CHAPTER 11

Variable Selection

11.1 Introduction

Frequently we start out with a fairly long list of independent variables that we suspect have some effect on the dependent variable, but for various reasons we would like to cull the list. One important reason is the resultant parsimony: It is easier to work with simpler models. Another is that reducing the number of variables often reduces multicollinearity. Still another reason is that it lowers the ratio of the number of variables to the number of observations, which is beneficial in many ways.

Obviously, when we delete variables from the model we would like to select those which by themselves or because of the presence of other independent variables have little effect on the dependent variable. In the absence of multicollinearity, as, say, with data from a well-designed experiment, variable search is very simple. One only needs to examine the b_j's and their standard errors and take a decision. Multicollinearity makes such decisions more difficult, and is the cause of any complexity in the methods given in this chapter. Notice that in the last chapter, when we examined the possibility of relationships among the independent variables, we ignored the effects any of them might have on the dependent variable.

Ideally, given the other variables in the model, those selected for removal have *no* effect on the dependent variable. This ideal situation is not likely to occur very often, and when it does not we could bias the regression (as we shall see in Section 11.2). Moreover, as we shall also show in Section 11.2, on the average, s^2 will tend to increase with a reduction in the variable list. Thus the practice of variable search is often a matter of making the best compromise between keeping s^2 and bias low and achieving parsimony and reducing multicollinearity.

During variable selection, one frequently finds, clustered around the chosen model, other models which are nearly 'as good' and not 'statistically distinguishable'. As with many other decisions in the practice of regression, the decisions involved in variable selection are seldom obvious. More often there is no unique choice and the one that is made reflects the analyst's best judgment at the time.

There is yet another problem with variable search procedures. Suppose we apply such a procedure to twenty independent variables constructed entirely of random numbers. Some of these variables, by sheer chance, may appear to be related to the dependent variable. Since we are picking the 'best' subset from our list, they appear in our short list and we end up with a nonsense relationship. An illustration can be found in Wilkinson (1979), and his findings are startling.

Despite these problems, variable search procedures can be invaluable if carried out judiciously. However, some further caveats are in order. We should be careful not to drop an important variable. For example, if the purpose of the study is to determine the relationship between the price of something and its sales, it would be silly to drop price just because our search procedure recommends it unless we are sure that price has virtually no effect on sales. In fact, the lack of effect might be the key finding here or may only be due to a poor design matrix where the corresponding column had a very short length or was highly related to some other column. Researchers whose primary interest is forecasting should also make sure that they are not (without very strong reasons) dropping an easy-to-predict independent variable in favor of a more troublesome one. Finally, if theoretical considerations or intuitive understanding of the underlying structure of the relationship suggest otherwise, the results of the mechanical procedures given below in Section 11.3 should play second fiddle. Ultimately, it is the researcher who should choose the variables — not the 'computer'!

11.2 Some Effects of Dropping Variables

Assume that

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \tag{11.1}$$

is the correct model and consider

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i(p-1)} + \epsilon_i^{(p)} \tag{11.2}$$

which includes only the first $p-1 < k$ independent variables from (11.1). In this section we discuss the effects of considering the incorrect abbreviated model (11.1). Since we have 'starred' several of the subsections of this section, we describe below some of the principal results obtained in the subsections.

As mentioned earlier, deleting some independent variables usually biases the estimates of the parameters left in the model. However, no bias occurs if the values of the deleted variables are orthogonal to those of the remaining variables, and then the estimates of $\beta_0, \ldots, \beta_{p-1}$ are exactly the same whether $x_{ip}, \ldots, x_{ik}$ are included or not. Deletion of variables usually increases the value of the expectation of s^2 and decreases (in the sense that the difference is non-negative definite) the covariance matrix of the estimates of $\beta_0, \ldots, \beta_{p-1}$. Note that we are referring to the covariance matrix, which is defined in terms of σ^2, not an estimate of it, which would frequently behave differently. Because estimates of $\beta_0, \ldots, \beta_{p-1}$ are biased, it is not surprising that predicted values usually become biased. One measure of this bias is called Mallows' C_p. While a definition of C_p is postponed to

Section 11.2.4, the key property for applications is that if (11.2) does not lead to much bias in the predicteds, then

$$\mathrm{E}(C_p) \approx p.$$

Therefore, if one is considering several candidate models, one can look at the corresponding s_p^2, R_p^2 and C_p values (the first two are the familiar s^2, R^2 measures applied to possibly truncated p variable model; all are provided by most packages) and judge which models are relatively bias-free. These become the ones from which a selection can be made.

Since coefficients become biased, the deletion of variables can cause residuals of the abbreviated model to have non-zero expectations (Section 11.2.4). This leads to plots of residuals against predicteds which sometimes show a pattern as if the residuals were related to the predicted values. When one sees such a plot, it is reasonable to suspect that a variable has been left out that should have been included. However, not much value should be placed on an apparent absence of pattern since even in quite biased models, such patterned plots do not occur with any regularity.

11.2.1 Effects on Estimates of β_j

Assume that the correct model is

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{11.3}$$

where $\mathrm{E}(\boldsymbol{\epsilon}) = 0$ and $\mathrm{cov}(\boldsymbol{\epsilon}) = \sigma^2 I$. Let $X = (X_1, X_2)$, and $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1' \ \boldsymbol{\beta}_2')$ where X_1 is $n \times p$ dimensional, $\boldsymbol{\beta}_1$ is a p-vector and the other dimensions are chosen appropriately. Then

$$\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}. \tag{11.4}$$

While this is the correct model, suppose we leave out $X_2\boldsymbol{\beta}_2$ and obtain the estimate $\boldsymbol{b}_1$ of $\boldsymbol{\beta}_1$ by least squares. Then $\boldsymbol{b}_1 = (X_1'X_1)^{-1}X_1'\boldsymbol{y}$, which not only is usually different from the first p components of the estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$ obtained by applying least squares to the full model (11.3), but also is usually biased, since

$$\begin{aligned}\mathrm{E}(\boldsymbol{b}_1) &= (X_1'X_1)^{-1}X_1'\,\mathrm{E}(\boldsymbol{y})\\ &= (X_1'X_1)^{-1}X_1'(X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2) = \boldsymbol{\beta}_1 + (X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_2.\end{aligned}$$

Thus, our estimate of $\boldsymbol{\beta}_1$ obtained by least squares after deleting $X_2\boldsymbol{\beta}_2$ is biased by the amount $(X_1'X_1)^{-1}X_1'X_2\boldsymbol{\beta}_2$.

Notice that the bias depends both on $\boldsymbol{\beta}_2$ and X_2. For example, if X_2 is orthogonal to X_1, that is, if $X_1'X_2 = 0$, then there is no bias. In fact if $X_1'X_2 = 0$,

$$(X'X)^{-1} = \begin{pmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{pmatrix}^{-1} = \begin{pmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{pmatrix}$$

(see Example A.8, p. 276) and since $X'\boldsymbol{y} = (X_1'\boldsymbol{y}\ X_2'\boldsymbol{y})$, it follows that $\boldsymbol{b}_1$ is actually the same as the first p components of $\boldsymbol{b}$.

11.2.2 *Effect on Estimation of Error Variance

The estimate of σ^2 based on the full model is given by

$$s^2 = \text{RSS}_{k+1}/(n-k-1) \equiv (n-k-1)^{-1}\boldsymbol{y}'[I-H]\boldsymbol{y}.$$

When we delete $X_2\boldsymbol{\beta}_2$, an estimate of σ^2 based on p independent variables will be given by

$$s_p^2 = \text{RSS}_p/(n-p) = (n-p)^{-1}(\boldsymbol{y}-\hat{\boldsymbol{y}}_p)'(\boldsymbol{y}-\hat{\boldsymbol{y}}_p) = (n-p)^{-1}\boldsymbol{y}'[I-H_1]\boldsymbol{y}$$

where $H_1 = X_1(X_1'X_1)X_1'$ and $\hat{\boldsymbol{y}}_p = X_1\boldsymbol{b}_1 = H_1\boldsymbol{y}$ is the predicted value of $\boldsymbol{y}$ based on the first p independent variables. While $\text{E}(s^2) = \sigma^2$, we need to calculate $\text{E}(s_p^2)$. Since $\text{tr}(I-H_1) = n-p$ and from (11.3) and Theorem B.3, p. 288, we can easily show that $\text{E}(\boldsymbol{y}\boldsymbol{y}') = \sigma^2 I + X\boldsymbol{\beta}\boldsymbol{\beta}'X'$, we get, using various properties of the trace of a matrix (Section A.6, p. 271),

$$\begin{aligned}(n-p)\,\text{E}(s_p^2) &= \text{E}[\boldsymbol{y}'(I-H_1)\boldsymbol{y}] = \text{E}[\,\text{tr}((I-H_1)\boldsymbol{y}\boldsymbol{y}')]\\ &= \text{tr}[(I-H_1)\,\text{E}(\boldsymbol{y}\boldsymbol{y}')] = \text{tr}[(I-H_1)(\sigma^2 I + X\boldsymbol{\beta}\boldsymbol{\beta}'X')]\\ &= (n-p)\sigma^2 + \text{tr}[(I-H_1)X\boldsymbol{\beta}\boldsymbol{\beta}'X'] = (n-p)\sigma^2 + \boldsymbol{\beta}'X'(I-H_1)X\boldsymbol{\beta}\end{aligned}$$

Hence,

$$\text{E}[(n-p)^{-1}\text{RSS}_p] = \text{E}(s_p^2) = \sigma^2 + (n-p)^{-1}\boldsymbol{\beta}'X'(I-H_1)X\boldsymbol{\beta} \qquad (11.5)$$

and, since $\text{E}(s^2) = \sigma^2$, it follows that

$$\text{E}(s_p^2 - s^2) = (n-p)^{-1}\boldsymbol{\beta}'X'(I-H_1)X\boldsymbol{\beta} \geq 0.$$

Therefore, s_p^2 is usually a biased estimator of σ^2 and $\text{E}(s_p^2)$ increases when variables are deleted. On the other hand, as shown in the next subsection, the covariance of $\boldsymbol{b}_1$ is less than or equal to the covariance of the estimate of $\boldsymbol{\beta}_1$ based on the full model. Therefore, practical choices involve determination of the trade-offs between improvements in one aspect and deterioration in another, and the reconciliation of these trade-offs with the aims of the analysis.

11.2.3 *Effect on Covariance Matrix of Estimates

Since $\boldsymbol{b}_1 = (X_1'X_1)^{-1}X_1'\boldsymbol{y}$ and $\text{cov}(\boldsymbol{y}) = \sigma^2 I$, we get

$$\text{cov}(\boldsymbol{b}_1) = \sigma^2(X_1'X_1)^{-1}.$$

However, based on the full model

$$\operatorname{cov}(\boldsymbol{b}) = \sigma^2(X'X)^{-1} = \sigma^2 \begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}^{-1}.$$

Therefore, the covariance of the vector $\boldsymbol{b}^{(1)}$ containing the first p components of $\boldsymbol{b}$ is, using Example A.8, p. 276,

$$\operatorname{cov}(\boldsymbol{b}^{(1)}) = \sigma^2[X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1]^{-1}. \qquad (11.6)$$

Similarly, the vector $\boldsymbol{b}^{(2)}$ of the last $k+1-p$ components of $\boldsymbol{b}$ has covariance

$$\sigma^2[X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2]^{-1}. \qquad (11.7)$$

Since $X_1'X_2(X_2'X_2)^{-1}X_2'X_1$ is positive semi-definite,

$$X_1'X_1 \geq X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1,$$

where the inequality signs are as defined in Section A.13, p. 279. It follows that

$$(X_1'X_1)^{-1} \leq [X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1]^{-1}.$$

Hence, $\operatorname{cov}(\boldsymbol{b}_1) \leq \operatorname{cov}(\boldsymbol{b}^{(1)})$; i.e., the covariance matrix of the estimate of first p components of $\boldsymbol{\beta}$ decreases when the remaining components, along with the corresponding columns of X, are deleted.

11.2.4 *Effect on Predicted Values: Mallows' C_p

Since $\mathrm{E}[\hat{\boldsymbol{y}}_p] = H_1\,\mathrm{E}[\boldsymbol{y}] = H_1X\beta$ it follows that $\hat{\boldsymbol{y}}_p$ is biased, unless $X\beta - H_1X\beta = [I - H_1]X\beta = \mathrm{E}[\boldsymbol{e}_p] = 0$, where $\boldsymbol{e}_p$ is the vector of the residuals from the abbreviated model. Define Bias$\,(\hat{\boldsymbol{y}}_p)$ as $\mathrm{E}[\hat{\boldsymbol{y}}_p - \hat{\boldsymbol{y}}]$. Then

$$\text{Bias}\,(\hat{\boldsymbol{y}}_p) = \mathrm{E}(\hat{\boldsymbol{y}}_p) - \mathrm{E}(\hat{\boldsymbol{y}}) = H_1X\boldsymbol{\beta} - X\boldsymbol{\beta} = -(I - H_1)X\boldsymbol{\beta}$$

and

$$\sum_{i=1}^{n}[\text{Bias}\,(\hat{y}_{pi})]^2 = [\mathrm{E}(\hat{\boldsymbol{y}}_p) - \mathrm{E}(\hat{\boldsymbol{y}})]'[\mathrm{E}(\hat{\boldsymbol{y}}_p) - \mathrm{E}(\hat{\boldsymbol{y}})] = \boldsymbol{\beta}'X'[I - H_1]X\boldsymbol{\beta}$$

where Bias$\,(\hat{y}_{pi})$ is the ith component of Bias$\,(\hat{\boldsymbol{y}}_p)$. To make this last expression scale-free, we standardize it by σ^2. Thus a standardized sum of squares of this bias is given by

$$\boldsymbol{\beta}'X'[I - H_1]X\boldsymbol{\beta}/\sigma^2,$$

which, on using (11.5), becomes

$$\mathrm{E}[\mathrm{RSS}_p]/\sigma^2 - (n-p).$$

Hence, an estimate of this standardized bias is

$$\mathrm{RSS}_p/s^2 - (n-p). \tag{11.8}$$

If the bias in $\hat{\boldsymbol{y}}_p$ introduced by dropping variables is negligible, this quantity should be close to zero.

An alternative way to examine bias in $\hat{\boldsymbol{y}}_p$ is to look at the mean square error matrix $\mathrm{MSE}(\hat{\boldsymbol{y}}_p)$, or, more conveniently, at its trace $\mathrm{TMSE}(\hat{\boldsymbol{y}}_p)$ (see equation (B.6), p. 288). Since, $\mathrm{cov}(\hat{\boldsymbol{y}}_p) = H_1 \,\mathrm{cov}(\boldsymbol{y}) H_1 = \sigma^2 H_1$,

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{y}}_p) &= \mathrm{E}[(\hat{\boldsymbol{y}}_p - X\boldsymbol{\beta})(\hat{\boldsymbol{y}}_p - X\boldsymbol{\beta})'] \\ &= \mathrm{cov}(\hat{\boldsymbol{y}}_p) + \mathrm{Bias}\,(\hat{\boldsymbol{y}}_p)\mathrm{Bias}\,(\hat{\boldsymbol{y}}_p)' = \sigma^2 H_1 + (I - H_1)X\boldsymbol{\beta}\boldsymbol{\beta}'X'(I - H_1).\end{aligned}$$

Therefore, the trace $\mathrm{TMSE}(\hat{\boldsymbol{y}}_p)$ of the MSE matrix, which is the sum of the mean square errors of each component of $\hat{\boldsymbol{y}}_p$, is

$$\mathrm{TMSE}(\hat{\boldsymbol{y}}_p) = \sigma^2 p + \boldsymbol{\beta}'X'(I - H_1)X\boldsymbol{\beta},$$

since $\mathrm{tr}(H_1) = p$.

Let

$$J_p = \mathrm{TMSE}(\hat{\boldsymbol{y}}_p)/\sigma^2 = p + \boldsymbol{\beta}'X'(I - H_1)X\boldsymbol{\beta}/\sigma^2,$$

which, using (11.5), can be estimated by

$$C_p = \frac{\mathrm{RSS}_p}{s^2} - (n-p) + p = \frac{\mathrm{RSS}_p}{s^2} - (n - 2p).$$

This is known as Mallows' C_p statistic (see Mallows, 1973). From the discussion of (11.8) it follows that, if bias is close to zero, C_p should usually be close to p.

11.3 Variable Selection Procedures

The purpose of variable selection procedures is to select or help select from the total number k of candidate variables a smaller subset of, say, $p-1$ variables. There are two types of such procedures: those that *help* choose a subset by presenting several if not all possible combinations of variables with corresponding values of C_p, s_p^2, R_p^2 and possibly other statistics, and those that pretty much *do* the selecting by presenting to the analyst very few (frequently one) subsets of variables for each value of $p-1$. As we have already mentioned, in many situations there is rarely one obviously best equation and the near winners are almost as good. Therefore, we prefer the former approach, which we have called the *search over all possible subsets.* It has also been called the *best subset regression.* However, such methods have voracious appetites for computer time, so that when computer time is at a premium, particularly if there are a large number of variables to select from, other methods might be necessary. Of these, the most popular is the *stepwise procedure* discussed in Section 11.3.2. Draper and Smith (1981) give a fuller discussion of variable search than we do but their ultimate recommendation is somewhat different.

11.3.1 Search Over All Possible Subsets

As the name implies, the search over all possible subsets of independent variables allows us to examine all regression equations constructed out of a given list of variables, along with some measure of fit for each. In our opinion, this procedure is the most useful, particularly if the number of variables is not too large. At the present time, a search over 20 variables is easily feasible on a mainframe computer, although new developments in compiler technology and in supercomputers will soon make it possible to computationally handle much larger numbers of variables. Even now, for large numbers of variables, one may force inclusion of a predetermined list of the variables in all models and search over only the remaining variables. (It is perhaps worth mentioning that computer packages do not actually fit each possible model separately; they use a 'short-cut' method, frequently one based on a procedure given in Furnival and Wilson, 1974, or see Seber, 1977, Chapter 11.)

A difficulty with this procedure stems from the prospect of having to examine huge computer outputs. For example, if even one line is devoted to each combination of variables, 20 variables would necessitate over a million lines. Therefore, several of the packages at least allow the user to use some criterion to eliminate combinations of variables that can be ruled out *a priori*. For example, SAS PROC RSQUARE allows one to choose to be printed for each p only the 'best' (based on R^2) m models and to put bounds on the number of variables p. In BMDP[1] (see Dixon, 1985) the user can choose among R^2, R_a^2 and C_p as the determinant of 'best' and ask that only the 'best' m models of any specified size $p-1$ along with the 'best' model of each size be printed. The Linear Least Squares Curve Fitting Program, which is a companion to Daniel and Wood (1980), uses C_p as the only means for culling and shows a plot of C_p's against p.

The PRESS (acronym for PREdiction Sum of Squares) statistic, first presented by Allen (1971), is another statistic that might be used to compare different models. It is defined as $\sum_{i=1}^n e_{i,-1}^2$ where $e_{i,-1}$ is as in equation (8.12) on p. 161. For each combination of variables, this provides a composite measure of how well it would predict each of the observations had the observation been left out when parameters were estimated. Several other measures also exist — see, for example Amemiya (1980), Judge *et al.* (1985), and Hocking (1976). However, nearly all of them eventually reduce

[1] BMDP Statistical Software Package is a registered trademark of BMDP Statistical Software Inc., Los Angeles, CA

to relatively simple functions of n, p, s_p^2, s^2 and R_p^2 (see SAS, 1985b, p. 715–16) — as, indeed, does C_p.

As will be apparent from Section 11.4, we somewhat favor C_p as a criterion for an initial selection. However, two points should be noted about it. First, when using it, we need to assume that the model with the entire list of independent variables included is unbiased. Second, C_p measures the bias in predicteds $\hat{y}_p$ from the abbreviated model, and these predicteds may not reveal the extent of the bias in estimates of certain future observations (see Example 10.1, p. 220). Whatever criterion statistic is used, in practice one sets bounds in such a way that the subset of models presented includes all those one would seriously consider.

Boyce *et al.* (1974) describe a very flexible program which can be used for the search over all possible subsets, although the primary purpose of the program is to search through all possible combinations of a given number of variables and identify the one with the highest R^2. The program can be obtained by writing to the authors of that monograph.

11.3.2 Stepwise Procedures

Of the stepwise procedures, the only one commonly used in actual applications is the stepwise procedure. Lest this sound silly, we point out that among stepwise procedures, there is one called *the* stepwise procedure. We also discuss the backward elimination procedure and the forward selection procedure, but primarily as an aid to the discussion of the stepwise procedure. Some stepwise procedures are not discussed here. Among the more interesting ones are the MAXR and the MINR procedures given in SAS PROC STEPWISE (see also Myers, 1986).

The Backward Elimination Procedures

Backward elimination procedures start with all variables in the model and eliminate the less important ones one by one. A partially manual version consists of removing one or two variables with low t-values, rerunning, removing some more variables, etc. Such a manual method does not work too badly when the researcher has a good understanding of the underlying relationship. However, it is tedious, and an automated version is available. Mechanically it works the same way but, as with most automated procedures, we pay for the convenience of automation by having to use preset selection criteria. The procedure computes the partial F's corresponding to each variable, given the list of variables included in the model at that step. The partial F statistic (sometimes called 'F to remove') is the square of the t statistic corresponding to each variable. Hence the probabilities obtained and the decisions taken are identical to using the t. If the lowest F value falls below a preset number (the 100 × α per cent point for the F distribution with the appropriate degrees of freedom, where α is either set by the analyst or by the computer package) the corresponding variable is deleted.

After each variable is deleted, partial F's are recomputed and the entire step is repeated with the variables still remaining in the model. The procedure stops when no partial F falls below the appropriate preset number.

While most users pay little attention to these preset numbers and use only the default values supplied by the computer package (for SAS α is .1), it is perhaps appropriate to match the number to the purpose of the analysis. At each step, the minimum value of partial F over all variables still in the model is computed. Hence, if the test is performed at the $100 \times \alpha$ per cent level, the actual probability of including one variable when in fact it has no effect on the dependent variable is much higher than α. Therefore, if one wishes to be particularly careful about not including inappropriate variables, one might wish to set very low α's. On the other hand, if one wishes the model to lean towards inclusivity rather than exclusivity, as one would if prediction was the main purpose for the model, a higher value of α is desirable (see also Forsythe, 1979, p. 855).

Apart from the problem of providing an inadequate list of models for the analyst to choose from, there is one further problem with backward elimination. Suppose we have three independent variables x_1, x_2, x_3, where x_1 is highly correlated with x_2 and x_3 and also with y and we would like to have x_1 in the final model — at least for parsimony. But being highly correlated with both x_2 and x_3, x_1 would have a large standard error and consequently a low t-value and a low partial F-value. As a result, it may get deleted early and we would never see it again.

The Forward Selection Procedure

The forward selection procedure works in the opposite way to the backward elimination procedures. It starts with no variable in the model and first selects that x_j which has the highest correlation with y. Subsequent selections are based on partial correlations, given the variables already selected. The *partial correlation* of y and x_j *given* $x_{j_1}, \ldots, x_{j_s}$, written as $r_{yx_j, x_{j_1} \ldots x_{j_s}}$, is the correlation between

1. the residuals obtained after regressing y on $x_{j_1}, \ldots, x_{j_s}$, and
2. the residuals obtained after regressing x_j on $x_{j_1}, \ldots, x_{j_s}$.

Clearly, the partial correlation measures the relationship between y and x_j after the linear effects of the other variables have been removed.

At every step, the partial F-value is computed for the variable just selected, given that variables previously selected are already in the model (such a partial F is called a 'sequential F' or sometimes 'F to enter'). If this sequential F-value falls below a preset number (e.g., the α-point of the appropriate F distribution — the default value of α in SAS is .5) the variable is deleted and another one is sought. If no suitable variable is found or if all the variables are in the model, the procedure stops. SAS uses a variation in which, instead of the partial correlations, the partial F's are computed for each variable not in the model. If the highest of the F's computed is high enough, the variable is included; otherwise the procedure stops.

A problem with forward selection is the reverse of the one for the backward selection. Suppose that of the highly correlated variables x_1,x_2 and x_3, we want x_2 and x_3 in the model because together they provide a better fit. But x_1 may enter the model first and prevent the others from getting in. Because of such problems, these procedures are now of primarily pedagogical or historical interest, having been replaced in actual use by the stepwise procedure and the all possible subsets search.

The Stepwise Procedure

The stepwise procedure is actually a combination of the two procedures just described. Like the forward selection procedure, it starts with no independent variable and selects variables one by one to enter the model in much the same way. But after each new variable is entered, the stepwise procedure examines every variable already in the model to check if it should be deleted, just as in a backward elimination step. Typically, the significance levels of F for both entry and removal are set differently than for the forward selection and backward elimination methods. It would be counter-productive to have a less stringent criterion for entry and a more stringent criterion for removal, since then we would constantly be picking up variables and then dropping them. SAS uses a default value of .15 for both entry and exit. As for the forward selection procedure, SPSS-X[2] (SPSS, 1986) permits, as an additional criterion, a tolerance level (e.g., $\text{TOL}_j \geq .01$) to be specified which needs to be satisfied for a variable to be considered (see Section 10.3.1, p. 222 for a definition of tolerance).

It is generally accepted that the stepwise procedure is vastly superior to the other stepwise procedures. But if the independent variables are highly correlated, the problems associated with the other stepwise procedures can remain (see Example 11.1 below; also see Boyce *et al.*, 1974). Like the forward selection and backward elimination procedures, usually only one equation is presented at each step. This makes it difficult for the analyst to use his or her intuition, even though most stepwise procedures allow the user to specify a list of variables to be always included.

[2] SPSS-X is a trademark of SPSS, Inc., Chicago, IL.

11.3.3 Stagewise and Modified Stagewise Procedures

In stagewise regression the decision to append an additional independent variable is made on the basis of plots of the residuals (from a regression of the dependent variable against all variables already included) against variables which are candidates for inclusion. In the modified stagewise procedure the plot considered is that of

1. the residuals obtained after regressing y on $x_{j_1}, \ldots, x_{j_s}$ against

2. the residuals obtained after regressing x_j on $x_{j_1}, \ldots, x_{j_s}$,

where $x_{j_1}, \ldots, x_{j_s}$ are the variables already in the model. Plots of this latter kind are called *added variable plots*, *partial regression plots*, *partial regression leverage plots* or simply *partial plots.*

In the case of the stagewise procedure, the slope obtained from applying least squares to the residuals is not a least squares estimate in the sense that, if the candidate variable is included in the model and least squares is applied to the resultant multiple regression model, we would get a different estimate for its coefficient. In the case of the modified stagewise procedure (without intercept) the estimates are LS estimates (see Exercise 11.1, also Mosteller and Tukey, 1977, p. 374 *et seq.).*

Modified stagewise least squares might appear to resemble a stepwise technique. But actually they are very different largely because of the way they are practiced. Stagewise and modified stagewise methods are essentially manual techniques — perhaps computer aided but nonetheless manual in essence. At every stage, transformations may be made and outliers dealt with and perhaps even weighting performed. Several examples of what we have called a modified stagewise approach can be found in Mosteller and Tukey (1977, see chapter 12 *et seq.*).

It might be mentioned in passing that some analysts find *partial plots* valuable for the identification of outliers and influential points (see Chatterjee and Hadi, 1986, Cook and Weisberg, 1982, Belsley, Kuh and Welsch, 1980).

11.4 Examples

Example 11.1

The data shown in Exhibit 11.1 are essentially made up by the authors. The independent variable x_1 has values which are the same as an independent variable in a data set in the authors' possession, except that they have been divided by 10 to make them more compatible with the size of x_2, which consists of pseudo-random numbers between 0 and 1. x_3 is $x_1 + x_2$ with