Road to the Fall Classic: an Investigation in which Regular Season Statistics are the Greatest Indicators of World Series Success

Devin Sciarrino

Abstract

The pursuit of World Series glory involves a range of strategies, talent evaluations, and statistical analysis in Major League Baseball. This study explores whether regular season statistics can reliably predict a team's success in the World Series. By analyzing a comprehensive dataset of team statistics spanning multiple MLB seasons, we investigate the correlation between various regular season performance metrics and a team's ability to win the World Series. Through thorough statistical analysis and predictive modeling, we discover compelling evidence that specific regular season statistics can forecast a team's journey to the Fall Classic. My research illuminates the intricate relationship between regular and post-season performance dynamics, providing valuable insights for baseball enthusiasts, analysts, and team management. As the competition for World Series victory continues to captivate fans and professionals, recognizing the importance of regular season statistics is crucial for unraveling the path to baseball's most prestigious trophy.

Introduction

Baseball, often referred to as Americas pastime, has grown to be a sport appreciated by audiences around the world. With over 150 years of rich history, intricate strategies, and stories that sound almost too good to be true, Baseball in steeped in tradition. But a new tradition has emerged in the past couple years, a tradition of data analytics. Behind every triumph on the diamond lies a complex interlay of skill, athleticism, strategy, and statistical analysis, making baseball a fertile ground for exploring the relationship between performance metrics and championship success. At the heart of this sport lies the quest for the championship known more commonly as “The World Series”. Every year 30 teams from all over the country compete in a grueling 162 game season to reach the post season, and then fight for a championship.

Starting in the early 2000’s, the rise in popularity of advanced statistical analysis and sabermetrics have revolutionized past understandings of what baseball is. With the ever-expanding array of metrics, analyst have sought to uncover which statistics most accurately equate to success and failure. They lead to answer questions about what the optimal solution in on field situations are, where and when players should be placed on the field and in the batting lineup. But among all the data and analytics, one staggering question remains: which regular season statistics serve as the most reliable predictor of World Series success?

This paper seeks to delve into this question by doing a comprehensive analysis of regular season baseball statistics and their predictive power in determining World Series outcomes. By inspecting historical team statistics starting in the year 2000, we aim to identify key performance metrics that exhibit a strong correlation with postseason success. We will investigate a wide range of statistics, encompassing offensive, defensive, and pitching statistics and compare them the team’s world series success in that year.

By examining the relative importance of each metric and their collective impact on championship success, we aim to provide valuable insights into the dynamics of baseball excellence and inform strategic decision-making at both the team and organizational levels. Ultimately, this research represents not only a quest for knowledge but also a celebration of the timeless charm of baseball and its everlasting capacity to inspire, captivate, and unite fans across generations. Through a deeper understanding of the statistical underpinnings of championship success, we hope to enrich the discourse surrounding America's favorite pastime and contribute to the ongoing evolution of baseball analytics in the modern era.

Literature Review

The interest of advanced statistics in the baseball community really caught its footing After the Oakland A’s 2002 season. Money Ball, a book based on this team, is the true story of how A’s general manager Billy Beane started to look at baseball in a nontraditional way. After a successful 2001 season, the Athletics lost their star players to large market teams like the Redsox and Yankee’s since they could not afford to pay them. To remain competitive, Beane and the rest of the front office looked to analytical ideas of Bill James, namely the statistic of On Base Percentage for answers. Using this new strategy, the Athletics put a successful season together which included a record 20 game winning streak, and a playoff berth. Although by most measures the A’s 2002 season went beyond expectations, Beane was still disappointed in the fact that even though they got a lot of wins, they did not win “The last one”, meaning the World Series. This is seen as the main criticism for the money ball theory, because although wins get you to the postseason, winning the world series takes more.

Houser’s (2005) investigates the theory from money ball. Along with on base percentage he also investigated batting average (BA), slugging percentage (SLG), steels, walk/hits per innings pitched (WHIP), Strikeouts per nine innings (K’s/9), and fielding percentage. In his findings he found that on base percentage is a good predictor of wins, along with WHIP.

A 2013 Bleacher Report article by Jason Catania investigated whether pitching or hitting had a greater impact in winning championships. His research investigated runs scored, batting average, OBP, OPS, and weighted on base average. For pitching he investigated runs allowed, earned run average, earned run average plus, fielding independent pitching, WHIP, and strikeout percentage. His findings where that neither pitching nor hitting where conclusively better at predicting the champion. In some years like 2009 Yankees and 2011 Cardinal, the team that one did so primarily with their abilities at the plate, while other teams like 2005 White Sox and 2010 Giants won primarily through pitching excellence.

As was mentioned earlier, the problem with the money ball theory is that it fails to convert its regular season success to the post season. In Houser (2005) he finds that WHIP is also a good predictor of regular season success, along with reaffirming the theories in money ball about on base percentage. An issue with this paper is that it only analyzes the 2000 through 2004 seasons and fails to investigate championship success. Catania’s article does delve into regular season statistics as predictor but only as an offensive versus pitching statistics. It also fails to investigate the economic impacts of winning a world series. Finally, all previous investigations fail to investigate more advanced “Plus” statistics which compare the team to the rest of the league. All aforementioned issues will be investigated in this paper.

Objectives

This paper aims to answer the question of what statistic is the best predictor of World Series success, to find if OBP can be a good predictor of championship success, and to create a model capable of forecasting the World Series Champion. Using OLS logistic regressions, this study will test these hypotheses: Teams with better raw statistics will have greater championship success. Teams with better calculated statistics with have greater championship success. The Statistics the Houser 2005 found to be the best predictor of wins will also predict world series success. Teams with better regular season statistics compared to the rest of league will have greater championship success.

Proposed Models

This paper will investigate four different models. The various models are as follows:

**Table 1**

**Empirical model 1**

**Dependent Variable**

WSWin(d) Dummy variable with 1 being a world series win, 0 being a non-win

**Explanatory Variables**

R (+) Runs

x1B (+) Singles

x2B (+) Doubles

x3B (+) Triples

HR (+) Homeruns

BB (+) Walks

SO (-) Strikeouts

SB (+) Stolen Basses

CS (-) Caught Stealing

HBP (+) Hit by Pitch

SF (+) Sacrifice Fly

RA (+) Runs Allotted to opposing teams

CG (+) Complete Games

SHO (+) Shutout, Games where opposing team failed to score

SV (+) Saves, Award to relief pitcher who completes win when team is winning by 3 or less runs

HA (-) Hits Allotted to opposing teams

HRA (-) Homeruns Allotted to opposing teams

BBA (-) Walks Allotted to opposing teams

K (+) Strikeouts Allotted to opposing team

E (-) Errors, failure to convert an out on play an out should be made, i.e. dropped ball

DP (+) Double Plays, Making two outs on same continuous play

Empirical Model 1 explores the primary regular season raw statistic that exerts the most significant influence on World Series success. Each explanatory variable in the model represents the cumulative performance of a team throughout the regular season. Omission of variables such as Hits (H) and Earned Runs (ER) was deemed necessary due to their inherent mechanical relationship with other variables within the model.

**Table 2**

**Empirical Model 2**

**Dependent Variable**

WSWin(d) Dummy variable with 1 being a world series win, 0 being a non-win

**Explanatory Variable**

BA (+) Batting Average, Function of total hits per total at bats

OBP (+) On Base Percentage, Function of total times reached base safely per plate appearances

SLG (+) Slugging Percentage, Batting Average weighted by total bases per hit

ERA (-) Earned Run Average, Mean runs per nine innings

KPN (+) Strikeout’s per Nine, mean strikeouts per nine innings

WHIP (-) Walks/Hits Innings Pitched, Average baserunners allowed per inning

FP (+) Fielding Percentage

Empirical model 2 investigates which regular season calculated statistic has the greatest impact on world series success. Each one of the explanatory variables is a metric that is calculated based on the raw statistics in model 1.

**Table 3**

**Empirical Model 3**

**Dependent Variable**

WSWin(d) Dummy variable with 1 being a world series win, 0 being a non-win

**Explanatory Variable**

BA (-)

OBP (+)

SLG (+)

x3B (+)

SB (+)

WHIP (-)

KPN (-)

FP (+)

Empirical model 3 investigates the findings of Houser (2005) and if they remain the same when predicting world series success. The hypothesized relationship between the variables is based on the results Houser’s model of predicting wins. In his model he was surprised to find that Batting Average had a negative influence on the amount of wins a team achieved. I have added the variable for x3B since it had the greatest correlation in Model 1 and I need to a replacement for “Payroll” in Houser’s model.

**Table 4**

**Empirical Model 4**

**Dependent Variable**

WSWin(d) Dummy variable with 1 being a world series win, 0 being a non-win

**Explanatory Variable**

OPSP (+) On Base Plus Slugging Plus

WHIPP (+) Walks/Hits Inning Pitched Plus

FPP (+) Fielding Percentage Plus

Empirical model 4 investigates the comparison statistics. All three explanatory variables are based on calculated statistic from empirical model 2. The comparison aspect comes from that each variable comes from the basic formula of with some variation for each variable. OPS+ is a metric commonly used around the baseball community, while WHIP+ and FP+ are two metrics I created for this paper. More information on how these metrics where calculated can be found in the data cleaning steps.

In Money Ball, Beane and the rest of the Athletics front office hypothesized that increasing On Base Percentage is the fastest way to increasing the number of wins in a season. Houser (2005) investigated this hypothesis and concluded the OBP is an important predictor of wins, as well as Slugging Percentage (SLG). The problem with each of these metrics OBP overlooks the added benefit of getting extra base hits and how getting more bases per at bat and SLG does not consider walks and hit by pitches. Thankfully there is a statistic that combines the best of OBP and SLG. This calculated statistic is called On Base Plus Slugging (OPS). This metric does exactly what it sounds like and adds together OBP and SLG. This statistic paints a more definitive picture of a team’s offensive success.

Houser (2005) also concluded that Walks/Hits per Innings Pitched (WHIP) is a significant predictor of wins. The way WHIP works is by taking the total amount of baserunners a team or player allows and divides it by the number of innings played.

Data

To successfully test my hypotheses, data will be compiled for every team in Major Leagues Baseball from the year 2000 to 2015. The data is the Baseball Databank from the website *Kaggle.com.*  The specific table I used for the data is the Team sheet. This sheet had all the raw data necessary for me to calculated various other metrics I would need.

The type of cleaning performed was getting rid of the extra, unnecessary data like ballparks ID’s and other statistics that would not be used in or to calculate other metrics for the models. The largest cut of data was in deleting all the data from 1871 through 1999. This data was deleted for multiple reasons including it being incomplete and the game being very different than it is today. An example of this is the fact that the world series did not exist until the 1884 season and did not have the playoff format which includes the wildcard series until 1995.The reason for only using data after the 1999 season is because it and was completely detailed. Before this, some gaps in the data where due to the statistic not being tracked until a certain point. For example, the statistics Hit by Pitch and Sac Fly where not tracked until the year 2000, so metrics like On Base Percentage could not be calculated.

The other kind of cleaning performed was the calculation of more advance metrics. Using the raw on field data, metrics like Batting Average and Slugging percentage could be calculated. And from these metrics even more advanced like OPS+ and WHIP+.

*Data cleaning steps*

1. Upload Baseball Databank Team.csv into excel
2. Add a new works sheet
   1. Name it “2000”
3. Copy Rows 2327 through 2806 and paste in cell A2 on sheet “2000”
4. Copy column titles from “Teams” sheet and paste in cell A1 on sheet “2000”
5. In sheet 2000
   1. Delete columns teamID, WCWin, name, Park, Attendance, teamIDBR, teamIDLahm, and TeamIDretro.
   2. Rename columns 2B and 3B to x2B andx3b
   3. Add Column between “H” and “2B”
      1. Name column “1B”
   4. Add four columns between “SF” and “RA”
      1. Name columns “BA”, “OBP”, “SLG”, and “OPS+”
6. In column “x1B” (Column J)
   1. Create singles calculations
      1. In cell R: I2-(K2+L2+M2)
   2. Copy to rest of column
7. In column “BA” (Column T)
   1. Create Batting Average calculations
      1. In cell T2: I2/H2
   2. Copy to rest of Column
   3. Convert data to 3 decimal places
8. In column “OBP” (Column U)
   1. Create On Base Percentage calculation
      1. In Cell U2: =(I2+N2+R2)/(H2+N2+R2+SS2)
   2. Copy to rest of column
   3. Convert data to 3 decimal places
9. In column “SLG” (Column V)
   1. Create Slugging Percentage calculation
      1. In cell V2: =(J2+(K2\*2)+(L2\*3)+(M2\*4))/H2
   2. Copy to rest of column
   3. Convert data to 3 decimal places
10. In column “OPS” (column W)
    1. Create On Base Plus Slugging calculation
       1. In cell W2: =U2+V2
    2. Copy to rest of column
    3. Convert data to 3 decimal places if necessary
11. Create Innings Column (Column AE)
    1. Insert column between IPouts and HA
    2. Label column “Innings”
    3. Create innings calculation
       1. In cell AE2: =AD2/3
    4. Copy to rest of column
    5. Convert data to 2 decimal places
12. Create KPN column
    1. Inset column between columns “SOA” and “E”(column AJ)
    2. Label column “KPN”
    3. Create Strikeouts per 9 calculations
       1. In cell AJ1: =9\*(AI2/AE2)
    4. Copy to rest of column
    5. Convert data to 2 decimal places
13. Create BBPN column
    1. Insert column between columns “KPN” and “E”(Column AK)
    2. Label column “BBPN”
    3. Create walks per 9 calculations
       1. In Cell AK1: =9\*(AH2/AE2)
    4. Copy to rest of column
    5. Convert data to 2 decimal places
14. Create WHIP column
    1. Insert column between columns “BBPN” and “E” (Column AL)
    2. Label column “WHIP”
    3. Create Walks Hits Innings Pitched calculations
       1. In Cell AL1: =(AH2+AF2)/AE2
    4. Copy to rest of column
    5. Convert data to 2 decimal places
15. Create HRPN column
    1. Insert column between columns “WHIP” and “E” (Column AM)
    2. Label column “HRPN”
    3. Create Homeruns per 9 calculations
       1. In Cell AU1: =9\*(AG2/AE2)
    4. Copy to rest of column
    5. Convert data to 2 decimal places
16. Create EPN column
    1. Insert column After column “FP” (Column AQ)
    2. Label column “EPN”
    3. Create errors per 9 calculations
       1. In Cell AQ1: =9\*(AN2/AE2)
    4. Copy to rest of column
    5. Convert data to 2 decimal places
17. Create DPPN column
    1. Insert column After columns “EPN” (Column AR)
    2. Label column “DPPN”
    3. Create double plays per 9 calculations
       1. In Cell AR1: =9\*(AO2/AE2)
    4. Copy to rest of column
    5. Convert data to 2 decimal places
18. Create 2 more sheets and name them “2000 raw” and “2000 calc
19. Import only raw statistics to “2000 raw”, raw statistics are stats that have not found using a calculation
20. Import all other statistics to “2000 Calc” sheet
21. Create WSWin dummy variable in sheet 2000 raw
    1. Insert column between “WSWin” and “BA” (G)
    2. Name “WSWin(d))
    3. In cell G2: =IF(F2="Y",1,0)
    4. Copy to rest of column
    5. Repeat process in sheet 2000 Calc
22. Create “plus” sheet
    1. Copy over data from “2000 Calc”
    2. Delete all columns except for: WSWin(d), OBP, SLG, WHIP, and FP
    3. Insert 3 columns between SLG and WHIP, 2 between WHIP and FP, and two after FP
    4. In order, name new columns lgOBP, lgSLG, OPS+, lgWHIP, WHIP+, lgFP, and FP+
    5. In cell G2 write lgOBP formula: =IF(A2=2000,AVERAGE(E$2:E$31),IF(A2=2001,AVERAGE(E$32:E$61),IF(A2=2002,AVERAGE(E$62:E$91),IF(A2=2003,AVERAGE(E$92:E$121),IF(A2=2004,AVERAGE(E$122:E$151),IF(A2=2005,AVERAGE(E$152:E$181),IF(A2=2006,AVERAGE(E$182:E$211),IF(A2=2007,AVERAGE(E$212:E$241),IF(A2=2008,AVERAGE(E$242:E$271),IF(A2=2009,AVERAGE(E$272:E$301),IF(A2=2010,AVERAGE(E$302:E$331),IF(A2=2011,AVERAGE(E$332:E$361),IF(A2=2012,AVERAGE(E$362:E$391),IF(A2=2013,AVERAGE(E$392:E$421),IF(A2=2014,AVERAGE(E$422:E$451),IF(A2=2015,AVERAGE(E$452:E$481),""))))))))))))))))
    6. Copy to rest of column
    7. Use the same formula to calculate the rest of lg\_\_\_ statistics converting the letter of the column using chat gpt (switch E to F, etc.…)
    8. In cell I2 write formula: =100\*((E2/G2)+(F2/H2)-1) and copy to rest of column
    9. In cell L2 write formula: =100\*(J2/K2) and copy to rest of column
    10. In cell O2 write formula: =100\*((M2/N2)-1) and copy to rest of column

Empirical Methodology

For all four regression tests, the explanatory variable, coefficients, standard error, z-value, and significance level are presented in the following tables.

**Table 5**

**Regression Results: Model 1**

**Explanatory Variable â Std. Error Z. Value Sig**

(Intercept) -3.071623 15.613689 -0.197 0.8440

R 0.010398 0.013154 0.790 0.4292

x1B 0.004541 0.009664 0.470 0.6385

x2B 0.020900 0.014813 1.411 0.1583

x3B 0.071880 0.04024 1.786 0.0741 .

HR 0.006865 0.021309 0.322 0.7473

BB -0.001695 0.006064 -0.279 0.7799

SO -0.003025 0.003294 -0.918 0.3586

SB -0.007930 0.012060 -0.658 0.5108

CS 0.027193 0.032852 0.828 0.4078

HBP 0.011215 0.023665 0.474 0.6356

SF -0.062294 0.042395 -1.469 0.1417

RA -0.006448 0.013668 -0.472 0.6371

CG 0.068882 0.098912 0.696 0.4862

SHO -0.089053 0.101111 -0.881 0.3785

SV 0.060150 0.053840 1.117 0.2639

HA -0.009321 0.008979 -1.038 0.2993

HRA -0.003524 0.020051 -0.176 0.8605

BBA -0.001673 0.007130 -0.235 0.8145

K 0.002051 0.003603 0.569 0.5692

E -0.018658 0.022838 -0.817 0.4140

DP 0.001454 0.020682 0.070 0.9440

N = 480 Deviance = 103.9006 MSE = 0.02801215

Table 5 shows the regression results for Model 1. Model 1 investigated which regular-season raw statistics show the greatest correlation with post-season success. According to the regression the number of triples hit and the amount of sacrifice flyballs hit had the greatest significance to winning the World Series. One reason for this could be that 19% of the World Series’ won from 2000 to 2015 were won by the San Francisco Giants. The Giants are regularly in the top third of triples hit in a season. This is due to the configuration of their stadium, Oracle Park. The stadium has a section of right-center field that is colloquially named “Triples Alley”. If a ball is hit in this area, it is almost a guaranteed triple for even players of average speed. One interesting thing about sacrifice fly’s is even though it is the second most significant variable, it has a negative relationship with predicted world series success. This means for every successful sac fly hit, your chances of winning the World Series decrease by 6.23%. This relationship may be because that even though it may score a run or advance a baserunner, the additional out it caused has a greater effect. One more variable relationship that caught me by surprise was the effect of a walk. Walks ended up having a negative relationship with championship success. Theoretically, the added baserunner of a walk should increase the chances of winning a World Series, but the regression points to another conclusion.

**Table 6**

**Regression Results: Model 2**

**Explanatory Variable â Std. Error Z. Value Sig**

(Intercept) -134.1037 120.4216 -1.114 0.265

BA 66.4524 44.9416 1.479 0.139

OBP -0.6783 37.1342 -0.018 0.985

SLG 22.5364 17.8507 1.262 0.207

ERA -1.8883 1.4658 -1.288 0.198

KPN 0.0324 0.4164 0.078 0.938

WHIP -3.1439 7.8741 -0.399 0.690

FP 117.2293 121.2764 0.967 0.334

N = 480 Deviance = 113.1074 MSE = 0.03018791

Model 2 investigates the impact of calculated statistics on the probability of winning a World Series. According to the regression, none of the variables were statistically significant at an alpha level of 10 percent. The three most significant variables were Batting Average, ERA, and slugging percentage. One surprising item to come from this regression is the numbers associated with On Base Percentage. Not only does it have a negative correlation with winning the World Series, but it is also the least significant variable in the model. This shows that while OBP may be a good predictor of wins, it is a terrible predictor of World Series success according to the model. The reason for this could be because OBP is also a measure of a hitter’s competition. Batting Average and Slugging Percentage are more reflection of a hitter’s skill at the plate, while OBP also considers a pitcher’s ability, or lack thereof, to throw strikes and not hit the batters.

**Table 7**

**Regression Results: Model 3**

**Explanatory Variable â Std. Error Z. Value Sig**

(Intercept) -1.225e+02 1.225e+02 -1.000 0.31739

BA 5.098e+01 4.543e+01 1.122 0.26178

OBP 4.937e+00 3.704e+01 0.133 0.89396

SLG 2.046e+01 1.793e+01 1.141 0.25394

x3B 4.491e-02 3.219e-02 1.395 0.16296

SB -2.091e-03 9.108e-03 -0.230 0.81843

KPN 5.586e-02 4.282e-01 0.130 0.89620

WHIP -1.289e+01 4.369e+00 -2.950 0.00318 \*\*

FP 1.126e+02 1.236e+02 0.911 0.36239

N=480 Deviance = 112.8139 MSE = 0.03033286

Model 3 test’s Houser (2005) model of predicting wins with the addition of the statistic Triples instead of payroll. Instead of predicting wins, the dependent variable is World Series success. Houser’s model was very good at predicting the number of wins in a season a team gets with an R-squared value of .849. Differences between the two models can be found with the variable Batting Average and Base Percentage. In Houser’s model batting average had a negative correlation and was not as statistically significant while OBP had a positive correlation and a significance level of 0. The results of Model 3 are the opposite of this with batting average and OBP keeping similar figures as they did in Model 2. One thing that was similar between Houser (2005) and model 3 was the significance of WHIP.

**Table 8**

**Regression Results: Model 4**

**Explanatory Variable â Std. Error Z. Value Sig**

(Intercept) 2.13992 6.43612 0.332 0.73952

OPSP 0.10398 0.03300 3.150 0.00163 \*\*

WHIPP -0.16809 0.06079 -2.765 0.00569 \*\*

FPP 1.15762 1.17008 0.989 0.32249

N=480 Deviance = 116.5193 MSE = 0.03019563

Model 4 investigates comparative statistics. All three explanatory variables account for not only what each team does, but also for what the rest of the league does in its given year. In the regression, OPS+ and WHIP+ were both statistically significant, with the offensive statistics having the most significance. This was to be expected as most opinions surrounding these theories agree with this

**Table 9**

**Prediction Results: 2023 Texas Rangers**

**Model Prediction %Chance of Championship**

Model 1 0.430306 43.03%

Model 2 0.08280767 8.28%

Model 3 0.1261997 12.62%

Model 4 0.1359617 13.60%

Table nine is a comparison between all four models. The prediction was based on the season statistics of the 2023 World Series champions, the Texas Rangers. The data for the 2023 Rangers’ season was sourced from baseballreference.com Model 1 gave them the highest chance of success with 43.03%.

**Table 10**

**Prediction Results: Measure of Model Fit**

**Model Deviation MSE**

Model 1 103.9006 0.02801215

Model 2 113.1074 0.03018791

Model 3 112.8139 0.03033286

Model 4 116.5193 0.03019563

Of the four models, model 1 was the greatest predictor of World Series success. Model 3 has the best measure of fit figures with deviation and MSE being 103.9 and .028 respectively. It was also the best model for predicting the 2023 World Series champion giving the Rangers a 43% chance of winning by using only in-season statistics. The reason for model 1’s accuracy likely has to do with the number of variables used in the model. Model 1 utilized 20 different raw variables. The other three models use 7, 8, and 3 each.

Even though model 1 may be the best predictor of world series success, it is not as good at showing which statistic is the most important. The best model for this would be model 4. Model 4 had two explanatory variables with significant levels below the alpha level of 10%. The two metrics being the best individual predictors were OPS+ and WHIP+.

Summary

The objectives of this paper were to find the best predictor of World Series success. Utilizing team data from the seasons 2000 through 2015, we investigated various offensive, pitching, and defensive statistics and measured their importance to championship success through logistic regression. We also investigated whether the theories and findings surrounding On Base Percentage and its relationship with winning still held which World Series Success.

Regarding which regular season individual statistics have the greatest significance on championship win probability, the calculated statistics OPS+ and WHIP+ were found to be the most important. Both statistics encompass offense and pitching respectively and give the added benefit of also being comparison ratios to the rest of the league. The theory of Base Percentage being a good predictor of wins may be true, however, it fails to be significant went analyzing World Series success. According to the regressions in this paper, On Base Percentage has almost no significance. Because of this Major League Baseball Teams should stop trying to pad this statistic and instead focus more on maximizing Base Plus Slugging and minimizing Walks/Hits per Innings Pitched.

Potential challenges that were faced in this study are the limitations on data, specifically time-wise. With data being traceable back to 1874, it is a shame that so much is lost because it failed to be recorded. Investigating this topic further would be very interesting. With more historical data, one could investigate how which statistics had the greatest importance throughout the different eras of baseball.  More investigation could also be done to find an even more accurate model to predict the World Series Champion. This could be done with more trial and error with different statistics. I predict that model 4 is a good start to finding an optimal model and more explanatory variables could further increase its accuracy.

References

“2023 Major League Baseball Team Statistics.” *Baseball*, www.baseball-reference.com/leagues/majors/2023.shtml. Accessed 1 May 2024.

“2023 Texas Rangers Statistics.” *Baseball*, www.baseball-reference.com/teams/TEX/2023.shtml#all\_standard\_fielding. Accessed 1 May 2024.

Catania, Jason. “Statistical Look at Whether Pitching or Hitting Really Wins Championships.” *Bleacher Report*, Bleacher Report, 3 Oct. 2017, bleacherreport.com/articles/1786978-statistical-look-at-whether-pitching-or-hitting-really-wins-championships.

Houser '05, Adam (2005) "Which Baseball Statistic Is the Most Important When Determining Team Success?," The Park Place Economist: Vol. 13 Available at: <https://digitalcommons.iwu.edu/parkplace/vol13/iss1/12>

Luca, De Michael, et al. *Moneyball*. Sony Pictures Entertainment, 2011.

Majors, Peter. “How Truly Anyone Can Calculate League Normalized Statistics: OPS+ and WRC+.” *Fordham SAS*, Fordham SAS, 1 Feb. 2021, fordhamsas.wixsite.com/website/post/how-anyone-can-calculate-league-normalized-statistics-ops-and-wrc#:~:text=The%20formula%20for%20OPS%2B%20is,lgSLG)%20%2D%201)%20%2F%20BPF.

Sports, Open Source. “Baseball Databank.” *Kaggle*, 17 Nov. 2019, www.kaggle.com/datasets/open-source-sports/baseball-databank?resource=download&select=Teams.csv.

“Statcast Park Factors.” *Baseballsavant.Com*, baseballsavant.mlb.com/leaderboard/statcast-park-factors?type=year&year=2003&batSide=&stat=index\_wOBA&condition=All&rolling=. Accessed 1 May 2024.