Was the basis of Moneyball Correct?

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

Introduction

In 2003, Michael Lewis released 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.

This data analysis was conducted using R (R Core Team (2019)), and in particular the packages Tidyverse (Wickham et al. (2019)) and Lahman (Friendly et al. (2020)) and was compiled using R markdown (Xie, Allaire, and Grolemund (2018)).

Data

Model

```
##
## Call:
## glm(formula = playoffs ~ wOBA + FIP + log(payroll) + BA + ERA,
##     family = binomial(), data = all_years)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.52304  -0.37238  -0.11565  -0.01555   3.01252
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -29.51677    2.38311 -12.386  < 2e-16 ***
## wOBA         137.61756    9.82966  14.000  < 2e-16 ***
## FIP           -0.24111    0.37114  -0.650  0.51592
## log(payroll)   0.43368    0.04986   8.699  < 2e-16 ***
## BA           -29.51632   10.70731  -2.757  0.00584 **
```

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

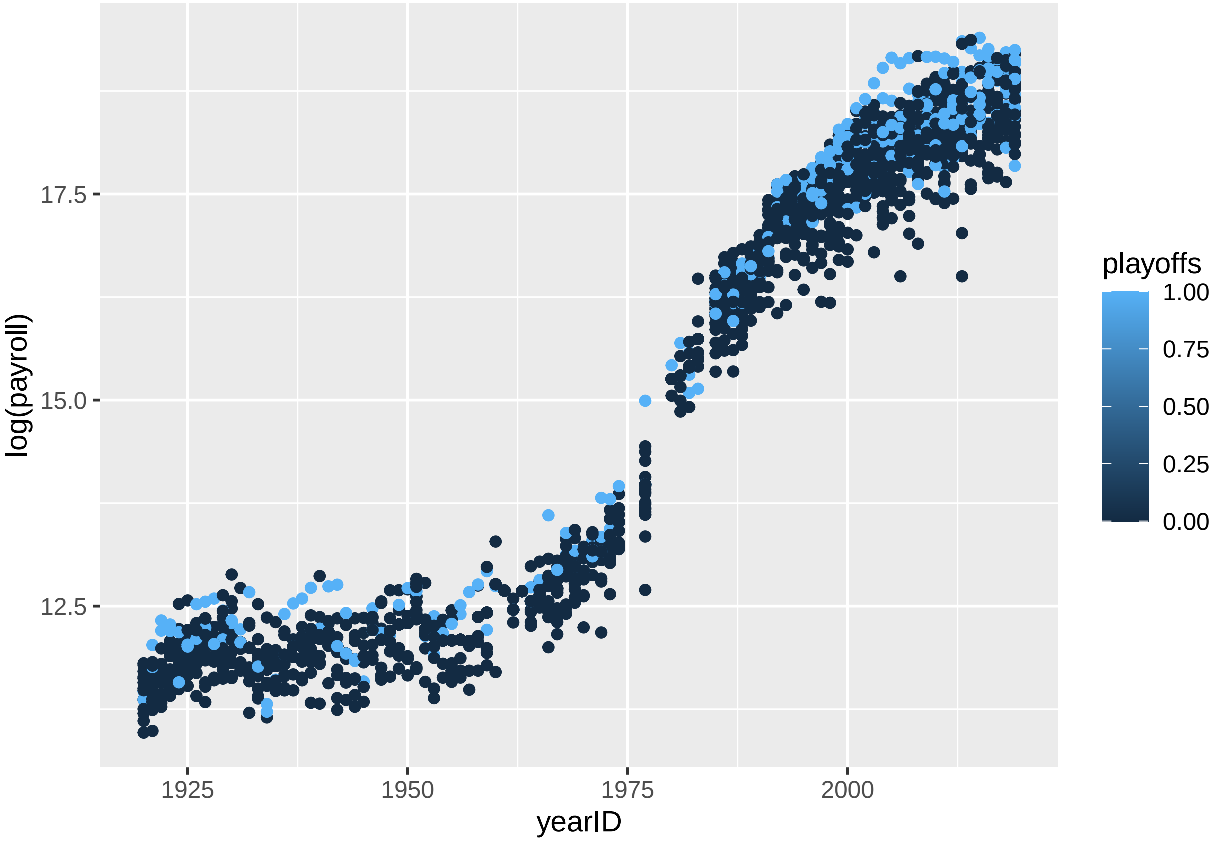

Figure 1: The increase in Payroll over the Years

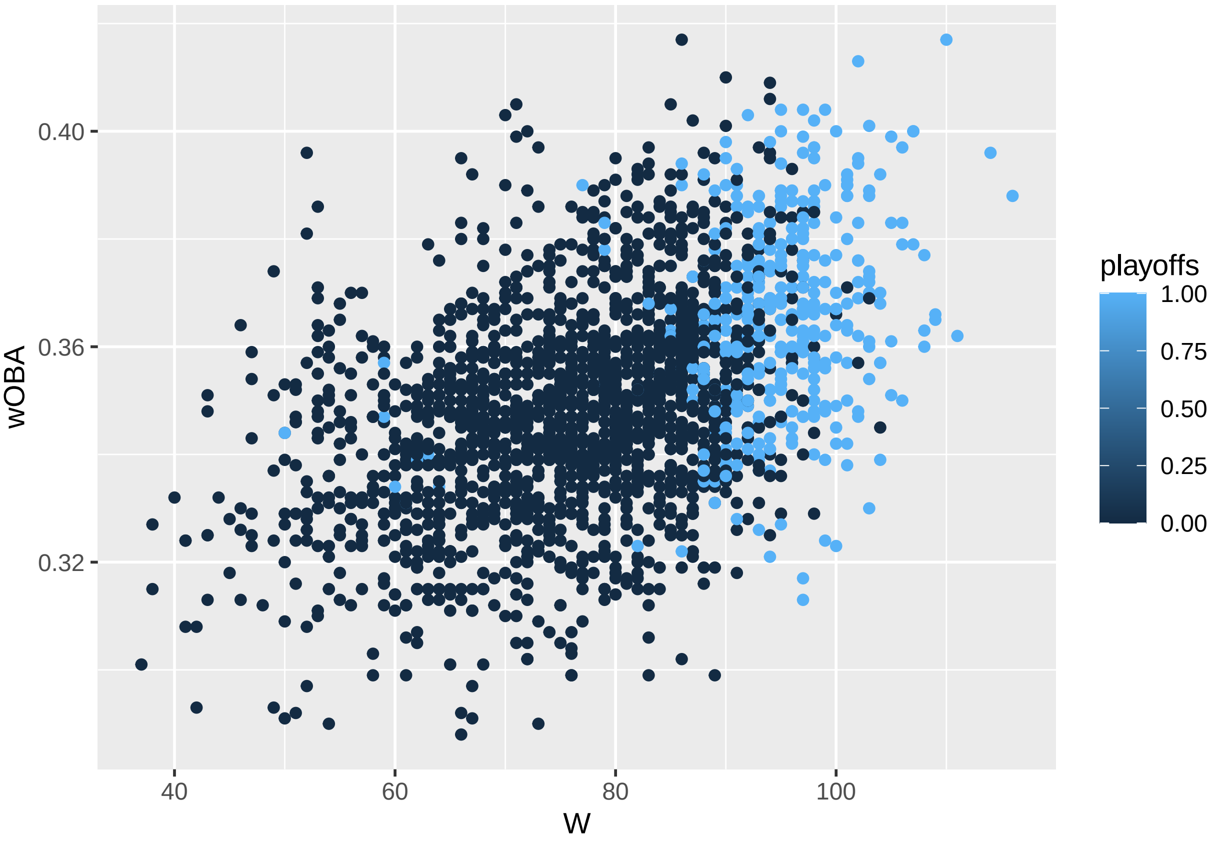

Figure 2: Relationship between wOBA and Wins

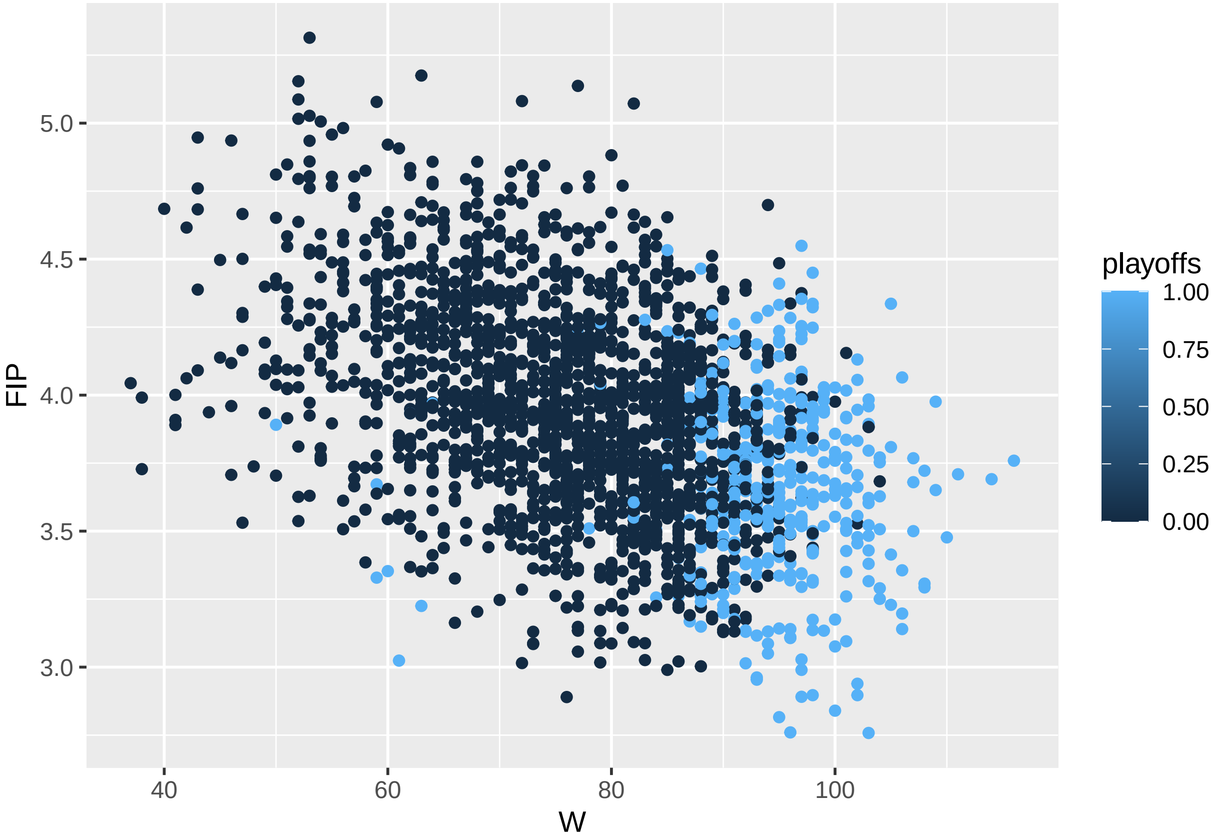

Figure 3: Relationship between FIP and Wins

```
## ERA           -4.97746    0.35239 -14.125  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2037.6  on 2095  degrees of freedom
## Residual deviance: 1029.8  on 2090  degrees of freedom
## AIC: 1041.8
##
## Number of Fisher Scoring iterations: 7
```

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.

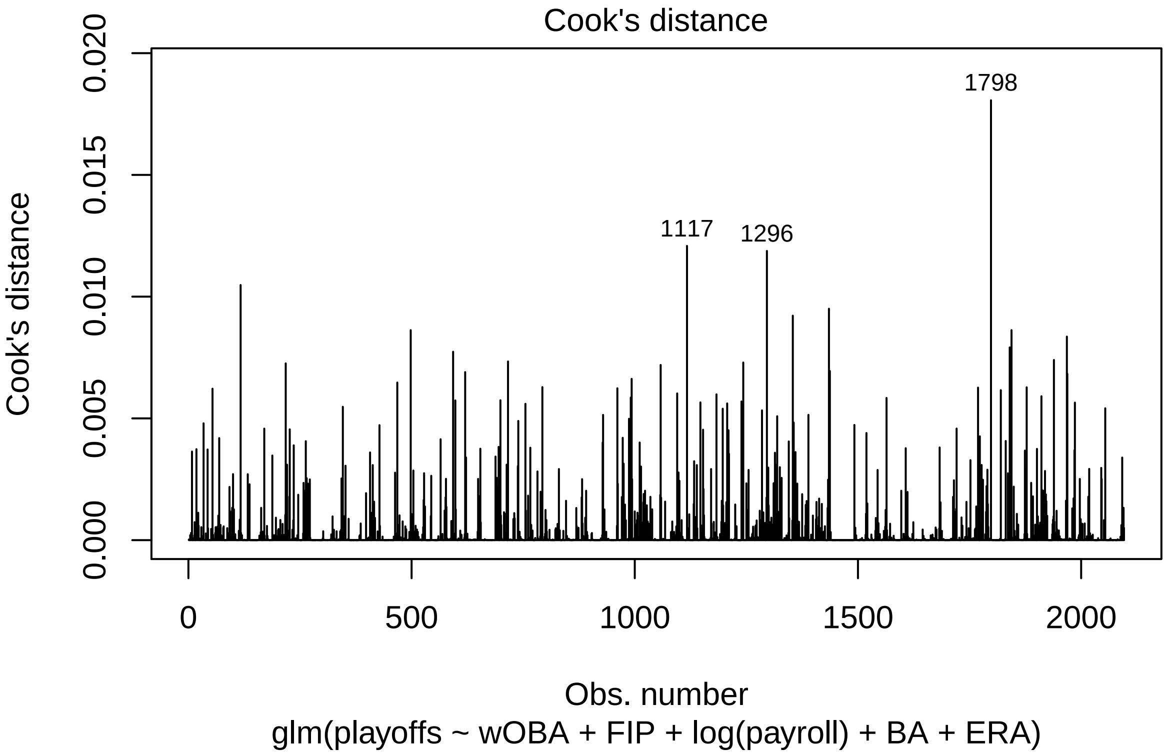

Figure @ref(fig:cooks) shows us the observed cook's distances for the model. Cook's distance gives us the difference between the ... . We see that these cook's distances are very low (most are under 0.01), since our distances are fairly low, we can conclude that the model is fitting the data fairly well.

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

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

Discussion

Weaknesses and Next Steps

References

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

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

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.