Chapter 10

Model Selection

For all but the simplest cases we are confronted with a choice of possible regression models for our data. We may even have expanded the choice of possible models by introducing new variables derived from those available originally by making transformations, creating interactions or adding polynomial terms. In this chapter, we consider the problem of selecting the "best" subset of predictors.

More information should only be helpful so one might wonder why we do not simply include all the available variables in the model. However, we may wish to consider a smaller model for several reasons. The principle of Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is best. Applied to regression analysis, this implies that the smallest model that fits the data adequately is best.

Another consideration is that unnecessary predictors will add noise to the estimation of other quantities that interested us. Degrees of freedom will be wasted. More precise estimates and predictions might be achieved with a smaller model. In some cases, collecting data on additional variables can cost time or money so a smaller prediction model may be more economical.

Model selection is a process that should not be separated from the rest of the analysis. Other parts of the data analysis can have an impact. For example, outliers and influential points can do more than just change the current model — they can change the model we select. It is important to identify such points. Also transformations of the variables can have an impact on the model selected. Some iteration and experimentation are often necessary to find better models.

Although Occam's Razor is a compelling heuristic, we must focus our effort on the main objective of regression modelling. We might obtain better predictions by using larger models so although smaller models might be appealing, we do not wish to compromise on predictive ability. For investigations that focus on the explanatory effect of the predictors, one should be cautious about the use of automated variable selection procedures. In such cases, attention is put on just a few predictors of interest while the remaining predictors are not of primary interest but must be controlled for. It would be unwise to expect an automated procedure to do this reliably.

When comparing potential models, we might use hypothesis testing methods to make a choice or use some criterion-based method on the relative fit to decide. We consider both these approaches in this chapter.

Even with a moderate number of potential predictors, the possible combinations of variables can become very large. Procedures that consider all possible combinations may not be practical and we must step through the space of possible models in

an incremental way. In some situations, the model space is structued hierarchically which constrains the reasonable choice of model.

10.1 Hierarchical Models

Some models have a natural hierarchy. For example, in polynomial models, x^2 is a higher order term than x. When selecting variables, it is important to respect the hierarchy. Lower order terms should not usually be removed from the model before higher order terms in the same variable. There are two common situations where this can arise:

Consider the polynomial model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Suppose we fit this model and find that the regression summary shows that the term in x is not significant but the term in x^2 is. If we then remove the x term, our reduced model would become:

$$y = \beta_0 + \beta_2 x^2 + \varepsilon$$

However, suppose we make a scale change $x \to x + a$; then the model would become:

$$y = \beta_0 + \beta_2 a^2 + 2\beta_2 ax + \beta_2 x^2 + \varepsilon$$

The first order x term has now reappeared. Scale changes should not make any important change to the model, but in this case an additional term has been added. This is not desirable. This illustrates why we should not remove lower order terms in the presence of higher order terms. We would not want interpretation to depend on the choice of scale. Removal of the first-order term here corresponds to the hypothesis that the predicted response is symmetric about and has an optimum at $x = 0$. Usually this hypothesis is not meaningful and should not be considered. Only when this hypothesis makes sense in the context of the particular problem could we justify the removal of the lower order term.

For models with interactions, consider the example of a second-order response surface model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

We would not normally consider removing the $x_1 x_2$ interaction term without simultaneously considering the removal of the x_1^2 and x_2^2 terms. A joint removal would correspond to the clearly meaningful comparison of a quadratic surface and a linear one. Just removing the $x_1 x_2$ term would correspond to a surface that is aligned with the coordinate axes. This is harder to interpret and should not be considered unless some particular meaning can be attached. Any rotation of the predictor space would reintroduce the interaction term and, as with the polynomials, we would not ordinarily want our model interpretation to depend on the particular basis for the predictors.

10.2 Testing-Based Procedures

Backward Elimination is the simplest of all variable selection procedures and can be easily implemented without special software. In situations where there is a complex hierarchy, backward elimination can be run manually while taking account of what variables are eligible for removal.

We start with all the predictors in the model and then remove the predictor with highest p-value greater than α_{crit}. Next refit the model and remove the remaining least significant predictor provided its p-value is greater than α_{crit}. Sooner or later, all "nonsignificant" predictors will be removed and the selection process will be complete.

The α_{crit} is sometimes called the "p-to-remove" and does not have to be 5%. If prediction performance is the goal, then a 15 to 20% cutoff may work best, although methods designed more directly for optimal prediction should be preferred.

Forward Selection just reverses the backward method. We start with no variables in the model and then for all predictors not in the model, we check their p-values if they are added to the model. We choose the one with lowest p-value less than α_{crit}. We continue until no new predictors can be added.

Stepwise Regression is a combination of backward elimination and forward selection. This addresses the situation where variables are added or removed early in the process and we want to change our mind about them later. At each stage a variable may be added or removed and there are several variations on exactly how this is done.

We illustrate backward elimination on some data on the 50 states from the 1970s. The data were collected from U.S. Bureau of the Census. We will take life expectancy as the response and the remaining variables as predictors:

```
> data(state)
> statedata <- data.frame(state.x77,row.names=state.abb)
> lmod <- lm(Life.Exp ~ ., statedata)
> sumary(lmod)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.09e+01   1.75e+00   40.59  < 2e-16
Population   5.18e-05   2.92e-05    1.77    0.083
Income      -2.18e-05   2.44e-04   -0.09    0.929
Illiteracy   3.38e-02   3.66e-01    0.09    0.927
Murder      -3.01e-01   4.66e-02   -6.46  8.7e-08
HS.Grad      4.89e-02   2.33e-02    2.10    0.042
Frost       -5.74e-03   3.14e-03   -1.82    0.075
Area        -7.38e-08   1.67e-06   -0.04    0.965

n = 50, p = 8, Residual SE = 0.745, R-Squared = 0.74
```

The signs of some of the coefficients match plausible expectations concerning how the predictors might affect the response. Higher murder rates decrease life expectancy as one might expect. Even so, some variables such as income, are not significant, contrary to what one might expect.

At each stage we remove the predictor with the largest p-value over 0.05. `Area` is the first to go:

```
> lmod <- update(lmod, . ~ . - Area)
```

```
> sumary(lmod)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.09e+01   1.75e+00   40.59  < 2e-16
Population   5.18e-05   2.92e-05    1.77    0.083
Income      -2.18e-05   2.44e-04   -0.09    0.929
Illiteracy   3.38e-02   3.66e-01    0.09    0.927
Murder      -3.01e-01   4.66e-02   -6.46  8.7e-08
HS.Grad      4.89e-02   2.33e-02    2.10    0.042
Frost       -5.74e-03   3.14e-03   -1.82    0.075
Area        -7.38e-08   1.67e-06   -0.04    0.965

n = 50, p = 8, Residual SE = 0.745, R-Squared = 0.74
> lmod <- update(lmod, . ~ . - Illiteracy)
> sumary(lmod)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.0657509  1.0289415   69.07  < 2e-16
Population   0.0000511  0.0000271    1.89    0.066
Income      -0.0000248  0.0002316   -0.11    0.915
Murder      -0.3000077  0.0370418   -8.10  2.9e-10
HS.Grad      0.0477580  0.0185908    2.57    0.014
Frost       -0.0059099  0.0024678   -2.39    0.021

n = 50, p = 6, Residual SE = 0.728, R-Squared = 0.74
> lmod <- update(lmod, . ~ . - Income)
> sumary(lmod)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.0271285  0.9528530   74.54  < 2e-16
Population   0.0000501  0.0000251    2.00    0.052
Murder      -0.3001488  0.0366095   -8.20  1.8e-10
HS.Grad      0.0465822  0.0148271    3.14    0.003
Frost       -0.0059433  0.0024209   -2.46    0.018

n = 50, p = 5, Residual SE = 0.720, R-Squared = 0.74
> lmod <- update(lmod, . ~ . - Population)
> sumary(lmod)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.03638    0.98326   72.25   <2e-16
Murder      -0.28307    0.03673   -7.71    8e-10
HS.Grad      0.04995    0.01520    3.29    0.002
Frost       -0.00691    0.00245   -2.82    0.007

n = 50, p = 4, Residual SE = 0.743, R-Squared = 0.71
```

The final removal of the `Population` variable is a close call. We may want to consider including this variable if interpretation is made easier. Notice that the R^2 for the full model of 0.736 is reduced only slightly to 0.713 in the final model. Thus the removal of four predictors causes only a minor reduction in fit.

It is important to understand that the variables omitted from the model may still be related to the response. For example:

```
> sumary(lm(Life.Exp ~ Illiteracy+Murder+Frost, statedata))
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.55672    0.58425  127.61   <2e-16
Illiteracy  -0.60176    0.29893   -2.01   0.0500
Murder      -0.28005    0.04339   -6.45    6e-08
Frost       -0.00869    0.00296   -2.94   0.0052
```

```
n = 50, p = 4, Residual SE = 0.791, R-Squared = 0.67
```

We see that illiteracy does have some association with life expectancy. It is true that replacing illiteracy with high school graduation rate gives us a somewhat better fitting model, but it would be insufficient to conclude that illiteracy is not a variable of interest. This demonstrates one failing of the method in that it cannot reliably distinguish between important and unimportant predictors.

Testing-based procedures are relatively cheap computationally and easy to understand, but they do have some drawbacks:

1. Because of the "one-at-a-time" nature of adding/dropping variables, it is possible to miss the "optimal" model.
2. The p-values used should not be treated too literally. There is so much multiple testing occurring that the validity is dubious. The removal of less significant predictors tends to increase the significance of the remaining predictors. This effect leads one to overstate the importance of the remaining predictors.
3. The procedures are not directly linked to final objectives of prediction or explanation and so may not really help solve the problem of interest. With any variable selection method, it is important to keep in mind that model selection cannot be divorced from the underlying purpose of the investigation. Variable selection tends to amplify the statistical significance of the variables that stay in the model. Variables that are dropped can still be correlated with the response. It would be wrong to say that these variables are unrelated to the response; it is just that they provide no additional explanatory effect beyond those variables already included in the model.
4. Stepwise variable selection tends to pick models that are smaller than desirable for prediction purposes. To give a simple example, consider the simple regression with just one predictor variable. Suppose that the slope for this predictor is not quite statistically significant. We might not have enough evidence to say that it is related to y but it still might be better to use it for predictive purposes.

Except in simple cases where only a few models are compared or in highly structured heirarchical models, testing-based variable selection should not be used. We include it here because the method is still used but should be discouraged.

10.3 Criterion-Based Procedures

If we have some idea about the purpose for which a model is intended, we might propose some measure of how well a given model meets that purpose. We could choose that model among those possible that optimize that criterion.

It would be natural to pick a model g, parameterized by θ, that is close to the true model f. We could measure the distance between g and f by

$$I(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x|\theta)}\right) dx$$

This is known as the Kullback-Leibler information (or distance). It will be positive except when $g = f$ when it will be zero. Unfortunately, it is impractical for direct implementation because we do not know f. We can substitute in the MLE of θ and rearrange to obtain:

$$\hat{I}(f,g) = \int f(x)\log f(x)dx - \int f(x)\log g(x|\hat{\theta})dx$$

The first term is a constant that doesn't depend on the model we choose g. Akaike (1974) showed that $E\hat{I}(f,g)$ can be estimated by

$$-\log L(\hat{\theta}) + p + \texttt{constant}$$

where p is the number of parameters in the model and the constant depends on the unknown true model. For "historical" reasons, Akaike multiplied this by two to obtain "an information criterion" (AIC):

$$AIC = -2L(\hat{\theta}) + 2p$$

For linear regression models, the -2 max log-likelihood, $-2L(\hat{\theta}) = n\log(RSS/n)+$ another constant. Since the constants are the same for a given data set and assumed error distribution, they can be ignored for regression model comparisons on the same data. Additional care is necessary for other types of comparisons. See Burnham and Anderson (2002) for more detail.

We choose the model which minimizes the AIC. The first term in the AIC is based on RSS which is made smaller by improving the fit. Adding more predictors, provided they are not collinear, will achieve this but we are dissuaded from going too far with this by the second term, called the *penalty term*, $2p$. Hence, we see that AIC naturally provides a balance between fit and simplicity in model selection.

Many other criteria have been proposed. Most well-known among the alternatives is the Bayes information criterion (BIC) which is -2 max log-likelihood + $p\log n$. BIC penalizes larger models more heavily and so will tend to prefer smaller models in comparison to AIC. AIC and BIC are often used as selection criteria for other types of models too. There is some debate regarding the relative merits of these two criteria although AIC is generally considered better when prediction is the aim.

We demonstrate the use of AIC on the `state` dataset. The `leaps` package exhaustively searches all possible combinations of the predictors. For each size of model p, it finds the variables that produce the minumum RSS.

```
> require(leaps)
> b <- regsubsets(Life.Exp~.,data=statedata)
> rs <- summary(b)
> rs$which
  (Intercept) Population Income Illiteracy Murder HS.Grad Frost  Area
1        TRUE      FALSE  FALSE      FALSE   TRUE   FALSE FALSE FALSE
2        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE FALSE FALSE
3        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
4        TRUE       TRUE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
5        TRUE       TRUE   TRUE      FALSE   TRUE    TRUE  TRUE FALSE
6        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE FALSE
7        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE  TRUE
```

Here we see that the best one predictor model uses `Murder` and so on. We compute and plot the AIC:

```
> AIC <- 50*log(rs$rss/50) + (2:8)*2
> plot(AIC ~ I(1:7), ylab="AIC", xlab="Number of Predictors")
```

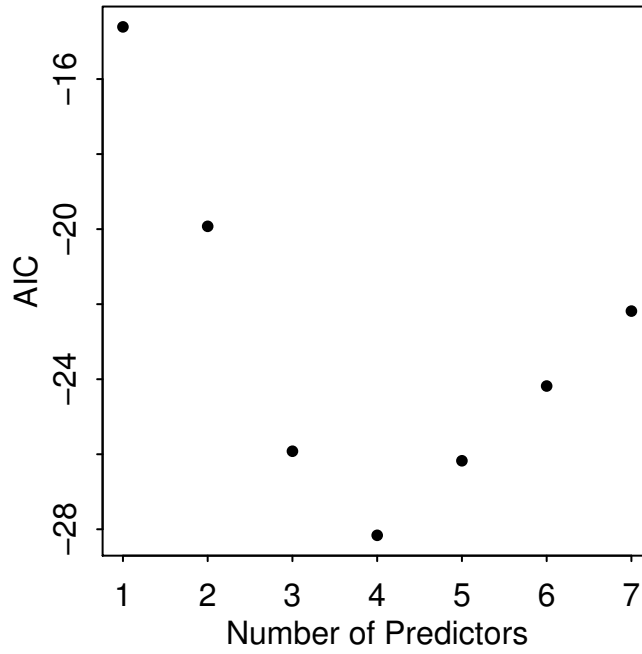

Figure 10.1 AIC for models with varying numbers of predictors using the state data.

We see in Figure 10.1 that the AIC is minimized by a choice of four predictors, namely population, murder, high school graduation and frost as determined by the logical matrix above.

Another commonly used criterion is adjusted R^2, written R^2_a. Recall that $R^2 = 1 - RSS/TSS$. Adding a variable to a model can only decrease the RSS and so only increase the R^2. Hence R^2 by itself is not a good criterion, because it would always choose the largest possible model:

$$R^2_a = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2) = 1 - \frac{\hat{\sigma}^2_{model}}{\hat{\sigma}^2_{null}}$$

Adding a predictor will only increase R^2_a if it has some predictive value. We can see this because minimizing the standard error for prediction means minimizing $\hat{\sigma}^2$ which in term means maximizing R^2_a.

Now let's see which model the adjusted R^2 criterion selects using the plot shown in the first panel of Figure 10.2:

```
> plot(2:8,rs$adjr2,xlab="No. of Parameters",ylab="Adjusted R-square")
> which.max(rs$adjr2)
[1] 4
```

We see that the population, frost, high school graduation and murder model has the largest R_a^2.

Our final criterion is Mallow's C_p statistic. A good model should predict well, so the average mean square error of prediction might be a good criterion:

$$\frac{1}{\sigma^2}\sum_i E(\hat{y}_i - Ey_i)^2$$

which can be estimated by the C_p statistic:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n$$

where $\hat{\sigma}^2$ is from the model with all predictors and RSS_p indicates the RSS from a model with p parameters. For the full model $C_p = p$ exactly. If a p predictor model fits, then $E(RSS_p) = (n-p)\sigma^2$ and then $E(C_p) \approx p$. A model with a bad fit will have C_p much bigger than p. It is usual to plot C_p against p. We desire models with small p and C_p around or less than p. C_p, R_a^2 and AIC all trade-off fit in terms of RSS against complexity (p).

The C_p plot can be constructed as:

```
> plot(2:8,rs$cp,xlab="No. of Parameters",ylab="Cp Statistic")
> abline(0,1)
```

as seen in the second panel of Figure 10.2. The competition is between the four-parameter, three-predictor, model including frost, high school graduation and murder and the model also including population. Both models are on or below the $C_p = p$ line, indicating good fits. The choice is between the smaller model and the larger model, which fits a little better. Some even larger models fit in the sense that they are on or below the $C_p = p$ line, but we would not opt for these in the presence of smaller models that fit.

If there are q potential predictors, then there are 2^q possible models. For larger q, this might be too time consuming and we may need to economize by limiting the search. In such cases, the `step()` function is a cheaper alternative. The function does not evaluate the AIC for all possible models but uses a search method that compares models sequentially. Thus it bears some comparison to the stepwise method described above, but only in the method of search — there is no hypothesis testing.

```
> lmod <- lm(Life.Exp ~ ., data=statedata)
> step(lmod)
Start:  AIC= -22.18
 Life.Exp ~ Population + Income + Illiteracy + Murder +
    HS.Grad +  Frost + Area

             Df Sum of Sq   RSS   AIC
- Area        1    0.0011  23.3 -24.2
- Income      1    0.0044  23.3 -24.2
```

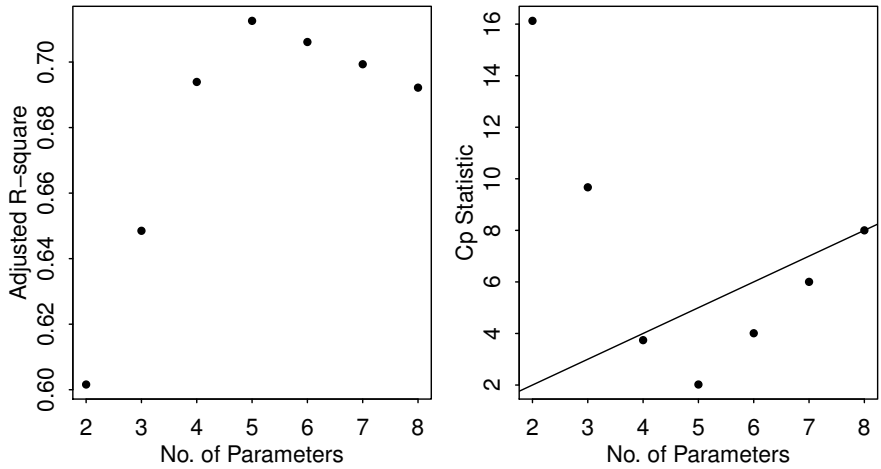

Figure 10.2 The adjusted R^2 for the state data models is on the left and the C_p plot for the same data at the right.

```
- Illiteracy  1    0.0047  23.3 -24.2
<none>                     23.3 -22.2
- Population  1       1.7  25.0 -20.6
- Frost       1       1.8  25.1 -20.4
- HS.Grad     1       2.4  25.7 -19.2
- Murder      1      23.1  46.4  10.3

Step:  AIC= -24.18
 Life.Exp ~ Population + Income + Illiteracy + Murder +
    HS.Grad + Frost

.. intermediate steps omitted ..

Step:  AIC= -28.16
 Life.Exp ~ Population + Murder + HS.Grad + Frost

             Df Sum of Sq   RSS   AIC
<none>                     23.3 -28.2
- Population  1       2.1  25.4 -25.9
- Frost       1       3.1  26.4 -23.9
- HS.Grad     1       5.1  28.4 -20.2
- Murder      1      34.8  58.1  15.5

Coefficients:
(Intercept)   Population       Murder      HS.Grad        Frost
 71.0271285    0.0000501   -0.3001488    0.0465822   -0.0059433
```

The sequence of variable removal is the same as with backward elimination and the model selected is the same as for AIC.

Variable selection methods are sensitive to outliers and influential points. Let's check for high leverage points:

```
> h <- lm.influence(lmod)$hat
> names(h) <- state.abb
> rev(sort(h))
      AK       CA       HI       NV       NM       TX       NY
0.809522 0.408857 0.378762 0.365246 0.324722 0.284164 0.256950
```

We can see that Alaska has high leverage. Let's try excluding it:

```
> b<-regsubsets(Life.Exp~.,data=statedata,  subset=(state.abb!="AK"))
> rs <- summary(b)
> rs$which[which.max(rs$adjr),]
(Intercept)  Population      Income  Illiteracy      Murder
       TRUE        TRUE       FALSE       FALSE        TRUE
    HS.Grad       Frost        Area
       TRUE        TRUE        TRUE
```

We see that `Area` now makes it into the model. Transforming the predictors can also have an effect. Take a look at the variables:

```
> stripchart(data.frame(scale(statedata)), method ="jitter", las=2,
    vertical=TRUE)
```

Jittering adds a small amount of noise (in the horizontal direction in this example). It is useful for moving apart points that would otherwise overprint each other.

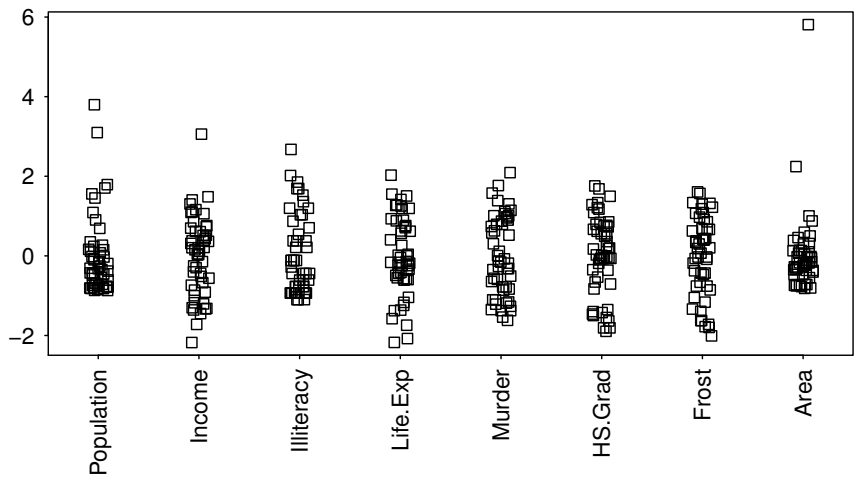

Figure 10.3 Strip charts of the state data; all variables have been standardized.

In Figure 10.3, we see that population and area are skewed — we try transforming them:

```
> b<-regsubsets(Life.Exp ~ log(Population)+Income+Illiteracy+ Murder+
    HS.Grad+Frost+log(Area),statedata)
> rs <- summary(b)
> rs$which[which.max(rs$adjr),]
    (Intercept) log(Population)          Income      Illiteracy
           TRUE            TRUE           FALSE           FALSE
```

```
      Murder          HS.Grad            Frost        log(Area)
        TRUE             TRUE             TRUE            FALSE
```

This changes the "best" model again to log(population), frost, high school graduation and murder. The adjusted R^2 of 71.7% is the highest among models we have seen so far.

10.4 Summary

Variable selection is a means to an end and not an end itself. The aim is to construct a model that predicts well or explains the relationships in the data. Automatic variable selections are not guaranteed to be consistent with these goals. Use these methods as a guide only.

Hypothesis testing-based methods use a restricted search through the space of potential models and use a dubious method for choosing between models when repeated many times. Criterion-based methods typically involve a wider search and compare models in a preferable manner. For this reason, we recommend that you use a criterion-based method.

Accept the possibility that several models may be suggested which fit about as well as each other. If this happens, consider:

1. Do the models have similar qualitative consequences?
2. Do they make similar predictions?
3. What is the cost of measuring the predictors?
4. Which has the best diagnostics?

If you find models that seem roughly comparable, but lead to quite different conclusions, then it is clear that the data cannot answer the question of interest unambiguously. Be alert to the possibility that a model contradictory to the tentative conclusions might be out there.

Exercises

1. Use the `prostate` data with `lpsa` as the response and the other variables as predictors. Implement the following variable selection methods to determine the "best" model:
   (a) Backward elimination
   (b) AIC
   (c) Adjusted R^2
   (d) Mallows C_p
2. Using the `teengamb` dataset with `gamble` as the response and the other variables as predictors, repeat the work of the first question.
3. Using the `divusa` dataset with `divorce` as the response and the other variables as predictors, repeat the work of the first question.