WAR

Juan Valencia

5/12/2021

Introduction

I wanted to show which advanced baseball stats most correlate with WAR, which in baseball is the best indicator of success. My analysis will show which statistics can be used to find the best fit model. I acquired my data though baseball-reference.com, and the data I am using is the Team Player Value for Batters in the 2017 baseball season.

Packages and Data Preparation

The dataset contains 12 columns of different baseball statistics with 30 rows representing the 30 teams in the MLB. The column names and definitions are the following:

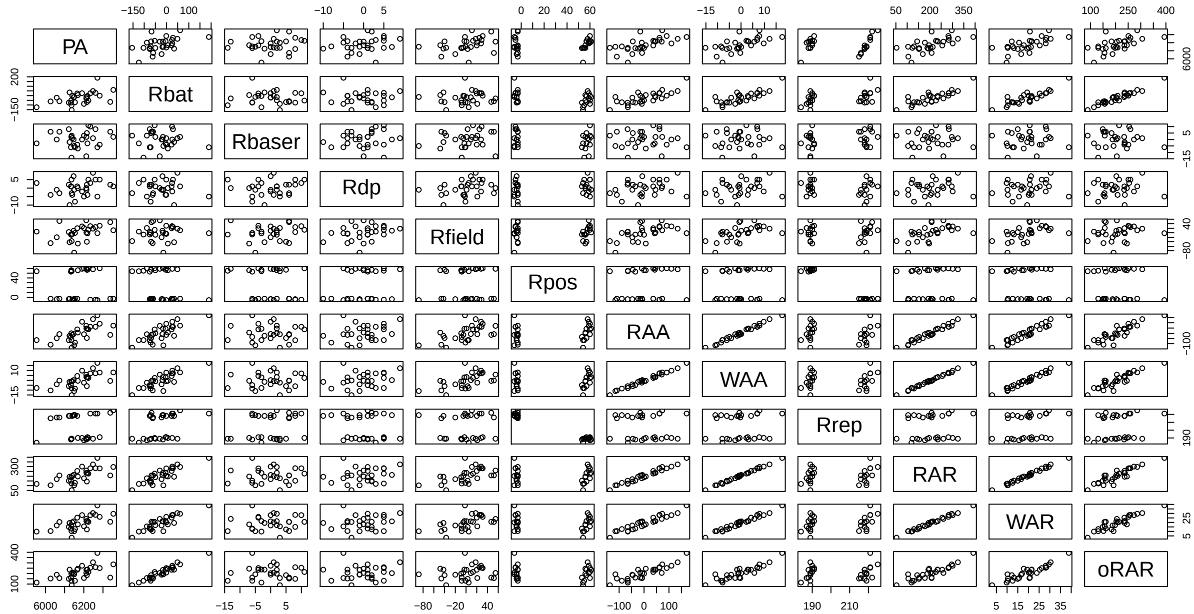
- PA – Plate Appearances
- Rbat – Runs Batting Number of runs better or worse than average the player was as a hitter. This is based on a modified version of wRAA.
- Rbaser – Runs from Baserunning Number of runs better or worse than average the player was for all baserunning events. SB, CS, PB, WP, Defensive Indifference.
- Rdp – Runs Grounded into Double Plays Number of runs better or worse than average the player was at avoiding grounding into double plays.
- Rfield – Runs from Fielding Number of runs better or worse than average the player was for all fielding. Fielding of balls in play, turning double plays, outfield arms and catcher defense are all included.
- Rpos – Runs from Positional Scarcity Number of runs above or below average due to positional differences. Positions like C, SS, and 2B get a bonus. Positions like 1B, DH, LF get a penalty.
- RAA – Runs better than Avg It is the number of runs this player is better than a league average player.
- WAA – Wins Above Avg This is the wins added by this player above that of an average player. I compute the waaW-L% using a PythagenPat conversion and then subtract .500 and multiply by the number of games played.
- Rrep – Runs from Replacement Level Number of runs an average player is better than a replacement player. Replacement is set for a .294 team winning percentage. Stronger leagues may get a larger bonus.
- RAR – Runs above Replacement Level Total of other columns It is the number of runs this player is better than a replacement player. Replacement is set for a .294 team winning percentage.
- WAR – Wins Above Replacement A single number that presents the number of wins the player added to the team above what a replacement player (think AAA or AAAA) would add. Scale for a single-season: 8+ MVP Quality, 5+ All-Star Quality, 2+ Starter, 0-2 Reserve, < 0 Replacement Level

- waaWL% – Win-Loss% w/ Avg. Team This is the win-loss of an otherwise average team in ONLY the games this player played in. For example, for a pitcher this would only consider the games the pitcher threw in and ignoring games they did not play in. 162WL% – Win-Loss% w/ Avg. Team Season This is the win-loss of an otherwise average team for an entire season giving them credit for only the games this player played in.
- oRAR – Offensive Runs above Replacement Level

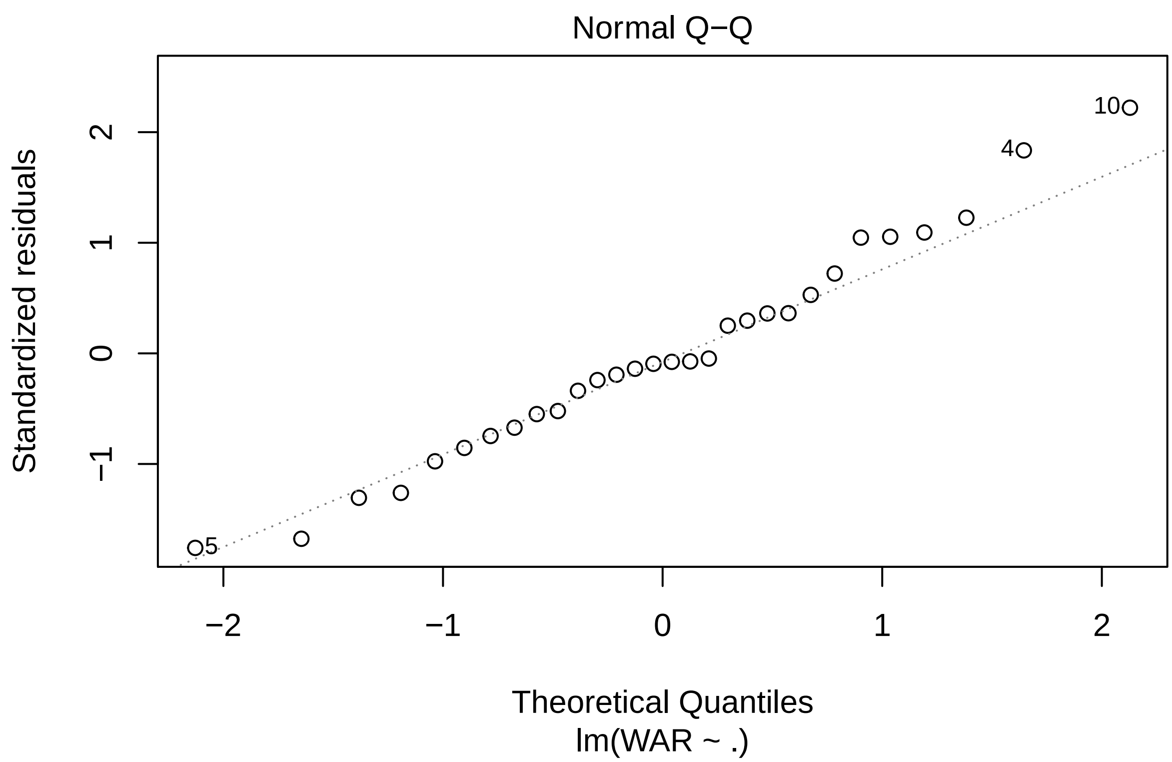
A Data Cleaning

Loading the data produces seven columns and one row with only NA values. I removed those columns and the row and plotted the data.

```
##
## Call:
## lm(formula = WAR ~ ., data = war)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09496 -0.03523 -0.00389  0.02250  0.12649
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.6008224  1.5322170  -1.045   0.3100
## PA          -0.0002477  0.0004840  -0.512   0.6150
## Rbat         0.0281665  0.0417055   0.675   0.5080
## Rbaser       0.0270824  0.0431010   0.628   0.5377
## Rdp          0.0271757  0.0423288   0.642   0.5290
## Rfield       0.1021257  0.0535830   1.906   0.0728 .
## Rpos         0.0316314  0.0411934   0.768   0.4525
## RAA         -0.0469030  0.0304490  -1.540   0.1409
## WAA          0.8769182  0.1617059   5.423 3.75e-05 ***
## Rrep         0.0801483  0.0397482   2.016   0.0589 .
## RAR         -0.0434207  0.0367635  -1.181   0.2529
## oRAR         0.0744591  0.0444001   1.677   0.1108
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0625 on 18 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:  0.9999
## F-statistic: 4.074e+04 on 11 and 18 DF,  p-value: < 2.2e-16
```




```r
plot(lm(WAR~., war),which = 2)
```



My model shows that I have eleven models that stand out as being candidates for the best model, so my analysis will be on these eleven models to find the best model.

```r
ols_step_best_subset(lm(WAR~.,war))
##                          Best Subsets Regression
## --------------------------------------------------------------------------
## Model Index    Predictors
## --------------------------------------------------------------------------
##      1         RAR
##      2         WAA Rrep
##      3         Rbat WAA Rrep
##      4         Rdp WAA Rrep oRAR
##      5         Rfield WAA Rrep RAR oRAR
##      6         Rbat Rfield RAA WAA Rrep oRAR
##      7         Rbat Rfield RAA WAA Rrep RAR oRAR
##      8         Rbat Rfield Rpos RAA WAA Rrep RAR oRAR
##      9         PA Rbat Rfield Rpos RAA WAA Rrep RAR oRAR
##     10         Rbat Rbaser Rdp Rfield Rpos RAA WAA Rrep RAR oRAR
##     11         PA Rbat Rbaser Rdp Rfield Rpos RAA WAA Rrep RAR oRAR
## --------------------------------------------------------------------------
##
##                                                          Subsets Regression Summary
## -----------------------------------------------------------------------------------------------------
##                          Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)          AIC          SBIC          SBC         MSEP
## -----------------------------------------------------------------------------------------------------
##   1        0.9848       0.9843      0.9826   6768.8709      87.4587       -3.6423      91.6623     28.43
##   2        0.9999       0.9999      0.9999      3.5978     -75.7269     -160.3561     -70.1221      0.12
##   3        0.9999       0.9999      0.9999      2.0268     -77.8838     -161.1549     -70.8779      0.10
##   4        0.9999       0.9999      0.9999      2.6430     -77.6634     -159.7618     -69.2563      0.10
##   5        1.0000       0.9999      0.9999      3.6580     -76.9977     -157.8087     -67.1893      0.10
##   6        1.0000       0.9999      0.9999      4.2867     -76.9600     -155.8516     -65.7505      0.10
##   7        1.0000       0.9999      0.9999      5.2129     -76.5915     -153.3750     -63.9807      0.10
##   8        1.0000       0.9999      0.9999      6.7541     -75.3166     -150.3784     -61.3046      0.10
##   9        1.0000       0.9999      0.9999      8.4213     -73.8537     -147.2092     -58.4405      0.10
##  10        1.0000       0.9999      0.9999     10.2620     -72.1143     -143.9363     -55.3000      0.11
##  11        1.0000       0.9999      0.9999     12.0000     -70.5478     -140.5730     -52.3322      0.11
## -----------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

For the PRESS statistic I can rule out PRESS(lm(WAR~RAR,war)) because it is much larger than the other values.

```r
PRESS(lm(WAR~RAR,war))
## [1] 30.45763
PRESS(lm(WAR~WAA+ Rrep,war))
## [1] 0.1311349
PRESS(lm(WAR~Rbat+ WAA+ Rrep,war))
## [1] 0.1231245
PRESS(lm(WAR~Rdp+ WAA+ Rrep+ oRAR,war))
```

```
## [1] 0.1260137
PRESS(lm(WAR~Rfield+ WAA+ Rrep+ RAR+ oRAR,war))
## [1] 0.1206167
PRESS(lm(WAR~Rbat+ Rfield+ RAA+ WAA+ Rrep+ oRAR,war))
## [1] 0.1386966
PRESS(lm(WAR~Rbat+ Rfield+ RAA+ WAA +Rrep+ RAR+ oRAR,war))
## [1] 0.1372765
PRESS(lm(WAR~Rbat+ Rfield+ Rpos +RAA+ WAA+ Rrep +RAR +oRAR,war))
## [1] 0.1446987
PRESS(lm(WAR~PA +Rbat+ Rfield +Rpos+ RAA+ WAA+ Rrep+ RAR+ oRAR,war))
## [1] 0.1503246
PRESS(lm(WAR~Rbat +Rbaser+ Rdp+ Rfield +Rpos +RAA +WAA+ Rrep +RAR +oRAR,war))
## [1] 0.1600664
PRESS(lm(WAR~PA+ Rbat+ Rbaser+ Rdp +Rfield+ Rpos+ RAA+ WAA+ Rrep+ RAR+ oRAR,war))
## [1] 0.1687674
```

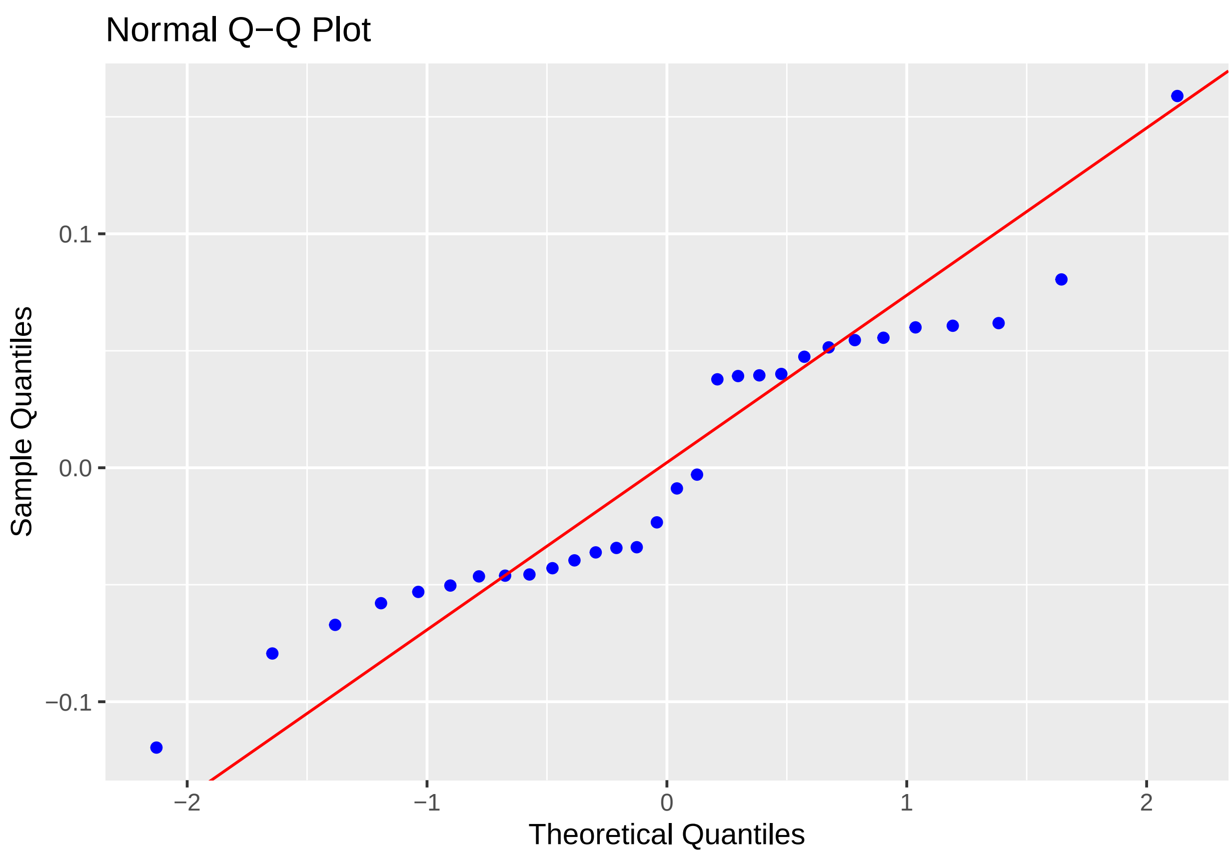
The models vif(lm(WAR~WAA+ Rrep,war)) and vif(lm(WAR~Rbat+ WAA+ Rrep,war)) show good VIF scores, since their VIFs are lower than 5, so they have no potential multicollinearity problems. The rest of the models show multicollinearity.

```
vif(lm(WAR~WAA+ Rrep,war))
##      WAA     Rrep
## 1.007144 1.007144
vif(lm(WAR~Rbat+ WAA+ Rrep,war))
##     Rbat      WAA     Rrep
## 4.767387 4.038240 1.531149
vif(lm(WAR~Rdp+ WAA+ Rrep+ oRAR,war))
##      Rdp      WAA     Rrep     oRAR
## 1.170181 6.092071 1.109990 5.977784
vif(lm(WAR~Rfield+ WAA+ Rrep+ RAR+ oRAR,war))
##      Rfield         WAA        Rrep         RAR        oRAR
##  8295.30498  3135.35338    16.74038 44269.96046 32276.33944
vif(lm(WAR~Rbat+ Rfield+ RAA+ WAA+ Rrep+ oRAR,war))
##        Rbat      Rfield         RAA         WAA        Rrep        oRAR
##    54.05785  5475.29554 26513.48156  3227.27244  1168.47352 20945.54350
vif(lm(WAR~Rbat+ Rfield+ RAA+ WAA +Rrep+ RAR+ oRAR,war))
##        Rbat      Rfield         RAA         WAA        Rrep         RAR
##    54.47627 11050.59713 27748.77824  3245.53958  1228.09843 46432.27361
##        oRAR
## 42820.92768
vif(lm(WAR~Rbat+ Rfield+ Rpos +RAA+ WAA+ Rrep +RAR +oRAR,war))
##        Rbat      Rfield        Rpos         RAA         WAA        Rrep
##    81.42187 11140.13726   167.87581 28029.97406  6184.02669  1376.49783
##         RAR        oRAR
## 46695.84086 43388.92528
vif(lm(WAR~PA +Rbat+ Rfield +Rpos+ RAA+ WAA+ Rrep+ RAR+ oRAR,war))
##          PA        Rbat      Rfield        Rpos         RAA         WAA
##    11.20831    87.59191 11608.18096   300.61049 28887.90510  8520.76345
##        Rrep         RAR        oRAR
##  1383.65370 47492.66994 45433.77972
vif(lm(WAR~Rbat +Rbaser+ Rdp+ Rfield +Rpos +RAA +WAA+ Rrep +RAR +oRAR,war))
##       Rbat     Rbaser        Rdp     Rfield       Rpos        RAA        WAA
## 55635.8030   521.4949   250.5802 21471.6612 12304.7448 37069.1303  8033.8647
```

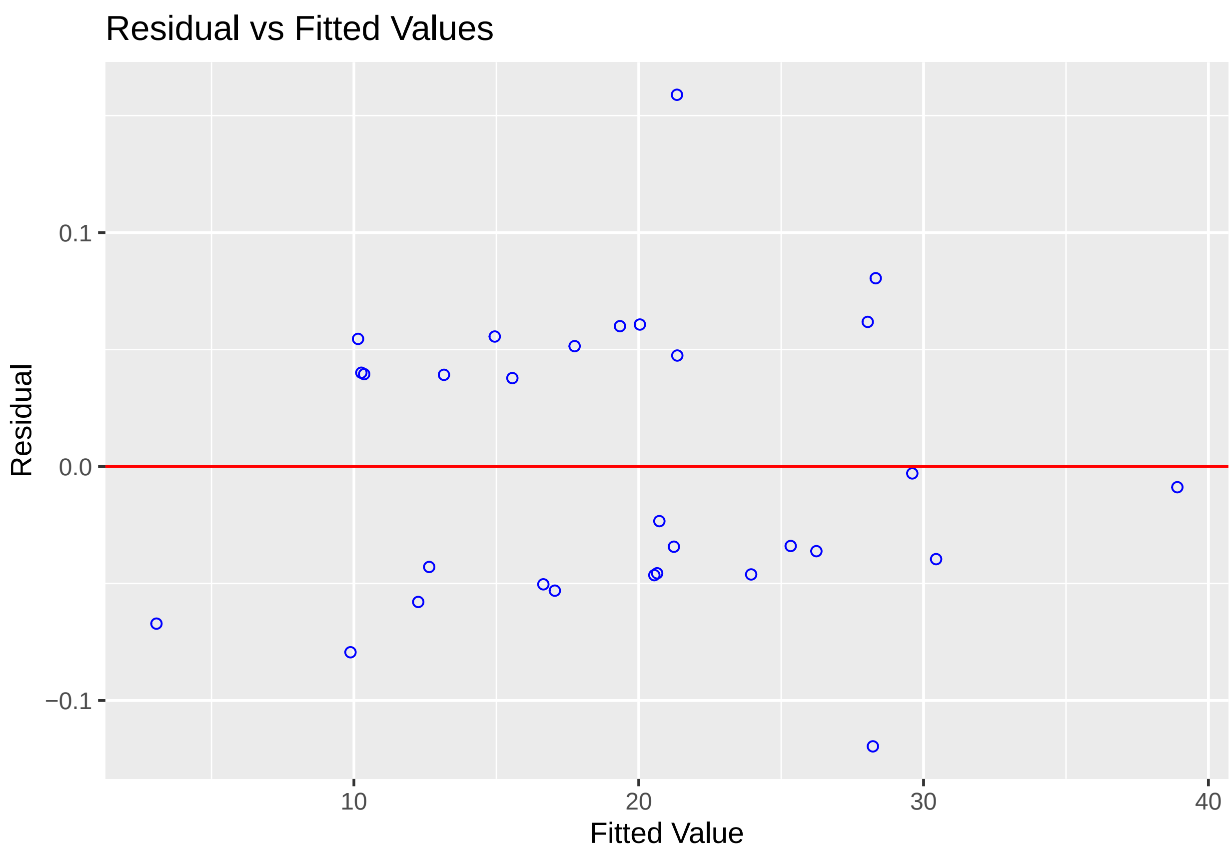
```
##       Rrep        RAR       oRAR
##  2782.8277 56935.3092 52418.4491
vif(lm(WAR~PA+ Rbat+ Rbaser+ Rdp +Rfield+ Rpos+ RAA+ WAA+ Rrep+ RAR+ oRAR,war))
##          PA         Rbat       Rbaser          Rdp       Rfield         Rpos
##    12.88696 56561.30238   530.31644   252.75222 22245.18429 12306.28014
##         RAA          WAA         Rrep          RAR         oRAR
## 38845.66100 10942.74844  2822.28320 57964.11345 59572.02362
```

```
m1 = lm(WAR~WAA+ Rrep,war)
m2 = lm(WAR~Rbat+ WAA+ Rrep,war)
```

```
ols_plot_resid_qq(m1)
```

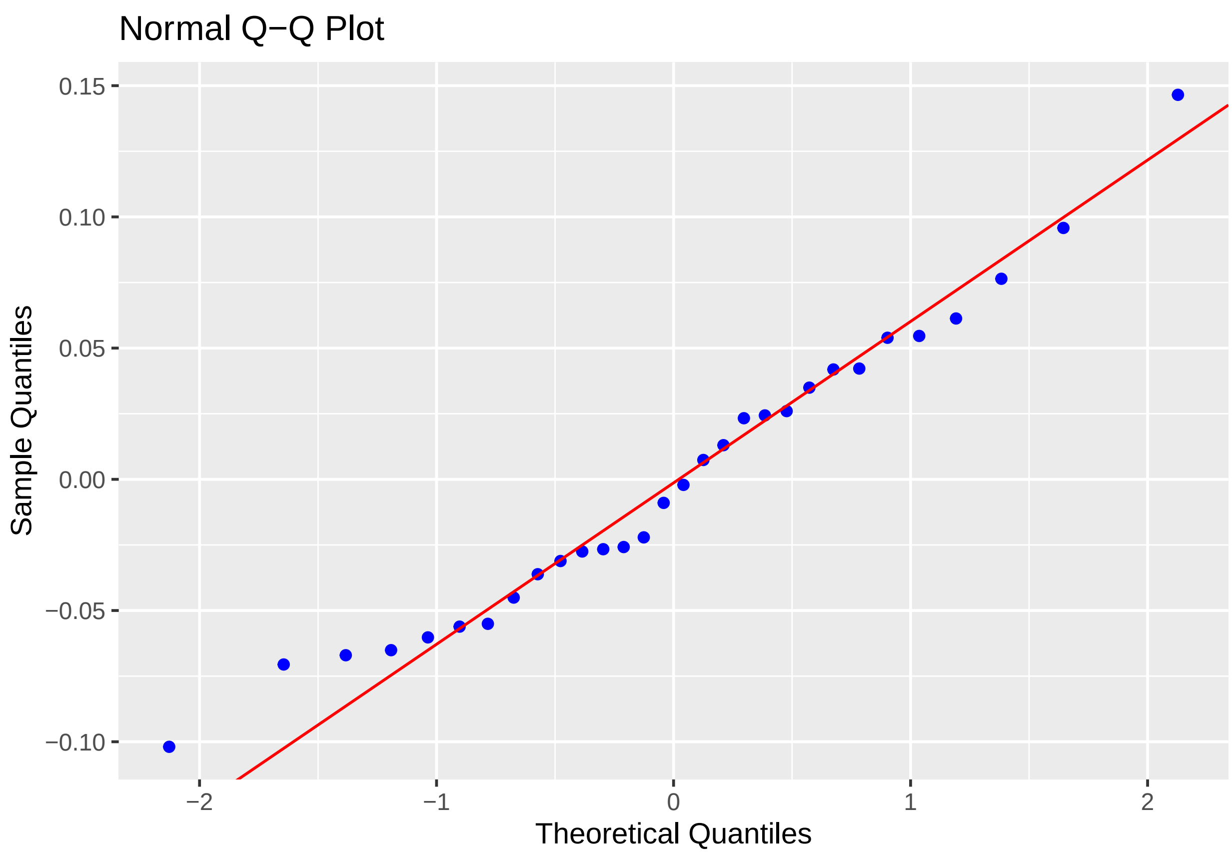



```
ols_test_normality(m1)
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9346         0.0650
## Kolmogorov-Smirnov        0.1779         0.2652
## Cramer-von Mises          8.7518         0.0000
## Anderson-Darling          0.9703         0.0125
## -----------------------------------------------
ols_plot_resid_fit(m1)
```

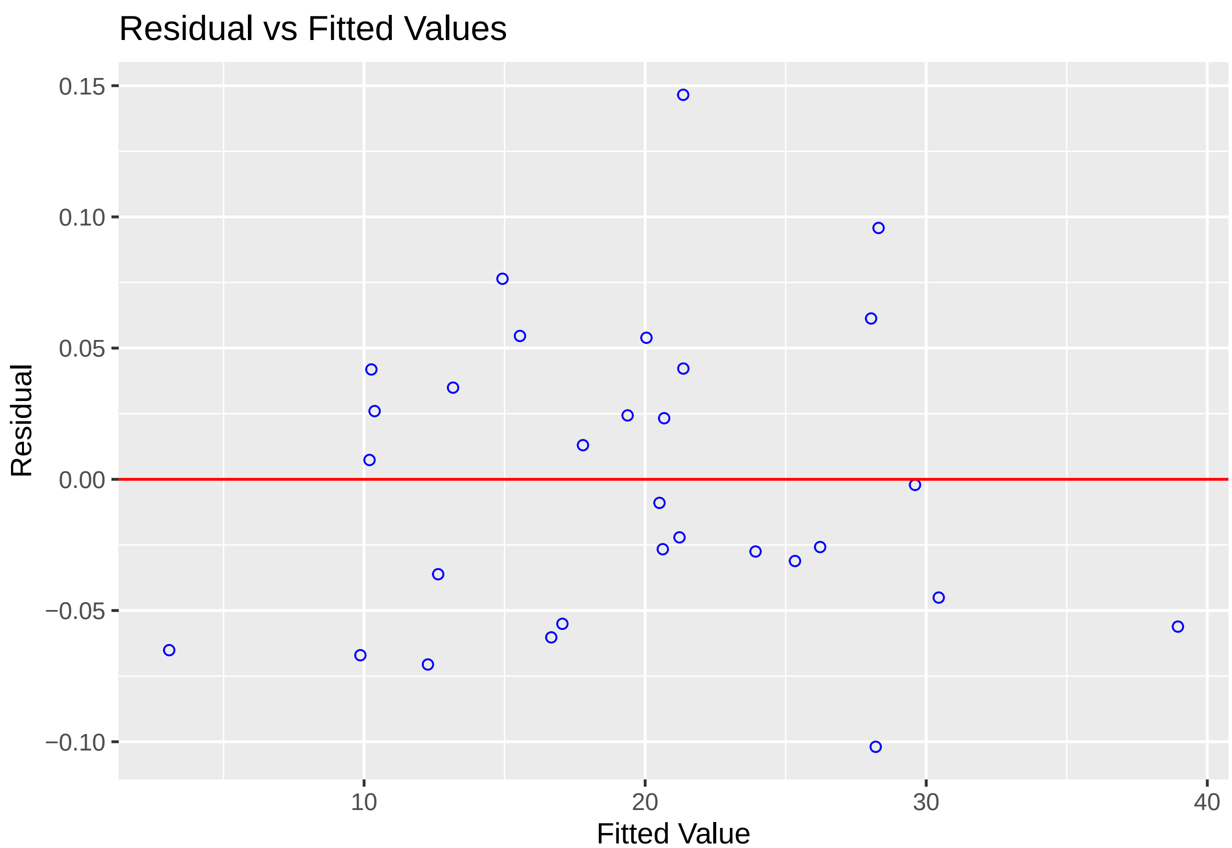



```
anova1 = anova(m1)
sst1 = sum(anova1$'Sum Sq')
1-PRESS(m1)/(sst1)
## [1] 0.9999251
```

```
ols_plot_resid_qq(lm(WAR~Rbat+ WAA+ Rrep,war))
```




```
ols_test_normality(lm(WAR~Rbat+ WAA+ Rrep,war))
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9725         0.6084
## Kolmogorov-Smirnov        0.118          0.7545
## Cramer-von Mises          8.8553         0.0000
## Anderson-Darling          0.2909         0.5856
## -----------------------------------------------
ols_plot_resid_fit(lm(WAR~Rbat+ WAA+ Rrep,war))
```




```
anova2 = anova(m2)
sst2 = sum(anova2$'Sum Sq')
1-PRESS(m2)/(sst2)
## [1] 0.9999297
```

- For model 1, the PRESS statistic is 0.1311349 and the predictive R-squared is 0.9999251.
- For model 2, the PRESS statistic is 0.1231245 and the predictive R-squared is 0.9999297.
- According to the above result, model 2 has a larger value of the predictive R-square, so model 2 is better than model 1.

```
summary(lm(WAR~Rbat+ WAA+ Rrep,war))
##
## Call:
## lm(formula = WAR ~ Rbat + WAA + Rrep, data = war)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.101928 -0.042838 -0.005556  0.040130  0.146509
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.3174571  0.1869771   1.698   0.1015
## Rbat        0.0007235  0.0003681   1.966   0.0601 .
## WAA         0.9939082  0.0029862 332.833   <2e-16 ***
## Rrep        0.0950607  0.0008900 106.810   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06008 on 26 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 1.616e+05 on 3 and 26 DF,  p-value: < 2.2e-16
```

- I can now take the summary of model 2 to get

$$\hat{y} = 0.3174571 + (0.0007235)Rbat + (0.9939082)WAA + (0.0950607)Rrep$$

Conclusion

With my analysis I was able to indicate which baseball statistics could be used to determine the best model. I was able to do this by using PRESS and VIF. Thus, with a combination of runs batting, wins above replacement and runs from replacement player statistics, I am able to find the best model to determine WAR, versus if it were done with individual baseball statistics or with a combination of other statistics, in which I would not be able to achieve the same results. Since WAR is the best metric to determine which team is most likely to be successful, then my model,

$$\hat{y} = 0.3174571 + (0.0007235)Rbat + (0.9939082)WAA + (0.0950607)Rrep$$

should also be a good indicator of team success. Thus, my model will look at multiple variables and could potentially determine which teams could have the best success in the playoffs for a given season.

Reference

1. 2017 MLB Value. (2017). Retrieved from Baseball Reference: https://www.baseball-reference.com/leagues/MLB/2017-value-batting.shtml