Regression Final Project

Alan Flint and Rushil Sheth

October 2019

1 Data set Description

Our data is from MLB Pitch Data Kaggle.

The data consists of information on every pitch thrown in Major League Baseball from the 2015 to 2018 seasons. It consits of 2.87 million rows and 40 columns. The columns include information on the outcome of the pitch, pitch location (x and z location), pitch speed (start and end), type of pitch, and the game situation (balls, strikes, innings, runners on base, etc). The data description table is below.

Variable	**Description**
px	x-location of the pitch
pz	z-location of the pitch
start-speed	speed of the pitch as it's thrown
end-speed	speed of the pitch when it reaches the plate
spin-rate	pitch's spin rate in RPM
spin-dir	pitch spin direction in degrees
break_angle	angle of the pitch break
break_length	length of the pitch break in inches
break_y	break in the y direction
ax	n/a
ay	n/a
az	n/a
sz_bot	n/a
sz_top	n/a
type_confidence	confidence in pitch_type classification
vx0	n/a
vy0	n/a
vz0	n/a
x	n/a
x0	n/a
y0	n/a
z0	n/a

pfx_x	n/a
pfx_z	n/a
nasty	n/a
zone	n/a
code	result of the pitch
type	simplified code, S (strike) B (ball) X (in play)
pitch_type	type of pitch
event_num	identification number
b_score	batting team score
ab_id	ID number of ab-bat
b_count	balls in current count
s_count	strikes in current count
outs	number of outs
pitch_num	pitch number of at-bat
on_1b	True if runner on 1st base
on_2b	True if runner on 2nd base
on_3b	True if runner on 3rd base

2 Statement of Research Problem and Methods

Our goal was to predict the outcome of pitches that were swung at, either a swing and a miss, or the ball was put in play. This is a binary classification problem at its core.

The main method used in this report is logistic regression. For model selection, we used recursive feature elimination (RFE) to find the best variables for a given number of predictors, then chose a model based on AIC, BIC, and prediction accuracy in a testing data sample.

We used prediction accuracy as one of our criterion because the goal of our problem is to predict. AIC and BIC are typical criterion. We also took note of the number of predictors and checked for multicollinearity in the models returned by RFE.

3 Explanatory Analysis

We first reduced our data set to only the events we were interested in, swinging strike (swing and a miss) and ball in play. We created a binary boolean variable in-play(0 for swinging strike and 1 for in play) which is our dependent variable.

Next, we began our analysis by limiting our possible predictors to only relevant and interpretable variables. We removed variables that didn't have descriptions in the data table, identification number variables, and variables that don't influence if the pitch was hit or not. We were left with px, pz, start speed, end speed, spin rate, spin direction, break angle, break length, break y, and pitch type.

In this process, we noticed that variable 'pitch-type' was categorical with

15 levels. To simplify the model, we limited this to only two categories, four-seam fastball and two-seam fastball. This simplification also helps us avoid multicollinearity, as pitch-type = curve ball is most likely correlated to the variables about spin and break of the pitch.

We also examined the influence of the variable px, the x-location of the pitch, with px = 0 being a pitch right down the middle. This is obviously hard to interpret for left handed versus right handed batters, so we only included right handed at bats since those were the majority of the at bats. Additionally, we only looked at plays where the batter swung the bat–either the ball hit in play or missed.

In the end, our model data set was 201,076 rows with 10 predictor variables and no missing values. There were 58,737 swinging strikes and 142,339 in play balls, 70.8% and 29.2% of the entire data set respectively.

To check the relationship between pitch type (four seam / two seam fastball) and in play boolean, we cross tabulated the frequencies.

		Pitch Type	
		Four Seam	Two Seam
In-Play	1	10,857	47,880
	0	41,775	100,564

The Chi-Square test for independence has a p-value of 0.0, meaning that there is a strong relationship between the type of fastball and the in play boolean variable.

4 Model Selection

Once we had our 10 possible predictors, we used a model selection process to find the best model. Our method used recursive feature elimination to find the best subset of predictors for a given number of predictors. For example, if we specified that we wanted two predictors, the RFE would return the two best predictors for our model. We did this RFE process ten times, for desiring 1 through 10 predictors (where 10 predictors is the full model). This gave us 10 possible models to then choose between using other model selection methods like AIC, BIC, and percent correctly predicted using a training and testing data. The table below is the result of the RFE process, an 'x' denotes that variable was selected for the model given the subset size.

	Subset Size	1	2	3	4	5	6	7	8	9	10
Features											
px			x	x	x	x	x	x	x	x	x
pz		x	x	x	x	x	x	x	x	x	x
start_speed					x	x	x	x	x	x	x
end_speed						x	x	x	x	x	x
spin_rate											x
spin_dir										x	x
break_angle									x	x	x
break_length							x	x	x	x	x
break_y				x	x	x	x	x	x	x	x
FF								x	x	x	x

A quick observation: Pz (pitch location in the z-axis, or height) is the best predictor, followed by Px (pitch location in the x-axis, or width). This makes sense as MLB batters are much better at hitting a pitch that is waist height and center of the plate, than a pitch that is low and away.

To continue our model selection, we ran these models using a training and test sample from our data frame. We ran the model using the training data then used the test data set to compute percent correctly predicted using untrained data. We also compared AIC and BIC values of the individual models.

From this analysis, we determined that the model with six predictors: px, pz, start-speed, end-speed, break-length, and break-y was the best. This model has the highest percent correctly predicted at 74.5% and an AIC and BIC comparable to the full model (154,252 versus 153,616 for AIC, 154,321 versus 153,725 for BIC). We thought this 600 point difference in AIC/BIC is relatively small considering the values are in the 150,000, so the simpler model with slightly more predictive power is better.

The model output and summary is in the table below.

Variable	Coef.	Std. Error	z	P $>\|z\|$	[0.025	0.975]
const	137.6839	3.9211	35.1137	0.0000	129.9987	145.3691
px	-0.3645	0.0109	-33.4748	0.0000	-0.3859	-0.3432
pz	-0.8087	0.0108	-74.8285	0.0000	-0.8299	-0.7875
start_speed	-0.6169	0.0112	-55.0772	0.0000	-0.06389	-0.5950
end_speed	0.6069	0.0122	49.8740	0.0000	0.5831	0.6208
break_length	0.1315	0.0057	23.1301	0.0000	0.1203	0.1426
break_y	-5.447	0.1650	-33.0259	0.0000	-5.7710	-5.1244

5 Model Diagnosis

Before fitting the model we looked at the distribution of the all potential predictors for high leverage points. We did not identify any high leverage points.

Next we examined the predictors in relation to the other predictors in the model itself.

A correlation matrix of our selected predictors is presented below.

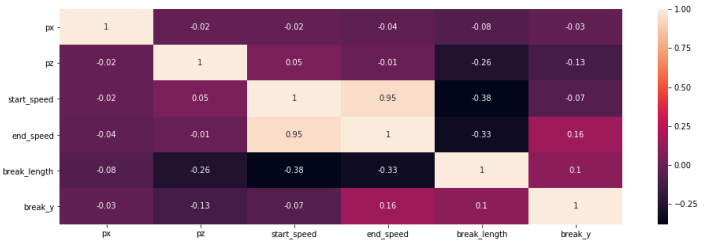

Figure 1: Correlation matrix of Best Model Predictors

From this, we can see that start and end speed are very highly correlated at 0.95. This makes sense as a pitch with a higher start speed will likely have a high end speed as well. The two variables essentially encode the same information.

Since the two predictors are highly correlated, it's likely that the model suffers from multicollinearity. To alleviate this issue, the easy fix is to remove one of these predictors. We decided to remove end speed.

We checked for influential points using Pearson's residuals. Of the 140,753 observations used in the training data set, only 2465 of them had a Pearson's residual statistic outside the ideal range [-3,3]. This is only 1.75%. If the fitted model is true, most of the residuals would fall inside this band, as they do in our case. Due to this and our large sample size, we determined that influential points were not an issue in our analysis.

6 Final Model of Choice and Interpretation

Our final model of choice is regression in-play on px, pz, start speed, break length and break y. The regression equation is $in.play = -6.98 - 0.39 * px - 0.86 * pz - 0.07 * start.speed + 0.13 * break.length + 0.68 * break.y$. The summary table is listed below.

Variable	Coef.	Std. Error	z	P >| z |	[0.025	0.975]
const	-6.9759	2.6109	-2.6719	0.0075	-12.0931	-1.8588
px	-0.3968	0.0108	-33.8199	0.0000	-0.4179	-0.3757
pz	-0.8693	0.0107	-81.4576	0.0000	-0.8902	-0.8484
start_speed	-0.0729	0.0024	-29.9095	0.0000	-0.0777	-0.0681
break_length	0.1376	0.0057	24.2719	0.0000	0.1265	0.1487
break_y	0.6872	0.1089	6.3104	0.0000	0.04737	0.9006

Interpretation of the Coefficients:

β_0: $e^{-6.9759} = 0.0009$ the odds of success when all other predictors are 0 is 0.0009
β_1: $e^{-0.3968} = 0.672$ when px increases by 1, the odds of hitting the ball in play decrease 32.7%
β_2: $e^{-0.8693} = 0.419$ when pz increases by 1, the odds of hitting the ball in play decrease 58.1%
β_3: $e^{-0.0729} = 0.9297$ when start speed increases by 1, the odds of hitting the ball in play decrease 7.0%
β_4: $e^{0.1376} = 1.1475$ when break length increases by 1, the odds of hitting the ball in play increase 14.75%
β_5: $e^{0.6872} = 1.9881$ when break y increases by 1, the odds of hitting the ball in play increase 98.81%

We can predict the outcome of a pitch using our regression equation. For example, if we use the mean of the independent variables:

px: -0.0328
pz: 2.5982
start-speed: 92.765
break-length: 4.5025
break-y: 23.7972,
the probability of the batting putting the ball in play is 72.8%.

7 Summary of Findings

Going into this research we expected the two seam or four seam to be one of the biggest indicators for putting a ball in play. Based on our final regression model we were at too high of a level with our initial thinking. There are more granular aspects of each pitch that ultimately determine if it is put in play when swung at.

Pitch location, break, and speed are the best indicators for this. For right handed batters you want to pitch as outside as possible. This makes sense, since batters will have much less power when they have to reach for a pitch and can not use their whole body power, and could possibly be off balance as well. Also throwing an above 90 fastball is pretty important as a pitcher if you want a swinging strike.

It seems like fastball swing and misses are pretty rare the last few years. If you want nasty pitches, where batters swing and come up empty handed, you should focus on pitch location, speed and break with the biggest emphasis on location.