BMI713 Problem Set 4

Instructions:

Please submit this problem set before class on Tuesday, November 21. Problem sets may be submitted within a week past the due date at a 20% penalty; each person is allowed to submit one problem late (within a week) without penalty. Please include comments in your code to show your work and to clearly indicate your answers. Commented code tends to receive more partial credit!

If you have any questions, please post on the piazza site. This problem set was prepared by Eric Bartell and Jacob Luber, so they will be most prepared to answer questions.

1. Correlations (40 points)

Load in the US census population data posted on piazza under the resources tab. The first column is the year the census was taken and second column is the US population. We want to investigate how the passage of time is correlated with the US population.

(a) Pearson (10 points)

Calculate the Pearson correlation coefficient. This should be coded "by hand" (use the formulation from lecture; do not use the cor() function). Report r between time and US population.

(b) P-value (10 points)

Calculate p-value using both the test statistic t and Fisher's Z-transformation, using the formulas described in class. What do these p-values tell us?

(c) Spearman (10 points)

Calculate r using the spearman correlation (you do not need to do this by hand). Comment on the difference between correlation coefficients.

(d) Add an outlier. (10 points)

The data from 2010 got corrupted, and the replacement value is negative the original value! Set the 2010 value to -1*(old value), and re-perform both Pearson and Spearman correlations (you may use built in functions). Comment on the different outputs.

```
#example corruption code
corruptedCensus <- census
corruptedCensus$Population[corruptedCensus$Year==2010] <- -corruptedCensus$Population[corruptedCensus$Y
```

2. Basic Linear Regression (60 points)

We will perform linear regression on the heights of father-son pairs. Use the following code to load the dataset.

```r
install.packages("UsingR")
library(UsingR)
data(father.son)
```

The first column is fheight for father's height, and the second column is sheight for son's height.

(a) Plot the data (5 points)

On a scatterplot, plot father height vs son height. Does the data appear to be roughly linearly correlated?

(b) Least squares (10 points)

Calculate the least squares line with father height as the independent variable manually.

Comment your code. You do not need to perform any calculus; only calculate the slope and intercept.

(c) Built in (5 points)

Calculate the least squares line using the built in function, and verify that your answer to part (b) is correct.

(d) Plot 2 (5 points)

Add your least squares line to your scatterplot in part (a).

(e) R^2 (5 points)

How well does your line fit the data? Calculate r^2 (manually).

(f) Residuals (5 points)

Plot the residuals for different values of x, and comment on their distribution.

(g) Significance (10 points)

How significant is our slope? Calculate both the t-stat and the p-value, and comment on the correlation's significance.

(h) CI (10 points)

What is the 95% confidence interval for b?

(i) Simple Multivariate (5 points)

Here we add a random variable as a covariate to our dataset:

```r
#add a random column
set.seed(1)
father.son$random <- rnorm(dim(father.son)[1])
```

Use the lm() function to predict son height using both father height and the random covariate. Comment on the significance (p-value) of the to predictive variables.

(j) Extra (5 points)

Given a father's height, we can use a simulation method to construct the 100*(1 - alpha)% confidence interval for the mean of his son's height. First draw 1000 samples each of size 1078 with replacement from the 1078 pairs of father-son heights, then from each sample fit a linear regression model by the method of least squares, and compute the estimated mean of son's height. What are the mean and standard deviation of these 1000 simulated values? Sort these 1000 estimated means in ascending order. Denote the 25th largest as h_25 and the 975th largest as h_975 , which are our estimates of the 0.025 and 0.975 quantiles of the sampling distribution for the mean of son's height. Then the 100(1 - alpha)% confidence interval for the mean of the son's height is (h_25 , h_975). Compute the 95% confidence interval for the mean of son's height if his father is 72 inches tall.