Choosing the Optimal Model

Algorithm 6.1 (p. 205), and similarly Algorithms 6.2 (p. 207) and Algorithm 6.3 (p. 209), describe the process of model selection in three steps. Step 2 involves, for each \( k = 0, 1, 2, \ldots, p \), selecting a model \( M_k \) which, among some collection of models using \( k \) of the available \( p \) predictors, is best. It is appropriate to compare models with a fixed number \( k \) of parameters via the size of RSS, calling one model fit to training data better than another if it has smaller RSS. Because RSS cannot rise with the inclusion of more predictor variables, it is not an appropriate measure for comparing a model \( M_k \) with \( M_j \) when \( j \neq k \).

Using a statistic adjusted for model size

To compare models \( M_k \) and \( M_j \), with \( j \neq k \), we have several other statistics which have been developed to adjust to number of predictor variables used. Whereas \( R^2 \) is always bigger for \( M_k \) than for \( M_j \) if \( k > j \), the \( C_p \), AIC, and BIC are model statistics which can be computed using training data for both models, and will allow meaningful conclusions about the training error for each. More specifically, the model, between \( M_k \) and \( M_j \), which has the smaller value of one of these will generally have the smaller test error as well. One other statistic, the adjusted \( R^2 \), serves the same purpose, but its maximum, rather than minimum, is sought.

The `summary()` command, when carried out on the results from `regsubsets()`, gives us access to these adjusted statistics.

```R
require(leaps) # package contains regsubsets() function

% Loading required package: leaps

Credit = read.csv("http://www-bcf.usc.edu/~gareth/ISL/Credit.csv",row.names=1)
regfit.fwd = regsubsets(Balance ~ ., data=Credit, nvmax=11, method="forward")
fwd.model.summary = summary(regfit.fwd)
fwd.model.summary$cp

The last line above reveals the various \( C_p \) values for the best models obtained using `regsubsets()`. We might just as well have asked for values of other adjusted statistics:

```R
names(fwd.model.summary)
%    [1] "which" "rsq" "rss" "adjr2" "cp" "bic" "outmat" "obj"
fwd.model.summary$adjr2
```

```R
```
We identify the minimum

\[
\text{which.min}(\text{fwd.model.summary$cp})
\]

[1] 6

\[
\text{coef(regfit.fwd, 6)}
\]

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>Income</th>
<th>Limit</th>
<th>Rating</th>
<th>Cards</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>-493.7341870</td>
<td>-7.7950824</td>
<td>0.1936914</td>
<td>1.0911874</td>
<td>18.2118976</td>
<td>-0.6240560</td>
</tr>
</tbody>
</table>

which suggests the best model considered here (i.e., out of all linear models involving some collection of the \( p = 11 \) predictors) is

\[
\text{Balance} = -493.7 - 7.8(\text{Income}) + 0.193(\text{Limit}) + 1.09(\text{Rating}) + 18.2(\text{Cards}) - 0.62(\text{Age}) + 425.6(\text{StudentYes}).
\]

For visual confirmation, we plot the \( C_p \) values and see that the minimum dot occurs (plausibly) at 6 predictors.

\[
\text{plot(} \text{fwd.model.summary$cp, pch=20, cex=.6, col="blue"})
\]

\[
\text{points(6, fwd.model.summary$cp[6], col="red", pch=20, cex=1)}
\]
Similar commands that, instead of $C_p$, make use of BIC:

```r
regfit.full = regsubsets(Balance ~ ., data=Credit, nvmax=13)
full.model.summary = summary(regfit.full)
plot(full.model.summary$bic, pch=20, cex=.6, col="blue")
lines(full.model.summary$bic, col="red")
which.min(full.model.summary$bic)
points(4, full.model.summary$bic[4], col="red", pch=20, cex=1)
coef(regfit.full, 4)
```

If you carry out the previous code snippet, you realize we do not arrive at the same model size for both: by $C_p$ standards, we want a model with 6 predictors, but BIC suggests 4 is best. What is reassuring is that there is little change in either $C_p$ or BIC through the range $k = 3, 4, 5, \ldots, 11$ predictors.

**Using validation and cross-validation**

We did not delve into this topic, found in Section 6.1.3 (specifically pp. 213-214).