Homework #2
Linear Models

ENGM 182
Data Analytics
Due 5pm Monday April 15, 2019

1 Tests for normality

1.1 Are body temperatures normally distributed?

Download the data on body temperature from a paper in the *Journal of the American Medical Association* from Canvas (bodytemp.csv) and load it into R. Males are coded 0, females are coded 1.

1. Using R, create a qqnorm plot of body temperature. What do you see? Does the data look normally distributed?
2. Use the shapiro.test function to conduct a Shapiro-Wilk normality test (recall that low p-values reject the null that the data are normally distributed.

2 t-test review: test whether body temperatures are 98.6 degrees Fahrenheit

1. Use R to conduct a t-test for men and women separately to test the null hypothesis that temperature is 98.6 degrees Fahrenheit.
2. Use R to conduct a t-test to test the null hypothesis that men and women have the same body temperature.

3 Effects of gender and heart rate on body temperature

Use the linear regression function in R to determine the relationship between n explanatory variables $x_1, x_2, ...x_n$ and an outcome variable y.
For the body temperature dataset, assume the following linear relationship:

$$\text{temp} = \beta_0 + \beta_1 \cdot \text{sex} + \beta_2 \cdot \text{bpm} + \text{residuals},$$

where "residuals" are uncorrelated error terms. Answer the following questions:

1. Find estimates for the intercept β_0 and the slopes β_1, β_2.
2. At the confidence level $\alpha = 0.05$, which of the coefficients are significant?
3. Do the residuals (errors) look normally distributed?
4. Conduct a Shapiro Wilkes test on the residuals.

4 New Orleans Airbnb price analysis

Download a subset of the data on Airbnb prices from Canvas in file

`NOLAlistingsJune2016_subset2.csv.`

1. Examine whether price is normally distributed. If not, make an appropriate transform prior to carrying out the rest of the analysis. The following paper might provide some useful background:

   `http://support.sas.com/resources/papers/proceedings12/430-2012.pdf`

2. Create an OLS model of price as a function of neighourhood, room type, and availability.

3. Interpret the model output. What coefficients are significant? What kinds of impact do the explanatory variables have? Are the residuals normally distributed?

5 Cooking Procedures

1. Download the file GBDOE-French Fry DOE.csv from Canvas. The data set contains information on the following variables: Potato Type, Cooking Oil Type, Cooking Temp, Cooking Time, Taste Rating. Build a linear model to explain Taste Rating as a function of the other variables in the data set. How well does your model explain taste?

2. BONUS: find a way to present the results of your analysis in graphical form and/or a well-formatted table.

6 Food Enjoyment

Download the file foodenjoyment.csv from Canvas. The data set contains information on the enjoyment of different foods and condiments.

1. Build a linear model to examine the effect of these foods and condiments on enjoyment. What is driving enjoyment in this data set?

2. BONUS: find a way to present the results of your analysis in graphical form and/or a well-formatted table.

7 Homework Time

How many hours did you spend on this assignment (including question 8 if attempted)?

8 Challenge Problem (Optional)

Download the file NYCairbnb.csv from canvas. This dataset contains information on Airbnb listings in New York City. Use this dataset to build a linear model that predicts price. You may use any variables that appear in the dataset for the regression but you must justify your model. Some possible things you might want to look at include normality of the residuals, significance of the coefficients, AIC, and R^2. Present your results using any appropriate text, figures, and/or tables.