Predicting MLB Runs Scored for the 2023 Season using Various Linear Regression Techniques

Author: Sawyer Hunt
Publish Date: March 9, 2024

Introduction

In this analysis, different linear regression techniques were used to predict runs scored for MLB players during the 2023 season. The data used for this analysis are multiple offensive stats from the 2023 season sourced from the Baseball Savant website. The linear regression techniques that were used were linear regression, multiple linear regression, ridge regression, and lasso regression, all of which are outlined in the sections below.

Exploratory Data Analysis

This data set contains various offensive stats for 133 different players across MLB for the 2023 season. Initially, stolen bases were of interest to see how they related to runs scored. Exploratory data analysis (EDA) was performed on the stolen bases and runs scored fields of the data set. The average number of stolen bases was 11.75, however, the distribution is skewed right as few players had more than 20 stolen bases, see the histogram in Figure 1. In addition, the 75th percentile of stolen bases was 16 and the 90th percentile was 28, which is visualized by the boxplot in Figure 2. This leads to the assumption that the players with the most stolen bases tend to have more runs scored than players with the least amount of stolen bases. This generally seems to be the case when sorting the data by stolen bases and quickly comparing runs scored.

Figure 1: Histogram of stolen bases.

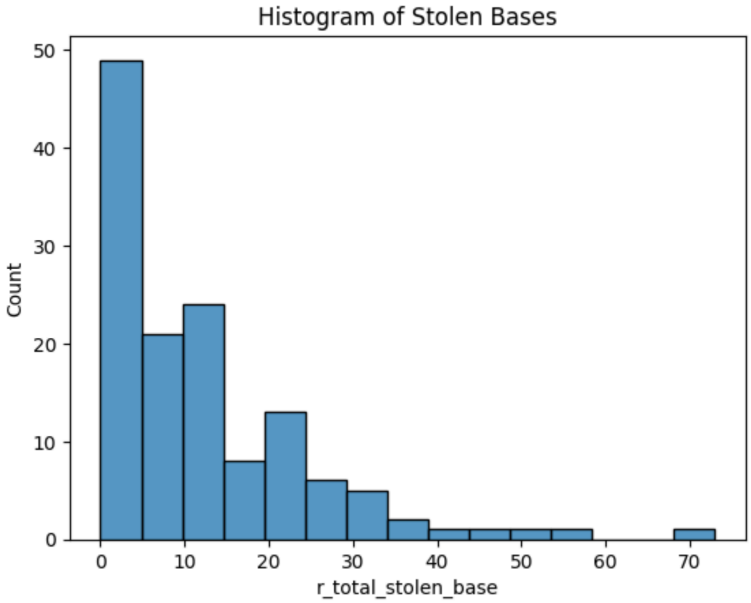
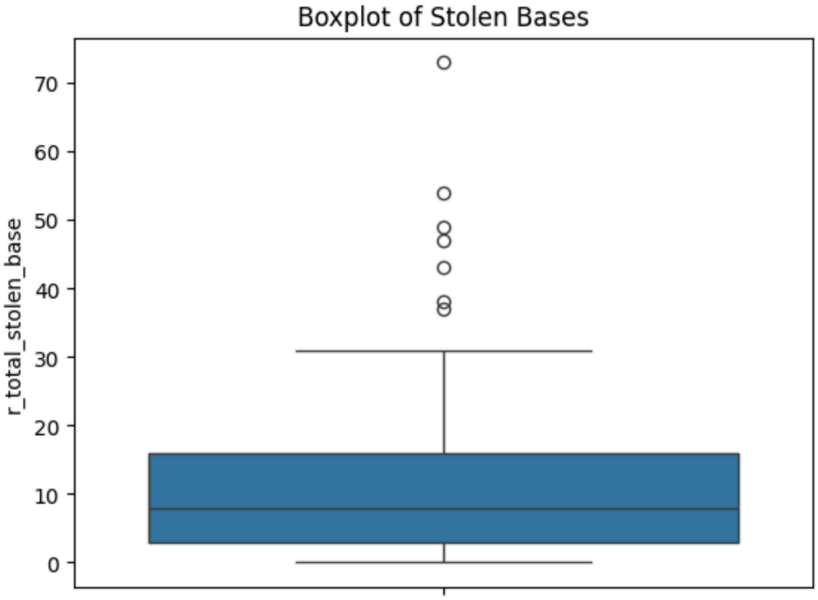


Figure 2: Boxplot of stolen bases.



The average runs scored was 79.56, with most players scoring between 60 and 80 runs which can be seen in the histogram in Figure 3. The upper and lower outlier limits were 126 and 30, respectively, with the 75th percentile at 90 runs and the 90th percentile at 102 runs, which is visualized by the boxplot in Figure 4.

Figure 3: Histogram of runs scored.

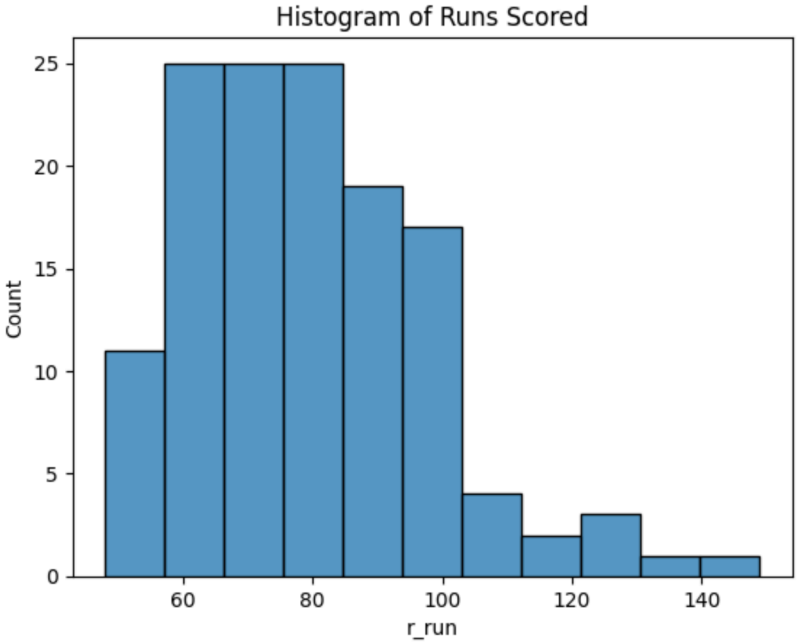
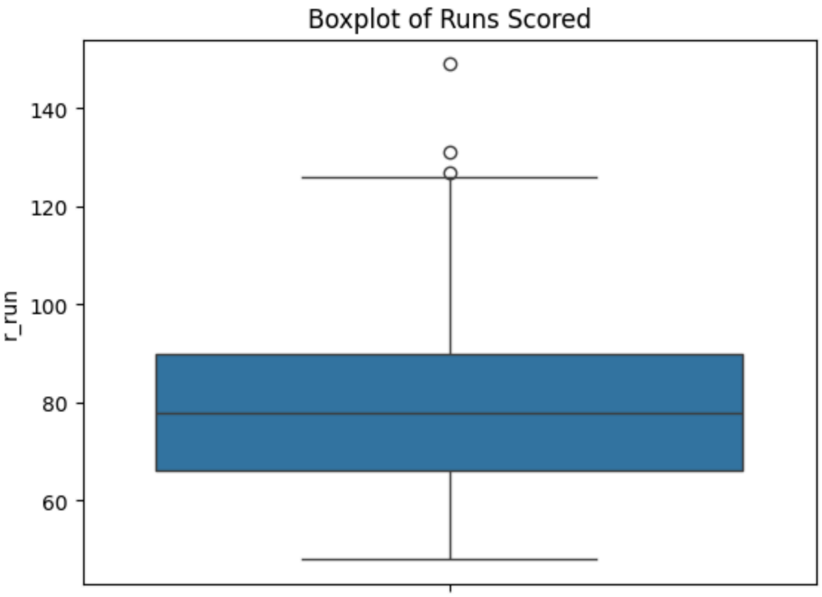


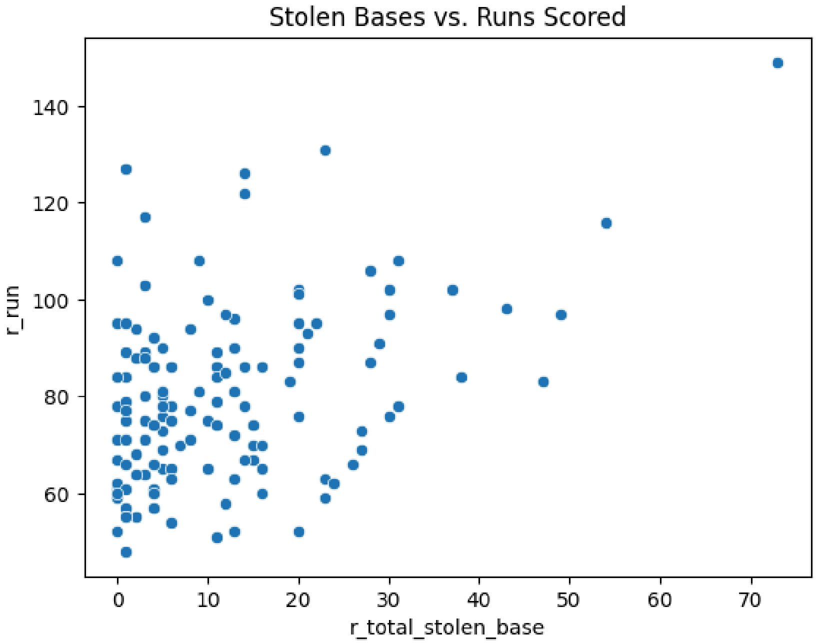
Figure 4: Boxplot of runs scored.



Linear Regression

Stolen bases were used as the independent variable in the linear regression model. It was originally thought that stolen bases and runs scored would have a strong relationship, however, the model suggests otherwise. Looking at a scatter plot of the two variables, depicted in Figure 5, it's apparent that there is not much of a linear relationship. In addition, the two variables only have a correlation value of 0.43. The model determined an intercept of 70.73 and an independent variable coefficient value of 0.71, meaning each stolen base led to 0.71 runs. The mean squared error (MSE) of the model was 418.31. Given that the range of target variable of runs scored was 101, the model did not perform well and more variables are needed to obtain accurate predictions.

Figure 5: Scatter plot of stolen bases vs. runs scored.



Multiple Linear Regression

The variables that were used in the multiple linear regression model were stolen bases, on base percentage (OBP), batting average, on base plus slugging percentage (OPS), hits, walks, and slugging percentage. This model performed much better than the single variable linear regression model with OBP, OPS, and slugging percentage having the largest coefficients of 50.53, 91.94, and 41.41, respectively, and produced an acceptable MSE of 79.42. Notice that the coefficients of OBP and slugging percentage equal that of OPS when summed together. This is because OPS is the sum of OBP and slugging percentage and all three are regarded as highly important stats when evaluating a players offensive metrics. The reason that the coefficients are so large, especially when compared to the other independent variables (the next highest coefficient value is 0.54 from hits), is because each of these stats are mostly less than one. If a player is performing well offensively, OPS can climb above one. So a one unit increase in these variables would lead to a large amount of runs scored.

Ridge Regression

The ridge regression analysis used the same variables as the multiple linear regression analysis. Initially, the ridge regression was performed untuned with an alpha value of one. The independent variables with the largest coefficient values were hits, OPS, and stolen bases, and produced an MSE of 57.69. Cross validation was then used to determine the optimal value of alpha, which was found to be 0.1, and the model was reanalyzed using this alpha value. Results of the tuned model returned a similar output as the untuned model. An MSE of 59.74 was obtained along with hits, OBP, and OPS as the independent variables with the largest coefficient values.

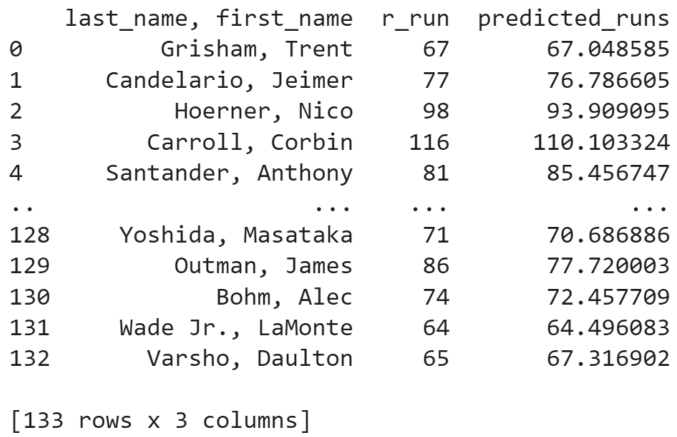
Lasso Regression

The same independent variables were used in the lasso model as in the ridge and multiple linear regression models. Cross validation was used to determine the optimal alpha value, which was found to be 0.001. The independent variables with the largest coefficients were OPS, OBP, and hits, and an acceptable MSE of 58.58 was obtained.

Predictions

Each model, except the single variable linear regression, performed well and produced similar results. The multiple linear regression, ridge regression, and lasso regression models were applied to the data frame. See figures 6-8 for predicted runs scored compared to actual runs scored for a handful of players.

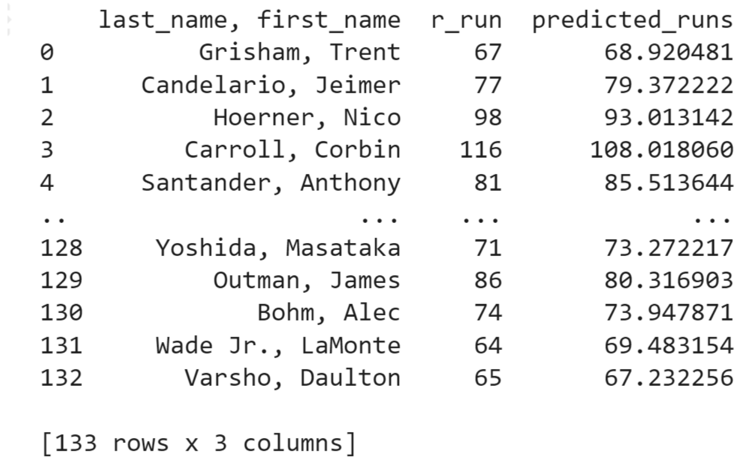
Figure 6: Predicted runs scored from the multiple linear regression model (predicted_runs) compared to actual runs scored (r_runs) during the 2023 season.



```
    last_name, first_name  r_run  predicted_runs
0          Grisham, Trent     67       67.048585
1      Candelario, Jeimer     77       76.786605
2            Hoerner, Nico     98       93.909095
3          Carroll, Corbin    116      110.103324
4      Santander, Anthony     81       85.456747
..                    ...    ...             ...
128     Yoshida, Masataka     71       70.686886
129          Outman, James     86       77.720003
130             Bohm, Alec     74       72.457709
131     Wade Jr., LaMonte     64       64.496083
132        Varsho, Daulton     65       67.316902

[133 rows x 3 columns]
```

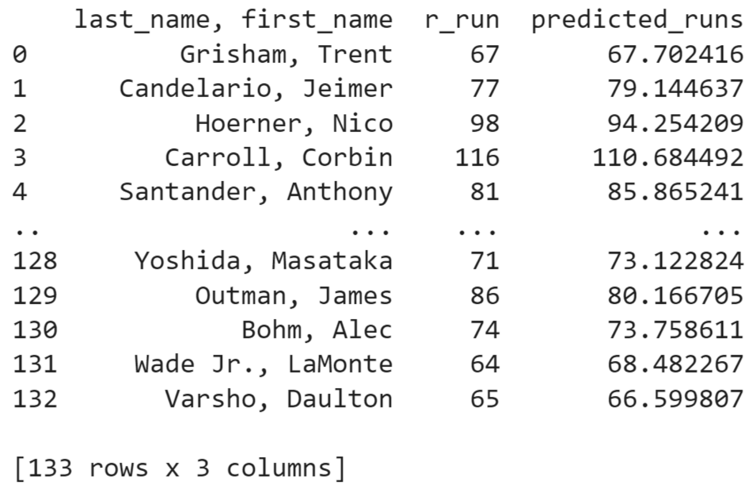
Figure 7: Predicted runs scored from the tuned ridge regression model (predicted_runs) compared to actual runs scored (r_runs) during the 2023 season.



```
     last_name, first_name  r_run  predicted_runs
0          Grisham, Trent     67       68.920481
1      Candelario, Jeimer     77       79.372222
2            Hoerner, Nico     98       93.013142
3          Carroll, Corbin    116      108.018060
4      Santander, Anthony     81       85.513644
..                    ...    ...             ...
128      Yoshida, Masataka     71       73.272217
129          Outman, James     86       80.316903
130             Bohm, Alec     74       73.947871
131      Wade Jr., LaMonte     64       69.483154
132        Varsho, Daulton     65       67.232256

[133 rows x 3 columns]
```

Figure 8: Predicted runs scored from the tuned lasso regression model (predicted_runs) compared to actual runs scored (r_runs) during the 2023 season.



```
     last_name, first_name  r_run  predicted_runs
0          Grisham, Trent     67       67.702416
1      Candelario, Jeimer     77       79.144637
2            Hoerner, Nico     98       94.254209
3          Carroll, Corbin    116      110.684492
4      Santander, Anthony     81       85.865241
..                    ...    ...             ...
128      Yoshida, Masataka     71       73.122824
129          Outman, James     86       80.166705
130             Bohm, Alec     74       73.758611
131      Wade Jr., LaMonte     64       68.482267
132        Varsho, Daulton     65       66.599807

[133 rows x 3 columns]
```

Conclusions

The single variable linear regression model using stolen bases as the independent variable did not produce acceptable results, so more variables were used in subsequent models to predict runs scored. The multiple linear regression, ridge regression, and lasso regression models all produced similar predictions. Ultimately, the lasso regression model is believed to be the best model for this particular analysis for different reasons. The first being that lasso regression penalizes the coefficients whereas the multiple linear regression model does not. Many of the offensive stats used in the models are not weighed equally (i.e. percentages vs. counting stats vs. averages); therefore, it is wise to use a modeling technique that penalizes the coefficients to prevent overfitting. The next reason is that this is a fairly sparse model with only seven features used to estimate the predictions. A ridge model is best suited for many correlated predictors to avoid multicollinearity.

A model such as this one could be used as proof of concept to showcase accurate predictions. More data could be added to the model to make predictions for future seasons.