Statcast Analysis - MA678 Midterm Project Report

Andrew Sisitzky

10 December 2021

ABSTRACT

In 2015, Major League Baseball (MLB) implemented the Statcast system with the purpose of measuring previously unquantifiable aspects of the game. A significant and unprecedented aspect of Statcast is its ability to measure statistics related to the quality of contact a batter makes with the baseball. In recent seasons, however, reports have surfaced about the MLB making changes to the composition of the baseballs used in their games. With the shifting weight, size, and density of the baseball over the previous six years of competetion, it would be helpful to account for these year to year changes when using quality of contact Statcast stats to predict traditional baseball statistics. In this report, I will use a multilevel model to predict BABIP (batting average on balls in play) using Statcast data from 2015 to 2021. To address the issue of the changes in the ball, I will be grouping by year and using varying slopes and varying intercepts for all statistics that could be affected by these changes.

INTRODUCTION

BABIP, or batting average on balls in play, is a statistic in baseball which measures a player's batting average exclusively on balls hit into the field of play. Historically, many have considered luck to play a large role in this stat; once the ball leaves the player's bat, numerous factors which the batter has no control over come into play, such as defensive positioning and the range/ability of the opposing fielders. Both of these factors, among others, can inflate or deflate a batter's BABIP. However, it has long been asserted that the quality of contact by the batter and their running ability could be effective in predicting this stat, although prior to the 2015 season, there was no large scale effort to quantify and/or track these statistics. With the implementation of the Statcast system, such prediction became possible.

METHOD

Data Processing

The data that I used for this project was gathered from Baseball Savant's custom leaderboards. I made the decision to not include the data from the shortened 2020 season as the distribution of some statistics during this season appeared to differ greatly from the others I observed. Therefore, the final dataset contained 835 observations of 38 variables, 4 of which would remain in my final model. The variables included in the BABIP model are listed in the table below.

Statistic	Description
exit_velocity_avg	How fast, in miles per hour, a ball was hit by a batter.
launch_angle_avg	How high/low, in degrees, a ball was hit by a batter.
sweet_spot_percent	How often a player produces a batted-ball event in the launch angle sweet-spot zone of 8-32 degrees.
sprint_speed	A measurement of a player's top running speed, expressed in "feet per second in a player's fastest one-second window."

Before getting further into the model itself, I will take a step back to show my exploratory data analysis and highlight how each variable and the model itself was chosen.

Exploratory Data Analysis

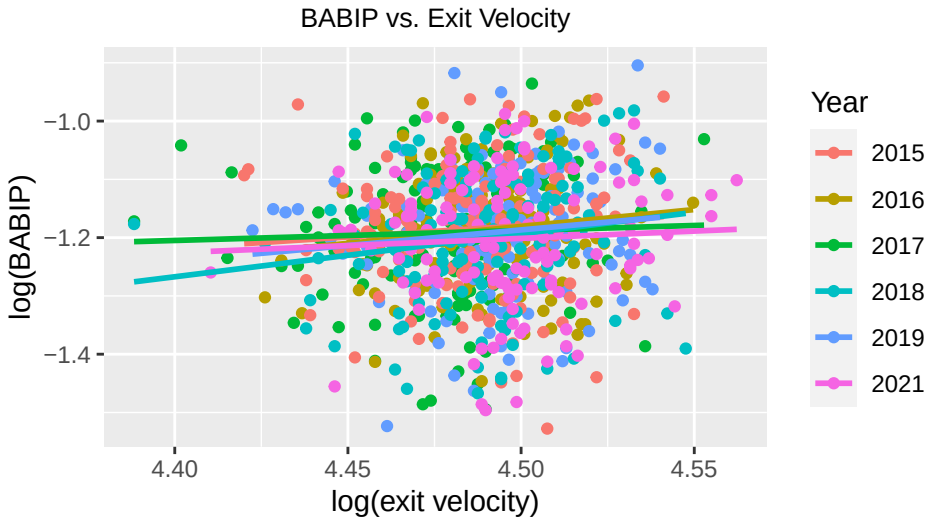

Figure 1