INFO 4604-5604

Applied Machine Learning

Spring 2023

Project Report

Student Name	Alex Ojemann
Student ID	109722375
Project Title	ERA Estimator Regression and Stepwise Feature Selection
Date Submitted	5/8/2023

Table of Contents

1. Overall context for your project

For over 100 years after the inception of professional baseball in America, pitchers have been evaluated based on whether their team won the game and how many runs the team gave up, both of which are extremely dependent on the rest of the team. That finally changed in the late 1990s when Voros McCracken's research uncovered that a pitcher's batting average on balls in play had no consistency from year to year. This implied that a pitcher had little control over whether a ball in play was a hit or an out. The only pitching statistics that were correlated from year to year were home runs allowed, walks and strikeouts, which became known as the three true outcomes. That gave rise to Fielding Independent Pitching, or FIP, which is a pitching statistic designed to be on the same scale as Earned Run Average, or ERA, but only takes into account the three true outcomes. Over time, more advanced pitching statistics evolved that used the three true outcomes to try to evaluate a pitcher's performance. One of these is called SIERA, which accounts for more complex trends involving the three true outcomes and incorporates whether a batted ball was a grounder, popup or fly ball as well. One example of this is that it includes a walks squared term because pitchers that have high walk rates increasing their walk rate slightly means more runs will score than pitchers with low walk rates increasing their walk rate slightly because the pitchers with high walk rates are more likely to already have runners on base. My project builds on SIERA in a number of ways but will make some changes to try to improve its RMSE when estimating ERA.

2. Problem definition

Like the other metrics described above, my goal is to create a metric that is on the same scale as ERA but uses the three true outcomes. Similar to SIERA, I want to incorporate whether a batted ball was a grounder, popup or fly ball as well. Pitchers are thought to at least have some level of influence on this as opposed to whether a batted ball is a hit or an out. In addition, I want to incorporate exit velocity into my model. This is another metric for which it's questionable how much control a pitcher has in its outcome, but it would make sense that a pitcher has more control over it than whether a batted ball is a hit because it's directly representative of the quality of contact whereas a softly hit ball may drop for a hit while a hard line drive hit right at a fielder could be an out.

3. Project motivation – why should we care?

Research into improved player evaluation metrics is valuable because Major League Baseball set a record revenue in 2022 of nearly $11 billion so any incremental increase in ability to evaluate players better could be extremely valuable. For example, the Oakland Athletics of the early 2000s had a budget that was a fraction the size of their competitors but they still won over 100 games in multiple seasons due to their superior evaluations of players using data analytics.

In addition, value metrics for hitters are much better than those for pitchers. While WAR, the most popular baseball cumulative value statistic, is very in depth for hitters, for pitchers it simply uses a cumulative form of either RA/9, the number of runs a pitcher allows per nine innings, or FIP. A better metric of pitcher performance on a per nine innings basis could make WAR a better cumulative value statistic for pitchers.

4. Project methodology

My plan is to use linear regression to create an equation on the same scale as ERA that incorporates the outcomes of which pitcher is believed to have the most control. These features include the three true outcomes other than home run rate because many believe that the rate of home runs per fly ball is random, rate of ground balls, line drives and fly balls, combinations of two of these variables multiplied together as used in SIERA (i.e. walks squared), and average exit velocity, with ERA as the target variable.These features are inspired by SIERA in that SIERA used some products and squares of walk rate, strikeout rate, and the rates of the three batted ball types but my method is to let the model select which ones to use through backward stepwise feature selection. Average exit velocity was not considered for SIERA but it is an interesting variable because it considers batted balls but isn't as luck based as batting average on balls in play because it represents how hard the hitter hit the ball.

5. Data source

MLB's Baseball Savant is a web-based tool that provides a wealth of baseball data, including detailed statistics on every pitch thrown in MLB games. It allows users to access data on individual players, teams, games, and seasons, and provides customizable filters to help users analyze the data in a variety of ways. I used Baseball Savant to get a csv file with a pitcher's batters faced, walk rate, strikeout rate, ground balls outs, ground ball hits, fly ball outs, fly ball hits, line drive outs, line drive hits, and average exit velocity for each pitcher in each season from 2015-2022 because 2015 was the first season where statcast data was collected. Further mutations were performed to get columns for ground ball rate, line drive rate, and fly ball rate and to get the squares and products of each of the columns other than exit velocity.

Analysis of the data showed that there were some pitchers who faced very few batters in a given season so the data was filtered to only contain pitchers who faced more than 200 batters. This is roughly equivalent to throwing 40 innings as was the requirement for the development of SIERA.

The data was split into training and test data. The test data will be used to evaluate the RMSE and adjusted R squared of the model to see if the model is overfitting the training data.

6. ML model design

The model I will use is linear regression with backward stepwise feature selection. I want to exclusively use a linear regression model because my desired outcome is an equation that can be easily interpreted to roughly equate to ERA using the features in our data set rather than just achieving the highest possible predictive accuracy. Backward stepwise feature selection is a technique used in statistical modeling to select a subset of predictor variables that are most important in predicting the response variable. It works by starting with a model that includes all of the available predictor variables and then iteratively removing the least significant variable until the desired level of model parsimony is reached. For linear regression it removes the variable with the highest resulting adjusted R squared value until adjusted R squared can no longer be reduced by removing a variable.

The following is the final model once backward stepwise feature selection was finished.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.64131    1.21558  16.158   <2e-16 ***
FB_pct       -26.33530    2.33519 -11.278   <2e-16 ***
GB_pct       -17.32146    1.39233 -12.441   <2e-16 ***
p_k_percent   -0.23788    0.01389 -17.122   <2e-16 ***
p_bb_percent  -0.04385    0.02205  -1.989   0.0476 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

7. Key insights/findings and ML model

The model is ultimately very similar to FIP but with ground ball rate and fly ball rate used instead of home run rate. This is in accordance with the belief that home runs are mostly the product of how often the pitcher allows the ball to be hit in the air and how far it goes is relatively random. The model ultimately did not find any of the square or product features or exit velocity to be useful in terms of reducing adjusted R squared.

The final adjusted R squared of the model on the training set was 0.549. The adjusted R squared on the test set was 0.47 so there may be some overfitting but not enough to be concerned. The RMSE on the test set was 0.571.

8. Potential real-world applications of project

Potentially, further research could go into understanding why certain predictors are deemed useful or unuseful by our feature selection method. Additionally, other models could be used to estimate ERA that may not output an interpretable equation but achieve a superior accuracy.