Is The Performance Of Major League Players Affected By Missing Time They Had In The Minor Leagues?

UCLA

Client: Matt (Money) Marks

Stats 141XP: Practice of Statistical Consulting

Multivariate Marlins

Full Name
Rene Delgadillo
Merve Dumlu
Rainier Ho
Jack Keeton
Daniel Smith
Arnav Talukder

Prof. Dave Zes

Department of Statistics
University of California Los Angeles

August 22, 2023

Contents

1 | Abstract

In this study, we aimed to quantify the effect of time lost by baseball players in the minor leagues and the effect that it had on their performance in the major leagues. To find this we generated variables to quantify the time lost by players in the minors. Initially, to examine the relationships between these variables and other relevant baseball statistics, we constructed correlation plots. This analysis allowed us to assess the degree of association between the time lost variables and performance indicators such as WAR (Wins Above Replacement) and other variables.

To further investigate these relationships, we conducted ANOVA tests comparing different baseball statistics with our time lost variables. These tests allowed us to establish the significance of distinct groups of players with varying levels of time lost. Our hypothesis was that the means of these groups, in terms of WAR and other variables, would exhibit statistically significant differences. The results of the ANOVA tests confirmed this hypothesis for some tests and rejected it for others, indicating that time lost may or may not have an effect on different performance indicators.

Motivated by the identified effects, we proceeded to perform multiple linear regression analysis. The purpose of this regression was to delve into the specific impact of time lost on player performance. Through this analysis, we aimed to determine whether the effect was positive or negative and to uncover any other insights regarding the relationship between time lost and performance.

We furthered our analysis by examining the relationship between various award winners and their missing time in the minor leagues. For this we also ran an ANOVA test and conducted multiple linear regression to gain insights on these extraordinary players.

Overall, our study provides a comprehensive examination of the quantified time lost variables and their associations with various baseball statistics. The findings suggest that time lost in the minor leagues has a significant effect on player performance, as demonstrated by the ANOVA tests and multiple linear regression analysis.

2 | Introduction

In the field of Baseball analytics, there is an increasing incentive to search for ever more evanescent data. When every team has the same access to common statistics derived directly from games, be it RBI, strikeout percentage, or bases stolen, the team that is able to identify non-obvious trends or otherwise parse through nebulous data will come out with a distinct advantage. To these ends, we have endeavored here to investigate one such non-obvious trend. Given that many professional baseball players advance through the minor leagues before their career in Major League Baseball, we wanted to identify if players missing time from their minor league career would go on to find their major league careers affected.

We define this search through two essential questions, whether there is a significant effect of missing minor league time on major league player performance, and, if such an effect exists, how does it effect player performance metrics and to what extent? For our purposes, we define missing time in the minor leagues to include time missed for whatever reason, including injuries or personal motivations. We calculated this metric based on how many games a minor league player missed relative to the median number of games played in their league and whether in a particular season they missed an extraordinary amount of time relative to their own career. We find that by using these metrics to statistically quantify missing time, we are able to effectively describe what otherwise might be a highly variable and difficult to define player experience into understandable terms, and most importantly, using easily accessible data that does not require time-consuming or expensive additional data gathering methods like player surveys.

The advantages of being able to use players' missing time during their minor league careers as data to predict their major league performance are self-evident. When every player recruitment represents a significant financial and logistical investment for a team, being able to better predict which minor league players would make the best recruits is an essential edge over other teams in the recruitment process.

3 | Exploratory Data Analysis

We scraped the data used in this project from the baseball statistics and analysis website "FanGraphs". Our data primarily focused on baseball players who played in the Minor League and Major League between the years 2006 and 2019. For each, we included separate datasets for batters and pitchers, with 1,770 and 210 players/rows respectively. Each row of the datasets corresponds to a players' average baseball statistics for one year of play.

Interested in answering how missing time in the minor leagues affects later major league performance, we initiated our exploratory data analysis by coming up with two ways to measure missed time. The first measures the player against the rest of the league by counting the number of years they played below the league median of games. The latter measures the player against themselves by checking if they had any years where they played a significantly lower number of games compared to the other years they played. To find these outlier years, we constructed an interquartile range and found all years below the 25% mark.

3.1 | Batting Performance

Our dataset consisting of batters has 26 variables, four of which were constructed by our team. The variables included are: Player ID, Number of Games Played, Plate Appearances, Home Runs, Runs, Runs Batted In, Stolen Bases, Rate of Base-on-Balls, Strikeout Rate, Isolated Power, Batting Average on Balls in Play, Batting Average, On-base Percentage, Slugging Percentage, Weighted On-base Average, Expected Weighted On-base Average, Weighted Runs Created Plus, Baserunning Runs, Offensive Runs Above Average, Defensive Runs Above Average, Wins Above Replacement, and Year. We added the four following variables: "Years Below Median", the number of years in the minors where the player played below the median amount in their league; "Outliers", the number of years in the minors where the player played fewer games (compared to their past); "Outliers No Rookie", the number of years in the minors where the player played fewer games (comparing to their past, rookie year removed); and the number of years played in the minors. Our focus variables for exploring missing time in the minor leagues are "Outliers" and "Years Below Median".

To visualize whether our measures have any significant relationships with batting performance, we constructed the following correlation plot.

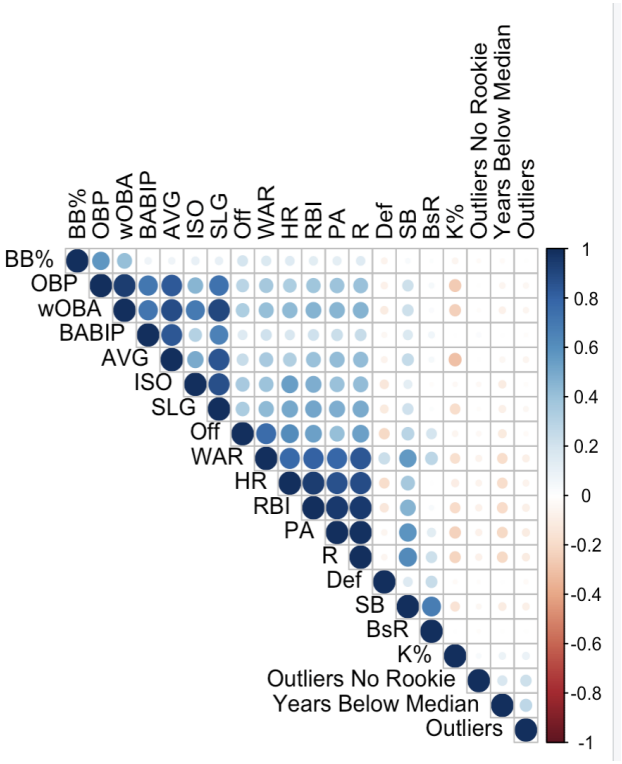


Figure 3.1: Correlation Plot of Batting Performance Variables

The correlation plot indicates that the highest correlation among our measures and other variables exists for 'Years Below Median' with homerun (HR), runs batted in (RBI), and strikeout percentage (K%).

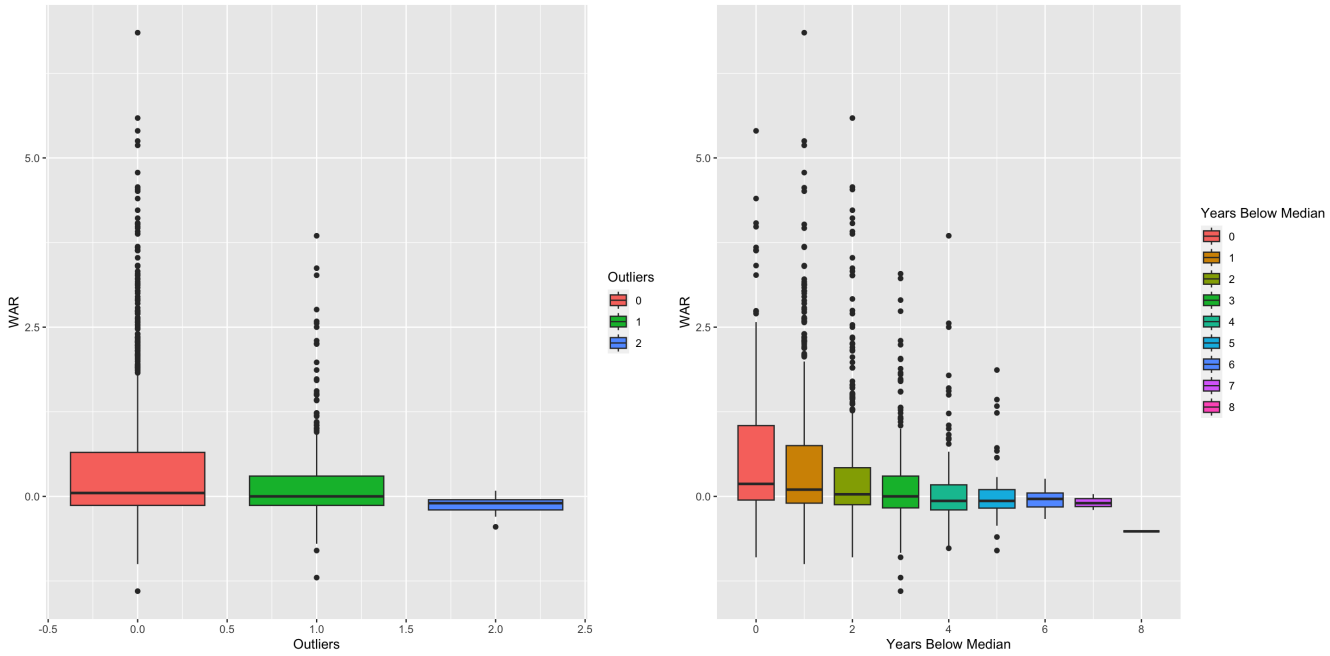
However, acknowledging that we could find a significant relationship despite the low correlation, we created an ANOVA model with our measures and wins above replacement (WAR) as our response variable. The table is provided below.

Table 3.1: ANOVA Table for WAR

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Outliers	1	10.56	10.56	13.20	0.0003
Number Below the Median	1	40.48	40.48	50.61	0.0000
Residuals	1767	1413.09	0.80		

The ANOVA results show that both 'Outliers' and 'Years Below Median' are statistically significant measures for the batting performance. The high mean square value of 'Years Below Median ' means that a baseball player's overall value compared to a replacement-level player is highly affected by the number of years in the minor league where the player played below the median amount in their league. Furthermore, the number of years in the minors where the player played fewer games when compared to the amount of games they played in other years also affects the player's batting performance.

Additionally, we made boxplot visualizations to show the distribution of all 1,770 players on their WAR vs. Outliers and WAR vs 'Years Below Median' values.



(a) Box plot for WAR against Outliers

(b) Box plot for WAR against YBM

The majority of the baseball players maintained a similar number of games they played in minors throughout the years, shown in plot (a) in red with Outliers = 0. These players tend to have a higher batting performance. Plot (b) shows that players who played below the median amount in their minor league for less than 3 years have a similar and much higher batting performance when compared to players with 'Years Below Median' greater than or equal to 3 years.

3.2 | Pitching Performance

Our dataset consisting of pitchers has 25 variables including our measures, namely: Player ID, Wins, Losses, Saves, Games, Games Started, Innings Pitched, Strikeout Average Every Nine Innings Pitched, Walks per Nine Innings, Home Runs per Nine Innings, Batting Average on Balls in Play, Left on Base Percentage, Ground-ball Rate, Home-Run-to-Fly-ball, Fastball Velocity, Earned Run Average (ERA),

Expected ERA, Fielding Independent Pitching (FIP), Expected FIP, Wins Above Replacement, Year, Years Below Median, Outliers, Outliers No Rookie, Number of Years Played in the Minors.

Again, we constructed a correlation plot to visualize whether our measures have any significant relationships with pitching performance.

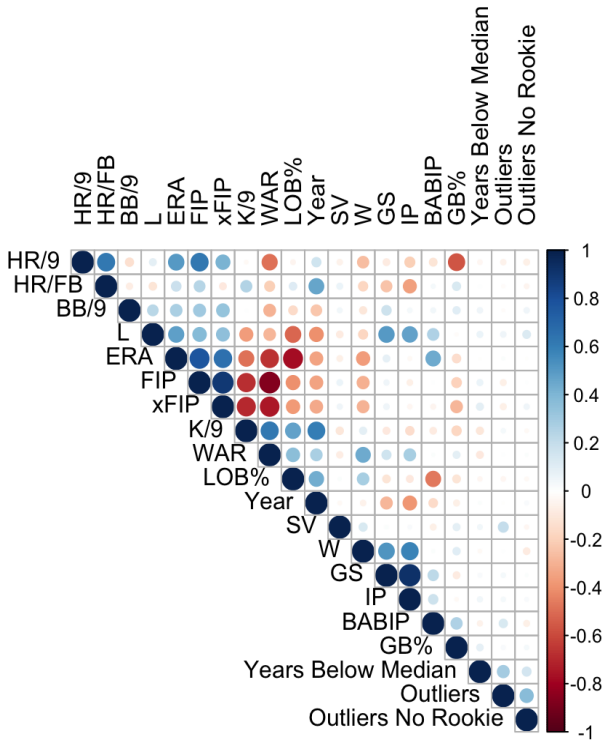


Figure 3.3: Correlation Plot of Pitching Performance Variables

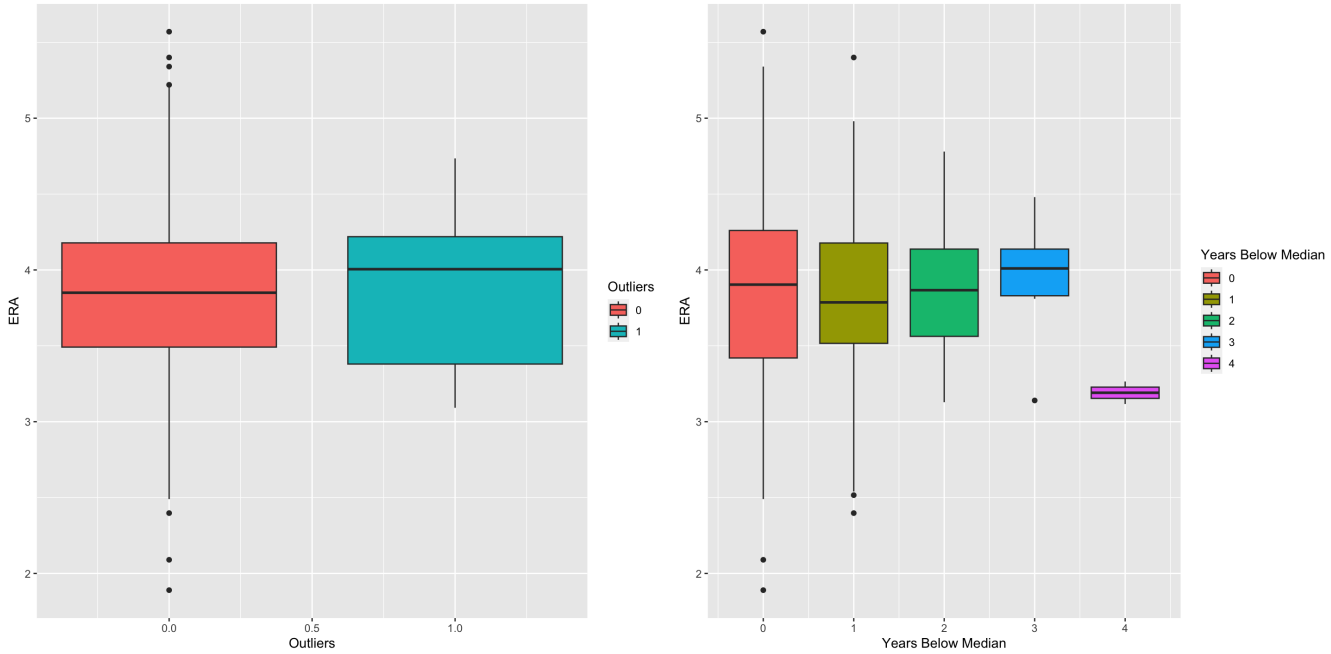
The above plot does not show a correlation between the explanatory variables and any of the performance statistics. Once again, we may be able to find a significant relationship between the inputs and the results, but we cannot anticipate it based on this plot. Therefore we created an ANOVA model with our measures and earned run average (ERA) as our response variable.

Table 3.2: ANOVA Table for ERA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Outliers	1	0.01	0.01	0.04	0.8461
Number Below the Median	1	0.00	0.00	0.01	0.9142
Residuals	207	78.78	0.38		

The ANOVA results show that both 'Outliers' and 'Years Below Median' are not statistically significant measures for the pitching performance. This means that the number of earned runs a pitcher allows per nine innings is not affected by the player's performance in the minor leagues.

To further visualize the insignificance, we created the following boxplots visualizations for all 210 pitchers.



(a) Box plot for ERA against Outliers

(b) Box plot for ERA against YBM

As shown on graph (a), the distribution of the players' ERA versus similar number of games they played in minors throughout the years are not related, as the boxplots seem to fall in the similar ERA interval. Graph (b) yields a comparable result with the exception that players who played below the median amount in their minor league for 4 years have a lower ERA, and hence a lower pitching performance.

3.3 | Exploring Potential Relationships for Batting Performance

Given the patterns we observed in the correlation plot, we decided to first assess the connection between missing time in the minor leagues and those variables most apparently correlated, these being HR (Home Runs), RBI (Runs Batted In), and Strikeout Percentage (K%). Given below are the ANOVA tables between these three variables and time lost in the minors.

Table 3.3: ANOVA Table for Home runs

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Outliers	1	242.32	242.32	7.50	0.0062
Number Below the Median	1	1859.05	1859.05	57.52	0.0000
Residuals	1767	57114.32	32.32		

Table 3.4: ANOVA Table for RBI

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Outliers	1	4908.67	4908.67	14.92	0.0001
Number Below the Median	1	20083.50	20083.50	61.04	0.0000
Residuals	1767	581334.87	329.00		

Table 3.5: ANOVA Table for Strikeout Percentage

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Outliers	1	1866.82	1866.82	14.79	0.0001
Number Below the Median	1	804.48	804.48	6.37	0.0117
Residuals	1767	223018.68	126.21		

Given these ANOVA tables which all demonstrate statistical significance, we conclude that the connection between time lost in the minor leagues and these variables is a valid one and one worth paying attention to. While we still prefer WAR (Wins Above Replacement) as our overall metric for determining what constitutes a change in player performance, we note these peripheral relationships as additional metrics to describe what effect minor league experience may be having on players' later careers. Though it is not our place to speculate given this data alone why minor league play may effect major league performance, the observation that minor league play is specifically connected with players' later success hitting homeruns or striking out may provide an important clue to guide later study.

3.4 | Games Played per League

Different MILB leagues tend to play different amounts of games per season. While Rookies tend to play 60 games in a full season, players from higher leagues may play upwards of 132 (Single-A, high-A), 138 (Double-A), and 150 (Triple-A) in a season. If we plot the games played across all players, we obtain a bimodal bell distribution with peaks around 50 and 130 (which are expected given the typical games per season and league) as well as some variation. Decomposing G by league helps visualize these differences.

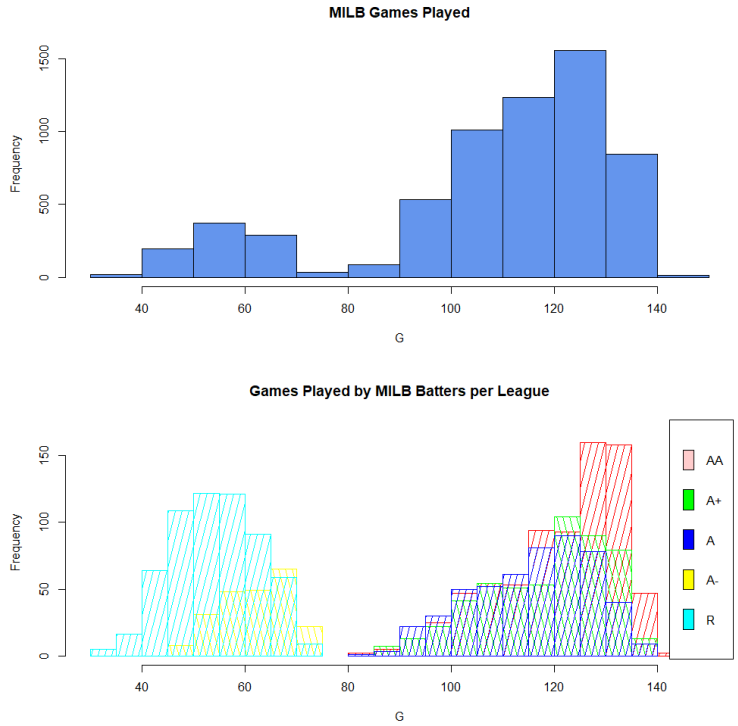


Figure 3.5: Plots of the Number of Games Played by MILB Batters

The variation in these histograms may be due to multiple factors. Players may earn demotions to lower MILB leagues or promotions to higher leagues or to the MLB which may affect the games they play in a season. Moreover, a team's roster may change per team depending on player's skill or team strategy. And of most importance to our research question, receiving injuries and rest may produce significant time loss in a season (some players even missing whole seasons).

4 | Modeling

For the modeling of our problem, we decided to focus on batting performance. The ANOVA and exploration of the pitching performance as it relates to our input variables yielded insignificant results. Choosing our model, we wanted the ability to see the effect of missing time as an effect that we could measure. We considered a simple t-test, but the findings would be somewhat irrelevant given that ANOVA had already shown there is significant difference between populations that missed time and those that did not. So we chose to approach modeling with multiple linear regression using our two created inputs

"Outliers" and "Years Below Median" to evaluate multiple performance metrics. For batting performance, we decided to highlight three measures that we believe were key to understanding a players overall quality and could be explained by our inputs.

First and most importantly, when looking at the relationship between Wins Above Replacement, Outliers, and Years Below Median we found the following results.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5960	0.0356	16.76	0.0000
Outliers	-0.0863	0.0565	-1.53	0.1269
'Years Below Median'	-0.1168	0.0164	-7.11	0.0000

The estimates for both Outliers and Years Below Median indicated a negative effect on WAR as missed time increases by a magnitude of .1. While this result is what we intuitively might have expected, the effect seems to be minimal overall at low values of missed time. Considering 75% of the players in our set had 3 or less years below the median number of games played this result may not be practically significant for the majority of players. However, for players with exceedingly high, four or more, years below the median the effect is quite large.

Looking at one of the building blocks of WAR, OFF, we see a mixed relationship. For reference, according to FanGraphs OFF is a "statistic that combines a position player's total context-neutral value at the plate and on the bases. OFF is a combination of our park adjusted Batting Runs Above Average and our Baserunning Runs above Average and credits a player for the quality and quantity of their total offensive performance during a given period of time" [1].

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.4333	0.2184	-1.98	0.0474
Outliers	0.0600	0.3472	0.17	0.8629
'Years Below Median'	-0.4971	0.1009	-4.93	0.0000

Looking at our table, having a player specific down year seems to have a positive effect, but looking at the significance level it seems as if Outliers should not be considered in the relationship. Years Below Median seems to have a much stronger relationship with a negative effect on offensive output. Given that OFF typically lies between -3 and -.1 we can see Years Below Median would have a practical effect on offensive rating.

Finally, looking at strikeout percentage we see a similar relationship to the previous variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.0327	0.4466	53.81	0.0000
Outliers	2.1294	0.7101	3.00	0.0027
'Years Below Median'	0.5208	0.2063	2.52	0.0117

As players had down years or played below the median number of games, they tended to increase their strikeout percentage. With the majority of strikeout percentage sitting between 18% and 30% over a player's career, we can easily see how missed time could have a practically significant effect on a players performance.

5 | Further Exploration

After analyzing the effect of missed time in the minors on player performance in the majors, we wanted to further analyze a selective group of players with exceptionally noteworthy performance on the majors. We specifically decided to further investigate players that were MVP or Cy Young winners in their leagues. After filtering out the players that won these awards we ended up with 21 players that were MVP's that had minor league records from Fangraphs and 17 Cy Young winners that had minor league records from Fangraphs.

Once we filtered the data we plotted the award winners games played in the minors with a barplot

to analyze if there were any trends or particular distributions that caught the eye. It seemed as if the MVP players had no clear trend in the number of games played in their minor league careers whereas the Cy Young winners have all played the average (30) or higher number of games compared to all of the Major Leagues.

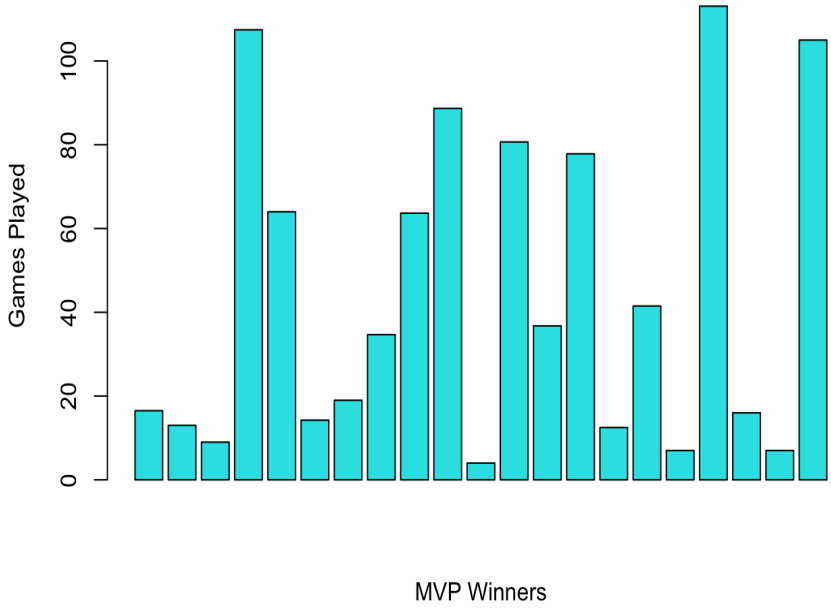


Figure 5.1: Plot of the Number of Games Played by MVP Award Winners

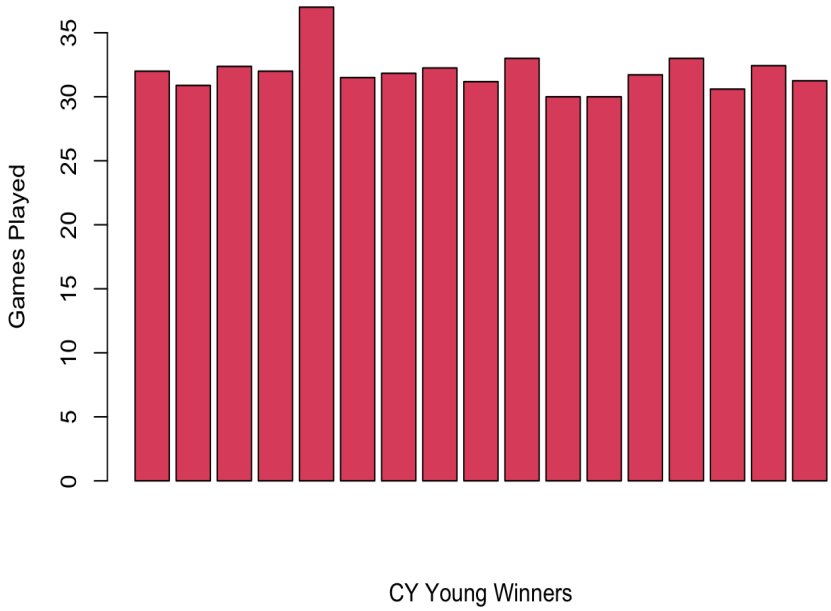


Figure 5.2: Plot of the Number of Games Played by Cy Young Award Winners

By analyzing the barplots of the games played in the minors by the award winners, we wanted to further examine the trends in these exceptional players and decided to with multiple linear regression using our two created inputs "Outliers" and "Years Below Median" to evaluate multiple performance metrics.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
MVP winners Outliers	1	0.02	0.02	0.04	0.8376
MVP winners Years Below Median‘	1	0.17	0.17	0.34	0.5663
Residuals	18	8.71	0.48		

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CY Young Outliers	1	1.85	1.85	0.73	0.4079
CY Young Years Below Median‘	1	2.41	2.41	0.95	0.3469
Residuals	14	35.55	2.54		

After running our multiple regression using our two created inputs "Outliers" and "Years Below Median" we saw p-values of 0.8376 and 0.5663 for the MVP winners and 0.4079 and 0.3469 for the Cy Young winners respectively. Given these results we could conclude that missed time is insignificant in those that go on to become MVP or Cy Young winners.

Once we completed running the multiple regression we ran a test to see if the mean games played in the minors was significantly different from the rest of the league by running a two sided t-test.

Table 5.1: T-Test for MVP Winners

Welch Two Sample T-Test					
Test Statistic	df	P Value	Alternative Hypothesis	Mean of X	Mean of Y
-0.735	20.43	0.4722	Two-sided	44.36	50.43

Table 5.2: T-Test for Cy Young Winners

Welch Two Sample T-Test					
Test Statistic	df	P Value	Alternative Hypothesis	Mean of X	Mean of Y
2.295	51.63	0.02582	Two-sided	31.26	30.29

From our t-tests we got p-values of 0.4722 and 0.02582 for MVPs and Cy Young Winners respectively. Seeing that the t-test gave a p-value of 0.4722, it can be seen that missing time in the minors has no significant difference in means from the rest of the league's missed time in the minors. On the other hand we see that the t-test gave a p-value of 0.02582 for Cy Young winners, and hence rejects that the mean games played in the minors are not significantly different then the mean games played in the minors for the rest of the league. From this finding we see that Cy Young winners on average have a higher number of games played in the minors from the rest of the league. As a result it is seen that the on average more games played in the minors can increase the chances of someone winning the Cy Young.

6 | Limitations

Although our analysis has provided some valuable insights in the relationship between minor league experience and major league performance, there are some important considerations and limitations that should be taken into account when interpreting the findings.

Firstly, it's critical to acknowledge that baseball performance is highly complex and multidimensional. Numerous factors, many of which were not taken into account in this study, such as physical characteristics, the caliber of training, mental toughness, and even luck, all have an impact on it. Therefore, even though our analysis concentrated on the role of missed time in the minors, it is important to recognize that other influencing factors do exist.

Our study's main flaw is its reliance on observational data, which limits our capacity to draw conclusions about causality. It is difficult to establish clear cause-and-effect links, even while we can identify correlations between performance after missing time in the minor leagues and later events. For instance, missed time could be a result of a player's injury, their personal situation, or even management choices, all of which have varying effects on their performance going forward.

Secondly, the study makes the inevitable assumption that circumstances and levels of competition are constant among minor league teams and seasons. Our study does not take into consideration variations in aspects like coaching, infrastructure, team dynamics, and opponents' talent levels, which can have an impact on a player's minor league experiences and later major league performances. Additionally, our analysis relied on aggregate career statistics, which might have overlooked player performance variations from year to year. This strategy could cover up the immediate consequences of lost time, which might be easier to spot through a season-by-season examination.

Furthermore, while the study's concentration on anomalies like MVP and Cy Young winners was illuminating, it also means that our findings are most applicable to elite players. It is not specifically addressed if these findings apply to players who are ordinary or below average. By including more factors in the analysis, and expanding the player pool, future research may be able to overcome these drawbacks. Baseball player development plans and talent management will benefit greatly from a deeper knowledge of the relationships between minor league and major league performance.

7 | Conclusions

Our study aimed to investigate the relationship between baseball players' minor league performance, specifically in terms of missing time, and their following major league performance. Our research produced some fascinating findings. In the first phase of our study, we used multiple linear regression to model the issue utilizing two newly formed inputs, "Outliers" and "Years Below Median." Using this technique, we examined a number of performance metrics, such as offensive rating, strikeout percentage, and WAR.

The results of this preliminary investigation revealed that a player's batting success in the main leagues is affected by missed time in the minor leagues. However, it showed that the impact was only of a moderate size, with players who missed a lot of time in the minors showing a slightly stronger impact. We further investigated this relationship by focusing on two specific groups of major league players: MVP and Cy Young winners. According to the research, failure in the major leagues for MVP winners was not significantly predicted by missing time in the lower leagues. The multiple linear regression analysis's high p-values and the non-significant difference in the means of games played in the minors for MVP winners and the rest of the league were evidence of this.

Contrarily, our examination of Cy Young winners indicated that these players frequently play more minor league games than the rest of the league. The t-test produced a significant p-value, indicating that participating in more minor league games would boost a player's chances of winning the Cy Young award.

In conclusion, our analysis shows that depending on the individual set of players being studied, the association between minor league experiences, particularly in terms of missing time, and major league performance can vary. While playing in the lower leagues does appear to have an effect on performance in major league baseball, the strength of this influence varies. This highlights the requirement for more thorough and diversified research to better comprehend these linkages and the contributing elements. We can only hope that these studies will help us comprehend player growth, performance forecasting, and talent management in baseball on a deeper level.

8 | References

[1] Neil Weinberg. Off. *Sabermetrics Library*, 2014.