S.J. Sheather, A Modern Approach to Regression with R (Chapter 3, Exercises 1,3,4)

Soodabeh Ramezani

September 22

Question 1

1. The data file airfares.txt on the book web site gives the one-way airfare (in US dollars) and distance (in miles) from city A to 17 other cities in the US. Interest centers on modeling airfare as a function of distance. The first model fit to the data was $Fare = \beta 0 + \beta 1 Distance + e$

(a) Based on the output for model (3.7) a business analyst concluded the following:The regression coefficient of the predictor variable, Distance is highly statistically significant and the model explains 99.4% of the variability in the Y-variable, Fare. Thus model (1) is a highly effective model for both understanding the effects of Distance on Fare and for predicting future values of Fare given the value of the predictor variable, Distance. provide a detailed critique of this conclusion.

Answer-a: The numerical regression output is not enough to ensure the appropriateness of the fitted model. It should be supplemented by an analysis and needs additional tools. The plot of residuals is one tool to validate a regression model. The plot of residuals here does not indicate an appropriate fit to the data.

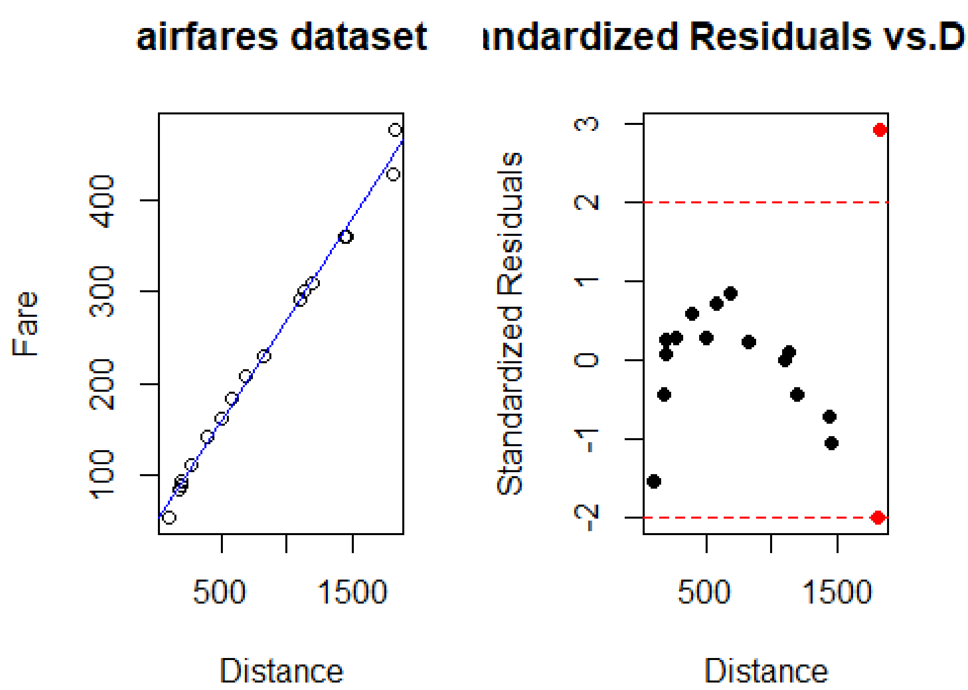
(b) Does the ordinary straight line regression model (3.7) seem to fit the data well? If not, carefully describe how the model can be improved. Given below and in Figure 3.41 is some output from fitting model (3.7).

Answer-b: It is clear that the mode does not fit to data. In the scatter plot of standardized residual, a non-random pattern is evident, so this indicates that the relationship between airfare and distance is in fact non-linear. We can also identify non-constant variances in the errors, and the errors are not normally distributed. There are two bad leverage points which need more attention .

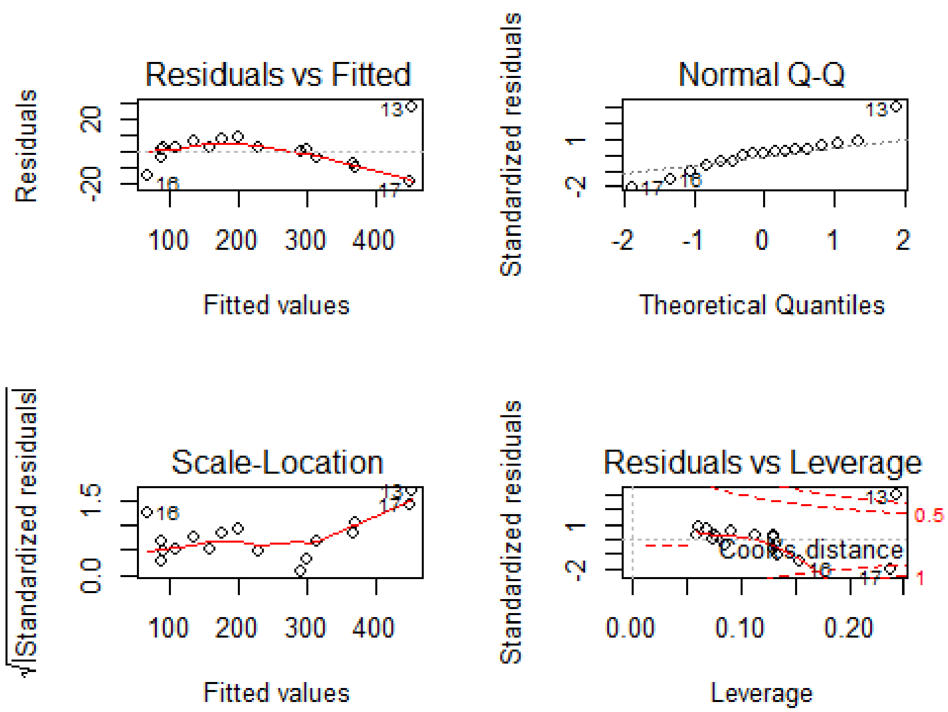
At first, the two bad leverage points should be investigated to see if there was any reason why they do not follow the pattern of the other points. Secondly, we have to consider if there are any other predictor variables (e.g., destination cities) affecting the airfare. Finally, if we could not fix the model, we will use transformation to overcome the problem of nonlinearity. as an example, a box-cox transformation has been applied to both x and y datapoints as shown below. The standardized resiudals shows more or less no patterns after the transformation applied

```
setwd("C:/Users/")
airfares <- read.delim("airfares.txt", header = T, sep = "\t")
#head(airefares)
```

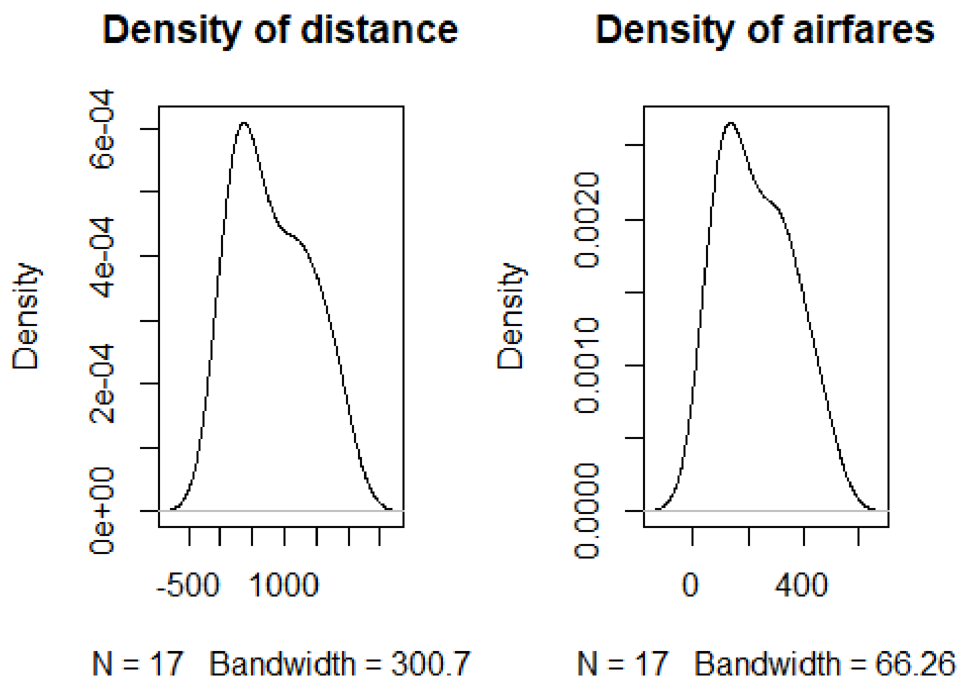
```r
X_AF=airfares$Distance
Y_AF=airfares$Fare
model1_AF=lm(Y_AF~X_AF)
#plot the first model
par(mfrow = c(1,2))
plot(X_AF,Y_AF,xlab = "Distance", ylab="Fare", main="airfares dataset")
abline(model1_AF, col="blue")

#levergage points
l=length(X_AF)
lev <- hatvalues(model1_AF)
high<- 4/l
co= ifelse(lev <= high ,1,2)
d=which(high <= lev)

#look at the standardized residuals
#rstudent(model1_AF)
sdres_AFM1 <- rstandard(model1_AF)
plot(sdres_AFM1~X_AF,xlab="Distance", ylab = "Standardized Residuals", main =
"Standardized Residuals vs.Distance",pch = 16, col=co)
abline(h = 2, lty = 2, col = 2)
abline(h = -2, lty = 2, col = 2)
```




```r
#The 4 diagnosis plots in R
par(mfrow = c(2,2))
plot(model1_AF)
```

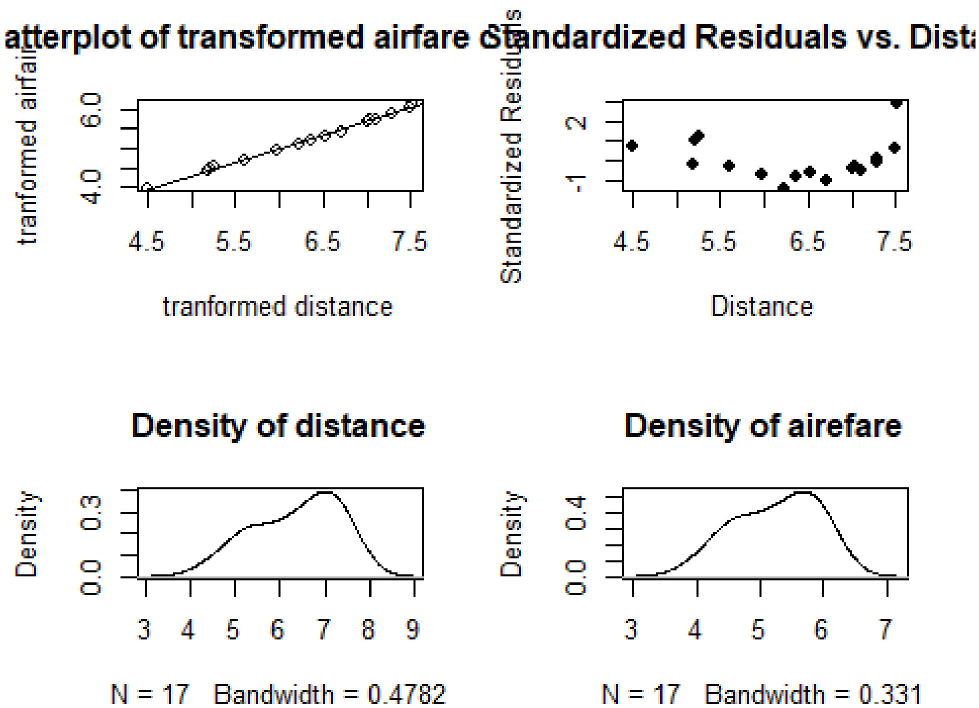



```
#check the normality of variables
par(mfrow = c(1,2))
plot(density(X_AF), main ="Density of distance")
plot(density(Y_AF), main="Density of airfares")

#transformming x and y.
library(alr4)

## Loading required package: car

## Loading required package: carData

## Loading required package: effects

## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```




```
#Find out Lambda
Lambda_AF <- powerTransform(cbind(X_AF,Y_AF))
summary(Lambda_AF)

## bcPower Transformations to Multinormality
##      Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## X_AF    0.1098           0      -0.2315       0.4512
## Y_AF   -0.0207           0      -0.4549       0.4135
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                             LRT df      pval
## LR test, lambda = (0 0) 11.73688  2 0.0028273
##
## Likelihood ratio test that no transformations are needed
##                             LRT df       pval
## LR test, lambda = (1 1) 19.99211  2 4.5579e-05

X_AF_ln <- log(X_AF)
Y_AF_ln <- log(Y_AF)
lm2_transform <- lm(Y_AF_ln ~ X_AF_ln)
par(mfrow = c(2,2))
plot(Y_AF_ln ~ X_AF_ln, xlab = "tranformed distance", ylab = "tranformed
airfair", main = "Scatterplot of transformed airfare dataset")
abline(lm2_transform)

sdres_AFM1 <- rstandard(lm2_transform)
```

```
plot(sdres_AFM1~X_AF_ln,xlab="Distance", ylab = "Standardized Residuals",
main = "Standardized Residuals vs. Distance",pch = 16)

#check the normality of variables
plot(density(X_AF_ln), main ="Density of distance")
plot(density(Y_AF_ln), main="Density of airefare")
```



Question 3

3. The price of advertising (and hence revenue from advertising) is different from one consumer magazine to another. Publishers of consumer magazines argue that magazines that reach more readers create more value for the advertiser. Thus, circulation is an important factor that affects revenue from advertising. In this exercise, we are going to investigate the effect of circulation on gross advertising revenue. The data are for the top 70 US magazines ranked in terms of total gross advertising revenue in 2006. In particular we will develop regression models to predict gross advertising revenue per advertising page in 2006 (in thousands of dollars) from circulation (in millions). The data were obtained from Http://adage.com and are given in the file AdRevenue.csv which is available on the book web site. Prepare your answers to parts A, B and C in the form of a report.

Part A

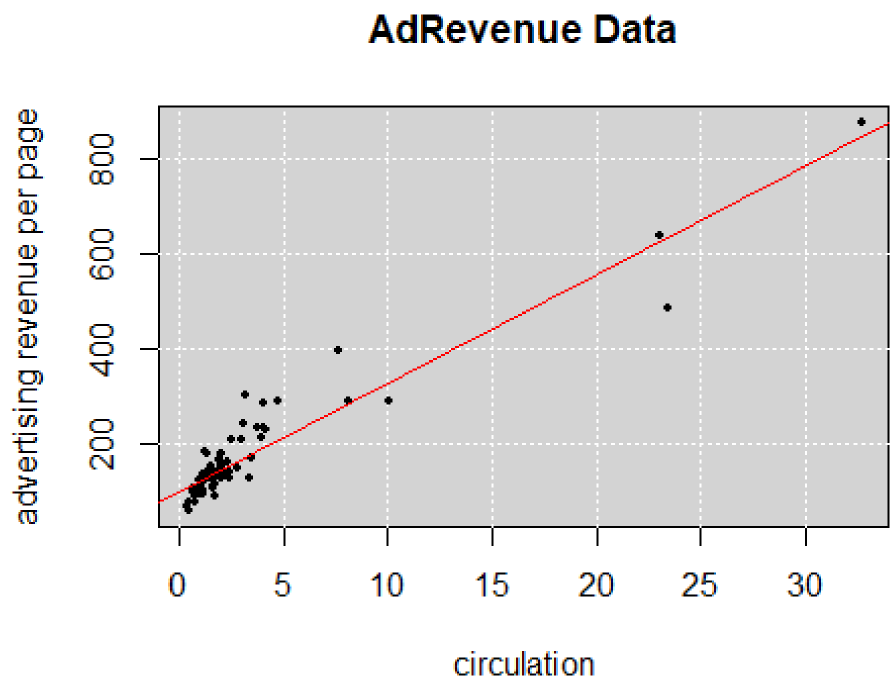
(a) Develop a simple linear regression model based on least squares that predicts advertising revenue per page from circulation (i.e., feel free to transform either the predictor or the response variable or both variables). Ensure that you provide justification for your choice of model. (b)Find a 95% prediction interval for the advertising revenue per page for magazines with the following circulations: (i)0.5 million (ii)20 million

(c)Describe any weaknesses in your model.

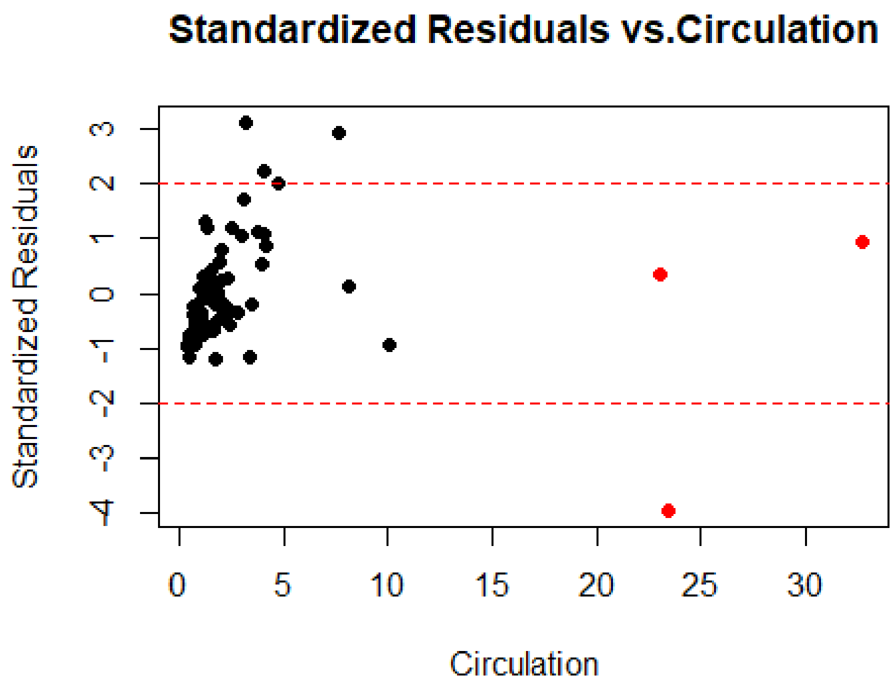
Answer: The final model has a segnificant R_square close to 1.from the scatter plot of standardized residual, a non-random pattern is evident .The normal Q-Q plot is not a straight line thus violating the assumption normality of errors. There are also some outliers.

95% prediction intervals are shown below

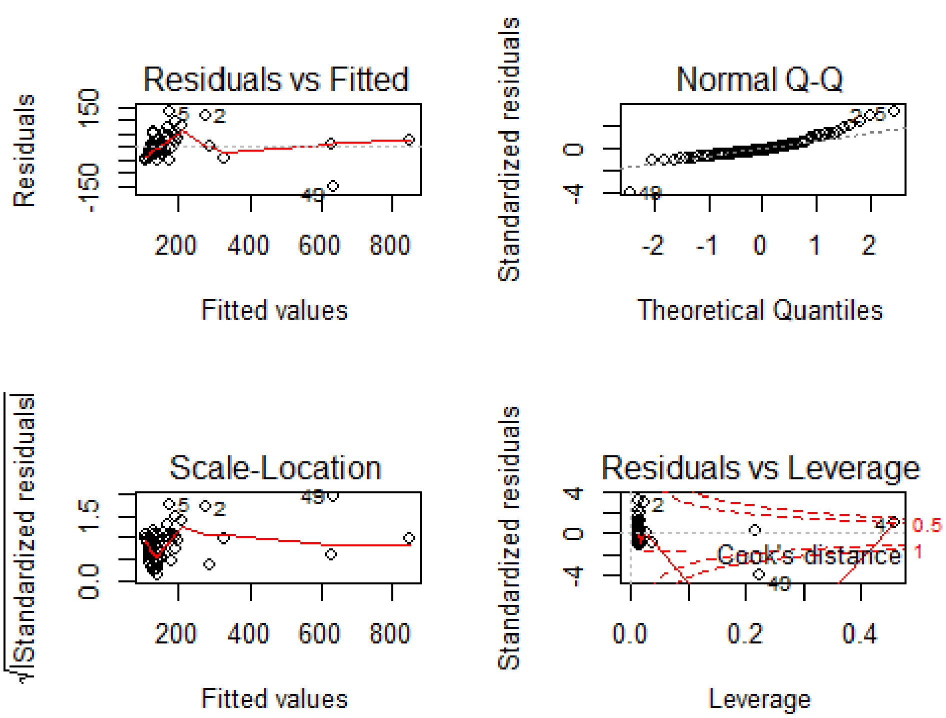
```
setwd("C:/Users/")
AdRevenue <- read.csv("AdRevenue.csv ", header = T, stringsAsFactors = F)
#head(AdRevenue)
X_AD=AdRevenue$Circulation
Y_AD=AdRevenue$AdRevenue

#trying  simple linear regression---------------------------------------------
------------------
model1_AD=lm(Y_AD~X_AD)
summary(model1_AD)

##
## Call:
## lm(formula = Y_AD ~ X_AD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -147.694  -22.939   -7.845   13.810  131.130
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  99.8095     5.8547   17.05   <2e-16 ***
## X_AD         22.8534     0.9518   24.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.22 on 68 degrees of freedom
## Multiple R-squared:  0.8945, Adjusted R-squared:  0.8929
## F-statistic: 576.5 on 1 and 68 DF,  p-value: < 2.2e-16

plot(X_AD,Y_AD,ylab ="advertising revenue per page" ,type = "n" ,xlab
="circulation",main="AdRevenue Data")
rect(par("usr")[1],par("usr")[3],par("usr")[2],par("usr")[4],col = "light
gray")
grid(col = "white", lty = "dotted",lwd = par("lwd"))
points(X_AD,Y_AD,pch = 20, cex = 0.8 )
abline(model1_AD, cex = 25, col = "red")
```



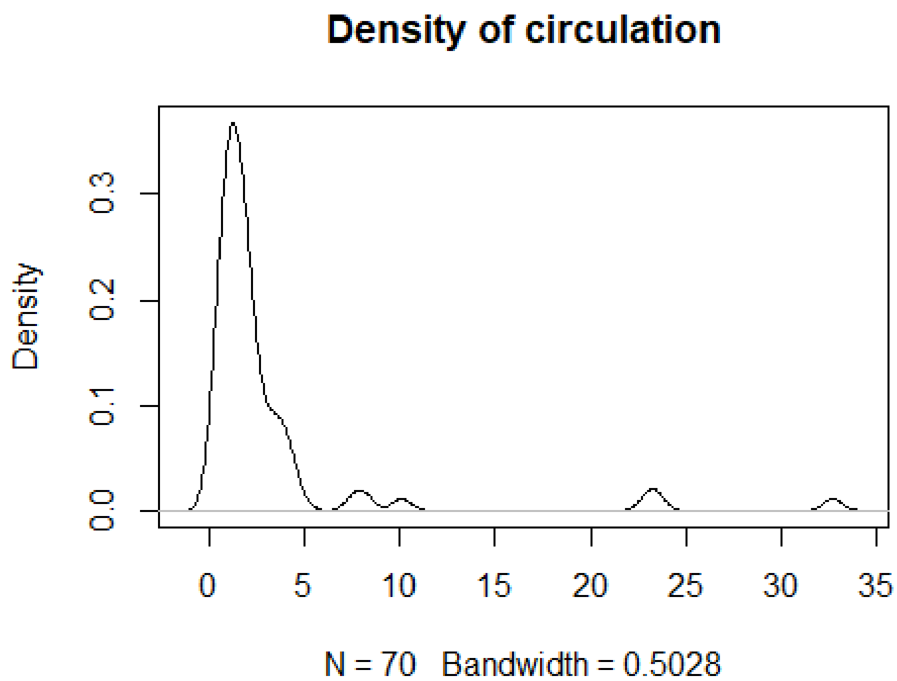

```
#levergage points
l=length(X_AD)
lev <- lm.influence(model1_AD)$hat
high<- 4/l
co= ifelse(lev <= high ,1,2)
d=which(high <= lev)
cat("leverage:",d)

## leverage: 4 8 49

#look at the standardized residuals
sdres_ADM1 <- rstandard(model1_AD)
plot(sdres_ADM1~X_AD,xlab="Circulation", ylab = "Standardized Residuals",
main = "Standardized Residuals vs.Circulation",col=co,pch = 16)
abline(h = 2, lty = 2, col = 2)
abline(h = -2, lty = 2, col = 2)
```



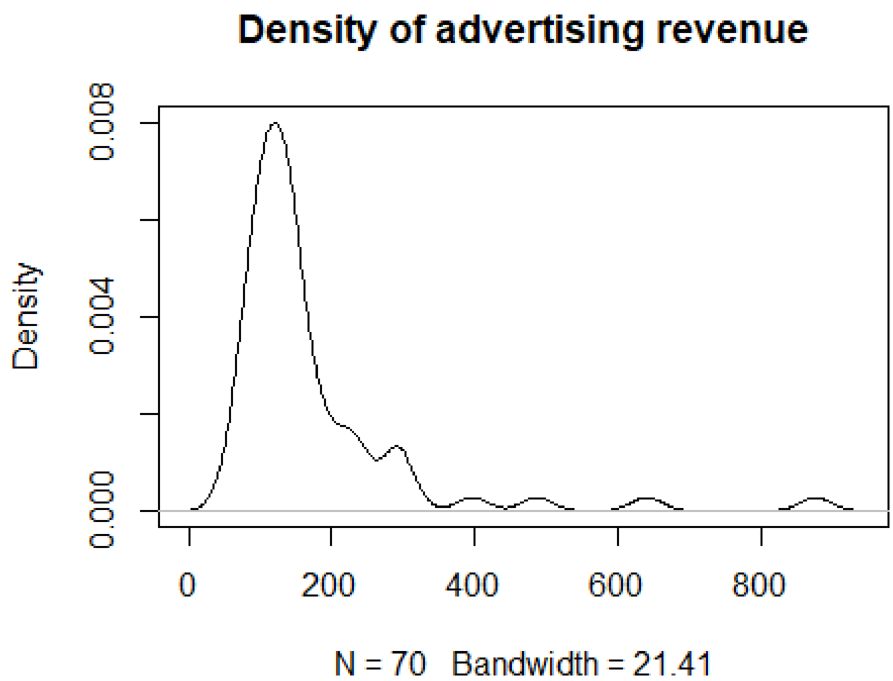

```r
#The 4 diagnosis plots in R
par(mfrow = c(2,2))
plot(model1_AD)
```



```r
#check the normality of variables
par(mfrow = c(1,1))
plot(density(X_AD), main ="Density of circulation")
```



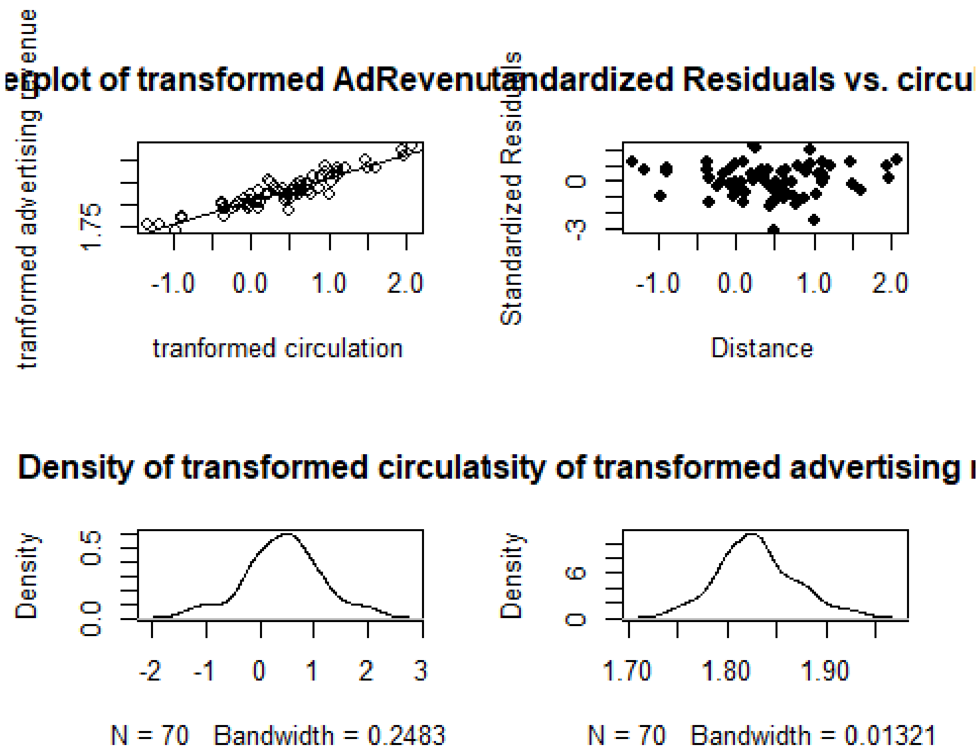

```r
plot(density(Y_AD), main="Density of advertising revenue")
```



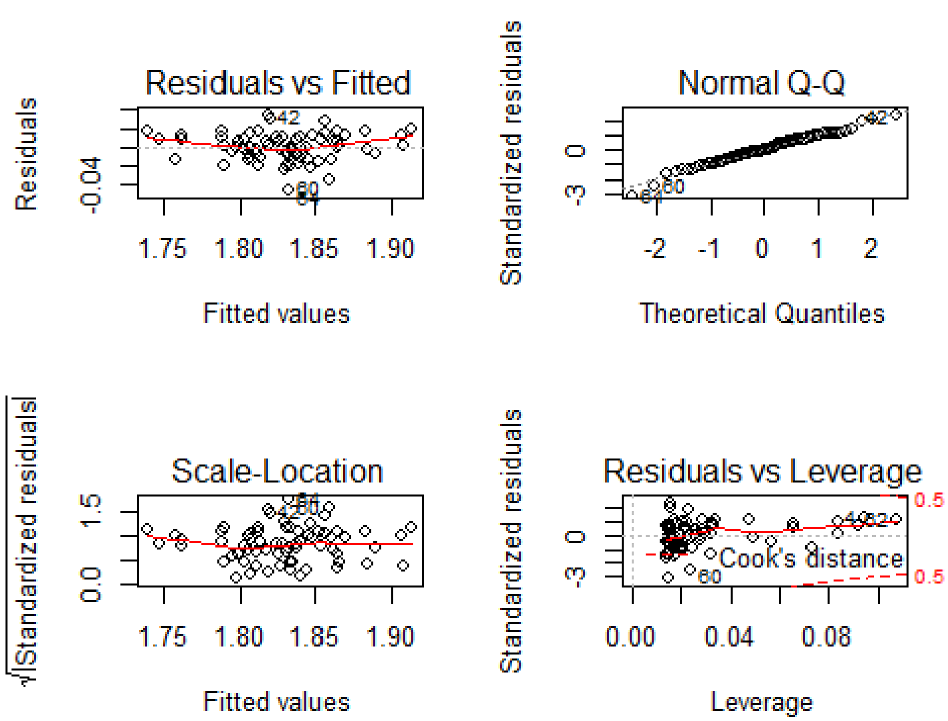

```
#transforming variables--------------------------------------------------------------------

#Lambda

Lambda_AD <- powerTransform(cbind(X_AD,Y_AD))
summary(Lambda_AD)

## bcPower Transformations to Multinormality
##      Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## X_AD   -0.2428       -0.33      -0.4051      -0.0805
## Y_AD   -0.5873       -0.50      -0.9222      -0.2524
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                               LRT df     pval
## LR test, lambda = (0 0) 13.80957  2 0.001003
##
## Likelihood ratio test that no transformations are needed
##                               LRT df       pval
## LR test, lambda = (1 1) 249.2147  2 < 2.22e-16

X_AD_t <- bcPower(X_AD, -.33)
Y_AD_t <- bcPower(Y_AD, -.50)
lm_AD_t <- lm(Y_AD_t ~ X_AD_t)
summary(lm_AD_t)
```

```
##
## Call:
## lm(formula = Y_AD_t ~ X_AD_t)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.044776 -0.008991 -0.000452  0.010003  0.033169
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.807685   0.002057  878.96   <2e-16 ***
## X_AD_t      0.051047   0.002524   20.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0146 on 68 degrees of freedom
## Multiple R-squared:  0.8574, Adjusted R-squared:  0.8553
## F-statistic:   409 on 1 and 68 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(Y_AD_t ~ X_AD_t, xlab = "tranformed circulation", ylab = "tranformed
advertising revenue", main = "Scatterplot of transformed AdRevenue dataset")
abline(lm_AD_t)


sdres_ADMt <- rstandard(lm_AD_t)
plot(sdres_ADMt~X_AD_t,xlab="Distance", ylab = "Standardized Residuals", main
= "Standardized Residuals vs. circulation",pch = 16)

#check the normality of variables

plot(density(X_AD_t), main ="Density of transformed circulation")
plot(density(Y_AD_t), main="Density of transformed advertising revenue")
```




```r
par(mfrow = c(2,2))
plot(lm_AD_t)
```



```r
#-----------------------------------------------------------------------------
---

m=predict(lm_AD_t,data.frame(X_AD_t= bcPower(c(0.5,20),-.33)), interval =
"prediction",conf.int = 0.95)
in1= bcnPowerInverse(m, lambda = -0.5, gamma = 0)
```

Part B

(a) Develop a polynomial regression model based on least squares that directly predicts the effect on advertising reve-nue per page of an increase in circulation of 1 million people (i.e., do not transform either the predictor nor the response variable). Ensure that you provide detailed justification for your choice of model. [Hint: Consider polynomial models of order up to 3.]

(b) Find a 95% prediction interval for the advertising page cost for magazines with the following circulations: (i)0.5 million (ii)20 million

(c)Describe any weaknesses in your model.

Answer: This model has a segnificant R_squared . from the scatter plot of standardized residual, a non-random pattern is evident .The normal Q-Q plot is not a straight line, so the assumption that the errors are normally distributed is not true. There are some outliers.

a polynomial of degree 4 was fitted to to data to help with better fitting the data and removing non-random patterns in standardized residuals

95% prediction intervals are shown below

```r
#polynomial models-----------------------------------------------------------
--------------

#First way using poly() function
lm_AD_p=lm(Y_AD~poly(X_AD,degree=3,raw = TRUE))
summary(lm_AD_p)

##
## Call:
## lm(formula = Y_AD ~ poly(X_AD, degree = 3, raw = TRUE))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -83.75 -13.56  -2.16  11.46 104.82
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          59.17037    8.34505   7.090 1.12e-09
## poly(X_AD, degree = 3, raw = TRUE)1 51.23582    4.71123  10.875 2.33e-16
## poly(X_AD, degree = 3, raw = TRUE)2 -2.50538    0.41141  -6.090 6.48e-08
## poly(X_AD, degree = 3, raw = TRUE)3  0.05223    0.00923   5.658 3.57e-07
##
```

```
## (Intercept)                         ***
## poly(X_AD, degree = 3, raw = TRUE)1 ***
## poly(X_AD, degree = 3, raw = TRUE)2 ***
## poly(X_AD, degree = 3, raw = TRUE)3 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.06 on 66 degrees of freedom
## Multiple R-squared:  0.9333, Adjusted R-squared:  0.9303
## F-statistic: 308.1 on 3 and 66 DF,  p-value: < 2.2e-16
```

```
cbind(coef(lm_AD_p))
```

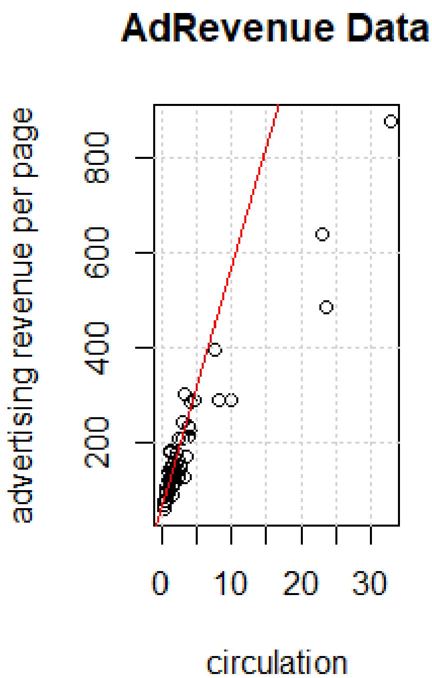
```
##                                            [,1]
## (Intercept)                         59.17036829
## poly(X_AD, degree = 3, raw = TRUE)1 51.23581639
## poly(X_AD, degree = 3, raw = TRUE)2 -2.50537894
## poly(X_AD, degree = 3, raw = TRUE)3  0.05222479
```

```
#Second way------------------------------
x1=X_AD
x2=X_AD^2
x3=X_AD^3
pol_AD <- lm(Y_AD ~ x1 +x2+x3)
summary(pol_AD)
```

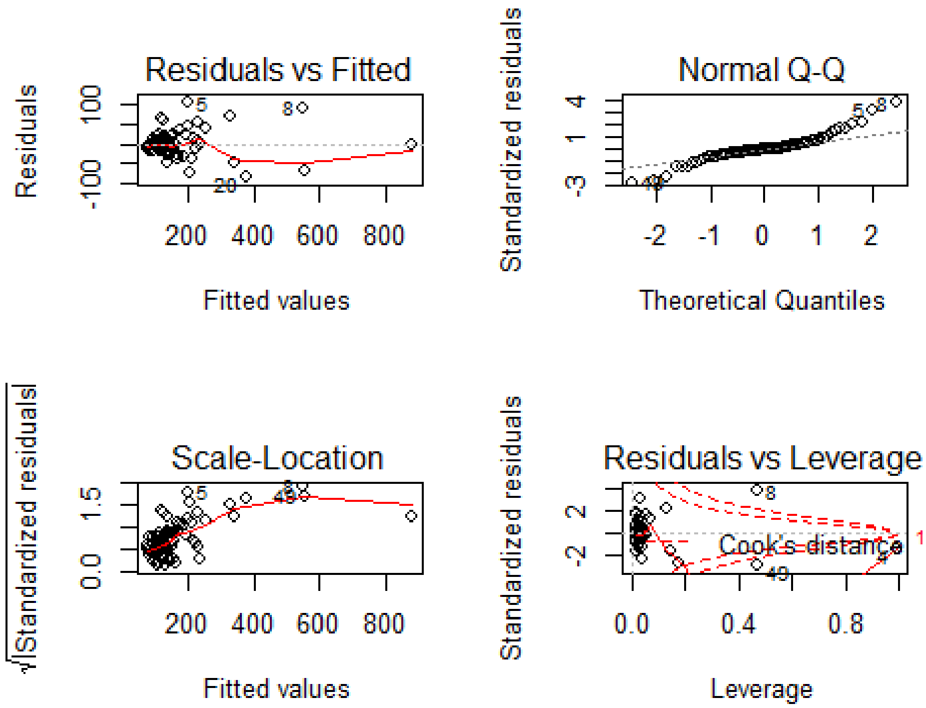
```
##
## Call:
## lm(formula = Y_AD ~ x1 + x2 + x3)
##
## Residuals:
##    Min      1Q Median      3Q     Max
## -83.75 -13.56  -2.16  11.46 104.82
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.17037    8.34505   7.090 1.12e-09 ***
## x1          51.23582    4.71123  10.875 2.33e-16 ***
## x2          -2.50538    0.41141  -6.090 6.48e-08 ***
## x3           0.05223    0.00923   5.658 3.57e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.06 on 66 degrees of freedom
## Multiple R-squared:  0.9333, Adjusted R-squared:  0.9303
## F-statistic: 308.1 on 3 and 66 DF,  p-value: < 2.2e-16
```

```
#plot the model--------------------------
par(mfrow=c(1,2))
plot(X_AD,Y_AD,ylab ="advertising revenue per page",xlab
="circulation",main="AdRevenue Data")
```

```
grid()
abline(lm_AD_p, cex = 25, col = "red")

## Warning in abline(lm_AD_p, cex = 25, col = "red"): only using the first
two
## of 4 regression coefficients

#check if the model is valis--------------
par(mfrow=c(2,2))
```




```
plot(lm_AD_p)

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```




```
#prediction-------------------------------
a=0.5
b=20
in2=predict(lm_AD_p,data.frame(X_AD=c(a,b)), interval = "prediction",conf.int
= 0.95)
```

Part C

(a) Compare the model in Part A with that in Part B. Decide which provides a better model. Give reasons to justify your choice.

Answer-a: Model in part A is better. However, the model in part B has a higher Adjusted R-squared, the plot of standard residuals in part A is more acceptable. Another reason to choose the model in part A is that it is not as complex as that in Part B.

(b)Compare the prediction intervals in Part A with those in Part B. In each case, decide which interval you would recommend. Give reasons to justify each choice.

Answer-b: At point 0.5 the predicition intervals in part A is narrower , but for 20 million point model A gives rise to slightly wider prediction interval indicating that the model A better describes the original data (in the original data variance increases as x-variable increases).

```
#coparing intervals
in1
```

```
##         fit       lwr       upr
## 1  74.26997  58.25424  97.92731
## 2 441.48250 254.17394 948.96743
```

```
in2
```

```
##         fit       lwr      upr
## 1  84.16846  14.92314 153.4138
## 2 499.53342 418.17903 580.8878
```

Question 4

Tryfos (1998, p. 57) considers a real example involving the management at a Canadian port on the Great Lakes who wish to estimate the relationship between the volume of a ship's cargo and the time required to load and unload this cargo. It is envisaged that this relationship will be used for planning purposes as well as for making comparisons with the productivity of other ports. Records of the tonnage loaded and unloaded as well as the time spent in port by 31 liquid-carrying vessels that used the port over the most recent summer are available. The data are available on the book website in the file glakes.txt. The first model fit to the data was

$Time = \beta 0 + \beta 1 Tonnage + e$

On the following pages is some output from fitting model (3.8) as well as some plots of Tonnage and Time (Figures 3.42 and 3.43).

(a)Does the straight line regression model (3.8) seem to fit the data well? If not, list any weaknesses apparent in model (3.8).

Answer-a: Looking at the plot of the model we see that the fitted model does not adequately describe the data specially after point x=10,000.The plot of standardized residuals shows that two points stand out of the other points ,and the variance tends to increase as x increases and the distribution is not quite normal.

(b)Suppose that model (3.8) was used to calculate a prediction interval for Time when Tonnage = 10,000. Would the interval be too short, too long or about right (i.e., valid)? Give a reason to support your answer.

Answer-b: The prediction interval is just a little bit longer than the points smaller than x=10000.It should have been wider because the variance of the original data increases as the x-variable increases .

The second model fitted to the data was $log(Time) = \beta 0 + \beta 1 Tonnage^{0.25} + e$

Output from model (3.9) as well as some plots (Figures 3.44 and 3.45) appears on the following pages.

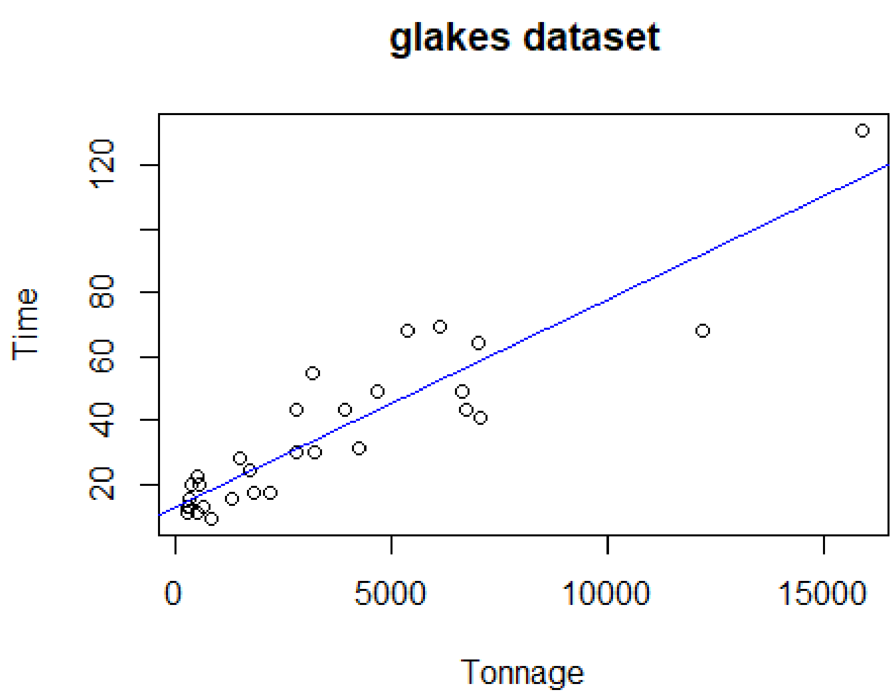
(a)Is model (3.9) an improvement over model (3.8) in terms of predicting Time? If so, please describe all the ways in which it is an improvement.

Answer-a: The prediction interval is wider after point 10,000 which means this model describes better the original data. In original scale the data has increasing variance along x-variable meaning that the realistic prediction intervals will get wider as x-variable increases. Therefore, this model will predict time better than the previous model. Also, here the variance of standard residuals are constant and have a normal distribution.

(b)List any weaknesses apparent in model (3.9).

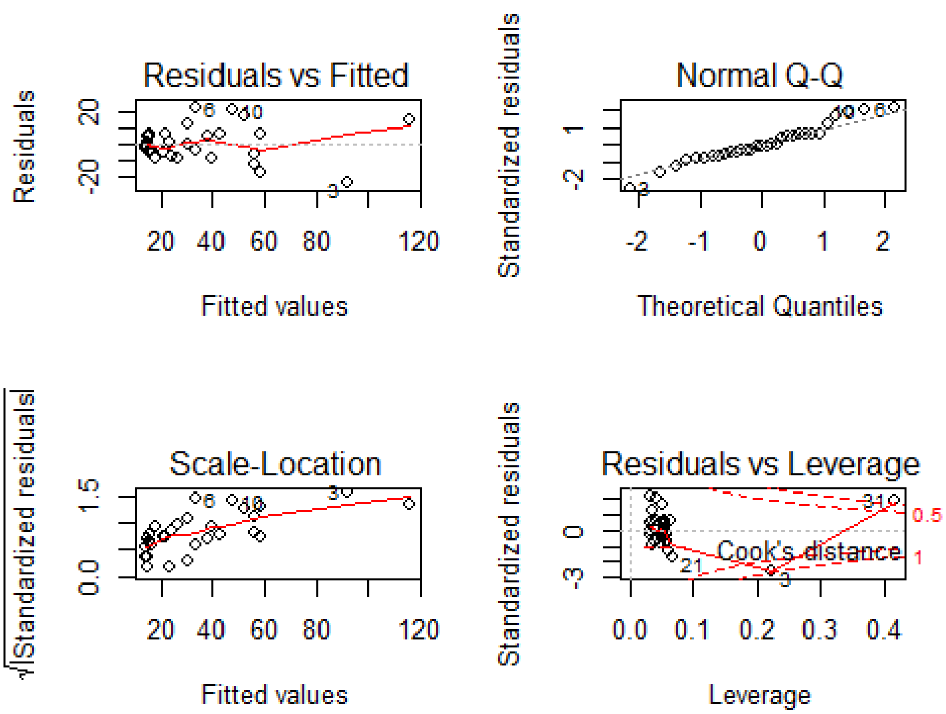
Answer-b: We have still an outlier.The R-squared decreased slightly in this model compared to that in the previous model.

```
setwd("C:/Users/")
glakes <- read.delim("glakes.txt", header = T, sep = "\t")

#head(airefares)
X_G=glakes$Tonnage
Y_G=glakes$Time
model1_G=lm(Y_G~X_G)
#plot the first model
par(mfrow = c(1,1))
plot(X_G,Y_G,xlab = "Tonnage", ylab="Time", main="glakes dataset")
abline(model1_G, col="blue")
```

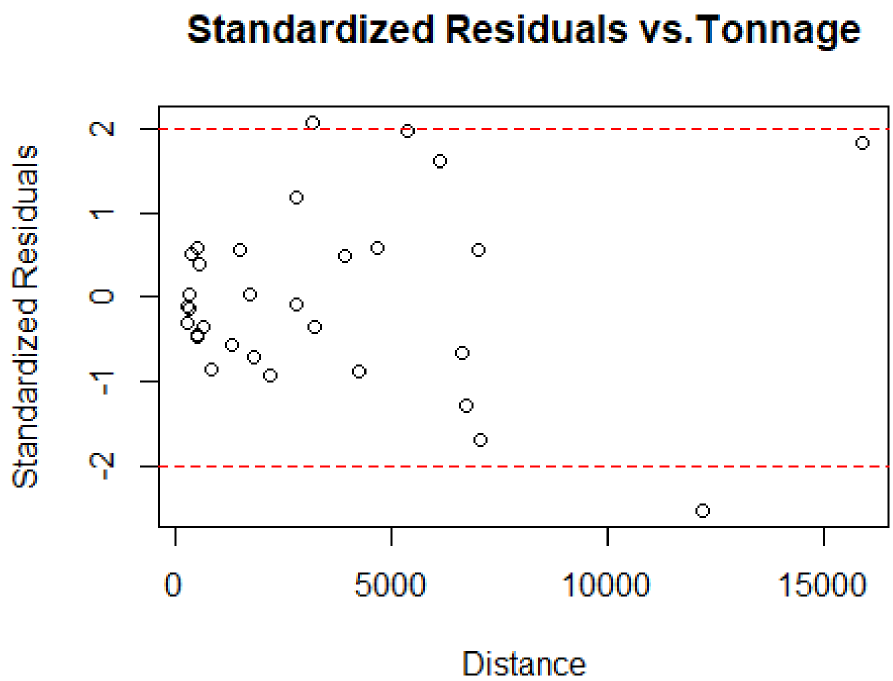



```
summary(model1_G)

##
## Call:
```

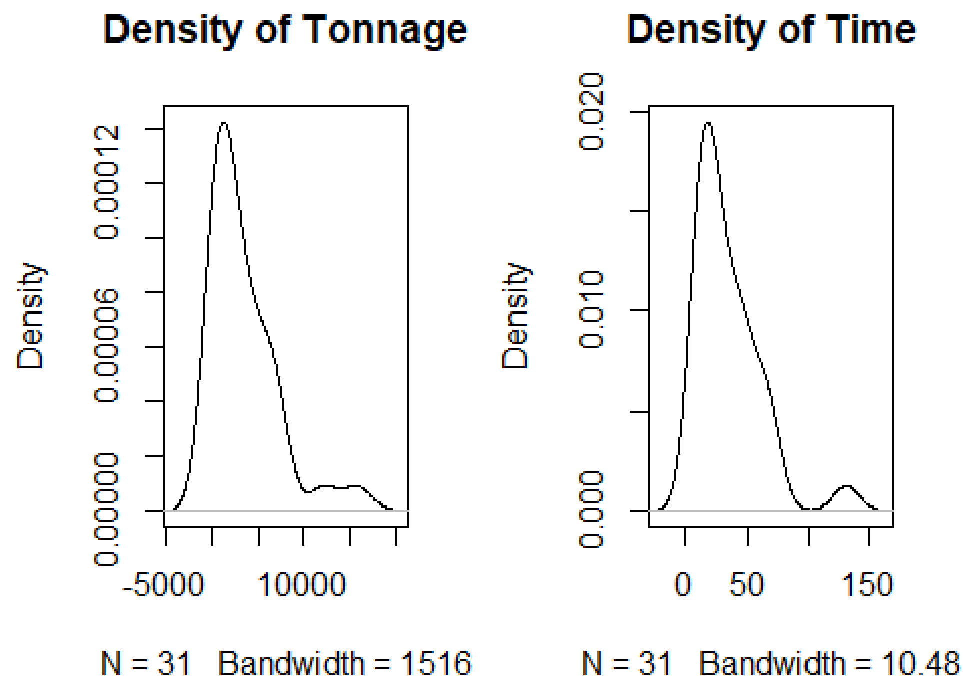
```
## lm(formula = Y_G ~ X_G)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.882  -6.397  -1.261   5.931  21.850
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.344707   2.642633   4.671 6.32e-05 ***
## X_G          0.006518   0.000531  12.275 5.22e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.7 on 29 degrees of freedom
## Multiple R-squared:  0.8386, Adjusted R-squared:  0.833
## F-statistic: 150.7 on 1 and 29 DF,  p-value: 5.218e-13

par(mfrow = c(2,2))
plot(model1_G)
```




```
#look at the standardized residuals
sdres_GM1 <- rstandard(model1_G)
par(mfrow = c(1,1))
plot(sdres_GM1~X_G,xlab="Distance", ylab = "Standardized Residuals", main =
"Standardized Residuals vs.Tonnage")
abline(h = 2, lty = 2, col = 2)
abline(h = -2, lty = 2, col = 2)
```

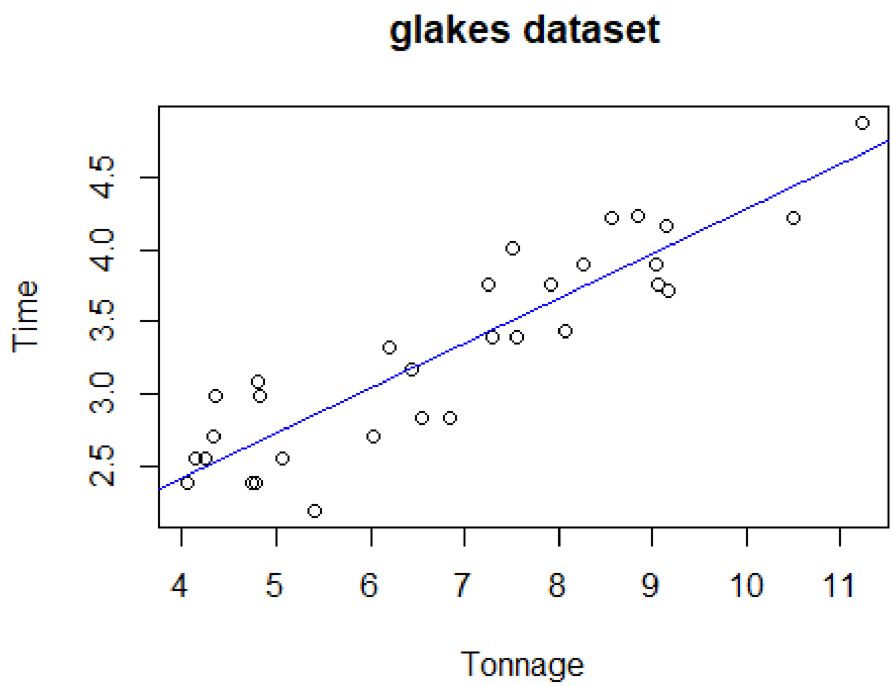



```
#check the normality of variables
par(mfrow = c(1,2))
plot(density(X_G), main ="Density of Tonnage")
plot(density(Y_G), main="Density of Time")
```

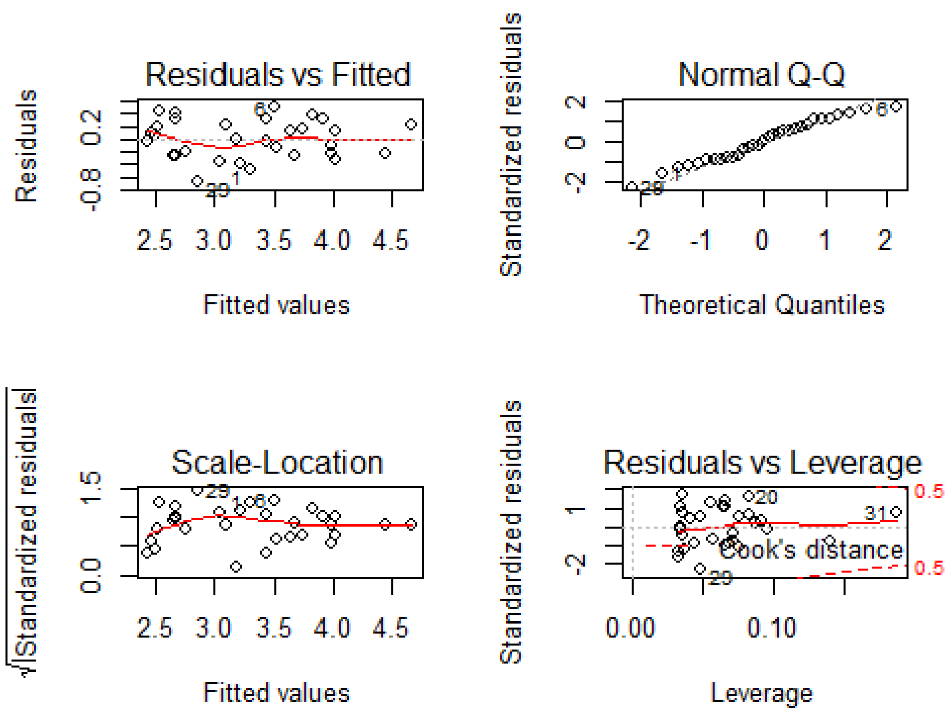



```
predict(model1_G,data.frame(X_G =10000), interval = "prediction",conf.int =
0.95)

##        fit      lwr      upr
## 1 77.5234 54.17047 100.8763

predict(model1_G,data.frame(X_G =4000), interval = "prediction",conf.int =
0.95)

##         fit      lwr      upr
## 1 38.41619 16.17566 60.65671

101-54

## [1] 47

61-16

## [1] 45

#-----------------------------------------------------------------------------
------

#head(airefares)
X_G1=(glakes$Tonnage)^0.25
Y_G1=log(glakes$Time)
model2_G=lm(Y_G1~X_G1)
#plot the first model
```

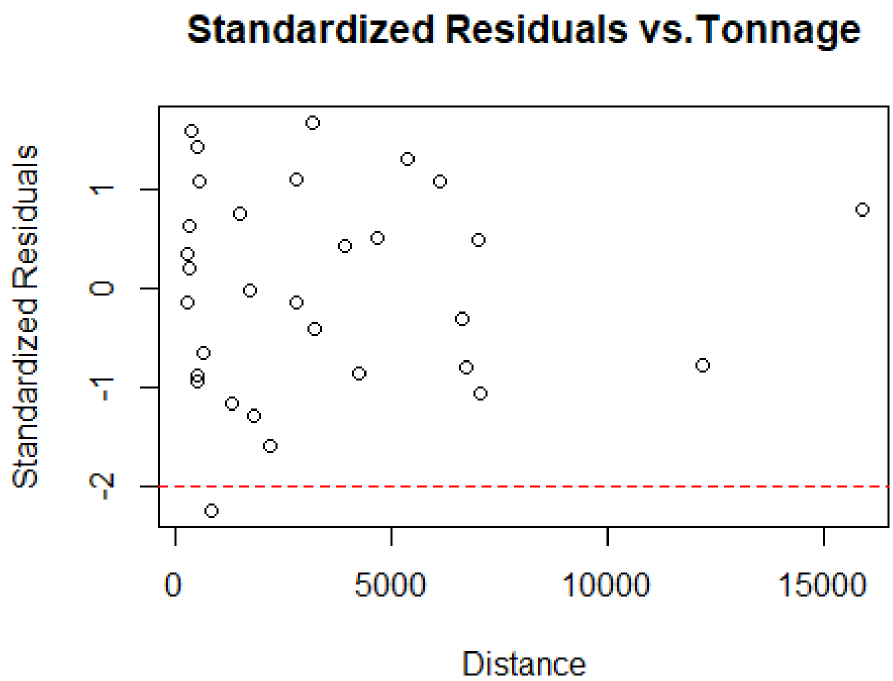
```r
par(mfrow = c(1,1))
plot(X_G1,Y_G1,xlab = "Tonnage", ylab="Time", main="glakes dataset")
abline(model2_G, col="blue")
```



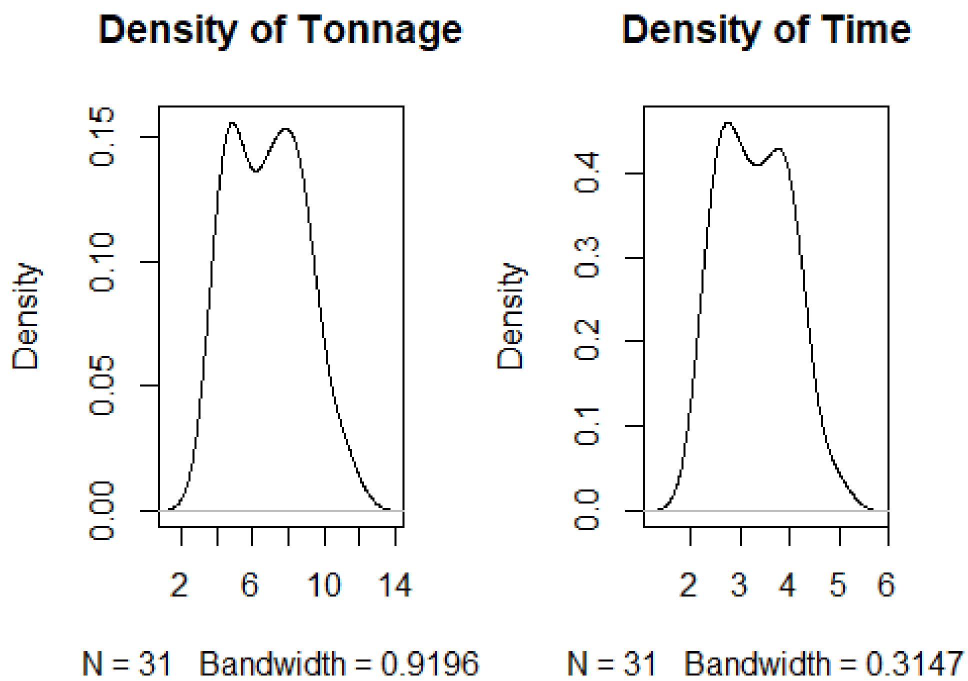

```r
summary(model2_G)

## 
## Call:
## lm(formula = Y_G1 ~ X_G1)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.6607 -0.2410 -0.0044  0.2203  0.4956 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.18842    0.19468   6.105  1.2e-06 ***
## X_G1         0.30910    0.02728  11.332  3.6e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3034 on 29 degrees of freedom
## Multiple R-squared:  0.8158, Adjusted R-squared:  0.8094 
## F-statistic: 128.4 on 1 and 29 DF,  p-value: 3.599e-12

par(mfrow = c(2,2))
plot(model2_G)
```




```r
#look at the standardized residuals
sdres_GM2 <- rstandard(model2_G)
par(mfrow = c(1,1))
plot(sdres_GM2~X_G,xlab="Distance", ylab = "Standardized Residuals", main =
"Standardized Residuals vs.Tonnage")
abline(h = 2, lty = 2, col = 2)
abline(h = -2, lty = 2, col = 2)
```




```
#check the normality of variables
par(mfrow = c(1,2))
plot(density(X_G1), main ="Density of Tonnage")
plot(density(Y_G1), main="Density of Time")
```




```
10000^0.25

## [1] 10

predict(model2_G,data.frame(X_G1 =10), interval = "prediction",conf.int =
0.95)

##        fit      lwr      upr
## 1 4.279393 3.624927 4.933859

exp(4.933859)-exp(3.624927)

## [1] 101.3926
```