Modeling Baseball Probability of Winning the League

Connor Demorest

2/16/2021

Introduction

Baseball data are the low hanging fruit for statisticians; there is the result of every inning going all the way back to 1871, every play since the 1980's, and every pitch since the early 2000's. The regular season in baseball is 162 games long, long enough that large sample asymptotic theory and law of large numbers start to make an impact. In other sports, such as football, the regular season is only (currently) 16 games, which may not be long enough to find out if a team is truly good enough to be a playoff team or if they are truly average and got lucky. According to Reddit, the probability of a football team making the playoffs with a 10-6 record is around 97% (source: https://www.reddit.com/r/theydidthemath/comments/2o0zuj/requestnflhow_likely_is_it_for_a_team_to_go_106/, so very likely. If a football team is truly average with an expected probability of winning each game is 50%, there is a 22.72% probability that a team wins 10 or more games and makes the playoffs just by random chance. In contrast, the chance that as baseball team with the same observed winning percentage is actually a .500 team is 0.1%.

All that was to say that baseball is very nice to model as a statistician, and when you think about the long run averages, they will tend to occur over the course of a season. This is portrayed in the book Moneyball by Michael Lewis and the movie made about it in 2011 starring Brad Pitt and Jonah Hill, which describes the buy-in from baseball into statistics and player evaluations. In the movie, the main idea is that there is a fundamental misunderstanding about how teams win games: "Your goal shouldn't be to buy players, your goal should be to buy wins. And in order to buy wins, you need to buy runs." A current baseball fan would know that Jonah Hill portraying Peter Brand is talking about sabermetrics, the application of statistics and mathematical modeling to baseball.

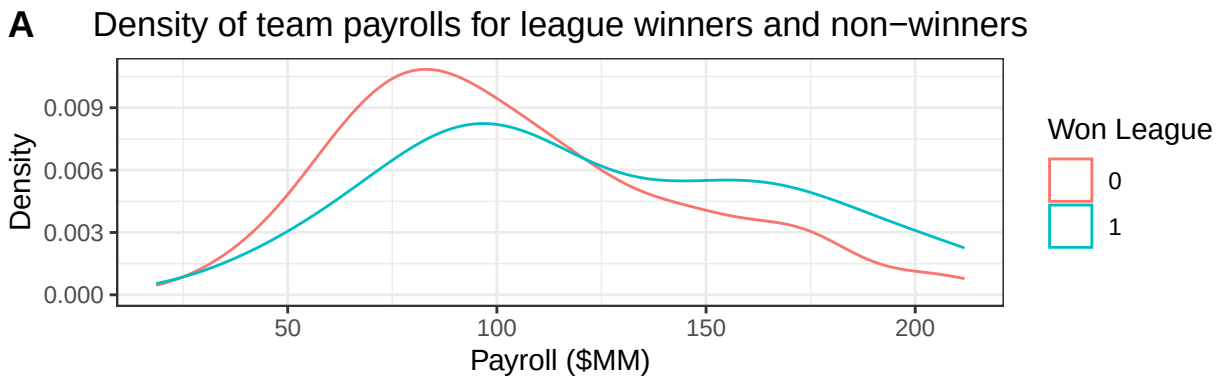
In 2021, there is essentially universal agreement from baseball executives that this is the ideal way to evaluate players, using these so-called "advanced statistics". Presumably, the way that each team uses statistics to create their teams is proprietary, but the goal for this project is to make a simple model to see what should be valued for a team to try to win their division and make it into the playoffs.

Data Overview

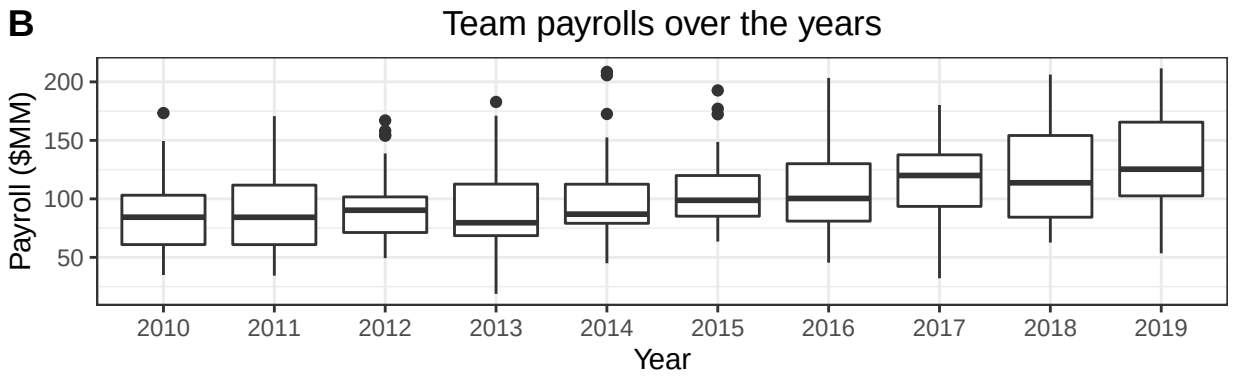
These data come from 3 sources: The Lahman Database, USAToday, and Baseball-Reference.com

The Lahman Database gives me the main portion of the data that contains counting statistics and the indicator for the response - if a team makes the World Series or not. It's a database maintained by volunteers on GitHub and exported to R occasionally. It has basic statistics for both player and team level and lots of tables.

The Lahman database does have a table for player and team salaries, but it's only updated through the 2016 season. I expected the team payroll to be an important predictor of if a team was successful or at least something that I would want to control for, so I scraped the payrolls for the 2010-2019 seasons off USAToday's website. Figure A shows the distribution of payrolls for teams that won their league and teams that didn't. Figure B shows that the average salary has been increasing since 2010 but the maximum salary has stayed around the 200 million dollar mark since 2014.



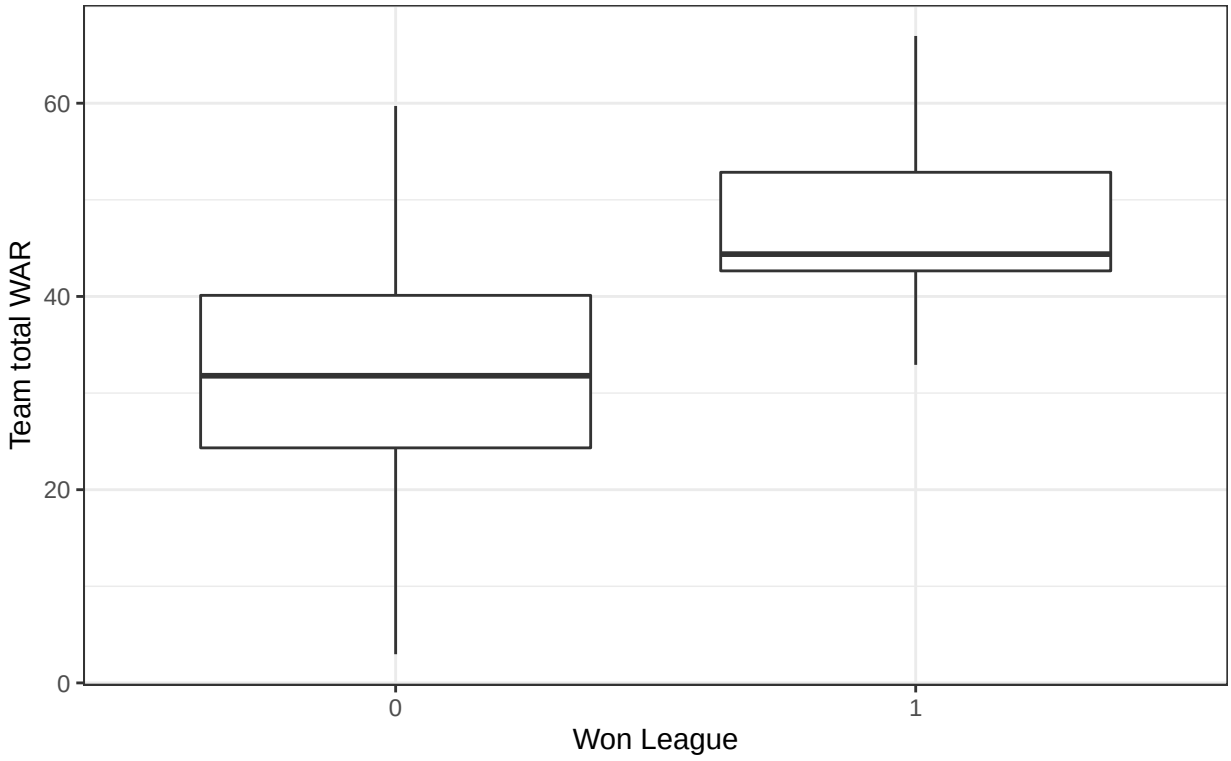
Team payrolls are right skewed and more right skewed for league winners



Average payrolls have increased over time but max payrolls have stayed the same since 2014 ($p < 0.001$)

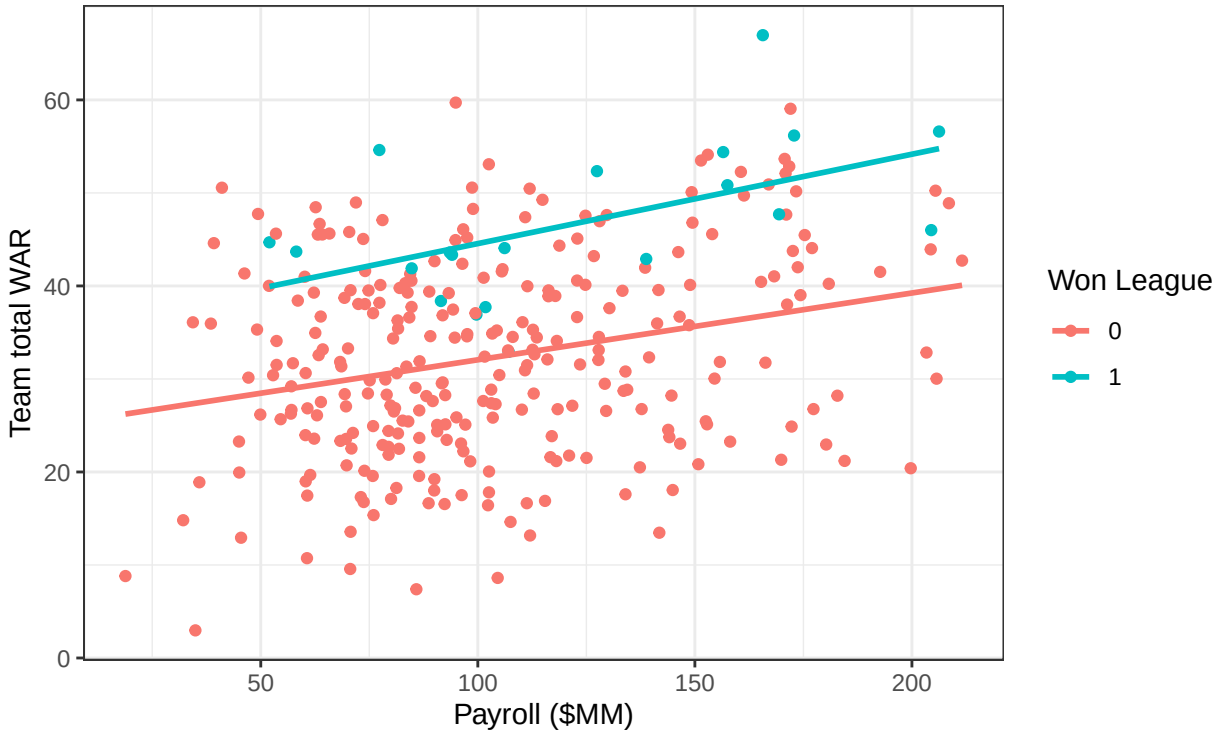
Baseball-Reference.com contains advanced statistics for each team, which are essentially estimators of player performance, typically adjusted for positional value, scaled to the league average or "replacement level", and controlling for other variables like where a team plays. The most common of these advanced statistics is Wins Above Replacement, or WAR. Plot C shows that there is higher WAR for teams that won their league compared to those that didn't. Plot D shows that that there is a weak association between payroll and WAR, and that there is no interaction between the two, indicating that teams who win their league don't get more WAR out of their players at the same payroll compared to teams that don't win their league.

C Team total WAR for league winners and non−winners



Teams that won their league have a much higher average WAR ($p < 0.001$)

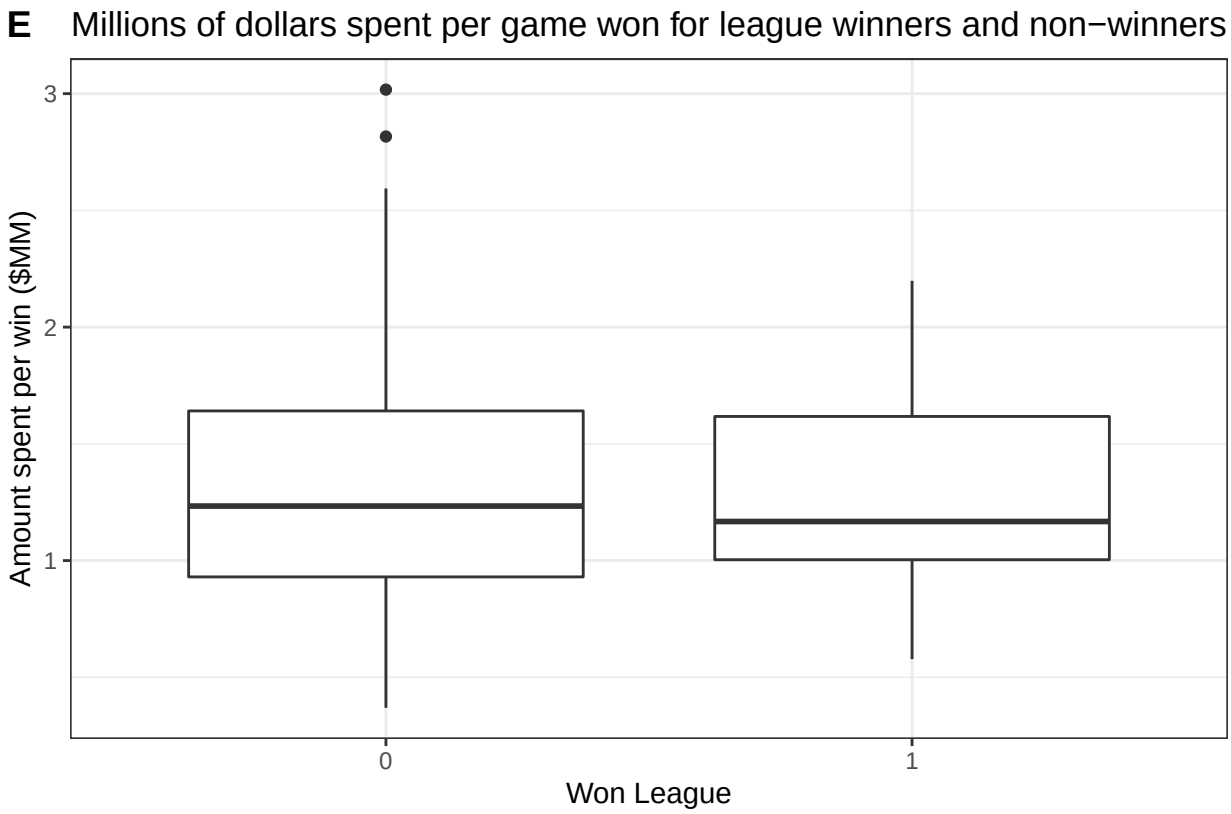
D Plot of Payroll and WAR for league winners and non−winners

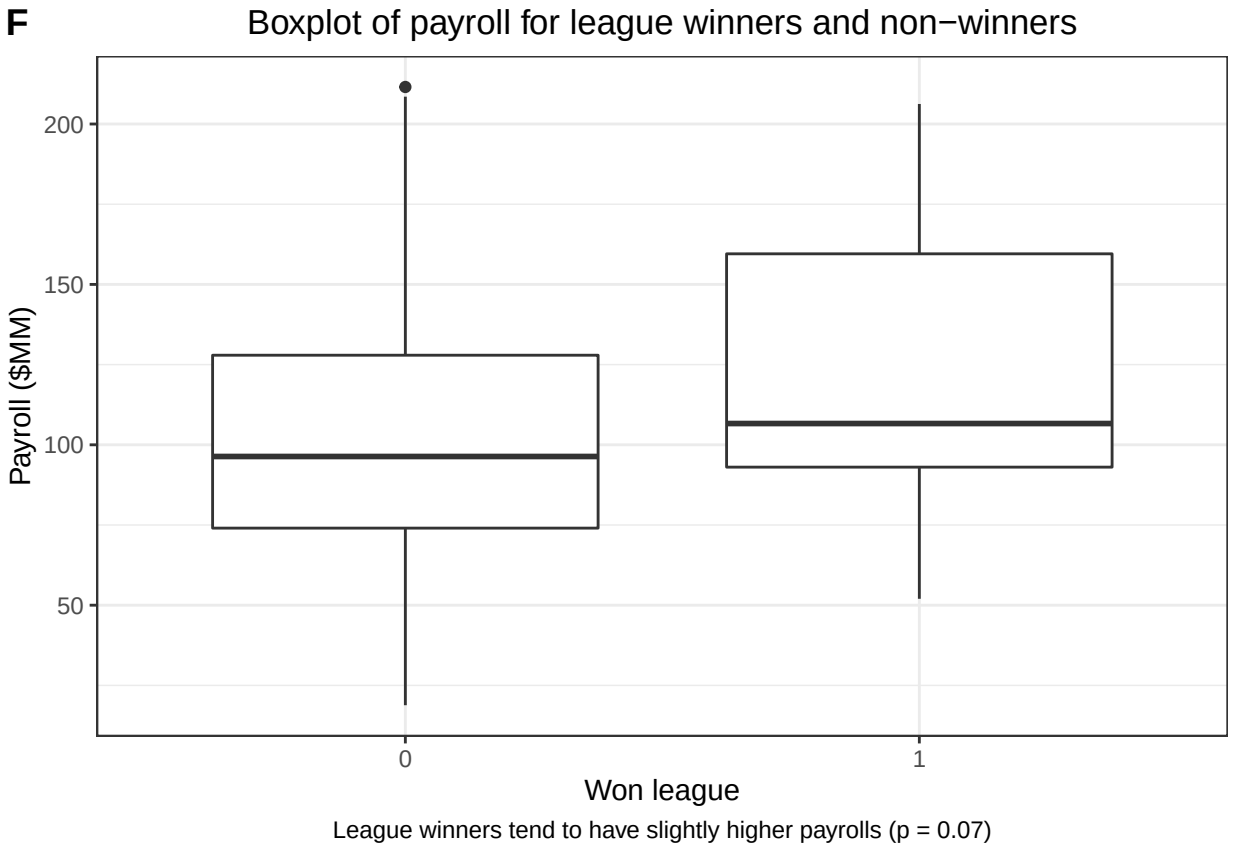


Teams with a higher payroll have a higher total WAR on average (corr = 0.3) and higher WAR is associated with winning the league. There is no evidence of different slopes indicating an interaction.

WAR comes from the idea that players can create wins essentially by themselves by creating runs on offense or preventing them on defense. This is in comparison to a "replacement player" which is the average player that a team could get for free from the minor leagues. A team full of replacement level players is expected to win around 48 games in a 162 game season. A player who gets 1 WAR in a season is a solid contributor, 2-5 is a starter, and 5+ is one of the best players in the league. WAR can be interpreted as the number of wins that they can be attributed with over the course of the season above the 48 that is replacement level. If you take the sum of a team's player's WAR, you have an estimated number of wins above replacement level of 48 wins for that team. More reading about WAR and how it is calculated is found at baseball-reference.com.

The data are filtered to only include the 2010-2019 seasons for a couple reasons. I don't have 2020 data in the Lahman database yet, and I don't know if the shortened COVID season is representative of the future. There were several calculated statistics that I used and did not use. Instead of using payroll in dollars, it was scaled to millions of dollars and centered around the median. A metric for the "efficiency" of a team's payroll were calculated, price per win which was Payroll/# of Wins. I expected the efficiency of how a team spends their money to be a good predictor of if they are competitive. The price per win could be nicely interpreted as for every extra million dollars a team spends, how many more wins did they get. Higher price per win would indicate that teams are wasting their money on players that are overpaid compared to their contributions, and lower would be getting more value from the players than they are probably worth on the open market. The problem is that teams with high price per win and low price per win both win their league. Usually low-budget teams will find talented players and make a run at trying to win every few years, then their best players will leave and sign with the high-budget teams in free agency. Also, price per win tracks really closely with payroll. Figure E shows that the price per win is not associated with winning the league, but plot F shows that payroll alone is slightly associated with increased chance of winning the league. So basically, the price per win is just not an important predictor of a team winning their league.





Statistical Procedures

To model the probability of a team winning their league, Bayesian generalized linear models with a Bernoulli family and a logit link were fit using payroll, batting WAR, pitching WAR, and price per win as potential predictors and a default prior. That model was chosen because the outcome is a binary T/F response, either a team won their league or it did not.

They were compared using leave-one-out cross validation, with the model with the lowest estimated LOO information criterion being the best model possible. Payroll was kept in the model as a control variable because of concerns that it could be a potential confounder.

The models to compare are $logit(Won\ League) = Payroll + Batting\ WAR + Pitching\ WAR$ and $logit(Won\ League) = Payroll + Batting\ WAR + Pitching\ WAR + Price\ per\ Win$. Those predictors were chosen due to the exploratory data analysis in the data overview section.

Results

Under this model, the outcome (if a team wins their league or not) is assumed to have a Bernoulli distribution with a independent probability of success every year for each team. That assumption may be violated because there may be some association between if a team is successful in one year and if they were successful in the next or previous year. This could possibly be corrected by a different model choice, one that would take the correlation structure into account.

Table 1: Table of LOO information criteria for models

Model	Looic
Payroll + Batting WAR + Pitching WAR	120.9219
Payroll + Batting WAR + Pitching WAR + Price per Win	117.0889

Table 1 is the result of the LOO information criteria for both models, and shows that the best model is: $logit(Won\ League) = Payroll + Batting\ WAR + Pitching\ WAR + Price\ per\ Win$. The estimated coefficients of that model are in Table 2.

However, there are issues with fitting the model that includes the price per win term so the model that doesn't include that term makes the most sense.

Table 2: Estimated coefficients and 95% CI for model

Coefficients	Estimates	Std. Error	95% lower	95% upper
Intercept	-8.300	1.363	-11.027	-5.573
Payroll ($MM, centered)	-0.001	0.006	-0.013	0.011
Batting WAR	0.140	0.039	0.063	0.217
Pitching WAR	0.145	0.043	0.058	0.232

Discussion

Table 2 shows the 95% intervals for the coefficients in the final model. When controlling for the batting and pitching WAR, the odds of a team winning their league with a 1 million dollar increase in the budget does not change (95% interval of odds: {0.99, 1.01}). This makes sense because so much of the variability in the payrolls is accounted for within the amount of batting and pitching WAR on the team. The team's payroll is used to buy good players, and paying players that aren't performing doesn't lead to any higher chance of winning the league.

The odds that a team wins the league is nearly the exact same for both a 1 WAR increase in batting and in pitching, which indicates to me that there is no preference between emphasizing one over the other. This makes sense, because most competitive teams have a balance between batting and pitching on their rosters.

Approximately 2/3s of the way through the season, the trade deadline comes into play, after which no teams can trade their players to other teams. By that point, most teams know if they have a chance of making it to the playoffs or not, and this affects their decision making. Teams that are not going to compete start trading their good players that will leave soon for free agency in exchange for money and young players. Teams that will compete will buy these good players in a last ditch effort to improve their team. After 2/3rds of a long season, most players (barring injury) will have shown how good they are for that season, and can project that out to the end of the season and the playoffs.

Using this model, we can calculate the estimated effect on the odds of a team winning their league based on the amount of WAR that they add at the trading deadline. If a team can add a 5 WAR (total, over the season) player, their odds double if that player is a batter (95% interval: {1.37, 2.93}) or increase by 2.05 times if the player is a pitcher (95% interval: {1.35, 3.11})

I think this is the best application of this model or a model like this using current season WAR data, because it assumes that the WAR is known for each player, which wouldn't be known until after the season.

An alternative approach that could be taken to model these data is use each player on the team's WAR data for the previous year. Then the model could be used to predict the probability of a team winning the league based on the players it has, with the hope that those players produce at the level they're expected to based on the previous year and hoping no important players get injured or traded.

As it is, it's tricky to try to generalize these data to the next season's chances of winning the league, since the WAR for a player isn't constant, and players move around in the offseason each year.

Sources and Acknowledgements:

Thank you to the anonymous Redditor who did the calculations for the probability of a football team making the playoffs, and to USAtoday.com for the salary data that was not available in the Lahman package.

Reddit post about chances of making football playoffs: https://www.reddit.com/r/theydidthemath/comments/2o0zuj/requestnflhow_likely_is_it_for_a_team_to_go_106/

USAtoday baseball team salary database: https://www.usatoday.com/sports/mlb/salaries/2019/team/all/

Baseball-Reference.com WAR explained article: https://www.baseball-reference.com/about/war_explained.shtml

Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Goodrich B, Gabry J, Ali I & Brilleman S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1 https://mc-stan.org/rstanarm.

Brilleman SL, Crowther MJ, Moreno-Betancur M, Buros Novik J & Wolfe R. Joint longitudinal and time-to-event models via Stan. StanCon 2018. 10-12 Jan 2018. Pacific Grove, CA, USA. https://github.com/stan-dev/stancon_talks/

Michael Friendly, Chris Dalzell, Martin Monkman and Dennis Murphy (2021). Lahman: Sean 'Lahman' Baseball Database. R package version 8.0-1. https://CRAN.R-project.org/package=Lahman

Hadley Wickham (2020). forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.0. https://CRAN.R-project.org/package=forcats

Claus O. Wilke (2020). cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. R package version 1.1.0. https://CRAN.R-project.org/package=cowplot

Yihui Xie (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.30.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

```
knitr::opts_chunk$set(echo = TRUE, eval = FALSE, message = FALSE, warning = FALSE)
library(tidyverse)
library(rstanarm)
library(Lahman)
library(forcats)
library(cowplot)
library(knitr)

# Lahman database salaries only go through 2016!?!
# Data from https://www.usatoday.com/sports/mlb/salaries
Salaries = read_csv("PlayerSalaries.csv", col_types = "cdnc")

# WAR and WAA are "advanced metrics" aimed at measuring the number of "wins" each player adds compared
WAR_bat = read_csv("https://www.baseball-reference.com/data/war_daily_bat.txt") %>%
  filter(year_ID >= 2010 & year_ID < 2020) %>%
  mutate(battingWAR = as.numeric(WAR), battingWAA = as.numeric(WAA)) %>%
```

```r
  select(year_ID, team_ID, battingWAR, battingWAA) %>% group_by(year_ID, team_ID) %>%
  summarize_if(is.numeric, sum, na.rm = T)

WAR_pitch = read_csv("https://www.baseball-reference.com/data/war_daily_pitch.txt") %>% filter(year_ID

Teams = Lahman::Teams %>%
  filter(yearID >= 2010 & yearID != 2020) %>%
  full_join(Salaries, by = c("yearID", "franchID")) %>%
  full_join(WAR_bat, by = c("yearID" = "year_ID", "teamIDBR" = "team_ID")) %>%
  full_join(WAR_pitch, by = c("yearID" = "year_ID", "teamIDBR" = "team_ID")) %>%
  select(-c(divID,Rank,Ghome,WCWin,SB,CS,HBP,CG,SHO,SV,IPouts,HA,HRA,SOA,FP,park,attendance,BPF,PPF,tea
  mutate(yearID = forcats::as_factor(yearID),
         totalWAR = pitchingWAR + battingWAR,
         Payroll_mil = Payroll/1000000,
         Payroll_cnt_mil = (Payroll - median(Payroll))/1000000,
         # Number of wins above replacement level divided by amount of money spent
         Payroll_eff = ((W)/Payroll_mil - mean((W)/Payroll_mil))/sd((W)/Payroll_mil),
         WAE = W - (48 + totalWAR),
         # Money spent per win
         PPW = Payroll_mil/W,
         DivWin = factor(x = DivWin, labels = c(0,1)),
         WSWin = factor(x = WSWin, labels = c(0,1)),
         LgWin = factor(x = LgWin, labels = c(0,1)))

g = ggplot(data = Teams) + theme_bw() + theme(plot.title = element_text(hjust = 0.5), plot.caption = el

p1 = g + geom_density(aes(x = Payroll_mil, col = LgWin)) +
  labs(x = "Payroll ($MM)",
       y = "Density",
       title = "Density of team payrolls for league winners and non-winners",
       caption = "Team payrolls are right skewed and more right skewed for league winners",
       col = "Won League") # Payroll is right-skewed

p2 = g + geom_boxplot(aes(x = yearID, y = Payroll_mil)) +
  labs(x = "Year",
       y = "Payroll ($MM)",
       title = "Team payrolls over the years",
       caption = "Average payrolls have increased over time but max payrolls have stayed the same since
 # Payroll increases over time

p3 = g + geom_boxplot(aes(x = LgWin, y = Payroll_mil)) +
  labs(x = "Won league",
       y = "Payroll ($MM)",
       title = "Boxplot of payroll for league winners and non-winners",
       caption = "League winners tend to have slightly higher payrolls (p = 0.07)")
# League winners have slightly higher payrolls (p = 0.07)

p4 = g + geom_boxplot(aes(x = LgWin, y = PPW)) +
  labs(x = "Won League",
       y = "Amount spent per win ($MM)",
       title = "Millions of dollars spent per game won for league winners and non-winners",
       caption = "There is no difference in amount of money spent per win for league winners vs non-win
  # Basically no difference in price per win for league winners and non-winners
```

```r
p5 = g + geom_boxplot(aes(x = LgWin, y = totalWAR)) +
  labs(x = "Won League",
       y = "Team total WAR",
       title = "Team total WAR for league winners and non-winners",
       caption = "Teams that won their league have a much higher average WAR (p < 0.001)")
# Teams that won the league have a much higher total WAR (p < 0.0001)

p6 = g + geom_point(aes(x = Payroll_mil, y = totalWAR, col = LgWin)) +
  geom_smooth(aes(x = Payroll_mil, y = totalWAR, col = LgWin), method = "lm", formula = y~x, se = F) +
  labs(x = "Payroll ($MM)",
       y = "Team total WAR",
       col = "Won League",
       title = "Plot of Payroll and WAR for league winners and non-winners",
       caption = "Teams with a higher payroll have a higher total WAR on average (corr = 0.3) and highe
# Teams with higher payroll have higher WAR on average, and higher WAR is associated with winning the l

plot_grid(ncol = 1, p1, p2, labels = c("A", "B"))

plot_grid(p5, labels = "C")
plot_grid(p6, labels = "D")

plot_grid(p4, labels = c("E"))
plot_grid(p3, labels = "F")

# Things you can control about your team: how much you spend, what you focus on, how good you are at fi
# Things you can't control: how lucky your team is, how your players play over a short term
mod = stan_glm(data = Teams, family = binomial(link = "logit"),
               formula = LgWin ~ Payroll_cnt_mil + battingWAR + pitchingWAR, refresh = 0)
mod2 = stan_glm(data = Teams, family = binomial(link = "logit"),
               formula = LgWin ~ Payroll_cnt_mil + PPW + battingWAR + pitchingWAR,
               refresh = 0)

x = data.frame(`Model` = c("Payroll + Batting WAR + Pitching WAR",
                           "Payroll + Batting WAR + Pitching WAR + Price per Win"),
               `Looic` = c(loo(mod)$estimates[3,1], loo(mod2)$estimates[3,1]))
kable(x, caption = "Table of LOO information criteria for models")

x2 = data.frame(Coefficients = c("Intercept", "Payroll ($MM, centered)","Batting WAR", "Pitching WAR"),
                Estimates = mod$coefficients,
                `StdErr` = mod$ses,
                `95%lower` = mod$coefficients + -2*mod$ses,
                `95%upper` = mod$coefficients + 2*mod$ses)
kable(x2,
      digits = 3,
      row.names = F,
      caption = "Estimated coefficients and 95% CI for model",
      col.names = c("Coefficients", "Estimates", "Std. Error", "95% lower", "95% upper"))
```