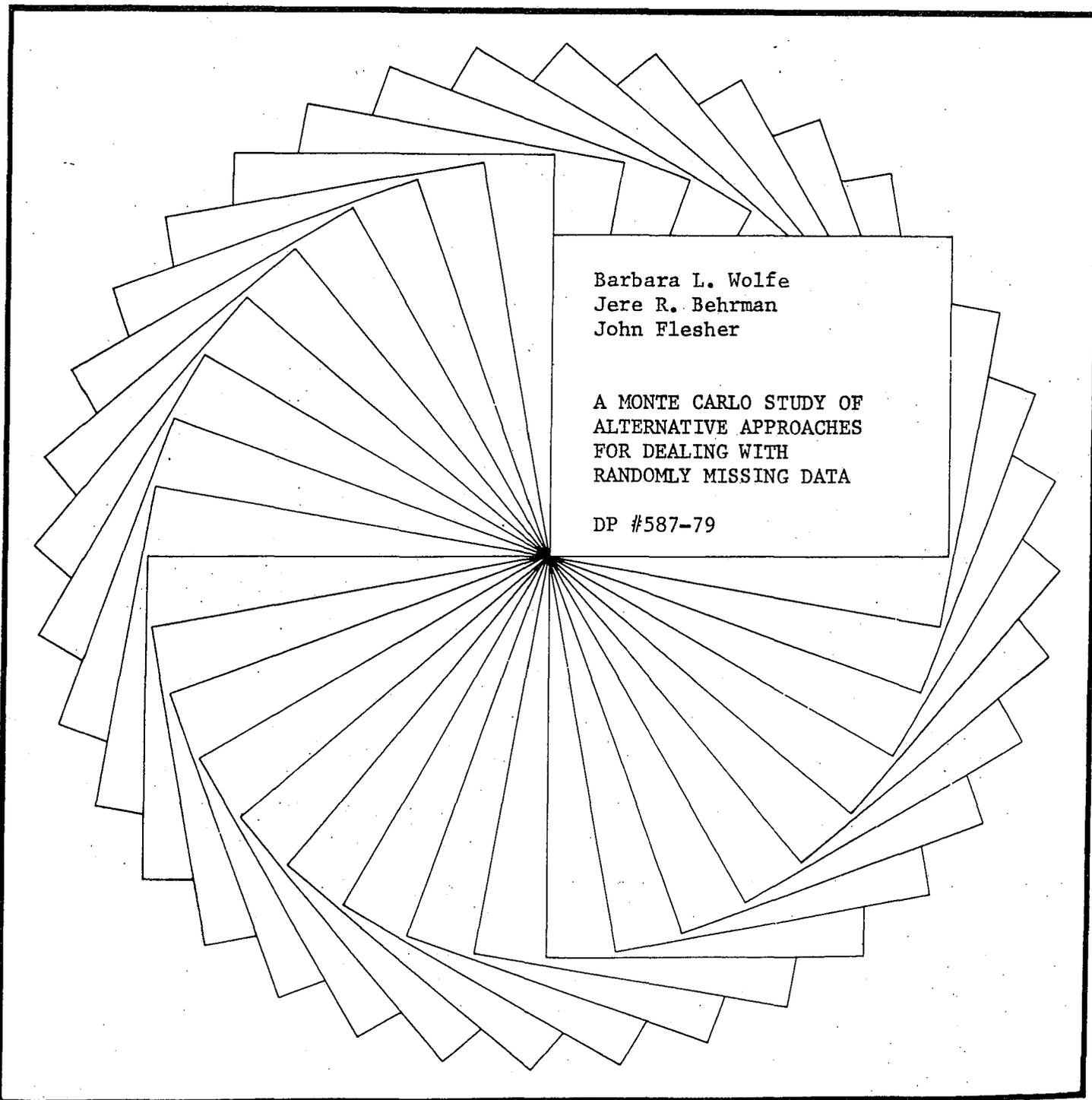




Institute for Research on Poverty

Discussion Papers



Barbara L. Wolfe
Jere R. Behrman
John Flesher

A MONTE CARLO STUDY OF
ALTERNATIVE APPROACHES
FOR DEALING WITH
RANDOMLY MISSING DATA

DP #587-79

A Monte Carlo Study of Alternative
Approaches for Dealing with Randomly
Missing Data

Barbara L. Wolfe
Jere R. Behrman
John Flesher

December 1979

The authors are assistant professor of preventive medicine and of economics and research affiliate of the Institute for Research on Poverty at the University of Wisconsin; professor of economics and research associate of the Population Studies Center at the University of Pennsylvania; and specialist in Computer Services of the Institute for Research on Poverty at the University of Wisconsin. This paper is one in a series resulting from a survey and research project to investigate the social, economic, and demographic roles of women in the developing country of Nicaragua, which has been funded by a variety of sources, including the Ford and Rockefeller Foundations, AID (Contract AID/otr-C-1571), and a J.S. Guggenheim Fellowship for Behrman. The authors would like to thank, but not implicate, the funding agencies, our co-principal investigators, Humberto Belli (director of the Centro de Investigaciones Sociales Nicaragüense) and Antonio Ybarra (Head, Division of Social Studies and Infrastructure, Banco Central de Nicaragua), various colleagues at the University of Pennsylvania and Wisconsin (particularly David Crawford and Robert Summers at the former and Arthur Goldberger at the latter), and project research associates at the University of Wisconsin, especially David Blau, Kathleen Gustafson, Michael Watts and Nancy Williamson. Behrman and Wolfe share equally in the major responsibility for this paper.

ABSTRACT

Applied social scientists often have missing values of some variables that they assume are random in their data sets. Yet the guidance for choosing among a number of methods that have been proposed to deal with randomly missing values is not very clear and often inconsistent. We present the results of a Monte Carlo experiment to test the alternative methods for dealing with different degrees of randomly missing observations in a set of social science data with about 1200 observations. Our Monte Carlo results and cost considerations lead us to several conclusions: (1) The use of means as a proxy for missing observations is unsatisfactory (2) The use of proxies estimated from auxiliary regressions may be best if the correlations in those regressions are sufficiently high and if enough emphasis is placed on small dispersions around the true values, but often instruments for the auxiliary regressions that result in sufficiently high correlations are not available. (3) Frequently, therefore, various methods of moments are preferable. (4) In many cases the simplest "listwise" method of moments, in which incomplete observations are simply dropped, may be the best choice because the cheapness of implementing such a strategy offsets its slight inadequacies relative to other more efficient methods of moments.

1. INTRODUCTION

Economists and other social scientists often encounter missing data problems. These may be grouped into at least three major categories: (1) Some desired variables simply may not be observable (e.g., permanent income, expected prices, natural ability and motivation).¹ (2) Some variables may not be observable for part of the actual (or conceptual) sample because of selection rules about inclusion in the sample or in a certain activity (e.g., one cannot determine market wages for individuals who select out of paid participation in the labor force, or the returns to university education for individuals who select not to go to universities).² (3) Some values of some variables may be missing randomly (e.g., random nonresponses to survey questions).

In this paper we focus on the third of these missing data problems. The discussion of this problem has a long history in statistics and a more limited history in the economic literature. Most of it is concerned with maximum likelihood solutions for dealing with incomplete data sets because of randomly missing values for some variables.³ We became interested in this problem because we are involved in a large study in which, we believe, there are randomly missing responses from respondents in a survey. Conversations with other applied researchers led to several suggestions about alternative approximate methods that have been adopted in previous empirical work to deal with data that are hypothesized to have some randomly missing values of some variables. However, as practitioners, we could find very little to guide us in choosing among these methods for a sample of our size (about 1200), given different proportions of missing observations and different apparent correlations between variables.

Therefore we designed an experiment to give us some insight. We report its results in hopes that they will be useful to other researchers with randomly missing values in their data sets. In section 2 we define alternative means of dealing with randomly missing observations that we consider. These encompass five methods-of-moments approaches that use different subsets of the available data (including perhaps the most used "listwise" option of disregarding all incomplete observations) and two proxy methods (using means or estimates from auxiliary regressions). In section 3 we comment on the costs of the alternative methods. In section 4 we describe our Monte Carlo experiment. In section 5 we present our results, and in section 6, our conclusions.

2. ALTERNATIVE PROCEDURES FOR DEALING WITH MISSING OBSERVATIONS

We focus on correlation coefficients between pairs of series, because they are the most commonly used measure of association between two series. The square of the correlation coefficient between x and y is the measure of the proportion of the variation in the dependent variable that is "explained" by the bivariate regression of x on y (or vice versa). If there were no missing values among our n observations, we could use the standard formula to calculate the correlation coefficient between two n -element vectors, x and y :

$$\begin{aligned}
 (1) \quad r &= \frac{\text{cov}(x,y)}{(\text{var}(x) * \text{var}(y))^{1/2}} = \frac{\Sigma(x-\bar{x})(y-\bar{y})/(n-1)}{(\Sigma(x-\bar{x})^2 * \Sigma(y-\bar{y})^2)^{1/2}/(n-1)} \\
 &= \frac{(\Sigma xy - \bar{x}\Sigma y - \bar{y}\Sigma x + n\bar{x}\bar{y})/(n-1)}{((\Sigma x^2 - n\bar{x}) * (\Sigma y^2 - n\bar{y}))^{1/2}/(n-1)}
 \end{aligned}$$

where subscripts referring to individual elements are suppressed to increase clarity, all sums are over the n observations, and \bar{x} and \bar{y} , respectively, refer to the means of x and y over n observations.

However, we are interested in a situation in which we do not have a complete set of n observations on both x and y , and in which, therefore, the standard formula in relation (1) is not directly applicable. Instead we consider the situation that is represented schematically in Figure 1. We have p observations on x , q observations on y , s observations on other relevant variables (z), m overlapping complete observations on x and y (but not necessarily on z), and r overlapping observations on x , y , and z . There are n total observations, $n-r$ of which are incomplete, in that at least the observation on x , y or z is missing. All missing observations are random.⁴ We explore the impact of selecting alternative estimators of the correlation between x and y under these conditions. The estimators that we consider can be grouped into two broad categories: methods of moments and proxies.

Methods of Moments

These alternatives use subsets of complete observations on x and y to estimate the equivalent of the moments in the numerator and denominator in relation (1). They do not use proxies for missing observations. They differ in the extent of information that they use to estimate relevant variances, covariances and means. We list them in order of increasing use of information:⁵

MML Method of Moments Based on Complete Overlapping Observations for All Variables, or Listwise Approach. This method drops all incomplete observations and calculates the correlation between x and y on the base of the r remaining totally complete observations.

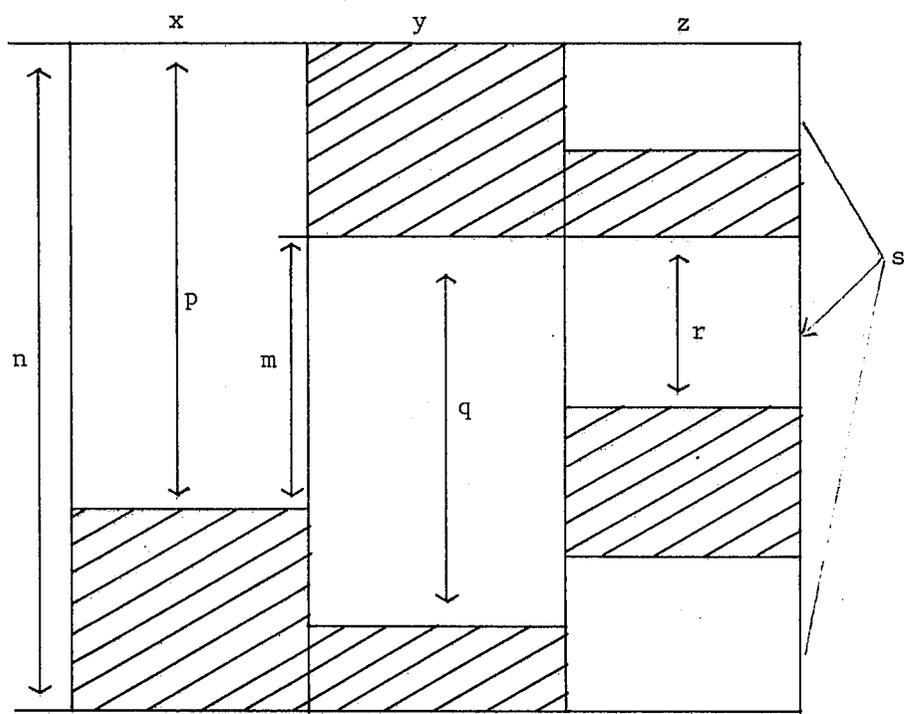


Figure 1

Shaded areas represent randomly missing observations

MM2 Method of Moments Based on Complete Overlapping Observations for the Two Relevant Series. This method uses all m complete overlapping observations on x and y , whether or not these observations are complete for z . If these observations are complete for z , m equals r and MM1 and MM2 are identical.

MM3 Method of Moments Based on Complete Overlapping Observations for the Two Relevant Series for Covariance in Numerator and All Available Observations for Variances in Denominator. This method uses the m complete overlapping observations on x and y for the numerator and the p complete observations for x and the q complete observations for y , respectively, for the variances in the denominator. If p equals q equals m so that either both x and y are observed or both are missing, MM3 reduces to MM2.⁶

MM4 Method of Moments Based on All Available Observations for Means in Covariance in Numerator and for Variances in Denominator. This method uses the p complete observations for x to estimate the mean of x that is used to calculate the covariance in the numerator and to estimate the variance of x in the denominator; it does the same for y . The covariance in the numerator then is calculated for the m complete overlapping observations for x and y , but the discrepancies from the means refer to the means based on p and q observations for x and y , respectively, not just to the m overlapping observations as in MM3. If p equals q equals m , MM4 reduces to MM2.

MM5 Method of Moments Based on All Available Observations for Means in Covariance in Numerator and for Variance in Denominator, with Scaling-Up. This method uses all of the available observations for each component of the correlation as does MM4, but each sum is "scaled-up" to represent n observations (e.g., $\sum x$ over p observations is multiplied by n/p , $\sum y$ over q observations is multiplied by n/q , etc.).

Proxy Methods

These methods replace the randomly missing observations with proxies, and then use relation (1) to estimate the correlation between x and y as if there were n complete observations. They differ in the choice of proxies.

PM1 Proxy Method Based on Means, or Zero-Order Method. In this alternative the proxy for missing values of x is the mean of x estimated from the p observations on x ; the proxy for missing values of y is the mean of y estimated from the q observations on y .

PM2 Proxy Method Based on Estimates from Auxiliary Regressions, or First-Order Method. This method uses estimates of x from auxiliary regressions that are based on existing observations of x with some regression(s) to fill in for missing observations of x ; it does the same for y .⁷

General Observations

Before we turn to our discussion of relative costs and of our experimental procedure and results, we make five observations about these approaches to estimating correlation coefficients if there are randomly missing observations.

First, these methods are widely used, because values of variables are very often missing in data sets, and researchers are willing to assume (and occasionally test) that the missing values are random.

Perhaps the most widely used (although not always acknowledged) procedure is the listwise approach or method of moments in which only complete observations for all variables (MMI) are used. Quite often, incomplete observations are simply dropped from the samples, and standard statistical programs are used to analyze the remaining complete observations as if they comprised the total data set.

The other methods of moments, with one exception, are available on widely used standard statistical packages. For example SPSS includes MM2 ("pairwise"), and BMD includes MM2 ("corpair"), MM3 ("copair"), and MM4 ("allvalue"). The exception is MM5, the one that uses all available observations and "scales-up;" this is an alternative that we devised early in our concern about randomly missing values for some variables.

The proxy methods probably are second only to the listwise method of moments (MML) in frequency of use. The use of the mean as a proxy (PML) is an option in canned statistical programs at many centers of such analysis (e.g., STAT JOB at the University of Wisconsin). If the use of such proxies is not an option in canned programs, researchers frequently "fill-in" the missing observations by means or by regression estimates, and then use the completed data matrix with standard computer programs as if there were not any missing values.

Second, the methods clearly differ in their efficiency. As we noted above, the five methods of moments use different subsets of data to estimate the correlations, and only the last two (MM4 and MM5) use all of the available information on x and y . The other methods of moments, which discard some of this information, are therefore less efficient. Since we present the methods of moments in order of their use of information, we also present them in order of increasing efficiency.

The proxy method that uses means (PML) is more efficient than the first three methods of moments in that all available information on x and y , respectively, is used to estimate the means. But it is less efficient because it does not use the available information about associations between deviations from the mean for x and for y in the m observations that are complete for

both. Given this trade-off, the relative efficiency of PM1 versus the methods of moments is not clear.

The proxy method that uses auxiliary regressions to obtain estimates for missing values (PM2) is more efficient than the one that uses means (PM1) in that it uses the information about the association(s) between the variable of interest and other variable(s). If the only right-hand-side variable that is in the auxiliary regression is the other variable that enters into the correlation of interest, then this method uses exactly the same information as do the most efficient of the methods of moments and must be equally efficient.⁸ If, in addition, other right-hand-side variables (say z) are used, this method uses more information and appears to be more efficient than any of the others. However, unless there are no additional missing values of the other right-hand-side variables (z), this appearance may be misleading, since the missing value problem is only pushed back a step to the auxiliary regression and there may be an efficiency trade-off in dealing with it.⁹ If r is less than m so that there are some missing values of z and these incomplete observations are dropped from the auxiliary regression, relative efficiency hinges upon whether or not more information is gained from using z than is lost by disregarding some information on the associations between x and y . If, on the other hand the auxiliary regression does not include the other correlates of interest (e.g., y , if the purpose is to estimate values for missing values of x ; as Hester [1976] and others have done using this method), then relative efficiency again hinges on the gains from using z versus the losses from not using the information on the association between x and y .

Third, the methods that use different subsets of observations to estimate the different components of the correlation coefficient are not guaranteed by

construction to give a correlation estimate that is less than or equal to one in absolute value. The three more efficient methods of moments (MM3, MM4, and MM5) all fall into this category. We note, however, that a correlation estimate less than or equal to one in absolute value is a mixed blessing. The first two methods of moments and the two proxy methods give such correlation estimates even if the maintained hypothesis--that the missing observations are random--is patently false. In such circumstances, methods MM3, MM4, and MM5 may give estimates greater than one in absolute value--a result that would warn the user that the missing observations are indeed not random.

In the multivariate extension of the missing value problem to the estimation of a correlation matrix, the analogue to a correlation coefficient greater than one in absolute value is a variance-covariance matrix that is not positive definite. The second method of moments (MM2), like MM3, 4, and 5, also may give this result even if all of the correlation coefficients calculated by it are less than one since the elements of the variance-covariance terms in general are estimated from different subsets of observations.¹⁰ We do not believe that the guarantee by construction of a positive definite covariance matrix for the listwise method of moments (MM1) and the two proxy methods (PM1 and PM2) is a strong recommendation for their use instead of the other methods of moments (MM2, MM3, MM4, and MM5).

Fourth, we suspect that the proxy methods (PM1 and PM2) are biased towards zero; we believe that, the use of imperfect proxies for randomly missing values may be analogous to introducing, into the right-hand-side variable in a bivariate regression, a random error that biases the estimated coefficient of that variable towards zero. If missing values of x are replaced by estimates, the estimated coefficient of x in the regression of y on x is biased downward

in absolute magnitude (and vice versa). Thus such procedures result in a downward bias in the product of the two coefficient estimates, and thereby in its square root, the correlation coefficient estimate. For the auxiliary regression proxy method (PM2) this "errors in the variable" bias is smaller, the more highly correlated the proxy is with the time series.

Fifth, the few previous related studies and conjectures do not lead us to a strong a priori ranking of all of these alternatives, although most of these studies agree in placing the use of the means as a proxy (PML) below a number of other alternatives.

Hester (1976) conducts Monte Carlo experiments with different sample sizes (50, 100, 500) for three methods to deal with randomly missing observations within a multivariate framework. On the basis of his results he prefers the method of moments (MM2) over the auxiliary regression proxy method (PM2), and prefers that over the use of means (PML). He also refers to Glasser's (1964) proof of the consistency of the multivariate extension of MM2.¹¹ However he finds that this method often does not lead to a positive definite variance-covariance matrix, particularly with smaller samples (see point three above).

Griliches, Hall and Hausman (1978) present some estimates that suggest that for their sample of 2200-2300 observations, the gains in efficiency from using the greater information in the second method of moments (MM2) if r is less than m do not lead to different substantive interpretations of their regression results than when they use the listwise procedure in the first method of moments (MM1) (an option that Hester did not consider). Haitovsky (1968) also concludes that the listwise method of moments (MM1) is preferable to the use of the means as proxies (PML).

But Griliches, Hall and Hausman (1978, p. 178) also explicitly criticize Hester's ranking of the method of moments (MM2) above the auxiliary regression proxy alternative (PM2). They characterize Hester's results as "somewhat strange and biased" because he excludes, from the eligible instruments for his estimated replacement for missing observations, the other right-hand-side variables in the multivariate regression. We note further that Hester's proxies are no more correlated with the variables that they are used to estimate than all of the variables in his regressions are correlated with each other; this, too, seems a strange restriction to use. Frane (1976), incidentally, advocates the use of some variant of the auxiliary regression proxy procedure as a practical alternative when there are not too many missing observations and when the correlations between the missing variable and other variables are sufficiently high, although he makes no explicit comparison with the methods of moments.

To complete the circle, Maddala (1977, p. 205) suggests that "in common-sense considerations" the auxiliary regression proxy method (PM2) "should do better than the classical least squares methods which discard all observations with gaps in data" (i.e., the listwise method of moments, MMI). At the risk of some oversimplification, one can summarize this literature as being **circular**: Maddala conjectures that PM2 is better than MMI that Griliches, Hall and Hausman illustrate in a particular case is as good as is MM2 that Hester concludes is superior (if usable) to PM2 that Maddala...

3. COSTS OF ALTERNATIVES

Before turning to our experimental procedure we make comments about the relative costs of the various methods. The objective is to obtain a correlation matrix from a set of variables that have randomly missing observations. Costs may involve: (1) computer programming, (2) computer time, and (3) other features of the already existing computer programs.

(1) If the computer programming is costly to a particular user and the BMD program (or some canned statistical package with the same features for dealing with randomly missing observations) is readily available, the last method of moments with scaling-up (MM5) is most expensive. The listwise method of moments (MM1) requires trivial programming or hand sorting to eliminate the incomplete observations; then a standard correlation program can be used. The other three methods of moments (MM2, MM3, MM4) are available as BMD options. The replacement of missing observations by means (PM1) requires trivial programming to fill the gaps; a standard correlation program can then be used (and some standard statistical packages have this whole procedure as an option). The auxiliary regression proxy approach (PM2) requires trivial programming or hand-sorting to identify the subset of observations for the estimation of the auxiliary regressions by a standard statistical package, trivial programming for the estimation of the proxies from the estimated regressions, and the use of a standard correlation program with the completed matrix once again, some standard statistical packages have this whole procedure as an option).¹²

If the BMD or some similar program is not available or is too costly to use, the programming involved in the middle three methods of moments (MM2, MM3, MM4) is on the same level as for MM5--somewhat more than trivial for a

programmer with limited experience, but not all that complicated. If however, the correlation matrix is not an end in itself, but an intermediate step that is used in an already existing program that a programmer finds complicated, the last four methods of moments may be relatively costly.

(2) In terms of computer time, the listwise method of moments (MM1) is the fastest and the use of means as proxies (PM1) is next. Among the other methods of moments, MM2 is slowest, because the relevant subsample for the mean and variance for a given variable generally changes for every correlation coefficient involving that variable since the missing observations for the other variables in general change. The required time for MM2 increases more than proportionally with the number of variables. For a similar reason in computing the means, MM3 is somewhat slower than MM4 and MM5. In the auxiliary regression proxy method (PM2) computing the auxiliary regressions themselves is time consuming; whether this method is quicker than the last four methods of moments depends on the number of variables that are used in the auxiliary regression in comparison to the number of variables in the overall correlation matrix, the proportion of missing observations, and the speed of the regression algorithm versus the specialized correlation coefficient algorithms.

(3) Packaged programs differ in their degree of accuracy, and in the specialized options (some of them costly and infrequently used) that they offer. Partly for this reason, more satisfactory discipline-specific and often institution-specific programs have been developed. If alternative methods of dealing with randomly missing variables do not already exist in such programs (or cannot be readily introduced), switching to a program like BMD to deal with the problem of randomly missing values of some variables means that the user must forego the advantages of the preferred program--including familiarity.

This is most likely to make the middle three methods of moments (MM2, MM3, and MM4) relatively costly.

We cannot aggregate all of these elements of cost since they obviously vary among users. For most economists at major research institutions, however, the listwise method of moments (MML) is probably cheapest, with the use of means as proxies (PML) very close behind. Because of familiarity with specialized programs, the auxiliary regression proxy method (PM2) comes next. The middle three methods of moments (MM2, MM3, MM4) follow because of their availability (especially MM2) in general statistical packages, but if programming has to be undertaken to use them, the last of the methods of moments (MM5) is no more costly. Although the relative differences in speed may be large, the absolute differences are not likely to be a major factor in choosing among the last four methods of moments unless the selected option must be heavily used or the data sets are very large.

4. OUR EXPERIMENTAL PROCEDURE

We begin with a set of six correlations that we assume are true. Monte Carlo studies often use correlations among constructed variables, but we use as our standard for comparison the actual correlations among data from every third family in the 1976 Wisconsin SIE data set for the following four variables: family income, family earnings, number of family members under age 65, number of family members over age 65. We use this standard of comparison because we believe that it is representative of the type of data that economists and other social scientists often use. The complete sample size (n) is 1238.

We present the complete sample correlations among these four variables in Table 1. Note that this set of variables has a wide range of "true"

correlations. - This enables us to see if the ordering of our seven alternative methods depends on the degree of the true underlying correlations.

We also are interested in whether or not the proportion of missing observations affects the relative merits of the alternative methods. Therefore, we consider the results for six different ranges of randomly missing observations of each variable: 0 to 5 percent, greater than 5 to 10 percent, greater than 10 to 15 percent, greater than 15 to 20 percent, greater than 20 to 25 percent, and greater than 25 to 30 percent.

We consider all seven methods. For the auxiliary regression proxy method (PM2), however, there is a critical question: how correlated is the actual series with the estimated one that replaces the missing observations? We consider three alternatives, in which this correlation is about 0.33 (low), about 0.67 (medium); and about .95 (high), respectively. These imply that the coefficients of determination between the estimated and the actual series are about .11, .45, and .90, respectively. In all three cases we assume that the right-hand-side variable in the auxiliary regression has no missing values.

For each combination of correlation, range of missing observations, and method, we conduct 20 independent random drawings for missing observations and then estimate the correlation coefficient by the indicated method. This implies a total of 6 correlations * 6 ranges of missing observations * 9 methods (including all three alternatives for PM2) * 20 independent drawings = 6480 correlation coefficients. Of course these can be grouped in various ways: 1080 alternative estimates for each of the 6 correlations, 720 estimates for each of the 9 methods, 1080 estimates for each of the 6 ranges, etc. We summarize these distributions by giving the difference between the true correlation for a group and the mean of the estimates in that group, as well as the standard deviations of the distributions of these differences.

Table 1

"True" Correlations Among Data on
 Every Third Family in 1976 Wisconsin SIE Data
 for n = 1238.

	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>
1. Family income	1.000			
2. Family earnings	0.898	1.000		
3. Number of family members under 65	0.475	0.592	1.000	
4. Number of family members over 65	-0.156	-0.370	-0.538	1.000

5. RESULTS

Table 2 summarizes the results for different methods across the 6 true correlation values. Table 3 summarizes the results for different methods across the 6 ranges of proportions of missing observations. In both of these tables in each cell we give the difference between the true value and the mean of the estimated values, and the standard deviation of the distribution of the differences between the true value and the estimated values. We first consider the deviations of the means of the distributions from the true value and the sensitivity of these deviations to alternative methods, true correlations, and proportions of missing observations. We then consider the standard deviations of these distributions of discrepancies from the true values and their sensitivity to the same characteristics. Table 4 illustrates the ratings by these criteria.

Deviations of Distribution Means from True Values

Smaller absolute deviations between the means and the true values clearly are desirable, ceteris paribus.

Across methods: The last row of Table 2 summarizes the results across methods on an overall basis -- across all 6 true correlation coefficients and all 6 ranges of missing observations. By this criterion the methods of moments are preferable to the proxy methods. But the absolute differences among the overall deviations between the means and the true values are all within one standard deviation of the distribution of discrepancies from the true values for the method (MM4) that has the smallest absolute deviation between the means and the true values.

Table 2

Means and Standard Deviations of Distributions of Discrepancies for
Alternative Methods of Estimating Correlation Coefficients with
Randomly Missing Data and Alternative "True" Correlations.

"True" correlation (from table 1)	Method of Moments					Proxy Methods				Total Across Methods
	Listwise MM1	MM2	MM3	MM4	MM5	Means PM1	Auxiliary Regressions PM2-low correlation PM2-medium correlation PM2-high correlation			
X2X1 .898	.00157 .026	.00022 .018	.00079 .041	-.00085 .041	-.00094 .061	.13901 .086	.12053 .073	.07923 .047	.01366 .008	.03907 .074
X3X1 .475	-.00311 .039	-.00112 .027	.00093 .024	.00089 .024	-.00099 .042	.07496 .044	.06798 .041	.03932 .025	.00886 .006	.02086 .044
X3X2 .592	-.00135 .033	-.00179 .018	-.00066 .019	-.00067 .019	-.00005 .037	.09152 .052	.08524 .049	.05242 .030	.00902 .006	.025967 .049
X4X1 -.156	.00035 .039	.00049 .025	.00032 .024	.00036 .024	.00157 .029	-.02374 .024	-.01974 .022	-.01238 .014	-.00303 .004	-.00620 .026
X4X2 -.370	.00020 .025	.00057 .016	.00094 .017	.00099 .017	.00025 .019	-.05676 .036	-.05158 .033	-.03300 .022	-.00882 .006	-.01635 .032
X4X3 -.583	-.00279 .020	-.00054 .011	-.00048 .018	-.00047 .018	-.00135 .015	-.09122 .053	-.08259 .048	-.04879 .030	-.01087 .007	-.02659 .045
Total across correlations	-.00086 .031	-.00036 .020	.00005 .025	.00004 .025	-.00025 .037	.02230 .099	.01998 .089	.01277 .056	.00147 .012	.00613 .053

Note: Each cell contains the mean discrepancy and the standard deviation of the discrepancies. See Section 2 for the definition of the methods. See Section 3 for a description of the experimental procedure. Each method correlation cell is based on 120 observations. The totals across methods are based on 1080 observations. The totals across correlations are based on 720 observations. The overall total is 6480 observations.

Table 3

Means and Standard Deviations of Distributions of Discrepancies
for Alternative Methods of Estimating Correlation Coefficients
with Randomly Missing Data for Alternative Proportions of
Total Data Set.

Range of Percentage of Missing Observations	Methods of Moments					Proxy Methods				Total
	Listwise MM1	MM2	MM3	MM4	MM5	Means PM1	Auxiliary Regressions			
							PM2-low correlation	PM2-medium correlation	PM2-high correlation	
0 to 5	.00123 .010	.00173 .007	.00125 .009	.00125 .009	.00091 .013	.00568 .022	.00523 .020	.00327 .013	.00054 .003	.00234 .013
5+ to 10	-.00302 .018	.00004 .014	.00109 .016	.00109 .016	.00280 .022	.01257 .048	.01050 .042	.00715 .028	.00052 .006	.00364 .027
10+ to 15	.00068 .025	.00038 .017	.00052 .020	.00053 .020	-.00048 .026	.01845 .075	.01744 .067	.01065 .040	.00138 .009	.00550 .040
15+ to 20	.00138 .032	.00243 .018	.00076 .027	.00076 .027	-.00339 .040	.02626 .097	.02330 .087	.01618 .055	.00207 .012	.00775 .053
20+ to 25	.00028 .037	-.00304 .024	-.00519 .028	-.00522 .028	-.00453 .044	.02899 .128	.02661 .115	.01756 .072	.00231 .014	.00636 .068
25+ to 30	-.00512 .048	-.00369 .029	.00185 .038	.00185 .038	.00317 .057	.04184 .158	.03674 .137	.02180 .087	.00200 .018	.01116 .084

Note: Each cell contains the mean discrepancy and the standard deviation of the discrepancies. See Section 2 for the definition of the methods. See Section 3 for a description of the experimental procedure. Each method correlation cell is based on 120 observations. The totals across methods are based on 1080 observations. The totals across correlations are based on 720 Observations. The overall total is 6480 observations.

Table 4

Ratings of Alternative Methods

Method	Criteria		
	Mean discrepancy from true value	Standard deviation of discrepancies	Probable costs
MM1	good	good	best
MM2	good	good	moderate
MM3	best	good	moderate
MM4	best	good	moderate
MM5	good	good	worst
PM1	worst	worst	best
PM2-low	worst	poor	moderate
PM2-medium	fair	fair	moderate
PM2-high	good	best	moderate

Among the methods of moments, there is a weak suggestion that those that use more information rank higher than do those that use less. The overall ordering, by this criterion, is the same as that for the overall discrepancy between the means and true values, except that MM5 and MM3 are reversed. This suggests that the more extended pairwise methods of moments (MM3, MM4 and MM5) may be somewhat better than the one based on complete observations for the two variables under consideration (MM2), and the pairwise one based on complete observations for the two variables under consideration (MM2) may be better than just using observations that are complete for all variables as in the listwise option (MM1).

Among the proxy methods the ordering is not surprising: the auxiliary regression proxies (PM2) with high, then medium, and then low correlations, and finally the use of the mean as a proxy (PM1).

Across true correlations: The columns of Table 2 give some indication of the sensitivity of the results to different true correlations. Both the general overall ranking of the methods of moments above the proxy method and the ranking among the proxy methods themselves are preserved. However the overall ranking among the methods of moments is not. In fact, each of the five methods of moments has the smallest absolute difference between the mean and the true value for at least one true coefficient.

For all methods of moments and proxy methods, the absolute deviations tend to be relatively small for smaller true correlations. But this is a tendency, not a tight pattern. For example for the methods of moments with full use of information and scaling up (MM5), the largest discrepancy is for the smallest true correlation ($X4X1 = -0.156$).

For the auxiliary regression and the use of the mean proxy methods (PM2, PM1), the patterns are as we anticipated in Section 2, above. The signs indicate that the biases are toward zero. The absolute magnitudes are systematically inversely associated with the correlation between the proxy and the variable that is estimated by it.

Across proportions of missing observations: The columns of Table 3 indicate the sensitivity of the results to changing proportions of randomly missing observations. The overall ranking among the proxy methods (PM1, PM2) is preserved as the proportion of missing observation changes, but the overall ranking among the methods of moments is not. With one exception, each of the methods of moments has the smallest absolute difference between the mean and the true correlation for at least one of the 6 ranges of missing observations.¹³

Of more significance is the fact that the methods of moments are not better than PM2 - high correlation for all ranges of missing observations. In fact, PM2-high correlation is better by this criterion than all of the methods of moments for 0 to 5 percent missing observations, better than all but one for greater than 5 to 10 percent and greater than 20 to 25 percent missing observations, and better than 2 or 3 of the methods of moments for greater than 15 to 20 and greater than 25 to 30 percent missing observations (see Table 3). On the other hand, the medium- and low-correlation versions of PM2 and the use of means in PM1 do not lead to better results, by this criterion, than do the methods of moments for any of the 6 ranges of missing observations (although in some cases the order of magnitude of the discrepancy between the distribution mean and the true correlation is the same).

Not surprisingly, the smaller absolute discrepancies between the means and true values tend to be for smaller proportions of missing observations.

This pattern is quite tight for the proxy methods, with only one small and probably statistically insignificant inversion.¹⁴ It is much less tight for the methods of moments. The cases with 0 to 10 percent of the observations missing contain the smallest absolute discrepancy only for MM2 (although the second smallest discrepancy for MM5 and the third smallest one for all of the methods of moments). The cases with 20 to 30 percent missing observations include the largest absolute discrepancy for all of the methods of moments and the next largest for 3 of them, but also the smallest one for the list-wise method of moments (MM1).

Standard Deviations of Distributions of Discrepancies from True Values

A tight distribution and small standard distribution are preferred, ceteris paribus. But there may be a trade-off between the position of the mean of a distribution relative to its true value and the degree of dispersion of the estimates.

Across methods: The last row of Table 2 gives the overall summary statistics across methods. The smallest overall standard deviation is for the auxiliary regression with high correlation (PM2-high), followed by the various methods of moments (MM2, MM3, MM4, MM1, MM5), and the other proxy methods (PM2-medium, PM2-low, PM1). Among the proxy methods the ranking is in order of the correlation of the proxies with the variables that they are estimating.

Among the methods of moments, the reason for this ordering is not transparent. It does not seem to reflect the amount of information that each method uses.

Across true correlations: The columns of Table 2 indicate the sensitivity of the standard deviations of the results to different true correlations. The auxiliary regression proxy with high correlation (PM2-high) is best for each true correlation. The proxy methods tend to be better, relative to the methods of moments, by the standard deviation criterion for lower absolute values of the true correlations. For the smallest true correlation ($X4X1 = .156$), for example, all of the proxy methods have smaller standard deviations than do any of the methods of moments.

Among the methods of moments, MM2, based on complete observations both series, has the smallest standard deviation for 4 of the 6 true correlations while MM3 and MM4, which use all available observations, have almost identical standard deviations that tend to be the next smallest (and smallest for two cases). The listwise option (MM1) and the scaling-up method (MMS) tend to be worst, although each has the second smallest standard deviation among the methods of moments for one true correlation.

Within methods, finally, there is a weak association between the absolute sizes of the true correlation coefficients and these standard deviations.

Across proportions of missing observations: The columns of Table 3 reveal the sensitivity of the standard deviations to the proportions of missing observations. The overall ordering among methods is preserved exactly among all of these ranges. Within methods there is a strong positive association between the proportions of missing observations and the sizes of the standard deviations.

6. CONCLUSION

Our most confident conclusions are that the use of the mean as a proxy (PM1) is the least satisfactory alternative, that the mean and auxiliary regression methods (PM1 and PM2) tend to give estimates of the correlation coefficient biased towards zero, that the auxiliary regression proxy procedure (PM2) is more satisfactory the higher is the correlation in the auxiliary regression, and that our scaling-up method of moments (MM5) tends to be the least satisfactory among the methods of moments.

Among the methods of moments, those that use more information (MM3, MM4 and MM5) tend to be somewhat better by the criterion related to the average size of the discrepancy. However, this result is not very robust, as the size of the true correlations and the proportions of missing observations vary. The more weight that is placed on the dispersion of the estimates, moreover, the better is the method of moments that uses only overlapping observations between the two series (MM2) and the worse is our scaling-up method (MM5) which uses considerable information. By either criterion, just using complete observations for all variables (MM1) ranks relatively low among the alternative methods of moments. But the relatively low ranking of this frequently used (but not always explicitly acknowledge) listwise alternative should not be overstressed. Under some conditions, particularly if the criterion related to the average discrepancy from the true value is emphasized, this approach does better or about as well as the alternatives, the differences in ranking may not be statistically significant and it may be much easier to implement than the other methods of moments.

The results for the auxiliary regression proxy method (PM2) seem to warrant neither Maddala's optimism nor Hester's pessimism (see Section 2).

This method is preferable to the methods of moments if sufficient emphasis is placed on smaller dispersions, the correlation in the auxiliary regression is large enough, the true correlations are close enough to zero, and perhaps if the proportion of missing observations is small enough. However, in many cases instruments that are correlated nearly as highly with the variables to be estimated as in our PM2-high alternative are simply not available.

Thus we have conducted Monte Carlo experiments for 7 different methods of dealing with randomly missing observations for social science data with a range of "true" correlations for about 1200 observations. We conclude that unless very good proxies are available for the auxiliary regression method (PM2), the methods of moments seem to be the best alternatives -- and they may be best even if very good proxies are available. Among the various methods of moments, in light of both bias and dispersion criteria, MM3 and MM4 appear slightly more satisfactory -- but the more that dispersion is emphasized, the better is the MM2, based on overlapping observations between two series. And the often followed listwise method of using complete observations only (MM1), although less satisfactory, is not all that bad -- particularly once the ease of computation is considered.¹⁵ In fact our results suggest that for our sample of about 1200, nothing substantive is likely to result from using this listwise option instead of other more efficient method of moments -- a conclusion the Griliches, Hall and Hausman (1978) also reach for a particular model and a particular sample (not a Monte Carlo study) with about twice as many observations.

APPENDIX: Formulae for Alternative
Correlation Estimates

As in Section 2, let:

p = number of observations on x ,

q = number of observations on y ,

s = number of observations on z ,

m = number of observations with values present for x and y ,

r = number of observations with values present for x , y
and z ,

n = total number of observations ($n-r$ of which are incomplete
in sense that the value of x , y , and/or z is randomly
missing).

We use these letters as superscripts to designate which observations
are used in a particular sum (e.g., \bar{x}^m is the mean value of x over the
 m observations for which both x and y are observed, Σx^{p-q} is the sum
of x over the observations for which x is observed but y is not observed).

We first define the following covariances and variances:

$$\text{cov } (x,y)1 = \Sigma (x^r - \bar{x}^r)(y^r - \bar{y}^r)/(r-1) = (\Sigma (xy)^r - r \bar{x}^r \bar{y}^r)/(r-1)$$

so that only the r complete observations for all
variables in the larger set (i.e., x , y , and z)
under analysis are used.

$$\text{var } (x)1 = \Sigma (x^r - \bar{x}^r)^2/(r-1) = (\Sigma (x^r)^2 - r (\bar{x}^r)^2)/(r-1)$$

so that all sums are defined over the r complete ob-
servations for the larger set (i.e., x , y , and z)
of variables under analysis (and likewise for var
(y)1).

$$\text{cov } (x,y)2 = \Sigma (x^m - \bar{x}^m)(y^m - \bar{y}^m)/(m-1) \text{ so that all sums are defined}$$

over m complete overlapping observations for x and y .

var (x)2 = $\Sigma (x^m - \bar{x}^m)^2 / (m-1)$ so that all sums are defined over m complete overlapping observations for x and y (and likewise for var (y) 2).

cov (x,y)3 = $\Sigma (x^m - \bar{x}^p) (y^m - \bar{y}^q) / (m-1)$
 = $(\Sigma (xy)^m - \bar{x}^p \Sigma y^m - \bar{y}^q \Sigma x^m + m \bar{x}^p \bar{y}^q) / (m-1)$
 so that the mean for x is calculated over all p available observations, the mean for y is calculated over all q available observations, and the deviations from the means are for the m overlapping complete observations for x and y.

var (x)3 = $\Sigma (x^p - \bar{x}^p)^2 / (p-1) = (\Sigma (x^p)^2 - p(\bar{x}^p)^2) / (p-1)$

so that all sums are calculated over all p available observations of x (and likewise for all q observations of y for var (y)3).

cov (x,y)4 = $((n/m) \Sigma (xy)^m - n \bar{x}^p \bar{y}^q) / (n-1)$ so that the crossproduct term is calculated over all m complete overlapping observations for x and y and then scaled up by n/m to represent n observations and the means for x and y, respectively, are calculated over the maximum available number of observations (p and q, respectively) and then scaled up (by multiplying Σx^p and Σy^q by n/p and n/q, respectively) to represent n observations.

var (x)4 = $((n/p) \Sigma (x^p)^2 - n(\bar{x}^p)^2)/(n-1)$ so that the squared term and the mean are calculated over the p available observations on x and then scaled up to represent n observations by multiplying by n/p (and likewise by using all q observations on y and scaling up by n/q for var (y)4).

cov (x,y)5 = $(\Sigma (xy)^m + \bar{y}^q \Sigma x^{p-q} + \bar{x}^p \Sigma y^{q-p} + (n-m-(p-q))-(q-p)) \bar{x}^p \bar{y}^q - n \bar{x}^p \bar{y}^q)/(n-1)$ so that all missing observations are replaced by the respective mean values that are defined over the largest possible set of observations (i.e., p for the mean of x and q for the mean of y).

var (x)5 = $(\Sigma (x^p)^2 + (n-p) (\bar{x}^p)^2 - n(\bar{x}^p)^2)/(n-1)$ so that all missing observations are replaced by the mean of x defined over all p observed values of x (and likewise for var (y)5 with all missing values replaced by mean of y over all q observed values of y).

cov (x,y)6 = $(\Sigma (xy)^m + \Sigma (x\hat{y})^{p-q} + (x\hat{y})^{q-p} + (\hat{x}\hat{y})^{n-p-q} - n(\Sigma x^p + \Sigma \hat{x}^{n-p}) (\Sigma y^q + \Sigma \hat{y}^{n-q})/n^2)/(n-1)$ so that each missing value of x is replaced by an estimate, \hat{x} , and each missing value of y is replaced by an estimate, \hat{y} , (which reduces to cov (x,y)5 if means \bar{x}^p and \bar{y}^q are used as the estimates \hat{x} and \hat{y} , respectively).

$$\text{var (x)6} = (\Sigma(x^p)^2 + \Sigma(\hat{x}^{n-p})^2 - n(\Sigma x^p + \Sigma \hat{x}^{n-p})/n)/(n-1)$$

so that each of the n-p missing values of x is replaced by an estimate of x (which reduces to var(x)5 if \hat{x} equals \bar{x}^p) -- and likewise for var (y)6 in which each of the n-q missing values of y is replaced by an estimate \hat{y} .

We now define the alternatives that are presented in Section 2.

Methods of Moments

$$\text{MM1:} \quad r1 = \frac{\text{cov (x,y)1}}{(\text{var (x)1} * \text{var (y)1})^{1/2}}$$

$$\text{MM2:} \quad r2 = \frac{\text{cov (x,y)2}}{(\text{var (x)2} * \text{var (y)2})^{1/2}}$$

$$\text{MM3:} \quad r3 = \frac{\text{cov (x,y)2}}{(\text{var (x)3} * \text{var (y)3})^{1/2}}$$

$$\text{MM4:} \quad r4 = \frac{\text{cov (x,y)3}}{(\text{var (x)3} * \text{var (y)3})^{1/2}}$$

$$\text{MM5:} \quad r5 = \frac{\text{cov (x,y)4}}{(\text{var (x)4} * \text{var (y)4})^{1/2}}$$

Proxy Methods:

$$\text{PM1:} \quad r6 = \frac{\text{cov (x,y)5}}{(\text{var (x)5} * \text{var (y)5})^{1/2}}$$

$$\text{PM2:} \quad r7 = \frac{\text{cov (x,y)6}}{(\text{var (x)6} * \text{var (y)6})^{1/2}}$$

NOTES

¹In applied work such variables sometimes are represented by proxies. The various representations of permanent income--by weighted averages of actual income, or by mean income or asset value for a group that is defined by occupation or by area of residence--provide examples. See Friedman's (1957) original study or, for a recent effort, Wolfe (1979).

An alternative approach is to posit that such variables can be represented by a latent variable structure. Recent estimates of earnings functions in which latent variables are used to control for unobserved abilities and motivation are illustrations of this approach. See Behrman, Hrubec, Taubman, and Wales (1980), Behrman and Wolfe (1979a,b,c), Chamberlain (1978), and Chamberlain and Griliches (1977).

²In recent years the modeling of such selectivity has been considerably developed, in large part, originally, in regard to women's labor force participation following the suggestions of Gronau (1974) and Lewis (1974). Methodology has been developed for maximum likelihood estimation (and approximations thereof) of related models with limited dependent variables. This literature is still expanding rapidly, as recent issues of Econometrica make clear. Among the most active contributors are Amemiya, Heckman, Lee and Maddala. For references and recent surveys, see Maddala (1978) and Wales and Woodland (1979). As part of the larger project of which this paper is a part, we provide some extensions and applications for women in a developing country in Behrman and Wolfe (1979a, 1979c, 1980a, 1980b) and Behrman, Wolfe, and Tunali (1979).

³Afifi and Elashoff (1966, 1967, 1969a, 1969b) and Maddala (1977) provide recent surveys of proposed models. Dempster, Laird and Rubin (1977) give an integrated technical account of the current state of the art in this area.

⁴To obtain Figure 1 the n observations have been reordered so that all of those that are complete only x and z are first, all of those that are complete only for x are next, all of those that are complete for x , y , and z are next, etc.

⁵In the appendix we give formulae for all of the alternatives that we consider.

⁶Given that the missing observations are random, the probability that $m = p = q$ is small.

⁷The auxiliary regressions need not be linear. For example, Gleason and Staelin (1975) propose a procedure in which the variables are partitioned into sets of missing and available variables, and then the missing variables are regressed on the principal components of the available variables in order to obtain estimates. This approach generally is more difficult to implement and more expensive than the use of linear auxiliary regressions.

⁸In regard to the question of using the dependent variable (say y_b) to fill in missing values of an independent variable (say x_{1b}), we find it useful to quote Griliches, Hall, and Hausman (1978, p. 172):

"We have not discussed explicitly using y_b to estimate the missing x_{1b} . The full-information maximum-likelihood procedure would do so implicitly. To econometricians using y to estimate missing x values looks suspiciously like an invitation to simultaneity bias. But a complete maximum likelihood procedure which assumes that both y and all the x 's are multi-variate normal, would use all the information in the sample (see, e.g., the description of the E-M algorithm in Dempster et al. [1977], but it would not use the constructed

\hat{x}_{1b} directly in a regression of y on x_1 and x_2 . Rather it would use such an \hat{x}_{1b} to fill in the covariances in the $X'X$ matrix, where the fact that \hat{x}_{1b} may depend on e (the disturbance in y) does not matter and rely on a more elaborate procedure for getting an estimate of its variance, where it does matter."

⁹As long as m equals r , adding z to the auxiliary regression unequivocally does not reduce (and may increase) efficiency and this method is at least as efficient as are any of the others. Still more efficient estimates might be obtained, however, if any other observations on the variables of interest are incorporated (e.g., the association between the first observations on x and z in Figure 1).

¹⁰It is possible, but not very likely, that all of the overlap among the three (or more) relevant variables is for the same m complete observations. In this case MM2 is guaranteed to give a positive definite variance-covariance matrix, since only these m observations are used to construct the matrix.

¹¹But Dagenais (1973) also derives a consistent generalized least square estimator based on the generalization of the auxiliary-regression proxy approach (PM2) to the multivariate regression case.

¹²If there is missing information at this stage too, then the process becomes more complicated or the partial observations are discarded with less efficiency. We ignore such a possibility in our discussion of costs.

¹³The exception is that MM5 does not have the smallest absolute difference for any of the 6 ranges of missing observations, although it does have the second smallest absolute difference for 2 ranges (i.e., 0 to 5 and 10 to 15 percent, see rows 1 and 3 in Table 3).

¹⁴For PM2-high correlation the smallest absolute value is for the 5 to 10 percent range and the second smallest is for the 0 to 5 percent range.

¹⁵Obviously this conclusion is dependent on the assumption that there are a reasonable number of complete observations. In our case it seems to hold if each of four variables have up to 30 percent of randomly missing observations out of 1200.

BIBLIOGRAPHY

- Afifi, A.A. and R.M. Elashoff, 1966. "Missing Observations in Multi-variate Statistics - I: Review of the Literature," Journal of the American Statistical Association 61:315 (September), pp. 595-605.
- . 1967. "Missing Observations in Multi-variate Statistics - II: Point Estimations in Simple Linear Regression," Journal of the American Statistical Association 62:317 (March), pp. 10-29.
- . 1969a. "Missing Observations in Multi-variate Statistics - III: Large Sample Analysis of Simple Linear Regression," Journal of the American Statistical Association 64:325 (March), pp. 357-58.
- . 1969b. "Missing Observations in Multi-variate Statistics - IV: A Note on Simple Linear Regression," Journal of the American Statistical Association 64:326 (March), pp. 359-65.
- Behrman, Jere R., Zdenek Hrubec, Paul Taubman and Terence J. Wales. 1979. Socioeconomic Success: A Study of the Effects of Genetic Endowments, Family Environment and Schooling, Amsterdam: North-Holland Publishing Company.
- Behrman, Jere R., and Barbara L. Wolfe. 1979a. "Important Early Life Cycle Socioeconomic Decisions for Women in a Developing Country: Years of Schooling, Age of First Cohabitation, and Early Labor Force Participation," Philadelphia: University of Pennsylvania, mimeo.
- . 1979b. "The Returns to Schooling in Terms of Adult Health, Occupational Status, and Earnings in a Developing Country: Omitted Variable Bias and Latent Variable-Variance Components Estimates." Philadelphia: University of Pennsylvania, mimeo.

- . 1979c. "Wage Rates for Adult Family Farm Workers in a Developing Country and Human Capital Investments in Health and Schooling." Philadelphia: University of Pennsylvania, mimeo.
- . 1980a. "Determinants of Health Utilization in a Developing Country." Philadelphia: University of Pennsylvania, mimeo.
- . 1980b. "The Demand for Household Nutrition in a Developing Country." Philadelphia: University of Pennsylvania, mimeo.
- Behrman, Jere R., Barbara L. Wolfe, and Insan Tunali. 1979. "Women's Earnings in a Developing Country: Double Selectivity, a Broader Definition of Human Capital to Include Health and Nutrition Status, Discrimination, Pluralism, and Family Status and Childcare," Institute for Research on Poverty Discussion Paper. 596-79.
- Chamberlain, Gary. 1978. "Omitted Variable Bias in Panel Data: Estimating the Returns to Schooling," Annals de L'insée 30-31 (April-September), pp. 49-82.
- , and Zvi Griliches. 1977. "More on Brothers," in Paul Taubman (ed.), Kinometrics: The Determinants of Socioeconomic Success Within and Between Families, Amsterdam: North-Holland Publishing Company.
- Dagenais, Marcel G. 1973. "The Use of Incomplete Observations in Multiple Regression Analysis: A Generalized Least Squares Approach," Journal of Econometrics 1:4 (December), pp. 317-328.
- Dempster, A.P., N.M. Laird, and D.P. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the E-M Algorithm," Journal of the Royal Statistical Society, Series B, 39:1, pp. 1-38.

- Frane, James W. 1976. "Some Simple Procedures for Handling Missing Data in Multivariate Analysis," Psychometrika 41:3 (September), pp. 409-415.
- Friedman, Milton. 1957. A Theory of the Consumption Function, Princeton: Princeton University Press.
- Glasser, M. 1964. "Linear Regression Analysis with Missing Observations Among the Independent Variables", Journal of the American Statistical Association 59:307 (September), pp. 834-844.
- Gleason, T.C. and R.A. Staelin. 1975. "A Proposal for Handling Missing Data," Psychometrika 40, pp. 229-252.
- Goldberger, Arthur S. 1973. "Structural Equation Models: An Overview," in Arthur S. Goldberger and Otis Dudley Duncan (eds.), Structural Equation Models in the Social Sciences, New York: Seminary Press.
- Griliches, Zvi, Bronwyn H. Hall and Jerry A. Hausman. 1978. "Missing Data and Self-Selection in Large Panels," Annals de L'insee 30-31 (April-September), pp. 137-176.
- Gronau, R. 1974. "Wage Comparisons: A Selectivity Bias," Journal of Political Economy, 82:5, pp. 1119-1143.
- Haitovsky, Yule. 1968. "Missing Data in Regression Analysis," Journal of the Royal Statistical Society, Series B, 30, pp. 67-81.
- Hester, Donald, D. 1976. "Estimation from Incomplete Samples," Madison: University of Wisconsin, mimeo.
- Lewis, H.G. 1974. "Comments on Selectivity Biases in Wage Comparisons," Journal of Political Economy, 82:5, pp. 1145-1156.
- Maddala, G.S. 1977. Econometrics, New York: McGraw-Hill Book Company.
- . 1978. "Selectivity Problems in Longitudinal Data," Annals de L'insée 30-31 (April-September), pp. 423-450.

Wales, Terence J. and A.D. Woodland. 1980. "Sample Selectivity and the Estimation of Labour Supply Functions," International Economic Review, forthcoming.

Wolfe, Barbara L. 1979. "Income Measures in Empirical Research: Results with Family Size and Value of Home," Madison: University of Wisconsin, mimeo.

Zellner, Arnold. 1970. "Estimation of Regression Relations Containing Unobservable Independent Variables," International Economic Review, (October).