STA 101 Spring 2018
Homework 01 - Due Friday, April 13^{th}

Note: You do not have to use R Markdown to turn in the homework, but the homework must be turned in in a reasonable format. The answers to the questions should be in the body of the homework, and the code used to obtain those answers should be in an appendix. There should be no code in the body of the homework. You can accomplish this in R, Word, LaTex, Google Docs, etc.

1. On Canvas you will find the file `poverty.csv`, which has the following columns:

   Column 1: `Location`: The state.

   Column 2: `PovPct`: The percent of the states population living in poverty (according to the federal definition).

   Column 3: `Brth15to17`: The birth rate per 1000 females 15 to 17 years old

   This is a description of poverty for the listed states. The data was collected in 2002.
   The source of the data is: `Mind On Statistics, 3rd edition, Utts and Heckard`.
   The goal is to estimate the birth rate per 1000 females aged 15 to 17 years old.

   (a) Create a scatter plot of PovPCt and Brth15to17. Does this suggest a linear relationship?

   (b) Find and report the estimated correlation. Would you say this suggests a weak, moderate, or strong linear relationship between PovPct and Brth15to17?

   (c) Find the five number summary of your explanatory variable.

   (d) Would you say there are any outliers in your dataset (unusually small or large vales)? Why or why not?

   (e) What is an appropriate range to predict Brth15to17 using PovPct?

2. Continue with the `poverty.csv` dataset.

   (a) Find and report the estimated linear regression line, using PovPct as your explanatory variable (do not use the information about state).

   (b) Interpret the estimated slope in terms of the problem.

   (c) Interpret the estimated intercept in terms of the problem (if appropriate). If it is not appropriate, explain why.

   (d) Predict the value of Brth15to17 for a state with poverty level 10%.

   (e) Find and interpret the value of R^2.

3. On Canvas you will find the file `rehab.csv`, which has the following columns:

   Column 1: `fitness`: The prior physical fitness of the subject, as `BelowAve`, `Ave`, `AboveAve`.

   Column 2: `days`: The number of days required for a successful completion of therapy.

   This data set describes patient records of 36 male subjects, who all underwent a similar knee surgery.
   The source of this data was *Applied Linear Statistical Models, Fifth Edition, by Kutner, Nachtsheim, Neter, and Li.*
   Assume the goal is to estimate the number of days required for a successful completion of therapy.

   (a) Make a grouped boxplot of days by fitness level. Does it appear that there is a different between the groups?

   (b) Does there appear to be any outliers in `days`?

   (c) How many β's will there be that correspond to the variable `fitness`? List them, and what the corresponding X_i's will be.

   (d) Find the five number summary of `days`.

   (e) Would it make sense for there to be an outlier for `days`? Explain.

4. Continue with the `rehab.csv` dataset.

   (a) Find the estimated regression line, using `days` as your response variable, and `fitness` as your explanatory variable.

   (b) Interpret the value of $\hat{\beta}_1$ in terms of the problem.

   (c) Interpret the value of $\hat{\beta}_2$ in terms of the problem.

   (d) Interpret the value of $\hat{\beta}_0$ in terms of the problem.

   (e) Write down the separate linear models that result from the fact we have one numeric response variable, and one categorical explanatory variable.

5. On Canvas you will find the file `hospital.csv`, which has the following columns:

   Column 1: `InfctRsk`: The percentage of patients who get a secondary infection during their hospital stay.

   Column 2: `MedSchool`: If the hospital was associated with a teaching school (`Yes`) or not (`No`). (X_1)

   Column 3: `Stay`: The average length of stay for patients in days. (X_2)

   The data was collected from 113 hospitals in the US.
   The source of this data was *Applied Linear Statistical Models, Fifth Edition, by Kutner, Nachtsheim, Neter, and Li.*
   The goal is to estimated the percentage of patients who get a secondary infection during their stay.

   (a) Plot the scatter plot of `InfctRsk` and `Stay`. Does there appear to be a linear relationship between the variables?

   (b) Plot a grouped box plot of `InfctRsk` and `MedSchool`. Does there appear to be a difference in the percentage of infection?

(c) Find the estimated correlation between your response variable and the appropriate explanatory variable.

(d) Are there any outliers present? Explain, and either find what they are exactly or estimate the values from the plots if there are outliers.

(e) Based on this dataset, what is a reasonable range of average patient stay to use to predict your response variable?

6. Continue with the `hospital.csv` dataset.

(a) Fit and report the estimated regression function, assuming there is no interaction term.

(b) Write down the two resulting regression equations based on your categorical explanatory variable, simplifying as much as you can.

(c) Interpret the value of $\hat{\beta}_1$.

(d) Interpret the value of $\hat{\beta}_2$.

(e) Predict the infection percentage of a hospital with no medical school affiliation, that has an average patient stay length of 7 days.

7. Continue with the `hospital.csv` dataset.

(a) Fit and report the estimated regression function, assuming there **is** an interaction term.

(b) Write down the two resulting regression equations based on your categorical explanatory variable, simplifying as much as you can.

(c) Interpret the slope in each of your regression models from (b).

(d) Interpret the intercept in each of your regression models from (b) (if appropriate). If it is not appropriate, explain why.

(e) Predict the infection percentage of a hospital with no medical school affiliation, that has an average patient stay length of 7 days.

8. Answer the following questions with TRUE or FALSE. It is good practice for exams to explain your answer, whether your answer is TRUE or FALSE.

(a) Correlation measures the strength of any relationship between X and Y.

(b) Adding an interaction term will not (in general) change the value of a prediction.

(c) The sign of the slope and the sign of the sample correlation should match.

(d) If an error for a regression line is positive, that means we underestimated the value in our dataset.