Assignment 5 - Multiple Linear Regression

your name goes here

NOT Due: Wednesday, February 21, 2018, noon, to Sakai

Summary

The tasks in this homework assignment focus on creating, interpreting, and performing inference on the multiple regression model.

Assignment

1. Set up the $\mathbf{X}$ matrix and β vector for each of the following regression models (assume $i = 1, 2, 3, 4$).

   (a)
   $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1} X_{i2} + \epsilon_i$$

   (b)
   $$\log Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

   (c)
   $$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1}^2 + \epsilon_i$$

   (d)
   $$\sqrt{Y_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 \log_{10} X_{i2} + \epsilon_i$$

2. Why is it not meaningful to attach a sign to the coefficient of multiple correlation ($R = \sqrt{R^2}$), even though we do attach a sign in the simple linear regression case?

3. Consider a dataset collected from `https://www.cars.com/` on Toyota Corollas. A group of students scraped the website for information on Toyota Corollas within a particular geographic region and collected information on model, price, year, mileage, and location.

```
cars = read.csv("http://www.amherst.edu/~nhorton/stat135/carscollated.csv")
cars = mutate(cars, Mileage = Mileage/1000) # mileage now in thousands of miles
```

   (a) Plot y=Price vs x=Mileage.

   (b) Construct linear models using the following three sets of explanatory variables:

      i. Mileage
      ii. Mileage and Mileage^2 (to inlude a squared term, use `I(Mileage^2)`)
      iii. Mileage and Mileage^2 and Location

   For each of the models report the coefficients, R^2 values, and residual plots. To save paper, consider the following code:

```
mylm <- lm(response ~ explanatory variables, data=...)
glance(mylm)
tidy(mylm)
augment(mylm) %>%
  ggplot(aes(x=.fitted, y=.resid)) +
  geom_hline(yintercept = 0)
```

(c) Which one of the three models do you think is best? Why? Interpret the output you produced in the previous question.

(d) Interpret the `LocationLos Angeles` coefficient from the model. (Hint: you need to first figure out the reference group.)

(e) Do the coefficients seem unusual at all to you? Can you hypothesize any additional variables that might seem to be driving the location differences?

(f) What can you say about the coefficient on the squared term?

(g) Using the model with the squared coefficient and the location variable, test whether or not there is a difference in average price in San Francisco versus Detroit (conditional on mileage).

4. Data were collected on the volume of users on the Northampton Rail Trail in Florence, Massachusetts. Variables in the dataset include the number of crossings on a particular day (measured by a sensor near the intersection with Chestnut Street, volume), the average of the min and max temperature in degrees Fahrenheit for that day (avgtemp), and a dichotomous indicator of whether the day was a weekday or a weekend/holiday (weekday).

```
data(RailTrail)
```

In predicting volume, do average temperature and weekday interact?

In answering the question: (a) provide appropriate models / figures, and (b) explain in your own words what it means for variables to interact.