Homework 6

Due October 22 at 3:00.

Searching for solutions is forbidden.

(1) For some regression methods, we can use the shortcut formula for estimating prediction risk. But that is not always the case. Here is a more general way to estimate the prediction error:

(i) Split the data into two groups, the training set $\mathcal{D}_0$ and the test set $\mathcal{D}_1$.

(ii) Use $\mathcal{D}_0$ to fit several models. Let's call them $\widehat{m}_1, \ldots, \widehat{m}_k$.

(iii) Estimating the prediction error using the test set:

$$\widehat{R}_j = \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} (Y_i - \widehat{m}_j(X_i))^2 = \frac{1}{n_1} \sum_{i \in \mathcal{D}_1} B_{ij}$$

where n_1 is the number of observations in $\mathcal{D}_1$ and $B_{ij} = (Y_i - \widehat{m}_j(X_i))^2$. This is an estimate of the prediction risk R_j for model $\widehat{m}_j$.

(iv) A confidence interval for R_j is $\widehat{R}_j \pm z_{\alpha/2} s_j / \sqrt{n}_1$ where s_j^2 is the standard deviation of $B_{1j}, \ldots, B_{n_1 j}$. In other words,

$$s_j^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (B_{ij} - \widehat{R}_j)^2.$$

(v) A confidence interval for $R_j - R_k$ is $\widehat{R}_j - \widehat{R}_k \pm z_{\alpha/2} s_{jk} / \sqrt{n}_1$ where s_{jk}^2 is the standard deviation of $B_{1j} - B_{1k}, \ldots, B_{n_1 j} - B_{n_1 k}$.

Let's try this.

(a) Generate $n = 1000$ observations as follows:

```
n = 1000
X = runif(n,-1,1)
epsilon = rnorm(n,0,.1)
Y = 2*X^2 + epsilon
```

Divide the data into a training set of size 800 and a test set of size 200. Plot the data from the training set.

(b) Using the traning data, fit ten models $\widehat{m}_1, \ldots, \widehat{m}_{10}$ where $\widehat{m}_j$ is a j^{th} degree polynomial. Plot the ten fitted models. Your plot should have the training data, and ten curves, one for each model.

(c) Estimate the prediction error using the test data. Plot the estimated prediction error along with confidence intervals. One way (but not the only way) to do that is as follows. Suppose that `Rhat` is your estimate of prediction error, `up` is your upper confidence limit for R and `low` is your lower confidence limit for R. In other words, `[low[j],up[j]]` is your confidence interval for model j. Then you might do the following:

```
a = min(low)
b = max(up)
```

```
plot(1:10,Rhat,ylim=c(a,b))
for(i in 1:10){
     segments(i,low[i],i,up[i],lwd=3)
     }
points(1:10,Rhat)
```

You can use this code to make the plot but feel free to create your own way of creating the plot.

(d) Based on your results, what model do you think is best. Why?

(2) Read the data from the file `nuclear.txt`. There are 254 observations on the following 2 variables: Temperature and Toughness. The dataset, from the US Electric Power Research Institute (EPRI), consists of 254 toughness results obtained from non-irradiated representative steels. For each steel, the fracture toughness and temperature were measured. The goal is to understand the relationship between fracture toughness for the reactor pressure vessel materials as a function of the temperature. So X = temperature and Y = toughness.

(a) Fit a linear regression. Plot the data with the fitted line and plot the residuals. Comment on your findings.

(b) Now estimate the regresion function using three methods: kernel regression, local linear regression and smoothing splines. In each case, use cross-validation to choose the tuning parameters. Plot the data and the three curves on one plot. Comment on any differences between the methods.

For the rest of the question, use your local linear fit.

(c) Plot the data, the fitted function and a 95 percent confidence band.

(d) Estimate $\sigma^2(x)$ and plot the estimate.

(e) It is believed that the true regression function should be an increasing function. Does your analysis support this?