STATS 201 Assignment 1

Liu Siyuan 2019210173

Due Date: 10.10

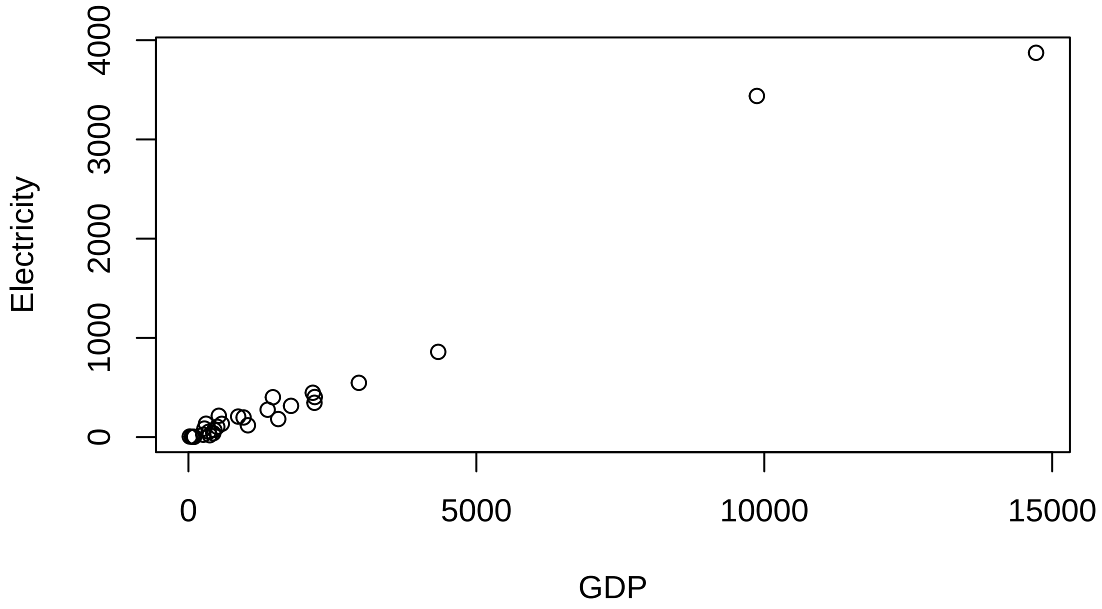
Question 1

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:

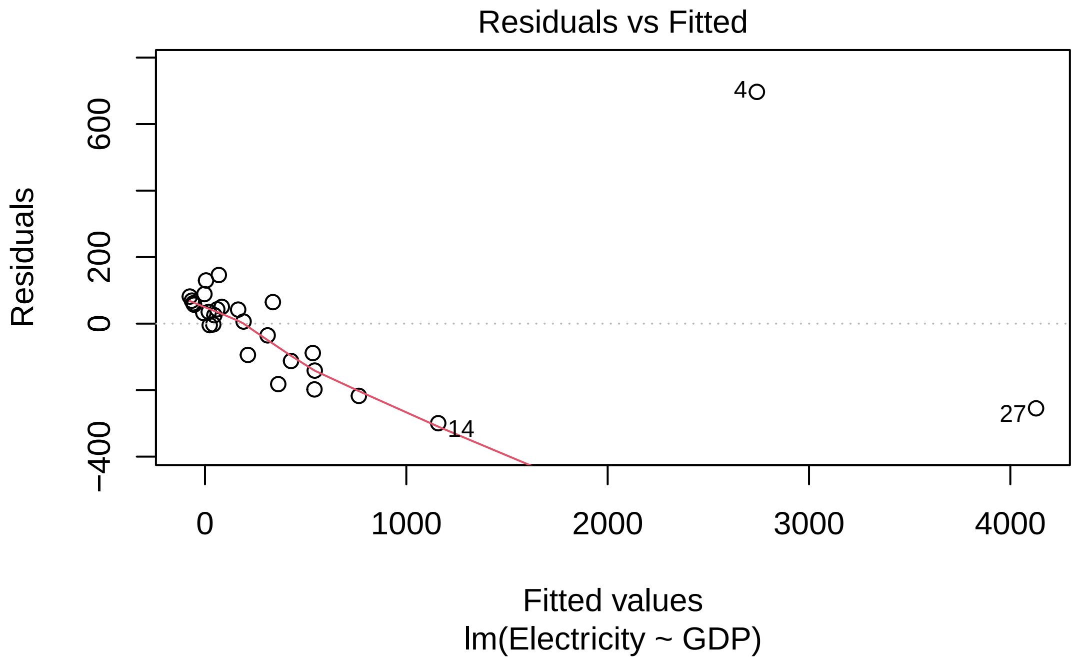
```
elec.df<-read.csv("electricity.csv")
plot(Electricity~GDP, data=elec.df)
```



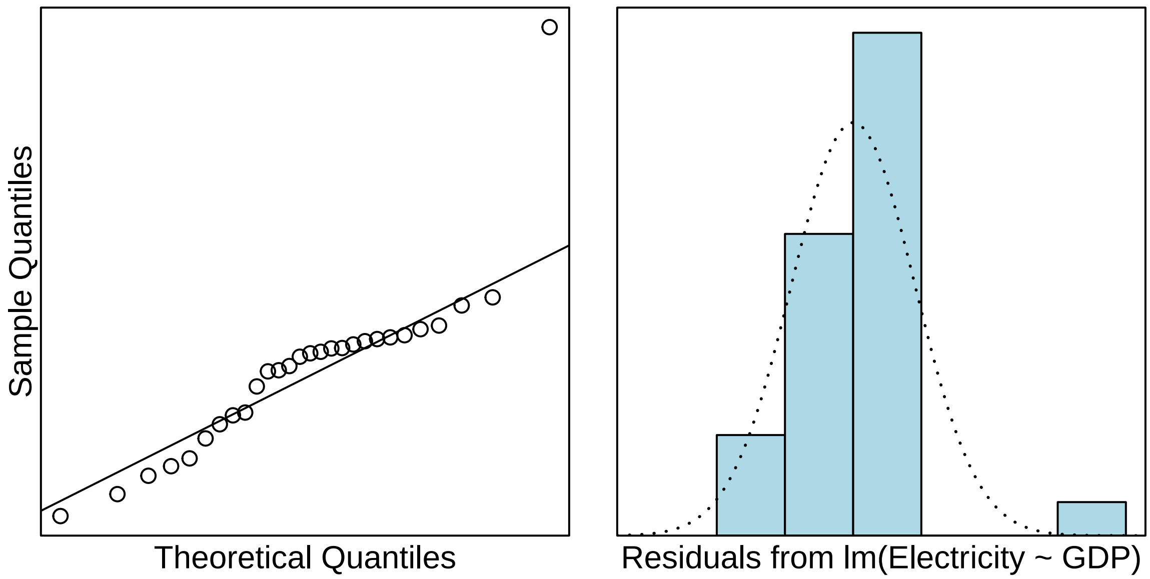
It is observed from the figure that most of the points show a linear distribution, except two points whose GDP are greater than 5000.

Fit an appropriate linear model, including model checks.

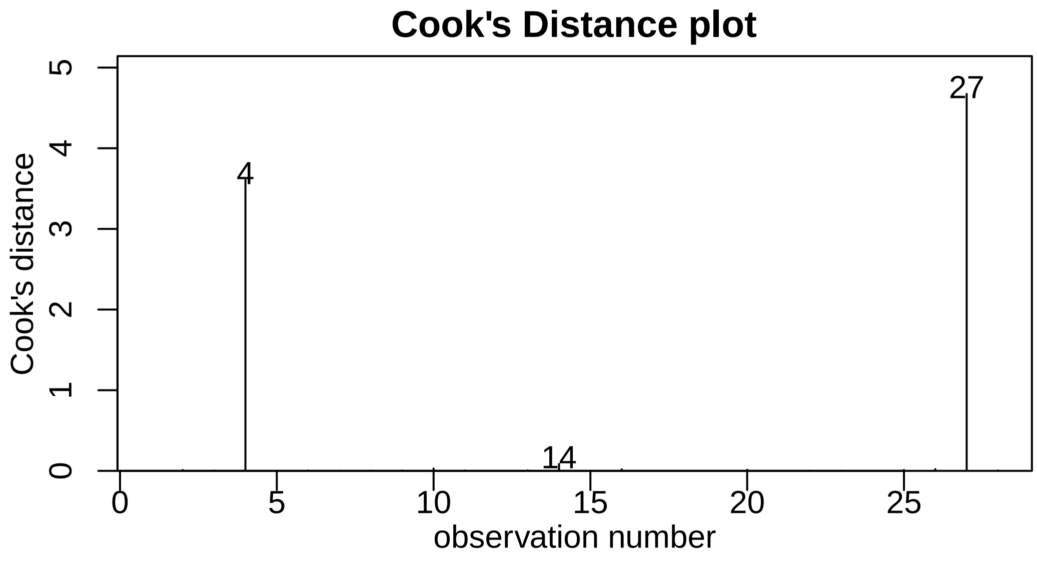
```
elecfit1=lm(Electricity~GDP,data=elec.df)
plot(elecfit1,which=1)
```



```
normcheck(elecfit1)
```



```
cooks20x(elecfit1)
```



Identify the two countries with GDP greater than 6000.

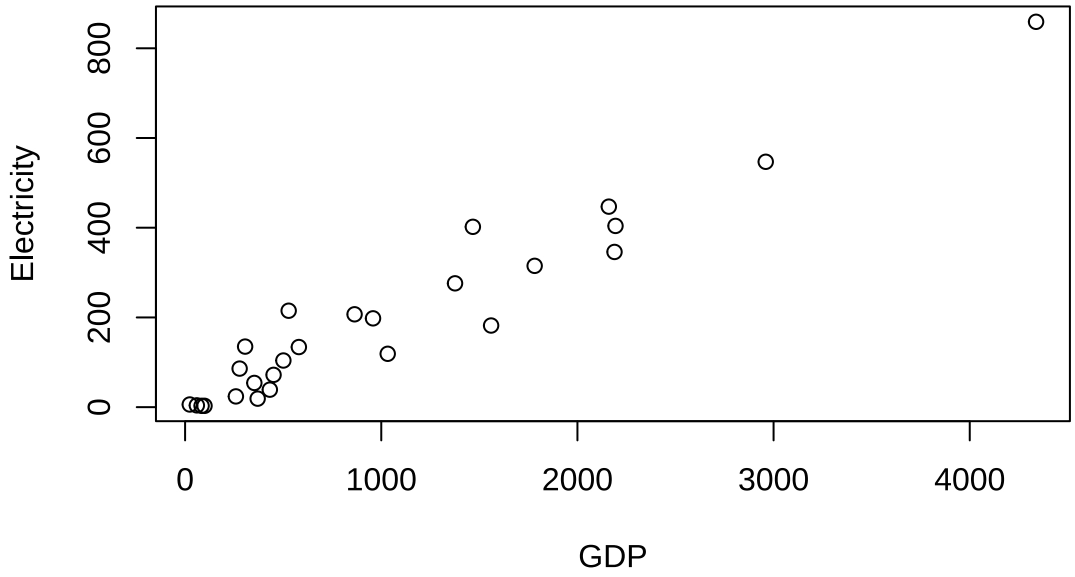
```
# could use some R code to do this
elec.df[elec.df$GDP>6000, ]
```

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

China and UnitedStates are the two countries whose GDP are greater than 6000. We need to eliminate these two points.

Replot data eliminating countries with GDP greater than 6000.

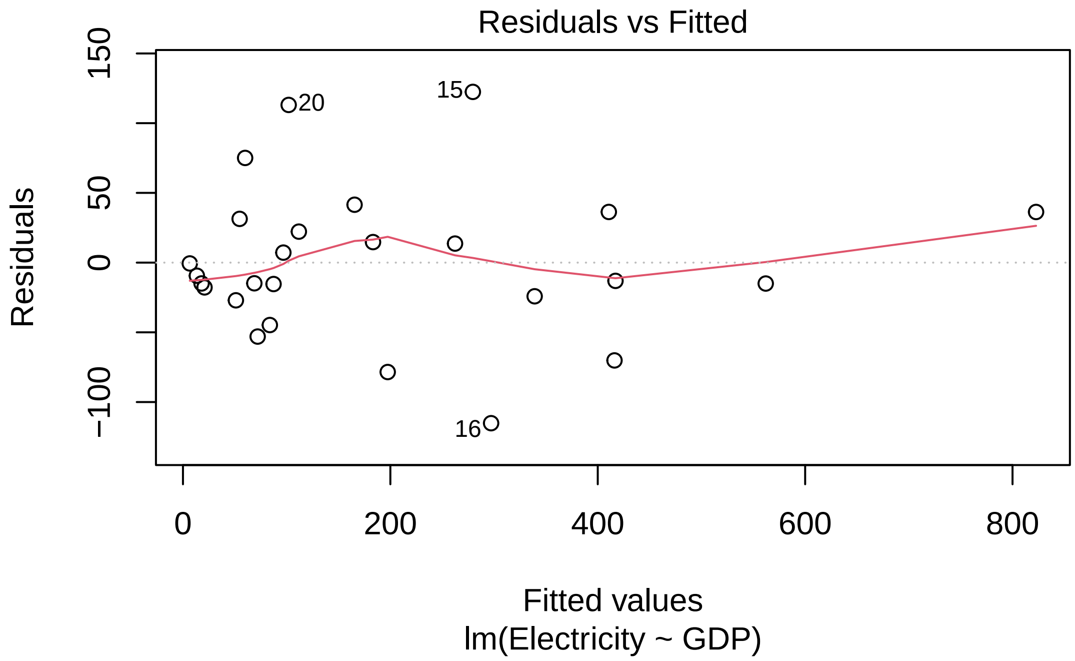
```
# Hint: If you want to limit the range of the data, do so in the data statement. E.G. something similar
elecdata.df = elec.df[elec.df$GDP<6000, ]
plot(Electricity~GDP, data=elecdata.df)
```



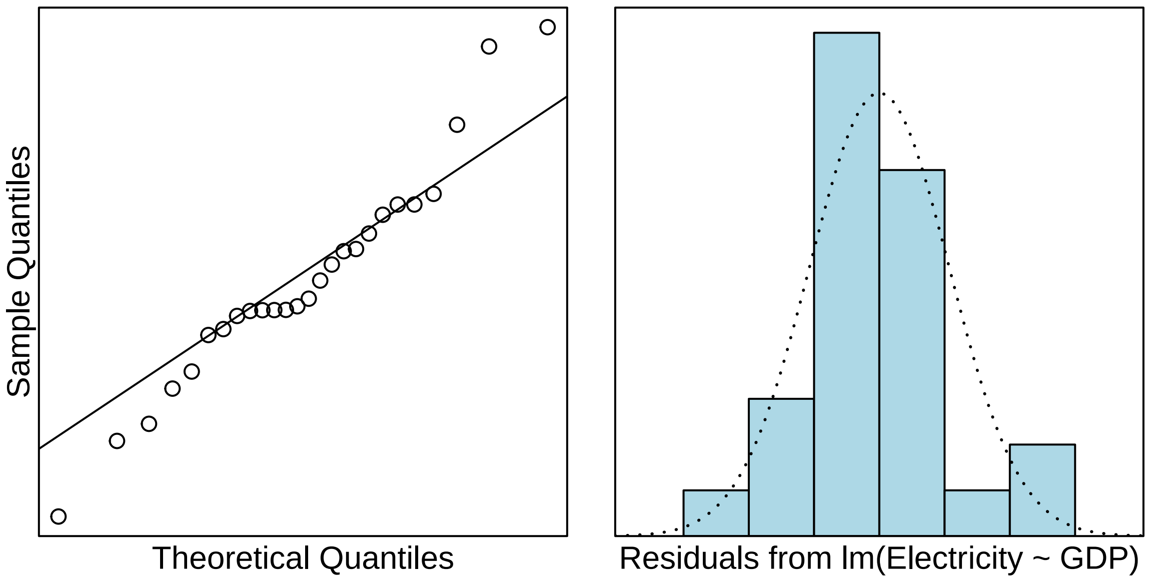
After eliminating the two countries, all the countries' GDP are less than 6000 and the data are linearly distributed from observing the figure.

Fit a more appropriate linear model, including model checks.

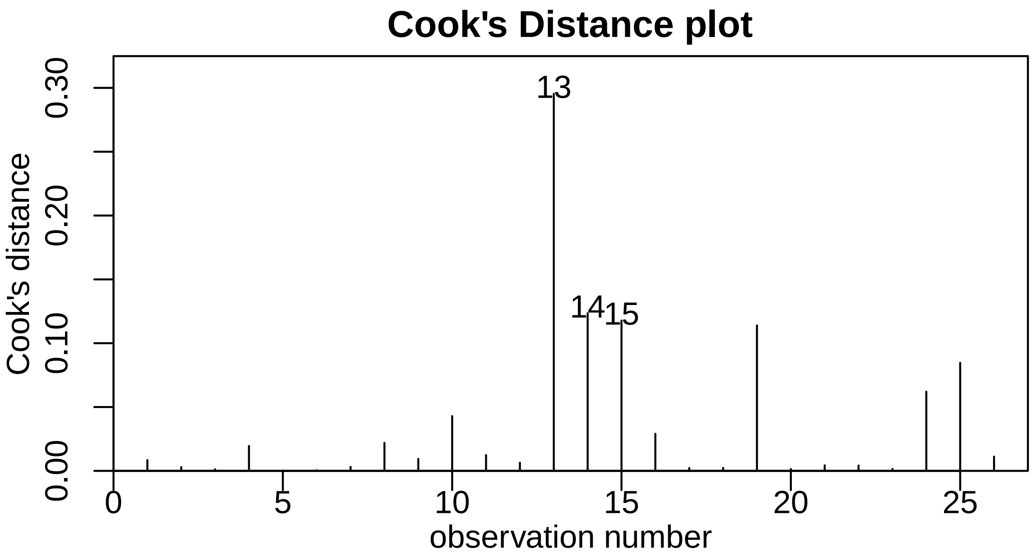
```
elecfit2=lm(Electricity~GDP,data=elecdata.df)
plot(elecfit2,which=1)
```



```
normcheck(elecfit2)
```

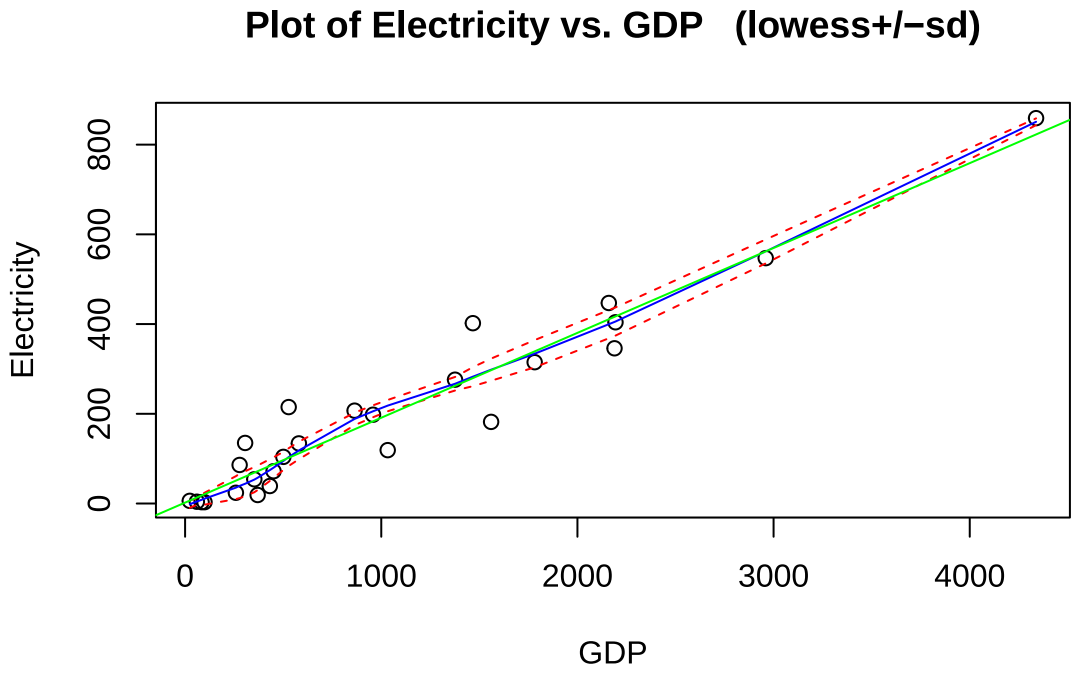


```
cooks20x(elecfit2)
```



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

```
trendscatter(Electricity~GDP, data=elecdata.df)
abline(coef(elecfit2), col = "green")
```



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

```
summary(elecfit2)
```

```
##
## Call:
## lm(formula = Electricity ~ GDP, data = elecdata.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
```

Abort this analysis, we wish to investigate the relationship between electricity consumption and the gross domestic product (GDP) for countries of the world, and in order to the accuracy of the model, we eliminated two countries that have a great impact on the model. We actually have the evidence that a relationships exists, it likes a linear relationship. Because of p-value is too small, such that 1.561e-15, so we have strongly evidence to explain that there does exist a linear relationship. And the β_1 represent the GDP, and we find that the estimated mean value of β_1 is equal to 0.18917 which is between 0.1676863 and 0.2106611. So we have 95% of the confidence intervals (0.1676863, 0.2106611) will contain the true value. In addition, we can obtain the Multiple R-squared is equal to 0.9322, which is very close to 1. In conclusion, our model explains 93% of the total variation in the response variable, and so will be reasonable for prediction. Between electricity consumption and the gross domestic product (GDP) has a strong linear relationship.

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

```
preds.df = data.frame(GDP = c(1000))
predict(elecfit2, preds.df, interval = "prediction")
```

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

Interpret the prediction and comment on how useful it is.

We calculated fitted values for GDP which got 1000 billion. And then we get the fixed value is that 191.2253. We actually know the range of the electricity usage when its GDP reaches 1000 billion. Though the width of the forecast may be a little big, it is still useful to predict the electricity usage through the value of GDP.

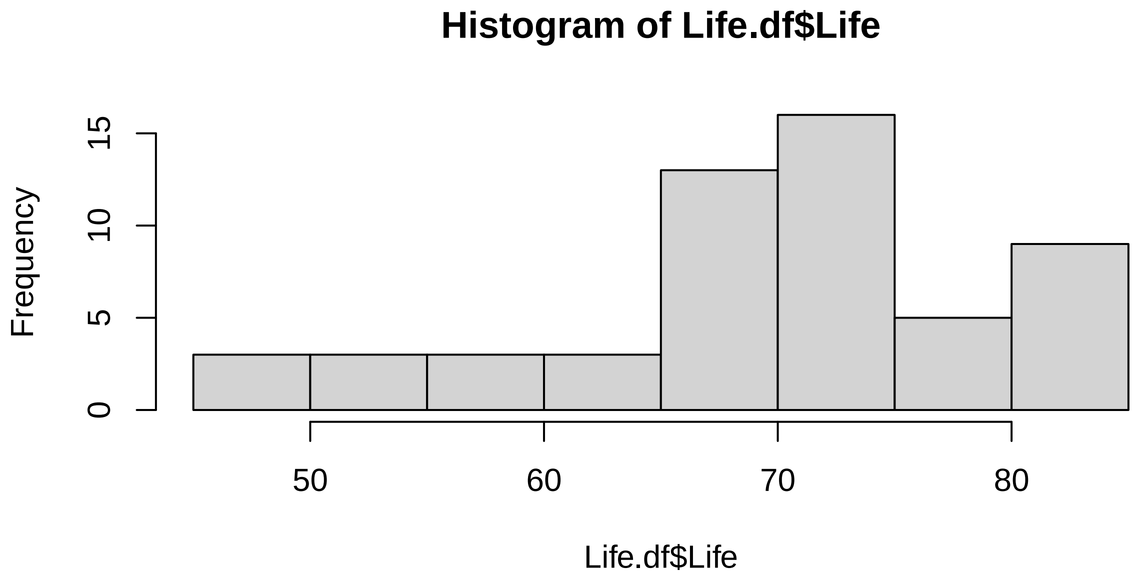
Question 2

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
```

Through the summary of the data we had, we obtain the mean of the life is 69.79. It is much closer to the expected number,68. And from observing the histogram, the peak is in the range of about 70.

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.

```
# Calculate the t-statistic
n = nrow(Life.df)
df = n - 1
mean_life_pre = mean(Life.df$Life)
mean_life_hyp = 68
se_life = sd(Life.df$Life) / sqrt(n)
t = (mean_life_pre - mean_life_hyp) / se_life
paste("t = ", t)
```

```
## [1] "t =  1.43268389415761"
```

```
#Calculate the CI
t_df_0.975 = qt(0.975, df)
df = n - 1
CI = mean_life_pre + c(-1, 1) *  t_df_0.975 * se_life
paste(CI, collapse = ",")
```

```
## [1] "67.2862880801168,72.2877482835196"
```

Using the t.test function

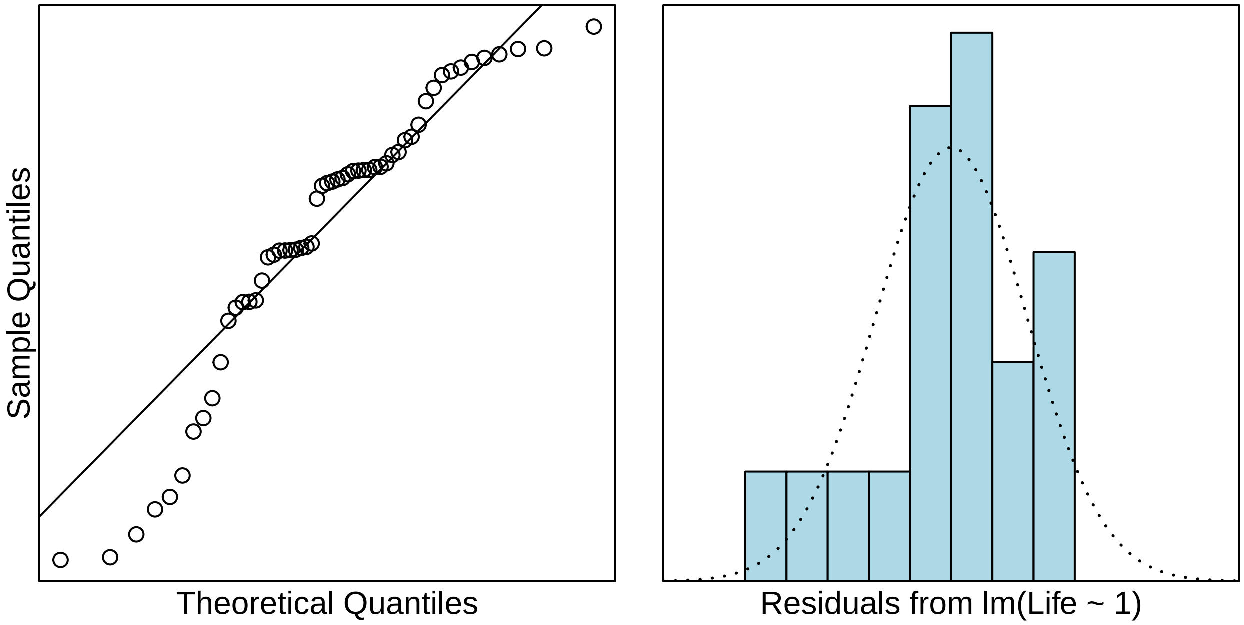
```
t.test(Life.df$Life, mu=68)
```

```
##
##  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
```

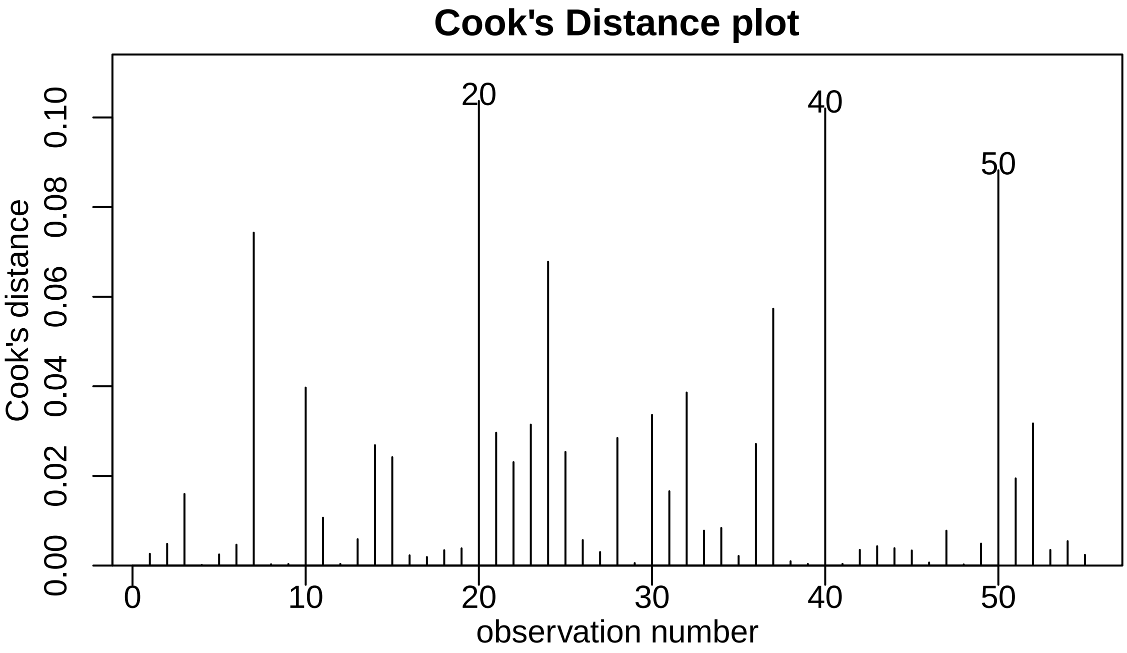
Note: You should get exactly the same results from the manual calculations and using the *t.test* function. Doing this was to give you practice using some R code.

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?

First, I think t-test and null linear fitting are two different concepts.

In the t-test, the null hypothesis is that the value of the true mean is equal to 68. And the alternative hypothesis is that the value of the true mean is not equal to 68. So the p-value means that if the mean value is equal to 68, the possibility of the data we get is the same or or even more extreme as we assumed.

In the null linear fitting model, the null hypothesis is that the value of the true mean is equal to 0. And the alternative hypothesis is that the value of the true mean is not equal to 0. So the p-value means that if the mean value is equal to 0, the possibility of the data we get is the same or or even more extreme as we assumed. Of course, the null linear model's p-value is much smaller than t-test's

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

To begin with, we want to estimate whether the mean value of the life in the world is equal to 68. First, we plot the histogram of the data and have a summary. The result is that the observing mean value is equal to 69.79. Then, we have a t-test. We find that the p-value is equal to 0.1577 which is too big, so we cannot refuse the null hypothesis that the mean value of is 68. In other words, the null hypothesis is valid. And we have 95% of the confidence intervals (67.28629, 72.28775) will contain the true value. It's more useful to predict the more closer mean value of life. And we don't have a R-square because we have a single-variable prediction.

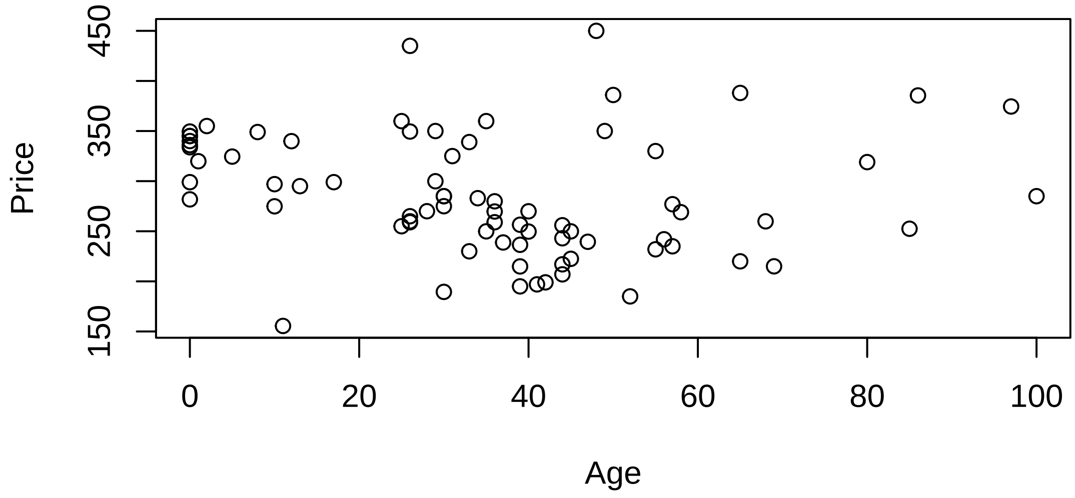
Question 3

Question of interest/goal of the study

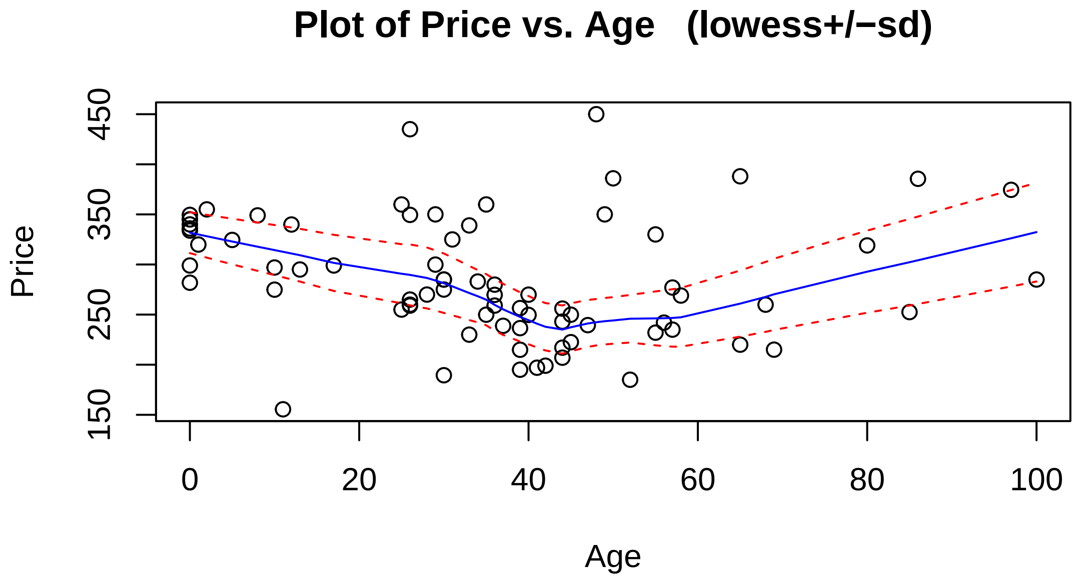
The question of the goal of the analysis is that figuring out the relationship between the price and age.

Read in and inspect the data:

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



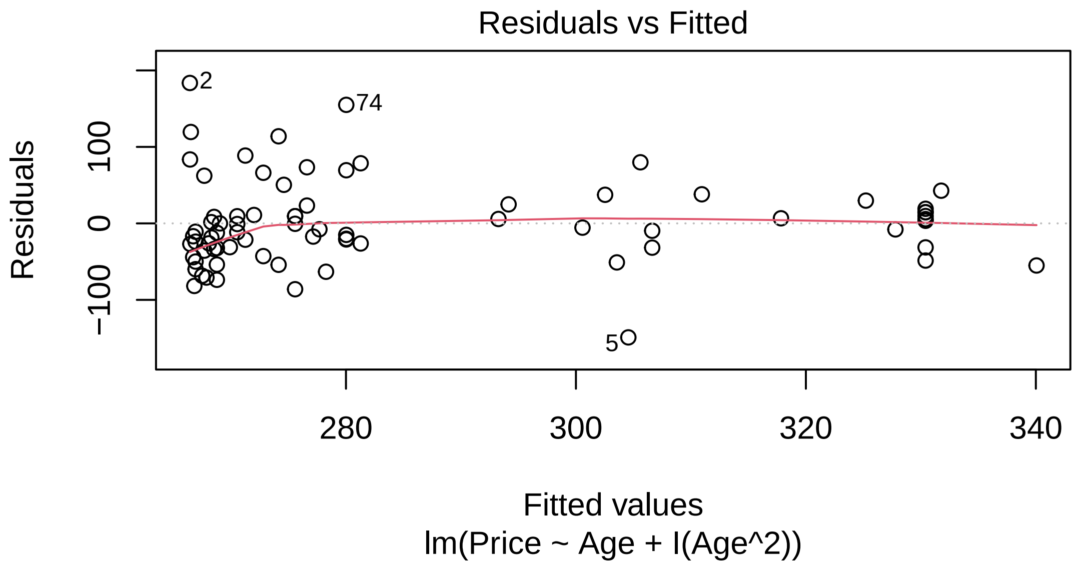
```
trendscatter(Price~Age,data=home.df)
```



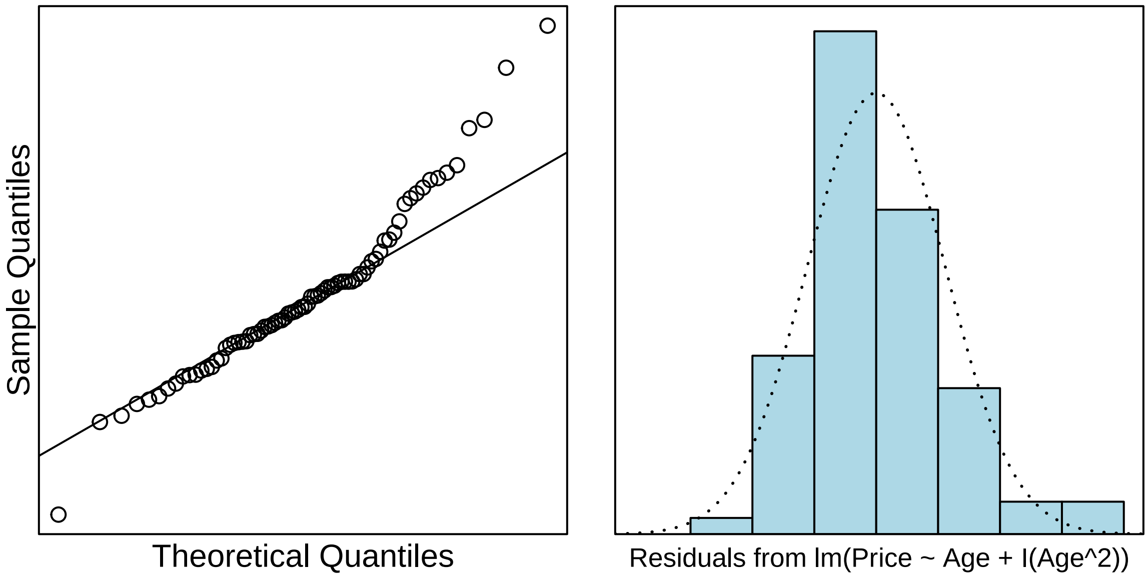
We plot the data and the trend scatter. We find that the price and age do not have a linear relationship. It's like quadratic. So we try to fit it with a quadratic regression model.

Fit an appropriate linear model, including model checks.

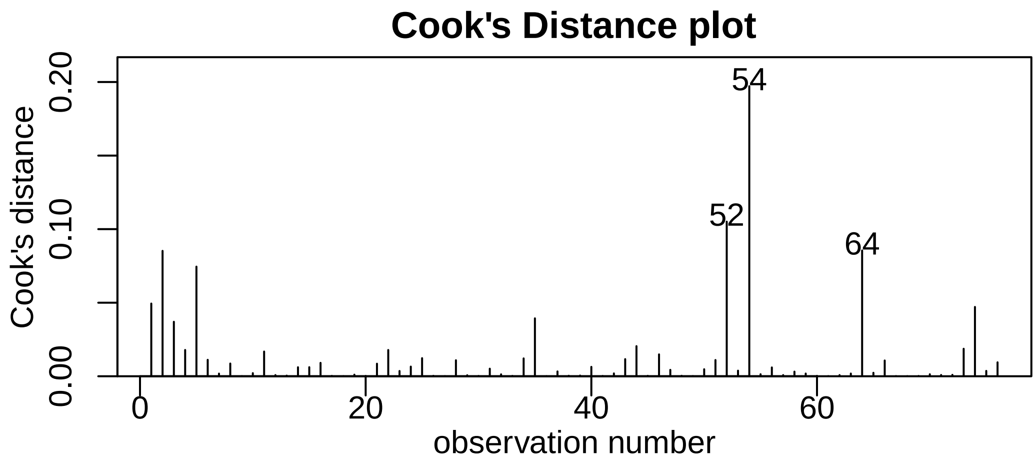
```
home.fit = lm(Price~Age+I(Age^2), data = home.df)
plot(home.fit, which = 1)
```



```
normcheck(home.fit)
```

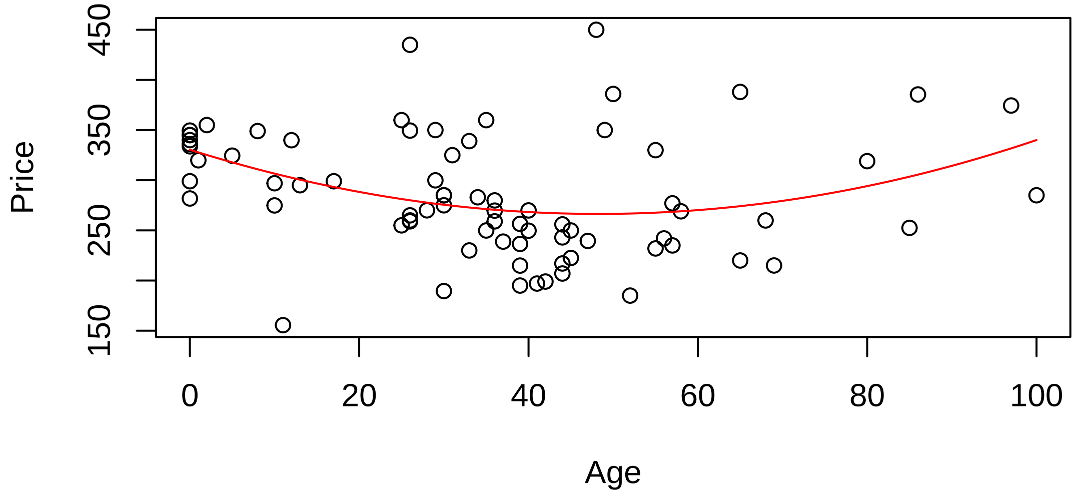


```
cooks20x(home.fit)
```

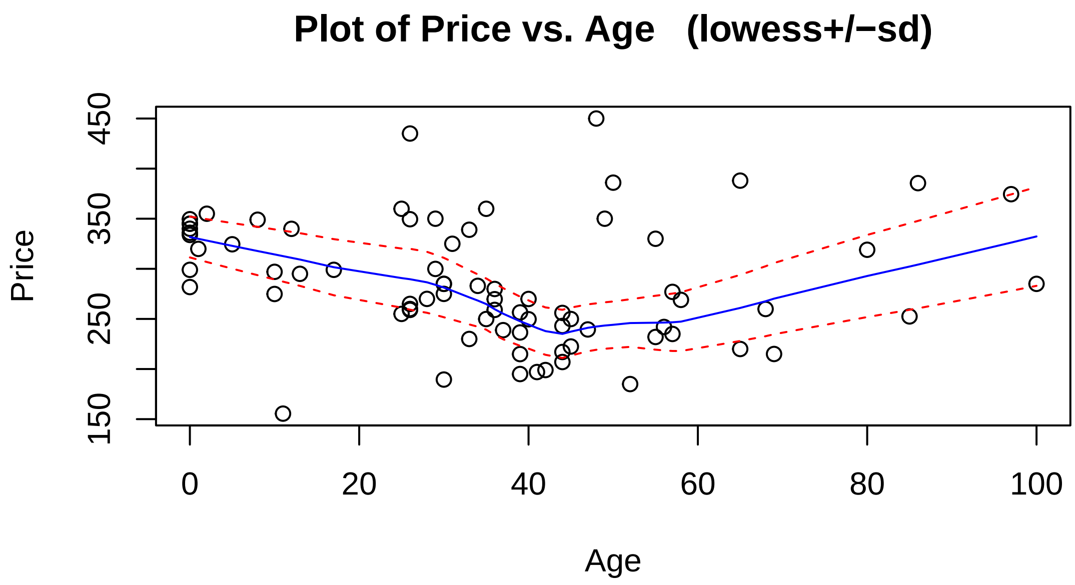


Plot the data with your appropriate model superimposed over it.

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



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



Method and Assumption Checks

```
nrow(home.df)
```

```
## [1] 76
```

```
summary(home.fit)
```

```
##
## 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(home.fit)
```

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

Since we have a quadratic model to fit the data, because the data we observed is not a linear model, and it has a curve, so we choose the quadratic model. We have 76 of the random single-family homes, and because of random, we could not doubt about the independence. These are likely to be minor, with a bigger concern being how representative the data is of a wider group of single-family homes. First, we plot the data and find it more suitable for quadratic model. Then, we plot the residual diagram for the EOV. The residuals show patternless scatter with fairly constant variability - so no problems. After this, 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: $Price_i = \beta_0 + \beta_1 \times Age_i + \beta_2 \times Age_i^2 + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$

Our model explains 14.68% of the total variation in the response variable, and so will be reasonable for prediction.

Executive Summary

We want to know how the sale price of a house is influenced by the age of the house. Then we fit the quadratic model to the data we obtain. We have evidence to explain that there does exist a quadratic relationship, because the p-value is equal to 0.001773. And the β_2 represent the the square of age, and we find that the estimated mean value of β_2 is equal to 0.02749 which is between 0.01060739 and 0.04437476. So we have 95% of the confidence intervals (0.01060739, 0.04437476) will contain the true value. In addition, we can obtain the Multiple R-squared is equal to 0.1468. In conclusion, our model explains 14.68% of the total variation in the response variable, and so will be reasonable for prediction. Between price and age has a quadraic relationship.