices of $\mathbf{A}$ to define the norm include appropriate submatrices of the expected information (Hartley and Hocking 1971) or of the observed information (Louis 1982). These may lead to different ordering of cases on influence, but the approach is similar to that given here.

6. NUMERICAL EXAMPLES

First, we present simulated examples to explore the agreement between one-step and fully iterated estimates and between internal and external norms. This example is part of a larger simulation study by Shih (1981). A variety of regression situations were studied, all with $p = 3$ and $n = 50$, using a design also used by Little (1979). Here we report typical results for two extreme cases, one in which the one-step estimators match the fully iterated estimates very closely and one in which they do not match as well.

In the first set, data were generated from the model

$$y = x_1 + .5x_2 + .4x_3 + e,$$

where

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \sim N_3 \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.00 & .90 & .86 \\ .90 & 1.00 & .78 \\ .86 & .78 & 1.00 \end{pmatrix} \right],$$

$e \sim N(0, .5)$, with $\mathbf{x}$ and e independent. After generation of the data, x_1 was deleted from cases 31–40 and x_2 was deleted from cases 41–50.

For this setup, agreement between fully iterated and one-step distance measures is excellent. Figures 2 and 3 summarize the results for external scaling and internal scaling, respectively. In each figure, the crosses indicate the fully iterated estimates and the circles indicate the one-step estimates. The longer the line segment joining the cross to the circle, the worse the agreement. Case 30, the candidate for the most influential case, is clearly indicated by either measure. We would be led to consider this case carefully.

The second example is more severely incomplete and correlations are lower. The model is the same as for the first example, except $\sigma^2 = 2$;

$$\boldsymbol{\Sigma}_{xx} = \begin{pmatrix} 1.00 & .14 & -.15 \\ .14 & 1.00 & .25 \\ -.15 & .25 & 1.00 \end{pmatrix};$$

and cases 1–10 are complete, 11–20 are missing x_1, 21–30 are missing x_2, 31–40 are missing x_1 and x_3, and 41–50 are missing x_2 and x_3. The results are summarized in Figures 4 and 5. Agreement of one-step and fully iterated measures is not nearly as close in these figures, especially for cases with relatively large influence. For detecting such cases, however, the one-step measures are satisfactory. For example, in Figure 4, case 2 has the largest $D_i(\mathbf{A}_W)$, and although $D_2^1(\mathbf{A}_W)$ is much less than $D_2(\mathbf{A}_W)$, it is the largest one-step influence measure. Both figures would suggest more or less the same cases for further analysis, namely cases 2, 7, 14, and 26.

In both examples, the complete cases, 1–30 in Figures 2 and 3 and 1–10 in Figures 4 and 5, generally show larger values for the influence statistics than the incomplete cases. Because of the methods of generating these data, no overly influential cases are expected and none seem to be apparent here. Yet these examples suggest that complete cases will tend to be more influential in fitting models, and one-step measures will generally detect the most influential ones.

For a final example, we consider the following problem. Endogenous creatinine (CR) clearance is an important measure of renal function. Although

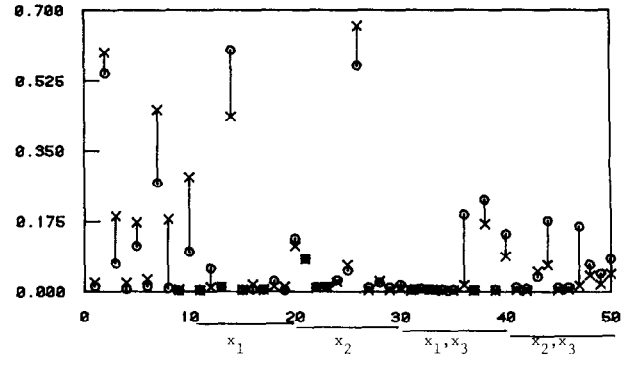

Figure 5. $\times = D_i(\mathbf{A}_I)$ *and* $\circ = D_i^1(\mathbf{A}_I^1)$ *for the second simulated example.*

Table 1. Data for the Creatinine Clearance Example

	Weight	SC	Age	CR
1	71.0	.71253	38	132.0
2	69.0	1.48161	78	53.0
3	85.0	2.20545	69	50.0
4	100.0	1.42505	70	82.0
5	59.0	.67860	45	110.0
6	73.0	.75777	65	100.0
7	63.0	1.11969	76	68.0
8	81.0	.91611	61	92.0
9	74.0	1.54947	68	60.0
10	87.0	.93873	64	94.0
11	79.0	.99528	66	105.0
12	93.0	1.07445	49	98.0
13	60.0	.70122	43	112.0
14	70.0	.71253	42	125.0
15	83.0	.99528	66	108.0
16	70.0	2.52212	78	30.0
17	73.0	1.13100	35	111.0
18	85.0	1.11969	34	130.0
19	68.0	1.37982	35	94.0
20	65.0	1.11969	16	130.0
21	53.0	.97266	54	59.0
22	50.0	1.60602	73	38.0
23	74.0	1.58339	66	65.0
24	67.0	1.40244	31	85.0
25	80.0	.67860	32	140.0
26	67.0	1.19886	21	80.0
27	68.0	7.60001	81	4.3
28	72.2	2.10001	43	43.2
29	NA	1.35719	78	75.0
30	NA	1.05183	38	41.0
31	107.0	NA	62	120.0
32	75.0	NA	70	52.0
33	62.0	NA	63	73.0
34	52.0	NA	68	57.0

NOTE: NA—not available.

measurement of this quantity is inexpensive in humans, it is difficult to use in a clinical setting because it requires 24-hour urine collections. Consequently, it is usual to model CR as a function of easily collected information, typically *serum* creatinine (SC) concentration, in mg/deciliter, body weight (WT) in kg, age in years, and sex. For estimating CR, many pharmacokinetics textbooks recommend using a model of the form

$$E(\log(\mathrm{CR})) = \beta_0 + \beta_1 \log(\mathrm{WT}) + \beta_2 \log(\mathrm{SC}) + \beta_3 \log(140 - \mathrm{age}),$$

where β_0 will be different for males and for females.

Table 2. $D_i(\boldsymbol{A}_w)$ for All i With Relatively Large Values

Case	n = 34	n = 33
20	.20	.32
22	.15	.06
27	4.76	—
28	.004	.41
30	.29	.49

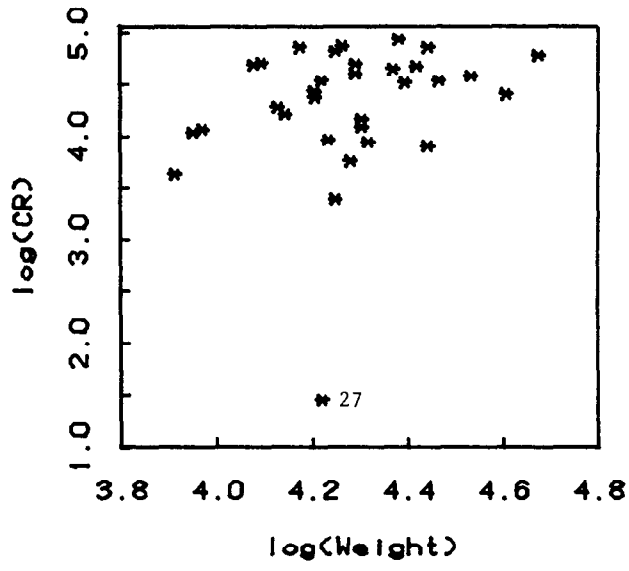

Figure 6. Scatterplot for Creatinine Clearance Example.

The data in Table 1 are from a clinical trial conducted overseas by Merck Sharp and Dohme Research Laboratories. Of 34 male patients, 2 had no record of WT, and 4 were missing SC. No apparent reason for these data to be missing was available. Because the fraction of data with incomplete information, .18, is so large, incorporation of the incomplete data into the analysis seems desirable. If we can make the assumption of joint multivariate normality of the response and the three predictors, then the incomplete data methodology of this article can be applied.

For these data, almost identical inferences are obtained if one-step or fully iterated measures are com-

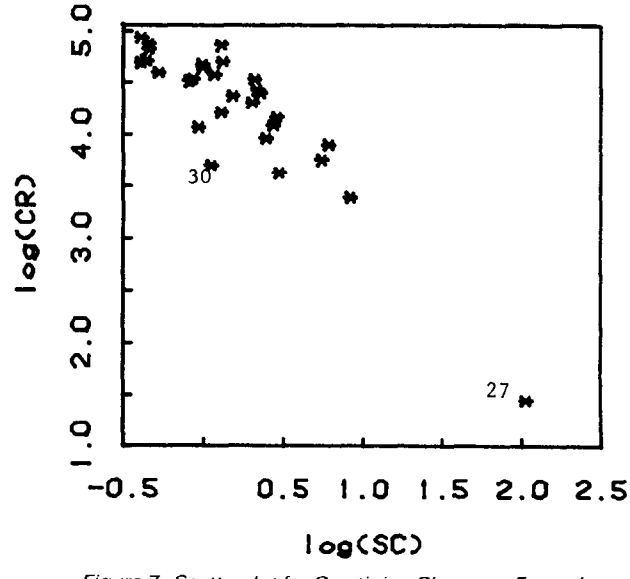

Figure 7. Scatterplot for Creatinine Clearance Example.

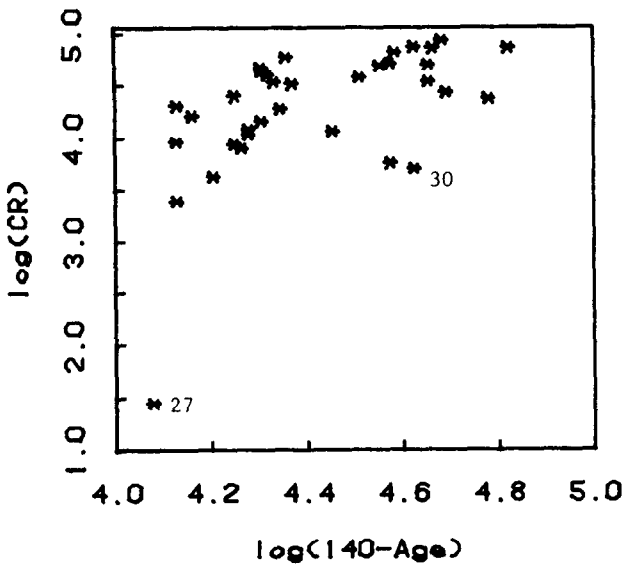

Figure 8. Scatterplot for Creatinine Clearance Example.

puted, for either choice of norm, so for clarity, in Table 2 we present only $D_i(\mathbf{A}_W)$; for all cases not given in the table, $D_i(\mathbf{A}_W) < .11$. One case, 27 in Table 1, is clearly far removed from the others in the data and is strongly influential. The second most influential case, 30, is one of the partially observed cases that is missing WT.

Figures 6, 7, and 8 are scatterplots of the response versus each of the predictors, with cases 27 and 30 marked on the plot. The importance of case 27 is easily predictable from the plot. The importance of case 30, however, may have been overlooked without this analysis; Figure 6 does not contain case 30.

Later discussion with the supervising physician revealed that case 27 had high-grade renal impairment and could therefore be expected to be quite different from the other patients. Deletion of case 27 seems to be appropriate for this analysis. When the inappropriate case 27 is deleted from the data, the resulting fitted model is shown in Table 3. The largest change seems to be in estimated residual variance and in the coefficient for log(SC). Influence values for selected cases are given in Table 2 and show that the incomplete case 30 is now the most influential, but it is only of modest influence. This case was low for each of its observed predictors. Of course we do not know the third predictor, and this may well have "explained" the unusual response for case 30. The careful investigator, however, would be alerted to the possible special interest in this case.

7. DISCUSSION

Detection of influential cases in regression analysis has been proven important in principle and useful in practice in the past for complete data problems. In the situation of incomplete data, there is no reason for the usefulness to diminish. Detecting influential cases becomes more interesting because of additional new questions. For example, one might ask whether an incomplete case can be influential. Since the missing values are replaced by conditional expectations and tend to move cases to the center of the data, the lack of influence of incomplete cases can be expected in general. From the real data example, however, we also have seen that incomplete cases can be relatively influential.

Automatic deletion of incomplete cases has been argued to be not desirable when the portion of incomplete cases is relatively large or in situations such as clinical trials in which deletion of any cases may involve regulatory discussion. The example suggests that automatic deletion of incomplete cases may even lose important information in the sense of modeling, since they may be influential.

We have presented two choices for influence norms—external and internal. Although the two norms provided similar ordering of cases with respect to influence in our examples, one should not expect this similarity to hold in general. We have found it helpful to look at both norms and view the information from them as complementary. If only one norm is to be chosen, we tentatively recommend the external norm because of the agreement between fully iterated and one-step procedures using it. The

Table 3. Estimates and Standard Errors

	Complete Data Only				*All Data*			
	All cases		*Case 27 deleted*		*All cases*		*Case 27 deleted*	
	Estimate	*SE*	*Estimate*	*SE*	*Estimate*	*SE*	*Estimate*	*SE*
Intercept	−2.97	1.51	−2.68	.99	−3.33	1.20	−3.24	1.08
log(WT)	.99	.26	.90	.17	1.17	.21	−1.07	.15
log(SC)	−1.08	.09	−.78	.08	−1.08	.31	−.77	.10
log(140-age)	.73	.21	.75	.14	.65	.21	−.71	.18
$\hat{\sigma}^2$	.044		.019		.045		.019	

NOTE: SE is standard error.

choice of influence norms in this and other diagnostic problems, however, is not a completely settled issue.

The techniques discussed here extend influence analysis to incomplete data problems. The extension is one more step toward making regression analysis with incomplete data useful for the data analyst. Because of the additional complexity of these problems, there are many problems for further research, such as calibration with respect to some external standard. Other problems are inherited from the maximum likelihood estimation, including filling in missing values and the normality assumption. For filling in missing values, a possible problem is the likely sensitivity (nonrobustness) of the filled-in values, since these are conditional expectations. This may not be much of a problem, however, because the filling in is just a convenient computational procedure; in fact, we integrate out the unobserved values, suggesting that this nonrobustness is more apparent than real.

The normality assumption for the missing X's is used explicitly in several places. If the X's were nonnormal, one might ask if the approach taken here will give reasonable results. The answer depends on how well the true log-likelihood is approximated by the normal log-likelihood. If the true log-likelihood for $\boldsymbol{\beta}$ is approximately quadratic, one might expect this method to work acceptably; if it is not quadratic, it is likely to do poorly. In either case, we have seen that the one-step norms tend to point out influential cases.

ACKNOWLEDGMENTS

We thank the editor, the associate editor, and the referees for their helpful comments. A computer program for the methods outlined in this article may be obtained from W. J. Shih.

APPENDIX: DERIVATION OF EQUATION (3.3)

To derive (3.3), we first write

$$\hat{\boldsymbol{\beta}} = (\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}})^{-1}\hat{\mathbf{X}}^T\mathbf{Y} = (\hat{\mathbf{X}}_{(i)}^{1T}\hat{\mathbf{X}}_{(i)}^{1} + \hat{\mathbf{C}}_{(i)}^{1} + \hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^T + \hat{\mathbf{C}}_i)^{-1}(\hat{\mathbf{X}}_{(i)}^{1T}\mathbf{Y}_{(i)} + \hat{\mathbf{x}}_i y_i). \quad (A.1)$$

Let $\hat{\mathbf{C}}_i = \mathbf{M}_i\mathbf{M}_i^T$ be a square root decomposition of the symmetric matrix $\hat{\mathbf{C}}_i$. We write

$$\hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^T + \hat{\mathbf{C}}_i = (\hat{\mathbf{x}}_i, \mathbf{M}_i)(\hat{\mathbf{x}}_i, \mathbf{M}_i)^T$$

and then apply the following matrix identity to the first factor on the right side of (A.1):

$$(\mathbf{E} + \mathbf{G}^T\mathbf{H})^{-1} = \mathbf{E}^{-1} - \mathbf{E}^{-1}\mathbf{G}^T(\mathbf{I} + \mathbf{G}\mathbf{E}^{-1}\mathbf{H}^T)^{-1}\mathbf{H}\mathbf{E}^{-1}.$$

[A proof and history of a generalization of this identity was given by Henderson and Searle (1981).] Then

$$(\hat{\mathbf{X}}_{(i)}^{1T}\hat{\mathbf{X}}_{(i)}^{1} + \hat{\mathbf{C}}_{(i)}^{1} + \hat{\mathbf{x}}_i\hat{\mathbf{x}}_i^T + \hat{\mathbf{C}}_i)^{-1} = (\hat{\mathbf{X}}_{(i)}^{1T}\hat{\mathbf{X}}_{(i)}^{1} + \hat{\mathbf{C}}_{(i)}^{1})^{-1} - (\hat{\mathbf{X}}_{(i)}^{1T}\hat{\mathbf{X}}_{(i)}^{1} + \hat{\mathbf{C}}_{(i)}^{1})^{-1}(\hat{\mathbf{x}}_i, \mathbf{M}_i)(\mathbf{I} + \mathbf{U})^{-1}(\hat{\mathbf{x}}_i, \mathbf{M}_i)^T \times (\hat{\mathbf{X}}_{(i)}^{1T}\hat{\mathbf{X}}_{(i)}^{1} + \hat{\mathbf{C}}_{(i)}^{1})^{-1}, \quad (A.2)$$

where

$$\mathbf{U} = (\hat{\mathbf{x}}_i, \mathbf{M}_i)^T(\hat{\mathbf{X}}_{(i)}^{1T}\hat{\mathbf{X}}_{(i)}^{1} + \hat{\mathbf{C}}_{(i)}^{1})^{-1}(\hat{\mathbf{x}}_i, \mathbf{M}_i).$$

Substituting (A.2) into (A.1) and writing

$$\hat{\mathbf{x}}_i y_i = (\hat{\mathbf{x}}_i, \mathbf{M}_i)\begin{pmatrix} y_i \\ \mathbf{0} \end{pmatrix}$$

in (A.1), we have

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{(i)}^{1} - (\hat{\mathbf{X}}_{(i)}^{1T}\hat{\mathbf{X}}_{(i)}^{1} + \hat{\mathbf{C}}_{(i)}^{1})^{-1}(\hat{\mathbf{x}}_i, \mathbf{M}_i) \times \left[(\mathbf{I} + \mathbf{U})^{-1}\begin{pmatrix} \hat{\mathbf{x}}_i^T \\ \mathbf{M}_i^T \end{pmatrix}\hat{\boldsymbol{\beta}}_{(i)}^{1} - [\mathbf{I} - (\mathbf{I} + \mathbf{U})^{-1}\mathbf{U}]\begin{pmatrix} y_i \\ \mathbf{0} \end{pmatrix}\right]. \quad (A.3)$$

Denote $\hat{r}_i^* = y_i - \hat{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}}_{(i)}^{1}$. The following facts will be shown immediately after the result:

$$\mathbf{I} - (\mathbf{I} + \mathbf{U})^{-1}\mathbf{U} = (\mathbf{I} + \mathbf{U})^{-1}, \quad (A.4)$$

and

$$(\mathbf{I} + \mathbf{U})^{-1}\begin{pmatrix} \hat{r}_i^* \\ -\mathbf{M}_i^T\hat{\boldsymbol{\beta}}_{(i)}^{1} \end{pmatrix} = \begin{pmatrix} \hat{r}_i \\ -\mathbf{M}_i^T\hat{\boldsymbol{\beta}} \end{pmatrix}. \quad (A.5)$$

Applying (A.4) and (A.5), (A.3) can be simplified to

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{(i)}^{1} + (\hat{\mathbf{X}}_{(i)}^{1T}\hat{\mathbf{X}}_{(i)}^{1} + \hat{\mathbf{C}}_{(i)}^{1})^{-1}(\hat{\mathbf{x}}_i\hat{r}_i - \hat{\mathbf{C}}_i\hat{\boldsymbol{\beta}}).$$

Result (A.4) follows by multiplying both sides of (A.4) by $(I + U)$ and simplifying. Result (A.5) requires two intermediate results, which follow from (A.2) and (A.4):

$$(\hat{\mathbf{x}}_i, \mathbf{M}_i)^T(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}})^{-1} = (\mathbf{I} + \mathbf{U})^{-1}(\hat{\mathbf{x}}_i, \mathbf{M}_i)(\hat{\mathbf{X}}_{(i)}^{1T}\hat{\mathbf{X}}_{(i)} + \mathbf{C}_{(i)}^{1})^{-1} \quad (A.6)$$

and

$$\mathbf{I} - (\hat{\mathbf{x}}_i, \mathbf{M}_i)^T(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}})^{-1}(\hat{\mathbf{X}}_i, \mathbf{M}_i) = (\mathbf{I} + \mathbf{U})^{-1}. \quad (A.7)$$

Then (A.5) is proved as follows:

$$\begin{pmatrix} \hat{r}_i \\ -\mathbf{M}_i^T\hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} y_i - \hat{\mathbf{x}}_i^T\hat{\boldsymbol{\beta}} \\ -\mathbf{M}_i^T\hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} y_i \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \hat{\mathbf{x}}_i^T \\ \mathbf{M}_i^T \end{pmatrix}[(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}})^{-1}\hat{\mathbf{X}}^T\mathbf{Y}].$$

Write $\hat{\mathbf{X}}^T\mathbf{Y} = \hat{\mathbf{X}}_{(i)}^{1T}\mathbf{Y}_{(i)} + \hat{\mathbf{x}}_i y_i$, $\hat{\mathbf{x}}_i y_i = (\hat{\mathbf{x}}_i, \mathbf{M}_i)\binom{y_i}{0}$. Then

$$\begin{pmatrix} \hat{r}_i \\ -\mathbf{M}_i^T\hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} y_i \\ 0 \end{pmatrix} - (\hat{\mathbf{x}}_i, \mathbf{M}_i)^T(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}})^{-1}\hat{\mathbf{X}}_{(i)}^{1T}\mathbf{Y}_{(i)}$$

$$
\begin{aligned}
&- (\hat{\mathbf{x}}_i, \mathbf{M}_i)^T(\hat{\mathbf{X}}^T\hat{X} + \hat{\mathbf{C}})^{-1}(\hat{\mathbf{x}}_i, \mathbf{M}_i)\begin{pmatrix} y_i \\ 0 \end{pmatrix} \\
&= -(\hat{\mathbf{x}}_i, \mathbf{M}_i)^T(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}})^{-1}\hat{\mathbf{X}}_{(i)}^{1T}\mathbf{Y}_{(i)} \\
&\quad + [\mathbf{I} - (\hat{\mathbf{x}}_i, \mathbf{M}_i)^T(\hat{\mathbf{X}}^T\hat{\mathbf{X}} + \hat{\mathbf{C}})^{-1}(\hat{\mathbf{x}}_i, \mathbf{M}_i)]\begin{pmatrix} y_i \\ 0 \end{pmatrix} \\
&= -(\mathbf{I} + \mathbf{U})^{-1}(\hat{\mathbf{x}}_i, \mathbf{M}_i)^T(\hat{\mathbf{X}}_{(i)}^{1T}\hat{\mathbf{X}}_{(i)}^{1} \\
&\quad + \mathbf{C}_{(i)}^{1})^{-1}\hat{\mathbf{X}}_{(i)}^{1T}\mathbf{Y}_{(i)}[\mathbf{I} + \mathbf{U}]^{-1}\begin{pmatrix} y_i \\ 0 \end{pmatrix} \\
&\qquad\qquad \text{[from (A.6) and (A.7)]} \\
&= (\mathbf{I} + \mathbf{U})^{-1}\left[\begin{pmatrix} y_i \\ \mathbf{0} \end{pmatrix} - (\hat{\mathbf{x}}_i, \mathbf{M}_i)^T\hat{\boldsymbol{\beta}}_{(i)}^{1}\right] \\
&= (\mathbf{I} + \mathbf{U})^{-1}\begin{pmatrix} \hat{r}_i^* \\ -\mathbf{M}_i^T\hat{\boldsymbol{\beta}}_{(i)}^{1} \end{pmatrix}.
\end{aligned}
$$

[*Received March 1984. Revised January 1986.*]

REFERENCES

Beale, E. M. L., and Little, R. J. A. (1975), "Missing Values in Multivariate Analysis," *Journal of the Royal Statistical Society*, Ser. B, 37, 129–146.

Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18 (additional correspondence, 348–350).

——— (1979), "Influential Observations in Linear Regression," *Journal of the American Statistical Association*, 74, 169–174.

Cook, R. D., and Weisberg, S. (1980), "Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression," *Technometrics*, 22, 495–508.

——— (1982), *Residuals and Influence in Regression*, New York: Chapman & Hall.

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

Fox, T., Hinkley, D., and Larntz, K. (1980), "Jackknifing in Nonlinear Regression," *Technometrics*, 22, 29–33.

Hartley, H. O., and Hocking, R. R. (1971), "The Analysis of Incomplete Data" (with discussion), *Biometrics*, 27, 783–823.

Henderson, H. V., and Searle, S. R. (1981), "On Deriving the Inverse of a Sum of Matrices," *SIAM Review*, 23, 53–60.

Hinkley, D. V. (1977), "Jackknifing in Unbalanced Situations," *Technometrics*, 19, 285–292.

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

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, 76–88.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 44, 226–233.

Press, S. J., and Scott, A. (1974), "Missing Variables in Bayesian Regression," in *Studies in Bayesian Econometrics and Statistics*, eds. S. Fienberg and A. Zellner, Amsterdam: North-Holland, pp. 259–272.

Rubin, D. B. (1974), "Characterizing the Estimation of Parameters in Incomplete Data Problems," *Journal of the American Statistical Association*, 69, 467–474.

——— (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592.

Shih, W. J. (1981), "Case Analysis of Multiple Linear Regression With Incomplete Data," unpublished Ph.D. thesis, University of Minnesota, Dept. of Statistics.

Weisberg, S. (1983), "Some Principles for Regression Diagnostics and Influence Analysis," discussion of "Developments in Linear Regression Methodology: 1959–1982," by R. R. Hocking, *Technometrics*, 25, 240–244.

Wilks, S. S. (1963), "Multivariate Statistical Outliers," *Sankhya*, Ser. A, 25, 407–426.