What determines WAR in Baseball

Sam Eichel
Mentor: Jebessa Mijena
Department of Mathematics
Georgia College & State University 2020

Content

Acknowledgements

I would like to thank Jebessa Mijena for mentoring me these last two semesters. I would also like to thank my family for their constant support.

Abstract

2020 has been a very weird year. The 2020 Baseball season was no different. Instead of the normal 162 game season, it was a very short season of 60 games. In this research, we used variable selection to determine what predictors have the highest effect on War(Wins Above Replacement). We also used several other techniques from multiple regression analysis to find the best fit for War.

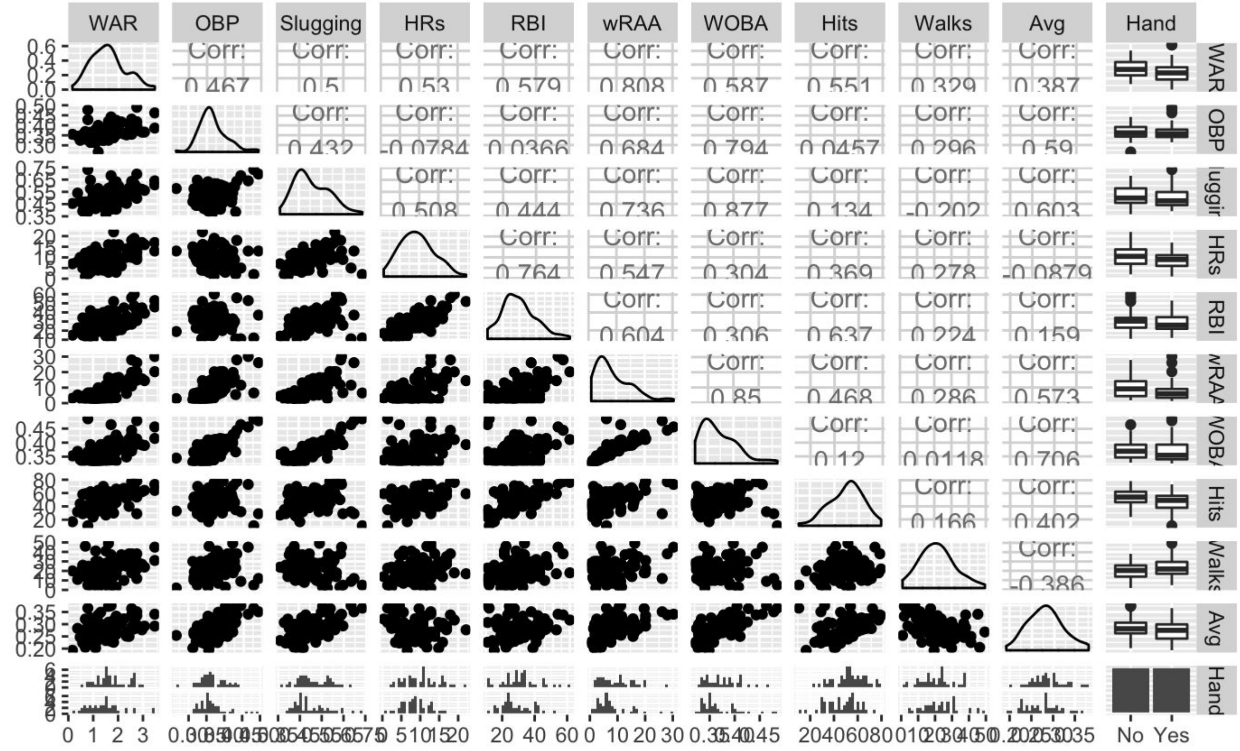
Introduction

2020 was a very weird baseball Season. The motivation behind this project is to see what kind of effect 2020 had on a player’s WAR and how WAR differs from a normal length season. For this project I used several statistics: WAR(Wins Above Replacement), OBP(On Base Percentage), Slugging(Slugging Percentage), HRs (Home runs), RBI(Runs Batted In), wRAA(Weighted Runs Above Average), wOBA(Weighted On base Percentage), Hits, Walks, Avg(Batting Average), and Handedness. In general, WAR assesses how valuable a player is. For example, one player could hit a lot of HRs, but his WAR isn't very high because he struggles in other areas like wRAA, or wOBA. OBP measures how often a player reaches base via hit(this includes HRs) or walk. Slugging percentage measures a player’s productivity. WRAA assesses how many runs a player creates for his team, relative to the league average. WOBA measures a player’s overall performance per plate appearance. This research uses Variable Selection to determine what statistics had the highest influence on WAR. It also uses techniques from multiple regression to determine the best fit. For the data analysis, I used R statistical software.

Data Analysis

I started this project by collecting data from the 2020 mlb season. The data I collected included several different statistics for 98 players(49 left handed batters and 49 right handed batters). These statistics were: WAR(Wins Above Replacement), OBP(On Base Percentage), Slugging(Slugging Percentage), HRs (Home runs), RBI(Runs Batted In), wRAA(Weighted Runs Above Average), wOBA(Weighted On base Percentage), Hits, Walks, Avg(Batting Average), and Handedness.

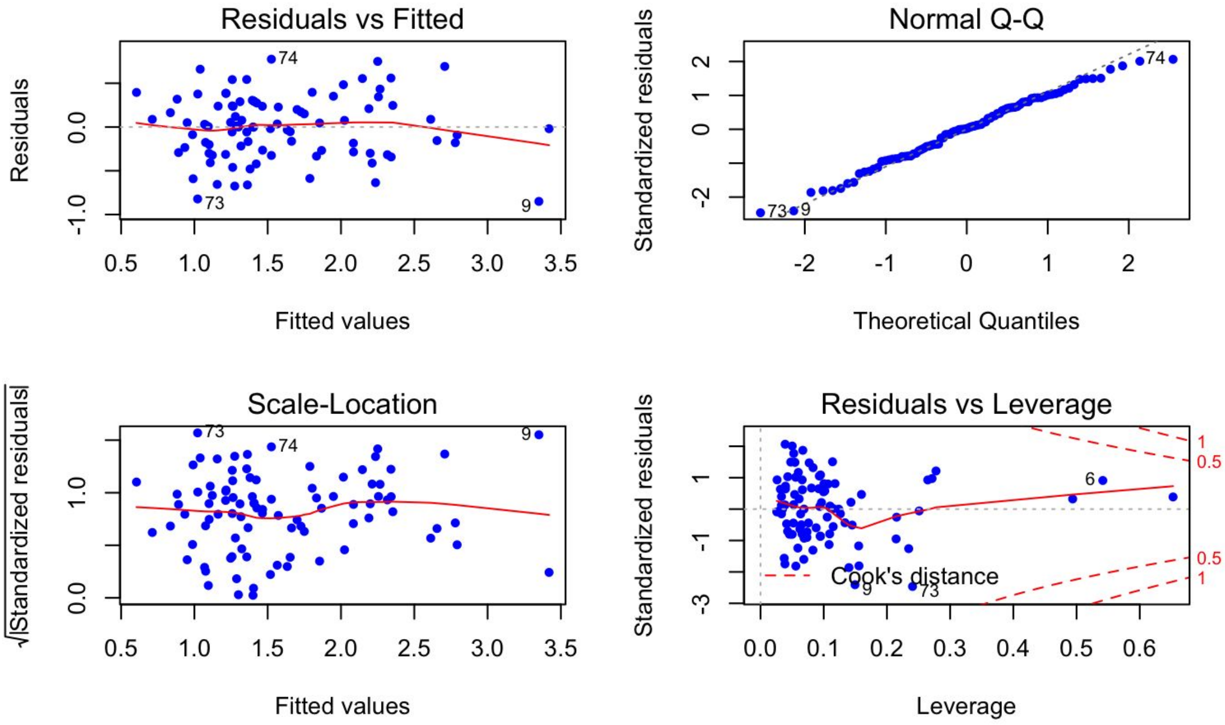
After recording the data, I imported the data table into RStudio. To determine the effect of Handedness, I created two separate variables, one named Righthanded, and the other Lefthanded. After doing this, I created the scatterplot matrix below:



The scatterplot matrix shows that right handed hitters had a higher war in 2020 than left handed batters(This differs from what we would expect from a normal length season). Next, I created the original fit for WAR. The fit and the diagnostic plots are as follows:

```
Call:
lm(formula = WAR ~ OBP + Slugging + HRs + RBI + wRAA + WOBA +
    Hits + Walks + Avg, data = MLB4)


Residuals:
     Min       1Q   Median       3Q      Max
-0.85003 -0.27211  0.00845  0.25336  0.77485


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.212026   1.463139   0.145  0.88514
OBP          0.331037   5.498938   0.060  0.95214
Slugging    -1.618973   2.822587  -0.574  0.56782
HRs          0.011577   0.024792   0.467  0.64176
RBI         -0.005218   0.008085  -0.645  0.52046
wRAA         0.078327   0.024644   3.178  0.00209 **
WOBA         5.575667   9.663463   0.577  0.56553
Hits         0.017816   0.008431   2.113  0.03764 *
```

```
Walks       -0.006460   0.010877  -0.594  0.55417
Avg         -4.965705   4.393963  -1.130  0.26172
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.383 on 82 degrees of freedom
Multiple R-squared:  0.7129,  Adjusted R-squared:  0.6814
F-statistic: 22.63 on 9 and 82 DF,  p-value: < 2.2e-16
```



The regression model for the original fit is: WAR=0.331(OBP)-1.62(Slugging)+0.0116(HRs)-0.00522(RBI)+0.0783(wRAA)+5.58(wOBA)+ 0.0178(Hits)-0.00646(Walks)-4.97(Avg)+0.212. The R squared value is 0.6814. We can tell by looking at the plots, that this original fit has problems with constant variance and cook's distance.

Now I wanted to see which predictors had the high influence on WAR. To do so, I used variable selection, a technique where you remove predictors one by one until each of the remaining one are statistically significant(p<0.05). As seen below, The final model included wRAA, Hits, and Avg. The R squared value for this model was 0.6973, which meant that 69.73% of the variation in WAR could be explained by these 3 predictors.

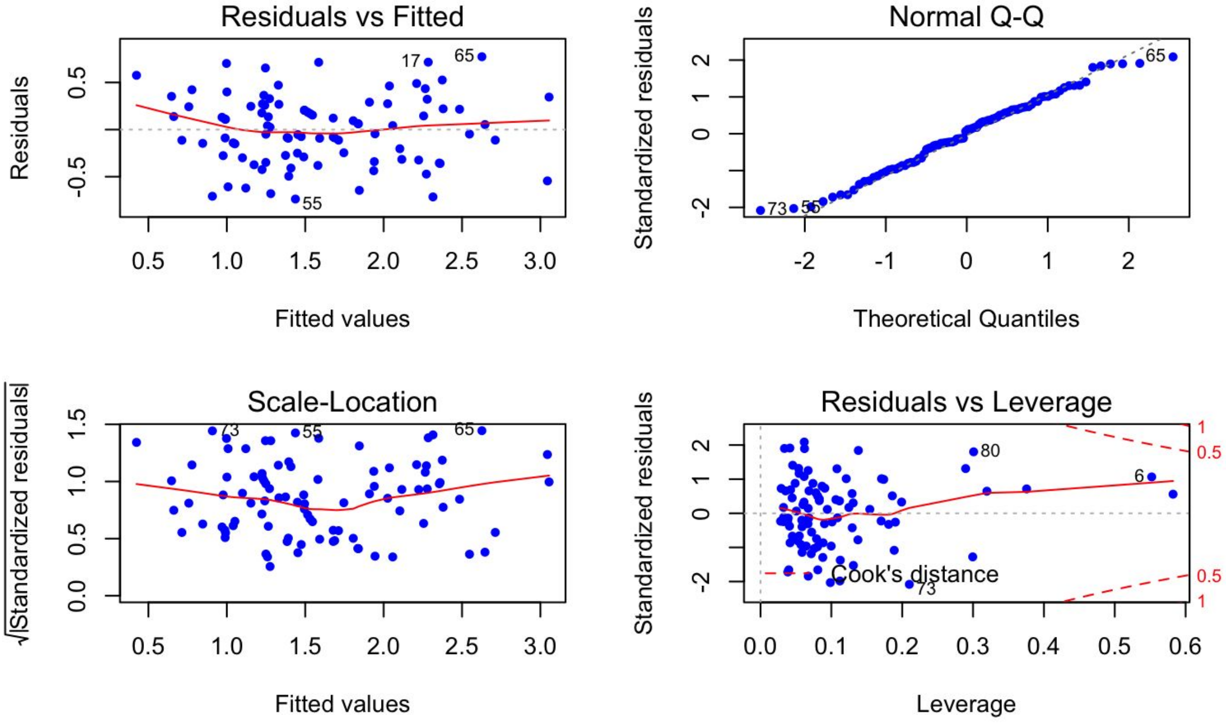
```
Call:
lm(formula = WAR ~ wRAA + Hits + Avg, data = MLB4)


Residuals:
     Min       1Q   Median       3Q      Max
-0.89261 -0.26349  0.01905  0.25225  0.79657


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.992633   0.317249   3.129 0.002380 **
wRAA         0.085455   0.008092  10.561  < 2e-16 ***
Hits         0.012354   0.003291   3.754 0.000311 ***
```

```
Avg         -2.767697   1.229618  -2.251 0.026886 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3733 on 88 degrees of freedom
Multiple R-squared:  0.7073,  Adjusted R-squared:  0.6973
F-statistic: 70.89 on 3 and 88 DF,  p-value: < 2.2e-16
```

Now I wanted to try to improve the fit for WAR. To do this, I tried two different transformations. The first transformation I tried was transforming all the variables. To do this I used the powerTransform function with all the variables. The power transform along with the resulting model, and the diagnostic plots are below:

	Estimated	Rounded	Lower	Upper
WAR	0.8666	1	0.5772	1.156
OBP	0.008854	0	-0.5537	0.5714
Slugging	1.143	1	0.5192	1.766
HRs	0.5727	0.5	0.3721	0.7734
RBI	0.3334	0	-0.02299	0.6898
wRAA	0.5597	0.5	0.4247	0.6947
WOBA	0.4261	1	-0.4361	1.288
Hits	0.4791	0.5	0.1935	0.7647
Walks	0.45	0.5	0.3057	0.5942
Avg	1.472	1	0.9485	1.995

```
Call:
lm(formula = WAR ~ tOBP + Slugging + tHRs + tRBI + twRAA + WOBA +
    tHits + tWalks + Avg)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73741 -0.28000  0.03256  0.24949  0.77325

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3950     4.9721   0.281   0.7798
tOBP          1.4412     2.2780   0.633   0.5287
Slugging     -1.2327     2.9962  -0.411   0.6818
tHRs          0.1361     0.1522   0.894   0.3737
tRBI         -0.1245     0.2271  -0.548   0.5852
twRAA         0.4307     0.1600   2.692   0.0086 **
WOBA          3.8807    10.4567   0.371   0.7115
tHits         0.2674     0.1287   2.078   0.0408 *
tWalks       -0.1085     0.1170  -0.927   0.3565
Avg          -6.2309     5.0520  -1.233   0.2210
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3825 on 82 degrees of freedom
Multiple R-squared:  0.7137,  Adjusted R-squared:  0.6823
F-statistic: 22.71 on 9 and 82 DF,  p-value: < 2.2e-16
```

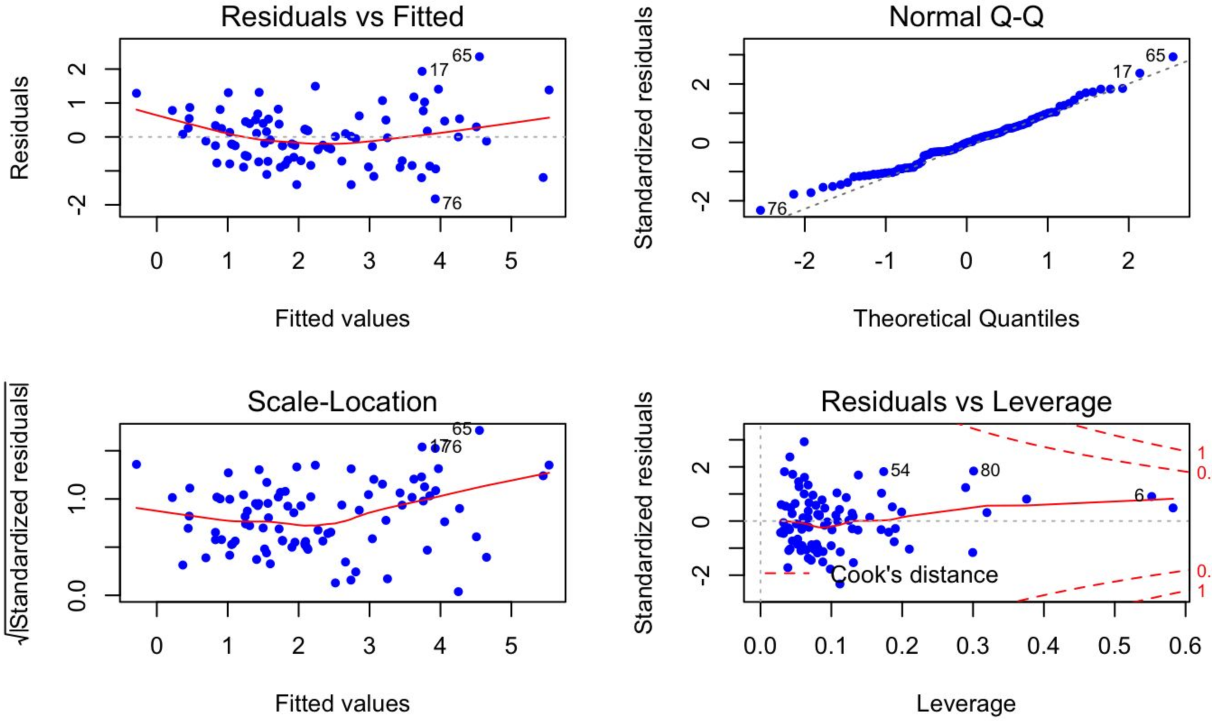


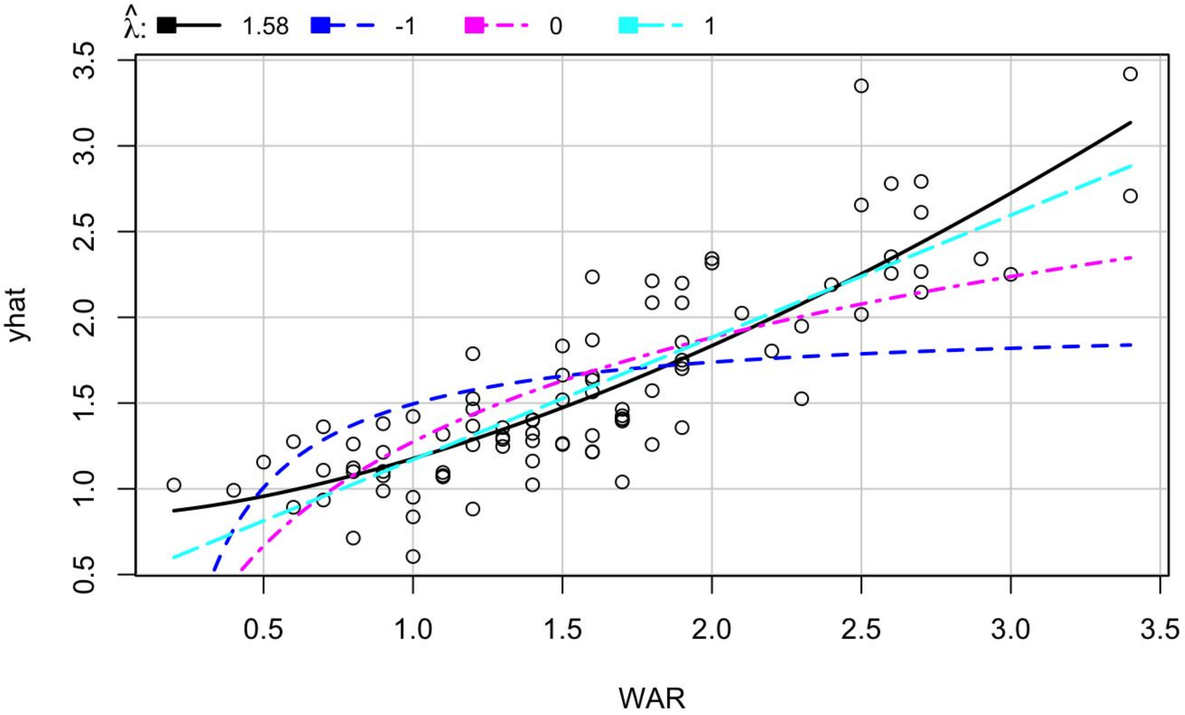
The regression model for full power transformation fit is: WAR=1.44*log(OBP)-1.23(Slugging)+0.136*(HRs)^(½)-0.125*log(RBI)+0.431*(wRAA)^(½) +3.88(wOBA)+0.267*(Hits)^(½)-0.109*(Walks)^(½)-6.23(Avg)+1.40. While this fit has a higher R squared value than the original one, it still has problems with cook’s distance and constant variance. The second transformation I tried was inverse response:

	Estimated	Rounded	Lower	Upper
OBP	-0.0389	0.0	-0.6095	0.5317
Slugging	1.1274	1.0	0.5024	1.7523
HRs	0.5771	0.5	0.3759	0.7782
RBI	0.3382	0.0	-0.0180	0.6944
wRAA	0.5348	0.5	0.3953	0.6743
WOBA	0.3309	1.0	-0.5432	1.2049
Hits	0.4699	0.5	0.1794	0.7604
Walks	0.4538	0.5	0.3090	0.5986
Avg	1.4219	1.0	0.8929	1.9509

```
Call:
lm(formula = tWAR ~ tOBP + Slugging + tHRs + tRBI + twRAA2 +
    WOBA + tHits + tWalks + Avg)

Residuals:
     Min       1Q   Median       3Q      Max
-1.82559 -0.68809 -0.01112  0.47404  2.36418
```

```
Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)   1.4021    10.8372   0.129  0.89738

tOBP          2.9589     4.9651   0.596  0.55285

Slugging     -1.4830     6.5306  -0.227  0.82092

tHRs          0.2347     0.3317   0.708  0.48123

tRBI         -0.2015     0.4951  -0.407  0.68508

twRAA2        0.9334     0.3487   2.677  0.00897 **

WOBA          7.6179    22.7915   0.334  0.73905

tHits         0.5330     0.2805   1.900  0.06090 .

tWalks       -0.1781     0.2550  -0.698  0.48687

Avg         -13.5993    11.0113  -1.235  0.22035

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.8336 on 82 degrees of freedom

Multiple R-squared:  0.7142,  Adjusted R-squared:  0.6828

F-statistic: 22.77 on 9 and 82 DF,  p-value: < 2.2e-16
```


Residuals vs Fitted
Residuals
Fitted values
-2
0
1
2
0
1
2
3
4
5
65
17
76
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-2
0
2
-2
-1
0
1
2
65
17
76
Scale-Location
√|Standardized residuals|
Fitted values
0.0
1.0
0
1
2
3
4
5
65
17
76
Residuals vs Leverage
Standardized residuals
Leverage
-2
0
2
0.0
0.1
0.2
0.3
0.4
0.5
0.6
54
80
6
1
0.5
0.5
1
Cook's distance



The model for the inverse response transformation is: (WAR)^1.58 =2.959*log(OBP)-1.483(Slugging)+0.2347*(HRs)^(0.5)-0.2015*log(RBI)+0.9334*(wRAA)^(0.5)+7.618(wOBA)+0.5530*(Hits)^(0.5)-0.1781*(walks)^(0.5)-13.60(Avg)+1.402.

Not only does this fit have a higher R Squared value than the original fit, it also fixes the cook's distance problem(the normal power transformation model failed to do this). Inverse response still fails to account for problems with constant variance.

Conclusion

The inverse response transformation gave us the best fit for WAR. The inverse response model had an R Squared value of 0.6828, vs 0.6823 for the normal power transformation model. The inverse response model also fixed issues with cook’s distance. One thing Inverse response still did not fix is the issue with constant variance. In order to further improve the fit for WAR, I could have tried other models such as the tree based model. There were also a couple lurking variables that affected my model. For example, defense and baserunning factor into WAR as well, but I didn't include statistics such as BsR(Base Running Runs) or UZR(Ultimate Zone Rating): A player could be really good offensively, but still have a low WAR because he is not a great base runner or fielder. Having said that, had I used Offensive WAR instead of just WAR, the model would have been a lot better, because offense WAR does not include defense.

References

“Major League Leaderboards " 2020 " Batters " Dashboard: FanGraphs Baseball.” *Major League Leaderboards " 2020 " Batters " Dashboard | FanGraphs Baseball*, www.fangraphs.com/leaders.aspx?pos=all.

Barret Schloerke, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg and Joseph Larmarange (2018). GGally: Extension to 'ggplot2'. R package version 1.4.0. https://CRAN.R-project.org/package=GGally

Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr

Sanford Weisberg (2005). Applied Linear Regression, Third Edition. Hoboken NJ: Wiley. URL: http://www.stat.umn.edu/alr