Fall 2022 STAT 706 HW 2

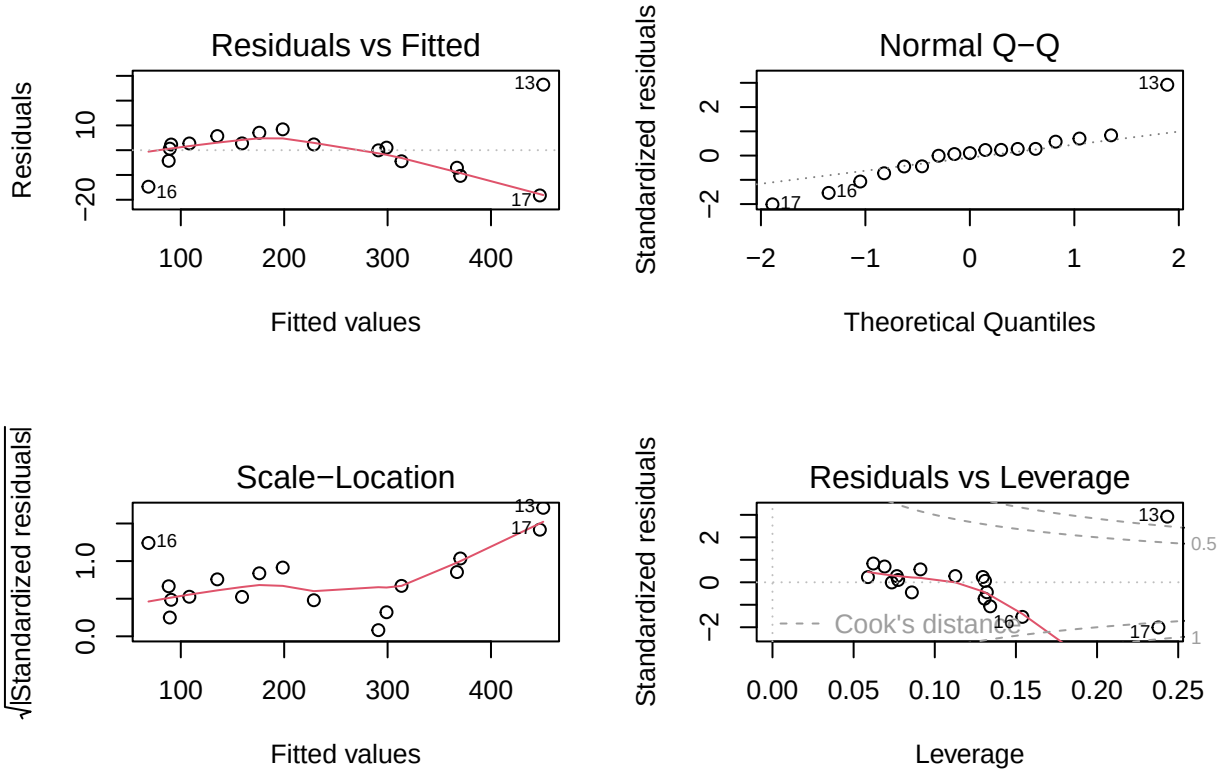
Reina Li

Chapter 3.4 : 1a-b, 2, 3 PtA a-c, PtB a-c, PtC a-b, 4a-b

1.

```
# Import data set-----
df_1 <- read.table("airfares.txt", header = TRUE, sep = "")

# Fit the model to the data-----
model_q1 <- lm(Fare ~ Distance,
               data = df_1)
par(mfrow = c(2,2))
plot(model_q1)
```



(a)

Based on the output for model $Fare = \beta_0 + \beta_1 Distance + e$, when looking at the Residual vs Fitted and Standardized Residuals vs Distance plots, there seems to be a non-linear trend in the residuals. Looking at the Residuals vs Leverage plot, points 13 and 17 are leverage points. Leverage points reduce the confidence in OLS assumptions. I agree that Distance explains most of the variability in the Y-variable Fare, but I do not think the model $Fare = \beta_0 + \beta_1 Distance + e$ is a highly effective model to (1) understand the effects of Distance on Fare and (2) predict future values of Fare given Distance.

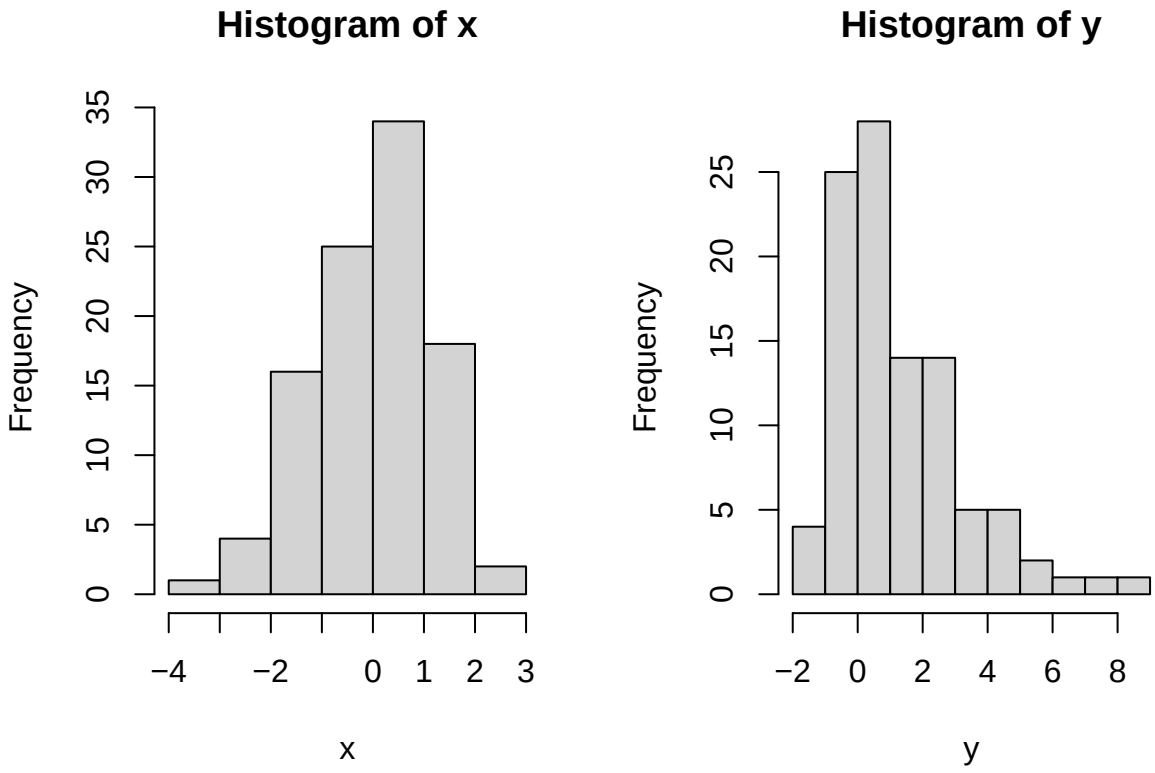
(b)

No, the ordinary straight line regression model does not seem to fit the data well. Looking at the Residuals vs Leverage plot, there is an outlier/leverage point that is in Cook's distance greater than 1, which is point 13. Looking at the Residual vs Fitted and Standardized Residuals vs Distance plots, there seems to be a non-linear trend in the residuals. Therefore, the model can be improved by applying a transformation or removing the leverage points.

2.

Suppose that a straight line regression model has been fit to bivariate data set of the form $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Furthermore, suppose that the distribution of X appears to be normal while the Y variable is highly skewed. A plot of standardized residuals from the least squares regression line produce a quadratic pattern with increasing variance when plotted against $(x_1, x_2, \ldots, x_n)$. In this case, one should consider adding a quadratic term in X to the regression model and thus consider a model of the form $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$.

```
#test case
x <- rnorm(n = 100, mean = 0, sd = 1)
err <- rnorm(n = 100, mean = 0, sd = 1)
y <- x + x^2 + err
par(mfrow = c(1,2))
hist(x)
hist(y)
```



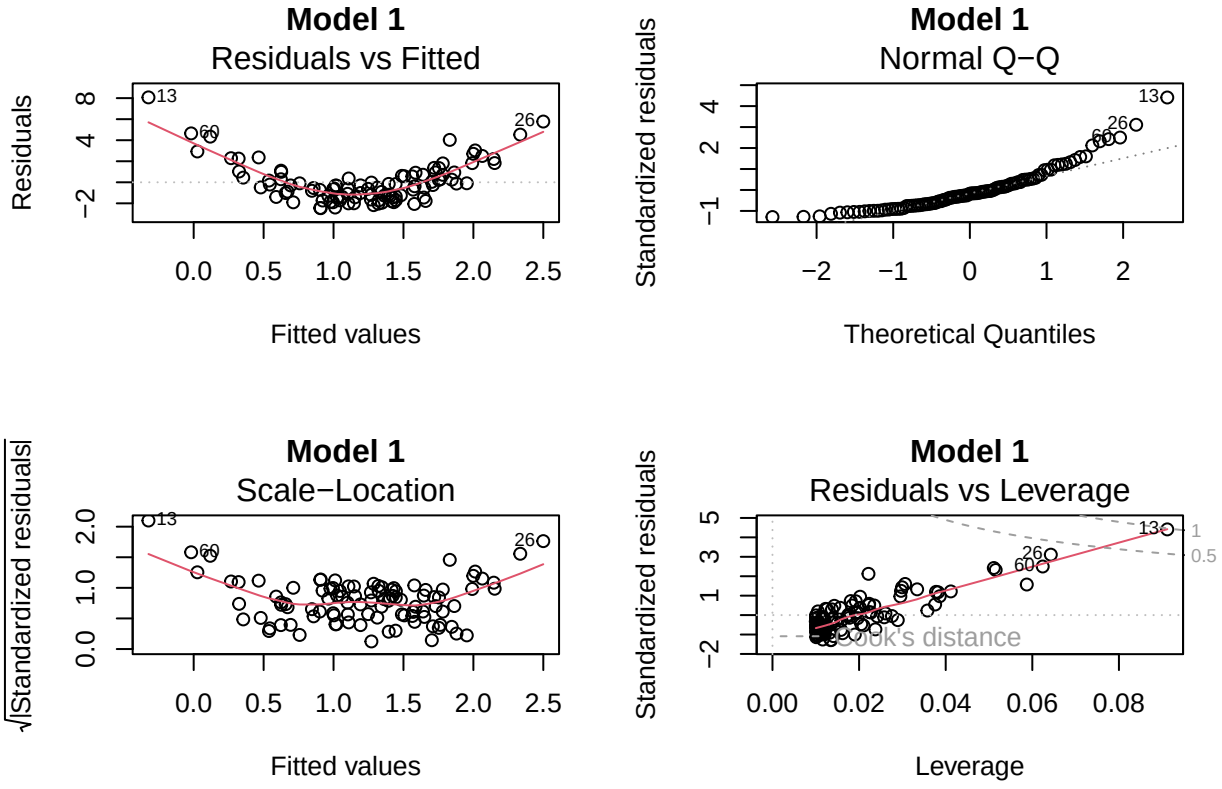

```
model1_q2 <- lm(y ~ x)
summary(model1_q2)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4608 -1.3907 -0.3449  0.7644  8.0707
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2458     0.1920   6.487 3.58e-09 ***
## x             0.4709     0.1658   2.840  0.00548 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.919 on 98 degrees of freedom
## Multiple R-squared:  0.07606,    Adjusted R-squared:  0.06663
## F-statistic: 8.067 on 1 and 98 DF,  p-value: 0.005483
```

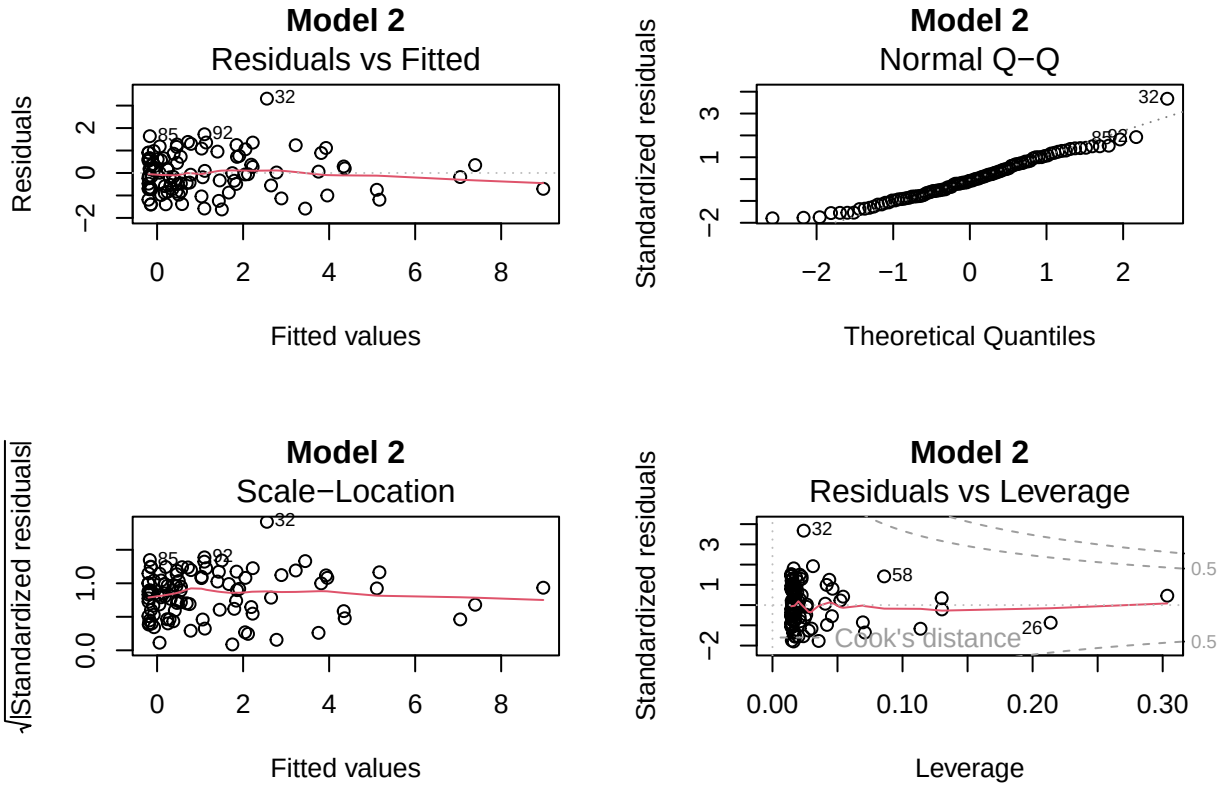
```
model2_q2 <- lm(y ~ x + I(x^2))
summary(model2_q2)
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6228 -0.7044 -0.0710  0.6562  3.3088
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.01050    0.11322   0.093    0.926
## x            0.88732    0.08188  10.837   <2e-16 ***
## I(x^2)       0.93137    0.05065  18.387   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.911 on 97 degrees of freedom
## Multiple R-squared:  0.794,  Adjusted R-squared:  0.7898
## F-statistic:   187 on 2 and 97 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(model1_q2, main = "Model 1") # model 1: Y = beta0 + beta1*x + e
```




```
par(mfrow = c(2,2))
plot(model2_q2, main = "Model 2") # model 2: Y = beta0 + beta1*x + beta2*x^2 + e
```




```
anova(model1_q2, model2_q2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ x + I(x^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     98 361.06
## 2     97  80.50  1    280.57 338.09 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

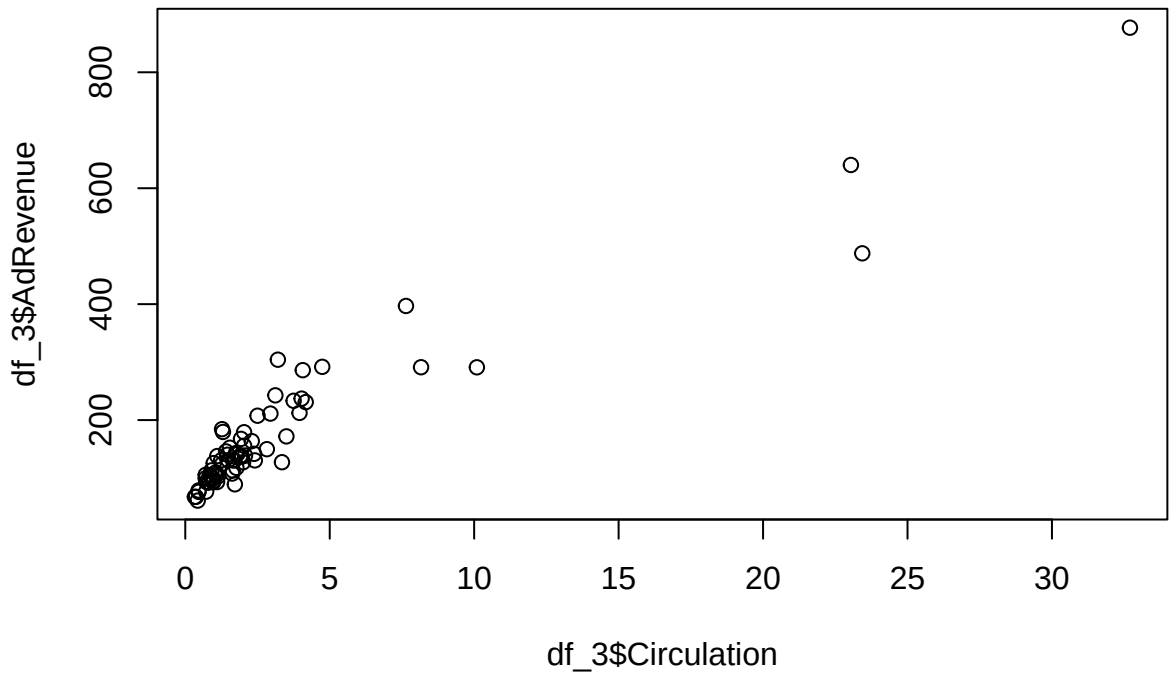
I believe the statement above is true. Here in this test case, the distribution of X appears to be normal while the distribution of Y is skewed, as shown in the histograms. When we look at Model 1's Residual vs Fitted plot, we see a quadratic line. When we look at Model 2's Residual vs Fitted plot, we see a linear line. Also, when we look at the ANOVA table, we can conclude that model 2 is statistically significant. That is why I think the statement is true.

3. Part A

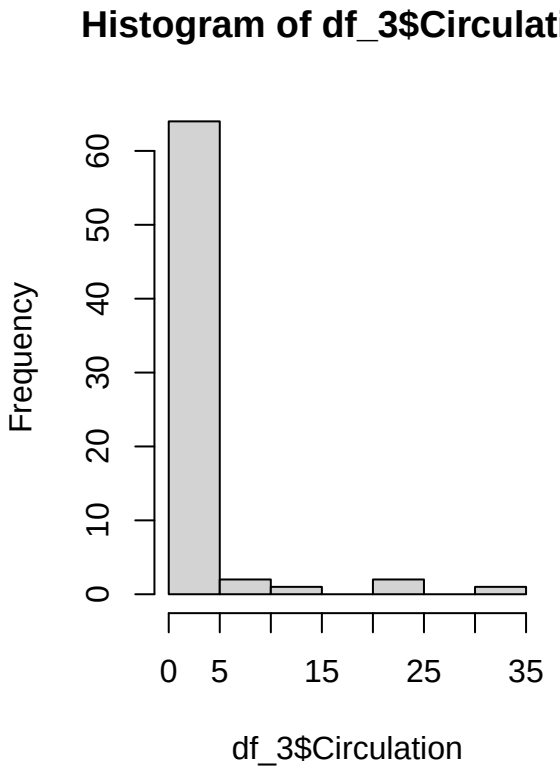
```
# Import data set-----
df_3 <- read.csv("AdRevenue.csv")
```

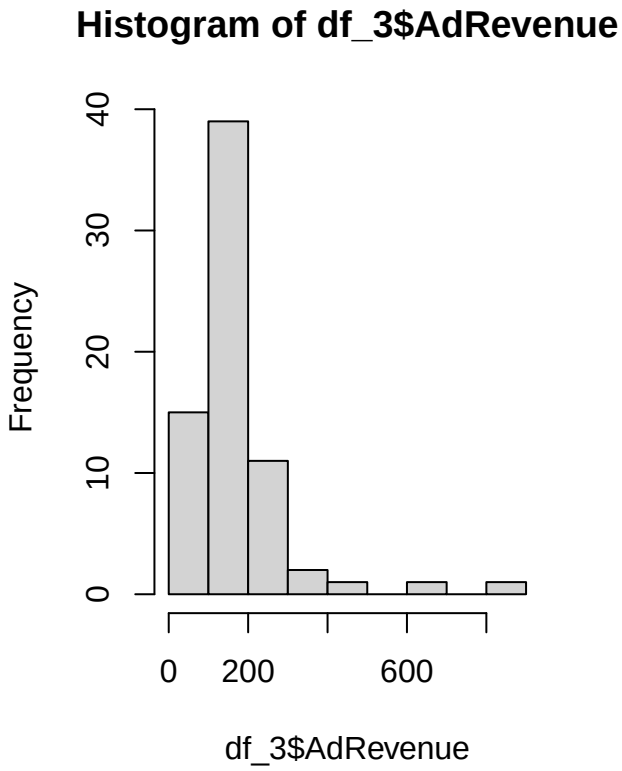
(a)

```
# Fit the model to the data-----
plot(df_3$Circulation,df_3$AdRevenue)
```




```
par(mfrow = c(1,2))
hist(df_3$Circulation)
hist(df_3$AdRevenue)
```






```
# Load library
library(MASS)

# Find optima lambda for box cox transformation
bc <- boxcox(AdRevenue ~ Circulation,
             data = df_3)
(lambda <- bc$x[which.max(bc$y)])
```

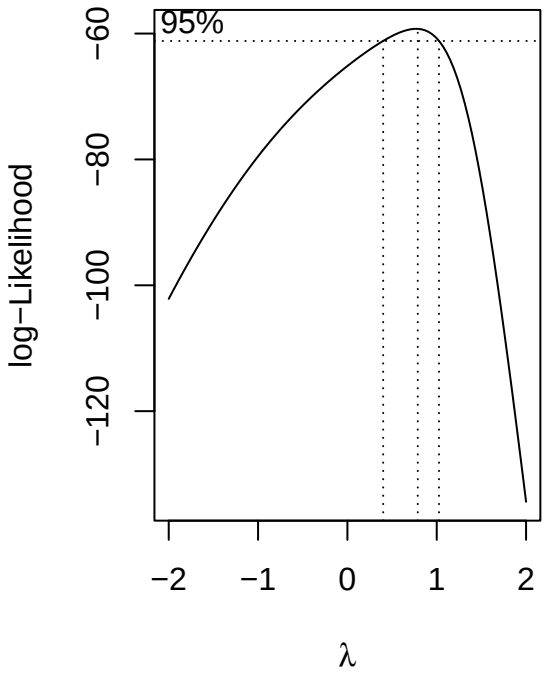
```
## [1] 0.7878788
```

```
model_3a <- lm(((AdRevenue^lambda-1)/lambda) ~ Circulation,
               data = df_3)
summary(model_3a)
```

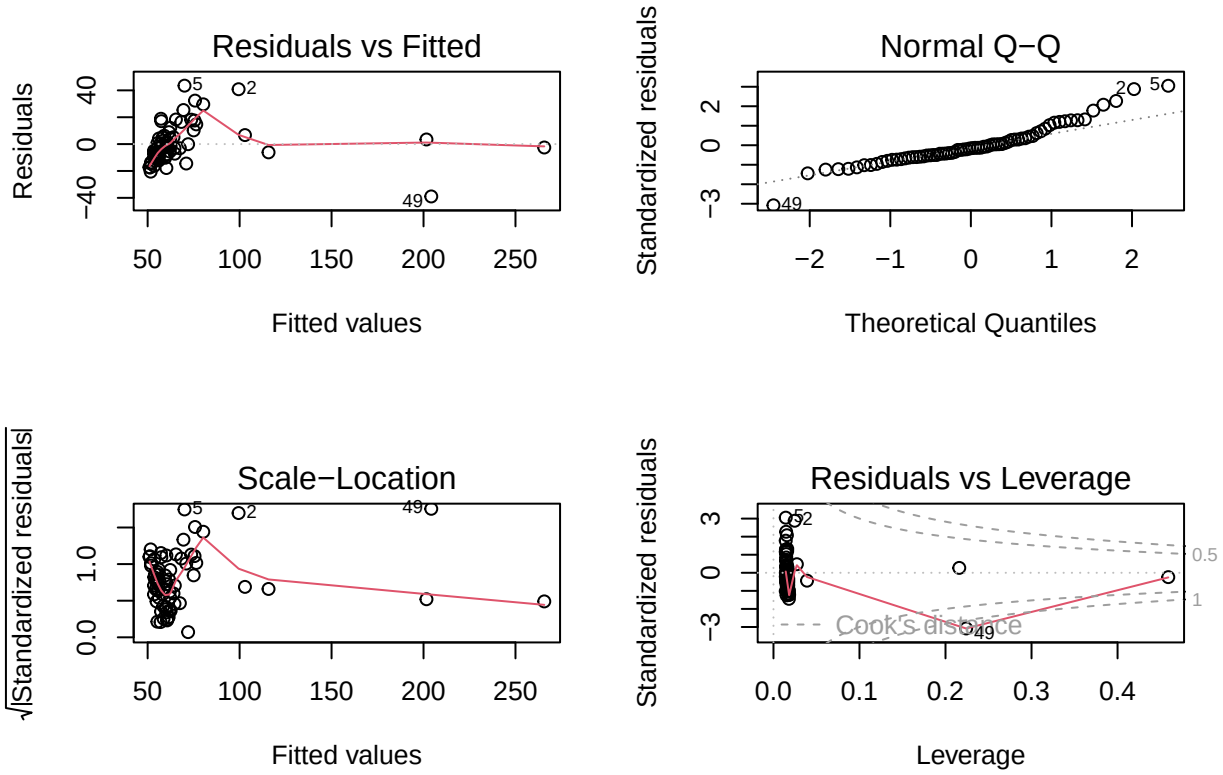
```
##
## Call:
## lm(formula = ((AdRevenue^lambda - 1)/lambda) ~ Circulation, data = df_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.958  -8.563  -2.243   5.068  43.479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.8002     1.9905   24.52   <2e-16 ***
```

```
## Circulation   6.6315     0.3236   20.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.35 on 68 degrees of freedom
## Multiple R-squared:  0.8607, Adjusted R-squared:  0.8586
## F-statistic:   420 on 1 and 68 DF,  p-value: < 2.2e-16
```

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




```
plot(model_3a)
```



Looking at the scatter plot of the data, it seems like there is a non-linear trend. Also, looking at the histograms, the distributions of AdRevenue and Circulation are highly skewed. So, both of the variables will be transformed. I performed a box-cox transformation in R using the `boxcox()` function from the `MASS()` library. The optimal lambda for box-cox transformation was found to be 0.7878788. So, the new regression model replaced the original response variable with the variable $AdRevenue = (AdRevenue^{0.788} - 1)/0.788$

(b)

```
predict(model_3a, data.frame(Circulation = 0.5), interval = "prediction", level = 0.95)
```

(i)

```
##        fit      lwr      upr
## 1 52.11597 23.21844 81.01349
```

A 95% prediction interval for the advertising revenue (in thousands of dollars) per page for magazines with 0.5 million circulations is (23.2184412, 81.0134943).

```
predict(model_3a, data.frame(Circulation = 20), interval = "prediction", level = 0.95)
```

(ii)

```
##       fit      lwr      upr
## 1 181.431 150.5922 212.2698
```

A 95% prediction interval for the advertising revenue (in thousands of dollars) per page for magazines with 20 million circulations is (150.5921506, 212.2697851).

(c)

The weaknesses in my model are that there are leverage points.

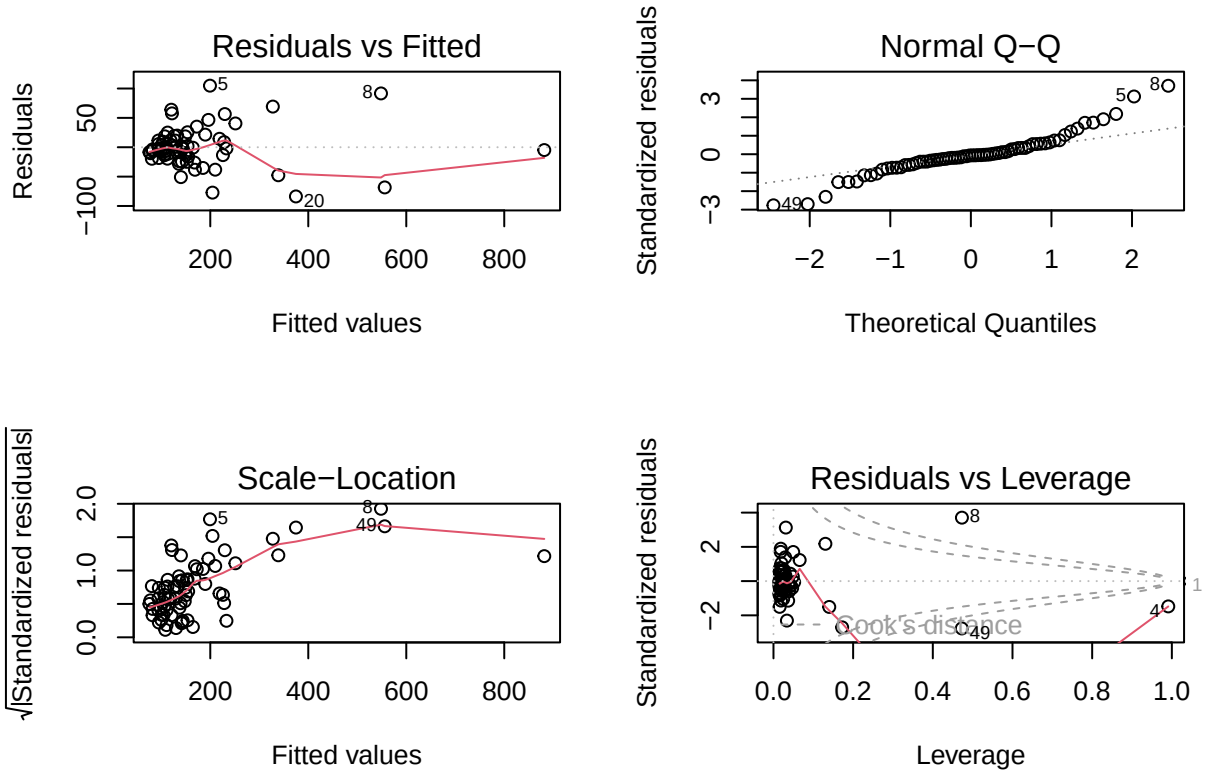
3. Part B

(a)

```
model_3b <- lm(AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3),
               data = df_3)
summary(model_3b)
```

```
##
## Call:
## lm(formula = AdRevenue ~ Circulation + I(Circulation^2) + I(Circulation^3),
##     data = df_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
```

```
par(mfrow = c(2,2))
plot(model_3b)
```



(b)

```
predict(model_3b, data.frame(Circulation = 0.5), interval = "prediction", level = 0.95)
```

(i)

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

A 95% prediction interval for the advertising revenue (in thousands of dollars) per page for magazines with 0.5 million circulations is (14.9231416, 153.4137781).

```
predict(model_3b, data.frame(Circulation = 20), interval = "prediction", level = 0.95)
```

(ii)

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

A 95% prediction interval for the advertising revenue (in thousands of dollars) per page for magazines with 20 million circulations is (418.1790316, 580.8878085).

(c)

In contrast to Part A, for this Part B, I transformed neither the predictor nor the response variable. Instead, I considered a polynomial model of order up to 3. The weakness I see in my model is that there are leverage points.

3. Part C

(a)

In Part A, I performed a box-cox transformation, while in Part B I considered a polynomial model of order 3. Looking at the plots created from both models, I decide that the model from Part XYZ provides a better model because when we look at the diagnostic plots from model A, we can see that model A fits better than model B. However, I see weaknesses in both models because there are still leverage points. Additionally, for both models, the Normal Q-Q plot points fall along a line in the middle of the graph, but curve off in the extremities, but more so for model B, which may indicate that the dataset likely does not follow a normal distribution.

(b)

For 0.5 million circulations, the prediction interval using model A was (23.2184412, 81.0134943). The prediction interval using model B was (14.9231416, 153.4137781). The prediction interval for model A is narrower than that of model B. In this case, I recommend model A's prediction interval because on the original scale, the data has variance which increases as the x-variable increases.

For 20 million circulations, the prediction interval using model A was (150.5921506, 212.2697851). The prediction interval using model B was (418.1790316, 580.8878085). The prediction interval for model A is narrower than that of model B. In this case, I recommend model A's prediction interval because on the original scale, the data has variance which increases as the x-variable increases.

4.

(a)

The textbook page 106 shows some output from fitting model $Time = \beta_0 + \beta_1 Tonnage + e$ as well as some plots of Tonnage and Time. Looking at the outputted plots, the straight line regression model does not seem to fit the data well. When tonnage is lower, the model seems to fit the data, but as tonnage increases, the model does not fit the data. The residual also increases as tonnage increases. When we look at the Residuals vs Fitted plot, we can see that as tonnage increases, more values are not predicted well, because they are far away from the line at $x = 0$. Looking at the box and whisker plots, we can see that there are outliers.

(b)

If the model $Time = \beta_0 + \beta_1 Tonnage + e$ was used to calculate a prediction interval for Time when Tonnage = 10,000, I think the interval would be too short/narrow because the prediction interval will be too long/wide when tonnage increases (or low tonnage), but the prediction interval will be too short/narrow when tonnage decreases (or high tonnage).