Exam 1

Biplav Timalsina

April 01, 2018

STAT 757 Applied Regression Analysis

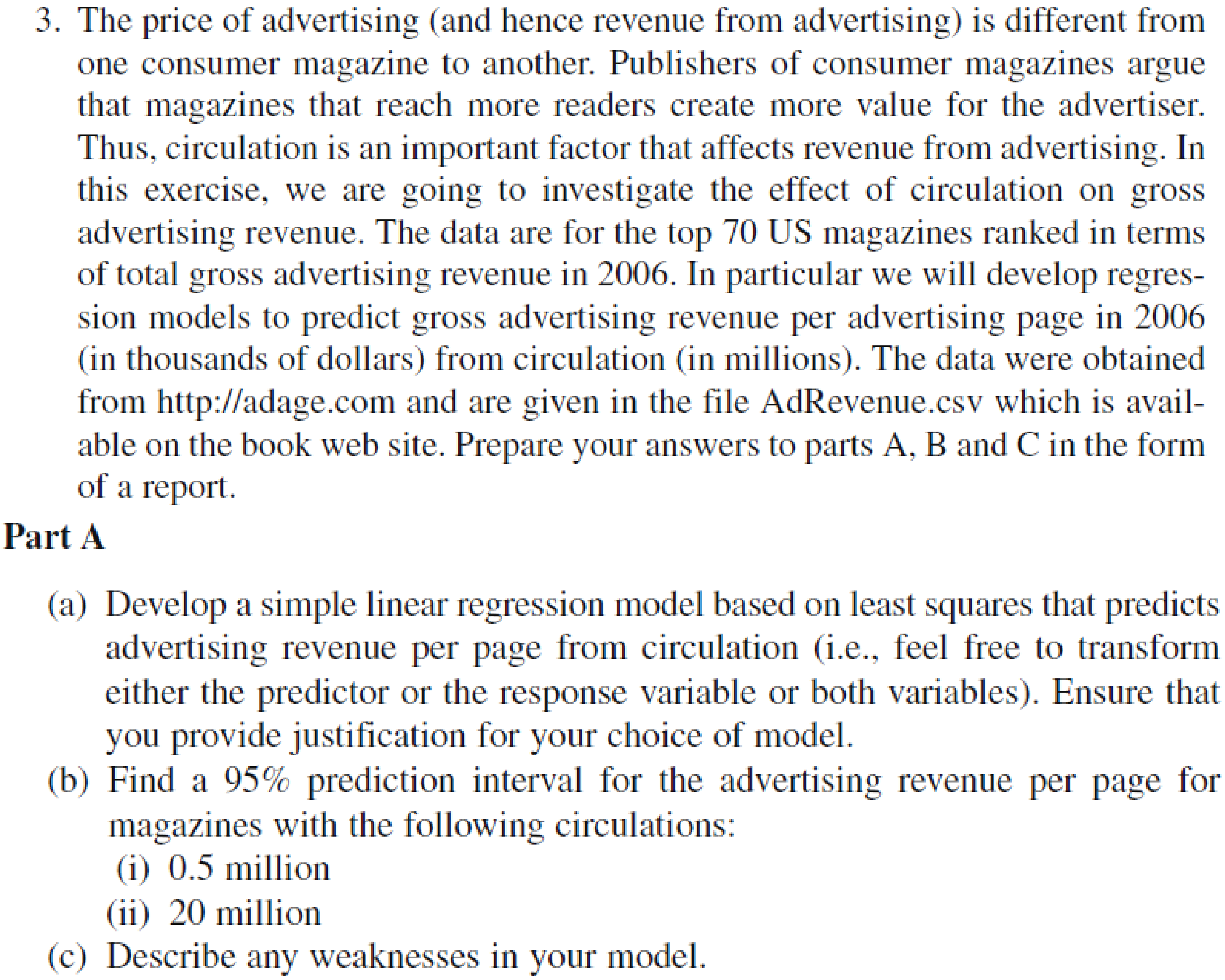
Take-home Exam Procedures

You may use your notes, textbook, and R while taking this exam. You are free to look online for definitions, however searching directly for answers to the given questions will be treated as cheating and dealt with according to UNR's policies regarding Academic Dishonesty. You are encouraged to ask questions to the instructor during the exam (but only hints will be given!) Partial credit WILL be awarded where sufficient details have been provided. This exam must be completed **individually** (you may not discuss any aspect of the exam with your classmates).

Modify this `.Rmd` file to produce a **report** that addresses the parts indicated in Exercise 3.4.3 in sheather2009 for `AdRevenue.csv` dataset. You can find some helpful code here: http://www.stat.tamu.edu/~sheather/book/docs/rcode/Chapter10.R. Please email **both** your .Rmd (or roxygen .R) and one of the following either .HTML, .PDF, or .DOCX using the format `SURNAME-FIRSTNAME-Exam1.Rmd` and `SURNAME-FIRSTNAME-Exam1.pdf`.

By a *report*, I mean that disjointed answers are not acceptable. Synthesize the parts into a coherent story. Be sure, however, to address all parts indicated in the question as you are responsible for everything that is asked. Make your report reader friendly - **concise**, but **informative** and conveying **mastery** of course concepts and skills. Please see `STAT_757_Exam1_rubric.pdf` for how you will be assessed.

References



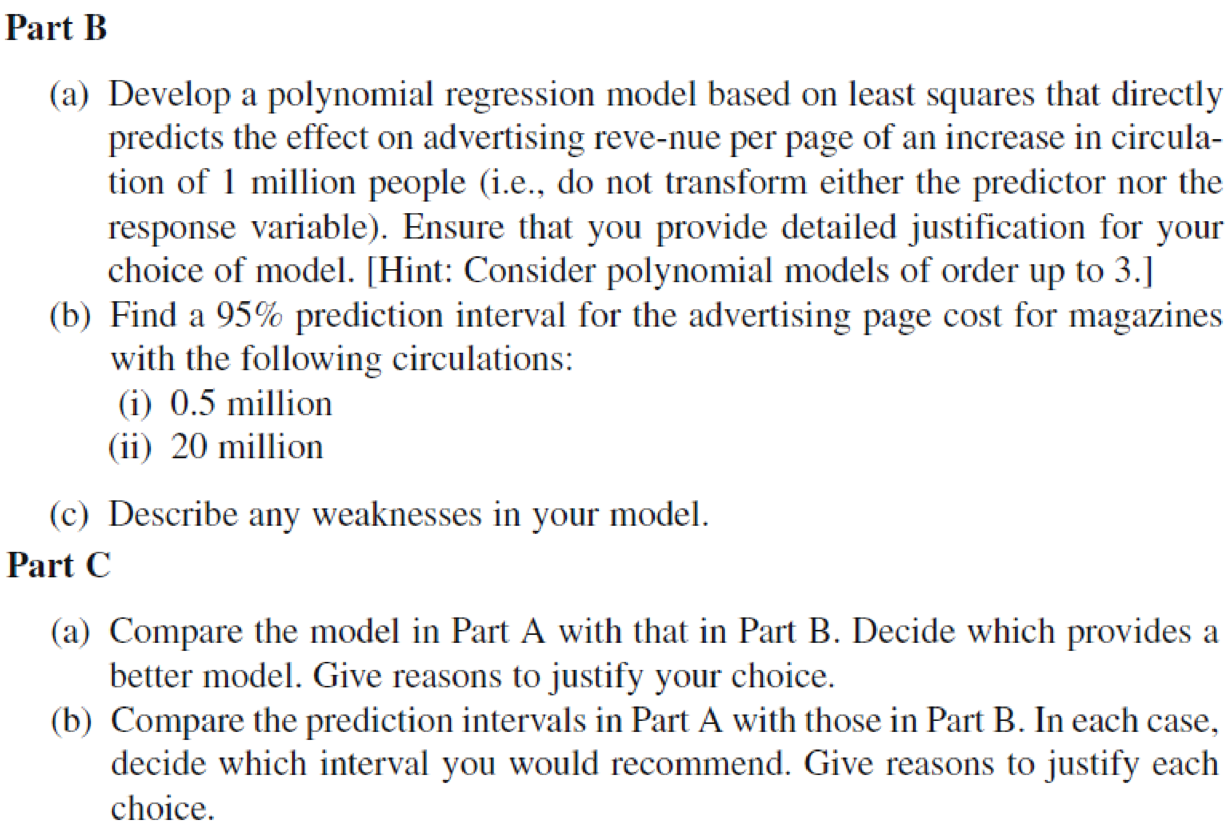
3. The price of advertising (and hence revenue from advertising) is different from one consumer magazine to another. Publishers of consumer magazines argue that magazines that reach more readers create more value for the advertiser. Thus, circulation is an important factor that affects revenue from advertising. 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.



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.

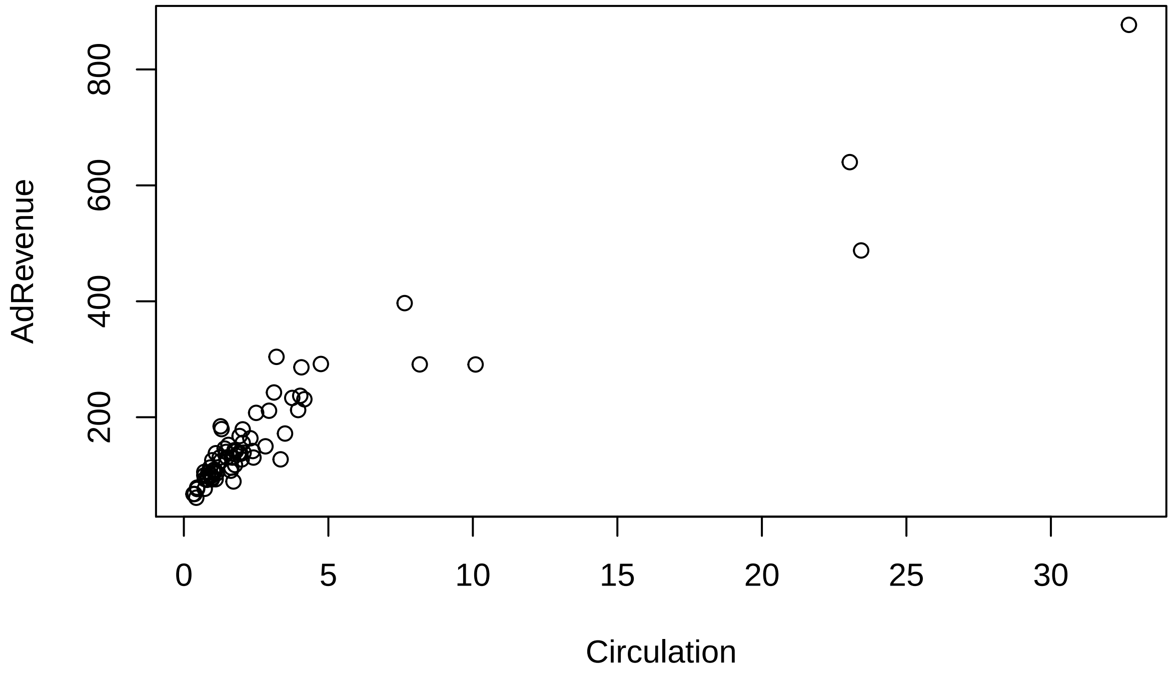
Part C

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

(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.

In any data analysis, we should start by first plotting our data in order to have an idea of any obvious relationship present. Below we can see scatter plot and line of best fit.

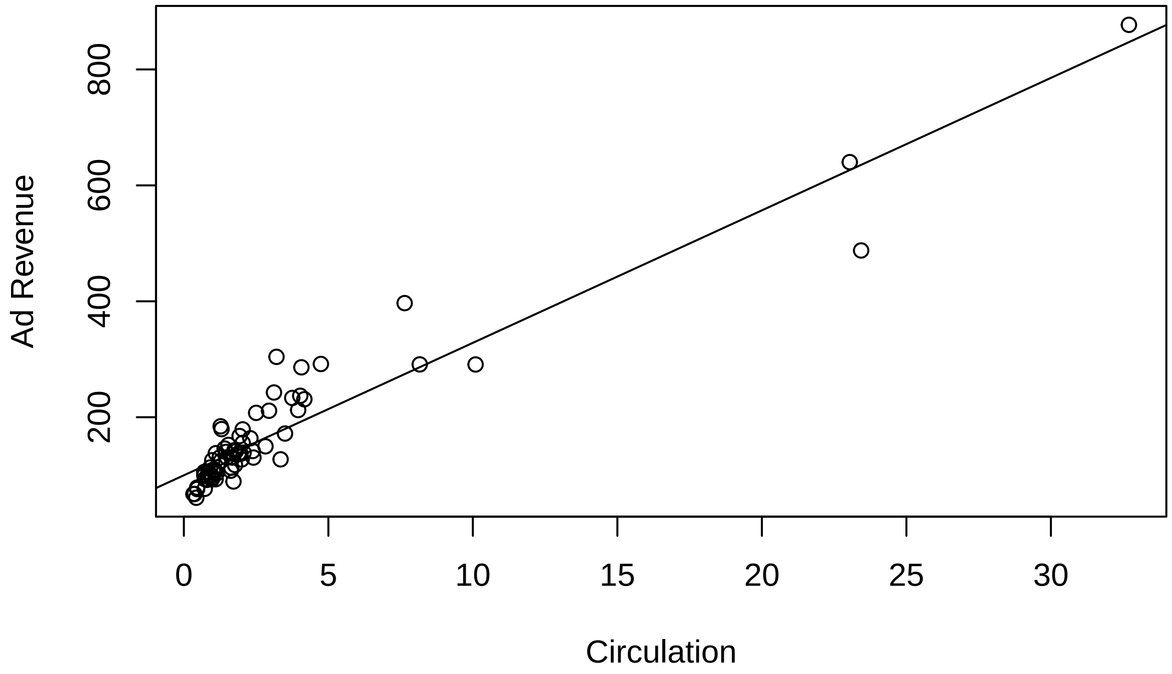
```r
my_data_path <- "F:/unr/4th sem/applied regression analysis/Assignments/Exam1"
adData <- read.csv(file.path(my_data_path,"AdRevenue.csv"),header=TRUE)
attach(adData)

#Figure 2.6 on page 38
plot(adData$Circulation,adData$AdRevenue,xlab="Circulation",
ylab="AdRevenue")
```



From the figure above, we can see that the data points are highly concentrated around a fixed area, and there are other data points that are very scattered. If we draw a linear regression line, we will have a few leverage points. I think that will not be good. Therefore, we might need to perform some transformation.

Let us now try a linear model. We can use functions available at R to construct linear models.

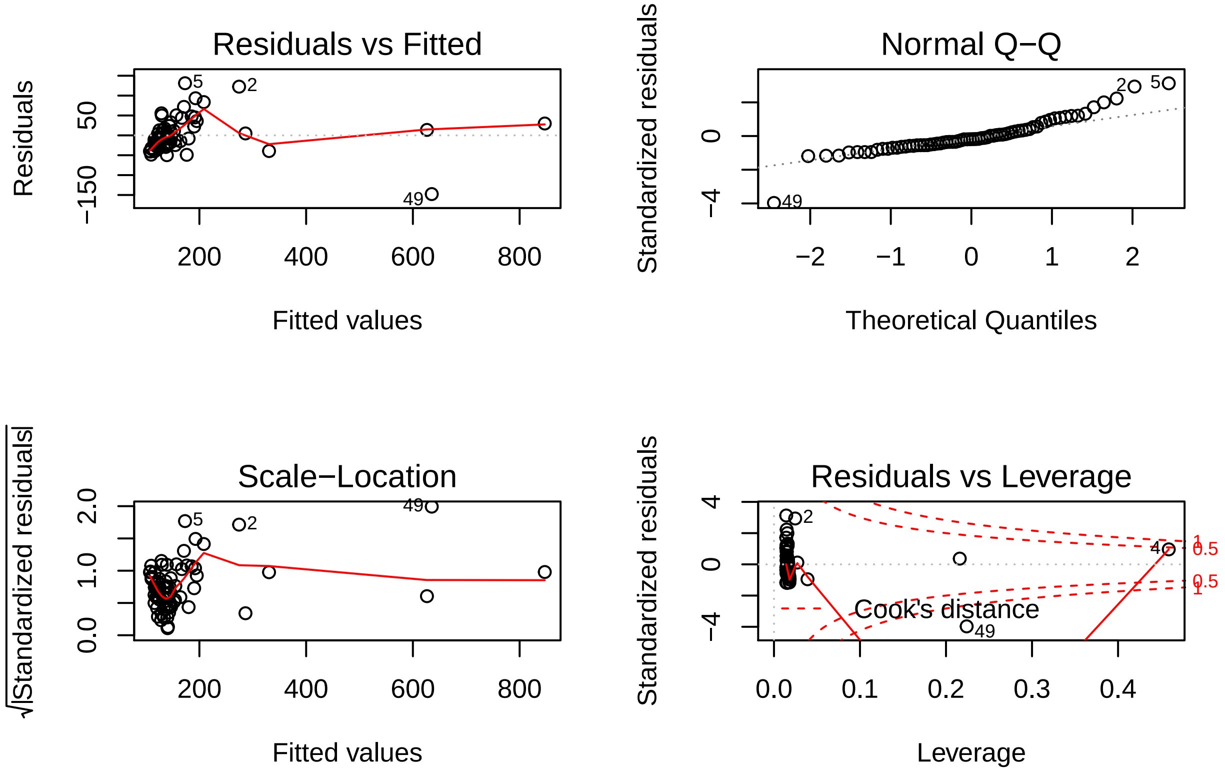
```
par(mfrow=c(1,1))
plot(adData$Circulation,adData$AdRevenue,xlab="Circulation",ylab="Ad Revenue")
abline(lsfit(adData$Circulation,adData$AdRevenue))
```



```
model<-lm(adData$AdRevenue~adData$Circulation)
summary(model)
```

```
##
## Call:
## lm(formula = adData$AdRevenue ~ adData$Circulation)
##
## 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 ***
## adData$Circulation  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
```

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



From the above analysis, we can see abrupt change in the standardized residuals in the bottom left hand figure. Similarly, we can see that the residuals in the top left are changing. This violates our assumption of constant variance. Also the Q-Q plot indicates that the residuals are not normally distributed. Thus, this model might not be appropriate model for our data.

From our observations so far, we derived that we may need to perform some transformation to our data in order to generate a good model.

Next, I have compiled some statistics for different models, log and square root transformation,and combinations of them. The exact formula and figures can be found at Appendix 1.

From Appendix 1,

	Model	Residual	Multiple R-Squared	F-statistic
1	$y = -29.45 + 132.91\sqrt{X}$	41.2	0.901	304
2	$\sqrt{y} = 10.55 + 0.633\sqrt{x}$	1.737	0.79	261.4
3	$\sqrt{y} = 6.66 + 3.88\sqrt{x}$	1.23	0.89	579.6
4	$log(y) = 4.67 + 0.528log(x)$	0.17	0.881	503.6
5	$y = 98.63 + 123.55log(x)$	65.24	0.748	202
6	$log(y) = 4.74 + 0.07x$	0.307	0.639	120.9
7	$log(\sqrt{y}) = 2.33 + 0.52log(\sqrt{x})$	0.088	0.88	503.6

It seems that log-log transformation and log(Sqrt)-log(sqrt) transformation are almost similar. But since the Residual Standard error for Log(sqrt)-log(sqrt) is lower than that of Log-log transformation, this means that the residual errors in log(sqrt) is smaller than that of log transformation. Hence, numerically I would prefer logSqrt. log(Sqrt(AdReveue))= 2.34 + 0.52 *log(Sqrt(Circulation))

Although, logSqrt transformation is numerically slightly better, for ease of interpretation, I might go with simply log-log transformation too. log(AdRevenue)= 4.67 + 0.52 * log(Circulation)

Let us now find Prediction interval for 0.5 and 20 million circulation (using only log-log transformation model) that we just selected. From Appendix 1,

For 0.5 (=-0.691 in log scale) million spent in circulation, we have in log scale (3.94, 4.66) which when changed to normal scale is (51.41,105.63) in thousands.

For 20 million(=2.99 in log scale) million spent in circulation, we have in log scale (5.88, 6.62) which when changed to normal scale is (357.8,749.94) in thousands.

Let us now analyse the model in terms of weakness. Although log-log model was the model that we selected, and it is best that we have found so far among different combinations of transformation we have tested, but we can still see that variance is slightly increasing for this model too. There are a few outliers which have not been addressed seperately in the model. We could infact do a closer analysis to reveal any important insight on these outlier cases. Addressing these issues by removing them altogether or creating a seperate model for them could infact lead to better model.

How about we try some polynomial regression model? It could be a good fit to our model. Let us now try to develop polynomial regression model upto order of 3. We can develop the following polynomial models. Refer Appendix 2 for the exact code and figures.

From Appendix 2,

	Model	Residual	Multiple R-Squared	F-statistic
1	$y = 88.13 + 29.5x - 0.23x^2$	41.2	0.901	304
2	$y = 59.17 + 51.23x - 2.5x^2 + 0.05x^3$	34.06	0.93	308.1
3	$y = 95.93 + 24x - 0.0024x^3$	42.25	0.8959	288.3
4	$y = 136.57 + 1.61x^2 - 0.02x^3$	56.49	0.813	146.5
5	$y = 150 + 0.02x^3$	69.83	0.7114	167.6
6	$y = 143.12 + 0.73x^2$	60.57	0.7829	245

After comparing all the models and the plots of sqroot of standardized residuals, we observe that none of the models produce stable variance. But if we have to select one among them, because the second model is the one with most stable variance in the region of maximum points, I would select it.

From Appendix 2,

The 95% prediction interval adRevenue when circulation= 0.5 million, is (14.92, 153.41) thousands.

The 95% prediction interval for adRevenue when circulation=20 million is (418.179,580.8878)

Again, lets try to analyse this model in terms of its weakness. We can see from the plot of Sqroot of Standardized residual that the variance is not constant for any of the models, and we just the best among these models. So, it violates our assumption of constant variance, which in itself is sufficient to not regard this model as proper model to predict any prediction.

There are a few outliers and leverage points in this case which has not been addressed by the model by either removing them or studying them seperately.

Conclusion:

From the above analysis, we can see that the variance for Log-log transformed model is more stable than that of the polynomial case. From this reason only, we should be able to say that log-log transformed model is better than polynomial regression.

Also F-statistic for Log-log transformed model is larger than its counterpart, hence log-log transformed model is better.

Also,

The prediction interval for Log-log transformed model is: For 0.5 circulation, (51.41,105.63), and for 20 million circulation, (357.8,749.94)

For polynomial regression, For 0.5 M circulation (14.92,153.41) and for 20 million circulation, (418.179, 580.78)

Since the prediction interval for 0.5 million is larger for polynomial regression (and the larger the prediction interval, the greater the chance our predicted value lies inside the range), I would select polynomial regression range for 0.5 Million case.

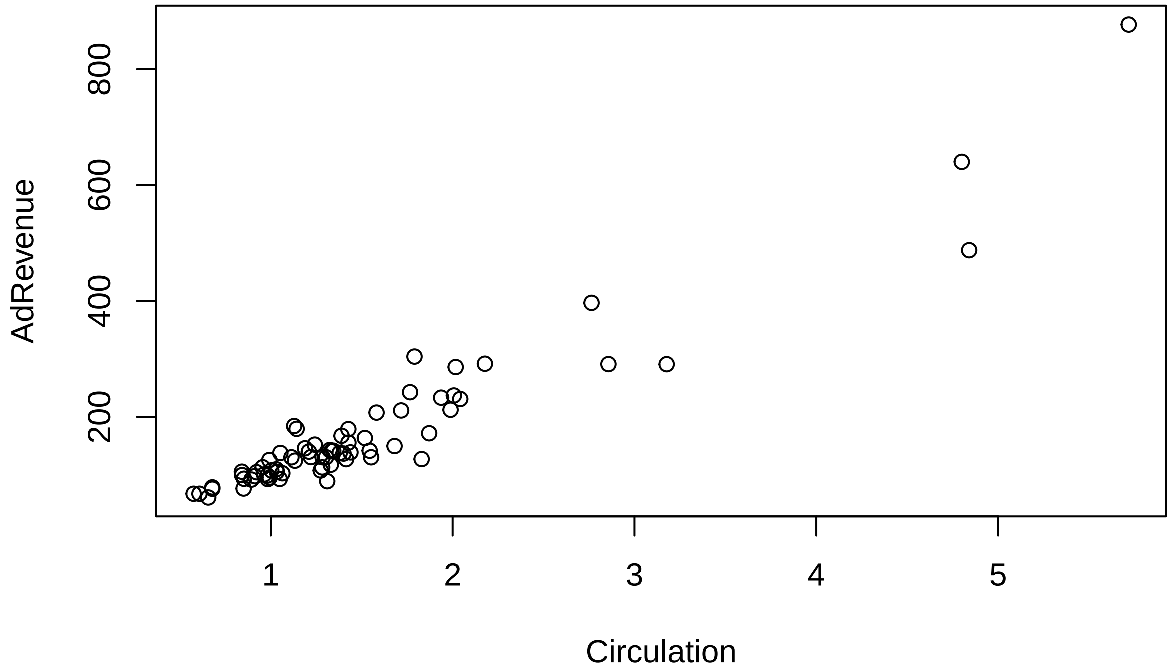
Since the prediction interval for 20 million is larger for log-log transformation (and for the same reason as above), I would select log-log transformed model for 20 million circulation.

Appendix 1

Since the units of measurement is in dollars for both dependent and independent variables, we might use square-root transformation.

Let us first apply square-root transformation to circulation.

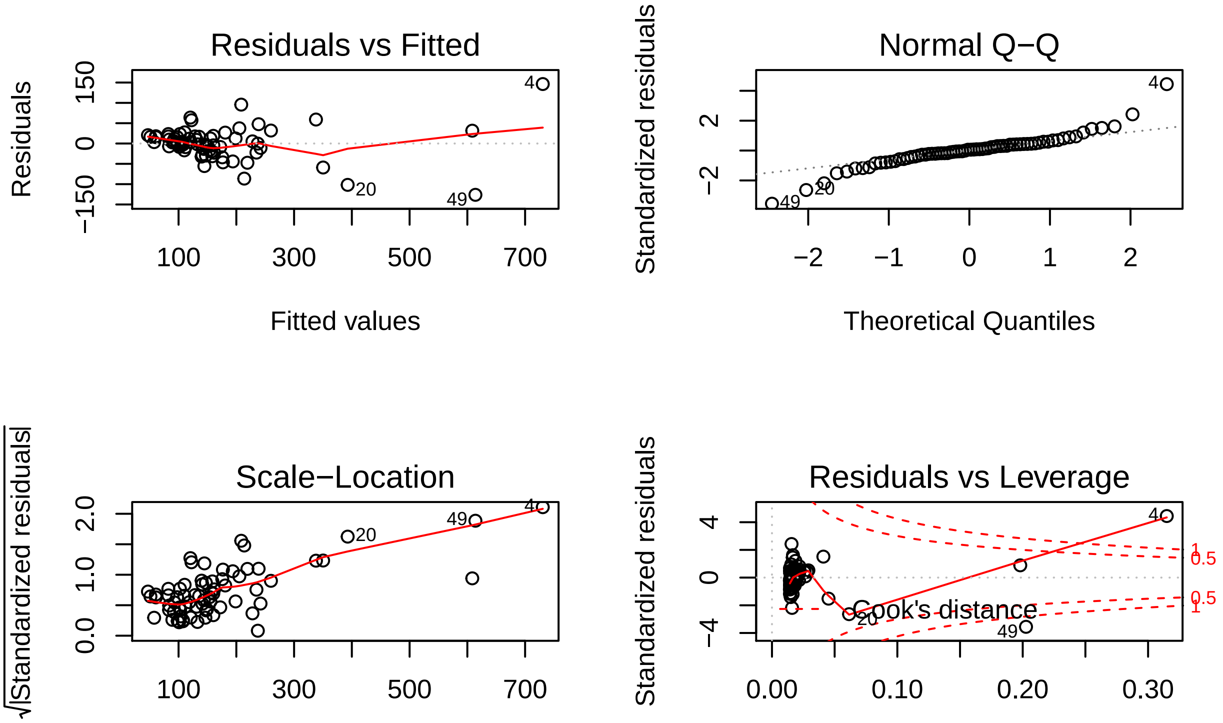
```
sqrtCirculation <- sqrt(adData$Circulation)
plot(sqrtCirculation,adData$AdRevenue,xlab="Circulation",
ylab="AdRevenue")
```



```
#sqrtrooms <- sqrt(Rooms)
m2 <- lm(adData$AdRevenue~sqrtCirculation)
summary(m2)
```

```
##
## Call:
## lm(formula = adData$AdRevenue ~ sqrtCirculation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.300  -15.177    2.232   17.077  146.310
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -29.458      9.149   -3.22  0.00197 **
## sqrtCirculation  132.914      5.181   25.66  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.78 on 68 degrees of freedom
## Multiple R-squared:  0.9064, Adjusted R-squared:  0.905
## F-statistic: 658.2 on 1 and 68 DF,  p-value: < 2.2e-16
```

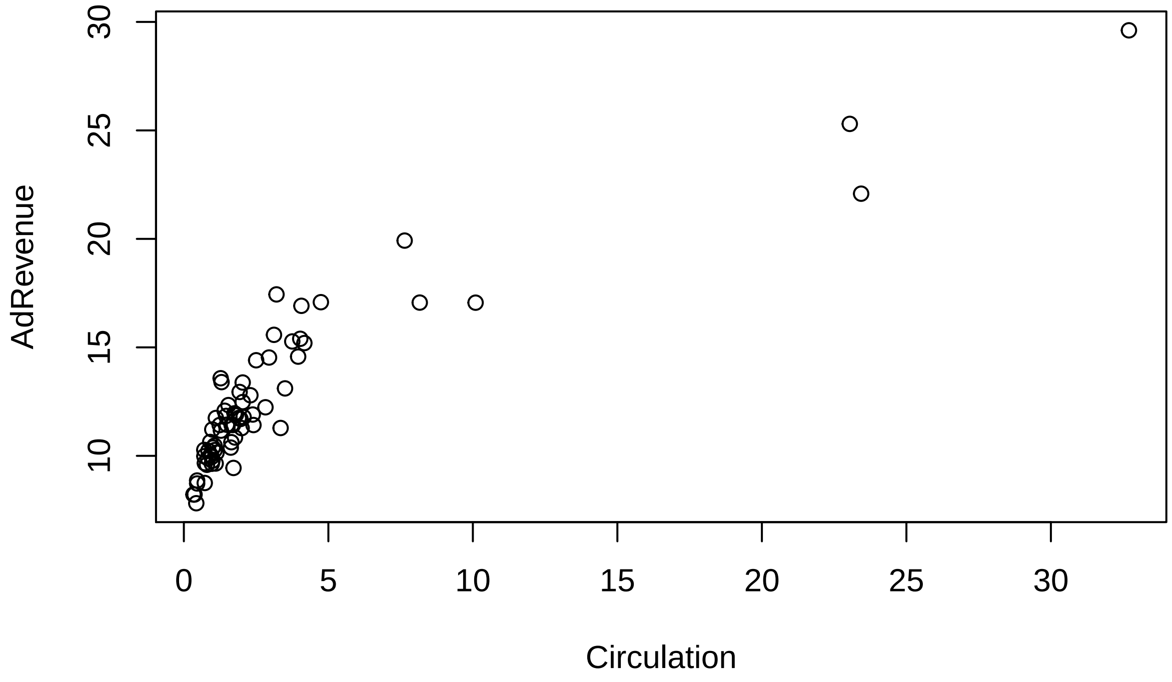
```
par(mfrow=c(2,2))
plot(m2)
```



We can see from the bottom left figure that the variance has further been stabilized in the (0-200) region of fitted values, but has been destabilized in bigger fitted values.

Lets us now try square root transformation of adRevenue and original values of Circulation.

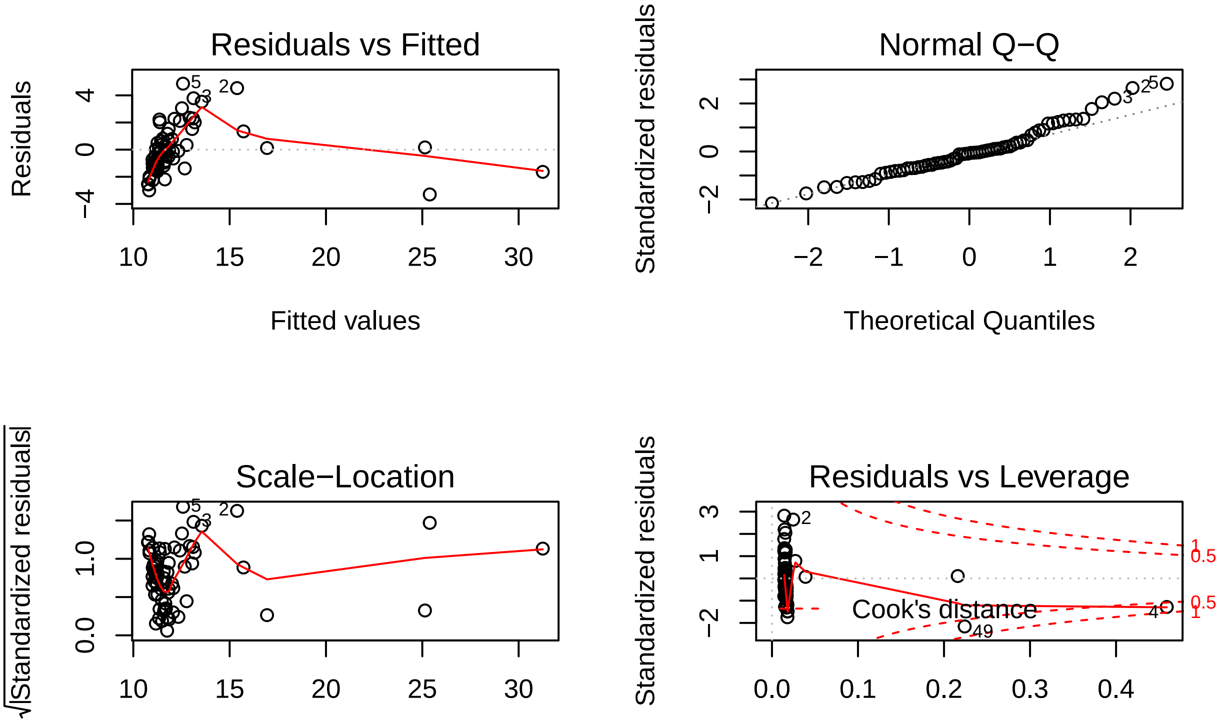
```
sqrtAdRevenue <- sqrt(adData$AdRevenue)
plot(adData$Circulation,sqrtAdRevenue,xlab="Circulation",
ylab="AdRevenue")
```



```
#sqrtrooms <- sqrt(Rooms)
m3 <- lm(sqrtAdRevenue~adData$Circulation)
summary(m3)
```

```
##
## Call:
## lm(formula = sqrtAdRevenue ~ adData$Circulation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3035 -1.1714 -0.1274  0.7487  4.8610
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        10.55089    0.24085   43.81   <2e-16 ***
## adData$Circulation  0.63308    0.03916   16.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.737 on 68 degrees of freedom
## Multiple R-squared:  0.7936, Adjusted R-squared:  0.7905
## F-statistic: 261.4 on 1 and 68 DF,  p-value: < 2.2e-16
```

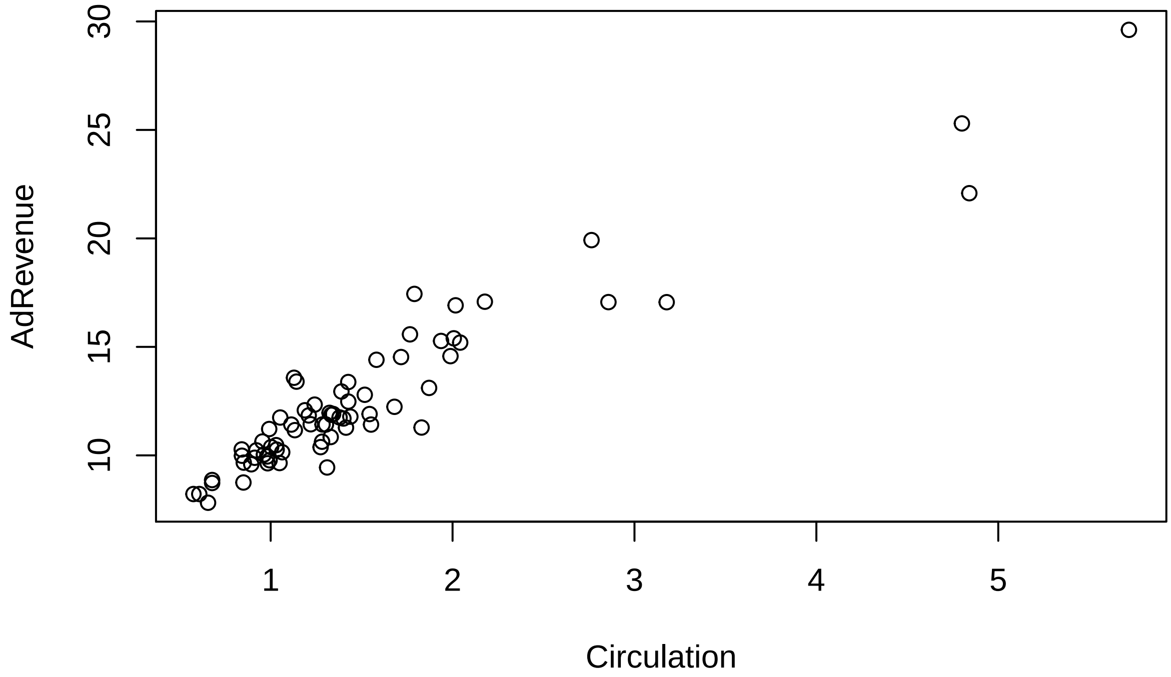
```
par(mfrow=c(2,2))
plot(m3)
```



It seems the variance has been further destabilized from this transformation as can be seen from plot of Sqrt Standardized residuals and fitted values at bottom left. So, we don't favor this model.

Now, lets try to transfrom both variables by sqrt transformation.

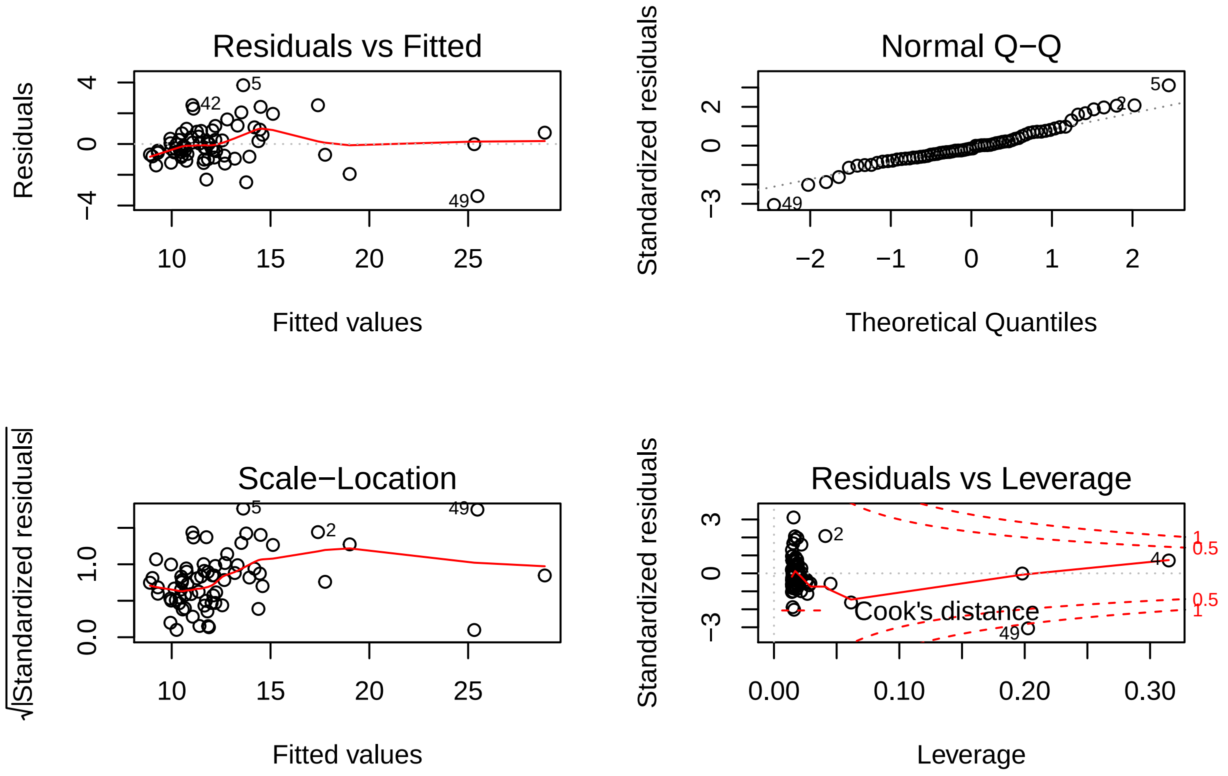
```
sqrtAdRevenue <- sqrt(adData$AdRevenue)
sqrtCirculation <- sqrt(adData$Circulation)
plot(sqrtCirculation,sqrtAdRevenue,xlab="Circulation",
ylab="AdRevenue")
```



```
#sqrtrooms <- sqrt(Rooms)
m4 <- lm(sqrtAdRevenue~sqrtCirculation)
summary(m4)
```

```
##
## Call:
## lm(formula = sqrtAdRevenue ~ sqrtCirculation)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3857 -0.7370 -0.1861  0.6731  3.8223
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       6.6643     0.2849   23.39   <2e-16 ***
## sqrtCirculation   3.8845     0.1614   24.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.239 on 68 degrees of freedom
## Multiple R-squared:  0.895,  Adjusted R-squared:  0.8935
## F-statistic: 579.6 on 1 and 68 DF,  p-value: < 2.2e-16
```

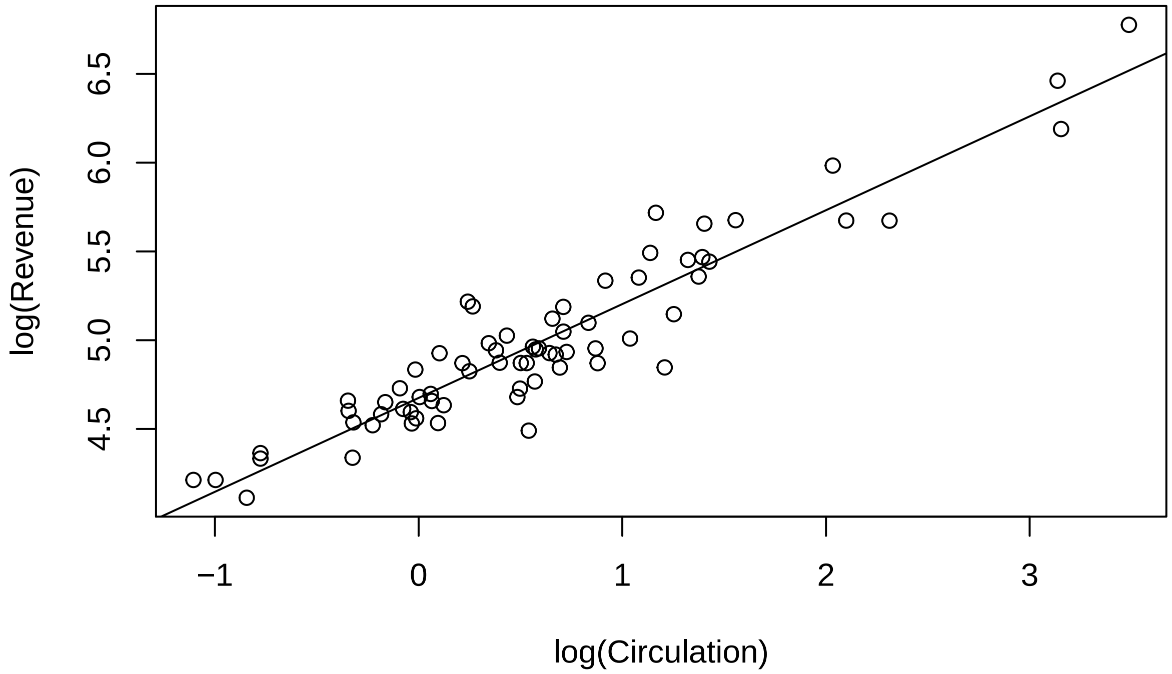
```
par(mfrow=c(2,2))
plot(m4)
```



It is comparatively better than the previous models.

Let us now use Log log transformation.

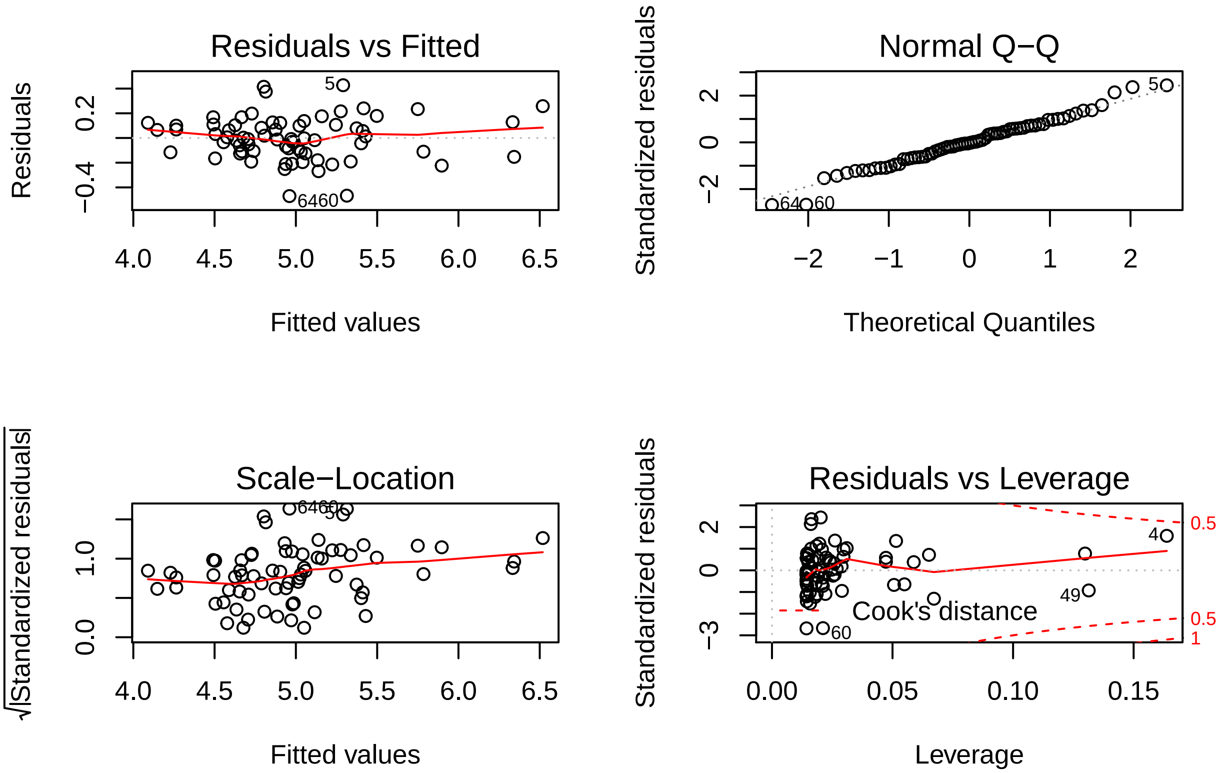
```
 #Regression output on page 82
plot(log(adData$Circulation),log(adData$AdRevenue),xlab="log(Circulation)",ylab="log(Revenue)")
abline(lsfit(log(adData$Circulation),log(adData$AdRevenue)))
```



```r
m5 <- lm(log(adData$AdRevenue)~log(adData$Circulation))
summary(m5)
```

```
##
## Call:
## lm(formula = log(adData$AdRevenue) ~ log(adData$Circulation))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47022 -0.11142 -0.00532  0.10835  0.42705
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.67473    0.02525  185.16   <2e-16 ***
## log(adData$Circulation)  0.52876    0.02356   22.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1768 on 68 degrees of freedom
## Multiple R-squared:  0.881,  Adjusted R-squared:  0.8793
## F-statistic: 503.6 on 1 and 68 DF,  p-value: < 2.2e-16
```

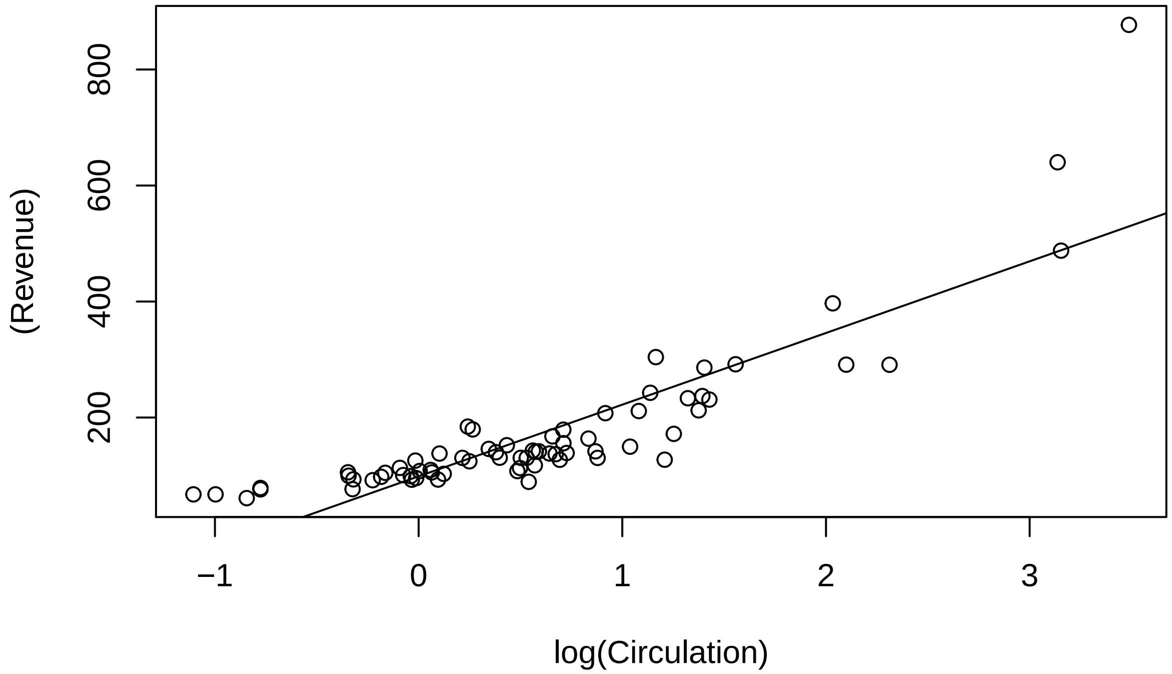
```r
par(mfrow=c(2,2))
plot(m5)
```



We can see that the data points are not distributed randomly across the line of fit. Also, now this is tremendous stabilization of our variance so far.

Let us try with only one variable log-transformed.

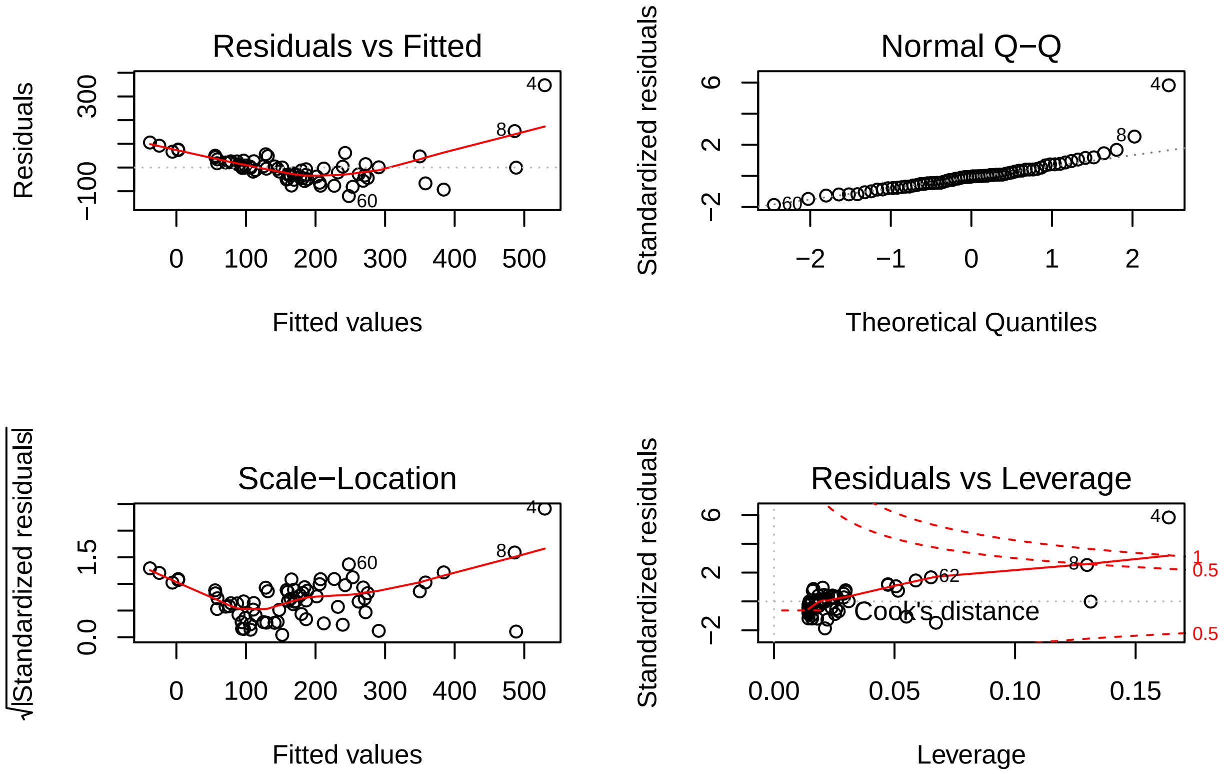
```
 #Regression output on page 82
plot(log(adData$Circulation),(adData$AdRevenue),xlab="log(Circulation)",ylab="(Revenue)")
abline(lsfit(log(adData$Circulation),(adData$AdRevenue)))
```



```
m6 <- lm((adData$AdRevenue)~log(adData$Circulation))
summary(m6)
```

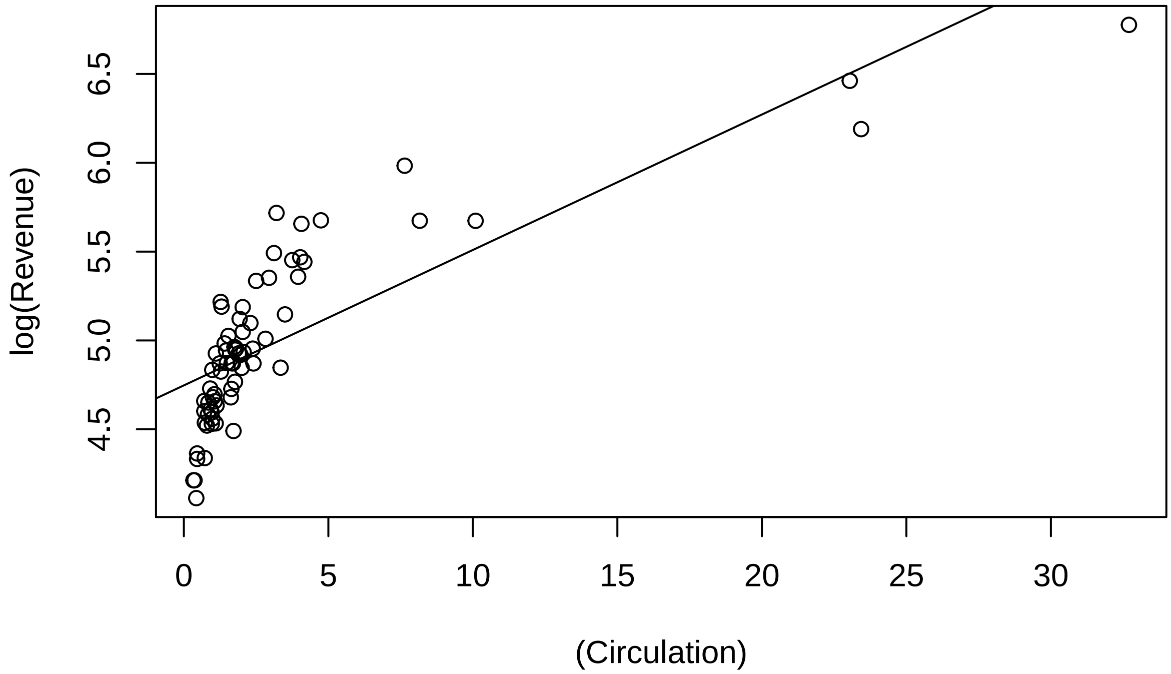
```
##
## Call:
## lm(formula = (adData$AdRevenue) ~ log(adData$Circulation))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -120.58  -36.92   -3.00   24.94  347.40
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)               98.631      9.316   10.59 5.04e-16 ***
## log(adData$Circulation)  123.554      8.694   14.21  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.24 on 68 degrees of freedom
## Multiple R-squared:  0.7481, Adjusted R-squared:  0.7444
## F-statistic:   202 on 1 and 68 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m6)
```



It is not better compared to both variable transformed by log transformation.

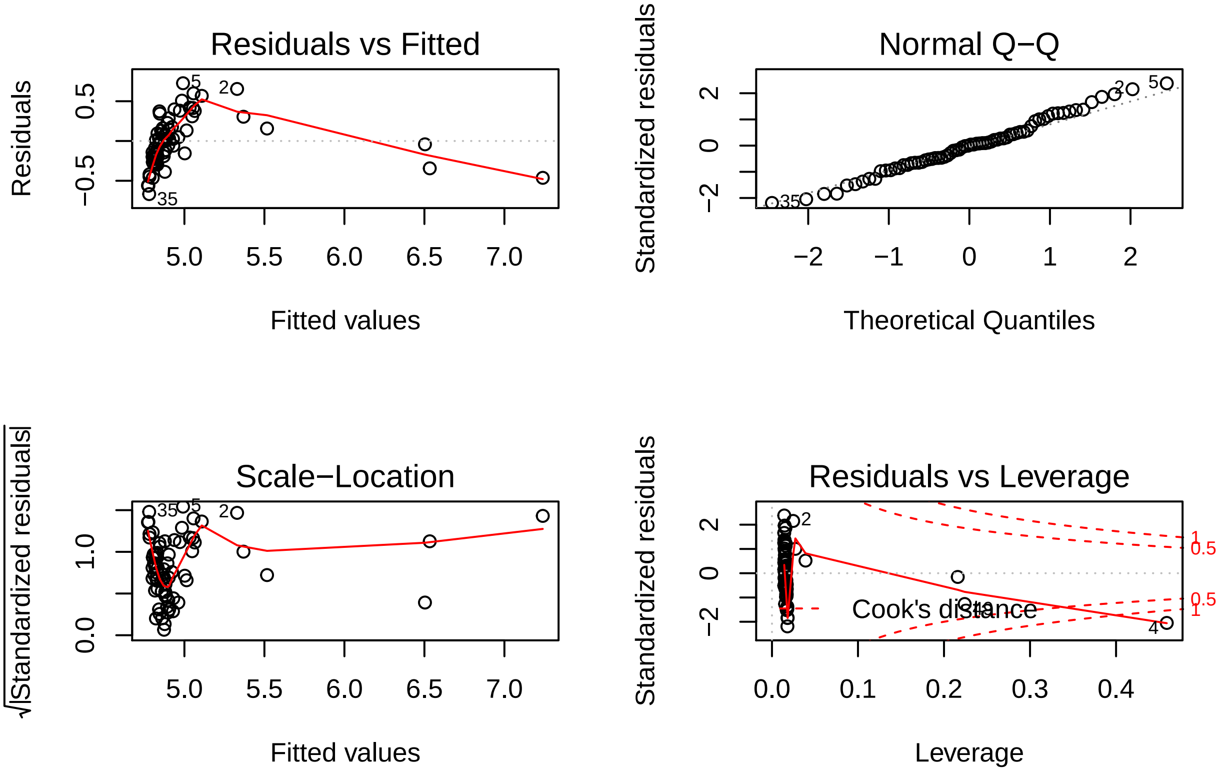
```
 #Regression output on page 82
plot((adData$Circulation),log(adData$AdRevenue),xlab="(Circulation)",ylab="log(Revenue)")
abline(lsfit((adData$Circulation),log(adData$AdRevenue)))
```



```
m7 <- lm(log(adData$AdRevenue)~(adData$Circulation))
summary(m7)
```

```
##
## Call:
## lm(formula = log(adData$AdRevenue) ~ (adData$Circulation))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6673 -0.1994  0.0055  0.1600  0.7263
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.747059   0.042652  111.30   <2e-16 ***
## adData$Circulation 0.076228   0.006934   10.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3076 on 68 degrees of freedom
## Multiple R-squared:  0.6399, Adjusted R-squared:  0.6346
## F-statistic: 120.9 on 1 and 68 DF,  p-value: < 2.2e-16
```

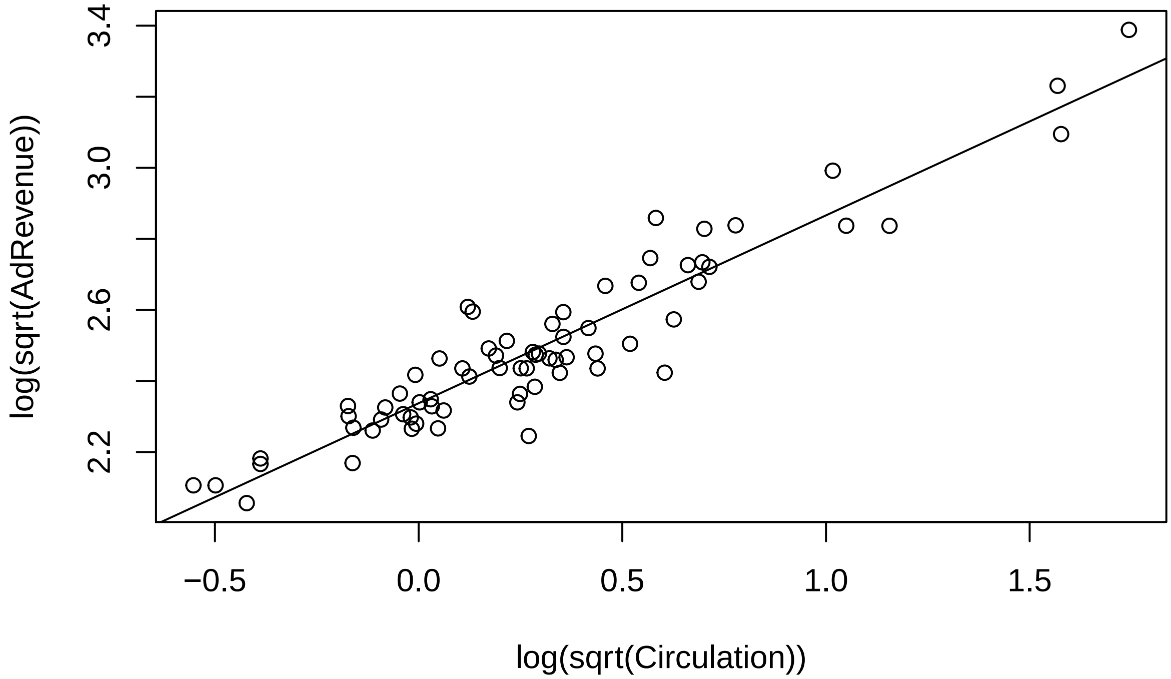
```
par(mfrow=c(2,2))
plot(m7)
```



This is also really bad. So, up until this point our best fit is log-log transformation.

Let us finally try log transformation of sqrt transformation.

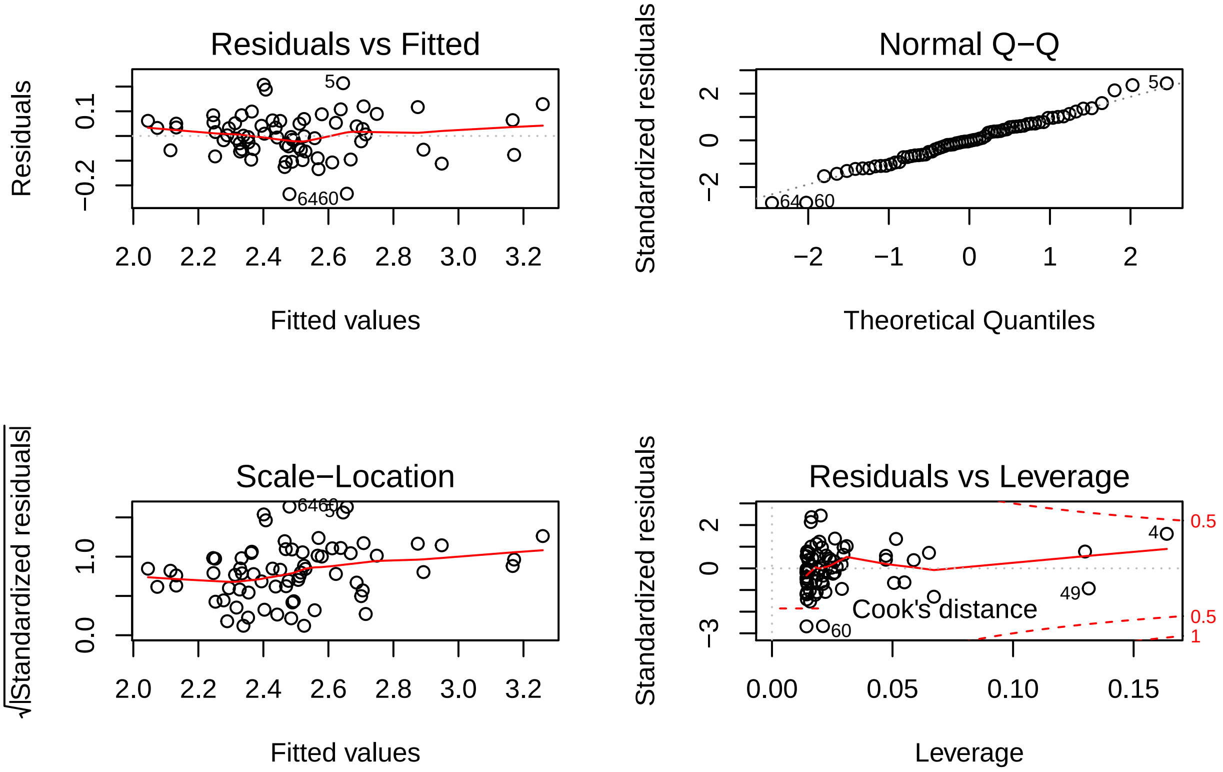
```
sqrtAdRevenue <- sqrt(adData$AdRevenue)
sqrtCirculation <- sqrt(adData$Circulation)
plot(log(sqrtCirculation),log(sqrtAdRevenue),xlab="log(sqrt(Circulation))",
ylab="log(sqrt(AdRevenue))")
abline(lsfit((log(sqrtCirculation)),log(sqrtAdRevenue)))
```



```
#sqrtrooms <- sqrt(Rooms)
m8 <- lm(log(sqrtAdRevenue)~log(sqrtCirculation))
summary(m8)
```

```
##
## Call:
## lm(formula = log(sqrtAdRevenue) ~ log(sqrtCirculation))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.235108 -0.055711 -0.002662  0.054173  0.213525
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.33737    0.01262  185.16   <2e-16 ***
## log(sqrtCirculation)  0.52876    0.02356   22.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0884 on 68 degrees of freedom
## Multiple R-squared:  0.881,  Adjusted R-squared:  0.8793
## F-statistic: 503.6 on 1 and 68 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m8)
```



For Prediction Interval at 0.5 million,

```
#my_data_path <- "F:/unr/4th sem/applied regression analysis/Assignments/Exam1"
#adData <- read.csv(file.path(my_data_path,"AdRevenue.csv"),header=TRUE)
#attach(adData)
#m5 <- lm(log(adData$AdRevenue)~log(adData$Circulation))
#summary(m5)
#newData1=data.frame(log(adData$Circulation)=-0.6931)
lRevenue<-log(adData$AdRevenue)
lCirculation<-log(adData$Circulation)
m5<-lm(lRevenue~lCirculation)
newData<-data.frame(lCirculation=-0.691)     #log(.5)=-0.691, natural log (ln) is considered
predict(m5,newData,interval="prediction",level=0.95)
```

```
##         fit      lwr      upr
## 1 4.309363 3.949007 4.669718
```

Similarly, for 20 million, the prediction interval is:

```
lRevenue<-log(adData$AdRevenue)
lCirculation<-log(adData$Circulation)
m5<-lm(lRevenue~lCirculation)
newData<-data.frame(lCirculation=2.99)
#log(20)=2.99, natural log (ln) is considered
predict(m5,newData,interval="prediction",level=0.95)
```

```
##         fit      lwr      upr
## 1 6.255721 5.882865 6.628576
```

```
detach(adData)
```

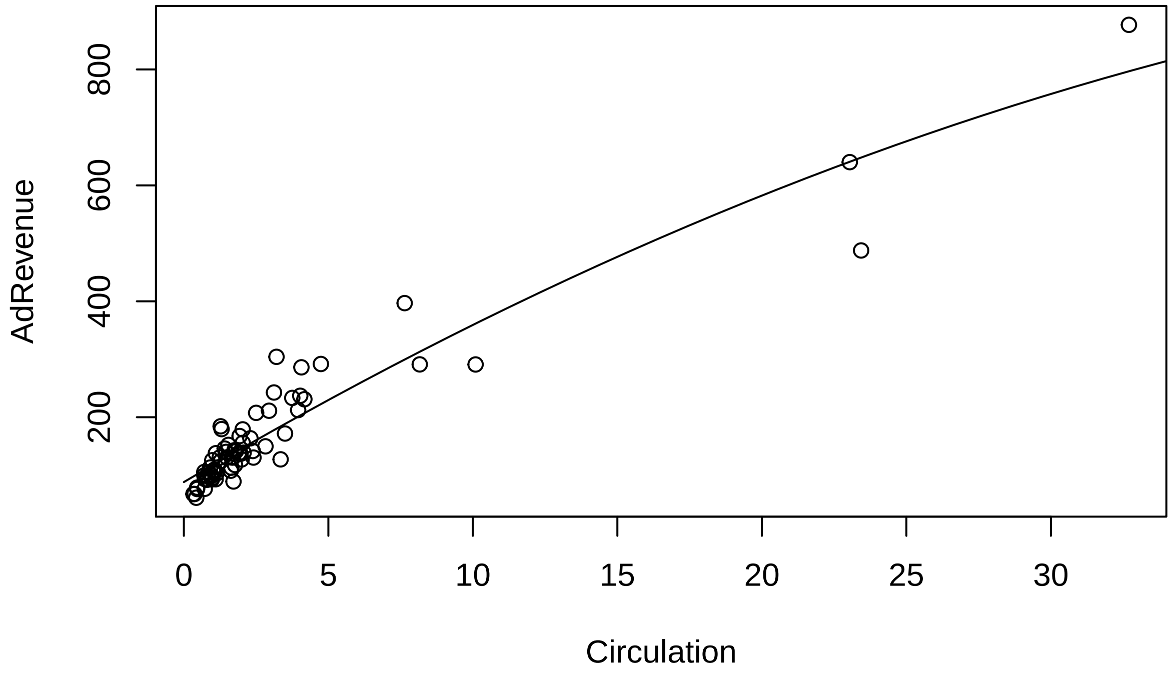
Appendix 2

1st Try: adRevenue = Circulation + Circulation ^2

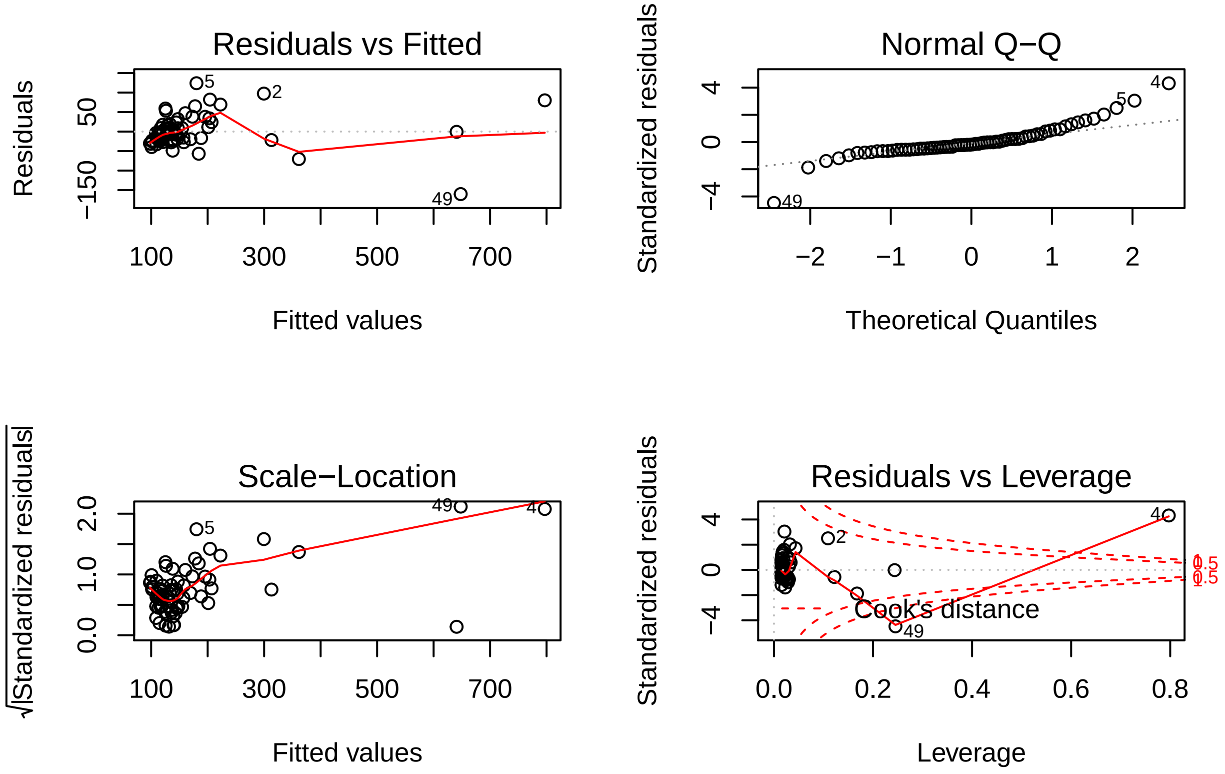
```
my_data_path <- "F:/unr/4th sem/applied regression analysis/Assignments/Exam1"
adData <- read.csv(file.path(my_data_path,"AdRevenue.csv"),header=TRUE)
attach(adData)
Circulation=adData$Circulation;
AdRevenue=adData$AdRevenue;
polym1 <- lm(AdRevenue~ Circulation + I(Circulation^2))
summary(polym1)
```

```
##
## Call:
## lm(formula = AdRevenue ~ Circulation + I(Circulation^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -160.32  -21.05   -7.65   15.30  123.96
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       88.1390     7.9708  11.058  < 2e-16 ***
## Circulation       29.5006     3.2992   8.942 4.87e-13 ***
## I(Circulation^2)  -0.2394     0.1140  -2.100   0.0395 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.2 on 67 degrees of freedom
## Multiple R-squared:  0.901,  Adjusted R-squared:  0.8981
## F-statistic: 304.9 on 2 and 67 DF,  p-value: < 2.2e-16
```

```
plot(Circulation,AdRevenue,xlab="Circulation")
CirculationNew <- seq(0,40,len=40)
lines(CirculationNew,predict(polym1,newdata=data.frame(Circulation=CirculationNew)))
```



```
par(mfrow=c(2,2))
plot(polym1)
```



We can see that the variance is highly varying which indicates that this model is a bad model for our data.

2nd Try: adRevenue = Circulation + Circulation ^2 + Circulatio^3

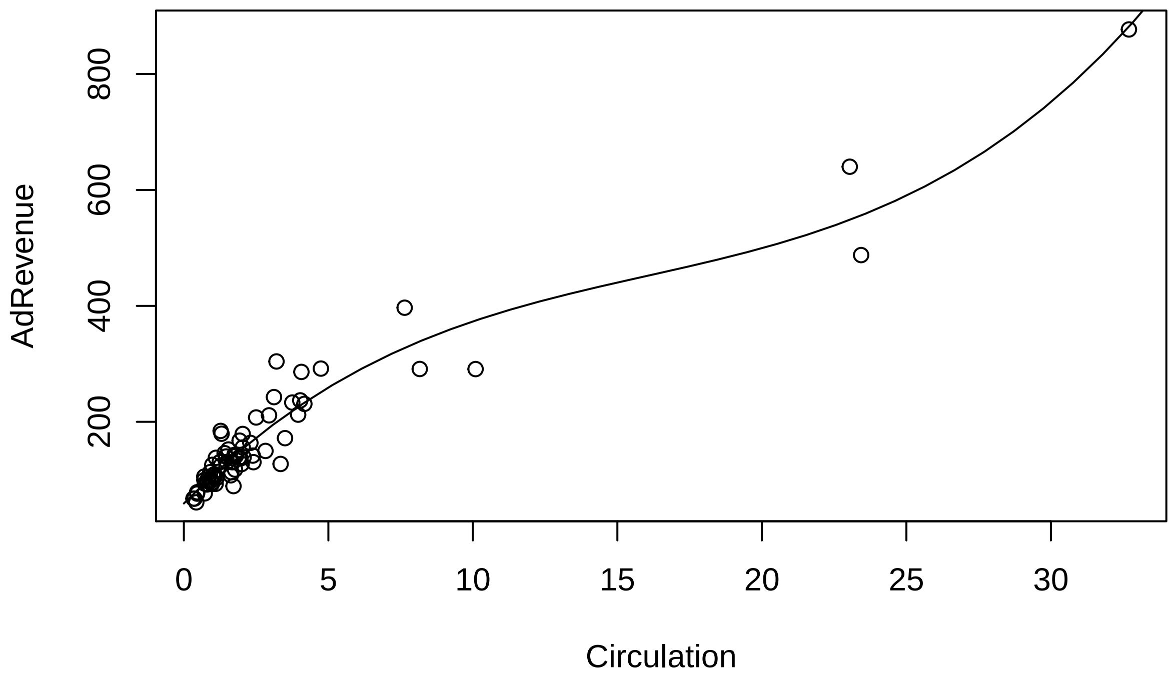
```
my_data_path <- "F:/unr/4th sem/applied regression analysis/Assignments/Exam1"
adData <- read.csv(file.path(my_data_path,"AdRevenue.csv"),header=TRUE)
attach(adData)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     AdRevenue, Circulation

## The following objects are masked from adData (pos = 3):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY
```

```
Circulation=adData$Circulation;
AdRevenue=adData$AdRevenue;
polym1 <- lm(AdRevenue~ Circulation + I(Circulation^2)+ I(Circulation^3))
summary(polym1)
```

```
##
## Call:
## lm(formula = AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3))
##
## 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 ***
## Circulation      51.23582    4.71123  10.875 2.33e-16 ***
## I(Circulation^2) -2.50538    0.41141  -6.090 6.48e-08 ***
## I(Circulation^3)  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
```

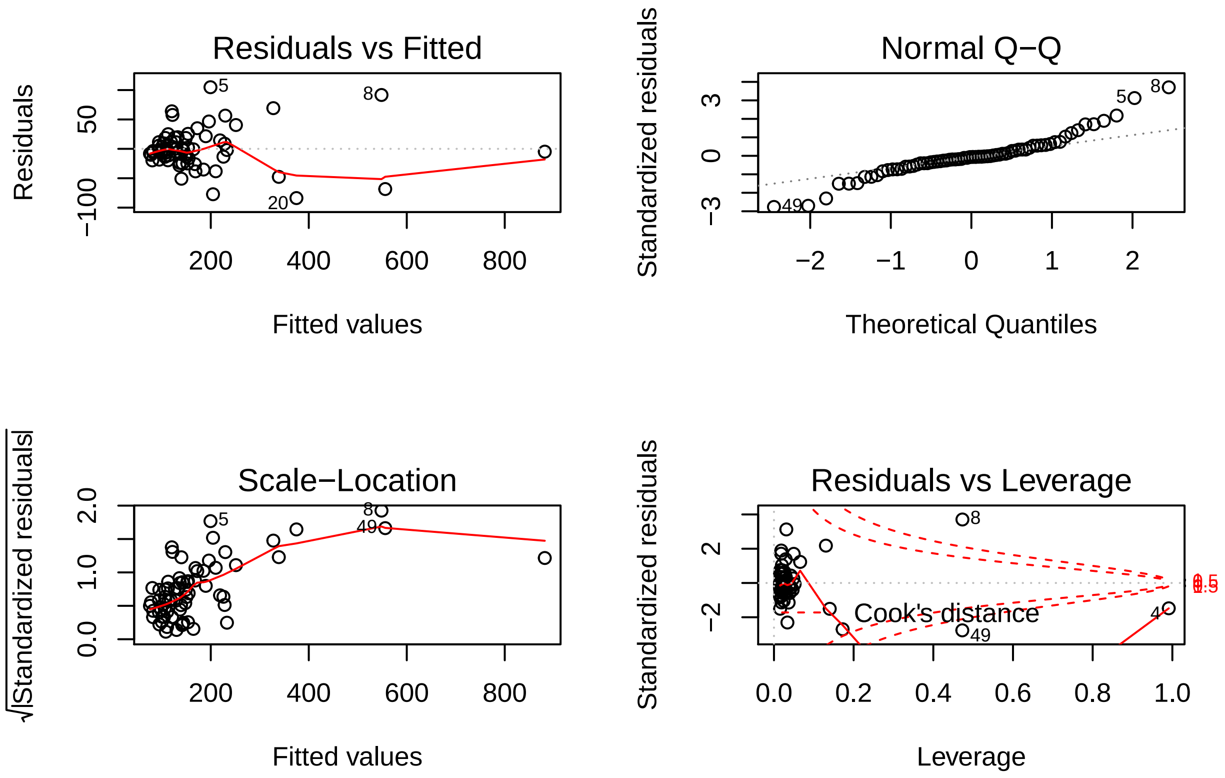
```
plot(Circulation,AdRevenue,xlab="Circulation")
CirculationNew <- seq(0,40,len=40)
lines(CirculationNew,predict(polym1,newdata=data.frame(Circulation=CirculationNew)))
```



```
par(mfrow=c(2,2))
plot(polym1)
```

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

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



3rd Try: adRevenue = Circulation + Circulation^3

```
my_data_path <- "F:/unr/4th sem/applied regression analysis/Assignments/Exam1"
adData <- read.csv(file.path(my_data_path,"AdRevenue.csv"),header=TRUE)
attach(adData)
```

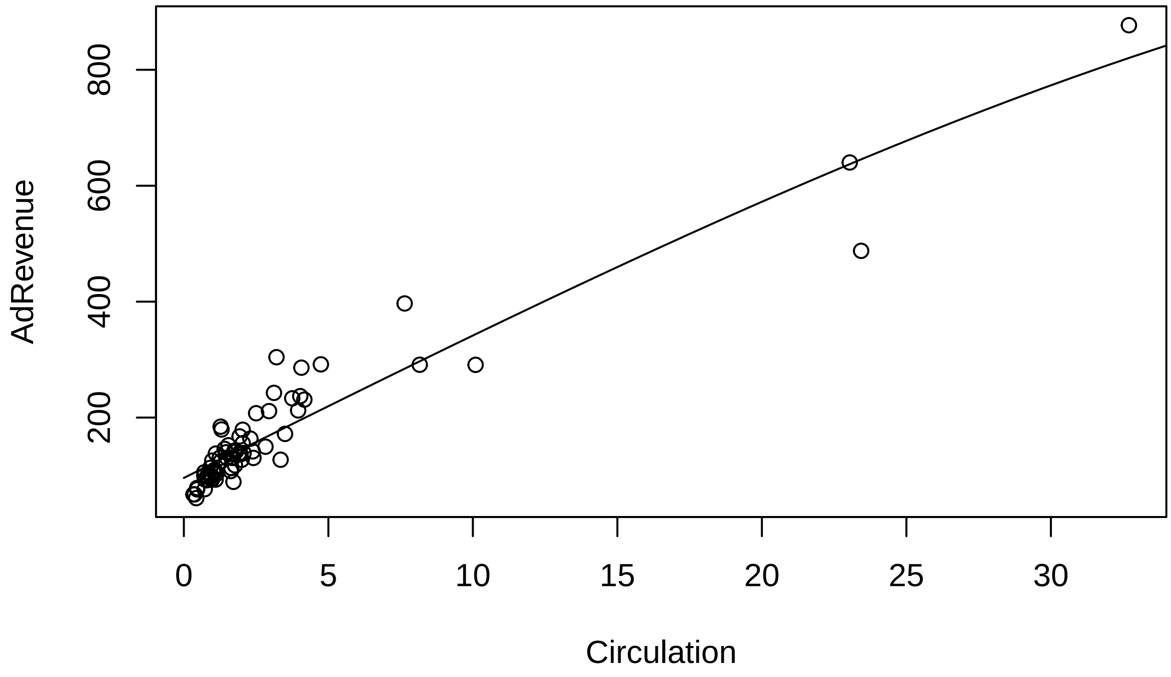
```
## The following objects are masked _by_ .GlobalEnv:
##
##     AdRevenue, Circulation

## The following objects are masked from adData (pos = 3):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY

## The following objects are masked from adData (pos = 4):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY
```

```
Circulation=adData$Circulation;
AdRevenue=adData$AdRevenue;
polym1 <- lm(AdRevenue~ Circulation +  I(Circulation^3))
summary(polym1)
```

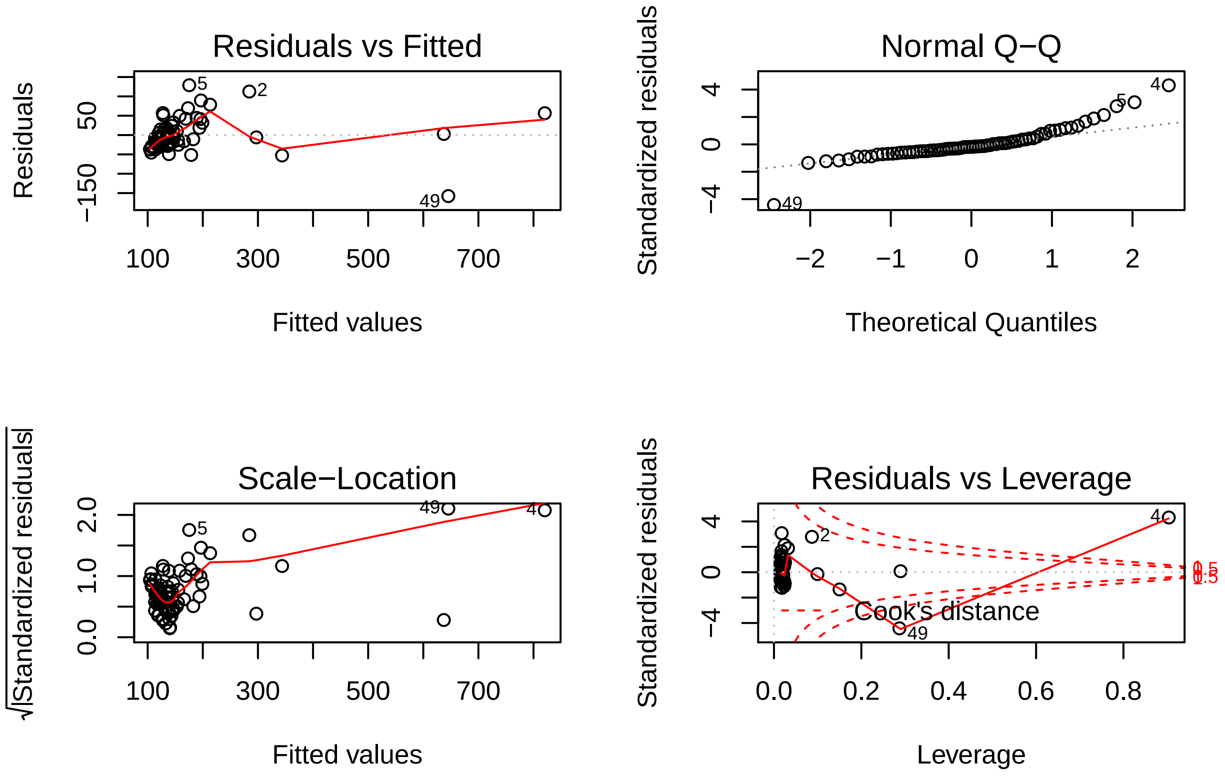
```
##
## Call:
## lm(formula = AdRevenue ~ Circulation + I(Circulation^3))
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -157.759  -21.924   -7.655   14.852  128.809
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      95.930241   7.147283  13.422   <2e-16 ***
## Circulation      24.813688   2.277217  10.896   <2e-16 ***
## I(Circulation^3) -0.002486   0.002623  -0.948    0.347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.25 on 67 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8928
## F-statistic: 288.3 on 2 and 67 DF,  p-value: < 2.2e-16
```

```
plot(Circulation,AdRevenue,xlab="Circulation")
CirculationNew <- seq(0,40,len=40)
lines(CirculationNew,predict(polym1,newdata=data.frame(Circulation=CirculationNew)))
```



```
par(mfrow=c(2,2))
plot(polym1)
```



4th Try: adRevenue = Circulation ^2 + Circulation^3

```
my_data_path <- "F:/unr/4th sem/applied regression analysis/Assignments/Exam1"
adData <- read.csv(file.path(my_data_path,"AdRevenue.csv"),header=TRUE)
attach(adData)
```

```
## The following objects are masked _by_ .GlobalEnv:
##
##     AdRevenue, Circulation
```

```
## The following objects are masked from adData (pos = 3):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY
```

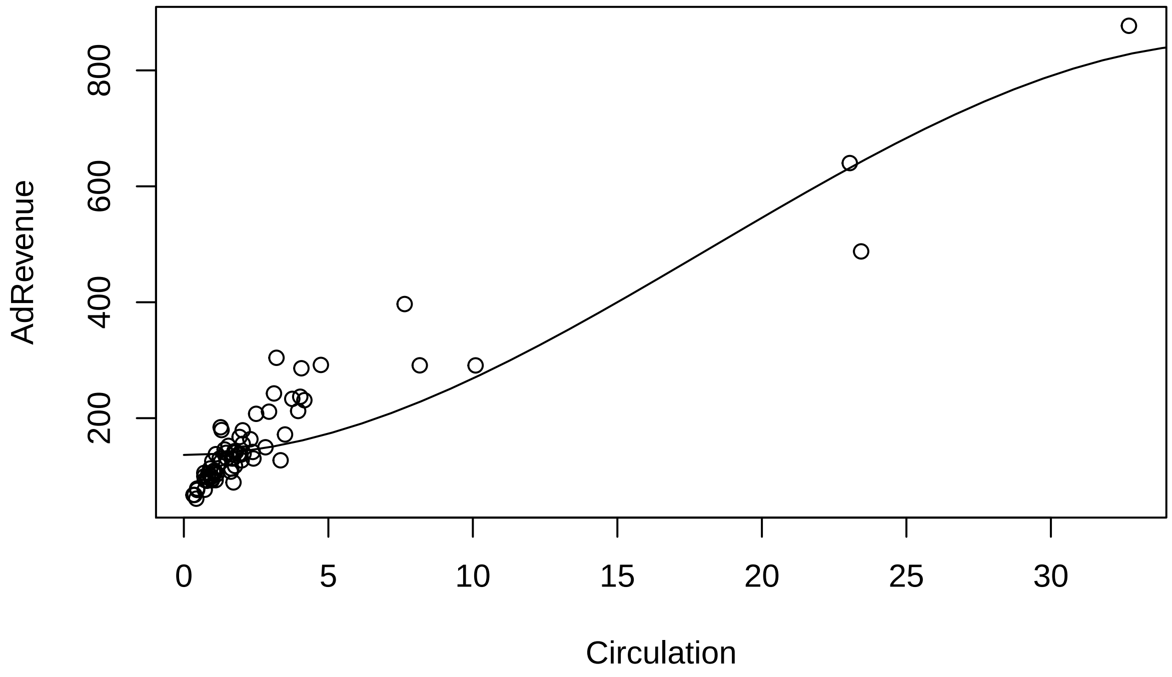
```
## The following objects are masked from adData (pos = 4):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY
```

```
## The following objects are masked from adData (pos = 5):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY
```

```
Circulation=adData$Circulation;
AdRevenue=adData$AdRevenue;
polym1 <- lm(AdRevenue~ I(Circulation^2)+ I(Circulation^3))
plot(Circulation,AdRevenue,xlab="Circulation")
summary(polym1)
```

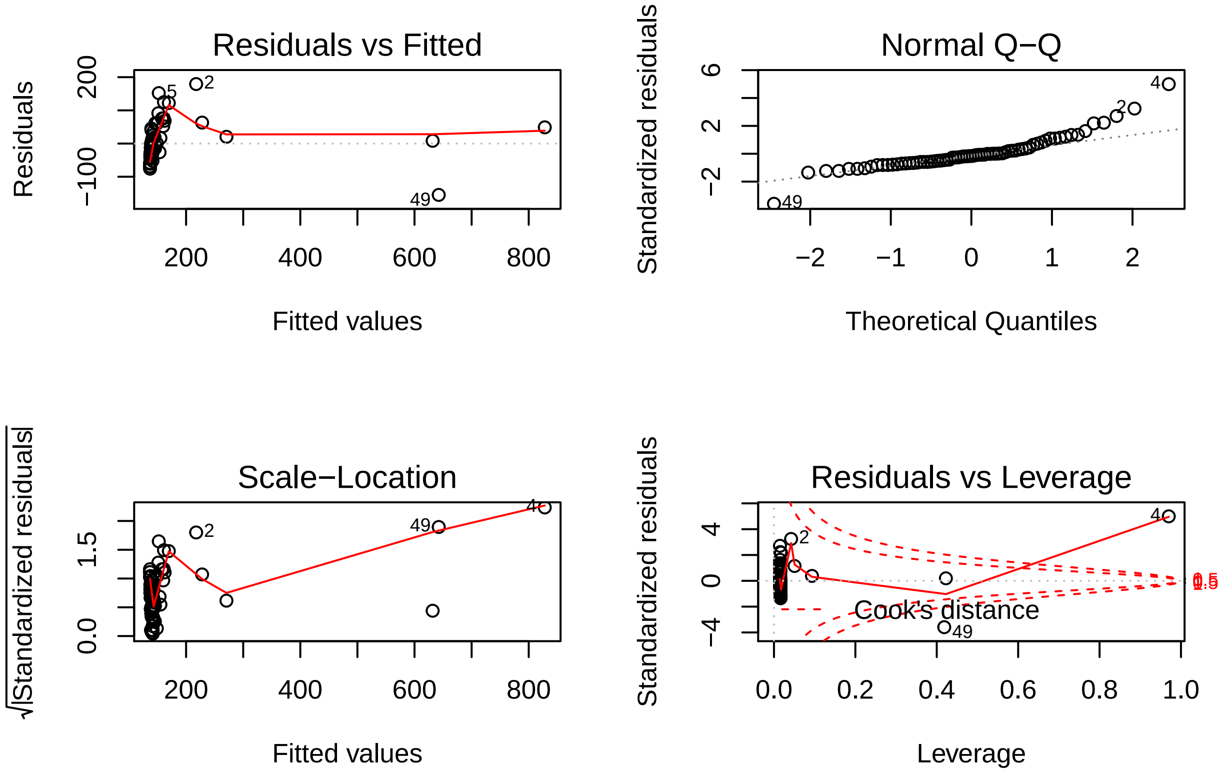
```
##
## Call:
## lm(formula = AdRevenue ~ I(Circulation^2) + I(Circulation^3))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -154.750  -35.027   -8.955   19.989  179.241
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)      136.575853   7.224973  18.903  < 2e-16 ***
## I(Circulation^2)   1.615131   0.265875   6.075 6.57e-08 ***
## I(Circulation^3)  -0.029615   0.008862  -3.342  0.00136 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.49 on 67 degrees of freedom
## Multiple R-squared:  0.8139, Adjusted R-squared:  0.8083
## F-statistic: 146.5 on 2 and 67 DF,  p-value: < 2.2e-16
```

```
CirculationNew <- seq(0,40,len=40)
lines(CirculationNew,predict(polym1,newdata=data.frame(Circulation=CirculationNew)))
```



```
par(mfrow=c(2,2))
plot(polym1)
```

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

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



5th Try: adRevenue = Circulationˆ3

```r
my_data_path <- "F:/unr/4th sem/applied regression analysis/Assignments/Exam1"
adData <- read.csv(file.path(my_data_path,"AdRevenue.csv"),header=TRUE)
attach(adData)
```

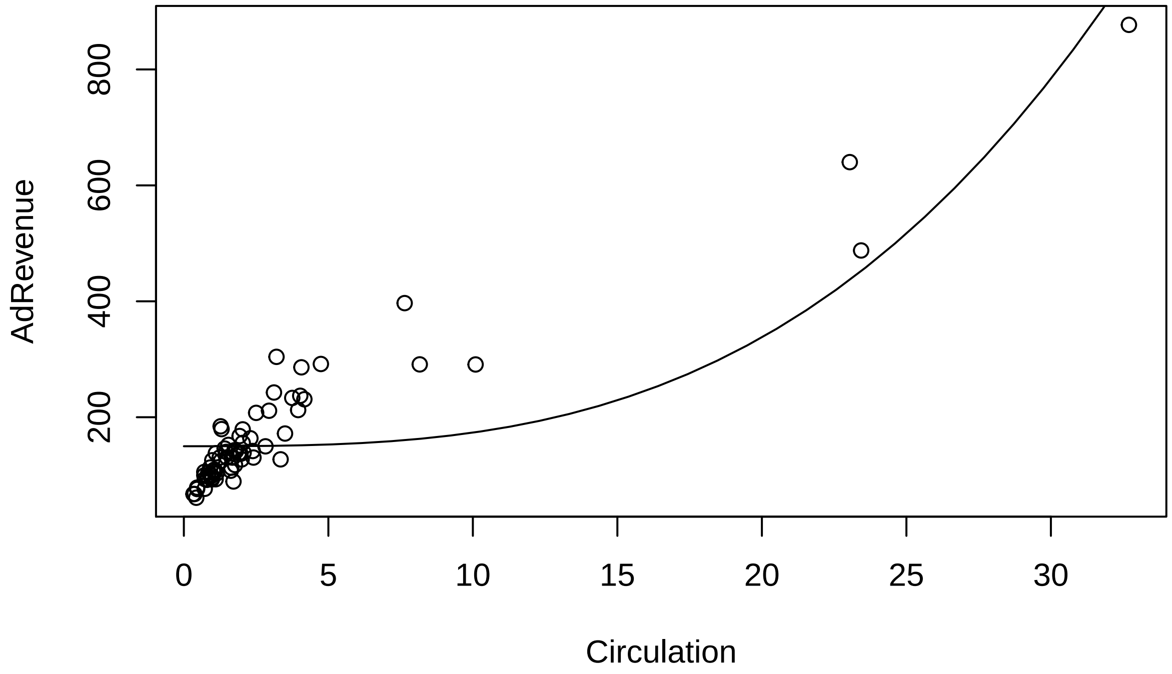
```
## The following objects are masked _by_ .GlobalEnv:
##
##     AdRevenue, Circulation

## The following objects are masked from adData (pos = 3):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY

## The following objects are masked from adData (pos = 4):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY

## The following objects are masked from adData (pos = 5):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY

## The following objects are masked from adData (pos = 6):
```

```
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY
```

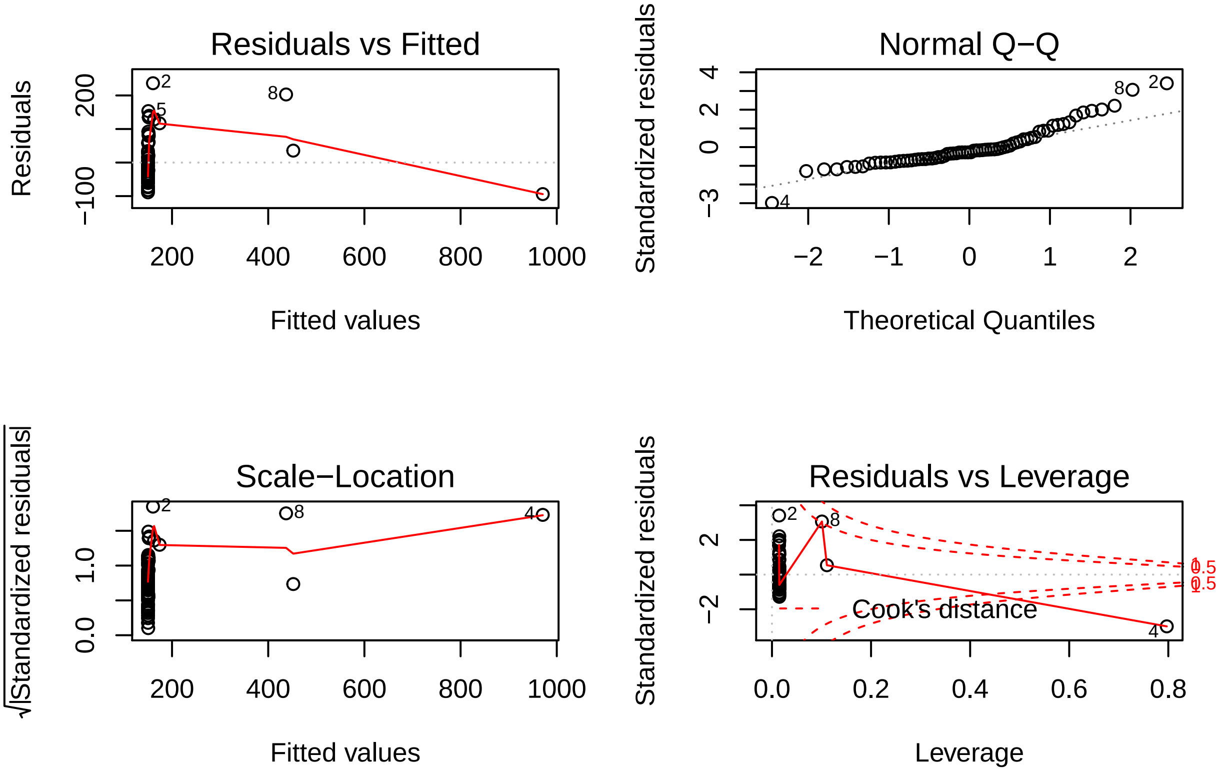
```
Circulation=adData$Circulation;
AdRevenue=adData$AdRevenue;
polym1 <- lm(AdRevenue~  I(Circulation^3))
summary(polym1)
```

```
##
## Call:
## lm(formula = AdRevenue ~ I(Circulation^3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -93.99 -46.66 -19.45  26.85 236.42
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.500e+02  8.504e+00   17.64   <2e-16 ***
## I(Circulation^3) 2.348e-02  1.813e-03   12.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.83 on 68 degrees of freedom
## Multiple R-squared:  0.7114, Adjusted R-squared:  0.7071
## F-statistic: 167.6 on 1 and 68 DF,  p-value: < 2.2e-16
```

```
plot(Circulation,AdRevenue,xlab="Circulation")
CirculationNew <- seq(0,40,len=40)
lines(CirculationNew,predict(polym1,newdata=data.frame(Circulation=CirculationNew)))
```



```
par(mfrow=c(2,2))
plot(polym1)
```



6th Try: adRevenue = Circulation ^2

```
my_data_path <- "F:/unr/4th sem/applied regression analysis/Assignments/Exam1"
adData <- read.csv(file.path(my_data_path,"AdRevenue.csv"),header=TRUE)
attach(adData)
```

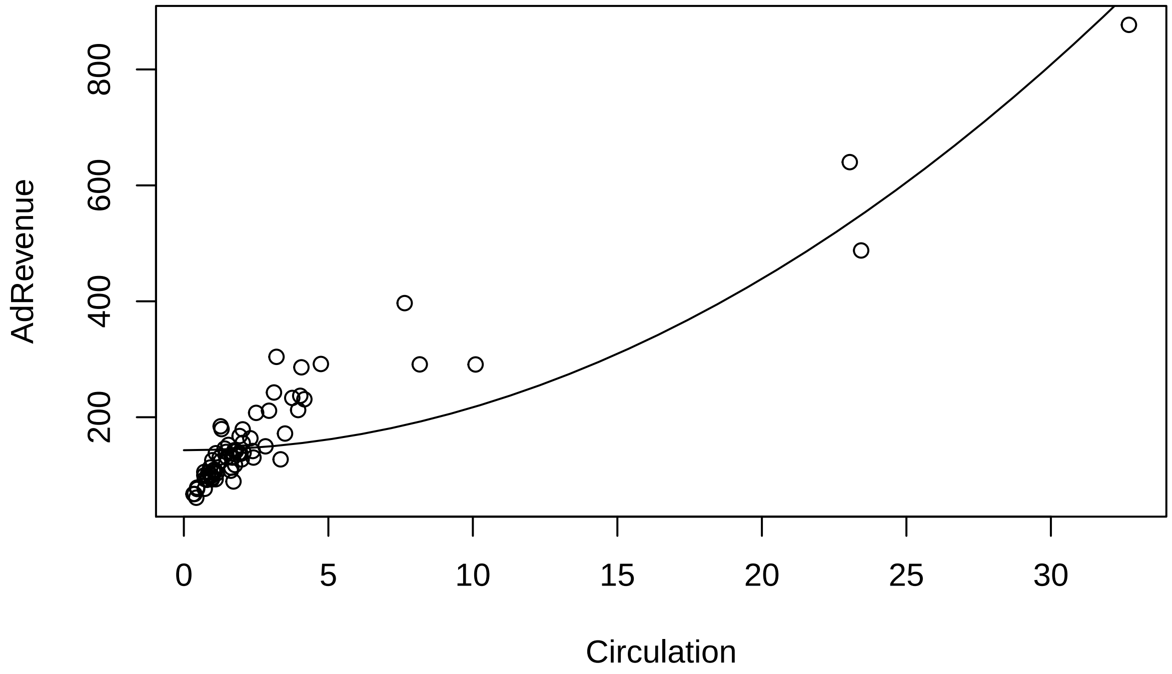
```
## The following objects are masked _by_ .GlobalEnv:
##
##     AdRevenue, Circulation

## The following objects are masked from adData (pos = 3):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY

## The following objects are masked from adData (pos = 4):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY

## The following objects are masked from adData (pos = 5):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY

## The following objects are masked from adData (pos = 6):
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY

## The following objects are masked from adData (pos = 7):
```

```
##
##     AdRevenue, Circulation, Magazine, PARENT.COMPANY..SUBSIDIARY
```

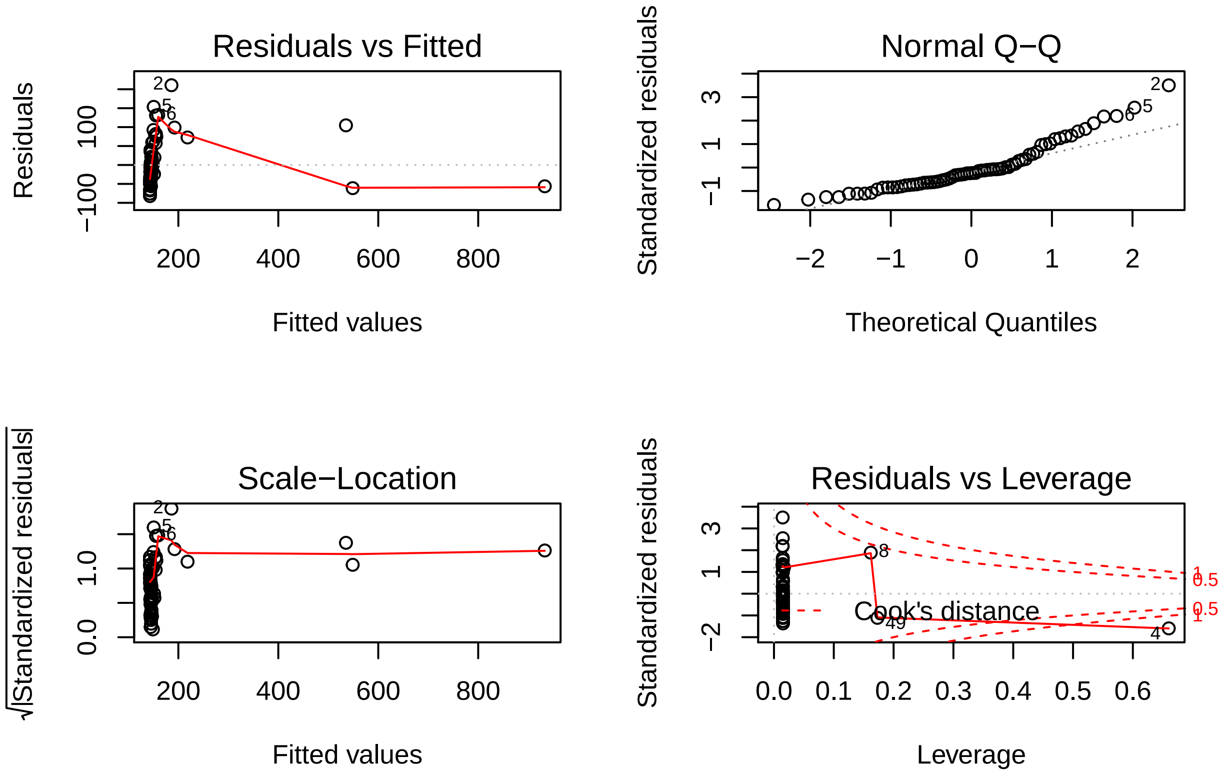
```
Circulation=adData$Circulation;
AdRevenue=adData$AdRevenue;
polym1 <- lm(AdRevenue~ I(Circulation^2))
summary(polym1)
```

```
##
## Call:
## lm(formula = AdRevenue ~ I(Circulation^2))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -82.16 -42.52 -14.32  21.17 210.63
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      143.11980    7.45636   19.19   <2e-16 ***
## I(Circulation^2)   0.73889    0.04719   15.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.57 on 68 degrees of freedom
## Multiple R-squared:  0.7829, Adjusted R-squared:  0.7797
## F-statistic: 245.2 on 1 and 68 DF,  p-value: < 2.2e-16
```

```
plot(Circulation,AdRevenue,xlab="Circulation")
CirculationNew <- seq(0,40,len=40)
lines(CirculationNew,predict(polym1,newdata=data.frame(Circulation=CirculationNew)))
```



```
par(mfrow=c(2,2))
plot(polym1)
```



For prediction interval:

```
my_data_path <- "F:/unr/4th sem/applied regression analysis/Assignments/Exam1"
adData <- read.csv(file.path(my_data_path,"AdRevenue.csv"),header=TRUE)
#attach(adData)
Circulation=adData$Circulation;
AdRevenue=adData$AdRevenue;
polym1 <- lm(AdRevenue~ Circulation + I(Circulation^2)+ I(Circulation^3))

newData<-data.frame(Circulation=0.5)
predict(polym1,newData,interval="prediction",level=0.95)
```

```
##        fit      lwr      upr
## 1 84.16846 14.92314 153.4138
```

```
#detach(adData)
```

For prediction interval:

```
my_data_path <- "F:/unr/4th sem/applied regression analysis/Assignments/Exam1"
adData <- read.csv(file.path(my_data_path,"AdRevenue.csv"),header=TRUE)
#attach(adData)
Circulation=adData$Circulation;
AdRevenue=adData$AdRevenue;
polym1 <- lm(AdRevenue~ Circulation + I(Circulation^2)+ I(Circulation^3))

newData<-data.frame(Circulation=20)
predict(polym1,newData,interval="prediction",level=0.95)
```

```
##        fit     lwr      upr
## 1 499.5334 418.179 580.8878
```

```
#detach(adData)
```