STOR 455 Homework #2

40 points - Due Wednesday 2/9 at 5:00pm

Situation: Suppose that you are interested in purchasing a used vehicle. How much should you expect to pay? Obviously the price will depend on the type of vehicle that you get (the model) and how much it's been used. For this assignment you will investigate how the price might depend on the vehicle's year and mileage.

Data Source: To get a sample of vehicles, begin with the UsedCars CSV file. The data was acquired by scraping TrueCar.com for used vehicle listings on 9/24/2017 and contains more than 1.2 million used vehicles. For this assignment you will choose a vehicle *Model* from a US company for which there are at least 100 of that model listed for sale in North Carolina. Note that whether the companies are US companies or not is not contained within the data. It is up to you to determine which *Make* of vehicles are from US companies. After constructing a subset of the UsedCars data under these conditions, check to make sure that there is a reasonable amount of variability in the years for your vehicle, with a range of at least six years.

Directions: The code below should walk you through the process of selecting data from a particular model vehicle of your choice. Each of the following two R chunks begin with {r, eval=FALSE}. eval=FALSE makes these chunks not run when I knit the file. Before you knit these chunks, you should revert them to {r}.

```
library(readr)

# This line will only run if the UsedCars.csv is stored in the same directory as this notebook!
UsedCars <- read_csv("UsedCars.csv")

StateHW2 = "NC"

# Creates a dataframe with the number of each model for sale in North Carolina
Vehicles = as.data.frame(table(UsedCars$Model[UsedCars$State==StateHW2]))

# Renames the variables
names(Vehicles)[1] = "Model"
names(Vehicles)[2] = "Count"

# Restricts the data to only models with at least 100 for sale
# Vehicles from non US companies are contained in this data
# Before submitting, comment this out so that it doesn't print while knitting
Enough_Vehicles = subset(Vehicles, Count>=100)
Enough_Vehicles
```

```
# Delete the ** below and enter the model that you chose from the Enough_Vehicles data.
ModelOfMyChoice = "**"

# Takes a subset of your model vehicle from North Carolina
MyVehicles = subset(UsedCars, Model==ModelOfMyChoice & State==StateHW2)

# Check to make sure that the vehicles span at least 6 years.
range(MyVehicles$Year)
```

MODEL #1: Use Mileage as a predictor for Price

1. Calculate the least squares regression line that best fits your data using *Mileage* as the predictor and *Price* as the response. Interpret (in context) what the slope estimate tells you about prices and mileages of your used vehicle model. Explain why the sign (positive/negative) makes sense.
2. Produce a scatterplot of the relationship with the regression line on it.
3. Produce appropriate residual plots and comment on how well your data appear to fit the conditions for a linear model. Don't worry about doing transformations at this point if there are problems with the conditions.
4. Find the five vehicles in your sample with the largest residuals (in magnitude - positive or negative). For these vehicles, find their standardized and studentized residuals. Based on these specific residuals, would any of these vehicles be considered outliers? Based on these specific residuals, would any of these vehicles possibly be considered influential on your linear model?
5. Determine the leverages for the vehicles with the five largest absolute residuals. What do these leverage values say about the potential for each of these five vehicles to be influential on your model?
6. Determine the Cook's distances for the vehicles with the five largest absolute residuals. What do these Cook's distances values say about the influence of each of these five vehicles on your model?
7. Compute and interpret in context a 95% confidence interval for the slope of your regression line. Interpret (in context) what the confidence interval for the slope tells you about prices and mileages of your used vehicle model.
8. Test the strength of the linear relationship between your variables using each of the three methods (test for correlation, test for slope, ANOVA for regression). Include hypotheses for each test and your conclusions in the context of the problem.
9. Suppose that you are interested in purchasing a vehicle of this model that has 50,000 miles on it (in 2017). Determine each of the following: 95% confidence interval for the mean price at this mileage and 95% prediction interval for the price of an individual vehicle at this mileage. Write sentences that carefully interpret each of the intervals (in terms of vehicles prices).
10. Experiment with some transformations to attempt to find one that seems to do a better job of satisfying the linear model conditions. Include the summary output for fitting that model and a scatterplot of the original data with this new model (which is likely a curve on the original data). Explain why you think that this transformation does or does not improve satisfying the linear model conditions.
11. According to your transformed model, is there a mileage at which the vehicle should be free? If so, find this mileage and comment on what the "free vehicle" phenomenon says about the appropriateness of your model.
12. Again suppose that you are interested in purchasing a vehicle of this model that has 50,000 miles on it (in 2017). Determine each of the following using your transformed model: 95% confidence interval for the mean price at this mileage and 95% prediction interval for the price of an individual vehicle at this mileage. Write sentences that carefully interpret each of the intervals (in terms of vehicle prices).

MODEL #2: Again use Mileage as a predictor for Price, but now for new data

13. Select a new sample from the UsedCar dataset using the same *Model* vehicle that was used in the previous sections, but now from vehicles for sale in a different US state. You can mimic the code used above to select this new sample. You should select a state such that there are at least 100 of that model listed for sale in the new state.
14. Calculate the least squares regression line that best fits your new data and produce a scatterplot of the relationship with the regression line on it.
15. How does the relationship between *Price* and *Mileage* for this new data compare to the regression model constructed in the first section? Does it appear that the relationship between *Mileage* and *Price* for your *Model* of vehicle is similar or different for the data from your two states? Explain.

16. Again suppose that you are interested in purchasing a vehicle of this model that has 50,000 miles on it (in 2017) from your new state. How useful do you think that your model will be? What are some possible cons of using this model?

MODEL #3: Use Year as a predictor for Price

17. What proportion of the variability in the *Mileage* of your North Carolina vehicles' sale prices is explained by the *Year* of the vehicles?

18. Calculate the least squares regression line that best fits your data using *Year* as the predictor and *Price* as the response. Produce a scatterplot of the relationship with the regression line on it.

19. Produce appropriate residual plots and comment on how well your data appear to fit the conditions for a simple linear model. Don't worry about doing transformations at this point if there are problems with the conditions.

20. Experiment with some transformations to attempt to find one that seems to do a better job of satisfying the linear model conditions. Include the summary output for fitting that model and a scatterplot of the original data with this new model (which is likely a curve on the original data). Explain why you think that this transformation does or does not improve satisfying the linear model conditions.

21. How do the transformed models, using either *Year* or *Mileage* as the predictor for your model of vehicle for sale in North Carolina compare? Does one of the models seem "better" or do they seem similar in their ability to predict *Price*? Explain.