# The Impact of Player Performance Metrics on Team Wins in Major League Baseball (MLB)

## I. Introduction

Baseball is a game driven by statistics. The industry and its culture have embodied this in so many ways. Movies like Moneyball, fantasy baseball leagues that rely on player stats, sports betting on MLB teams and even major contract negotiations are all filled with statistical decision-making. These things, coupled with the increasing desire from coaches to have meaningful data to make strategic decisions throughout games, clearly demonstrate the fact that baseball is a game riddled with statistical wisdom. The purpose for this research is to examine some of these well known statistics and determine their relevance in predicting the number of wins a team will achieve throughout the course of a season as well as how large of an effect they each have individually. The overarching question is: “What statistics matter most in accurately predicting the number of wins a team will achieve?” This research is important because statistics mean so much for the culture of baseball, but outside of that it is an important topic because it tells us how different organizations operate. It will give insight into what aspects of the game they believe to be important as well as how they think economically. As part of the research for this project the model was run for 10 different MLB seasons from 2014-2024 excluding 2020 because of the shortened season due to COVID-19. The reason this is also relevant is because there are trends in most major sports surrounding how strategy adapts over time and this seemed relevant to understand. It also is a relevant time period because in the last 10 years there have

been lucrative contracts signed in the baseball industry. Some examples would include Shohei Ohtani signing for $700 million with the Dodgers in 2023, Yoshinobu Yamamoto signing for $325 million in 2023 also with the Dodgers and Aaron Judge signing for $360 million in 2023 with the Yankees. These contracts signify the deep appreciation for these players that their respective organizations have for them as well as their skills that they bring to the team. They wouldn't receive contracts like these without respectable, top-notch skills inside the game. Another interesting point to note is that all these players play different roles on their teams. Some are offensive and some are defensive. This is relevant to this paper's question because the statistics will look at both offensive and defensive stats for the team. The research draws on previous literature and aims to improve upon some of the work that has been done towards answering this question in the past. While previous literature towards this topic is very respectable, there seems to be a gap in some of the findings that hopefully will be connected throughout this paper. Section II will summarize the relevant literature to this paper. It will focus on works such as Adam Houser's 2005 paper titled "Which Baseball Statistic is the Most Important When Determining Team Success?" as well as Sarah Sult's 2021 paper "Common MLB Statistics: Which Stats Determine a Team's Win Percentage?" Section III will examine the theoretical model and lay the groundwork for how the model will be built. Section IV will discuss the data related to the question. Section V will evaluate the model's results and provide empirical evidence. Finally Section VI will conclude.

## II. Literature Review

The question of what performance metrics matter most in professional baseball has been a long and comprehensively debated topic. Many people have spent countless hours researching and analyzing what makes teams successful, contributing to a plethora of valuable research surrounding what qualities make a successful team, and furthermore what qualities makes a successful player. This research has surely been crucial in driving data-driven decision making, surrounding management and player personnel within professional organizations. While this topic is clearly relevant to many people it isn't completely answered. This literature review aims to highlight some of the strengths in the available research, as well as some of the potential opportunities for improvement. This will be done by critically examining published works such as those of Houser (2005), Sult (2021) and Ballou-Crawford (n.d.). All of these relevant works will be attached below and provide contextual insight into the answer of the question, "What statistics matter most in accurately predicting the number of wins a team will achieve?"

The theory behind this project is simple: certain measured statistics will have a strong correlation with the number of wins a team will achieve over a given season. This theory will be examined using ordinary least squares (OLS) regression which will determine a linear relationship between the chosen variables and the number of wins a team achieved for the given season. This is one of the most commonly accepted practices for research like this. This data was taken from the Official Major League Baseball website as well as Baseball Reference's website. The data contained 80+ different relevant statistics that could provide insight into the question. Once the data has been prepared it will be read into a statistical software package that is able to run

the regression and provide all the necessary information for answering this hypothesis. Multiple different variables and combinations of variables were used throughout this process in the attempt to find the most statistically significant model. The values that the model finds can be used to predict future outcome values but also to infer the relationship between each feature of the model and the dependent variable. For this inference to be valuable the model must be sound and follow the Gauss-Markov assumptions for linear regression.

The purpose of drawing on previous literature for this paper is that it gives insight, as well as opportunities for improvement on the research surrounding this topic. The reason these need improvement is because a lot of the work directly violates the critical assumptions of linear regression. The first paper in question will be Houser's 2005 work, "Which Baseball Statistic is the Most Important When Determining Team Success?" This research is strong and gives a useful model with a strong R-squared value, however, this model inherently holds some flaws. For starters, two of the first features of Houser's model are average (AVG) and on-base percentage (OBP). This is a significant mistake in a model that is being used for interpretability. Average is found by dividing the number of hits a player gets by the number of at bats they have had. This gives a percentage that alludes to how often that player will get a hit. It can be done for a whole team as well by summing the hits they get and dividing it by the sum of at-bats they have. On-base percentage is a combination of hits, walks and hit by pitches that a player or team gets, divided by the total number of at-bats they have. These two features are highly correlated because of the fact that they both have hits in their numerator over at-bats in their denominator. While they are different, they generally are

within 5-10 percentage points of each other; for example, a player with an average of .300 might have an on-base percentage of .360. This high level of collinearity is an issue for a model that is being used for inference purposes and is likely the reason that his model found a negative parameter value for the average feature. There isn't any reasonable logic that could explain why getting more hits would make a team less successful. On-base percentage has higher values and explains the outcome variable a little more due to also including walks and hit by pitch which gives the team runners on base, so that is probably why it received a positive value by the model while average received a negative value to offset it. This error makes Houser's model unreliable for purposes of inferring what metrics matter most. Houser also makes the point throughout his research that ERA wouldn't be a useful statistic because it is subject to human error when being assessed. The point that they make is that some players will be able to get to a ball and might be able to make an error while others wouldn't even get to it so it could be a hit. This could be said for most of the features in this model, for example, if a player hits a weak ball not very far in the infield but is very speedy, they still may have a chance of getting a hit, which would increase their average, whereas some players might not have the speed to do so. This doesn't seem to be a significant argument for not including ERA. The most significant feature in Houser's model ends up being WHIP, which stands for Walks and Hits per Innings Pitched. This feature is directly related to human scoring decisions that are made, just the same as ERA is, because the scorers decide what is a hit and what isn't. The last point of Houser's model that needs further exploration is the point made that what works for one team will surely work for another. While this model is trying to predict what overall works best for each team based on

certain metrics, there is also an element to each team that makes them unique. Each team has its own unique identity that they play with and this is what contributes to the amount of wins or losses they have, some teams steal a lot of bases while some teams hit a lot of home runs, hopefully this claim can be examined further in this research. The next relevant paper to be examined is Sult (2021). Sult developed a significant model that produced an even stronger R-Squared value compared to Houser. While this would indicate Sult's model improved upon Houser's as far as predictability, it struggled with the same downfalls as Houser's due to the features that were selected for the model. Some of the features selected also contained high levels of multicollinearity similar to Houser's such as average squared, as well as slugging squared. Slugging will inherently contain average inside of it because when a player gets an extra base hit they are also contributing towards their average so these two stats will move together. While these errors don't directly violate Gauss-Markov's classical assumptions for linear regression, they undermine the model's reliability for inference purposes.

This research project aims to correct the gaps in the current available research on answering the question, "What statistics matter most in accurately predicting the number of wins a team will achieve?" The methodology to be used is listed above and will be the guiding force throughout this project. Tests will be conducted to prevent multicollinearity and uphold the critical assumptions of linear regression by Gauss-Markov. Through filling these gaps the question will be answered to a higher and more accurate extent furthering the useful knowledge in the world of baseball on what metrics are most important to a team's success. The research that will be drawn on is sound, useful, strong and has laid the foundation for this project.

## III. Modeling

The underlying question for this research is “What statistic matters most in predicting team wins?” This question takes on multiple hypotheses that are necessary to understand whether it has been answered properly. The first hypothesis is that in a standardized model, there will be features that are statistically significant different from zero. In other words, these variables calculated t-statistics, will be larger than the critical t-statistics at the chosen confidence levels for this model. These features will be directly comparable as far as their significance in predicting the number of wins as well as the magnitudinal effect they each have on the number of wins. Failing to reject the null would mean that features are discovered through this research that prove to be significant in predicting the number of wins. The second hypothesis that will be tested is that OPS and WHIP will be the most significant features in predicting the number of wins a team will achieve. This will also be measured by observing these features’ calculated t-statistics and comparing them to the critical t-statistic. Failing to reject the null would mean that OPS and WHIP will be the two most significant features in the standardized model, compared by their t-statistics.

The first hypothesis will be examined by running the OLS regression model on different sets of features. Some of the features that were examined were average, slugging, on-base percentage, on-base plus slugging, walks or hits per innings pitched and also stolen bases. As more variables were introduced to the model the Adjusted R-Squared value consistently decreased even if the R-Squared value was going which strongly indicated that penalties were being incurred due to including irrelevant

variables. It became clear early on that for this model's parameters to be statistically significant different than zero as well as useful for interpretability, it would have to favor the path of omitting relevant variables rather than including irrelevant ones. This was to protect the variance of the model and in turn protect the t-statistics this model would produce. If this model's purpose was to predict the outcome it would be relatively easy to pick features from the list of 83 that would increase the R-Squared value without incurring major penalties but this wasn't the purpose of the model. There are however almost certainly omitted variables in this model that could increase its Adjusted R-Squared value if they were determined.

The second hypothesis will be tested by running a regression that uses just OPS and WHIP in its model. It will also be run as a standardized linear regression model as these two variables are on entirely different scales. This will allow the model's parameter estimates to be significant in determining which feature has a stronger effect on the number of wins a team will achieve in a season. To determine if OPS and WHIP are statistically significant different than zero the feature's t-statistics will need to be evaluated. If these prove to be statistically significantly different than zero and have the highest calculated t-statistics compared to other features then the null will fail to be rejected and it will be accepted that OPS and WHIP are the two most significant features in explaining the number of wins a team will achieve. Throughout the research pairwise correlations were also used, examining the magnitude of the relationship between different features and the number of wins teams achieved by the end of the season. This was another important factor in determining whether this null hypothesis should be accepted.

## IV. Data

Data was not a major issue for this research. Baseball is a naturally statistically driven sport so there is ample data out there on team statistics as well as other new measures that have been coming out that are only going to further the game's skill level and bring in new strategies for how to manage games. This was very nice to have and the only conflicts that needed to be addressed was how to merge different data sets together. This wasn't too strenuous and the data ended up being in a very useful format. However, if there was something that could be adjusted in this research it would have been having access to more data. One example that would have been very useful would be data on stolen base success percentage (Houser, 2005). While there isn't direct evidence that this variable would be statistically significant, in theory it seems that there are teams that succeed as much as they do because of their ability to steal bases. Take the 2024 Brewers for example. They had a slightly better average than the league average, however they had one of their best seasons in a long time. Outside of that feature there are likely others that would be very interesting to explore under this topic.

The data that was used included 30 observations, one for every team in the league as well as 83 different features that could be chosen from. These features included hitting, fielding and pitching statistics so that the model had the opportunity to be tested with a comprehensive set of features. Asdf

## V. Evidence

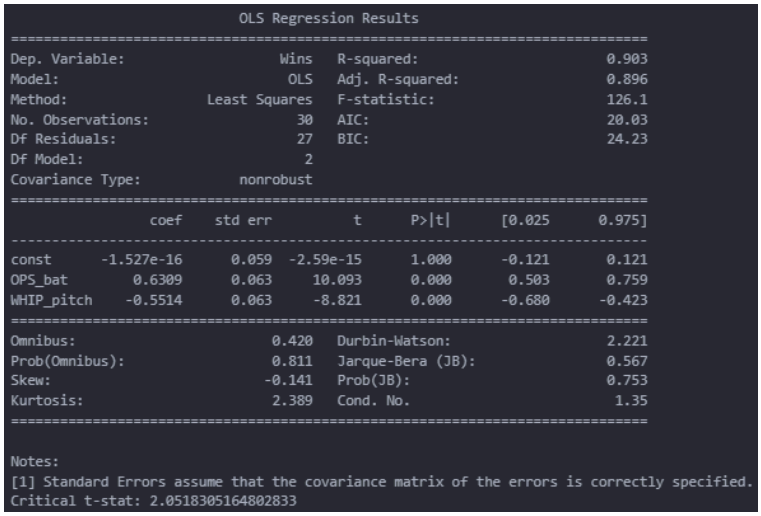
The model concluded that there were variables that were statistically significant in predicting the number of wins a team would achieve throughout the course of a regular season. The model also found OPS and WHIP to both be statistically significant different than zero and both had high calculated t-stats indicating that they were very significant in predicting this dependent variable. They also had very low p-values which further assured that these variables were significant in this model. The model found OPS to have a stronger effect on the number of wins a team would achieve compared to WHIP but it was a close comparison. As the model was run for each season going back to 2014 there was also a noticeable trend moving away from this and instead finding that WHIP had a stronger effect on the number of wins a team would achieve. This was surprising and wasn't an anticipated result going into the research. What this could signal is that there has been a shift in the way players play the game and the playstyles they play with. It could be interpreted that pitching used to have a much stronger impact on the outcome of games and that now there has been a shift in which offense is having a stronger effect. This could be because players are focusing on hitting for power more nowadays or could also not have much significance at all and just be due to different seasons having randomly different statistics but it was an interesting result.

**Table 1: Tests for Heteroskedasticity, Autocorrelation and Multicollinearity**

```
Bresuch-Pagan P-value: 0.15739918495147448
Durbin-Watson Statistic: 2.220753048186621
     Variable       VIF
0       const  1.000000
1     OPS_bat  1.091209
2  WHIP_pitch  1.091209
```

This model went through lots of testing to prove that it doesn't violate the classical assumptions for linear regression. First to test for heteroskedasticity and autocorrelation, the Bruesch-Pagan test as well as the Durbin-Watson test were performed. What these tests both did was ensure that the model wasn't incurring any heteroskedasticity or autocorrelation which would mess with the standard deviation for the model and in turn, make the features less reliable. The model achieved a BP test score of 0.157 (rounded) as well as a DW test score of 2.221 (rounded). This indicated that the model didn't show signs for serious heteroskedasticity or autocorrelation as the BP test was evaluated at a 5% level of significance and it passes that. The DW stat, although wasn't exactly 2, doesn't give evidence of serious autocorrelation and therefore also wasn't a concern for this model. Finally the variance inflated factors, or VIFs, were calculated and this test also demonstrated reassurance in the model similar to the BP test and DW test. Both variables had VIFs around 1.09. What this test showed was that both of the features didn't demonstrate signs of multicollinearity meaning this model was a reliable model for inference purposes, because the model will be able to determine where the variance in the dependent variable is coming from all else held equal as one of the features changes. This was a very encouraging sign for the purposes of this research.

**Table 2: Regression Results: Predicting Number of Wins**



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Wins   R-squared:                       0.903
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     126.1
No. Observations:                  30   AIC:                             20.03
Df Residuals:                      27   BIC:                             24.23
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1.527e-16      0.059  -2.59e-15      1.000      -0.121       0.121
OPS_bat        0.6309      0.063     10.093      0.000       0.503       0.759
WHIP_pitch    -0.5514      0.063     -8.821      0.000      -0.680      -0.423
==============================================================================
Omnibus:                        0.420   Durbin-Watson:                   2.221
Prob(Omnibus):                  0.811   Jarque-Bera (JB):                0.567
Skew:                          -0.141   Prob(JB):                        0.753
Kurtosis:                       2.389   Cond. No.                         1.35
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Critical t-stat: 2.0518305164802833
```

The final model only includes two features. They are OPS and WHIP. They both proved to be statistically significant in literature review as well as in models that were run for this and because of this they stood out as the most significant features. They also both contained very high pairwise correlation with the number of wins a team will get. At the 0.05 level of significance both features proved to be statistically significant different than zero and also had p-values of 0 meaning they are very significant. The F-statistic is high indicating the model is significant as a whole and useful for interpretation. Finally the model has a very high R-Squared value as well as a high Adjusted R-Squared value which would signal that this model can explain a lot of the variation in the number of wins a team gets.