Homework 4: Outliers and Hypothesis Testing

Ruonan Zhao

September 19, 2024

Instructions:

The purpose of this assignment is to investigate outliers and their impact on hypothesis testing. Follow the instructions closely, ensuring that your code produces the exact output requested. Avoid modifying data or plots unless instructed. Pay attention to the comments and make sure to remove any unnecessary # signs before adding your code. If a code block contains #DO NOT CHANGE, follow the instructions accordingly. You should include all code used (do not hide code cells). **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 Homework 4.

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. We 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.

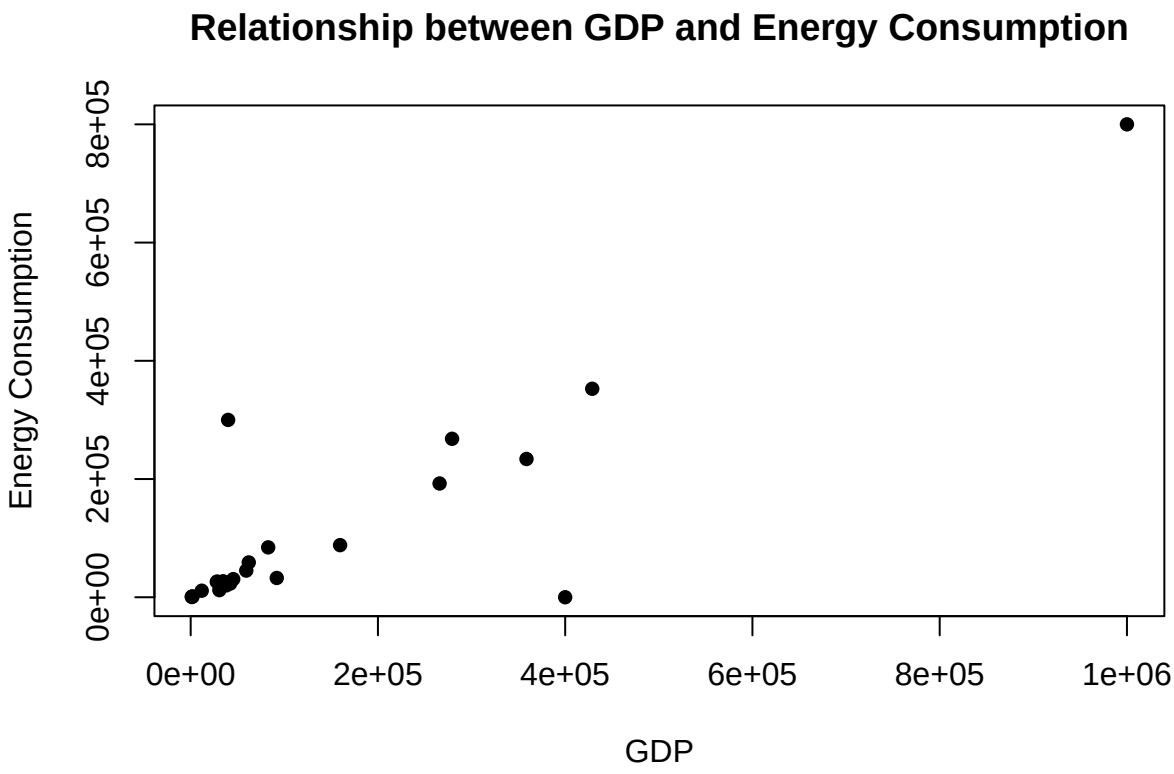
We want to investigate the relationship between *gdp* (predictor) and *energy* (response). Create a scatterplot using **energy2** to investigate this relationship. Use `pch=16` to change the type of points. The only output should be the plot.

```
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

plot(energy2$gdp, energy2$energy,
     main = "Relationship between GDP and Energy Consumption",
     xlab = "GDP",
     ylab = "Energy Consumption",
     pch = 16)
```



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 12 countries of this sorted **energy2** dataset with the lowest leverage.

```
avg.x = mean(energy2$energy)
dev_squared = (energy2$energy - avg.x)^2

leverage = 1/nrow(energy2) + dev_squared / sum(dev_squared)

energy2$lev = leverage

head(energy2[order(energy2$lev, decreasing = F), ], 12)
```

```
##                gdp energy        lev
## Spain       159602  88148 0.04442165
## Netherlands  82804  84416 0.04470663
## Belgium      62049  58894 0.04766224
## Sweden       59350  45132 0.04998491
## Italy       265863 192453 0.05165080
## Turkey       91946  32619 0.05254008
## Austria      45451  30633 0.05298445
## Denmark      34540  27049 0.05381329
## Finland      28388  26405 0.05396590
```

```
## Norway       27914  26086 0.05404190
## Switzerland  42238  23234 0.05473362
## Greece       38039  20119 0.05551418
```

Q3 (4 Points)

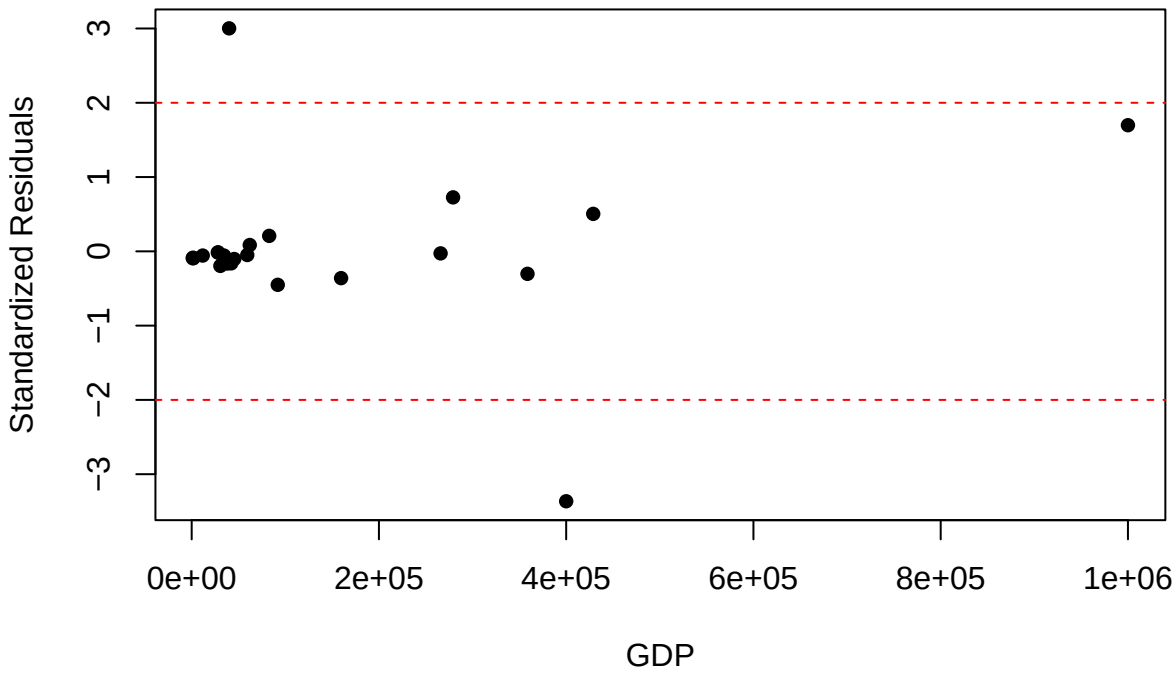
Fit the appropriate simple linear regression model and add the standardized residuals (use the updated formula from 4.4) into the dataset **energy2** as a new variable called *std.res*. Then, plot the variable **std.res** versus the **gdp** using a scatterplot with the `pch=16` argument. Plot dashed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

```
model <- lm(energy ~ gdp, data = energy2)

energy2$std.res = rstandard(model)

plot(energy2$gdp, energy2$std.res, pch=16, xlab="GDP", ylab="Standardized Residuals")
abline(h=2, col="red", lty=2)
abline(h=-2, col="red", lty=2)
```



Q4 (2 Points)

Add a variable to **energy2** named *CookD* that contains Cook's Distance for each point according to the formula in the textbook. Then, 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.

```
energy2$CookD <- cooks.distance(model)

final_energy = energy2 %>% filter(CookD < 1)

str(final_energy)
```

```
## 'data.frame':    22 obs. of  5 variables:
##  $ gdp    : num  45451 62049 2003 34540 28388 ...
##  $ energy : num  30633 58894 1211 27049 26405 ...
##  $ lev    : num  0.053 0.0477 0.0608 0.0538 0.054 ...
##  $ std.res: num  -0.1033 0.0849 -0.0906 -0.0568 -0.0149 ...
##  $ CookD  : num  3.00e-04 1.92e-04 2.76e-04 9.48e-05 6.69e-06 ...
```

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()` should be your only output.

Finally, based off the output, conduct a t-Test for the slope. What decision would you make regarding the hypotheses and why? First, write your decision is written for statisticians. Then, I want you 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.

```
model_final <- lm(energy ~ gdp, data = final_energy)
summary(model_final)
```

```
##
## Call:
## lm(formula = energy ~ gdp, data = final_energy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -228755  -25482  -17461   12664  253696
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.603e+04  2.400e+04   1.084  0.29109
## gdp         5.069e-01  1.365e-01   3.713  0.00137 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 86110 on 20 degrees of freedom
## Multiple R-squared:  0.4081, Adjusted R-squared:  0.3785
## F-statistic: 13.79 on 1 and 20 DF,  p-value: 0.001375
```

Response in Complete Sentences for Statistician: Reject the null hypothesis as p-value is less than .05, and the t-test for slope *Coef / SE Coef* is far from zero(3.713...). Given this, we can reject the null hypothesis that the true slpoe is zero. **Response in Complete Sentences for a General Audience:** Based on the output, there's a relationship between the country's GDP and its energy usage. As the *GDP* increases or decreases, the *energy* will increases or decreases, which means *GDP* significantly affects *energy*.