5-1-2023

An Examination into Statistical Evaluations of Major League Baseball Coaches and Managers

Jacob Hamilton Zimmerman
Butler Univeristy

BUTLER UNIVERSITY HONORS PROGRAM

Honors Thesis Certification

Please type all information in this section:

Applicant Zimmerman, Jacob Hamilton
(Name as it is to appear on diploma)

Thesis title An Examination into Statistical Evaluations of Major League Baseball Coaches and Managers

Intended date of commencement 05/05/2023

Read, approved, and signed by:

Thesis adviser(s) Mohammad Ehaha A. Patwary — 04/27/2023 (Date)

Reader(s) John E. Hess — 4-28-2023 (Date)

(Date)

Certified by — Director, Honors Program — 5/1/23 (Date)

5-22-2023

An Examination into Statistical Evaluations of Major League Baseball Coaches and Managers

Jacob H. Zimmerman

An Examination into Statistical Evaluations of Major League Baseball Coaches and Managers

Jacob Zimmerman
April 22, 2023

An Undergraduate Thesis

Presented to The Honors Program

of Butler University

Supervised by Mohammad Shaha Patwary, PhD

In Partial Fulfillment

of the Requirements for Graduation Honors

ABSTRACT

An Examination into Statistical Evaluations of Major League Baseball Coaches and Managers

Jacob H. Zimmerman

April 22, 2023

This exploration attempts to create statistical procedure capable of defining and comparing a Major League Baseball Manager's performance in respect to player development. Using a supervised learning method of multiple linear regression, we examine how improvements or deterioration of certain player skills predict an increase in runs scored in a season. Major League Baseball teams (N = 30) over the span of eight seasons since the beginning of the Statcast era in 2015 were evaluated on their ability to improve their team's batters' patience in selecting hittable balls, or pitch selection, and quality of contact with the baseball, and graded on how their team's variation predicts either runs scored, or additional, more traditional offensive metrics that already have well established statistics and research into their run predictiveness. Given the millions of dollars spent on team Manager contracts a year, much less the hundreds of millions spent on player assets said Manager is responsible for developing, more research and data is needed to create more understandable metrics and analytics to better judge a manager's performance.

Contents

Chapter 1

Introduction

During my work with the Butler University Softball team, I recognized that despite the copious available information regarding the production of the players on the field, there are very few analytics that evaluate the manager of these respective teams. Furthermore, the limited statistics we have are centered around in-game management. While this is certainly an important aspect of any team's manager, it is arguably not the most important responsibility. That distinction, for the purposes of this thesis, goes to player development. This thesis is intended to open and explore the conversation around further potential statistical evaluations of player development as a mark of coaching performance.

Hitting a baseball is not easy, its complex biomechanics alone, requiring ideal stride, extension, rotation, bat guidance, and correctly timed weight shifts to generate the amount of force necessary to succeed at even the lower levels of Baseball are daunting (Welch et al, 1995). On top of that, learning to properly swing a bat is merely the bare minimum when attempting to hit even semi-competent pitching. With just 60.5 feet between the pitcher's mound and the batter, the speed at which the pitch arrives, and the time a batter needs to extend his bat over the plate, the MLB hitter typically has just 0.3-0.4 seconds on average to decide whether he wants to swing at a given pitch or not (Quinton, 2017). The ability to assist in not only developing these skills in players at any level, much less create substantial improvement among the world's best hitters against their opponents' best players, who are also receiving development and support from some of the world's best coaches, is an ability unsurprisingly uncommon, difficult, and valuable.

Articles like "Is finding a star nothing but luck?: Quantifying the effectiveness of MLB Player Development" (Aucoin, 2021) mimic much of what I want to accomplish, as it aims statistically to assess a player's development to judge the organization, relative to others, its ability to develop their prospects. This contains concepts that this thesis draws inspiration from, such as tracking player growth or decline from year to year to derive ranking systems for organizations. However, there are several differences. Aucoin refers his grading system to Fangraphs (*Major League leaderboards*) and Baseball America and does so to predict the future player value and assign dollar values to the players. By contrast, I want to evaluate growth from year to year using our own chosen statistics, attempting to link back to coaches' influence on the player's development while assigning run amounts as the grade, as runs are the primary goal of on-the-field results.

This thesis intends to focus on hitting development, this has been explored in the past, using established run creation formulas to analyze how the change in offensive production (Singles, Doubles, Triples, Home Runs, Walks, and Hit By Pitches) reflects on the performance of the manager (Hill, 2009). My thesis intends to do something similar, but rather than using result based statistics (statistics regarding the end result of plate appearances), such as singles, doubles, and home runs, the focus will be on the improvement of skills like plate discipline (being able to determine what pitch will be a ball or strike), rate of contact (the rate at successfully hitting the ball when attempting to hit), and quality of contact (exit velocities of the ball after contact is made, as well as the angle the ball is being hit at) to predict the manager's development influence on the team's performance of said result statistics, or total runs scored by the team over the season.

While it is hoped these statistics can further be applied to minor league development down the road,

and it perhaps could become applicable soon regarding upper-level minor leagues, current research shows that, "a [minor league] player's overall body of work does not become truly telling of Major League potential until he has progressed to the higher levels." (Chandler & Stevens, 2012). Furthermore, further research into alternate evaluations would need extensive care beyond the scope of this thesis, to ensure prospects are not being evaluated on molds and archetypes that do not fit their skill set or potential. In a philosophy hammered home and emphasized greatly in the watershed book Moneyball, with pitchers like Barry Zito, Tim Hudson, and Chad Bradford all were doubted for seemingly relevant measurements like pitch speed, size, and pitch arsenal, but were nevertheless capable of dominating in the major leagues (Lewis, 2003).

With such little apparent correlation to lower-level development, as well as the limited number of seasons many players who make it to the upper-levels play there, the problems and hurdles needed to be solved to apply to this thesis are best left to further development and research beyond the scope of this thesis. However, Conklin's work did provide one piece of very helpful information regarding treating high school draft picks with college draft picks: "The data showed there was not a clear difference between high school and college athlete performance as rookies for both the hitting and pitching categories " (Conklin, 2014). This allows limited concern over accounting for this difference in development patterns of the player once they reach the majors, as the differences between High School draft picks and college draft picks are insignificant. "The fact that the two groups were so equally productive when they arrived in the major leagues shows that in either case, the minor leagues are sufficiently preparing players if they either spent significant time out of high school or a shorter time for the college players" (Conklin, 2014). Thus little attention will be paid to this when attempting to evaluate rookie players' development once reaching the majors.

1.1 Research Question

There is one overarching research question I intend to answer in the pursuit of this thesis. "How can we apply statistical methods to isolate the influence the training staff had on runs scored?" To answer this, I seek to create the best prediction models possible (relative to the limited Statcast era sample size available at this time) to evaluate the relationships between the development/regression of skill indicative statistics and on-field run production. The hope in this is to establish a correlation between the off-field development the organization's coaches and managers provide to the on-field success the team has. Once these provisions are met, either by this thesis or another intending to improve upon it, it will open a vast set of possibilities and opportunities for evaluation and discourse, not just for coaches and managers, but will wrap around to reflecting upon the players once again as well, as new statistics will allow us to better contextualize the statistics we already know and rely on for evaluation.

1.2 Background on Major League Baseball

Baseball is a bat-to-ball sport in which two teams compete to score more points (runs) than their opponent over a set number of turns to try (innings). Runs are scored for each player to safely complete a full counter-clockwise rotation around 4 bases set in a diamond-shaped formation. Baseball is widely considered one of the most analytically inclined sports in the world. With two major paradigm shifts in the role of analytics in Major League Baseball within just the 21st century, those being the early 2000's Moneyball movement and, most relevant to this thesis, the mid-2010 introduction of the Statcast era, it is clear there has never been a more influential era for analytics in baseball.

There are four core methods in Baseball for scoring or preventing runs: Hitting, Pitching, Fielding, and Baserunning. Hitting, which marks the player's ability in a bat-to-ball sport to effectively and productively make contact with the ball using his bat, will be the primary focus of this thesis as Hitting, and its run-preventing counterpart Pitching, are both substantially more impactful in scoring than their other two contemporaries and contain substantially deeper, established and trustworthy analytics and data.

1.3 Definitions of Key Terms

Balls and Strikes are cumulative counters that take place during an at-bat or interaction between a batter and pitcher. A strike zone is an invisible square-shaped area over home plate, to get a strike call without the batter swinging and missing, the pitcher must throw a pitch into some part of this zone. Balls are accumulated whenever the batter does not swing at a pitch that never enters the strike zone. Baseball's rules for strikes are complicated, but for this thesis, can be simplified to whether the batter swings and misses at a ball, doesn't swing at a ball that lands in the strike zone, or swings and makes contact with the ball, but the ball does not land within the playable (fair) territory. If a hitter accumulates four balls throughout the same at-bat, they are granted a 'free pass' to first base. Meanwhile, if a pitcher can accumulate 3 strikes during the same at-bat, the at-bat is over and the batter is out. Outs are the measure used to determine when it is time for teams to switch their turn hitting. Once a team accumulates 3 outs in the same inning, their turn to bat is over and the two teams switch. Given only the current hitting team can score, it is advantageous for hitters to avoid creating outs, while advantageous for pitchers to create as many, as quickly, as possible.

Swinging Strike % is the percentage of total pitches seen that result in a batter attempting to hit the ball by swinging and completely missing the ball. This is not to be confused with Whiff%, which tracks the miss percentage of just pitches swung at and was not used in data models to avoid multicollinearity and redundancy among the variables. Inside Zone Swing% and Outside Zone Swing% mark the percentage of balls inside and outside the strike zone, that the batter attempted to swing at. These are not to be confused with Outside Zone Contact% and Inside Zone Contact%, which mark the percentage of swing attempts that occurred inside and outside the strike zone respectively that resulted in contact made with the pitch. Ball In Play%, are the percentage of at-bats that end with the baseball being hit and put into play. Called Strike% is the percentage of total pitches thrown that were not swung at by the batter but were called a strike by the umpire. Mean Exit Velocity is the mean speed at which the ball leaves the bat upon contact, with Max Exit Velocity being the hardest ball hit that season according to Exit Velocity. Mean Exit Velocity should not be confused with Hard Hit%, which tracks the percentage of batted balls that met or exceeded an Exit Velocity of 95 mph. Finally, Barrel% is the percentage of batted balls that were Barreled. A Barrel is a batted ball event where the contacted ball is hit with an ideal combination of Exit Velocity and Launch Angle to predict a high offensive production outcome, based on the outcome of similarly hit baseballs in the Statcast era.

Chapter 2

Data and Methodology

2.1 Data

2.1.1 Data for Provision 1

Number of observations: 240 (30 teams, over 8 usable Statcast seasons)

Response Variable: Total runs scored per 6,138 plate appearances (6,138 being the league average number of plate appearances per team)

Type of Response Variable: Numerical and Continuous

Predictor Variables: Swinging Strike Percentage, Inside Zone Swing Percentage, Outside Zone Swing Percentage, Inside Zone Contact Percentage, Outside Zone Contact Percentage, Called Strike Percentage, Ball in Play Percentage, Barrell Percentage, Mean Exit Velocity, Max Exit Velocity, Hard Hit Percentage.

Type of Predictor Variables: All listed predictor variables are numerical

2.1.2 Data for Provision Two

Number of observations: 240 (30 teams, over 8 usable Statcast seasons)

Response Variable: Total singles per 6,138 plate appearances, total doubles per 6,138 plate appearances, total triples per 6,138 plate appearances, total home runs per 6,138 plate appearances, total walks per 6,138 plate appearances, and total hit by pitches per 6,138 plate appearances (6,138 being the league average number of plate appearances per team)

Type of Response Variable: Numerical and Continuous

Predictor Variables: Swinging Strike Percentage, Inside Zone Swing Percentage, Outside Zone Swing Percentage, Inside Zone Contact Percentage, Outside Zone Contact Percentage, Called Strike Percentage, Ball in Play Percentage, Barrell Percentage, Mean Exit Velocity, Max Exit Velocity, Hard Hit Percentage.

Type of Predictor Variable: All listed predictor variables are numerical

2.2 Preprocessing

While traditionally, data preprocessing, which is the craft of pulling, cleaning, and formatting data from its source for proper use, tends to be one of the most challenging and time-consuming parts of data analysis, baseball's treasure-trove of deep, detailed, and organized statistics makes this step relatively painless, and all the data needed to study and evaluate could be found on several different free websites. Beyond just the data available, the process of organizing it was painless, FanGraphs was capable of easily organizing and exporting the exact yearly data needed for this thesis. This is not to say the data collection process was convenient, as several issues needed addressing.

Just as COVID-19 managed to affect many datasets inconveniently, the MLB's shortened 2020 season of just sixty games (as opposed to the typical 162) created problems when comparing the cumulative stats among seasons. To account for this, each season's run-scoring, as well as all other predicted offensive production, were weighted on a scale of 6,138 Plate

Appearances, which is the league's mean number of plate appearances per team in a season (excluding 2020's shortened season). While this does duplicate 2020's data, it does not have any effect on the linear regression beyond encouraging the data to be viewed as rate data rather than cumulative. Data also need to be altered by removing commas on data eclipsing one thousand to correctly be processed in the data frame.

Ultimately, the biggest conundrum in the data preprocessing stage was the very lack of data to preprocess. With just thirty teams to monitor and eight seasons since the creation of many of the variables being used, that leaves a mere 240 responses to base the advanced analytical models on. While said model will likely possess enough data to create models of some tangible merit, the size of the dataset prevents us from better considering how different ballpark environments affect the model by not providing enough data to allow for models tailored to each team. Dataset size is certainly something to be and will be improved upon with time. Given that players cannot be directly linked to total runs scored in the way teams are, the only short-term solution to increase the volume of the dataset would be to break up individual seasons into their respective sections. While this approach does have potential upsides, it comes with too strong risk of both duplication of the database as well as compromising an already small sample size by exposing each data point to inconsistent noise variables such as weather, temperature, player fatigue, or any other additional variable unique or more prominent to one half of a season that may significantly affect performance.

2.3 Methodology Description

Once the data preprocessing phase was complete, I reviewed and analyzed the two hypotheses described in the following chapter. This was accomplished using multiple linear regression (Mendenhall & Sincich, 2020). In this technique, the response variable, the runs scored per 6138 plate appearances, as well as the other response variables in their respective model, was plotted against the eleven predictor variables in the model. Multiple linear regression, predicting Y, can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \boldsymbol{\varepsilon}$$

Here, Y represents each respective response variable within their model, where X_1 represents Swinging Strike Percentage, X_2 represents Inside Zone Swing Percentage, X_3 represents Outside Zone Swing Percentage, X_4 represents Inside Zone Contact Percentage, X_5 represents Outside Zone Contact Percentage, X_6 represents Called Strike Percentage, X_7 represents Ball in Play Percentage, X_8 represents Barrell Percentage, X_9 represents Mean Exit Velocity, X_{10} represents Max Exit Velocity, X_{11} represents Hard Hit Percentage and the epsilon (ε) is an additional factor for model error. Assumptions of epsilon (ε) include the independence of both individual observations of ε as well as X, the epsilon is normally distributed and with a mean of zero.

These variables are then whittled down to just those that have a t-value probability that did not exceed (was not greater then) .10, or 10%, through supervised learning. 'Supervised' means there was a response variable used against the tested variables to help determine relationships between said variables. This model assumes the four assumptions regarding the multi-linear regression models are met, which will be confirmed as true three chapters from now.

Multiple linear regression models are used in relevant statistical analysis. The model is known for its ease of both execution and interpretation. Its various summary statistics, that support or challenge its predictive effectiveness are also easily understandable to those both with and without extensive backgrounds in statistical models. Furthermore, it can accurately create and explain linear relationships between a variable(s) within the given dataset without compromising its ability to predict unseen data outside the provided sample, also called overfitting.

As opposed to factors like kNN nearest neighbor approaches, and other tests considered early on, linear regression models will more easily apply to the numeric, scalable, decimal data desirable to achieve the desired estimated run statistics the data is being created for.

Lastly, with our new prediction models for Provision 2 secured, we will incorporate established run-creating statistics, in this case, wOBA (weighted On Base Average) and its related descendent statistic wRAA (weighted Runs Above Average), to create our run estimator regarding runs created through player development. From there, we can use changes in the predictor variables from year to year of both overall team changes, as well as changes among individual players, to create run estimations for player development.

Chapter 3

Hypotheses

3.1 Hypotheses for Provision One

Statement of Hypothesis One: *There is no significant linear relationship between any of the eleven predictor variables and Runs Scored per 6138 Plate Appearances.*

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{10} = \beta_{11} = 0$

$H_1 :$ *At least one* $\beta_i \neq 0; i = 1, \ldots, 11$

3.2 Hypothesis for Provision Two

This Hypothesis will be repeated for each of the six variable models (Singles, Doubles, Triples, Home Runs, Walks, and Hit By Pitches)

Statement of Hypothesis Two: *There is no significant linear relationship between any of the eleven predictor variables and the given response variable.*

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{10} = \beta_{11} = 0$

$H_1 :$ *At least one* $\beta_i \neq 0; i = 1, \ldots, 11$

Chapter 4

Statistical Analysis Results

4.1 Statistical Analysis Results for Provision One

Using the base model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \boldsymbol{\varepsilon}$$

Using the t-test, we can test the probability of set variables influencing the regression line of the model. By taking the observed Coefficient Estimate, and dividing it by the observed Standard Error, we obtain a t-value that, when used on a normal probability "Bell Curve", we obtain the likelihood of such Coefficient variance being random. Using 10% significance (0.10), we reject variables that contain $P(t>|t_{obs}|)$ values of less than 0.10 as not statistically significant.

For this model, there are a total of four predictor variables, each of which being quantitative, that were statistically significant to the model, based on each variable's p-value of their outputted t-tests found when running multiple regression models in R-Studio. At 10% significance, these four predictor variables had substantial evidence to support a linear relationship with the response variable.

Variables	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1901.4	492.2	3.863	0.000154

X_2	-1865.6	494.4	-3.774	0.000216
X_6	-2570.7	1006.5	-2.554	0.011439
X_7	444.5	192.9	2.304	0.022303
X_8	3567.2	390.3	9.140	$< 2e^{-16}$

$R^2 = 0.3663, \quad R^2_{Ad} = 0.3527, \quad F = 27.02, \quad p - val = < 2.2e^{-16}$

Therefore, drawing from hypothesis one, the conclusion drawn is that, at 10% significance, we reject the null hypothesis that there is not a significant linear relationship between any of the predictor variables and Runs Scored per 6138 Plate Appearances.

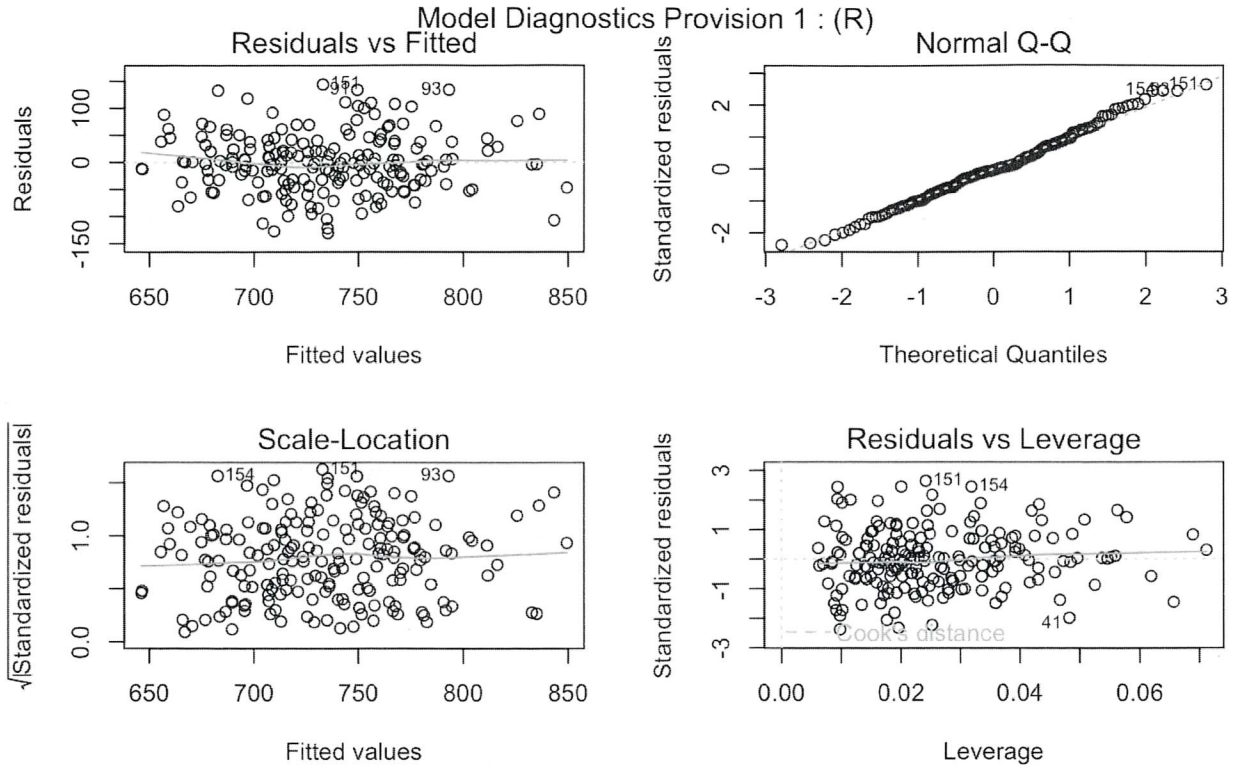
The final model, with the parameter estimates included, are as follows:

$\hat{Y}_{Runs} = 1901.4 - 1865.6*X_2 - 2570.7*X_6 + 444.5*X_7 - 3567.2*X_8$

However, while these four variables did show substantial significance in relation to the response variable, using an eighty-twenty Train-Test split to test the model's effectiveness at predicting the response variable, we found that after analyzing the r-squared value of the model, only approximately 24% of the variation in the response variable could be explained by the model. For this reason, this model will not be included in our further statistics development.

4.2 Model Diagnostics for Provision One

Before acting further on my findings, I first confirmed that my models meet the four assumptions of a multiple linear regression model. These four assumptions are critical to checking and confirming our overall assumption that a linear relationship between our response and explanatory variables is a valid depiction of said interaction. Here is each assumption we made when creating and running the model.



- The mean of the residuals is approximately zero. Residuals are the differences between the fitted values and observed values of the response variable. Checking by generating a residual plot and plotting the residuals against the individual values of the response variable, this assumption was met.

- The variance of the residuals is constant over all fitted values of the response variable. This means the residuals do not fluctuate substantially as the value of the response variable increases. The term for this assumption is homoscedasticity, in which the variance of the residuals is constant over all fitted values of the response variable. This assumption was checked utilizing a scale-location plot, which demonstrates whether residuals are spread across the range of predictors equally, with this test, we see the assumption is met.

- The distribution of the residuals follows a normal distribution. In other words, the residuals

are normally distributed with a mean of zero and a constant variance over all fitted response variable values. Using a normal probability plot in which a linear pattern of data points indicates a normally distributed set of residuals and where the axes of the plot are the standardized quantiles and the theoretical quantiles, this assumption was met.

- There is no multicollinearity present in the model. Multicollinearity is a situation where one or more predictor variables have their own independent linearly predictive relationship. While this does not necessarily jeopardize the predictive power and trustworthiness of the complete model, the validity and accuracy of individual predictor variables within the model will be under severe scrutiny. To check this assumption, I utilized the Variance Inflation Factor (VIF) between each set of predictor variables. Using the commonly accepted VIF limit of ten, which means any variable exceeding said VIF value indicates multicollinearity present in the model. I was able to conclude that since no predictor variables exceeded the VIF limit, no multicollinearity was present in the model.
- Cook's distance was also utilized to confirm there were no individual data points that substantially affected the regression model.

4.3 Statistical Analysis Results for Provisions Two

Using the base model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \boldsymbol{\varepsilon}$$

With Y representing the respective model being tested (Singles, Doubles, Triples, Home Runs, Walksm, and Hit By Pitches)

Using the t-test, we can test the probability of set variables influencing the regression line of the model. By taking the observed Coefficient Estimate, and dividing it by the observed Standard Error, we obtain a t-value that, when used on a normal probability "Bell Curve", we obtain the likelihood of such Coefficient variance being random. Using 10% significance (0.10), we reject variables that contain Pr(>|t|) values of less than 0.10 as mere chance.

For this model, there are a total of three predictor variables for Singles, two predictor variables for Doubles, two predictor variables for Triples, five predictor variables for Home Runs, six predictor variables for Walks, and two predictor variables for Hit By Pitches, each of which is quantitative, that were statistically significant to their respective model, based on each variable's p-value of their outputted t-tests found when running multiple regression models in R-Studio. At 10% significance, these predictor variables had substantial evidence to support a linear relationship with their respective response variable.

Singles:

Variables	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1098.4	344.0	3.193	0.00165
X_2	-1333.3	325.2	-4.099	$6.16e^{-05}$

X_6	-3183.7	609.152	-4.638	$6.58e^{-06}$
X_7	1832.7	101.8	18.000	$< 2e^{-16}$

$R^2 = 0.6645,$ $R^2_{Ad} = 0.6591,$ $F = 124.1,$ $p - val = < 2.2e^{-16}$

Doubles:

Variables	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	148.14	29.56	5.012	$1.24e^{-06}$
X_8	-962.11	244.20	-3.940	0.000115
X_{11}	523.10	117.99	4.434	$1.57e^{-05}$

$R^2 = 0.09432,$ $R^2_{Ad} = 0.08474,$ $F = 9.842,$ $p - val = 8.588e^{-05}$

Triples:

Variables	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.376	19.079	0.334	0.7386
X_2	60.602	29.080	2.084	0.0385
X_8	-329.202	39.697	-8.293	$2.03e^{-14}$

$R^2 = 0.2676,$ $R^2_{Ad} = 0.2598,$ $F = 34.52,$ $p - val = 1.663e^{-13}$

Home Runs:

Variables	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	823.967	350.523	2.351	0.01979
X_2	-1063.142	228.670	-4.649	$6.30e^{-06}$
X_6	-1909.289	465.896	-4.098	$6.21e^{-05}$
X_7	-271.057	93.121	-2.911	0.00405
X_8	1411.784	228.825	6.170	$4.18e^{-09}$
X_9	5.602	3.196	1.753	0.08131

$R^2 = 0.5665, \quad R^2_{Ad} = 0.5548, \quad F = 48.6, \quad p - val = < 2.2e^{-16}$

Walks:

Variables	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4171.572	600.946	6.942	$6.34e^{-11}$
X_2	-945.932	310.380	-3.048	0.002644
X_3	-2126.086	149.178	-14.252	$< 2e^{-16}$
X_6	-2345.012	660.806	-3.549	0.000491
X_7	-849.038	108.420	-7.831	$3.65e^{-13}$
X_9	-19.549	5.915	-3.305	0.001141
X_{11}	908.557	183.113	4.962	$1.58e^{-06}$

$R^2 = 0.7373, \quad R^2_{Ad} = 0.7288, \quad F = 86.55, \quad p - val = < 2.2e^{-16}$

Hit By Pitches:

Variables	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	585.345	149.287	3.921	0.000123
X_8	580.584	95.684	6.068	$6.94e^{-09}$
X_9	-6.330	1.733	-3.653	0.000336

$$R^2 = 0.1634, \quad R^2_{Ad} = 0.1545, \quad F = 18.46, \quad p - val = 4.766e^{-08}$$

Therefore, drawing from hypothesis two, the conclusion drawn is that, at 10% significance, we reject the null hypothesis that there is not a significant linear relationship between any of the predictor variables and one of the six tested response variables (Singles, Doubles, Triples, Home Runs, Walks and Hit By Pitches).

The final models, with the parameter estimates included, are as follows:

$\hat{Y}_{Singles} = 1098.4 - 1333.3*X_2 - 3183.7*X_6 + 1832.7*X_7$

$\hat{Y}_{Doubles} = 148.14 - 962.11*X_8 + 523.10*X_{11}$

$\hat{Y}_{Triples} = 6.376 + 60.602*X_2 - 329.202*X_8$

$\hat{Y}_{Home\ Runs} = 823.967 - 1063.142*X_2 - 1909.289*X_6 - 271.057*X_7 + 1411.784*X_8 + 5.602*X_9$

$\hat{Y}_{Walks} = 4171.572 - 945.932*X_2 - 2126.086*X_3 - 2345.012*X_6 - 849.038*X_7 - 19.549*X_9 + 908.557*X_{11}$

$\hat{Y}_{Hit\ By\ Pitches} = 585.345 + 580.584*X_8 - 6.330*X_9$

However, while each model contained predictor variables with a substantial significance to the response variable. Using an eighty-twenty Train-Test split to test the model's effectiveness at predicting the response variable, we found that after analyzing the r-squared value of the model, the variation in the response variable that could be explained by the model varied with each model. For models such as Singles, Home

Runs, and Walks, 54%, 51%, and 74% of the data respectively was explained by the model, while models such as Doubles, Triples, and Hit By Pitches, 0%, 16% and 6% of the data respectively was explained by the model. For this reason, further statistical development will only include models of singles, home runs, and walks.

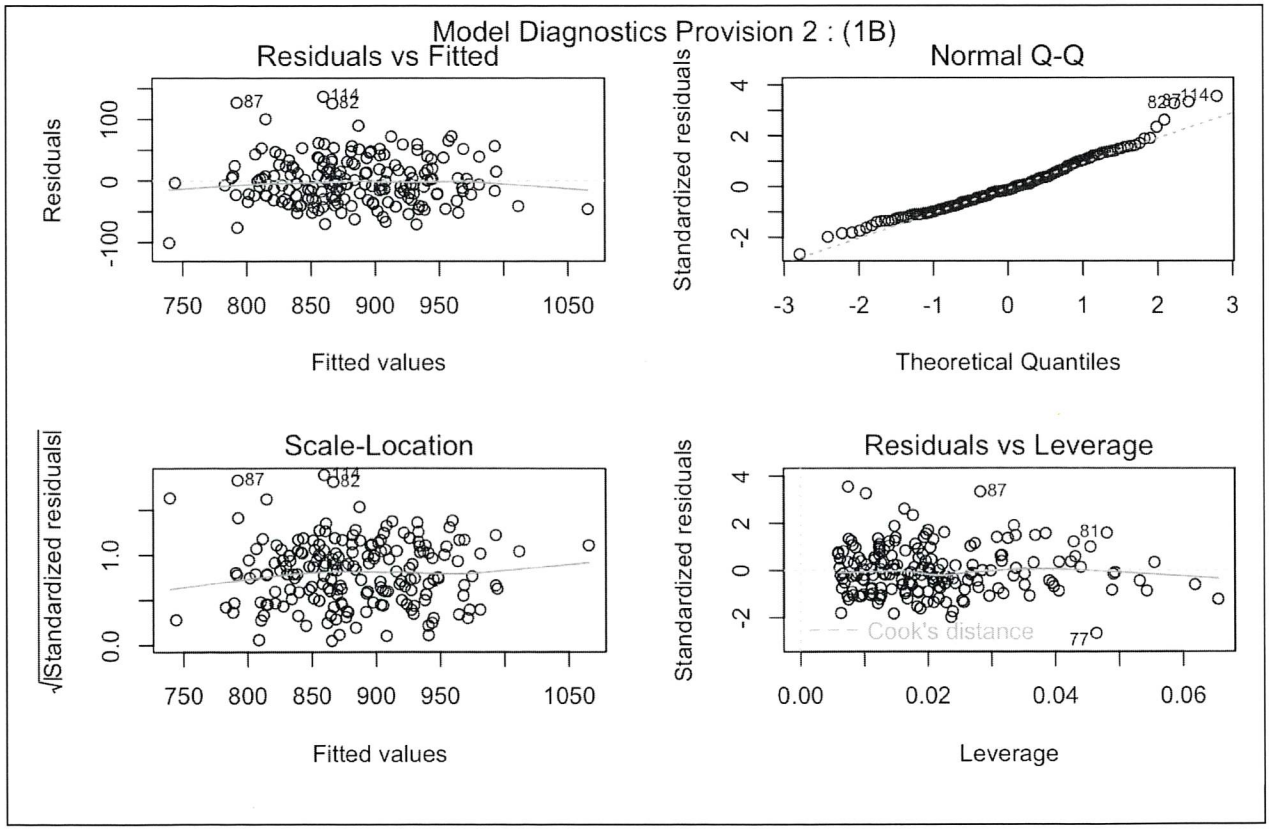
The models predicting Triples and Hit By Pitches fails to account for a substantial percent of their data is unsurprising. This is possibly due to a variety of factors such as: A low sample of occurrences per season, large year-to-year variance among those occurrences, and additional factors like footspeed or position the batter places themselves relative to home plate, (or in other words, do they stand exceptionally close to, or 'crowd', the plate). Doubles, however, have less obvious explanations for their failure.

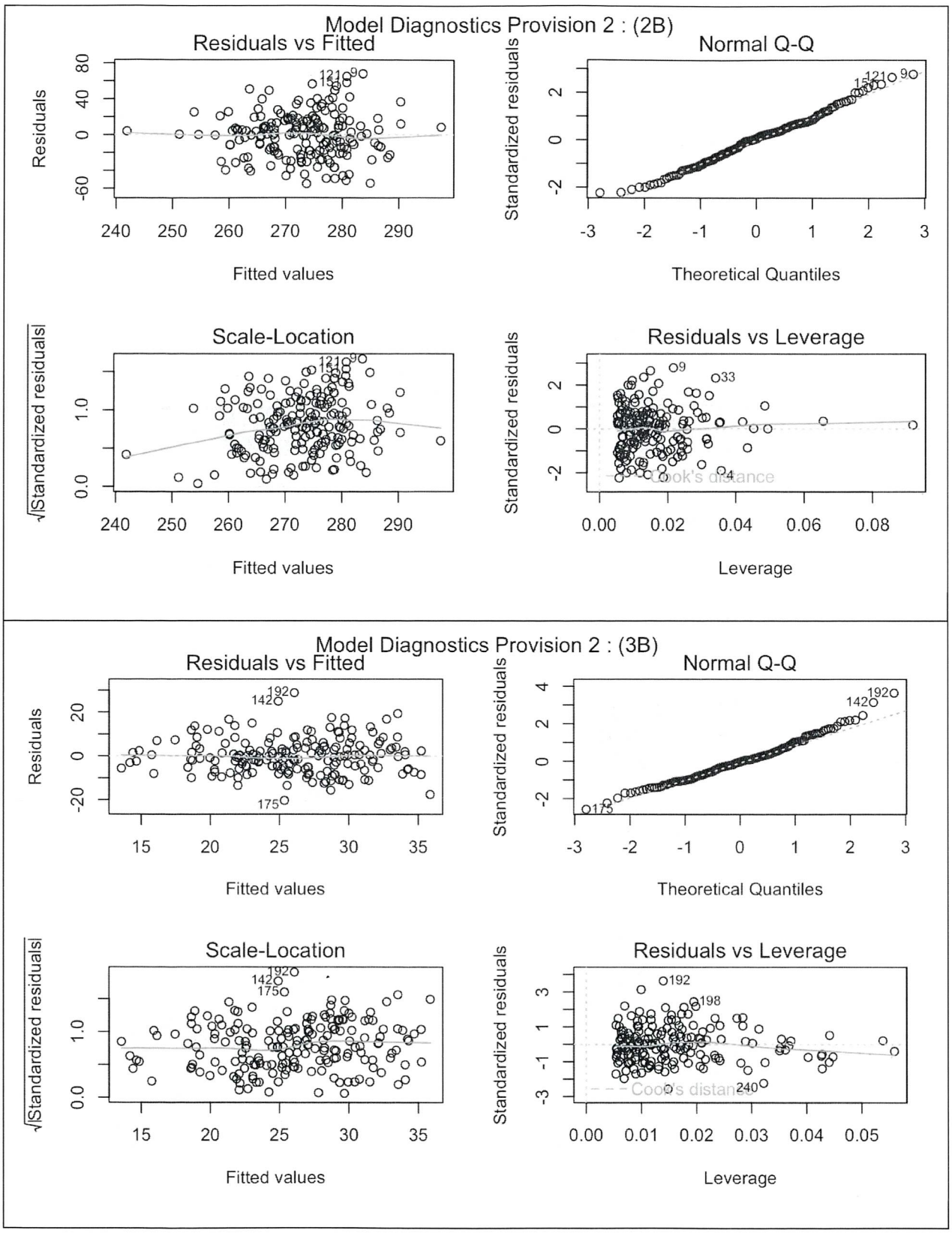
One potential explanation is the rate at which doubles are hit could be strongly linked to balls hit into the outfield that were hit notable distances away from said outfielders. While this model takes into account the vertical angle at which the ball was hit with power with Barrel%, the model did little to account for horizontal angles at which the ball would be hit. Further factors outside the hitting skills such as batter agility and fielder agility could have a far more substantial influence on this statistic than previously thought (Barker, 2013).

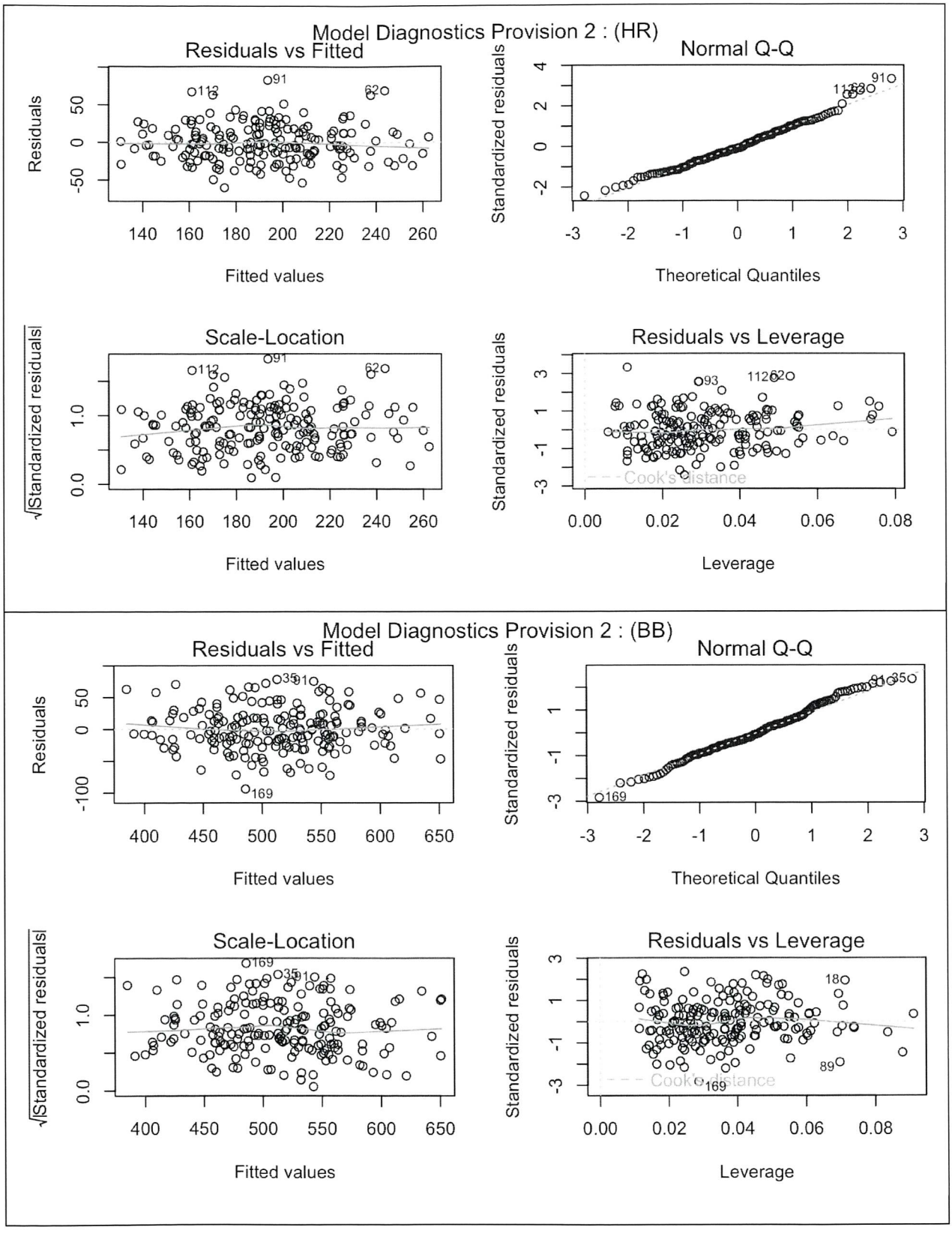
4.4 Model Diagnostics for Provisions Two

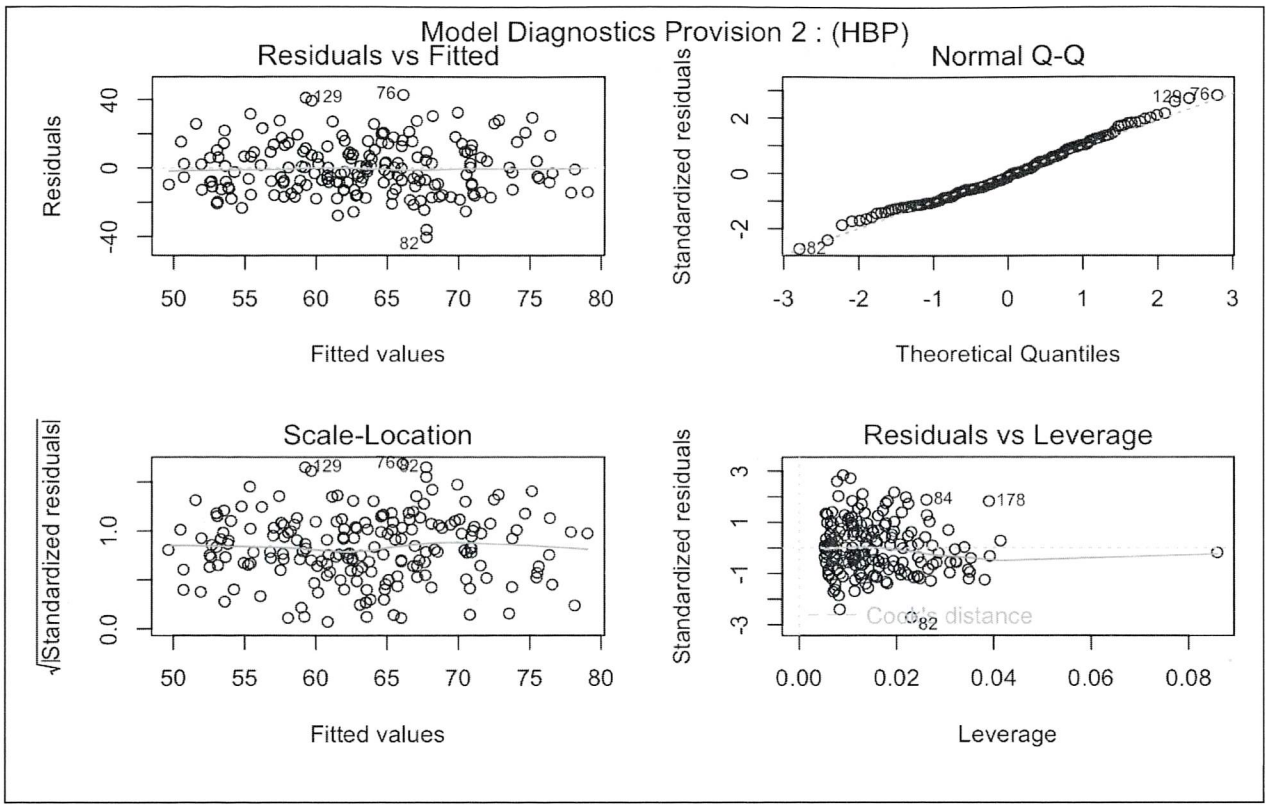
Before acting further on my findings, I first confirmed that my models meet the four assumptions of a multiple linear regression model. These four assumptions are critical to checking and confirming our overall assumption that a linear relationship between our response and explanatory variables is a valid depiction of said interaction. Here is each assumption we made when creating and running the model.

Below are each of the graphs for the Provision 2 models, each of the six models were tested and found to have met the same assumptions.




Model Diagnostics Provision 2 : (2B)
Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
Model Diagnostics Provision 2 : (3B)
Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Model Diagnostics Provision 2 : (HR)
Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
Model Diagnostics Provision 2 : (BB)
Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage



- The mean of the residuals is approximately zero. This assumption was met.
- The variance of the residuals is constant over all fitted values of the response variable. The assumption of homoscedasticity was met.
- The distribution of the residuals follows a normal distribution. This assumption was met in all six models.
- Hard Hit % in both the Triples Model and Home Runs model exceeded the VIF limit (10) of acceptable Multicollinearity and thus had to be removed. Once this action was taken, the assumption of no multicollinearity present in all the models was met.
- Cook's distance was also utilized to confirm there were no individual data points that substantially affected the regression models.

4.5 Application of Provision Two

Using the 3 predictor models (1B, HR, and BB) that met our threshold, we will examine the projected run creation from the expected increase or decrease of these run-producing statistics. To accomplish this, we can turn to established and accredited formulas: wOBA (weighted On Base Average) and wRAA (weighted Runs Above Average) to convert our projected changes in Singles, Home runs, and Walks respectively into runs scored on the field.

As previously discussed, wOBA is a run estimation tool that projects the total run production of a player/team/league on a per-PA basis. When paired with wRAA, we can estimate the number of runs above or below average throughout the total Plate Appearances taken. wOBA takes all 6 previously established run-producing stats and creates weights for how well that statistic predicts runs in a given season and, with the help of the wRAA statistic, transforms the data into an estimate of how many runs were created above or below average compared to a league average hitter.

The wOBA formula for 2022 (individual scales for each variable adjusts every year) was as follows:

wOBA = (0.689*uBB + 0.720*HBP + 0.884*1B + 1.261*2B + 1.601*3B + 2.072*HR) / (AB + BB – IBB + SF + HBP)

By isolating the 3 aforementioned variables, we acquire each season's league average wOBA from just these three respective variables. By having this baseline, we can use the wRAA formula to convert deviations from this mean to project runs scored above or below the league average.

The wRAA formula is as follows:

wRAA = ((wOBA – league wOBA) / wOBA scale) * PA

wOBA scale is merely a year-to-year value that sets wOBA on the same scale as the league average OBP (On Base Percentage), no additional work was needed beyond updating the pre-established statistics for the respective tested season.

Now that we have the projected runs created from our projected Singles, Home runs and Walks, we can monitor how those runs above or below league average grow or shrink over time. This way, we can see the team's development or regression of our 11 original predictor variables led to an increase or decrease in projected runs scored in the season.

This process is applicable not just to the team statistics. Given judging a team doesn't account for the changing personnel on the team from one year to the next, a far superior method would be to evaluate individual players that took substantial plate appearances (five hundred being the chosen number, as this is roughly the number of plate appearances needed to qualify to lead the league in rate stats) from one year to the next. While this process does not account for everything, such as how a typical player aging would affect their development and rate of decline (Schulz et al, 1994), or potential injuries affecting on-field performance, both making it difficult to compare various coaching performances, it does supply us with strong individual trackers that may be able to validate or contradict team-wide trend of development.

Due to COVID-19, the 2020 baseball season only having sixty games opposed to its typical 162 created problems when applying wRAA, as this is a metric that takes the difference between the league average wOBA, the individual or teams, and multiplies it by the number of plate appearances (or in similar terms, takes how many more or fewer runs you create per plate appearance, and multiplies it by the number of plate appearances). This proves to problematically skew the data,

something extra troublesome when the wRAA must be compared to previous and following seasons to derive value. The solution for the team statistic was the same as the previous solution for dealing with the COVID-19 data, adjusting all the wRAA stats to a per 6,138 plate appearances basis. For the individual player development evaluations, since no players were able to reach the five-hundred plate appearance threshold, their seasons were deemed too small a sample size and were excluded from consideration.

Chapter 5

Future Research

With just eight seasons since the start of baseball's Statcast era, each season will provide new, crucial data to better develop these models. With this new information, however, there is also potential for said models to become obsolete due to upcoming 2023 rule changes that look to shift game dynamics permanently moving forward. Starting in this upcoming 2023 season, the MLB's ban on the use of defensive shifting will take effect. Shifting is the practice of strategically realigning fielders away from a 'typical' placement, to increase the likelihood of fielding batted balls from hitters with strong batted ball tendencies. This represents both a massive change to the recent dynamics of the sport, as well as a potential to render this thesis obsolete. Variables such as Hard Hit %, Mean Exit Velocity, and Maximum Exit Velocity could see a substantial increase in influence upon the models created for this Thesis, particularly for our Singles regression model, as shifting was a common tactic to counter many hard-hitting left-handed hitters that had extreme tendencies to pull the batted ball towards their side of the batter's box. Future research should keep a close eye on how these upcoming changes affect the model, and whether the current data is usable moving forward. Additional rule changes like the institution of the universal DH, a rule that requires both the National and American league to use Designated Hitters as a replacement for pitchers in team lineups, leading to a higher offensive run environment in leagues with DHs, in 2020, 2022 and also present problems as only 25% of seasons examined in the study reflect this environment, which is expected to continue in baseball into the foreseeable future.

It should be noted that while this development focuses on on-field production, some studies have

suggested team development and the rise of star power contains "some intangible element beyond the performance measures." (Lewis & Yoon, 2016). This suggests that the development of players, particularly once they enter the upper tier of stardom in the sport, creates additional run-based benefits for their team beyond their performance on paper. This is to suggest there are potential additional factors that could be incorporated into this thesis that could explore these underlying runs created and lost as part of an evaluation of the organization's coaching development.

Further development into this application on pitching will allow for more appropriate samples of data and statistical insights to be found. However, it should be taken cautiously. Pitching injuries have gained the reputation of having a much higher rate of derailing careers, and data used to try to account for such should heed careful consideration and scrutiny, and it is "recommend[ed] against utilizing nonvalidated statistical measures to assess performance after injury, as they demonstrated unacceptably high variability even among healthy, non-injured professional baseball pitchers" (Pareek et al, 2021). This wrinkle and struggle to account for instances of outliers and biased sampling unsuitable for its intended use will present a challenging endeavor. Future research into pitching applications of this process would potentially see even more beneficial results as surface level pitching statistics have shown to be poor predictors themselves of quality pitching performance, "The ERA estimators that were tested (xFIP, FIP, SIERA and tERA) all did a better job of predicting future ERA than actual ERA" (McDaniel, 2012).

Beyond this, additional factors that were avoided specifically due to their immediate redundance created by the upcoming rule changes can now be introduced, such as horizontal launch angles, which could both benefit all models, but particularly help improve doubles and triples. Ultimately, being able to create and monitor these stats in the minor leagues would be another logical next step, as player development is even more relevant and pronounced at the levels attempting to improve to reach the Major Leagues.

Additional avenues of research for the betterment of this thesis would be to examine the potential adverse effect of MLB's various undisclosed baseball (the ball itself) alterations. While the MLB has repeatedly denied intentionally 'juicing' the balls, to aid in additional league-wide offense, several times within the eight-season span of this thesis they have been accused of doing so (Passan, 2019)(Rymer, 2022), as well as admitting themselves to using two different baseballs in 2021, something that the "players claim they had no idea [about]" (Cwik, 2021). These factors additionally jeopardize the validity of this study as it pertains to seasons going forward, as well as present a frustrating burden to anyone attempting to do future work as well, assuming these periodic 'alterations' with ball dynamics should continue.

Furthermore, as previously mentioned, further development into applying this process to minor league development would provide a substantial boost to its potential application to Major League Baseball.

Bibliography

Aucoin, D. (2021, July 11). *Is finding a star nothing but luck?: Quantifying the effectiveness of MLB Player Development*. Driveline Baseball. Retrieved March 4, 2023, from https://www.drivelinebaseball.com/2019/04/finding-star-nothing-luck-quantifying-effectiveness-mlb-player-development/

Barker, D. G. (2013, March 17). *The factor structure of Major League Baseball Records*. Taylor & Francis. Retrieved April 3, 2023, from https://www.tandfonline.com/doi/abs/10.1080/10671188.1964.10613280

Chandler, G., & Stevens, G. (2012, November 12). *An exploratory study of Minor League Baseball statistics*. De Gruyter. Retrieved April 3, 2023, from https://www.degruyter.com/document/doi/10.1515/1559-0410.1445/html

Conklin, K. (2014). *The role of a player development system in Major League Baseball*. Fisher Digital Publications. Retrieved April 3, 2023, from https://fisherpub.sjf.edu/cgi/viewcontent.cgi?article=1005&context=sport_undergrad

Cwik, C. (2021, November 30). *MLB reportedly used two different baseballs last season*. Yahoo! Sports. Retrieved April 3, 2023, from https://sports.yahoo.com/mlb-reportedly-secretly-used-two-different-baseballs-last-season-215050389.html

Hill, G. (2009). *The effect of frequent managerial turnover on organizational performance: A study of professional baseball managers.* American Psychological Association. Retrieved April 3, 2023, from https://psycnet.apa.org/record/2009-11894-009

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

Lewis, M., & Yoon, Y. (2016, March 7). *An empirical examination of the development and impact of star power in ...*Sage Journals. Retrieved April 3, 2023, from https://journals.sagepub.com/doi/10.1177/1527002515626220

Major League leaderboards " 2022 " batters " Custom statistics: Fangraphs baseball. FanGraphs. (n.d.). Retrieved April 3, 2023, from https://www.fangraphs.com/leaders.aspx?pos=all&stats=bat&lg=all&qual=0&type=c,6,12,110,103,102,106,105,313,212,308,305,309,311,8,9,10,11,14,17,5,15,18&season=2022&month=0&season1=2022&ind=0&team=0,ts&rost=0&age=0&filter=&players=0&startdate=&enddate=

McDaniel, R. J. (2012, September 19). *Should we be using era estimators during the season?* The Hardball Times. Retrieved April 3, 2023, from https://tht.fangraphs.com/should-we-be-using-era-estimators/

Mendenhall, W., & Sincich, T. (2020). *A second course in statistics: Regression Analysis*. Pearson.

Pareek, A., Parkes, C. W., Leontovich, A. A., Krych, A. J., Conte, S., Steubs, J. A., Wulf, C. A., & Camp, C. L. (2021, November 18). *Are baseball statistics an appropriate tool for assessing return to play in injured pitchers? analysis of statistical variability in healthy players*. Orthopaedic journal of sports medicine. Retrieved April 3, 2023, from https://pubmed.ncbi.nlm.nih.gov/34820461/

Passan, J. (2019, July 8). *Verlander: MLB juicing balls for more offense*. ESPN. Retrieved April 3, 2023, from https://www.espn.com/mlb/story/_/id/27149029/verlander-mlb-juicing-balls-more-offense

Quinton, S. (2017, March 31). *Don't blink: The science of a 100-MPH fastball*. The Seattle Times. Retrieved April 3, 2023, from https://projects.seattletimes.com/2017/mariners-preview/science/

Rymer, Z. D. (2022, April 29). *MLB's ball controversy could define the 2022 season*. Bleacher Report. Retrieved April 3, 2023, from https://bleacherreport.com/articles/10034259-mlbs-ball-controversy-could-define-the-2022-season

Schulz, R., Musa, D., Staszewski, J., & Siegler, R. S. (1994). *The relationship between age and major league baseball performance: Implications for development.* American Psychological Association. Retrieved April 3, 2023, from https://psycnet.apa.org/record/1994-39406-001

Welch, C. M., Banks, S. A., Cook, F. F., & Draovitch, P. (1995). *Hitting a baseball: A biomechanical description*. The Journal of orthopaedic and sports physical therapy. Retrieved April 3, 2023, from https://pubmed.ncbi.nlm.nih.gov/8580946/