Exam 2

* 1. Two of the VIFs are >10, which suggests multicollinearity. There are several ways to address multicollinearity. One is to collect more data, another is to respecify the model (perhaps through variable elimination), and another is to perform ridge regression to introduce bias into the model.
  2. Because smaller values of Cp are generally desirable, especially when Cp is less than p+1, the best model appears to be that which includes rbi, contract, ko, and err. Because Cp=3.1181 is less than p+1=5, this estimate is considered unbiased. The question mentions performing stepwise selection on all 13 potential predictor variables, but the table shown only shows 4 of the variables and the Cp value from adding each of them to the model. Typically, stepwise selection would be performed with a value of Cp found for each potential predictor variable, each combination of 2, each combination of 3, etc.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source of Variation | Sum of Square | Degree of Freedom | Mean Square | F | p-value |
| Regression | 5550.8166 | 2 | 2775.4083 | 261.2419 | 4.6862e-16 |
| Residual | 233.726 | 22 | 10.6239 |  |  |
| Total | 5784.5426 | 24 |  |  |  |

Because p<0.05, we reject the null hypothesis; there is enough evidence to conclude that at least one of the regressors contributes significantly to the model.

* 1. The dimensions of the matrix would be 2 x 2. The top left element would be n. The rest of the elements would be n2.
  2. The dimensions of the matrix would be 2 x 1. The top element of the matrix would be the sum of all the elements in the **Y** matrix. The bottom element of the matrix would be the sum of the y elements that were successes in the **Y** matrix.
  3. The dimensions of this matrix would be 2 x 1. The top element of the matrix would be . The bottom element of the matrix would be .

Since :

Since **X** is invertible:

* 1. (calculated in attached R file)

β0 refers to the clarity of the lake in ecoregion A. 90% confidence interval: (36.978, 43.822)

* 1. (calculated in attached R file)

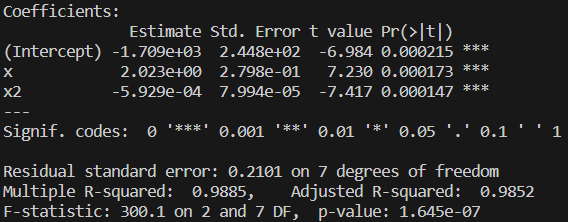
p-value=4.857e-07

Since p<0.05, we reject the null hypothesis; there is enough evidence to conclude that β1, β2, or both does not equal 0.

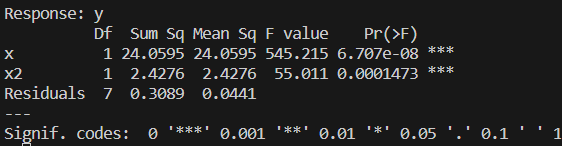
* 1. (calculated in attached R file)

p-value=9.584e-06

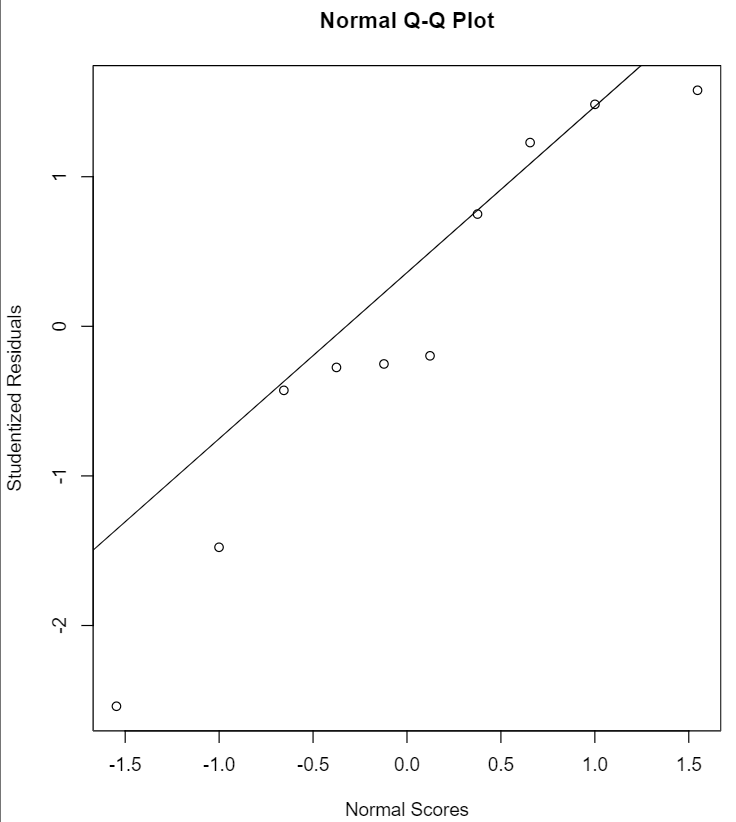
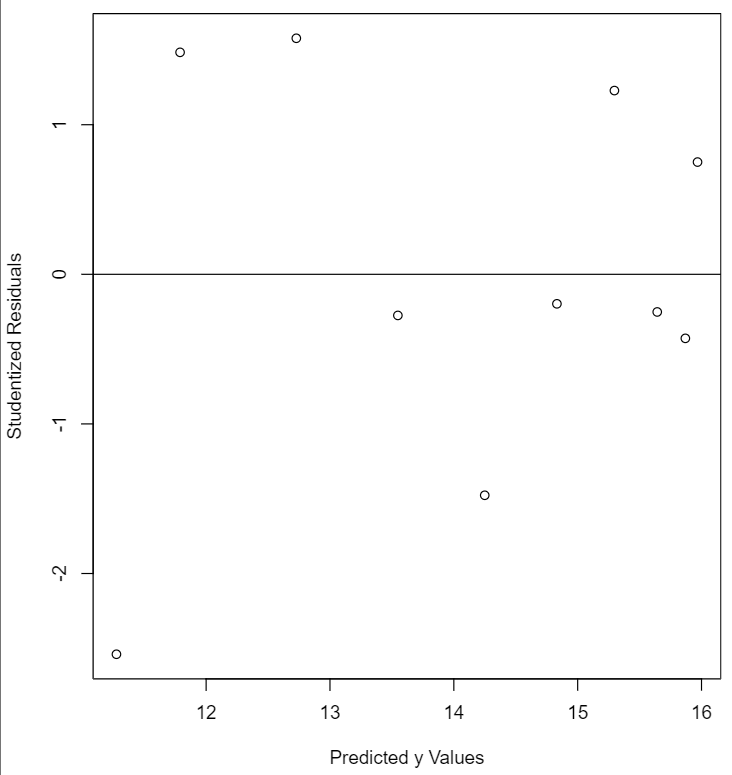
Since p<0.05, we reject the null hypothesis; there is enough evidence to conclude that .

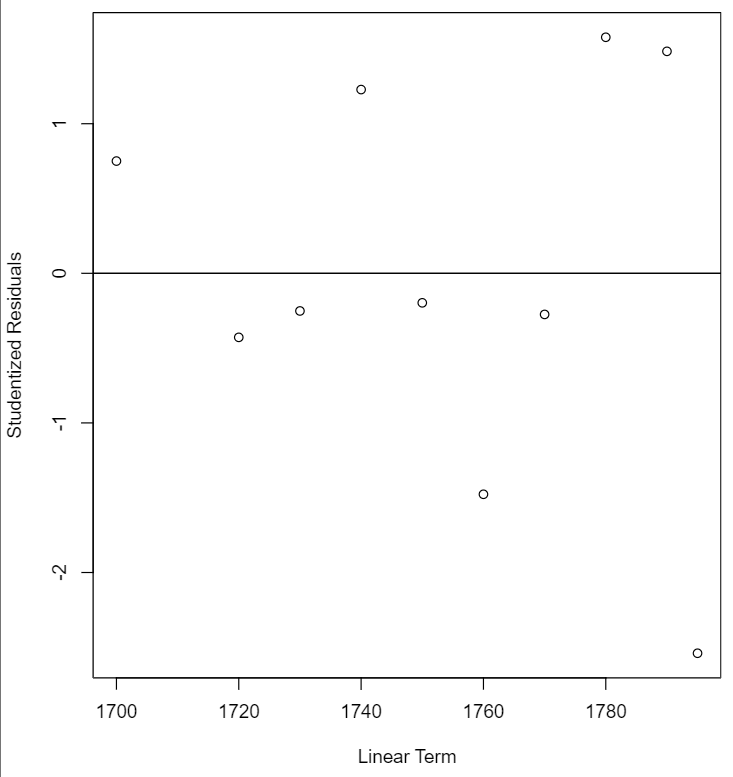
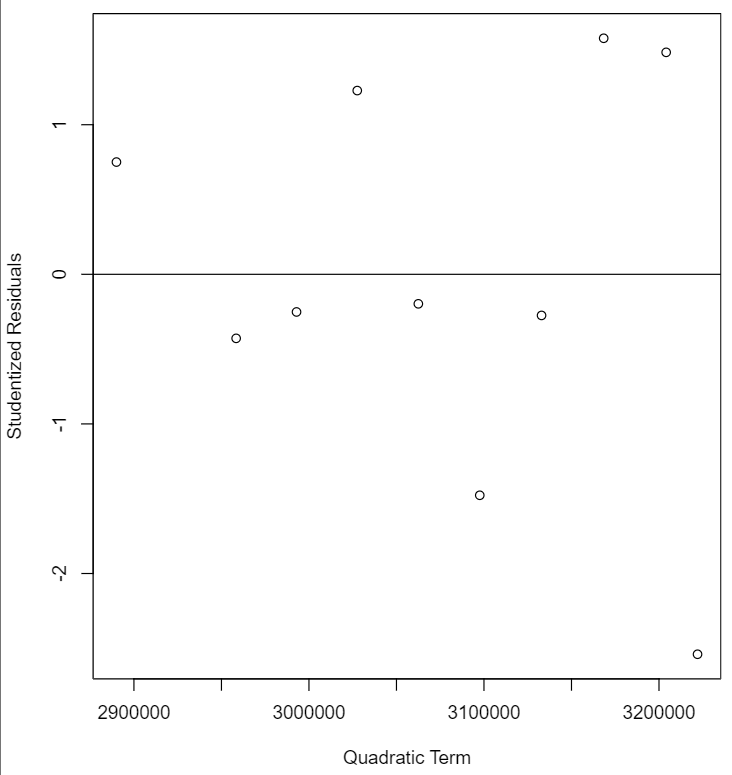
* 1. 

The p-value of the F-statistic is <0.05, so we reject the null hypothesis. We have enough evidence to conclude that at least one of the regressors contributes significantly to the model.

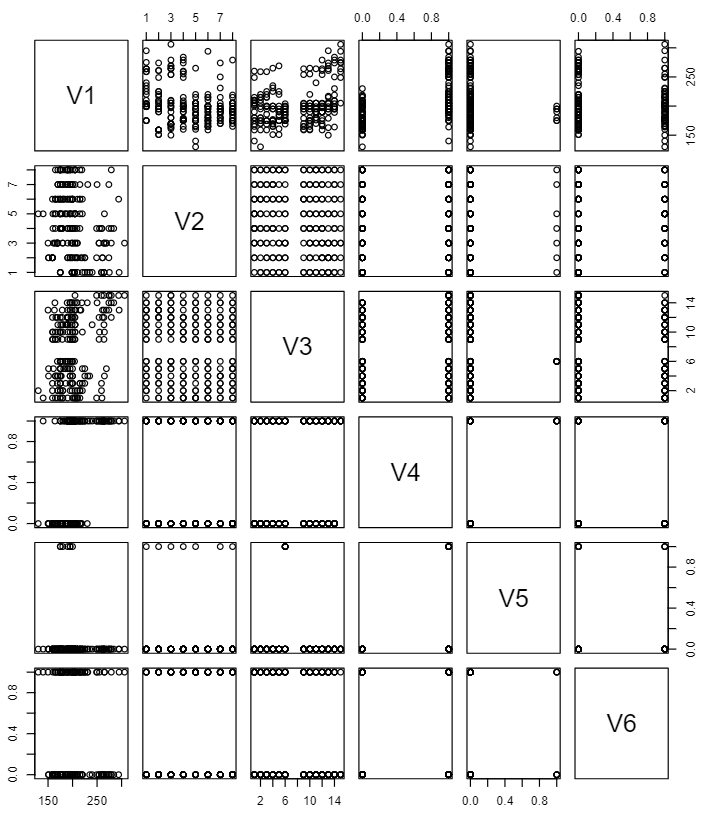
* 1. 

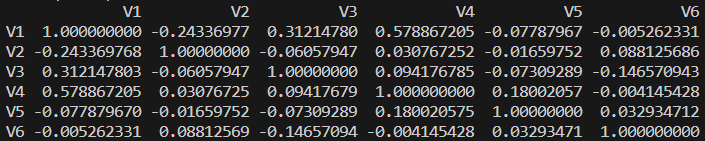
Both terms contribute significantly to the model. The linear term contributes more.

* 1.  

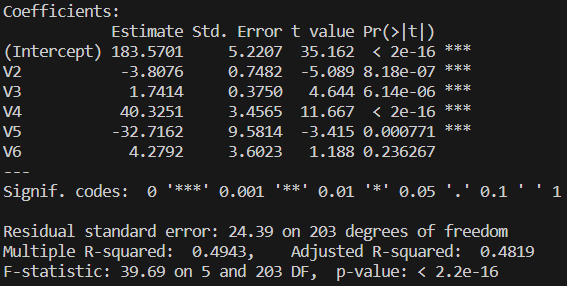
 

The model fit seems accurate!

* + 1. The relationship between the response y and the regressors is linear, at least approximately.
    2. The error term ε has 0 mean.
    3. The error term ε has constant variance σ2.
    4. The errors are uncorrelated.
    5. The errors are normally distributed.
  1. 



I don’t see any evidence of multicollinearity in these matrices.

* 1. 

Since the p-value for the F-statistic is <0.05, we reject the null hypothesis; at least one of these variables contributes significantly to the model. Therefore, I would call the model useful.

* 1. Based on the p-values for each t-test conducted on the regressors, floor, distance from elevator, ocean view, and end unit contribute significantly to the model.
  2. 

The VIFs do not suggest multicollinearity.

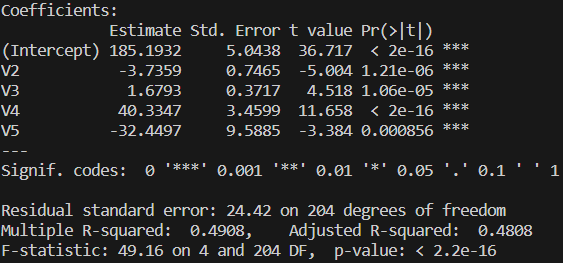


The eigenvalues do not suggest multicollinearity.



The kappa value does not suggest multicollinearity.

The low value of R2, the significant F-statistic p-value, and the fact that most of the regressors are found to be significant does not suggest multicollinearity.

* 1. 

By using the stepAIC function in R, I found that the best subset of the full model includes all variables except for furnished.

* 1. No new variables were introduced in the reduced model, so the matrix of scatter plots and the correlation matrix would show the same lack of multicollinearity.



The VIFs do not suggest multicollinearity



The eigenvalues do not suggest multicollinearity.



The kappa value does not suggest multicollinearity.

* 1. Outliers:



Influential observations:

11: dffits, dfbetas

37: dffits, dfbetas

79: hat

85: hat

92: hat

101: hat, dfbetas

114: hat, dfbetas

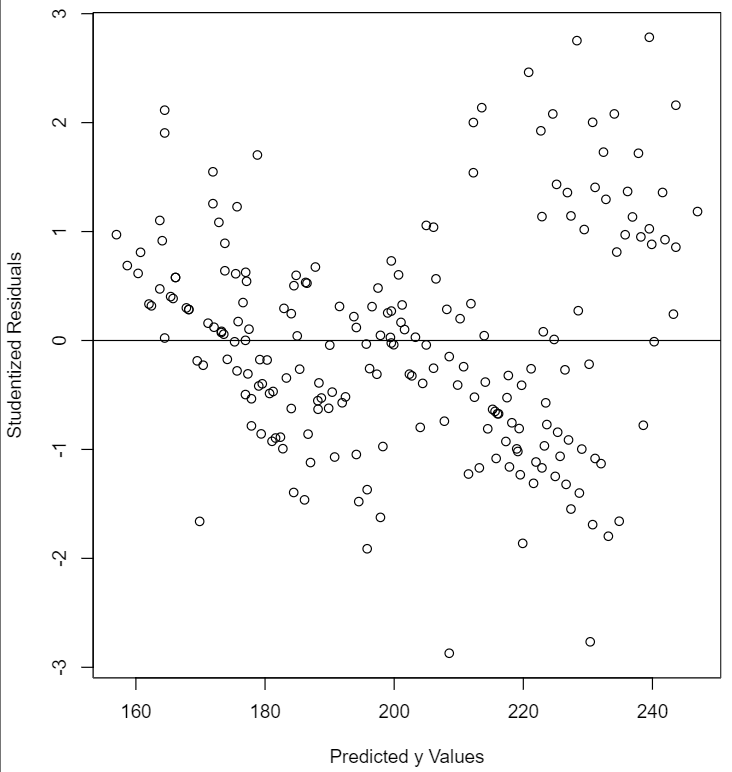
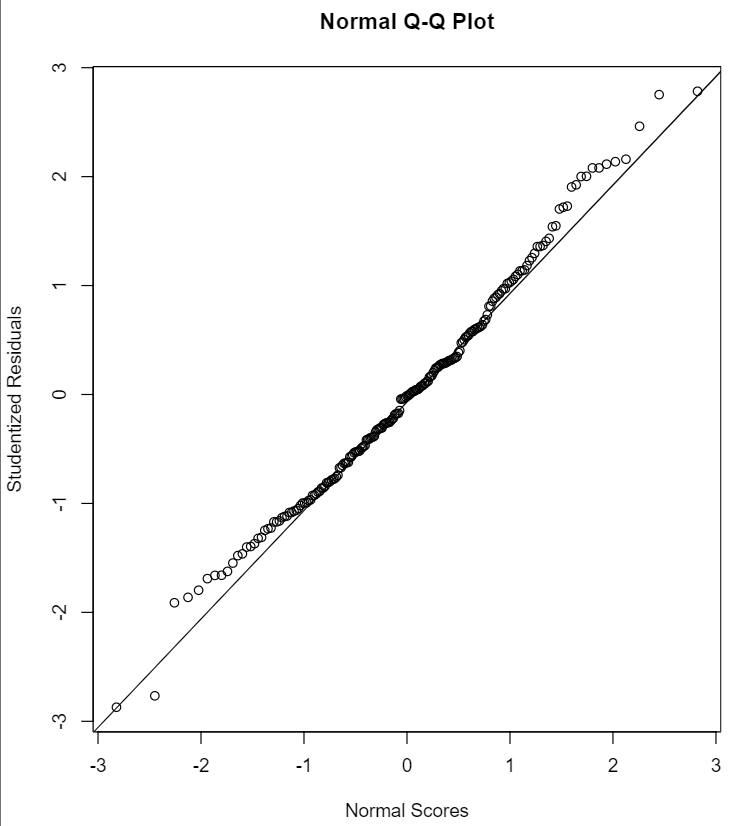
153: hat, dfbetas

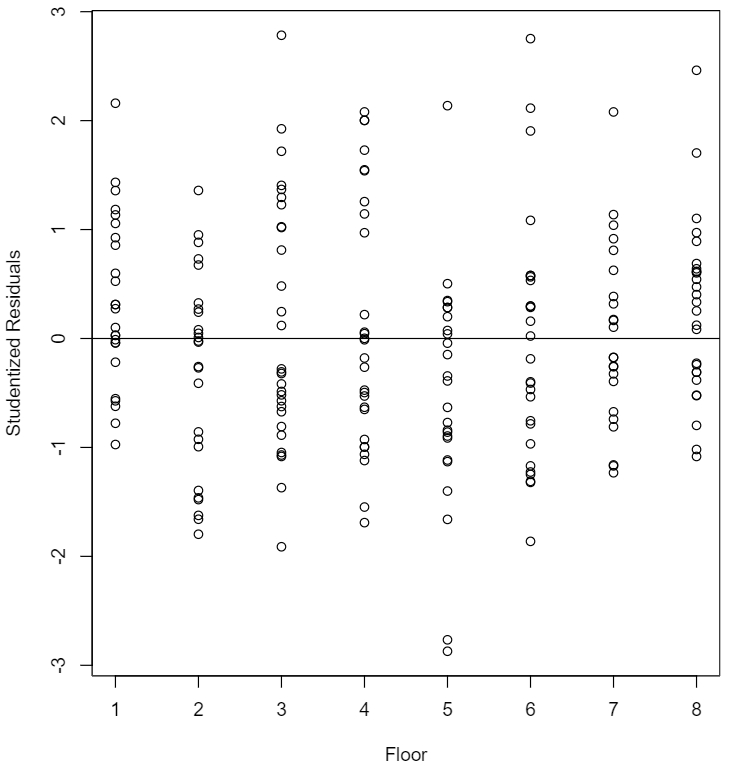
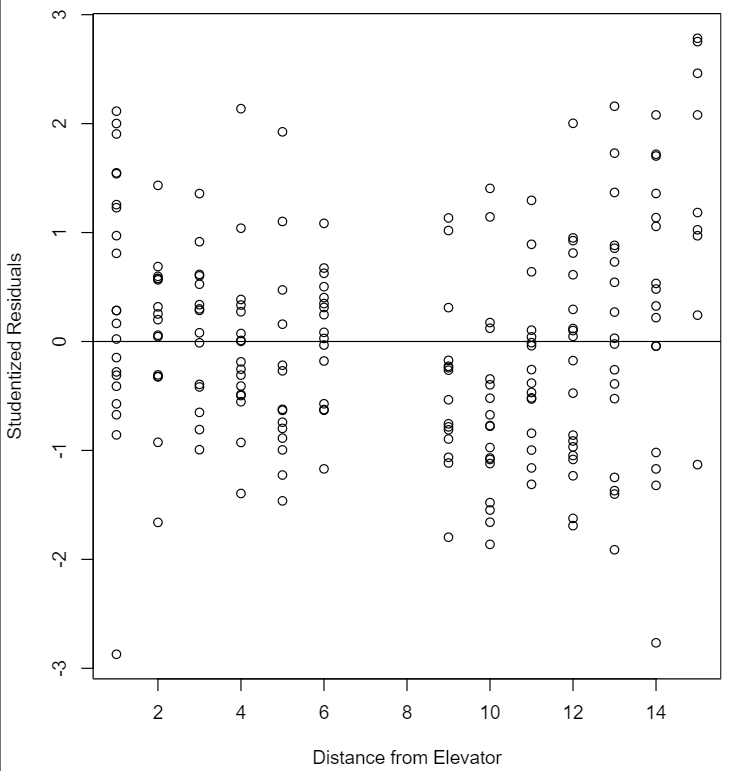
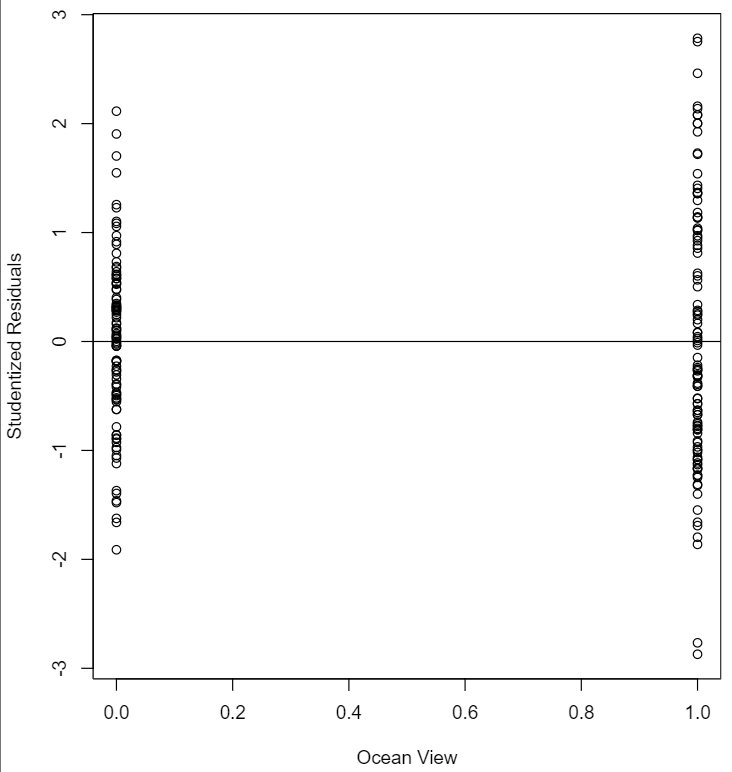
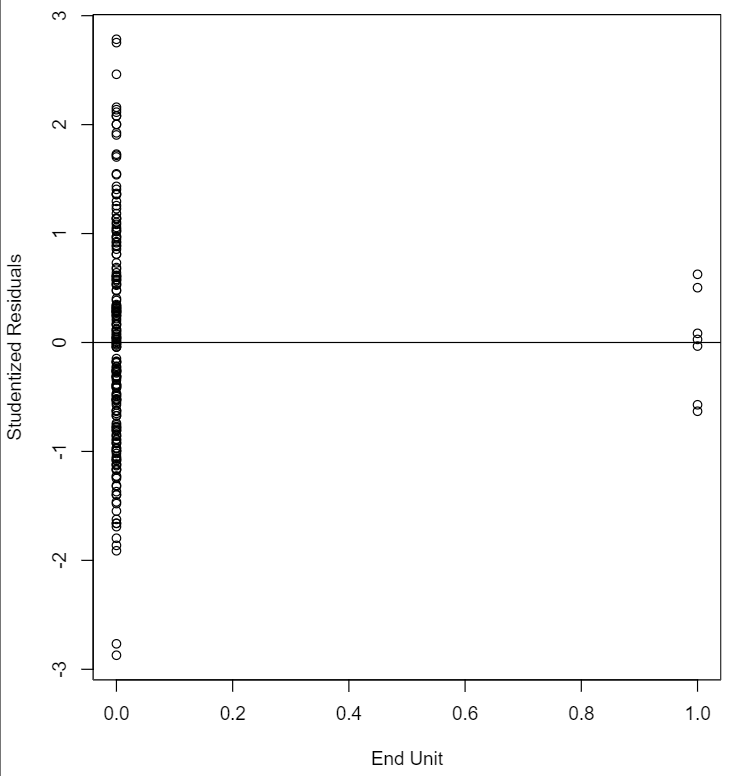
173: hat

184: dffits, dfbetas

188: dffits, dfbetas

208: dffits, dfbetas

* 1. 

The residuals are normally distributed with relatively constant variance (except noted below). None of these plots lead me to believe that the relationship between y and the regressors is (no bows). The residuals are also all centered around 0.

* 1. Perhaps the plot of the studentized residuals against the end unit variable could be considered funnel shaped, indicating nonconstant variance. Per the textbook, this could be addressed by applying a suitable transformation to either the regressor of the response variable. However, I think the true problem is an imbalance in data. There are far more non-end units than end units. I think gathering more data on end units would be the solution here.