Brian Holliday

WAR Data Regression Analysis

Professor Natov

16 October 2019

Abstract

WAR is a baseball statistic that measures a player’s value against a replacement level player. A replacement level player being a player that will give you the production value of a minimum salary baseball player. This is an important statistic because it’s a way of measuring a player’s value against their peers in the number of extra wins that they will give you.

We looked to do a regression analysis of this statistic to see what measures are integral to the WAR statistic. Traditionally players are evaluated on classic baseball statistics such as batting average, home runs, stolen bases, and runs batted in, but we know that these stats do not tell the whole story. For example, those statistics do not tell you how often a player is getting on base, or strikeout rate. We want to get down to what statistic are the best indicators for the WAR statistic.

Regression Analysis a good tool to figure this out because we can account for multiple factors that help us determine WAR score. With a standard deviation score for each factor, it helps us determine how statistically significant each variable is. With the Regression Analysis we are trying to limit the number of outliers, outliers will make our adjusted R squared score worse, so we will try to make it as high as possible.

For the Regression Analysis, I used R and ran it in a CoCalc Jupyter notebook. This way made it easier to add and delete data, while making it look clean. For this analysis, I started out with 25 different factors and was able to narrow it down to 12 by adding and deleting different data points.

Over the course of the project, it became clearer that finding a fit that I was comfortable with was going to be an art and a science. This was the case because there were some statistics that are not great indicators for how good of a batter someone is. This means that were getting high t – value scores, but they not good indicators for successful at bats. Keeping these would have yielded a better fit, meaning that the R – squared score would have been higher, but I could not keep them in good faith.

My idea was a start will all the relevant statistics, which turned out to be 25 and try to get the best fit from there. The first couple that were deleted were obvious bad fits. These were Total Bases, On-Base percentage, Stolen Bases, etc. This did help me get a better fit. I was able to go from a fit of 0.5737 to almost a one to one fit of 0.909. Although that is a great fit, like I mentioned before there were some stats that needed to be removed from the model. I will talk about two examples from ones that I deleted and got a worse fit.

The first example of these statistics is Hit by Pitch. This stat only tells us how many times a player gets unintentionally hit by a pitch. At the major league level getting hit by a pitch is not a common occurrence. The top five players had only between 8 – 10 in this category, and most players had over 600 plate appearances. Players should not expect to be hit by a pitch when they step into a Major League batters’ box. Yet, hit by pitch had a -3.698 t – score, indicating that it is a significant data point.

Another example of one of this Sacrifice Hits. Out of the 26 players with the highest WAR scores, 18 players had a zero in this category. Sacrifice Hits are situational and not great indictors for success per plate appearance. Sacrifice Hits happen when a player attempts to move baserunners into scoring position by bunting. The problem with this is that if we are looking at the top 26 players when it comes to WAR. This means that they probably provide valuable offense to their team. If they are among the best players, why would you not just let them try to get a hit. Also, a player could provide more value by just getting a hit because the player at the plate and the baserunners could get on base. Despite this, Sacrifice Hits had a high t – value score of -4.834.

There were more stats like this that I removed, but I feel that I was able on a model that included stats like Batting Average, Slugging Percentage, On – Base plus Slugging. The thing that I like about these stats is that they are telling you what a player is doing on an at – bat basis. For example, On – Base plus Slugging is taking into account many aspects of great hitting. Slugging percentage take into account