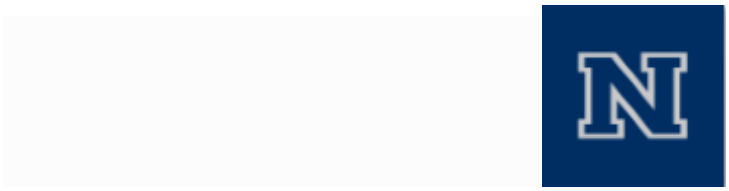
Assignment 3

Biplav Timalsina

March 02, 2018



STAT 757 Applied Regression Analysis

Instructions [20 points]

Modify this file to provide responses to the Ch.3 Exercises in @sheather2009. 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.

Exercise 3.4.1 [20 points]

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 Fare = b0 + b1Distance + e (3.7) (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 significant 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.

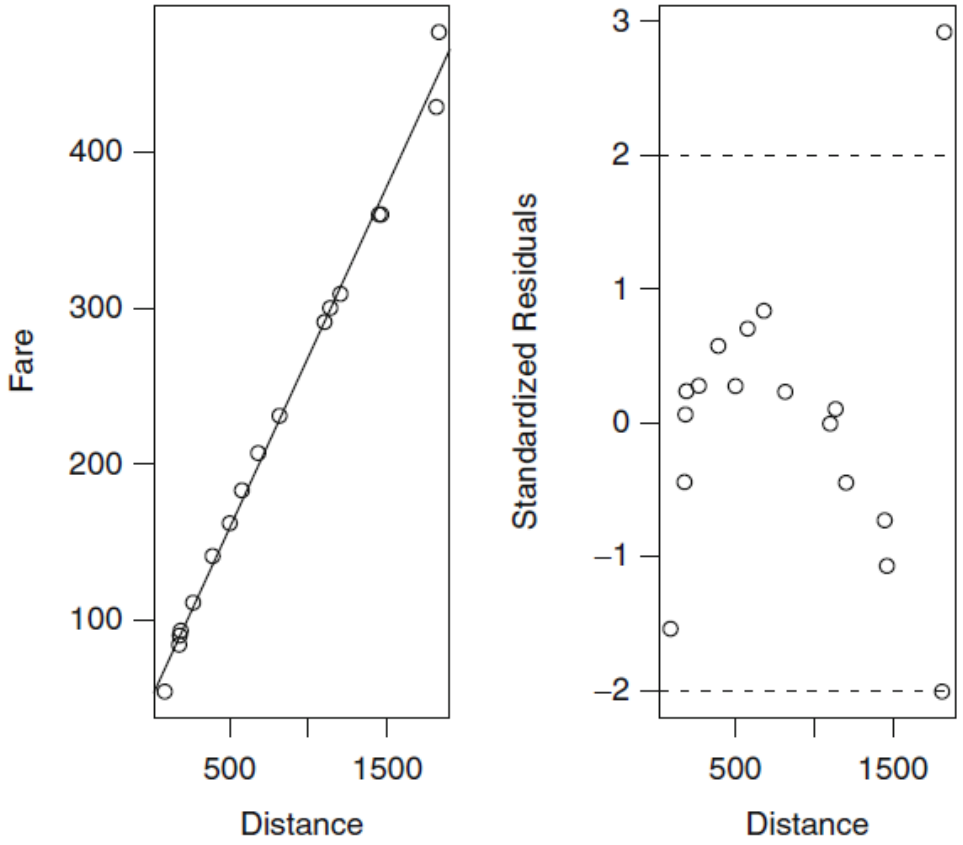


Figure 3.41 Output from model (3.7)

Provide a detailed critique of this conclusion.

(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).

Ans to both a) and b):

We are asked to test the validity of the model presented in Fig. 3.41. At first glance, we can observe that the scatter plot on left seems pretty linear. Hence a linear model can be used to model the relationship between distance and Fare variable considered in this dataset. Also, from the right diagram, we can see that residuals for most of the variables lie within range [-1,+1], which is very good. But, we also see a pattern(quadratic) in the standardized residuals which means that our model might not explain the data properly. While we can model it using a quadratic function to see how accurate it is, we are not sure we can generate better model using that. As mentioned in the question, 99.4% variability is explained by this model, hence this model seems pretty legitimate to predict the Fare with respect to distance.

On further observation, we can also see that for distance much less than 500 (probably <100) and for distance larger than 1500, the prediction seems a little off as seen in the scatter plot and the residuals plot. That means they can be considered at outliers. The point on the right extreme can be considered 'bad' leverage points. Hence, if we consider the data without such points we might be able to generate comparatively accurate improved model for points between (250, 1500) approximately. This might be because there are less number of observations for range outside it. On another note, however, this model is useful to predict fare for range outside it (better than not having any models for those regions).

Another better way to improve the model is to use a quadratic model as the residual plot show quadratic pattern. So, we could use it to see if we can generate a model with constant variance.

Exercise 3.4.4 [20 points]

Tryfos (1998, p. 57) considers a real example involving the management at a Canadian port on the Great Lakes who wish to estimate the relationship between the volume of a ship's cargo and the time required to load and unload this cargo. It is envisaged that this relationship will be used for planning purposes as well as for making comparisons with the productivity of other ports. Records of the tonnage loaded and unloaded as well as the time spent in port by 31 liquid-carrying vessels that used the port over the most recent summer are available. The data are available on the book website in the file glakes.txt. The first model fit to the data was

$$Time = \beta_0 + \beta_1 Tonnage + \epsilon$$

On the following pages is some output from fitting model (3.8) as well as some plots of Tonnage and Time (Figures 3.42 and 3.43).

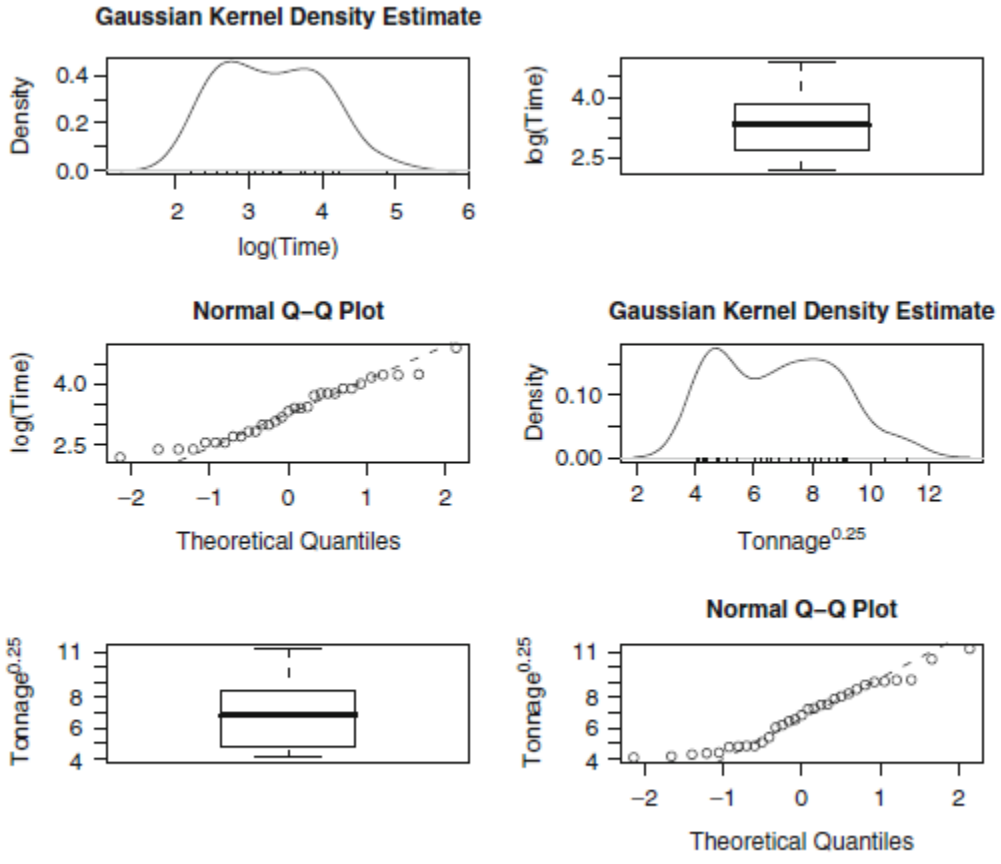


Figure 3.45 Density estimates, box plots and Q–Q plots of log(Time) and Tonnage$^{0.25}$

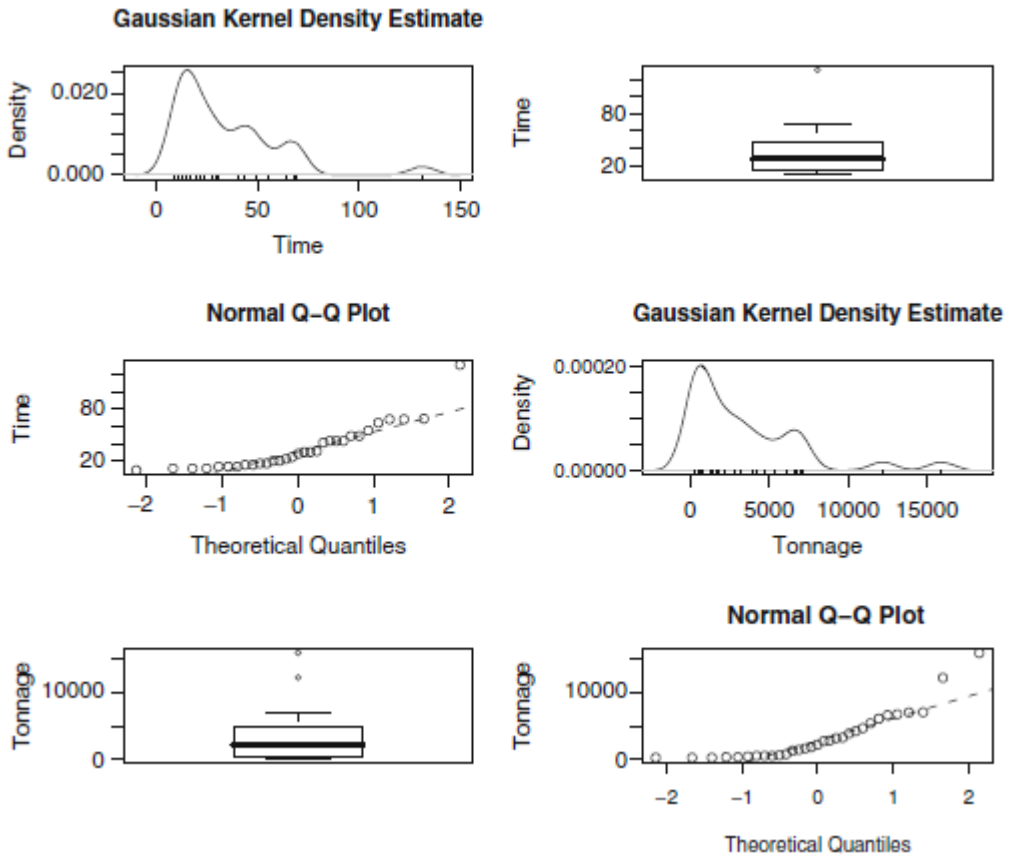


Figure 3.43 Density estimates, box plots and Q–Q plots of Time and Tonnage

Output from model (3.9) as well as some plots (Figures 3.44 and 3.45) appears on the following pages.

(a) Is model (3.9) an improvement over model (3.8) in terms of predicting Time? If so, please describe all the ways in which it is an improvement.
(b) List any weaknesses apparent in model (3.9).

Regression output from R for model (3.8)

```
Call:
lm(formula = Time ~ Tonnage)

Coefficients:
             Estimate Std. Error t value   Pr(>|t|)
(Intercept) 12.344707   2.642633   4.671   6.32e-05 ***
Tonnage      0.006518   0.000531  12.275   5.22e-13 ***
---
Residual standard error: 10.7 on 29 degrees of freedom
Multiple R-Squared: 0.8386, Adjusted R-squared: 0.833
F-statistic: 150.7 on 1 and 29 DF, p-value: 5.218e-13
```

(a) Does the straight line regression model (3.8) seem to fit the data well? If not, list any weaknesses apparent in model (3.8).

Ans:

Looking at fig 3.42 (a), we can observe that a lot of data is found near 0, and data gets sparse for larger tonnage cargo due to which data cannot be closely scrutinized. But from what we see from fig (a), it seems data are more or less well represented in the tonnage range (0,5000), and data gets random for larger tonnnage. It makes sense as the time taken to handle large ton cargo might vary due to various circumstances and dealing with them is hard. As for whether the model is a good fit, we can see in plot c, that the variance is increasing which violates our assumption of constant variance. Hence, this model is not very good to be used for prediction. The bad leverage points may be responsible for such behavior of the model.

(b) Suppose that model (3.8) was used to calculate a prediction interval for Time when Tonnage = 10,000. Would the interval be too short, too long or about right (i.e., valid)? Give a reason to support your answer.

For Tonnage=10000, if we look at fig (a) and (b), we can see that there are only few data observation, and that data observation is way far from the fitted model. From (b), we can see that standard residual is larger around 10000 tonnage cargo. So, we have to say that the interval predicted using this model would be too short or perhaps even valid.

The second model fitted to the data was

$$log(Time) = \beta_0 + \beta_1 Tonnage^{0.25} + \epsilon$$

Output from R

```
box.cox Transformations to Multinormality
         Est.Power    Std.Err.   Wald(Power=0)    Wald(Power=1)
Time        0.0228      0.1930          0.1183         -5.0631
Tonnage     0.2378      0.1237          1.9231         -6.1629
                                         LRT     df      p.value
LR test, all lambda equal     0    3.759605      2 1.526202e-01
LR test, all lambda equal     1   45.315290      2 1.445140e-10
```

Output from R for model (3.9)

```
Call:
lm(formula = log(Time) ~ I(Tonnage^0.25))

Coefficients:
                Estimate   Std. Error  t value  Pr(>|t|)
(Intercept)      1.18842      0.19468    6.105  1.20e-06 ***
I(Tonnage^0.25)  0.30910      0.02728   11.332  3.60e-12 ***
---
Signif. codes: 0'***' 0.001 '**' 0.01 '*' 0.05'.' 0.1' '1

Residual standard error: 0.3034 on 29 degrees of freedom
Multiple R-Squared: 0.8158, Adjusted R-squared: 0.8094
F-statistic: 128.4 on 1 and 29 DF, p-value: 3.599e-12
```

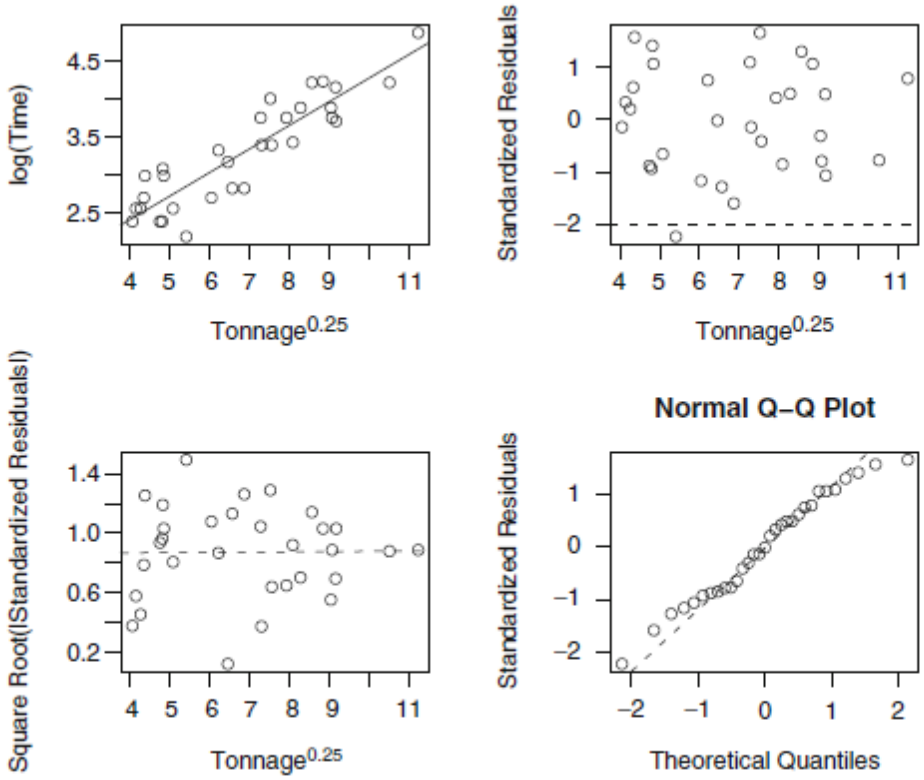


Figure 3.44 Output from model (3.9)

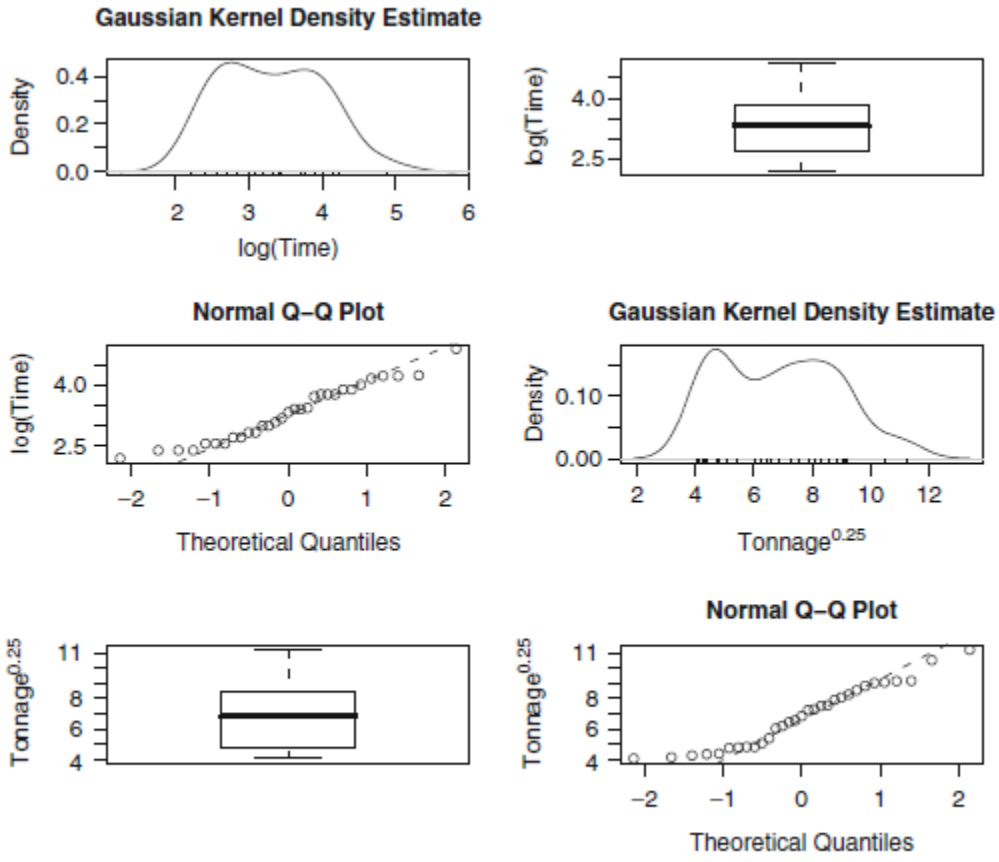


Figure 3.45 Density estimates, box plots and Q–Q plots of log(Time) and Tonnage$^{0.25}$

Output from model (3.9) as well as some plots (Figures 3.44 and 3.45) appears on the following pages.

(c) Is model (3.9) an improvement over model (3.8) in terms of predicting Time? If so, please describe all the ways in which it is an improvement.

Ans:

It is an improvement. It seems that with Log of the observation, the data points are distributed well along the regression line, fig(a), and that standard residuals are more normally distributed along 0 line in fig(b). Also, the data is normally distributed according to fig (d). Improvement 1) We can better predict the time for handling larger tonnage cargo using this model, as there seems to be larger number of data points as $Tonnage^0.25$ increases. Improvement 2) It gets rid of bad leverage point of linear model. Improvement 3) Constant variance is generated in the log model.

(d) List any weaknesses apparent in model (3.9).

Ans:

The residuals' variance decreases for higher tonnages. The distribution is heavy-tailed towards right, which indicates that the data might be better represented using some other function of x.

Exercise 3.4.5 [20 points]

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 function 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 0 1 Suggested Retail Price = b + b Dealer Cost + e (3.10) On the following pages is some output from fitting model (3.10) as well as some plots (Figure 3.46).

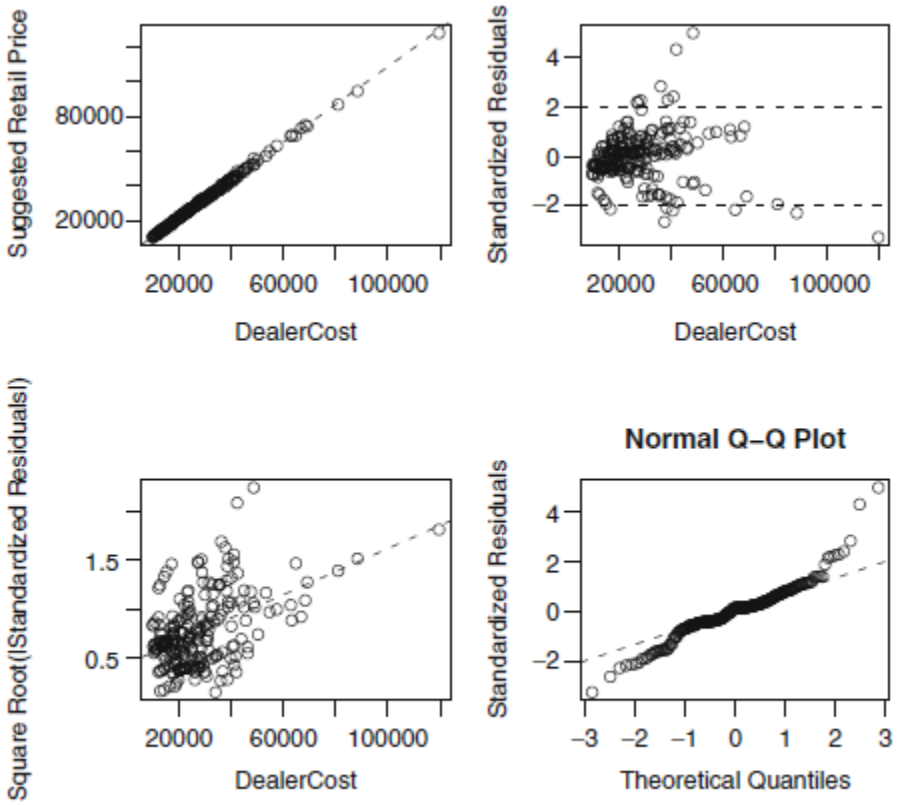


Figure 3.46 Output from model (3.10)

```
Coefficients:
               Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  -61.904248    81.801381    -0.757       0.45
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
```

Output from R for model (3.11)

```
Call:
lm(formula = log(SuggestedRetailPrice) ~ log(DealerCost))

Coefficients:
                Estimate   Std.Error   t value    Pr(>|t|)
(Intercept)    -0.069459    0.026459    -2.625     0.00924 **
log(DealerCost) 1.014836    0.002616   387.942      <2e-16 ***
---
Signif.codes:0'***'0.001 '**'0.01 '*'0.05'.' 0.1 ' '1

Residual standarderror: 0.01865 on232 degreesof 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
```

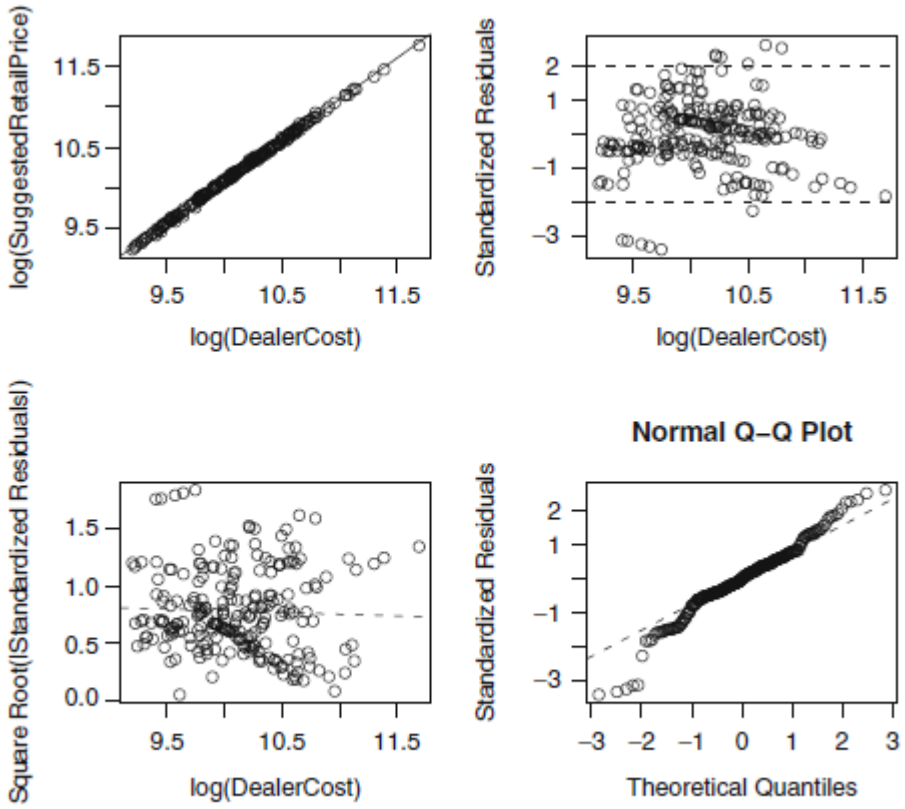


Figure 3.47 Output from model (3.11)

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.

Ans:

It seems that the model explains 99.8% variability in the suggested Retail Price and t-value is >412. We want t-value to be very very >0 because that would mean that we could reject null hypothesis, and thus make sure that there is association between suggested retail price and dealer cost. And, being able to explain 99.8% variability means to predict suggested retail price, we need to include dealer cost as input variable. Although the numbers seems nice, since the interpretation is solely based on general hypothesis test and coefficients which might be misleading and says little about the assumptions of regression, the conclusion may be wrong.

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

Ans:

Problem 1: We can see that linear model has clear evidence of increasing trend for Square root (Standardized residuals) and that the log model doesn't have it. Therefore, the linear model violates our assumption that the variance is constant with X. SOlution 1: We can use weighted least squares to account for the changing variance (with weights equal to 1/variance) or try to transform the data so that the non-constant variance problem disapears.

Problem 2: We can observe from the plot 2 between Standardized residuals and Dealer Cost, that there are some data points that has standardized residuals outside the range (-2,2); and thus they can be classified as outliers. There is one specific point that lies in the far left bottom, which is a bad leverage point, since it is far from usual data and is also an outlier. There are yet some other outliers near it, which could be considered bad leverage points. Solution: We have to analyse why these points are so far away from the normal points, and we might encounter they may represent some special cases of Dealer cost and suggested retail Price. Thus, we might need to remove them from our analysis as a whole, and treat such special cases seperately.

*** The second model fitted to the data was* $ log(Suggested Retail Price) = b + b log(Dealer Cost) + e (3.11)$***
*## Output from model (3.11) and plots (Figure 3.47) appear on the following pages.***(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.* #### Ans: To derive any meaningful conclusions, we need to carefully compare both models and the plots. If we compare linear and log models, we can observe significant difference in two plots, namely Standardized Residuals VS DealerCost and Square Root(S.R) VS DealerCost.**

Standardized Residuals VS DealerCost: We can see that linear model has large number of values concentrated at one area, and it has some bad leverage points and some outliers. But in log model, it gets rid of bad leverage points but still has some outliers. This is an improvement as we hate bad leverage points more than outliers, and bad leverage points significantly change our model. Log model improves by solving the problem of bad leverage points prevalent in linear model. That means better distributed errors.

Square Root(standardized Residuals) Vs DealerCost: We can see while linear model has increasing residuals with x, log model has constant variance over x. This is a huge thing as linear model violates our assumption,

and is not suitable to make promising predictions from. But, log model satisfies our assumption and is useful for predictions. Thus, log model is an improvement because it satisfies our constant variance assumption.

Pattern in residuals are not as apparent as previous.

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

It means that for every 1% increase in the Dealer Cost, the Suggested Retail Price will yield a 1.014% increase.

(e) List any weaknesses apparent in model (3.11).

It seems that it still has some outliers. So, if we could remove such outliers and generate a new model, we might generate better results. Also, it looks like there is a pattern in the third plot with negative slope. We might have to further study why it is dense in that region.

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).

Done With Dhurba Neupane! He has attached the Project Milestone part. ## References