Null Behavior of Best Subset Selection

March 3, 2017

In this exercise, you will examine how best subset selection works when there is no relation at all between predictors and the response.

Preparation

Use the code below to make a random matrix of 20 predictors and 200 observations together with a random response, drawn from $N(0, 1)$. Then fit a linear model to predict y from the other 20 columns and examine it with **summary()**. Verify that none of the predictors are significant.

```
set.seed(107)
X = as.data.frame(matrix(rnorm(4200), ncol = 21))
names(X)[1] <- "y"
```

Your Tasks

(a) Use **regsubsets()** to find the best subsets with 1, 2, ..., 20 predictors. Make plots of the adjusted R^2, Mallows Cp, and the Bayes Information Criterion and determine the overall best model size for each of these criteria.

(b) Use tenfold cross validation to examine these 20 models. Carry this out at least three times to assess the variability of the error estimate. What is the conclusion?

(c) Compare all four approaches to select the overall best model. Which recommendations make sense?