Homework 10

(1) Download the Ozone data:

```
library(mlbench)
data(Ozone)
attach(Ozone)
names(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. (Throwing away data points with missing values is not necessarily good practice but we will do it here to simplify things.) Throw away the variables 'Day of Month' and 'Day of Week.' Also, 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.

We are going to do some variable selection. It will be easier to use `lars` or `glmnet` if you create a vector `y` for the outcome and a matrix `x` for the covariates. (You don't have to use the names `y` and `x`. Use whatever names you like). 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 10 extra columns to `x` that are just extra, unrelated variables. The commands are:

```
n = nrow(x)
fake = rnorm(10*n)
fake = matrix(fake,n,10)
x = cbind(x,fake)
```

(a) Use forward stepwise selection to select the variables. Summarize the analysis. 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) Repeat part (a) but use the lasso.

(2) Generate some data as follows:

```
n = 100
x = runif(n,-1,1)
m = sin(5*x)
y = m + rnorm(n,0,.3)
```

(a) Fit a kernel regression estimators for bandwidths $h = .01$, $h = .1$, $h = .5$. Plot the data with the fitted functions.

(b) Estimate the prediction error for a grid of bandwidths h. For example, you might use:

```
h = seq(.01,.5,length=20)
```

Plot the cross-validation error versus h.

(c) Find the bandwidth h that minimizes the cross-validation error. Plot the corresponding estimator and the residuals.

(3) We are going to use the Ozone data again but this time we will do some nonparametric regression. Do the same pre-processing as before but don't add the fake variables. Again, the goal is to predict ozone from the other variables.

(a) Make a pairs plot. You will probably find it hard to make sense of the pairs plot. As an alternative, plot each covariate versus ozone. Then add a nonparametric smooth to each of these plots. The R command `scatter.smooth` will plot the data and add a nonparametric fit. For example:

```
scatter.smooth(x,y)
```

Comment on the plots.

(b) Fit a linear model using all the covariates to predict ozone. Summarize the fitted model and the residual plots.

(c) Estimate the predictive error of your model using leave-one-out-cross-validation.

(d) Now fit a nonparametric additive model. I suggest you use the library `mgcv` and the command `gam`. Summarize the fitted model. Plot the fitted functions and comment on the plots.

(e) Estimate the predictive error of your model. You can get the diagonal elements of your fitted model using:

```
out$hat
```

where `out` is the name of the output of your model. How does the predictive error of your model compare to the predictive error of your linear model?

(f) Let's do some (somewhat subjective) variable selection. If you look at the plots of the fitted functions, some of the fitted functions are nearly constant. Remove those variables. Fit an additive model based on the remaining variables. Summarize the fit and estimate the predictive error of your model.

(4) Suppose we have data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $X_i \in [0, 1]$. Let $m(x) = \mathbb{E}[Y|X = x]$. Furthermore, suppose that $X_i = i/n$. We are treating the X_i's as fixed (non-random). Let $h > 0$ and consider the the following kernel estimator:

$$\widehat{m}(x) = \frac{1}{k} \sum_{i \in B} Y_i$$

where $B = \{i : \ |X_i - x| \leq h\}$ and k is the number of points in B.

Let us fix some point $x \in (0, 1)$. You can make the following assumptions:

1. $x = j/n$ for some integer j.
2. $0 < x - h < x + h < 1$.
3. $h = r/n$ for some integer r.
4. $Y_i = m(X_i) + \epsilon_i$ where $\mathbb{E}[\epsilon_i] = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$.

(a) Find $\text{Var}(\widehat{m}(x))$.

(b) Find an expression for the bias of $\widehat{m}(x)$. You may use the following Taylor series approximation: for any y near x,

$$m(y) \approx m(x) + (y - x)m'(x) + \frac{(y - x)^2}{2} m''(x).$$

Hint: You should find that the bias is of the form Ch^2 for some $C > 0$. You should have an explicit expression for C.

(c) Using your expression for the bias and variance, find an explicit formula for the bandwidth h that minimizes the integrated mean squared error.