Chapter 14
Variable Selection

Suppose we have a set of variables $y, x_1, \ldots, x_s$ and observations on these variables $y_1, x_{i1}, \ldots, x_{is}$, $i = 1, \ldots, n$. We want to identify which of the independent variables x_j 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 $x_1, \ldots, x_s$. There are 2^s of these. Even if one has the time and 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

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_s x_{is} + e_i \tag{1}$$

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 compute only the best fitting models. Typically, one would want to identify several of the best fitting models and investigate them further. The computing effort for this "best subset regression" method is still considerable.

An older group of methods is stepwise regression. These methods consider the efficacy of adding or deleting individual variables to a model that is currently under consideration. These methods have the flaw 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 followed by the one variable that adds the most to this augmented model. The flaw of stepwise procedures is also their virtue. Because computations go one variable at a time, they are relatively easy.

In this book, the term "mean squared error" (MSE) has generally denoted the quantity $Y'(I-M)Y/r(I-M)$. This is a sample quantity, a function of the data. In Chapters 6 and 12, when discussing prediction, we needed a theoretical concept of the mean squared 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 of dealing with collinearity, we will need both

R. Christensen, *Plane Answers to Complex Questions: The Theory of Linear Models*,
Springer Texts in Statistics, DOI 10.1007/978-1-4419-9816-3_14,

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.

14.1 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 a waste of money to compute all possible regressions. One should only compute those regressions that consist of the best subsets of the predictor 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's C_p. These criteria are discussed in the subsections that follow.

14.1.1 R^2

The coefficient of determination, R^2, was discussed in Section 6.4. It is computed as

$$R^2 = \frac{SSReg}{SSTot - C}$$

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

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + e_i \tag{1}$$

must be less than the R^2 for the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \beta_{p+1} x_{i\,p+1} + \cdots + \beta_q x_{iq} + e_i\,. \tag{2}$$

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

$$SSReg(1) \leq SSReg(2)$$

and

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

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 predictor 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.

14.1.2 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 predictor variables. For a model with $p-1$ predictor variables plus an intercept, the adjusted R^2 is defined as

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

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

Define $s_y^2 = (SSTot - C)/(n-1)$. Then s_y^2 is the sample variance of the y_is ignoring any regression structure. It is easily seen (Exercise 14.1) that

$$\text{Adj } R^2 = 1 - \frac{RMS}{s_y^2}.$$

The best models based on the Adj R^2 criterion are those models with the smallest residual mean squares.

As a method of identifying sets of good models, this is very attractive. The models with the smallest residual mean squares 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 1, then deleting the variable will increase the residual mean square. By the adjusted R^2 criterion, the variable should not be deleted. However, unless the F value is substantially greater than 1, the variable probably should be deleted. The Adj R^2 criterion tends to include too many variables in the model. (See Exercise 14.2.)

Exercise 14.1 Show that Adj $R^2 = 1 - (RMS/s_y^2)$.

Exercise 14.2 Consider testing the regression model (1) against (2). Show that $F > 1$ if and only if the Adj R^2 for model (1) is less than the Adj R^2 for model (2).

14.1.3 Mallows's C_p

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

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

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

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

$$M_0 = X_0(X_0'X_0)^- X_0'$$

gives

$$\begin{aligned}(X_0\hat{\gamma} - X\beta) &= M_0Y - X\beta,\\ \mathrm{E}(X_0\hat{\gamma} - X\beta) &= M_0X\beta - X\beta = -(I - M_0)X\beta,\\ \mathrm{Cov}(X_0\hat{\gamma} - X\beta) &= \sigma^2 M_0.\end{aligned}$$

From Theorem 1.3.2

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

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

$$\mathrm{E}\left[Y'(I - M_0)Y\right] = \sigma^2 \mathrm{tr}(I - M_0) + \beta'X'(I - M_0)X\beta;$$

so

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

With $p = \mathrm{tr}(M_0)$, an unbiased estimate of the mean squared error is

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

Mallows's C_p statistic simply rescales the estimated mean squared error,

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

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

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 $\mathrm{E}(C_p) = p + 2(s-p)/(n-s-2)$. To do this you need to know that if $W \sim F(u,v,0)$, then $\mathrm{E}(W) = v/(v-2)$ for $v > 2$. For large values of n (relative to s and p), what is the approximate value of $\mathrm{E}(C_p)$?

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

14.2 Stepwise Regression

Stepwise regression methods involve adding or deleting variables one at a time.

14.2.1 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

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \\ y_i &= \beta_0 + \beta_1 x_{i1} + \beta_3 x_{i3} + e_i. \end{aligned}$$

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 $r_{y2\cdot 1}$ and $r_{y3\cdot 1}$ 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

$$y_i = \beta_0 + e_i.$$

14.2.2 Tolerance

Regression assumes that the model 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

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

If the R^2 from this model is high, the column vectors, say $J, X_1, \ldots, X_p$, are nearly linearly dependent. The tolerance of x_p (relative to $x_1, \ldots, x_{p-1}$) 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.

14.2.3 Backward Elimination

Backward elimination sequentially deletes variables from the model. At any stage in the selection process, it deletes the variable with the smallest absolute t or F statistic. Backward 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 backward elimination procedure is the model with all of the predictor variables included,

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

Backward 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.

14.2.4 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 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 can be 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 ever added if its tolerance is too small.

14.3 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 that includes all the predictor variables. This 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. That 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. Even though the Adjusted R^2 and Mallows's C_p methods define a unique best model and could be subject to the same problem, best subset regression programs generally present several of the best models.

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 toward being too small. Almost any measure of the fit of a model is related to the RMS, so the fit of the model will appear to be better than it is. If one could sample the data over again and fit the same model, the RMS would almost certainly be larger, perhaps substantially so.

When using Mallows's 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 squared 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 predictor variables the three 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^2s).

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 15.4.) The test 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 similar to 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 measuring the predictive ability of a model. R^2 is also a good statistic for comparing models. That is what we used it for here. But the actual value of R^2 should not be overemphasized when it is being used to identify correct models (rather than models that are merely useful for prediction). 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 for models based on different transformations. For example, it is possible for a transformation to increase 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.

To compare models based on different transformations of y, say $y_1 = f_1(y)$ and $y_2 = f_2(y)$, fit models to the transformed data to obtained predicted values $\hat{y}_1$ and $\hat{y}_2$. Return these to the original scale with $\tilde{y}_1 = f_1^{-1}(\hat{y}_1)$ and $\tilde{y}_2 = f_2^{-1}(\hat{y}_2)$. Finally, define R_1^2 as the squared sample correlation between the ys and the $\tilde{y}_1$s and define R_2^2 as the squared sample correlation between the ys and the $\tilde{y}_2$s. These R^2 values are comparable.

Exercise 14.6 Mosteller and Tukey (1977) reported data on verbal test scores for sixth graders. They used a sample of 20 Mid-Atlantic and New England schools taken from *The Coleman Report*. The dependent variable y was the mean verbal test score for each school. The predictor variables were: x_1 = staff salaries per pupil, x_2 = percent of sixth grader's fathers employed in white collar jobs, x_3 = a composite score measuring socioeconomic status, x_4 = the mean score on a verbal test administered to teachers, and x_5 = one-half of the sixth grader's mothers' mean number of years of schooling. Compare the results of using the various model selection techniques on the data.

Obs.	x_1	x_2	x_3	x_4	x_5	y
1	3.83	28.87	7.20	26.60	6.19	37.01
2	2.89	20.10	−11.71	24.40	5.17	26.51
3	2.86	69.05	12.32	25.70	7.04	36.51
4	2.92	65.40	14.28	25.70	7.10	40.70
5	3.06	29.59	6.31	25.40	6.15	37.10
6	2.07	44.82	6.16	21.60	6.41	33.90
7	2.52	77.37	12.70	24.90	6.86	41.80
8	2.45	24.67	−0.17	25.01	5.78	33.40
9	3.13	65.01	9.85	26.60	6.51	41.01
10	2.44	9.99	−0.05	28.01	5.57	37.20
11	2.09	12.20	−12.86	23.51	5.62	23.30
12	2.52	22.55	0.92	23.60	5.34	35.20
13	2.22	14.30	4.77	24.51	5.80	34.90
14	2.67	31.79	−0.96	25.80	6.19	33.10
15	2.71	11.60	−16.04	25.20	5.62	22.70
16	3.14	68.47	10.62	25.01	6.94	39.70
17	3.54	42.64	2.66	25.01	6.33	31.80
18	2.52	16.70	−10.99	24.80	6.01	31.70
19	2.68	86.27	15.03	25.51	7.51	43.10
20	2.37	76.73	12.77	24.51	6.96	41.01