Assignment - 4

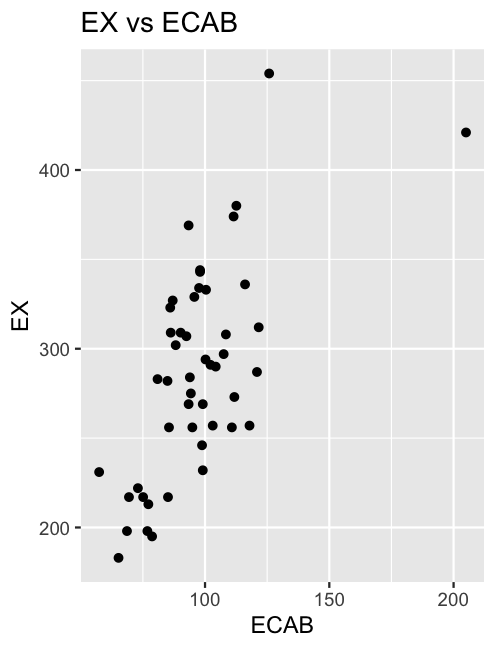
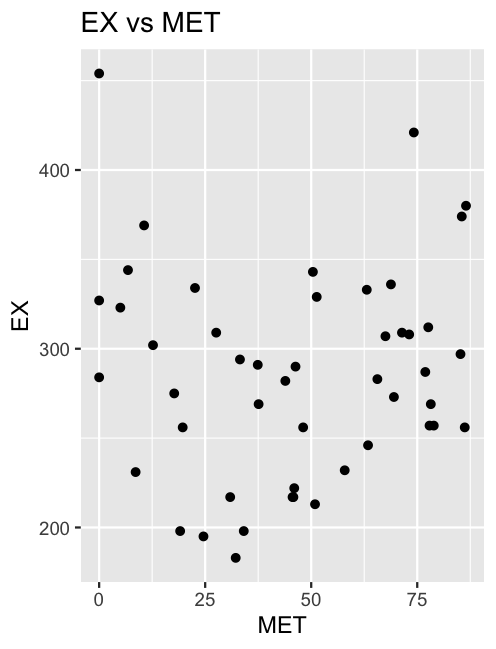
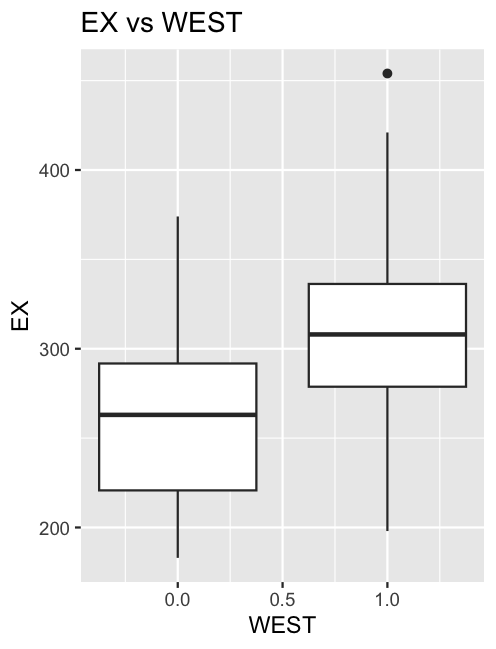
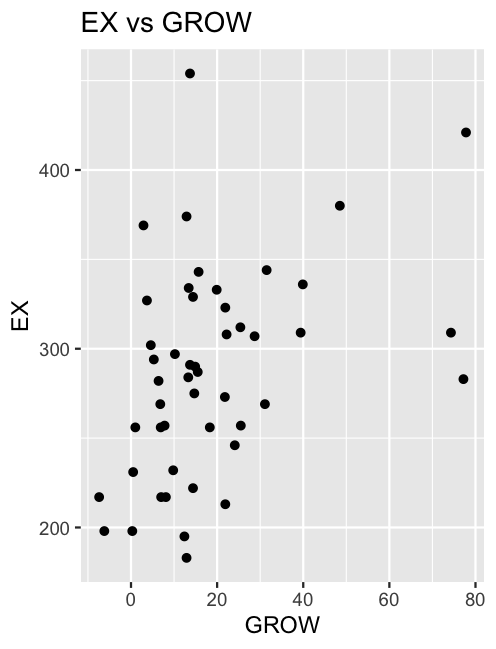
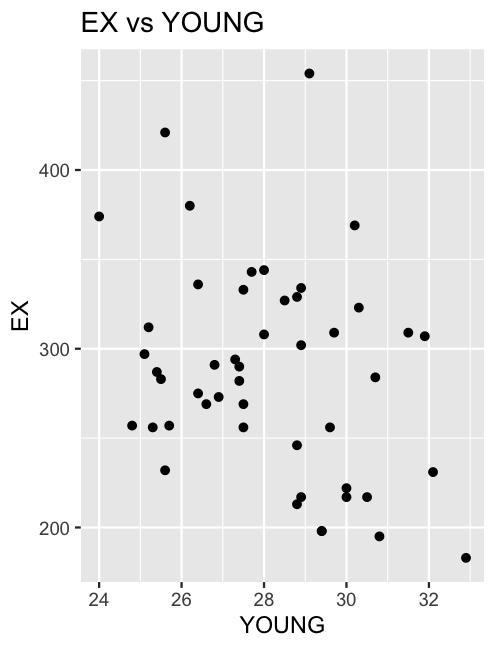
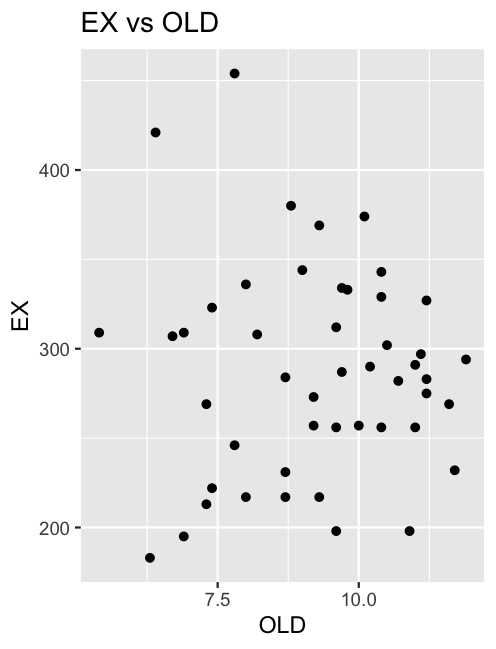
BSAN 450 (Spring 2023)

**This assignment is due on February 16, 2023 at 9:00 PM Central. The total points possible are 100 and there are two (2) problems, each carrying equal points. You can form groups to attempt these problems. Each group will submit one copy of the assignment on Canvas, either in word or pdf, and the assignment should clearly include the names of the group members**.

1. The purpose of this exercise is to find a model to predict the per capita state and local public expenditures for different states. This data is from 1960. The variables are as follows.

* EX: Per capita state and local public expenditures (in dollars)
* ECAB: Economic ability index, in which income, retail sales, and the value of output (manufactures, mineral, and agricultural) per capita are equally weighted.
* MET: Percentage of population living in standard metropolitan areas
* GROW: Percent change in population, 1950-1960
* YOUNG: Percent of population aged 5-19 years
* OLD: Percent of population over 65 years of age
* WEST: Western state (1) or not (0)

The data is on a file named Expend.csv. Use this data to do the following.

1. Read the data into R Studio. Plot the expenditures versus each of the independent variables and comment on the plots. Describe the relationship between the expenditures and each of the independent variables.

ECAB - There is a positive, mostly linear relationship between ECAB and EX. However, there may be an outlier near (220, 440).

MET - There is no pattern in the plot between MET and EX. The mean looks to hover around 300 with an equal number of observations equidistant above and below the mean.

GROW - There may be a slight positive relationship between GROW and EX. However, it does not look like it would be significant in a model.

YOUNG - There may be a slight decreasing relationship between YOUNG and EX. However, it does not appear to be a strong relationship.

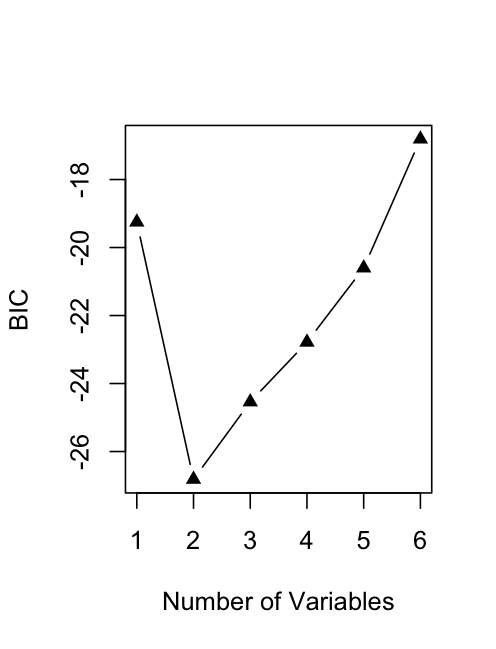
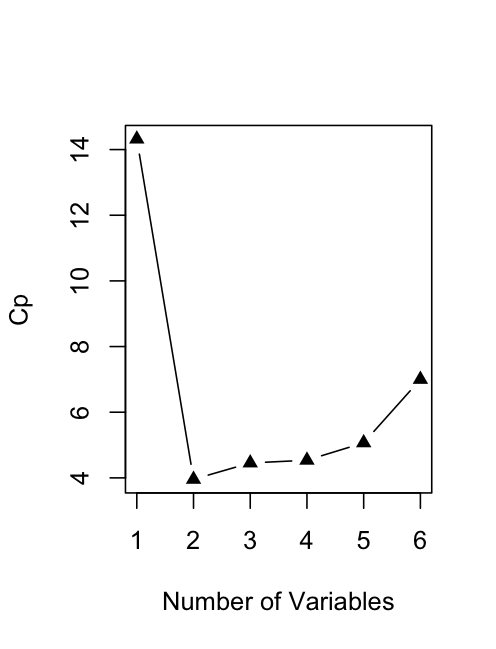
OLD - There does not appear to be a relationship between OLD and EX.

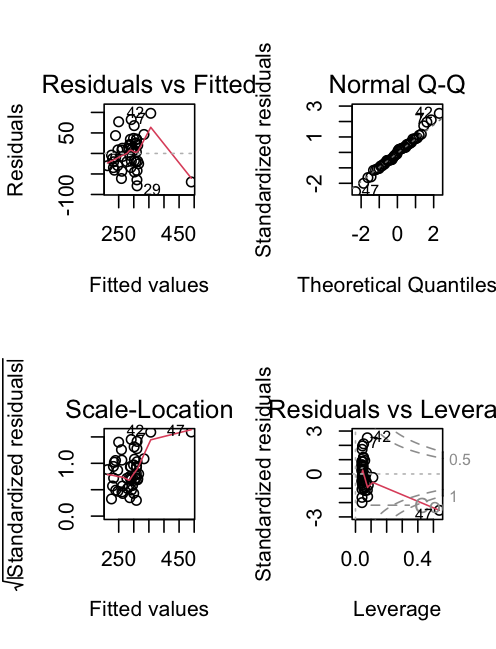
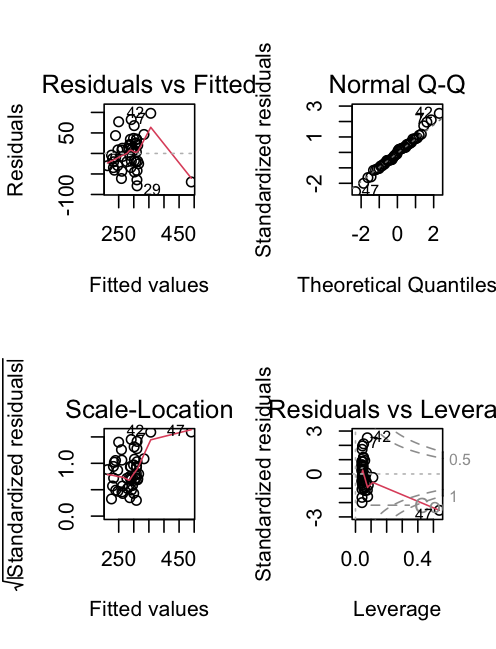
WEST - Places that are WEST have a higher median EX than non WEST places. The mean of WEST is above the 3rd quartile for non WEST places.

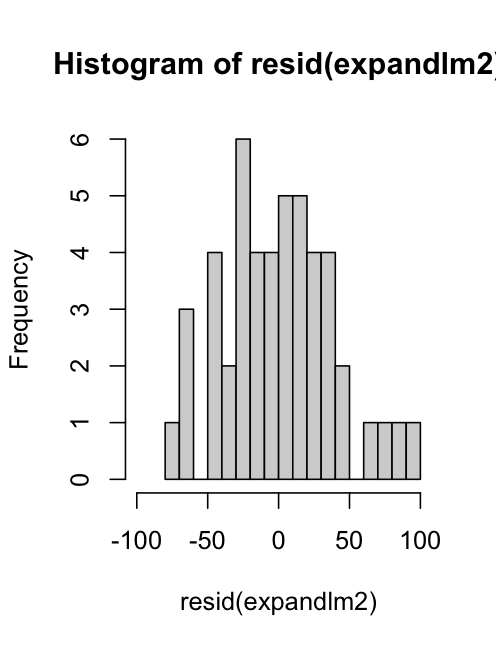
[*Hint: Do you notice that variables YOUNG and OLD do not seem to have much of a relationship with EX?. Also, what do you think about the variable MET and the nature of the relationship with EX?*]

MET and EX - states that have a higher metropolitan area tend to have a greater need for infrastructure and other expenditures; however, the data does represent this.

1. Based on the plots in part a, fit an appropriate regression model. Perform the diagnostic checks for the model you fit. If appropriate make adjustments to your model by either dropping variables that are not needed, adding variables or otherwise changing the model. If you fit a new model perform the diagnostic checks for the new model.
2. Fit a saturated model
   1. ECAB and WEST were significant predictors at the 95% confidence level and MET at the 90%.
3. We fit a few more models containing significant predictors from the saturated model and that made the most intuitive sense. Namely,
   1. Model 1: EX = ECAB + WEST
   2. Model 2: EX = ECAB + WEST + MET
      1. MET was not a significant predictor in this model

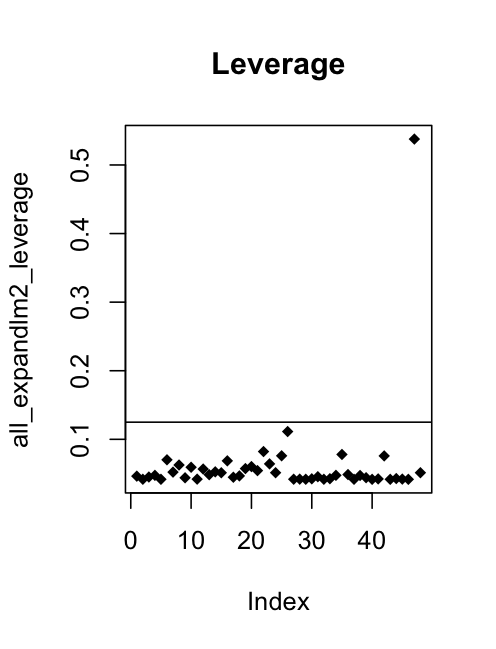
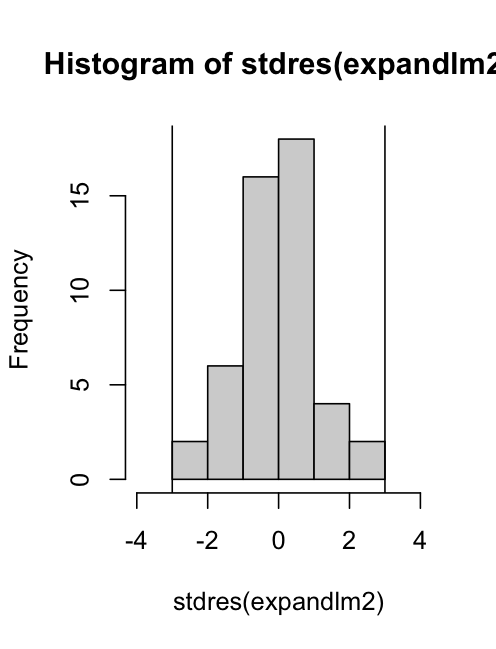
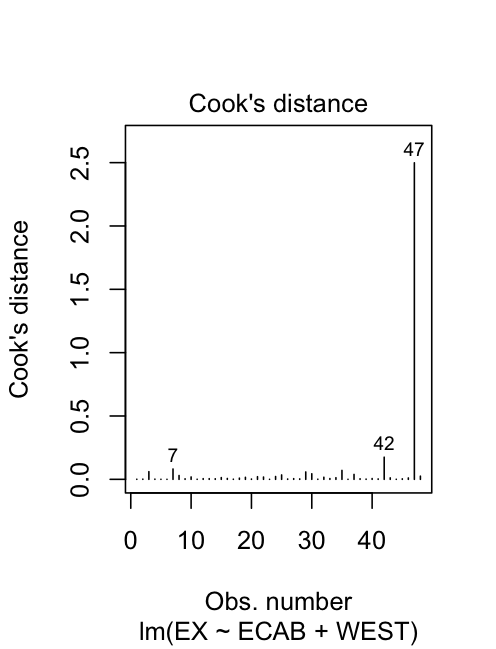
* We also fit a few polynomial models, but concluded adding a polynomial did not lead to significant improvements. 

1. Which model should we use?
   1. Using Mallow’s Cp and BIC, a model with two predictors should be used, and those predictors should be ECAB and WEST according to best subset.
   2. 5-fold CV:
      1. Model 1 RMSE: 43.92613
      2. Model 2 RMSE: 46.77509
2. Adding MET consistently produces a worse model, so we are going with Model 1 (2 predictors).
3. Model Diagnostics

* There is an issue with the variance of the residuals (bptest p-value: 0.04384).
* No issues with normality (shapiro.test p-value: 0.7557).
* There is an obvious high leverage point at point 47 and potentially one at point 42.
* The issues of non-constant variance will likely be fixed after taking into consideration any influential points.

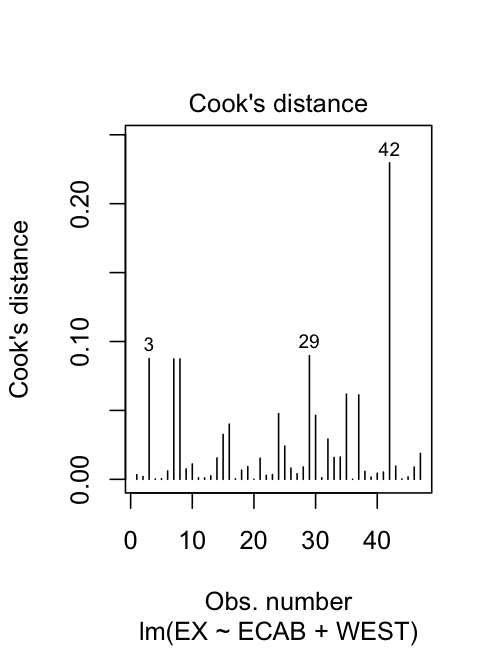
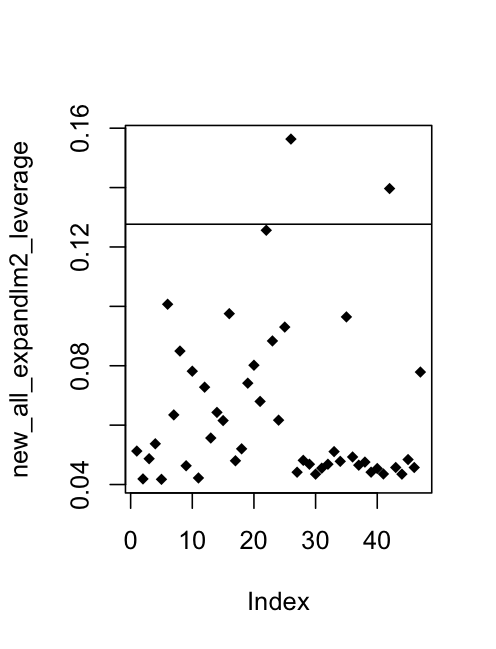
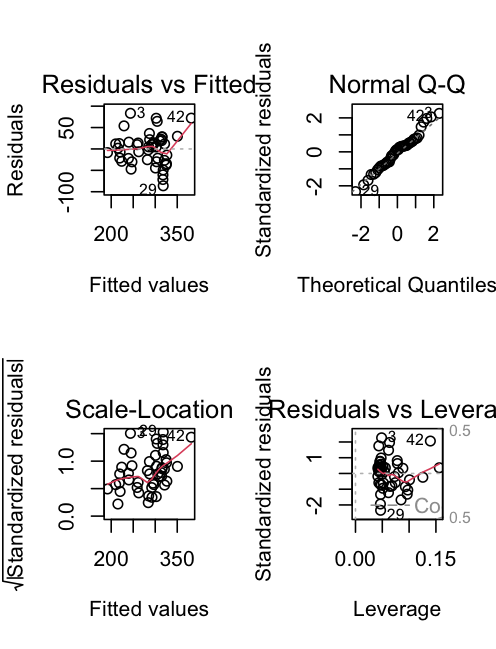
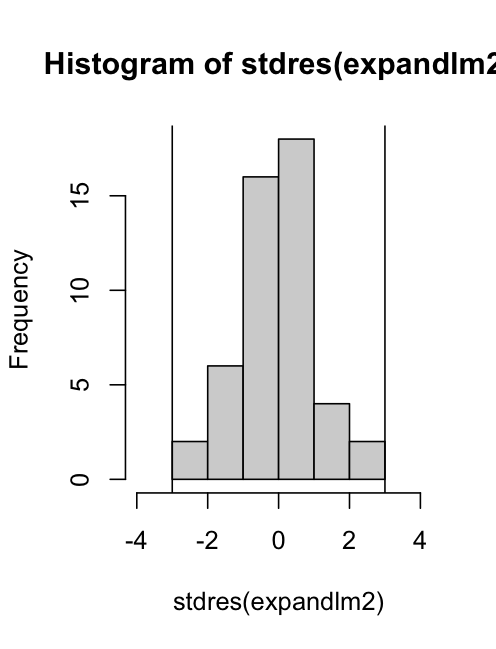
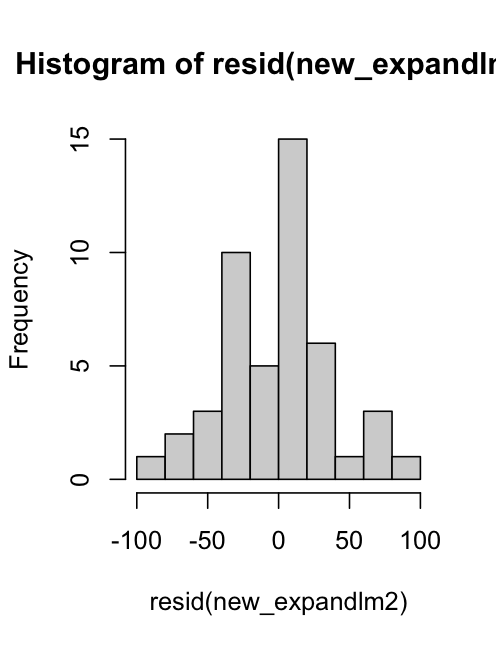
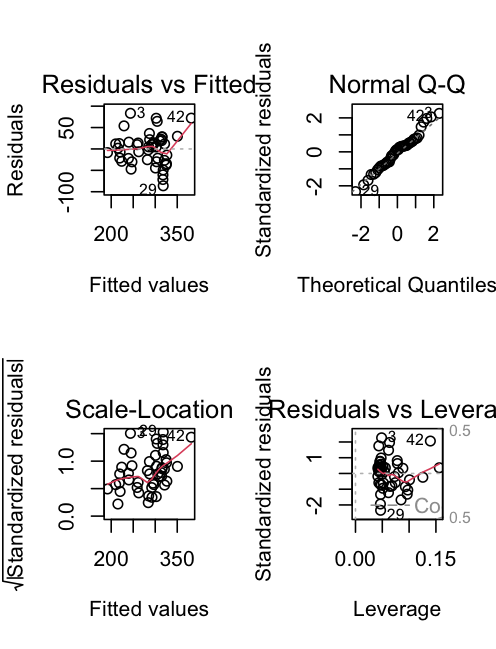
1. For the model that you fit in part b compute and plot the values of the leverage and Cook’s distance.

Recall that if the fitted model has *p* + 1 *β* parameters, then the average leverage is (*p* + 1)*/n* and cases in which the leverage exceeds twice its average are considered high-leverage cases. Are there any high leverage cases in this example?

1. Current model
   1. The Cook’s Distance for Point 47 is 2.498663 so it is highly influential and needs to be removed. Point 42’s Cook’s Distance is only .17, so we left it in for now.
   2. There are no points with a standardized residual outside of +- 3.
   3. Point 47 is also a high leverage point (0.53776 > 6/48).

Also recall that cases where Cook’s distance is greater than 1 are of great concern and cases where Cook’s distance between .5 and 1 should be looked into. Do the values of Cook’s distance suggest that any cases should be flagged for study?

1. Identify the cases which have high Cook’s distance. Refit the model by removing the case with the high Cook’s distance. Compare the parameter estimates of the model that includes the case with the large Cook’s distance and without the case that includes the large Cook’s distance.
2. We removed Point 47 and refit the model.
3. Old model: EX = 102.30 + 1.70(ECAB) + 40.48(WEST)
4. New model: EX = 35.70 + 2.39(ECAB) + 45.52(WEST)
5. New diagnostics:
   1. The Cook’s Distance for Point 42 is 0.2296599, so it is not highly influential.
   2. There are no points with a standardized residual outside the absolute value of 3.
   3. Point 42 (leverage = 0.1563342 > 6/47) and Point 26 (0.1396709 > 6/47) are high leverage points. However, given that their Cook’s Distance is well under .5, we decided to leave them, especially since this dataset is so small.
   4. There are no issues with the residuals in general; non-constant variance has been resolved; normality is still good.
   5. Additionally, 5-fold CV RMSE for new model: 38.76007 (was 43.92613 before)



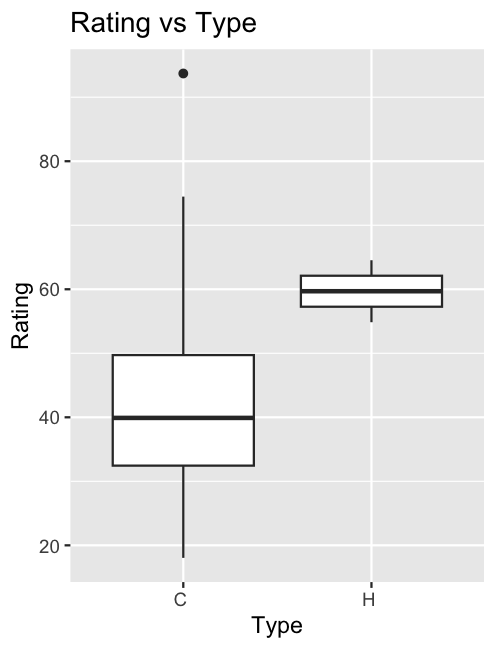
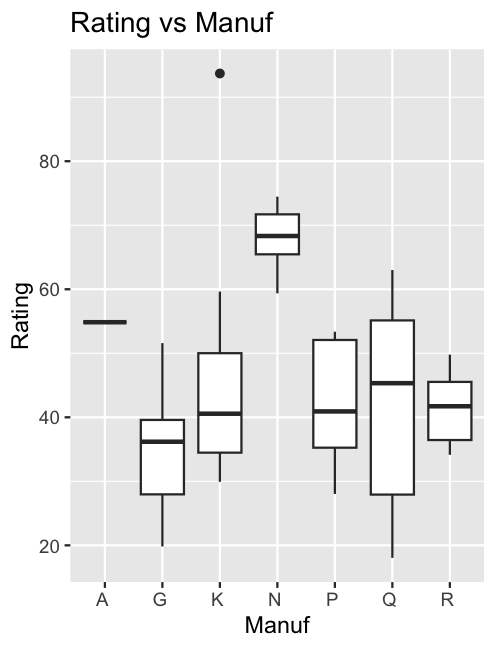
1. Make predictions for the model including and excluding the data points you dropped. Use the following newdata for predictions.

| newdata= data.frame(ECAB=(c(100,100,100,150,150,150,175,175,175)),  MET=(c(20,40,60,20,40,60,20,40,60)),  WEST=(c(0,0,0,0,0,0,0,0,0))) |
| --- |

* Old: 271.91, 271.91, 271.91, 356.72, 356.72, 356.72, 399.13, 399.13, 399.13
* New: 274.78, 274.78, 274.78, 394.32, 394.32, 394.32, 454.09, 454.09, 454.09
  + The new model has a greater emphasis on ECAB and WEST, but a lower y-int.

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.

1. Perform a preliminary analysis of the data to determine which independent variables if any are related to the variable Rating.

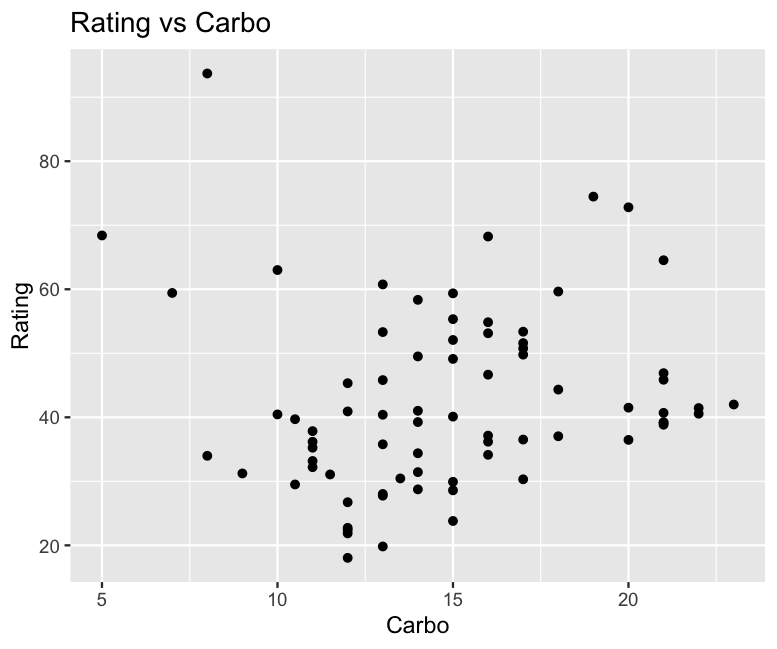
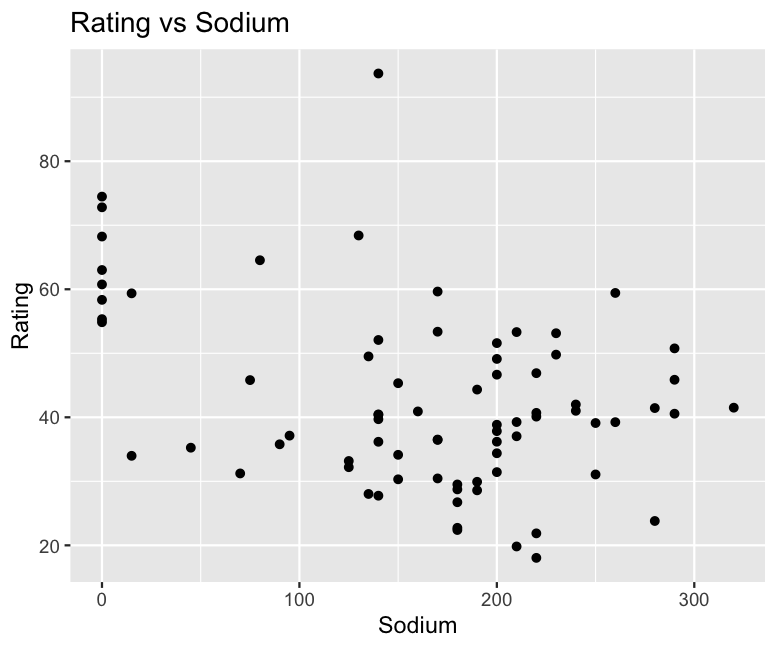
Type - While Type H looks to have a considerably higher 1st quartile, median, and 3rd quartile than Type C, there is not enough data to truly determine if there is a difference between the two types. There are 74 observations for Type C and only two for Type H.

Manuf - Similar issues reside with Manuf and Rating. Manufacturers G and K have over twenty cereals while no other manufacturer has more than nine. While there looks like significant differences in the ratings between manufacturers, the sparse number of observations will likely lead to non-significant differences in a model.

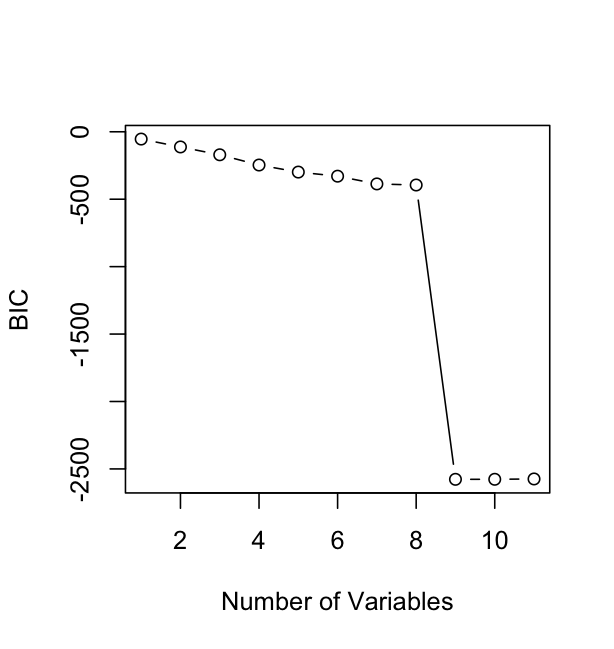
Correlations between continuous variables and Rating:

* *CalCor -0.68876328* - moderate, negative
* *ProteinCor 0.47030522* - moderate, positive
* *FatCor -0.42052619* - moderate, negative
* *SodiumCor -0.39713585* - moderate, negative
* *FiberCor 0.58390015* - moderate, positive
* *CarboCor 0.08871244* - very weak
* *SugarsCor -0.76390195* - strong, negative
* *PotassCor 0.41578244* - moderate, positive
* *VitaminsCor -0.23377517* - weak, negative

[*Hint: What do you think about Sodium and Carbo?*]

Based on the correlation of the two variables with rating, they are quite different. Graphically, though, sodium and carbs look somewhat similar (multicollinear).

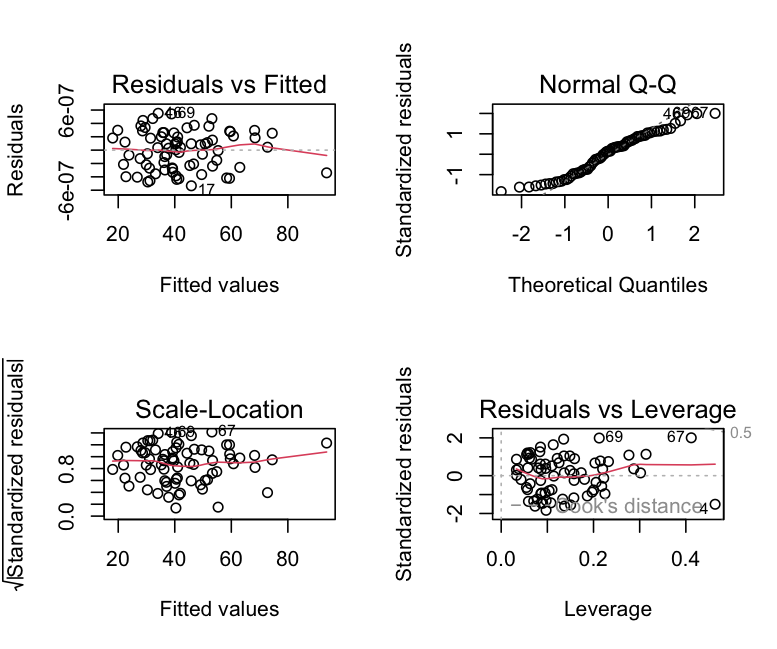
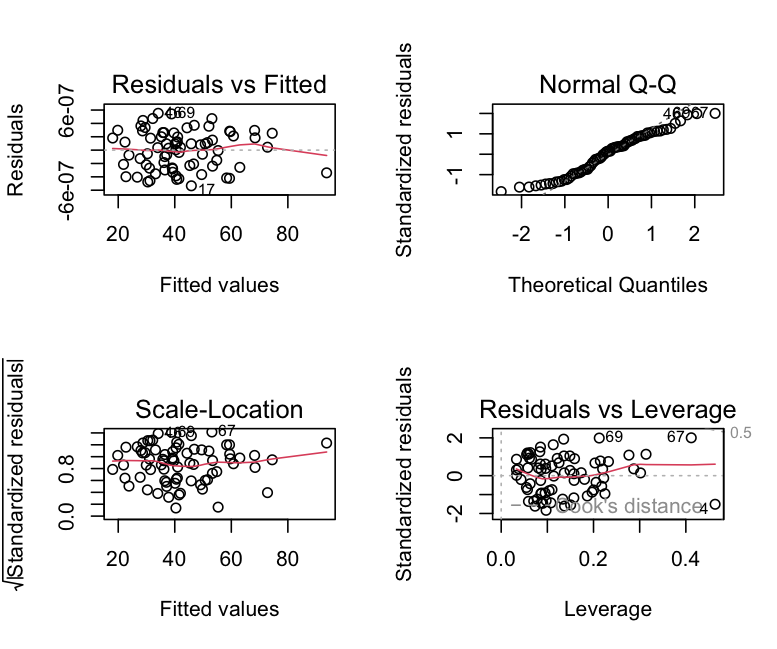
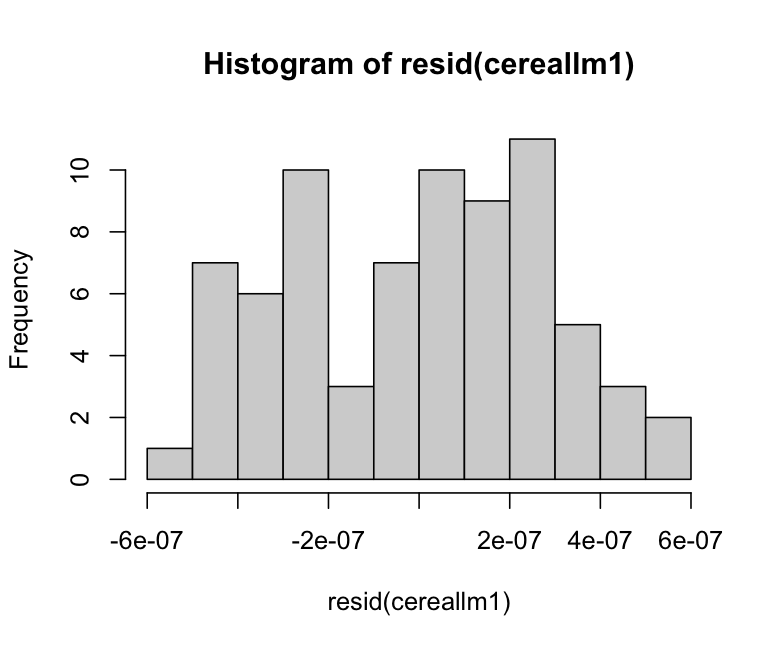
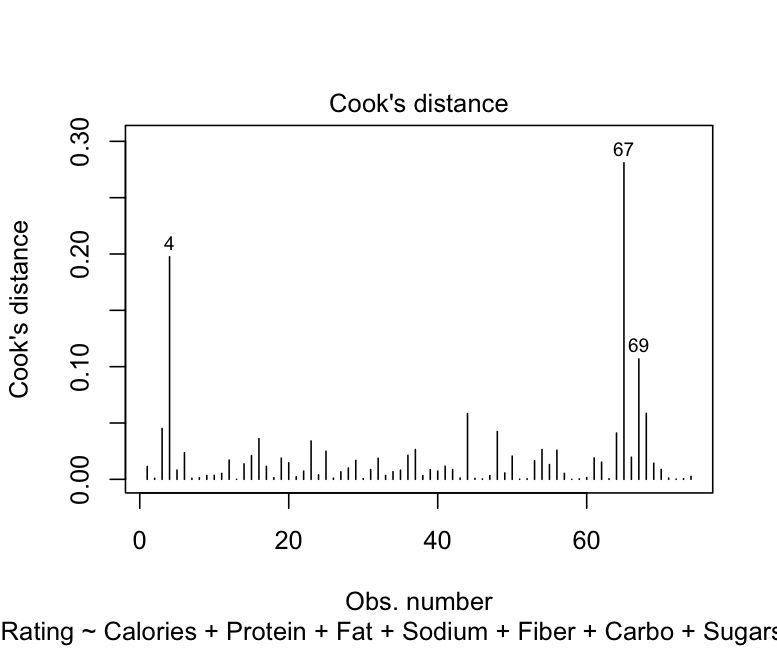
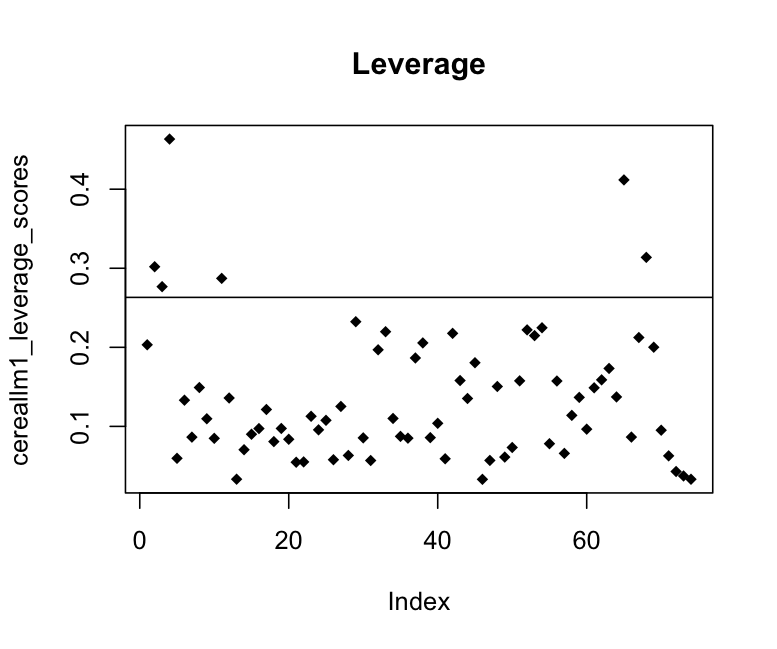
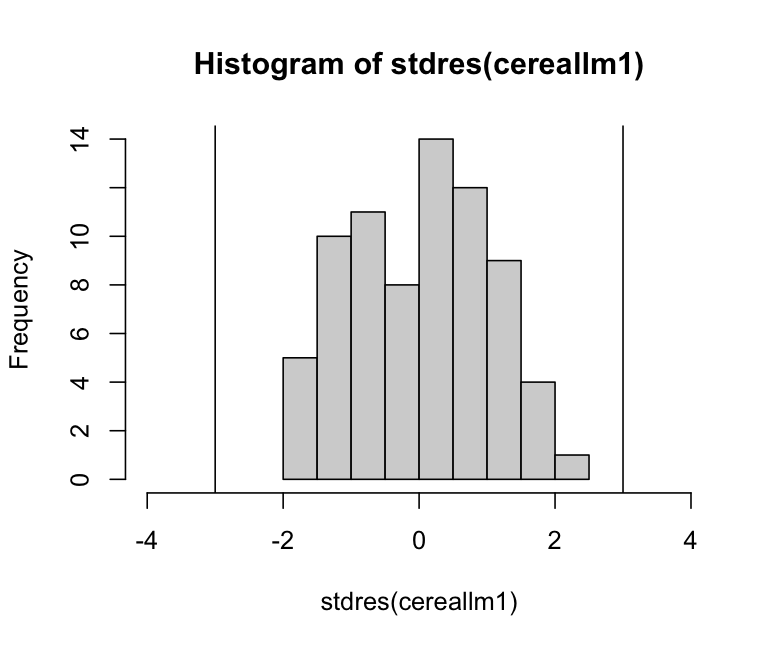
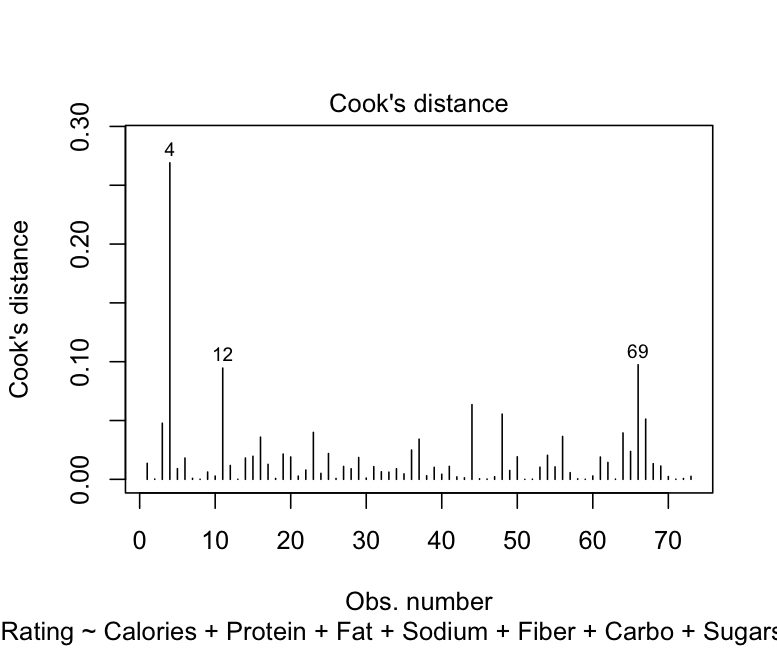
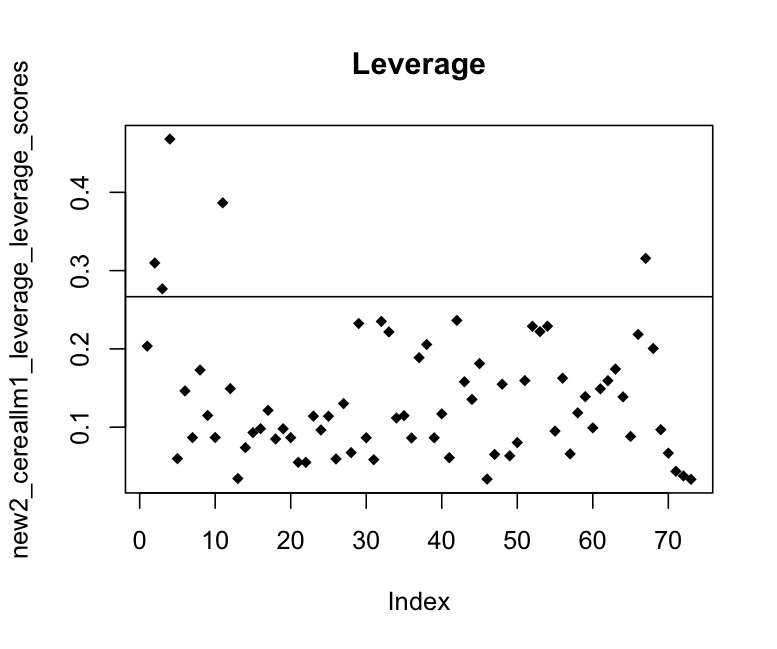
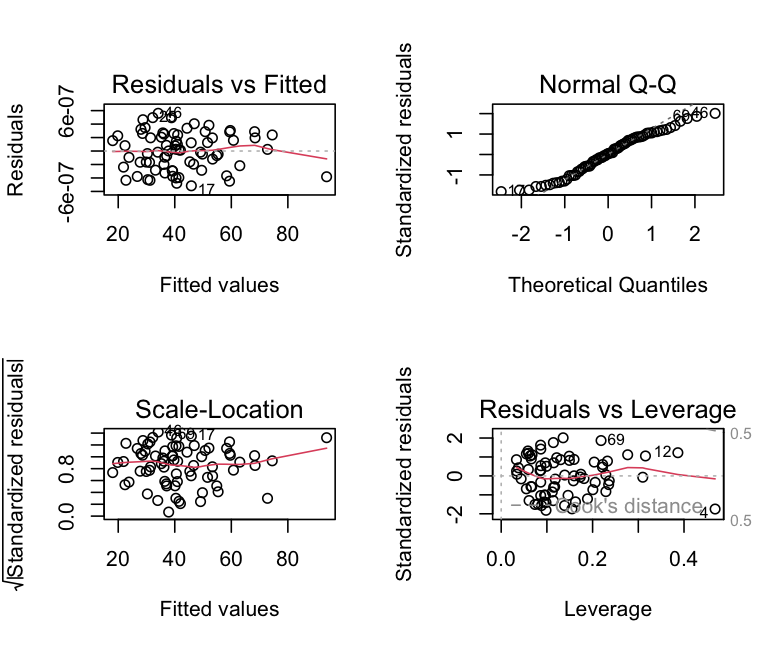
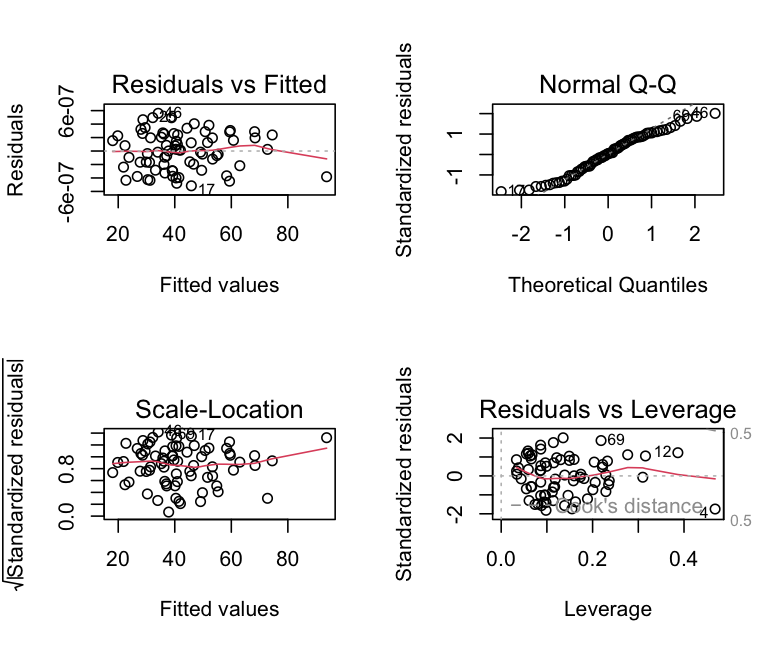
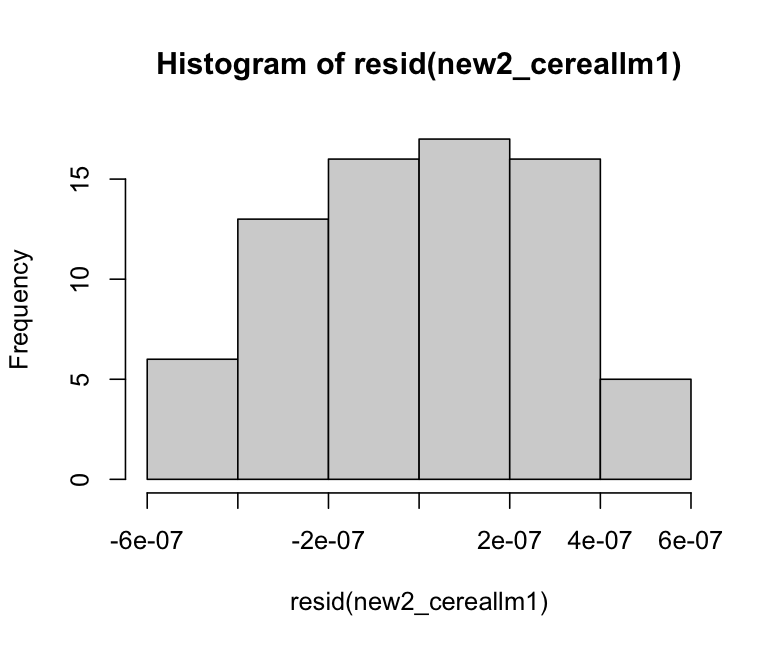
1. Based upon part a) determine a multiple regression model that you could fit to explain the variable Rating. Fit this model. Make any modifications to this model that you believe are appropriate.

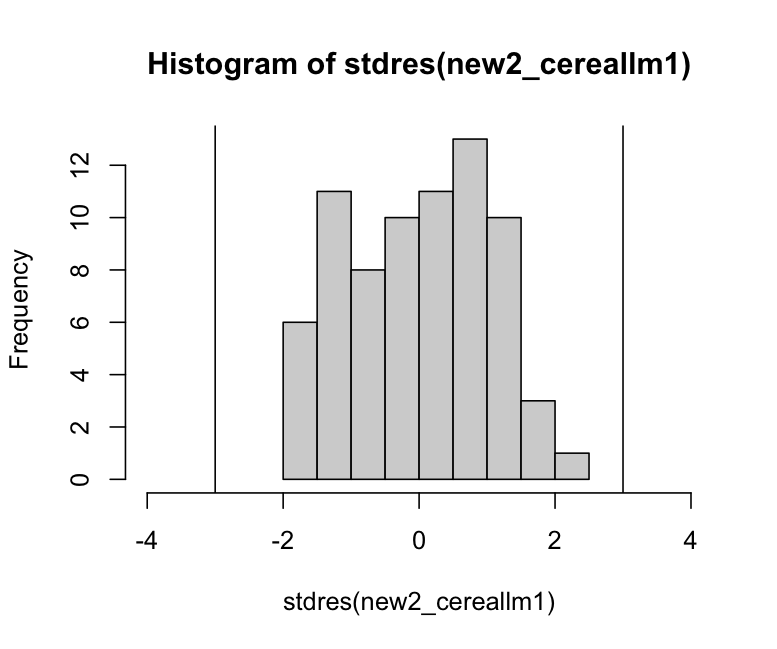
* Selecting a model:
  + In a saturated model, all variables except Name, Manufacturer, and Type were significant predictors.
  + Additionally, using BIC, we can see that a model with 9 predictors is likely the best.
  + Using best subset to choose predictor variables, we created a model.
* Original model:
  + Rating = 54.93 - 0.22(Calories) + 3.27(Protein) - 1.691(Fat) - .054(Sodium) + 3.44(Fiber) + 1.09(Carbo) - 0.72(Sugars) - 0.03(Potass) - .05(Vitamins)
  + The model returns an adj. r^2 value of 1, which seems almost too good.

Using those nine variables, we also did Ridge and Lasso Regression to deal with any multicollinearity. We intentionally left out Manuf and Type. The results are fairly similar to our original model.

* Ridge:
  + Rating = 58.45 - 0.16(Calories) + 2.65(Protein) - 2.60(Fat) - .048(Sodium) + 2.01(Fiber) + .58(Carbo) - 1.11(Sugars) - 0.009(Potass) - .047(Vitamins)
* Lasso:
  + Rating = 56.40 - 0.17(Calories) + 2.85(Protein) - 2.29(Fat) - .053(Sodium) + 2.93(Fiber) + .78(Carbo) - 1.01(Sugars) - 0.02(Potass) - .05(Vitamins)
    - If we had run Lasso Regression using all potential predictor variables, it would have removed Manuf but kept Type in the model.

1. 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.

* Using our original model.
  + There is no issue with non-constant variance, but there is with normality amongst residuals (shapiro.test = .03).
* Check for outliers, high leverage, and Cook’s Distance
  + No standardized residuals > 3, a few high leverage points, no CD’s above .5. 
  + From the plots, we can see that Point 4 and Point 67 are the most influential points. Ultimately, we decided to remove Point 67 since it has the highest Cook’s Distance.
* Point 67: Special\_K
* Fit the model again without Point 67 and run diagnostics:
  + No issues with non-constant variance and not as major an issue with normality of residuals (Shapiro-Wilks test p-value = .06).



* + We could remove Point 4 to potentially create a better model. However, we did not think there was sufficient evidence (CD < .5) to warrant removing another point in an already small dataset. There are also multiple high-leverage points that could be further explored.

1. 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)?

Using LOOCV, the RMSE for the model in Part C is 3.379842e-07. After removing Point 67, the RMSE is 3.269119e-07, suggesting a better model. Additionally, the RMSE after removing both Points 4 and 67 is 3.165115e-07, which means we might want to reconsider removing Point 4 as well.