STATS 201/8 Assignment 1

Anish Hota ahot228

Due Date: 3pm, Tuesday 30th July 2024

1 Question 1

We wish to investigate the relationship between electricity consumption and the gross domestic product (GDP) for countries of the world. GDP is an indicator of a country's economic performance adjusted for purchasing power parities to account for between-country differences in price levels. Information was obtained for a selection of 26 of the most populous countries in the world.

The data is stored in the file *electricity.csv* and contains the variables:

Variable	Description
Electricity	electricity consumption (in billions of kilowatt-hours),
GDP	gross domestic product (GDP) in billions of dollars (US),
Country	name of the country.

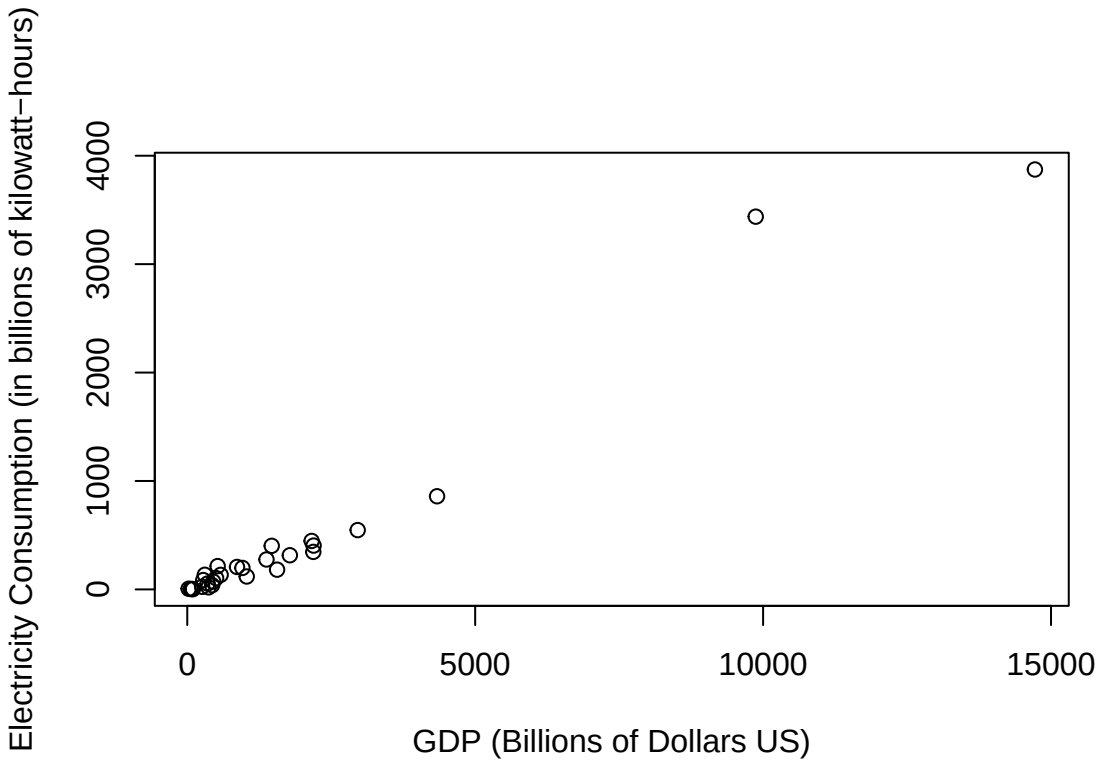
- Make sure you change your name and UPI/ID number at the top of the assignment.
- Comment on the two scatter plots of the data.
- A simple linear regression model has been fitted and output provided (for both the original data and the reduced data set as shown in the second plot). Stick with the simple linear regression model. (DO NOT CHANGE THE FITTED MODEL.)
- Create a scatter plot with the fitted line from the fitted model superimposed over it.
- Complete the equation of the fitted model as part of the Method and Assumption Checks.
- Complete the "% of the variability" statement as part of the Method and Assumption Checks.
- Complete the Executive Summary by adding two more sentences. One sentence should be interpreting the relevant strength of evidence in context and the other should be estimating, in context, the effect of a 100 billion dollar increase in GDP on electricity consumption.

1.1 Question of interest/goal of the study

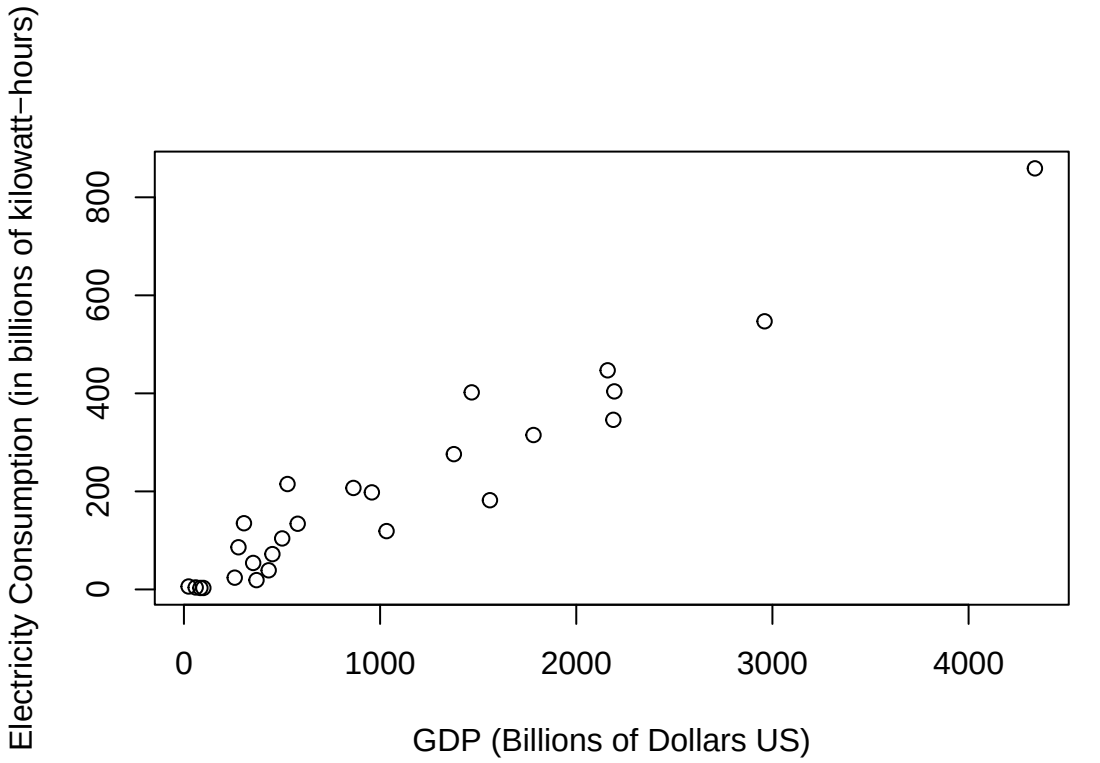
We are interested in using a country's gross domestic product to predict the amount of electricity that they use.

1.2 Read in and inspect the data:

```
elec.df<-read.csv("electricity.csv")
plot(Electricity~GDP, data=elec.df,xlab = "GDP (Billions of Dollars US)", ylab = "Electricity Consumptio
```




```
plot(Electricity~GDP, data=elec.df[elec.df$GDP<6000,],xlab = "GDP (Billions of Dollars US)", ylab = "El
```

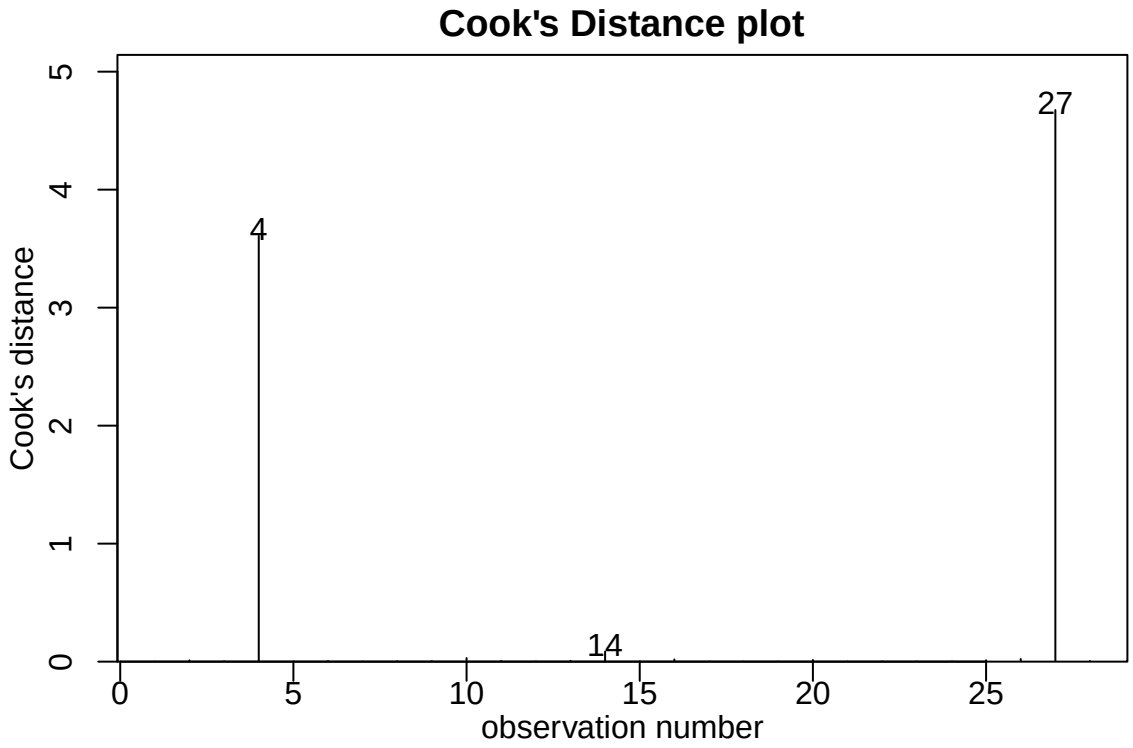


1.3 Comment on the plots

The first plot is linear but we have two outliers with China and USA being much larger than the rest so we can't see the other countries cluttered at the bottom. The second plots have removed these countries to get a better view of the spread of the other countries. With this plot we can see that in general there is a linear relationship and that as GDP increases the electrciity consu,ption increases as well.

1.4 Fit an appropriate linear model, including model checks and relevant output.

```
elecfit1.lm=lm(Electricity~GDP,data=elec.df)
cooks20x(elecfit1.lm)
```

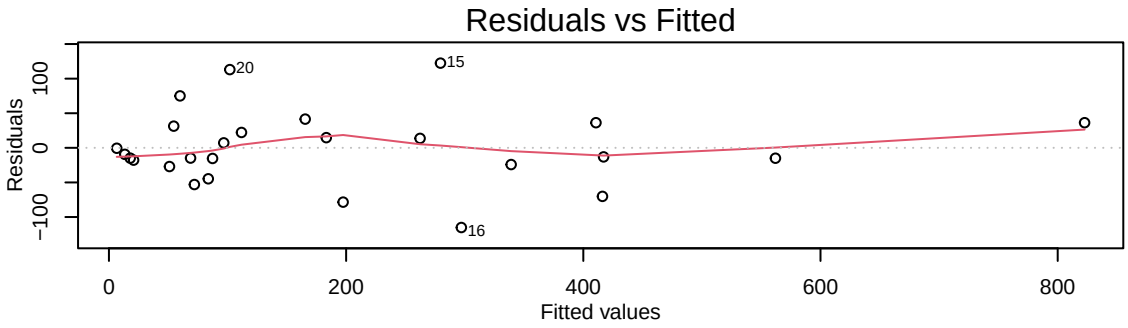


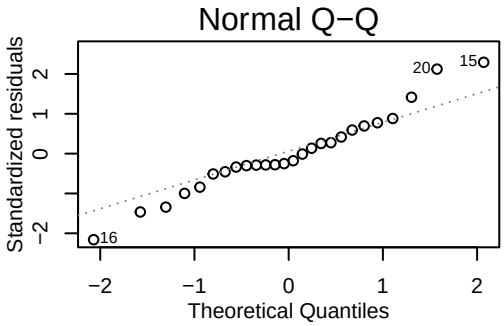

```
elec.df[elec.df$GDP>6000,]
```

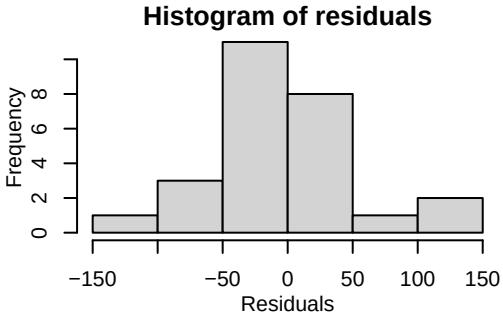
```
##         Country Electricity   GDP
## 4         China        3438  9872
## 27 UnitedStates        3873 14720
```

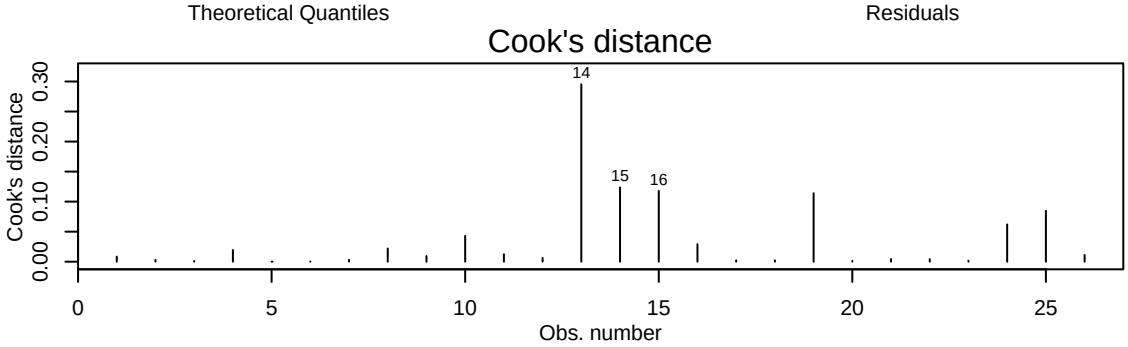
```
elecfit2.lm=lm(Electricity~GDP,data=elec.df[elec.df$GDP<6000,])

modelcheck(elecfit2.lm)
```










```
summary(elecfit2.lm)
```

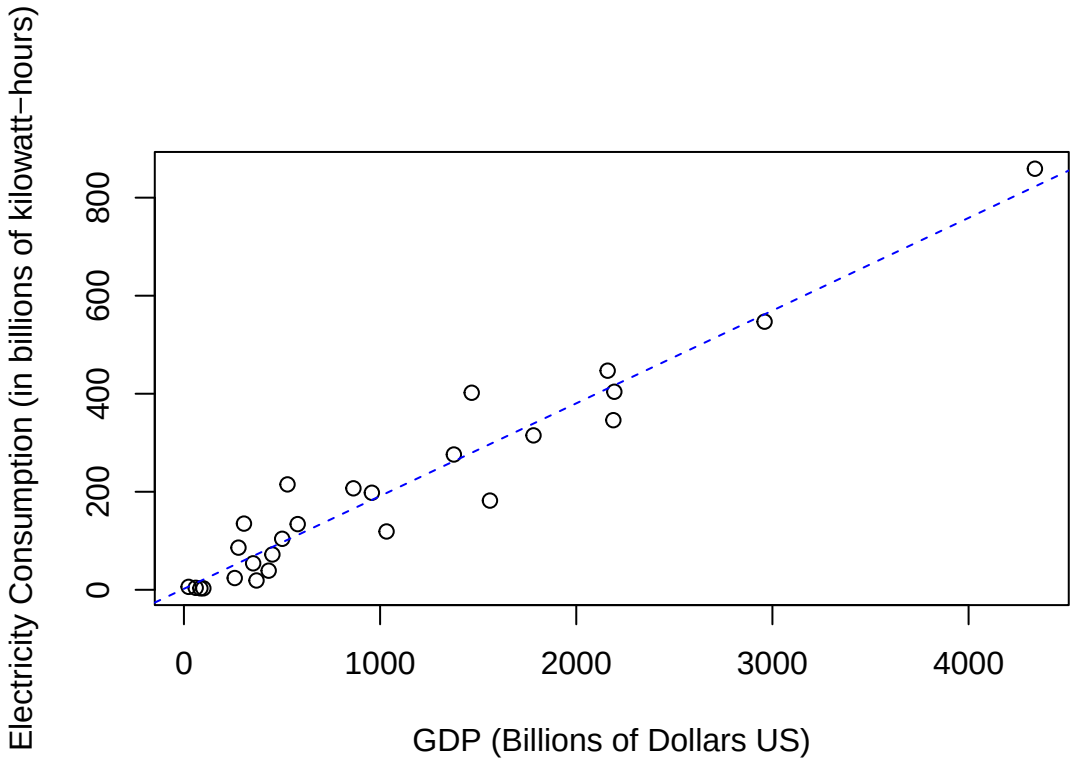
```
##
## Call:
## lm(formula = Electricity ~ GDP, data = elec.df[elec.df$GDP <
##     6000, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -115.16  -22.56  -11.25   29.08  122.43
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.05155   15.28109   0.134    0.894
## GDP          0.18917    0.01041  18.170 1.56e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.64 on 24 degrees of freedom
## Multiple R-squared:  0.9322, Adjusted R-squared:  0.9294
## F-statistic: 330.2 on 1 and 24 DF,  p-value: 1.561e-15
```

```
confint(elecfit2.lm)
```

```
##                   2.5 %     97.5 %
## (Intercept) -29.4870645 33.5901674
## GDP           0.1676863  0.2106611
```

1.5 Create a scatter plot with the fitted line from your model superimposed over it.

```
plot(Electricity~GDP, data=elec.df[elec.df$GDP<6000,],xlab = "GDP (Billions of Dollars US)", ylab = "El
abline(elecfit2.lm, lty = 2, col = "blue")
```



1.6 Method and Assumption Checks

Since we have a linear relationship between GDP and electricity consumption, we have fitted a simple linear regression model to our data. We have 28 of the most populous countries, but have no information on how these were obtained. As the method of sampling is not detailed, there could be doubts about independence. These are likely to be minor, with a bigger concern being how representative the data is of a wider group of countries. The initial residuals and Cooks plot showed two distinct outliers (USA and China) who had vastly higher GDP than all other countries and therefore could be following a totally different pattern so we limited our analysis to countries with GDP under 6000 (billion dollars). After this, the residuals show patternless scatter with fairly constant variability - so no problems. The normality checks don't show any major problems (slightly long tails, if anything) and the Cook's plot doesn't reveal any further unduly influential points. Overall, all the model assumptions are satisfied.

1.6.1 Complete the equation below:

Our model is:

$y = \beta_0 + \beta_0 x + \epsilon$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$ $ELectricity = 2.05 + 0.19 * GDP + \epsilon$

1.6.2 Complete the statement

Our fitted model explains 93.22% of the variability in the data.

1.7 Executive Summary

It was of interest to see if there is a relationship between electricity consumption and gross domestic product (GDP) for countries.

We restricted our analysis to countries with GDP less than 6,000 billion dollars.

As seen by the data, our p-value is very small. This means that we have very strong evidence in the fact that an increase in GDP does increase Electricity consumption in a given country.

Since our GDP is very a 100 billion dollar difference won't make much difference to the Electricity Consumption but there is still an increase.