Homework 5: Outliers and Hypothesis Testing

FIRSTNAME LASTNAME

September 18, 2023

Instructions:

The purpose of this homework assignment is to look more into inspecting for outliers and investigating the impact that outliers have in hypothesis testing. 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 HW5.

Questions

Q1 (2 Points)

The dataset **EuroEnergy** from the **AER** package (renamed as **energy**) contains the GDP and energy consumption of 20 different countries in Europe from the year 1980. Unfortunately, some of the most powerful but forgotten European countries were left out. For this reason, I added the GDP and energy consumption of the countries: Richenclean, Poorendirti, and Bankenstank. The dataset **energy2** was created by using the `rbind()` function to combine the original dataset **energy** with the missing dataset **missing.data**. This function is useful for stacking one dataset on top of another datset. Notice that you can assign names to the rows of a data frame in R.

We plan on building models to understand the relationship between *energy* and *gdp*. For all following questions, *energy* will be the response variable and *gdp* will be the predictor variable. Create a scatterplot using the data in **energy2** to investigate this relationship. Use the option `pch=16` to change the type of points. The only output should be the plot.

```
data("EuroEnergy") #DO NOT CHANGE
energy = EuroEnergy #DO NOT CHANGE

missing.data=data.frame(gdp=c(400000,40000,1000000),energy=c(30,300000,800000)) #DO NOT CHANGE
row.names(missing.data)=c("Richenclean","Poorendirti","Bankenstank") #DO NOT CHANGE

energy2=rbind(energy,missing.data) #DO NOT CHANGE

#
```

Q2 (3 Points)

Calculate the leverage for each of these 23 countries and create a new variable in **energy2** called *lev* to store the leverage. Then, sort the data in **energy2** from low leverage to high leverage. Only show the top 10

countries of this sorted **energy2** dataset. These would be the 10 countries with the lowest leverage.

```
#
```

Q3 (4 Points)

Fit the appropriate simple linear regression model and add the standardized residuals (use the updated formula from 4.4) into the dataset **energy2** as a new variable called *std.res*. Then, plot the variable **std.res** versus the **gdp** using a scatterplot with the `pch=16` argument. Also, I want dashed red lines at -2 and +2.

Standardized residuals outside the -2 and +2 lines indicate unusual residuals that could be resulting from rare situations or they could be caused by bad data. The plot should be your only output from this code.

```
#
```

Q4 (2 Points)

I want you to add a variable to **energy2** named *CookD* that contains Cook's Distance for each point according to the formula in the textbook. Then, I want you to create a new dataset **final_energy** that contains all the data in **energy2** except for countries that have a very unusual Cook's distance. Finally, use the `str()` function to show a preview of **final_energy**. This should be the only output.

```
#
```

Q5 (6 Points)

Refit the linear model on the dataset **final_energy**. Use the `summary()` function on your model to output the model information to your audience. The output from `summary()`

Finally, based off the output, conduct a t-Test for the slope. What decision would you make regarding the hypotheses and why? Your decision is written for statisticians. Then, I want you to try to interpret this decision to an audience who may have very little understanding of math or stats. For example, you should be able to explain the slope without using the word slope. Put your response to this prompt below in the appropriate space.

```
#
```

Response in Complete Sentences: REPLACE THIS SENTENCE IN ALL CAPS WITH YOUR ANSWER IN COMPLETE SENTENCES