Salaries vs. Wins in the MLB

Kevin Corcoran
School of Computing and Information
University of Pittsburgh
Pittsburgh, PA, USA
kjc100@pitt.edu

Drew Whiteside
School of Computing and Information
University of Pittsburgh
Pittsburgh, PA, USA
daw196@pitt.edu

Joseph Dynoske
School of Computing and Information
University of Pittsburgh
Pittsburgh, PA, USA
jwd36@pitt.edu

***Abstract*—The following paper will use a combination of Excel and R to formulate and analyze MLB data in a way to determine the extent of a relationship between team salaries and wins in the league. Graphs will be formed that illustrate the relationship between salary and win percentage, along with league-wide buying power per year.**

Keywords—MLB, baseball, professional, sports, salaries, payroll, consumer price index (CPI), inflation, data, analytics, University of Pittsburgh, INFSCI 0310, final project

I. INTRODUCTION

Since the 19th century, teams have competed in Major League Baseball (MLB) for the chance to win the World Series. At the start of every season, each team has an equal chance to win enough regular season games to make the playoffs and ultimately the World Series. However, when observing overall achievements over time, it is evident that not every team has had equal success during their time in the league. Out of the current 30 MLB teams, 6 of them have not won a championship, and one team, the Seattle Mariners, has not reached the World Series at all. On the contrary, teams such as the New York Yankees have won 27 titles in their history. While every team has not existed in the league for the same amount of time, the disparity of success calls into the question of how such an inequality has come to exist.

A potential explanation for the imbalance of wins is the amount of money team executives put into the teams themselves, specifically in the form of player salaries. Without the existence of a salary cap to regulate the amount spent on players each year, it is plausible that a "pay to win" scenario could exist in some respect. For example, leagues such as the NBA and NHL institute a limit to the total amount of money each team can spend on player salaries, which is intended to create a more even playing field and limit the possibility of dominant teams emerging simply because they can pay more for the most talented players. However, there are mixed opinions over whether spending more money on payroll can increase success in the MLB in the first place. The following report aims to determine the extent to which salaries affect the frequency of a team's wins in the regular season.

II. GOALS/PROBLEM STATEMENT

The goal of the report is to establish whether there is a correlation between payroll and winning percentage in the MLB. If there is a positive correlation between with an r-squared value of at least 0.9, then we will determine that salary does in fact influence seasonal success. If there is a weak correlation or no correlation, then we will conclude that salaries do not have a significant effect on success. We will also create a variable called "Buying Power," which will represent an average team's win percentage per dollar paid in salary. Using Buying power, we will be able to view how the ability to influence win percentage with salary has changed year-by-year from 1985 to 2016.

III. METHODS

To analyze the impact of total salaries on team performance, we decided to use the R Library for Sean Lahman's Baseball Database. While there are plenty of other sources for data, we found that this was the most organized and usable for our circumstances. It is also convenient that it can be installed as an R package, allowing us to manipulate and represent data easily. This library provides convenient and comprehensive data on all aspects of professional baseball, including salaries by player for each MLB team since 1985, as well as win/loss statistics for each team. As we proceeded though the study, we selected R packages that we felt were useful in collecting, representing, and observing the data. The results of these packages indicated which variables and trends to study, ultimately leading conclusive evidence in the relationship between salary and win percentage in the MLB. The following describes which packages were used, how the data was collected and studied, and what led us to make different choices throughout the study.

A. R Packages and Usages

The following indicates the most important R Studio packages that were use, as well as how they were implemented in our study:

- "Lahman" – Sean 'Lahman' Baseball Database: the primary source of MLB data. This package supplied information on salary and wins from 1985 to 2016.
- "quantmod" – Quantitative Financial Modeling Framework: used to record CPI data from the FRED

database over time, creating a conversion factor to account for inflation. This factor made sure that comparisons in salary data were accurate and unbiased by inflation.

- "dplyr" – A Grammar of Data Manipulation: this package made extracting relevant fields from the Lahman database quick and easy. It was primarily used to filter, group, and mutate large amounts of data with minimal code.
- "tidyr" – Tidy Messy Data: used to separate and combine column data by keywords/expressions.
- "ggplot2" - Create Elegant Data Visualizations Using the Grammar of Graphics: used to create all visual representations of data. Allowed us to make plots with meaningful titles, highlight relevant datapoints, and perform regressions.
- "stats" – The R Stats Package: used to perform linear and curvilinear regressions on the data.

B. Initial Data Processing in R

Lahman's database provides a vast amount of data, much of which was irrelevant to the study. In order to make clearer observations, the data had to be filtered and compiled into new dataframes. The "Teams" dataframe in the package contains statistics on individual players, teams, years, and parks. To make this data useful for the study, we had to average the salaries and wins of all players by team and year. A "WinPct" column was added to show the percentage of wins by each team per year. After pulling all the data we needed from Lahman's library, we were left with the average salary and win percentage per team from 1985 to 2016.

C. Accounting for Inflation

While doing some initial tests and rough analysis, it quickly became apparent that the effect of inflation on our salary data could potentially skew the results. To counteract this, we used the "quantmod" R package to obtain financial history data. This package allowed us to pull monthly consumer price index data from the FRED database. We then found the mean CPI pear year and collected that into a new dataframe. Using this data to adjust the yearly salaries, we divided every yearly CPI by the CPI in 2020, providing us with a conversion factor for every year. By dividing each team salary by the conversion factor for their corresponding year, the data was finally adjusted to dollar amounts relative to 2020.

D. Team Salaries vs. Win Percentage Observations

After the data was normalized to the same year, we were able to begin some observations. Since all the salary data was above one million dollars per year, we created a new variable, "SalaryAdjustedMil", which is the adjusted team salary divided by 1,000,000 and made our graphs and data more readable. To meet our goal of finding a relationship between wins and salary, we started by plotting the average team salaries by year on the x-axis, versus the corresponding win percentage on the y-axis. A linear regression was also performed and imposed on the plot, as well as shading for the confidence interval. This was done using the 'ggplot' R package, which allowed us to easily plot and format a visual representation of the data, as well as perform statistical operations. The resulting R-squared and p-values were not sufficient to come to a firm conclusion, so we had to try to find some other possible trends in the data. After seeing the results on the plot, we added color indicators to each point based on the year. By doing this, we were able to see another relationship that we could study, the effect of salary on win percentage over time. Adding this metric to the graph allowed us to see a potential trend, so we decided to study this further.

E. Studying Salary vs. Win Percentage Over Time

To study the change in the effect of team salary on win percentage over time, we had to create a new variable and gather data from our original table. The influence of salary on win percentage can be studied simply by dividing win percentage by salary. We decided to call this variable "Buying Power", as it shows how much of an effect an increase in spending has on the success of the team. Since we no longer needed to differentiate between teams and rather wanted to study the league as a whole, we created a new table which showed the average buying power in the league per year. By plotting this data, we were able to see the change in buying power over time, which visually showed a logarithmic curve. To check this, we performed a regression on the data using the log10 of buying power as the y variable, and the year as the x variable. This was then added to our plot by adding the x and y values of this regression to another dataframe and plotting those values as a dashed line on the plot of buying power over time. The resulting R-squared and p-values of this regression showed promise in the relationship, allowing us to come to some final conclusions about the decrease in the effect of salary on win percentage over time.

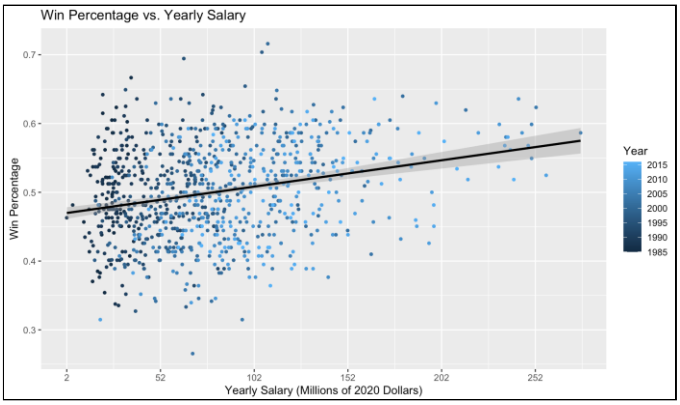
IV. Results

After all our data processing and observations were complete, we were able to come up with some promising correlation statistics. The following results show the process of data collection and visualization and indicate how and why choices were made throughout the extent of the study. All data processing was completed in R Studio, using the packages indicated in the "Methods" section of this document.

A. Graphs, Plots, and Regression Data

1) Wins vs. Salary Data: see appendix.

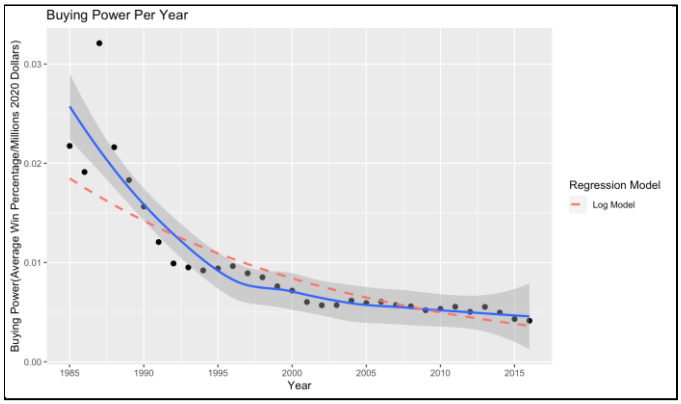
2) Wins vs. Salary Plot:



3) Buying Power Per Year:

Year	Buying Power
1985	0.02174861517
1986	0.01912177131
1987	0.03209848654
1988	0.021618501
1989	0.01830870851
1990	0.01563657719
1991	0.01206614113
1992	0.009906571692
1993	0.009507743618
1994	0.009203355016
1995	0.00940780057
1996	0.009638995108
1997	0.008916373831
1998	0.008514367563
1999	0.007605900891
2000	0.007177267098
2001	0.006023976583
2002	0.00569558889
2003	0.005706577868
2004	0.006164059098
2005	0.00592147452
2006	0.006052142594
2007	0.005717231401
2008	0.005591454572
2009	0.005195924601
2010	0.005337350427
2011	0.005546188514
2012	0.005033056548
2013	0.005527920521
2014	0.00495322603
2015	0.004300642108
2016	0.004136938033

4) Buying Power Per Year Plot:



5) Wins vs. Salary Regression

```
Call:
lm(formula = WinPct ~ SalaryAdjusted, data = winSal)

Residuals:
      Min        1Q    Median        3Q       Max
-0.230219 -0.047910  0.000668  0.048884  0.205036

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.691e-01  4.347e-03 107.924  < 2e-16 ***
SalaryAdjusted 3.833e-10  4.688e-11   8.176 9.72e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06633 on 916 degrees of freedom
Multiple R-squared:  0.06801,   Adjusted R-squared:  0.067
F-statistic: 66.85 on 1 and 916 DF,  p-value: 9.719e-16
```

6) Buying Power Log Regression:

```
Call:
lm(formula = log(BuyingPower) ~ Year, data = BuyingPowerPerYear)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28480 -0.15086 -0.01647  0.11967  0.65742

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 100.141211   8.020092   12.49 2.05e-13 ***
Year         -0.052460   0.004009  -13.09 6.22e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2094 on 30 degrees of freedom
Multiple R-squared:  0.8509,    Adjusted R-squared:  0.8459
F-statistic: 171.2 on 1 and 30 DF,  p-value: 6.221e-14
```

B. Findings/Discussion

Initially, we expected to find a stronger correlation between salary and win percentage from the original data. However, as Wins vs. Salary plot and regression show, there is not a strong relationship between the two. This is mostly explained by the R-squared value of only 0.067. However, by adding color indicators to the points based on their year, in the earlier years the correlation between the variables in data seems to be much stronger. To study this, we created the Buying Power variable, which is defined as the win percentage of each team per year

divided by the average team salary for that year (in 2020 millions of dollars). We then averaged this variable by year to study its change over time.

As seen in the Buying Power Per Year table and plot, there seems to be a stronger logarithmic trend, implied by the blue trend line and shaded confidence interval. By running logarithmic regression on the data, we found a strong R-squared value of 0.8459 with a confident p-value of 6.221e-14. The similarities in the curve of this regression to our data can be seen on the Buying Power Per Year Plot, as we can determine that the influence of salary on win percentage has decayed significantly since the 1980s.

When observing the "Buying Power per Year" graph, a possible reason for the downward trend in buying power could be the MLB's institution of a luxury tax. The first stage of the tax occurred from 1997 to 1999, where the 5 teams with the highest payroll had to pay 34% of every dollar spent above that of the sixth most spending team in the league. The introduction of the tax may have reduced buying power because the graph dips slightly from years 1996 to 1997 and continues to decrease gradually through 1999, when there had previously been a moderate trend upward from 1994 to 1997. However, a counterpoint to the existence of an effect of luxury tax on buying power is the period from 2000 to 2002, when the luxury tax was lifted temporarily. Despite no teams being limited by a tax, the buying power continued to decrease during those years. Finally, the luxury tax, renamed as the Competitive Balance Tax, was reinstated in 2003 and has continued to the present day. The system remained the same, except that a tax threshold would be set for each season, and teams with payrolls above that threshold would pay a percentage of every dollar above that set value. Teams who exceeded the threshold for a second consecutive year would have to pay a higher percentage than that for first-time offenders, and the rate would be further for the third and fourth straight years. In the period 2003 to 2016 in the graph, there is a slight overall downward trend, but there is no consistent decrease. Overall, the lack of distinct changes in the trajectory of the buying power graph as the luxury tax was instated and lifted suggests that the MLB's luxury tax does not have a significant impact on yearly buying power. It may be one of multiple reasons that has caused the reduction in buying power, but it is likely not the sole reason.

V. Conclusion

In conclusion, our team had the task of finding out whether a baseball team having higher player salaries correlates to that team being able to win more than a team with lower player salaries. We found that the average player salary actually does not have an impact on a team's chance to win in more recent years. A team's chance to win is based on the skill of the players on it, and not how much said players are getting paid. However, we found that player salary used to have a much bigger impact in the 1980s; teams with bigger salaries would often perform better. This influence decays heavily after 1980, though.

Even though our original assumption was wrong, we still went further into the data to see if we could find any relationships with teams having a high salary and win percentage over time. In order to do this we created the variable "Buying Power" which represents win percentage of each team per year divided by the average team salary for that year. Buying Power allowed us to see how the potential to influence wins with salary has changed on a yearly basis. We found that the effect that average salaries have on getting wins has decreased over time, and probably will continue to stay in that trend.

We accomplished this project using R and the R library for Sean Lahman's Baseball Database as well as various other R packages such as quantmod, dplyr, tidyr, ggplot2, and stats. Lahman's database was filled with information that did not concern our study, so we had to filter and compile the data into dataframes. We also created a new column in the data frames that detailed the win percentage of each team. Our data was also affected by inflation. Using quantmod, we used the CPI in 2020 to turn the dollar amounts of all salaries to be close to 2020.

Overall, the goal of this project was achieved because we were able to use Lahman's data set to answer the question of whether player salaries affect win percentage or not. After answering this question, we explored the relationship more by making the Buying Power variable and analyzing the effect that average salaries has on winning. Through our findings, we have determined that average player salaries in the MLB has had a lower and lower effect on teams win percentages as time has gone on.

Acknowledgment

Thank you to Professor David Tipper for teaching the INFSCI 310 course and providing us the skills to be able to complete this report to the most complete extent possible. Additionally, thank you to Sean Lahman, who developed the Lahman database and the R package for which we based our data on.

References

References are made to Sean Lahman's Baseball Database and other R packages used in gathering and processing data. Past studies that helped guide us are also referenced, as they gave us a baseline of what variables to study.

[1] Michael Friendly, Chris Dalzell, Martin Monkman and Dennis Murphy (2021). Lahman: Sean 'Lahman' Baseball Database. R package version 9.0-0. https://CRAN.R-project.org/package=Lahman

[2] Jeffrey A. Ryan and Joshua M. Ulrich (2020). quantmod: Quantitative Financial Modelling Framework. R package version 0.4.18. https://CRAN.R-project.org/package=quantmod

[3] Kleinbard, Martin. "Can't Buy Much Love: Why money is not baseball's most valuable currency." MIT Sloan Sports Conference. 28 February 2014. Accessed 20 September 2021. https://www.sloansportsconference.com/research-papers/cant-buy-much-love-why-money-is-not-baseballs-most-valuable-currency

[4] Edwards, Craig. "In 2019, Team Payroll and Wins Are Closely Linked." Fangraphs. 16 August 2019. Accessed 12 November 2021. https://blogs.fangraphs.com/in-2019-team-payroll-and-wins-are-closely-linked/