

**Regression Analysis Report: The Impact of Common Advanced Baseball Statistics on Weighted On-Base**

\

I. Introduction to Theme and Motivation

II. Data Introduction

III. Model Selection and Model Diagnostics

IV. Model Remediation

V. Model Interpretation and Summary

**Group 12**

410978004 統計二 郭依璇(YI-HSUAN KUO)

410378011 統計二 譚靖蓉

410978034 統計二 李海寧

410978044 統計二 謝瑋芸

410875019 統計三 林哲兆

410875024 統計三 李亮頤

# I. Introduction to Theme and Motivation

Baseball is a popular sport in Taiwan, and watching the players' effort on the field is just one aspect of the game's appeal. From a data perspective, the sport offers its own unique intrigue. Furthermore, with the advancement of technology, many teams have begun to adjust their training methods for players based on data analysis. Therefore, we aim to study the impact of batting quality data on weighted on-base percentage (wOBA).

# II. Data Introduction

The data source is an international website for advanced statistical analysis of Major League Baseball (MLB) in the United States, savant (<https://baseballsavant.mlb.com/>). The sample consists of advanced data for hitters in the MLB from 2015 to 2022, who have a certain number of plate appearances. This includes the hitters' weighted on-base percentage (wOBA), age, average exit velocity, average launch angle, sweet spot percentage, and barrel rate. Since the data includes entries for the same hitters in different years, there may be issues with the data not being independent of each other.

Variable Introduction：

1. woba (Weighted On-Base Average): A metric that combines the insights of on-base percentage and slugging percentage to evaluate a hitter's offensive abilities. It is structured somewhat like a combination of the formulas for both, taking into account various hitting outcomes and assigning them more rational weights for calculation.
2. player\_age: We categorize age into two groups, with ages 25 to 29 being marked as 1, and all others as 0, because the age range of 25 to 29 is commonly recognized as the peak performance period for MLB athletes. Besides, we rename player\_age as age0.
3. exit\_velocity\_avg (Average Exit Velocity): Represents the initial speed of the ball after it has been hit by the bat
4. launch\_angle\_avg (Average Launch Angle): The launch angle refers to the angle between the trajectory of the ball after being hit and the ground, at the moment the ball leaves the bat.
5. sweet\_spot\_percent (Sweet Spot)(%): The proportion of a hitter's batted ball events where the launch angle falls within the sweet spot.
6. barrel\_batted\_rate(Barrel Rate)(%): The proportion of batted balls that qualify as barrels, explicitly defined as hits with a launch angle and exit velocity that result in a batting average of at least .500 and a slugging percentage of at least 1.500. This represents instances of hitting the ball solidly, with both the launch angle and exit velocity within an optimal range.
7. Interactions: x6 = launch\_angle\_avg \* barrel\_batted\_rate

x7 = sweet\_spot\_percent \* barrel\_batted\_rate

# III. Model Selection and Model Diagnostics

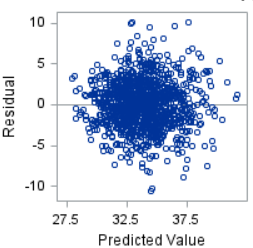
The following establishes two models to explain the relationship between wOBA and other variables.

## Model 1: Without Interaction

𝑤𝑜𝑏𝑎 = 𝛽0 + 𝛽1 𝑎𝑔𝑒0 + 𝛽2 𝑒𝑥𝑖𝑡\_𝑣𝑒𝑙𝑜𝑐𝑖𝑡𝑦\_𝑎𝑣𝑔 + 𝛽3 𝑙𝑎𝑢𝑛𝑐ℎ\_𝑎𝑛𝑔𝑙𝑒\_𝑎𝑣𝑔

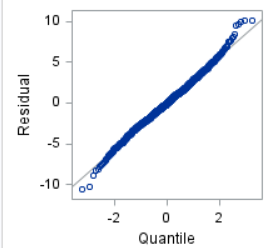
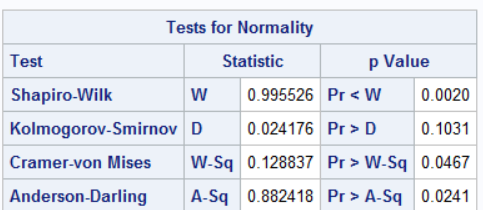
+ 𝛽4 𝑠𝑤𝑒𝑒𝑡\_𝑠𝑝𝑜𝑡\_𝑝𝑒𝑟𝑐𝑒𝑛𝑡 + 𝛽5 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒 + 𝜀𝑖

1. Linear Tendency：Passed



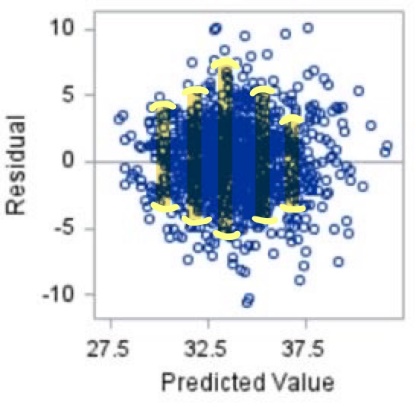
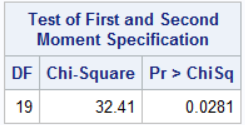
The residual plot shows that at the location marked by the red box, there are more positive residuals (the distribution is not symmetric above and below the horizontal line r = 0). However, we consider that the proportion of data in this location is extremely low relative to the overall dataset, and therefore still determine that this data conforms to linear tendency.

1. Error Normality：Passed



The Normal Q-Q plot shows residuals deviating from the straight line at the tails, but still largely conforms to the 45-degree line, leading to the determination that the residuals are normally distributed. Among the four normality tests conducted, except for the Kolmogorov-Smirnov Test, the rest do not reject the hypothesis of error normality. In summary, we believe that the deviations at the tails of the Q-Q plot are the reason for not passing some normality tests, but since the vast majority of the data aligns with the 45-degree line, we ultimately conclude that the residuals are normally distributed.

1. Constant Variance：Not Passed



1. The residual plot shows that around the predicted value () of approximately 33, the distribution of residuals appears to be wider than at the locations of around 30 and 37.5. This indicates 1that the variance of the residuals changes with the change in

(2) The result of the White Test leads to the rejection of constant error variance.

## Model 2: With Interaction

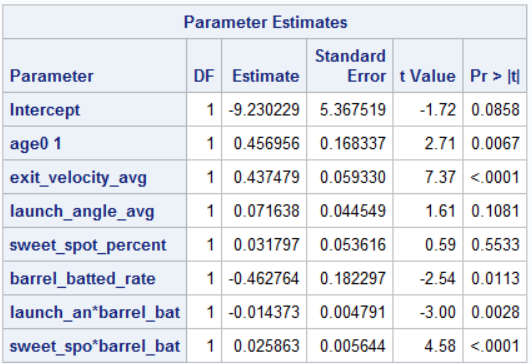
Using Backward Elimination, a model is constructed retaining independent variables and main effects of interactions with a p-value < 0.05.

x6 = launch\_angle\_avg \* barrel\_batted\_rate x7 = sweet\_spot\_percent \* barrel\_batted\_rate

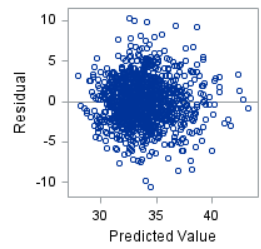
𝑦(𝑤𝑜𝑏𝑎) = 𝛽0 + 𝛽1 𝑎𝑔𝑒0 + 𝛽2 𝑒𝑥𝑖𝑡\_𝑣𝑒𝑙𝑜𝑐𝑖𝑡𝑦\_𝑎𝑣𝑔 + 𝛽3 𝑙𝑎𝑢𝑛𝑐ℎ\_𝑎𝑛𝑔𝑙𝑒\_𝑎𝑣𝑔 +

𝛽4 𝑠𝑤𝑒𝑒𝑡\_𝑠𝑝𝑜𝑡\_𝑝𝑒𝑟𝑐𝑒𝑛𝑡 + 𝛽5 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒 + 𝛾1 𝑥6 + 𝛾2 𝑥7 + 𝜀𝑖



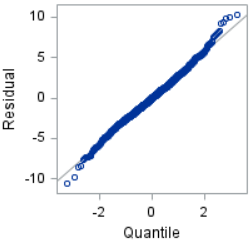
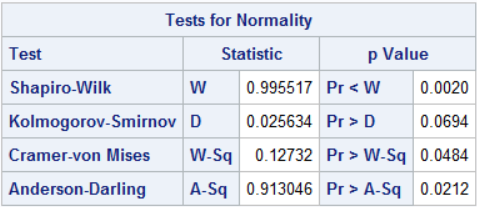


1. Linear Tendency：Passed



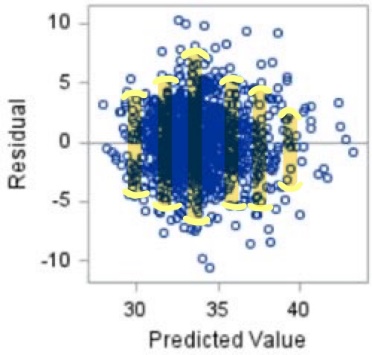
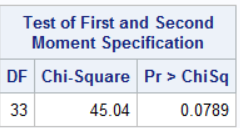
Using Backward Elimination, a model is constructed retaining independent variables and main effects of interactions with a p-value < 0.05.

1. Error Normality：Passed



The Normal Q-Q plot shows residuals deviating from the straight line at the tails, but still largely conforms to the 45-degree line, leading to the determination that the residuals are normally distributed. Among the four normality tests conducted, except for the Kolmogorov-Smirnov Test, the rest do not reject the hypothesis of error normality. In summary, we believe that the deviations at the tails of the Q-Q plot are the reason for not passing some normality tests, but since the vast majority of the data aligns with the 45-degree line, we ultimately conclude that the residuals are normally distributed.

1. Constant Variance：Passed



1. The residual plot shows that the distribution of residuals appears wider around the estimated value () of approximately 35 compared to when is around 30 or 40, indicating that the variance of residuals changes with the estimated values.
2. The result of the White Test is "do not reject constant error variance." Despite the residuals' variance seeming less constant in the residual plot, we choose to trust the outcome of the White Test.

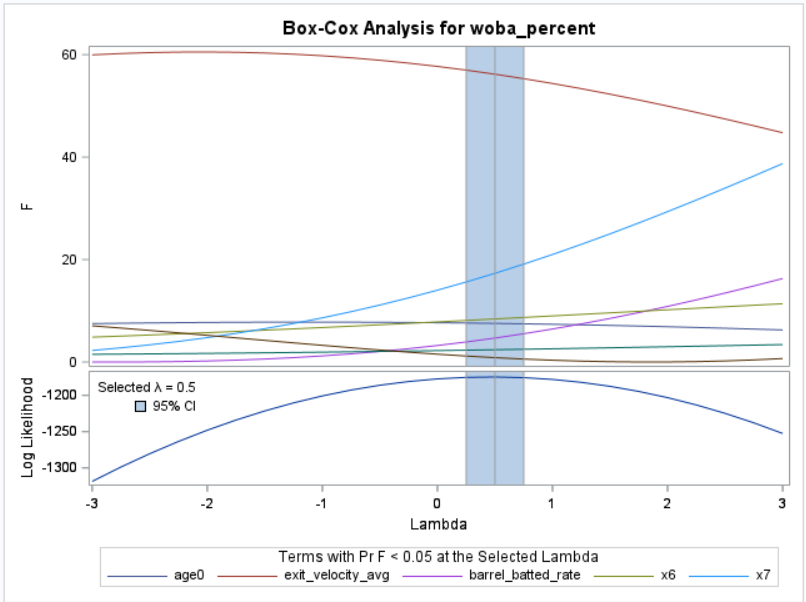
Conclusion:

|  |  |  |
| --- | --- | --- |
|  | **Model 1** | **Model 2** |
| *Linear Tendency* | Passed | Passed |
| *Error Normality* | Passed | Passed |
| *Constant Variance* | Not Passed | Passed |

The observation suggests that the interaction has addressed the issue of failing the Constant Variance test, hence Model 2 has been selected for analysis.

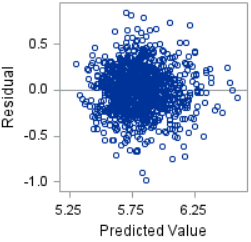
# IV. Model Remediation

Given that the p-value for Model 2's White Test is 0.0789, only slightly above 0.05, we have decided to attempt model remediation in hopes of improvement.

Following the model diagnostics, it is evident that Constant Variance is not met, thus opting to transform the response variable (y). Referring to the results of the Box-Cox Transformation (as shown in the figure below), we choose to use 𝑔(𝑦𝑖) =

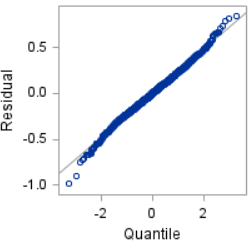
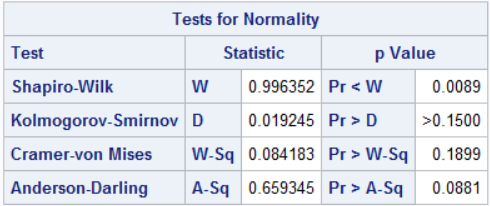
The following is the diagnosis after remediation:

1. Linear Tendency：Passed



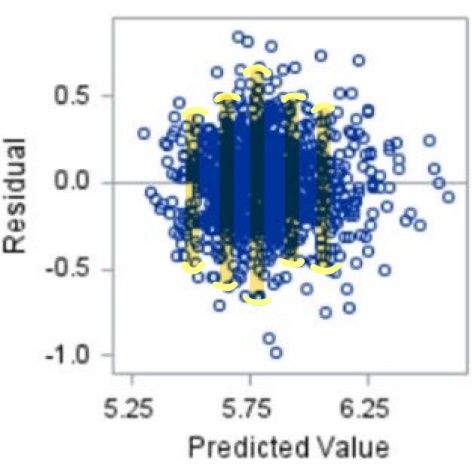
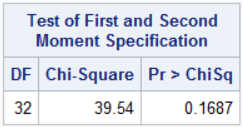
The residual plot shows that the distribution of is symmetric above and below the horizontal line r = 0 , thus determining that this data conforms to linear tendency.

1. Error Normality：Passed



The Normal Q-Q plot shows that the residuals deviate from the straight line at the tails but still largely conform to the 45-degree line, leading to the determination that the residuals are normally distributed. Among the four normality tests conducted, except for the Shapiro-Wilk Test, the rest do not reject error normality. In summary, it is ultimately determined that the residuals conform to a normal distribution.

1. Constant Variance：Passed



1. The residual plot shows that around approximately at 5.75, the distribution of residuals appears to be wider than at around 5.3 and 6.2, indicating that the variance of the residuals changes with the change in.
2. The result of the White Test is to not reject constant error variance. From the residual plot, although the variance of the residuals seems to be more variable, we choose to trust the result of the White Test

From the residual plot, it is observed that the variance of the residuals seems to be more variable, but we choose to trust the results of the White Test.

Conclusion:

Although the test results improved after remediation, the extent of improvement was minimal and the transformed response variable is not conducive to interpretation. Therefore, we still opt to use Model 2 before remediation.

# V. Model Interpretation and Summary

# 

* The significance of the intercept estimate:

Here, the intercept refers to the parameter in the model without any variables: 𝛽0. Its significance is: when the age is outside of 25~29 years, and the average exit velocity is 0 MPH, the average launch angle is 0°, and both the barrel rate and sweet spot percentage are 0%, the mean weighted on-base percentage is -9.23023%. However, such data would not occur in reality.

* The significance of age grouping:

With other independent variables held constant, we have 95% confidence that, on average, players aged 25~29 have a weighted on-base percentage that is 0.12667% to 0.78724% higher than players of other age groups.

* The significance of the average exit velocity estimate:

With other independent variables held constant, we have 95% confidence that, on average, for every 1 MPH increase in average exit velocity, the weighted on-base percentage will increase by 0.32107% to 0.55389%.

* The significance of the average exit velocity estimate:

With other independent variables held constant, we have 95% confidence that, on average, for every 1 MPH increase in average exit velocity, the weighted on-base percentage will increase by 0.32107% to 0.55389%.

* Significance of the Sweet Spot Percentage:

With other variables held constant and the barrel rate at 0, we have 95% confidence that, on average, for every 1% increase in barrel rate, the weighted on-base percentage (wOBA) will increase by -0.07340% to 0.13700%. Since the p-value of the test result is > 0.05, this indicates that this variable has a minimal impact on wOBA.

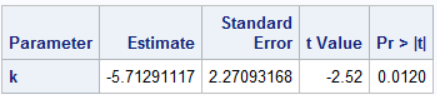
* Significance of the Barrel Rate:

With other variables held constant and both the average launch angle and sweet spot percentage at 0, we have 95% confidence that, on average, for every 1% increase in barrel rate, the weighted on-base percentage will decrease by 0.10509% to 0.82044%.

* Interaction between Average Launch Angle and Barrel Rate:

With other variables held constant, on average, for every 1° increase in average launch angle, the weighted on-base percentage will increase by

𝛽3 + 𝛾1 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒 = (0.07164 − 0.01437 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒)%. The marginal effect of the average launch angle changes with the value of the barrel rate.



Since the barrel rate for all hitters in the dataset falls between 0% and 25.7%, we set the barrel rate at 12.5% and substitute it into 𝛽3 + 𝛾1 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒. We have 95% confidence that, on average, for every 1° increase in average launch angle, the weighted on-base percentage will decrease by 1.24545% to 10.1804%. With other variables held constant, on average, for every 1% increase in barrel rate, the weighted on-base percentage will increase by 𝛽5 + 𝛾1 𝑙𝑎𝑢𝑛𝑐ℎ\_𝑎𝑛𝑔𝑙𝑒\_𝑎𝑣𝑔 = (−0.46276 − 0.01437 𝑙𝑎𝑢𝑛𝑐ℎ\_𝑎𝑛𝑔𝑙𝑒\_𝑎𝑣𝑔)% . The marginal effect of the barrel rate changes with the value of the average launch angle.

* Interaction between Sweet Spot Percentage and Barrel Rate:

With other variables held constant, on average, for every 1% increase in sweet spot percentage, the weighted on-base percentage will increase by 𝛽4 + 𝛾2 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒 = (0.03180 + 0.02586 𝑏𝑎𝑟𝑟𝑒𝑙\_𝑏𝑎𝑡𝑡𝑒𝑑\_𝑟𝑎𝑡𝑒)%. The marginal effect of the sweet spot percentage changes with the value of the barrel rate.

With other variables held constant, we have 95% confidence that, on average, for every 1% increase in barrel rate, the weighted on-base percentage will increase by

𝛽5 + 𝛾2 𝑠𝑤𝑒𝑒𝑡\_𝑠𝑝𝑜𝑡\_𝑝𝑒𝑟𝑐𝑒𝑛𝑡 = (−0.46276 + 0.02586 𝑠𝑤𝑒𝑒𝑡\_𝑠𝑝𝑜𝑡\_𝑝𝑒𝑟𝑐𝑒𝑛𝑡)%. The marginal effect of the barrel rate changes with the value of the sweet spot percentage.

Summary:

Based on the p-values from the model tests, the variables that significantly impact the weighted on-base percentage are average exit velocity, the interaction between sweet spot percentage and barrel rate, age, and the interaction between average launch angle and barrel rate. The possible reasons for these findings are speculated as follows:

* **Average Exit Velocity:** A higher exit velocity reduces the reaction time for defensive players, thus increasing the probability of successfully getting on base.
* **Interaction between Sweet Spot Percentage and Barrel Rate:** The positive interaction between sweet spot percentage and barrel rate could be due to both metrics incorporating launch angle into their definitions. Theoretically, players with higher barrel rates tend to have higher sweet spot percentages, hence the significant and positive interaction.
* **Age:** The age range of 25-29 is widely recognized as the peak performance period for MLB athletes, hence its significant impact on the weighted on-base percentage.
* **Interaction between Average Launch Angle and Barrel Rate:** The interaction between average launch angle and barrel rate is negative, suggesting that an excessively high launch angle does not constitute a barrel hit. Moreover, since the definition of a barrel includes launch angle, there is a significant negative interaction between these two variables.

From the p-values of the model tests, the variables that have a less significant impact on the weighted on-base percentage are average launch angle and sweet spot percentage. The possible reasons for these findings are speculated as follows:

* **Average Launch Angle:** The speculated reason is that a launch angle that is too high or too low is unlikely to result in a hit that gets the player on base. The launch angle needs to be within a specific range to easily produce a solid hit that results in getting on base, hence the average launch angle does not have a particularly significant impact on the weighted on-base percentage.
* **Sweet Spot Percentage:** Since the sweet spot percentage is an estimate of the quality of the hit, sometimes a solidly hit fly ball may be caught by an outfielder, resulting in an out, which leads to the sweet spot percentage having an insignificant impact on the average weighted on-base percentage.