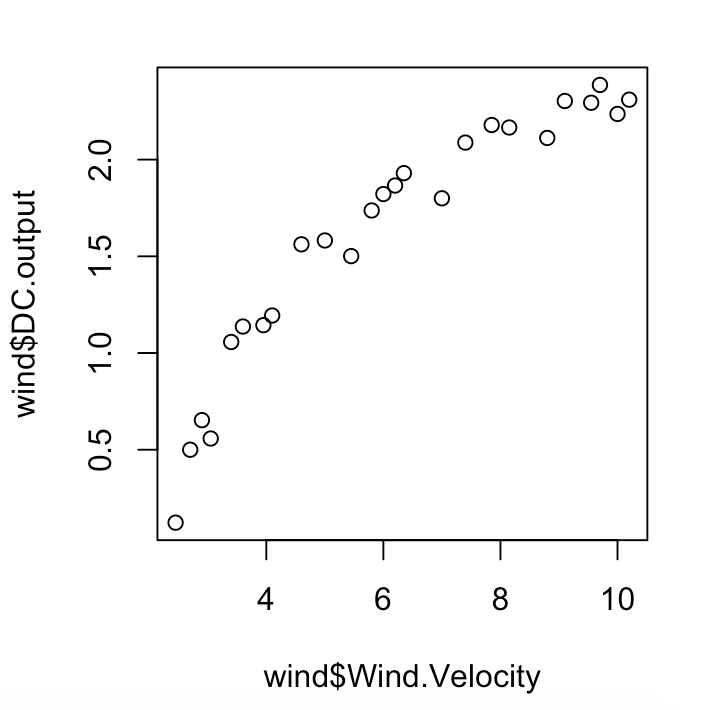
BSAN 450 Assignment 2

1) Direct current output from a windmill generator is measured against the wind velocity (miles per hour). Data from 25 observations are available. The goal is to find a regression model that can be used to predict the direct current output from a given wind velocity. The data for this example are in a file named “Windmill.csv”. The variable names are DC.output = the direct current output and Wind.Velcity = the wind velocity. After you have set the working directory that this file is stored on your computer, the following command will read the data into R Studio.

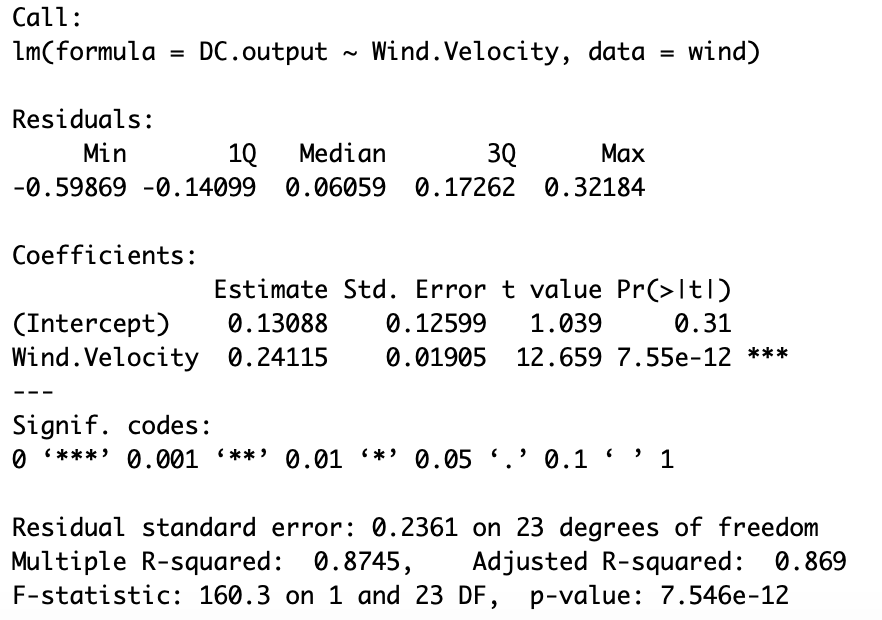
wind=read.csv("Windmill.csv")

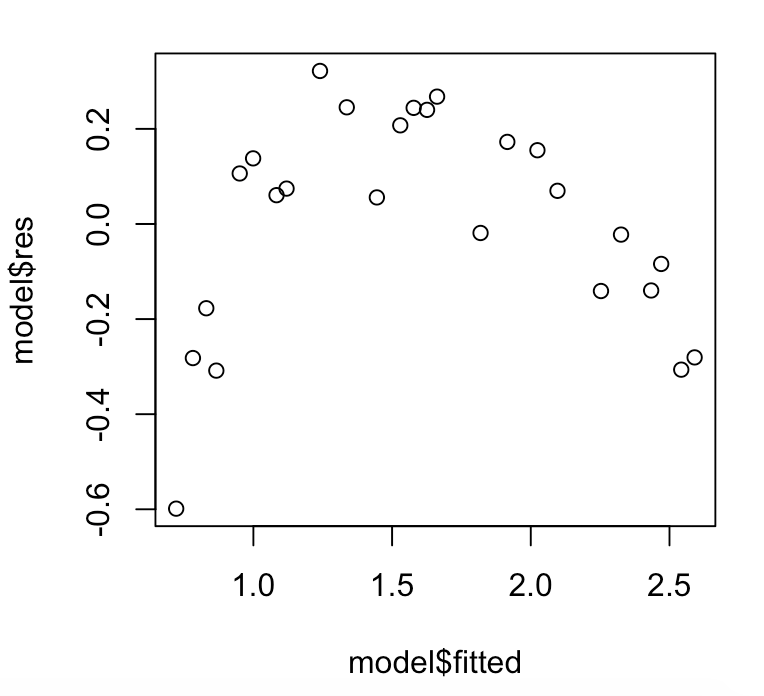
a) Plot a scatter plot of the variable DC.output on the horizontal axis versus the variable Wind.Velocity on the horizontal axis. Comment on this plot. Is there a relationship between these two variables? What is the nature of this relationship?

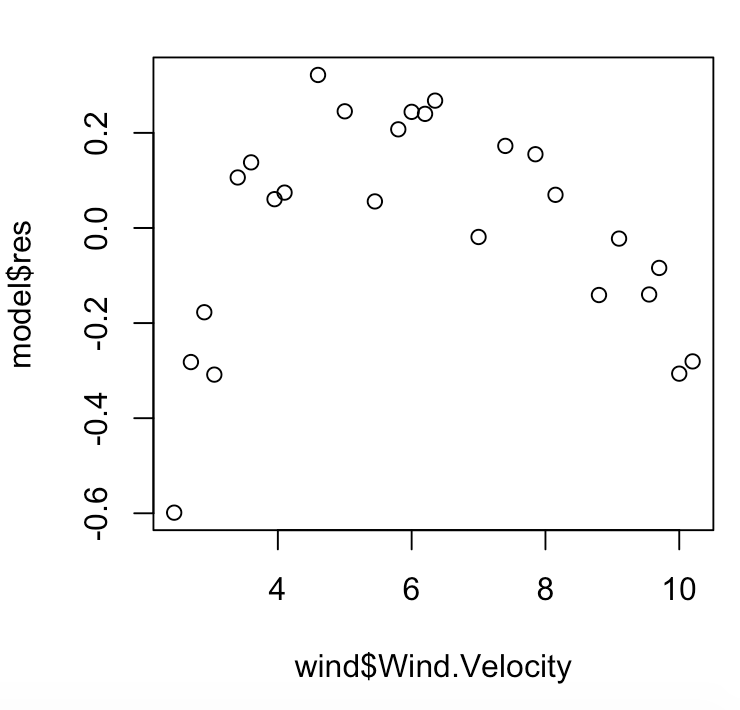


There is a positive relationship between DC.output and Wind.Velocity. The relationship looks potentially linear.

b) Fit a simple linear model with DC.output as the dependent variable and Wind.Velocity as the independent variable. Print out the summary of this model. Plot the residuals versus the fitted values and plot the residuals versus the Wind.Velocity. Comment on these plots.







Both of the residual graphs look problematic. There appears to be a downward trend in the data. The mean of the residuals also is not 0.

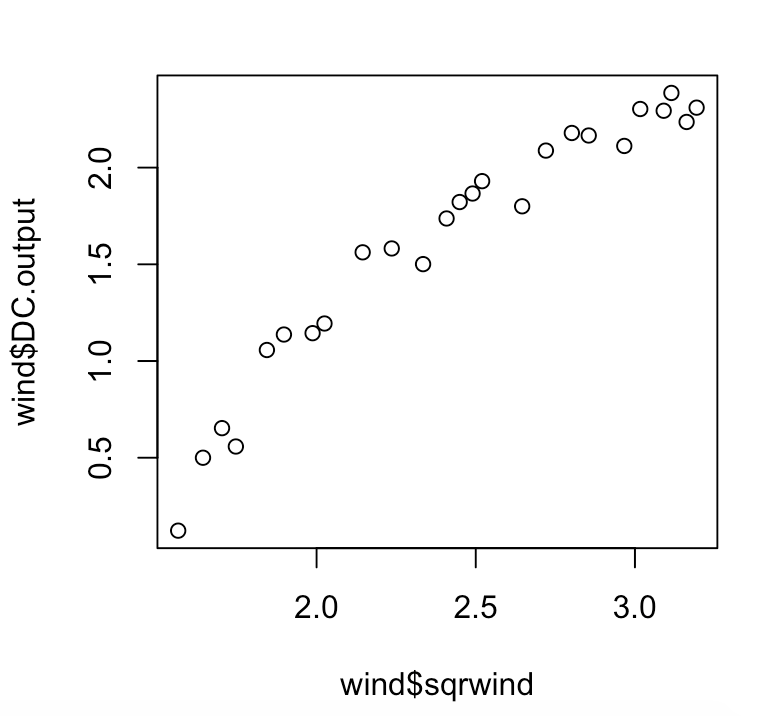
c) For the model you fit in part b) the plot of the residuals versus the fitted values has a curved pattern. This is not consistent with what we would expect if the model assumptions are correct for this data. There is a problem with the model because the mean of the residuals is not zero for all the values of the fitted values. Since there is a curved pattern, a different model needs to be tried.

Consider the plot in part a) of the DC.output verses Wind.Velocity. This plot has a slight curve and the plot is monotonically increasing as the Wind.Velocity increases. In addition, the variation around this slight curve is consistent as the Wind.Velocity increases.

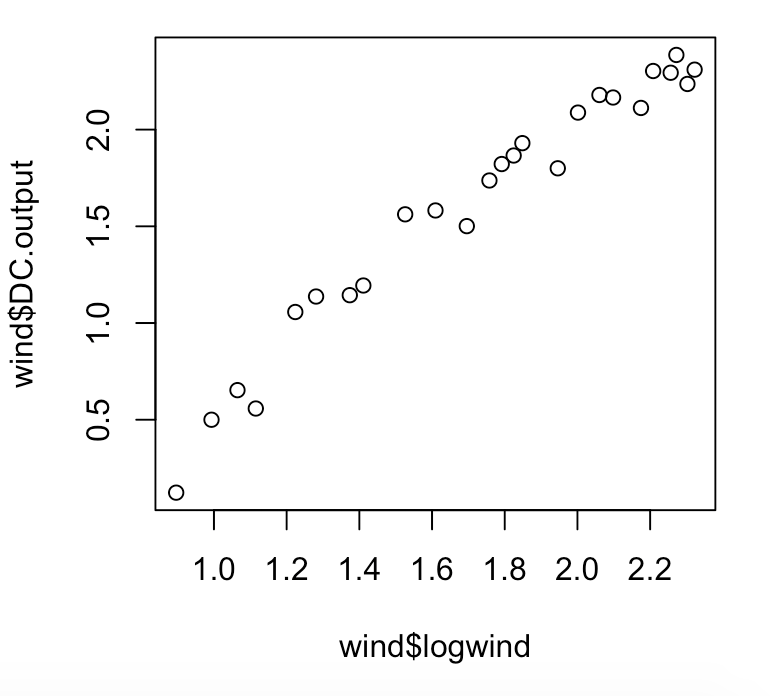
The plot of DC.output versus Wind.Velocity has the following characteristics: (1) There is a curved relationship. (2) The curved relationship is monotonic. (3) The variation around the curve is relatively constant. When these characteristics are present, one possible thing to do is to transform the X variable, in other words transform the Wind.Velocity.

Execute the following R commands (these all assume that you read the data into R using the command above). These commands transform the variable Wind.Velocity into 4 new variables: sqrwind = the square root of Wind.Velocity, logwind = the log of Wind.Velocity, recsq = the reciprocal of the square root of the Wind.Velocity and rec = the reciprocal of the Wind.Velocity.

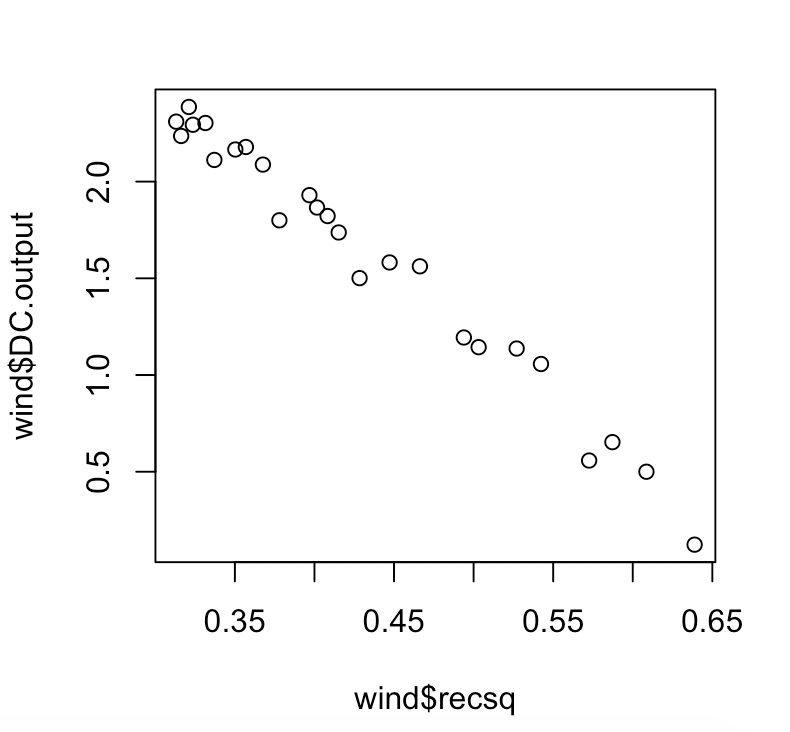
Now produce scatter plots of DC.output versus the 4 new variables that were computed by executing the following R commands.



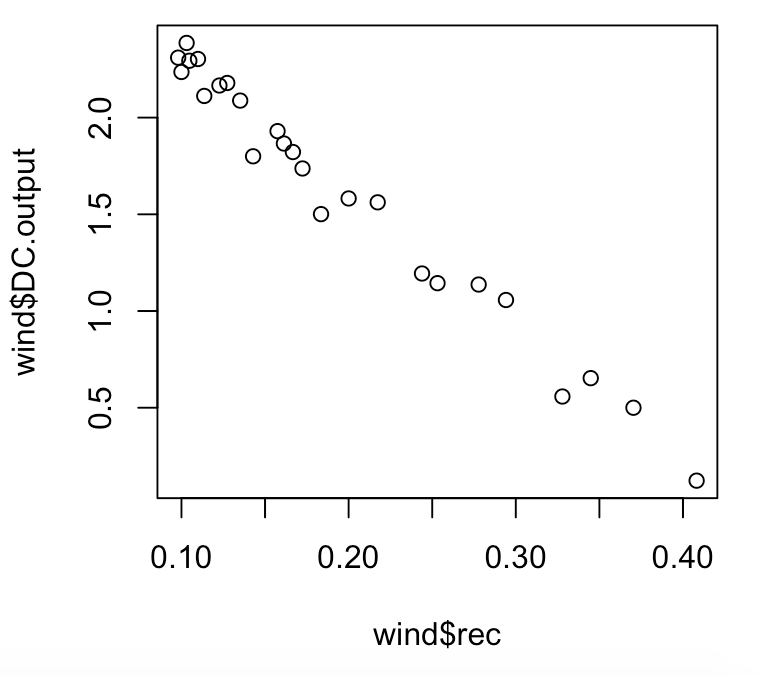
The square root graph appears as if it might be slightly curved still.



The logarithmic graph appears close to a straight line. There is a little cluster at the top of the graph.



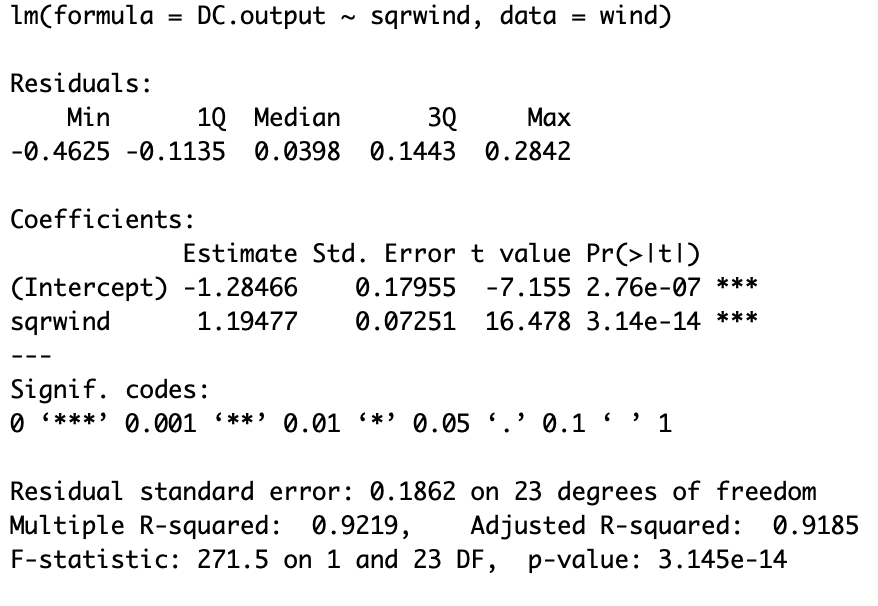
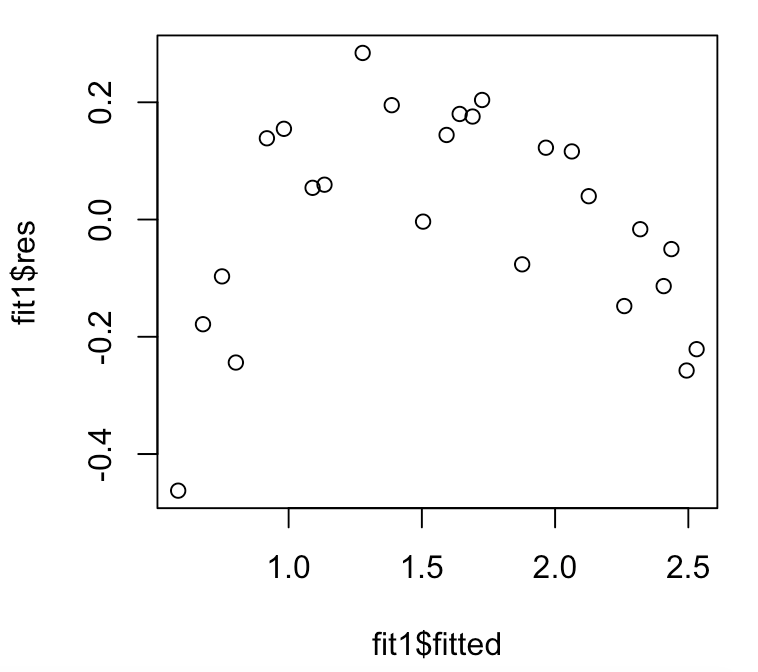
The square root reciprocal graph looks close to a straight line. There is a small cluster at the top of the graph.

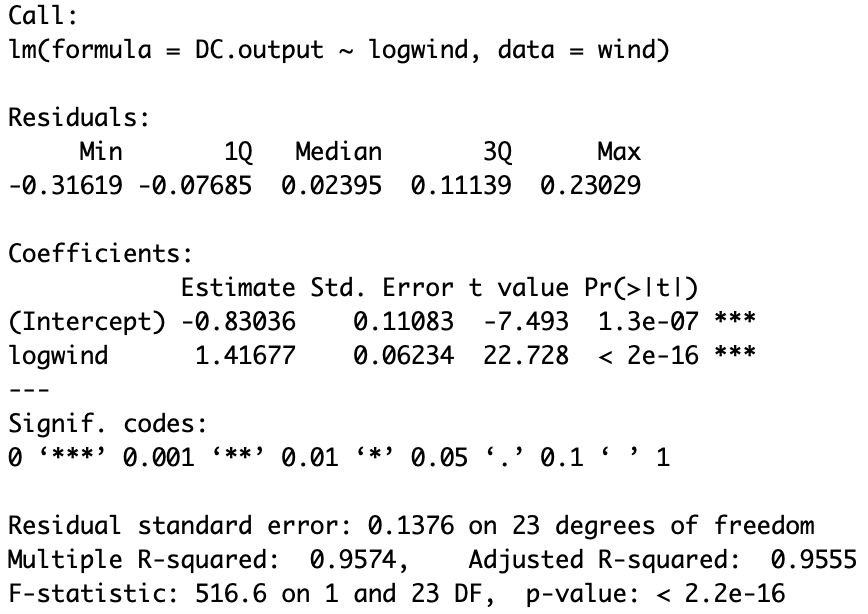
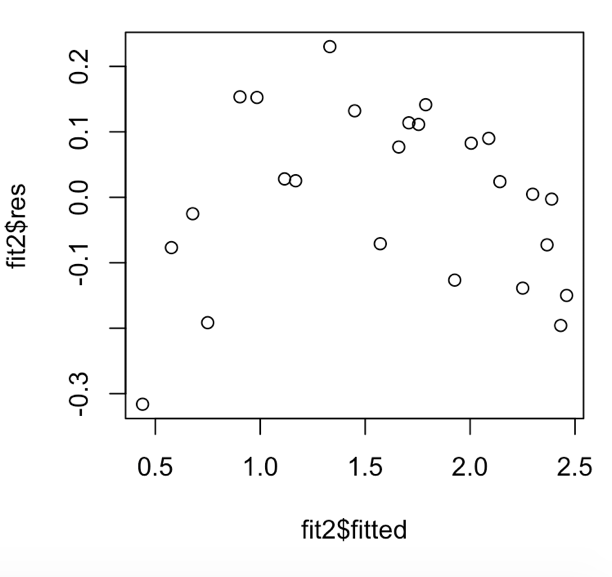


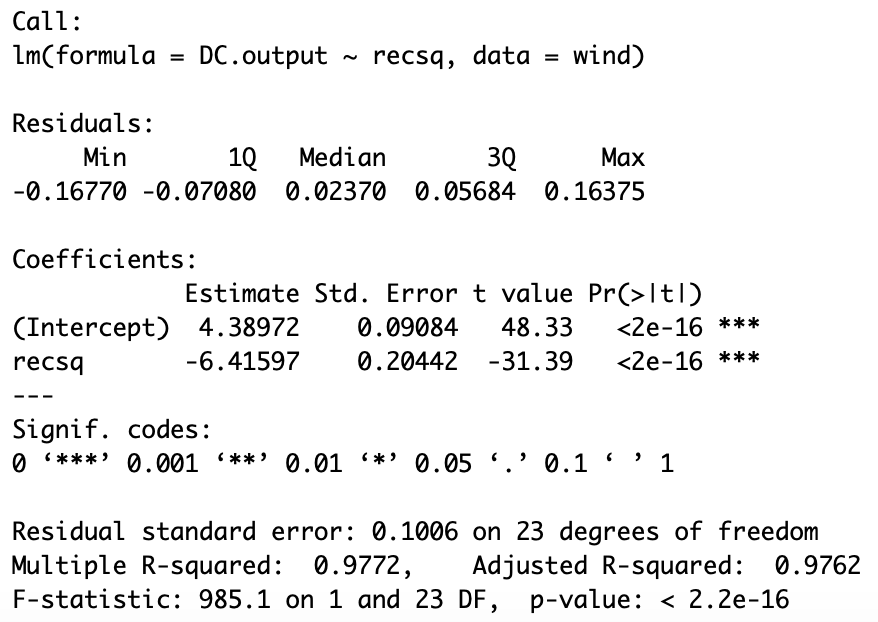
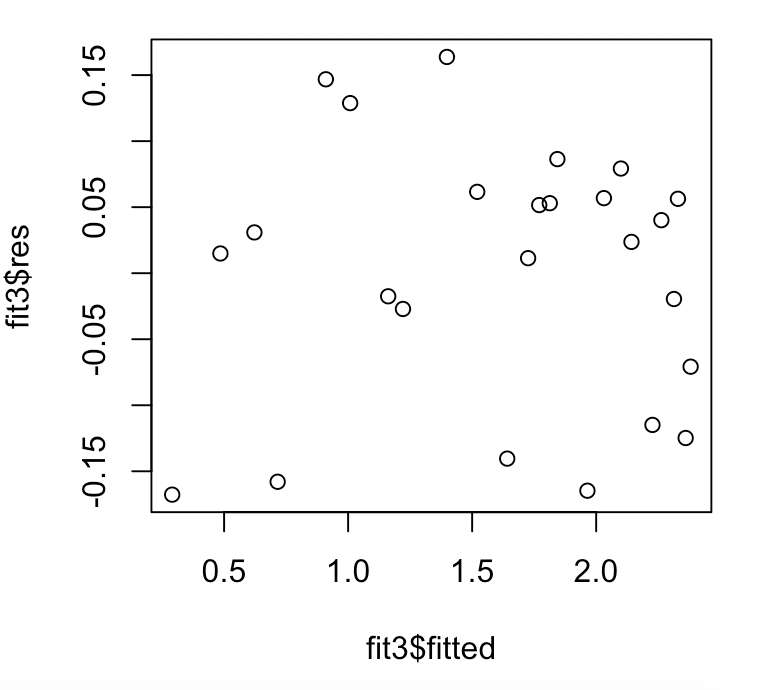
The reciprocal graph is approximately a straight line. There is a little bit of clustering at the top of the graph.

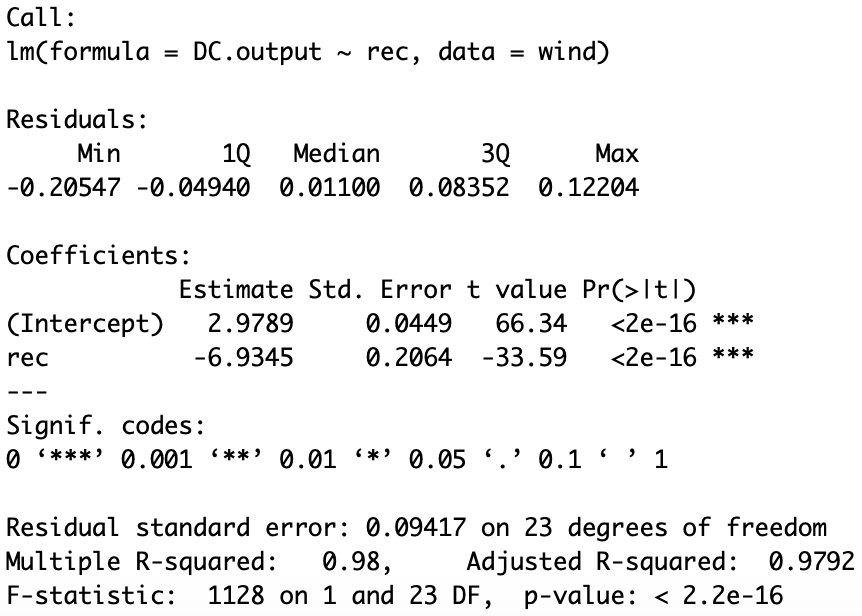
The purpose of these plots is to find a transformation of Wind.Velocity so that the scatter plot of the DC.output and the transformed variable exhibits a straight line. Comment on these plots, which of the transformations seem to show a straight line relationship?

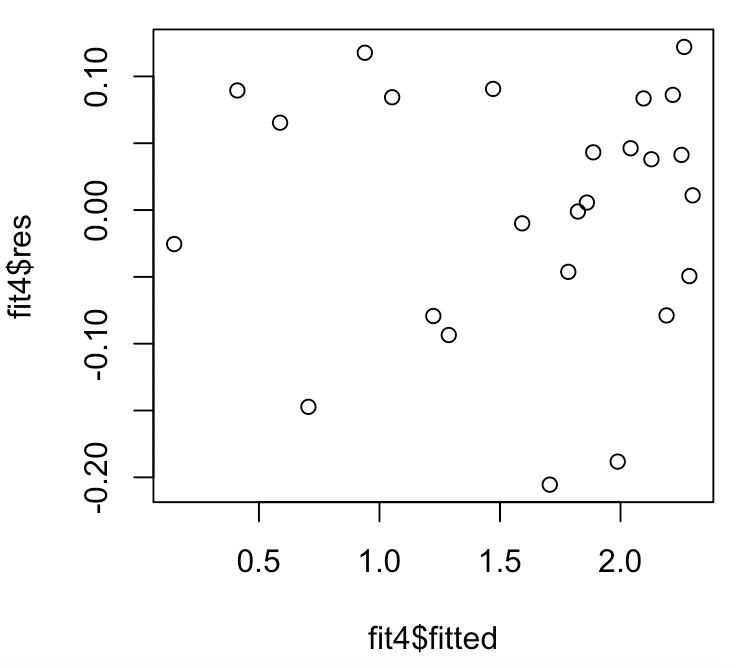
d) It can sometimes be difficult to tell which transformation is best from the scatter plots. Another thing that can be done is to try to fit a simple linear regression model for all of the 4 transformed variables and examine the residual plots. Execute the following R commands and comment on the results, in particular does the plot of the residuals versus the fitted values suggest a problem?











Fit1 isn’t the best model because the residual graph has a downward curve.

Fit2 isn’t the best model because the residual graph has a slight downward curve in the data.

Fit3 isn’t the best fit because the mean of the data is not approximately 0.

Fit4 is the best fit. The data is well spread and the mean is approximately 0.

Since the dependent variable (Y) is the same in all of the regression models that were fit, it is appropriate to compare the values of R-squared and the values of the residual standard error for the different models. Verify that the values in the table below are correct. Note the higher values of R-squared values and lower residual standard error values are preferred.

|  |  |  |
| --- | --- | --- |
| Model | R-squared value | Residual standard error |
| fit – Wind.Velocity | .8745 | .2361 |
| fit1 – sqrwind | .9219 | .1862 |
| fit2 – logwind | .9574 | .1376 |
| fit3 – recsq | .9772 | .1006 |
| fit4 – rec | .9800 | .09417 |

e) Based upon all the analysis of the results for the possible X variables that could be used in a simple linear regression for which DC.output is the dependent variable, which of the possible X variables should be used? Justify your answer.

The reciprocal of the Wind.Velocity is the best independent variable to use because it has the best r-squared value and the lowest residual standard error. In addition, the graph of the residual and fitted values looks the best.

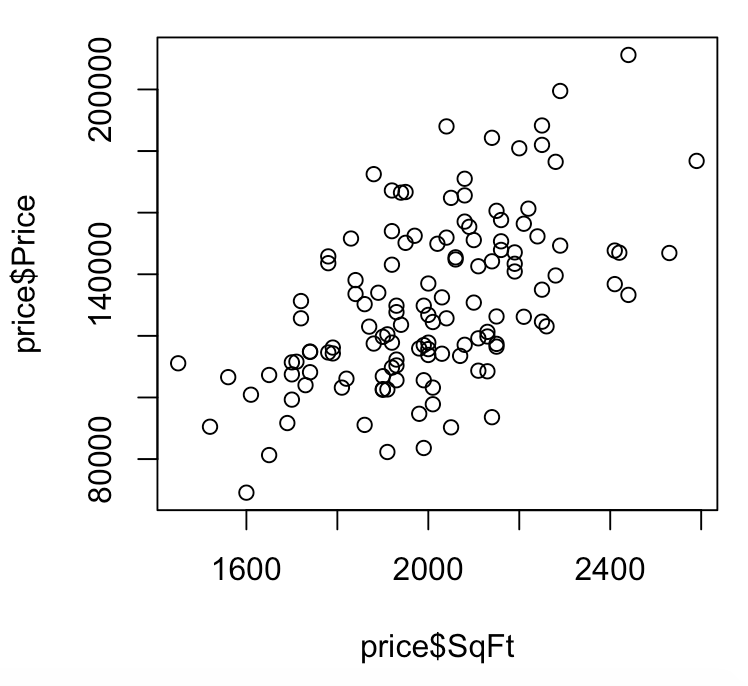
2) House Prices. This data set includes prices and characteristics of 128 houses in a major metropolitan area. The variables include Price (sales price in dollars), SqFt (size in square feet), Bed (number of bedrooms), Bath (number of bathrooms), Offers (number of offers the house has received while on the market), Brick (whether it is brick construction: Yes/No) and Nbrhood(East/North/West). The objective is to explain the sale price of a house as a function of its characteristics. The data is in a file named HousePrices.csv.

For this example note that there are two categorical variables: Brick and Nbrhood. Thus, if the variable Brick is included in a multiple regression model, R will create an indicator variable that takes on the value 1 if the variable Brick = “Yes” and 0 if the variable Brick = “No”.

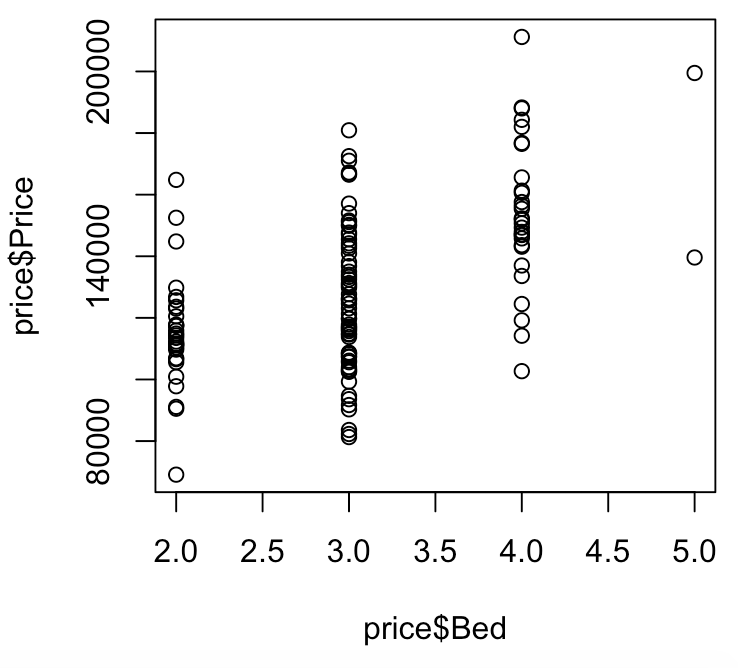
The variable Nbrhood is more complex because there are three groups and as a result there must be 2 indicator variables created for the regression. R creates the variable NbrhoodNorth which is equal to 1 if the house is in the North neighborhood and 0 if the house is not in the North neighborhood. It also creates a second variable NbrhoodWest which is equal to 1 if he house is in the West neighborhood and 0 if the house is not in the West neighborhood. If the house is in the East neighborhood then NbrhoodNorth = 0 and NbrhoodWest = 0. Thus, knowing the values of these two indicator variables is the same as knowing the neighborhood.

a) Read the data into R Studio. Plot the following scatter plots: Price vs. SqFt, Price vs. Bed, Price vs. Bath, and Price vs. Offers. Comment on these plots. Use the following command to read in the data:

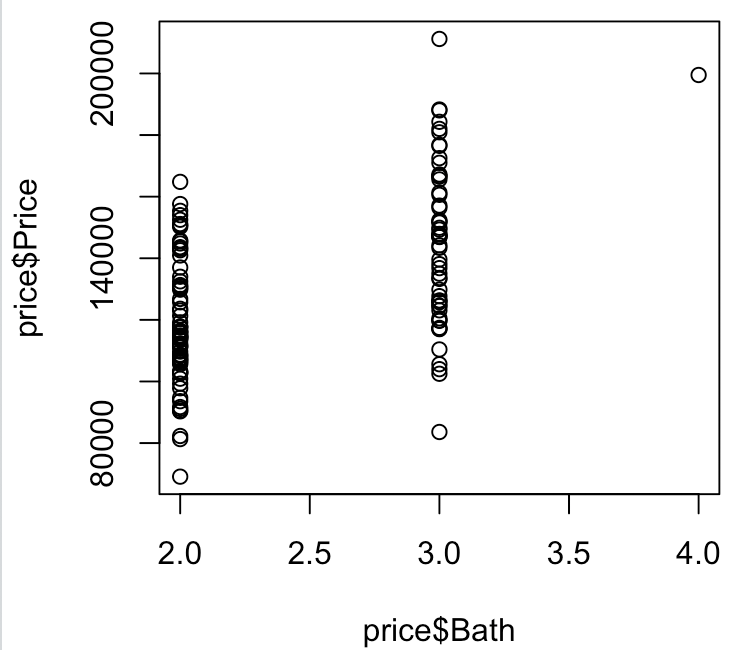
price=read.csv("HousePrices.csv")



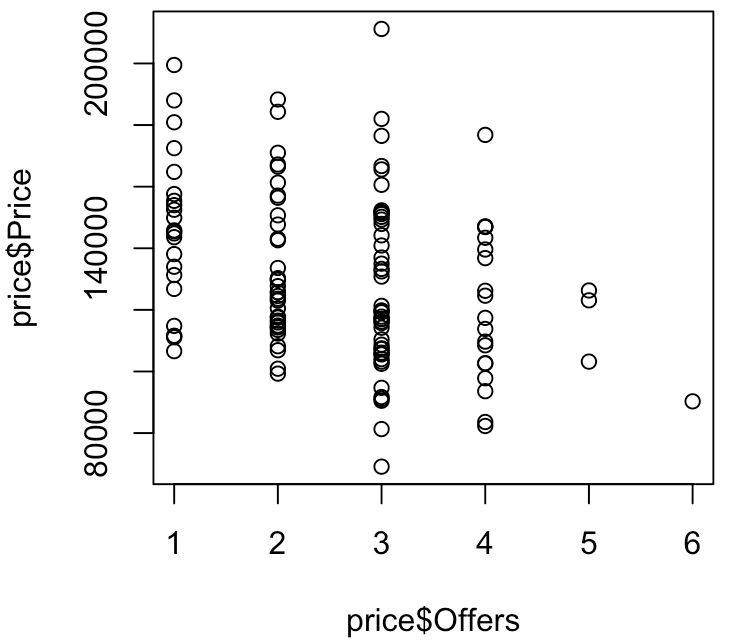
The SqFt graph appears to have a positive trend that looks close to linear.



The Bed graph has a slight positive trend. The trend is difficult to determine with the limited option of x-values.

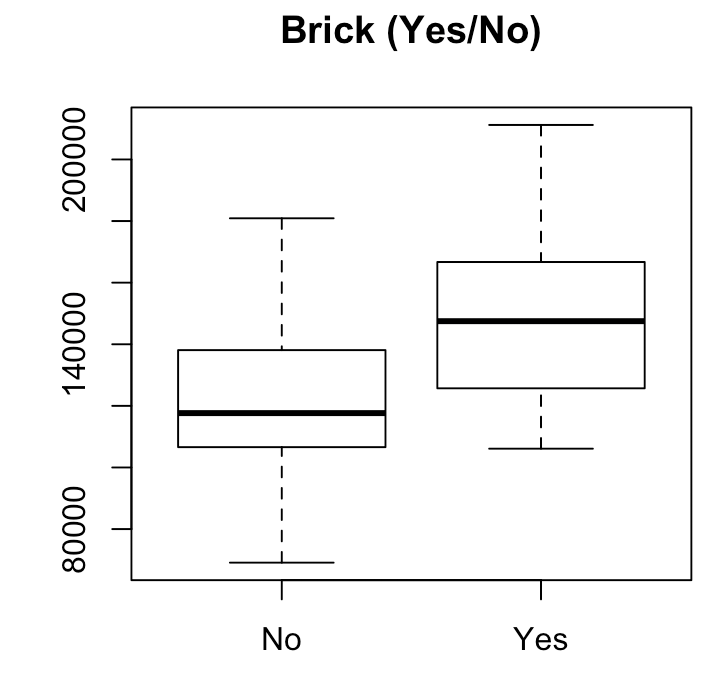


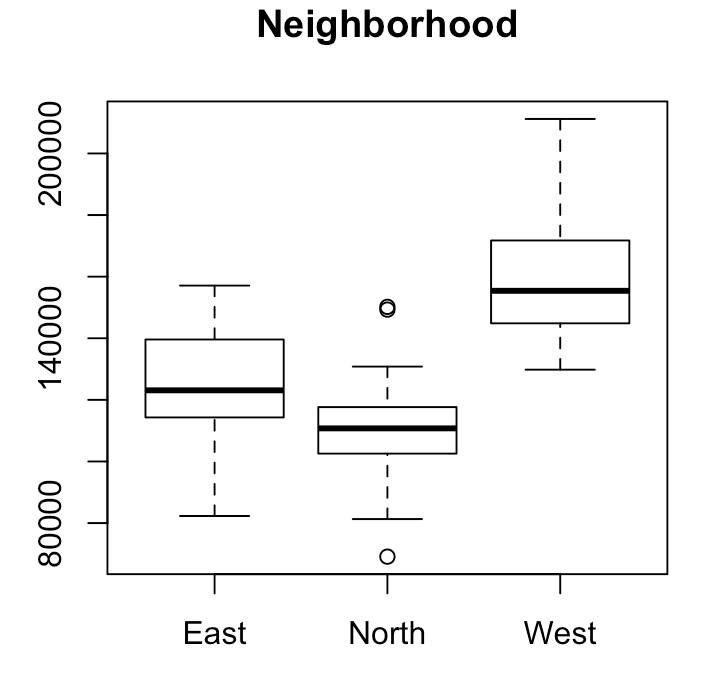
The bath graph appears to have a slight positive trend, but it is not very clear.



The offers graph seems to have a negative trend.

b) Since the variables Brick and Nbrhood are categorical variables, the appropriate plots are boxplots. The commands to plot boxplots in R are the following.





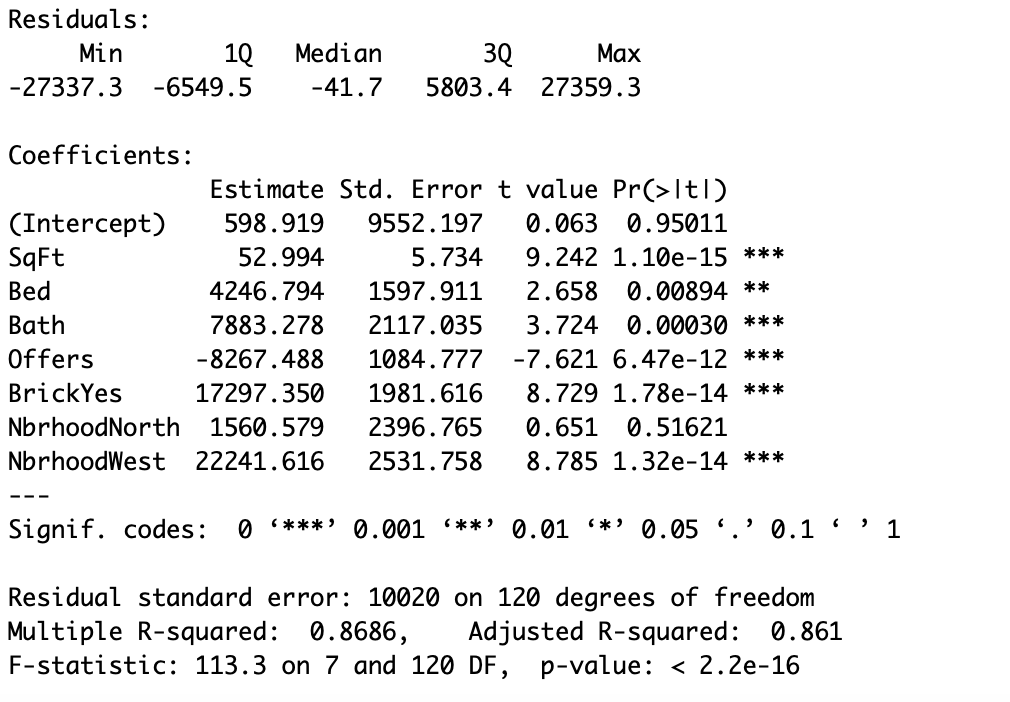
In boxplots, the median of the data is the bold middle line, the lower edge of the box is the 25th percentile (25 percent of the data is below this value), and the upper edge of the box is the 75th percentile (75 percent of the data is below this value). The lines extending vertically from the boxes are called whiskers. If there are no outliers or potential outliers, then the ends of the whiskers are respectively the minimum and the maximum values in the data.

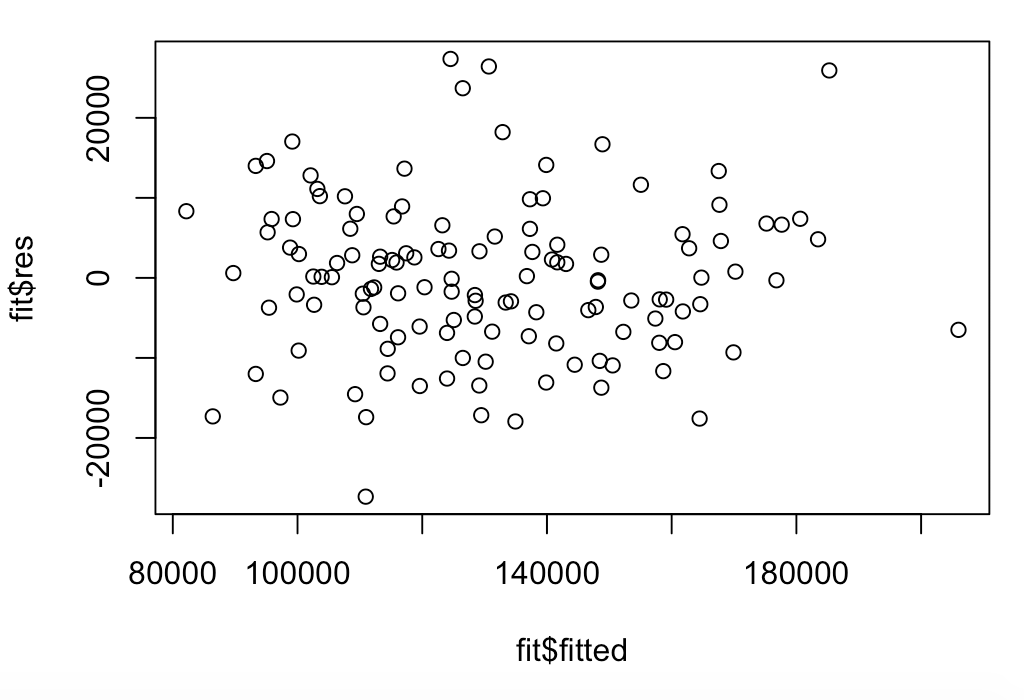
Plot box plots of Price grouped by Brick and Price grouped by Nbrhood.

Based on these plots, which variables appear to have a relationship with Price? Do the model assumptions for the multiple regression model appear to be correct? Do these plots suggest any problems?

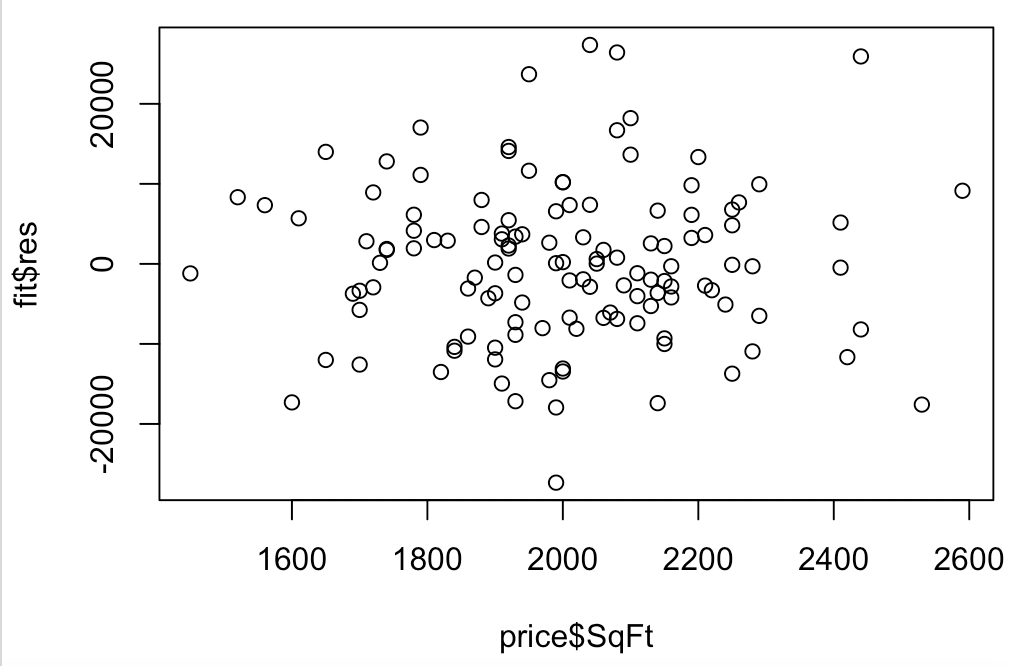
Brick and neighborhood both have an impact on the price. The model assumption for the multiple regression appears to be correct. The relationship between multiple of the variables appear to have an impact on the price. The neighborhood graph indicated that there are a couple of outlier data points.

c) Fit a multiple regression model with Price as the dependent variable and all the other variables as independent variables. Print out a summary of the model and perform the diagnostic checks for this model.

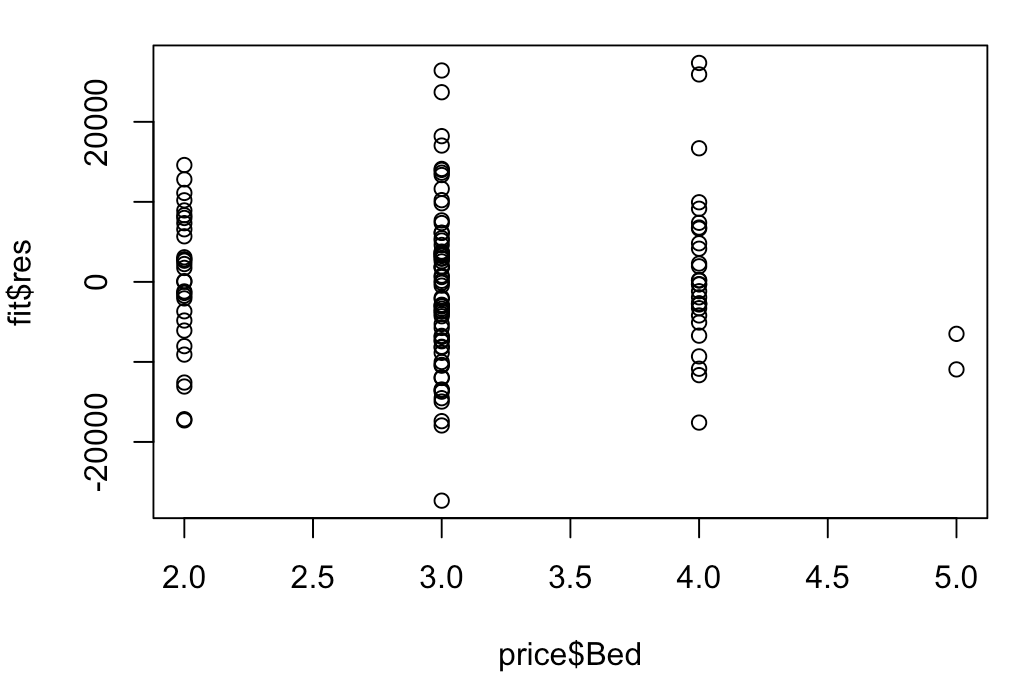




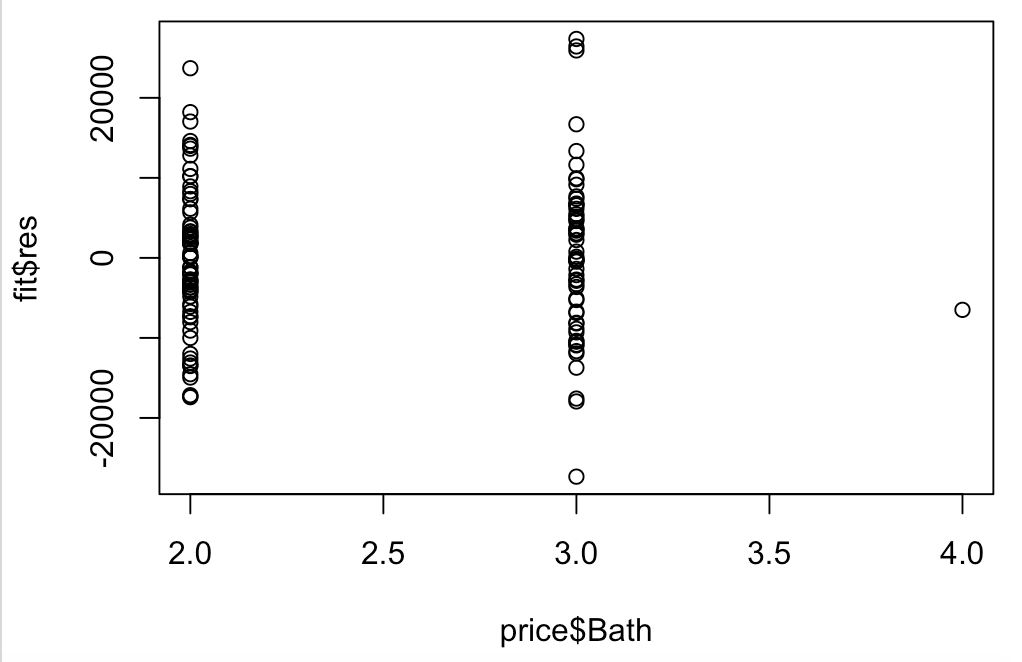
The graph above the data looks well distributed and has a mean approximately 0.



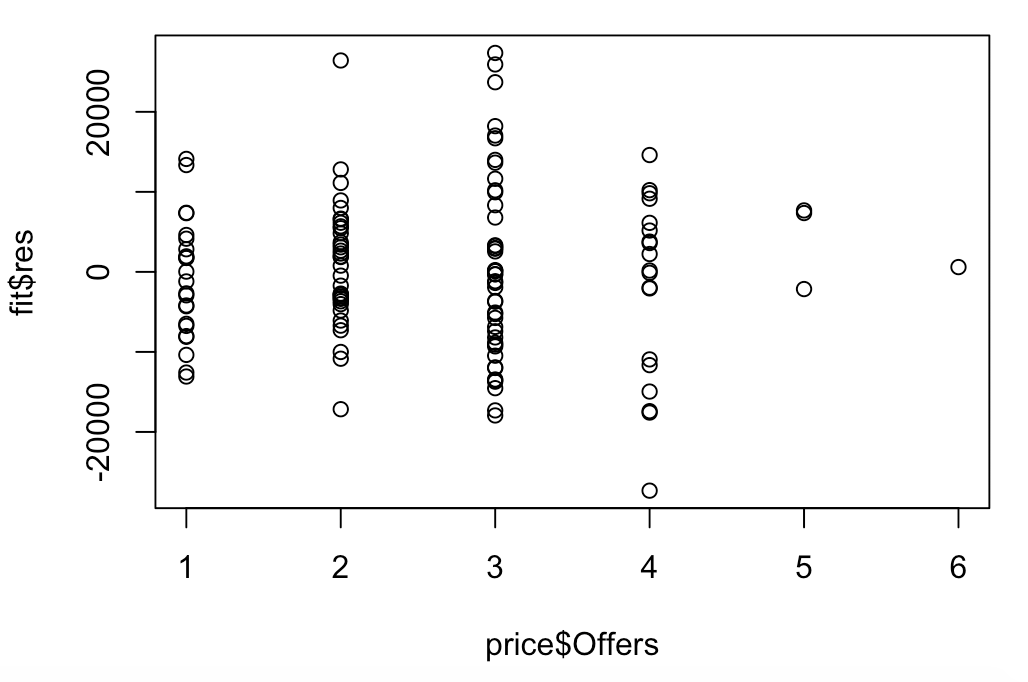
The graph above the data looks well distributed and has a mean approximately 0.



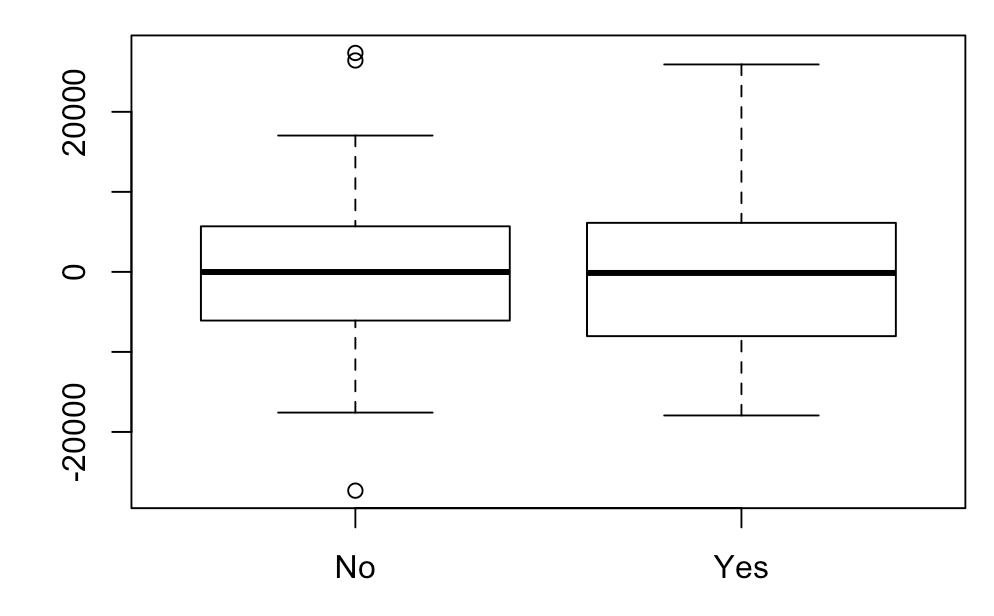
The graph above the data looks well distributed and has a mean approximately 0. There is a lack of data for larger x-values.



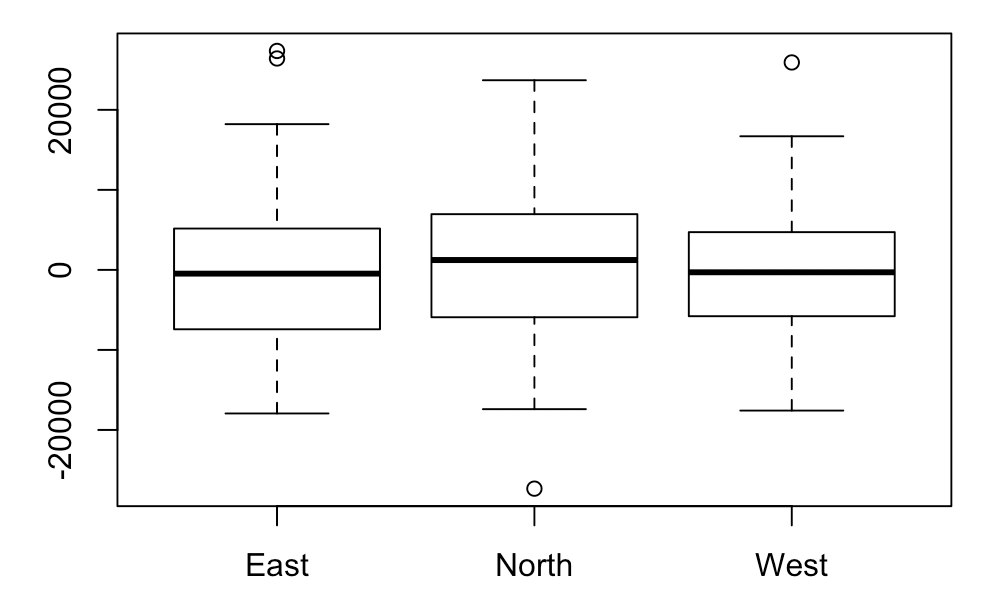
The graph above the data looks well distributed and has a mean approximately 0. There is only two real x-values that are represented in the graph.



The graph above the data looks well distributed and has a mean approximately 0. There is a slight lack of data for larger values of offers.



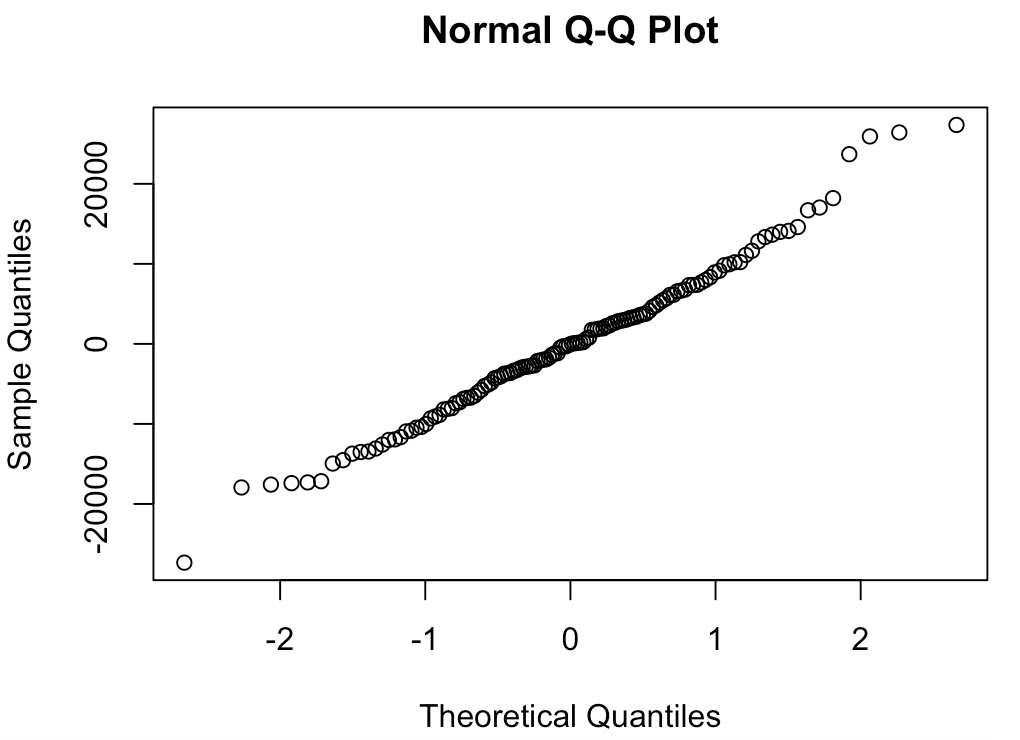
The graph above the data looks relatively well distributed and has a mean right about 0. The data is a little clustered around 0 but stays pretty consistent between values.



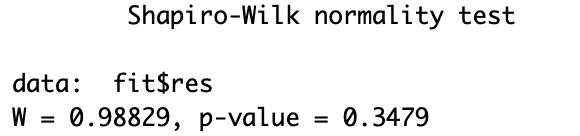
The graph above the data looks relatively well distributed and has a mean right about 0. The data is a little clustered around 0 but stays pretty consistent between values.



This looks approximately like a bell curve.



The graph is mostly a straight line but is out of line at the beginning and end of the line.



The p-value in the Shapiro-Wilk test is larger than .05 so the data follows a normal data pattern.

All of the diagnostics suggest no issues with the model.

Do the diagnostic checks suggest any problems with this model? Justify your answer.

There is some data that might be outliers, but overall this seems like a relatively good fit for this data.

d) For the model you fit in part c), answer the following questions.

Are all the variables needed in the model? Why or why not?

No, not all the variables are needed in the model. The variable neighborhood north has a high p-values of 0.51625 indicating that it is not significantly impacting the price.

What is the difference between the average selling price of house with brick construction compared to the average selling price of a house without brick construction?

The average selling price of a house with brick is over 20,000 more dollars than houses without brick.

What is the difference in the average selling price between houses in the North neighborhood compared to houses in the East neighborhood?

There is a difference of about 10,000 dollars with East being the higher of the two.

What is the difference in the average selling price between houses in the West neighborhood compared to houses in the East neighborhood?

West neighborhoods have an average selling price that is about 40,000 more dollars.

What is the difference in the average selling price between houses in the North neighborhood compared to houses in the West neighborhood?

North neighborhoods have an average selling price that is about 50,000 dollars less than West neighborhoods.

3) An experiment was conducted to understand the behavior of worsted yarn under conditions of repeated loading in production conditions. The experiment was conducted to understand the relationship between Y = cycles to failure and three predictor variables  = the length of the test specimen,  = the amplitude of the loading cycle, and  = the load. The data is in the file “textile.csv”, the variable names are: Cycles = cycles to failure, Length = the length of the test specimen, Amplitude = amplitude of the loading cycle, and Load = the load.

The data is in the file “textile.csv”, the variable names are: Cycles = cycles to failure, Length = the length of the test specimen, Amplitude = amplitude of the loading cycle, and Load = the load.

Recall that this data was analyzed in class. When the model was estimated the output from R was:

Call:

lm(formula = Cycles ~ Length + Amplitude + Load, data = textile)

Residuals:

Min 1Q Median 3Q Max

-644.6 -279.1 -150.1 199.6 1268.0

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4521.333 1621.778 2.788 0.010457 \*

Length 13.200 2.301 5.736 7.66e-06 \*\*\*

Amplitude -535.889 115.061 -4.657 0.000109 \*\*\*

Load -62.156 23.012 -2.701 0.012751 \*

---

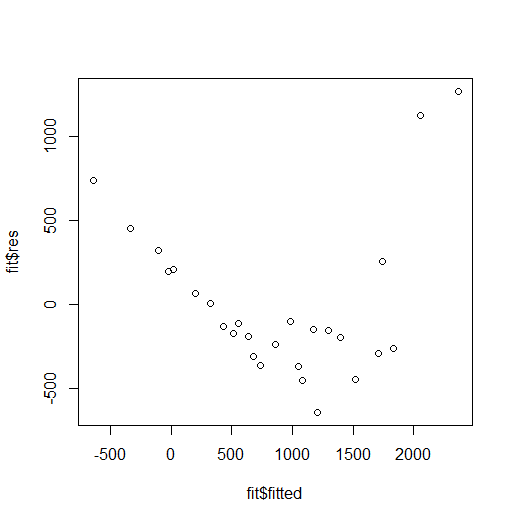
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 488.2 on 23 degrees of freedom

Multiple R-squared: 0.7291, Adjusted R-squared: 0.6937

F-statistic: 20.63 on 3 and 23 DF, p-value: 1.029e-06

The plot of the residuals versus the fitted values for this model was:

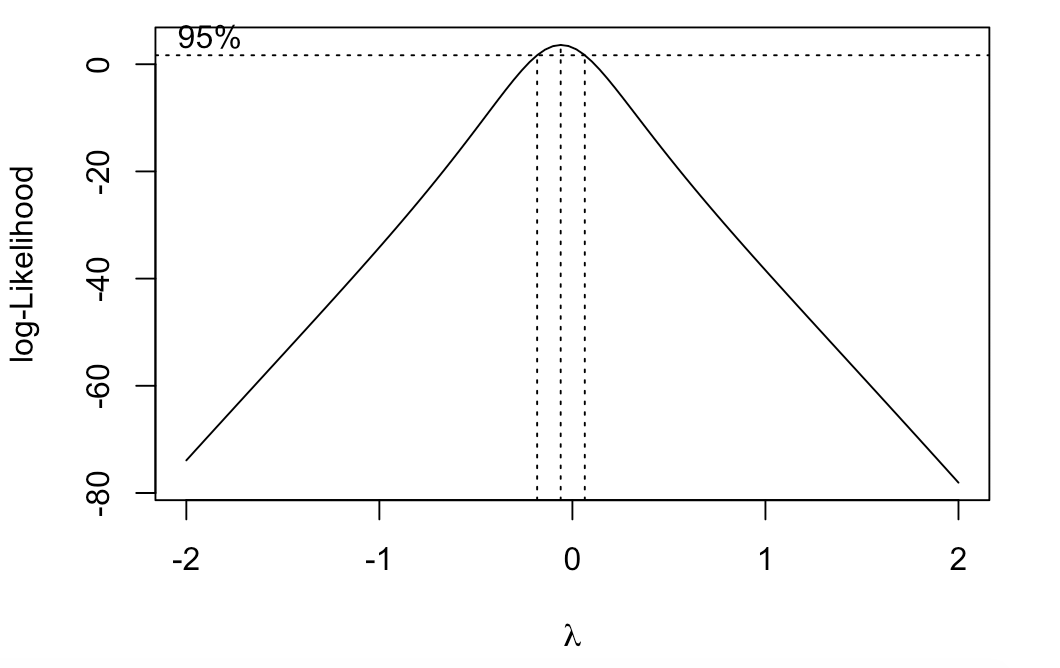


This plot suggests that the assumptions of the multiple regression model are not satisfied, in particular the variance of the errors is not constant rather the variance increases as the fitted values increase. This pattern in the plot of the residuals versus fitted suggests that an appropriate change is a nonlinear transformation of the Y variable.

a) The following R commands provide information about the appropriate transformation. If the package MASS has not been installed on your computer, you will need to install it before running the library command. This assumes that you read the data into a dataframe named textile in R. Execute these commands. What transformation of the Y variable does this output suggest?

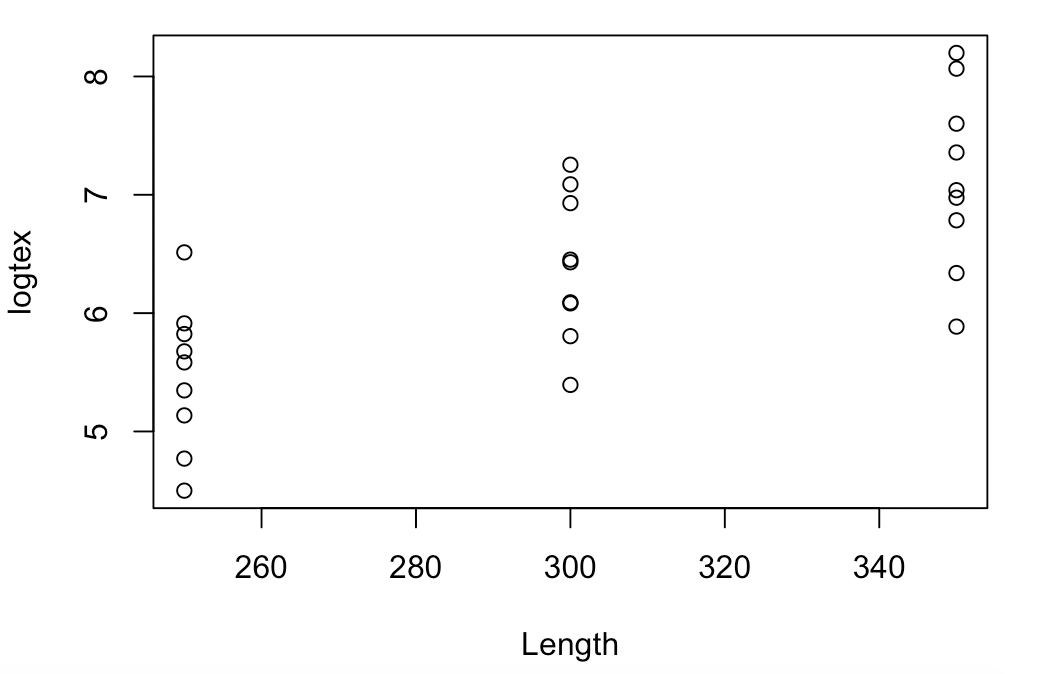
library(MASS)

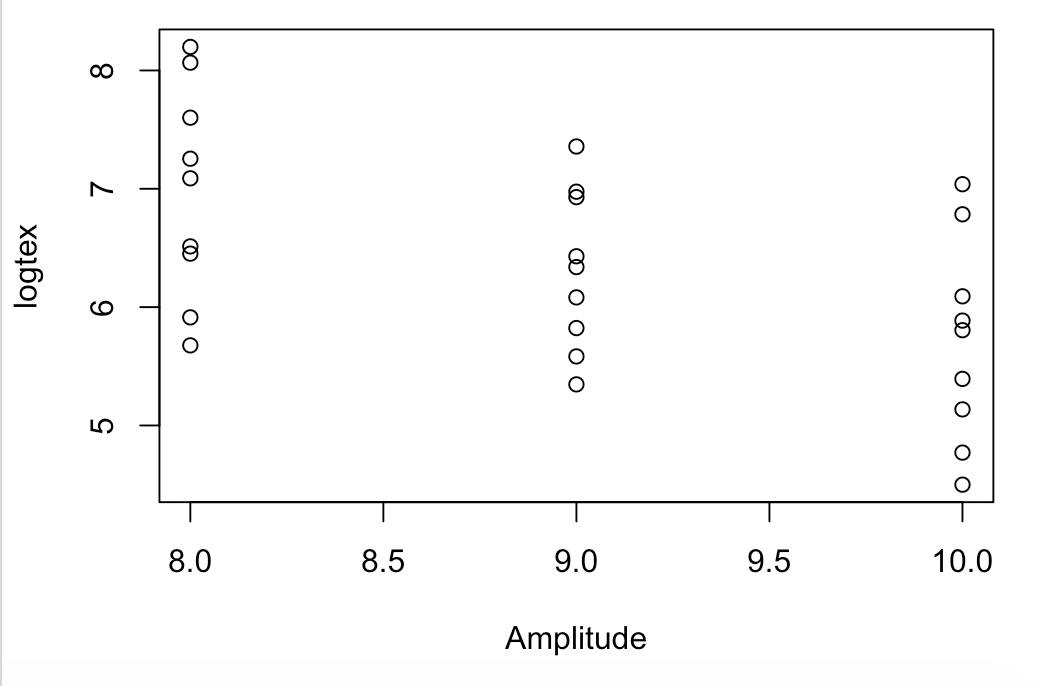
boxcox(Cycles~Length+Amplitude+Load,data=textile)

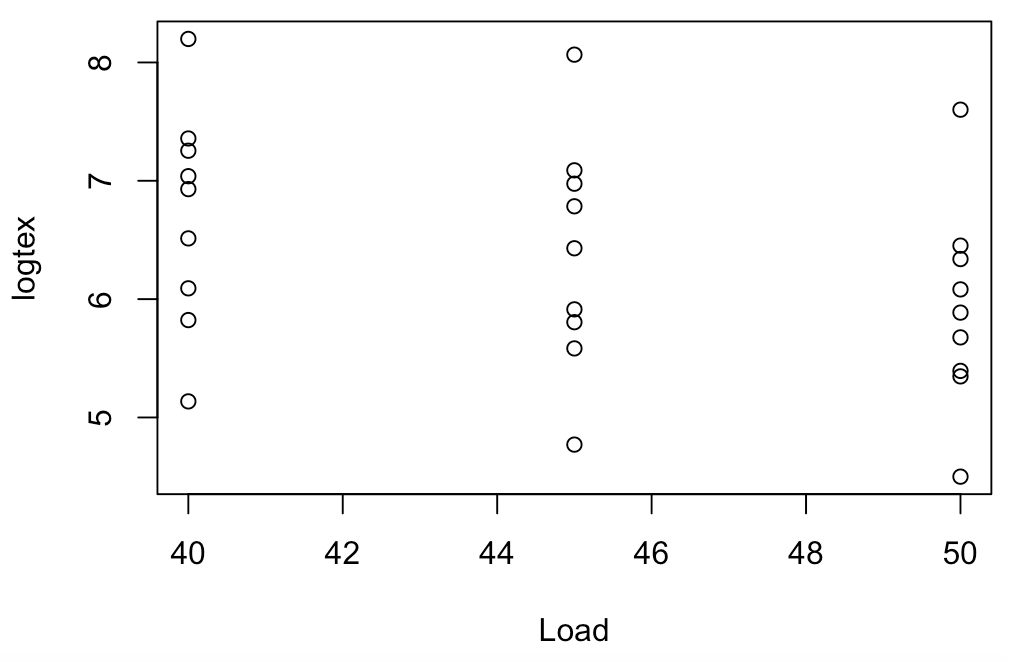


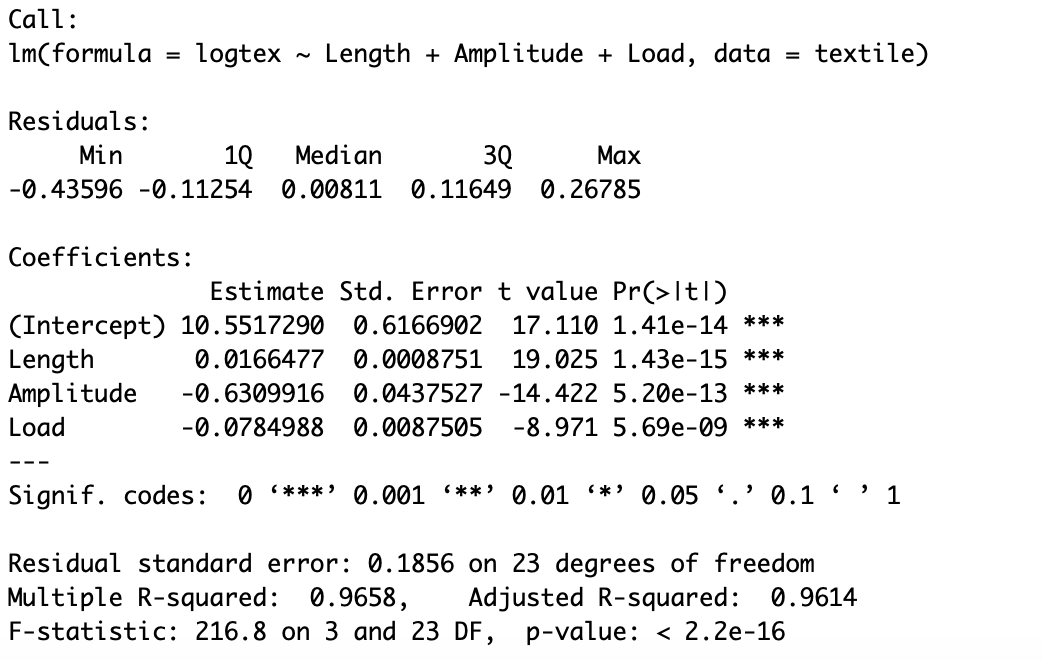
The shape of the graph is comparable to that of a negative parabola. Since the graph reaches 95% around lambda = 0 the best Y transformation is then logarithmic.

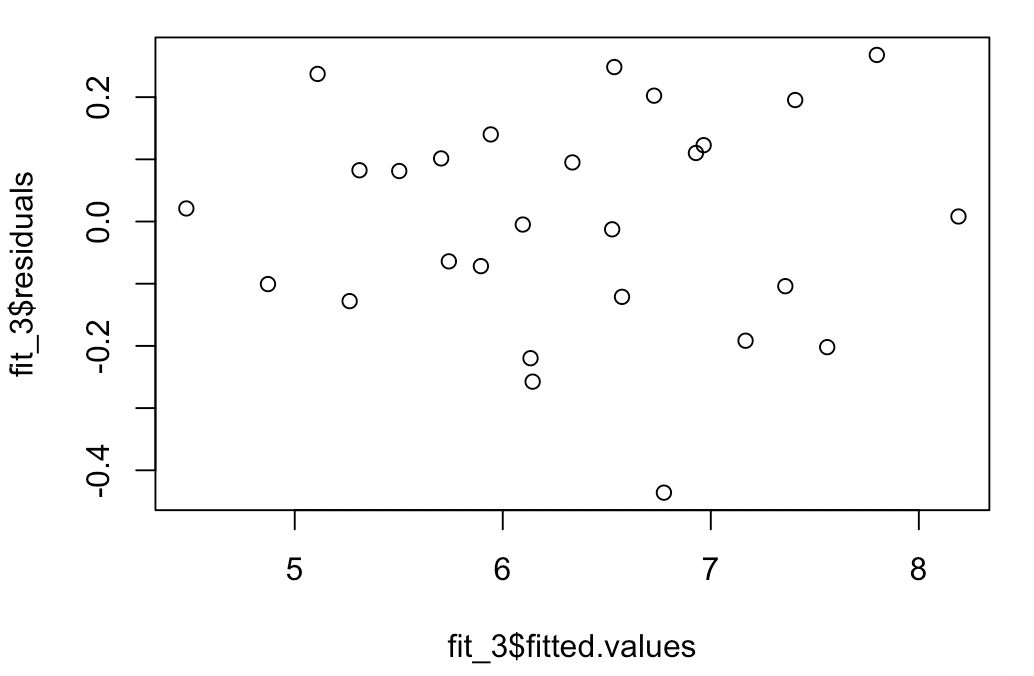
b) Based upon your answer in part a) transform the Y data and redo the analysis of this data. This would include scatter plots of the transformed Y versus the three X variables, estimation of a model and doing the diagnostic checks for the new model. Comment on this new model compared to the model with Cycles as the dependent variable.

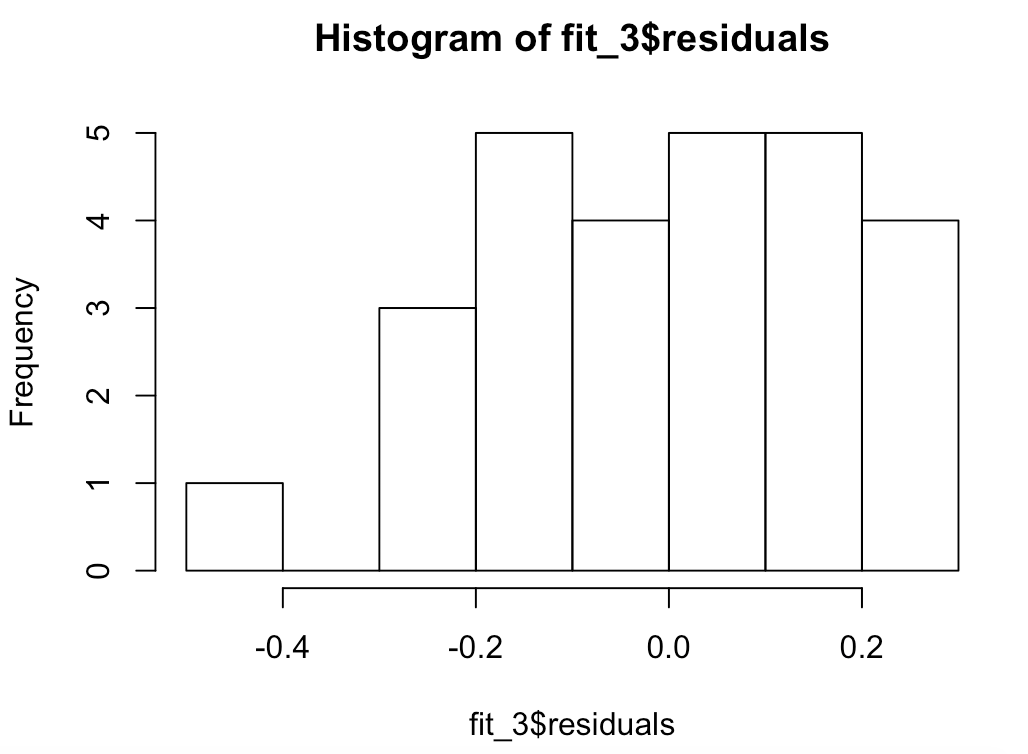


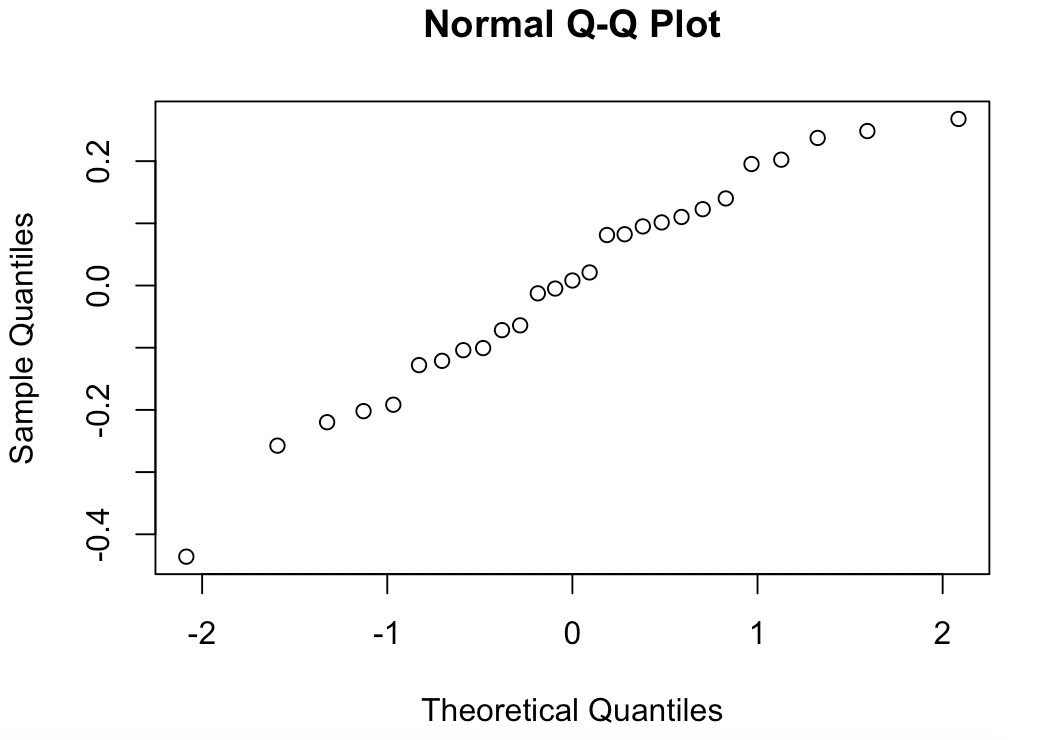


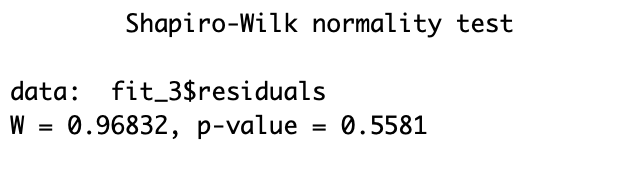












The new model is a much better fit than the old. The new r squared values is .96 whereas the old one was .72 All the diagnostics look good except for that the histogram of the residual is not that of a standard bell curve, which could be explained in the limited variety of values within the data. The diagnostic of the residual v. fitted values looks much better. The data is well distributed and suggests no issue.