AN INVESTIGATION OF SOME MULTIPLE REGRESSION METHODS FOR INCOMPLETE SAMPLES

by

C. F. Federspiel, R. J. Monroe, and B. G. Greenberg

This research was supported in part by U. S. Public Health Service Training Grant 2G-38 and by the Revolving Research Fund of the Institute of Statistics of the Consolidated University of North Carolina through a grant from the General Education Board for research in basic statistical methodology.

Institute of Statistics
Mimeo Series No. 236
August, 1959

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
LIST OF FIGURES	ix
CHAPTER I INTRODUCTION	1
CHAPTER II METHODS	5
2.1 Generation of Multivariate Samples	5
2.2 Censoring	7
2.3 Empirical Measures of Efficiency	8
2.4 Description of Experiments	10
2.5 The Computing Programs	14
CHAPTER III THE DELETION OF INCOMPLETE OBSERVATIONS	16
3.1 Discussion	16
3.2 Bias and Variance of Regression Coefficients	17
3.3 Efficiency of Prediction Equations	23
CHAPTER IV THE METHOD OF PAIRED CORRELATION COEFFICIENTS	27
4.1 Definition	27
4.2 Bias and Variance of Regression Coefficients	29
4.3 Error Variance	29
4.4 Efficiency of Prediction Equations	32
CHAPTER V THE GREENBERG PROCEDURE	49
5.1 Definition	49
5.2 Bias and Variance of Regression Coefficients	52
5.3 Error Variance	56
5.4 Efficiency of Prediction Equations	61
5.5 Convergence	71
CHAPTER VI MAXIMUM LIKELIHOOD	79
6.1 Discussion	79
6.2 Bias and Variance of Regression Coefficients	81
6.3 Error Variance	81
6.4 Efficiency of Prediction Equations	83
CHAPTER VII AN EXAMPLE	85
7.1 The Data	85
7.2 Solutions	87
7.3 Discussion	90

TABLE OF CONTENTS (Continued)

	Page
CHAPTER VIII SUMMARY AND CONCLUSIONS	94
8.1 Summary	94
8.2 Conclusions	103
LIST OF REFERENCES	106
APPENDIX A. GENERATION OF MULTIVARIATE NORMAL DEVIATES . . .	109
APPENDIX B. THE NORMAL DEVIATES USED IN SAMPLING	113

LIST OF TABLES

Page

1. Statistics of regression coefficients estimated by the deletion of incomplete observation vectors 18

2. Means and variances of errors of estimate obtained by the deletion of incomplete observations 22

3. Means of $\hat{V}_w(b_i)$ estimated by the deletion of incomplete observations 24

4A. Values of $\bar{\bar{E}}$ and their standard errors estimated by the deletion of incomplete observations 25

4B. Values of $\bar{\bar{\Delta}}$ and their standard errors estimated by the deletion of incomplete observations 25

5. Statistics of regression coefficients estimated by the Method of Paired Correlation Coefficients 30

6. Means and variances of errors of estimate obtained by the Method of Paired Correlation Coefficients 33

7A. Values of $\bar{\bar{E}}$ and their standard errors estimated by the Method of Paired Correlation Coefficients 34

7B. Values of $\bar{\bar{\Delta}}$ and their standard errors estimated by the Method of Paired Correlation Coefficients 34

8. Analysis of variance of ln E for the trivariate populations of Experiment 1 35

9. Values of $\bar{\bar{E}}$ and residuals from the response equation, $\hat{E}_{A+B}$, of Experiment 2 36

10. Analysis of variance of values of E 37

11. Values of $\bar{\bar{E}}$ and their standard errors at symmetrical points in the trivariate population space 38

12. Analysis of variance for the model $\hat{E}_{A+B}$ of Experiment 2 . . . 42

13. Values of $\bar{\bar{E}}$ estimated by the Method of Paired Correlation Coefficients 44

14. Statistics of regression coefficients estimated by the Greenberg Procedure; Experiment 1 53

15. Statistics of regression coefficients estimated by the Greenberg Procedure; Experiment 4 55

LIST OF TABLES (continued)

Page

16. Statistics of regression coefficients from Experiment 7 . . . 58

17A. Values of $\bar{\bar{E}}$ and their standard errors estimated by the Greenberg Procedure; Experiment 1 63

17B. Values of $\bar{\bar{\Delta}}$ and their standard errors estimated by the Greenberg Procedure, Experiment 1 63

18. Values of $\bar{\bar{E}}$ and their standard errors estimated by the Greenberg Procedure; Experiment 4 64

19. Analyses of variance for values of E from the data of Experiment 4 65

20. Analyses of variance for models fitted in Experiment 4 . . 66

21. Values of E and Δ from the twenty samples of Experiment 8 . 70

22. Mean number of iterations required for convergence . . . 72

23. Analysis of variance for number of iterations required for convergence in Experiment 4B 73

24. Iterations required for convergence using the Greenberg procedure on Carver's data with different initial estimates of missing values 74

25. Statistics of regression coefficients estimated by maximum likelihood 82

26. Means and variances of $\hat{\sigma}^2_{y.x}$ estimated by maximum likelihood 83

27A. Means of E and their standard errors estimated by maximum likelihood 84

27B. Means of Δ and their standard errors estimated by maximum likelihood 84

28. Biochemical determinations from Welt data 86

29. Sample correlation coefficients and their confidence limits based on the n_{ij} available pairs of x_i and x_j in Table 28 89

30. Some extreme muscle potassium values predicted from the Welt data by various methods 92

31A. Values of $\bar{\bar{E}}$ from Experiment 1 95

LIST OF TABLES (continued)

Page

31B. Values of $\bar{\Delta}$ from Experiment 1 96

32A. Values of $\bar{\bar{E}}$ from Experiment 4B 98

32B. Values of $\bar{\Delta}$ from Experiment 4B 98

33A. Values of $\bar{\bar{E}}$ and their standard errors, Experiment 6 99

33B. Values of $\bar{\Delta}$ and their standard errors, Experiment 6 100

34. Values of $\bar{\bar{E}}$, $\bar{\Delta}$ and their standard errors from Experiment 7 . 101

35A. Values of $\bar{\bar{E}}$ and their standard errors, Experiment 8 102

35B. Values of $\bar{\Delta}$ and their standard errors, Experiment 8 102

APPENDIX Table B-1. Frequency tabulation of ten thousand N(0,1) deviates 113

LIST OF FIGURES

Page

1. Histograms of regression coefficients estimated by deletion of incomplete observations 20
2. Histograms of regression coefficients estimated from complete observations 21
3. Histograms of regression coefficients estimated by the Paired Correlation Method 31
4. Contours of $\hat{E}_A$ in trivariate population space 40
5. Contours of $\hat{E}_B$ in trivariate population space 41
6. Associated values of $\bar{\bar{E}}$ and avg $(\hat{E}_{A+B})$ 45
7. Histograms of E; Experiment 6 47
8. Histograms of regression coefficients estimated by the Greenberg Procedure 57

9A. Mean pseudo-variance of Greenberg Procedure plotted against true variance; from data of Experiment 4B 60

9B. Mean variance estimated from uncensored data plotted against true variance; from data of Experiment 4B 60

10. Histograms of estimated error variances 62
11. Histograms of E; Experiment 7 69
12. Histograms of Δ; Experiment 7 68
13. Histogram of number of iterations required for convergence . 75
14. Convergence curves applying the Greenberg Procedure to the five-variate sample, number 107 77
15. Convergence curves applying the Greenberg Procedure to the five-variate sample, number 106 78

CHAPTER I

INTRODUCTION

A common problem in statistics is that of obtaining a multiple regression equation from data which have missing values among the independent variates. Examples of this situation occur in attempting to get prediction equations in such diverse fields as psychology, economics, and biochemistry. In psychology, some subjects may have failed to take all the tests in a series. In economics, production records of a given commodity may not have been kept consistently over a long period of time. In biochemistry, the analysis of some component of a serological sample may have been vitiated.

Several investigators have been concerned with the general problem of missing values in multivariate samples, They have, for the most part, dealt with maximum likelihood estimators for specific cases and were not directly concerned with prediction equations.

Wilks (1932) considered cases where values of x and y are both missing in samples from a bivariate normal and gave maximum likelihood estimates for some of the parameters, given others. He also evaluated the large sample efficiency of some simpler alternative estimates. In considering an extension of Wilks' work to the trivariate case, Matthai (1951) derived maximum likelihood estimators for the three means given the other six parameters. He suggested a modified form of estimation for the case in which it is necessary to estimate the variances and covariances from the data. In addition, he discussed applications of this work to the design of sample surveys.

Lord (1955) solved maximum likelihood equations to estimate the parameters for the special case of three variables, x_1, x_2, and x_3, in which x_1 and x_3 are distributed bivariate normally, as are x_2 and x_3. With no loss of generality, the data could be arranged:

$$\begin{array}{llllll} x_{11} & x_{12} & \cdots & x_{1n} & x_{1(n+1)} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2n} & & & \\ & & & & x_{3(n+1)} & \cdots & x_{3N} \end{array}$$

where the primary subscript denotes the variable, and the secondary, the individual. He also showed that, although the estimates of some of the parameters are improved by using all the data, estimates of the regression coefficients and errors of estimate, $\beta_2 = \frac{\sigma_2\rho_{21}}{\sigma_1}$, β_1, $\sigma^2_{2.1}$, and $\sigma^2_{3.1}$ are not. In considering this special case, Lord pointed out that "the general maximum likelihood equations have proved rather intractable and no simple formulae for the maximum likelihood estimators are available in the general case, even for a sample from a bivariate population."

Edgett (1956) dealt more directly with the prediction problem in considering the special case of the trivariate normal in which observations are missing from one variate only. Data would have the form:

$$\begin{array}{lllllll} x_{11}, & x_{21}, & \cdots & x_{1n} & x_{1(n+1)} & \cdots & x_{1N} \\ x_{21}, & x_{22}, & \cdots & x_{2n} & & & \\ x_{31}, & x_{32}, & \cdots & x_{3n} & x_{3(n+1)} & \cdots & x_{3N} \end{array}$$

After some rather tedious algebra, he obtained solutions in closed form for the maximum likelihood equations to estimate the parameters μ_1, μ_2,

μ_3, σ_1^2, σ_2^2, σ_3^2, σ_{12}, σ_{13}, σ_{23}. From these, coefficients for the regression of x_1 on x_2 and x_3 were estimated, considering the latter two as independent variates.

The problems of Edgett and Lord were also considered by T. W. Anderson (1957) who indicated an approach which somewhat simplifies the mathematical manipulation necessary to get the maximum likelihood solutions. The procedure involves expressing the joint densities of the variates as the product of a marginal density and the remaining conditional densities. Further reference will be made to the papers of Edgett and Anderson in the chapter on maximum likelihood.

In a paper dealing particularly with the prediction problem, Nicholson (1957) discussed the multivariate case for which observations are missing in the dependent variable. He gave maximum likelihood estimates for the parameters based on the "incomplete sample" and for the complete portion of the sample, and concluded that, for prediction purposes, no use can be made of data which are incomplete in the dependent variable.

This thesis will deal primarily with methods for determining prediction equations when observations among the independent variates are not available. These methods include: ordinary least squares, using only those observation vectors which are complete in every variate; maximum likelihood for certain special cases; and two other procedures which have been proposed on the basis of their intuitive appeal and general applicability.

One of these methods involves finding the matrix of paired correlation coefficients for all variables, then solving for the standardized regression coefficients, and finally, converting these to the ordinary regression coefficients by making use of the means and variances calculated from all observations present for a given variable. This will be referred to as the Method of Paired Correlation Coefficients, or simply, the Paired Correlation Method.

The other approach consists of an iterative computing scheme whereby missing values of a given variable are successively predicted from the other variables until the coefficients for the regression of y on the x_i converge simultaneously. This is called the Greenberg Procedure.

The objective of this study is to appraise these methods. An empirical sampling approach, or what is sometimes referred to as the Monte Carlo method, is used in this investigation. This approach has been found useful in problems which seem to defy strictly mathematical treatment, and has been particularly successful in studies of estimation procedures for a wide variety of models (Aitchison and Brown, 1957; Morrison, 1956; Neiswanger and Yancey, 1959; Orcutt and Cochrane, 1949). Its use here involves application of the estimation procedures described above to simulated fragmentary multivariate normal samples. Measures of prediction efficiency are calculated for each sample estimate and are appropriately summarized. The calculations of estimates and their efficiencies were carried out on an IBM 650 electronic computer.

In the following chapter, notation, definitions of efficiency, and the general method of approaching the problem are discussed. In Chapters III to VI, respectively, each method and the results of investigation using that method are described in detail. Chapter VII illustrates the use of the different procedures with an example and finally, in Chapter VIII, comparisons of the methods and their applicability under different circumstances are considered.

CHAPTER II

METHODS

2.1 Generation of Multivariate Samples

All the estimation procedures to be considered will assume samples from an underlying multivariate normal population. In order to carry out empirical studies, it was necessary to generate samples from multivariate normal populations with given characteristics. For p small, a convenient way of generating samples from a p-variate normal population with given parameters is given by applying a series of transformations to p independent sets of normal deviates. Finding the transformations involves getting the conditional distribution for x_2, given x_1, then, that for x_3, given x_1 and x_2, and so on, finally getting the conditional distribution for x_i, given the remaining p-1 variates. The transformation for the bivariate case is mentioned in the Rand Tables (1955) and that for the trivariate case is developed in Appendix A, section A.1. Actually, explicit expressions for the transformations are only of academic interest, since, for a given population, numerical values of the parameters may be substituted in the distribution function and the necessary conditional distributions found directly. This procedure, which has been described elsewhere by Moonan (1957), is illustrated in detail in Appendix A, section A.2 in finding the transformations necessary to generate samples from a four-variate normal population used in succeeding chapters.

Multivariate normal deviates having a mean of zero and variance equal to one have been generated, since we are not interested in trying to simulate any special population with regard to these parameters. In this way, different populations to be considered are characterized completely by their correlation matrices.

A convenient notation to indicate a given population consists of writing the correlation matrix for that population in a single line between brackets, omitting the diagonal and corresponding symmetrical elements, and setting off the rows of the correlation matrix by commas. For example, a population having the correlation matrix

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{bmatrix}$$

would be designated $[\rho_{12}\rho_{13}\rho_{14}, \rho_{23}\rho_{24}, \rho_{34}]$.

Since the Rand Tables (1955) afforded a ready source of N(0,1) deviates, they were used in punching a deck of two thousand cards, five deviates per card, for input in generating the multivariate samples needed. A description of the way the deviates were selected and a frequency tabulation of their values is given in Appendix B.

Henceforth, samples of these N(0,1) independent deviates will be referred to as "original samples". Their transformed counterparts will be called "generated samples".

To demonstrate the efficacy of the generation program, the matrix of sums, sums of squares, and sums of cross products for four hundred observations of a four-variate population is included in Appendix A, section A.3.

2.2 Censoring

Patterns of missing values to be considered do not include cases of observations with characters missing from every independent variate, that is, cases where an observation vector consists of only the dependent variable. It is clear that such observations can contribute nothing to a prediction equation for y from the x's.

A description of the notation used to designate patterns of missing values in general follows. For the trivariate case, $(c_1 c_2)$ is used to denote the number of characters, c_1, missing from x_1 alone, and c_2, the number of characters missing from x_2 alone. For the four-variate case $(c_1 c_2 c_3,\ c_{12} c_{13} c_{23})$ indicates $c_i (i = 1,2,3)$ as the number of characters missing in x_i alone and c_{ij} $(i \neq j)$ as the number of observation vectors having characters missing from both x_i and x_j. For example, a sample of fifteen observations designated (303,211) might appear as follows:

x_1	*	*	-	-	*	-	*	*	*	-	-	*	*	-	*
x_2	*	*	-	*	*	*	*	*	-	-	*	*	*	*	*
x_3	-	*	*	*	*	-	*	*	-	*	*	-	-	*	*
y	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

where a dash represents a missing value and an asterisk represents an observed value. The number of missing characters is given by $\sum c_i + 2 \sum c_{ij} = 14$ and the number of complete observations is given by $n = N - \sum_{i,ij} c = 5$. The extension to higher variate cases is obvious.

In the sampling work which follows, the artificial deletion of certain values among the independent variates to simulate missing measurements will be referred to as censoring.

The censoring schemes considered fall into two classes. One of these is the special case where values are missing from one variable only. The other is the case in which missing values occur with equal frequency among all the independent variables, with only one value missing from each observation vector. By the general notation indicated above, this latter type of censoring would be written $(c_1c_2c_3,000)$ for the four-variate case and $(c_1c_2c_3c_4,000000,0000)$ for the five-variate case. For these four- and five-variate cases the general notation has been abbreviated to $(c_1c_2c_3)$ and $(c_1c_2c_3c_4)$, respectively.

It should be emphasized, however, that except for the method of maximum likelihood, the estimation procedures discussed are not restricted in any way by the pattern of missing values.

2.3 Empirical Measures of Efficiency

The usefulness of a regression equation is usually judged by the magnitude of the bias and the variance of the estimated regression coefficients. The t-statistic, $t = \frac{\bar{b}_i - \beta_i}{s_{\bar{b}_i}}$, has been taken as a criterion for bias, and in the tables of biases in succeeding chapters, values which exceed the appropriate tabulated values of t at the five per cent level have been designated with an asterisk. However, it should be noted that we are interested primarily in prediction, and that estimates of regression coefficients may be biased in opposite directions so as to be mutually compensatory in nature as far as prediction is concerned. For this reason, as well as for general convenience, certain single measures of efficiency for a prediction equation have been proposed and will also be used to evaluate the methods discussed. One such measure defined by Nicholson (1948) is

$$E_Z = \frac{\overset{N}{\underset{\alpha=1}{S}} (y_\alpha - Y_\alpha)^2}{\overset{N}{\underset{\alpha=1}{S}} (y_\alpha - Z_\alpha)^2}$$

where y_α represents the observed value of y for the α^{th} observation; Y_α, the estimate of y_α based on the least squares regression equation of y on the x_i for the N uncensored observations; and Z_α, the estimate of y_α based on the prediction equation, Z, whose efficiency is being tested. Both sums are taken over the N observation vectors which give rise to Y. Thus $0 \leq E \leq 1$, assuming the maximum value when Y = Z.

Since the parameters are known, another important measure of efficiency is

$$\Delta_Z^2 = \frac{\overset{N}{\underset{1}{S}} (\xi_\alpha - Z_\alpha)^2}{N}$$

where $\xi = \overset{p-1}{\underset{i=1}{S}} \beta_i x_i$ is the true regression equation. Z_α is defined as above, and the sum is taken over the same N observation vectors as indicated above.

The measures E_Z and Δ_Z^2 are of interest when we consider Z as a prediction equation estimated from incomplete data.

Although both are measures of "prediction efficiency" and are usually referred to throughout the text by the word "efficiency", it is understood that they are merely convenient empirical devices for comparison, and, as such, are to be distinguished from orthodox theoretical efficiency which can be defined for an estimate, θ', as

$$E_f = \frac{\sigma^2_{\hat{\theta}}}{\sigma^2_{\theta'}}$$

where $\sigma^2_{\hat{\theta}}$ is the variance of $\hat{\theta}$, the minimum variance unbiased estimator of θ.

It may be noted that E is an index which measures the goodness of the equation in question <u>relative</u> to the "best" estimate from the uncensored sample. It should be pointed out in this connection that E may be very high although neither estimate is good in the sense of giving a prediction equation close to the true one. On the other hand, Δ^2_Z, which is a measure of the average squared deviation of the Z from values predicted by the true regression equation over a given sample, is dependent on the population from which samples were generated. Hence, it is not useful in comparing methods applied to samples from different populations.

In the tables which follow, the values of Δ^2 are considered in terms of Δ, since sampling distributions of Δ appear to be more symmetrical than those of Δ^2.

2.4 Description of Experiments

As was indicated in Chapter I, an empirical sampling approach was used. This approach took the form of a series of eight sampling experiments which were performed in sequence and were dictated in large part by successive results. The principal factors considered were the characteristics of the populations from which samples were taken and the intensity of censoring. Except for one experiment, sample size was confined to N = 20. Since we have some knowledge of how sample size affects estimates, this does not seem to be a very objectionable restriction. The populations, intensity of censoring, and number of samples on which calculations were made varied from experiment to experiment. The number of samples considered in a given experiment or on which a set of means and variances was based is denoted by k.

A brief description of the experiments conducted is set forth in the following paragraphs for the twofold purpose of (1) providing some perspective before the presentation of detailed results in chapters concerned with the respective methods of estimation and (2) identifying the experiments so that they need not be described repeatedly in various places throughout the text.

Experiment 1 was the most comprehensive experiment undertaken. It consisted of estimating regression equations for an arrangement in which twenty original samples were transformed to produce samples from the trivariate populations, [.25 .50, .25] and [.80 .60, .80], and the four-variate population [.25 .50 .75, .25 .50, .25]. Each of the resulting forty generated samples from the trivariate populations was censored in ten different ways, (11), (22), (33), (44), (55), (66), (04), (40), (08), (80). The twenty generated samples from the four-variate population were each censored (111), (222), (333), (444), and (555). Estimates of prediction equations were made from the samples at each of these population-censoring combinations using the uncensored data, complete observations, the Greenberg Procedure, and the Method of Paired Correlation Coefficients.

In Experiment 2, samples from thirty-nine different trivariate populations were censored (55) and regression statistics calculated using the Method of Paired Correlation Coefficients. The main purpose was to find out how different populations affected the efficiency of the prediction equations determined by using this method. The experiment can be subdivided into two parts. In Experiment 2A, the same ten original samples were used to provide ten generated samples from each of the thirty-nine populations. In Experiment 2B, 390 different sets of original samples were used in generating ten samples from each of the thirty-nine populations. The rationale dictating the use of this type of design is based on the premise that the mean efficiencies at each point in the population space, relative to other points, may be better determined by using the same set of original samples, but that the entire configuration in the space would be better determined by using different random samples at each point. This feature of the sampling plan represents a unique facet of the general empirical sampling approach. It is also used in Experiment 4.

Experiment 3 was performed to determine whether or not conclusions based on Experiment 2 also held in the four-variate case. Twenty original samples were used to generate twenty samples from each of six different four-variate populations. The samples were censored (555) and regression equations determined using the Method of Paired Correlation Coefficients.

In Experiment 4, the Greenberg Procedure was used on samples censored (55) from sixteen different trivariate populations. The populations selected were dictated by a response surface design in the space of admissible sets of correlation coefficients. Like Experiment 2, this experiment consisted of two parts: 4A, involving the use of ten original samples to generate ten samples from each population, and 4B, making use of different sets of original samples for each population.

Experiment 5 was designed to investigate the convergence of successive estimates made by the Greenberg Procedure. It was limited in scope and is described in detail in the chapter dealing with the Greenberg Procedure.

The Paired Correlation Method is applied to samples from two five-variate populations in Experiment 6. The populations were chosen to generate samples simulating those arising in two "live problem" situations. The purpose was to demonstrate cases where the Paired Correlation Method would or would not be appropriate. The method was used on two hundred samples censored (3333), one hundred from each of the populations examined.

Experiment 7 involved the use of complete observations, the Greenberg Procedure, and the Paired Correlation Method to find regression equations from one hundred four-variate samples censored (333). The population chosen was one for which the Method of Paired Correlation Coefficients was not expected to provide efficient results. The experiment aimed at determining, under such circumstances, the relative "goodness" of prediction with the use of complete observations, The Greenberg Procedure, and the Paired Correlation Method.

In Experiment 8, the Greenberg Procedure was used on twenty samples from the two five-variate populations considered in Experiment 6. The primary function of this work was to study convergence properties under the Greenberg Procedure, however, the results in terms of prediction efficiency were tabulated and discussed.

Values of Δ_Z and E_Z, as well as estimates of the regression parameters, were computed for each of the cases in the experiments described above.

2.5 The Computing Programs

There are a few points about the computing programs which merit discussion. The input for every program consisted of sets of p independent N(0,1) deviates which were transformed into multivariate normal deviates and stored in a data table on the drum. After twenty vectors had been generated, the computations were carried out and the process repeated, the transformed values never existing outside the machine. The transforming coefficients for a given population were computed manually and stored on the drum to be used repeatedly for a series of samples.

Although the data consisted of a complete table of N(N=20) p-dimensional vectors, they were accompanied by a censoring code comprised of Np binary digits, each one denoting the presence or absence of its corresponding deviate value. Censoring codes, like the transforming coefficients mentioned above, were punched on cards and stored on the drum to be used repeatedly for a series of samples.

One device used to check the correctness of a program for different methods of estimation consisted of using a censoring code which indicated no missing values. All methods of estimation should then yield identical results.

Symbolic Optimum Assembly Programming (SOAP) was used in writing all programs.

CHAPTER III

THE DELETION OF INCOMPLETE OBSERVATIONS

3.1 Discussion

The most obvious solution to the problem of missing values is simply to discard observation vectors which are not complete in every variate. We will refer to this procedure as estimation from complete data. Actually, such treatment hardly can be called a solution, and, at best, must be considered a trivial one. The main purposes of including a section devoted to this procedure are to point out objections to it, and to include results given by it which will serve as standards of comparison for results obtained by other methods.

The objection to such a procedure is that the efficiency of the prediction equation might be decreased by casting out important information. In deciding whether to use only data complete in every variate or whether to employ some other method making use of all the information, the first factor to be considered is the amount and proportion of complete data available. If a large amount of data is available, one usually would not hesitate to discard a part of it in order to apply a straightforward analysis.

On the other hand, there might be enough complete observation vectors available so that the others could be discarded, but by so doing, the variance of predicted values would be increased or estimates biased. Many times values missing from incomplete observation vectors occur near the extremes of the range of the variable in question. The occurrence of missing values is, in fact, often correlated with their extreme size for one reason or another. It is likely that observed values in an incomplete vector fall at very high or very low levels also, and by discarding such values the range of the x's is decreased, resulting in diminished precision of the regression estimate. For example, in the field of anthropometry, Münter (1936) and his associates, in measuring skeletal remains, found that bones of smaller specimens were often broken and could not be measured. If a prediction equation for some specific characteristic were to be determined, or if a linear discriminator were to be established, then discarding all the measurements of the specimens with at least one broken bone would clearly be undesirable. Another example applicable to the present discussion arises in trying to predict earnings based on questionnaire information regarding an employee's experience, education, place of employment, and other variables. Non-response on education is known to be correlated with low earnings.

It follows from the above discussion that consideration should be given to the levels of the x's at which one expects to apply the prediction equation. In predicting y for values in the middle range of the x's, there would be less concern about discarding incomplete observations having extreme x values than would otherwise be the case.

3.2 Bias and Variance of Regression Coefficients

The data from Experiment 1, computed by using only complete observation vectors, is summarized in Table 1 which gives the biases and variances of the regression coefficients for different combinations of censoring and population. Since the estimates obtained by deleting incomplete observation vectors are ordinary least squares estimates based on $N - \sum_i c_i = n$ observations, they are expected to be unbiased

Table 1. Statistics of regression coefficients estimated by the deletion of incomplete observation vectors; Experiment 1, N = 20, k = 20

Censoring	[.25 .50 , .25]						[.80 .60 , .80]					
	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$	$\bar{b}_o-\beta_0$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$
(00)	.113*	.055	.051	.058	-.033	.025	.078*	.026	.060	.052	-.037	.032
(11)	.125*	.055	.069	.072	.014	.042	.087*	.027	.038	.076	.015	.053
(22)	.145*	.050	.082	.093	.037	.046	.101*	.024	.030	.101	.042	.058
(33)	.193*	.062	.056	.135	.064	.046	.134*	.030	-.007	.123	.072	.058
(44)	.150*	.055	.048	.150	.086	.075	.104*	.026	-.029	.150	.096	.096
(55)	.124*	.068	.009	.195	.053	.095	.086*	.033	-.032	.197	.059	.120
(66)	.194*	.144	.018	.243	.016	.208	.135*	.070	.001	.312	.018	.262
(04)	.098	.063	.034	.076	-.001	.041	.069	.030	.014	.071	-.001	.052
(40)	.150*	.055	.059	.121	.048	.040	.104	.017	.007	.109	.054	.050
(08)	.086	.082	-.029	.074	-.053	.053	.060	.040	.018	.078	-.060	.067
(80)	.125	.072	.117	.209	.021	.050	.087	.035	.067	.168	.023	.063

Censoring	[.25 .50 .75 , .25 .50 , .25]							
	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$	$\bar{b}_3-\beta_3$	$\hat{V}(b_3)$
(000)	.028	.016	-.019	.030	.035	.020	.050	.021
(111)	.068*	.014	-.001	.037	.056	.018	.036	.028
(222)	.064	.032	.009	.058	.051	.038	.002	.040
(333)	.042	.035	-.001	.092	.006	.048	.017	.055
(444)	.024	.112	.069	.279	.030	.100	-.022	.099
(555)	-.023	.398	.090	1.008	.371	2.763	.112	.866

[a] The symbol * indicates $t = \dfrac{\bar{b}_i - \beta_i}{s_{\bar{b}_i}} > t_{.05,\,19}$, where $s^2_{\bar{b}_i} = \dfrac{\hat{V}(b_i)}{20}$

with variance increasing as n decreases. Table 1 seems to bear out these expectations. None of the biases of b_1 and b_2 are significant by the t criterion given in Chapter II. The bias of b_0 which is indicated significant for most of the three variate cases here may be dismissed as a sampling phenomenon on the basis of similar unbiased statistics calculated from other sets of original samples censored (55) in Experiment 4. Recall that the data of Table 1 for each censoring population combination is based on the same set of twenty original deviate samples.

Further evidence confirming the expected normal character of the regression coefficients based on complete data can be seen in the histograms of Figures 1 and 2. These data were taken from experiments concerned with higher variate cases. Each figure is based on different sets of one hundred original samples. For the data in Figure 1, the values of t for testing bias in b_0, b_1, b_2, b_3, and b_4 were -.818, -.391, -.473, .212 and .760, respectively, while the t values for bias in the data represented by Figure 2 were -.521, .886, -.068, and .904 for b_0, b_1, b_2, and b_3, respectively.

The mean values and variances of the $s^2_{y.x}$ from Experiment 1 have been computed and are tabulated in Table 2. As in Table 1, the tendency for the variances of the estimates to increase with increased censoring may be noted. A similar tendency with regard to the bias is not evident. Figure 10(B) shows the histogram of values of $s^2_{y.x}$ from one hundred samples censored (333). The shape of the histogram is consistent with the theoretical distribution of such a statistic, which is proportional to χ^2 with seven degrees of freedom.

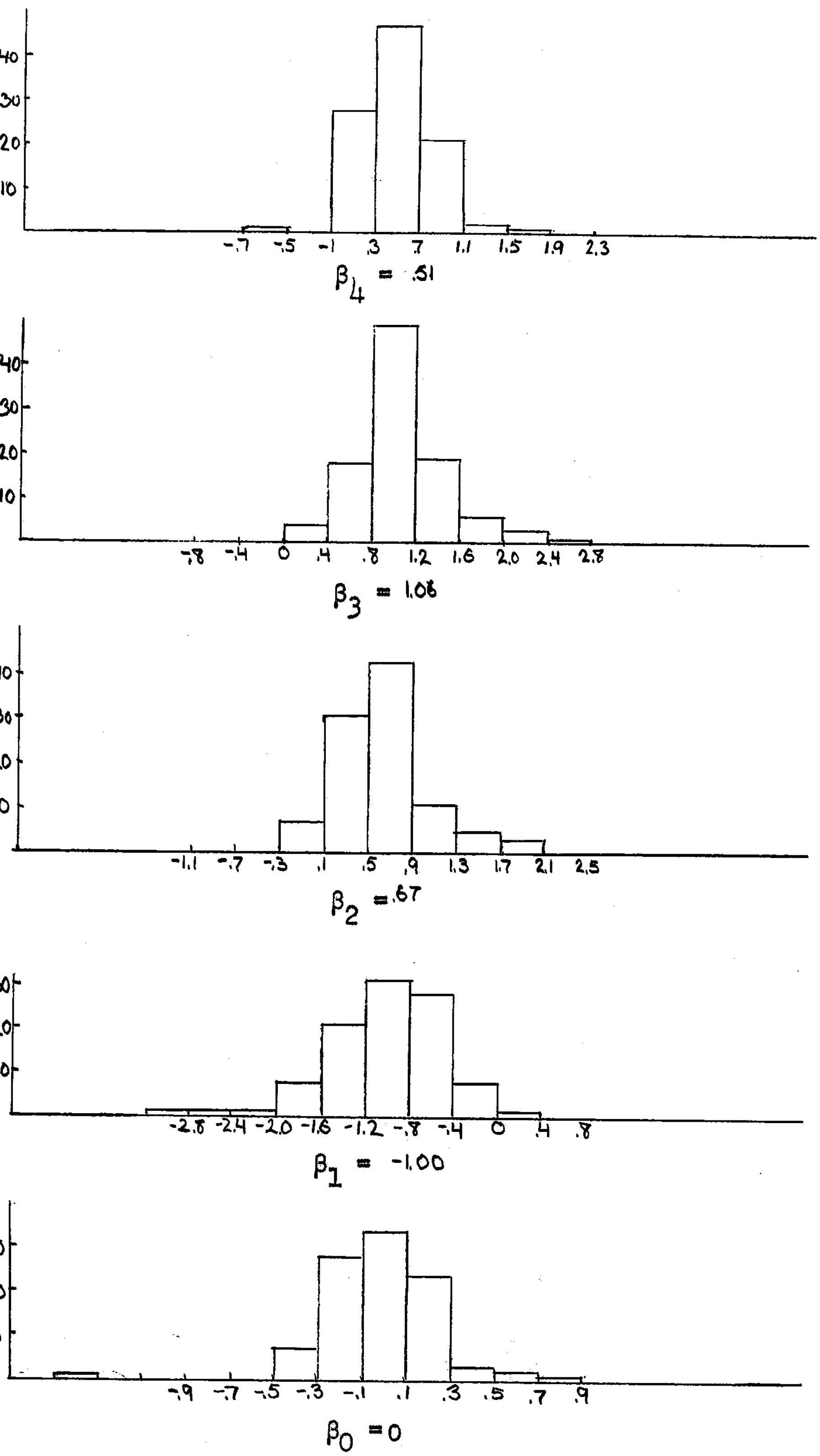


Figure 1. Histograms of regression coefficients estimated by deletion of incomplete observations; [0.80 .29 .40, .50 -.20 .50, .20 .70 .30] censoring (3333), N = 20, k = 100

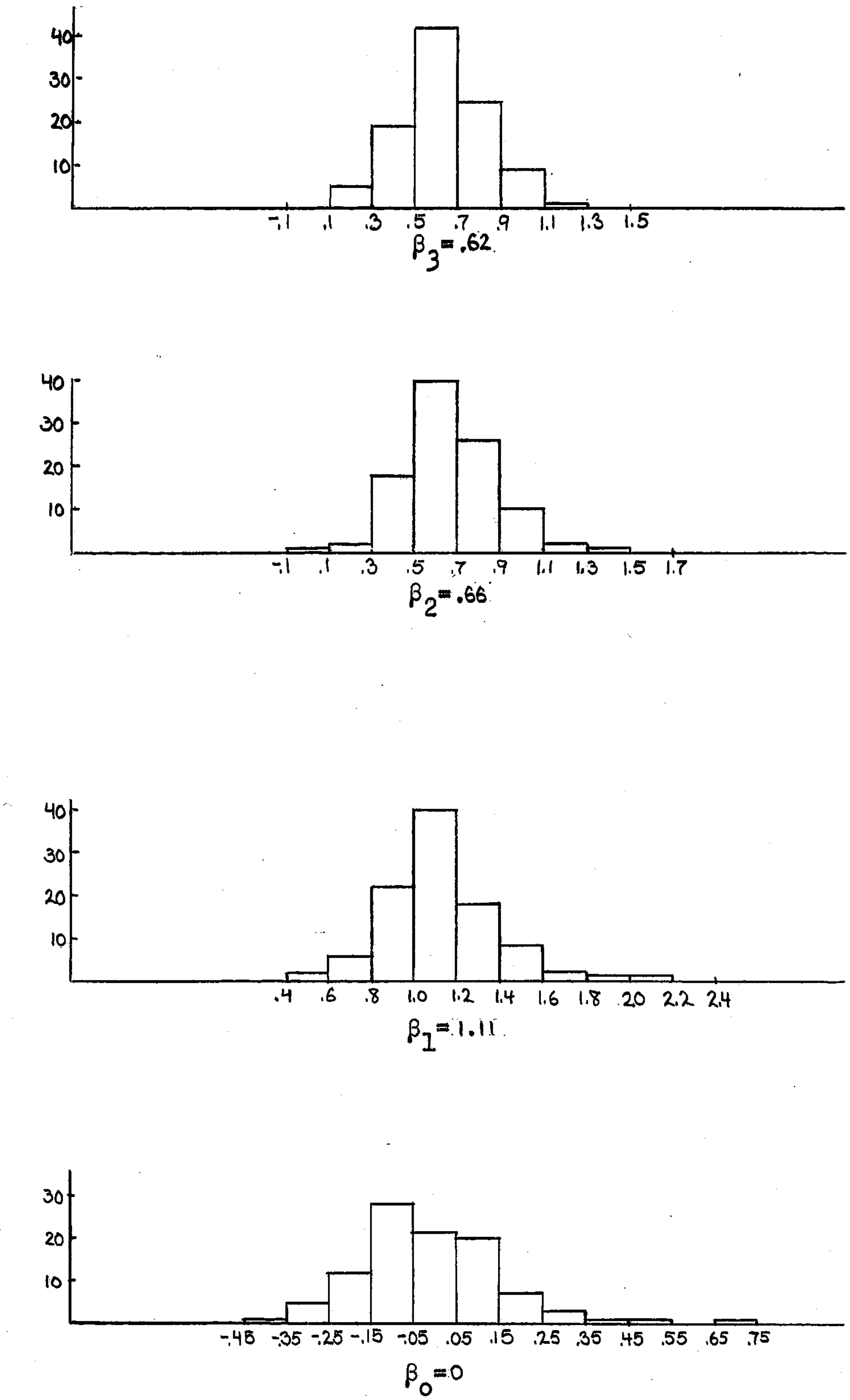


Figure 2. Histograms of regression coefficients estimated from complete observations; [-.70 -.52 .32, .60 .26, .44], censoring (333), N = 20, k = 100

Table 2. Means and variances of errors of estimate obtained by the deletion of incomplete observation; Experiment 1, N = 20, k = 20

Censoring	[.25 .50 , .25] $\sigma^2_{y.x} = .7333$		[.80 .60 , .80] $\sigma^2_{y.x} = .3556$		Censoring	[.25 .50 .75 , .25 .50 , .25] $\sigma^2_{y.x} = .3239$	
	$\overline{s^2_{y.x}}$	$\hat{V}(s^2_{y.x})$	$\overline{s^2_{y.x}}$	$\hat{V}(s^2_{y.x})$		$\overline{s^2_{y.x}}$	$\hat{V}(s^2_{y.x})$
(00)	.742	.082	.360	.019	(000)	.349	.007
(11)	.734	.094	.356	.022	(111)	.361	.009
(22)	.736	.076	.357	.018	(222)	.345	.015
(33)	.756	.110	.366	.026	(333)	.370	.026
(44)	.740	.143	.358	.034	(444)	.315	.064
(55)	.770	.208	.373	.049	(555)	.351	.310
(66)	.724	.106	.351	.025			
(04)	.737	.086	.357	.020			
(40)	.750	.118	.364	.028			
(08)	.701	.097	.340	.032			
(80)	.763	.136	.370	.023			

As mentioned above, the values in Tables 1 and 2 are included primarily as standards for comparison with similar tables based on statistics calculated by methods which utilize incomplete observation vectors.

The estimated variances of regression coefficients referred to above are the "between sample" or external estimates of variances, denoted $\hat{V}(b_i)$. Although the estimated variances in succeeding discussions will be of this type, it seems appropriate to mention that "within sample" or in Bliss' (1952) terms, the internal variances $\hat{V}_w(b_i)$, are available here. The inconsistency between estimates of these two types has long been of concern in bioassay applications, but as an experimental rather than a statistical problem (Bliss, 1952; Sheps and Munson, 1957). An examination of the consistency of these two ways of estimating the variance of regression coefficients is of interest as a by-product of this research. Table 3 gives mean values of the $\hat{V}_w(b_i)$. A comparison of these with their counterparts in Table 2 reveals no startling discrepancies.

3.3 Efficiency of Prediction Equations

Table 4A gives the means of E and their standard errors for the censoring schemes of Experiment 1. When estimates are obtained by deleting incomplete observation vectors, E is invariant for samples representing different populations, but generated from the same original samples.

The means of Δ and their standard errors are shown in Table 4B.

Table 3. Means of $\hat{V}_w(b_i)$ estimated by the deletion of incomplete observations; Experiment 1, N = 20, k = 20

Censoring	[.25 .50, .25]			[.80 .60, .80]			Censoring	[.25 .50 .75, .25 .50, .25]			
	$\overline{\hat{V}_w(b_o)}$	$\overline{\hat{V}_w(b_1)}$	$\overline{\hat{V}_w(b_2)}$	$\overline{\hat{V}_w(b_o)}$	$\overline{\hat{V}_w(b_1)}$	$\overline{\hat{V}_w(b_2)}$		$\overline{\hat{V}_w(b_o)}$	$\overline{\hat{V}_w(b_1)}$	$\overline{\hat{V}_w(b_2)}$	$\overline{\hat{V}_w(b_3)}$
(00)	.042	.052	.048	.020	.064	.060	(000)	.021	.034	.023	.031
(11)	.047	.060	.058	.023	.075	.073	(111)	.028	.043	.032	.040
(22)	.055	.075	.063	.027	.086	.080	(222)	.036	.058	.040	.048
(33)	.069	.093	.081	.033	.106	.102	(333)	.061	.090	.068	.073
(44)	.078	.107	.095	.038	.122	.120	(444)	.080	.128	.109	.117
(55)	.108	.159	.122	.052	.166	.155	(555)	.798	1.402	3.994	1.635
(66)	.151	.229	.191	.073	.251	.241					

Table 4A. Values of $\bar{E}$ and their standard errors estimated by the deletion of incomplete observations; Experiment 1, N = 20, k = 20

Censoring	Trivariate $\bar{E}$	Trivariate S.E.($\bar{E}$)	Censoring	Four Variate $\bar{E}$	Four Variate S.E.($\bar{E}$)
(11)	.965	.009	(111)	.967	.007
(22)	.939	.011	(222)	.888	.018
(33)	.885	.018	(333)	.795	.033
(44)	.855	.023	(444)	.625	.056
(55)	.803	.028	(555)	.335	.045
(66)	.722	.039			
(04)	.944	.011			
(40)	.939	.011			
(08)	.881	.020			
(80)	.859	.025			

Table 4B. Values of $\bar{\Delta}$ and their standard errors estimated by the deletion of incomplete observations; Experiment 1, N = 20, k = 20

Censoring	[.25 .50, .25] $\bar{\Delta}$	[.25 .50, .25] S.E.($\bar{\Delta}$)	[.80 .60, .80] $\bar{\Delta}$	[.80 .60, .80] S.E.($\bar{\Delta}$)	Censoring	[.25 .50 .75, .25 .50, .25] $\bar{\Delta}$	[.25 .50 .75, .25 .50, .25] S.E.($\bar{\Delta}$)
(00)	.303	.028	.211	.020	(000)	.222	.012
(11)	.340	.031	.237	.021	(111)	.252	.016
(22)	.375	.032	.261	.022	(222)	.301	.019
(33)	.437	.039	.304	.027	(333)	.348	.029
(44)	.460	.037	.320	.026	(444)	.524	.074
(55)	.488	.039	.340	.027			
(66)	.607	.074	.423	.051			
(04)	.353	.030	.245	.021			
(40)	.385	.039	.268	.027			
(08)	.381	.030	.266	.021			
(80)	.468	.055	.326	.038			

These tables demonstrate an expected relationship, that greater censoring decreases efficiency of prediction. By so doing, the results lend support to the use of E_Z and Δ_Z as effective measures of efficiency.

The histograms of Figures 7A, 11A, and 12B indicate distributions of E and Δ when using only complete observations for certain cases considered in Experiments 6 and 7.

Chapter VIII includes tables in which mean efficiencies and their standard errors based on estimates from complete observations, are compared with their counterparts based on other methods.

CHAPTER IV

THE METHOD OF PAIRED CORRELATION COEFFICIENTS

4.1 Definition

Although the procedure described below seems to have been used widely to deal with the problem at hand, there is no apparent reference to it in the literature, either describing it or even referring to it as a technique employed in analyzing incomplete data. It is not clear whether this is due to the rather simple nature of the scheme or whether it is a reflection of its limited usefulness. At any rate, it will be described below and some of its limitations will be brought out in connection with results given later in the chapter.

Suppose we have a set of data such as that schematically represented on page 6. From this we can form the matrix of paired correlation coefficients:

$$C_{pxp} = \left[\begin{array}{ccccc|c} 1 & r_{12} & r_{13} & \cdots & r_{1(p-1)} & r_{1y} \\ & 1 & r_{23} & \cdots & r_{2(p-1)} & r_{2y} \\ & & & & 1 & r_{(p-1)y} \\ \hline & & & & & 1 \end{array}\right] = \left[\begin{array}{c|c} C_{(p-1)(p-1)} & \underline{r}_{iy} \\ \hline \underline{r}'_{iy} & 1 \end{array}\right] \qquad (4.1)$$

where the r_{ij} are correlation coefficients estimated from the existing n_{ij} pairs of observations in the sample of N observation vectors. The standardized regression coefficients, $b_1^* \; b_2^* \; \ldots \; b_{(p-1)}^*$, are then obtained as

$$\underline{b}^* = C^{-1}_{(p-1)(p-1)}\underline{r}_{iy} \qquad (4.2)$$

The b_i^* may then be converted to ordinary regression coefficients by substituting the appropriate statistics for the parameters in the well-known formulae

$$\beta_i = \beta_i^* \frac{\sigma_y}{\sigma_i} \tag{4.3}$$

and

$$\beta_0 = \mu_y - \beta_1\mu_1 - \beta_2\mu_2 - \cdots - \beta_{p-1}\mu_{p-1}. \tag{4.4}$$

The sample values for the mean and variance of a variable are taken over all existing values of that variable; that is,

$$\bar{x}_i = \frac{\sum_1^{n_i} x_i}{n_i} \tag{4.5}$$

and

$$s_i^2 = \frac{\sum_1^{n_i} (x_i - \bar{x}_i)^2}{n_i - 1} \tag{4.6}$$

Generally, $n_i \geq n_{ij}$. As mentioned in Chapter I, the restriction, $n_y = N$, will be maintained here.

An estimate of the error variance $\sigma^2_{y.x}$ may also be obtained by substituting sample values given above for parameters in the formula

$$\sigma^2_{y.x_1x_2 \ldots x_{p-1}} = \sigma_y^2 - \underline{\beta}'V\underline{\beta} \tag{4.7}$$

where $\underline{\beta}$ is the vector of regression coefficients, β' is its transpose, and V is the variance-covariance matrix of the x_i. It is clear that this method of estimation reduces to ordinary least squares regression when $n_{ij} = N$; that is, where no values are missing.

4.2 Bias and Variance of Regression Coefficients

The biases, $\bar{b}_i - \beta_i$, and variances, $\hat{V}(b_i)$, for the population-censoring schemes of Experiment 1 are indicated in Table 5. There is no marked tendency for bias to increase with increased censoring, and only the biases in b_0 appear to be significant by the t criterion adopted to assess this bias. The remarks in Chapter III, pointing out that the significant bias of b_0 is due to sampling, are also applicable here. There were no significant biases among the b_1 and b_2 in Table 5. In the four-variate situation, at higher levels of censoring, the values of $\bar{b}_i - \beta_i$ are very large, although not significantly so by the t tests. This draws our attention to the large variances at these levels. In certain situations, which will be discussed in succeeding sections, the biases, $\bar{b}_i - \beta_i$, become very large as do the variances of the $\bar{b}_i$. Consequently, one may be estimating the coefficients very badly in using this procedure, although the t criterion for bias would not indicate this. Examples of these large variances are illustrated in the results of Experiments 6 and 7, which will be discussed in the section on efficiencies later in this chapter. The statistics relating to the bias and variances of regression coefficients computed by the Paired Correlation Method in Experiment 7 are given in the second column of Table 16 and serve to point up the statements made above. Histograms of these regression coefficients may be seen in Figure 3.

4.3 Error Variance

The error variance, $s^2_{y.x}$, was predicted for each sample as described in section 4.1, equation (4.7). The mean values and

Table 5. Statistics of regression coefficients estimated by the Method of Paired Correlation Coefficients; Experiment 1, N = 20, k = 20

Censoring	[.25 .50 , .25]						[.80 .60 , .80]					
	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$
(00)	.113*[a]	.055	.051	.058	-.033	.025	.078*	.027	.060	.052	-.037	.032
(11)	.121*	.050	.066	.064	-.001	.034	.082	.031	.040	.082	-.002	.059
(22)	.124*	.050	.072	.067	.017	.033	.078	.032	.044	.108	.000	.083
(33)	.120*	.055	.047	.084	.026	.031	.089	.042	-.020	.157	.039	.105
(44)	.147*	.058	.042	.102	.052	.029	.092	.046	-.048	.198	.060	.137
(55)	.187*	.083	.068	.136	.066	.067	.091	.050	.027	.250	.021	.152
(66)	.202*	.078	.141	.166	.009	.132	.086	.046	.080	.485	-.027	.298
(04)	.125*	.049	.045	.055	.017	.039	.096*	.036	.031	.089	.006	.080
(40)	.135*	.065	.048	.102	.002	.028	.079	.030	-.091	.170	.041	.090
(08)	.108	.058	.048	.077	-.108	.076	.085	.049	-.061	.182	-.079	.130
(80)	.128	.083	.075	.225	-.046	.058	.092*	.038	.038	.354	-.047	.162

Censoring	[.25 .50 .75 , .25 .50 , .25]							
	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$	$\bar{b}_3-\beta_3$	$\hat{V}(b_3)$
(000)	.028	.016	-.019	.030	.035	.020	.049	.021
(111)	.051	.020	.017	.041	.043	.034	.025	.033
(222)	.054	.023	.041	.055	.076	.041	-.005	.042
(333)	.051	.026	.032	.079	.087	.047	-.017	.064
(444)	.261	.646	-.270	2.244	.449	2.310	.613	9.190
(555)	.329	.973	.239	.609	.475	.749	.157	.928

[a] The symbol * indicates $t = \dfrac{\bar{b}_i - \beta_i}{s_{\bar{b}_i}} > t_{.05,\ 19}$, where $s^2_{\bar{b}_i} = \dfrac{\hat{V}(b_i)}{20}$

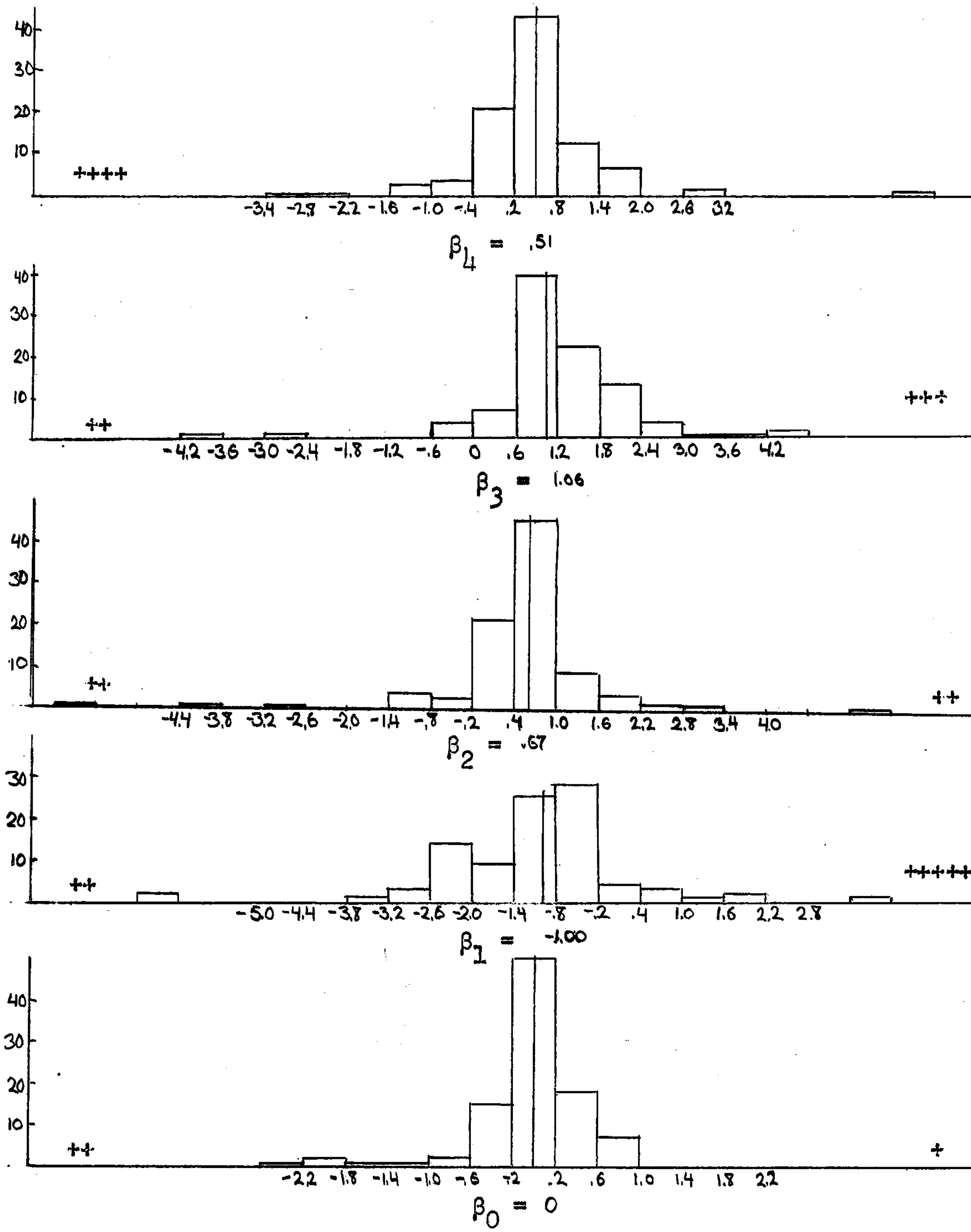


Figure 3. Histograms of regression coefficients estimated by the Paired Correlation Method;
[.60 .80 .29 .40, .50 -.20 .50, .20 .70, .30]
censoring (3333), N = 20, k = 100

The symbol + indicates a value outside the range of scale.

variances of these estimates from Experiment 1 are given in Table 6. Table 6 reveals the negative bias in the estimates attendant with increased censoring, and suggests that the variance of the estimates increases as censoring becomes heavy. An examination of the individual $s^2_{y.x}$ for the four-variate population of Experiment 7, censored (333), points up this bias in a rather startling manner by revealing negative estimates of variance. The distribution of $s^2_{y.x}$ at this particular point is indicated in Figure 10D by a histogram of the estimates from one hundred samples. It becomes apparent that it is possible for the negative estimates to arise when we realize that the correlation matrix, $C_{(p-1)(p-1)}$, is not guaranteed positive definite, since the r_{ij} are based on different n_{ij} pairs.

4.4 Efficiency of Prediction Equations

Table 7A gives the means of E and their standard errors for the various population-censoring combinations of Experiment 1. Table 7B shows the mean values of Δ and their associated standard errors. There appears to be a very definite population effect influencing the efficiencies. It is clearly demonstrated by testing the hypothesis that geometric mean values of E representing the two populations are equal, assuming the usual linear model. An analysis of variance of the logarithms of E for the three variate populations censored (11) to (66) appears in Table 8.

Table 6. Means and variances of errors of estimate obtained by the Method of Paired Correlation Coefficients; Experiment 1, N = 20, k = 20

Censoring	[.25 .50 , .25] $\sigma^2_{y.x} = .7333$		[.80 .60 , .80] $\sigma^2_{y.x} = .3556$	
	$\overline{s^2_{y.x}}$	$\hat{V}(s^2_{y.x})$	$\overline{s^2_{y.x}}$	$\hat{V}(s^2_{y.x})$
(00)	.741	.082	.360	.019
(11)	.715	.084	.349	.017
(22)	.715	.093	.340	.012
(33)	.710	.098	.332	.016
(44)	.696	.096	.308	.015
(55)	.652	.095	.308	.015
(66)	.611	.102	.289	.032
(04)	.718	.087	.341	.015
(40)	.723	.092	.326	.015
(08)	.718	.082	.333	.021
(80)	.628	.109	.296	.034

Censoring	[.25 .50 .75 , .25 .50 , .25] $\sigma^2_{y.x} = .3239$	
	$\overline{s^2_{y.x}}$	$\hat{V}(s^2_{y.x})$
(000)	.349	.007
(111)	.340	.006
(222)	.320	.012
(333)	.319	.008
(444)	.194	.219
(555)	.064	.364

Table 7A. Values of $\bar{E}$ and their standard errors estimated by the Method of Paired Correlation Coefficients; Experiment 1, N = 20, k = 20

Censoring	[.25 .50, .25] $\bar{E}$	[.25 .50, .25] S.E.($\bar{E}$)	[.80 .60, .80] $\bar{E}$	[.80 .60, .80] S.E.($\bar{E}$)	Censoring	[.25 .50 .75, .25 .50, .25] $\bar{E}$	[.25 .50 .75, .25 .50, .25] S.E.($\bar{E}$)
(11)	.983	.005	.962	.010	(111)	.967	.008
(22)	.972	.006	.943	.013	(222)	.907	.021
(33)	.958	.009	.901	.025	(333)	.870	.028
(44)	.948	.010	.874	.028	(444)	.784	.053
(55)	.918	.021	.819	.038	(555)	.606	.072
(66)	.857	.035	.769	.046			
(04)	.983	.005	.931	.012			
(40)	.967	.008	.939	.019			
(08)	.936	.024	.833	.027			
(80)	.906	.029	.880	.042			

Table 7B. Values of $\bar{\Delta}$ and their standard errors estimated by the Method of Paired Correlation Coefficients; Experiment 1, N = 20, k = 20

Censoring	[.25 .50, .25] $\bar{\Delta}$	[.25 .50, .25] S.E.($\bar{\Delta}$)	[.80 .60, .80] $\bar{\Delta}$	[.80 .60, .80] S.E.($\bar{\Delta}$)	Censoring	[.25 .50 .75, .25 .50, .25] $\bar{\Delta}$	[.25 .50 .75, .25 .50, .25] S.E.($\bar{\Delta}$)
(00)	.303	.028	.211	.020	(000)	.222	.012
(11)	.321	.028	.246	.020	(111)	.262	.016
(22)	.329	.028	.251	.024	(222)	.296	.019
(33)	.352	.029	.297	.034	(333)	.330	.024
(44)	.376	.033	.317	.036	(444)	.807	.454
(55)	.422	.049	.360	.039	(555)	.774	.202
(66)	.484	.054	.408	.057			
(04)	.329	.025	.253	.023			
(40)	.355	.035	.270	.030			
(08)	.360	.029	.335	.031			
(80)	.430	.063	.338	.057			

Table 8. Analysis of variance of ln E for the trivariate populations of Experiment 1

Variation due to	d.f.	S.S.	M.S.	F
Censoring (C)	5	1.2691	.2538	8.29
Population (P)	1	.3902	.3902	12.75**
C x P	5	.1162	.0232	NS
Error	228	6.9693	.0306	
Total	239	8.7448		

In order to define more clearly the regions of differential efficiency in the three-variate case, Experiment 2 was set up to evaluate efficiency at a large number of points in the population space at a fixed level of censoring (55). The aim of the experiment was to find a response surface to describe efficiency for different types of populations. Thirty-nine populations were chosen to represent regions in the space of admissible ρ_{12}, ρ_{1y}, ρ_{2y} values for which ρ_{1y} and ρ_{2y} were positive. The points chosen are given in the first column of Table 9. As mentioned in Chapter II, the experiment was divided into two parts; part A, in which the same set of ten original samples was used to generate the samples on which estimates were made at each point, and part B, in which 390 different samples were used to generate ten samples at each of the thirty-nine points. The purpose of choosing this type of sampling scheme was discussed in Chapter II. Table 10 shows that the use of the same set of ten original samples at each point in Experiment 2A makes for an expected reduction in the estimated error variance of the efficiencies.

Table 9. Values of $\overline{E}$ and residuals from the response equation, $\hat{E}_{A+B}$, of Experiment 2

Population			$\hat{E}_{A+B}$	$\overline{E}_A$	$\overline{E}_B$	$\overline{E}_A - \hat{E}_{A+B}$	$\overline{E}_B - \hat{E}_{A+B}$
[.80	.60,	.80]	.8047	.8046	.8819	-.0001	.0772
[-.70	.30,	.40]	.5454	.4835	.3218	-.0619	-.2236
[-.60	.30,	.20]	.7735	.7937	.8599	.0202	.0864
[-.30	.30,	.30]	.9012	.8387	.8949	-.0625	-.0063
[-.30	.30,	.60]	.7159	.7884	.7988	.0725	.0829
[.10	.30,	.20]	1.0164	.8977	.9612	-.1187	-.0552
[-.10	.30,	.70]	.7663	.8090	.7382	.0421	-.0281
[.50	.30,	.60]	.9368	.9159	.9311	-.0209	-.0057
[.70	.30,	.70]	.8776	.8057	.8220	-.0719	-.0556
[.30	.30,	.80]	.8575	.8825	.8943	.0250	.0368
[.50	.30,	.20]	.9219	.9026	.9268	-.0193	.0049
[.70	.50,	.20]	.7670	.7428	.8102	-.0242	.0432
[.60	.50,	.80]	.8556	.8694	.8219	.0138	-.0337
[-.50	.50,	.20]	.7434	.7806	.8562	.0372	.1128
[-.20	.50,	.40]	.8265	.8245	.8403	-.0020	.0138
[.20	.50,	.10]	.9293	.8863	.9332	-.0430	.0039
[.40	.50,	.30]	.9217	.8936	.9195	-.0281	-.0022
[.60	.50,	.50]	.8883	.8908	.9418	.0025	.0535
[.60	.70,	.80]	.8030	.8734	.8342	.0704	.0312
[.90	.70,	.50]	.7006	.5468	.6907	-.1538	-.0099
[-.50	.70,	.10]	.6468	.6998	.5719	.0530	-.0749
[.30	.80,	.30]	.7400	.8218	.7197	.0318	-.0703
[.75	.20,	.30]	.8212	.8829	.8579	.0617	.0367
[.69	.30,	.40]	.8753	.9003	.8273	.0250	-.0480
[.87	.50,	.55]	.7841	.8318	.8843	.0477	.1002
[-.20	.70,	.40]	.7189	.7499	.6603	.0310	-.0586
[.20	.70,	.20]	.8426	.8460	.8497	.0034	.0069
[.90	.70,	.87]	.7480	.6889	.6515	-.0591	-.0965
[.30	.70,	.70]	.7995	.8377	.8007	.0382	.0012
[.60	.70,	.10]	.6924	.6994	.7153	.0070	.0229
[.60	.70,	.50]	.8208	.8431	.9058	.0223	.0850
[.18	.30,	.50]	.9665	.8972	.9359	-.0693	-.0306
[-.55	.21,	.30]	.7804	.8046	.8332	.0242	.0528
[-.32	.31,	.50]	.7729	.8096	.8006	.0367	.0277
[.89	.67,	.70]	.7576	.8062	.7646	.0484	.0070
[.30	.50,	.63]	.8947	.8925	.8973	-.0022	.0026
[-.75	.20,	.30]	.6186	.6533	.7118	.0347	.0932
[.20	.60,	.77]	.7799	.7989	.6877	.0190	-.9022
[-.69	.30,	.40]	.5551	.5163	.4746	-.0388	-.0805

Table 10. Analysis of variance of values of E

Variation due to	d.f.	S.S.	M.S.	
		Experiment 2A		
Population (P)	38	4.2437	.1117	
Samples (S)	9	9.3076	1.0342	
P x S	342	9.0269	.0264	$s_{\bar{E}}^2 = .0026$
Total	389	22.5782		
		Experiment 2B		
Population	38	6.7964	.1789	
Samples in populations	351	13.4488	.0383	$s_{\bar{E}}^2 = .0038$
Total	389	20.2452		

It is necessary to digress briefly at this point to explain what is meant by admissible sets of the ρ_{ij} and to justify choosing the points indicated in the first column of Table 9 instead of those given by some more orderly response surface design. It is obvious that some sets of ρ_{ij} cannot be considered, since they lead to values of the multiple correlation coefficient outside the range zero to unity. The region of admissible values is defined by the restriction that any value of ρ_{1y} must lie in the range $\rho_{12}\rho_{2y} \pm (1 - \rho_{12}^2 - \rho_{2y}^2 + \rho_{12}^2\rho_{2y}^2)^{1/2}$. Geometrically, this region may be visualized as the space inside a circular cylinder with axis ρ_{1y}, which has been pinched closed at each end and twisted until the ends are at right angles to each other. An orthodox response surface design did not seem to be desirable at this point, since it would not have fitted into such an irregularly shaped space so as to provide points falling into extreme regions.

It was mentioned that among the populations considered, there were no sets having negative values for ρ_{1y} or ρ_{2y}. Such a selection was based on the symmetry which exists in the relationship of ρ_{ij} and R^2, the multiple correlation coefficient. If both ρ_{1y} and ρ_{2y} are negative, R^2 is the same as if they were positive. If ρ_{1y} and ρ_{2y} are of opposite sign, the value of R^2 is the same as if the sign of ρ_{12} were changed and the signs of ρ_{1y} and ρ_{2y} were considered the same. A similar sort of symmetry has been assumed to exist for the efficiencies. Some empirical evidence to indicate the validity of this assumption is offered in Table 11.

Table 11. Values of $\bar{E}$ and their standard errors at symmetrical points in the trivariate population space; censoring (55), N = 20, k = 10

Population [ρ_{12}	ρ_{1y},	ρ_{2y}]	$\bar{E}$	S.E.($\bar{E}$)
[-.20	.50,	.40]	.824	.090
[.20	.50,	-.40]	.869	.021
[-.75	.20,	.30]	.653	.093
[.75	.20,	-.30]	.635	.095
[.18	.30,	.50]	.897	.038
[.18	-.30,	-.50]	.940	.011
[-.55	.21,	.30]	.805	.094
[.55	.21,	-.30]	.815	.044
[-.55	-.21,	-.30]	.854	.054

Differences in mean efficiency of the magnitude indicated above can be explained on the basis of differential censoring. Recall that the dependent variable is not censored, whereas the independent variables are censored (55).

General second degree equations were fitted to the mean values of E indicated in Table 9 and resulted in the following equations to describe the response surfaces. For Experiment 2A,

$$\hat{E}_A = 1.1019 - .0047\rho_{12} - .3935\rho_{1y} - .0622\rho_{2y} - .4637\rho_{12}^2 - .0173\rho_{1y}^2 - .3608\rho_{2y}^2 - .0314\rho_{12}\rho_{1y} + .5711\rho_{12}\rho_{2y} + .3488\rho_{1y}\rho_{2y} \tag{4.8}$$

and for Experiment 2B

$$\hat{E}_B = 1.0566 - .2630\rho_{12} + .2271\rho_{1y} + .1049\rho_{2y} - .5965\rho_{12}^2 - .8344\rho_{1y}^2 - .7245\rho_{2y}^2 + .4691\rho_{12}\rho_{1y} + .9301\rho_{12}\rho_{2y} + .1519\rho_{1y}\rho_{2y} \tag{4.9}$$

Although the coefficients of these equations are quite different, the contours which they represent are very similar in the regions where the sampling was done. Figures 4 and 5 show contours for $\hat{E}_A$ and $\hat{E}_B$, respectively. The contour levels of $\hat{E}$ for .7, .8, and .9 are shown at three levels of ρ_{1y} in each diagram. The subspace of admissible ρ_{12} and ρ_{2y} is bounded by heavy lines and the shaded areas near these borders indicate that the quadratic models, $\hat{E}_A$ and $\hat{E}_B$ fit very poorly in these peripheral regions. For example, an additional population considered, [.65 .30, .90], gave values of $\bar{E}_A$ = .401 and $\bar{E}_B$ = .451. For the thirty-nine points fitted, $\hat{E}_A$ explained about eighty-two per cent of the variation in the values of $\bar{E}_A$, and $\hat{E}_B$ explained about seventy-eight per cent of the variation in the values of $\bar{E}_B$. In both cases the variation accounted for by the model was highly significant.

The seventy-eight points of the two experiments were combined to fit the single general quadratic model

$$\hat{E}_{A+B} = 1.0793 - .1338\rho_{12} - .0832\rho_{1y} + .0214\rho_{2y} - .5301\rho_{12}^2 - .4258\rho_{1y}^2 - .5431\rho_{2y}^2 + .2188\rho_{12}\rho_{1y} + .7506\rho_{12}\rho_{2y} + .2503\rho_{1y}\rho_{2y}. \tag{4.10}$$

The analysis of variance for this model is given in Table 12 following.

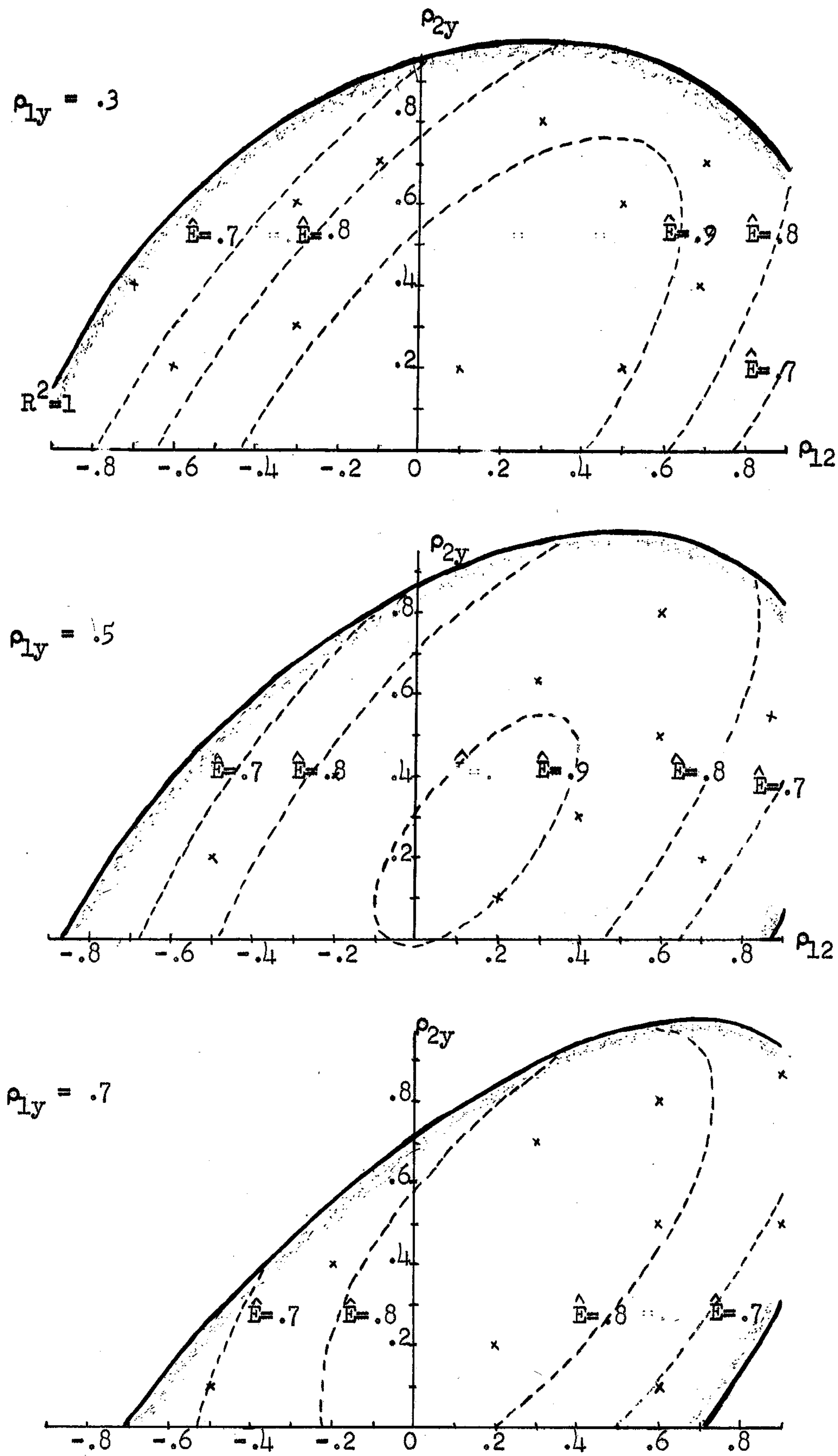


Figure 4. Contours for $\hat{E}_A$ in trivariate population space

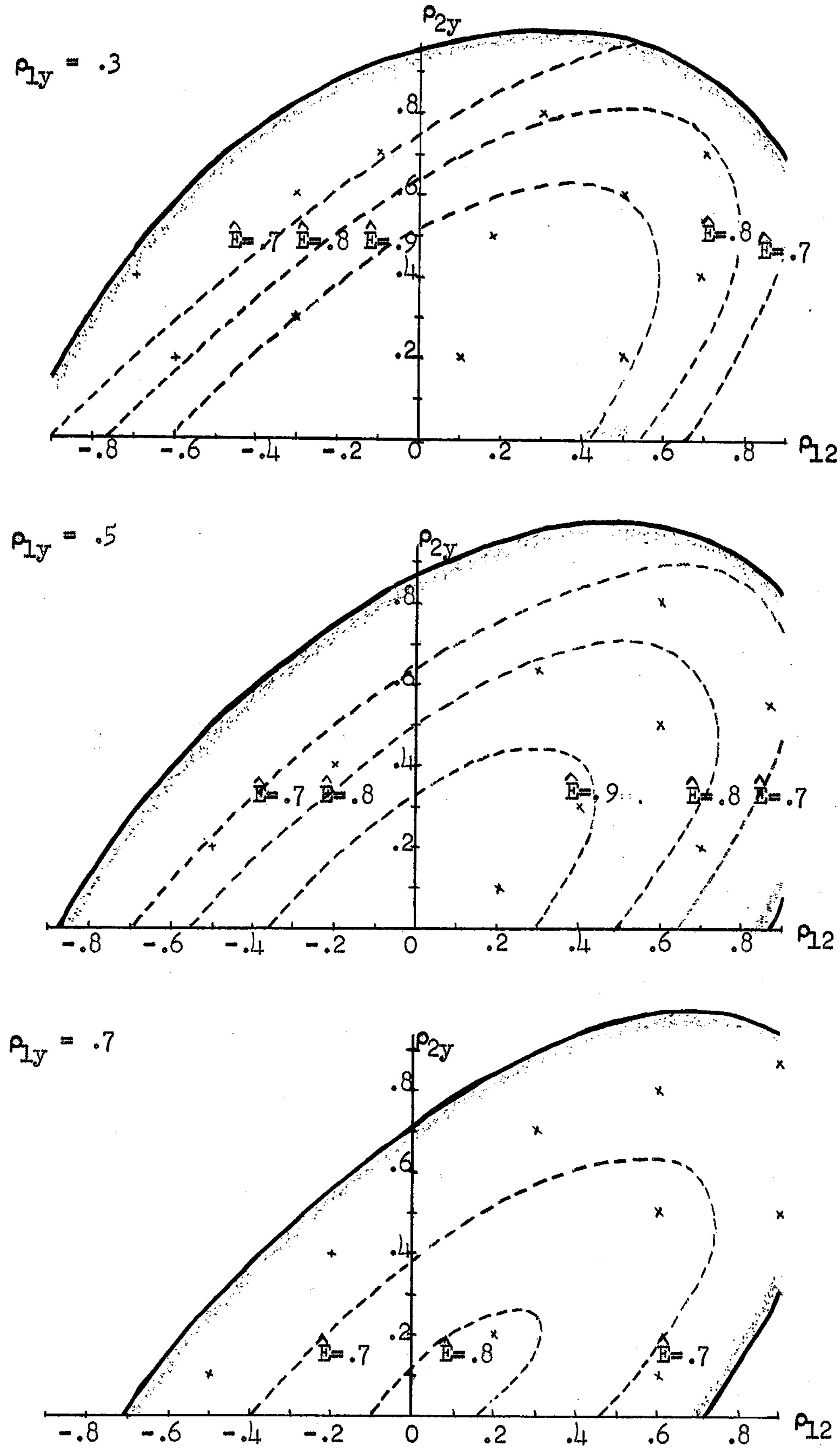


Figure 5. Contours for $\hat{E}_B$ in trivariate population space

Table 12. Analysis of variance for the model $\hat{E}_{A+B}$ of Experiment 2

Variation due to	d.f.	S.S.	M.S.	F
Model $\hat{E}_{A+B}$	9	.8339	.0927	14.24*
Lack of fit	29	.1887	.0065	3.10
Error[a]	39	.0813	.0021	
Total	77	1.1039		

[a] Failure of duplicates at the same points in Experiment 4A and 4B to agree.

Note the similarity between the estimates of experimental error for $\bar{E}$ given in Table 11 and that determined by duplicates at each point as given in Table 12. The residuals from the single general quadratic are given in Table 9 with the values of $\bar{E}$ at each point. These data and the contour diagrams of Figures 4 and 5 might reasonably be interpreted to indicate that this method of estimation is least suitable when applied to populations in which zero order correlation coefficients are of unlike magnitude or when their absolute values are large.

The approach used here to indicate how the type of population affects efficiency would be very cumbersome in considering more than three variates. In the four-variate case, six factors are required to characterize a population, hence, the fitting of a general second degree surface would require twenty-eight points, plus those needed to obtain an estimate of error.

An alternative approach to higher variate cases might consist of an extension of the results obtained for the trivariate case. In the four-variate case, for example, we could consider all possible subsets of three correlation coefficients as representing three-variate populations and could get some appraisal of expected efficiency for each of these. Then, if the results for the trivariate case held for four variates, some composite measure of these expected efficiencies would be highly correlated with the efficiency obtained for the four-variate population.

The simplification of this idea resorted to here consisted of considering the four-variate population $[\rho_{12}\ \ \rho_{13}\ \ \rho_{1y},\ \ \rho_{23}\ \ \rho_{2y},\ \ \rho_{3y}]$ in terms of the three trivariate populations made up of all sets of two (of the three) predictors with the predictand; specifically, $[\rho_{12}\ \ \rho_{1y},\ \ \rho_{2y}]$, $[\rho_{13}\ \ \rho_{1y},\ \ \rho_{3y}]$, and $[\rho_{23}\ \ \rho_{2y},\ \ \rho_{3y}]$. The contention is then that the efficiency of the four-variate population will be related to a combined measure of the efficiencies of the three subsets as determined by the kind of relationship indicated in equation (4.10) above.

Experiment 3 provided some data which, with this approach, confirms the extension of the results for the trivariate case to that for four variates. Table 13 gives the mean efficiencies for six four-variate populations, each censored (555) with the means based on twenty samples. The population correlation matrices for all possible subsets of two predictors with the predictand are indicated for each four-variate population. For these trivariate subset populations, the predicted efficiencies based on $\hat{E}_{A+B}$ were computed and are given in the last column of Table 13. The means of these $\hat{E}_{A+B}$, denoted avg $(\hat{E}_{A+B})$, and the $\bar{E}$ for each four-variate population may be seen plotted against each other in Figure 6.

Table 13. Values of $\bar{E}$ estimated by the Method of Paired Correlation Coefficients; Experiment 3, censoring (555) N = 20, k = 20

Population $[\rho_{12}$	ρ_{13}	ρ_{1y},	ρ_{23}	ρ_{2y},	$\rho_{3y}]$	$\bar{E}$	Possible subsets of two predictors and the predictand $[\rho_{ij}$	ρ_{iy},	$\rho_{jy}]$	$\hat{E}_{A+B}$ [a]
[.20	.29	.30,	.59	.40,	.51]	.757	.20	.20,	.40	1.008
							.29	.30,	.51	.971
							.59	.40,	.51	.913
[.10	.20	.40,	.48	.50,	.60]	.710	.10	.40,	.50	.930
							.20	.40,	.60	.915
							.48	.50,	.60	.906
[.60	.80	.29,	.50	-.20,	.20]	.686	.60	.29,	-.20	.778
							.80	.29,	.20	.741
							.50	-.20,	.20	1.026
[.20	.31	.60,	.25	.50,	.77]	.545	.20	.60,	.50	.879
							.31	.60,	.77	.814
							.25	.50,	.77	.827
[.80	.90	.30,	.70	.35,	.53]	.342	.80	.30,	.35	.800
							.90	.30,	.53	.788
							.70	.35,	.53	.882
[-.70	-.52	.32,	.60	.26,	.44]	.271	-.70	.32,	.26	.647
							-.52	.32,	.44	.674
							.60	.26,	.44	.923

[a] These values were determined by substituting the values of ρ_{ij}, ρ_{iy}, and ρ_{jy} into equation (4.10) as ρ_{12}, ρ_{1y}, and ρ_{2y}, respectively.

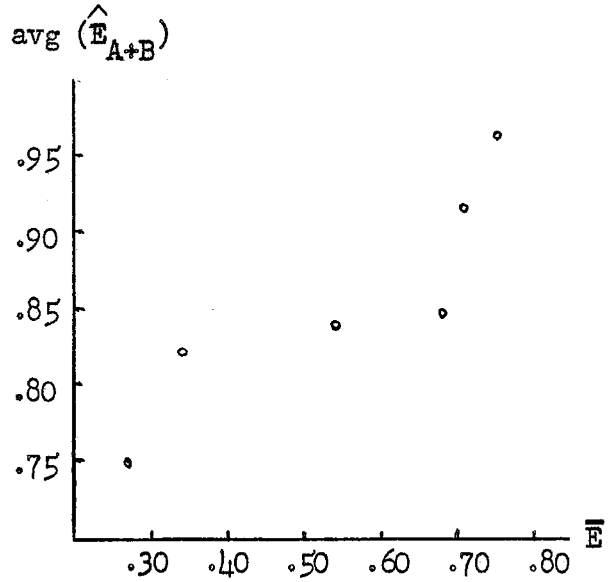


Figure 6. Associated values of $\bar{E}$ and avg ($\hat{E}_{A+B}$) from the data of Table 13. $R^2 = .78$

Although the evidence here is meager, it appears that the Paired Correlation Method in the four-variate case also yields poor prediction equations for populations with heterogeneously correlated variables. This hypothesis is further supported by the data from the five-variate cases of Experiment 6, which will be discussed presently, and by the data of Experiment 7, which will be summarized in Chapter VIII.

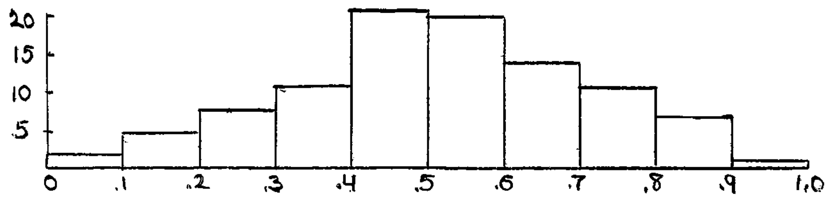
Experiment 6 involved use of the Paired Correlation Method with data simulated to represent two situations, one for which this method would be expected to yield reliable results, and another for which it would not.

A type of population which often arises in biological situations involving variates governed by feedback mechanisms is exemplified by [.60 .80 .29 .40, .50 -.20 .50, .20 .70, .30]. Because of the heterogeneous character of the correlation coefficients it would not seem advisable to apply the Paired Correlation Method to incomplete samples from such a population.

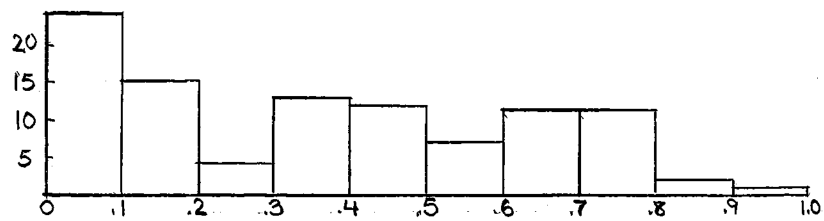
An example of the type of population frequently encountered in predicting from batteries of tests is [.36 .41 .40 .32, .24 .24 .30, .24 .25, .22]. This correlation matrix is, in fact, based on results of tests administered to trainees in attempting to predict their success in pilot training (Du Bois, 1958). Since, in this case, correlations between pairs of variates are similar, the Paired Correlation Coefficient Method would seem to be a satisfactory tool to use in getting a prediction equation where values among the independent variables are missing.

Two hundred samples, censored (3333), were generated to simulate one hundred from each of the two populations above. The Paired Correlation Method was then applied to each sample to determine a prediction equation. Histograms of the resulting values of E from these equations are shown in Figure 7, along with the efficiencies obtained by the use of complete observations. The results rather strikingly confirm the expectations stated above. The mean values of E and Δ from Experiment 6 and their respective standard errors are given in Tables 33A and 33B which are included among the summary results of Chapter VIII.

In the absence of <u>a priori</u> knowledge of the population from which a fragmentary sample is taken, it may be worthwhile to set confidence limits on the sample correlation coefficients from all existing pairs. This may be of help in deciding whether or not to use the method of this chapter in forming a prediction equation.

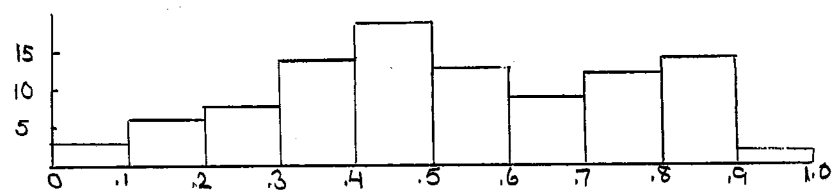


(A) Complete observations

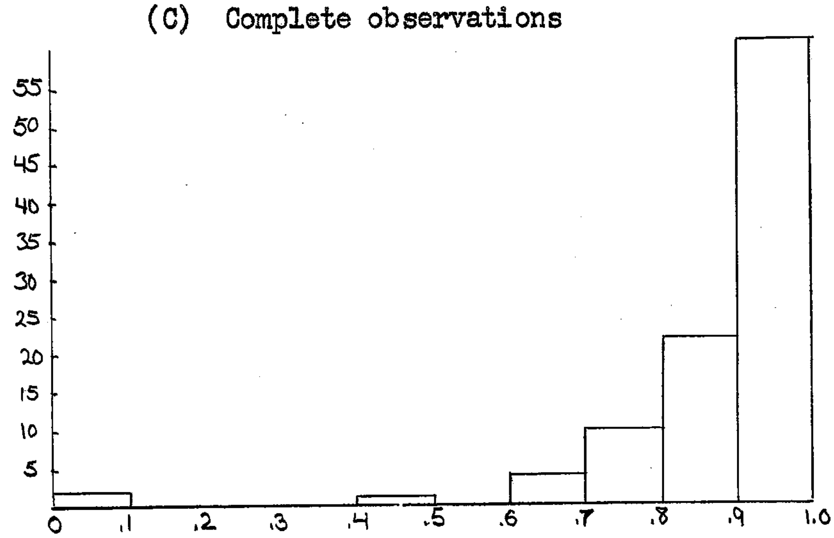


(B) Paired Correlation Method

[.60 .80 .29 .40, .50 -.20 .50, .20 .70, .30]



(C) Complete observations



(D) Paired Correlation Method

[.36 .41 .40 .32, .24 .24 .30, .24 .25, .22]

Figure 7. Histograms of E; Experiment 6
censoring (3333), N = 20, k = 100

4.5 Related Methods

It is well known that r is a biased estimate of ρ. Accordingly, one might substitute the approximately unbiased estimate,

$$\check{\rho} = r\,[1 + \frac{1}{2(N-1)}(1 - r^2)]\ ,$$

in the correlation matrix and then proceed to find the prediction equation as outlined in section 4.1.

A type of maximum likelihood estimate of ρ is found by maximizing $\ln f(r)$ with respect to ρ. It is approximately

$$\hat{\rho} = r\,[1 - \frac{1}{2(N-1)}(1 - r^2)]\ .$$

This estimate of ρ may also be used in forming the correlation matrix to be used in estimating the regression equation.

Since neither of these schemes gives regression equations consistent with the least squares estimate of the regression equation when no observations are missing, they have not been investigated here. However, it is interesting to note that in two populations where the method of paragraph 4.1 proved inefficient, neither of the alternatives provided a significant improvement. In the populations [-.69 .30, .40] and [.90 .70, .87] of Experiment 2A, the method of section 4.1 yielded mean efficiencies of .474 and .649, respectively. Use of the unbiased estimates of ρ gave mean efficiencies of .441 and .636, respectively, while substitution of the maximum likelihood estimates yielded efficiencies of .504 and .660, respectively.

CHAPTER V

THE GREENBERG PROCEDURE

5.1 Definition

The method to be discussed in this chapter was first suggested by Greenberg (Institute of Statistics, 1957) in connection with data from a perinatal mortality study in which hospital records were not complete in the detail desired. In a comprehensive analysis of these data, Wells (1959) defined and applied the method.

A more specific notation than that used in previous chapters will facilitate a description of the procedure. Suppose we assume the p variates, x_1, x_2, ..., $x_{(p-1)}$, y to be distributed multivariate normally. Consider a sample of N independent observation vectors, having a subset of n observations for which each of the p characteristics is measured; then (N - n) observations are incomplete for at least one x_i. The value of the i^{th} variate (i = 1, ..., p - 1) for the α^{th} observation (α = 1, ..., N) may be denoted as $x_{i\alpha}$.

The method suggested by Greenberg predicts y by the regression equation

$$^{m}Y = {}^{m}b_0 + {}^{m}b_1x_1 + \dots + {}^{m}b_{(p-1)}x_{(p-1)} \qquad (5.1)$$

which is determined as follows:

a. Estimate all missing characters by some reasonable initial value such as the sample mean of the particular variate from which the character comes. Having done this, one can proceed to find the ordinary least squares solution for predicting y. It may be written

$$ {}^{0}Y = {}^{0}b_0 + {}^{0}b_1 {}^{0}x_1 + {}^{0}b_2 {}^{0}x_2 + \dots + {}^{0}b_{(p-1)} {}^{0}x_{(p-1)} \qquad (5.2) $$

where the 0_{x_i} indicate that both observed values and initial estimates for missing characters were used to find 0_Y by least squares.

b. Using the same data, determine

$$ {}^{1}X_1 = {}^{1}b_{10} + {}^{1}b_{12} {}^{0}x_2 + {}^{1}b_{13} {}^{0}x_3 + \dots + {}^{1}b_{1(p-1)} {}^{0}x_{(p-1)} + {}^{1}b_{1y} y, \qquad (5.3) $$

the least squares equation obtained by taking x_1 as the dependent variable and the remaining x_i and y as independent variables.

Replace the previous estimates for missing values of x_1 by those given using equation (5.3); for example, to estimate the value of x_1 missing from the α^{th} observation,

$$ {}^{1}X_{1\alpha} = {}^{1}b_{10} + {}^{1}b_{12} {}^{0}x_{2\alpha} + {}^{1}b_{13} {}^{0}x_{3\alpha} + \dots + {}^{1}b_{1y} y_\alpha $$

c. With the data now including revised estimates for missing values of x_1, determine

$$ {}^{1}X_2 = {}^{1}b_{20} + {}^{1}b_{21} {}^{1}x_1 + {}^{1}b_{23} {}^{0}x_3 + \dots + {}^{1}b_{2(p-1)} {}^{0}x_{(p-1)} + {}^{1}b_{2y} y \qquad (5.4) $$

by ordinary least squares, where x_2 has been regarded as the dependent variable and the other variables as independent.

Replace the previous estimates of x_2 by those given using equation 5.4.

d. Continue to successively re-estimate the missing $x_{j\alpha}$ using

$$x_{j\alpha} = \sum_{i \neq j} {}^{k}b_{ji}\,{}^{h}x_{i\alpha} + b_{jy}y_{\alpha}$$

where ${}^{k}b_{ji}$ represents the coefficient of ${}^{h}x_{i}$ used in predicting x_j for the k^{th} iteration, and ${}^{h}x_{i}$ indicates either observed values of the x_i or those estimated on the h^{th} iteration by continuing the procedure outlined above ($h = k$ if $j > i$ and $h = k - 1$ if $j < i$). The successive estimating equations are:

$$
\begin{aligned}
{}^{1}X_{3} &= {}^{1}b_{30} + {}^{1}b_{31}\,{}^{1}x_{1} + {}^{1}b_{32}\,{}^{1}x_{2} + {}^{1}b_{34}\,{}^{0}x_{4} + \ldots + {}^{1}b_{3y}y \\
&\;\;\vdots \\
{}^{1}X_{(p-1)} &= {}^{1}b_{(p-1)0} + {}^{1}b_{(p-1)1}\,{}^{1}x_{1} + \ldots + {}^{1}b_{(p-1)(p-2)}\,{}^{1}x_{(p-2)} + {}^{1}b_{(p-1)y}y \\
{}^{1}Y &= {}^{1}b_{(p-1)0} + {}^{1}b_{(p-1)1}\,{}^{1}x_{1} + \ldots + {}^{1}b_{(p-1)}\,{}^{1}x_{(p-1)} \\
{}^{2}X_{1} &= {}^{2}b_{10} + {}^{2}b_{12}\,{}^{1}x_{2} + {}^{2}b_{13}\,{}^{1}x_{3} + \ldots + {}^{2}b_{1(p-1)}\,{}^{1}x_{(p-1)} + {}^{2}b_{1y}y \\
{}^{2}X_{2} &= {}^{2}b_{20} + {}^{2}b_{21}\,{}^{2}x_{1} + {}^{2}b_{23}\,{}^{1}x_{3} + \ldots + {}^{2}b_{2(p-1)}\,{}^{1}x_{(p-1)} + {}^{2}b_{2y}y \\
&\;\;\vdots \\
{}^{m}Y &= {}^{m}b_{0} + {}^{m}b_{1}\,{}^{m}x_{1} + \ldots + {}^{m}b_{(p-1)}\,{}^{m}x_{(p-1)}
\end{aligned}
$$

where m is the number of iterations required for convergence; convergence being defined by the condition ${}^{m}Y = {}^{m-1}Y$, (equation 5.1).

If characters are missing from one variate only, say x_{ji}, no iteration is required. In this case, missing entries among the x_j are initially estimated by

$$X_{j\alpha} = \sum_{i \neq j} b_{ji} x_{i\alpha} + b_{jy} y_{\alpha}$$

where the b_{ji} and b_{jy} are determined by ordinary least squares regression methods using the n complete observations, regarding x_j as the dependent variate. This condition is made clear by considering that when x_j is the dependent variate, the criterion used for the best fitting hyperplane (in terms of the remaining variates) consists of minimizing the sums of squares $\sum_{\alpha=1}^{n} (x_{j\alpha} - X_{j\alpha})^2$. Since no x_j exist for the N - n incomplete observation vectors, the hyperplane fitted to all N observations is identical to that for the n complete observations.

5.2 Bias and Variance of Regression Coefficients

Use of the Greenberg Procedure in Experiment 1 yielded biases, $\bar{b}_i - \beta_i$, and variances, $\hat{V}(b_i)$, for the various population-censoring combinations as given in Table 14. These data illustrate the tendency for both bias and variance of the regression coefficients to increase with increased censoring.

The bias in b_0 which, from Table 14, appears to be significant or nearly significant in most cases can probably be regarded as a sampling phenomenon. This statement appears reasonable when we recall that corresponding values of Table 1, based on complete observations, also yielded t tests indicating bias. It is known that those coefficients are theoretically unbiased. The remarks in Chapter III in regard to these apparently significant values of b_0 may be cited as applicable

Table 14. Statistics of regression coefficients estimated by the Greenberg Procedure; Experiment 1, N = 20, k = 20

Censoring	[.25 .50 , .25]						[.80 .60 , .80]					
	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$
(00)	.113*	.055	.051	.058	-.033	.025	.078*	.026	.060	.052	-.037	.032
(11)	.121*	.049	.082	.069	-.066	.036	.091*	.027	.009	.073	.032	.053
(22)	.124*	.047	.107	.081	.010	.041	.096*	.025	-.044	.104	.084	.068
(33)	.143*	.054	.102	.104	.029	.042	.112	.025	-.123	.133	.155*	.087
(44)	.159*	.058	.138	.126	.055	.093	.107*	.026	-.211*	.184	.240*	.136
(55)	.222*	.085	.159	.233	.057	.181	.111*	.035	-.300	.264	.303	.192
(66)	.286*	.160	.238	.573	.006	.539	.164*	.081	-.377*	.489	.350*	.359
(04)	.122*	.050	.045	.055	.029	.056	.091*	.027	-.062	.066	.115	.062
(40)	.132*	.060	.126	.124	-.005	.029	.084*	.027	-.006	.155	.026	.077
(08)	.095	.062	.022	.097	.016	.163	.080	.035	-.240*	.098	.284*	.086
(80)	.120	.072	.230	.326	.059	.052	.078	.029	.049	.352	-.037	.161

Censoring	[.25 .50 .75 , .25 .50 , .25][b]							
	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$	$\bar{b}_3-\beta_3$	$\hat{V}(b_3)$
(000)	-.104	.016	.026	.026	.011	.026	-.095	.019
(111)	-.171*	.014	-.037	.060	-.045	.034	-.054	.048
(222)	-.177*	.017	-.113	.073	-.073	.048	.040	.067
(333)	-.206	.029	-.095	.063	-.210	.189	.205	.156
(444)	-.166*	.015	-.222	.057	-.264	.214	.330	.298
(555)	-.491	.261	-.047	.568	-.756	1.348	.435	.774

[a] The symbol * indicates $\frac{\bar{b}_i - \beta_i}{s_{\bar{b}_i}} > t_{.05}, k - 1$, where $s^2_{\bar{b}_i} = \frac{\hat{V}(b_i)}{k}$, and k is the number of samples on which the b_i are based.

[b] For this four-variate case, k = 5.

here also. In addition, it may be noted that there are no significant t values for the biases, $\bar{b}_0 - \beta_0$, in Table 15, which shows the biases and variances of the regression coefficients from Experiment 4 where samples from several different populations were censored (55).

Significant biases in b_1 and b_2 are indicated in Table 14 at the higher levels of censoring in population [.80 .60, .80]. From Table 15, it can be seen that, although the Greenberg procedure does not generally produce biased regression coefficients, more than the expected number of coefficients, b_1 and b_2, in Experiment 4 fail to meet the criterion for unbiasedness. From the limited amount of data available, it is not possible to determine the exact nature of the bias. There does not seem to be a pronounced relationship between bias and population, as is indicated by the disparity between results obtained for the same populations in Experiments 4A and 4B.

In both Experiment 1 and Experiment 4, estimates of the regression coefficients were made using complete observations as well as by the Greenberg Procedure. It has been mentioned that coefficients determined in this way are unbiased and consequently provide a kind of standard with which one can compare the Greenberg Procedure. Except for the significant values of $\bar{b}_0 - \beta_0$, corresponding to those in Table 14, none of the biases of regression coefficients calculated from complete observations were found to be significant. This suggests that the Greenberg Procedure may produce biased estimates of β_1 and β_2.

Experiment 7 provides us with still more data which we may examine in commenting on the bias and variance of regression coefficients obtained by use of the Greenberg Procedure. The one hundred samples from the population [-.70 -.52 .32, .60 .26, .44], censored (333),

Table 15. **Statistics of regression coefficients estimated by the Greenberg Procedure; Experiment 4, censoring (55), N = 20, k = 10**

Population	Experiment 4A						Experiment 4B					
	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$
[-.2 .6 , .6]	.069	.016	.015	.037	.116	.440	.043	.014	.063	.011	-.013	.011
[-.2 .6 , -.2]	.101	.068	.072	.107	.120	.345	.133	.116	.403*	.179	-.245	.496
[-.2 -.2 , -.2]	.090	.088	-.251	.457	.035	.536	-.036	.060	-.116	.128	-.179	.277
[-.2 -.2 , -.6]	.148	.081	-.159	.314	.180	1.303	-.078	.090	.022	.255	.262	.142
[.6 .6 , .6]	.146	.064	-.254	.449	.120	.528	.033	.036	.002	.239	.001	.292
[.6 .6 , -.2]	.030	.008	.080	.107	-.198*	.072	.002	.010	.130*	.014	-.166*	.021
[.6 -.2 , -.2]	.128	.127	-.299	.665	.210	1.037	-.068	.165	.029	.562	-.071	.748
[.6 -.2 , .6]	.076	.023	-.259*	.032	.233*	.063	.043	.013	-.138	.054	.085	.079
[.873 .2 , .2]	.165	.159	-.574	2.763	.610	3.221	.174	.239	.043	3.059	-.117	3.677
[-.473 .2 , .2]	.147	.114	.008	.503	.300	.539	.037	.073	.307	.380	.279	.278
[.2 .873, .2]	.072	.036	.052	.028	.028	.209	.007	.017	.042	.034	-.062	.074
[.2 -.473, .2]	.098	.091	-.366*	.179	.230	.229	-.033	.067	-.112	.105	-.004	.154
[.2 .2 , .873]	.102	.031	-.140	.108	.108*	.036	.092	.022	-.068	.092	.106	.033
[.2 .2 , -.473]	.077	.045	-.195	.374	-.164	.219	-.018	.019	.246*	.041	-.225*	.063
[.2 .2 , .2]	.171	.074	-.133	.369	.100	.413	.053	.040	.185	.873	.380	.970

* Indicates $\dfrac{\bar{b}_i - \beta_i}{s_{\bar{b}_i}} > t_{.05,\ 9\ d.f.}$ where $s^2_{\bar{b}_i} = \dfrac{\hat{V}(b_i)}{10}$

yielded estimates of regression coefficients, the biases and variances of which are shown in the third column of Table 16. Histograms of the coefficients obtained using the Greenberg Procedure are given in Figure 8. For this particular case, with evidence available from a large number of samples, all the regression coefficients except b_0 are clearly shown to be biased.

The tabulated variances of the regression coefficients call for only brief comment. In Tables 14 and 15 marked differences may be noted in the variances for different populations, just as one would expect. For example, the variances in the regression coefficients are large for population [.8728 .2, 2], which has a large covariance between the independent variables and large error of estimate. On the other hand, the variances of regression coefficients are small for population [-.2 .6, .6], where the covariance between the independent variables and the error of estimate are both small. The tendency for the $\hat{V}(b_i)$ to increase with increased censoring in Table 14 was also noted in discussing similar data calculated by other methods.

5.3 Estimate of Error

An estimate of error, $s^{2\prime}_{y.x}$ may be calculated in the usual way after estimating the missing values and determining the regression equation by the Greenberg Procedure. Such a statistic grossly underestimates the true variance, since the degrees of freedom used in its computation are based on some values estimated from the data.

In similar cases it has been possible to estimate "effective degrees of freedom" based on the comparison of like moments. This approach is not applicable here, however, since the mathematical form

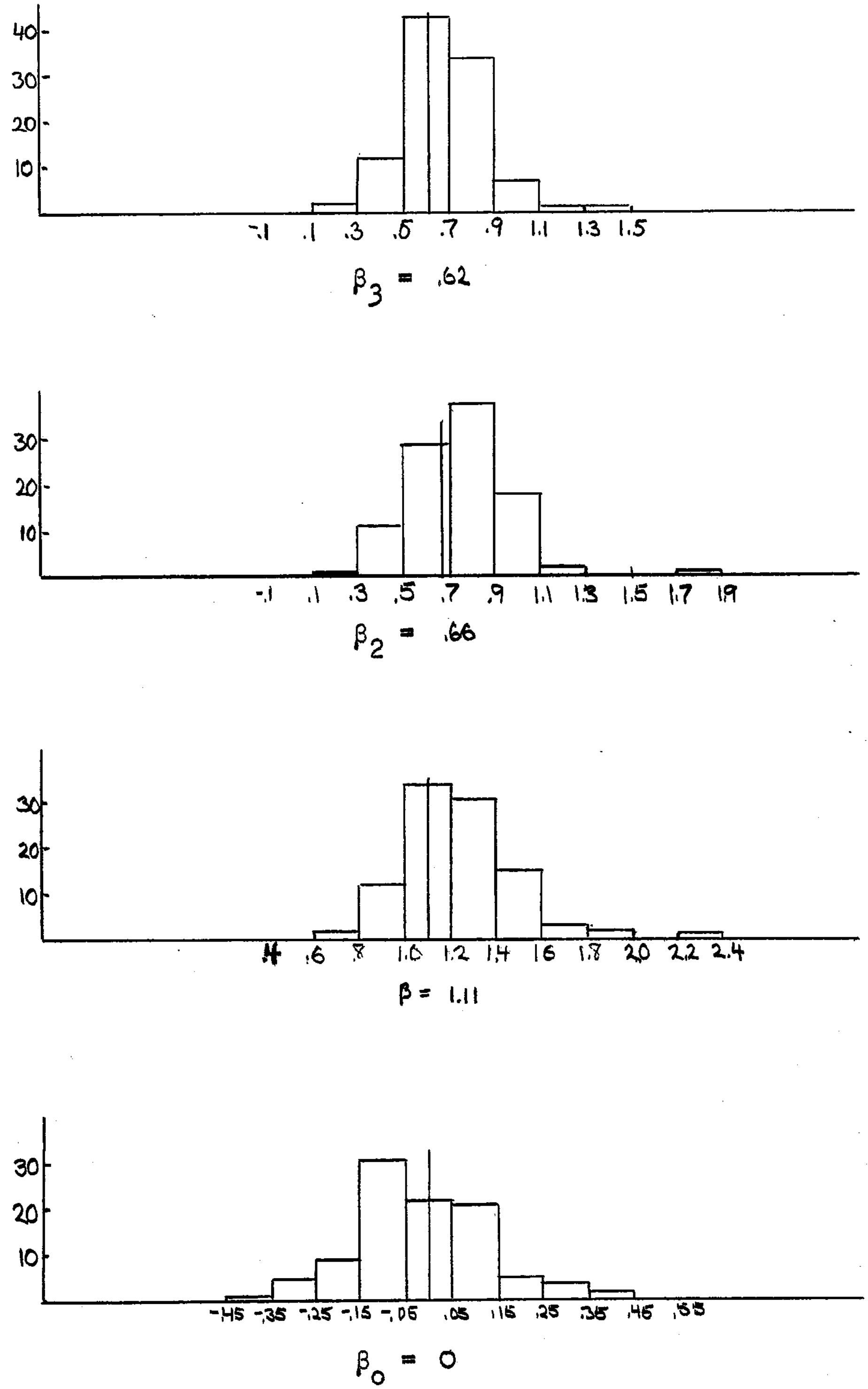


Figure 8 Histograms of regression coefficients estimated by the Greenberg Procedure; population [-.70 -.52 .32, .60 .26, .44], censoring (333), N = 20, k = 100

Table 16. Statistics of regression coefficients from Experiment 7; population [-.70 -.52 .32, .60 .26, .44], censoring (333), N = 20, k = 100

Statistic	Estimated by			
	Complete Observations	Paired Correlation	Greenberg Procedure	Uncensored Data
$\bar{b}_0 - \beta_0$	-.009	.048	-.014	-.007
$\hat{V}(b_0)$	.030	.242	.023	.012
$\frac{\bar{b}_0 - \beta_0}{s_{\bar{b}_0}}$	-.521	.970	-.940	-.644
$\bar{b}_1 - \beta_1$	.022	.177	.132	.001
$\hat{V}(b_1)$	.061	.706	.063	.035
$\frac{\bar{b}_1 - \beta_1}{s_{\bar{b}_0}}$	.886	2.109	5.270	.050
$\bar{b}_2 - \beta_2$	-.002	.212	.084	-.002
$\hat{V}(b_2)$	.049	2.564	.046	.028
$\frac{\bar{b}_2 - \beta_2}{s_{\bar{b}_2}}$	-.068	1.324	3.887	-.100
$\bar{b}_3 - \beta_3$	.018	-.087	.063	.010
$\hat{V}(b_3)$	.038	1.300	.032	.020
$\frac{\bar{b}_3 - \beta_3}{s_{\bar{b}_3}}$	.904	-.764	3.565	.696

of the distribution or its moments is not known. A simple rule of thumb sufficient to produce reasonable estimates of $\sigma^2_{y.x}$ is given by defining "effective degrees of freedom" for error as

$$f_e = N[1 - \frac{c}{N(p-1)}] - p = N - p - \frac{c}{p-1} . \qquad (5.5)$$

where c = the total number of missing variate values among the p-1 independent variates. The resulting "pseudo-variance" then is

$$s^{2*}_{y.x} = \frac{(N-p)\ s^{2'}_{y.x}}{N-p - \frac{c}{p-1}} \qquad (5.6)$$

where $s^{2'}_{y.x}$ is calculated in the usual way assuming that the estimated characters were observed values. The rationale for choosing $N - p - \frac{c}{p-1}$ as "effective degrees of freedom" is that the proportion of the independent variate values estimated is $\frac{c}{N(p-1)}$.

To illustrate the effectiveness of equation (5.6) as a measure of variance, we will consider the "pseudo variances" arising from the use of the Greenberg Procedure in Experiment 4B. The mean of the "pseudo-variances" for each of the fifteen populations is plotted against the true variance, $\sigma^2_{y.x}$, in Figure 9A. A weighted least squares line, assumed to go through the origin, was fitted to these points and gave $\sigma^2_{y.x} = 1.05\ s^{2*}_{y.x}$. The slope is not significantly different from unity although values of $s^{2*}_{y.x}$ still appear to be underestimated for low values of $\sigma^2_{y.x}$. Differences between $\sigma^2_{y.x}$ and $s^{*}_{y.x}$ can be accounted for to some extent by sampling variation. This is evinced by comparing Figure 9A with Figure 9B, where the values of $s^2_{y.x}$, calculated from the uncensored samples, are plotted against the values of $\sigma^2_{y.x}$.

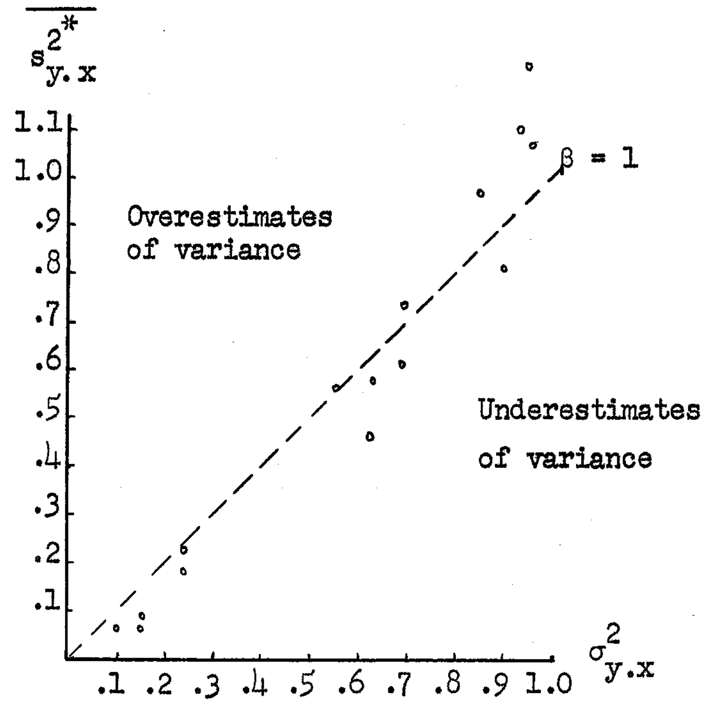


Figure 9A. Mean pseudo-variance of Greenberg Procedure plotted against the true variance; from data of Experiment 4B

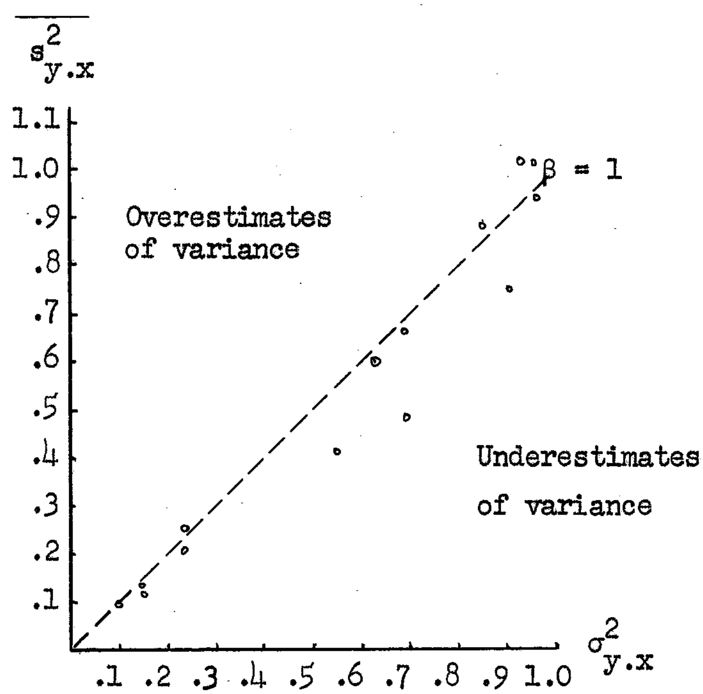
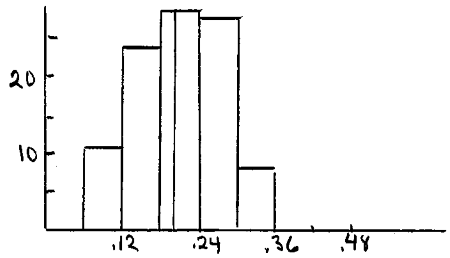


Figure 9B. Mean variance estimated from uncensored data plotted against true variance; from data of Experiment 4B

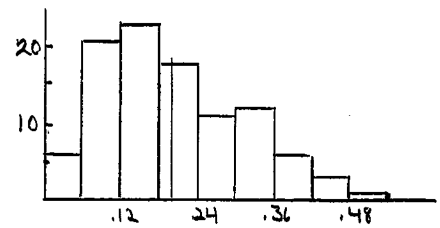
Additional evidence concerning the use of $s^{2*}_{y.x}$ as a measure of residual error is given by considering its sampling distribution for the data from population [-.70 -.52 .32, .60 .26, .44] when samples are censored (333). Figure 10C shows a histogram for values of $s^{2*}_{y.x}$ based on one hundred such samples. Figure 10A represents the sampling distribution for errors of estimate from corresponding uncensored samples and Figure 10B shows the sampling distribution for errors of estimate based on complete observations. Since $f_e = 13$ here, the histogram of Figure 10C would be expected to assume a form intermediate in shape between that given by Figure 10A (theoretically proportional to χ^2 with 16 degrees of freedom) and that shown in Figure 10B (theoretically proportional to χ^2 with 7 degrees of freedom). Thus, it can be seen that $s^{2*}_{y.x}$ underestimates $\sigma^2_{y.x}$ here. This observation supports a statement made in the preceding paragraph; namely, that $s^{2*}_{y.x}$ appears to underestimate $\sigma^2_{y.x}$ for low values of $\sigma^2_{y.x}$. For the four-variate population from which the one hundred samples were selected, $\sigma^2_{y.x} = .20$. It must be pointed out that this apparent underestimation may not occur for higher values of $\sigma^2_{y.x}$.

5.4 Efficiency of Prediction Equations

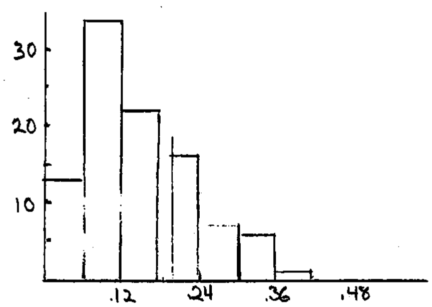
Table 17A gives the means of E and their standard errors for the various population-censoring schemes of Experiment 1. Table 17B shows the corresponding means of Δ and their standard errors. There does not appear to be a population effect influencing the efficiencies here, as was the case with the Paired Correlation Coefficient Method. A further evaluation of the Greenberg Procedure was considered desirable, however, so Experiment 4 was set up to evaluate the method over a wider range of



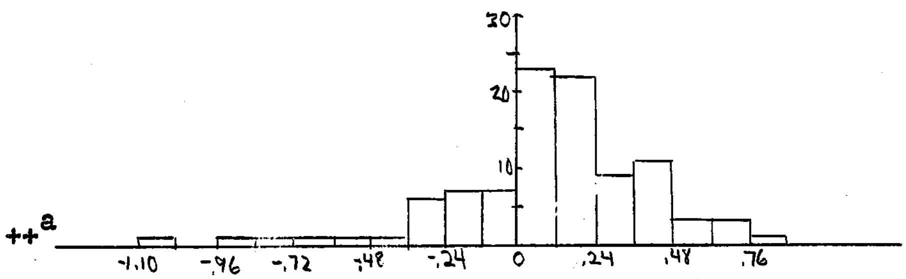
(A) No Censoring



(B) Complete Observations



(C) Pseudo Variance Greenberg Procedure



(D) Paired Correlation Coefficients

Figure 10 Histograms of estimated error variances; Experiment 7 [-.70 -.50 .32, .60 .26, .44], censoring (333), N = 20, k = 100

---

[a] The symbol + indicates a value outside the range of scale.

Table 17A. Values of $\bar{E}$ and their standard errors estimated by the Greenberg Procedure; Experiment 1, N = 20, k = 20

Censoring	[.25 .50, .25]		[.80 .60, .80]		Censoring	[.25 .50 .75, .25 .50, .25]	
	$\bar{E}$	S.E.($\bar{E}$)	$\bar{E}$	S.E.($\bar{E}$)		$\bar{E}$	S.E.($\bar{E}$)
(11)	.976	.009	.978	.007	(111)	.949	.018
(22)	.956	.011	.960	.011	(222)	.858	.032
(33)	.939	.013	.910	.021	(333)	.736	.129
(44)	.894	.023	.836	.030	(444)	.627	.136
(55)	.810	.041	.764	.041	(555)	.445	.159
(66)	.614	.047	.664	.051			
(04)	.967	.009	.944	.014			
(40)	.956	.012	.953	.012			
(08)	.875	.036	.845	.027			
(80)	.825	.032	.860	.033			

Table 17B. Values of $\bar{\Delta}$ and their standard errors estimated by the Greenberg Procedure; Experiment 1, N = 20, k = 20

Censoring	[.25 .50, .25]		[.80 .60, .80]		Censoring	[.25 .50 .75, .25 .50, .25]	
	$\bar{\Delta}$	S.E.($\bar{\Delta}$)	$\bar{\Delta}$	S.E.($\bar{\Delta}$)		$\bar{\Delta}$	S.E.($\bar{\Delta}$)
(00)	.303	.028	.211	.020	(000)	.222	.012
(11)	.330	.028	.239	.020	(111)	.324	.017
(22)	.351	.028	.256	.022	(222)	.365	.019
(33)	.386	.033	.291	.024	(333)	.482	.097
(44)	.447	.036	.343	.024	(444)	.576	.117
(55)	.564	.055	.401	.029	(555)	1.114	.350
(66)	.833	.099	.518	.062			
(04)	.347	.027	.253	.019			
(40)	.379	.036	.262	.026			
(08)	.431	.037	.332	.021			
(80)	.430	.069	.345	.044			

populations and a larger set of random samples. Since the Greenberg Procedure is quite time consuming, only a limited number of populations could be sampled. And, just as in Chapter IV, the coordinates of the points were required to be admissible sets of ρ_{ij}. The configuration of points chosen was a central composite type of response surface design selected so that the extreme points would fall within the region of admissible values. As mentioned in Chapter II, Experiment 4 was composed of two parts; part A, based on samples from different populations generated from the same set of original deviates, and part B, based on different sets of original samples at each point. The points chosen and the means of efficiencies with their standard errors at these points are shown in Table 18. Mean efficiency at each point was based on ten samples.

Table 18. Values of $\bar{\bar{E}}$ and their standard errors estimated by the Greenberg Procedure; Experiment 4, censoring (55), N = 20, k = 10

Population	Experiment 4A $\bar{\bar{E}}$	Experiment 4A S.E.($\bar{\bar{E}}$)	Experiment 4B $\bar{\bar{E}}$	Experiment 4B S.E.($\bar{\bar{E}}$)
[-.2 .6 , .6]	.710	.074	.810	.044
[-.2 .6 , -.2]	.772	.067	.760	.081
[-.2 -.2 , -.2]	.716	.083	.808	.039
[-.2 -.2 , .6]	.720	.081	.820	.053
[.6 .6 , .6]	.750	.086	.820	.061
[.6 .6 , -.2]	.716	.056	.837	.032
[.6 -.2 , -.2]	.765	.080	.783	.042
[.6 -.2 , .6]	.714	.069	.769	.040
[.8728 .2 , .2]	.711	.082	.727	.085
[-.4728 .2 , .2]	.712	.084	.812	.037
[.2 .8728, .2]	.692	.073	.840	.039
[.2 -.4728, .2]	.774	.064	.830	.035
[.2 .2 , .8728]	.764	.060	.798	.059
[.2 .2 , -.4728]	.765	.061	.822	.035
[.2 .2 , .2]	.777	.072	.727	.085

The use of the same set of original samples at each point in Experiment 4A has the same effect, although not as marked, as that which was noted in Experiment 2. This effect is that the mean square error for part A becomes smaller than that for part B after differences among samples have been accounted for in the analysis of variance of E. This outcome is seen in Table 19 below.

Table 19. Analyses of variance for values of E from the data of Experiment 4.

Variation due to	d.f.	S.S.	M.S.	
		Experiment 4A		
Population (P)	14	.1246	.0089	
Samples (S)	9	4.1677	.4631	
P x S	126	3.1720	.0252	s_E^2 = .0025
Total	149	7.4643		
		Experiment 4B		
Population	14	.1886	.0135	
Samples in populations	135	3.9861	.0295	s_E^2 = .0030
Total	149	4.1747		

The general second degree equation fitted to the data of Experiment 4A yielded the response surface described by

$$\hat{E}_A = .7718 + .0658\rho_{12} + .0268\rho_{1y} - .0113\rho_{2y} - .1458\rho_{12}^2 - .0982\rho_{1y}^2 - .0287\rho_{2y}^2 - .0468\rho_{12}\rho_{1y} + .0315\rho_{12}\rho_{2y} + .0134\rho_{1y}\rho_{2y} \qquad (5.5)$$

Experiment 4B gave

$$\hat{E}_B = .7538 - .0668\rho_{12} - .1088\rho_{1y} - .0601\rho_{2y} + .0792\rho_{12}^2 + .2239\rho_{1y}^2$$
$$+ .1687\rho_{2y}^2 - .1278\rho_{12}\rho_{1y} - .0716\rho_{12}\rho_{2y} + .0273\rho_{1y}\rho_{2y} \quad (5.6)$$

The analyses of variance for these fitted surfaces are given in Table 20.

Table 20. Analyses of variance for models fitted in Experiment 4.

Variation due to	d.f.	S.S.	M.S.	F
		Experiment 4A		
Model $\hat{E}_A$	9	.0065	.0007	< 1
Error	5	.0059	.0012	
Total	14	.0124		
		Experiment 4B		
Model $\hat{E}_B$	9	.0156	.0017	2.65 N.S.
Error	5	.0033	.0007	
Total	14	.0189		

The variation accounted for by the models $\hat{E}_A$ and $\hat{E}_B$ is not significant. Thus, it does not appear that, in trivariate cases, the population sampled affects the efficiency of prediction equations estimated by the Greenberg Procedure. These analyses confirm the findings from Experiment 1.

Experiment 7, which was mentioned in section 5.2, lends some additional information on the efficiency of the Greenberg Procedure. From the four-variate population [-.70 -.52 .32, .60 .26, .44], one hundred samples, censored (333), were selected and regression equations estimated by various methods. Histograms of the resulting sampling distributions of E and Δ are shown in Figures 11C and 12C respectively. This particular case exemplifies the sort of missing data situation to which the Greenberg Procedure might be advantageously applied if there were some basis for not wanting to discard the fragmentary observations.

The computing time required in using the Greenberg Procedure for higher variate cases precluded extensive sampling work in this area. However, it was necessary to ascertain that such computations were possible. In addition to this, the results of the four-variate work in Experiment 7 were encouraging enough to warrant applying the Greenberg Procedure to some of the five-variate samples of Experiment 6. Consequently, Experiment 8 was carried out with twenty samples, the results of which gave the efficiencies E and Δ shown in Table 21. The values of the statistics given by using the other methods are also presented for comparison. In population [.60 .80 .29 .40, 50 -.20 .50, .20 .70, .30], the Greenberg Procedure yielded better results than were obtained with the other two methods. This was true for eight of the ten samples considered in comparing it with the use of complete observations, and for nine of the ten samples considered in comparing it with the Paired Correlation Method. In population [.36 .40 .41 .32, .24 .24 .30, .24 .25, .22], there appeared to be little difference between the Greenberg Procedure and the use of complete observations, while the Paired Correlation Method yielded significantly better results than either of these.

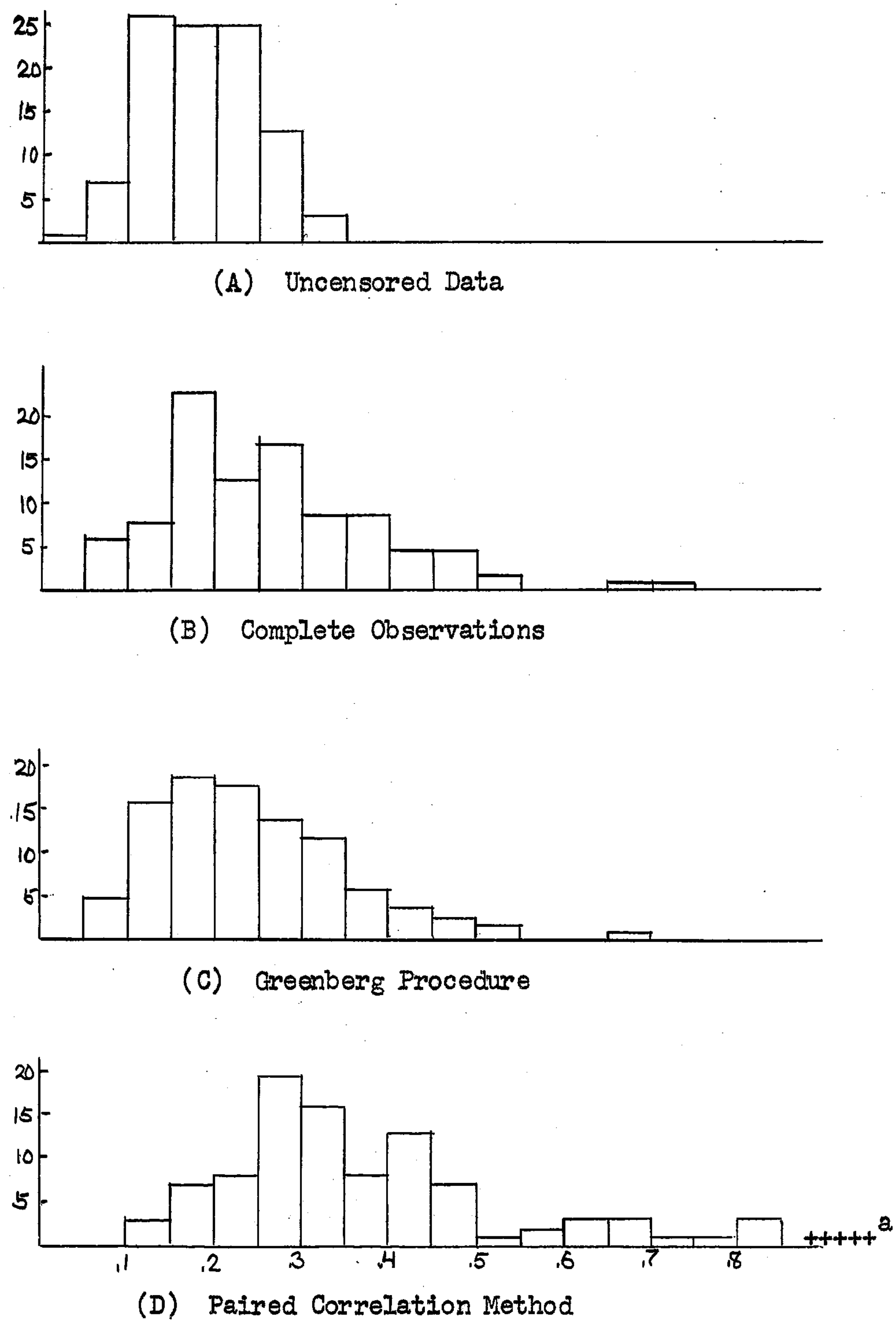
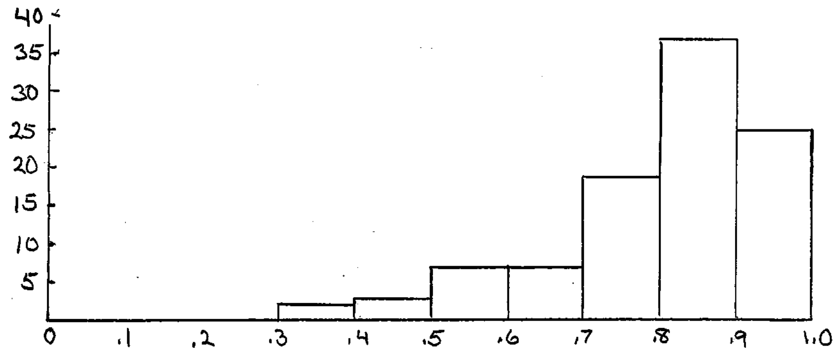
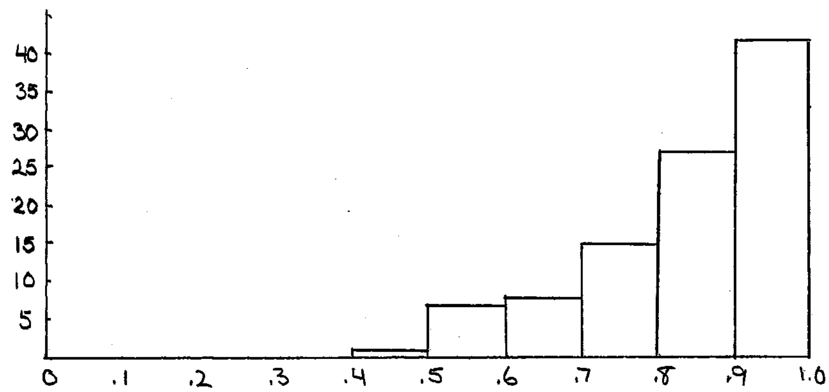


Figure 12. Histograms of Δ; Experiment 7
population [-.70 -.52 .32; 60 .26, .44]
censoring (333), N = 20, k = 100

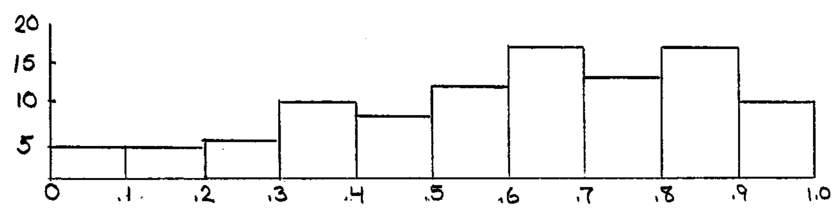
[a] The symbol + indicates a value outside the range of scale.



(A) Complete observations



(B) Greenberg Procedure



(C) Paired Correlation Method

Figure 11. Histograms of E; Experiment 7.
population [-.70 -.52 .32, .60 .26, .44]
censoring (333), N = 20, k = 100

Table 21. Values of E and Δ from the twenty samples of Experiment 8; censoring (3333), N = 20

Sample Number	E: Complete Observations	E: Paired Correlation	E: Greenberg Procedure	Δ: Complete Observations	Δ: Paired Correlation	Δ: Greenberg Procedure	Δ: Uncensored Data
Population [.60 .80 .29 .40, .50 -.20 .50, .20 .70, .30]							
91	.681	.487	.898	.211	.381	.158	.119
92	.817	.319	.356	.358	.676	.490	.325
93	.634	.094	.756	.283	1.414	.250	.152
94	.041	.676	.743	2.144	.464	.405	.227
95	.168	.835	.317	.940	.228	.510	.175
96	.415	.165	.475	.496	.721	.379	.195
97	.704	.089	.835	.405	.856	.378	.301
98	.496	.587	.640	.535	.388	.446	.199
99	.509	.753	.870	.409	.303	.129	.188
100	.612	.391	.589	.397	.533	.410	.179
Mean	.508	.440	.648	.618	.596	.356	.206
Std. Error	.077	.086	.066	.181	.110	.042	.020
Population [.36 .40 .41 .32, .24 .24 .30, .24 .25, .22]							
105	.865	.871	.802	.897	.935	1.093	.803
106	.036	.936	.868	3.297	.282	.291	.331
107	.773	.869	.746	.448	.312	.450	.333
108	.539	.878	.981	1.096	.555	.625	.538
109	.878	.950	.852	.794	.848	1.009	.721
110	.796	.774	.326	.546	.705	1.424	.244
111	.270	.707	.386	1.258	.357	1.173	.390
112	.726	.974	.353	.655	.398	1.112	.331
113	.710	.953	.781	.644	.423	.632	.354
114	.423	.978	.549	1.223	.247	1.000	.214
Mean	.601	.889	.664	1.086	.506	.881	.426
Std. Error	.089	.028	.076	.261	.077	.114	.063

5.5 Convergence

In describing the Greenberg Procedure, convergence was defined by the condition that the respective b_i of the prediction equation be equal on two successive iterations. Except for differences caused by rounding, this is tantamount to the requirement that successive estimates of missing values be equal. In view of the fact that several of these estimates must converge simultaneously, it is not surprising that a large number of iterations are required in certain cases.

In Experiment 1, the Greenberg Procedure was applied to twenty random samples from each of two trivariate populations which were censored (11), (22), (33), (44), (55), (66), respectively. The mean number of iterations required for each of these censoring schemes in the order indicated is 5.0, 6.7, 10.3, 13.6, 17.1, and 24.2, respectively. An analysis of variance for these data indicated that censoring was the outstanding factor affecting speed of convergence.

The same sort of relationship between intensity of censoring and number of iterations follows for higher variate cases. For example, in the four-variate case of Experiment 1 with censoring (222), which is comparable to (33) above, the mean number of iterations required for convergence was 11.8. Likewise, in censoring (444), comparable to (66) above, an average of 28.4 iterations were required.

In Experiment 8, five-variate samples were generated from each of two populations. Ten different original samples, censored (3333), were evaluated for each. The mean number of iterations per sample required for convergence was 27.6 for one population, and 23.7 for the other. The standard errors were 8.9 and 8.2, respectively.

A reasonable question to ask at this point is whether or not the number of iterations required for convergence depends more on the number or the proportion of observations missing. An answer to this question is indicated by the results from Experiment 5, a 2 x 3 factorial involving samples of varying size from the population [.6 -.2, .6]. Estimation using the Greenberg Procedure was carried out, using five randomly selected samples for each combination, where 6, 10, and 16 values were censored representing proportions of twenty and fifty percent of the total number of observation vectors. The sample size, N, and the mean number of iterations, $\bar{k}$, required for convergence is shown in Table 22 below. The time required per iteration is also given for the various sample sizes.

Table 22. Mean number of iterations required for convergence[a] in sampling from [.6 -.2, .6]; k = 5

	Proportion of Observation Vectors Censored							
	.2				.5			
Censoring	$\bar{m}$	s_m	N	seconds per iteration	$\bar{m}$	s_m	N	seconds per iteration
(33)	5.2	1.30	30	37	18.0	5.79	12	22
(55)	4.4	.55	50	54	10.3	1.34	20	31
(88)	4.4	.55	80	79	10.6	2.79	32	43

[a] In this experiment the criterion for convergence was taken as agreement of the respective regression coefficients to three decimal places on successive iterations.

The most striking feature of Table 22 is that the number of iterations required seems to depend almost entirely on the proportion of data missing, rather than the amount. It should be noted, however, that for both proportions, a trifle more iteration was necessary at the lower levels than at the higher levels. Although this phenomenon could easily be explained by sampling variation (note the s_k), one is led to speculate that a certain amount of complete data is necessary before the rate of convergence for a given proportion becomes constant.

The data from Experiment 1 indicate that the difference in speed of convergence between populations is very small. Experiment 4B affords more illuminating data in this connection. In that experiment, the Greenberg Procedure was applied to 150 samples of independent normal deviates grouped into fifteen sets, each of which was used to generate ten multivariate normal samples representing a different population. The number of iterations required for convergence in these 150 cases was used to test the hypothesis that population affects the number of iterations necessary for convergence. The analysis of variance shown in Table 23 reveals the unimportance of population as an effect influencing convergence.

Table 23. Analysis of variance for number of iterations required for convergence in Experiment 4B; $N = 20$, $k = 10$

Variation due to	d.f.	S.S.	M.S.	F
Population	14	871.4	62.2	1.0
Samples in populations	135	8421.4	62.4	
Total	149	9292.8		

The initial estimates of missing values also seem to be of relatively minor importance. In section 5.1, it was stated that missing characters would initially be estimated by some "reasonable" value. Two kinds of initial estimates were tried in a preliminary investigation using trivariate samples of size twenty from Carver's Anthropometric Data (1954). In each of seven samples, ten of the forty characters among the independent variates were censored. The Greenberg Procedure was carried out, using as initial estimates the sample variate mean; zero; and a bivariate regression estimate, $X_{i\alpha} = a + by_{\alpha}$, where a and b were estimated by least squares, using all available pairs of x_i and y. The number of iterations required using these initial estimates is given in Table 24.

Table 24. Iterations required for convergence using the Greenberg Procedure on Carver's data with different initial estimates of missing values.

Sample Number	Initial estimates by: Sample variate means	Bivariate regression	Zero
1	11	10	13
2	14	11	21
3	6	5	22
4	5	5	22
5	5	8	35
6	7	11	24
7	4	4	49

Zero as an initial estimate is clearly inferior. Since the true means of the variates, x_1 and x_2, were 139 and 282, respectively, the use of zero as an initial estimate could hardly be regarded as reasonable. There is little basis for preference between the sample variate mean and the bivariate regression estimate as initial values, except that the variate means are simpler to calculate.

In using the Greenberg Procedure with artificially generated samples, initial estimates were made using the true mean, the true regression equation, and the random character actually designated as missing. Obviously such estimates could not be used in practice, but their use shortened convergence time somewhat in the sampling work.

In discussing convergence time, we have talked in terms of the <u>number</u> of iterations required for convergence. It must also be remembered that the time <u>per iteration</u> increases as the number of variates increase. In adding an additional variate, another matrix must be inverted with each iteration, and the order of the matrices to be inverted increases by one. Inversion time is roughly proportional to the square of the dimension of the matrix involved. The time required for one iteration by the IBM 650, with N = 20, is approximately thirty seconds for the three-variate case, seventy seconds for the four-variate case, and about one hundred and fifty seconds for the five-variate case.

The distribution of the number of iterations is obviously skewed and is indicated by the histogram of Figure 13 for the work done using the Greenberg Procedure in Experiment 7.

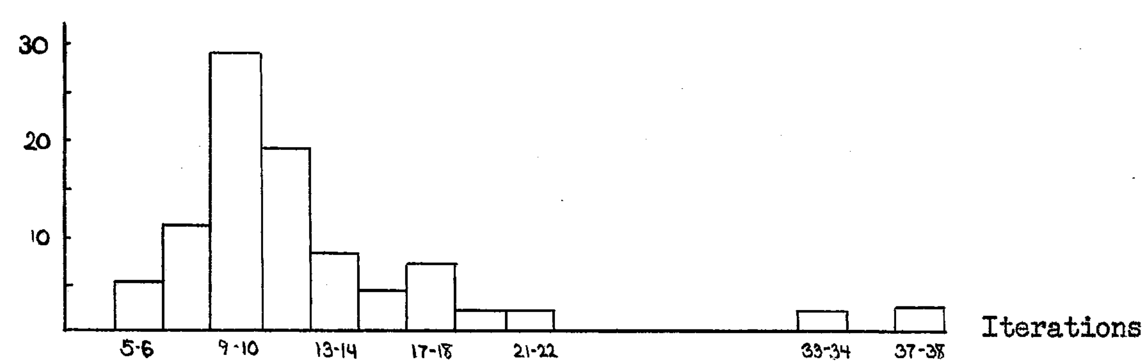


Figure 13. Histogram of number of iterations required for convergence; [-.70 -.52 .32, .60 .26, .44], censoring (333), N = 20, k = 100

The regression coefficients in the three-variate case usually converge monotonically, almost invariably so after the first two or three iterations. As the number of variates increase, the pattern of convergence becomes a bit more erratic. Examples of the convergence curves for two five-variate cases from Experiment 8 can be seen in Figures 14 and 15.

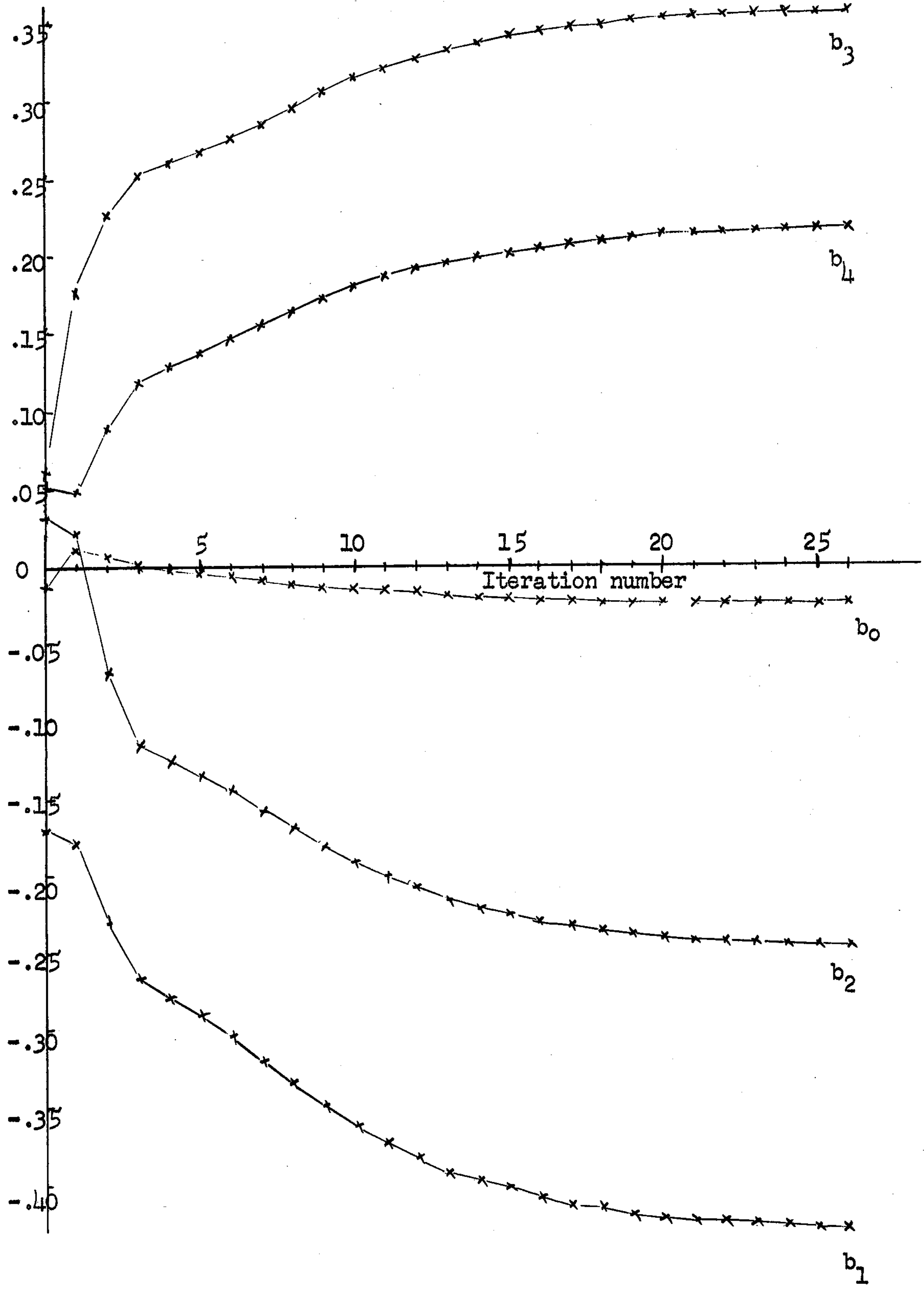


Figure 14 Convergence curves applying the Greenberg Procedure to the five-variate sample, number 107, censored (3333)

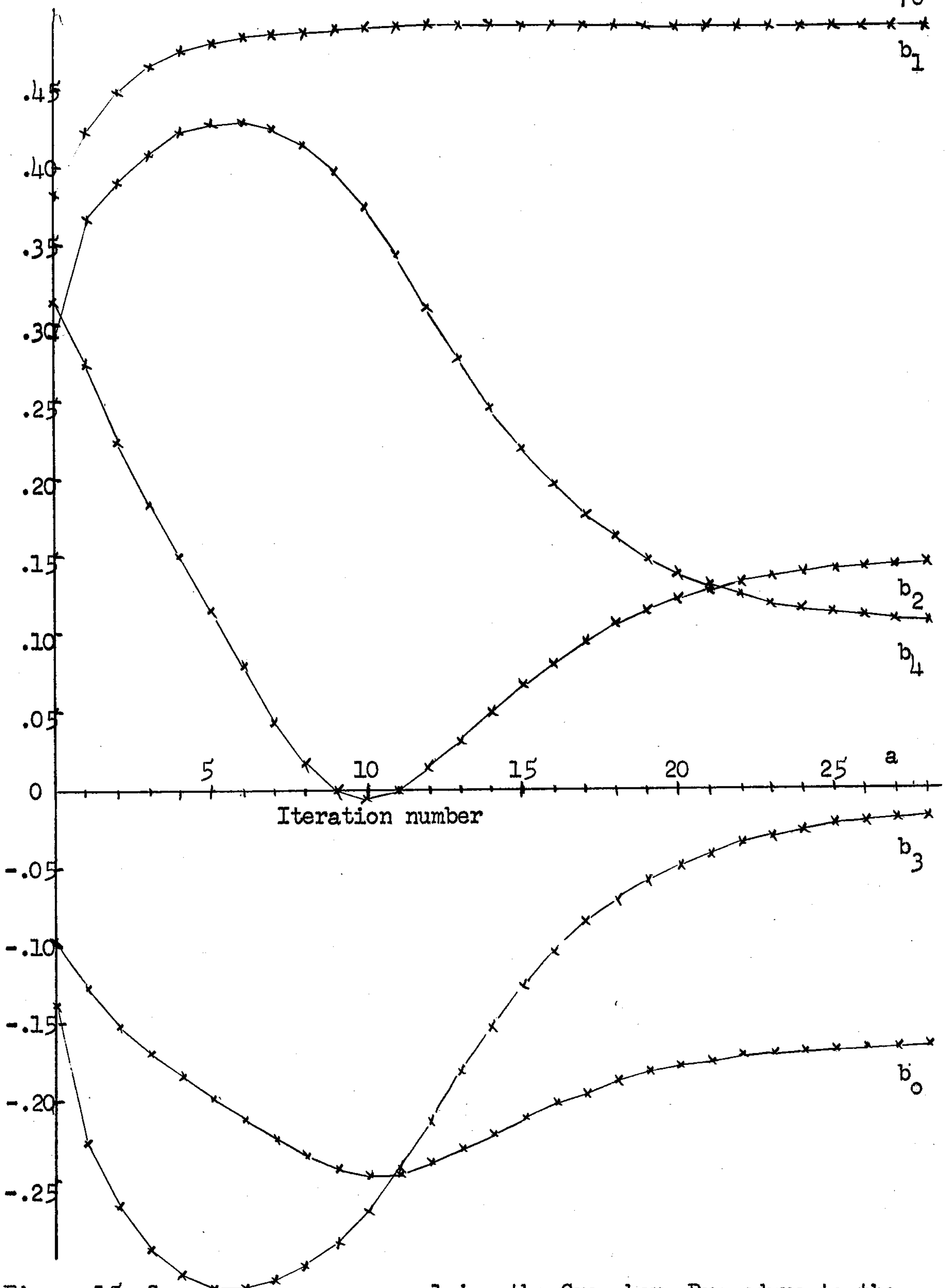


Figure 15 Convergence curves applying the Greenberg Procedure to the five-variate sample, number 106, censored (3333)

---

[a] Converges after 33 iterations.

CHAPTER VI

MAXIMUM LIKELIHOOD

6.1 Discussion

Application of the principle of maximum likelihood to the problem discussed in this thesis is practical for only a few special cases. For the most part, the mathematics involved in the maximizing process becomes extremely unwieldy.

One case in the literature which is dealt with in detail is that in which values are missing from only one of the variables in a sample from a trivariate normal population (Edgett, 1956 and Anderson, 1957).

Using the notation indicated for data of the form given on page 2, Edgett sets up the likelihood function

$$\prod_{\alpha=1}^{n} f_\alpha (x_1, x_2, x_3 \mid \mu_1, \mu_2, \mu_3;\ \sigma_1^2, \sigma_2^2, \sigma_3^2;\ \rho_{12}, \rho_{13}, \rho_{23})$$
$$\prod_{\alpha=n+1}^{n} f_\alpha (x_1, x_3 \mid \mu_1, \mu_3;\ \mu_1, \mu_3;\ \sigma_1^2, \sigma_3^2;\ \rho_{13}) \qquad (6.1)$$

and differentiates with respect to the μ_i and the σ^{ij}. Then, expressing the σ^{ij} in terms of σ_{ij}, he proceeds through some cumbersome algebra and indicates the solutions for the μ_i and σ_{ij} in a series of formulae. A thorough examination of Edgett's paper is most enlightening in that it points up the difficulties of extending the maximum likelihood method to more general cases for the problem at hand.

The approach used by Anderson in solving the same problem is worthy of mention here by virtue of its simplicity and the fact that it can be used to get maximum likelihood estimates easily for certain other special cases. Considering the same form of data referred to in the preceding paragraph, Anderson expressed the likelihood function (6.1) as

$$\prod_{\alpha=1}^{n} f_{\alpha}(x_1, x_3 \mid \mu_1, \mu_3; \sigma_1^2, \sigma_3^2; \rho_{13})$$
$$\prod_{\alpha=1}^{n} f_{\alpha}(x_2 \mid \beta_{20} + \beta_{21.3}x_1 + \beta_{23.1}x_3; \sigma_{2.13}^2) \tag{6.2}$$

where

$$\begin{aligned} \beta_{20} &= \mu_2 - \beta_{21.3}\mu_1 - \beta_{23.1}\mu_1 \\ \sigma_1^2\beta_{21.3} + \sigma_1\sigma_3\rho_{13}\beta_{23.1} &= \sigma_1\sigma_2\rho_{12} \\ \sigma_1\sigma_2\rho_{13}\beta_{21.3} + \sigma_3^2\beta_{23.1} &= \sigma_3\sigma_2\rho_{23} \\ \sigma_{2.13}^2 &= \sigma_2^2 - (\beta_{21.3}^2\sigma_1^2 + \beta_{23.1}^2\sigma_3^2 + 2\,\beta_{21.3}\beta_{23.1}\sigma_1\sigma_3\rho_{13}) \end{aligned} \tag{6.3}$$

If any other values of the parameters give a higher value to (6.1), they would give a higher value to (6.2), with the β's and $\sigma_{2.13}^2$ given by (6.3).

The maximum likelihood estimates of μ_1, μ_3, σ_1^2, σ_3^2, and ρ_{13} are based on all N observations, while the estimates of the regression parameters for x_2, given x_1 and x_3, are computed from the complete observations only. The values of μ_2, σ_2^2, ρ_{12} and ρ_{23} are then found from the relations (6.3). If we consider $x_3 = y$ as the dependent variate, estimates of the coefficients for the prediction of y from x_1 and x_2 may be calculated by formulae similar to 6.3. The results are identical with those of Edgett and, in addition to being more easily derived, are more simply expressible. Another example of this approach will be considered in discussing the example of Chapter VII. It should be noted, however, that this sort of manipulation does not provide a general solution. For example, it is not applicable to the trivariate cases we have considered in previous chapters in which censoring was of the form (c_1c_2).

The trivariate cases of Experiment 1, in which observations were missing from one variable only, have been considered using the maximum likelihood estimates described above.

6.2 Bias and Variance of Regression Coefficients

Table 25 indicates the biases and variances of the regression coefficients for the cases of Experiment 1 where maximum likelihood estimation was carried out as described above.

As was the case in Chapters III, IV, and V, only the biases of the constant coefficient are significant by the t criterion, and this has been explained as a sampling phenomenon. As one would expect, there is a tendency for the variance of the coefficients to be higher when eight values are censored than when only four are absent.

6.3 Estimate of Error

It is well known that maximum likelihood estimates of variance are biased. In the case of the estimates of $\sigma^2_{y.x}$ above, three parameters (the β_i) were estimated from twenty (fragmentary) observations, so that an estimate corrected for this bias would be obtained by using $20/17\ \hat{\sigma}^2_{y.x}$.

There will also be some bias due to the fragmentary character of the data. A rough adjustment for both types of bias might consist of multiplying the maximum likelihood estimate by a factor

$$k = \frac{N - \frac{c}{p-1}}{N-p- \frac{c}{p-1}}$$

Table 25. Statistics of regression coefficients estimated by maximum likelihood; Experiment 1, N = 20, k = 20

Censoring	[.25 .50, .25]						[.80 .60, .80]					
	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$	$\bar{b}_o-\beta_o$	$\hat{V}(b_o)$	$\bar{b}_1-\beta_1$	$\hat{V}(b_1)$	$\bar{b}_2-\beta_2$	$\hat{V}(b_2)$
(04)	.118*[a]	.051	.049	.053	.002	.037	.087*	.029	.026	.062	.006	.057
(40)	.131*	.063	.049	.103	.009	.029	.083*	.027	.022	.107	-.001	.057
(08)	.096	.061	.042	.076	-.033	.096	.070	.036	-.039	.096	.045	.081
(80)	.115	.073	.070	.208	-.032	.051	.078*	.027	.065	.186	-.054	.085

[a] The symbol * indicates $t = \frac{\bar{b}_i - \beta_i}{s_{\bar{b}_i}} > t_{.05,\ 19}$, where $s^2_{\bar{b}_i} = \frac{\hat{V}(b_i)}{20}$.

where c equals the total number of missing characters among the p-1 independent variates.

The mean values and variances of errors of estimate, calculated from the twenty samples for each case in Experiment 1, are presented in Table 26. The mean values multiplied by the correction factor, k, above are also given.

Table 26. Means and Variances of $\hat{\sigma}^2_{y.x}$ estimated by maximum likelihood

	[.25 .50, .25] $\sigma^2_{y.x} = .7333$			[.80 .60, .80] $\sigma^2_{y.x} = .3556$		
Censoring	$\overline{\hat{\sigma}^2_{y.x}}$	$\hat{V}(\hat{\sigma}^2_{y.x})$	$k\,\overline{\hat{\sigma}^2_{y.x}}$	$\overline{\hat{\sigma}^2_{y.x}}$	$\hat{V}(\hat{\sigma}^2_{y.x})$	$k\,\overline{\hat{\sigma}^2_{y.x}}$
(04)	.617	.063	.741	.296	.012	.355
(40)	.622	.069	.746	.294	.012	.353
(08)	.605	.061	.745	.284	.009	.350
(80)	.562	.080	.692	.280	.012	.345

6.4 Efficiency of Prediction Equations

Means of E and their standard errors are indicated in Table 27A and the means of Δ and their associated standard errors are shown in Table 27B.

Those tables do not indicate marked differences between the two populations and show an expected decrease in efficiency with increased censoring. A comparison of the efficiencies obtained using maximum likelihood with those obtained by using other methods is discussed in Chapter VIII.

Table 27A. Means of E and their standard errors estimated by maximum likelihood; Experiment 1, N = 20, k = 20

	[.25 .50, .25]		[.80 .60, .80]	
Censoring	$\bar{E}$	S.E.($\bar{E}$)	$\bar{E}$	S.E.($\bar{E}$)
(04)	.981	.005	.970	.008
(40)	.969	.008	.973	.008
(08)	.918	.029	.930	.014
(80)	.017	.027	.927	.024

Table 27B. Means of Δ and their standard errors estimated by maximum likelihood; Experiment 1, N = 20, k = 20

	[.25 .50, .25]		[.80 .60, .80]	
Censoring	$\bar{\Delta}$	S.E.($\bar{\Delta}$)	$\bar{\Delta}$	S.E.($\bar{\Delta}$)
(04)	.325	.025	.237	.020
(40)	.354	.034	.241	.023
(08)	.369	.032	.278	.017
(80)	.424	.053	.280	.035

CHAPTER VII

AN EXAMPLE

7.1 The Data

Difficulties involved in the determination of muscle potassium severly limit the physician in prescribing proper therapy for patients suffering a depletion of this ion. The actual measurement of muscle potassium requires the use of a radioactive isotope, which is an expensive, exacting, and time consuming procedure.

Working with laboratory animals, Welt and his associates [1958] recently carried out studies aimed at developing methods to facilitate measurement of muscle potassium in humans. Their approach involved predicting the concentration of muscle potassium from a number of other more easily obtainable measurements. The ability to determine the concentration of serum potassium, serum carbon dioxide, and red cell potassium from the analysis of a single blood sample provided such measurements. Data available for such determinations on a series of fifty-four rats is given in Table 28. The animals were sacrificed to obtain the associated values of muscle potassium.

It will be noted that determinations for serum carbon dioxide, (X_2), and red cell potassium, (X_3), are not available for every animal. Thus the question of methodology in forming an appropriate prediction equation arises.

Table 28. Biochemical determinations from Welt data

Rat Number	X_1 Serum K (mM./L.)	X_2 Serum CO_2 (mM./L.)	X_3 Red Cell K (mM./L.)	$X_4 = Y$ Muscle K (mM/100 gr.)
1	2.61	--[a]	--	44.7
2	3.61	24.6	--	47.7
3	2.94	25.7	--	47.3
4	3.58	--	--	45.5
5	4.65	19.7	101.4	47.2
6	4.53	23.4	102.0	47.8
7	3.41	25.6	100.8	47.2
8	3.41	22.4	99.6	46.9
9	5.00	24.7	98.8	45.7
10	5.38	21.6	101.3	44.9
11	3.54	--	106.0	44.9
12	4.73	--	94.3	47.5
13	5.05	--	96.8	46.0
14	5.00	--	100.3	46.5
15	2.84	22.8	--	43.5
16	3.11	--	--	38.1
17	2.87	25.1	--	42.9
18	3.32	--	108.0	45.0
19	2.79	17.4	101.5	42.4
20	3.10	25.4	105.9	42.3
21	2.61	--	107.3	42.3
22	2.94	19.3	102.5	39.1
23	2.32	--	106.0	38.1
24	2.74	23.4	101.2	41.0
25	2.70	23.6	100.1	34.6
26	2.43	21.5	100.1	41.0
27	2.60	26.9	96.8	34.3
28	2.60	25.1	96.7	38.6
29	2.85	22.0	100.9	38.6
30	1.84	26.2	100.1	35.5
31	2.05	29.8	104.2	34.8
32	1.61	25.1	96.9	33.4
33	1.92	30.6	100.3	34.5
34	2.01	29.5	100.2	38.8
35	1.66	33.3	106.3	31.5
36	2.02	31.7	103.2	33.0
37	1.50	33.8	106.8	30.6
38	1.72	34.4	96.0	31.3
39	1.73	38.7	111.0	33.1
40	1.97	42.2	98.8	31.4
41	1.67	39.4	91.1	29.5
42	1.68	--	89.4	25.2
43	1.72	27.4	92.8	27.7
44	1.84	--	106.3	28.2
45	1.74	34.3	104.1	30.0
46	2.20	--	99.3	27.7
47	1.50	44.3	109.6	25.0
48	2.16	43.8	86.5	24.2
49	2.81	--	66.5	20.6
50	4.81	--	103.0	29.6
51	1.62	39.2	105.2	23.4
52	2.73	--	83.2	30.1
53	2.65	--	98.0	23.1
54	2.38	--	90.4	29.9

[a] The symbol-- indicates a missing datum.

7.2 Solutions

The methods indicated in previous chapters will be applied here and the results given. A discussion of these results will be given in section 7.3.

Equation (7.1) below was determined by ordinary least squares, using the thirty-two complete observations.

$$Y_c = 23.93 + 3.53x_1 - .45x_2 + .16x_3 \qquad (7.1)$$

It was mentioned in the previous chapter that the approach of Anderson (1957) would be used in obtaining a maximum likelihood prediction equation from the Welt data. To do so, we delete values of x_2 from rats number two, three, fifteen, and seventeen, so that the remaining data may be arranged in the form:

$$\begin{array}{llllll}
x_{11} & x_{12} & x_{13} & \cdots\cdots\cdots\cdots & & x_{1N} \\
x_{21} & x_{22} & x_{23} & \cdots\cdots & x_{2n_2} & \\
x_{31} & x_{32} & x_{33} & \cdots\; x_{3n_1} & & \\
x_{41} & x_{42} & x_{43} & \cdots\cdots\cdots\cdots & & x_{4N}
\end{array}$$

where the primary subscript denotes the variable; the secondary subscript, the observation; and $x_4 \equiv y$ is the dependent variable. In our particular case, there are $n_1 = 32$ observations complete on all variates, $n_2 = 47$ complete on all except x_3, and $N = 54$ complete on x_1 and y. The assumption of multivariate normality is reasonable here and the likelihood function can be written:

$$\prod_{\substack{\alpha=1 \\ i,j=1,2,3,4}}^{n_1} f_\alpha(x_i|\mu_i,\sigma_i^2,\rho_{ij}) \prod_{\substack{\alpha=n_1+1 \\ i,j=1,3,4}}^{n_2} f_\alpha(x_i|\mu_i,\sigma_i^2,\rho_{ij}) \prod_{\substack{\alpha=n_1+n_2+1 \\ i=1,4}}^{N} f_\alpha(x_i|\mu_i,\sigma_i^2,\rho_{14}) \qquad (7.2)$$

and may be rewritten

$$= \prod_{\substack{\alpha=1\\ i=1,4}}^{N} f_\alpha(x_i|\mu_i,\sigma_i^2,\rho_{14}) \prod_{\alpha=1}^{n_2} f_\alpha(x_3|\beta_{30.0} + \beta_{31.4}x_1 + \beta_{34.1}x_4, \sigma^2_{3.14})$$

$$\prod_{\alpha=1}^{n_1} f_\alpha(x_4|\beta_{20.00} + \beta_{21.34}x_1 + \beta_{23.14}x_3 + \beta_{24.13}x_4) \tag{7.3}$$

where

$$\begin{aligned}
\beta_{30.0} &= \mu_3 - \beta_{31.2}\mu_1 - \beta_{34.1}\mu_4\\
\sigma^2_{3.14} &= \sigma_3^2 - (\beta^2_{31.2}\sigma_1^2 + \beta^2_{34.1}\sigma_4^2 + 2\beta_{31.4}\beta_{34.1}\sigma_1\sigma_4\rho_{14})\\
\sigma_1^2\beta_{31.4} + \sigma_1\sigma_4\rho_{14}\beta_{34.1} &= \sigma_1\sigma_3\rho_{13}\\
\sigma_1\sigma_4\rho_{14}\beta_{31.4} + \sigma_4^2\beta_{34.1} &= \sigma_3\sigma_4\rho_{34}
\end{aligned} \tag{7.4}$$

and

$$\begin{aligned}
\beta_{20.00} &= \mu_2 - \beta_{21.43}\mu_1 - \beta_{23.14}\mu_3 - \beta_{24.23}\mu_4\\
\sigma^2_{2.134} &= \sigma_2^2 - \beta' V\beta\\
V\beta &= \gamma
\end{aligned} \tag{7.5}$$

where

$$V = \begin{bmatrix} \sigma_1^2 & \sigma_{13} & \sigma_{14}\\ \sigma_{31} & \sigma_3^2 & \sigma_{34}\\ \sigma_{41} & \sigma_{43} & \sigma_4^2 \end{bmatrix} \qquad \gamma = \begin{bmatrix} \sigma_{12}\\ \sigma_{23}\\ \sigma_{24} \end{bmatrix} \qquad \text{and } \beta = \begin{bmatrix} \beta_{21.43}\\ \beta_{23.14}\\ \beta_{24.13} \end{bmatrix}$$

Thus, maximum likelihood estimates of μ_1, μ_4, σ_1^2, σ_4^2, and ρ_{14} are obtained in the usual way from all $N = 54$ observations. Estimates of μ_3, σ_3^2, ρ_{13}, and ρ_{23} are made from these foregoing values combined with estimates of the regression statistics from the n_2 observations using the relationships of (7.4). The remaining parameters are based on those already calculated plus regression statistics from the n_1 complete observations by substituting these in the parametric forms of (7.5) and solving. The estimates for the regression parameters β_{40}, $\beta_{41.23}$, $\beta_{42.13}$, $\beta_{43.12}$, and $\sigma^2_{4.123}$ may then be derived from the maximum likelihood estimates of the μ_i, σ^2_i, and ρ_{ij}. The resulting equation is

$$Y_\ell = 21.38 + 2.77x_1 - .58x_2 + .25x_3 \tag{7.6}$$

All the data available were used in obtaining a prediction equation by the method of paired correlation coefficients. The correlation matrix, the number of values upon which each coefficient was based, and the confidence interval for each coefficient is given in Table 29.

Table 29. Sample correlation coefficients and their confidence limits based on the n_{ij} available pairs of x_i and x_j in Table 28

ij	n_{ij}	r_{ij}	ρ_{ij} 95% confidence interval
12	36	-.625	-.79 to -.34
13	47	-.020	-.29 to .28
23	32	.008	-.36 to .37
1y	54	.672	.50 to .79
2y	36	-.792	-.89 to -.62
3y	47	.342	.06 to .57

Further reference will be made to this in the discussion section below. The equation determined by this method is

$$Y_p = 13.11 + 2.24x_1 - .67x_2 + .37x_3 \tag{7.7}$$

The Greenberg method also utilized all the data available. The sample variate means were used as initial estimates for the missing values and the iteration procedure described in Chapter V was carried out. Convergence was reached after fourteen iterations, each of which took two minutes and seventeen seconds. Slightly over a half hour of IBM 650 time was required for all the calculations which yielded the prediction equation

$$Y_g = 17.94 + 2.40x_1 - .69x_2 + .32x_3 \tag{7.8}$$

7.3 Discussion of the results

The solutions given in section 7.2 will be discussed briefly here with the purpose of pointing out the advantages of and objections to applying the various methods to the set of data at hand. The broader aim of indicating factors, which would be of importance in extending these methods and selecting the appropriate ones for other practical situations, is implied.

Equation (7.1) was determined by ordinary least squares using only the thirty-two complete observations. Although a sample of this size can be expected to give fairly reliable results, it seems more than a little wasteful to discard data from over forty percent ($\frac{22}{54}$) of the animals used; or, on a measurement basis, over one-third ($\frac{63}{187}$) of the total number of determinations available. A researcher would have serious reservations about disregarding so great a portion of data even when observations are easy to obtain. If the determinations were from human subjects rather than rats, the research worker would be less inclined to discard incomplete data.

Another point to be considered here is the one discussed generally in paragraph 3.1. Inspection of Table 28 reveals that the data to be discarded would include many of the observations with very high and very low muscle potassium values. Ignoring such portions of the data would seem unwise if the equation were to be used to predict muscle potassium values which were expected to be at the extremes of the sample. With only the data of Table 28 it is not possible to test whether or not one of the other methods might be preferable in this regard. However, some of the results using each prediction equation are tabulated in Table 30 below for the three complete observations having muscle potassium determinations at either end of the muscle potassium scale ($25.0 > y > 47.0$). The sums of the absolute deviations and sums of the squared deviations from the true values for these six observations are given at the foot of the table. From this examination no dramatic differences in the methods are indicated as far as predictability at the extremes is concerned.

In obtaining prediction equation (7.6) by the method of maximum likelihood, only four values were discarded, two of which were included in observation vectors having low muscle potassium values. However, the serum and muscle potassium values (x_1 and x_4) in these vectors were retained so that not all of the information on any one animal was ignored; and, on a measurement basis, only slightly over two percent ($\frac{4}{187}$) of the available information was discarded. Considering the sample size and known desirable properties of maximum likelihood estimators, prediction equation (7.6) could be regarded as a reasonable first choice. It must be remembered, however, that this is a special case of maximum likelihood estimation, and that a slightly different pattern of missing observations would have precluded its use here.

Table 30. Some extreme muscle potassium values predicted from the Welt data by various methods

For rat number	Value predicted by Y_c	Y_ℓ	Y_p	y_g	Actual value y
5	47.87	48.18	47.61	47.85	47.2
6	45.89	45.90	45.09	45.21	47.8
7	40.76	41.23	40.67	40.63	47.2
47	27.11	27.37	27.11	26.00	25.0
48	25.95	23.72	20.43	20.57	24.2
51	29.09	29.53	29.17	28.39	23.4
$\sum \lvert Y - y \rvert$	18.56	17.83	21.31	19.43	
$\sum (Y - y)^2$	85.40	83.62	102.20	89.38	

Comments on prediction equation (7.7), determined by the Method of Paired Correlation Coefficients, call for reference to Table 29. The sample correlation matrix given there with the confidence limits for each coefficient, indicate that this is a sample from a population for which this type of estimation is not particularly well suited;

that is, one in which the correlations between variables are of different magnitudes. Even though we are dealing with a different type of censoring and have larger numbers of pairs than were considered in the empirical sampling work of Chapter IV, the Method of Paired Correlation Coefficients would appear to be the least appropriate of the four methods considered here.

The Greenberg Procedure, which also utilizes all the data available yields the prediction equation (7.8). There are no a priori reasons for not applying the Greenberg Procedure to these data, nor would its use be contraindicated on the basis of the empirical sampling work discussed in previous chapters. One objection to employing it here might be the rather lengthy computations involved. These would seem to be of minor importance considering the availability of high speed computing machinery and the accompanying program. In this case, the Greenberg Procedure permitted a solution which would certainly be preferable to that obtained by the Method of Paired Correlation Coefficients. If the missing characters had been more evenly distributed among the independent variables, so that the method of maximum likelihood could not have been easily applied to the bulk of the data, it would have presented a still more appealing alternative.

CHAPTER VIII

SUMMARY AND CONCLUSIONS

8.1 Summary

In the previous chapters, alternative methods have been described for dealing with the problem of prediction by multiple regression when values are missing from among the independent variates. The results of empirical studies designed to evaluate the efficiency of the respective methods were discussed. The objectives of this chapter are first, to compare these methods by bringing together results in which all the methods were used with the same or similar data and secondly, to make suggestions regarding the use of the methods. Finally, comments will be made on the need for further research to answer some of the questions arising from the results of this work.

The eight sampling experiments described in Chapter II were sequential in nature. Experiment I was the broadest in scope but perhaps the narrowest in depth. A summary of the statistics of prediction efficiency for all the sampling done in this experiment is given in Tables 31A and 31B. These tables indicate that estimation by maximum likelihood, where it can be easily employed, is superior to the other schemes. For the three variate cases where maximum likelihood estimation is not easily carried out, neither the Paired Correlation Method nor the Greenberg Procedure appear to yield results overwhelmingly better than those obtained by merely deleting incomplete observations. However, a population differential was noted in results given by the Paired Correlation Method.

Table 31A. Values of $\overline{E}$ from Experiment 1;

N = 20, k = 20

Censoring	[.25 .50, .25]				[.80 .60, .80]			
	Complete Observations	Paired Correlation	Greenberg Procedure	Maximum Likelihood	Complete Observations	Paired Correlation	Greenberg Procedure	Maximum Likelihood
(04)	.944	.983	.967	.981	.944	.931	.944	.970
(40)	.939	.967	.956	.969	.939	.939	.953	.973
(08)	.881	.936	.875	.918	.881	.833	.845	.930
(80)	.858	.906	.825	.917	.858	.880	.860	.927
(11)	.964	.983	.976		.964	.962	.978	
(22)	.939	.972	.956		.939	.943	.960	
(33)	.885	.958	.939		.885	.901	.910	
(44)	.855	.948	.894		.855	.874	.836	
(55)	.803	.918	.810		.803	.819	.764	
(66)	.722	.857	.614		.722	.769	.664	
	[.25 .50 .75, .25 .50, .25]							
(111)	.967	.967	.949[a]					
(222)	.888	.907	.858[a]					
(333)	.795	.870	.736[a]					
(444)	.625	.784	.627[a]					
(555)	.335	.606	.445[a]					

[a] These means were based on k = 5.

Table 31B. Values of $\overline{\Delta}$ from Experiment 1;

N = 20, k = 20

Censoring	[.25 .50, .25] Complete Observations	Paired Correlation	Greenberg Procedure	Maximum Likelihood	[.80 .60, .80] Complete Observations	Paired Correlation	Greenberg Procedure	Maximum Likelihood
(04)	.353	.329	.347	.325	.246	.253	.253	.237
(40)	.385	.355	.379	.354	.268	.270	.262	.241
(08)	.381	.360	.431	.369	.266	.335	.322	.278
(80)	.468	.430	.518	.424	.326	.338	.345	.280
(11)	.340	.321	.330		.237	.246	.239	
(22)	.375	.329	.351		.261	.251	.256	
(33)	.437	.352	.386		.304	.292	.291	
(44)	.459	.376	.447		.320	.316	.343	
(55)	.488	.422	.564		.340	.360	.401	
(66)	.607	.484	.833		.423	.408	.518	
	[.25 .50 .75, .25 .50, .25]							
(111)	.252	.262	.324[a]					
(222)	.301	.296	.365[a]					
(333)	.348	.330	.482[a]					
(444)	.524	.807	.576[a]					
(555)	1.292	.774	1.114[a]					

[a] These means were based on k = 5.

This population differential led to the second and third experiments which were designed to investigate it further. In Experiment 2, it was found that a general quadratic relationship would describe the mean efficiencies of samples from trivariate populations with different correlation matrices. The fitted quadratic relationship, based on 780 samples from thirty-nine different populations, revealed that the Method of Paired Correlation Coefficients was grossly inefficient for samples from populations in which the variables were highly and heterogeneously correlated. Likewise, it was eminently satisfactory for cases in which the variables were not so related.

Experiment 3 indicated in a rather cursory way that the results of Experiment 2 were also applicable to four-variate situations.

Experiment 4 was planned to confirm the results of Experiment 1 over a wider range of populations as far as the Greenberg Procedure was concerned. It was also intended to check on whether or not a population differential of the type indicated in Experiment 2 existed for this method of estimation. Tables 32A and 32B show that although the efficiency of the Greenberg Procedure is insensitive to the trivariate population being sampled, it does not yield results superior to those obtained by deleting incomplete observations. In Experiment 4, the Greenberg Procedure also yielded more than the expected number of biased regression coefficients by the t criterion defined in Chapter II. The biases were compensating in nature, however, as far as prediction was concerned. These results, combined with the extensive computations necessary for solutions by the Greenberg Procedure, seem to dictate against using it in trivariate cases.

Table 32A. Values of $\bar{E}$ from Experiment 4B; censoring (55), N = 20, k = 10

Population	Complete Observations	Paired Correlation	Greenberg Procedure
[-.2 .6 , .6]	.778	.558	.810
[-.2 .6 , -.2]	.845	.867	.760
[-.2 -.2 , -.2]	.788	.891	.808
[-.2 -.2 , .6]	.812	.961	.820
[.6 .6 , .6]	.871	.876	.820
[.6 .6 , -.2]	.797	.464	.837
[.6 -.2 , -.2]	.756	.866	.783
[.6 -.2 , .6]	.804	.474	.769
[.8728 .2 , .2]	.758	.753	.727
[-.4728 .2 , .2]	.908	.863	.812
[.2 .8728, .2]	.858	.854	.840
[.2 -.4728, .2]	.761	.803	.830
[.2 .2 , .8728]	.819	.803	.798
[.2 .2 , -.4728]	.890	.882	.822
[.2 .2 , .2]	.751	.906	.727
$s^2_{\bar{E}}$	.002601	.025107	.001349

Table 32B. Values of $\bar{\Delta}$ from Experiment 4B; censoring (55), N = 20, k = 10

Population	Complete Observations	Paired Correlation	Greenberg Procedure	Uncensored Data
[-.2 .6 , .6]	.199	.294	.180	.095
[-.2 .6 , -.2]	.542	.383	.687	.315
[-.2 -.2 , -.2]	.492	.356	.485	.287
[-.2 -.2 , .6]	.475	.373	.550	.322
[.6 .6 , .6]	.275	.346	.382	.258
[.6 .6 , -.2]	.192	.512	.191	.140
[.6 -.2 , -.2]	.689	.510	.689	.363
[.6 -.2 , .6]	.245	.416	.242	.160
[.8728 .2 , .2]	.600	.523	.766	.332
[-.4728 .2 , .2]	.401	.499	.532	.315
[.2 .8728, .2]	.219	.268	.269	.172
[.2 -.4728, .2]	.389	.458	.426	.312
[.2 .2 , .8727]	.288	.288	.306	.196
[.2 .2 , -.4728]	.320	.344	.428	.238
[.2 .2 , .2]	.722	.452	.831	.324

Experiment 5 showed that the number of iterations required for convergence in using the Greenberg Procedure was influenced more by the proportion than by the amount of missing data.

In Experiment 6, samples censored (3333) were selected from two different five-variate populations. Prediction equations were computed by deleting incomplete observations and by applying the Paired Correlation Method. The five-variate populations were chosen to represent cases in which the Paired Correlation Method was expected to yield diverse results in terms of prediction efficiency. The resulting means and standard errors of E and Δ are given in Tables 33A and 33B, respectively. These tables represent further evidence to confirm, that, in higher variate cases, the Paired Correlation Method is suitable for use with fragmentary samples from populations with homogeneously correlated variables; but that it is definitely unsatisfactory for samples from populations with heterogeneously correlated variables. The histograms representing the distribution of E for both methods of estimation in the two populations are given in Figure 7.

Table 33A. Values of $\bar{\bar{E}}$ and their standard errors; Experiment 6, censoring (3333), N = 20, k = 100

	Complete Observations	Paired Correlation
	[.60 .80 .29 .40, .50	-.20 .50, .20 .70, .30]
$\bar{\bar{E}}$	.515	.357
S.E.($\bar{\bar{E}}$)	.020	.027
	[.36 .40 .41 .32, .24	.24 .30, .24 .25, .22]
$\bar{\bar{E}}$	.518	.874
S.E.($\bar{\bar{E}}$)	.023	.016

Table 33B. Values of $\bar{\Delta}$ and their standard errors; Experiment 6, censoring (3333), N = 20, k = 100

	Complete Observations	Paired Correlation	Uncensored Data
	[.60 .80 .29 .40, .50 −.20 .50, .20 .70, .30]		
$\bar{\Delta}$	.440	3.040	.183
S.E.($\bar{\Delta}$)	.026	1.560	.006
	[.36 .40 .41 .32, .24 .24 .30, .24 .25, .22]		
$\bar{\Delta}$	1.001	.650	.444
S.E.($\bar{\Delta}$)	.059	.091	.015

Although Experiments 1 and 4 did not indicate the Greenberg Procedure to be worthwhile in any of the trivariate cases considered, it was not shown to be clearly inferior to the other methods as far as prediction efficiency was concerned. An examination of the relative effectiveness of the Greenberg Procedure in higher variate cases seemed to be warranted and was undertaken in Experiment 7. Complete observations, the Paired Correlation Method, and the Greenberg Procedure were used to obtain prediction equations and measures of their efficiency for one hundred samples from a population for which the Paired Correlation Method was not thought to be especially well suited. The results in terms of the sampling distributions of E and Δ are shown in Figures 11 and 12, respectively. The mean values of E and Δ, with their standard errors, for the different methods of estimation are given in Table 34. According to these criteria, both the Greenberg Procedure and the deletion method are better than the Paired Correlation Method; and results using the Greenberg Procedure appear to be at least slightly better than those obtained by omitting the incomplete observations.

It will be recalled from Chapter V, however, that the regression coefficients obtained by the Greenberg Procedure in this experiment showed significant bias by the t criterion.

Table 34. Values of $\bar{E}$, $\bar{\Delta}$ and their standard errors from Experiment 7; population [.70 -.51 .31, .60 .26, .44], censoring (333) N = 20, k = 100

	Estimation by			
	Complete Observations	Paired Correlation	Greenberg Procedure	Uncensored Data
$\bar{E}$	.794	.603	.837	
S.E.($\bar{E}$)	.014	.024	.013	
$\bar{\Delta}$	.266	.482	.245	.189
S.E.($\bar{\Delta}$)	.013	.063	.015	.006

The computing time involved precluded an extensive amount of sampling with the Greenberg Procedure for five-variate cases; but the results of its use with four variates, relative to that with three, suggested some evaluation at the five-variate level. Also, the need for some convergence information at higher variate levels demanded investigation with at least a few samples. Accordingly, ten samples, each censored (3333), were selected from each of the two populations used in Experiment 6; and the Greenberg Procedure was used to find prediction equations from them. A summary of the resulting efficiencies is given by the means and standard errors of E and Δ in Tables 35A and 35B, respectively. In population [.60 .80 .29 .40, .50 -.20 .50, .20 .70, .30], the Greenberg Procedure yielded better results than those obtained by the use of complete observations. In population [.36 .40 .41 .32, .24 .24 .30, .24 .25, .22],

Table 35A. Values of $\bar{\bar{E}}$ and their standard errors; Experiment 8, censoring (3333), N = 20, k = 10

	Complete Observations	Paired Correlation	Greenberg Procedure
	[.60 .80 .29 .40, .50 -.20 .50, .20 .70, .30]		
$\bar{\bar{E}}$	.508	.440	.648
S.E.($\bar{\bar{E}}$)	.077	.086	.066
	[.36 .40 .41 .32, .24 .24 .30, .24 .25, .22]		
$\bar{\bar{E}}$	.601	.889	.664
S.E.($\bar{\bar{E}}$)	.089	.028	.076

Table 35B. Values of $\bar{\Delta}$ and their standard errors; Experiment 8, censoring (3333), N = 20, k = 10

	Complete Observations	Paired Correlation	Greenberg Procedure	Uncensored Data
	[.60 .80 .29 .40, .50 -.20 .50, .20 .70, .30]			
$\bar{\Delta}$	.618	.596	.356	.206
S.E.($\bar{\Delta}$)	.181	.110	.042	.020
	[.36 .40 .41 .32, .24 .24 .30, .24 .25, .22]			
$\bar{\Delta}$	1.066	.506	.881	.426
S.E.($\bar{\Delta}$)	.261	.077	.114	.063

there appeared to be little difference between the Greenberg Procedure and the use of complete observations, while the Paired Correlation Method yielded significantly better results than either of these.

8.2 Conclusions

The findings summarized above support the following conclusions and suggestions regarding the use of the methods described here:

1. The use of the Paired Correlation Method is advisable in the trivariate case only if it is reasonably certain that the sample has been drawn from a population in which the variables are homogeneously correlated. This provision also holds in higher variate cases; and it must be kept in mind that, as more variates are introduced, there is less possibility of homogeneity existing among their zero order correlations.
2. If there are enough complete observations in three-and four-variate cases, and unless there is some *a priori* reason for not discarding incomplete observation vectors, the use of only the complete observations gives results which usually will not be much less efficient than those obtained by the Greenberg Procedure and Paired Correlation Method.
3. In the trivariate case, the Greenberg Procedure generally is not worthwhile.
4. In addition to these conclusions, the research carried out provides some basis for the two suggestions below:
   a) If all or nearly all the data can be used to obtain maximum

likelihood estimates, similar to the manner indicated in Chapters VI and VII, that method is to be recommended. This suggestion is based on the limited amount of sampling done in Experiment 1 and the well-known desirable properties of maximum likelihood estimates.

b) The Greenberg Procedure seems to give best results in cases of four or more variates where the Paired Correlation Method would not be expected to perform efficiently. The former technique is to be recommended for higher variate cases in which most of the observations are incomplete, but in which the total number of missing values is small and evenly distributed among the different variates. These cases could be described in the notation of section 2.2 as ones in which the c_i are large and nearly equal, and in which the c_{ij} and other c's with multiple subscripts are small or equal to zero.

The example of Chapter VII illustrates the applicability of some of these suggestions to a biological problem.

The limited amount of investigation which led to the last two recommendations (4a and 4b) points up the need for further research to confirm and supplement that done here. Possibly the most obvious need for further investigation is with the use of the Greenberg Procedure in higher variate cases. Further studies appear to be in order to determine its prediction efficiency as well as the extent and nature of bias arising with its use. The apparent effectiveness in the four-variate case should be verified for similar cases, and the rather cursory examination of the two five-variate cases carried out here certainly invites further research. However, the sampling approach in any very large investigation of this kind would probably require computing facilities of capacity greater than the IBM 650.

Another suggestion for further research is that some sampling studies be done using maximum likelihood estimates obtained by iterative methods for the three-variate cases. It is also conceivable that maximum likelihood estimates could be obtained for some particular types of censoring in four-variate cases, either by explicit solutions, as pointed out in the example of Chapter VII, or by iterative solutions. Empirical investigations involving these estimates and comparing them with the ones discussed here would be enlightening.

LIST OF REFERENCES

Aitchison, V. and J. A. C. Brown. 1957. The Lognormal Distribution. Cambridge University Press, Cambridge.

Anderson, T. W. 1957. Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. Journal of the American Statistical Association 52:200-203.

Bliss, C. I. 1952. The Statistics of Bioassay. The Academic Press, Inc. New York.

Carver, H. C. 1941. Anthropometric Data. Edwards Brothers. Ann Arbor.

Du Bois, P. H. 1957. Multivariate Correlational Analysis. Harper Brothers. New York.

Duncan, D. B. 1957. Introduction to Statistical Theory. Mimeographed classnotes. University of North Carolina, Chapel Hill.

Dwyer, P. S. 1951. Linear Computations. John Wiley and Sons, New York.

Edgett, G. L. 1956. Multiple regression with missing observations among the independent variables. Journal of the American Statistical Association 51:122-131.

Foster, W. D. 1950. On the selection of predictors: two approaches. Unpublished Ph.D. Thesis. North Carolina State College, Raleigh.

Institute of Statistics. 1957. A Record of Research IV:17. Consolidated University of North Carolina.

Lord, F. M. 1955. Estimation of parameters from incomplete data. Journal of the American Statistical Association 50:870-876.

Matthai, A. 1951. Estimation of parameters from incomplete data with application to design of sample surveys. Sankhya 11:145-152.

Moonan, W. J. 1957. Linear transformation to a set of stochastically dependent normal variables. Journal of the American Statistical Association 52:247-252.

Morrison, R. D. 1956. Some studies on the estimates of the exponents in models containing one and two exponentials. Unpublished Ph.D. Thesis. North Carolina State College, Raleigh.

Munter, A. H. 1936. A study of the lengths of long bones of the arms and legs in man, with special reference to Anglo-Saxon skeletons. Biometrika 28:258.

Neiswanger, W. A. and T. A. Yancey. 1959. Parameter estimates and autonomous growth. Journal of the American Statistical Association 54: 389-402.

Nicholson, G. E., Jr. 1948. The application of a regression equation to a new sample. Unpublished Ph.D. Thesis. University of North Carolina, Chapel Hill.

Nicholson, G. E., Jr. 1957. Estimation of parameters from incomplete multivariate samples. Journal of the American Statistical Association 52: 523-526.

Orcutt, G. H. and D. Cochrane. 1949. A sampling study of the merits of autoregressive and reduced form transformations in regression analysis. Journal of the American Statistical Association 44: 356-372.

Rand Corporation. 1955. A Million Random Digits with 100,000 Normal Deviates. The Free Press, Glencoe, Illinois.

Rao, C. R. 1956. Analysis of dispersion with incomplete observations on one of the characters. Journal of the Royal Statistical Society, Series B, 18: 259-264.

Sheps, M. C. and P. L. Munson. 1957. The error of replicated potency estimates in a Biological Assay method of the parallel line type. Biometrics 13: 131-132.

Wells, H. B. 1959. North Carolina perinatal mortality study: investigation of methods for analyzing data. Unpublished Ph.D. Thesis. University of North Carolina, Chapel Hill.

Welt, L. G. et al. 1958. The prediction of muscle potassium from blood electrolytes in potassium depleted rats. Transactions of the Association of American Physicians 10: 250-259.

Wilks, S. 1932. Moments and distributions of estimates of population parameters from fragmentary samples. Annals of Mathematical Statistics 3: 163-203.

APPENDIX

APPENDIX A

GENERATION OF MULTIVARIATE NORMAL DEVIATES

A.1 Transformations to Generate Trivariate Normal Deviates

Consider a p-variate normal distribution, $\underline{x}_p \sim N(\underline{\mu}_p, V_{pp})$, $\underline{\mu}_p$ being the vector of means (p dimensional) and V the p x p variance-covariance matrix. If $\underline{x}_r$ is taken as the vector of any r elements of $\underline{x}_p$ arranged in the last r positions, and if the remaining p - r = q elements are denoted $\underline{x}_q$, then the conditional distribution

$$\underline{x}_r \mid \underline{x}_q \sim N(\underline{\mu}_{r|q}, V_{rr|q})$$

where

$$\underline{\mu}_{r|q} = \underline{\mu}_r - S_{rr}S^{rq}(\underline{x}_q - \underline{\mu}_q)$$

S_{pp} is a lower triangular matrix such that $S_{pp}^{-1} = V^{1/2'^{-1}}$

$$\begin{bmatrix} \underline{\mu}_q \\ \underline{\mu}_r \end{bmatrix} = \underline{\mu}_p$$

$$\begin{bmatrix} S^{qq} & 0 \\ S^{rq} & S^{rr} \end{bmatrix} = \begin{bmatrix} S_{qq} & 0 \\ S_{rq} & S_{rr} \end{bmatrix}^{-1} = S_{pp}^{-1}$$

$$V_{rr|q} = S_{rr}^2$$

This theorem is stated and proved by Duncan (1957).

Given independent N(0,1) deviates x_1, x_2, x_3 and wishing to find the transformations necessary to generate trivariate normal deviates x_1', x_2', x_3' with $\mu_1' = \mu_2' = \mu_3' = 0$,

$$V = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}$$

we take $x_1' = x_1$, and, in a manner similar to that described below find $x_2' = \rho_{12}x_1 + \sqrt{1 - \rho_{12}^2} \quad x_2$.

We want $f(x_3 \mid x_1x_2)$

$$V = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}$$

$$S' = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ 0 & 1 - \rho_{12}^2 & (\rho_{23} - \rho_{12}\rho_{13})(1 - \rho_{12}^2)^{-1/2} \\ 0 & 0 & k(1 - \rho_{12}^2)^{-1/2} \end{bmatrix}$$

where $k^2 = 1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23}$

$$S^{-1'} = \begin{bmatrix} 1 & 0 & 0 \\ -\rho_{12}(1 - \rho_{12}^2)^{-1/2} & (1 - \rho_{12}^2)^{1/2} & 0 \\ \dfrac{(\rho_{12}\rho_{23} - \rho_{13})}{k(1 - \rho_{12}^2)^{1/2}} & -\dfrac{\rho_{23} - \rho_{12}\,\rho_{13}}{k(1 - \rho_{12}^2)^{1/2}} & \dfrac{(1 - \rho_{12}^2)}{k} \end{bmatrix}$$

$$\underline{\mu}_3 \Big|_{1,2} = (1 - \rho_{12}^2)^{-1} \quad [(\rho_{13} - \rho_{12}\rho_{23})\, x_1 + (\rho_{23} - \rho_{12}\rho_{13})\, x_2]$$

$$V_3 \Big|_{1,2} = S_{33}^2 = k^2\, (1 - \rho_{12}^2)^{-1} .$$

So

$$f(x_3 \mid x_1 x_2) = C\, e^{-1/2 \left[x_3 - \frac{\rho_{13} - \rho_{12}\rho_{23}}{1 - \rho_{12}^2} x_1 - \frac{\rho_{23} - \rho_{12}\rho_{13}}{1 - \rho_{12}^2} x_2 \right]^2 \left[\frac{1 - \rho_{12}^2}{k^2}\right]}$$

therefore

$$x_3' = \frac{\rho_{13} - \rho_{12}\rho_{23}}{1 - \rho_{12}^2} x_1' + \frac{\rho_{23} - \rho_{12}\rho_{13}}{1 - \rho_{12}^2} x_2' + \frac{k}{(1 - \rho_{12}^2)^{1/2}} x_3$$

A.2 An Example of the Procedure Used to Generate Four-variate Normal Deviates

Given independent N(0,1) deviates x_1, x_2, x_3, x_4, we wish to generate multivariate normal deviates x_1', x_2', x_3', x_4' such that $\mu_1' = \mu_2' = \mu_3' = \mu_4' = 0$

$$V = \begin{bmatrix} 1 & .25 & .50 & .75 \\ .25 & 1 & .25 & .50 \\ .50 & .25 & 1 & .25 \\ .75 & .50 & .25 & 1 \end{bmatrix}$$

Using computing methods described by Dwyer (1951) we find directly

$$(S_{ii}) \qquad S' = \begin{bmatrix} 1 & .25 & .50 & .75 \\ 0 & .96824584 & .12909944 & .32274861 \\ 0 & 0 & .85634884 & .09731237 \\ 0 & 0 & 0 & .56909018 \end{bmatrix}$$

$$(S^{ic}) = S^{-1'} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -.25819889 & 1.03279556 & 0 & 0 \\ -.54494926 & -.15560978 & 1.16774841 & 0 \\ -1.07827614 & -.55910614 & -.19968077 & 1.75719074 \end{bmatrix}$$

Take $x_1' = x_1$ and using the relationship

$$\underline{x}_r' = -S_{rr} S^{rq} \underline{x}_q' + S_{rr} x_r \text{ 1/}$$

successively for r = 2, 3, 4, q < r, we get

$$x_2' = .25x_1' + .9682458 4x_2$$
$$x_3' = .46666667x_1' + .13333333x_2' + .85534884x_3$$
$$x_4' = .61363636x_1' + .31818181x_2' + .11363636x_3' + .56909018x_4$$

A.3 Some Results Obtained by Using the Transformations of A.2

Four hundred sets of multivariate deviates were generated using the transformations above and the first four hundred input cards. These deviates yielded the following matrix of sums, sums of squares, and sums of cross products:

$$\begin{bmatrix} N & \sum x_1 & \sum x_2 & \sum x_3 & \sum x_4 \\ & \sum x_1^2 & \sum x_1x_2 & \sum x_1x_3 & \sum x_1x_4 \\ & & \sum x_2^2 & \sum x_2x_3 & \sum x_2x_4 \\ & \text{Sym.} & & \sum x_3^2 & \sum x_3x_4 \\ & & & & \sum x_4^2 \end{bmatrix} = \begin{bmatrix} 400 & 3.70 & -26.21 & 39.44 & 4.79 \\ & 377.86 & 109.08 & 205.01 & 297.28 \\ & & 408.86 & 105.29 & 219.09 \\ & \text{Sym.} & & 411.46 & 116.62 \\ & & & & 418.19 \end{bmatrix}$$

which reduces to the correlation matrix

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ & 1 & \rho_{23} & \rho_{24} \\ & \text{Sym.} & 1 & \rho_{34} \\ & & & 1 \end{bmatrix} = \begin{bmatrix} 1 & .279 & .521 & .748 \\ & 1 & .265 & .532 \\ & \text{Sym.} & 1 & .281 \\ & & & 1 \end{bmatrix}.$$

1/ It may be noted that the $S_{rr}S^{rq}$ are vectors of partial regression coefficients.

APPENDIX B

THE NORMAL DEVIATES USED IN SAMPLING

Ten thousand N(0,1) deviates were selected at random from the one hundred thousand listed in the Rand Tables (1955). A full description of these deviates, their origin, and the tests for randomness which were applied to them is given with that listing.

It will suffice here to note that the deviates were punched, five per card, in two non-overlapping groups. The starting point for each group was selected at random, one group being punched serially from rows, the other from columns. The cards were then shuffled and numbered in order. A frequency tabulation of these N(0,1) deviates is given for the ten equal intervals from -5.000 to +5.000 in Table B-1 below.

Table B-1. Frequency tabulation of ten thousand N(0,1) deviates

From	-4.000	-3.000	-2.000	-1.000	.000	.000	1.000	2.000	3.000	4.000
To	-4.999	-3.999	-2.999	-1.999	-.999	.999	1.999	2.999	3.999	4.999
Frequency	1	13	200	1335	3410	3454	1368	212	7	0
Number Expected	.3	13	214	1359	3413	3413	1359	214	13	.3