**Nearly 20 Years After Moneyball: An Analysis**

Abstract:

*Moneyball: The Art of Winning an Unfair Game* by Michael Lewis was published in 2003. The book detailed the Oakland Athletics’ strategy for winning an inordinate number of MLB games in the early 2000s. This study is an investigation *Moneyball*’s effect on Major League Baseball and seeks to evaluate the role of money and Moneyball-isms in professional baseball using 21 years of hitting, pitching, and payroll data from 1999-2019. While side-by-side boxplots for Walks and OBP showed, if anything, a downward trend contrary to Moneyball thought, a possible confounding factor was pitching – specifically strikeouts which have dramatically increased over time. I found a significant decrease in the variance of adjusted payroll per run suggesting teams pay for similar run creating metrics in recent years in contrast to less rigorous traditional metrics. Next, with multiple linear regression I found the best predictor of wins was runs scored and runs allowed with an adjusted r-squared of .8854. Finally, logistic regression analysis showed playoff berths could be modeled strongly and payroll does not clearly affect playoff potential except in the extremes of high and low payroll relative to the league.

Introduction:

In 1977 Bill James wrote his first in a series of baseball abstracts; short books containing essays criticizing the paradigm of baseball statistics and his own exotic, diligently collected baseball statistics. Many intellectuals have studied baseball and quietly criticized its insistence on batting average and pitcher wins among other traditional statistics, but Bill James popularized it. However, his work did not influence Major League Baseball until A’s owner and baseball philanthropist Walter A. Haas, Jr died and the team was sold to businessmen who specifically set out to profit from the venture. A’s general manager Sandy Alderson having read Bill James’ baseball abstracts sought to improve efficiency by installing three rules: “1) Every batter needs to behave like a leadoff man and adopt as his main goal getting on base. 2) Every batter should also possess the power to hit home runs, in part because home run power forced opposing pitchers to pitch more cautiously, and led to walks, and high on base percentages. 3) To anyone with the natural gifts to become a professional baseball player, hitting was less a physical than a mental skill. Or, at any rate, the aspects of hitting that could be taught were mental” (Lewis, 59). Conventional baseball wisdom derided walks preferring hits meaning Sandy Alderson and later Billy Beane could acquire players who walked more than usual for cheaper than players who hit more than usual even though a single and a walk produced the same outcome. Nearly 20 years after Moneyball’s publication, have more teams adopted the Moneyball model or are there other factors at play?

Team data (called TeamStats in R code) is a csv of 630 observations (one for each of the 30 teams and 21 seasons from 1999 to 2019) and 89 variables collected from baseball-reference.com and thebaseballcube.com (for payroll statistics). In the following analysis I used 25 of the 89 variables collected since many are derivative or luck dependent.

Definitions and Terms:

In the R-code, “.x” generally refers to offensive statistics and “.y” generally refers to pitching statistics. OBP.x; Offensive On-Base Percentage. WLP; Win-Loss Percentage. Adjusted Payroll = Team Payroll / MLB Payroll. Stealing Rate = Stolen Bases / (Stolen Bases + Caught Stealing). SLG; Slugging Percentage (Average number of bases gained per at bat). PA; Plate Appearances (The number of times batters face pitchers). BB.x; Offensive Walks. SB; Stolen Bases. SO.x; Offensive Strikeouts. W; Team wins. R.y; Runs allowed. H.x; Offensive Hits. GDP; Offensive number of double plays grounded into. HBP.x; Offensive batters hit by pitch. LOB.x; Offensive runners left on base when an inning ends. SecA; Offensive Secondary Average [(Total Bases – Hits + Walks + Stolen Bases – Caught Stealing)/(At Bats)]. AIR; Offensive (Measures the offensive level of the leagues and parks the team played in relative to an all-time average of .335 OBP and .400 SLG). BAbip (Measures how effectively the defense turned balls-in-play into outs). HR.y; Home runs allowed. BF; Batters faced by pitchers (pitching statistic). WHIP; (Pitching) walks and hits per inning pitched. LOB.y; Defensive – runners left on base when an inning ends. BB9; Walks allowed per 9 innings. SO.y; Pitching strikeouts. Playoffs; binary – 1 indicates the team made the playoffs

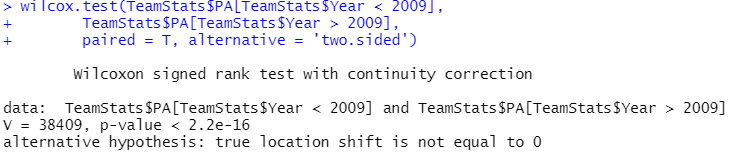
Methods:

I start with a hypothesis test for a possible change in plate appearances over time since comparing counting statistics such as walks and strikeouts by season would be inappropriate if there were a different number of plate appearances per season. Next, we see walks per PA, on-base percentage, and strikeouts per PA through time using side-by-side boxplots and when there is a clear trend, use a hypothesis test to confirm it. In simple linear regression I analyzed adjusted payroll vs wins and runs uncovering small differences over time. With multiple linear regression I estimate wins with runs scored and runs allowed then use multiple linear regression on runs scored and runs allowed separately. Finally, I use logistic regression to create a model for playoff berths and find what role payroll plays in playoff potential.

Analysis:

One thing I must be particularly careful of in the analysis is counting statistics. Examples include home runs, hits, and strikeouts. These statistics are counted from zero and reset at the end of the regular season. Thus, a change in plate appearances (the number of times batters face opposing pitchers) could result in inappropriate changes in counting statistics. As such, an analysis of the change in walks over time must use a statistic for walks per plate appearance to correct for this possibility. As it happens, the following wilcoxon test shows significant evidence that the amount of plate appearances in baseball has increased before and after 2009, the median year in this dataset.

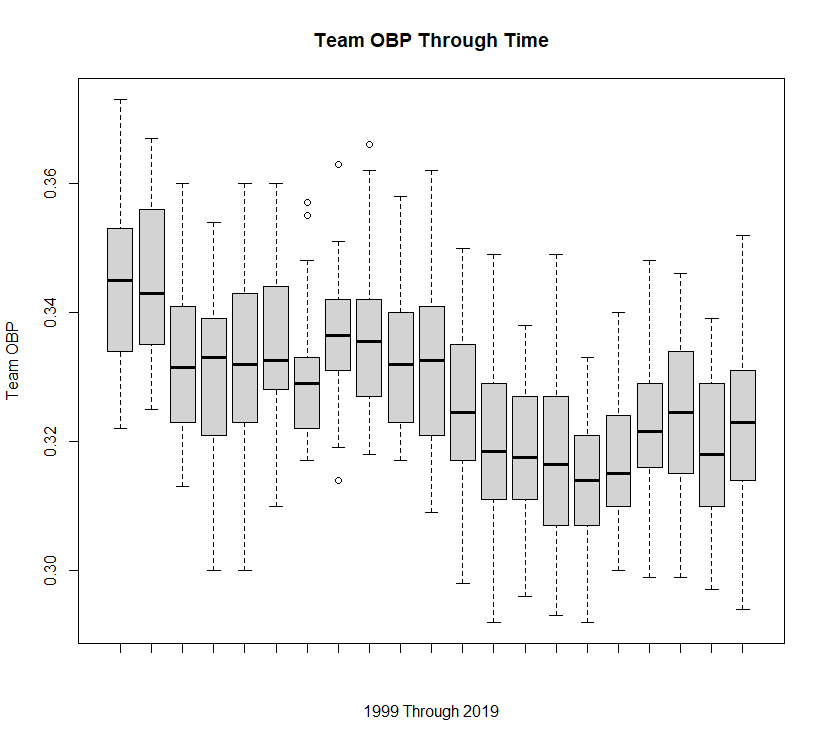
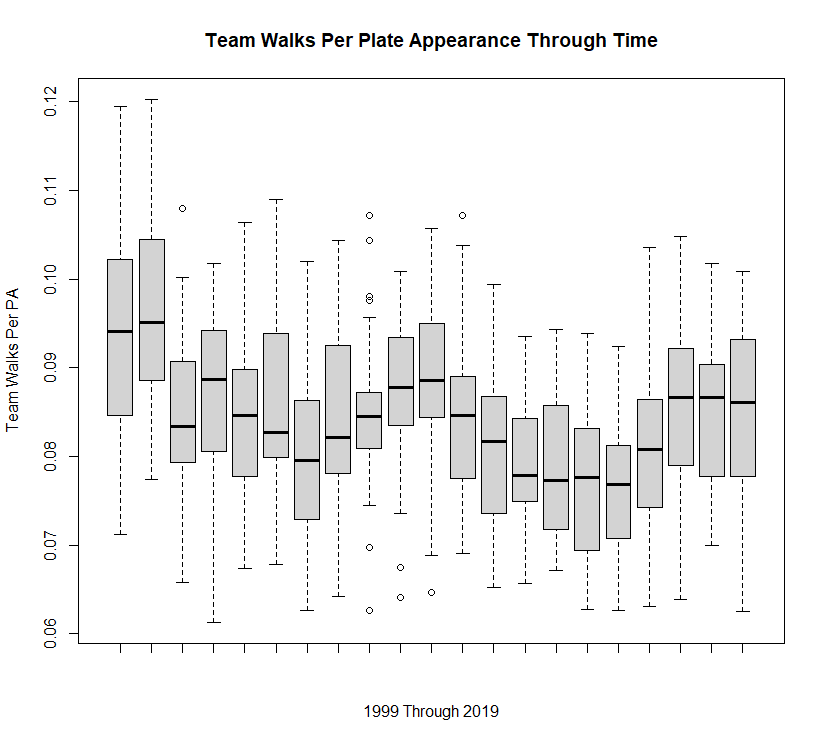
I.



With a p value less than 2.2x10^-16, we reject the null hypothesis and conclude that there is significant evidence that there were more plate appearances after 2009 than before (With the number of seasons being equal to ten in each cohort)

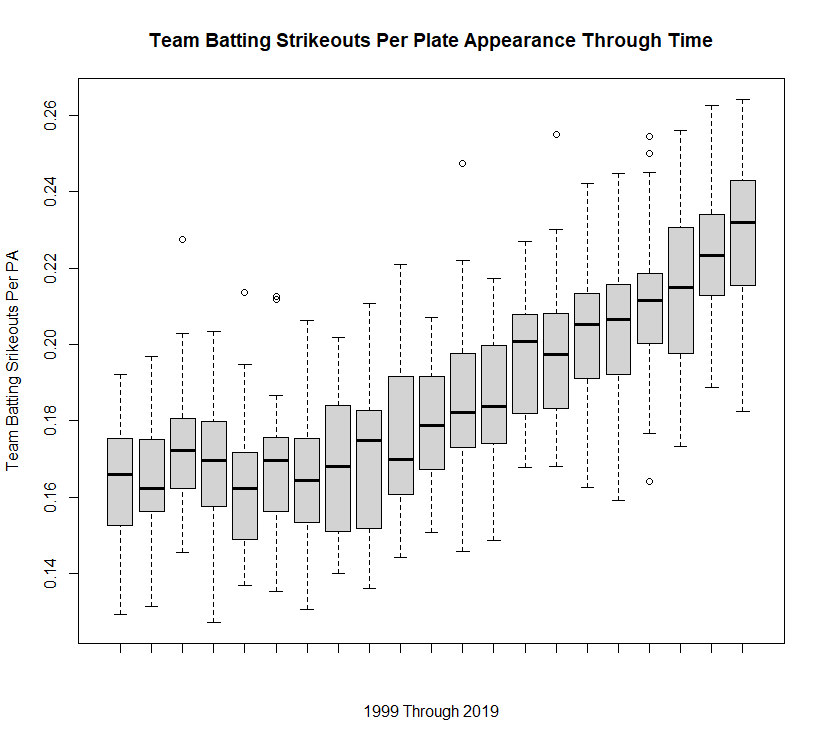
As such, it would be inappropriate to select a counting statistic when comparing different seasons.

II.

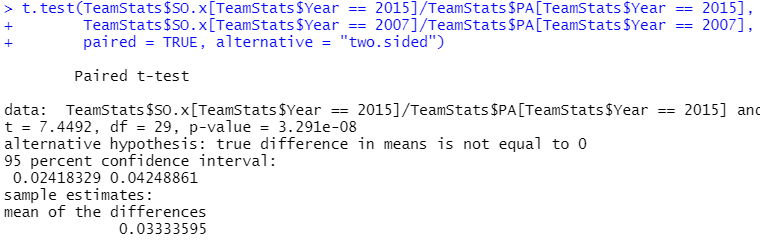


These plots contradict my hypothesis that walks and on-base percentage would increase at least immediately after Moneyball’s mid-season 2003 publication. Walks and On-Base Percentage to some degree depend on pitching, so it may be fruitful to check pitching statistics to see if we are missing something on the pitching side.

III.

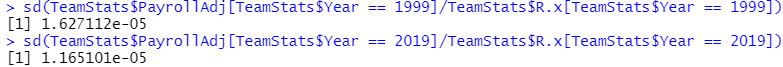


The strikeout plot may show some insight as it has a clear trend. *The MVP Machine: How Baseball's New Nonconformists Are Using Data to Build Better Players* by Ben Lindbergh and Travis Sawchik, a 2019 book focused heavily on the role of new technology on pitching and tracked the early career of starting pitcher Trevor Bauer mentioned a newly commercially available global shutter camera which could capture a baseball at a pitcher’s release point in slow motion with no distortion, perfect for analysis and tweaking of breaking pitches. A hypothesis for future testing might be that it is easier to teach effective pitching to a pitcher than plate discipline to a batter.

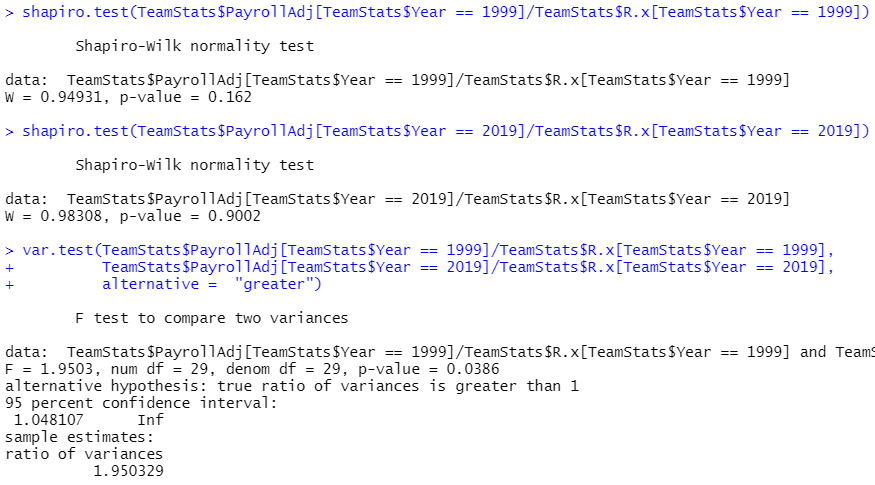


Here I ran a paired t test for the difference in batting strikeouts per batting plate appearance, the same metric as the previous side-by-side boxplot plot. The test shows that there is significant evidence that there is a difference in strikeouts per plate appearance in 2015 as opposed to 2007. Further, the 95% confidence interval for the change is (2.418%, 4.248%) since both sides of the confidence interval are positive, the evidence shows a significant increase in the strikeout rate.

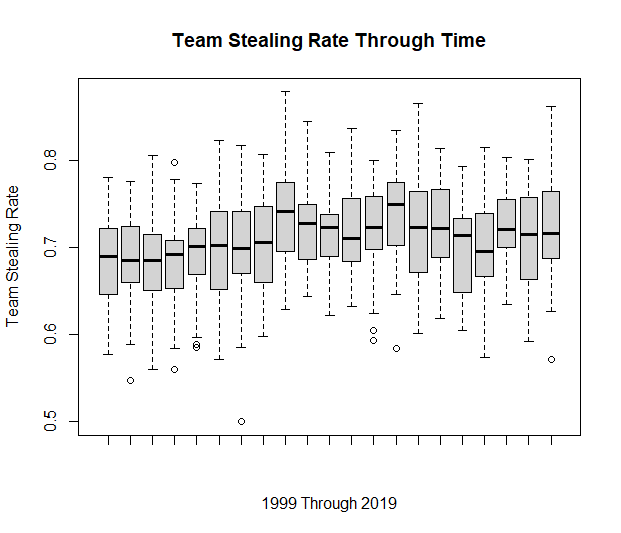
IV.



The standard deviation of the proportion of total payroll in all of baseball that year each team uses per run decreased from 1.627x10^-5 to 1.165x10^-5 from 1999 to 2019 perhaps suggesting more teams are paying for run-producing statistics over traditional metrics. The Shapiro-Wilk test below might suggest that cost per run is better modelled with a normal distribution than it was 21 years prior and the F-Test shows a significant decrease in the 2019 distribution variance compared to the 1999 distribution.



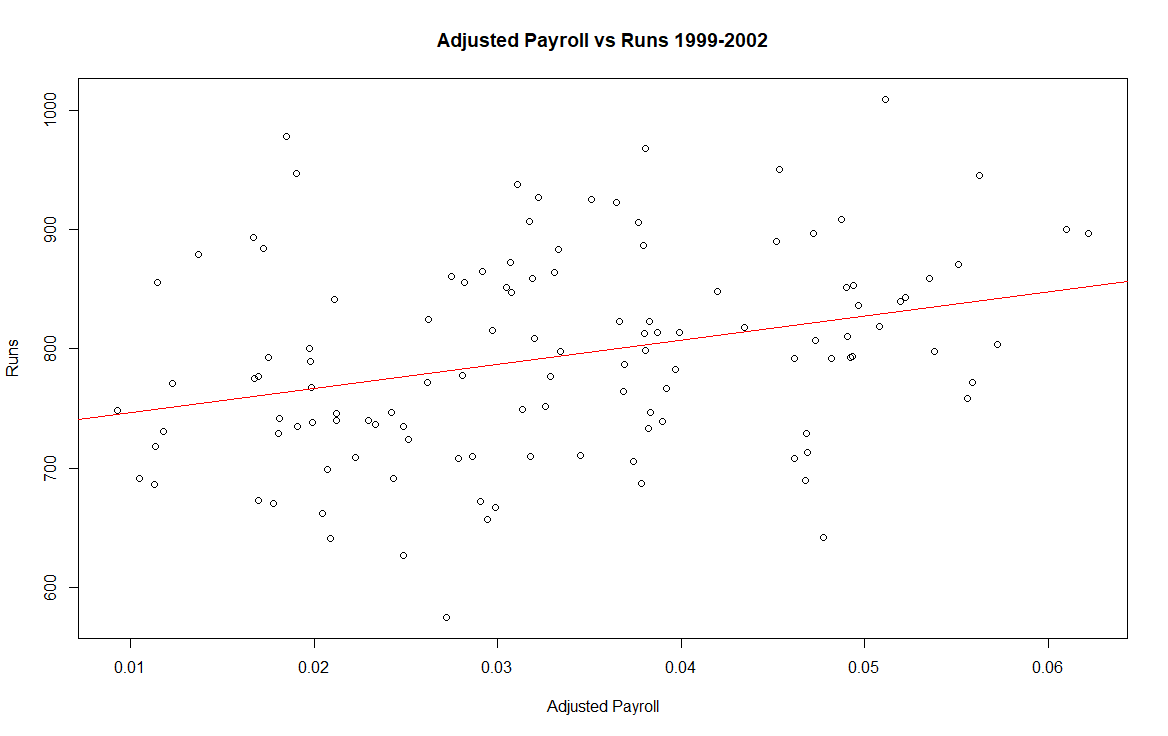
V.



Another Moneyball-ism is that stealing bases is unfavorable if the runner is successful less than about 70% of the time (Lewis, 129). As the chart shows, the worst teams in terms of this statistic abruptly increased their stealing rate in 2006/2007 from around .58 to around .63 and mostly stayed elevated. The median however has remained stable over time. Limitation: this chart also reports steals of third and home which likely have a different percentage of success required to have a positive expected value.

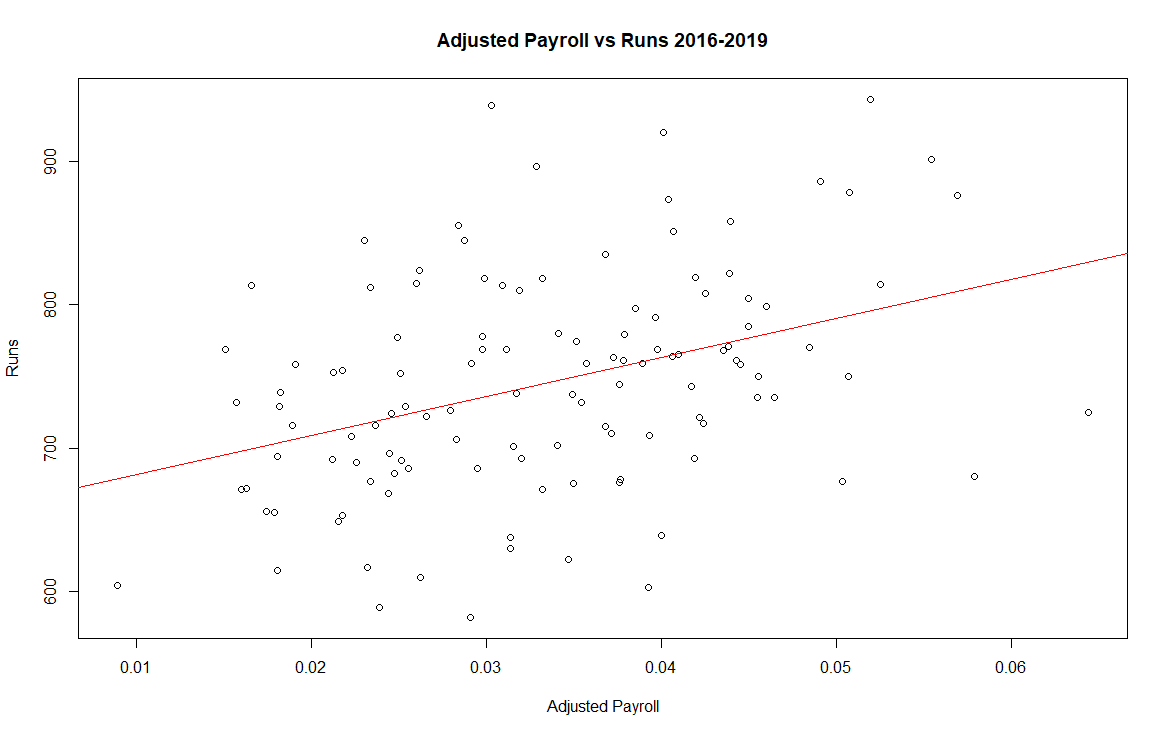
VI.

Simple Linear Regression

cor = .3104

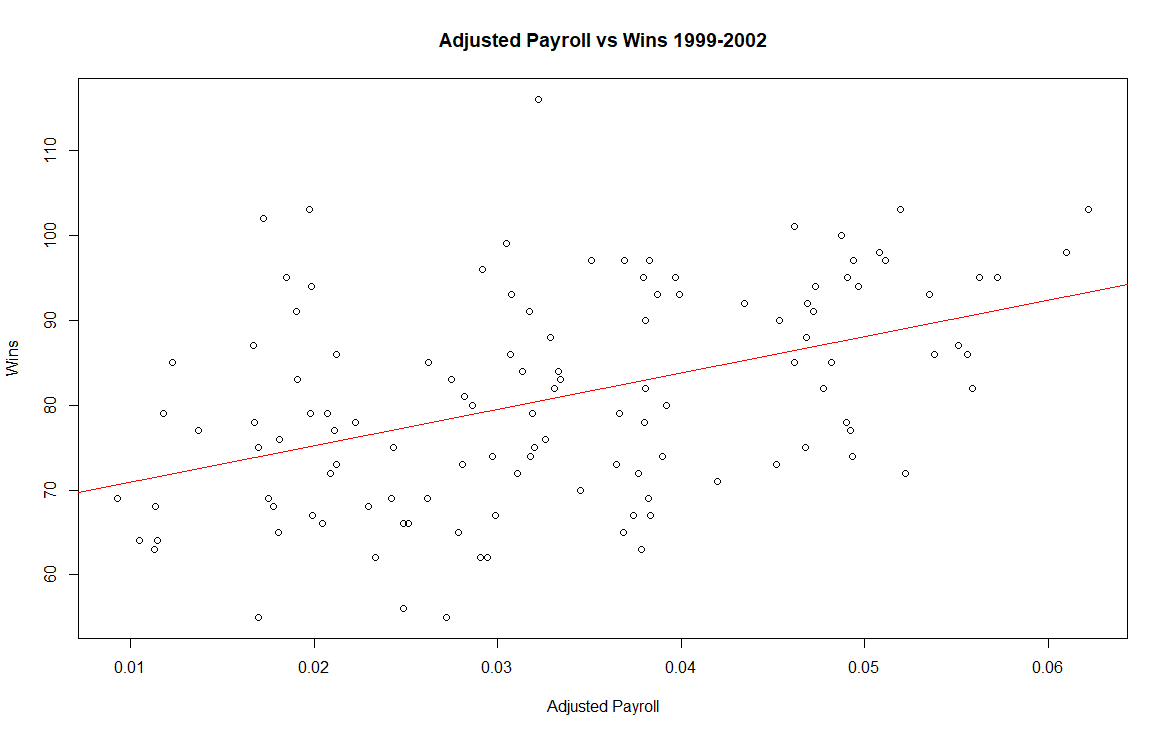
***Runs = .00004758 (Adjusted Payroll) - .004441***

For each additional .00004758 in adjusted payroll a team spends, the team would expect to get one more run per season with a coefficient of determination of .0963

cor = .3850

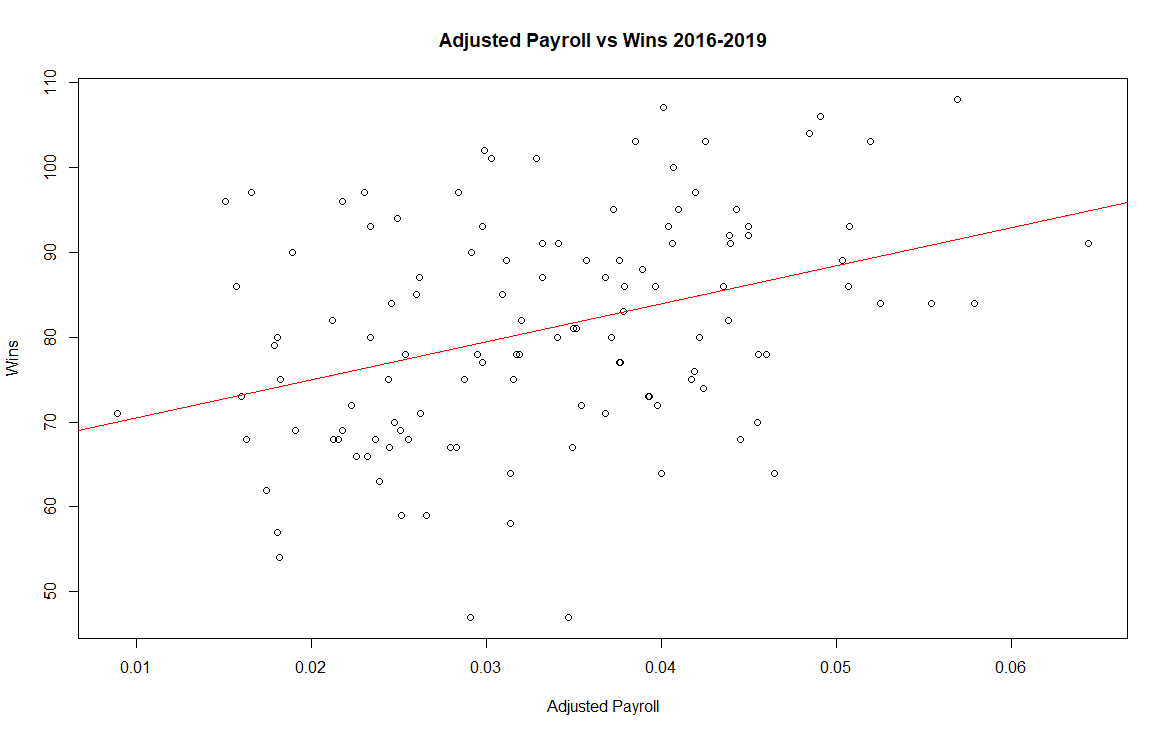
***Runs = .00005445 (Adjusted Payroll) - .007246***

For each additional .00005445 in adjusted payroll a team spends the team would expect to get one more run per season with a coefficient of determination of .1482

cor = .4465

***Wins = .0004647 (Adjusted Payroll) - .0042572***

For each additional .0004647 in adjusted payroll a team spends the team would expect an additional win with .1353 coefficient of determination

cor = .3679

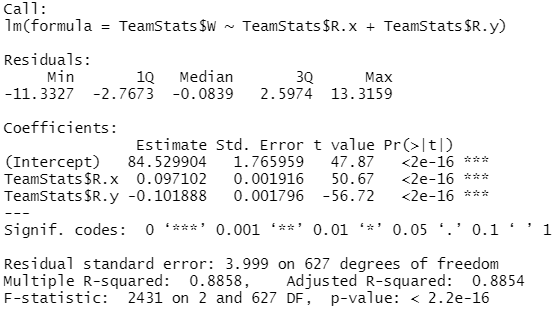
***Wins = .0003017 (Adjusted Payroll) - .008906***

For each additional .0003017 in adjusted payroll a team spends the team would expect an additional win with .1993 coefficient of determination

Interestingly, the correlation between adjusted payroll and runs increased in the 2016-2019 cohort but the correlation decreased between adjusted payroll and wins. I might argue this is evidence that teams are willing to pay for runs but spending is no longer the best way to acquire talent. The Athletics under Billy Beane won games by out-drafting and out-trading since they could not afford to out-spend. Currently, *The MVP Machine* argues, out-developing is best way to acquire talent.

VII.

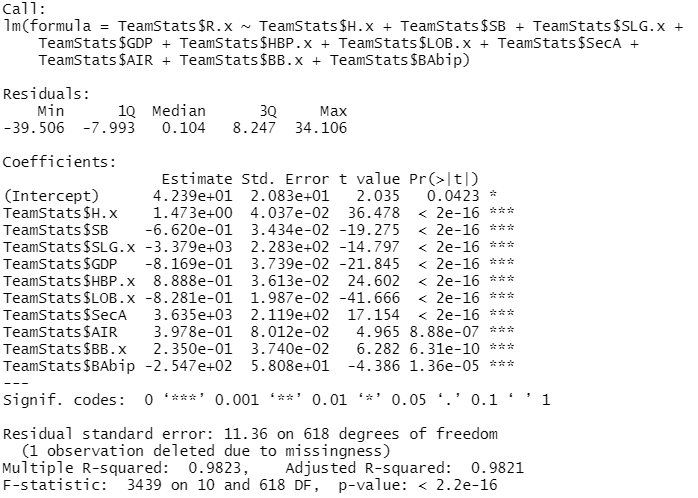
**Multiple Linear Regression**



***Wins = 84.529904 + .097102(Offensive Runs) - .101888(Runs Allowed)***

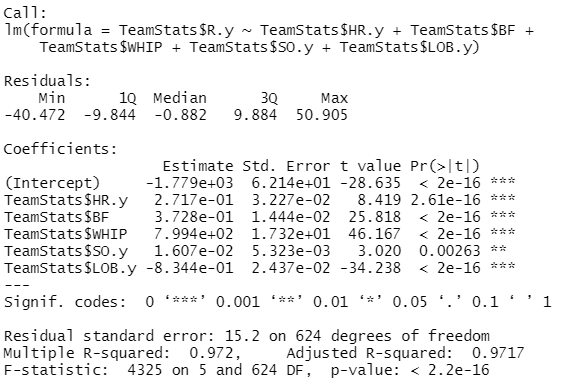
Adjusted R-Squared = .8854

What do models for offensive runs and runs allowed look like?



***Offensive Runs = 42.39 + 1.473(Hits) - .662(Stolen Bases) - 3379(SLG%) – .8169(Double Plays Grounded into) + .8888(Hit By Pitch) – .8281 (Left on Base) + 3635(Secondary Average) + .3978(Hitting AIR) + .235 (Walks) – 254.7(Batting Average on Balls in Play)***

Adjusted R-Squared = .9821

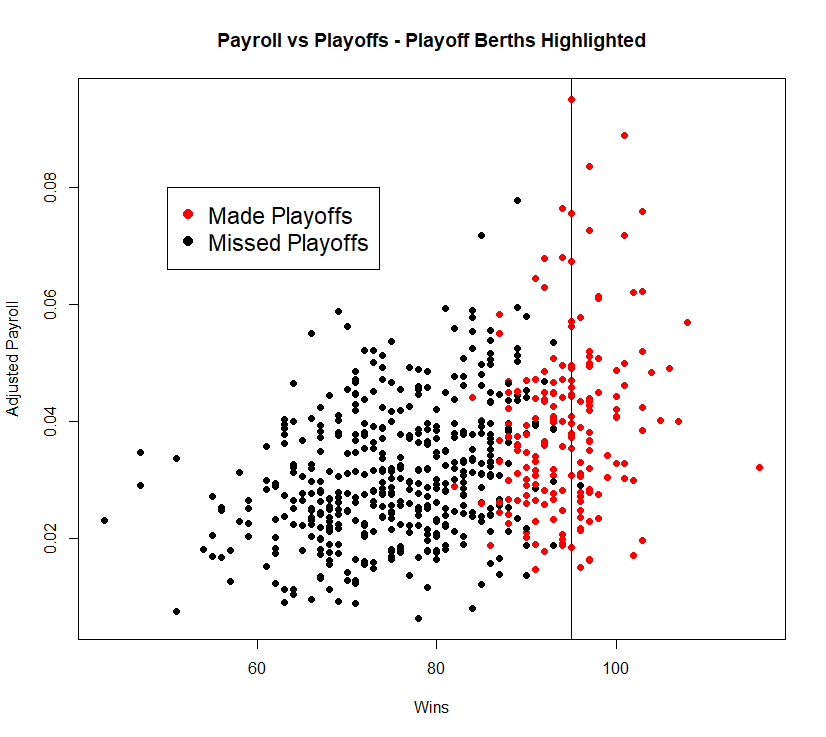


***= -1779 + .2717(Home Runs Allowed) + .3728(Batters Faced) + 799.4(Walks and Hits Per Inning Pitched) + .01607(Pitching Strikeouts) - .8344(Left on Base)***

Adjusted R-Squared = .9717

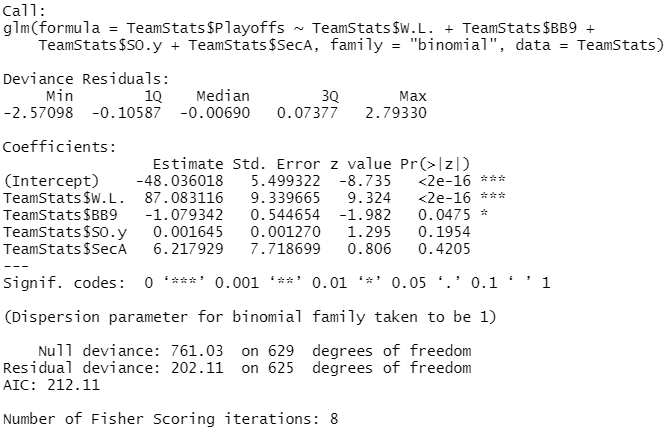
VIII.

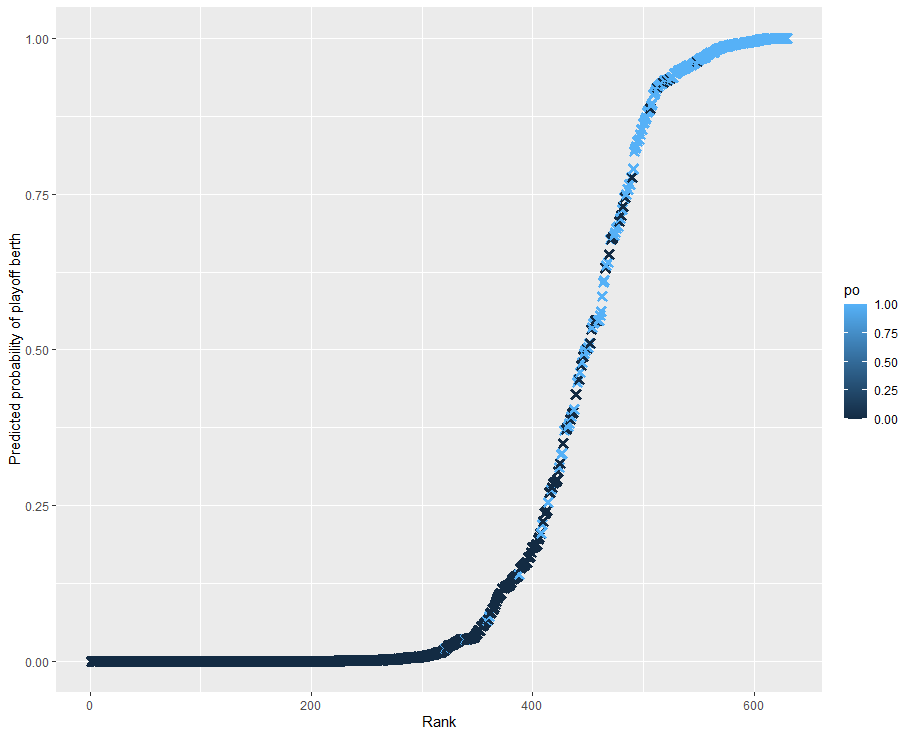
**Logistic Regression**

****

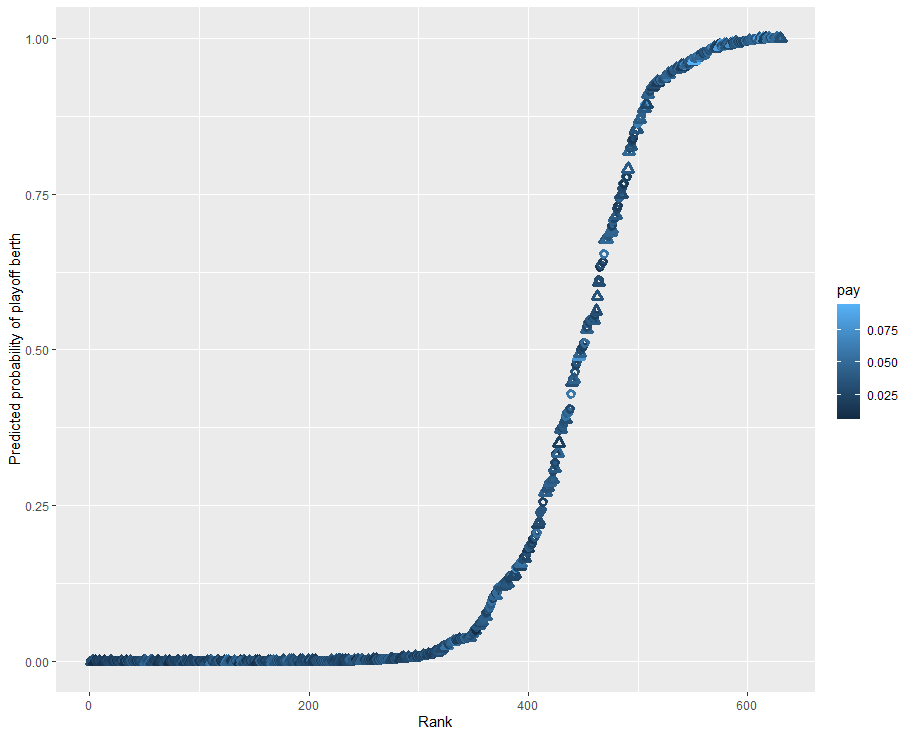
The Athletics aimed for 95 wins to secure a playoff berth.

Let’s use logistic regression to find an equation for making the playoffs.

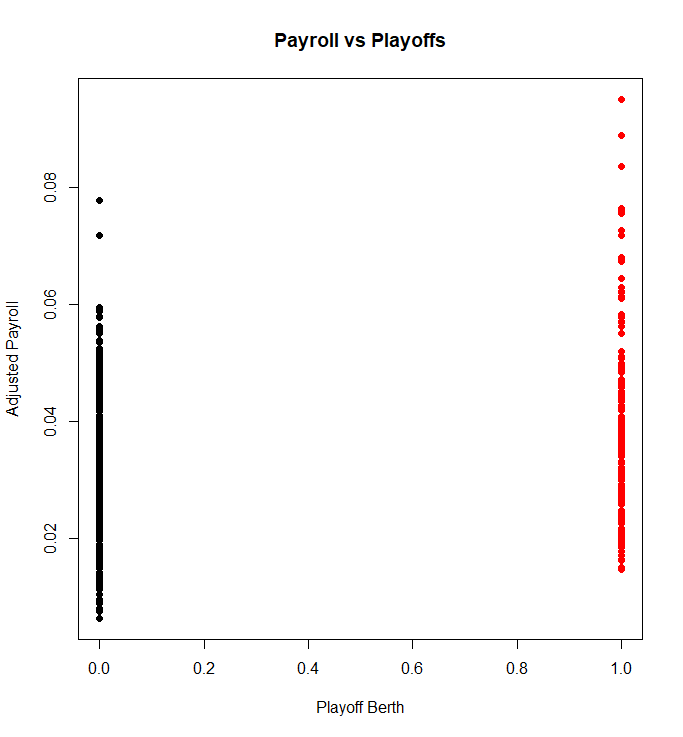
 ***= 87.083116(WLP) – 1.079342(Walks Allowed per 9 Innings) + .001645(Pitching Strikeouts) + 6.217929 (Secondary Average)***



All 630 observations ranked by predicted probability of reaching the playoffs colored by the observed playoff status. Most of the dark points are on a low level of predicted probability of playoff berth and most of the blue points are on a high level of predicted probability of a playoff berth indicated that the model is reasonably predictive.



On this version of the plot, triangles reached the playoffs and circles did not. Lighter shapes indicate larger adjusted payroll spent by the team. There appears to be a mix of light and dark shapes on all levels of the model indicating that payroll may not be a significant predictor of playoff potential.



This plot shows the relationship between Playoff Berths and Payroll more clearly. There are 19 teams in 21 years that spent more than 6% of the total MLB Payroll on their team. Only two of them missed the playoffs. Similarly, 27 teams in the 21-year sample spent less than 1.5% of the total MLB payroll on their team. Only one of them made the playoffs.

Concluding observations:

This study began as an analysis of Moneyball principles—uncovering a possible shift in baseball from theory to practice as Jamesian thought became mainstream. However, immediately in section II, my expectation was contradicted with walks per plate appearance and on-base percentage falling, if anything, in the seasons following the midseason 2003 publication of Michael Lewis’ *Moneyball*.

After seeing the plot in section III and wanting to understand the trend, I picked up *The MVP Machine* and discovered a wave of technological and psychological innovation on the pitching side possibly contributing to the sky-high strikeout rate and the decreased walk rate.

Section IV explores the relationship between adjusted payroll per run in 1999 and 2019 discovering that the variance of the recent distribution has significantly decreased compared to the 1999 adjusted payroll per run distribution suggesting that more teams are focused on Jamesian run-generating statistics over traditional statistics which bare a weaker relationship to runs scored.

In *Moneyball,* Oakland’s front office was aware of the risk-benefit of base stealing. “[B]roadly speaking, an attempted steal had to succeed about 70 percent of the time before it contributed positively to run totals.” (Lewis, 129) More narrowly speaking, the risk-benefit changes according to the inning situation i.e., the number of outs, another baserunner, the following batters, etc. Given that, we would expect teams to change their behavior and we do to some degree among the lower bound and the upper bound, but the median remains similar and often below the 70 percent threshold post-*Moneyball*

The simple linear regression analyses show positive relationships between adjusted payroll and both wins and runs, however the correlation coefficient differs when comparing 1999-2002 and 2016-2019 though to a nonsignificant degree.

In the multiple linear regression models, I try to answer the question of what determines a team’s success. Measuring success in terms of wins unsurprisingly reveals that runs scored and runs allowed are the strongest factors similar to Bill James’ rudimentary “Pythagorean theorem of baseball’ which is WLP = (Runs scored)^2 / [(Runs scored)^2 + (Runs Allowed)^2].

The strangest model was section VIII, logistic regression. In the model of course WLP dominates as WLP compared to other teams is how a team makes the playoffs. The strange part is nearly nothing else can significantly predict a playoff berth but walks allowed. In the first regression plot, most of the teams that did not make the playoffs indeed plotted near the bottom left and most of the teams that did make the playoffs plotted near the top right indicating the strength of the regression. The second regression plot shows a different story since the color of the shape indicates the team’s adjusted payroll. Contrary to my hypothesis, there are a variety of hues all over the chart, particularly not including the extreme high and low budget teams. The final plot, a simple binary scatter plot, shows the differences at the extremes more clearly. In conclusion, payroll is not a significant indicator of playoff potential unless the team payroll is below around 1.5% of the total MLB payroll or above around 6% of the total MLB payroll.

References

Law, K. (2017). *Smart baseball: The story behind the old stats that are ruining the game, the new ones that are ruining it, and the right way to think about baseball*. New York, NY: William Morrow, an imprint of HarperCollins.

Lindbergh, B., & Sawchik, T. (2019). *MVP machine: How baseball's New nonconformists are using data to build better players*. New York: Basic Books.

Lewis, M. (2003). *Moneyball: The art of winning an unfair game*. New York: W.W. Norton.

The baseball cube - research site for pro + college stats ... (n.d.). Retrieved April 27, 2021, from http://thebaseballcube.com/

MLB stats, Scores, History, & Records. (n.d.). Retrieved April 27, 2021, from https://www.baseball-reference.com/