THE AMERICAN UNIVERSITY IN CAIRO

FIFA

PLAYER RATINGS

Malak Gaballa

Masa Tantawy

Dr Ali Hadi
Applied Regression Methods
Fall Semester 2022

TABLE OF CONTENTS

Introduction

Football is one of the most popular and competitive sports worldwide. The Fédération Internationale de Football Association, known as FIFA, constantly analyzes each player's performance and statistical attributes (such as strength, stamina, work rate on the field, etc.) and accordingly gives a rating to every player. However, some attributes have a heavier influence on a player's overall rating than others. Therefore, the aim of this project is to find whether or not all player attributes are significant or affect the overall rating that FIFA decides.

Data Preparation

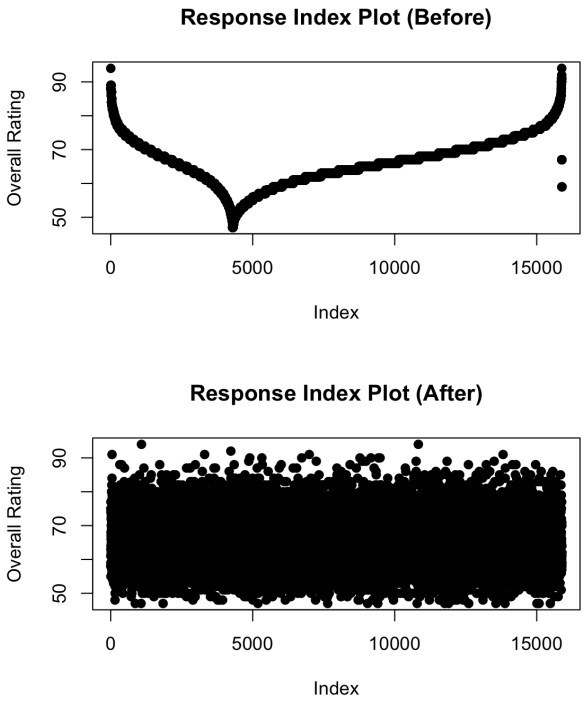
The data obtained from the internet requires some preparation to be ready for usage. For a detailed description of the data and the attributes, please refer to FIFA Project Description.pdf. The script named *data preparation.R* contains the R codes for the following steps.

First, some attributes in the dataset were unexplained (their names were abbreviations and no description was found for them online), so these columns were omitted from the dataset. Likewise, we decided to focus on football players who are not goalkeepers since the latter group has different criteria for measuring their performance and what affects their *overall rating*, so the attributes for goalkeepers were also omitted as also observations for goalkeepers. The variable *name* was redundant due to the presence of *full name* and *birthdate* due to the presence of *age*, so these 2 columns were omitted. The *ID* was only omitted as it is only an identifier for each player and the variable *positions* because some players have multiple positions or can play in any position.

Second, the dataset contained 9 columns with missing values; 15137 out of 15889 observations had at least 1 missing value. To address this issue, kNN was used to impute missing

values in the numeric columns, and categorical variables with missing values, constituting 4 attributes, were omitted from the dataset.

Third, the index plot of the response variable *overall rating*, which was made the 1st column in the data frame, showed a necklace-like pattern;this implies that players were ordered in that way for a certain reason which is the index plots for predictors has to be checked.Looking at the index plots of predictor variables, it turned out that some pedictors such as composure and reactions(found below in the Appendix A) resembled the same necklace-like pattern as the response variable. Even though we unable to identify a specific reason as to why the observations were ordered in such a manner, it can be inferred that some attributes are influencing the order or placement of the observations.Thus, random sampling was used which resulted in the index plot of Y being a random scatter of points. The seed was set to ensure that the same order of observations is yielded every time the code is run. This method of resampling, however, will not affect the regression results in any way because the same data is used; it just ensures the randomness of the response and predictors' indices plot.



Fourth, each categorical variable with more than 2 levels was divided into multiple indicator variables. *Preferred foot* had only 2 levels indicated by 1 (Left foot) and 2 (Right foot) respectively, so 1 was subtracted from this column to make it binary. As for *international*

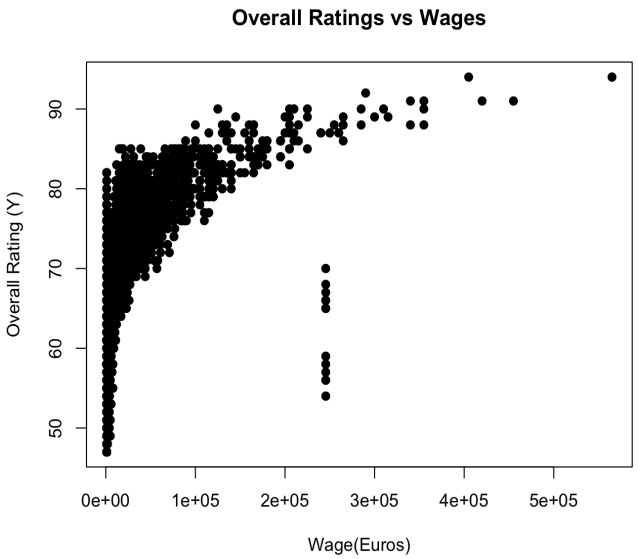
reputation, it was divided into 5 indicator variables (one for each level), and likewise for *weak foot; skill moves* were divided into 4 indicator variables, and *work rate* into 9 indicator variables. This is because in football, the change in *overall rating* (which is the response variable) is not equal for each unit increase/decrease in any categorical variable. In other words, the regression coefficient of each categorical variable does not represent the changes equally.

Now the data is ready to begin working with. It was saved into a new file named *FIFA Cleaned.csv*.

Step 1: Validation of Assumptions before Fitting

The script named *Project.R* contains the R codes for the following steps.

The first step before fitting a model to the data is to validate the assumptions. The index plot of each predictor variable (shown in Appendix A) indicated that some variables might need interaction variables in later iterations as there are clusters of points, and they also showed that some variables are collinear. To validate the assumption of linearity in regression coefficient, the response variable versus each predictor variable was plotted (shown in Appendix A). It was clear that the variable *value_euro, wage_euro,* and *release_clause_euro* needed transformation (as shown on the right). This issue was dealt with in the next step. Next, for an overview of the dataset, the histogram and boxplot of the response variable were plotted and a numeric summary of each predictor variable.



Step 2: Adjustments to Match Assumptions

As concluded from the previous step, some predictors needed transformation. The optimal lambda for the power transformation of each variable that needs transformation was determined after plotting the predictor versus the response for several values of lambda. The numeric columns were standardized following transformation.

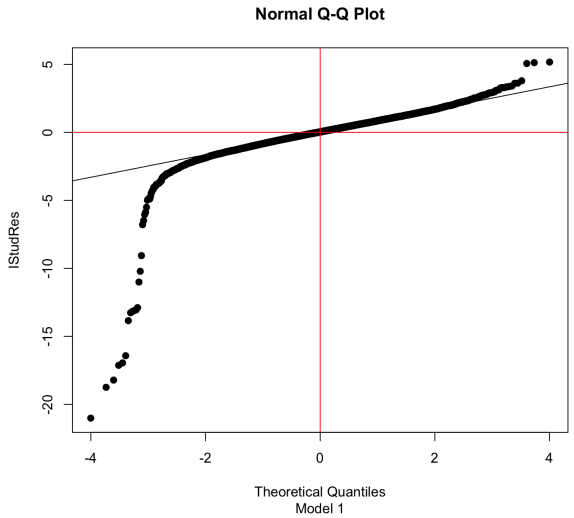
Step 3: Fitting the Model

Since each categorical variable with *c* levels was divided into *c* indicator variables, this makes the variables perfectly collinear since their sum must equal 1. Hence, the 1st level represented by the 1st indicator variable was chosen to be the base and dropped from the dataset. These variables were: *ir1* (international reputation 1), *weak_foot_1, skills_2* (skill moves 1), and *work_rate_LL* (work rate low-low 1).

Iteration 1

For the 1st iteration, a linear model was fit for the overall rating against all other predictors. Before making any conclusions, the assumptions must be verified, so the following plots were constructed: QQ-plot of studentized residuals, internally studentized residuals versus each predictor variable, internally studentized residuals versus response variable, index plot of internally studentized residuals and plots to check influence which are Cook's distance, index plot of leverage values, Hadi's influence, and potential residual plot. These are the plots used in each of the following iterations to verify the assumptions.

(Please refer to the appendix titled according to each iteration, for example, Appendix Iteration 1.)

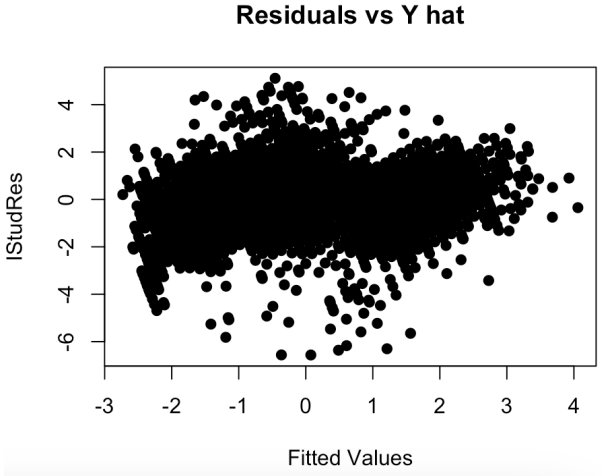


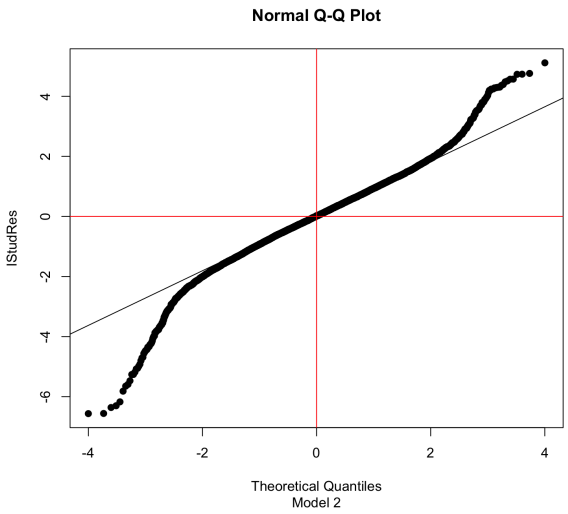
The most prominent issue was that the QQ plot of the residuals (shown above) did not meet the assumption of being a straight line through the origin. Points that were skewed from the straight line were identified; these points were visibly clear outliers in the index plots of several predictor variables. **Hence, these points were removed from the dataset.** The summary of model 1 is:

```
Residual standard error: 0.1553 on 15830 degrees of freedom
Multiple R-squared:  0.976,     Adjusted R-squared:  0.9759
F-statistic: 1.109e+04 on 58 and 15830 DF,  p-value: < 2.2e-16
```

Iteration 2: Outlier Detection

In model 2, the response variable was regressed on all other predictor variables after removing the outliers identified from model 1. The plots for this model showed a great improvement compared to model 1, particularly the QQ-plot which was closer to a straight line than the previous model (shown below), and the plot of residuals versus the fitted values which had fewer outlier points (shown below).

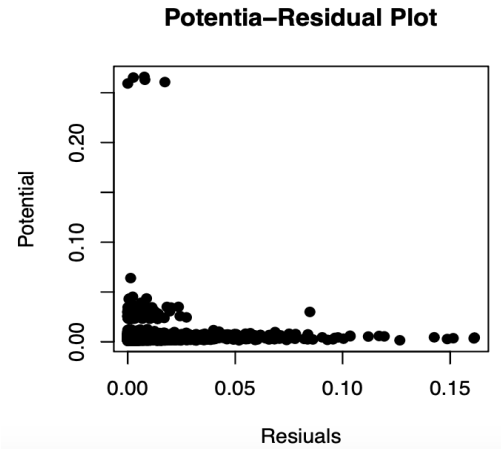




Model 2 also had an increased adjusted R squared. Although the F-statistic decreased, the degrees of freedom also changed so the p-value of the F-statistic is used to compare the models, which remained the same. The summary of model 2 is:

```
Residual standard error: 0.1337 on 15807 degrees of freedom
Multiple R-squared:  0.9822,    Adjusted R-squared:  0.9821
F-statistic: 1.504e+04 on 58 and 15807 DF,  p-value: < 2.2e-16
```

In the plots to validate the assumptions, 5 points were detected as outliers in the x-space due to their very high potential in the potential-residual plot (shown to the right). These 5 observations were the only ones in the dataset with an international reputation equal to 5, which is the highest; this means that these 5 players had the highest reputation worldwide - the reason for which they were outliers.



Iteration 3

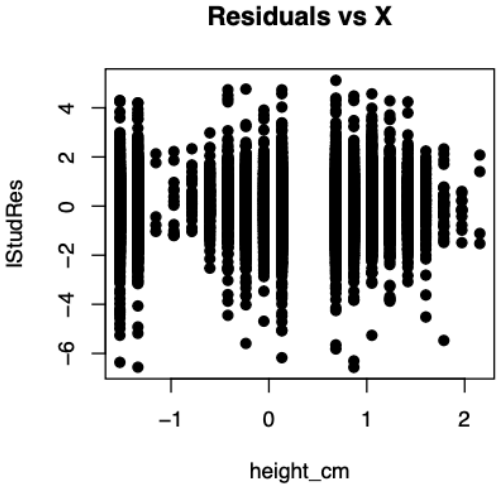
The outliers detected in the x-space from iteration 2 were removed from the dataset while fitting model 3. Since the omitted observations were the only ones with variable *ir5*=1, the new dataset has all values of *ir5*=0, so this attribute was also removed from the dataset. *Overall rating* was fit against all other predictors. This iteration resulted in no difference in comparison to the previous one; all outcomes and graphs were almost identical. It was also visible from the influence plots (Cook's distance, leverage values, Hadi's influence, potential-residual plot) that new points appeared as outliers; this technique of omitting observations will result in an outlier fountain. Therefore, these observations and the variable *ir5* should not be omitted from the dataset and we will continue working with model 2. The coefficient of *ir5* in the model is positive, so it is the increase in overall rating due to having an international reputation equal to 5.

It has been noticed as well that the observations are divided into clusters since the plot of residuals versus *height* was divided into three groups (shown below). Because of the existence of

these clusters, the observations are significant to the dataset and shouldn't be removed as outliers since they represent a certain group of players that share common attributes.

Iteration 4: interaction variables: height

The residuals versus predictors plots of model 2 indicated that some variables required interaction with others because they were divided into clusters unlike the expected outcome: a random scatter of points as the plots of other predictors. One of these variables is *height* (plot shown to the right). In this iteration, interaction variables were added between *height* and *skill moves* (all 3 indicator variables). However, there was no improvement in any of the plots, the p-value value of the F-statistic, and the adjusted R squared increased by 0.001, which is extremely small. Thus, we went back to the previous model - model 2.



Iterations 5, 6, 7, 8

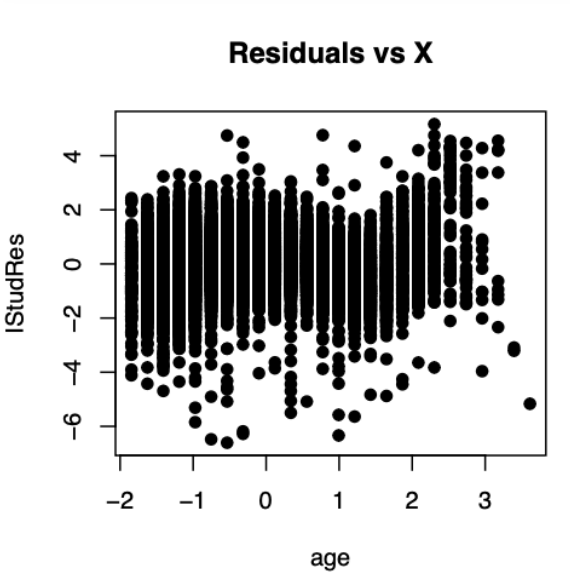
In the next 4 iterations, interaction variables were added to model 2 between *height* and all the categorical variables other than with *skill moves* as it was added in model 4. Each category, *preferred foot, international reputation, weak foot, and work rate,* were added in an iteration respectively. As in the previous model, there was no improvement to the results, so none of these interaction variables were needed. Interaction variables between *age* and other numeric variables were also added to the model in several iterations (these were omitted from the R script to avoid it being very long), yet the plots and the statistics did not enhance.

Iteration 9

Another proposed solution was to create a new categorical variable to differentiate between the different clusters of points in the plot of residuals versus *height*, which are also present in the index plot of *height*. An indicator variable *height_bin* was created; observations with height less than 0.5 were represented as 0 and the rest 1. The overall rating of the players was regressed on all the predictors including the new variable, but there was no change in any of the plots or any of the statistics for this model. This variable was also multiplied by many other predictors to create interaction variables in multiple iterations (these were omitted from the R script to avoid it being very long), yet no improvement occurred to the model, so they were removed and the *height_bin* variable created.

Iteration 10: interaction variables: age

The variable *age* also needed interaction with other variables as clear for the plot of residuals versus this predictor (shown to the right). We decided to interact *value_euro* with *age*, for it is expected that as a player's age increases, they become more experienced, so their value increases. Adding this variable indeed enhanced the model, although not to a great extent. The plots of the residuals versus age and residuals versus fitted values became more random, and the new summary for the model is:




```
Residual standard error: 0.1238 on 15806 degrees of freedom
Multiple R-squared:  0.9847,    Adjusted R-squared:  0.9847
F-statistic: 1.727e+04 on 59 and 15806 DF,  p-value: < 2.2e-16
```

Iteration 11

After trying several interaction variables between *age* and the other predictors (these iteration were omitted from the R script to avoid it being very long), it was found that adding an interaction variable between *age* and *potential* enhanced the adjusted R squared while ensuring the same p-value of the F-statistic and the randomness of the scatter plots of the residuals versus the predictors and the fitted values. The summary for model 11 is:

```
Residual standard error: 0.1117 on 15805 degrees of freedom
Multiple R-squared:  0.9876,    Adjusted R-squared:  0.9875
F-statistic: 2.091e+04 on 60 and 15805 DF,  p-value: < 2.2e-16
```

Iteration 12: Principle component regression

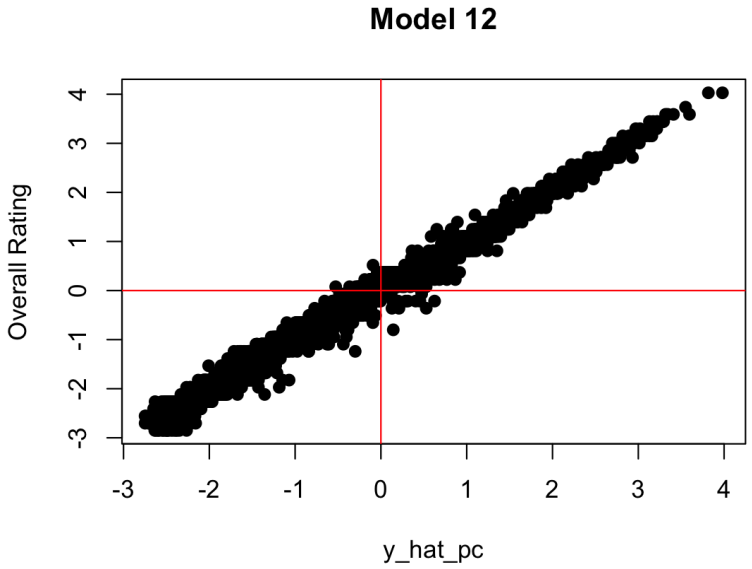
While validating the assumptions before fitting the model, it was visible that the data contains collinearity due to the similarity of point patterns in the index plots of the predictors. To check for collinearity, the variance inflation factor method was used. Many VIFs were very large (greater than 10) indicating the existence of collinearity. Next, the condition indices were computed; the condition number was 56.36, which is very large, and a total of 12 indices were greater than 10, so there are 12 sets of collinear variables in the dataset. This violates the assumption that the predictors are independent from each other.

To solve this problem, principle component regression was used. In model 12, the *overall rating* was regressed on the new columns which were all orthogonal to one another. After fitting this model, 10 columns appeared to be insignificant due to having a p-value greater than 0.05. These columns were removed, and *overall rating* was fit against all the adjusted columns. Unlike expected, the adjusted R squared and the p-value of the F-statistic remained the same (shown below). Yet, this is reasonable since the adjusted R squared is already extremely high (98.75%), and all the coefficients are now significant.

```
Residual standard error: 0.1118 on 15815 degrees of freedom
Multiple R-squared:  0.9876,    Adjusted R-squared:  0.9875
F-statistic: 2.51e+04 on 50 and 15815 DF,  p-value: < 2.2e-16
```

Conclusion (Final Model Interpretation)

According to the model of *overall rating* versus the adjusted predictor variables shown above (model 12), the model is almost perfectly linear since it is a straight line that passes through the origin with slope 1 (graph shown below). This means that all predictors used are significant, which eliminates the problem of collinearity since the dataset contains more than 50 predictor variables, all affecting the response, *overall rating*.



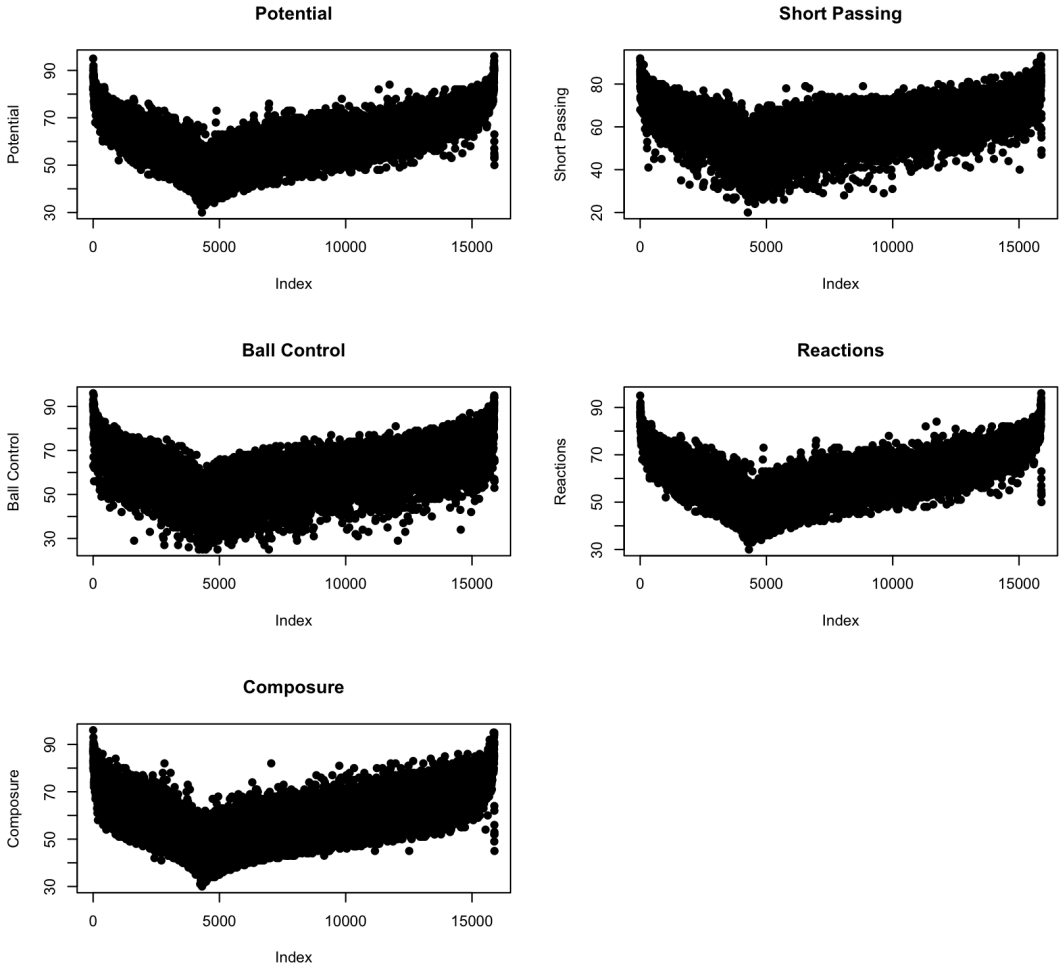
After going through the iterative nature of linear regression and undergoing more than 12 iterations on the dataset, we realized that many attributes affect the players' overall rating. After transforming the variables, creating indicator and interaction variables, and using principal component regression, many performance statistics were a promising additive to the model. For future purposes, the data may need to undergo more analysis before applying linear regression to understand the reason behind the place of the observations that made the response variable resemble a necklace-like pattern and prepare the data even further to include , for example, a unique playing position for each player instead of the 3 or 4 positions that were labelled for each

player in the dataset at first. Positions could've helped identify the clusters that exist in the dataset and would've facilitated the whole regression process.

Appendix A

Predictor Index Plots

Index Plots that show necklace-like pattern similar to the response (overall rating) variable's index plot.

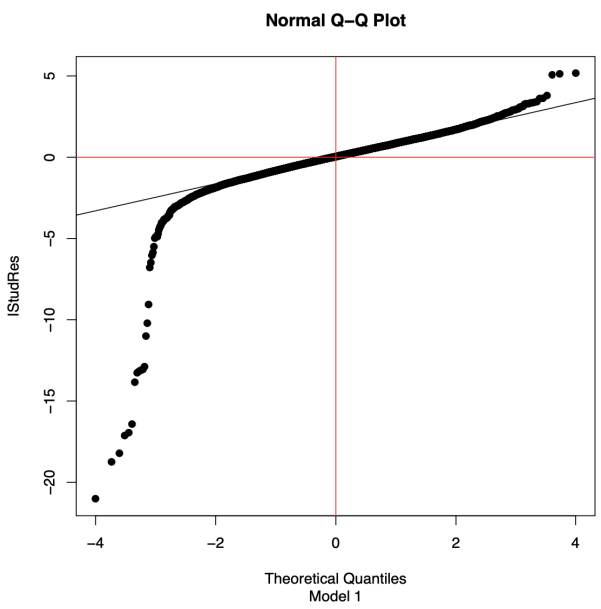


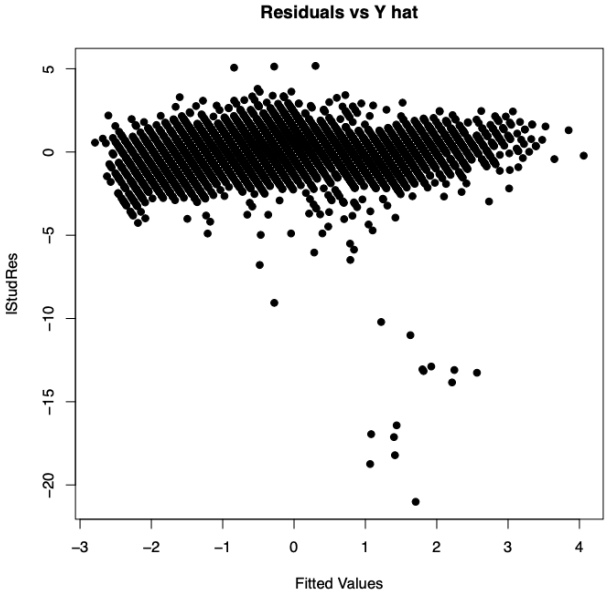
Appendix B

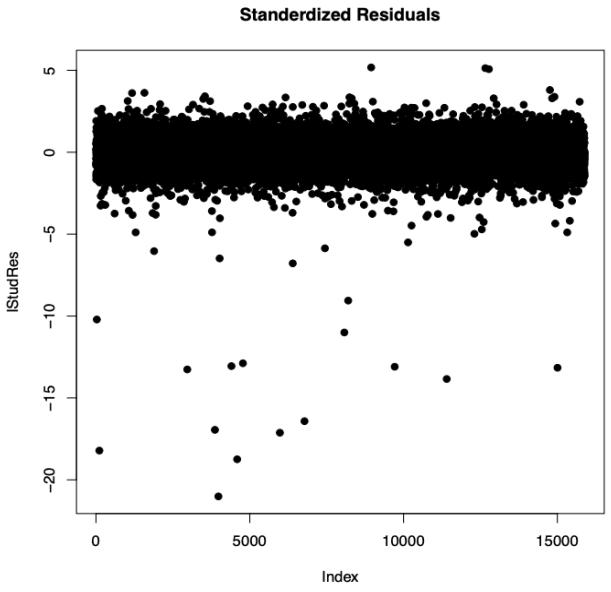
This appendix contains the following plots for each iteration (1-11):

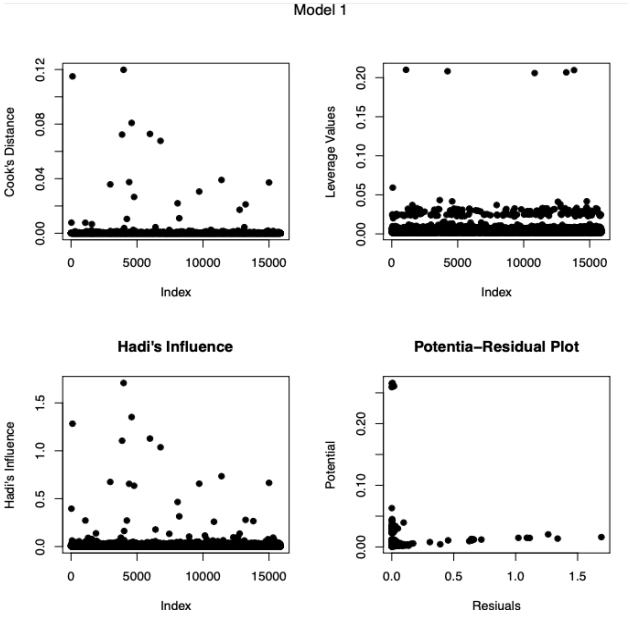
1. QQ-Plot of studentized residuals
2. Plot of studentized residuals versus the fitted values
3. Index plot of studentized residuals
4. Influence plots: index plot of Cook's distance, index plot of leverage values, index plot of Hadi's influence, and the potential-residual plot

Model 1

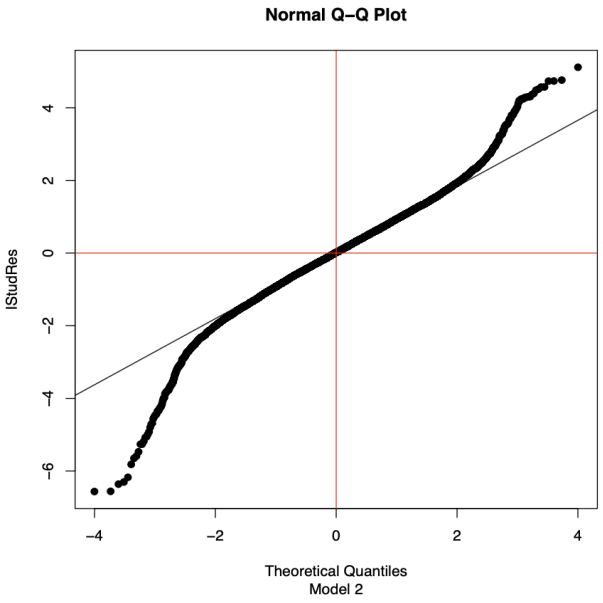

Normal Q–Q Plot
lStudRes
Theoretical Quantiles
Model 1

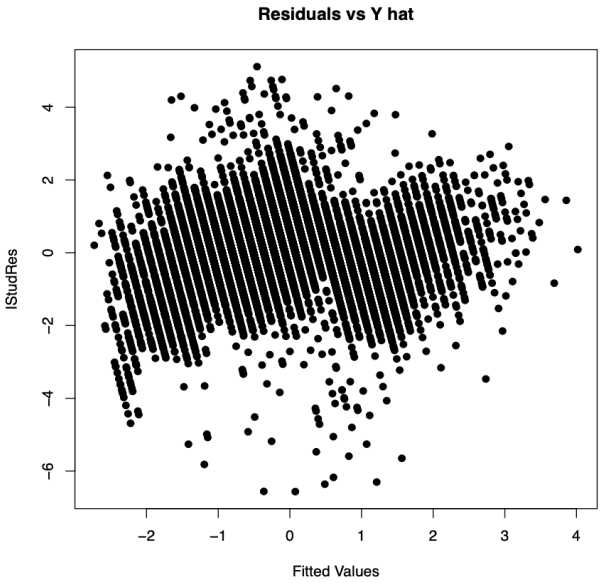

Residuals vs Y hat
lStudRes
Fitted Values

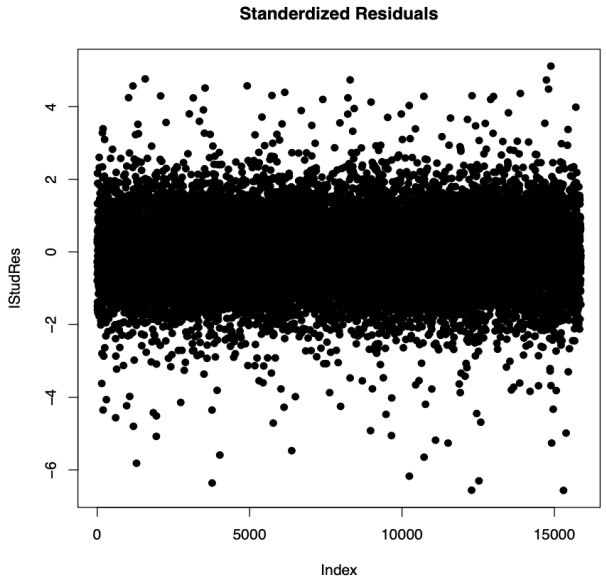

Standerdized Residuals
lStudRes
Index

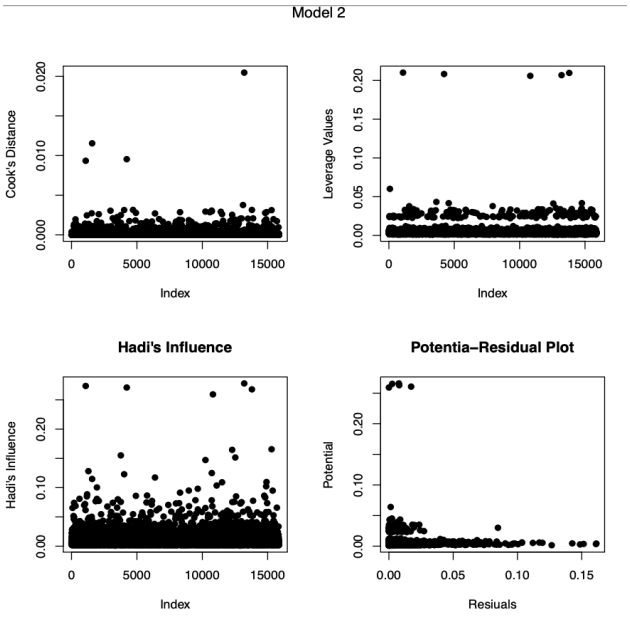

Model 1
Cook's Distance
Index
Leverage Values
Index
Hadi's Influence
Hadi's Influence
Index
Potentia–Residual Plot
Potential
Resiuals

Model 2

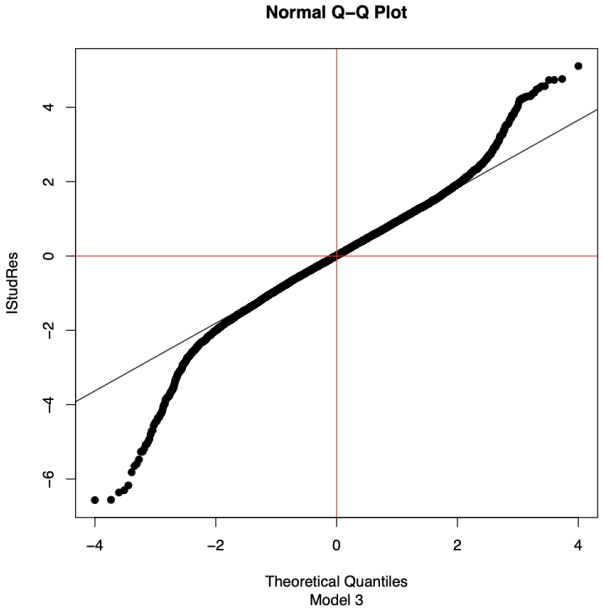

Normal Q–Q Plot
lStudRes
Theoretical Quantiles
Model 2

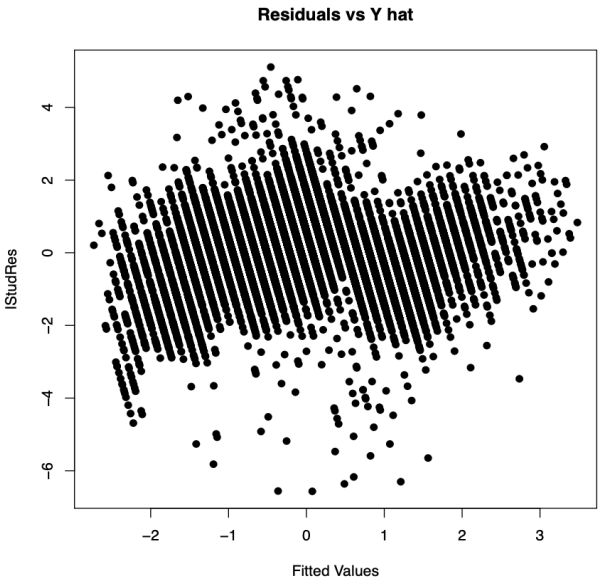

Residuals vs Y hat
lStudRes
Fitted Values

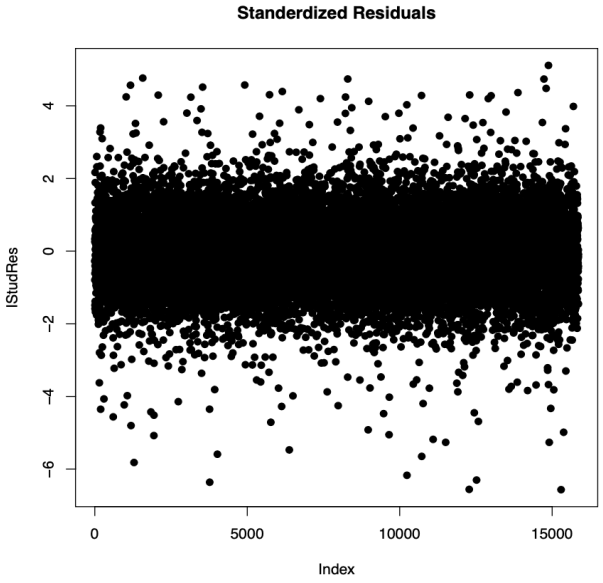

Standerdized Residuals
lStudRes
Index

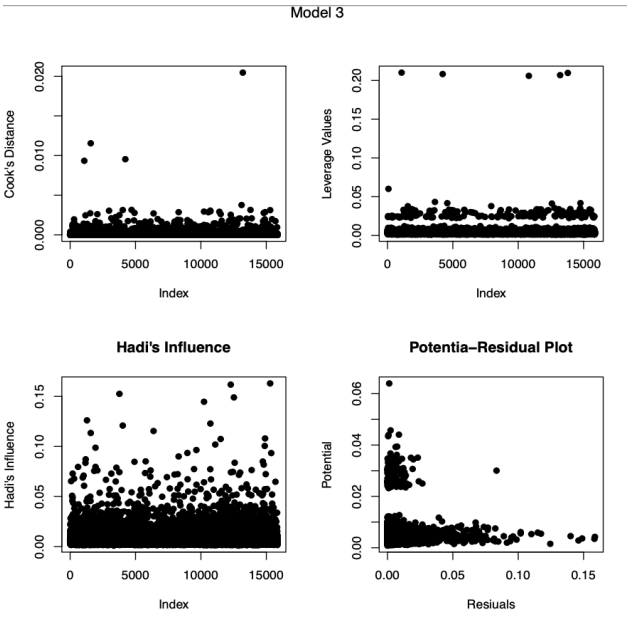

Model 2
Cook's Distance
Index
Leverage Values
Index
Hadi's Influence
Hadi's Influence
Index
Potentia–Residual Plot
Potential
Resiuals

Model 3

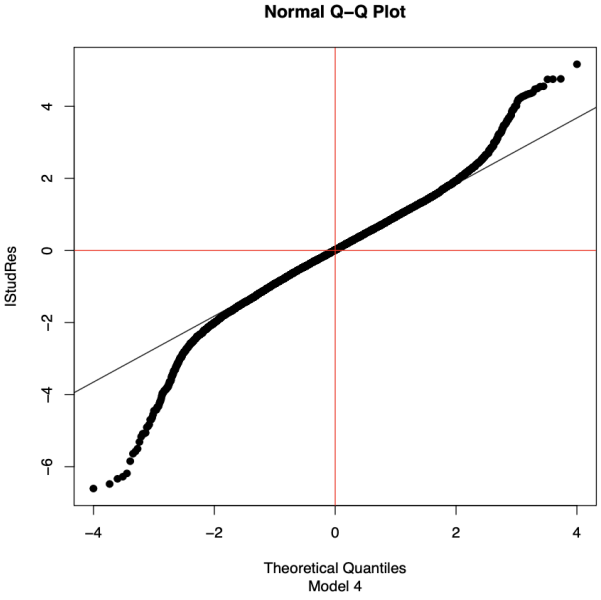

Normal Q–Q Plot
lStudRes
Theoretical Quantiles
Model 3

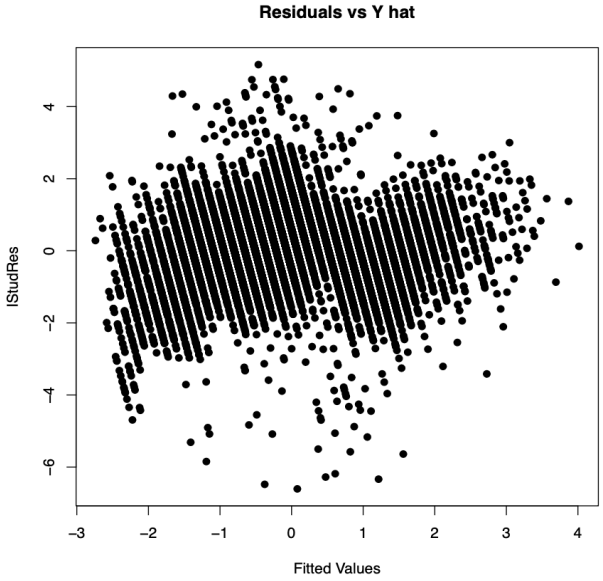

Residuals vs Y hat
lStudRes
Fitted Values

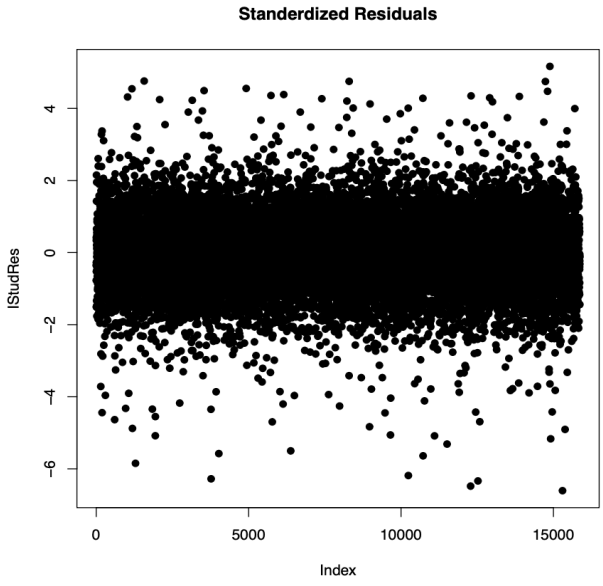

Standerdized Residuals
lStudRes
Index


Model 3
Cook's Distance
Index
Leverage Values
Index
Hadi's Influence
Hadi's Influence
Index
Potentia–Residual Plot
Potential
Resiuals

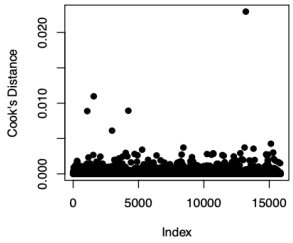
Model 4

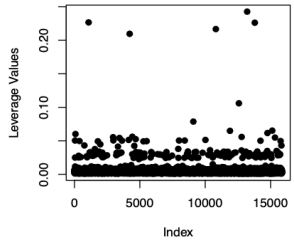


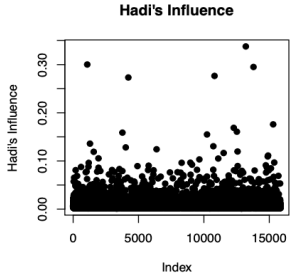


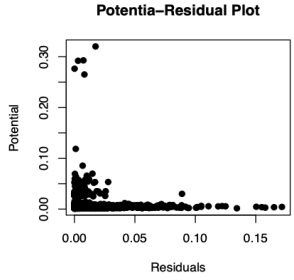


Model 4

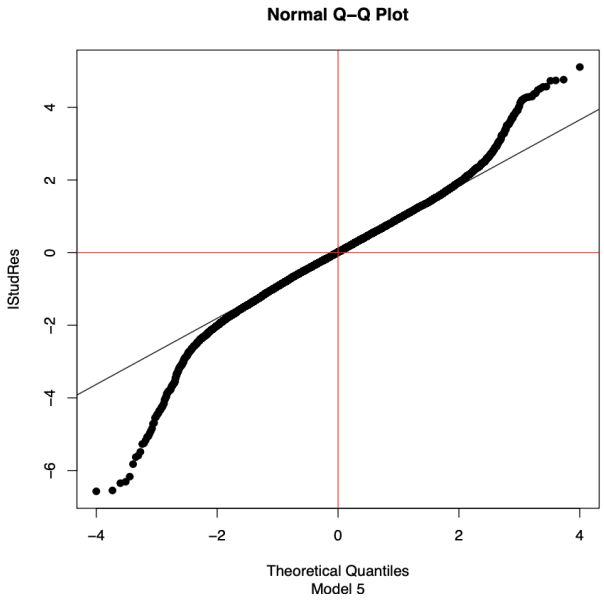


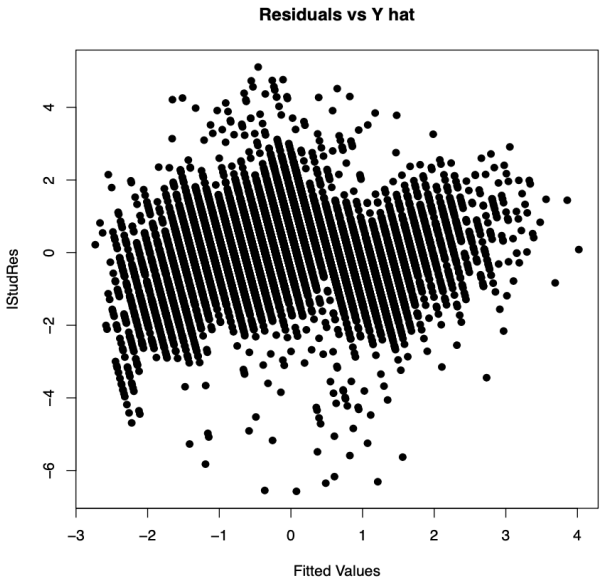


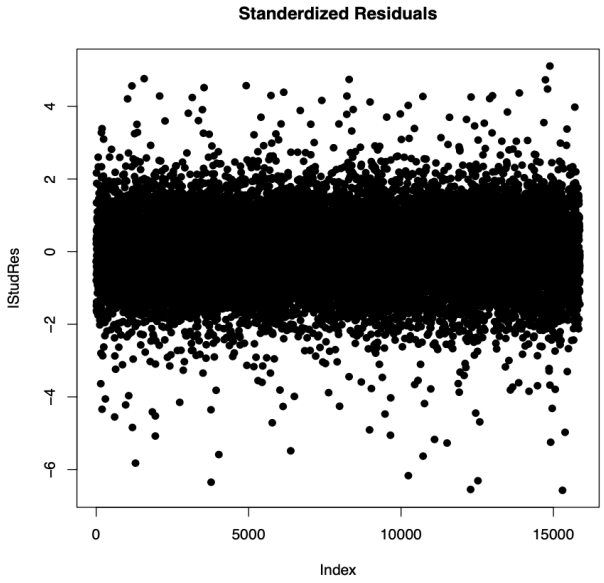


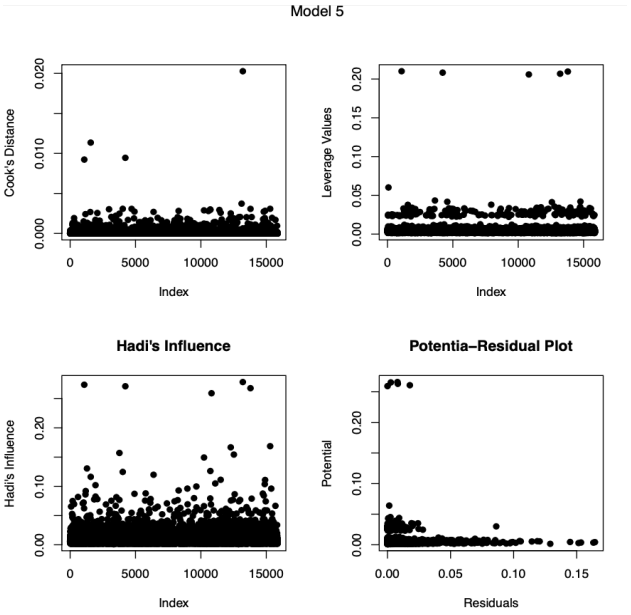


Model 5

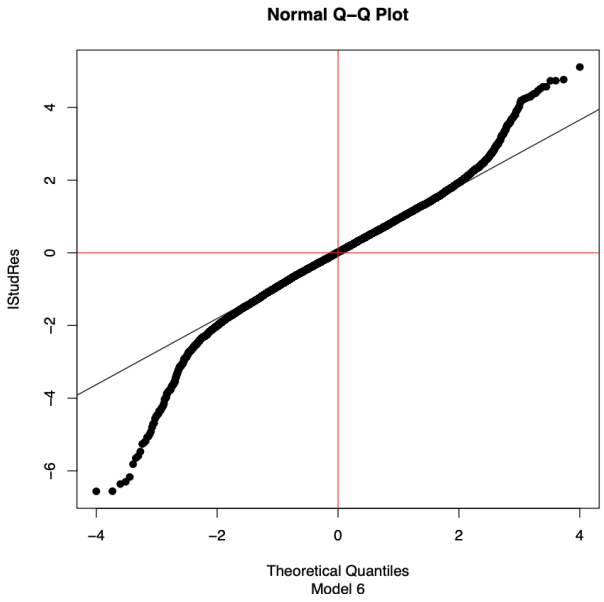

Normal Q–Q Plot
lStudRes
Theoretical Quantiles
Model 5

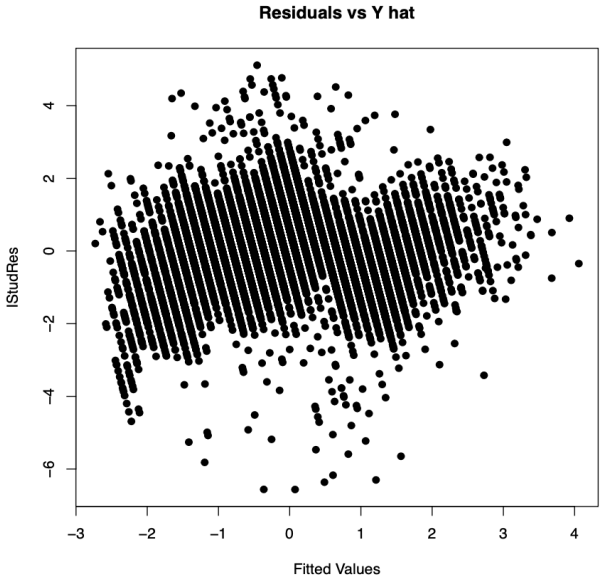

Residuals vs Y hat
lStudRes
Fitted Values

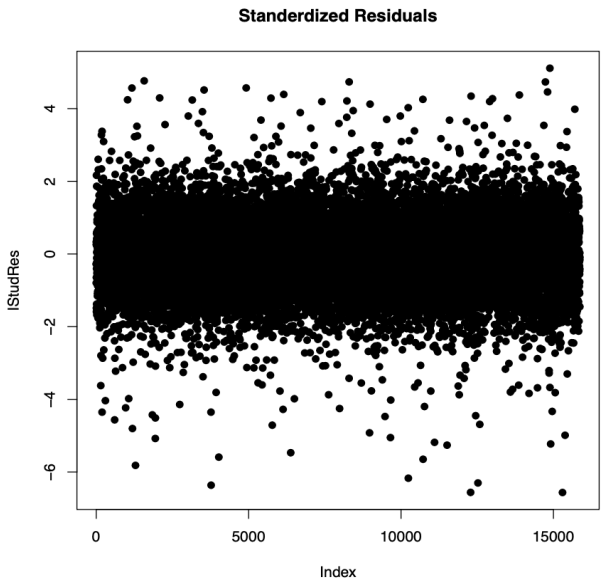

Standerdized Residuals
lStudRes
Index

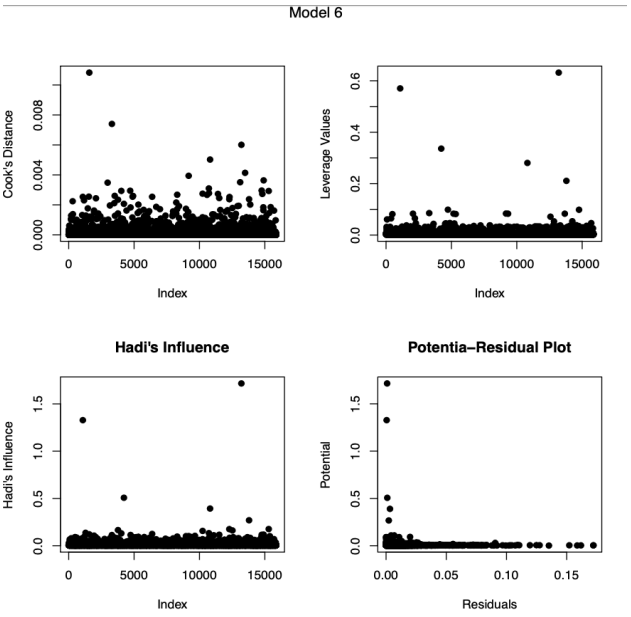

Model 5
Cook's Distance
Index
Leverage Values
Index
Hadi's Influence
Hadi's Influence
Index
Potentia–Residual Plot
Potential
Residuals

Model 6

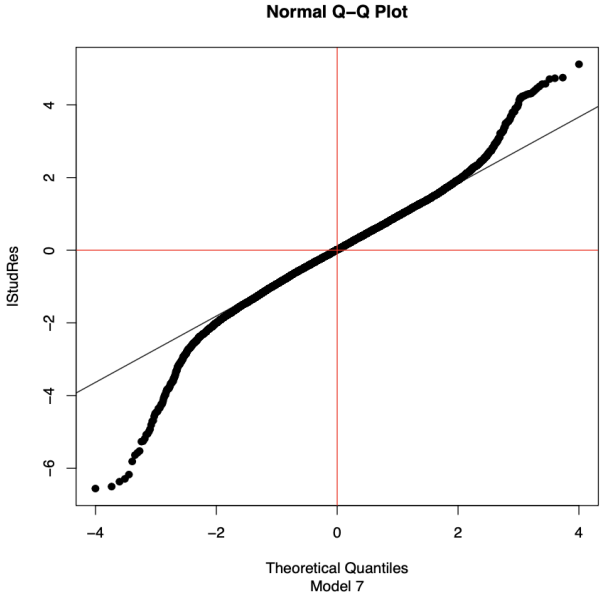

Normal Q–Q Plot
lStudRes
Theoretical Quantiles
Model 6

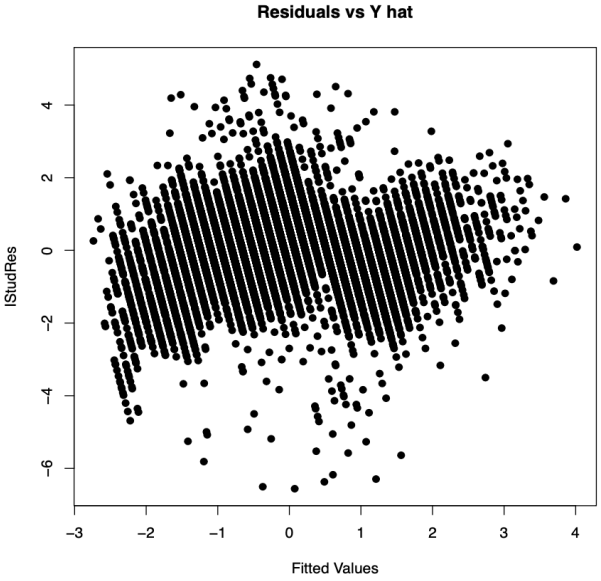

Residuals vs Y hat
lStudRes
Fitted Values

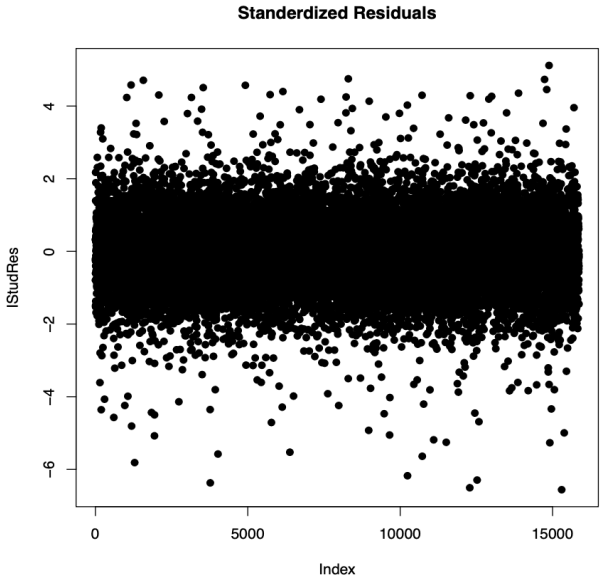

Standerdized Residuals
lStudRes
Index

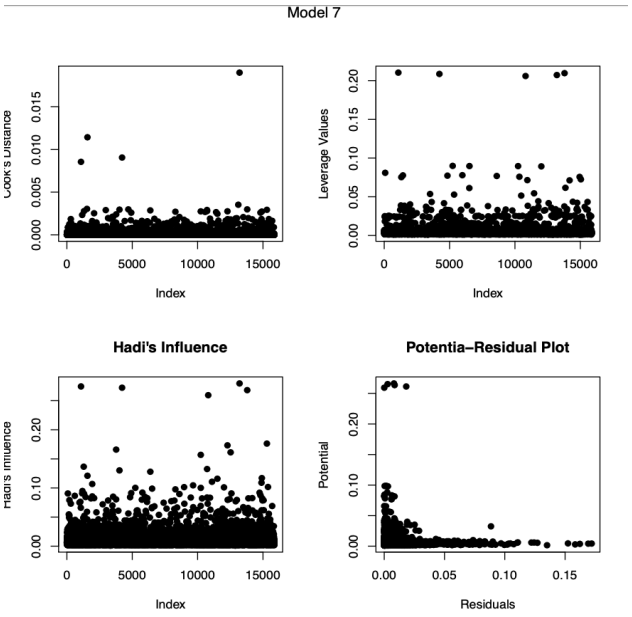

Model 6
Cook's Distance
Index
Leverage Values
Index
Hadi's Influence
Hadi's Influence
Index
Potentia–Residual Plot
Potential
Residuals

Model 7

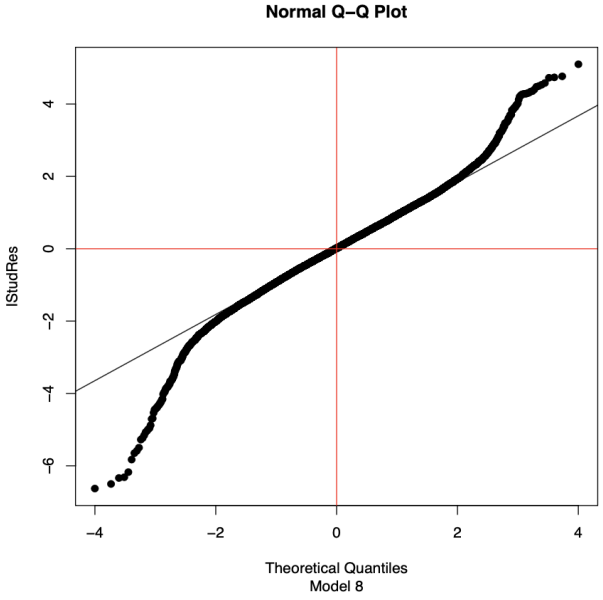

Normal Q–Q Plot
lStudRes
Theoretical Quantiles
Model 7

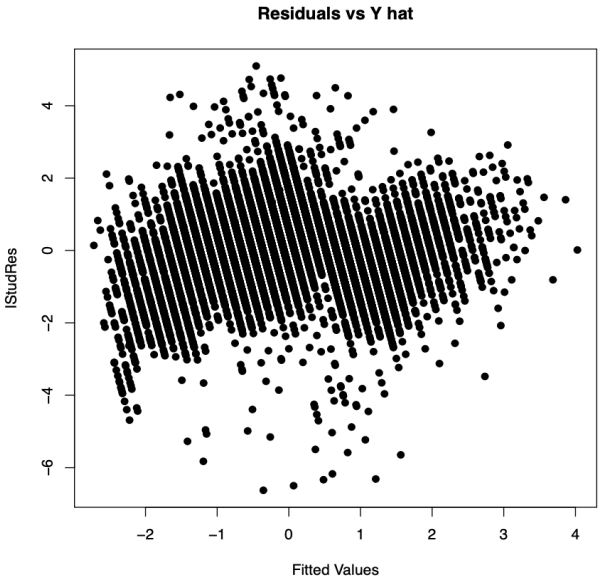

Residuals vs Y hat
lStudRes
Fitted Values

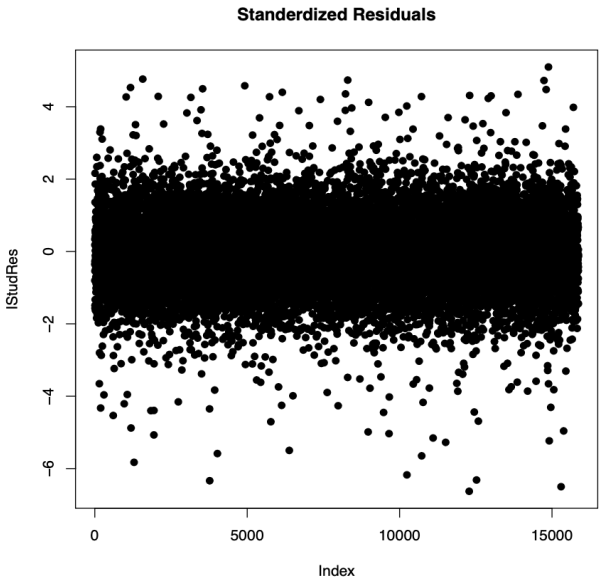

Standerdized Residuals
lStudRes
Index

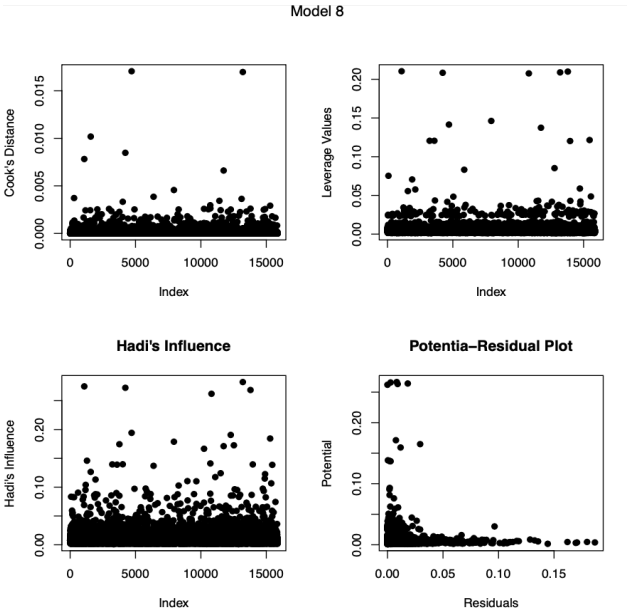

Model 7
Index
Index
Hadi's Influence
Index
Potentia–Residual Plot
Potential
Residuals

Model 8

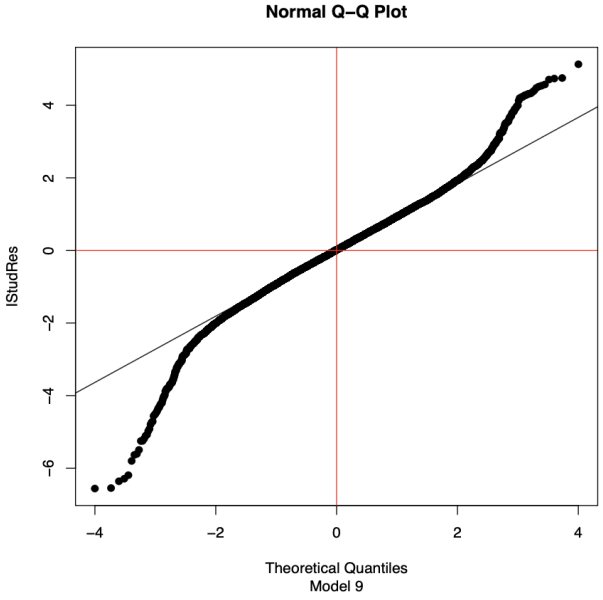

Normal Q–Q Plot
lStudRes
Theoretical Quantiles
Model 8

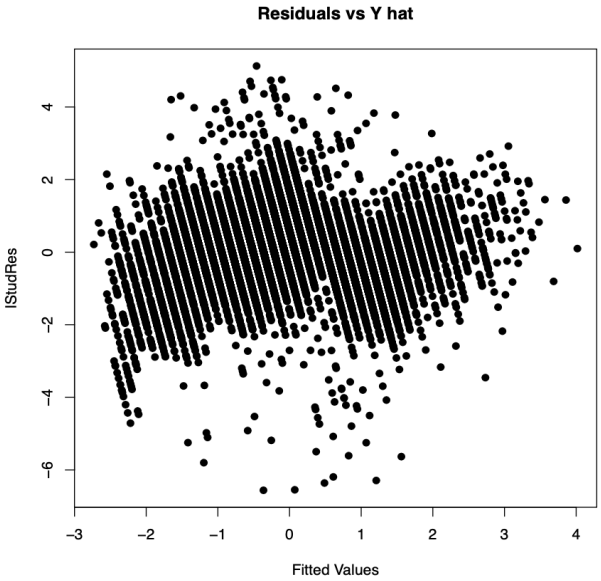

Residuals vs Y hat
lStudRes
Fitted Values

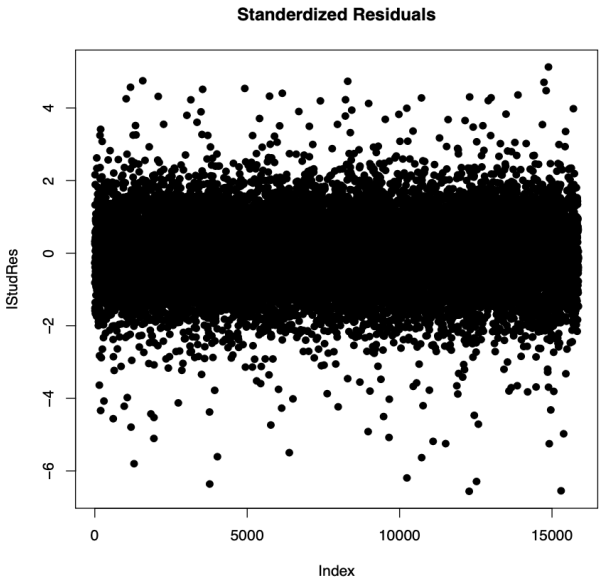

Standerdized Residuals
lStudRes
Index

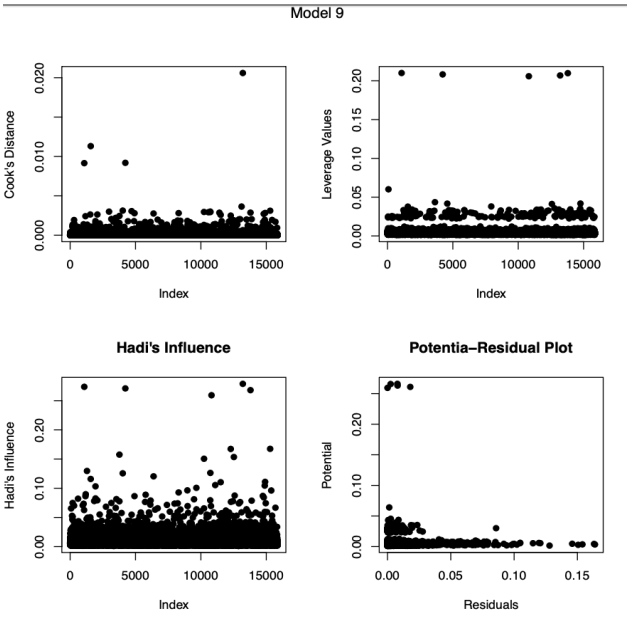

Model 8
Cook's Distance
Index
Leverage Values
Index
Hadi's Influence
Hadi's Influence
Index
Potentia–Residual Plot
Potential
Residuals

Model 9

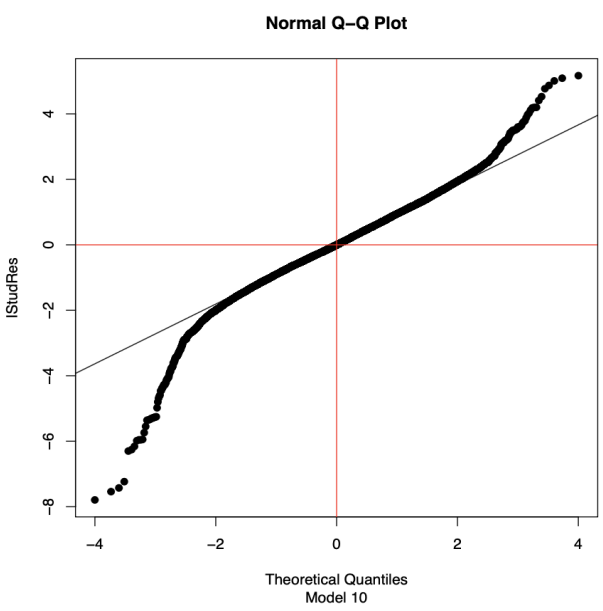

Normal Q–Q Plot
lStudRes
Theoretical Quantiles
Model 9

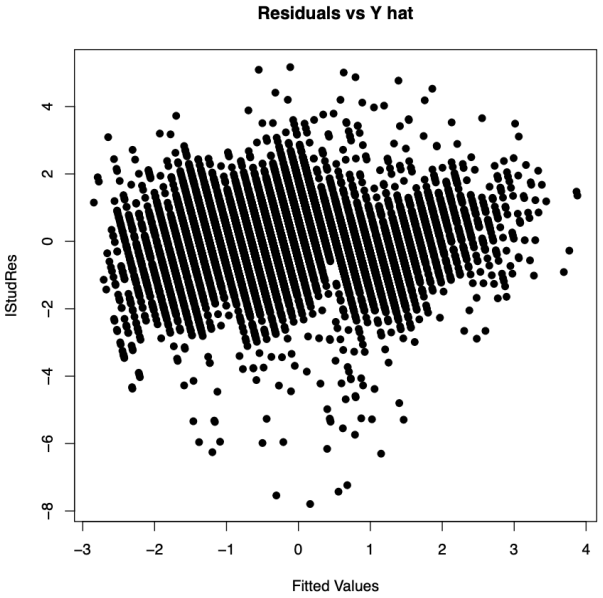

Residuals vs Y hat
lStudRes
Fitted Values

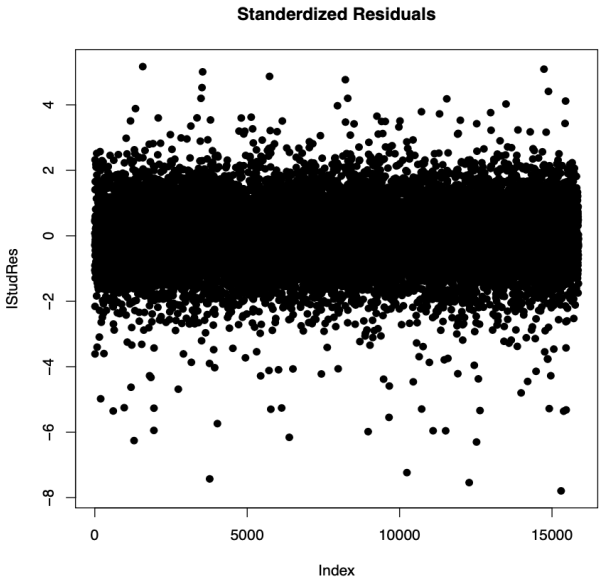

Standerdized Residuals
lStudRes
Index

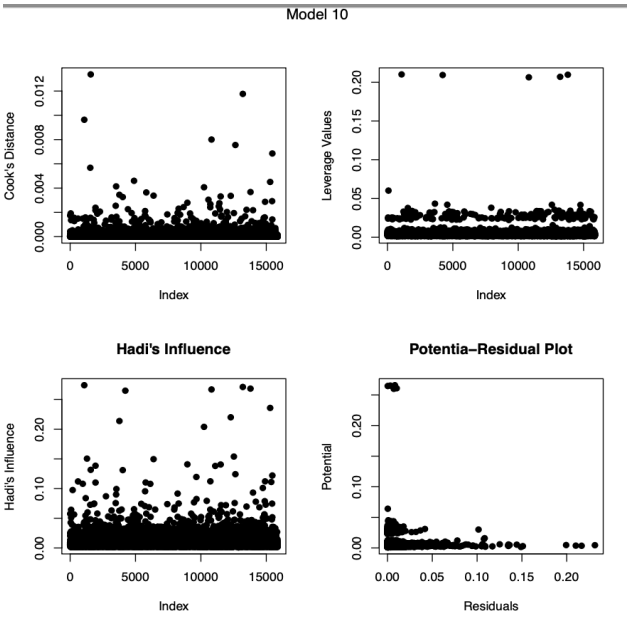

Model 9
Cook's Distance
Index
Leverage Values
Index
Hadi's Influence
Hadi's Influence
Index
Potentia–Residual Plot
Potential
Residuals

Model 10

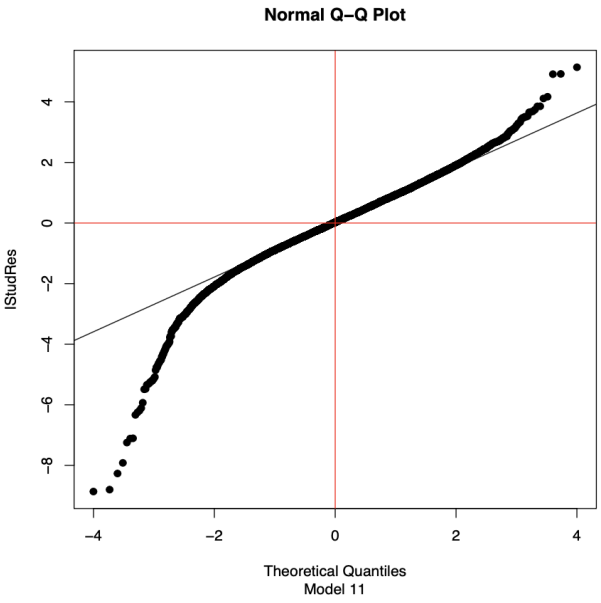

Normal Q–Q Plot
lStudRes
Theoretical Quantiles
Model 10

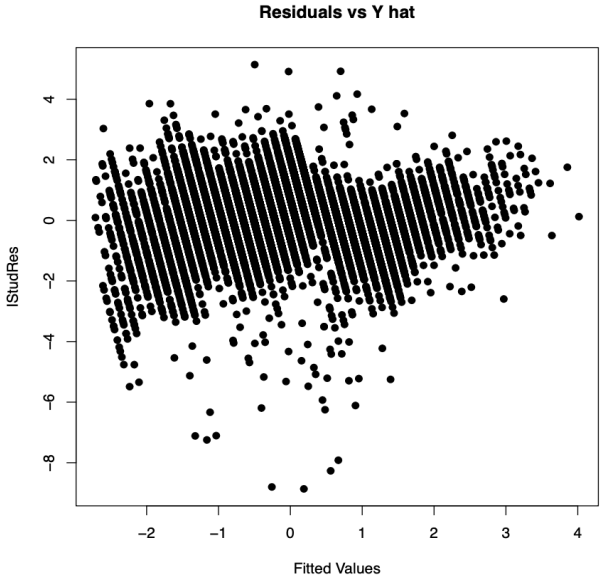

Residuals vs Y hat
lStudRes
Fitted Values

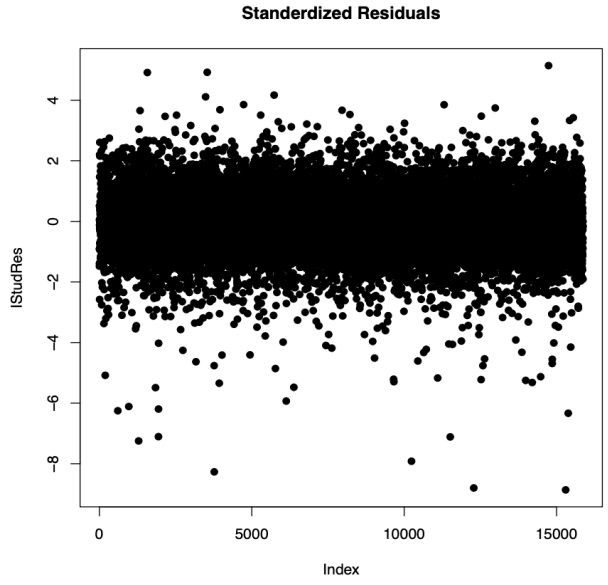

Standerdized Residuals
lStudRes
Index

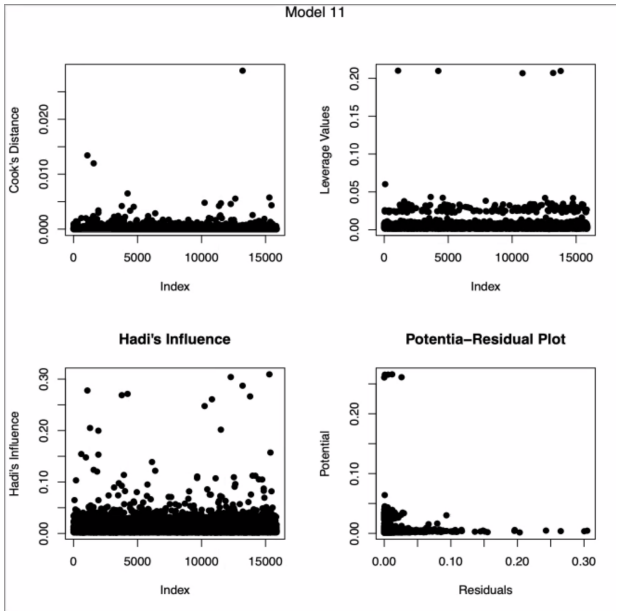

Model 10
Cook's Distance
Index
Leverage Values
Index
Hadi's Influence
Hadi's Influence
Index
Potentia–Residual Plot
Potential
Residuals

Model 11


Normal Q–Q Plot
lStudRes
Theoretical Quantiles
Model 11


Residuals vs Y hat
lStudRes
Fitted Values


Standerdized Residuals
lStudRes
Index


Model 11
Cook's Distance
Index
Leverage Values
Index
Hadi's Influence
Hadi's Influence
Index
Potentia–Residual Plot
Potential
Residuals