HW2

Getong Zhong

2022-10-14

Problem 1

(a)

From the Figure 3.41, Distance vs. Standardized Residuals, we can clearly observe a nonlinear pattern for the variance. Besides, in the same figure we can see there is a leverage point over the absolute value of 2 of the graph, which is concerned to be an outlier, e.g bad leverage point. Therefore this model is not fit to the data.

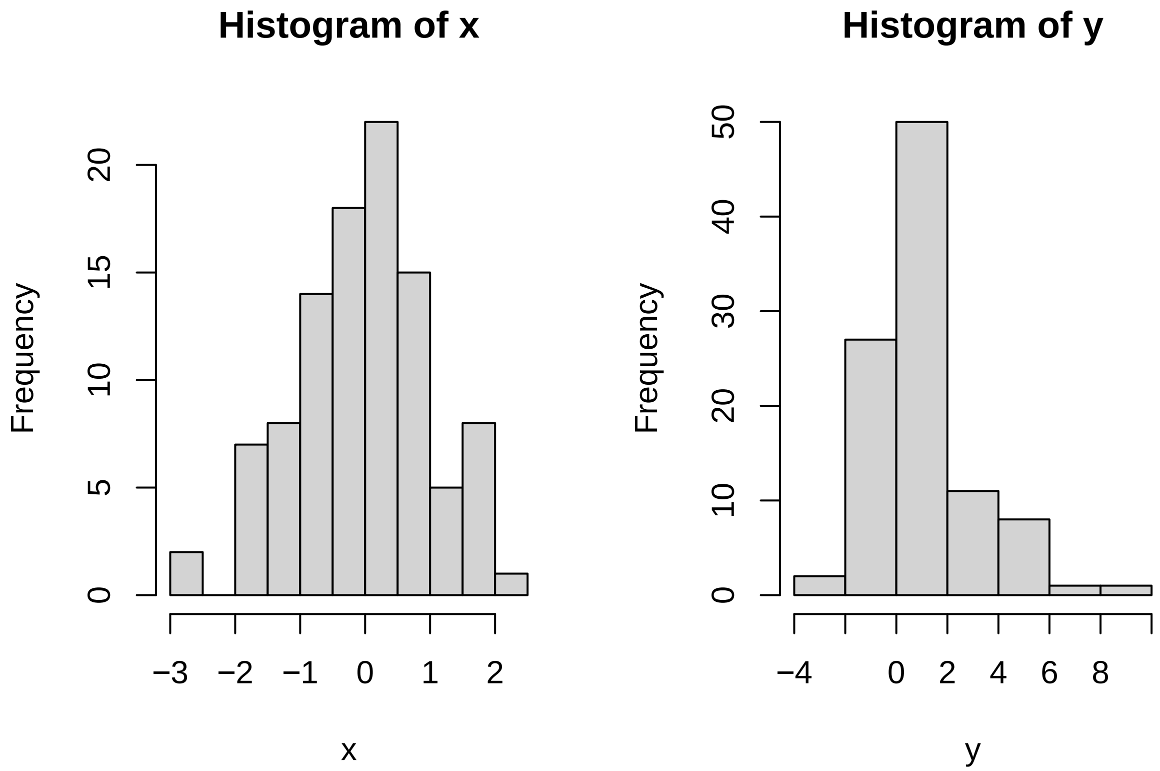
(b)

From the Figure 3.41, Distance vs. Fare, we can see a strong linear relationship between the distance and fare. However the from the plot of Distance vs. Standardized Residuals, it shows a non linear pattern which suggest the data might better fit in a nonlinear model. To improve the model, we can make a log transformation on the distance and remove the outiers from the data, e.g bad leverage point, and then fit the model.

Problem 2

I think the statement is TRUE. From the below results, I create x and y based on the context, and from the test we can see that mod2 has a better performance than mod1.

```r
x <- rnorm(100)
y <- x + x ^ 2 + rnorm(100)
par(mfrow = c(1, 2))
hist(x)
hist(y)
```



```
par(mfrow = c(1, 2))
mod1 <- lm(y ~ x)
mod2 <- lm(y ~ x + I(x ^ 2))
anova(mod1, mod2)
```

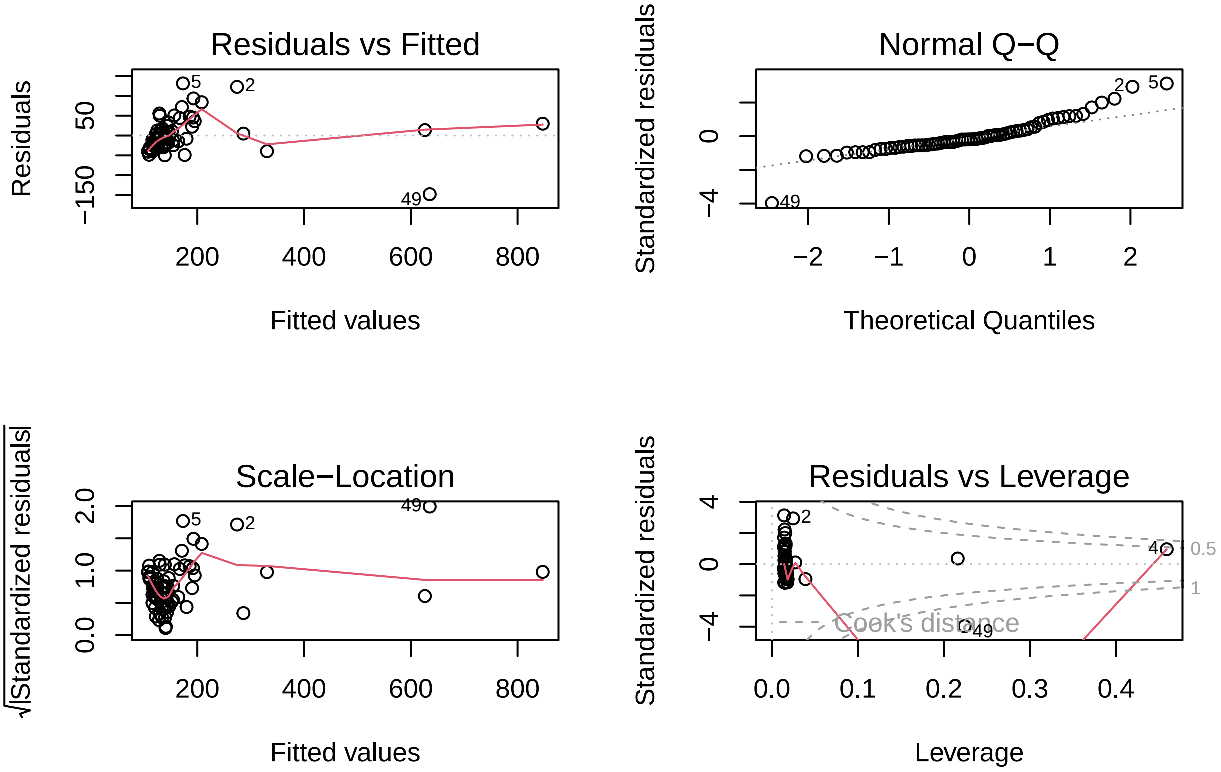
```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ x + I(x^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     98 321.01
## 2     97 102.92  1    218.09 205.54 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Problem 3

Part A

```

adr<-read.csv("C:/Users/tonyg/Desktop/Academic/Grad/HUDM 5126/AdRevenue.csv")
mod1<-lm(AdRevenue ~ Circulation, data = adr)
par(mfrow = c(2, 2))
plot(mod1)
```



(a)

From the result above, we observe: a potentially nonlinear relationship, non-constant variance of the residuals, and outliers and bad leverage data points. These all indicates the inappropriate of the model. Therefore, we are going to address some transformation to the data, by add square root or log to the response or independent variable. Then, we'll test the residuals for constant variance by performing a Breusch–Pagan test,

```
Adjmod1 <- lm(log(AdRevenue) ~ Circulation,data = adr)
adjmod2 <- lm(sqrt(AdRevenue) ~ sqrt(Circulation),data = adr)
adjmod3 <- lm(sqrt(AdRevenue) ~ Circulation,data = adr)
adjmod4 <- lm(AdRevenue ~ sqrt(Circulation),data = adr)

bptest(Adjmod1)$p.value
```

```
##        BP
## 0.3039185
```

```
bptest(adjmod2)$p.value
```

```
##          BP
## 0.005239347
```

```
bptest(adjmod3)$p.value
```

```
##        BP
## 0.1715771
```

```
bptest(adjmod4)$p.value
```

```
##           BP
## 1.75537e-10
```

only model 1 and model 3 shows they have a constant variance. We will then test the assumption of the normality of the errors for Models 1 and 3 by addressing the Shapiro-Wilk Test.

```
set.seed(5126)
shapiro.test(resid(Adjmod1))$p.value
```

```
## [1] 0.7483936
```

```
shapiro.test(resid(adjmod3))$p.value
```

```
## [1] 0.02504224
```

By the results, we will select adjmod1.

```
summary(Adjmod1)
```

```
##
## Call:
## lm(formula = log(AdRevenue) ~ Circulation, data = adr)
##
## 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 ***
## 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
```

```
# (i) for 0.5 million
pred1 = predict(Adjmod1,
        newdata = data.frame(Circulation = 0.5),
        interval = 'confidence')

exp(pred1)
```

(b)

```
##        fit      lwr      upr
## 1 119.7221 110.3163 129.9298
```

```
# (ii) for 20 million
pred2 = predict(Adjmod1,
        newdata = data.frame(Circulation = 20),
        interval = 'confidence')
exp(pred2)
```

```
##        fit      lwr     upr
## 1 529.3341 414.3819 676.175
```

For 0.5 million, the 95% confidence interval for the advertising revenue per page for magazines is (110.3163, 129.9298). For 20 millions, the 95% confidence interval for the advertising revenue per page for magazines is (414.3819, 676.175).

(c) From the plot in (a), we can see there are outliers and bad leverage points, which still not solved in my adjusted model.

Part B

(a) Going through the similar process as Part A. The final adjust model I got is adj mod4

```
adjmod1 <- lm(AdRevenue ~ sqrt(Circulation), data = adr)
adjmod2 <- lm(sqrt(AdRevenue) ~ Circulation, data = adr)
adjmod3 <- lm(AdRevenue ~ Circulation, data = adr)
Adjmod4 <- lm(AdRevenue ^ (1/3) ~ Circulation, data = adr)


bptest(adjmod1)$p.value
```

```
##          BP
## 1.75537e-10
```

```
bptest(adjmod2)$p.value
```

```
##        BP
## 0.1715771
```

```
bptest(adjmod3)$p.value
```

```
##          BP
## 0.002271491
```

```
bptest(Adjmod4)$p.value
```

```
##        BP
## 0.2209658
```

```
set.seed(5126)
shapiro.test(resid(adjmod2))$p.value
```

```
## [1] 0.02504224
```

```
shapiro.test(resid(Adjmod4))$p.value
```

```
## [1] 0.08657822
```

```
summary(Adjmod4)
```

```
##
## Call:
## lm(formula = AdRevenue^(1/3) ~ Circulation, data = adr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9690 -0.3428 -0.0100  0.2399  1.3629
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.83719    0.07135   67.80   <2e-16 ***
## Circulation  0.16387    0.01160   14.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5145 on 68 degrees of freedom
## Multiple R-squared:  0.7459, Adjusted R-squared:  0.7422
## F-statistic: 199.6 on 1 and 68 DF,  p-value: < 2.2e-16
```

```
# (i) for 0.5 million
pred1 = predict(Adjmod4,
        newdata = data.frame(Circulation = 0.5),
        interval = 'confidence')

pred1^3
```

(b)

```
##        fit    lwr      upr
## 1 119.0317 109.37 129.2463
```

```
# (ii) for 20 million
pred2 = predict(Adjmod4,
        newdata = data.frame(Circulation = 20),
        interval = 'confidence')
pred2^3
```

```
##        fit      lwr      upr
## 1 534.3267 457.4391 619.3805
```

For 0.5 million, the 95% confidence interval for the advertising revenue per page for magazines is (109.37, 129.2463). For 20 millions, the 95% confidence interval for the advertising revenue per page for magazines is (457.4391, 619.3805).

(c) From the plot in (a) from Part A, we can see there are outliers and bad leverage points, which still not solved in my adjusted model.

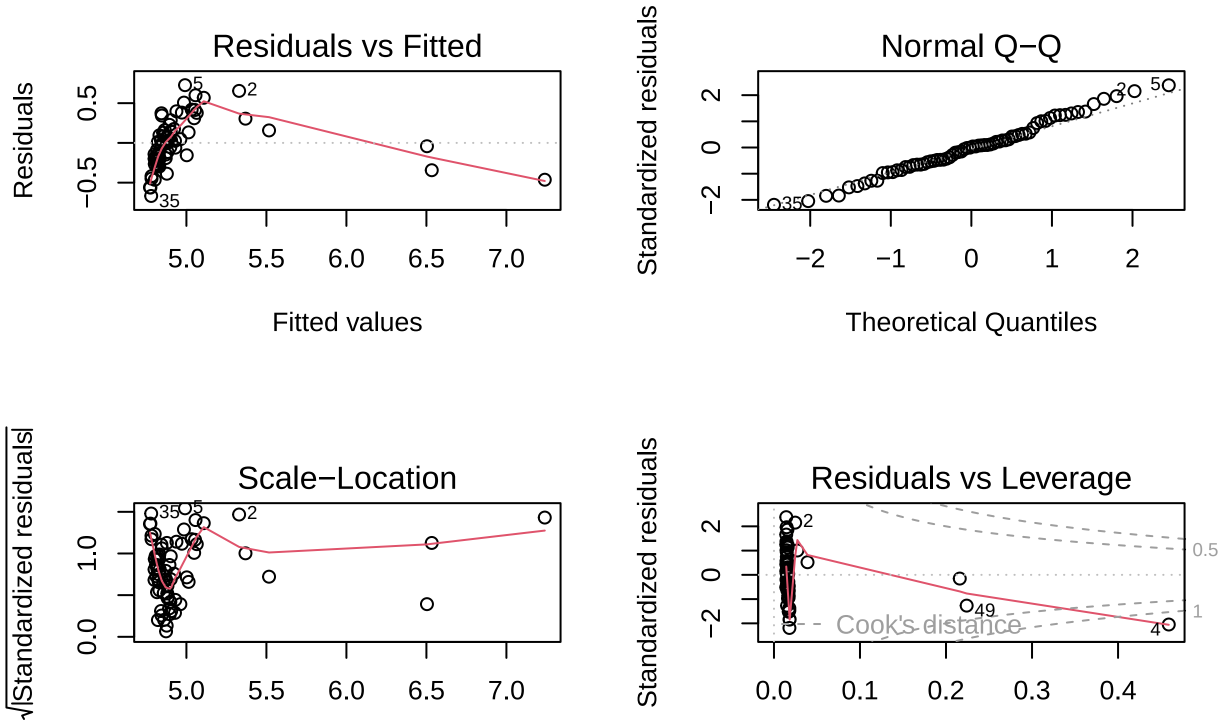
Part C

(a) The R-squared for model from Part A is 0.6399, and the R-squared for the model from Part B is 0.7459. Since both the graph have similar plot results, I would choose the model from Part B as the better on based on the R-squared value

```
summary(Adjmod1)
```

```
## 
## Call:
## lm(formula = log(AdRevenue) ~ Circulation, data = adr)
## 
## 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 ***
## 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
```

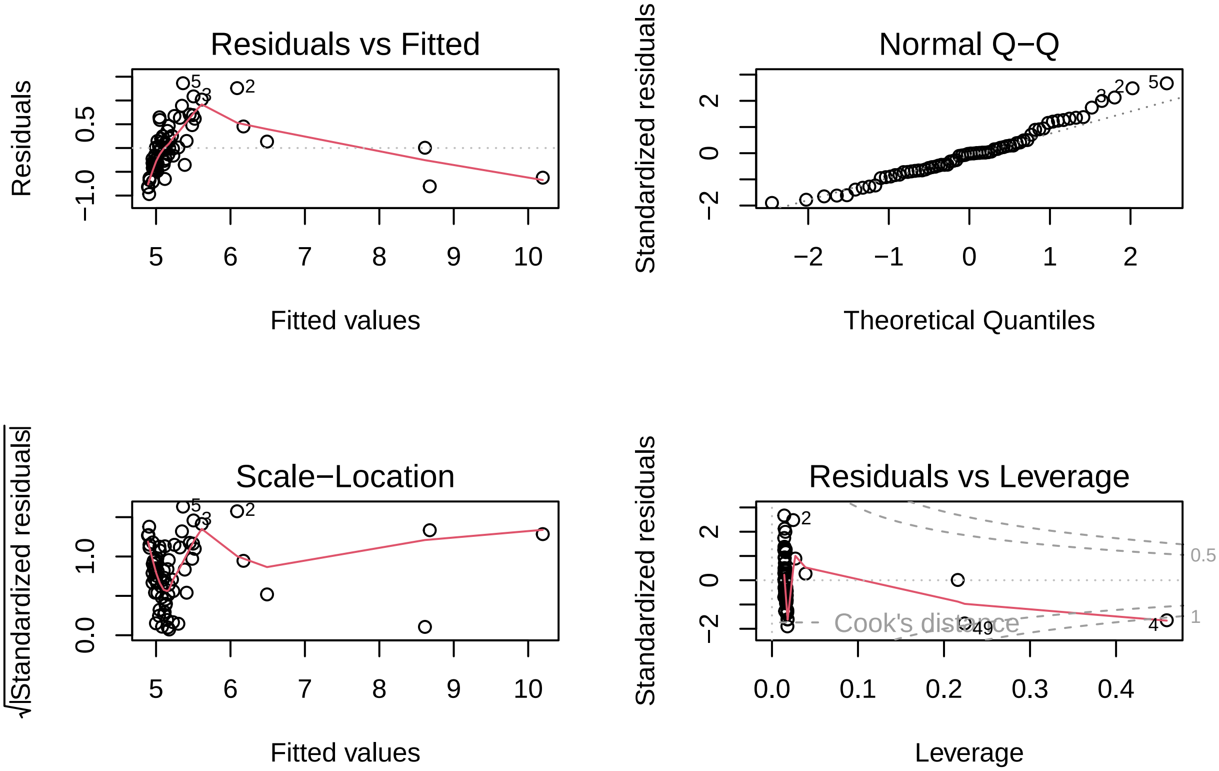
```
par(mfrow = c(2, 2))
plot(Adjmod1)
```



```
summary(Adjmod4)
```

```
##
## Call:
## lm(formula = AdRevenue^(1/3) ~ Circulation, data = adr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9690 -0.3428 -0.0100  0.2399  1.3629
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.83719    0.07135   67.80   <2e-16 ***
## Circulation  0.16387    0.01160   14.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5145 on 68 degrees of freedom
## Multiple R-squared:  0.7459, Adjusted R-squared:  0.7422
## F-statistic: 199.6 on 1 and 68 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(Adjmod4)
```



(b) For 0.5 million, the 95% confidence interval for the advertising revenue per page for magazines are respectively (110.3163, 129.9298) and (109.37, 129.2463) for part a and b; and the 95% confidence interval for the advertising revenue per page for magazines is(414.3819, 676.175) and (457.4391, 619.3805) for part a and b. Although the have very similar 95% confidence intervals, strictly we should recommend the smaller confidence interval. Therefore I recommend the confidence intervals both from Part B model.

Problem 4

(a)

No. Since as the value of Tonnage grows, there are some outliers come up, and one outlier is already get out of the range of absolute value of 2 from the standardized residuals, I think the model does not fit well in the straight line regression model.

(b)

I think the interval might be too large for Tonnage equals to 10000 since many point clustered when Tonnage is small and outliers come up when Tonnage get larger. Also, aswe can observe from the plot the variance of the data grows rapidly as the value of Tonnage grows, therefore as the value reach to 10000, the high variance of the data might leads to a longer interval.