Exercise Sheet 9 — Model Selection

Problem 1. F-test distribution after model selection
The aim of this exercise is to show through a small simulation study that the distribution of the p-value of the global F-test — which tests whether the model is significantly different than a model with intercept only — after variable selection is not the one advertised.

(a) We simulate data as in slides 5 and 6 of the lecture, i.e., Y is normally distributed with mean 0 and variance 1, as are the 100 explanatory variables $X_1, \ldots, X_{100}$. Repeat the simulation presented in the lecture 1000 times and store the p-values of the F-tests in a vector. The F-statistic, numerator and denominator degrees of freedom can be found in

```
summary(mdl.2)$fstatistic
```

For some simulation runs, no variable will be selected. What do you propose to do?

(b) Display an histogram of the p-value. Comment (Hint: Under the null hypothesis, the p-values should be uniformly distributed)

Problem 2. bias-variance tradeoff and optimism of the training error
Consider the data `College` available in package **ISLR**. The aim is to predict the log of the number of applications received (`Apps`) using the other variables as explanatory variables.

(a) Split the data set into a training set and a test set of roughly equal size.

(b) Fit a linear model with forward stepwise selection using the `regsubsets` function, setting `nvmax` equal to 17.

(c) Compute the mean squared error in the training set and the test set as a function of the number of variables included in the model (a `predict` function for `regsubsets` objects is available in the R file accompanying the lecture.)

(d) Plot the results and comment