Missing Data in Regression Analysis
Author(s): Yoel Haitovsky
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 30, No. 1 (1968), pp. 67-82
Published by: Blackwell Publishing for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2984459
Accessed: 01/11/2010 12:35

Missing Data in Regression Analysis

By Yoel Haitovsky

Technion—Israel Institute of Technology

[Received October 1966. Revised January 1967]

Summary

Two alternative methods for dealing with the problem of missing observations in regression analysis are investigated. One is to discard all incomplete observations and to apply the ordinary least-squares technique only to the complete observations. The alternative is to compute the covariances between all pairs of variables, each time using only the observations having values of both variables, and to apply these covariances in constructing the system of normal equations. The former is shown to be equivalent to the Fisher–Yates method of assigning "neutral" values to missing entries in experimental design.

The investigation is carried out by means of simulation. Eight sets of regression data were generated, differing from each other with respect to important factors. Various deletion patterns are applied to these regression data. The estimates resulting from applying the two alternative methods to the data with missing entries are compared with the known regression equations. In almost all the cases which were investigated the former method (ordinary least squares applied only to the complete observations) is judged superior. However, when the proportion of incomplete observations is high or when the pattern of the missing entries is highly non-random, it seems plausible that one of the many methods of assigning values to the missing entries should be applied.

1. Introduction

The problem of non-response to one or more questions in budgetary studies may be very troublesome when the data are to be used in regression analysis. Similar problems have been encountered in other branches of statistics where destruction of some parts of the experimental unit, or its removal from the experiment in a later stage, prevents the investigator from taking measurements.

Several ways—both pragmatic and statistically sophisticated—have been devised to cope with the missing data problem depending on their nature and proportion. Three general cases are distinguishable:

(*a*) randomly missing observations;

(*b*) missing categories, meaning that no answer is available owing to the fact that the question refers to some non-existing category in the responding unit;

(*c*) non-randomly missing observations—this is a case in which the researcher has evidence to believe that neither (*a*) nor (*b*) is true and that some special reason exists for non-response to a particular question.

A missing reply in a budgetary study to the question "What is your annual income?" provides an example for the latter. The researcher is justified in assuming that the responding unit has an annual income—therefore it does not fall into our case (*b*)—and that the missing datum is probably the result of an intentional refusal to respond (case *c*), rather than carelessness on the part of the respondent (case *a*).

It is obvious that the solution will differ with the case on hand. The solution to case (*a*), which has occupied most, if not all, of the published papers concerning the problem, is dealt with in this paper.

A solution to case (*b*) was offered by Taylor (1964). Regarding case (*c*), Albert E. Beaton, in conversation, has suggested assigning dummy variables for the missing values, and adding interactions between the dummy and the explanatory variables, to account for different slopes for the different groups of non-random missing observations or categories. (For the use of dummy variables and interactions between the dummy and the explanatory variables in regression analysis, see Suits, 1957.)

2. THE PROBLEM AND TWO POSSIBLE SOLUTIONS

Suppose the experimenter wishes to estimate the linear regression equation:

$$\hat{y}_j = \sum_{i=0}^{p} \beta_i x_{ij} \quad (j = 1, \ldots, n), \tag{1}$$

where $x_0 \equiv 1$.

The ordinary least-squares solution (OLS) requires that all measurements on each observation be included in the computations of the covariance matrices, namely, that for each j all x's and y's will be measured.

In the event that some (but not all) observations are incomplete in the sense that not all measurements on them exist, one can discard all incomplete observations and apply OLS to estimate the β's in (1) on the basis of the remaining observations. We shall refer to this procedure as Method 1.

An alternative, referred to as Method 2, is to compute the normal equations

$$\mathbf{cov}(x_i x_j)\hat{\boldsymbol{\beta}} = \mathbf{cov}(x_i y) \quad (i,j = 1, \ldots, p), \tag{2}$$

where $\mathbf{cov}(x_i x_j)$ is the $p \times p$ covariance matrix in which the (i,j)th element $(i,j = 1, \ldots, p)$ is computed from the measurements common to both x_i and x_j $(i \neq j)$ as well as from all the existing measurements on x_i for $i = j$, and similarly for $\mathbf{cov}(x_i y)$ $(i = 1, \ldots, p)$. $\hat{\boldsymbol{\beta}}$ is the $p \times 1$ vector of estimators. The $_j$ subscript from equation (1) is dropped in the process of summation in computing the **cov** matrices; lower case letters denote variables.

Method 2 appears to be consistent with the maximum-likelihood solutions applied to specific patterns of missing observations by Wilks (1932), Matthai (1951), Rao (1952, pp. 161–165; 1956), Lord (1955), Edgett (1956), Anderson (1957) and Nicholson (1957). Glasser (1964) compared the two procedures mentioned above for $p = 2$ and concluded that the smaller the *geometric* value of the correlation between the two independent variables and the smaller the proportion of observed values on x_j (with few exceptions when $|r|$ is large), the more efficient will be the estimates of β_i $(i,j = 1, 2)$. A number of his conclusions will be challenged later.

It can be shown (see Appendix) that the procedure of discarding all incomplete observations is analogous to the classical procedure in experimental design of inserting "neutral" values in place of the missing ones.

The covariance matrix of the estimated partial regression coefficients in OLS is known to be:

$$\mathrm{V}(\hat{\boldsymbol{\beta}}) = (\sigma^2/n)\{\mathbf{cov}(x_i x_j)\}^{-1} \quad (i,j = 1, \ldots, p), \tag{3}$$

where n is the number of observations minus unity, and σ^2 is the error variance, estimated by the following formula

$$\hat{\sigma}^2 = n\,[\mathrm{var}(y) - \{\mathbf{cov}(x_i y)\}'\{\mathbf{cov}(x_i x_j)\}^{-1}\mathbf{cov}(x_i y)]/(n-p) \tag{4}$$

(where a prime is understood to denote transposition in any matrix operation). These are the estimates applicable in Method 1 with n the number of complete observations minus unity.

For Method 2 it is easy to show that

$$\mathbf{cov}(x_i x_j)\, E\{\hat{\boldsymbol\beta}\} = E\{(n_{ij}/n_{iy})\,\mathbf{cov}(x_i x_j)\,\boldsymbol\beta + \mathbf{cov}(x_i \epsilon)\},$$

hence

$$E\{\hat{\boldsymbol\beta}\} = \{\mathbf{cov}(x_i x_j)\}^{-1}\{(n_{ij}/n_{iy})\,\mathbf{cov}(x_i x_j)\}\,\boldsymbol\beta \quad (\mathrm{i}, j = 1, \ldots, p), \tag{5}$$

where n_{ij} and n_{iy} are the number of observations common to both x_i and x_j, and x_i and y, respectively, minus unity.

It is immediately seen that the vector $\hat{\boldsymbol\beta}$ is an unbiased estimator of $\boldsymbol\beta$ only when all n_{ij}'s and n_{iy}'s are equal.

$$\mathbf{V}(\hat{\boldsymbol\beta}) = [\{\mathbf{cov}(x_i x_j)\}^{-1}\{(n_{ij}/n_{iy})\,\mathbf{cov}(x_i x_j)\} - \mathbf{I}]\,\boldsymbol\beta\boldsymbol\beta'[\{(n_{ji}/n_{jy})\,\mathbf{cov}(x_i x_j)\}\{\mathbf{cov}(x_i x_j)\}^{-1} - \mathbf{I}]$$
$$+\sigma^2\{\mathbf{cov}(x_i x_j)\}^{-1}\{(n_{ij}/n_{iy}\, n_{jy})\,\mathbf{cov}(x_i x_j)\}\{\mathbf{cov}(x_i x_j)\}^{-1} \quad (i, j = 1, \ldots, p). \tag{6}$$

Here, once more, $\mathbf{V}(\hat{\boldsymbol\beta})$ reduces to the OLS form only if $n_{ij} = n_{iy} = n_{jy}$ for all $i, j = 1, \ldots, n$.

For computing the intercept we have

$$\hat\beta_0 = \bar{y} - \sum_{i=1}^{p} x_i \hat\beta_i. \tag{7}$$

Two alternative methods for computing (7) are available. In the event of most observations on x_i existing, we can compute $\bar{X}_i$ directly; however, if there is a substantial number of missing observations in x_i we can apply the maximum-likelihood estimators proposed by Lord (1955).

The formula for computing the variance of residuals is

$$\hat\sigma^2 = \frac{\mathrm{SSE}}{n_y - p} = \{\mathrm{var}(y) - \sum_i \hat\beta_i \,\mathrm{cov}(y x_i)\}\, n_y/(n_y - p), \tag{8}$$

where $\mathrm{var}(y)$ is computed from all $(n_y + 1)$ existing observations on y. Finally,

$$\bar{R}^2 = 1 - \hat\sigma^2/\mathrm{var}(y) \tag{9}$$

is the so-called "corrected" multiple correlation coefficient R^2.

3. A Monte Carlo Experiment

Since Method 2 does not have optimal statistical properties, and since the derivation of its distribution theory is intractrable, the problem has been tackled by the Monte Carlo technique.

An IBM 7094 computer program has been written for constructing regression data with correlated independent variables generated from either normal or uniform random generator subroutines. The regression error term is normally distributed. An option to pre-specify all means, variances, and correlations of the independent variables in the regression is available in this program. The dependent variable is constructed as a linear combination of the independent variables and the error term, with optionally pre-specified weights. The OLS estimates are also computed. A deletion subroutine then "creates" the missing observations. (A more realistic approach would have been to regard the generated data as the population from which samples are drawn. However, it was felt that this unduly complicates both the program and the analysis.) Two options are available. Firstly, the number of elements to be deleted in each series can be pre-specified, and the elements will then be randomly deleted; the actual number of deleted elements might fall short of the specified number because of "ties" and "zeros". Secondly, the user can pre-specify the actual elements to be deleted. The former method will be referred to as random deletion, whilst the latter will be referred to as systematic deletion. Another available option is designed to investigate the problem of misspecification; it is carried out by omitting any number of independent variables in the estimation stage.

On the basis of the created incomplete observations, equations (2) and (6) to (9) are computed. In the original work four versions of Method 2 were proposed, the major differences being: (*a*) Lord's correction for the mean is applied to the computation of the covariance matrix and intercept, and (*b*) $\text{cov}(x_i x_j)$ $(i = j)$ is computed from the observations of x_i which have their counterparts in the y vector. However, the differences between all four versions were found to be insignificant, and the simplest of the four was selected for the present paper. For a detailed description of the four versions and the results, see Haitovsky (1966). Finally, all incomplete observations are discarded and OLS estimates are computed from the remaining data. This last procedure is our Method 1.

Eight sets of regression data were generated, differing in the number of independent variables in the regression, the distribution of independent variables, the correlation between them, the relationship between the highly correlated independent variables and their relative weights in the regression, and the amount of variability in the various independent variables as compared to that of the error term. The sensitivity of the two methods for estimating regression from incomplete data to all these factors can thus be investigated.

Different patterns of artificially created missing observations were applied to the eight sets of data. Here again, the various patterns of deletion were designed to investigate the sensitivity of the estimation methods to factors such as the proportion of missing observations in each variable and in the whole "sample", the proportion of complete observations in the "sample" and the effect of high (or low) proportion of deletion on important (or non-important) variables, as measured by their relative weights, and on highly (or insignificantly) correlated variables. The patterns of deletion applied to the last two sets were designed to represent a "typical" missing observation case. For that purpose we conducted a poll amongst consultants at the Harvard Computing Centre, whose results are reflected in these two cases. Finally, a systematic pattern of deletion was applied to a subset of the next to last set of data. This was chosen to reflect the case where, for example, high (or low) income units are more reluctant to report their income than the remainder of the participants in the survey. Therefore the data were sorted on the leading independent variable and 40 per cent of the upper 25 per cent values recorded on that variable were deleted, while only 20 per cent of the remaining 75 per cent of the elements were deleted.

In most cases, three deletion patterns were applied to every set of data and 10 computer runs per deletion pattern were performed. In four out of the nine experiments both correct specification and misspecification models were tried. All regressions which exhibited anomalies like zero- and higher-order correlation coefficients outside their feasible limits, negative variances, and negative determinants of the proper dispersion matrices were discarded. The means and standard deviations of the estimated partial regression coefficients and their respective estimated standard errors were computed and are summarized in Tables 1–8 for the eight sets of regression data, and in Table 9 for the "systematic deletion" experiment. The computed means and standard deviations can thus be compared with the true parameters.

TABLE 1

Summary of Monte Carlo Experiment

(True model: $\hat{Y} = 10{\cdot}0+1{\cdot}5X_1+0{\cdot}5X_2$. 1,000 observations.)

	Y	X_1	X_2	*Number of runs*
Correlation matrix	1.0	0·9817	0·9722	
		1·0	0·9697	
			1·0	
OLS	10·5005	1·5220	0·4863	($R^2 = 0{\cdot}971$)
		(0·0518)	(0·0319)	
Deletion pattern	30	75	50	10
Method 1	10·4376	1·5179	0·4887	10
	(0·9117)	(0·0308)	(0·0190)	
Method 2	14·2714	1.6713	0·3913	10
	(15·0362)	(0·5276)	(0·3382)	
Deletion pattern	50	90	135	10
Method 1	10·6771	1·5063	0·4930	10
	(0·7532)	(0·0313)	(0·0179)	
Method 2	18·2974	1·7983	0·3109	10
	(16·7671)	(0·6506)	(0·4136)	
Deletion pattern	50	30	75	10
Method 1	10·5134	1·5212	0·4862	10
	(1·1254)	(0·0260)	(0·0169)	
Method 2	10·1451	1·5381	0·4819	10
	(8·5397)	(0·3281)	(0·2056)	
Deletion pattern	1	100	100	10
Method 1	10·1037	1·5294	0·4839	10
	(0·5526)	(0·0157)	(0·0086)	
Method 2	21·2702	1·9250	0·2286	9
	(12·9767)	(0·5172)	(0·3291)	
Deletion pattern	1	200	200	10
Method 1	10·7599	1·5232	0·4847	10
	(1·5000)	(0·0444)	(0·0276)	
Method 2	1·8596	1·1013	0·7443	10
	(13·0240)	(0·6752)	(0·4044)	
Deletion pattern	100	400	400	9
Method 1	10·5131	1·5067	0·4945	9
	(1·3334)	(0·0634)	(0·0350)	
Method 2	8·8926	1·2079	0·6594	6
	(17·4673)	(0·6512)	(0·4060)	

Notes to Tables 1–9

Deletion pattern refers to the number of observations deleted in each series. The discrepancies between "number of runs" in the "deletion pattern" line and that in the other lines reflect the discarding of regression equations which failed to meet basic statistical requirements.

The estimated values in the Y-column are the intercept terms.

The figures in brackets are the estimated standard errors of the coefficients above them in OLS and the computed standard deviations of the coefficients elsewhere.

Independent variables are normally distributed in Tables 1, 2, 4, 7, and 9, and uniformly distributed in Tables 3, 5, 6, and 8.

TABLE 2

Summary of Monte Carlo Experiment

(True model: $\hat{Y} = 10{\cdot}0 + 1{\cdot}5X_1 + 0{\cdot}5X_2$. 1,000 observations.)

	Y	X_1	X_2	*Number of runs*
Correlation matrix	1·0	0·7291	0·4783	
		1·0	0·0083	
			1·0	
OLS	12·3125	1·5085	0·4834	(R^2 = 0·755)
	(4·0516)	(0·0326)	(0·0161)	
Deletion pattern	0	100	100	10
Method 1	12·0440	1·5155	0·4812	10
	(2·3716)	(0·0199)	(0·0084)	
MSE	10·2666	0·00066	0·00046	
Method 2	12·6139	1·5099	0·4801	10
	(2·9636)	(0·0225)	(0·0087)	
MSE	16·3745	0·00062	0·00052	
Deletion pattern	0	200	200	10
Method 1	10·6729	1·5141	0·4903	10
	(2·6223)	(0·0181)	(0·0094)	
MSE	7·3796	0·000549	0·000193	
Method 2	10·6369	1·5145	0·4904	10
	(3·3213)	(0·0248)	(0·0118)	
MSE	11·4677	0·000849	0·000242	
Deletion pattern	0	400	400	10
Method 1	12·1707	1·5001	0·4907	10
	(3·5183)	(0·0343)	(0·0156)	
MSE	17·6139	0·001165	0·00034	
Method 2	15·2847	1·4867	0·4785	10
	(4·7512)	(0·0311)	(0·0276)	
MSE	53·6048	0·001164	0·001275	

TABLE 3

Summary of Monte Carlo Experiment

(True model: $\hat{Y} = 10{\cdot}0 + 1{\cdot}5X_1 + 0{\cdot}5X_2$. 1,000 observations.)

	Y	X_1	X_2	*Number of runs*
Correlation matrix	1·0	0·8217	0·5596	
		1·0	0·0136	
			1·0	
OLS	12·4920	1·4895	0·4980	($R^2 = 0{\cdot}976$)
	(1·8980)	(0·0090)	(0·0045)	
Deletion pattern	0	100	100	10
Method 1	12·7134	1·4870	0·4987	10
	(0·9873)	(0·0046)	(0·0022)	
MSE	9·1553	0·000209	0·000007	
Method 2	13·1304	1·4851	0·4986	10
	(2·4680)	(0·0115)	(0·0056)	
MSE	16·9791	0·000379	0·000034	
Deletion pattern	0	200	200	10
Method 1	12·3978	1·4896	0·4982	10
	(1·1077)	(0·0068)	(0·0027)	
MSE	7·6152	0·000166	0·000008	
Method 2	11·8884	1·4939	0·4993	10
	(6·6687)	(0·0224)	(0·0146)	
MSE	48·4338	0·000543	0·000214	
Deletion pattern	0	400	400	10
Method 1	11·7376	1·4934	0·4990	10
	(1·5218)	(0·0082)	(0·0037)	
MSE	5·6706	0·000116	0·000015	
Method 2	11·5506	1·4809	0·5052	10
	(8·4217)	(0·0276)	(0·0187)	
MSE	73·6299	0·001172	0·000380	

TABLE 4

Summary of Monte Carlo Experiment

(True model: $\hat{Y} = 50{\cdot}0+1{\cdot}5X_1-3{\cdot}0X_2+5{\cdot}0X_3-0{\cdot}5X_4+0{\cdot}2X_5$. 500 observations.)

	Y	X_1	X_2	X_3	X_4	X_5	*Number of runs*
Correlation matrix	1·0	−0·4896	−0·6735	−0·6066	0·1583	0·0289	
		1·0	0·9152	0·8603	0·2394	0·1654	
			1·0	0·9450	0·2501	0·1791	
				1·0	0·2784	0·1972	
					1·0	0·6008	
						1·0	
OLS	25·8895	1·2929	−2·6789	4·6083	−0·5009	0·2131	(R^2 = 0·596)
	(40·6786)	(0·1153)	(0·1745)	(1·3079)	(0·1239)	(0·0333)	
Deletion pattern	1	50	50	50	50	50	10
Method 1	21·2784	1·2487	−2·6257	4·5505	−0·4972	0·2157	10
	(55·9812)	(0·2562)	(0·3004)	(1·7291)	(0·1484)	(0·0330)	
Method 2	0·6977	1·4856	−3·0253	6·1898	−0·5755	0·2196	9
	(116·2508)	(0·4087)	(0·7473)	(4·4161)	(0·1355)	(0·0521)	
Deletion pattern	1	100	100	100	100	100	10
Method 1	20·8727	1·2707	−2·6699	4·9535	−0·5290	0·2151	10
	(59·7559)	(0·1462)	(0·2772)	(2·1991)	(0·1918)	(0·0390)	
Method 2	18·6949	1·2311	−2·6270	4·5052	−0·3102	0·1851	8
	(132·5279)	(0·4551)	(0·5663)	(4·0922)	(0·2410)	(0·0618)	
Deletion pattern	30	50	50	100	100	100	10
Method 1	43·4851	1·2662	−2·5771	3·8733	−0·4963	0·2165	10
	(58·2549)	(0·1074)	(0·1745)	(1·8452)	(0·1474)	(0·0338)	
Method 2	−43·3055	1·6161	−3·3232	8·0376	−0·5896	0·2159	9
	(174·3591)	(0·4588)	(1·0899)	(6·9988)	(0·2097)	(0·0570)	
Deletion pattern	50	60	10	280	400	5	10
Method 1	−53·5387	0·0183	−1·4685	3·7497	−0·4476	0·2385	10
	(225·5116)	(0·1135)	(0·6900)	(6·8821)	(0·4086)	(0·1089)	
MSE	62766·8725	2·4522	3·0822	49·1002	0·1700	0·0135	
Method 2	−59·0650	1·6270	−3·3390	8·2347	−0·4840	0·2171	10
	(269·5687)	(0·9381)	(1·9248)	(11·2670)	(0·6212)	(0·1622)	
MSE	85884·0121	0·8980	3·8325	138·5710	0·3862	0·0266	
Misspecification Model							
OLS	36·1668	1·2825	−2·6681	4·7026	−0·0412	—	(R^2 = 0·563)
		(0·1198)	(0·1815)	(1·3600)	(0·1123)	—	
Deletion pattern	0	50	50	50	50	—	10
Method 1	39·7647	1·3061	−2·6790	4·6240	−0·0439	—	10
	(19·9731)	(0·0398)	(0·0734)	(0·6590)	(0·0501)	—	
Method 2	26·8669	1·4992	−2·9986	5·7379	−0·0159	—	10
	(152·7071)	(0·4294)	(0·8846)	(5·8500)	(0·1145)	—	
Deletion pattern	0	100	100	100	100	—	10
Method 1	40·7382	1·2562	−2·6716	4·7778	−0·1088	—	10
	(33·1093)	(0·1391)	(0·2297)	(1·1500)	(0·0782)	—	
Method 2	143·6100	1·0682	−2·0144	0·2899	0·0200	—	10
	(458·4927)	(1·2876)	(3·0416)	(18·6421)	(0·1567)	—	
Deletion pattern	0	50	50	50	100	—	10
Method 1	23·8771	1·2450	−2·6606	4·9172	0·0040	—	10
	(32·6449)	(0·0598)	(0·1505)	(1·2392)	(0·1011)	—	
Method 2	18·7662	1·4150	−2·8945	5·6078	0·0076	—	10
	(146·1931)	(0·3104)	(0·7967)	(5·5907)	(0·1147)	—	

TABLE 5

Summary of Monte Carlo Experiment

True model: $\hat{Y} = 50{\cdot}0+1{\cdot}5X_1-3{\cdot}0X_2+5{\cdot}0X_3-0{\cdot}5X_4+0{\cdot}2X_5$

Ordinary least squares: $\hat{Y} = 178{\cdot}17+2{\cdot}37X_1-3{\cdot}53X_2+2{\cdot}51X_3-0{\cdot}18X_4+0{\cdot}17X_5$

(0·389) (0·619) (5·02) (0·45) (0·12)

(R^2 = 0·161. 500 observations.)

	Y	X_1	X_2	X_3	X_4	X_5	*Number of runs*
			Misspecification Model				
Correlation matrix	1·0	−0·1783	−0·3035	−0·2846	−0·0707	0·0221	
		1·0	0·9188	0·8723	0·2573	0·1804	
			1·0	0·9525	0·2857	0·1692	
				1·0	0·2987	0·1639	
					1·0	0·5793	
						1·0	
OLS	210·6691	2·4160	−3·5373	2·1822	0·1521	—	(R^2 = 0·158)
	(148·9808)	(0·3907)	(0·6168)	(4·7123)	(0·3424)	—	
Deletion pattern	0	50	50	50	50	—	10
Method 1	236·6640	2·5025	−3·5069	1·3233	0·1813	—	10
	(59·0673)	(0·2941)	(0·3542)	(2·0793)	(0·1908)	—	
Method 2	147·1110	2·6329	−4·0358	5·0102	0·1477	—	10
	(350·4664)	(0·5329)	(1·5768)	(12·2994)	(0·2047)	—	
Deletion pattern	0	100	100	100	100	—	10
Method 1	269·7148	2·5678	−3·6114	1·3207	−0·0227	—	10
	(123·4965)	(0·1515)	(0·4770)	(4·1234)	(0·2795)	—	
Method 2	362·0251	2·4756	−3·0939	−2·3094	−0·0735	—	10
	(305·0858)	(1·0008)	(2·0063)	(12·1168)	(0·2497)	—	
Deletion pattern	0	50	50	50	100	—	10
Method 1	222·2397	2·4906	−3·4983	1·5536	0·2091	—	10
	(57·3730)	(0·2876)	(0·4150)	(1·7485)	(0·3022)	—	
Method 2	139·5214	2·4948	−3·8423	4·3920	0·2946	—	10
	(353·3617)	(0·5912)	(1·7308)	(12·7298)	(0·2315)	—	

TABLE 6

Summary of Monte Carlo Experiment

(True model: $\hat{Y} = 50{\cdot}0+1{\cdot}5X_1-3{\cdot}0X_2+5{\cdot}0X_3-0{\cdot}5X_4+0{\cdot}2X_5$. 500 observations.)

	Y	X_1	X_2	X_3	X_4	X_5	*Number of runs*
Correlation matrix	1·0	−0·5782	−0·8298	−0·7604	−0·1806	0·0732	
		1·0	0·9188	0·8730	0·2570	0·1801	
			1·0	0·9529	0·2851	0·1687	
				1·0	0·2977	0·1626	
					1·0	0·5791	
						1·0	
OLS	60·8824	1·5725	−3·0436	4·7868	−0·4461	0·1972	($R^2 = 0{\cdot}957$)
		(0·0326)	(0·0509)	(0·3946)	(0·0691)	(0·0099)	
Deletion pattern	50	50	50	50	50	30	
Method 1	57·3972	1·5912	−3·0694	4·9063	−0·4457	0·1974	10
	(7·1547)	(0·0205)	(0·0407)	(0·2750)	(0·0661)	(0·0126)	
Method 2	96·6995	1·4267	−2·7569	3·2719	−0·4281	0·1900	7
	(74·8982)	(0·1645)	(0·1666)	(1·8266)	(0·5633)	(0·0635)	
Deletion pattern	100	100	100	100	100	70	10
Method 1	62·2983	1·5618	−3·0323	4·8114	−0·4608	0·1936	10
	(27·2005)	(0·0698)	(0·0974)	(0·5903)	(0·0733)	(0·0165)	
Method 2	128·9724	1·5023	−2·7888	2·4545	−0·6257	0·2655	7
	(222·5091)	(0·3856)	(0·8017)	(6·7911)	(0·6049)	(0·0611)	
Deletion pattern	50	100	230	70	80	50	10
Method 1	43·6974	1·5442	−3·0526	5·1670	−0·4760	0·2036	10
	(15·5713)	(0·0429)	(0·0732)	(0·4342)	(0·1186)	(0·0207)	
Method 2	47·2974	1·4882	−3·0299	4·4583	−0·1295	0·2041	4
	(486·0464)	(0·5589)	(1·1688)	(12·5221)	(0·5448)	(0·0893)	
			Misspecification	*Model*			
OLS	118·3746	1·6288	−3·0525	4·3784	0·3349	—	($R^2 = 0{\cdot}922$)
	(21·8775)	(0·0437)	(0·6910)	(0·5299)	(0·0766)	—	
Deletion pattern	30	50	50	50	50	—	20
Method 1	121·6952	1·6187	−3·0328	4·2670	0·3314	—	20
	(15·3276)	(0·0381)	(0·0646)	(0·4674)	(0·0450)	—	
Method 2	197·4687	1·5049	−2·6554	1·5215	0·3916	—	18
	(133·2780)	(0·2159)	(0·5536)	(4·0238)	(0·2476)	—	
Deletion pattern	70	100	100	100	100	—	20
Method 1	110·6905	1·6246	−3·0678	4·5307	0·3580	—	20
	(41·3545)	(0·0661)	(0·1172)	(0·9546)	(0·1137)	—	
Method 2	175·0418	1·4755	−2·7340	2·6763	0·2119	—	14
	(135·4735)	(0·3422)	(0·6561)	(4·3549)	(0·3207)	—	
Deletion pattern	50	100	230	70	80	—	20
Method 1	125·0418	1·6226	−3·0279	4·1185	0·3706	—	20
	(37·3602)	(0·0872)	(0·0964)	(0·7249)	(0·1405)	—	
Method 2	171·3962	1·2813	−2·4591	1·3167	0·5537	—	14
	(318·7019)	(0·4235)	(0·9113)	(8·3915)	(0·5257)	—	

TABLE 7

Summary of Monte Carlo Experiment

(True model: $\hat{Y} = 150\cdot0+5\cdot0X_1-2\cdot0X_2+0\cdot3X_3+3\cdot0X_4$. 400 observations.)

	Y	X_1	X_2	X_3	X_4	*Number of runs*
Correlation matrix	1·0	0·7522	0·5958	0·6979	0·8232	
		1·0	0·8385	0·4596	0·3618	
			1·0	0·6077	0·4706	
				1·0	0·7962	
					1·0	
OLS	142·4500	4·9613	−1·9938	0·7690	2·9559	($R^2 = 0\cdot991$)
	(8·6144)	(0·0479)	(0·0371)	(0·3070)	(0·0331)	
Deletion pattern	30	50	260	0	80	10
Method 1	148·4747	4·9304	−1·9575	0·6322	2·9452	10
	(12·2894)	(0·0918)	(0·0644)	(0·4045)	(0·0394)	
Method 2	161·4807	4·7150	−1·8105	0·4517	2·9095	8
	(71·5068)	(1·2035)	(0·7198)	(2·9106)	(0·3787)	
			Misspecification Model			
OLS	44·6090	5·0597	−2·1100	20·7272	—	($R^2 = 0\cdot800$)
	(39·3901)	(0·2206)	(0·1711)	(0·9722)	—	
Deletion pattern	30	50	260	0	—	10
Method 1	35·0046	4·9686	−2·0214	20·7550	—	10
	(69·9230)	(0·3526)	(0·2859)	(1·8529)	—	
Method 2	17·7651	5·5651	−2·4758	21·7058	—	10
	(162·5780)	(2·0014)	(1·3855)	(4·5679)	—	

TABLE 8

Summary of Monte Carlo Experiment

True model: $\hat{Y} = 150{\cdot}0+5{\cdot}0X_1-2{\cdot}0X_2+0{\cdot}3X_3+3{\cdot}0X_4$

Ordinary least squares: $\hat{Y} = 151{\cdot}22+5{\cdot}00X_1-2{\cdot}00X_2+0{\cdot}36X_3+2{\cdot}99X_4$

(0·0129) (0·0096) (0·0868) (0·0973)

(R^2 = 0·998. 400 observations.)

	Y	X_1	X_2	X_3	X_4	*Number of runs*
Correlation matrix	1·0	0·8743	0·4570	0·3765	0·3705	
		1·0	0·8255	0·5181	0·4575	
			1·0	0·6080	0·5309	
				1·0	0·8261	
					1·0	
			Misspecification Model			
OLS	974·4923	5·0064	−1·9963	2·3547	—	(R^2 = 0·997)
	(6·8654)	(0·0238)	(0·0177)	(0·1061)	—	
Deletion pattern	30	50	260	0	—	10
Method 1	974·2385	5·0032	−2·0029	2·3972	—	10
	(14·0030)	(0·0395)	(0·0384)	(0·1822)	—	
Method 2	975·5982	4·4183	−1·5508	1·4878	—	6
	(116·2091)	(0·3137)	(0·2132)	(2·1054)	—	

TABLE 9

Summary of Monte Carlo Experiment: Systematic deletion

(True model: $\hat{Y} = 150{\cdot}0+5{\cdot}0X_1-2{\cdot}0X_2+0{\cdot}3X_3+3{\cdot}0X_4$. 100 observations.)

	Y	X_1	X_2	X_3	X_4	*Number of runs*
Correlation matrix	1·0	0·7852	0·6137	0·6389	0·8323	
		1·0	0·8738	0·5166	0·4267	
			1·0	0·6314	0·4650	
				1·0	0·7119	
					1·0	
OLS	150·7320	4·9683	−1·9217	0·5143	2·9224	(R^2 = 0·992)
	(17·7730)	(0·0960)	(0·0710)	(0·5530)	(0·0590)	
Deletion pattern	6	25	15	0	10	10
Method 1	147·3267	5·0059	1·9684	0·7748	2·9073	10
	(9·7994)	(0·0935)	(0·0422)	(0·3182)	(0·0214)	
Method 2	414·4428	4·1160	−0·6600	−6·5819	2·6987	7
	(128·5933)	(0·5613)	(0·5911)	(2·1446)	(0·3197)	
			Misspecification Model			
OLS	169·8529	5·6002	−2·2497	17·4946	—	(R^2 = 0·778)
	(92·4250)	(0·4970)	(0·3680)	(2·2640)	—	
Deletion pattern	10	25	15	0	—	10
Method 1	199·9281	5·6932	−2·3708	17·1867	—	10
	(64·8925)	(0·3204)	(0·1750)	(1·2481)	—	
Method 2	340·0518	4·8219	−1·1133	11·3555	—	10
	(151·9317)	(0·6615)	(0·3423)	(4·4350)	—	

A word about the goodness-of-fit criteria is now in order. Since Method 2 is not unbiased, we feel that the proper criterion is the mean square error (MSE). The MSE criterion, however, is not satisfactory when one estimation method is not uniformly better, with respect to all partial regression coefficients, than the others. (For a discussion of goodness-of-fit criteria with special application to Monte Carlo studies, see, for example, Summers, 1965, pp. 5, 11–14.) A way out of the difficulty in non-uniform cases is to apply the mean square error of prediction $E(Y-\hat{Y})^2$, where $\hat{Y}$ is the predicted Y. However, the latter is not applicable to the analysis of incomplete observations since $\hat{Y}$ is not always estimatable. Two alternatives are open in this case, (*a*) re-inserting the artificially omitted values, and (*b*) devoting a part of the "sample" solely to this purpose; none of these alternatives were provided by the computer program. Fortunately, there are very few non-uniform cases.

Inspection of Tables 1–9 will immediately establish the surprising superiority of Method 1, based on least-squares analysis of complete observations, to its alternative Method 2, based on estimating the covariance matrix from all available observations. In most cases Method 1 is so superior, with respect to both the unbiasedness and the efficiency criteria, that there is no need to report the MSE's. The MSE's are listed in Tables 2, 3 and part of Table 4, wherever the situation is otherwise ambiguous. The MSE criterion in these tables discloses that there is only one case in which Method 2 is significantly superior to Method 1; this occurs when the fourth deletion pattern is applied to the fourth set of data (Table 4), where Method 2 produces better results in estimating β_1. The explanation lies in the fact that for this case only 9–10 per cent of the observations are complete and hence available for use in Method 1. A possible explanation of why only $\hat{\beta}_1$ was affected, while all other estimated coefficients were not, is that neither X_1 nor a variable highly correlated with X_1 has a high proportion of missing observations. This explanation is not without reservations, since, according to it, Method 2 should estimate β_5 better than Method 1.

Another unfavourable feature of Method 2 is the presence of "nuisance parameters" in the estimation of the variances of the partial coefficients. The inclination in practical situations is to ignore the term involving the nuisance parameters. As the term ignored is always non-negative, the variance is usually underestimated. Our study indicates that a twenty-fold underestimation of the coefficients' variances is typical. This provides further evidence in favour of Method 1.

Further inspection of the tables leads to the following conclusions:

(*a*) Method 1 is relatively better for misspecified models than for correctly specified ones; this observation makes Method 1 even more attractive since in practice misspecified models are probably the rule rather than the exception.

(*b*) The trend discovered by Glasser (1964), that the relative (and, after a point, absolute) efficiency of Method 2 increases as the correlation between the independent variables reduces, is confirmed. However, even if the correlation is not significantly different from zero (Tables 2 and 3) Method 1 is still preferable, contrary to Glasser. Apparently, Glasser's assumption that the first term on the right-hand side of (6) vanishes for large samples is not justified even for samples of size 1,000.

It might be interesting to identify the major source of trouble in the rejected method. The MSE can be decomposed into two additive terms: one term accounts for the bias and the other one is the variance when the bias is ignored. Comparison of the two terms clearly shows that the latter is far more important. We therefore conclude that, although the bias affects the relevance of the inference, the real trouble is caused by inconsistency introduced into the system of normal equations given by (2).

4. THE ALTERNATIVES

Another avenue by which one can approach the problem is to assign dummy variates for those that are missing. We have shown in the Appendix that Method 1 is equivalent to the classical method of assigning "neutral" values in place of the missing ones. It is obvious, however, that when the subset of complete data is small relative to the whole sample, one loses valuable information by ignoring the incomplete observations. What are the alternatives?

Pragmatic procedures have been used by experimental statisticians. One is to assign the missing values at the mean. This procedure might bias the estimates badly. However, it can be argued that the gain in precision might over-compensate the bias. Kosobud (1963) proposes to use the existing pairs of two correlated series to establish the correlation between them, and to apply it to fill the missing records. He claims to have proved that the residual variance will be smaller in the case of this assignment than in the case where the incomplete questionnaires are dropped altogether. The proof is based on Kosobud's unjustified practice of granting the assigned values degrees of freedom as if they were genuine observations.

Obvious extensions of Kosobud's technique were proposed by Walsh (1959) and Buck (1960).

A more complicated method was proposed by Dear (1959), using the well-known property of principal component analysis that the original data are obtainable from the vector of factor scores and factor loadings. Dear estimates the latter by factoring the cross-products matrix based only on the complete part of the sample. Thereafter, he estimates the missing values by using the principal component transformation.

Finally, Afifi and Elashoff (1966), who investigated the simple regression case, used a weighted sum of all the existing observations on the dependent and independent variables respectively. By minimizing the SSE they obtained a simple regression coefficient which differs from that of the classical missing-observations method by the fact that the denominator contains an additional term which is the sum of squares of deviations from the mean of X which corresponds to the missing subset of Y. Afifi and Elashoff also carried out an intensive investigation on the four methods applicable to simple regressions, namely assigning values to the missing ones: (i) by the classical method, (ii) at the mean, (iii) by simple regression and (iv) by their own method.

They derived their distribution theory and evaluated their efficiencies using numerical examples. They concluded that no estimation technique was uniformly best (Afifi and Elashoff, 1966). Generally, they found that Method (ii) is best for very lowly correlated series, Method (iv) is best for low correlations, Method (i) is best for moderate correlations, and Method (iii) is best for highly correlated series.

It seems plausible that Afifi and Elashoff's conclusion that "no estimation technique is uniformly best" is true also for multiple regressions. For each particular case the distribution theory should be determined and compared for the various methods. However, a rule of thumb can be formulated for data with a high proportion of incomplete observations:

(*a*) One should use the classical method when the proportion of missing values is not large and when they are not too scattered among the multivariate observations; in addition, when none of the pairwise correlation coefficients is reasonably large.

(*b*) One should use the simple regression method when a variable with missing entries is highly correlated with one variable with no (or a small proportion of) missing values.

(*c*) One should use the weighted predicted values method when none of the variables stands out for its high correlation with variables with missing observations.

All other methods should be applied in very special designs. (In this context, see also Trawinski and Bargmann, 1964.)

ACKNOWLEDGEMENTS

This paper is based on an excerpt from Chapter 4 of the author's Ph.D. thesis submitted to the Department of Economics, Harvard University. The paper resulted from a problem presented to the author by Professor J. R. Meyer, to whom the author is grateful for his advice and assistance. The author also benefited from discussions with Professor Houthakker, Professor L. D. Taylor, Dr M. L. Ingbar and Dr A. E. Beaton. The computations were carried out at the Harvard Computing Centre.

REFERENCES

AFIFI, A. A. and ELASHOFF, R. M. (1966). Missing observations in multivariate statistics. I: Review of the literature. *J. Amer. Statist. Ass.*, **61**, 595–604.

ANDERSON, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *J. Amer. Statist. Ass.*, **52**, 200–203.

BUCK, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. R. Statist. Soc.* B, **22**, 302–306.

DEAR, R. E. (1959). A principal-component missing-data method for multiple regression models. (System Development Corporation, Report SP-86.)

EDGETT, G. L. (1956). Multiple regression with missing observations among the independent variables. *J. Amer. Statist. Ass.*, **51**, 122–131.

GLASSER, M. (1964). Linear regression analysis with missing observations among the independent variables. *J. Amer. Statist. Ass.*, **59**, 834–844.

HAITOVSKY, Y. (1966). Estimation of multivariate statistics from grouped and missing data. (Ph.D. dissertation, Harvard University.)

KOSOBUD, R. (1963). A note on a problem caused by assignment of missing data in sample surveys. *Econometrica*, **31**, 562–563.

LORD, F. M. (1955). Estimation of parameters from incomplete data. *J. Amer. Statist. Ass.*, **50**, 870–876.

MATTHAI, A. (1951). Estimation of parameters from incomplete data with application to design of sample surveys. *Sankhyā*, **11**, 145–152.

NICHOLSON, G. E., JR (1957). Estimation of parameters from incomplete multivariate samples. *J. Amer. Statist. Ass.*, **52**, 523–526.

RAO, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. New York: Wiley.

—— (1956). Analysis of dispersion with incomplete observations on one of the characters. *J. R. Statist. Soc.* B, **18**, 259–264.

SUITS, D. B. (1957). Use of dummy variables in regression equations. *J. Amer. Statist. Ass.*, **52**, 548–551.

SUMMERS, R. (1965). A capital intensive approach to the small sample properties of various simultaneous equation estimators. *Econometrica*, **33**, 1–41.

TAYLOR, L. D. (1964). A note on the problem of missing observations in cross section. (Mimeographed, Department of Economics, Harvard University.)

TRAWINSKI, I. M. and BARGMANN, R. E. (1964). Maximum likelihood estimation with incomplete multivariate data. *Ann. Math. Statist.*, **35**, 647–657.

WALSH, J. E. (1959). Computer-feasible general method for fitting and using regression functions when data incomplete. (System Development Corporation, Report SP-71.)

WILKS, S. S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *Ann. Math. Statist.*, **3**, 163–195.

APPENDIX

In order to show that Method 1 is equivalent to the classical method of assigning neutral values for the missing ones we minimize the SSE by differentiating it partially with respect to x_{ij} and y_j respectively and equating to zero. Using matrix notations we have

$$\begin{aligned}\partial\mathrm{SSE}/\partial x_{ij} &= \partial[\mathbf{Y}'\mathbf{Y}-\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}]/\partial x_{ij}\\ &= -2y_j\,\mathbf{e}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}+\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}[\mathbf{E}'_{ij}\mathbf{X}+\mathbf{X}'\mathbf{E}_{ij}](\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}=0\\ &\qquad (i=1,\ldots,p;\ j=i,\ldots,n) \qquad (10)\end{aligned}$$

and

$$\partial\mathrm{SSE}/\partial y_j = 2y_j-2\mathbf{x}_j(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}=0 \quad (j=1,\ldots,n), \qquad (11)$$

where x_{ij} and y_j indicate the elements of $\mathbf{X}$ and $\mathbf{Y}$ respectively, $\mathbf{x}_j$ is the jth row of $\mathbf{X}$, $\mathbf{e}_i$ is an $n\times 1$ unit vector with unity in the ith position, and $\mathbf{E}_{ij}$ is an $n\times p$ matrix of zeros with unity in the (i,j)th position.

A typical $\mathbf{E}'_{ij}\mathbf{X}+\mathbf{X}'\mathbf{E}_{ij}$ matrix is

$$\left[\begin{array}{ccc} & x_{1j} & \\ 0 & x_{2j} & 0\\ & \vdots & \\ x_{i1}\ldots & 2x_{ij} & \ldots x_{ip}\\ 0 & \vdots & 0\\ & x_{pj} & \end{array}\right],$$

where the non-zero elements occur at the ith row and jth column. Substituting

$$\hat{\boldsymbol\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

equation (10) reads

$$y_j\hat\beta_i = \sum_k x_{kj}\hat\beta_k\hat\beta_i = \hat\beta_i\sum_k x_{kj}\hat\beta_k,$$

where the subscripted $\hat\beta$'s are the elements of the $\hat{\boldsymbol\beta}$ vector. Hence

$$y_j = \sum_k x_{kj}\hat\beta_k \quad (j=1,\ldots,n). \qquad (12)$$

The proof is completed by observing that (12) implies that the assigned value for x_{ij} lies on the hyperplane defined by $\hat{\boldsymbol\beta}$.

Equation (12) is derived immediately from (11).

Similarly, one can show that all other statistics, particularly the standard errors of the partial regression coefficients and of the prediction, computed from the classical method of assigning values for missing ones are identical to those computed from the complete data subset.