Chapter 10

Variable Selection

Variable selection is intended to select the "best" subset of predictors. But why bother?

1. We want to explain the data in the simplest way — redundant predictors should be removed. 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 is best.

2. Unnecessary predictors will add noise to the estimation of other quantities that we are interested in. Degrees of freedom will be wasted.

3. Collinearity is caused by having too many variables trying to do the same job.

4. Cost: if the model is to be used for prediction, we can save time and/or money by not measuring redundant predictors.

Prior to variable selection:

1. Identify outliers and influential points - maybe exclude them at least temporarily.

2. Add in any transformations of the variables that seem appropriate.

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 be removed from the model before higher order terms in the same variable. There two common situations where this situation arises:

- Polynomials models. Consider the 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 removed the x term, our reduced model would then become

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

but suppose we then made a scale change $x \rightarrow x + a$, 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 good. 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 $x = 0$. Often 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.

- Models with interactions. Consider the second order response surface model:

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

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 linear one. Just removing the $x_1 x_2$ term would correspond to a surface that is aligned with the coordinate axes. This is hard 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 Stepwise Procedures

Backward Elimination

This 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.

1. Start with all the predictors in the model
2. Remove the predictor with highest p-value greater than α_{crit}
3. Refit the model and goto 2
4. Stop when all p-values are less than α_{crit}.

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 15-20% cut-off may work best, although methods designed more directly for optimal prediction should be preferred.

10.2.1 Forward Selection

This just reverses the backward method.

1. Start with no variables in the model.
2. For all predictors not in the model, check their p-value if they are added to the model. Choose the one with lowest p-value less than α_{crit}.
3. Continue until no new predictors can be added.

10.2.2 Stepwise Regression

This 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.

Stepwise procedures are relatively cheap computationally but they do have some drawbacks.

1. Because of the "one-at-a-time" nature of adding/dropping variables, it's 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 these variables are unrelated to the response, it's 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.

We illustrate the variable selection methods on some data on the 50 states - the variables are population estimate as of July 1, 1975; per capita income (1974); illiteracy (1970, percent of population); life expectancy in years (1969-71); murder and non-negligent manslaughter rate per 100,000 population (1976); percent high-school graduates (1970); mean number of days with min temperature $<$ 32 degrees (1931-1960) in capital or large city; and land area in square miles. The data was collected from US Bureau of the Census. We will take life expectancy as the response and the remaining variables as predictors - a fix is necessary to remove spaces in some of the variable names.

```
> data(state)
> statedata <- data.frame(state.x77,row.names=state.abb,check.names=T)
> g <- lm(Life.Exp ~ ., data=statedata)
> summary(g)
Coefficients:
             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
```

```
Residual standard error: 0.745 on 42 degrees of freedom
Multiple R-Squared: 0.736,      Adjusted R-squared: 0.692
F-statistic: 16.7 on 7 and 42 degrees of freedom,       p-value: 2.53e-10
```

Which predictors should be included - can you tell from the p-values? Looking at the coefficients, can you see what operation would be helpful? Does the murder rate decrease life expectancy - that's obvious a priori, but how should these results be interpreted?

We illustrate the backward method - at each stage we remove the predictor with the largest p-value over 0.05:

```
> g <- update(g, . ~ . - Area)
> summary(g)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.10e+01   1.39e+00   51.17  < 2e-16
Population   5.19e-05   2.88e-05    1.80    0.079
Income      -2.44e-05   2.34e-04   -0.10    0.917
Illiteracy   2.85e-02   3.42e-01    0.08    0.934
Murder      -3.02e-01   4.33e-02   -6.96  1.5e-08
HS.Grad      4.85e-02   2.07e-02    2.35    0.024
Frost       -5.78e-03   2.97e-03   -1.94    0.058

> g <- update(g, . ~ . - Illiteracy)
> summary(g)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.11e+01   1.03e+00   69.07  < 2e-16
Population   5.11e-05   2.71e-05    1.89    0.066
Income      -2.48e-05   2.32e-04   -0.11    0.915
Murder      -3.00e-01   3.70e-02   -8.10  2.9e-10
HS.Grad      4.78e-02   1.86e-02    2.57    0.014
Frost       -5.91e-03   2.47e-03   -2.39    0.021

> g <- update(g, . ~ . - Income)
> summary(g)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.10e+01   9.53e-01   74.54  < 2e-16
Population   5.01e-05   2.51e-05    2.00   0.0520
Murder      -3.00e-01   3.66e-02   -8.20  1.8e-10
HS.Grad      4.66e-02   1.48e-02    3.14   0.0030
Frost       -5.94e-03   2.42e-03   -2.46   0.0180

> g <- update(g, . ~ . - Population)
> summary(g)
```

```
Coefficients:
            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   0.0020
Frost       -0.00691    0.00245   -2.82   0.0070

Residual standard error: 0.743 on 46 degrees of freedom
Multiple R-Squared: 0.713,      Adjusted R-squared: 0.694
F-statistic:   38 on 3 and 46 degrees of freedom,       p-value: 1.63e-12
```

The final removal of the Population variable is a close call. We may want to consider including this variable if interpretation is aided. 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.

10.3 Criterion-based procedures

If there are p potential predictors, then there are 2^p possible models. We fit all these models and choose the best one according to some criterion. Clever algorithms such as the "branch-and-bound" method can avoid actually fitting all the models — only likely candidates are evaluated. Some criteria are

1. The Akaike Information Criterion (AIC) and the Bayes Information Criterion (BIC) are some other commonly used criteria. In general,

   $$AIC = -2\mathtt{log} - \mathtt{likelihood} + 2p$$

   while

   $$BIC = -2\mathtt{log} - \mathtt{likelihood} + p\log n$$

   For linear regression models, the -2log-likelihood (known as the *deviance* is $n\log(RSS/n)$. We want to minimize AIC or BIC. Larger models will fit better and so have smaller RSS but use more parameters. Thus the best choice of model will balance fit with model size. BIC penalizes larger models more heavily and so will tend to prefer smaller models in comparison to AIC. AIC and BIC can be used as selection criteria for other types of model too.

   We can apply the AIC (and optionally the BIC) to the state data. 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 with the advantage that no dubious p-values are used.

   ```
   > g <- lm(Life.Exp ~ ., data=statedata)
   > step(g)
   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
- 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
   7.10e+01     5.01e-05    -3.00e-01     4.66e-02    -5.94e-03
```

The sequence of variable removal is the same as with backward elimination. The only difference is the the Population variable is retained.

2. Adjusted R^2 — called 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 so 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 value. Do you see the connection to $\hat{\sigma}^2$? Minimizing the standard error for prediction means minimizing $\hat{\sigma}^2$ which in term means maximizing R^2_a.

3. Predicted Residual Sum of Squares (PRESS) is defined as $\sum_i \hat{\varepsilon}^2_{(i)}$ where the $\hat{\varepsilon}_{(i)}$ are the residuals calculated without using case i in the fit. The model with the lowest PRESS criterion is then selected. This tends to pick larger models (which may be desirable if prediction is the objective).

4. Mallow's C_p Statistic. A good model should predict well so average MSE 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

$$C_p = \frac{\text{RSS}_p}{\hat{\sigma}^2} + 2p - n$$

where $\hat{\sigma}^2$ is from the model with all predictors and RSS_p indicates the RSS from a model with p parameters.

(a) C_p is easy to compute

(b) It is closely related to R_a^2 and the AIC.

(c) For the full model $C_p = p$ exactly.

(d) If a p predictor model fits then $E(\text{RSS}_p) = (n-p)\sigma^2$ 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.

Now we try the C_p and R_a^2 methods for the selection of variables in the State dataset. The default for the `leaps()` function is the Mallow's Cp criterion:

```
> library(leaps)
> x <- model.matrix(g)[,-1]
> y <- statedata$Life
> g <- leaps(x,y)
> Cpplot(g)
```

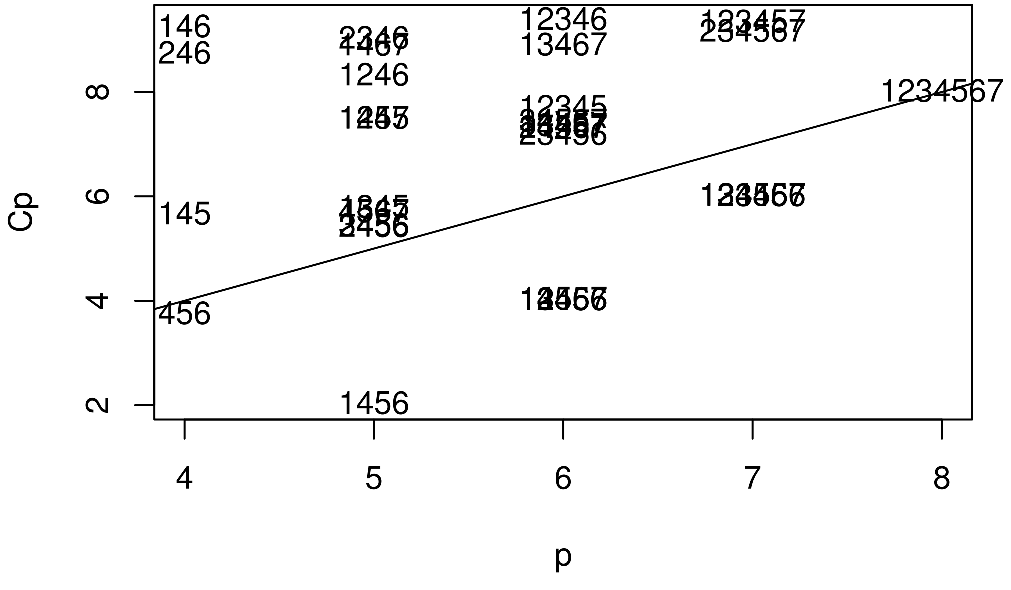

Figure 10.1: The Cp plot for the State data

The models are denoted by indices for the predictors. The competition is between the "456" model i.e. the Frost, HS graduation and Murder model 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. Smaller models with 1 or 2 predictors are not shown on this plot because their C_p plots are so large.

Now let's see which model the adjusted R^2 criterion selects.

```
> adjr <- leaps(x,y,method="adjr2")
> maxadjr(adjr,8)
  1456  12456  13456  14567 123456 134567 124567    456
 0.713  0.706  0.706  0.706  0.699  0.699  0.699  0.694
```

We see that the Population, Frost, HS graduation and Murder model has the largest R_a^2. The best three predictor model is in eighth place but the intervening models are not attractive since they use more predictors than the best model.

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

```
> h <- hat(x)
> names(h) <- state.abb
> rev(sort(h))
      AK       CA       HI       NV       NM       TX       NY       WA
0.809522 0.408857 0.378762 0.365246 0.324722 0.284164 0.256950 0.222682
```

Which state sticks out? Let's try excluding it (Alaska is the second state in the data).

```
> l <- leaps(x[-2,],y[-2],method="adjr2")
> maxadjr(l)
 12456   1456 123456
 0.710  0.709  0.707
```

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

```
> par(mfrow=c(3,3))
> for(i in 1:8) boxplot(state.x77[,i],main=dimnames(state.x77)[[2]][i])
```

In Figure 10.3, we see that Population, Illiteracy and Area are skewed - we try transforming them:

```
> nx <- cbind(log(x[,1]),x[,2],log(x[,3]),x[,4:6],log(x[,7]))
```

And now replot:

```
> par(mfrow=c(3,3))
> apply(nx,2,boxplot)
```

which shows the appropriately transformed data.

Now try the adjusted R2 method again.

```
> a <- leaps(nx,y,method="adjr2")
> maxadjr(a)
 1456 12456 13456
0.717 0.714 0.712
```

This changes the ”best” model again to log(Population), Frost, HS graduation and Murder.

The adjusted R^2 is the highest models we have seen so far.

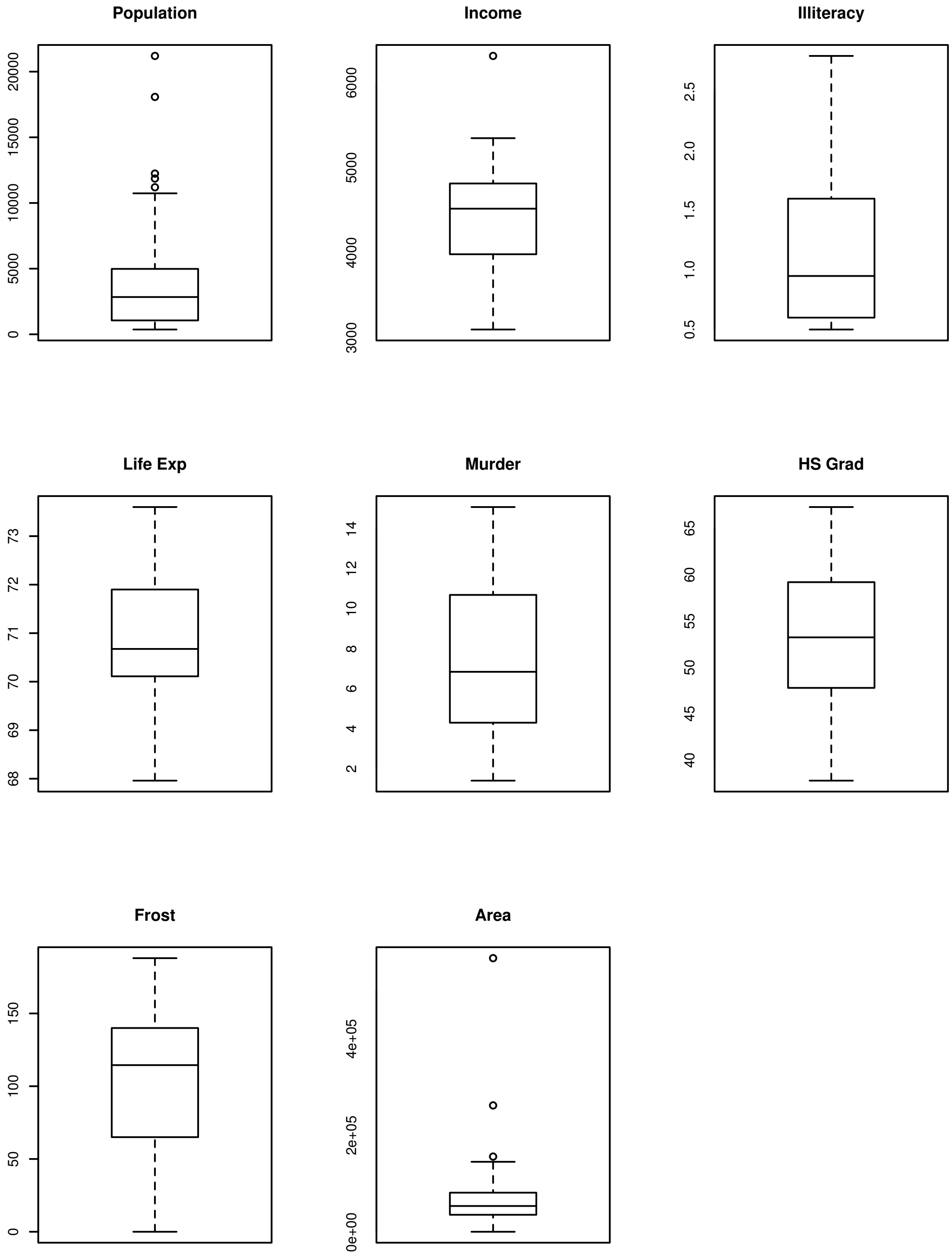

Figure 10.2: Boxplots of the State data

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.

Stepwise methods use a restricted search through the space of potential models and use a dubious hypothesis testing based method for choosing between models. Criterion-based methods typically involve a wider search and compare models in a preferable manner. For this reason, I 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 equally as good but lead to quite different conclusions then it is clear that the data cannot answer the question of interest unambiguously. Be alert to the danger that a model contradictory to the tentative conclusions might be out there.