STATS 201 Assignment 1

Model Answers

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

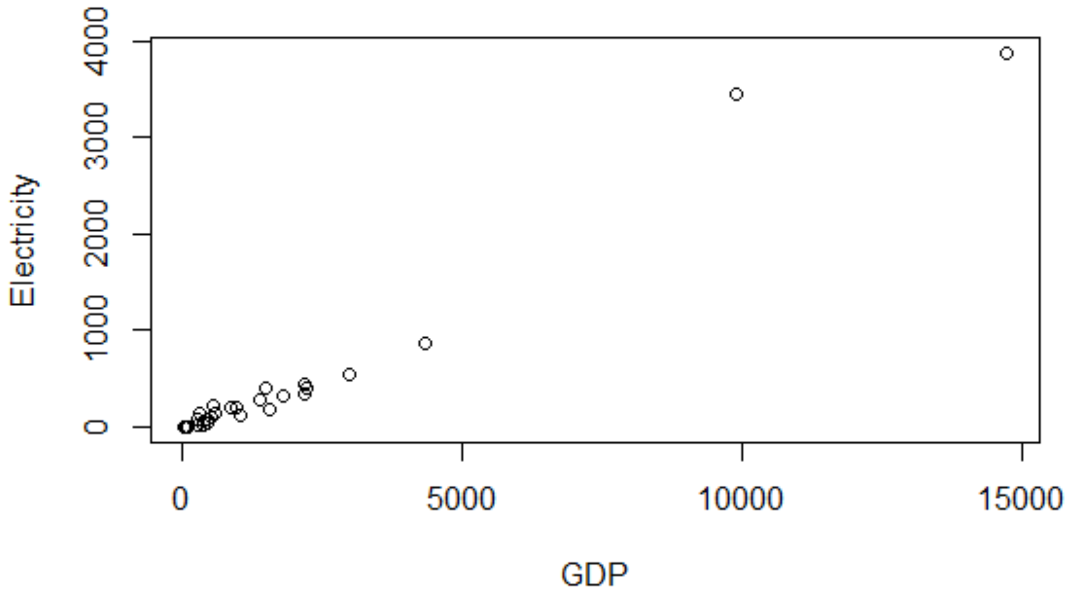
Question 1

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:

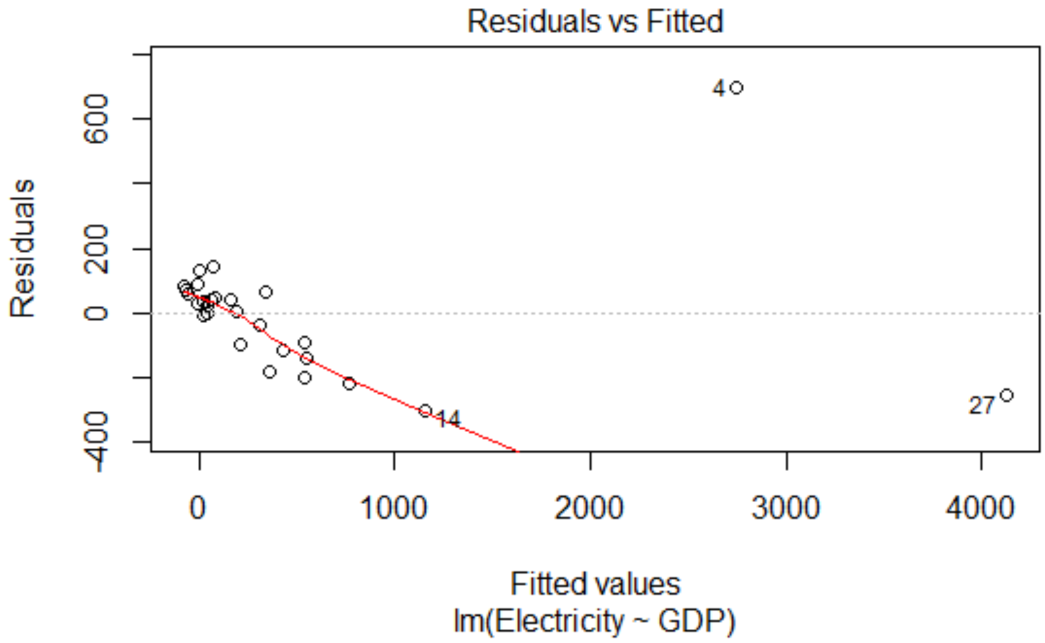
```
elec.df<-read.csv("electricity.csv")
plot(Electricity~GDP, data=elec.df)
```



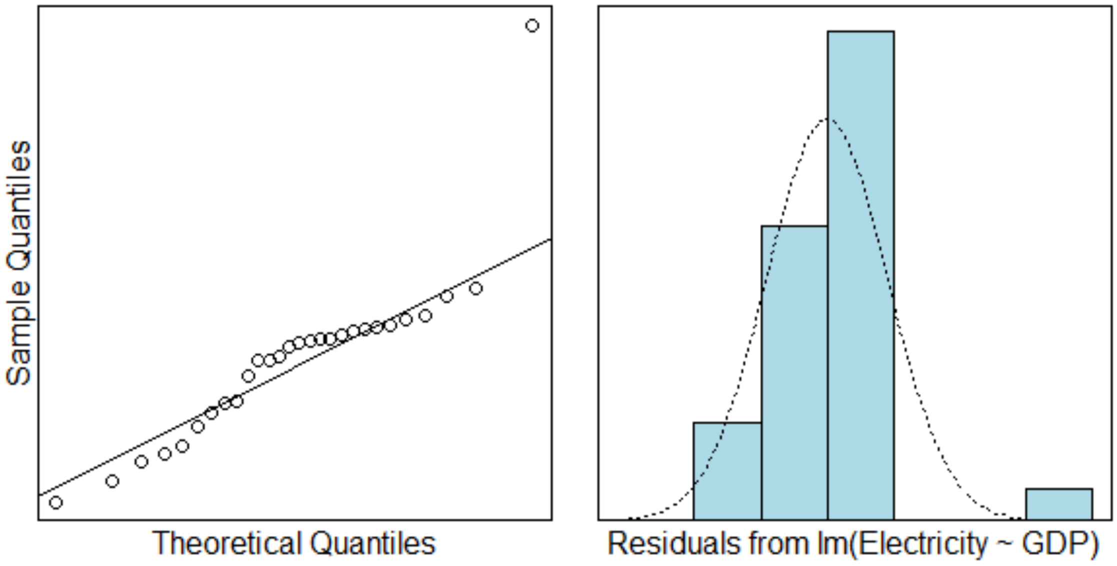
There is a positive linear trend in the data. There are two observations with very large GDP's that don't seem to lie exactly on the trend line and dominate the plot.

Fit an appropriate linear model, including model checks.

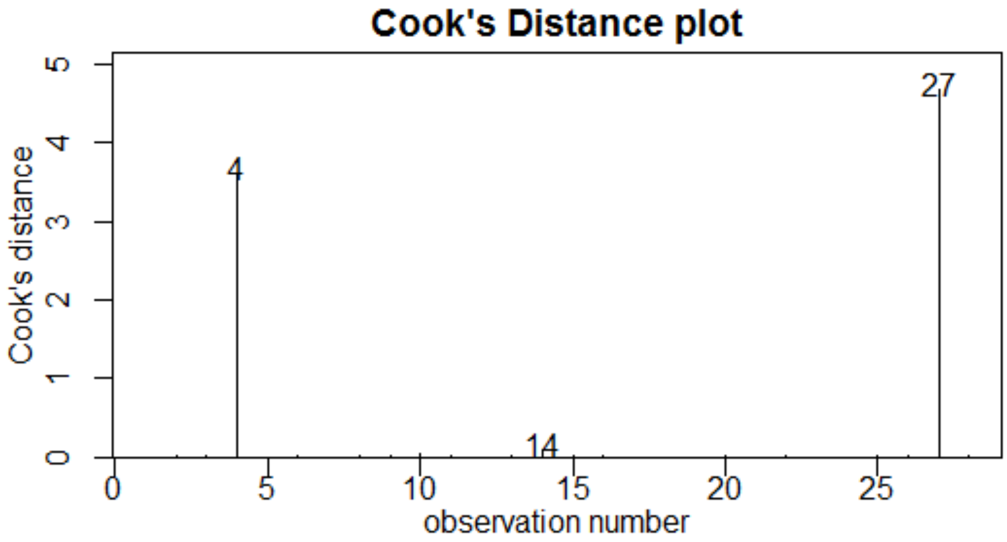
```
elecfit1=lm(Electricity~GDP,data=elec.df)
plot(elecfit1,which=1)
```



```
normcheck(elecfit1)
```



```
cooks20x(elecfit1)
```



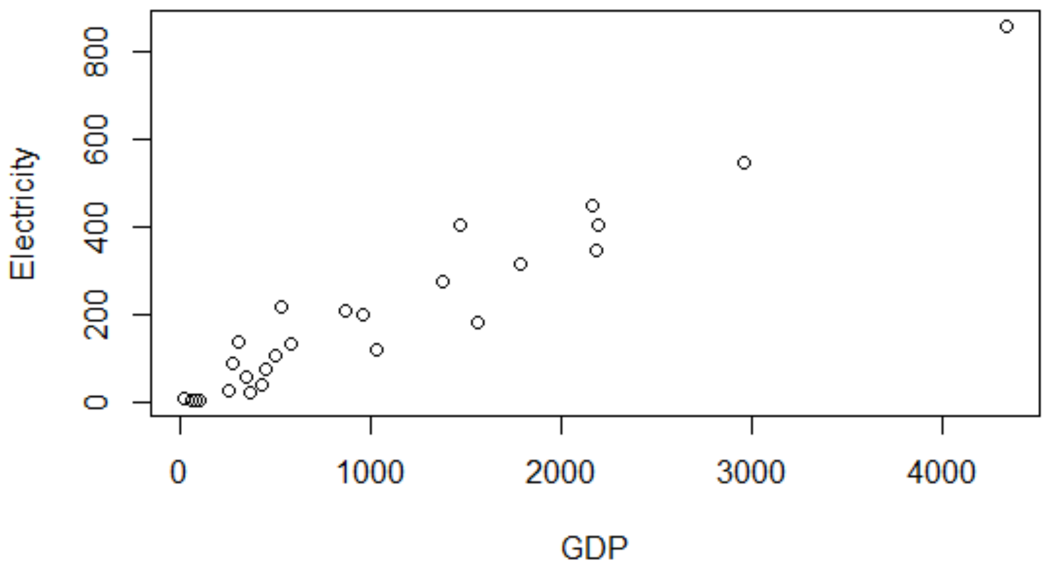
Identify the two countries with GDP greater than 6000.

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

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

The two countries with extremely high GDP are USA and China.

Replot data eliminating countries with GDP greater than 6000.

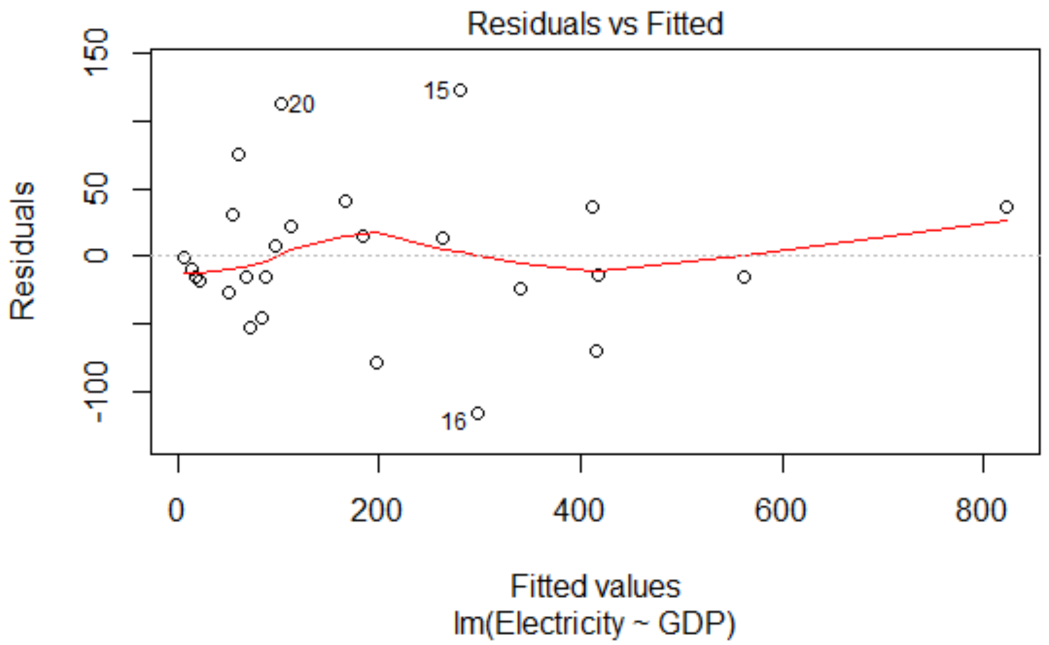
```
# Hint: If you want to limit the range of the data, do so in the data
statement. E.G. something similar to data=elec.df[elec.df$GDP>2000,]
plot(Electricity~GDP, data=elec.df[elec.df$GDP<6000,])
```



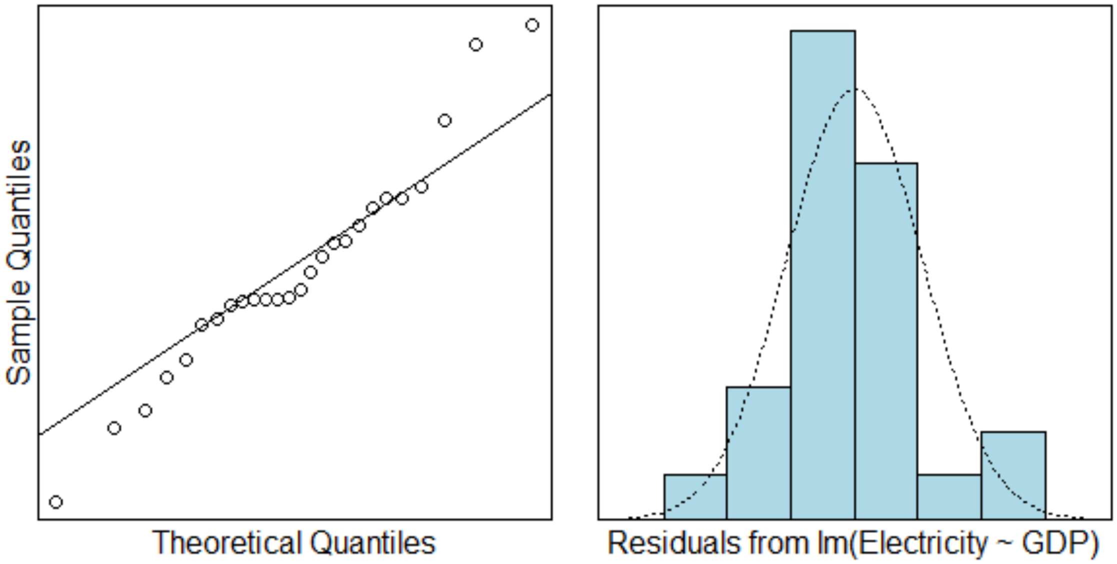
The plot still looks linear, but the relationship looks weaker. Most countries have GDP under 1000 billion dollars.

Fit a more appropriate linear model, including model checks.

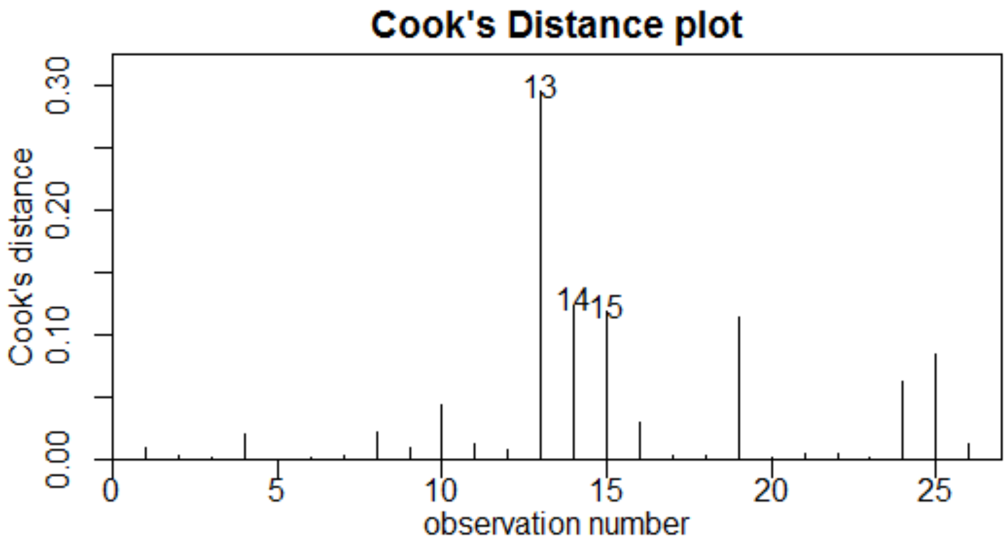
```
elecfit2=lm(Electricity~GDP,data=elec.df[elec.df$GDP<6000,])
plot(elecfit2,which=1)
```



```
normcheck(elecfit2)
```



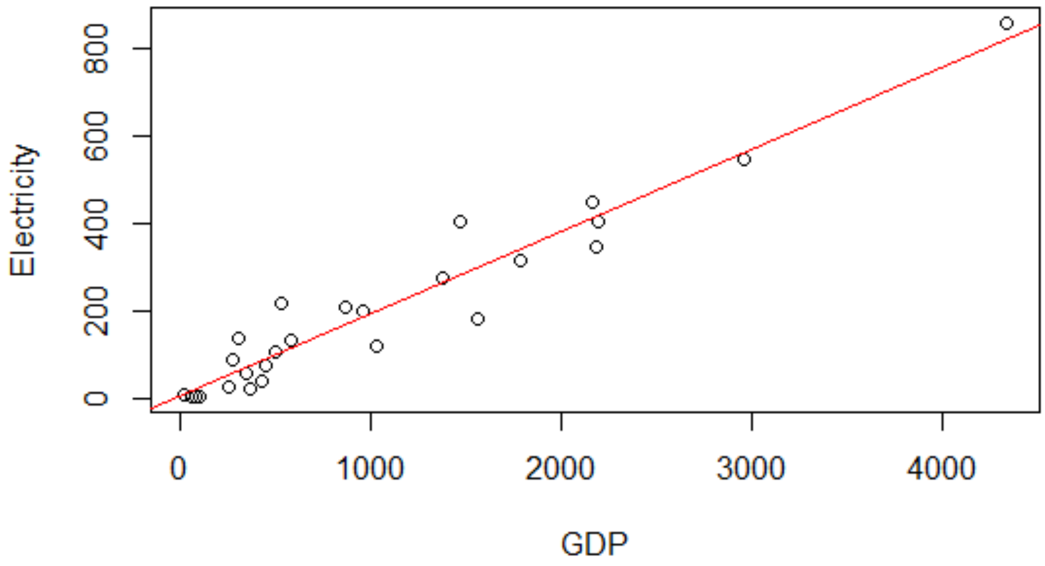
```
cooks20x(elecfit2)
```



```
summary(elecfit2)

##
## 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)

##                   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. Add additional slope 1 intercept 0 line.

```
plot(Electricity~GDP, data=elec.df[elec.df$GDP<6000,])
abline(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: $Electricity_i = \beta_0 + \beta_1 \times GDP_i + \epsilon_i$ 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

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.

We have strong evidence suggesting that electricity consumption is positively related to the GDP.

For each additional 100 billion dollars increase in GDP, the average electricity consumption increased by somewhere between 17 and 21 billion kilowatt-hours.

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

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

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

Interpret the prediction and comment on how useful it is.

We predict that a country with a gross domestic product of 1000 billion dollars will use between 76.3 and 306.2 billion kilowatt-hours of electricity.

The original range of electricity usages are between 0 and 850 billions of kilowatt-hours (so a range of 850). We have narrowed this down to 76-306, so a range of 230: about a quarter of the original range. So this is a good improvement although it is still not very precise.

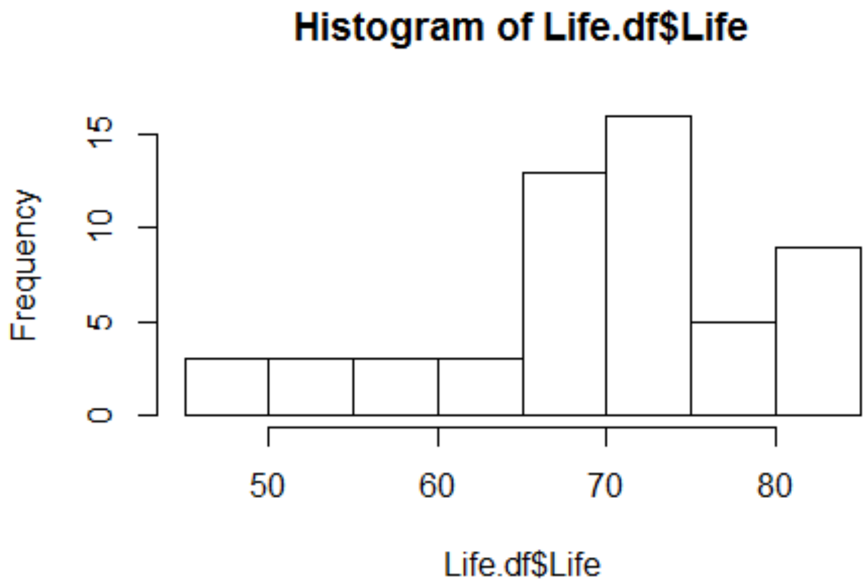
Question 2

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
```

The life expectancies appear to be centred around 70 and negatively (left) skewed.

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

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

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

```
n=length(Life.df$Life)
Tstat=(mean(Life.df$Life)-68)/(sd(Life.df$Life)/sqrt(n))
Tstat

## [1] 1.432684

mean(Life.df$Life)+c(-1,1)*qt(0.975,n-1)*sd(Life.df$Life)/sqrt(n)

## [1] 67.28629 72.28775
```

Using the t.test function

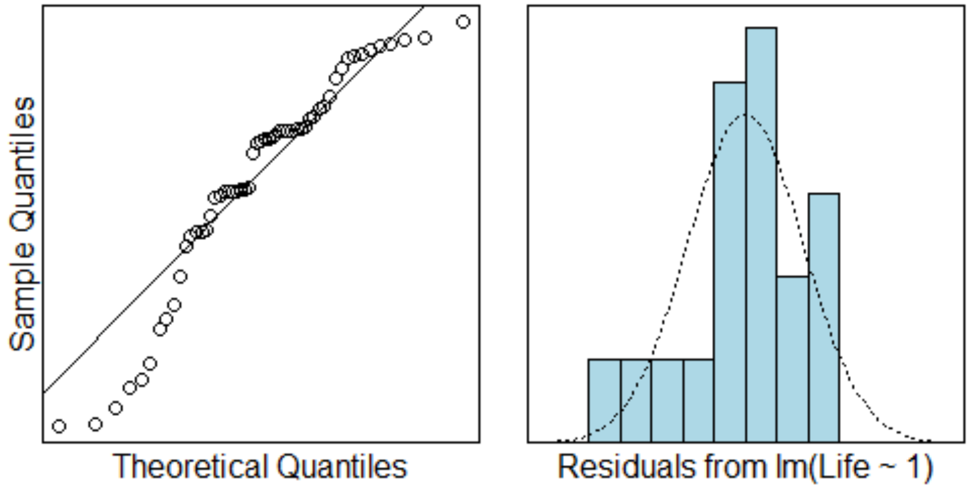
```
t.test(Life.df$Life, mu=68)
```

```
## 
##  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
```

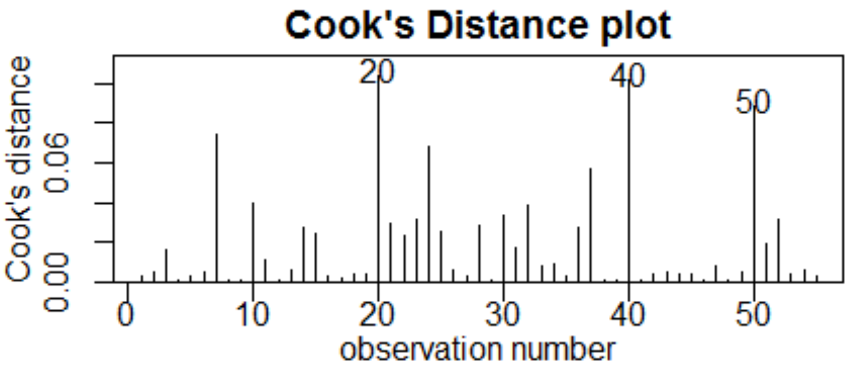
Note: You should get exactly the same results from the manual calculations and using the *t.test* function. Doing this was to give you practice using some R code.

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?

The t-test was modified to test for a hypothesised value = 68, however in the null linear model the default hypothesis test of hypothesised value = 0 was tested, giving a somewhat useless piece of information (we have extremely strong evidence that the average life expectancy of humans is greater than 0).

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

We are interested in estimating the mean human life expectancy among countries of the world and seeing if it is consistent with a value of 68 years.

We estimate that the mean life expectancy is somewhere between 67.3 and 72.3 years. This is consistent with the mean life expectancy being 68 years old. (Or we do not have any evidence to suggest that the mean life expectancy is different from 68 years.)

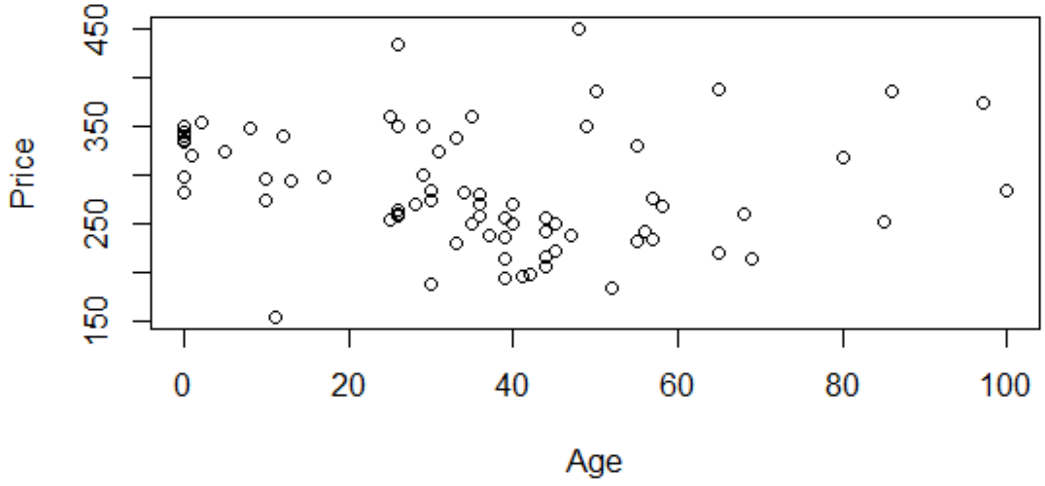
Question 3

Question of interest/goal of the study

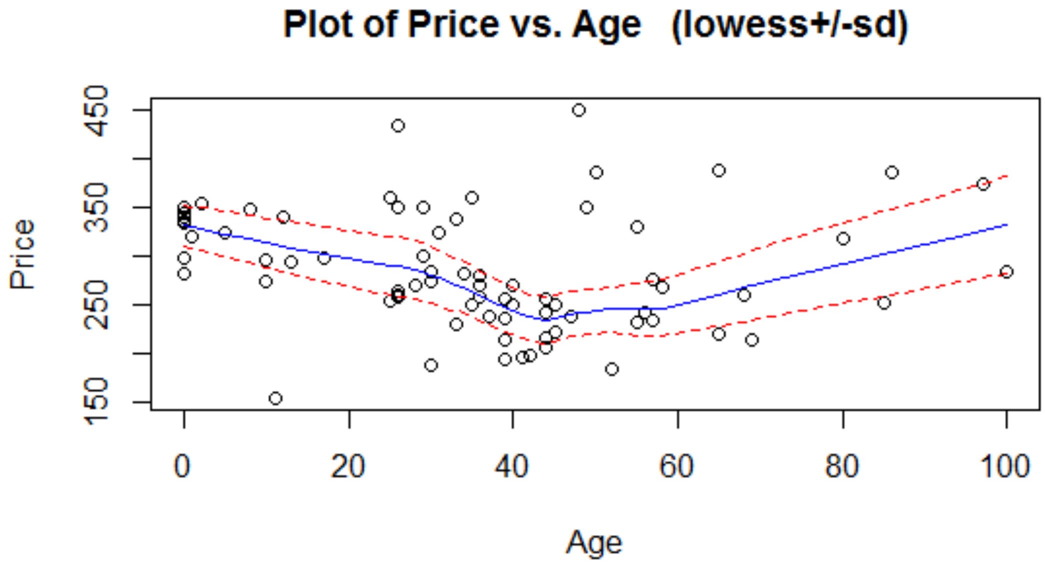
We are interested in how the sale price of a house is influenced by the age of the house (for the city of Eugene, Oregon).

Read in and inspect the data:

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



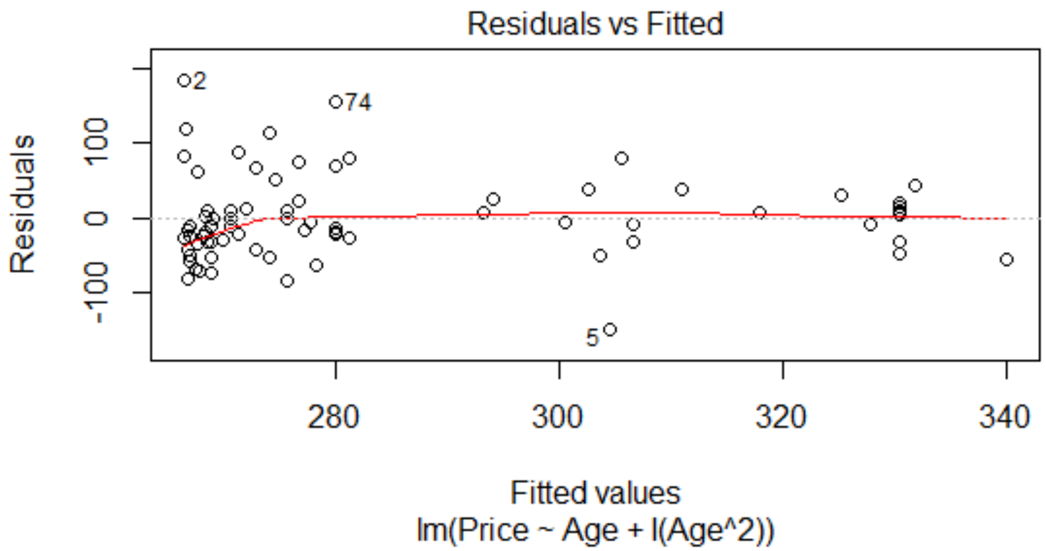
```
trendscatter(Price~Age,data=home.df)
```



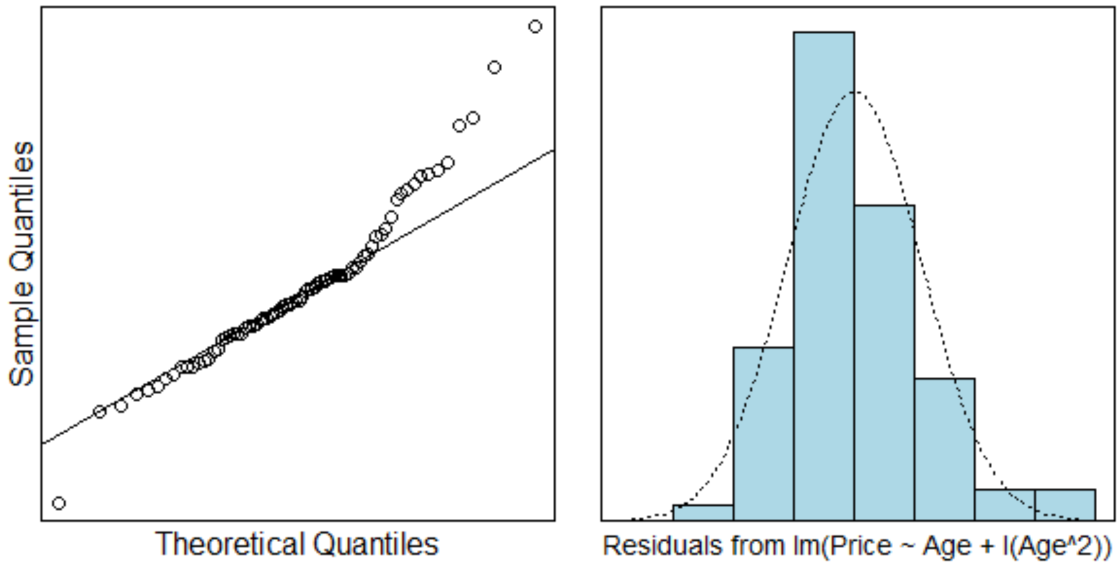
From the plot, we see there is some curvature in the plots. The newest and oldest houses are the most expensive with houses around 40 years old being cheaper. The scatter looks reasonably constant based on the smoother lines. A model with quadratic term is suggested.

Fit an appropriate linear model, including model checks.

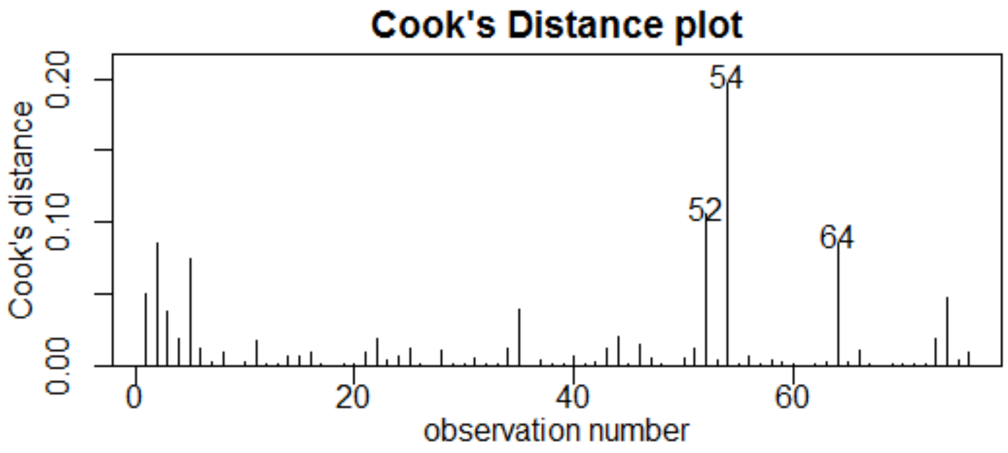
```
homefit1=lm(Price~Age+I(Age^2),data=home.df)
plot(homefit1,which=1)
```



```
normcheck(homefit1)
```



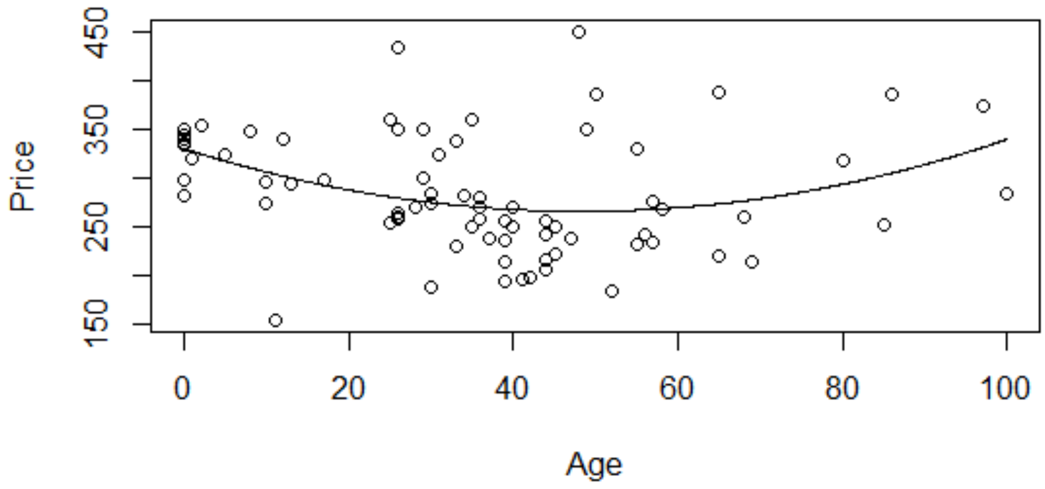
```
cooks20x(homefit1)
```



```
summary(homefit1)

##
## 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
```

Plot the data with your appropriate model superimposed over it.

```
pred.Age = data.frame(Age = seq(0, 100, 0.1))
pred = predict(homefit1, newdata = pred.Age)
plot(Price~Age,data=home.df)
lines(pred ~ pred.Age[, 1])
```



Method and Assumption Checks

We fitted a linear model with quadratic term as exploratory plots revealed some curvature. After fitting the quadratic, the residuals were fine, there were no problems with normality and no unduly influential points. We have independence from taking a random sample.

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

Our model only explained 15% of the variation in the data.

Executive Summary

It was of interest to model the relationship between sale price of a house and the age of a house.

We found strong evidence suggesting there is a relationship between house prices and their ages. In particular, we found the housing price decreases as the age of house increases when the house is younger than about 50 years. The housing price increases as the age of the house increases for houses older than about 50 years old.