STATS 201/8 Assignment 2

Anish Hota ahot228

Due Date: 3pm, Tuesday 6th August 2024

1 Question 1 [10 Marks]

A lecturer is interested in experiments where the "Wisdom of the Crowd" is put to the test. In these cases, combined answers from a large group usually proves better than the individuals. A simple example is to get people to guess the number of jelly beans in a jar. Individual guesses vary widely and are usually all wrong, but the average can be quite close to the actual answer.

An internet experiment was carried out where people where asked to guess the weight of a cow (in pounds) that was pictured. The actual weight of the cow in the picture was 1355 pounds. A large number of people responded with their estimated weight. For this question, we will use a random subset of the respondents.

We will treat the random sample of guesses as representative of group wisdom. What we are interested in are is was the group wisdom consistent with the actual weight of the cow or was there evidence that it differed? If there was evidence that the group wisdom estimate differed, what was the group wisdom estimate? (Remember, we want an interval for the estimate.)

The data on the 500 guesses of the cows weight are in the file *Cow.csv*, which contains the variable:

Variable	Description
Weight	the guess of the weight of the cow (in pounds)

Instructions:

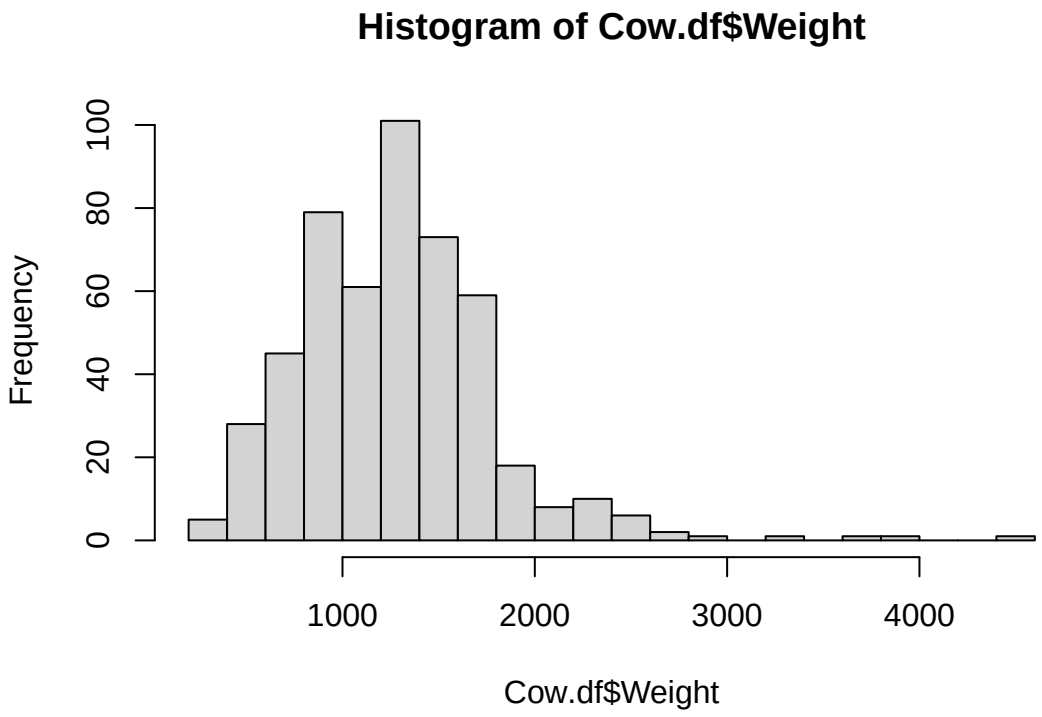
- Make sure you change your name and UPI/ID number at the top of the assignment.
- Comment on the plot of the data.
- One concern that the lecturer has is that, since the respondents were self-selected from an internet website with readers from all around the world, some would be more used to estimating weights in kilograms rather than pounds and tend to guess low. Looking at the plot, comment if there seem to be any sign of this?
- Manually calculate the t-statistic for comparing the mean height to 1355 pounds and the corresponding 95% confidence interval.
- Write an appropriate **Executive Summary**.

1.1 Question of interest/goal of the study

We are interested in seeing if group wisdom can guess the average weight of a cow that is shown in a picture on the internet.

1.2 Read in and inspect the data:

```
Cow.df=read.csv("Cow.csv", header=T)
hist(Cow.df$Weight,breaks=20)
```




```
summary(Cow.df$Weight)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     300     950    1276    1286    1532    4450
```

1.3 Comment on the plot/exploratory data analysis

The plot appears to be right skewed with heavy frequencies around the 1000 kg mark with a few outliers around the 3000kg - 4000kg mark.

1.4 One concern that the lecturer has is that, since the respondents were self-selected from an internet website with readers from all around the world, some would be more used to estimating weights in kilograms rather than pounds and tend to guess low. Looking at the plot, comment if there seem to be any sign of this?

There seems to be some sign of this as most of the plots are quite low and they may have thought it was in kgs as it is lower in pounds. The weigth of the cow would be higher in pounds but people who are used to kg would think its lower than usual.

1.5 Manually calculate the t-statistic for testing if the underlying mean is 1355, and the 95% confidence interval for the mean.

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

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

```
# t-statistic for H0: mu=1355 :
y = (Cow.df$Weight)
```

```
(t_stat = (mean(y)-1355)/(sd(y)/sqrt(length(y))))
```

```
## [1] -3.026792
```

```
# 95% confidence interval for the mean:

(confidence_interval_1 = mean(y)-qt(0.975,length(y)-1)*(sd(y)/sqrt(length(y))))
```

```
## [1] 1241.15
```

```
(confidence_interval_2 = mean(y)+qt(0.975,length(y)-1)*(sd(y)/sqrt(length(y))))
```

```
## [1] 1330.776
```

1.6 Repeat the same calculation using the t.test function (done for you):

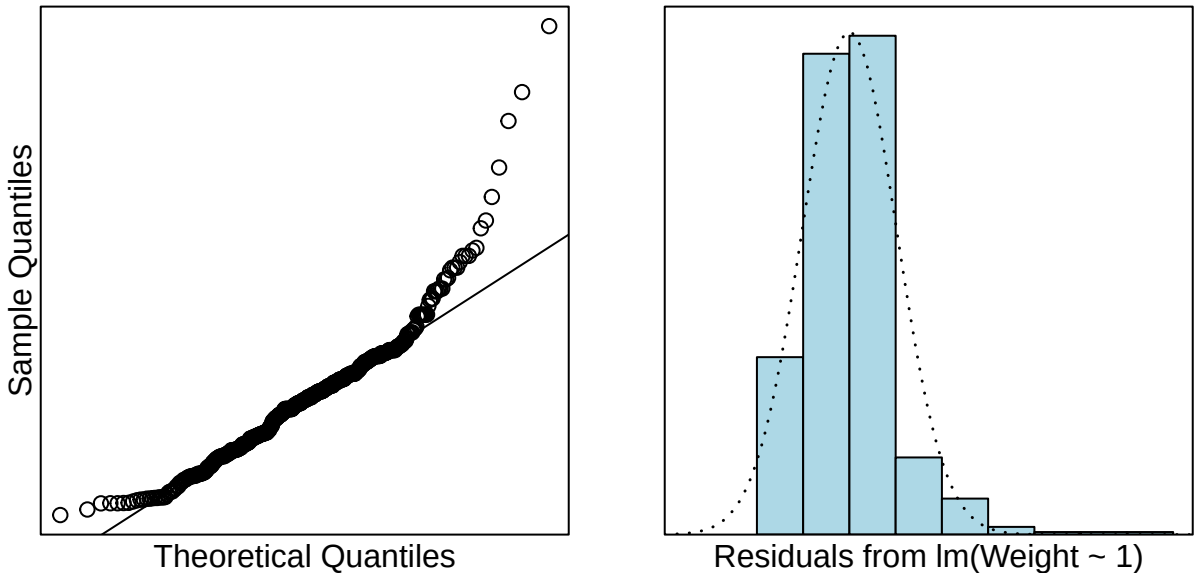
```
t.test(Cow.df$Weight, mu=1355)
```

```
##
##  One Sample t-test
##
## data:  Cow.df$Weight
## t = -3.0268, df = 499, p-value = 0.002599
## alternative hypothesis: true mean is not equal to 1355
## 95 percent confidence interval:
##  1241.150 1330.776
## sample estimates:
## mean of x
##  1285.963
```

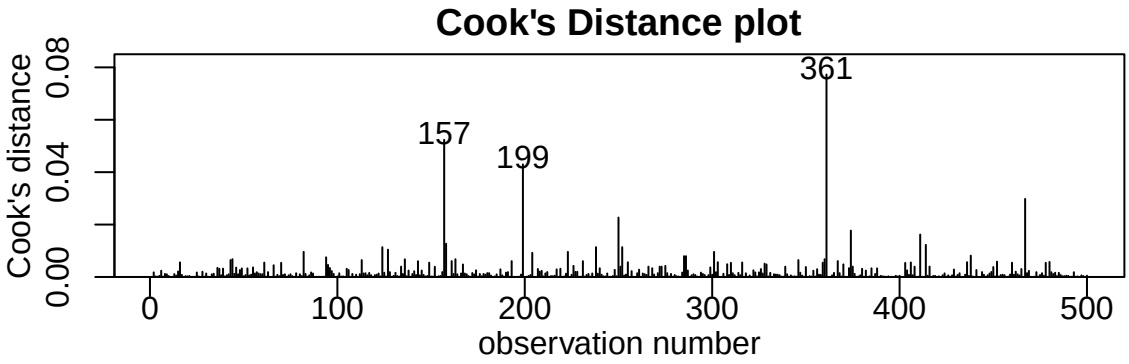
Note: You should get exactly the same results from the manual calculations and using the *t.test* function. Doing this was to give you practice using some R code. The *t.test* function also delivers the p-value that we did not calculate above.

1.7 Fit and check the null model (done for you):

```
Cow.fit=lm(Weight~1,data=Cow.df)
normcheck(Cow.fit)
```




```
cooks20x(Cow.fit)
```




```
summary(Cow.fit);
```

```
##
## Call:
## lm(formula = Weight ~ 1, data = Cow.df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -986.0 -336.0   -9.5  246.0 3164.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1285.96      22.81   56.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 510 on 499 degrees of freedom
```

```
confint(Cow.fit)
```

```
##               2.5 %   97.5 %
## (Intercept) 1241.15 1330.776
```

1.8 Method and Assumption Checks

As this data consists of 500 guesses (of a cows weight). 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 500 guesses, and we wished to see if their average guess is consistent with the actual cows weight of 1355 pounds. The guesses should be independent of each other. Though the data is skewed, we are happy with the normality assumption (see answer to previous question). There were no unduly influential points.

Our model is: $Weight_i = \mu + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$

1.9 Executive Summary

We are interested in whether the group wisdom was consistent with the actual weight of the cow or was there evidence that it differed. We estimate that the cow weight they the group of wisdom predicted somewhere between 1241 pounds and 1331 pounds. We have strong evidence that the estimation differs to the actual weight of the cow as 1355 pounds is outside the 95% confidence interval (P-value = 0.0026).

---

2 Question 2 [16 Marks]

A manufacturer of electric bikes wants to investigate how power consumption of their model of bike increases with speed. 100 independent measurements of speed vs power consumption were recorded for their bikes.

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

Variable	Description
kph	Speed (kilometres per hour)
watts	Power consumption (watts)

Instructions:

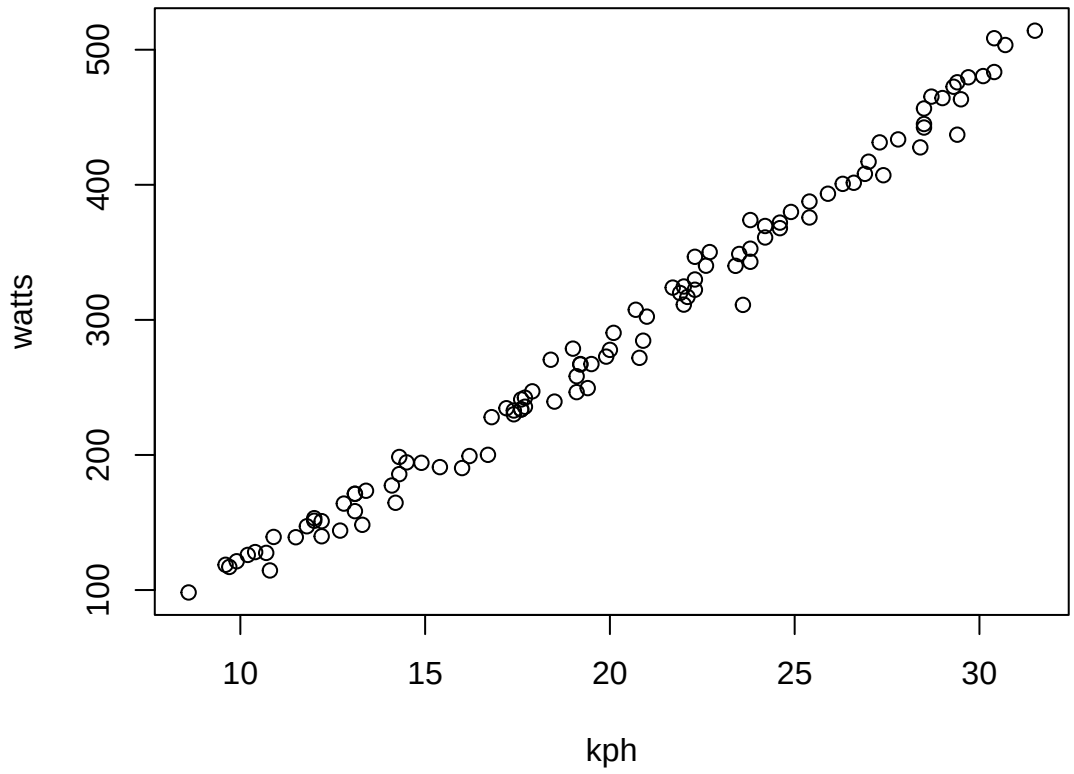
- Comment on the initial plot of the data.
- Fit *BOTH* a simple linear model and a second linear model with an appropriate quadratic term, including model checks.
- Create predicted power consumptions for *BOTH* models for kph ranging from 9 to 32 kph (changing in steps of 1 kph.) You do not need to list these.
- Create and list the differences between the two sets of predictions.
- Plot the data with *BOTH* models superimposed over it.
- Which of the two models is the most correct according to the assumptions and would be the most accurate for prediction? Justify your answer.
- Write the equation of the model you chose above as if for Methods and Assumption Checks.
- In one sentence, explain the relationship between speed and power consumption from the simple linear model.
- Describe how the quadratic model further adjusts this relationship from the simple linear relationship. (You will find looking at your plot and the list of differences helpful here.)
- If you wanted to just have a simple rule of thumb description for the relationship, which model would you use? Justify your answer.

2.1 Question of interest/goal of the study

We are interested in investigating how the power consumption of an electric bike changes with the bikes speed.

2.2 Read in and inspect the data:

```
Ebike.df=read.csv("CyclePower.csv")
plot(watts~kph,data=Ebike.df)
```

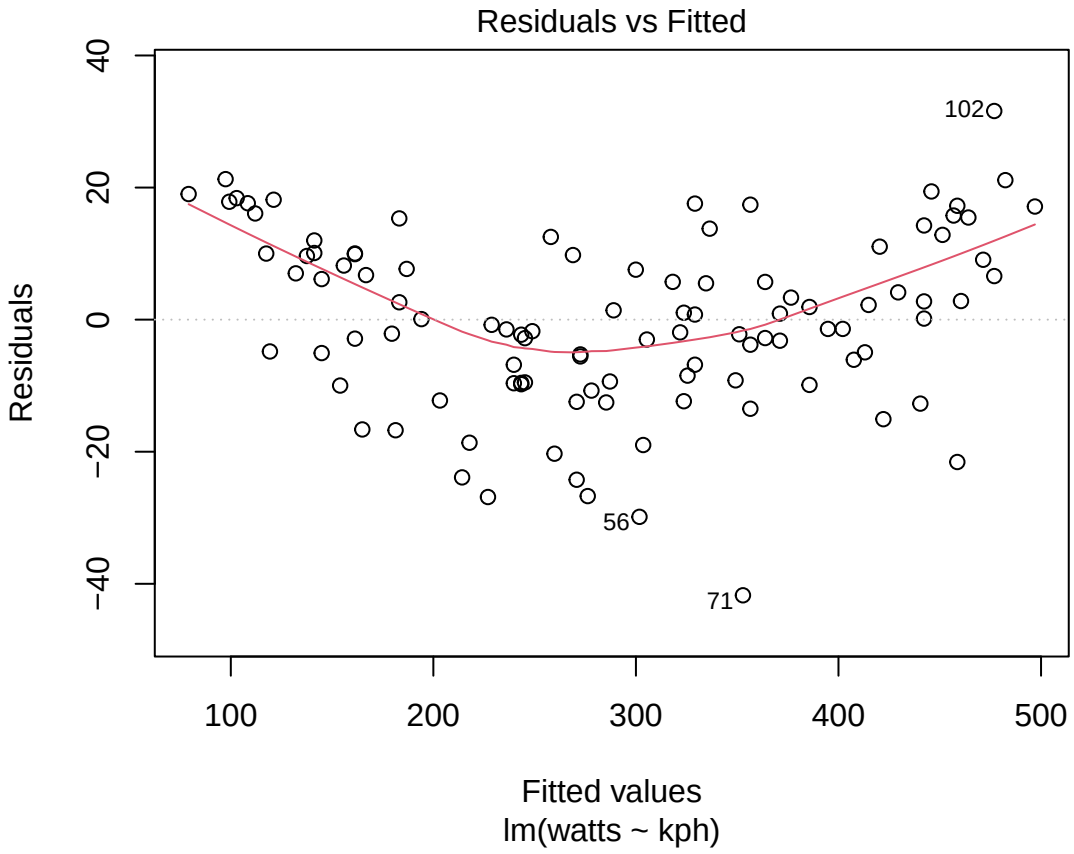


2.3 Comment on the plot

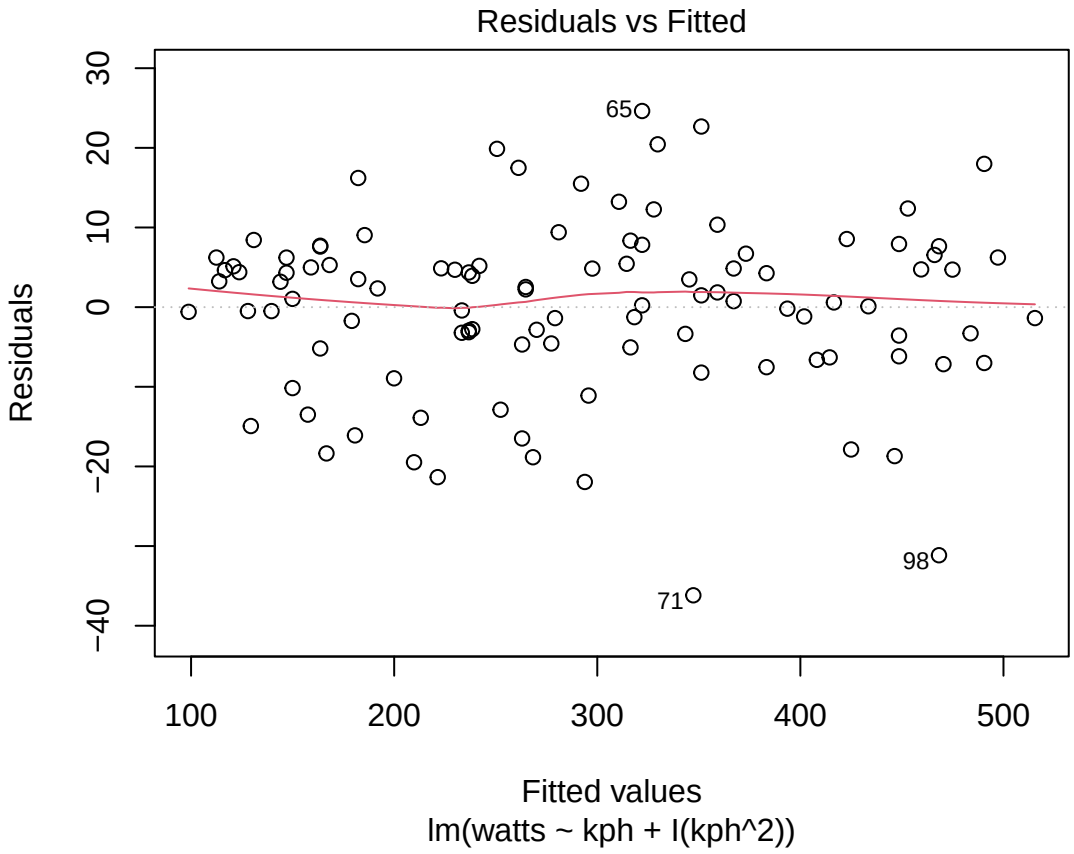
We are seeing an increasing relationship between speed and power consumption. This relationship seems toe be very strong with a very linear scatter. Meaning there is near to none curvature in this plot.

2.4 Fit a linear model with an appropriate quadratic term, including model checks.

```
## Fitting the simple linear model first.
cyclepower.fit = lm(watts~kph, data = Ebike.df)
plot(cyclepower.fit, which = 1)
```




```
## Fit a quadratic relationship.
cyclepower.fit2 = lm(watts~kph + I(kph^2), data = Ebike.df)
plot(cyclepower.fit2, which = 1)
```



3 Prediction of power consumption ranging from 9 to 32 kph for BOTH models and find differences.

```
Pred.df=data.frame(kph=9:32)
(y =predict(cyclepower.fit, Pred.df))
```

```
##         1         2         3         4         5         6         7         8
##  86.47943 104.72360 122.96778 141.21195 159.45612 177.70029 195.94446 214.18863
##         9        10        11        12        13        14        15        16
## 232.43280 250.67698 268.92115 287.16532 305.40949 323.65366 341.89783 360.14200
##        17        18        19        20        21        22        23        24
## 378.38618 396.63035 414.87452 433.11869 451.36286 469.60703 487.85120 506.09538
```

```
(x =predict(cyclepower.fit2, Pred.df))
```

```
##        1        2        3        4        5        6        7        8
## 104.2155 118.0582 132.3135 146.9814 162.0620 177.5551 193.4608 209.7791
##        9       10       11       12       13       14       15       16
## 226.5100 243.6535 261.2097 279.1784 297.5597 316.3537 335.5602 355.1793
##       17       18       19       20       21       22       23       24
## 375.2110 395.6554 416.5123 437.7819 459.4640 481.5587 504.0661 526.9860
```

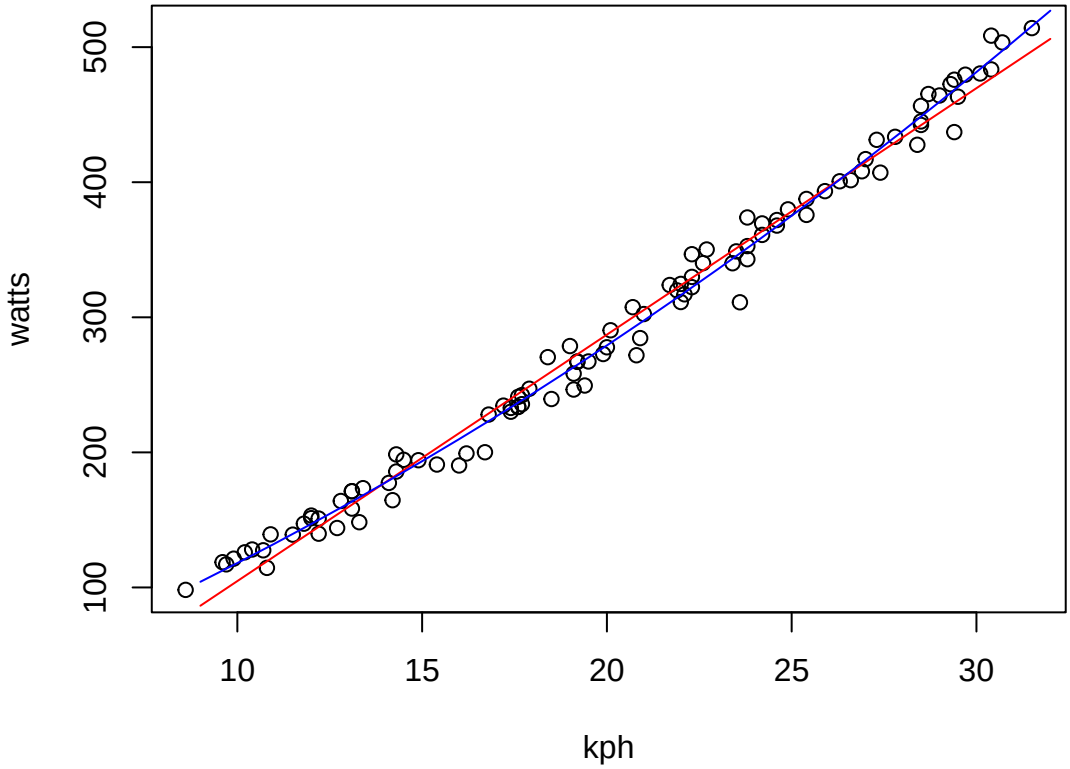
```
(y-x)
```

```
##           1           2           3           4           5           6
## -17.7360920 -13.3346268  -9.3457635  -5.7695020  -2.6058424   0.1452153
```

```
##           7           8           9          10          11          12
##   2.4836712   4.4095252   5.9227773   7.0234276   7.7114760   7.9869226
##          13          14          15          16          17          18
##   7.8497673   7.3000101   6.3376511   4.9626902   3.1751275   0.9749628
##          19          20          21          22          23          24
##  -1.6378036  -4.6631720  -8.1011421 -11.9517142 -16.2148881 -20.8906639
```

3.1 Plot the data with BOTH model superimposed over it.

```
# Have already predicted values over a range for the model above so can use the lines command to add th
plot(watts~kph,data=Ebike.df)
x=9:32
lines(x, predict(cyclepower.fit, Pred.df), col = "red")
lines(x, predict(cyclepower.fit2, Pred.df), col = "blue")
```



3.2 Which of the two models is the most correct according to the assumptions and would be the most accurate for prediction? Justify your answer.

The quadratic model is better (the blue line). This because it coresponds to data more and is more accurate with the plots. The linear model is also quite accurate but the quadratic model is more accurate.

3.3 Write the equation of the model you chose above as if for Methods and Assumption Checks.

$PowerConsumption = \beta_0 + \beta_1 * Speed_i + \beta_2 * Speed_i^2 + \epsilon_i$ where $\epsilon_i \sim iid\ N(0, \sigma^2)$

3.4 In one sentence, explain the relationship between speed and power consumption from the simple linear model.

There is an increasing linear relationship between speed and power consumption, as kph increase so does the watts.

3.5 Describe how the quadratic model further adjusts this relationship from the simple linear relationship. (You will find looking at your plot and the list of differences helpful here.)

It shows how there is a bigger increase in some areas than others. The middle has a bigger increase than the ends. The curvature allows us to better understand the relaionship between the two with more detail on the rate of increase.

3.6 If you wanted to just have a simple rule of thumb description for the relationship, which model would you use? Justify your answer.

The simple linear model as it is much easier to determine and plot. It also is a much easier equation to deal with and still is quite accurate in determining the relative increasing relationship between the power consumption and speed.