Regression Analysis Project

Justin Reising

December, 7 2018

1 Introduction

At the time beginning the composition of this document, the Boston Red Sox and Los Angeles Dogders are in Game 2 of 2018 World Series with a full count, 2 runners on base, and 2 outs in the bottom of the fifth inning at 10:07pm EST on October 24, 2018 in Boston with a temperature of 42 degrees Fahrenheit as Boston tries to take back the lead with the score 2-1. What is remarkable about the game in the age of "Big Data Analytics" is that the details in the opening statement are now variables to consider as a manager of a team for decision making in game situations. Major League Baseball uses state of the art tracking systems installed in every stadium called Statcast and PITCHf/x. Statcast and PITCHf/x are powered by cloud computing services provided by Amazon Web Services (AWS). The tracking technology records the incredibly precise data such as spin rate, velocity, spin direction, horizontal and vertical break, and location relative to the strike zone of a single pitch to name a few. Consequently, every game thousands of different data points recorded at every instance of game play. There are massive amounts data being analyzed simultaneously and analytics created to inform decisions from on the field, to the corporate office. This project will focus on a subset of the type of data gathered to build a multiple regression model to predict an outcome.

The data set provided is a subset collected from the PITCHf/x system and is comprised of all pitches thrown on Mondays during the 2016 MLB regular season, excluding intentional walks (79,931 observations and 35 variables). There are numerous categorical variables which are recorded for every pitch, such as batter name, pitcher name, umpire name, bats (L/R), hits (L/R), etc. For the purposes of this project, features will be selected to omit the consideration of "identifier" variables such as player names and focus primarily on pitch metrics. Although, we are equipped to look at specific match-ups between particular pitchers and hitters.

2 Potential Predictor Variables and Response Variable

Inning: (*Integer*) The inning number.

Balls: (*Integer*) The number of balls (pitches outside the strike zone) that the pitcher has thrown in the at-bat, calculated before the pitch is thrown.

Strikes: (*Integer*) The number of strikes that the batter has obtained in the at-bat, calculated before the pitch is thrown.

ProbCalledStrike: (*Float*) Estimated probability that the umpire will call the pitch a strike, if the batter does not swing, based on TruMedia's model.

ReleaseVelocity: (*Float*) Pitch velocity (mph).

SpinRate: (*Float*) Pitch spin rate (rpm).

SpinDir: (*Float*) From the catcher's perspective, the angle (from 0 to 360) between the pole around which the ball is rotating and the positive x-axis.

HorizLocation: (*Float*) Distance in feet from the horizontal center of the plate as the ball crosses the front plane of the plate, negative values are inside to right handed batters.

VertLocation: (*Float*) Height in feet above the ground as the ball crosses the front plane of the plate.

HorizMovement: (*Float*) The horizontal movement, in inches, of the pitch between the release point and home plate, as compared to a theoretical pitch thrown at the same speed with no spin-induced movement.

VertMovement: (*Float*) The vertical movement, in inches, of the pitch between the release point and home plate, as compared to a theoretical pitch thrown at the same speed with no spin-induced movement.

3 Research Question

One of the most prolific attributes of the pitcher-hitter dialogue is the exploitation of weaknesses between the two. The best hitters are ones that are best at timing pitches and the best pitchers are those that are best at upsetting timing of hitters. Baseball is a unique game that combines the reliance of team members to play with synergy around frequent one-on-one match-ups between hitters and pitchers where strategy is hidden from plain sight from people new to the game. On top of the on-field chess match that occurs between every pitch, the speeds and movement of pitches thrown can be missed by a blink of an eye. Physically speaking, the ability for a hitter to track, decide whether or not to swing, and if so, actually hit a 3" diameter ball traveling 95 MPH from 60 feet away, while swinging a 34" bat weighing approximately 31 ounces with a 2.5" barrel, and strike the ball within 1/4" of the center of mass seems impossible. With all of the physics stacked against the hitters, they can use all of the help they can get. In addition, baseball players are very habitual in their actions. When something is working for them, they tend to not deviate from that behavior. During the course of this project I will investigate the variation in pitch speeds from pitchers to try to build a model to predict the release velocity of a pitch given certain metrics measured of the pitch from the Statcast System in place at major league stadiums.

4 Background and Motivation

The evolution of the game of baseball has experienced substantial changes in the past decade since the beginning stages of the implementation of the Statcast system and other technologies in Major League Baseball. With this data being collected, we are able to try to make predictions to inform pitchers about their pitching tendencies and pitch characteristics from the raw data collected of their pitches. We can also compare pitchers' pitches to other pitchers in the league as well as learn how to advise pitchers to try new grips for pitches to achieve greater spin rates or breaking action on curve balls, sliders, etc. This pitch data being collected is currently being analyzed by some of the best statisticians and physicist working for Major League teams to help them understand this data in a more comprehensible way. For this reason, the frame in which this project will focus on is checking our predictive ability of release velocity. One of the main reasons is that the ability to predict release velocity can help determine the pitchers' ability to change speeds. In baseball, even slight changes in pitch speed can be the difference between a key double-play in a one-run game with bases loaded in the bottom of the 9th or a walk off home run.

5 Checking Model Assumptions & Model Selection

In this section, we will check our model assumptions. After further investigation of the linear model, I found that the variable spin direction had a curvilinear relationship with release velocity. This is because of the way spin direction is defined and measured. The data showed pitches with similar spin direction could either be classified as curve balls or fastballs which could add unnecessary noise to our model. After various transformation attempts using the Box Cox Transformation, the curvilinear shape of the plot between release velocity and spin direction could not be corrected with current techniques I have been exposed to. Additionally, this unique occurrence caused the concern for multicolinearity in our predictors. In fact, removing the spin direction predictor from the model had no noticeable effect on the parameter estimates. My final reduced model is as follows:

$$\hat{Y} = 0.07709 + 0.0091X_1 + 0.0028X_2 - 0.1114X_3 + 0.4361X_4 - 0.1343X_5 + 0.7661X_6$$

Note that in the R output in Table 3, we can see that the each of the predictor variables have significance and that approximately 55.47% of the variation in release velocity can be explained by the linear relationship between release velocity and probability of called strike, spin rate, horizontal location, vertical location, horizontal movement, and vertical movement of the pitch. The estimated mean change in mph of release velocity is approximately .91 for each 0.1 probability of called strike. The estimated mean change in mph of release velocity is 0.0028 per rpm unit. The estimated mean change in mph of release velocity is approximately -0.1114 per foot of horizontal location. The estimated mean change in mph of release velocity is approximately 0.4361 per foot of vertical location. The estimated mean change in mph of release velocity is approximately -0.1343 per inch of horizontal movement. The estimated mean change in mph of release velocity is approximately 0.766 per inch of vertical movement. Note that these mean estimates are assuming all other variables held constant.

In the model selection process, our focus was primarily on pitch metrics as oppose to situational variables such as inning, balls, strikes, or outs. The scope of our model considered all pitches thrown on Mondays in the 2016 MLB season. After removing incomplete cases and anomalies from the data before selecting the model, I needed to check our assumptions. First, I checked Relationships between the predictor variables and the response variable. Given the correlation plot in Figure 1, I confirmed the results of having slight linear relations between Release Velocity with Probability Called Strike, Spin Rate, Horizontal and Vertical Location, and Horizontal Movement. Also, there was a slight positive relationship between Release Velocity and Vertical Movement. This is expected because pitches that drop vertically tend to have top spin and are

thrown at lower velocities described by the Magnus Effect.

Next, I needed to check for Multicolinearity in my predictor variables formally. The correlation plot in Figure 1 gave me an indication there was no concern for multicolinearity and checking the Variance Inflation Factors for each of the predictors as seen in Table 4 confirm this hypothesis with VIF values for each predictor approximately 1. With the number of observations being 73,723, the assumption of the errors terms having a standard normal distribution is satisfied by the Central Limit Theorem. In fact, the histogram of the residuals in Figure 2 verify this assumption. However, there are observation with significant deviations from the mean of zero on the low end. These types of pitches are referred to as "eephus" pitches, which are extremely slow pitches as to significantly upset the timing of a hitter. While they are very uncommon, it is important to note their existence in the game.

This leads to the final assumption to check, which is constant variance of the error terms. Figure 3 shows the plot of the residuals v.s the fitted values and can see these small number of residual outliers from such "eephus" pitches described above. Note that deviations tend to occur below the predicted value as physical limitations of release velocity should be considered. A formal test for non-constant variance by the Breusch-Pagan Test indicated that there was non-constant variance with a p-value of practically zero. In an attempt to correct this, I initialed a weighted least squares procedure, but to no avail. Table 4 shows the summary of the weighted least squares regression model which shows very little change in the parameter estimates. Next, I considered removing influential outliers to improve the model. In doing so, I found 4,319 outliers using the DFFITS criteria, in which the DFFITS value of $2(\sqrt{\frac{p}{n}}) \approx 0.02$. Table 5 shows the summary output of the regression model fit after the removal of influential outliers for this criteria and again, there is practically no change in the estimators. Because of the ratio outliers to the total number of observations in this data set, these influential outliers are not "pulling down" the regression line in any significant manner.

6 Model Validation and Prediction Accuracy

In this project, I am fitting the training model to the Monday pitch sample and testing the model on a Tuesday pitch sample of size 925 after removing incomplete cases. To measure the predictive ability of the model, I computed the Mean Square Prediction Error to compare with the Mean Square Error of the training model. Comparing the ratio of MSPR:MSE yielded approximately 0.9138, which is a good indicator of predictive ability without evidence of the model over fitting the training data. Figure 4 shows the plot of the predicted values and actual values for Release Velocity of the Tuesday data set. This plot shows that the predictive ability of velocities between the low to mid 90's are fairly reliable where as pitches below 90mph

are more scattered. I believe this could because of the variation of pitch speeds and spin rates of secondary pitches. Curve balls, sliders, and change-ups are pitches designed to upset the timing of the hitter where as fastballs are more consistent and easier to time as hitter. In fact, the accuracy of the model is approximately 96.74%. Further analysis can be done on individual pitchers with this same model if we wish to do so. In conclusion, variation of pitch release velocities play a key role in a pitcher's effectiveness. While this model is an adequate model for predicting release velocity based on the pitch metrics collected by Statcast, we can further investigate the anatomy of effective pitches. All statistical results in the form of tables and figures are found in the Appendix.

Appendix

Table 1: MLB 2016 Batted Balls on Mondays

inning	balls	strikes	probCalledStrike	releaseVelocity	spinRate
1	0	0	0.975	94.2	2044.22
1	0	1	0.745	97.1	1966.32
1	0	2	0.968	96.5	2127.17
1	0	0	1.000	95.6	1947.11
1	0	1	1.000	95.6	1903.08
1	0	2	0.321	98.3	2038.06

spinDir	locationHoriz	locationVert	movementHoriz	movementVert
205.477	-0.374	2.933	-6.93	8.28
220.143	0.342	3.223	-7.48	7.35
198.816	0.389	2.266	-5.22	9.79
198.734	-0.004	2.380	-7.24	8.40
205.503	0.272	2.421	-6.79	9.37
206.732	-0.206	1.430	-8.30	7.96

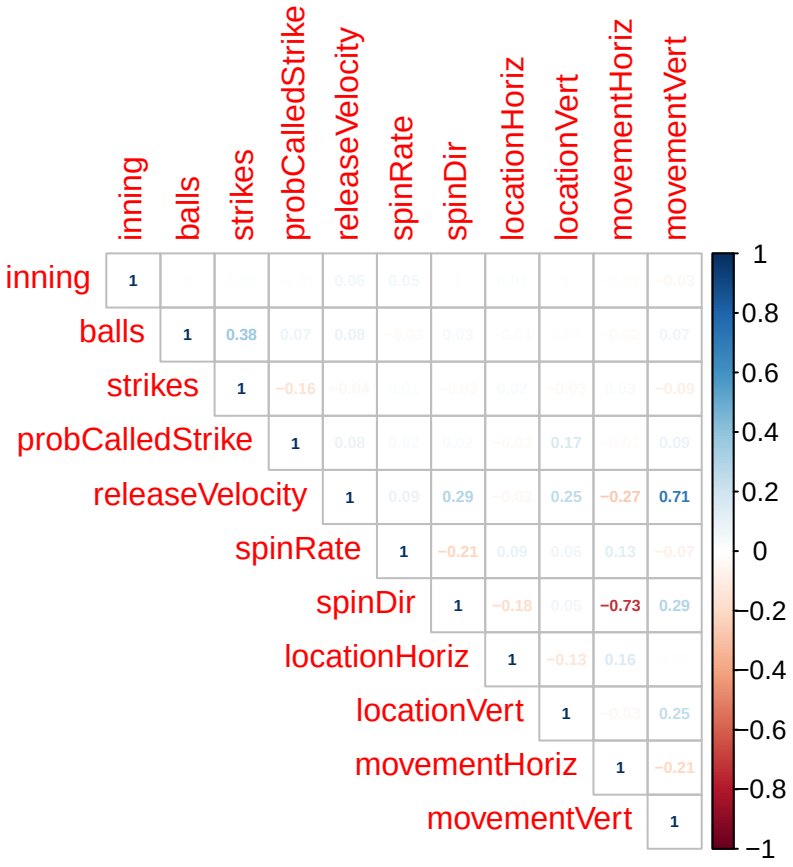

Figure 1: Correlation Plot

Table 3: Full Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	77.0882	0.1080	713.5684	0.0000
probCalledStrike	0.0911	0.0348	2.6174	0.0089
spinRate	0.0028	0.0000	60.9600	0.0000
locationHoriz	-0.1114	0.0173	-6.4202	0.0000
locationVert	0.4361	0.0167	26.1919	0.0000
movementHoriz	-0.1343	0.0024	-56.5908	0.0000
movementVert	0.7661	0.0029	260.1708	0.0000

Table 4: Variance Inflation Factors

	VIF
probCalledStrike	1.031896
spinRate	1.034268
locationHoriz	1.056402
locationVert	1.120689
movementHoriz	1.085177
movementVert	1.125403

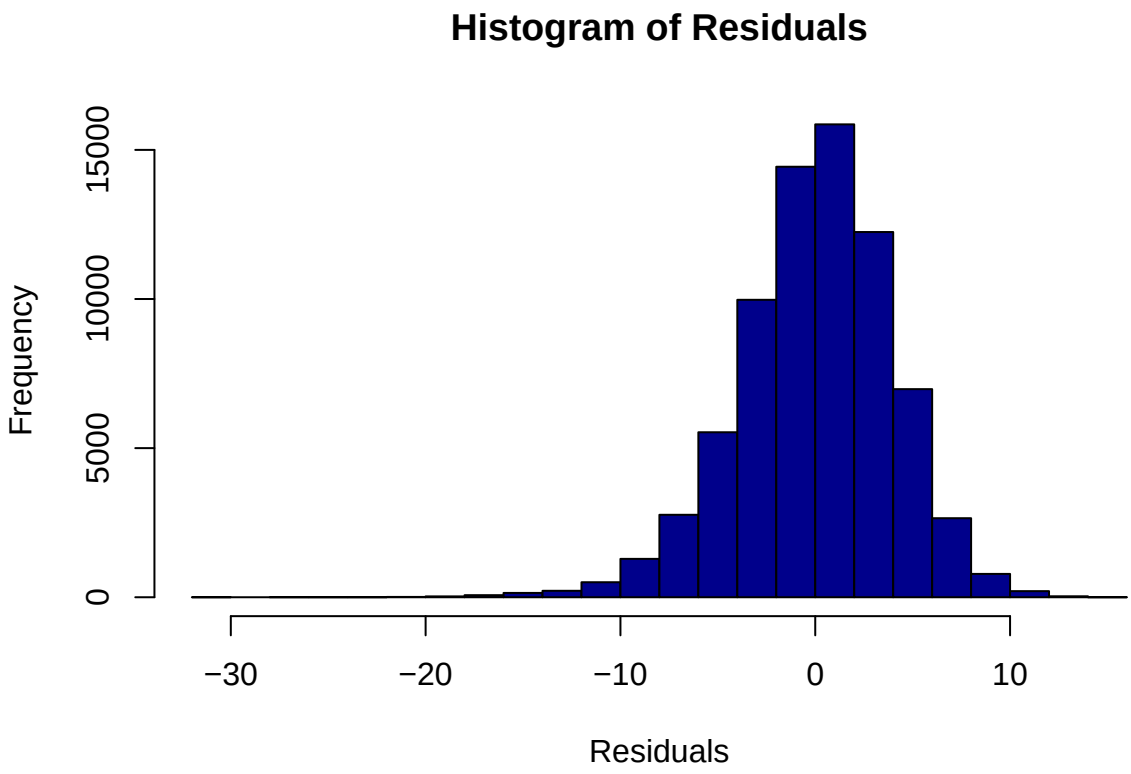

Figure 2: Residual Histogram

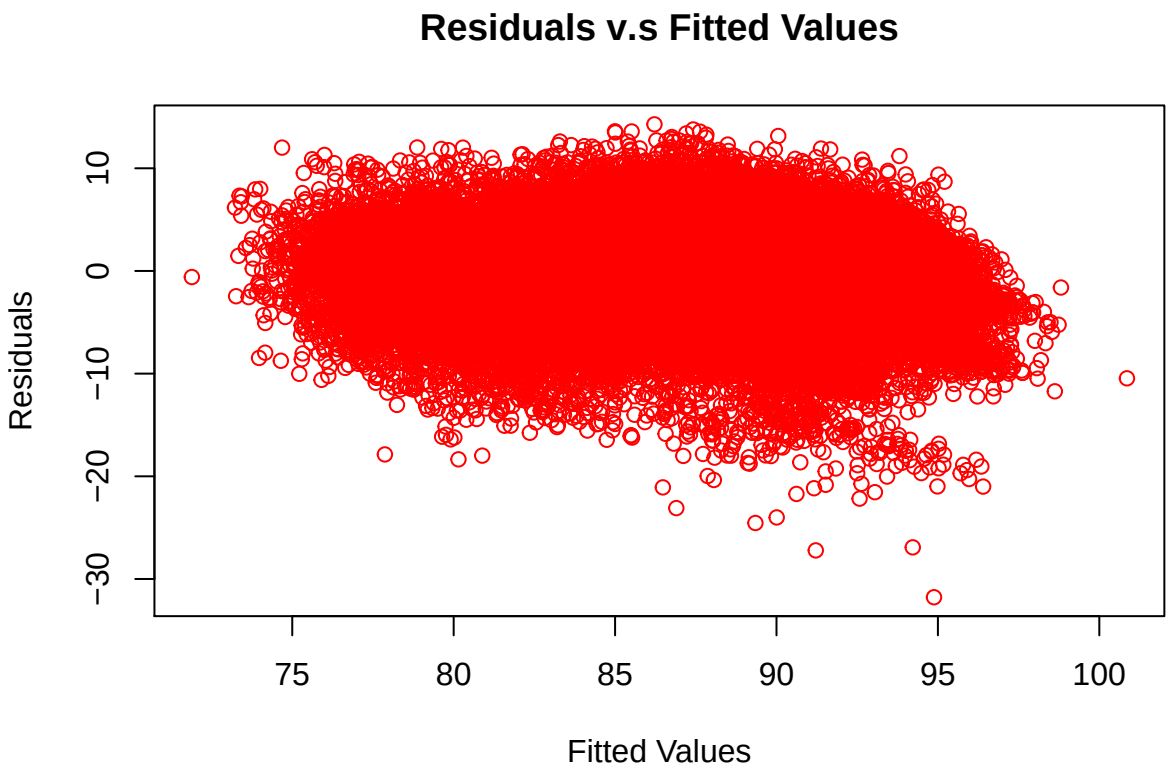

Figure 3: Residuals v.s Fitted Values Plot

Table 5: Weighted Least Squares Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	77.2707	0.1190	649.5547	0e+00
probCalledStrike	0.1177	0.0340	3.4642	5e-04
spinRate	0.0028	0.0001	54.7939	0e+00
locationHoriz	-0.1120	0.0170	-6.5874	0e+00
locationVert	0.4670	0.0164	28.4694	0e+00
movementHoriz	-0.1392	0.0024	-58.8123	0e+00
movementVert	0.7422	0.0030	244.4715	0e+00

Table 6: DFFITS Citerion Reduced Regression Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	77.0882	0.1080	713.5684	0.0000
probCalledStrike	0.0911	0.0348	2.6174	0.0089
spinRate	0.0028	0.0000	60.9600	0.0000
locationHoriz	-0.1114	0.0173	-6.4202	0.0000
locationVert	0.4361	0.0167	26.1919	0.0000
movementHoriz	-0.1343	0.0024	-56.5908	0.0000
movementVert	0.7661	0.0029	260.1708	0.0000

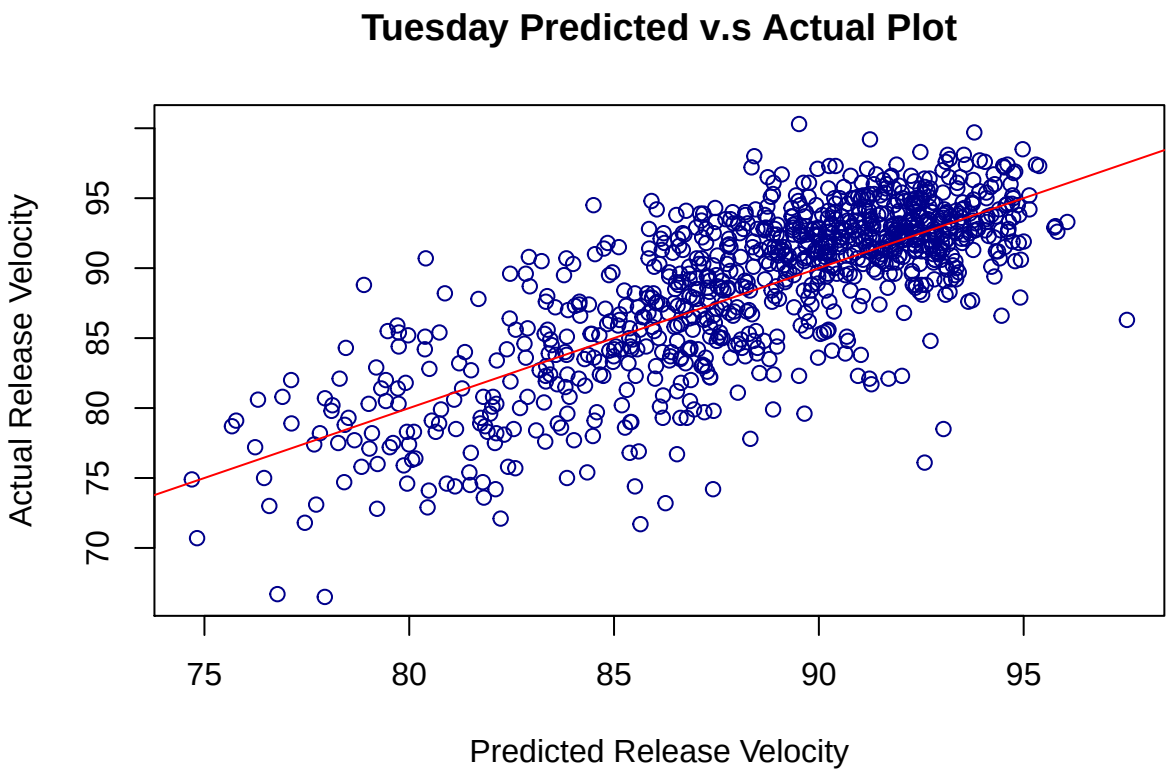

Figure 4: Tuesday Predicted v.s Actual Plot