Logistic Regression shows that Moneyball was right

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, I will use logistic regression to determine whether the advanced stats that stemmed from this revolution are useful for predicting a team's success. According to the model (I still have to do this). The model correctly predicted that (insert) teams would make the playoffs using these metrics.

Keywords: Moneyball, Baseball, Sabermetrics, Logistic Regression

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

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 (n.d.)). 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, Allaire, and Grolemund (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, 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,

[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.

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!

Data

The data collected was from the `Lahman` package in R (Friendly et al. (2020)) and Baseball-Reference (LLC (n.d.)), Fangraphs (Fangraphs (n.d.)) 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 Batting Average, Earned Run Average, Weighted On Base Average and Fielding Independent Pitching. Batting average is the rate at which a player will get a hit, it is calculated by $Hits/Atbats$. Earned Run Average is the amount of runs a player allows, per 9 innings (1 whole game), it is calculated by $Runs/InningsPitched$. Next, we get to the advanced statistics, 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, Lichtman, and Dolphin (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] Lastly, 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.

Model

term	estimate	std.error	statistic	p.value
(Intercept)	-29.5168	2.3831	-12.3858	0.0000
wOBA	137.6176	9.8297	14.0002	0.0000
FIP	-0.2411	0.3711	-0.6496	0.5159
log(payroll)	0.4337	0.0499	8.6986	0.0000
BA	-29.5163	10.7073	-2.7567	0.0058
ERA	-4.9775	0.3524	-14.1248	0.0000

Table @ref(tab:model) shows us the coefficients from our logistic regression model. We see that teams with a higher wOBA have better chances of making the playoffs, along with team with higher payrolls. This makes sense because when teams will attempt to spend the most money to get the best players and when they score more runs they'll win more games. We can also see that the coefficient for FIP is negative, this does not mean that FIP brings down the chances of making the playoffs. The best values for FIP are the ones closer to 0 and as the value begins to increase, it shows that the pitcher is not that good. What the coefficient is telling us is that lower FIP's give higher chances of making the playoffs while higher FIP's bring down the chances of making the playoffs.

[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).

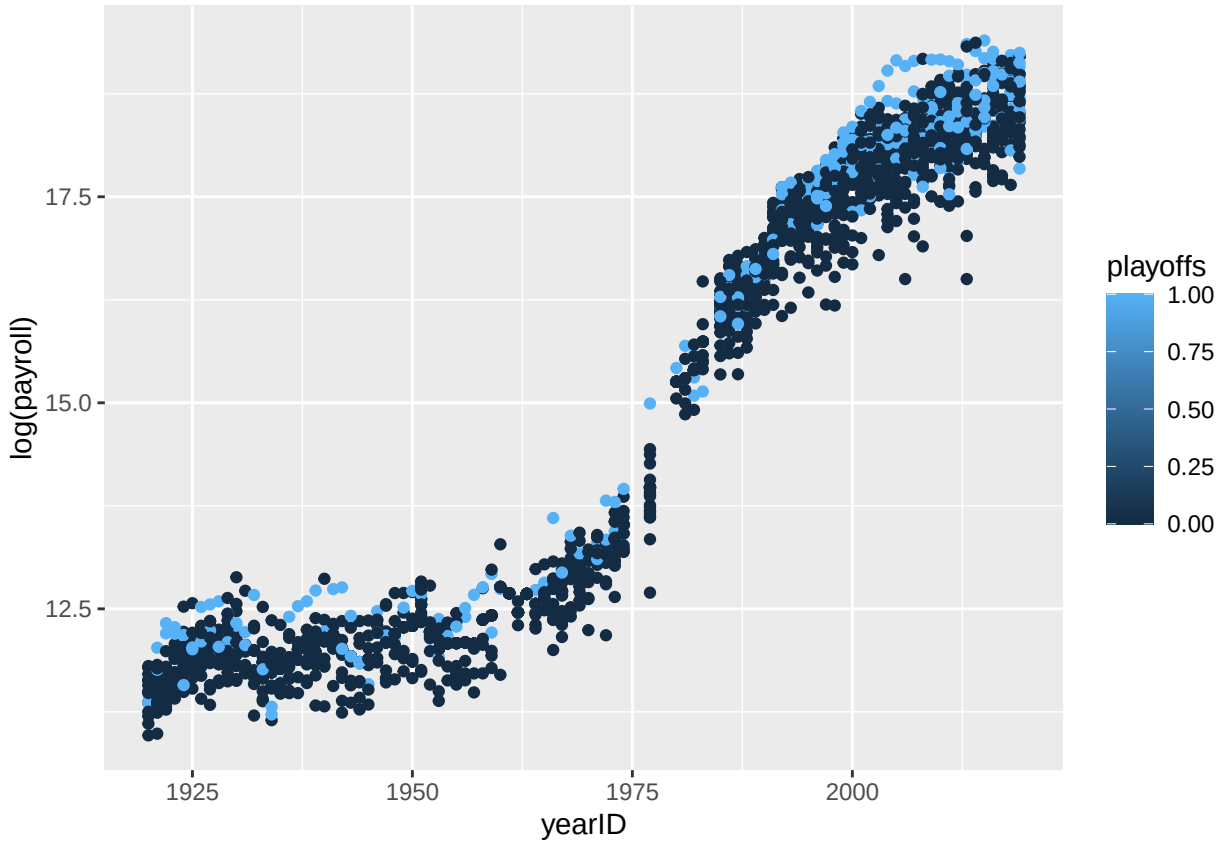

Figure 1: The increase in Payroll over the Years

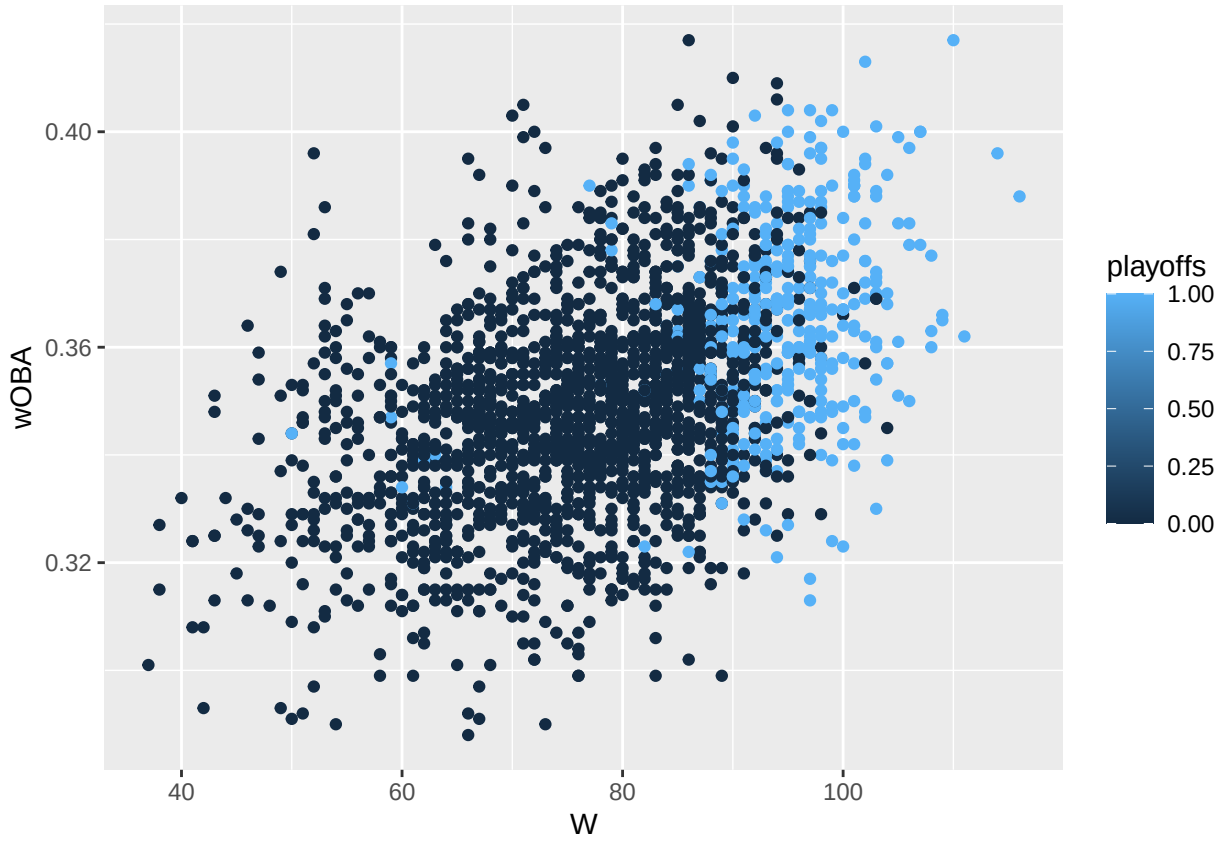

Figure 2: Relationship between wOBA and Wins

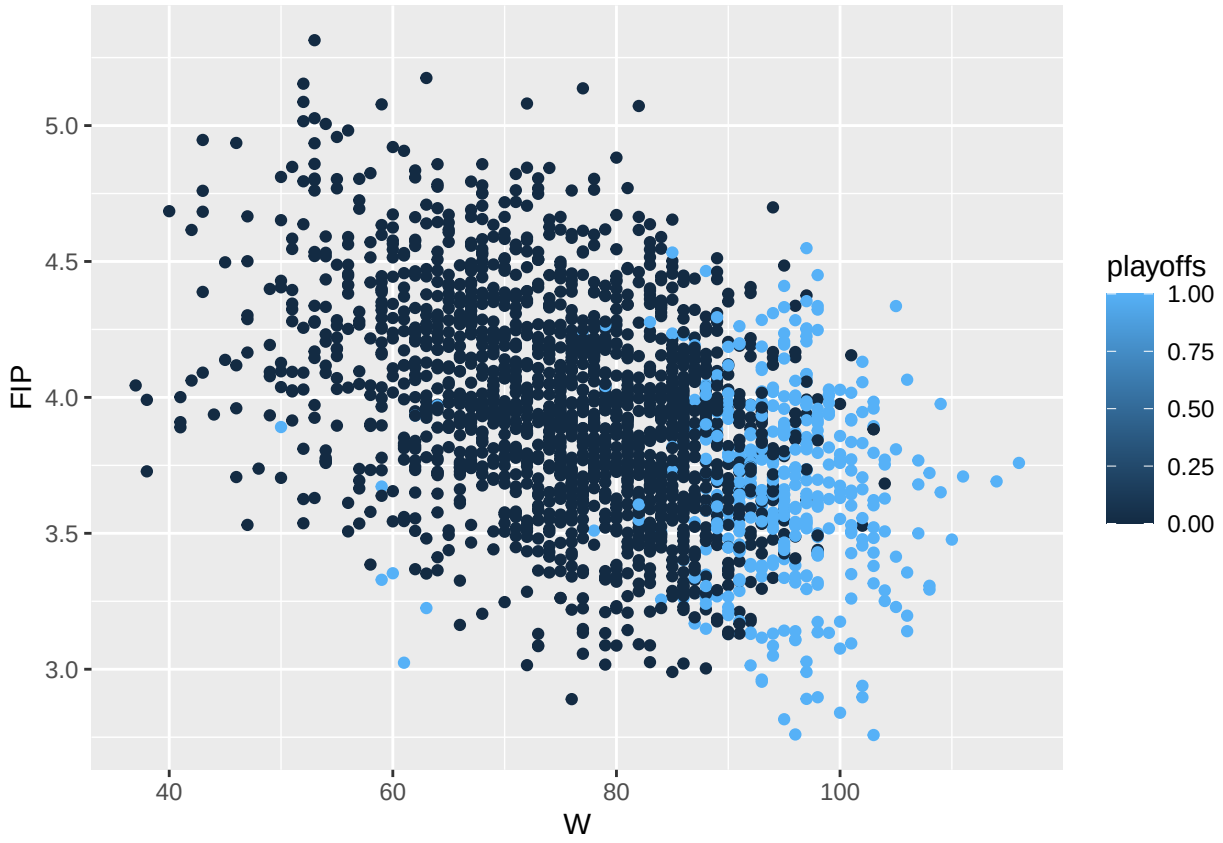

Figure 3: Relationship between FIP and Wins

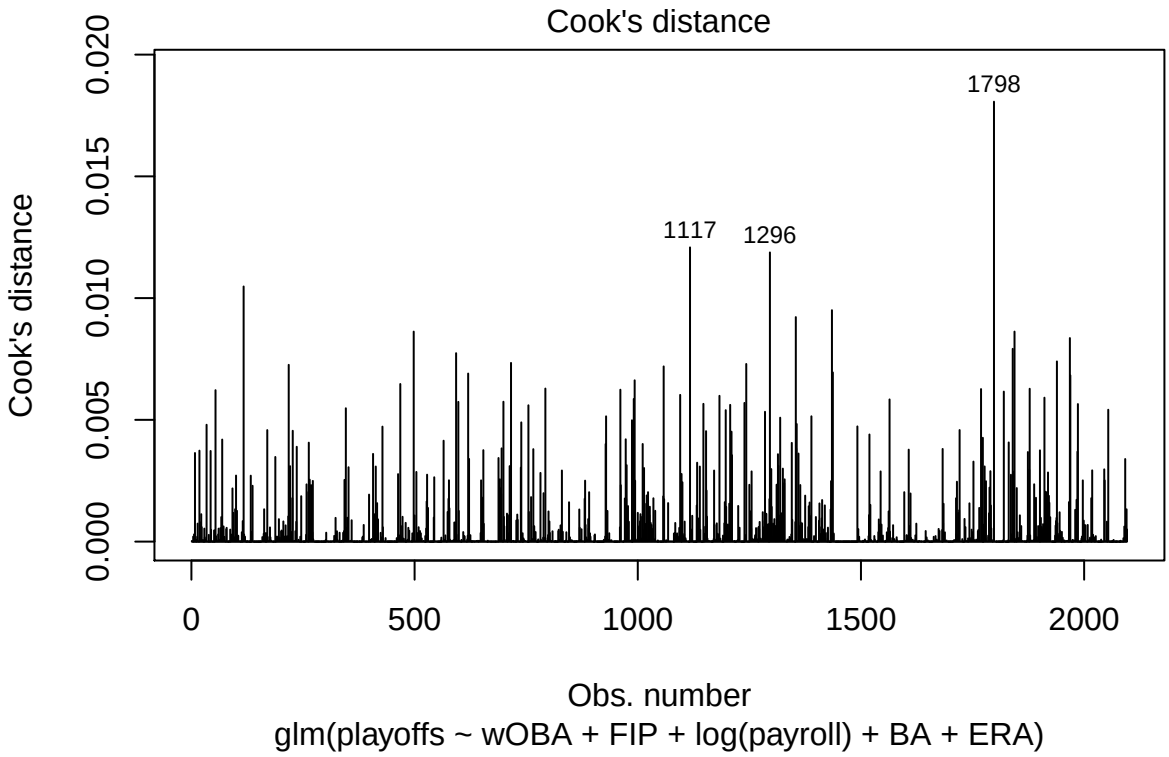

Figure @ref(fig:cooks) shows us the observed cook's distances for the model. Cook's distance gives us the difference between the . . . . We see that these cook's distances are very low (most are under 0.01), since our distances are fairly low, we can conclude that the model is fitting the data fairly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

```
## [1] 0.1898855
```

```
## [1] 0.1593511
```

Results

Discussion

Weaknesses and Next Steps

References

Fangraphs. n.d. "Guts!: FanGraphs Baseball." https://www.fangraphs.com/guts.aspx?type=cn.

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

LLC, Sports Reference. n.d. *Baseball-Reference.com*.

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

Tango, Tom M., Mitchel G. Lichtman, and Andrew E. Dolphin. 2014. *The Book: Playing the Percentages in Baseball*. TMA Press.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

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