Chapter 9

Variable Selection and Model Building

In the preceding chapters we have assumed that the predictors included in the model are known to be important. Our focus was on techniques to ensure that the functional form of the model was correct and that the underlying assumptions were not violated. In this chapter, we consider methods for choosing "best" model from a class of multiple regression models using variable selection methods.

Finding an appropriate subset of predictors for the model is often called the variable selection problem. To provide motivation for variable selection we will start with brief review of the consequences of incorrect model specification.

9.1 Consequences of model misspecification

Before criteria are presented, we will consider what the data analyst faces if he/she underfits (i.e., ignores or deletes relatively important terms) or overfits (i.e., includes model terms that have only marginal, or no, contribution). In what follows, the results are based on use of the procedure of least squares with the assumption that the random errors ϵ_i are uncorrelated with homogeneous variance σ^2.

Assume that there are K candidate predictors $X_1, \ldots, X_K$ and $n(\geq K+1)$ observations on these predictors and the response Y. The full model, containing all K predictors, is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} + \epsilon_i, \quad i = 1, \ldots, n$$

or equivalently

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

where $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$. The full model can be written as

$$\mathbf{y} = \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\epsilon}$$

where the $\mathbf{X}$ matrix has been partitioned into $\mathbf{X}_p$, an $n \times p$ matrix whose columns represent the intercept and the $k(= p-1) < K$ predictors to be retained in the subset model, and $\mathbf{X}_r$,

an $n \times r$ matrix whose columns represent the predictors to be deleted from the full model. Let $\boldsymbol{\beta}$ be partitioned conformably into $\boldsymbol{\beta}_p$ and $\boldsymbol{\beta}_r$. Notice that $K = k + r = p + r - 1$.

Let r be the number of predictors that are deleted from the full model. The subset model retaining the k predictors out of K potential predictors is

$$\mathbf{y} = \mathbf{X}_p\boldsymbol{\beta}_p + \boldsymbol{\epsilon}_*,$$

where $\mathrm{E}[\boldsymbol{\epsilon}_*] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}_*) = \sigma^2\mathbf{I}$.

Impact of under-fitting

We now suppose that the full model is true model, but the model fit to the data was the subset model. Remind that the least-squares estimate of $\boldsymbol{\beta}_p$ is

$$\hat{\boldsymbol{\beta}}_p = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^T = (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{X}_p^T\mathbf{y}$$

and estimate of σ^2 is

$$S_p^2 = \frac{\mathbf{y}^T(\mathbf{I} - \mathbf{H}_p)\mathbf{y}}{n - p}$$

with $\mathbf{H}_p = \mathbf{X}_p(\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{X}_p^T$.

The properties of the estimates $\hat{\boldsymbol{\beta}}_p$ and S_p^2 from the subset model can be summarized as follows:

- $\mathrm{E}[\hat{\boldsymbol{\beta}}_p] = \boldsymbol{\beta}_p + (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{X}_p^T\mathbf{X}_r\boldsymbol{\beta}_r = \boldsymbol{\beta}_p + \mathbf{A}\boldsymbol{\beta}_r$ with $\mathbf{A} = (\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{X}_p^T\mathbf{X}_r$; i.e., $\hat{\boldsymbol{\beta}}_p$ is a biased estimate of $\boldsymbol{\beta}_p$ unless the regression coefficients corresponding to the deleted variables ($\boldsymbol{\beta}_r$) are zero or the retained variables are orthogonal to the deleted variables ($\mathbf{X}_p^T\mathbf{X}_r = 0$).

- $\mathrm{E}[S_p^2] = \sigma^2 + \dfrac{\boldsymbol{\beta}_r^T\mathbf{X}_r^T(\mathbf{I} - \mathbf{H}_p)\mathbf{X}_r\boldsymbol{\beta}_r}{n - p}$. That is, S_p^2 is generally biased upward as an estimate of σ^2.

Suppose that we wish to predict the response at the point $\mathbf{x}_0^T = (\mathbf{x}_{0,p}^T, \mathbf{x}_{0,r}^T)$. If the subset model is used to predict response at the point $\mathbf{x}_{0,p}$, $\hat{y}_{0,p} = \mathbf{x}_{0,p}^T\hat{\boldsymbol{\beta}}_p$, with mean

$$\mathrm{E}[\hat{y}_{0,p}] = \mathbf{x}_{0,p}^T\boldsymbol{\beta}_p + \mathbf{x}_{0,p}^T\mathbf{A}\boldsymbol{\beta}_r$$

and prediction variance

$$\mathrm{Var}(\hat{y}_{0,p}) = \sigma^2(1 + \mathbf{x}_{0,p}^T(\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{x}_{0,p}).$$

However, we know that at the point $\mathbf{x}_0^T$, the mean response, i.e., what one is truly attempting to estimate, is $\mathrm{E}[Y_0] = \mathbf{x}_{0,p}^T\boldsymbol{\beta}_p + \mathbf{x}_{0,r}^T\boldsymbol{\beta}_r$. Thus, the bias in the prediction at $\mathbf{x}_0$ is given by

$$\mathrm{Bias}(\hat{y}_{0,p}) = \mathrm{E}[\hat{y}_{0,p}] - \mathrm{E}[Y_0] = (\mathbf{x}_{0,p}^T\mathbf{A} - \mathbf{x}_{0,r}^T)\boldsymbol{\beta}_r.$$

After some manipulation, it can be shown that

$$\mathrm{E}[S_p^2] = \sigma^2 + \frac{1}{n-p}\sum_{i=1}^{n}\left\{\mathrm{Bias}(\hat{y}_i)\right\}^2.$$

It is obvious that the impact of an underspecified model is fixed bias in the important quantities, the regression coefficients, the predicted responses, and the estimate of error variance. Essentially when we underfits or, say, ignore important predictors, we deposit the variation accounted for by the ignored predictors in the residual sum of squares and hence inflate the residual mean square. This inflation may be viewed as reflecting bias in prediction at the data points.

Impact of over-fitting

We now suppose that the subset model is true model, but the model fit to the data was the full model. Then, for the full model the least-squares estimate of $\boldsymbol{\beta}$ is

$$\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_k, \tilde{\beta}_{k+1}, \ldots, \tilde{\beta}_K)^T = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

and an estimate of the residual variance σ^2 is

$$\tilde{\sigma}^2 = \frac{\mathbf{y}^T(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-K-1}.$$

The important results of model overspecification are summarized as follows:

- For a model containing the predictors $X_1, \ldots, X_k$ with least squares estimates given by $\hat{\boldsymbol{\beta}}_p = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)^T$, if the predictors $X_{k+1}, \ldots, X_K$ are added to the model producing the new regression coefficients $\tilde{\beta}_0, \tilde{\beta}_1, \ldots, \tilde{\beta}_K$, then

  $$\mathrm{E}[\tilde{\beta}_j] = \beta_j,\ j = 0, 1, \ldots, k, \quad \mathrm{E}[\tilde{\beta}_j] = 0,\ j = k+1, \ldots, K,$$

  and

  $$\mathrm{Var}(\tilde{\beta}_j) \geq \mathrm{Var}(\hat{\beta}_j), \quad j = 0, 1, \ldots, k.$$

- $\mathrm{E}[\tilde{\sigma}^2] = \sigma^2$; i.e., even if one overfits, the residual mean square is unbiased for σ^2.

- Suppose that we wish to predict the response at the point $\mathbf{x}_0^T = (\mathbf{x}_{0,p}^T, \mathbf{x}_{0,r}^T)$. For the full model, the predicted response at the point $\mathbf{x}_0^T$ is given by $\tilde{y}_0 = \mathbf{x}_0^T\tilde{\boldsymbol{\beta}} = \tilde{\beta}_0 + \sum_{j=1}^{K}\tilde{\beta}_j x_{0,j}$. For the subset model, the predicted response at the point $\mathbf{x}_{0,p}^T$ is given by $\hat{y}_0 = \mathbf{x}_{0,p}^T\hat{\boldsymbol{\beta}}_p = \hat{\beta}_0 + \sum_{j=1}^{k}\hat{\beta}_j x_{0,j}$. Then,

  $$\mathrm{Var}(\tilde{y}_0) \geq \mathrm{Var}(\hat{y}_0).$$

From the results above, we should realize that the overspecified model produces results involving variances that are larger than those for the simpler model while the underspecified model results in important estimated quantities being biased. That is, in general, the more predictors included in a valid model the lower the bias of the predictions but the higher variance.

The process of finding a model that is a compromise between overfitting and underfitting is called the variable selection problem. The two key aspects of variable selection methods are generating the subset models and deciding if one subset is better than another. We begin by discussing criteria for evaluating and comparing subset regression models.

9.2 Criteria for evaluating subset regression models

We shall discuss various criteria for evaluating subsets of predictors.

Criterion 1 : R^2_{adj}

Recall that R^2, the coefficient of determination of the regression model, is defined as the proportion of variation in the observations Y's explained by the regression model, that is,

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}.$$

Remind that R^2 increases as the number of predictors increases and hence adding irrelevant predictors to the model increases R^2. Thus, it is not straight forward to use R^2 as a criterion for choosing the number of predictors to include in the model. However, R^2_p can be used to compare the models including a fixed number of predictors p. So, it is difficult to judge whether an increase in R^2 is really telling us anything important. To avoid the difficulties of interpreting R^2, some analysts prefer to use the adjusted R^2 statistic, defined for a model with $k(= p-1)$ predictors as

$$R^2_{adj,p} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)}.$$

The $R^2_{adj,p}$ statistic does not necessarily increase as additional predictors are introduced into the model. In fact, it can be shown that if s predictors are added to the model, $R^2_{adj,p+s}$ will exceed $R^2_{adj,p}$ if and only if the partial F statistic for testing the significance of the s additional predictors exceeds 1. Consequently, one criterion for selection of an optimum subset model is to choose the model that has a maximum $R^2_{adj,p}$.

Criterion 2 : MS_{Res}

Let $MS_{Res}(p) = \frac{SS_{Res}(p)}{n-p}$ denote the residual mean square for a subset regression model with p terms. Because $SS_{Res}(p)$ always decreases as p increases, $MS_{Res}(p)$ initially decreases, then

stabilizes, and eventually may increase. One criterion for selection of an optimum subset model is to choose the model that has a minimum $MS_{Res}(p)$. It can be shown that the criteria minimum $MS_{Res}(p)$ and maximum $R^2_{adj,p}$ are equivalent.

Criterion 3 : Mallow's C_p

Mallow (1964) proposed a criterion that is related to the mean square error of a fitted value, that is,

$$\mathrm{E}[\hat{y}_i - \mathrm{E}[y_i]]^2 = (\mathrm{E}[y_i] - \mathrm{E}[\hat{y}_i])^2 + \mathrm{Var}(\hat{y}_i).$$

Let the total squared bias for a p-term regression model be

$$SS_B(p) = \sum_{i=1}^{n} (\mathrm{E}[y_i] - \mathrm{E}[\hat{y}_i])^2$$

and define the standardized total mean square error as

$$\begin{aligned}
\Gamma_p &= \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}(\mathrm{E}[y_i] - \mathrm{E}[\hat{y}_i])^2 + \sum_{i=1}^{n}\mathrm{Var}(\hat{y}_i)\right\} \\
&= \frac{SS_B(p)}{\sigma^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n}\mathrm{Var}(\hat{y}_i) \\
&= \frac{\mathrm{E}[SS_{Res}(p)]}{\sigma^2} - n + 2p
\end{aligned}$$

since $\sum_{i=1}^{n}\mathrm{Var}(\hat{y}_i) = p\sigma^2$ and $\mathrm{E}[SS_{Res}(p)] = SS_B(p) + (n-p)\sigma^2$. Then, Mallow's C_p statistic is defined as

$$C_p = \frac{SS_{Res}(p)}{\hat{\sigma}^2} - n + 2p.$$

If the p-term model has negligible bias, then $SS_B(p) = 0$. Consequently, $\mathrm{E}[SS_{Res}(p)] = (n-p)\sigma^2$ and

$$\mathrm{E}[C_p|\mathrm{Bias} = 0] = p.$$

One criterion for selection of an optimum subset model is to choose the model with a small value of C_p. See Figure 9.3 in page 269 of textbook. When using the C_p criterion, it can be helpful to visualize the plot of C_p as a function of p for each regression model.

Criterion 4 : PRESS

The criterion that we use for model selection should certainly be related to the intended use of the model. There are several possible uses of regression, including (1) data description, (2) prediction and estimation, (3) parameter estimation, and (4) control. Frequently, regression

equations are used for prediction of future observations or estimation of the mean response. If the objective is in prediction, the PRESS statistic can be used:

$$\text{PRESS}_p = \sum_{i=1}^{n}(y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2.$$

One then select the subset regression model based on a small value of PRESS_p.

Criterion 5 : AIC (Akaike's Information Criterion)

Suppose that the observations $y_1, \ldots, y_n$ are independently and normally distributed with mean $\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ and variance σ^2. Then, the likelihood function is given by

$$L(\beta_0, \beta_1, \ldots, \beta_k, \sigma^2 | \mathbf{X}, \mathbf{y}) = \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 - \frac{n}{2}\log(2\pi\sigma^2)\right]$$

and hence the log-likelihood function is given by

$$\log(L(\beta_0, \beta_1, \ldots, \beta_k, \sigma^2 | \mathbf{X}, \mathbf{y})) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 - \frac{n}{2}\log(2\pi\sigma^2).$$

Substituting the maximum likelihood estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k, \hat{\sigma}^2$ for $\beta_0, \beta_1, \ldots, \beta_k, \sigma^2$ into the log likelihood the log likelihood associated with the maximum likelihood estimates is given by

$$\log(L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k, \hat{\sigma}^2 | \mathbf{X}, \mathbf{y})) = -\frac{n}{2} - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{SS_{Res}}{n}\right).$$

The most popular motivation of Akaike's information criterion (AIC) seems to be based on balancing goodness of fit (measured by SS_{Res}) and a penalty for model complexity (measured by m). **AIC is defined such that the smaller the value of AIC the better the model**. AIC is defined to be

$$\text{AIC} = 2\left[-\log(L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k, \hat{\sigma}^2 | \mathbf{X}, \mathbf{y})) + m\right]$$

where $m = k + 2$ since $\beta_0, \beta_1, \ldots, \beta_k, \sigma^2$ are estimated in the fitted model. The measure of complexity is necessary since adding irrelevant predictors to the regression model can increase the log likelihood. Then,

$$\text{AIC} = 2\left[\frac{n}{2} + \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\left(\frac{SS_{Res}}{n}\right) + (k+2)\right] = n\log\left(\frac{SS_{Res}}{n}\right) + 2k + \text{other terms}.$$

Criterion 6 : BIC (Bayesian Information Criterion)

Schwarz (1968) proposed the Bayesian Information criterion as

$$\text{BIC} = -2\log(L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k, \hat{\sigma}^2|\mathbf{X}, \mathbf{y})) + m\log(n)$$

where $m = k + 2$, the number of parameters estimated in the model. **BIC is defined such that the smaller the value of BIC the better the model**. BIC is similar to AIC except that the factor 2 in the penalty term is replaced by $\log(n)$. When $n \geq 8$, $\log(n) > 2$ and so the penalty term in BIC is greater than that in AIC. Thus, in this circumstances, BIC penalizes complex models more heavily than AIC, thus favoring simpler models than AIC.

9.3 Computational Techniques for Variable Selection

To find the subset of variables to use in the final model, it is natural to consider fitting models with various combinations of the candidate predictors. In this section we will discuss several computational techniques for generating subset regression models and illustrate criteria for evaluation of these models.

9.3.1 All possible regressions

This procedure requires that the analyst fit all the regression models involving one candidate predictor, two candidate predictors, and so on. The models are evaluated according to some suitable criterion (R^2_{adj}, MS_{Res}, C_p, AIC, BIC, etc.) and the "best" regression model selected. If there are K candidate predictors then there are 2^K total models to be estimated and examined.

Example

See the Hald Cement Data in Ex 9.1 of textbook.

Hald (1952) presents data concerning the heat evolved in calories per gram of cement (Y) as a function of the amount of each of four ingredients in the mix: tricalcium aluminate (X_1), tricalcium silicate (X_2), tetracalcium alumino ferrite (X_3), and dicalcium silicate (X_4).

SAS Program

```
option ls=90 ps=75;
title 'Hald Cement Data';
/* all possible regression method */
proc reg data=cement;
model y = x1-x4 / selection=adjrsq; /* criterion: adj R square */
run;
```

```
proc reg data=cement;
model y = x1-x4 / selection=cp mse; /* criterion: Mallow Cp */
run;
proc reg data=cement;
model y = x1-x4 / selection=adjrsq mse aic bic;
run;
```

Output

Adjusted R-Square Selection Method

Number of Observations Read 13
Number of Observations Used 13

Number in Model	Adjusted R-Square	R-Square	Variables in Model
3	0.9764	0.9823	x1 x2 x4
3	0.9764	0.9823	x1 x2 x3
3	0.9750	0.9813	x1 x3 x4
2	0.9744	0.9787	x1 x2
4	0.9736	0.9824	x1 x2 x3 x4
2	0.9670	0.9725	x1 x4
3	0.9638	0.9728	x2 x3 x4
2	0.9223	0.9353	x3 x4
2	0.8164	0.8470	x2 x3
1	0.6450	0.6745	x4
1	0.6359	0.6663	x2
2	0.6161	0.6801	x2 x4
1	0.4916	0.5339	x1
2	0.4578	0.5482	x1 x3
1	0.2210	0.2859	x3

C(p) Selection Method

Number of Observations Read 13
Number of Observations Used 13

Number in Model	C(p)	R-Square	MSE	Variables in Model
2	2.6782	0.9787	5.79045	x1 x2
3	3.0182	0.9823	5.33030	x1 x2 x4
3	3.0413	0.9823	5.34562	x1 x2 x3
3	3.4968	0.9813	5.64846	x1 x3 x4
4	5.0000	0.9824	5.98295	x1 x2 x3 x4
2	5.4959	0.9725	7.47621	x1 x4
3	7.3375	0.9728	8.20162	x2 x3 x4

```
         2      22.3731      0.9353       17.57380    x3 x4
         2      62.4377      0.8470       41.54427    x2 x3
         2     138.2259      0.6801       86.88801    x2 x4
         1     138.7308      0.6745       80.35154    x4
         1     142.4864      0.6663       82.39421    x2
         2     198.0947      0.5482      122.70721    x1 x3
         1     202.5488      0.5339      115.06243    x1
         1     315.1543      0.2859      176.30913    x3
-------------------------------------------------------------------------------

                    Adjusted R-Square Selection Method

                  Number of Observations Read          13
                  Number of Observations Used          13

Number in  Adjusted
  Model    R-Square  R-Square         AIC         BIC          MSE  Variables in Model

      3     0.9764    0.9823     24.9739     31.1723      5.33030  x1 x2 x4
      3     0.9764    0.9823     25.0112     31.1839      5.34562  x1 x2 x3
      3     0.9750    0.9813     25.7276     31.4057      5.64846  x1 x3 x4
      2     0.9744    0.9787     25.4200     29.2437      5.79045  x1 x2
      4     0.9736    0.9824     26.9443     34.4130      5.98295  x1 x2 x3 x4
      2     0.9670    0.9725     28.7417     30.9805      7.47621  x1 x4
      3     0.9638    0.9728     30.5759     32.9997      8.20162  x2 x3 x4
      2     0.9223    0.9353     39.8526     37.8866     17.57380  x3 x4
      2     0.8164    0.8470     51.0371     46.8392     41.54427  x2 x3
      1     0.6450    0.6745     58.8516     55.5401     80.35154  x4
      1     0.6359    0.6663     59.1780     55.8498     82.39421  x2
      2     0.6161    0.6801     60.6293     55.5085     86.88801  x2 x4
      1     0.4916    0.5339     63.5195     60.0035    115.06243  x1
      2     0.4578    0.5482     65.1167     59.7425    122.70721  x1 x3
      1     0.2210    0.2859     69.0674     65.3850    176.30913  x3
```

9.3.2 Stepwise regression methods

Because evaluating all possible regressions can be burdensome computationally, various methods have been developed for evaluating only a small number of subset regression models by either adding or deleting predictors one at a time. These methods are generally referred to as stepwise-type procedures. They can be classified into three broad categories:

- Forward selection,
- Backward elimination,
- Stepwise regression : a combination of forward selection and backward elimination.

Forward selection

The forward selection technique begins with no variables in the model. For each of the predictors, this method calculates F statistics that reflect the predictor's contribution to the model if it is included. The P-values for these F statistics are compared to a preselected significance level α_{IN}. If no F statistic has a significance level greater than α_{IN}, the forward selection stops. Otherwise, the forward method adds the predictor that has the largest F statistic to the model. The forward method then calculates F statistics again for the predictors still remaining outside the model, and the evaluation process is repeated. Thus, predictors are added one by one to the model until no remaining predictor produces a significant F statistic. Once a variable is in the model, it stays.
The first predictor selected for entry into the model is the one that has the largest correlation coefficient with the response variable Y.

Note that for forward selection the default for preselected significance level α_{IN} is .5 in SAS.

Backward elimination

The backward elimination technique begins by calculating F statistics for a model including all potential predictors. Then the predictors are deleted from the model one by one until all the predictors remaining in the model produce F statistics significant at a preselected significance level α_{OUT}. At each step, the predictor showing the smallest contribution to the model is deleted.

Note that for backward elimination the default for preselected significance level α_{OUT} is .1 in SAS.

Stepwise regression

The stepwise regression technique is a modification of the forward selection technique and differs in that predictors already in the model do not necessarily stay there. As in the forward selection method, predictors are added one by one to the model, and the F statistic for a predictor to be added must be significant at α_{IN}. After a predictor is added, however, the stepwise method looks at all the predictors already included in the model and deletes any predictor that does not produce an F statistic significant at α_{OUT}. Only after this check is made and the necessary deletions are accomplished another predictor can be added to the model. The stepwise process ends when none of the predictors outside the model has an F statistic significant at α_{IN} and every predictor in the model is significant at α_{OUT}, or when the predictor to be added to the model is the one just deleted from it.

Note that for stepwise regression the defaults for preselected significance levels α_{IN} and α_{OUT} are .15 in SAS.

Note: Forward selection and backward elimination consider at most $K + (K-1) + (K-2) + \cdots + 1 = K(K+1)/2$ of the 2^K possible predictor subsets. Thus, forward selection and

backward elimination do not necessarily find the optimum subset model across all 2^K possible predictor subsets. In addition, there is no guarantee that forward selection, backward elimination, and stepwise regression will produce the same final model.

Example

Recall the Hald Cement Data in textbook.

SAS Program

```
/* stepwise method */
proc reg data=cement;
model y = x1-x4 / selection=forward;
run;
proc reg data=cement;
model y = x1-x4 / selection=backward;
run;
proc reg data=cement;
model y = x1-x4 / selection=stepwise;
run;
```

Output

```
/* Forward Selection */
                              Forward Selection: Step 1

               Variable x4 Entered: R-Square = 0.6745 and C(p) = 138.7308

                                 Analysis of Variance

                                          Sum of           Mean
     Source                   DF         Squares         Square    F Value    Pr > F

     Model                     1      1831.89616     1831.89616      22.80    0.0006
     Error                    11       883.86692       80.35154
     Corrected Total          12      2715.76308

                         Parameter     Standard
           Variable       Estimate        Error   Type II SS  F Value  Pr > F

           Intercept     117.56793      5.26221        40108   499.16  <.0001
           x4             -0.73816      0.15460   1831.89616    22.80  0.0006

                            Bounds on condition number: 1, 1
--------------------------------------------------------------------------------
```

Forward Selection: Step 2

Variable x1 Entered: R-Square = 0.9725 and C(p) = 5.4959

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2641.00096	1320.50048	176.63	<.0001
Error	10	74.76211	7.47621		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	103.09738	2.12398	17615	2356.10	<.0001
x1	1.43996	0.13842	809.10480	108.22	<.0001
x4	-0.61395	0.04864	1190.92464	159.30	<.0001

Bounds on condition number: 1.0641, 4.2564

Forward Selection: Step 3

Variable x2 Entered: R-Square = 0.9823 and C(p) = 3.0182

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2667.79035	889.26345	166.83	<.0001
Error	9	47.97273	5.33030		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	71.64831	14.14239	136.81003	25.67	0.0007
x1	1.45194	0.11700	820.90740	154.01	<.0001
x2	0.41611	0.18561	26.78938	5.03	0.0517
x4	-0.23654	0.17329	9.93175	1.86	0.2054

Bounds on condition number: 18.94, 116.36

No other variable met the 0.5000 significance level for entry into the model.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4	1	0.6745	0.6745	138.731	22.80	0.0006
2	x1	2	0.2979	0.9725	5.4959	108.22	<.0001
3	x2	3	0.0099	0.9823	3.0182	5.03	0.0517

```
/* Backward Elimination */
```

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.9824 and C(p) = 5.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2667.89944	666.97486	111.48	<.0001
Error	8	47.86364	5.98295		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	62.40537	70.07096	4.74552	0.79	0.3991
x1	1.55110	0.74477	25.95091	4.34	0.0708
x2	0.51017	0.72379	2.97248	0.50	0.5009
x3	0.10191	0.75471	0.10909	0.02	0.8959
x4	-0.14406	0.70905	0.24697	0.04	0.8441

Bounds on condition number: 282.51, 2489.2

Backward Elimination: Step 1

Variable x3 Removed: R-Square = 0.9823 and C(p) = 3.0182

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2667.79035	889.26345	166.83	<.0001
Error	9	47.97273	5.33030		
Corrected Total	12	2715.76308			

	Parameter	Standard

Variable	Estimate	Error	Type II SS	F Value	Pr > F
Intercept	71.64831	14.14239	136.81003	25.67	0.0007
x1	1.45194	0.11700	820.90740	154.01	<.0001
x2	0.41611	0.18561	26.78938	5.03	0.0517
x4	-0.23654	0.17329	9.93175	1.86	0.2054

Bounds on condition number: 18.94, 116.36

Backward Elimination: Step 2

Variable x4 Removed: R-Square = 0.9787 and C(p) = 2.6782

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2657.85859	1328.92930	229.50	<.0001
Error	10	57.90448	5.79045		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	52.57735	2.28617	3062.60416	528.91	<.0001
x1	1.46831	0.12130	848.43186	146.52	<.0001
x2	0.66225	0.04585	1207.78227	208.58	<.0001

Bounds on condition number: 1.0551, 4.2205

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x3	3	0.0000	0.9823	3.0182	0.02	0.8959
2	x4	2	0.0037	0.9787	2.6782	1.86	0.2054

```
/* Stepwise regression */
```

Stepwise Selection: Step 1

Variable x4 Entered: R-Square = 0.6745 and C(p) = 138.7308

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1831.89616	1831.89616	22.80	0.0006
Error	11	883.86692	80.35154		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	117.56793	5.26221	40108	499.16	<.0001
x4	-0.73816	0.15460	1831.89616	22.80	0.0006

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable x1 Entered: R-Square = 0.9725 and C(p) = 5.4959

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2641.00096	1320.50048	176.63	<.0001
Error	10	74.76211	7.47621		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	103.09738	2.12398	17615	2356.10	<.0001
x1	1.43996	0.13842	809.10480	108.22	<.0001
x4	-0.61395	0.04864	1190.92464	159.30	<.0001

Bounds on condition number: 1.0641, 4.2564

Stepwise Selection: Step 3

Variable x2 Entered: R-Square = 0.9823 and C(p) = 3.0182

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2667.79035	889.26345	166.83	<.0001

Error	9	47.97273	5.33030	
Corrected Total	12	2715.76308		

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	71.64831	14.14239	136.81003	25.67	0.0007
x1	1.45194	0.11700	820.90740	154.01	<.0001
x2	0.41611	0.18561	26.78938	5.03	0.0517
x4	-0.23654	0.17329	9.93175	1.86	0.2054

Bounds on condition number: 18.94, 116.36

Stepwise Selection: Step 4

Variable x4 Removed: R-Square = 0.9787 and C(p) = 2.6782

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2657.85859	1328.92930	229.50	<.0001
Error	10	57.90448	5.79045		
Corrected Total	12	2715.76308			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	52.57735	2.28617	3062.60416	528.91	<.0001
x1	1.46831	0.12130	848.43186	146.52	<.0001
x2	0.66225	0.04585	1207.78227	208.58	<.0001

Bounds on condition number: 1.0551, 4.2205

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4		1	0.6745	0.6745	138.731	22.80	0.0006
2	x1		2	0.2979	0.9725	5.4959	108.22	<.0001
3	x2		3	0.0099	0.9823	3.0182	5.03	0.0517
4		x4	2	0.0037	0.9787	2.6782	1.86	0.2054