Assignment One Solutions

Question One

1.

```
> setwd("F:/STAT2008/STAT2008/2013/Assignments")

> rugby<-read.csv("rugby.csv",header=T)

> attach(rugby)

> rugby.lm<-lm(Attendance~Temp)

> summary(rugby.lm)


Call:

lm(formula = Attendance ~ Temp)


Residuals:

   Min      1Q  Median      3Q     Max

-4.7740 -2.6775 -0.2893  2.2336 13.4458


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  7.76642    0.53706   14.46   <2e-16 ***

Temp         1.40152    0.04119   34.03   <2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 3.335 on 98 degrees of freedom

Multiple R-squared: 0.922,      Adjusted R-squared: 0.9212

F-statistic:  1158 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
> coeffdet<-cor(Attendance,Temp)^2

> coeffdet

[1] 0.9219658
```

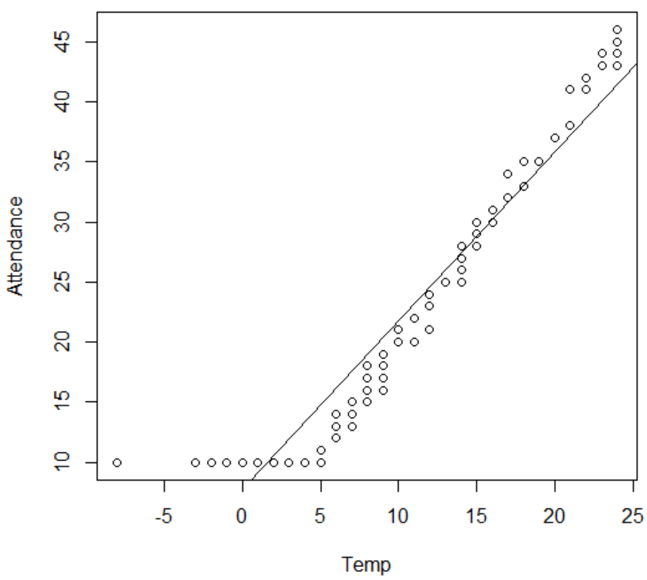
The coefficient of determination is therefore 92.1966%. 92.1966% of the variation in attendance is explained by the variation in temperature.

The fitted regression line is:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = 7.7664 + 1.4015 X_i$$
$$Attendance = 7.7664 + 1.4015 Temp$$

2.

```
>plot(Temp,Attendance)

> abline(rugby.lm)
```



3.

```
> cor.test(Attendance,Temp)

        Pearson's product-moment correlation

data:  Attendance and Temp
t = 34.0273, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9413006 0.9730858
sample estimates:
      cor
0.9601905

> anova(rugby.lm)
Analysis of Variance Table

Response: Attendance
          Df  Sum Sq Mean Sq F value    Pr(>F)
Temp       1 12880.0 12880.0  1157.9 < 2.2e-16 ***
Residuals 98  1090.2    11.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Test One

$$H_0 : \rho = 0$$
$$H_A : \rho \neq 0$$

Test Statistic = T=34.0273

Compare to the T distribution on 98 degrees of freedom. We will reject the Null if P(|T|>34.0273) is less than alpha.

SKETCH

Since the p-value is small, (p-value < 2.2e-16) then we reject the null and conclude there is a significant linear relationship between temperature and attendance.

Test Two

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

Test Statistic = F=1157.9

Compare to the F distribution on 1 numerator and 98 denominator degrees of freedom. We will reject the Null if P(F>1157.9) is less than alpha.

SKETCH

Since the p-value is small, (p-value < 2.2e-16) then we reject the null and conclude there is a significant linear relationship between temperature and attendance.

Test Three

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

Test Statistic = T= $\pm\sqrt{1157.9} = \pm 34.0279$

Compare to the T distribution on 98 degrees of freedom. We will reject the Null if P(|T|>34.0279) is less than alpha.

$$\{t_{n-2}(1-\alpha/2)\}^2 = F_{1,n-2}(1-\alpha)$$

Since the p-value is small, (p-value < 2.2e-16) then we reject the null and conclude there is a significant linear relationship between temperature and attendance.

4.

$$H_0 : \beta_0 = 0$$
$$H_A : \beta_0 \neq 0$$

Test Statistic = T=14.46

Compare to the T distribution on 98 degrees of freedom. We will reject the Null if P(|T|>14.46) is less than alpha, say 5%.

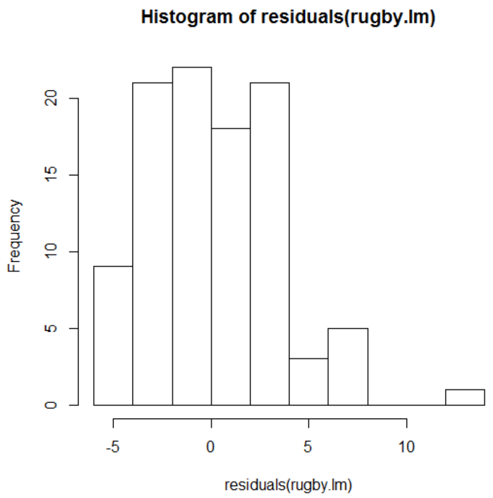
Since the p-value is small, (p-value < 2.2e-16) then we reject the null and conclude that the intercept is significant.

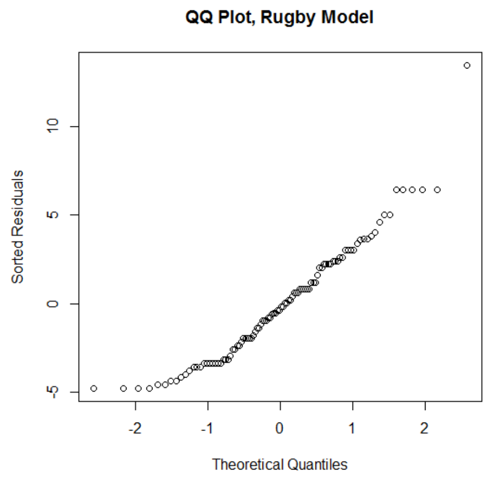
Given the scatter plot, it doesn't seem sensible for (0,0) to be reflected in a real-world model for attendance and temperature. In fact, the scatterplot indicates that there is a different relationship between attendance and temperature for temperatures below 5 degrees. We may be able to fit a superior model by re-fitting a linear regression model only for temperatures above 5.

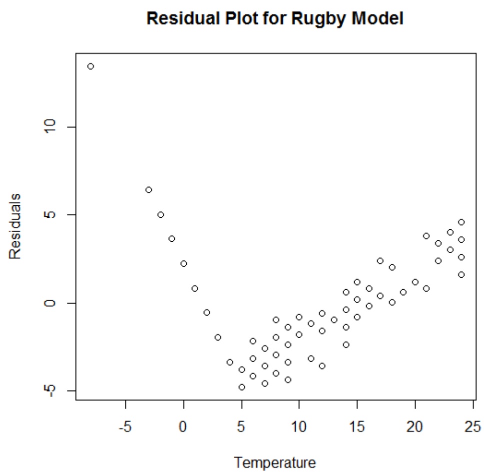
Additionally, given the small p-value in the hypothesis test above, a model through the origin seems inappropriate.

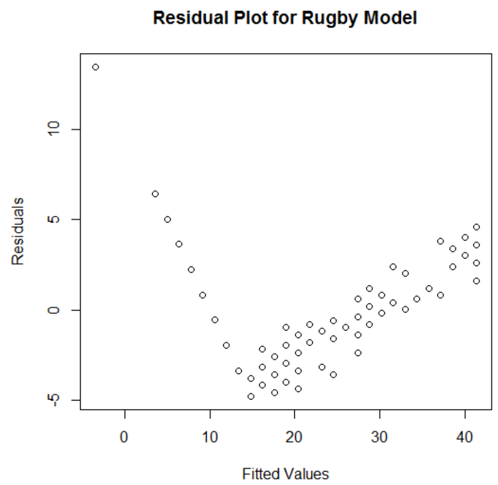
5.

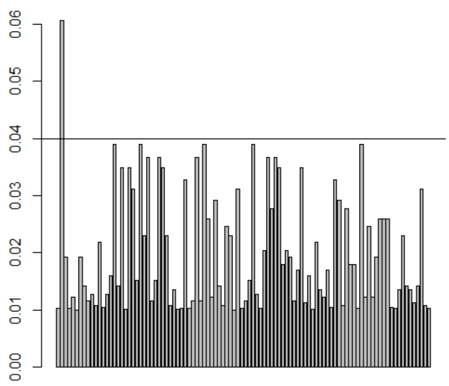
```
> plot(fitted( rugby.lm),residuals( rugby.lm),xlab="Fitted Values", ylab="Residuals",main="Residual Plot for Rugby Model")

>plot(Temp,residuals( rugby.lm),xlab="Temperature", ylab="Residuals",main="Residual Plot for Rugby Model")

> hist(residuals(rugby.lm))

> qqnorm(residuals( rugby.lm),ylab="Sorted Residuals",main="QQ Plot, Rugby Model")

>  barplot(hat( Temp))

>  abline(h=4/length( Temp))
```


Histogram of residuals(rugby.lm)
Frequency
residuals(rugby.lm)


QQ Plot, Rugby Model
Sorted Residuals
Theoretical Quantiles


Residual Plot for Rugby Model
Residuals
Temperature


Residual Plot for Rugby Model
Residuals
Fitted Values



Comments:

The assumptions of linear regression are (from lectures):

- In our fitting we assume the errors have a particular distribution – that is, $\varepsilon \sim N(0,\sigma_\varepsilon^2)$
    - Normal distribution
    - Mean = 0
    - Constant variance = σ_ε^2
        - If σ_ε^2 is small, then small spread of observations around fitted line
        - If σ_ε^2 is large, then observations have wide spread around fitted line
    - Errors are independent

The qq-plot and histogram show a violation of the normality assumption. Both plots indicate the residuals are positively skewed.

The versus fitted values and versus predictors show a violation of the independence and constant variance assumption. Both plots show a series of increasing and decreasing residuals and don't appear to be random. There appears to be a linear relationship between the residuals and the predictor and the fitted values. Variation of the residuals appears to be non-constant, the size of the residual (and therefore the spread around zero) appears to be a function of the fitted value/temperature.

There is one point with very high leverage which has the potential to be influential as identified by the barplot of leverages. The scatterplot also indicates that the fitted line may be overly influenced by this point.

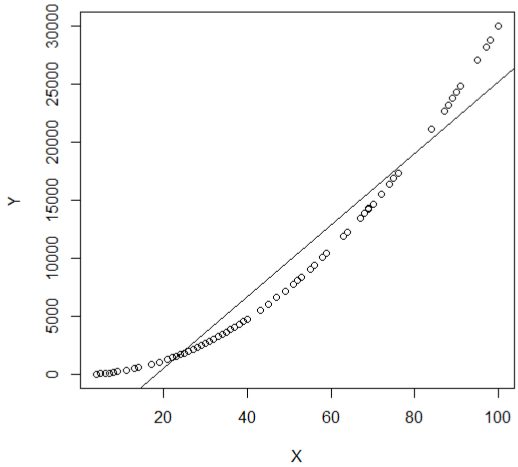
The assumptions of linear regression have been violated (normality, independence and constant variance). There also appears to be a possible influential observation.

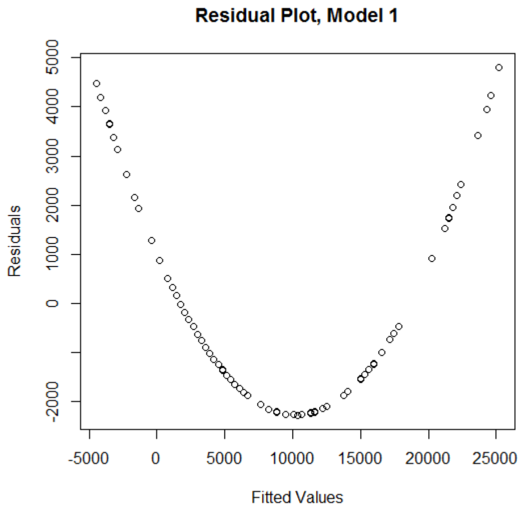
Question Two

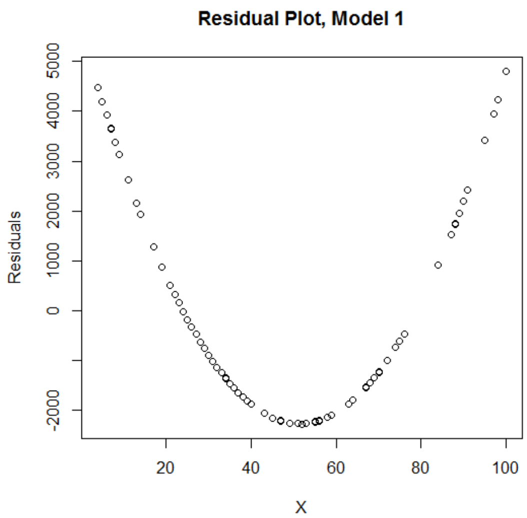
1.

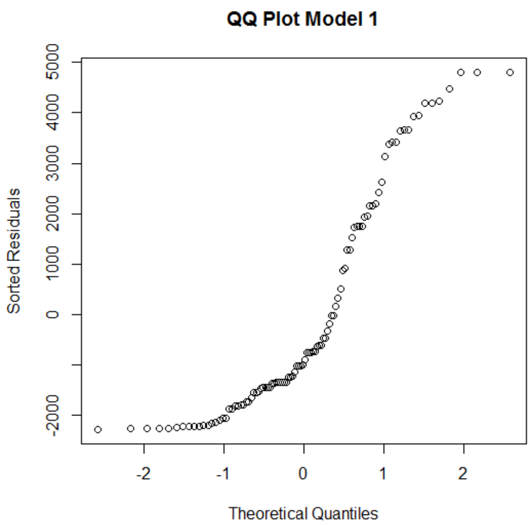
```
> assign1<-read.csv("ass1q2.csv",header=T)

> attach(assign1)

> names(assign1)

[1] "X" "Y"

> reg1<-lm(Y~X)

> summary(reg1)
```

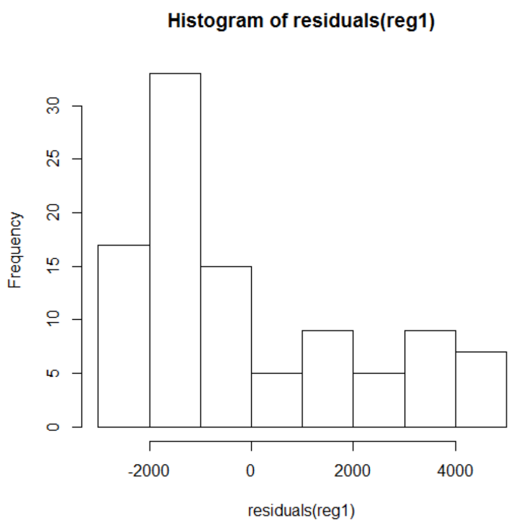
```
Call:
lm(formula = Y ~ X)

Residuals:
   Min      1Q  Median      3Q     Max
-2277.1 -1663.2  -951.9  1739.8  4801.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5660.4      448.9  -12.61   <2e-16 ***
X              308.6        8.0   38.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2192 on 98 degrees of freedom
Multiple R-squared: 0.9382,     Adjusted R-squared: 0.9376
F-statistic:  1488 on 1 and 98 DF,  p-value: < 2.2e-16

> plot(X,Y)
> abline(reg1)
> plot(fitted(reg1),residuals(reg1),xlab= "Fitted Values",ylab="Residuals",main="Residual Plot, Model 1")
> plot(X,residuals(reg1),xlab= "X",ylab="Residuals",main="Residual Plot, Model 1")
> qqnorm(residuals(reg1),ylab="Sorted Residuals",main="QQ Plot Model 1")
> hist(residuals(reg1))
> barplot(hat(X))
> abline(h=4/length(X))
```

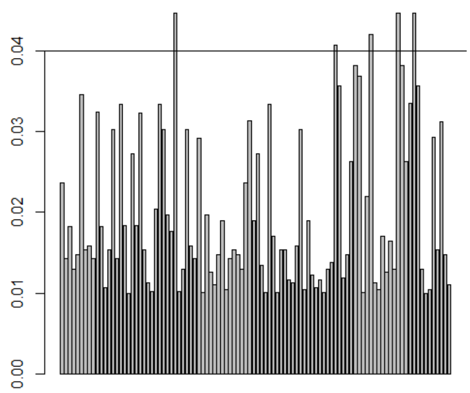

Y
0
5000
10000
15000
20000
25000
30000
20
40
60
80
100
X


Residual Plot, Model 1
Residuals
-2000
0
1000
2000
3000
4000
5000
-5000
0
5000
10000
15000
20000
25000
Fitted Values


Residual Plot, Model 1
Residuals
-2000
0
1000
2000
3000
4000
5000
20
40
60
80
100
X


QQ Plot Model 1
Sorted Residuals
-2000
0
1000
2000
3000
4000
5000
-2
-1
0
1
2
Theoretical Quantiles


Histogram of residuals(reg1)
Frequency
0
5
10
15
20
25
30
-2000
0
2000
4000
residuals(reg1)


0.00
0.01
0.02
0.03
0.04

The fitted model is: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \text{-}5660.4 + 308.6 X_i$

The residuals vs fitted values show a violation of the independence assumption as there appears to be a relationship between the residuals and the fitted values. The constant variation assumption seems to be plausible but there are bigger residuals for very small fitted values and the larger fitted values. Overall, the assumption of constant variance is acceptable.

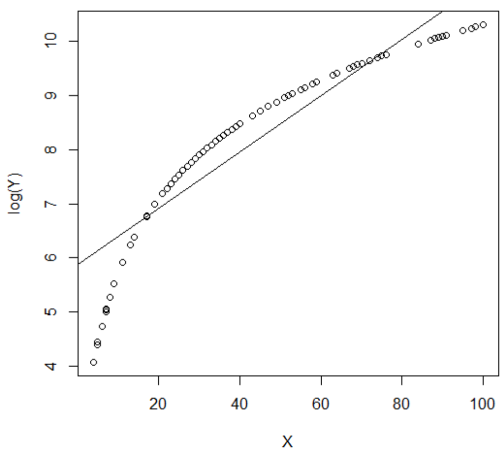
The qq plot and histogram show a violation of the assumption of normality as both show positive skew.

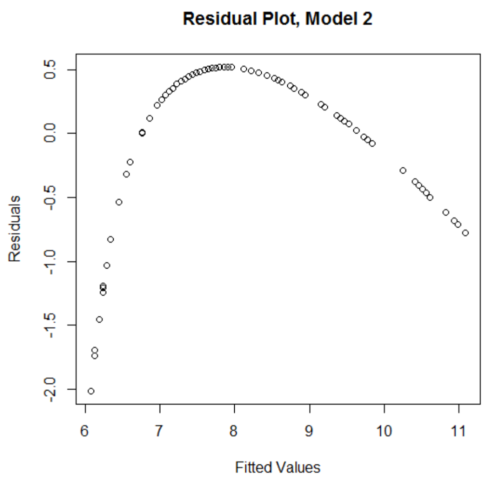
There are a few points with high leverage, but none appear to be a real issue in terms of influence.

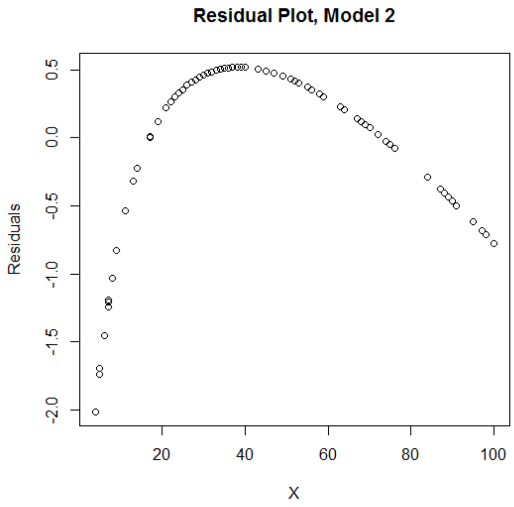
2.

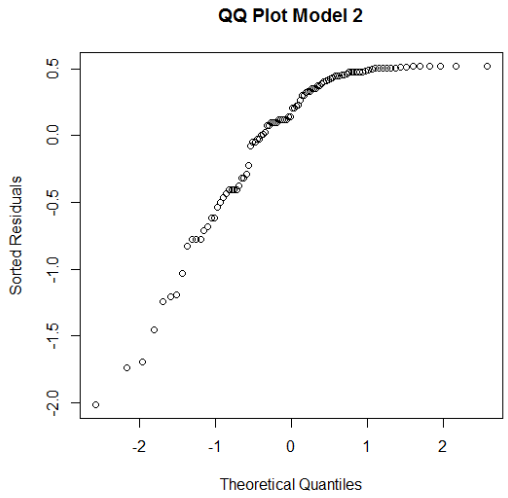
```
> reg2<-lm(log(Y)~X)

> summary(reg2)

Call:

lm(formula = log(Y) ~ X)


Residuals:

    Min      1Q  Median      3Q     Max

-2.0128 -0.3344  0.1745  0.4539  0.5202



Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.876032   0.119659   49.11   <2e-16 ***

X           0.052098   0.002132   24.43   <2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.5844 on 98 degrees of freedom
```

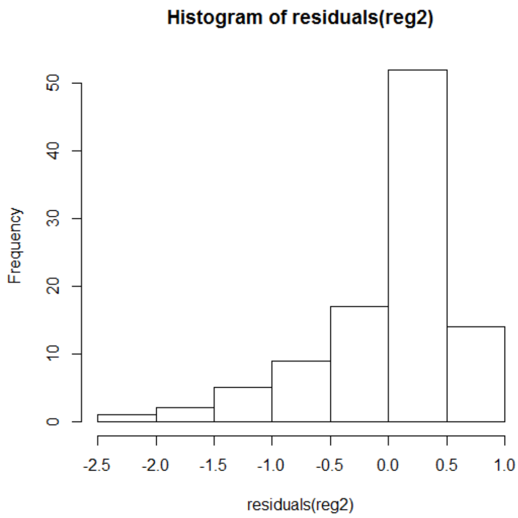
```
Multiple R-squared: 0.859,      Adjusted R-squared: 0.8575

F-statistic:   597 on 1 and 98 DF,  p-value: < 2.2e-16


> plot(X,log(Y))

> abline(reg2)

> plot(fitted(reg2),residuals(reg2),xlab= "Fitted Values",ylab="Residuals",main="Residual Plot, Model 2")

> plot(X,residuals(reg2),xlab= "Fitted Values",ylab="Residuals",main="Residual Plot, Model 2")

> plot(X,residuals(reg2),xlab= "X",ylab="Residuals",main="Residual Plot, Model 2")

>

> qqnorm(residuals(reg2),ylab="Sorted Residuals",main="QQ Plot Model 2")

> hist(residuals(reg2))

> barplot(hat(X))

> abline(h=4/length(X))
```

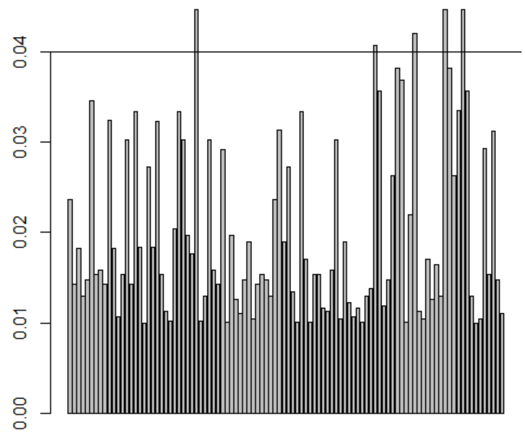












The fitted model is: $\hat{Y}_i = e^{5.8760+.0521X_i}$

The assumption of independence is violated as seen in the residuals vs fitted plot. There appears to be a non linear relationship between the residuals and the fitted values. The constant variation assumption seems to be plausible.

The normal qq plot and the histogram show a violation of the assumption of normality as they both indicated negative skew.

There are a few points with high leverage, but none appear to be a real issue in terms of influence.

3.

```
> W<-X^2

> reg3<-lm(Y~W)

> summary(reg3)


Call:

lm(formula = Y ~ W)


Residuals:

   Min     1Q Median     3Q    Max

-8.405 -2.378 -0.516  2.826 11.420


Coefficients:

            Estimate Std. Error   t value Pr(>|t|)

(Intercept) 4.9267910  0.5593973     8.807 4.67e-14 ***

W           2.9999148  0.0001305 22996.371  < 2e-16 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 3.797 on 98 degrees of freedom

Multiple R-squared:     1,     Adjusted R-squared:     1

F-statistic: 5.288e+08 on 1 and 98 DF,  p-value: < 2.2e-16


> plot(W,Y)

> abline(reg3)

> plot(fitted(reg3),residuals(reg3),xlab= "Fitted Values",ylab="Residuals",main="Residual Plot, Model 3")

> plot(W,residuals(reg3),xlab= "W",ylab="Residuals",main="Residual Plot, Model 3")
```

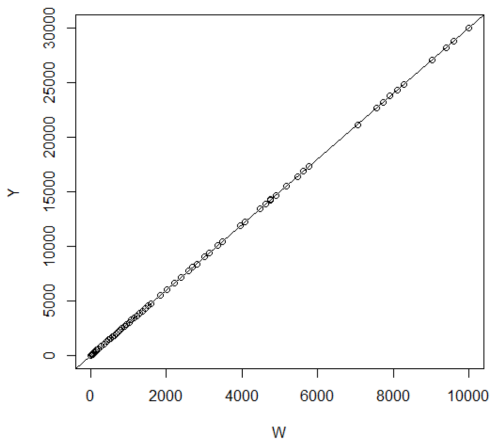
>

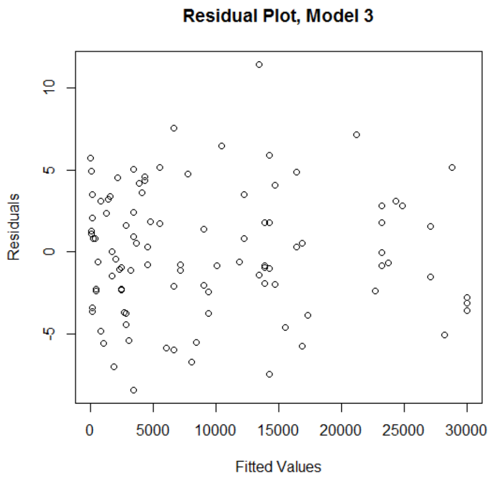
> qqnorm(residuals(reg3),ylab="Sorted Residuals",main="QQ Plot Model 3")

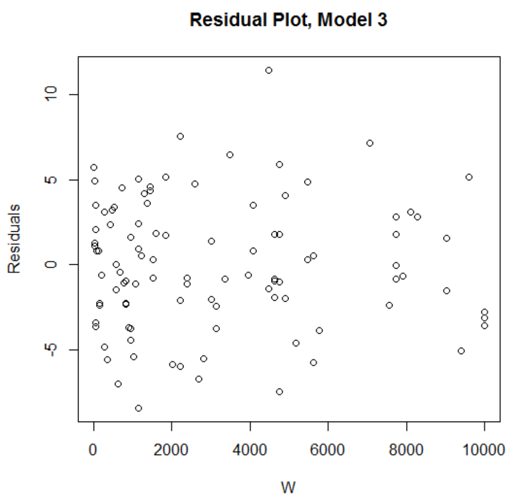
> hist(residuals(reg3))

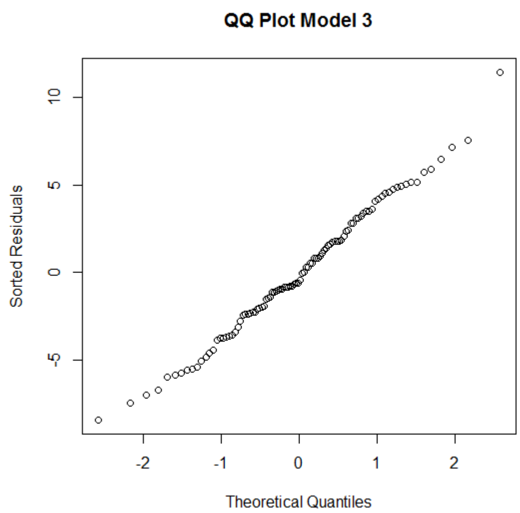
> barplot(hat(W))

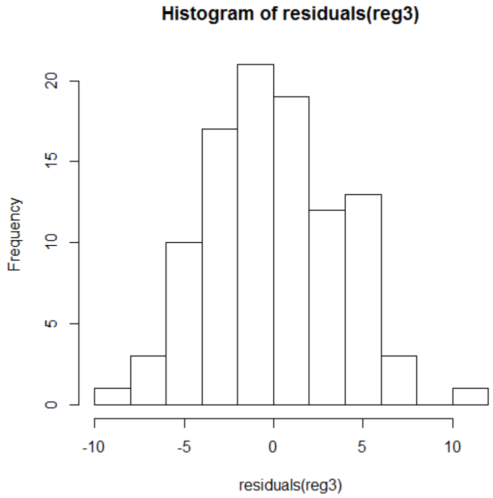
> abline(h=4/length(W))

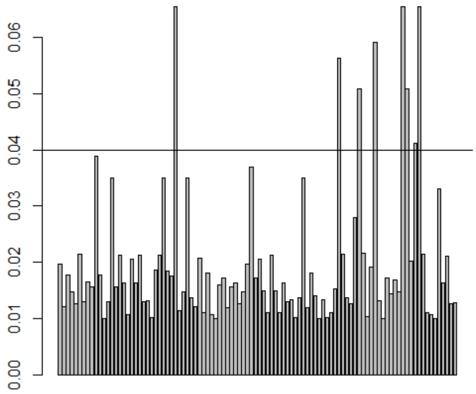












All assumptions of linear regression seem plausible.

There are a few points with high leverage, but none appear to be a real issue in terms of influence.

NB, recall that we usually test whether points are influential by removing them from the data and re-fitting.

The fitted model is: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i^2 = 4.9268 + 2.9999 X_i^2$

4.

There are a number of ways the solution could be arrived at for this question. The commands below are just one of the possibilities.

```
> X<-2.3

> pred1pi<-predict.lm(reg1,as.data.frame(X),se.fit=T, interval =  "prediction" ,level=0.99)

> pred1pi

$fit

      fit       lwr      upr

1 -4950.617 -10821.15 919.9198


$se.fit

[1] 432.9692
```

```
$df
[1] 98
```

The fitted value for model 1 is: -4950.617. The prediction interval is (-10824.15, 919.9198)

```
> pred1pi2<-predict.lm(reg2,as.data.frame(X),se.fit=T, interval =  "prediction" ,level=0.99)
> pred1pi2
$fit
      fit     lwr     upr
1 5.995858 4.43112 7.560596

$se.fit
[1] 0.115404

$df
[1] 98

$residual.scale
[1] 0.5843661
```

The fitted value for model 2 is exp(5.9959)=401.7781, and the prediction interval is (exp(4.4311),exp(7.5606))=(84.0255, 1920.9901)

```
> W<-2.3*2.3
> pred1pi3<-predict.lm(reg3,as.data.frame(W),se.fit=T, interval =  "prediction" ,level=0.99)
> pred1pi3
$fit
      fit      lwr      upr
1 20.79634 10.71471 30.87797
```

$se.fit

[1] 0.5588907

$df

[1] 98

$residual.scale

[1] 3.796884

The fitted value is 20.79634 and the prediction interval is (10.7147, 30.8779).

Note that making predictions so far away from the centre and the bulk of the data can be problematic. We should also be careful about making predictions that are outside of the range of the data (extrapolation). The minimum X value is 4.