Fall 2022 STAT 706 HW 3

Reina Li

Chapter 3.4 : 5a-e, 6, 7, 8 Pt1a-b, 8 Pt2a-b, 8 Pt3

5 (a)

This conclusion is not good because dependent on the R^2 value ($R^2 = 0.9986$) and the t-value ($t - value = 412.768$), so it is doesn't give us anything information about the assumptions of regression.

5 (b)

In model 3.10, when we look at the plotted data points with a regression line, it seems like a linear model fit well. However, when we look at the Scale-Location plot, we can see an increasing trend, which means that there is a non-constant error variance. Lastly, when we look at the Normal Q-Q plot, the middle of the points almost fit well on a straight line, but begin to curve at the extremities. We can also see that there are several leverage points. To overcome these shortcomings, I recommend possibly removing leverage points, transforming the model to get a constant error variance or adding a non-linear term to the model.

5 (c)

For model 3.11, when we look at the plotted data points with a regression line, it seems like the new model fit better than model 3.10. When we look at the Scale-Location plot, we can see almost no trend, which is good, because it implies that there is a constant error variance. So, I think that model 3.11 is an improvement over model 3.10 in terms of predicting Suggested Retail Price. However, when we look at the Normal Q-Q plot, the middle of the points almost fit well on a straight line, but begin to curve at the extremities, and there are still several leverage points.

5 (d)

The estimated coefficient of $log(DealerCost)$ in model 3.11 is 1.014836. First, the coefficient is statistically significant because it is associated with a p-value < 0.05. So, the model 3.11 suggests that log(suggested retail price) is a function of the log(dealer cost). The coefficient is positive, so we can say that dealer cost increases suggested retail price.

5 (e)

For model 3.11, when we look at the Normal Q-Q plot, the middle of the points almost fit well on a straight line, but begin to curve at the extremities, and there are still several leverage points. To improve model 3.11, I recommend removing leverage points and transforming the model further.

6

The inverse response plot fails to produce an estimate value of λ close to the correct value of λ in this situation because the distribution of x is skewed.

7

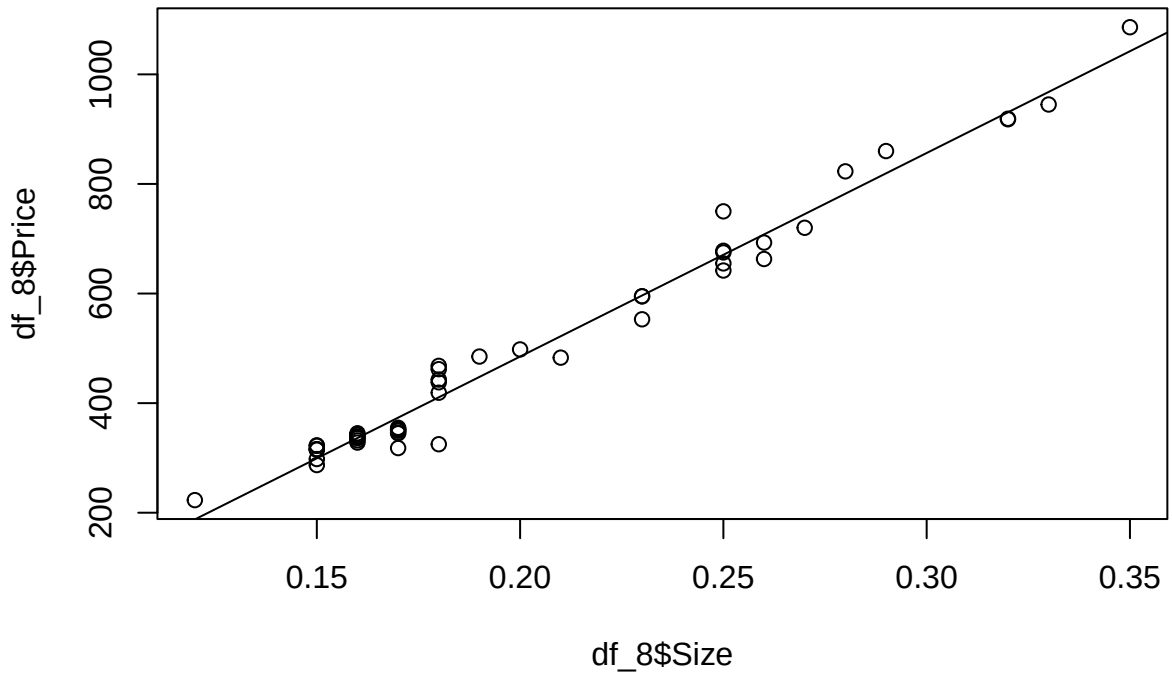
$$Var(f(Y)) = [f'(E(Y))]^2 Var(Y)$$
$$Var(f(Y)) = [f'(\mu)]^2 \mu^2$$
$$Var(f(Y)) = c$$
$$[f'(\mu)]^2 \mu^2 = c$$
$$[f'(\mu)]^2 = c/\mu^2$$
$$\sqrt{[f'(\mu)]^2} = \sqrt{c}/\sqrt{\mu^2}$$
$$f'(\mu) = \sqrt{c}/\mu$$
$$\int f'(\mu) = \int \sqrt{c}/\mu$$
$$f(\mu) = \sqrt{c} \int 1/\mu$$
$$f(\mu) = \sqrt{c} * log(\mu)$$

8 Part 1 (a)

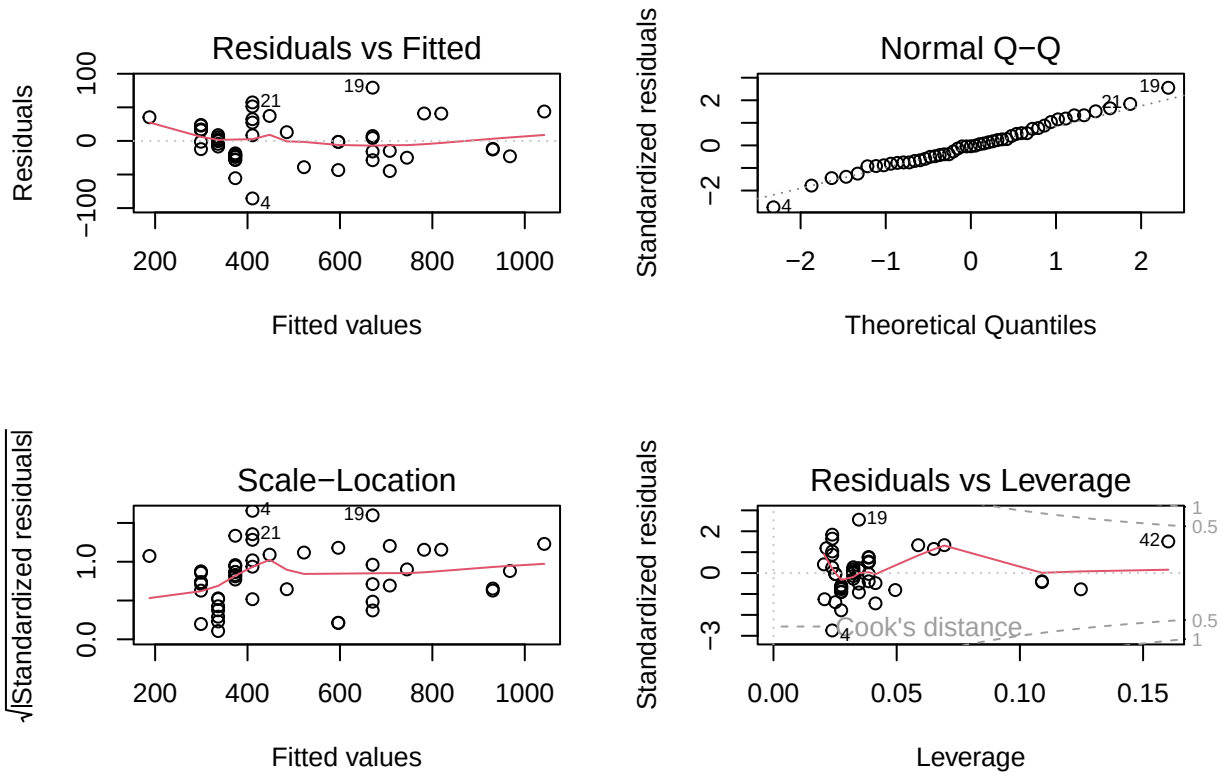
```
# Import data set-----
df_8 <- read.table("diamonds.txt", header = TRUE, sep = "")

# Fit the model to the data-----
model_q8 <- lm(Price ~ Size,
               data = df_8)
summary(model_q8)
```

```
##
## Call:
## lm(formula = Price ~ Size, data = df_8)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.654 -21.503  -1.203  16.797  79.295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -258.05      16.94  -15.23   <2e-16 ***
## Size         3715.02      80.41   46.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.6 on 47 degrees of freedom
## Multiple R-squared:  0.9785, Adjusted R-squared:  0.978
## F-statistic:  2135 on 1 and 47 DF,  p-value: < 2.2e-16
```

```r
# Plot of data and regression line
par(mfrow=c(1,1))
plot(df_8$Size, df_8$Price)
abline(model_q8)
```




```r
# Plot four graphs for lm
par(mfrow=c(2,2))
plot(model_q8)
```



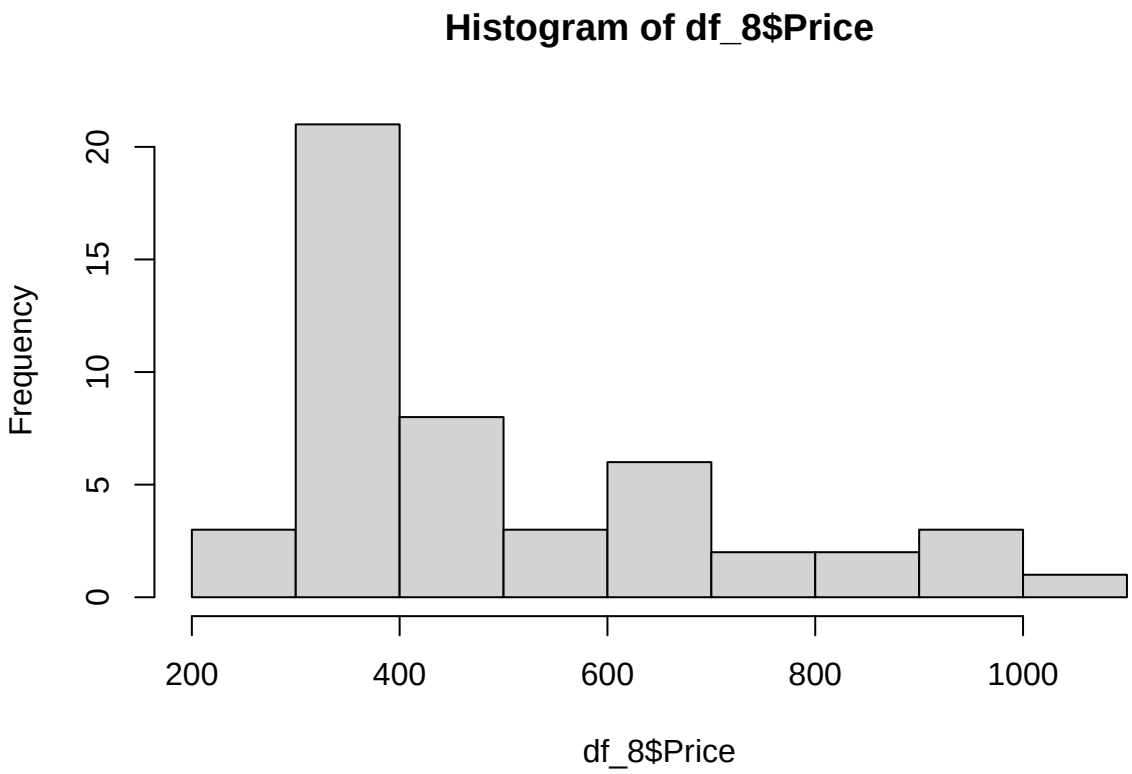
A simple linear regression model based on least squares that directly predicts Price from Size, without transforming either the predictor nor the response variable, is given by y =-258.0503726 + 3715.0219356x. The justification for my choice of model is because when you plot the data, the data seems to have a positive linear trend.

8 Part 1 (b)

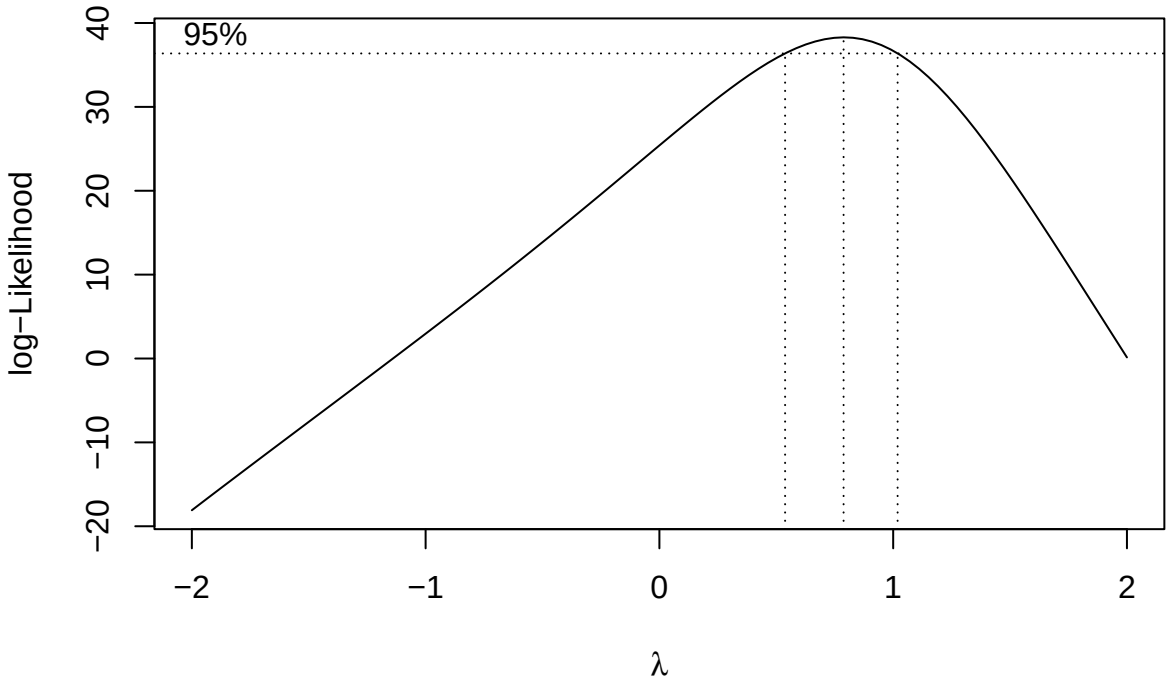
Even though the calculated R^2 value is really close to 1, (which is 0.9785), when we look at the Residuals vs. Fitted plot, there seems to be a non-linear trend. When we look at the Normal Q-Q plot, we can see points that leverage points (points 4, 19, and 21). When we look at the Scale-Location plot, there is a slight increasing trend. Looking at the plotted data points with the regression line, we can see that a linear model seems to fit well. If we apply a transformation to get a more constant error variance, we can improve this model. We can also improve this model by adding a non-linear trend to the model.

8 Part 2 (a)

```
# Histogram
hist(df_8$Price)
```




```
# Load library
library(MASS)

# Find optima lambda for box-cox transformation
bc <- boxcox(Price ~ Size,
             data = df_8)
```




```
(lambda <- bc$x[which.max(bc$y)])
```

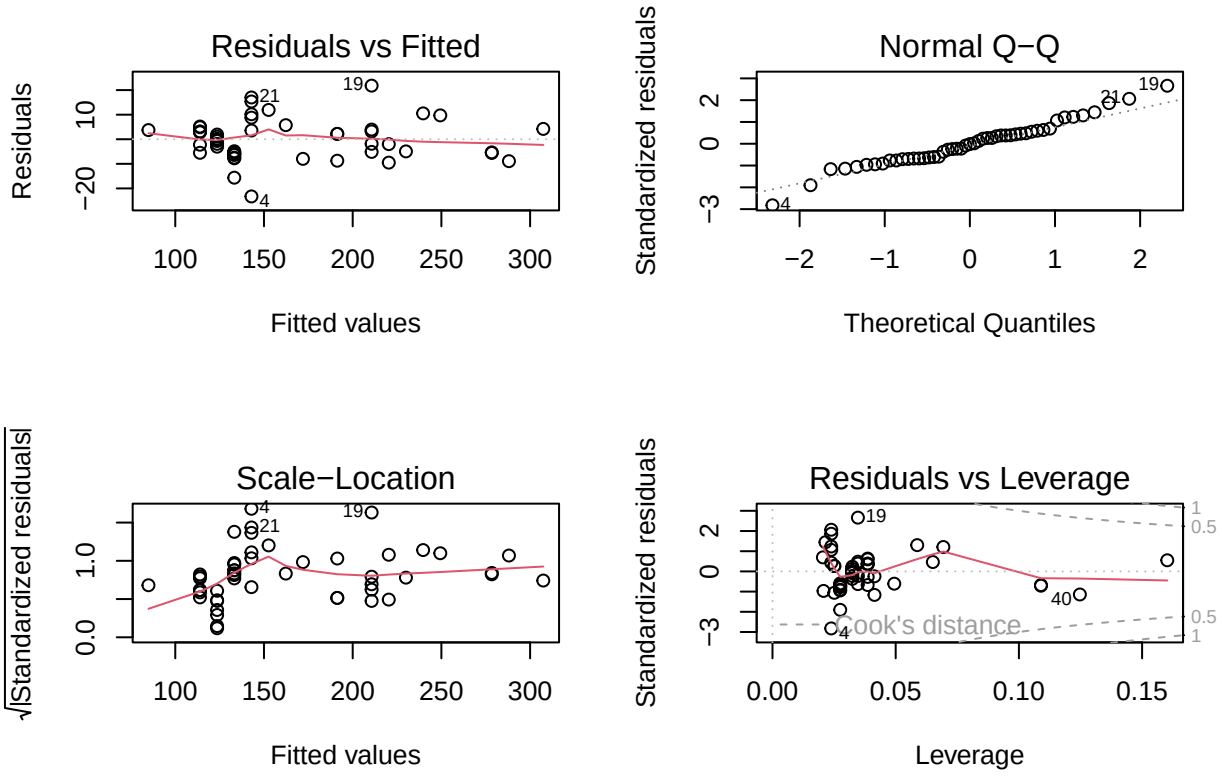
```
## [1] 0.7878788
```

```
model_8_2a <- lm(((Price^lambda-1)/lambda) ~ Size,
                 data = df_8)
summary(model_8_2a)
```

```
##
## Call:
## lm(formula = ((Price^lambda - 1)/lambda) ~ Size, data = df_8)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.2615  -5.5219  -0.1146   3.9383  21.8052
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -31.22       4.47  -6.985 8.62e-09 ***
## Size          967.55      21.22  45.606  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.338 on 47 degrees of freedom
```

```
## Multiple R-squared:  0.9779, Adjusted R-squared:  0.9774
## F-statistic:  2080 on 1 and 47 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(model_8_2a)
```



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 $\lambda = 0.7878788$. So, the new regression model replaced the original response variable with the variable $Price = (Price^{\lambda} - 1)/\lambda$ The reason why I did a Box-Cox transformation was because when I looked at the histogram of the data, the data was right skewed, implying that the assumption that the data is normally distributed was violated, and the relationship between the variables were not linear.

8 Part 2 (b)

Although the Box-Cox transformation I performed transformed the data to closely resemble a normal distribution, there are some weaknesses I observe in my model.When we look at the Normal Q-Q plot, we can see points that leverage points (points 4, 19, and 21). When we look at the Scale-Location plot, there is an increasing trend until fitted values=150, and then a slight increasing trend from there.

8 Part 3

I think Model B provides a better model because even though the R^2 value in Model B (R^2=0.9779) is slightly less than the R^2 value in Model A (R^2=0.9785), when we compare the Residuals vs Fitted plots,

the residuals behave better in Model B than in Model A. In Model B, most of the residuals are around the residual=0 line. However, there are some residuals that stand out from the basic pattern of residuals, suggesting that there are outliers in the data.