B565: Homework 5

1. Consider the Vocab.csv data on Canvas (in the data folder), containing the number of years of education and the performance on a vocabulary test for a number of individuals. For this problem we want to perform simple linear regression by modeling the relationship between the response (the test score) and the predictor (the education level). As with the simple linear regression done in class, case we want to model

$$\hat{y}_i = ax_i + b$$

However, we want to do this according to the generic method for regression presented in class in which we solve the normal equations, $X^tXw = X^ty$ using an appropriate choice of X matrix.

   (a) In R, read in the data and create the X matrix that would be appropriate for estimating a and b.

   (b) Solve the normal equations with your choice of X and report the values you get for a and b.

   (c) Does it appear that people with more education tend to have larger vocabularies?

   (d) It is usually hard to make quantitative statements about the value of a year of education, however, you can do so here in the context of this particulary vocabulary test. Make such a quantitative statement here.

2. This problem deals with the ais.csv data (also in the Canvas data folder) describing various measurements taken on a collection of Austrailian athletes. These data can be read in using the

```
dat = read.csv("ais.csv",stringsAsFactors=FALSE, sep=",")
```

command. For each athlete, the data contain a number of numeric variables we will consider, though we will ignore the gender and sport variables. We are interested in trying to predict the red blood cell count (rcc) from the other numeric variables.

   (a) In R, create the X matrix using variables 3 through 12 as the predictors, while creating the response variable from the 2nd variable, rcc. Solve the normal equations and report the values you get for the regression coefficients.

   (b) Compute your predicted values of the rcc level, $\hat{y}$, the errors ($e = y - \hat{y}$), and the give the sum of squared errors (sse) , defined by $\sum_i e_i^2$.

   (c) In a loop, perform a regression by omitting a single variable from the collection of predictors used above. The first time through the loop you should omit variable 3, the 2nd time through the loop you should include 3 but omit 4, etc. In each case compute your sum of squared errors. Which variable's omission causes the greatest increase in sse? Which variable appears to be the most important in predicting rcc?

3. This problem uses the Nottingham beer sales data which you can include into your R program with

```
data(nottem)
y = nottem
n = length(y)
x = 1:n;
```

Here y is a vector containing the montly beer sales from the Nottingham company for 20 consecutive years ($n = 240$ months), numbered by the vector x.

   (a) Plot the sales data for the various months using both lines and points, as below

```
plot(x,y,type="b")  # b is for "both"
```

(b) We will model this periodic component by including a sine wave and a cosine wave, both n points long, oscillating with 12 points per period (there are 12 months in a year). Specifically, we use the model:

$$\hat{y}_i = a\cos(2\pi x_i/12) + b\sin(2\pi x_i/12) + c$$

Find the values for a, b, c that minimize the sum of squared errors in predicting the $\{y_i\}$. Plot your fitted model on the same plot as the original data, using a different color for each. Be sure to use both lines and points in your plot.

(c) The company hopes that, in addition to the obvious seasonal pattern, their sales grow over time. Consider the model

$$\hat{y}_i = a\cos(2\pi x_i/12) + b\sin(2\pi x_i/12) + c + dx_i$$

that accons for a possible linear trend in sales. Fit this model to that data and plot the fitted results as you did in the previous part, showing both original data and fitted model. Do the results suggest that the company is experiencing growth in sales? Explain your answer in terms of the estimated regression coefficients.

4. The class website has two data sets for this homework assignment. The first has an $n \times p = 1000 \times 50$ data matrix (X) "pred1.dat" with a 1000×1 response vector (y) "resp1.dat." The second has a 1000×500 data matrix "pred2.dat" with a response vector "resp2.dat." These data sets were generated according to the standard linear regression model.

   (a) For each data set use the first half of the data (observations $i = 1, \ldots, n/2$, all p predictors) to get the estimate of w produced by solving the normal equations, $\hat{w}$.

   (b) For each data set, use your estimate of w on the 2nd half of the data set $(n/2 + 1, \ldots, n)$, to get your estimated response variables, $\hat{y}$ and compute and report your total squared error:

$$SSE = \sum_{i=n/2+1}^{n} (\hat{y}_i - y_i)^2$$

5. In *variable selection* we iteratively grow a collection of variables that are used as predictor variables. We begin by finding the single predictor variable that gives the smallest value of $SSE = \sum(y_i - \hat{y}_i)^2$, making this our first predictor variable. We then add another predictor variable (column of our data matrix) by again choosing the variable that, when combined with our first variable, gives the smallest value of SSE. In our simple implementation we implement this greedy search until we have have a fixed number of variables.

   (a) Implement variable selection on the first half of the first data set to identify the 3 best predictor variables. Report the three variables you get and the three decreasing values of SSE they produce.

   (b) Compute the value for SSE on the 2nd half of the data set using the model you have learned. Compare this SSE with that obtained using all predictor variables. Which approach gives a better SSE and why?

6. Use the first half of the 2nd data set to perform ridge regression on w using a parameter of $\lambda = 20$ to get a new $\hat{w}$.

   (a) Using your new $\hat{w}$, compute the estimated response, $\hat{y}$, for the 2nd half of the dataset.

   (b) Compare the resulting sum of squared errors on the 2nd half of the dataset, using both ridge regression and plain regression. If one of the two methods performs better, explain why.

   (c) In general, we do not know what the best choice of the smoothing parameter, λ will be. One way to choose the parameter would be to try a variety of values estimated using the first half of the dataset, choosing the value that gives the best performance on the 2nd half of the dataset. This is cross validation. Use this idea to estimate your best choice of smoothing parameter, λ.

7. A time series is a sequence of observations taken over time, usually with constant time between measurements. The data on the website "time_series.dat" was taken from a time series model

$$x_i = \alpha_1 x_{i-1} + \alpha_2 x_{i-2} + e_i$$

where the $\{e_i\}$ are modeled as independent $N(0, \sigma^2)$ random variables. Estimate the parameters $\alpha_1, \alpha_2, \sigma^2$ from these data.