STAT 757 Assignment 3

DUE X/XX/2018 11:59PM

AG Schissler

2/14/2018

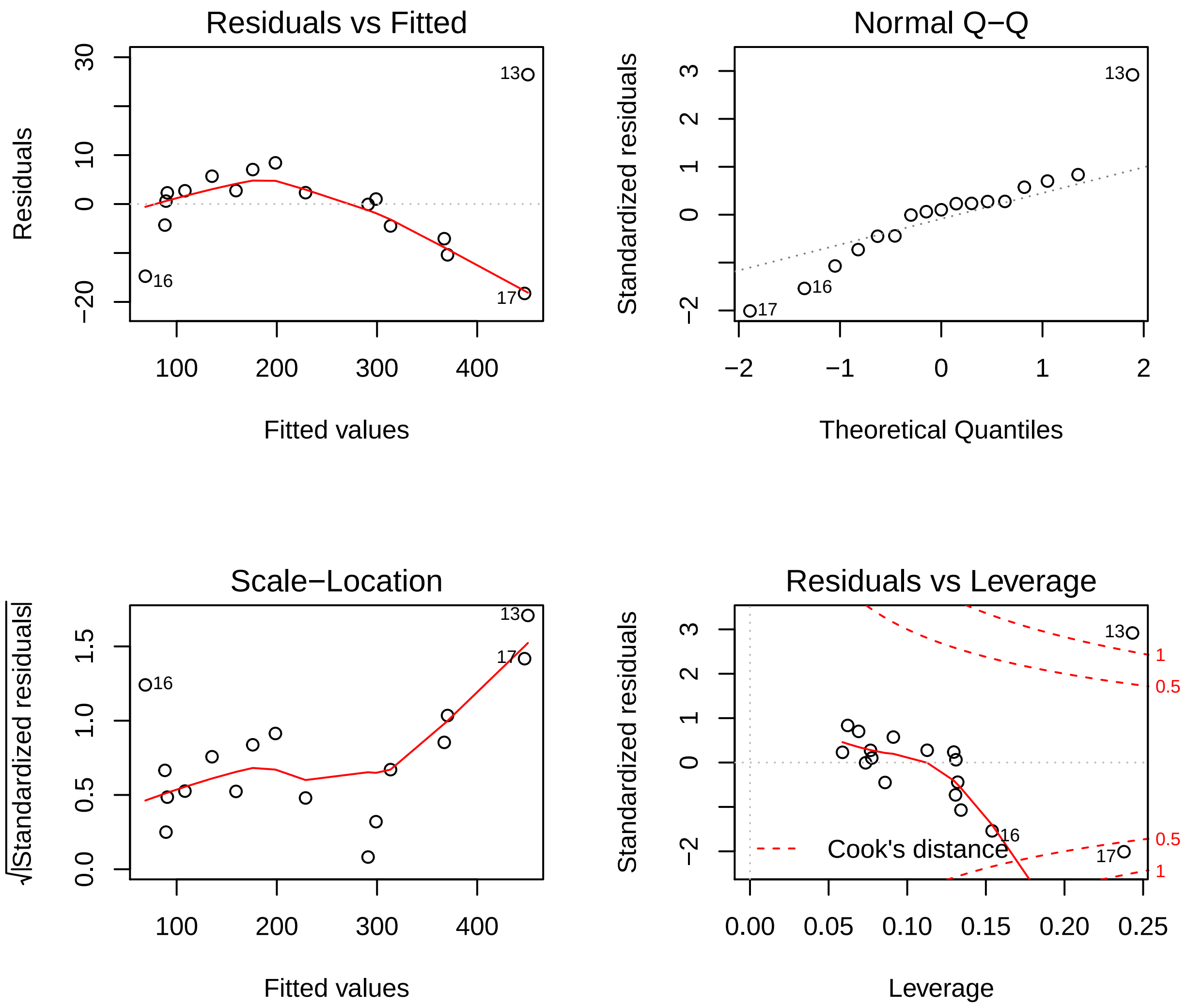
Instructions [20 points]

Modify this file to provide responses to the Ch.3 Exercises in Sheather (2009). You can find some helpful code here: http://www.stat. tamu.edu/~sheather/book/docs/rcode/Chapter3NewMarch2011.R. Also address the project milestones indicated below. Please email **both** your .Rmd (or roxygen .R) and one of the following either .HTML, .PDF, or .DOCX using the format SURNAME-FIRSTNAME-Assignment3.Rmd and SURNAME-FIRSTNAME-Assignment3.pdf.

```
data_dir <- "/Users/alfred/OneDrive - University of Nevada, Reno/Teaching/STAT_757/Sheather_data/Data"
```

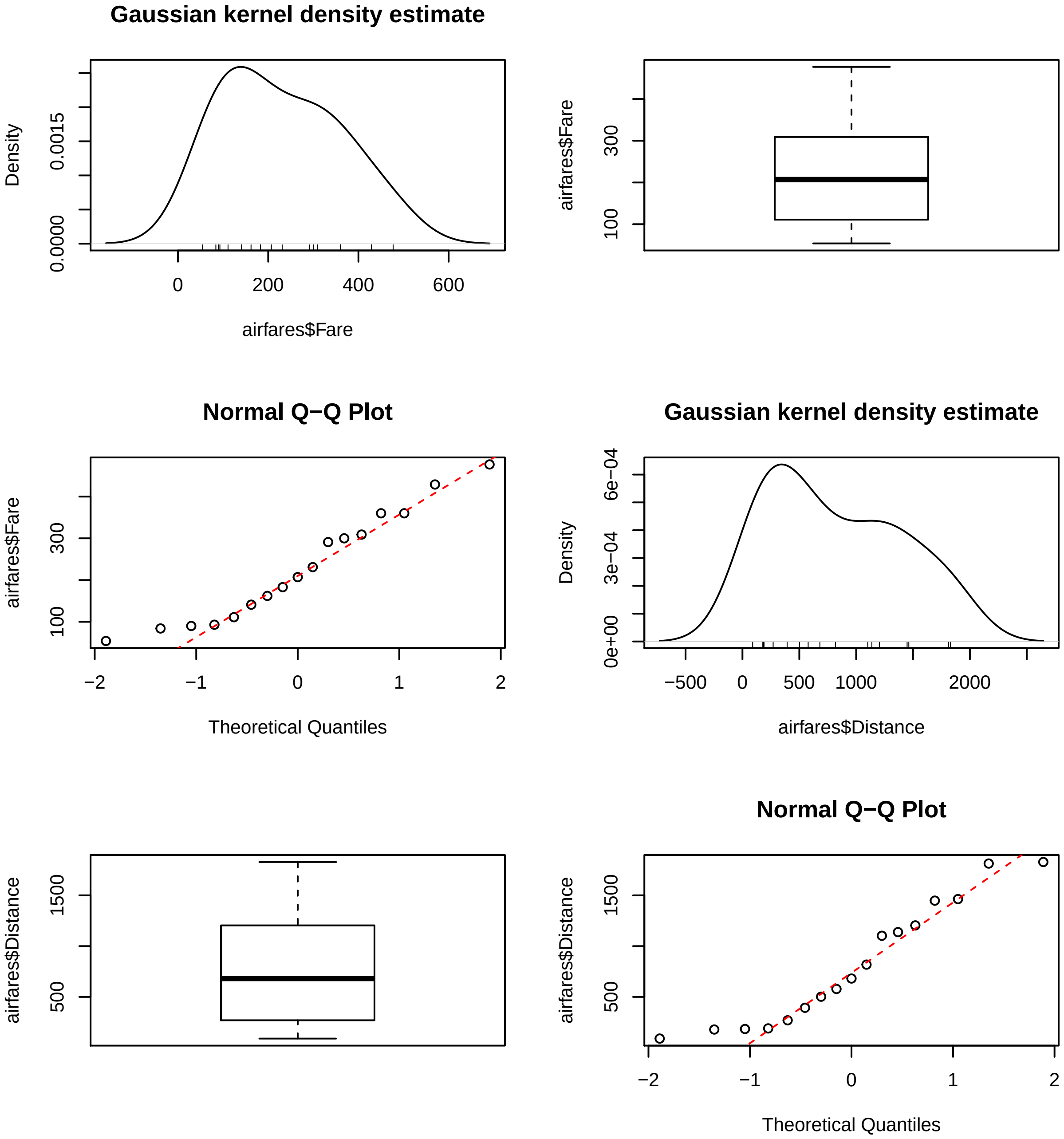
Exercise 3.4.1 [20 points]

```
airfares <- read.table(file.path(data_dir,"airfares.txt"), header = T)

## plot(x = airfares$Distance, y = airfares$Fare)
m1 <- lm(Fare ~ Distance, data = airfares)
## summary(m1)

par(mfrow = c(2,2))
plot(m1)
```



The residual and influencial point analysis reveals a distinct nonlinear pattern in the residuals and two high leverage (and "bad") points that greatly reduce confidence in the ordinary least squares assumptions. It is true that the 'Distance' explain a large about of the overall variation in the 'Fare', leading to a large R^2 statistic. However, the inferential devices are in question including prediction intervals and hypothesis testing on the coefficients or analysis of variance due to the lack of fit.

```
##

par(mfrow=c(3,2))
plot(density(airfares$Fare,bw="SJ",kern="gaussian"),type="l",
main="Gaussian kernel density estimate",xlab="airfares$Fare")
rug(airfares$Fare)
boxplot(airfares$Fare,ylab="airfares$Fare")
qqnorm(airfares$Fare, ylab = "airfares$Fare")
qqline(airfares$Fare, lty = 2, col=2)
plot(density(airfares$Distance,bw="SJ",kern="gaussian"),type="l",
main="Gaussian kernel density estimate",xlab="airfares$Distance")
rug(airfares$Distance)
boxplot(airfares$Distance,ylab="airfares$Distance")
qqnorm(airfares$Distance, ylab = "airfares$Distance")
qqline(airfares$Distance, lty = 2, col=2)
```



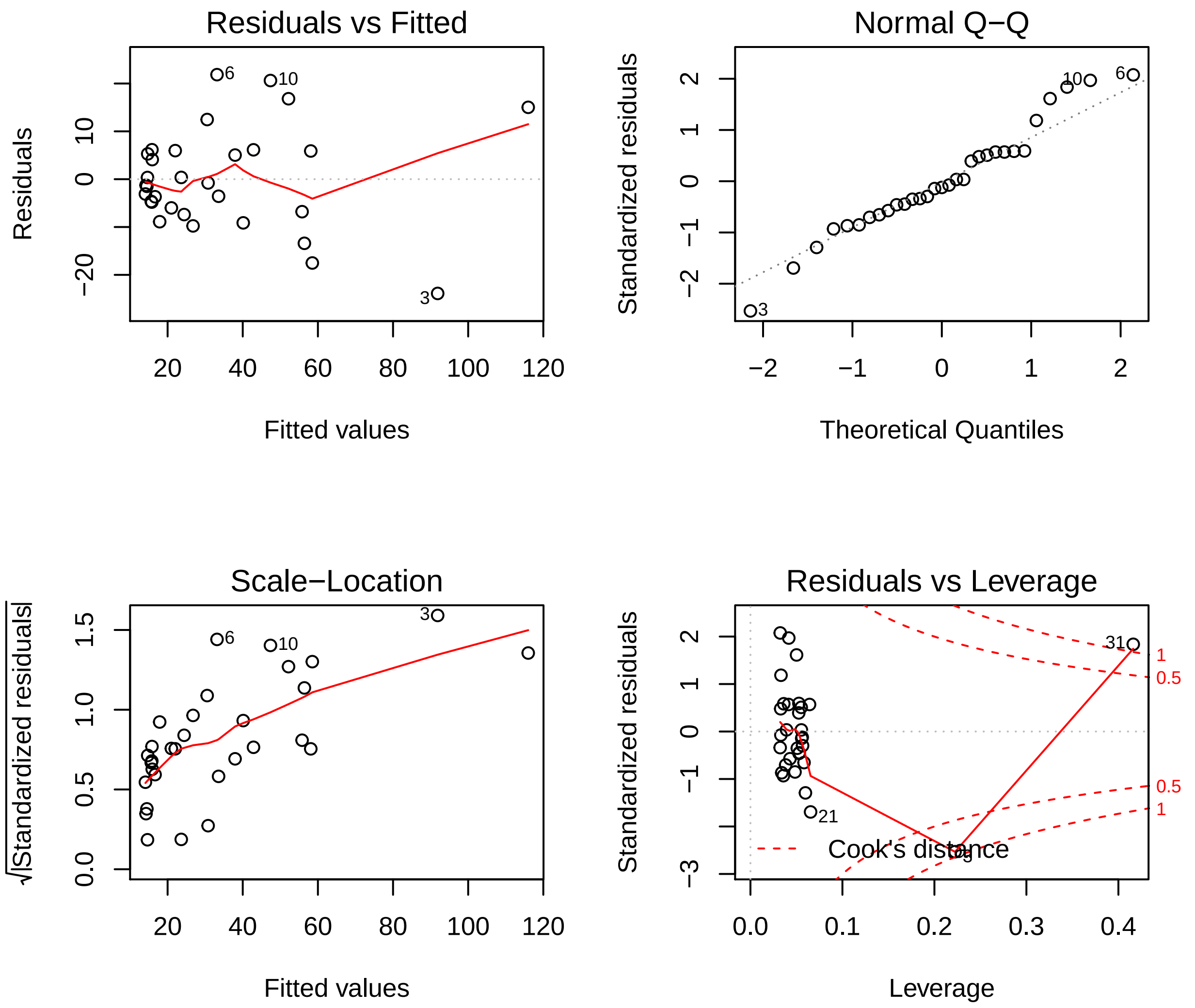
```
## library(alr3)
## summary(powerTransform(cbind(Fare, Distance)~1, data= airfares))

## so try log-log transformation
## plot(x = log(airfares$Distance), y = log(airfares$Fare))
## identify(log(airfares$Distance), log(airfares$Fare), Case)

## m2 <- lm(log(Fare) ~ log(Distance), data = airfares)

## remove outlier
## m2 <- update(m2, subset=(1:nrow(airfares))[-c(13)])

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

## quadratic
## m2 <- lm(Fare ~ Distance + I(Distance^2), data = airfares)
```

The shapes of the two variables exhibit bimodality. Moreover, the residual were nonlinear indicating, among several possibilities, that missing covariates are the reason for a lack of fit. These two facts combine to hint that a group variable may be missing, perhaps 'economy' vs 'business' fares? Moreover, two observations (case 13 and 17) exhibit high leverage and have a large Cook's distance and greatly influence the regression estimation. I would advise to speak with the domain expertise for missing covariate information and the validity of the outlying points.

Exercise 3.4.4 [20 points]

Part A

```
glakes <- read.table(file.path(data_dir,"glakes.txt"), header = T)
## head(glakes)

## plot(x = glakes$Tonnage, y = glakes$Time)
m1 <- lm(Time ~ Tonnage, data = glakes)
## summary(m1)

par(mfrow = c(2,2))
plot(m1)
```

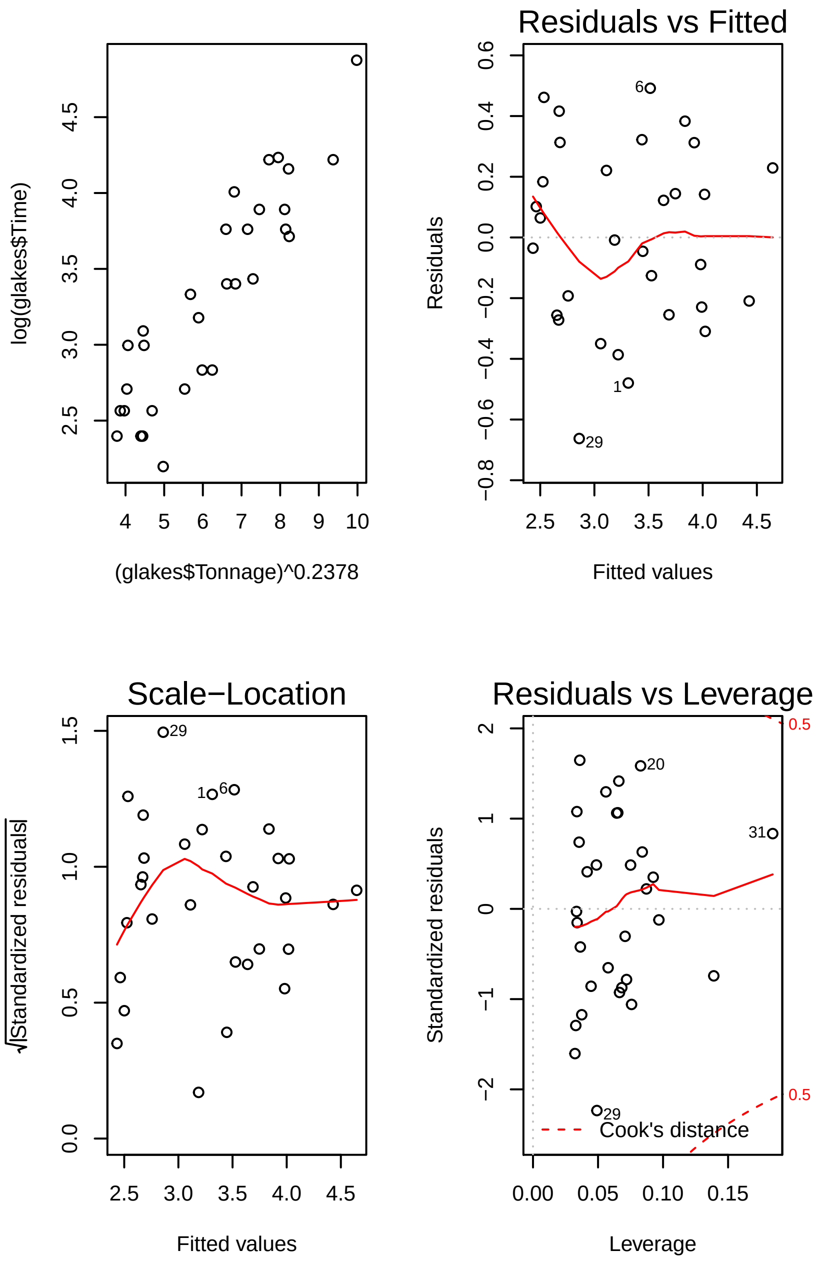


The residual analysis shows clear issues with inconstant variances and bad leverage points. Further the data are highly skewed. Transformation via multivariate Box-Cox and/or inverse-responses are recommended. For example,

```
## bcPower Transformations to Multinormality
##          Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## Time        0.0228           0      -0.3554       0.4011
## Tonnage     0.2378           0      -0.0046       0.4802
##
## Likelihood ratio tests about transformation parameters
##                                LRT df        pval
## LR test, lambda = (0 0)  3.7596  2 1.5262e-01
## LR test, lambda = (1 1) 45.3153  2 1.4451e-10
```

```
##
## Call:
## lm(formula = log(Time) ~ I(Tonnage^0.2378), data = glakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6626 -0.2421 -0.0086  0.2249  0.4917
##
```

```
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.0843     0.2041    5.31  1.1e-05
## I(Tonnage^0.2378)    0.3569     0.0316   11.30  3.8e-12
##
## Residual standard error: 0.304 on 29 degrees of freedom
## Multiple R-squared:  0.815,  Adjusted R-squared:  0.809
## F-statistic:  128 on 1 and 29 DF,  p-value: 3.84e-12
```



Part B

Under the original, untransformed model the inferences are invalid. I'd bet that the empirical coverage rates for prediction interval when Tonnage would be too narrow, giving false confidence in predictive ability. This is better of lack of data in this region with only two points more extreme in Tonnage and constant variance assumed thoroughout the model (ignoring the fact that variance in Time increases with Tonnage).

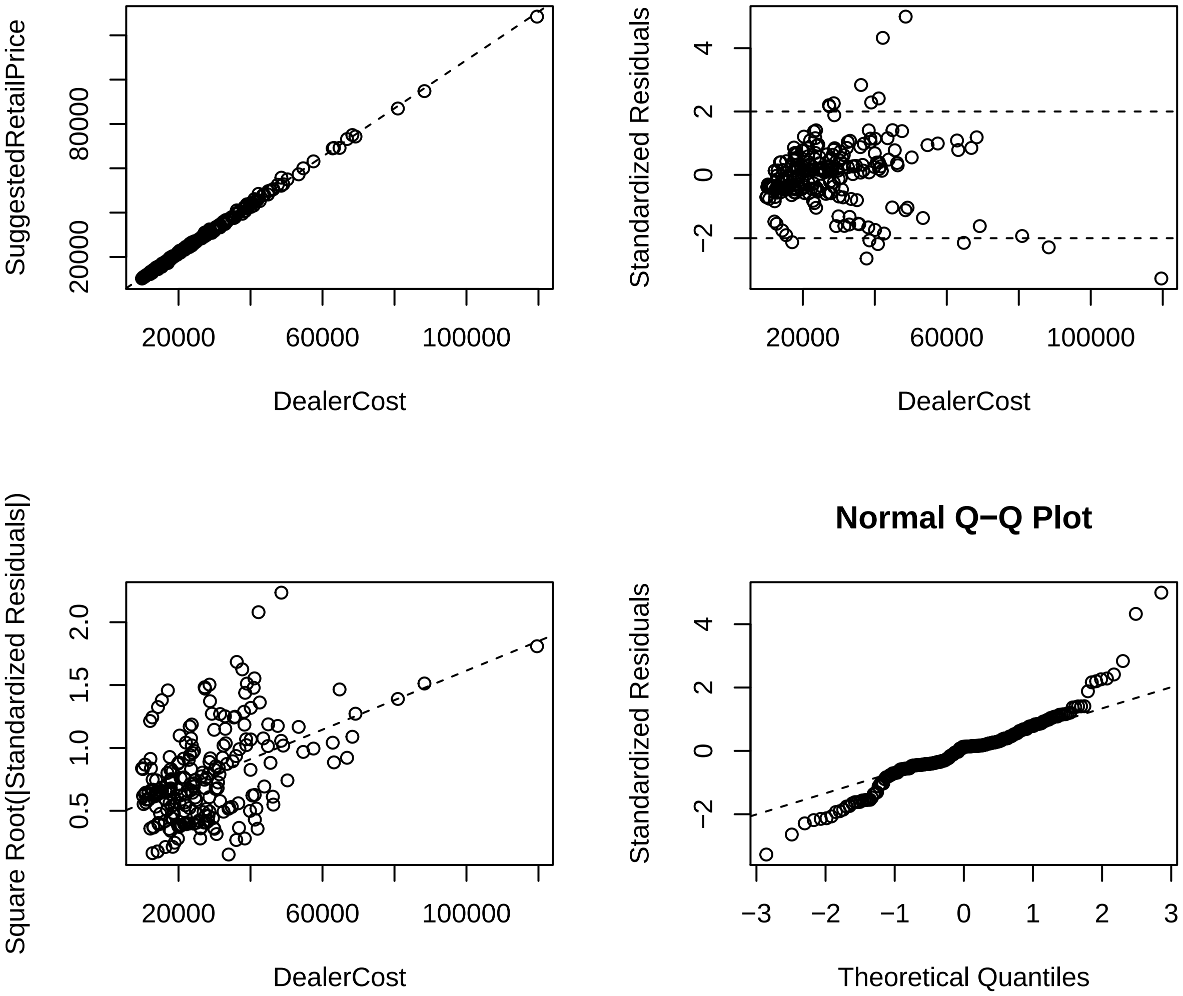
Exercise 3.4.5 [20 points]

Part A

```
cars04 <- read.csv(file.path(data_dir,"cars04.csv"),header=TRUE)
attach(cars04)

#Output from R on pages 110 and 111
m1 <- lm(SuggestedRetailPrice~DealerCost)
summary(m1)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ DealerCost)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1743.5  -262.6    74.9   266.0  2912.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -61.90425   81.80138   -0.76     0.45
## DealerCost    1.08884    0.00264  412.77   <2e-16
##
## Residual standard error: 587 on 232 degrees of freedom
## Multiple R-squared:  0.999,  Adjusted R-squared:  0.999
## F-statistic: 1.7e+05 on 1 and 232 DF,  p-value: <2e-16
```

```
#Figure 3.46 on page 110
par(mfrow=c(2,2))
plot(DealerCost,SuggestedRetailPrice)
abline(lsfit(DealerCost,SuggestedRetailPrice),lty = 2, col=1)
leverage1 <- hatvalues(m1)
StanRes1 <- rstandard(m1)
absrtsr1 <- sqrt(abs(StanRes1))
residual1 <- m1$residuals
plot(DealerCost,StanRes1, ylab="Standardized Residuals")
abline(h=2,lty=2)
abline(h=-2,lty=2)
plot(DealerCost,absrtsr1,ylab="Square Root(|Standardized Residuals|)")
abline(lsfit(DealerCost,absrtsr1),lty=2,col=1)
qqnorm(StanRes1, ylab = "Standardized Residuals")
qqline(StanRes1, lty = 2, col=1)
```



Inconstant variance in residual is evident that greatly reduce confidence in the ordinary least squares assumptions. It is true that 'DealCost' explain a large about of the overall variation in the 'SuggestedRetailPrice', leading to a large R^2 statistic. However, the inferential devices are in question including prediction intervals and hypothesis testing on the coefficients or analysis of variance due to the lack of fit.

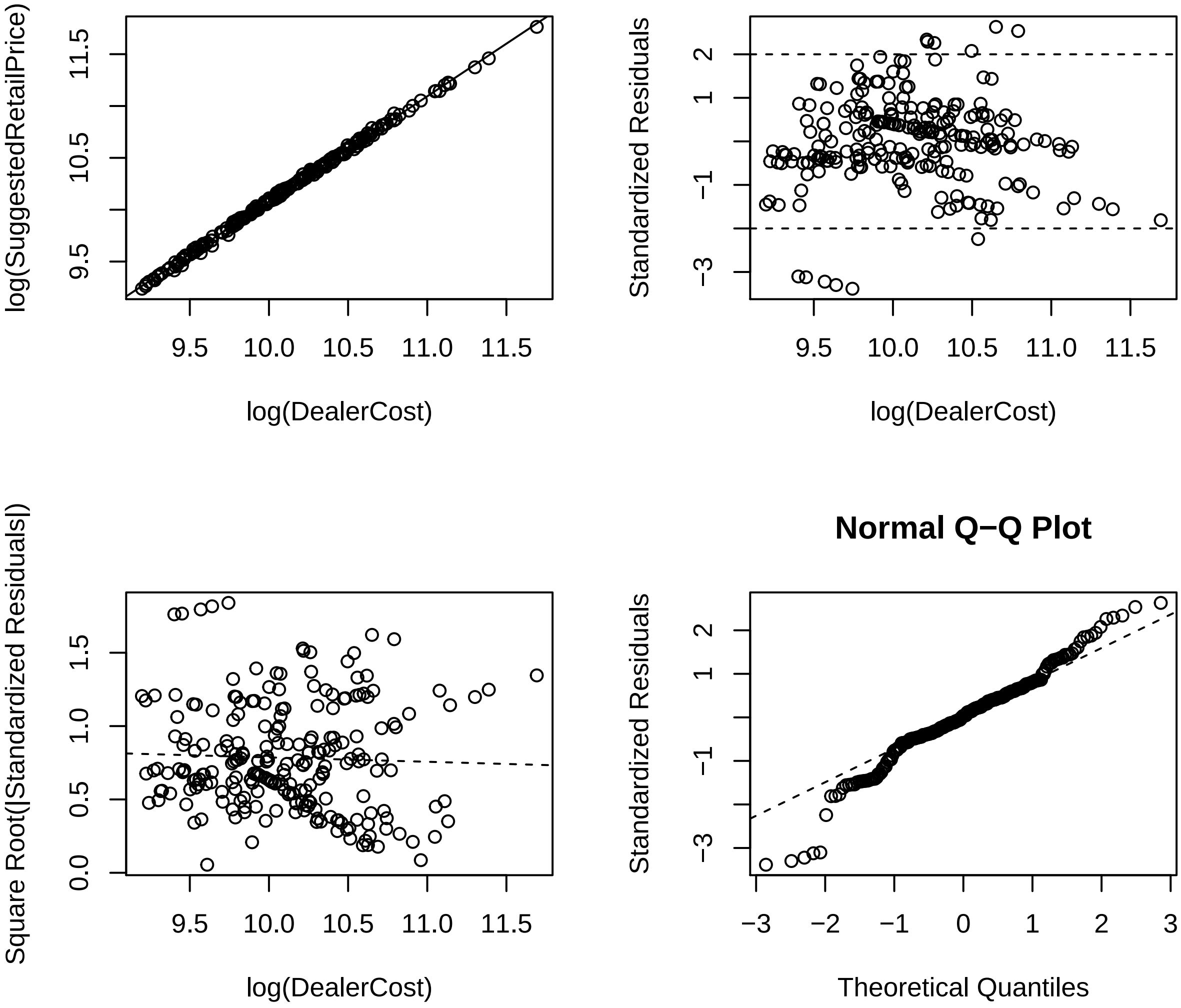
Part B

As mentioned above, the inconstant variance in the standardized residuals is most glaring modeling assumption violation. There is skewness in both the respones and predictor variables as well. Several residuals points lie outside (-2,2), but with n=234 this often happens. These data seem like good candidates for transformation to stabilize variance. The transformations could be estimated or, similar to the economic example in the chapter, a log-log transformation could be employed to find change in the percentage rates.

Part C

```r
#Output from R on page 111
m2 <- lm(log(SuggestedRetailPrice)~log(DealerCost))

#Figure 3.47 on page 111
par(mfrow=c(2,2))
```

```r
plot(log(DealerCost),log(SuggestedRetailPrice))
abline(lsfit(log(DealerCost),log(SuggestedRetailPrice)),lty = 1, col=1)
leverage2 <- hatvalues(m2)
StanRes2 <- rstandard(m2)
absrtsr2 <- sqrt(abs(StanRes2))
residual2 <- m2$residuals
plot(log(DealerCost),StanRes2, ylab="Standardized Residuals")
abline(h=2,lty=2)
abline(h=-2,lty=2)
plot(log(DealerCost),absrtsr2,ylab="Square Root(|Standardized Residuals|)")
abline(lsfit(log(DealerCost),absrtsr2),lty=2,col=1)
qqnorm(StanRes2, ylab = "Standardized Residuals")
qqline(StanRes2, lty = 2, col=1)
```



```r
detach(cars04)
```

It is clear that the new model fits much better with respect to the constant variance assumption (see the flat trend line in bottom-left plot - $\sqrt{|Standardizedresiduals|}$ vs log(DealerCose)).

Part D

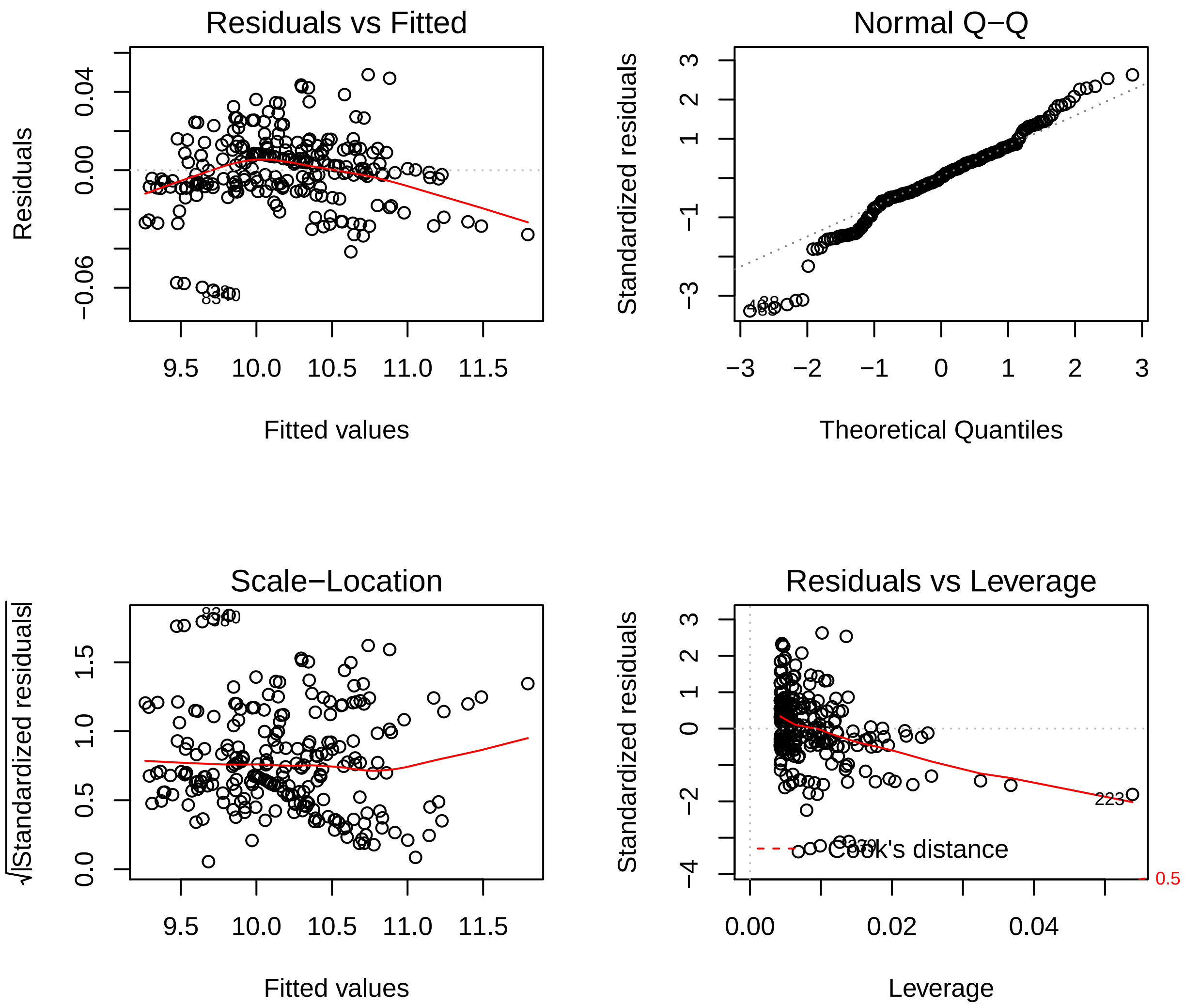
```
summary(m2)
```

```
##
## Call:
## lm(formula = log(SuggestedRetailPrice) ~ log(DealerCost))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.06292 -0.00869  0.00062  0.01062  0.04880
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.06946    0.02646   -2.63   0.0092
## log(DealerCost)  1.01484    0.00262  387.94   <2e-16
##
## Residual standard error: 0.0187 on 232 degrees of freedom
## Multiple R-squared:  0.998,  Adjusted R-squared:  0.998
## F-statistic: 1.5e+05 on 1 and 232 DF,  p-value: <2e-16
```

Approximately, a 1% increase in the DealerCost corresponds to 1.015% increase in SuggestedRetailPrice.

Part E

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



There appears to be some slight nonlinearity in the Residual vs Fitted plot, but not enough to consider the model completely invalid. Perhaps using Box-Cox or inverse-response plots would improve the fit slightly. Moreover, there is one moderate influencial point that is slightly outlying (case 223). A discussion on the validity of this point is warranted.

Project milestones [20 points]

1. Announce your project team of 1 to 3 members.
2. Submit or describe the dataset you will analyze.

- Include a data dictonary: What (and what type of data) are your variables? What are the observational units? How many observations and variables? Which variable will likely by your response (Y) and which variables will be your predictors, X?

3. Pose a preliminary research question(s) and a research hypothesis(es).

References

Sheather, Simon. 2009. *A Modern Approach to Regression with R*. Springer Science & Business Media.