THE DETERMINATION OF REGRESSION RELATIONSHIPS USING STEPWISE REGRESSION TECHNIQUES

by

D. JOHN PAYNE, B.Sc. M.Sc.(Econ)

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy
July, 1973.

ABSTRACT

Stepwise regression routines are rapidly becoming a standard feature of large-scale computer statistical packages. They provide, in particular, a certain degree of flexibility in the selection of 'optimum' regression equations when one has available a large set of potential regressor variables.

A major problem in the use of such routines is the determination of appropriate 'cut-off' criteria for terminating the procedures. There is a tendency in practice for standard F or t- statistics to be calculated at each step of the procedure, and for this value to be compared with conventional critical values.

In this thesis an attempt has been made to provide a more satisfactory rationale for (single-step) stepwise procedures. The approach taken is to assume that a 'true' model exists (the regressors in which are a subset of those available) and to investigate the distribution of statistics which, at each stage, seem relevant to the termination decision. This leads to the consideration of alternative tests at each step to those usually employed.

In the presence of considerable analytical complexity a simulation approach is used to obtain a comparison of the relative performances of various procedures. This study encompasses the use of forward, backward and mixed forward/backward procedures in both orthogonal and non-orthogonal set-ups. Procedures are evaluated both in terms of the 'closeness' of the finally selected model to the true one, and also in terms of prediction mean square-error.

The study ends with an investigation into the usefulness of stepwise regression in identifying and estimating stochastic regression relationships of the type encountered in the analysis of time series.

ACKNOWLEDGEMENTS

I should like to express my appreciation and thanks to those from whom I have received assistance in the writing of this thesis : to my advisor Professor Clive Granger I wish to extend my particular thanks both for the interest he has shown and also for the valuable advice he has profferred throughout; to Mr. John Bates of the Economics Department for reading much of the draft version of the thesis, and also for making some useful suggestions regarding details of exposition; to Dr. Paul Newbold both for some general discussions which took place and also for his, and Professor Granger's, permission for full access to the data which features in Chapter 10; to Mrs B. Page and Mrs M. Clough for their patience and efficiency in the typing of the thesis.

Finally I wish to record a special note of gratitude to my wife Jenny for the understanding she has shown and the support she has given throughout the period I have spent working on this thesis.

CONTENTS

INTRODUCTION

The initial motivation for this thesis stems from an attempt to obtain practical content for some ideas on 'causality' which were put forward by Granger [28]. A promising approach to that problem seemed to be offered by the technique of 'stepwise regression'. In particular it was felt that the application of such a procedure to time series data might reveal the underlying lag structure relating observations on several different series. However, on trying to find information on the practical use of stepwise regression, it became evident that little work had been done in establishing its validity or usefulness. Indeed there seemed to be a proliferation of stepwise procedure variants with hardly any indication, theoretical or empirical, of their relative merits. It was for these reasons that a closer look at stepwise regression itself became the main interest.

The thesis begins with a summarized account of some basic results from the field of classical regression analysis, followed in chapter 2 by a brief account of stepwise regression itself in relation to other existing procedures of a comparable nature. In chapter 3 some extensions are made to the classical theory which are required in the later discussion. Cpater 4 is concerned with the problem of establishing a formal framework for stepwise regression in terms of both the identification and prediction objectives which are postulated there. This is followed in chapters 5 and 6 by a fairly detailed look at the situation of orthogonal regression, ending in chapter 7 with the presentation and discussion of the results of a fairly extensive simulation study. Chapters 8 and 9 are similar to the previous three chapters except t at the discussion is now turned to the non-orthogonal case. In the final main chapter, chapter 10, the investigation is extended to that of stochastic regression with special emphasis on autoregressive relationships, the chapter culminating in comparative studies with some other suggested approaches.

Chapter 1 Some Basic Results in Regression Analysis

1.1 Definition of model

In this section a model is defined which will serve as the basic framework for a large section of the subsequent discussion.

Suppose n observations are available on each of the variables $Y, X_1, \ldots\ldots X_k$. The classical linear regression model (or linear hypothesis model) which relates the regressand Y to the k regressor variables $X_j, j = 1, \ldots . k$, is:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_n X_{in} + \varepsilon_i, \qquad i = 1, \ldots . n.$$

The following assumptions are also made:

(i) Each $\varepsilon_i (i = 1, \ldots . n)$ is a random variable with zero mean, variance σ^2, and is uncorrelated with ε_j for $j \neq 1$.

(ii) The observations on the X-variables are regarded as fixed numbers.

(iii) No exact linear relationships exist amongst the X-variables.

(iv) The ε_i are normally distributed.

Assumption (ii) stems from the early application of regression analysis to the results from controlled experiments. In such circumstances one is justified in attempting to make inferences about the conditional distribution of Y with the X's held fixed since one can contemplate experimental replication. The assumption is of course untenable in non-experimental situations of the kind encountered in the field of economics, for example. In this latter case it is necessary for the X-variables to be regarded as stochastic, i.e. the model becomes a stochastic regression model. Complications then arise according to the nature of the joint distribution of ε with the set of regressors. Such problems are discussed later in Chapter 10.

There is however a class of stochastic regression relationship which can, in a certain sense, be included within the fixed regressor model. This occurs when the error terms ε_i are statistically independent of the X-variables in all n equations. One can then consider a conditional model using the set of X's actually observed. Inferences made on this conditional model can then be applied to the more general unconditional model. Of special interest is the case in which Y and the X's are joint observations from a (k+1)-dimensional normal distribution. In this case one can then properly identify many of the distribution problems of regression with those of correlation analysis for normal random variables.

The assumption (iv), though not essential for much of classical regression analysis, becomes necessary for performing tests on estimated coefficients (at least for small samples) and is needed later in the treatment of stepwise regression.

1.2 Matrix formulation of model

Before proceeding further it will be useful to reformulate our model in matrix notation.

Matrices (and vectors) will be denoted by underlined letters, e.g. $\underset{\sim}{A}$, $\underset{\sim}{b}$.

The transpose of a matrix $\underset{\sim}{A}$ will be denoted by $\underset{\sim}{A}'$.

The matrix inverse of a square, non-singular matrix $\underset{\sim}{A}$ will be denoted by $\underset{\sim}{A}^{-1}$.

The expectation of a matrix whose elements are random variables will be taken to mean the corresponding matrix of expectations, and will be denoted by $E[\underset{\sim}{A}]$, for example.

We now let

$$\underset{\sim}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix} \qquad \underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} \qquad \underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

$$\text{and } \underset{\sim}{X} = \begin{bmatrix} X_{11}\ldots X_{12}\ldots\ldots X_{1n} & \\ X_{21} & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ X_{n1} & X_{nk} \end{bmatrix}$$

Our model then becomes

$$\underset{\sim}{Y} = \underset{\sim}{X}\, \underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \qquad (1)$$

where (i) $E[\underset{\sim}{\varepsilon}] = \underset{\sim}{0}$, $E[\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'] = \sigma^2 \underset{\sim}{I}$

(ii) X is a matrix of fixed constants

(iii) The rank of $\underset{\sim}{X}$ is equal to k ($k \leq n$).

(i.e. our model is of <u>full-rank</u>)

(iv) $\underset{\sim}{\varepsilon}$ has multivariate normal distribution

i.e. $\underset{\sim}{\varepsilon}$ is $N(\underset{\sim}{0}, \sigma^2 \underset{\sim}{I})$.

In the above specifications $\underset{\sim}{0}$ represents a vector whose elements are all zero, $\underset{\sim}{I}$ is the identity matrix, i.e. all diagonal terms are unity, all off-diagonal terms are zeros.

It should be noted that our model specification allows a constant term to be included in the equation. This is achieved by taking the first column of $\underset{\sim}{X}$ to consist entirely of elements equal to 1. Throughout the later analysis it will be assumed that a constant term is automatically fitted in a regression equation (and hence will not enter into the stepwise selection process).

1.3 <u>Statistical inference in regression</u>

A first step in the inferential problem is to obtain estimates of the unknown β coefficients having desirable properties. The accepted estimation procedure for classical regression is that of least squares. If we let

$$\underset{\sim}{b} = \begin{bmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ b_k \end{bmatrix},$$

the least squares principle then impliesthat we minimize, with respect to the elements of $\underset{\sim}{b}$, the sum of the squared deviations of the Y observations from the fitted equation.

In matrix terms we have to minimize

$$S = (\underset{\sim}{Y}-\underset{\sim}{X}\underset{\sim}{b})'(\underset{\sim}{Y}-\underset{\sim}{X}\underset{\sim}{b})$$

with respect to $\underset{\sim}{b}$.

Differentiating partially with respect to the elements of $\underset{\sim}{b}$ and equating to zero the resulting expressions we obtain the normal equations

$$\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{b} = \underset{\sim}{X}'\underset{\sim}{Y}$$

and hence the least squares solution

$$\underset{\sim}{b} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y} \qquad (1)$$

This estimator can be shown to have the optimal property of being <u>best linear unbiased</u>. (This is the well-known Gauss-Markov result. For references see the end of this section.) It also follows, using assumption (iv) of the basis model, that $\underset{\sim}{b}$ is the maximum likelihood estimator of $\underset{\sim}{\beta}$. In the stochastic version of our model in which the X-variables are multivariate normal (see section 1.1) it can be shown that $\underset{\sim}{b}$ is in fact <u>minimum variance unbiased</u>.

The other main result concerning $\underset{\sim}{b}$ is that its covariance matrix $\underset{\sim}{V}$ is given by

$$\underset{\sim}{V} = E[(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})'] = \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1} \qquad (2)$$

If we then use assumption (iv) it follows that

$$\underset{\sim}{b} \text{ is } N\left(\underset{\sim}{\beta},\ \sigma^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\right). \qquad (3)$$

This result then allows significance tests on null hypotheses of the form

$$H_o \equiv \beta_j = \beta_j^* \quad (1 \leq j \leq k) \text{ for some specified } \beta_j^*.$$

If v_{jj} is the j^{th} diagonal element of $\underset{\sim}{V}$ then the quantity

$$\frac{b_j-\beta_j^*}{\sqrt{v_{jj}}}$$

will be distributed as N(0,1) if H_o is true.

Since v_{jj} requires knowledge of σ^2 one has to estimate σ^2 by

$$\hat{\sigma}^2 = \frac{RSS}{(n-k)}$$

where RSS is the residual sum of squares from the estimated regression equation. It can be shown, using standard results for the distribution of quadratic forms in normal variables that RSS, and hence $\hat{\sigma}^2$, is distributed independently of each b_j $(j = 1,...k)$. Further, $\frac{(n-k)\hat{\sigma}^2}{\sigma^2}$ is distributed as chi-square with (n-k) degrees of freedom. Hence it follows that, defining $\hat{v}_{jj}$ as the j^{th} diagonal term of $\hat{\sigma}^2(\underset{\sim}{X}'\underset{\sim}{X})^{-1}$, the statistic

$$t = \frac{b_j - \beta_j^*}{\sqrt{\hat{v}_{jj}}}$$

is, under H_o, distributed as Student's t with (n-k) degrees of freedom.

The use of the above test for investigating the significance of (partial) regression coefficients can lead to a problem of interpretation when applied to more than one b_j. This is particularly so in non-orthogonal regression situations. A regression model is orthogonal if

$$\sum_{i=1}^{n} X_{ip}X_{iq} = 0 \text{ for } p \neq q.$$

Many simplifications arise in such a case. In particular one is able to isolate sum of squares "contributions" for each of the k regressor variables, the presence of the other regressors having no influence on the "explanatory power" of any individual variable. In non-orthogonal set-ups (i.e. in which the regressors possess some degree of multicollinearity) the explanatory power of regressors does depend very much on which other variables have also been fitted. Consequently, when performing more than one test of a regression coefficient, the equation should be re-estimated each time a non-significant result leads to the dropping off of one regressor.

A more general, and more fruitful, approach to the testing of regression coefficients is provided by the Analysis of Variance technique. This emphasizes the aspect of decomposing a total sum of squares into components attributable to specific "factors". We shall see that stepwise regression can in fact be thought of as a method for performing this decomposition in a meaningful way.

The analysis of variance approach provides the following important generalization of the t-test above (4):-

Suppose we partition $\underset{\sim}{X}$ as

$$\underset{\sim}{X} = [X_* X_{**}]$$

where $\underset{\sim}{X}_*$ is $n \times (k-q)$, $\underset{\sim}{X}_{**}$ is $n \times q$.

Corresponding to this, partition $\underset{\sim}{\beta}$ as

$$\underset{\sim}{\beta}' = [\underset{\sim}{\beta}'_* \quad \underset{\sim}{\beta}'_{**}]$$

Consider the hypothesis $H_o \equiv \underset{\sim}{\beta}_{**} = 0$. (This being a special case of the more general hypothesis that $\underset{\sim}{\beta}_{**}$ takes any prescribed value.) Let

RSS = Residual sum of squares after fitting $\underset{\sim}{X}$

RSS_* = Residual sum of squares after fitting $\underset{\sim}{X}_*$ only.

The statistic

$$F = \frac{(RSS_*-RSS)(n-k)}{RSS\ (q)} \qquad (5)$$

will in general be distributed as a non-central F, collapsing to a central F under H_o. This test can be shown to be the likelihood-ratio test of the specified hypothesis. We may note the following points:

1. In the case in which q = 1 the test statistic F is computationally the square of the t value calculated at (4) (with $\beta_j^* = 0$). Since $F_{(1,\nu)}$ is distributed exactly as $(t_{(\nu)})^2$, the two tests are equivalent.
2. Again with q = 1 but allowing $\underset{\sim}{X}$ to be stochastic and normal (as described in (1.1)) the test is identical to that of a partial correlation coefficient. In fact, if X_{**} is the variable corresponding to $\underset{\sim}{X}_{**}$, we are testing $H_o \equiv \rho_{YX**.\underset{\sim}{X}_*} = 0$, where $\rho_{YX**.\underset{\sim}{X}^*}$ is the correlation between Y and X_{**} holding fixed the set of regressors contained in $\underset{\sim}{X}_*$. Whilst significance tests in the two basically different situations of fixed and stochastic regressors are performed in essentially the same way, with the same levels of significance holding, the <u>powers</u> of the tests will differ.

This concludes a very brief summary of only a small part of statistical inference in regression. Many of the results, and others, will appear again in the more general stepwise context, and for this reason it does not seem necessary to enter into explicit derivations at this stage. There are many excellent references on this topic. Among those which have been found particularly useful are Graybill[30], Johnston[35] and Goldberger[27]

1.4 The basic stepwise regression algorithm

Before presenting a description of the computational "mechanics" of stepwise regression a very brief introduction to the context of its use is perhaps appropriate. The sort of problem envisaged can be said to be that which precedes the classical analysis which has been described in the previous three sections. For the discussion up to now assumes that one knows, or has a pretty good idea, as to which regressor variables should be included in the regression equation. In practice it is not unusual, however, to have available a large number of possible explanatory variables, a decision having to be made as to which ones to select. Stepwise regression is a procedure which attempts to aid such a decision by generating a sequence of calculated regression equations, and terminating when a subset of "optimal" regressors has been found. Much of the subsequent discussion in this thesis is concerned with the problems of specifying the optimality criterion in an acceptable way, and in steering the stepwise sequence towards this optimal objective.

The origins of stepwise regression can be said to lay more in the field of computational theory than in statistics. Indeed, most of the published work on stepwise and related methods is also predominantly orientated towards computational aspects. This is certainly true of the pioneer paper for the technique due to Efroymson[24]. In that paper no attempt was made to impose any particular underlying structure on the observed data, nor were any precise objectives formulated apart from the vague one of finding an "optimal" predictive equation.

The algorithm underlying the stepwise technique will now be described, along with its implications in what should, at this stage, properly be referred to as "descriptive" regression. The algorithm itself can be identified with the Gaussian elimination method for the inversion of a square, non-singular matrix (see Orden[64]).

We begin by writing

$$\underset{\sim}{Z} = [\underset{\sim}{X} \,\vdots\, \underset{\sim}{Y}]$$

i.e. $\underset{\sim}{Z}$ is the augmentation of $\underset{\sim}{X}$ and $\underset{\sim}{Y}$, and therefore has dimensions $n \times (k+1)$.

We suppose, here and henceforth, that a constant term is always fitted automatically in the regression equation, our interest being focussed entirely on the fitting of "true" regressor variables (these being the k variables present in $\underset{\sim}{X}$). This assumption is in no way essential but it facilitates the subsequent algebraic treatment. With this in mind we lose no generality by supposing the n observations on each of the k + 1 variables in $\underset{\sim}{Z}$ are measured in terms of deviations from their corresponding sample means.

The stepwise regression algorithm consists of a sequence of pivotal operations performed successively on a matrix, starting with $\underset{\sim}{Z}'\underset{\sim}{Z}$. At any particular stage we are able to identify the elements of this matrix, $\underset{\sim}{A}$ say, with quantities relevant to a certain set of fitted regression equations. That this is true will be demonstrated below using an inductive argument. The regressions referred to above will be on a subset of the so-called <u>eligible</u> variables which for our purposes will always be taken to be the complete set $\underset{\sim}{X}$. The variable Y, for reasons which become obvious, will always be regarded as ineligible.

For the moment suppose that at a certain stage we can write

$$\underset{\sim}{A} = \begin{bmatrix} \underset{\sim}{A}_{11} & \underset{\sim}{A}_{12} \\ \underset{\sim}{A}_{21} & \underset{\sim}{A}_{22} \end{bmatrix}$$

where

$$\text{(i)}\quad \underset{\sim}{A}_{11} = (\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}$$

$$\underset{\sim}{A}_{12} = -\underset{\sim}{A}_{21} = (\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'\underset{\sim}{Z}_2$$

$$\underset{\sim}{A}_{22} = \underset{\sim}{Z}_2'\underset{\sim}{Z}_2 - \underset{\sim}{Z}_2'\underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'\underset{\sim}{Z}_2$$

and (ii) $\underset{\sim}{Z} = [\underset{\sim}{Z}_1 \vdots \underset{\sim}{Z}_2]$, where $\underset{\sim}{Z}_1$ is n × q

$\underset{\sim}{Z}_2$ is n × (k+1−q)

and where it is supposed throughout that $\underset{\sim}{Y}$ is the last column of $\underset{\sim}{Z}_2$.[†]

We note the following facts concerning $\underset{\sim}{A}$ and the k + 1 - q different regression equations obtainable by regressing each member of $\underset{\sim}{Z}_2$ against the whole $\underset{\sim}{Z}_1$ set:-

(a) $\underset{\sim}{A}_{11}$ is, up to a multiplying constant σ^2, the covariance matrix for the regression coefficients in these regressions.

(b) $\underset{\sim}{A}_{12}$ contains the estimated regression coefficients.

(c) $\underset{\sim}{A}_{22}$ is, apart from a divisor n, the sample partial covariance matrix for the set $\underset{\sim}{Z}_2$, holding fixed $\underset{\sim}{Z}_1$.

For each of the k + 1 - q such regressions it is then possible to calculate from $\underset{\sim}{A}$ the following quantities:-

1. The estimated regression coefficients, estimated standard errors and the t (or F) values for each member of $\underset{\sim}{Z}_1$ ——— the set of so-called <u>included</u> variables.

[†] Throughout much of the subsequent analysis use is made of inverses of various sub-matrices of $\underset{\sim}{Z}'\underset{\sim}{Z}$ which are centred on the main diagonal. That these inverses exist is guaranteed by the presupposition that no exact linear relationship exists amongst the data, together with the positive definiteness of moment matrices.

2. The partial correlation coefficient between Y and each other member of $\underset{\sim}{Z}_2$ (with $\underset{\sim}{Z}_1$ fixed), the t (or F) value for each eligible variable in $\underset{\sim}{Z}_2$ which would result if this variable were to be included in the regression, and a so-called "tolerance" value for each eligible variable. This tolerance value is in fact one minus the partial correlation coefficient between the variable concerned and Y, and is checked as a precaution against the inversion of an ill-conditioned matrix. The set of eligible variables in $\underset{\sim}{Z}_2$ will be referred to as the <u>excluded</u> variables.

It was asserted previously that a certain pivotal operation produced a sequence of matrices typified by $\underset{\sim}{A}$. In fact this operation, to be defined below, will either introduce an excluded variable into the regression or remove an included one depending on the pivotal element chosen. That this operation produces the appropriate matrix $\underset{\sim}{A}$ for the new regressions is the essence of the stepwise algorithm.

<u>The pivotal operation</u>:

(i) Select an element a_{cc} of $\underset{\sim}{A}$, $(1 \leq c \leq k)$.

(ii) Form a new matrix $\underset{\sim}{A}^*$ as follows:

$$a^{*}_{ij} = \begin{cases} a_{ij} - \dfrac{a_{ic}\,a_{cj}}{a_{cc}} & \text{if } i,j \neq c \\[2ex] \dfrac{a_{cj}}{a_{cc}} & \text{if } i = c,\ j \neq c \\[2ex] -\dfrac{a_{ic}}{a_{cc}} & \text{if } i \neq c,\ j = c \\[2ex] \dfrac{1}{a_{cc}} & \text{if } i = j = c \end{cases}$$

Two theorems are now stated which are sufficient to establish the main stepwise property. Since the proofs of these theorems in a statistical context do not seem to be readily available in the published literature they are given in Appendix I at the end. (A paper by Lutjohann[53] presents similar results for a more general algorithm in which sets of regressors are pivoted into or out of the equation. However, the proofs given seem unnecessarily lengthy).

Before stating these theorems another theorem which is essential to their proofs (and is also used extensively later) is given.

Theorem 1.4.1

Suppose a square, non-singular matrix B is partitioned in the form

$$\underset{\sim}{B} = \begin{bmatrix} \underset{\sim}{E} & \underset{\sim}{F} \\ \underset{\sim}{G} & \underset{\sim}{H} \end{bmatrix}, \quad \text{where } \underset{\sim}{E} \text{ and } \underset{\sim}{H} \text{ are square.}$$

Then $\underset{\sim}{B}^{-1} = \begin{bmatrix} \underset{\sim}{E}^{-1}(\underset{\sim}{I}+\underset{\sim}{F}\underset{\sim}{D}^{-1}\underset{\sim}{G}\underset{\sim}{E}^{-1}) & -\underset{\sim}{E}^{-1}\underset{\sim}{F}\underset{\sim}{D}^{-1} \\ -\underset{\sim}{D}^{-1}\underset{\sim}{G}\underset{\sim}{E}^{-1} & \underset{\sim}{D}^{-1} \end{bmatrix}$

assuming $\underset{\sim}{E}^{-1}$ and $\underset{\sim}{H}^{-1}$ exist, and where $\underset{\sim}{D} = \underset{\sim}{H} - \underset{\sim}{G}\underset{\sim}{E}^{-1}\underset{\sim}{F}$.

<u>Proof</u>. Using the usual rules for the multiplication of partitioned matrices we immediately obtain $\underset{\sim}{B}\underset{\sim}{B}^{-1} = \underset{\sim}{I}$ as required.

<u>Theorem 1.4.2.</u>

If the pivotal element a_{cc} is selected such that $1 \leq c \leq q$ then $\underset{\sim}{A}^{*}$ has the same properties as $\underset{\sim}{A}$ but related to

$\underset{\sim}{Z}_1^{*} = \underset{\sim}{Z}_1$ with the column corresponding to X_c excluded.

$\underset{\sim}{Z}_2^{*} = \underset{\sim}{Z}_2$ with the X_c column inserted.

<u>Theorem 1.4.3.</u>

If $q + 1 \leq c \leq k$ then $\underset{\sim}{A}^{*}$ has the same properties as $\underset{\sim}{A}$ but related to

$\underset{\sim}{Z}_1^{*} = \underset{\sim}{Z}_1$ with the X_c column inserted.

$\underset{\sim}{Z}_2^{*} = \underset{\sim}{Z}_2$ with the X_c column excluded.

Since the first pivotal operation, performed on $\underset{\sim}{Z}'\underset{\sim}{Z}$, is easily shown to yield the appropriate matrix $\underset{\sim}{A}$ the stepwise algorithm is thereby verified.

In practice it is customary to perform stepwise proceduresusing the correlation matrix initially instead of $\underset{\sim}{Z}'\underset{\sim}{Z}$. By normalizing in this way the matrix elements will be of a similar order of magnitude, thus minimizing the extent of round-off error in calculation. When printing out the various quantities of interest at any stage a simple re-scaling is needed using the original standard deviations.

There is of course no need to rearrange the matrix $\underset{\sim}{A}$ to conform to the partition used above at each stage. A vector of Boolean variables can, for example, be used to carry the information on the included/excluded regressor sets.

Chapter 2

Practical Procedures for Selecting Optimal Regression Equations.

2.1 The Variants of stepwise regression

In Chapter 1 the basic stepwise algorithm was presented in its simplest form. There are several ways in which this algorithm can be incorporated into a sequential procedure for selecting a regression equation. However, one can distinguish three fundamental types of procedure which can be said to generate most of the others, and these are now considered

(a) The Forward Selection Procedure

In this procedure the equation is built up one variable at a time. At each pivotal step the partial F values for each excluded eligible variable are calculated and the maximum such value determined. This is then compared with a pre-selected critical value which, if exceeded, causes the appropriate regressor to be pivoted into the equation. Otherwise the procedure terminates.

(b) The Backward Elimination Procedure

Here one first calculates the complete equation involving all k regressors. At each subsequent stage of the procedure the minimum partial F value amongst the included variables is compared with a critical value. If this value is not attained the corresponding variable is excluded from the equation, otherwise the procedure terminates.

(c) The General Stepwise Regression Procedure

This is the procedure envisaged by Efroymson in his original paper, and is also the one most commonly encountered in statistical computer packages. In particular the widely used 'BIOMED' stepwise routine [11] is of this type. The procedure begins like a forward procedure, but at each stage variables are first checked for possible deletion against an appropriate critical value. If no variable qualifies for deletion the excluded variables are searched for the one yielding the maximum F value. This is then compared with another critical value as in a forward procedure. The whole procedure terminates when no more variables can be deleted or entered.

The general procedure is motivated by the fact that it is possible for previously entered variables to become 'insignificant' at a later stage when further variables have been introduced. This is really just a consequence of the presence of multicollinearity amongst the regressors, and could not occur in an orthogonal regression set-up for example.

A particularly disturbing feature of the procedures described above is that, notwithstanding the problem of choosing suitable critical levels, the procedures cannot in general be expected to produce the same final equations. This is so even in the relatively more manageable situation of orthogonality. Discussions in the literature on the relative merits of the various procedures tend to place more emphasis on the amounts of computation needed rather than on the attainment of properly formulated objectives. There is also a notable lack of discussion on the problem of the choice of critical values for the F values calculated. Mostly, when the subject is referred to at all, it is merely to imply that conventional critical values are appropriate i.e. as for regression tests in a non-sequential situation. The validity of this approach (or rather its invalidity) underlies much of the later discussion.

We now briefly look at a few other procedures which, whilst they are not based on the stepwise algorithm, relate to the problem of choosing optimal regression equations. For further discussion on the use of stepwise regression variants in practice see Draper and Smith [20].

2.2. Other methods for selecting optimal regression equations.

Although the centre of interest of the present investigation is concerned with procedures stemming from the basic stepwise algorithm a brief review of some alternative approaches is perhaps not out of place.

(i) Stagewise Regression

In this procedure one obtains the residuals from the regression equation at any stage and correlates these with each of the excluded variables. The variable having maximum correlation is then entered. A theoretical treatment of this method is to be found in Goldberger [27] under the heading of 'stepwise' regression. The main disadvantage of the method is that the resulting least squares coefficient estimates will be biased. From the point of view of constructing a meaningful inferential basis for such a procedure this drawback would seem to be insurmountable. It might be remarked however that the preliminary removal of trend or seasonal components by regression methods applied to time series data is in fact an application of stagewise regression. The subsequent analysis of the residuals thus obtained is often carried out as though no such prior adjustment had been performed. If the residuals analysis just involves the fitting of a regression equation of some kind then both phases of the analysis could be incorporated into a single stepwise run. Any variables which on a priori grounds are thought to be essential members of an equation can easily be forced into the regression initially before the stepwise selection process is initiated. The introduction of these variables will of course have to be accompanied by the corresponding operations on the matrix $\underline{A}$.

(ii) All possible regressions

The feasibility of performing all possible regressions involving k variables is obviously going to depend on the availability of large computing facilities. Since there are $2^k - 1$ possible equations a value of k as small as 10 will involve as many as 1023 different models. A paper by Garside [26] gives a simplified procedure for calculating all possible regressions which even then is only practicable for values of k up to about 12. Because of this limitation Beale, Kendall and Mann [10] developed a computational algorithm which eliminates the calculation of regressions which cannot possibly be better (in terms of R^2) than other previously calculated equations of the same order. Since there is no guarantee as to the amount of time this saves in any particular situation it is suggested that the procedure is terminated after a pre-determined number of regressions have been calculated. A final decision is then made from this subset on purely subjective grounds.

An almost identical procedure for computing optimal regression subsets is given by Hocking and Leslie [34]. They also suggest the use of Mallows C_p statistic as a decision criterion for the final equation (this statistic is discussed later).

A further paper, by Schatzoff, Tsao and Fienberg [71], described a more efficient algorithm for the calculation of all possible regressions on the lines of Garside's paper.

In each of the above approaches the emphasis is very definitely on the computational rather than the inferential side of the problem. Whilst they have the advantage over stepwise techniques of yielding, for a given order of equation, the regression with highest R^2 they do not admit to a sequential decision treatment so easily. For this reason, and also because the use of these more general techniques is as yet nowhere near as widespread as that of stepwise methods, they will not be pursued further in this study.

(iii) The Newton and Spurrell Method

The problem of selecting multiple regression relationships is attacked in a direct way by Newton and Spurrell [62] by a proposed disentanglement of the effect on the regression of correlations between the regressors. They suggest that a prime objective, whether in a prediction or control context, is to seek regressors which have strong 'independent' effects on the regressand. This gives rise to the consideration of so-called 'elements' which can be interpreted as being representative of the amounts of individual information which variables bring into the model.

To illustrate their approach we can consider the simple model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where X_1 and X_2 are possibly highly correlated. It is well-known that a test of the hypothesis

$$H_o \equiv \beta_1 = \beta_2 = 0$$

can be found to be significant while, at the same time, the two sub-hypotheses

$$H_o' \equiv \beta_1 = 0$$

$$H_o'' \equiv \beta_2 = 0$$

are individually not significant.

The Newton-Spurrell method involves calculating the <u>Primary Elements</u>, x_1 and x_2 say, where x_1 is the extra sum of squares due to X_1 given X_2 has been fitted and similarly for x_2. If X_1 alone is fitted one gets the so-called <u>Secondary Element</u> $x_1 x_2$, where $x_1 + x_1 x_2$ is the sum of squares due to X_1 alone, omitting X_2.

This generalises to higher order regressions in which any non-primary element is referred to as a secondary element. Newton and Spurrell suggest a heuristic procedure in which one looks for variables with small secondary elements compared with primary ones. Whilst it is recognised that such an approach could possibly be very useful when carried out by experienced users, it does not readily lend itself to routine application. The two authors do indeed admit that an objective statistical treatment of the procedure seems out of the question.

(iv) The use of principal components

The use of principal components is especially appealing in regression situations due to the othogonality present. There have been some suggestions made for performing regression analysis along these lines, and these will be discussed more conveniently later on.

Chapter 3.

Some Extensions of Classical Regression Theory

3.1 Development of main results

The whole of this section is devoted to the extension of results of classical regression to a wider context than is usually considered in the literature. These results will be of fundamental interest in the development of sequential tests in stepwise regression.

We consider again the basic underlying model

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \qquad (1)$$

where in particular $\underset{\sim}{\beta}$ is a k-dimensional vector of unknown coefficients. We suppose that the model has a 'true' order p in the sense that k − p of the elements of $\underset{\sim}{\beta}$ are zero. We correspondingly partition $\underset{\sim}{\beta}$ in the form

$$\underset{\sim}{\beta}' = [\underset{\sim}{\beta_1}' \ \underset{\sim}{B_2}']$$

where $\underset{\sim}{\beta_2} = \underset{\sim}{0}$, and without loss of generality the p true regressor variables are taken to be the first p columns of $\underset{\sim}{X}$. We can therefore partition $\underset{\sim}{X}$ in the form

$$X = [\underset{\sim}{X_1} \ \underset{\sim}{X_2}]$$

where $\underset{\sim}{X_1}$ is $n \times p$

$\underset{\sim}{X_2}$ is $n \times (k-p)$

Suppose that, using a stepwise procedure, the stage has been reached at which r regressors have been includedin a fitted model. We will later be concerned with investigating the hypothesis that all the true variables havealready been entered but for the time being we will derive some results of a more general nature.

Corresponding to their included variables is a matrix $\underset{\sim}{Z}_1$ consisting of a subset of r columns of $\underset{\sim}{X}$. We can then write, again after a suitable re-ordering of the columns of $\underset{\sim}{X}$,

$$\underset{\sim}{X} = [\underset{\sim}{Z}_1 \quad \underset{\sim}{Z}_2]$$

where $\underset{\sim}{Z}_1$ is $n \times r$

$\underset{\sim}{Z}_2$ is $n \times (k-r)$

Suppose that, for each of the $k-r$ variables in $\underset{\sim}{Z}_2$, we calculate the partial regression coefficient of Y on this variable, holding fixed the $\underset{\sim}{Z}_1$ variables. We will denote these calculated values by d_i, $i = 1,\ldots,(k-r)$. It follows immediately from standard least squares theory that

$$d_i = [(\underset{\sim}{Z}^{*\prime} \underset{\sim}{Z}^{*})^{-1} \underset{\sim}{Z}^{*\prime} \underset{\sim}{Y}]_{(r+1)}$$

where $\underset{\sim}{Z}^{*} = [\underset{\sim}{Z}_1 \quad \underset{\sim}{Z}_2^{(i)}]$

and where we use the notation that $\underset{\sim}{B}_{(q)}, \underset{\sim}{B}^{(q)}$ denote respectively the q^{th} row and column of a matrix $\underset{\sim}{B}$.

Substituting $\underset{\sim}{Y}$ from (1) gives, recalling $\underset{\sim}{\beta}_2 = \underset{\sim}{0}$,

$$d_i = [(\underset{\sim}{Z}^{*\prime}\underset{\sim}{Z}^*)^{-1}\underset{\sim}{Z}^{*\prime}(\underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon})]_{(r+1)}$$

$$= (\underset{\sim}{Z}^{*\prime}\underset{\sim}{Z}^*)^{-1}_{(r+1)}\underset{\sim}{Z}^{*\prime}(\underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon}) \qquad (2)$$

Now

$$\underset{\sim}{Z}^{*\prime}\underset{\sim}{Z}^* = \begin{bmatrix} \underset{\sim}{Z}_1'\underset{\sim}{Z}_1 & \underset{\sim}{Z}_1'\underset{\sim}{Z}_2^{(i)} \\ \underset{\sim}{Z}_2^{(i)\prime}\underset{\sim}{Z}_1 & \underset{\sim}{Z}_2^{(i)\prime}\underset{\sim}{Z}_2^{(i)} \end{bmatrix}$$

Using Theorem 1.4.1 it follows that

$$(\underset{\sim}{Z}^{*\prime}\underset{\sim}{Z}^*)^{-1}_{(r+1)} = f_i^{-1}[-\underset{\sim}{Z}_2^{(i)\prime}\underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1} \quad 1] \qquad (3)$$

where the scalar $f_i = \underset{\sim}{Z}_2^{(i)\prime}[\underset{\sim}{I} - \underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1']\underset{\sim}{Z}_2^{(i)}$

$$= \underset{\sim}{Z}_2^{(i)\prime}\underset{\sim}{M}\underset{\sim}{Z}_2^{(i)}, \text{ say.} \qquad (4)$$

Note that $f_i = \underset{\sim}{Z}_2^{(i)\prime}\underset{\sim}{M}'\underset{\sim}{M}\underset{\sim}{Z}_2^{(i)}$, since $\underset{\sim}{M}$ is symmetric and idempotent. Writing $\underset{\sim}{u} = \underset{\sim}{M}\,\underset{\sim}{Z}_2^{(i)}$ it follows that $f_i = 0$ if and only if $\underset{\sim}{u} = \underset{\sim}{0}$ which implies that

$$\underset{\sim}{Z}_2^{(i)} = \underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'\underset{\sim}{Z}_2^{(i)} = \underset{\sim}{Z}_1\underset{\sim}{C}, \text{ say.}$$

This contradicts the full-rank assumption for the matrix $\underset{\sim}{X}$, hence it follows that $f_i > 0$. In fact f_i is, up to a factor n, just the residual variance of $\underset{\sim}{Z}_2^{(i)}$ taking out the effect of $\underset{\sim}{Z}_1$.

It now follows that, using (3), we can rewrite (2) as

$$d_i = f_i^{-1}\{-\underset{\sim}{Z}_2^{(i)\prime}\underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'(\underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon}) + \underset{\sim}{Z}_2^{(i)\prime}(\underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon})\}$$

$$= f_i^{-1}\underset{\sim}{Z}_2^{(i)\prime}\underset{\sim}{M}(\underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon})$$

If we let $\underset{\sim}{d}$ be the vector such that

$$\underset{\sim}{d}' = [d_1 \, d_2 \ldots d_{k-r}]$$

then

$$\underset{\sim}{d} = \underset{\sim}{F}^{-1} \underset{\sim}{Z}_2' \underset{\sim}{M}(\underset{\sim}{X}_1 \underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon}) \tag{5}$$

where $\underset{\sim}{F}^{-1}$ is the diagonal matrix with i^{th} diagonal term f_i^{-1}.

We now turn to consideration of the properties of $\underset{\sim}{d}$. In the most general sense the distribution of $\underset{\sim}{d}$ will depend very much on how the various fitted regressions are arrived at. In particular we are only really justified in regarding $\underset{\sim}{d}$ as being multivariate normal in the case where $r = k$. However while it is conceded that stepwise regression, being a sequential solution to a multiple decision problem, should properly involve considerations of this type the practicality of the problem forces on us a more restricted approach. The approach used is in fact to treat the stepwise process as a sequence of conditional hypothesis tests. Within this framework it is argued later (in context) that the normality assumption for $\underset{\sim}{d}$ is reasonable. For the purposes of the remainder of this chapter we should regard the r fitted regressors as having been chosen arbitrarily from the full set of k regressors available.

Since $E[\underset{\sim}{\varepsilon}] = \underset{\sim}{0}$ it follows from (5) that

$$E[\underset{\sim}{d}] = \underset{\sim}{F}^{-1} \underset{\sim}{Z}_2' \underset{\sim}{M} \, \underset{\sim}{X}_1 \underset{\sim}{\beta}_1 = \underset{\sim}{\mu}_d, \text{ say.}$$

Also, the covariance matrix of $\underset{\sim}{d}$ is given by

$$\underset{\sim}{V}_d = E[(\underset{\sim}{d}-\underset{\sim}{\mu}_d)(\underset{\sim}{d}-\underset{\sim}{\mu}_d)']$$

$$= E[\underset{\sim}{F}^{-1}\underset{\sim}{Z}_2'\underset{\sim}{M}\underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}'\underset{\sim}{M}'\underset{\sim}{Z}_2\underset{\sim}{F}^{-1}{}']$$

$$= \sigma^2\underset{\sim}{F}^{-1}\underset{\sim}{Z}_2'\underset{\sim}{M}\underset{\sim}{Z}_2\underset{\sim}{F}^{-1},$$ since $\underset{\sim}{M}$ is symmetric and indempotent. It follows that the d_i are uncorrelated, and hence independent, if and only if,

$$\underset{\sim}{V}^* = \underset{\sim}{Z}_2'\underset{\sim}{M}\,\underset{\sim}{Z}_2$$

is diagonal.

Since the matrix $\underset{\sim}{V}^*$ is, apart from a multiplying constant, the sample partial covariance matrix of the $\underset{\sim}{Z}_2$ variables holding fixed $\underset{\sim}{Z}_1$ it will be convenient to refer to the above condition as'partial orthogonality'.

It has been established above that $\underset{\sim}{d}$ is $N(\underset{\sim}{\mu}_d, \underset{\sim}{V}_d)$ where

$$\underset{\sim}{\mu}_d = \underset{\sim}{F}^{-1}\underset{\sim}{Z}_2'\underset{\sim}{M}\,\underset{\sim}{X}_1\underset{\sim}{\beta}_1 \qquad (6)$$

$$\underset{\sim}{V}_d = \sigma^2\underset{\sim}{F}^{-1}\underset{\sim}{Z}_2'\underset{\sim}{M}\,\underset{\sim}{Z}_2\underset{\sim}{F}^{-1} \qquad (7)$$

It will now be determined under what conditions $\underset{\sim}{\mu}_d = \underset{\sim}{0}$. Suppose that we sub-partition $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$ in the form

$$\underset{\sim}{X}_1 = [\underset{\sim}{X}_{11}\ \underset{\sim}{X}_{12}]$$

$$\underset{\sim}{X}_2 = [\underset{\sim}{X}_{21}\ \underset{\sim}{X}_{22}]$$

where $\underset{\sim}{Z}_1 = [\underset{\sim}{X}_{12}\ \underset{\sim}{X}_{22}]$, $\underset{\sim}{Z}_2 = [\underset{\sim}{X}_{11}\ \underset{\sim}{X}_{21}]$.

i.e. $\underset{\sim}{X}_{12}$ and $\underset{\sim}{X}_{11}$ represent the 'true' regressor variables belonging to the fitted set $\underset{\sim}{Z}_1$ and unfitted set $\underset{\sim}{Z}_2$ respectively.

Writing $Z_1'Z_1 = \begin{bmatrix} \underset{\sim}{X}_{12}'\underset{\sim}{X}_{12} & \underset{\sim}{X}_{12}'\underset{\sim}{X}_{22} \\ \underset{\sim}{X}_{22}'\underset{\sim}{X}_{12} & \underset{\sim}{X}_{22}'\underset{\sim}{X}_{22} \end{bmatrix}$

and using Theorem 1.4.1 we obtain

$$(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1} =$$

$$\begin{bmatrix} (\underset{\sim}{X}_{12}'\underset{\sim}{X}_{12})^{-1}(\underset{\sim}{I}+\underset{\sim}{X}_{12}'\underset{\sim}{X}_{22}\underset{\sim}{D}^{-1}\underset{\sim}{X}_{22}'\underset{\sim}{X}_{12}(\underset{\sim}{X}_{12}'\underset{\sim}{X}_{12})^{-1}) & -(\underset{\sim}{X}_{12}'\underset{\sim}{X}_{12})^{-1}\underset{\sim}{X}_{12}'\underset{\sim}{X}_{22}\underset{\sim}{D}^{-1} \\ -\underset{\sim}{D}^{-1}\underset{\sim}{X}_{22}'\underset{\sim}{X}_{12}(\underset{\sim}{X}_{12}'\underset{\sim}{X}_{12})^{-1} & \underset{\sim}{D}^{-1} \end{bmatrix}$$

where $\underset{\sim}{D} = \underset{\sim}{X}_{22}'(\underset{\sim}{I}-\underset{\sim}{X}_{12}(\underset{\sim}{X}_{12}'\underset{\sim}{X}_{12})^{-1}\underset{\sim}{X}_{12}')\underset{\sim}{X}_{22}$

$= \underset{\sim}{X}_{22}'\underset{\sim}{M}^{*}\underset{\sim}{X}_{22}$, say.

Hence $\underset{\sim}{M} = \underset{\sim}{I} - \underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'$

$= \underset{\sim}{I} - \underset{\sim}{X}_{12}(\underset{\sim}{X}_{12}'\underset{\sim}{X}_{12})^{-1}\underset{\sim}{X}_{12}' - \underset{\sim}{X}_{12}(\underset{\sim}{X}_{12}'\underset{\sim}{X}_{12})^{-1}\underset{\sim}{X}_{12}'\underset{\sim}{X}_{22}\underset{\sim}{D}^{-1}\underset{\sim}{X}_{22}'\underset{\sim}{X}_{12}$

$(\underset{\sim}{X}_{12}'\underset{\sim}{X}_{12})^{-1}\underset{\sim}{X}_{12}' + \underset{\sim}{X}_{12}(\underset{\sim}{X}_{12}'\underset{\sim}{X}_{12})^{-1}\underset{\sim}{X}_{12}'\underset{\sim}{X}_{22}\underset{\sim}{D}^{-1}\underset{\sim}{X}_{22}'$

$+ \underset{\sim}{X}_{22}\underset{\sim}{D}^{-1}\underset{\sim}{X}_{22}'\underset{\sim}{X}_{12}(\underset{\sim}{X}_{12}'\underset{\sim}{X}_{12})^{-1}\underset{\sim}{X}_{12}' - \underset{\sim}{X}_{22}\underset{\sim}{D}^{-1}\underset{\sim}{X}_{22}'$.

After factorisation this reduces to

$$\begin{aligned} \underset{\sim}{M} &= \underset{\sim}{M}^* - \underset{\sim}{M}^*\underset{\sim}{X}_{22}\ \underset{\sim}{D}^{-1}\ \underset{\sim}{X}_{22}'\ \underset{\sim}{M}^* \\ &= \underset{\sim}{M}^*(\underset{\sim}{I}-\underset{\sim}{X}_{22}\underset{\sim}{D}^{-1}\ \underset{\sim}{X}_{22}')M^*, \text{ since } M^* \text{ is idempotent.} \end{aligned} \qquad (8)$$

Now suppose that $\underset{\sim}{Z}_1$ incorporates all the columns of $\underset{\sim}{X}_1$, i.e. all the true variables have been fitted. Then $\underset{\sim}{X}_{12} = \underset{\sim}{X}_1$, and

$$\underset{\sim}{M}^* = \underset{\sim}{I} - \underset{\sim}{X}_1\ (\underset{\sim}{X}_1'\underset{\sim}{X}_1\ )^{-1}\ \underset{\sim}{X}_1'$$

Hence $\underset{\sim}{M}\ \underset{\sim}{X}_1 = \underset{\sim}{M}^*(\underset{\sim}{I}-\underset{\sim}{X}_{22}\ \underset{\sim}{D}^{-1}\ \underset{\sim}{X}_{22}')(\underset{\sim}{I}-\underset{\sim}{X}_1\ (\underset{\sim}{X}_1'\underset{\sim}{X}_1\ )^{-1}\ \underset{\sim}{X}_1')\underset{\sim}{X}_1$

$$= \underset{\sim}{0}$$

and, from (6), it follows that $\underset{\sim}{\mu}_d = \underset{\sim}{0}$.

Conversely, now suppose that not all the true variables have been included i.e. $\underset{\sim}{X}_{11}$ exists. Noting that $\underset{\sim}{\mu}_d = \underset{\sim}{F}^{-1}\ \underset{\sim}{Z}_2'\underset{\sim}{MX}_1\ \underset{\sim}{\beta}_1 = 0$ if and only if $\underset{\sim}{Z}_2'\underset{\sim}{MX}_1\ \underset{\sim}{\beta}_1 = \underset{\sim}{0}$ we can partition $\underset{\sim}{\beta}_1$ in the form

$$\underset{\sim}{\beta}_1' = [\underset{\sim}{\beta}_{11}'\ \ \underset{\sim}{\beta}_{12}'\ ]$$

where $\underset{\sim}{\beta}_{11}$ contains the regression coefficients corresponding to $\underset{\sim}{X}_{11}$, and $\underset{\sim}{\beta}_{12}$ corresponds to $\underset{\sim}{X}_{12}$.

We can then write

$$\underset{\sim}{Z}_2'\ \ \underset{\sim}{MX}_1\ \underset{\sim}{\beta}_1 = \begin{bmatrix} \underset{\sim}{X}_{11}' \\ \underset{\sim}{X}_2' \end{bmatrix} \quad \underset{\sim}{M}[\underset{\sim}{X}_{11}\ \ \underset{\sim}{X}_{12}\ ]\begin{bmatrix} \underset{\sim}{\beta}_{11} \\ \underset{\sim}{\beta}_{12} \end{bmatrix}$$

Now $\underset{\sim}{M}[\underset{\sim}{X}_{11}\underset{\sim}{X}_{12}] = [\underset{\sim}{MX}_{11}\ \ \underset{\sim}{MX}_{12}]$. Further, from (8),

$$\utilde{MX}_{12} = \utilde{M}^*(\utilde{I}-\utilde{X}_{22}\;\utilde{D}^{-1}\;\utilde{X}_{22}')\utilde{M}^*\utilde{X}_{12}$$

But $\utilde{M}^*\utilde{X}_{12} = (\utilde{I}-\utilde{X}_{12}\;(\utilde{X}_{12}'\;\utilde{X}_{12})^{-1}\,\utilde{X}_{12}'\,)\utilde{X}_{12} = \utilde{0}$

Hence,

$$\utilde{Z}_2'\utilde{MX}_1\,\utilde{\beta}_1 = \begin{bmatrix} \utilde{X}_{11}' \\ \\ \utilde{X}_{21}' \end{bmatrix} \utilde{MX}_{11}\;\utilde{\beta}_{11} = \begin{bmatrix} \utilde{X}_{11}'\;\utilde{MX}_{11} \\ \\ \utilde{X}_{21}'\;\utilde{MX}_{11} \end{bmatrix} \utilde{\beta}_{11} \tag{9}$$

Now $\utilde{X}_{11}'\;\utilde{MX}_{11} = \utilde{X}_{11}'\;(\utilde{I}-\utilde{Z}_1\,(\utilde{Z}_1'\utilde{Z}_1\,)^{-1}\,\utilde{Z}_1')\utilde{X}_{11}$ is, apart from a multiplying constant, the sample partial covariance matrix for the $\utilde{X}_{11}$ variables holding fixed $\utilde{X}_{12}$ and $\utilde{X}_{22}$. It follows that this matrix is non-singular, and since $\utilde{\beta}_{11} \neq \utilde{0}$, we have

$$\utilde{X}_{11}'\;\utilde{MX}_{11}\;\utilde{\beta}_{11} \neq \utilde{0} \tag{10}$$

Thus the elements of $\utilde{\mu}_d$ corresponding to the excluded set of variables $\utilde{X}_{11}$ are not all zero. It is important to note that some such elements of $\utilde{\mu}_d$ can in fact be zero, as is illustrated by the following simple example:

Suppose $\utilde{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$

and take $\beta_1 = \beta_4 = 0$, $\beta_2 = -1$, $\beta_3 = \frac{3}{2}$. Then, taking $\utilde{Z}_1$ to consist of the single variable X_1 , we obtain

$$\utilde{X}_{11}'\;\utilde{MX}_{11} = \begin{bmatrix} \frac{3}{2} & 1 \\ \\ 1 & 2 \end{bmatrix}$$

The expression (10) in this case is equal to

$$\begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

The implications of this result (in its generality) will be discussed later in relation to stepwise regression applied to a non-orthoginal set of regressors.

With regard to the other component in (9), the matrix $\underset{\sim}{X}_{21}' \underset{\sim}{MX}_{11}$ is not so easily dealt with. For it is in fact the matrix of(calculated) partial covariances between the sets $\underset{\sim}{X}_{21}$ and $\underset{\sim}{X}_{11}$,holding fixed the $\underset{\sim}{Z}_{1}$ set, and is not even square in general. That the vector itself given by (9) is not necessarily null is easily demonstrated using the above example. Perhaps more importantly some or all of its elements can exceed any of those in the vector at (10). We find, indeed, that in the above example

$$\underset{\sim}{X}_{21}' \; \underset{\sim}{MX}_{11} \; \underset{\sim}{\beta}_{11} = 3 \qquad \text{'11)}$$

We return to a discussion of this point in a later chapter.

One final point can be made concerning the expression at (9). This is that it is quite easy to show that, in the special case in which the columns of $\underset{\sim}{X}$ are mutually orthogonal, the component expressions of (9) are given by

$$\underset{\sim}{X}'_{11}\ \underset{\sim}{M}\ \underset{\sim}{X}_{11}\beta_{11} = \underset{\sim}{F}\ \beta_{11}$$

$$\underset{\sim}{X}_{21}\ \underset{\sim}{M}\ \underset{\sim}{X}_{11}\beta_{11} = \underset{\sim}{0}.$$

We therefore merely obtain the standard result for orthogonal regression,

$$E[\underset{\sim}{d}] = \underset{\sim}{F}^{-1}\underset{\sim}{Z}'_2\underset{\sim}{M}\ \underset{\sim}{X}_1\beta_1 = \begin{bmatrix} \beta_{11} \\ \underset{\sim}{0} \end{bmatrix}$$

It will be useful at this stage to standardize each d_i by dividing by its standard error. Noting that, from (7), the variance of d_i is

$$\sigma_{d_i} = \sigma^2_{d_i} f_i^{-2}\underset{\sim}{Z}_2^{(i)'}\ \underset{\sim}{M}\ \underset{\sim}{Z}_2^{(i)}$$

$$= \sigma^2 f_i^{-1}\ \text{, from (4)}$$

then

$$d_i = d_i/\sigma_{d_i} = (\sigma^2 f_i)^{-\frac{1}{2}}\underset{\sim}{Z}^{(i)'}\underset{\sim}{M}\ (\underset{\sim}{X}_1\beta_1 + \underset{\sim}{\varepsilon}) \qquad (12)$$

We can then write

$$\underset{\sim}{d}^* = \sigma^{-1}\ \underset{\sim}{F}^{\frac{1}{2}}\ \underset{\sim}{d}.$$

We see that

$$\text{Var}[\underset{\sim}{d}^*] = \underset{\sim}{F}^{\frac{1}{2}}\ \underset{\sim}{F}^{-1}\underset{\sim}{Z}'_2\underset{\sim}{M}\ \underset{\sim}{Z}_2\underset{\sim}{F}^{-1}\ \underset{\sim}{F}^{\frac{1}{2}}$$

$$= \underset{\sim}{F}^{-\frac{1}{2}}\ \underset{\sim}{Z}'_2\ \underset{\sim}{M}\ \underset{\sim}{Z}_2\ \underset{\sim}{F}^{-\frac{1}{2}}$$

$$= \underset{\sim}{\Omega}\text{, say} \qquad (13)$$

It is shown presently that selection of the $\underset{\sim}{Z}_2^{(i)}$ variable having largest contribution to explained sum of squares is equivalent to selecting the variable with highest d_i^* in modulus. It turns out easier, at least in the orthogonal case, to work in terms of the maximum d_i^{*2}, but the results developed above will be of fundamental importance to most of the subsequent discussion.

We return now to consideration of the increase in sum of squares due to a $\underset{\sim}{Z}_2^{(i)}$ variable being pivoted into the regression equation.

First, consider the residual sum of squares after $\underset{\sim}{Z}_1$ has been fitted. If we let $\underset{\sim}{C}$ be the vector of estimates of the coefficients in this equation we can write the residual sum of squares as

$$(\underset{\sim}{Y}-\underset{\sim}{Z}_1\underset{\sim}{C})'(\underset{\sim}{Y}-\underset{\sim}{Z}_1\underset{\sim}{C}) = \underset{\sim}{P}'\underset{\sim}{P}, \text{ say,}$$

where $\underset{\sim}{P} = (\underset{\sim}{X}_1\underset{\sim}{\beta}_1+\underset{\sim}{\varepsilon}-\underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'(\underset{\sim}{X}_1\underset{\sim}{\beta}_1+\underset{\sim}{\varepsilon}))$

$$= (\underset{\sim}{I}-\underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1')(\underset{\sim}{X}_1\underset{\sim}{\beta}_1+\underset{\sim}{\varepsilon})$$

$$= \underset{\sim}{M}(\underset{\sim}{X}_1\underset{\sim}{\beta}_1+\underset{\sim}{\varepsilon}).$$

Hence the residual sum of squares is

$$(\underset{\sim}{X}_1\underset{\sim}{\beta}_1+\underset{\sim}{\varepsilon})'\underset{\sim}{M}(\underset{\sim}{X}_1\underset{\sim}{\beta}_1+\underset{\sim}{\varepsilon}).$$

In the same way the residual sum of squares after fitting $\underset{\sim}{Z}_1$ and $\underset{\sim}{Z}_2^{(i)}$ is

$$(\underset{\sim}{X}_1\underset{\sim}{\beta}_1+\underset{\sim}{\varepsilon})'\underset{\sim}{M}^{**}(\underset{\sim}{X}_1\underset{\sim}{\beta}_1+\underset{\sim}{\varepsilon})$$

where $\underset{\sim}{M}^{**} = \underset{\sim}{I} - \underset{\sim}{Z}^*(\underset{\sim}{Z}^{*\prime}\underset{\sim}{Z}^*)^{-1}\underset{\sim}{Z}^{*\prime}$, and $\underset{\sim}{Z}^*$ is as defined previously.

The <u>extra</u> sum of squares contributed by $\underset{\sim}{Z}_2^{(i)}$ is then

$$(\underset{\sim}{X}_1 \underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon})'(\underset{\sim}{M}-\underset{\sim}{M}^{**})(\underset{\sim}{X}_1 \underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon})$$

Using the partitioned form of $\underset{\sim}{Z}^*$ and Theorem 1.4.1 we can now write

$$\begin{aligned}
\underset{\sim}{M} - \underset{\sim}{M}^{**} &= \underset{\sim}{Z}^*(\underset{\sim}{Z}^{*\prime}\underset{\sim}{Z}^*)^{-1}\underset{\sim}{Z}^{*\prime} - \underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1' \\
&= f_i^{-1}\{\underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'\underset{\sim}{Z}_2^{(i)}\underset{\sim}{Z}_2^{(i)\prime}\underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1' \\
&\qquad - \underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'\underset{\sim}{Z}_2^{(i)}\underset{\sim}{Z}_2^{(i)\prime} \\
&\qquad - \underset{\sim}{Z}_2^{(i)}\underset{\sim}{Z}_2^{(i)\prime}\underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1' + \underset{\sim}{Z}_2^{(i)}\underset{\sim}{Z}_2^{(i)\prime}\} \\
&= f_i^{-1}\underset{\sim}{M}\,\underset{\sim}{Z}_2^{(i)}\underset{\sim}{Z}_2^{(i)\prime}\underset{\sim}{M} = \underset{\sim}{E}_i, \quad \text{say.}
\end{aligned}$$

Reverting to expression (12) for d_i^* we see that

$$(d_i^*)^2 = (\sigma^2)^{-1}(\underset{\sim}{X}_1 \underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon})'\underset{\sim}{E}_i(\underset{\sim}{X}_1 \underset{\sim}{\beta}_1 + \underset{\sim}{\varepsilon})$$

i.e.

$\sigma^2(d_i^*)^2$ is the extra sum of squares due to fitting $\underset{\sim}{Z}_2^{(i)}$, $= \underset{\sim}{S}_i^2$, say.

We are now in a position to write down the joint density function of the S_i^2, for $i = 1, \ldots, (k-v)$, in termsof the known distribution of d^*. In particular we can note some special results :-

(i) If $E[d_i] = 0$ then S_i^2 is distributed as $\sigma^2\chi^2_{(1)}$. (where $\chi^2_{(\nu)}$ denotes a random variable having chi-square distribution with ν degree of freedom).

(ii) If $E[d_i] = E[d_j] = E[d_i d_j] = 0$ then S_i^2 and S_j^2 are independent random variables each being distributed as $\sigma^2 \chi^2_{(1)}$.

(iii) If $E[d_i] = 0$ then S_i^2/σ^2 is distributed as a non-central chi-square random variable.

3.2 Estimation of the error variance

Before bringing together the main results developed in the previous section, and interpreting their relevance to stepwise regression, it is apparent that any test of the hypothesis that $E[d_i] = 0$ will require knowledge of σ^2, or at least a suitable estimate of it. We now show that such an estimate is provided by the residual variance v_k obtained from fitting all k regressors.

The actual distribution of v_k is known from standard theory to be that of $\sigma^2 \chi^2_{(n-k-1)}/(n-k-1)$.
We now show that v_k and S_i^2 are independent $(i = 1, \ldots, (k-r)$; $r = 0, \ldots, (k-1))$. Again this result is virtually immediate from standard regression theory, but it is felt that a more general proof is not out of place here.

It is sufficient to show that the explained sum of squares due to fitting <u>any</u> subset $\underset{\sim}{Z}$ of columns of $\underset{\sim}{X}$ is independent of v_k. The increase in sum of squares due to fitting $\mathbf{X}_2^{(i)}$ will then be independent of v_k since it can be written as the difference between the explained sums of squares due to $\mathbf{Z}_1$ and $\mathbf{Z}^*$ respectively.

Without loss of generality we again use the partition

$$\mathbf{X} = [\mathbf{Z}_1\, \mathbf{Z}_2].$$

The sum of squares due to $\mathbf{Z}_1$ is then

$$(\mathbf{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon})'\mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1'(\mathbf{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon})$$

$$= (\mathbf{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon})'\mathbf{A}(\mathbf{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}), \text{ say.}$$

Now $(n-k-1)v_k = \boldsymbol{\varepsilon}'(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}'\mathbf{B}\boldsymbol{\varepsilon}$, say.

Using the above partition for $\mathbf{X}$ we obtain

$$\mathbf{B} = \mathbf{I} - \mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1' - \mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1'\mathbf{Z}_2\mathbf{G}^{-1} \mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1'$$

$$+ \mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1'\mathbf{Z}_2\mathbf{G}^{-1} \mathbf{Z}_2' + \mathbf{Z}_2\mathbf{G}^{-1} \mathbf{Z}_2'\mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1 /$$

$$- \mathbf{Z}_2\mathbf{G}^{-1} \mathbf{Z}_2'$$

where $\mathbf{G} = \mathbf{Z}_2'(\mathbf{I}-\mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1')\mathbf{Z}_2$

Therefore

$$\mathbf{AB} = \mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1' - \mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1' - \mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1'\mathbf{Z}_2\mathbf{G}^{-1}$$

$$\mathbf{Z}_2'\mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1' + \mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1'\mathbf{Z}_2\mathbf{F}^{-1} \mathbf{Z}_2' +$$

$$\mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1'\mathbf{Z}_2\mathbf{G}^{-1} \mathbf{Z}_2'\mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1' - \mathbf{Z}_1 (\mathbf{Z}_1'\mathbf{Z}_1)^{-1} \mathbf{Z}_1'\mathbf{Z}_2\mathbf{G}^{-1} \mathbf{Z}_2'$$

$$= 0$$

Hence, by Craig's Theorem on idempotent quadratic forms in normal variables, it follows that v_k and S_i^2 are independent.

3.3 Some fundamental distributions

To conclude this chapter the main results developed above are brought together in a form relevant to stepwise regression applications. The basic procedure structure described below will be discussed further in later chapters, our interest here being of a preliminary nature.

It will be convenient to assume that only a strict uni-directional procedure is being considered, i.e. either forward or backward. The distributional aspects which arise will however be seen to carry over to the context of mixed forward/backward procedures of the general type.

In the case of a forward procedure suppose that the stage has been reached at which r variables have been fitted. The fundamental problem is regarded as being one of making the decision as to either

1. Enter another variable

or 2. Terminate the procedure.

If decision 1. is made then, in line with conventional stepwise procedures, the excluded variable which yields the maximum increase in explained sum of squares will be pivoted into the regression equation. As shown in section 1 above, this variable will be $\underline{Z}_2^{(i)}$ for which $|d_i^*|$ is a maximum.

Alternatively, in a backward procedure, decision 1. now becomes 'Delete another variable', it being natural to then sslect the variable which results in the least decrease in the explained sum of squares. Suppose that this variable is pivoted out of the equation, and consider the now enlarged set of excluded variables. We now immediately see a difficulty which could easily be expected to arise in the case of a non-orthoginal regression. This is that the variable just excluded will not necessarily be the one which would then yield the maximum contribution to the regression sum of squares if it were to be re-entered. This will however be so in the case of an orthogonal regression situation. In such a case the decision problem involved can be identified exactly with that arising in a forward procedure.

The distinction between the orthogonal and non-orthogonal cases, alluded to above, will continue to be necessary throughout the whole of the remaining discussion. This reflects both the theoretical complexity involved, and also perhaps the different degrees of faith one should have in using single-step procedures in the two cases. For the time being we restrict the discussion to uni-directional procedures applied to athogonal regressor variables or forward procedures in non-orthogonal situations.

We suppose that r variables have been fitted, and consider the problem of whether any excluded variables are worth entering. Within the conditional decision framework being contemplated the relevant information for making our decision is summarised by the vector $\underset{\sim}{d}^*$ and its correlation matrix $\underset{\sim}{\Omega}$. Under the assumption that no further variables are worth entering (either in isolation or in groups) the results of section 1 above imply that $\underset{\sim}{d}^*$ has zero mean. The decision to proceed or not can then be identified with the test of the hypothesis that $\underset{\sim}{\mu}_d = \underset{\sim}{0}$. The tendency of the resulting complete procedure to incorrectly identify the underlying model will then depend on the power and significance levels of the tests employed.

It is apparent therefore that the distribution of $\underset{\sim}{d}^*$, under the hypothesis that $\underset{\sim}{\mu}_d = \underset{\sim}{0}$, will be of fundamental interest. This distribution will in fact be one of two types:

1. If σ^2 is known, or at least can be estimated accurately by v_k (implying $n-k$ is large), then $\underset{\sim}{d}^*$ will be multivariate normal with correlation matrix $\underset{\sim}{\Omega}$. Correspondingly, the set of d_i^{*2}, $i = 1,\ldots,k-r$, will have a joint density which is termed in the literature as the Multivariate Chi-Square Distribution. (e.g. see Krishnaiah[41]) Such a distribution is characterised by its degrees of freedom (in our case unity) and the correlation matrix of the associated multivariate normal distribution (i.e. $\underset{\sim}{\Omega}$).

2. If σ^2 is unknown and is estimated by an independent quantity S_o^2, say, with ν degrees of freedom, then the d_i^* have a Multivariate t-Distribution

(2.g. see Krishnaiah[41],Dunnett and Sobel [22]). Again the distribution of the d_i^{*2} will be the Multivariate F-Distribution, with 1 and ν degrees of freedom.

Further discussion on these distributions is more conveniently postponed till a later chapter.

Chapter 4 Formalisation of the Problem

4.1 Basic objectives

While we have already gone some way towards specifying formal objectives for stepwise regression by postulating the existence of a 'true' underlying model it is necessary at this junctûre to give the matter rather more thought. Before doing this a few comments will be made relating to some remarks made by Anscombe [6] on the stepwise regression technique in general.

Anscombe takes the view that stepwise regression should be regarded entirely as a descriptive procedure, and one which should be contemplated only if ample computing resources are available. He suggests that as much computer output as possible is obtained at each step, and that several different regression paths should be explored by over-riding any automatic pivotal mechanism present in the program being used. The final decision as to the 'best' equation is then based on such devices as the examination of residuals using data plots against excluded variables, Durbin-Watson or periodogram tests for serial correlation, and also subjective arguments, as to the reasonableness of the regressors obtained.

Whilst being indisputable on general statistical grounds Anscombe's suggestions would seem to rule out the possibility of anyone but an experienced statistician from using such a procedure. Given that stepwise regression is an increasingly widely used technique, it does seem worthwhile investigating whether currently used automatic stopping criteria can be improved upon in any way. Whenever possible of course as much auxiliary knowledge and diagnostic checking should be brought to bear on the problem as the situation warrants. With these points in mind we now turn to a rather more formal treatment of stepwise regression.

4.2. Stepwise regression as a multiple decision problem

It is true to say that much of the standard theory of inference in regression is restricted in application to two-decision (or hypothesis test) problems. Amongst the various tests available of this type possibly the most commonly used are those for the multiple correlation coefficient and the individual (partial) regression coefficients. It is usual for a sequence of such tests to be performed on the same data set without paying much heed to either the induced overall level of significance or to the power in picking up various alternatives. When we also reflect that in stepwise regression the particular sequence of tests performed is dictated by the actual sample data we see that the problem calls for closer investigation.

A general way of formulating the stepwise regression problem is in terms of :

1. A parameter space $\mathcal{B}$ for the unknown vector $\underset{\sim}{\beta}$ (i.e. the 'true state of nature'). For our purposes this space will throughout be taken to be k-dimensional Euclidean space R^k.

2. A sample space $\mathcal{Y}$ on which is defined, for each element $\underset{\sim}{\beta} \in \mathcal{B}$, a probability distribution $F_y(\beta)$. The sample space will in fact consist of all possible values of the n-dimensional regressand vector $\underset{\sim}{Y}$ and is therefore taken to be R^n. The probability structure on $\mathcal{Y}$ has already been implied by the model specification in Chapter 1.

3. A decision class D consisting of decision rules d(Y) mapping Y into an action space A. D will throughout be restricted to be the class of what will be termed 'stepwise procedures'. By this is meant procedures using the basic pivotal algorithm and involving a decision at each step of either introducing or deleting a regressor on the grounds of its respective maximum or minimum contribution to explained sum of squares. The action space A will of course depend on the particular motivation for using the procedure.

4. A loss function $L(a,\underset{\sim}{\beta})$ defined on the product space $A \times \mathcal{B}$. This represents the cost incurred in taking action a when $\underset{\sim}{\beta}$ is the true coefficient vector. For each decision procedure $d \in D$ a probability distribution is induced over A for each $\underset{\sim}{\beta} \in \mathcal{B}$ by identifying actions

with decisions d(y) and then invoking $F_y(\underset{\sim}{\beta})$. This leads to a re-characterization of L( ) which now becomes a random variable depending on $\underset{\sim}{\beta}$. The expectation of this stochastic loss,

$$R_d(\underset{\sim}{\beta}) = E[L(d(\underset{\sim}{Y}), \underset{\sim}{\beta})],$$

is called the *risk function*, and is of prime interest in decision theory.

Leaving aside for the moment the non-trivial matter of the evaluation of $R_d(\underset{\sim}{\beta})$ for stepwise-type procedures there remains the problem of how to use such a measure in the pursuit of an optimal procedure. Two concepts of a very general nature which occur in the theory of decision making are those of admissibility and completeness, which are now defined.

Definition 1 A decision rule $d \epsilon D$ is said to be *Admissible* if (in the present context) there is no other rule $d' \in D$ such that

$$R_d(\underset{\sim}{\beta}) > R_{d'}(\underset{\sim}{\beta}),$$

with strict inequality holding for some $\underset{\sim}{\beta} \in \mathcal{B}$.

Definition 2 A class C of decision rules ($C \subset D$) is *Complete* if for any decision rule $d' \in D - C$[†] there exists $d \epsilon C$ such that

$$R_d(\underset{\sim}{\beta}) \leqslant R_{d'}(\underset{\sim}{\beta})$$

with strict inequality holding for some $\underset{\sim}{\beta} \in \mathcal{B}$

[†] D − C denotes the symmetric difference between the sets D and C.

Definition 3 A class C of decision rules ($C \subset D$) is Minimal Complete if no proper subset of C is complete.

While it is possible to use these notions in special restricted classes of problem to arrive at optimal decision rules (see for example Ferguson [25]) the form of the decision class D of stepwise procedures would appear to rule out such an approach here.† Instead, we restrict our attention now to the problem of making objective comparisons between procedures which are suggested on intuitive (but arbitary) grounds, with a view to obtaining a meaningful ranking criterion.

Two basic principles which occur in decision theory for the determination of ranking orders for decision rules are those of Minimax and Bayes. On the minimax basis a decision rule d is preferred to another rule d' if

$$\sup_{\underline{\beta} \in \mathfrak{B}} R_d(\underline{\beta}) < \sup_{\underline{\beta} \in \mathfrak{B}} R_{d'}(\underline{\beta})$$

Using this criterion it is then sometimes possible (depending on the nature of D) to obtain a Minimax Rule d_o, say, for the complete class D. Such a procedure d_o would be given by

$$\sup_{\underline{\beta} \in \mathfrak{B}} R_{d_o}(\underline{\beta}) = \inf_{d \in D} \sup_{\underline{\beta} \in \mathfrak{B}} R_d(\underline{\beta})$$

The use of the principle of Bayes in evaluating decision rules requires recourse to a completely different

† A situation where this approach is possible is that of orthogonal regression using a multiple decision class discussed by Lehmann [50]. This is discussed later.

and controversial philosophical approach to the classical concept of probability. The essential point insofar as it impinges on the present discussion is that one can contemplate a probability distribution defined over β. This can be interpreted as either reflecting nature's own mechanism for choosing its particular state or, perhaps more realistically, this 'prior' distribution can reflect the statistician's own beliefs (objective or subjective) as to the true state. With this generalisation of the problem the risk function itself becomes a random variable for each d $\in$ D.

Using a Bayes formulation, and with a utility interpretation of the loss function, procedures can be ranked entirely on the basis of expected risk (<u>Bayes Risk</u>) i.e. decision rule d is preferred to d' if

$$\underset{\beta}{E}\,[R_d(\underline{\beta})] < \underset{\beta}{E}\,[R_{d'}(\underset{\sim}{\beta})]$$

In analogy with minimax procedures it may be possible to find a decision rule $d_o \in D$ possessing <u>Minimum Bayes Risk</u>, i.e. if there exists $d_o \in D$ such that

$$\underset{\beta}{E}[R_{d_o}(\underset{\sim}{\beta})] = \inf_{d\in D} \underset{\beta}{E}\,[R_d(\beta)]$$

We have outlined above a formal framework for stepwise regression, and mentioned some basic techniques and criteria which arise at a general level in decision theory. Apart from a particular case to be discussed below no attempt will be made to develop an optimal stepwise procedure on such a formal basis.

As far as the use of the minimax and Bayes principles is concerned the former of these at least does not seem to have any outstanding claim to be the appropriate one for ranking purposes. Although no use is made of Bayesian philosophy either in the subsequent analysis it is appreciated that its use would be helpful in overcoming some of the formal difficulties of interpretation which arise. However, from the viewpoint of finding reasonable practical procedures the added difficulties of specifying prior information for specific applications will be avoided. In any case it will become evident that for purposes of ranking procedures overwhelming computational problems are faced for the sequential types of procedure envisaged. In passing we do however mention a Bayesian treatment of the general problem of choosing variables for regression given by Lindley [51] using a prediction mean square-error loss function. The decision class considered in that paper is however of a much broader nature than the class of stepwise procedures at present being studied.

We now conclude this section by describing briefly a class of decision problem for which Lehmann [50] gave an optimal solution. Each decision procedure in this class derives from the simultaneous application of a set of two-decision (or hypothesis test) procedures. Within this class the proviso is of course made that no inconsistencies can occur regarding the actions to be taken, i.e. the component tests must lead to <u>compatibility.</u> Using an extended definition of Neyman-Pearson unbiasedness for significance tests, and assuming a certain additive property for the loss function (specifically, the losses for each component test are additive), Lehmann derives a procedure which uniformly minimizes the risk amongst all such unbiased procedures. A particularly relevant feature of this approach is that it allows one to determine the optimal set of (in general) different significance levels to apply to each component test.

The possibility of formulating stepwise regression in this way is looked into later when the two distinct situations of orthogonal and non-orthogonal regressors are examined. A major impediment towards this end arises from the data-induced nature of the sequence of hypothesis tests, especially in the non-orthogonal case. One can however handle situations in which an *a priori* sequence of tests is known as, for example, when a natural ordering of the regressor variables exists. Such a selection procedure was looked at by Anderson [3] in determining the degree of a polynomial regression relationship.

4.3. The multiple comparisons approach

An important field of study within the area of simultaneous statistical inference is that concerning multiple comparison problems. This can be said to have had its origins in attempts to detect the presence of outliers in sample data. Later work, much of it attributable to

Duncan, Scheffe and Tukey in particular, was slanted towards analysis of variance type applications. As pointed out by Miller [58] an underlying constraint in the derivation of such procedures is that of protection for a certain null hypothesis. In terms of overall procedure optimality this in turn imposes a specific loss structure which might not always be the right one. In order to compare the merits of completing procedures under various alternative hypothesis Duncan proposed the use of so-called <u>p - mean significance levels</u>. To illustrate the meaning of these consider a problem involving the unknown parameters $\theta_1, \theta_2, \ldots, \theta_m$. Denoting these by a vector $\underset{\sim}{\theta}$, suppose there are available a number of similar tests of the hypothesis

$$H_o \equiv \underset{\sim}{\theta} = \underset{\sim}{0}.$$

Now let $\underset{\sim}{\theta}_p$ be the vector representing a given (arbitrary) set of p of the parameters, and let $\underset{\sim}{\theta}_p^*$ be the vector of the remaining m-p elements of $\underset{\sim}{\theta}$. The p-mean significance level for $\underset{\sim}{\theta}_p$ is then defined as

$$\sup_{\underset{\sim}{\theta}_p^*} \text{Prob } \{D(\underset{\sim}{\theta}_p \neq \underset{\sim}{0}) / \underset{\sim}{\theta}_p = \underset{\sim}{0}, \underset{\sim}{\theta}_p^*)$$

where $D(\underset{\sim}{\theta}_p \neq \underset{\sim}{0})$ denotes the decision that $\underset{\sim}{\theta}_p \neq \underset{\sim}{0}$.

While such criteria could, at least in theory, be applied to a sequential decision procedure such as stepwise regression (indeed Duncan did so in substantiating his sequential multiple-range procedure) we will not formally do so here for two main reasons. Firstly, although it is apparent that the evaluation of such quantities amounts to only a small part of the larger problem of risk evaluation the computation involved is still exceedingly formidable, This is particularly so in the case of general forward/backward procedures applied to non-orthogonal regressions. Secondly, such criteria are designed for protection against only a testricted class of incorrect decisions which are not necessarily the appropriate ones to consider in stepwise regression.

Although stepwise regression procedures will not be evaluated formally as multiple comparison procedures the various procedures put forward later, albeit mostly on the basis of intuition combined with the results of Chapter 3, have close similarity with such techniques. It seems difficult to avoid this approach to stepwise regression due to its very nature of construction.

Before proceeding to a discussion of the two types of loss structure mentioned earlier we briefly take a look at the similarity between stepwise regression and the seemingly unrelated problem of detecting the presence of outliers in an observed sample.

4.4 Relationship of stepwise regression to the problem of testing for outliers.

For simplicity we will suppose that the value of the error variance σ^2 is known in the stepwise regression context, and also assume the regressors are orthogonal. Referring to the results of Chapter 3 we note that we initially have available k quantities d_i^{*2} which in general are independently distributed with different non-central chi-square distributions. The stepwise regression problem then is equivalent to the detection of which, if any, of these variables have central chi-square distributions (or zero non-centrality parameters). In other words, we wish to detect the non-central chi-square outliers in a random sample from a central chi-square distribution. It therefore, seems worth investigating if the theory or outlier detection is of help to us.

Most of the theory of tests for outliers relates to normally distributed variables for which means and variances are unknown. In stepwise regression we do know these two quantities (in the sense that the mean should be zero and we can at least obtain an independent estimate of σ^2), and the situation falls into the first of four categories discussed by David [18 , Chapter 8]. It is of interest to note that the two main types of stepwise procedure considered later use test statistics which are virtually equivalent to the two which David suggests as being appropriate for outlier detection in either direction in normal samples.

Littlework has been carried out on the power properties of outlier detection procedures. Even then what has been done relates almost exclusively to the case in which only one outlier at most is suspected of being present. David does however suggest that tests based exclusively on extremes should be expected to be more efficient. This assertion will be examined later in the context of some proposed stepwise regression procedures. The problem of detecting an unknown number of outliers has received very little attention in the published literature. The suggestion by David (p.191) that sequential tests should be performed on samples of reducing size till insignificance is first obtained is in agreement with the philosophy underlying the stepwise procedures developed later (and indeed concurs with most multi-stage multiple comparison procedures).

Although a very special situation was examined above for illustrative purposes the ideas carry over to more general regression set-ups. In particular the general non-orthogonal case leads to consideration of the detection of outliers in a non-random sample with a known dependence structure. As far as is known no work has been done on this latter problem.

We now turn to a discussion of the two basic types of stepwise regression objective referred to earlier.

4.5. The identification problem

By this will be taken to mean the problem of deciding which regressor variables enter into the true model with a non-zero regression coefficient. The action space A is then of finite dimension 2^k, the number of different model formulations possible. Each action in fact corresponds to a decision that $\underline{\beta}$ lies in a set B_i, where the sets $B_i (i = 1,2,\ldots,2^k)$ form a partition of $\boldsymbol{\beta}$.

It is important here to differentiate between the two different objectives of identification and control insofar as a single stepwise regression analysis is contemplated for both situations. If by 'control' is meant the evaluation of the marginal effects due to controllable regressor variables the nature of the action space, and also the associated loss function, will be considerably changed. In particular it should be noted that although stepwise regression has a least squares basis none of the standard results, such as the Gauss-Markov properties, for example, can be expected to hold. This point is returned to below when the prediction problem is discussed. One possible interpretation of the identification problem as considered here is as a data-reducing technique in a regression situation. One might for instance carry out such an analysis as a preliminary to the estimation of a regression model using a further set of data. In this sense our aims can perhaps best be described as those of model specification.

Reverting to the formal framework established earlier we can express the risk function, for $\underline{\beta} \in B_i$ and $d \in D$, in the form

$$R_d(\underline{\beta}) = \sum_{j=1}^{2^k} C_{ij} \text{ Prob}_d\{a_j/\underline{\beta}\} \qquad (1)$$

where C_{ij} is the cost involved in taking action a_j (corresponding to $\underline{\beta} \in B_j$) when $\beta \in B_i$. The actual form of the cost function can be considerably simplified by supposing that only the <u>underfitting</u> and <u>overfitting</u> characteristics of procedures are involved. However, one still has to resolve difficulties such as whether or not the two shortcomings are equally disadvantageous, and also the manner in which the loss increases with the degree of under and overfit. In the subsequent comparison of various procedures using a similation approach no attempt is made to impose such a strict cost structure. Instead, evaluations are made on more general grounds of 'closeness' to the underlying true model.

4.6 <u>The prediction problem</u>

The identification problem as described above leads to regarding stepwise regression as only an exploratory technique, the main emphasis being placed on trying to identify the true underlying model structure.

In this section we will be looking at the different objective of minimization of the mean square-error of prediction(MSEP) with regard to a further set of regressor values. More specifically, if $\underset{\sim}{x}$ is a (K+1)-dimensional vector of regressor values (including the mean element), and if $\underset{\sim}{b}$ is the estimator of $\underset{\sim}{\beta}$ obtained from the stepwise procedure used, we wish to minimize

$$\text{MSEP} = E[(\hat{y} - y)^2] \tag{1}$$

where $\hat{y} = \underset{\sim}{x}'\underset{\sim}{b}$, $y = \underset{\sim}{x}'\underset{\sim}{\beta} + \varepsilon$.

Rewriting (1) as

$$\text{MSEP} = E[\{\underset{\sim}{x}'(\underset{\sim}{b}-\underset{\sim}{\beta})-\varepsilon\}^2]$$

and noting that ε is (by assumption) independent of $\underset{\sim}{b}$ with zero mean, we obtain

$$\text{MSEP} = \sigma^2 + E[\{\underset{\sim}{x}'(\underset{\sim}{b}-\underset{\sim}{\beta})\}^2] \tag{2}$$

$$= \sigma^2 + \text{Var}(\underset{\sim}{x}'\underset{\sim}{b}) + B^2 \tag{3}$$

where $B = E[\underset{\sim}{x}'\underset{\sim}{b}] - \underset{\sim}{x}'\underset{\sim}{\beta}$ is the bias in the estimation of $\underset{\sim}{x}'\underset{\sim}{\beta}$.

Since σ^2 is a constant for all procedures the problem is equivalent to the minimization of the mean square-error of $\underset{\sim}{x}'\underset{\sim}{b}$ as an estimator of $\underset{\sim}{x}'\underset{\sim}{\beta}$.

Using (2) we see that

$$\begin{aligned} E[\{\underset{\sim}{x}'(\underset{\sim}{b}-\underset{\sim}{\beta})\}^2] &= E[\underset{\sim}{x}'(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})'\underset{\sim}{x}] \\ &= \underset{\sim}{x}'E[(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})']\underset{\sim}{x} \\ &= \underset{\sim}{x}'\underset{\sim}{P}\underset{\sim}{x}, \text{ say,} \end{aligned} \tag{4}$$

where $\underset{\sim}{P} = E[(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})']$

In terms of the risk function as defined in section 2 above we can write

$$R_d(\underset{\sim}{\beta},\underset{\sim}{x}) = \underset{\sim}{x}'\underset{\sim}{P}\underset{\sim}{x} \tag{5}$$

where risk now depends on the value of $\underset{\sim}{x}$ as well as $\underset{\sim}{\beta}$. Consider now the possibility of using (5) as a ranking criterion for stepwise procedures in a preduction context. A procedure d_1 is uniformly as good as d_2 if, for all $\underset{\sim}{\beta} \in \mathcal{B}$ and all $\underset{\sim}{x}$,

$$R_{d_1}(\underset{\sim}{\beta},\underset{\sim}{x}) \leqslant R_{d_2}(\underset{\sim}{\beta},\underset{\sim}{x})$$

From (4) this implies we require

$$\underset{\sim}{x}'(\underset{\sim}{P}_2 - \underset{\sim}{P}_1)\underset{\sim}{x} \geqslant 0$$

for all $\underset{\sim}{\beta}$ and all $\underset{\sim}{x}$ (where $\underset{\sim}{P}_1$ and $\underset{\sim}{P}_2$ have obvious meanings).

Hence the required condition is that

$\underset{\sim}{P}_2 - \underset{\sim}{P}_1$ is positive semi-definite for all $\underset{\sim}{\beta}$.

Since, as will later become evident, the theoretical evaluation of $\underset{\sim}{P}_2 - \underset{\sim}{P}_1$ is intractable (at least for the procedures to be considered) one is forced to resort to less objective means. Such an approach, that of using simulation methods, is left over to a later chapter. We can however throw a little light on the situation by considering a very simple example. This will also serve in examining a conjecture made by Allen [2] in a more general paper on the determination of optimal prediction models. On the basis of some numerical work (not presented in his paper) Allen suggests that different subsets of regressors are optimal at the least squares estimation stage depending on the variables which appear in the $\underset{\sim}{x}$ vector used at the prediction stage. In particular he offers, by way of explanation, the argument that if one is interested in estimating β_1 for example (i.e. $\underset{\sim}{x} = [1\ 0\ 0 \ldots 0]$) then one feels, a priori, that X_1 should appear in the subset of regressors at the estimation stage. The plausibility of this intuitive argument will in fact be shown to be questionable.

Consider the model

$$Y = X_1 \beta_1 + X_2 \beta_2 + \varepsilon$$

where the usual assumptions are made. Let b_{y1} and b_{y2} be the calculated regression coefficients obtained by regressing Y on X_1 and X_2 respectively in single regressor models, and let $b_{y1.2}$ and $b_{y2.1}$ be the estimators obtained from the full model. Consider then the estimation of a linear function $\lambda_1 \beta_1 + \lambda_2 \beta_2$ of the true regression coefficients using each of the following estimators.

1. $\lambda_1 \ b_{y1}$
2. $\lambda_2 \ b_{y2}$
3. $\lambda_1 \ b_{y1.2} + \lambda_2 b_{y2.1}$

Wenow derive the mean square-error values for each of these in turn. In each case we need only determine the matrix $\underset{\sim}{P}$ as in (4). We do so in terms of the elements $p_{ij}^{(k)}$ for k = 1,2,3, where $p_{ij}^{(k)}$ is the (i,j)-element of $\underset{\sim}{P_k}$.

1. $p_{11}^{(1)} = E[b_{y1}-\beta_1)^2]$. Now $b_{y1} = \Sigma YX_1/\Sigma X_1^2$

$$= \Sigma(X_1\beta_1 + X_2\beta_2 + \varepsilon)X_1/\Sigma X_1^2$$

$$= \beta_1 + \beta_2\Sigma X_1 X_2/\Sigma X_1^2 + \Sigma X_1\varepsilon/\Sigma X_1^2 .$$

Hence we obtain, on taking expectations. and letting b_{12} be the calculated (non-stochastic) regression coefficient for X_2 on X_1,

$$p_{11}^{(1)} = \beta_2^2\, b_{12}^2 + \sigma^2/\Sigma X_1^2$$

Similarly, $p_{12}^{(1)} = p_{21}^{(1)} = E[-\beta_2(b_{y1}-\beta_1)] = -\beta_2^2\, b_{12}$,

and

$$p_{22}^{(1)} = E[\beta_2^2] = \beta_2^2$$

2. By symmetry we have immediately

$$p_{11}^{(2)} = \beta_1^2 b_{21}^2 + \sigma^2/\Sigma X_2^2$$

$$p_{12}^{(2)} = p_{21}^{(2)} = -\beta_1^2\, b_{21}$$

$$p_{22}^{(2)} = \beta_1^2$$

3. In this case $b_{y1.2}$ and $b_{y2.1}$ are unbiased estimators and $\underset{\sim}{P_3}$ is just the usual covariance matrix for least squares estimators i.e.

$p_{11}^{(3)} = \sigma^2/\{\Sigma X_1^2(1-r_{12}^2)\}$, where r_{12} is the calculated correlation coefficient between X_1 and X_2 (using raw moments).

$p_{12}^{(3)} = p_{21}^{(3)} = -\sigma^2 b_{12}/\{\Sigma X_1^2(1-r_{12}^2)\}$

$p_{22}^{(3)} = \sigma^2/\{\Sigma X_2^2(1-r_{12}^2)\}$.

We first take a look at Allen's assertion referred to previously. In particular, suppose we put $\lambda_2 = 0$. Using the above results it follows that Method 2 is preferable to Method 1 if

$$\beta_1^2 < \beta_2^2 b_{12}^2 + \sigma^2/\Sigma X_1^2, \qquad (6)$$

i.e. in order to estimate $\lambda_1\beta_1$ (for any λ_1) it may be better to ignore X_1 completely at the estimation stage. It is seen that, except perhaps in the case where X_1 and X_2 are orthogonal, this result does not in general add weight to Allen's reasoning.

Without going into explicit detail it can easily be demonstrated using the above results that, whilst a condition like (6) holds for particular values of λ_1 and λ_2, the condition is not that for positive - definiteness of $\underset{\sim}{P}_2 - \underset{\sim}{P}_1$ over all values of β_1 and β_2. In fact there are values of (λ_1, λ_2) for which, even for fixed values of β_1 and β_2, both Method 1 and Method 3 are optimal with condition (6) still holding. Since, for fixed values of $\underset{\sim}{\beta}$, stepwise regression procedures will have different associated probabilities of terminating at the various possible model structures it follows that procedure optimality will depend on $\underset{\sim}{x}$.

A natural way to overcome this dependence is to select a representative set of such values $\underset{\sim}{x}_i, i=1,\ldots,q$, say, and then consider a straightforward average

$$R_d(\underset{\sim}{\beta}) = \frac{1}{q} \sum_{i=1}^{q} R_d\,(\underset{\sim}{\beta},\underset{\sim}{x}_i) \qquad (7)$$

This does at least permit a ranking of procedures for fixed values of $\underset{\sim}{\beta}$.[†] Alternatively one might perhaps be prepared to specify a probability structure for $\underset{\sim}{x}$, i.e. treat it as a stochastic quantity $\underset{\sim}{X}^*$ say, and then use

$$R_d(\underset{\sim}{\beta}) = E[R_d(\underset{\sim}{\beta},\underset{\sim}{X}^*)] \qquad (8)$$

where the expectation is now with respect to the joint distribution of $\underset{\sim}{X}^*,\underset{\sim}{X}$ and $\underset{\sim}{\varepsilon}$ (and where we now sensibly also assume $\underset{\sim}{X}$ is stochastic). A particular case of interest is that of <u>forecasting</u> using time series models, in which case (8) leads to the standard criterion of forecast mean square-errpr. In this most general situation we must acknowledge the joint dependence of $\underset{\sim}{X}^*,\underset{\sim}{X}$ and $\underset{\sim}{\varepsilon}$ and rewrite (8) as

$$R_d(\underset{\sim}{\beta}) = \underset{\underset{\sim}{X}^*,\underset{\sim}{X},\underset{\sim}{\varepsilon}}{E} [\underset{\sim}{X}^{*\prime}(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})'\underset{\sim}{X}^*] \qquad (9)$$

Of course in each of (7),(8) and (9) one is still faced with the problem of obtaining a ranking criterion not dependent on $\underset{\sim}{\beta}$. Again, as in the identification case looked at earlier, $\underset{\sim}{\beta}$ will be retained as a nuisance parameter in the simulation investigations conducted

[†] We may note that by selecting $\underset{\sim}{x}_i, i=1\ldots,k$, to be the k columns of the k-dimensional identity matrix then we convert the problem to one of mean square error <u>estimation</u> of $\underset{\sim}{\beta}$.

later.

Before concluding this section a brief look will be taken at the problem of evaluating the matrix $\underset{\sim}{P}$ which occurs in (4). We first note that $\underset{\sim}{P}$ can be written in the form

$$\underset{\sim}{P} = \text{Var}\ (\underset{\sim}{b}) + \underset{\sim}{B}$$

where Var($\underset{\sim}{b}$) is the coveriance matrix of the estimator $\underset{\sim}{b}$ and $\underset{\sim}{B}$ is what can perhaps be termed a generalized bias matrix, having (i,j)-element equal to

$$(E[b_i]-\beta_i)\ (E[b_j]-\beta_j) \qquad \text{for } i,j = 1,\ldots,k.$$

A natural approach is then to consider separately the bias and variance properties of stepwise estimation procedures.

The first study of the effects of performing preliminary tests of significance in a situation of what has come to be termedin the literature an incompletely specified model seems to be that of Bancroft [7]. In his paper Bancroft considered the standard practice of performing t-tests sequentially on an estimated regression model, and, in particular, pointed out that the usual unbiasedness property of least squares no longer holds. Kitogawa [40] developed the argument by obtaining, for a simple two regressor model, the cumulative distribution function and associated moments for the coefficient of the variable remaining after such a variable deletion.

In two papers Larson and Bancroft [46,47] and also later Kennedy and Bancroft [39], extended the argument to estimates obtained specifically at the termination of sequential procedures using mean square-error of prediction as the criterion of interest. In all cases however the study was restricted to situations in which there exists a natural order of importance for the regressors involved. Such prior knowledge might arise, it is suggested, from theoretical considerations or from previous experience in similar applications. Although one would concede there might well be situations in which some variables seem more realistic as regressors than others it would be unusual if such a complete ordering were to be available. There are also occasions in which what might seem to be a natural ordering is not in fact so. An example of this occurs in the determination of the lag structure of autoregressive models. It has been implied by some authors (e.g. Kendall and Stuart [38,p.476]) that a natural ordering is obtained by assuming that variables decrease in importance in inverse relation to the length of their associated time lag. This does however impose severe restrictions on the underlying 'causal' structure of the system, and could give rise to misleading conclusions in a dynamic control set-up.

Although the three papers referred to above restrict the analytical treatment to the situation of orthogonal regressors the point is made that the results still apply to non-orthogonal set-ups. The essential argument is that the order of introducing variables into the equation is still fixed and determines the extra sum of squares decomposition of the total explained variation. Since this is identical to performing an orthogonalizing transformation on the original set of regressors one only has to show that one can re-transform the finally selected equation back to this original set without affecting the values obtained for bias and variance. The truth of this is demonstrated quite easily (e.g. see Allen [2,p.1282]).

The restricted nature of the class of sequential decision procedures considered in the three studies referred to above considerably simplifies the analysis involved In particular it is reasonable in such a situation to apply a fixed critical value to the F-statistic obtained at each stage, i.e. since there is no data-induced order for variable entry one does not have to deal with the more complicated aspect of order statistic distributions. Further, the non-orthogonal case is uniquely orthogonalized once a natural priority ordering is given for the regressors. Whilst orthogonalization is a desirable feature and is still possible in the more general class of stepwise procedures being considered the non-uniqueness of this gives rise to further difficulties. Suggestions have been made (e.g. see Kendall [36], Daling and Tamura [17] and Wickens and Ord [63]) that an appropriate transformation is provided by the principal component transformation of the regressor set. Although this has the advantage of being widely available as a computer procedure it seems difficult to regard it, in a general sense, as anything but an arbitrary transformation. This point is not however pursued any further for now since it is more conveniently discussed in the light of some ideas developed later.

To conclude this section we now obtain formal expressions for $R_d(\underset{\sim}{\beta})$ on the lines of that given in (5.1). Consider again the general case given by (9),

$$R_d(\underset{\sim}{\beta}) = \underset{\underset{\sim}{X}^*,\underset{\sim}{X},\underset{\sim}{\varepsilon}}{E} [\underset{\sim}{X}^{*\prime}(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})'\underset{\sim}{X}^*]$$

which can be rewritten in the form

$$\underset{\underset{\sim}{X}^*,\underset{\sim}{X}}{E} \{X^* \underset{\underset{\sim}{\varepsilon}/\underset{\sim}{X}^*,\underset{\sim}{X}}{E} [(\underset{\sim}{b}-\underset{\sim}{\beta})(\underset{\sim}{b}-\underset{\sim}{\beta})'/\underset{\sim}{X},\underset{\sim}{X}^*]\underset{\sim}{X}^*\} \qquad (10)$$

In the case of non-stochastic $\underset{\sim}{X}^*$, or the stochastic case in which $\underset{\sim}{\varepsilon}$ is independent of $\underset{\sim}{X}$ and $\underset{\sim}{X}^*$, the inner expectation is with respect to the unconditional normal distribution of $\underset{\sim}{\varepsilon}$. In the general stochastic case however the inner expectation, whilst being over a normal distribution (at least if $\underset{\sim}{X}^*$ and $\underset{\sim}{X}$ are jointly normal), is no longer spherical normal. In such a situation one has to resort to a simultaneous evaluation with respect to the overall joint distribution as in (9).

Restricting our interest to the former of the two situations described above, and using the notation as in (4), we can consider the evaluation of the inner expectation. This essentially reduces to the determination of the first two moments of $\underset{\sim}{x}'\underset{\sim}{b}$, i.e. $E[\underset{\sim}{x}'\underset{\sim}{b}]$ and $E[(\underset{\sim}{x}'\underset{\sim}{b})^2]$. Conditional on $\underset{\sim}{x}$ and $\underset{\sim}{X}$ we then have, using the action space A of section 5 above,

$$E[\underset{\sim}{x}'\underset{\sim}{b}] = \underset{\sim}{x}'E[\underset{\sim}{b}] = \underset{\sim}{x}' \sum_{j=1}^{2^k} E[\underset{\sim}{b}/a_j]P(a_j)$$

$$E[(\underset{\sim}{x}'\underset{\sim}{b})^2] = E[\underset{\sim}{x}'\underset{\sim}{bb}'\underset{\sim}{x}] = \underset{\sim}{x}'\left\{\sum_{j=1}^{2^k} E[\underset{\sim}{bb}'/a_j]P(a_j)\right\}\underset{\sim}{x}$$

It will be apparent that the formal evaluation of both theP(a_j) and the conditional expectations is a formidable task. Even in the considerably more simple situation which is contemplated by Bancroft,Larson and Kennedy (see earlier) one is still faced with intractable expressions. Again, as in the identification case above, the approach will be taken of constructing procedures on what appear to be reasonable grounds within the basic stepwise structure.

Finally it should be remarked that throughout this section it has been assumed that the underlying cost structure is adequately represented by the mean square-error of prediction. It must however be recognised that in practice other criteria might be relevant. In particular one might well be prepared to forego some predictive efficiency if data collection or processing costs can be reduced enough by way of compensation. This point is especially significant in the comparison of procedures which involve the calculation of the complete (k-variable) equation with those that do not. Such cost considerations could quite easily be incorporated into the previous formulations but would not of course make them any more susceptable to solution. One must however pay heed to them in any conclusions which may be drawn on less formal grounds such as the empirical studies performed later.

Chapter 5 Identification with Orthogonal Regressors

5.1 Special features of orthogonality

In any attempt to determine acceptable stepwise regression procedures a sensible starting point would seem to be that of orthogonal regressors. This will of course restrict the application to data obtained from controlled experiments, but hopefully some light will also be thrown on the more complex situation of non-orthogonality which is discussed later. In any event the orthogonal situation is of some interest in its own right.

A major simplication which arises is that partial contributions to explained sums of squares are also overall contributions. This ensures that the equation sequences produced by any of the basic stepwise techniques will be identical provided they are carried out for all k stages. In particular no real motivation exists for performing general forward/backward routines. Indeed the problem can in fact be treated entirely as one of non-sequential simultaneous inference, such an approach being considered later. We can throughout confine our interest to the independently distributed quantities

$$S_1^2, S_2^2, \ldots\ldots, S_k^2$$

which, in the notation of Chapter 3, are the quantities $\sigma^2 d_j^{*^2}$, $j = 1, \ldots., k$. It follows from the results of that chapter that each S^2_j is distributed as non-central chi-square with one degree of freedom and non-centrality parameter $\beta^2_j \sum_{i=1}^{n} X^2_{ij}/2\sigma^2$ (i.e. $\chi^2\ (1,\beta^2_j \sum_{i=1}^{n} X^2_{ij}/2\sigma^2)$). In addition to these k quantities we also have the residual sum of squares

$$S^2_o = (n-k-1)v_k$$

calculated from the complete equation. As demonstrated in Chapter 3 S^2_o is also distributed independently of each S^2_j, $j = 1,\ldots.,k$.

5.2. Expected order of variable entry

Before proceeding to the consideration of a suitable conditional hypothesis structure for stepwise procedures it is important to look at the expected order of entry (deletion) of variables to (from) the fitted equation. We shall throughout continue to assume that the order of variable entry is determined by the magnitudes of the S^2_j, $j = 1,\ldots,k$. Since each S^2_j can be written as $b^2_j \sum_{i=1}^{n} X^2_{ij}$ (where b_j is just the simple regression coefficient estimate for Y on X_j), and since also b_j is, with suitable general restrictions on the X values selected, a consistent estimator of β_j, then the order of entry is essentially determined by the values of $\lambda_j = \beta^2_j \sum_{i=1}^{n} X^2_{ij}$, $j = 1,\ldots.,k$.

More specifically we have

$$E[S^2_j] = \sum_{i=1}^{n} X_{ij}E[b^2_j]$$

Using the fact that b_j is distributed as $N\left(\beta_j, \sigma^2 \Big/ \sum_{i=1}^{n} X^2_{ij}\right)$ it follows that

$$E[b^2_j] = \beta^2_j + \sigma^2 \Big/ \sum_{i=1} X^2_{ij},$$

and hence

$$E[S^2_j] = \beta^2_j \sum_{i=1}^{n} X^2_{ij} + \sigma^2$$

$$= \lambda_j + \sigma^2 \qquad (1)$$

Further, considering the variance of S^2_j, we have

$$\mathrm{Var}[S^2_j] = \left(\sum_{i=1}^{n} X^2_{ij}\right)^2 \mathrm{Var}[b^2_j]$$

Now $\mathrm{Var}[b^2_j] = E[b^4_j] - E^2[b^2_j]$. Using the moment generating function for b_j we find that

$$E[b^4_j] = 3\sigma^4 \Big/ \left(\sum_{i=1}^{n} X_{ij}\right)^2 + 6\beta^2\sigma^2 \Big/ \sum_{i=1}^{n} X^2_{ij} + \beta^4_j$$

Combining this with the expression above for $E[b^2_j]$ we obtain

$$\mathrm{Var}[S^2_j] = 2\sigma^4 + 4\beta^2_j\sigma^2 \sum_{i=1}^{n} X^2_{ij} \qquad (2)$$

While from (1) it follows that the values of λ_j determine the order of variable entry from an expectation standpoint, the form of (2) is not very informative as to the probability of such an ordering being obtained. To this end let us consider the regressor variables X_j and $X_{j'}$ such that

$$\lambda_j < \lambda_{j'} \qquad (\lambda_{j'} \neq 0) \qquad (3)$$

Then

$\mathrm{Prob}[S_j^2 - S_{j'}^2 < 0] = \mathrm{Prob}[b_j^2 \sum_{i=1}^{n} X_{ij}^2 - b_{j'}^2 \sum_{i=1}^{n} X_{ij'}^2 < 0]$

$= \mathrm{Prob}[pb_j^2 - b_{j'}^2 < 0]$,

where, for all n, we take $p = \sum_{i=1}^{n} X_{ij}^2 \Big/ \sum_{i=1}^{n} X_{ij'}^2$.

Putting $U = pb_j^2 - b_{j'}^2$, we see that

$$E[U] = p\left(\beta_j^2 + \sigma^2 \Big/ \sum_{i=1}^{n} X_{ij}^2\right) - \beta_{j'}^2 - \sigma^2 \Big/ \sum_{i=1}^{n} X_{ij'}^2$$

$$= p\beta_j^2 - \beta_{j'}^2$$

$$= (\lambda_j - \lambda_{j'}) \Big/ \sum_{i=1}^{n} X_{ij'}^2 < 0 \qquad (4)$$

Also, $\mathrm{Var}[U] = p^2 \mathrm{Var}[b_j^2] + \mathrm{Var}[b_{j'}^2]$.

On using a slight modification of (2), and with some simplification, we find that

$$\mathrm{Var}[U] = 2\sigma^4\left(1 + 1 \Big/ \sum_{i=1}^{n} X_{ij'}^2\right) \Big/ \sum_{i=1}^{n} X_{ij'}^2 + 4\beta_j^2\sigma^2 p \Big/ \sum_{i=1}^{n} X_{ij'}^2$$

$$+ 4\beta_{j'}^2\sigma^2 \Big/ \sum_{i=1}^{n} X_{ij'}^2 . \qquad (5)$$

Hence, as $n \longrightarrow \infty$ and imposing the realistic constraint that $\sum_{i=1}^{n} X_{ij}^2 \Big/ \sum_{i=1}^{n} X_{ij'}^2 = p$, we see that

$$\mathrm{Var}[U] \longrightarrow 0 \qquad (6)$$

Using (4) and (6) in conjunction with Chebychev's Inequality it follows that

$$\lim_{n \to \infty} \mathrm{Prob}[pb_j^2 - b_{j'}^2 < 0] = 1$$

for all j, j' such that (3) is true.

Hence we can now assume, with confidence which increases with n, that "true" regressors will be entered first, the actual order being governed by the magnitudes of λ, λ_j, $j = 1,....,k$. This provides the basis for treating stepwise regression as a series of conditional hypothesis tests, as will now be described in more detail.

5.3. Hypothesis structure for stepwise regression

Suppose, without loss of generality, that the r variables included in the equation at the stage in question are $X_1,.....,X_r$. Then for a forward procedure we can formulate the null hypothesis

$$H_o \equiv \beta_j \begin{cases} \neq 0 & (j = 1,....,r) \\ = 0 & (j = r+1,....,k) \end{cases}$$

with the alternative hypothesis

$$H_1 \equiv \begin{cases} \beta_j \neq 0 & (j = 1,.....,r) \\ \text{Not all } \beta_j = 0 & (j = r+1,....,k) \end{cases}$$

Acceptance of H_o then corresponds to the decision to terminate the procedure while H_1 implies that not all significant variables have yet been included. Conversely, in a backward procedure, a possible hypothesis test is

$$H_o \equiv \beta_j \begin{cases} \neq 0 & (j = 1,....,r) \\ = 0 & (j = r+1,....,k) \end{cases}$$

$$H_1 \equiv \begin{cases} \text{Some } \beta_j = 0 & (j = 1,....,r) \\ \beta_j = 0 & (j = r+1,...,k) \end{cases}$$

H_1 here implies that further variables can still be deleted.

While no such basic formal structure is specified in the existing literature on stepwise regression it certainly seems to be implicit in most instances. Having established this framework we now turn to the problem of the appropriate choice of test statistic to use. This is done by looking first of all at quantities which arise in the context of completely specified regression models. Modifications of such statistics are then suggested as being appropriate in stepwise regression contexts. The discussion will throughout relate only to the identification problem defined at 4.5. While there is the added restriction of orthogonality of regressors many of the points made are equally valid in the more general non-orthogonal case.

5.4. Possible test statistics for stepwise regression

We suppose the stage has been reached at which r regressors appear in the fitted equation, and therefore $q = k - r$ variables are excluded. Various test statistics are considered in the light of both their conditional and overall implications when incorporated into stepwise regression procedures.

(a) The conventional F statistic

By this is meant the square of the t-value calculated for a regressor in a completely specified model. In such a context this statistic gives the likelihood ratio test for the regression coefficient involved. When used within a stepwise set-up we can write

$$F = S^2_{(q)}(\nu+q-1)\Big/\left(\sum_{i=1}^{q-1} S_{(i)}+S^2_o\right) \qquad (1)$$

where $S^2_{(i)}$ is the i^{th} smallest of the set of values S^2_i $(i = 1,....,k)$ and $\nu = n-k-1$. Evidently this quantity can in no way be regarded as having standard F-distribution with 1 and $\nu+q-1$ degrees of freedom (see Pope and Webster [66], who have independently investigated the problem). The actual distribution of F was investigated by Draper, Guttman and Kanemasu [19] who attempted to find critical points of F for various values of q. Making the assumption that $S^2_{(i)}/\sigma^2$ $(i = 1,....,q)$ is an ordered random sample from a $\chi^2_{(1)}$ distribution these authors derive a recurrence relation for determining these critical values. However, since this still requires the by no means simple task of evaluating 2^{q-1} (q-1)-fold integrals of incomplete beta functions, the authors were only able to present results for $q = 1(1)4$.

Although they do not investigate the effectiveness of the use of F in practice, the authors do admit to two objections against its use. These are that:

(i) in the early stages (of forward procedures) the <u>numerator</u> will be a biased estimator of σ^2.

(ii) in general regressors are not orthogonal.

Point (ii) is of course of fundamental importance, and is discussed in more detail later. In regard to objection (i), the numerator $S^2_{(q)}$ is necessarily biased as an estimator of σ^2 since it is at best the maximum of a $\sigma^2\chi^2_{(1)}$ random sample. But, even allowing for such bias, one still expects bias under the alternative kypothesis that the corresponding variable is significant. The extent of this bias will in fact determine the "power" of the overall procedure. It is possible that the authors really mean to refer to the denominator as being biased in early stages since $S^2_{(k-1)}$, $S^2_{(k-2)}$,......, etc., will in general have (different) non-central chi-square distributions. Whilst intuitively one feels this could lead to premature truncation of procedures in certain instances such conclusions would seem difficult to establish on a theoretical basis.

Turning to backward stepwise procedures suppose that the stage has been reached where $r + 1$ variables appear in the fitted equation. On removing the variable which yields the smallest contribution to explained sum of squares we are then in essentially the same situation as in the forward case. The sum of squares due to the regressor involved can now be denoted by $S^2_{(q)}$ and can be combined with the previously obtained values of $S^2_{(i)}$, $i = 1,....,(q-1)$. A different kind of problem can now be seen to arise.

For suppose the true order of equation is p(< k) and that q < k - p. Then strictly we should regard the test statistic in (1) as being distributed like

$$S^2_{(q,k-p)}(\nu+q-1)/\left(\sum_{i=1}^{q-1} S^2_{(i,k-p)}+S^2_0\right) \qquad (2)$$

where $S^2_{(i,k-p)}$ is the i^{th} smallest observation of a random sample of sixe k - p from a $\sigma^2\chi^2_{(1)}$ distribution. Since it is assumed that p is unknown one could not of course hope to use critical levels from such a distribution even if such values could be determined. Again one is left to conjecture as to the effects of using critical levels derived from (1) instead of the appropriate form of (2).

(b) The residual variance estimate

An intuitively reasonable procedure, at least when used in conjunction with a forward approach, is to plot the residual variance estimate at each stage. One feels that a flattening out of the graph thus obtained indicates correct model identification. If r variables have been fitted then we have

$$v_r = RSS_r/(n-r-1)$$

where RSS_r is the residual sum of squares from the fitted equation. We can write v_r in the form

$$(n-r-1)v_r = \sum_{i=1}^{q} S^2_{(i)} + S^2_0$$

from which we can infer that in general only when r = n can v_r be regarded as an unbiased estimator of σ^2. We can however say that asymptotically (i.e. conditional on the p true regressors being entered first) V_p will also be an unbiased estimator of σ^2. Thus a rough practical approach would be to plot v_r against r and choose the equation order r such that v_r approximately equals v_k. We can in fact be more precise than this by noting that under these stated conditions

$$(n-p-1)v_p - (n-k-1)v_k = \sum_{i=1}^{k-p} S^2(i).$$

It follows that, dividing this quantity by $(n-k-1)v_k$, we have the ratio of independent chi-square variates with (k-p) and (n-k-1) degrees of freedom respectively. Hence the statistic

$$\left(\frac{n-p-1}{k-p}\right) \frac{v_p}{v_k} - \frac{(n-k-1)}{(k-p)} \qquad (3)$$

will be distributed as central F with ((k-p) and (n-k-1) degrees of freedom. For all values of r less than p this quantity will be inflated according to the non-central F distribution which then holds. Conversely, for r > p, the statistic can be seen to be what is in a sense a negatively biased central F variate. The statistic is in fact the same one as is used in the likelihood ratio test for a subset of regression parameters in a completely specified model. In that conditional sense it can be shown to possess uniform maximum power for alternatives specified in terms of the overall non-centrality parameter of the excluded variables i.e.

$$\lambda = \sum_{j=r+1}^{k} \beta_j^2 \ {}_{i=1} \ X_{ij}^2/2\sigma^2$$

To distinguish this statistic being contemplated from the standard calculated value of F given in (1) we will denote it by F', i.e.

$$F' = \left(\frac{n-r-1}{k-r}\right) \frac{v_r}{v_k} - \frac{(n-r-1)}{(k-r)} \qquad (4)$$

The incorporation of F' into a forward procedure seems straightforward, one merely stops the procedure as soon as a test is found to be insignificant. Likewise, in a backward procedure one stops at the first significant result. Even in the orthogonal regression situation being considered it is evident that the final equations obtained will in general be different in the two cases, a backward process being expected to retain more variables. The backward case is in fact equivalent to stopping at the <u>last</u> non-significant step of the corresponding forward procedure. To resolve this ambiguity one needs to have in mind the overall implications. If one uses a fixed percentage level of α for all the significance tests then, in a forward procedure, one is building in this degree of protection against <u>overfitting</u>. If r_1 is the number of variables occurring in the finally selected equation then we have

$$\text{Prob}[r_1 > p] \leq \alpha \tag{5}$$

To demonstrate the validity of (5) consider the k independent quantities $S_1^2, S_2^2, \ldots\ldots, S_k^2$ as defined in section 1. These quantities, on division by σ^2, can be regarded as being the union of p independent observations from each of p possibly different non-central chi-square distributions, together with a random sample of size k - p from a central chi-square distribution with one degree of freedom. Denote this latter sample by the random variables

$$W_1, W_2, \ldots\ldots, W_{k-p}.$$

Since we have

$$\underset{i=1,\ldots\ldots,k-p}{\text{Max}}\ S^2_{(i)} \leq \underset{i=1,\ldots,k-p}{\text{Max}}\ W_i$$

and

$$\sum_{i=1}^{k-p} S^2_{(i)} \leq \sum_{i=1}^{k-p} W_i$$

(where $S^2_{(i)}$ denotes, as before, the i^{th} smallest value in the S^2_j sample, $j = 1,\ldots,k$)
it follows that

$$\left\{\begin{array}{l} \text{Prob}\ \underset{i=1,\ldots,k-p}{[\text{Max}}\ S^2_{(i)}/v_k \leq \text{FMAX}_\alpha] \leq \text{Prob}\ \underset{i=1,\ldots,k-p}{[\text{Max}}\ W_i/v_k \\ \qquad\qquad \leq \text{FMAX}_\alpha] = \alpha \\ \text{Prob}\ [\sum_{i=1}^{k-p} S^2_{(i)}/((k-p)v_k) \leq F'_\alpha] \leq \text{Prob}\ [\sum_{i=1}^{k-p} W_i/((k-p)v_k \\ \qquad\qquad \leq F'_\alpha] = \alpha \end{array}\right. \tag{6}$$

where v_k is the independent estimate of σ^2 based on (n-k-1) degrees of freedom and given by

$$S_0^2/(n-k-1),$$

and $FMAX_\alpha$, F'_α denote respectively the α-level critical values for FMAX and F' appropriate to the stage where p variables have been entered.

Hence it follows from the statements at (6) that, even if the procedure succeeds in entering p variables, the probability of proceeding further is bounded by the value of α.

Further to the result at (5), since the tests involved at each step are consistent†, strict equality will be obtained in this expression as the sample size n becomes large, i.e.

$$\lim_{n\to\infty} \text{Prob}[r_1 > p] = \alpha \qquad (7)$$

More specifically the consistency property ensures a zero asymptotic probability of omitting "true" regressors (i.e. underfitting) i.e.

$$\lim_{n\to\infty} \text{Prob}[r_1 < p] = 0 \qquad (8)$$

†This in turn follows from the consistency of least squares estimators in the linear model. Lehmann refers to this property as procedure consistency.

Backward procedures are more difficult to evaluate in this way. It is no longer possible to relate the value of α to the final order of equation, r_2 say, as in (5) or (7). The consistency property does however carry over so that

$$\lim_{n \to \infty} \text{Prob}[r_2 < p] = 0 \qquad (9)$$

and hence large sample protection against underfitting is obtained.

The above points do serve to illustrate the sort of limitations imposed on the use of stepwise regression procedures in general. It is argued however that, in view of the scope of the problem being faced, even some kind of asymptotic optimality is better than the unpredictable consequences resulting from the conventional use of stepwise regression. Proceeding along these lines it would seem that, at least in the orthogonal case, forward procedures have more to recommend them than backward ones if n is large (or, strictly, if n-k is large). In the finite case the situation is by no means as clear-cut. Further discussion on these general lines is however more conveniently postponed till after some other possible stopping criteria have been discussed.

(c) The goodness of fit statistic

By this is meant the calculated value of the square of the multiple correlation coefficient, R^2. This is merely the ratio of the explained to the total sum of squares. Denoting this quantity at the r^{th} stage by R_r^2 it is evident that its distribution will depend on the values of the non-zero β coefficients, i.e.

$$R_r^2 = \sum_{i=r+1}^{k} S_{(i)}^2 \Big/ \left(\sum_{i=1}^{k} S_{(i)}^2 + S_0^2 \right).$$

Direct use of this statistic does not therefore seem a promising possibility.

For the same reason the so-called "Corrected" R_r^2, defined by

$$\bar{R}_r^2 = 1 - \left(\frac{n-1}{n-r-1}\right)(1-R_r^2),$$

does not seem to be very useful either. We might note however that we can write

$$R_r^2 = 1 - \left(\frac{n-r-1}{n-1}\right)\left(\frac{v_r}{v_o}\right)$$

$$\bar{R}_r^2 = 1 - \left(\frac{v_r}{v_o}\right)$$

from which it follows that the underlying distributional theory can in fact be discussed in terms of that of v_r.

(d) The Mallows C_r Statistic

In two unpublished papers [54 , 55] C.L. Mallows proposed the use of the "Standardized Total Squared Error" for the comparison of two regressions. As an estimate of this quantity he suggested

$$C_r = RSS_r / \hat{\sigma}^2 - (n-2r-1)$$

where r is the number of fitted variables.

RSS_r is the residual sum of squares

$\hat{\sigma}^2$ is an estimate of σ^2, and is usually taken as the residual variance estimate v_k.

Mallows demonstrates that regressions with "small bias" have C_r approximately equal to r. Hence a possible criterion is the nearness of C_r to r together with the actual magnitude of C_r.

It might be noted that, in a non-stepwise context,

$$E[RSS_r] = (n-r-1)\sigma^2$$

(provided $r \geq p$) from which it follows that

$$E[C_r] \doteq (n-r-1) - (n-2r-1) = r.$$

In fact we may also write

$$C_r = (n-r-1)v_r/v_k - (n-2r-1)$$

which is essentially the F' statistic given in (3).

(e) Change in R^2

We now consider the quantity

$$\Delta R_r^2 = R_{r+1}^2 - R_r^2$$

i.e. the change in R^2 which results by including the variable with highest contribution. In the previous notation we have

$$\Delta R_r^2 = S_{(q)}^2 \Big/ \left(\sum_{i=1}^{k} S_{(i)}^2 + S_0^2 \right)$$

or, alternatively we can write

$$\Delta R_r^2 = \left\{ (n-r-1)v_r - (n-r-2)v_{r+1} \right\} \Big/ TSS$$

where TSS is the overall sample sum of squares.

To avoid the complexities arising from the denominator in these expressions it is better to consider

$$\Delta R^2_r . TSS$$

which is just the quantity $S^2_{(q)}$. The conventional F statistic described in (a) above is essentially based on this quantity but, notwithstanding the objections to its use on more general grounds, the appropriate tables for its implementation were in any case stated to be not readily available. A natural way round the problem is to use a similar approach to that of F' in (b). This leads directly to the quantity

$$S^2_{(q)}/S^2_0$$

which, when multiplied by n - k - 1, gives the statistic

$$FMAX = S^2_{(q)}/v_k \qquad (10)$$

We leave aside the problem of the distribution of FMAX until the next section but one. The incorporation of FMAX into both types of stepwise procedure will follow the lines used for F'. Again only constant significance levels are contemplated thus avoiding the admittedly more flexible but nevertheless more complicated possibilities of varying levels. The expressions (5), (7), (8) and (9) will still hold but with different stochastic quantities r_1 and r_2. The main difference will arise out of (6), (7) and (8) in terms of the rate of the stochastic convergence.

This in turn can be expected to depend on, amongst other things, the value of the elements of $\underset{\sim}{\beta}$.

5.5. F' and FMAX as test statistics of a conditional hypothesis

The two most plausible test criteria arising out of the previous section, F' and FMAX, can be compared in general terms with respect to their expected performance by using the framework of Chapter 3. For in the notation of (3.1.6) and (3.1.7) we are essentially contemplating a test of the conditional hypothesis

$$H_o \equiv \underset{\sim}{\mu} = \underset{\sim}{0}$$

against the alternative

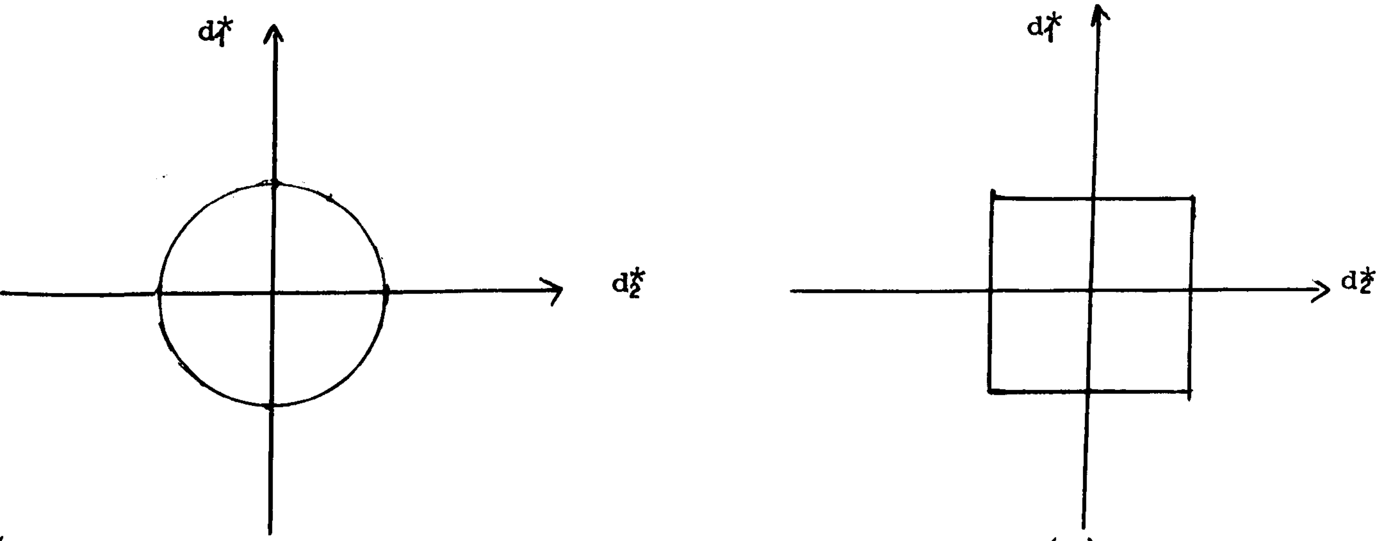
$$H_1 \equiv \underset{\sim}{\mu} \neq \underset{\sim}{0} \, ,$$

using the information contained by the values of $\underset{\sim}{d}$. The conditional basis arises from the assumption that the regressors already fitted are "true" ones i.e. they appear in the underlying model with non-zero coefficients. The orthogonality property considerably simplifies the covariance matrix $\underset{\sim}{V}_d$ of $\underset{\sim}{d}$ by making it diagonal.

Little work has been done on the power of the statistic FMAX apart from that of Ramachandran [68] who showed that the power function is monotonic in each element of $\underset{\sim}{\mu}$. The power of the F' test is of course well established and was discussed in section 2 part (b) above. The main difference between the two tests can be illustrated using the simple two-dimensional case. If we write

$$\underset{\sim}{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

then the acceptance region for F' is circular as in diagram (a) below, whilste that for FMAX is square as in diagram (b). (Note: These diagrams relate to the standardized vector $\underset{\sim}{d}^*$. The appropriate regions for $\underset{\sim}{d}$ will be, respectively, elliptical and rectangular).



(a)

(b)

Although these regions generalize in a straightforward way to hyperspheres and hypercubes respectively it is not easy to come to precise conclusions for when these tests are used in stepwise regression.

It is interesting to note that the use of F' is also, like FMAX, equivalent to a test of a maximal quantity, namely (in the general case)

$$\underset{\Phi}{\text{Max}} \sum_{i=1}^{q} \phi_i d_i^*$$

where Φ is the space of $(\phi_1, \phi_2, \ldots, \phi_q)$ subject to the normalizing constraint

$$\sum_{i=1}^{q} \phi_i^2 = C, \text{ an arbitrary constant.}$$

This is in fact just a special case of a more general result concerning tests on linearly independent estimable functions, a proof being given by Scheffé [72 , p.70]. A similar interpretation will also apply in the non-orthogonal regression situation which is discussed later.

We can perhaps gain some kind of qualitative feeling for the potential relative performances of F' and FMAX by comparing (1) with the corresponding expression for F', i.e.

$$\underset{i=1,\ldots,q}{\text{Max}} |d_i^*|$$

In the orthogonal case at lease one might, in later stages, expect FMAX to be more sensitive in detecting the relatively few true regressors which remain among the excluded set. In the non-orthogonal situation however even this possible slight advantage of FMAX over F' would seem to be lost. Again we shall have to rely on the results of empirical investigation to throw some light on the problem.

5.6. The distribution of FMAX

We return now to consider the distribution of the statistic FMAX defined at (2.9). FMAX is just the largest of the q quantities

$$d_i^{*2} = S_i^2/v_k, \qquad i = 1,\ldots\ldots,q.$$

The underlying distribution is the quasi-independent form of the multivariate F-distribution, the dependence only arising out of the common presence of the denominator v_k. The problem of the distribution of the extremum was first looked at by Hartley [33] who investigated the case where one has q independent estimates of σ^2, S_i^2 $(i = 1,\ldots,q)$, each having m degrees of freedom. Given a further independent estimate S_o^2 with ν degrees of freedom Hartley derived an iterative method for the distribution function of

$$\text{FMAX} = \text{Max}(S_o^2/S_o^2)\ , \qquad i = 1,\ldots,q.$$

The approximation was not too good however in the upper tails, especially when m = 1. This is of course the case of interest here.

Nair [58] used a more efficient approximation procedure and tabulated the upper 5% and 1% points of FMAX for q = 1(1)10 and ν = 10,12,15,20,30,60,∞. These tables are perhaps more readily available in Pearson and Hartley's Biometrika tables [65]. Some slightly more extensive tables have been produced by Krishnaiah and Armitage [42] covering values of $q = 1(1)12$ and $\nu = 5(1)45$. There are a few differences between these latter tables and Nair's on the overlapping sections but on the whole these appear to be negligible.

It might be noted that the case $\nu = \infty$ produces a multivariate chi-square distribution, corresponding to knowledge of σ^2. The critical points for a fixed value of q are then easily determined using another approach, i.e.,

$$\text{Prob}[\text{Largest } \chi^2_{(1)} \leq c] = \alpha$$

implies

$$[2\ \Phi(\sqrt{c})]^q = \alpha$$

and hence

$$[2\ \Phi(\sqrt{c}) = 0.5\alpha^{1/q}$$

where $\Phi(x)$ is the area enclosed between 0 and x (> 0) by a standard normal curve. Using a standard Algol procedure [16] which gives $\Phi(x)$ to 11 decimal places the 1% and 5% critical values of FMAX were determined by interpolation from a fine grid of values of c. Values for $q = 2(1)40$ are presented in Appendix 2. There is complete agreement with Nair's results on the overlapping section.

5.7. A simultaneous inference approach

It was mentioned earlier in Chapter 4 that Lehmann [50] has developed the general theory for arriving at optimal procedures in a certain class of multiple decision problems. Since the basic structure of the stepwise identification problem for orthogonal regressors lends itself to a treatment in this way it is worthwhile investigating this further. Substituting our established notation into Lehmann's more abstract formulation we can list the main requirements of the class considered:

(i) The decision procedure must consist of the simultaneous application of a number of different hypothesis tests, each of the form

$$H_o \equiv \underset{\sim}{\beta} \in B^*_j, \qquad j = 1,\ldots.,m,$$

where B^*_j is a subset of the parameter space $\mathcal{B}$. In each case the alternative hypothesis involves the entire complement set of B^*_j.

(ii) The individual tests must be <u>compatible</u>, i.e., they must not lead to conflicting decisions.

(iii) The overall loss function is defined <u>additively</u> over the m component tests.

(iv) Only <u>unbiased</u> procedures are contemplated, i.e. procedures for which

$$E[L(d^{**},\underset{\sim}{\beta})] \geq E[L(d^{*},\underset{\sim}{\beta}^{*})]$$

for all d^* and d^{**}, and all $\underset{\sim}{\beta}^*$, where
d^* is the decision that $\underset{\sim}{\beta} = \underset{\sim}{\beta}^*$
d^{**} " " " " " $= \underset{\sim}{\beta}^{**}$
and $\underset{\sim}{\beta}^*$ is the true value of $\underset{\sim}{\beta}$.
This definition of procedure unbiasedness is again due to Lehmann [49]. Descriptively it amounts to the requirement that one should come closer to the correct decision on average than any incorrect one.

For problems falling within the above class Lehmann presents the general theory for arriving at a procedure which uniformly minimizes the risk.

We now consider the possibility of using such a decision structure for the stepwise identification problem. The natural hypotheses to employ in (i) are the set

$$H_o \equiv \beta_j = 0$$

with alternatives $H_1 \equiv \beta_j \neq 0,\qquad j = 1,\ldots.,k.$
The simultaneous application of such tests will certainly satisfy the compatibility requirement (ii). With regard to the loss structure we need to specify, for each $j = 1,\ldots,k$, the losses a_j and b_j incurred in falsely rejecting or accepting $H_o \equiv \beta_j = 0$. If, a priori, all regressors are valued equally, and if over- and under-fitting are regarded as being equally

disadvantageous, then we can take $a_j = b_j = 1$ for $j = 1,\ldots,k$.

With this particular specification it follows (see Lehmann [50,p.72]) that the individual tests must also be unbiased. This in turn requires that each such test must be carried out at the level of significance

$$\alpha_j = b_j \big/ (a_j+b_j), \qquad j = 1,\ldots,k.$$

In the case where $a_j = b_j = 1$ it follows that each test is performed at the 50% level. This amounts in fact to the use of FMAX with a constant critical value. We might in fact more profitably consider the general class of procedure which use FMAX with some constant critical value. To each such value will correspond a constant marginal significance level $\alpha_j = \alpha$, say, for the individual tests. While this class of procedures will of course be biased in general they do possess the advantage of avoiding the ambiguity associated with forward and backward approaches based on other stopping criteria. Thus we may now write $r_1 = r_2 = r$, say. Although it is no longer possible to make simple probability statements for the case of finite n as in (2.5) the consistency property of least squares estimation still ensures that

$$\lim_{n\to\infty} \text{Prob}[r < p] = 0 \qquad (1)$$

as in (2.8). The analogous expression to (2.7) is however not so readily obtainable. In fact, if c is the upper $\alpha\%$ critical value for the F-distribution with 1 and n - k - 1 degrees of freedom, then

$$\lim_{n\to\infty} \text{Prob}[r > p] = \gamma \qquad (2)$$

where γ is the probability that FMAX exceeds c, and where the q parameter of FMAX has the value k - p.

A procedure characteristic which it is possible to find in the class being considered is that of the expected number of redundant variables appearing in the selected equation. For if U_i, $i = 1,\ldots\ldots,k-p$, are indicator variables associated with each of the k - p variables concerned then the required expectation is immediately given by

$$\sum_{i=1}^{k-p} E[U_i] = (k-p)\alpha \qquad (3)$$

For reference purposes we henceforth refer to the above-described use of FMAX as F*.

5.8. Summary

In this chapter have been discussed various possible test statistics for incorporation in stepwise regression procedures, and it has been seen that the particular quantities F', FMAX and perhaps F* offer some degree of plausibility. Whilst ideally one would now like to investigate the relative performance characteristics in terms of the analytical framework of chapter 4, it has already been remarked that it is not possible to follow through at such a level of rigour. Instead of this approach, therefore, some less objective evaluations will be attempted later based on the results of empirical studies.

Chapter 6 Prediction with Orthogonal Regressors

6.1 General considerations

Having examined possible test criteria for the identification problem we now turn to the much more ambitious task of prediction which was discussed in general terms in Chapter 4. Again the decision rules considered will be constrained by the basic one-step stepwise algorithm, decisions being made at each stage as to whether to terminate or proceed to enter (or delete)another variable. We will restrict the discussion to that of prediction in the context of the classical linear model thus avoiding many of the not inconsiderable difficulties mentioned in(4.6). It is in any case unrealistic to impose the orthogonality condition in a stochastic regression set-up. In particular we will be concerned, at least initially with the prediction of the 'response' to a single set of regressor values $\underset{\sim}{x}$.

We suppose the stage has beenreached at which r regressors have been fitted. We are then faced with the decision as to whether it is worth adding extra variables when judged in terms of the expected change in mean square prediction error which would result. A major difficulty which arises is that the vector estimator $\underset{\sim}{b}_{(r)}$ corresponding to the r included variables will not have the classical least squares properties of unbiasedness and normality. This is because the induced distribution is conditioned by the fact that the r^{th} stage has actually been reached. More particularly, even in the orthogonal case being considered, we cannot in general regard the elements of $\underset{\sim}{b}_{(r)}$ as being distributed independently either of each other or of coefficient estimates obtained at later stages. The consequences of this will become apparent below.

Corresponding to the vector $\underset{\sim}{b}_{(r)}$ will be a set of r of the k possible regressors. To avoid what appear to be complexities of an intractable nature it will now be assumed that the r variables entered correspond to the first r largest values of $\beta_j^2 \sum_{i=1}^{n} X_{ij}^2$, $j = 1,\ldots,k$. We will also suppose that these r values are non-zero (i.e. $\beta_j \neq 0$), and without loss of generality the corresponding regressors will be taken to be $X_1,\ldots,X_r$. If we denote by $\underset{\sim}{b}^*_{(r)}$ the augmented estimator obtained by setting the remaining k-r regression coefficients equal to zero, i.e.

$$b^*_{(r)} = \begin{bmatrix} \underset{\sim}{b}_{(r)} \\ \underset{\sim}{0} \end{bmatrix} ,$$

then the mean square-error of prediction associated with the r^{th} stage is given by

$$MSE_{(r)} = \sigma^2 + \mathrm{Var}\ [\underset{\sim}{x}'\underset{\sim}{b}^*_{(r)}] + B^2_{(r)} \qquad (1)$$

where $B(r) = \underset{\sim}{x}'E[\underset{\sim}{b}^{*}_{(r)}-\underset{\sim}{\beta}]$.

Although a formal verification would be difficult it does seem reasonable to suppose that, for large n, the vector estimator $b_{(r)}$ converges in distribution to the marginal (non-sequential) estimator $\underset{\sim}{b}_r$ of classical least squares. Such a conclusion is in fact suggested by the discussion in (5.2). With such an assumption we can then write (1) in the form

$$MSE_{(r)} \doteq \sigma^2 + \sum_{j=1}^{r} x_j^2 \ Var[b_j] + B^2_{(r)} \tag{2}$$

where $B_{(r)} = - \sum_{j=r+1}^{k} x_j\beta_j$.

We might just pause to note an important difference between the identification problem previously looked at and the prediction problem now under investigation. For in addition to the 'nuisance' aspect of deciding on an appropriate value of $\underset{\sim}{x}$, a further major difficulty is the non-additive contributions of individual regressors to the overall bias component $B^2_{(r)}$. Hence one is no longer able to evaluate regressors in isolation from the remaining ones as in the identification case.

A point of crucial importance which has to be considered is whether the order of variable introduction or deletion is justifiably related to the S^2_j quantities. To examine this suppose, without loss of generality, that X_{r+1} is introduced into the regression equation. The resulting change in mean square prediction error is given by

$$\begin{aligned} MSE_{(r)} - MSE_{(r+1)} &\doteq \Big(\sum_{j=r+1}^{k} x_j\beta_j\Big)^2 - \Big(\sum_{j=r+2}^{k} x_j\beta_j\Big)^2 - x^2_{r+1}\mathrm{Var}[b_{r+1}] \\ &= x^2_{r+1}\beta^2_{r+1} + 2x_{r+1}\beta_{r+1}\sum_{j=r+2}^{k} x_j\beta_j - x^2_{r+1}\mathrm{Var}[b_{r+1}] \end{aligned} \tag{3}$$

It is now that we come up against the difficulty of choosing $\underset{\sim}{x}$. For suppose that X_{r+2} were to be entered at the r^{th} stage instead of X_{r+1} (which is now also assumed to correspond to the maximum S^2_j, $j=r+1,\ldots,k$). The corresponding expression to (3) is then

$$x^2_{r+2}\beta^2_{r+2} + 2x_{r+2}\beta_{r+2}\sum_{\substack{j=r+1 \\ j\neq r+2}}^{k} x_j\beta_j - x^2_{r+2}\mathrm{Var}[b_{r+2}] \tag{4}$$

Forming the difference between (3) and (4) we obtain

$$\begin{aligned} &x^2_{r+1}[\beta^2_{r+1}-\mathrm{Var}[b_{r+1}]] - x^2_{r+2}[\beta^2_{r+2}-\mathrm{Var}[b_{r+2}]] \\ &\quad + 2[x_{r+1}\beta_{r+1} - x_{r+2}\beta_{r+2}]\sum_{j=r+3}^{k} x_j\beta_j \end{aligned} \tag{5}$$

It is evident that, unless some severe restrictions are imposed on $\underset{\sim}{\beta}$, there are no grounds for supposing (5) to be a positive definite quadratic form in the x's.

Although, for a specific choice of $\underset{\sim}{x}$, one could contemplate introducing regressors according to their maximum contribution in terms of (3) it does not in fact seem a realistic course to pursue. In any case the distributional problems are of a much greater order of complexity than in the analogous use of FMAX for the identification situation. We might however attempt to circumvent the nuisance aspect of $\underset{\sim}{x}$ by relating our prediction objective to a typical set of such values. We are then effectively led into dealing with expectations (over $\underset{\sim}{x}$ values) of quantities like (1), (3), (4) and (5). In the absence of more specific information we might then be prepared to assume that

(i) over the area of interest for $\underset{\sim}{x}$ the individual regressors are orthogonal (or independent).

and(ii) the sample ranges of variation at the fitting stage reflect potential future ranges.

Thus, taking $E[x_j x_\ell] = 0 \; (j \neq \ell)$ and $E[x_j^2] = A \sum_{i=1}^{n} X_{ij}^2$ (where A is an arbitrary positive constant of proportionality) we can rewrite (1) as

$$\begin{aligned} MSE_{(r)} &= \sigma^2 + A \sum_{j=1}^{r} \left(\sum_{i=1}^{n} X_{ij}^2 \right) \left(\sigma^2 / \sum_{i=1}^{n} X_{ij}^2 \right) \\ &\quad + A \sum_{j=r+1}^{k} \beta_j^2 \left(\sum_{i=1}^{n} X_{ij}^2 \right) \\ &= \sigma^2 (1+Ar) + A \sum_{j=r+1}^{k} \lambda_j \qquad (6) \end{aligned}$$

where we retain the notation $\lambda_j = \beta_j^2 \sum_{i=1}^{n} X_{ij}^2$.

Consider now the reduction in expected prediction square error resulting from the extra inclusion at the fitting stage of an arbitrary regressor X_s, say, to an equation already containing r arbitrary regressors. From (6) we have

$$MSE_{(r)} - MSE_{(r+1)} = A(\lambda_s - \sigma^2) \qquad (7)$$

We can regard (7) as representing the predictive contribution offered by the use of X_s at the fitting stage. Further, if we consider the total predictive contribution obtained be using all k regressors, we have

$$MSE_{(0)} - MSE_{(k)} = A \sum_{j=1}^{k} \lambda_j - Ak\sigma^2 \qquad (8)$$

Hence we obtain a decomposition of the 'total predictive potential' into contributions from individual regressors.

We might consider for a moment whether the expected order of entry, which from (5.2.1.) is govenned by the magnitudes of the λ_j values, is in agreement with the prediction objective. That this is so follows immediately from the result at (7).

Having thus presented some degree of justification for continuing to use the maximum sum of squares criterion we can now turn to the problem of choosing an appropriate test structure. Thus will now be approached by way of seeking analogues to the statistics FMAX and F' which arose in the identification situation.

6.2. The FMAX approach

The reasonableness of testing, for a single specific determination $\underset{\sim}{x}$ of the regressor variables, the extra predictive contribution obtained by including another variable has already been seriously questioned. But even apart from such considerations one is still faced with a non-trivial problem of distribution theory. For the expression at (6.1.3), on substitution of sample estimates for the unknown regression coefficients, does not yield a distributional form which is in any sense recognisable.

We might instead, however, consider the possibility of using the simplified version of (6.1.3) given at (6.1.7), especially since the implications of this were shown to be more consistent with the use of the S_j^2 criterion for variable entry. We are thus led to consider the hypothesis:

$$H_o = \lambda_{r+1} \sigma^2 / \leq 1 \qquad (1)$$

On substituting b_{r+1} for β_{r+1} and v_k for σ^2 we obtain the test quantity

$$FMAX(p) = b_{r+1}^2 \sum_{i=1}^{n} X_{i,r+1}^2 \Big/ v_k \qquad (2)$$

While computationally this is just the conventional F-value for the estimate b_{r+1} its null distribution (taking equality in (1)) is, in a marginal sense, non-central F(1,n-k-1, 1). Hence, in order to implement such a statistic at the r^{th} stage of a stepwise procedure, one would require the use of Studentised largest non-central chi-square statistics. It can be said with some degree of certainty that such tables do not at present exist.

What we are able to infer from the above considerations is that treatment of the quantity FMAX(p) (or equivalently FMAX) as though it were distributed as FMAX in an identification context will lead to non-conservative conditional tests. While admittedly this may or may not be serious in the context of a complete procedure (in the sense that the optimum significance levels to emply are not known in any case) it does indicate that an underfitting tendency in an identification procedure may be desirable from a prediction viewpoint. This is something which again will have to be examined in the light of the empirical results to be presented later.

6.3 The F' approach

We now turn to the possibility of adapting the F' statistic to the prediction case. The natural quantity to consider here is, in the case of prediction for a single determination $\underset{\sim}{x}$ of the regressors, then given by

$$\mathrm{MSE}_{(r)} - \mathrm{MSE}_{(k)} = \mathrm{Var}[\underset{\sim}{x}'\underset{\sim}{b}^*_{(r)}] + B^2_{(r)} - \mathrm{Var}[\underset{\sim}{x}'\underset{\sim}{b}] \qquad (1)$$

(where we note that $B_{(k)} = 0$ by assumption).

The quantity given in (1) is in fact essentially the same as that considered by Toro-Vizcarrondo and Wallace [74]. These authors are concerned with the different problem as to whether the imposition of a <u>given</u> set of linear restrictions is desirable in terms of the mean square-error estimation of any (and all) linear functions of $\underset{\sim}{\beta}$. If however we continue to assume that the order of variable entry is conditioned by the magnitudes of $\beta^2_j \sum_{i=1}^{n} X^2_{ij}$, and if we also assume that $\underset{\sim}{b}_{(r)}$ is distributed as $\underset{\sim}{b}_r$, the two problems are basically the same. Hence, making use of the orthogonality present, we may write (1) as

$$\mathrm{MSE}_{(r)} - \mathrm{MSE}_{(k)} = \left(\sum_{j=r+1}^{k} x_j\beta_j\right)^2 - \sum_{j=r+1}^{k} x^2_j\mathrm{Var}[b_j] \qquad (2)$$

The relevant criterion which determines whether it is worth adding extra variables at the r^{th} stage can be written as

$$\left(\sum_{j=r+1}^{k} x_j\beta_j\right)^2 \Big/ \sum_{j=r+1}^{k} x^2_j\mathrm{Var}[b_j] \leq 1 \qquad (3)$$

Unlike in the case of FMAX(p), substitution of least squres estimates of the unknown regression parameters does now yield a fairly reasonable distribution. For $\sum_{j=r+1}^{k} x_j b_j$ will be distributed as

$$N\left(\sum_{j=r+1}^{k} x_j\beta_j, \sum_{j=r+1}^{k} x_j^2 Var[b_j]\right)$$

Thus we can form the test statistic

$$F'(p) = \left(\sum_{j=r+1}^{k} x_j b_j\right)^2 \Big/ \left(v_k \sum_{j=r+1}^{k} [x_j^2 \Big/ \sum_{i=1}^{n} X_{ij}^2]\right) \quad (4)$$

Taking equality to hold in (3), and noting the independence of v_k and $\underset{\sim}{b}$, it follows that F'(p) is distributed as non-central F(1,n-k-1, 1). Although tables of this distribution do exist (see for example Toro-Vizcarrondo and Wallace [74]) its use in stepwise algorithms will not be investigated in this study. This is because it is considered more realistic and useful to investigate the effect of taking expectations over $\underset{\sim}{x}$ in the sense described in the opening section of this chapter. On doing this the resulting modification of (3) becomes

$$\sum_{j=r+1}^{k} \beta_j^2 \sum_{i=1}^{n} X_{ij}^2 \Big/ \left(\sum_{j=r+1}^{k} Var[b_j] \sum_{i=1}^{n} X_{ij}^2\right)$$

$$= \sum_{j=r+1}^{k} \lambda_j \Big/ \{(k-r)\sigma^2\} \leq 1 \quad (5)$$

Substituting the estimates of the β parameters and σ^2 we recognize the numerator as the usual explained sum of squares quantity due to the k - r fitted regressors. The statistic which results is in fact computationally identical to F' in the identification case. The difference now is that, under the relevant null hypothesis, we need to refer to tables of the non-central F(k-r, n-k-1, k-r) distribution. While again the lack of tabulated critical values precludes the use of such a test at present we might infer, as in the case of FMAX(p), that F' identification procedures which tend to underfit may be desirable for prediction purposes.

At this juncture we might pause to notice a change in the underlying hypotheses which FMAX and F' are testing as compared with the identification case. For in the latter situation the null hypotheses are <u>identical</u> in that they specify all the λ_j ($j = r+1,\ldots,k$) to be zero. In the present situation of prediction we see that, while the null hypothesis for FMAX(p) is that

$$\lambda_j \leq \sigma^2 \quad \text{for } j = r+1,\ldots,n,$$

from (5) we see that the corresponding null hypothesis for F'(p) requires that the <u>average</u> λ_j is less than σ^2. The null hypothesis parameter space for FMAX(p) is thus a strict subset of that for F'(p). The difference arises out of the fact that F'(p) is concerned with the decision whether or not to enter <u>all</u> the $k-r$

excluded variables, and thereofe involves a balancing out of large and small predictive contributions. FMAX(p) on the other hand concentrates on the most promising *single* potential new entrant to the equation. However, despite this seeming advantage of the FMAX(p) approach, one has also to take into account the relative power characteristics of the two test quantities involved. We must again await the outcome of the simulation studies in order to pass any kind of judgment on this.

It is interesting to note another type of conditional hypothesis which leads to the same test statistic given at (5) but whose null distribution is non-central F(k-r, n-k-1, 1). For suppose we seek a condition on $\underset{\sim}{\beta}$ such that the expression at (2) is negative semi-definite, or equivalently that (3) holds for all $\underset{\sim}{x}$. If we let $\underset{\sim}{\beta}^*$ and $\underset{\sim}{x}^*$ represent the sub-vectors consisting of the last k - r elements of $\underset{\sim}{\beta}$ and $\underset{\sim}{x}$ respectively, and also if $\underset{\sim}{X}^*$ is the sub-matrix of the last k - r columns of $\underset{\sim}{X}$, we can write the left-hand side of (3) as

$$(\underset{\sim}{\beta}'^*\underset{\sim}{x}^*)^2/(\sigma^2\underset{\sim}{x}'^*(\underset{\sim}{X}'^*\underset{\sim}{X}'^*)^{-1}\underset{\sim}{x}^*) \qquad (6)$$

This expression satisfies the conditions of a form of the Cauchy-Schwarz Inequality, from which it follows that a maximum is attained when

$$\underset{\sim}{x}^* = (\underset{\sim}{X}'^*\underset{\sim}{X}^*)\underset{\sim}{\beta}^* \qquad (7)$$

(See Rao[69 ,p.48]).

Substituting (7) into (6) it follows that (2) is negative semi-definite if and only if

$$\underset{\sim}{\beta}'^*(\underset{\sim}{X}^{*\prime}\underset{\sim}{X}^*)\beta^*/\sigma^2 < 1 \qquad (8)$$

Replacement of the unknown quantities by the usual estimates leads to the non-central F(k-r,n-k,1) distribution (i.e. when the estimated left-hand side of (8) is divided by (k-r).

As is reasonable on intuitive grounds the resulting procedure will tend to overfit in comparison with the previous version based on (5). It is questionable however as to whether the particular hypothesis structure is appropriate from an overall viewpoint. One could for instance argue equally well for a test of the positive definiteness of (2), though such a test is not so readily available.

6.4 Limitations of the approach

Since, in an identification context, the choice of significance level α was seen to furnish some kind of asymptotic control over procedure performance we might consider whether such guidelines are available in the prediction case. It must however, be remarked that the whole character and scope of the problem is different to that of identification, especially from an asymptotic viewpoint. For the consistency property of least squares when applied to the complete regression equation ensures that, with r = k in (6.1.1)

$$\lim_{n\to\infty} MSE_{(k)} = \sigma^2$$

One must first of all, therefore, be prepared to judge procedure performance placing more emphasis on non-asymptotic relative efficiency criteria.

In endeavouring to optimise in some way the choice of α, or at least in trying to establish the precise nature ofits role, we immediately come up against some major difficulties. For suppose we look again, for example, at the application at each procedure step of the FMAX(p) criterion given at (6.2.1) and (6.2.2). We must of course retain the assumption of large samples in order that the orthogonality property of the stepwise induced elements of the estimator $\underset{\sim}{b}_{(r)}$ is still approximately tenable. This asymptotic framework then carries with it the almost certain inclusion of all regressors with non zero coefficients. This will occur as a consequence of the monotonically increasing component $\sum_{i=1}^{n} X_{ij}^2$ of the associated non-centrality parameter λ_j. The success with which this is achieved will however, for moderate and small sample sizes at least, depend a great deal on the level of α. At the same time of course onestill wishes to guard against too high a level of α causing regressors to enter which will have a negative net contribution to future predictive performance.

Although it will not be pursued here there is again the possibility of using a simultaneous decision approach along the lines of F*. Such a procedure could presumably be derived from the simultaneous application of non-central F tests on hypotheses based on (6.1.7). However, it does seem preferable, for the rest of this study at least, to concentrate on the more easily applicable test criteria such as conventional F, FMAX and F'. This is especially desirable in the more general situation of non-orthogonal regression in which the concept of simultaneous testing would seem to have little place.

In the next chapter we proceed to describe an empirical study of various procedures which have arisen for use with orthogonal regressor set-ups.

Chapter 7 An Empirical Study of the Orthogonal Case

7.1 Scope of study

The motivation for carrying out empirical investigations using simulated data has already been established in much of the preceding discussion. However, despite the advantages of being able to avoid considerable analytical complexity, the "nuisance" parameter space is still such as to prevent anything but a cursory examination of the relative performances of procedures. The scope of the study will in fact be restricted to a comparison only of the FMAX, F' and conventional F sequential procedures, and will be directed mainly at the relatively less demanding objective of identification as described in chapter 4. Predictive performances will however also be examined in the light of some of the results of chapter 6. Procedures based on a simultaneous inference approach are not looked at since they are contrary to the basic conditional test philosophy which seems to underly stepwise regression (a feature which is perhaps more apparent in the non-orthogonal case to be discussed later).

Since the three test statistics investigated can each be used in both a forward and backward manner we have six different procedures in all.

For reference purposes we number these procedure as follows:

1. Forward procedure using FMAX
2. Forward procedure using F'
3. Forward procedure using conventional F

$\left.\begin{array}{l}4.\\5.\\6.\end{array}\right\}$ As 1, 2 and 3 but using a backward approach

Another procedure which was briefly entertained was that of using conventional F-tests with a critical value of unity. The motivation for this stems from the fact that it can be shown that an equation of order r has larger corrected R^2 (equivalent to smaller residual variance) than the $(r+1)^{th}$ order model obtained by adding another variable if and only if the (partial) F-value of this new variable is less than unity (see Haitovsky [31]). It was asserted by Lott [52] that econometricians use largest corrected R^2 as an optimality criterion in model selection, and this author goes on to use it in a stepwise regression analysis with orthogonal regressors. The results obtained in the simulation runs using such a criterion resulted, as expected, in extreme overfitting and were not thought to be worth reporting in any detail.

We can now turn to the problem of deciding on appropriate model formulations to use in the investigation.

The main factors which would seem to influence procedure performance (within the framework envisaged) are:-

(i) The significance level α

(ii) The value of $n - k$

(iii) The value of k

(iv) The values of $|\beta_i|$, or β_i^2, for i = 1,....k.

(v) The sample variance of each X variable

(vi) The error variance σ^2.

One could perhaps incorporate the last three of these into the non-centrality parameter $\lambda = \beta^2 \; \Sigma \; X^2/\sigma^2$ which is relevant to the conditional tests used. However, it is not absolutely clear that this follows for procedures involving the sequential use of an undetermined number of such tests.

Since the basic aims of the exercise were to substantiate some of the arguments propounded in favour of FMAX and F' as opposed to the use of conventional F, and also to indicate possible areas for further study, a full grid coverage of values of the influencing parameters was not attempted. In particular only one level of α (5%) was used throughout the investigation. In the case of conventional F, although it is common to use a fixed critical value of 4 throughout (i.e. the large sample 5% point), the values actually used were the ones taken from F tables corresponding to the actual degrees of freedom.

Another limitation arises in the choice of k due to the restricted nature of Studentised maximum chi-square tables. For this reason a value of k = 10 was used in almost all cases. The main study then involved combinations of the three levels of n:

31, 71, 150

with five specifications of $\underset{\sim}{\beta}$

(a) [0 0 0 0 0 0 0 0 0 0]'
(b) [1 0.5 0.5 0.5 0.5 0.25 0.25 0 0 0]'
(c) [3 3 3 0 0 0 0 0 0 0]'
(d) [10 9 8 7 6 5 0 0 0 0]
(e) [3 3 0.125 0.125 0.125 0 0 0 0 0]'

Fifteen different configurations were thus obtained, in each case both the regressor variances and error variance being taken as unity. Since it was not thought entirely realistic for the error variance to be as large as regressor variances three further models were determined by using the specification (b) for $\underset{\sim}{\beta}$ with each of the levels of n, but with regressor variances equal to 9. For reference purposes this latter configuration is denoted by the letter (f). We thus have the six different set-ups (a), (b), (c), (d), (e) and (f) each investigated at the three given levels of n. It was felt that the six chosen specifications of $\underset{\sim}{\beta}$ covered a reasonable range of variation in the underlying model formulation. It might be remarked that case (e) is of especial interest from a prediction viewpoint. For, in the notation of the previous chapter, we see that the λ values associated with each regressor variable are, in the case when n = 31, given by:

279 279 0.48 0.48 0.48 0 0 0 0 0

We thus see that that three of the non-zero λ coefficients are less than the unit value taken for σ^2. Out of all the models considered theory suggests that in this case alone can we expect an underfitted model to yield smaller mean square prediction error than the estimated version of the true model.

Finally, since tables are available for FMAX in the special case in which n - k is very large (see (5.6) and Appendix 2), four further simulation runs were performed for the case n = 180, k = 30 and regressor and error variances equal to unity. The four cases are distinguished according to the $\underset{\sim}{\beta}$ specification as follows:

(g) $\underset{\sim}{\beta} = \underset{\sim}{0}$

(h) $\underset{\sim}{\beta}' = [10 \quad 10 \quad 10 \quad 5 \quad 5 \quad 5 \quad 5 \quad 5 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0.5 \quad 0.5 \quad \underset{\sim}{0}]$

(i) $\underset{\sim}{\beta}' = [10 \quad 10 \quad 10 \quad 10 \quad 10 \quad 5 \quad 5 \quad 5 \quad 5 \quad 5 \quad \underset{\sim}{0}]$

(j) $\underset{\sim}{\beta}' = [1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0.5 \quad 0.5 \quad 0.5 \quad 0.5 \quad 0.5 \quad 0.25 \quad 0.25 \quad \underset{\sim}{0}]$

(where each vector is of dimension 30).

7.2 Description of program for identification

For the purposes of the study described in the previous section a computer program was written which began by generating a $n \times k$ matrix of regressor values. These were then held fixed for the succeeding iterations. By allowing a general linear transformation of this matrix it was possible (in theory) to impose an arbitrary "correlation" structure. The case of orthogonal regressors presented a problem in that variables generated completely randomly will not in general have diagonal sample correlation matrix. Since, for small values of n at least, this proved to be a serious problem a different approach was required. One possibility was to use the values of orthogonal polynomials given by Pearson and Hartley [65]. However, this restricts the range of values of n which can be considered, and would also have involved a fair amount of data punching. The method actually used was the Gram-Schmidt orthogonalisation procedure applied to columns of a randomly generated matrix of standard normal deviates.† The

The simulations were mainly carried out on the Nottingham University ICL 1906A computer (a few initial runs being performed on an English Electric KDF9). All programs were written in Algol 60.

†In this and all succeeding simulation studies the random numbers were generated using Algorithm G05ADA of the Nottingham Algorithms Library [60]. This procedure employs two independent sequences generated by the multiple congruential method, normality being obtained by using the standard"Box-Muller" transformation (see Box and Muller [15], Neave [61]).

procedure incorporated the usual Gram-Schmidt practice of normalizing the matrix columns so that they had unit length and zero mean.

Given the k-dimensional vector of regressor coefficients n "conditional" means were calculated and stored. At each subsequent iteration of the program independent standardized normal random variables were added to the means as residuals, thus generating the Y values. The actual number of iterations chosen for each run was taken as 500 in the identification case, this taking approximately 45 minutes of computing time on KDF9 in the case where n = 71. Whilst the choice of the number of iterations obviously has a bearing on the accuracy of the summary statistics to be described below it was felt that the exploratory nature of the exercise did not warrant the use of more rigour in choosing this number. In any case limitations of computer resources would have prevented a much larger study.

A major problem was faced in deciding how best to summarize the simulation results. Two types of summarization were in fact decided upon. The first of these was a table showing the number of times each regressor variable appeared in the final selected equation. The second type of summarization was a table for each of the six methods showing the percentage number of occasions on which the method over- or underfitted the correct model, and by how many variables. This second type of table was thought to be the most informative, and will be used almost exclusively to describe the results obtained. In their full version each table is a (k+1) × (k+1) matrix in which the rows represent the number of variables overfitted and the columns the number underfitted. All entries in cells other than in either the first row or first column relate to what will be referred to henceforth as "mixed cases", i.e., cases in which incorrect variables are present at the expense of "true" ones.

7.3 Description of program for prediction

As was stated previously predictive error of fitted equations depends very much on the value of $\underset{\sim}{x}$ used at the prediction stage. It is therefore important that procedures be compared using the same $\underset{\sim}{x}$, or more realistically using a set of $\underset{\sim}{x}$ values which are typical of future applications. It was decided therefore to generate, in exactly the same way as for the original $\underset{\sim}{X}$ matrix, a 100 × k matrix of regressor values from which a hundred values of the (exact) conditional mean were calculated. For each model selected by the various stepwise procedures a set of corresponding estimated conditional means was obtained, and a hundred values of squared error were thus determined. These values were then averaged, and again averaged over all iterations of the program to produce what should be fairly reliable figures with which to compare procedure predictive efficiency.

Since the absolute values of mean square error thus obtained could be changed merely by a rescaling of the Y and X data it was thought best to record the results in percentage efficiency form. This was effected by taking the ratio of the optimal value (smallest mean square error) with each of the other values obtained and expressing these figures as percentages. Strictly speaking one should add the residual vaiance σ^2 to each mean square error value to obtain genuine predictive measures. Though this would effect the actual absolute percentage values the ranking of procedures would be unchanged. The major reasons for leaving out the σ^2 contribution were, firstly, that it represents the inherently unpredictable component of variation in the regressand and, secondly, the wider range of prediction error ratios which results from its omission facilitates the subsequent evaluation of prodecures.

Although the basic program structure needed is the same as in the identification case it was decided to write a separate program for prediction evaluations. This has the effect of making both the $\underset{\sim}{X}$ values and the number of iterations different in the two cases of identification and prediction. This was not however thought to be crucial in relation to the kind of inferences which were to be drawn from the results. The number of iterations obtained in each prediction run was in fact 250.

7.4 Presentation of results

Although the results obtained from the simulation runs were already in summarized form as described in the previous two sections, it is possible to draw up considerably more simplified tables whilst still retaining most of the features of interest. Particularly noticeable was that there is a neligible difference in performance between forward and backward procedures of the same type. In fact the only occasions when a difference occurred at all were those for which n = 31. For this reason it was decided to focus attention on the three basic procedure types by looking only at those labelled 1, 2 and 3 in section 1. Also since, in this orthogonal case at least, the so-called 'mixed' cases account for a relatively small proportion of the equations fitted, it did not seem necessary to present a detailed breakdown of all such cases.

The following Tables 7.1 to 7.6 relate respectively to the cases (a), (b),(c),(d),(e) and (f) as specified previously, and Table 7.7 presents the results for cases (g), (h), (i) and (j). Each table is sub-divided into separate tables labelled A and B, relating respectively to the identification and prediction criteria. Type A tables are essentially condensed versions of the identification tables of the second kind described in section 2 and show the percentage number of occasions (iterations) on which various final equations were selected.

Thus the type A tables indicate the distribution of the various degrees of under- and overfitting which occurred and also the proportion of mixed cases. The final column of the tables contains a 'score' which reflects the degree of departure of a fitted equation from the correct one. At each iteration incorrectly omitted and included variables are counted, and these counts are averaged over all the iterations to give the score value recorded.

The second kind of table, type B, records the percentage efficiency ratios based on the mean square prediction errors as described in the previous section. In addition to the three basic methods (1,2 and 3) referred to above two further ones are also featured in this prediction situation. These are:

7. Prediction using the estimated complete equation (i.e. using all K regressors).

8. Prediction using the estimated 'true' model (i.e. only estimating the (known) non-zero coefficients).†

Method 7 should provide an indication of the seriousness of the overfitting aspect associated with not attempting to reject non-informative regressors. Method 8, on the other hand, should indicate the consequences of the

†For reference purposes the various procedures used in this and subsequent simulation studies are listed in Appendix 3.

underfitting tendencies which are anticipated for stepwise procedures based on FMAX and F' in particular.

Figures 7.1 to 7.6 serve as a supplement to the type B tables in the evaluation of the predictive performances of procedures. Each graph shows, for each specification of $\underset{\sim}{\beta}$, how each of the five methods improves in predictive performance as the value of n increases. In order to construct these graphs the underlying <u>absolute</u> predictive efficiency values of the type B tables were used. The smallest of the 15 X k values obtained was then expressed as a percentage of each of the remaining values thus yielding overall percentage predictive efficiency ratios.

7.5 <u>Conclusions</u>

Considering firstly the identification aspect it would seem fair to say that the results just presented substantiate the theoretical arguments put forward earlier. In particular the use of FMAX and F' in methods 1 and 2 respectively demonstrates the asymptotic controlling effect of the choice of α. This is to be contrasted to the use of the conventional F approach, which clearly has no such property. It is remarkable that, in addition to the almost exact agreement between forward and backward procedures of the same type, hardly any difference is apparent between the

TABLE 7.1

($\underset{\sim}{\beta} = \underset{\sim}{0}$; Regressor variances = 1)

TABLE A

Method	n	Number of variables overfitted					score
		0	1	2	3	4	
1	31	94.8	4.6	0.6	-	-	0.06
	71	96.4	3.4	0.2	-	-	0.04
	150	95.4	4.4	0.2	-	-	0.05
2	31	94.4	4.8	0.6	0.1	0.1	0.07
	71	95.8	3.6	0.6	-	-	0.05
	150	94.8	4.6	0.6	-	-	0.06
3	31	56.4	29.2	12.0	2.0	0.4	0.60
	71	59.6	30.8	7.8	1.6	0.2	0.54
	150	60.2	30.2	8.0	1.6	-	0.51

TABLE B

Method	n		
	31	71	150
1	80.6	100	100
2	100	78.2	81.1
3	10.0	9.2	12.9
7	2.5	2.9	3.8
8	-	-	-

TABLE 7.2

($\underset{\sim}{\beta}$'= [1 0.5 0.5 0.5 0.5 0.25 0.25 0 0 0];
Regressor variances = 1)

TABLE A

Method	n	Number of variables under-/overfitted -7	-6	-5	-4	-3	-2	-1	0	1	2	Mixed Cases	Score
1	31	0.4	15.8	25.4	25.0	18.2	8.6	3.2	0.8	0.2	-	2.4	4.1
	71	-	-	0.2	2.0	14.8	37.4	33.4	10.2	0.2	-	1.8	1.66
	150	-	-	-	-	-	10.6	36.8	49.0	2.0	0.2	1.4	0.63
2	31	0.2	5.6	14.0	27.4	25.4	13.2	5.6	1.0	0.4	-	7.2	2.7
	71	-	-	-	0.2	9.2	37.4	39.8	8.0	0.4	-	5.0	1.57
	150	-	-	-	-	-	7.0	45.0	41.4	2.4	0.4	3.8	0.68
3	31	0.2	1.8	5.6	15.4	24.2	24.8	11.8	3.0	0.4	-	12.8	2.7
	71	-	-	-	0.2	1.0	21.2	37.8	26.2	5.0	0.4	8.2	1.06
	150	-	-	-	-	-	1.6	21.4	63.4	9.6	0.4	3.6	0.4

TABLE B

Method	n 31	 71	 150
1	29.2	50.6	67.0
2	34.1	53.3	63.7
3	46.1	69.1	75.3
7	75.4	69.5	68.0
8	100	100	100

TABLE 7.3

$(\underset{\sim}{\beta}' = [3 \quad 3 \quad 3 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$

Regressor variances = 1)

TABLE A

Method	n	Number of variables overfitted				Score
		0	1	2	3	
1	31	93.4	6.4	0.2	-	0.07
	71	95.8	4.2	-	-	0.04
	150	95.6	4.4	-	-	0.04
2	31	94.4	5.4	0.2	-	0.06
	71	95.6	4.2	0.2	-	0.05
	150	95.4	4.2	0.4	-	0.05
3	31	65.4	28.6	6.0	-	0.41
	71	70.0	25.4	4.4	0.2	0.35
	150	71.6	24.6	3.8	-	0.32

TABLE B

Method	n		
	31	71	150
1	87.3	79.1	87.5
2	90.0	81.9	88.5
3	61.6	53.2	63.5
7	29.7	26.3	34.2
8	100	100	100

TABLE 7.4

($\underset{\sim}{\beta}'$ = [10 9 8 7 6 5 0 0 0 0];
Regressor variances = 1)

TABLE A

Method	n	Number of variables overfitted				Score
		0	1	2	3	
1	31	94.8	4.8	0.4	-	0.06
	71	95.0	4.8	0.2	-	0.05
	150	96.0	4.0	-	-	0.04
2	31	95.2	4.0	0.6	0.2	0.06
	71	96.0	4.0	-	-	0.04
	150	96.8	3.2	-	-	0.03
3	31	81.4	16.8	1.6	0.2	0.20
	71	82.0	17.4	0.6	-	0.19
	150	79.8	19.6	0.6	-	0.21

TABLE B

Method	n		
	31	71	150
1	96.1	96.7	96.5
2	95.8	97.0	98.3
3	86.3	87.3	87.3
7	61.9	69.0	69.4
8	100	100	100

TABLE 7.5

($\underset{\sim}{\beta}'$ = [3 3 0.125 0.125 0.125 0 0 0 0 0];
Regressor variances = 1)

TABLE A

Method	n	Number of variables under/overfitted					Mixed Cases	Score
		-3	-2	-1	0	1		
1	31	88.0	8.2	0.4	-	-	3.4	2.9
	71	84.0	10.0	0.8	-	-	5.2	2.9
	150	67.0	25.0	4.2	0.4	-	3.4	2.7
2	31	89.4	6.6	0.4	-	-	3.6	2.9
	71	81.8	10.0	1.2	-	-	7.0	3.0
	150	61.0	26.2	7.0	0.4	-	5.4	2.6
3	31	52.0	19.2	2.2	0.4	-	26.2	2.9
	71	41.4	28.2	7.0	0.8	-	22.6	2.7
	150	28.0	34.0	15.0	4.0	1.4	17.6	2.2

TABLE B

Method	n		
	31	71	150
1	100	74.8	55.9
2	96.6	73.1	54.8
3	63.6	59.1	53.8
7	37.4	43.9	52.2
8	81.5	100	100

TABLE 7.6

$(\underset{\sim}{\beta}' = [\ 1\quad 0.5\quad 0.5\quad 0.5\quad 0.5\quad 0.25\quad 0.25\quad 0\quad 0\quad 0\ ]$
Regressor variances $= 9)$

TABLE A

Method	n	Number of variables under-/overfitted						Mixed Cases	Score
		-2	-1	0	1	2	3		
1	31	1.4	13.2	80.2	4.4	0.4	0.2	0.2	0.22
	71	-	-	95.4	4.4	0.2	-	-	0.05
	150	-	-	95.8	4.2	-	-	-	0.04
2	31	1.0	12.4	81.0	4.6	0.6	0.2	0.4	0.22
	71	-	-	95.6	4.2	0.2	-	-	0.05
	150	-	-	95.2	4.6	0.2	-	-	0.05
3	31	0.2	3.4	82.2	12.4	1.2	0.2	0.4	0.20
	71	-	-	85.8	13.2	1.0	-	-	0.15
	150	-	-	85.8	13.8	0.4	-	-	0.15

TABLE B

Method	n		
	31	71	150
1	66.1	84.3	81.7
2	58.4	84.5	81.4
3	71.3	79.2	76.1
7	59.8	59.8	60.0
8	100	100	100

TABLE 7.7*

(For specification of cases see the end of section 1)

TABLE A

Method	Case	Number of variables under-/ overfitted								Mixed Cases	Score
		-2	-1	0	1	2	3	4	5		
1	g	-	-	94.6	5.4	-	-	-	-	-	0.05
	h	-	-	95.6	4.4	-	-	-	-	-	0.04
	i	-	-	90.2	8.8	1.0	-	-	-	-	0.11
	j	14.2	34.8	42.6	5.4	-	-	-	-	3.0	0.75
2	g	-	-	92.6	3.2	4.2	-	-	-	-	0.12
	h	-	-	96.8	3.2	-	-	-	-	-	0.03
	i	-	-	92.4	5.4	1.2	1.0	-	-	-	0.11
	j	14.2	47.8	34.8	1.2	-	-	-	-	2.0	0.83
3	g	-	-	18.2	31.2	27.0	14.0	8.6	1.0	-	1.67
	h	-	-	46.6	36.6	11.2	5.6	-	-	-	0.76
	i	-	-	33.6	32.6	17.4	12.0	4.4	-	-	1.21
	j	1.2	4.4	41.4	34.8	10.8	1.2	-	-	6.2	0.81

TABLE B

Method	Case			
	g	h	i	j
1	36.5	97.2	90.8	69.4
2	100	99.2	93.8	65.9
3	2.5	71.8	61.4	90.3
7	0.8	47.6	34.5	57.1
8	-	100	100	100

* Tables A and B are based on 90 and 50 iterations respectively.

FIGURE 7.1: Prediction Efficiencies for case (a)

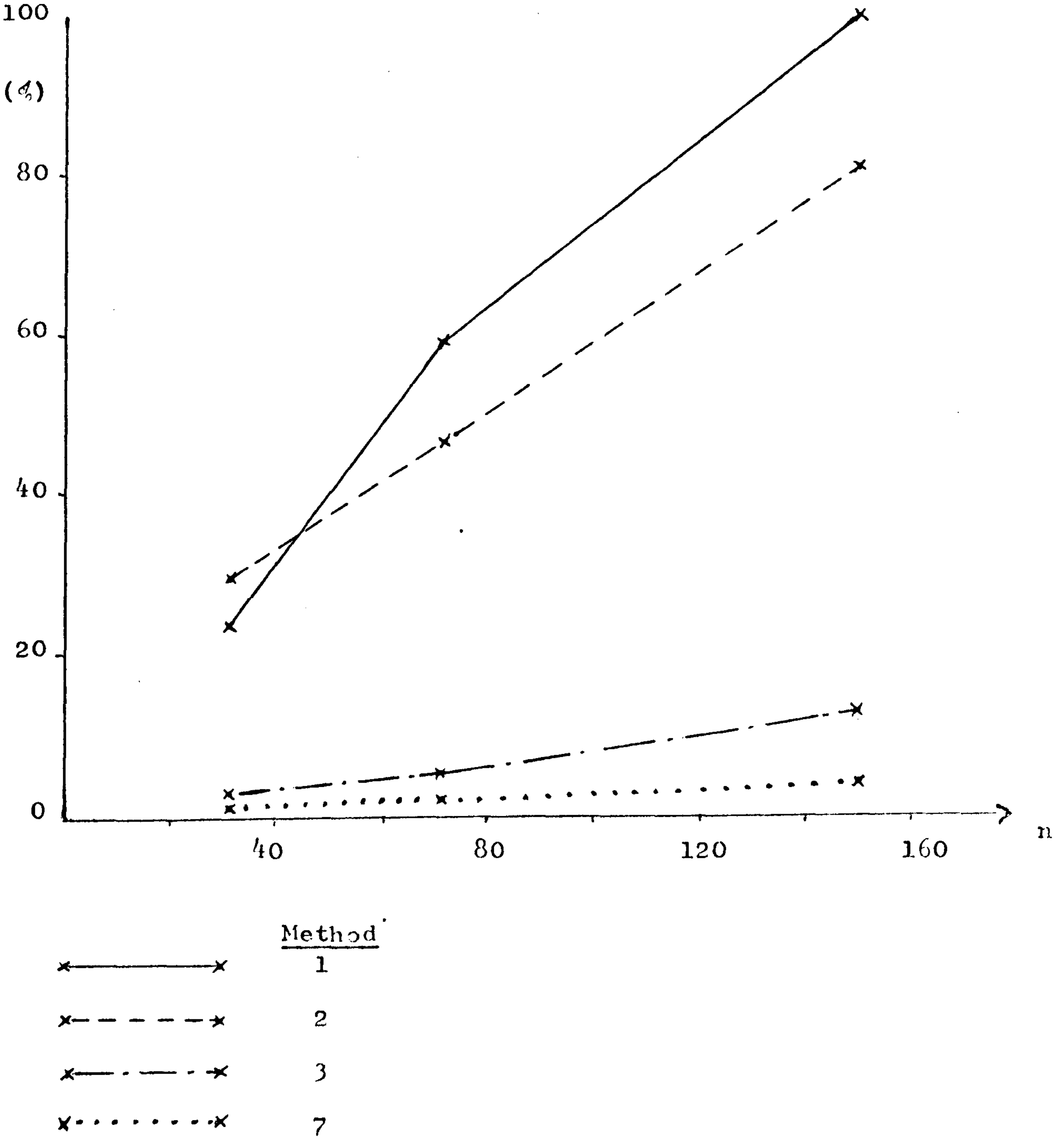


FIGURE 7.2: Prediction Efficiencies for case (b)

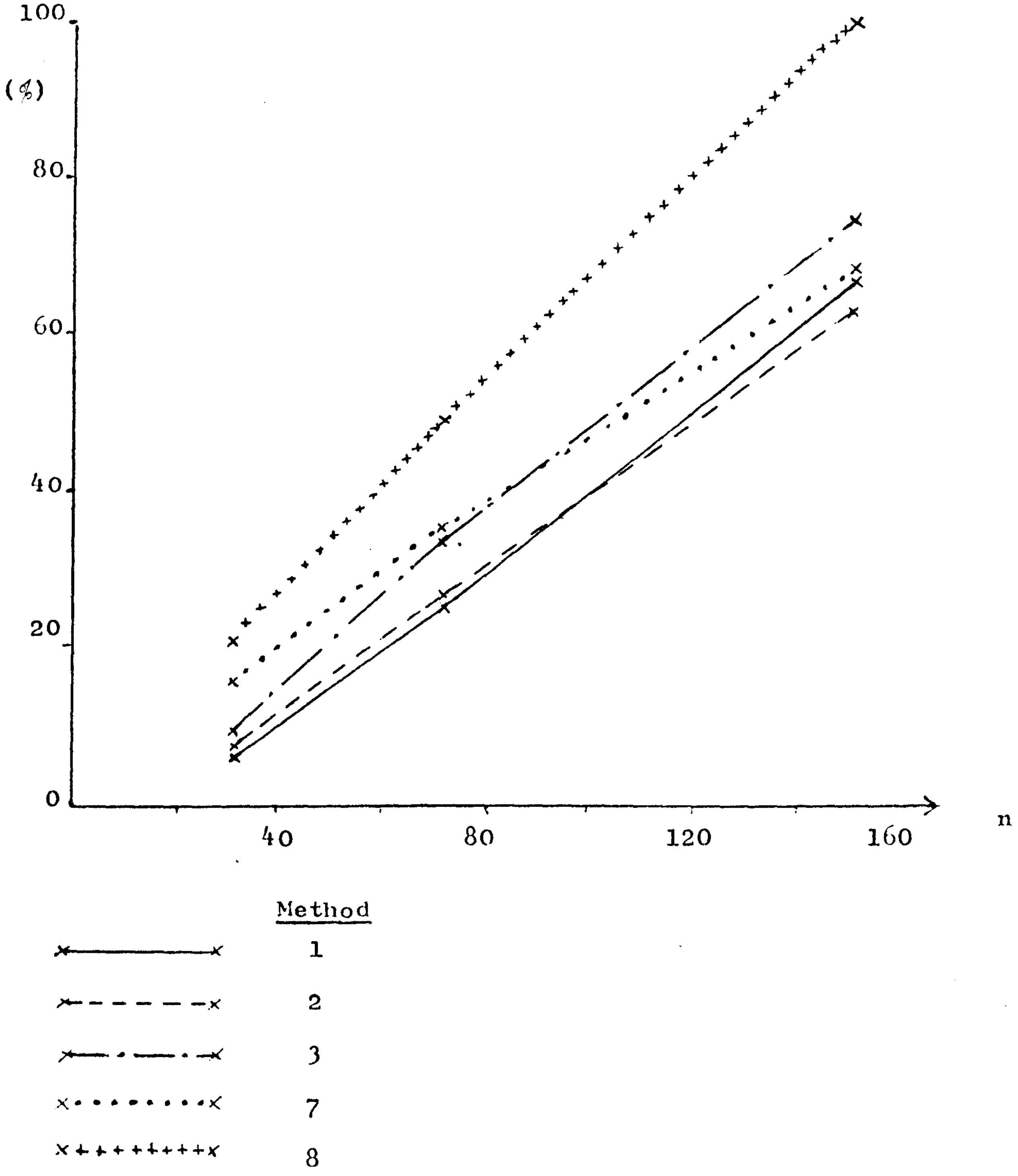


FIGURE 7.3: Prediction Efficiencies for case (c)

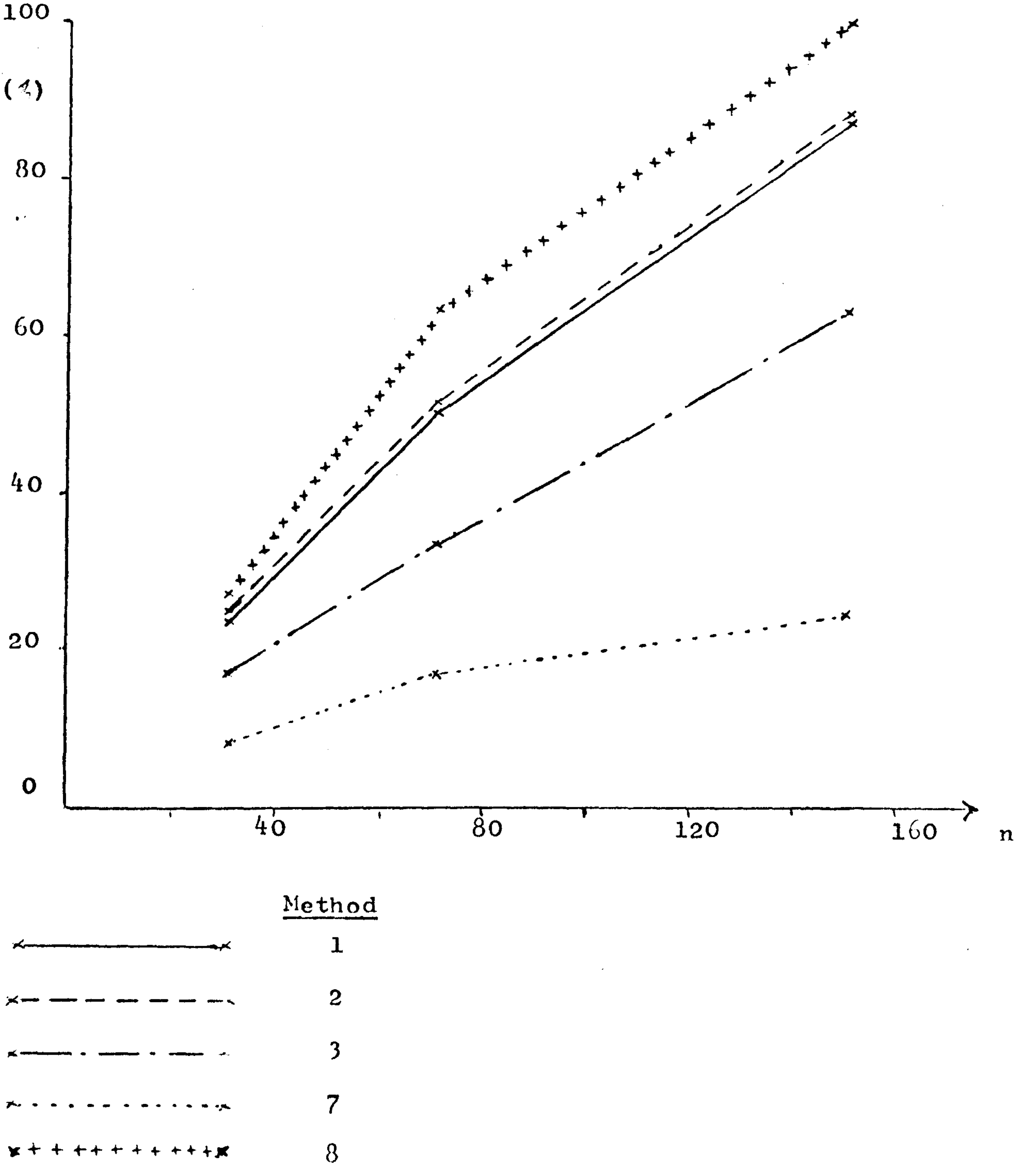


FIGURE 7.4: Prediction Efficiencies for case (d)

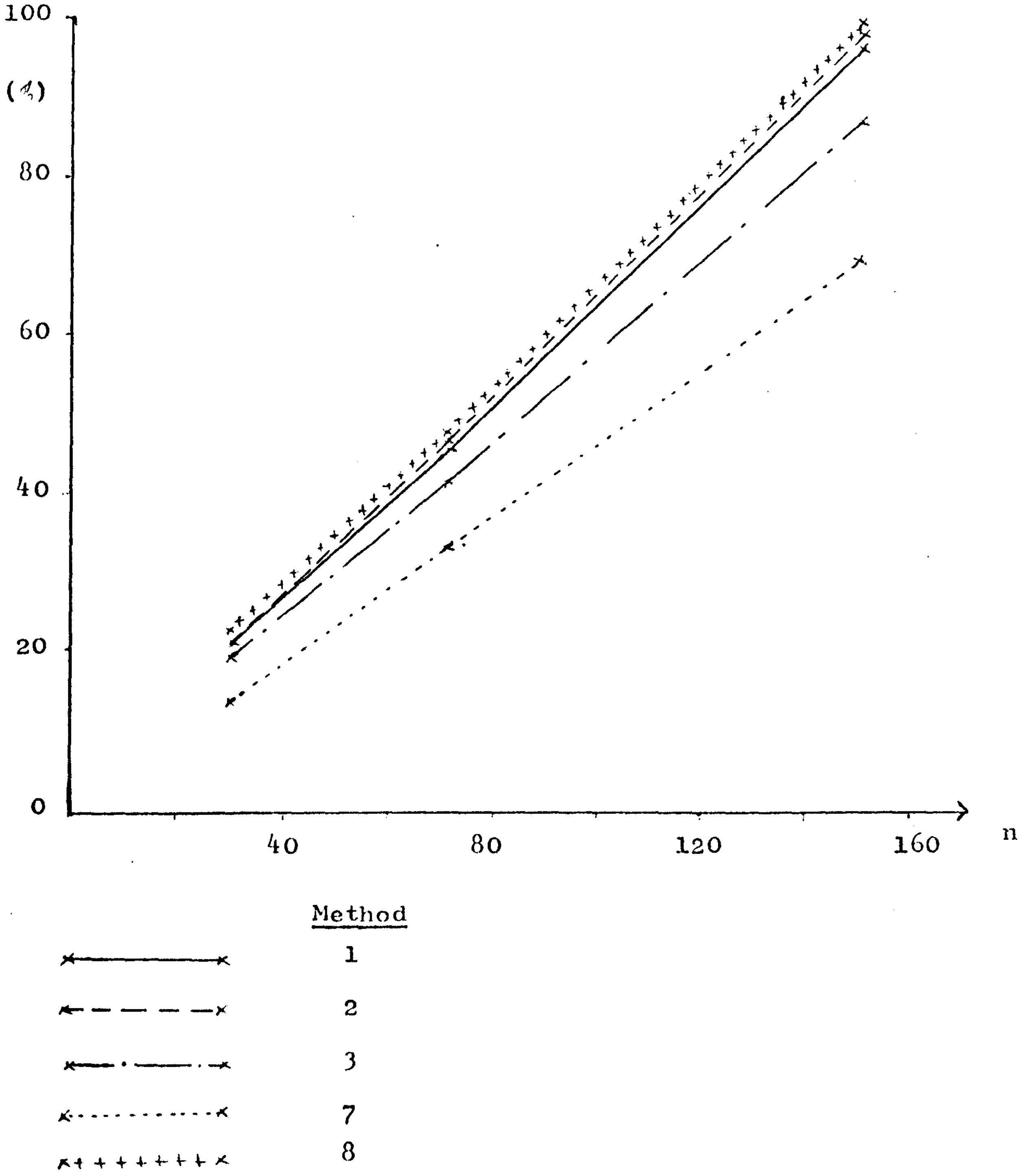


FIGURE 7.5: Prediction Efficiencies for case (e)

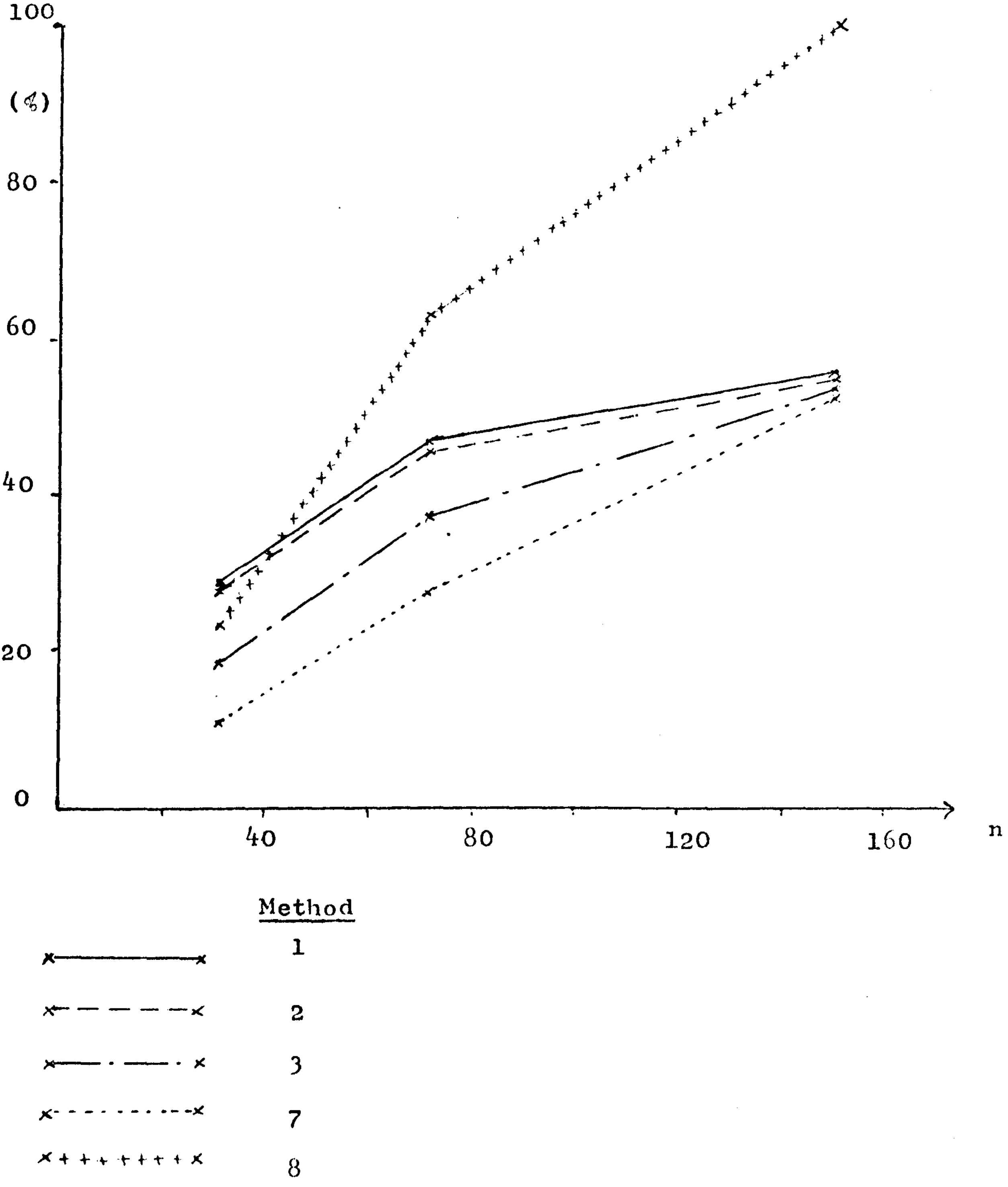
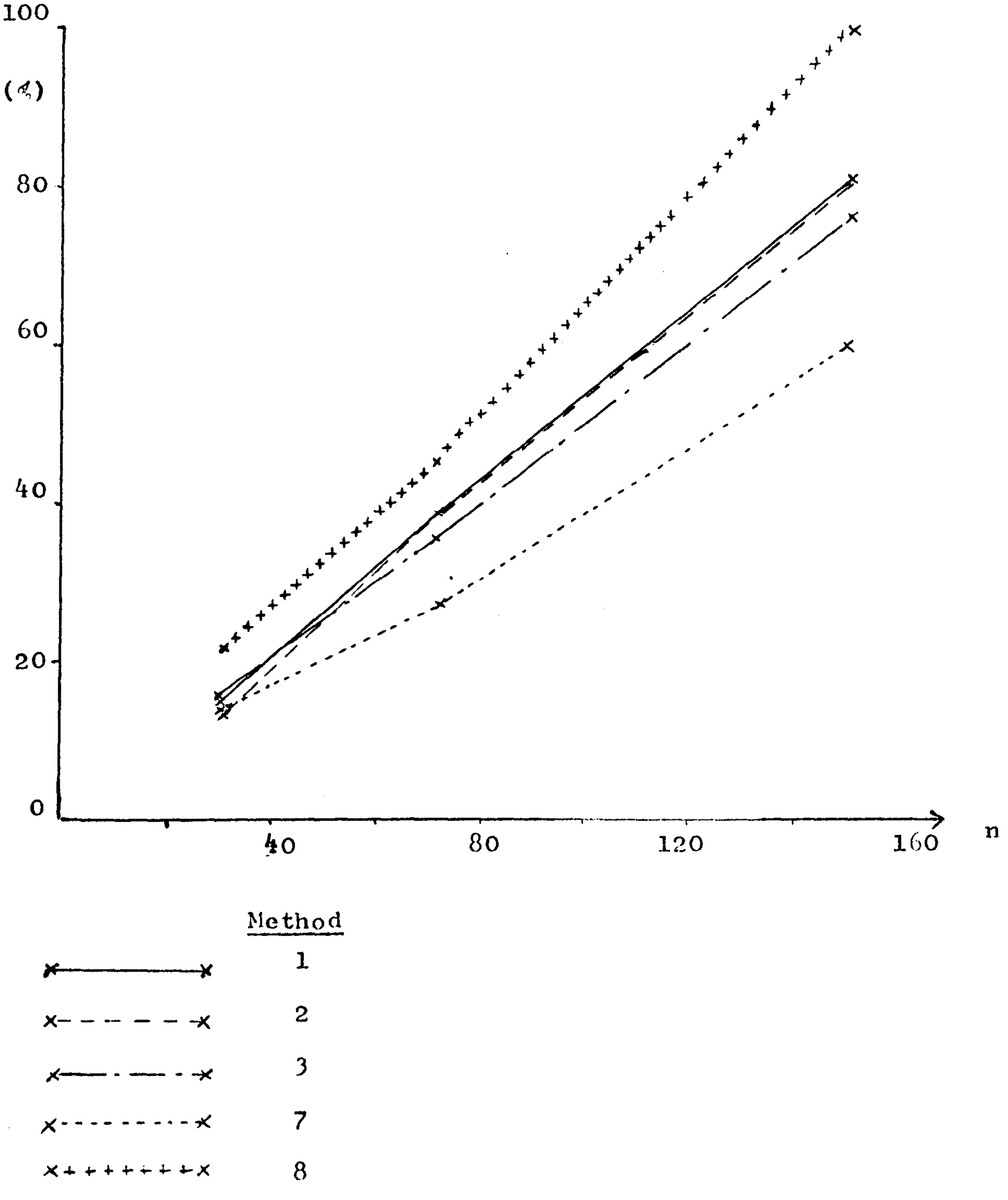


FIGURE 7.6: Prediction Efficiencies for case (f)



performances of FMAX and F'. Neither of these criteria can be said to demonstrate a definite superiority over the other. For while F' (method 2) appears marginally better in the situation of Table 7.2, for example, the similar configuration of case (j) in Table 7.7 can be said to be more favourable to the use of FMAX.

If we turn to the performance of the procedures using conventional F one can only really find support again in the two similar situations of Table 7.2 and case (j) of Table 7.7. In such instances the seemingly inherent tendency to overfit of such procedures appears to compensate for the small magnitudes of the λ_j parameters, this same characteristic causing the FMAX and F' criteria to have reduced power. The situation of case (j) does however also indicate that even in such instances this compensation can ultimately (as n gets large) be undesirable. What is perhaps one of the most dangerous aspects of the conventional F approach is revealed very clearly in Table 7.1 and case (g) of Table 7.7. For in such circumstances one would often be led to conclude that significant regressors do exist whereas in fact the contrary is true. This characteristic of conventional F will be encountered again later when time series situations are investigated. One consequence there will be that an autocorrelation structure is likely to be deduced for what is in fact a completely random process.

While it is obviously impossible to arrive at completely hard and fast conclusions on the basis of what are only exploratory investigations, methods 1 and 2 do certainly demonstrate a desirable asymptotic property. If, in a particular application, one also has a priori knowledge as to the number of regressors possessing non-zero coefficients there is then scope for improving procedure efficiency in the general (finite sample) case. In a situation like that of Table 7.2, for example, this might be achieved by automatically fitting the first few variables without invoking test criteria. Alternatively the same effect could be achieved by using a higher level of significance at each testing stage. Much will depend of course on the kinds of situation one expects to meet in the practical field of study.

The above discussion has related to the problem of identification of regression models and, as was mentioned before, one is faced with difficult problems concerning the specification of an appropriate loss structure. When we turn to the prediction aspect procedure evaluation is considerably simplified. If we look firstly at the performance of method 8 we find that knowledge of the true model specification does, in general, lead to optimal predictive equations. The only exception to this is when n = 31 in Table 7.5, and this is precisely the situation where theory leads us to expect such an occurrence. It is interesting to note that, although the best procedures in this circumstance (methods 1 and 2) are associated with an underfitting tendency as expected, the procedure based on conventional F does not do as well as method 8. The explanation would seem to be that while conventional F does tend to underfit, in so far as it omits true regressors, it does also give rise to a large proportion of mixed cases.

Looking now at the performance of method 7 we find, as might be expected, that the use of all available potential regressors is certainly not desirable. Such a procedure does in fact only avoid being worst in the situation of Table 7.2, and even then its performance can be seen to be declining relative to the other methods as the sample size increases. In what is perhaps a more realistic version of this same situation we see in Table 7.6 (or from Fig.7.6) that method 7 becomes firmly entrenched at the bottom of the overall ranking.

There remains the problem of evaluating the three variants of the stepwise approach. If we glance at Figs.7.1 to 7.6 we can see that our conclusions have to be essentially as they were for the identification case, except that there is slightly more evidence now in favour of F' and FMAX. Again it is hard to detect any real difference in performance between the use of FMAX and F' (the striking result in case (g) of Table 7.7 being accounted for as perhaps not an unexpected value of the ratio of two very small quantities). Although faring well in situations in which there is a large proportion of regressors with small associated non-centrality parameters

$$\lambda = \beta^2 \ \Sigma \ X^2/\sigma^2$$

this advantage is eroded as we either increase the regressor variances or allow the sample size n to increase.

In overall conclusion it seems that, on asymptotic grounds at least, the use of either FMAX or F' as test criteria is to be preferred to that of conventional F both from an identification and prediction viewpoint. While one cannot be so definite in the finite (small) sample situation, there is still much evidence in support of the former two procedures. Further to this, one always has the knowledge that at worst the final selected equation will most likely be an underfitted one unlike in the use of conventional F.

Having investigated the orthogonal regression situation in some detail in chapters 5, 6 and the present one we go on in the next chapter to broaden our discussion to the non-orthogonal situation.

Chapter 8 Stepwise Regression with Non-Orthogonal Regressors

8.1. Special features of the non-orthogonal case

The discussion so far has mainly related to situations in which orthogonal regressors are available, and this has led to considerable simplification in our argument. This has not however reflected the area of application of stepwise regression in practice. For indeed the very existence of mixed forward/backward routines must suggest its intended use in the more general non-orthogonal case. Particularly noticeable in this direction has been the recent use of the stepwise approach in a number of econometric studies, and in the associated field of time series analysis. We therefore now take a look at the extra difficulties which arise in this more general situation of non-orthogonality. Specifically, in this chapter we shall be generalising some of the points made and proposals suggested in the previous chapters, 3, 5 and 6. Initially the discussion will relate mainly to the objective of identification i.e. the determination of which regressors occur in the underlying model with non-zero coefficients. The concluding section will however deal with the aspect of prediction.

We will continue to assume that the regressor can be regarded as a set of fixed constants. This means that the circumstances are either such as to permit replication of these values, or at least that they are stochastically independent of the errors in the underlying model. The main point of departure is that the matrix $\underset{\sim}{X}'\underset{\sim}{X}$ is now no longer assumed to be strictly diagonal. The first point to emphasize is that, in general, different equation sequences will now be produced by forward and backward procedures. One can in fact construct examples in which the first variable to enter in a forward procedure is also the first to be deleted in a backward approach. This stems from the fact that, unlike in the orthogonal case, the contributions due to particular regressors depend very much on which other variables have already been entered. Hence one can, for a fixed order of equation r, obtain two vastly different sets of regressors and two different residual sums of squares values. Further, neither of these residual sums of squares need be smallest possible amongst all fitted equations involving r regressors. This contrasts sharply with the orthogonal case previously considered and is a feature to be taken into account when comparing single-step procedures with 'all equation' procedures of the kind described in chapter 2. It is nevertheless still constructive to examine the source of these limitations and ambiguities within a theoretical framework if only so that stepwise regression results are regarded with appropriate caution in practice.

8.2 Comparison of forward and backward approaches

The choice between using a forward or backward approach now has an added dimension in the non-orthogonal case.

For whilst previously in the orthogonal case interest could be centered on the underfitting/ overfitting characteristics of procedures, we now have to face the very real possibility of spurious relationships leading to what can be described as severe mixed cases. This can occur when a particular regressor, although not present in the true model, has a high correlation with a set of regressors which are. It is then quite possible, in a forward approach at least, for such a regressor to be fitted at the complete exclusion of the set concerned. This was in fact an argument put forward by Mantel [57] in favour of the use of backward approaches. The same kind of argument was however also employed by Beale [9] in support of the forward approach. Beale argues that a regressor which could considerably decrease the residual sum of squares if added to the final equation selected might already have been irretrievably lost due to a nonsense correlation with variables which are later eliminated. While this kind of occurence can indeed be demonstrated to be plausible on theoretical grounds (though being less likely to occur as n increases) it is precisely this eventuality which mixed forward/backward procedures are designed to overcome.

The above points can be more properly demonstrated using the theoretical basis developed in chapter 3. In particular we can, as in the orthogonal vase, consider the expected order of entry or deletion of variables to the fitted equation. Again, at the stage where r variables have already been fitted, the decision to enter or delete a variable will still be based on the k - r extra sum of squares quantities

$$S_j^2 = \sigma^2 (d_i^*)^2$$

where $d_i^* = d_i/\sigma_{d_i}$ as at (3.1.12).

Continuing to use the notation of chapter 3, we first consider the expected values of these k - r random variables in a marginal sense (i.e. disregarding for the moment the fact that a stepwise procedure induces a conditional distribution at each step.). Thus, recalling from (3.1.12) that $\underset{\sim}{d}^* = \sigma^{-1}\underset{\sim}{F}^{\frac{1}{2}}\underset{\sim}{d}$, it follows that we can focus attention on the diagonal elements of the matrix

$$\underset{\sim}{F}^{\frac{1}{2}}\underset{\sim}{V}_d\underset{\sim}{F}^{\frac{1}{2}} + \underset{\sim}{F}^{\frac{1}{2}}\underset{\sim}{\mu}_d\underset{\sim}{\mu}_d'\underset{\sim}{F}^{\frac{1}{2}} \qquad (1)$$

On substituting for $\underset{\sim}{\mu}_d$ and $\underset{\sim}{V}_d$ in terms of the expressions given at (3.1.6) and (3.1.7) we can write (1) as

$$\sigma^2\underset{\sim}{F}^{-\frac{1}{2}}\underset{\sim}{Z}_2'\underset{\sim}{M}\underset{\sim}{Z}_2\underset{\sim}{F}^{-\frac{1}{2}} + (\underset{\sim}{F}^{-\frac{1}{2}}\underset{\sim}{Z}_2'\underset{\sim}{M}\underset{\sim}{X}_1\underset{\sim}{\beta}_1(\underset{\sim}{F}^{-\frac{1}{2}}\underset{\sim}{Z}_2'\underset{\sim}{M}\underset{\sim}{X}_1\underset{\sim}{\beta}_1)' \qquad (2)$$

The first term of this expression is immediately seen to have all its diagonal elements equal to σ^2. This follows from the definition of $\underset{\sim}{F}$ given at (3.1.4) and (3.1.5). The diagonal terms of the second matrix in (2) are seen to be just the squares of the elements of the vector $\underset{\sim}{F}^{-\frac{1}{2}}\underset{\sim}{Z}_2'\underset{\sim}{M}\underset{\sim}{X}\underset{\sim}{\beta}_1$, the magnitudes of which were investigated in some detail in the latter part of section 1 of chapter 3. In the light of the results obtained there some comments of a general nature can be made concerning the expected behaviour of various procedure types.

Firstly, with regard to purely forward procedures, (3.1.11) indicates that we can no longer necessarily expect variables with non-zero true coefficients to be entered before those whose true coefficients are zero. This is essentially the "spurious correlation"aspect referred to earlier. However (3.1.10) does suggest we can expect, subject to the sensitivity of the test criterion used, that procedures will continue to enter variables until all the correct ones have been included. When such a stage has been arrived at the fact that $\underset{\sim}{\mu}_d = \underset{\sim}{0}$ should then have the effect of terminating the procedure quite quickly. We note that the facility of allowing variable deletions to take place is still desirable in order to eliminate variables whose earlier spurious significance has ultimately disappeared. Although this could admittedly sometimes cause true regressors to be deleted it can be argued that on balance the overall effect of allowing deletions to occur should be advantageous.

If we turn to consideration of a strict backward approach we can now expect procedures to begin by deleting the truly redundant ones. The extent to which true regressors are then also deleted is very dependent on the associated test sensitivity. Again a substantial case can be put forward for invoking a mixed procedure approach by allowing, in this instance, a variable entry facility to be present.

Although the details will not be given here we can, as in the orthogonal case at (5.2), strengthen the above arguments for large sample sizes with a procedure consistency property. The essential point in the argument is that, as we increase the sample size n, the corresponding moment matrices $\underset{\sim}{X}'\underset{\sim}{X}$ must bear a scalar proportionality to each other. This ensures that we preserve the expected equation sequences which are produced irrespective of the value of n.

Finally, although we have not yet investigated possible test criteria for use at each step of a procedure in the non-orthogonal case, we might briefly contemplate the possibility of making statements similar to those in (5.4) concerning control on under- or overfitting. As far as a strict forward procedure is concerned we can obviously no longer expect protection against overfitting as is implied by

(5.4.5) and (5.4.7) in the orthogonal case. Indeed the best that we can say for any reasonable procedure will be that there will be a zero asymptotic probability that underfitting occurs. This is again just a consequence of the consistency property referred to above.

In the following two sections we consider the possibility of again using the test criteria of F' and FMAX respectively. The use of a simultaneous inferential approach along the lines of F* will not be entertained. For such an approach cannot, by its very nature, utilize the extra sensitivty obtained by making decisions conditionally on the outcomes of previously tested hypotheses. Thus, henceforth we concentrate exclusively on the class of stepwise decision procedures.

8.3 The use of F'

It was stated earlier in (5.4) that the statistic

$$F' = \left(\frac{n-r-1}{k-r}\right)\frac{v_r}{v_k} - \frac{n-k-1}{k-r}$$

is distributed as F(k-r,n-k-1) under the hypothesis that the included variables are the true ones and the excluded variables are the unwanted ones. That this result continues to hold in the non-orthogonal situation is a standard result of regression theory.

However we can demonstrate the validity of this in a way which also throws light on the connection between F' and the composite hypothesis $H_o \equiv \underset{\sim}{\mu}_d = \underset{\sim}{0}$ which was discussed in (3.3).

We take as our starting point the result that, if $\underset{\sim}{d}$ is distributed as $N(\underset{\sim}{\mu},\underset{\sim}{V})$, then

$$Q = (\underset{\sim}{d}-\underset{\sim}{\mu})'\underset{\sim}{V}^{-1}(\underset{\sim}{d}-\underset{\sim}{\mu})$$

is distributed as chi-square with k - r degrees of freedom (where k-r is the number of excluded variables at the r^{th} stage). We will suppose that the p "true" regressors are included in the r that have already been entered i.e. that in the notation of chapter 3 $\underset{\sim}{Z}_2$ consists only of unwanted variables. Since $\underset{\sim}{\mu} = \underset{\sim}{0}$, we have

$$Q = \underset{\sim}{d}'\underset{\sim}{V}^{-1}\underset{\sim}{d} = \underset{\sim}{Y}'\underset{\sim}{M}Z_2\underset{\sim}{F}\{\sigma^2\underset{\sim}{F}Z_2'\underset{\sim}{M}Z_2\underset{\sim}{F}\}^{-1}\underset{\sim}{FZ_2'MY}$$

where $\underset{\sim}{M} = \underset{\sim}{I} - \underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'$

Since Q involves the unknown value of σ^2 it is necessary to use the independent estimate given by v_k. As v_k in turn is distributed as $\sigma^2\ (n-k-1)^{-1}$ times a chi-square variable with n - k - 1 degrees of freedom it follows that

$$Q' = \frac{Q\sigma^2}{(k-r)v_k}$$

is distributed as F[k-r, n-k-1]. Moreover,

$$\sigma^2 Q = \underset{\sim}{Y}'\underset{\sim}{MZ_2}\underset{\sim}{F}(F\underset{\sim}{Z}_2'\underset{\sim}{MZ_2F})^{-1}\underset{\sim}{FZ_2'MY} \qquad (1)$$

and hence Q' can be calculated from the sample data. It now remains to show that Q' = F'.

To see this we write

$$F' = \frac{(n-r-1)v_r - (n-k-1)v_k}{(k-r)v_k}$$

Further, from chapter 3, we have

$$(n-r-1)v_r - (n-k-1)v_k = \underset{\sim}{\varepsilon}'[\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}' - \underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'$$

Using the partitioned form of $\underset{\sim}{X} = [\underset{\sim}{Z}_1 \;\; \underset{\sim}{Z}_2]$ we can write this as

$$\underset{\sim}{\varepsilon}'[\underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1' + \underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'\underset{\sim}{Z}_2\underset{\sim}{E}^{-1}\underset{\sim}{Z}_2'\underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'$$

$$- \underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'\underset{\sim}{Z}_2\underset{\sim}{E}^{-1}\underset{\sim}{Z}_2' - \underset{\sim}{Z}_2\underset{\sim}{E}^{-1}\underset{\sim}{Z}_2'\underset{\sim}{Z}_1(\underset{\sim}{Z}_1\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'$$

$$+ \underset{\sim}{Z}_2\underset{\sim}{E}^{-1}\underset{\sim}{Z}_2'\underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1']\underset{\sim}{\varepsilon}$$

where $\underset{\sim}{E} = \underset{\sim}{Z}_2'\ [\underset{\sim}{I} - \underset{\sim}{Z}_1(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1']\underset{\sim}{Z}_2$

This simplifies to

$$\underset{\sim}{\varepsilon}'\underset{\sim}{M}\underset{\sim}{Z}_2(\underset{\sim}{Z}_2'\underset{\sim}{M}\underset{\sim}{X}_2)^{-1}\underset{\sim}{Z}_2'\underset{\sim}{M}\underset{\sim}{\varepsilon}$$

$$= \underset{\sim}{\varepsilon}'\underset{\sim}{M}\underset{\sim}{Z}_2\underset{\sim}{F}(\underset{\sim}{F}\underset{\sim}{Z}_2'\underset{\sim}{M}\underset{\sim}{Z}_2\underset{\sim}{F})^{-1}\underset{\sim}{F}\underset{\sim}{Z}_2'\underset{\sim}{M}\underset{\sim}{\varepsilon} \qquad (2)$$

The required result is obtained by putting $\underset{\sim}{Y} = \underset{\sim}{Z}_1\underset{\sim}{\beta}_1^* + \underset{\sim}{\varepsilon}$ in (1) (where $\underset{\sim}{\beta}_1^* = \begin{bmatrix}\underset{\sim}{\beta}_1 \\ \underset{\sim}{0}\end{bmatrix}$), and showing equality with (2).

We may again, as in the orthogonal case at (5.5.1), observe that the use of F' is equivalent to the test of a maximal quantity. Again, following Scheffé [72], we are essentially testing the hypothesis

$$H_o \equiv \underset{\hat{c}}{\text{Max}} \sum_{i=1}^{q} \emptyset_i E[d_i^*] = 0$$

where in this non-orthogonal situation Φ is the space of vectors $\phi = [\phi_1 \ldots \phi_q]'$ subject to the normalizing constraint

$$\underset{\sim}{\phi}' \, \underset{\sim}{\Omega} \, \underset{\sim}{\phi} = C \; ,$$

and where C is arbitrary (but fixed) constant.

In a corresponding manner the acceptance region for F', which in the orthogonal case was spherical, now becomes ellipsoidal.

8.4 The use of FMAX

Some remarks have already been made in section 3 of Chapter 3 concerning the distribution of FMAX in the general case. It was stated there that the relevant distribution theory is that of certain multivariate generalizations of the chi-square, t and F distributions. The study of such distributions might be said to have begun with a paper by Krishnamoorthy and Parthasarathy [46] who looked at what is generally referred to as the multivariate gamma or multivariate chi-square distribution. The purpose of their paper was to obtain an expression for the joint density function of what are essentially the diagonal elements of a Wishart distributed matrix. This they managed to achieve in terms of an infinite series of Laguerre polynomials, the validity of such a representation depending on certain convergence conditions on the correlation matrix of the underlying associated normal distribution.

Of more direct relevance to the problem at hand are studies concerned with the actual evaluation of probability integrals of Studentised versions of such multivariate distributions. Since the expression derived by Krishmamoorthy and Parthasarathy is not sufficiently tractable to permit such computations other approaches have been suggested. A starting point in this direction consists of two papers by Dunnett and Sobel [21 , 22] in which they defined what is referred to as the multivariate t distribution. This distribution is in fact obtained by Studentising the multivariate chi-square variates referred to above, thus leading to a distribution having probability density function given by :

$$f(t_1, \ldots t_q) = \frac{|\Omega|^{-\frac{1}{2}}\Gamma[\frac{1}{2}(\nu_o+q)]}{(\nu_o\pi)^{q/2}\,\Gamma\left[\frac{\nu_o}{2}\right]} \left[1 + \frac{1}{\nu_o} \sum_{i=1}^{q} \sum_{j=1}^{q} \omega^*_{ij} t_i t_j\right]^{\frac{1}{2}(\nu_o+q)} \tag{1}$$

where $\underset{\sim}{\Omega}$ is the correlation matrix of the underlying multivariate normal distribution (and corresponds to $\underset{\sim}{\Omega}$ as defined at (3.1.13)), ω^*_{ij} is the (i, j) - element of $\underset{\sim}{\Omega}^{-1}$ and ν_o are the degrees of freedom of the denominator estimator of σ^2.

Dunnett and Sobel then proceed to look at the evaluation of the general probability integral:

Prob$[t_i \leq h_i;\ i = 1, \ldots, q]$

$$\int_{-\infty}^{h_1} \int_{-\infty}^{h_2} \cdots \int_{-\infty}^{h_q} f(t_1, \ldots, t_q)\ dt_1, \ldots,\ dt_q \quad \ldots (2)$$

While such evaluations are possible (but by no means straightforward) in the case where q = 2, severe difficulties are faced in the general case when q > 2. However, in the second of their two papers referred to above Dunnett and Sobel suggested a transformation whereby, in the special case in which $\omega_{ij} = c_i c_j$ for $i, j = 1, \ldots, q$ (and $0 \leq c_i > 1$), the problem can be converted to one involving q + 1 independent standard normal variables. This transformation has subsequently been exploited in several other investigations into related problems.

It might be noted at this point that the problem as far as stepwise regression is concerned is the determination of h such that, for given α,

$$\int_{-h}^{h} \int_{-h}^{h} \cdots \int_{-h}^{h} f(t_1, \ldots, t_q)\ dt_1, \ldots, dt_q = 1 - \alpha \quad ..(3)^{\dagger}$$

† We are here supposing that equi-co-ordinate probability points are the right ones. While this was probably the correct approach in the orthogonal case the situation is now less clear. For, given a specified alternative hypothesis, it is very likely that a more general acceptance region should be determined from

$$\int_{-h_1}^{h_1} \int_{-h_2}^{h_2} \cdots \int_{-h_q}^{h_q} f(t_1, \ldots, t_q) dt_1, \ldots, dt_q = 1 - \alpha$$

where the h_i, $i = 1, \ldots q$, are now dependent on $\underset{\sim}{\Omega}$. Problems of specification of appropriate hypotheses will however preclude further investigation of this point in this study.

Evaluation of the integral on the left-hand side of (3) is seen in fact to correspond to finding

$$\text{Prob} \left[\max_{i=1,\ldots,q} |t_i| \leq h \right]$$

or equivalently,

$$\text{Prob} \left[\max_{i=1,\ldots,q} t_i^2 \leq h^2 \right]$$

While it is true (e.g. see Krishnaiah [43]) that evaluation of the integral at (3) depends only on the absolute values of the elements of $\underset{\sim}{\Omega}$, one is still faced with the same basic difficulties which occur in the more general case given at (2). Thus, as far as is known, the most extensive tabulations of integrals which are relevant to our proposed use of FMAX are those given by Krishnaiah [43,44,45]†

The first of these sets of tables (i.e. Krishnaiah [43]) presents the values of h^2 in (3) corresponding to $\nu_o = 5\ (1).\ 35$, $q = 1(1)10$ and $\alpha = 0.10, 0.05, 0.025$ and 0.01. An equi-correlation structure is assumed throughout with $\omega_{ij} = \rho$ taking the values 0.05 (0.05) 0.9. In Krishnaiah [45] the roles of h and α are interchanged and q taken to be 2, thus giving values of α corresponding to $h = 1.0(0.1)5.5$ and the same grid of values for ν_o and ρ

† Krishnaiah has produced a number of (joint) reports which comprise mainly tabulations of various multivariate probability integrals useful in simultaneous inference applications. His motivation seems to stem from Krishnaiah [41] in which the main interest is in post analysis of variance tests in a particular experimental design set-up. These reports contain many references to the published literature on the evaluation of multivariate probability integrals of the type considered above.

as above.Finally, in Krishnaiah [44], tables similar to the two described above are presented for the case when $\nu_0 = \infty$ i.e. in the case where the multivariate F distribution is equivalent to the multivariate chi-square.

Having examined the situation regarding the distribution of FMAX it must be remarked that the prospects of its implementation within stepwise procedures in the general non-orthogonal case are pretty daunting. For it is extremely implausible that the correlation matrices encountered in practice will be anything like the equi-correlated versions for which tables of FMAX are available.Nor is it at all feasible to contemplate probability evaluations corresponding to the actual correlation structures actually obtained. At best only some kind of approximation might be attempted. One possibility is to take the average of the q(q-1)/2 different values of ω_{ij} and to use this as if it were a common correlation coefficient. However, there seems to be no theoretical justification at all for doing this and its implementation would be rather cumbersome.

A technique which is often of use in similar situations in which there is a dependent structure is that of the Bonferroni Inequalities. If we let A_i denote the event that $|t_i| > h$ then we have, using Boole's fundamental equality,

$$\text{Prob}[\bigcup_{i=1,q} A_i] = \sum_{i=1}^{q} \text{Prob}[A_i] - \sum_{i<j}^{q}\sum^{q} \text{Prob}\ [A_i \cap A_j] + \\ \ldots + (-1)^{q-1}\ \text{Prob}\ [\bigcap_{i=1,\ldots,q} A_i] \tag{4}$$

It follows from (4) that

$$\text{Prob}[\max_{l=1,\ldots q} |t_i| \leq h] = 1 - \text{Prob}[\bigcup_{i=1,\ldots,q} A_i]$$

$$= 1 - \sum_{i=1}^{q} \text{Prob}\ [A_i] + \sum_{i<j}^{q}\sum^{q} \text{Prob}[A_i \cap A_j] \ldots + (-1)^q \text{Prob} \\ [\bigcap_{i=1,\ldots q} A_i] \tag{5}$$

The partial sums obtained by including successively more of the terms on the right-hand side of (5) give successively sharper upper and lower bounds to the required probability and are known as the Bonferroni Inequalities. One such inequality is immediately obtainable using the fact that

$$\sum_{i=1}^{q} \text{Prob}[A_i] = q\ P_1$$

where P_1 is the marginal probability that a t - variate with ν_o degrees of freedom falls outside the interval [-h,h]. One can then proceed to obtain an <u>upper bound</u> to the required probability on the left-hand side of (5) by again invoking Boole's Eqality on each of the terms $\text{Prob}[A_i \cap A_j]$, i.e.,

$$\text{Prob}[A_i \cap A_j] = 2\ P_1 - \text{Prob}\ [A_i \cup A_j]$$

$$= 2\ P_1 - 1 + P_{2\,(i,j)}$$

where $P_{2\,(i,j)} = \text{Prob}\ [\ \overline{A_i \ \cup A_j}\,]$

$$= \text{Prob}\ [|t_i|, |t_j| \le h; \omega_{ij}].$$

The $q(q-1)/2$ such values of $P_{2\,(i,j)}$ which are required can then be obtained from the tables of Krishnaiah for each of the values of ω_{ij}. Thus, from (5), we have the inequality :

$$\text{Prob}[\max_{i=1,\dots,q} |t_i| \le h] \le 1\ q(q-1)/2 + q(q-2)P_1 + \sum_{i<j}^{q} \sum^{q} P_{2\,(i,j)} \qquad (6)$$

A way in which the above procedure could be implemented in a stepwise regression context would be to find the particular upper bound associated with the <u>observed</u> maximum value of $|t_i|$, $i = 1,\dots,q$. Using a desired rejection level of α this then leads to an actual rejection level which is greater than α. The subsequent effect on procedure performance will be a tendency to introduce further variables into the regression equation than is consistent with the desired significance level α.

An attempt was made to gauge the accuracy of the upper bound at (6) by making comparisons in a situation in which the correct probabilities are known. Such a situation is that of the equi-correlated case where, in order also to allow comparisons to be made with the effect of ignoring the correlation structure altogether, the particular case was chosen in which the denominator degrees of freedom are infinite. Using Krishnaiah's tables [44] for the distribution of the maximum of correlated chi-square variates the true critical values h on the left-hand side of (6) were found for $\rho = 0.0(0.1)0.8$, $q = 10$ and for a fixed probability of 0.95. Then, for each such value of h, the upper bound for the probability was found as on the right-hand side of (6), the resulting values being given in column 2 of Table 8.1. Finally, in column 3 of the same table are recorded the probabilities corresponding to each value of h but this time supposing the correlations $\omega_{ij}(i,j=1,\ldots,10)$ are all zero.

Table 8.1

(Evaluation of accuracy of Bonferroni Inequality approximation)

ρ	Bonferroni Upper Bound	Probability taking $\underset{\sim}{\Omega} = \underset{\sim}{I}$
0.0	0.95000	0.95000
0.1	0.95000	0.94971
0.2	0.95008	0.94836
0.3	0.95133	0.94628
0.4	0.95415	0.94188
0.5	0.95721	0.93588
0.6	0.97241	0.92678
0.7	1.00069	0.91249
0.8	1.06673	0.88859

Table 8.1 cannot be said to be very encouraging towards the use of the Bonferroni bound. Although there is still the possibility that the approximation might be more worthwhile in the more general case where correlations are

not all equal the indications are that one might equally as well suppose the correlations are all zero. The main point is of course as to whether it is thought worthwhile to go to a great deal of trouble in incorporating extensive tables in a stepwise routine in order to achieve what promises to be only a very crude probability bound. The main purpose of the above exercise has been to demonstrate in fact that there are now overwhelming advantages in favour of the use of the F' approach of the previous section. Further, the experiences in the orthogonal case (which should be no less favourable to FMAX) seem to indicate that even precisely evaluated probabilities for FMAX would be an unnecessary luxury. Thus our concentration will henceforth be mainly focussed on procedures using the F' criterion.

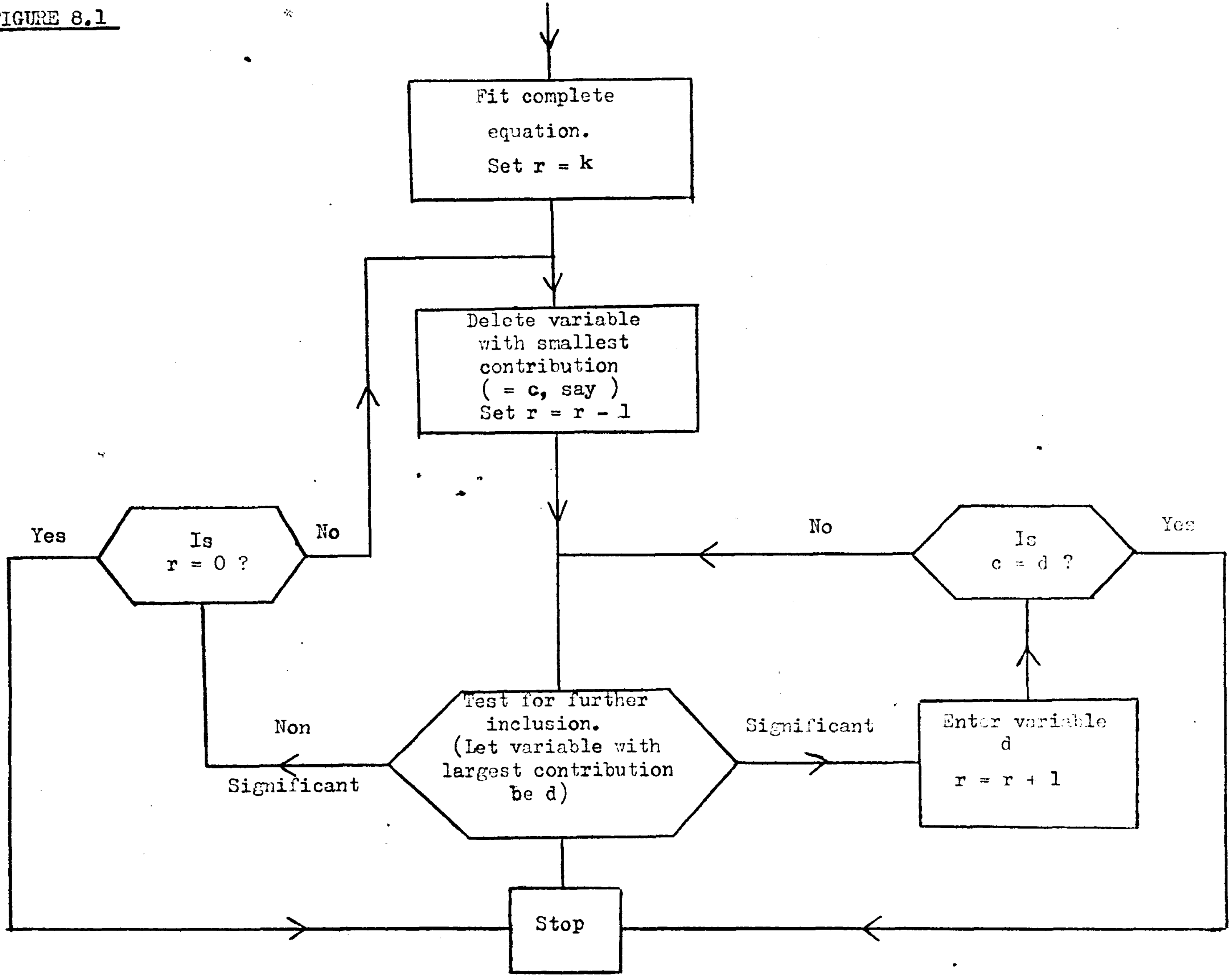
8.5 A general procedure using F'

It has been seen that a characteristic of the general non-orthogonal situation is that a variable which contributes the minimum to the explained sum of squares, when included in an equation does not in general, when deleted, yield the maximum such quantity amongst the resulting excluded set. This will in fact be so irrespective of whether the variable concerned is a 'true' one or not. In the same way the excluded variable which gives the largest contribution will not, in general, give the smallest contribution amongst the included set when it is entered. As a consequence of this some care needs to be taken in specifying the practical implementation of the mixed forward/backward procedure now appropriate in non-orthogonal contexts.

Our proposed procedure will be 'backward orientated' in the sense that the full equation involving k variables will be fitted initially. The reason for choosing this backward bias, rather than beginning from a zero order equation, is motivated by the observed tendency of procedures to underfit in the orthogonal case. A backward approach also has the property that the variables which are delated first are more likely than not to be truly unwanted ones. This contrasts with a procedure which starts out with a forward approach where we have seen that entering variables may very well be ones which we eventually wish to delete. The procedure will be described in general terms to allow for the possible implementation of FMAX- type test criteria.

The procedure begins by testing for a variable deletion in the usual way. At each subsequent stage one then first of all tests whether introduction of an excluded variable is significant. If this is so then the variable with largest contribution is entered, and the same question is asked again. If no variable is entered in this way a search is then made of the included variables for the one giving smallest explanatory

FIGURE 8.1


Fit complete
equation.
Set r = k
Delete variable
with smallest
contribution
(= c, say)
Set r = r - 1
Yes
Is
r = 0 ?
No
No
Is
c = d ?
Yes
Non
Significant
Test for further
inclusion.
(Let variable with
largest contribution
be d)
Significant
Enter variable
d
r = r + 1
Stop

sum of squares, this variable then being automatically deleted. A test is again made for possible variable entry. If this test if not significant one continues to look for another possible deletion. If however the test is significant one enters the variable yielding the largest contribution as before except that now if this entering variable is also the one which has just been automatically deleted, the procedure then terminates. The question arises of course as to whether ultimate termination is certain. It would seem that one could possibly construct cases in which the termination exit is never encountered. However, such an eventuality could not in any case be ruled out in the use of mixed stepwise procedures based on conventional F. In view of the fact that cycling has never occurred in any of the numerous stepwise applications which have been performed in this study, both on artificial and real data, it would seem that this feature constitutes an extremely remote possibility.

To aid understanding of the mixed procedure described above a flow diagram is given in Figure 8.1. The procedure is henceforth referred to as Method 10,and is described again in Appendix 3 for ease of reference.

8.6 Prediction with non-orthogonal regressors

We conclude this chapter by generalizing some results obtained in Chapter 6 relating to the prediction objective for stepwise procedures. We continue to use the

'$\underset{\sim}{Z}$' notation whereby $\underset{\sim}{Z}_r$ denotes the n x r matrix corresponding to the r included variables and $\underset{\sim}{Z}_{k-r}$ similarly relates to the excluded set. Again we will be forced to assume that the sequentially obtained coefficient estimators behave as they would in the estimation of the corresponding completely specified model (albeit a possibility mis-specified model). The consistency property referred to earlier in section 2 does in fact at least provide an asymptotic justification for such an assumption. As before our objective is to obtain a final equation which minimizes the subsequent prediction mean square-error when applied to a future set of regressor values. Initially we relate the problem to prediction of a single future value of Y given a single determination $\underset{\sim}{x}$ of regressors (where $\underset{\sim}{x}$ is a k x 1 vector).

We begin, as at (6.1.1), with the mean square error of prediction which results from using the equation obtained at the rth stage (or more accurately the stage when r regressors occur in the equation). As before we have

$$MSE_{(r)} = \sigma^2 + Var\ [\underset{\sim}{x}'\underset{\sim}{b}^*_{(r)}] + B^2_{(r)} \qquad (1)$$

where $B_{(r)} = \underset{\sim}{x}'E\ [\underset{\sim}{b}^*_{(r)} - \underset{\sim}{\beta}]$, and where the notation follows that of (6.1). We now obtain expressions for the last two components on the right-hand side of (1).

Firstly, consider the variance component

$$Var[\underset{\sim}{x}'\underset{\sim}{b}^*_{(r)}] = Var[\underset{\sim}{b}^{*\prime}_{(r)}\underset{\sim}{x}]$$

$$= E[(\underset{\sim}{b}^*_{(r)} - E[\underset{\sim}{b}^*_{(r)}])'\ \underset{\sim}{x}\ \underset{\sim}{x}'(\underset{\sim}{b}^*_{(r)} - E[\underset{\sim}{b}^*_{(r)}])] \qquad (2)$$

Recalling that $\underset{\sim}{b}^{*}_{(r)} = \begin{bmatrix} \underset{\sim}{b}_{(r)} \\ \underset{\sim}{0} \end{bmatrix}$ and noting that,

with ourassumption that $\underset{\sim}{b}_{(r)}$ can be regarded as the ordinary least squares estimator $\underset{\sim}{b}_{r}$,

$$\underset{\sim}{b}_{(r)} = (\underset{\sim}{Z}'_{r}\ \underset{\sim}{Z}_{r})^{-1}\ \underset{\sim}{Z}'_{r}\ \underset{\sim}{Y} \qquad (3)$$

we can then write (2) as

$$E[\underset{\sim}{\varepsilon}'\underset{\sim}{Z}_{r}(Z'_{r}\ Z_{r})^{-1}\ \underset{\sim}{x}_{r}\ \underset{\sim}{x}'_{r}(\underset{\sim}{Z}'_{r}\underset{\sim}{Z}_{r})^{-1}\underset{\sim}{Z}'_{r}\ \underset{\sim}{\varepsilon}] \qquad (4)$$

(Here $\underset{\sim}{x}_{r}$ represents the vector of those elements in $\underset{\sim}{x}$ which correspond to the r included variables).

A similar expression can also be derived for the bias component $B^{2}_{(r)}$ in (1). We have in this case

$$B^{2}_{(r)} = \{(E[\underset{\sim}{b}^{*}_{(r)}] - \underset{\sim}{\beta})'\ \underset{\sim}{x}\ \underset{\sim}{x}'\ (E[\underset{\sim}{b}^{*}_{(r)}] - \underset{\sim}{\beta})\} \qquad (5)$$

Noting that, on using (3), we have

$$E[\underset{\sim}{b}^{*}_{(r)}] - \underset{\sim}{\beta} = \begin{bmatrix} (\underset{\sim}{Z}'_{r}\underset{\sim}{Z}_{r})^{-1}\underset{\sim}{Z}'_{r}\underset{\sim}{Z}_{k-r}\underset{\sim}{\beta}_{k-r} \\ \\ -\underset{\sim}{\beta}_{k-r} \end{bmatrix}$$

then (5) becomes, after some simplification (and also using the partitioning of $\underset{\sim}{x}\ \underset{\sim}{x}'$ which results from writing $\underset{\sim}{x}' = [\underset{\sim}{x}'_{r}\ \ \underset{\sim}{x}'_{k-r}]$), equal to

$$\underset{\sim}{\beta}'_{k-r}\underset{\sim}{Z}'_{k-r}(\underset{\sim}{Z}'_{r}\underset{\sim}{Z}_{r})^{-1}\underset{\sim}{x}_{r}\underset{\sim}{x}'_{r}\ \underset{\sim}{M}_{r}\underset{\sim}{Z}_{k-r}\underset{\sim}{\beta}_{k-r} \qquad (6)$$

where $\underset{\sim}{M}_{r} = \underset{\sim}{I} - \underset{\sim}{Z}_{r}(\underset{\sim}{Z}'_{r}\underset{\sim}{Z}_{r})^{-1}\underset{\sim}{Z}'_{r}$.

Having thus obtained an expression for (1) in terms of σ^2, (4) and (6) we will now proceed to 'average' the value of $\text{MSE}_{(r)}$ over a typical set of future $\underset{\sim}{x}$ determinations. As in the orthogonal case we will now suppose that the fitting stage covariance matrix $\underset{\sim}{X}'\underset{\sim}{X}$ typifies the future covariance pattern of the elements of $\underset{\sim}{x}$. Hence we assume that

$$E[\underset{\sim}{x}\ \underset{\sim}{x}'] = A(\underset{\sim}{X}'\underset{\sim}{X})$$

where A is an arbitrary constant of proportionality. Using this assumption in (4) and (6), and not distinguishing between the now averaged version of MSE_r and the previous version relating to a single specific $\underset{\sim}{x}$, we obtain

$$\begin{aligned}\text{MSE}_{(r)} &= \sigma^2 + A\ E[\underset{\sim}{\varepsilon}'\underset{\sim}{Z}_r(\underset{\sim}{Z}_r'\underset{\sim}{Z}_r)^{-1}\underset{\sim}{Z}_r'\underset{\sim}{\varepsilon}] \\ &\quad + \underset{\sim}{A}\ \underset{\sim}{\beta}_{k-r}'\underset{\sim}{Z}_{k-r}'\underset{\sim}{M}_r\underset{\sim}{Z}_{k-r}\underset{\sim}{\beta}_{k-r}\end{aligned} \qquad (7)$$

Looking at the variance component on the right-hand side of (7) we have

$$\begin{aligned}&A\ E[\underset{\sim}{\varepsilon}'\underset{\sim}{Z}_r(\underset{\sim}{Z}_r'\underset{\sim}{Z}_r)^{-1}\underset{\sim}{Z}_r'\underset{\sim}{\varepsilon}] \\ &= A\ \sigma^2 \text{trace}\ [\underset{\sim}{Z}_r(\underset{\sim}{Z}_r'\underset{\sim}{Z}_r)^{-1}\underset{\sim}{Z}_r'] \\ &= A\ \sigma^2\ \text{trace}\ [\underset{\sim}{Z}_r'\underset{\sim}{Z}_r(\underset{\sim}{Z}_r'\underset{\sim}{Z}_r)^{-1}] \\ &= A\ r\ \sigma^2\end{aligned}$$

Hence, we finally obtain

$$\text{MSE}_{(r)} = \sigma^2(1 + Ar) + A\underset{\sim}{\beta}_{k-r}'\underset{\sim}{Z}_{k-r}'\underset{\sim}{M}_r\underset{\sim}{Z}_{k-r}\underset{\sim}{\beta}_{k-r} \qquad (8)$$

Using (8) we can now proceed to obtain analogues for quantities already derived in chapter 6 specifically for the orthogonal case. We do so by looking in turn at

(a) $MSE_{(0)} - MSE_{(k)}$

(b) $MSE_{(r)} - MSE_{(r+1)}$

(c) $MSE_{(r)} - MSE_{(k)}$

(a) $\underline{MSE_{(0)} - MSE_{(k)}}$

In this case we see immediately that the 'total predictive potential', $MSE_{(0)} - MSE_{(k)}$, is given by

$$A(\underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{X}\,\underset{\sim}{\beta} - k\,\sigma^2).$$

Comparing this with (6.1.8) in which $\underset{\sim}{X}'\underset{\sim}{X}$ was diagonal we see that we can no longer effect a decomposition of this quantity into additive contributions from individual regressors.

(b) $\underline{MSE_{(r)} - MSE_{(r+1)}}$

The main difficulty here is in finding a simple expression for the quantity

$$\underset{\sim}{\beta}'_{k-r}\underset{\sim}{Z}'_{k-r}\underset{\sim}{M}_r\underset{\sim}{Z}_{k-r}\underset{\sim}{\beta}_{k-r} - \underset{\sim}{\beta}'_{k-(r+1)}\underset{\sim}{Z}'_{k-(r+1)}\underset{\sim}{M}_{r+1}\underset{\sim}{Z}_{k-(r+1)}\underset{\sim}{\beta}_{k-(r+1)} \qquad (9)$$

which is essentially the change in subsequent expected bias due to including a further variable Z_{r+1} in the equation at the fitting stage.

We now let $\underset{\sim}{Z}_{r+1}$ denote the matrix of regressor values obtained by augmenting the matrix $\underset{\sim}{Z}_r$ containing

Using (8) we can now proceed to obtain analogues for quantities already derived in chapter 6 specifically for the orthogonal case. We do so by looking in turn at

(a) $MSE_{(0)} - MSE_{(k)}$

(b) $MSE_{(r)} - MSE_{(r+1)}$

(c) $MSE_{(r)} - MSE_{(k)}$

(a) $MSE_{(0)} - MSE_{(k)}$

In this case we see immediately that the'total predictive potential', $MSE_{(0)} - MSE_{(k)}$, is given by

$$A(\underset{\sim}{\beta}'\underset{\sim}{X}'\underset{\sim}{X}\,\underset{\sim}{\beta} - k\,\sigma^2).$$

Comparing this with (6.1.8) in which $\underset{\sim}{X}'\underset{\sim}{X}$ was diagonal we see that we can no longer effect a decomposition of this quantity into additive contributions from individual regressors.

(b) $MSE_{(r)} - MSE_{(r+1)}$

The main difficulty here is in finding a simple expression for the quantity

$$\underset{\sim}{\beta}'_{k-r}\underset{\sim}{Z}'_{k-r}\underset{\sim}{M}_r\underset{\sim}{Z}_{k-r}\underset{\sim}{\beta}_{k-r} - \underset{\sim}{\beta}'_{k-(r+1)}\underset{\sim}{Z}'_{k-(r+1)}\underset{\sim}{M}_{r+1}\underset{\sim}{Z}_{k-(r+1)}\underset{\sim}{\beta}_{k-(r+1)} \qquad (9)$$

which is essentially the change in subsequent expected bias due to including a further variable Z_{r+1} in the equation at the fitting stage.

We now let $\underset{\sim}{Z}_{r+1}$ denote the matrix of regressor values obtained by augmenting the matrix $\underset{\sim}{Z}_r$ containing the regressors already fitted with the observation vector $\underset{\sim}{z}_{r+1}$ corresponding to the newly entered regressor Z_{r+1}. Thus

$$\underset{\sim}{Z}_{r+1} = [\underset{\sim}{Z}_r \; \underset{\sim}{z}_{r+1}].$$

Following an identical argument to that which led to (3.1.8) we find that

$$\underset{\sim}{M}_{r+1} = \underset{\sim}{M}_r(\underset{\sim}{I} - d^{-1} \, \underset{\sim}{z}_{r+1}\underset{\sim}{z}'_{r+1})\underset{\sim}{M}_r \tag{10}$$

where d is the scalar quantity $\underset{\sim}{z}'_{r+1} \, \underset{\sim}{M}_r\underset{\sim}{z}_{r+1}$.

If we also partition $\underset{\sim}{\beta}_{k-r}$ and $\underset{\sim}{Z}_{k-r}$ in the form

$$\underset{\sim}{\beta}_{k-r} = \begin{bmatrix} \beta_{r+1} \\ \underset{\sim}{\beta}_{k-(r+1)} \end{bmatrix} \quad \text{and} \quad \underset{\sim}{Z}_{k-r} = [\underset{\sim}{z}_{r+1} \; \underset{\sim}{Z}_{k-(r+1)}]$$

we find, on using the expression for $\underset{\sim}{M}_{r+1}$ at (6), that (9) can be written (with some simplification) as :

$$\begin{aligned}&\underset{\sim}{\beta}'_{r+1}\underset{\sim}{z}'_{r+1}\underset{\sim}{M}_r[\underset{\sim}{z}_{r+1}\beta_{r+1} + \underset{\sim}{Z}_{k-(r+1)}\underset{\sim}{\beta}_{k-(r+1)}] \\ &\quad -\underset{\sim}{\beta}'_{k-(r+1)}\underset{\sim}{Z}'_{k-(r+1)}\underset{\sim}{M}_r\underset{\sim}{z}_{r+1}[\beta_{r+1}+d^{-1}\underset{\sim}{z}'_{r+1}\underset{\sim}{M}_r\underset{\sim}{Z}_{k-(r+1)}\underset{\sim}{\beta}_{k-(r+1)}]\end{aligned} \tag{11}$$

It follows from (8) that a variable Z_{r+1} will be worth entering on our mean square-error of prediction criterion provided the expression at (11) exceeds σ^2. We now show that the introduction at the fitting stage of the variable Z_{r+1} which maximizes the explained sum of squares (out of the k-r possible choices) corresponds in expectation to the variable subsequently yielding the biggest decrease in mean square prediction error. Further, the same argument shows that the appropriate test statistic for the hypothesis implied above is in fact given by FMAX. The only difference in its use in the present context is that the appropriate null distribution will, in the general case, be a non-central multivariate F distribution. For obvious reasons the possible practical implementation of such a test criterion will not be pursued here.

To verify the above statements we need to refer back to the general results of Chapter 3. In particular we need the expression given towards the end of section 1 of that chapter for the extra explained sum of squares due to introducing a variable $\underset{\sim}{Z}_2^{(i)}$ into the equation. This expression was there given as

$$(\underset{\sim}{X}_1 \; \underset{\sim}{\beta}_1 + \varepsilon)' \, \underset{\sim}{E}_i \; (\underset{\sim}{X}_1 \; \underset{\sim}{\beta}_1 + \varepsilon) \qquad (12)$$

In our present notation the matrix $\underset{\sim}{E}_i$ will be seen to be identical to

$$d^{-1} \underset{\sim}{M}_r \; \underset{\sim}{z}_{r+1} \; \underset{\sim}{z}'_{r+1} \; \underset{\sim}{M}_r$$

On writing $\underset{\sim}{X}_1 \underset{\sim}{\beta}_1 = \underset{\sim}{Z}_r \underset{\sim}{\beta}_r + \underset{\sim}{z}_{r+1} \beta_{r+1} + \underset{\sim}{z}_{k-(r+1)} \underset{\sim}{\beta}_{k-(r+1)}$ and taking expectation (w.r.t. $\underset{\sim}{\varepsilon}$) in (12) we obtain, apart from a σ^2 term, an expression which is the sum of 9 terms similar to those in (11). Five of these terms combine to give exactly expression (11) while the other four terms embody either a $\underset{\sim}{M}_r \; \underset{\sim}{Z}_r$ or $\underset{\sim}{Z}'_r \underset{\sim}{M}_r$ matrix product. Such latter terms are therefore identically zero.

Hence we have demonstrated the validity of the previous assertion, which also completely generalizes a similar result obtained in the orthogonal case.

It is interesting to note that we have also shown that, still retaining the assumptions made regarding future values of $\underset{\sim}{x}$, the introduction of <u>any</u> variable at <u>any</u> stage can be expected to improve the bias component of subsequent mean square error. However this still has to be weighed against the extra σ^2 contribution to instability which is incurred, and one also has to guard against variables included at an earlier stage becoming redundant.

(c) <u>MSE$_{(r)}$ – MSE$_{(k)}$</u>

Finally, we briefly look at the quantity which essentially indicates whether there is any predictive capability left in the remaining set of excluded regressors at any stage. We have immediately from (8) that

$$MSE_{(r)} - MSE_{(k)} = A\underset{\sim}{\beta}'_{k-r}\underset{\sim}{Z}'_{k-r}\underset{\sim}{M}_{r}\underset{\sim}{Z}_{\kappa-r}\underset{\sim}{\beta}_{\kappa-r} - A(k-r)\sigma^2 \qquad (13)$$

It is easy to show that, replacing the β coefficients by their estimates (obtained from the complete equation), we arrive at exactly the test quantity F' discussed in the context of identification earlier. Again, as in the orthogonal case at (6.3.5), we need to refer this statistic to tables of non- central F[κ-r,n-k-1,k-r].

Having generalized much of the theory which was previously obtained strictly in the context of orthogonal regression we go on in the next chapter to present the results of some simulation studies relating to the non-orthogonal case.

Chapter 9 An Empirical Study of the Non-orthogonal Case

9,1 Scope of study

The reasons for carrying out an empirical investigation in a non-orthogonal situation are mainly twofold. Firstly, it is of interest to compare the performance of the proposed forward/backward procedure using F' (method10) with the three procedures based on conventional F which are now possible. Secondly, it will be informative to compare procedure performance in general with the results already obtained in the orthogonal case.

Altogether five different procedures will be investigated. Firstly, we again include the procedures denoted as methods 3 and 6, which are just the forward and backward versions of stepwise procedures based on the use of conventional F. In addition we also include a general forward/backward procedure based on conventional F similar to that described under (c) in (2.1), except that now a backward orientation is imposed. Essentially, beginning with the complete fitted equation, the included variables are searched and tested for a possible deletion according to the conventional F criterion. If, at any stage, no deletions are indicated the excluded variables are then examined with a view to introducing a variable. The procedure terminates when no variables are deleted or entered at a particular stage. For reference purposes the procedure is henceforth referred to as method 9 and, like all other procedures contemplated in this study, is listed and described in Appendix 3. The fourth procedure looked at is the one based on F' which was fully described in (8.5), and is listed as method 10.

The fifth procedure investigated (method 11) is identical to method 10 in all respects except that the FMAX criterion replaces that of F'. The FMAX critical values are however taken to be those strictly applicable to the orthogonal regression situation, i.e., the non-diagonality of $\underset{\sim}{\Omega}$ is completely ignored throughout. The two sets of probability approximations given in Table 8.1 do seem to indicate that the above procedure is as good an approximation to the exact use of FMAX as is afforded by recourse to the rather inelegant Bonferroni approximation.

As a digression, we might just contemplate the possibility of using an exact FMAX approach by simulating a situation which exhibits an equi-correlation structure at all its stages. The actual generation of such an initial set of regressors presents no problem using an observation made by Dunnett and Sobel [22]. Specifically, if $V_i (i = 0, 1, \ldots, n)$ are a set of independent random variables each having zero mean and unit variance, then (for $\rho = 0$)

$$U_i = \sqrt{\rho}\, V_o + \sqrt{1-\rho}\, V_i, \; i = 1, \ldots, n$$

are a set of equi-correlated random variables with common correlation coefficient ρ. A similar transformation applies to the case where ρ is negative, such possibilities being limited by the restriction that $\rho \geq - 1/(n-1)$

(see David [18 ,p.85]). There is however a major impediment to the usefulness of such an approach in a stepwise simulation study. For it is quite easy to show using an inductive argument that, beginning with n equally correlated random variables with correlation coefficients ρ and equal (arbitrary) variances, the common partial correlation coefficient between any n - r variables, holding fixed the remaining r variables, is given by

$$\rho_{(r)} = \frac{\rho}{1+r\rho}$$

In the case where ρ is positive we see that $\rho_{(r)}$ declines monotonically to zero as r increases. For example, when $\rho = 0.5$ we obtain the sequence:

0.5, 0.33, 0.25, 0.2, 0.16,...,

and even for ρ as large as 0.8 we just obtain the sequence

0.8, 0.44, 0.3, 0.23, 0.19,....

It is evident that, starting with a positive valued correlation coefficient, equi-correlation non-orthogonal designs converge very rapidly to orthogonal type configurations when used in a stepwise context. Thus we can expect little light to be thrown on the performance arising from the exact use of FMAX in the general non-orthogonal case. This view was supported by a few simulation runs which were performed with $\rho = 0.5$, negligible differences being revealed between the operation of exact FMAX, approximate

FMAX (taking $\rho = 0$) and F' criteria. Similar remarks apply to the case in which ρ is negative. For suppose we take the lowest possible value of ρ, $\rho = -1/(n-1)$, corresponding to a sample size of n. Also, suppose we think that a subsequent value of $\rho_{(r)}$ equal to $-\frac{1}{3}$ is a desired correlation level in terms of exhibiting the potential use of FMAX. Then we easily find that such a value of $\rho_{(r)}$ only occurs for $r \geq n-4$. Again such a configuration would be of little value for the use which we wish to make of it.

The five different procedures described above (methods 3,6,9,10 and 11) were investigated simultaneously for each of the 18 configurations generated by the 6 specifications of $\underset{\sim}{\beta}$ given in (7.1) each taken at the same three levels of n as before. In the present context the dimensionality of the space of influencing parameters is of course considerably increased by the non-orthogonality now permitted. It was however decided to select (arbitrarily) an initial $\underset{\sim}{X}$ matrix exhibiting a reasonable degree of non-orthogonality at all of its stages in a stepwise sequence. To this end a transformation was applied to an initial matrix which was generated completely randomly from a normal distribution with zero mean and unit variance. By appropriately normalizing the columns of the post-multiplying transformation matrix the initial regressor variables still retained the variance values chosen in the orthogonal case.

Appendix 4 shows the sequence of expected partial correlation matrices generated by introducing variables into the equation according to their (known) magnitudes of β. This gives some idea of the degree of non-orthogonality retained throughout a typical stepwise sequence.

Unlike in the orthogonal case the empirical investigation here only covers the cases for which k = 10. The main reason for limiting the study in this way was to avoid the handling of the cumbersome 30 x 30 transformation matrices which would be necessary in situations like those designated as (g), (h), (i) and (j) in (7.1). The cases which are investigated should however be quite adequate for the type of inferences we wish to draw.

9.2 Description of program and presentation of results

The workings of the programs for both the identification and prediction studies have already been described fully in (7.2) and (7.3). The only alteration now arises in the generation of the initial $\underset{\sim}{X}$ matrix, and this was described in the previous section. As before the number of iterations performed was fixed at 500 and 250 in the two respective cases.

While the form of the summarized tables which follow remains the same as in the orthogonal case in chapter 7 it must be remarked here that the columns for'mixed cases' now incorporate a much greater degree of summarization than before. In particular it frequently happens that the proportion of fitted equations falling into this category actually increases as the value of n gets larger. This is of course no more than is to be expected in situations which, for very small n, possess a strong tendancy to underfit. While a breakdown of mixed cases in such situations usually reveals a reassuring tendency for the 'quality' to improve (in the sense that they get closer to the true model), it was thought unnecessary here to present such a breakdown in detail. Instead it is felt that the 'score' column effectively summarizes such features. As far as the prediction aspect is concerned there are really no new statements to be made. In these cases, as before, we also incorporate the 'control' procedures (methods 7 and 8) into the study.

We now present the results of the simulation study in Tables 9.1 to 9.6. As mentioned above these results run parallel to those already looked at in Tables 7.1 to 7.6.

TABLE 9.1

($\underset{\sim}{\beta} = \underset{\sim}{0}$; Regressor variances = 1)

Table A

Method	n	Variables overfitted										Score
		0	1	2	3	4	5	6	7	8	9	
3	31	64.8	27.2	7.2	0.8	-	-	-	-	-	-	0.44
3	71	65.2	27.0	6.6	1.2	-	-	-	-	-	-	0.44
3	150	66.0	27.2	6.2	0.6	-	-	-	-	-	-	0.41
6	31	64.8	18.8	7.6	2.8	1.2	1.6	1.8	0.6	0.8	-	0.77
6	71	67.6	19.6	5.8	2.6	0.8	1.4	1.2	0.6	0.2	0.2	0.64
6	150	68.6	17.8	7.0	2.2	1.4	0.4	1.8	0.4	0.4	-	0.63
9	31	59.2	23.0	8.8	3.0	1.2	1.6	1.8	0.6	0.8	-	0.84
9	71	61.6	24.6	6.8	2.6	0.8	1.4	1.2	0.6	0.2	0.2	0.69
9	150	61.4	23.6	7.8	2.8	1.4	0.4	1.8	0.4	0.4	-	0.72
10	31	96.2	2.8	0.4	-	0.2	-	-	-	0.2	0.2	0.08
10	71	93.8	4.6	0.4	0.6	-	0.2	0.4	-	-	-	0.11
10	150	96.0	2.6	0.6	0.4	-	-	0.4	-	-	-	0.07
11	31	94.0	5.2	0.2	-	0.2	-	0.2	-	0.2	-	0.09
11	71	96.8	3.0	0.2	-	-	-	-	-	-	-	0.03
11	150	92.6	6.6	0.4	-	-	-	0.4	-	-	-	0.10

Table B

Method	n		
	31	71	150
3	6.2	14.1	27.4
6	4.1	10.8	22.5
9	3.7	9.8	19.8
10	100	100	99.7
11	91.6	67.2	100
7	1.0	2.7	6.1

TABLE 9.2

$\beta' = [1\ 0.5\ 0.5\ 0.5\ 0.5\ 0.25\ 0.25\ \ 0\ \ 0\ \ 0]$;

Regressor variances = 1

Table A

Method	n	Variables under/overfitted									Mixed Cases	Score
		-6	-5	-4	-3	-2	-1	0	1	2		
3	31	0.4	36.8	26.2	12.0	2.2	-	-	-	-	22.4	4.29
3	71	-	8.2	21.6	13.2	11.2	0.8	-	-	-	45.0	4.48
3	150	-	0.6	10.6	14.8	22.2	2.6	0.2	-	-	49.0	3.85
6	31	0.2	9.8	13.2	18.0	7.6	5.0	-	0.2	0.4	45.6	3.83
6	71	-	2.8	6.6	8.6	11.8	6.2	0.4	0.2	-	63.4	3.58
6	150	-	0.2	4.0	7.2	21.8	6.6	2.6	0.2	0.2	56.4	2.88
9	31	-	10.4	13.0	17.2	7.6	4.8	-	0.2	0.4	45.6	3.84
9	71	-	2.4	7.2	6.4	12.8	5.8	0.6	0.2	-	64.6	3.60
9	150	-	0.2	3.6	6.2	22.2	6.2	3.0	0.2	0.2	58.2	2.93
10	31	6.2	51.4	17.0	6.8	1.2	0.6	-	-	0.2	16.6	4.72
10	71	-	16.6	21.4	14.6	7.8	3.0	-	-	-	36.8	4.17
10	150	-	0.8	13.8	21.4	20.8	7.2	-	-	-	36.0	3.15
11	31	3.6	49.0	18.2	7.8	2.4	0.4	-	-	-	18.6	4.59
11	71	-	17.4	18.2	11.6	8.6	4.0	-	-	-	40.2	4.13
11	150	-	2.2	14.4	15.6	23.2	7.0	0.2	-	-	37.4	3.26

Table B

Method	n		
	31	71	150
3	68.0	54.5	54.4
6	54.5	53.3	60.0
9	56.5	55.6	62.4
10	47.1	44.5	47.5
11	50.6	45.8	50.9
7	52.3	66.0	64.0
8	100	100	100

TABLE 9.3

($\beta' = [3\ 3\ 3\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$;
Regressor variances = 1)

Table A

Method	n	Variables overfitted								Mixed Cases	Score
		0	1	2	3	4	5	6	7		
3	31	30.6	11.6	20.2	2.6	10.8	1.8	1.2		21.2	2.47
3	71	10.2	4.0	7.6	2.2	1.4	1.6	51.8	9.2	12.0	5.53
3	150	0.8	1.2	0.6	-	63.4	6.2	26.4	1.4	-	4.55
6	31	65.0	15.8	7.2	2.4	1.8	0.6	0.4	0.2	6.6	0.93
6	71	71.6	16.0	5.8	2.0	1.8	1.6	0.2	-	1.0	0.58
6	150	71.0	18.6	4.4	3.2	0.4	1.6	0.8	-	-	0.51
9	31	61.4	19.2	7.4	2.4	1.8	0.6	0.4	0.2	6.6	0.97
9	71	68.6	18.8	6.2	1.8	1.8	1.6	0.2	-	1.0	0.61
9	150	68.8	20.2	5.0	3.2	0.4	1.6	0.8	-	-	0.54
10	31	90.6	3.4	0.4	0.4	0.2	-	0.2	0.2	4.6	0.37
10	71	93.4	3.2	1.2	0.4	-	0.6	0.2	-	1.0	0.18
10	150	94.0	3.8	0.8	-	0.2	0.8	0.4	-	-	0.13
11	31	88.8	4.0	1.2	-	0.2	-	0.2	0.2	5.4	0.42
11	71	91.8	4.0	1.8	0.2	0.4	0.6	0.2	-	1.0	0.21
11	150	93.0	5.2	0.6	-	-	0.8	0.4	-	-	0.13

Table B

Method	n		
	31	71	150
3	18.9	19.0	32.8
6	48.7	49.1	60.7
9	47.6	47.9	58.4
10	72.3	76.6	86.6
11	67.8	75.7	81.7
7	23.5	24.5	29.5
8	100	100	100

TABLE 9.4

(β' = [10 9 8 7 6 5 0 0 0 0];
Regressor variances = 1)

Table A

Method	n	Variables overfitted					Mixed Cases	Score
		0	1	2	3	4		
3	31	-	-	10.8	32.8	46.2	10. 2	3.68
3	71	7.6	1.6	0.2	83.6	7.0	-	2.81
3	150	-	-	14.6	82.4	3.0	-	2.88
6	31	83.0	12.2	4.0	0.8	-	-	0.23
6	71	83.2	11.2	4.0	1.4	0.2	-	0.24
6	150	83.8	9.6	5.2	1.4	-	-	0.24
9	31	82.2	13.0	4.0	0.8	-	-	0.23
9	71	82.8	11.4	4.2	1.4	0.2	-	0.25
9	150	83.8	9.6	5.2	1.4	-	-	0.24
10	31	96.6	2.0	1.2	0.2	-	-	0.05
10	71	94.4	3.8	1.2	0.4	0.2	-	0.08
10	150	94.4	3.0	1.8	0.8	-	-	0.09
11	31	95.8	3.0	1.0	0.2	-	-	0.06
11	71	94.2	4.4	0.8	0.4	0.2	-	0.08
11	150	94.2	3.6	1.4	0.8	-	-	0.09

Table B

Method	n		
	31	71	150
3	11.5	55.7	85.6
6	76.9	79.9	77.1
9	76.9	79.7	77.0
10	86.7	89.1	89.2
11	85.9	88.8	88.6
7	50.0	51.9	50.2
8	100	100	100

TABLE 9.5

($\beta' = [3 \quad 3 \quad 0.125 \quad 0.125 \quad 0.125 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$;
Regressor variances = 1)

Table A

Method	n	Variables under/overfitted									Mixed Cases	Score
		-3	-2	-1	0	1	2	3	4	5		
3	31	58.2	21.2	1.8	-	-	-	-	-	-	18.8	2.92
3	71	34.8	42.8	3.2	0.2	-					19.0	2.61
3	150	13.8	53.6	12.4	-	0.2					20.0	2.26
6	31	51.2	11.6	1.8	-	0.2	0.2	1.0	0.6	0.2	33.2	3.31
6	71	37.4	25.2	3.2	-	-	0.2	1.0	-	-	33.0	2.94
6	150	13.6	37.6	13.2	-	-	0.6	-	0.6	0.2	34.2	2.63
9	31	48.2	14.8	2.0	-	0.2	0.2	1.0	0.6	0.2	32.8	3.23
9	71	30.0	31.0	3.8	-	-	0.2	1.0	-	-	34.0	2.83
9	150	11.0	39.0	14.2	-	-	0.6	-	0.6	0.2	34.4	2.57
10	31	88.4	2.2	-	-	-	0.2	-	0.2	0.2	8.8	3.11
10	71	79.6	12.6	0.4	-	-	-	0.4	0.2	-	6.8	2.85
10	150	48.6	36.8	3.8	-	-	0.2	-	-	0.2	10.4	2.65
11	31	85.6	4.4	-	-	-	-	0.2	0.2	0.2	9.4	3.11
11	71	73.4	17.6	0.4	-	-	-	0.4	-	-	8.2	2.9
11	150	41.6	41.4	5.4	-	-	-	-	0.4	0.2	11.0	2.6

Table B

Method	n		
	31	71	150
3	69.1	64.9	62.0
6	59.7	56.7	56.3
9	57.5	56.8	58.4
10	100	60.4	52.0
11	91.2	60.7	52.6
7	32.2	39.6	56.0
8	63.0	100	100

TABLE 9.6

(β' = [1 0.5 0.5 0.5 0.5 0.25 0.25 0 0 0] ;
Regressor variances = 9)

Table A

Method	n	Variables under/overfitted								Mixed Cases	Score
		-5	-4	-3	-2	-1	0	1	2		
3	31	1.6	3.8	3.6	34.4	1.8	1.0	0.6	-	53.2	3.32
3	71	-	-	-	27.2	2.6	5.4	2.8	1.4	59.0	2.91
3	150	-	-	-	1.6	0.2	8.2	19.8	2.2	68.0	1.96
6	31	-	0.2	1.8	22.2	6.0	9.4	0.6	0.4	59.4	2.80
6	71	-	-	-	15.4	4.6	32.6	2.8	0.4	44.2	1.71
6	150	-	-	-	0.6	0.2	66.2	3.0	0.2	29.8	0.79
9	31	-	0.2	1.4	23.8	4.8	10.2	0.6	0.4	58.6	2.76
9	71	-	-	-	14.4	4.6	32.6	2.8	0.4	45.2	1.71
9	150	-	-	-	0.6	0.2	66.2	3.0	0.2	29.8	0.79
10	31	0.2	2.6	16.0	35.0	9.4	2.4	-	-	34.4	2.96
10	71	-	-	-	36.2	4.2	20.4	0.4	-	38.8	2.29
10	150	-	-	-	3.2	2.0	63.8	0.8	-	30.2	0.87
11	31	0.8	2.4	9.2	38.6	8.8	4.6	-	-	35.6	2.82
11	71	-	-	-	33.8	2.8	25.6	0.4	0.2	37.2	1.88
11	150	-	-	-	3.4	0.4	65.2	1.2	-	29.8	0.82

Table B

Method	n		
	31	71	150
3	12.4	54.9	62.3
6	21.7	62.1	75.5
9	22.3	62.6	75.5
10	14.9	54.8	68.7
11	16.6	56.1	71.6
7	23.6	72.6	70.1
8	100	100	100

Figure 9.1 : Prediction Efficiencies for case (a)

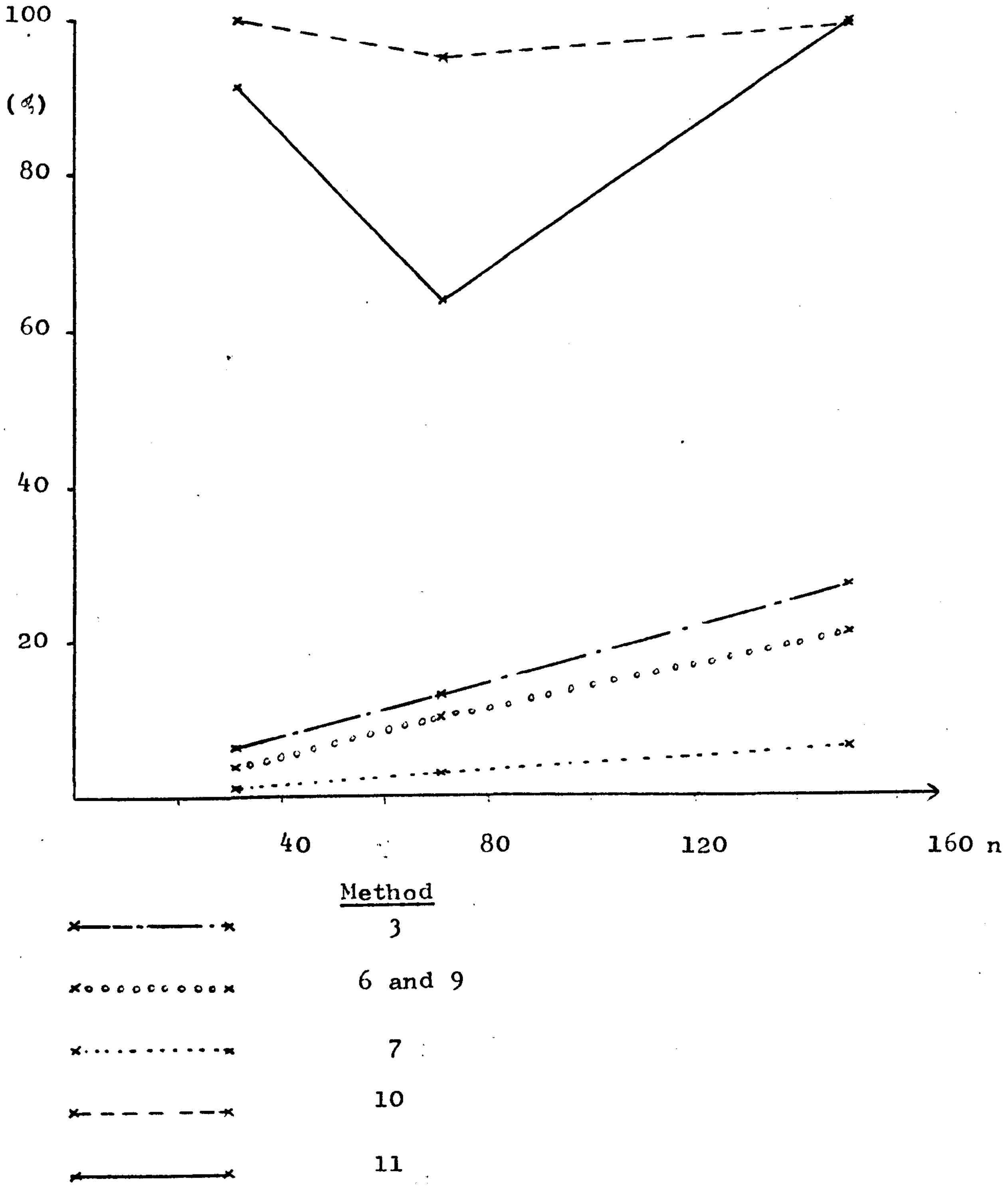


Figure 9.2 : Prediction Efficiencies for case (b)

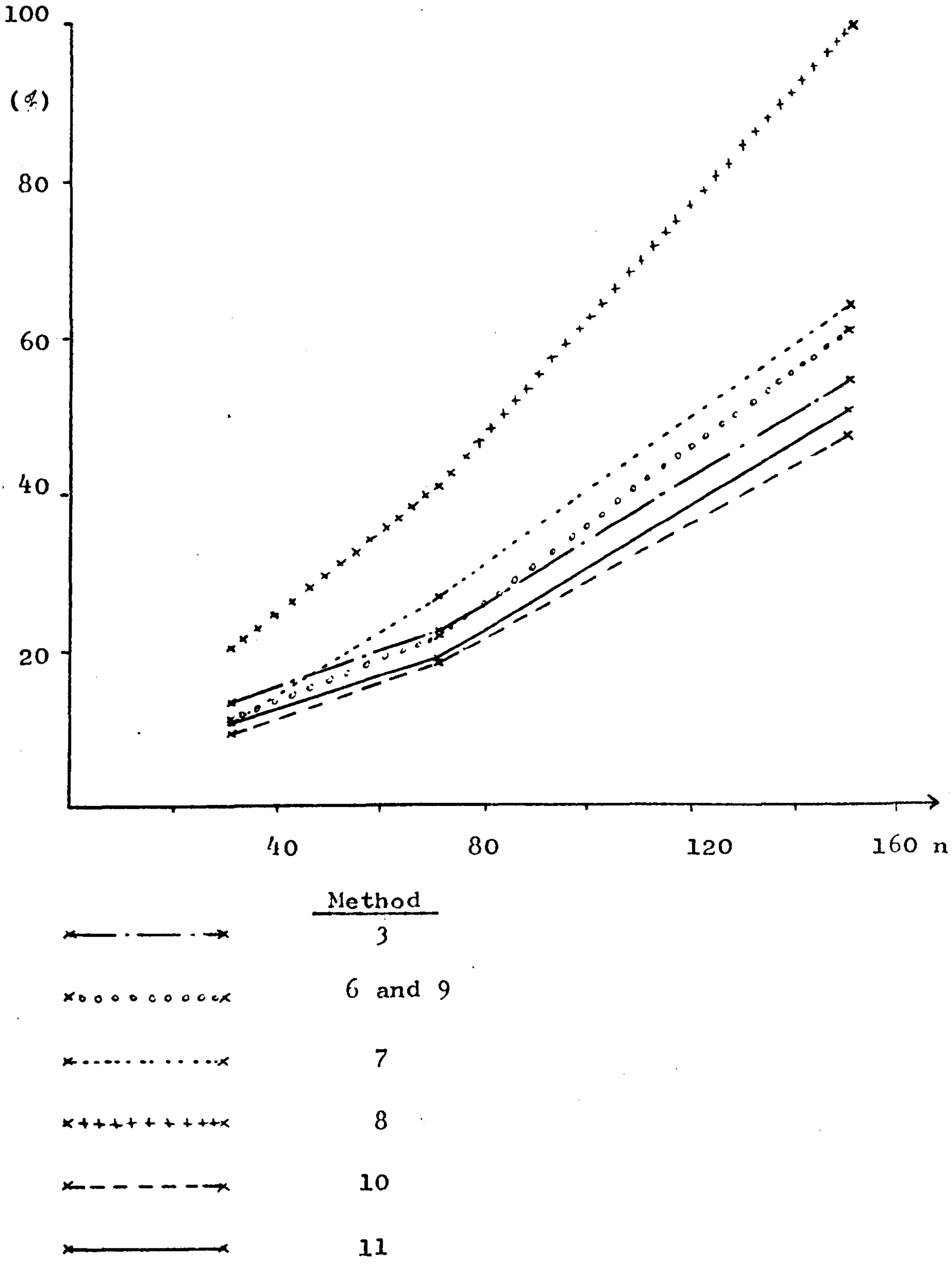


Figure 9.3 : Prediction Efficiencies for case (c)

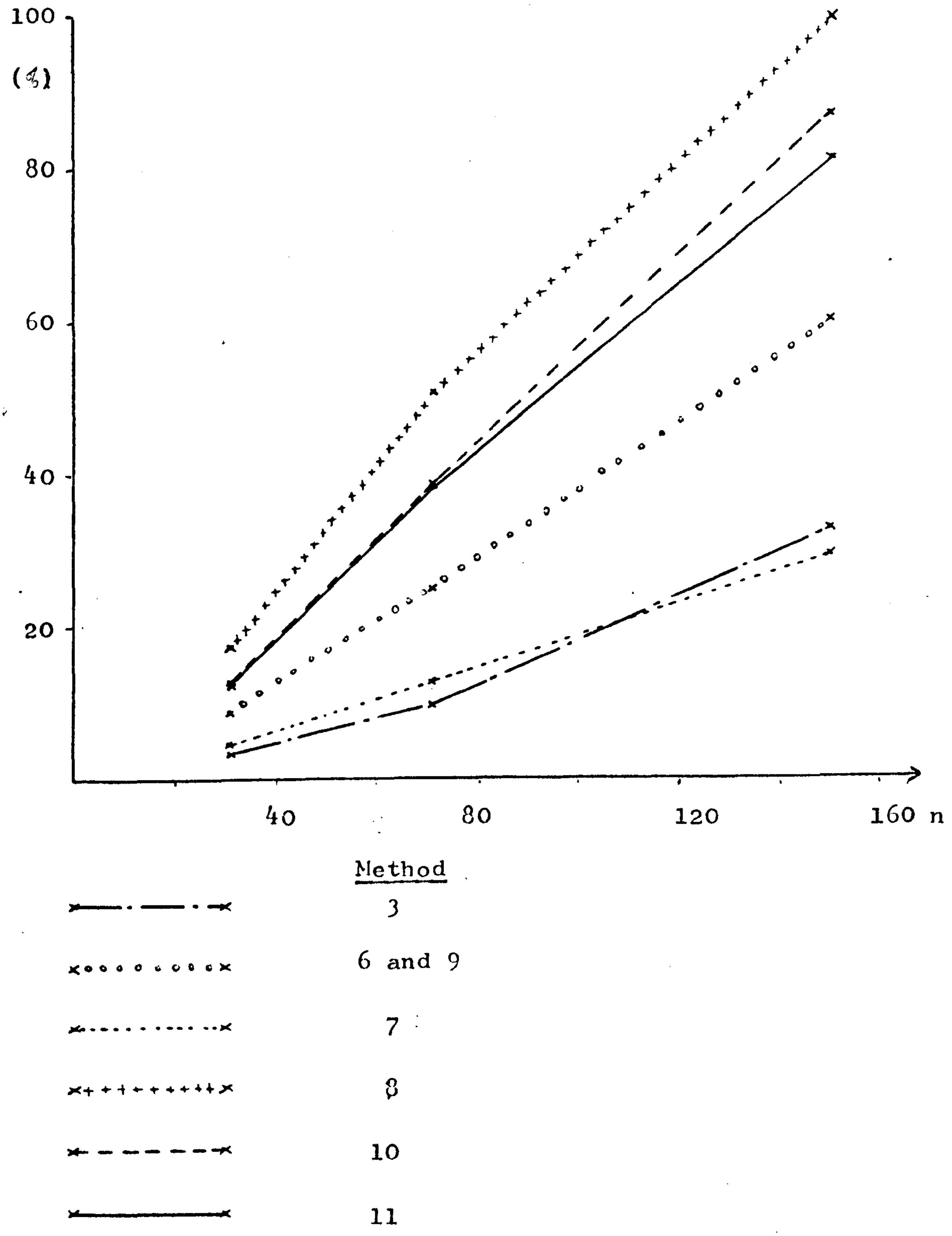


Figure 9.4 : Prediction Efficiences for case (d)

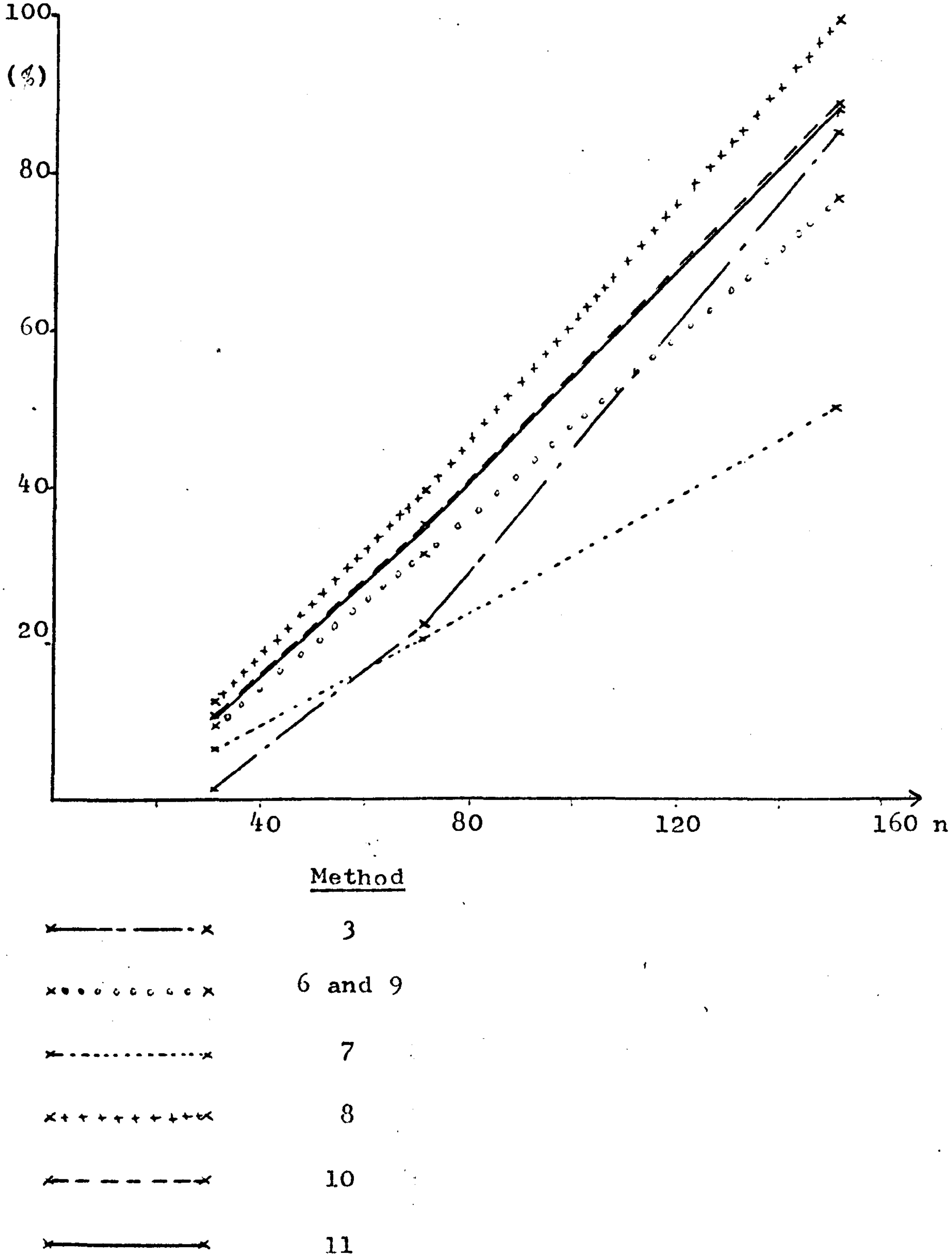


Figure 9.5 : Prediction Efficiencies for case (e)

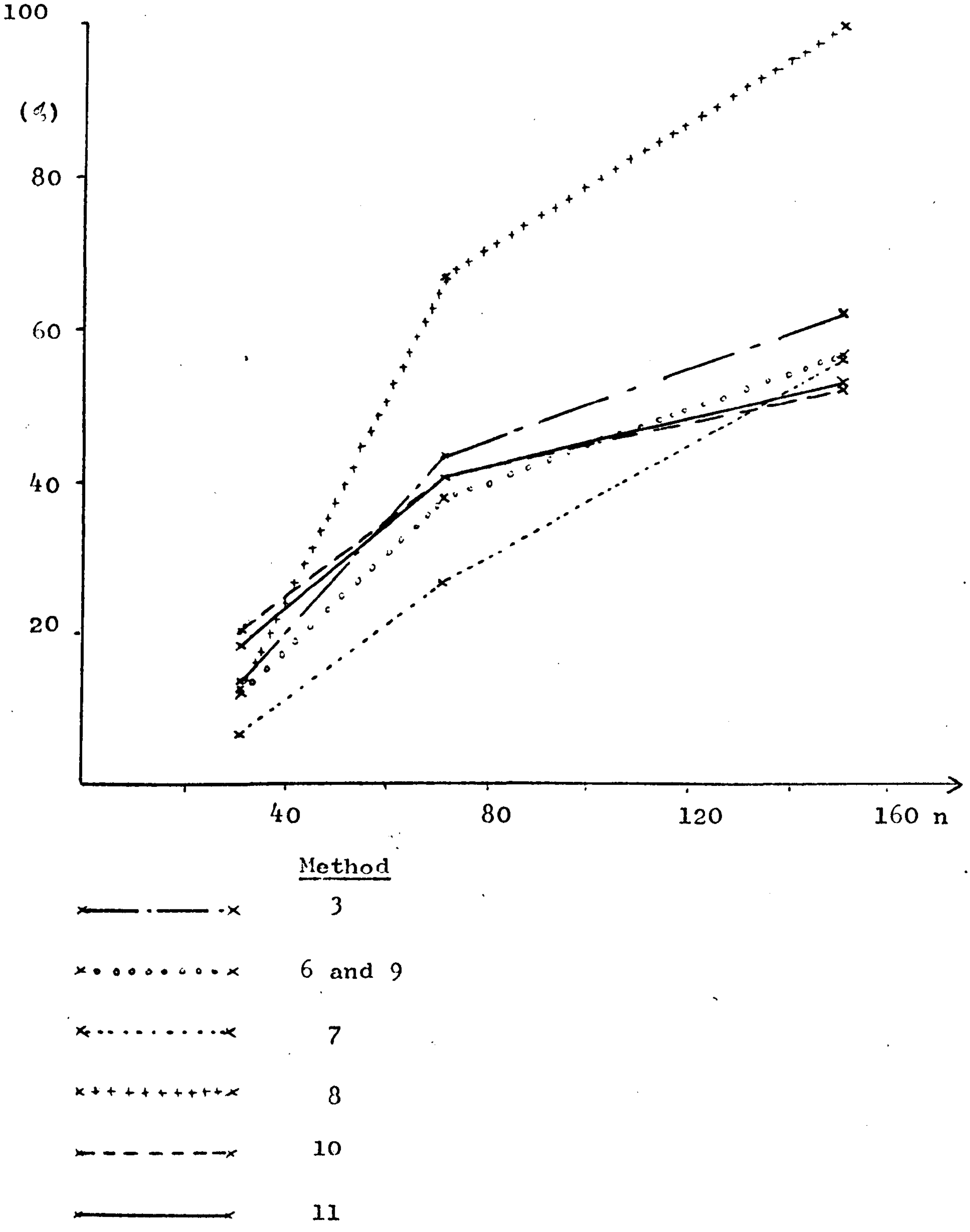
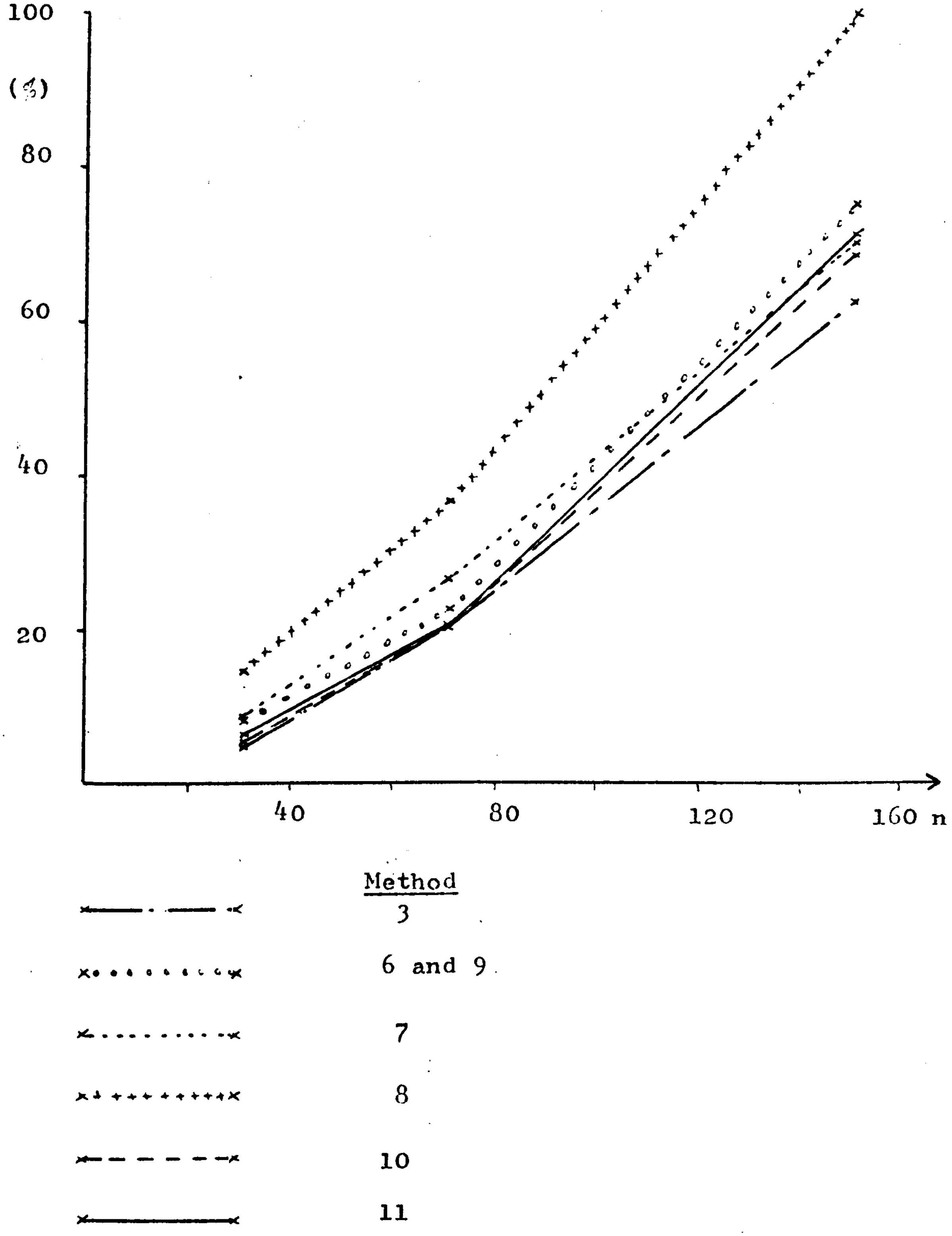


Figure 9.6 : Prediction Efficiencies for case (f)



9.3 Conclusions

If we begin, as in chapter 7, by looking at the results which relate to the identification aspect a very noticeable feature is the sharp increase in the proportion of mixed cases obtained. Indeed situations like those investigated in Tables 9.3 and, to a lesser extent. 9.4 now exhibit this phenomenon for the first time. A general look at the type A tables reveals, not surprisingly, that the procedures based on the conventional F criterion (methods 3, 6 and 9) have a greater tendency to result in mixed cases than do the other two methods. An explanation for this would seem to be the previously observed tendency of such procedures to overfit in comparison with the F' and FMAX procedures. While in the orthogonal situation such a characteristic can be an advantage in the present circumstances it manifests itself in an inability to distinguish between true and spurious regressors. Unfortunately it seems that the hoped for ability of the general forward/backward method to eventually drop such spurious regressors in favour of true ones has not materialized. Thus the indications are that considerably larger sample sizes are required before asymptotic considerations become properly effective. It must be noted that a contributory factor to the above situation is the fact that the conditional variances of excluded variables holding fixed the included set, can be expected to decrease in non-orthogonal set-ups. This will be accompanied by a corresponding reduction in procedure sensitivity.

We might now turn to an examination of the relative merits of the conventional F methods. Here it is somewhat surprising to note that there is hardly any difference in performance between the strict backward approach (method 6) and the general forward/backward version (method 9). However, it would be extremely dangerous to infer that this similarity should be expected to hold true in general. A more complicated correlation structure together with a considerably larger sample size might well lead to an increased disparity between these two approaches. A comparison of the performance of the strictly forward approach of method 3 with that of the other two methods using conventional F reveals some interesting features. In particular, on a strict 'score' basis, the former method comes out best only in Table 9.1 and (more marginally) in Table 9.5. In both cases this superiority can be attributed to the relative overfitting tendency of the latter two procedures. But it is then noted that methods 6 and 9 do not consistently overfit in comparison with method 3, prime counter-examples being the situations of Tables 9.3 and 9.4. We can only remark once again on the extreme range and variability of the performance characteristics of procedures based on conventional F, As far as the F' and FMAX procedures are concerned the outstanding feature is again, as in the orthogonal case, their close similarity in performance. It is very unlikely too that this would have changed even if FMAX could have been used with exact critical values. From the point of view of finding a best overall procedure there continues to be a close agreement with the orthogonal case. For only in Table 9.2 can a conventional F procedure be said to be better than F' on a pure score basis. However, in parallel with the orthogonal case, the very similar set-up of Table 9.6 again demonstrates the coming into play of the desirable asymptotic properties of F'.

Switching now to the prediction aspect our conclusions are now almost exactly as they were in the orthogonal set-up. For direct comparisons of Figure 9.1 to 9.6 with their counterparts Figures 7.1 to 7.6 reveals a striking similarity in performance. The only change worth remarking on is the slightly greater efficiency of method7 in the similar situations of Figure 9.2 and 9,6. This is entirely consistent with the reduced sensitivity of the other stepwise procedures in detecting regressor influences in the present non-orthogonal context. Finally, we again note in Fig. 9..5 the small sample superiority of the F' procedure over that of method 8 which uses knowledge of the specified model.

It cannot be overstressed, in overall summary of the investigations carried out here, that the results must relate very specifically to the correlation structure actually used, and one must not expect too much by way of generalization to other non-orthogonal situations. There does however seem to be sufficient evidence to suggest that the F' and FMAX approaches do behave in accordance with the underlying principles from which their use is derived. This contrasts with the other three approaches which, while they sometimes fortuitously appear to do slightly better, on the whole suffer from having no underlying rationality for their use.

In the next chapter we proceed to the possibly even more ambitious task of using stepwise routines in the detection of time series models.

Chapter 10 Stepwise Regression in Stochastic Regression Models.

10.1 The independent stochastic regression model

Up to this point our discussion has been based entirely on the assumption that the matrix $\underset{\sim}{X}$ consists only of fixed constants. This is in fact the context in which classical regression theory is usually discussed, and is indeed in harmony with early areas of application of the technique. However in more recent times a demand has developed for techniques applicable to more general models than the classical version. This has been especially true in the field of non-experimental science which has been (and still is) undergoing what can almost be described as a "quantitative revolution". A prime example in which this has occurred has been in the study of economics. While it is true to say that quantitative formulations of economic theory have a long history the appropriate techniques of testing and verification for such models using actual data were not fully studied until comparatively recently.

The main extension which we need to make regarding the model given at (1.2.1) is to allow the X variables to be stochastic in addition to ε. The point is that since the deliberate selection of representative X values is often not practicable, and since also for similar reasons replication of such values in an experimental sense is impossible, the inferential basis has now to be generalised to incorporate an underlying population for $\underset{\sim}{X}$. The simplest form of such an assumption which can be made is to suppose that the $\underset{\sim}{X}$ matrix constitutes a sample of size n from a k-dimensional multivariate random variable with density function $h(\underset{\sim}{x})$, and also that the residual vector $\underset{\sim}{\varepsilon}$ is independent of $\underset{\sim}{X}$. With such an assumption it follows (e.g. see Goldberger [27 ,p.271]) that the least squares estimator of $\underset{\sim}{\beta}$ in (1.2.1) is unbiased and also, under general conditions on $h(\underset{\sim}{x})$, consistent. With the slightly stronger condition that $h(\underset{\sim}{x})$ does not involve $\underset{\sim}{\beta}$ or σ^2 it is easy to show (e.g. see Johnston [35 ,p.29]) that the least squares estimator $\underset{\sim}{b}$ is also the maximum likelihood estimator. Thus least squares retains the desirable properties of unbiasedness, consistency, efficiency and sufficiency in this more general formulation of (1.2.1), though of course the linearity of the estimator $\underset{\sim}{b}$ is now lost. Finally, by stipulating that $h(\underset{\sim}{x})$ is the multivariate normal density, the stronger property of minimum variance then holds for $\underset{\sim}{b}$ (see Graybill [30,p.198]).

From the viewpoint of stepwise regression, so far as it has been formulated for model identification objectives, the crucial point (apart from the unbiasedness property) is that the hypothesis tests used continue to be valid. That this holds true follows automatically from the independence assumption for $\underset{\sim}{X}$ and $\underset{\sim}{\varepsilon}$ (see Johnston [35 ,p.31]). The situation will however be changed regarding the power characteristics of the various tests used. For instead of requiring distributions of non-central F or t as in the case of fixed $\underset{\sim}{X}$ the necessary distribution theory becomes that of partial correlation analysis from normal samples (assuming $\underset{\sim}{X}$ is multivariate normal). If we turn to the prediction objective which was discussed in chapters 6 and 8 things become less straightforward. For, apart from having to re-specify the forms of the various hypotheses, such as (6.2.1) for example, the relevant distribution theory even under these null hypotheses can be expected to be extremely forbidding. Since, in a stepwise regression context, the precision that would result from such an exact treatment can only really be justified in asymptotic terms the matter is not thought to be worth pursuing here. In any case one expects the unconditional theory to converge to that of the conditional case as n gets large due to the consistency property of the sample covariance (see for example Kendall and Stuart [37 ,p.340] who demonstrate this result for the unconditional and conditional distributions of R^2 in multivariate normal samples).

10.2 The application of stepwise regression to time series models

In this section we briefly discuss the applicability of the stepwise regression technique to an important class of results arising in the analysis of time series. We follow this up in the next section with an exploratory simulation study of procedure performance in such applications.

We are interested in obtaining models which tell us something about the behaviour of random variables Y_t which are observed at various discrete time points indexed by the parameter t. Data plots of time series samples more often than not reveal noticeable non-stationary † characteristics in the sense that there are evident deterministic time dependencies present.

† It is not thought appropriate here to proceed with a detailed exposition of time series analysis. Thus terms will often be used without going into lengthy formal definitions and explanations of their meaning. Many excellent textbooks now exist which deal with the various concepts referred to above. Amongst those found to be particularly useful are Anderson [5], Box and Jenkins [14] and Hannan [32]

However such deterministic components can very often be effectively eliminated leaving a stationary residual component. Unless these non-deterministic residuals are already in the form of an uncorrelated process it will be advantageous to reduce their apparent unpredictability by attempting to fit some kind of explanatory model. The most general class of model which might be considered here is the moving average model. For it was shown by Wold [77] that any purely non-deterministic stationary process $\{Y_t\}$ can be represented as:

$$Y_t = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j} \qquad (1)$$

where the sequence $\{\beta_j; j = 0,1,\ldots\ldots\}$ is a set of constants satisfying $\sum_{j=0}^{\infty} \beta_j^2 < \infty$, and where $\{\varepsilon_t; t = 0, \pm 1,\ldots\ldots\}$ is a sequence of zero mean uncorrelated random variables with common variance σ^2 (i.e. $\{\varepsilon_t\}$) is a "white noise" process. Defining a generating function $G(Z) = \sum_{j=0}^{\infty} \beta_j Z^j$ for the $\{\beta_j\}$ sequence it then follows that, provided $G^{-1}(Z)$ is convergent for $|Z| \leq |$, an alternative autoregressive representation exists for Y_t, i.e.

$$Y_t = \sum_{j=1}^{\infty} \alpha_j Y_{t-j} + \varepsilon_t \qquad (2)$$

The uncorrelated nature of the $\{\varepsilon_t\}$ sequence can be used to demonstrate that in (2) ε_t is uncorrelated with Y_{t-j} for $j = 1,2,.....$.

We shall, henceforth, concentrate on models of the form of (2) since, unlike models of type (1), they lend themselves to the application of linear least squares stepwise regression routines. In practice of course one has to truncate the infinite limit on the number of lags involved in (2) but, provided the truncation point is chosen sufficiently large to include all potentially important terms, such an approximation should be of little consequence. Thus, given a sample of observations Y_t, $t = 1,....,T$, and deciding on a maximum lag p, the model at (2) can be formally used to obtain a standard set of equations as represented by (1.2.1).

Hence, proceeding with the relationship

$$Y_t = \sum_{j=1}^{p} \alpha_j Y_{t-j} + \varepsilon_t \qquad (3)$$

we write

$$\underset{\sim}{Y} = \begin{bmatrix} Y_{p+1} \\ Y_{p+2} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ Y_T \end{bmatrix}, \quad \underset{\sim}{X} = \begin{bmatrix} Y_p & Y_{p-1} \cdots\cdots & Y_1 \\ Y_{p+1} & Y_p \quad \cdots\cdots & Y_2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ Y_{T-1} & Y_{T-2} \cdots\cdots & Y_{T-p} \end{bmatrix},$$

$$\underset{\sim}{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \alpha_p \end{bmatrix} \quad \text{and } \underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_{p+1} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_T \end{bmatrix}$$

whereby we are then enabled to write the estimation problem in the form

$$\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\alpha} + \underset{\sim}{\varepsilon} \qquad (4)$$

similar to (1.2.1). The problem now is to justify the use of ordinary least squares in obtaining an estimator of $\underset{\sim}{\alpha}$ in (4).

We might note first of all that the stochastic regression relationship at (4) violates the assumption made in the previous section with regard to the complete independence of $\underset{\sim}{X}$ with $\underset{\sim}{\varepsilon}$. For the j^{th} element of $\underset{\sim}{\varepsilon}$ will necessarily be correlated with elements in the $(j+1)^{th}$ and succeeding rows of the matrix $\underset{\sim}{X}$. The resulting effect on the usual least squares estimator $\underset{\sim}{a}$ is that bias is obtained. For we have

$$\begin{aligned} \underset{\sim}{a} &= (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'(\underset{\sim}{X}\underset{\sim}{\alpha}+\underset{\sim}{\varepsilon}) \\ &= \underset{\sim}{\alpha} + (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{\varepsilon} \qquad (5) \end{aligned}$$

and, on taking expectation with respect to $\underset{\sim}{X}$ and $\underset{\sim}{\varepsilon}$, the second term on the right-hand side will now no longer vanish. Though the presence of this bias element alone is sufficient to seriously undermine the validity of stepwise regression procedures this is not the only difficulty that has to be contended with. For the usual testing procedures of the classical approach which were discussed in (1.3) will also no longer apply. One way of viewing why this is so is to recognize that the classical decompositions of sums of squares will not now behave as independent chi-square variates. While this problem can be effectively side-stepped in the independent stochastic regression model discussed in section 1 such simplifications are not now available to us.

In view of the above considerations the only hope left to us is that ordinary least squares will at least be viable in an asymptotic sense. That this is indeed so is a consequence of what Goldberger [27] refers to as the "contemporaneous uncorrelation" property possessed jointly by $\underset{\sim}{X}$ and $\underset{\sim}{\varepsilon}$. For provided the error term ε_j in the j^{th} equation of (4) is uncorrelated with regressors occurring in the same and preceding equations it follows, by taking probability limits in (5), that $\underset{\sim}{a}$ is a consistent estimator of $\underset{\sim}{\alpha}$. This is in fact a general result for stochastic regression models where, in the general case, we would need to make appropriate stationarity assumptions about the form of the distribution of the regressor variables. Despite the consistency property referred to above we still need to consider the asymptotic distribution of the estimator $\underset{\sim}{a}$ in order to arrive at appropriate asymptotic test procedures. The definitive paper dealing with this problem is that of Mann and Wald [56]. These authors showed that, with a normality assumption for the distribution of $\underset{\sim}{\varepsilon}$ in (4), the conditional maximum likelihood estimator of $\underset{\sim}{\alpha}$ obtained by holding fixed the sample values $Y_1, Y_2, \ldots\ldots, Y_p$ is the same as the ordinary least squares estimator $\underset{\sim}{a}$. Moreover, they demonstrated the important result that the asymptotic distribution of $\underset{\sim}{a}$ converges to the full (unconditional) maximum likelihood estimator. The consequence is that the least squares estimator of $\underset{\sim}{\alpha}$ is, asymptotically, efficient and normally distributed with the standard least squares covariance properties.†

† Anderson [5 , Chapt.5] gives a full account of the derivation of these results, which he also generalizes to models involving general stochastic regressors in addition to the lagged dependent variables. The same conclusions are shown also to apply when the error terms have some general distribution other than the normal, although the efficiency property in this case no longer holds.

Having satisfied ourselves at least as to the asymptotic validity of least squares in the estimation of autoregressive models, we must now see where this leaves us in relation to the intended use of stepwise regression procedures in this context. It is evident that we shall find ourselves in the same situation as described in chapter 8 relating to non-orthogonal set-ups. The only difference here is that we can now interpret the $\underset{\sim}{\Omega}$ matrix which arose in that chapter as exhibiting the partial autocorrelation structure of the excluded lagged variables, holding fixed the included lagged values.

Before proceeding with the presentation of an empirical study relating to the present situation some comments should perhaps be made concerning the motivation for applying stepwise regression atall in such circumstances. Perhaps the major justification for doing so is to identify which lagged terms really are directly related to the present value of a variable. Much consideration has been paid in the published literature on determining the 'order' of an autoregressive process, in the sense of finding the highest lagged term having a non-zero coefficient. Such procedures which have been presented then usually proceed to fit successively higher order equations until the result of some decision process implies one should terminate (see for example Quenouille [67], Bartlett and Diananda [8] Whittle [75] and Anderson [4]). While this approach has certain desirable aspects† we have already seen that it is quite possible to end up with an equation implying a very spurious lag structure.

10.3 An empirical study

In parallel with the orthogonal and non-orthogonal cases of classical regression previously looked at it was decided to gauge the effectiveness of the various stepwise regression procedures when applied to data generated from known time series models. A change did of course have to be made to the simulation program previously described to allow for the different data generation procedure which is now required, and also to deal with the different way in which the initial correlation matrix has to be constructed. As far as the identification case is concerned everything else remains as before. In the prediction case however (which should perhaps now be referred to as forecasting) a further 50 observations were generated at each iteration on top of each

†A particular simplifying feature is that one does not encounter problems arising from the data-induced selection of the maximum sum of squares regressors. Also, simple recursive procedures are available for calculating the particular partial autocorrelation coefficients which are needed in such approaches (e.g. see Durbin [23])

fitting set obtained, and one-step forecast errors were evaluated using these extra values.

Altogether five different models were investigated, these being as follows:-

(a) $Y_t = \varepsilon_t$

(b) $Y_t = 0.125\ Y_{t-1} + \varepsilon_t$

(c) $Y_t = 0.5\ Y_{t-1} + \varepsilon_t$

(d) $Y_t = 0.25\ Y_{t-3} + 0.5\ Y_{t-7} - 0.125\ Y_{t-10} + \varepsilon_t$

(e) $Y_t = 0.8\ Y_{t-1} - 0.8\ Y_{t-2} + \varepsilon_t$

All of these models can be shown to satisfy the stationarity condition for an autoregressive process (i.e. that the roots of the characteristic polynomial $A(Z) = \sum_{j=0}^{p} \alpha_j Z^j$ $(\alpha_o=1)$ fall outside the unit circle in the complex plane).†

Each of the above models was then used in simulation investigations corresponding to sample

†It should be pointed out that although none of the models investigated incorporated a constant (mean) term allowance was in fact made for such a term in forming the correlation matrix. This should not however seriously effect the subsequent comparative performances of the various procedures used.

sizes T = 41,81 and 200 respectively. Since in all cases the lag truncation point p was taken as before to be 10 this means that the effective sample sizes were in fact 31,71 and 190 respectively.

The procedures themselves which were investigated were exactly the same as those looked at for the classical non-orthogonal case in chapter 9, i.e., methods 3,6,9,10 and 11. In addition, in the prediction studies, the control methods 7 and 8 were again incorporated.

In all cases the first twenty observations produced by the generation procedure were discarded in order to avoid possible influences arising from starting up effects. One further change from the previous studies is in the number of iterations produced for each configuration, it being found convenient to choose 150 and 250 iterations in the identification and prediction cases respectively.

The following Tables 10.1 to 10.5 summarize the simulation results in exactly the same way as in the previous empirical studies. Thus tables of type A show the distribution of the various equations arrived at by the various procedures, the score value being the average number of variables incorrectly included or omitted at each iteration. Similarly the type B tables show, for each value of T, the ratio of mean square prediction (forecast) errors using as a base the procedure yielding the smallest such value. As before the prediction errors used exclude the common σ^2 component associated with the residual terms ε_t. The tables are followed in the usual way by graphs illustrating the comparative predictive performances as the sample size T increases. In this case the overall optimum observed mean square prediction error for any T is used as base.

Finally, although the various stepwise methods are described collectively in Appendix 3, it might be of some help to give here a brief reminder that methods 3, 6 and 9 are the forward, backward and forward/backward procedures using conventional F whilst methods 10 and 11 involve the forward/backward use of F' and FMAX respectively.

TABLE 10.1

$(Y_t = \epsilon_t)$

Table A

Method	T	Variables overfitted					Score
		0	1	2	3	4	
3	41	69.3	25.3	4.7	0.7	-	0.37
	81	69.3	23.3	5.3	1.3	0.7	0.41
	200	60.0	31.3	6.7	2.0	-	0.51
6	41	68.7	23.3	6.7	0.7	0.7	0.42
	81	68.7	22.7	5.3	1.3	2.0	0.45
	200	60.0	30.0	6.7	3.3	-	0.53
9	41	68.0	24.0	6.7	0.7	0.7	0.42
	81	68.0	23.3	5.3	1.3	2.0	0.46
	200	60.0	30.0	6.7	3.3	-	0.53
10	41	98.7	1.3	-	-	-	0.01
	81	94.7	4.0	0.7	0.7	-	0.08
	200	94.0	6.0	-	-	-	0.06
11	41	98.0	2.0	-	-	-	0.02
	81	96.0	4.0	-	-	-	0.04
	200	95.3	4.7	-	-	-	0.05

Table B

	T		
Method	41	81	200
3	4.0	10.1	21.2
6	2.9	9.0	20.2
9	2.9	8.8	20.2
10	100	100	100
11	38.7	57.3	84.4
7	0.6	2.3	5.9

TABLE 10.2

$$(Y_t = 0.125 \ Y_{t-1} + \epsilon_t)$$

Table A

Method	T	Variables under/overfitted					Mixed Cases	Score
		-1	0	1	2	3		
3	41	63.6	5.6	1.2	0.8	-	28.0	1.29
	81	59.6	11.2	2.8	-	-	26.4	1.18
	200	43.6	22.0	10.0	3.2	0.8	20.4	1.08
6	41	60.0	5.6	2.4	1.2	-	30.8	1.38
	81	58.4	10.8	2.8	0.4	0.4	27.2	1.23
	200	42.8	22.0	9.6	3.6	1.2	20.8	1.10
9	41	59.6	5.6	2.4	1.2	-	31.2	1.38
	81	58.0	10.8	2.8	0.4	0.4	27.6	1.23
	200	42.8	22.0	9.6	3.6	1.2	20.8	1.10
10	41	96.0	0.8	0.4	-	-	2.8	1.02
	81	96.8	1.2	0.4	-	-	1.6	1.00
	200	83.6	9.6	1.6	-	-	5.2	0.96
11	41	95.2	-	0.4	-	-	4.4	1.06
	81	92.8	3.2	0.4	-	-	3.6	1.00
	200	81.6	14.0	0.4	-	-	4.0	0.90

Table B

Method	T		
	41	81	200
3	30.6	28.8	23.9
6	23.5	27.5	23.2
9	23.1	27.5	23.2
10	100	64.4	32.5
11	93.8	58.3	32.5
7	4.7	9.1	9.7
8	71.5	100	100

TABLE 10.3

$$(Y_t = 0.5 Y_{t-1} + \epsilon_t)$$

Table A

Method	T	Variables under/overfitted						Mixed Cases	Score
		-1	0	1	2	3	4		
3	41	17.6	44.0	14.8	1.6	-	-	22.0	0.84
	81	1.6	68.4	21.6	6.0	0.4	-	2.0	0.40
	200	-	67.2	27.2	4.8	0.8	-	-	0.39
6	41	18.0	39.2	14.4	4.4	2.0	0.4	21.6	1.00
	81	1.2	65.2	20.0	8.8	2.4	-	2.4	0.52
	200	-	64.0	25.2	9.2	1.6	-	-	0.48
9	41	17.2	39.6	14.4	4.4	2.0	0.4	22.0	1.00
	81	1.2	64.8	20.4	8.8	2.4	-	2.4	0.53
	200	-	63.6	25.6	9.2	1.6	-	-	0.49
10	41	79.6	14.4	0.8	0.4	-	-	5.2	0.92
	81	23.6	70.8	3.6	-	-	-	2.0	0.32
	200	-	95.6	4.0	0.4	-	-	-	0.05
11	41	59.2	31.2	1.6	-	-	-	8.0	0.77
	81	9.6	84.4	4.0	0.4	-	-	1.6	0.18
	200	-	95.6	4.0	0.4	-	-	-	0.05

Table B

Method	T		
	41	81	200
3	17.5	24.9	26.8
6	15.6	24.0	24.5
9	15.6	24.0	24.5
10	14.0	14.4	65.1
11	16,0	23.7	57.8
7	7.8	8.7	8.7
8	100	100	100

TABLE 10.4

$$(Y_t = 0.25 \quad Y_{t-3} + 0.5 \quad Y_{t-7} - 0.125\ Y_{t-10} + \epsilon_t)$$

Table A

Method	T	Variables under/overfitted						Mixed Cases	Score
		-3	-2	-1	0	1	2		
3	41	13.2	30.8	20.0	-	-	-	36.0	2.44
	81	-	14.0	50.4	5.2	0.4	-	30.0	1.58
	200	-	-	45.2	19.2	10.4	0.4	24.8	1.09
6	41	12.0	25.6	14.8	-	-	-	47.6	2.76
	81	-	12.8	46.4	2.4	0.4	-	37.8	1.78
	200	-	-	44.4	17.2	10.0	2.4	26.0	1.16
9	41	11.6	24.4	16.4	-	-	-	47.6	2.73
	81	-	12.8	45.6	2.8	0.4	-	38.0	1.80
	200	-	-	44.4	17.2	10.0	2.4	26.0	1.16
10	41	52.0	29.6	4.0	-	-	-	14.4	2.76
	81	5.6	50.4	36.0	-	-	-	8.0	1.78
	200	-	2.8	84.0	6.8	0.8	-	5.6	1.02
11	41	43.2	32.8	8.0	-	-	-	16.0	2.67
	81	3.2	43.2	46.8	-	-	-	6.8	1.64
	200	-	1.6	86.8	3.2	0.4	-	8.0	1.06

Table B

Method	T		
	41	81	200
3	30.6	24.6	41.6
6	18.0	22.9	39.6
9	19.6	23.0	39.5
10	19.1	18.7	43.3
11	20.4	23.6	41.9
7	12.5	13.7	24.1
8	100	100	100

TABLE 10.5

$(Y_t = 0.8\ Y_{t-1} - 0.8\ Y_{t-2} + \epsilon_t)$

Table A

Method	T	Variables under/overfitted									Mixed Cases	Score
		-2	-1	0	1	2	3	4	5	6		
3	41	-	0.4	8.8	20.8	8.4	0.8	-	-	-	60.8	2.43
	81	-	-	0.8	56.8	23.2	4.4	-	-	-	14.8	1.65
	200	-	-	-	73.2	23.2	2.4	1.2	-	-	-	1.32
6	41	-	0.4	70.0	12.0	6.0	3.6	2.0	-	0.4	5.6	0.62
	81	-	-	73.2	15.6	7.2	2.8	0.8	0.4	-	-	0.44
	200	-	-	72.4	13.2	7.6	4.8	1.2	0.8	-	-	0.52
9	41	-	0.4	66.8	14.8	6.0	3.2	2.4	-	0.4	6.0	0.66
	81	-	-	68.4	19.2	8.4	2.8	0.8	0.4	-	-	0.50
	200	-	-	71.2	14.4	7.2	5.2	1.2	0.8	-	-	0.53
10	41	3.2	3.6	82.0	0.8	-	0.4	-	-	0.4	9.6	0.45
	81	-	-	97.6	1.6	0.4	0.4	-	-	-	-	0.04
	200	-	-	97.6	1.2	0.4	0.4	0.4	-	-	-	0.05
11	41	2.4	1.2	86.8	1.6	-	0.4	0.4	-	0.4	6.8	0.34
	81	-	-	97.2	2.4	-	0.4	-	-	-	-	0.04
	200	-	-	94.0	3.6	2.0	0.4	-	-	-	-	0.09

Table B

Method	T		
	41	81	200
3	10.7	12.5	74.4
6	36.6	35.9	45.9
9	35.7	35.4	45.0
10	29.7	83.2	86.8
11	46.1	59.1	86.1
7	15.5	15.7	12.5
8	100	100	100

Figure 10.1 : Prediction Efficiencies for case (a)

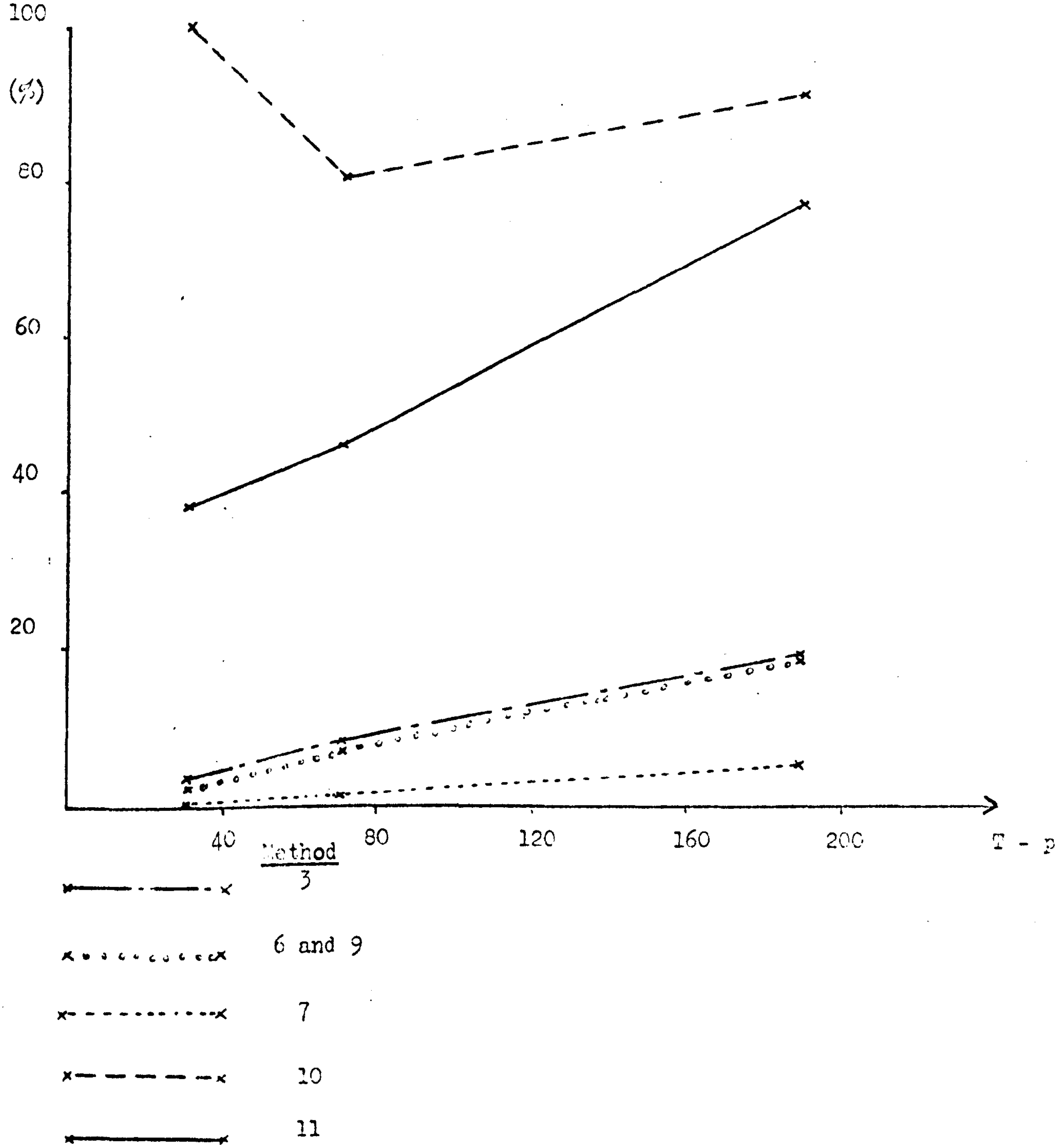


Figure 10.2 : Prediction Efficiencies for case (b)

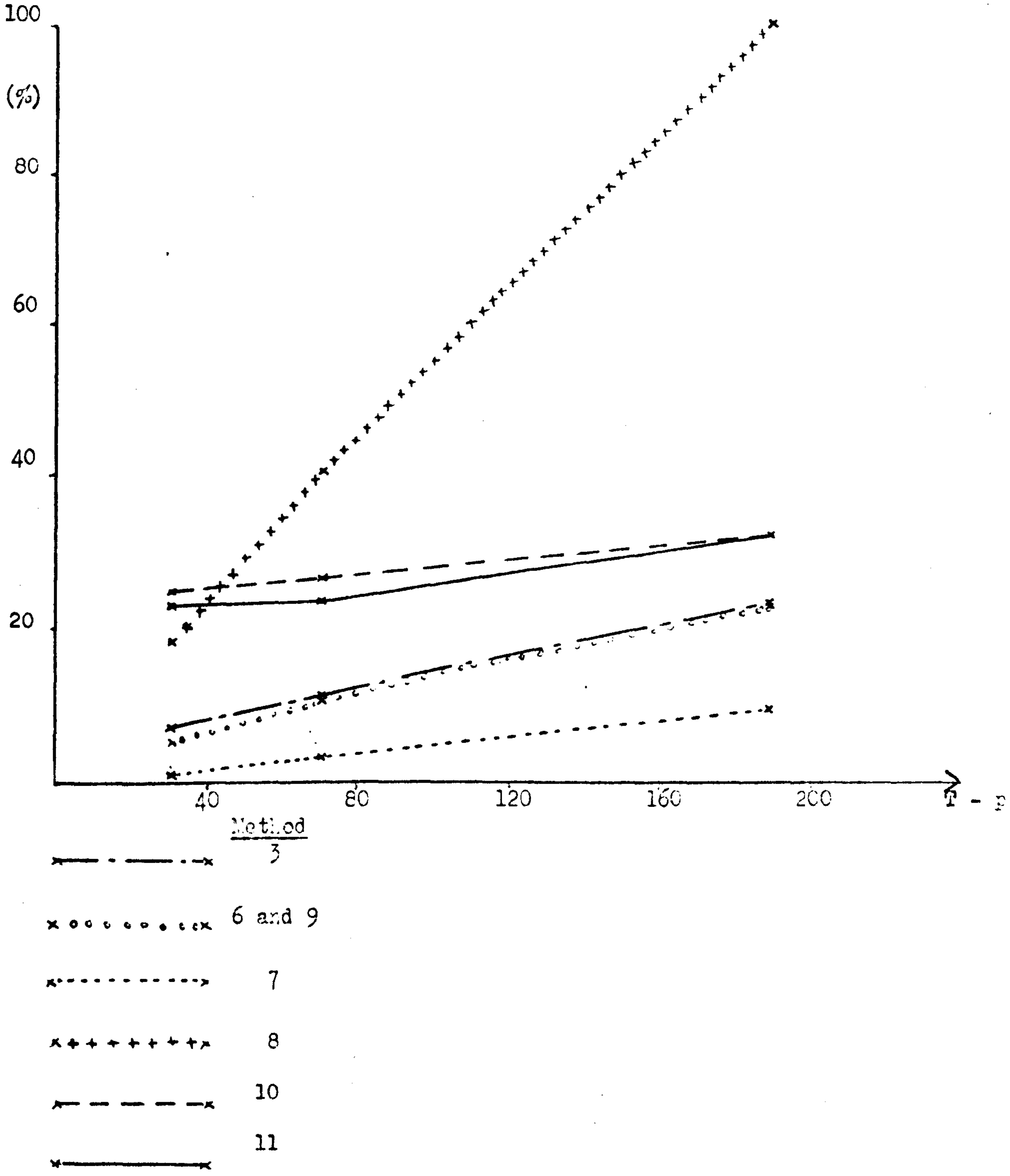


Figure 10.3 : Prediction Efficiencies for case (c)

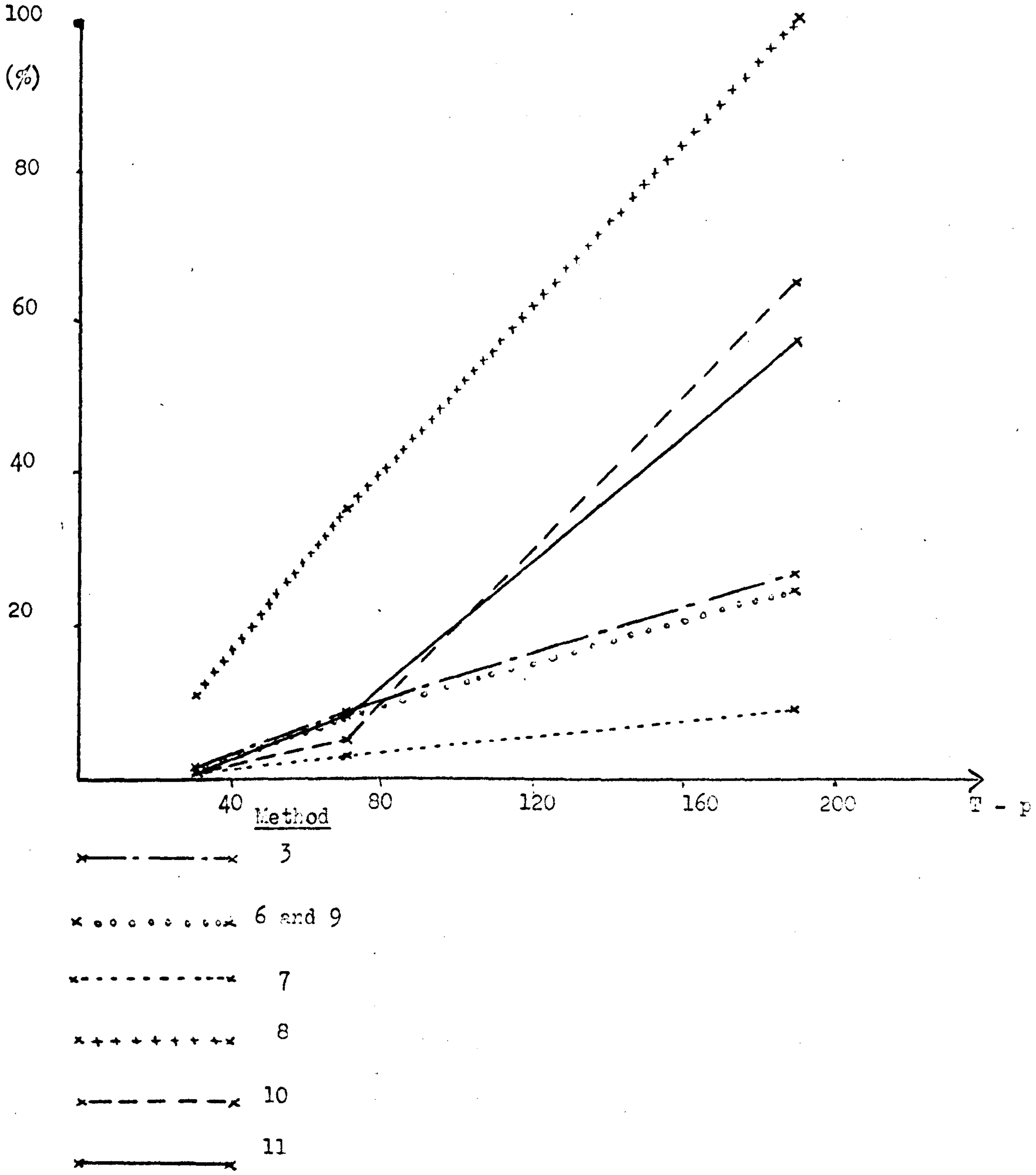


Figure 10.4 : Prediction Efficiencies for case (d)

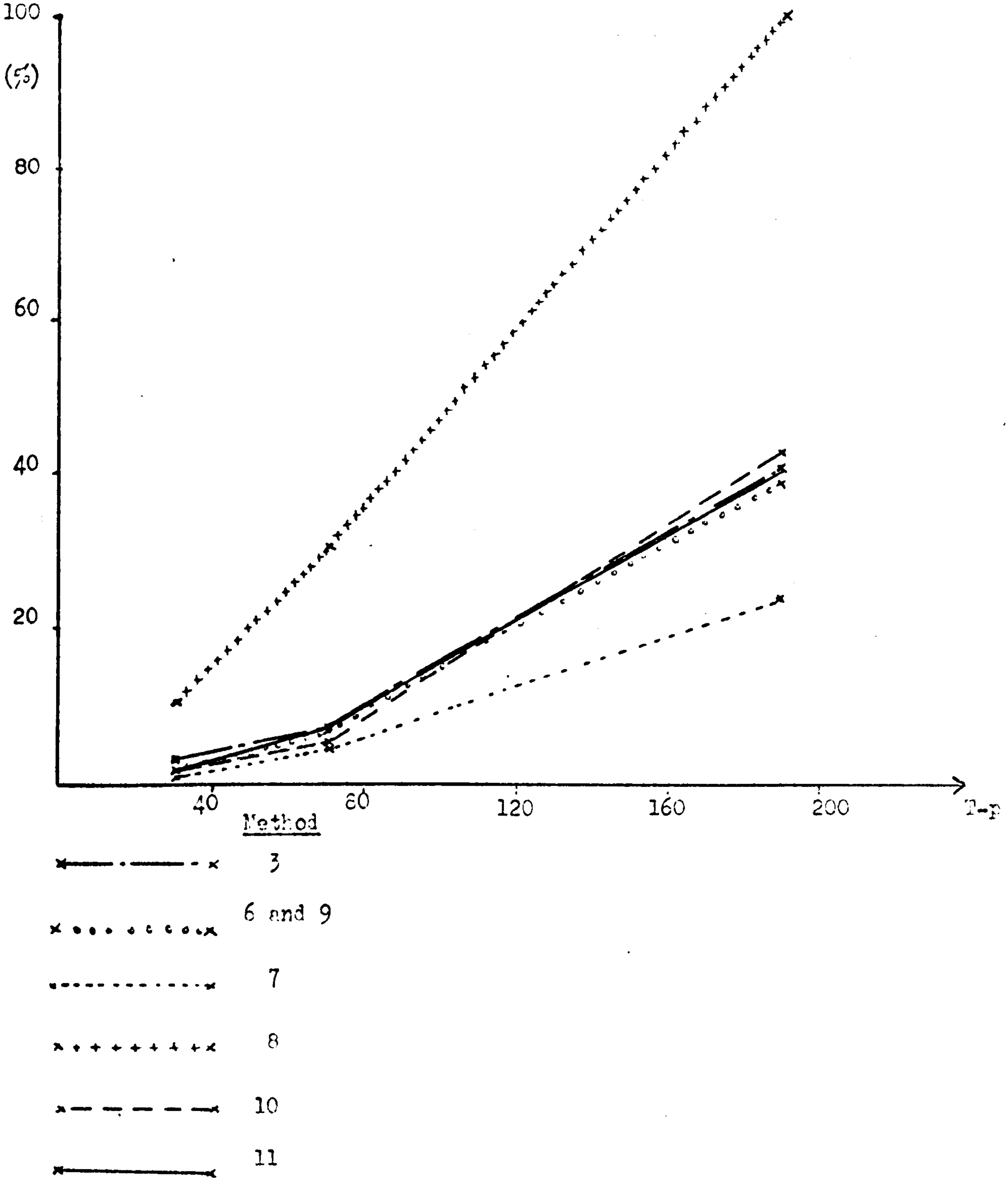
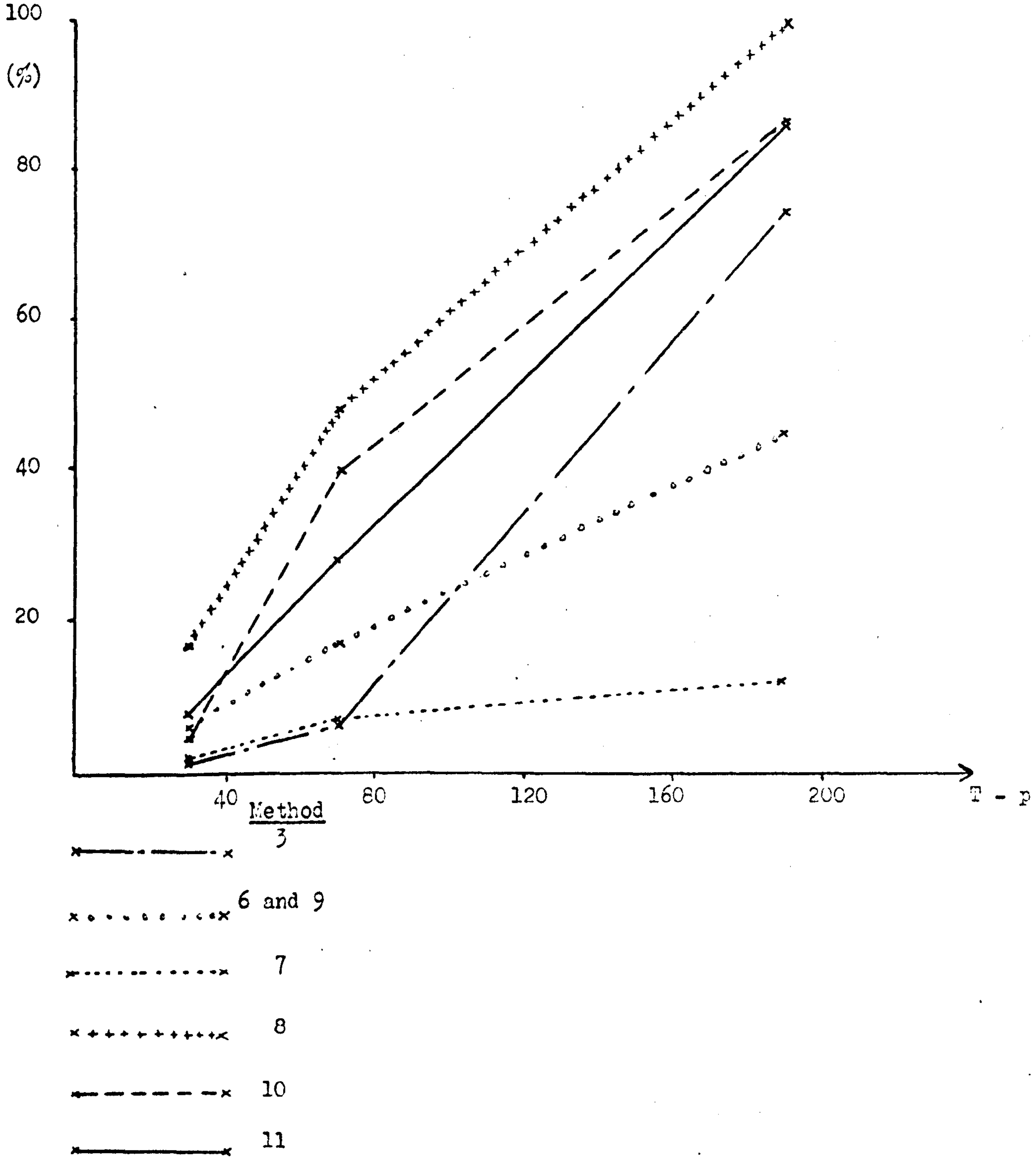


Figure 10.5 : Prediction Efficiencies for case (e)



10.4 Conclusions

Starting with an overall view we see that, with the particular models investigated, the F' and FMAX procedures (methods 10 and 11) are almost always superior to the other three procedures based on conventional F. This statement is equally true in the two situations of identification and prediction. We also note that, similar to the classical non-orthogonal case studied in chapter 9, the backward and forward/backward versions based on conventional F (methods 6 and 9) again behave almost identically. Because, however, the different model specifications individually reveal features of special interest each situation will now be looked at case by case.

Looking first at the white noise model, case (a), we see as expected that F' and FMAX are far superior in performance to the other three procedures. Bearing in mind also the lack of theoretical support for least squares in smaller size samples the performances in these particular cases is especially encouraging. On the other hand the results for the conventional F approaches are very disturbing in that they are increasingly likely to pick up a spurious autoregressive structure.

In case (b), which represents a very slight departure from a white noise process, the F' and FMAX procedures only begin to detect the proper structure in the large sample situation. Nevertheless the other three procedures can be said to be no more successful in the smaller sample cases insofar as they only succeed in producing a large proportion of mixed cases. Again this seems to be a consequence of the usual overfitting tendency of these procedures. Again it is noticeable from Fig. 10.2 that the underfitting bias of the F' and FMAX gives rise to a marginal advantage even over method 8 in terms of prediction efficiency, this being a result comparable to that obtained in the non-orthogonal regression situation in Fig. 9.5.

Case (c) which is represented by Table 10.3, again relates to a first-order Markov process but this time with a stronger regressor relationship. The results obtained here are extremely consistent with theoretical expectations of the asymptotic performances of F' and FMAX. We note however, that, for the smaller sample sizes, these two procedures diverge in performance with FMAX seeming to be slightly better on a score basis. This perhaps indicates that the situation calls for F' to be tested at a higher level than 5%, but this of course being wise after the event. If a higher level were to be used in the situation of Table 10.1, for example, the consequence then would be a higher propensity to overfit. The other three procedures based on conventional F now definitely begin to exhibit serious overfitting characteristics, this again conforming to our expectations. Again the predictive performances indicated in Fig.10.3 are undoubtedly in favour of methods 10 and 11.

Turning now to Table 10.4 we see a similar pattern to that in Table 10.2. For although methods 10 and 11 show only a relatively slow reduction in underfitting as T increases, the other three procedures only suceed in generating more mixed cases in trying to overcome this effect. From a prediction viewpoint the outcomes of these two alternative effects seem, from. 10.4, to be roughly the same.

Finally, boking at Table 10.5, a very prominent feature is the weak performance of the forward F procedure (method 3) as compared with the two other conventional F procedures of methods 6 and 9. The explanation for this lies in the form of the underlying autocorrelation function for Y_t. For, on solving the first three Yule-Walker equations, we find

$$\rho_1 = 0.44$$
$$\rho_2 = -0.44$$
$$\rho_3 = -0.7$$

Since stepwise procedures introduce lags according to the squares of these cor elation coefficients it follows that lag 3 will usually be the first lag to enter. Procedures which have a backward deletion facility, particularly if starting from a fully fitted initial equation, will avoid the consequences of this since the partial autocorrelation coefficients for <u>any</u> lag at least as great as 3 (holding fixed Y_{t-1} and Y_{t-2}) will be zero. This feature is substantiated by the performances of all of the other methods investigated. Apart from this aspect there are no other points to comment on here except that F' and FMAX again demonstrate a superiority over the other types of procedure.

We have seen then that, at least for the admittedly restricted group of models investigated here, the F' and FMAX procedures perform very favourably in comparison with the conventional F procedures. It would be dangerous however to extrapolate this apparent superiority to more general situations, especially those involving much more complicated lag structures. What can be said though is that the F' and FMAX approaches do seem to keep within the limits of what information is available, while the other methods are not restrained in this way. The question of whether one should prefer a routine which has a tendency to underfit rather than one which, while it gains on the underfitting criterion, does so at the expense of an increase in unwanted variables calls for an answer based on more subjective considerations. What can be said is that on the objective criterion of prediction performance it would appear that the former procedure characteristic is the more desirable one.

10.5 Some comparisons with other approaches

In this concluding section of the present chapter the performance of the stepwise regression technique will be compared with that of other approaches which have been tried by other people in three particular instances.

(a) Analysis of sunspot data

Data on sunspot intensity has been subjected to several statistical investigations over the years since Schuster [73], using periodogram analysis, detected a periodicity of 11.125 years. The series is in fact particularly appealing for analysis since it comprises a very long data record from what can be regarded as a fairly stable generating mechanism. The idea of a strict periodicity existing in the process, as implied by Schuster's approach, is however not entirely consistent with observed data plots. For while the series certainly does fluctuate with peaks occurring approximately every eleven years there is a noticeable variation about this frequency together with changes in the amplitudes attained by different cycles. Such a phenomenon seems therefore to imply a stationary but not strictly deterministic underlying structure of the type associated with autoregressive schemes, for example. For this reason several attempts have been made to fit such models using both time and frequency domain techniques, and also employing various types of data series. One such analysis is that due to Whittle [76] who was particularly concerned with testing a hypothesis postulated by H. Alfvén. This is the study which will be followed up here.

Whittle's main concern was in deciding which of two competing autoregressive representations most adequately describes the observed process. On the one hand there is the second-order model fitted by Yule [78] which yielded a peak in the spectrum at a frequency corresponding to a period of about 10.6 years, such a process also implying what Yule described as a 'disturbed pendulum' effect. On the other hand however is a model incorporating an eleven year lag term, this relationship in turn being in line with Alfvén's 'reflection' theory for sunspot activity. Using a series of data collected for specific solar latitudes every six months over the period 1886 to 1945 Whittle proceeded to fit a model of the form

$$Y_t = \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \alpha_{22} Y_{t-22} + \alpha_{44} Y_{t-44} + \varepsilon_t \qquad (1)^{\dagger}$$

† We note that Whittle used the Yule-Walker equations in obtaining these estimates, these being asymptotically equivalent to the least squares estimates. The overall mean was first subtracted from the data in this and the later investigations.

Using a test derived in Whittle [75] he then finds that α_{22} is significant in (1) while α_2 and α_{44} are not, thus ending up with a model containing lags 1 and 22 (and thereby supporting Alfvén's theory).

Since the above problem can be regarded as falling into the identification aspect of stepwise regression it seems a good idea to apply the general forward/backward procedure (method 10) which came out well in the simulation studies i.e. the procedure based on F'. Before doing so however we note that a stepwise-type procedure has been applied to this data by Schaerf [70]. Her approach has a stopping rule based on the partial autocorrelation coefficients between the excluded variables and the regressand, and invokes the asymptotic independence property between these estimates proved earlier in her paper, Schaerf then proceeds to apply the stepwise principle in a strictly forward manner and obtains, to her admitted surprise, the model:

$$Y_t = 0,647Y_{t-1} - 0.280Y_{t-9} \qquad (2)$$

Since it seems very plausible that the lag of order 9 here arises spuriously out of the underlying auto-correlation structure the application of method 10 of this paper should be able to overcome this. Using a maximum lag of 25 this method was duly applied and resulted in the equation:

$$Y_t = 0.632Y_{t-1} + 0.336Y_{t-20} \qquad (3)$$

Although not actually producing a lag of 22 the result obtained is certainly much more appealing than Schaerf's model at (2). In addition, the residual variances of the three models (1), (2) and (3) (each now recalculated on the same effective sample of size 95 using ordinary least squares) turned out to be:-

(1) 3.08×10^6

(2) 3.145×10^6

(3) 3.05×10^6

Again what evidence there is here supports the approximate eleven year dependence structure of the process. Thus, while all three models possess remarkably similar associated transfer functions, all having pronounced peaks around the eleven year frequency band, the F' stepwise procedure has selected the model which is most in agreement with conjectured theory.

(b) A further prediction study

In a research report based on a Ph.D. thesis ([12]) Bhansali [13] presents a Monte Carlo comparison of the prediction performances of autoregressive models fitted by, respectively, frequency and time domain techniques. Using two variations of a method which employs the Fourier inversion of the log-spectrum Bhansali sets out to demonstrate that, at least in situations in which the true order p of an autoregression is unknown, better predictions can thus be obtained than by using standard regression methods. This he succeeds in doing, albeit for a sample size T of 1000, with the aid of three selected model specifications. A crucial point in his argument however is his dismissal of the stepwise regression approach to model identification (apparently on the grounds of imprecise stopping criteria) in favour of the following procedure proposed by Akaike [1]. Akaike's method is based on the result that the one-step asymptotic mean square error of prediction using an estimated p^th order autoregressive model is given by

$$\sigma^2 (1+\frac{p}{T}) \qquad (4)$$

where T is the sample size and σ^2 is the (unknown) error variance. Deciding on a value of L (which corresponds to the use of k in stepwise regression) Akaike suggests that one then calculates L autoregressive equations having successively increasing lagged terms, each time recording the residual variance estimate

$$\hat{\sigma}^2 p = (T-p)^{-1}\left[\sum_{t=p+1}^{T}\left(Y_t - \sum_{j=1}^{p} a_j Y_{t-j}\right)^2\right].$$

$$(p = 1,2,\ldots\ldots,L).$$

The value of p, p_o say, which on substituting $\hat{\sigma}^2 p$ in (4) gives an overall minimum is then taken as the true order of autoregression.

Two particular comments can be made concerning the above procedure. Firstly, it ignores completely the possibility that some intermediate lags of order less than p might be unnecessary. Secondly, as is demonstrated by Bhansali's results, the approach is exceedingly likely to produce serious overfitting†. Though Bhansali presents a very detailed description of his investigations we shall only be concerned here with comparing the performances of the F' and conventional F procedures (methods 10 and 9 respectively) with three of the procedures looked at by Bhansali. The particular procedures concentrated on are those referred to by Bhansali as

(i) The Regression - Akaike method (R.A)

(ii) The Suggested - Akaike method (S.A)

(iii) The Jones-Akaike method (J.A).

The first of these involves the standard use of Akaike's approach in conjunction with the estimates from the Yule-Walker equations. For details of the other two (spectral) approaches the reader is referred to Bhansali's original paper.

†A model with true order p = 2, for example, produced (using 100 iterations) an average fitted order of 10.6 using L = 25

The three model structures investigated by Bhansali are as follows:-

Experiment I

$$Y_t = 0.55Y_{t-1} + 0.05Y_{t-2} + \varepsilon_t$$

Experiment II

$$Y_t = 0.5\varepsilon_{t-1} + 0.5\varepsilon_{t-2} + \varepsilon_t$$

Experiment III

$$Y_t = 0.5Y_{t-1} - 0.06Y_{t-2} + 0.45Y_{t-15} + \varepsilon_t$$

Experiment II here is especially interesting since it corresponds to an infinite order autoregressive model. For each of these three cases Bhansali generated samples of size 1100, the first 1000 observations being used for estimation purposes and the last 100 serving as a prediction set. In all cases the residual variance was taken as unity. As well as the three estimation procedures referred to above (R.A., S.A. and J.A.) estimates were also obtained for the true order equations in the case of Experiments I and III. We shall refer to these non-identification procedures as merely R, S and J.

Since the above described field of application obviously lends itself to a stepwise regression approach it was decided to replicate Bhansali's experiments as closely as possible, but this time using the F' and conventional F stepwise procedures. Thus proceeding exactly as is described in Bhansali's report mean square prediction errors were obtained for one, two and three-step ahead forecasts for each of the three model specifications. A maximum lag k equal to 25 was chosen in all cases, corresponding to the value of L used in the Akaike approach. Table 10.6 below presents the results obtained. Tables 10.7 and 10.8 are reproduced from Bhansali's report for comparison purposes. Table 10.8 also includes the expected mean square errors as given in Bhansali [13] for the case where knowledge of p is used in the estimation. Apart from such expected values all figures are averages over 100 iterations.

Table 10.6: Mean square error of prediction using the F' and conventional F procedures

Step ahead	Experiment I		Experiment II		Experiment III	
	F'	Conv.F	F'	Conv.F	F'	Conv.F
1	0.99	1.00	1.03	1.03	1.01	1.01
2	1.30	1.30	1.28	1.28	1.25	1.26
3	1.42	1.43	1.55	1.56	1.28	1.29

Table 10.7: Mean square error of prediction using Akaike's identification procedure with the R, S and J methods

Step ahead	Experiment I			Experiment II			Experiment III		
	R.A.	S.A.	J.A.	R.A.	S.A.	J.A.	R.A.	S.A.	J.A.
1	1.17	0.90	0.89	0.91	0.93	0.91	1.05	1.01	0.97
2	1.65	1.60	1.53	2.05	1.66	1.47	2.36	1.47	1.52
3	2.23	2.69	2.41	2.98	2.21	2.21	2.86	1.53	1.54

Table 10.8: Mean square error of prediction by fitting the true order model

Step ahead	EXPERIMENT I				EXPERIMENT III			
	R	S	J	Expected	R	S	J	Expected
1	0.88	0.90	0.89	1.002	0.96	1.02	0.99	1.015
2	1.51	1.50	1.50	1.31	1.51	1.50	1.51	1.27
3	1.58	2.07	2.06	1.43	1.52	1.51	1.53	1.31

A particularly striking feature of Bhansali's results which calls for comment is the degree of divergence in Table 10.8 for the case of Experiment I between the observed and expected prediction variances. For some reason Bhansali chooses to comment only on the three-step ahead cases where he states that the observed differences are significant at the one percent level using the chi-square distribution. By way of explanation he suggests that the theoretical expression for the expected value 'may tend to underestimate' the actual value for small values of p. The evidence of the results obtained in Table 10.6 for the present study does however seem to rule out this possibility, there being an almost exact agreement between observed and expected values. In any case an inspection of Bhansali's results in the one and two-step ahead cases of Table 10.8 again reveals a significant departure of observed from expected. Bearing in mind that in each case the figures for R, S and J were obtained independently of each other one is led to conjecture that something is amiss in Bhansali's simulation program.

The point which can be made here regarding the results for the stepwise approaches is that they agree very closely with the optimum values obtainable. If we ignore the suspiciously low values obtained by Bhansali in the one-step ahead cases we see also that the rate of increase of prediction variance (as a function of the step-ahead forecast) is much lower in the case of Table 10.6 than in either of Tables 10.7 or 10.8. From this one may still draw the tentative conclusion that a stepwise identification procedure is much preferable to the Akaike approach (at least in the models investigated).

As a final remark here it must be stated that the results obtained for the stepwise application do not tell us anything relevant to the question of whether the spectral approach is preferable to the regression one for such a large sample size. In order to be in a position to answer this one would need to use the spectral estimates corresponding to the stepwise identified model. It is felt however that the obvious tendency of Akaike's approach to produce overfitted equations, and its associated inclusion of unnecessary lags of order lower than the true order of equation p, must weight more heavily against the regression approach than the spectral one. For the problem of unstable estimates due to the underlying multicollinearity of regressors is known to be more predominant in time domain approaches to estimation rather than frequency domain alternatives.

(c) A comparison between stepwise regression and the Box-Jenkins approach to forecasting

In an extensive empirical study into a number of univariate forecasting techniques Granger and Newbold [29] make comparisons between several competing approaches requiring various degrees of sophistication in their operation. One such procedure which was looked at and observed to perform quite well, especially when used in combination with other methods, was that of stepwise regression. It is the intention in this section to investigate whether the particular stepwise version which was used in the above-mentioned study is in fact capable of any improvement. We shall do so by again comparing, as do Granger and Newbold, the stepwise predictive performance against the yardstick of the Box-Jenkins approach. No attempt will be made here to give specific details of the 106 sets of real data which were used in the original study, nor will details be given as to how the samples were divided into the two separate parts required for the respective purposes of fitting and forecasting.

A full discussion on all these aspects is in any case, to be found in the original source. It suffices to say that the samples used were from a wide field of seasonal and non-seasonal series at both micro- and macro-economic levels.

In both the Box-Jenkins and stepwise applications of the original study the data were all initially first-differenced to eliminate the presence of any non-stationary random walk type of behaviour such as often occurs with economic series. In the case of the Box-Jenkins procedure a further differencing operation was often carried out when a seasonal component was evident from the data plot. As far as the Box-Jenkins approach is concerned a standard type of analysis was followed by Granger and Newbold, as much as is allowed by the subjective element of course. In particular a constant term was only very occasionally included in the general ARIMA model being entertained, such as a decision being made entirely on the basis of first-difference data plots over the fitting period. In the case of the stepwise regression procedure, which was in fact a strict forward approach based on the conventional F criterion, any seasonality which might be present was left in the data for the possible detection by the procedure itself of the appropriate lagged terms.

The most surprising aspect of the procedure used, apart from its strict forward orientation, was however that a constant term was fitted on all but a few occasions. Since such terms were not usually thought to be called for in the Box-Jenkins applications (nor indeed do Box and Jenkins expect the presence of such terms in general - see [14, p.93]) it seemed likely that the inclusion of such terms might be a disadvantage on two possible counts. For, firstly, one has to take into account the stability of such a term when estimated from only a moderate sized (non-random) time series sample. And secondly, and perhaps even more importantly, the automatic inclusion of a constant is tantamount to a considerable assumption being made as to the existence of a deterministic trend structure in the original undifferenced series.

In view of the doubts expressed above concerning the stepwise approach used in the Granger/Newbold study it was decided to re-run the investigations using instead now the stepwise approach based on F' (method 10) and, in addition, only incorporating a constant term whenever this was thought to be necessary for the Box-Jenkins procedure. Thus the 104 series which were still readily available from the original set of 106 were re-run with this alternative stepwise program, and the corresponding one-step ahead forecast mean square-errors were duly obtained. We can thus make a direct comparison of these results with those for the Box-Jenkins forecasts, and in this way we can, at the same time, see if an improvement has been made on the original stepwise results. We find in fact that the number of occasions on which Box-Jenkins has smaller forecast variance than stepwise now falls from 70 to 63 out of the 104 cases being considered. Further, the (geometric) mean of the ratios of mean square forecast errors of Box-Jenkins relative to stepwise now increases from 0.86 to 0.90. The increase is particularly high in the case of the 25 quarterly series which were investigated, the mean ratio here changing from 0.87 to 0.94. It would however be rash to try to generalize on the basis of this result which is based on a fairly small number of cases.

The main point then that arises out of these forecast procedure evaluations is really a point which is made in the Granger/Newbold study i.e. that stepwise regression can do exceedingly well even in comparison with a relatively more ambitious approach such as that of Box and Jenkins. The above investigation does however demonstrate that it is still possible to achieve albeit marginal improvements using a stepwise procedure based on something more than just an ad hoc argument.

Chapter 11 Summary and Conclusions

11.1 A general review

It is at this juncture that one has to stand back and assess what has been achieved in terms of the original objective, namely that of investigating whether one can fruitfully impose a formal structure on what has so far tended to be regarded as a rather 'rule-of-thumb' type of procedure. In common with several other techniques which have undergone rapid escalations in popularity over recent years stepwise regression can be said to have created a sizeable foothold in the armoury of the applied statistician without there having occurred a comparable growth in the understanding of its theoretical basis. As often happens when one strives to establish a plausible formal framework for such procedures, this particular study has encountered many unforeseen complexities which have sometimes made it necessary to resort to tactics involving a reduction in the desired level of rigour. A prime example of this has been the recourse which has had to be made to a simulation approach in demonstrating some facets of the theory which has been presented. Despite these problems, however,it is felt that the exercise has been useful in indicating both the capabilities and limitations of the whole concept of stepwise regression.

By looking at situations of an increasing degree of complexity, from the relatively simple situation of orthogonality to the considerably more demanding and ambitious applications in time series analysis, it is felt that the theoretically viable approaches of the F' and FMAX procedures in particular have been shown to perform in accordance with the expectations of that theory. On the other hand the conventional use of the stepwise regression technique, using what has been referred to throughout this thesis as the conventional F criterion, has been shown to be inconsistent with any reasonable underlying theoretical basis. While it is not of course an essential prerequisite that any statistical technique must behave well in accordance with some idealised formal structure, only that it should in fact provide fairly consistent answers of a useful kind, it is nonetheless of interest to be able to judge how far such answers might depart from reality. It is in this sense then that the F' and, to a lesser extent, the FMAX procedures are considered to be superior to the other approaches. Whether or not other factors enter into a particular proposed application, such as considerations of computational cost for example, is something else which has to be taken into account of course.

11.2 Possible areas of further investigation

In this final section of the discussion one must raise the natural question as to what remains to be done by way of extending the results so far obtained. The short answer to this is that a very considerable amount remains to be done, not only in stepwise regression but also in the whole field of similar techniques which contain an element of data-induced model formulation. As far as stepwise regression is concerned some possible generalisations have already been alluded to previously. Foremost of these is perhaps the possibility of using varying significance levels in accordance with such prior information which might be available regarding the underlying model. In such a situation involving a partially specified model it might however still be preferable to treat stepwise regression entirely as an automatic objective procedure, only invoking subjective information at the end stage in the light of the regression sequences which have been obtained.

Apart from the above considerations there is however still the unsettled problem, even if one decides to use a fixed significance level throughout, as to what this level should be. For while the choice of level has been demonstrated (at least in the cases of F' and FMAX) to provide an asymptotic controlling effect on the degree of overfitting obtained the finite sample size case presents some difficulties. For although the empirical studies showed that a 5% level was fairly reasonable for models containing only a small number of non-zero regression coefficients, considerable underfitting could occur in more complicated situations. Thus a lower significance level might possibly have led, in such instances, to an improvement in the selection performance. However, particularly in situations exhibiting non-orthogonality, one needs to be careful not to merely be substituting an underfitted model by one which is just what has been termed previously as a mixed case. Such problems do however admit to an extended investigation of the type that has already been carried out.

Another more specific area which calls for a closer study is that which was briefly entered in part (c) of (10.5), namely that of obtaining forecasting models for economic time series. Now it must be stated first of all that a great attraction of the stepwise approach, like that of Box and Jenkins for example, is its easy and almost automatic mode of application. With this in mind though it was seen that even a simple qualitatively determined adjustment of not fitting a constant term in such models could lead to a measurable improvement in forecast performance. It might therefore be possible to find other qualitative elements, possibly depending on some specific characteristics of the type of series being investigated, which leads to further improvements. While in the context of forecasting we might just remark that, in principal, there is no reason why one should not comtemplate using stepwise regression for models incorporating lagged terms of other potentially useful series as well as those of the regressand, i.e. general distributed lag models. An incentive for performing such investigations is that, in such a multivariate situation, there are fewer existing competitive procedures than in the univariate case.

Finally, a mention should be made regarding the possibility of using an a priori othogonalizing transformation on the regressrr matrix $\underset{\sim}{X}$ in order to overcome some of the extra difficulties encountered in the non-orthogonal situation. Thus we can contemplate re-writing the model at (1.2.1) in the form

$$\underset{\sim}{Y} = (\underset{\sim}{XT})\ \underset{\sim}{T}^{-1}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

where $\underset{\sim}{T}$ is a square non-singular matrix of dimension k chosen such that $\underset{\sim}{T}'\underset{\sim}{X}'\underset{\sim}{XT}$ is strictly diagonal. Such an approach seems to have first been put forward by Kendall [36], who suggested that one chose the particular transformation matrix used in obtaining the principal components of the regressor variables in $\underset{\sim}{X}$. In fact stepwise regression itself can be shown to correspond to using an upper triangular matrix $\underset{\sim}{T}$ to ultimately orthogonalize $\underset{\sim}{X}$, but such a transformation is of course only induced in an a posteriori manner by the sequence of decisions made. It must of course be noted that the possibility of using the above approach at all really only arises in the prediction context. For, denoting the new regressor matrix $\underset{\sim}{X}\ \underset{\sim}{T}$ by $\underset{\sim}{U}$, the set of regressors

$$U_1\ ,\ U_2\ , \ldots ,\ U_{k'},\qquad (k' \le k)$$

which is subsequently selected will in general transform back into an equation involving all k original X variables. This does seem to rule out such a procedure for the purpose of model identification as was specified earlier in this study.

Concentrating therefore on the prediction aspect alone one must question the relevance of using the principal component transformation in preference to any other orthogonalizing matrix. For although some authors (e.g. Wickens and Ord [63], Daling and Tamura [17]) conjecture that one can safely omit orthogonalised regressors having small variances there really seems to be no a priori grounds for making such an assumption. Neither in fact has there been any real evidence presented in support of this theory. The point is however not of direct relevance in any case to stepwise regression since such a procedure is itself designed to detect which is the set of significant regressors. Indeed the use of principal component regressors might in fact be said to be desirable since it provides a readily available method for producing such an initial set of orthogonal variables. This again is something which calls for further investigation.

Appendix 1

The proofs are given here of Theorems 1.4.2. and 1.4.3. Without loss of generality we can take c to be q and q + 1 respectively in the two theorems, in which case the following partition of A can be used throughout

$$\underset{\sim}{A} = \begin{bmatrix} \underset{\sim}{a}_{11} & \underset{\sim}{a}_{12} & \underset{\sim}{a}_{13} \\ \underset{\sim}{a}_{21} & a_{22} & \underset{\sim}{a}_{23} \\ \underset{\sim}{a}_{31} & \underset{\sim}{a}_{32} & \underset{\sim}{a}_{33} \end{bmatrix}$$

where a_{22} is a scalar. Suppose the stepwise pivotal operation is performed with a_{22} as pivotal element. This will transform $\underset{\sim}{A}$ into $\underset{\sim}{A}^*$ where, with the same partitioning structure,

$$\underset{\sim}{a}^*_{11} = \underset{\sim}{a}_{11} - a_{22}^{-1}\ \underset{\sim}{a}_{12}\ \underset{\sim}{a}_{21}$$

$$\underset{\sim}{a}^*_{12} = -a_{22}^{-1}\ \underset{\sim}{a}_{12}$$

$$\underset{\sim}{a}^*_{13} = \underset{\sim}{a}_{13} - a_{22}^{-1}\ \underset{\sim}{a}_{12}\ \underset{\sim}{a}_{23}$$

$$\underset{\sim}{a}^*_{21} = a_{22}^{-1}\ \underset{\sim}{a}_{21}$$

$$\underset{\sim}{a}^*_{22} = a_{22}^{-1}$$

$$\underset{\sim}{a}^*_{23} = a_{22}^{-1}\ \underset{\sim}{a}_{23}$$

$$\underset{\sim}{a}^*_{31} = \underset{\sim}{a}_{31} - a_{22}^{-1}\ \underset{\sim}{a}_{32}\ \underset{\sim}{a}_{21}$$

$$\underset{\sim}{a}^*_{32} = -a_{22}^{-1}\ \underset{\sim}{a}_{32}$$

$$\underset{\sim}{a}^*_{33} = \underset{\sim}{a}_{33} - a_{22}^{-1}\ \underset{\sim}{a}_{32}\ \underset{\sim}{a}_{23}$$

Consider Theorem 1.4.2.

We have to show:-

$$\text{(i)} \quad \underset{\sim}{a}^*_{11} = (\underset{\sim}{Z}^{*\prime}_1 \underset{\sim}{Z}^*_1)^{-1}$$

$$\text{(ii)} \quad [a^*_{12} \ \vdots \ a^*_{13}] = (Z^{*\prime}_1 Z^*_1)^{-1} Z^{*\prime}_1 Z^*_2$$

$$\text{(iii)} \quad \begin{bmatrix} a^*_{22} & \underset{\sim}{a}^*_{23} \\ a^*_{32} & a^*_{33} \end{bmatrix} = \underset{\sim}{Z}^{*\prime}_2 (\underset{\sim}{I} - \underset{\sim}{Z}^*_1 (\underset{\sim}{Z}^{*\prime}_1 \underset{\sim}{Z}^*_1)^{-1} \underset{\sim}{Z}^{*\prime}_1) \underset{\sim}{Z}^*_2$$

(The results for $\underset{\sim}{a}^*_{21}$ and $\underset{\sim}{a}^*_{31}$ will then follow immediately from symmetry considerations.)

(i) Consider $\underset{\sim}{a}^*_{11} = \underset{\sim}{a}_{11} - a_{22}^{-1}\ \underset{\sim}{a}_{12}\ \underset{\sim}{a}_{21}$.

$$\text{Now } \underset{\sim}{A}_{11} = \begin{bmatrix} \underset{\sim}{a}_{11} & \underset{\sim}{a}_{12} \\ \underset{\sim}{a}_{21} & a_{22} \end{bmatrix} = (\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}$$

Writing $\underset{\sim}{Z}_1 = [\underset{\sim}{Z}_1^* \vdots \underset{\sim}{X}_q]$ and using Theorem 1.4.1. we find that

$$\underset{\sim}{a}_{11} = (\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{Z}_1^*)^{-1}(\underset{\sim}{I}+d^{-1}\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{X}_q\underset{\sim}{X}_q'\underset{\sim}{Z}_1^*(\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{Z}_1^*)^{-1})$$

where $d = \underset{\sim}{X}_q'(\underset{\sim}{I}-\underset{\sim}{Z}_1^*(\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{Z}_1^*)^{-1}\underset{\sim}{Z}_1^{*\prime})\underset{\sim}{X}_q$ is a scalar,

$$\underset{\sim}{a}_{12} = -d^{-1}(\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{Z}_1^*)^{-1}\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{X}_q$$

and $a_{22} = d^{-1}$.

Since $\underset{\sim}{a}_{21} = \underset{\sim}{a}_{12}'$ it follows directly that $\underset{\sim}{a}^*_{11} = (\underset{\sim}{Z}_1^{*\prime}Z_1^*)^{-1}$.

(ii) First note that, using the above results,

$$\underset{\sim}{a}^*_{12} = -a_{22}^{-1}\ \underset{\sim}{a}_{12} = (\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{Z}_1^*)^{-1}\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{X}_q.$$

Hence we must show that

$$\underset{\sim}{a}^*_{13} = (\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{Z}_1^*)^{-1}\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{Z}_2.$$

$$\text{Now } \underset{\sim}{a}^*_{13} = \underset{\sim}{a}_{13} - a_{22}^{-1}\ \underset{\sim}{a}_{12}\ \underset{\sim}{a}_{23}.$$

Since $(\underset{\sim}{Z}_1'\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}_1'\underset{\sim}{Z}_2 = \underset{\sim}{A}_{12} = \begin{bmatrix} \underset{\sim}{a}_{13} \\ \underset{\sim}{a}_{23} \end{bmatrix}$ it follows that

$$\underset{\sim}{a}_{13} = (\underset{\sim}{a}_{11}\underset{\sim}{Z}_1^{*\prime}+\underset{\sim}{a}_{12}\underset{\sim}{X}_q')\underset{\sim}{Z}_2$$

$$\underset{\sim}{a}_{23} = (\underset{\sim}{a}_{21}\underset{\sim}{Z}_1^{*\prime}+a_{22}\underset{\sim}{X}_q')\underset{\sim}{Z}_2$$

Therefore $\underset{\sim}{a}_{13} = \underset{\sim}{a}_{11}\ \underset{\sim}{Z}_1^{*\prime}\underset{\sim}{Z}_2 + \underset{\sim}{a}_{12}\underset{\sim}{X}_q'\underset{\sim}{Z}_2 - d\underset{\sim}{a}_{12}\underset{\sim}{a}_{21}\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{Z}_2 - d\underset{\sim}{a}_{12}a_{22}\underset{\sim}{X}_q'\underset{\sim}{Z}_2$

which, using the results obtained in (i), reduces to

$$(\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{Z}_1^*)^{-1}\underset{\sim}{Z}_1^{*\prime}\underset{\sim}{Z}_2$$

(iii) First consider $a^{*}_{22} = a^{-1}_{22} = d = \underset{\sim}{X}'_q(\underset{\sim}{I}-\underset{\sim}{Z}^{*}_1(\underset{\sim}{Z}^{*\prime}_1\underset{\sim}{Z}^{*}_1)^{-1}\underset{\sim}{Z}^{*\prime}_1)\underset{\sim}{X}_q$

Also, $\underset{\sim}{a}_{23} = (\underset{\sim}{a}_{21}\underset{\sim}{Z}^{*\prime}_1+a_{22}\underset{\sim}{X}'_q)\underset{\sim}{Z}_2$

$$= d^{-1}\underset{\sim}{X}'_q(\underset{\sim}{I}-\underset{\sim}{Z}^{*}_1(\underset{\sim}{Z}^{*\prime}_1\underset{\sim}{Z}^{*}_1)^{-1}\underset{\sim}{Z}^{*\prime}_1)\underset{\sim}{Z}_2$$

and therefore $\underset{\sim}{a}^{*}_{23} = a^{-1}_{22}\ \underset{\sim}{a}_{23}$

$$= \underset{\sim}{X}'_q(\underset{\sim}{I}-\underset{\sim}{Z}^{*}_1(\underset{\sim}{Z}^{*\prime}_1\underset{\sim}{Z}^{*}_1)^{-1}\underset{\sim}{Z}^{*\prime}_1)\underset{\sim}{Z}^{*\prime}_2.$$

It only remains to show that

$$\underset{\sim}{a}^{*}_{33} = \underset{\sim}{Z}'_2(\underset{\sim}{I}-\underset{\sim}{Z}^{*}_1(\underset{\sim}{Z}^{*\prime}_1\underset{\sim}{Z}^{*}_1)^{-1}\underset{\sim}{Z}^{*\prime}_1)\underset{\sim}{Z}_2.$$

Now $\underset{\sim}{a}^{*}_{33} = \underset{\sim}{a}_{33} - a^{-1}_{22}\ \underset{\sim}{a}_{32}\ \underset{\sim}{a}_{23}$. But $\underset{\sim}{a}_{33} = \underset{\sim}{A}_{22} = \underset{\sim}{Z}'_2$ $(\underset{\sim}{I}-\underset{\sim}{Z}_1(\underset{\sim}{Z}'_1\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}'_1)\underset{\sim}{Z}_2$. Writing $\underset{\sim}{Z}_1 = \left[\underset{\sim}{Z}^{*}_1 \vdots \underset{\sim}{X}_q\right]$, using Theorem 1.4.1 and recalling that $(\underset{\sim}{Z}'_1\underset{\sim}{Z}_1)^{-1} = \underset{\sim}{A}_{11}$ we obtain

$\underset{\sim}{Z}_1(\underset{\sim}{Z}'_1\underset{\sim}{Z}_1)^{-1}\underset{\sim}{Z}'_1 = \underset{\sim}{Z}^{*}_1\ \underset{\sim}{a}_{11}\underset{\sim}{Z}^{*\prime}_1 + \underset{\sim}{Z}^{*}_1\ \underset{\sim}{a}_{12}\underset{\sim}{X}'_q + \underset{\sim}{X}_q\underset{\sim}{a}_{21}\underset{\sim}{Z}^{*\prime}_1 + \underset{\sim}{X}_q\ a_{22}\underset{\sim}{X}'_q.$

Also, $a^{-1}_{22}\ \underset{\sim}{a}_{32}\ \underset{\sim}{a}_{23} = -a^{-1}_{22}\ \underset{\sim}{Z}'_2(\underset{\sim}{Z}^{*}_1\underset{\sim}{a}'_{21}\underset{\sim}{a}_{21}\underset{\sim}{Z}^{*\prime}_1+\underset{\sim}{Z}^{*}_1\underset{\sim}{a}'_{21}a_{22}\underset{\sim}{X}'_q+\underset{\sim}{X}_q a_{22}\underset{\sim}{a}_{21}\underset{\sim}{Z}^{*\prime}_1$

$$+\underset{\sim}{X}_q\ a_{22}\underset{\sim}{X}'_q)\underset{\sim}{Z}_2.$$

On substituting for $\underset{\sim}{a}_{11}$, $\underset{\sim}{a}_{12}$ and a_{22} in terms of $\underset{\sim}{Z}^{*}_1$, $\underset{\sim}{X}_q$ and d a direct (but lengthy) calculation shows that

$$\underset{\sim}{a}^{*}_{33} = \underset{\sim}{Z}'_2(\underset{\sim}{I}-\underset{\sim}{Z}^{*}_1(\underset{\sim}{Z}^{*\prime}_1\underset{\sim}{Z}^{*}_1)^{-1}\underset{\sim}{Z}^{*\prime}_1)\underset{\sim}{Z}_2$$

Hence Theorem 1.4.2. is proved.

To prove Theorem 1.4.3. it is only necessary to re-pivot on a^{*}_{22} and show that $\underset{\sim}{A}^{*}$ is transformed back to $\underset{\sim}{A}$ (i.e. that the resulting transformation is in fact the inverse of the previous one. This is immediate e.g. $\underset{\sim}{a}^{**}_{11} = \underset{\sim}{a}^{*}_{11} - \underset{\sim}{a}^{*-1}_{22}\ \underset{\sim}{a}^{*}_{12}\ \underset{\sim}{a}^{*}_{21} = (\underset{\sim}{a}_{11} - a^{-1}_{22}\ \underset{\sim}{a}_{12}\ \underset{\sim}{a}_{21}) +$ $a_{22}(a^{-1}_{22}\ \underset{\sim}{a}_{12}\ a^{-1}_{22}\ \underset{\sim}{a}_{21}) = \underset{\sim}{a}_{11}$ etc.

Appendix 2.

Critical Values for Studentized Maximum Chi-Square Distribution

q	α=0.05	α=0.01	q	α=0.05	α=0.01
2	5.002	7.875	22	9.271	12.284
3	5.701	8.609	23	9.352	12.367
4	6.205	9.134	24	9.430	12.447
5	6.598	9.542	25	9.505	12.523
6	6.922	9.877	26	9.577	12.596
7	7.197	10.161	27	9.646	12.667
8	7.436	10.407	28	9.712	12.735
9	7.648	10.624	29	9.777	12.801
10	7.838	10.819	30	9.839	12.864
11	8.010	10.996	31	9.899	12.925
12	8.167	11.157	32	9.958	12.985
13	8.313	11.305	33	10.014	13.042
14	8.446	11.443	34	10.069	13.098
15	8.572	11.571	35	10.123	13.153
16	8.689	11.691	36	10.175	13.205
17	8.800	11.804	37	10.225	13.257
18	8.904	11.911	38	10.275	13.307
19	9.002	12.011	39	10.322	13.355
20	9.096	12.107	40	10.369	13.403
21	9.185	12.198			

Appendix 3

This appendix brings together the procedures used in various parts of the discussion.

Method 1: Strictly forward procedure using the FMAX stopping criterion.

Method 2: Strictly forward procedure using the F' stopping criterion.

Method 3: Strictly forward procedure using the conventional F criterion.

Method 4
Method 5
Method 6 } As 1, 2 and 3 but using a strict backward approach

Method 7: Estimation of the complete equation involving all k regressors.

Method 8: Estimation of the equation containing only the regressors having true non-zero coefficients.

Method 9: General forward/backward procedure using the conventional F criterion. The procedure structure is the same as for Method 10(below) except that, at each stage, the included variables are examined first for a possible deletion.

Method 10: General forward/backward procedure using the F' criterion. The procedure starts by fitting the complete equation (i.e. is backward orientated) and at each subsequent stage first tests for possible inclusion of an extra variable. If such a test is negative a variable deletion is contemplated. (For a full description see (8.5) and Fig.8.1).

Method 11: General forward/backward procedure using the FMAX criterion. In all other respects this procedure is identical to Method 10.

APPENDIX 4

This appendix shows, for the given initial regressor matrix $\underset{\sim}{X}$ as discussed in (9.1), the sequence of partial correlation matrices pertaining to the excluded variables at each stage of a forward stepwise application. (upper triangular sections only are recorded).

Stage 1

1.0	0.2	0.0	0.0	0.0	0.0	-0.2	0.3	-0.3	-0.3
	1.0	0.3	0.3	0.3	-0.3	-0.2	-0.1	0.1	0.3
		1.0	0.3	0.5	0.5	0.0	0.0	0.0	0.3
			1.0	0.3	-0.5	0.3	0.0	0.3	0.0
				1.0	0.0	0.3	0.0	0.0	0.0
					1.0	0.6	0.0	-0.3	0.3
						1.0	0.2	0.6	-0.6
							1.0	0.3	0.0
								1.0	0.5
									1.0

Stage 2

1.0	0.3	0.3	0.3	-0.3	-0.2	-0.2	0.7	0.3
	1.0	0.3	0.5	0.5	0.0	0.0	0.0	0.3
		1.0	0.3	-0.5	0.3	0.0	0.3	0.0
			1.0	0.3	0.0	0.0	0.0	0.0
				1.0	0.6	0.0	-0.3	0.3
					1.0	0.1	0.6	-0.7
						1.0	0.2	-0.1
							1.0	0.5
								1.0

Stage 3

1.0	0.2	0.7	0.4	-0.1	-0.1	0.0	0.4
	1.0	0.2	-0.4	0.4	0.1	0.2	-0.1
		1.0	0.1	0.4	0.1	0.0	-0.1
			1.0	0.7	-0.1	-0.2	0.4
				1.0	0.1	0.6	-0.6
					1.0	0.0	0.0
						1.0	0.6
							1.0

Stage 4

1.0	0.4	-0.4	0.4	0.1	0.2	0.0
	1.0	0.3	0.6	0.2	-0.1	-0.5
		1.0	0.8	0.0	-0.3	0.2
			1.0	0.1	0.6	-0.7
				1.0	0.0	0.0
					1.0	0.6
						1.0

Stage 5

1.0	0.1	0.5	0.1	-0.2	-0.5
	1.0	0.7	0.0	-0.2	0.3
		1.0	0.1	0.6	-0.7
			1.0	0.0	0.0
				1.0	0.6
					1.0

Stage 6

1.0	0.7	0.0	-0.2	0.2
	1.0	0.0	0.8	-0.6
		1.0	0.0	0.1
			1.0	0.8
				1.0

Stage 7

1.0	0.0	0.9	-0.7
	1.0	0.0	0.1
		1.0	0.8
			1.0

Stage 8

1.0	0.2	0.1
	1.0	0.7
		1.0

Stage 9

1.0	0.7
	1.0

BIBLIOGRAPHY

[1] AKAIKE, H. 'Fitting autoregressive models for prediction'. Ann. Inst. Statist. Math., Vol. 21 (1969).

[2] ALLEN, D.M. 'Mean square error of prediction as a criterion for selecting variables'. Technometrics, Vol. 13, No. 3 (1971).

[3] ANDERSON, T.W. 'Choice of the degree of a polynomial regression as a multiple decision problem'. Ann. Math. Stat., Vol. 32 (1962).

[4] ANDERSON, T.W. 'Determination of the order of dependence in normally distributed time series'. Proceedings of the Symposium in Time Series Analysis, Brown Univ. (ed. M. Rosenblatt) Wiley (1963).

[5] ANDERSON, T.W. 'The statistical analysis of time series' Wiley (1971).

[6] ANSCOMBE, F.J. 'Topics in the investigation of linear relations fitted by the method of least squares'. J.R.S.S. (B), Vol. 29, No. 1 (1967).

[7] BANCROFT, T.A. 'On biases in estimation due to the use of preliminary tests of significance'. Ann. Math. Stat., Vol. 15 (1944).

[8] BARTLETT, M.S. and DIANANDRA, P.H. 'Extensions of Quenouille's test for autoregressive schemes'. J.R.S.S. (B), Vol. 12 (1950).

[9] BEALE, E.M.L. 'Note on procedures for variable selection in multiple regression'. Technometrics, Vol. 12 (1970).

[10] BEALE, E.M.L. KENDALL, M.G. and MANN, D.W. 'The discarding of variables in multivariate analysis'. Biometrika, Vol. 54 (1967).

[11] BIOMED program BMDO2R, The Biomedical Computer Programs (Univ. of Calif.) (1971).

[12] BHANSALI, R.J. 'Asymptotic properties of the Wiener-Kolmogorov predictor' Unpublished Ph.D. thesis, London Univ. (1971).

[13] BHANSALI, R.J. 'A Monte Carlo comparison of the regression method and the spectral methods of prediction'. Univ. of Liverpool Technical Report (1972).

[14] BOX, G.E.P. and JENKINS, G.M. 'Time Series analysis: forecasting and control'. Holden-Day (1970).

[15] BOX, G.E.P. and MULLER, M.E. 'A note on the generation of random normal deviates'. Ann. Math. Stat., Vol. 29 (1958).

[16] C.A.C.M. Manual of Computer Programs, Algorithm 299.

[17] DALING, J.R. and TAMURA, H. 'Use of orthogonal factors for selection of variables in a regression equation - an illustration'. App. Stat., Vol. 19 (1970).

[18] DAVID, H.A. 'Order statistics'. Wiley (1970).

[19] DRAPER, N.R., GUTTMAN, I. and KANEMASU, H. 'The distribution of certain regression statistics'. Biometrika, Vol. 58 (1971).

[20] DRAPER, N.R. and SMITH, H. 'Applied regression analysis'. Wiley (1966).

[21] DUNNETT, C.W. and SOBEL, M. 'A bivariate generalization of students' t-distribution, with tables for certain special cases'. Biometrika, Vol. 41 (1954).

[22] DUNNETT, C.W. and SOBEL, M. 'Approximations to the probability integral and certain percentage points of a multivariate analogue of students' t-distribution Biometrika, Vol. 42 (1955).

[23] DURBIN, J. 'The fitting of time series models'. Rev. Int. Inst. Stat., Vol. 28 (1960).

[24] EFROYMSON, M.A. 'Multiple regression analysis', in 'Mathematical methods for digital computers' Part I (ed. A. Ralston and H.S. Wilf) Wiley (1960).

[25] FERGUSON, T.S. 'Mathematical statistics: a decision theoretic approach'. Academic Press (1967).

[26] GARSIDE, M.J. 'The best subset in multiple regression analysis'. App. Stat., Vol. 14 (1965).

[27] GOLDBERGER, A.S. 'Econometric theory'. Wiley (1964).

[28] GRANGER, C.W.J. 'Investigating causal relations by econometric models and cross-spectral methods'. Econometrika, Vol. 37 (1969).

[29] GRANGER, C.W.J. and NEWBOLD, P. 'Experience with forecasting univariate time series and the combination of forecasts'. Nottingham Forecasting Report No. 12 (1973).

[30] GRAYBILL, F.A. 'An introduction to linear statistical models'. Vol. I, McGraw-Hill (1961).

[31] HAITOVSKY, Y. 'A note on the maximization of $\bar{R}^2$'. The American Statistician, Vol. 23 (1969).

[32] HANNAN, E.J. 'The analysis of multiple time series' Wiley (1972).

[33] HARTLEY, H.O. 'Studentization and large-sample theory' J.R.S.S. Supp. 5 (1938).

[34] HOCKING, R.R. and LESLIE, R.N. 'Selection of the best subset in regression analysis'. Technometrics, Vol. 9, No. 4 (1967).

[35] JOHNSTON, J. 'Econometric methods'. Second Edition, McGraw-Hill (1972).

[36] KENDALL, M.G. 'Multivariate analysis'. Griffin (1957).

[37] KENDALL, M.G. and STUART, A. 'The advanced theory of statistics Vol. II. Griffin (1961).

[38] KENDALL, M.G. and STUART, A. 'The advanced theory of statistics' Vol. III. Griffin (1966).

[39] KENNEDY, W.J. and BANCROFT, T.A. 'Model building for prediction in regression based upon repeated significance tests'. Ann. Math. Stat., Vol 42 (1971).

[40] KITAGAWA, T. 'Successive process of statistical inference applied to linear regression analysis and its specialization to response surface analyses' Bull. Math. Statist., Vol. 8 (1959).

[41] KRISHNAIAH, P.R. 'Simultaneous tests and the efficiency of generalised balanced incomplete block designs'. Aerospace Research Laboratories Report. (1963).

[42] KRISHNAIAH, P.R. and ARMITAGE, J.V. 'Tables for the Studentized largest chi-square distribution and their applications'. Aerospace Research Laboratories Report (1964).

[43] KRISHNAIAH, P.R. and ARMITAGE, J.V. 'Probability integrals of the multivariate F distribution, with tables and applications'. Aerospace Research Laboratories Report (1965).

[44] KRISHNAIAH, P.R. and ARMITAGE, J.V. 'Tables for the distribution of the maximum of correlated chi-square variates with one degree of freedom'. Aerospace Research Laboratories Report (1965).

[45] KRISHNAIAH, P.R., ARMITAGE, J.V. and BREITER, M.C. 'Tables for the bivariate $|t|$ distribution'. Aerospace Research Laboratories Report (1969).

[46] KRISHNAMOORTHY, A.S. and PARTHASARATHY, M. 'A multivariate gamma-type distribution'. Ann. Math. Stat., Vol. 22 (1951).

[47] LARSON, H.J. and BANCROFT, T.A. 'Biases in prediction by regression for certain incompletely specified models'. Biometrika, Vol. 50 (1963).

[48] LARSON, H.J. and BANCROFT, T.A. 'Sequential model building for prediction in regression'. Ann. Math. Stat., Vol 34 (1963).

[49] LEHMANN, E.L. 'A general concept of unbiasedness'. Ann. Math. Stat., Vol. 22 (1951).

[50] LEHMANN, E.L. 'A theory of some multiple decision problems' (I and II). Ann. Math. Stat., Vol 28 (1957).

[51] LINDLEY, D.V. 'The choice of variables in multiple regression'. J.R.S.S. (B), Vol. 30, No. 1 (1968).

[52] LOTT, W.F. 'The optimum set of principal component restrictions on a least squares regression'. Dept. of Economics Research Report, University of Connecticut (1971).

[53] LÜTJOHANN, H. 'The stepwise regression algorithmseen from the statistician's point of view' Research Memorandum No. 11, Institute for Advanced Studies, Vienna (1968).

[54] MALLOWS, C.L. 'Choosing variables in a linear regression: a graphical aid'. Paper presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas (1964).

[55] MALLOWS, C.L. 'Choosing a subset regression'. Paper presented at the Joint Statistical Meeting, Los Angeles, Calif. (1966).

[56] MANN, H.B. and WALD, A. 'On the statistical treatment of linear stochastic difference equations'. Econometrika Vol. 11 (1943).

[57] MANTEL, N. 'Why stepdown procedures in variable selection?' Technometrics, Vol. 12 (1970).

[58] MILLER, R.G. 'Simultaneous statistical inference'. McGraw-Hill (1966).

[59] NAIR, K.R. 'The distribution of the extreme deviate from the sample mean and its studentized form'. Biometrika, Vol. 35 (1948).

[60] N.A.G. Library Manual, Nottingham Algorithms Group, I.C.L. 1900 System (1972).

[61] NEAVE, H.R. 'A random number package'. Computer applications in the natural and social sciences (Nottingham University), Vol. 14 (1972).

[62] NEWTON, R.G. and SPURRELL, D.J. 'A development of multiple regression for the analysis of routine data'. App. Stat., Vol. 16, No. 1 (1967).

[63] ORD, J.K. and WICKENS, M.R. 'The Use of principal components to counteract near-multicollinearity among the regressor variables'. Discussion paper in economics No. 38, Dept. of Economics, University of Bristol (1970).

[64]9 ORDEN, A. 'Matrix inversion and related topics by direct methods', in 'Mathematical methods for digital computers' Part I (ed. by A. Ralston and H.S. Wilf). Wiley (1960).

[65] PEARSON, E.S. and HARTLEY, H.O. 'Biometrika tables for statisticians' C.U.P. (1954).

[66] POPE, P.T. and WEBSTER, J.T. 'The use of an F-statistic in stepwise regression procedures'. Technometrics, Vol. 14 (1972).

[67] QUENOUILLE, M.H. 'A large sample test for the goodness of fit of autoregressive schemes'. J.R.S.S. Vol. 110 (1947).

[68] RAMACHANDRAN, K.V. 'On the simultaneous analysis of variance test'. Ann. Math. Stat. Vol. 27 (1956).

[69] RAO, C.R. 'Linear statistical inference and its applications'. Wiley (1965).

[70] SCHAERF, M.C. 'Estimation of the covariance and auto-regressive structure of a stationary time series'. Stanford University Technical Report No. 12 (1964).

[71] SCHATZOFF, M., TSAO, R. and FIENBERG, S. 'Efficient calculation of all possible regressions'. Technometrics, Vol. 10, No. 4 (1968).

[72] SCHEFFE, H. 'The analysis of variance'. Wiley (1959).

[73] SCHUSTER, A. 'On the periodicities of sunspots'. Phil. Trans. Royal Soc., A206 (1906).

[74] TORO-VIZCARRONDO, C. and WALLACE, T.D. 'A test of the mean square error criterion for restrictions in linear regression'. J.A.S.A., Vol. 63 (1968).

[75] WHITTLE, P. 'Tests of fit in time series analysis'. Biometrika, Vol. 39 (1952).

[76] WHITTLE, P. 'A statistical investigation of sunspot observations with special reference to H. Alfven's sunspot model'. The Astrophysical Journal Vol. 120 (1954).

[77] WOLD, H. 'A study in the analysis of stationary time series'. Almquist and Wiksell (1954).

[78] YULE, G. 'On the method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot series'. Phil. Trans. Royal Soc. A226 (1927).