STATS 201 Assignment 1

Li Ruoqi 2019220113

Due Date: 2021-10-10

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

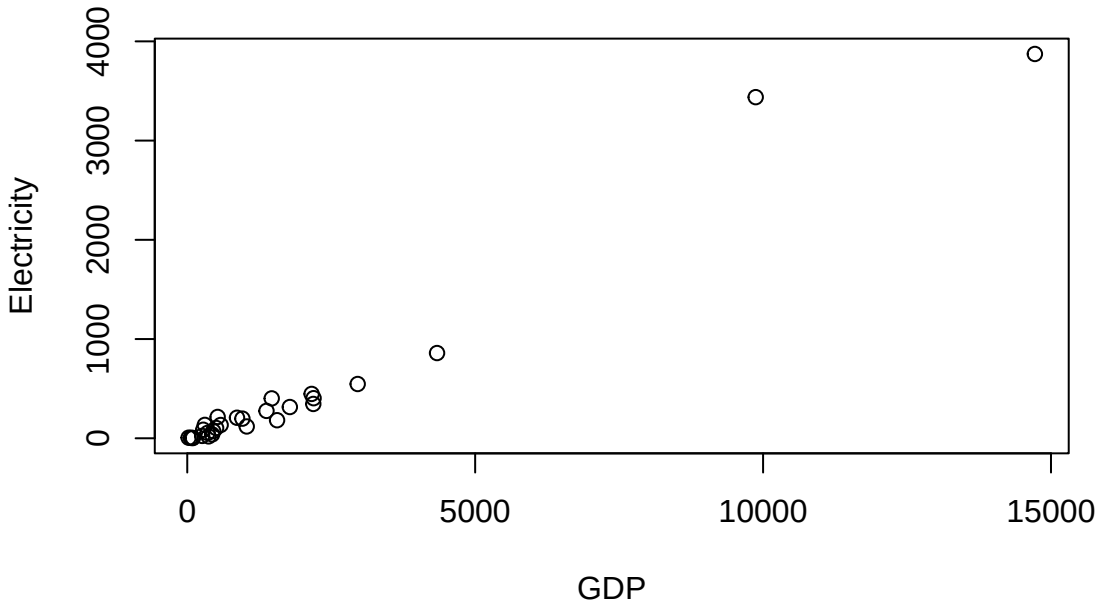
Question 1

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)
```

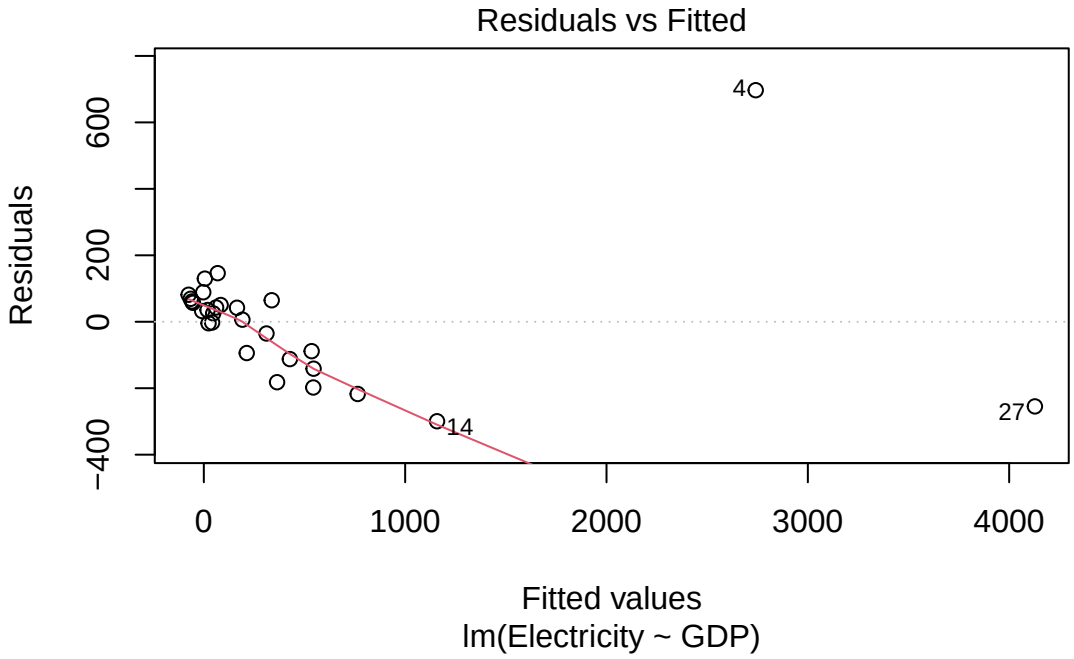


WRITE COMMENT HERE
Basically that is a scatter plot. Electricity on the y-axis and GDP on the x-axis. Graph is pretty clear that Electricity and GDP are related, but we can also see two countries with high GDP in this plot and are clearly

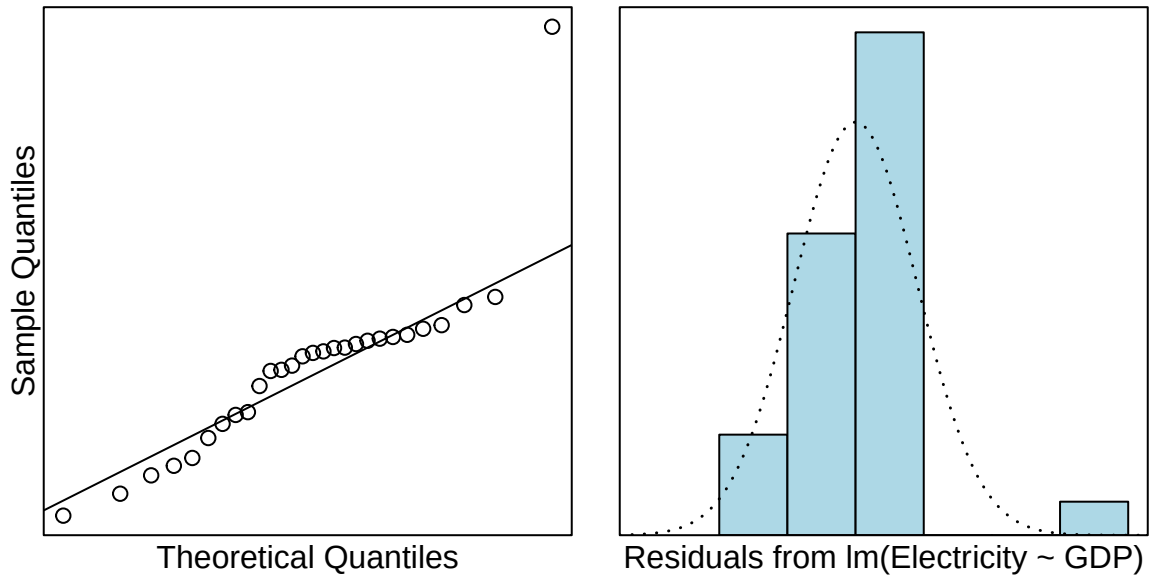
identified as being high influential. It is also clear that the relationship is linear(positive correlation). We probably construct a simple linear model which is fitted in a straight line. The higher the GDP for country gets, the higher electricity the country tend to consume.

Fit an appropriate linear model, including model checks.

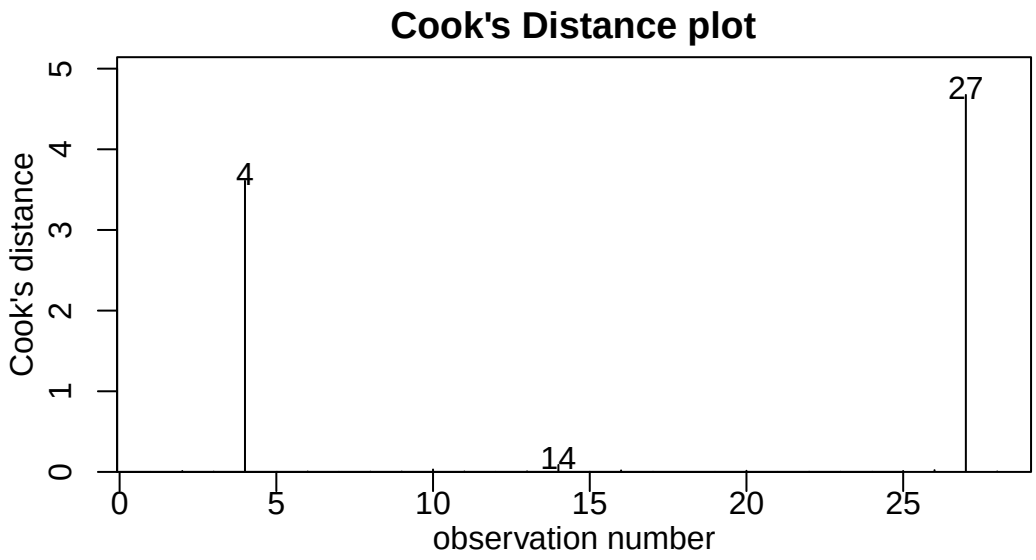
```
elecfit1=lm(Electricity~GDP,data=elec.df)
plot(elecfit1,which=1)
```




```
normcheck(elecfit1)
```



```
cooks20x(elecfit1)
```



Identify the two countries with GDP greater than 6000.

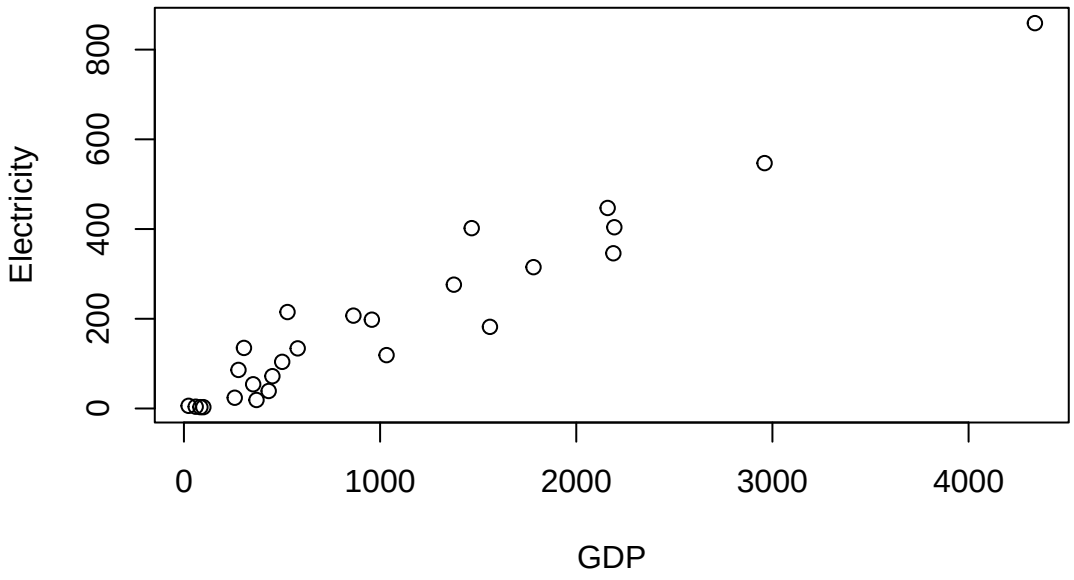
```
# could use some R code to do this
elec.df[elec.df$GDP>6000,]$Country
```

```
## [1] "China"         "UnitedStates"
```

WRITE COMMENT HERE
We have fitted a linear(straight line) model and did model check. There were two non-constant scatters in the residual plot, dramatically causing the residual fail to be around the line 0 roughly. From the Cook's distance plot, the two countries dominates this plot and is clearly identified as being highly influential(their Cook's distance is greater than 0.4). We identified these two countries and will eliminate them to fit a more appropriate linear model.

Replot data eliminating countries with GDP greater than 6000.

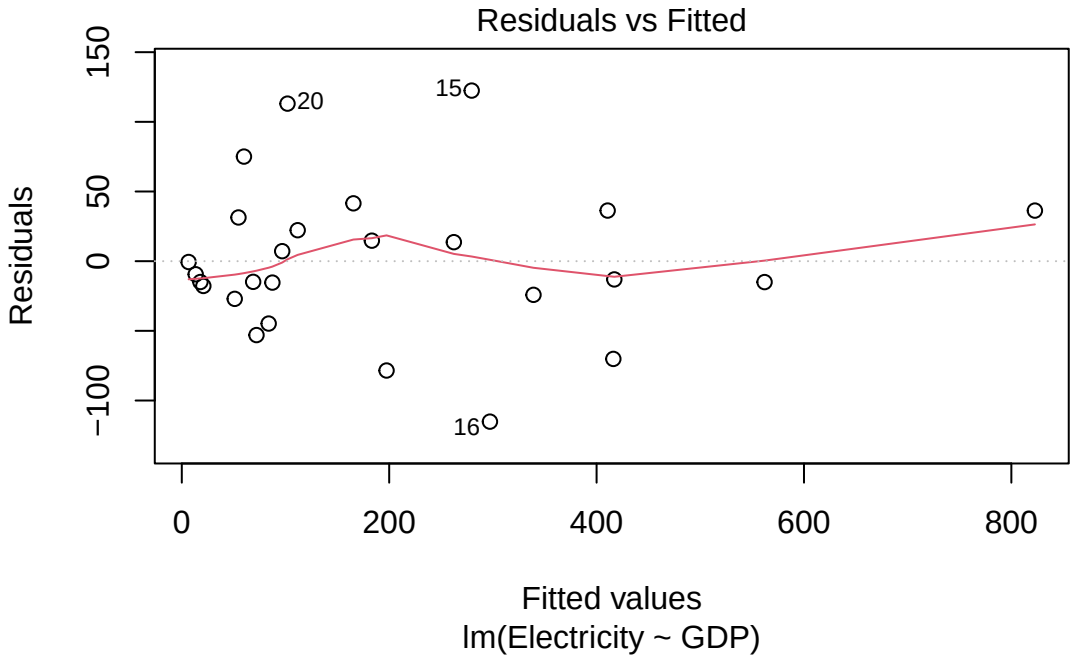
```
# Hint: If you want to limit the range of the data, do so in the data statement. E.G. something similar
newelec.df=elec.df[elec.df$GDP<6000,]
plot(Electricity~GDP, data=newelec.df)
```



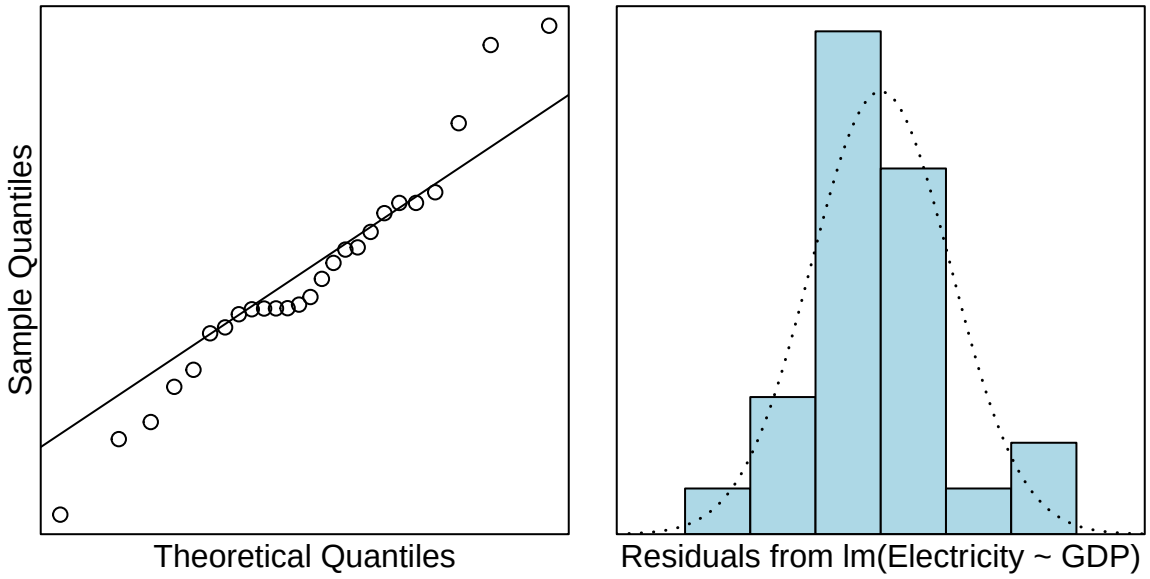
WRITE COMMENT HERE
Looking at this new plot, it is definitely clear that relationship between GDP and electricity is linear(positive correlation). The higher the GDP for country gets, the higher electricity the country tend to consume. We may now conclude that we can (mostly) trust our new simple linear model, and we do not have any unduly influential data points.

Fit a more appropriate linear model, including model checks.

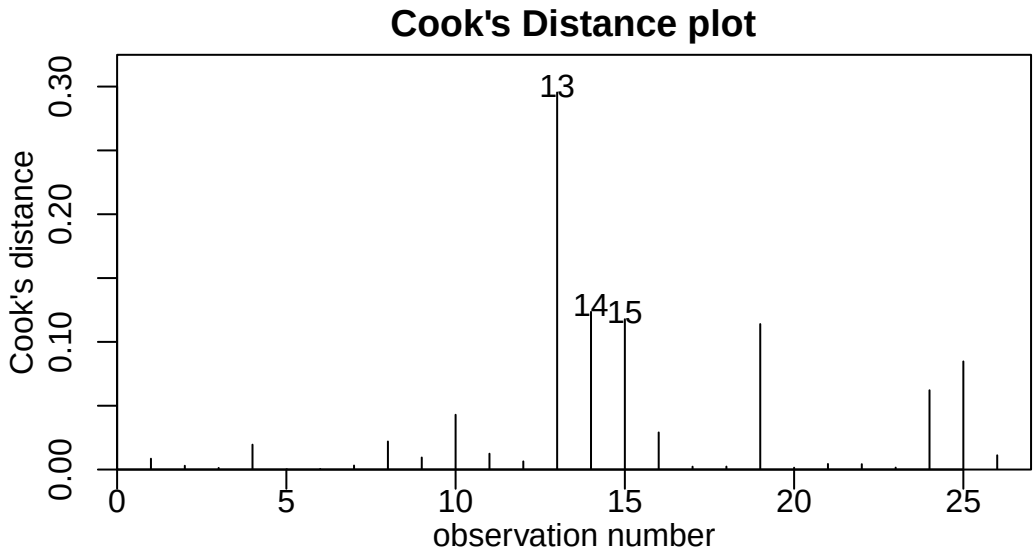
```
elecfit2=lm(Electricity~GDP,data=newelec.df)
plot(elecfit2,which=1)
```



```
normcheck(elecfit2)
```




```
cooks20x(elecfit2)
```




```
summary(elecfit2)
```

```
##
## Call:
## lm(formula = Electricity ~ GDP, data = newelec.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -115.16  -22.56  -11.25   29.08  122.43
##
## Coefficients:
```

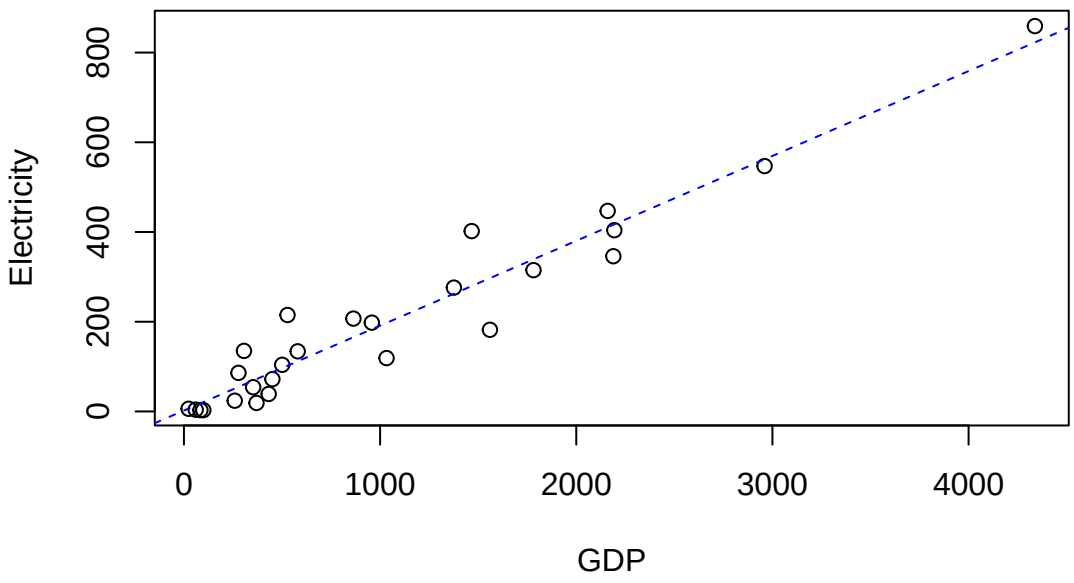
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.05155   15.28109   0.134    0.894
## GDP          0.18917    0.01041  18.170 1.56e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.64 on 24 degrees of freedom
## 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=newelec.df)
abline(elecfit2, lty=2, col= "blue")
```



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 $\epsilon_i \sim iid\ N(0, \sigma^2)$

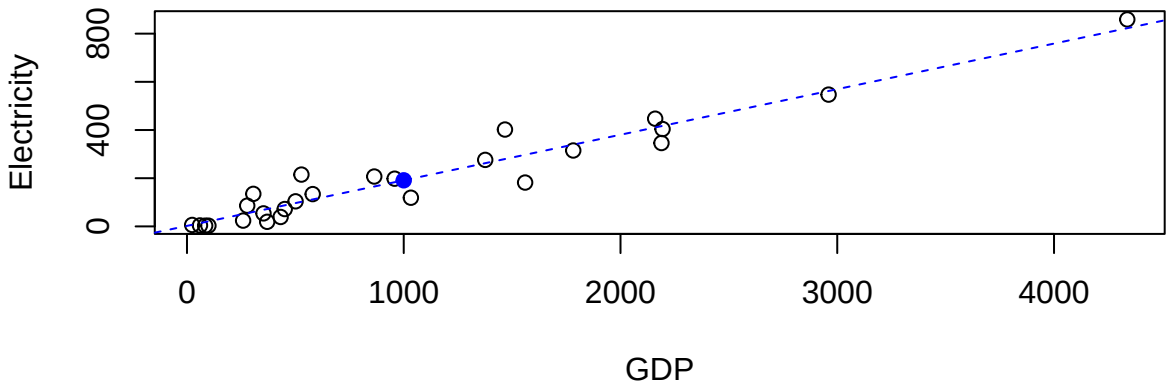
Our model explains 93% of the total variation in the response variable, and so will be reasonable for prediction.

Executive Summary

WRITE EXEC SUMMARY HERE
Our aim is to investigate the relationship between country's gross domestic product(GDP) and the electricity they consume. We have extremely strong evidence that suggest the positive linear relationship between country's gross domestic product(GDP) and the electricity they consume exists. We estimated every unit increases in GDP is associated with an increase value in the mean of electricity consumption in the range of 0.1676863 and 0.2106611. We estimated the average electricity consumption for countries of GDP = 1000 is between 169.086 and 213.3645. We also estimate a typical(not mean) individual country of GDP = 1000, the electricity consumption is between 76.29873 and 306.1518. By comparing the range of original interval and prediction interval, we can conclude our model is useful for prediction.

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

```
plot(Electricity~GDP, data=newelec.df)
abline(elecfit2, lty=2, col= "blue")
elec_pre.df = data.frame(GDP = 1000)
points(1000, predict(elecfit2, elec_pre.df), col = "blue", pch = 19)
```




```
##predict PI
predict(elecfit2, elec_pre.df, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 191.2253 76.29873 306.1518
```

```r
##predict CI
predict(elecfit2, elec_pre.df, interval = "confidence")
```

```
##        fit     lwr      upr
## 1 191.2253 169.086 213.3645
```

Interpret the prediction and comment on how useful it is.

WRITE COMMENTS HERE
From the prediction, we predicted and pointed fitted value for GDP = 1000 is 191.2253. We estimated the average electricity consumption for countries of GDP = 1000 is between 169.086 and 213.3645. We also estimate a typical(not mean) individual country of GDP = 1000, the electricity consumption is between 76.29873 and 306.1518.

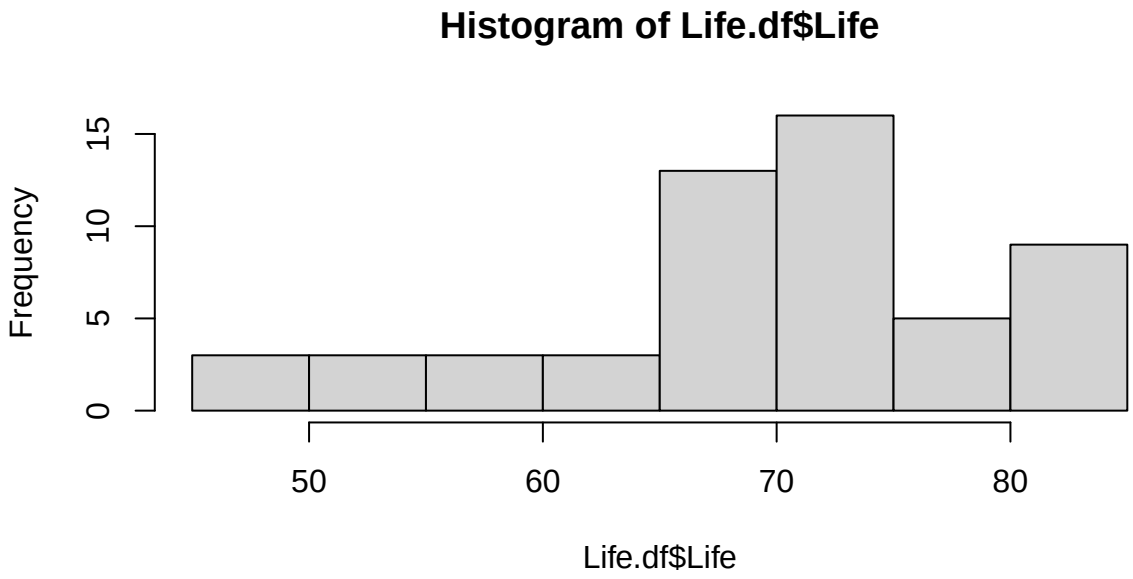
Question 2

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:

```r
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
The histogram provides a convenient graphical representation of the distribution of the overall values. For the total of 55 countries, we obtain some numerical summaries, the mean of life expectancy is 69.79.

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.

```
y = Life.df$Life
n = nrow(Life.df[1])
tmult = qt(1 - 0.05/2, df = n - 1)
##the t-statistic
tstat = (mean(y)-68)/(sd(y)/sqrt(n))
tstat
```

```
## [1] 1.432684
```

```
##the corresponding 95% confidence interval.
##both the lower and upper bounds of the CI
mean(y) + c(-1, 1) * tmult * sd(y)/sqrt(n)
```

```
## [1] 67.28629 72.28775
```

Using the t.test function

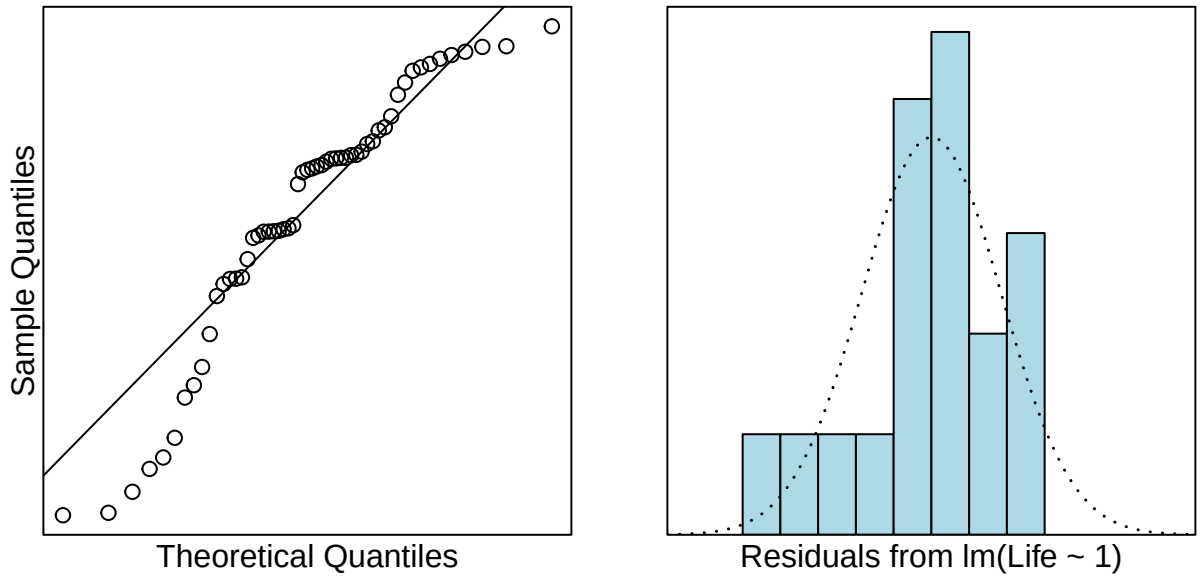
```
t.test(Life.df$Life, mu=68)
```

```
##
##  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
```

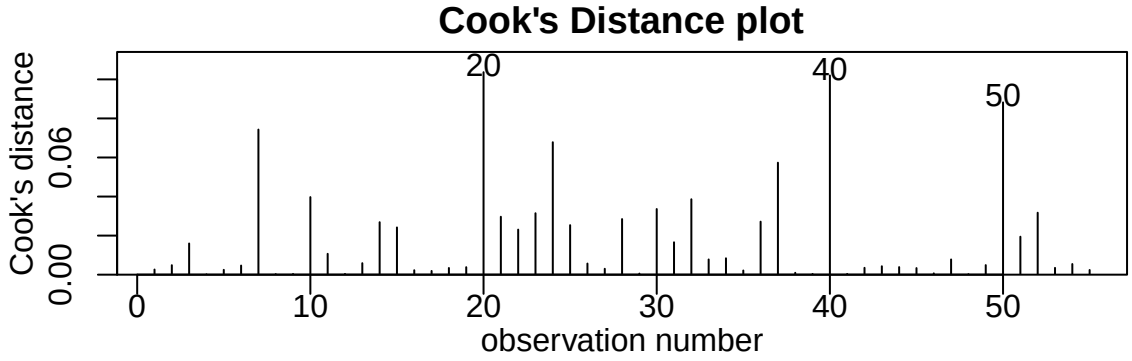
Note: You should get exactly the same results from the manual calculations and using the *t.test* function. Doing this was to give you practice using some R code.

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
Comparing t-test output with null linear model, the p-value from the t-test output is the calculated t-statistic for hypothesis equals to 68, we can not refuse the null hypothesis. In the null linear model, it is assumed that the average life expectancy equals to 0.The fitted linear model of t-test and the null linear model have different possibilities to predict the actual situation, we can conclude that the P-values from the t-test output and null linear model different.

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

Executive Summary

WRITE EXEC SUMMARY HERE
Our aim is to find out whether the mean life expectancy of people in the world is consistant with a mean value of 68 years. We use the t-test and fit a null model. In t-test, we can not refuse the hypothesis the mean life expectancy if people in the world is 68 years. And we also have strong evidence that suggest the mean value is not 0. And we have the 95% confidence that the mean life expectancy is in the range of 67.28629 to 72.28775. We can conclude that the mean life expectancy of people in the world is 68 years.

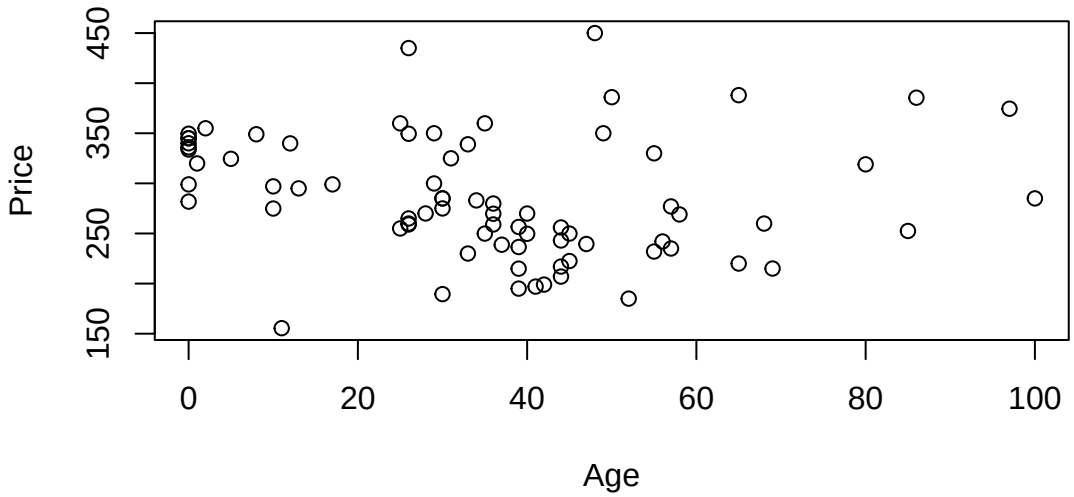
Question 3

Question of interest/goal of the study

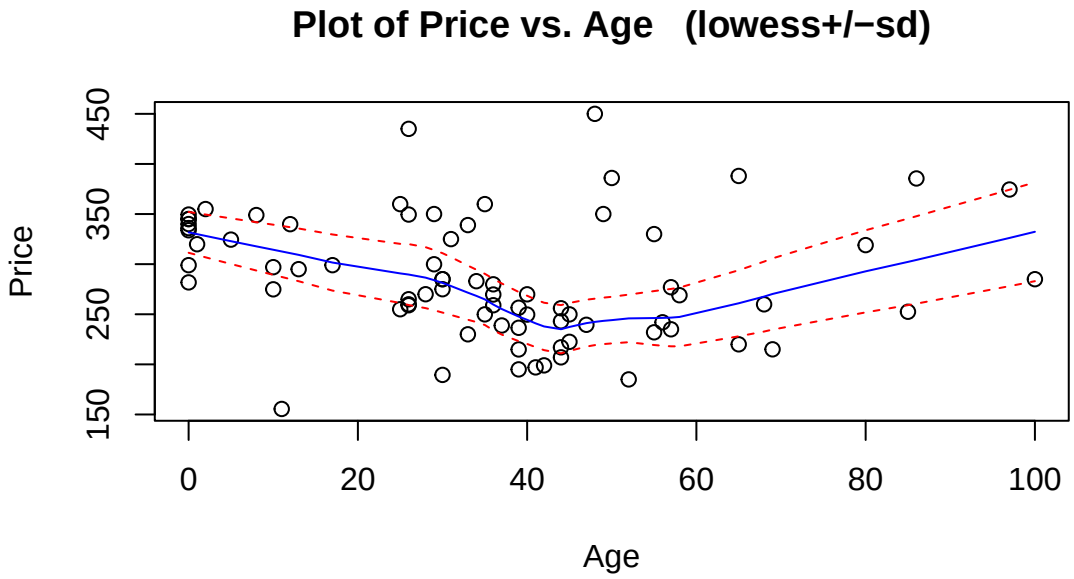
WRITE COMMENT HERE
We are interested in identifying if the age of home is associated with the home price.

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)
```

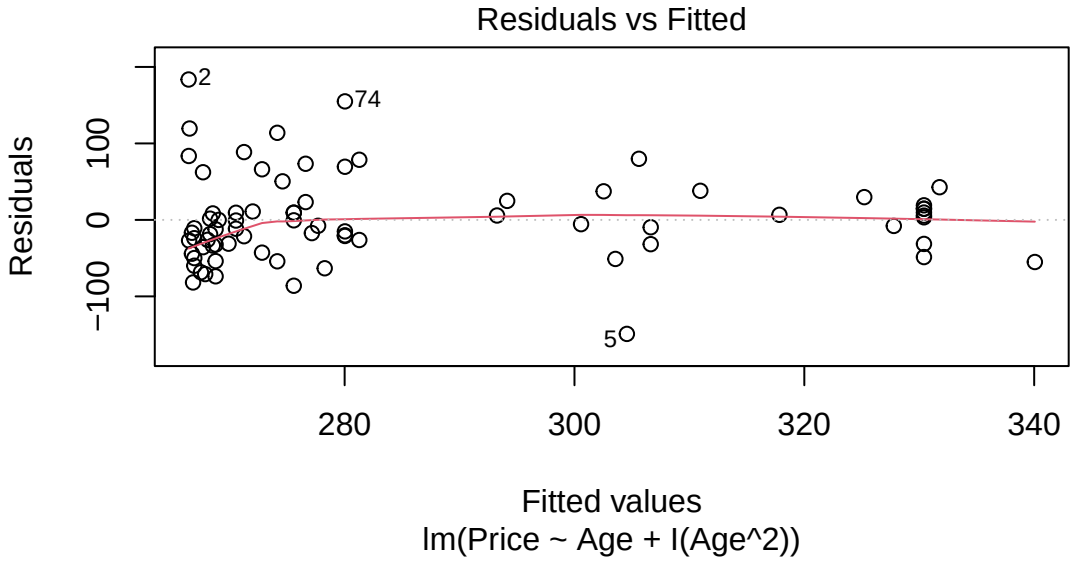


WRITE COMMENT HERE
From the scatter plot of the data, it is not quite a straight line, it could be some curvature with a lot of

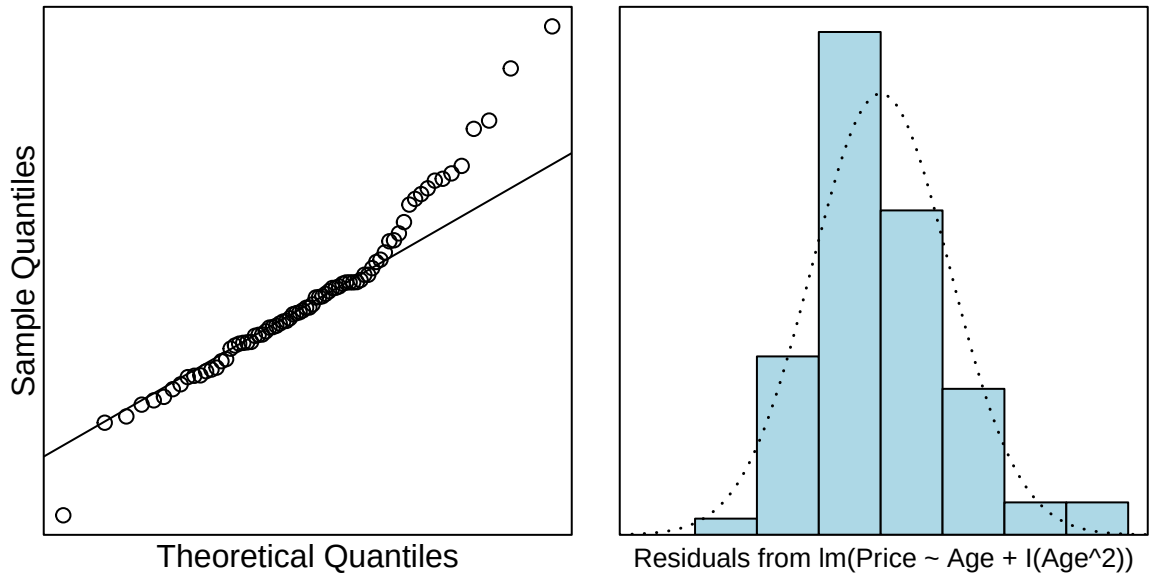
scatters. Scatterplot with trend line paints a clearer picture and confirms it looks like some curvature. We may fit a simple linear model to these data and see the relationship between Price and Age. if it still have a curved relationship, we will add a quadratic term for Age.

Fit an appropriate linear model, including model checks.

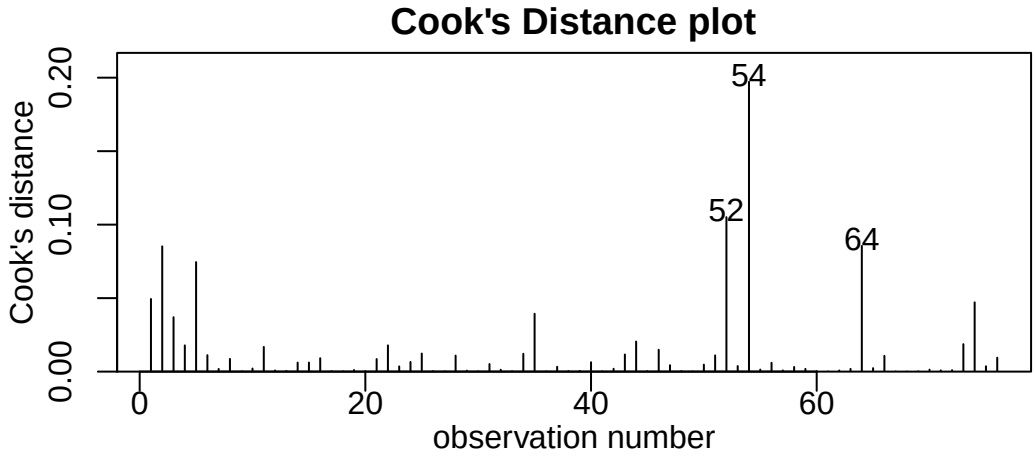
```
##add a quadratic(squared term) for Age.
priceage.fit1 = lm(Price~Age + I(Age^2), data=home.df)
plot(priceage.fit1, which = 1)
```




```
normcheck(priceage.fit1)
```



```
cooks20x(priceage.fit1)
```




```
summary(priceage.fit1)
```

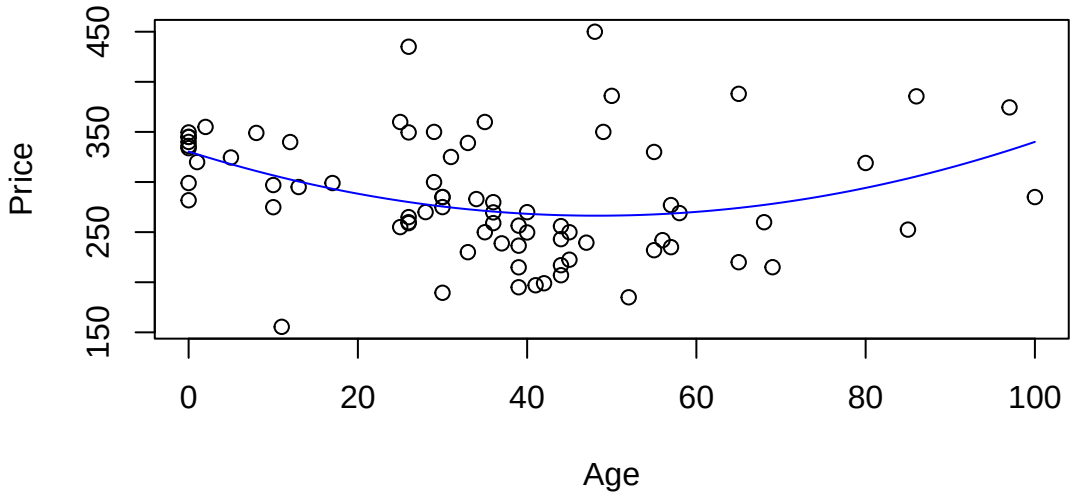
```
##
## Call:
## lm(formula = Price ~ Age + I(Age^2), data = home.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -149.058  -31.868   -7.788   20.141  183.576
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 330.410440  15.103397  21.877  < 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(priceage.fit1)
```

```
##                    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.

```r
plot(Price~Age,data=home.df)
x = 0:100 #Age values at which to predict the price
lines(x, predict(priceage.fit1, data.frame(Age = x)), col = "blue")
```

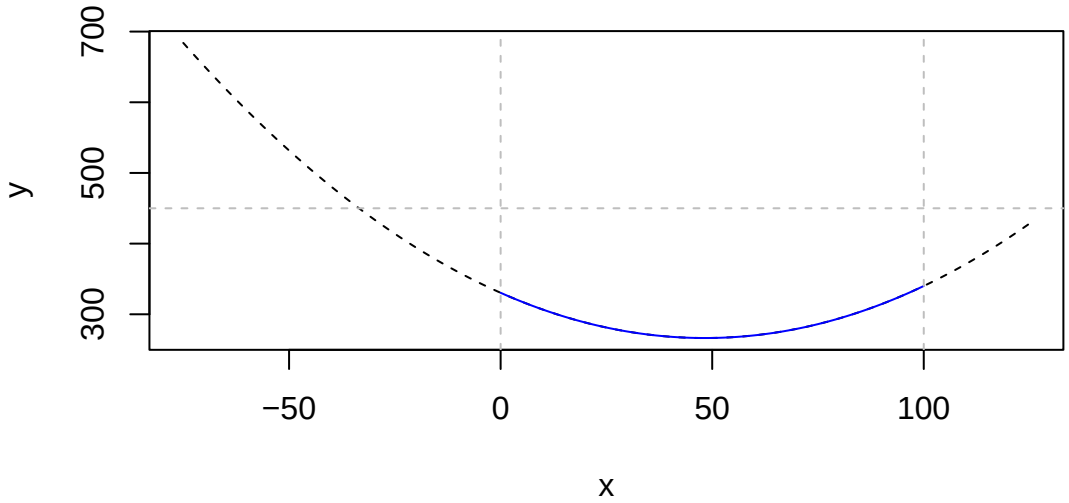



```r
##plot the quadratic over a wider range of Age
x = seq(-75, 125, by = 0.1)

y = predict(priceage.fit1, newdata = data.frame(Age = x))
plot(y ~ x, type = "l", lty = 2)

##The bits we want
lines(x[x >= 0 & x <= 100], y[x >= 0 & x <= 100], col = "blue")

##The range of Age & Price respectively
abline(v = range(home.df$Age), lty = 2, col = "grey")
abline(h = c(0, 450), lty = 2, col = "grey")
```



Method and Assumption Checks

WRITE M & A CHECKS HERE
Since data seem to suggest between Price and Age is a quadratic trend, so we have fitted a linear model with explanatory x and xˆ2 (quadratic model). We have a random sample of 76 homes so we can assume they form an independent and representative sample. The residual plot shows patternless residual. Althrough there wae a hint of non-constant scatters in the residual, they not change crazily. The normaility checks show slightly long tails, and sample quantiles not matches theoretical quantiles, but it roughly still matches, we may not have a normal data. And there are some slightly influential points according to the Cook's distance. So we still use our model and trust our estimate., but we need to be careful with our confidence interval. Our model is:
$Price_i = \beta_0 + \beta_1 \times Age_i + \beta_2 \times Age_i^2 + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$

Our model explains 15% of the total variation in the response variable. We have done a better job of modelling these data by adding the quadratic term, because it explained another 13% of the total variation.

Executive Summary

WRITE EXEC SUMMARY HERE
Our aim is to find out the relationship between home age and home price. We have strong evidence that suggest the quadratic relationship between home age and home price exists. We estimated every unit increases in home age is associated with an increasing value in the mean of home price in the range of 0.01060739 and 0.04437476. We have done a better job of modelling these data by using the quadratic model instead of simple linear model. We can conclude our model is reasonable.