Analyzing the Key Drivers of Success in Major League Baseball

Jordan Badstuebner

"Baseball is 90% mental. The other half is physical."
- Yogi Berra

I.   ABSTRACT

This study examines which factors correlate most strongly with winning in Major League Baseball (MLB). Through a quantitative approach, we analyze statistical variables such as On-Base Percentage (OBP), Slugging Percentage (SLG), and Runs Scored (RS) to understand how they influence a team's success. While the idea that scoring more runs than the opponent leads to victory is intuitive, this paper provides a structured mathematical foundation that shows how specific metrics contribute to winning. By utilizing linear regression models, we not only find the relationships but quantify the extent to which they impact success.

II.   INTRODUCTION

In baseball, statistics are extensively used to evaluate team and player performance. An essential question for analysts and fans alike is determining which statistics best predict a team's success. To address this, we can use linear regression, a mathematical model that helps us understand the relationship between different variables.

In our analysis, we use the general linear regression formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon, \tag{1}$$

where Y represents the response variable (wins, in our case), X_i are the predictor variables (such as OBP and SLG), β_i are the coefficients that represent the weight or influence of each predictor, and ϵ is the error term, accounting for variability not explained by the model.

III.   DATA COLLECTION AND METHODOLOGY

The data spans MLB teams from 1962 to 2012, encompassing statistics like Wins (W), Runs Scored (RS), Runs Allowed (RA), On-Base Percentage (OBP), and Slugging Percentage (SLG). These variables give insight into a team's offensive and defensive capabilities.

To determine which factors correlate most with winning, we applied linear regression. In particular, we looked at RS as a function of OBP and SLG, using:

$$RS = \beta_0 + \beta_1 \text{OBP} + \beta_2 \text{SLG} + \epsilon. \tag{2}$$

Additionally, we explored a multiple linear regression model that includes RS and RA to explain Wins:

$$W = \beta_0 + \beta_1 \text{RS} + \beta_2 \text{RA} + \epsilon. \tag{3}$$

This equation allows us to see how runs scored and runs allowed affect wins, giving a more comprehensive picture of team success.

IV.   RESULTS

Our analysis revealed that OBP and SLG are indeed significant in predicting RS. The model for RS with OBP and SLG had an R^2 value of 0.92, indicating that these variables explain 92% of the variance in RS. This strong result shows how crucial these offensive metrics are to generating runs.

For Wins, we found that:

$$W = 0.1 \times RS - 0.1 \times RA. \tag{4}$$

This suggests that each additional run scored increases wins by approximately 0.1, while each run allowed decreases wins by the same amount. This model, with an R^2 of 0.88, highlights the importance of both scoring runs and limiting opponents' runs.

V. DISCUSSION

While it might seem obvious that scoring more than the opponent results in a win, quantifying this relationship provides deeper insights. For example, the coefficients from our model show that offense and defense are equally important in baseball—both need to be managed for consistent success.

We can further explore these relationships through calculus. If we define the rate of change of Wins W with respect to Runs Scored RS as $\frac{dW}{dRS}$, then:

$$\frac{dW}{dRS} = \beta_1 \approx 0.1. \tag{5}$$

This tells us that small increases in scoring have a measurable impact on winning. Similarly, we define the rate of change with respect to Runs Allowed RA as:

$$\frac{dW}{dRA} = -\beta_2 \approx -0.1. \tag{6}$$

This reinforces the significance of defense, as each run prevented is just as valuable as a run scored.

VI. CONCLUSION

In this paper, we demonstrated that while OBP and SLG are indirect predictors of winning by boosting RS, it is the combination of RS and RA that directly correlates with wins. Our findings not only validate the importance of both offense and defense but also provide a quantitative measure of their impact. Future studies could include additional variables like fielding percentage to see if they enhance the predictive power of the model.

REFERENCES

[1] James, B. (1985). *The Bill James Historical Baseball Abstract*. Villard Books.