

CHAPTER 6

SELECTING THE "BEST" REGRESSION EQUATION

6.0. Introduction

We shall defer discussion of the general model building process to Chapter 8, and in this chapter deal only with the use of 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 "independent" or predictor variables X_1, X_2, \dots, X_k . Suppose further that Z_1, Z_2, \dots, 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 want our model to include as many Z 's as possible so that reliable fitted values can be determined.
2. 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 compromise between these extremes is what is usually called *selecting the best regression equation*. There is no unique statistical procedure for doing this. 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 never in this position, so a great deal of personal judgment will be a necessary part of any of the methods discussed. In this chapter we shall describe several

procedures which have been proposed; all of these appear to be in current use. 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 Mallows' C_p ; (2) best subset regressions using R^2 , R^2 (adjusted), and C_p ; (3) backward elimination, (4) stepwise regression, (5) some variations on previous methods, (6) ridge regression, (7) PRESS, (8) principal components regression, (9) latent root regression, and (10) stagewise regression. After each discussion, we state our personal opinion.

Some Cautionary Remarks on the Use of Unplanned Data

When we do regression calculations on unplanned data (that is, data arising from continuing operations and not from a designed experiment) some potentially dangerous possibilities can arise, as discussed by G. E. P. Box in "Use and abuse of regression," *Technometrics*, 8, 1966, 625-629. 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. (He calls these *latent* or *turking* variables.) Due to the possibilities of bias in the estimates, discussed in Section 2.12, 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 "nonsignificant," a conclusion which 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 (for example "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.

6.1. All Possible Regressions

This procedure is a rather cumbersome one and is quite impossible without access to a high-speed computer. Thus it has come into use only since fast computers have become generally available. The procedure first requires the fitting of every possible regression equation which involves Z_0 plus any number of the variables Z_1, \dots, 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 $Z_i, i = 1, 2, \dots, r$ ($r=Z$'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 we shall discuss are

1. the value of R^2 achieved by the least squares fit,
2. the value of s^2 , the residual mean square, and
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 an example. We shall use the data in a four-variable ($k = 4$) problem given by A. Hald on p. 647 of his book *Statistical Theory with Engineering Applications*, published by Wiley, New York, in 1952. This particular problem has been chosen because it illustrates some typical difficulties which occur in regression analysis. The data are given on the (retyped) computer sheets in Appendix B. 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 $Y = X_5$. A β_0 term is *always* included. Thus there are $2^4 = 16$ possible regression equations, which involve X_0 and the $X_i, i = 1, 2, 3, 4$. These all appear in Appendix B. We now apply the procedures mentioned above.

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 $E(Y) = \beta_0$).

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 .

3. Examine the leaders and see if there is any consistent pattern of variables in the leading equations in each set. For this example, the leaders in each set are:

Set	Variables in Equation	100R ² %
B	$\hat{Y} = f(X_4)$	67.5%
C	$\hat{Y} = f(X_1, X_2)$	97.9%
	$\hat{Y} = f(X_1, X_4)$	97.2%
D	$\hat{Y} = f(X_1, X_2, X_4)$	98.234%
E	$\hat{Y} = f(X_1, X_2, X_3, X_4)$	98.237%

(Notice that in Set C there are two leaders with practically the same size R^2 value.) If we view these results 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 B) reveals that (X_1 and X_3) and (X_2 and X_4) are highly correlated, since (rounding to three decimal places)

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

Thus the addition of further variables when X_1 and X_2 or when X_3 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 the slight increase in R^2 from 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 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 always be added to enable a decision to be made.

The (Algol 60) Algorithm AS 38, "Best subset search," by M. J. Garside, *Applied Statistics*, 20, 1971, 112-115 "provides a fast search through all possible subsets of a regression model to find those subsets with the largest coefficient of multiple correlation. The method is described fully" by Garside in the same issue pp. 8-15.

Use of the Residual Mean Square, s^2

If all regressions are done on a large problem, an assessment of the average magnitude 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. 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 sheets in Appendix B.

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*	47.00
4	5.35, 5.33, 5.65, 8.20	6.13
5	5.98	5.98

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

When the number of potential variables in the model is large, r greater than ten 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, and the number of observations greatly exceeds the number of variables in the fitted equation—five to ten times as many as indicated above. This situation is illustrated in Figure 6.1.

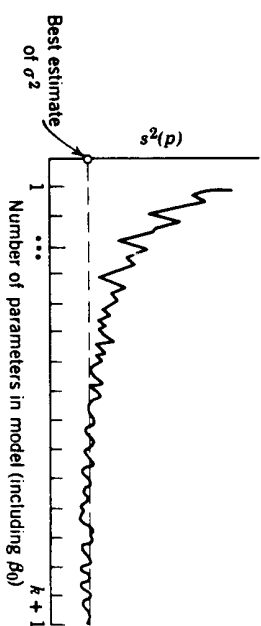


Figure 6.1 Overfitting, showing typical stabilization of s^2 .

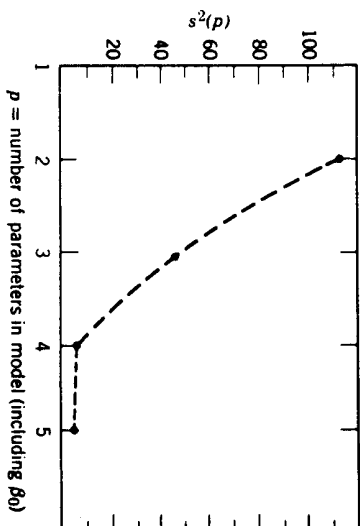


Figure 6.2 Plot of average residual mean square against p .

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 6.2. 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.

Use of the Mallows C_p Statistic

An alternative statistic which has gained popularity in recent years is the C_p statistic, initially suggested by C. L. Mallows. This has the form

$$C_p = \text{RSS}_p / s^2 - (n - 2p) \tag{6.1.1}$$

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 . As R. W. Kennard 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; see Eqs. (6.1.1), (2.6.11b), and (2.6.11a). Now, if an equation with p parameters is adequate, that is, does not suffer from lack of fit, $E(\text{RSS}_p) = (n - p)\sigma^2$. 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 (that is $C_p \neq p$) 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 judgement that apply in such cases will be found in *Fitting Equations to Data*, by C. Daniel and F. S. Wood, assisted by J. W. Gorman, 2nd edition, Wiley, New York, 1980, and "Selection of variables for fitting equations to data," by J. W. Gorman and R. J. Toman, *Technometrics*, 8, 1966, 27-51. Read also "Some comments on C_p " by C. L. Mallows, *Technometrics*, 15, 1973, 661-675. 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 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.

EXAMPLE OF USE OF THE C_p STATISTIC. For the Hald data (see Appendix B) 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 6.1. Note that, for the equation with all predictors in, $C_p = p$, as must be true by definition, because in this case $RSS_p = (n - p)s^2$.

A plot of the smaller C_p statistics is given in Figure 6.3. 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. 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 would have 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.

GENERAL REMARKS. It was mentioned earlier that the data used in this example contained a theoretical restriction $X_1 + X_2 + X_3 + X_4 = \text{constant}$. This means that X_4 is, theoretically, dependent on X_1 , X_2 , and X_3 . Thus if all four X 's are included in the model, the $X'X$ matrix would, theoretically,

Table 6.1 Values of C_p and p for the Hald Data Equations

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

Example. For the equation with predictors X_2 and X_4 we look for 24 in left column: the corresponding values of C_p and p are 138.2 and 3, respectively.

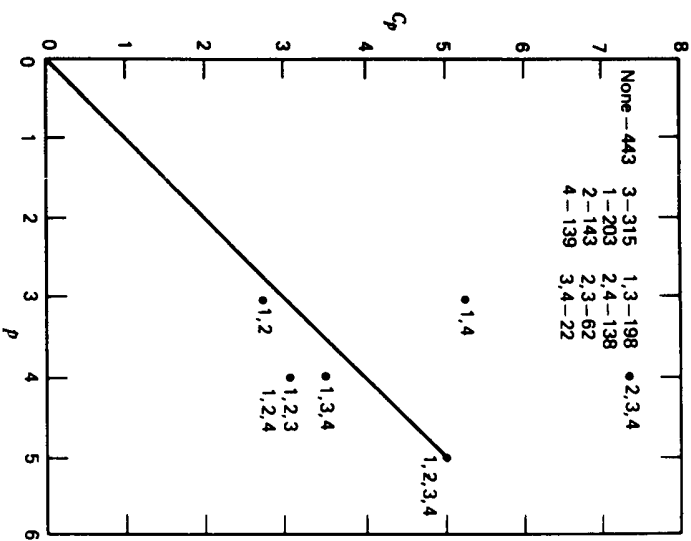


Figure 6.3 C_p plot for Hald's data.

be singular and have zero determinant, both before and after transformation. As we see from the regression printout, p. 659 the transformed determinant actually has the very small value 0.0010677. When the 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. (See Section 5.5.)

Selected references on various aspects of the all regressions procedure are listed in the bibliography.

OPINION. In general the analysis of all regressions is quite unwarranted. While it means that the statistician has "looked at all possibilities" it also means he has examined a large number of regression equations which 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 which shortens this task is preferable.

6.2. "Best Subset" Regression

There now exist excellent computer algorithms for selecting best subsets of predictor variables in regression. A popular one is that given by G. M. Furnival and R. W. Wilson in "Regression by leaps and bounds." *Technometrics*, 16, 1974, 499-511, which computes only a fraction of all possible regressions in determining the "best K" subsets. Three criteria may be applied for determining these "best K" subsets, namely,

1. Maximum R^2 ,
2. Maximum adjusted R^2 (see Eqn. [2.6.11b]).
3. Mallows C_p statistic.

In the BMDP set of programs (see p. 344) is one entitled P9R, *All Possible Subsets Regression*. The user specifies the number, K, of best subsets wanted, and the criterion, and the program then produces the "best K" subsets out of all possible regressions. (The printout shows all three statistics but chooses the subsets on the basis of the criterion selected.) The program also provides the "best K" subsets with one predictor variable, the "best K" subsets with two predictor variables, and so on up to the single equation "subset" with all predictors included. In each of these subsets, an equation which belongs to the "best K" overall subset is so designated. If the selected value of K exceeds the number of equations available for a subset, then all the ($< K$) possible equations will be displayed. This will be clear from the Hald data (see Appendix B) example below, where we specify $K = 5$. The values of all three criteria are shown, but the subsets are chosen on the basis of the values of C_p . Finally, the program points out the statistics of the "best" of all the "best K" subsets.

Using the Hald data, specifying $K = 5$, the following reproduced printouts show the BMDP9R results.

NOTE THAT THE CRITERIA (R-SQUARED, ADJUSTED R-SQUARED AND CP), THE REGRESSION COEFFICIENTS, AND THEIR T-STATISTICS ARE REPORTED FOR THE 5 BEST SUBSETS. THE CRITERIA ARE EVALUATED FOR MANY OTHER SUBSETS, SOME OF WHICH MAY ALSO BE QUITE GOOD. THESE OTHER SUBSETS ARE NOT NECESSARILY BETTER THAN ANY SUBSET WHICH HAS NOT BEEN PRINTED.

```

***** REGRESSIONS WITH 1 VARIABLES *****
R-SQUARED  ADJ R-SQ  CP  VARIABLES IN SUBSET
0.67454      0.64495  138.73  4
0.66627      0.63593  142.49  2
0.53395      0.49158  202.55  1
0.28587      0.22095  315.15  3
    
```

***** REGRESSIONS WITH 2 VARIABLES *****

R-SQUARED ADJ R-SO CP VARIABLES IN SUBSET
 0.97868 0.97441 2.68 1 2

THIS IS ONE OF THE 5 BEST SUBSETS.
 VARIABLE COEFFICIENT T-STATISTIC
 1 X1 0.146831D 01 12.10
 2 X2 0.662250D 00 14.44
 INTERCEPT 0.525773D 02

0.97247 0.96697 5.50 1 4
 0.93529 0.92235 22.37 3 4
 0.68006 0.61607 138.23 2 4
 0.54823 0.45787 198.07 1 3

***** REGRESSIONS WITH 3 VARIABLES *****

R-SQUARED ADJ R-SO CP VARIABLES IN SUBSET
 0.98234 0.97645 3.02 1 2 4

THIS IS ONE OF THE 5 BEST SUBSETS.
 VARIABLE COEFFICIENT T-STATISTIC
 1 X1 0.145194D 01 12.41
 2 X2 0.416110D 00 2.24
 4 X4 -0.236540D 00 -1.37
 INTERCEPT 0.716483D 02

0.98228 0.97638 3.04 1 2 3

THIS IS ONE OF THE 5 BEST SUBSETS.
 VARIABLE COEFFICIENT T-STATISTIC
 1 X1 0.169589D 01 8.29
 2 X2 0.656915D 00 14.85
 3 X3 0.250017D 00 1.35
 INTERCEPT 0.481936D 02

0.98128 0.97504 3.50 1 3 4

THIS IS ONE OF THE 5 BEST SUBSETS.
 VARIABLE COEFFICIENT T-STATISTIC
 1 X1 0.105185D 01 4.70
 3 X3 -0.410043D 00 -2.06
 4 X4 -0.642796D 00 -14.43
 INTERCEPT 0.111684D 03

0.97282 0.96376 7.34 2 3 4

***** REGRESSIONS WITH 4 VARIABLES *****

R-SQUARED ADJ R-SO CP VARIABLES IN SUBSET
 0.98238 0.97356 5.00 1 2 3 4

THIS IS ONE OF THE 5 BEST SUBSETS.
 VARIABLE COEFFICIENT T-STATISTIC
 1 X1 0.155110D 01 2.08
 2 X2 0.510170D 00 0.70
 3 X3 0.101911D 00 0.14
 4 X4 -0.144059D 00 -0.20
 INTERCEPT 0.624052D 02

10 REGRESSIONS COMPUTED IN THE PROCESS OF FINDING BEST SUBSETS.
 38 MULTIPLICATIONS AND DIVISIONS WERE USED IN FINDING BEST SUBSETS
 (NOT INCLUDING COMPUTATION OF COVARIANCE MATRIX).

STATISTICS FOR 'BEST' SUBSET		2.68					
MALLOWS' CP		0.97868					
SQUARED MULTIPLE CORRELATION		0.98928					
MULTIPLE CORRELATION		0.97441					
ADJUSTED SQUARED MULT. CORR.		0.579044D 01					
RESIDUAL MEAN SQUARE		0.240633D 01					
STANDARD ERROR OF EST.		229.50					
F-STATISTIC		2					
NUMERATOR DEGREES OF FREEDOM		10					
DENOMINATOR DEGREES OF FREEDOM		0.0000					
SIGNIFICANCE							
VARIABLE NO.	NAME	REGRESSION COEFFICIENT	STANDARD ERROR	STAND. COEF.	T-STAT.	2TAIL SIG.	TOLERANCE
1	INTERCEPT	0.525773D 02	0.228617D 01	3.495	23.00	0.000	0.947751
1	X1	0.146831D 01	0.121301D 00	0.574	12.10	0.000	0.947751
2	X2	0.662250D 00	0.458547D-01	0.685	14.44	0.000	0.947751

OPINION. There are some drawbacks to this procedure: (1) It tends to provide (in 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 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 of great value, and we recommend its use in conjunction with the stepwise method, if examination of equations "around" the "best" is desired.

6.3 The Backward Elimination Procedure

The backward elimination method is more economical than the "all regressions" method in the sense that 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 variable¹ treated as though it were the last variable to enter the regression equation.

¹ 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.

3. The lowest partial F -test value, F_L , say, is compared with a prespecified significance level F_0 , say.

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.

First, do the complete regression on all predictor variables. In the example referred to in Section 6.1, we thus find the least squares equation $\hat{Y} = f(X_1, X_2, X_3, X_4)$. The analysis for this model is shown in Appendix B, p. 659. As long as the $X'X$ matrix is not singular, the residual error variance obtained here is a good estimate of σ^2 in the "asymptotic" sense discussed in relation to Figure 6.1. The backward elimination procedure essentially attempts to remove all unneeded X -variables without substantially increasing the size of this "asymptotic" estimate of σ^2 .² In order to check the variables at this stage, one must determine the contribution of each of X_1 , X_2 , X_3 , and X_4 to the regression sum of squares as if each were in the last position. The partial F -values shown in the last column of this printout provide measures of these contributions.

We now choose the smallest partial F -value and compare it to a critical value of F based on a predetermined α -risk. In this case, the critical F -value for, say, $\alpha = 0.10$ is $F(1, 8, 0.90) = 3.46$. The smallest partial F is for variable, X_3 ; namely calculated $F = 0.018$. Since the calculated F is smaller than the critical value 3.46 we reject X_3 .

Next, find the least squares equation, $\hat{Y} = f(X_1, X_2, X_4)$. This is shown on p. 653. 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 $F(1, 9, 0.90) = 3.36$. Because the partial F associated with X_4 is 1.86 (which is less than 3.36), we remove X_4 .

We now find the least squares equation $\hat{Y} = f(X_1, X_2)$, shown on p. 639. This provides a statistically significant overall equation with an F -value of 229.50 which exceeds $F(2, 10, 0.999) = 14.91$. Both variables X_1 and X_2 are

significant regardless of position, that is, each partial F -value exceeds 14.91. Thus the backward elimination selection procedure is terminated and yields the equation:

$$\hat{Y} = 52.58 + 1.47X_1 + 0.66X_2$$

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 time and manpower than the "all regressions" method. However, if the input data yields an $X'X$ matrix which 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 recognize that once a variable has been eliminated in this procedure it is gone forever. Thus all alternative models using eliminated variables are not available for possible examination.

Note. We recapitulate points made earlier in the text which are relevant here.

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 $F(1, v)$ is an F -variable with 1 and v degrees of freedom, and if $t(v)$ is a t -variable with v degrees of freedom, then $F(1, v) = t^2(v)$. (See p. 102.)
2. Some programs use the words "F to remove" in their computer printouts. This is exactly the same as the partial F . (See p. 102.)

6.4. The Stepwise Regression Procedure

The backward elimination method begins with the largest regression, using all variables, and subsequently reduces the number of variables in the equation until a decision is reached on the equation to use. The stepwise selection procedure is an attempt to achieve a similar conclusion working from the other direction, that is, to insert variables in turn until the regression equation is satisfactory. The order of insertion is determined by using the partial correlation coefficient as a measure of the importance of variables not yet in the equation. 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 $Y = \bar{Y}$ as best; otherwise we search for the second predictor variable to enter regression. We examine the partial

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.

correlation coefficients³ of all the predictors not in regression at this stage, namely $Z_j, j \neq 1$, with Y ; that is, Y and Z_j are both adjusted for their straight-line relationships with Z_1 , and the correlation between these adjusted variables is calculated for all $j \neq 1$. Mathematically this is equivalent to finding the correlations between (1) the residuals from the regression $\hat{Y} = f(Z_1)$ and (2) the residuals from each of the j regressions $\hat{Z}_j = f_j(Z_1)$ (which we have not actually performed). The Z_j with the highest partial correlation coefficient with Y is now selected, (suppose this is Z_2) 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 entered⁴) are examined. The lower of these two partial F 's is then compared with an appropriate F percentage point, 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 pre-selected percentage point of the appropriate F -distribution. This provides a judgement 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 correlation with Y 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

³ Many regression computing packages do not evaluate and print out the partial correlation coefficients or their squares, as we discuss above. Instead, they compute the analogous statistic, F to enter, for each predictor not in the model at any stage. This list yields essentially identical information, the largest F to enter being associated with the next entry candidate.

⁴ 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.

cycling effect⁵), when no variables in the current equation can be removed and the next best candidate variable cannot hold its place in 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. (Refer to the printout where indicated and recall that $Y = X_5$ and $Z_i = X_i$ here for $i = 1, 2, 3, 4$.) Both entry and exit tests will be made at the level $\alpha = 0.10$.

1. Calculate the correlations of all the predictor variables with the response. Select, as the first variable to enter the regression, the one most highly correlated with the response. Examination of the correlation matrix in Appendix B shows that X_4 is most highly correlated with the response Y or X_5 ; $r_{45} = -0.821$. Thus X_4 is the first variable to enter the regression equation.

2. Regress Y on X_4 and obtain the least squares equation shown on p. 637. The overall F -test shows that the regression equation is significant. We retain X_4 .

3. Calculate the partial correlation coefficients of all variables not in regression with the response. Their squares are shown at the bottom of p. 637. Choose as the next variable to enter into the regression the one with the highest partial correlation coefficient. This is variable X_1 ; $r_{15.4}^2 = 0.915$.

4. With X_1 as well as X_4 in the regression, the least squares equation $Y = f(X_4, X_1)$ is that shown on p. 643. This equation has a percentage R^2 of 97.2% and is significant, since we have an overall $F = 176.63$ which exceeds $F(2, 10, 0.90) = 4.10$; that the new variable X_1 provides a significant decrease in the residual sum of squares is shown by its partial F -value, 108.22, which exceeds $F(1, 10, 0.90) = 4.96$. We retain X_1 . We also check the contribution X_4 would have made if X_1 had been entered first and X_4 entered second. Because the value of the partial F (shown on p. 643) is 159.295, which greatly exceeds $F(1, 10, 0.90) = 4.96$, X_4 is retained. (In practice, most programs do not test both variables quite like this, but instead look for the smallest of the partial F -values and test that. The decision is made to either reject or retain the corresponding predictor and the equation is recomputed or the next candidate is sought, respectively.)

⁵ It is usually best to choose the same significance levels for the entry and exit tests. If a smaller α is chosen for the exit test than 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. See also pp. 310-312.

CHAPTER XIV

Additional Topics in Regression: Variable Selection and Collinearity

Christman, R. (1987).

*Plane Answers to Complex Questions: The Theory
of Linear Models.*
Springer

Suppose we have a set of variables y, x_1, \dots, x_s and observations on these variables $y_i, x_{i1}, x_{i2}, \dots, x_{is}, i = 1, \dots, n$. We want to identify which of the independent variables, x_i , are important for a regression on y . There are several methods available for doing this.

Obviously, the most complete method is to look at all possible regression equations involving x_1, \dots, x_s . There are 2^s of these. Even if one has the money to compute all of them it may be very difficult to assimilate that much information. Tests for the adequacy of various models can be based on general linear model theory, assuming of course that the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_s x_{is} + e_i \quad (1)$$

is an adequate model for the data.

A more efficient procedure than computing all possible regressions is to choose a criterion for ranking how well different models fit and computing only the best fitting models. Typically, one would want to identify several of the best fitting models and investigate them further. Computing costs for this "best subset regression" method are still considerable.

An older group of methods are the stepwise regression methods. These methods consider the efficacy of adding or deleting individual variables to a model that is currently under consideration. These methods have the fault of considering variables only one at a time. For example, there is no reason to believe that the best two variables to add to a model are the one variable that adds most to the model and then the one variable that adds the most to this augmented model. The fault of stepwise procedures is also their virtue. Because computations go one variable at a time, they are relatively cheap.

Collinearity or multicollinearity refers to the problem, in regression analysis, of having the columns of the design matrix nearly linearly dependent. Ideally,

this is no problem at all. There are numerical difficulties associated with the actual computations, but there are no theoretical difficulties. If, however, one has any doubts about the accuracy of the design matrix, the analysis could be in deep trouble.

Section 4 discusses what collinearity is and what problems it can cause. Four techniques for dealing with collinearity are examined. These are regression in canonical form, principal component regression, generalized inverse regression, and classical ridge regression. The methods, other than classical ridge regression, are essentially the same. The chapter closes with some additional comments of the potential benefits of biased estimation.

Generally in this book, the term mean square error (MSE) has denoted the quantity $Y'(I - M)Y/(I - M)$. This is a sample quantity; a function of the data. In Chapters VI and XII, when discussing prediction, we needed a theoretical concept of the mean square error. Fortunately, up until this point we have not needed to discuss both the sample quantity and the theoretical one at the same time. To discuss variable selection methods and techniques for dealing with collinearity, we will need both concepts simultaneously. To reduce confusion, we will refer to $Y'(I - M)Y/r(I - M)$ as the residual mean square (RMS) and $Y'(I - M)Y$ as the residual sum of squares (RSS). Since $Y'(I - M)Y = [(I - M)Y]'[(I - M)Y]$ is the sum of the squared residuals, this is a very natural nomenclature.

XIV.1. All Possible Regressions and Best

Subset Regression

There is very little to say about the all possible regressions technique. The efficient computation of all possible regressions is due to Schatzoff, Tsao, and Fienberg (1968). Their algorithm was a major advance. Further advances have made this method obsolete. It is just a waste of money to compute all possible regressions. One should only compute those regressions that consist of the best subsets of the independent variables.

The efficient computation of the best regressions is due to Furnival and Wilson (1974). "Best" is defined by ranking models on the basis of some measure of how well they fit. The most commonly used of these measures are R^2 , adjusted R^2 and Mallows' C_p . These criteria will be discussed in the subsections below.

R^2

The coefficient of determination, R^2 , was discussed in Section VI.4. It is defined as

$$R^2 = \frac{SS_{Reg}}{SSTot - C},$$

and is just the ratio of the variability in y explained by the regression to the total variability of y . R^2 measures how well a regression model fits the data as compared to just fitting a mean to the data. If one has two models with, say, p independent variables, other things being equal, the model with the higher R^2 will be the better model.

Using R^2 is not a valid way to compare models with different numbers of independent variables. The R^2 for the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + e_i \quad (1)$$

must be less than the R^2 for the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \beta_{p+1} x_{i,p+1} + \dots + \beta_q x_{iq} + e_i. \quad (2)$$

The second model has all the variables in the first model plus more so

$$\text{SSRReg}(1) \leq \text{SSRReg}(2),$$

and

$$R^2(1) \leq R^2(2).$$

Typically, if the R^2 criterion is chosen, a program for doing best subset regression will print out the models with the highest R^2 for each possible value of the number of independent variables. It is the use of the R^2 criterion in best subset regression that makes computing all possible regressions obsolete. The R^2 criterion fits all the good models one could ever want. In fact, it probably fits too many models.

Adjusted R^2

The adjusted R^2 is a modification of R^2 so that it can be used to compare models with different numbers of independent variables. For a model with $p - 1$ independent variables plus an intercept, the adjusted R^2 is defined as

$$\text{Adj } R^2 = 1 - \frac{n-1}{n-p} (1 - R^2).$$

(With the intercept there are a total of p variables in the model.)

Define $s_e^2 = (\text{SSTot} - C)/(n - 1)$, then s_e^2 is just the sample variance of the y_i 's ignoring any regression structure. It is easily seen (see Exercise 14.1) that

$$\text{Adj } R^2 = 1 - (\text{RMS}/s_e^2).$$

The best models based on the $\text{Adj } R^2$ criterion are those models with the smallest residual mean square.

As a method of identifying sets of good models, this is very attractive. The models with the smallest residual mean square should be among the best models. However, the model with the smallest residual mean square may very well not be the best model.

Consider the question of deleting one variable from a model. If the F for

testing that variable is greater than one, then deleting the variable will increase the residual mean square. By the adjusted R^2 criterion the variable should not be deleted. In fact, unless the F value is substantially greater than one, the variable probably should be deleted. The $\text{Adj } R^2$ criterion tends to include too many variables in the model.

Mallows' C_p

Suppose we have a model that is assumed to be correct, say $Y = X\beta + e$. In the regression set up, this is the model with all the independent variables $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + e_i$. Our problem is that some of the β_i 's may be zero. Rather than merely trying to identify which β_i 's are zero, Mallows suggested that the appropriate criterion for evaluating a reduced model $Y = X_0 \gamma + e$ is by its mean square error for estimating $X\beta$, i.e.,

$$E[(X_0 \hat{\gamma} - X\beta)'(X_0 \hat{\gamma} - X\beta)].$$

As mentioned earlier, to distinguish between this use of the term mean square error and the estimate of the variance in a linear model with $E(Y) = X\beta$, we refer to $Y'(I - M)Y$ as the residual sum of squares ($\text{RSS}(\beta)$) and $Y'(I - M)Y/(I - M)$ as the residual mean square ($\text{RMS}(\beta)$). The statistics $\text{RSS}(\hat{\gamma})$ and $\text{RMS}(\hat{\gamma})$ are the corresponding quantities for the model $Y = X_0 \hat{\gamma} + e$.

The quantity

$$(X_0 \hat{\gamma} - X\beta)'(X_0 \hat{\gamma} - X\beta)$$

is a quadratic form in the vector $(X_0 \hat{\gamma} - X\beta)$. Writing

$$M_0 = X_0(X_0'X_0)^{-1}X_0'$$

gives

$$(X_0 \hat{\gamma} - X\beta) = M_0 Y - X\beta,$$

$$E(X_0 \hat{\gamma} - X\beta) = M_0 X\beta - X\beta = -(I - M_0)X\beta,$$

$$\text{Cov}(X_0 \hat{\gamma} - X\beta) = \sigma^2 M_0.$$

From Theorem 1.3.2

$$E[(X_0 \hat{\gamma} - X\beta)'(X_0 \hat{\gamma} - X\beta)] = \sigma^2 \text{tr}(M_0) + \beta'X'(I - M_0)X\beta.$$

We do not know σ^2 or β , but we can estimate the mean square error. First note that

$$E[Y'(I - M_0)Y] = \sigma^2 \text{tr}(I - M_0) + \beta'X'(I - M_0)X\beta,$$

so

$$\begin{aligned} E[(X_0 \hat{\gamma} - X\beta)'(X_0 \hat{\gamma} - X\beta)] \\ = \sigma^2 [\text{tr}(M_0) - \text{tr}(I - M_0)] + E[Y'(I - M_0)Y]. \end{aligned}$$

With $p = tr(M_0)$, a natural estimate of the mean square error is

$$RMS(\beta)[2p - n] + RSS(y).$$

Mallows' C_p statistic simply rescales the estimated mean square error,

$$C_p = \frac{RSS(y)}{RMS(\beta)} - (n - 2p).$$

The models with the smallest values of C_p have the smallest estimated mean square error and should be among the best models for the data.

EXERCISE 14.1 Show that $\text{Adj } R^2 = 1 - (RMS/s_e^2)$.

EXERCISE 14.2 Consider the two regression models (1) and (2). If F is the F statistic for testing (1) against (2), show that $F > 1$ if and only if the $\text{Adj } R^2$ for model (1) is less than the $\text{Adj } R^2$ for model (2).

EXERCISE 14.3 Give an informal argument to show that if $Y = X_0\gamma + e$ is a correct model, then the value of C_p should be around p . Provide a formal argument for this fact. Show that if $n - s > 2$ then $E(C_p) = p + 2(s - p)/(n - s - 2)$. To do this you need to know that if $W \sim F(u, v, 0)$, then $E(W) = v/(v - 2)$ for $v > 2$. For large values of n (relative to s and p), what is the approximate value of $E(C_p)$?

EXERCISE 14.4 Consider the F statistic for testing model (1) against model (14.0.1): (a) show that $C_p = (s - p)F + (2p - s)$; (b) show that, for a given value of p , the R^2 , $\text{Adj } R^2$, and C_p criteria all induce the same rankings of models.

XIV.2. Stepwise Regression

Forward Selection

Forward selection sequentially adds variables to the model. Since this is a sequential procedure, the model in question is constantly changing. At any stage in the selection process, forward selection adds the variable that:

- (1) has the highest partial correlation,
- (2) increases R^2 the most,
- (3) gives the largest absolute t or F statistic.

These criteria are equivalent.

EXAMPLE 14.2.1. Suppose we have variables y , x_1 , x_2 , and x_3 and the current model is

$$y_i = \beta_0 + \beta_1 x_{i1} + e_i.$$

We must choose between adding variables x_2 and x_3 . Fit the models

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i;$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + e_i.$$

Choose the model with the higher R^2 . Equivalently one could look at the t (or F) statistics for testing $H_0: \beta_2 = 0$ and $H_0: \beta_3 = 0$ and choose the model that gives the larger absolute value of the statistic. Finally, one could look at $r_{y,2,1}$ and $r_{y,3,1}$ and pick the variable that gives the larger absolute value for the partial correlation.

EXERCISE 14.5 Show that these three criteria for selecting a variable are equivalent.

Forward selection stops adding variables when one of three things happens:

- (1) p^* variables have been added,
- (2) all absolute t statistics for adding variables not in the model are less than t^* ,
- (3) the tolerance is too small for all variables not in the model.

The user picks the values of p^* and t^* . Tolerance is discussed in the next subsection. No variable is ever added if its tolerance is too small, regardless of its absolute t statistic.

The forward selection process is often started with the initial model

$$y_i = \beta_0 + e_i.$$

Tolerance

Regression assumes that the design matrix in $Y = X\beta + e$ has full rank. Mathematically, either the columns of X are linearly independent or they are not. In practice, computational difficulties arise if the columns of X are nearly linearly dependent. By nearly linearly dependent, we mean that one column of X can be nearly reproduced by the other columns.

Suppose we have the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + e_i,$$

and we are considering adding variable x_p to the model. To check the tolerance, fit

$$x_{ip} = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_{p-1} x_{i,p-1} + e_i. \quad (1)$$

If the R^2 from this model is high, the column vectors, say, J , X_1, \dots, X_p are nearly linearly dependent. The tolerance of x_p (relative to x_1, \dots, x_{p-1}) is defined as the value of $1 - R^2$ for fitting model (1). If the tolerance is too small, variable x_p is not used. Often in a computer program, the user can define which values of the tolerance should be considered too small.

Backwards Elimination

Backwards elimination sequentially deletes variables from the model. At any stage in the selection process, it deletes the variable with the smallest t or F statistic.

Backwards elimination stops deleting variables when:

- (1) p_* variables have been eliminated,
- (2) the smallest absolute t statistic for eliminating a variable is greater than t_* .

The user can usually specify p_* and t_* in a computer program.

The initial model in the backwards elimination procedure is the model with all of the independent variables included,

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_s x_{is} + e_i.$$

Backwards elimination should give an adequate model. We assume that the process is started with an adequate model, and so only variables that add nothing are eliminated. The model arrived at may, however, be far from the most succinct. On the other hand, there is no reason to believe that forward selection gives even an adequate model.

Other Methods

Forward selection is such an obviously faulty method that several improvements have been recommended. These consist of introducing rules for eliminating and exchanging variables. Four rules for adding, deleting, and exchanging variables follow.

- (1) Add the variable with the largest absolute t statistic if that value is greater than t_* .
- (2) Delete the variable with the smallest absolute t statistic if that value is less than t_* .
- (3) A variable not in the model is exchanged for a variable in the model if the exchange increases the R^2 .
- (4) The largest R^2 for each size model considered so far is saved. Delete a variable if the deletion gives a model with R^2 larger than any other model of the same size.

These rules are used in combination. For example,

- 1 then 2,
 - 1 then 2 then 3,
 - 1 then 4,
- or
- 1 then 4 then 3.

Again, no variable is even added if its tolerance is too small.

XIV.3. Discussion of Variable Selection Techniques

Stepwise regression methods are fast, easy, cheap, and readily available. When the number of observations, n , is less than the number of variables, $s + 1$, forward selection or a modification of it is the only available method for variable selection. Backward elimination and best subset regression assume that one can fit the model with all the independent variables included, which is not possible when $n < s + 1$.

There are serious problems with stepwise methods. They do not give the best model (based on any of the criteria we have discussed). In fact, stepwise methods can give models that contain none of the variables that are in the best regressions. This is because, as mentioned earlier, they handle variables one at a time. Another problem is nontechnical. The user of a stepwise regression program will end up with one model. The user may be inclined to think that this is *the* model. It probably is not. In fact, *the* model probably does not exist. Best subset regression programs usually present several of the best models, although the Adjusted R^2 and Mallows' C_p methods do define a unique best model and could be subject to the same problem.

A problem with best subset selection methods is that they tend to give models that appear to be better than they really are. For example, the Adjusted R^2 criterion chooses the model with the smallest RMS. Because one has selected the smallest RMS, the RMS for that model is biased towards being too small. The fit of the model (almost any measure of the fit of a model is related to the RMS) will appear to be better than it is. If one could sample the data over again and fit the same model, the RMS will almost certainly be larger. Perhaps substantially so.

When using Mallows' C_p statistic, one often picks models with the smallest value of C_p . This can be justified by the fact that the model with the smallest C_p is the model with the smallest estimated expected mean square error. However, if the target value of C_p is p (as suggested by Exercise 14.3) it seems to make little sense to pick the model with the smallest C_p . It seems that one should pick models for which C_p is close to p .

The result of Exercise 14.4, that for a fixed number of independent variables all best regression criteria are equivalent, is very interesting. The Adj R^2 and C_p criteria can be viewed as simply different methods of penalizing models that include more variables. The penalty is needed because models with more variables necessarily explain more variation (have higher R^2 's).

Influential observations are a problem in any regression analysis. Variable selection techniques involve fitting lots of models, so the problem of influential observations is multiplied. Recall that an influential observation in one model is not necessarily influential in a different model.

Some statisticians think that the magnitude of the problem of influential observations is so great as to reject all variable selection techniques. They argue that the models arrived at from variable selection techniques depend almost exclusively on the influential observations and have little to do with any real world effects. Most statisticians, however, approve of the judicious

use of variable selection techniques. (But then, by definition, everyone will approve of the *judicious* use of anything.)

John W. Tukey, among others, has emphasized the difference between exploratory and confirmatory data analysis. Briefly, exploratory data analysis (EDA) deals with situations in which you are trying to find out what is going on in a set of data. Confirmatory data analysis is for proving what you already think is going on. EDA frequently involves looking at lots of graphs. Confirmatory data analysis looks at things like tests and confidence intervals. Strictly speaking, you cannot do both exploratory data analysis and confirmatory data analysis on the same set of data.

Variable selection is an exploratory technique. If you know what variables are important you do not need it and should not use it. When you do use variable selection, if the model is fitted with the same set of data that determined the variable selection, then the model you eventually decide on will give biased estimates and invalid tests and confidence intervals. This sounds a lot worse than it is. The biased estimates may very well be better estimates than you could get by refitting with another data set. (This is related to James-Stein estimation, see also Section XIV.7.) The tests and confidence intervals, although not strictly valid, are often reasonable.

One solution to this problem of selecting variables and fitting parameters with the same data is to divide the data into two parts. Do an exploratory analysis on one part and then a confirmatory analysis on the other. To do this well requires a lot of data. It also demonstrates the problem of influential observations. Depending on where the influential observations are, you can get pretty strange results. The PRESS statistic was designed to be used in procedures like this. However, as we have seen, the PRESS statistic is highly sensitive to influential observations.

Finally, a word about R^2 . R^2 is a good statistic for comparing models. That is what we have used it for here. The actual value of R^2 should not be overemphasized. If you have data with a lot of variability, it is possible to have a very good fit to the underlying regression model without having a high R^2 . For example, if the SSE admits a decomposition into pure error and lack of fit, it is possible to have very little lack of fit while having a substantial pure error, so that R^2 is small while the fit is good.

If transformations of the dependent variable y are considered, it is inappropriate to compare R^2 's for models based on different transformations. For example, it is possible for a transformation to increase the R^2 without really increasing the predictive ability of the model. One way to check whether this is happening is to compare the width of confidence intervals for predicted values after transforming them to a common scale.

XIV.4. Defining Collinearity

In this section we define the problem of collinearity. The approach taken is by quantifying the idea of having columns of the design matrix that are "nearly linearly dependent". The effects of near linear dependencies are examined. The

section concludes by establishing the relationship between the definition given here and other commonly used concepts of collinearity.

Suppose we have a regression model

$$Y = X\beta + e, \quad E(e) = 0, \quad \text{Cov}(e) = \sigma^2 I, \quad (1)$$

where Y is $n \times 1$, X is $n \times p$, β is $p \times 1$, and $\text{rank}(X) = p$. The essence of model (1) is that $E(Y) \in C(X)$. Suppose that the design matrix consists of some independent variables, say x_1, x_2, \dots, x_p , that are measured with some small error. A near linear dependence in the observed design matrix X could mean a real linear dependence in the underlying design matrix of variables measured without error. Let X_* be the underlying design matrix. If the columns of X_* are linearly dependent, there exist an infinite number of least squares estimates for the true regression coefficients. If X is nearly linearly dependent, the estimated regression coefficients may not be meaningful and may be highly variable.

The real essence of this particular problem is that $C(X)$ is too large. Generally, we hope that in some sense, $C(X)$ is close to $C(X_*)$. Regression should work well precisely when this is the case. However, when X_* has linearly dependent columns, X typically will not. Thus $r(X) > r(X_*)$. $C(X_*)$ may be close to some proper subspace of $C(X)$, but $C(X)$ has extra dimensions. By pure chance, these extra dimensions could be very good at explaining the Y vector that happened to be observed. In this case we get an apparently good fit that has no real world significance.

The extra dimensions of $C(X)$ are due to the existence of vectors b such that $X_*b = 0$ but $Xb \neq 0$. If the errors in X are small, then Xb should be approximately zero. We would like to say that a vector w in $C(X)$ is ill-defined if there exists b such that $w = Xb$ is approximately zero. w is approximately the zero vector if its length is near zero. Unfortunately, multiplying w by a scalar can increase or decrease the length of the vector arbitrarily, while not changing the direction determined within $C(X)$. To rectify this, we can restrict attention to vectors b with $b'b = 1$ (i.e., length of b is one), or equivalently we make

Definition 14.4.1. A vector $w = Xb$ is said to be ϵ -ill-defined if $w'w/b'b = b'X_*b/b'b < \epsilon$. The matrix X is ϵ -ill-defined if any vector in $C(X)$ is ϵ -ill-defined. We use the terms ill-defined and ill-conditioned interchangeably.

The assumption of a real linear dependence in the X_* matrix is a strong one. We now indicate how that assumption can be weakened. Let $X = X_* + \Delta$ where the elements of Δ are uniformly small errors. Consider the vector Xb . (For simplicity assume $b'b = 1$.) The corresponding direction in the underlying design matrix is X_*b .

Note that $b'X_*Xb = b'X_*X_*b + 2b'\Delta X_*b + b'\Delta\Delta b$. The vector Δb is short, so if Xb and X_*b are of reasonable size they have about the same length. Also

$$b'X_*Xb = b'X_*X_*b + b'X_*\Delta b, \quad (2)$$

Choice of explanatory variables in multiple regression

A2.1 Introduction

In Section 2.1 we set out procedures for analysing a general linear logistic regression in which dependence of a probability of success on explanatory variables is expressed via a linear relation on the logistic scale. Implementation of these methods is straightforward once a set of explanatory variables is chosen for inclusion. In applications, however, a crucial aspect is precisely that choice, it being fairly rarely the case that, for example, theoretical considerations indicate unambiguously the equation to be fitted. The points at issue are similar to those arising with 'ordinary normal theory' empirical multiple regression based on the method of least squares and indeed in other forms of empirical linear regression in generalized linear models. In the present Appendix we discuss these issues in fairly broad terms.

One special aspect of the normal-theory case that does affect the strategy of tackling the analysis of highly balanced sets of data is that exact or nearly exact orthogonality in normal theory implies that estimates of certain parameters are unaffected by the inclusion or exclusion of some other parameters. This makes it feasible to begin the analysis of a balanced design by inspection of a 'full' analysis of variance in which possibly large numbers of main effects and interactions are included. In linear logistic regression however a balanced design leads to only approximate orthogonality of the estimated parameters and it is not always possible to see immediately the precise effect of such inclusion or exclusion. For this reason it is commonly sensible to begin with some relatively simple model and then to examine the need to amplify or indeed simplify the initial model. The criteria for the choice of that starting model as well as for modifying the model in the light of the data become of more pressing concern.

Sections A2.2 and A2.3 deal with type and formation of explanatory variables rather than with the strategy for choice of explanatory variables but these ideas are nevertheless important in the analysis and interpretation of regression models. Also much of the material in Sections A2.4 and A2.5 is not specific to binary data but is given here for completeness of discussion. Reference is made to examples discussed in Chapter 2 of this book.

A2.2 Types of explanatory variable

It is convenient to classify potential explanatory variables in several different ways.

First for purposes of interpretation we may classify explanatory variables as in Section 2.8, namely as

1. treatment or quasi-treatment variables representing aspects which can in principle at least be manipulated (Example 2.10);
2. intrinsic variables measuring aspects characterizing an individual under study or the environment in which the study on an individual is carried out, for example age, socio-economic class (Example 2.18);
3. non-specific variables characterizing broad groupings of individuals; often such groupings are described by names such as blocks, strata and so on (Example 2.16).

The object of study is normally the assessment of the effect of treatments and of possible interaction of treatment effects with variables of type (2) or (3) (Example 2.17).

Of course this division into three types depends on the context and may not be clear-cut. Especially in observational studies some intrinsic variables, such as socio-economic class of individuals, may be surrogates for other more specific properties, such as educational background, and wherever possible more specific variables should, of course, be used.

In a randomized experiment, treatment variables are randomized and intrinsic variables are those measurements made on the individuals before randomization.

In an observational study, the treatments are typically aspects that ideally have been investigated via a randomized experiment, but which in fact were determined in a way outside the investigator's control. Thus in a study of the effect of alcohol consumption during pregnancy on some feature of the infant, randomization is obviously

not feasible with human subjects. Treatment, perhaps better called quasi-treatment, variables are thus measures of alcohol consumption and other matters, such as diet, necessary to define the treatment effect under study, whereas intrinsic variables are mother's age and parity, socio-economic class, etc. If the study were replicated in a number of centres, centres would form a non-specific variable.

A second classification of explanatory variables, relevant in analytical formulation, is by their mathematical structure, according to whether they take

1. a number of qualitatively different levels, such as one of a number of regions of residence;
2. a number of ordered levels, such as the description of the severity of some condition as slight, moderate, severe and very severe;
3. values specified by a reasonably well-defined quantitative scale.

A third classification is into

1. directly measured variables;
2. derived variables, by which we mean both composite variables obtained by taking combinations of measurements or variables such as squares and products of more directly observed quantities.

A2.3 Formation of explanatory variables

In some situations explanatory variables may be entered into a multiple regression equation either in exactly the form in which they are measured or after rescaling: a simple change of units to make all variables have approximately the same standard deviation in the data, and in some cases a change of origin to produce means that are not too large may help avoid numerical instability (Example 2.11). For essentially positive variables a log transformation may be wise (Examples 2.10, 2.11). Care is, however, needed with variables that have a very wide range, especially where very non-linear effects are likely. Thus if in a clinical study age at entry ranged from 60 to 70 years, direct introduction of, say, age - 65 as a quantitative variable would be reasonable, and non-linearity could, if necessary, be checked via a squared term. But if age ranged from 20 to 80 years some grouping of age into a fairly small number of groups and their treatment initially as qualitatively different, as explained below, would protect against strong non-linearity. Again for alcohol con-

sumption in litres per week it would usually be better to work initially with none, slight, moderate, heavy rather than directly with the quantitative measurement; later analysis could refine the initially arbitrary subdivision, if that seemed likely to be fruitful.

For qualitative variables at l levels, the construction of $l-1$ explanatory variables will be needed if the main effect of such a variable is to be represented without prior constraint (Example 2.12). The method of construction is in one sense arbitrary so long as we make the $l-1$ variables linearly independent but the following considerations are helpful.

1. The marginal frequencies of the different levels should be inspected, in particular to avoid giving prominence to levels that occur with very low frequency.
2. If there is a level, say 1, with a very low frequency and its possible merging with another level, say 2, appears possibly sensible, it will be useful to define one variable, say $x_1 = 1$ (level 1), -1 (level 2), 0 (all other levels), so that the resulting estimated parameter provides a test of the reasonableness of the proposed merging. In defining the other $l-2$ variables, the two levels 1 and 2 can then be treated identically.
3. If one level, say 1, is a control, or other natural reference level, or occurs with especially high frequency, it may be sensible to define all the x s relative to 1, i.e. to define x_1, \dots, x_{l-1} by $x_j = 1$ (level $j+1$), -1 (level 1), 0 (otherwise).
4. If the levels are ordered and are of very roughly equal frequency, it may be sensible to define x 's via the standard orthogonal polynomials (Pearson and Hartley, 1966, Table 47) for l equally spaced points, using thus for three levels $-1, 0, 1$; $1, -2, 1$ to define respectively x_1, x_2 .
5. There should be some rough check that the x s are not defined so as to be nearly linearly dependent.
6. Any special arguments indicating contrasts that are likely to be particularly important should, of course, be used in defining the x s.

Interactions are normally best studied in this context by defining products of the x s defining the main effects in question. In exploratory work, the principle that large main effects are on the whole more likely to generate appreciable interactions than small main effects is often helpful. Thus if two qualitative variables have

l_1 and l_2 levels respectively, leading to the definition of $l_1 - 1$ and $l_2 - 1$ explanatory variables, the set of all products of these variables defines the two-factor interaction with $(l_1 - 1)(l_2 - 1)$ degrees of freedom. If it is required to extract a few degrees of freedom from the interaction this can be done via products of component x 's with especially strong interpretations or, failing that, via components that happen to be large.

In some applications to calculate *a priori* combinations of the explanatory variables may be useful. For example: near orthogonality may be achieved by replacing diastolic and systolic blood pressure by the sum and difference of the logs of the two measurements; if a particular feature has been measured in several different ways a composite score may initially be tried. In these cases it will often be wise to test from the data whether the indicated combination appears to have sacrificed information about the response variable under study.

A2.4 Small numbers of explanatory variables

In some applications there may be a reasonably small number of explanatory variables, corresponding to say at most five or six parameters. Unless a treatment effect of primary interest is substantially confounded with variables of no direct interest, there seems little point in trying to simplify the resulting equation by omitting explanatory variables merely on grounds of statistical insignificance; it may be a useful quick check on the potential for improving precision of treatment effects to compare the standard error under the full fit with that achieved by omission of *all* other variables.

The model with all explanatory variables in linear form may be augmented by adding non-linear functions, e.g. squares of quantitative variables, and interaction terms (Example 2.11). A simple strategy is to begin by adding such terms one at a time, concentrating on interactions of the treatment effects of primary interest with intrinsic and non-specific explanatory variables and on possible non-linearity of response to important quantitative variables.

In judging statistical significance it is important to make allowance whenever the largest from among many possible contrasts is chosen for interpretation. One way to do this, when a variable is chosen for inclusion out of a block of variables, is to examine the change in 2 log (maximized likelihood) when all the variables in the block are fitted; if

this is not statistically significant at an interesting level, there is a danger that the variable selected is an artefact.

A2.5 Large numbers of explanatory variables

A much more difficult situation arises if there are so many potential explanatory variables that some reduction from the full fit is essential either to achieve understandable interpretation or reasonable precision in the primary comparisons. Now many computer packages contain automatic algorithms for variable selection. We strongly recommend against reliance on these algorithms, except occasionally in the very restricted context set out below. This is because the choices they force are often of a very arbitrary character and are often not the most appropriate for the purpose either of prediction or of interpretation; to end with one set of variables when there are other quite different choices having virtually as good a fit to the data invites misinterpretation.

We suggest a procedure along broadly the following lines.

1. List those explanatory variables which it is essential to include either because they are treatment variables of primary concern, or because it is known from previous studies that they are important.
2. Consider whether certain subsets of variables (e.g. measurements of the same feature) should be treated separately or whether some preliminary reduction across subsets, such as by the formation of totals, might be fruitful.
3. Check for the influence of other variables, at first one at a time as in Section A2.4, or by cautious use of a computerized selection algorithm.
4. Iterate this procedure, i.e. repeat both phases with the initial variables, those from the initial list supplemented by and/or replaced by other variables found empirically from the data.
5. When one or more apparently adequate fits have been obtained check for the addition of further variables, interactions and so on as outlined in Section A2.4.

If, as is likely, there are different choices in the later phases that give adequate fits it is important, as far as is feasible, to give *all* fits consistent with the data, making any choice between alternative choices on subject-matter grounds.

A crucial aspect is the behaviour under alternative models of the aspects of primary interest, i.e. parameters representing treatment effects and their potentially important interactions with intrinsic variables. If these are reasonably stable, choice of other aspects of the model is probably not of critical importance.

Note especially that important variables, especially those representing treatments, should not be excluded solely because the corresponding estimates are insignificant statistically; their estimation is likely to be of direct interest and the inclusion of estimates and standard errors is in any case likely to be essential in a final report on the data, if only to allow comparison with subsequent related studies.

Where there are several rather similar sets of data for analysis it will usually be wise to use the same explanatory variables for all sets. Thus if the number of such variables is large a cautious procedure is to aim to choose explanatory variables separately for each set and then to re-analyse using the set of *all* variables so chosen, before possibly attempting some common reduction. Incautious use of automatic selection algorithms is quite likely in these contexts to throw up different choices of variables in the different sets, with consequent dangers of misinterpretation.

A final note of caution concerns the interpretation of significance tests and confidence intervals when a complex sequence of data-dependent choices has been made. If there is really very little or no explanatory power in a block of variables and one or two are selected on the basis of the data as showing the largest apparent effect, there is a clear possibility of considerable exaggeration of significance; this stresses the need for some protection from global tests of blocks of parameters.

We do not think it feasible to specify probability properties of complex sequences of data-dependent choices. Note, however, that if *all* sufficiently simple equations consistent with the data at a specified standard significance level are listed, any 'correct' such specification will be included with the specified confidence coefficient.

We consider, however, that by following the broad guidelines above and concentrating on the treatment effects of primary concern, these puzzling difficulties are to a large extent bypassed. If all that is required is a well-fitting empirical equation and no interest attaches to individual effects, again questions of the significance of individual terms are essentially irrelevant, although we regard such totally empirical prediction equations as of rather limited interest.