Stats 101A Homework 3

Damien Ha

2023-02-02

Contents

Problem 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

$\text{Fare} = \beta_0 = \beta_1 \text{Distance} + \ e$

Loading Data

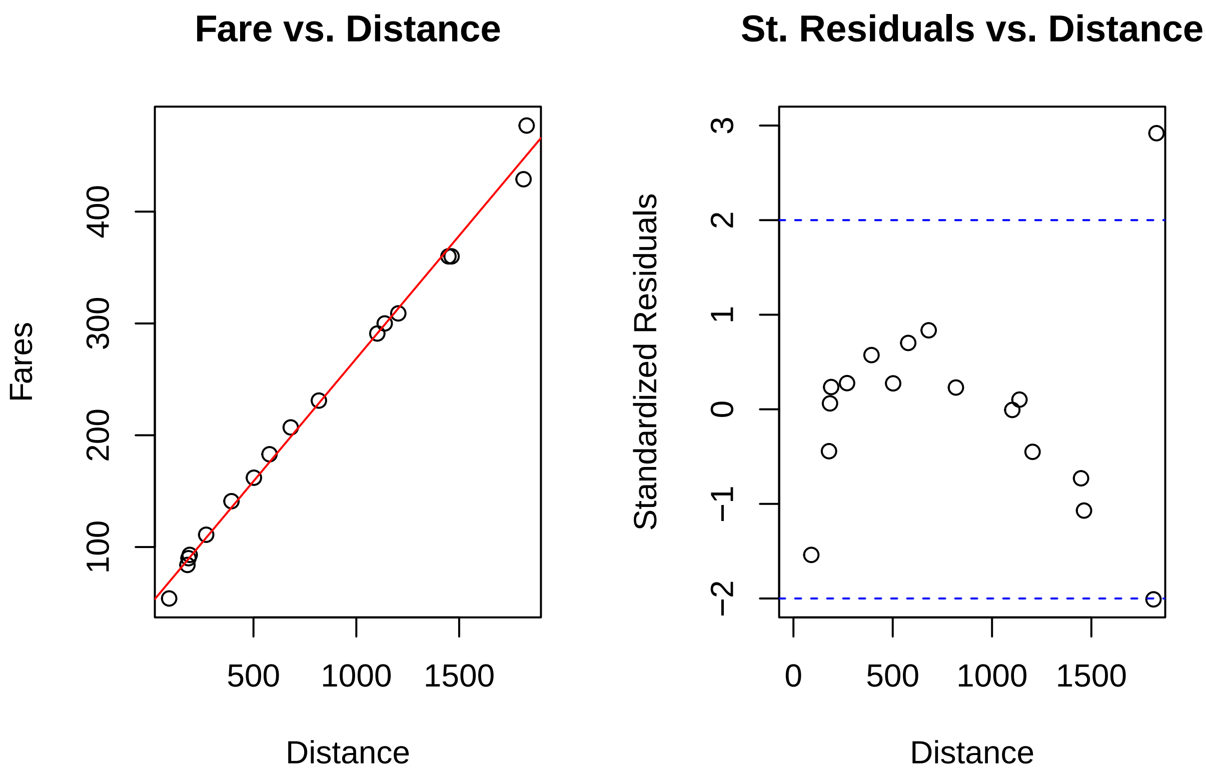
```
airfare <- read.table("airfares.txt", header = T)
```

Part 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 signifi- cant 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.

```
par(mfrow = c(1, 2))
plot(airfare$Distance, airfare$Fare, main = "Fare vs. Distance", xlab = "Distance",
     ylab = "Fares", pch = 1)
abline(lm(Fare ~ Distance, data = airfare), col = "red")

plot(airfare$Distance, rstandard(lm(Fare ~ Distance, data = airfare)),
     xlim = c(0, 1800), ylim = c(-2, 3), ylab = "Standardized Residuals",
     xlab = "Distance", main = "St. Residuals vs. Distance")

abline(h = c(-2, 2), col = "blue", lty = 2)
```



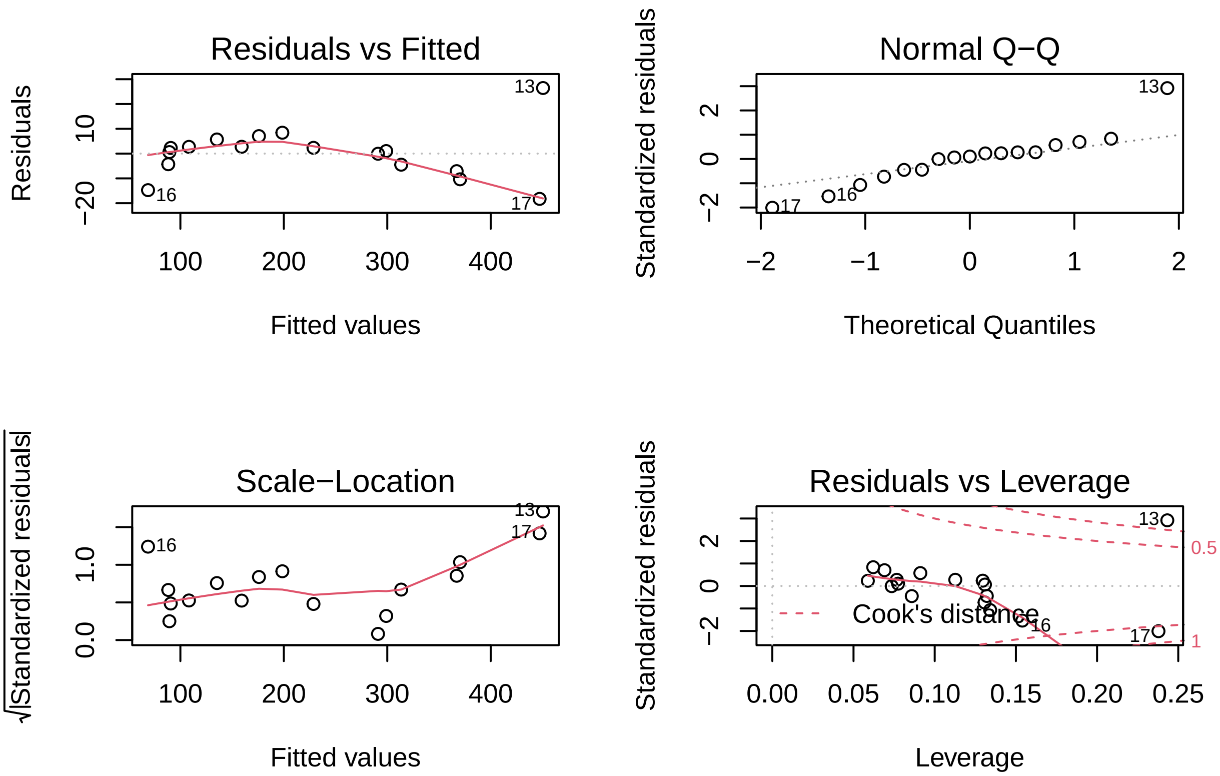
Provide a detailed critique of this conclusion.

```
summary(lm(Fare ~ Distance, data = airfare))
```

```
##
## Call:
## lm(formula = Fare ~ Distance, data = airfare)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.265  -4.475   1.024   2.745  26.440
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.971770   4.405493   11.12 1.22e-08 ***
## Distance     0.219687   0.004421   49.69  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.41 on 15 degrees of freedom
## Multiple R-squared:  0.994,  Adjusted R-squared:  0.9936
## F-statistic:  2469 on 1 and 15 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(lm(Fare ~ Distance, data = airfare))
```



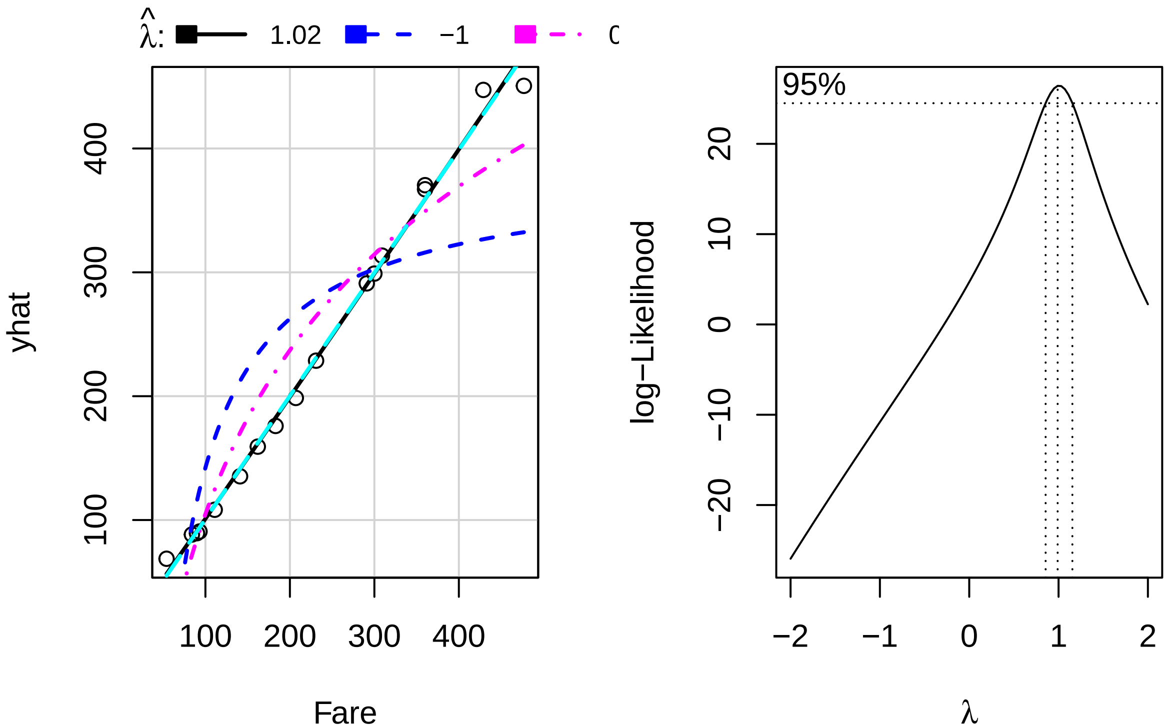

```
suppressWarnings(library(car))
```

```
## Loading required package: carData
```

```
par(mfrow=c(1,2))
inverseResponsePlot(lm(Fare ~ Distance, data = airfare))
```

```
##       lambda       RSS
## 1  1.024061  1605.994
## 2 -1.000000 81066.642
## 3  0.000000 22925.898
## 4  1.000000  1616.388
```

```
suppressWarnings(library(MASS))
boxcox(lm(Fare ~ Distance, data = airfare))
```




```
summary(powerTransform(lm(Fare ~ Distance, data = airfare)))
```

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    1.0069           1       0.8661       1.1477
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                            LRT df        pval
## LR test, lambda = (0) 43.51064  1 4.2166e-11
##
## Likelihood ratio test that no transformation is needed
##                               LRT df    pval
## LR test, lambda = (1) 0.009082514  1 0.92407
```

```
powerTransform(lm(Fare ~ Distance, data = airfare))$roundlam
```

```
## Y1
##  1
```

```
summary(powerTransform(cbind(airfare$Fare, airfare$Distance) ~ 1))
```

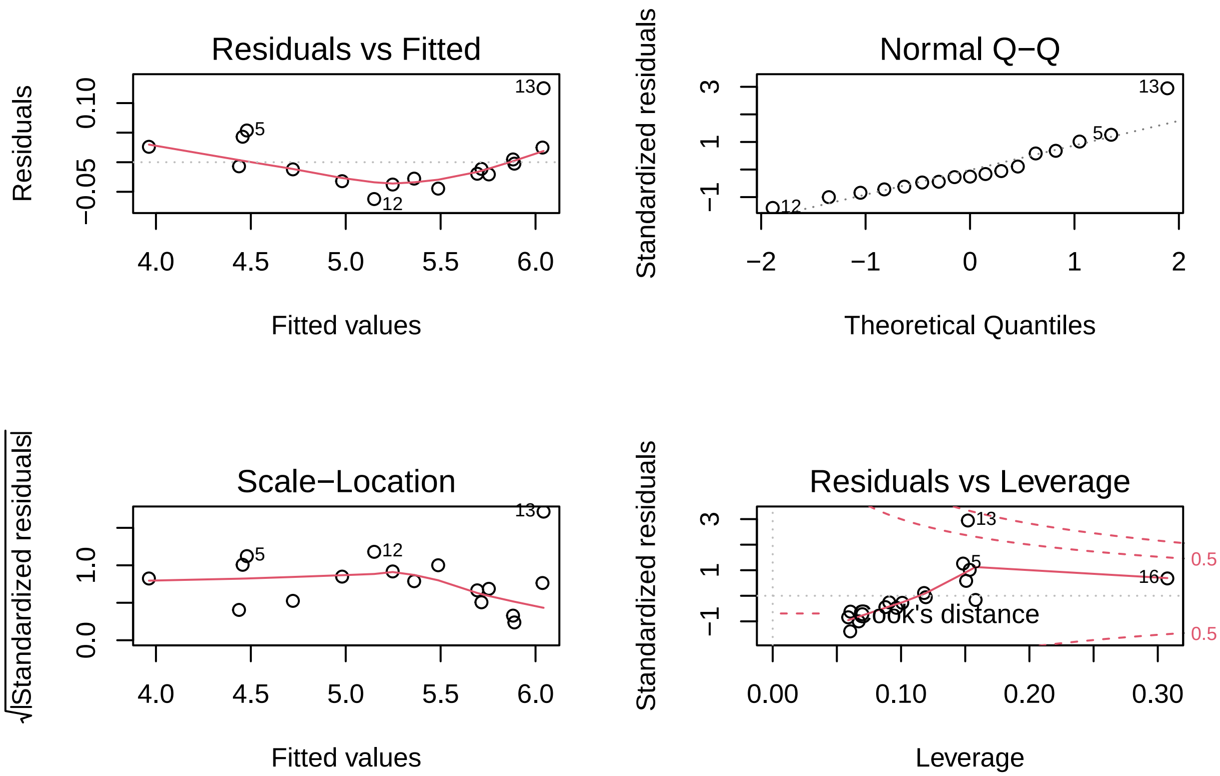
```
## bcPower Transformations to Multinormality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1   -0.0207           0      -0.4549       0.4135
## Y2    0.1098           0      -0.2315       0.4512
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                              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
```

```r
powerTransform(cbind(airfare$Fare, airfare$Distance) ~ 1)$roundlam
```

```
## Y1 Y2
##  0  0
```

```r
par(mfrow=c(2,2))
plot(lm(log(airfare$Fare) ~ log(airfare$Distance)))
```



The conclusion is partially correct. The predictor variable looks to have high statistical significance and the R^2 value of 99.4% suggests a strong relationship. That said, there are other variables that could be affecting the data, and the residual plot looks to have a slight curved pattern which shouldn't be the case. See some transformations above

Part 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).

```r
summary(lm(Fare ~ Distance, data = airfare))
```

```
##
## Call:
## lm(formula = Fare ~ Distance, data = airfare)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -18.265  -4.475   1.024   2.745  26.440
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.971770   4.405493   11.12 1.22e-08 ***
## Distance     0.219687   0.004421   49.69  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.41 on 15 degrees of freedom
## Multiple R-squared:  0.994,  Adjusted R-squared:  0.9936
## F-statistic:  2469 on 1 and 15 DF,  p-value: < 2.2e-16
```

From this summary and what we showed earlier in part a, the straight line regression model has some merits, but we seem to achieve a better fit if we transform both x and y

Problem 2

An analyst for the auto industry has asked for your help in modeling data on the prices of new cars. Interest centers on modeling suggested retail price as a func- tion of the cost to the dealer for 234 new cars. The data set, which is available on the book website in the file cars04.csv, is a subset of the data from http://www.amstat.org/publications/jse/datasets/04cars.txt

The first model fit to the data was

Suggested Retail Price $= \beta_0 + \beta_1$Dealer Cost $+ e$

Loading Data

```r
cars <- read.csv("cars04.csv")
```

Part A

Based on the output for model (3.10) the analyst concluded the following:

Since the model explains just more than 99.8% of the variability in Suggested Retail Price and the coefficient of Dealer Cost has a t-value greater than 412, model (1) is a highly effective model for producing prediction intervals for Suggested Retail Price.

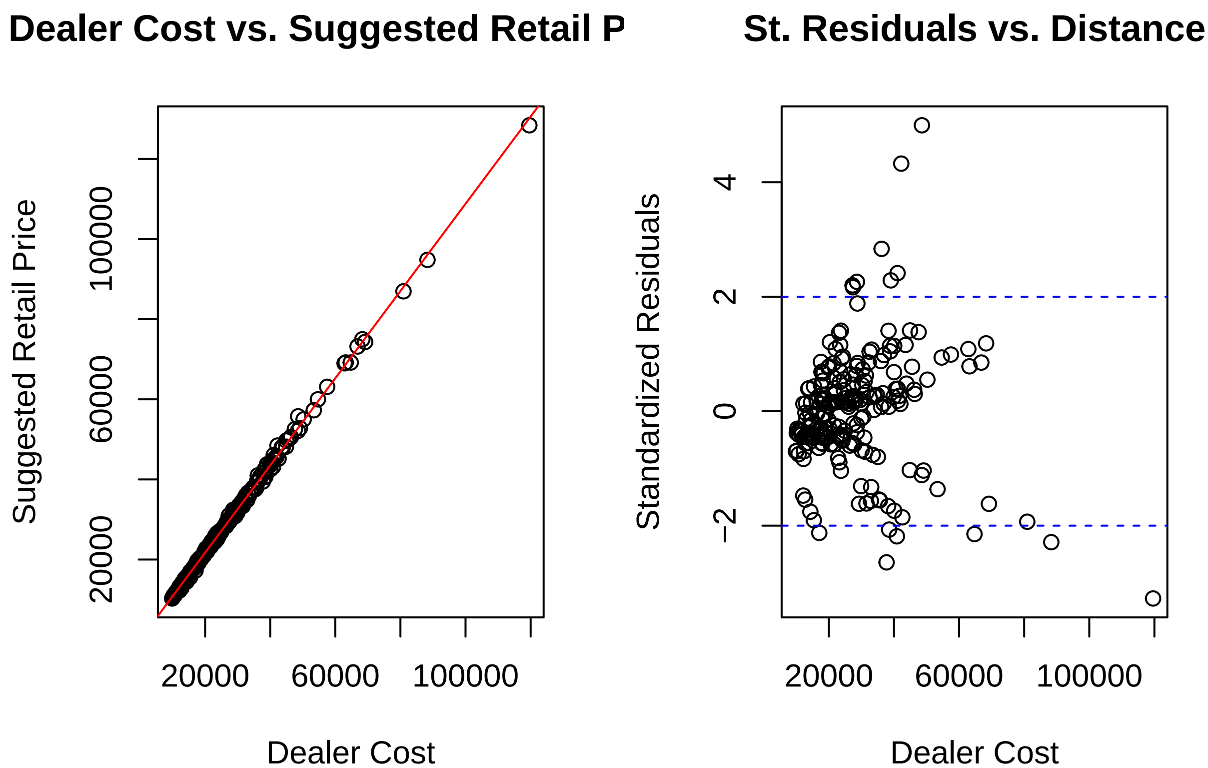
Provide a detailed critique of this conclusion.

```r
summary(lm(cars$SuggestedRetailPrice ~ cars$DealerCost))
```

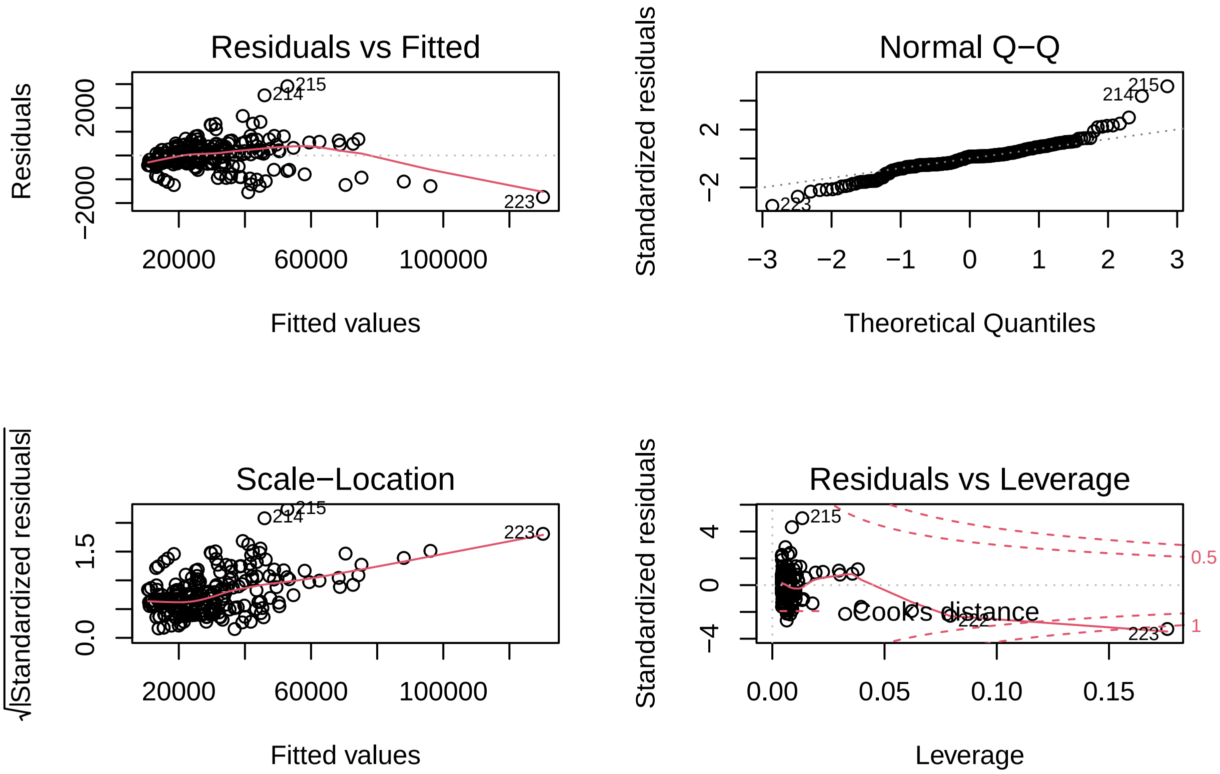
```
##
## Call:
## lm(formula = cars$SuggestedRetailPrice ~ cars$DealerCost)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1743.52  -262.59    74.92   265.98  2912.72
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -61.904248  81.801381  -0.757     0.45
## cars$DealerCost   1.088841   0.002638 412.768   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 587 on 232 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 1.704e+05 on 1 and 232 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(1, 2))
plot(cars$DealerCost, cars$SuggestedRetailPrice,
     main = "Dealer Cost vs. Suggested Retail Price", xlab = "Dealer Cost",
     ylab = "Suggested Retail Price", pch = 1)
abline(lm(cars$SuggestedRetailPrice ~ cars$DealerCost), col = "red")

plot(cars$DealerCost, rstandard(lm(cars$SuggestedRetailPrice ~ cars$DealerCost)),
     ylab = "Standardized Residuals",
     xlab = "Dealer Cost", main = "St. Residuals vs. Distance")
abline(h = c(-2, 2), col = "blue", lty = 2)
```




```
par(mfrow = c(2,2))
plot(lm(cars$SuggestedRetailPrice ~ cars$DealerCost))
```



It is true that based on the R^2 value, the model explains just more than 99.8% of the variability in Suggested Retail Price and the relationship appears fairly linear. From the qqplot, the data seems like it would be close to a normal distribution.

Part B

Carefully describe all the shortcomings evident in model (3.10). For each shortcoming, describe the steps needed to overcome the shortcoming.

There appear to be fairly large Cook's distance(s) in the residuals versus leverage plot, and these points might warrant further analysis. There may also be points that could be considered outliers, which could violate the regression assumptions.

The second model fitted to the data was

$log(\text{SuggestedRetailPrice}) = \beta_0 + \beta_1 log(\text{DealerCost}) + e$

Part C

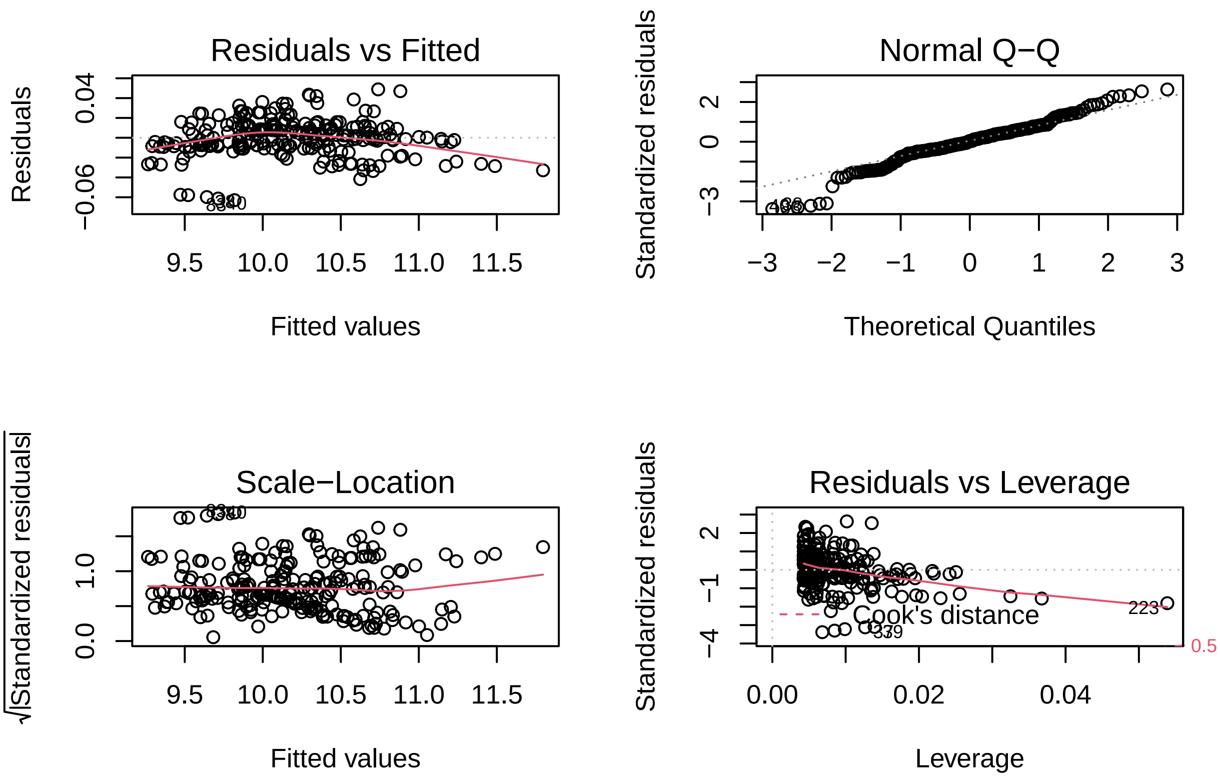
Is model (3.11) an improvement over model (3.10) in terms of predicting Suggested Retail Price? If so, please describe all the ways in which it is an improvement.

```
summary(lm(log(cars$SuggestedRetailPrice) ~ log(cars$DealerCost)))
```

```
##
## Call:
## lm(formula = log(cars$SuggestedRetailPrice) ~ log(cars$DealerCost))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.062920 -0.008694  0.000624  0.010621  0.048798
##
```

```
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.069459   0.026459  -2.625  0.00924 **
## log(cars$DealerCost)   1.014836   0.002616 387.942  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01865 on 232 degrees of freedom
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9985
## F-statistic: 1.505e+05 on 1 and 232 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(lm(log(cars$SuggestedRetailPrice) ~ log(cars$DealerCost)))
```



This model does look to be an improvement over 3.10. R^2 is still over 0.998. There seems to be smaller Cook's distances, the distribution should still be fairly normal from the qqplot, and based on the lines in Residuals vs Fitted and Scale-Location plots, the residuals appear to more closely averaging 0.

Part D

Interpret the estimated coefficient of log(Dealer Cost) in model (3.11).

Since the log transformation indicates percent changes, the value 1.014836 indicates that a 1% change in dealer cost is a 1.014836% increase in suggested retail price.

Part E

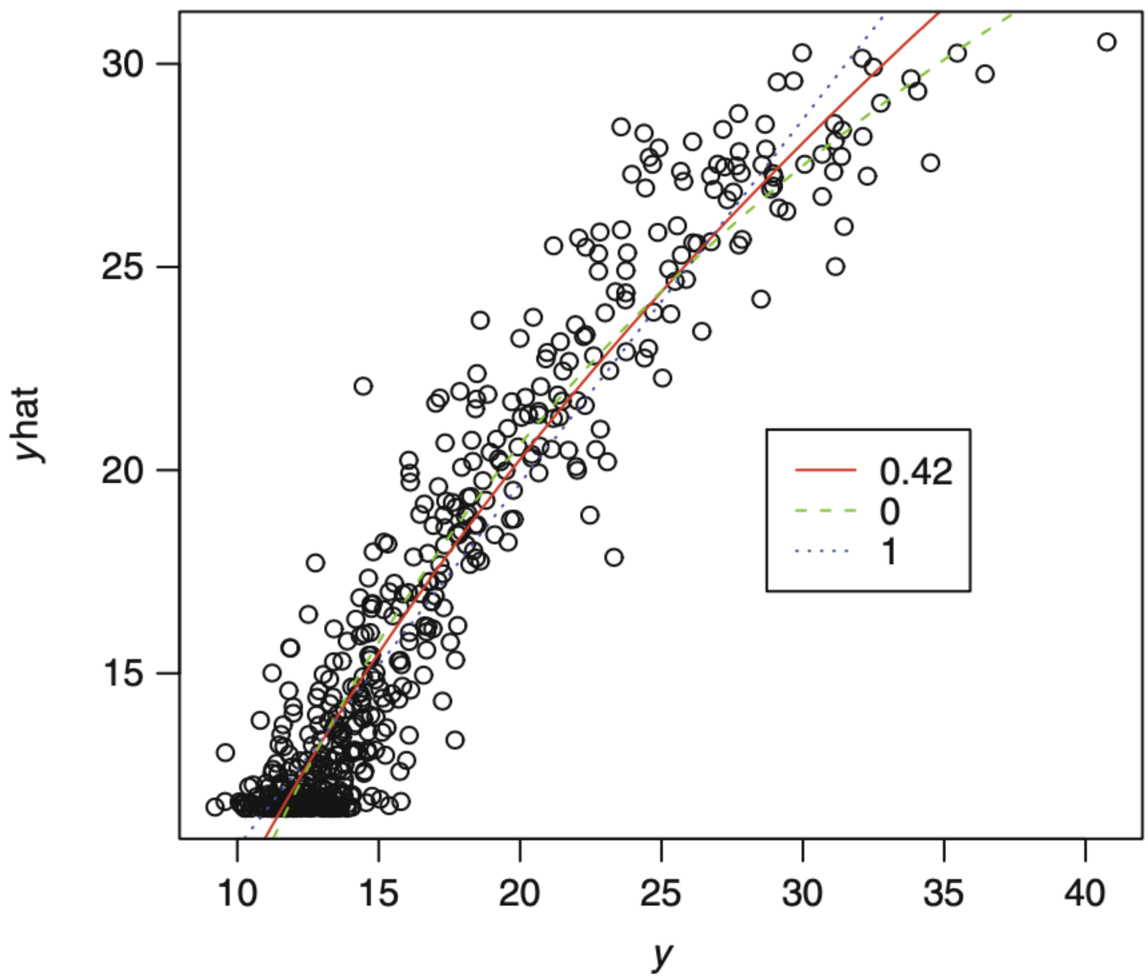
List any weaknesses apparent in model (3.11).

There are still values to the far right that look to be extreme and could be outliers in the data. The R^2 value is also very slightly lower than 3.10, though it seems to be a negligible amount

Problem 3

Chu (1996) discusses the development of a regression model to predict the price of diamond rings from the size of their diamond stones (in terms of their weight in carats). Data on both variables were obtained from a full page advertisement placed in the Straits Times newspaper by a Singapore-based retailer of diamond jewelry. Only rings made with 20 carat gold and mounted with a single diamond stone were included in the data set. There were 48 such rings of varying designs. (Information on the designs was available but not used in the modeling.)

```r
suppressWarnings(library(knitr))
knitr::include_graphics("ss_hw3_101a_p3.png")
```



The weights of the diamond stones ranged from 0.12 to 0.35 carats (a one carat diamond stone weighs 0.2 gram) and were priced between $223 and $1086. The data are available on the course web site in the file diamonds.txt.

Loading Data

```r
diamonds <- read.table("diamonds.txt", header = T)
```

Part 1

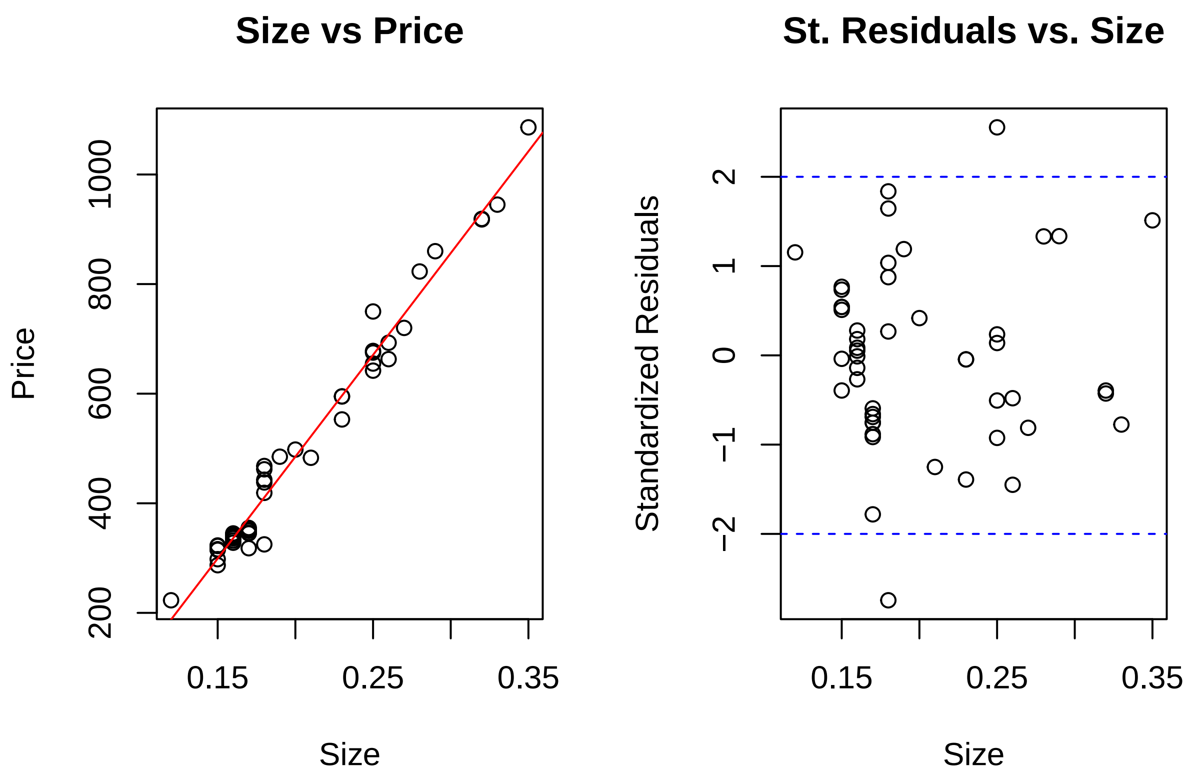
Part A

Develop a simple linear regression model based on least squares that directly predicts Price from Size (that is, do not transform either the predictor nor the response variable). Ensure that you provide justification for your choice of model.

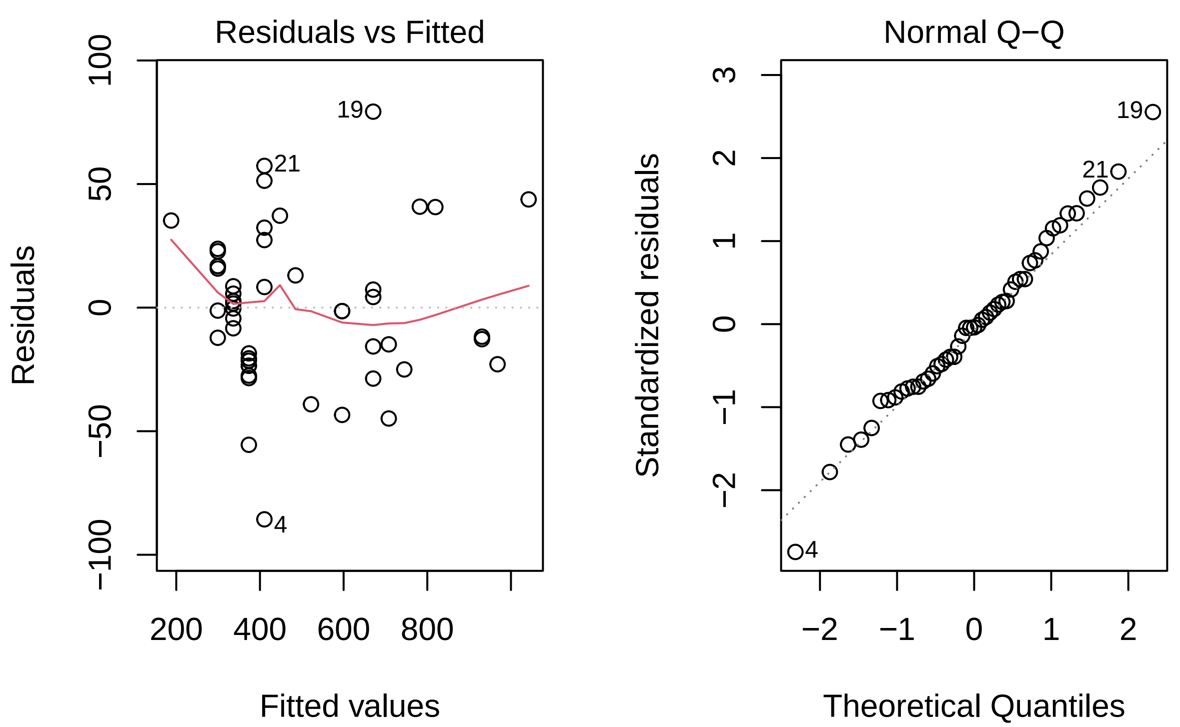
```r
summary(lm(Price ~ Size, data = diamonds))
```

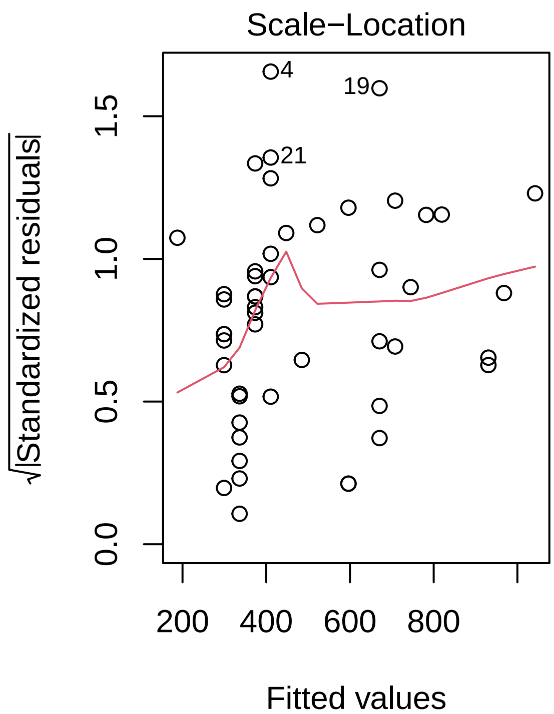
```
##
## Call:
## lm(formula = Price ~ Size, data = diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.654 -21.503  -1.203  16.797  79.295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -258.05      16.94  -15.23   <2e-16 ***
## Size         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
```

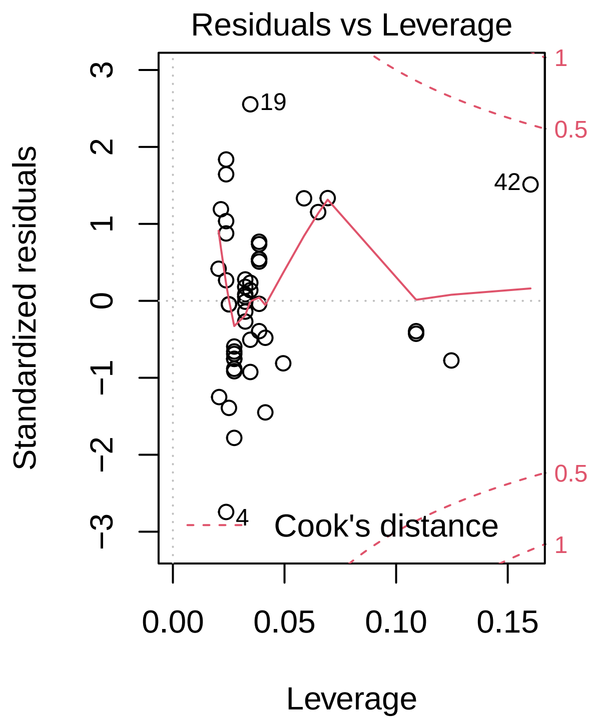
```r
par(mfrow = c(1, 2))
plot(diamonds$Size, diamonds$Price, main = "Size vs Price", xlab = "Size",
     ylab = "Price", pch = 1)
abline(lm(Price ~ Size, data = diamonds), col = "red")

plot(diamonds$Size, rstandard(lm(Price ~ Size, data = diamonds)),
     ylab = "Standardized Residuals",
     xlab = "Size", main = "St. Residuals vs. Size")
abline(h = c(-2, 2), col = "blue", lty = 2)
```




```
plot(lm(Price ~ Size, data = diamonds))
```







Part B

Describe any weaknesses in your model.

From the residuals vs fitted plot, there might not be constant variance based on the stacking of values to the left of the plot with less on the right. The residuals vs leverage appear to have some decent sized Cook's differences, and the lines in the residual graphs don't seem to be quite as flat as they should be for residuals averaging 0.

Part 2

Part A

Develop a simple linear regression model that predicts Price from Size (i.e., feel free to transform either the predictor or the response variable or both variables). Ensure that you provide detailed justification for your choice of model.

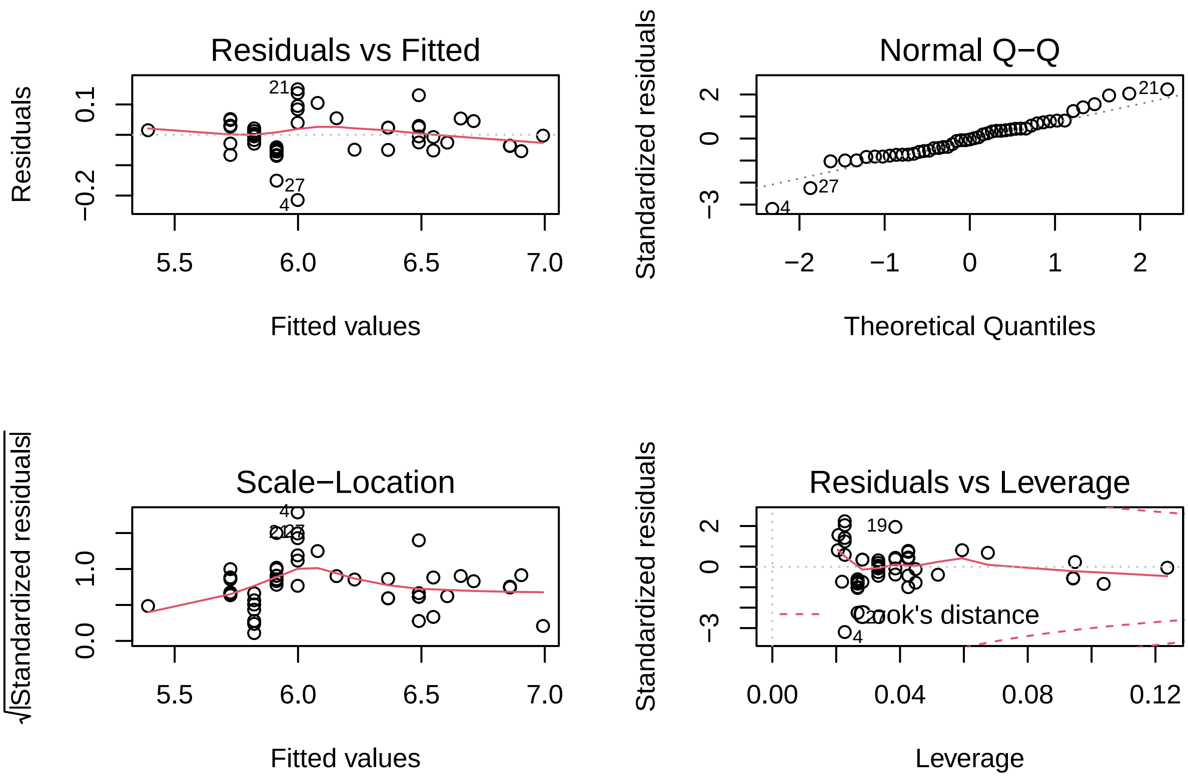
Logarithms can be used to observe percent effects. I'll transform this data by taking log of both x and y to observe how much as % increase in diamond size effects price

```
summary(lm(log(diamonds$Price) ~ log(diamonds$Size)))
```

```
##
## Call:
## lm(formula = log(diamonds$Price) ~ log(diamonds$Size))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.21460 -0.04646 -0.00274  0.03001  0.15005
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         8.56317    0.06221  137.65   <2e-16 ***
```

```
## log(diamonds$Size)  1.49566    0.03772   39.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06796 on 47 degrees of freedom
## Multiple R-squared:  0.971,  Adjusted R-squared:  0.9704
## F-statistic:  1572 on 1 and 47 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(lm(log(diamonds$Price) ~ log(diamonds$Size)))
```



Part B

Describe any weaknesses in your model.

The R^2 value is marginally lower than in the non-transformed plot. The normal QQ Plot seems to have a few off trend points suggesting a relationship that isnt perfectly linear

Part 3

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

Both p-values and R^2 values suggest a strong correlation, but the log transformed data seems to have more constant variance. Despite its few points off, the QQ Plot seems to still suggest a normal distribution, so the regression assumptions aren't violated.