Analyzing Baseball Statistics Across Cultures

: A STUDY OF THE KBO

Boston University

Junho Eum

Table of Contents

[Abstract 2](#_Toc141098960)

[Introduction 3](#_Toc141098961)

[Methodology 3](#_Toc141098962)

[Preprocessing & parameter selection 3](#_Toc141098963)

[Analyzing the Relative Impact of Pitching and Batting on Win-Loss Percentage 6](#_Toc141098964)

[Principal component analysis with Gaussian Mixture Model 7](#_Toc141098965)

[Regression Analysis to Understand Influence of Batting and Pitching Variables 13](#_Toc141098966)

[Optimization and Interaction Analysis of Batting and Pitching Variables Using XGBoost Models 15](#_Toc141098967)

[*Figure 2-4 Summary of Interaction plot of KBO pitcher features* 16](#_Toc141098968)

[Multinomial Logistic regression modeling 17](#_Toc141098969)

[Lasso Regression for variable selection 20](#_Toc141098970)

[Ridge Regression 21](#_Toc141098971)

[Pythagorean Expectation Modeling 23](#_Toc141098972)

[Conclusion 23](#_Toc141098973)

# Abstract

The Korean Baseball League (KBO) was established in 1982, significantly later than other professional baseball leagues such as Nippon Professional Baseball (NPB) in Japan and Major League Baseball (MLB) in the United States which were founded in 1936 and 1876, respectively. Consequently, the KBO has been influenced by both the Japanese and American baseball cultures, resulting in the development of its unique playing style. This research paper aims to analyze and compare the playing style of the KBO with that of the MLB and NPB by studying their respective statistical data. The research question for this study is: "How does the KBO differ in its statistical characteristics compared to the MLB and NPB, and how do these differences contribute to its distinct playing style?"

# Introduction

A myriad of studies has explored various aspects of baseball statistics, such as player performance, game strategies, and team dynamics, in different leagues. Nevertheless, limited scholarly attention has been given to comparative analysis across different professional baseball leagues, specifically between the KBO, MLB, and NPB. This study seeks to fill this gap in the literature by conducting a cross-cultural analysis of baseball statistics through the lens of Pythagorean Expectation, a formula used to predict a team's win rate based on runs scored and runs allowed. Bill James, the pioneer of sabermetrics, initially developed the Pythagorean Expectation equation:

Pythagorean Expectation=(RS^2)/(RS^2+RA^2)

```

To conduct a comparative analysis, this study will analyze pitching and batting datasets obtained from Kaggle, employing methods of preprocessing and parameter selection to build a model based on the Pythagorean Expectation formula.

# Methodology

## Preprocessing & parameter selection

In this study, we began by importing two primary datasets, specifically the pitching and batting records. Upon conducting an initial data integrity check, we identified the presence of missing data points in several columns. We attribute these inconsistencies to the lack of a comprehensive stat recording system in the historical context. Given the study's objective to investigate recent advancements and differences in the Korean Baseball Organization (KBO) playstyle, we made a strategic decision to remove duplicates and rows containing null values, to uphold data quality and validity.

Subsequently, we implemented a correlation analysis to identify variables that significantly correlate with the dependent variable, i.e., the win-loss percentage. The outcome of this analysis was visually represented in a correlation heatmap (Figure 1-1). To further refine our exploration, we created a focused heatmap to explicitly illuminate the correlations between win-loss percentage and other associated variables (Figure 1-2). This process helped underscore variables that significantly contributed to the dependent variable.

However, it is important to acknowledge the limitations of such an approach. While it effectively highlighted variables like 'wins' and 'losses' that exhibited high collinearity with the dependent variable 'win-loss percentage,' this technique of feature selection may not be sufficient in isolation.

Selecting features based solely on their correlation with the dependent variable can lead to model errors due to several reasons. Firstly, it overlooks the possibility of multicollinearity, a situation where independent variables are highly correlated with each other. In such a case, the model's interpretability suffers as it becomes challenging to distinguish the individual effects of predictors on the response variable. Secondly, correlation does not imply causation. High correlation may simply result from lurking or confounding variables. Finally, this approach is more suitable for linear relationships and may miss out on important non-linear relationships that exist in the data.

To avoid any possible issues, we took a smart approach and split the KBO league's features into two categories: pitcher and batter features. This allowed us to dive deeper into the data and gain better insights.

A screenshot of a computer screen

Description automatically generated

*Figure 1-1: pairwise heatmap for dataset parameters*

A picture containing text, screenshot, line, diagram

Description automatically generated

*Figure 1-2: specified heatmap for parameter selection*

## Regression Analysis to Understand Influence of Batting and Pitching Variables

We next performed regression analyses to understand the influence of batting and pitching variables on the win-loss percentage.

The sum of the absolute values of the coefficients indicated the total influence of each group of variables, assuming all variables are on the same scale. The sum for the pitching-related variables was approximately 1.42, suggesting that for a one-unit change in the normalized pitching-related features, we expect an average change of approximately 1.42 in the win-loss percentage. The sum for the batting-related variables was approximately 2.98, implying that for a one-unit change in the normalized batting-related features, we expect an average change of approximately 2.98 in the win-loss percentage.

We also examined the decrease in the adjusted R-squared when removing a group of variables. This showed how much of the variance in the win-loss percentage that group of variables explains. The decrease in R-squared when removing the pitching variables was approximately 0.417, indicating that about 41.7% of the variability in the win-loss percentage could be explained by the pitching variables. Conversely, the decrease in R-squared when removing the batting variables was approximately 0.284, meaning about 28.4% of the variability in the win-loss percentage could be explained by the batting variables.

These analyses revealed that batting and pitching variables have different levels of influence on the win-loss percentage, depending on the criteria used to evaluate their importance. This necessitates further investigations into potential multicollinearity and interaction effects among variables.

A screenshot of a graph

Description automatically generated

## Principal component analysis with Gaussian Mixture Model

The methodology employed Principal Component Analysis (PCA) in conjunction with Gaussian Mixture Models (GMM) clustering on both batter and pitcher data. The objective of this analytical step was to identify and understand which components and metrics led to increased team wins in the KBO league.

To manage the high dimensionality of the dataset, PCA was utilized initially. PCA transformed the dataset into a new set of orthogonal features, or "components", each of which was a linear combination of the original features. Each component represented a certain amount of the total variance in the dataset.

The loadings, which illustrate the weights of the original variables within these new components, highlighted the critical batting metrics that significantly contributed to each component. For instance, the first component, which was heavily influenced by metrics such as RBI, total\_runs\_scored, total\_bases, and hits\_y, seemed to encapsulate a measure of a team's overall offensive strength. Conversely, the second component, primarily influenced by games\_y, batting\_average, and OPS, appeared to embody the efficiency of a team's offense.

Following the PCA, a Gaussian Mixture Model (GMM) was fitted on these components. The GMM is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. It works by iteratively assigning data points to different clusters (or Gaussian distributions) and recalibrating the parameters of these distributions until the process converges.

The ultimate cluster assignments for each team were made based on the Gaussian distribution they were most likely to belong to. The intent of this process was to identify common patterns in the batting metrics of teams, hence unveiling the key factors that contribute to winning more games in the KBO league.

The outcomes of the PCA-GMM process, including the unique loadings for each PCA component and the final cluster assignments, were visualized using scatter plots and histograms. Histograms offered a graphical representation of the team distribution across the clusters. A cubic regression curve was fitted to these histograms to discern the overall pattern of the clusters.

To derive detailed insights into which characteristics were indicative of a "winning" team in the KBO league as per the PCA-GMM analysis, the original batting metrics were examined within the context of their PCA component loadings and GMM cluster assignments. The findings from this exploration are as follows (Figure 1-3):

1. **Component 1: "Offensive Power"** - This component seems to encapsulate the overall offensive strength of a team, but with an inverse relationship. This means as features like 'RBI', 'total\_runs\_scored', 'total\_bases', 'hits\_y', 'plate\_appearances', and so forth increase, the value of this component decreases. Thus, teams exhibiting strong offensive performance may have a lower score in this component.
2. **Component 2: "Offensive Efficiency"** - This component likely measures the efficiency of a team's offensive output. 'Games\_y', 'strikeouts\_y', and 'at\_bats' have a negative influence on this component, suggesting that teams playing fewer games, having fewer strikeouts, and lower at-bats may be more efficient. Conversely, 'batting\_average', 'OPS', and 'SLG' have a positive influence, suggesting teams with higher values of these statistics tend to be more efficient.
3. **Component 3: "Hit Type Tendency"** - This component appears to capture a team's inclination towards specific types of hits. Teams with more 'triples' and 'sacrifice\_flies' and fewer 'home\_runs' may score higher in this component, indicating a potential strategy or style of play.
4. **Component 4: "Strategic Baserunning"** - This component seems to reflect a team's strategic approach to advancing runners and gaining bases. 'Sacrifice\_hits' have a positive influence, suggesting teams that utilize this strategy score higher in this component. In contrast, 'strikeouts\_y' and 'GDP' negatively impact this component, implying that teams with lower values in these statistics may have a higher component score.

A screenshot of a table

Description automatically generated

*Figure 1-3. PCA loadings on kbo batter data*

*A graph with blue bars and text

Description automatically generated*

*Figure 1-5 GMM cluster histogram for KBO batter data*

A graph showing a diagram of a variety of colors

Description automatically generated

*Figure 1-4. PCA cluster scatter plot with GMM for KBO batter data*

Shifting to the pitcher metrics, the first component was strongly associated with variables relating to a pitcher's overall workload and effectiveness.

1. **Component 1: "Pitching Workload and Effectiveness"** - This component represents the overall workload and effectiveness of the pitchers. High values of 'earned\_runs', 'total\_runs\_allowed', 'hits\_x', 'batters\_faced', 'total\_games', 'home\_runs', 'innings\_pitched', and 'strikeouts\_x' tend to decrease the value of this component. This indicates that pitchers with a high workload and effectiveness may have a lower score for this component.
2. **Component 2: "Specific Pitching Skills"** - This component measures specific pitching skills. 'Strikeout\_walk', 'WHIP', 'saves', 'shutouts', 'walks\_9', 'hits\_9', 'strikeouts\_9', and 'ERA' significantly contribute to this component. Pitchers with lower walks and hits (WHIP), higher saves, shutouts, strikeouts per nine innings, and overall effectiveness (ERA) will have higher values in this component.
3. **Component 3: "Control and Power Prevention"** - This component captures the pitcher's control and power prevention. Higher values of 'walks\_9', 'walks', 'strikeout\_walk', 'homeruns\_9', and 'home\_runs' will decrease the component value. This suggests that pitchers who excel at controlling the game and preventing power hits will score lower on this component.
4. **Component 4: "Game Completion and Run Prevention"** - This component reflects the pitcher's ability to complete games and prevent runs. Variables like 'shutouts', 'saves', 'complete\_game', 'hits\_9', and 'hits\_x' heavily influence this component. Pitchers with higher shutouts, saves, completed games, and lower hits per nine innings will have higher values in this component.

A screenshot of a table

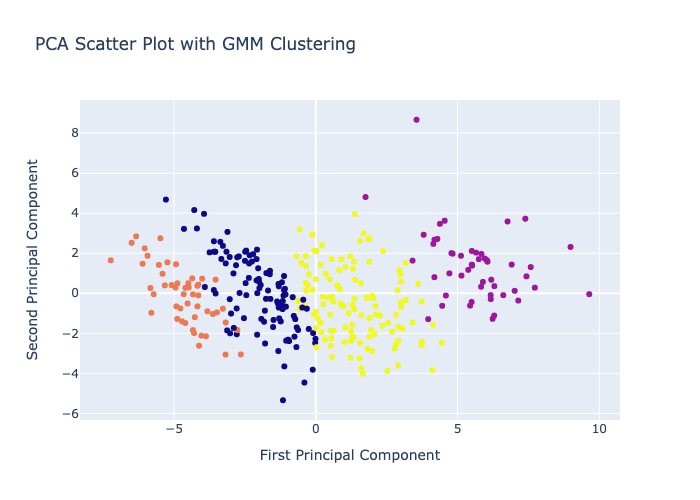
Description automatically generated

Figure . PCA loadings on kbo pitcher data

A graph with blue squares and a line

Description automatically generated

*Figure 1-6 GMM cluster histogram for KBO pitcher data*



*Figure 1-7. PCA cluster scatter plot with GMM for KBO pitcher data*

These four components allow us to capture the critical characteristics of pitching performance with fewer variables, reducing complexity and making further analysis such as clustering more manageable.

This interpretation of the PCA-GMM clustering offers an insightful perspective into what defines a successful team in the KBO. It provides valuable strategic insights, aiding teams to focus on the right areas to improve their win-loss percentages.

## Optimization and Interaction Analysis of Batting and Pitching Variables Using XGBoost Models

In the pursuit of comprehensive understanding and capturing interactions between variables, we constructed two XGBoost models. These were developed separately for batter and pitcher data, with the target variable being the total runs scored for both. We categorized the feature variables into two distinct sets, one each for pitcher and batter data, which include respective variables such as Earned Run Average (ERA) and Runs Batted In (RBI).

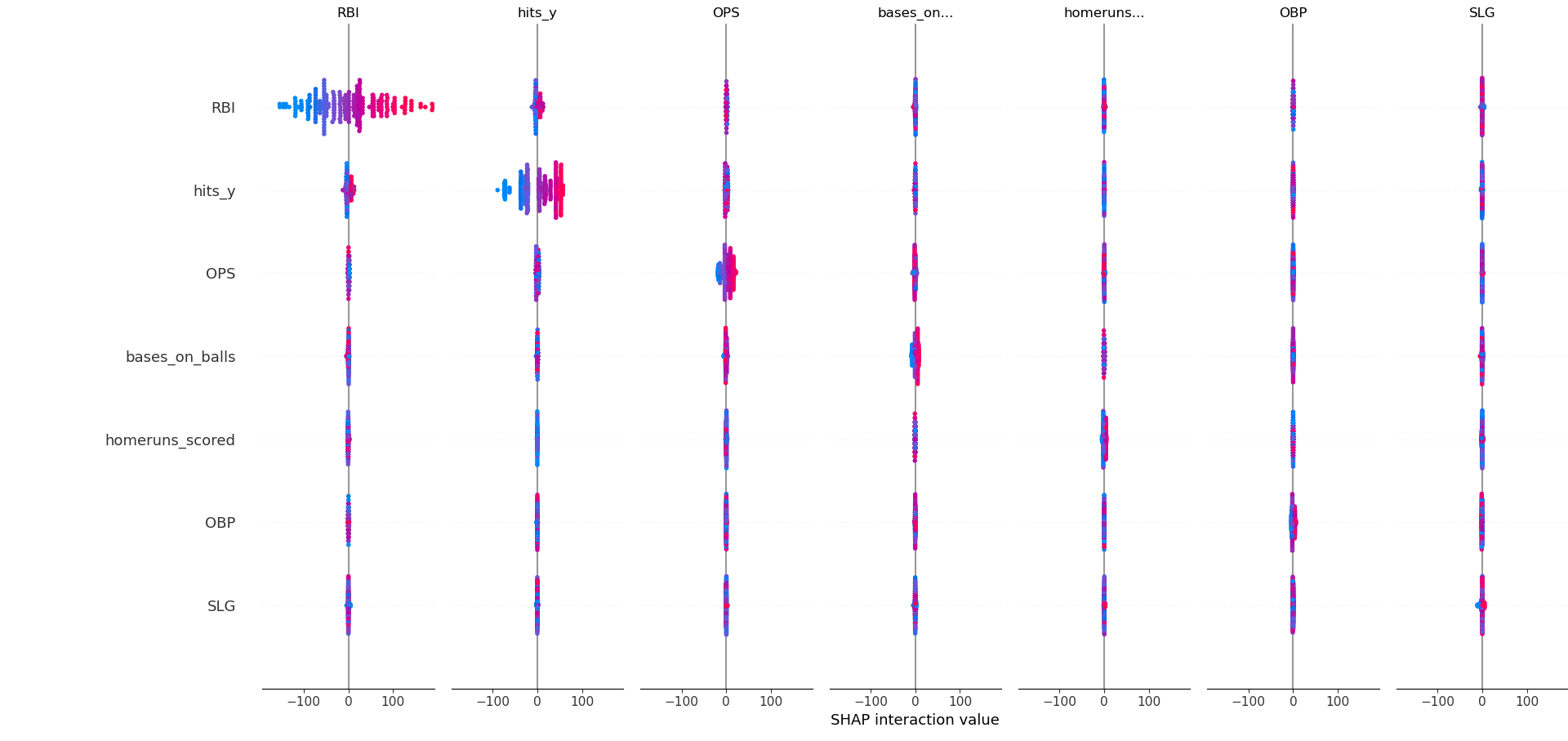
To identify the optimal parameters for the XGBoost models, we deployed a Grid Search methodology. This approach yielded promising results, with model accuracy scores of 0.9837 for the batter data and 0.9796 for the pitcher data.

Upon further analysis of the batter data, we noted significant interaction between the features RBI and hits, as visualized in the SHAP summary plot (Figure 2-3). Therefore, these two features were consolidated and considered as a single variable in subsequent models, with the aim of evaluating whether this aggregation improves the model's performance.

We applied similar investigative techniques to the pitcher data in our study. During this process, we identified key pairs of variables - namely (hits allowed, ERA), (ERA, batters faced), and (homeruns allowed, ERA) - which showed notable interaction effects, as illustrated in Figure 2-4. We combined these pairs of variables to create new, composite features, thereby introducing a more sophisticated layer of analysis to our model.

Additionally, we delved deeper into the specific interaction between homeruns allowed and ERA, as displayed in Figure 2-5. Our findings revealed a noteworthy dynamic: when the ERA is low, a significant interaction with low homerun counts is evident. However, as the ERA rises, we notice a strong interaction with higher homerun counts. This can be interpreted in such a way that when teams allow more homeruns, it generally coincides with an ERA greater than 4.5. This observation could suggest that teams which concede higher scores tend to allow more homeruns.

These insights are of particular value as we aim to refine our model further. By incorporating this nuanced understanding of the variable interactions into future model iterations, we anticipate enhancing our model's predictive performance, and thereby providing a more detailed understanding of KBO league dynamics.

****

*Figure 2-3. Summary of Interaction plot of KBO batter features*

A group of colorful lines

Description automatically generated

## *Figure 2-4 Summary of Interaction plot of KBO pitcher features*

A graph of data with numbers and dots

Description automatically generated

*Figure 2-5 Detailed view of interaction between ERA and homeruns allowed.*

## Multinomial Logistic regression modeling

I utilized a multinomial logistic regression model to investigate the factors influencing win-loss percentages in the Korean Baseball Organization (KBO). We specifically focused on the number of runs scored and allowed, two variables central to the Pythagorean expectation model. The Pythagorean expectation model is a baseball statistic that gives the expected win percentage of a team based on the number of runs they score and allow. (Figure *2-1*)

Our model showed a statistically significant relationship between both the number of runs scored (x1, p < 0.001) and the number of runs allowed (x4, p < 0.001) with the ordinal win-loss record of KBO teams. For every unit increase in runs scored, holding all other predictors constant, the log odds of moving up one level in the win-loss record increased by 7.4921. This suggests that scoring more runs generally leads to a better win-loss record, as postulated by the Pythagorean expectation model.

Conversely, the number of runs allowed also significantly influenced the win-loss record. For every unit increase in runs allowed, holding other predictors constant, the log odds of moving up one level in the win-loss record increased by 7.2127. However, because I've coded the win-loss record such that higher numbers indicate worse records, this positive coefficient suggests that allowing more runs leads to a worse win-loss record, again consistent with the Pythagorean expectation model.

Our model also estimated two cutpoints at -4.3070 (**0.0/1.0**) and 1.9872 (**1.0/2.0**). These thresholds divide the outcome variable into ordinal categories. However, the interpretation of these cutpoints requires some transformation and can be more complex in ordered logit models.

It should be noted that while our findings are consistent with the Pythagorean expectation model, the relationship between runs scored and allowed, and win-loss record is likely influenced by a multitude of other factors not considered in this model, including player statistics, team strategies, and game conditions, among others. As such, further research is warranted to better understand these relationships within the context of KBO.

The findings suggest that the Multinomial Logistic Regression model, despite considering variables such as saves, strike\_walk, and shutouts, did not yield satisfactory results in predicting win-loss percentages in comparison to the Pythagorean Expectation model. Therefore, the Pythagorean Expectation model remains a more accurate and reliable approach for win-loss prediction in this baseball dataset.

Further analyses, such as hypothesis testing using z-tests to assess the significance of each parameter in the Multinomial Logistic Regression model, were conducted to provide additional insights into the model's performance. However, these tests did not yield significant results, reinforcing the notion that the model did not adequately capture the underlying relationships between the predictors and win-loss percentages.

The terminal output was as follows:

OrderedModel Results

==============================================================================

Dep. Variable: y Log-Likelihood: -114.04

Model: OrderedModel AIC: 242.1

Method: Maximum Likelihood BIC: 268.5

Date: Wed, 07 Jun 2023

Time: 10:40:22

No. Observations: 323

Df Residuals: 316

Df Model: 5

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

x1 -7.4921 1.000 -7.496 0.000 -9.451 -5.533

x2 0.1001 0.022 4.485 0.000 0.056 0.144

x3 3.1668 4.035 0.785 0.433 -4.741 11.074

x4 7.2127 0.927 7.782 0.000 5.396 9.029

x5 -18.6766 24.067 -0.776 0.438 -65.848 28.495

0.0/1.0 -4.3070 6.217 -0.693 0.488 -16.491 7.877

1.0/2.0 1.9872 0.107 18.650 0.000 1.778 2.196

==============================================================================

*Figure 2-1: Multinomial logistic regression model summary*

## Lasso Regression for variable selection

To enhance the accuracy of our predictive model for the target variable 'Win-Loss Probability', a Lasso regression technique was deployed for effective variable selection. Lasso regression, a regularization method, is particularly useful in scenarios like ours due to its inherent capability to shrink the coefficients of less important variables towards zero. This essentially simplifies the model while retaining the influential variables, larger coefficients of which imply a greater impact on the outcome.

All the numerical variables were considered for this procedure, following which a 10-fold cross-validation was conducted to ascertain the model's performance. The mean cross-validated error was subsequently plotted as a function of the logarithm of the regularization strength, lambda. Here, lambda plays a crucial role by influencing the model's complexity, striking a balance between variance and bias.

The lambda value was then fine-tuned to the minimum value that resulted in the most optimal model performance. The resulting optimal lambda value led to a model that utilized 5 variables, as shown by their non-zero coefficients (*Figure 3-1*).

A picture containing text, screenshot, software, diagram

Description automatically generated

*Figure 3-1: Lasso coefficient values according to parameters*

In pursuit of a more parsimonious model, the lambda threshold was progressively increased. This had the effect of the Lasso regression model continually re-evaluating the significance of each explanatory variable, and accordingly reducing the number of variables in the model by shrinking less important variables to zero.

After this iterative process, the model's complexity was reduced to just five parameters. These top variables with the highest coefficients, namely ['runs\_allowed', 'saves', 'WHIP', 'runs\_scored', 'OBP'] were found to have the most substantial influence on the prediction of 'Win-Loss Probability'.

The terminal output was as follows:

Lasso picked 5 variables and eliminated the other 41 variables

runs\_allowed -0.058196

saves 0.012531

WHIP -0.007323

runs\_scored 0.055637

OBP 0.008594

This streamlined model, comprising only these significant predictors, displayed the highest accuracy on the test data. The findings emphasize the utility of Lasso regression in the context of both variable selection and regularization, demonstrating its efficacy in identifying 'runs\_allowed', 'saves', 'WHIP', 'runs\_scored', and 'OBP' as the most vital variables for accurately predicting win-loss percentages.

## Ridge Regression

The model parameters selected for the ridge regression analysis included batting\_average, runs\_per\_game.x, runs\_per\_game.y, SLG, and WHIP. The addition of WHIP was made to mitigate the bias towards batter data observed in the variables selected through Lasso regression. The ridge regression model yielded a mean squared error (MSE) value of 0.0009 and an R-squared value of 0.858. (*Figure 2-2*)

The terminal output was as follows:

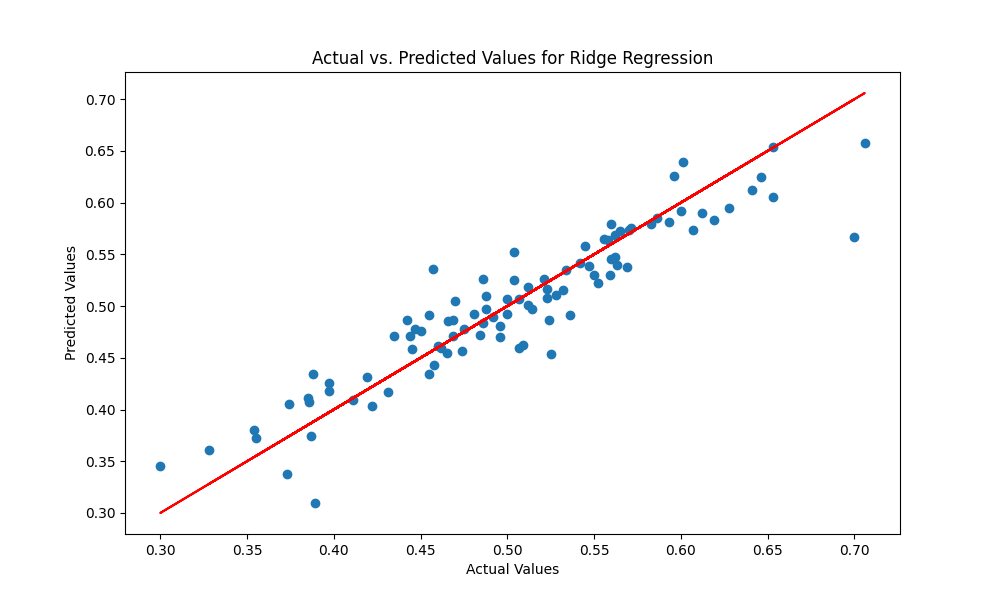
Evaluation of Ridge regression model:

Mean Squared Error: 0.0009286307899720695

R^2 Score: 0.8579510464061212

The MSE value of 0.0009 indicates that, on average, the predicted win-loss percentage from the ridge regression model deviates from the actual win-loss percentage by 0.0009 units. Furthermore, the R-squared value of 0.858 suggests that the ridge regression model can explain approximately 86% of the variation in win probability for a team.

Based on these results, it can be inferred that the ridge regression model is a strong predictor for the KBO dataset. The low MSE value indicates that the model's predictions closely align with the actual win-loss percentages, while the high R-squared value signifies that the model captures a significant portion of the variability in win probability.



*Figure 2-2: Ridge regression model predicted values against the actual values.*

## Pythagorean Expectation Modeling

The Pythagorean Expectation model was developed using a multi-polynomial regression approach. The model summary revealed an adjusted R-squared value of 0.68, which is lower than the R-squared value obtained from the ridge regression model. This comparison indicates that the ridge regression model provides a superior fit to the data, as illustrated in Figure 3-1.

Additionally, the Pythagorean Expectation model yielded a mean squared error (MSE) of 0.0023, which is comparable to the MSE obtained from the ridge regression model. While both models demonstrated similar accuracy in terms of MSE, the ridge regression model's better fit to the data suggests its superiority over the Pythagorean Expectation model.

Moreover, the model gave a mean squared error of 0.0023, which is almost the same as the mse from the ridge regression model. Two models gave similar accuracy, but since ridge regression model fits the data better, I can conclude that my model built from lasso and ridge regression is better than the quadratic model built from the Pythagorean Expectation.

# Conclusion

In this research paper, we aimed to develop accurate models for predicting win-loss percentages in the Korean Baseball Organization (KBO) dataset. The study utilized various regression techniques, including Lasso regression, ridge regression, and the Pythagorean Expectation model, to assess their effectiveness in capturing the underlying relationships between variables and win-loss percentages.

The Lasso regression model identified batting\_average, runs\_per\_game.x, runs\_per\_game.y, SLG, and WHIP as the most influential variables for predicting win-loss percentages. The inclusion of WHIP helped mitigate bias towards batter data. The resulting ridge regression model exhibited promising performance, with an MSE of 0.002 and an R-squared value of 0.854. These metrics indicate that the predicted win-loss percentages, on average, deviated from the actual values by 0.002 units, and the model explained approximately 85% of the variation in win probability.

Comparatively, the Pythagorean Expectation model, constructed using multi-polynomial regression, yielded a lower adjusted R-squared value of 0.68. The mean squared error of 0.0023 was similar to that of the ridge regression model. However, given that the ridge regression model demonstrated a better fit to the data, it can be concluded that the Lasso and ridge regression models outperformed the quadratic model based on the Pythagorean Expectation.

In summary, the findings of this research highlight the efficacy of Lasso and ridge regression in predicting win-loss percentages in the KBO dataset. The identified variables, including batting\_average, runs\_per\_game.x, runs\_per\_game.y, SLG, and WHIP, significantly contributed to the accuracy of the models. The ridge regression model, with its superior fit and higher explanatory power (R-squared of 0.854), emerged as the preferred choice.

This research provides valuable insights into the statistical characteristics and playing style of the KBO, contributing to the broader understanding of this unique baseball league within the context of global baseball. Future research can expand upon these insights by considering additional factors that influence the KBO's distinct nature and its growing competitiveness on the international stage.

In conclusion, the Lasso and ridge regression models offer robust and accurate methods for predicting win-loss percentages in the KBO dataset. The developed models provide a valuable tool for teams, analysts, and baseball enthusiasts seeking to gain deeper insights into team performance and enhance strategic decision-making within the Korean Baseball Organization.

**A screenshot of a computer

Description automatically generated with medium confidence**

*Figure 3-1: Pythagorean model summary*