STATS 201 Assignment 1

My name : WangYingpai My ID number : 2019210179

Due Date: —-

```
## Loading required package: s20x
```

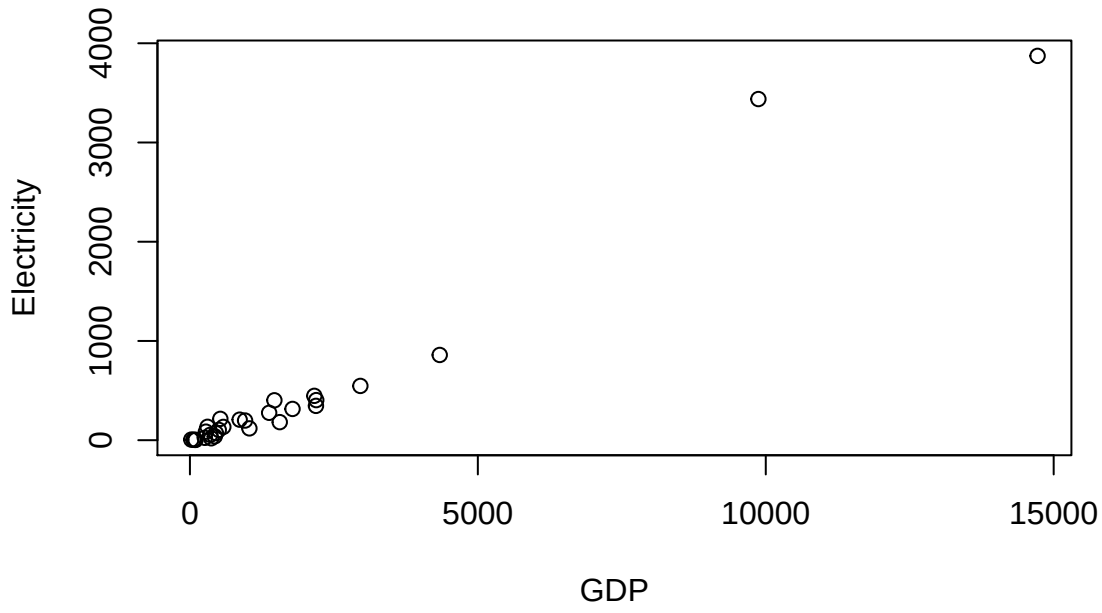
Question 1

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)
```

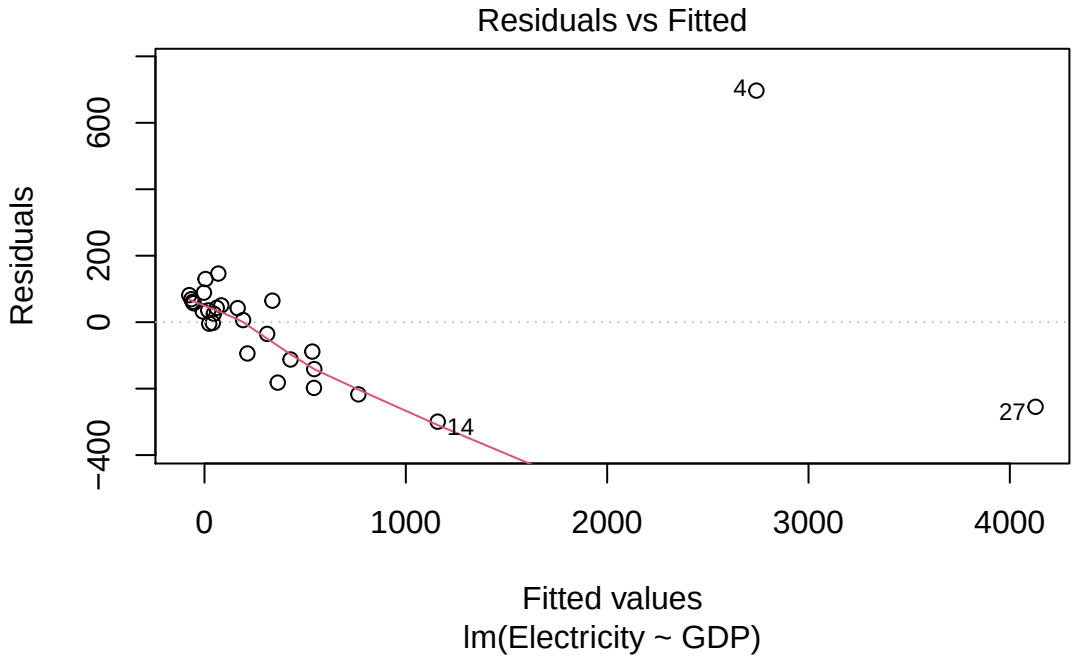


WRITE COMMENT HERE

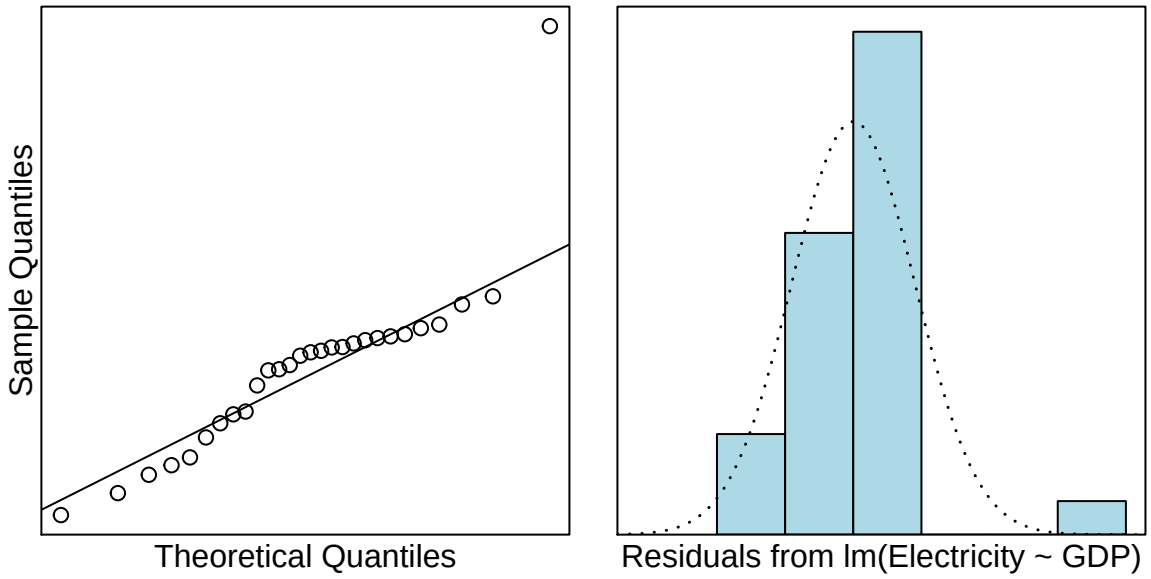
Through this initial plot of the data.It is clear that there is some relationship between GDP and electricity.It's almost linear.We can use a linear model to fit them.

Fit an appropriate linear model, including model checks.

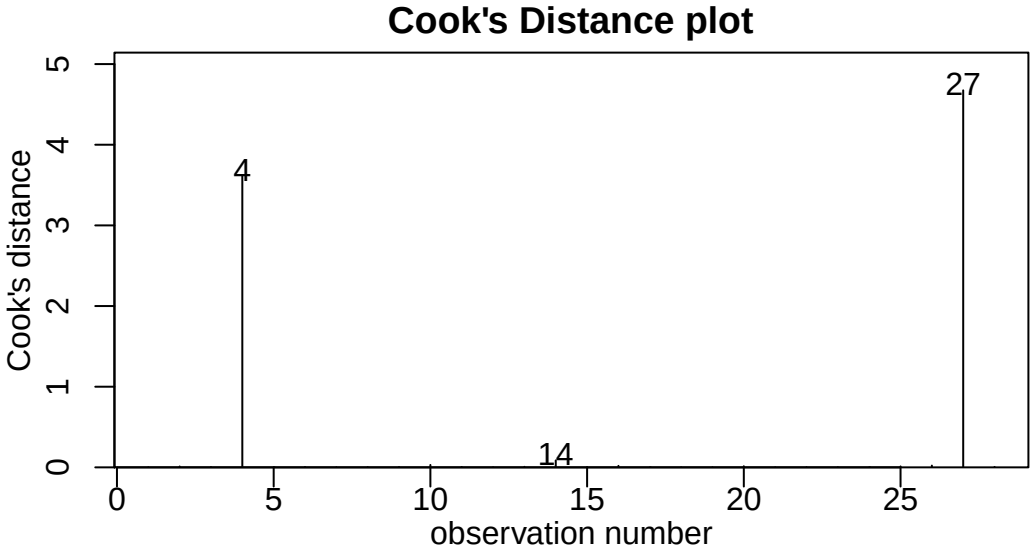
```
elecfit1=lm(Electricity~GDP,data=elec.df)
plot(elecfit1,which=1)
```




```
normcheck(elecfit1)
```




```
cooks20x(elecfit1)
```



Identify the two countries with GDP greater than 6000.

```
# could use some R code to do this
data1=elec.df[elec.df$GDP>6000,]
data1
```

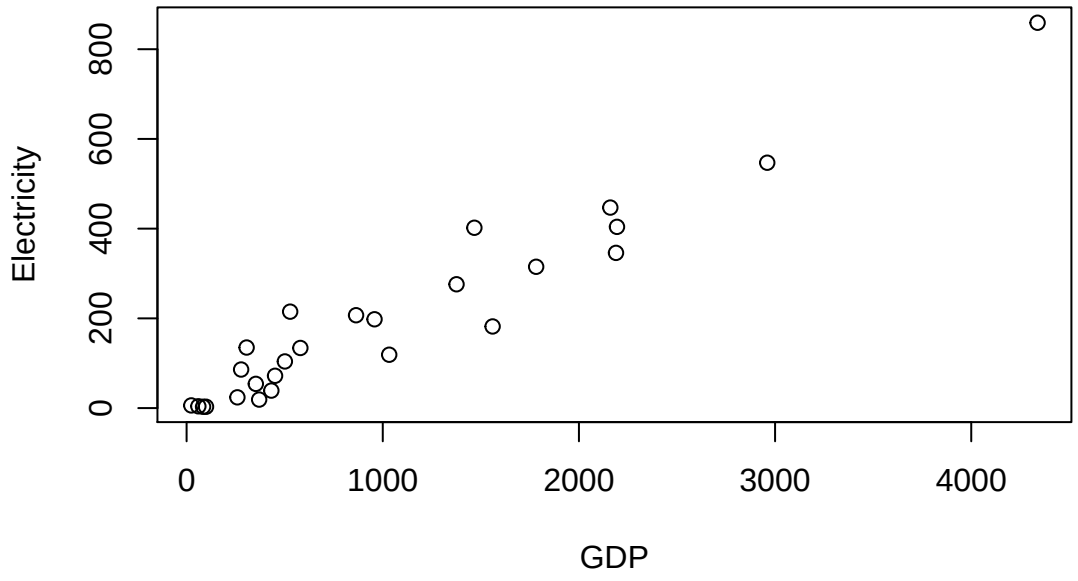
```
##         Country Electricity   GDP
## 4         China        3438  9872
## 27 UnitedStates        3873 14720
```

WRITE COMMENT HERE

We have fitted an initial linear model.Through model checking, we found that the residuals vs fitted plot seems not to confirm reasonably constant scatter. There were issues with the two countries with the highest GDP.They are China and United States,They use far more electricity and GDP than other countries. The residuals are not reasonable and cook's distance are very large.It states these tow countries make a big difference on our model.Thus we need to refit the linear regression model without these two countries.

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
data2 = elec.df[elec.df$GDP<=6000,]
plot(Electricity~GDP,data = data2)
```

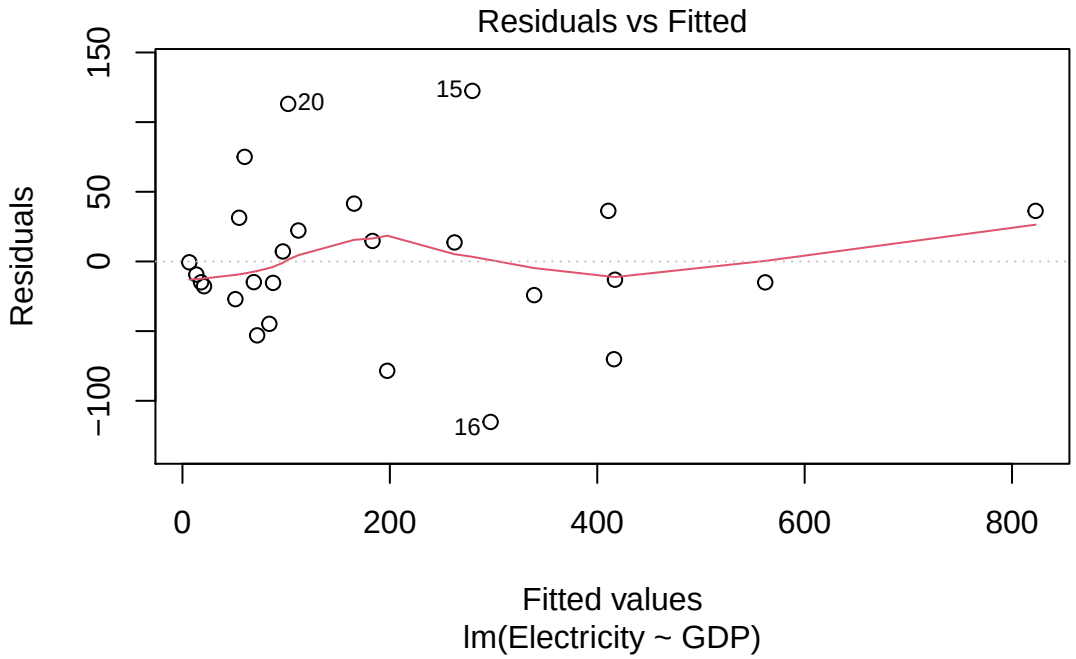


WRITE COMMENT HERE

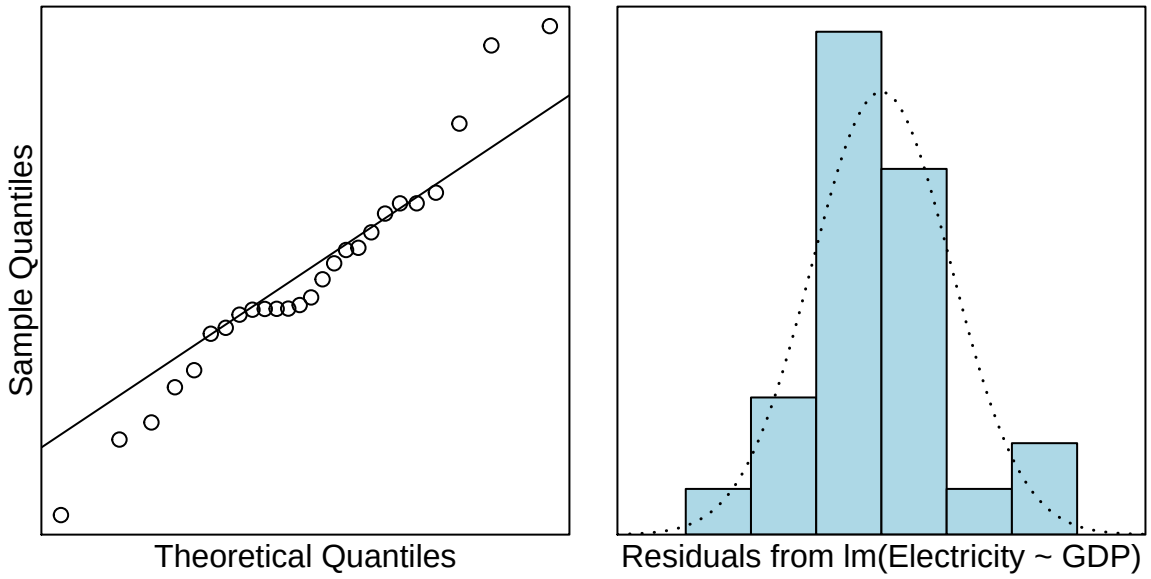
We replot data eliminating countries with GDP greater than 6000.It almost linear.It's more scatter and smooth than our first plot.It's more evenly distributed.We need to refit a more appropriate linear model.And then recheck the model.

Fit a more appropriate linear model, including model checks.

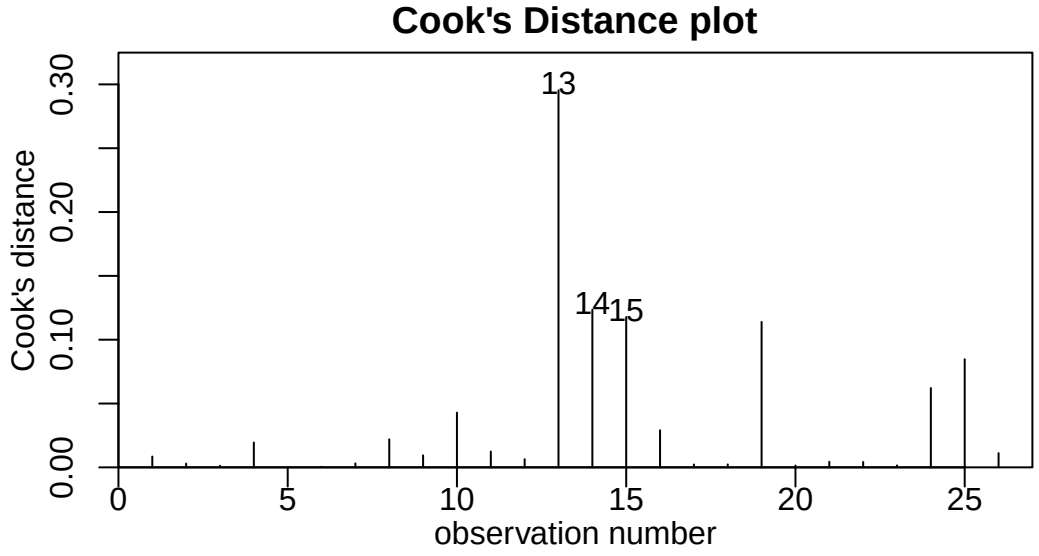
```
elecfit2=lm(Electricity~GDP,data=data2)
plot(elecfit2,which=1)
```




```
normcheck(elecfit2)
```




```
cooks20x(elecfit2)
```




```
summary(elecfit2)
```

```
##
```

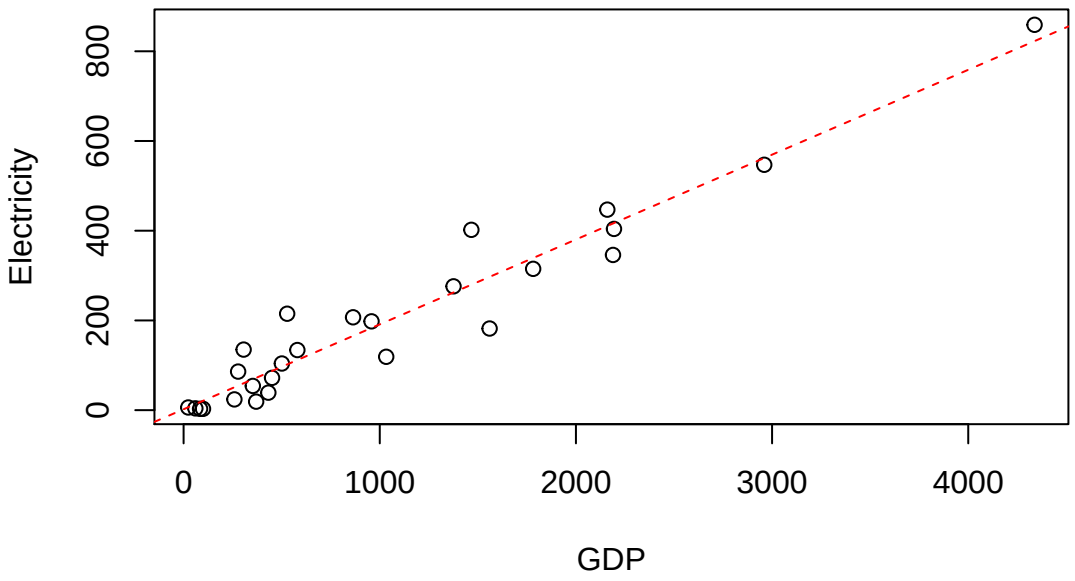
```
## Call:
## lm(formula = Electricity ~ GDP, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -115.16  -22.56  -11.25   29.08  122.43
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.05155   15.28109   0.134    0.894
## GDP          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=data2)
abline(elecfit2,lty=2,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: $Electricity_i = \beta_0 + \beta_1 \times GDP_i + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$

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

Executive Summary

WRITE EXEC SUMMARY HERE

The purpose of the analysis is to investigate the relationship between electricity consumption and GDP.The relationship is linear. We have strong evidence that there exists linear relationship between electricity and GDP because the p-value of the GDP's coefficient is far less than 0.001. We have a positive relationship between electricity and GDP because the estimate value of slope is positive. We get the estimate mean value of electricity varies with GDP is 0.18917.We are 95% confident that the range (0.1676863,0.2106611) contains the true value of each increase of GDP lead to the mean increase of GDP. We can explain approximately 93.22% of the total variation of Electricity by fitting a straight line model using GDP. It's useful for the prediction.

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

```
predict(elecfit2,newdata=data.frame(GDP=1000))
```

```
##        1
## 191.2253
```

```
#predict(examtest.fit,newdata=data.frame(Test=0))
```

Interpret the prediction and comment on how useful it is.

WRITE COMMENTS HERE

Through our linear regression model to predict the electricity usage for a country with GDP 1000 billion dollars(It's less than 6000). We predict it's Electricity is 191.2253 billions of kilowatt-hours. Our model is reasonable and reliable.

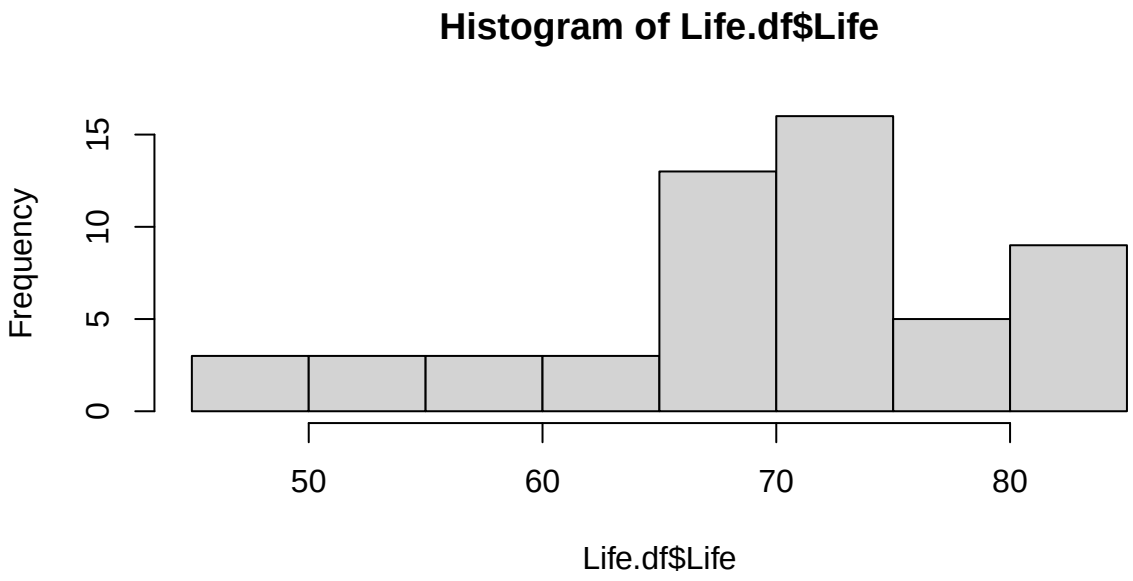
Question 2

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
```

WRITE COMMENT HERE

We plot the histogram of Life and summary them. We know that the mean number is 69.79 and the median is 72.90.

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

Formula: $T = \frac{\bar{y}-\mu_0}{se(\bar{y})}$ and 95% confidence interval $\bar{y} \pm t_{df,0.975} \times se(\bar{y})$

NOTES: The R code `mean(y)` calculates $\bar{y}$, `sd(y)` calculates s, the standard deviation of y, and the degrees of freedom, $df = n - 1$. The standard error, $se(\bar{y}) = \frac{s}{\sqrt{n}}$ and `qt(0.975,df)` gives the $t_{df,0.975}$ multiplier.

```
avg_life = mean(Life.df$Life)
mu = 68
s = sd(Life.df$Life)
n = length(Life.df$Life)
df = n - 1
se_life = s/sqrt(n)
t = qt(0.975,df)

T = (avg_life-mu)/se_life
confidence_interval1 = avg_life - t*se_life
confidence_interval2 = avg_life + t*se_life
```

Using the t.test function

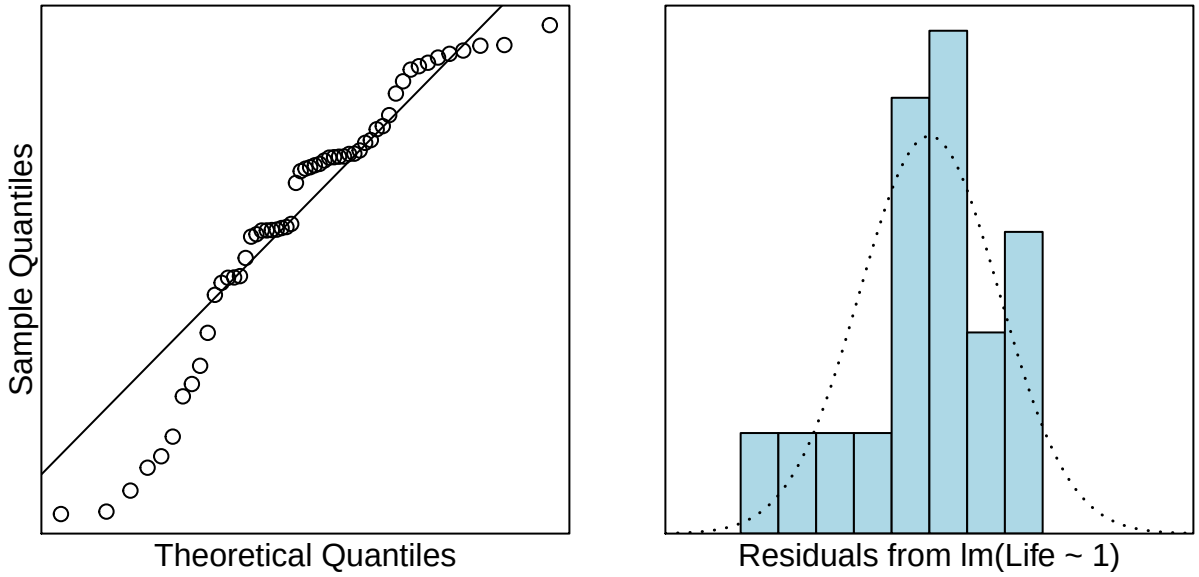
```
t.test(Life.df$Life, mu=68)
```

```
##
##  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
```

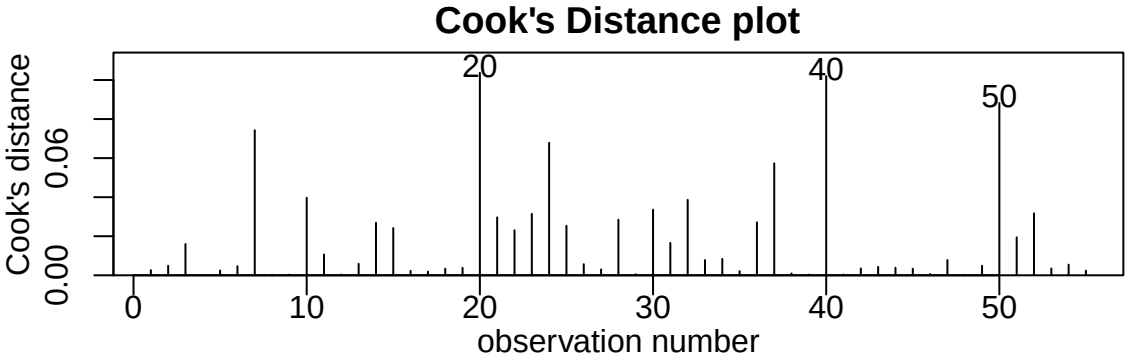
Note: You should get exactly the same results from the manual calculations and using the *t.test* function. Doing this was to give you practice using some R code.

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?

WRITE COMMENT HERE

In the t-test. The null hypothesis is that the true mean is equal to 68. The alternative hypothesis is the true mean is not equal to 68. The p-value of t-test means that if the mean value is 68,the probability that we get our data–Life.

And in the null linear model. The null hypothesis is that value of Life is zero. The alternative hypothesis is that the value of Life is not zero. The p-value of null linear model means that if the value of Life is zero, the probability we get our data–Life. It's different means in different null hypothesis.

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 independant 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: $Life_i = \mu_{Life} + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$

Executive Summary

WRITE EXEC SUMMARY HERE

We want to estimate if the mean Life of people in the world is 68. Firstly we plot the histogram of Life and summary it.We found the mean value of data is 69.79.The p-value is less than 0.001,so we have strong evidence that the mean value of Life is not zero. We are 95% confident the range (67.28629 72.28775) contain the true mean value of Life. It's useful to predict the mean true value of Life. It concludes our estimation. Then we fit a null model and summary it. We get p-value is very small. Our null hypothesis is not reliable. Finally, we get confidence interval,it is the same as we get from t-test.

Question 3

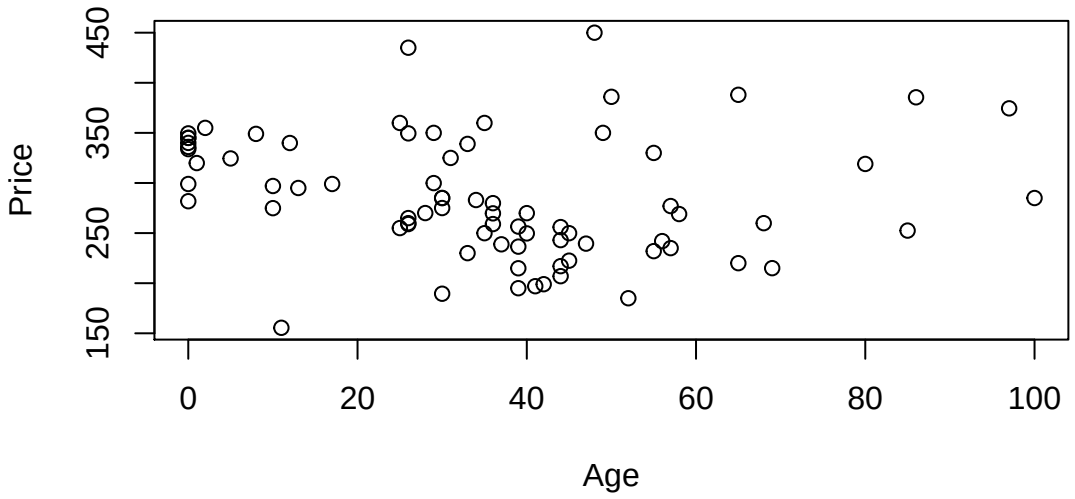
Question of interest/goal of the study

WRITE COMMENT HERE

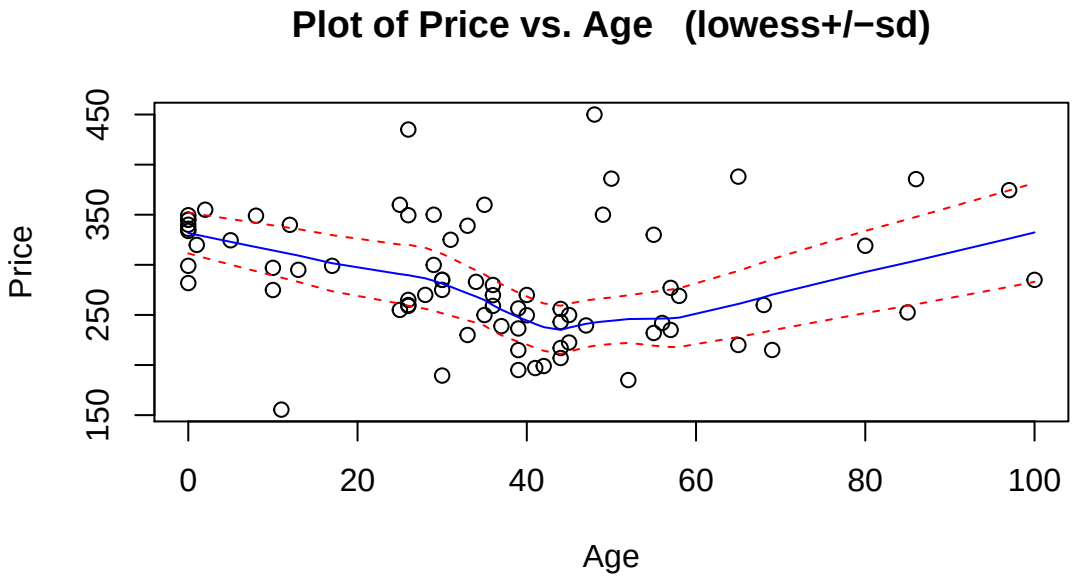
We are interested in the relationship between the sale price of a house and the age of the house.

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)
```

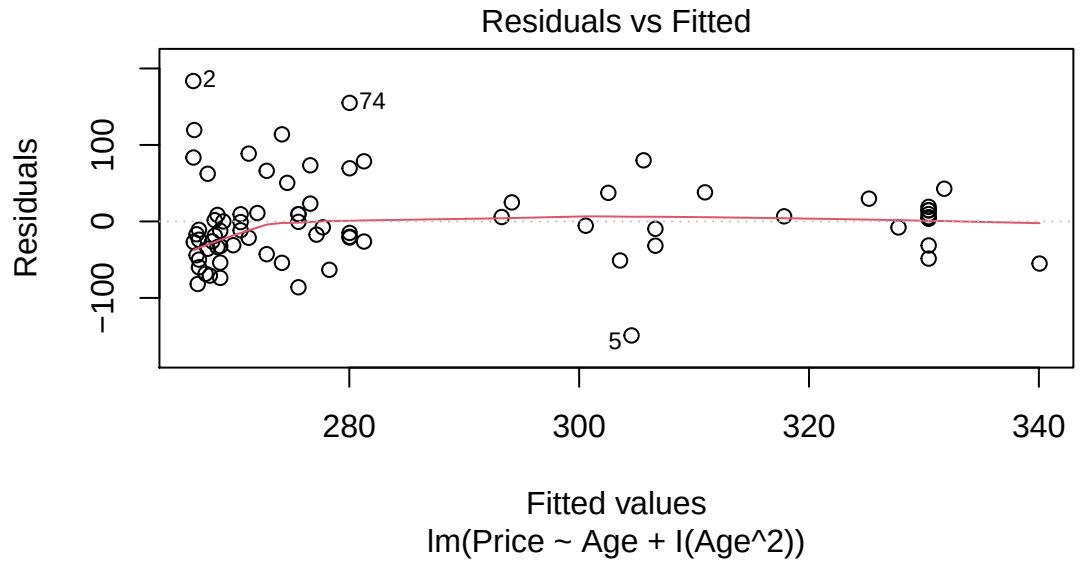


WRITE COMMENT HERE

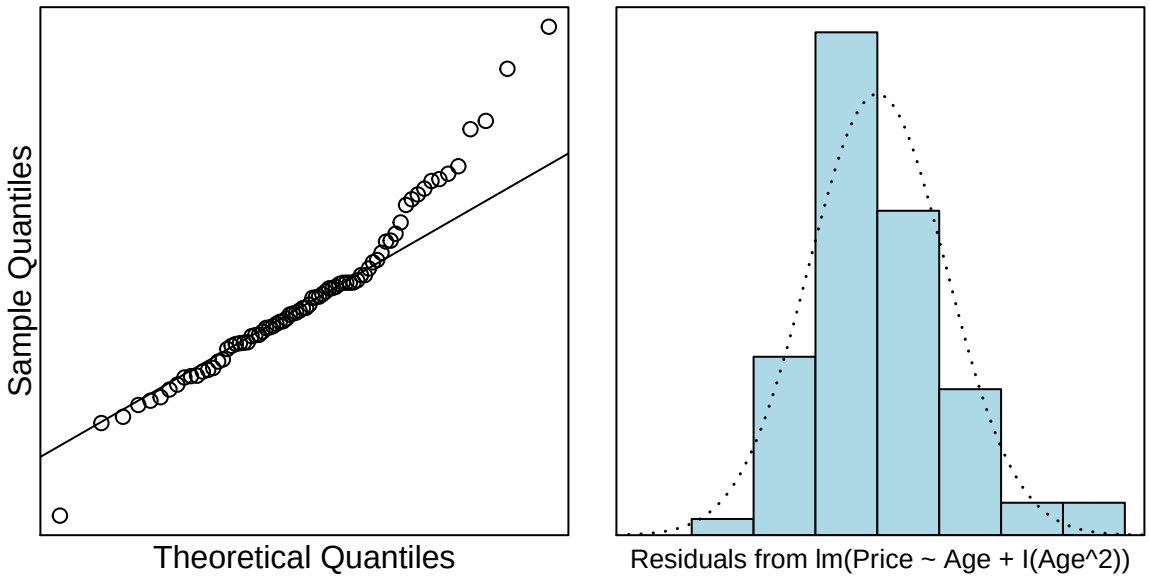
We plot the data and trend scatter it.It's not linear. It's quadratic on the whole.So we fit it with a quadratic model.

Fit an appropriate linear model, including model checks.

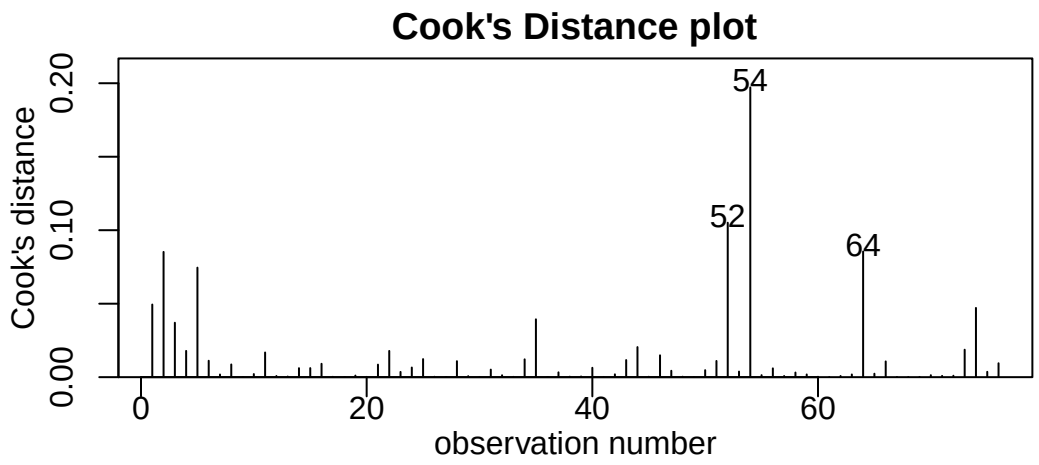
```
homefit =lm(Price~Age+I(Age^2),data=home.df)
plot(homefit,which=1)
```




```
normcheck(homefit)
```

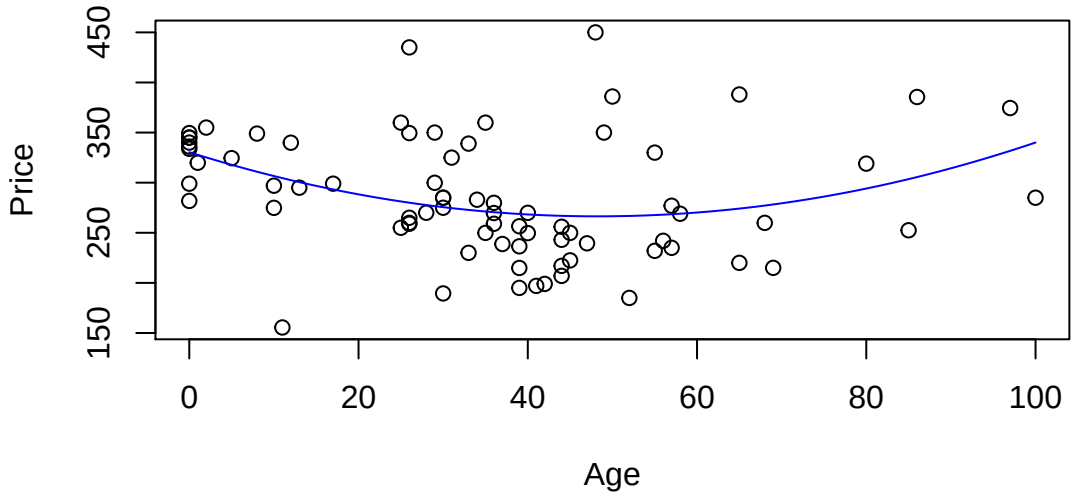



```
cooks20x(homefit)
```



Plot the data with your appropriate model superimposed over it.

```
plot(Price~Age,data=home.df)
x = 0:100
lines(x,predict(homefit,data.frame(Age=x)),col="blue")
```




```
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
```

Method and Assumption Checks

WRITE M & A CHECKS HERE

Since We fitted a quadratic regression model to the data.Through making trend scatter to our data,we found it's quadratic trend on the whole.We have 76 single family homes situation,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 houses.Firstly, we fit a quadratic regression model to our data.Then we go through model check. The residuals show patternless scatter with constant variability,so no problems.The normaility checks don't show any major problems, the cook's plot are normal and no problems.So all the assumptions are statisfied.

Our model is: $Price_i = \beta_0 + \beta_1 \times Age_i + \beta_2 \times Age_i^2 + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$ output: beamer_presentation

Our model explains 14.68% of the total variation in the response variable,it is reasonable for prediction.

Executive Summary

WRITE EXEC SUMMARY HERE

We want to know the relationship between the sale price of a house and its age. We fit a quadratic regression model to our data. Then we get the p-value of quadratic term's coefficient is 0.001773, so we have evidence there exists a quadratic relationship between age and price in our data. We get the estimate mean value of price varies with Age^2 is 0.027491. And We are 95% confident the range (0.01060739 0.04437476) contains each increase of Age^2, the mean value of Price increase. We can explain approximately 14.68% of the variation of Price by fitting a quadratic regression model. So it is reasonable to predict with this model.