# Predicting Winning Percentage in Major League Baseball Using Regression Analysis (2000-2022)

A Python-based study of offensive and defensive metrics

Kyla Igawa
Computing and Data Science
Wentworth Institute of Technology
Boston, MA.
igawak@wit.edu

## ABSTRACT

This project applies multiple linear regression to Major League Baseball team statistics from 2000 to 2022 to identify which offensive and defensive metrics best predict winning percentage. Using Python's scikit-learn library, the models were fit in different ways: with all predictors combined, offense only, defense only, as well as era subsets. Standardization techniques were used throughout to ensure that coefficients were comparable across variables. The results confirm that run differential (runs scored minus runs allowed) is the single strongest predictor of winning percentage, while defensive metrics explain slightly more variance than offensive metrics. ERA-specific regression models revealed shifts in predictor importance across decades, reflecting changes in baseball strategy and performance trends. Visualizations, including coefficient bar charts and scatterplots, illustrate the findings clearly.

## KEYWORDS

Linear Regression, Baseball Analytics Metrics, Offense vs. Defense, MLB Performance Analysis.

## 1 Introduction

Baseball has always been a sport where statistics derive both strategy and evaluation. From the early adoption of batting averages to the modern emphasis on advanced metrics, analysts and managers have tried to quantify what leads to more wins. One of the most consistent findings in baseball research is that run differential (the difference between runs scored and runs allowed) is a great predictor of win percentage [3]. However, debates continue over the contributions of offense and defense, and whether their importance has shifted across eras.

This project uses Python-based regression modeling to explore these questions. By fitting linear regression models with offensive and defensive predictors separately, and by analyzing subsets of seasons grouped into eras (2000s, 2010-2015, 2016-2022), the study seeks to find how different metrics contribute to winning percentage. The approach used builds on established sabermetric research but emphasizes reproducibility and clarity through code, visualization, and standardized coefficients. The goal of this paper is to provide interpretable results that highlight both the enduring importance of run differential and the evolving role of other predictors over time.

## 2 Data

There was a single dataset used for this analysis. It included team level information for the seasons 2000-2022.

### 2.1 Source of dataset

The dataset used in this project is the Lahman Baseball Database (Teams.csv), covering Major League Baseball statistics from 2000-2002 [1]. The database was originally created by Sean Lahman in the 1990s to make baseball statistics freely available to the public and has since been expanded and maintained by a team of researchers under the Society for American Baseball Research (SABR).

### 2.2 Characters of the datasets

The dataset is taken from the Lahman Baseball Database (Teams.csv), which provided season level statistics for MLB teams. For this project, we focused on 2000-2022, selecting variables that were relevant to offensive and defensive performance. The dataset was stored in csv format and was updated to an xlsx format which contains one row per team season.

Key variables used in the regression analysis are summarized below:

- Runs scored by team (R: offensive)
- Runs allowed by team (RA: defensive)
- Home runs hit (HR: offensive)
- On-base percentage (OBP: offensive)

[*] This report was prepared as part of the Data Science Foundations course project at Wentworth Institute of Technology.
[†] Author Kyla Igawa is a student in the Department of Computer and Data Science, Wentworth Institute of Technology, Boston, MA.
This report is prepared for academic purposes only. The data used comes form the Lahman Baseball Database (SABR, Creative Commons Attribution-ShareALike 3.0 License)

- Slugging percentage (SLG: offensive)
- Batting average (AVG: offensive)
- Earned run average (ERA: defensive)
- Errors committed (E: defensive)
- Run differential (RunDiff: combined)
- Winning percentage (WinPct: target)
- Year (year: used for era grouping)
- Era (ERA: used for era models)

# 3 Methodology

The analytical framework for this paper is multiple linear regression, implemented in Python using the ***LinearRegression*** class from ***scikit-learn*** [2]. Regression was chosen because it provides interpretable coefficients that quantify the importance of predictors in explaining winning percentage. The general form of the model is:

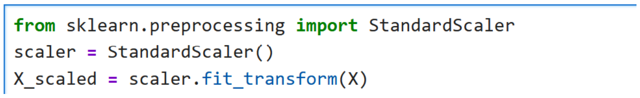
$$y_{pred} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (1)$$

Where y is the predicted winning percentage, $x_i$ are the predictors (such as Runs, ERA, or OBP), and $\beta_i$ are the estimated coefficients. To evaluate model performance, two metrics were used: the coefficient of determination R-squared, which measures the proportion of variance explained, and the mean squared error (MSE), which measures average prediction error:

$$R^2 = 1 - \frac{\Sigma(y_i - y_{pred,i})^2}{\Sigma(y_i - y_{bar,i})^2} \qquad (2)$$

$$MSE = \frac{1}{n}(y_i - y_{pred,i})^2 \qquad (3)$$

Because the predictors are measured on different scales (for example, runs are in the hundreds while batting average is between 0 and 1), all variables were standardized using scikit-learns ***StandardScaler.*** This transformation centers each predictor at zero mean and unit variance, which makes sure that coefficients are directly comparable.

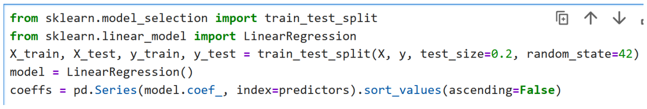


```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

**Figure 1: Python code using StandardScaler to standardize predictors, so coefficients are directly comparable across variables.**

## 3.1 Experiment 1: Full Model

The first experiment fit a regression model using all predictors simultaneously: Runs (R), Runs Allowed (RA), Home Runs (HR), On-Base Percentage (OBP), Slugging Percentage (SLG), Batting Average (AVG), Earned Run Average (ERA), Errors (E), and Run Differential (RunDiff). After splitting the dataset into training and testing sets (80/20), the model was trained and evaluated. Coefficients were extracted and sorted to reveal which predictors had the strongest positive or negative influence on winning percentage.

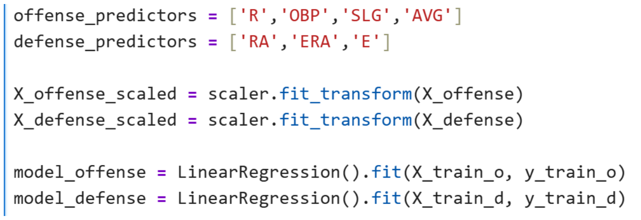


```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LinearRegression()
coeffs = pd.Series(model.coef_, index=predictors).sort_values(ascending=False)
```

**Figure 2: Code fitting a multiple linear regression model with all predictors, extracting and sorting standardized coefficients.**

## 3.2 Experiment 2: Offense vs. Defense

To look at the contributions of different aspects of baseball, the predictors were divided into offensive metrics (R, HR, OBP, SLG, AVG) and defensive metrics (RA, ERA, E). Separate regression models were fit for each group, again using standardized predictors and an 80/20 train-test split. Comparing the resulting R-squared values allowed us to understand whether scoring runs or preventing runs explained more variance in winning percentage.

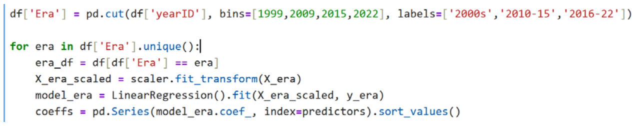


```
offense_predictors = ['R','OBP','SLG','AVG']
defense_predictors = ['RA','ERA','E']

X_offense_scaled = scaler.fit_transform(X_offense)
X_defense_scaled = scaler.fit_transform(X_defense)

model_offense = LinearRegression().fit(X_train_o, y_train_o)
model_defense = LinearRegression().fit(X_train_d, y_train_d)
```

**Figure 3: Code separating offensive and defensive predictors, fitting independent regression models, and preparing R-Squared comparison.**

## 3.3 Experiment 3: Era-Specific Models

Finally, to explore how the importance of predictors has shifted over time, seasons were grouped into three eras: 2000s (2000-2009(, 2010-2015, and 2016-2022. For each era, a regression model was fit using the same set of predictors as before, standardized as well. Coefficients were extracted and compared across eras and visualized in side-by-side bar charts to highlight changes in predictor importance.



```
df['Era'] = pd.cut(df['yearID'], bins=[1999,2009,2015,2022], labels=['2000s','2010-15','2016-22'])

for era in df['Era'].unique():
    era_df = df[df['Era'] == era]
    X_era_scaled = scaler.fit_transform(X_era)
    model_era = LinearRegression().fit(X_era_scaled, y_era)
    coeffs = pd.Series(model_era.coef_, index=predictors).sort_values()
```

**Figure 4: Code looping through defined eras, fitting regression models for each subset, and storing coefficients for comparison.**

## 3.4 Summary of Experiments

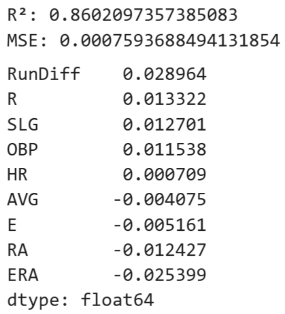
Across all the experiments, multiple linear regression provided interpretable results that quantified the impact of offensive and defensive metrics on winning percentage. Standardization was a critical preprocessing step, ensuring that coefficients could be compared across variables, a key piece of our analysis. By structuring our analysis into a full model, offensive vs defensive comparison, and era-specific regressions, the methodology allowed for broad and specific insights into how team statistics relate to team success.

# 4 Results

The results of the regression experiments are presented in this section. Each highlights different aspects of the relationship between team statistics and winning percentage. Numerical outputs (R-squared, coefficients) and visualizations (bar charts, scatterplots) are included to illustrate the findings.

## 4.1 Full Model (All Predictors)

The regression model using all predictors explained a substantial portion of the variance in winning percentage. The standardized coefficients revealed that **Run Differential** was the strongest positive predictor, while **ERA** and **Runs Allowed (RA)** were the strongest negative predictors



```
R²: 0.8602097357385083
MSE: 0.0007593688494131854

RunDiff    0.028964
R          0.013322
SLG        0.012701
OBP        0.011538
HR         0.000709
AVG       -0.004075
E         -0.005161
RA        -0.012427
ERA       -0.025399
dtype: float64
```

**Figure 5: Output from script showing value of standard coefficients from the full regression model.**

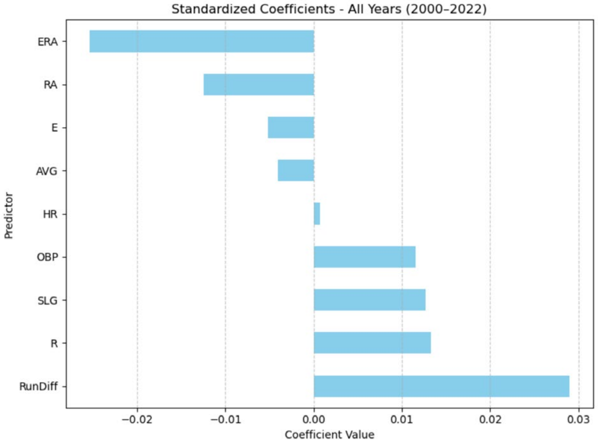
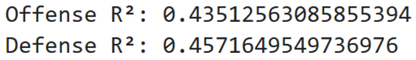


**Figure 6: Horizontal bar chart of standardized coefficients from the full regression model, showing relative predictor importance.**

This confirms that run differential captures the combined effect of scoring and preventing runs, making it the most reliable single predictor of success.

## 4.2 Offense vs. Defense Models

When predictors were separated into offense and defensive groups, the defensive model achieved a slightly higher R-squared value than the offensive model. This suggests that preventing runs (through ERA, RA, and limiting errors) explains slightly more variance in winning percentage than scoring runs.



```
Offense R²: 0.43512563085855394
Defense R²: 0.4571649549736976
```

**Figure 7: Printed R-Squared values comparing offense-only and defense-only models, highlighting defense's slightly stronger explanatory power.**

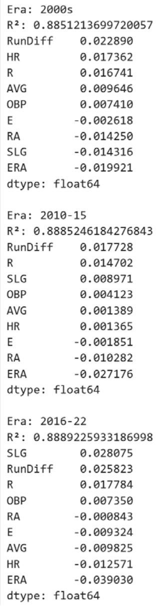
The result highlights the importance of pitching and defense in determining team success, even in an era often associated with offensive power.

This confirms that run differential captures the combined effect of scoring and preventing runs, making it the most reliable single predictor of success.

## 4.3 Era-Specific Models

Regression models fit separately for each era revealed shifts in predictor importance over time.

- In the 2000s, offensive metrics such as Home Runs and On-Base Percentage had stronger coefficients
- In 2010-2015, defensive metrics such as ERA and Runs Allowed dominated
- In 2016-2022, Slugging Percentage and Run differential emerged as the most influential predictors



```
Era: 2000s
R²: 0.8851213699720057
RunDiff    0.022890
HR         0.017362
R          0.016741
AVG        0.009646
OBP        0.007410
E         -0.002618
RA        -0.014250
SLG       -0.014316
ERA       -0.019921
dtype: float64

Era: 2010-15
R²: 0.8885246184276843
RunDiff    0.017728
R          0.014702
SLG        0.008971
OBP        0.004123
AVG        0.001389
HR         0.001365
E         -0.001851
RA        -0.010282
ERA       -0.027176
dtype: float64

Era: 2016-22
R²: 0.8889225933186998
SLG        0.028075
RunDiff    0.025823
R          0.017784
OBP        0.007350
RA        -0.000843
E         -0.009324
AVG       -0.009825
HR        -0.012571
ERA       -0.039030
dtype: float64
```

**Figure 8: Output showing regression coefficients for different eras (2000-2009, 2010-2015, 2016-2022).**

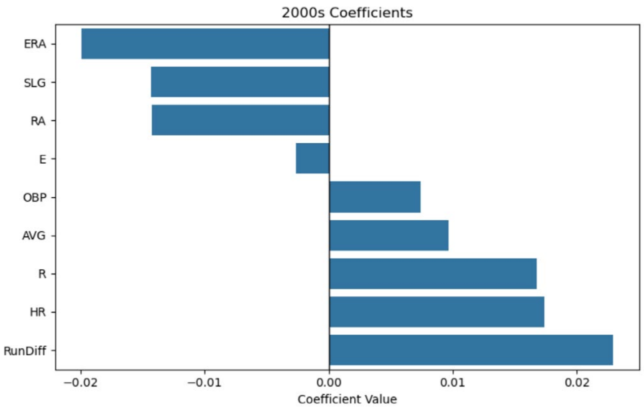


**Figure 9: Individual bar chart of regression coefficients for the 2000s, illustrating offensive metrics as stronger positive predictors.**

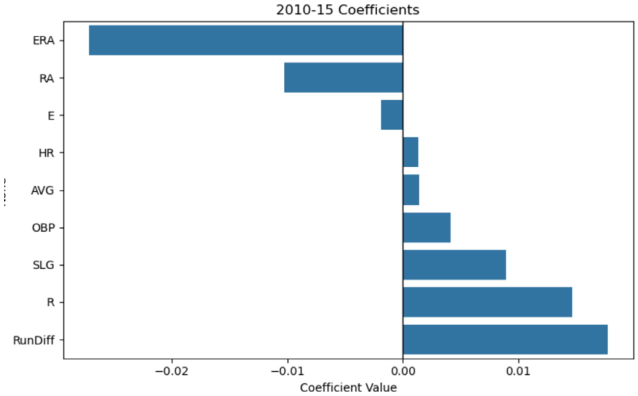


**Figure 10: Individual bar chart of regression coefficients for 2010-2015, showing defensive metrics dominating positive predictor importance.**

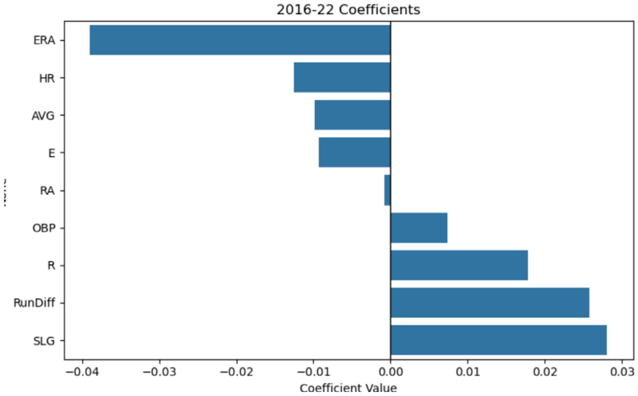


**Figure 11: Individual bar chart of regression coefficients for 2016-2022, highlighting slugging percentage and run differential as key predictors.**

These shifts reflect changes in baseball strategy, including the rise of power hitting in the early 2000s, the dominance of pitching in the mid-2010s, and the emphasis on slugging and overall efficiency in recent years.

### 4.5 Run Differential vs. Winning Percentage

A scatterplot of Run Differential against Winning Percentage shows a clear upward trend: teams with higher run differentials consistently achieve higher-winning percentages. The relationship is nearly linear, visually confirming run differential as the strongest predictor of success.

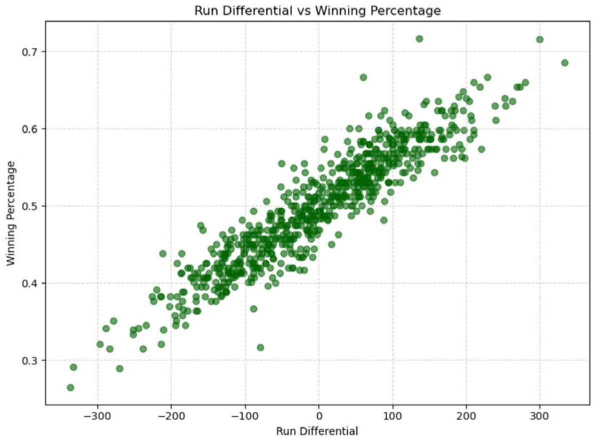


**Figure 12: Scatterplot showing the nearly linear relationship between run differential and winning percentage, confirming run differential as the strongest single predictor of win percentage.**

## 5 Discussion

The regression models provided have clear insights but also reveal some limitations. Linear regression assumes simple, independent relationships, yet many baseball metrics overlap (like OBP and AVG), this creates multicollinearity that can weaken coefficient accuracy. Using season-level aggregates also hides in-season variability and contextual factors like injuries or bullpen usage. Era-specific models highlighted shifts in predictor importance, but smaller sample sizes reduce statistical power and might exaggerate certain trends. If you were to extend the analysis to include decades, that approach might make space for more concrete takeaways of how the game has progressed.

Future work could improve results by applying regularized methods (Ridge or Lasso) to stabilize coefficients, incorporating advanced metrics such as WAR or FIP for richer analysis, and exploring game level or non-linear models to capture interactions that linear regression cannot. Despite these limits, the study confirms that run differential as the strongest predictor and shows how offense and defense contributions evolve across eras.

## 6 Conclusion

This project used multiple linear regression in Python to analyze MLB team statistics from 2000-2022, testing full models, offense-only and defense-only models, and era specific regressions. Across all experiments, Run Differential proved to be the strongest predictor of winning percentage, while defensive metrics explained slightly more variance than offensive ones. Era comparisons showed shifts in importance, with home runs and OBP more influential in the 2000s, pitching dominating in the early 2010s, and slugging rising in recent years. These findings reinforce sabermetric principles and demonstrate how statistical modeling can guide real-world strategy by highlighting the evolving balance between offense and defense in baseball success.

## ACKNOWLEDGMENTS

I would like to express my gratitude to my professor for her guidance and constructive feedback throughout the semester. I am also indebted to the Society for American Baseball Research (SABR) and Sean Lahman for making the Lahman Baseball Database freely available, which provided the foundation for this analysis. Their commitment to preserving and sharing baseball statistics made it possible to explore the relationship between team performance metrics and winning percentage in a rigorous, reproducible way. Finally, I acknowledge the broader sabermetric community, whose pioneering work continues to inspire and inform projects such as this one.

## REFERENCES

[1] Lahman Baseball Database (no date) Society for American Baseball Research. Available at: https://sabr.org/lahman-database/ (Accessed: 28 November 2025).

[2] *Learn* (no date) *scikit*. Available at: https://scikit-learn.org/stable/index.html (Accessed: 28 November 2025).

[3] Franks, D. and Holmes, D. (2019) The most important stat in baseball, The Baseball Scholar. Available at: https://thebaseballscholar.com/2017/02/07/the-most-important-stat-in-baseball/ (Accessed: 28 November 2025).