STOR 455 Homework #8

25 points - Due 4/13 at 5:00pm

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

Data Source: Your sample of cars will again be taken from the UsedCar CSV file on Sakai. The data was acquired by scraping TrueCar.com for used car listings on 9/24/2017 and contains more than 1.2 million used cars.

For this assignment, you will need to select six new samples, each with *exactly* 50 vehicles, for six different *Model* of used vehicles for sale in North Carolina from the UsedCar dataset. There will likely be more than 50 of your selected models for sale in North Carolina, so you should randomly select those 50 vehicles from the larger number that are available. The six models of vehicles should be selected such that three models of vehicles are selected from Japanese companies, and another three from US companies (i.e. *Make*; It does not matter where the cars were actually manufactured). Within each country, you should select a compact car, a mid-sized car, and a SUV (Note that the country and types of vehicles are not given in the data and are for you to determine). You should add new variables to the dataframes for the country of the company and type of vehicle (compact vs mid-sized vs SUV) and combine these six samples into one dataframe (just as rbind was used in a previous assignment). When selecting these samples make sure to use set.seed(). This will select the same sample each time that you run (and knit) your code.The code below is an example of how you could select a random sample of 50 cars for a given model:

```
# Suppose that NCUsedCars_Honda is a subset of
# only Hondas sold in NC from the Usedcars dataset.

library(dplyr)
set.seed(8675309)
Civic = sample_n(subset(NCUsedCars_Honda, Model=='Civic'), 50)
```

One Way ANOVA

1. Produce a set of side-by-side boxplots to compare the price distributions of your three types of vehicles (not the models). Comment on any obvious differences in the distributions.

2. Produce summary statistics (mean and standard deviation) for each of the groups (vehicle types) AND the entire sample of vehicle prices.

3. Based on just what you see in the boxplots and summary statistics comment on whether you think there are significant differences in the mean prices among your three vehicle types. Also comment on any concerns you see about the conditions for the ANOVA for means model.

4. Construct an ANOVA model for the mean price by vehicle type. Include the output showing the ANOVA table; state hypotheses, and provide a conclusion in the context of your data.

5. Produce plots and/or summary statistics to comment on the appropriateness of the following conditions for your data: normality of the residuals, and equality of the variances. If you find that the conditions are *not* met, You can still continue with analysis of your data for this homework. We will soon discuss how to deal with violations of these conditions.

6. If your ANOVA model indicates that there are significant differences among the vehicle type price means, discuss where the significant differences occur using Tukey HSD methods. If your ANOVA indicates there are not significant differences among the vehicle type price means, determine how different your means prices would need to be in order to find a significant difference using the Tukey HSD methods.

Two Way ANOVA

7. Construct an ANOVA model for the mean price using the country of the company and the type of vehicle as predictors (without an interaction). Include the output showing the ANOVA table; state hypotheses and provide a conclusion in the context of your data. If your ANOVA model indicates there are significant differences among the vehicle price means: Discuss where the significant differences occur using Tukey HSD methods.

8. Produce plots and/or summary statistics to comment on the appropriateness of the following conditions for your data: normality of the residuals, and equality of the variances.

9. Construct an ANOVA model for the mean price using the country of the company and the type of vehicle as predictors with the interaction. Include the output showing the ANOVA table; state hypotheses and provide a conclusion in the context of your data. If your ANOVA indicates that there are significant differences among the car price means: Discuss where the significant differences occur using Tukey HSD methods.

10. Produce two interaction plots for the previous model. If you found significant interactions in your hypothesis test, comment on how these interactions are shown in the plot. If you did not find significant interactions in your hypothesis test, comment on how the (lack of) interactions are shown in the plot.

Additional Topics

11. Recall that we can also handle a categorical predictor with multiple categories using ordinary multiple regression if we create indicator variables for each category and include all but one of the indicators in the model. Run an ordinary multiple regression to predict *Price* using the country of the company, the type of vehicle, and the interaction between the two as predictors. Interpret each of the coefficients in the "dummy" regression by what they mean in the context of mean prices.

12. One possible drawback of the analysis for this assignment is that different people might have chosen vehicles with quite different mileages when collecting their samples. Thus an apparent "difference" between two countries or vehicle types might be due to one sample having considerably more higher mileage vehicles in it than another. Construct a model that allows you to check for mean price differences between your vehicles from the model constructed in question 11 after accounting for variability due to the mileage of the vehicles. Explain how you use the output from the model to address this question.