Using Logistic Regression to Predict Baseball Playoff Probabilities

Matthew Wankiewicz

26/11/2020

Abstract

The book moneyball highlighted a statisitcs revolution that was occuring in baseball in the early 2000's. In this report, logistic regression to determine whether the advanced stats that stemmed from this revolution are useful for predicting a team's success. After running the model, we find that it successfully determines if a team will make the playoffs (or not) 88 percent of the time. This is a significant finding as it shows that these advanced metrics are extremely useful and teams should fully shift their focus on these metrics.

Keywords: Moneyball, Baseball, Sabermetrics, Logistic Regression

Code supporting this analysis can be found at: https://github.com/matthewwankiewicz/moneyball

1 Introduction

In 2003, Michael Lewis wrote a book called Moneyball [1]. Moneyball is the story of the Oakland Athletics who, after losing players to the richer teams in the league, decided to focus on using the misfits of baseball to try to win a World Series. These "misfits" were players who never received support from teams because their traditional stats didn't look good but they excelled in the stats that mattered.

In order to test if Moneyball's basis was correct, I will be using logistic regression to predict whether or not a team will make the playoffs. The seasons that I will analyze go all the way back to 1920, with some missing years in the 70s because of missing values in the contract data. In order to predict if a team makes the playoffs or not, the variables I plan to use are the team's payroll, the team's batting average (BA), the team's earned run average (ERA), their weighted on base average (wOBA) and lastly their fielding independent pitching (FIP). These stats may sound scary but they are fairly simple and will be exaplined in the data section below.

The data collected is from the `Lahman` package in R (Friendly et al. [2020]). The `Lahman` package contains yearly player statistics going all the way back to the late 1800s. For this project, I will only use the data from 1920 and onward because those years have the most observations for stats like payroll data. In the `Lahman` package there are many different datasets present, for this project I will be using the `Teams` dataset. In order to collect the payroll data, I created a scraper that collected yearly Team stats from Baseball Reference (LLC). The scraper takes the data from each year and saved it into a larger dataset [2]. The payroll data will be crucial for my report because the basis of Moneyball is fielding the best team by using the least amount of money.

This data analysis was conducted using R (R Core Team [2019]), and in particular the packages Tidyverse (Wickham et al. [2019]) and was compiled using R markdown (Xie et al. [2018]). I have 4 sections not including the introduction. The first section covers the data I used for my report, including a few plots to show how the distributions and relationships between various stats. The next section is about the model I used, this will include a in-depth breakdown of logistic regression. Next, I will display the results of my model,

[1] Moneyball was actually the main reason I decided to get into statistics. Applying baseball with numbers seemed like an absolute win to me

[2] This scraper can be found in the scripts folder of my GitHub repo.

this will include the coefficients of the model and an interpretation of the model, along with an application of the model to the results of the 2020 season to see how accurate its predictions are. Lastly, there is a section discussing what we learn from this model and some next steps. Was Moneyball right? Does money impact a team's ability to make the playoffs? Keep reading to find out!

2 Data

The data collected was from the `Lahman` package in R (Friendly et al. [2020]) and Baseball-Reference (LLC), Fangraphs (Fangraphs) was also used to help calculate one of the statistics.

The `Lahman` package contains the `Teams` dataset which was used for the calculation of the basic statistics along with the more advanced ones. The basic statistics used for this analysis were Fielding Independent Pitching (FIP), Weighted on base average (wOBA), batting average balls in play of the opponents (BABIPA), walks plus hits divided by innings pitched (WHIP) and the strikeout rates (K%).

- wOBA: Weighted on base average is considered to be a "better" version of batting average. wOBA was created by Tom Tango and he writes more in depth about it in "The Book" Tango et al. [2014].[3] wOBA is a weighted average of each of the ways a player can get on base (walking, getting a single, double, triple and a home run) and is divided by the ways a player gets a chance to hit (at bats, walks and sacrifice flies). Usually the coefficients of wOBA change by year but for this report, I used a manipulation of other advanced statistics that were described in "The Book", $(2 * OBP + SLG)/3$ [4]
- FIP: Fielding independent pitching estimates a player's run prevention if they did not have a defence, as opposed to ERA which includes defence. The equation for FIP is simple, $(13 * HRA + 3 * BBA - 2 * SOA)/IP$ where HRA is home runs allowed, BBA is walks allowed and SOA is the number of strikeouts a player gets. As Fangraphs says, "FIP is a measurement of a pitcher's performance that strips out the role of defense, luck, and sequencing, making it a more stable indicator of how a pitcher actually performed over a given period of time" (Slowinski).
- BABIPA: Batting average on balls in play against is pretty much the opposite of FIP. It looks at the batting average of only the balls that the defence has to field, as opposed to FIP which only looks at outcomes that the defence cannot control. The equation for BABIPA is: $(HA - HRA)/(IPouts - SOA - HRA)$ where HA is hits allowed, HRA is home runs allowed, IPouts is the number of outs a team gets while pitching and SOA is the number of strikeouts a team gets while pitching.
- WHIP: Walks plus hits divided by innings pitched is kind of self explanatory. It takes the amount of walks a team gives up, plus the number of hits and divides it by the number of innings pitched: $(BBA + HA)/IP$. WHIP is useful because it tells us how good a team (or pitcher) is at preventing runners from getting on base. If a team allows more people to get on base, they are more likely to give up runs and losing a game.
- SO%: Strikeout rates tell us how often a team (or batter) strikes out when they are batting. Some players have a tendency to only hit the ball far or not hit it at all while other players hit the ball almost all of the time. SO% helps show what type of gameplay a team relies on and it will be interesting to see how it affects a team's playoff chances. To calculate SO% we take the number of strikeouts a team got while batting and divide that by the number of times they bat SO/AB

[3] The Book is seen as the best book for aspiring baseball analysts or "saberists" to get into the MLB analysis game.

[4] OBP is on base percentage, the rate a player gets on base (includes walks). SLG is slugging percentage, the total bases (TB) a player gets divided by at-bats (a single = 1 TB, double = 2, triple = 3 and Home Run = 4).

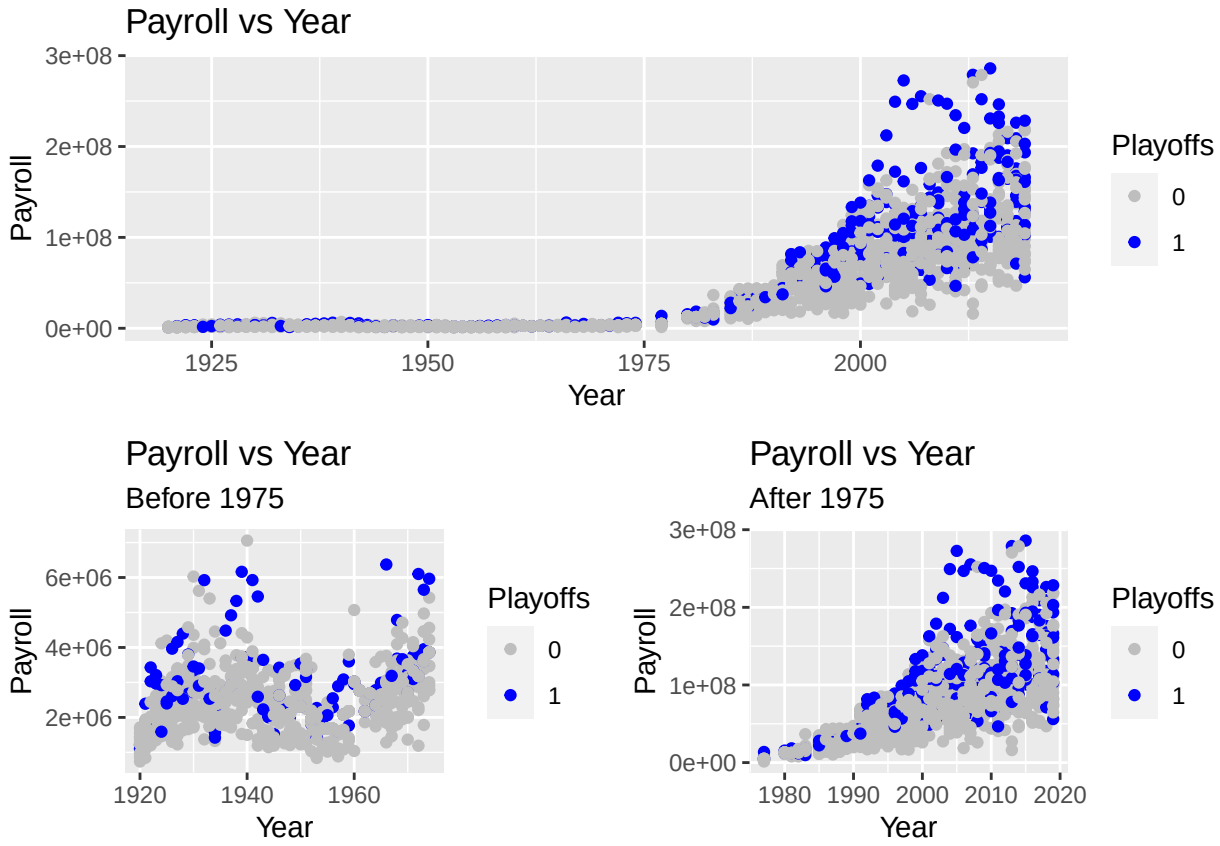

Figure 1: The increase in Payroll over the Years

Figure 1 shows us the increase in payroll over the years, starting at 1920 and ending at 2019. We see that up until 1960, payrolls were fairly constant but after the 60s we have seen a rapid increase in payroll. We can also see colours which represent if a team made the playoffs or not (light blue representing yes and dark blue meaning no). It is clear that in most seasons, teams that spend more make the playoffs. The salaries have been adjusted to represent the dollar in terms of what it would be in 2019 using inflation rates from the US Inflation Calculator (cit [2020]). Clearly, the adjustment does not do much as we see that teams were likely just spending more compared to the past.

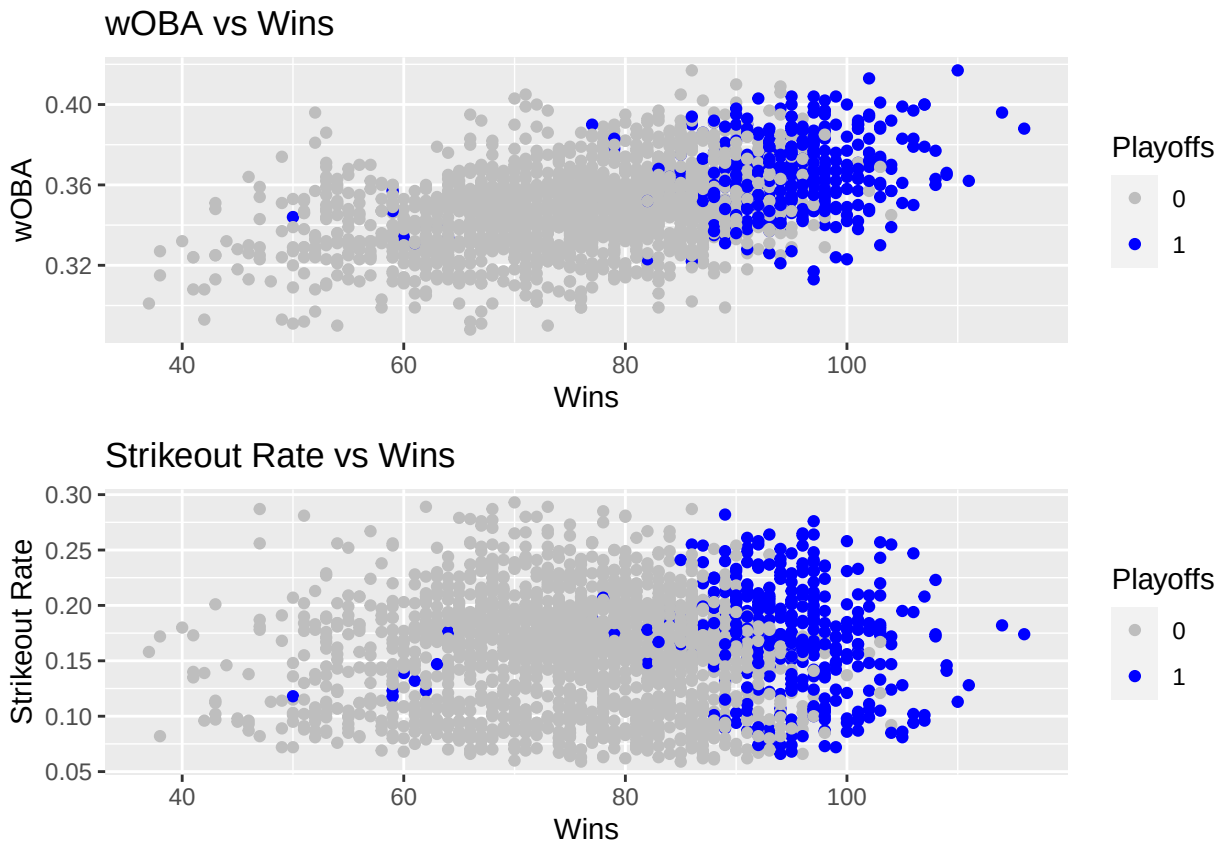

Figure 2: Relationship between Advanced Offense Metrics and Wins

Figure 2 shows the relationship between offensive advanced metrics and the amount of teams a win gets, along with colours representing a team's ability to make the playoffs. From the first graphs, we see that as wOBA increases, the amount of wins increases and the number of teams making the playoffs increases as well. We also see that there does not appear to be too much of a relationship between the number of wins a team gets and their strikeout rate. This makes sense because some teams are built on hitting home runs and strikeout a lot, versus teams who play more strategically and don't hit the ball as hard.

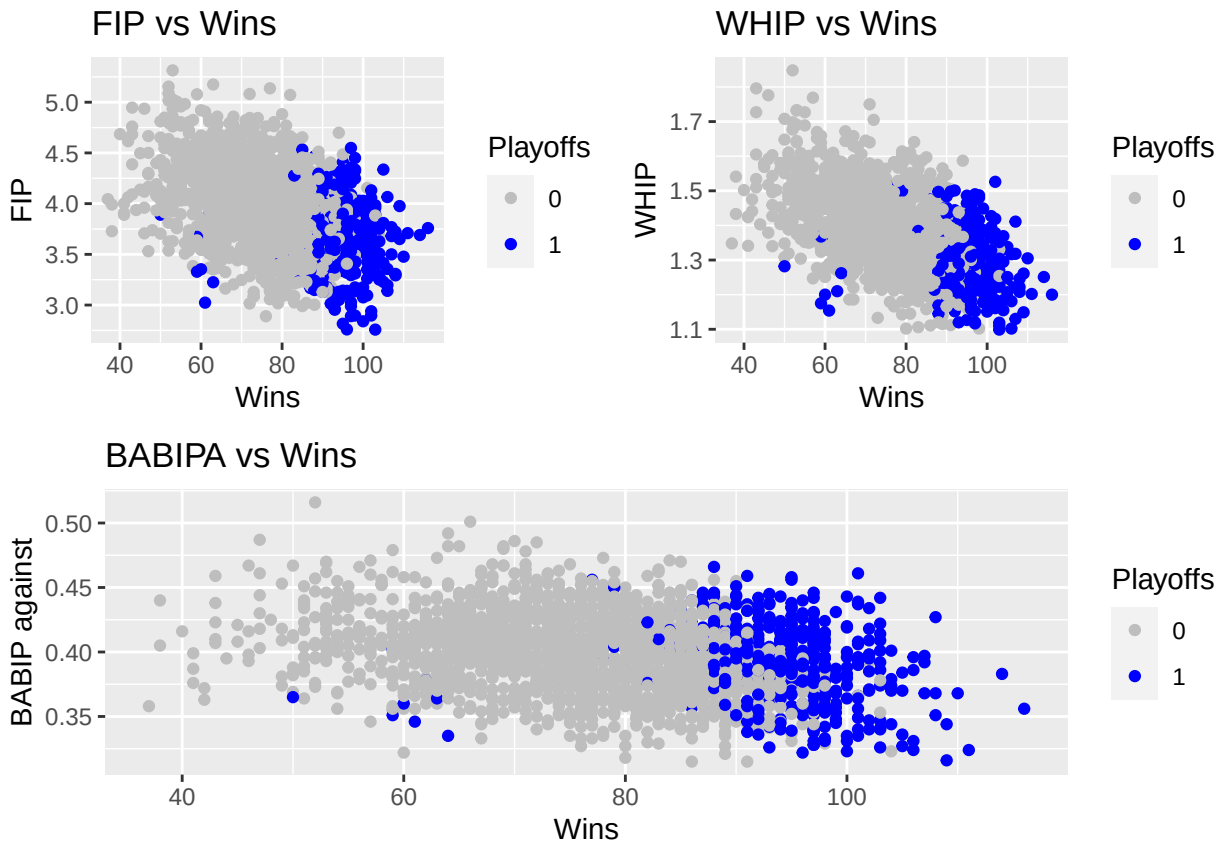

Figure 3: Relationship between FIP and Wins

Figure 3 shows the relationship between some advanced pitching metrics and the amount of wins a team gets in a season. We see that WHIP appears to have a very strong impact on wins and if a team can keep their WHIP as low as possible, they're more likely to make the playoffs. It appears that FIP has the next strongest relationship with wins, better teams have lower FIPs. Lastly, it appears that BABIP against does have an impact on the amount of wins a team gets, just not as significant as FIP and WHIP.

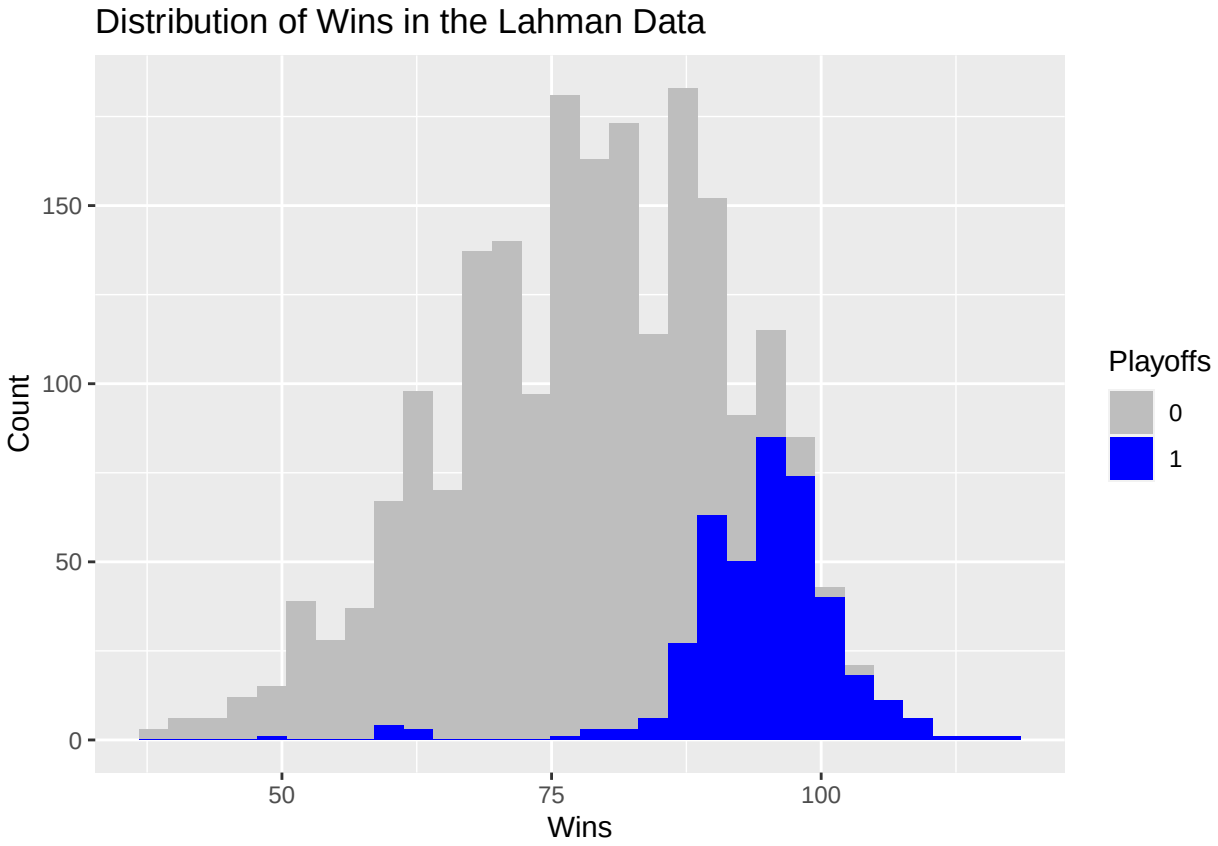

Figure **??** shows the distribution of wins for teams present in the Lahman data. We see that majority of team make the playoffs when they win about 85 games although there is some spread in the distribution of wins for playoff teams. There are also some teams present which made the playoffs after winning less than 75 games. This occured because of dispute between the players and the league which led to a 50 day strike and led to teams winning less games in total (Bumbaca [2020])[5].

3 Model

In order to get playoff probabilites, I will use logistic regression. Logistic regression takes a sum of various factors and then after manipulating the equation, we will get a probability that tells how likely something is to happen. In our case, logistic regression is used to determine the probability that a team will be making the playoffs. The logistic regression will be in the form:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 x_{FIP} + \beta_2 x_{wOBA} + \beta_3 x_{BABIPA} + \beta_4 x_{SOrate} + \beta_5 x_{before75} \quad (1)$$

where x_{FIP} represents the team's FIP, x_{wOBA} represents the team's wOBA, x_{BABIPA} represents the team's BABIP against, x_{SOrate} represents the team's strikeout rate and $x_{before75}$ represent whether the team in question was playing before 1975 or not.

In Equation (1), the β values represent a coefficient determined by the `glm` function, the will either be positive or negative, depending on if a higher x values increases or decreases a team's chances of making the playoffs. Each β value will be multiplied by it's corresponding x, for example, teams with higher wOBA's tend to make the playoffs more often, so the β value for x_{wOBA} will likely be positive. Once all of the coefficients are multiplied we take the sum of the equation and insert it into the equation below:

[5]The players eventually returned and the playoffs continued

$$\frac{e^{sum}}{1+e^{sum}} \tag{2}$$

Equation (2) is just a manipulation of equation (1), where e is the exponential equation and sum is the sum of the right side of equation (1). We see that as the sum of the equation increases, the probability of a team making the playoffs increases as well. The logistic regression model is run using the `glm()` function in R (R Core Team [2019]). The decision to run this model over other models like linear regression was made by the fact that we were predicting a binary variable about a voter's decision. Since there are only two possible options our data will likely follow an S shape and a straight line equation will not be helpful to model this relationship.

One weakness present with a logistic regression model is that since the response variable must be binary (either make the playoffs or don't), we cannot expand it to determine if a team wins their division or just scrapes into the playoffs.

Table **??** shows us the coefficients from our logistic regression model. We can see that all variables except the intercept have statistically significant p-values, meaning that we know for sure that the variables have an effect on a team's playoff chances. The intercept's insignificance is not bad in our situation because it means that we don't know for sure if the intercept is not 0.

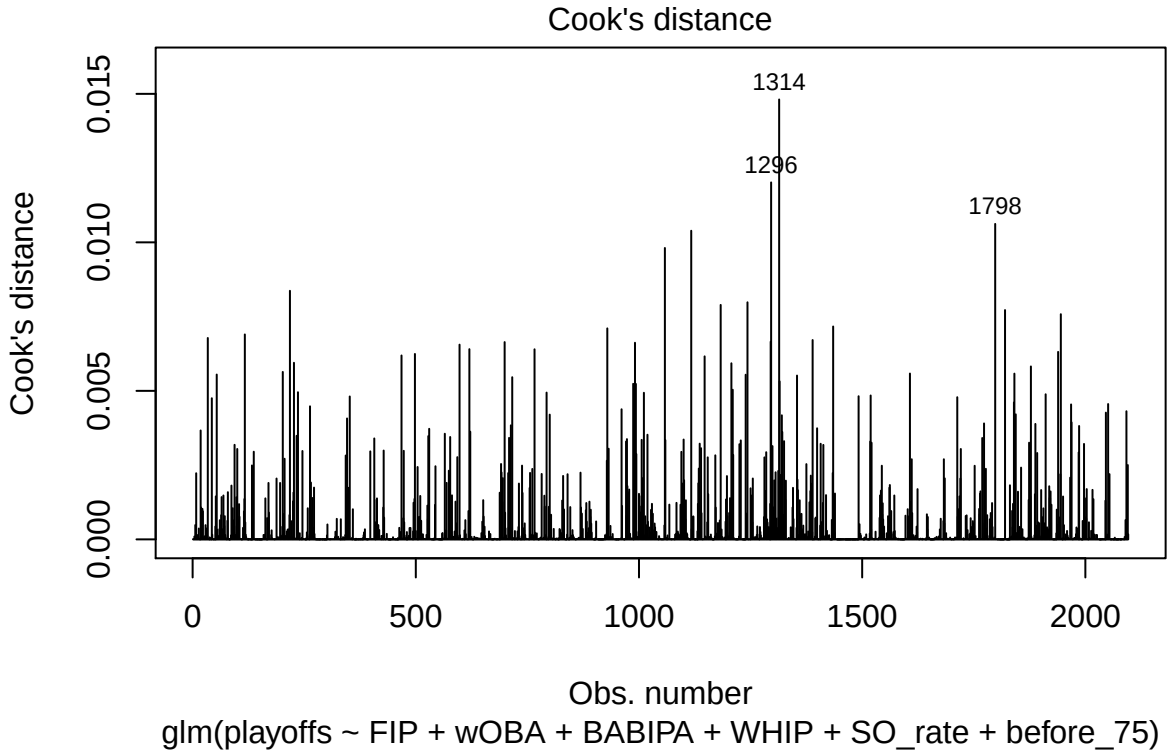

Figure 4: Cook's Distance for the Model

Figure 4 shows us the observed cook's distances for the model. Cook's distance gives us the difference between the ... . We see that these cook's distances are very low (most are under 0.01), since our distances are fairly low, we can conclude that the model is fitting the data fairly well.

Table 1: Regression Model Coefficients

term	estimate	std.error	statistic	p.value
(Intercept)	-8.33451	1.78681	-4.66448	0.00000
FIP	-2.80547	0.43729	-6.41554	0.00000
wOBA	117.03659	6.65185	17.59460	0.00000
BABIPA	-32.95231	5.86867	-5.61496	0.00000
WHIP	-9.78865	2.29391	-4.26724	0.00002
SO_rate	12.28808	2.92559	4.20020	0.00003
before_75	-0.58262	0.22962	-2.53730	0.01117

4 Results

Table 1 shows us the coefficients from our logistic regression model. We can see that all variables except the intercept have statistically significant p-values, meaning that we know for sure that the variables have an effect on a team's playoff chances. The intercept's insignificance is not bad in our situation because it means that we don't know for sure if the intercept is not 0.

We see that using the logistic regression equation, the lowest possible probability of making the playoffs is $4.23 * 10^{-5}$ percent. This was calculated using the worst possible stats present in the data used and it is extremely unlikely a team would be this bad! The highest possible probability of making the playoffs is 99.99 percent. This was taken from the maximum stats present in the data, and although it would be tough for a team like this to exist, there have been some teams that were close.

Data Type	Accuracy (%)	result
Data	-	18.989
Model	87.9770992366412	15.267

Table **??** displays the proportion of teams which made the playoffs from the observed data vs the results of our model's predictions. We can see that the model is very strict with its results, only predicted about 15 percent of teams to make the playoffs, less than the observed 18.9 percent.

We can also look at the overall accuracy of the model. In order to find the accuracy of the model, I compared the predicted results to the actual results and we find that the model succesfully predicted if a team would make the playoffs about 88 percent of the time. These results are very significant because it shows that the model can fairly succesfully determine what a playoff team is. This also bodes well for our attempt to see if the results for the 2020 predictions will be correct.

5 Discussion

Now that the model has been created (and tested) and we have our results, what can we make of them? Well, the first major point we can see is that when using advanced statistics, we can accurately predict if a team will make the playoffs 88 percent of the time. This result is truly significant because it means that if a team put their focus on acquiring players who excel in advanced statistics but are average in the 'regular' ones, they can drastically improve their chances of making the playoffs.

When looking at our batting stats, we see that wOBA is very influential in determining a team's playoff status. Taking a quick look at Fangraphs' wOBA leaderboard from 2019[6] we see that 6 of the top ten players in wOBA were not in the top 10 for batting average. If teams are still evaluating players using traditional statisitics, they have a much higher chance of paying for a player who may not have a significant impact

[6]We're going to focus on 2019 because the 2020 season was just a really small sample size

on their playoffs chances. If teams make the shift to stats like wOBA, they certainly won't be hurting their chances at making the playoffs.

Next, we can also take a look at FIP, now that we know it is significant. FIP is much more difficult to evaluate compared to the traditional stats because of how it is calculated. As we saw in the data section, FIP is calculated only by true outcomes: home runs, strikeouts and walks. Since FIP only deals with these "true outcomes", it tends to overlook players who don't rely on strikeouts but instead rely on their teamates to help them out. Regardless, finding players with an above average FIP[7] and below average traditional is fairly simple. Once again, from fangraphs we find that there are many pitchers with high ERAs and above average FIPs. This is significant because FIP predicts success better than ERA. In a Sporting News article from 2017, John Edwards talks about a player named Michael Pineda's 2016 season. He started off the season with a high ERA but a low FIP and as the season went on, his ERA began to decrease (Edwards [2017]). This is one of many examples where FIP outperforms ERA and this shows...

6 Weaknesses and Next Steps

[7] under about 4.2

References

Us inflation calculator, Dec 2020. URL https://www.usinflationcalculator.com/inflation/consumer-price-index-and-annual-percent-changes-from-1913-to-2008/.

Chris Bumbaca. Explaining the 1981 mlb season: How baseball survived shortened year, Mar 2020. URL https://www.usatoday.com/story/sports/mlb/2020/03/15/1981-mlb-season-coronavirus-delay-baseball/5054780002/.

John Edwards. Stat to the future: Move over, era, it's time for fip, Sep 2017. URL https://www.sportingnews.com/us/mlb/news/what-is-fip-era-pitching-baseball-mlb-stats-statistics-advanced-sabermetrics/d5p48p6z6us51lqbo0wvgigto.

Fangraphs. Guts!: Fangraphs baseball. URL https://www.fangraphs.com/guts.aspx?type=cn.

Michael Friendly, Chris Dalzell, Martin Monkman, and Dennis Murphy. *Lahman: Sean 'Lahman' Baseball Database*, 2020. URL https://CRAN.R-project.org/package=Lahman. R package version 8.0-0.

Sports Reference LLC. *Baseball-Reference.com*. Major League Statistics and Information. https://www.baseball-reference.com/.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

Steve Slowinski. Fip. URL https://library.fangraphs.com/pitching/fip/.

Tom M. Tango, Mitchel G. Lichtman, and Andrew E. Dolphin. *The book: playing the percentages in baseball*. TMA Press, 2014.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

Yihui Xie, J.J. Allaire, and Garrett Grolemund. *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida, 2018. URL https://bookdown.org/yihui/rmarkdown. ISBN 9781138359338.