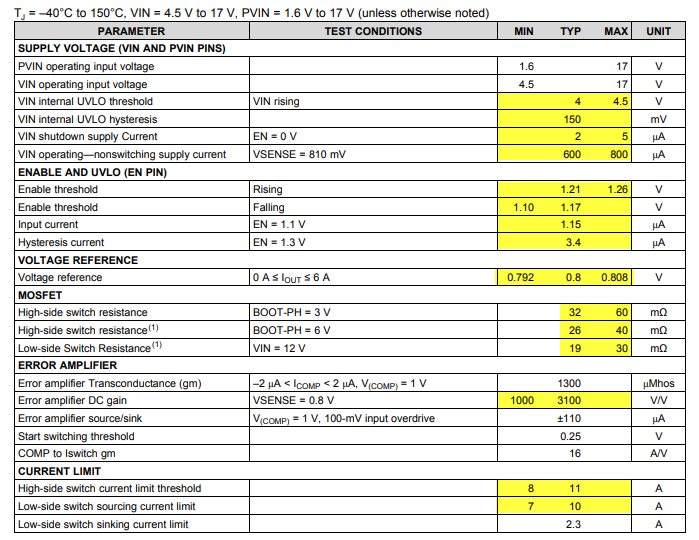
**Integrated Circuits Manufacturing Process Model**

**Joanna Duran, Nikhil Gupta, Max Moro**

**Introduction**

Semiconductor manufacturing is a variable process and outcomes depend on several factors. In order to meet target specifications, some parameters are controlled by engineers. However, some parameters are beyond human control. The output variables are typically measured after the manufacturing process and at times can lead to issues if the outputs are outside target specifications. Due to time, resource and cost constraints not all values can be practically measured.

Below is an example specification sheet. Each column is a single output and their respective values but not all values are populated because they cannot be practically measured.

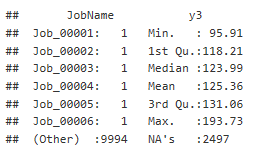
Like in most manufacturing environments, the question is “Can we predict the limits before the integrated circuits are manufactured to preemptively make changes when specs are expected to be out of range?” The answer is yes. Current practice is to use electrical simulation (+ running Monte Carlo simulations). However, this is very resource and time intensive as each electrical simulation can take several hours.

Our objective is to build a model to predict the performance (min, typ, max) of an output variable. Data has been sponsored and approved for use by Texas Instruments Inc who provided two files. The data were collected randomly from Jan1, 2019 to Jan2,2019. Due to proprietary information, we cannot disclose the variables measured or the department. The data consists of:

* Engineer Controlled variables (x1 – x23). Values differ, some are between 1 to 100 while others are in Nano or micro range.
* Process variation variables ( stat1 – stat217). These parameters are beyond human control. They represent various statistical parameters whose values represent the sigma variation around the mean. Range is from -4 (sigma) to 4 (sigma)
* Output Variables (y1 - y19) which represent various output variables

We will intent to pick at least 1 output variable and build a model to predict the mean value and the statistical variation of the output with respect to the process. Our target accuracy of +/- 10% is desired, but +/- 15% would be acceptable.

**Data Preparation** **and Descriptive statistics**

Like most data, it needed cleaning. We began by checking correlations to evaluate removal of redundant features. Basic descriptive statistics revealed how many NA values were present. 

We cleaned up labels and removed NA values. Once the data was cleaned, the datasets were merged. It was necessary to transform data.

**Analysis**

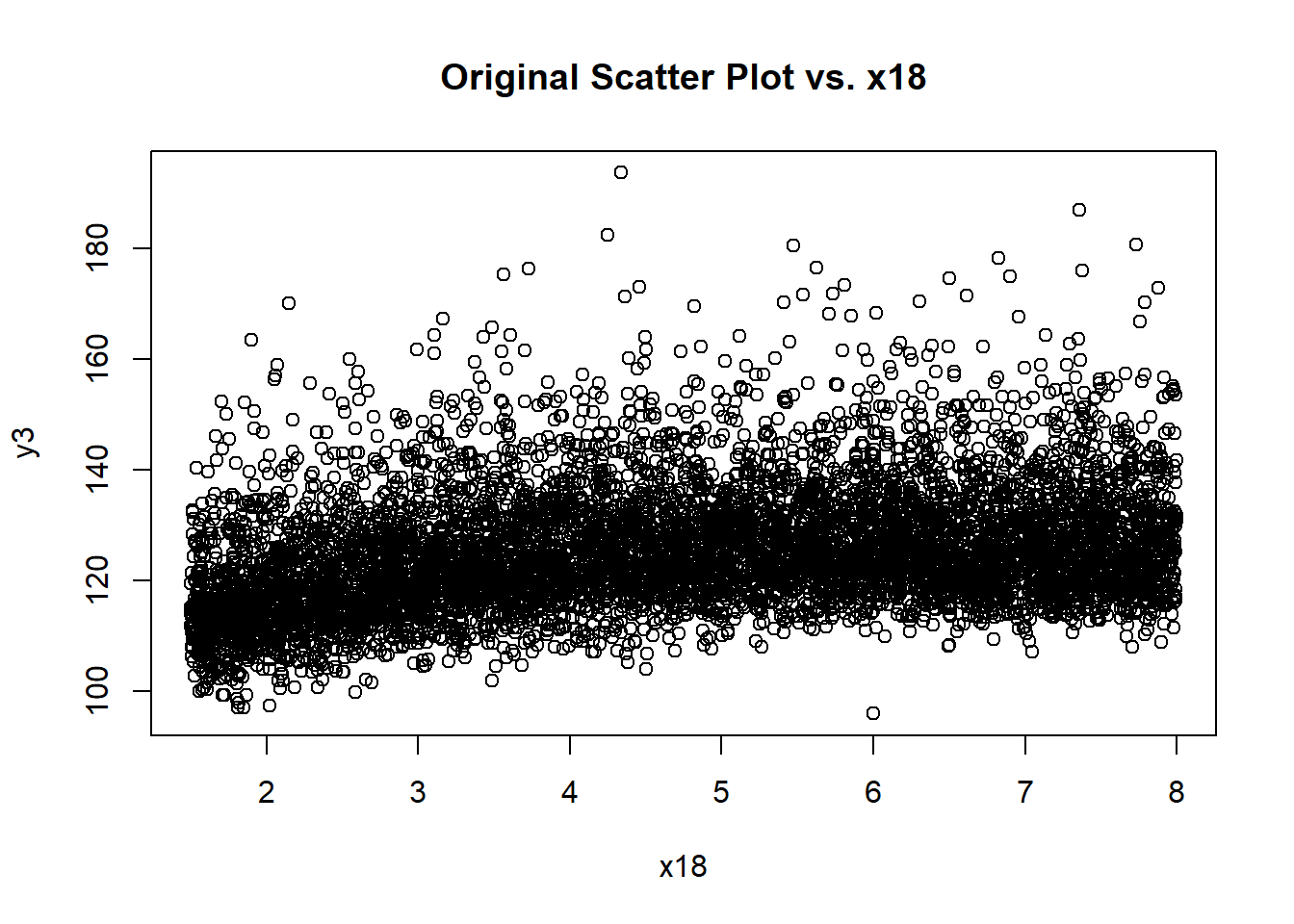
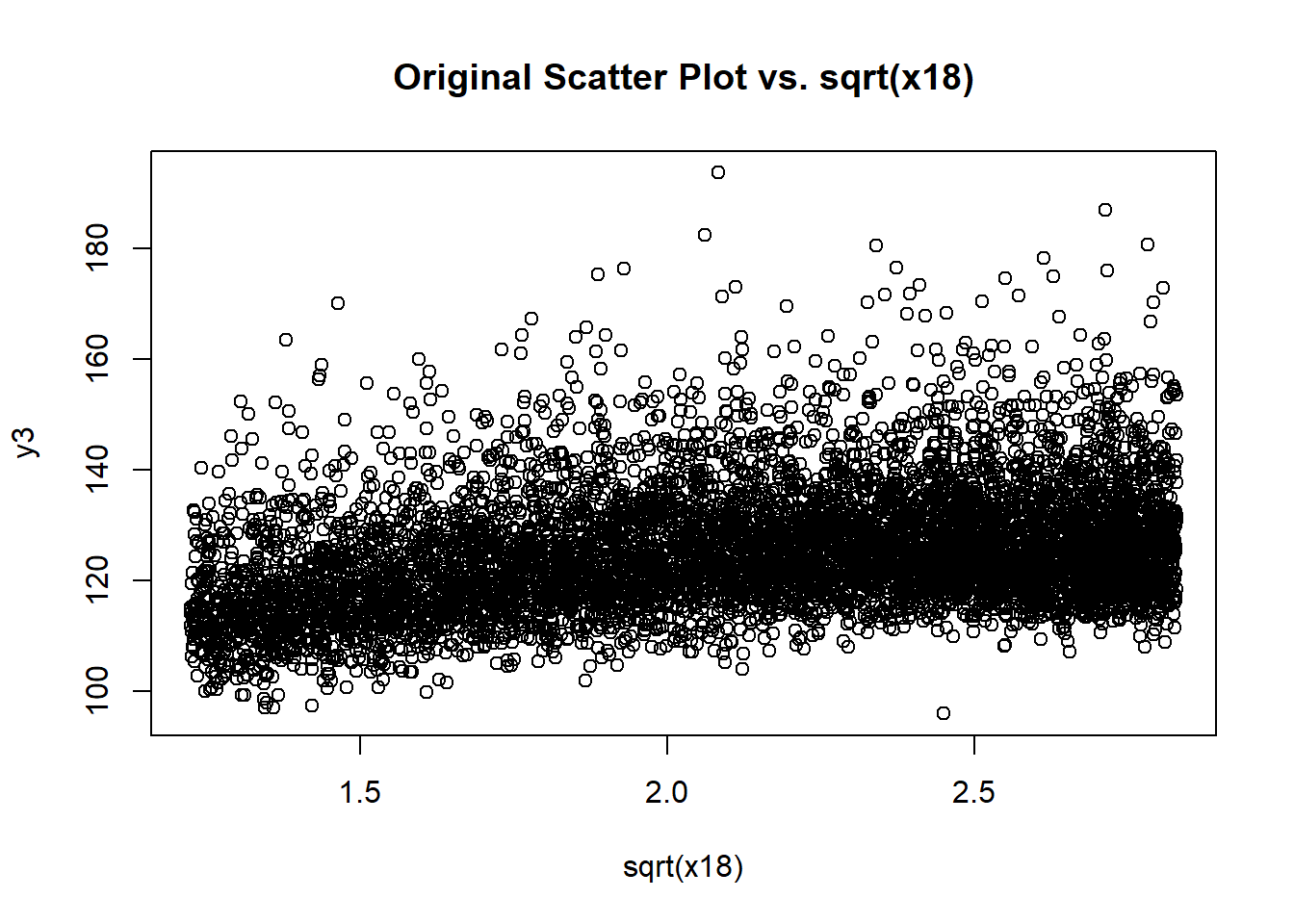
* Exploratory Data Analysis
* Check correlation of Label with Featires
* Multicollinearity - VIF
* Scatterplots
* Feature Engineering
* Modeling
  + Train Test Split
  + Common Functions
  + Setup Formulae
  + Full & Grand Means Model
  + Just trial removing all high influence points
  + Variable Selection
    - Forward Selection
    - Backward Elimination
    - Stepwise Selection
    - LASSO Selection
    - LASSO (w/ full train)
    - LASSO (w/ filtered train)
  + LASSO calls

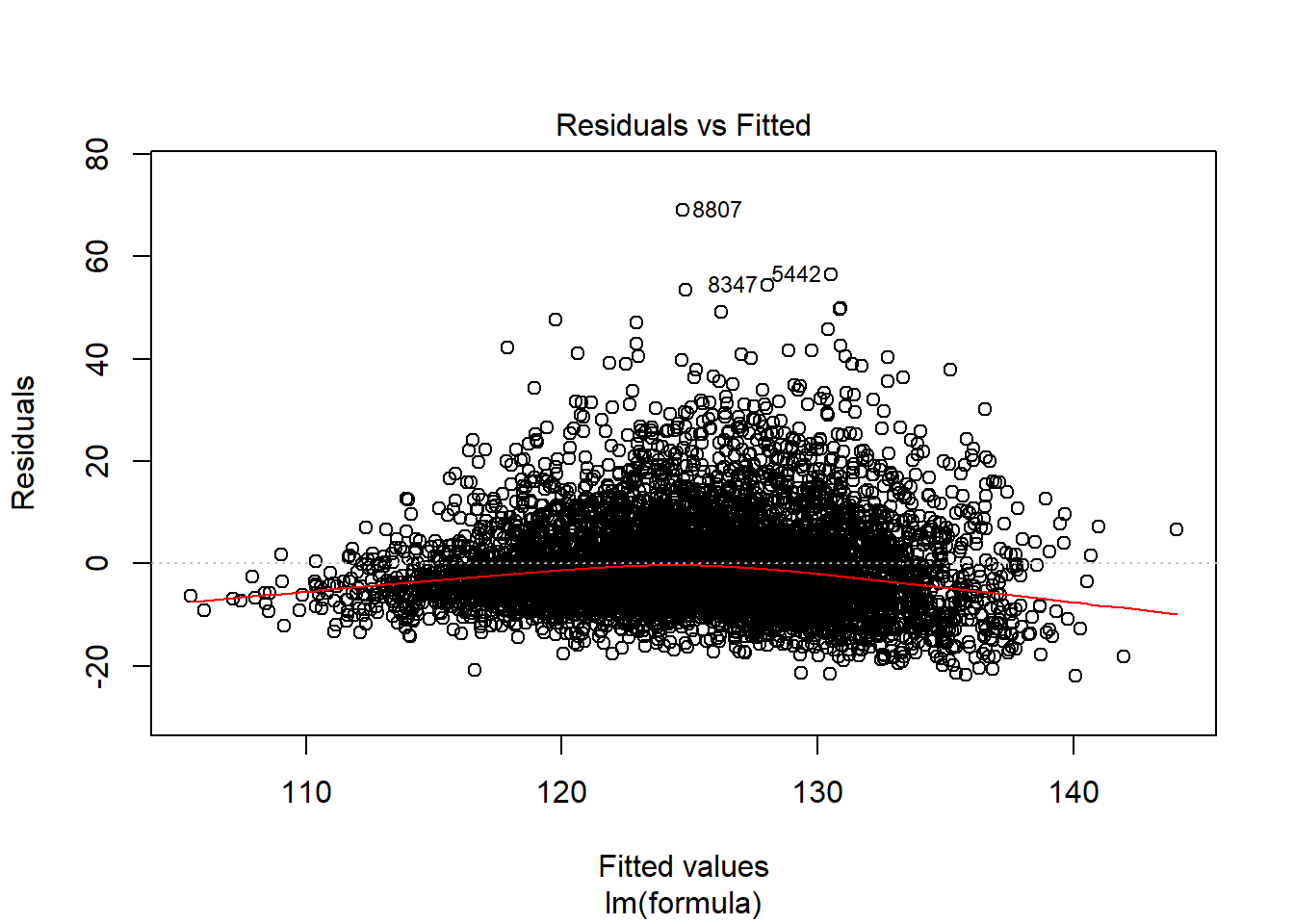
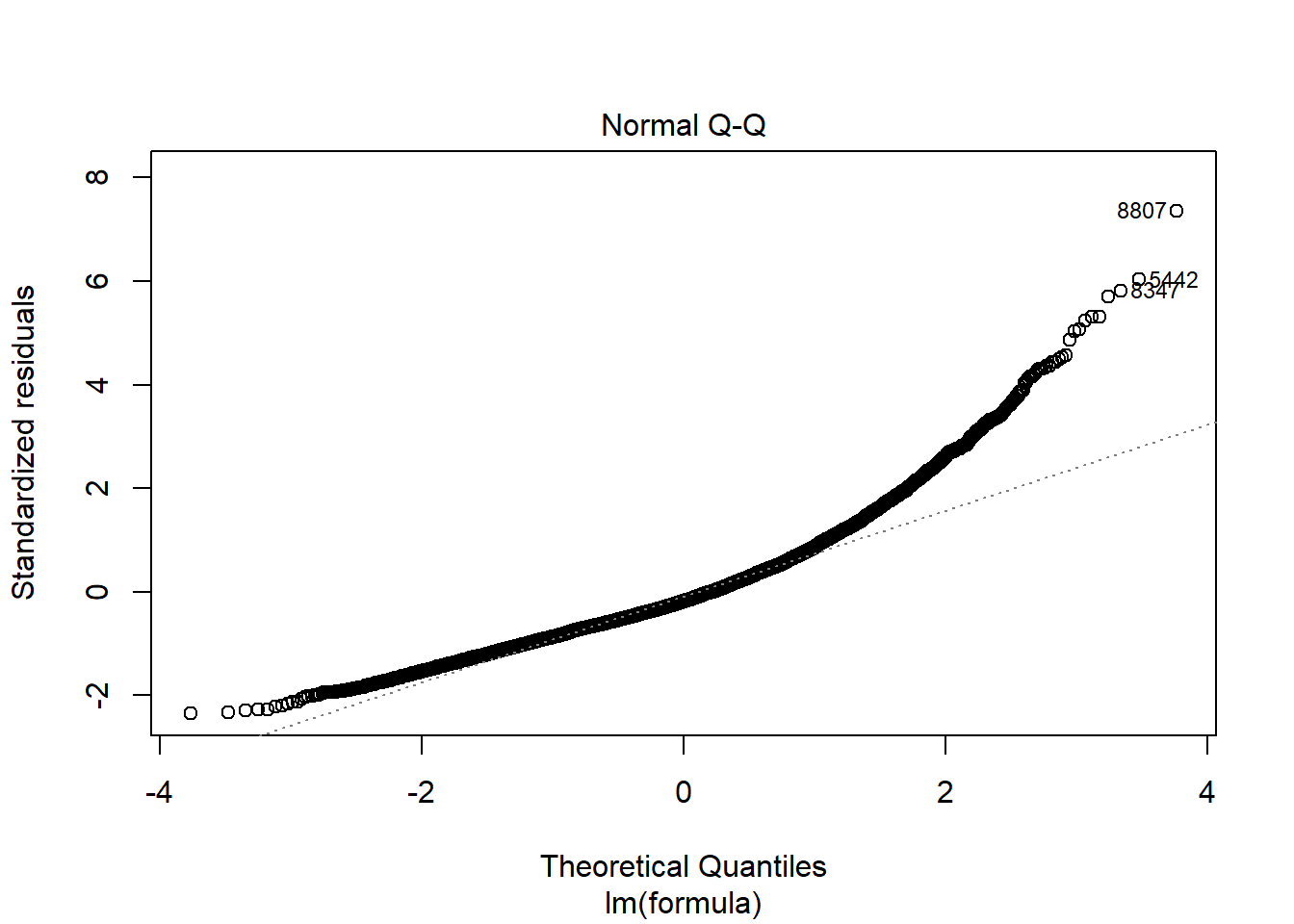
**Interpretation**

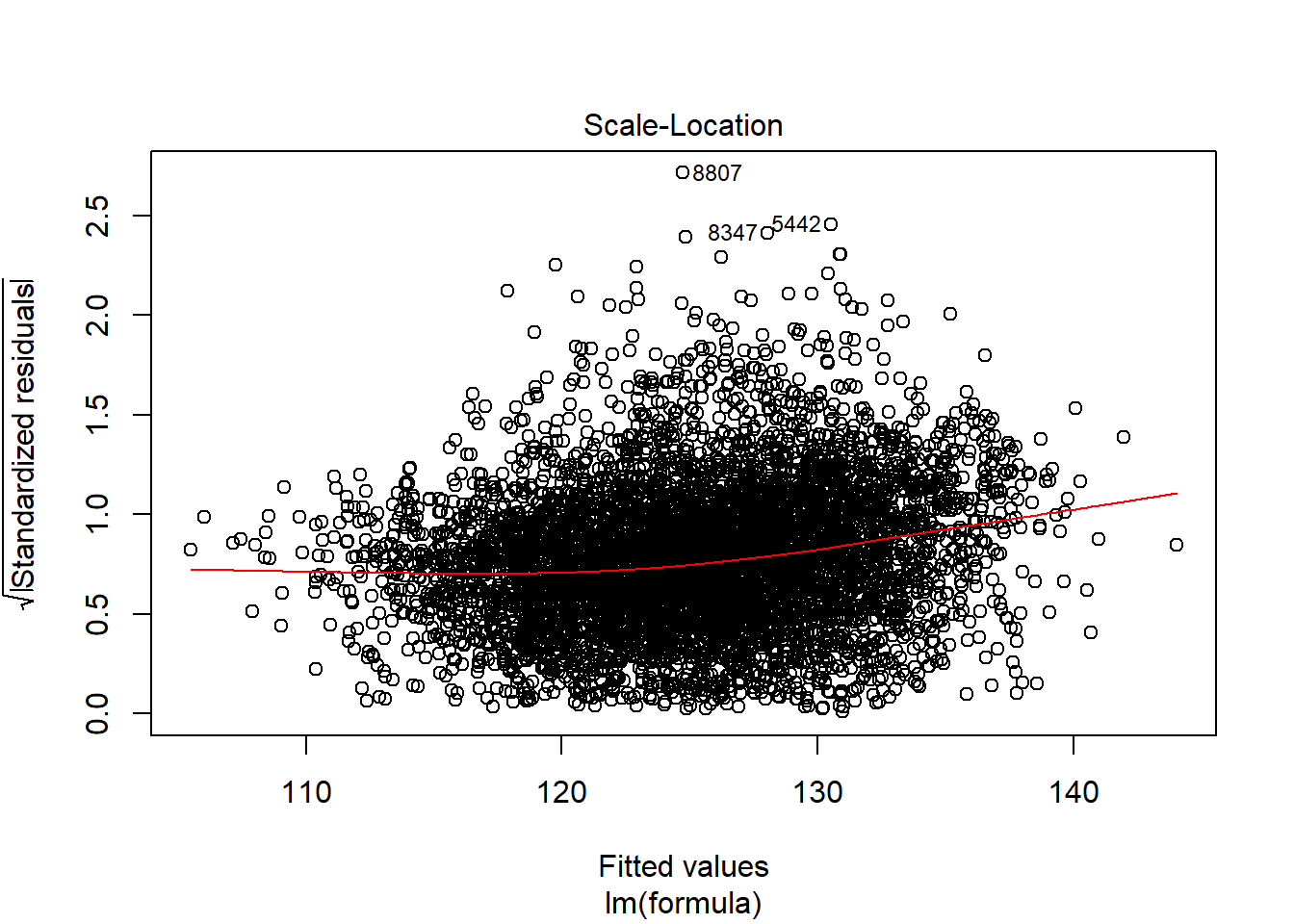
**Conclusion**

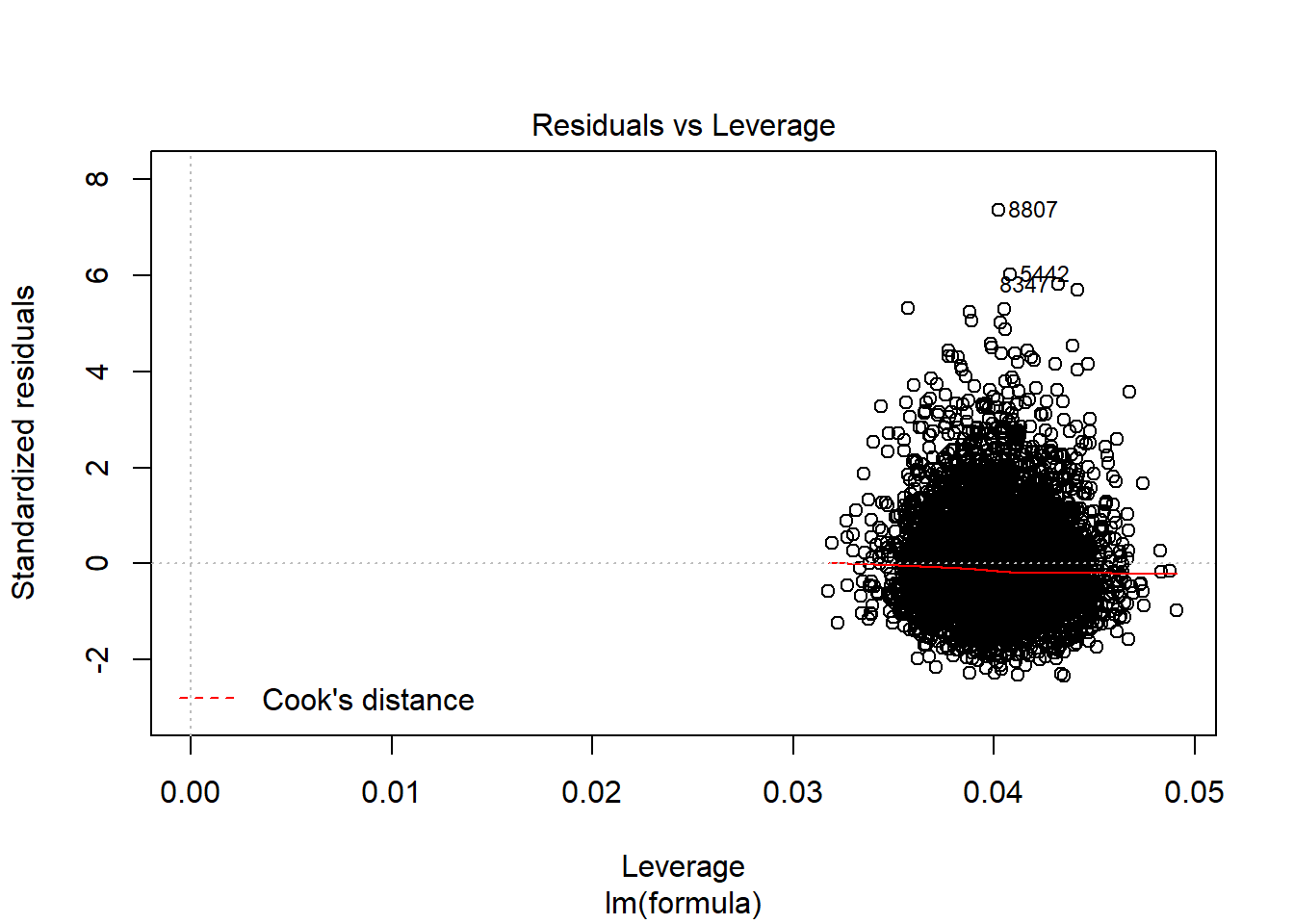
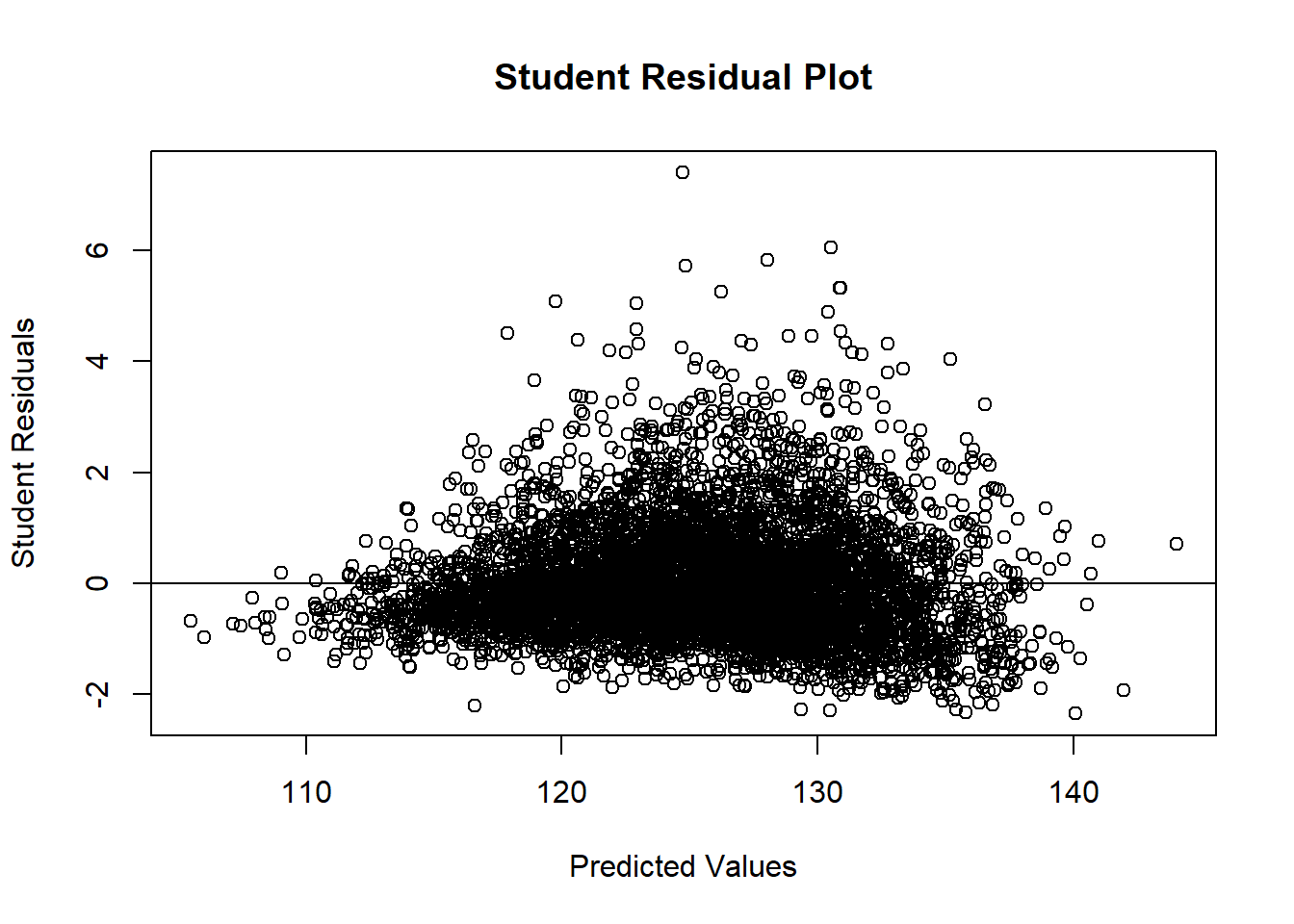
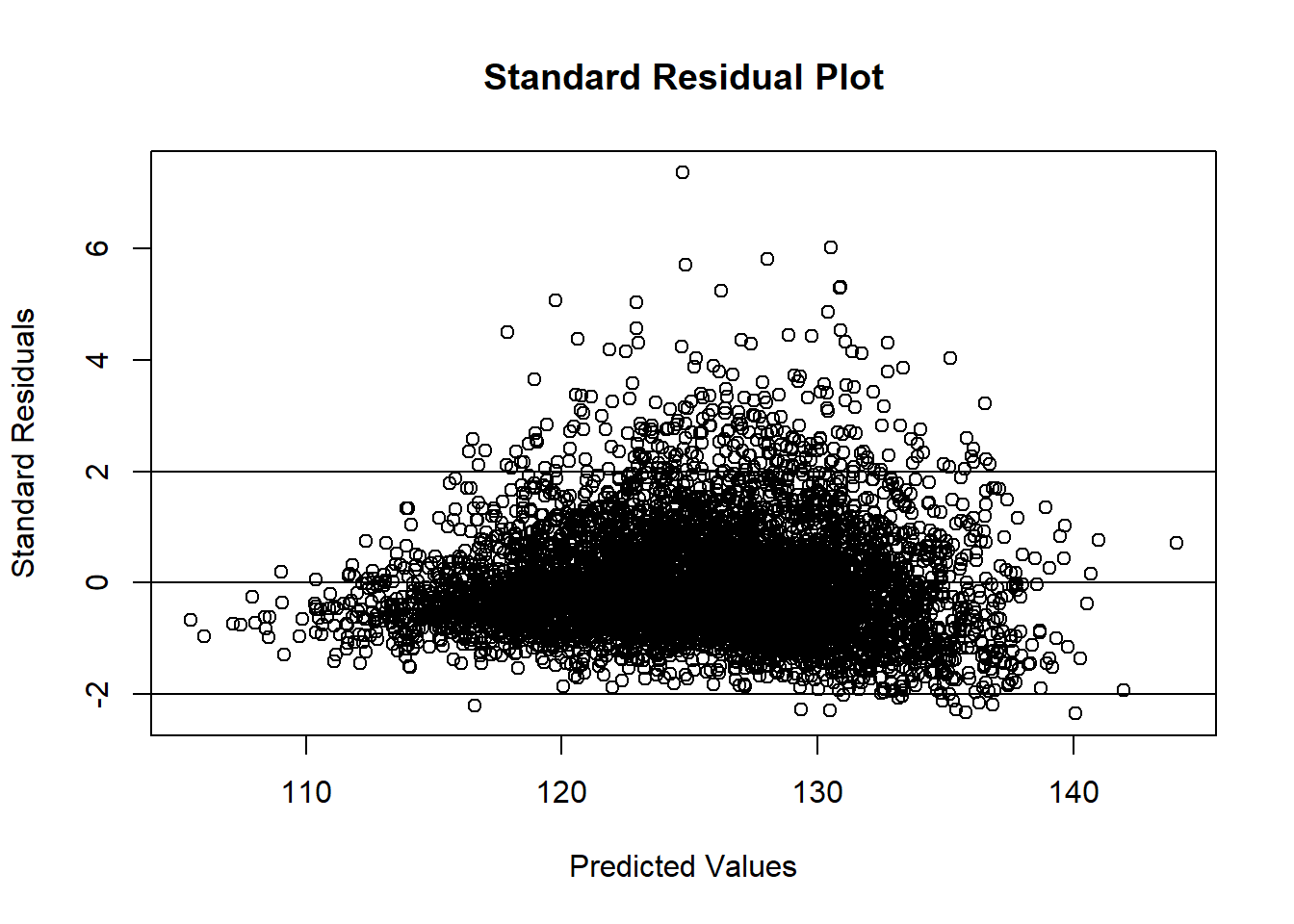
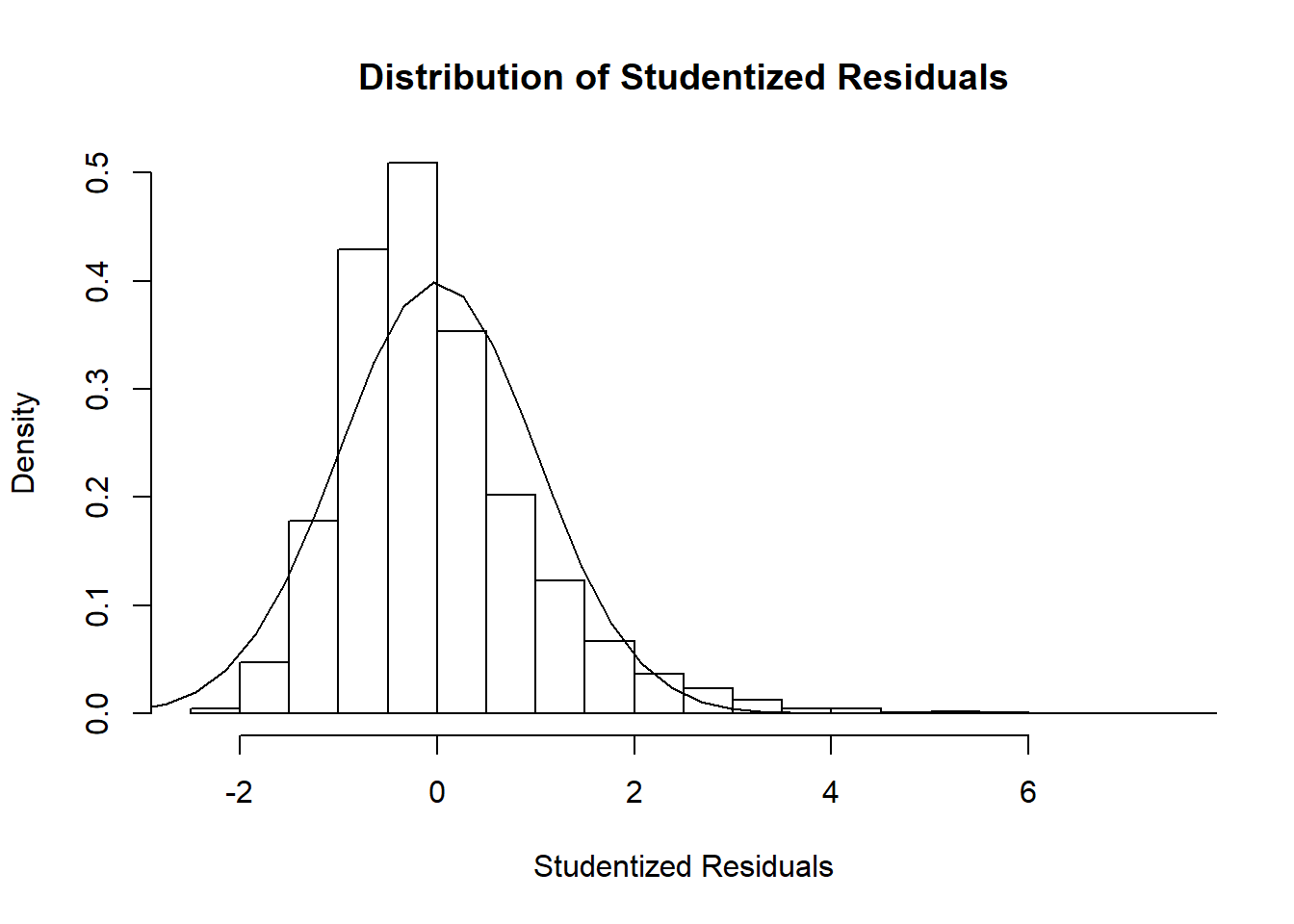
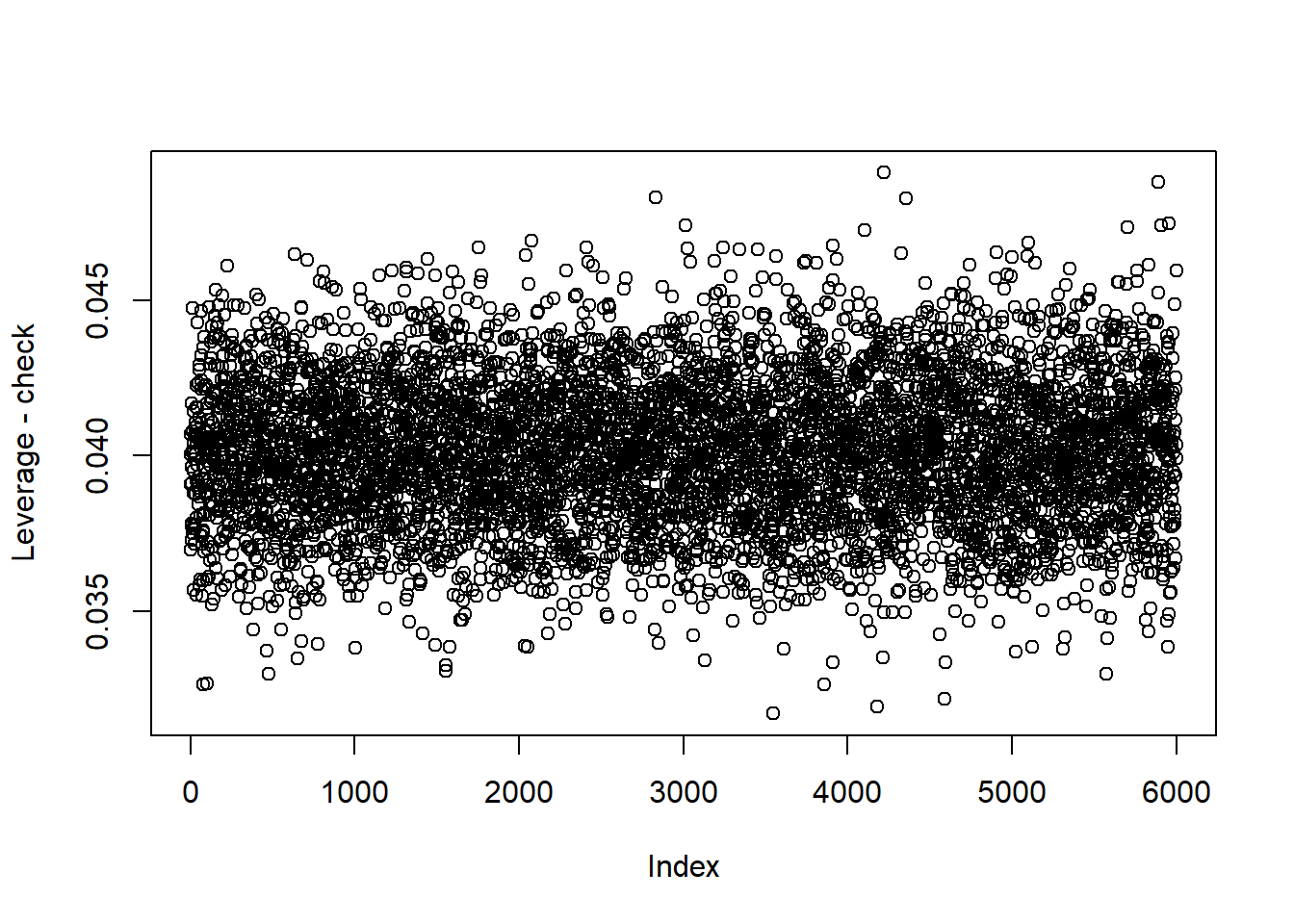
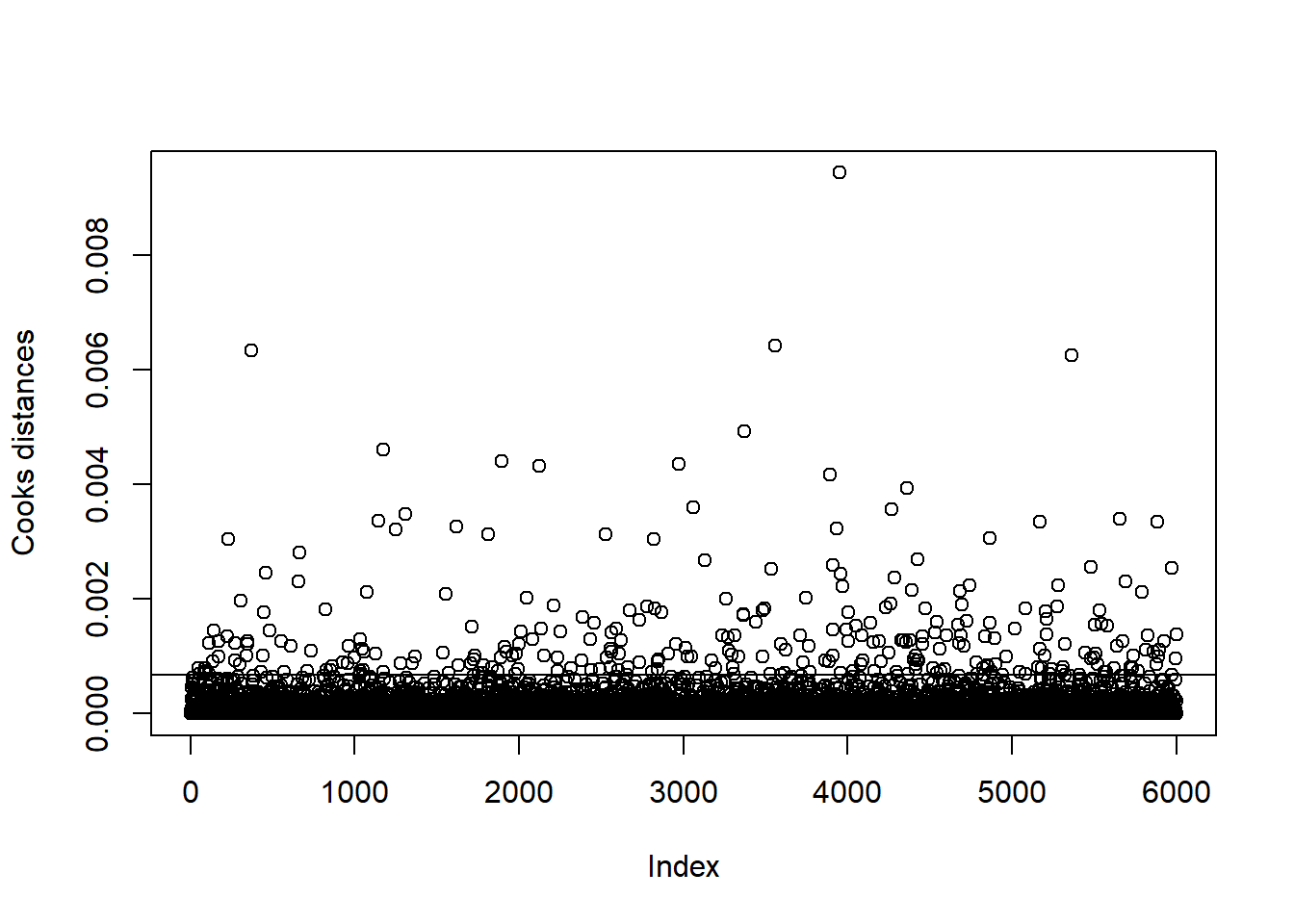
**APPENDIX**

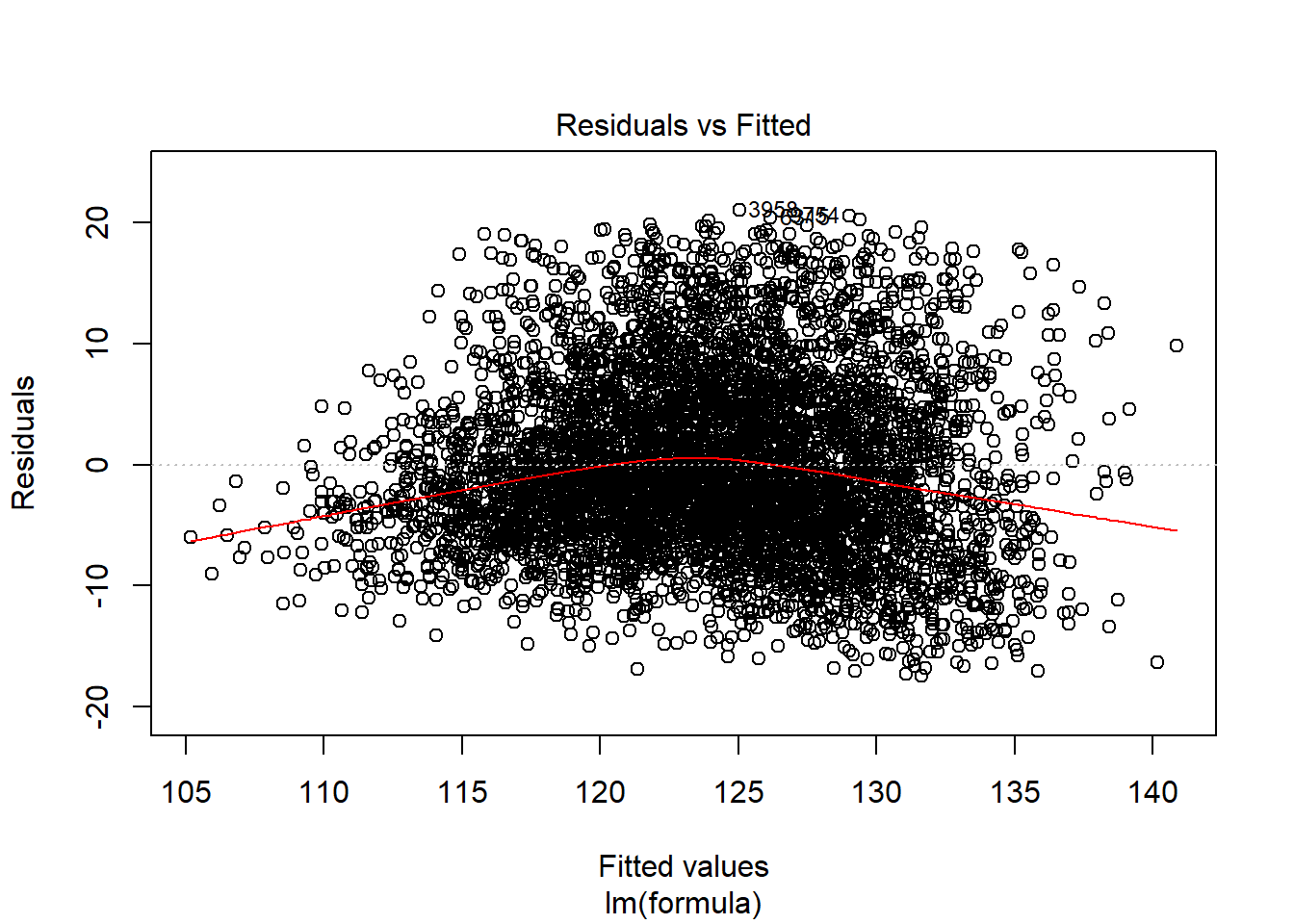
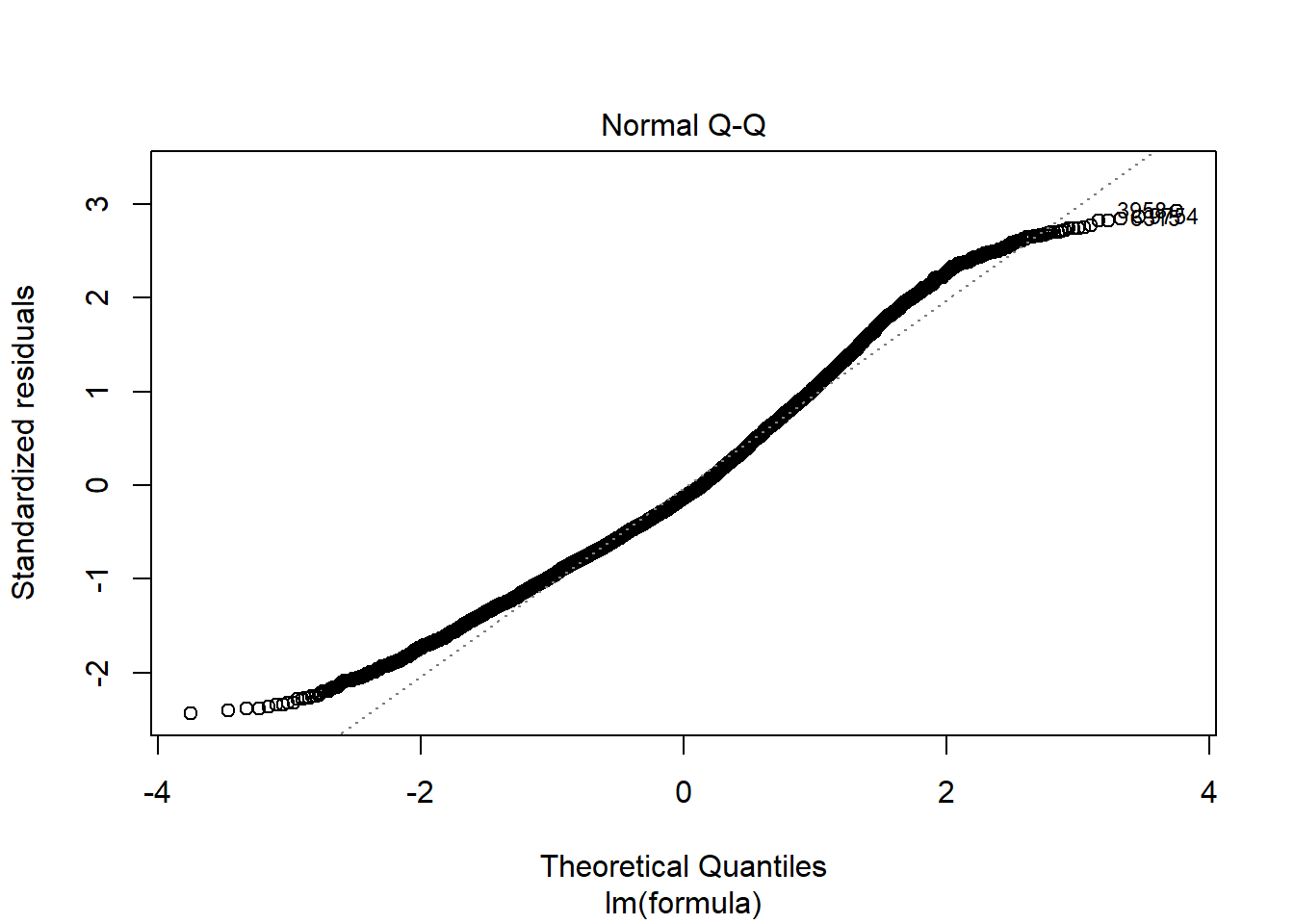
**Feature Engineering**

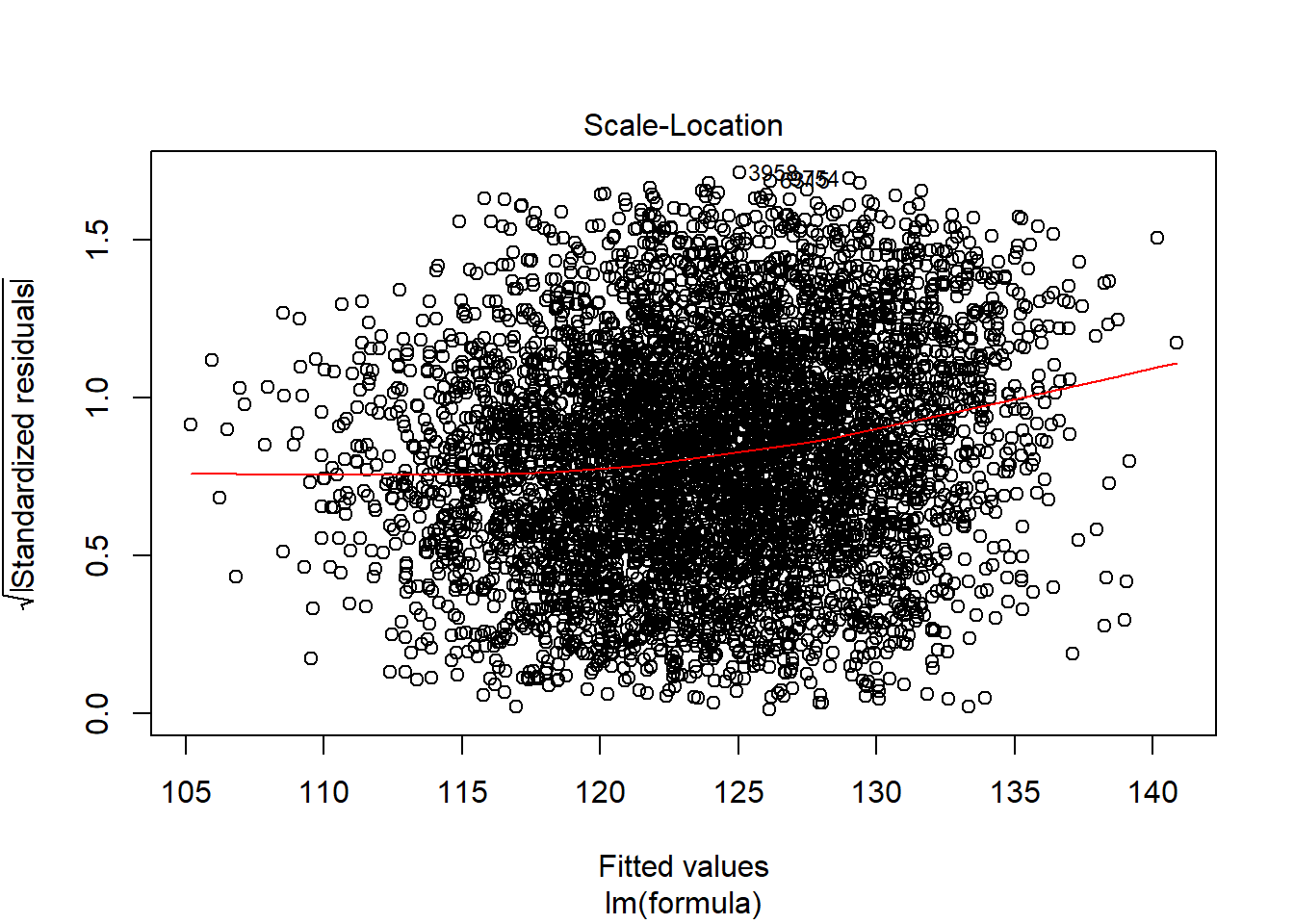
** **

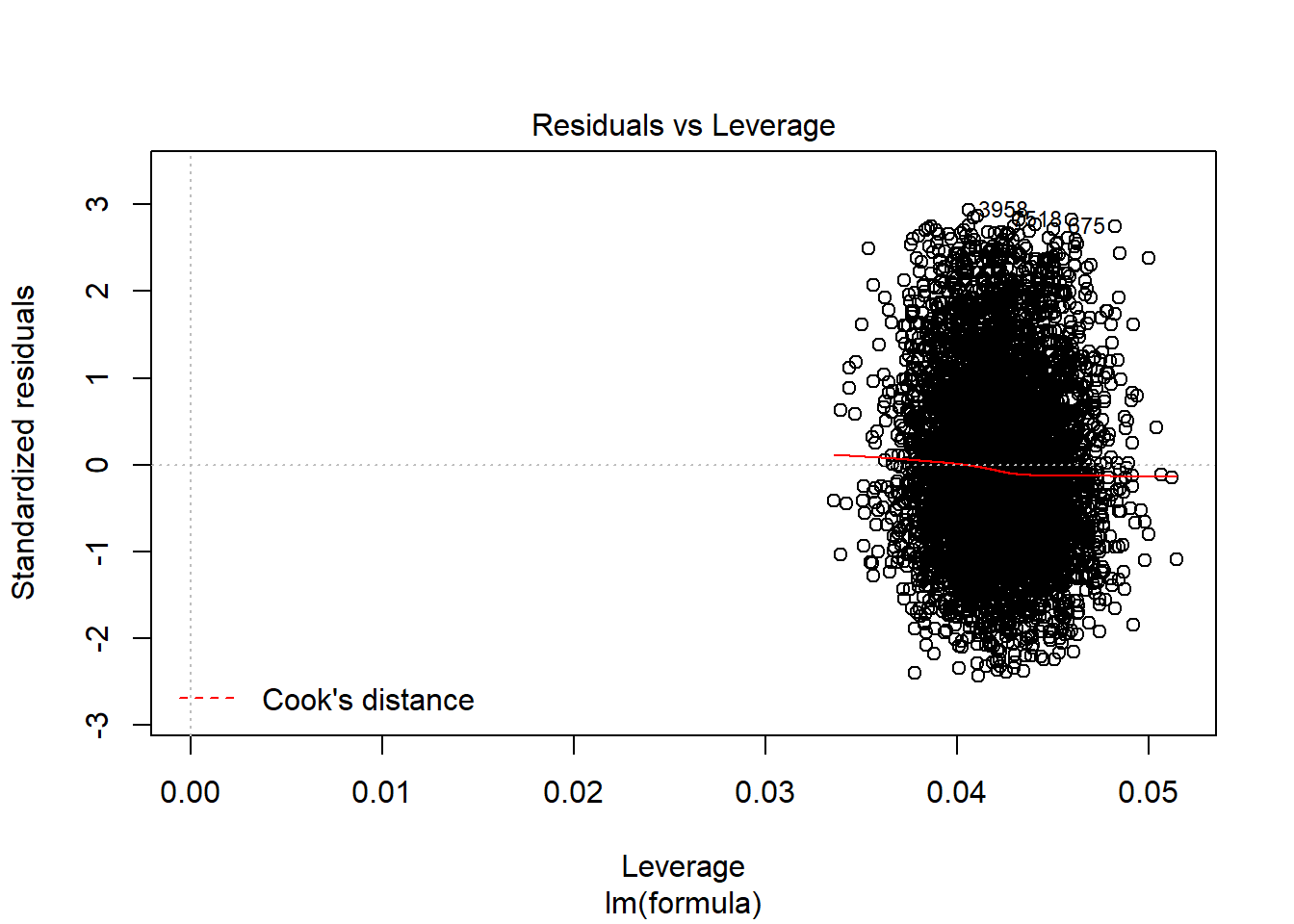
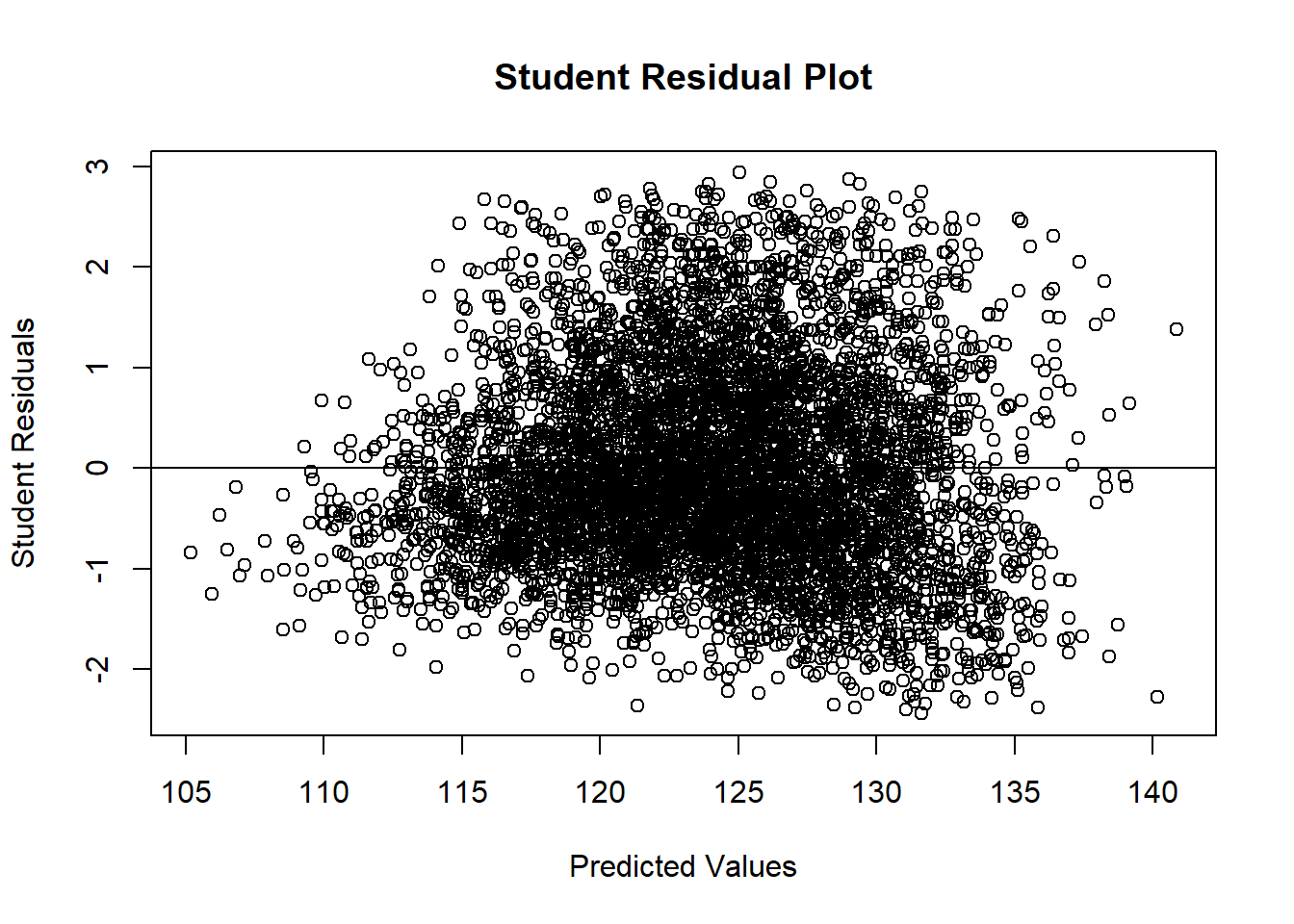
Full & Grand Means Model** **

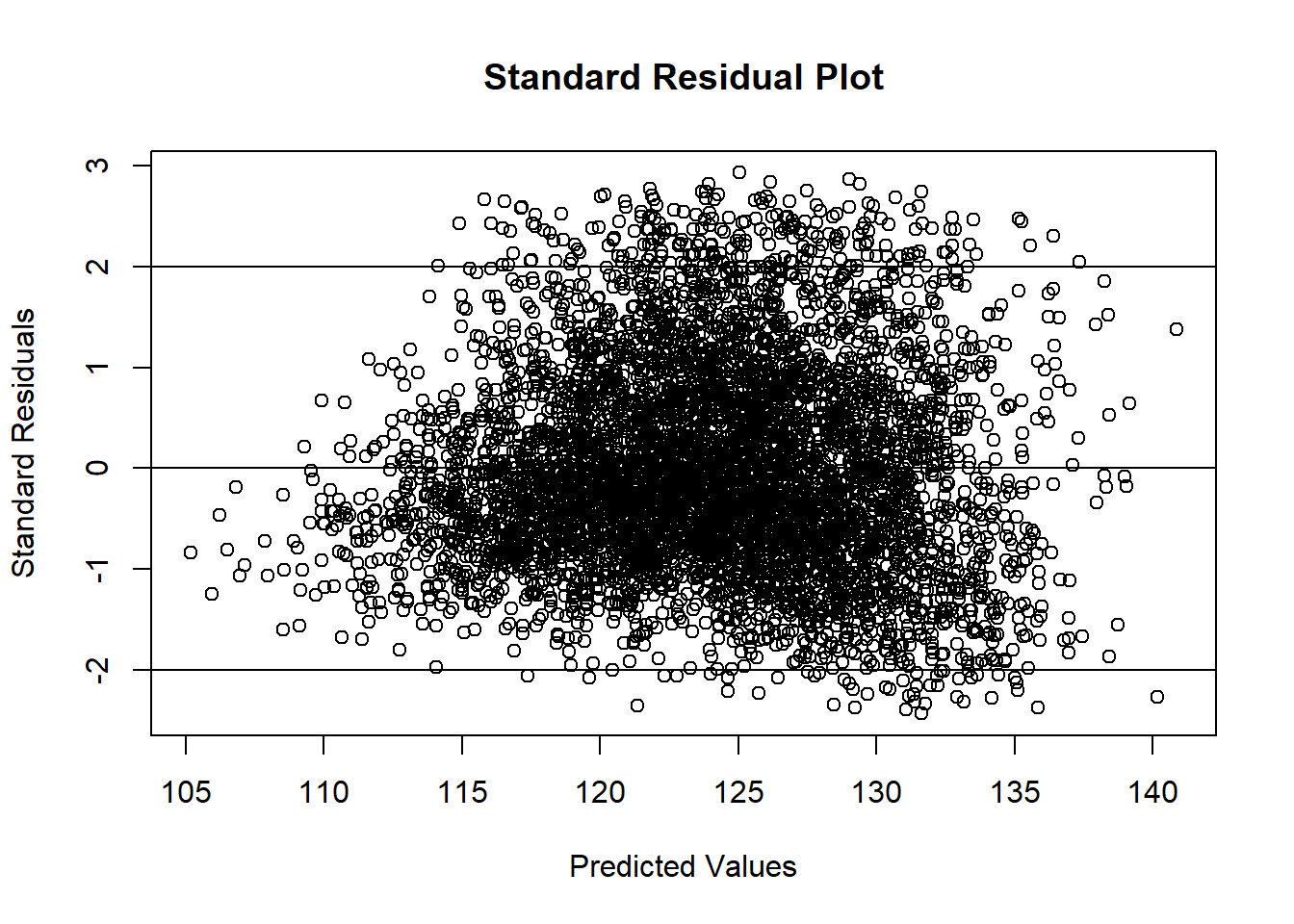
****

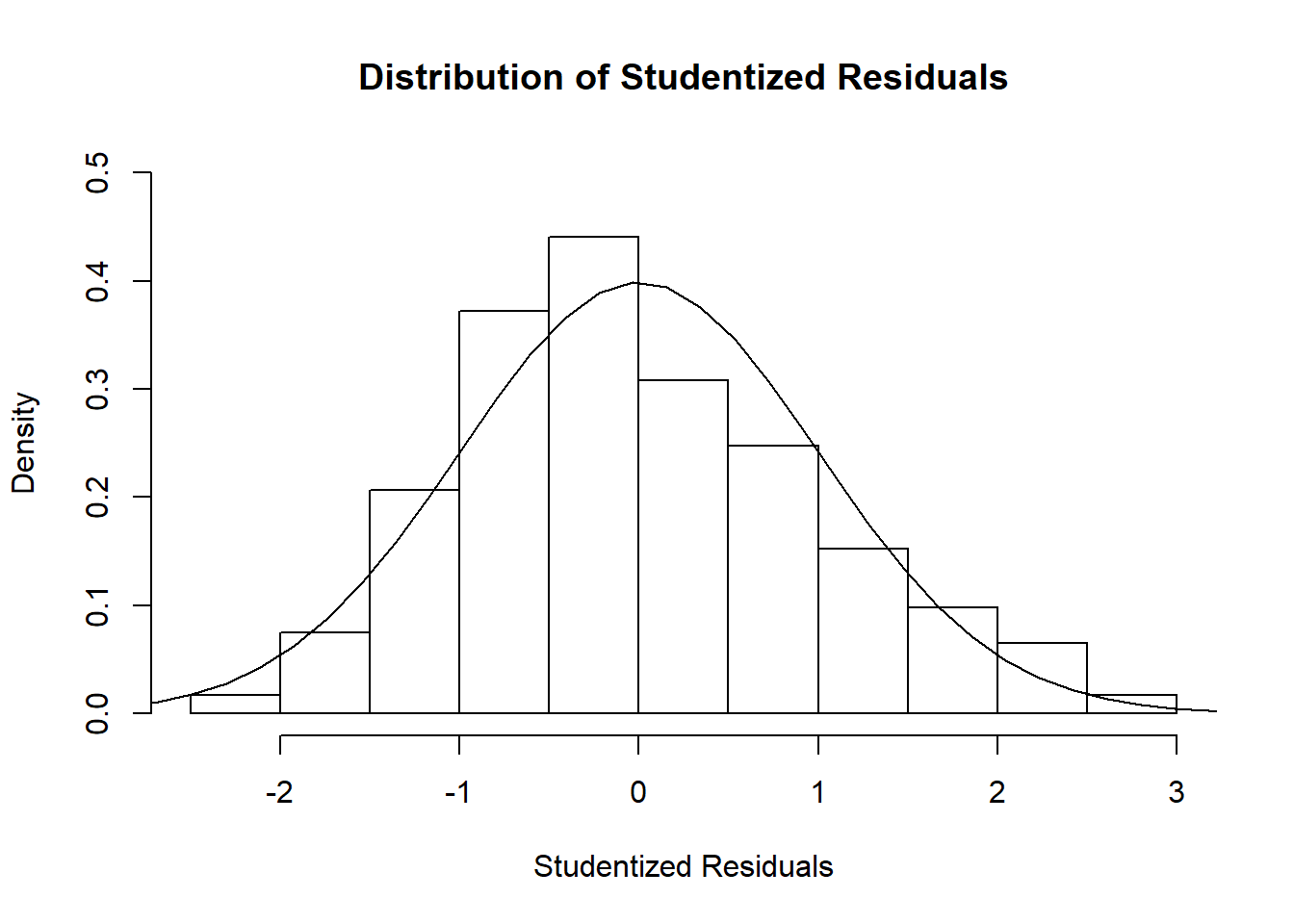
**     **

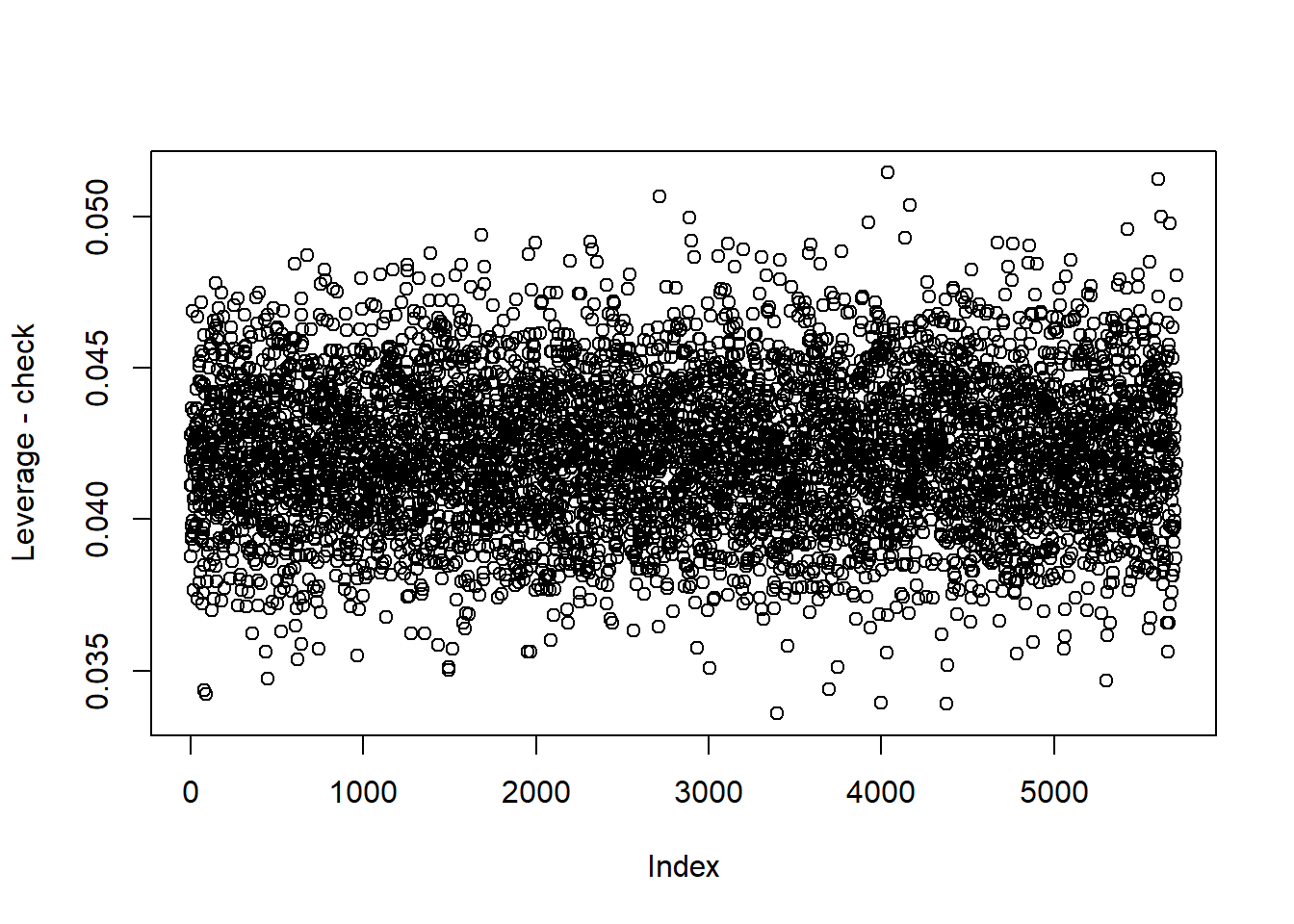
Just trial removing all high influence points  

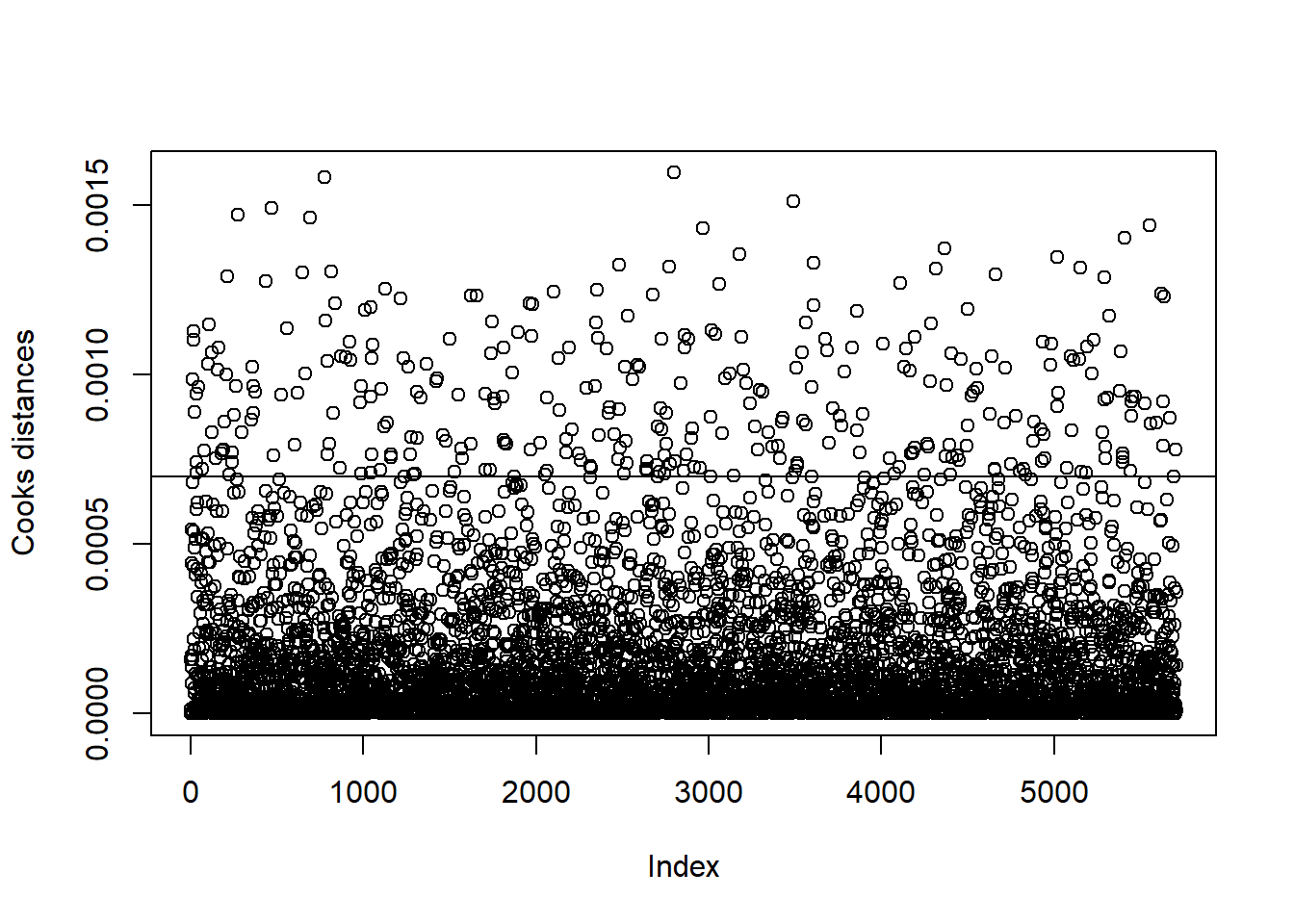
****

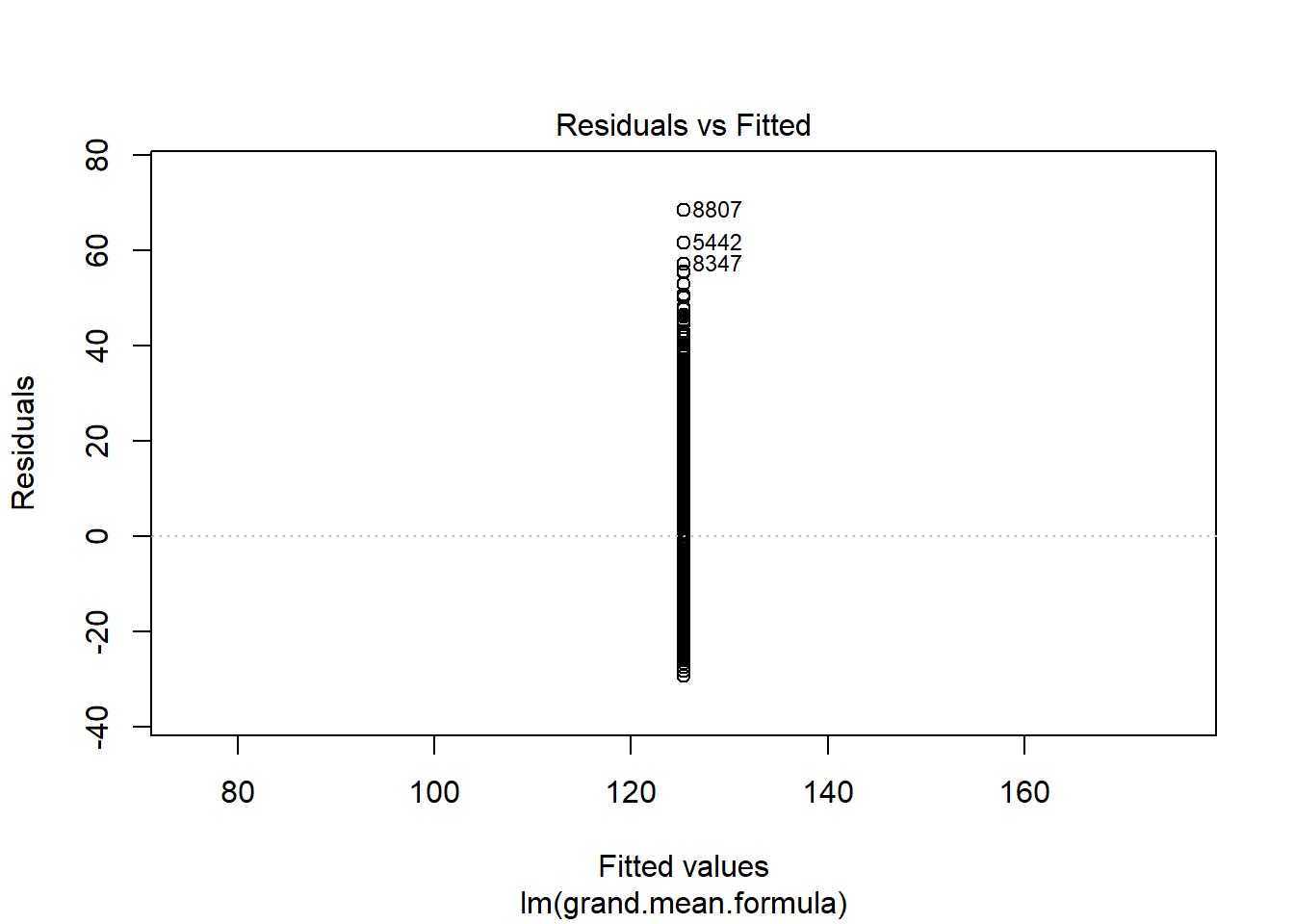


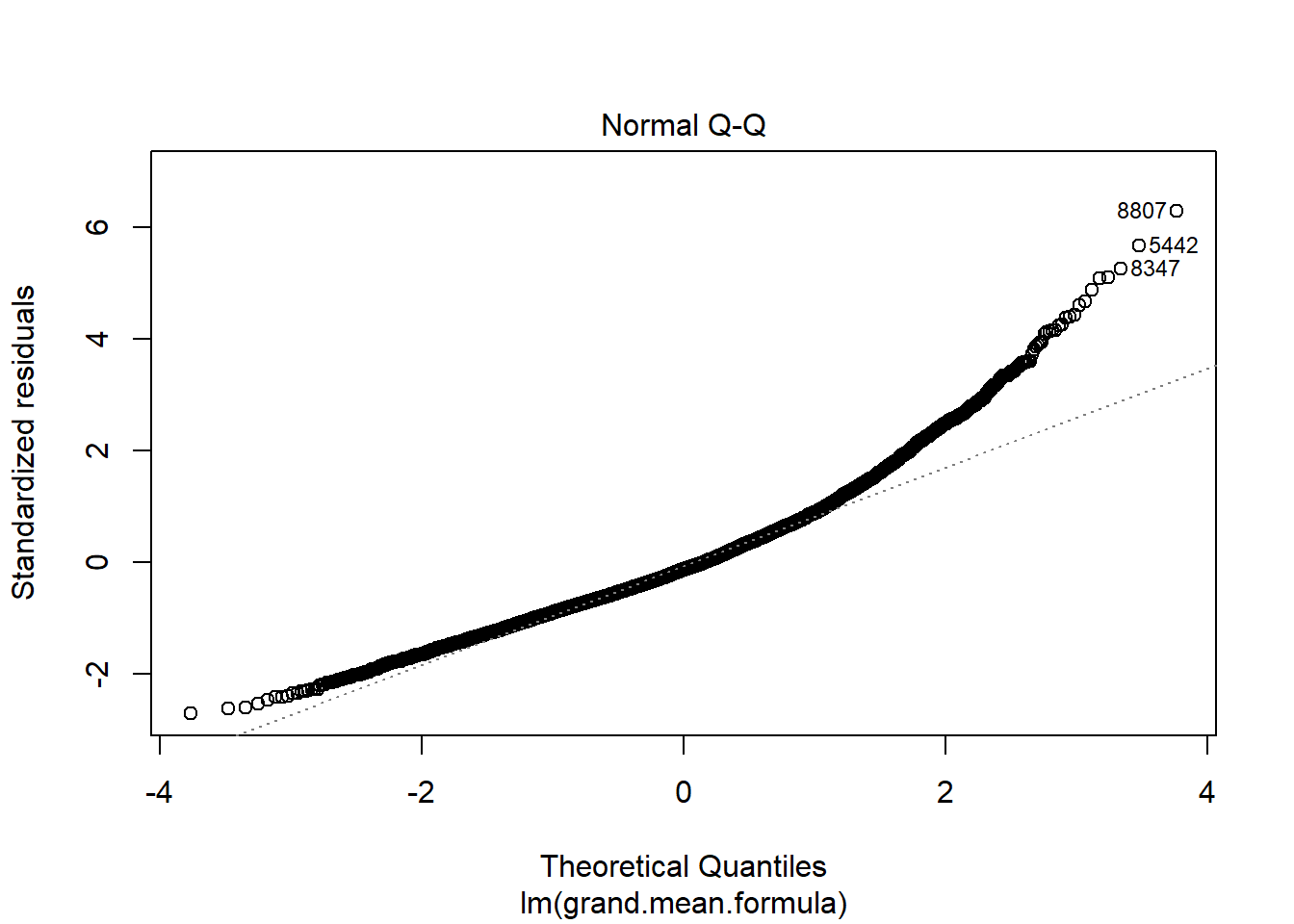
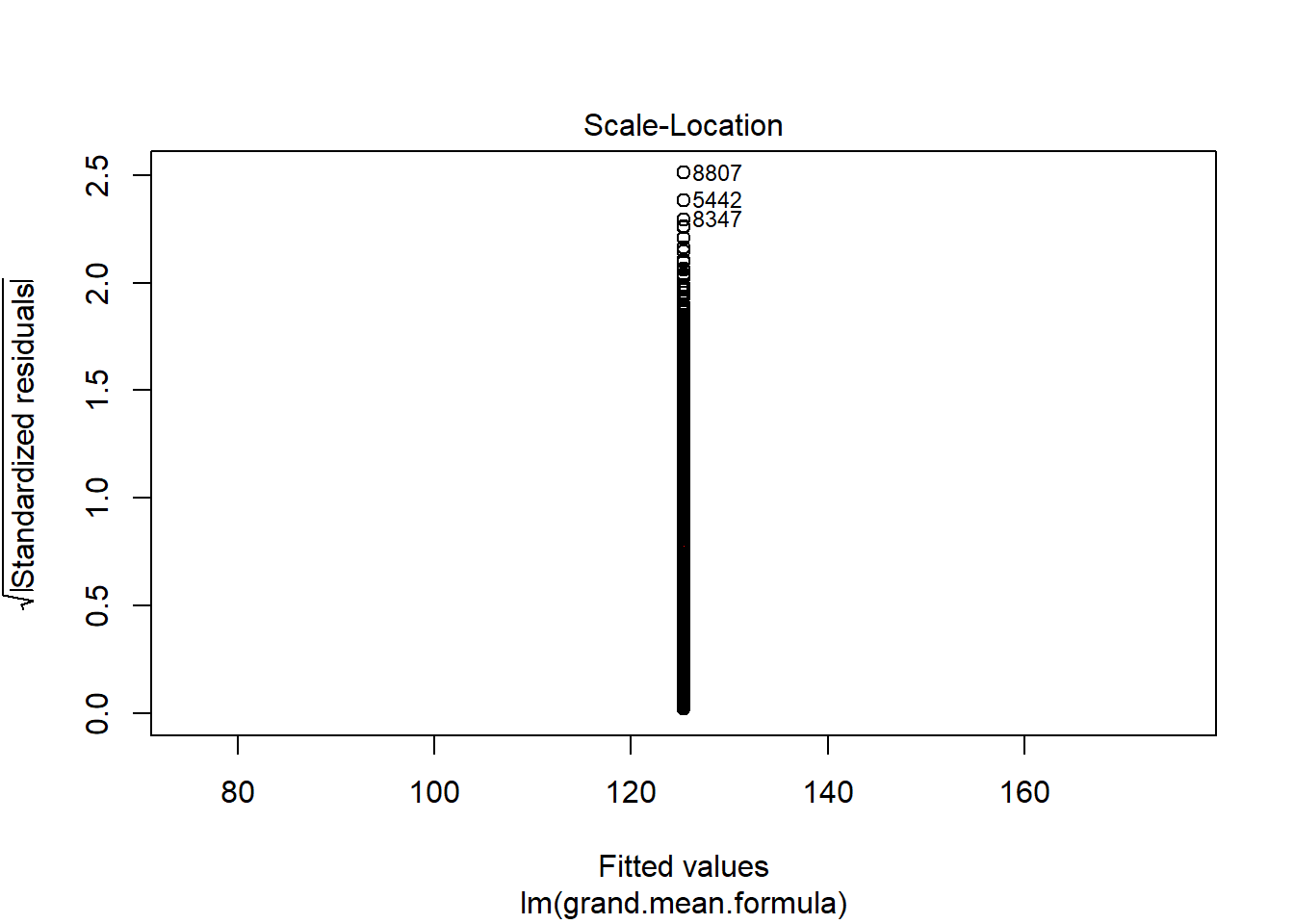


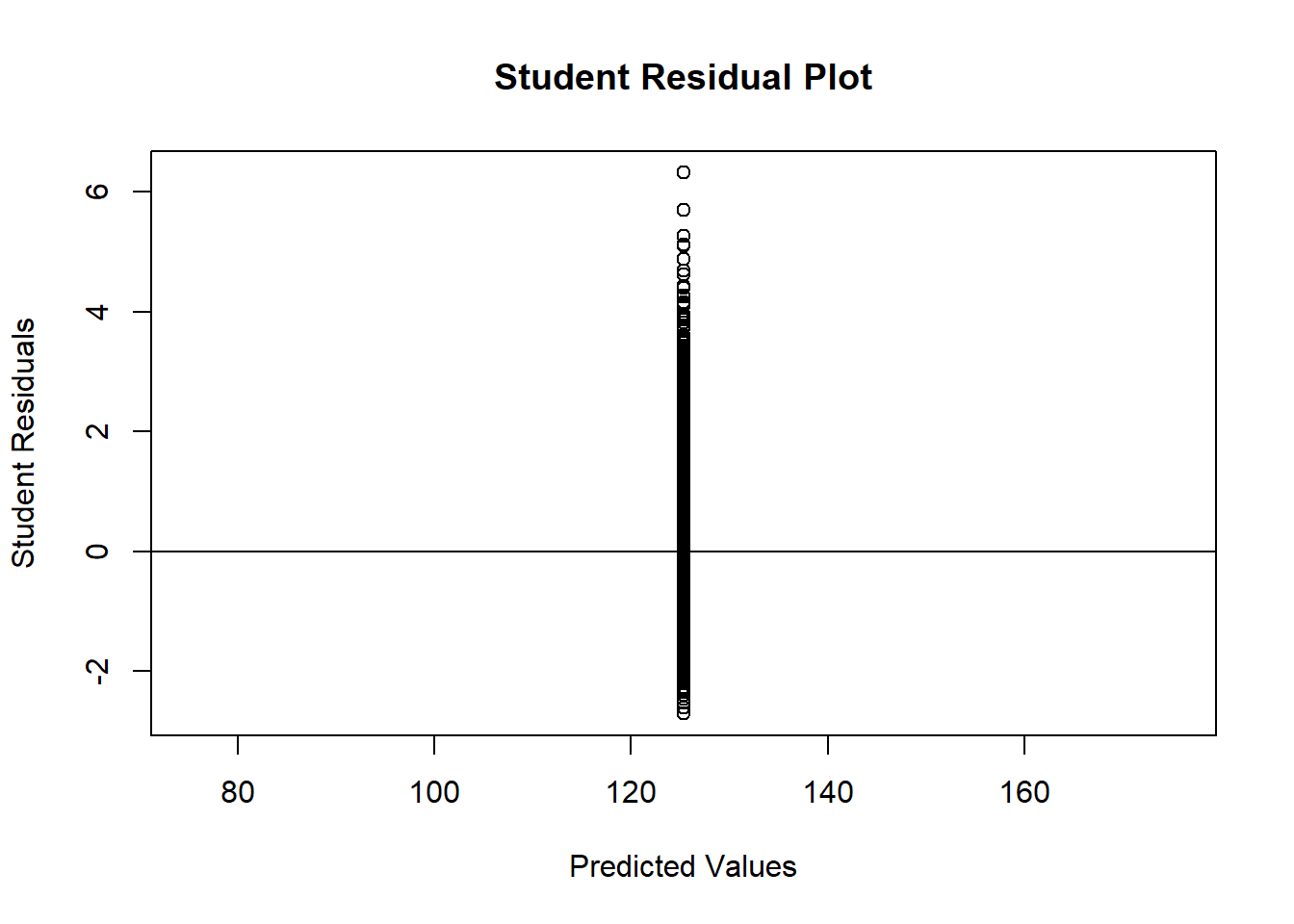
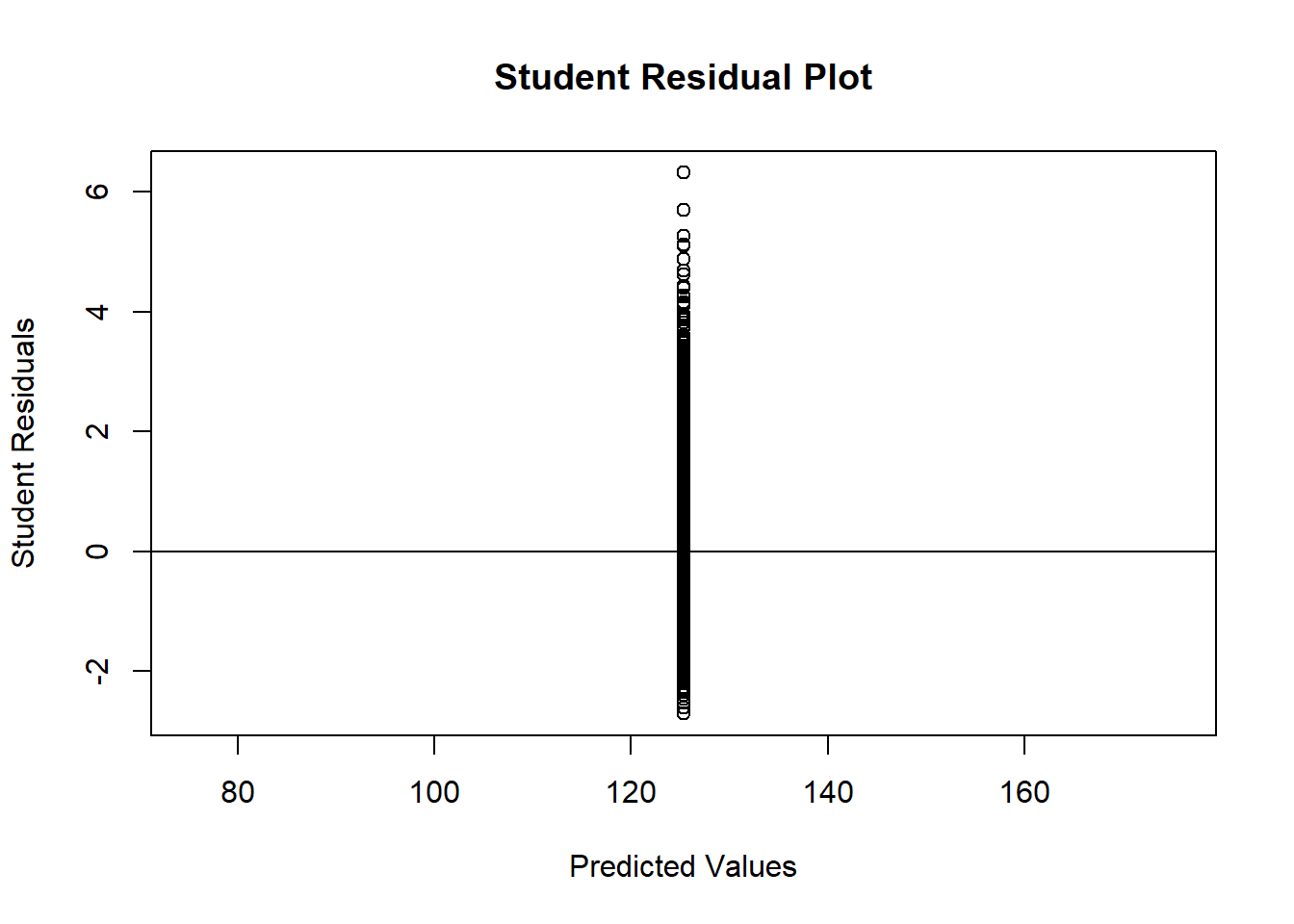
## Potentially influential observations of ## lm(formula = formula, data = data.train2) : 

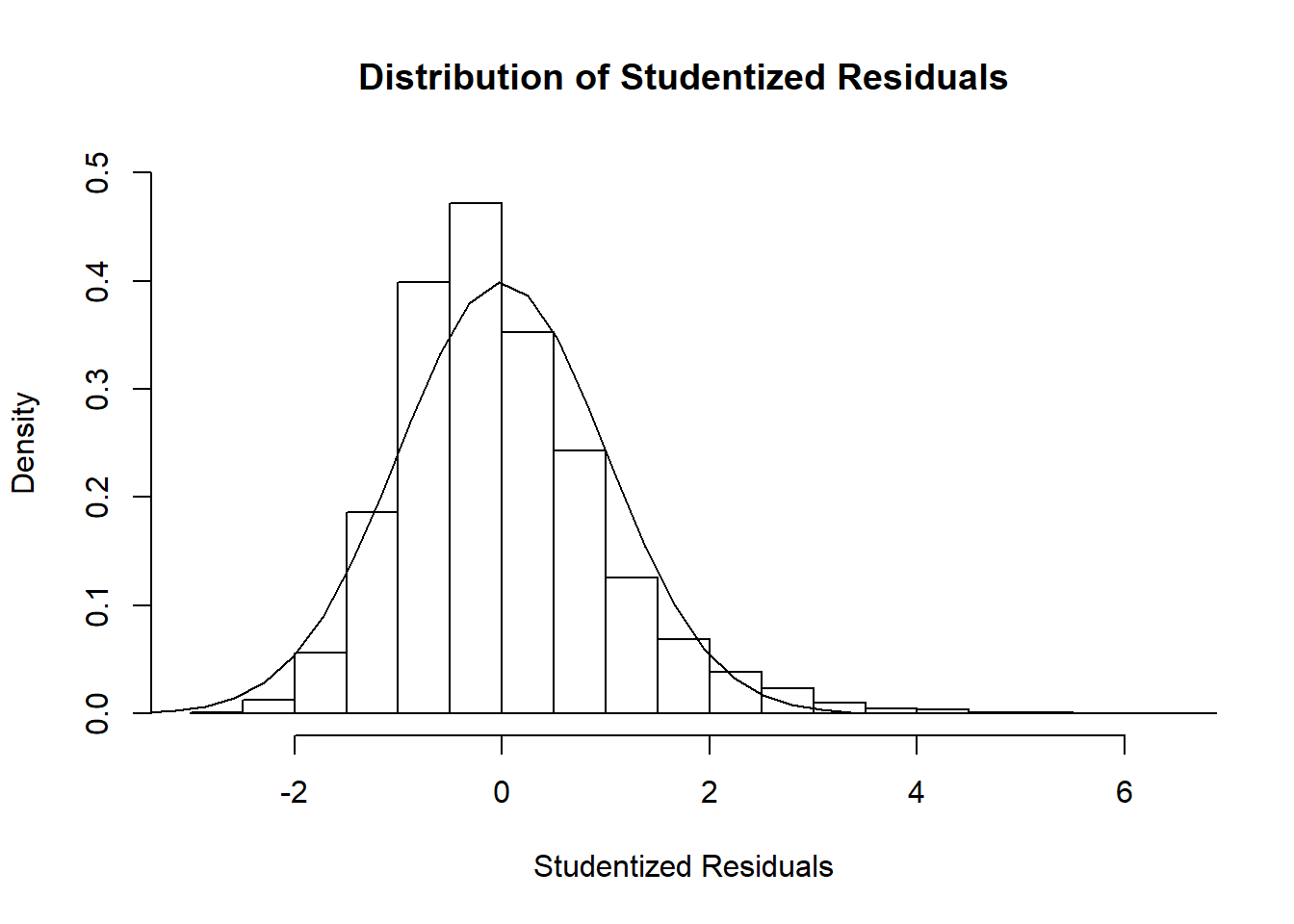


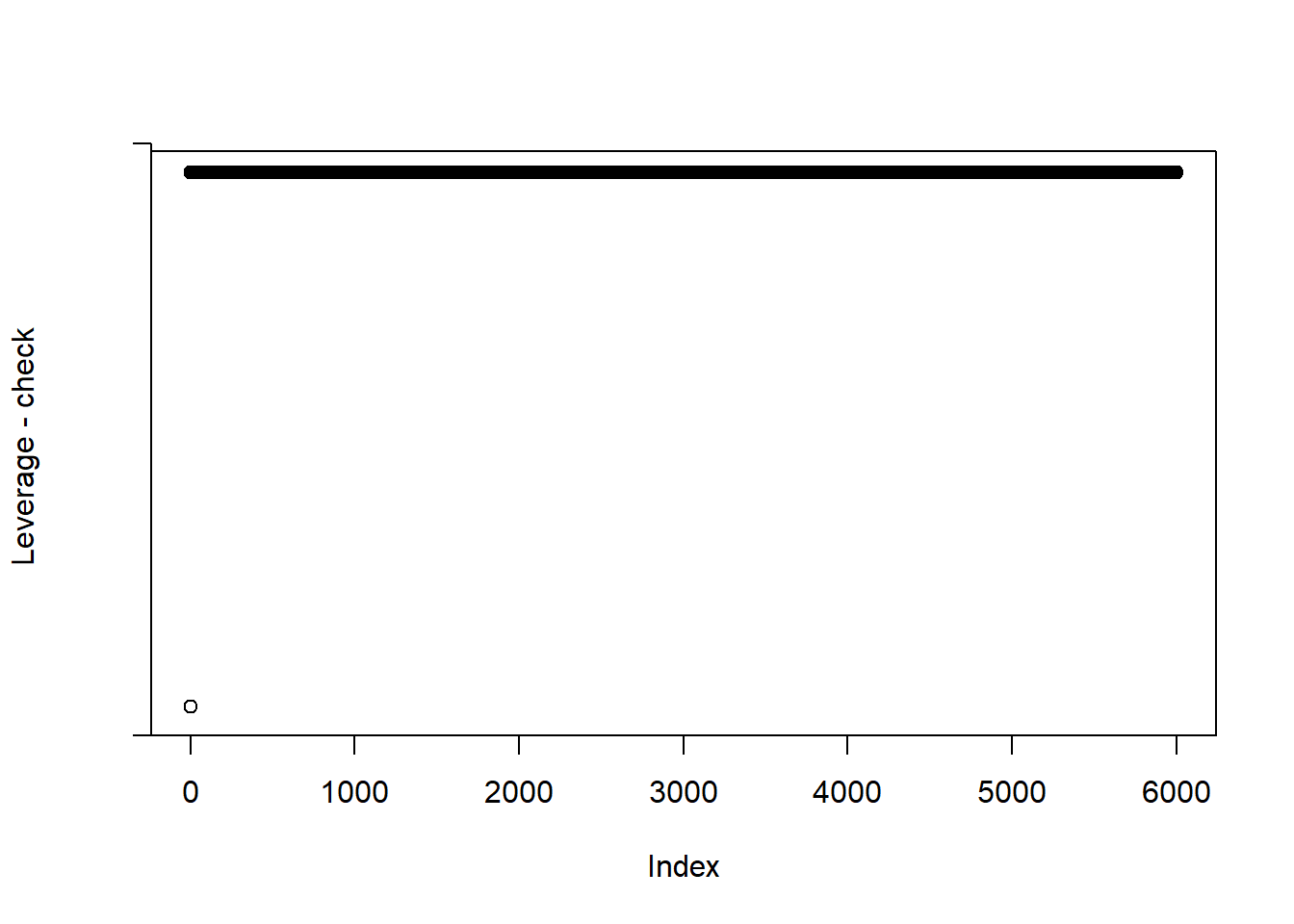
*# much more normal residuals than before. # See if you can check the distribution (boxplots) of the high leverage points and the other points*

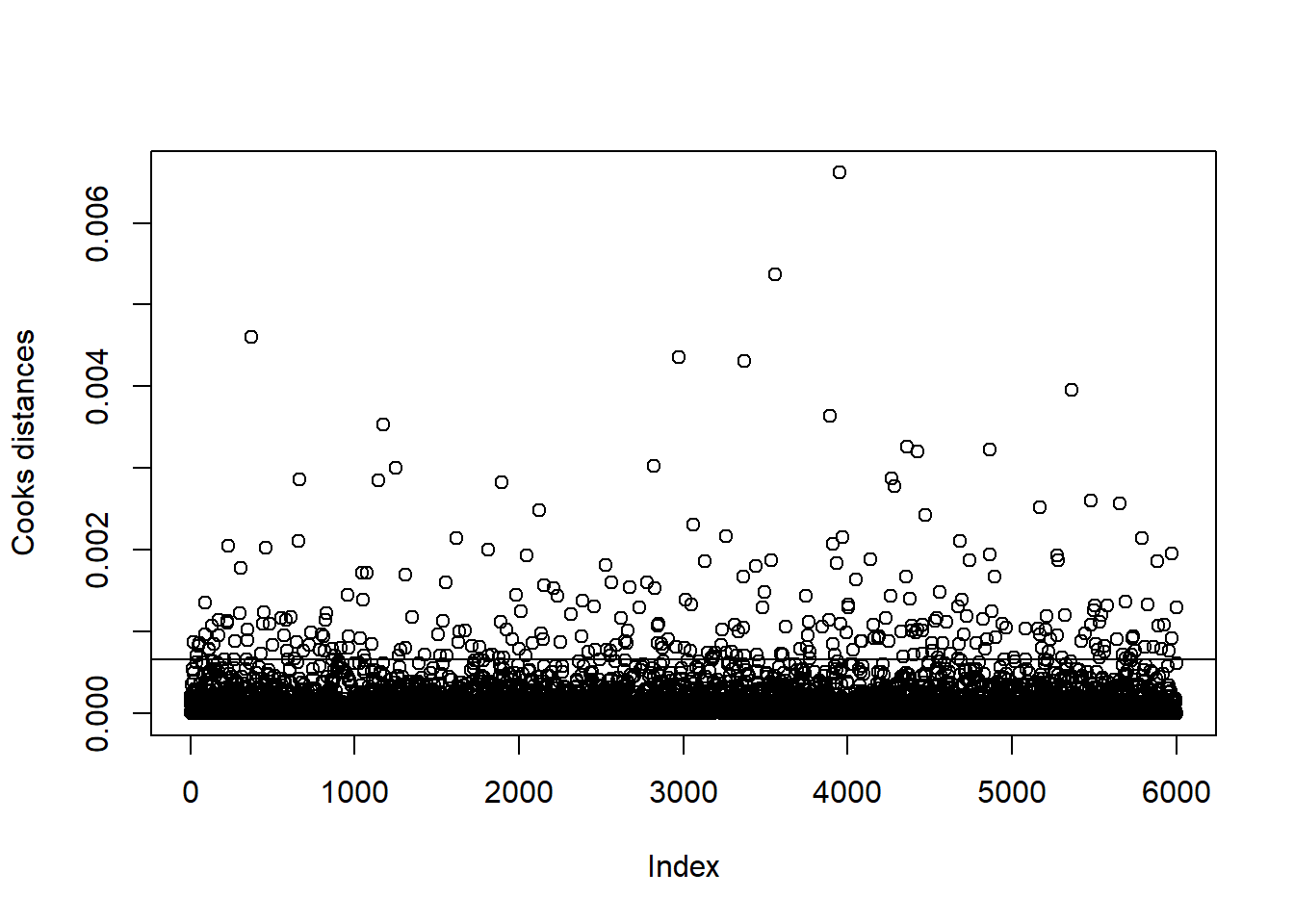








Variable Selection

