Homework 3

Chapter 6

Question 2

A

```
women.marathon = marathon[marathon$Gender == 'female',]
```

B

```
t.test(women.marathon$Age, mu = 36.1)
```

```
##
##  One Sample t-test
##
## data:  women.marathon$Age
## t = 6.1735, df = 106, p-value = 1.249e-08
## alternative hypothesis: true mean is not equal to 36.1
## 95 percent confidence interval:
##  39.80704 43.31446
## sample estimates:
## mean of x
##  41.56075
```

C

```
wilcox.test(women.marathon$Age, mu = 36.1)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  women.marathon$Age
## V = 4522, p-value = 3.879e-07
## alternative hypothesis: true location is not equal to 36.1
```

D

```
confint(lm(women.marathon$Age ~ 1), level = 0.90)
```

```
##                  5 %     95 %
## (Intercept) 40.09296 43.02854
```

Question 3

A

```
men.marathon = marathon[marathon$Gender == 'male',]
test = t.test(
  x = men.marathon$Age,
  y = women.marathon$Age,
  alternative = 'greater',
  conf.level = 0.90
  )
print(test)
```

```
##
##  Welch Two Sample t-test
##
## data:  men.marathon$Age and women.marathon$Age
## t = 2.4519, df = 252.66, p-value = 0.007443
## alternative hypothesis: true difference in means is greater than 0
## 90 percent confidence interval:
##  1.420086      Inf
## sample estimates:
## mean of x mean of y
##  44.54438  41.56075
```

Ordinarily I prefer the formula method for describing tests, but if using a one-sided alternative, I prefer to have greater control going into the test as to which side is being compared to which.

B

```
test$conf.int
```

```
## [1] 1.420086      Inf
## attr(,"conf.level")
## [1] 0.9
```

Question 5

```
url = paste(
  'http://personal.bgsu.edu/~mrizzo',
  'Rx/Rx-data/buffalo.cleveland.snowfall.txt',
  sep = '/'
  )
snowdata = read.table(url, header = TRUE)
```

A

```
snowdata$diff = snowdata$Buffalo - snowdata$Cleveland
summary(snowdata$diff)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -14.90   12.40   29.70   33.51   43.50  136.00
```

It appears that on average, Buffalo gets more snow than Cleveland.

B

```
snowtest = t.test(snowdata$Buffalo, snowdata$Cleveland, paired = TRUE)
print(snowtest)
```

```
##
##  Paired t-test
```

```
##
## data:  snowdata$Buffalo and snowdata$Cleveland
## t = 7.5692, df = 40, p-value = 3.061e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  24.56221 42.45731
## sample estimates:
## mean of the differences
##                33.50976
```

Clearly, Buffalo gets significantly more snow than Cleveland.

C

```
snowtest$conf.int
```

```
## [1] 24.56221 42.45731
## attr(,"conf.level")
## [1] 0.95
```

The 95% Confidence Interval pegs the average difference between roughly 24.5 and 42.5 inches, in each season.

Question 7

```
t.test(
  x = pres.data$height_win,
  y = pres.data$height_lose,
  paired = TRUE
  )
```

```
##
##  Paired t-test
##
## data:  pres.data$height_win and pres.data$height_lose
## t = 1.3108, df = 15, p-value = 0.2097
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.369631  5.744631
## sample estimates:
## mean of the differences
##                  2.1875
```

We do not find sufficient evidence to reject the null hypothesis, that the height of the victor and loser are on average equal.

Chapter 7

Question 1

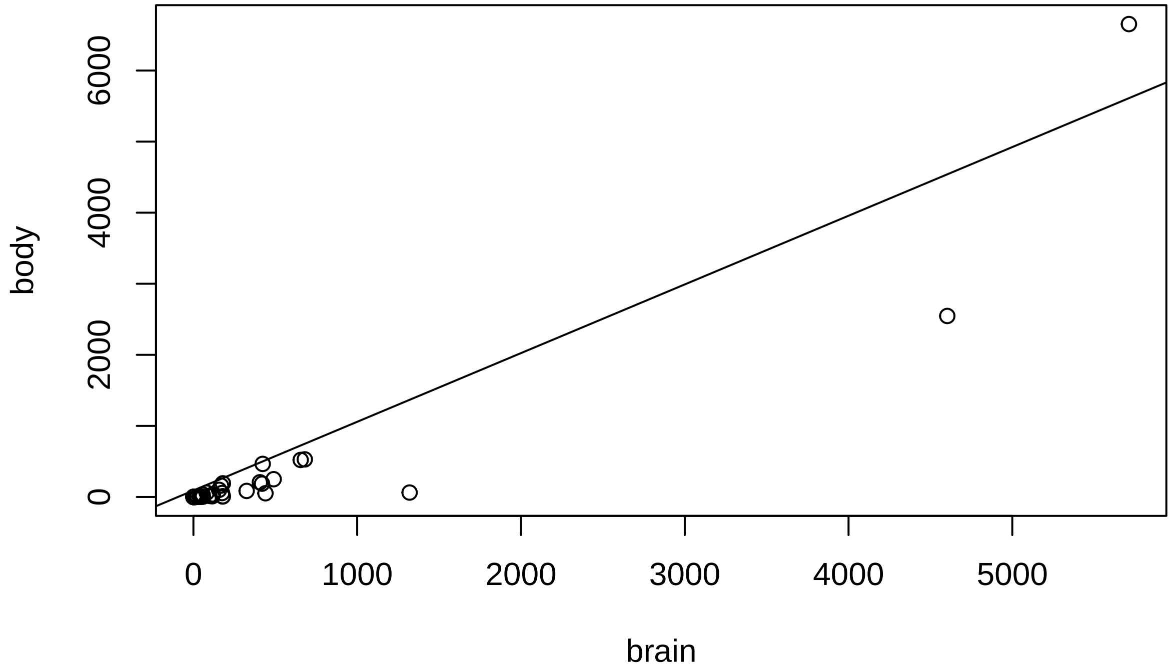
```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.4.2
```

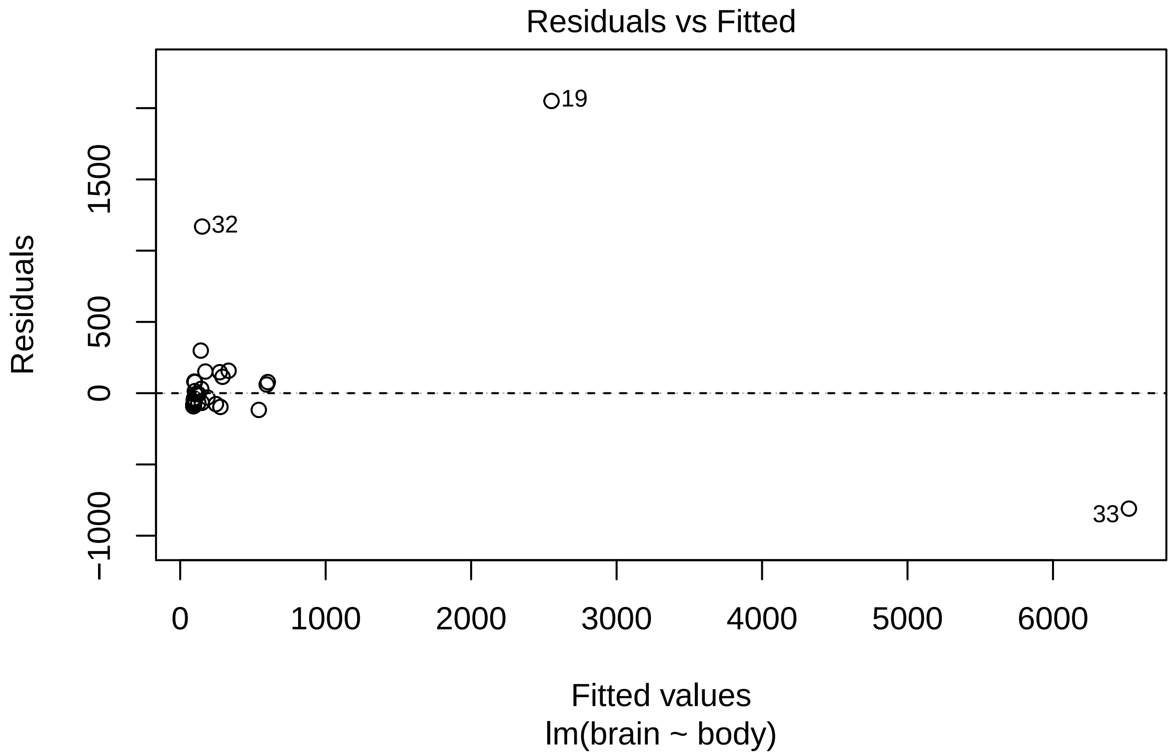
```
data(mammals)
attach(mammals)
F1=lm(brain ~ body)
print(F1)
```

```
##
## Call:
## lm(formula = brain ~ body)
##
## Coefficients:
## (Intercept)         body
##     91.0044       0.9665
```

```
plot(brain,body)
abline(lm(brain ~ body))
```



```
plot(F1, which=1, add.smooth=FALSE)
abline(h=0,lty=2)
```



```
mammals[c(19,32,33), ]
```

```
##                   body brain
## Asian elephant    2547  4603
## Human               62  1320
## African elephant 6654  5712
```

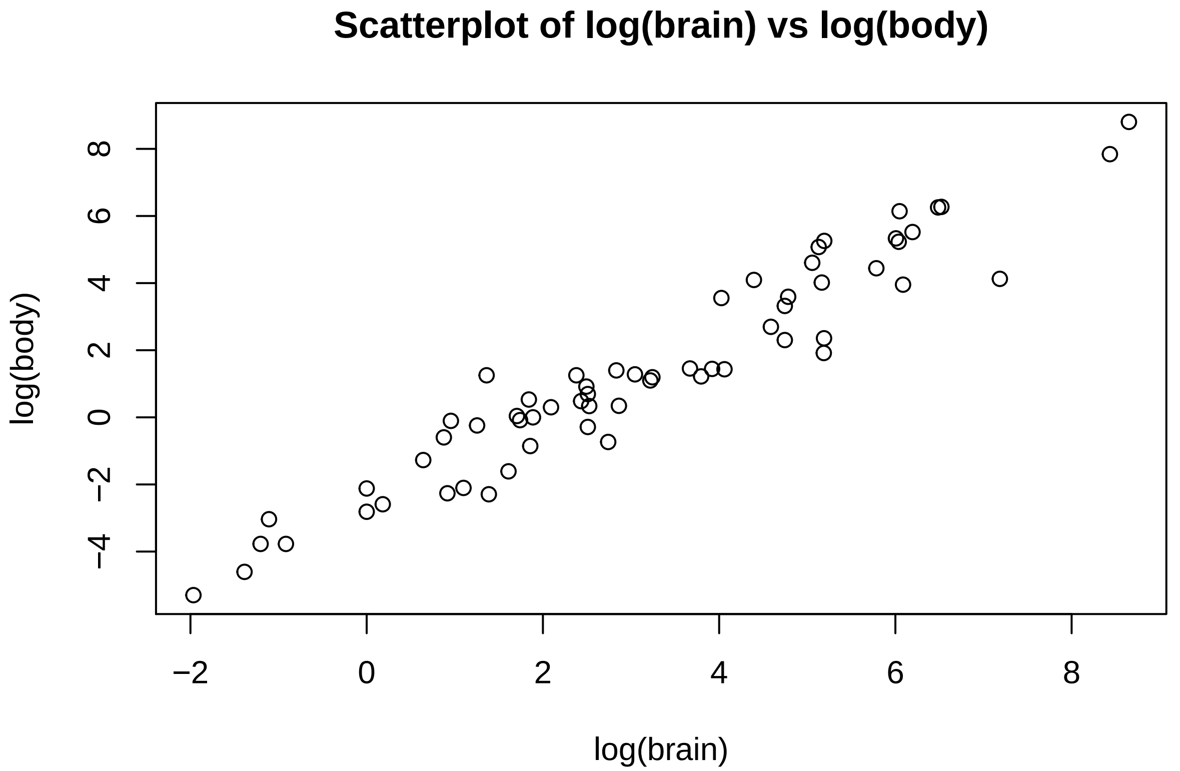
```
F1$residuals[c(19,32,33)]
```

```
##        19        32        33
## 2050.3294 1169.0728 -810.0712
```

The observations 19, 32 and 33 have the largest residuals (in absolute value), which correspond to the mammals: Asian elephant, Human and African elephant. Thus, the observation that has the largest residual (in absolute value) is the 33, corresponding to the African elephant.

Question 2

```
plot(log(brain),log(body),main="Scatterplot of log(brain) vs log(body)")
```

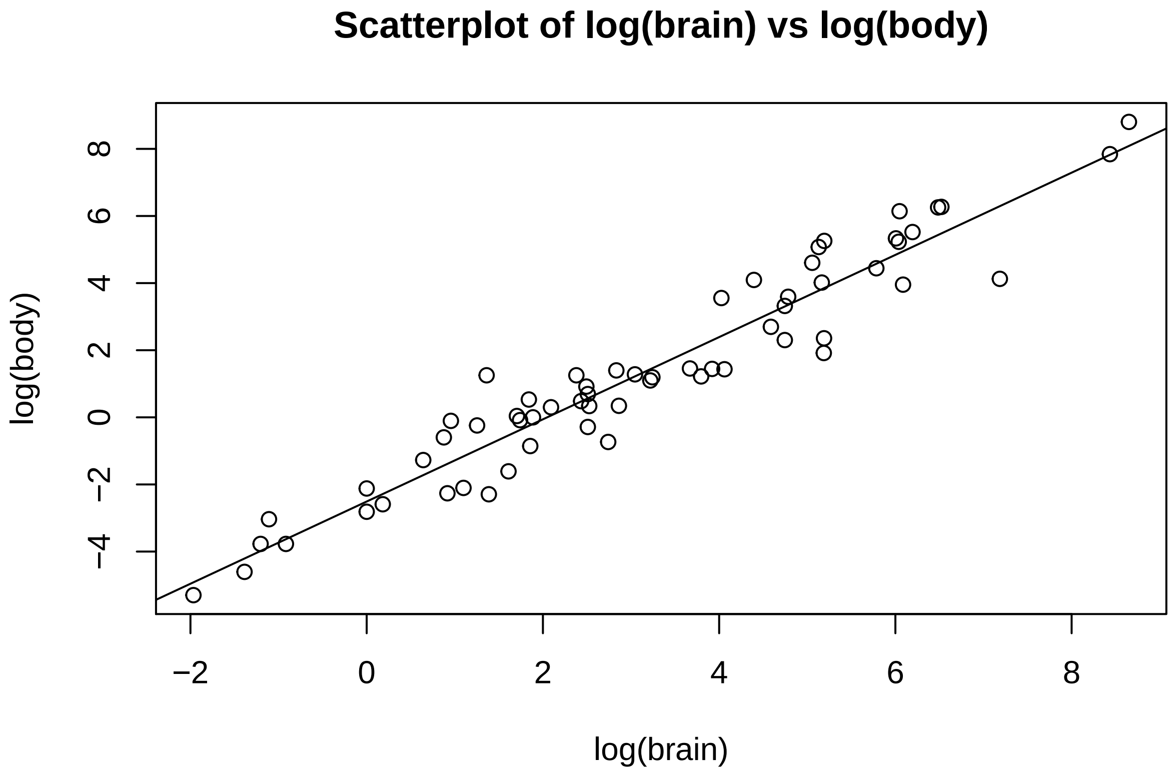


```
F2=lm(log(body) ~ log(brain))
print(F2)
```

```
##
## Call:
## lm(formula = log(body) ~ log(brain))
##
## Coefficients:
## (Intercept)   log(brain)
##      -2.509        1.225
```

The equation of the fitted model is $\log(brain) = 2.1248 + 0.7517 * \log(body)$. Which means that for every one percent increase in *body*, would yield a 0.7517% increase in the average of *brain*. We can noticed that there are not outliers anymore after the transformation.

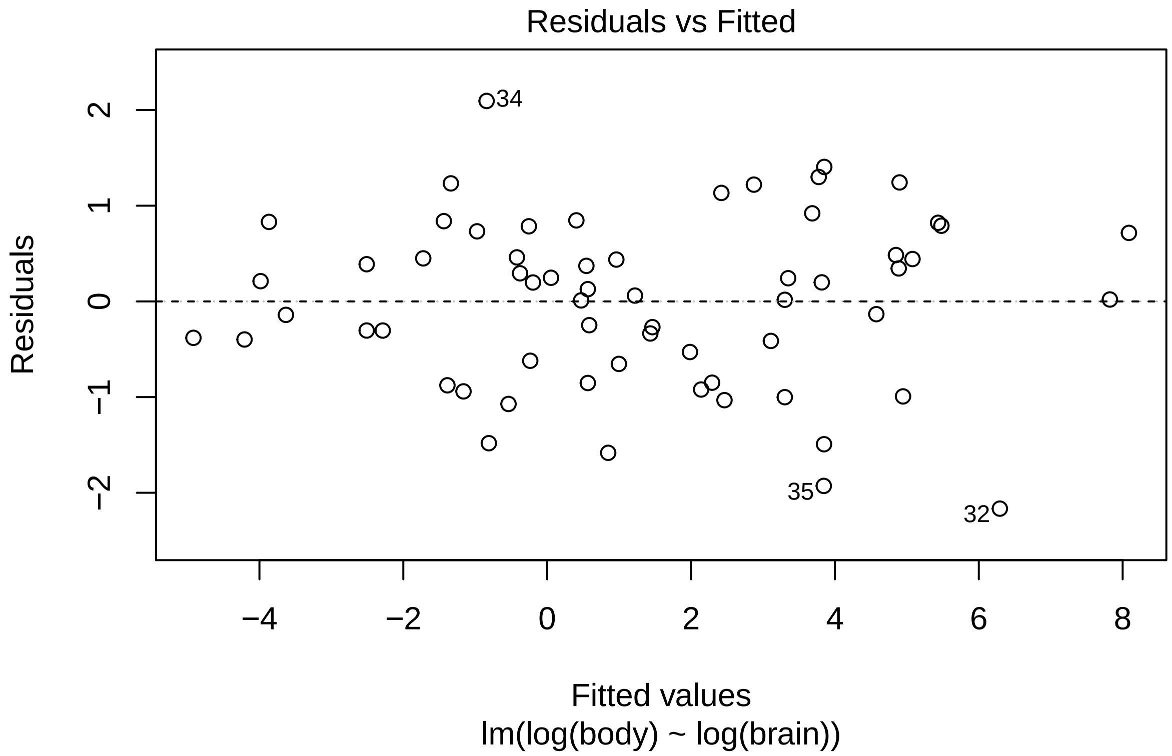
```
plot(log(brain),log(body),main="Scatterplot of log(brain) vs log(body)")
abline(lm(log(body) ~ log(brain))) #add fitted line
```



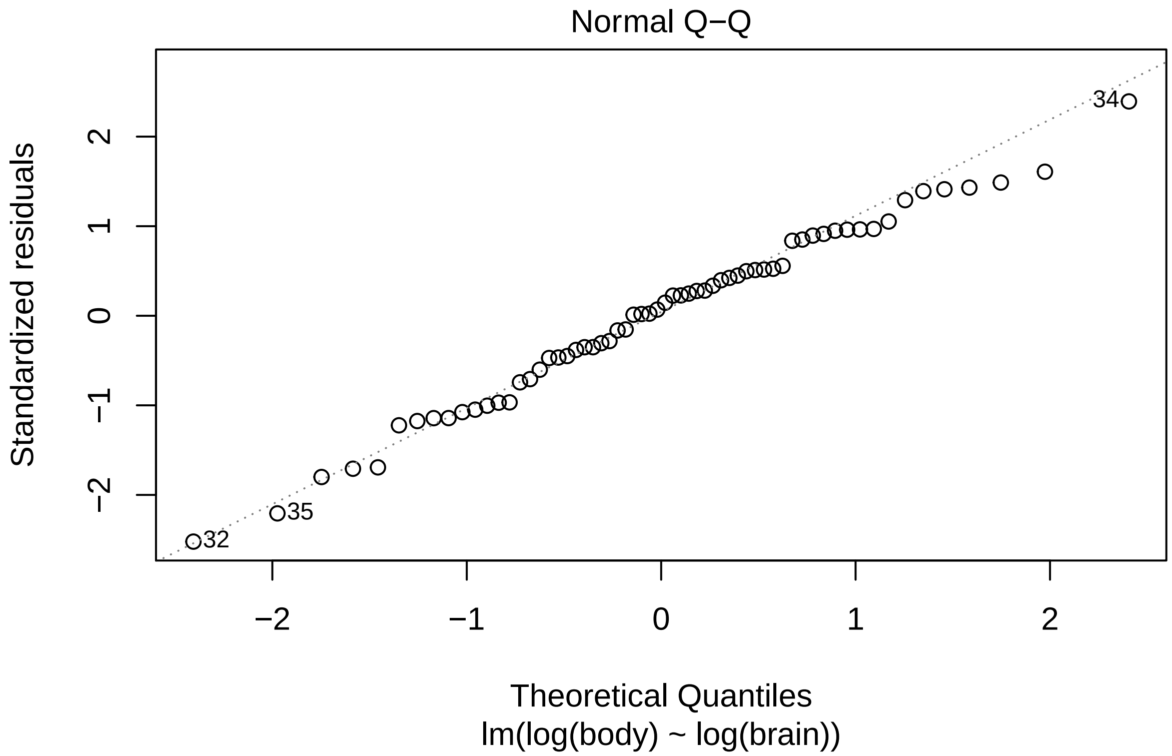
By evaluating only the fitted line seems to fit the data well since the data is close to the line. This graph was different from the scatterplot of the Exercise 7.1, which concentrated the values around a certain point, and also had three outliers, thus not producing a good fit.

Question 3

```
plot(F2, which=1, add.smooth=FALSE)
abline(h=0,lty=2)
```



```
plot(F2, which=2, add.smooth=FALSE)
```



We can observe in the residual plot that the residuals are spread around zero, and the variance is a little constant along the $\log(brain)$. In addition, by the normal QQ-plot the residuals are close to the reference line on the plot. Thus the errors may be iid with a $\text{Normal}(0, \sigma^2)$ distribution.

Question 4

```
summary(F2)
```

```
##
## Call:
## lm(formula = log(body) ~ log(brain))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.16559 -0.59763  0.09433  0.65789  2.09470
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.50907    0.18408  -13.63   <2e-16 ***
## log(brain)   1.22496    0.04638   26.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8863 on 60 degrees of freedom
## Multiple R-squared:  0.9208, Adjusted R-squared:  0.9195
## F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

```r
cor(log(brain), log(body))
```

```
## [1] 0.9595748
```

The error variance is 0.7855. The coefficient of determination is 0.9195, and the square of the correlation between the response and predictor is 0.9595. The adjust R^2 value indicates that more than 91% of the total variaton in brain size is explained by the linear associantion with body size.

Question 7

```r
attach(cars)
L1=lm(dist ~ speed)
summary(L1)
```

```
##
## Call:
## lm(formula = dist ~ speed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```r
#intercept being 0
L2 = lm(dist ~ 0 + speed)
summary(L2)
```

```
##
## Call:
## lm(formula = dist ~ 0 + speed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.183 -12.637  -5.455   4.590  50.181
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## speed   2.9091     0.1414   20.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.26 on 49 degrees of freedom
## Multiple R-squared:  0.8963, Adjusted R-squared:  0.8942
## F-statistic: 423.5 on 1 and 49 DF,  p-value: < 2.2e-16
```

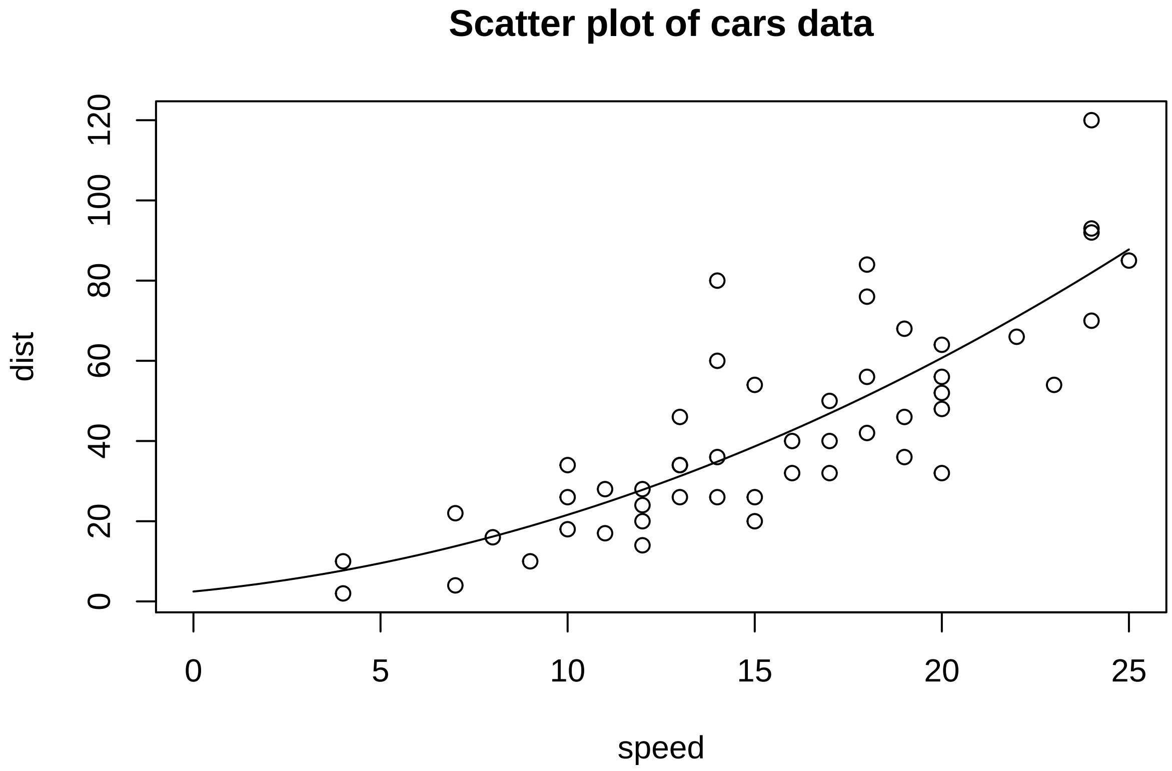
The adjust R^2 for the first model is 0.6438, and for the second model is 0.8942. For the first model the R^2 indicates that more than 64% of the total variaton in the speed is explained by the linear associantion with stopping distance. And for the second model, the R^2 indicates that more than 89% of the total variaton in the speed is explained by the linear associantion with stopping distance, which we can conclude that this model is a better fit for the data.

Question 8

```
speed2=speed^2
L3=lm(dist ~ speed + speed2)
print(L3)
```

```
##
## Call:
## lm(formula = dist ~ speed + speed2)
##
## Coefficients:
## (Intercept)        speed       speed2
##     2.47014      0.91329      0.09996
```

```
plot(cars, main="Scatter plot of cars data", xlim=c(0, 25))
curve(2.47014 + 0.91329*x + 0.09996*x^2, add=TRUE)
```



The model adding a quadratic variable seems to fit better the data, given that the data are more spread along the fitted line.

Question 9 NEED COMMENTS

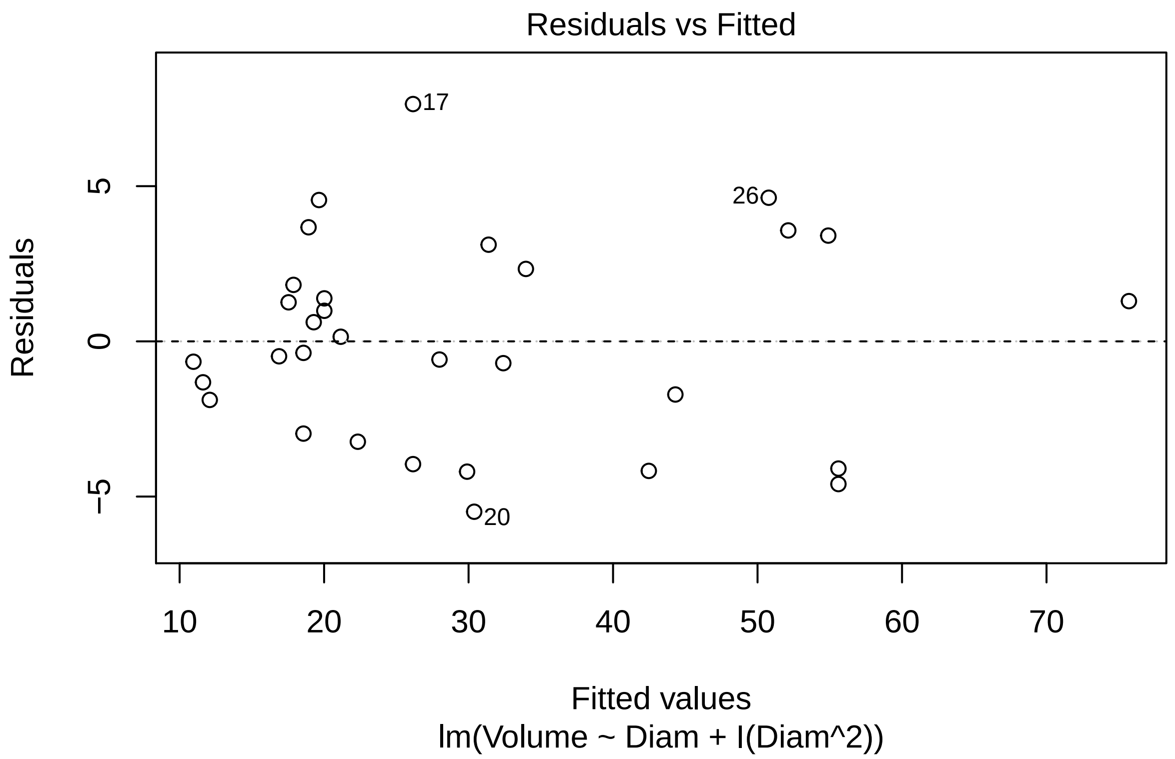
```
Trees = trees
names(Trees)[1] = "Diam"
attach(Trees)

M1 = lm(Volume ~ Diam + I(Diam^2))
print(M1)
```

```
##
## Call:
## lm(formula = Volume ~ Diam + I(Diam^2))
##
## Coefficients:
## (Intercept)         Diam    I(Diam^2)
##     10.7863      -2.0921       0.2545
```

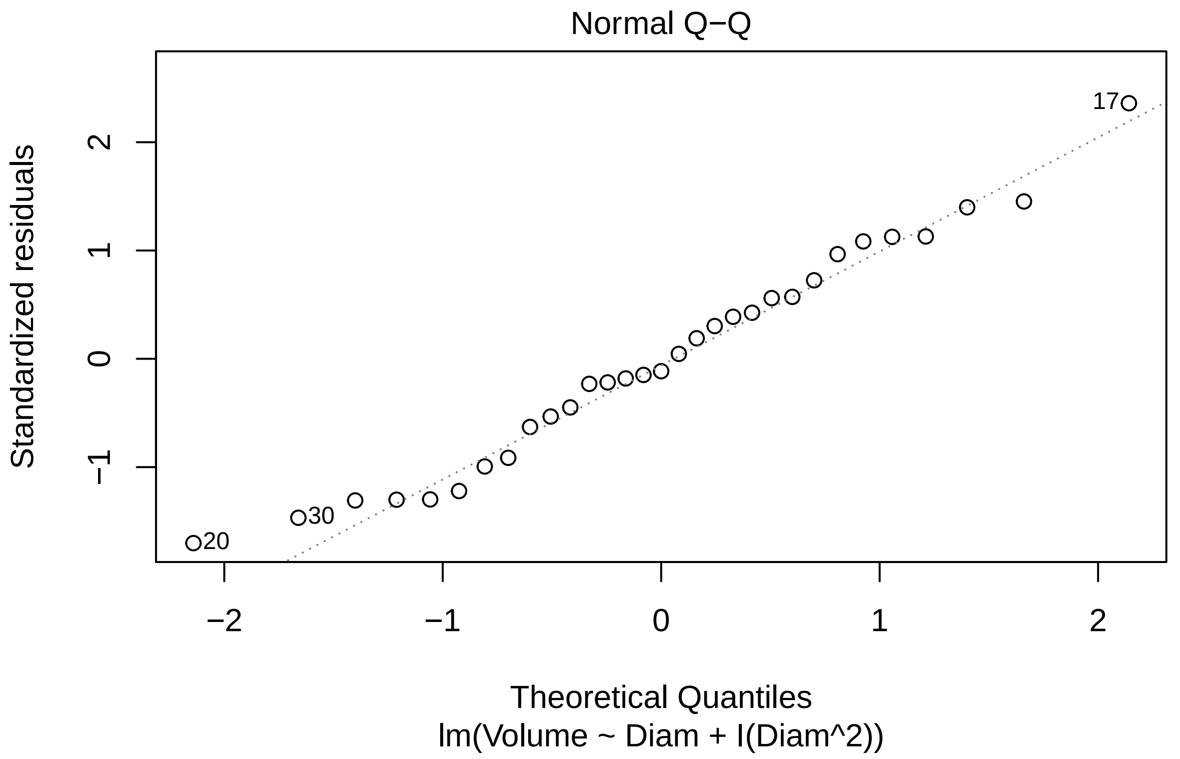
```
new = data.frame(Diam=16)
predict(M1, new)
```

```
##        1
## 42.47365
```

```
plot(M1, which=1, add.smooth=FALSE)
abline(h=0,lty=2)
```



```r
plot(M1, which=2, add.smooth=FALSE)
```



We can observe in the residual plot that the residuals are spread around zero, with one observation having a higher residual value. In addition, the normal QQ-plot the residuals seems to be close to the reference line on the plot. Thus the errors may be iid with a Normal$(0, \sigma^2)$ distribution.

```r
summary(M1)
```

```
##
## Call:
## lm(formula = Volume ~ Diam + I(Diam^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4889 -2.4293 -0.3718  2.0764  7.6447
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.78627   11.22282   0.961 0.344728
## Diam        -2.09214    1.64734  -1.270 0.214534
## I(Diam^2)    0.25454    0.05817   4.376 0.000152 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.335 on 28 degrees of freedom
## Multiple R-squared:  0.9616, Adjusted R-squared:  0.9588
## F-statistic: 350.5 on 2 and 28 DF,  p-value: < 2.2e-16
```

The adjusted R^2 value of 0.9588 indicates that more than 95% of the total variation in Volume about its mean is explained by the linear association with the predictors Diam and Diam2. The residual standard error is 3.335.

In additon, the p-values of the Diam variable are not under significant level 0.05, which we can conclude that Diam are not significant.