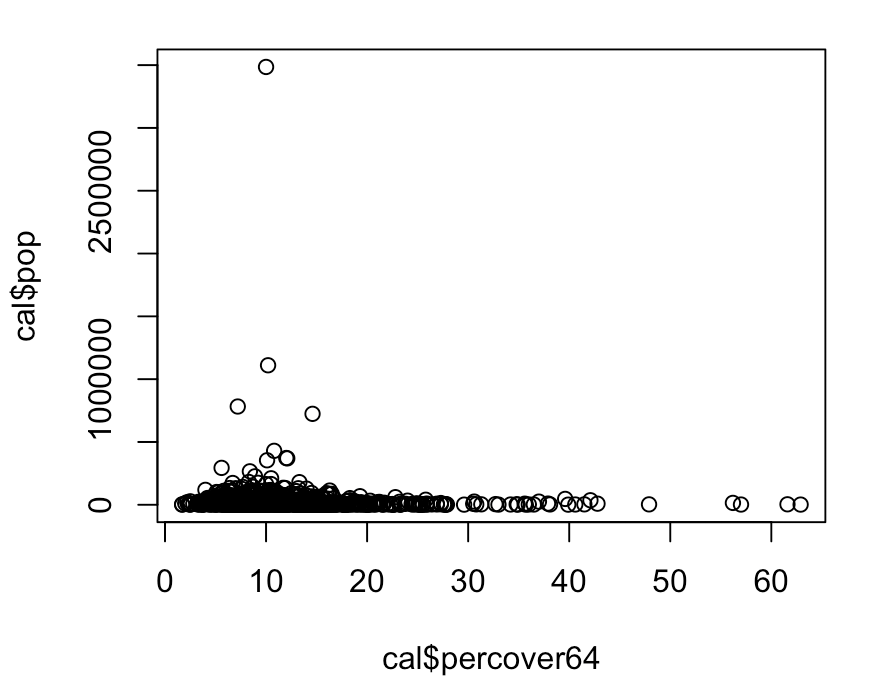
BSAN 450 Assignment 4

1) The file California.csv contains some census data on 858 cities in California. In this problem you are to examine the relationship between the variable pop (the total population of the city) and the variable percover64 (the percentage of the population that is over age 64). Read the data into R Studio using the following command.

cal=read.csv("California.csv")

1. Plot a scatter plot of pop versus percover64. Is there a relationship between these two variables?

There does not appear to be a relationship looking at the graph; however, the data a bunched together and it hard to determine a relationship.

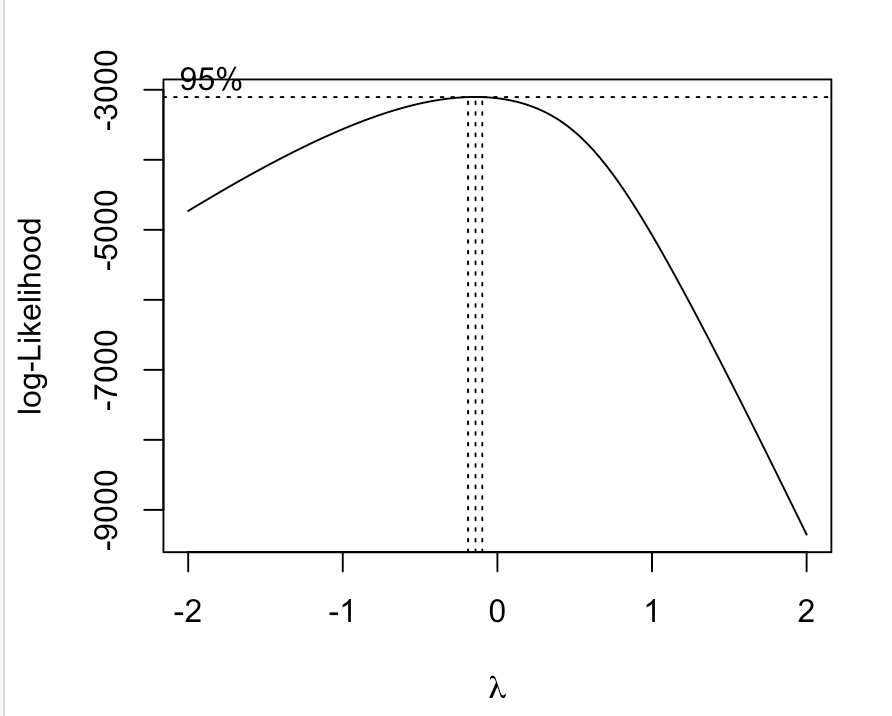


b) The source that I got this data from suggests that it would be appropriate to use the log of pop as the dependent variable. The source suggests that the application of this transformation is due solely to the skewness inherent in the variable pop and not the result of any regression diagnostics.

Since the variance of the pop is not constant for different values of percover64, you might consider a nonlinear transformation of pop. Execute the boxcox command in R to get an idea of a possible nonlinear transformation of pop. To execute this command you need to input a model. Use a simple linear regression model with pop as the dependent variable and percover64 as the independent variable.

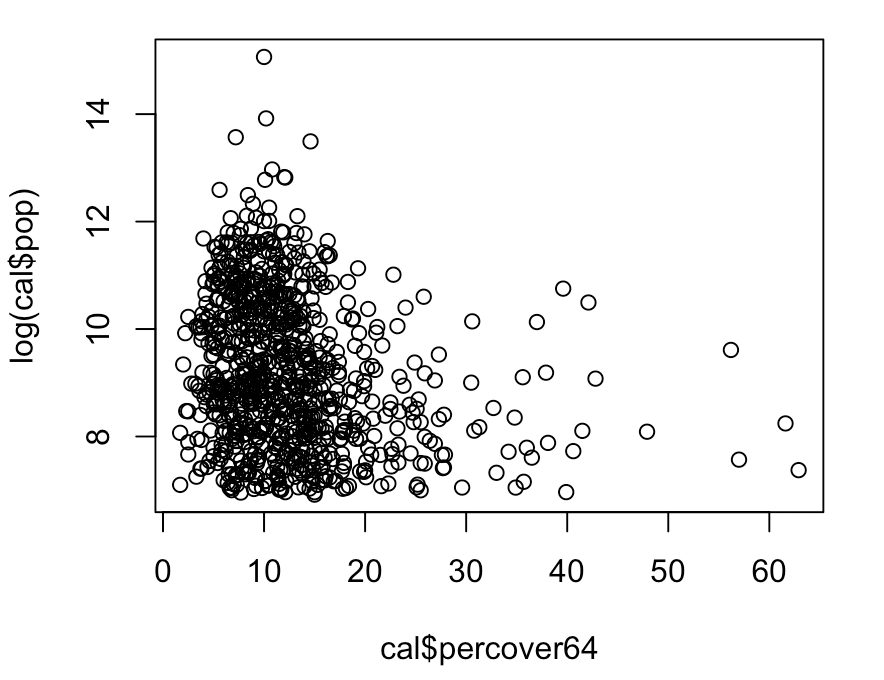
Does the log transformation seem reasonable?

Yes, the log transformation seems reasonable, because running the boxcox command resulted in a lambda of about 0, which indicates a log transformation.



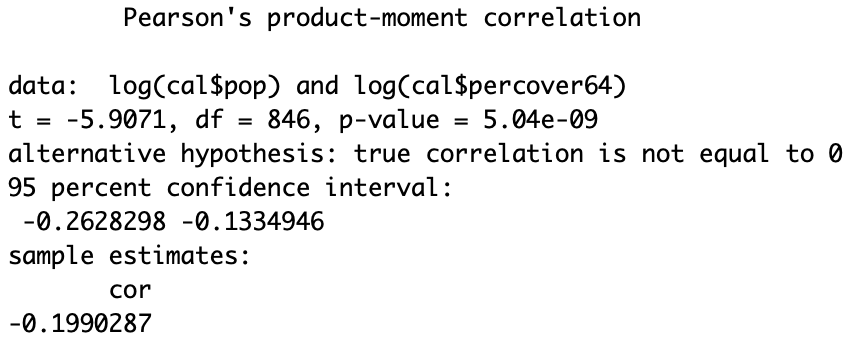
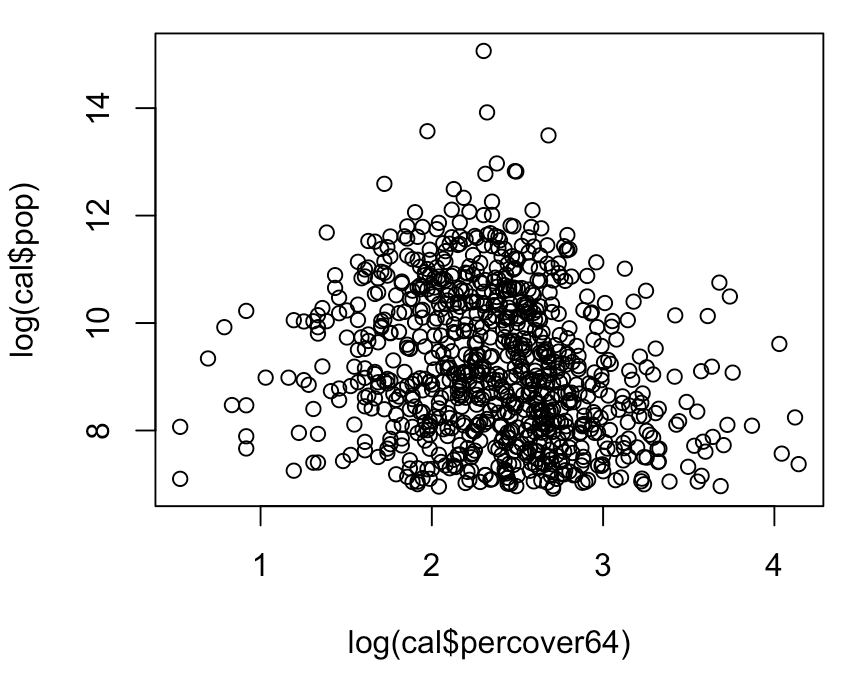
1. Plot a scatter plot of the log of pop versus percover64. Is there a relationship between these two variables?

There does not appear to be any realtionship between these two variables. Although that data is more spread out, there does not appear to be any apparent trend.

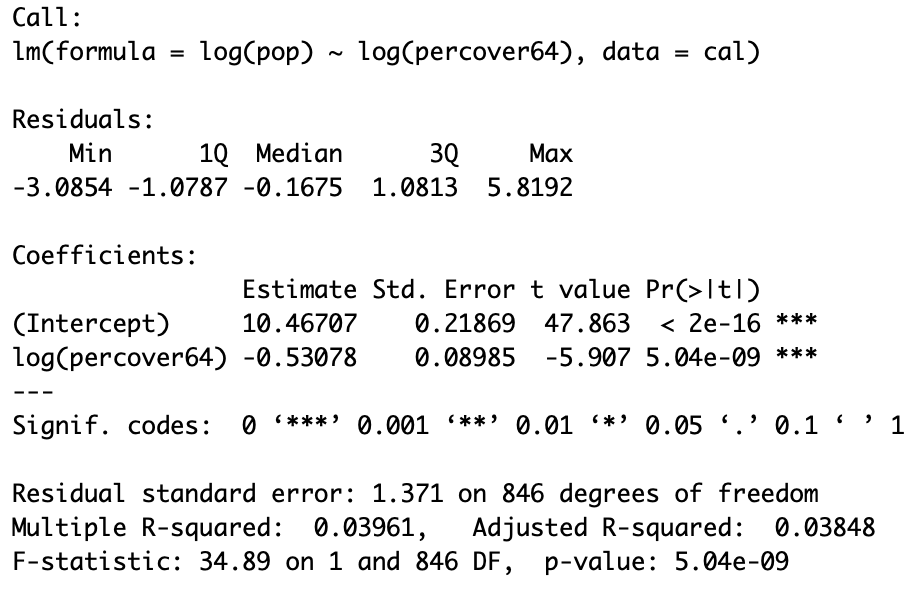
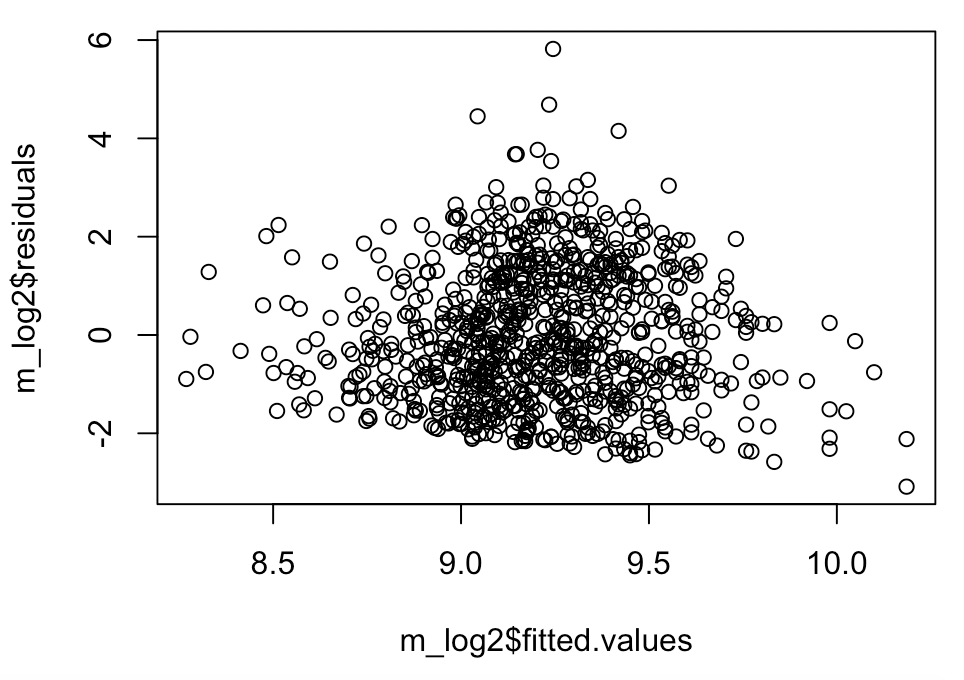


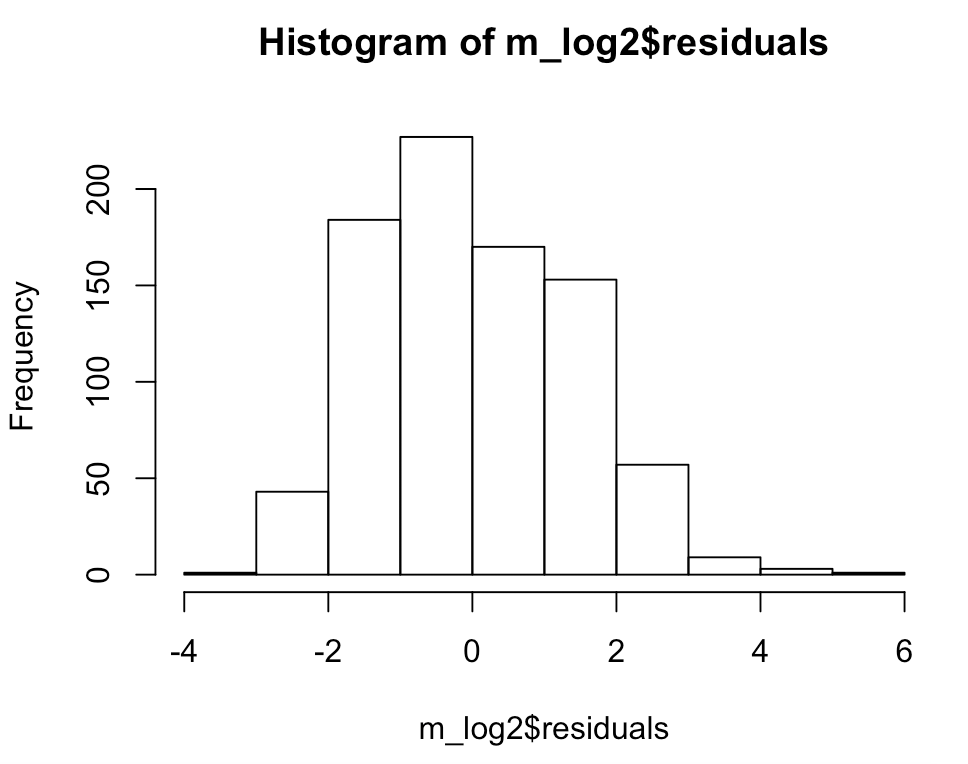
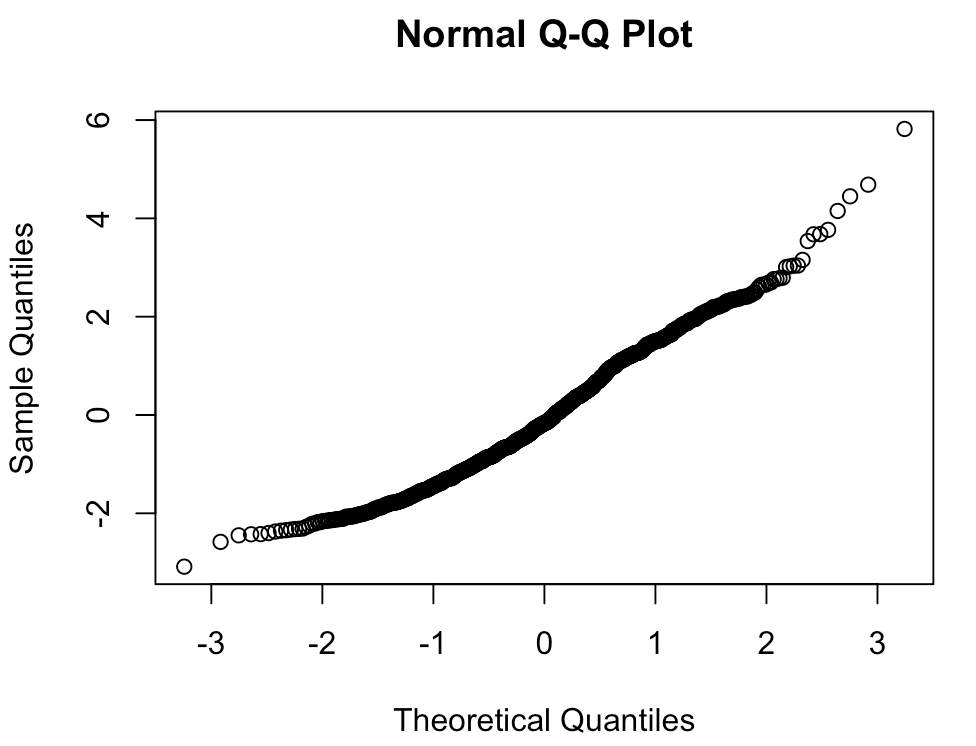
d) Plot a scatter plot of the log of pop versus the log of percover64. Is there a relationship between these two variables? Use the following R command to compute the correlation between the log of pop and log of percover64 and to test the hypothesis that this correlation is equal to 0.

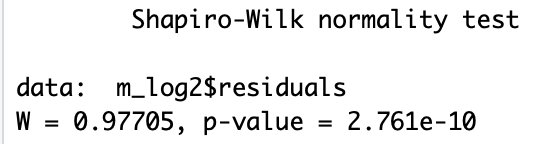
There does not appear to be any realtionship between these two variables. Although that data is more spread out, there does not appear to be any apparent trend.



e) Estimate a simple linear regression model with the dependent variable being log of pop and the independent variable being log of percover64. Perform the diagnostic checks of the model. Is there any indication of a problem?

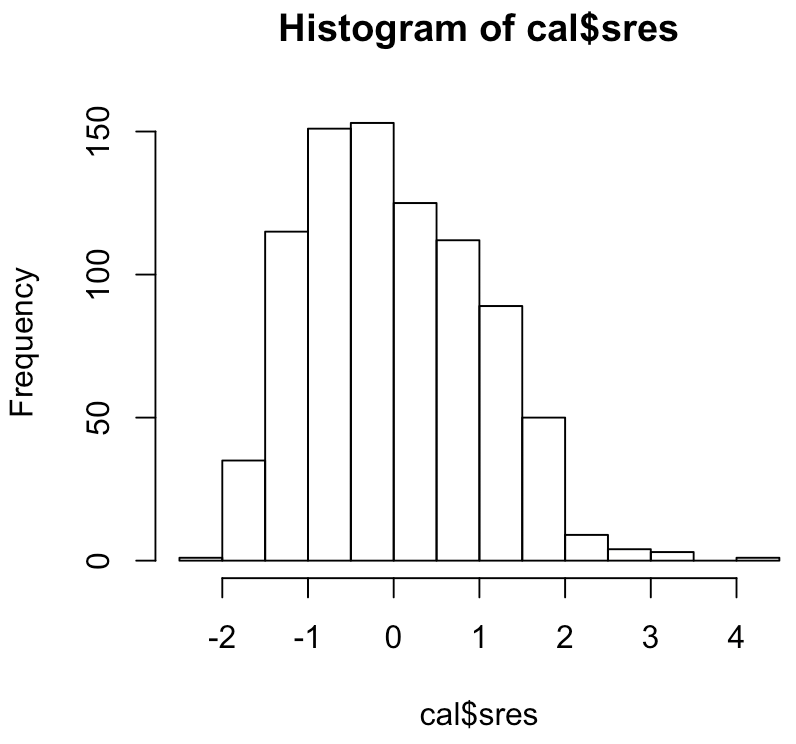
 



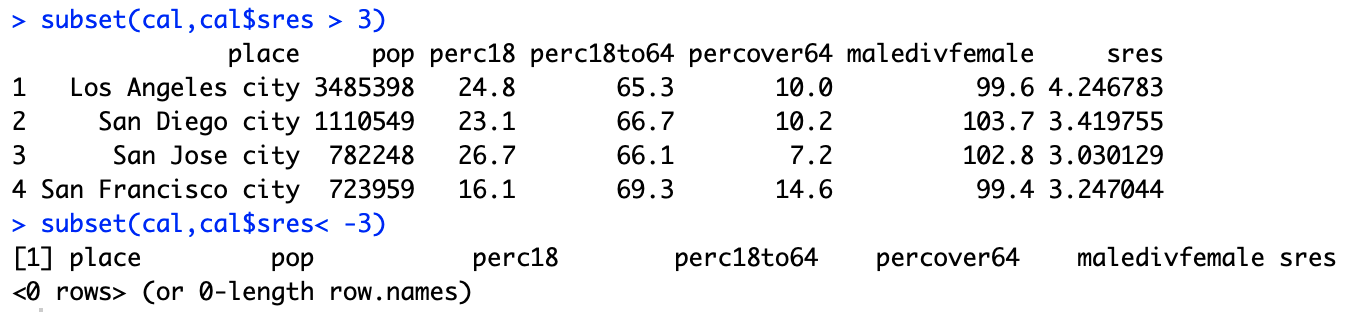
The p-value in the shapiro-wilk test is low indicating that the data does not follow a normal distribution. The variance in the residual v fitted values looks small in the middle and then large on the outer edges. The variance also looks like it is decreasing. This indicates a problem with our current model.

f) It may be that there are some outliers in this data. To check for outliers it is easiest to use the standardized residuals, these are the residuals divided by their standard deviation. If the absolute value of a standardized residual is larger than 3 then the associated data is an outlier. The following R commands can be used to compute the standardized residuals. In the second command the standardized residuals are stored in the data frame named cal in a variable named sres.

Plot a histogram of the standardized residuals. Does there appear to be any outliers.

Yes it looks like there is some outliers above 3.

g) The following R commands will print out the standardized residuals that are smaller than -3 and which are larger than 3. How many outliers are there in this data. Which cities are outliers?

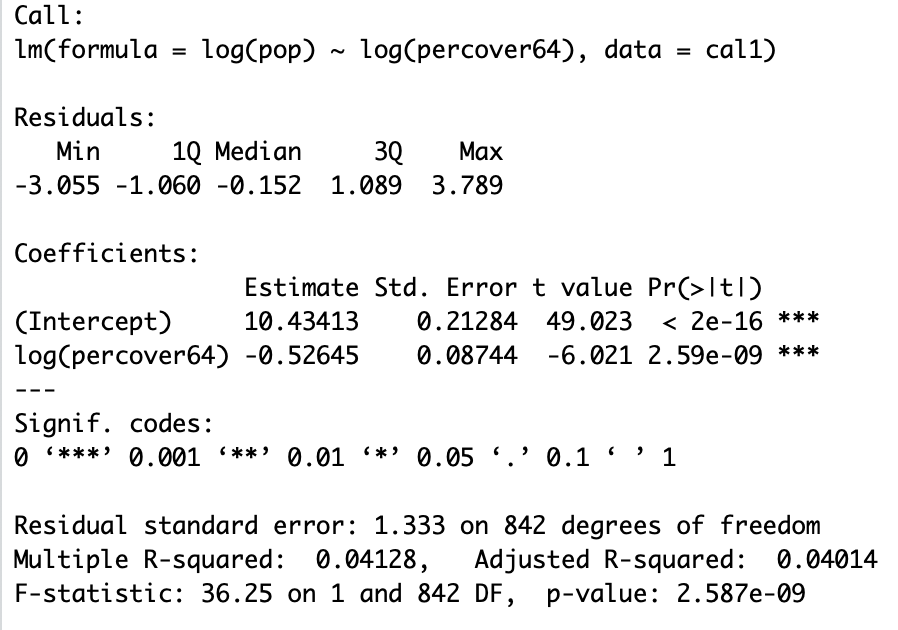
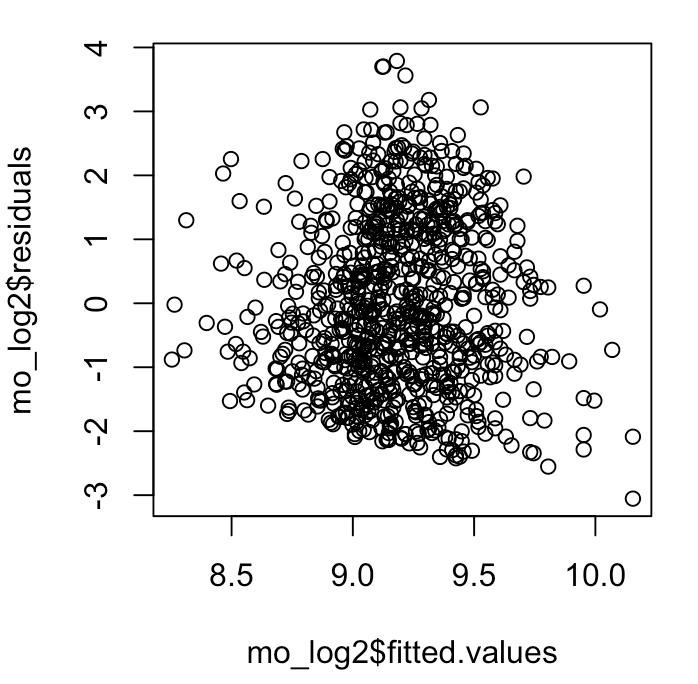
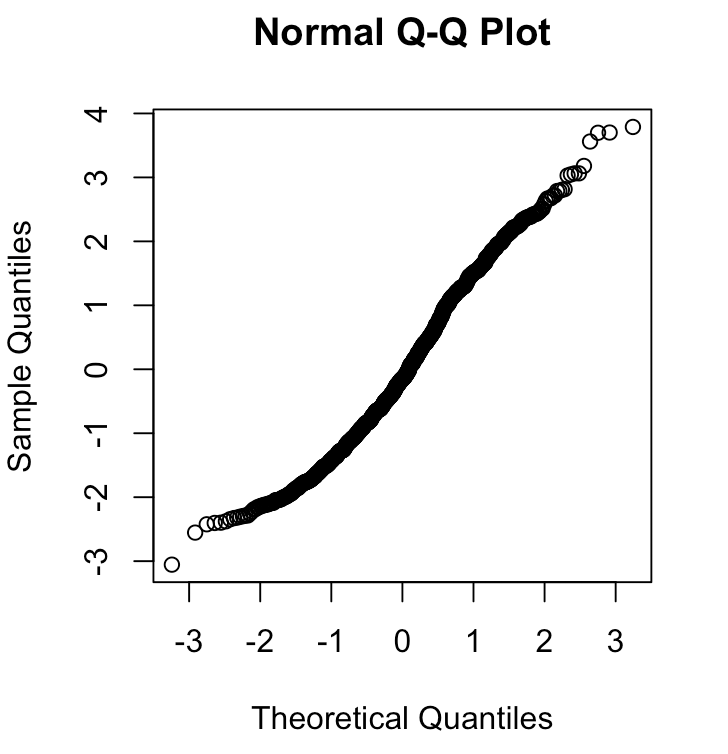
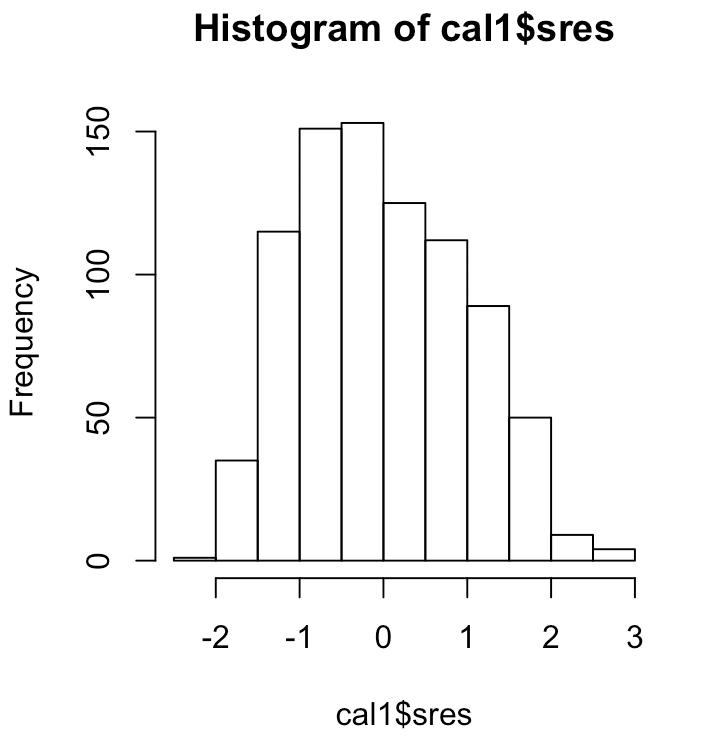
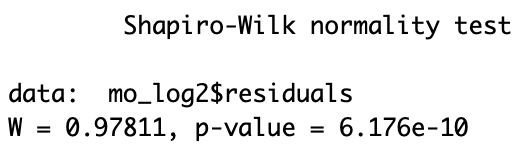


There are 4 outliers that have a standard residual larger than 3

The following R command will keep the observations for which the standardized residuals are less than 3 and put these observations in a new data frame named cal1. Execute this command.

cal1=subset(cal,abs(cal$sres) < 3)

h) Re-estimate the model with the outliers removed. This can be done by using the data frame cal1. Plot a histogram of the standardized residuals from this model. Are there any more outliers in the new model?

All of the data now fits within 3 standard deviations of the residual, which indicates no outlier in the data.

i) For the model you obtained in part h) answer the following questions. Is the log of percover64 statistically significant? What is the value of R-squared? What does the value of R-squared suggest about the model?

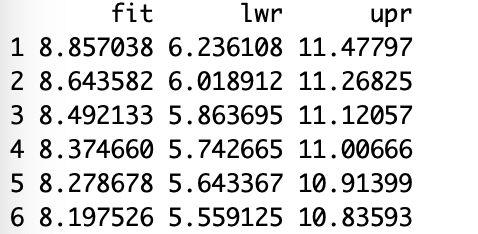
The log of percover64 is significant. The p-value is small, indicating a relationship. The r-squared value is .04128 which suggests that this model is not the most precise model for the data.

It is useful to examine the predictions that can be made from the model estimated in part h). The following R commands compute the predictions for the values of percover64 equal to 20, 30, 40, 50, 60, and 70. In these commands the expression logo64 needs to be changed to the name that you used for the variable corresponding to the log of percover64 and the expression fit1 needs to be changed to the name you gave to the model you estimated in part h). Make the changes required in these commands and execute these commands.

newdata=data.frame(logo64=log(c(20,30,40,50,60,70)))

pred=predict(fit1,newdata,interval="prediction")

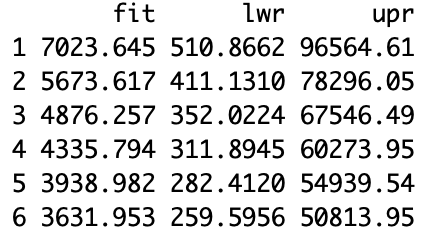
pred



j) The values computed in part i) are the predictions of the log of pop, thus we want to convert these back to the original scale. This is done by reversing the transformation, exponentiating the results. The following R command converts the predictions and the lower and upper prediction values to the original scale. Execute these commands.

pred=exp(pred)

pred



Suppose that you wanted to predict the population of a city in California that has 20 percent of the population over age 64, what is a 95% prediction interval for this population of this city? Compare this to the prediction interval for a city that has 70 percent of the population over age 64.

20 percent: (510.8662, 96564.61)

70 percent: (259.5956, 50813.95)

Both of these intervals are very large and considering that the percent values are significantly different the prediction intervals are similar.

k) The prediction intervals obtained for the variable pop in part j) may not be that great; however it is useful to compare those to the predictions that could be made about pop without using the variable percover64. If the only data that is available is the population then the following prediction can be made about the population of a city in California.

Since the data for pop is extremely skewed, it is best to transform the data by taking the logs. The mean of the log of pop (excluding the 4 outliers) is: 9.1828. The standard deviation of the log of pop (excluding the 4 outliers) is: 1.3610.

A 95% prediction interval for the log of pop has the following lower and upper bounds

Lower bound = 9.1828 – (1.96)\*(1.3610) = 6.5152

Upper bound = 9.1828 + (1.96)\*(1.3610) = 11.8504

Converting the lower and upper bounds to the original scale:

Lower bound: exp(6.5152) = 675.3

Upper bound: exp(11.8504) = 140,136.3

Thus, the 95% prediction interval for the population of a city in California (excluding the 4 largest cities) is (675.3 to 140,136.3).

Compare this result to the results you obtained in part j) where the percover64 variable was used.

Given the 95% prediction interval for the population, it shows that in part j that the percover64 does impact the prediction, narrowing the population intervals for a given percent.

2) The data file Cereals.csv contains information about the characteristics of different types of cereals. The first column of the file is the name of the cereal and should not be used in the analysis. Suppose the purpose is to find a regression model to predict the variable Rating based upon the other characteristics of the cereal. The possible dependent variables are:

Manuf: A categorical variable indicating the manufacturer of the cereal

Type: A categorical variable indicating if the cereal is cold or hot.

Calories: The number of calories in a serving.

Protein: The amount of protein is a serving.

Fat: The amount of fat in a serving.

Sodium: The amount of sodium in a serving.

Fiber: The amount of fiber in a serving

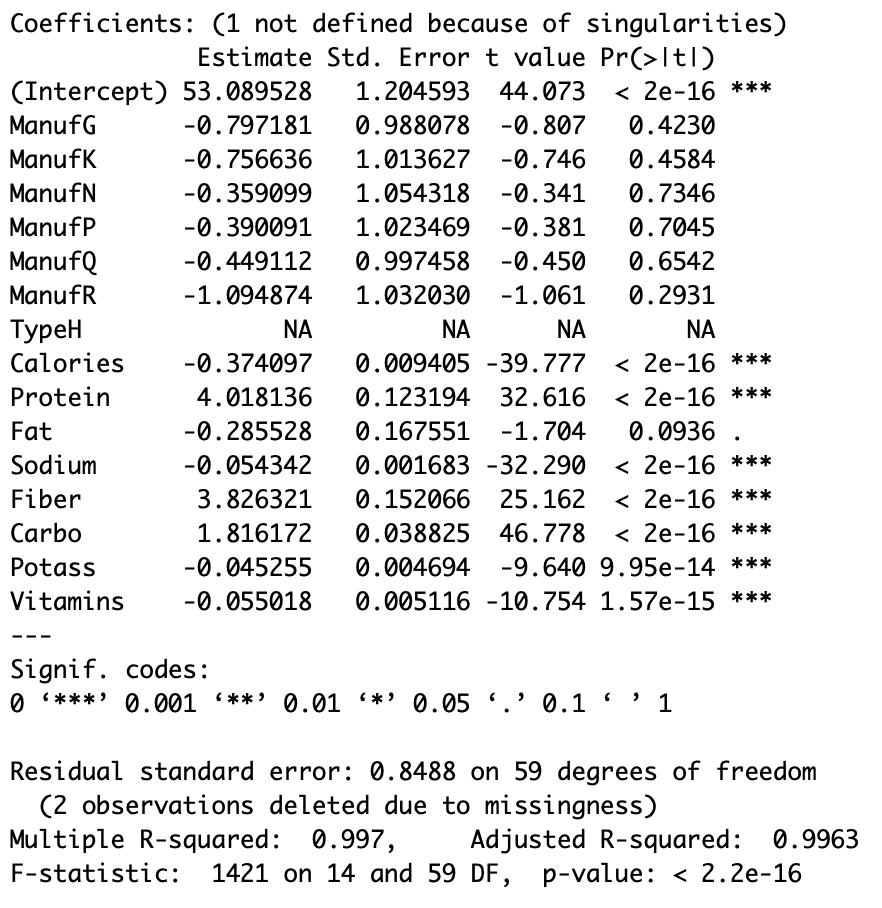
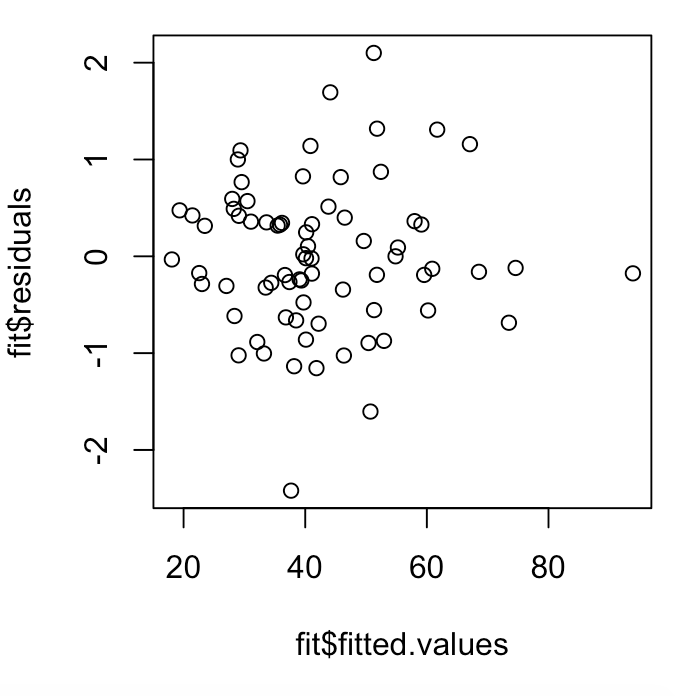
Carbo: The amount of carbohydrates in a serving.

Sugars: The amount of sugar in a serving

Potass: The amount of Potassium in a serving

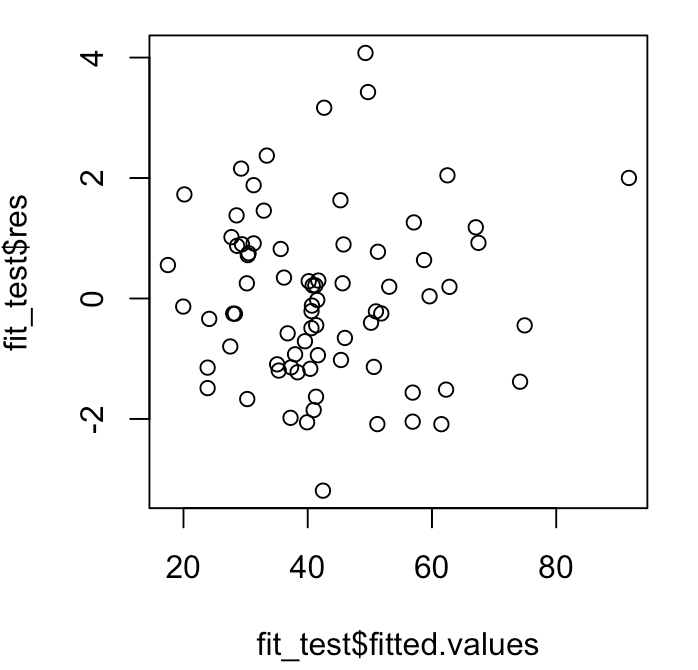
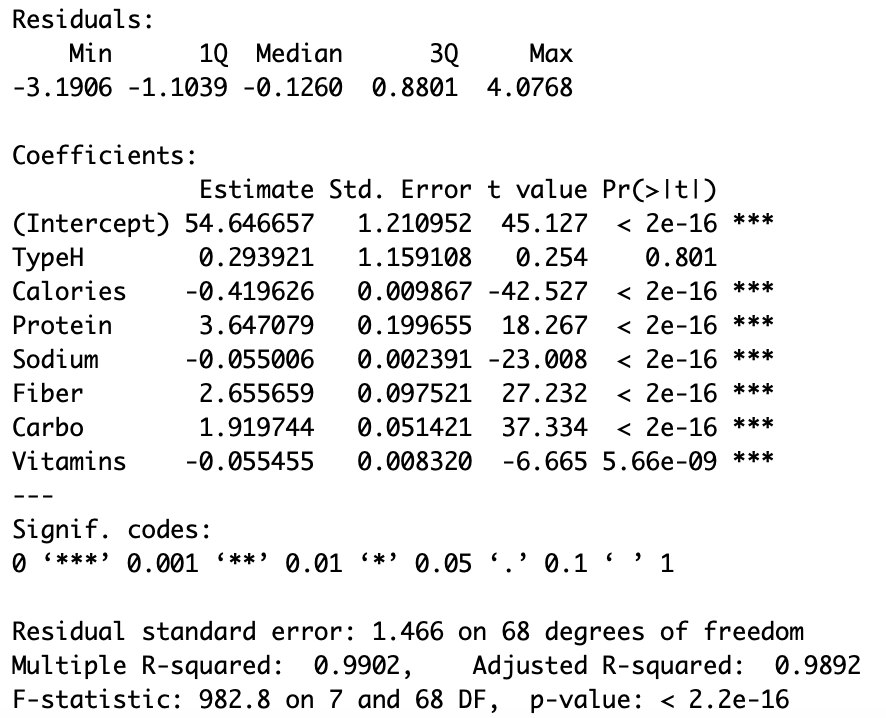
Vitamins: The percent of required daily vitamins in a serving.

a) Perform a preliminary analysis of the data to determine which independent variables if any are related to the variable Rating. Is there evidence of a need to transform any of the variables?

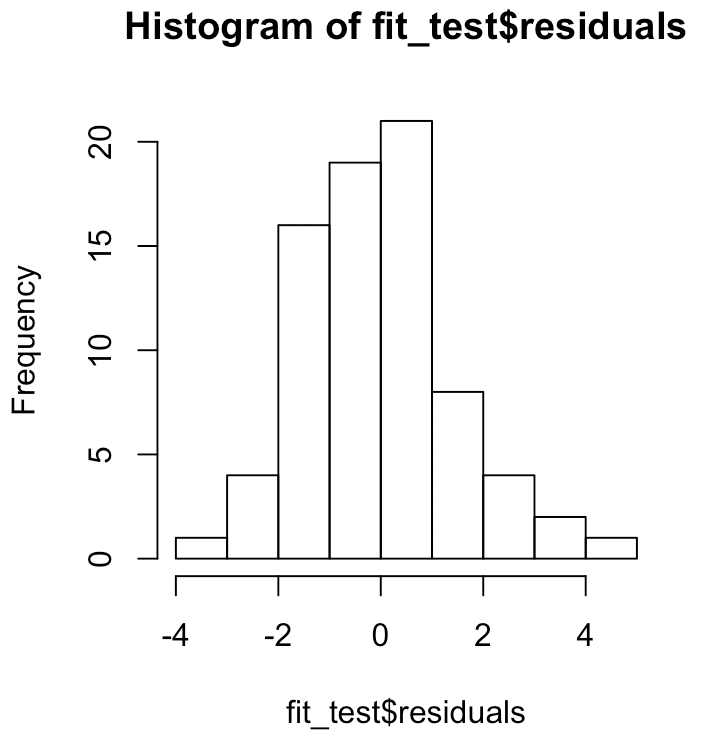
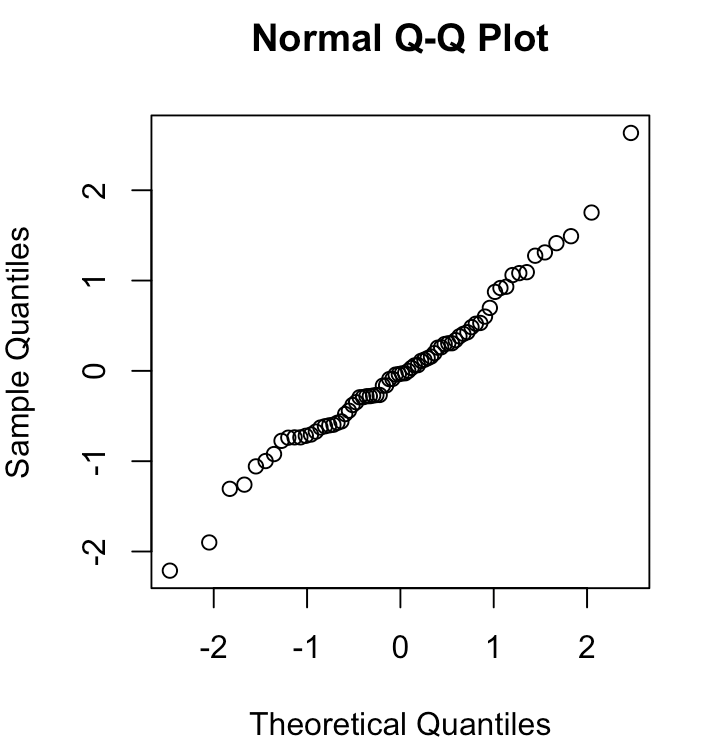
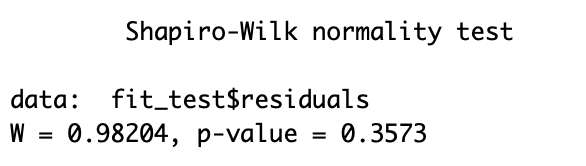
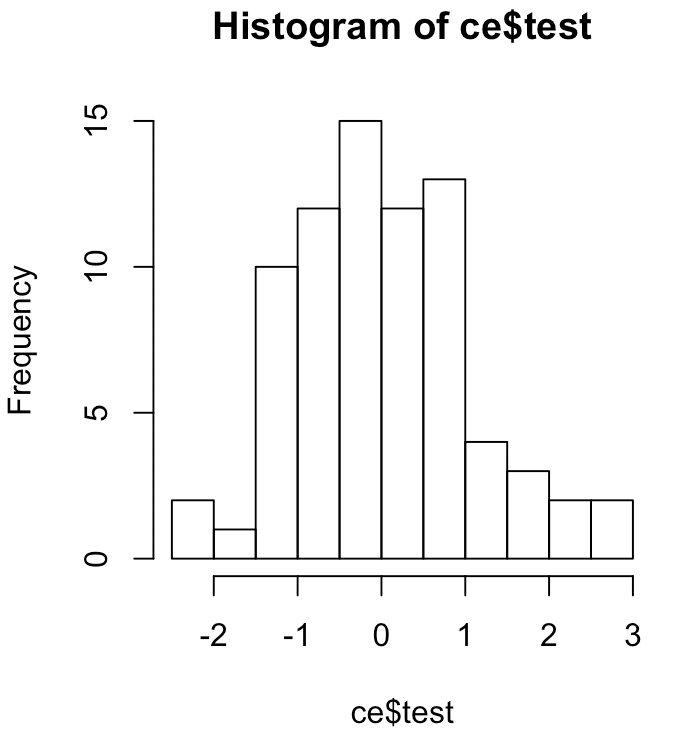
 

There does not look like a transformation is needed. The residual graph has ok variance. However, there are some variables that are not needed in the model: Manuf, Fat. I did not remove typeH since it is a part of a variable and TypeC appears to be ok.

b) Based upon part a) determine a multiple regression model that you could fit to explain the variable Rating. Fit this model. Make any modifications to this model that you believe are appropriate.

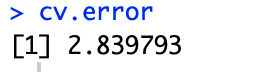
 

c) Perform diagnostic checks of the final model you came up with in part b). Make sure you examine the standardized residuals to check to see if there are any outliers. Do these checks suggest any changes that should be made? If yes, make those changes and refit the model. Is there an indication of any outliers? If yes identify the cereal corresponding to the outlier.

The graphs look good. There is no indication for a need to change the model. There also appears to be no outliers.

d) Once you find a final model, use leave one out cross validation to determine the cross validation standard error for this model. How does this cross validation standard error compare to the residual standard error that was estimated in the model you found in part c)?



This error is larger than the residual standard error of 1.466.