Minor League Batting Success and wOBA

Michael Rabayda, Kyle Beach
Department of Computer Science
Haverford College
Ardmore, PA 19041
mrabayda@haverford.edu, kjbeach@haverford.edu

February 20, 2023

Abstract

We intend to model the wOBA of Minor League batters with given variables such as AVG, BB%, K%, SLG, OBP, ISO, BABIP, GB%, LD%, FB%, IFFB% and other statistics found on Fangraphs through a linear/polynomial regression model, using RMSE as our indicator for model accuracy. Since wOBA is largely used by baseball analysts as a better indicator of individual at-bat success, this analysis can be useful as a self-diagnostic tool for prospective batters to see how their own individual statistics compare to professional baseball players, and see which parts of their game they can work to adjust to produce the wOBA needed to take them to the next level.

1 Purpose and Hypothesis

Explanation for the purpose, application, and hypothesis for our study.

1.1 Purpose

The topic that we chose to research for this project was modeling of wOBA Minor League Baseball Players using relevant variables in regression models and checking our work using R^2 values. wOBA, which stands for "Weighted On-Base Average," measures how often a batter reaches base and accounts for how he reaches base, not just whether or not he does. This particular stat is more comprehensive than more well-known stats such as OBP, which only measures how often a batter reaches base. Our intention with this project was to provide batters with a self-diagnostic tool to evaluate themselves and their skillset relative to their peers as well as players in higher levels of the professional baseball system. This project will allow batters to evaluate their strengths and weaknesses, and see how much of an impact that trait has on their overall value as a player.

1.2 Hypothesis

The wOBA calculation values every method of reaching base differently, with more potential "damage" as an indicator of being more valuable. "Damage" in this context is defined as the potential for the batter to advance himself and other baserunners further along the bases. For example, home runs are weighted more heavily than doubles, and doubles are weighted more heavily than walks. Therefore, we hypothesize that extra base hit-oriented metrics such as SLG, ISO, and FB% will have a strong correlation with wOBA. We also hypothesize that BB%, AVG, BABIP, and OBP will correlate with wOBA as well since it values reaching base in any capacity, though they may not correlate as strongly as the first set of metrics. Finally, we hypothesize that metrics such as K%, GB%, and IFFB% will correlate negatively, since these events are negative outcomes for a batter.

We will be using various methods of feature selection to find appropriate variable combinations in order to model wOBA, since we do not want to have independent variables correlating with each other or having multicollinearity while running a multiple regression.

2 Approach and First Steps

We first visited fangraphs.com, a website which contains a plethora of stats for every player in MLB and MiLB, and downloaded a csv file containing all of the metrics that we decided were relevant.

Relevant Metrics:

- AVG- Batting Average. Measures how frequently a batter gets a hit.
- BB%- Walk Rate. Measures how frequently a batter draws a walk.
- K%- Strikeout Rate. Measures how frequently a batter strikes out.
- SLG- Slugging Percentage. Similar to AVG, but uses Total Bases instead of Hits when calculating, where a single is 1 total base, a double is 2, a triple is 3, and a home run is 4.
- OBP- On-Base Percentage. Combines Hits, Walks, and Hit-by-Pitches to generate the frequency at which a batter reaches base.
- ISO- Isolated Power. ISO is calculated simply by subtracting AVG from SLG. The objective is to see how often a player's hits are of the extra-base variety.
- BABIP- Batting Average on Balls in Play. The same as AVG, but BABIP excludes strikeouts and home runs, both outcomes which do not involve the defense.
- GB%- Ground Ball Rate.

- LD%- Line Drive Rate.
- FB%- Fly Ball Rate.
- IFFB%- Infield Fly Ball Rate. Same exact thing as Popup Rate.

The first issue that we encountered in the data was that several players had played at multiple levels of the Minor Leagues during the year, and the data didn't reflect how they performed under constant circumstances. We addressed this issue by cleaning the data so that any players who played in the DSL and/or the CPX as well as any A level would be removed from the study. A total of 36 entries were removed. We did not change anything for players who played at multiple levels of A since those levels all operate very similarly and an overall evaluation of those players was valuable, nor did we remove any players who only played in the DSL or the CPX. The cleaned dataset was stored in a new csv file called forPairPlot.csv.

Using this new csv file, we generated two heatmaps: one which showed numerical correlations among the variables and one which displayed those correlations visually through a single gradient color showing how much of an impact the metrics had on each other.

Figure 1: Numerical Heatmap

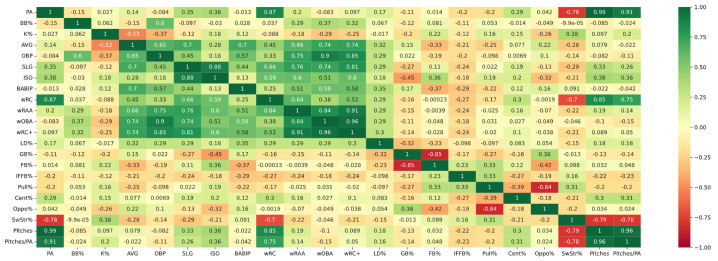
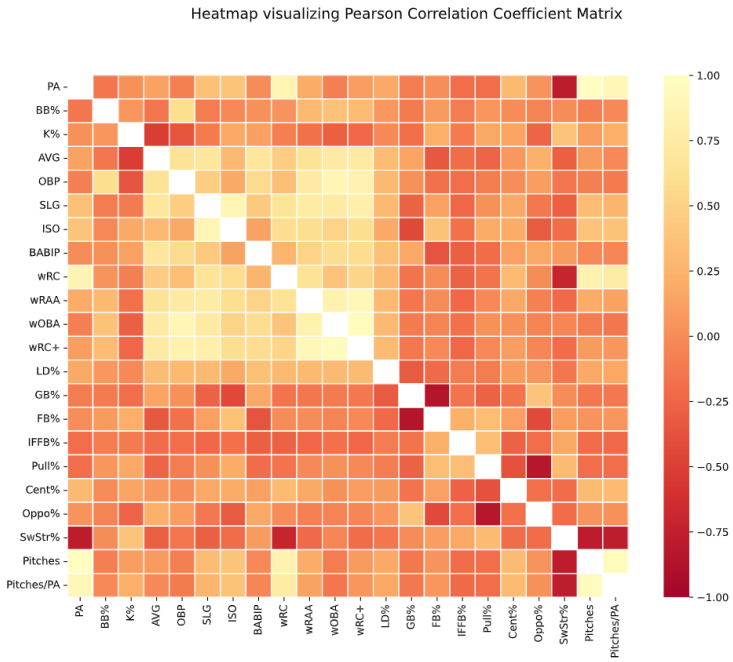


Figure 2: Visual Heatmap



Based on these figures, we conclude that metrics that positively correlate (0.7 or greater) include AVG, OBP, SLG, wRAA, and wRC+. We also conclude that metrics that negatively correlate (-0.1 or less) include K%, GB%, IFFB%, and Pitches/PA. Pitches also fall into that category, but we decided that Pitches/PA was more relevant than Pitches since Pitches/PA will allow us to see trends between itself and outcomes while Pitches will not.

3 Feature Selection Methods

3.1 Univariate

Next, we implemented the feature_selection function from sklearn, specifically SelectKBest and chi2. We used data from forPairPlot.csv to generate scatter plots of all the variables and set the hue to level of play (A, A+, AA, AAA, DST, CPX). These plots illustrate the univariate distribution for all the variables and plot the variables against each other.

3.2 Feature Selection

We then used recursive feature selection from scikit-learn and used F-score statistical analysis to find variables with high F-scores and low p-values in order

to see if any were useful for predicting the model. We wrote another script for feature selection, but this time the output included the F-score and p-value for every independent variable. We concluded that high F-scores and low p-values are the best predictors for wOBA.

3.3 Results

Figure 3: Recursive Feature Selection Output

```
Index(['LD%', 'GB%', 'FB%', 'Cent%', 'Oppo%'], dtype='object')
```

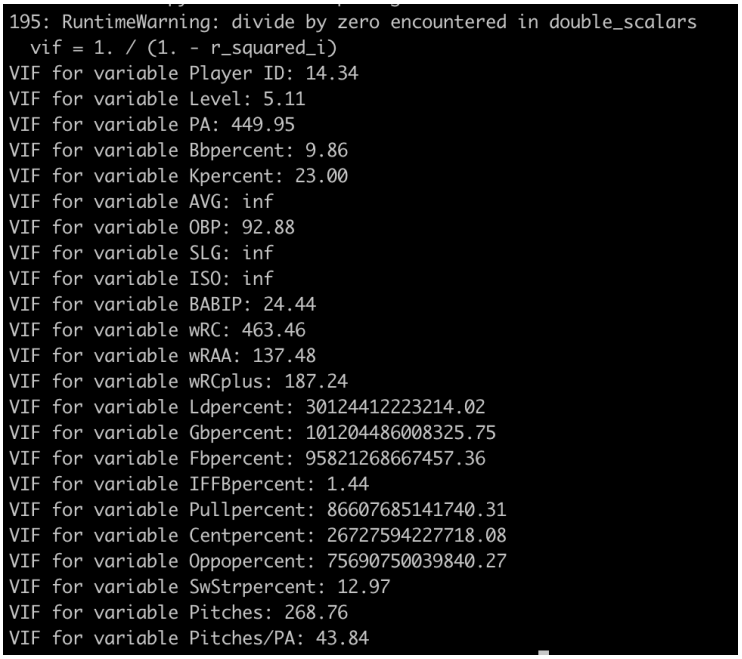
Figure 4: F-Score Statistical Analysis

```
Player ID: F-score=1.475, p-value=0.225
Level: F-score=10.378, p-value=0.001
PA: F-score=6.904, p-value=0.009
BB%: F-score=160.567, p-value=0.000
K%: F-score=92.498, p-value=0.000
AVG: F-score=1204.448, p-value=0.000
OBP: F-score=4041.567, p-value=0.000
SLG: F-score=1220.713, p-value=0.000
ISO: F-score=353.564, p-value=0.000
BABIP: F-score=528.385, p-value=0.000
wRC: F-score=163.599, p-value=0.000
wRAA: F-score=2426.363, p-value=0.000
wRC+: F-score=10468.699, p-value=0.000
LD%: F-score=94.068, p-value=0.000
GB%: F-score=13.087, p-value=0.000
FB%: F-score=2.311, p-value=0.129
IFFB%: F-score=34.657, p-value=0.000
Pull%: F-score=0.943, p-value=0.332
Cent%: F-score=0.738, p-value=0.391
Oppo%: F-score=2.393, p-value=0.122
SwStr%: F-score=2.146, p-value=0.143
Pitches: F-score=10.333, p-value=0.001
Pitches/PA: F-score=23.949, p-value=0.000
```

We conclude from this data that the strongest predictors of wOBA are AVG, OBP, SLG, BABIP, wRAA, and wRC+. This supports our hypothesis that slugging and getting on base were strong predictors of wOBA. Based on the F-scores, OBP is more significant than we thought, and SLG and ISO are not as important as we thought.

An issue that we ran into while running multiple regression was cases of high collinearity. In order to address this issue, we wrote a script to locate VIF values indicating high collinearity to decide what should and should not be used. After trying a variety of different combinations and discarding over half of the variables, we concluded that the best combination of variables to predict wOBA was OBP + BABIP + wRAA.

Figure 5: VIF Scores



These VIF, which stands for Variance Inflation Factor, values were important because variables like ISO are impacted greatly by others, specifically AVG and SLG. A batter's ISO may be high, but his AVG and SLG might be very low, so his ISO might look good, but he is not as valuable as an inference from his ISO might suggest. The reason ISO has such a high VIF value is it's dependence on other variables. We see values like OBP and wRAA with low VIF values, which is what we want in order to find the best independent predictors.

OBP, or On-Base Percentage, shows how often a batter reaches base. BABIP, or Batting Average on Balls in Play, shows how frequently a player's

batted balls result in hits, excluding home runs, since fielders are not involved. wRAA, or Weighted Runs Above Average, is a metric that attempts to quantify how many runs a player is contributing relative to an average player (wRAA for an average player is 0). wRAA is also a cumulative stat, so there is the potential to attain a higher wRAA by getting more opportunities. Therefore, in order for a player to maximize wOBA, they should focus on reaching base by any means necessary and maximizing the quality of his batted balls. Both of these focuses are very player specific, since every player is different and opponents have different strategies for different players.

Note: wOBA is the y axis on Figures 6, 7, and 8
Figure 6: OBP vs wOBA

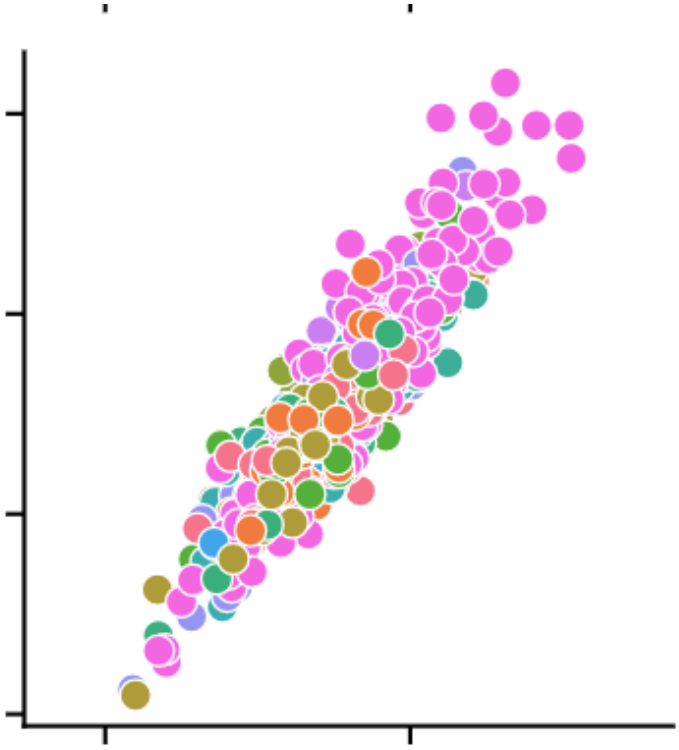


Figure 7: BABIP vs wOBA

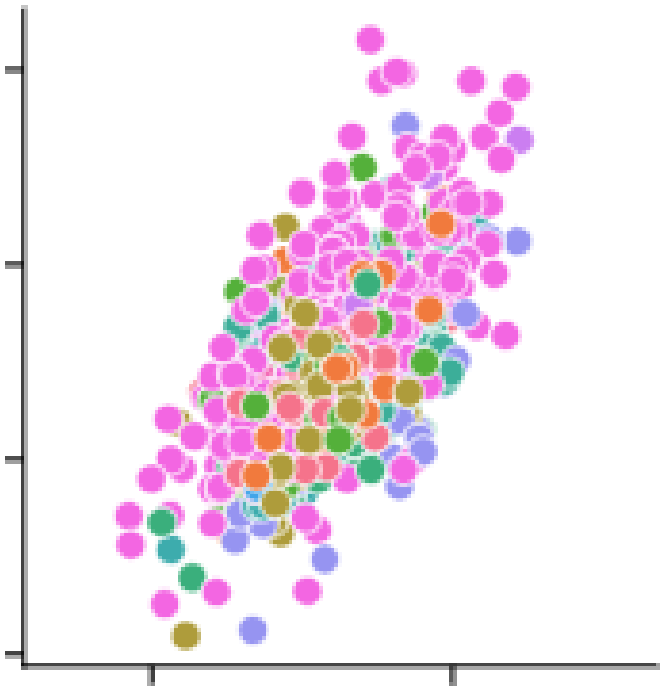


Figure 8: wRAA vs wOBA

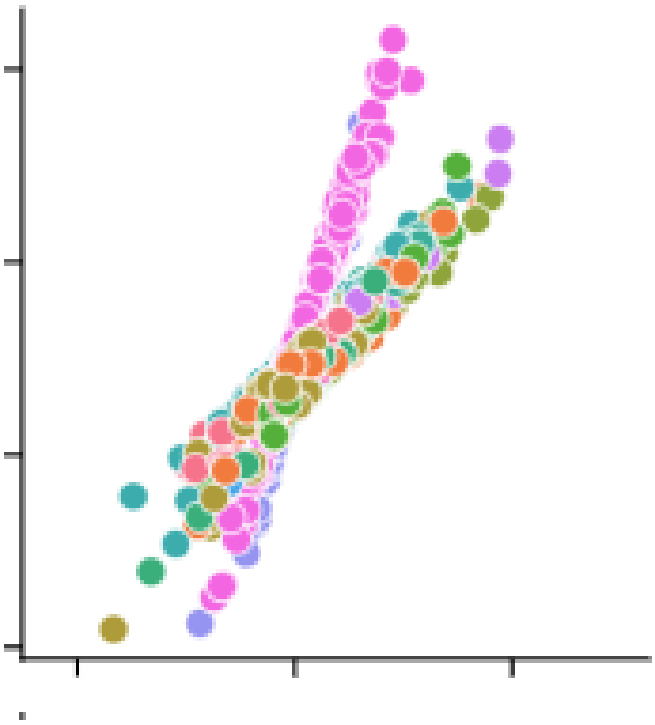


Figure 9: Key

4 Regressions

We used the player data as well as the formula "OBP + wRAA + BABIP" in order to create a multiple regression model for wOBA. We utilized ols from the statsmodel library in order to accomplish this. The regression results explained that OBP had a coefficient of 0.5901, BABIP had a coefficient of 0.0703, and wRAA had a coefficient of 0.0013. We used 80% of our data to train the model, and we used the remaining 20% to test against the model. R^2 is a metric to measure how well the model explains the variance, with values ranging from 0 to 1 and 1 meaning the model fully explains the variance. The R^2 value was 0.874, indicating that the model is a strong fit for the test data.

Figure 10: OLS Regression Results

OLS Regression Results

Dep. Variable:	wOBA	R-squared:	0.874
Model:	OLS	Adj. R-squared:	0.873
Method:	Least Squares	F-statistic:	2292.
Date:	Thu, 15 Dec 2022	Prob (F-statistic):	0.00
Time:	13:13:30	Log-Likelihood:	2804.0
No. Observations:	999	AIC:	-5600.
Df Residuals:	995	BIC:	-5580.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.1184	0.006	18.597	0.000	0.106	0.131
OBP	0.5901	0.018	32.071	0.000	0.554	0.626
BABIP	0.0703	0.014	5.080	0.000	0.043	0.097
wRAA	0.0013	6.12e-05	22.057	0.000	0.001	0.001

Omnibus:	135.665	Durbin-Watson:	1.828
Prob(Omnibus):	0.000	Jarque-Bera (JB):	283.334
Skew:	0.794	Prob(JB):	2.98e-62
Kurtosis:	5.069	Cond. No.	511.

Next, we used sklearn.linear_model from scikit-learn to do multiple regression and build an additional final model. The final model was saved to multiplereg.svg and the code outputs the R^2 value, the coefficients, and the intercepts.

Figure 11: Multiple Regression Model for wOBA

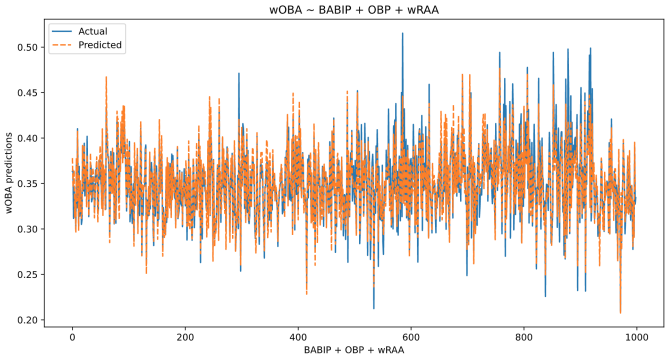


Figure 12: Coefficients, Intercepts, and R^2

```
Coefficients: [0.00134905 0.59007996 0.07029971]
Intercept: 0.11835938445969466
R2 Score: 0.8735844641988169
```

4.1 Conclusion

Our intention with this report was to provide a self-diagnostic tool for players seeking to move onto the next level. Based on our conclusion, players need to focus on three things: getting on base by any means necessary, maximizing the quality of their batted balls, and maximizing their opportunities to contribute. Players should evaluate their strengths and weaknesses in each category and adjust accordingly. Players who struggle to get on base should work on their ability to draw walks. Players who need to raise their level of batted ball should focus on hitting the ball with the sweet spot of their bat as often as possible and only swing at pitches they know they can hit well. Both of these can be accomplished by being more selective when batting and not swinging at every pitch. Players who need more opportunities need to evaluate why they aren't getting them and focus on improving whatever skill they lack. These lacking skills may not be related to batting. It could be poor defense, poor baserunning, or just a poor attitude.

In conclusion, we determined that the best indicators of wOBA are OBP, BABIP, and wRAA. Two of these metrics support our original hypothesis. We hypothesized that OBP and BABIP would have correlations to wOBA, though they were much higher than anticipated. We did not consider how wRAA, a cumulative stat, would impact wOBA when writing our original hypothesis. OBP, as illustrated in Figure 3, was highly correlated to wOBA, and players in AAA were most successful at obtaining high values for both numbers. AAA players also had the most success with BABIP, indicating that they had the most efficient batted balls of any level, as shown in Figure 4. Figure 5 is fascinating because of how differently the levels act on the graph. We interpreted the behavior to mean that despite having fewer opportunities relative to other levels, AAA players were able to attain higher wOBA in less time than players in lower levels. Overall, these observations make the skill gap between players at different levels of the Minor Leagues abundantly clear, especially comparing AAA players to any other level. In general, AAA players had higher OBP, higher BABIP, and higher wRAA relative to their number of opportunities. Given that they are at the highest level before the Major Leagues, it may not be surprising to see a trend like this, but it should be noted that they are facing the best competition as well. Overall, the best batters in the world are the best at reaching base, hitting the ball well consistently, and produce in the majority of their opportunities.

4.2 Replication

In order for one to replicate this type of analysis on similar statistics for a given sport/industry, we have generalized our methodology below to give an example of how one could accomplish this type of modeling.

- Identify a dependent variable from a dataset that you want to study using multiple regression, download datasets from web/collect organic data
- Clean your dataset, this will be an ongoing part of your analysis, but if you clean your dataset from the beginning to include which variables you would like to consider and make sure it is properly formatted to the various tests you will put them through (i.e making sure your variables are quantitative or categorical) it can save a lot of time
- Research your dataset and perform exploratory data analysis on univariate distributions and single variable regressions to find correlation statistics of each independent variable vs. the dependent variable. We used seaborn, matplotlib, and scikit-learn for this.
- Decide which feature selection method you believe to be appropriate for the independent variables you are looking at, as categorical and quantitative variables have different methods for identifying statistical significance in relation to a dependent variable. We used recursive feature selection in scikit-learn in addition to looking at statistical F-scores and p-values from the statsmodels library to get two different perspectives on our data. These methods do not guarantee final variables for the model you would like to produce, in our case we had the issue of high multicollinearity that could be seen in high condition value numbers when creating our first models in addition to low R^2 values that indicated low model fit with the test data. For high multicollinearity, we utilized VIF(variance inflation factor) values to spot any problematic independent variables using the stasmodels library. For low R^2 values, we used previous feature selection methods to help direct us to safer variable combinations.
- Create the model, we used both the ols from the statsmodels library and scikit-learn to produce multiple regression models of the same formula to ensure what the accurate model found could be replicated in multiple trials

4.3 Related Works

- Fangraphs.com: Used to find the necessary data
- Predicting-Run-Production-and-Run-Prevention-in-Baseball-The-Impact-of-Sabermetrics.pdf (researchgate.net): wOBA is the best predictor for runs scored by a team, wOBA is estimated to be able to explain 90% of the variability in runs scored by a team

- An Exploratory Study of Minor League Baseball Statistics (degruyter.com): Since hitters are less prone to injury than pitchers, it may be easier to predict how they will succeed at higher levels and therefore a safer investment
- How to Plot Multiple Linear Regression in Python - Javatpoint: Referenced to develop the different methods we used, especially PairPlot
- Feature Selection in Python - A Beginner's Reference - AskPython: Referenced to develop the different methods we used, especially Univariate Selection and Feature Selection
- Python Machine Learning Multiple Regression (w3schools.com): Referenced to develop multiple regression methods and to address a couple technical issues

GitHub Link Available Here