Homework 10: ANOVA

Mario Giacomazzo

November 22, 2023

Instructions:

The purpose of this homework assignment is to practice ANOVA. Make sure you read each question carefully. In each question, I will give you a task to do, and I will tell you what I want you to output. You can write as much code as you want in each code chunk, but make sure you complete the task and only print the output I asked you to print. Don't sort the data unless you are told to sort the data. You should remove the "#" sign in each code chunk before writing your code. Also, if you see the comment "#DO NOT CHANGE", then I don't want you to make any modifications to that code. **You should knit your RMD file to a PDF after you answer every question.**

After you are done, knit the RMarkdown file to PDF and submit the PDF to Gradescope under HW10.

For this assignment, you will be working with a dataset containing information on vehicle sales on Craigslist. The data was acquired by scraping Craigslist for vehicles for sale across the southeastern United States. In this dataset, there are 73,033 different observations and 26 variables. In this assignment, we will be focused on only the three variables: *price*, *manufacturer*, and *condition*. The data frame **vehicle2** shows a preview of this data.

```
vehicle = read_csv("vehiclesSE.csv")
vehicle2 = vehicle[,c("price","manufacturer","condition")]
head(vehicle2)
```

```
## # A tibble: 6 x 3
##   price manufacturer condition
##   <dbl> <chr>        <chr>
## 1 33590 gmc          good
## 2 22590 chevrolet    good
## 3 39590 chevrolet    good
## 4 30990 toyota       good
## 5 15000 ford         excellent
## 6 27990 gmc          good
```

We want to investigate if the average price of cars is different for different groups. Therefore, for all future models, the variable *price* will be the response variable.

Questions

Q1 (3 Points)

First, I want you to investigate the *manufacturer* variable in **vehicle2**. Write code to find the number of unique manufacturers in the dataset and the names of manufacturers represented at least 2,500 times in this dataset. Your output should show how you got the answers you give below the code chunk.

- Number of Unique Manufacturers: REPLACE WITH NUMBER
- Manufacturers with At Least 2,500 Observations: REPLACE WITH MANUFACTURERS

Q2 (3 Points)

Second, I want you to investigate the *condition* variable in **vehicle2**. Write code to find the number of unique conditions in the dataset and the names of the two conditions that occur the least. Your output should show how you got the answers you give below the code chunk.

- Number of Unique Conditions: REPLACE WITH NUMBER
- Two Conditions that Occur the Least: REPLACE WITH CONDITIONS

Q3 (2 Points)

Now that we have investigated both the manufacturer and condition variable a little. I want you to choose 3 manufacturers that have at least 2,500 observations. You can pick any of the manufacturers listed in your answer to Q1.

Then, I want you to create a data frame named **vehicle3** that contains all the data in **vehicle2** for those 3 manufacturers that DO NOT HAVE the two conditions that occur the least.

Be very careful here and then use the `unique()` function for the *manufacturer* and *condition* variables in **vehicle3** to confirm that your code worked. This should be your only output from the code chunk.

All future questions will only involve the data in **vehicle3**.

Q4 (4 Points)

Construct side-by-side boxplots to investigate the distribution of price for different manufacturers. Then, calculate the sample standard deviation of price for each of the different manufacturers. Then, use the rule-of-thumb mentioned on page 220 of the textbook to determine if we should be worried about the homoscedasticity assumption being violated. I should see the boxplot, the standard deviations, and a calculation of a ratio based on page 220 of the textbook.

Below the code chunk I want you to write 1-3 sentences about whether or not we should be worried about the homoscedasticity assumption? Use your output to make your argument on why or why not we should be worried.

- Comment about Homoscedasticy: REPLACE WITH COMPLETE SENTENCES

Q5 (4 Points)

This question is identical to Q4 except I want you to look at the *condition* variable instead of the *manufacturer* variable.

- Comment about Homoscedasticy: REPLACE WITH COMPLETE SENTENCES

Q6 (4 Points)

Read Chapter 8.1 in the textbook about *Levene's Test for Homogeneity of Variances*. In Q4 and Q5, we used a simple rule-of-thumb to determine if we were concerned about the homoscedasticy assumption. Levene's test is a hypothesis test to test the following hypotheses:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$$
$$H_a : \sigma_i^2 \neq \sigma_j^2 \text{ for some } i \neq j$$

where K is the number of groups.

I want you to conduct Levene's Test twice, once for looking at the variance of price for different manufacturers and once for looking at the variance of price for different conditions. Look at the link https://statmethods.com/home/levenes-test-for-homogeneity-of-variance/ to learn how to do this in R using the `leveneTest()` function from the **car** package. I want to see the output from this test for both *manufacturer* and *condition* variables.

Note: I am suppressing a warning in the PDF output. Do not be worried about the warning when you run `leveneTest()` on your data.

Finally, compare the results of this hypothesis test to Q4 and Q5. Below the code chunk, write 1 to 4 sentences discussing if there is agreement or disagreement between using the rule-of-thumb from page 220 and Levene's test for homogeneity of variances. Reference elements from the output from Levene's test when making your argument.

- Comment about Comparison of Results: REPLACE WITH COMPLETE SENTENCES

Q7 (6 Points)

Assume that the homoscedasticity assumption is reasonable regardless of what you discovered from previous questions.

Fit a one-way ANOVA model for the relationship between *price* and *manufacturer* using the data in **vehicle3**. Print out the ANOVA table for this model using the `summary()` function.

Also, in this output I want you to find the confidence intervals for every pairwise comparison of mean price using a Bonferroni adjustment so the family-wise error rate you are targeting is 0.05. Since there are three manufacturers you chose, there should be 3 confidence intervals calculated. Your output should show your calculations of the lower and upper bounds for the three confidence intervals. I want you to calculate the confidence intervals by hand hand using functions like `qt()` to find the critical value,`tapply()` to find group means, etc

Finally, I want you to replace the word REPLACE below with the manufacturer names and write the confidence interval using interval notation like (3,10). Round your lower and upper bounds in all confidence intervals to 2 decimal places.

The output should contain the ANOVA table and all the work you did to calculate the confidence intervals. Show output as much as possible on the confidence intervals so it is clear to the grader that your work is reasonable and correct.

- CI for REPLACE versus REPLACE: (COMPLETE,COMPLETE)
- CI for REPLACE versus REPLACE: (COMPLETE,COMPLETE)
- CI for REPLACE versus REPLACE: (COMPLETE,COMPLETE)

Q8 (5 Points)

Assume that the homoscedasticity assumption is reasonable regardless of what you discovered from previous questions.

Fit a one-way ANOVA model for the relationship between *price* and *condition* using the data in **vehicle3**. Print out the ANOVA table for this model using the `summary()` function.

Also, I want you to use Tukey's HSD to find all of the confidence intervals for every pairwise comparison of the average price between different conditions. Use the `TukeyHSD()` function to do this but target a familywise

error rate of 0.01 instead of 0.05. The `TukeyHSD()` function targets 0.05 by default. You will need to figure out how to do this by looking at the documentation of `TukeyHSD()`. Print the output from `TukeyHSD()`.

Based off the confidence intervals in the output, which pairs of conditions do we not have evidence that there is a difference in average price? Below the code chunk answer this question in 1 to 4 sentences citing the confidence intervals in the output associated with these pairs and why those intervals do not provide evidence. If you believe all intervals provide evidence that there is a significant difference in the average price for every pair of conditions, then comment on this with a reason why all intervals provide evidence that there is a significant difference.

- Comment about CI's: REPLACE WITH COMPLETE SENTENCES