DS6371_kgPrjClean

Introduction

The purpose of this exercise is to use linear regression techniques on a dataset pertaining to the real estate market in Ames, Iowa. There are two sections to this project, the first will outline using multiple linear regression, and the second will use more advanced model selection techniques.

Data Description

This dataset was obtained through a kaggle competition with the same general purpose as what is stated in the introduction. Develop the best linear regression techniques to predict housing prices. More info can be found on kaggle's website.

Analysis Question 1:

Restatement of Problem

Century 21 Ames would like an estimate of how the SalePrice of the house is related to the square footage of the living area of the house and if the SalesPrice depends on which neighborhood the house is located in.

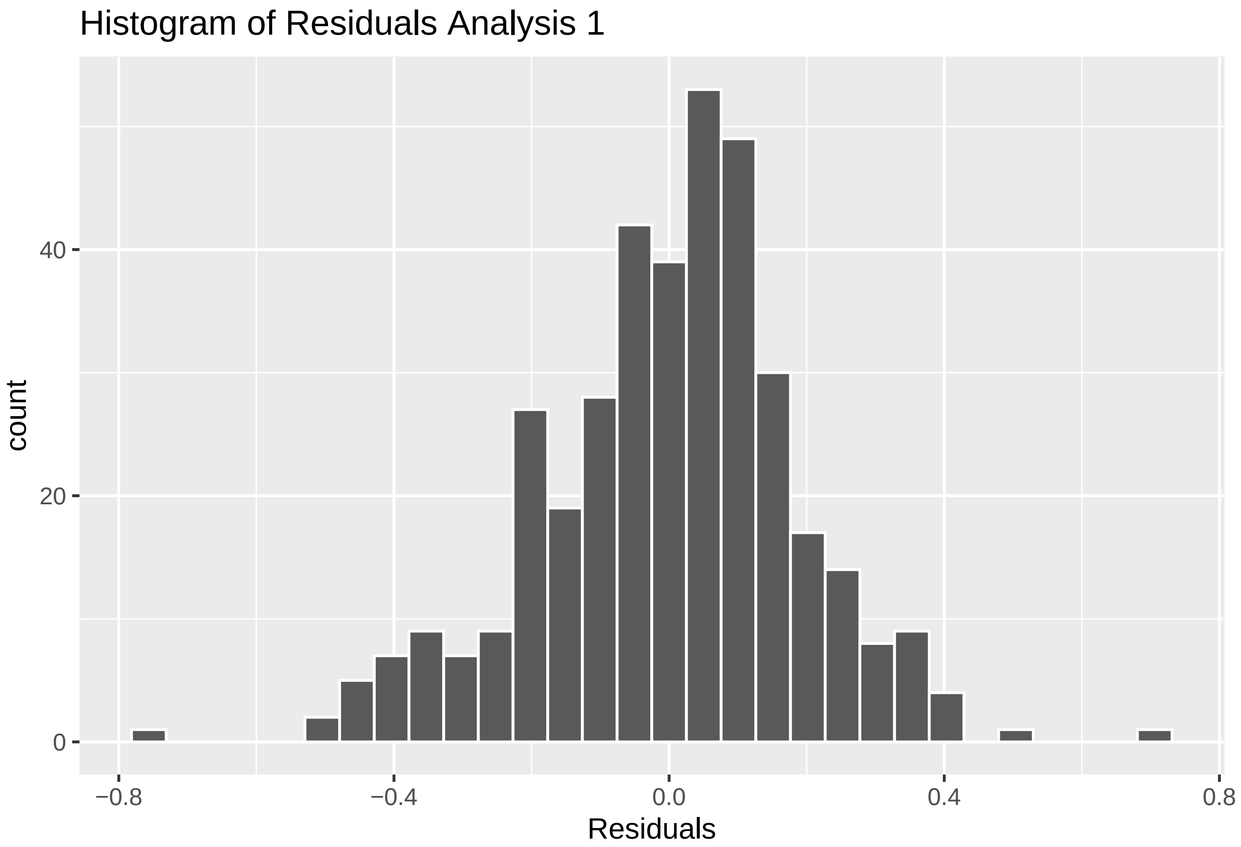
Build and Fit the Model

From initial models, it was found that from these 3 neighborhoods in Century 21 Ames district, the only neighborhood that was statistically significant was North Ames after the outliers were accounted for.

log(salePriceEstimate) = β_0 + β_1 log(GrLIvArea) + β_2 NAmes + β_3 log(GrLIvArea) * NAmes

Checking Assumptions

Normality: Judging from scatter plot, q-q plot, and histogram of residuals there is not strong enough evidence against normality in the log transformed data.



Here we can see the histogram plotted from the residuals that they are very normally distributed. See the appendix for the scatter plots of the data, as well as the qq plot for the residuals.

Linear Trend:

The transformed data appears to have a linear trend, this assumption seems to be satisfied. Please see the appendix for the scatterplots of the data.

Equal SD:

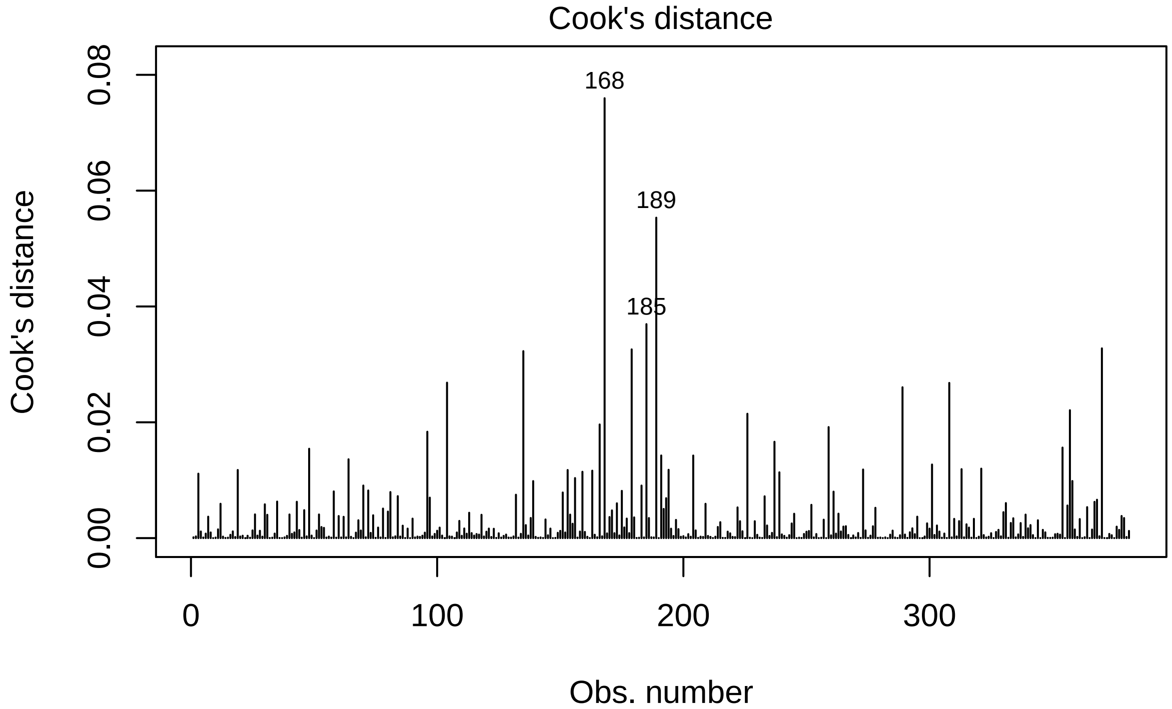
This is assumption is satisfied, though it seems that smaller values of x have somewhat of a right skew for values of y in the NAmes neighborhood. We will proceed with this in mind, but this assumption seems like it is satified.

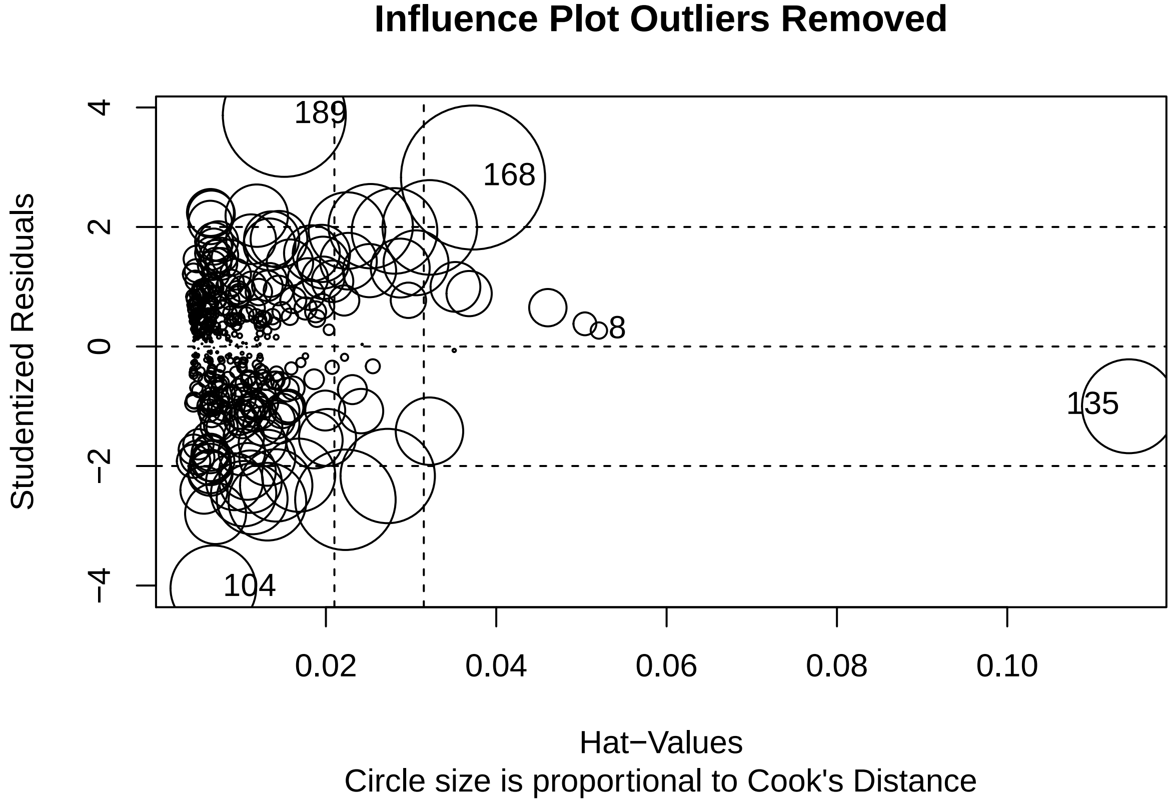
Independence:

The data is assumed to be independent.

Influential point analysis (Cook's D and Leverage)

These plots show the entire dataset for the neighborhoods of interest for Century 21 Ames. There is a point that has very high leverage, and it can be considered to be an outlier. These points were examined to be partially built houses, which were dropped from certain analysis, and the model was run with both datasets for comparison. After removing the partially built houses from the dataset, the leverage and influence for each point is far more balanced.





```
##       StudRes         Hat        CookD
## 8    0.266763 0.052050021 0.0009792602
## 104 -4.046479 0.006778209 0.0268413527
## 135 -1.000419 0.114289929 0.0322863846
## 168  2.827397 0.037273151 0.0759667333
## 189  3.867570 0.015112686 0.0553328038
```

Comparing Competing Models

Adj R2

lm with partial houses - adj.r.squared: 0.4931

lm without partial houses - adj.r.squared: 0.5206

From the 2 adjusted rˆ2 we can see that the model without the large partial builds performs slightly better.

Internal CV Press

lm with partial houses - PRESS: 14.8988

lm without partial houses - PRESS: 13.86

From the 2 PRESS stats, we can see that the model without the large partial builds performs slightly better again.

Parameters

Estimates:

β_0 = 8.4927276 β_1 = 0.4730236 β_2 = -2.0422171 β_3 = 0.2680962

Interpretation:

Both logGrLIvArea and NorthAmes/NotNorthAmes were found to be significant in this model. This indicates that each doubling of the GrLIvArea, holding all other variables constant, results in a 2ˆ(0.473)= 1.39 multiplicative change in the median of the SalePrice. That translates to a 39% increase in the median SalePrice with every doubling of GrLIvArea.

Confidence Intervals:

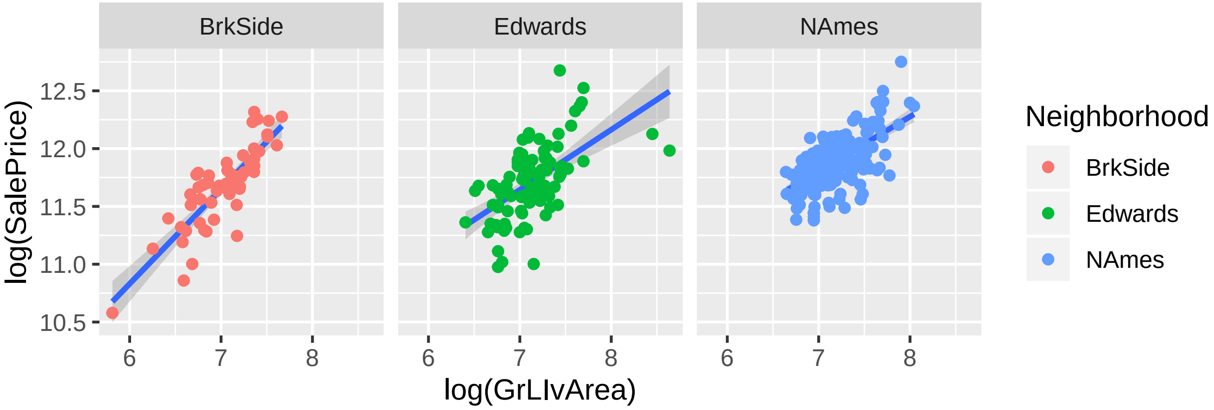
Conclusion:

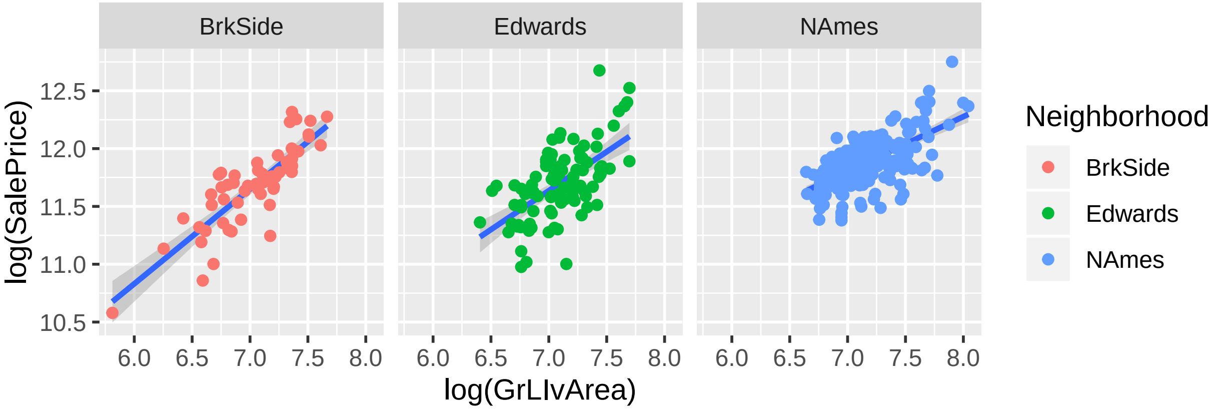
The influence of square footage of a house has a bearing on the final sale price from these analysis, we have found this to be statistically significant(p-value $<$ 2.2e-16). The neighborhood in which the house resides is also a factor in deciding the final sale price, and it was found that the North Ames part of the sales district to be associate with higher predicted sale price relative to the median GrLIvArea.

APPENDIX

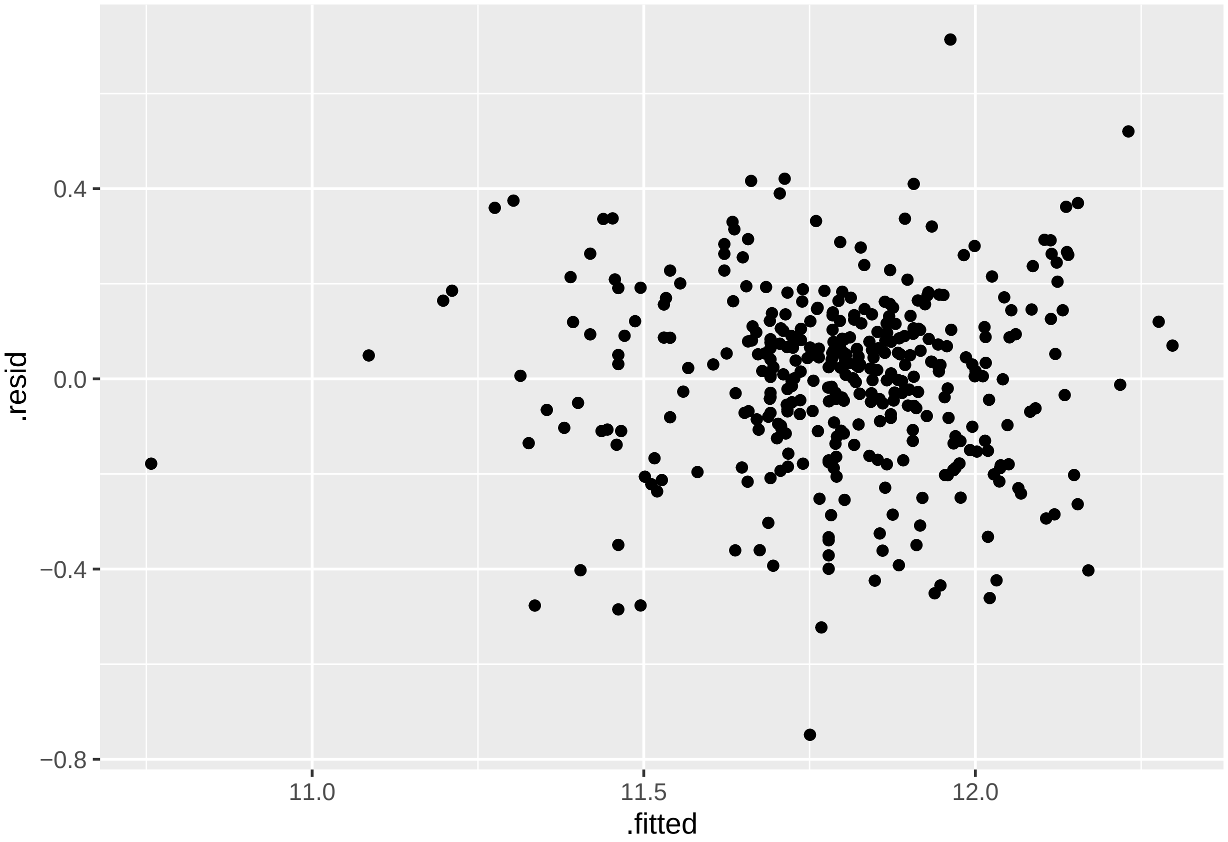
```
## Scatter plot with only two neighborhood variables

aOne_scat_plot <-
  aOne_df_noLargePartials %>%
  ggplot() +
  geom_smooth(
    method = lm,
    mapping =
      aes(
        x = aOne_df_noLargePartials$logGrLivArea,
        y = aOne_df_noLargePartials$logSalePrice,
        color=BlendNgbrhd
      )
  ) +
  geom_point(
    aes(
      x=logGrLivArea,
      y=logSalePrice,
      color=BlendNgbrhd
    )
  ) +
  xlab("log(GrLIvArea)") +
  ylab("log(SalePrice)")
```

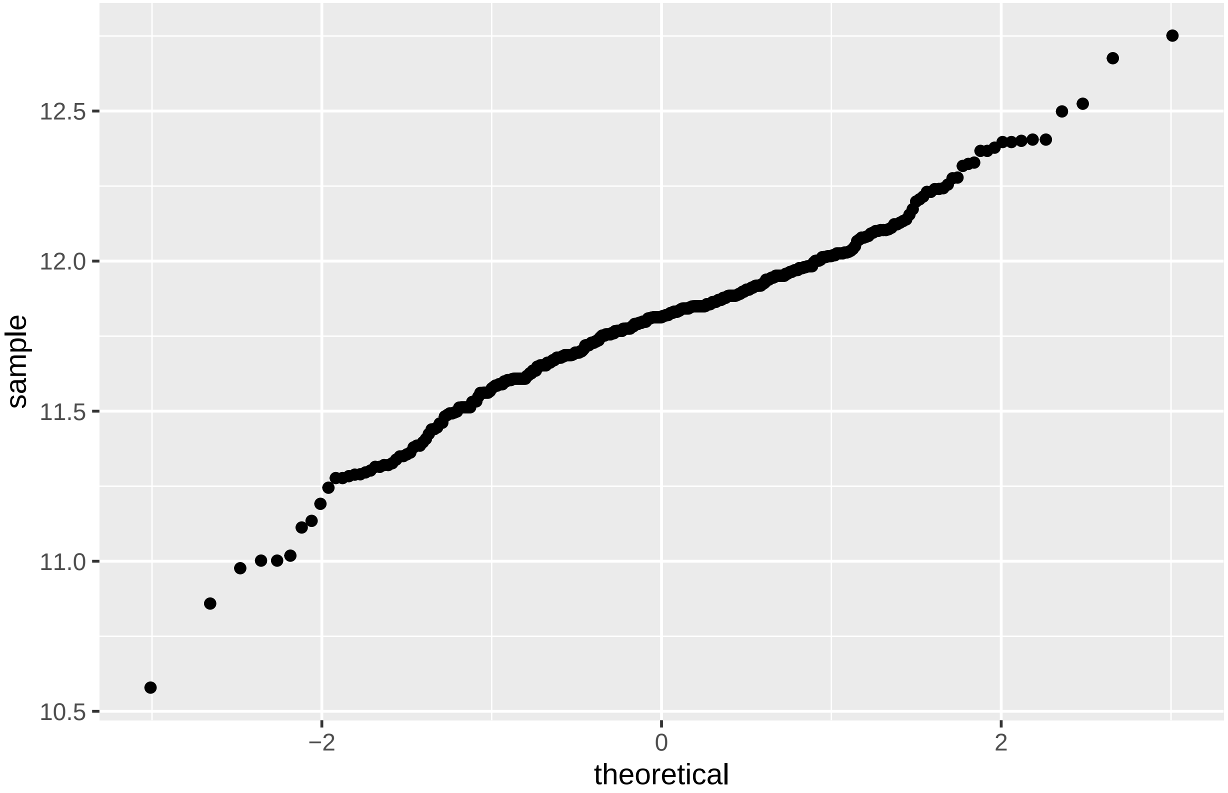


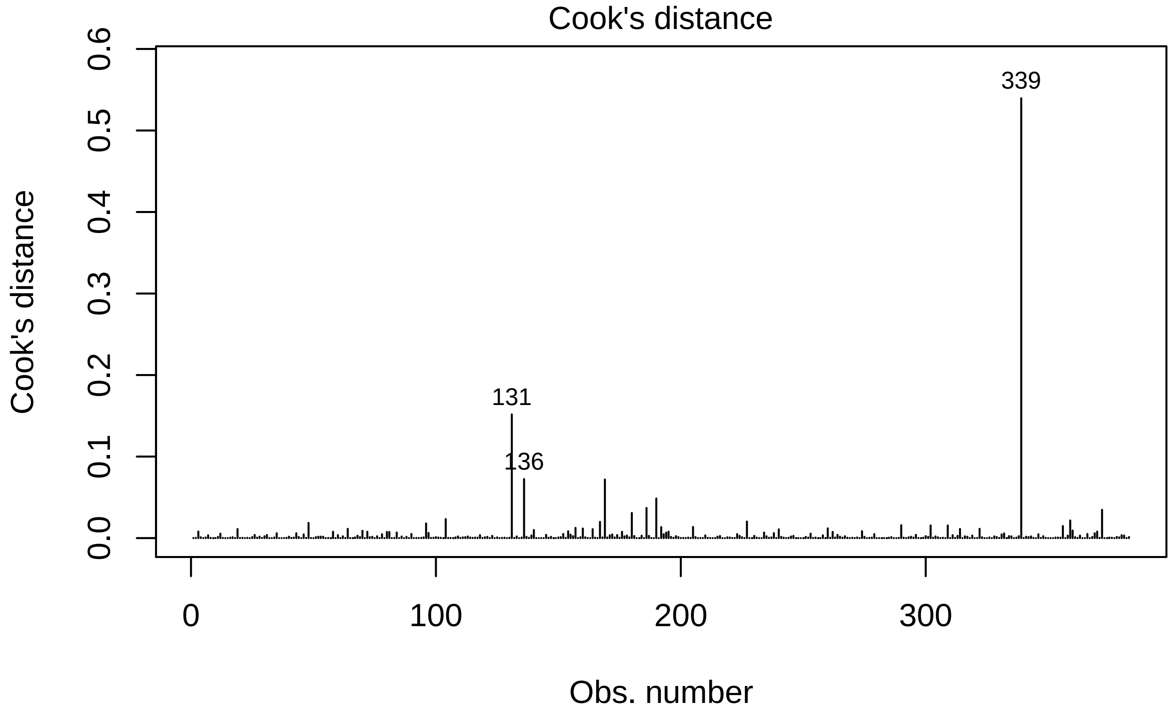


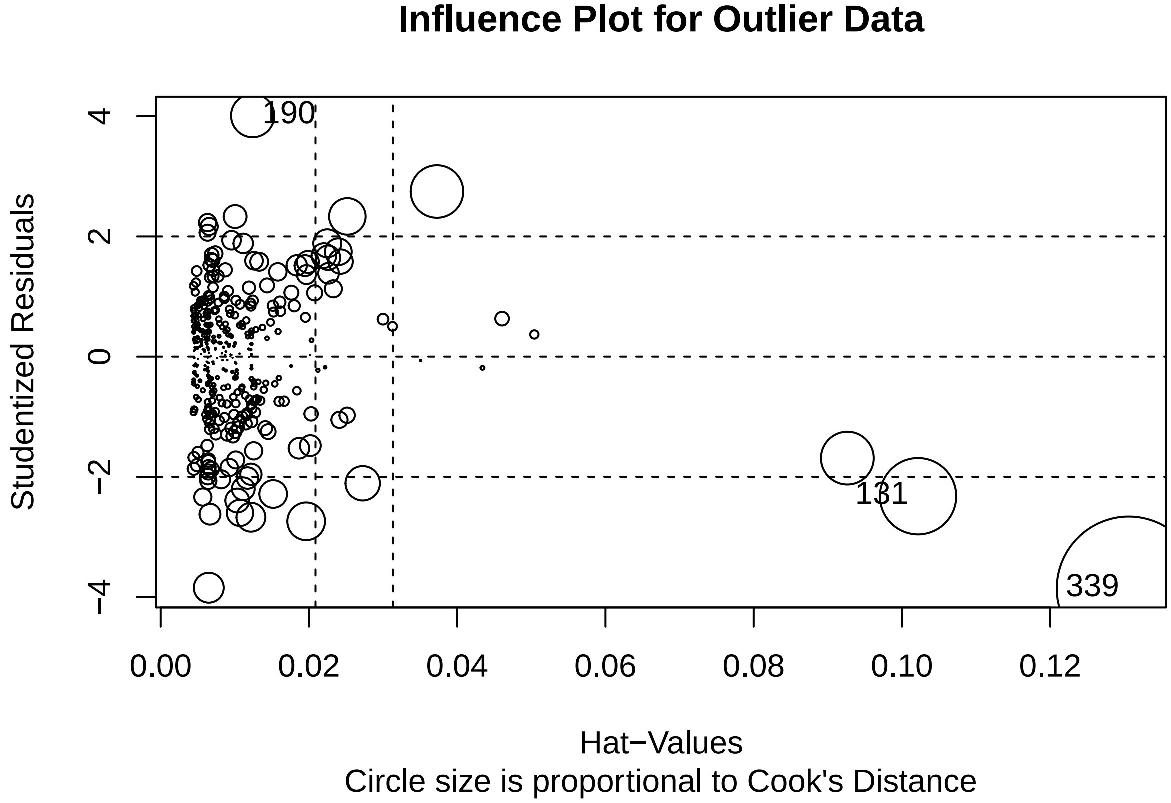
Residual Plots



Residual QQ Plot Analysis 1







```
##       StudRes        Hat      CookD
## 131 -2.322948 0.10217513 0.15176235
## 190  4.010642 0.01241146 0.04860308
## 339 -3.859488 0.13060358 0.53963326
```