STATS 201/8 Assignment 3

Anish Hota ahot228

Due Date: 3pm, Thursday 22nd August

1 Question 1 [18 Marks]

The manager of a company wants to investigate which section they should advertise their sport equipment in local newspapers, and they want to know how many inquiries they will receive resulting from the advertisement. They advertised their products in 200 local newspapers around America. They randomly allocated the newspapers into two groups. In one group they advertised in the business section of the paper, while in the other they advertised in the sports section. The company recorded the number of inquiries resulting from the advertisement in each of the areas.

The dataset is stored in *advertise.csv* and includes variables:

Variable	Description
Inquiries	The number of inquiries resulting from the advertisement in an area.
Section	The section of the local newspaper (Business or Sports) the advertisement was placed in for that area.

The questions the manager is particularly interested in are: does there tend to be a difference in the number of inquiries depending on which section they advertise in? If so, how big is the difference? Also, they want to estimate the average amount of inquiries generated when advertising in the section that gives the best results (give both estimates if you can't decide).

Instructions:

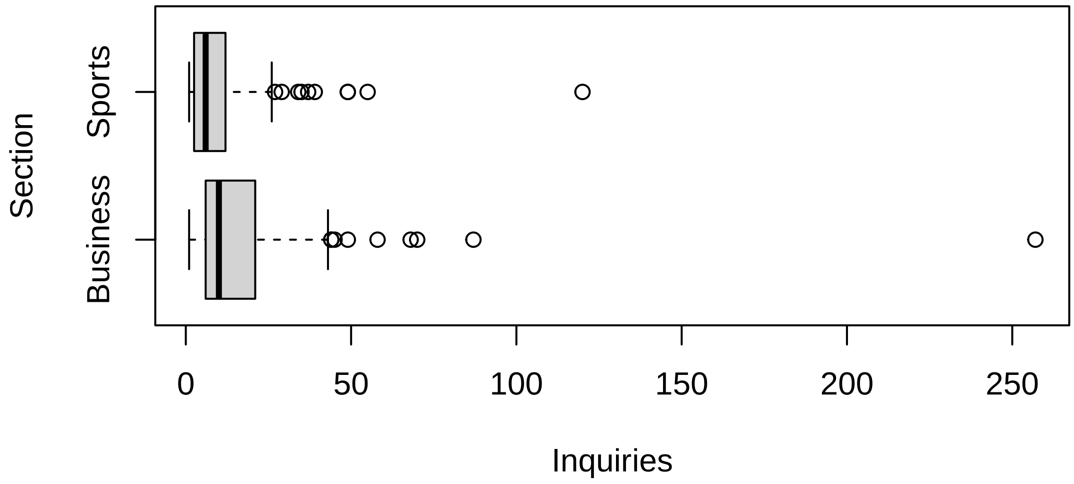
- Make sure you change your name and UPI/ID number at the top of the assignment.
- Comment on the plots and summary statistics of the data.
- Comment why it is more appropriate to use a log model for this data. (Consider the shape of the data here and discuss what is a likely explanation for this distribution.)
- Fit an appropriate model to the data. Check the model assumptions.
- Write appropriate Methods and Assumption Checks.
- Write an appropriate **Executive Summary**. (Remember to answer ALL the questions asked.)

1.1 Question of interest/goal of the study

It was of interest to learn where to advertise the company's product. We also want to know how many inquiries the company tended to received resulting from the advertisement.

1.2 Read in and inspect the data:

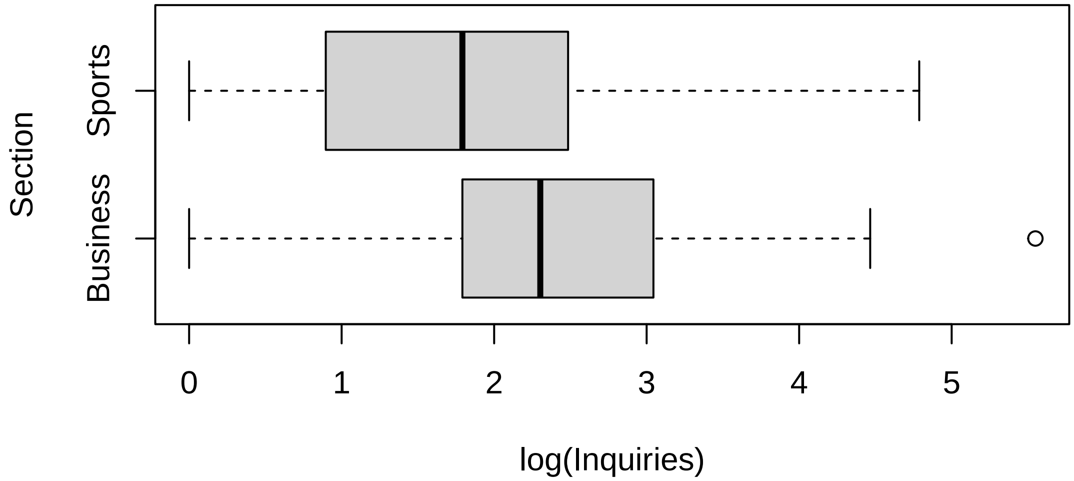
```
advertise.df=read.csv("advertise.csv",header=T, stringsAsFactors = TRUE)
boxplot(Inquiries~Section,horizontal=TRUE,data=advertise.df)
```



```
summaryStats(Inquiries~Section,data=advertise.df)
```

```
##          Sample Size  Mean Median  Std Dev Midspread
## Business         100 18.22     10 28.91841     15.00
## Sports           100 10.78      6 15.67381      9.25
```

```
boxplot(log(Inquiries)~Section,horizontal=TRUE,data=advertise.df)
```



```
summaryStats(log(Inquiries)~Section,data=advertise.df)
```

```
##          Sample Size     Mean   Median   Std Dev Midspread
## Business         100 2.376020 2.302585 0.9827192  1.252763
## Sports           100 1.755483 1.791759 1.1066796  1.487661
```

1.3 Comment on plots and summary statistics

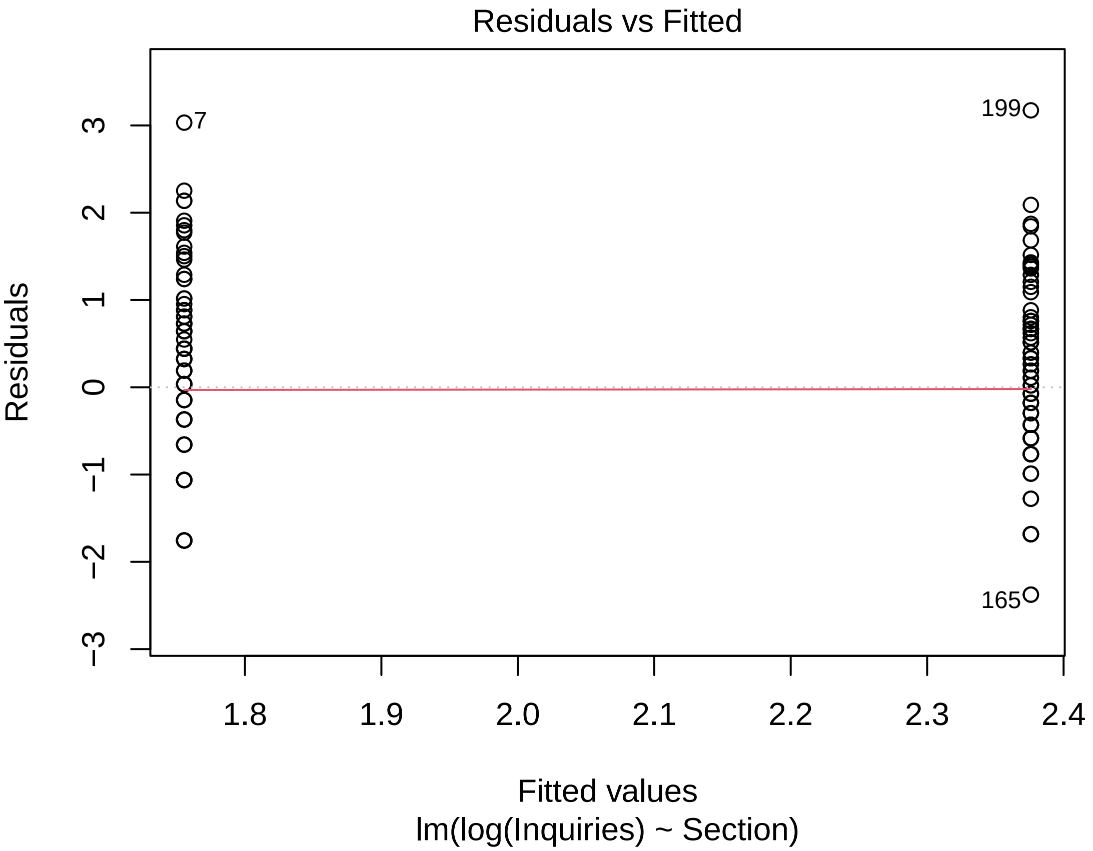
There seems to be more inquiries the business section than the sport section. The data seems to be slightly right skewed. There also doesn't seem to be an equality of variance.

1.4 Comment why it is more appropriate to use a log model for this data. (Consider the shape of the data here and discuss what is a likely explanation for this distribution.)

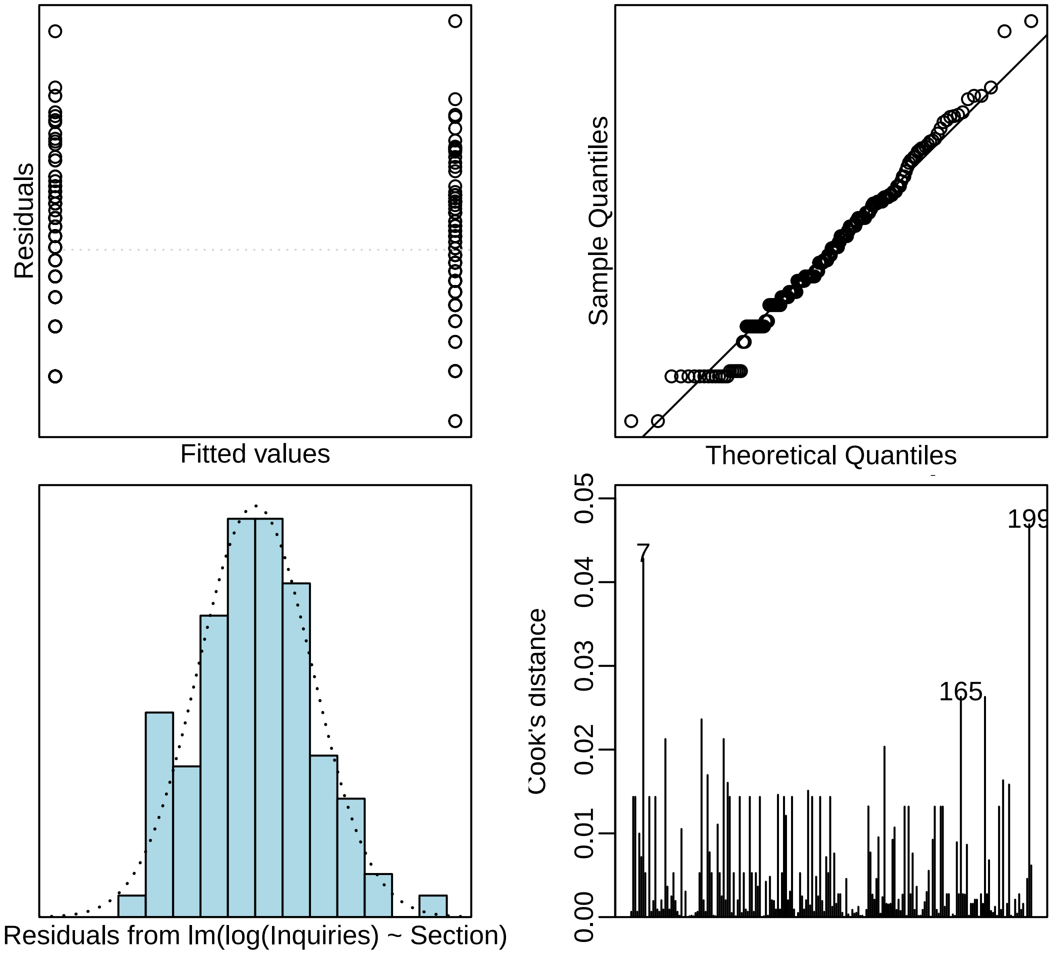
The sections don't have equality of variance (the business boxplot is much wider than the sports section) and the data was right skewed. The box plots are also quite small and hard to see. This is probably why logs of the graph have been taken, it has better variance and we can examine the data much better.

1.5 fit model and check assumptions

```
advertise.fit = lm(log(Inquiries)~Section,data=advertise.df)
plot(advertise.fit, which = 1)
```



```
modcheck(advertise.fit)
```



```r
summary(advertise.fit)
```

```
##
## Call:
## lm(formula = log(Inquiries) ~ Section, data = advertise.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3760 -0.6569 -0.0258  0.6685  3.1731
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.3760     0.1047  22.704  < 2e-16 ***
## SectionSports  -0.6205     0.1480  -4.193 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.047 on 198 degrees of freedom
## Multiple R-squared:  0.08154,    Adjusted R-squared:  0.07691
## F-statistic: 17.58 on 1 and 198 DF,  p-value: 4.154e-05
```

```r
exp(confint(advertise.fit))
```

```
##                   2.5 %    97.5 %
## (Intercept)   8.755140 13.228831
## SectionSports 0.401559  0.719878
```

1.6 Method and Assumption Checks

The boxplot of section vs inquiries revealed that the data was right-skewed and there was no equality of variance, that meant that we logged inquiries. So we fitted the a linear model with log(Inquiries) and Section.

Our histogram of the residuals are reasonably symmetric and all other assumptions are satisfied.

Our final model is:

$Section_i = \beta_0 + \beta_1 * log(Inquiries)_i + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$

Our model explained 8.15% of the variability in each section with logged inquiries. ## Executive Summary (Remember to answer ALL the questions asked.)

We have strong evidence that we should advertise the company's product in the business section.

We estimate that sports section has between 40% and 72% the amount of inquiries as the business section.

The newspaper recieved 200 inquiries regarding the advertisements during the study.

2 Question 2 [9 Marks]

'Earning time' is a measure of purchasing power in cities around the world. For a worker earning the average wage in a given city, it measures how much time must be worked in order to pay for a commodity such as a Big Mac burger or an iPhone in that city. The earning time of an item is therefore a measure of the effective price of the item relative to earnings.

In 2009 the UBS bank compiled data on earning times for various commodities in 72 cities around the world. We wish to investigate how the earning time for an expensive commodity (the iPhone 4S) compares with that of a cheap commodity (a Big Mac burger). In particular, we wish to investigate how iPhone earning times change as Big Mac earning times increase: what is the percentage change in earning time for the iPhone, for every 50% increase in Big Mac earning times?

The resulting data is in the file *EarningTimes.csv*, which contains the variables:

Variable	Description
City	The city name,
iPhone	The earning time for the iPhone 4S in that city, measured in hours,
BigMac	The earning time for a BigMac burger in that city, measured in minutes.

For this question we have provided you all the relevant output you need AND some **incorrect** output. You do not need to provide any additional output. Just answer the questions.

Instructions:

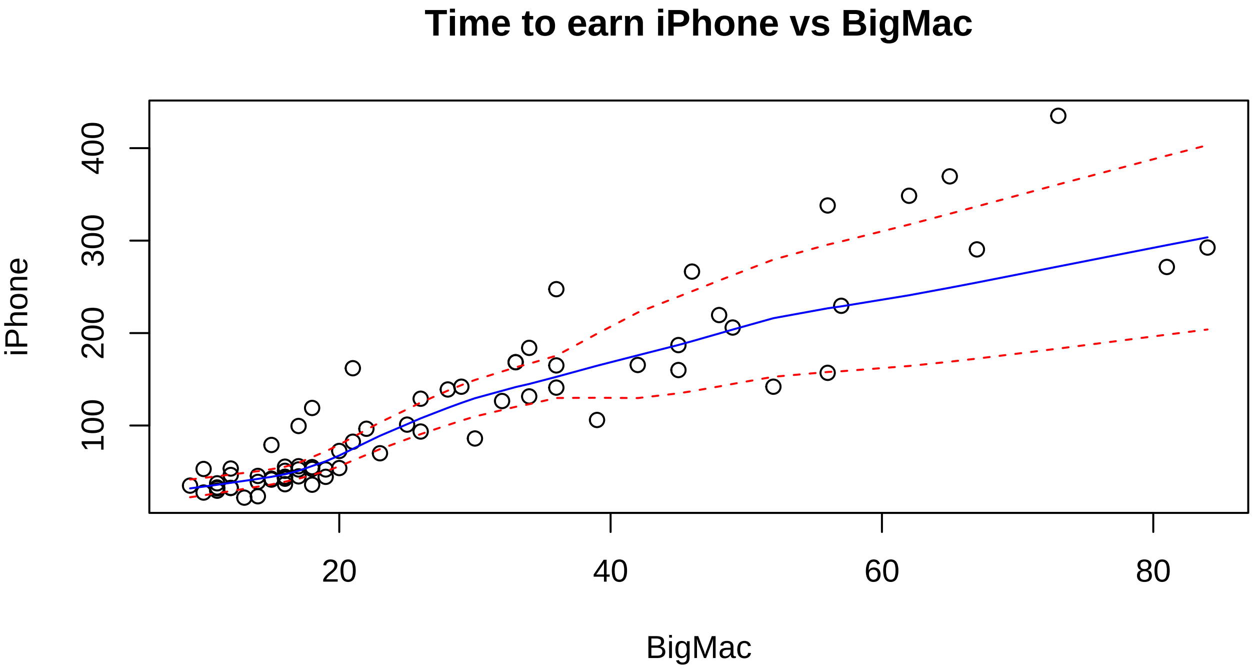
- Look at the three initial plots of the data and comment on them.
- State why a log-log (power) model is appropriate here.
- Do we have evidence that earning time for one of these items (iPhone or BigMac) grows more quickly than that of the other as we range from cheaper to more expensive cities? If so, which item? If not, justify your answer with a relevant *P-value*.
- Write a sentence (as if for an **Executive Summary**) quantifying the percentage change in earning time for the iPhone, for every 50% increase in Big Mac earning times.

2.1 Question of interest/goal of the study

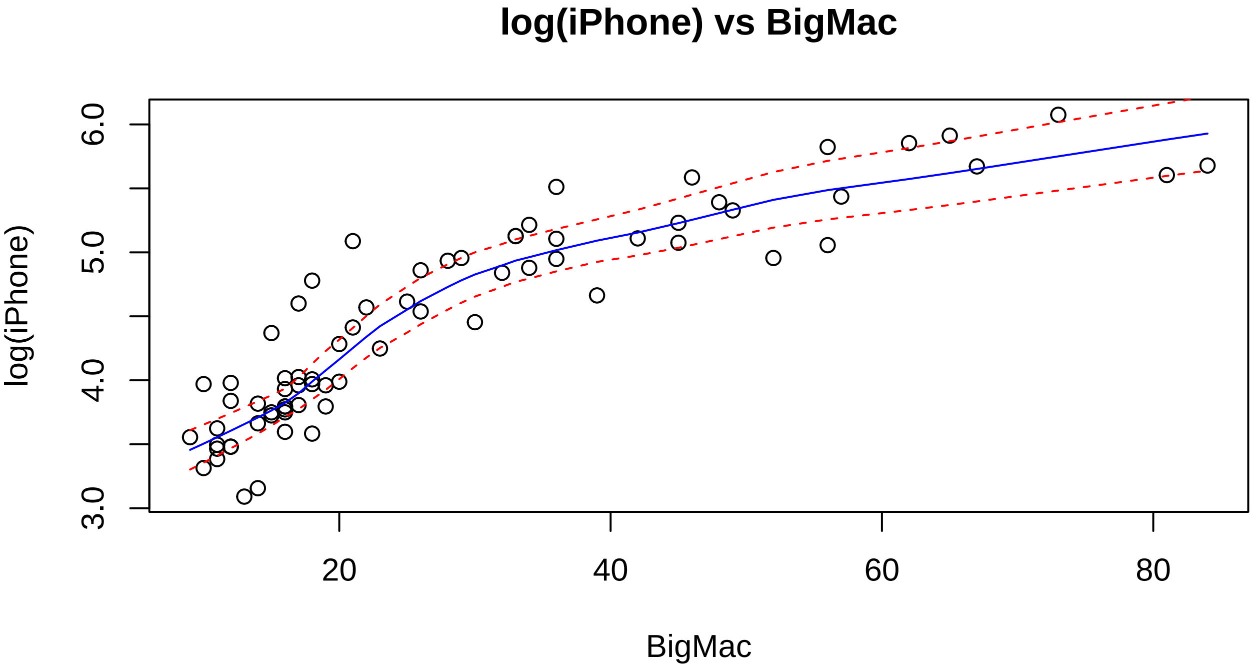
We wish to model the relationship between earning time for iPhones and Big Mac burgers in major cities of the world. In particular, we wish to investigate the percentage change in earning time for the iPhone, for every 50% increase in Big Mac earning time.

2.2 Read in and inspect the data:

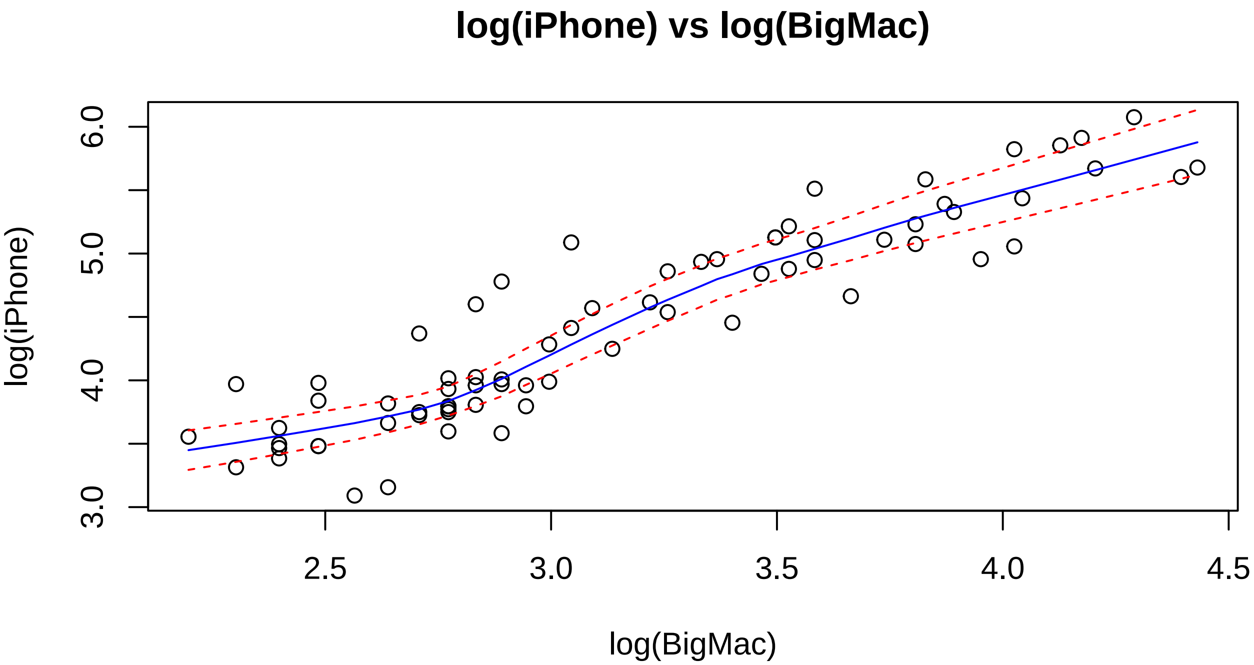
```r
earnings.df=read.csv("EarningTimes.csv",header=T, stringsAsFactors = TRUE)
trendscatter(iPhone~BigMac, main="Time to earn iPhone vs BigMac", data=earnings.df)
```



```r
trendscatter(log(iPhone)~BigMac, main="log(iPhone) vs BigMac", data=earnings.df)
```



```
trendscatter(log(iPhone)~log(BigMac),main="log(iPhone) vs log(BigMac)",data=earnings.df)
```



2.3 Comment on plot and summary statistics.

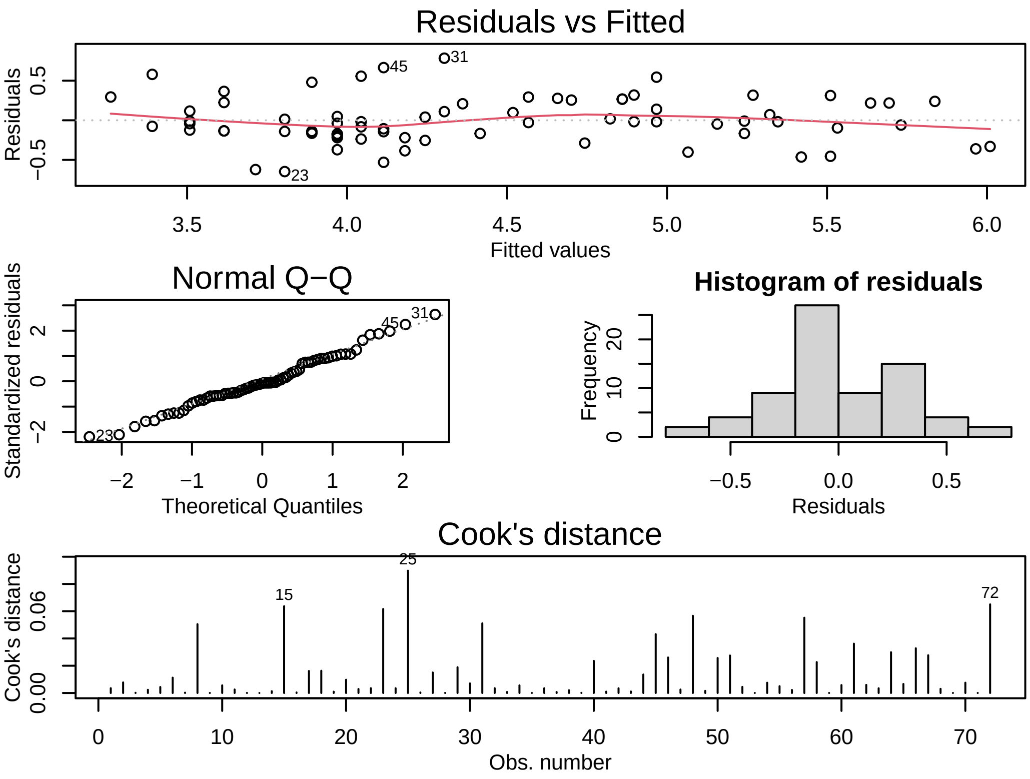
There is an increase in the relationship between the price of a BigMac and Iphone. There is a big spread in variability and the data is right skewed, so log was taken on iPhone. There was still a big spread and right skew so a log was taken on BigMac as well.

2.4 State why a log-log (power) model is appropriate here

The first was log was taken to decrease spread in variability and make it not right skewed. It was still not good so another log to make the graph more linear and not be as skewed and good equality of variance.

2.5 Fit an appropriate linear model and Check Assumptions

```
earnings.fit <- lm(log(iPhone) ~ log(BigMac), data=earnings.df)
modelcheck(earnings.fit)
```



```r
summary(earnings.fit)
```

```
##
## Call:
## lm(formula = log(iPhone) ~ log(BigMac), data = earnings.df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64795 -0.16895 -0.02369  0.21950  0.78371
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.55752    0.19314   2.887  0.00518 **
## log(BigMac)  1.23053    0.06002  20.502  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2991 on 70 degrees of freedom
## Multiple R-squared:  0.8572, Adjusted R-squared:  0.8552
## F-statistic: 420.3 on 1 and 70 DF,  p-value: < 2.2e-16
```

```r
confint(earnings.fit)
```

```
##                 2.5 %    97.5 %
## (Intercept) 0.1723022 0.9427307
## log(BigMac) 1.1108250 1.3502319
```

```r
100*( 0.50*confint(earnings.fit)[2,] - 1 )
```

```
##     2.5 %    97.5 %
## -44.45875 -32.48841
```

```
100*( 1.50*confint(earnings.fit)[2,] - 1 )
```

```
##     2.5 %    97.5 % 
##  66.62375 102.53478
```

```
100*( 0.50^confint(earnings.fit)[2,] - 1 )
```

```
##     2.5 %    97.5 % 
## -53.69708 -60.77710
```

```
100*( 1.50^confint(earnings.fit)[2,] - 1 )
```

```
##    2.5 %   97.5 % 
## 56.89409 72.88767
```

2.6 Write a sentence (as if for an Executive Summary) quantifying the percentage change in earning time for the iPhone, for every 50% increase in Big Mac earning times.

The increase of percentage for every 50% increase is increase of 111.06 % to 135.02% increase and then a 120.32% to 163% increase recurring.

2.7 Do we have evidence that earning time for one of these items (iPhone or BigMac) grows more quickly than that of the other as we range from cheaper to more expensive cities? If so, which item? If not, justify your answer with a relevant *P-value.*

We have very strong evidence that the iPhone increase more quickly than a BigMac as we range from cheaper to more expensive cities. This is shown by a p-value of 2,2*10^16 which is very, very small.

3 Question 3 [19 Marks]

Electricity distribution companies need to keep a close watch on demand patterns during the winter, to ensure that power lines and other infrastructure are not overloaded. A researcher was interested in how electricity demand was affected by temperature and whether this differed between Auckland and Christchurch. He studied the electricity demand for central Auckland and central Christchurch, taking a random sample of days and times from the winter of 2023 and then recording the temperature and electricity use for that day at that time.

The resulting data is in the file *Power3.csv*, which contains the variables:

Variable	Description
Demand	The electric power consumed over an hour (kilowatts),
Temperature	The local air temperature measured at a nearby weather station (degrees Celsius),
City	The city at which the recording was made (either Auckland or Christchurch).

For this question we are particularly interested in:

- Do the effects of temperature on electricity usage differ between the two cities.
- Estimating the effects of a 1 degree change in temperature on electricity demand for each city.

Instructions:

- Comment on plot and summary statistics.
- Fit an appropriate model to the data. Check the model assumptions.
- Comment why we do not need to transform the response in this data.
- Plot the data with your appropriate model superimposed over it.
- Write appropriate Methods and Assumption Checks.
- Does the relationship between power use and temperature depend on the city? Justify your answer, including a relevant *P-value*.
- Write sentences (as if for an **Executive Summary**) quantifying the the estimated effect of a one degree increase in temperature on the amount of power used for **both** Auckland and Christchurch.

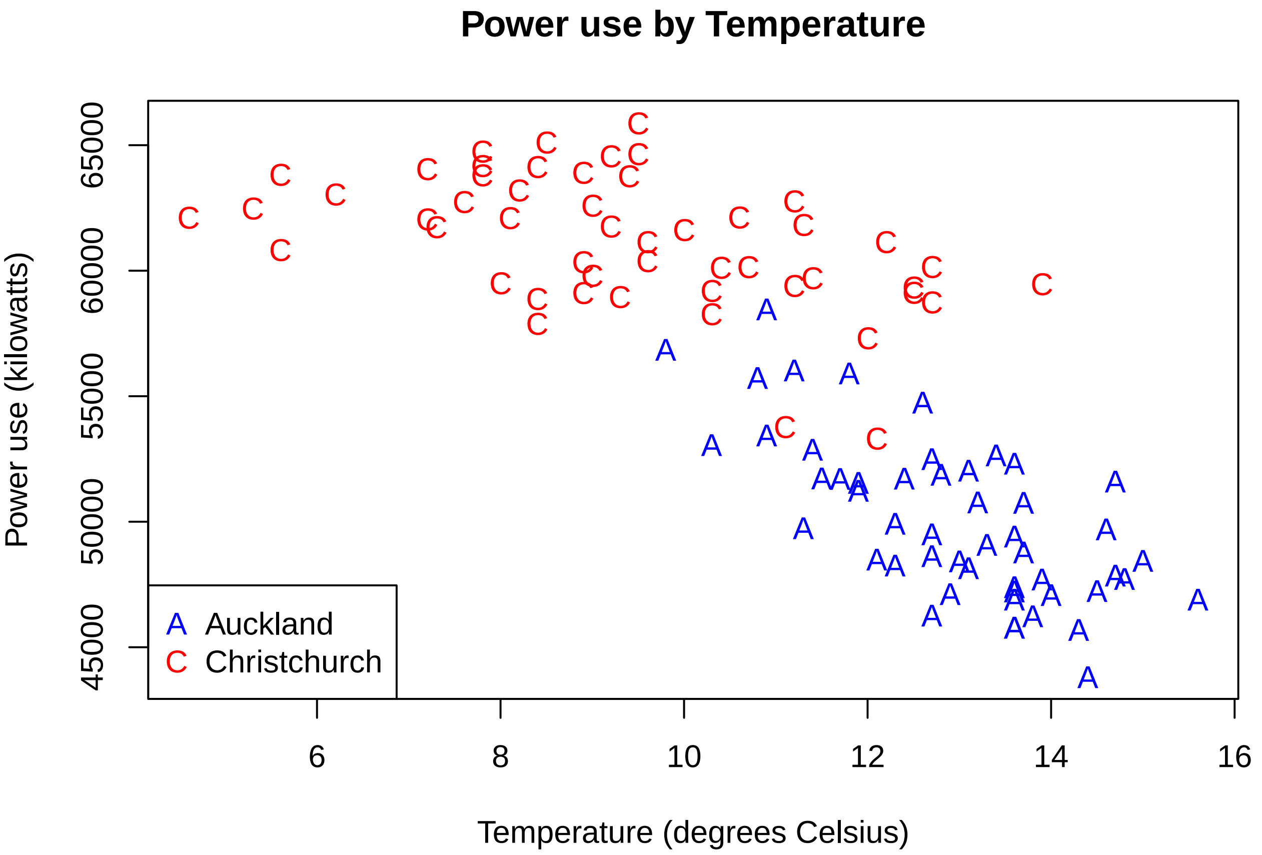
3.1 Question of interest/goal of the study

Do the effects of temperature on electricity usage differ between the Auckland and Christchurch? What are the estimated effects of a 1 degree change in temperature on electricity demand for each city.

3.2 Read in and inspect the data:

```
Power.df=read.csv("Power3.csv",header=T, stringsAsFactors = TRUE)

plot(Demand~Temperature,pch=ifelse(City == "Auckland", 'A', 'C'),
     col=ifelse(City == "Auckland", 'blue', 'red'), main="Power use by Temperature",
     xlab="Temperature (degrees Celsius)",ylab="Power use (kilowatts)",data=Power.df)
legend("bottomleft",pch=c("A","C"),col=c('blue','red'),legend=c("Auckland","Christchurch"))
```

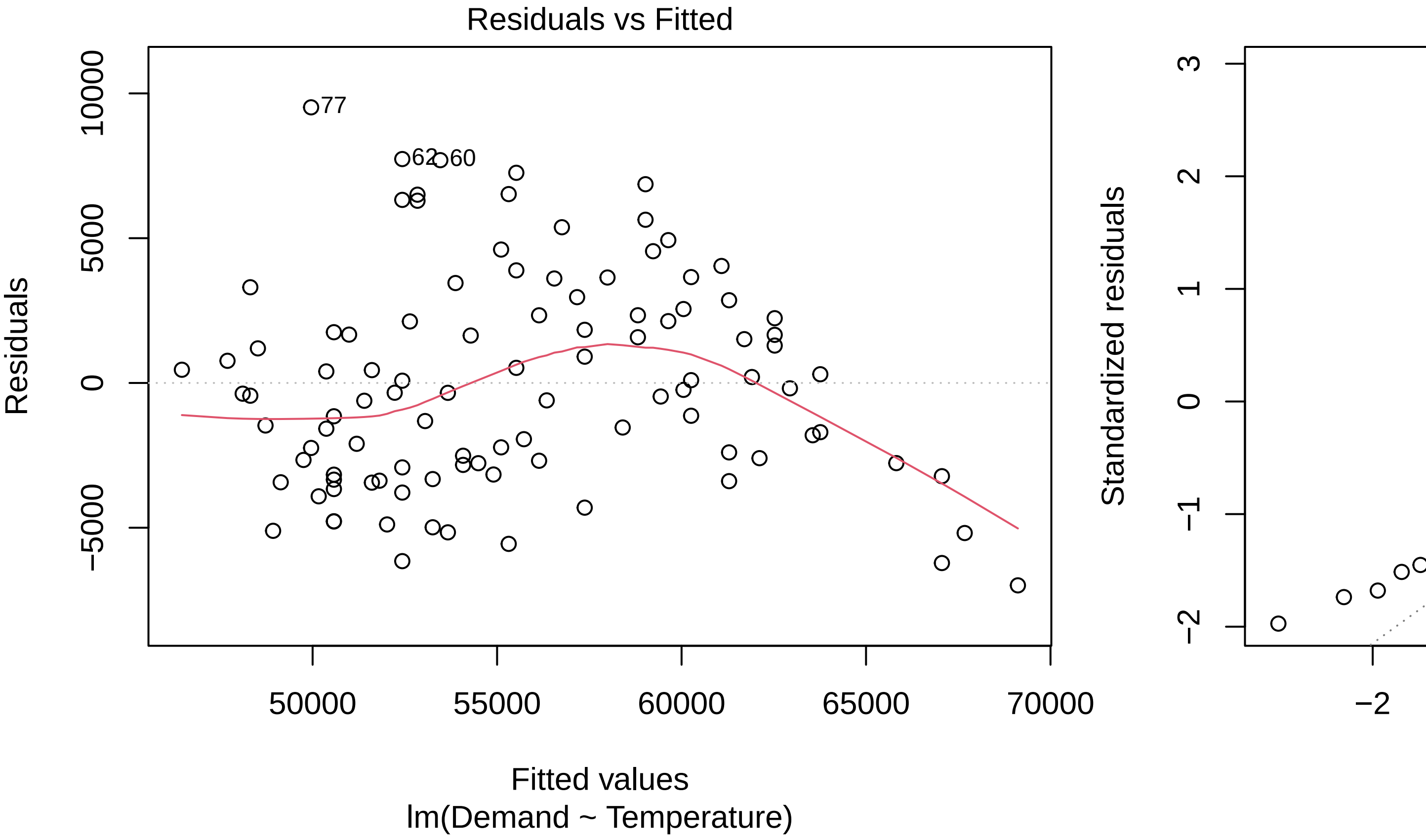


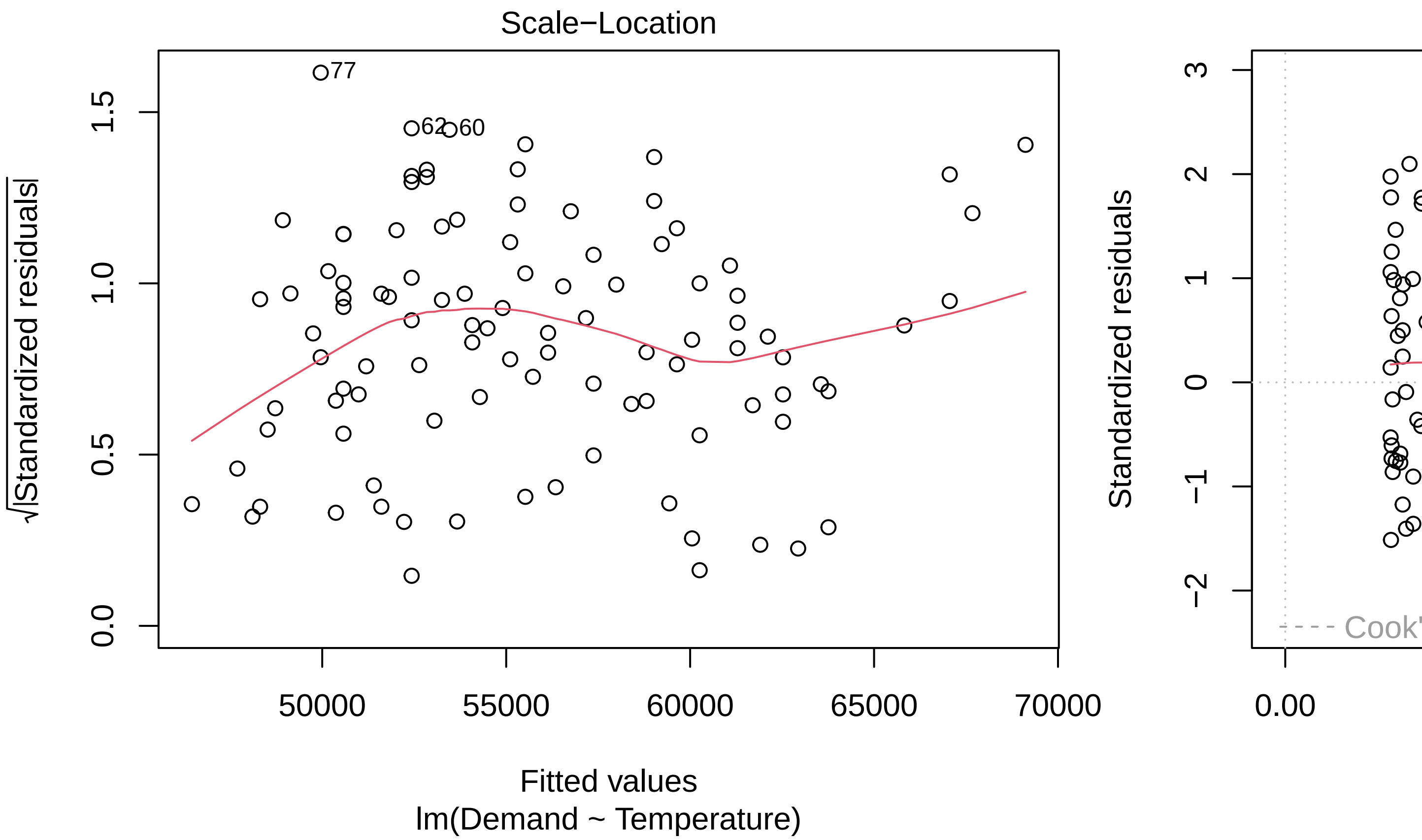
3.3 Comment on plot

Both Aucland and Christchurch are decreasing in temperature with the increase in power use. Auckland is decreasing at a much higher rate than Christhcurch is however.

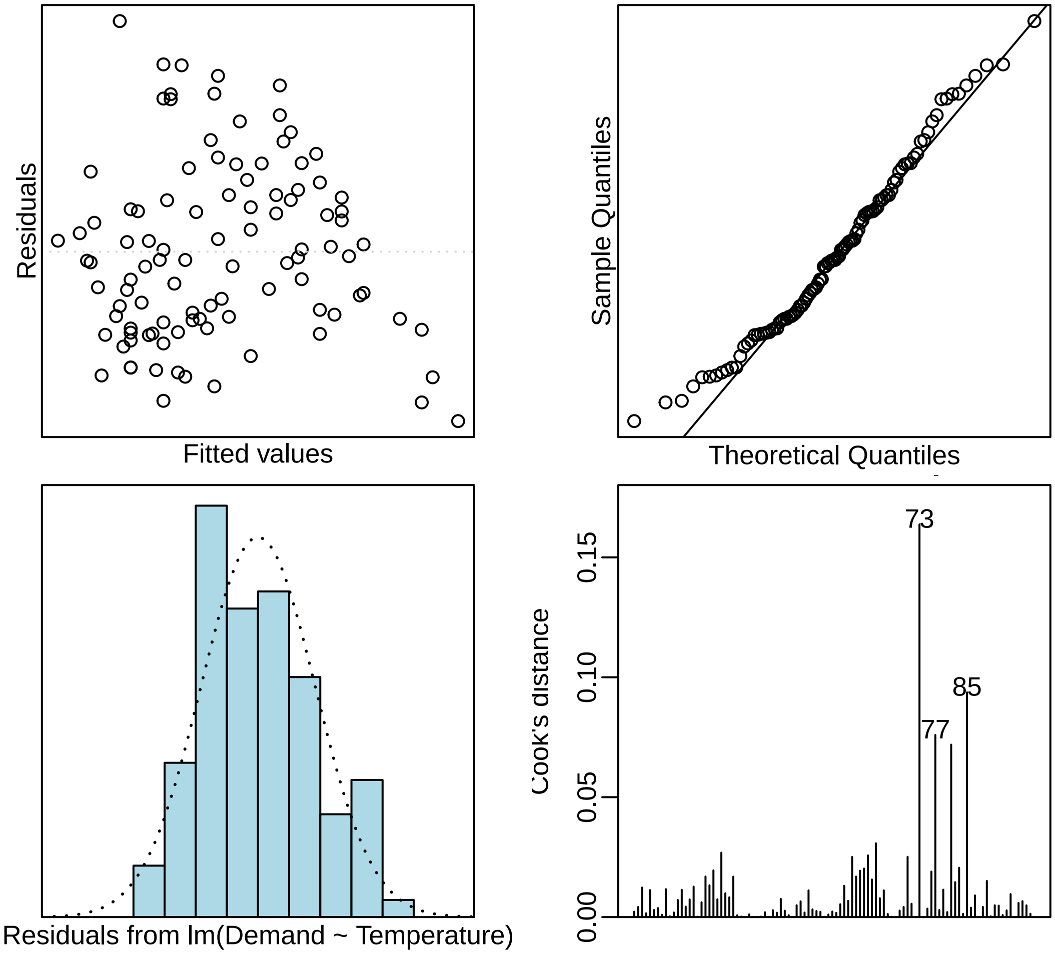
3.4 Fit an appropriate linear model and Check Assumptions

```r
demandTemp.fit=lm(Demand~Temperature,data=Power.df)
plot(demandTemp.fit)
```





```
modcheck(demandTemp.fit)
```



```r
summary(demandTemp.fit)
```

```
##
## Call:
## lm(formula = Demand ~ Temperature, data = Power.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6990.3 -2814.8  -339.1  2311.5  9519.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  78592.1     1676.5   46.88   <2e-16 ***
## Temperature  -2060.0      146.8  -14.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3690 on 100 degrees of freedom
## Multiple R-squared:  0.6631, Adjusted R-squared:  0.6598
## F-statistic: 196.8 on 1 and 100 DF,  p-value: < 2.2e-16
```

```r
confint(demandTemp.fit)
```

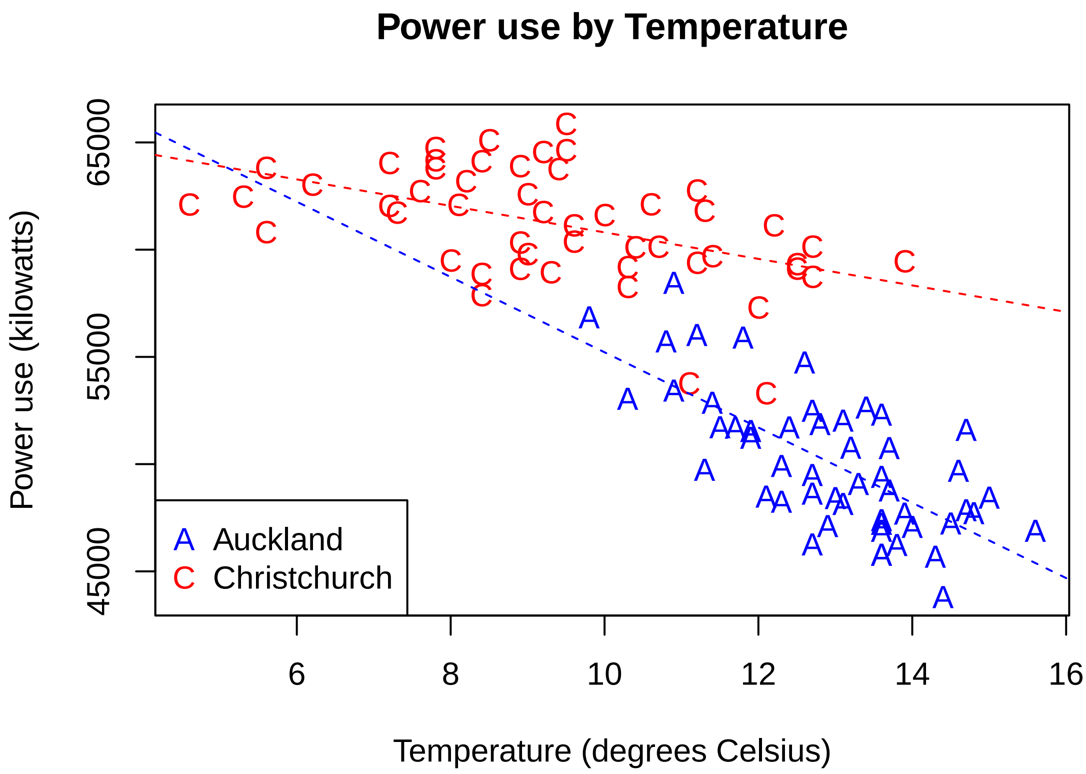
```
##                 2.5 %    97.5 %
## (Intercept) 75265.949 81918.266
## Temperature -2351.254 -1768.663
```

3.5 Comment why we do not need to transform the response in this data.

There is good equality in variance and the data is not right-skewed. The data is also pretty spread out and we can see clearly the difference in change of temperature between the two cities without transformation.

3.6 Plot the data with your appropriate model superimposed over it

```
plot(Demand~Temperature,pch=ifelse(City == "Auckland", 'A', 'C'),
     col=ifelse(City == "Auckland", 'blue', 'red'), main="Power use by Temperature",
     xlab="Temperature (degrees Celsius)",ylab="Power use (kilowatts)",data=Power.df)
legend("bottomleft",pch=c("A","C"),col=c('blue','red'),legend=c("Auckland","Christchurch"))
abline(lm(Demand ~ Temperature, data = subset(Power.df, City == "Auckland")), col = "blue",lty=2)
abline(lm(Demand ~ Temperature, data = subset(Power.df, City == "Christchurch")), col = "red",lty=2)
```



3.7 Method and Assumption Checks

The residuals follow the diagonal line reasonably well, suggesting that the normality assumption is satisfied.

Our final model: $Power_i = \beta_0 + \beta_1 * Temperature_i + \beta_2 * Auckland_i + \beta_3 * Christchurch_i + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$

Our model explains 63% of the variability.

3.8 Does the relationship between power use and temperature depend on the city? Justify your answer, including a relevant ***P-value.***

Yes we have very strong evidence that Auckland has a bigger decrease than Christchurch in Temperature. We can see this as the p-value is very small.

3.9 Write sentences (as if for an Executive Summary) quantifying the the estimated effect of a one degree increase in temperature on the amount of power used for both Auckland and Christchurch.

In this study, we examined how electricity demand in Auckland and Christchurch was affected by temperature during the winter of 2023.

A 1-degree increase in Auckland is approximately 1971 kW while in Christurch it is approximately 606kw at a 95% cofindence interval.