Baseball Player Investment: How to Import Foreign Hitter Correctly from Major League Baseball to Nippon Baseball League

Ying-Wei, Hung (n10733850)

IFN703/4 Assessment 3

Executive Summary

Baseball player investment is always a difficult job in every baseball club around the world. The goal of the front office is to invest in the right players to assist the team to win the championship every year. In the report, the main objective is to identify which type of hitters is likely to be imported from the Major League Baseball (MLB) to Nippon Baseball League (NPB). To achieve the goal, clustering analysis and regression predictions are conducted in the project. There are 203 hitters in the dataset and the time range is from 2000 to 2020. In clustering analysis, the result indicated that one group of players tends to have higher run earned in the NPB. This suggested that the front office should investigate the characteristics of these players. In regression predictions, linear regression and lasso regression are deployed. The best model indicated that it is more likely to predict the strikeout percentage in NPB by using MLB data. The insights from the regression prediction have shown that although the players are likely to get more strikeouts, these players are powerful hitters potentially due to the impact of exploratory variables. Therefore, the front office should consider this type of player if the team is not good at scoring runs.

Introduction

The baseball game has been popularised across many countries such as the United States, Japan, Korea, Taiwan, or central American countries. For any baseball organisation, their mutual goal is to win the championship in the post-season without any doubt. Therefore, in order to win the games, these teams try to buy the best players around the world.

A few decades before, In general, the rich teams won the games often. However, the concept of Moneyball comes in and changes the way to win [11]. One of the most popular notions is called sabermetrics or baseball statistics, which was invented by Bill James. Sabermetrics is an advanced data analytics tool to measure player performances [6]. The baseball organisations apply these statistical measures to evaluate the players. For instance, when we try to decide whether to sign this hitter, we are likely to look into the on-based percentage even though this player is underestimated [13].

As the Moneyball concept becomes popular, it has been embraced across the domains such as businesses and government departments. Also, it has also influenced other sports indeed. There has been an increasing number of front offices deploying the concept and data analytics tool to assist and to improve both the standing in the league and business perspectives [7].

Furthermore, since 1970, people have started to do baseball prediction due to the advent of free agency and the invention of fantasy baseball [12]. Especially, player performance is under the spotlight. Professional baseball clubs often analyse player performance to improve their team roster [2]. Due to the long history of baseball, abundant data collection has been analysing by using statistical models and mathematical analysis. In nowadays, there are many advanced models to apply to improve player performance or team performance due to the emerging technology. For instance, Houston Astros acquired a league-average pitcher, Gerrit Cole, from Pittsburgh Pirates. The team analysed his pitching data and noticed that he had to increase the spin rate of his fastball and the location of his pitches. After the training, Gerrit Cole has become one of the best starting pitchers in the league in 2019 [14]. Again, the huge amount of baseball data, combined with advanced statistical analysis models, can create opportunities for baseball teams to find market inefficiencies and field the most competitive, cost-efficient teams.

In the study, the main objective is to find which type of foreign hitter is likely to survive in the NPB. To be more specific, by using statistical models and clustering analysis, we try to identify if there are any predictive or correlations between MLB performance and NPB performance. In the meant time, we can also identify which type of hitter can import to NPB.

In the rest of the paper, some previous related works will be presented in the literature review. The analysis methods will be included in the approach section, including data collection, clustering analysis, and statistical prediction methods. The important and interesting results will be presented in the findings section. In the reflection, the conclusions and future recommendations will be presented. Also, personal reflection will be included as well.

The following table is the specialist terms for the report:

Term	Description
G	Game played
PA	Plate appearance
AB	At bats
R	Run Scored
H	Hits included 1B, 2B, 3B, and HR
2B	2 bases hit
3B	3 bases hit
HR	Home run
RBI	Run batted in
SB	Stolen bases
CS	Caught stealing
BB	Walks
SO	Strikeouts
BA	Batting average
OBP	On base percentage

SLG	Slugging percentage
OPS	On base + slugging percentage
TB	Total bases
GDP	Double play grounded info
HBP	Hit by pitch
SH	Sacrifice hits/bunts
SF	Sacrifice flies
IBB	Intentional base on balls
Hit per AB	Number of hits divided by total at bats
Two Base Hit per AB	Number of 2 bases hit divided by total at bats
Three Base Hit per AB	Number of 3 base hits divided by total at bats
HR per AB	Number of home runs divided by total at bats
Stolen Base percentage	Successful stolen bases divided by total attempts
Base on Ball percentage	Base on balls divided by total at bats
Strikeouts percentage	Strikeouts divided by total at bats
OPS plus	A normalized measure of the frequency with which a player reaches base plus the average number of bases the player records per plate-appearance.
WAR (Win Above Replacement)	A non-standardized sabermetric baseball statistic developed to sum up a player's total contributions to his team.
WOBA (Weighted On-Base Average)	Measure a player's overall offensive contributions per plate appearance.

Table 1. Specialist Terminology

Literature Review

Overview of Statistical Measurements

Baseball prediction has been prevailing since the 1900s [11]. The reason why organisations and fans are fascinated is that baseball has extremely abundant historical data. As a result, there are tons of statistical models out there to make accurate performance predictions. In the baseball game, players focus on how to earn as many as scores they can to help the team to win. This is the nature of the baseball game without any doubt. Because of the nature, people focus on some offensive measurements such as hits or home runs. Moreover, BA (batting average) is also a good measurement for a position player [5]. However, there are emerging debates about which statistics are the most representative measurements for the offensive player. Bennett and Flueck indicate that BA should be weighted because it counts all hits equally. For instance, if a player hits a home run, it only counts for a hit. Therefore, SLG (slugging percentage) appears to measure how well an offensive player produces at least 2 base hits. In modern baseball analytics, people not just focus on how many hits or home runs a player-generated but also put a spotlight on how well they generate runs efficiently. Therefore, advanced statistics have been created. For instance, run created which was invented by James, and linear weight base metrics like wOBA (weight on-base average) were invented by Tango, Lichtman, and Dolphin [17]. To sum up, there are too many useful statistics measurements in modern and will keep inventing and improving in the future. To sum up, in the project, we will incorporate some important baseball statistics into the models.

Statistical Modelling

With regard to the regression method, for example, there is a study from Stanford University. The project team used linear regression and support vector regression with different variants such as unregularised, lasso, and ridge [4]. The project has chosen some important modern statistics such as Win Above Replacement (WAR), OPS, and OPS plus, to be the response variables. The two regression models were player-specific and trained for 14 years of data. The predictive performance with response variables as OPS for linear regression and support vector regression was poor due to the gaps in data. On the other hand, the predictive performance with response variable as OPS plus for support vector regression was relatively better than others. However, in the real world, we don't have the best model. There are always some limitations. For example, the study from Stanford University indicated that they were unable to predict some issues such as injury and player personal life. The issue has also been revealed in Gow's study as well [1]. The suggestion from the study is to scrape some tweets from Twitter APIs to analyse. Nevertheless, it is time-consuming to accomplish it in a few weeks in this project.

Predicting the batting average also plays an important role. In Bailey's study [3], the main objective is to use the play-by-play data provided by the Statcast system in an attempt to predict batting averages. In the study, logistic regression was used to estimate the probability of a hit based on the Statcast data in 2015. Then, the prediction combined with 2016 PECOTA (Player Empirical Comparison and Optimization Test Algorithm) and actual batting average to get the prediction for 2017 by using linear regression. Interestingly, some variables such as exit velocity, launch angle, the distance the ball was hit, handedness of the hitter, and the footspeed of the player. These variables are not linear related but have some

interaction with the probability of hits. The result indicates that the Statcast dataset can improve the prediction with the PECOTA. However, the study mentions that age and injuries are likely to improve the prediction.

Machine Learning Algorithms

Besides statistical analysis, people also deploy some machine learning models to forecast player performance. A study states that most machine learning applications separate into three categories, including regression, binary classification, and multiclass classification [10]. With regard to a binary classification problem, for instance, Ganeshapillai and Guttag have researched about predict the next pitch type is a fastball or not [8]. They apply linear SVM (support vector machine) to classify pitches. Some scholars use other techniques such as K nearest neighbours or linear discriminant analysis (LDA). In the review from Koseler and Stephan, they conclude that SVM is the relatively easy method to conduct although linear discriminant analysis has slightly higher accuracy.

With regard to Multiclass classification, Sidle has a solid work [16]. Sidle uses several techniques such as LDA, SVM, and bagged random forest of classification. The results indicate that the forest of classification trees has the best prediction, but LDA is more efficient than tree-based classification in terms of running time. Therefore, it still depends on the data and goal of the prediction.

Predicting hitter performance based on some baseball statistics is our main goal. In Lyle's study [12], SVM, artificial neural networks, and other tree-based models predict different offensive statistics. Lyle compared the models with two baseball prediction systems, PECOTA and Szymborski Projection System (ZiPS). The result indicated that only three bases hit outperformed against the two systems.

Another interesting work from Panda, using penalised regression models to reduce the number of features to do the player performance prediction [15]. At first, 31 variables are selected. After the implementation, the final model only contained 9 important variables.

Bayesian methods have also played a major role in predicting the batting average. Jiang and Zhang applied the Empirical Bayes method to predict the batting average in 2006 by using MLB data from 2005 [9]. The result indicated that Empirical Bayes methods are likely substantially to improve upon the least-squares predictor. Moreover, Empirical Bayes methods are likely to capture a portion of the effects of missing covariables in the linear model.

Summary of Available Techniques and Materials

To sum up, there are many methods to do player performance prediction and do not have the best way to tackle. In statistical methods, logistic regression and linear regression are often used to do the prediction. On the machine learning approach, there are various methods to apply depending on the problem. For instance, in a classification problem, support vector machine and tree-based method often apply. On the other hand, penalised regression models and neural networks are often used to predict continuous variables. In the project, we will try not only statistical methods but also machine learning approaches to make the valid prediction. To be more specific, multiple linear regression and lasso regression will be performed in the project.

Data Analysis

Data Collection

In this project, we are going to use the data from baseball reference.com. Appendix A explains the code, data availability, and the example of the data. Furthermore, we are only interested in hitter data because pitchers in Japan are usually hard to attack. As a result, most ball clubs are eager to find powerful hitters to win the game. For the players' data, we will only consider former MLB players and exclude the players who were from Japan and also been to MLB then returned back. For instance, former Kansas Royals outfielder Nori Aoki. In terms of season, in the project, we will exclude the post-season data and only include regular season data. The reason is that everything is hard to predict in the post-season because of psychological issues or others. Regarding the time range, the project includes the data from 2000 to 2020 with both Central league and Pacific league.

Figure 1 presents the whole data collection process. All the data is collected from the website. The method to select the correct hitter is that manually search each team roster in each year. After finalising the data for MLB and NPB, merging and dropping missing values are conducted. The missing values are not significant so we choose to drop them. As a result, the final dataset contains 203 players' data. Table 2 presents the selected attributes from baseball reference.com. The reason why I selected these variables is that they are all important measurements for a hitter. Appendix 2 shows the codes for data cleaning.

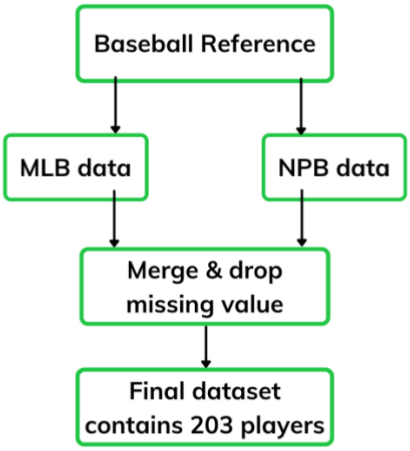

Baseball Reference
MLB data
NPB data
Merge & drop missing value
Final dataset contains 203 players

Figure 1. Data collection process

Term	Description
G	Game played
PA	Plate appearance
AB	At bats
R	Run Scored
H	Hits included 1B, 2B, 3B, and HR
2B	2 bases hit
3B	3 bases hit
HR	Home run
RBI	Run batted in
SB	Stolen bases
CS	Caught stealing
BB	Walks
SO	Strikeouts
BA	Batting average
OBP	On base percentage
SLG	Slugging percentage
OPS	On base + slugging percentage
TB	Total bases
GDP	Double play grounded info
HBP	Hit by pitch
SH	Sacrifice hits/bunts
SF	Sacrifice flies
IBB	Intentional base on balls
Hit per AB	Number of hits divided by total at bats
Two Base Hit per AB	Number of 2 bases hit divided by total at bats
Three Base Hit per AB	Number of 3 base hits divided by total at bats
HR per AB	Number of home runs divided by total at bats
Stolen Base percentage	Successful stolen bases divided by total attempts

Base on Ball percentage	Base on balls divided by total at bats
Strikeouts percentage	Strikeouts divided by total at bats

Table 2. Attributes for the dataset

Approach

Programming Languages

There are two programming languages involved in the project, python and R programming. For the exploratory data analysis and clustering analysis, python programming is used. It is easier to implement the algorithms by using the scikit learn package. For the linear and lasso regression, R programming is used. The reason is that there are many statistical packages such as broom and glmnet packages. Some of the inbuilt functions in R assist us to do and interpret the analysis.

Exploratory Data Analysis

To identify which type of players are likely to survive in NPB, a simple exploratory data analysis is conducted at first. The project has compared all variables between MLB and NPB data by using a scatter plot. The majority of the variables show no trends. However, base on ball percentage and strikeouts percentage show some degrees of positive relationships. Figures 2 and 3 illustrate that there are some positive relationships between MLB and NPB data in base on ball percentage and strikeouts percentage. Further investigation is required for these two variables.

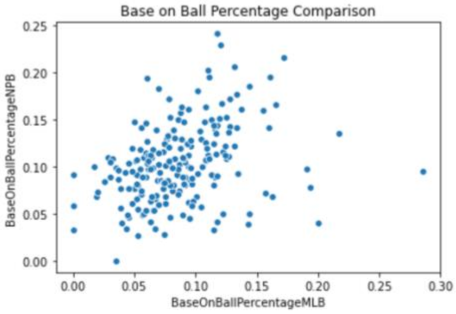


Figure 2. Base on ball percentage comparison

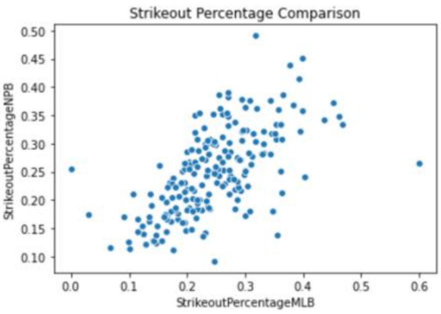


Figure 3. Strikeout percentage comparison

Clustering Analysis

To observe whether there are any trends in the NPB data, the project conducts k means clustering. Appendix 3 shows the process of conducting the clustering. Our aim is to identify which cluster has what kind of characteristics. The reason why we use k means clustering is that the variables are all numerical variables. Before we implement the algorithm, the number of clusters has to specify first. To determine the number of clusters, there are two methods to use, the elbow method and silhouette score. Figures 4 and 5 demonstrate that the ideal number of clusters is 4 clusters.

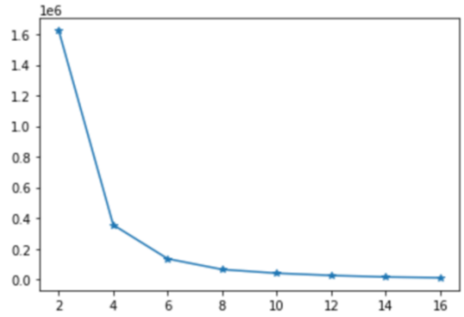
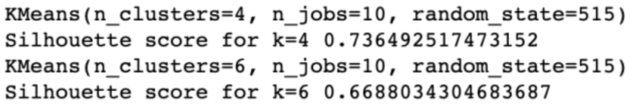


Figure 4. Elbow method



```
KMeans(n_clusters=4, n_jobs=10, random_state=515)
Silhouette score for k=4 0.736492517473152
KMeans(n_clusters=6, n_jobs=10, random_state=515)
Silhouette score for k=6 0.6688034304683687
```

Figure 5. Silhouette score

The results of the k means clustering will be presented in the findings part.

Multiple Linear Regression

In the project, we are going to predict BA, OBP, SLG, RBI, Stolen base percentage, Base on ball percentage, Strikeout percentage, 2B per AB, 3B per AB, HR per AB, Game played, PA, AB, R, and CS in NPB as response variables by using MLB data as covariates.

By using multiple linear regression, we can predict one response variable with multiple covariates. For instance, to predict the batting average in NPB, we use other variables from MLB as covariates to predict the value. The formula for the multiple linear regression is presented below:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta 2 x_2 + \cdots + \varepsilon_i$$

Y_i is the continuous response variable while X1, X2, X3, …, etc as the predictor variables. β_0 is the intercept term and β_1 is the coefficient. Last, ε_i is the residual value. Appendix 4 illustrates one of the examples of conducting the linear regression.

In the meantime, the project will conduct polynomials models to check whether there are some other unexplained curvatures. In this project, we use the quadratic model to check the variability. If the R-squared value is better than the linear model, the quadratic model is preferred. The quadratic model equation is presented below:

$$Y_i = \beta_0 + \beta_1 x_i + \beta 2 x_i^2 + \cdots + \varepsilon_i$$

This is still a linear regression because we have a linear function of the β parameter. Appendix 5 illustrates one of the examples of conducting the quadratic model.

Moreover, the three assumptions of the linear regression have been checked in the project. There are linearity, homogeneity of errors, and normality of errors. Only a few models meet the assumptions. The example is presented in appendix 7.

Lasso regression

The reason why we choose the least absolute shrinkage and selection operator (Lasso) regression is that Lasso regression not only can assist to prevent overfitting but also can help to do the feature selection [panda]. The estimate of Lasso is presented below:

$$\hat{\beta}^{lasso} = argmin|y - X\beta|^2 + \lambda|\beta|_1$$

Lasso regression is a penalised least squares regression that applies ℓ_1 penalty on the coefficients. λ is a tuning parameter. The constraint term, $|\beta|_1$, is called ℓ_1 the penalty. The penalization is used on the regression coefficients, which tends to continuously shrink some coefficients toward zero as λ increases, and sets other coefficients exactly equal to zero if λ is sufficiently large. The penalty generates both continuous shrinkage and automatic variable selection. Eventually, the rest of the feature importance will explain the model. Therefore, after selecting the best models, it is easier to interpret the results and give solid insights. One of the examples is presented in appendix 6.

Findings

After we implement the k means clustering, Figure 6 shows the result. There are several findings in the clustering analysis.

1. The majority of the cluster overlap with each other, meaning that there is no clear trend in the data.
2. Only the first row and column have a slightly clear boundary between the 4 clusters. The results indicated that the green cluster tends to have a high run earned value.
3. There are some linear relationships between some variables. There is batting average, on-base percentage, slugging percentage, and on-base plus slugging percentage. The result is understandable because these baseball statistics are correlated with each other.

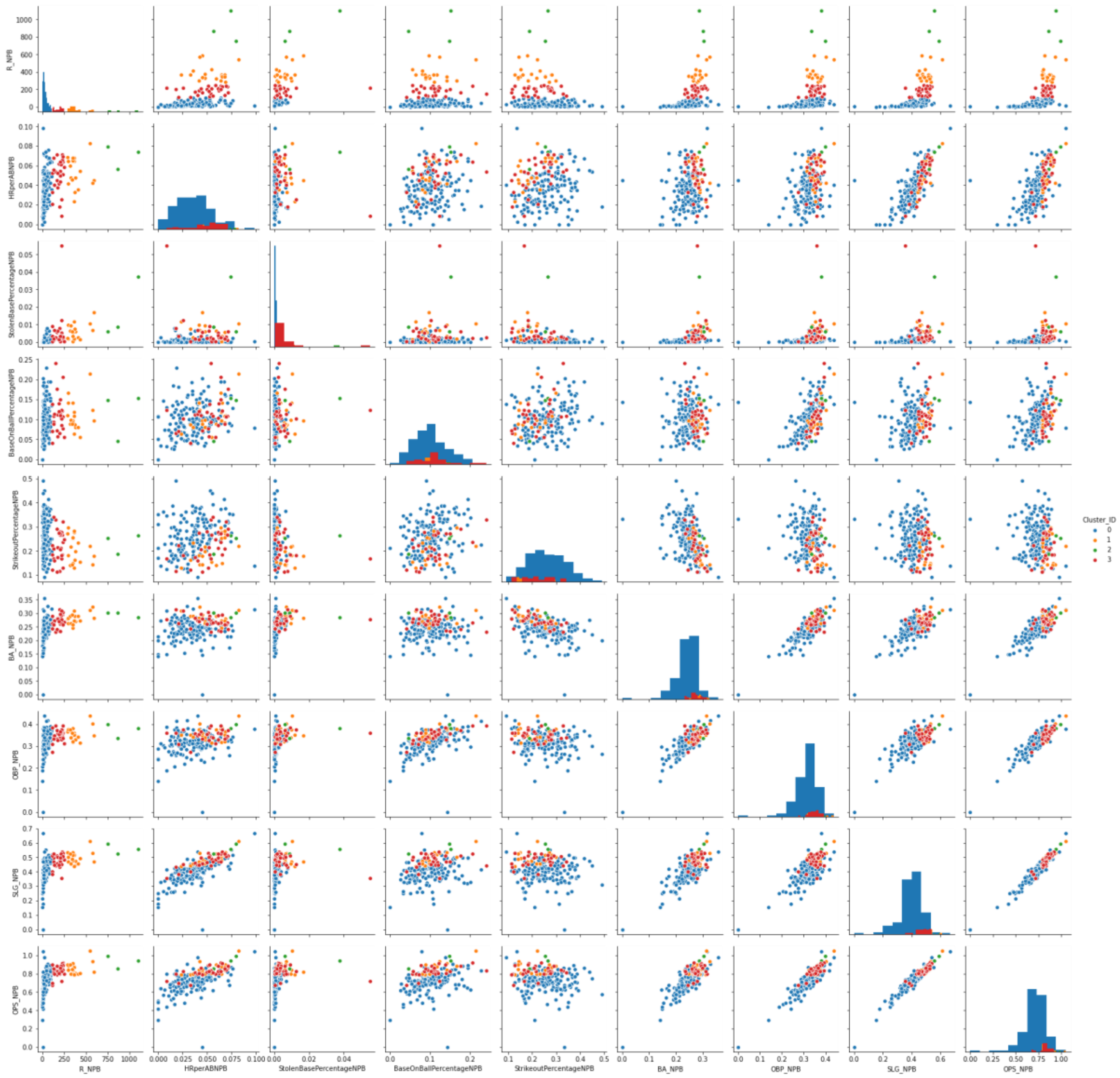


Figure 6: K means clustering result

For the linear and lasso regression, we use the R-squared value to determine the goodness of fit of the models. The coefficient of determination, or R-squared, is how much of the original variability still remains. If the value is close to one, the model explains nearly all the variations. On the other hand, if the value is close to zero, the model no better than the null model. Table 3 presents the R-squared for all models. There are several findings from table 3.

1. Quadratic models do not explain the variability well
2. Strikeout percentage has the highest value in the lasso model
3. Overall, the models cannot explain over 50 percent of the variability.

The top three models are strikeout percentage, base on ball percentage, and HR per AB. The table 4, 5, and 6 present the feature importance for each lasso model. Although the R-squared value is not convincing, the prediction of the batting performance in either MLB or NPB is extremely difficult [3]. Therefore, we can say that the top three R-squared values are explained relatively well in the context. In particular, the Strikeout percentage explained nearly 50% of the variability, so the model appears to have some degree of robustness.

R-squared	Linear Model	Quadratic Model	Lasso Model
BA	3.56%	0.29%	0%
OBP	4.31%	2.51%	0%
SLG	11.23%	2.26%	0%
OPS	10.27%	2.42%	0%
RBI	3.75%	0.49%	0%
SB%	4.95%	0.31%	0.68%
BB%	23.56%	3.76%	29.26%
K%	44.48%	4.65%	45.25%
2B per AB	3.88%	0.28%	0%
3B per AB	14.02%	0.11%	11%
HR per AB	22.65%	3.07%	25.71%
Game played	5.02%	0.57%	0%
PA	4.61%	0.36%	0%
AB	4.96%	0.36%	0%
R	3.57%	0.35%	0%
CS	6.41%	0.08%	1.74%

Table 3. R-squared table

Coefficient	Impact on the response variable
Strikeouts	Positive
OPS	Positive
3B per AB	Positive
HR per AB	Positive
BB%	Positive
K%	Positive
GDP	Negative
SH	Negative

Table 4. Feature importance with strikeout percentage as the response variable

Coefficient	Impact on the response variable
CS	Positive
BB	Positive
HBP	Positive
HR per AB	Positive
BB%	Positive
K%	Positive
H	Negative
2B	Negative
OBP	Negative
SH	Negative
IBB	Negative
2B per AB	Negative
3B per AB	Negative

Table 5. Feature importance with base on ball percentage as the response variable

Coefficient	Impact to the response variable
CS	Positive
HR	Positive
SF	Positive
HR per AB	Positive
BB%	Positive
K%	Positive
G	Negative
2B	Negative
3B	Negative
OBP	Negative
HBP	Negative
SH	Negative
IBB	Negative
2B per AB	Negative
3B per AB	Negative

Table 6. Feature importance with HR per AB as the response variable

Reflection

In the section, the conclusion, recommendation, limitation, and self-reflection will be explained clearly here.

Conclusion

After the analysis, there are two insights for the NPB front office.

1. Based on the clustering analysis, there is a group of players that tends to have higher run earned. In general, these players have high HR per AB, Stolen base percentage, BA, OBP, SLG, and OPS.
2. In the regression prediction, the lasso models are likely to predict the strikeout percentage, base on ball percentage, and HR per AB in the NPB.
    - Based on the coefficients to predict the strikeout percentage, the higher the strikeouts, OPS, 3B per AB, HR per AB, base on ball percentage, and strikeout percentage in MLB, the more likely to have a relatively higher strikeout percentage in NPB.
    - Based on the coefficients to predict the base on ball percentage, the higher the number of base on balls, CS, HBP, HR per AB, base on ball percentage, and strikeout percentage in MLB, the more likely to have a relatively higher base on ball percentage in NPB.
    - Based on the coefficients to predict the HR per AB, the higher the HR, CS, SF, HR per AB, base on ball percentage, and strikeout percentage, the more likely to have relatively higher HR per AB.

Recommendation

There are two suggestions to the NPB ball club front offices.

1. Based on the clustering result, the front office should investigate why these players tend to have high-run earned. For instance, The office can investigate the age, race, and past experiences within these players.

2. Based on the coefficients for the lasso models with strikeout percentage as response variables, the type of player tends to be a power hitter. There are several reasons.
    - High OPS means that the players have high OBP and SLG.
    - High 3B per AB means that the players have good speed and hitting techniques.
    - High HR per AB means that the players are powerful.
    - High base on ball percentage means that the players have great plate discipline.

    From the above reasons, the players are likely to be powerful hitters who can generate more runs for the team although they are more likely to get strikeouts. As a result, if the teams are not good at scoring runs, the front office can consider investing this type of player.

Limitation

There are three limitations in the analysis.

1. League differences
    i) Different playing style
        (1) In Japan, the playing style is more focused on the efficiency of scoring runs. When a hitter gets on base, the next hitter is more likely to sacrifice himself to let the runner move to second base or third base. Therefore, when foreign hitters come to NPB, they have to adapt to the environment in order to succeed.
        (2) In MLB, the hitters are more likely to hit by themselves rather than sacrifice.
    ii) Different ballpark factor
        (1) Pitcher favour or hitter favour ballpark
        For example, the Tokyo dome is likely to hit a homerun compared to other stadiums.
        (2) Weather factors
        The wind direction is different from place to place. For instance, the ZOZO Marine stadium is windy in general because it locates around the bay area.
2. Lack of scout perspectives
    Baseball is not all about statistics but is human-oriented. There are so many factors to evaluate a player. The scouts have to observe the players from many aspects including their bodies, families, educations, previous injuries, etc. There are so many hidden places to evaluate whether to sign a player. Therefore, we have to combine the scout perspectives to decide whether to invest the foreign players.
3. Imbalanced data
    The data we collected have an imbalanced issue. That is, every player has a different game played, plate appearance, and at bat.

There are still many aspects to discuss because sports is human-oriented. There are some suggestions for future work in order to make an accurate analysis.

1. Adding age factor as a variable
2. Adding the team the players belong to as a variable
3. Doing the injury prediction
4. Investigating the player off-the-fields background

Self-reflection

Before choosing the project, I have been watching baseball for 12 years. To be frank, I have relatively strong domain knowledge compared with others. With my passion, I really enjoy the whole semester. However, there are some obstacles indeed. First, during the data collection phase, there are too much data to collect. Therefore, I have tried web scraping with Selenium on the baseball reference website. Due to the lack of experience, it took me many days to debug. Unfortunately, the structure of the website is too messy to scrape and organise. Therefore, I collected the data manually in the end. At least, I accomplished my first web scrapping in the semester. Second, lack of peer discussion is also a downside of the project. I think the reason is that not many people familiar with baseball in the data hub. Another reason might be that the topic I chose is different than others. Fortunately, I attend the weekly meeting with my supervisor Dimitri. He always gives me some solid suggestions to complete my analysis via meeting or on slack. Last, I have never done the lasso regression. Because linear regression is not suitable to predict the response variables, I have to investigate other techniques. Before implementing the lasso regression, I have read some articles to get familiar with it. Fortunately, when I proposed this method to my supervisor, I got positive feedback. Overall, I am really happy working with my own collected dataset and with professor Dimitri.

Reference

[1] A. Gow, "Using Machine Learning to predict MLB success Based on MILB performance," p. 2.

[2] Michael Hamilton, Phuong Hoang, Lori Layne, Joseph Murray, David Padgett, Corey Stafford, Hien Tran, "Applying Machine Learning Techniques to Baseball Pitch Prediction:," in *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods*, ESEO, Angers, Loire Valley, France, 2014, pp. 520–527. doi: 10.5220/0004763905200527.

[3] S. R. Bailey, J. Loeppky, and T. B. Swartz, "The prediction of batting averages in major league baseball," *Stats*, vol. 3, no. 2, pp. 84–93, 2020.

[4] S. Benavidez, S. Brito, D. McCreight, and P. McEvoy, "Prediction of Future Offensive Performance of MLB Position Players," 2019.

[5] J. M. Bennett and J. A. Flueck, "An evaluation of major league baseball offensive performance models," *The American Statistician*, vol. 37, no. 1, pp. 76–82, 1983.

[6] R. Elitzur, "Data analytics effects in major league baseball," *Omega*, vol. 90, p. 102001, Jan. 2020, doi: 10.1016/j.omega.2018.11.010.

[7] M. J. Fry and J. W. Ohlmann, "Introduction to the special issue on analytics in sports, part I: General sports applications," 2012.

[8] G. Ganeshapillai and J. Guttag, "Predicting the next pitch," presented at the Sloan Sports Analytics Conference, 2012.

[9] W. Jiang and C.-H. Zhang, “Empirical Bayes in-season prediction of baseball batting averages,” in *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, Institute of Mathematical Statistics, 2010, pp. 263–273.

[10] K. Koseler and M. Stephan, “Machine learning applications in baseball: A systematic literature review,” *Applied Artificial Intelligence*, vol. 31, no. 9–10, pp. 745–763, 2017.

[11] M. Lewis, *Moneyball: The art of winning an unfair game*. WW Norton & Company, 2004.

[12] A. Lyle, “Baseball prediction using ensemble learning,” 2007.

[13] A. Milgram, “Moneyballing criminal justice,” *The Atlantic, June*, vol. 20, 2012.

[14] D. T. Pagliarini, “Winning Investment: The Strategic Implications of the Growth of Sabermetrics in Japanese Baseball,” 2021.

[15] M. L. Panda, “Penalized regression models for major league baseball metrics,” 2014.

[16] G. Sidle and H. Tran, “Using multi-class classification methods to predict baseball pitch types,” *Journal of Sports Analytics*, vol. 4, no. 1, pp. 85–93, 2018.

[17] T. M. Tango, M. G. Lichtman, and A. E. Dolphin, *The book: Playing the percentages in baseball*. Potomac Books, Inc., 2007.

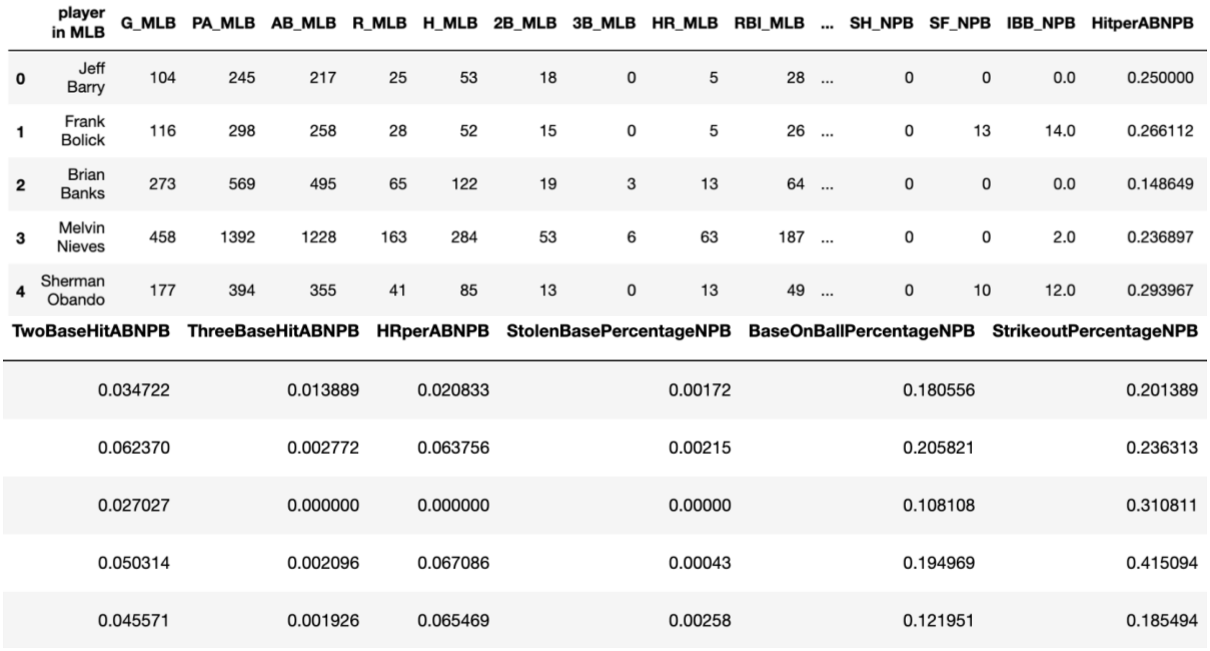
Appendix

Appendix 1: Code and Data Availability

The data set is from baseball-reference.com. This website is publicly available therefore there is no data ethics problem. The reader can visit this website for reproducibility.

For the code, r markdown and python screenshots presented in the following appendix are only for the professors to mark the project's performance and submit in Turnitin system for the Queensland University of Technology.

Example of data



	player in MLB	G_MLB	PA_MLB	AB_MLB	R_MLB	H_MLB	2B_MLB	3B_MLB	HR_MLB	RBI_MLB	...	SH_NPB	SF_NPB	IBB_NPB	HitperABNPB	TwoBaseHitABNPB	ThreeBaseHitABNPB	HRperABNPB	StolenBasePercentageNPB	BaseOnBallPercentageNPB	StrikeoutPercentageNPB
0	Jeff Barry	104	245	217	25	53	18	0	5	28	...	0	0	0.0	0.250000	0.034722	0.013889	0.020833	0.00172	0.180556	0.201389
1	Frank Bolick	116	298	258	28	52	15	0	5	26	...	0	13	14.0	0.266112	0.062370	0.002772	0.063756	0.00215	0.205821	0.236313
2	Brian Banks	273	569	495	65	122	19	3	13	64	...	0	0	0.0	0.148649	0.027027	0.000000	0.000000	0.00000	0.108108	0.310811
3	Melvin Nieves	458	1392	1228	163	284	53	6	63	187	...	0	0	2.0	0.236897	0.050314	0.002096	0.067086	0.00043	0.194969	0.415094
4	Sherman Obando	177	394	355	41	85	13	0	13	49	...	0	10	12.0	0.293967	0.045571	0.001926	0.065469	0.00258	0.121951	0.185494

Appendix 2: Data Cleaning Process

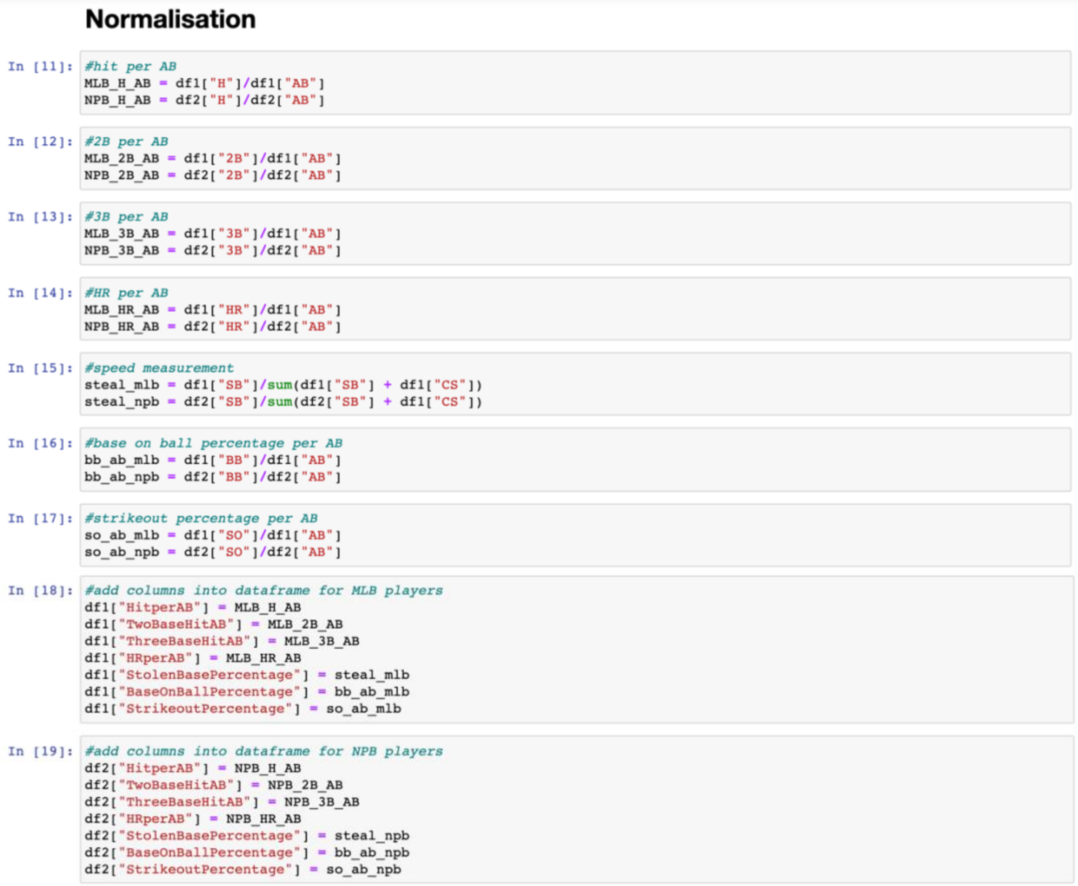
Data cleaning

```
#library

library(tidyverse)

library(broom)

library(DHARMa)

#import data

df_final <- read_csv("baseball_data_cleaned.csv")

df_MLB <- read_csv("baseball_data_MLB.csv")

df_NPB <- read_csv("baseball_data_NPB.csv")

#drop the useless column

df_final %>%

 select (-c(`player in NPB`))
```

```
#check the final dataframe

colnames(df_final)

#drop missing value

df_final <- na.omit(df_final)
```

Normalisation

This part of codes normalises the baseball statistics before doing clustering analysis.

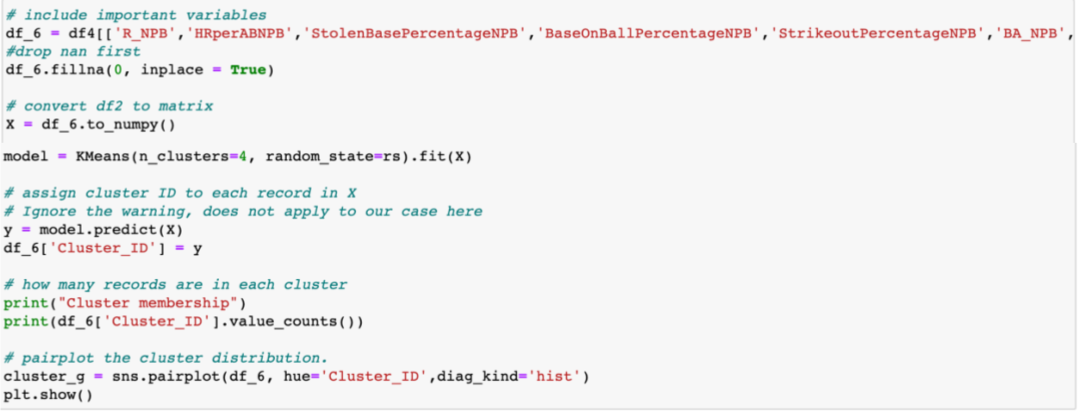


Normalisation

```
In [11]: #hit per AB
         MLB_H_AB = df1["H"]/df1["AB"]
         NPB_H_AB = df2["H"]/df2["AB"]

In [12]: #2B per AB
         MLB_2B_AB = df1["2B"]/df1["AB"]
         NPB_2B_AB = df2["2B"]/df2["AB"]

In [13]: #3B per AB
         MLB_3B_AB = df1["3B"]/df1["AB"]
         NPB_3B_AB = df2["3B"]/df2["AB"]

In [14]: #HR per AB
         MLB_HR_AB = df1["HR"]/df1["AB"]
         NPB_HR_AB = df2["HR"]/df2["AB"]

In [15]: #speed measurement
         steal_mlb = df1["SB"]/sum(df1["SB"] + df1["CS"])
         steal_npb = df2["SB"]/sum(df2["SB"] + df1["CS"])

In [16]: #base on ball percentage per AB
         bb_ab_mlb = df1["BB"]/df1["AB"]
         bb_ab_npb = df2["BB"]/df2["AB"]

In [17]: #strikeout percentage per AB
         so_ab_mlb = df1["SO"]/df1["AB"]
         so_ab_npb = df2["SO"]/df2["AB"]

In [18]: #add columns into dataframe for MLB players
         df1["HitperAB"] = MLB_H_AB
         df1["TwoBaseHitAB"] = MLB_2B_AB
         df1["ThreeBaseHitAB"] = MLB_3B_AB
         df1["HRperAB"] = MLB_HR_AB
         df1["StolenBasePercentage"] = steal_mlb
         df1["BaseOnBallPercentage"] = bb_ab_mlb
         df1["StrikeoutPercentage"] = so_ab_mlb

In [19]: #add columns into dataframe for NPB players
         df2["HitperAB"] = NPB_H_AB
         df2["TwoBaseHitAB"] = NPB_2B_AB
         df2["ThreeBaseHitAB"] = NPB_3B_AB
         df2["HRperAB"] = NPB_HR_AB
         df2["StolenBasePercentage"] = steal_npb
         df2["BaseOnBallPercentage"] = bb_ab_npb
         df2["StrikeoutPercentage"] = so_ab_npb
```

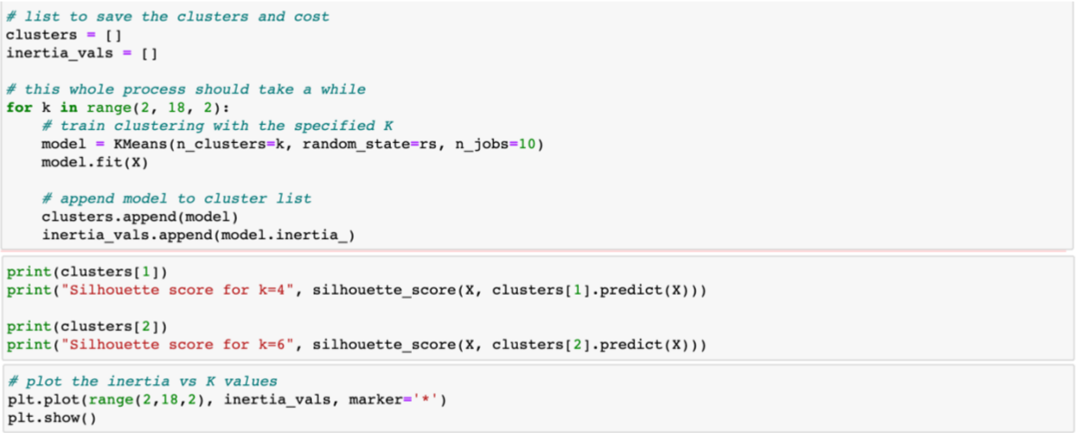
Appendix 3: Clustering analysis process

Here, the codes demonstrate the process of k means clustering.



```
# include important variables
df_6 = df4[['R_NPB','HRperABNPB','StolenBasePercentageNPB','BaseOnBallPercentageNPB','StrikeoutPercentageNPB','BA_NPB',
#drop nan first
df_6.fillna(0, inplace = True)

# convert df2 to matrix
X = df_6.to_numpy()

model = KMeans(n_clusters=4, random_state=rs).fit(X)

# assign cluster ID to each record in X
# Ignore the warning, does not apply to our case here
y = model.predict(X)
df_6['Cluster_ID'] = y

# how many records are in each cluster
print("Cluster membership")
print(df_6['Cluster_ID'].value_counts())

# pairplot the cluster distribution.
cluster_g = sns.pairplot(df_6, hue='Cluster_ID',diag_kind='hist')
plt.show()
```

Specify the number of clusters



```
# list to save the clusters and cost
clusters = []
inertia_vals = []

# this whole process should take a while
for k in range(2, 18, 2):
    # train clustering with the specified K
    model = KMeans(n_clusters=k, random_state=rs, n_jobs=10)
    model.fit(X)

    # append model to cluster list
    clusters.append(model)
    inertia_vals.append(model.inertia_)
```

```
print(clusters[1])
print("Silhouette score for k=4", silhouette_score(X, clusters[1].predict(X)))

print(clusters[2])
print("Silhouette score for k=6", silhouette_score(X, clusters[2].predict(X)))
```

```
# plot the inertia vs K values
plt.plot(range(2,18,2), inertia_vals, marker='*')
plt.show()
```

Appendix 4: Example of doing the linear regression

```
#fit the linear model

BB_npb_lm <-lm(data=df_final, BaseOnBallPercentageNPB ~ G_MLB + PA_MLB + AB_MLB + 
R_MLB +     HR_MLB + RBI_MLB +

            BA_MLB +

            OBP_MLB +

            SLG_MLB +

            OPS_MLB +

            CS_NPB +

            TB_MLB +

            SH_MLB +

            SF_MLB +

            StolenBasePercentageMLB +

            BaseOnBallPercentageMLB +

            StrikeoutPercentageMLB)


summary(BB_npb_lm)

#check the homogeneity of error

fortify_linear_BB <-fortify(BB_npb_lm)


ggplot(data=fortify_linear_BB,aes(x=.fitted, y=.resid))+

 geom_point()+
 theme_bw()+
 xlab("Fitted values")+
 ylab("Residuals")+
 geom_smooth()
#check the normality of error
ggplot(data=fortify_linear_BB, aes(sample=.stdresid))+
 stat_qq(geom="point")+
 geom_abline()+
 xlab("Theoretical (Z ~ N(0,1))")+
 ylab("Sample")+
 coord_equal()+
 theme_bw()
```

Appendix 5: Example of the linear regression with quadratic model

```
BB_npb_quadratic <-lm(data= df_final, BaseOnBallPercentageNPB ~ poly(G_MLB + PA_MLB
+ AB_MLB + R_MLB +     HR_MLB + RBI_MLB +
            BA_MLB +
            OBP_MLB +
            SLG_MLB +
            OPS_MLB +
            CS_NPB +
            TB_MLB +
            SH_MLB +
            SF_MLB +
            StolenBasePercentageMLB +
            BaseOnBallPercentageMLB +
            StrikeoutPercentageMLB, 2, raw=T))
summary(BB_npb_quadratic)
#check the homogeneity of error
fortify_quadratic_BB <-fortify(BB_npb_quadratic)
```

```
ggplot(data=fortify_quadratic_BB,aes(x=.fitted, y=.resid))+
  geom_point()+
  theme_bw()+
  xlab("Fitted values")+
  ylab("Residuals")+
  geom_smooth()
#check the normality of errors
ggplot(data=fortify_quadratic_BB, aes(sample=.stdresid))+
  stat_qq(geom="point")+
  geom_abline()+
  xlab("Theoretical (Z ~ N(0,1))")+
  ylab("Sample")+
  coord_equal()+
  theme_bw()
```

Appendix 6: Example of the lasso regression

```
#response variable
y <- df_final$BaseOnBallPercentageNPB


#define matrix of predictor variables
x <- data.matrix(df_final[, c("G_MLB", "PA_MLB", "AB_MLB", "R_MLB", "H_MLB" ,
  "2B_MLB", "3B_MLB", "HR_MLB","RBI_MLB", "SB_MLB", "CS_MLB", "BB_MLB",
  "SO_MLB", "BA_MLB", "OBP_MLB", "SLG_MLB", "OPS_MLB","TB_MLB", "GDP_MLB",
  "HBP_MLB", "SH_MLB", "SF_MLB" , "IBB_MLB", "HitperABMLB", "TwoBaseHitABMLB",
  "ThreeBaseHitABMLB","HRperABMLB", "StolenBasePercentageMLB",
  "BaseOnBallPercentageMLB","StrikeoutPercentageMLB")])
#perform k-fold cross-validation to find optimal lambda value
cv_model <- cv.glmnet(x, y, alpha = 1)
```

```
#find optimal lambda value that minimizes test MSE
best_lambda <- cv_model$lambda.min
best_lambda
## [1] 0.001123102
#produce plot of test MSE by lambda value
plot(cv_model)
#lamda value
best_lambda
## [1] 0.001123102
#find coefficients of best model
best_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)
coef(best_model)
#use fitted best model to make predictions
y_predicted <- predict(best_model, s = best_lambda, newx = x)

#find SST and SSE
sst <- sum((y - mean(y))^2)
sse <- sum((y_predicted - y)^2)

#find R-Squared
rsq <- 1 - sse/sst
rsq
## [1] 0.2616034
```

Appendix 7: Example of linear regression assumptions

Homogeneity of errors

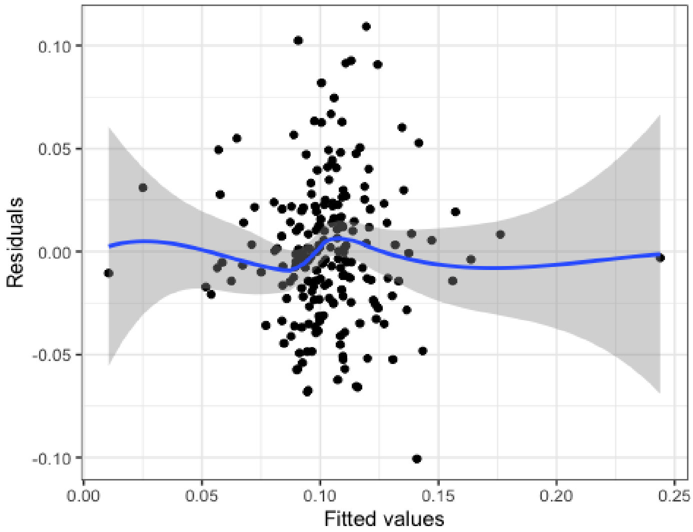


Figure 7: Homogeneity of errors with the base on ball percentage

Normality of errors

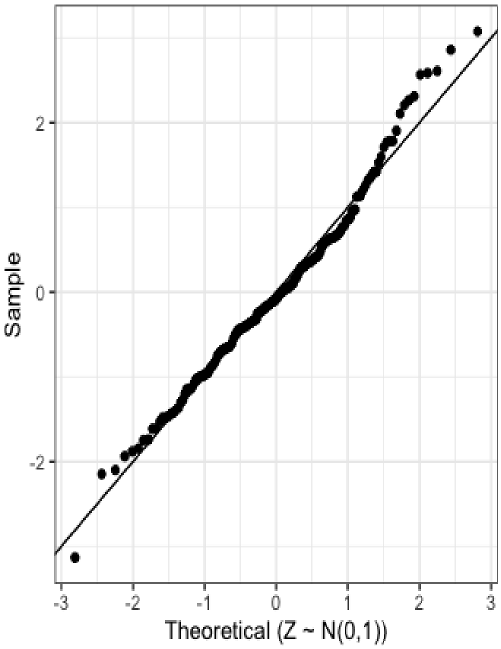


Figure 8: Normality of errors with the base on ball percentage

The above example indicates that the residuals are not homogeneous and normally distributed.