UCLA Statistics Series No. 38 May, 1989

An Imputation Method for Dealing with Missing Data in Regression

Lee G. Cooper [1]

Jan de Leeuw [2]

Aram G. Sogomonian [3]

UCLA Classification: Theory and Methods. Social Sciences-Education.

[1] Anderson Graduate School of Management, UCLA.
[2] Departments of Psychology and Mathematics, UCLA.
[3] Anderson Graduate School of Management, UCLA.

An Imputation Method for Dealing with Missing Data in Regression

Lee G. Cooper, Jan de Leeuw, and Aram G. Sogomonian

Abstract

In this paper we describe a method for imputing the missing values in regression situations. We examine the standard fixed-effects linear regression model $\mathbf{y} = \mathbf{X}\beta + \epsilon$, where the regressors $\mathbf{X}$ are fixed and ϵ is the error term. This research focuses on the problem of missing values in $\mathbf{X}$. While a particular component of market-share analysis has motivated this research (where the columns of $\mathbf{X}$ represent various marketing variables), the techniques we suggest have a greater scope of applicability. Using influence functions, from robust statistics we develop two *loss functions* each of which is a function of the missing and existing values in $\mathbf{X}$. These loss functions turn out to be sums of ratios of low order polynomials. The minimization of either *loss function* is an unconstrained nonlinear-optimization problem. The solution to this nonlinear optimization leads to imputed values which have minimum 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.

We would like to thank Jim Brandon and John Cripps for their help on this project.

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 modeling brand sales. As described in Cooper and Nakanishi (1988) brand sales are often modeled as a function of a market share and a total category volume component. Specifically, brand sales is often modeled by the following relationship:

$$\text{Sales for brand } i = \text{Market share for brand } i \times \text{Total-category volume}$$

Typically, share and category volume are modeled as two separate processes. The appropriate technique for handling missing data in total-category volume models has been a source of concern in estimating these models. We propose a new technique that differs 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 total category-volume model.

Within the context of estimating a total category-volume model, the missing data problem can be summarized as follows:

> The estimation of a brand sales model typically involves the use of a retail scanner data source (eg. Nielsen). These data sources provide up to N weekly observations for J brands. An observation consists of the levels of all brand instruments (eg. promotion, price) in a given store in a given week. In general, all J brands are associated with K marketing instruments in any weekly observation. Some of these marketing instruments (variables) are continuous (eg. price) and some are categorical (eg. display). An example of the layout of the data base is provided in figure 1.

> Although the data base provides (space) for J brands in each store, it is possible that in some weeks and some stores, not all brands are available. It is also possible that some stores may not stock particular brands at any time. Thus there are often instances of missing data points within the weekly observations. However, 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, typically 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 we are throwing out valuable information. An example of missing data (marketing instruments) for brand J, in week 5, is provided in figure 1. We note here, that missing data are not necessarily data which take on the value of zero. For example, if brand 1 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!

It is the above problem that we are considering here (i.e. when some observations are missing some relevant values). We seek to develop a satisfactory method for handling missing values which does not involve discarding useful information. We achieve this by replacing missing values with quantities which minimize the "influence" of the imputed quantities on the estimates of the model parameters.

The balance of this paper is organized as follows: In §1.1 we develop the general notation. In §2 we review relevant research on the treatment of missing data in estimating statistical models. §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 §4. §5 is the conclusion

obs. no.	**Brand 1**	**Br. 2**	...	...	**Br. J**	**Br. 1**	...	...
	Instrument 1	**Inst. 1**	...	...	**Inst. 1**	**Inst. 2**	...	...
1	x_{111}	x_{121}			x_{1J1}	x_{211}		
2								
3								
4								
5					?			
.								
.								
.								
n	x_{11n}	x_{12n}			x_{1Jn}	x_{21n}		x_{KJn}

? - represents locations of missing value

Figure 1: **Layout of Database**

and suggested areas for further investigation.

1.1 Some Notation

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

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

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

The specific total-category volume model we are considering (for example see Cooper and Nakanishi (1988, p. 152.)) is as follows:

$$\log T(t) = \sum_{i=1}^{m} \beta_{p_i} \log p(i,t) + \sum_{k=1}^{K} \sum_{i=1}^{m} \beta_{k_i} x(k,i,t) + e_t \tag{2}$$

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

$$\log T(t) = \sum_{i=1}^{m} \beta_{p_i} \log p(i,t) \tag{3}$$

where $T(t)$ is the total-category volume at week t;
m is the number of brands;
k is the number of marketing instruments;
$p(i,t)$ is the price of the i^{th} brand at time t;
$x(k,i,t)$ is 1 if the k^{th} marketing instrument is used for the i^{th} brand in week t and 0 otherwise;
β's are the corresponding regression coefficients;
e_t is an error term.

2 RELATED LITERATURE

In this section we summarize the previous work which has been done in the study of incomplete data problems. In addition, the concept of an influence function is introduced. While influence functions have been commonly used in data analysis, the use of influence measures as a method of replacing missing data has not been considered previously to Cooper (1987). The primary contribution of this research is to implement and extend the work suggested by Cooper (1987) and to carry out some empirical work indicating the usefulness of our approach for some missing data problems. In positioning this paper it is important to discuss the 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. Some of these assumptions about the missing data generation mechanism, are too restrictive for our application. 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 which deals with missing data problems in model estimation. Much of the preliminary work is described in Afifi and Elashoff (1966a, 1966b). These authors list several approaches, each of which attempts to provide a single estimate of β. The simplest method is to drop (delete) observations which 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 methods which avoid dropping observations are 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 nonmissing values. Then the parameters of interest are estimated with least squares.

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 the observed or unobserved values. As mentioned estimates can be severely biased if this assumption is violated (Little and Rubin, 1987; Simon and Simonoff, 1986). Current statistical computing software makes wide use of these methods, however Little and Rubin (1987) provide more general techniques and recommend not using the methods proposed by Afifi and Elashoff (1966a, 1966b) 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-based procedures. These procedures fill in the missing values and the resulting completed data are analyzed by standard methods. Two commonly used techniques are the following:

i) Imputing unconditional means, where missing values in a column of $\mathbf{X}$ 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 that case).

See Little and Rubin (1987) for a thorough discussion of missing data techniques when the missing observations are in the dependent variable.

To summarize, one of the key assumptions which must be made when using these types of procedures (i.e., those in Little and Rubin, 1987) is that the missing 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 (1987) the authors discuss linear regression with missing values in the predictor variables. This corresponds to our problem in that the **X** matrix (equation 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 (equation 1) and the corresponding variances. No assumption is made about multivariate normality between **y** and **X**, which allows for dummy variables (i.e. categorical data) and interactions. By partitioning the **X** matrix into portions with and without missing data a mechanism is provided to estimate the covariance matrix. Recall that modifications must be made in this scheme because of the missing data. These modifications make use of the data we do have as well as changes in the EM algorithm.

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 missing values assuming new parameter estimates are correct, and 4) re-estimate 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 was no missing data.

One unfortunate aspect of this presentation is that Little and Rubin (1987) spend little time discussing the analysis when missing values occur in the **X** matrix. The authors instead focus 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[1]. Also, Little and Rubin (1987) note the the EM algorithm converges very slowly when many data points are missing.

2.3 Attempts to Relax Assumptions

In an article by Simon and Simonoff (1986), the authors derive limits for the values of the least-squares estimates of the coeffecients, β, and the associated t statistics when there are missing observations in one column of the **X** matrix. Extensions are also discussed for 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 the missing 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 which 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

[1] The issue of missing values occuring in X is discussed in chapter 10 for logistic regression and for categorical X's.

the usual fixed-effects linear-regression model as we do in our investigation.

The 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 nonrandomness 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 $\mathbf{X}$ 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 (as described in §5) 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's 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 x 187 matrix (i.e. the dimension of $X^T X$). This approach is impractical.

A second way to implement a dummy-variable scheme (see model 2, §5) would be to add a single column with several 1's corresponding to rows which have observations missing and 0's elsewhere. We have several pragmatic reasons for critiqueing this scheme. First, this dummy-variable scheme focuses on the entire observation which has a missing value (i.e. $X_{i\cdot}$, the corresponding row in X). We would prefer an approach which considers the influence of each 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} which 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 log of the price for missing values equal to 0, which corresponds to a price of \$1). In 3-space, this data would look like a scatter of points in the X-Y plane and a single line of points in the Y-Z plane, given by Z= c. 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 c, all of them will be represented by the same regression weight as shown in β. Intuitively, this would 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 occured 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 in the work described in Belsley, Kuh and Welsch (1980), Welsch (1982), Welsch (1985), Hoaglin and Welsch (1978), Pregibon (1981), Cook (1979), Cook (1977). In a paper by Welsch (1982), 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 which Belsley, Kuh and Welsch (1980) develop are concerned with determining whether an observation is having a disproportionately large impact on the analysis. The authors also 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, is introduced by Welsch (1982). 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's procedures use data deletion to measure influential points. Influential data are then flagged and carefully examined. Alternative fits, judgment or external information may be needed to reconcile the situation. While there are many ways to measure influence, the authors conceptually describe one way as follows: We can think of all the data but the i^{th} observation as "good" and the i^{th} observation as "strange." We would like the influence measure we use to ascertain whether the i^{th} observation is really a cause for concern. A useful measure to do this is the influence function:

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

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

2.5 The Hat Matrix

Finally, Hoaglin and Welsch (1978) and Cook (1977, 1979) identify H, the hat matrix, as the key component in terms of understanding the influence of an observation. $H = X(X^TX)^{-1}X^T$ is a function of the explanatory variable matrix or design matrix, X, only[2]. From the equation $\hat{y} = Hy$, we see that H maps the

[2] From a computational point of view, both Belsley, Kuh and Welsch (1980) and Hoaglin and Welsch (1978) mention computing H as the product, LRL^T where L is orthogonal (obtained using Householder transformations) and R is upper trianglar. Alternatively, they suggest using a singular value decomposition of X into $U\Sigma V^T$. This leads to computing H as UU^T.

observed values y into the fitted values $\hat{y}$. Hoaglin and Welsch (1978) note that this relationship allows us to directly interpret H as indicator of how much influence a particular observation has on the fit of a model. In articles by Cook (1977, 1979) suggests that H can be used to detect nonhomogeneous 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, Pergibon (1981) 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 MODEL DEVELOPMENT

The two methods we propose make fundamental use of influence functions similar to the one described by Belsley, Kuh and Welsch (1980), but our objective is to obtain "good" estimates of β. This objective adheres more closely to the ideas described in Little and Rubin (1987) and Simon and Simonoff (1986). Welsch, et al. are mainly interested in the problem of identifying the influential data and presenting the information in a way which will be useful to the analyst. However, their work assumes that while the data may be anamolous it is not missing.

To aid in the intuitive understanding of our technique de Leeuw (1988), draws an analogy with both the jackknife method and modern cross-validation. Recall that in the jackknife estimate (see Miller, 1974 and Miller, 1986) of a parameter θ we systematically delete each observation of the population of size n and recompute $\theta = f($ original data less one observation $)$. The bulk of the remaining analysis (e.g., parameter estimation, interval estimation) is carried out with the "pseudo-values", θ_k (the value of θ obtained with all but the k^{th} observation) $k = 1, ..., n$. In cross-validation (for example see Weisberg, 1985), the data is 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[3] (PREdiction Sum of Squares) (see Allen 1971, 1974) which is related to our "loss functions."

Both the jackknife and cross-validation require computations with a subset of the available data (see §5 for a brief discussion and the results of some jackknife estimates). Given small changes in the data it is desirable that the results of our analysis do not change drastically. One example of a small change in the data is the deletion of a single point. Using this idea, de Leeuw (1988) has suggested subtle, yet intuitively appealing, modifications to the loss function suggested in Cooper (1987). We elaborate on both of these loss functions next.

3.1 The 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 total-category volumes):

$$\begin{aligned} C(X) &= X^TX & (5)\\ D(X) &= C^{-1}(X) = (X^TX)^{-1} & (6)\\ G(X) &= D(X)X^T = (X^TX)^{-1}X^T & (7)\\ H(X) &= XG(X) = X(X^TX)^{-1}X^T & (8)\end{aligned}$$

Let h_i is the i^{th} diagonal entry of matrix $H(X) = H_{ii}$.

In chapter 2 of Belsley, Kuh and Welsch (1980) the authors define $DFBETA_i$ as the expression in (4). They show that the j^{th} component of $DFBETA_i$ can be written,

[3] PRESS is defined as the sum of the squared differences between the observed value and the prediction of this value without the i^{th} observation. The formula for PRESS is related to our loss functions, particularly P. Both Allen (1974) and Weisberg (1985) have commented on the usefulness of PRESS as an important diagnostic statistic in regression analysis.

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

where g_{ji} is the ji^{th} entry of the matrix G above. Thus, our first loss function is obtained by taking the sum of the square of the expectations of the $DFBETA_{ij}$, that is,

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

$$Q = \sigma^2 \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{g_{ji}^2}{(1 - h_i)} \tag{12}$$

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: 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,

$$Q(X) = \sum_{r=1}^{n}\sum_{l=1}^{m} \frac{g_{lr}^2}{(1 - h_r)} \tag{13}$$

using the rule for derivatives of a quotient, we have

$$\frac{\partial Q(x_{ij})}{\partial x_{ij}} = \sum_{r=1}^{n}\sum_{l=1}^{m} \frac{2g_{lr}(1 - h_r)(\frac{\partial c_{lr}}{\partial x_{ij}}) + g_{lr}^2(\frac{\partial h_r}{\partial x_{ij}})}{(1 - h_r)^2} \tag{14}$$

We represent $\partial h_r/\partial x_{ij}$ and $\partial g_{lr}/\partial x_{ij}$ in terms of basic components in Appendix 2.

While minimizing Q seems intuitively reasonable, de Leeuw (1988) conjectures that the loss function Q can be improved upon because of the following unattractive features: First, 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 we will minimize Q by making the missing data dominant, which is not what we desire. Second, 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 was 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 avoid these potential problems we consider the second loss function.

3.2 The Loss-Function: P

Let

$$b(X,y) = (X^TX)^{-1}X^Ty \tag{15}$$

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

$$z(X,y) = X(X^TX)^{-1}X^Ty \tag{16}$$

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

$$\mathbf{z}_i = z_i + (1-h_i)^{-1}(z_i - y_i)h_i \tag{17}$$

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

$$z_{k(i)} - z_k = (1-h_i)^{-1}(z_i - y_i)H_{ik} \tag{18}$$

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

$$P = \sum_{i=1}^{n}\sum_{k=1}^{n}\mathbf{E}(z_{k(i)} - z_k)^2 \tag{19}$$

$$P = \sum_{i=1}^{n}\sum_{k=1}^{n}(\frac{H_{ik}}{1-h_i})^2\mathbf{E}(z_i - y_i)^2 \tag{20}$$

Recalling the formula for the variance of a residual $\mathbf{E}(z_i - y_i)^2 = \sigma^2(1 - h_i)$ we rewrite (20):

$$P = \sigma^2 \sum_{i=1}^{n} \sum_{k=1}^{n} (\frac{H_{ik}}{1 - h_i}) \tag{21}$$

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

$$P = \sigma^2 \sum_{i=1}^{n} \frac{h_i}{1 - h_i} \tag{22}$$

Huber (1981) uses the equation,

$$z(x_i, y_i) = (1 - h_i) x_i{}^T b(i) + h_i y_i \tag{23}$$

(where x_i is the i^{th} row of X and y_i is the i^{th} 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 value, $x_i b(i)$.

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,

$$P(X) = \sum_{r=1}^{n} \frac{h_r}{(1 - h_r)} \tag{24}$$

$$\frac{\partial P(x_{ij})}{\partial x_{ij}} = \frac{\partial P(x_{ij})}{\partial h_r} \frac{\partial h_r}{\partial x_{ij}} \tag{25}$$

$$\frac{\partial P(x_{ij})}{\partial x_{ij}} = \sum_{r=1}^{n} \frac{1}{(1 - h_r)^2} \frac{\partial h_r}{\partial x_{ij}} \tag{26}$$

We represent $\partial h_r / \partial x_{ij}$ in terms of its basic components in Appendix 2.

There are several points worth emphasizing. First, we can see from equations (13) and (14) that the only terms which will contribute to the minimization of $Q(X)$ are those x_{ij} which correspond to missing values in X. Similarly, we can see from equations (24) and (25) that the only terms which will contribute to the minimization of $P(X)$ are those x_{ij} which correspond to missing values in X. Recall that these loss functions[4] are aggregate measures of the influence that the missing values have on the estimates b of β. Second, 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. Third, matrix derivatives (for example, see Graybill, 1969 or Tatsuoka, 1971) 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[5]. In appendix 1 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 nonlinear optimization of the loss function. Consider the data in a matrix, **X**. The locations where data is 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}:

$$x_{ij} = \begin{cases} \sqrt[n]{\prod_{j=1}^{n} x_{ij}} & \text{if } j^{th} \text{ element of column } i \text{ missing} \\ x_{ij} & \text{if } j^{th} \text{ element of column } i \text{ not missing} \end{cases}$$

The geometric means serve as initial values used by the nonlinear 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 nonzero

[4] 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 log or sine function, for example).

[5] If the second derivatives of P or Q with respect to missing values could be obtained analytically and represented convienently, then it may be possible to investigate the performance of second-order methods.

derivatives (i.e. elements which can be perturbed to allow us to make gains in the objective-function value) will correspond to positions in $\mathbf{X}$ which have missing values. Once the data are read in the two basic components of the nonlinear-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 nonlinear-optimization software is used to modify the missing values so as to minimize the objective function. We have investigated using two types of optimization software. The first is a conjugate-gradient algorithm (see Shanno and Phua, 1980) and the second is an algorithm based on solving a sequence of local linear programs (LLP- algorithm of Professor Glenn Graves, UCLA, 1988). The fundamental difference between these two approaches is in how the derivatives are computed. The conjugate-gradient approach uses the formulas we have developed to evaluate the exact derivatives at any point. In contrast, the LLP method relies on numerical derivatives.

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 the a linear combination of the current gradient (vector of derivatives) and the preceding direction vector. The algorithm can be summarized as follows (see Luenberger, 1984): Starting at any $\mathbf{x}_0$ in $\Re^n$ let $\mathbf{d}_0 = -\mathbf{g}_0$ (initial vector of derivatives).

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k \tag{27}$$

$$\alpha_k = -\frac{\mathbf{g}_k{}^T \mathbf{d}_k}{\mathbf{d}_k{}^T \mathbf{Q} \mathbf{d}_k} \tag{28}$$

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \gamma_k \mathbf{d}_k \tag{29}$$

$$\gamma_k = -\frac{\mathbf{g}_{k+1}{}^T \mathbf{Q} \mathbf{d}_k}{\mathbf{d}_k{}^T \mathbf{Q} \mathbf{d}_k} \tag{30}$$

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:

$$\begin{array}{l}\text{minimize } g^m(y),\\ \text{subject to } g^i(y) \leq 0 \quad i = 1, ..., m-1\end{array}$$

where y is a vector in $\Re^n$, and $g^i(y)$ $(i = 1, ..., m)$ 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^o in the domain of the $g's$, a correction vector Δy is obtained and the new point $y = y^o + k\Delta y$ 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 behavior in a "small" neighborhood of the current point y^o. 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 detector for small changes in the objective function, a detector for small changes in the model variables which enter the objective function, a detector for the maximum number of iterations.

5 RESULTS

In this section we compare various parameter estimates and statistics obtained from five different models. In addition, we present the results of jackknife estimates obtained for four of the models. The fundamental model used in this section is:

$$\log T(t) = \sum_{j=1}^{8} \beta_{p_j} \log p(j,t) \tag{31}$$

In Model 1 observations are deleted if the observation has a missing value. 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 dataset to only 55 observations each 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. Model 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 data is available but not used in the statistical analysis!

Model 2, suggested by Little (1987), hypothesizes that the missing values are not truly missing. A more appropriate model would include an dummy variable which would set the log of the price of the brand to one when the brand did not appear in a store for a given week and zero otherwise. This would increase the number of columns in the X matrix by m_v (where m_v is equal to the number of brands which have a missing value). Thus, model 2 adds three dummy variables to equation (31) which leads to:

$$\log T(t) = \sum_{j=1}^{8} \beta_{p_j} \log p(j,t) + \delta_4 D(4) + \delta_6 D(6) + \delta_7 D(7) \tag{32}$$

where D(i) is an dummy variable such that

$$D(i) = \begin{cases} 1 & \text{if the data is absent} \\ 0 & \text{if the data is present.} \end{cases}$$

In Model 3 all the missing values are replaced by the geometric mean of the remaining non-missing data within a particular column. In Model 4 the missing values are replaced by those values which minimize the loss function P. In Model 5 the missing values are replaced by those values which minimize the loss function Q. Parameter estimates and diagnostic statistics are shown in figures (2), (3) and (4).

Besides 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, Kuh and Welsch, 1980):

$$\begin{aligned} DFBETAS_{ij} &= \frac{b_j - b_j(i)}{s(i)\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}} && (33) \\ &= \frac{g_{ji}}{\sqrt{\sum_{k=1}^{n} g^2{}_{jk}}} \frac{e_i}{s(i)(1-h_i)} && (34) \end{aligned}$$

where $s(i)$ is the sample standard deviation computed after deleting the i^{th} observation and the other terms are as defined earlier. $DFBETAS_{ij}$ represents the influence of the i^{th} observation in the determination of the j^{th} coefficient. In our analysis we considered $DFBETAS_{ij}$ for $j = 1$ to 8 and only those 179 i's which have 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 of the observations with missing values only, on the coefficients. We should be aware that by imputing for missing values in $\mathbf{X}$ we may alter the influence of observations without missing values on the coefficients. To emphasize this point, suppose we replace a missing value by 10^6 (when the average of the remaining data in $\mathbf{X}$ was 10) then the corresponding observation would have a large influence on the coefficient estimates.

Two important points are worth noting. Recalling that the loss functions are the sums of the two influence measures (P or Q) for each observation, neither of these influence measures is the same as $DFBETAS_{ij}$, shown in equation (33, 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 the $DFBETAS_{ij}$ over the observations. Second, 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 (by taking expectations, as is done in the development of Q, e_i falls out). We prefer that the dependent variable 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: While the dependent variable does appear in either of the loss functions (see equations (18) and (19)) 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 $\mathbf{Y}$ to P).

In figure (2) we see that the sign and magnitude of the coefficient estimates remain stable. Two noticeable exceptions are at Price 6 and Price 7 (the two brands corresponding to those prices containing over 90% of the missing values). An interesting aspect of the market which these price data were obtained from, is that brands 6 and 7 are very small market share brands capturing 0.2% and 0.3% of the market, respectively. It would be very unusual for such small brands to have a large impact on total-category volume. For price 6, Model 1 (row deletion), Model 2 (dummy variables) and Model 3 (the geometric mean substitution method) lead to parameter estimates which are insignificant at the 5% level. For Model 4 (criterion P) the price 6 coefficient is negative and significant ($p< .01$). For model 5 (criterion Q) the parameter is extremely close to zero and not significant. For price 7 both model 1 and 2 yield insignificant parameter estimates. Model 3 and 4 give significant positive parameter estimates, however the estimate given by model 3 is about 10 times the estimate given by model 4. For model 5 the parameter value is the closest to zero, but still statitically significant. Finally, the parameter estimates

VARIABLE	MODEL 1		MODEL 2		MODEL 3		MODEL 4		MODEL 5	
	parm. est.	**std. err.**	**parm. est.**	**std. err.**	**parm. est.**	**std. err.**	**parm. est.**	**std. err.**	**parm. est.**	**std. err.**
Intercept	8.79	1.93	8.91	1.03	6.56	1.37	10.81	.71	10.22	.7
Log Price 1	-1.54	.28	-1.45	.21	-1.87	.3	-1.66	.28	-2.0	.28
Log Price 2	-1.8	.29	-1.89	.2	-1.9	.3	-1.79	.28	-2.01	.28
Log Price 3	-1.48	.72	-.11	.38	-.14	.55	-.06	.52	-.29	.53
Log Price 4	-.92	.58	-.24	.22	-.53	.31	-.62	.26	.02	.006
Log Price 5	-1.92	.5	-2.0	.23	-.89	.31	-1.08	.29	-1.08	.29
Log Price 6	1.11	.91	1.4	.74	1.11	.92	-.33	.06	.002	.005
Log Price 7	2.65	1.81	-.9	.7	2.4	.92	.18	.04	-.03	.005
Log Price 8	2.3	1.09	3.04	.15	2.3	.2	1.53	.17	2.0	.16
Dummy 4			-.1	.21						
Dummy 6			.84	.87						
Dummy 7			-.68	.57						
Model Statistics										
Root MSE	.227		.246		.358		.336		.340	
R-square	.72		.8		.58		.63		.62	
RSS[6]	2.48		13.47		28.87		25.41		25.94	
no. data pts.	55		234		234		234		234	

[6] RSS is the sum of the squared residuals.

Figure 2: **Comparison Statistics for Models I-V**

VARIABLE	MODEL 1		MODEL 2		MODEL 3		MODEL 4		MODEL 5	
	T[7]	Prb[8]	T	Prb	T	Prb	T	Prb	T	Prb
Intercept	4.55	.0001	8.62	.0001	4.8	.0001	15.24	.0001	14.66	.0001
Log Price 1	-5.47	.0001	-6.89	.0001	-6.26	.0001	-5.93	.0001	-7.07	.0001
Log Price 2	-6.29	.0001	-9.26	.0001	-6.45	.0001	-6.47	.0001	-7.18	.0001
Log Price 3	-2.06	.045	-.28	.7797	-.26	.7922	-.11	.9104	-.55	.5854
Log Price 4	-1.59	.1176	-1.07	.2862	-1.7	.0908	-2.34	.0204	2.545	.0116
Log Price 5	-3.81	.0004	-8.76	.0001	-2.83	.005	-3.75	.0002	-3.677	.0003
Log Price 6	1.23	.2265	1.90	.0588	1.21	.2296	-5.83	.0001	.405	.6858
Log Price 7	1.46	.1512	-1.28	.2028	2.61	.0096	4.58	.0001	-5.735	.0001
Log Price 8	2.11	.0398	20.0	.0001	11.48	.0001	9.16	.0001	12.706	.0001
Dummy 4			-.49	.6245						
Dummy 6			.97	.3331						
Dummy 7			-1.19	.2337						

7. T statistic for H_0: parameter = 0.
8. Significance probability, i.e., Probability that $|T|$ is so big when H_0 is true.

Figure 3: **Comparison Statistics for Models I-V (part 2)**

VARIABLE	MODEL 2	MODEL 3	MODEL 4	MODEL 5
RMSI[9]				
Price 1	.054059	.046269	.04852	.04218
Price 2	.047313	.040795	.039017	.03673
Price 3	.062984	.068833	.063117	.06243
Price 4	.072068	.077229	.071222	.07015
Price 5	.077167	.077564	.076191	.07354
Price 6	.025989	.06068	.083582	.05608
Price 7	.067914	.07629	.077823	.05817
Price 8	.083258	.060265	.068601	.05679
Total	.49075	.50792	.52807	.45607
dummy P4	.076797			
dummy P6	.02573			
dummy P7	.067501			
Total	.17003			

9. RMSI is defined to be the root mean square influence computed for all coefficients over the 179 observations which had at least 1 missing value.

Figure 4: **Comparison Statistics for Models I-V (part 3)**

obtained from the model 4 seem to be more "clear cut," leading to 7 parameter estimates which are significant at .0001 level and another significant at the .02 level (only the parameter for brand 3 price was insignificant, as in all the models).

In an aggregate sense, using R^2, RSS and RMSE measures presented in figure (2), model 2 (which has the most parameters) and model 1 (which has the least data points) are best, followed by models 4 and 5 (which have nearly identical R^2), and then model 3. The SAS influence measures (figure 4) suggest that in the aggregate, models 2, 3 and 4 all perform about the same while model 5, which was designed to minimize the influence of the missing values, performs best on this criterion. In the analysis using only the observations which had missing values the sum of the RMSI was lowest for model 5, followed by model 2 and 3 then model 4. Decomposing the sum of the RMSI over eight coefficients, model 4 had four out of eight coefficients with lower influence than in models 1, 2 or 3. In addition, most of the increase in the sum of the RMSI for model 4 over models 2 and 3 is due to the contribution of RMSI from the sixth coefficient. Model 5 outperforms model 4 in all cases and compared to all other models is superior 23 out of 24 times.

5.1 Jackknife estimates

In this section we discuss the results of the jackknife estimates of β, using models 2, 3, 4 and 5 (model 1 was not analyzed because there were too few data points). One of the main purposes for jackknifing has been for bias reduction (Miller, 1974). In addition, the method is sensitive to outliers, in that outliers tend to lead to poorer estimates, with much higher variance. Thus, the results of figure 5 can be interpreted as indicating the magnitude of the bias in our original parameter estimates shown in figure 2. Stability of the parameter estimates (i.e. close values when comparing estimates obtained in standard fashion versus estimates obtained by jackknifing), accompanied by a small variance for the parameter estimates can be interpreted as the original estimates not being severely biassed. Also, we can make inferences based on the information in figure 5, such as: what are the sensitivity characteristics of the "influence method" for missing-value replacement, to small perturbations in the data.

For models 2 and 3 the jackknife estimate of β, β_J is obtained as follows: For model 2, each row of the design matrix (X) and the independent 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 (= to the average of the β_{-i}'s) and σ_{β_J}, the estimated standard deviation. For model 3, the missing values are replaced with the geometric mean of the corresponding column. β_J is then computed by constructing the β_{-i}'s in the same fashion as was just described. For models 4 and 5 the jackknife 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 jackknife estimates for models 4 and 5 requires running the optimization algorithm(s) n times. Figure 5 shows the jackknife estimate, β_J, for models 2, 3, 4 and 5. Also, reported are the root mean squared errors which are based on the difference between β_J and the standard least-squares estimate of β.

Comparing the tables in figure 2 and figure 5 we can see all the models show stable parameter estimates. We might expect, a priori, that models 2 and 3, might lead to smaller RMSE's, than models 4 and 5 because models 2 and 3 are static with respect to how the missing values are handled (i.e. to construct the jackknife estimate for models 4 and 5, the missing value replacement algorithm must be run to obtain the β_{-i}'s). We can see that model 3 leads to the smallest RMSE followed by models 5, 4 and then 2. Also, we can see that models 2 and 3 lead to comparitively larger estimates for the regression coefficients associated with brands 6 and 7. The coefficients associated with these brands, estimated using models 4 and 5, seem more reasonable given that these are brands associated with small market shares.

VARIABLE	MODEL 2		MODEL 3		MODEL 4		MODEL 5	
	Mean[10]	**S.D.**[11]	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
Intercept	8.29	.06	6.56	.094	10.67	.067	10.22	.08
Log Price 1	-1.43	.018	-1.87	.02	-1.79	.025	-1.98	.025
Log Price 2	-1.87	.014	-1.90	.018	-1.89	.019	-1.98	.026
Log Price 3	-.091	.028	-.144	.044	-.198	.047	-.31	.075
Log Price 4	-.299	.016	-.534	.024	-.49	.028	.038	.008
Log Price 5	-1.94	.017	-.892	.025	-.986	.025	-1.09	.033
Log Price 6	1.31	.05	1.11	.069	-.364	.031	.005	.002
Log Price 7	-.182	.008	2.4	.066	.17	.022	-.072	.015
Log Price 8	3.09	.011	2.3	.011	1.84	.015	2.0	.016
Dummy 4	-.157	.016						
Dummy 6	.776	.059						
Dummy 7	-.1	.006						
Root MSE[12]	.324		.002		.145		.021	

[10] Jackknife estimate over all observations.

[11] Standard deviation of jackknife estimate.

[12] Root mean square error, where errors are differences between jackknifed estimate and standard least-square estimate. These are averaged over all the parameters.

Figure 5: **Jackknife Statistics for Models II-V**

6 CONCLUSION

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).

While these new methods seem intuitively appealing, they have their critics. It has been mentioned that the analysis we propose is ad hoc, for the following:

1) What is the justification for our choice of the objective function?

2) The imputations we obtain for the missing data and b (the estimate of β) do not correspond to those from maximum-likelihood based methods.

In response to some of these comments, we feel that minimizing these particular influence functions we have developed will best serve our purpose. Our intention is to obtain representative coefficient estimates, b, of the market we are analyzing. These coefficient estimates represent price elasticities. The influence-function approach emphasizes obtaining representative coefficient estimates over obtaining good estimates of the missing data values as emphasized by ML techniques. Because of this it seems that the ML based techniques discussed in the imputation-based literature may skew the coefficient estimates. Our intention is to make the best use of the data that is available while meeting the assumptions of the statistical techniques used(i.e. having a full X matrix without the missing-value holes required to use least squares). We require the modified dataset which we obtain in order to meet these assumptions, to have a minimal amount of influence on our results.

APPENDIX 1

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 formulas are developed in appendix 2). To make these calculations without the special formulas 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

$$C(X) = X^T X$$

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

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

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

h_i is i^{th} 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

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

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

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

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

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

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

Note: C_{ij} is ij^{th} cofactor

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

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

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

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

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

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

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

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

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

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

Note: let $k = 1/det\ C(X)$

$$g_{11} = k(2xy^2 + 2y^2 + 2x - 4y + 6)$$

$$g_{12} = k(x^2y^2 + x^2y + xy^2 + 2x^2 - 4xy + 3x + 3y + 9)$$

$$g_{13} = k(2x^2 - 2xy - 2x - 2y + 12)$$

$$g_{14} = k(x^2y^2 - x^2y + xy^2 + 2x^2 - 2xy + x - 5y + 3)$$

$$g_{21} = k(2xy^2 + 2y^2 + 6x - 4y + 2)$$

$$g_{22} = k(-xy^2 - y^2 + x + 8y - 7)$$

$$g_{23} = k(-2xy - 2x - 2y + 14)$$

$$g_{24} = k(-xy^2 - y^2 + 2xy - 5x - 2y - 9)$$

$$g_{31} = k(-4xy - 4y + 8)$$

$$g_{32} = k(-x^2y - 3x^2 + y - 13)$$

$$g_{33} = k(2x^2y + 4xy - 4x + 2y - 4)$$

$$g_{34} = k(-x^2y + 3x^2 + 4x + y + 9)$$

$$H(X) = XG(X) = X(X^TX)^{-1}X^T$$

$$h_{11} = k(2x^2y^2 + 4xy^2 + 6x^2 + 2y^2 - 8xy + 4x - 8y + 14)$$

$$h_{21} = k(4xy - 4x + 4y - 4)$$

$$h_{31} = k(8x + 8)$$

$$h_{41} = k(-4xy - 4x - 4y + 12)$$

$$h_{12} = h_{21}$$

$$h_{22} = k(x^2y^2 + 2x^2y + 2xy^2 + 5x^2 + y^2 - 4xy + 2x - 6y + 29)$$

$$h_{32} = k(-2x^2y + 2x^2 - 4xy + 4x - 2y + 2)$$

$$h_{42} = k(x^2y^2 + 2xy^2 - x^2 + y^2 - 4xy + 2x - 4y + 3)$$

$$h_{13} = h_{31}$$

$$h_{23} = h_{32}$$

$$h_{33} = k(2x^2y^2 + 4xy^2 + 2x^2 + 2y^2 - 8xy - 4x - 8y + 26)$$

$$h_{43} = k(2x^2y + 2x^2 + 4xy - 4x + 2y - 6)$$

$$h_{14} = h_{41}$$

$$h_{24} = h_{42}$$

$$h_{34} = h_{43}$$

$$h_{44} = k(x^2y^2 - 2x^2y + 2xy^2 + 5x^2 + y^2 - 4xy + 10x - 2y + 21)$$

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 which minimize the influence measure. Since H_{ij} is a function of both x and y we can write the following:

$$P(x,y) = \sum_{i=1}^{n} \frac{h_i}{(1 - h_i)} \tag{35}$$

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

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

$$= \sum_{i=1}^{n} \frac{1}{(1-h_i)^2} \frac{\partial h_i}{\partial x} \tag{36}$$

similarly,

$$\frac{\partial P}{\partial y} = \sum_{i=1}^{n} \frac{1}{(1-h_i)^2} \frac{\partial h_i}{\partial y} \tag{37}$$

Using the formula for $\partial h_i / \partial x_{12}$ derived in appendix 2, we can write out equation 36 as

$$\frac{\partial P}{\partial x} = \frac{1}{(1-h_1)^2} 2g_{21}(1-h_1) - \frac{1}{(1-h_2)^2} 2g_{22}h_{21} - \frac{1}{(1-h_3)^2} 2g_{23}h_{31} \tag{38}$$

To confirm the formulas developed in appendix 2 we can compare $\partial h_1 / \partial x$ computed using the formula versus normal differentiation.

Formula:

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

Normal Differentiation:

$$\frac{\partial}{\partial x} \frac{2x^2y^2 + 4xy^2 + 6x^2 + 2y^2 - 8xy + 4x - 8y + 14}{2x^2y^2 + 4xy^2 + 6x^2 + 2y^2 - 8xy + 4x - 8y + 30} = \tag{39}$$

$$\begin{aligned} &= \frac{(4xy^2 + 4y^2 + 12x - 8y + 4)(16)}{(2x^2y^2 + 4xy^2 + 6x^2 + 2y^2 - 8xy + 4x - 8y + 30)^2} \\ &= 2k^2(2xy^2 + 2y^2 + 6x - 4y + 2)(16) \end{aligned}$$

The remaining partial derivatives can be validated similarly. There are 3 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 intial 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 which minimize our influence measure. Also, examination of the objective function (equation 35), reveals that the highest order term in h_i is of the form $\alpha x_{ij}^2 x_{kl}^2$, for a problem which has missing values at the (i, j) and (k, l) 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.

APPENDIX 2

In this appendix we derive:

$$\frac{\partial h_r}{\partial x_{ij}} \tag{40}$$

Since $H = XG$ we can write (from equation 40)

$$\begin{aligned} h_r &= \sum_{l=1}^{n} x_{rl} g_{lr} \\ \frac{\partial h_r}{\partial x_{ij}} &= \sum_{l=1}^{n} (\delta_{ri}\delta_{jl} g_{lr} + x_{rl}\frac{\partial g_{lr}}{\partial x_{ij}}) \end{aligned} \tag{41}$$

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 $G = (X^T X)^{-1} X^T$. In Cooper (1987), the author shows that

$$\frac{\partial (X^T X)^{-1}}{\partial x_{ij}} = -(X^T X)^{-1} Q (X^T X)^{-1} \tag{42}$$

and

$$Q = J^T X + X^T J \tag{43}$$

where J is a $(N \times K)$ 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

$$\begin{aligned} \frac{\partial G}{\partial x_{ij}} &= \frac{\partial ((X^T X)^{-1} X^T)}{\partial x_{ij}} && (44) \\ &= \frac{\partial (X^T X)^{-1}}{\partial x_{ij}} X^T + (X^T X)^{-1} \frac{\partial X^T}{\partial x_{ij}} \\ &= -(X^T X)^{-1} Q (X^T X)^{-1} X^T + (X^T X)^{-1} \frac{\partial X^T}{\partial x_{ij}} \\ &= -(X^T X)^{-1} J^T X (X^T X)^{-1} X^T - (X^T X)^{-1} X^T J (X^T X)^{-1} X^T && (45) \\ &\quad + (X^T X)^{-1} \frac{\partial X^T}{\partial x_{ij}} \\ &= -D J^T H - G J G + D \delta_{ri} && (46) \end{aligned}$$

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

$$\frac{\partial g_{lr}}{\partial x_{ij}} = -d_{lj}h_{ri} - g_{li}g_{jr} + \delta_{ri}d_{lj} \tag{47}$$

where the lower case variables represent particular elements in the respective matrices. Equation (47) can now be substituted into equation (41) which yields,

$$\begin{aligned} \frac{\partial h_r}{\partial x_{ij}} &= \sum_{l=1}^{n}(\delta_{ri}\delta_{jl}g_{lr}) + x_{rl}(\delta_{ri}d_{lj} - d_{lj}h_{ri} - g_{li}g_{jr}) && (48) \\ &= \delta_{ri}\sum_{l=1}^{n}\delta_{jl}g_{lr} + \sum_{l=1}^{n}x_{rl}(\delta_{ri}d_{lj} - d_{lj}h_{ri} - g_{li}g_{jr}) \\ &= \delta_{ri}g_{jr} + \sum_{l=1}^{n}\delta_{ri}x_{rl}d_{lj} - \sum_{l=1}^{n}x_{rl}d_{lj}h_{ri} - \sum_{l=1}^{n}x_{rl}g_{li}g_{jr} && (49) \end{aligned}$$

CASE 1: When $i = r$ (49) leads to

$$\begin{aligned} &= \delta_{ri}g_{jr} + (\delta_{ri} - h_{ri})\sum_{l=1}^{n}x_{rl}d_{lj} - g_{jr}h_{ri} \\ &= g_{ji} + (1 - h_i)g_{ji} - g_{ji}h_i \\ &= g_{ji} + g_{ji} - h_i g_{ji} - g_{ji}h_i \\ &= 2(g_{ji} - h_i g_{ji}) \\ &= 2g_{ji}(1 - h_i) && (50) \end{aligned}$$

CASE 2: When $i \neq r$ (49) leads to

$$\begin{aligned} &= \delta_{ri}\sum_{l=1}^{n}x_{rl}d_{lj} - h_{ri}\sum_{l=1}^{n}x_{rl}d_{lj} - g_{jr}\sum_{l=1}^{n}x_{rl}g_{li} \\ &= -g_{jr}h_{ri} + (\delta_{ri} - h_{ri})\sum_{l=1}^{n}x_{rl}d_{lj} \\ &= -h_{ri}g_{jr} - g_{jr}h_{ri} \\ &= -2g_{ji}h_{ri} && (51) \end{aligned}$$

Thus, combining (50) and (51) we have the following formula for $\partial h_r/\partial x_{ij}$:

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

References

[1] A.A. Afifi and R.M. Elashoff. Missing observations in multivariate statistics i: review of the literature. *Journal of the American Statistical Association*, 61:p. 595–604, 1966a.

[2] A.A. Afifi and R.M. Elashoff. Missing observations in multivariate statistics ii: point estimation in simple linear regression. *Journal of the American Statistical Association*, 62:p. 10–29, 1966b.

[3] D.M. Allen. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(13):p. 469–475, 1971.

[4] D.M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16(1):p. 125–127, 1974.

[5] D.A. Belsley, E. Kuh, and R.E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley, 1980.

[6] R.D. Cook. Detection of influential observations in linear regression. *Technometrics*, 19(1):p. 15–18, 1977.

[7] R.D. Cook. Influential observations in linear regression. *Journal of the American Statistical Association*, 74(365):p. 169–174, 1979.

[8] L.G. Cooper. The minimum-influence approach to missing data in regression. June 1987. Marketing Science Conference, France.

[9] L.G. Cooper and M. Nakanishi. *Market-Share Analysis: Evaluating Competitive Marketing Effectiveness.* Kluwer Academic Press, 1988.

[10] J. de Leeuw. Derivatives of regression functions. 1988. Departments of Psychology and Mathematics, UCLA.

[11] G.W. Graves. A nonlinear programming algorithm. 1988. Anderson Graduate School of Management, UCLA.

[12] F.A. Graybill. *Introduction to Matrices with Applications in Statistics.* Wadsworth Publishing Co., 1969.

[13] D.C. Hoaglin and R.E. Welsch. The hat matrix in regression and anova. *The American Statistician*, 32(1):p. 17–22, 1978.

[14] P.J. Huber. *Robust Statistics.* John Wiley, 1981.

[15] R.J.A. Little. Personal communication. 1987. Department of Biomathematics, UCLA.

[16] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data.* John Wiley, 1987.

[17] D.G. Luenberger. *Linear and Nonlinear Programming.* Addison–Wesley Publishing Company, 2nd edition, 1984.

[18] R.G. Miller. *Beyond ANOVA, Basics of Applied Statistics.* John Wiley, 1986.

[19] R.G. Miller. The jackknife–a review. *Biometrika*, 61:p. 1–15, 1974.

[20] D. Pregibon. Logistic regression diagnostics. *The Annals of Statistics*, 9(4):p. 705–724, 1981.

[21] D.F. Shanno and K.H. Phua. Remark on algorithm 500. *ACM Transactions on Mathematical Software*, 6(4):p. 618–622, 1980.

[22] G.A. Simon and J.S. Simonoff. Diagnostic plots for missing data in least squares regression. *Journal of the American Statistical Association*, 81(394):p. 501–509, 1986.

[23] M.M. Tatsuoka. *Multivariate Analysis: Techniques for Educational and Psychological Research.* John Wiley, 1971.

[24] S. Weisberg. *Applied Linear Regression.* John Wiley and Sons, second edition, 1985.

[25] R.E. Welsch. An introduction to regression diagnostics. In *Proceedings of the Thirtieth Conference on the Design of Experiments in Army Research Development and Testing*, The Army Mathematics Steering Committee, 1985. ARO Report 85–2.

[26] R.E. Welsch. *Modern Data Analysis*, chapter Influence Functions and Regression Diagnostics, pages 148–169. Academic Press, New York, 1982.