STA 101 Spring 2018
Homework 2 - Due Friday, April 20^{th}

1. Recall the data `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

   (a) Using the simple linear regression model, assess normality of the errors using plots and a test. Report back any relevant p-values, and state your final conclusion on if the errors are approximately normal.

   (b) Using the simple linear regression model, assess constancy of error variance using plots and a test.

   (c) Find and remove any outliers, using the criteria of your choice. If there are no outliers, simply state this.

   (d) Compare the values of the slopes for the models with, and without, outliers. What was the absolute difference? If you had no outliers, there would be no difference.

2. Continue with the `poverty.csv` dataset. Use the model with no outliers (if you found any).

   (a) Find the 90% confidence intervals for all values of β.

   (b) Interpret the confidence interval for β_1 in terms of the problem.

   (c) Does your confidence interval from (a) suggest that X_1 has a significant linear relationship with Y? Explain.

   (d) Find the test-statistic and the p-value for testing if there is a significant linear relationship between X_1 and Y.

   (e) Interpret the p-value in (d) in terms of the problem.

3. Recall the datset `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)

   (a) Using the linear regression model without interactions, assess normality of the errors using plots and a test. Report back any relevant p-values, and state your final conclusion on if the errors are approximately normal.

   (b) Using the linear regression model without interactions, assess constancy of error variance using plots and a test.

   (c) Find and remove any outliers, using the criteria of your choice. If there are no outliers, simply state this.

   (d) Compare the values of the slopes for the models with, and without, outliers. What was the absolute difference? If you had no outliers, there would be no difference.

4. Continue with the `hospital.csv` dataset. Use the model with no outliers (if you found any).

   (a) Find the 90% confidence intervals for all values of β.

   (b) Interpret the confidence interval for β_2 in terms of the problem.

   (c) Does your confidence interval from (a) suggest that X_2 has a significant linear relationship with Y? Explain.

   (d) Find the test-statistic and the p-value for testing if there is a significant linear relationship between X_2 and Y.

   (e) Interpret the p-value in (d) in terms of the problem.

5. Continue with the `hospital.csv` dataset. Use the model with no outliers (if you found any).

   (a) Fit the model with an interaction term.

   (b) Find the confidence intervals for all of the β_i.

   (c) Based on your confidence intervals, which β's should be retained in the model? Explain.

   (d) Based on your confidence intervals, what is the largest change in Y we can expect when an X changes by one unit (you may exclude the interaction term)?

6. On Canvas there is the dataset `fish.csv`, which has the following columns:

   Column 1: `Age`: The age of the fish in days. (X_1)

   Column 2: `Temp`: Temperature of four tanks (X_2)

   Column 3: `Length`: The length of the fish in mm. (Y)

   The goal is mainly to see if temperature changes how fast the fish grow. Age is recorded as well,

   (a) Using the linear regression model without interactions, assess normality of the errors using plots and a test. Report back any relevant p-values, and state your final conclusion on if the errors are approximately normal.

   (b) Using the linear regression model without interactions, assess constancy of error variance using plots and a test.

   (c) Find and remove any outliers, using the criteria of your choice. If there are no outliers, simply state this.

(d) Compare the values of the slopes for the models with, and without, outliers. What was the absolute difference? If you had no outliers, there would be no difference.

7. Continue with the `fish.csv` dataset. Use the model with no outliers (if you found any).

   (a) Find the 90% confidence intervals for all values of β.

   (b) Interpret the confidence interval for β_1 in terms of the problem.

   (c) Does your confidence interval from (a) suggest that X_2 has a significant linear relationship with Y? Explain.

   (d) Find the test-statistic and the p-value for testing if there is a significant linear relationship between X_2 and Y.

   (e) Interpret the p-value in (d) in terms of the problem.

8. Continue with the `fish.csv` dataset. Use the model with no outliers (if you found any).

   (a) Fit the model with an interaction term.

   (b) Find the confidence intervals for all of the β_i.

   (c) Based on your confidence intervals, which β's should be retained in the model? Explain.

   (d) Based on your confidence intervals, what is the largest change in Y we can expect when an X changes by one unit (you may exclude the interaction term)?

9. 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) The model with the most X variables will have the largest R^2 value.

   (b) When a confidence interval for a β_i does not contain zero, the corresponding X_i has a significant linear relationship with Y (you may exclude interaction terms).

   (c) Outliers may significantly change the value/s of β/s

   (d) The smaller the value of AIC, the better a model fits according to this criteria.