Eric Cartaya, Tyler O'Connor, Vincent Pepe Final Project

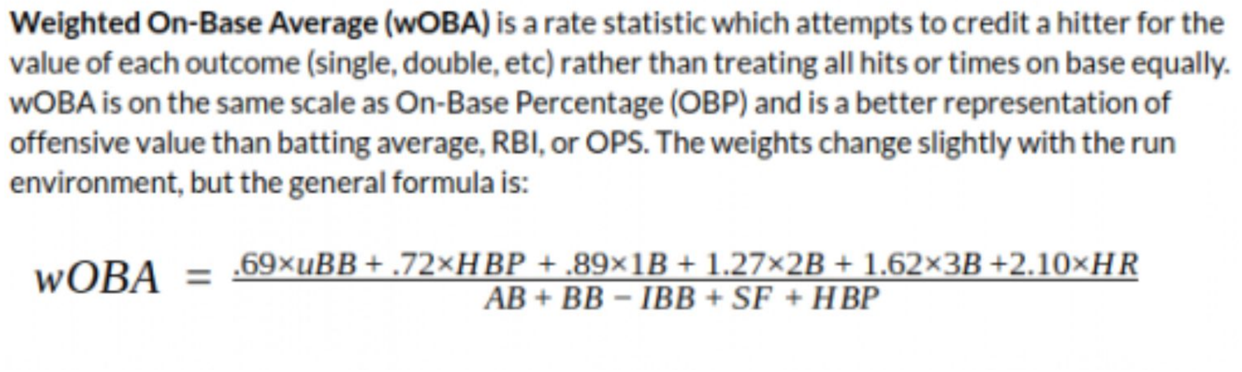
MSDS 596: Regression and Time Series 11/20/19

Analysis of Estimated Weighted On-Base Average In Baseball

I. Introduction:

Currently, MLB baseball is amidst a data-revolution. Beginning in the late 1990s and early 2000s, small-market MLB teams (that could not afford to pay players at the scale that large-market teams could) began to seek out additional ways to gain a competitive edge to even the playing field. Specific organizations, inspired by statisticians like Bill James, began to emphasize “sabermetrics” (the empirical analysis of baseball) to collect and summarize data in various ways to get a better fundamental understanding of the game and individual players' value. Over the last two decades, the data revolution has taken over baseball. Every MLB organization has sabermetric departments consisting of engineers, data scientists, analysts, etc. and each year additional baseball developments occur. We hope to analyze baseball in a similar way that these professionals do to leave us with a better understanding of the game and also hope to derive insights that are not solved.

For our project, the problem we addressed was evaluating what the significant contributors were to “EstimatedWOBA” in the statcast data set. Weighted On-Base Average is a relatively new rate statistic and was created to provide a better indication of an offensive player's “true-value” he is providing his team. A brief description of the statistic and the equation for it can be found below.



Weighted On-Base Average (wOBA) is a rate statistic which attempts to credit a hitter for the value of each outcome (single, double, etc) rather than treating all hits or times on base equally. wOBA is on the same scale as On-Base Percentage (OBP) and is a better representation of offensive value than batting average, RBI, or OPS. The weights change slightly with the run environment, but the general formula is:

$$wOBA = \frac{.69 \times uBB + .72 \times HBP + .89 \times 1B + 1.27 \times 2B + 1.62 \times 3B + 2.10 \times HR}{AB + BB - IBB + SF + HBP}$$

Our intellectual curiosity motivates us to discover whether new-age statistics such as Swing Percentage, Exit Velocity, and pitcher statistics such as Horizontal Location, Vertical Location, Spin Rate, and Velocity have a profound significance on a batter’s ability to provide offensive value to his team.

II. Datasets Utilized.

There are three main datasets that we used. Two of the datasets that we analyzed came from the Major League Baseball owned website: baseballsavant.mlb.com. On this website, we exported data into CSVs to read in R. All of baseball savants data is an extension of the MLB statcast data that we utilized further. However, the data is organized, cleaned up, and summarized in an efficient and easy to digest way for quick analytical purposes. The first dataset was comprised of the velocity information for each pitcher throughout the 2019 season. The dataset contained keys about each pitcher's last name, first name, and a uniquely, identifiable primary-key ID number for each pitcher. It then consisted of information about the average velocity of all the unique pitches that each pitcher threw.

The second dataset that we downloaded for our analysis from Baseball Savant, was given in a similar format as the velocity one. The only difference is that instead of containing velocity information, the dataset consisted of spin rate information about the pitches that each pitcher threw. Due to the similarities, we quickly merged these datasets into a “SpinVelocity” data frame that had information about both the average velocity and average spin rate for each pitch a pitcher threw this season.

The third, and largest dataset we utilized, involved scraping additional data off of MLB’s main Statcast data set in R. To clarify what Statcast is, it is important to know that Statcast data is utilized within every single team’s front office. Statcast is a high-speed, high-accuracy automated tool developed to analyze individual players’ movements, athletic abilities, and performance. Statcast collects this data from TrackMan, a Doppler radar technology which is implemented in all 30 stadiums, and records 3D characteristics of the baseball. The Statcast publicly shared data set was incredibly large because of the fact it tracks every-single play (pitch) in baseball. The dataset consists of 738,029 rows and 90 different variable columns utilized for analysis. We were able to use the package dplyr, to wrangle and clean this data set for easy access and to generate additional variables for analysis.

III. Analysis and Methods

One of the first things we did was mutate a swing column and a miss column. One of the columns in the Statcast data set is called “description”. This column contains information on what happened to the pitch (whether it was a swinging strike, ball, foul ball, out, etc.). We ran an if-else statement through the code to include a column called “swing percentage” and a column called “miss percentage”. We tracked down “Swing Percentage” by summing up “Total Swings” dividing it by “Total Pitches”), tracking “Miss Percentage” followed the same tool (summing up “Total Misses” and dividing it over “Total Pitches”. Due to our experience watching and playing baseball, we understood that by conventional wisdom both of these things should play a role in a batters' effectiveness and we wanted to make sure to include them in the data set to test our model. Other common analytical measures such as exit velocity, batting average, hit distance, barrels, horizontal and vertical location of the ball crossing the plate, we averaged from the Statcast data. Now after grabbing this information, we grouped the data by "Pitcher_ID" and "Pitch_Type" and included all these measures we created or averaged. This allowed us to gauge the performance each pitcher had when throwing specific pitches. We then ignored all other columns that statcast tracked that were not relevant to our project (such as fielder location) and we were left with a smaller Statcast data set containing 2580 observations and 15 variables. We then were able to join and merge the Statcast data sets and Spin Velocity data set together to get a "complete" functional data-set consisting of 2249 observations and 18 variables which incorporated everything we were looking for in our hypothesis. The cumulative performance a batter had towards a specific pitch by a specific pitcher, and the average velocity and spin-rate of the pitch.

One of our first issues we had to figure out was how to filter the amount of "Total Pitches" in the newly-formed "complete" data set. While the number of times a certain pitch a pitcher threw over the season could get as large as 1800 or 2000 approximately, there were many rows in the data set that contained cumulative information about a certain pitch that was thrown under 10 times by a specific pitcher. Using conventional wisdom, we believed that that was not a significant enough number of pitches to analyze the data. We then filtered out all pitches by a specific pitcher that was not throw more than 150 times this year (which is around 5-6 times a game for a starter or 3-4 times a game for a full time reliever). This compressed our data set to 1350 observations.

One of our largest problems when working with these data sets was model selection. The code itself may not show it, but there was a lot of behind the scenes work done to get this data even readable in R. We used multiple linear regression to shrink our already reduced data set from 18 predictors to nine. We employed various techniques such as hybrid selection, the boxcox method for transformations, outlier analysis using the bonferroni correction, as well as multiple different types of residual analysis. Originally, we wanted to see the comparisons of WOBA by pitch as each pitch moves differently both horizontally and vertically. As we delved deeper into the project we realized that pitch would have to be categorized as a factor variable which we couldn't use as our response variable. After performing our hybrid search with WOBA as our response variable, we realized that pitch type wasn't even a significant predictor in determining a batters WOBA, yet horizontal and vertical movement themselves were very significant. After completing our first hybrid search, we ran the VIF function to check for heteroskedasticity and realized we had 2 variables, total pitches and total swings, that had variances >10, so we removed those predictors and ran a new linear model with just the significant predictors.

IV Results / Transformations:

We then tested for normality using the Shapiro-Wilks, as well as made graphs of the residuals versus fitted values, a histogram of the residuals, and a QQ-plot. From these methods, we determined the data was not normal and needed a transformation. We also ran Cook's distance to test for influential observations and an outlier test using the Bonferroni correction and determined that the three outliers were significant to the model and could not be removed. We then ran a Boxcox test and determined the best transformation to run was a square root transformation. We created a new data frame with just the significant predictors, performed a square root transformation, and re-ran the linear model. All the normality plots looked much better and the p-value for the Shapiro-Wilk test got better, but the data was still not normal. After computing numerous more transformations, we realized that our second model (first transform) was our best model, so we went ahead and did our analysis on that model.

V: Conclusion & Future Projects:

There are various different things we could potentially analyze in the future. Our main limitation when exploring this data was that there simply is still a plethora of privately owned data that MLB teams utilize that the general public does not have access too. This likely caused underlying trends that our model simply was not able to account for.

In the future we might divide the model based on the pitch type seeing how the WOBA would be affected by different variables based on the pitch (i.e. curveballs being more dependent on spin rate). Our current population consisted of all pitches where there was a frequency greater than 150 pitches thrown. As we previously mentioned, perhaps in the future we could just look at the performance of batters facing a certain pitch. Or we could look at statistics that only a pitcher can control (pitch-location, pitch velocity, and spin-rate) and see if that provides any impact to EstimatedAverageWOBA. We just simply did not have access to enough pitcher-centric variables to conduct the experiment like this.

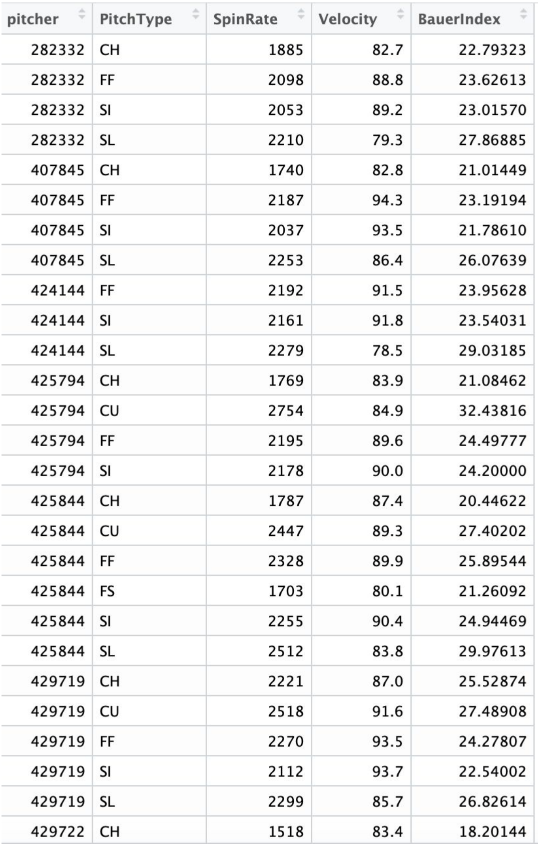
Going off that, we could potentially try to look over the course of several seasons. We explored just data from 2019 in this analysis because we felt it would give us the most consistent set of data to analyze. Now that we are learning Time Series in class, it would be exciting to attempt to analyze trends in WOBA performance from year to year.

References:

Statcast Pitch Movement Leaderboard. (n.d.). Retrieved from

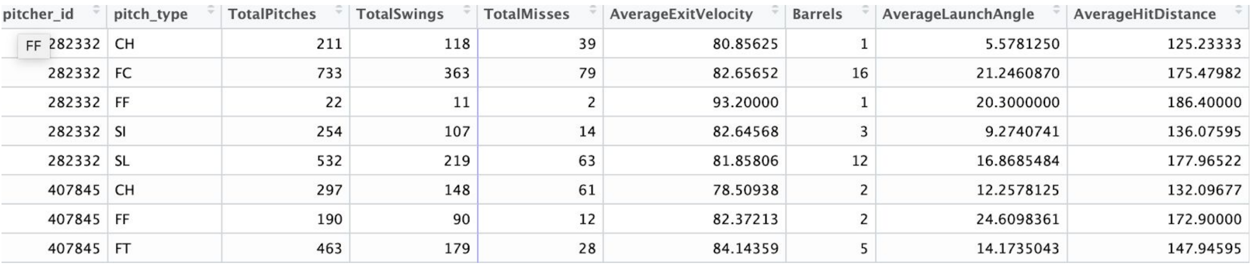
https://baseballsavant.mlb.com/pitch-movement?year=2019&team=&min=q&pitch_type=ALL&hand=&x=diff_x_hidden&z=diff_z_hidden.

The Wrangled "SpinVelocity" Table



pitcher	PitchType	SpinRate	Velocity	BauerIndex
282332	CH	1885	82.7	22.79323
282332	FF	2098	88.8	23.62613
282332	SI	2053	89.2	23.01570
282332	SL	2210	79.3	27.86885
407845	CH	1740	82.8	21.01449
407845	FF	2187	94.3	23.19194
407845	SI	2037	93.5	21.78610
407845	SL	2253	86.4	26.07639
424144	FF	2192	91.5	23.95628
424144	SI	2161	91.8	23.54031
424144	SL	2279	78.5	29.03185
425794	CH	1769	83.9	21.08462
425794	CU	2754	84.9	32.43816
425794	FF	2195	89.6	24.49777
425794	SI	2178	90.0	24.20000
425844	CH	1787	87.4	20.44622
425844	CU	2447	89.3	27.40202
425844	FF	2328	89.9	25.89544
425844	FS	1703	80.1	21.26092
425844	SI	2255	90.4	24.94469
425844	SL	2512	83.8	29.97613
429719	CH	2221	87.0	25.52874
429719	CU	2518	91.6	27.48908
429719	FF	2270	93.5	24.27807
429719	SI	2112	93.7	22.54002
429719	SL	2299	85.7	26.82614
429722	CH	1518	83.4	18.20144

The Unfiltered Grouped "StatCast" Table

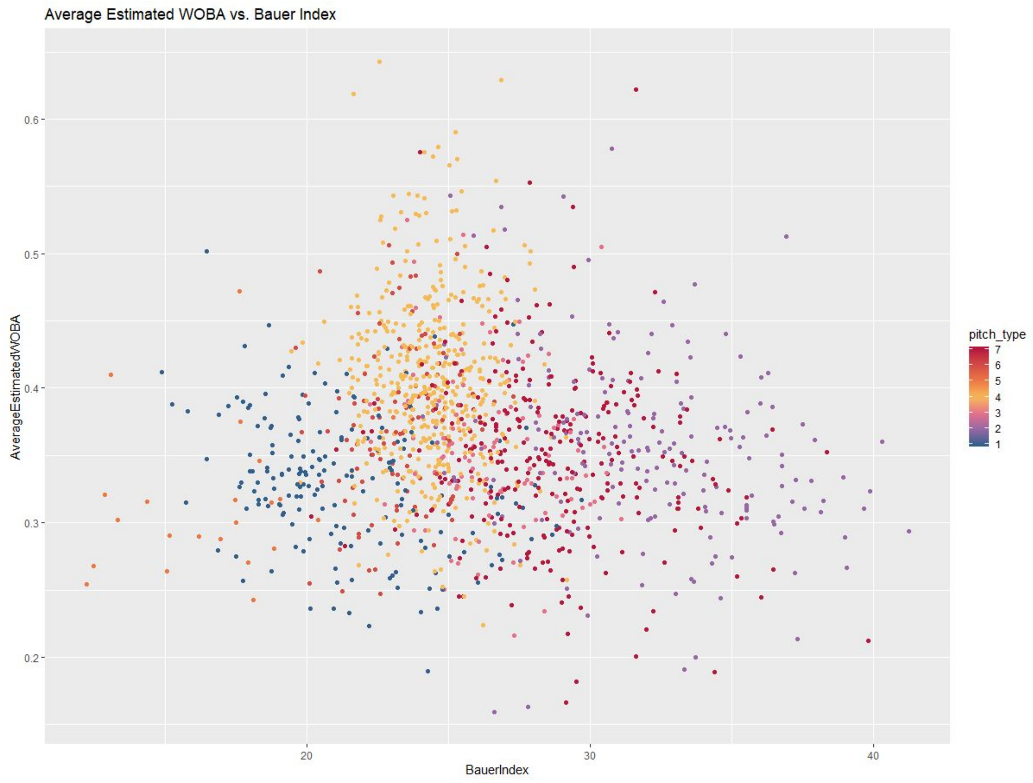


pitcher_id	pitch_type	TotalPitches	TotalSwings	TotalMisses	AverageExitVelocity	Barrels	AverageLaunchAngle	AverageHitDistance
282332	CH	211	118	39	80.85625	1	5.5781250	125.23333
282332	FC	733	363	79	82.65652	16	21.2460870	175.47982
282332	FF	22	11	2	93.20000	1	20.3000000	186.40000
282332	SI	254	107	14	82.64568	3	9.2740741	136.07595
282332	SL	532	219	63	81.85806	12	16.8685484	177.96522
407845	CH	297	148	61	78.50938	2	12.2578125	132.09677
407845	FF	190	90	12	82.37213	2	24.6098361	172.90000
407845	FT	463	179	28	84.14359	5	14.1735043	147.94595

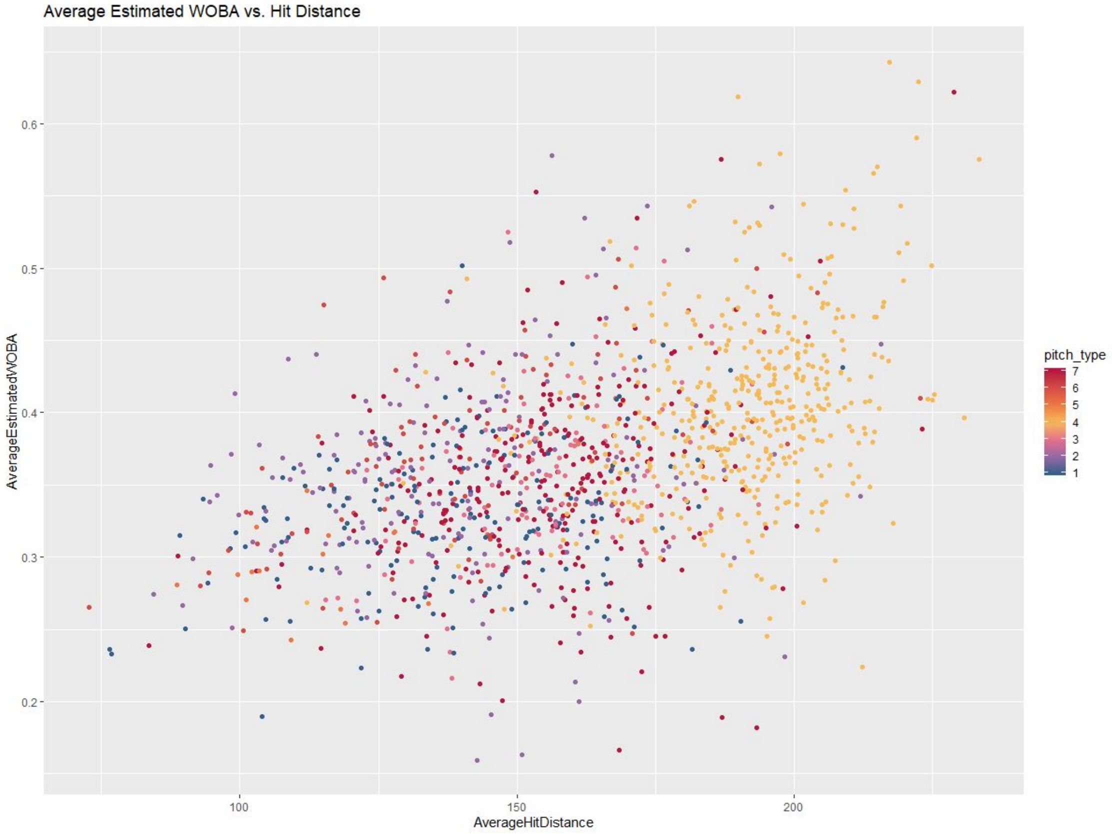
The First Hybrid Select Model

```
##
## Call:
## lm(formula = AverageEstimatedWOBA ~ TotalPitches + TotalSwings +
##     HorizontalLocationBall + AverageExitVelocity + AverageHitDistance +
##     AverageEstimatedBattingAverage + SwingPercentage + MissPercentage +
##     BarrelPercentage + SpinRate + BauerIndex, data = complete2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.060188 -0.010492 -0.000492  0.009666  0.081971
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    4.130e-02  2.101e-02   1.966  0.04951 *
## TotalPitches                  -2.125e-05  1.425e-05  -1.492  0.13605
## TotalSwings                    4.919e-05  2.988e-05   1.647  0.09986 .
## HorizontalLocationBall        -1.077e-02  2.466e-03  -4.367 1.36e-05 ***
## AverageExitVelocity           -5.955e-04  2.297e-04  -2.593  0.00963 **
## AverageHitDistance             2.649e-04  2.102e-05  12.603  < 2e-16 ***
## AverageEstimatedBattingAverage  1.090e+00  1.090e-02  99.998  < 2e-16 ***
## SwingPercentage               -1.524e-01  1.447e-02 -10.533  < 2e-16 ***
## MissPercentage                 1.511e-01  1.217e-02  12.419  < 2e-16 ***
## BarrelPercentage               2.283e+00  6.735e-02  33.905  < 2e-16 ***
## SpinRate                       1.720e-05  4.079e-06   4.218 2.64e-05 ***
## BauerIndex                    -1.590e-03  3.123e-04  -5.093 4.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01662 on 1338 degrees of freedom
## Multiple R-squared:  0.9378, Adjusted R-squared:  0.9373
## F-statistic:  1835 on 11 and 1338 DF,  p-value: < 2.2e-16
```

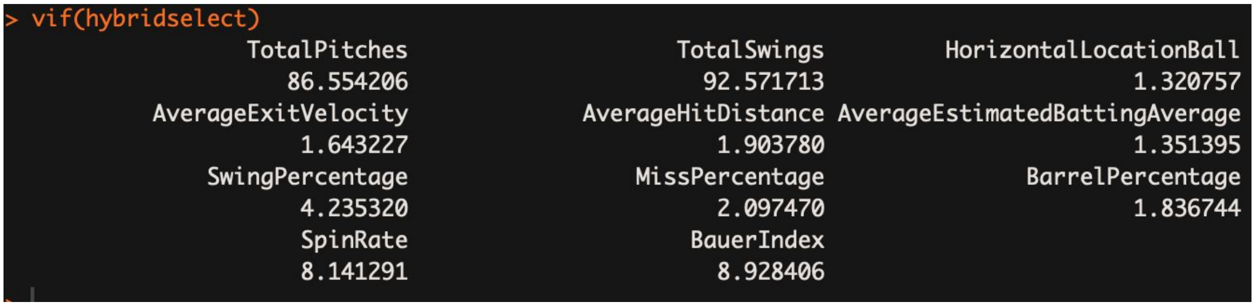
Average Estimated WOBA vs. Bauer Index



Average Estimated WOBA vs. Hit Distance

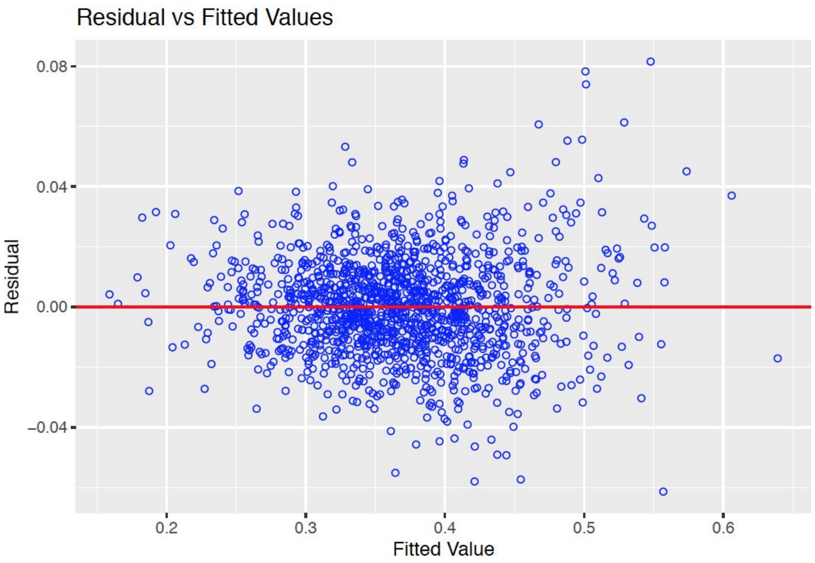


VIF Results Of That Model

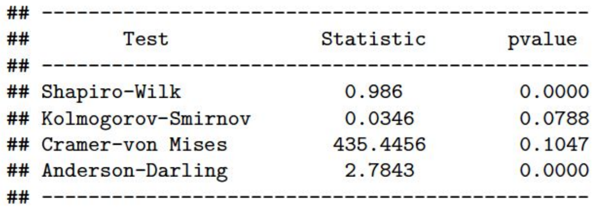


```
> vif(hybridselect)
                  TotalPitches                    TotalSwings         HorizontalLocationBall
                     86.554206                      92.571713                       1.320757
           AverageExitVelocity             AverageHitDistance AverageEstimatedBattingAverage
                      1.643227                       1.903780                       1.351395
               SwingPercentage                 MissPercentage               BarrelPercentage
                      4.235320                       2.097470                       1.836744
                      SpinRate                     BauerIndex
                      8.141291                       8.928406
>
```

Residuals vs. Fitted Values of Model 1

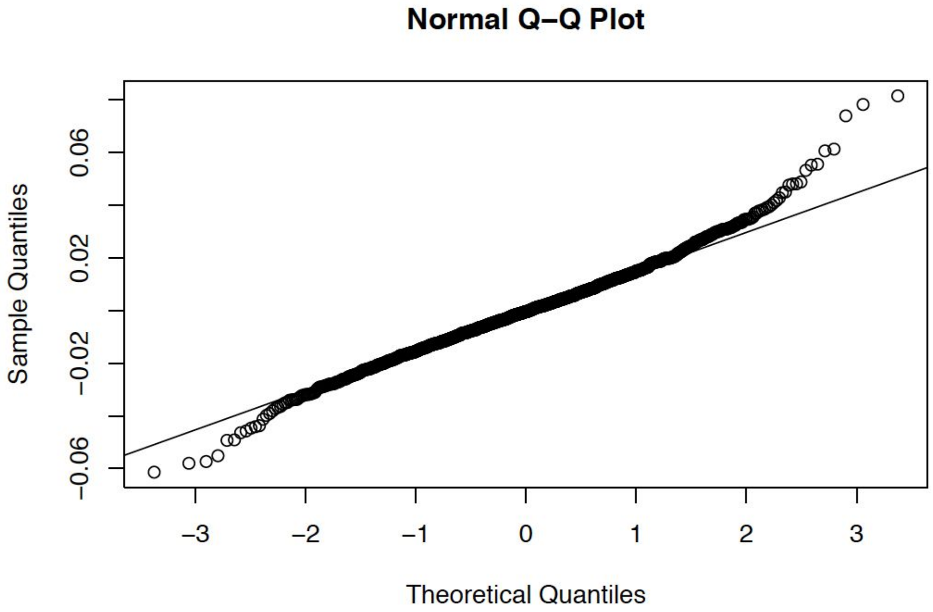


The Normality Test Results

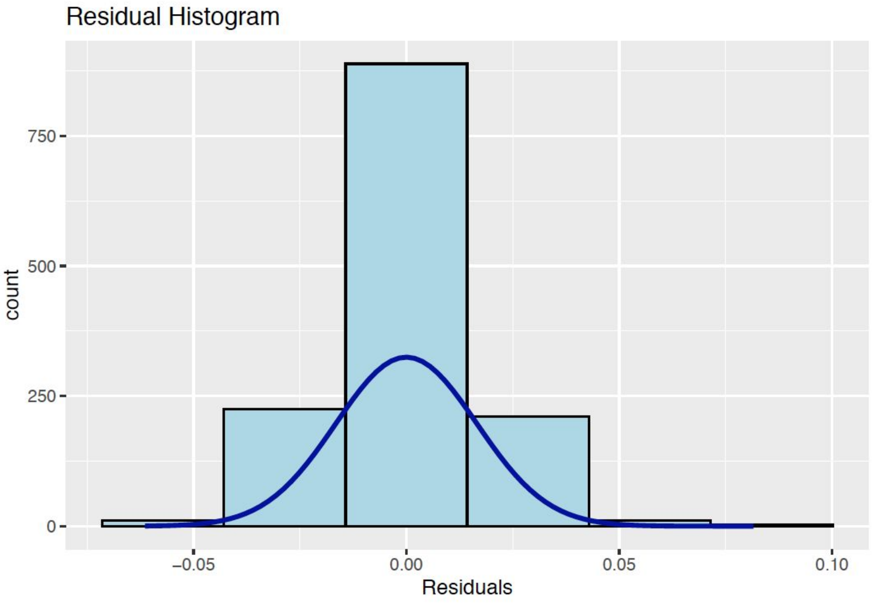


```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.986          0.0000
## Kolmogorov-Smirnov        0.0346         0.0788
## Cramer-von Mises        435.4456         0.1047
## Anderson-Darling          2.7843         0.0000
## -----------------------------------------------
```

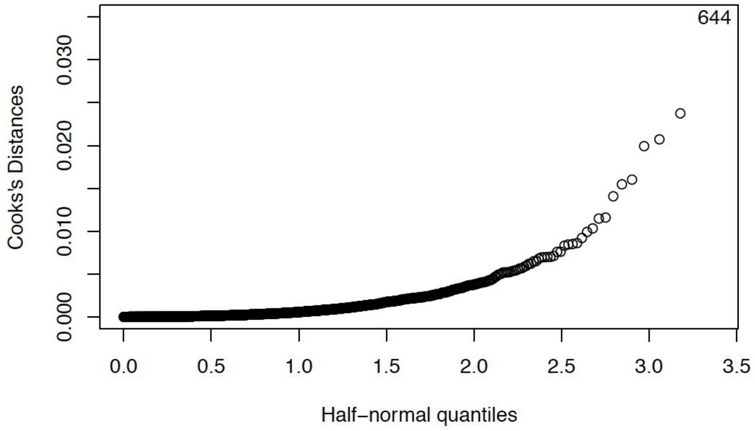
The QQ Plot of Model 1



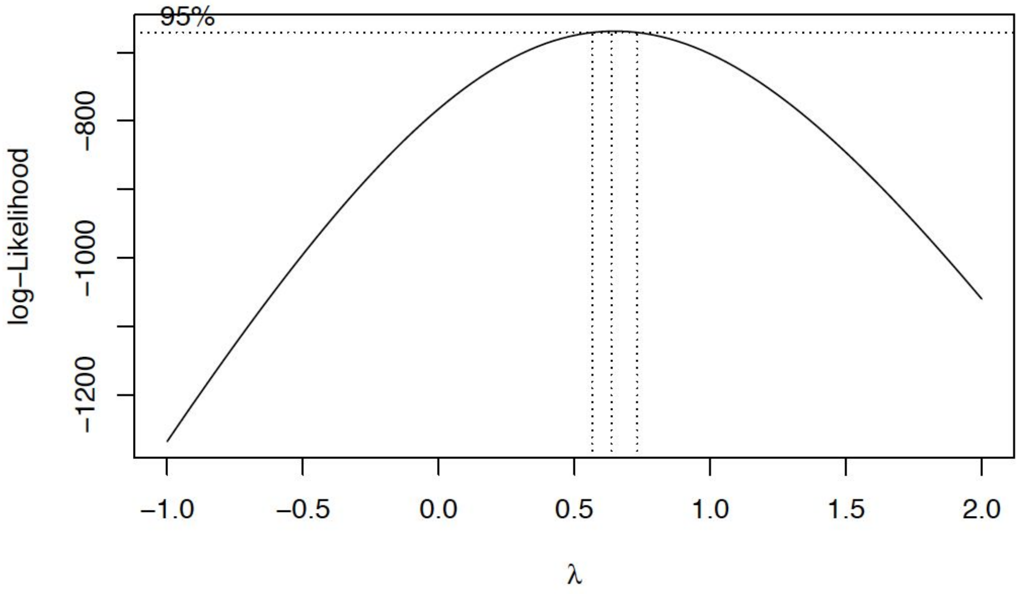
The Histogram of the Residuals for Model 1



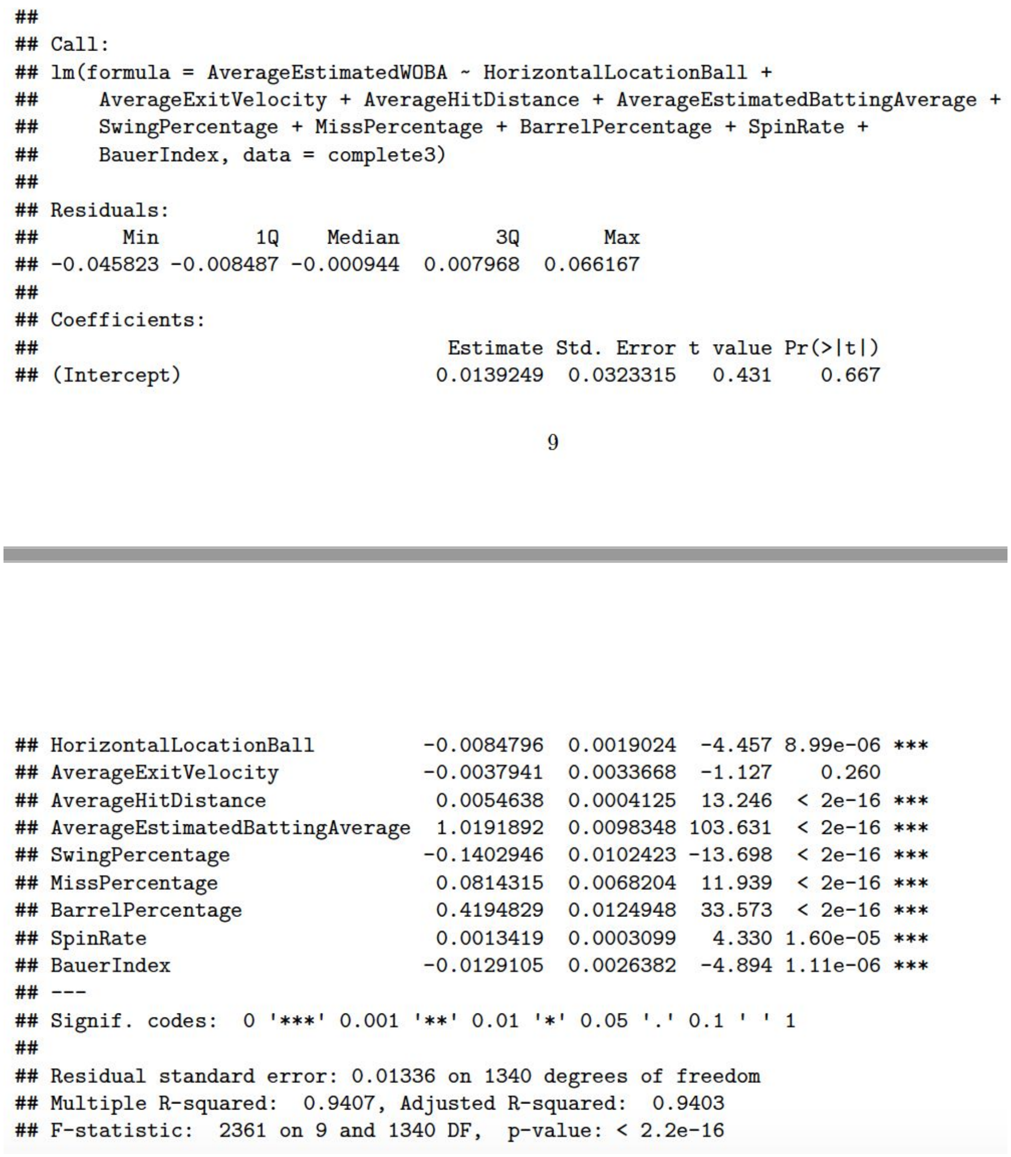
Cook's Distance Test On First Model



Box-Cox Test



Summary of Second Model



```
##
## Call:
## lm(formula = AverageEstimatedWOBA ~ HorizontalLocationBall +
##     AverageExitVelocity + AverageHitDistance + AverageEstimatedBattingAverage +
##     SwingPercentage + MissPercentage + BarrelPercentage + SpinRate +
##     BauerIndex, data = complete3)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.045823 -0.008487 -0.000944  0.007968  0.066167
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.0139249  0.0323315   0.431    0.667
```

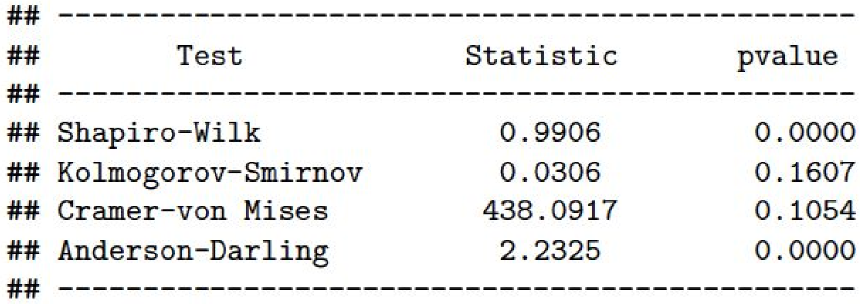
9

```
## HorizontalLocationBall        -0.0084796  0.0019024  -4.457 8.99e-06 ***
## AverageExitVelocity           -0.0037941  0.0033668  -1.127    0.260
## AverageHitDistance             0.0054638  0.0004125  13.246  < 2e-16 ***
## AverageEstimatedBattingAverage  1.0191892  0.0098348 103.631  < 2e-16 ***
## SwingPercentage               -0.1402946  0.0102423 -13.698  < 2e-16 ***
## MissPercentage                 0.0814315  0.0068204  11.939  < 2e-16 ***
## BarrelPercentage               0.4194829  0.0124948  33.573  < 2e-16 ***
## SpinRate                       0.0013419  0.0003099   4.330 1.60e-05 ***
## BauerIndex                    -0.0129105  0.0026382  -4.894 1.11e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01336 on 1340 degrees of freedom
## Multiple R-squared:  0.9407, Adjusted R-squared:  0.9403
## F-statistic:  2361 on 9 and 1340 DF,  p-value: < 2.2e-16
```

VIF of Model 2

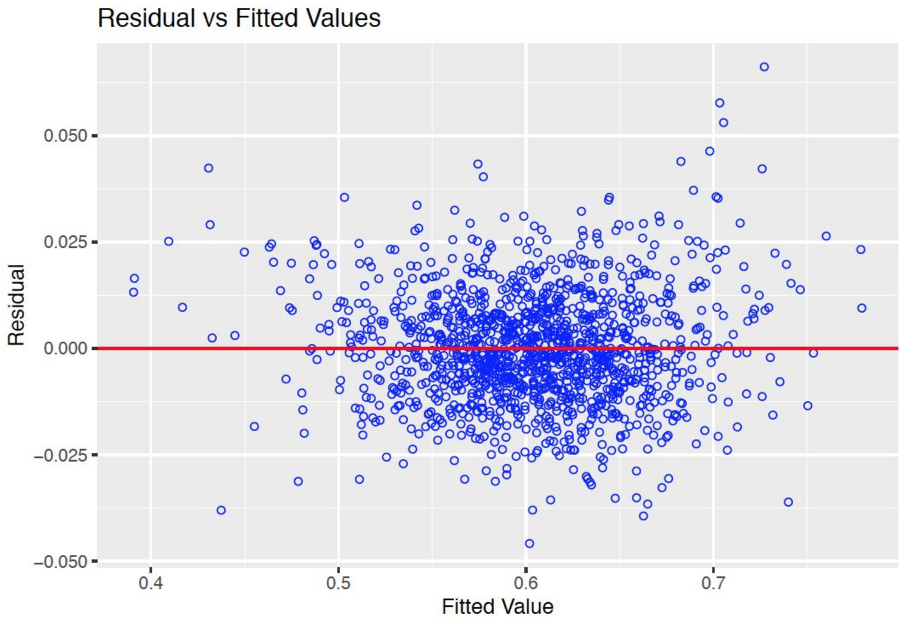
```
##          HorizontalLocationBall            AverageExitVelocity
##                        1.255609                       1.646473
##              AverageHitDistance AverageEstimatedBattingAverage
##                        1.827139                       1.349012
##                 SwingPercentage                 MissPercentage
##                        1.869348                       2.141510
##                BarrelPercentage                       SpinRate
##                        1.795160                       8.577754
##                      BauerIndex
##                        9.423643
```

Normality Tests For Model 2

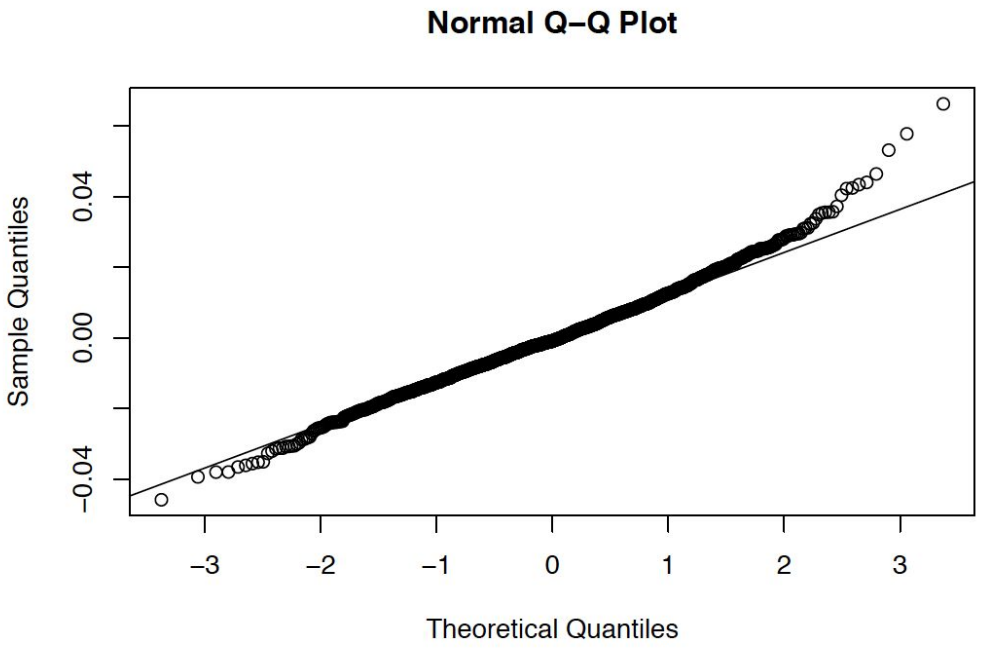


```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9906         0.0000
## Kolmogorov-Smirnov        0.0306         0.1607
## Cramer-von Mises        438.0917         0.1054
## Anderson-Darling          2.2325         0.0000
## -----------------------------------------------
```

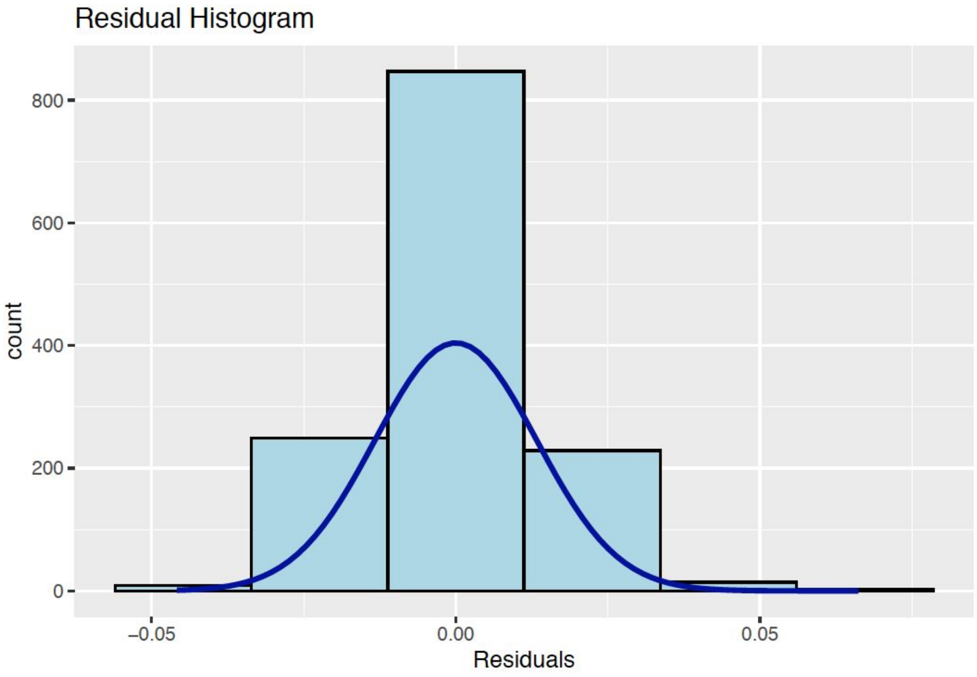
Residuals vs. Fitted Values For Model 2



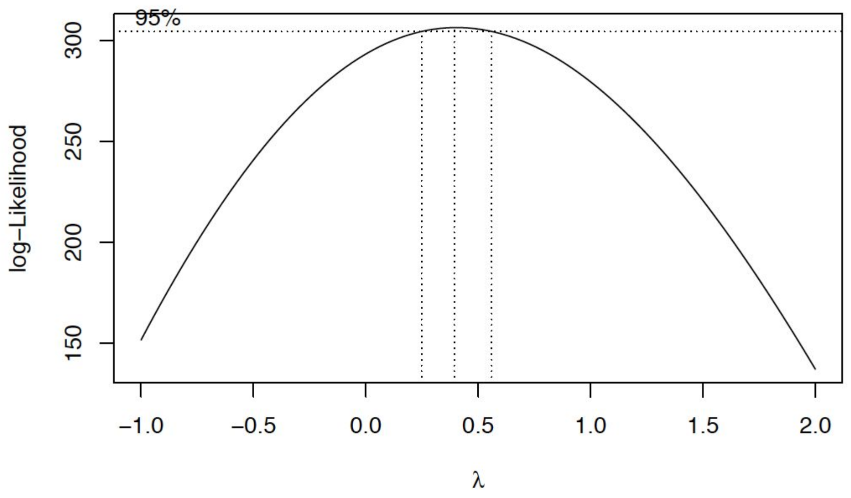
QQ Plot For Model 2



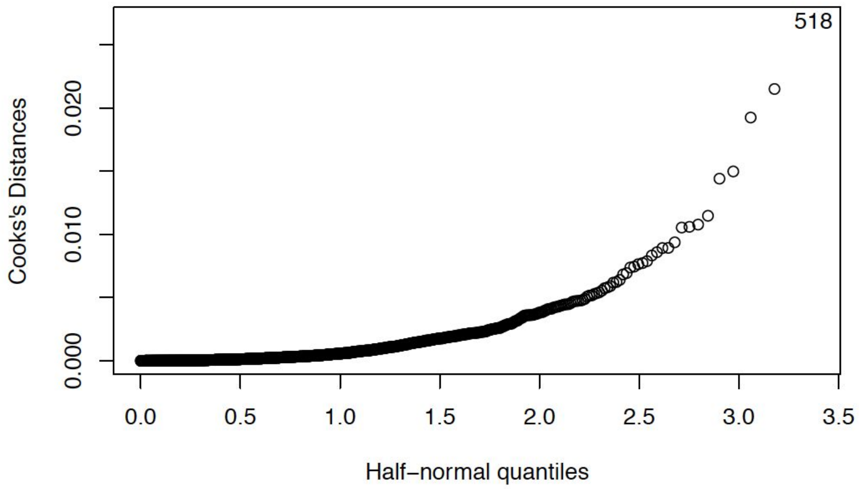
Histogram of Residuals For Model 2



Box-Cox Test For Model 2



Cook’s Distance Test On Model 2


Cooks's Distances
Half−normal quantiles