CHAPTER 15

Selecting the "Best" Regression Equation

15.0. INTRODUCTION

In this chapter we discuss some specific statistical procedures for selecting variables in regression. Suppose we wish to establish a linear regression equation for a particular response Y in terms of the basic predictor variables $X_1, X_2, \ldots, X_k$. Suppose further that $Z_1, Z_2, \ldots, Z_r$, all functions of one or more of the X's, represent the complete set of variables from which the equation is to be chosen and that this set includes any functions, such as squares, cross-products, logarithms, inverses, and powers, thought to be desirable and necessary. Two opposed criteria of selecting a resultant equation are usually involved:

1. To make the equation useful for predictive purposes we should like our model to include as many Z's as are necessary to keep bias errors small, so that reliable fitted values can be determined.

On the other hand:

2. (a) To keep the variance of the predictions reasonably small (recall that the average variance of the $\hat{Y}_i$ is $p\sigma^2/n$, where p is the number of parameters in the model and n is the number of observations), and (b) because of the costs involved in obtaining information on a large number of Z's and subsequently monitoring them, we should like the equation to include as few Z's as possible.

The practical compromise between these extremes is what is usually called *selecting the best regression equation.* There is no unique statistical procedure for doing this and many procedures have been suggested. If we knew the magnitude of σ^2 (the true random variance of the observations) for any single well-defined problem, our choice of a best regression equation would be much easier. Unfortunately, we are rarely in this position, so a great deal of personal judgment will be a necessary part of any of the methods discussed. We shall describe several procedures that have been proposed; all of these have their supporters. To add to the confusion they do not all necessarily lead to the same solution when applied to the same problem, although for many problems they will achieve the same answer. We shall discuss: (1) all possible regressions using three criteria: R^2, s^2, and the Mallows C_p; (2) best subset regression using R^2, R^2 (adjusted), and C_p; (3) stepwise regression; (4) backward elimination; and (5) some variations on previous methods. After each discussion we state a personal opinion. These opinions may stimulate the reader to agree or disagree.

Some Cautionary Remarks on the Use of Unplanned Data

When we do regression calculations on unplanned data (i.e., data arising from continuing operations and not from a designed experiment), some potentially dangerous possibilities can arise, as discussed by Box (1966). The error in the model may well not be random but may result from the joint effect of several variables not incorporated in the regression equation nor, perhaps, even measured. (Box calls these *latent* or *lurking* variables.) Due to the possibilities of bias in the estimates, discussed in Section 10.1, an observed false effect of a visible variable may, in fact, be caused by an unmeasured latent variable. Provided the system continues to run in the same way as when the data were recorded, this will not mislead. However, because the latent variable is not measured, its changes will not be seen or recorded, and such changes may well cause the predicted equation to become unreliable. Another defect in unplanned data is that, often, the most effective predictor variables are kept within quite a small range to keep the response(s) within specification limits. These small ranges will then frequently cause the corresponding regression coefficients to be found "non-significant," a conclusion that practical workers will interpret as ridiculous because they "know" the variable is effective. Both viewpoints are, of course, compatible; if an effective predictor variable is not varied much, it will show little or no effect. A third problem with unplanned data is that the operating policy (e.g., "if X_1 goes high, reduce X_2 to compensate") often causes large correlations between the predictors. This makes it impossible to see if changes in Y are associated with X_1, or X_2, or both. A carefully designed experiment can eliminate all the ambiguities described above. The effects of latent variables can be "randomized out," effective ranges of the predictor variables can be chosen, and correlations between predictors can be avoided. Where designed experiments are not feasible, happenstance data may still be analyzed via regression methods. However, the additional possibilities of jumping to erroneous conclusions must be kept in mind.

All of the cautions above imply that the use of *any* mechanical selection procedure is fraught with possible dangers for the unwary user. Selection procedures are valuable for quickly producing regression equations worth further consideration. However, common sense, basic knowledge of the data being analyzed, and considerations related to polynomial formation (Section 12.3) cannot ever be set aside.

Example Data

We use, for illustration, the data in a four-variable ($k = 4$) problem given by Hald (1952, p. 647). These are shown in Table 15.1. This particular problem has been chosen because it illustrates some typical difficulties that occur in regression analysis, and yet it has only four predictor variables and thirteen observations. There are only 15 possible regressions containing one or more of the X's and all of these 15 are shown in Appendix 15A, together with the original source of the data, some descriptive details, and a correlation matrix. Details from these printouts will be used in the various discussions that follow.

Comments

The data in the original reference have an additional predictor variable and one more observation. We have chosen to omit these so as not to become entangled here with intricacies not related to the selection procedures but related to the data. The omitted

TABLE 15.1. The "Hald" Regression Data (Also See Appendix 15A)

X_1	X_2	X_3	X_4	Y	Row Sum of X's
7	26	6	60	78.5	99
1	29	15	52	74.3	97
11	56	8	20	104.3	95
11	31	8	47	87.6	97
7	52	6	33	95.9	98
11	55	9	22	109.2	97
3	71	17	6	102.7	97
1	31	22	44	72.5	98
2	54	18	22	93.1	96
21	47	4	26	115.9	98
1	40	23	34	83.8	98
11	66	9	12	113.3	98
10	68	8	12	109.4	98

predictor variable varied only slightly. The data are actually from a mixtures experiment on cement in which the predictors were intended to add to 100%. Contrary to some views we have heard expressed, this does not make the data ineligible for a selection procedure, particularly as we employ only first-degree terms. We shall also make use of these data in the discussion of Chapter 19, where the special requirements of linear models for mixtures must be understood, particularly for quadratic and higher-order models.

15.1. ALL POSSIBLE REGRESSIONS AND "BEST SUBSET" REGRESSION

The first procedure is a rather cumbersome one. It requires the fitting of every possible regression equation that involves Z_0 plus any number of the variables $Z_1, \ldots, Z_r$ (where we have added a dummy variable $Z_0 = 1$ to the set of Z's as usual). Since each Z_i can either be, or not be, in the equation (two possibilities) and this is true for every Z_i, $i = 1, 2, \ldots, r$ (r Z's), there are altogether 2^r equations. (The Z_0 term is always in the equation.) If $r = 10$, a not unusually excessive number, $2^r = 1024$ equations must be examined! Each regression equation is assessed according to some criterion. The three criteria most used are:

1. The value of R^2 achieved by the least squares fit.
2. The value of s^2, the residual mean square.
3. The C_p statistic.

(All these are related to one another, in fact.) The choice of which equation is best to use is then made by assessing the patterns observed, as we describe via our example. The predictor variables here are X_1, X_2, X_3, and X_4. In this particular problem, there are no transformations, so that $Z_i = X_i$, $i = 1, 2, 3, 4$. The response variable is Y. A β_0 term is *always* included. Thus there are $2^4 = 16$ possible regression equations, which involve X_0 and the X_i, $i = 1, 2, 3, 4$. One of these is the fit of $Y = \beta_0 + \epsilon$. The other 15 fits appear in Appendix 15A. We now look at the R^2, s^2, and C_p statistics and assess the various equations.

TABLE 15.2. R^2 Values for All Possible Regressions, Hald Data

B	C	D	E
0.675 (4)	0.979 (1, 2)	0.98234 (1, 2, 4)	0.98237 (1, 2, 3, 4)
0.666 (2)	0.972 (1, 4)	0.98228 (1, 2, 3)	
0.534 (1)	0.935 (3, 4)	0.98128 (1, 3, 4)	
0.286 (3)	0.847 (2, 3)	0.97282 (2, 3, 4)	
	0.680 (2, 4)		
	0.548 (1, 3)		

Use of the R^2 Statistic

1. Divide the runs into five sets:

Set *A* consists of the run with only the mean value [model $E(Y) = \beta_0$]; this is not shown.

Set *B* consists of the four 1-variable runs [model $E(Y) = \beta_0 + \beta_i X_i$].

Set *C* consists of all the 2-variable runs [model $E(Y) = \beta_0 + \beta_i X_i + \beta_j X_j$].

Set *D* consists of all the 3-variable runs (and so on . . .).

Set *E* consists of the run with 4 variables.

2. Order the runs within each set by the value of the square of the multiple correlation coefficient, R^2. The parenthetical numbers in Table 15.2 are the subscripts of the X's in each equation. Later, the separating commas will be dropped but we put them in for this first display.

3. Examine the leaders and see if there is any consistent pattern of variables in the leading equations in each set. We see that after two variables have been introduced, further gain in R^2 is minor. Examination of the correlation matrix for the data (Appendix 15A) reveals that the pairs (X_1 and X_3) and (X_2 and X_4) are highly correlated, since

$$r_{13} = -0.824 \quad \text{and} \quad r_{24} = -0.973.$$

Thus the addition of further variables when X_1 and X_2 or when X_1 and X_4 are already in the regression equation will remove very little of the unexplained variation in the response. This is clearly shown by comparing Set *C* to Set *D*. The gain in R^2 from Set *D* to Set *E* is extremely small. This is simply explained by the observation that the X's are four (of five) mixture ingredients and the sum of the X-values for any specific point is nearly a constant (actually between 95 and 99).

What equation should be selected for further attention? One of the equations in Set *C* is clearly indicated but which one? If $f(X_1, X_2)$ is chosen, there is some inconsistency because the best single-variable equation involves X_4. For this reason many workers would prefer to use $f(X_1, X_4)$. The examination of all possible regressions does not provide a clear-cut answer to the problem. Other information, such as knowledge of the characteristics of the product studied and the physical role of the X-variables, must as always be added to enable a decision to be made.

Use of the Residual Mean Square, s^2

If all regressions are done on a large problem, an assessment of the average magnitude, $s^2(p)$, say, of the residual mean square as the number of variables in regression increases sometimes indicates the best cutoff point for the number of variables in regression.

TABLE 15.3. Residual Mean Squares for Hald Data Regressions

p	Residual Mean Squares	Average $s^2(p)$
2	115.06, 82.39, 176.31, 80.35	113.53
3	5.79, 122.71, 7.48, 41.54, 86.89, 17.57[a]	47.00
4	5.35, 5.33, 5.65, 8.20	6.13
5	5.98	5.98

[a] For example, 17.57 is the residual mean square obtained when fitting the model containing X_3 and X_4.

For the Hald data, the various residual mean squares for all sets of "p" variables, where p is the number of parameters in the model *including* β_0, are read off the computer displays of Appendix 15A and are shown in Table 15.3. The subscripts of the variables are ordered 1, 2, 3, 4, 12, 13, . . . , 123, 124, . . . , 1234.

When the number of potential variables in the model is large, $r > 10$, say, and when the number of data points is much larger than r, say, $5r$ to $10r$, the plot of $s^2(p)$ is usually quite informative. The fitting of regression equations that involve more predictor variables than are necessary to obtain a satisfactory fit to data is called overfitting. As more and more predictor variables are added to an already overfitted equation, the residual mean square will tend to stabilize and approach the true value of σ^2 as the number of variables increases, provided that all important variables have been included.

For small sets of data, such as in our example, we cannot expect this idea to work as effectively, of course, but it may provide a helpful first guideline. A plot of the average $s^2(p)$ against p is shown in Figure 15.1. It appears from this that an excellent estimate of σ^2 is about 6.00, and that four parameters (i.e., three predictor variables) should be included. However, looking at the s^2 values in more detail, we see that one of the runs with $p = 3$ had a residual mean square of 5.79, indicating that there exists a better run with three parameters (i.e., two predictor variables) than was indicated by the average residual mean square for $p = 3$, namely, 47.00. This is, in fact, the fitted equation with variables X_1 and X_2 in it. The next best $p = 3$ is the (X_1, X_4) combination with $s^2 = 7.48$. Thus this procedure has given us an "asymptotic" estimate of σ^2 with which we can choose a model or models whose residual variance estimate is approximately 6 and which contain the fewest predictor variables to achieve that.

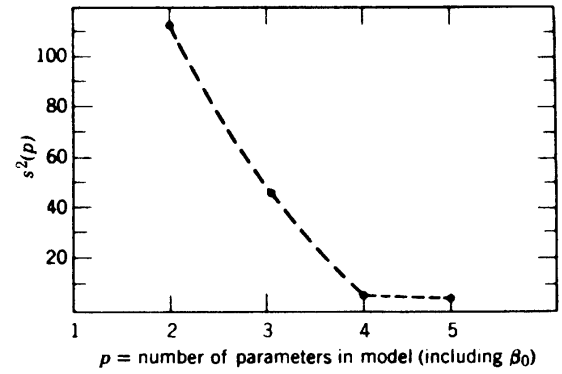

Figure 15.1. Plot of average residual mean square against p.

Use of the Mallows C_p Statistic

An alternative statistic is the C_p statistic, initially suggested by C. L. Mallows. This has the form

$$C_p = \mathrm{RSS}_p/s^2 - (n - 2p), \tag{15.1.1}$$

where RSS_p is the residual sum of squares from a model containing p parameters, p is the number of parameters in the model *including* β_0, and s^2 is the residual mean square from the largest equation postulated containing all the Z's and is presumed to be a reliable unbiased estimate of the error variance σ^2.

Note that the notation C_p is not completely explicit, as there are several C_p statistics for each p, except for the C_p value when all variables are included, C_{r+1}. The latter value is, in fact, not a random variable at all, but is

$$C_{r+1} = \{(n - r - 1)s^2\}/s^2 - \{n - 2(r + 1)\} = r + 1. \tag{15.1.2}$$

Essentially we are comparing the various models with p parameters to the full model with $r + 1$ parameters in it. Thus C_{r+1} is always "perfect" in the C_p versus p plot described below and must be discounted.

As Kennard (1971) has pointed out, C_p is closely related to the adjusted R^2 statistic, R_a^2, and it is also related to the R^2 statistic. Now, if an equation with p parameters is adquate, that is, does not suffer from lack of fit, $E(\mathrm{RSS}_p) = (n - p)\sigma^2$. Because we are also assuming that $E(s^2) = \sigma^2$, it is true, *approximately*, that the ratio RSS_p/s^2 has expected value $(n - p)\sigma^2/\sigma^2 = n - p$ so that, again approximately,

$$E(C_p) = p$$

for an adequate model. It follows that a plot of C_p versus p will show up the "adequate models" as points fairly close to the $C_p = p$ line. Equations with considerable lack of fit, that is, *biased equations,* will give rise to points above (often considerably above) the $C_p = p$ line. Because of random variation, points representing well-fitting equations can also fall below the $C_p = p$ line. The actual height C_p of each plotted point is also of importance because (it can be shown) it is an estimate of the overall total sum of squares of discrepancies (variance error plus bias error) of the fitted model from the true but unknown model. As terms are added to the model to reduce RSS_p, C_p usually increases. The "best" model is chosen after inspecting the C_p plot. We would look for a regression with a low C_p value about equal to p. When the choice is not clear-cut, it is a matter of personal judgment whether one prefers:

1. a biased equation that does not represent the actual data as well because it has larger RSS_p (so that $C_p > p$) but has a smaller estimate C_p of total discrepancy (variance error plus bias error) from the true but unknown model, or,

2. an equation with more parameters that fits the actual data better (i.e., $C_p \approx p$) but has a larger total discrepancy (variance error plus bias error) from the true but unknown model.

In other words, the smaller model has a smaller C_p value, but the C_p of the larger model (which has a larger value of p) is closer to its p.

Additional Reading. More detail on the considerations of judgment that apply in such cases will be found in Daniel and Wood (1980) and Gorman and Toman (1966). See also Mallows (1973). One quotation from the latter is worth repeating: "[C_p] cannot be expected to provide a single best equation when the data are intrinsically inadequate to support such a strong inference." Nor can *any* of the other selection

TABLE 15.4. Values of C_p and p for the Hald Data Equations[a]

Subscripts of Variables in Equation				C_p Values in Same Order				p
					443.2			1
1,	2,	3,	4	202.5,	142.5,	315.2,	138.7	2
12,	13,	14		2.7,	198.1,	5.5		3
23,	24,	34		62.4,	138.2,	22.4		3
123,	124,	134,	234	3.0,	3.0	3.5,	7.3	4
	1234				5.0			5

[a] For example, for the equation with predictors X_2 and X_4 we look for 24 in the left column; the corresponding values of C_p and p are 138.2 and 3, respectively.

procedures. All selection procedures are essentially methods for the orderly displaying and reviewing of the data. Applied with common sense, they can produce useful results; applied thoughtlessly, and/or mechanically, they may be useless or even misleading. See, further, the discussion by Mallows (1995); the dangers in using C_p when there are many competing "good" C_p values are described. Also relevant is Gilmour (1996).

Example of Use of the C_p Statistic

For the Hald data (see Appendix 15A) we have $n = 13$ and $s^2 = 5.983$ from the model fitted to all four predictor variables. Thus, for example, for the model $Y = \beta_0 + \beta_1 X_1 + \epsilon$ (for which $p = 2$ notice) we find a value

$$C_p = 1265.687/5.983 - (13 - 4) = 202.5.$$

This and all the remaining C_p values are given in Table 15.4. (Recall that, for the equation with all predictors in, $C_p = p$, as must be true by definition, because in this case $\text{RSS}_p = (n - p)s^2$; here $p = r + 1 = 5$.)

A plot of the smaller C_p statistics is given in Figure 15.2. The larger ones come from models that are clearly so biased (in comparison with the ones remaining) that we can eliminate them from consideration immediately. Since we are interested only in C_p values close to p, we can use the same scale on both vertical and horizontal axes and omit the points that will not be of interest. On the basis of the C_p statistic we see that the fitted equation with X_1 and X_2 is to be preferred over all others. It not only provides the smallest C_p value overall but has the edge over the fitted equation with X_1 and X_4, which exhibits signs of bias. The conclusion that the X_1 and X_2 equation is preferable is consistent with what we have already decided by checking both the R^2 and $s^2(p)$ values for the various equations as described above. However, the conclusion comes somewhat more easily from the C_p plot in the present example.

Opinion. In general, the analysis of all regressions is quite unwarranted. While it means that the investigator has "looked at all possibilities" it also means he has examined a large number of regression equation that intelligent thought would often reject out of hand. The amount of computer time used is wasteful and the sheer physical effort of examining all the computer printouts is enormous when more than a few variables are being examined. Some sort of selection procedure that shortens this task is preferable.

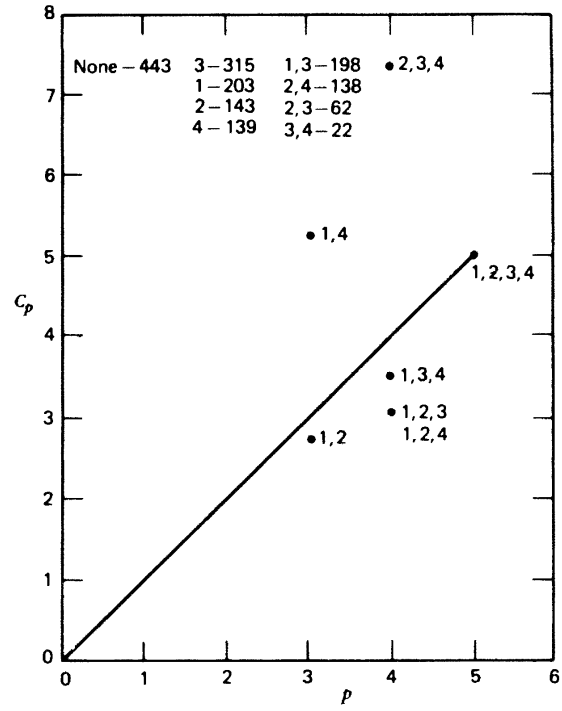

Figure 15.2. C_p plot for the Hald data.

Remark

We mentioned earlier (see Table 15.1, also) the fact that $X_1 + X_2 + X_3 + X_4$ is almost a constant. Thus any one X is almost linearly dependent on the other three. Thus if all four X's are put into the model, the $\mathbf{X'X}$ matrix has a small determinant value, 0.0010677. When an $\mathbf{X'X}$ determinant does have such a small value, it sometimes happens that the calculations involve primarily rounding errors and are meaningless. While this has not happened in the present case, the occurrence of a small determinant must always be regarded as a danger signal. Calculation of the Variance Inflation Factors (see Section 16.4) is one routine way to check for such problems.

"Best Subset" Regression

An alternative to performing all regressions is to use a program that provides a listing of the best K (you choose K) equations with one predictor variable in, two in, three in, . . . , and so on. Remember that β_0 is in all equations. See, for example, Furnival and Wilson (1974). "Best" is interpreted via (1) maximum R^2 value, or (2) maximum adjusted R^2 value, or (3) the Mallows C_p statistic, your choice. The best K equations overall are indicated in some programs also. Not all $2^r - 1$ equations are examined, so equations that should appear in a list sometimes do not. Table 15.5 shows a MINITAB display of the "best three" ($K = 3$) selections for the Hald data. The crosses indicate which predictors are in the regression given in each line. It is clear that we would reach the same conclusions as earlier, namely, that equations with (X_1, X_2) and (X_1, X_4) look reasonable.

The MINITAB instructions for producing such a printout, with the predictor variables stored in C1–C4 and the response in C5 are simply:

TABLE 15.5. Best Subsets Regression: MINITAB's breg Version, with K Selected as 3, for the Hald Data

Best Subsets Regression of C5

Vars	R-sq	Adj. R-sq	C-p	s	C 1	C 2	C 3	C 4
1	67.5	64.5	138.7	8.9639				X
1	66.6	63.6	142.5	9.0771		X		
1	53.4	49.2	202.5	10.727	X			
2	97.9	97.4	2.7	2.4063	X	X		
2	97.2	96.7	5.5	2.7343	X			X
2	93.5	92.2	22.4	4.1921			X	X
3	98.2	97.6	3.0	2.3087	X	X		X
3	98.2	97.6	3.0	2.3121	X	X	X	
3	98.1	97.5	3.5	2.3766	X		X	X
4	98.2	97.4	5.0	2.4460	X	X	X	X

```
breg  C5 C1-C4;
     best 3.
```

Opinion. There are some drawbacks to this type of procedure: (1) It tends to provide (in variations that supply the overall "best K" subset) equations with too many predictors included. (2) If K is chosen to be too small, the most sensible choice of fitted equation may not appear in the overall "best K" subset, though it will usually appear elsewhere in the printout. (3) No printed information is readily available concerning how the various subsets were obtained. However, provided these features are taken into account, this type of program can be useful as a quick alternative to doing all regressions.

15.2. STEPWISE REGRESSION

The stepwise regression procedure starts off by choosing an equation containing the single best X variable and then attempts to build up with subsequent additions of X's one at a time as long as these additions are worthwhile. The order of addition is determined by using the partial F-test values to select which variable should enter next. The highest partial F-value is compared to a (selected or default) F-to-enter value. After a variable has been added, the equation is examined to see if any variable should be deleted.

The basic procedure is as follows. First we select the Z most correlated with Y (suppose it is Z_1) and find the first-order, linear regression equation $\hat{Y} = f(Z_1)$. We check if this variable is significant. If it is not, we quit and adopt the model $Y = \overline{Y}$ as best; otherwise we search for the second predictor variable to enter regression.

We examine the partial F-values of all the predictor variables not in regression. The Z_j with the highest such value (suppose this is Z_2) is now selected and a second regression equation $\hat{Y} = f(Z_1, Z_2)$ is fitted. The overall regression is checked for significance, the improvement in the R^2 value is noted, and the partial F-values for

both variables now in the equation (not just the one most recently entered[1]) are examined. The lower of these two partial F's is then compared with an appropriate F percentage point, F-to-remove, and the corresponding predictor variable is retained in the equation or rejected according to whether the test is significant or not significant.

This testing of the "least useful predictor currently in the equation" is carried out at every stage of the stepwise procedure. A predictor that may have been the best entry candidate at an earlier stage may, at a later stage, be superfluous because of the relationships between it and other variables now in the regression. To check on this, the partial F criterion for each variable in the regression at any stage of calculation is evaluated, and the lowest of these partial F-values (which may be associated with the most recent entrant or with a previous entrant) is then compared with a preselected percentage point of the appropriate F-distribution or a corresponding default F-value. This provides a judgment on the contribution of the least valuable variable in the regression at that stage, treated as though it had been the most recent variable entered, irrespective of its actual point of entry into the model. If the tested variable provides a nonsignificant contribution, it is removed from the model and the appropriate fitted regression equation is then computed for all the remaining variables still in the model.

The best of the variables not currently in the model (i.e., the one whose partial F-value, given the predictors already in the equation, is greatest) is then checked to see if it passes the partial F entry test. If it passes, it is entered, and we return to checking all the partial F's for variables in. If it fails, a further removal is attempted. Eventually (unless the α-levels for entry and removal are badly chosen to provide a cycling effect[2]), when no variables in the current equation can be remoded and the next best candidate variable cannot hold its place in the the equation, the process stops. As each variable is entered into the regression, its effect on R^2, the square of the multiple correlation coefficient, is usually recorded and printed.

We shall use the Hald data once again to illustrate the workings of the stepwise procedure. Both entry and exit tests will be made at the level $\alpha = 0.05$.

Stepwise Regression on the Hald Data

1. The first variable in is the predictor most correlated with the response. It is also the one with the largest partial F-value (same as the regression equation F-value) of all those obtained with only one variable in regression; see the first line in Table 15.6. The choice falls on X_4; $r_{4Y} = -0.821$ is the largest correlation with Y and $F_{4|-} = 22.80$ is the largest of the partial F's. (The notation $F_{4|-}$ means the partial F-value for variable X_4 when no other X's are in the equation; β_0 is always in.)

2. Test for X_4. The F-value of 22.80 exceeds $F(1, 11, 0.95) = 4.84$. Retain X_4.

3. Seek the next best X. From Table 15.6 or Appendix 15A we see it is X_1 because $F_{1|4} = 108.22$ exceeds $F_{2|4} = 0.17$, and $F_{3|4} = 40.29$. Enter X_1 and compute the equation $\hat{Y} = f(X_1, X_4)$ as

[1] A simpler, and less effective, procedure in which only the most recent entrant is tested is called the *forward selection procedure.* This was described in the first edition and remains an option in many stepwise computer routines. Forward selection ensures that variables entered are not subsequently removed, which may be desirable for specific applications.

[2] It is usually best to choose the same significance levels for both the entry and exit test. If a smaller α is chosen for the exit test than for the entry test, a cycling pattern may occur. Use of a larger α for the exit test is conservative and may cause variables whose contributions have weakened to be retained. Some workers find this a desirable characteristic; it is a matter of personal preference.

TABLE 15.6. Partial F-Values for the Hald Data[a]

Subscripts of Variables in Regression	Partial F-Values of Variables Not in Regression			
	1	2	3	4
—	12.60	21.96	4.40	22.80
1	—	208.58	0.31	159.30
2	146.52	—	11.82	0.43
3	5.81	36.68	—	100.36
4	108.22	0.17	40.29	—
12	—	—	1.83	1.86
13	—	220.55	—	208.24
14	—	5.03	4.24	—
23	68.72	—	—	41.65
24	154.01	—	96.94	—
34	22.11	12.43	—	—
123	—	—	—	0.04
124	—	—	0.02	—
134	—	0.50	—	—
234	4.34	—	—	—

[a] For example, $F_{3|14} = 4.24$ is the partial F-value for b_3 given that (b_1 and b_4) are already in the regression equation. Remember that an intercept β_0 is fitted in all of the equations but is not specifically mentioned in the table.

$$\hat{Y} = 103.10 + 1.440X_1 - 0.614X_4.$$

This has an $R^2 = 0.972$ and has an overall $F = 176.63$, clearly significant. This completes this entry step. (The notation $F_{2|4}$ indicates the partial F-value for X_2 in an equation with X_2, X_4 and intercept, and so on.)

4. Check for a possible exit. The two partial F's are $F_{1|4} = 108.22$ and $F_{4|1} = 159.30$. The recently entered X_1 is the weaker, but it is significant since $108.22 > F(1, 10, 0.95) = 4.96$. Retain both X_4 and X_1.

5. The stepwise method now selects, as the next variable to enter, the one most highly partially correlated with the response (given that variables X_4 and X_1 are already in regression). This is seen to be variable X_2. $F_{2|14} = 5.03$ exceeds $F_{3|14} = 4.24$. The recomputed equation is

$$\hat{Y} = 71.65 + 1.452X_1 + 0.416X_2 - 0.237X_4,$$

with $R^2 = 0.98234$ and a significant overall $F = 166.83$.

6. Check for a possible exit. The three partial F's are

$$F_{1|24} = 154.01, \qquad F_{2|14} = 5.03, \qquad F_{4|12} = 1.86.$$

Clearly, $F_{4|12} = 1.86 < F(1, 9, 0.95) = 5.12$, so we must remove X_4. [*Note:* Although $F_{2|14} = 5.03$ is also less than 5.12, we take no action on X_2, because we first recompute the $\hat{Y} = f(X_1, X_2)$ equation. We remove only one X at a time and then "think again."]

7. The fitted equation with X_1 and X_2 is

$$\hat{Y} = 52.58 + 1.468X_1 + 0.662X_2.$$

The partial F's are $F_{1|2} = 146.52$ and $F_{2|1} = 208.58$ from Table 15.6, so both variables are retained. $R^2 = 0.979$ and overall $F = 229.50$.

8. We now look for a new candidate variable and check $F_{3|12} = 1.83$ and $F_{4|12} = 1.86$. Although X_4 is slightly the better, it is useless to enter it, as this just takes us back to Step 6. So the stepwise procedure terminates with the result given in Step 7.

Opinion. We believe this to be one of the best of the variable selection procedures and recommend its use. It makes economical use of computer facilities, and it avoids working with more X's than are necessary while improving the equation at every stage. However, stepwise regression can easily be abused by the "amateur" statistician. As with all the procedures discussed, sensible judgment is still required in the initial selection of variables and in the critical examination of the model through examination of residuals. It is easy to rely too heavily on the automatic selection performed in the computer.

A discussion of the method is given by M. A. Efroymson in an article in Ralston and Wilf (1962).

Remarks

1. Some programs based on this precedure use a t-test on the square root of the partial F-value instead of an F-test, making use of the fact that $F(1, \nu) = t^2(\nu)$, where $F(1, \nu)$ is an F-variable with 1 and ν degrees of freedom and $t(\nu)$ is a t-variable with ν degrees of freedom.

2. The test values for the partial F statistics are often called "F to enter" and "F to remove." (In MINITAB, Fenter and Fremove.)

3. Some programs use a fixed F-test value (e.g., $F = 4$) rather than look up individual percentage points as the df change. Typically the default values can be reset by the user. For example, F to enter can be set either higher to restrict entry or lower to allow more variables to enter.

Forward Selection Result

Suppose we had conducted the forward selection procedure mentioned in footnote 1. What would have been different? At Step 4 only X_1 (the last X entered) would have been tested, but the net result would be unchanged. Step 5 would also be unchanged. At Step 6, however, only X_2 would have been tested (and, just, rejected) and we would not have seen that X_4 is now the weakest variable. (Had we tested at the $\alpha = 0.10$ level, however, X_2 would have been retained.) We do not recommend use of Forward Selection unless it is specifically desired never to remove variables that were retained at earlier stages.

MINITAB Version of Stepwise Regression

Table 15.7 shows the progress of the MINITAB stepwise regression method as applied to the Hald data. The instructions that produced this printout with the predictors in columns C1–C4 and the response in column C5 were as follows:

```
stepwise C5 C1-C4;
fenter = 4;
fremove = 4.
```

Note that the default values (preset and operative unless they are altered) *are*

TABLE 15.7. Stepwise Method Applied to the Hald Data

STEPWISE REGRESSION OF C5 ON 4 PREDICTORS, WITH N = 13

STEP	1	2	3	4
CONSTANT	117.57	103.10	71.65	52.58
C4	−0.738	−0.614	−0.237	
T-RATIO	−4.77	−12.62	−1.37	
C1		1.44	1.45	1.47
T-RATIO		10.40	12.41	12.10
C2			0.416	0.662
T-RATIO			2.24	14.44
S	8.96	2.73	2.31	2.41
R-SQ	67.45	97.25	98.23	97.87

fenter = 4 and fremove = 4, so that there was no need to specify them. We can study this printout and see the effect of lowering the fenter and fremove values. Suppose, for example, that we had set both to 1. These would correspond to t-values of $1^{1/2} = 1$. We see that X_4 would not have been removed from the equation under those circumstances and the procedure would have terminated at step 3. Note that many selection procedures have special instructions allowing one to retain specific variables if desired, even if they are not large contributors to the regression.

15.3. BACKWARD ELIMINATION

The backward elimination method is also an economical procedure. It *tries* to examine only the "best" regressions containing a certain number of variables. The basic steps in the procedure are these:

1. A regression equation containing all variables is computed.
2. The partial F-test value is calculated for every predictor variable[3] *treated as though it were the last variable to enter the regression equation.*
3. The lowest partial F-test value, say, F_L, is compared with a preselected or default significance level, say, F_0.
   a. If $F_L < F_0$, remove the variable Z_L, which gave rise to F_L, from consideration and recompute the regression equation in the remaining variables; reenter stage (2).
   b. If $F_L > F_0$, adopt the regression equation as calculated.

We illustrate this procedure using the same data (Hald, 1952) as in the previous section. Since no transformations are used here, $Z_i = X_i$ and we refer to the variables as X's in discussing the example. Printouts of all the regression equations are in Appendix 15A. Also see Table 15.6.

The regression containing all four X's is

[3]Of course, the partial F-value is associated with a test of $H_0: \beta = 0$ versus $H_1: \beta \neq 0$ for any particular regression coefficient but this loose phrasing, in which we talk of the F-statistic for a particular predictor variable, is convenient and colloquial, and we use it frequently here and elsewhere.

$$\hat{Y} = 62.41 + 1.551X_1 + 0.510X_2 + 0.102X_3 - 0.144X_4.$$

The four partial F-values are

$$F_{1|234} = 4.34, \qquad F_{2|134} = 0.50,$$
$$F_{3|124} = 0.02, \qquad F_{4|123} = 0.04,$$

where the X-variable subscript before the vertical line is the one examined, and those after the vertical line are the others in the regression equation. The extra sum of squares formula is used to obtain all these F's from details given in Appendix 15A. (Example calculation: $F_{1|234} = [(73.815 - 47.863)/1]/5.983 = 4.34$. Note that our calculation is based on residual sums of squares, one of three ways of calculation that could be used.) Recall that a complete table of all the partial F-values is shown as Table 15.6. The ones quoted above can also be obtained as the squares of the t-values in the regression of Y onto X_1, X_2, X_3, and X_4.

We now choose the smallest partial F-value and compare it to a critical value of F based on a predetermined α-risk (or to a default value). In this case, the critical F-value for, say $\alpha = 0.05$ is $F(1, 8, 0.95) = 5.32$. The smallest partial F is for variable X_3; namely, calculated $F = 0.018$. Since the calculated F is smaller than the critical value 5.32, we reject X_3, dropping it from further consideration.

Next, find the least squares equation, $\hat{Y} = f(X_1, X_2, X_4)$. This is

$$\hat{Y} = 71.65 + 1.452X_1 + 0.416X_2 - 0.237X_4$$

(see Appendix 15A). The overall F-value for the equation is $F = 166.83$, which is statistically significant and in fact exceeds $F(3, 9, 0.999) = 13.90$. Examining this equation for potential elimination, one sees that X_4 has the smallest partial F and is a candidate for removal. The procedure for this elimination is similar to the preceding elimination with one change: namely, the critical F-value is $F(1, 9, 0.95) = 5.12$. Because the partial F associated with X_4 is 1.86 (which is less than 5.12), we remove X_4.

We now find the least squares equation $\hat{Y} = f(X_1, X_2)$, namely,

$$\hat{Y} = 52.58 + 1.468X_1 + 0.662X_2.$$

This provides a statistically significant overall equation with an F-value of 229.50, which exceeds $F(2, 10, 0.99) = 14.91$. Both variables X_1 and X_2 are significant regardless of position; that is, each partial F-value exceeds $F(1, 10, 0.95) = 4.96$, and also both exceed higher percentage points. Thus the backward elimination selection procedure is terminated and yields the equation

$$\hat{Y} = 52.59 + 1.47X_1 + 0.66X_2.$$

Opinion. This is a satisfactory procedure, especially for statisticians who like to see all the variables in the equation once in order "not to miss anything." It is much more economical of computer and personnel time than the "all regressions" method. However, if the input data yield an $\mathbf{X}'\mathbf{X}$ matrix that is ill conditioned—that is, nearly singular—then the overfitted equation may be nonsense due to rounding errors. With modern matrix inversion routines this is not usually a serious problem. One must also recognize that once a variable has been eliminated in this procedure *it is gone forever.* Thus all alternative models using the eliminated variable are not available for possible examination.

TABLE 15.8. Backward Elimination Method Applied to the Hald Data

STEPWISE REGRESSION OF C5 ON 4 PREDICTORS, WITH N = 13

STEP	1	2	3
CONSTANT	62.41	71.65	52.58
C1	1.55	1.45	1.47
T-RATIO	2.08	12.41	12.10
C2	0.510	0.416	0.662
T-RATIO	0.70	2.24	14.44
C3	0.10		
T-RATIO	0.14		
C4	−0.14	−0.24	
T-RATIO	−0.20	−1.37	
S	2.45	2.31	2.41
R-SQ	98.24	98.23	97.87

Remarks

1. Some programs based on this procedure use a t-test on the square root of the partial F-value instead of an F-test as given above, making use of the fact that, if $F(1, \nu)$ is an F-variable with 1 and ν degrees of freedom, and if $t(\nu)$ is a t-variable with ν degrees of freedom, then $F(1, \nu) = t^2(\nu)$.

2. Some programs use the words "F to remove" in their computer printouts. This is exactly the same as the partial F.

3. Some programs use a fixed F-test value (e.g., $F = 4$) rather than look up individual percentage points as the df change. Typically the user can reset the default value to a higher value, to finally include fewer variables, or to a lower value, to leave more variables in the equation.

4. As long as the $\mathbf{X'X}$ matrix is not singular, nor badly conditioned, the residual error variance obtained from the full model is a good estimate of σ^2 in the "asymptotic" sense discussed in Section 15.1. The backward elimination procedure essentially attempts to remove all unneeded X-variables without substantially increasing the size of the "asymptotic" estimate of σ^2. Note that the C_p method takes, basically, the same approach, comparing estimates of σ^2 from various models with the estimate of σ^2 from the full model. Satisfactory models are those for which the two estimates are of about the same size.

Table 15.8 shows a printout, for the Hald data, of the MINITAB version of the backward elimination method. The program is actually implemented through the stepwise program described in the foregoing section, via instructions like the following:

```
stepwise C5 C1-C4;
enter C1-C4;
fenter = 10000;
fremove = 4.
```

The "enter" subcommand puts in all the X's to begin. The large value of fenter prevents rejected variables being restored to the equation, while the moderate value of fremove permits the removal of predictors. If fremove had been set to a number

less than the square of the t-ratio for C4 at step 2, the procedure would have retained X_4 and terminated at step 2.

15.4. SIGNIFICANCE LEVELS FOR SELECTION PROCEDURES

Selecting Significance Levels in Stepwise Regression

In working the stepwise example for the Hald data in detail, we quoted the F percentage points for $\alpha = 0.05$. In Step 4, however, both F-values exceeded the 99.9% values, while in Step 5, the newly entering variable X_2 has an associated F-value that is significant at the $\alpha = 0.10$ level but not at the $\alpha = 0.05$ level. Once in, however, X_2 displaces X_4. Obviously it depends on how the program is written, and what α's are chosen for entering and removing variables whether or not a predictor even gets in. Thus choice of the α-level of the tests determines how the program will run.

1. If a small α (large F-test value) is selected, say, $\alpha \leq 0.05$, we reduce the chance of spurious variables entering. (Computer studies have shown that if we choose α as large as 0.15, columns of random numbers are often selected. There is good reason for this; see below.)
2. If a larger α (smaller F-test value) is selected, more predictor variables would typically be admitted. One can then look at the printout, however, and deduce what equation would have been offered if larger F-test values (smaller α's) had been inserted.

Clearly (2) is a more flexible approach. Most often, the entry and exit tests would be performed using the same α-value. It is also possible to use different levels for the entry and exit tests. If this is done, however, it is not wise to set the "exit α" smaller than the "entry α," or else one sometimes rejects predictors just admitted. Some workers like to set the "exit α" to be larger than the "entry α" to provide some "protection" for predictors already admitted to the equation. Such variations are a matter of personal taste, which, together with the α-values actually selected, have a great effect on the way a particular selection procedure behaves, and how many predictors are retained in the final equation. Some workers ignore the F-table altogether and simply compare the F-values with an arbitrary number such as 4. (This is used as the default value in MINITAB, for example.) To readers without strong personal opinions on this matter, we suggest setting $\alpha = 0.05$, or $\alpha = 0.10$ for both entry and exit tests, if the regression package being used allows this option (e.g., SAS). These levels can then be changed as experience dictates. (Alternatively, we can produce a larger printout by using smaller F-values, that is, larger α's.) As we discuss below the α-values are not accurate measures anyway, so that detailed agonizing over the precise choice of α is hardly worthwhile. Typically, the $\alpha = 0.05$ choice is conservative, that is, the actual value of α is much larger than 0.05 (some limited studies have been done and indicate this) and thus there will be a tendency to admit more predictors than the user might anticipate.

A Drawback to Understand But Not Be Overly Concerned About

Both the backward elimination and the stepwise procedures suffer from a difficulty that is perhaps not obvious at first sight. In the stepwise procedure, for example, the partial F-test made on the entry variable uses the largest partial F-value from all those partial F-values that might have been selected, arising from variables that are not in

the regression at that stage. The correct "null" sampling and testing distribution for this is *not* the ordinary F-distribution as we implicitly assume and is very difficult to obtain, except in certain simple cases. Studies have shown, for example, that, in some cases where an entry F-test was made at the α-level, the appropriate probability was roughly $q\alpha$, where there were q entry candidates at that stage. What can be done about this? One possibility is to find out the correct probability levels for any given case, while another is to make use of a different test statistic instead of the partial F. These possibilities have been discussed in various research papers but the overall problem has not, at the time of writing, been resolved to the extent that would warrant amending the procedures. Until it is resolved, we suggest that the reader use the procedures as given, not be too concerned with the actual probability levels, and simply regard the procedures as making a series of internal comparisons that will produce what appears to be the most useful set of predictors.

As we have already mentioned, some programs simply ask the user to provide numbers F-to-enter and F-to-remove for making the tests (e.g., such as $F = 4$). A. B. Forsythe put it this way in several editions of the BMDP program book *Biomedical Computer Programs, P-Series,* W. J. Dixon, Chief Editor, University of California Press, Berkeley:

> Several users of the stepwise regression and discriminant function analysis programs have asked why we use the terms F-to-enter and F-to-remove throughout the writeup and output, rather than simply calling them F values. Some have suggested that we could ask the user for the level of significance (α) and have the program convert it to the appropriate F values. The computer programming to do this is simple enough but the statistics are difficult. Since the program is selecting the "best" variable, the usual F tables do not apply. The appropriate critical value is a function of the number of cases, the number of variables, and, unfortunately, the correlation structure of the predictor variables. This implies that the level of significance corresponding to an F-to-enter depends upon the particular set of data being used. For example, with several hundred cases and 50 potential predictors, an F-to-enter of 11 would roughly correspond to $\alpha = 5\%$ if all 50 predictors were uncorrelated. The usual use of the F table would erroneously suggest a value of 4.

(We add to this that use of a value 4 would not be "wrong" but would simply imply a higher α-value, as we have discussed. Often a *substantially* higher value is implied, however. For the example in the quotation with nominal $\alpha = 0.05$, and $m = 50$ uncorrelated predictors, the "actual" α-value, given by $1 - (1 - \alpha)^m$, would be 0.923. This formula can be used as a rough guide for the consequences in general. Note that the BMDP program book is revised periodically; a current version should be consulted for possible changes in procedure.)

(The paper by Grechanovsky and Pinsker (1995) gives a computationally intensive way to obtain accurate p-values in forward selection. Forward selection, however, is not a procedure we recommend, and even with more accuracy, the final equation for the Hald data is unchanged from what we have already given, that is, X_4 and X_1 are successively entered and the next candidate X_2 is rejected. Recall that, if the stepwise method is used, it is X_4 that is rejected, not X_2, at this stage.)

15.5. VARIATIONS AND SUMMARY

Variations on the Previous Methods

While the procedures discussed do not necessarily select the absolute best model, they usually select an acceptable one. However, alternative procedures have been suggested

in attempts to improve the model selection. One proposal has been: Run the stepwise regression procedure with given levels for acceptance and rejection. When the selection procedure stops, determine the number of variables in the final selected model. Using this number of variables, say, q, do all possible sets of q variables from the r original variables and choose the best set.

Opinion. This procedure would reveal the situation pointed out in the discussion of the Hald data for the two-variable case: namely, that there are two close candidates for the model instead of one. When this happens, one could say that there is insufficient information in the data to make a clear-cut single choice. Other *a priori* considerations and the experimenter's judgment would be needed to select a final predictive model. This procedure also fails when a larger set of variables would be better but was never examined by the stepwise algorithm. In our experience, the added advantages of this procedure are minor.

Most selection programs permit the option of forcing variables into the equation, whether they would have been selected or not, ordinarily. This is often done to satisfy the fact that a variable is "known" to have an effect, even if such an effect is unlikely to be detectable (significant) in a "controlled" set of data. Forcing in variables is usually sensible and rarely does harm.

Summary

As we have seen, all the least squares selection procedures chose the fitted equation $\hat{Y} = 52.58 + 1.47X_1 + 0.66X_2$ as the "best" for the Hald data. The all-regressions method also provided $\hat{Y} = 103.10 + 1.44X_1 - 0.61X_4$ as a close second possibility, but subsequent procedures demonstrated it to be less desirable than the model involving X_1 and X_2. Although in many cases the same equation will arise from all least squares procedures, this is not guaranteed. Note that the selection of an equation with X_1 and X_2 in it means that we can explain the response values reasonably well by using only those two predictors. It does not imply that X_3 and X_4 are not needed in making the cement mixture.

In one sense the all-regressions procedure is best in that it enables us to "look at everything." The availability of best subsets routines eliminates the necessity of doing all regressions. However, for the Hald data, both the backward elimination and stepwise procedures also picked the same equation. Much would depend on the selection of rejection levels for the various F-tests, and also on the statistician's feeling about the level of increase desired in R^2 when all regressions are viewed. Inconsistent choices might lead to entirely different equations being reached by the various methods and this is not surprising.

Opinion. Our own preference for a practical regression method is the stepwise procedure. If exploration of equations "around" the stepwise choice is then desired, we prefer[4] the "best subsets" procedure, perhaps with the C_p statistic as a criterion for examination. To perform all regressions is not sensible, except when there are few predictors. If difficulties arise in the predictive interpretation of the fitted model or with the functional values of the estimated regression coefficients, then it is advisable to check the stability of the model in terms of the actual data to try to discover where the problem lies, rather than to blindly apply a rescue technique whose limitations may not be fully understood. Of course, with present day computer capabilities, one

[4]Some workers do the best subsets procedure first and follow it with stepwise or some alternative to stepwise.

can now do everything we have discussed with little effort. It is better to work with one selected technique and to master its particular idiosyncrasies; such an approach is probably more important than which technique *is* selected, in the long run. No technique will work well in all circumstances, no matter how good it may look on a particular example. No technique is always better than all the others. (If it were, this chapter would be shorter!) It must be continually kept in mind that, when the data are messy and not from a well-designed experiment, any fitted model is subject to constraints implied by both the patterns of the data and the practical limitations of the problem. The techniques discussed in this chapter can be useful tools. However, none of them can compensate for common sense and experience.

It is generally unwise to allow a selection procedure to choose from a set of polynomial terms; see Section 12.3 for a discussion. For an example that works out well, however, see Exercise P in "Exercises for Chapter 15."

Readers with a deeper interest in the topics of this chapter should consult the excellent and thought-provoking book by A. J. Miller (1990).

Another method of doing subset regression was proposed by Breiman (1995). In this paper, the least squares estimates are shrunk by factors whose sum is constrained. This produces a procedure "more stable" than the usual subset procedures and "less stable" than ridge regression (see Chapter 17), where stable means (here) that small changes in the data would not induce large changes in the equation selected. It is, of course, inevitable that permitting shrinkage and introducing restrictions would provide "stability." As long as these extra assumptions make practical sense, no harm is done. When the assumptions are wrong, however, a false sense of security is attained. See also Breiman (1996, e.g., p. 2374).

A more general constraint system, where the sum of the positive values of the β's to a power q could be used to restrict the parameter values, was suggested by Frank and Friedman (1993); see their p. 124. As $q \to 0$, the method tends to subset regression and $q = 1$ leads to the work of Tibshirani (1996). Again, a user must always consider whether such restrictions make practical sense. Ridge regression occurs when $q = 2$ (see Chapter 17). For other work look up "selection" in *Current Index to Statistics.*

15.6. SELECTION PROCEDURES APPLIED TO THE STEAM DATA

The steam data of Appendix 1A were subjected to selection procedure analysis by the MINITAB routines for stepwise regression, backward elimination, and best subsets regression using the following commands:

```
read C1-C10
[data from Appendix 1A]
end of data

stepwise C1 C2-C10;
fenter=1
fremove=1.

stepwise C1 C2-C10;
enter C2-C10;
fenter=10000;
fremove=1.

breg C1 C2-C10;
best 3.
```

Remarks

We have deliberately set three F-values to 1. This is to produce a larger printout than would typically arise with larger F-values.

The stepwise procedure passes through five steps and the final equation contains X_8, X_2, X_6, X_5 and X_{10}, entering in that order. If the default F-values of 4 were used, only X_8 and X_2 would enter because $F_{6|8,2} = (1.85)^2 = 3.42 < 4$, at stage 3.

The backward elimination procedure successively eliminates X_3, X_7, and X_5 and then stops at the point where the lowest F-value, for X_4, is 2.03. Note that, if we had set fremove = 4 rather than 1, the *same* equation, containing X_8, X_2, X_6, X_{10}, X_9, and X_4 would have been obtained! (The order of the X's is that of decreasing $|t|$ values; this facilitates comparison with the stepwise order above.) We see that the two procedures produce different equations but that X_8, X_2, X_6, and X_{10} occur in both sets.

The best subsets printout with a choice of $K = 3$ equations in each set provides some clarification. It shows that the best pair is X_8 and X_2. The best triple is X_8, X_6, and X_5, but the triple X_8, X_2, and X_6 is close behind in terms of R^2 value. Using four predictor variables contributes little to R^2, and the jump from two X's to three X's also adds little (about 2%). Both the equation with X_8 and X_2 and the equation with X_8, X_2, and X_6 seem like reasonable compromises here.

STEPWISE REGRESSION OF STEAM DATA C1 ON 9 PREDICTORS, C2-C10

STEP	1	2	3	4	5
CONSTANT	13.62299	9.47422	8.56626	0.09878	2.14331
C8	-0.0798	-0.0798	-0.0758	-0.0756	-0.0775
T-RATIO	-7.59	-10.59	-10.16	-10.50	-10.42
C2		0.76	0.49	0.30	0.43
T-RATIO		4.78	2.30	1.26	1.59
C6			0.108	0.142	0.150
T-RATIO			1.85	2.36	2.47
C5				0.29	0.22
T-RATIO				1.61	1.18
C10					-0.20
T-RATIO					-1.00
S	0.890	0.637	0.605	0.583	0.583
R-SQ	71.44	86.00	87.96	89.34	89.88

BACKWARD ELIMINATION ON STEAM DATA C1 USING 9 PREDICTORS, C2-C10

STEP	1	2	3	4
CONSTANT	1.8942	0.7743	-0.3667	4.3966
C2	0.71	0.48	0.54	0.66
T-RATIO	1.25	1.73	2.05	2.98
C3	-1.9			
T-RATIO	-0.46			
C4	1.13	1.28	1.27	1.30
T-RATIO	1.52	1.95	1.96	2.03
C5	0.12	0.15	0.16	
T-RATIO	0.58	0.79	0.87	

STEP	1	2	3	4	(continued)
C6	0.179	0.156	0.148	0.138	
T-RATIO	2.22	2.53	2.48	2.37	
C7	-0.018	-0.017			
T-RATIO	-0.74	-0.73			
C8	-0.077	-0.077	-0.068	-0.069	
T-RATIO	-4.67	-4.76	-6.39	-6.62	
C9	-0.086	-0.095	-0.094	-0.098	
T-RATIO	-1.65	-2.05	-2.06	-2.17	
C10	-0.35	-0.34	-0.34	-0.41	
T-RATIO	-1.64	-1.67	-1.69	-2.19	
S	0.570	0.555	0.548	0.544	
R-SQ	92.37	92.26	92.01	91.65	

==========

Best Subsets Regression of Steam Data, K = 3

Vars	R-sq	Adj. R-sq	C-p	s	C 2	C 3	C 4	C 5	C 6	C 7	C 8	C 9	C 10
1	71.4	70.2	35.1	0.89012							X		
1	41.0	38.5	94.9	1.2790						X			
1	28.7	25.6	119.1	1.4061					X				
2	86.0	84.7	8.5	0.63716	X						X		
2	84.9	83.5	10.7	0.66157					X		X		
2	84.7	83.3	11.1	0.66691		X					X		
3	88.5	86.8	5.6	0.59136				X	X		X		
3	88.0	86.2	6.7	0.60493	X				X		X		
3	86.5	84.6	9.5	0.64038	X						X		X
4	89.3	87.2	6.0	0.58318	X			X	X		X		
4	89.1	87.0	6.4	0.58877	X				X		X		X
4	89.0	86.8	6.6	0.59265				X	X	X	X		
5	90.0	87.3	6.7	0.58088	X	X			X		X		X
5	89.9	87.2	6.9	0.58304	X			X	X		X		X
5	89.8	87.1	7.1	0.58665			X	X	X		X	X	
6	91.7	88.9	5.4	0.54395	X		X		X		X	X	X
6	90.8	87.7	7.2	0.57264	X	X			X		X	X	X
6	90.7	87.6	7.4	0.57532	X		X	X	X		X	X	
7	92.0	88.7	6.7	0.54770	X		X	X	X		X	X	X
7	92.0	88.7	6.8	0.54917	X		X		X	X	X	X	X
7	91.9	88.5	7.0	0.55269	X	X	X		X		X	X	X
8	92.3	88.4	8.2	0.55548	X		X	X	X	X	X	X	X
8	92.2	88.3	8.3	0.55781	X	X	X		X	X	X	X	X
8	92.1	88.1	8.6	0.56168	X	X	X	X	X		X	X	X
9	92.4	87.8	10.0	0.56975	X	X	X	X	X	X	X	X	X

TABLE 15A.1. The Hald Data

X_1	X_2	X_3	X_4	Y	Row Sum of X's
7	26	6	60	78.5	99
1	29	15	52	74.3	97
11	56	8	20	104.3	95
11	31	8	47	87.6	97
7	52	6	33	95.9	98
11	55	9	22	109.2	97
3	71	17	6	102.7	97
1	31	22	44	72.5	98
2	54	18	22	93.1	96
21	47	4	26	115.9	98
1	40	23	34	83.8	98
11	66	9	12	113.3	98
10	68	8	12	109.4	98

APPENDIX 15A. HALD DATA, CORRELATION MATRIX, AND ALL 15 POSSIBLE REGRESSIONS

(*Source:* The data were first given in "Effect of composition of Portland cement on heat evolved during hardening," by H. Woods, H. H. Steinour, and H. R. Starke, *Industrial and Engineering Chemistry*, **24**, 1932, 1207–1214, Table I.) The variables shown in Table 15A.1 are:

X_1 = amount of tricalcium aluminate, $3\,CaO \cdot Al_2O_3$.

X_2 = amount of tricalcium silicate, $3\,CaO \cdot SiO_2$.

X_3 = amount of tetracalcium alumino ferrite, $4\,CaO \cdot Al_2O_3 \cdot Fe_2O_3$.

X_4 = amount of dicalcium silicate, $2\,CaO \cdot SiO_2$.

(Response) Y = heat evolved in calories per gram of cement.

X_1, X_2, X_3, and X_4 are measured as percent of the weight of the clinkers from which the cement was made.

Table 15A.2 shows the MINITAB version of the correlation matrix for the Hald data.

The following pages (MINITAB derived) show all 15 regressions containing an

TABLE 15A.2. Correlation Matrix for Hald Data (MINITAB Version)

```
read c1-c5
(enter Hald data now)
correlate c1-c5 m1
print m1
end
stop

                    Matrix M1
                      1.00000     0.22858    -0.82413    -0.24545     0.73072
                      0.22858     1.00000    -0.13924    -0.97295     0.81625
                     -0.82413    -0.13924     1.00000     0.02954    -0.53467
                     -0.24545    -0.97295     0.02954     1.00000    -0.82131
                      0.73072     0.81625    -0.53467    -0.82131     1.00000
```

intercept and one or more of the Hald predictors. In these, Y = column C50 and X_1–X_4 correspond to C1–C4.

```
MTB > regress c50 1 c1

The regression equation is
C50 = 81.5 + 1.87 C1

Predictor       Coef       Stdev     t-ratio         p
Constant      81.479       4.927       16.54     0.000
C1            1.8687      0.5264        3.55     0.005

s = 10.73      R-sq = 53.4%      R-sq(adj) = 49.2%

Analysis of Variance

SOURCE        DF         SS           MS          F          p
Regression     1     1450.1       1450.1      12.60      0.005
Error         11     1265.7        115.1
Total         12     2715.8

Unusual Observations
Obs.      C1       C50       Fit Stdev.Fit   Residual    St.Resid
 10     21.0    115.90    120.72      7.72      -4.82      -0.65 X

X denotes an obs. whose X value gives it large influence.


MTB > regress c50 1 c2

The regression equation is
C50 = 57.4 + 0.789 C2

Predictor       Coef       Stdev     t-ratio         p
Constant      57.424       8.491        6.76     0.000
C2            0.7891      0.1684        4.69     0.000

s = 9.077      R-sq = 66.6%      R-sq(adj) = 63.6%

Analysis of Variance

SOURCE        DF         SS           MS          F          p
Regression     1     1809.4       1809.4      21.96      0.000
Error         11      906.3         82.4
Total         12     2715.8

Unusual Observations
Obs.      C2       C50       Fit Stdev.Fit   Residual    St.Resid
 10     47.0    115.90     94.51      2.53      21.39       2.45R

R denotes an obs. with a large st resid.


The regression equation is
C50 = 110 - 1.26 C3

Predictor       Coef       Stdev     t-ratio         p
Constant    110.203       7.948       13.87     0.000
C3          -1.2558      0.5984       -2.10     0.060

s = 13.28      R-sq = 28.6%      R-sq(adj) = 22.1%
```

```
Analysis of Variance

SOURCE        DF          SS           MS          F          p
Regression     1       776.4        776.4       4.40      0.060
Error         11      1939.4        176.3
Total         12      2715.8


MTB > regress c50 1 c4

The regression equation is
C50 = 118 - 0.738 C4

Predictor       Coef       Stdev     t-ratio         p
Constant     117.568       5.262       22.34     0.000
C4           -0.7382      0.1546       -4.77     0.000

s = 8.964      R-sq = 67.5%      R-sq(adj) = 64.5%

Analysis of Variance

SOURCE        DF          SS           MS          F          p
Regression     1      1831.9       1831.9      22.80      0.000
Error         11       883.9         80.4
Total         12      2715.8

Unusual Observations
Obs.      C4       C50       Fit Stdev.Fit   Residual    St.Resid
 10     26.0    115.90     98.38      2.56      17.52        2.04R

R denotes an obs. with a large st. resid.


MTB > regress c50 2 c1,c2

The regression equation is
C50 = 52.6 + 1.47 C1 + 0.662 C2

Predictor       Coef       Stdev     t-ratio         p
Constant      52.577       2.286       23.00     0.000
C1            1.4683      0.1213       12.10     0.000
C2           0.66225     0.04585       14.44     0.000

s = 2.406      R-sq = 97.9%      R-sq(adj) = 97.4%

Analysis of Variance

SOURCE        DF          SS           MS          F          p
Regression     2      2657.9       1328.9     229.50      0.000
Error         10        57.9          5.8
Total         12      2715.8

SOURCE        DF      SEQ SS
C1             1      1450.1
C2             1      1207.8


MTB > regress c50 2 c1,c3

The regression equation is
C50 = 72.3 + 2.31 C1 + 0.494 C3
```

```
Predictor       Coef       Stdev     t-ratio         p
Constant       72.35       17.05        4.24     0.002
C1            2.3125      0.9598        2.41     0.037
C3            0.4945      0.8814        0.56     0.587

s = 11.08      R-sq = 54.8%      R-sq(adj) = 45.8%

Analysis of Variance

SOURCE        DF         SS           MS          F          p
Regression     2     1488.7        744.3       6.07      0.019
Error         10     1227.1        122.7
Total         12     2715.8

SOURCE        DF     SEQ SS
C1             1     1450.1
C3             1       38.6


MTB > regress c50 2 c1,c4

The regression equation is
C50 = 103 + 1.44 C1 - 0.614 C4

Predictor       Coef       Stdev     t-ratio         p
Constant     103.097       2.124       48.54     0.000
C1            1.4400      0.1384       10.40     0.000
C4          -0.61395     0.04864      -12.62     0.000

s = 2.734     R-sq = 97.2%      R-sq(adj) = 96.7%

Analysis of Variance

SOURCE        DF         SS           MS          F          p
Regression     2     2641.0       1320.5      176.63      0.000
Error         10       74.8          7.5
Total         12     2715.8

SOURCE        DF     SEQ SS
C1             1     1450.1
C4             1     1190.9

Unusual Observations
Obs.      C1       C50       Fit Stdev.Fit   Residual    St.Resid
  8      1.0    72.500    77.523     1.241     -5.023      -2.06R

R denotes an obs. with a large st. resid.


MTB > regress c50 2 c2,c3

The regression equation is
C50 = 72.1 + 0.731 C2 - 1.01 C3

Predictor       Coef       Stdev     t-ratio         p
Constant       72.075       7.383        9.76     0.000
C2            0.7313      0.1207        6.06     0.000
C3           -1.0084      0.2934       -3.44     0.006
```

```
s = 6.445      R-sq = 84.7%      R-sq(adj) = 81.6%

Analysis of Variance

SOURCE        DF         SS           MS          F          p
Regression     2     2300.3       1150.2      27.69      0.000
Error         10      415.4         41.5
Total         12     2715.8

SOURCE        DF     SEQ SS
C2             1     1809.4
C3             1      490.9

Unusual Observations
Obs.      C2       C50       Fit Stdev.Fit   Residual    St.Resid
 10     47.0    115.90    102.41      2.92      13.49       2.35R

R denotes an obs. with a large st. resid.


MTB > regress c50 2 c2,c4

The regression equation is
C50 = 94.2 + 0.311 C2 - 0.457 C4

Predictor       Coef       Stdev     t-ratio         p
Constant       94.16       56.63        1.66     0.127
C2            0.3109      0.7486        0.42     0.687
C4           -0.4569      0.6960       -0.66     0.526

s = 9.321      R-sq = 68.0%      R-sq(adj) = 61.6%

Analysis of Variance

SOURCE        DF         SS           MS          F          p
Regression     2     1846.88       923.44      10.63      0.003
Error         10      868.88        86.89
Total         12     2715.76

SOURCE        DF     SEQ SS
C2             1     1809.43
C4             1       37.46

Unusual Observations
Obs.      C2       C50       Fit Stdev.Fit   Residual    St.Resid
 10     47.0    115.90     96.89      4.46      19.01       2.32R

R denotes an obs. with a large st. resid.


MTB > regress c50 2 c3,c4

The regression equation is
C50 = 131 - 1.20 C3 - 0.725 C4

Predictor       Coef       Stdev     t-ratio         p
Constant     131.282       3.275       40.09     0.000
C3           -1.1999      0.1890       -6.35     0.000
C4          -0.72460     0.07233      -10.02     0.000

s = 4.192      R-sq = 93.5%      R-sq(adj) = 92.2%
```

```
Analysis of Variance

SOURCE        DF         SS           MS          F          p
Regression     2     2540.0       1270.0      72.27      0.000
Error         10      175.7         17.6
Total         12     2715.8

SOURCE        DF     SEQ SS
C3             1      776.4
C4             1     1763.7

Unusual Observations
Obs.      C3       C50        Fit Stdev.Fit   Residual    St.Resid
 10      4.0    115.90     107.64      1.89       8.26       2.21R

R denotes an obs. with a large st. resid.


MTB > regress c50 3 c1,c2,c3

The regression equation is
C50 = 48.2 + 1.70 C1 + 0.657 C2 + 0.250 C3

Predictor       Coef       Stdev     t-ratio          p
Constant      48.194       3.913       12.32      0.000
C1            1.6959      0.2046        8.29      0.000
C2           0.65691     0.04423       14.85      0.000
C3            0.2500      0.1847        1.35      0.209

s = 2.312      R-sq = 98.2%      R-sq(adj) = 97.6%

Analysis of Variance

SOURCE        DF         SS           MS          F          p
Regression     3     2667.65       889.22      166.34      0.000
Error          9       48.11         5.35
Total         12     2715.76

SOURCE        DF     SEQ SS
C1             1    1450.08
C2             1    1207.78
C3             1       9.79


MTB > regress c50 3 c1,c2,c4

The regression equation is
C50 = 71.6 + 1.45 C1 + 0.416 C2 - 0.237 C4

Predictor       Coef       Stdev     t-ratio          p
Constant       71.65       14.14        5.07      0.000
C1            1.4519      0.1170       12.41      0.000
C2            0.4161      0.1856        2.24      0.052
C4           -0.2365      0.1733       -1.37      0.205

s = 2.309      R-sq = 98.2%      R-sq(adj) = 97.6%

Analysis of Variance
```

```
SOURCE        DF          SS          MS          F          p
Regression     3     2667.79      889.26     166.83      0.000
Error          9       47.97        5.33
Total         12     2715.76

SOURCE        DF      SEQ SS
C1             1     1450.08
C2             1     1207.78
C4             1        9.93


MTB > regress c50 3 c1,c3,c4

The regression equation is
C50 = 112 + 1.05 C1 - 0.410 C3 - 0.643 C4

Predictor       Coef       Stdev     t-ratio         p
Constant     111.684       4.562       24.48     0.000
C1            1.0519      0.2237        4.70     0.000
C3           -0.4100      0.1992       -2.06     0.070
C4          -0.64280     0.04454      -14.43     0.000

s = 2.377      R-sq = 98.1%      R-sq(adj) = 97.5%

Analysis of Variance

SOURCE        DF          SS          MS          F          p
Regression     3     2664.93      888.31     157.27      0.000
Error          9       50.84        5.65
Total         12     2715.76

SOURCE        DF      SEQ SS
C1             1     1450.08
C3             1       38.61
C4             1     1176.24


MTB > regress c50 3 c2,c3,c4

The regression equation is
C50 = 204 - 0.923 C2 - 1.45 C3 - 1.56 C4

Predictor       Coef       Stdev     t-ratio         p
Constant      203.64       20.65        9.86     0.000
C2           -0.9234      0.2619       -3.53     0.006
C3           -1.4480      0.1471       -9.85     0.000
C4           -1.5570      0.2413       -6.45     0.000

s = 2.864      R-sq = 97.3%      R-sq(adj) = 96.4%

Analysis of Variance

SOURCE        DF          SS          MS          F          p
Regression     3     2641.95      880.65     107.38      0.000
Error          9       73.81        8.20
Total         12     2715.76

SOURCE        DF      SEQ SS
C2             1     1809.43
C3             1      490.89
C4             1      341.63

MTB > regress c50 4 c1,c2,c3,c4
```

```
The regression equation is
C50 = 62.4 + 1.55 C1 + 0.510 C2 + 0.102 C3 - 0.144 C4

Predictor       Coef      Stdev    t-ratio       p      VIF
Constant       62.41      70.07       0.89   0.399
C1            1.5511     0.7448       2.08   0.071     38.5
C2            0.5102     0.7238       0.70   0.501    254.4
C3            0.1019     0.7547       0.14   0.896     46.9
C4           -0.1441     0.7091      -0.20   0.844    282.5

s = 2.446      R-sq = 98.2%      R-sq(adj) = 97.4%

Analysis of Variance

SOURCE        DF         SS          MS         F         p
Regression     4    2667.90      666.97    111.48     0.000
Error          8      47.86        5.98
Total         12    2715.76

SOURCE        DF      SEQ SS
C1             1    1450.08
C2             1    1207.78
C3             1       9.79
C4             1       0.25
```

EXERCISES FOR CHAPTER 15

A. A production plant cost-control engineer is responsible for cost reduction. One of the costly items in his plant is the amount of water used by the production facilities each month. He decided to investigate water usage by collecting 17 observations on his plant's water usage and other variables. He had heard about multiple regression, but since he was quite skeptical he added a column of random numbers to his original observations. The complete set of data is shown in Table A.

1. Fit a suitable model to these data for the prediction of water usage.
2. Discuss the pros and cons of your selected equation.
3. Comment on the role of the random vector X_5.
4. Check the residuals.

TABLE A. Water Usage Data and Correlation Matrix

Data Code

X_1 = average monthly temperature (°F)
X_2 = average of production (M pounds)
X_3 = number of plant operating days in the month
X_4 = number of persons on the monthly plant payroll
X_5 = two-digit random number
X_6 = Y is the monthly water usage (gallons)

Original Data

	X_1	X_2	X_3	X_4	X_5	$X_6 = Y$
1	58.8	7107	21	129	52	3067
2	65.2	6373	22	141	68	2828
3	70.9	6796	22	153	29	2891
4	77.4	9208	20	166	23	2994
5	79.3	14792	25	193	40	3082

Original Data

	X_1	X_2	X_3	X_4	X_5	$X_6 = Y$
6	81.0	14564	23	189	14	3898
7	71.9	11964	20	175	96	3502
8	63.9	13526	23	186	94	3060
9	54.5	12656	20	190	54	3211
10	39.5	14119	20	187	37	3286
11	44.5	16691	22	195	42	3542
12	43.6	14571	19	206	22	3125
13	56.0	13619	22	198	28	3022
14	64.7	14575	22	192	7	2922
15	73.0	14556	21	191	42	3950
16	78.9	18573	21	200	33	4488
17	79.4	15618	22	200	92	3295

Correlation Matrix (Multiplied by 100)

	1	2	3	4	5	Y
1	100	−2	44	−8	11	29
2	−2	100	11	92	−11	63
3	44	11	100	3	4	−9
4	−8	92	3	100	−16	41
5	11	−11	4	−16	100	−6
Y	29	63	−9	41	−6	100

B. The demand for a consumer product is affected by many factors. In one study, measurements on the relative urbanization, educational level, and relative income of nine geographic areas were made in an attempt to determine their effect on the product usage. The data collected were as follows:

Area Number	Relative Urbanization X_1	Educational Level X_2	Relative Income X_3	Product Usage Y
1	42.2	11.2	31.9	167.1
2	48.6	10.6	13.2	174.4
3	42.6	10.6	28.7	160.8
4	39.0	10.4	26.1	162.0
5	34.7	9.3	30.1	140.8
6	44.5	10.8	8.5	174.6
7	39.1	10.7	24.3	163.7
8	40.1	10.0	18.6	174.5
9	45.9	12.0	20.4	185.7
Means	41.86	10.62	22.42	167.07
Standard deviations	s_1 4.1765	s_2 0.7463	s_3 7.9279	s_4 12.6452

The correlation matrix is

	X_1	X_2	X_3	Y
X_1	1	0.684	−0.616	0.802
X_2	0.684	1	−0.172	0.770
X_3	−0.616	−0.172	1	−0.629
Y	0.802	0.770	−0.629	1

Requirements

1. Use the stepwise procedure to determine a fitted first-order model using $F = 2.00$ for both entering and rejecting variables.

2. Write out the analysis of variance table and comment on the adequacy of the final fitted equation after examining residuals.

C. A parcel packing crew consists of five workers numbered 1 through 5 plus a foreman who works at all times. In the data below,

$X_j = 1$ if worker j is on duty, and 0 otherwise,

Y = number of parcels dispatched that day.

By employing a regression selection procedure of your choice, find an equation of form $Y = b_0 + \Sigma\, b_j X_j$, where the summation sign may *not* be over all j, which, in your opinion, suitably explains the data. Also, obtain the usual analysis of variance table and perform other useful analyses and comment as extensively as you can on your selected equation and on any other relationships you observe. In particular, having seen your results, what advice would you give?

Run	X_1	X_2	X_3	X_4	X_5	Y
1	1	1	1	0	1	246
2	1	0	1	0	1	252
3	1	1	1	0	1	253
4	0	1	1	1	0	164
5	1	1	0	0	1	203
6	0	1	1	1	0	173
7	1	1	0	0	1	210
8	1	0	1	0	1	247
9	0	1	0	1	0	120
10	0	1	1	1	0	171
11	0	1	1	1	0	167
12	0	0	1	1	0	172
13	1	1	1	0	1	247
14	1	1	1	0	1	252
15	1	0	1	0	1	248
16	0	1	1	1	0	169
17	0	1	0	0	0	104
18	0	1	1	1	0	166
19	0	1	1	1	0	168
20	0	1	1	0	0	148
Sum	9	16	16	9	9	3880
SS	9	16	16	9	9	795364

Repeat runs: (1, 3, 13, 14), (2, 8, 15), (4, 6, 10, 11, 16, 18, 19), (5, 7)

D. (*Sources*: "Selection of variables for fitting equations to data," by J. W. Gorman and R. J. Toman, *Technometrics,* **8**, 1966, 27–51. *Fitting Equations to Data,* 2nd edition, by C. Daniel and F. S. Wood, Wiley, New York, 1980, pp. 95–103, 109–117.) Gorman and Toman discussed an experiment in which the "rate of rutting" was measured on 31 experimental asphalt pavements. Five independent or predictor variables were used to specify the conditions under which each asphalt was prepared, while a sixth "dummy" variable was used to express the difference between the two separate "blocks" of runs into which the experiment was divided. The equation used to fit the data was

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon \qquad (1)$$

where

Y = log (change of rut depth in inches per million wheel passes),
X_1 = log (viscosity of asphalt),
X_2 = percent asphalt in surface course,
X_3 = percent asphalt in base course,
X_4 = dummy variable to separate two sets of runs,
X_5 = percent fines in surface course,
X_6 = percent voids in surface course.

You may assume that Eq. (1) is "complete" in the sense that it includes all the relevant terms. Your assignment is to select a suitable subset of these terms as the "best" regression equation in the circumstances.

TABLE D. Residual Sums of Squares (RSS) for All Possible Fitted Models, to Three Decimal Places

Variables[a]	RSS	Variables[a]	RSS	Variables[a]	RSS	Variables[a]	RSS
—	11.058	5	9.922	6	9.196	56	7.680
1	0.607	15	0.597	16	0.576	156	0.574
2	10.795	25	9.479	26	9.192	256	7.679
12	0.499	125	0.477	126	0.367	1256	0.364
3	10.663	35	9.891	36	8.848	356	7.678
13	0.600	135	0.582	136	0.567	1356	0.561
23	10.168	235	9.362	236	8.838	2356	7.675
123	0.498	1235	0.475	1236	0.365	12356	0.364
4	1.522	45	1.397	46	1.507	456	1.352
14	0.582	145	0.569	146	0.558	1456	0.553
24	1.218	245	1.030	246	1.192	2456	1.024
124	0.450	1245	0.413	1246	0.323	12456	0.313
34	1.453	345	1.383	346	1.437	3456	1.342
134	0.581	1345	0.561	1346	0.555	13456	0.545
234	1.041	2345	0.958	2346	0.995	23456	0.939
1234	0.441	12345	0.412	12346	0.311	123456	0.307

[a]Variables included in the equation; a β_0 term is included in all equations.

You will find the residual sums of squares for all the possible regressions in the accompanying table. This information will permit you to use any of the following procedures:

1. Backward elimination.
2. Forward selection.
3. Stepwise regression.
4. Examination of C_p.
5. Variations of 1–4.

You are welcome to use your own ingenuity instead of, or together with, any of these "standard" procedures.

Hint: The residual for the regression containing *no* predictors but only β_0 will give you the corrected total sum of squares.

E. (*Source*: "Sex differentials in teachers' pay," by P. Turnbull and G. Williams, *Journal of the Royal Statistical Society*, **A-137**, 1974, 245–258.) Note carefully: The data we provide were obtained by (first) a systematic selection of 100 observations (every 30th up to 3000 from an original set of 3414 observations) and (then) a weeding down to 90 observations by a removal of 10 handpicked observations. Two predictor variables were also deleted at this stage. Thus, while the data given below are in some sense representative, they do not necessarily reflect properly the behavior to be observed in the original, larger set. For this, the given reference should be consulted.

The data in the attached table consist of 90 observations on a number of characteristics of a selected group of British teachers in 1971.

Y = salary in pounds sterling.

X_1 = service in months, minus 12

X_2 = dummy variable for sex, 1 for a man, 0 for a women.

X_3 = dummy variable for sex, 1 for a man or single woman, 0 for a married women.

[Note that the combination $(X_2, X_3) = (1, 0)$ is never used. The remaining three combinations distinguish between a man, (1, 1); a single woman, (0, 1); and a married woman, (0, 0).]

X_4 = class of degree for graduates, coded 0 to 6.

X_5 = type of school in which employed, coded 0 or 1.

X_6 = 1 for trained graduate, 0 for untrained graduate or trained nongraduate teacher.

X_7 = 1 for a break in service of more than two years, 0 otherwise.

1. Using a selection procedure of your choice, fit the "best" model of the form

$$\log_{10} Y = \beta_0 + \Sigma \beta_j X_j + \epsilon$$

to the data, where the summation may not include all the possible X's.

2. Check the residuals. What other possible term for the model suggests itself?

3. Add this new term to the model and fit the "best" model now, using the same selection procedure.

4. What model would you, in fact, use if you wanted to explain a usefully large amount of the variation in the salary data? Why?

(We are grateful to Dr. P. Turnbull who kindly provided the original set of data.)

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
998	7	0	0	0	0	0	0
1015	14	1	1	0	0	0	0
1028	18	1	1	0	1	0	0
1250	19	1	1	0	0	0	0
1028	19	0	1	0	1	0	0
1028	19	0	0	0	0	0	0
1018	27	0	0	0	0	0	1
1072	30	0	0	0	0	0	0
1290	30	1	1	0	0	0	0
1204	30	0	1	0	0	0	0
1352	31	0	1	2	0	1	0
1204	31	0	0	0	1	0	0
1104	38	0	0	0	0	0	0
1118	41	1	1	0	0	0	0
1127	42	0	0	0	0	0	0
1259	42	1	1	0	1	0	0
1127	42	1	1	0	0	0	0
1127	42	0	0	0	1	0	0
1095	47	0	0	0	0	0	1
1113	52	0	0	0	0	0	1
1462	52	0	1	2	0	1	0
1182	54	1	1	0	0	0	0
1404	54	0	0	0	1	0	0
1182	54	0	0	0	0	0	0

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1594	55	1	1	2	1	1	0
1459	66	0	0	0	1	0	0
1237	67	1	1	0	1	0	0
1237	67	0	1	0	1	0	0
1496	75	0	1	0	0	0	0
1424	78	1	1	0	1	0	0
1424	79	0	1	0	0	0	0
1347	91	1	1	0	1	0	0
1343	92	0	0	0	0	0	1
1310	94	0	0	0	1	0	0
1814	103	0	0	2	1	1	0
1534	103	0	0	0	0	0	0
1430	103	1	1	0	0	0	0
1439	111	1	1	0	1	0	0
1946	114	1	1	3	1	1	0
2216	114	1	1	4	1	1	0
1834	114	1	1	4	1	1	1
1416	117	0	0	0	0	0	1
2052	139	1	1	0	1	0	0
2087	140	0	0	2	1	1	1
2264	154	0	0	2	1	1	1
2201	158	1	1	4	0	1	1
2992	159	1	1	5	1	1	1
1695	162	0	1	0	0	0	0
1792	167	1	1	0	1	0	0
1690	173	0	0	0	0	0	1
1827	174	0	0	0	0	0	1
2604	175	1	1	2	1	1	0
1720	199	0	1	0	0	0	0
1720	209	0	0	0	0	0	0
2159	209	0	1	4	1	0	0
1852	210	0	1	0	0	0	0
2104	213	1	1	0	1	0	0
1852	220	0	0	0	0	0	1
1852	222	0	0	0	0	0	0
2210	222	1	1	0	0	0	0
2266	223	0	1	0	0	0	0
2027	223	1	1	0	0	0	0
1852	227	0	0	0	1	0	0
1852	232	0	0	0	0	0	1
1995	235	0	0	0	0	0	1
2616	245	1	1	3	1	1	0
2324	253	1	1	0	1	0	0
1852	257	0	1	0	0	0	1
2054	260	0	0	0	0	0	0
2617	284	1	1	3	1	1	0
1948	287	1	1	0	0	0	0
1720	290	0	1	0	0	0	1
2604	308	1	1	2	1	1	0

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1852	309	1	1	0	1	0	1
1942	319	0	0	0	1	0	0
2027	325	1	1	0	0	0	0
1942	326	1	1	0	1	0	0
1720	329	1	1	0	1	0	0
2048	337	0	0	0	0	0	0
2334	346	1	1	2	1	1	1
1720	355	0	0	0	0	0	1
1942	357	1	1	0	0	0	0
2117	380	1	1	0	0	0	1
2742	387	1	1	2	1	1	1
2740	403	1	1	2	1	1	1
1942	406	1	1	0	1	0	0
2266	437	0	1	0	0	0	0
2436	453	0	1	0	0	0	0
2067	458	0	1	0	0	0	0
2000	464	1	1	2	1	1	0

F. Twenty (20) mature wood samples of slash pine cross sections were prepared, 30 μm in thickness. These sections were stained jet-black in Chlorozol E. The following determinations were made.

Data on Anatomical Factors and Wood Specific Gravity of Slash Pine

Number of Fibers/mm^2 in Springwood	Number of Fibers/mm^2 in Summerwood	Spring-wood (%)	Light Absorption Springwood (%)	Light Absorption Summerwood (%)	Wood Specific Gravity
573	1059	46.5	53.8	84.1	0.534
651	1356	52.7	54.5	88.7	0.535
606	1273	49.4	52.1	92.0	0.570
630	1151	48.9	50.3	87.9	0.528
547	1135	53.1	51.9	91.5	0.548
557	1236	54.9	55.2	91.4	0.555
489	1231	56.2	45.5	82.4	0.481
685	1564	56.6	44.3	91.3	0.516
536	1182	59.2	46.4	85.4	0.475
685	1564	63.1	56.4	91.4	0.486
664	1588	50.6	48.1	86.7	0.554
703	1335	51.9	48.4	81.2	0.519
653	1395	62.5	51.9	89.2	0.492
586	1114	50.5	56.5	88.9	0.517
534	1143	52.1	57.0	88.9	0.502
523	1320	50.5	61.2	91.9	0.508
580	1249	54.6	60.8	95.4	0.520
448	1028	52.2	53.4	91.8	0.506
476	1057	42.9	53.2	92.9	0.595
528	1057	42.4	56.6	90.0	0.568

Source: "Anatomical factors influencing wood specific gravity of slash pines and the implications for the development of a high-quality pulpwood," by J. P. Van Buijtenen, *Tappi,* **47**(7), 1964, 401–404.

Requirements. Develop a prediction equation for wood specific gravity. Examine the residuals for your model and state conclusions.

G. Fit to the steam data of Appendix 1A a model for predicting monthly steam usage and which is such that:

1. R^2 exceeds 0.8.
2. All b coefficients are statistically significant (take $\alpha = 0.05$).
3. No evidence exists of patterns in the residuals obtained from the fitted model.
4. The residual standard deviation expressed as a percentage of the mean response is less than 7%.

H. In the CAED Report 17, Iowa State University, 1963, the following data are shown for the state of Iowa, 1930–1962.

Year	X_1	X_2 Pre season Precipitation (in.)	X_3 May Temperature (°F)	X_4 June Rain (in.)	X_5 June Temperature (°F)	X_6 July Rain (in.)	X_7 July Temperature (°F)	X_8 August Rain (in.)	X_9 August Temperature (°F)	Y Corn Yield (bu/acre)
1930	1	17.75	60.2	5.83	69.0	1.49	77.9	2.42	74.4	34.0
	2	14.76	57.5	3.83	75.0	2.72	77.2	3.30	72.6	32.9
	3	27.99	62.3	5.17	72.0	3.12	75.8	7.10	72.2	43.0
	4	16.76	60.5	1.64	77.8	3.45	76.1	3.01	70.5	40.0
	5	11.36	69.5	3.49	77.2	3.85	79.7	2.84	73.4	23.0
	6	22.71	55.0	7.00	65.9	3.35	79.4	2.42	73.6	38.4
	7	17.91	66.2	2.85	70.1	0.51	83.4	3.48	79.2	20.0
	8	23.31	61.8	3.80	69.0	2.63	75.9	3.99	77.8	44.6
	9	18.53	59.5	4.67	69.2	4.24	76.5	3.82	75.7	46.3
	10	18.56	66.4	5.32	71.4	3.15	76.2	4.72	70.7	52.2
	11	12.45	58.4	3.56	71.3	4.57	76.7	6.44	70.7	52.3
	12	16.05	66.0	6.20	70.0	2.24	75.1	1.94	75.1	51.0
	13	27.10	59.3	5.93	69.7	4.89	74.3	3.17	72.2	59.9
	14	19.05	57.5	6.16	71.6	4.56	75.4	5.07	74.0	54.7
	15	20.79	64.6	5.88	71.7	3.73	72.6	5.88	71.8	52.0
	16	21.88	55.1	4.70	64.1	2.96	72.1	3.43	72.5	43.5
	17	20.02	56.5	6.41	69.8	2.45	73.8	3.56	68.9	56.7
	18	23.17	55.6	10.39	66.3	1.72	72.8	1.49	80.6	30.5
	19	19.15	59.2	3.42	68.6	4.14	75.0	2.54	73.9	60.5
	20	18.28	63.5	5.51	72.4	3.47	76.2	2.34	73.0	46.1
	21	18.45	59.8	5.70	68.4	4.65	69.7	2.39	67.7	48.2
	22	22.00	62.2	6.11	65.2	4.45	72.1	6.21	70.5	43.1
	23	19.05	59.6	5.40	74.2	3.84	74.7	4.78	70.0	62.2
	24	15.67	60.0	5.31	73.2	3.28	74.6	2.33	73.2	52.9
	25	15.92	55.6	6.36	72.9	1.79	77.4	7.10	72.1	53.9
	26	16.75	63.6	3.07	67.2	3.29	79.8	1.79	77.2	48.4
	27	12.34	62.4	2.56	74.7	4.51	72.7	4.42	73.0	52.8
	28	15.82	59.0	4.84	68.9	3.54	77.9	3.76	72.9	62.1
	29	15.24	62.5	3.80	66.4	7.55	70.5	2.55	73.0	66.0
	30	21.72	62.8	4.11	71.5	2.29	72.3	4.92	76.3	64.2
	31	25.08	59.7	4.43	67.4	2.76	72.6	5.36	73.2	63.2
	32	17.79	57.4	3.36	69.4	5.51	72.6	3.04	72.4	75.4
1962	33	26.61	66.6	3.12	69.1	6.27	71.6	4.31	72.5	76.0
Averages:	17	19.09	60.8	4.85	70.3	3.55	75.2	3.82	73.2	50.0

Requirements. Construct a predictive model for corn yield (bu/acre). Comment on the relative importance of the predictor variables concerned and suggest what further investigations could be made.

I. The density of a finished product is an important performance characteristic. It can be controlled to a large extent by four main manufacturing variables:

X_1 = the amount of water in the product mix,

X_2 = the amount of reworked material in the product mix,

X_3 = the temperature of the mix,

X_4 = the air temperature in the drying chamber.

In addition, the raw material received from the supplier is important to the process, and a measure of its quality is X_5, the temperature rise. The following data were collected:

X_1	X_2	X_3	X_4	X_5	Y
0	800	135	578	13.195	104
0	800	135	578	13.195	102
0	800	135	578	13.195	100
0	800	135	578	13.195	96
0	800	135	578	13.195	93
0	800	135	578	13.195	103
0	800	150	585	13.180	118
0	800	150	585	13.180	113
0	800	150	585	13.180	107
0	800	150	585	13.180	114
0	800	150	585	13.180	110
0	800	150	585	13.180	114
0	1000	135	590	13.440	97
0	1000	135	590	13.440	87
0	1000	135	590	13.440	92
0	1000	135	590	13.440	85
0	1000	135	590	13.440	94
0	1000	135	590	13.440	102
0	1000	150	590	13.600	104
0	1000	150	590	13.600	102
0	1000	150	590	13.600	101
0	1000	150	590	13.600	104
0	1000	150	590	13.600	98
0	1000	150	590	13.600	101
75	800	135	550	12.745	103
75	800	135	550	12.745	111
75	800	135	550	12.745	111
75	800	135	550	12.745	107
75	800	135	550	12.745	112
75	800	135	550	12.745	106
75	800	150	595	13.885	111
75	800	150	595	13.885	107
75	800	150	595	13.885	104
75	800	150	595	13.885	103
75	800	150	595	13.885	104
75	800	150	595	13.885	103
75	1000	135	530	11.705	116
75	1000	135	530	11.705	108
75	1000	135	530	11.705	104
75	1000	135	530	11.705	116
75	1000	135	530	11.705	116
75	1000	135	530	11.705	112
75	1000	150	590	13.835	111
75	1000	150	590	13.835	110
75	1000	150	590	13.835	115
75	1000	150	590	13.835	114
75	1000	150	590	13.835	114
75	1000	150	590	13.835	114

Requirements

1. Examine the data carefully and state preliminary conclusions.
2. Estimate the coefficients in the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \epsilon.$$

3. Is the above model adequate?
4. Would you recommend an alternative model?
5. Draw some general conclusions about this experiment.

J. An experiment was conducted to determine the effect of six factors on rate. From the data shown below, develop a prediction equation for rate, and use it to propose a set of operating values that may provide an increased rate.

X_1	X_2	X_3	X_4	X_5	X_6	Y (Rate)
149	66	−15	150	105	383	267
143	66	−5	115	105	383	269
149	73	−5	150	105	383	230
143	73	−15	115	105	383	233
149	73	−15	115	78	383	222
143	66	−15	150	78	383	267
143	73	−5	150	78	383	231
149	66	−5	115	78	383	260
149	73	−15	150	78	196	238
149	66	−5	150	78	196	262
143	73	−5	115	78	196	252
143	66	−15	115	78	196	263
143	66	−5	150	105	196	263
149	73	−5	115	105	196	236
149	66	−15	115	105	196	268
143	73	−15	150	105	196	242

K. The data below consist of 16 indexed observations on rubber consumptions from 1948 to 1963. Use these data to develop suitable fitted equations for predicting Y_1 and Y_2 separately in terms of the predictor variables X_1, X_2, X_3, and X_4.

Observation Number	Total Rubber Consumption Y_1	Tire Rubber Consumption Y_2	Car Production X_1	Gross national Product X_2	Disposable Personal Income X_3	Motor Fuel Consumption X_4
1	0.909	0.871	1.287	0.984	0.987	1.046
2	1.252	1.220	1.281	1.078	1.064	1.081
3	0.947	0.975	0.787	1.061	1.007	1.051
4	1.022	1.021	0.796	1.013	1.012	1.046
5	1.044	1.002	1.392	1.028	1.029	1.036
6	0.905	0.890	0.893	0.969	0.993	1.020
7	1.219	1.213	1.400	1.057	1.047	1.057
8	0.923	0.918	0.721	1.001	1.024	1.034
9	1.001	1.014	1.032	0.996	1.003	1.014
10	0.916	0.914	0.685	0.972	0.993	1.013
11	1.173	1.170	1.291	1.046	1.027	1.037
12	0.938	0.952	1.170	1.004	1.001	1.007
13	0.965	0.946	0.817	1.002	1.014	1.008
14	1.106	1.096	1.231	1.049	1.032	1.024
15	1.011	0.999	1.086	1.023	1.020	1.030
16	1.080	1.093	1.001	1.035	1.053	1.029

TABLE L. Student Questionnaire Averages for 12 Instructors

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y
Instructor No.	Course No.	Well Org.	Log. Devel.	Questions Helpful	Out of Class	Text book	Fair Exams	Overall Grade
1	201	4.46	4.42	4.23	4.10	4.56	4.37	4.11
2	224	4.11	3.82	3.29	3.60	3.99	3.82	3.38
3	301	3.58	3.31	3.24	3.76	4.39	3.75	3.17
4	301	4.42	4.37	4.34	4.40	3.63	4.27	4.39
5	301	4.62	4.47	4.53	4.67	4.63	4.57	4.69
6	309	3.18	3.82	3.92	3.62	3.50	4.14	3.25
7	311	2.47	2.79	3.58	3.50	2.84	3.84	2.84
8	311	4.29	3.92	4.05	3.76	2.76	4.11	3.95
9	312	4.41	4.36	4.27	4.75	4.59	4.41	4.18
10	312	4.59	4.34	4.24	4.39	2.64	4.38	4.44
11	333	4.55	4.45	4.43	4.57	4.45	4.40	4.47
3	351	3.71	3.41	3.39	4.18	4.06	4.06	3.17
4	411	4.28	4.45	4.10	4.07	3.76	4.43	4.15
9	424	4.24	4.38	4.35	4.48	4.15	4.50	4.33
12	424	4.67	4.64	4.52	4.39	3.48	4.21	4.61

L. (*Source*: Steven Lemire. We are grateful to Steve for making this study available to us.) Table L, reproduced by Associated Students of Madison at the University of Wisconsin, shows average scores on a 1–5 scale (5 is the most favorable response) for 12 teachers in a particular University of Wisconsin Department in a single semester, based on results from student questionnaires in certain low-numbered courses. (Note that three teachers taught two courses each.) The response Y is the average overall grade on the student questionnaires and is often quoted alone as the crucial number for the teacher's performance. The predictor X_1 is the course number, and X_2 to X_7 are scores relating to a specific question as follows:

X_2: Were the lectures well organized?

X_3: Was the subject matter developed logically?

X_4: Were responses to questions helpful?

X_5: Was out-of-class contact helpful?

X_6: Was the textbook useful?

X_7: Were examinations graded fairly?

For Y, the question was: Overall the teacher was ____? with 1 = poor, 3 = good, 5 = excellent. Only integer responses 1, 2, 3, 4, or 5 could be made to all questions.

1. Using a selection procedure of your choice, or otherwise, find a useful equation that relates Y to $X_1, X_2, \ldots, X_7$, or to some subset of these X's.
2. Do any X's appear to be superfluous in the questionnaire? Explain how you reached that conclusion.
3. Your final equation may or may not imply a causal relationship. Suppose it did, however. What advice would you give to the teachers, to enable them to improve their scores in the future?

M. Refer to the table of Exercise D and then answer the following questions.

1. If variables 4 and 5 were already in the model, which variable would be the best one to add? Does it make a "significant" (at $\alpha = 0.05$) contribution?
2. If variables 1, 2, 3, 4, and 5 were already in the model, which would be a candidate for removal. At the $\alpha = 0.05$ level, would you remove it?
3. What is R^2 for the model with variables 1, 2, 3, and 4 in it?

4. What is the C_p statistic value for the model with variables 3, 4, 5, and 6 in it?
5. What is s^2 for the model with variables 1, 2, 4, and 6 in it?

N. A set of data given on p. 454 of K. A. Brownlee's *Statistical Theory and Methodology in Science and Engineering* (2nd edition), 1965, Wiley, New York, is reproduced below. The figures are from 21 days of operation of a plant oxidizing NH_3 to HNO_3 and the variables are:

X_1 = rate of operation,
X_2 = temperature of the cooling water in the coils of the absorbing tower for the nitric oxides,
X_3 = concentration of HNO_3 in the absorbing liquid (coded by minus 50, times 10),
Y = the percent of the ingoing NH_3 that is lost by escaping in the unabsorbed nitric oxides ($\times 10$). This is an inverse measure of the yield of HNO_3 for the plant.

Data from Operation of a Plant for the Oxidation of Ammonia to Nitric Acid

Run Number	Air Flow X_1	Cooling Water Inlet Temperature X_2	Acid Concentration X_3	Stack Loss Y
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

Investigation of plant operations indicates that the following sets of runs can be considered as replicates: (1, 2), (4, 5, 6), (7, 8), (11, 12), and (18, 19). While the runs in each set are not *exact* replicates, the points *are* sufficiently close to each other in the X-space for them to be used as such.

Requirements

1. Use a selection procedure to pick a planar model for these data and check for lack of fit.
2. Examine the residuals and comment.

O. This is a practical exercise on selecting the "best" regression equation. Coffee extract sold in the stores is obtained from coffee beans via a certain industrial process. An important response variable is

$$Y = \text{coffee density},$$

and it is anticipated that it may depend on six predictor variables,

TABLE O. Residual Sums of Squares (RSS) for All Possible Fitted Models, to Three Decimal Places

Variables[a]	RSS	Variables[a]	RSS	Variables[a]	RSS
—	41.685	123	27.359	1234	27.199
1	40.085	124	31.242	1235	14.084
2	31.302	125	15.987	1236	27.305
3	32.663	126	30.080	1245	15.777
4	41.659	134	31.063	1246	29.747
5	33.817	135	24.746	1256	15.979
6	34.457	136	31.562	1345	23.176
12	31.276	145	31.066	1346	30.482
13	31.855	146	32.833	1356	24.680
14	39.425	156	28.017	1456	25.537
15	32.482	234	27.213	2345	13.885
16	34.437	235	14.688	2346	27.206
23	27.366	236	27.305	2356	12.695
24	31.245	245	16.081	2456	16.078
25	16.748	246	29.760	3456	23.886
26	30.251	256	16.684	12345	13.727
34	32.503	345	24.805	12346	27.189
35	25.395	346	31.211	12356	12.542
36	31.862	356	25.035	12456	15.773
45	33.517	456	26.050	13456	22.884
46	33.178			23456	12.587
56	28.042			123456	12.508

[a] In the equation; a β_0 term is included in all equations.

X_1 = static pressure,
X_2 = air inlet temperature,
X_3 = air outlet temperature,
X_4 = extract pump pressure,
X_5 = extract pump concentration,
X_6 = extract temperature,

through the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \epsilon. \tag{1}$$

A set of 26 observations on $(X_1, X_2, \ldots, X_6, Y)$ is available and all possible regressions have been run; see Table O.

You may assume that model (1) is "complete" in the sense that it includes all the relevant terms. Your assignment is to select a suitable subset of these terms as the "best" regression equation in the circumstances.

You may try up to 23 different equations. For each subset of variables that you specify, you will be given (only) the residual sum of squares for that regression. This information will permit you to use any of the following selection procedures: backward elimination, forward selection, stepwise regression, examination of C_p, or variations of the above. You should now decide on the method of your choice and proceed. Remember that, to test for entry or exit of a single variable at any stage, you can use this formula for the test statistic F:

$$\frac{\{(\text{Residual SS with that variable out}) - (\text{Residual SS with that variable in})\}/1}{s^2 \text{ from larger model with that variable in}}.$$

The appropriate degrees of freedom for the test are (1, df of s^2).

(*Hint*: The residual for the regression containing *no* predictors but only β_0 will give you the corrected total sum of squares.)

TABLE P. Kilgo Data

x_1 Pressure (bar)	x_2 Temperature (°C)	x_3 Moisture (% by wt.)	x_4 Flow Rate (liters/min)	x_5 Average Size (mm)	Y Solubility
−1	−1	−1	−1	−1	29.2
1	−1	−1	−1	1	23.0
−1	1	−1	−1	1	37.0
1	1	−1	−1	−1	139.7
−1	−1	1	−1	1	23.3
1	−1	1	−1	−1	38.3
−1	1	1	−1	−1	42.6
1	1	1	−1	1	141.4
−1	−1	−1	1	1	22.4
1	−1	−1	1	−1	37.2
−1	1	−1	1	−1	31.3
1	1	−1	1	1	48.6
−1	−1	1	1	−1	22.9
1	−1	1	1	1	36.2
−1	1	1	1	1	33.6
1	1	1	1	−1	172.6

Source: The table is adapted from pp. 48 and 49 of the source reference by courtesy of Marcel Dekker, Inc.

P. (*Source*: "An application of fractional factorial experimental design." by M. B. Kilgo, *Quality Engineering,* **1,** 1988–89, 45–54.) Results from a 2_V^{5-1} fractional factorial design are shown in Table P in so-called standard order. The levels of the predictor variables $x_1, x_2, \ldots x_5$ are all coded to ± 1 and the response Y is the "solubility" measured in an experiment involving the extraction of oil from peanuts. These data are probably best analyzed by the standard factorial analysis, but consider the following selection procedure approach. Consider the predictor variables $x_1, x_2, \ldots x_5, x_1x_2, x_1x_3, \ldots x_4x_5$ and fit a useful linear model via the selection procedure of your choice. When you have decided on a suitable model, check the residuals and then recheck the data themselves. Notice anything? If yes, suggest a remedy, and then fit your chosen model again. What are your conclusions? (*Hint*: One of the observations may be faulty.)

Q. (*Source*: Dr. Jim Douglas, University of New South Wales, Sydney, Australia.) Because of a number of road alterations in Sydney, Dr. Douglas was faced with the problem of choosing between alternative routes for driving into the University in the most attractive way. The criteria were not altogether well defined, but certainly involved short travel time, short total distance, small exposure to fast multi-lane traffic, and pleasant surroundings. Dr. Douglas chose Y = travel time in minutes as the response variable and his predictors were specially coded in the following manner: $X_1 = 1, 2$ for into and out of UNSW; $X_2 = 0, 1$ for no school holiday; X_3 = time of departure, coded to reflect the way accident reports relate to the time of day; X_4 = day of week, coded 0 for Sunday, 1.7 for Monday and Tuesday, 2.3 for Wednesday, 3.6 for Thursday, 7 for Friday and 3.9 for Saturday; these numbers were proportional to the average daily numbers of accidents; X_5 = coded weather; X_6 = route; X_7 = number of persons in car. Nineteen observations, shown on page 565, were obtained on these variables. The runs were not fully randomized, said Dr. Douglas; "If it's a beautiful day, how can I not drive past Botany Bay!" A perfectly reasonable decision! Analyze the data using any of the selection procedures or analyses you have learned, and state your conclusions.