APPLIED STOCHASTIC MODELS AND DATA ANALYSIS, VOL. 7, 213–235 (1991)

AN IMPUTATION METHOD FOR DEALING WITH MISSING DATA IN REGRESSION

LEE G. COOPER, JAN DE LEEUW AND ARAM G. SOGOMONIAN
John E. Anderson Graduate School of Management, U.C.L.A. Los Angeles, California 90024-1481, U.S.A.

SUMMARY

We develop two methods for imputing missing values in regression situations. We examine the standard fixed-effects linear-regression model $y = X\beta + \varepsilon$, where the regressors X are fixed and ε is the error term. This research focuses on the problem of missing X values. A particular component of market-share analysis has motivated this research where the price and other promotional instruments of each brand are allowed to have their own impact on the total sales volume in a consumer-products category. When a brand is not distributed in a particular week, only a few of the many measures occurring in that observation are missing. 'What values should be imputed for the missing measures?' is the central question this paper addresses. This context creates a unique problem in the missing-data literature, i.e. there is no true value for the missing measure. Using influence functions, from robust statistics we develop two *loss functions*, each of which is a function of the missing and existing X values. These loss functions turn out to be sums of ratios of low-order polynomials. The minimization of either loss function is an unconstrained non-linear-optimization problem. The solution to this non-linear optimization leads to imputed values that have minimal influence on the estimates of the parameters of the regression model. Estimates using the method for replacing missing values are compared with estimates obtained via some conventional methods.

KEY WORDS Imputation Missing data Influence statistics Regression Market-response models

1. INTRODUCTION

In this paper we investigate an alternative method for dealing with missing data in fixed-effects linear-regression models. This problem arises in a number of contexts, but we develop our solution within the context of modelling brand sales. As described by Cooper and Nakanishi[1] brand sales are often modelled as a function of a market share and a category-volume component—two separable processes:

$$\text{Sales for brand } i = \text{Market share for brand } i \times \text{Category volume}$$

The appropriate technique for handling missing data in category-volume models has been a source of concern. For a brand-planning effort we would like to know the ability of each brand to expand (or shrink) the total store sales in a consumer-products category. Since most consumers may easily choose among stores, it is reasonable to assume that the same category-volume model applies to all stores in a trading area. However, each store can have at least a slightly different portfolio of brands stocked. The number of observations would be approximately the number of stores times the number of weeks; but since only a few, if any, stores are likely to stock *all* the brands in a category, many of the observations will have at

8755–0024/91/030213–23$11.50

Received 19 September 1989
Revised 1 August 1990

least a few missing measures. In the coffee-market case, which motivated this research (cf. Cooper and Nakanishi[1]), there were 234 observations, 179 of which had at least one missing brand. These missing data points are not generated randomly, but rather are the result of (non-random) store policy. Whereas procedures exist to impute missing values when they are randomly generated, in a typical case missing values generated by non-random processes are simply 'dropped' from statistical analyses. With this treatment, the entire weekly observation, including the valid values of other brands' marketing instruments, is discarded. Obviously, valuable information is being thrown out.

Note that the missing data do not necessarily take the value zero. For example, if a brand is not distributed then it is not displayed and a value of zero for this categorical variable is fine. However, for a linear price term, a value of zero would incorrectly imply that a particular brand of coffee was sold for free during a given week! In this circumstance there is *no true price* that we can put in the missing datum's place. Similarly for an asymmetric market-share model (Cooper and Nakanishi[1]) in which we assess the sensitivity of brands A's market share to changes in brand B's price, if brand B is not distributed in a particular week it would have no *true* price that might be divined and put in the place of the missing price.

We propose new techniques that differ from previous procedures developed in the statistical literature. We believe that our treatment will best *minimize* the effect of missing data on the parameter estimates of the category-volume model or asymmetric market-share model, while not discarding useful information.

The balance of this paper is organized as follows: In Section 1.1 the general notation is developed. In Section 2 we review relevant research on the treatment of missing predictor variables in estimating statistical models. Section 3 discusses how the concept of influence functions can be used to develop an alternative approach to the missing-data problem. The computer implementation for our algorithm is described in Section 4. Section 5 presents the conclusion.

1.1. Notation

It will be convenient to make occasional reference to the following linear model:

$$y = X\beta + \varepsilon \tag{1}$$

Where y is an $n \times 1$ vector of observations, X is an $n \times m$ data matrix, β is a vector of m regression coefficients to be estimated and ε is an $n \times 1$ vector of errors. The category-volume model we are interested in is a form (when log-linearized) of this general linear model.

The specific category-volume model we are considering (for example, see p. 152 of Cooper and Nakanishi[1]) is as follows:

$$\log T_{st} = \sum_{j=1}^{J} \beta_{p_j} \log p_{jst} + \sum_{k=1}^{K} \sum_{j=1}^{J} \beta_{kj} x_{kjst} + e_{st} \tag{2}$$

Note if the only marketing instrument used is price then the category-volume model reduces to

$$\log T_{st} = \sum_{j=1}^{J} \beta_{p_j} \log p_{jst} + e_{st} \tag{3}$$

where T_{st} is the category volume at store s in week t ($st = 1, 2, \ldots, n$); $j = 1, 2, \ldots, J$ is the number of brands; $k = 1, 2, \ldots, K$ is the number of marketing instruments; p_{jst} is the price of the jth brand in store s at time t; x_{jkst} is 1 if the kth marketing instrument (e.g. newspaper

feature or in-store display) is used for the jth brand in observation st and 0 otherwise; the β's are the corresponding regression coefficients; and e_{st} is an error term.

2. RELATED LITERATURE

In this section we summarize the previous research on incomplete-data problems. In addition, the concept of an influence is discussed. While influence functions have been commonly used in data analysis, their use in methods for replacing missing data has not been considered previously. In positioning this paper it is important to discuss earlier missing-data-replacement techniques in some detail, so that the reader is aware of some of the restrictive assumptions and limitations of these techniques. We conclude this section with a description of influence functions and suggest how they may be applied in the analysis of missing data.

2.1. Early work

The statistics literature provides the main body of work that deals with missing-data problems in model estimation. Much of the preliminary work has been described by Afifi and Elashoff.[2,3] These authors list several approaches, each of which attempts to provide a single estimate of β. The simplest method is to drop (delete) observations that contain missing values. The other techniques avoid dropping observations with missing data points in order to retain as much information as possible. The latter methods are consistent with our approach.

Two types of method which avoid dropping observations have been discussed by Afifi and Elashoff. The first is a modified least-squares approach for substituting values for missing-data points. In this method, a multivariate-normal random-effects model is assumed. The second type of method uses maximum-likelihood techniques to estimate the covariance structure of X with non-missing values. Then the parameters of interest are estimated by least-squares procedures.

When using either of the last two mentioned procedures, one of the key underlying assumptions is that the data points are missing at random (MAR). This means that the pattern of missing values is assumed to be a random process and does not depend on observed or unobserved values. Estimates can be severely biased if this assumption is violated (Little and Rubin,[4] Simon and Simonoff[5]). Current statistical-computing software makes wide use of these methods, however Little and Rubin[4] provide more general techniques and recommend not using the methods proposed by Afifi and Elashoff[2,3] except when only a small amount of data is missing.

2.2. Imputation-based procedures

Another approach to the missing-data problem involves using imputation procedures. These procedures fill in the missing values and the resulting completed data are analysed by standard methods. Two commonly used techniques are the following:

(i) Imputing unconditional means, where missing values in a column of X are replaced by the average of the non-missing values in the same column.
(ii) Imputing conditional means (Buck's method) where the sample mean and covariance matrix are estimated from the present data. Next, these estimates are used to calculate the linear regressions of the missing variables on the present variables. The observed

values of the present variables are substituted in the regressions (case by case) which yields predictions for the missing values (in the case).

Missing-data techniques when the missing observations are in the dependent variable have been thoroughly discussed in the marketing literature (Malhotra[6]) and covered by Little and Rubin.[4]

To summarize, one of the key assumptions which must be made when using these types of procedure (i.e. those of Little and Rubin[4]) is that the missing predictor data is MAR. Little and Rubin also mention that it is preferable for the data to be 'completely missing at random'. CMAR data is composed of data that is both MAR and observed at random—OAR.

In Chapter 8, Section 4, of Little and Rubin[4] the authors discuss linear regression with missing values in the predictor variables. This corresponds to our problem in that the X matrix in (1) may have missing values in a certain column because the brand corresponding to that column was not sold in a particular store for a given week. The authors make use of an expectation-maximization (EM) algorithm to get maximum-likelihood (ML) estimates of the β vector in (1) and the corresponding variances. By partitioning the X matrix into portions with and without missing data, a mechanism is provided to estimate the covariance matrix. Recall that in our case we are concerned with many observations that each have a few missing values. The partitioning method would place a great emphasis on the few (atypical) observations that are complete.

Conceptually, the EM algorithm is a very general algorithm for maximum-likelihood estimation in incomplete-data problems. The algorithm is the formalization of an *ad hoc* approach to incomplete-data problems, which can be described as follows:

(1) replace the missing values by estimated values;
(2) estimate the parameters;
(3) re-estimate the missing values assuming the new parameter estimates are correct; and
(4) re-estimate the parameters and continue until convergence.

In the E step we find the conditional expectation of the missing data given the observed data and current estimated parameters. We then substitute the expectations for the missing data. In the M step we perform maximum-likelihood estimation of the parameters just as if there were no missing data.

Little and Rubin[4] spend little time discussing the analysis when missing values occur in the X matrix, focusing instead on missing values in the dependent variable. They mention that since levels of factors in an experiment are fixed by the experimenter, missing values, if they occur, do so far more frequently in the outcome variable, y, than in the factors, X. Thus, analyses of the case where there are missing observations in y dominate the text. (The issue of missing values occurring in X is discussed in Chapter 10 for logistic regression and for categorical and continuous X's) Also, Little and Rubin[4] note that the EM algorithm converges very slowly when many data points are missing.

2.3. Attempts to relax assumptions

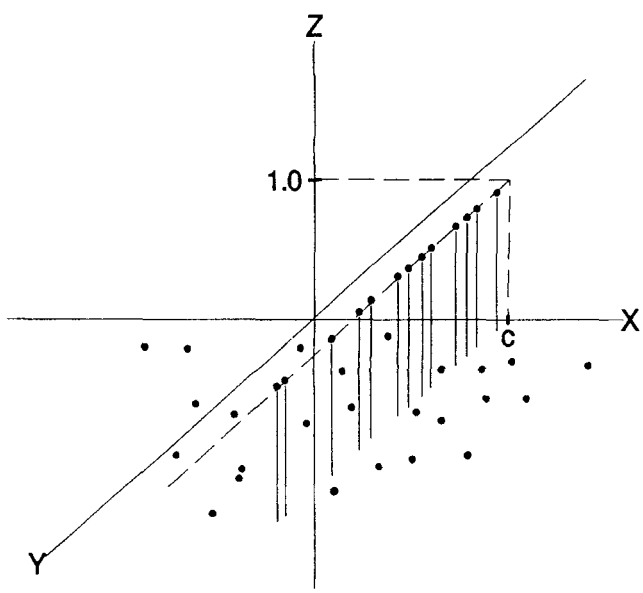
Simon and Simonoff[5] derive limits for the values of the least-squares estimates of the coefficients, β, and the associated t statistics when there are missing observations in one column of the X matrix. Extensions are also discussed to problems with missing observations in more than one column. These limits are developed subject to a constraint on the relationship of the missing data to the present data. The more restrictive MAR assumption is replaced by an

unknown mechanism (MUM) assumption. This assumption indicates that the missing values occur according to a probability mechanism that is a function of the data values. Ultimately, the authors develop a technique that makes no assumptions about the nature of the missing-value process and simply requires the use of ordinary least squares. In addition, the development is based upon examining the usual fixed-effects linear-regression model as we do in our investigation.

These authors suggest that their alternative considers the fact that the observed data have gone a long way toward providing results, regardless of the values the missing information assumes. This method provides upper and lower limits for the values in the β vector (and associated t statistics) as a function of the observed data and a measure of the non-randomness of the process that causes values to be missing. Unfortunately, while the authors do mention extensions of their work, the analysis is restricted to the case when only one column of X has missing values. Problems with their method include: the mathematical tractability of the proposed algorithm and numerical problems with respect to the algorithm's implementation on the computer.

Before proceeding, we note here that a common approach of adding dummy variables indicating when brands are not available is not an attractive alternative for our problem. For example, one way this approach could be implemented would be to add a new column in X with a '1' in the row corresponding to the observation with a missing value and '0' elsewhere in that column. We would add one new column in X corresponding to each observation with at least one missing value. In our example problem, this would require 179 additional columns and necessitate the inversion of a 187×187 matrix (i.e. the dimension of $X^{T}X$). This approach is impractical.

A second way to implement a dummy-variable scheme (see Method 2, Section 5) would be to add a single column for each brand with several 1's corresponding to rows that have values missing for that observations and 0's elsewhere. We have several pragmatic reasons for criticising this scheme. First, this dummy-variable scheme focuses on the entire observation



Clumping problem with the dummy-variable coding scheme

which has a missing value (i.e. X_i, the corresponding row in X). We would prefer an approach that considers the influence of each missing X_{ij} separately. Second, there will be a 'clumping' problem (which also occurs when one replaces the missing values with column means). The clumping problem occurs because, even though we include the dummy variable, we must still give values to the X_{ij} that are missing. For example, suppose we had only one variable in X and there were some missing observations. A second column would be added to X, which contained only 1's and 0's. In order to run a least-squares procedure to estimate β, we need to assign values to the X_{i1} which are missing, say c (in our problem we have assigned the logarithm of the price for missing values equal to 0, which corresponds to a price of \$1). In 3-space, these data would look like a scatter of points in the $X-Y$ plane and a single line of points in the $Y-Z$, given by $Z = 1 \cdot 0$ and $X = c$. This relation is shown in the Figure. We can see that the clump of points in the $Y-Z$ plane could skew our parameter estimates. Finally, by quantifying all the missing data with the same value of c, all of them will be represented by the same regression weight as shown in β. Intuitively, this seems to suggest that all the missing data are missing for the same reason. Even in the marketing problem we are considering the missing data could occur because:

(a) the store was out of stock;
(b) an accident occurred and the data was lost; or
(c) the brand of coffee was not distributed.

2.4. Making use of an influence function

A somewhat different approach to our missing data problem is suggested by the work described by Cook,[7,8] Hoaglin and Welsch,[9] Belsley *et al.*[10], Pregibon[11] and Welsch[12,13]. In the paper by Welsch[12], the author notes that regressions are constructed using prior knowledge, data, models and some form of estimation scheme. It is important to know whether our results depend significantly on prior knowledge, a small portion of the data or the estimation method we choose. Techniques that Belsley *et al.*[10] describe are concerned with determining whether an observation has a disproportionately large impact on the analysis: the authors use the idea of an 'influence function' in their work.

The purpose of an influence function, which is to measure what happens when a single observation is added to a sample, was introduced by Welsch.[12] An observation is called influential if its deletion would cause major changes in the various statistics constructed. Influential observations are usually outside the patterns set by the majority of the data in the context of a regression model. These observations usually arise from errors in observing or recording data, structural-model misspecification (e.g. using a linear model instead of non-linear) and legitimate extreme observations.

Welsch' procedures use data deletion to measure influential points. Influential data are then flagged and carefully examined. While there are many ways to measure influence, the authors conceptually describe one way as follows: we can think of all the data except the ith observation as 'good' and treat the ith observation as 'strange'. We should like the influence measure we use to ascertain whether the ith observation is really a cause for concern. A useful measure for this is the influence function:

$$b - b(i) \tag{4}$$

where b is the vector estimate of β in (1) and $b(i)$ is the vector of parameter estimates obtained by dropping the ith observation. The authors note that influential observations will lead to an influence measure greater than some magnitude (depending on the scaling used).

2.5. Hat matrix

Cook[7,8] and Hoaglin and Welsch[9] identify H, the hat matrix, as the key component in terms of understanding the influence of an observation. $H = X(X^{T}X)^{-1}X^{T}$ is a function of the explanatory variable matrix or design matrix, X, only.† From the equation $\hat{y} = Hy$, we see that H maps the observed values y into the fitted values $\hat{y}$. Hoaglin and Welsch[9] note that this relationship allows us directly to interpret elements in H as indicators of how much influence a particular observation has on the fit of a model. Cook[7,8] suggests that H can be used to detect non-homogeneous spacing in the observations which could lead to the identification of data deficiencies. While there is a consensus on the importance of the hat matrix as a diagnostic tool for detecting extreme points, Pregibon[11] points out that the usefulness for assessing the impact an observation has on various aspects of fit (e.g. parameter estimates, fitted values, goodness-of-fit measures) is not clear-cut. However, the author goes on to point out that various functions of H and the elements in H can be very useful in determining whether individual observations unduly influence the overall fit of a model.

3. METHOD DEVELOPMENT

The two methods we propose make fundamental use of influence functions similar to the one described by Belsley *et al.*[10] Welsch[12,13] and others are mainly interested in the problem of identifying the influential data and presenting the information in a way that will be useful to the analyst. Their work assumes that although the data may be anomalous, they are not missing. Our objective is to obtain 'good' estimates of β, influenced as little as possible by the missing values in an observation. This objective adheres more closely to the ideas described by Little and Rubin[4] and Simon and Simonoff.[5]

3.1. Loss-function: Q

In the development of our loss function it will be convenient to use the following definitions (with X an $n \times m$ matrix of n observations and m marketing variables, and y an $n \times 1$ vector of observed category volumes):

$$C(X) = X^{T}X \tag{5}$$

$$D(X) = C^{-1}(X) = (X^{T}X)^{-1} \tag{6}$$

$$G(X) = D(X)X^{T} = (X^{T}X)^{-1}X^{T} \tag{7}$$

$$H(X) = XG(X) = X(X^{T}X)^{-1}X^{T} \tag{8}$$

Let h_i be the ith diagonal entry of matrix $H(X) = H_{ii}$.

In Chapter 2 of Belsley *et al.*[10] the authors define $DFBETA_i$ as the expression in (4). They show that the jth component of $DFBETA_i$ can be written

$$b_j - b_j(i) = \frac{g_{ji}e_i}{1 - h_i} \tag{9}$$

where g_{ji} is the jith entry of the matrix G above. Thus, our first loss function is obtained by

† From a computational point of view, both Hoaglin and Welsch[9] and Belsley *et al.*[10] mention computing H as the product, LRL^{T} where L is orthogonal (obtained using Householder transformations) and R is upper triangular. Alternatively, they suggest using a singular value decomposition of X into $U\Sigma V^{T}$. This leads to computing H as UU^{T}.

taking the sum of the square of the expectations of the $DFBETA_{ij}$, that is

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{E}(b_j - b_j(i))^2$$

$$Q = \sigma^2 \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{g_{ji}^2}{(1 - h_i)} \tag{10}$$

where we have used the formula $\mathbf{E}(e_i^2) = \sigma^2(1 - h_i)$ for the residual variance. By taking expectations, the loss function Q avoids using information about the dependent measure to influence the treatment of the independent data. (Note that since σ^2 is constant it will not affect the optimization of the loss function Q.) In order to minimize this loss function with respect to the missing elements of X, we differentiate Q with respect to the missing elements. This leads to

$$Q(X) = \sum_{r=1}^{n} \sum_{l=1}^{m} \frac{g_{lr}^2}{(1 - h_r)} \tag{11}$$

using the rule for derivatives of a quotient, we have

$$\frac{\partial Q(x_{ij})}{\partial x_{ij}} = \sum_{r=1}^{n} \sum_{l=1}^{m} \frac{2g_{lr}(1 - h_r)\left(\dfrac{\partial g_{lr}}{\partial x_{ij}}\right) + g_{lr}^2\left(\dfrac{\partial h_r}{\partial x_{ij}}\right)}{(1 - h_r)^2} \tag{12}$$

We represent $\partial h_r / \partial x_{ij}$ and $\partial g_{lr} / \partial x_{ij}$ in terms of basic components in Appendix I.

While minimizing Q addresses our initial problem, its applicability to other problems is limited because of the following unattractive features: Firstly, suppose we have a case with only one x_{ij} missing and let the estimate of x_{ij} be large. The corresponding g_{ji} will be small, h_i will not change much and Q will become small. This implies that in this special case we can minimize Q by making the missing data dominant, which is not what we desire. Secondly, Q is not invariant under linear transformations. This could lead to a situation where, if the scale of a variable (i.e. a column in X) with no missing values were changed, Q would change and in turn the missing-data estimates would change. However, we note that the marketing problem which has motivated this work typically does not require the scaling of variables via linear transformations. In order to have alternatives that avoid these potential problems we consider a second loss function.

3.2. Loss-function: *P*

Let

$$b(X, y) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \tag{13}$$

be the least-squares estimates of the regression coefficients, and

$$z(X, y) = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \tag{14}$$

be the fitted or predicted values. The perturbed value of z may be written as

$$\mathbf{z}_i = z_i + (1 - h_i)^{-1}(z_i - y_i)h_i \tag{15}$$

The difference between the kth predicted value, z_k, and $z_{k(i)}$ the kth predicted value made without using the ith observation, can be written

$$z_{k(i)} - z_k = (1 - h_i)^{-1}(z_i - y_i)H_{ik} \tag{16}$$

To aid in the intuitive understanding of the second loss function, we can draw an analogy with both the jack-knife method and modern cross-validation. Recall that in the jack-knife estimate (see Miller[14,15]) of a parameter θ we systematically delete each observation of the population of size n and recompute $\theta = f$(the original data less one observation). The bulk of the remaining analysis (e.g. parameter estimation, interval estimation) is carried out with the 'psuedo-values', θ_i (the value of θ obtained with all but the ith observation) $i = 1, \ldots, n$. In cross-validation (for example, see Weisberg[16]), the data are divided into n overlapping subsets, each subset consisting of $n - 1$ cases. Estimates from the $n - 1$ cases can then be used to predict a value for the deleted point. This idea leads to PRESS† (PREdiction Sum of Squares) (see Allen[17,18]) which is related to our second 'loss function'.

Taking expectations of the squared change in predicted values we arrive at our loss function:

$$P = \sum_{i=1}^{n} \sum_{k=1}^{n} \mathbf{E}(z_{k(i)} - z_k)^2 \tag{17}$$

$$P = \sum_{i=1}^{n} \sum_{k=1}^{n} \left(\frac{H_{ik}}{1 - h_i}\right)^2 \mathbf{E}(z_i - y_i)^2 \tag{18}$$

Recalling the formula for the variance of a residual $\mathbf{E}(z_i - y_i)^2 = \sigma^2(1 - h_i)$ we rewrite (18):

$$P = \sigma^2 \sum_{i=1}^{n} \sum_{k=1}^{n} \left(\frac{H_{ik}}{1 - h_i}\right) \tag{19}$$

Since H is an idempotent matrix, i.e. $HH = H$, we have

$$P = \sigma^2 \sum_{i=1}^{n} \frac{h_i}{1 - h_i} \tag{20}$$

Huber[19] uses the equation

$$z(x_i, y_i) = (1 - h_i)x_i^{\mathrm{T}} b(i) + h_i y_i \tag{21}$$

(where x_i is the ith row of X and y_i is the ith observed value) to describe the term we are summing as the fraction of the fitted value, z_i, due to y_i divided by the fraction due to the predicted values, $x_i b(i)$.

In order to minimize the loss function with respect to the missing elements of X we differentiate P with respect to the missing elements. We can write

$$P(X) = \sum_{r=1}^{n} \frac{h_r}{(1 - h_r)} \tag{22}$$

$$\frac{\partial P(x_{ij})}{\partial x_{ij}} = \frac{\partial P(x_{ij})}{\partial h_r} \frac{\partial h_r}{\partial x_{ij}} \tag{23}$$

$$\frac{\partial P(x_{ij})}{\partial x_{ij}} = \sum_{r=1}^{n} \frac{1}{(1 - h_r)^2} \frac{\partial h_r}{\partial x_{ij}} \tag{24}$$

As mentioned above, we represent $\partial h_r / \partial x_{ij}$ in terms of its basic components in Appendix I.

There are several points worth emphasizing. Firstly, we can see from (11) and (12) that the

† PRESS is defined as the sum of the squared differences between the observed value and the prediction of this value without the ith observation. The formula for PRESS is related to our loss functions, particularly P. Both Allen[18] and Weisberg[16] have commented on the usefulness of PRESS as an important diagnostic statistic in regression analysis.

only terms which will contribute to the minimization of $Q(X)$ are those x_{ij} that correspond to missing values in X. Similarly, we can see from (22) and (23) that the only terms that will contribute to the minimization of $P(X)$ are those x_{ij} that correspond to missing values in X. Recall that these loss functions† are aggregate measures of the influence that the missing values have on the estimates b of β. Secondly, the terms h_r in the loss function and in the derivative of the loss function are complex non-linear functions of the missing values x_{ij} in X. Thirdly, matrix derivatives (for example, see Graybill,[20] Tatsuoka,[21] or Browne[22]) may be used to suggest a compact and computationally tractable representation of the objective function and, more importantly, of the analytic derivatives of this complex function.‡ In Appendix II we give an example which presents the objective function P and derivatives for a small missing-value problem. It is apparent that obtaining these derivatives without using matrix calculus is cumbersome. In the next section we discuss the software we are using to implement our procedure.

4. IMPLEMENTATION

This section discusses the computer implementation of the algorithm. The basic components include:

(1) missing-data initialization; and
(2) unconstrained non-linear optimization of the loss function.

Consider the data in a matrix X. The locations where data are missing are all replaced by the geometric mean of the *observed values* within the corresponding column (the geometric mean is used because the elements in the data matrix X are logged prices). That is, for each column of X we make the following assignment to x_{ij}:

$$x_{ij} = \begin{cases} (\Pi_{i=1}^{n} x_{ij})^{1/n} & \text{if the } i\text{th element of column } j \text{ missing} \\ x_{ij} & \text{if the element of column } j \text{ not missing} \end{cases}$$

The geometric means serve as initial values used by the non-linear optimization algorithms as they search for the replacement values of the x_{ij} which will minimize the objective function. These locations are also marked, because the only non-zero derivatives (i.e. elements that can be perturbed to allow us to make gains in the objective-function value) will correspond to positions in X that have missing values. Once the data are read in, the two basic components of the non-linear-optimization software begin to work. The first part is the function-generation component which evaluates the objective function and the derivatives at a particular point. The second part of the software is the component which does the optimization.

The non-linear-optimization software is used to modify the missing values so as to minimize the objective function. We have performed the investigation using two types of optimization software. The first is a conjugate-gradient algorithm (see Shanno and Phua[23]) and the second is an algorithm based on solving a sequence of local linear programs (the LLP- algorithm of Professor Glenn Graves, UCLA,[24]). The fundamental difference between these two approaches is in how the derivatives are computed. The conjugate-gradient approach used the formulae we have developed to evaluate the exact derivatives at any point. In contrast, the LLP method relies on numerical derivatives.

† With respect to the loss function P, it might be interesting to look at the imputed values obtained by considering the loss function, $\Phi(h_i/1 - h_i)$ (where Φ is the logarithm or sine function, for example).
‡ If the second derivatives of P or Q with respect to missing values could be obtained analytically and represented conveniently, then it may be possible to investigate the performance of second-order methods.

The conjugate-gradient algorithm uses the projection vector to evaluate the objective function and the derivative vector to obtain the direction of the search for new values. The new values will be those that minimize the influence function. This algorithm makes use of the analytic derivatives and at each step new values are obtained for the missing data. These new values are computed by modifying the current values using a linear combination of the current gradient (vector of derivatives) and the preceding direction vector. The algorithm can be summarized as follows (see Luenberger[25]). Starting at any $\mathbf{x}_0$ in $\mathbb{R}^n$ let $\mathbf{d}_0 = -\mathbf{g}_0$ (initial vector of derivatives).

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \tag{25}$$

$$\alpha_k = -\frac{\mathbf{g}_k^{\mathrm{T}} \mathbf{d}_k}{\mathbf{d}_k^{\mathrm{T}} \mathbf{Q} \mathbf{d}_k} \tag{26}$$

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \gamma_k \mathbf{d}_k \tag{27}$$

$$\gamma_k = \frac{\mathbf{g}_{k+1}^{\mathrm{T}} \mathbf{Q} \mathbf{d}_k}{\mathbf{d}_k^{\mathrm{T}} \mathbf{Q} \mathbf{d}_k} \tag{28}$$

In this algorithm, x_0 is obtained by replacing the missing values with the corresponding geometric mean, as described earlier. Also, $\mathbf{g}_k = \mathbf{Q}\mathbf{x}_k - \mathbf{b}$, $\mathbf{Q}$ is a symmetric matrix and $\mathbf{Q}\mathbf{x} = \mathbf{b}$. Thus, we can see that the first step is a steepest-descent step (i.e. in the direction of the gradient vector $-\mathbf{g}_0$) and the succeeding steps move in a direction $\mathbf{d}_{k+1}$ equal to a linear combination of the current gradient ($\mathbf{g}_{k+1}$) and the preceding direction vector $\mathbf{d}_k$.

The LLP software is a general-purpose algorithm that solves problems of the form

$$\begin{aligned} &\text{minimize} \quad g^m(y) \\ &\text{subject to} \quad g^i(y) \leqslant 0 \quad i = 1, \ldots, m-1 \end{aligned}$$

where y is a vector in $\mathbb{R}^n$, and $g^i(y)$ $(i = 1, \ldots, m)$ are differentiable functions.

The algorithm is referred to as a 'local gradient stepwise' correction descent algorithm. *Stepwise* refers to the fact that, given a y^0 in the domain of the g's, a correction vector Δy is obtained and the new point $y = y^0 + k\Delta y$ is used in the proceeding step. The method is *local* because the correction direction Δy and its length (determined by the scalar k) depend on the system's behaviour in a 'small' neighbourhood of the current point y^0. Finally, the algorithm is a *gradient* technique in that the gradients of the functions $g^i(y)$ play a major role in determining the correction direction.

Both algorithms will terminate based on user-supplied criteria including:

(a) a detector for small changes in the objective function;
(b) a detector for small changes in the model variables which enter the objective function; and
(c) a detector for the maximum number of iterations.

5. RESULTS

In this section we compare various parameter estimates and statistics obtained from five different methods. In addition, we present the results of jack-knife estimates obtained for four of the methods. The fundamental regression model used in this section is

$$\log T_{st} = \sum_{j=1}^{8} \beta_{p_j} \log p_{jst} + e_{st} \tag{29}$$

In Method 1, observations are deleted if they have any missing values. In this case, the original data set has a total of 234 observations (3 stores × 78 weeks) for 8 brands of coffee; 313 missing values occurred, for either brand 4, brand 6 or brand 7 (i.e. a single observation could have up to 3 missing elements), in 179 different observations. This reduces the data set to only 55 observations with price data for all 8 brands of coffee. We can quickly note that 313 missing values represents 17% of the total of 1872 (234 × 8) values. Method 1 analysis makes use of only 55 out of 234 observations, or 23.5% of the data. This means that 76·5 − 17 = 59·5% of the available data is not used in the statistical analysis!

Method 2, suggested by Little,[26] hypothesizes that the missing values are not truly missing. From this point of view, a more appropriate method would be to include a dummy variable which would be set to one if the brand did not appear in a store for a given week or set to zero otherwise. The corresponding value for the logarithm of price would be set to zero when the brand was absent. This would increase the number of columns in the X matrix by m_v (where m_v is the number of brands that have a missing value). Thus, Method 2 adds three dummy variables to (29), which leads to

$$\log T_{st} = \sum_{j=1}^{8} \beta_{p_j} \log p_{jst} + \delta_4 D_4 + \delta_6 D_6 + \delta_7 D_7 + e_{st} \tag{30}$$

where D_i is a dummy variable such that

$$D_i = \begin{cases} 1 & \text{if data are missing for brand } i \\ 0 & \text{if the data are present} \end{cases}$$

In Method 3, all the missing values are replaced by the geometric mean of the remaining non-missing data within a particular column. In Method 4 the missing values are replaced by those values which minimize the loss function P. In Method 5 the missing values are replaced by those values which minimize the loss function Q. Parameter estimates and diagnostic statistics are shown in Tables I, II and III.

In addition to reporting information about the parameter estimates and model statistics, we also present information about the influence of each observation. In particular, we present aggregate information about the following statistic (see Belsley[10]):

$$DFBETAS_{ij} = \frac{b_j - b_j(i)}{s(i)((X^{\mathrm{T}}X)^{-1}_{jj})^{1/2}} \tag{31}$$

$$= \frac{g_{ji}}{(\sum_{k=1}^{n} g_{jk}^2)^{1/2}} \frac{e_i}{s(i)(1-h_i)} \tag{32}$$

where $s(i)$ is the sample standard deviation computed after deleting the ith observation and the other terms are as defined earlier. $DFBETAS_{ij}$ represents the influence of the ith observation in the determination of the jth coefficient. In our analysis, we considered $DFBETAS_{ij}$ for $j = 1$ to 8 and only those 179 i's that had a missing value (for the price of brand 4, 6 and/or 7). RMSI is defined as the root-mean-square influence computed for all coefficients over those observations with a missing value. These statistics give us an impression of the influence on the coefficients of only the observations with missing values. We should be aware that by imputing for missing values in X we may indirectly alter the influence of observations without missing values on the coefficients. To emphasize this point, suppose we replace a missing value by 1,000,000 (when the average of the remaining data in X was 10) then the corresponding observations would have a large influence on the coefficient estimates.

Two important points are worth noting. Recalling that the loss functions are the sums of

Table I. Comparison statistics for Methods 1–5

Variable	Method 1		Method 2		Method 3		Method 4		Method 5	
	param. est.	std. err.	param. est.	std. err.	param. est.	std. err.	param. est.	std. err.	param. est.	std. err.
Intercept	8·79	1·93	8·91	1·03	6·56	1·37	10·81	0·71	10·22	0·70
Log Price 1	−1·54	0·28	−1·45	0·21	−1·87	0·30	−1·66	0·28	−2·00	0·28
Log Price 2	−1·80	0·29	−1·89	0·20	−1·90	0·30	−1·79	0·28	−2·01	0·28
Log Price 3	−1·48	0·72	−0·11	0·38	−0·14	0·55	−0·06	0·52	−0·29	0·53
Log Price 4	−0·92	0·58	−0·24	0·22	−0·53	0·31	−0·62	0·26	0·02	0·006
Log Price 5	−1·92	0·50	−2·00	0·23	−0·89	0·31	−1·08	0·29	−1·08	0·29
Log Price 6	1·11	0·91	1·40	0·74	1·11	0·92	−0·33	0·06	0·002	0·005
Log Price 7	2·65	1·81	−0·90	0·70	2·40	0·92	0·18	0·04	−0·03	0·005
Log Price 8	2·30	1·09	3·04	0·15	2·30	0·20	1·53	0·17	2·00	0·16
Dummy 4			−0·10	0·21						
Dummy 6			0·84	0·87						
Dummy 7			−0·68	0·57						
Method statistics: Root MSE	0·227		0·246		0·358		0·336		0·340	
Method statistics: R-square	0·72		0·80		0·58		0·63		0·62	
Method statistics: RSS†	2·48		13·47		28·87		25·41		25·94	
Method statistics: n	55		234		234		234		234	

† RSS is the sum of the squared residuals.

Table II. Comparison statistics for Methods 1–5 (part 2)

Variable	Method 1		Method 2		Method 3		Method 4		Method 5	
	T†	Prb‡	T	Prb	T	Prb	T	Prb	T	Prb
Intercept	4·55	0·0001	8·62	0·0001	4·80	0·0001	15·24	0·0001	14·66	0·0001
Log Price 1	−5·47	0·0001	−6·89	0·0001	−6·26	0·0001	−5·93	0·0001	−7·07	0·0001
Log Price 2	−6·29	0·0001	−9·26	0·0001	−6·45	0·0001	−6·47	0·0001	−7·18	0·0001
Log Price 3	−2·06	0·0450	−0·28	0·7797	−0·26	0·7922	−0·11	0·9104	−0·55	0·5854
Log Price 4	−1·59	0·1176	−1·07	0·2862	−1·70	0·0908	−2·34	0·0204	2·55	0·0116
Log Price 5	−3·81	0·0004	−8·76	0·0001	−2·83	0·0050	−3·75	0·0002	−3·68	0·0003
Log Price 6	1·23	0·2265	1·90	0·0588	1·21	0·2296	−5·83	0·0001	0·41	0·6858
Log Price 7	1·46	0·1512	−1·28	0·2028	2·61	0·0096	4·58	0·0001	−5·74	0·0001
Log Price 8	2·11	0·0398	20·00	0·0001	11·48	0·0001	9·16	0·0001	12·74	0·0001
Dummy 4			−0·49	0·6245						
Dummy 6			0·97	0·3331						
Dummy 7			−1·19	0·2337						

† T statistic for H_0: parameter = 0.
‡ Probability that $|T|$ is at least this extreme when H_0 is true.

the two influence measures (P or Q) for each observation, neither of these influence measures is the same as $DFBETAS_{ij}$ shown in (31); however, Q is the expectation of $DFBETAS_{ij}$. This means that the imputed values we find may not be those which minimize the sum of $DFBETAS_{ij}$ over the observations. Secondly, while $DFBETAS_{ij}$ might be a reasonable alternative loss function (i.e. the sum of the influence measures reported in the SAS output over the observations) we instead choose to use the expectation of this value. The main reason for this is that $DFBETAS_{ij}$ is a function of e_i, thus it is a function of the dependent variable

Table III. Comparison statistics for Methods 1–5 (part 3)

Variable RMSI†	Method 2	Method 3	Method 4	Method 5
Log Price 1	0·054059	0·046269	0·04852	0·04218
Log Price 2	0·047313	0·040795	0·039017	0·03673
Log Price 3	0·062984	0·068833	0·063117	0·06243
Log Price 4	0·072068	0·077229	0·071222	0·07015
Log Price 5	0·077167	0·077564	0·076191	0·07354
Log Price 6	0·025989	0·06068	0·083582	0·05608
Log Price 7	0·067914	0·07629	0·077823	0·05817
Log Price 8	0·083258	0·060265	0·068601	0·05679
Total	0·49075	0·50792	0·52807	0·45607
Dummy P4	0·076797			
Dummy P6	0·02573			
Dummy P7	0·067501			
Total	0·17003			

† RMSI is defined to be the root-mean-square influence computed for all coefficients over the 179 observations which had at least one missing value.

(by taking expectations, as is done in the development of Q, e_i falls out). We prefer that the dependent variable should not affect the values we impute for the independent variables. Intuitively, since we use a regression model to predict the dependent variable with a function of the independent variables, we should not use information from the dependent variables to impute the missing independent values. (Note that while the dependent variable does appear in either of the loss functions—see (16) and (17)—upon taking expectations it is a function of only the h_i terms, for P, or the h_i and g_i terms, for Q, which only depend on the independent variables. The constant σ^2 is a property of the dependent variables and represents the contribution of y to P).

The pattern of coefficients in Table I may be interpreted in the light of the considerable amount of study this market has received (cf. Cooper and Nakanishi[1]). The market is dominated by two national brands (brands 1 and 2) and a major regional brand (brand 5). These are the brands that have the ability to expand the market using their price and promotional policies. In addition, brand 8 is an aggregate of the premium label, private brands which tend to raise their prices during times of high demand and tend not to compete with the market leaders on the basis of price. Thus we expect to see a pattern with significant negative coefficients for brands 1, 2, and 5, and a significant positive coefficient for aggregate brand 8.

Method 1 results in parameters which seem to overstate the impact of the minor brands. Brand 3 has a significant negative coefficient that seems out of line, and only very large standard errors keep the other minor brands from also achieving an unwarranted significance. The great reduction in the available degrees of freedom that comes from deleting observations that have any missing values seems to have a very deleterious effect on the interpretability of the resulting parameters. Method 2 has an acceptable pattern of significance, but the 'clumping problem' may have contributed to values for brands 6 and 7 that seem too large (though not statistically significant). Replacing missing values with the geometric mean (Method 3) also seems to lead to overestimates of the impact of brands 6 and 7 on the market. Brand 7, in particular, is represented as having a significant effect in the wrong direction. Method 4 (criterion P) has a very reasonable set of parameters for the brands known to impact the market. The impact of brand 4 may be overstated, and even though the parameters for the

other minor brands seem within reason, their significance seems to be exaggerated by very small standard errors. The jack-knife estimates of these standard errors might be more realistic. Method 5 (criterion Q) produces the most *a priori* reasonable set of parameter estimates, only the very small standard errors for the minor brands seems questionable. Again, the jack-knife estimates of these standard errors should be better.

Using aggregate measures (R^2, RSS and RMSE) Method 2 does well (perhaps because it has the most parameters) and Method 1 does well (perhaps because it has the highest ratio of parameters to data points), followed by Methods 4 and 5 (which have nearly identical R^2), and Method 3 does the worst. The SAS influence measures—computed only for the observations with missing values (Table III), suggest that in the aggregate, Methods 2, 3 and 4 all perform about the same while Method 5, which was designed to minimize the influence of the missing values, performs best on this criterion. In the analysis using only the observations that had missing values, the sum of the RMSI was lowest for Method 5, followed by Methods 2 and 3, then Method 4. Decomposing the sum of the RMSI over eight coefficients, Method 4 had four out of eight coefficients with lower influence than in Methods 1, 2 or 3. In addition, most of the increase in the sum of the RMSI for Method 4 over Methods 2 and 3 is due to the contribution of RMSI from the sixth coefficient. Method 5 outperforms Method 4 in all cases and compared to all other methods is superior 23 out of 24 times. Only the influence of the missing components of an observation are directly affected by the imputation procedure. Thus minimizing the influence of an observation basically minimizes the influence of the missing components of that observation.

5.1. Jack-knife estimates

In this section we discuss the results of the jack-knife estimates of β, using Methods 2, 3, 4 and 5 (Method 1 was not analysed because there were too few data points). The main purpose of jack-knifing has been to obtain valid estimates of the standard errors (Miller[14]).

For Methods 2 and 3, the jack-knife estimates of β, β_J are obtained as follows: for Method 2, each row of the independent variable matrix X and the dependent variable vector y is deleted (one at a time) from the n ($=234$) total rows. With the $(n-1)$ remaining observations, β_{-i} is estimated using least squares. The n β_{-i}'s are used to obtain β_J (which equals the average of the β_{-i}'s) and σ_{β_J}, the estimated standard deviation. For Method 3, the missing values are replaced by the geometric mean of the corresponding column. β_J is then computed by constructing the β_{-i}'s in the same fashion as just described. For Methods 4 and 5, the jack-knife estimates are constructed in two steps. First, one of the n observations is deleted and the missing values are replaced by values which minimized the loss function (P or Q). Next, β_{-i} is obtained using least squares with the $(n-1)$ values. β_J and σ_{β_J} are then computed as previously described. Thus, constructing the jack-knife estimates for Methods 4 and 5 requires running each optimization algorithm n times. Table IV shows the jack-knife estimate β_J for Methods 2, 3, 4 and 5. Also reported are the root-mean-squared errors based on the difference between β_J and the standard least-squares estimate of β.

Comparing the entries in Table I and Table IV we can see the jack-knife provides much more reasonable standard errors for all the methods. The major brands have the expected significant effects using all the methods, but Methods 2 seems to overestimate the impact of brand 6 and Method 3 seems to overestimate parameters for both brands 6 and 7. Methods 4 and 5 have perfectly acceptable patterns of significance. Method 5 seems superior in that it gives larger estimates for the major brands and smaller estimates for the minor brands.

We might expect *a priori* that Methods 2 and 3 would lead to smaller RMSEs than Methods

Table IV. Jack-knife statistics for Methods 2–5

	Method 2		Method 3		Method 4		Method 5	
Variable	Mean†	S.E.‡	Mean	S.E.	Mean	S.E.	Mean	S.E.
Intercept	8·29	0·92	6·56	1·43	10·67	1·02	10·22	1·22
Log Price 1	−1·43	0·27	−1·87	0·31	−1·79	0·38	−1·98	0·38
Log Price 2	−1·87	0·21	−1·90	0·27	−1·89	0·29	−1·98	0·40
Log Price 3	−0·09	0·43	−0·14	0·67	−0·20	0·72	−0·31	1·14
Log Price 4	−0·30	0·24	−0·53	0·37	−0·49	0·43	0·04	0·12
Log Price 5	−1·94	0·26	−0·89	0·38	−0·99	0·38	−1·09	0·50
Log Price 6	1·31	0·76	1·11	1·05	−0·36	0·47	0·01	0·03
Log Price 7	−0·18	0·12	2·40	1·01	0·17	0·34	−0·07	0·23
Log Price 8	3·09	0·17	2·30	0·17	1·84	0·23	2·00	0·24
Dummy 4	−0·16	0·24						
Dummy 6	0·78	0·90						
Dummy 7	−0·10	0·09						
Root MSE§	0·324		0·002		0·145		0·021	

† Jack-knife estimate over all observations.
‡ Standard errors of the pseudo-values.
§ Root-mean-square error, where errors are differences between jack-knifed estimate and standard least-square estimate. These are averaged over all the parameters.

4 and 5 because Methods 2 and 3 are static with respect to how the missing values are handled (i.e. to construct the jack-knife estimate for Methods 4 and 5, the missing value replacement algorithm must be run to obtain the β_{-i}'s). We can see that Method 3 leads to the smallest RMSE followed by Methods 5, 4 and then 2.

6. CONCLUSIONS

The replacement of missing observations with values that minimize an influence function has been investigated. The missing-value problem we have considered is somewhat special, in terms of the mechanism which creates the missing values. This missing-value generation mechanism severely violates many of the assumptions required to use traditional missing-value replacement techniques. In order to address our problem, we have developed two intuitively appealing loss functions whose minimization provides imputed values to replace missing values. These methods will allow us to make more 'efficient' use of all available data. In addition, the methods lead to parameter estimates which have been minimally affected by the fact that in order to achieve greater efficiency we had to develop 'machinery' to allow us to carry out least-squares estimation (i.e. we had to impute values in order to use least-squares techniques). Method 5 (criterion Q) which was designed to address a specific problem in modelling category volume, seems to have resulted in the most appropriate parameter estimates for that problem. Method 4 (criterion P) comes very close to Method 4 and is designed to be applicable to a much broader class of missing-data problems, since it is not affected by linear transformations of the explanatory variables.

ACKNOWLEDGEMENT

The authors wish to thank Jim Brandon and John Cripps for their help on this project, and Roderick Little for his helpful comments on early drafts of this paper.

APPENDIX I

In this appendix we derive:

$$\frac{\partial h_r}{\partial x_{ij}} \tag{33}$$

Since $H = XG$ we can write from (33),

$$h_r = \sum_{l=1}^{n} x_{rl} g_{lr}$$

$$\frac{\partial h_r}{\partial x_{ij}} = \sum_{l=1}^{n} \left(\delta_{ri}\delta_{jl} g_{lr} + x_{rl} \frac{\partial g_{lr}}{\partial x_{ij}} \right) \tag{34}$$

where δ_{ri} is the Kronecker δ

$$\delta_{ri} = \begin{cases} 1 & \text{if } i = r \\ 0 & \text{if } i \neq r \end{cases}$$

Recall that $G = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. Browne[21] shows that

$$\frac{\partial (X^{\mathrm{T}}X)^{-1}}{\partial x_{ij}} = -(X^{\mathrm{T}}X)^{-1} Q (X^{\mathrm{T}}X)^{-1} \tag{35}$$

and

$$Q = J^{\mathrm{T}}X + X^{\mathrm{T}}J \tag{36}$$

where J is an $N \times K$ matrix with a '1' in the ij position and a '0' elsewhere. Taking the partial derivative of G with respect to x_{ij}, we have

$$\frac{\partial G}{\partial x_{ij}} = \frac{\partial ((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})}{\partial x_{ij}} \tag{37}$$

$$= \frac{\partial (X^{\mathrm{T}}X)^{-1}}{\partial x_{ij}} X^{\mathrm{T}} + (X^{\mathrm{T}}X)^{-1} \frac{\partial X^{\mathrm{T}}}{\partial x_{ij}}$$

$$= -(X^{\mathrm{T}}X)^{-1} Q (X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}} + (X^{\mathrm{T}}X)^{-1} \frac{\partial X^{\mathrm{T}}}{\partial x_{ij}}$$

$$= -(X^{\mathrm{T}}X)^{-1} J^{\mathrm{T}} X (X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}} - (X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}} J (X^{\mathrm{T}}X)^{-1} X^{\mathrm{T}} \tag{38}$$

$$+ (X^{\mathrm{T}}X)^{-1} \frac{\partial X^{\mathrm{T}}}{\partial x_{ij}}$$

$$= -DJ^{\mathrm{T}}H - GJG + D\delta_{ri} \tag{39}$$

For the partial derivative of the lrth element of G with respect to x_{ij} we have

$$\frac{\partial g_{lr}}{\partial x_{ij}} = -d_{lj}h_{ri} - g_{li}g_{jr} + \delta_{ri}d_{lj} \tag{40}$$

where the lower case variables represent particular elements in the respective matrices.

Equation (40) can now be substituted into (34), which yields

$$\frac{\partial h_r}{\partial x_{ij}} = \sum_{l=1}^{n} (\delta_{ri}\delta_{jl}g_{lr}) + x_{rl}(\delta_{ri}d_{lj} - d_{lj}h_{ri} - g_{li}g_{jr}) \tag{41}$$

$$= \delta_{ri} \sum_{l=1}^{n} \delta_{jl}g_{lr} + \sum_{l=1}^{n} x_{rl}(\delta_{ri}d_{lj} - d_{lj}h_{ri} - g_{li}g_{jr})$$

$$= \delta_{ri}g_{jr} + \sum_{l=1}^{n} \delta_{ri}x_{rl}d_{lj} - \sum_{l=1}^{n} x_{rl}d_{lj}h_{ri} - \sum_{l=1}^{n} x_{rl}g_{li}g_{jr} \tag{42}$$

Case 1

When $i = r$ (42) leads to

$$= \delta_{ri}g_{jr} + (\delta_{ri} - h_{ri}) \sum_{l=1}^{n} x_{rl}d_{lj} - g_{jr}h_{ri}$$

$$= g_{ji} + (1 - h_i)g_{ji} - g_{ji}h_i$$

$$= g_{ji} + g_{ji} - h_ig_{ji} - g_{ji}h_i$$

$$= 2(g_{ji} - h_ig_{ji})$$

$$= 2g_{ji}(-h_i) \tag{43}$$

Case 2

When $i \neq r$ (42) leads to

$$= \delta_{ri} \sum_{l=1}^{n} x_{rl}d_{lj} - h_{ri} \sum_{l=1}^{n} x_{rl}d_{lj} - g_{jr} \sum_{l=1}^{n} x_{rl}g_{li}$$

$$= -g_{jr}h_{ri} + (\delta_{ri} - h_{ri}) \sum_{l=1}^{n} x_{rl}d_{lj}$$

$$= -h_{ri}g_{jr} - g_{jr}h_{ri}$$

$$= -2g_{ji}h_{ri} \tag{44}$$

Thus, combining (43) and (44) we have the following formula for $\partial h_r/\partial x_{ij}$:

$$\frac{\partial h_r}{\partial x_{ij}} = \begin{cases} 2g_{ji}(1 - h_i) & \text{if } i = r \\ -2g_{jr}h_{ri} & \text{if } i \neq r \end{cases}$$

APPENDIX II

In this appendix we present the computations necessary for a small problem where X is 4×3 and has two missing values—at positions $(1, 2)$ and $(3, 3)$. The missing data are designated $x_{12} = x$ and $x_{33} = y$. Using the imputation technique we have developed, we find values for x and y which minimize the loss function. The reader should observe that the use of matrix calculus makes it straightforward to compute the analytic derivative of the objective function

when many missing values occur (the necessary formulae have been developed in Appendix I). To make these calculations without the special formulae would be a significant task for all but the smallest of missing data problems. Lower case letters represent particular elements in the respective matrices.

MATRIX EQUATIONS

Notation

$$C(X) = X^{\mathrm{T}}X$$
$$D(X) = C^{-1}(X) = (X^{\mathrm{T}}X)^{-1}$$
$$G(X) = D(X)X^{\mathrm{T}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$$
$$H(X) = XG(X) = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$$

where h_i is the ith diagonal entry of matrix $H(X) = H_{ii}$

Objective function

'C.D.I.M.'

$$P = \sigma^2 \sum_{i=1}^{N} \frac{h_i}{(1-h_i)}$$

Small example data matrix

$$X = \begin{bmatrix} 1 & x & 1 \\ 1 & -1 & -1 \\ 1 & 1 & y \\ 1 & -1 & 1 \end{bmatrix}$$

$$C(X) = X^{\mathrm{T}}X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ x & -1 & 1 & -1 \\ 1 & -1 & y & 1 \end{bmatrix} \begin{bmatrix} 1 & x & 1 \\ 1 & -1 & -1 \\ 1 & 1 & y \\ 1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & x-1 & y+1 \\ x-1 & x^2+3 & x+y \\ y+1 & x+y & y^2+3 \end{bmatrix}$$

$$D(X) = C^{-1}(X) = \frac{\text{adj } C}{\det C}$$

$$\det C(X) = 4 \begin{vmatrix} x^2+3 & x+y \\ x+y & y^2+3 \end{vmatrix} - (x-1) \begin{vmatrix} x-1 & x+y \\ y+1 & y^2+3 \end{vmatrix} + (y+1) \begin{vmatrix} x-1 & x^2+3 \\ y+1 & x+y \end{vmatrix}$$

$$\det C(X) = 2x^2y^2 + 4xy^2 + 6x^2 + 2y^2 - 8xy + 4x - 8y + 30$$

$$\text{adj } C = \begin{bmatrix} C_{11} & C_{21} & C_{31} \\ C_{12} & C_{22} & C_{32} \\ C_{13} & C_{23} & C_{33} \end{bmatrix}$$

Note that C_{ij} is ijth cofactor

$$C_{11} = \begin{vmatrix} x^2+3 & x+y \\ x+y & y^2+3 \end{vmatrix} = x^2y^2 + 2x^2 + 2y^2 - 2xy + 9$$

$$C_{12} = \begin{vmatrix} x-1 & x+y \\ y+1 & y^2+3 \end{vmatrix} = -xy^2 + 2y^2 + xy - 2x + y + 3$$

$$C_{13} = \begin{vmatrix} x-1 & x^2+3 \\ y+1 & x+y \end{vmatrix} = -x^2y + xy - x - 4y - 3$$

$$C_{23} = \begin{vmatrix} 4 & y+1 \\ x-1 & x+y \end{vmatrix} = xy - 3x - 5y - 1$$

$$C_{22} = \begin{vmatrix} 4 & y+1 \\ y+1 & y^2+3 \end{vmatrix} = 3y^2 - 2y + 11$$

$$C_{33} = \begin{vmatrix} 4 & x-1 \\ x-1 & x^2+3 \end{vmatrix} = 3x^2 + 2x + 11$$

$$D(X) = C^{-1}(X) = (X^{\mathrm{T}}X)^{-1} \frac{1}{\det(X^{\mathrm{T}}X)} \operatorname{adj}(X^{\mathrm{T}}X)$$

$$\operatorname{adj}(X^{\mathrm{T}}X) = \begin{bmatrix} x^2y^2 + 2x^2 + 2y^2 - 2xy + 9 & C_{21} & C_{31} \\ -xy^2 + 2y^2 + xy - 2x + y + 3 & 3y^2 - 2y + 11 & C_{32} \\ -x^2y + xy - x - 4y - 3 & -xy - 3x - 5y - 1 & 3x^2 + 2x + 11 \end{bmatrix}$$

$$G(X) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$$

$$G(X) = \frac{1}{\det\ X^{\mathrm{T}}X} \operatorname{adj}(X^{\mathrm{T}}X) \times \begin{bmatrix} 1 & 1 & 1 & 1 \\ x & -1 & 1 & -1 \\ 1 & -1 & y & 1 \end{bmatrix}$$

Note that if we let $k = 1/\det\ C(X)$, then

$$g_{11} = k(2xy^2 + 2y^2 + 2x - 4y + 6)$$
$$g_{12} = k(x^2y^2 + x^2y + xy^2 + 2x^2 - 4xy + 3x + 3y + 9)$$
$$g_{13} = k(2x^2 - 2xy - 2x - 2y + 12)$$
$$g_{14} = k(x^2y^2 - x^2y + xy^2 + 2x^2 - 2xy + x - 5y + 3)$$
$$g_{21} = k(2xy^2 + 2y^2 + 6x - 4y + 2)$$
$$g_{22} = k(-xy^2 - y^2 + x + 8y - 7)$$
$$g_{23} = k(-2xy - 2x - 2y + 14)$$
$$g_{24} = k(-xy^2 - y^2 + 2xy - 5x - 2y - 9)$$
$$g_{31} = k(-4xy - 4y + 8)$$
$$g_{32} = k(-x^2y - 3x^2 + y - 13)$$
$$g_{33} = k(2x^2y + 4xy - 4x + 2y - 4)$$
$$g_{34} = k(-x^2y + 3x^2 + 4x + y + 9)$$

$$H(x) = XG(X) = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$$
$$h_{11} = k(2x^2y^2 + 4xy^2 + 6x^2 + 2y^2 - 8xy + 4x - 8y + 14)$$
$$h_{21} = k(4xy - 4x + 4y - 4)$$
$$h_{31} = k(8x + 8)$$
$$h_{41} = k(-4xy - 4x - 4y + 12)$$
$$h_{12} + h_{21}$$
$$h_{22} = k(x^2y^2 + 2x^2y + 2xy^2 + 5x^2 + y^2 - 4xy + 2x - 6y + 29)$$
$$h_{32} = k(-2x^2y + 2x^2 - 4xy + 4x - 2y + 2)$$
$$h_{42} = k(x^2y^2 + 2xy^2 - x^2 + y^2 - 4xy + 2x - 4y + 3)$$
$$h_{13} = h_{31}$$
$$h_{23} = h_{32}$$
$$h_{33} = k(2x^2y^2 + 4xy^2 + 2x^2 + 2y^2 - 8xy - 4x - 8y + 26)$$
$$h_{43} = k(2x^2y + 2x^2 + 4xy - 4x + 2y - 6)$$
$$h_{14} = h_{41}$$
$$h_{24} = h_{42}$$
$$d_{34} = h_{43}$$
$$h_{44} = k(x^2y^2 - 2x^2y + 2xy^2 + 5x^2 + y^2 - 4xy + 10x - 2y + 21)$$

In order to find the minimum of the objective function, we differentiate P first with respect to x and then with respect to y. We are able to obtain the values of x and y that minimize the influence measure. Since H_{ij} is a function of both x and y we can write the following:

$$P(x, y) = \sum_{i=1}^{n} \frac{h_i}{(1 - h_i)} \tag{45}$$

$$\frac{\partial P}{\partial x} = \frac{\partial P}{\partial h_i}\frac{\partial h_i}{\partial x}$$

$$\frac{\partial P}{\partial x} = \sum_{i=1}^{n} \frac{(1 - h_i) + h_i}{(1 - h_i)^2}\frac{\partial h_i}{\partial x}$$

$$= \sum_{i=1}^{n} \frac{1}{(1 - h_i)^2}\frac{\partial h_i}{\partial x} \tag{46}$$

similarly,

$$\frac{\partial P}{\partial y} = \sum_{i=1}^{n} \frac{1}{(1 - h_i)^2}\frac{\partial h_i}{\partial y} \tag{47}$$

Using the formula for $\partial h_i/\partial x_{12}$ derived in Appendix I, we can write out (46) as

$$\frac{\partial P}{\partial x} = \frac{1}{(1 - h_1)^2}2g_{21}(1 - h_1) - \frac{1}{(1 - h_2)^2}2g_{22}h_{21} - \frac{1}{(1 - h_3)^2}2g_{23}h_{31} \tag{48}$$

To confirm the formulae developed in Appendix I, we can compare $\partial h_1/\partial x$ computed using the formula versus normal differentiation.

Formula

$$\frac{\partial h_1}{\partial x} = \frac{\partial h_1}{\partial x_{12}}$$
$$= 2g_{21}(1 - h_1)$$
$$= 2k^2(2xy^2 + 2y^2 + 6x - 4y + 2)(16)$$

Normal differentiation

$$\frac{\partial}{\partial x}\frac{2x^2y^2 + 4xy^2 + 6x^2 + 2y^2 - 8xy + 4x - 8y + 14}{2x^2y^2 + 4xy^2 + 6x^2 + 2y^2 - 8xy + 4x - 8y + 30}$$
$$= \frac{(4xy^2 + 4y^2 + 12x - 8y + 4)(16)}{(2x^2y^2 + 4xy^2 + 6x^2 + 2y^2 - 8xy + 4x - 8y + 30)^2} \tag{49}$$
$$= 2k^2(2xy^2 + 2y^2 + 6x - 4y + 2)(16)$$

The remaining partial derivatives can be validated similarly. There are three optimal solutions for this problem at $(x, y) = (1, 3)$, $(-3, -1)$ or $(1, -1)$ all with objective function values of 12. These solutions are obtained depending on the initial values of (x, y). Both optimization techniques gave identical results.

With the compact method for generating the objective function and its derivatives, we can use non-linear optimization software to search for values of the missing data that minimize our influence measure. Also, examination of the objective function in (45), reveals that the highest order term in h_i is of the form $\alpha x_{ij}^2 x_{kl}^2$ for a problem that has missing values at the (i, j) positions (where α is an arbitrary scalar constant). This function is a fourth-order polynomial (because the $D(X)$ term in $H(X)$ is a third-order polynomial). The functional form of our objective function is important when using non-linear optimization software. In particular, raising a small number (or large number) to a high power will lead to numerical/computational instabilities. Since the loss function under consideration behaves like a sum of low-order polynomials, our method should produce a good solution.

REFERENCES

1. L. G. Cooper and M. Nakanishi, *Market-Share Analysis: Evaluating Competitive Marketing Effectiveness*, Kluwer Academic, Boston, 1988.
2. A. A. Afifi and R. M. Elashoff, 'Missing observations in multivariate statistics I: Review of the literature', *J. Am. Statist. Assoc.*, **61**, 595–604 (1966a).
3. A. A. Afifi and R. M. Elashoff, 'Missing observations in multivariate statistics II: Point estimation in simple linear regression', *J. Am. Statist. Assoc.*, **62**, 10–29 (1966b).
4. R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley, New York, 1987.
5. G. A. Simon and J. S. Simonoff, 'Diagnostic plots for missing data in least squares regression', *J. Statist. Assoc.*, **81**, 394, 501–509 (1986).
6. N. K. Malhotra, 'Analyzing marketing research data with incomplete information on the dependent variable', *J. Mktng Res.*, **24** (February), 74–84 (1987).
7. R. D. Cook, 'Detection of influential observations in linear regression', *Technometrics*, **19**(1), 15–18 (1977).
8. R. D. Cook, 'Influential observations in linear regression', *J. Am. Statist. Assoc.*, **74**, 365, 169–174 (1979).
9. D. C. Hoaglin and R. E. Welsch, 'The hat matrix in regression and ANOVA', *Am. Statistn*, **32**(1), 17–22 (1978).
10. D. A. Belsley, E. Kuh and R. E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York, 1980.
11. D. Pregibon, 'Logistic regression diagnostics', *Ann. Statist.*, **9**(4), 705–724 (1981).

12. R. E. Welsch, *Modern Data Analysis: Influence Functions and Regression Diagnostics*, Academic Press, New York, 1982, pp. 148–169.
13. R. E. Welsch, 'An introduction to regression diagnostics', in *Proc. 30th Conf. on the Design of Experiments in Army Research Development and Testing*, The Army Mathematics Steering Committee, ARO Report 85–2, 1985.
14. R. G. Miller, 'The Jackknife – a Review', *Biometrika*, **61**, 1–15 (1974).
15. R. G. Miller, *Beyond ANOVA, Basics of Applied Statistics*, Wiley, New York, 1974.
16. S. Weisberg, *Applied Linear Regression*, 2nd end, Wiley, New York, 1985.
17. D. M. Allen, 'Means square error of prediction as a criterion for selecting variables', *Technometrics*, **13**(13), 469–475 (1971).
18. D. M. Allen, 'The relationship between variable selection and data augmentation and a method for prediction', *Technometrics*, **16**(1), 125–127 (1974).
19. P. J. Huber, *Robust Statistics*, Wiley, New York, 1981.
20. F. A. Graybill, *Introduction to Matrices with Applications in Statistics*, Wadsworth, Belmont, 1969.
21. M. M. Tatsuoka, *Multivariate Analysis: Techniques for Educational and Psychological Research*, Wiley, New York, 1971.
22. M. W. Browne, 'STA401—matrix methods in statistics: supplemental notes', Department of Statistics and Operations Research, University of South Africa.
23. D. F. Shanno and K. H. Phua, 'Remark on Algorithm 500', *ACM Trans. Math. Softw.*, **6**(4), 618–622 (1980).
24. G. W. Graves, 'A nonlinear programming algorithm', Anderson Graduate School of Management, U.C.L.A., 1988.
25. D. G. Luenberger, *Linear and Nonlinear Programming*. 2nd edn, Addison-Wesley, Reading, 1984.
26. R. J. A. Little, 'Personal communication', Department of Biomathematics, U.C.L.A.