BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-83/08

EDITING AND IMPUTATION FOR ECONOMIC SURVEY DATA

by

Roderick J.A. Little UCLA School of Medicine Los Angeles, CA 90024	and	Philip J. Smith U.S. Bureau of the Census Washington, D.C. 20233

This series contains research reports, written by or in cooperation with staff members of the Statistical Research Division, whose content may be of interest to the general statistical research community. The views reflected in these reports are not necessarily those of the Census Bureau nor do they necessarily represent Census Bureau statistical policy or practice. Inquiries may be addressed to the author(s) or the SRD Report Series Coordinator, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233.

Recommended by: Myron J. Katzoff

Report completed: December 9, 1983

Report issued: December 12, 1983

EDITING AND IMPUTATION FOR ECONOMIC SURVEY DATA

Roderick J.A. Little and Philip J. Smith*

*Roderick J.A. Little is Associate Professor of Biomathematics, UCLA School of Medicine, Los Angeles, California 90024. Philip J. Smith is a Mathematical Statistician in the Statistical Research Division of the Bureau of the Census, FOB 3, Room 3524, Washington, DC 20233. Professor Little's contribution to this research was conducted while he was an American Statistical Association Research Fellow at the Bureau of the Census, supported by the Census Bureau and by a grant from the National Science Foundation. A previous version of this paper was contributed to the Survey Research Methods Section of the 143rd Annual Meeting of the American Statistical Association, August 17, 1983, Toronto, Ontario, Canada. Whilst taking full responsibility for errors, the authors wish to thank Brian Greenberg and Preston J. Waite for guidance and constructive criticisms of the earlier draft.

ABSTRACT

At the U.S. Bureau of the Census, data in economic surveys may occasionally be missing as a result of a company's failure to respond to a certain question, for example. In addition, values of other variables may require editing because they are clearly implausible. Implausible (outlying) values may arise as a result of the failure of the respondent to understand the survey question. This paper develops a strategy for cleaning survey data with missing and outlying values in three stages: 1) detection of outlying cases; 2) detection of outlying values within outlying cases; and 3) imputation of likely values for missing and/or outlying and edited values. Methodological tools include distance measures, graphical procedures, and maximum likelihood and robust estimation for incomplete multivariate normal data. Data from the Annual Survey of Manufactures (ASM) are used to illustrate the method.

KEY WORDS: Beaton sweep operator; EM algorithm; graphical methods; incomplete data; Mahalanobis distance; maximum likelihood estimation; multivariate data; outlier detection; robust estimation.

1. INTRODUCTION

The quality control of data has become an increasingly important aspect of survey work. Nonsampling errors affecting the integrity of data are possible at virtually every junction in a survey where data are communicated or transcribed from one person or device to another. Without quality control of survey information, data intended for final analysis or tabulation and publication can be spurious or missing. In this case analysis and publication of such information may be of dubious value and may jeopardize the credibility of the organization conducting the survey and preparing the analysis and report: bad data must be edited and values imputed when they are missing or have been deleted during the editing process.

This paper discusses editing and imputation for a nxp rectangular data matrix $X = (X_1,...,X_n)^T$ containing n cases each with p variables, $X_i = (X_{i1},...,X_{ip})$, $i=1,...,n$. Some values in the matrix are missing because they were not recorded. Other values may be erroneous due to response errors or errors in coding or transcription.

We develop methods for editing and imputation for this incomplete data matrix by combining ideas from three areas of statistical research: multivariate robust estimation, multivariate outlier detection, and the analysis of incomplete data. The multivariate robust estimation problem has been discussed by Maronna (1976) and Huber (1977), and more recently by Campbell (1980), and Devlin, Gnanadesikan, and Kettering (1981). Research in outlying data has been recently reviewed by Beckman and Cook (1982). The literature on incomplete data is reviewed by Hartley and Hocking (1971), Orchard and Woodbury (1972), Dempster, Laird and Rubin (1977), and Little and Rubin (1983). Specific imputation methods for survey data are discussed by Kalton and Kasprzyk (1982), Little (1982), Sande (1982), Sedransk and Titterington (1980), and Chapman (1976).

Shih and Weisberg (1983) discuss the combination of outlier detection methods and methods for incomplete data in the regression context where one variable is dependent on the remaining variables. In contrast, we are concerned with methods which treat all the variables symmetrically.
Our approach builds on the work of Frane (1976) in which outlying cases are identified by large values of the Mahalanobis distance, with estimates of the mean and covariance matrix estimated by maximum likelihood for incomplete multivariate normal data, using the expectation-maximization (EM) algorithm. The use of the Mahalanobis distance in editing has been previously suggested by several authors (Wilks, 1963; Gnanadesikan and Kettering, 1972; Hawkins, 1980). The EM algorithm for normal data is discussed in Orchard and Woodbury (1972), Beale and Little (1975) and Dempster, Laird and Rubin (1977); as noted below, the algorithm is incorrectly stated in Frane's article. We extend Frane's basic approach by (a) providing a graphical procedure for detecting outlying cases; (b) modifying the EM algorithm to provide robust estimates of the mean and covariance matrix analogous to those given by Campbell (1980); and (c) providing an efficient stepwise procedure for identifying outlying values within outlying cases.
Our outlier identification procedure differs from (and in our view, improves on) the discriminant analysis suggested by Frane. Frane's procedures have been programmed and are available in the BMDPAM computer program (BMDP, 1981).

Our methods are illustrated using selected data for a particular industry from the Annual Survey of Manufactures (U.S. Department of Commerce, Bureau of the Census, 1981). Sixteen variables were selected for analysis, eight current year variables, and the eight corresponding variables from the previous year. The variables are the number of production workers (PW), the number of all other employees (OE), legally required fringe benefits paid (LE), voluntary payments to fringe benefit programs (VP), total man hours worked by production workers (MH), production workers wages (WW), all other salaries and wages (OW), and total wages paid (SW). When these variable labels are prefaced by an "A" they refer to current year information, and when prefaced by a "B" they refer to prior year information.

The next section presents our procedures and illustrates them with ASM data. Section 3 discusses statistical assumptions we make about the data and the mechanisms leading to missing values. Section 4 discusses specific requirements of editing ASM data that are not met by our methods. One requirement in particular, deserves attention because it applies in many editing contexts. This requirement arises from the presence of logical constraints between variables, such as a total adding to the sum of its parts. The analysis of a complex set of editing constraints is discussed in Fellegi and Holt (1976). A general synthesis of logical or mathematical approaches to editing with the statistical methods we propose seems a challenging task. As a step in this direction, we show in Section 4 how relatively simple logical constraints can be included in our editing procedures. Section 5 summarizes our recommendations and notes areas for future research.

Section 2. ESTIMATION, EDITING AND IMPUTATION

2.1 Preliminaries

As a preliminary to editing and imputation, rows and columns of the data matrix, X, are rearranged to group similar patterns of missing data together. Figure 1 illustrates the results of this operation. This step clarifies the pattern of missing data and reduces the computing time required in subsequent calculations. Another useful preliminary step is to display the marginal distributions of the observed values. Figure 2 presents the marginal distributions of two of the ASM variables, AWW and APW.

One might consider applying methods for univariable outlier detection to the marginal distributions of the variables. However, there are two reasons why this is not appropriate for economic data. First, the distributions of many of the variables are known to be skewed - see, for example, the marginal distributions in Figure 2. Thus, standard methods assuming normality cannot be applied without a preliminary transformation of the data. Secondly, even if an appropriate transformation can be applied, outliers based on transformed univariate data are plausibly valid members of the underlying population, at least in the context of ASM data: the crucial aspect of the data that univariate outlier detection methods ignore is association between the variables. It is these relationships (or lack thereof) that will indicate that a value is outlying rather than the value's location in the variable's marginal distribution.

In our illustration we transform the data to the natural log (ℓn) scale to remove the skewness of the marginal distributions as in Figure 2. A more fundamental reason for the ℓn transformation is that current imputation procedures for industrial data are often based on ratio estimates. For example, if it is required to impute for A on the basis of a correlated variable B and prior year values A' and B' of A and B, respectively, current imputation procedures impute for A as follows: $A = (A'/B')\,B$. Taking logarithms, this relation becomes linear: $\ell nA = \ell nA' - \ell nB' + \ell nB$. Our imputation procedure can be viewed as a generalization of this kind of edit where a linear relationship between ℓnA and $\ell nA'$, $\ell nB'$, ℓnB and other available predictors are empirically determined by regressions based on available data.

2.2 Identification of Outlying Cases: μ and $\sum$ Known.

Let $y_i = (y_{i1},\ldots,y_{ip})$ denote the ℓn transformed vector of values for case i, in the absence of missing values and contamination by response errors. Let $X_i = (X_{i1},\ldots,X_{ip})$ denote the corresponding vector of observed ℓn-transformed values for a completely recorded case, where the components are subject to error. Suppose that y_i are independently distributed with mean $\mu = (\mu_1,\ldots,\mu_p)$ and covariance matrix $\sum = \{\sigma_{jk}\}$. The Mahalanobis distance

$$D_i^2 = (X_i - \mu)^T \textstyle\sum^{-1}(X_i - \mu) \qquad (1)$$

is a natural measure of the distance of case i from the mean of the multivariate distribution. If y_i is multivariate normal and case i is not contaminated (that is, $X_i = y_i$), then it is well known that D_i^2 has a chi-squared distribution with p degrees of freedom. Large values of D_i^2 are evidence that one or more of the components of X_i are contaminated. If case i is incomplete, let $X_{(pi)}$ denote the vector of variables present in case i, and let p_i denote their number. An obvious adaptation of (1) is to restrict the distance D_i^2 to the present variables and their related parameters, only. That is, define

$$D_i^2 = (X_{(pi)} - \mu_{(pi)})^T \textstyle\sum_{(ppi)}^{-1} (X_{(pi)} - \mu_{(pi)}), \qquad (2)$$

where $\mu_{(pi)}$ and $\sum_{(ppi)}$ are the $(p_i x 1)$ vector of means and $(p_i x p_i)$ covariance matrix corresponding to the present variables $X_{(pi)}$. If y_i is multivariate normal, the observed values are not contaminated $(X_{(pi)} = y_{(pi)})$ and the missing data are missing at random (MAR) in the sense defined by Rubin (1976), then $D_i^2 \sim x_{p_i}^2$, and large values of D_i^2 suggest contamination in the observed components of $X_{(pi)}$.

Let $X_{(mi)}$ denote the missing components of case i with mean $\mu_{(mi)}$ and covariance matrix $\sum_{(mmi)}$. Dropping the subscript i for simplicity, we can decompose

$$(X- \mu)^T \; \Sigma^{-1} \; (X- \mu) = [X_{(p)}-\mu_{(p)}]^T \; \Sigma^{-1}_{(pp)} \; [X_{(p)}-\mu_{(p)}]$$
$$+ [X_{(m)} - \mu_{(m.p)}]^T \; \Sigma^{-1}_{(mm.p)} \; [X_{(m)}-\mu_{(m.p)}]$$

where $\mu_{(m.p)} = \mu_{(m)} + \sum_{(mp)} \sum_{(pp)}^{-1}(X_{(p)}- \mu_{(p)})$ is the best linear predictor of $X_{(m)}$ based on $X_{(p)}$, $\sum_{(mp)}$ is the covariance of $X_{(m)}$ and $X_{(p)}$ and $\sum_{(mm.p)}$ is the covariance matrix of the residual $X_{(m)} - \mu_{(m.p)}$.

If the missing values in case i, $X_{(mi)}$, are imputed from their best linear predictors, $\mu_{(m \cdot p)}$, then letting X_i^* denote the (px1) vector of present and imputed values for case i,

$$D_i^2 = (X_i^* - \mu)^T \; \Sigma^{-1} \; (X_i^* - \mu) \; . \qquad (3)$$

In this case (2) and (3) are identical and imputations formed from best linear predictors have no effect on the Mahalanobis distance, D_i^2. Our strategy is to replace μ and $\sum$ by estimates and to use (3) to determine outlying cases. In the next section we describe a method of obtaining robust estimates of μ and $\sum$ in the presence of missing data.

2.3 Estimation of μ and $\sum$.

In practice, μ and $\sum$ are unknown and must be estimated from available data. With a data matrix with no missing values, the standard procedure (Wilks, 1963) is to replace μ and $\sum$ by the sample mean and covariance matrix, yielding the Mahalanobis squared distance for case i. With missing data, Frane (1978) estimates μ and $\sum$ by maximum likelihood (ML), assuming the data are multivariate normal[1]. The ML estimates can be found by the iterative EM algo-

[1] The estimates are consistent for μ and $\sum$ under any underlying distribution with finite fourth moments (Beale and Little, 1975). Thus the multivariate normality assumption is not essential for the utility of the method.

rithm (Orchard and Woodbury, 1972; Beale and Little, 1975; Dempster, Laird and Rubin, 1977). To describe the EM algorithm let $\Theta^{(t)}=(\mu^{(t)}, \sum^{(t)})$ denote current estimates of μ and $\sum$ at iteration t. Each iteration of the algorithm involves an E-step and an M-step. The E-step calculates the expected values of the complete data sufficient statistics given the observed data and current estimates $\Theta^{(t)}$:

$$E\{\sum_{i=1}^{n} X_{ij}|X_{(pi)}, \Theta^{(t)}\} = \sum_{i=1}^{n} X_{ij}^{(t)} \qquad , j=1,\ldots,p,$$

$$E\{\sum_{i=1}^{n} X_{ij}X_{ik}|X_{(pi)}, \Theta^{(t)}\} = \sum_{i=1}^{n} \{X_{ij}^{(t)}X_{ik}^{(t)} + C_{jki}^{(t)}\}, \; j,k=1,..p, \tag{4}$$

where

$$X_{ij}^{(t)} = \begin{cases} X_{ij} & \text{if } X_{ij} \text{ is present, or} \\ E\{X_{ij}|X_{(pi)}, \Theta^{(t)}\} & \text{, if } X_{ij} \text{ is missing, and} \end{cases} \tag{5}$$

$$C_{jki}^{(t)} = \begin{cases} 0 & \text{, if } X_{ij} \text{ or } X_{ik} \text{ present, or} \\ \text{Cov }\{X_{ij}, X_{ik}|X_{(pi)}, \Theta^{(t)}\}, & \text{if } X_{ij} \text{ and } X_{ik} \text{ misssing.} \end{cases}$$

The imputed values $E\{X_{ij}|X_{(pi)}, \Theta^{(t)}\}$ and the adjustments $C_{jki}^{(t)}$ are found from the regression of the missing variables in case i on the observed variables, $X_{(pi)}$, by applying the Beaton sweep operator (Beaton, 1964; Goodnight, 1979; Clarke, 1982) to the current estimates of μ and $\sum$. The M-step of the algorithm computes new maximum likelihood estimates $\Theta^{(t+1)} = (\mu^{(t+1)}, \sum^{(t+1)})$ from the expected complete data sufficient statistics, as for complete data:

$$\mu_j^{(t+1)} = \sum_{i=1}^{n} X_{ij}^{(t)}/n \qquad (6)$$

$$\sigma_{jk}^{(t+1)} = n^{-1} \sum_{i=1}^{n} \{X_{ij}^{(t)} X_{ik}^{(t)} + C_{jki}^{(t)}\} - n\mu_j^{(t+1)}\ \mu_k^{(t+1)}$$

$$= n^{-1} \sum_{i=1}^{n} \{(X_{ij}^{(t)} - \mu_j^{(t+1)})(X_{ik}^{(t)} - \mu_k^{(t+1)}) + C_{jki}^{(t)}\} \ .$$

Frane's (1976) formula for computing the covariance matrix on page 161 is incorrect in that μ should denote the mean of observed and imputed cases, and not the mean using all available cases of each variable, as defined in the paper.

One possible procedure for identifying outlying cases is to estimate μ and Σ using the iterating equations (4) and (5), yielding estimates $(\hat{\mu},\hat{\Sigma})$ and then calculating distances

$$D_i^2 = (X_{(pi)} - \hat{\mu}_{(pi)})^T \hat{\Sigma}_{(ppi)}^{-1} (X_{(pi)} - \hat{\mu}_{(pi)}) \qquad (7)$$

from present variable i only, for each case i. By the same argument as that relating equations (2) and (3), this quantity can also be computed as

$$D_i^2 = (X_i^* - \hat{\mu})^T \hat{\Sigma}^{-1} (X_i^* - \hat{\mu}), \qquad (8)$$

where X_i^* is the vector of observed and imputed values from the final step of the EM algorithm. We define (7) or (8) to be the Mahalanobis distance for case i for an incomplete data set. Note that for complete data sets $\hat{\mu}$ and $\hat{\Sigma}$ are the sample mean and covariance matrix, and (7) and (8) reduce to the usual form of the Mahalanobis distance for complete data.

A drawback with this procedure is that $\hat{\mu}$ and $\hat{\Sigma}$ are calculated from data contaminated by outlying or erroneous values and hence are not consistent estimates of μ and Σ . Frane (1976) proposes the simple expedient of reestimating μ and Σ , excluding cases with large values of D_i^2. In contrast, we modify the M-step of the EM algorithm by downweighting extreme observations. Specifically, we propose the expectation-robust estimation

(ER) algorithm defined by replacing the M-step (6) by the R-step which yields robust estimates:

$$\mu_j^{(t+1)} = \sum_{i=1}^{n} w_i \; X_{ij}^{(t)} / \sum_{i=1}^{n} w_i$$

and

$$\sigma_{jk}^{(t+1)} = \frac{\sum_{i=1}^{n} w_i^2 \left\{ [X_{ij}^{(t)} - \mu_j^{(t+1)}][X_{ik}^{(t)} - \mu_k^{(t+1)}] + C_{jki}^{(t)} \right\}}{\sum_{i=1}^{n} w_i^2 - 1}$$

where

$$w_i = \omega\,(d_i) \,/\, d_i ,$$

and

$$d_i = \{[X_{(pi)}^{(t)} - \mu_{(pi)}^{(t)}]^T [\Sigma_{(pi)}^{(t)}]^{-1} [X_{(pi)}^{(t)} - \mu_{(pi)}^{(t)}]\}^{1/2} \tag{9}$$

is the square root of the Mahalanobis distance for present variables at iteration t .

Here ω denotes a two parameter bounded influence function (Hampel, 1974) defined by

$$\omega(d_i) = \begin{cases} d_i \text{ if } d_i < d_{oi} \\ d_{oi} \exp \{-(d_i - d_{oi})^2/2b_2^2\} \text{ if } d_i > d_{oi} \end{cases}$$

where $d_{oi} = \sqrt{p_i} + b_1/2$, p_i denotes the number of present variables for case i, and b_1 and b_2 are quantities to be specified by the data analyst.

In the R-step of the ER algorithm the square root Mahalanobis distance, d_i, represents the measure of proximity from case i to $\mu^{(t)}$. The influence function is designed to give a full weight of 1 to clearly inlying observations in the computations for the updated means $\mu_j^{(t+1)}$ and covariances $\sigma_{jk}^{(t+1)}$. However, observations whose square root distance exceeds the specified cut off, d_{oi}, receive decreasingly smaller weights in these computations as the distance increases beyond d_{oi}. Under standard assumptions, $d_i^2 \sim \chi_{p_i}^2$, and Fisher's square root approximation gives $d_i \stackrel{\cdot}{\sim} N(\sqrt{p_i}; \sqrt{1/2})$ (Kendall and Stuart, 1968). Consequently, square root distances beyond d_{oi} represent less likely observations since they lie beyond b_1 standard deviations of their mean $\sqrt{p_i}$. The choice of b_1 specifies this cut off. Also, b_2 specifies how rapidly the weights decrease beyond d_{oi}. Hampel (1973) has suggested that $b_1 = 2$ and $b_2 = 1.25$ are good choices.

A similar robust covariance matrix estimate (HUB) performed well in simulations by Devlin, Gnanadesikan and Kettering (1981), directed at comparing estimates of eigenvalues of the correlation matrix from contaminated normal data. Other robust estimates that did well in that paper such as the method MLT based on the p-variate elliptical t distribution, are also candidates for the R-step of the ER algorithm. However, comparisons of alternative robust procedures lie outside the scope of this paper.

The ER algorithm iterates between the E and R-step until the estimates converge. In the case of complete data ($p_i = p$, $i=1,\ldots, n$) the E-step of the algorithm is redundant, and the R-step corresponds to one step of the robust estimation procedure proposed by Campbell (1980). Thus the ER algorithm combines Campbell's robust method with the E-step of the normal EM algorithm for filling in missing data.

In so far as the EM algorithm is guaranteed to converge, and in the absence of missing values, Campbell's procedure is known to converge to robust estimates of μ and $\sum$, we conjecture that the ER algorithm converges. Further work is currently being conducted to substantiate this conjecture.

Finally, we suggest that initial estimates of μ and $\sum$ be obtained from complete observations. Tables 1 and 2 show the results of applying the EM algorithm and the ER algorithm respectively to the ASM data. The slightly reduced variance estimates in the second table compared with the first show the effect of downweighting the more outlying cases in the R step of the algorithm.

2.4 Identification of Outlying Cases: μ and $\sum$ estimated.

Our procedure identifies outlying cases as those that are improbably distant from the robustly estimated centroid of observations. We measure this distance for each case by the Mahalanobis distance, D_i^2, which is the square of the quantity d_i of equation (9) from the final iteration of the ER algorithm: $D_i^2 = d_i^2$. A statistical criterion may be developed to specify what is meant by "improbably distant." In the absence of missing or contaminated values and under normal assumptions, the Mahalanobis distance D_i^2 in (8) has the property that $(n-p)nD_i^2/[(n-1)(n+1)p]$ has an F distribution with p and n-p degrees of freedom, (Anderson, 1958; Hawkins, 1974). If the data are incomplete the exact distribution of D_i^2 in (7) is unknown, but it is clear that since the ML estimates of μ and $\sum$ are consistent, then asymptotically $D_i^2 \sim \chi_{p_i}^2$. Taking into account the robust estimation of μ and $\sum$, we conjecture that $(n_c-p_i)n_cD_i^2/[(n_c-1)(n_c+1)p_i]$ $(=F_i$, say) has approximately an F distribution with p_i and n_c-p_i degrees of freedom, where n_c is the number of the completely recorded cases. This conjecture has little theoretical justification at present, but the use of the number of complete cases to determine degrees of freedom has done well in simulation studies for a related problem in Little (1979).

Using F_i, we may compute a p-value for each case. This p-value may be interpreted as the probability of observing a more extreme observation than the one at hand, and describes the improbability of the case. Those cases with a p-value less than a specified significance level (e.g., 0.01) are designated as outlying. The choice of significance level, however, is somewhat arbitrary, and even if the approximate distribution theory of the previous paragraph is adequate, the F-test is based on normality assumptions that may be overrestrictive. Thus we advocate the use of an informal graphical procedure for detecting outlying cases which is more empirically-based.

Our procedure extends the graphical method of Gnanadesikan (1977) for multivariate data to incomplete multivariate observations. We noted in Section 2.1 that if case i is uncontaminated and the data are normal and missing values are missing at random, then $D_i^2 \sim \chi_{p_i}^2$. The Wilson-Hilferty (1931) transformation of the chi-squared distribution yields $(D_i^2/p_i)^{1/3} \stackrel{\cdot}{\sim} N(1-2/(9p_i),2/(9p_i))$. Consequently, a probability plot of

$$Z_i = [(D_i^2/p_i)^{1/3} - 1 + 2/(9p_i)] / [2/(9p_i)]^{1/2} \qquad (10)$$

versus standard normal order statistics should reveal atypical observations. Figure 3 gives a probability plot of the Z_i for the ASM data.

In the interval between .01 and .90 of this figure, the Z_i's plotted versus $\Phi^{-1}[(i-1/2)/n]$ exhibit a strong linear trend typical of normal data. However, beyond the expected normal statistic corresponding to the cumulative probability of .90, the Z_i's begin to deviate greatly from the linear trend. This departure suggests that these observations beyond .90 are atypical and may be considered to be outlying.

2.5 Selecting Outlying Values Within Cases: The Variable Selection Procedure

In the variable selection procedure each present variable in an outlying case is ranked according to the marginal decrease in Mahalanobis distance obtained by removing that variable and all more influential variables from the computation of the distance. Letting $\tilde{\mu}$ and $\tilde{\Sigma}$ denote the robust estimates of μ and Σ, we obtain this ranking via the following algorithm:

STEP 1: For outlying case i, compute for each present variable k

$$D_i^{(k)} = (X_{i(k)} - \tilde{\mu}_{(k)})^T \tilde{\Sigma}_{(k)}^{-1} (X_{i(k)} - \tilde{\mu}_{(k)}),$$

the Mahalanobis distance with variable k omitted. This distance shows the effect of eliminating k in computing the distance of the observation from the mean. If variable k is the only outlying value in this observation, then $D_i^{(k)}$ will be significantly smaller than D_i^2.

STEP 2: $\min_k D_i^{(k)}$ is determined:

The single most influential variable contributing to the extremity of observation i is found. Let us call this variable j_1. By removing j_1, the probability of case i's remaining attribute values is the greatest.

STEP 3: Compute $D_i^{(kj_1)}$, the Mahalanobis distance with both variable j_1 and k removed, for all present variables $k \neq j_1$.

STEP 4: Determine $\min_k D_i^{(kj_1)}$

The variable minimizing $D_i^{(kj_1)}$ is the next most influential variable, conditional on the removal of variable j_1 in Step 2. Let j_2 denote this variable. The algorithm then proceeds to find j_3, the next most influential variable, conditional on the prior removal of variables j_2 and j_1, and so on until all the present variables in observation i are exhausted.

Table 3 gives a summarization of this algorithm for one outlying case, number i=65. The total distance computed using μ and Σ for this case is D_{65}^2= 136.27 which corresponds to a p-value much less than 0.001.

Removal of the most influential variable, BWW, reduces the distance to 59.03, a 56.68% incremental decrease in distance. If BWW was the only outlying value, then by removing it the p-value associated with $D_{65}^{(BWW)}$ would be, at least, moderately large. However, on removing BWW, the p-value is still rather small (it is <0.001) and consequently we are led to search for other outlying variables in the case.

Conditional on removing BWW, BPW is the next most influential variable. Removing it yields a remaining distance of $D_{65}^{(BWW,BPW)} = 33.84$, again with a significant p-value (0.004) indicating that the case with both BWW and BPW removed is still unlikely and that other outliers must be imbedded in the case.

Conditional on removing BWW and BPW, ALE is the next most influential variable. Removing it yields a remaining distance of $D_{65}^{(BWW,BPW,ALE)} = 22.47$ with an insignificant p-value. Consequently, we stop our search here having identified three outlying variables, BWW, BPW, and ALE.

As with the selection of outlying cases, the appropriate choice of critical p-value above which this procedure is terminated is not at all clear, given the fact that the case has been preselected as extreme, and the choice of reference F distribution relies on normal assumptions. An alternative strategy to setting a critical value that is feasible for modest sized data sets is to determine graphically how deeply to edit the case. At each editing step for a given outlying case the Wilson-Hilferty transformation may be applied to the remaining Mahalanobis distance, with edited variables treated as missing. The transformed remaining distance for the case may then be graphed along with all other cases' transformed distances in a normal plot, as discussed in Section 2.4. If the outlying case lies along the diagonal line described by the body of well-behaved normal data, then the appropriate depth of editing for that case has been found. Otherwise, the next most influential variable should be edited also and the transformed remaining distance plotted as before. Editing further variables according to their rank of influence and plotting their remaining transformed distance would then continue until that case's point no longer lies off the diagonal line described by the body of well-behaved normal data.

Our procedure is in a sense analogous to backward elimination in linear regression. It shares with that procedure the property that it does not ensure that the best set of a given number of variables remains unedited. Elaborations of the procedure to stepwise selection, where previously deleted variables are allowed to reenter if they no longer add significantly to D^2, or to all possible subset selection, might be feasible for small data sets. However, we believe that the simple backward selection approach should eliminate distinctly outlying values in most practical applications.

Our procedure can be contrasted with Frane's (1976) method, which applies a two group backward elimination discriminant analysis, where the first group consists of the outlier and the second group is defined by the vector of means μ . This form of elimination removes the variable that adds the least to the discriminant function, that is, the most inlying variable. Our procedure eliminates outliers <u>first</u>, whereas the <u>last</u> variable removed in Frane's discriminant analysis is the least inlying. The latter is an outlier with respect to its marginal distribution, rather than its conditional distribution given other variables present, which forms the basis of our method. Consequently, in determining the most outlying variable Frane's procedure fails to exploit the associations between the edited variables and other observed variables, which as we noted in Section 2 are a key feature of our problem. We prefer our stepwise procedure since multivariate relationships are taken into account in determining the outlying variables.

A pragmatic test of the properties of our variable selection procedure, is to consider how well it works for particular cases. In our example of case 65 in Table 3, the procedure found the values for prior year production workers (BPW), prior year production workers' wages (BWW), and current year legally required payments to pension programs (ALE) to be outlying. Important ratios involving these variables are the average hours worked by a production worker in one year, BMH/BPW; the average hourly wage rate for production workers, BWW/BMH; the average yearly wage rate for production workers, BWW/BPW; the ratio of voluntary to legally required pension payments, BVP/BLE; and these ratios' current year versions. For case number 65 these ratios are given in Table 4.

From this table the prior year ratio for the average number of hours worked by a production worker is BMH/BPW = 1.1 thousand, or 1,100 hours per year. This corresponds roughly to a 23 hour work week for 48 working weeks per year. Full time employment throughout one year corresponds roughly to 1,730 hours, agreeing very closely with the computed ratio in Table 4 of 1.7 thousand for the current year. However, the prior year computed ratio of BMH/BPW = 1.1 would not seem inconceivable if most workers in this factory worked approximately half time for the entire year. But probing more deeply into the data, further doubt is cast on the BMH/BPW ratio: the average prior year yearly wage is BWW/BPW = 3.3 thousand dollars, and the average prior year hourly wage rate is BWW/BMH = $2.90 per hour. These figures are very low indeed: the hourly wage is well below the legal minimum hourly wage rate of $3.45 per hour, and if this job was the wage earner's sole source of income, the salary of $3,300 per year is below the poverty level of $4,620 per year (U.S. Department of Commerce, Bureau of the Census, 1982-1983). The current year values for these ratios are strikingly different and correspond to an average hourly wage of $6.30 per hour and an average annual salary of $11,000 per year. These values are much more concordant with our particular industry's average.

Also, the current to prior year voluntary legally required pension payments ratio is AVP/BVP = .2 . This value is known to be low for our particular industry and is caught by the variable selection procedure.

The adjusted (imputed) ratios corresponding to variables selected as outlier for case 65 are also given in Table 4 and will be discussed further in Section 2.4.

In spite of the evidence listed in Table 4 one could argue that the original data used to compute the alleged outlying ratios are valid. For example, one might plausibly contend that this factory improved salaries, hourly wages, and the total working hours between the prior and current year by lowering the current year voluntary payments to pension plans. Without excellent prior information vilification of the data is as plausible as its vindication. What our variable selection method offers is an empirically based and statistically principled procedure for the selection of unlikely values.

As a final check on the efficacy of the variable selection procedure we recompute the distances D_i^2 from equation (7) for each case accounting for present and unedited variables, only. Applying the Wilson-Hilferty transformation (10) to these distances and letting p_i denote the number of present and unedited variables for case i, a normal probability plot of the Z_i's will reveal atypical observations and serve as a check on how well the variable selection procedure worked: if the observations lie along a 45° line in the probability plot, then the variable selection procedure has removed the outlying values. However, if one selects a very large significance level for the tests of the variable selection procedure, one runs the very real risk of editing "good" data. In this case the transformed data will fall below the 45° line of inclination in the probability plot, indicating that the data has been overedited. Consequently, we advocate smaller significance levels for the tests in the variable selection procedure. Also, plots of these "edited" transformed distances may be used to help determine an appropriate level: starting with a very small significance level one may perform the variable selection procedure, and then draw the normal probability plot of the edited transformed distances. If too many values appear to be atypical from this plot then a larger significance level can be chosen. One may then iterate between the variable selection and probability plotting procedure and selection of increasingly larger significance levels until the analyst is satisfied with the final probability plot.

2.4 Imputation

In the final step of our procedure we "edit" values found to be outlying in Section 2.3: these values are treated as if they were originally missing at random. To impute for the edited and missing values, μ and $\sum$ are re-estimated via the EM algorithm (Orchard and Woodbury, 1972; Beale and Little, 1975; Dempster, Laird, and Rubin, 1977). The Beaton sweep operator is used in the E-step (5) of the final iteration of the algorithm to produce regressions of missing (or edited) variables on non-missing and unedited variables for each case. Missing values are then imputed from these regressions and a clean data set produced.

An ideal evaluation of this imputation procedure would be to attempt to obtain true values of the data from a reinterview, and then compare the edited values with the truth. In the absence of reinterview data, our evaluation is limited to looking at the data before and after editing and imputation, and checking whether the imputed values are substantively plausible. We present here the results of applying our method to two cases from the ASM data set.

Table 3 shows observed and edited values for case number 65. Values of four important ratios calculated from the data in Table 3 are presented in Table 4. In this example the imputed ratio for BMH/BPW is 2.0 thousand hours. This corresponds more closely than the observed ratio to the average number of hours a production worker works in one year as given by the current year ratio, 1.7 thousand hours. Also, this imputed ratio corresponds to approximately a 40 hour work week for 50 working weeks per year.

The imputed ratio for the prior year average hourly wage rate, BWW/BMM, is $5.70 per hour. This rate is very much in line with the industry average, is well above the unedited $2.90 per hour figure, and corresponds closely to the current year hourly wage rate of $6.30. Similarly, the imputed prior year average salary, BWW/BPW, $11.1 thousand, is in line with the industry norm, well above the unedited prior year ratio, $3.3 thousand and corresponds closely to the current year ratio, $11.0 thousand.

Finally, the procedure imputes .4 for the current year ratio of voluntary to legal payments to fringe benefit programs. This figure is twice that of the original ratio and is identical to the prior year figure.

This example and others not shown suggest that the edit/imputation method identifies cases with ratios that are implausible and imputes for these values in such a way so as to restore the ratios to reasonable values. Very often these imputed ratios correspond closely to the prior year (or if the prior year ratio is bad, the current year) value.

The original and imputed variate values for our second example, case number 78, appear in Table 5 and the ratios associated with selected outlying variables appear in Table 623 In this case the imputed ratios are in rough agreement from current to prior year. This is quite an interesting result considering that 6 of the 14 variables were either missing or edited.

In this example the ratio OW/OE, denotes the average annual salary for other employees. Usually "other" employees is understood to mean executive or professional staff as opposed to production staff. In this regard the prior year annual salary of $9.2 thousand is apparently low for professional staff but is brought closer into line (to $20.6 thousand) with the industry average and the current year value, $22.1 thousand.

Because the current year man hour value, AMH, is missing the average production man hour per year figure, MH/PW, is imputed as 1.7, very close to the industry average and concordant with the prior year ratio, 1.8.

The procedure likes neither prior nor current year values for the production worker average hourly wage, WW/MH (because data required to compute these ratios are outlying and missing, respectively). The prior year ratio value seems to be too high, $7.60 per hour. Consequently, the procedure imputes $5.90 and $5.60 for current and prior year hourly wage: these values are concordant with each other and other variables in the case not edited and present.

Both the current and prior year average production worker annual salary WW/PW, has been edited. The original current year ratio, $17.1 thousand, appears to be high and the prior year ratio appears to be high also, $13.8 thousand, although not as extreme. The procedure imputes ratios are nearly congruent: $10.3 and $10.1 thousand.

However, with regard to the ratio of voluntary to legally required payments to fringe benefit programs, VP/LE, our rule that year to year ratios are at least approximately preserved is broken: the current imputed year ratio is three times smaller than the prior year figure.

2.5 Addition of Random Residuals to Imputed Values

Our imputation method described thus far fills in a case's missing and edited values by their conditional means given the case's present and unedited variable values. This procedure is easily implemented using the Beaton sweep operator and the EM algorithm. Kalton (1981), Santos (1981), Kalton and Kish (1981), Sedransk and Titterington (1980), and Kalton and Kasprzyk (1982) have noted that this type of mean value imputation leads to efficient estimation of univariate item means but distorts the distribution of the item: the concentration of imputed values at their conditional means creates spikes in the distribution. The consequence of this is the artificial reduction and underestimation of item variance. Whereas this procedure is efficient as far as estimation of means is concerned, it may be highly unsuitable from the point of view of producing a "clean" data set that fairly represents the underlying distribution. This objective is highly desirable when a single clean data set must be produced for various and diverse statistical analyses, an important Census Bureau activity.

In our context, the distortions created by imputing conditional means for missing and edited variables can be corrected by adding perturbations to the predicted means. If the normality assumptions underlying the EM algorithm are accepted, then the perturbations for an observation with m missing or edited items should have a zero-centered m-variate normal distribution. The appropriate dispersion matrix is the residual covariance matrix of the m missing or edited items given the p-m present and unedited items. An estimate of this matrix is already available in the swept covariance matrix calculated in the final E-step of the EM algorithm.

These perturbations provide consistent estimates of the variances and covariances from the observed and imputed data. The distributions of imputed items may still be distorted because of departures from multivariate normality. An alternative procedure that places less reliance on the normality assumption involves matching each incomplete case with a complete case and calculating for the complete case a vector of residuals from the regression of missing and edited variables on present and unedited variables. This vector then serves as a set of perturbations for the incomplete case in the match. The choice of matching criterion is an interesting issue. In the context of univariate nonresponse, Little and Samuhel (1983) argue for matching on the predicted mean from the regression of the missing item on the observations. In our multivariate setting, a natural generalization is to match on the vector of predicted means for the set of missing items, scaled by their residual covariance matrix. Colledge, Johnson, Paré, and Sande (1978) present a simpler matching scheme in an applied setting.

The addition of noise to the predicted means has an attractive feature in our problem, where the variables are measured on the log scale. Exponentiating the predicted means yields estimates on the original scale which are downward biased, whereas exponentiating the imputations with noise added yields consistent estimates. Simple adjustments for the bias in the exponentiated means can be developed (see, for example, Eddy and Kadane, 1982; Little and Samuhel, 1983) but are not included in our illustrative example.

3. Model Assumptions

Any missing data analysis makes assumptions about the missing data mechanisms, that is, the processes leading to missing values. Rubin (1976) formalizes these mechanisms in terms of the distribution of missing value indicators and given the hypothetical complete data matrix X. If this distribution depends on observed values of X but not missing values, he calls the missing data mechanism missing at random (MAR). If the distribution of the indicators does not depend on observed or missing values, he calls the mechanism missing at random and observed at random, or missing completely at random (MCAR). We can apply Rubin's (1976) theory to determine conditions for our problem.

Our test procedures for detecting outlying cases and outlying values within cases make the strong MCAR assumption, since they fall in the framework of what Rubin calls sampling distribution inferences. The imputations from the EM algorithm, however, are conditioned on the observed, unedited values of each case and are consistent estimates under the model if the (weaker) MAR assumption is satisfied. A complication in applying these definitions to our problem is that the observed data is dependent on the editing process, since an observed value changes to missing when it is edited out of the data set.

The MAR assumption cannot be evaluated without knowing the true values of the missing variables. The MCAR assumption can be assessed, however, by comparing the distributions of observed variables classified by data pattern. Table 7 shows summary statistics of the observed values classified by whether the observation was complete or incomplete. More detailed breakdowns were considered inadvisable given the modest sample size.
From this analysis it seems that with the exception of BLE, variables from observations with complete data only have roughly the same means values as variables from observations having missing values. Also, for the most part, the variability of variables from observations with complete data only is roughly the same as for variables from cases having missing values. For variables where the variances for these two groups tested to be significantly different, the variability for variables from observations with incomplete data is somewhat greater. However, for only one of the 16 variables (BLE) does the mean and variance test to be significantly different. In view of these results we feel that the MCAR assumption is by and large reasonable.

A second assumption our analysis makes is that the ℓn transformed data is multivariate normally distributed (MVN). This assumption is required only in the variable selection procedure when hypothesis testing is employed. However, as noted previously in Section 2, an alternative is to employ a graphical method for identifying outlying variables. If the graphical procedures are used the MVN assumption can be somewhat relaxed.

Figure 2 depicts the typical shape of distributions of our variables both in their original and ℓn transformed scale: the distribution in the original scale is typically skewed whereas the ℓn transformed distribution has a much more symmetic shape.

For our example Figure 3 gives a probability plot of the Wilson-Hilferty transformed Mahalanobis distances for each case. Under the MVN assumption of the ℓn transformed data, these transformed distances should be roughly normal. Except for the outlying cases in the upper right hand corner of Figure 3, the transformed distances behave as typical normal data does: it is linear in the probability plot. In Figure 4 the outlying variables have been edited. The transformed distances calculated from the remaining unedited data are given in this figure and behave nearly throughout as normal data should. However, cases beyond the cummulative probability of .95 in Figure 4 fall below the 45° line of inclination, suggesting a slight overediting of the data.

Since this article is presented in the context of the analysis of survey data, some remarks are warranted on the impact of complex survey designs involving unequal probability sampling, stratification and clustering. Our methods are not formulated to allow explicitly for the survey design. However survey design variables, such as dummy variables indicating strata, can be included as variables in the data matrix, where they serve as predictors for missing or edited variables (cf. Little, 1982). Alternatively, large data sets may be disaggregated into separate strata and our methods applied separately within each stratum. For unequal probability designs, the appropriate role of selection probabilities is a matter of debate, as in other areas of multivariate analysis. For example, a recent discussion of the role of design weights in regression is given in DuMouchel and Duncan (1983). Alternative strategies include the following: a) ignore the design weights; b) apply the methods proposed here with cases weighted by the inverse of the selection probabilities; c) ignore the design weights for editing and imputation but weight final estimates of the mean and covariance matrix of the variables by the inverse of the selection probabilities; or d) form strata that are homogeneous in the selection probabilities and include dummy variables for these strata as variables in the analysis. Motivations for the latter strategy are given in Rubin (1983) and Little (1983a, 1983b). A theoretical discussion of the relative merits of these strategies lies outside the scope of this article. The practical expedient of comparing the results from weighted and unweighted analyses appears worthwhile.

4. Linear Constraints

In many industrial examples, a multivariate case may include variables representing totals of other variables in the case. A limitation of the procedures described so far is their failure to take into account linear constraints between variables. Barnett and Lewis (1978) and Fellegi (1975) comment on the presence of outliers in the editing of multivariate data where such "pre-identified" relationships must hold. For example, in our ASM example, the variables OW and WW sum to a third recorded variable, the wages and salary of all workers (SW) which is checked from an independent data source. If the procedure causes OW or WW to be changed, then the linear constraint WW + OW = SW will not be satisfied. Modifications of the basic procedure are required to produce imputations that satisfy such linear constraints.

The modifications should reflect the nature of the linear constraints in a particular problem. Two issues need particular attention: (a) does the fact that a linear constraint is satisfied by the recorded values increase one's confidence in their validity? If two or more independent data sources are involved in the recorded values, then the answer to this question is probably yes; on the other hand, if the total is obtained by summing the individual components, or one of the components is found by subtracting the other components from the total, then the satisfaction of the constraint simply confirms the arithmetic and does not confer any particular validity to the recorded values.

The second issue requiring particular attention is: (b) is the total more reliably recorded than its components? In the ASM context, the SW variable is checked against official records and is regarded as more trustworthy than the values of other variables: it has been previously reported to the Internal Revenue Service, and is imputed from IRS administrative records when it is missing from the ASM survey. Furthermore, it is felt that salary and wages for production workers, SW, is very well known by each industrial establishment (since it is usually a major expenditure of the firm) as compared to the number of production workers, PW, which changes over the course of one year and consequently is less clearly defined.

Regardless of the answers to (a) and (b), we suggest that one of the variables in each constraint is dropped from our algorithm, to avoid problems of near-collinearity. At the conclusion of the algorithm, the value of the omitted variable may be changed if necessary to satisfy the linear edit constraint. If the linear constraint is not satisfied by the edited and imputed variables we propose an additional editing procedure following the imputation step. In this procedure the variable in the linear constraint may be changed that results in the smallest Mahalanobis distance for the case. The total might change in this procedure; this would not be allowed to happen if the answer to (b) is yes.

If the answers to (a) or (b) are yes, then further improvements to the algorithm can be achieved by assigning priority levels to the variables in the stepwise variable selection procedure. If (a) is answered as yes and the linear constraint is satisfied by the unedited values, then the variables in the constraint may be assigned lower priority for selection than other variables. If (b) is answered as yes, then the total should be included as a variable in the algorithm and assigned low priority for selection. These rules require straightforward modifications to handle data where some of the components of the linear constraint are missing.

For the ASM we describe how a procedure may be implemented that accounts for the linear constraint that ASW = AWW + AOW. Since there is a great amount of confidence in the recorded value of ASW, it should be included in the analysis. One of the two addends, say AOW, would then be dropped from the procedure to avoid problems of collinearity with the other variables, ASW and AWW. Its removal does not remove information regarding it from our analysis, however: information aobut it is carried in AWW and ASW and, obviously, by the difference ASW-AWW.

With ASW, AWW and the remaining varibles, μ and $\sum$ may be estimated as before via the ER algorithm. However, since ASW is believed to be correct, all other variables present in the case are removed before it in the variables selection procedure. Because it is unknown whether AWW is obtained via subtraction this variable receives no particular editing priority and is entered in the variable selection procedure as all other variables: it is removed or retained according to its influence in reducing the Mahalanobis distance.

Following this modified variable selection procedure imputed values for missing and edited values may be obtained as before in the final iteration of the E-M algorithm.

Finally, each case is checked to determine whether the linear constraint ASW = AWW + AOW is satisfied by the edited data. We propose the following four distinct strategies for "correcting" AWW or AOW when the edited data fail to satisfy the additive constraint. Each strategy is specific to a particular pattern of missingness for AWW and AOW:

(i) Both AWW and AOW are present and unedited. In this case we advocate editing the least likely of the two addends and imputing the difference between ASW and the more likely addend for the least-likely variable. This strategy may be implemented as follows: $\widehat{AWW} = ASW - AOW$ is computed. Then the Mahalanobis distances $D_{(\widehat{AWW})}$ and $D_{(AWW)}$ may be computed. These distances represent $-2\ell n$ likelihood of the case based on the imputed value $\widehat{AWW}$ and the actual value AWW, respectively. If $D_{(\widehat{AWW})} < D_{(AWW)}$ then AOW is more likely than AWW and the imputed value, $\widehat{AWW}$, may be retained. Otherwise, the original value of AWW is retained and $\widehat{AOW} = ASW - AWW$ is imputed for AOW and the linear constraint is satisfied.

(ii) AWW is originally present and is unedited but AOW is missing. In this case we suggest imputing $\widehat{AOW} = ASW - AWW$ for AOW.

(iii) AWW has been imputed by $\widehat{AWW}$ but AOW is present. In this case we compute $\widehat{\widehat{AWW}} = ASW - AOW$. If $D_{(\widehat{AWW})} > D_{(\widehat{\widehat{AWW}})}$ then we revise the imputed value of $\widehat{AWW}$ to be $\widehat{\widehat{AWW}}$. Otherwise, we impute $\widehat{AOW} = ASW - \widehat{AWW}$ for AOW.

(iv) Finally, if AWW has been imputed by $\hat{AWW}$ and AOW is missing than we suggest imputing $\hat{AOW}$ = ASW - AWW.

5. Summary

This article draws together statistical methodology in robust estimation, graphical procedures, outlier detection, and multivariate analysis and extends and applies these methods to the problem of editing and imputation for survey data. The ER (expectation-robust estimation) algorithm is presented (Section 2.3) yielding a procedure for robust estimation of μ and $\sum$ from multivariate data where values may be missing or outlying. Graphical and testing methods are given to identify outlying cases (Section 2.4) and outlying values within outlying cases (Section 2.5). The EM (expectation-maximization) algorithm is used in conjunction with the Beaton sweep operator to impute values for edited or missing values (Section 2.4). Also, the problems of addition of random residuals to imputed values (Section 2.6) and satisfaction of special linearity constraints by imputed values (Section 4) are discussed.

Data taken from the Annual Survey of Manufactures are used to illustrate the statistical methods we advocate. This data set is composed of attributes from the 1981 survey and each firm's corresponding prior year 1980 data. Although our methods do not require current and prior year data as we have for our illustration, particular care for editing and imputation should be taken where this type of data is available. For example, when current and prior year data are available, a case may be identified as being outlying as a result of the firm's sudden growth or decline from one year to the next. In these cases, data may represent bonafide information about the transition of the firm although it may be highly discordant from year to year.

This transition may be confirmed by recontacting the firm. Clearly, if a firm's data appear to be outlying as a result of a sudden year-to-year transition, its data should not be edited.

Much work remains to be done in establishing the statistical properties of the techniques we advocate. Also, further applied work is needed to adapt our methods to the specific edit and imputation requirements of real surveys. We believe that the examples presented in this article and other examples we have seen indicate that our edit/imputation method works well: values that are clearly improbable are edited and reasonable values are imputed.

TABLE 1. Estimated Means and Covariances of ℓn Transformed Data Computed from EM Algorithm. ASM Data.

ESTIMATED MEANS

BPW	BWW	BLE	BVP	BMH	APW	AWW	BOW	ALE	AVP	AMH	BOE	AOE	AOW
5.13	7.63	5.72	5.48	5.78	5.08	7.65	6.15	5.77	5.57	5.71	3.19	3.15	6.23

ESTIMATED COVARIANCE MATRIX

	BPW	BWW	BLE	BVP	BMH	APW	AWW	BOW	ALE	AVP	AMH	BOE	AOE	AOW
BPW	.97													
BWW	.97	1.09												
BLE	.94	1.08	1.20											
BVP	1.03	1.24	1.28	1.81										
BMH	.87	.90	.91	.98	.85									
APW	.96	.97	.97	1.05	.87	1.04								
AWW	.91	1.02	1.02	1.17	.85	.95	1.01							
BOW	.63	.74	.86	1.05	.63	.64	.72	1.13						
ALE	.89	1.00	1.08	1.18	.84	.94	.98	.81	1.05					
AVP	1.03	1.26	1.31	1.79	.98	1.09	1.23	1.07	1.22	1.95				
AMH	.89	.92	.96	1.03	.86	.98	.92	.66	.91	1.07	1.01			
BOE	.69	.72	.83	.95	.66	.70	.69	1.00	.80	.97	.69	1.18		
AOE	.72	.78	.89	1.03	.71	.74	.75	1.05	.86	1.07	.73	1.16	1.27	
AOW	.71	.82	.93	1.11	.69	.73	.80	1.13	.90	1.18	.72	1.06	1.19	1.32

TABLE 2. Estimated Means and Covariances of ℓn Transformed Data Computed from ER Algorithm. ASM Data.

ESTIMATED MEANS

BPW	BWW	BLE	BVP	BMH	APW	AWW	BOW	ALE	AVP	AMH	BOE	AOE	AOW
5.14	7.67	5.77	5.55	5.79	5.12	7.70	6.20	5.82	5.65	5.77	3.19	3.18	6.28

ESTIMATED COVARIANCE

	BPW	BWW	BLE	BVP	BMH	APW	AWW	BOW	ALE	AVP	AMH	BOE	AOE	AOW
BPW	.95													
BWW	.94	1.03												
BLE	.91	1.00	1.09											
BVP	.99	1.14	1.16	1.67										
BMH	.88	.89	.87	.92	.84									
APW	.91	.91	.89	.97	.84	.91								
AWW	.89	.99	.97	1.12	.85	.89	.97							
BOW	.63	.68	.78	.96	.61	.63	.69	1.05						
ALE	.88	.96	1.03	1.12	.84	.88	.95	.78	1.03					
AVP	.99	1.18	1.20	1.70	.94	.99	1.17	1.00	1.16	1.83				
AMH	.83	.84	.83	.89	.79	.81	.82	.59	.82	.92	.77			
BOE	.67	.67	.78	.88	.64	.67	.67	.99	.78	.92	.63	1.14		
AOE	.70	.70	.81	.92	.67	.70	.70	1.00	.81	.96	.65	1.13	1.16	
AOW	.68	.73	.83	1.00	.65	.68	.73	1.06	.85	1.05	.64	1.02	1.05	1.15

TABLE 3. Variable Selection Procedure for Case Number 65

TOTAL DISTANCE = 136.27
P-VALUE = .000

Variable	Recorded Value (Raw Scale)	Rank	Imputed Value	Incremental Decrease in Distance	Distance Remaining	P Value
BWW	473	1	925	56.68	59.03	.000
BPW	143	2	83	75.17	33.84	.004
ALE	209	3	115	83.51	22.47	.044
BOE	3	4		90.96	12.32	.328
BOW	58	5		95.91	5.57	.809
BVP	36	6		97.47	3.45	.914
AMH	140	7		98.06	2.64	.924
AVP	49	8		98.25	2.38	.890
BLE	99	9		98.39	2.20	.830
AOW	117	10		98.52	2.01	.743
AOE	6	11		99.28	.97	.812
AWW	887	12		99.56	.60	.745
BMH	162	13		99.57	.58	.448
APW	81	14		100.00	.00	1.000

Table 4. Important Ratios Involving Edited Variables from Case Number 65

	Current Year		Prior Year	
Ratio	Original Value	Imputed Value	Original Value	Imputed Value
MH/PW	1.7	-	1.1	2.0
WW/PW	11.0	-	3.3	11.1
WW/MH	6.3	-	2.9	5.7
VP/LE	.2	.4	.4	-

TABLE 5. Variable Selection Procedure for Case Number 78

TOTAL DISTANCE = 169.47
P-VALUE = .000

Variable	Recorded Value (Raw Scale)	Rank	Imputed Value	Incremental Decrease in Distance	Distance Remaining	P Value
BWW	3198	1	2335	15.21	143.69	.000
APW	35	2	192	50.72	83.51	.000
BOE	36	3	16	62.89	62.89	.000
AWW	597	4	1976	73.39	45.10	.000
ALE	123	5	344	90.57	15.99	.065
AVP	139	6		99.16	1.42	.986
BVP	434	7		99.53	.79	.993
BOW	330	8		99.68	.54	.991
BPW	231	9		99.73	.46	.978
BMH	419	10		99.81	.32	.958
BLE	368	11		99.96	.07	.965
AOE	18	12		99.96	.06	.800
AOW	408	13		100.00	.00	1.000
AMH	MISSING		1643			

Table 6. Important Ratios Involving Edited Variables from Case Number 78

	Current Year		Prior Year	
Ratio	Original Value	Imputed Value	Original Value	Imputed Value
OW/OE	22.1	-	9.2	20.6
MH/PW	-	1.7	1.8	-
WW/MH	-	5.9	7.6	5.6
WW/PW	17.1	10.3	13.8	10.1
VP/LE	1.1	.4	1.2	-

Table 7. Summary Statistics for MAR Analysis. Data are measured on ℓn scale.

	Observations With Complete Data Only			Observations With Incomplete Data			p-values	
Variable	n_1	$\overline{X}_1$	S_1^2	n_2	$\overline{X}_2$	S_2^2	H: $\mu_1=\mu_2$	H: $\sigma_1=\sigma_2$
APW	130	5.0	1.4	25	5.3	1.0	.2	.4
BPW	120	5.1	1.2	35	5.2	1.2	.8	.5
AOE	125	2.8	2.1	30	3.4	1.7	.04	.26
BOE	115	2.9	2.1	40	3.1	1.6	.08	.17
ALE	80	5.5	2.5	71	5.9	1.0	.04	<.001
BLE	75	5.1	3.6	77	5.9	1.6	<.01	<.001
AVP	71	5.6	1.1	83	5.8	1.2	.26	.26
BVP	67	5.3	1.6	87	5.6	1.8	.08	.32
AWW	110	7.5	2.2	44	7.8	1.2	.32	.01
BWW	100	7.5	1.8	55	7.7	1.2	.5	.06
AOW	105	5.6	4.7	49	6.3	1.5	.04	<.001
BOW	95	5.6	3.9	59	6.2	1.3	.04	<.001
ASW	90	7.9	.8	65	8.0	1.0	.56	.14
BSW	85	7.8	.9	69	7.9	1.0	.52	.28
AMH	63	5.4	2.6	88	5.7	1.3	.10	<.001
BMH	59	5.8	.7	95	5.8	.9	.88	.19

FIGURE 1. Missing Data Pattern Analysis

Code	Variable Name
A	BPW
B	BWW
C	BLE
D	BVP
E	BMH
F	APW
G	AWW
H	BOW
I	ALE
J	AVP
K	AMH
L	BOE
M	AOE
N	AOW

Number	Pattern
	ABCDEFGHIJKLMN
31	--------M-----
121	---------M----
92	----------M---
44	----------M---
78	----------M---
133	----------M---
35	-----------M--
153	--------MM----
70	--------MM----
42	-----M--MM--M-
43	-------M---MMM
53	-------M---MMM
40	-------M---MMM
55	-------M---MMM
41	-------M---MMM
29	------M-MMM--M
25	------M-MMM--M
152	MMMMM--M---M--
146	-----MM-MMM-MM

FIGURE 2 : Histograms of Original Scale and ln Transformed AWW and APW

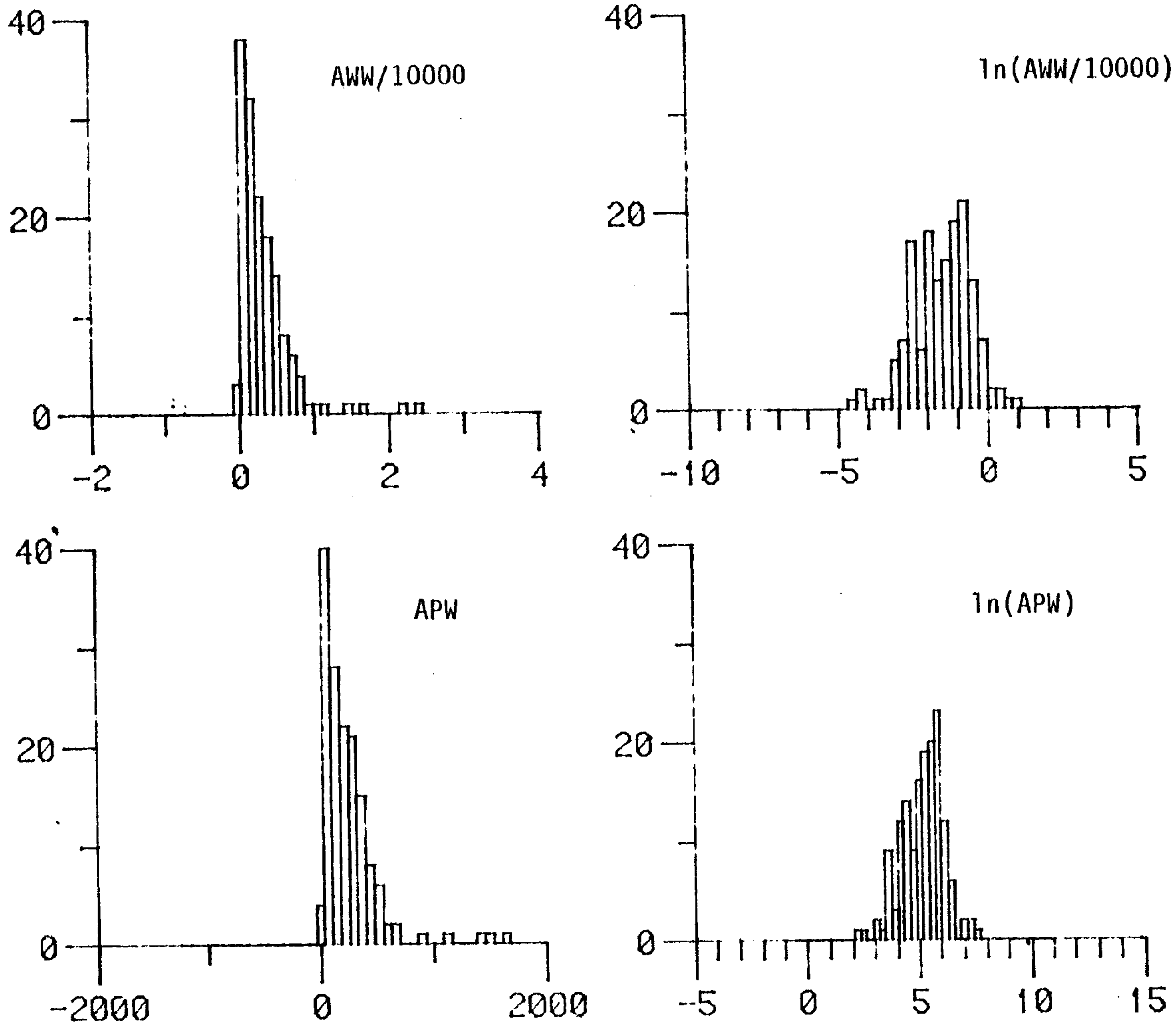


FIGURE 3. Normal Probability Plot of Transformed Distances Calculated from Contaminated Data

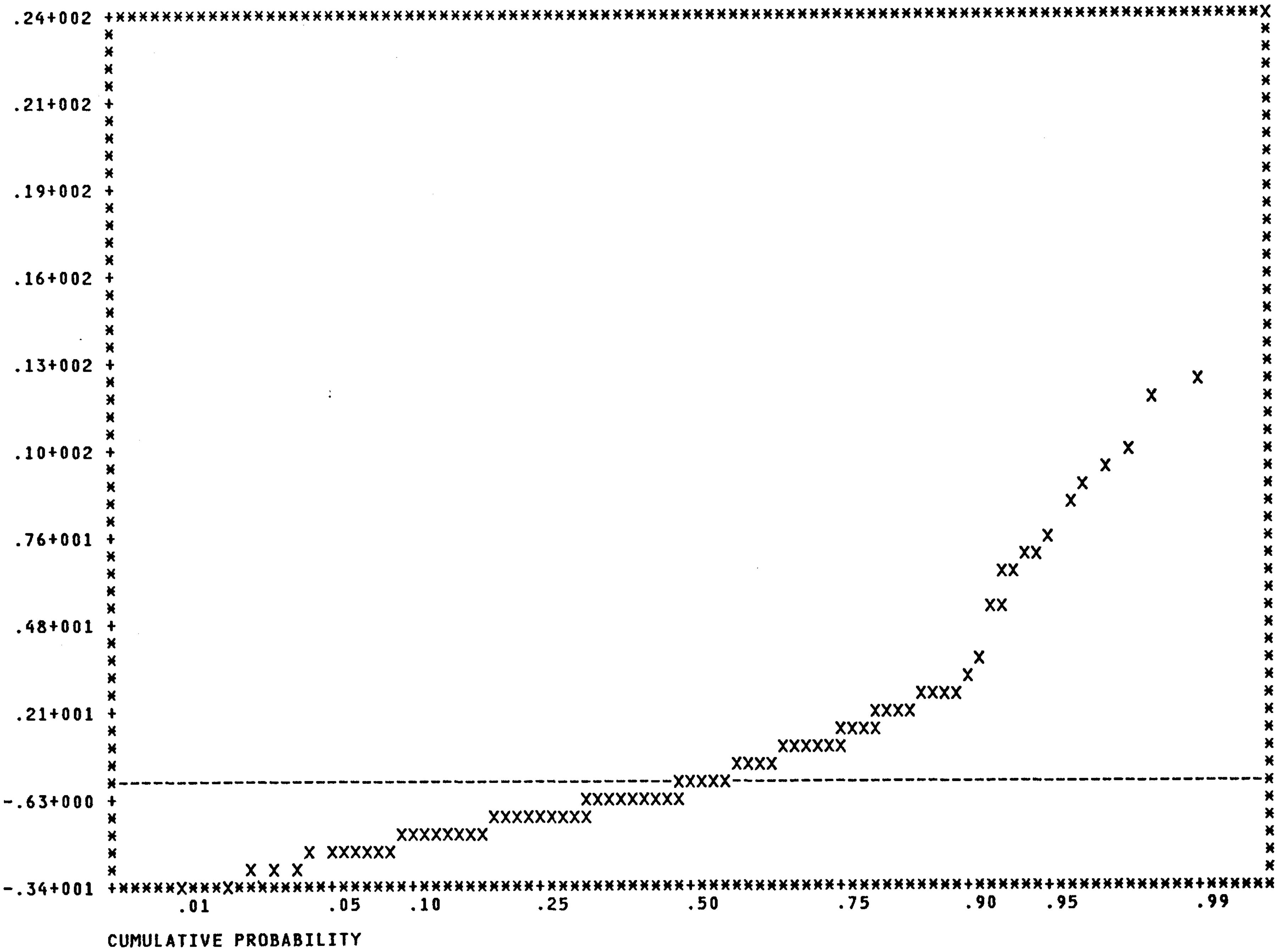
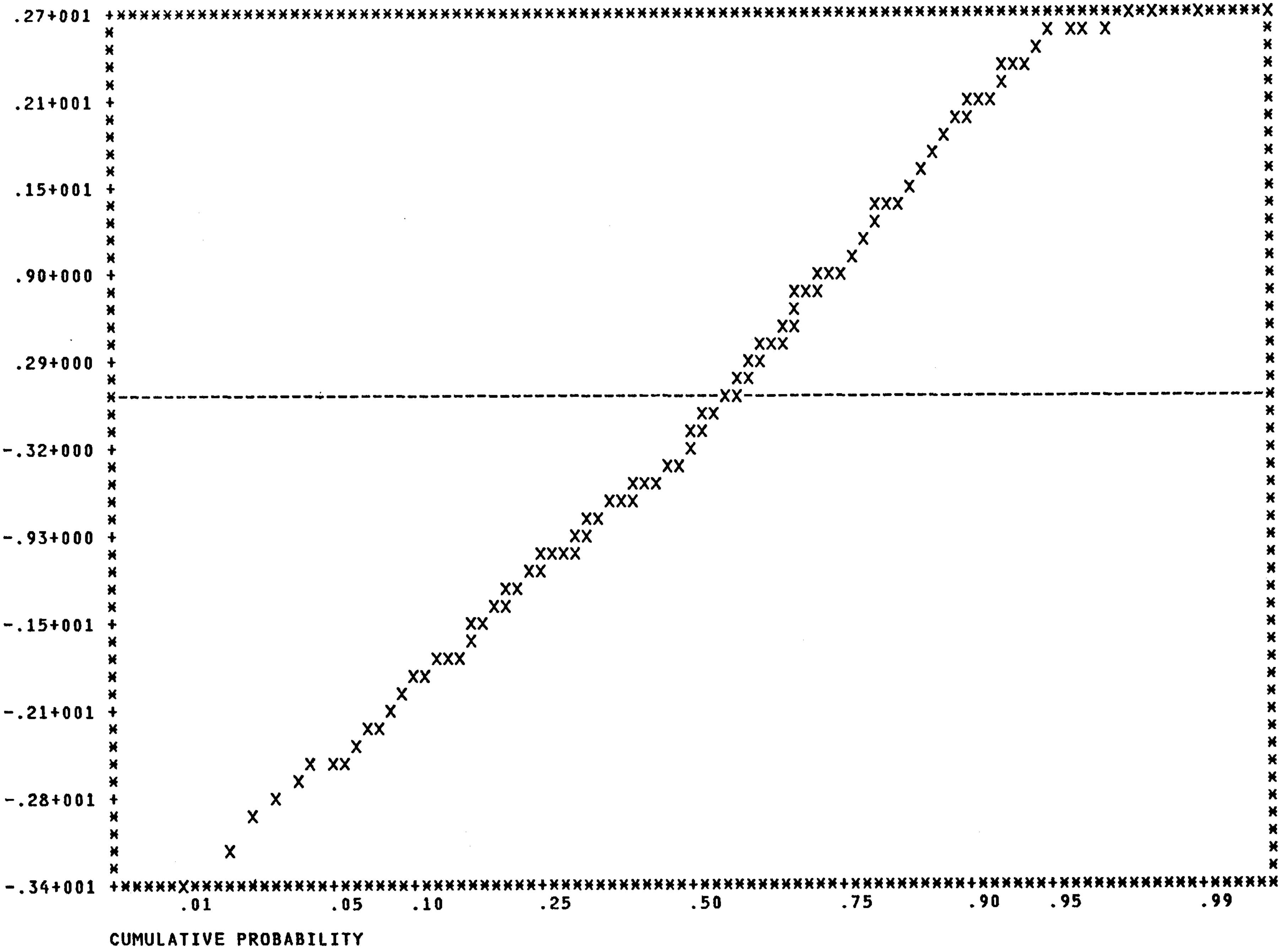


FIGURE 4. Normal Probability Plot of Transformed Distances Calculated from Cleaned Data



BIBLIOGRAPHY

Anderson, T.W. (1958). An Introduction to Multivariate Statistical Analysis. John Wiley and Sons: New York.

Barnett, V. and Lewis, T. (1978). Outliers in Statistical Data. John Wiley and Sons: New York.

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

Beaton, A.E. (1964). The Use of Special Matrix Operators in Statistical Calculus. Research Bulletin RB-64-51. Educational Testing Service: Princeton, New Jersey.

Beckman, R.J. and Cook, R.D. (1983). Outlier..............s, Technometrics, Vol. 25, No. 2 pp. 119-149.

BMDP (1981) BMDP Statistical Software, 1981 edition. Los Angeles: University of California Press.

Campbell, N.A. (1980). Robust Procedures in Multivariate Analysis I: Robust Covariance Estimation. Applied Statistics, 29, pp. 231-237.

Chapman, D.W. (1976). A Survey of Nonresponse Imputation Procedures, Proceedings of the Social Statistics Section, American Statistical Association, Washington, D.C.

Clarke, M.R.B. (1982). The Gauss-Jordon Sweep Operator with Detection of Collinearity. Applied Statistics, pp. 166-168.

Colledge, M.J., Johnson, J.H., Pare, R. and Sande, I.G. (1978). Large Scale Imputation of Survey Data. Proceedings of the Survey Research Methods Section, American Statistial Association, American statistical Association, Washington, D.C.

Dempster, A.P., Laird, M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. Journal of the Royal Statistical Society, Series B, 39, pp. 1-38.

Devlin, S.J., Gnanadesiken, R., and Kettering, J.R. (1981). Robust Estimation of Dispersion Matrices and Principle Components. Journal of the American Statistical Association, 76 (374), pp. 354-362.

DuMouchel, W.H. and Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples, Journal of the American Statistical Society, to appear.

Eddy, W.F. and Kadane, J.B. (1982). The Cost of Drilling for Oil and Gas: An Application of Constrained Robust Regression. Journal of the American Statistical Association, 77, 262-269.

Fellegi, I.P. and Holt, D. (1976). A Systematic Approach to Automatic Edit and Imputation. Journal of the American Statistical Association, 71 (353), pp. 17-35.

Fellegi, I.P. (1975). Automatic Editing and Imputation of Quantitative Data. (Summary). Bulletin of the International Statistical Institute, Vo. 46, pp. 249-253.

Frane, J.W. (1976). A New BMDP Program for the Identification and Parsimonious Description of Multivariate Outliers. Proceedings of the Statistical Computing Section, American Statistical Association, Washington, D.C.

Frane, J.W. (1978). Methods in BMDP for Dealing with Ill-Conditioned Data-Multicollinearity and Multivariate Outliers. Proceedings of the Computer Science and Statistics, A.R. Gallant and T.M. Gerig, editors. North Carolina State University, P.O. Box 5457, Raleigh, North Carolina.

Furnival, G.M. and Wilson, R.W.M. Jr., (1974). Regression by Leaps and Bounds. Technometrics, 16, pp. 449-551.

Gnanadesikan, R., and Kettering, J.R. (1972). Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. Biometrics, Vol. 28, pp. 81-124.

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. Wiley-Interscience: New York.

Goodnight, J.H. (1979). A Tutorial on the SWEEP Operator. The American Statistician, 33(b), pp. 149-158.

Hampel, F.R. (1973). Robust Estimation: A Condensed Partial Survey. Z. Wahr. verw. Geb., 27, pp. 87-104.

Hampel, F.R. (1974). The Influence Curve and Its Role in Robust Estimation. Journal of the American Statistical Association, 69, pp. 383-393.

Hartley, H.O. and Hocking, R.R. (1971). The Analysis of Incomplete Data (with discussion). Biometrics, 27, pp. 783-808.

Hawkins, D.A. (1974). The Detection of Errors in Multivariate Data Using Principal Components. Journal of the American Statistical Association, 69 (346), pp. 340-344.

Hawkins, D.A. (1980). Identification of Outliers. London: Chapman and Hall.

Huber, P.J. (1977). Robust Statistical Procedures. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania 19103.

International Mathematical and Statistical Libraries, Inc. (1982). The IMSL Library Reference Manual. Vols. 1-4. Houston, Texas.

Kalton, G. and Kasprzyk, D. (1982). Imputing for Missing Survey Responses. Procedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C.

Kalton, G. and Kish, L. (1981). Two Efficient Random Imputation Procedures. Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C. 20005.

Kalton, G. (1981). Compensating for Missing Survey Data. Income Survey Development Program. Department of Health and Human Services report. Survey Research Center, Ann Arbor, Michigan.

Kendall, M.G. and Stuart, A. (1968). The Advanced Theory of Statistics, Vols. I, II, and III. Griffin: London.

Little, R.J.A. (1979). Maximum Likelihood Inference for Multiple Regression with Missing Values: A Simulation Study. Journal of the Royal Statistical Society, Series B, 41, pp. 76-87.

Little, R.J.A. (1982). Models for Nonresponse in Sample Surveys. Journal of the American Statistical Association. Vol. 77, No. 378, pp. 237-250.

Little, R.J.A. (1983a). Estimating a Finite Population Mean from Unequal Probability Samples. Journal of the American Statistical Association, 78, 596-604.

Little, R.J.A. (1983b). Comment on "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," by M.H. Hansen, W.G. Madow and B.J. Tepping. Journal of the American Statistical Association, Vol. 78, No. 384, pp. 794-799.

Little, R.J.A., and Rubin, D.B. (1983). Incomplete data. Entry in Encyclopedia of the Statistical Sciences, Vol. 4. John Wiley: New York.

Little, R.J.A. and Samuhel, M.E. (1983). Alternative Models for CPS Income Imputation. Contributed paper to the Survey Methods Section of the 143rd Annual Meeting of the American Statistical Association, August 17, 1983.

Little, R.J.A. and Smith, P.J. (1983). Multivariate Edit and Imputation for Economic Data. Contributed paper to the Survey Methods Section of the 143rd Annual Meeting of the American Statistical Association, August 17, 1983.

Maronna, R.A. (1976). Robust M-Estimators of Multivariate Location and Scatter. Annals of Statistics, 4, pp. 51-67.

Orchard, T. and Woodbury, M.A. (1972). Missing Information Principle: Theory and Applications. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1. pp. 697-715.

Rubin, D.B. (1976). Inference and Missing Data. Biometrika, 63(3), pp. 581-592.

Rubin, D.B. (1983). Comment on "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," by M.H. Hansen, W.G. Madow and B.J. Tepping. Journal of the American Statistical Association, Vol. 78, No. 384, pp. 803-805.

Sande, J.G. (1982). Imputation in Surveys: Coping with Reality. The American Statistican, 36(3), part 1, pp. 145-152.

Santos, R.L. (1981). *Effects of Imputation on Complex Statistics*. Income Survey Development Program. Department of Health and Human Services report. Survey Research Center, Ann Arbor, Michigan.

Sedransk, J. and Titterington, D.M. (1980). *Nonresponse in Sample Surveys*. A report prepared for the Bureau of the Census under contract number JSA 80-12.

Shih, W.J. and Weisberg, S. (1983). Assessing influence in multiple linear regression with incomplete data. Contributed Paper, IMS/ENAR Biometric Society Meeting, Nashville, Tennessee.

U.S. Department of Commerce, Bureau of the Census, (1981). *Annual Survey of Manufactures.* Published in parts, beginning in 1983. U.S. Government Printing Office, Washington, D.C. 20402.

U.S. Department of Commerce, Bureau of the Census (1982-1983). *Statistical Abstract of the United States, National Databook, and Guide to Sources,* 103rd Annual Edition, Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

U.S. Department of Health, Education, and Welfare (1979). *Panel on Incomplete Data of the Committee on National Statistics*, Symposium on Incomplete Data: Preliminary Proceedings. Washington, D.C.

Wilks, S.S. (1963). Multivariate Statistical Outliers. *Sankhya*, Series A, Vol. 2, pp. 407-426.

Wilson, E.B. and Hilferty, M.M. (1931). Distribution of the Chi-Squared. *Proceedings of the National Academy of Science*, (U.S.A.), 17, p. 684.