STATS 201 Assignment 1

LiYuzhuo-2019210146

Due Date: 10/10/2021

## Loading required package: s20x

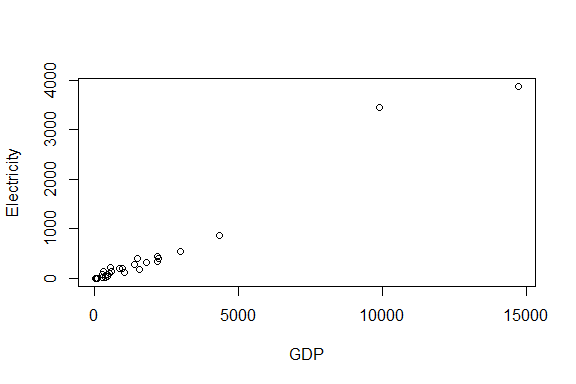
# Question 1

# 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.

# Read in and inspect the data:

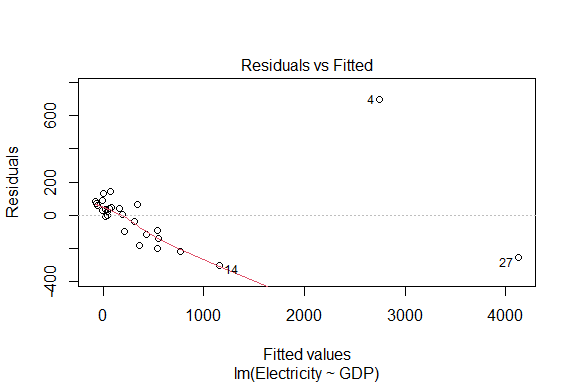
elec.df<-read.csv("electricity.csv")  
plot(Electricity~GDP, data=elec.df)



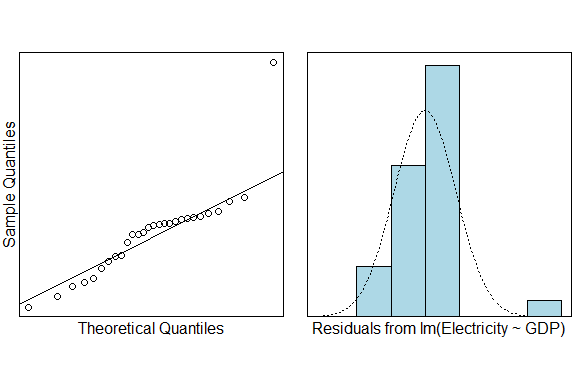
It seems to be a positive linear relationship between the GDP of a country and amount of electricity that they use. There are two observations with its responses are much greater than the rest of the data. These points probably have a big influence on the overall fit.

# Fit an appropriate linear model, including model checks.

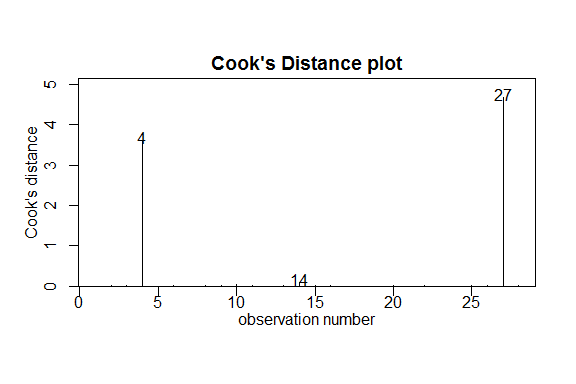
elecfit1=lm(Electricity~GDP,data=elec.df)  
plot(elecfit1,which=1)



normcheck(elecfit1)



cooks20x(elecfit1)



# Identify the two countries with GDP greater than 6000.

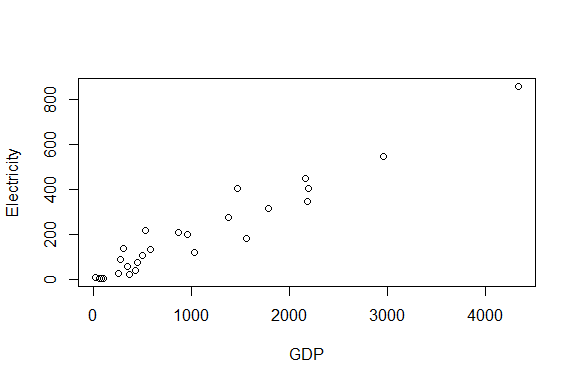
# could use some R code to do this  
high <- subset(elec.df, GDP > 6000)  
high

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

The two countries with GDP Greater than 6000 are China and the United States.

# Replot data eliminating countries with GDP greater than 6000.

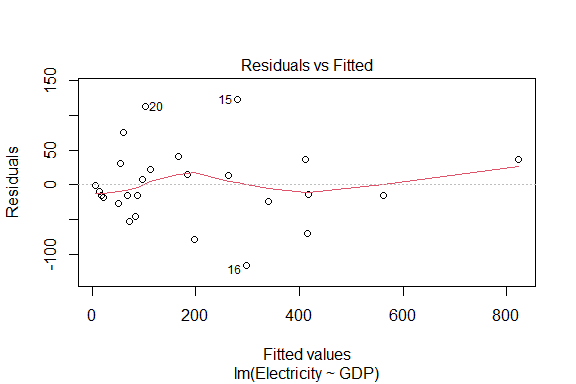
# Hint: If you want to limit the range of the data, do so in the data statement. E.G. something similar to data=elec.df[elec.df$GDP>2000,]  
data <- elec.df[elec.df$GDP<6000,]  
plot(Electricity~GDP, data=data)



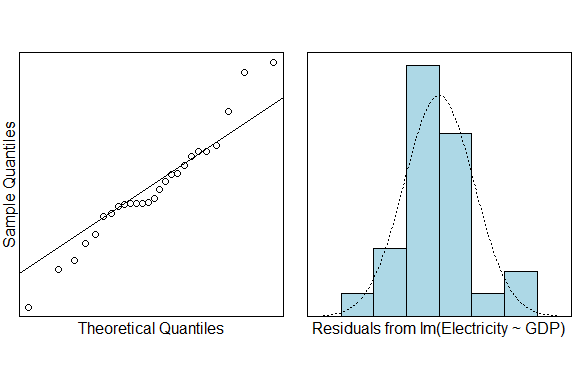
There still appears to be a positive linear relationship between the GDP of a country and amount of electricity that they use. In this plot, our fitted becomes better, and obviously the strength of the linear relation has gotten stronger.

# Fit a more appropriate linear model, including model checks.

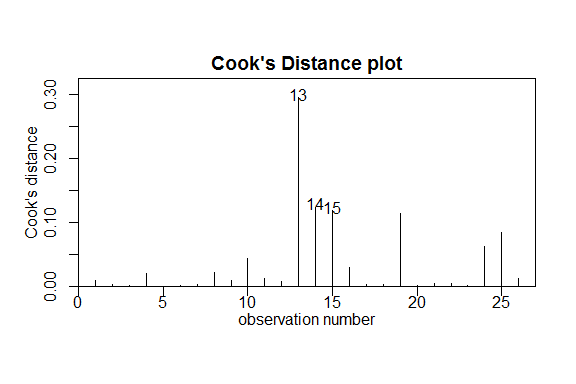
elecfit2 <- lm(Electricity~GDP,data=data)  
plot(elecfit2,which=1)



normcheck(elecfit2)



cooks20x(elecfit2)



summary(elecfit2)

##   
## Call:  
## lm(formula = Electricity ~ GDP, data = data)  
##   
## 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)

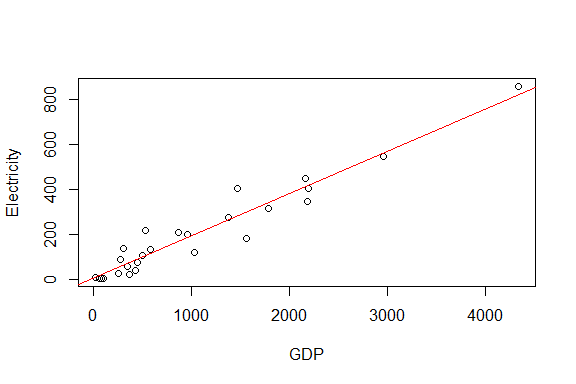
## 2.5 % 97.5 %  
## (Intercept) -29.4870645 33.5901674  
## GDP 0.1676863 0.2106611

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

plot(Electricity~GDP, data=data)  
coef(elecfit2)

## (Intercept) GDP   
## 2.0515514 0.1891737

abline(coef(elecfit2),col='red')



# Method and Assumption Checks

Since we have a linear relationship in the data, 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 normaility 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.

Our model is: where

Our model explains 93% of the total variation in the response variable, and so will be reasonable for prediction.

# Executive Summary

The purpose of this analysis was to investigate the relationship between electricity consumption and the gross domestic product (GDP) for countries of the world. Information was obtained for a selection of 28 of the most populous countries in the world. Of these 28 countries, China and the United States are excluded because they are regarded as outliers.

We found from our analysis that there exists a positive linear relationship between the GDP of a country and amount of electricity that they use, as .

The relationship has Residual standard error 54.64 on 24 degrees of freedom. And with 95% confidence we estimate that on average, for every 1 billion dollar increase in GDP there is between 0.168 to 0.211kWh increase in electricity consumption. As Adjusted R-squared is 0.9294, our fitted model is very accurate.

# Predict the electricity usage for a country with GDP 1000 billion dollars.

predict(elecfit2, newdata = data.frame('GDP'=1000), interval = 'prediction')

## fit lwr upr  
## 1 191.2253 76.29873 306.1518

# Interpret the prediction and comment on how useful it is.

The prediction says that a country with GDP of 1000 billion dollars will consume between 76.3 to 306.2kWh of electricity, with a mean consumption 191.2kWh.

Our model has explained nearly 93% of the variability in electricity consumption, so it is really good. The prediction interval accounts for about 29% of the original data range of 800kwh - which is quite good. However, from the perspective of the prediction range, our prediction of power consumption in countries with GDP far below 1000 billion maybe be wrong, because the lower limit of the prediction range will become negative. But in a word, for the most of the countries, our model is quite useful as a prediction.

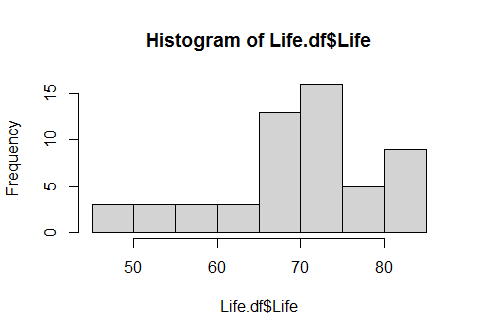
# Question 2

# Question of interest/goal of the study

We are interested in estimating the mean life expectancy of people in the world and seeing if the data is consistant with a mean value of 68 years.

## Read in and inspect the data:

Life.df=read.csv("countries.csv",header=T)  
hist(Life.df$Life)



summary(Life.df$Life)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 48.10 65.14 72.90 69.79 75.34 83.21

From the histogram of life expectancy, it seems that the average life expectancy of each country meets the normal distribution and is centered on the average age about 70 years. It also shows that life expectancy in a few countries is much lower than the average life expectancy. Result of the summary shows that the median and average life expectancy are also very similar, 72.9 and 69.8 years respectively.

## Manually calculate the t-statistic and the corresponding 95% confidence interval.

Formula: and 95% confidence interval

NOTES: The R code mean(y) calculates , sd(y) calculates , the standard deviation of , and the degrees of freedom, . The standard error, and qt(0.975,df) gives the multiplier.

sample.mean <- mean(Life.df$Life)  
sample.sd <- sd(Life.df$Life)  
sample.n <- length(Life.df$Life)  
t.mult <- qt(1-0.05/2, df = sample.n-1)  
CI.life <- sample.mean + t.mult \* c(-1, 1) \* sample.sd/sqrt(sample.n)  
sample.se <- sample.sd/sqrt(sample.n)  
t.stat <- (sample.mean - 68)/sample.se  
p.value <- 2\*(1-pt(abs(t.stat), df = sample.n - 1))  
c('sample mean'=sample.mean,'sample sd'=sample.sd,'sample size'=sample.n)

## sample mean sample sd sample size   
## 69.787018 9.250388 55.000000

c('standard error' = sample.se,'t value'=t.stat,'p value'=p.value)

## standard error t value p value   
## 1.2473220 1.4326839 0.1577122

## Using the t.test function

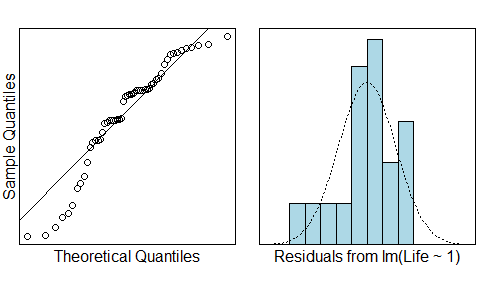
t.test(Life.df$Life, mu=68)

##   
## One Sample t-test  
##   
## data: Life.df$Life  
## t = 1.4327, df = 54, p-value = 0.1577  
## alternative hypothesis: true mean is not equal to 68  
## 95 percent confidence interval:  
## 67.28629 72.28775  
## sample estimates:  
## mean of x   
## 69.78702

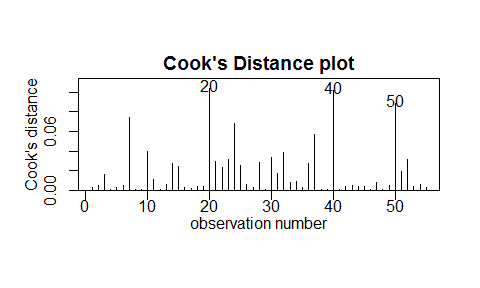
**Note:** You should get exactly the same results from the manual calculations and using the function. Doing this was to give you practice using some R code.

## Fit a null model

lifefit1=lm(Life~1,data=Life.df)  
normcheck(lifefit1)



cooks20x(lifefit1)



summary(lifefit1);

##   
## Call:  
## lm(formula = Life ~ 1, data = Life.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.688 -4.648 3.117 5.558 13.425   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69.787 1.247 55.95 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.25 on 54 degrees of freedom

confint(lifefit1)

## 2.5 % 97.5 %  
## (Intercept) 67.28629 72.28775

# Why are the P-values from the t-test output and null linear model different?

The null model compares the estimated average life expectancy (69.8 years) with the average life expectancy of 0 years. It is reasonable that the p value is basically 0, because if the real life expectancy is actually 0 years, it is impossible to get our estimated 69.8 years.

The t-test compares the estimated average life expectancy (69.8 years) with the assumed average life expectancy (68 years). The p value is much larger, because even if the average life expectancy we assume is indeed the real population average, we can actually get our estimate due to the random error in the sample.

# Method and Assumption Checks

As the data consists of one measurement - the life expectancy for each country - we have applied a one sample t-test to it, equivalent to an intercept only linear model (null model).

We have a random sample of 55 countries so we can assume they form an independent and representative sample. We wished to estimate their average life expectancy and compare it to 68 years. Checking the normality of the differences reveals the data is moderately left skewed. However, we have a large sample size of 55 and can appeal to the Central Limit Theorem for the distribution of the sample mean, so are not concerned. There were no unduly influential points.

Our model is: where

# Executive Summary

The purpose of this analysis was to see whether the average life expectancy among countries of the world is about 68 years. A sample of 55 countries was randomly selected.

We didn’t find evidence against the null hypothesis through the analysis . So we agree that the mean life expectancy of countries in the world is 68 years old. Our 95% confidence interval for this given sample ranges from 67.3 to 72.3 years. Therefore, we can conclude that 68 years old may be the real average life expectancy of all countries in the world.

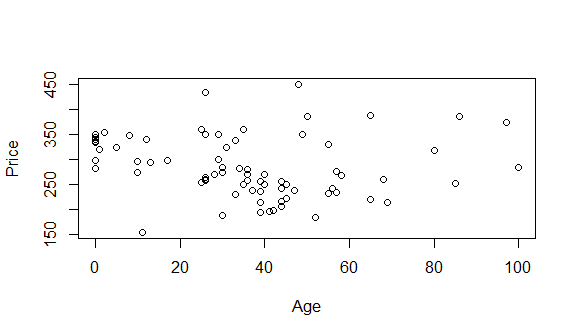
# Question 3

# Question of interest/goal of the study

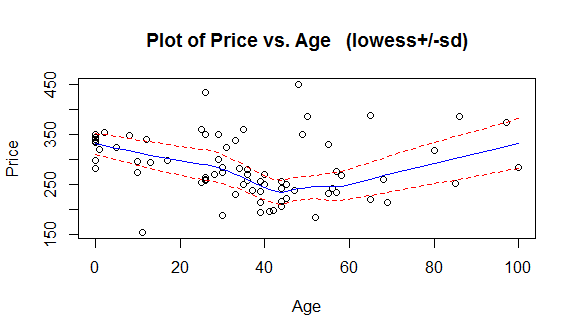
We are interested in investigate the relationship between the age of the house and the sale price of a house in the city Eugene in the state of Oregon, USA.

## Read in and inspect the data:

home.df=read.csv("homes.csv",header=T)  
plot(Price~Age,data=home.df)



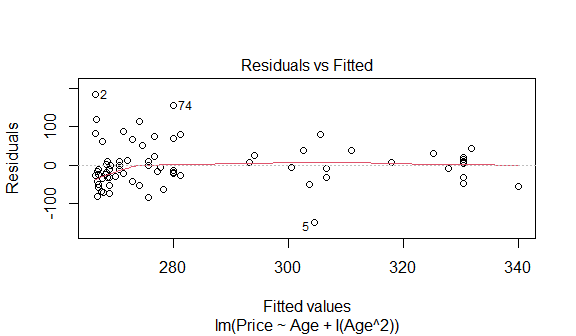
trendscatter(Price~Age,data=home.df)



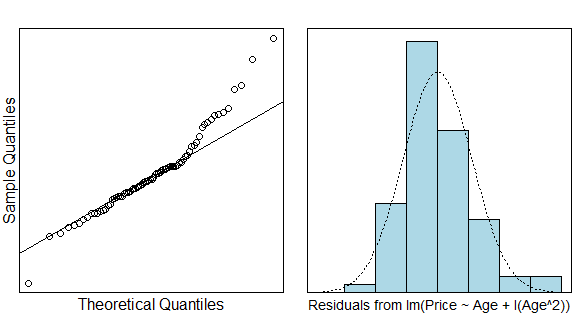
The data does not follow a simple linear trend, it seems like a quadratic functional relationship instead, so a quadratic model can be used to fit this data.

## Fit an appropriate linear model, including model checks.

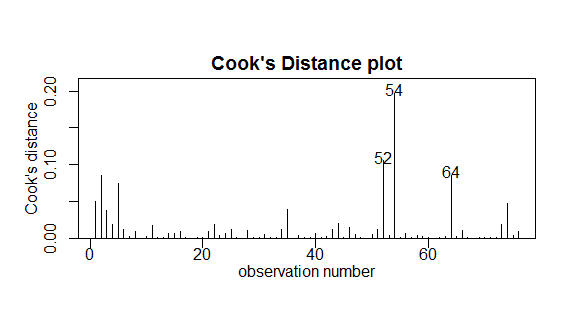
homefit <- lm(Price~Age + I(Age^2), data=home.df)  
plot(homefit,which=1)



normcheck(homefit)



cooks20x(homefit)



summary(homefit)

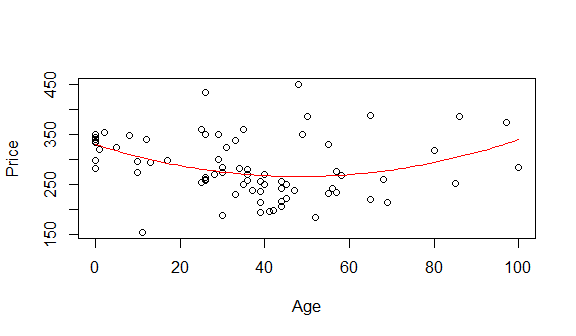
##   
## Call:  
## lm(formula = Price ~ Age + I(Age^2), data = home.df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -149.058 -31.868 -7.788 20.141 183.576   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 330.410440 15.103397 21.877 < 2e-16 \*\*\*  
## Age -2.652629 0.748807 -3.542 0.000695 \*\*\*  
## I(Age^2) 0.027491 0.008472 3.245 0.001773 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 56.49 on 73 degrees of freedom  
## Multiple R-squared: 0.1468, Adjusted R-squared: 0.1235   
## F-statistic: 6.282 on 2 and 73 DF, p-value: 0.003039

confint(homefit)

## 2.5 % 97.5 %  
## (Intercept) 300.30941199 360.51146738  
## Age -4.14499992 -1.16025848  
## I(Age^2) 0.01060739 0.04437476

## Plot the data with your appropriate model superimposed over it.

preds <- predict(homefit, data.frame("Age"=0:100), interval = "prediction")  
plot(Price ~ Age,data=home.df)  
lines(0:100, preds[,'fit'], col="red")



# Method and Assumption Checks

A quadratic linear model was fitted as the data seemed to follow a curved trend. The 76 single-family houses were chosen randomly, so we regard each data is independent. The initial trendscatter plot for the simple linear model showed fairly constant and patternless scatter, but the dataset followed a curved trend so a quadratic linear model was fitted. The residuals show patternless scatter with fairly constant variability - so no problems, except for a slight negative deviation on the left end of the graph. Although this deviation is a cause for some concern it is fairly moderate and the rest of the mean residuals are near constant around 0, so we have chosen to tolerate it. The normaility checks don’t show any major problems and the Cook’s plot doesn’t reveal any further unduly influential points. Overall, all the model assumptions are satisfied.

Our model is: where .

As Adjusted R-squared is 0.1235, so our model only explains 12.4% total variation in the response variable, it has low predictive power.

# Executive Summary

The purpose of this analysis was to investigate the relationship between the age of the house and the sale price of a house. We have found there to be a quadratic relationship between the age of a house and its sale price (). As Adjusted R-squared is 0.1235, our fitted model is not very accurate.