Economic Re-evaluation of the Moneyball Hypothesis

Master of Applied Economics

University of California, Los Angeles

Author: David Contento

Adviser: Dr. Randall Rojas

Table of Contents

1. Abstract

At the end of the 2001 regular season the Oakland Athletics were coming off a record-breaking season with 102 wins to 60 loses. Expectations were high in Oakland as they entered the 2001-2002 offseason. Unfortunately, the Athletics had lost three key free agents to larger market teams that offseason, and where forced to find suitable replacement. This would prove to be difficult especially when we consider the fact that Oakland possessed the fourth smallest payroll cap among all Major League teams at the time. Regardless, general manager Billy Beane was able to make several under-the-radar free agent signings that would help Oakland recover from its lack of star power. The team would go on to finish the 2002 season with a record of 103 wins to 59 loses breaking the previous seasons record. The team would also go on to win 20 consecutive games that season, resulting in an American League record that would last for over 15 years.

Oakland was able to achieve its repeated success by exploiting inefficiencies in the baseball labor market. The Athletics took advantage of a more analytical and evidenced-based approach to player evaluation, which resulted in the signing of several key undervalued players. As a result of this new approach Oakland was able to remain competitive despite losing a large amount of talent. This new approach involved looking at different indicators for offensive and defensive player success. This new system of leveraging player information and statistics would eventually spread to all other Major league teams issuing in the era which now referred to as the "Moneyball era".

This leads us to the main purpose of this paper, which is to identify whether baseball labor markets are still inefficient and determine which player statistics are currently overvalued or undervalued. In the paper we will look at player salary and statistics ranging back to 1985. We will also examine other unique questions such as how a team's winning percentage affects game attendance. We will also look at how a team's payroll size contributes to its winning percentage, as well as how these relationships change over time. Lastly, we will also create a model that accurately determines a player's salary based on their offensive statistics, and we will use this model to learn which players maybe currently over or undervalued.

2. Introduction

Coming into the 2002 season, no one expected Oakland to experience the amount of on field success that it achieved. So how exactly was Oakland able to compete against larger market teams after losing several key players, while simultaneously operating on a shoe-string budget? The answer can be found in general manager Billy Beane's exploitation of market inefficiencies, statistics, and econometrics. Billy Beane, along with economist Paul DePodesta, were able to identify and exploit labor market inefficiencies which were the byproduct of a subjective, flawed, and biased view of player worth. The team was able to leverage sabermetrics and took a more analytical approach to evaluating player performance instead of relying on the anecdotal advice of it scouts. Oakland argued that On-Base-Percentage and Slugging are better indicators of player performance than historically valued qualities like form or contact (Lewis, 2003). Oakland was able to acquire these new qualities cheaply on the open market and were therefore able to compete against teams that possessed payrolls three times larger the size of their own.

We would like to think that baseball is a fair game where all teams are made equal, but the reality is that modern baseball is plagued with inequality at multiple levels. Teams with larger markets and greater financial resources have a distinct advantage over others. Despite efforts by the MLB committee to even the playing field, there is still a tangible difference among team payrolls. Smaller teams are forced to think more critically about who they sign and how much they are willing to pay. That is why in this paper we will attempt to discern just how much a player's performance affects their salary, how the value of different player statistics have changed over time, and how much the market value of a player should be. It is also important to note that it has been over 15 years since the baseball world was introduced to the concept of "Moneyball". At this point in time almost all teams have adopted the idea of sabermetrics to some degree, and it stands to reason that baseballs labor market should no longer exhibit the "Moneyball" mispricing phenomenon. We will start by looking at player salary information and performance statistics going back to 1985. We will also examine how labor markets have adjusted over time to evaluating player performance.

3. Data

The datasets used in this project were gathered from two main sources. The first is Sean Lahmans Database for baseball statistics, which is a repository of baseball statistics gathered by the Sabermetrics community and compiled neatly into one website by baseball enthusiast, analyst, and sportswriter Sean Lahman. The dataset consists of several tables containing different player statistics such as batting, pitching, and salary. As well as other tables that contain the same statistics but for different time periods such as the regular season, post-season, and all-star game. The dataset goes back to the 1877 season and was used to calculate player batting statistics as well as salary statistics (Lahman, 2018).

The other source for data is baseball-reference.com, which is a statistics tracking website owned and operated by Sports Reference LLC. This website contains historical information dating as far back as the 1870s. This data set was used to gather information on overall team performance, such as winning percentage and offensive statistics in order to create time-series models and construct causality tests.

4. Methodology

4.1 Labor Market Efficiency

In Major League Baseball there are nine innings, in which each team has a chance to score runs on offense while the opposing team must prevent the other team from scoring by tagging players out or striking them out. In terms of offense, we can boil down the sole purpose of a player to scoring runs. This is accomplished through a combination of different skills, specifically the skill of hitting the ball and the ability to avoid getting tagged out. Up until the invention of sabermetrics, a player's offensive capabilities were evaluated on archaic qualities such as speed, contact, and batting form. Although these qualities may provide some insights on player performance, they are outdated relics of a 19th century view of the game. Up until 2002 teams relied heavily on their scouts to find and assess players. It wasn't until after the release of the book "Moneyball" by Michael Lewis, which focused on the modern biases imbedded in baseball and the benefits of an evidenced based approach, that MLB teams placed a heavy emphasis on an analytical approach (Lewis, 2003). It was Billie Beane and Paul DePodesta who argued that on-base-percentage and slugging percentage where the best

indicators for offensive success. Later, members of the Society for American Baseball Research (SABR) performed a study and determined that linear combinations of on-base-percentage and slugging percentage were very highly correlated with runs scored. It stands to reason that if these two statistics are highly correlated with runs scored, they should also help explain the variation in a player's salary, which I would argue is dependent on their ability to score runs. As a result, I will focus primarily on these two indicators in my analysis. We will examine how these statistics currently affect player's salary and how their influence has changed over time.

The dataset used for this section goes back to the 1890's, but unfortunately we only have data on player salaries for the 1985 season and beyond (Lahman, 2018). This diminishes the robustness of our model, but still provides enough observations to gain statistically significant results.

I would like to start off by looking at the distribution of salaries for all players after adjusting for inflation. I adjusted for inflation by taking the CPI index and multiplying each players salary by the appropriate inflation value.

Table 1: Salary Summary Statistics

	2016-2015	2014-2013	2012-2011	2010-2009	2008-2007
Mean	5869052	5590326	5222004	4996586	5044667
10th Percentile	537758	533045	527175.7	472663	466364
Median	3178563	2917164	2694809	2993774	2799712
90th Percentile	15892814	15174372	14129599	13647420	14539799

(Lahman, 2013)

We can see that since 2007 players salaries have been consistently growing across the board, which makes sense since modern players, both rookies and veterans, are consistently demanding greater contracts every season. If we look at the histogram of the distribution of salaries it does not appear to be normally distributed. In fact, there are several high-end outliers, which may skew our regression results. As a result, I ran a Box-Cox transformation test on the dependent variable, which determined that a log transformation was required to make salaries more normally distributed and lower the variance of our regressions.

I decided to regress player salaries on on-base-percentage (OBP) and slugging percentage (SLUG) along with other predictors. This includes a variable that denotes playing time, which in this

case is measured by plate appearances (AB), and dummy variables that control for a player's position (Catcher, Pitcher, Outfielder). In order to avoid the dummy variable trap an indicator for the infielder position was not included. It is also important to note that we ran a fixed effects regression which includes a factor variable that controls for fixed time effects. All of the factor variables were significant and allow us to control for effects likes inflations, and differences in labor market returns across seasons. I also lagged player statistics by one period so that their salary was determined by their performance in the previous season in the regression. The coefficients for these factor variables where suppressed in Figure 1 below.

Figure 1: Salary Regression

Dependent Variable: log(salary)			
Predictors	*Estimates*	*std. Error*	*p*
(Intercept)	11.13	0.08	**<0.001**
OBP	3.35	0.26	**<0.001**
Slugging	1.74	0.15	**<0.001**
AB	0.003	0.00	**<0.001**
Catcher	-0.12	0.03	**<0.001**
Pitcher	1.83	0.04	**<0.001**
Outfielder	-0.04	0.02	0.050
Observations	15364		
R^2 / adjusted R^2	0.336 / 0.334		

The regression above was generated using the entire dataset, and all the variables were significant at the 5% critical level. In order to make sure that our coefficients were accurate I also decided to bootstrap them. I ran 1000 random sampling bootstrap iterations and found that the coefficients were exactly the same, which confirms that the model is robust. I have included the output from the bootstrap function in the appendix. I would first like to note that both on-base-percentage and slugging are extremely significant and positively correlated with salary, which makes economic sense. I would also like to note that the effect on-base-percentage has on salaries is twice more than slugging. This goes against one of the assumptions made in the book Moneyball by Michael Lewis, who claims that on-base-percentage is undervalued relative to slugging even though it contributes more to team success. However, it is important to note that our results also cover the post Moneyball period, which is considered 2002 and beyond, and may thus be reflecting an

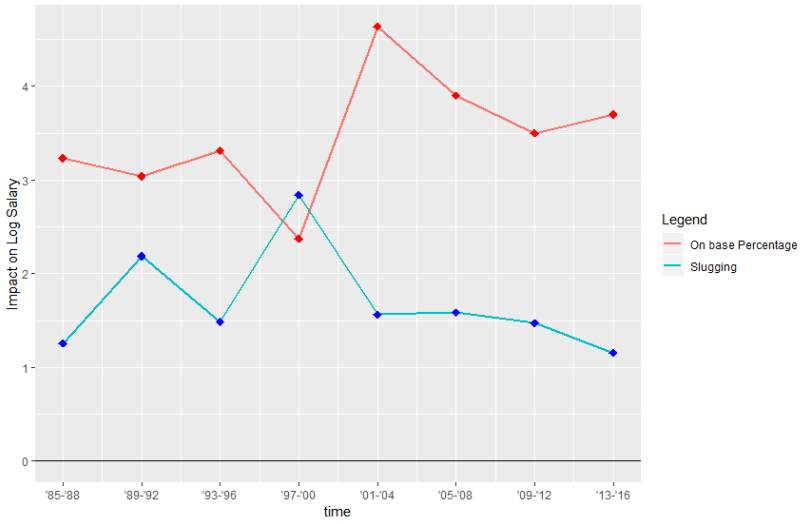
adjustment in the baseball labor market by valuing on-base-percentage more. In order to determine whether the ideas introduced by Michael Lewis and Billie Beane truly had an effect on player valuation I had to break up my dataset into different time periods. Below in Figure 2 I decided to split the dataset into 4-year periods in order to examine how returns have changed over time. If labor markets did adjust to the Moneyball phenomena I expect to see an increase in the coefficients for On-Base-Percentage and Slugging over time.

Figure 2: Salary Regression Overtime

	2011-2016		**2006-2010**		**1999-2005**		**1993-1998**		**1985-1992**	
Predictors	*Estimates*	*p-value*	*Estimates*	*p-value*	*Estimates*	*p-value*	*Estimates*	*p-value*	*Estimates*	*p-value*
(Intercept)	12.40	**<0.001**	12.24	**<0.001**	11.45	**<0.001**	10.96	**<0.001**	11.50	**<0.001**
OBP	3.71	**<0.001**	3.49	**<0.001**	4.08	**<0.001**	2.89	**<0.001**	3.03	**<0.001**
Slugging	1.65	**<0.001**	1.63	**<0.001**	2.75	**<0.001**	1.59	**<0.001**	1.28	**<0.001**
AB	0.00	**<0.001**	0.00	**<0.001**	0.00	**<0.001**	0.00	**<0.001**	0.00	**<0.001**
Catcher	-0.27	**<0.001**	-0.20	**0.002**	-0.14	**0.026**	0.03	0.589	-0.04	0.378
Pitcher	1.59	**<0.001**	1.73	**<0.001**	1.91	**<0.001**	2.24	**<0.001**	1.63	**<0.001**
Outfielder	-0.01	0.876	-0.14	**0.002**	-0.05	0.242	0.01	0.756	0.00	0.902
Observations	2522		2970		3151		3249		3472	
R^2 / adjusted R^2	0.191 / 0.188		0.200 / 0.197		0.275 / 0.272		0.342 / 0.339		0.304 / 0.302	

What I find is that On-base-percentage does grow in importance over time going from 3.03 to 3.71 at the end of the 32-year time frame, while the slugging coefficient experiences less significant change over the same period. This confirms the assumptions that on-base-percentage was undervalued before 2002, and as labor markets realized the importance of the statistic it grew in significance with respect to player salary. On the other hand, the market returns for slugging have remained stagnant, which suggests that the statistic may still be undervalued. I have also included a graph to illustrate the changes over time to the market returns for these two statistics. The graph is split into 3-year periods instead of 4.

Figure 3: Coefficients Overtime



From Figure 3 above we visually confirm that on-base-percentage has indeed seen an increase in coefficients overtime. Specifically, we can see that there is a dramatic increase from 1997 to 2004, which is likely due to the introduction of Sabermetrics and Moneyball. After this period on-base-percentage remains at a higher level then the pre-Moneyball periods. Confirming again that baseball did adopt a more analytical approach to player evaluation. We can also see how Slugging returns remain stagnant over time. Although there is a brief period in 1997-2000 where returns spiked, they quickly return to pre-Moneyball levels and have remained there since. This contradicts our assumption that baseball has completely accepted the driving ideals behind the Moneyball movement and suggests that Slugging is still undervalued in terms of player salary.

I would have also liked to add more regressors to my model to increase predictive power, unfortunately I was limited as a result of multicollinearity. Since many of these offensive statistics are constructed from each other and because there is a high degree of correlation among them, adding more can give rise to variance inflation. For example, if we decide to include RBI, which is a statistic that credits a player for making a play that allows a run to be scored, our variance rises dramatically. I have included Variance Inflation Factor (VIF) plots below to illustrate this issue.

Before RBI

	GVIF	Df
OBP	3.471247	1
slug	3.167525	1
AB	1.476515	1
catcher	1.171785	1
pitcher	1.853296	1
outfielder	1.215327	1
factor(yearID)	1.085626	31

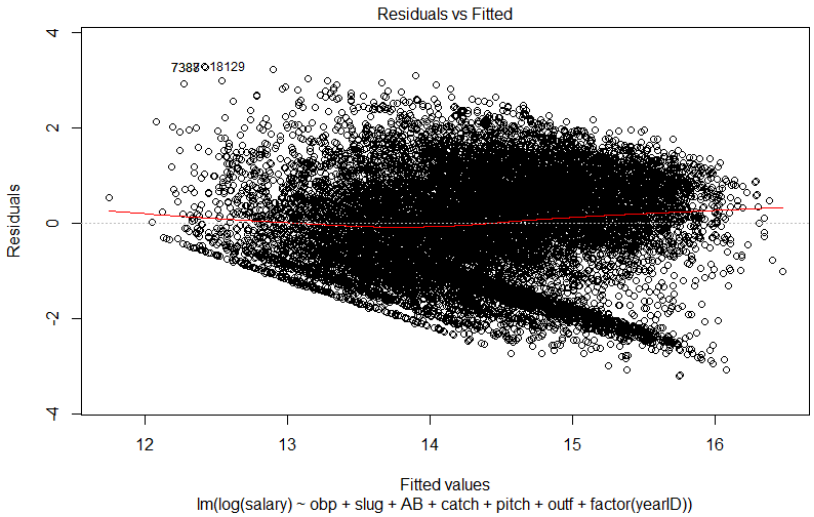
After RBI

	GVIF	Df
OBP	3.595312	1
slug	4.677508	1
AB	5.863429	1
catcher	1.172857	1
pitcher	1.953185	1
outfielder	1.216517	1
RBI	7.577714	1
factor(yearID)	1.097415	31

We can determine whether there is multicollinearity by looking at how the VIF factors change when we introduce new variables. If there is a high degree of multicollinearity then we will see an overall increase in the inflation factor values and spread. If we get a variable with a comparatively large inflation factor, then it is most likely introducing multicollinearity to our model. We can clearly see that RBI is excessively large when compared to the other factors in the model, we can therefore conclude that it should be excluded to avoid multicollinearity. I also ran a non-constant variance test in order to test whether we had heteroskedasticity in our model. The test determined that we did not have heteroskedasticity by rejecting the null of non-constant variance. I also looked at the fitted vs residual plot to confirm our results. If there is no heteroskedasticity we do not expect to see a spike in the residuals when we look at the fitted values plot. In other words, we hope to see the red line in the plot follow the white dotted line closely.

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 96.54062, Df = 1, p = < 2.22e-16
```

Figure 4: Residual vs Fitted Values



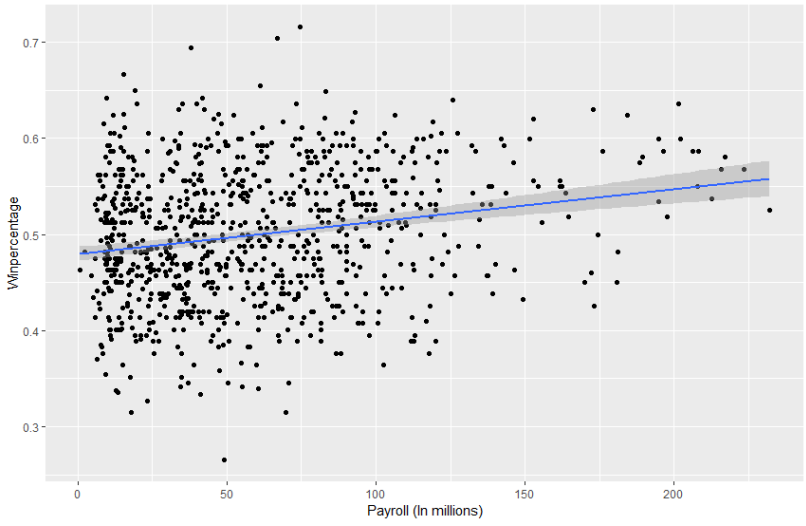
4.2 Payroll Effects on Winning

Another characteristic that is certainly advantageous to Major League teams is their financial resources. With greater financial resources comes the ability to sign multiple big-name players that demand expensive contracts. But how much is a team's success tied to the size of its market and payroll? I decided to aggregate player salaries by teams and years. This gave me each teams payroll for each season, which I then regressed on the teams winning percentage for that season. I calculated winning percentage as the number of wins over games played, which I gathered from the dataset provided in Sean Lahman's database for baseball statistics. Below is a table of the regression and a figure which illustrates the relationship between these two statistics.

Figure 5: Payroll Regression

Dependent Variable: Log(winning percentage)			
Predictors	*Estimates*	*std. Error*	*p*
(Intercept)	75.79	0.66	**<0.001**
Log(Payroll)	0.007	0.01	**<0.001**
Observations	907		
R^2 / adjusted R^2	0.12/ 0.11		

Figure 6: Effect of Payroll on Winning

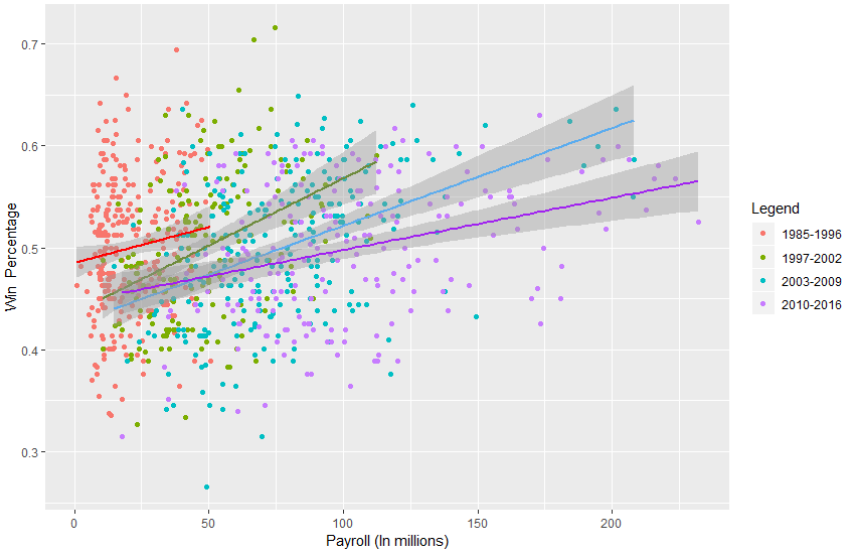


From the regression table above, we can see that there is a positive relationship between a teams payroll and their winning percentage for the season. This makes economic sense for the reasons

given earlier. Specifically, we can calculate that for each 1 percent increase in the payroll there is a .7% increase in the winning percentage. This may seem small, but it is important to note that the largest spending team last season was the Boston Red Sox at 224.1 million in payroll, while the smallest spending was the Tampa Rays with 61.9 million. This large difference in financial resources translates to a significant advantage for the Boston Red Sox. Based on our model, Boston's win percentage is 182% larger than Tampa's based solely on financial resources. It is important to note that while a larger payroll is certainly useful another important factor is the efficient use of team resources.

In an attempt to reduce the unfair advantage that larger market teams possessed the Major League Committee imposed a luxury tax on team payroll in 1997. The 1997 agreement taxed the top 5 salary teams 34% of every dollar they spent beyond the sixth salary team. This agreement proved to be unfair and ineffective resulting in the tax only lasting for 2 years. The luxury tax was reintroduced in 2002 with a threshold that teams could not pass without paying a fee. This allowed teams to control their fate rather than be punished for being in the top five spending teams. If the 2002 luxury tax agreement was truly effective, we would expect to see a decrease in the coefficient that payroll has on winning percentage. To test this hypothesis, I split the dataset into four time periods. The first being the pre-1997 luxury tax period, the second being the post 1997 luxury tax, the third being the post 2002 luxury tax period, and the last is the modern baseball period. I have included a visual figure of the data set below with lines denoting the fitted regressions.

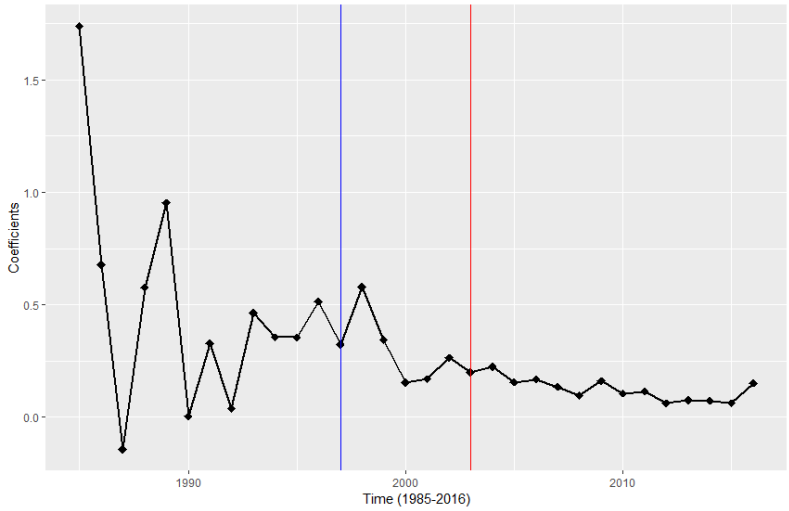
Figure 7: Effect of Payroll Overtime



From Figure 7 above we can see that over the four time periods there has been a steady decline in the value of the coefficients. It is important to note that the payroll values used in the regressions were adjusted for inflation after being aggregated to prevent our results from being biased. The declining slope of the regressions suggest that the 1997 and the 2002 luxury tax have had a negative effect on the advantage that large payrolls provide. I would like to note that although these events are correlated, we cannot confirm causality.

I have included another figure which illustrates the effect the luxury tax has had on payroll coefficients. For every year in the data set, 32 years total, I ran a regression on every teams' payroll against their winning percentage for that season. The figure shows the value of the payroll coefficients on the y-axis, and the two lines illustrate the 1997 and 2002 luxury tax respectively. We can see that before 1997 the coefficients fluctuated wildly and consistently broke .5 threshold. After 2002 we can see a dramatic decrease in the volatility of our coefficients as well as a drop in their consistent values. This reaffirms my belief that payroll advantages have decreased overtime due possibly to the introduction of the luxury tax.

Figure 8: Coefficients of Payroll Overtime



4.3 Winning and its effect on attendance

One of the greatest fears for the modern MLB community is the declining attendance to its games over the years. At the start of the 2019 season there are currently 11 teams that are unable to break the 20,000 average attendance mark, and 2018 marked the lowest average game attendance for the league in 15 years. Growing ticket, concession, and parking prices have stagnated attendance as

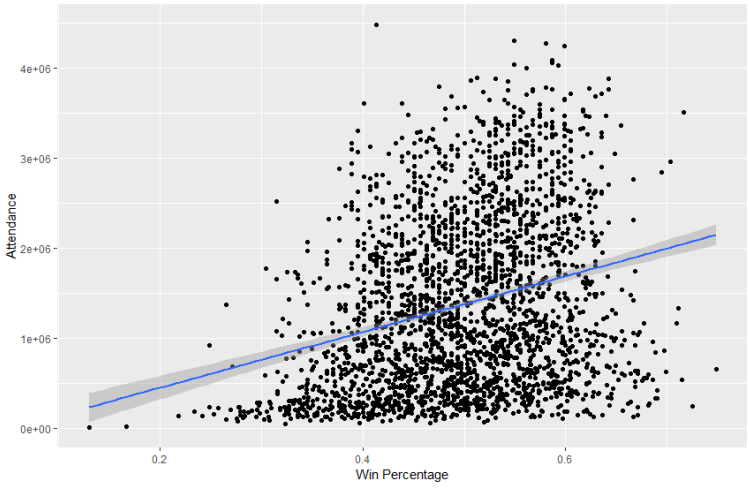
fans move to more convenient mediums to watch their favorite games. However, this decline in attendance has disproportionally affected teams that are rebuilding and lack the ability to contend on the field. Teams with big name stars that consistently win are more insulated from this dilemma. As a result, I wanted to see just how much a teams winning percentage contributed to its seasonal attendance. I was able to gather annual attendance data for all teams going back to 1871 and regressed these values on their winning percentage that season. Below is the figure that shows the regression output.

Figure 9: Attendance Regression

Dependent Variable: Log(Attendance)			
Predictors	*Estimates*	*std. Error*	*p*
(Intercept)	15.03	0.11	**<0.001**
Log(Win Percentage)	1.42	0.03	**<0.001**
Observations	2586		
R^2 / adjusted R^2	0.881/ 0.875		

I ran a Log-Log fixed effects regression which translates all variable inputs and coefficients to percentages. I would first like to note that all coefficients where significant at the 1% critical level. This confirms that winning percentage does indeed help explain the variation in park attendance. A 1% increase in a teams winning percentage results in a 1.42% increase in their attendance, which is certainly sizeable. I also included a factor variable for each year, 127 in total, so that I could control for unobserved time effects. The coefficients for these factor variables were suppressed in the final regression. To confirm that there is correlation between attendance and winning percentage I have also included a plot of both below.

Figure 10: Coefficients of Payroll Overtime



The figure above confirms our assumption that there is a positive relationship between these two variables. Out of curiosity I was also interested if there was a relationship between a teams payroll and attendance. If a team can afford to sign big players that can perform on the field and attract fans, it would make sense that payroll also has a correlation with attendance. To test this hypothesis, I ran a regression with payroll and winning percentage against a teams seasonal attendance. Unfortunately, since our salary data extends only to 1985 we are left with only 904 observations. I have included the output from the regression below and I have suppressed the factor variables that control for fixed times effects.

Figure 11: Attendance Regression with Payroll

Dependent Variable: Log(Attendance)			
Predictors	*Estimates*	*std. Error*	*P-value*
(Intercept)	13.80	0.08	**<0.001**
Log(Win%)	0.68	0.06	**<0.001**
Log(Payroll)	0.44	0.02	**<0.001**
Observations	907		
R^2 / adjusted R^2	0.569 / 0.553		

Again, we have extremely significant coefficients that are both positively correlated with attendance. Since this is a log-log regression it appears that for every 1% increase in attendance there is a .44% increase in attendance. It is important to note that payroll only indirectly affects attendance by allowing teams to sign star players.

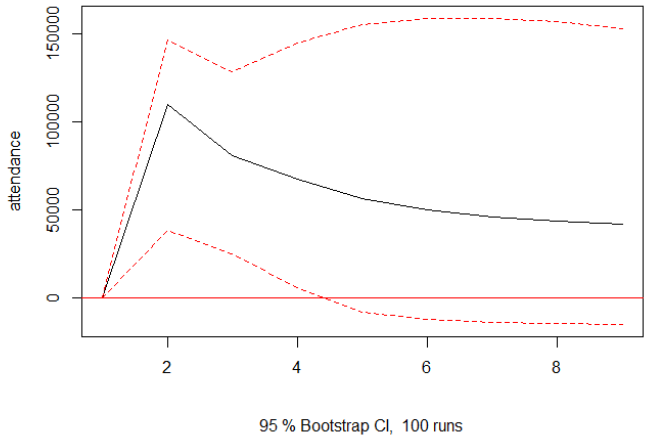
In order to determine causality with a stronger level of certainty I decided to create a VAR model where I regress attendance against lags of itself and lags of win percentage. Since VAR models must be fit using time series variables, I decided to follow one team across time and record their attendance and win percentage. I chose to follow the Chicago Cubs since they are one of the first MLB teams with 127 observations (seasons) in the data set. Since VAR models do not allow for simultaneous events, I also decided to lag winning percentage by one year. When running the VAR function I chose the best model according to the AIC comparison metric. Below is the final model that I reached.

```
Estimation results for equation attendance:
===========================================
attendance = attendance.l1 + winpercent.l1 + attendance.l2 +
winpercent.l2 + const

                Estimate Std. Error t value Pr(>|t|)
attendance.l1  1.013e+00  8.396e-02  12.064  < 2e-16 ***
winpercent.l1  1.677e+06  2.770e+05   6.054 1.66e-08 ***
attendance.l2 -2.419e-02  8.538e-02  -0.283    0.777
winpercent.l2 -1.366e+06  2.913e+05  -4.690 7.29e-06 ***
const         -1.178e+05  1.427e+05  -0.826    0.411
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can see that the VAR model includes one lag of attendance and win percentage that are extremely significant, and a second lag of the same variables, however, only win percentage is significant. According to the VAR model only two seasons, the current season and the previous season, have a significant effect on the attendance today. Attendance the previous season also has a significant effect on the attendance today. This may likely be due to some momentum factors; it makes sense that if there is a large attendance the previous season there will likely be a large attendance number the next season. This model also helps affirm our assumption that there is a relationship between win percentage and attendance. After constructing the VAR model I was able to calculate the Impulse Response Function, which illustrates how a shock in win percentage affects attendance. Below is the figure of the impulse response function.

Figure 11: IRF of Win percentage on Attendance



The impulse response figure above shows us what happens to the Chicago's Cubs attendance over time when we introduce a positive win percentage shock in period 1. We can see that there is an immediate increase in attendance for period 2, and a gradual decay in this shock over time. I

would like to note that even after 8 periods we can see that the shock persists, which means that the effects of these shocks last a long time. Although our VAR model results show us how these two variables interact overtime our results may be unreliable due to our lack of observations. We must therefore look at these results with some level of skepticism.

Next, we will look at the results from the granger-causality test between win percentage and attendance. A granger-causality test does not determine true causality, but instead determines predictive causality. We can determine predictive causality if one time-series is useful in forecasting another. If we conclude that one time-series granger-causes another then we can assume with a stronger degree of confidence that there is a strong interaction between the two variables. Below I have included the output from granger-test.

```
Granger causality test

Model 1: attendance ~ Lags(attendance, 1:2) + Lags(winpercent, 1:2)
Model 2: attendance ~ Lags(attendance, 1:2)
  Res.Df Df      F    Pr(>F)
1    120
2    122 -2 19.706 3.979e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

I selected the number of lags to be 2 based on the VAR model we calculated earlier. We can see that the P-value for the causality test is significant at the .1% critical level meaning we can reject null of hypothesis that win percentage does not granger cause attendance. In other words, we can conclude that win-percentage granger-causes attendance. Knowing this we can start to understand why teams that do well and are able to compete remain unaffected by the rampant decreases in game attendance that is affecting much of the Major League. Although it is much easier said than done, it seems that if teams want to increase game attendance one sure fire way is to start winning more games.

4.4 Effects of OBP and Slugging on Winning

Although we previously concluded that On-Base-Percentage and Slugging help explain the variation in player salary we have not discussed how effective these indicators are at scoring runs and ultimately winning games. If we want to determine the relationship between OBP/Slugging and

winning, we can run a Granger Causality test on each of the Variables. Similar to how we fit our previous VAR model, we can follow a team like the Chicago Cubs throughout their 130 year existence and look at how the variables interact over time. The data used for this example goes back to 1871 and has records of the number of games ever team won per season as well as the average on-base-percentage and slugging percentage for the entire team. First, I looked at the degree to which these variables can help explain the variation in the number of games won by running a fixed effect regression.

Figure 12: Wins Regression

Dependent Variable: Log(Wins Per Season)			
Predictors	*Estimates*	*std. Error*	*p*
(Intercept)	7.31	0.09	**<0.001**
Log(Slugging)	0.85	0.07	**<0.001**
Log(OBP)	1.74	0.10	**<0.001**
Observations	2690		
R^2 / adjusted R^2	0.513 / 0.488		

I suppressed the dummy coefficients that control for fixed time effect in the regression output above. The regression is Log-Log so that all betas and inputs are in terms of percent change. From the regression we can see that slugging and OBP are again extremely significant, and their coefficients are extremely sizeable. A 1% increase in either slugging or OBP result in a .85% or 1.74% increase in the number of games won respectively. The fact that OBP is two times more influential than slugging supports statements made by Michael Lewis in his novel. Lewis claimed that OBP and Slugging where the best indicators for offensive performance, but that OBP was the superior metric among the two (Lewis, 2003). I wanted to include other performance metrics such as RBI or the number of homeruns hit per season to see how the coefficients changed. Unfortunately, since many of these performance metrics are highly correlated we run into the issue of multicollinearity. I have included a figure below which shows the Variance Inflation Factor of two regressions, one with RBI and HR included and one without.

```
                GVIF  Df
log(SLG)     5.905778   1
log(OBP)     4.184805   1
factor(Year) 6.073533 130
```

```
                GVIF  Df
log(SLG)    17.753990   1
log(OBP)     8.505349   1
HR          13.661233   1
RBI         16.267843   1
factor(Year) 37.743222 130
```

The VIF for the variables in the first regression are clearly much lower and more concentrated, which indicates that none of the predictors are inflating the variance due to multicollinearity. The second model includes RBI and HR, which clearly drives up the VIF for each variable and increases the spread dramatically. This rapid increase in the VIF spread is a sign of multicollinearity, which is problematic because it can skew our results and make predictors less significant.

After looking at the OLS relationship between these statistics I wanted to determine whether these variables granger-caused one another. Specifically, I wanted to look at whether OBP or Slugging granger-caused wins. I again decided to follow the 130 seasons' that the Chicago cubs have played throughout their existence. I first had to gather all three variables and convert them into time series variables so that I could fit them into the VAR model. I then had to lag both Slugging and OBP by one time period, because the VAR model does not allow for simultaneous events to occur. After lagging the time-series variables I was then able to fit the two VAR models. I again used the AIC metric in order to determine what would be the best number of lags. I have posted both of the VAR models in the figure below.

```
Estimation results for equation wins:
=====================================
wins = wins.l1 + slug.l1 + wins.l2 + slug.l2 + const

          Estimate Std. Error t value Pr(>|t|)
wins.l1   0.56839    0.08831   6.436 2.42e-09 ***
slug.l1 -77.42399   28.43033  -2.723   0.0074 **
wins.l2   0.10543    0.08789   1.199   0.2326
slug.l2  64.45092   28.27524   2.279   0.0244 *
const    33.88241   17.67531   1.917   0.0575 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Estimation results for equation wins:
=====================================
wins = wins.l1 + obp.l1 + const

         Estimate Std. Error t value Pr(>|t|)
wins.l1  0.59718    0.07378   8.095 4.02e-13 ***
obp.l1  -1.13302   30.12931  -0.038  0.97006
const   30.98973   11.65706   2.658  0.00886 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the VAR models above that slugging has one significant lag, while OBP only has zero significant lags. This seems to suggest that neither variable has strong predicting power for wins. In addition to low significance lags the coefficients suggest that slugging has a negative effect on winning which does not make sense. In order to confirm this assumption, I ran two granger-causality tests for both slugging on OBP. The results for both these tests can be see below.

```
Granger causality test (Slugging on Wins)

Model 1: wins ~ Lags(wins, 1:1) + Lags(slug, 1:1)
Model 2: wins ~ Lags(wins, 1:1)
  Res.Df Df      F Pr(>F)
1    127
2    128 -1 2.1833  0.142
```

```
Granger causality test (On-base-percentage on Wins)

Model 1: wins ~ Lags(wins, 1:2) + Lags(obp, 1:2)
Model 2: wins ~ Lags(wins, 1:2)
  Res.Df Df      F Pr(>F)
1    124
2    126 -2 2.0665  0.131
```

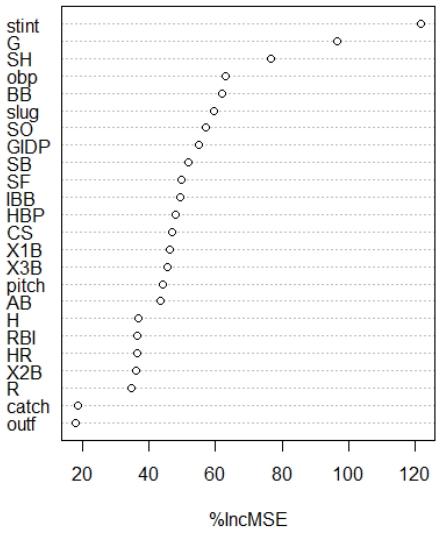
Both of these tests reached the same conclusion that OBP and Slugging does not granger-cause wins. Therefore, while there maybe a statistically significant OLS relationship between slugging/OBP and wins there is a lack of causality, despite the high R squared. It is important to note that VAR models and Granger-causality tests require a high amount of observations in order to be reliable. The small 130 observation time frame, which I am limited to maybe limiting the statistical power of our models. Therefore, there may still be a significant causal relationship between these variables even though the tests state otherwise.

4.5 Creating a predictive model

For the last section of the project I wanted to create a predictive model that could be used to determine whether a player was undervalued or overvalued based on their offensive capabilities. I first tried to fit a regression to the data using stepwise variable selection. I was able to determine the best model according to AIC and BIC, however, due to multicollinearity my coefficients were uninterpretable and counterintuitive. I instead decided to go with a random forest model since it could run through several regressions using different variables with varying weights in order to

determine which variables have the most predicting power. I decided to fit the forest to the inflation adjusted salary data and ran through 500 trees. I have included the variable importance plot below.

Figure 13: Random Forest Importance Plot



From the plot above we can see that stint, number of games played (G), OBP, and slugging are some of the most influential variables when it comes to predicting salary. This helps affirm the decision to regress OBP and Slugging on salary earlier. Using the random forest model, I then predicted player salaries based on offensive statistics and compared it to their actual salary. For example, the model predicted LA dodgers' pitcher Clayton Kershaw to have a salary of 19 million a season, when his actual contract is 34 million a year, signifying that he may be overvalued. The large discrepancy in predicted vs actual salary is likely due to factors not captured by our model. Kershaw is one of the most consistent and sought-after pitchers in the MLB, he has been with the Dodgers since 2008 which is extremely rare, and has not suffered any major injuries throughout his career. None of these factors are captured and are likely the main cause of our discrepancy. Another franchise player whose salary I predicted was Centerfielder Mike Trout for the Los Angeles Angels. According to the random forest model his salary should be 14 million a season, when in reality his contract is 17 million. Although less over valued than Clayton Kershaw the statistical discrepancy is much smaller. Lastly, I predicted Miami Marlins catcher Marcell Ozunas salary. The model determined that Marcell should have a salary of around 1.5 million, when his salary is closer to 6 hundred thousand. This means that Ozuna is likely undervalued based solely on his offensive

capabilities. It seems the model is inaccurate at predicting actual player salary, but it is important to remember that our model only takes into account offensive statistics and lacks variables for length of career, defensive capabilities, or injury. Therefore, we can still conclude that players may be overvalued or undervalued based solely on their batting skills, but we lack the ability to make more in-depth calculations.

5. Conclusion and Future Work

Based on the results, we can confidently draw conclusions about the nature of baseball salaries, player evaluation, and team performance. First, we can conclude that OBP and Slugging are strong predictors of offensive team performance and player compensation. From the regression results we saw that the coefficients were statistically significant and had a large impact on log salary. We were also able to conclude that On-Base-Percentage became more valued over time and that Slugging remained at consistent levels over time. We also looked at the effects of payroll on winning percentage and were able to conclude that payroll did have a statistically significant effect on winning percentage, despite the efforts by the MLB committee to try and reduce this relationship. I looked at the relationship on winning percentage and attendance as well and we were able to determine that winning percentage did Granger cause attendance. I was also able to determine the magnitude of the relationship between these variables through a simple regression. we looked at the time-series relationship between offensive statistics winning percentage. It was determined that these variables did not have enough predicative power to determine whether a team would have high or low winning percentage in the next period. Lastly, I tried to make a predictive model that could determine whether certain players are under or overvalued. It was determined that the model was ineffective at predicting real world examples due to the fact that it was limited in its scope.

Although we were able to reach these conclusions there is still other areas that can be expanded on. For example, we can add defensive capabilities to our models, look at the time-series relationship of variables for multiple teams, add more complex sabermetric performance indicators, or create different predictive models.

6. Appendix

Bootstrap Coefficients

```
BOOTSTRAP OF LINEAR MODEL  (method = rows)

Coefficients:
      (Intercept)              obp             slug               AB            catch
         11.13399          3.34722          1.73505          0.00240         -0.11670
             outf            pitch
         -0.03778          1.83476
```

References

Beneventano, P., Berger, P. D., & Weinberg, B. D. (2012). Predicting Run Production and Run Prevention in Baseball: The Impact of Sabermetrics. *International Journal of Business, Humanities and Technology,2*(4), 67-75. Retrieved from http://www.ijbhtnet.com/journals/Vol_2_No_4_June_2012/7.pdf

Chang, J., Zenilman, J., & Bishop, K. (2013). A Study of Sabermetrics in Major League Baseball: The Impact of Moneyball on Free Agent Salaries. Retrieved from https://olinblog.wustl.edu/wp-content/uploads/AStudyofSabermetricsinMajorLeagueBaseball.pdf.

Hakes, J. K., & Sauer, R. D. (2006). An Economic Evaluation of the Moneyball Hypothesis. *Journal of Economic Perspectives,20*(3), 173-186. Retrieved from https://www.aeaweb.org/articles?id=10.1257/jep.20.3.173.

Lahman, S., & Sabermetrics. (2018, May). Sean Lahman's Baseball Database. Retrieved from http://www.seanlahman.com/baseball-archive/statistics/

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

Sports Reference LLC. (n.d.). Baseball Reference Database. Retrieved from https://www.baseball-reference.com/