Homework 5

Due October 15 at 3:00.

Searching for solutions is forbidden.

(1) Get the Apple data.

```
library(alr4)
attach(allshoots)
names(allshoots)
help(allshoots)
```

The goal is to predict the number of 'stem units' on an apple tree based on days since dormancy. The variables are:

Day	days from dormancy
n	number of shoots sampled
ybar	average number of stem units
SD	within-day standard deviation
Type	1 if long shoots, 0 if shortshoots.

The data were collected on 106 days. But for each data, we do not have the raw data. Instead, we have the average value (ybar), the number of data points (n) and the standard deviation. Also, there are two types of shoots (this is the binary variable Type). The goal is to predict ybar from Days.

(a) Plot the data. Does the relationship between ybar and Days look different for the two different types?

(b) Perform a linear regression on ybar on Days and Type and their interaction. Summarize the fitted model. Check the residuals and report any problems you see (if any). Give 80 percent confidence intervals for the coefficients.

(c) Do a weighted regression using n_i (number of shoots) as weights. Give 95 percent confidence intervals for the coefficients and compare them to the confidence intervals from the unweighted regression.

(2) Download the following economic data:

```
library(alr4)
attach(BigMac2003)
names(BigMac2003)
help(BigMac2003)
```

The goal is to predict the variable FoodIndex (a measure of how expensive food is) from the other variables.

(a) Fit a linear regression model. Do all the usual diagnostics and comment on the diagnostics.

(b) There is a hypothesis that the price of a BigMac is a good predictor of FoodIndex. (The magazine *The Economist* has published a Big Mac index for many years.) Construct a 99 percent confidence interval for the BigMac variable in your regression. Interpret the confidence interval.

(c) To explore this further, fit a regression model that uses BigMac as the only covariate. Use an F-test to compare this to your previous model. What hypothesis are you testing? What is your conclusion?

(d) Now we want to address a different question: how good is each of your two models in terms of prediction error? Estimate the prediction error of both models. What is your conclusion?

(3) Load these data:

```
library(mlbench)
data(BreastCancer)
df = BreastCancer[complete.cases(BreastCancer), ]
```

The last command gets rid of cases with missing data. (Not necessarily a good idea but we'll do this to simplify things.) The goal is to predict cell "class" (benign or malignant) from the other variables. Use logistic regression to predict class from the other variables. What is the fitted model? How well does it predict (i.e. what is the proportion of misclassifications on the training data)?

(4) Download the Ozone data:

```
library(mlbench)
data(Ozone)
help(Ozone)
```

The goal is to predict ozone (variable 4) from the other variables. Some of the rows of the data frame have missing values. Find these rows and remove them. Throw away the variables 'Day of Month' and 'Day of Week.' Convert the variable 'Month' into a numeric variable (rather than a factor). The variables in the data frame don't have names. I suggest you give them meaningful names, such as: month, ozone, pressure, wind, etc. It will be easier to use `glmnet` if you create a vector `Y` for the outcome and a matrix `X` for the covariates. For example, if you have a data frame `D` and you want to create a matrix `X` that consists of the columns 7,8,9 and 10 you can do the following:

```
I = c(7,8,9,10)
X = D[,I]
X = data.matrix(x)
```

To make the problem even more interesting, we are going to add 20 extra columns to `X` that are just extra, unrelated variables. The commands are:

```
n = nrow(X)
fake = rnorm(20*n)
fake = matrix(fake,n,20)
X = cbind(X,fake)
```

(a) Fit the model using the lasso. Plot the cross-validation score versus the number of variables in the model. What variables are in the final selected model? Did the real variables enter the model before the fake variables?

(b) Construct a 90 percent confidence interval for the coefficient of variable 8 (Temperature at Sandburg, CA).