"What raw statistics have the greatest effect on wRC+ in Major League Baseball in 2017?"

Gavin D. Sanford
University of Minnesota Duluth
Honors Capstone Project

Abstract

Major League Baseball has different statistics for hitters, fielders, and pitchers. The game has followed the same rules for over a century and this has allowed for statistical comparison. As technology grows, so does the game of baseball as there is more areas of the game that people can monitor and track including pitch speed, spin rates, launch angle, exit velocity and directional break. The website QOPBaseball.com is a newer website that attempts to correctly track every pitches horizontal and vertical break and grade it based on these factors (Wilson, 2016). Fangraphs has statistics on the direction players hit the ball and what percentage of the time. The game of baseball is all about quantifying players and being able give a value to their contributions. Sabermetrics have given us the ability to do this in far more depth. Weighted Runs Created Plus (wRC+) is an offensive stat which is attempted to quantify a player's total offensive value (wRC and wRC+, Fangraphs). It is Era and park adjusted, meaning that the park and year can be compared without altering the statistic further. In this paper, we look at what 2018 statistics have the greatest effect on an individual player's wRC+.

Keywords: Sabermetrics, Econometrics, Spin Rates, Baseball,

Introduction

Major League Baseball has been around for over a century has given awards out for almost 100 years. The way that these awards are given out is based on statistics accumulated over the season. For so long Batting Average, Homeruns, and Runs Batted In were the numbers gauging a player's case for an award such as Most Valuable Player or Silver Slugger. Research has since suggested that the value of players should have less emphasis on the traditional statistics that are used in baseball. Runs Batted In is a famous baseball statistic but is too overvalued. Someone may not have as many opportunities to accumulate RBI's as another, making the statistic situational advantageous to some and a product of the team a player plays for (Weinberg, 2014).

Batting Average is another statistic that is getting less credit as a walk is equally as productive as a hit in most cases. It also doesn't distinguish between types of hits (Weinberg, 2015). "wRC+ is the most comprehensive rate statistic used to measure hitting performance because it takes into account the varying weights of each offensive action." With companies and teams tracking every play and game, we have advanced data for the 2017 season. There is available information on where each player hits the ball, how hard, and what type of pitch was hit from multiple different internet sources. Stats like RBI and Batting Average are just statistics that relay what happened on a certain play, but we have access to the raw data that comes before these numbers. The exit velocity, meaning the speed the ball leaves the bat is tracked as well as the angle and direction the ball goes. Strikeouts and walks are left out of the batted ball data because they don't involve a ball in play, but still have a vital role that a player has in being successful. It is common knowledge that hitting the ball Looking at all these factors, this paper examines the hypothesis that how hard you hit the ball leads to a better wRC+, and looks at the role other raw statistics play in effecting wRC+.

Literature Review

As listed in the Introduction, the most famous baseball stats are sometimes overvalued and given too much credit. Research amongst sabermatricians looking at tracked data like exit velocity and launch angle has been significant in increasing other performance statistics (Sapolsky, 2015). He concluded that for every mile per hour (MPH) a hitter is above the league average in exit velocity, he gains an additional eight points on his Weighted On Base Average (wOBA). wOBA is a statistic that has an aspect of run scoring added into its composition versus just measuring hits in the conventional style. Older statistics value a double as double a single, but this is not the case mathematically.(wOBA, Fangrpahs). wOBA is another statistic that is used to quantify a players production as a hitter as is wRC+. wRC+ uses wOBA as a main component in the

statistic as well as plate appearances. Scaling is done using park factors and league averages in other statistics. These numbers both are used to measure a hitter's value, but wOBA is included in wRC+ so intuitively, it makes sense to look for a positive correlation between these two statistics. The formula for wRC+ is

$$\text{wRC+} = \left(\frac{\left(\frac{\text{wRAA}}{\text{PA}} + \frac{\text{League R}}{\text{PA}}\right) + \left(\frac{\text{League R}}{\text{PA}} - \text{Park Factor} * \frac{\text{League R}}{\text{PA}}\right)}{\left(\text{AL or NL}\frac{\text{wRC}}{\text{PA}}\text{excluding pitchers}\right)}\right) * 100,$$

where wRAA/PA is to simply take a player's wOBA minus the League wOBA and divide it by the wOBA Scale (wRC and wRC+, Fangraphs). wRC is also in this statistic as it is the non-weighted version and is given by

$$\text{wRC} = \left(\left(\frac{\text{wOBA} - \text{League wOBA}}{\text{wOBA Scale}}\right) + \left(\frac{\text{League R}}{\text{PA}}\right)\right) * \text{PA}.$$

A change of one MPH in exit velocity relatively speaking is not too large and is attainable to a certain extent for each player if he is near the mean or league average. The further a player is above this average, the harder it can be to continually increase.

Statistics have been tracked by baseball teams and casual fans for years but as the game and technology grow closer together, more statistics are readily available that haven't been measured before. It is possible to look at a players sprint speed in feet per second throughout the game including while on the base paths or chasing down a ball in the outfield. His speed can be observed as the lead runner or as a hitter with a runner in front of him. Exit velocities and directional hitting in certain situations can be looked at in greater depth. (Arthur, 2016) He looked into exit velocity and how it affects the expected number of runs scored. "Hitting the ball harder doesn't always lead to more runs," is a headline in the article and talks about how the expected runs added decreases from exit velocity 73 MPH to 84 MPH and is negative from 80 MPH to 94MPH between 22 degrees and 28 degrees. This idea is based on those speeds being

caught by outfielders more often. It is important to look at exit velocities and launch angle together as a pair as they affect each other and success on the field. How do these factors affect wRC+ when player speed, batted ball direction, and other factors are included? Also according to this article, the expected runs added increases when the players are grouped into thirds and each tier gains .02 to .05 on expected runs added as the exit velocity varies. Do these correlations create more wRC+ or does it result in just expected runs added?

Hitting the ball hard is clearly important in Major League Baseball and statistically has proven to increase statistics include slugging percentage and batting average. Pulling the ball and having power to all fields are different strategies players may have. Tony Blengino examines pull ratios and the ratios between grounders and fly balls to look at success and what players are able to make this as a better approach to the plate. "In fact, almost all hitters fall into the same pattern; their ground ball pull ratio is higher than their liner pull ratio, which is higher than their fly ball pull ratio. (Blengino, 2016)" Of the 19 players that don't have the league average ratios, 3 players were in the top 10 in homeruns. These players have some strikeouts issues but being in the top 10 in homeruns in the league almost assures you are above a league average hitter meaning wRC+ would have been above 100. This article talks about a mix between success and failure in pulling the ball and going the opposite way. Certain players have success pulling if they have more air under it while those on the ground struggle more. Does this result extend to wRC+?

Model

This paper will examine the effect of several statistics across all of Major League Baseball and its players who accumulated 150 Batted Ball Events according to (Willman, Baseball Savant) and the relevance statistics have on explaining wRC+. This will be conducted by looking at a Regression Model and will include percentile grouping some of the statistics. Also observed will

be stepwise regression from select variables looking for a correlation in hitting variables that are most important. This will be done both forwards and backwards to align the valuable variables. The model used is as followed:

$$\mathbf{wRC+} = \beta_1 + \beta_2\mathbf{aev} + \beta_3\mathbf{ala} + \beta_4\mathbf{bbp} + \beta_5\,\boldsymbol{n}\mathbf{ss} + \beta_6\mathbf{nnfp} + \beta_7\mathbf{bbpa} + \beta_8\mathbf{ldp} + \beta_9\mathbf{kp} + u$$

As the algebraic representation of the model above indicates the dependent variable in this study will wRC+ across all players in the 2017 season. The main hypothesis is that Average exit velocity, walk percentage, and launch angle have a significant effect on Weighted Runs Created Plus. The higher these are, the higher wRC+ total should be. The significance that the remaining independent variables have on Weighted Runs Created Plus will also be examined in multiple different ways. Grouping the players into percentiles will also be looked at for far more generic trends.

The expected sign for the aev is positive because as exit velocity increases and the ball is hit harder it harder to catch as it may spend less time in the air or ground through the holes in the infield. Strike-out percentage (kp) should have a negative coefficient as it is making an out. It might not be as negative as some would think because a mentality that has become more popular is looking to drive the ball even with two strikes. Strike-outs are viewed as not as negative as they have in the past. Walk Percentage (bbp) should increase wRC+ as the value of a walk is gaining more equivalence to that of a hit. It is getting on base and in some cases moving a runner into scoring position or even in to score. Sprint speed (ss) is expected to have a positive connotation because the faster you are, the more infield hits you can beat out and the more extra bases you can take on hits into the gap or that a fielder is slow to receive. The expected sign of mph is positive because this measures balls hit over 95 miles per hour. This has been seen as above league average but also a point in certain launch angles that the expected runs added is

positive and increasing. Ala should be positive because it is the angle the ball leaves the bat. The higher it is the more likely it could be a homerun or a ball into the gap allowing for extra bases. Low exit velocities are grounders more often and if make they make it past an infielder will not normally make it past an outfielder. Barrels per plate appearance is expected to be positive as well because these are balls in play that have been contacted at a certain launch angle and exit velocity in combination. "A batted ball requires an exit velocity of at least 98 mph. At that speed, balls struck with a launch angle between 26-30 degrees are always a barrel. (What is a barrel?, MLB.com)" Then for an increase in exit velocity per mile per hour, the degrees add two or three to its range as a barrel classification. This continues until 116 miles per hour where the range is 8 and 50 degrees. Hits of this designation have a batting average over .500, so are clearly going to positively affect wRC+

One thing to be careful of is variables and their interactions. Barrels are a combination of launch angle and exit velocity. This is a designation when these are at their best, so there could be some multicollinearity but the main point of having these is to distinguish between someone who may have pop-ups and ground balls equally, resulting in an average exit velocity that is more in the middle with having very few hits resulting in this ideal launch angle and exit velocity. It also might help us identify if someone does hit line drives a lot, but not at the ideal exit velocity. The speed of the ball upon exit is what really gives value in the ideal launch angles and if you are below that, outs could be the ramifications.

The other problem to monitor is the interaction between ldp, fbp, and gbp, and pullp, centercp, and oppop. Each trio is percentages that add up to one. These are not count variables so there could be a problem as having two of the variables assumes what the third one is in value. It

could be valuable to have some of these in as it gives a breakdown of how players hit the ball and can be looked at in conjunction with average launch angle.

Data

Table 1: Original Variables considered in model

Variable name	Name in baseball world	Description	Where information was acquired
wRC+	Weighted Runs Created Plus (wRC+)	A rate statistic to quantify offense that is ballpark and era adjusted.	Fangraphs.com
aev	Average Exit Velocity	The average speed a ball leaves the bat.	MLBsavant.com
Kp	Strikeout Percentage	The percentage of at bats a hitter strikes out	Fangraphs.com
bbp	Walk Percentage	The percentage of at bats hitter gets a bases on balls	Fangraphs.com
(n)ss	Sprint Speed in FT/S	The speed measurement that quantifies speed of a player in his fastest one second	Mlbsavant.com
(n)nfp	Batted balls 95MPH+	Balls that are hitter faster than 95MPH	Mlbsavant.com
Barrels	Barrel divided by Plate appearance	Ball hit that has a Batting average of .500 and a slugging percentage of 1.500 in similarly hit balls.(Launch angle& exit velocity)	Mlbsavant.com
Ldp	Line Drive Percentage	The percentage of balls hit that are line drivers	Fangraphs.com
Fbp	Fly Ball Percentage	The percentage of balls hit that are fly balls	Fangraphs.com
Gbp	Ground Ball Percentage	The percentage of balls hit that are grounders	Fangraphs.com
(n)Pullp	Pull Percentage	The percentage of balls hit to the pull side of the field	Fangraphs.com
Centerp	Center Percentage	The percentage of balls hit up the middle	Fangraphs.com
Oppop	Opposite Percentage	The percentage of balls hit to opposite field	Fangraphs.com
Ala	Average launch angle	The average angle the ball is batted	Mlb.com

(n) is a created discrete variable that is categorical and making top percentage around 15% and bottom percentage around the same based on graphs and distributions. See appendix and code for specific values

The data used in this paper was obtained from Major League Baseball, MLBSavant.com, and Fangraphs. Other resources that I used in accumulating data were Baseball Reference and ESPN. The data used in this research is for all Major League Baseball Players in the 2017 season with 150 Batted Ball Events. This is a way to eliminate smaller sample sizes to skew the data. It is stated that in Baseball even a season can be a small sample size for certain players, but in this case it keeps the players in the sample the same. If we were to cross over seasons, at bats could be even more dramatic as some players did not play in the major leagues in 2016. The data set is balanced as all the players have observations for each variable that is up to date and accurate.

Table 2: Summary of Raw Data

Variable name	Average	Standard Deviation	Count	Minimum	Maximum
wRC+	101.4341317(100)	24.5331287	334	33	181
aev	87.0251497	2.595510139	334	77.2	94.9
kp	.20665296	0.058440453	334	.091	.378
bbp	.08596	0.032042853	334	.19	.021
(n)ss	27.0997006	1.178563484	334	23	30.2
(n)nfp	112.0838323	45.75198869	334	19	250
barrels	4.380838323	2.34247257	334	0	12.8
ldp	.204907186	0.031300845	334	.311	.134
Fbp	.35863173	0.067611334	334	.187	.542
gbp	. 436886228	0.064365109	334	.269	.627
(n)pullp	.401889222	0.056037544	334	.216	.543
Center	. 347164671	0.033596137	334	.244	.447
Oppop	.250949102	0.042927321	334	.148	.385
Ala	.12068207207(11.75)	4.180313091	334	.43	.023

Empirical Data

Working with all the variables above, a test for multicolinearity found that Pullp, oppop, and centerp were all coliner and related in some ways which was mentioned early on in the paper as a possibility. Also noted is the high Variance inflation on ground ball percentage and fly ball percentage. Oddly it is not as high on the line drive. This could point to similarities in results in

fly balls and ground balls. (Appendix A) Extra consideration was also be given to average launch angle and line drive percentage in the next procedure with less variables to look at, we can see more in-depth the value they bring. To offset this problem, pull percentage was turned into a discrete variable helping identify the top pull hitters based on a normal distribution. This was also done for the number of balls ninety five plus and sprint speed. (Appendix D)This is why the glm procedure is used as it give attention to discrete variables versus continuous.

A proc reg procedure looking at just aev, bppa, ala, ldp was also conducted as these all are attempting to measure very similar things. The angle the ball leaves the bat and the speed it leaves the bat so testing for multicolinearity was done and it was found that there was no multicolinearity between the variables. All the variables had low variance inflations which is a way to test for multicolinearity. Using the common threshold of 5, there is no multicolinearity between the variables meaning we included them all in the stepwise regression in finding the best model.(Appendix B).

The data was used on the normal model with all the original data using a forward and backward regression and Average exit velocity and launch angle were forced into the model as not at all percentages of entry and exit they were kept in the model. In most cases average exit velocity had the third highest criterion for entry, but was lowered to not as statistically significant for entry with the inclusion of barrels per plate appearance.(Appendix C)

There is more value in the discrete variables as they have they are showing who is elite in the variable groups. The fastest players, the players who hit the most balls above average exit velocity, and pull the ball the most should be considered as an area to look at. For this procedure, we will use glm because this is more powerful when identifying discrete variables

and the role they play in the regression equation. Glmselect is another way, similar to a forward and backward regression that can help identify the best variables for a model. The other important thing is to automatically keep the variables Aev and Ala because it has been proven in an abundance of papers above that exit velocity and launch angle increases have increased offensive output. Using glmselect to find the most efficient model adding variables into the equation, we find out again walk percentage, strikeout percentage, the discrete variable for sprint speed, the discrete variable ninety five plus, line drive percentage and barrels per plate appearance were all added (Appendix E) . Excluded again was ala and aev, so bppa is taken out as these were the variables that multicolinearity was looked for, as they measure similar things though it was not found. The difference in this next model is aev is now in the model as is fly ball percentage. This makes sense that fly ball is added versus groundball because fly ball represents the higher launch angle, which is how many people have altered there swing in favor of this movement (Appendix F). Using the two models with aev, bbp, and ala forced into the model, we keep the variables kp, nss, nnfp, bbpa, and ldp.

Table 3: Model including Bppa

Parameter	Estimate	Standard Error	T-Value	Pr > |t|
Intercept	66.1834938	47.434822	1.4	0.1639
AEV	-0.0209448	0.55293177	-0.04	0.9698
Ala	-0.1373571	0.21858149	-0.63	0.5302
Bbp**	227.20268	27.9982043	8.11	<.0001
Kp**	-174.35314	17.3155471	-10.07	<.0001
nss 0**	-11.355947	3.13939906	-3.62	0.0003
nss 1**	-7.2271393	2.58785107	-2.79	0.0055
nss 2	0	.	.	.
nnfp 0*	-9.8805593	4.05187092	-2.44	0.0153
nnfp 1	-0.3681699	2.70815033	-0.14	0.8919
nnfp 2	0	.	.	.
Bppa**	7.1840271	0.64032719	11.22	<.0001
Ldp**	158.308388	27.2845322	5.8	<.0001

* Means significant at 95% **means significant at 99%

Table 4: The GLM Procedure with bppa

Source	DF	Sum of Squares	Mean Square	F-Value	Pr > F
Model	10	128458.6709	12845.8671	57.18	<.0001
Error	323	72567.38	24.6668		
Corrected Total	333	201023.0509			

The first model that was run incorporated the discrete variables and included the variable bppa. As we can see from table 4, the f- value is high, meaning we have a significant model in determining wRC+. From table 3, we see bppa is the most significant variable in the model with a t-value of 11.22. This is valuable as bppa measures the percentage of barrels. Insightful also is that average launch angle and average exit velocity are deemed statistically insignificant to the model. This is due to the inclusion of bppa in the model. Walk percentage and strikeout percentage are the statistically significant. This is interesting as it incorporates what does happen on balls not in play and has identified the strikeout percentage and walk percentage as negative and positive respectively to wRC+ respectively when contributing to the model. Line drive percentage is statistically significant which is what we assumed because it helps identify the launch angles that are more ideal the higher they are. Also, this model has an R- squared value of 0.639015 meaning almost two thirds of the variation in the model is accounted for by this model. The discrete variables were included as classes meaning the first variable, nnfp 2 is the baseline as it is the hitters who hit the most balls ninety five miles per hour plus. The other two groups hit less than that and that is why they have negative values, and the nnfp 0 is statistically significant meaning the people in this category have the fewest, and it is negatively effecting the model. The same goes for nss as nss 2 is the baseline, but in

this case both variables are statistically significant meaning the slower you are, the lower the wRC+ you will have based on these groups. The intercept value also is not significant.

Table 5:Model without bppa

Parameter	Estimate	Standard Error	T-value	Pr > |t|
Intercept**	-226.9691456	46.59810704	-4.87	<.0001
AEV**	3.5507017	0.53212884	6.67	<.0001
Ala*	0.6103507	0.24502777	2.49	0.0132
Bbp**	250.163375	32.86669476	7.61	<.0001
Kp**	-95.6383827	18.63300806	-5.13	<.0001
nss 0*	-9.359436	3.68923955	-2.54	0.0117
nss 1	-5.6729816	3.04162271	-1.87	0.0631
nss 2	0	.	.	.
nnfp 0**	-17.7382312	4.69740546	-3.78	0.0002
nnfp 1	-5.0083421	3.15019426	-1.59	0.1128
nnfp 2	0	.	.	.
Ldp**	107.2872676	31.66561379	3.39	0.0008

* Means significant at 95% **means significant at 99%

Table 6:The GLM Procdure without bppa

Source	DF	Sum of Squares	Mean Square	F-Value	Pr > F
Model	9	100179.2202	11131.0245	35.75	<.0001
Error	324	100846.8307	311.2557		
Corrected Total	333	201023.0509			

After identifying the lack of value in average exit velocity and average launch angle in model one and knowing the importance of these in baseball research, I took out bppa to see if this added value to the other variables that we thought were extremely important. Walk Percentage now has the highest t-value as show in table 5. By removing bppa, the second highest t-value goes to average exit velocity meaning it is statistically significant and important to the model. Average launch angle also is significant at the 5% level. These items were tested

for multicolinearity before the model was decided and there was said to be no multicolinearity statistically, but when barrels per plate appearance is removed, both average exit velocity and average launch angle are significant. This model also shows significance in line drive percentage, and both the discrete variables nss 0 and nnfp 0. Also noteworthy is the intercept now being statistically significant. The strike out percentage is still highly statistically significant and has a negative parameter. The model itself is also significant with an F value of 35.75 as seen in table 6. This is lower than model one meaning it has less power put is still valuable in predicting wRC+.

The consistency between the two models outside of changing bppa is encouraging as the other parameters keep the same sign and significance in the models. Barrels per plate appearance can be seen as a more effective version of the combination of average launch angle and average exit velocity. This is no surprise as barrels have statistically produced higher batting averages and slugging percentages. The value is in average launch angle and average exit velocity still being significant in predicting wRC+.

Prediction

Though both models are significant we will use the model that incorporates barrels per plate appearance because this is a more efficient model and a crucial statistic in predicting wRC+. When looking at prediction, assuming the sample sizes are big enough and the players replicate their exact statistics from the 2017 season, these players are due for an increase in wRC+ as their residuals values are the furthest below the regression line. Meaning, the predicted values based on their performance statistics gave them a much higher predicted value than how they performed. The list of players who should enjoy more success in the coming season if they

replicate how well they performed is listed in table 7 based on studentized residuals outside of the absolute value 1.8.

Table 7:Predictions on increasing season

Name	wRC+	Resid	Predicted value	Stu_resid
D.J. Lamehieu	94	-58	152	-4.04
Miguel Cabrera	91	-47	138	-3.2
Mitch Moreland	98	-35	133	-2.33
Dansby Swanson	66	-34	100	-2.33
Brandon Moss	84	-34	118	-2.28
Alex Gordon	62	-31	93	-2.1
Austin Romine	49	-30	79	-2.05
Maikel Franco	76	-30	106	-2.04
Chris Herrman	58	-29	87	-2.01
Taylor Motter	57	-30	87	-1.99
Hyun Soo Kim	61	-28	89	-1.94
Pablo Sandoval	64	-29	93	-1.94
J.J. Hardy	50	-28	78	-1.9
Randal Grichuk	94	-27	121	-1.85
Tony Walters	49	-27	76	-1.81

Looking at the table above, a lot of these players are either veterans who had a bad season like Miguel Cabrera, Alex Grodon, Pablo Sandoval and Brandon Moss. Others are players in their first or second year and could be looking at a big jump in production like Randal Grichuk, Dansby Swanson, or Maikel Franco. This could be useful looking ahead into a new season and trying to identify weak spots in a roster or wondering if a player is going to be able to output at a certain level they had in the past. I know D.J. Lamehieu may struggle to reach a wRC+ of 152, but this could be beneficial, being he has room to grow his wRC+ and that it is likely to increase.

As some players underperformed their numbers, there is a group who over performs their numbers. This list is most likely going to be players at the top of the league as it is hard to be consistently way better than everyone else. These players also might be outliers in this model because they are the best players and can outplay the numbers. The reason at the top

this is more attainable is because the players at the bottom of the list are recycled out as people are filling in their spots and looking for replacements. The list of players who are outperforming statistically where they should be is in table 8.

Table 8: Predictions on who could Regress

Name	wRC+	Residual	Predicted Value	Stu-Resid
Jose Altuve	160	38	122	2.56
Mike Trout	181	35	146	2.42
Mitch Haniger	129	36	93	2.41
Marwin Gonzalez	144	35	109	2.38
Zack Cozart	141	32	109	2.12
Austin Jackson	131	31	100	2.08
Jose Rameriz	148	31	117	2.07
Eduardo Nunez	112	29	83	1.96
Marcell Ozuna	142	27	115	1.85
Scooter Gennett	124	27	97	1.84
George Springer	150	27	123	1.8

As you can see, some of the top players in the game are on this list as it seems like they can outperform predictions like Mike Trout, Jose Altuve, Jose Rameriz, George Springer and Marcell Ozuna. Say these guys do regress; they will still be above average players as will most on this list. Mitch Haniger, Marwin Gonzalez, Scooter Gennett, and Eduardo Nunez would be closer to an average player if they do indeed regress towards the mean. A lot of the players on both of these lists are also on a high cook's D list (Appendix H). The players on the high cook's D list have the most influence on the model because the residuals they have are far from their predicted value. They negatively affect the model because their residual value and leverage have an altering effect on the model. In some case, a residual may be so low that it tries to over compensate in the model to attribute to one value. A common cut off in mathematics is 4/n as

a cut off for this value to be excluded. This formula gives us .01197 as a test number. These 14 players could be removed as they have a negative effect on the model and have altered the regression line. After removing these players from the model, we see the value and accuracy of the model increase as it now has an F value of 69.82 which is up from 57.18. The r squared value also increased to 69.32 (Appendix I). Average Launch angle became a lot close to significant as it is now has a p value below .10 while before it was above .53. The players deleted could be outliers in average launch angle and that is why it is now more valuable because the extreme variance added to the model by those players is now gone and the true value of the parameter can be identified (Appendix J). None the less, the players deleted had a strong negative influence on the model as you can see; taking them out increased the accuracy of the model. This is proof should move more towards the mean this coming year.

Future Research

Miguel Cabrera was extremely unlucky last year as he had a career low in Batting Average on Balls in Play (Miguel Cabrera, Fangraphs). This could be another stat that is incorporated in this study for the future or looking at players with large residuals and if BABIP is affecting over performance or under performance. Another thing that could be looked at is this data over years and seeing if the people with the highest cook's D values are consistently the same. If they are you, could make the argument they are constantly over performing or underperforming and then create a dummy variable as to whether they are and how that effects the models for players who constantly over perform and under perform. These could be useful as other models for players who consistently effect the regular model found in this research paper.

Conclusion

Baseball is a game where statistics are playing a large role in the game than they ever have before. Teams are spending more money on front office analytic minds to make decisions and finding a recipe to get a good model to represent raw statistics that can project how players will perform. Based on the Empirical data above, we can conclude that Barrels per plate appearance, walk rate, strike out rate, Line-drive percentage, sprint speed, and the number of times you hit the ball 95 mph+ are all statistically significant in determining wRC+. We can conclude as well that average launch angle and average exit velocity are significant when replaced with barrels per plate appearance. This allows us to say that barrels per plate appearance is a more powerful and effective measure when determining wRC+. The positive values to barrels per plate appearance, walk percentage, line drive percentage, and sprint speed are all expected results. The more efficient you strike the ball the better as well as the more you walk or receive a “free pass”, the better. The faster a player is, the more effective they can be at beating out hits, turning singles into doubles, and doubles into triples. These added total bases to hits can add up over a season and added value to wRC+. The negative strike out rate is expected to be negative and is because a strike out is an out that doesn’t help the team. Certain at bats and situations a ball in play can be a productive out, scoring a run or moving a man over in a sacrifice fly or ground out to the right side of the diamond, but a strike out does not allow for that. wRC+ is an all-encompassing offensive stat that really allows for comparison across seasons. With the average being 100 and one over that is one percent better, it is an easy way to use this model for next year.

Works Cited

Arthur, Rob. “The New Science Of Hitting.” *FiveThirtyEight*, FiveThirtyEight, 18 Aug. 2016, fivethirtyeight.com/features/the-new-science-of-hitting

Blengino, Tony. “The Pros and Cons of Pulling the Baseball.” *FanGraphs Baseball*, 12 May 2015, www.fangraphs.com/blogs/the-pros-and-cons-of-pulling-the-baseball/

“Miguel Cabrera Statistics.” *FanGraphs Baseball*, www.fangraphs.com/statss.aspx?playerid=1744&position=1B%2F3B.

Sapolsky , William. “Improving Projections with Exit Velocity.” *The Hardball Times*, 2 May 2016, www.fangraphs.com/tht/improving-projections-with-exit-velocity/

Weinberg, Neil. “Stats To Avoid: Runs Batted In (RBI) .” *FanGraphs Sabermetrics Library*, 24 Oct. 2014, www.fangraphs.com/library/stats-to-avoid-runs-batted-in-rbi/.

Weinberg, Neil. “Stats to Avoid: Batting Average .” *FanGraphs Sabermetrics Library*, 20 Feb. 2015, www.fangraphs.com/library/stats-to-avoid-batting-average/.

“What Is a Barrel? | Glossary.” *Major League Baseball*, MLB, m.mlb.com/glossary/statcast/barrel.

Willman, Darren. “Statcast Hit Probability.” *Baseballsavant.com*, MLB, baseballsavant.mlb.com/statcast_hit_probability.

Wilson, Jason. “Measuring the Quality of a Pitch.” *The Hardball Times*, FanGraphs, 16 Mar. 2017, www.fangraphs.com/tht/measuring-the-quality-of-a-pitch/.

“WOBA .” *FanGraphs Sabermetrics Library*, www.fangraphs.com/library/offense/woba/.

“WRC and WRC+.” *Fangraphs* , www.fangraphs.com/library/offense/wrc/

Appendix

A)

Variable	Df	Parameter	Standard Error	T-Value	Pr> |t|	Tolerance	Variance inflation
Intercept	1	-22583	8732.65814	-2.59	0.0101	.	0
AEV	1	0.06191	0.58961	0.11	0.9164	0.27547	3.63012
Kp	1	-173.32801	17.85903	-9.71	<.0001	0.59226	1.68846
bbp	1	221.31681	27.33259	8.1	<.0001	0.84106	1.18897
Ss	1	3.72066	0.75041	4.96	<.0001	0.82479	1.21242
nfp	1	0.05543	0.02647	2.09	0.0371	0.43978	2.27384
bppa	1	7.34294	0.65741	11.17	<.0001	0.27204	3.67593
ldp	1	143.93216	108.80729	1.32	0.1868	0.05562	17.97939
fbp	1	-15.88854	110.12951	-0.14	0.8854	0.01164	85.9396
gbp	1	-22.33645	104.59183	-0.21	0.831	0.01423	70.24953
pullp	1	22542	8737.15894	2.58	0.0103	2.69E-06	371574
centerp	1	22568	8739.57876	2.58	0.0103	7.48E-06	133631
oppop	1	22518	8738.19452	2.58	0.0104	4.59E-06	218101
ala	1	-0.19976	0.66786	-0.3	0.7651	0.08277	12.08185

```
proc reg data=honors.project1;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
        /tol vif collin;
        run;
```

B)

Variable	Df	Parameter	Standard Error	T-Value	Pr> |t|	Tolerance	Variance inflation
Intercept	1	-119.74891	52.7302	-2.27	0.0238	.	0
AEV	1	1.85789	0.62455	2.97	0.0031	0.41626	2.40237
bppa	1	5.00018	0.74329	6.73	<.0001	0.3608	2.77162
ala	1	-0.08842	0.27423	-0.32	0.7473	0.8323	1.20148
ldp	1	188.68178	34.18759	5.52	<.0001	0.95518	1.04693

```
proc reg data=honors.project1;
        model wrcp=aev bppa ala ldp
        /tol vif collin;
        run;
```

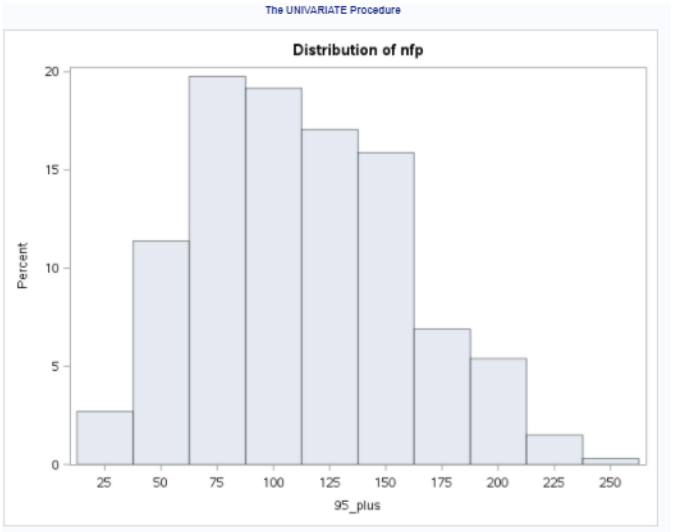
C) Criterion for entry before vs after the first variable is selected

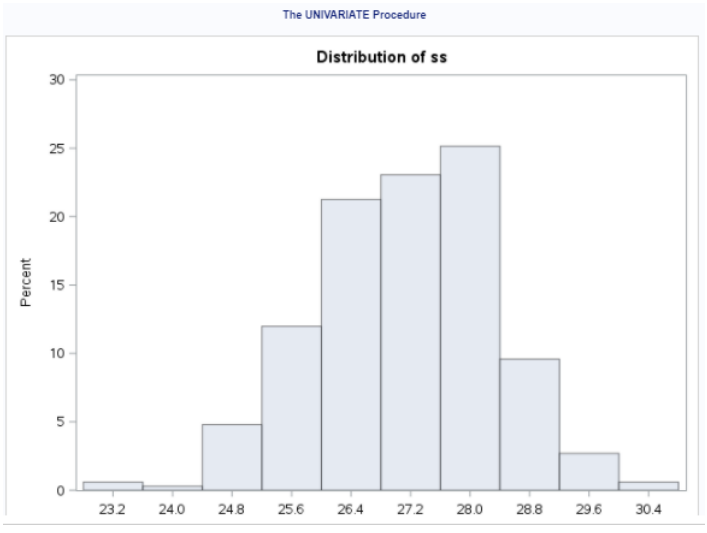
Variable	Tolerance	Model R- Square	F-value	Pr > F
AEV	1	0.2741	125.39	<.0001
kp	1	0.0077	2.59	0.1087
bbp	1	0.2208	94.1	<.0001
ss	1	0	0.02	0.8979
nfp	1	0.2917	136.72	<.0001
bppa	1	0.3278	161.93	<.0001
ldp	1	0.0144	4.85	0.0284
fbp	1	0.0444	15.41	0.0001
gbp	1	0.0757	27.19	<.0001
pullp	1	0.0025	0.82	0.365
centerp	1	0.0016	0.53	0.469
oppop	1	0.0092	3.09	0.0795
ala	1	0.0409	14.15	0.0002

Variable	Tolerance	Model R Square	F-value	Pr > F
AEV	0.425811	0.3468	9.58	0.0021
kp	0.798584	0.4768	94.25	<.0001
bbp	0.906203	0.4236	54.99	<.0001
ss	0.949983	0.3433	7.77	0.0056
nfp	0.742857	0.4118	47.24	<.0001
ldp	0.960195	0.385	30.73	<.0001
fbp	0.759272	0.3344	3.24	0.0729
gbp	0.83133	0.3298	0.95	0.3307
pullp	0.919809	0.3416	6.91	0.009
centerp	0.98504	0.3401	6.14	0.0137
oppop	0.924825	0.3319	1.99	0.1596
ala	0.853726	0.3282	0.16	0.6862

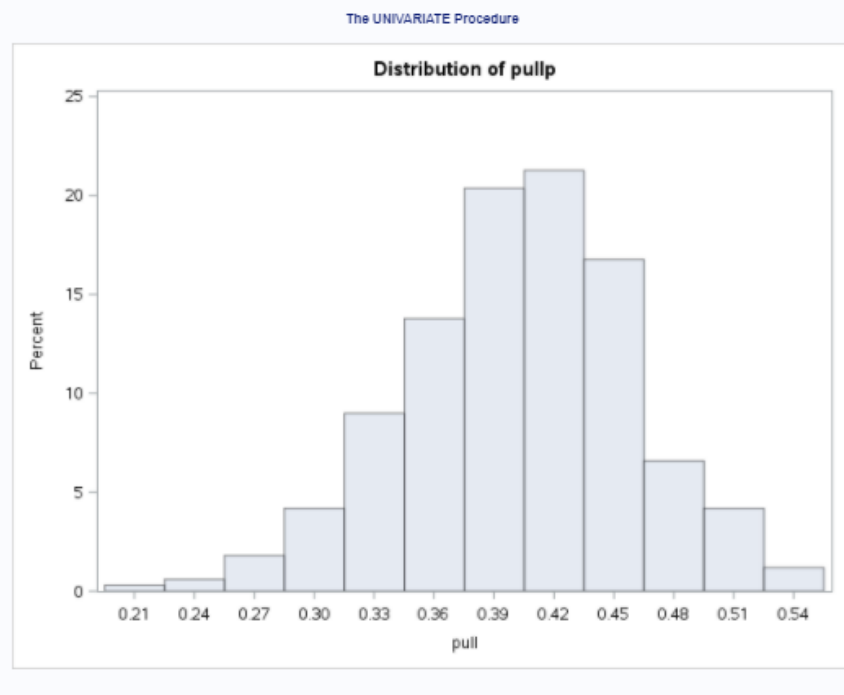
```
proc reg data=honors.project1;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
        /selection=stepwise slentry=.15 slstay=.15 details;
        Run;
```

D)





D Cont.)



Proc univariate data=honors.project;
var ss nfp pullp;
histogram;
run;

E)

Parameter	DF	Estimate	Standard Error	T-Value
Intercept	1	63.808225	7.283991	8.76
bbp	1	225.25092	27.232516	8.27
kp	1	-174.7705	17.260093	-10.13
nss 0	1	-11.504701	3.080927	-3.73
nss 1	1	-7.349718	2.554043	-2.88
nss 2	0	0	.	.
nnfp 0	1	-10.031789	3.698749	-2.71
nnfp 1	1	-0.511176	2.602511	-0.2
nnfp 2	0	0	.	.
bppa	1	7.07999	0.502381	14.09
ldp	1	157.52831	27.184866	5.79

```
proc glmselect data = honors.project1;
       class nss npullp nnfp;
       model wrcp=aev ala  bbp kp nss nnfp bppa ldp
fbp gbp npullp centerp oppop  /details= all selection = stepwise;
run;
```

F)

Parameter	DF	Estimate	Standard Error	T-Value
Intercept	1	-206.23151	45.76553	-4.51
AEV	1	3.101174	0.522086	5.94
bbp	1	251.78247	32.945971	7.64
kp	1	-91.072906	18.339211	-4.97
nnfp 0	1	-19.277274	4.677867	-4.12
nnfp 1	1	-5.706493	3.151135	-1.81
nnfp 2	0	0	.	.
ldp	1	129.44545	33.38017	3.88
fbp	1	42.522332	16.367487	2.6

```
proc glmselect data = honors.project1;
       class nss npullp nnfp;
       model wrcp=aev ala  bbp kp nss nnfp ldp
                      fbp gbp npullp centerp oppop  /details= all
selection = stepwise;
run;
```

G)

Paramater	DF	Estimate	Standard Error	T-Value
Intercept	1	66.183494	47.434822	1.4
AEV	1	-0.020945	0.552932	-0.04
ala	1	-0.137357	0.218581	-0.63
bbp	1	227.20268	27.998204	8.11
kp	1	-174.35315	17.315547	-10.07
nss 0	1	-11.355947	3.139399	-3.62
nss 1	1	-7.227139	2.587851	-2.79
nss 2	0	0	.	.
nnfp 0	1	-9.880559	4.051871	-2.44
nnfp 1	1	-0.36817	2.70815	-0.14
nnfp 2	0	0	.	.
bppa	1	7.184027	0.640327	11.22
ldp	1	158.30839	27.284532	5.8
*Forced into the model by the Include=option				

```
proc glmselect data = honors.project1;
       class nss npullp nnfp;
       model wrcp=aev ala  bbp kp nss nnfp bppa ldp
                      fbp gbp npullp centerp oppop  /include = 3
details= all selection = stepwise;
run;
```

H)

Obs	Name	Cooks D
1	D.J. Lemahieu	0.12631
2	Miguel Cabrera	0.05057
3	Mike Trout	0.03358
4	Chris Herrman	0.01886
5	Dansby Swanson	0.01834
6	Alex Presley	0.01713
7	Hyun Soo Kim	0.01666
8	Jose Altuve	0.01542
9	Buster Posey	0.01398
10	Randal Grichuk	0.01375
11	Brandon Moss	0.01309
12	Tony Wolters	0.01288
13	Ezequil Carrera	0.01287
14	Jose Rameriz	0.01271

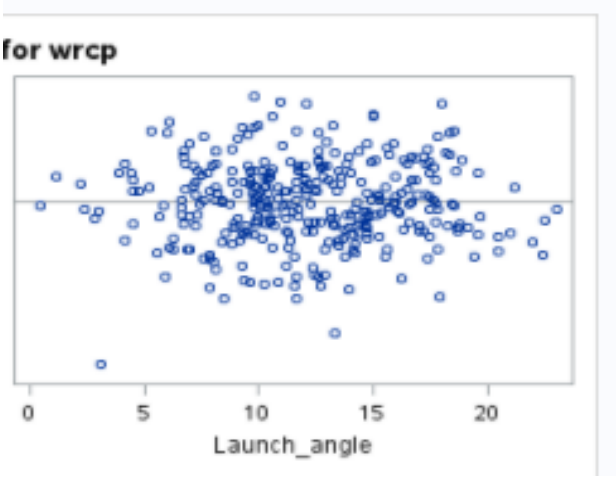
I)

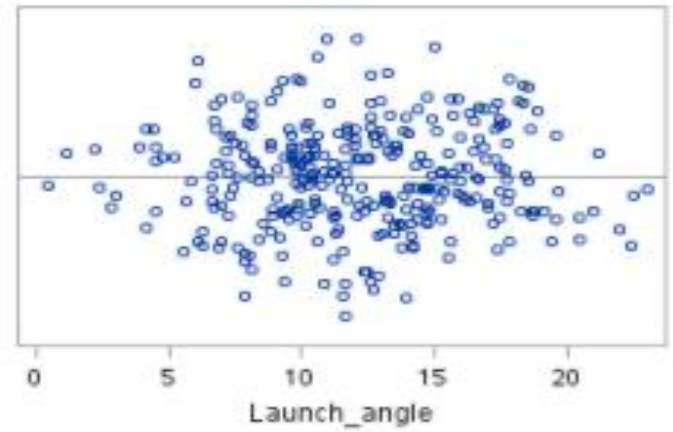
Parameter	Estimate	Standard error	T-Value	Pr > |t|
Intercept	52.9758367	44.35003082	1.19	0.2332
AEV	0.0759428	0.51648426	0.15	0.8832
ala	-0.3364781	0.20107571	-1.67	0.0953
Bbp**	219.4499824	25.60699269	8.57	<.0001
Kp**	-165.7002255	15.7922918	-10.49	<.0001
nss 0**	-11.2876926	2.865236	-3.94	0.0001
nss 1**	-6.7250057	2.37382198	-2.83	0.0049
nss 2	0	.	.	.
nnfp 0*	-9.1086093	3.76379614	-2.42	0.0161
nnfp 1	-1.4579365	2.47357815	-0.59	0.556
nnfp 2	0	.	.	.
Bppa**	7.6076802	0.60539329	12.57	<.0001
Ldp**	182.2024227	25.36021621	7.18	<.0001

Source	DF	Sum of Squares	Mean Square	F-Value	Pr
Model	10	124979.3212	12497.9321	69.82	<.0001
error	309	55310.228	178.9975		
Corrected Total	319	180289.5492			

```
data withoutdj;
       set res_infl;
       where cook lt 4/334;
run;
proc glm data= withoutdj;
       class nss npullp nnfp;
       model wrcp= AEV ala  bbp kp nss nnfp bppa ldp;
run;
```

Appendix J)





All Code)

```
/* Project */
/* CHecking COntents */
proc contents data=honors.projectdataset;
run;
/* Printing data set as a whole */
proc print data=honors.projectdataset;
run;
/* Renaming data */
data honors.Project;
	set honors.projectdataset (rename=(Average_exit_velocity=AEV wrc_plus=wrcp
k_pct=kp bb_pct=bbp Sprint_speed=ss _95_plus=nfp Barrels=bppa ld=ldp fb=fbp gb=gbp
pull=pullp center=centerp oppo=oppop launch_angle=ala));
run;

/* Checking for normality */
proc univariate data=honors.project normal plots;
var wrcp;
run;/* look at Shapiro-Wilk test   also */
/* Checking to make data categorical by distribution */
Proc univariate data=honors.project;
var ss nfp pullp;
histogram;
run;
/* Making new variables based on ones that can be come discrete */
data honors.project1;
	set honors.project;
/*	nss=0; */
/*	if ss ge 23 and ss lt 26 then nss=0; */
/*	if ss ge 26 and ss lt 28.4 then nss=1; */
/*	if ss ge 28.4 then nss=2; */
	if ss^= . then do;
	if ss lt 26 then nss=0;
	else if ss lt 28.4 then nss=1;
	else nss=2;
	end;
/*	npull=0; */
/*	if pullp*100 ge 19 and pullp*100 lt 34.5 then npullp=0; */
/*	if pullp*100 ge 34.5 and pullp*100 lt 46.5 then npullp=1; */
/*	if pullp*100 ge 46.5 then npullp=2; */
	if pullp^= . then do;
	if pullp lt .345 then npullp=0;
	else if pullp lt .465 then npullp=1;
```

```
            else npullp=2;
            end;
/*          nnfp=0; */
/*          if nfp ge 0 and nfp lt 62.5 then nnfp=0; */
/*          if nfp ge 62.5  and nfp lt 162.5 then nnfp=1; */
/*          if nfp ge 162.5 then nnfp=2; */
            if nfp^= . then do;
            if nfp lt 62.5 then nnfp=0;
            else if nfp lt 162.5 then nnfp=1;
            else nnfp=2;
            end;
run;

            proc print data=honors.project1;
            run;
data project1;
            set honors.project1;
            if Name = ' Bryce Harper' then name = 'Bryce Harper';
run;
proc sort data = project1 out =honors.project1;
            by name;
run;
            proc freq data=honors.project1;
            tables nss npullp nnfp;
run;
            proc means data=honors.project1;
            var nss npullp nnfp;
run;
proc sort data=honors.project;
by name;
run;
proc sort data=honors.project1;
by name;
run;
/* Standardizing the data the is continuous */
proc standard data=honors.Project mean=0 std=1 out=stdtest;
var aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala;
run;
/* Checking contents on new data set */
proc contents data = stdtest; run;
/* printing new data set */
proc print data=stdtest;
run;
proc corr data= honors.project1;
```

```
        var  wrcp aev kp bbp nss nnfp bppa ldp fbp gbp npullp centerp oppop ala;
        run;
/* Testing for Multicollinearity within the data */
proc reg data=honors.project1;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
        /tol vif collin;
        run;
proc reg data=honors.project1;
        model wrcp=aev kp bbp nss nnfp bppa ldp fbp gbp npullp centerp oppop ala
        /tol vif collin;
        run;
proc reg data=stdtest;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
        /tol vif collin;
        run;
proc reg data=honors.project1;
        model wrcp=aev kp bbp ss nfp bppa ldp  pullp  ala
        /tol vif collin;
        run;
proc reg data=honors.project1;
        model wrcp=aev kp bbp nss nnfp bppa ldp  npullp  ala
        /tol vif collin;
        run;
proc reg data=honors.project1;
        model wrcp=aev bppa ldp
        /tol vif collin;
        run;
        proc reg data=honors.project1;
        model wrcp=aev bppa ala ldp
        /tol vif collin;
        run;

/*      stepwise not included */
/*stepwise functions going forward standard*/
proc reg data= stdtest;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
        /selection=stepwise slentry=.15 slstay=.15 details;
        run;
        proc reg data= stdtest;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
        /selection=stepwise slentry=.10 slstay=.15 details include=1;
        run;
        proc reg data= stdtest;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
```

```
        /selection=stepwise slentry=.05 slstay=.15 details include=1;
        run;
        proc reg data= stdtest;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
        /selection=stepwise slentry=.05 slstay=.05 details include=1;
        run;
/*      stepwise forward not standard */
proc reg data=honors.project1;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
        /selection=stepwise slentry=.15 slstay=.15 details;
        run;
        proc reg data= honors.project1;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
        /selection=stepwise slentry=.10 slstay=.15 details include=1;
        run;
        proc reg data=honors.project1;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
        /selection=stepwise slentry=.05 slstay=.15 details include=1;
        run;
        proc reg data=honors.project1;
        model wrcp=aev kp bbp ss nfp bppa ldp fbp gbp pullp centerp oppop ala
        /selection=stepwise slentry=.05 slstay=.05 details;
        run;
/*      with newer variables */
        proc reg data=honors.project1;
        model wrcp=aev kp bbp nss nnfp bppa ldp fbp gbp npullp centerp oppop ala
        /selection=stepwise slentry=.15 slstay=.15 details include=1;
        run;
        proc reg data= honors.project1;
        model wrcp=aev kp bbp nss nnfp bppa ldp fbp gbp npullp centerp oppop ala
        /selection=stepwise slentry=.10 slstay=.15 details include=1;
        run;
        proc reg data=honors.project1;
        model wrcp=aev kp bbp nss nnfp bppa ldp fbp gbp npullp centerp oppop ala
        /selection=stepwise slentry=.05 slstay=.15 details include=1;
        run;
        proc reg data=honors.project1;
        model wrcp=aev kp bbp nss nnfp ldp fbp gbp npullp centerp oppop ala
        /selection=stepwise slentry=.05 slstay=.05 details ;
        run;

/*      Backward Selection */
proc reg data= honors.project1;
        model wrcp=aev kp bbp nss nnfp bppa ldp fbp gbp npullp centerp oppop ala
```

```
        /selection=backward slentry=.15 slstay=.15 details include=1;
        run;
        proc reg data= honors.project1;
        model wrcp=aev kp bbp nss nnfp bppa ldp fbp gbp npullp centerp oppop ala
        /selection=backward slentry=.10 slstay=.10 details ;
        run;

/* From Xuan */
proc reg data=honors.project1;
        model wrcp=aev ala bbp kp nss nnfp bppa ldp fbp gbp npullp centerp oppop
        /selection=stepwise  details include = 3;
        run;

/* FInal Model? */
proc reg data=honors.project1;
        model wrcp= aev ala bbp kp bppa ss nfp ldp/ influence R ;
        run;

proc reg data=stdtest;
        model wrcp= aev ala bbp kp bppa ss nfp ldp/ influence R ;
        run;
proc reg data=honors.project1;
        model wrcp= aev ala bbp kp bppa ss nfp ldp/ influence R ;
        run;

proc reg data=stdtest;
        model wrcp= aev ala bbp kp  ss nfp ldp/ influence R ;
        run;
proc glm data=stdtest;
        model wrcp= aev ala bbp kp bppa ss nfp ;
        run;



/* when fitting models, wrcp is the response */

proc glm data = honors.project1;
        class nss npullp nnfp;
        model wrcp=aev kp bbp nss nnfp bppa ldp
                        fbp gbp npullp centerp oppop ala /ss3;
run;
/* F = 36.44, p <.0001 */

/* variable selection */
```

```
proc glmselect data = honors.project1;
	class nss npullp nnfp;
	model wrcp=aev ala  bbp kp nss nnfp bppa ldp
				fbp gbp npullp centerp oppop  /details= all selection = stepwise;
run;

proc glmselect data = honors.project1;
	class nss npullp nnfp;
	model wrcp=aev ala  bbp kp nss nnfp bppa ldp
				fbp gbp npullp centerp oppop  /include = 3 details= all selection =
stepwise;
run;

/*
final model include
* AEV *   ala          * bbp (three forced to stay)
	kp nss nnfp bppa ldp (selected by default values of sls & sle)
*/proc glm data = honors.project1 PLOTS=(DIAGNOSTICS RESIDUALS);
	class nss npullp nnfp;
	model wrcp= AEV ala          bbp kp nss nnfp ldp /ss3 solution;
	output out=res_infl P=yhat R=resid STUDENT=stu_resid COOKD=cook H=lev
	run;


proc glm data = honors.project1 PLOTS=(DIAGNOSTICS RESIDUALS);
	class nss npullp nnfp;
	model wrcp= AEV ala          bbp kp nss nnfp bppa ldp /ss3 solution;
	output out=res_infl P=yhat R=resid STUDENT=stu_resid COOKD=cook H=lev
				PRESS=press RSTUDENT=rstd DFFITS=difts COVRATIO=covratios;
run;

/* F = 57.18, p <.0001 */
/* nnfp 1 and nnfp 2 might combine */

proc contents data = res_infl varnum; run;
proc print data = res_infl;
	var name yhat--covratios;
run;

/* for example, you want to look at Cook's D */

proc sort data =res_infl out=res_infl_cooksd;
	by descending cook ;
```

```
run;
proc print data = res_infl_cooksd (where=(cook gt 4/334));
        var name cook;
/* looking for outliers and predictions*/
proc sort data= res_infl;
BY STU_RESID;
run;
proc print data=res_infl;
        where Stu_resid gt 1.8;
run;
proc print data=res_infl;
        where Stu_resid lt -1.8;
run;


/* you can then study why "D.J. Lemahieu" is influential */

data withoutdj;
        set res_infl;
        where cook lt 4/334;
run;
proc glm data= withoutdj;
        class nss npullp nnfp;
        model wrcp= AEV ala          bbp kp nss nnfp bppa ldp;
run;
```