STA440 Individual Project

Amber Potter

2022-12-17

Introduction

Background

Despite Major League Baseball being one of the largest and most popular professional sports leagues in America, recent criticisms have grown in regard to the pace of the game and its relative lack of action compared to other sports. In recent years, the game has seen the growth of the three outcome at-bat, meaning that many batters are trending towards either hitting homeruns, striking out, or walking. This comes from batters prioritizing launch angle and the greater addition to expected runs contributed from homeruns than batted balls in play. Although these extra homeruns are exciting for fans, the number of strikeouts increasing has the opposite effect. With long games and increasingly rare action from batted balls in play, the MLB has started looking into potential solutions to fix the pace of game dissatisfaction from fans.

Already there are rules being added for the 2022 season in attempts to address the decrease in action, such as the elimination of the infield shift which was responsible for taking away many base hits by left-handed batters. MLB has also implemented a pitch clock in attempts to speed up the game. Still, much of the discussion about 'fixing baseball' has been made around the strike zone. Automated ball and strike calls seem to be where the game is headed, but there are also talks of changing the strike zone all together.

Changes to the strike zone have been made before in the past, with the last official update to the strike zone being in 1996 when it was defined as the "the midpoint between a batter's shoulders and the top of the uniform pants – when the batter is in his stance and prepared to swing at a pitched ball – and a point just below the kneecap" (MLB.com). Additionally, and especially with this indistinct wording, the strike zone as it has been called by umpires has shifted without updates to the rules. For example, from 2007 to 2017 umpires have narrowed but lengthened their strikes zones (Edwards 2018). But what associations does the strike zone have on batted balls in play? Although shifting the strike zone up or down would be a slowly implemented change that would have to begin with lower levels and younger players so as to not cause injury, are either of these options reasonable for increasing action?

Research Question

In this analysis, I aim to investigate the association between high or low strikes and the odds that hitting said strike into play results in getting on base. Specifically, I hypothesize that hitting a low strike decreases the odds of a successful batted ball.

Data

The `baseballr` Package

This analysis uses 2021 Statcast data from Bill Petti's `baseballr` package. This data contains pitch-level information, including a separate observation for each pitch during the season, including pre-season, regular season, and post-season games. In 2021, alone, this includes 797,290 pitches, each with 93 recorded variables. Statcast technology was introduced in all 30 MLB stadiums in 2015, and has since recorded information about each pitch, the game situation of that pitch, and other player/inning related context. Because this analysis focuses only on balls batted into play, this original dataset is filtered down greatly.

First the daata was filtered to only include regular season and post-season games (excluding exhibition and Spring training games). This was done to eliminate practice games and other games that may not represent a true competitive game. Next, the data was filtered to only included batters with at least 50 at-bats in order to increase the consistency in the quality of at-bats. This was done by counting players with at least 50 at-bat ending events (Appendix A) recorded in the data, then filtering the data for only those players. By using this threshold I eliminate any players that were only called to the majors briefly or players that may not have been good/healthy enough to start frequently. Because I was only interested in comparing the odds of a successful batted ball given that the ball was put in play, I filtered for observations of type "X" which represent batted balls events. This also handled the issue of making sure I don't consider multiple strikes during an at-bat as it is inherently only possible for a batter to hit one pitch they face into play and it is not necessary to hit every strike faced. I also filtered for only high strikes, which were originally labeled to be in zone 1, 2, or 3, or low strikes, which were originally labeled to be in zone 7, 8, or 9. I did this because pitchers generally avoid trying to pitch down the middle as it is easier to hit, but also because I am strictly interested in the difference in the relationships between high and low strikes and the odds of a successful hit. In order to further focus on this goal, I also recoded the zone variable to refer only to the vertical location, with zones 1, 2, and 3 relabeled as high strikes and zones 7, 8, and 9 relabeled as low strikes. I also made the decision to filter for only plays where there was no infield shift. This choice was made with consideration for the fact that the infield shift will no longer be allowed in future seasons. Even though I am not attempting to predict for the future, my goal is still to observe the relationship between pitch location and the odds of a successful ball in play in a way that might be insightful to the possibility of future strikezone changes. With as much data as this still leaves me, I do not believe it will hinder my results and will better allow me to examine the relevant associations on against standard defenses. Lastly, I chose to filter out Eephus pitches (EP), knuckle balls (KN), and screwballs (KC) due to their rarity or inconsistent pitch movement. I also removed pitches labeled as "FA" or "CS" as their occurrance was rare and it is not clear what type of pitch they represent (Visualization and detailed list of pitch types can be found in the Appendix B). I also simplified pitch type to just fastballs (4-seam fastballs, splittlers/split fastballs, sinkers/2-seam fastballs, and cutters), changeups, and breaking balls (curveballs, knuckle curves, and sliders) as this is often how they are examined by MLB and I do not wish to oversaturate my model with different pitch types.

Key Variables

For this analysis, the outcome variable being used is the pitch-level batting average on balls in play (BABIP). This value is 1 if the at bat resulted in a ball hit in play that allowed the batter to get on base, and 0 otherwise. Something to note, however, is that in regard to the definition of BABIP according to MLB and as will be used in this analysis, a homerun is not included in 'balls in play' because it does not involve action by the defense. This is appropriate for this analysis because I am considering the greater context of producing action for fans, and balls in play requiring action by the defense help decrease the limitations of the three outcome at-bat.

The primary predictors of interest are whether the pitch was a high strike or a low strike. However, I also control for the type of pitch thrown (fastball, changeup, or breaking ball), the speed of the pitch when released by the pitcher (in mph), the side of the plate the batter stands on (left or right), the handedness of the pitcher (left or right), the batted ball type (ground ball, pop up, line drive, or fly ball), and the launch

speed of the ball off the bat (in mph). I also considered including the projected distance the ball was hit, but as I include the type of batted ball which considers hit distance built in to the classification (eg. ground balls are hit into the dirt and have low distances and the only difference between a popup and a fly ball are the distance they travel), I did not think it would be ideal to include both variables. I chose to include batted ball type over distance because I believed that the type may incorporate more information on the action required by the defense to make a play compared to distance alone. All variables included in the model were less than 1% missing after filtering the data as described above (##################Figure 4 from Appendix B). Due to the low rates of missingness, I chose not to impute missing observations and records with missing data were dropped from the analysis.

EDA

Methodology

To explore the relationship between the vertical location of a pitch and whether the hit resulted in a successful batted ball, it is necessary to account for pitch characteristics that I expect may be associated with greater/reduced odds of the resulting batted ball being successful based on my existing knowledge of baseball and batting. First, I control for player relevant characteristics like pitcher handedness and batter stance side as these are characteristics relevant to each event of a batted ball and vary from pitcher to pitcher and batter to batter. I also control for the release speed of the pitch and the type of pitch as these characteristics are likely to have some association with the difficulty of hitting a given pitch. Controlling for the launch speed of the ball off the bat aims to account for the strength of the batter and caliber of contact (solid vs. weak) and also the assumption that the strength of the hit likely has some association with the success of the hit. Lastly, the type of hit is controlled for as different types of hits require different types of plays to be made by the defense and the difficulty of these plays potentially contribute to the odds that hit is successful. In our correlation plot (Appendix D), we see that there are generally low correlations between all of the covariates in our model. Note that low strikes and high strikes have a perfect correlation of -1 by design of the analysis, so we disregard this value. Also note that there are two other instances with correlations equal to or less than -0.5, but these either only show moderate correlation with a single level of a categorical variable (release speed and breaking balls) or events that would not occur at the same time (line drives and ground balls), so these variables are left in the model.

In addition to the main effects described above, I believe interaction effects between the type of batted ball and the pitch location (high/low) would also be interesting to examine. Specifically, I know from personal experience and background knowledge that it is more natural to hit high pitches up and more natural to hit low pitches down based on swing angles. This is why I suspect that different types of batted balls with different pitch locations may have different associations with the odds of that batted ball resulting in the batter successfully getting on base. For example if a high pitch is hit into the ground as a ground ball, this may be associated with a worse hit or weaker contact which would assumingly be less likely to result in successfully getting on base.

In this analysis, I consider each of the described pitcher and batter characteristics, pitch characteristics, hit characteristics, and interaction terms to be important covariates to account for in order to investigate whether there is an association between the pitch being a high or low strike and the odds of the resulting batted ball event being successful. I proceed with a logistic regression model for a few reasons. For starters, the logistic regression model is appropriate to use when the response variable is binary, as the pitch-level BABIP is. Second, I am interested in the association between these predictors and the odds of a batted ball event being successful and the logistic regression model can be easily interpreted in such a way. However, my confidence in the results of this model are potentially limited by the violation of certain model assumptions/diagnostics as I discuss in the following Assumptions and Diagnostics section.

$$\log(\frac{\hat{p_i}}{1-\hat{p_i}}) = \beta_0 + \beta_1(\text{Pitch Zone (Low)}_i) + \beta_2(\text{Pitch Type (Changeup)}_i)$$
$$+ \beta_3(\text{Pitch Type (Breaking Ball)}_i) + \beta_4(\text{Pitch Release Speed}_i) + \beta_5(\text{Batter Stance Side (Right)}_i)$$
$$+ \beta_6(\text{Pitcher Handedness (Right)}_i) + \beta_7(\text{Batted Ball Type (Ground Ball)}_i)$$
$$+ \beta_8(\text{Batted Ball Type (Line Drive)}_i) + \beta_9(\text{Batted Ball Type (Popup)}_i) + \beta_{10}(\text{Launch Speed}_i^5)$$
$$+ \beta_{11}((\text{Pitch Zone (Low)*Batted Ball Type (Ground Ball)})_i)$$
$$+ \beta_{12}((\text{Pitch Zone (Low)*Batted Ball Type (Line Drive)})_i)$$
$$+ \beta_{13}((\text{Pitch Zone (Low)*Batted Ball Type (Popup)})_i)$$

Where i refers to each pitch resulting in a batted ball event and $\hat{p}$ refers to the predicted probability of a successful batted ball.

Model Assumptions and Diagnostics

One assumtion of a logistic regression is the independence of obvservations included in the data. This independence assumption is violated due to there being multiple observations for the same batter and multiple observations for the same pitcher. Although the scope of the data is also limited to only a single pitch per at-bat, there are still multiple observations per game and team. Situational elements like score, outs, weather, defense, etc. also potentially contribute to non-independence. Still, considering the mentality of players that one play should not affect the next and the relative randomness of sports data, I chose to continue with this model under the assumption that the data may not be completely dependent on each other. However, this violation of independence will affect confidence in my results, as is discussed in Limitations.

The logistic model assumes that there is a linear relationship between the logit of the outcome and each predictor variable. To check the linearity assumption, empirical logit plots were observed to assess the relationships between the continuous predictor variables launch speed and release speed and the log odds of the outcome variable (Appendix E). In regard to the empirical logit plot for release speed, it appears that release speed is relatively linearly associated with the log odds of whether batted balls are successful or not. However, because the empirical logit plot for launch speed shows linearity being violated considering the presence of a moderately strong "U" shape, it is evident that launch speed does not have a linear relationship with the log odds of whether batted balls are successful or not. To avoid this linearity violation, I instead include a transformed version of launch speed (launch speed ^{5}) as it has a much more linear relationship with the log odds of whether batted balls are successful or not.

In assessing the fit of this model and its inferential purpose, I chose to examine the binned deviance residuals as a method of providing a rough idea of model fit (Appendix F). I decided to use deviance residuals rather than standardized residuals because the model is a generalized linear model (GLM) seeking to maximize the log likelihood function and thus deviance is a useful goodness-of-fit statistic. A small bin width of 0.01 was used to ensure many bins of average residuals vs average fitted value, allowing us to summarize our model fit while still showing more localized patterns in deviance residuals. Typically, we would expect these binned deviance residuals to be scattered around 0 to suggest good model fit. However, when examining this model, the binned deviance residuals show a oscillating pattern. Despite this just being one way to assess goodness-of-fit, it is important to acknowledge that there may be additional explanatory variables not included in the model that would help explain these patterns in deviance. Although this model is not a perfect fit, I will proceed to discuss the relationships suggested by the model, as they still offer some insight into my research question. Moving forward, however, it is necessary to keep in mind the presence of additional patterns unexplained by my model and how this potentially detracts from confidence in my results.

Results

Table 1: Results of logistic regression model

Variables	Odds Ratio Coefficient	95% Odds Ratio CI Lower Bound	95% Odds Ratio CI Upper Bound	P-Value (alpha = 0.05)
(Intercept)	0.024	0.012	0.049	<0.001
Pitch Zone (Low)	1.160	0.999	1.347	0.052
Pitch Type (Changeup)	0.953	0.857	1.059	0.374
Pich Type (Breaking Ball)	0.984	0.894	1.083	0.737
Release Speed	0.997	0.990	1.004	0.428
Batter Stance Side	0.966	0.907	1.029	0.278
Pitcher Handedness	0.984	0.927	1.045	0.607
Batted Ball Type (Ground Ball)	4.146	3.643	4.728	<0.001
Batted Ball Type (Line Drive)	14.333	12.584	16.361	<0.001
Batted Ball Type (Popup)	0.183	0.113	0.281	<0.001
Launch Speed^5	1.019	1.017	1.021	<0.001
Pitch Zone (Low)*Batted Ball Type (Ground Ball)	0.743	0.626	0.882	<0.001
Pitch Zone (Low)*Batted Ball Type (Line Drive)	0.917	0.769	1.093	0.335
Pitch Zone (Low)*Batted Ball Type (Popup)	0.906	0.400	1.932	0.805

In examining the results of the logistic regression model (Table 1 above), I was first interested in examining associations between the odds of a successful batted ball event and the location of the pitch, high or low. I found evidence of a significant association between the vertical pitch location and the odds of a successful batted ball, holding all other variables constant. That is, for pitches that resulted in a pop fly, I found that those that were low strikes were expected to have 1.164 times the odds of being a successful hit compared to a high strike. We found that the confidence interval for this estimate does not cross 1 ([1.002, 1.354]), so there is evidence of significant association between the odds of a successful hit and vertical pitch location, for these pitches.

However, I then proceeded to look closer into pitch location by looking at interaction effects between whether the pitch was a high or low strike and the type of batted ball. In doing so, I found evidence of a significant association between the pitch location and the batted ball being a ground ball, with all other variables held constant. That is, for pitches that are low strikes, batted balls that are ground balls are expected to have 1.907 [1.627, 2.237] times the odds of being a successful hit compared to those resulting from a high strike. That means that ground balls are associated with higher odds of a successful hit when the pitch is a low strike, but separately ground balls on low strikes do not seem to be associated with as great of a positive increase to the odds in comparison to other types of batted balls on low strikes. This is interesting because although the increase in the odds is expected according to my original justification of including this interaction effect, its comparison to the interaction effects of other types of batted balls on low strikes is not something I had expected.

Among the variables I controlled for, I noticed that oh high strikes, the type of batted ball has significant associations with the odds of a successful hit. Holding all other variables in the model constant, for high strikes resulting in line drives the odds of a successful hit are expected to increase by a factor of 14.215 ([12.470,16.239]) compared to fly balls. This made sense as line drives are typically hard hit balls but lack the elevation that potentially allows the defense higher chances of making a catch. Holding all other variables in the model constant, for high strikes resulting in ground balls the odds of a successful hit are expected to increase by a factor of 4.124 ([3.621, 4.706]) compared to fly balls. I was slightly surprised by this relationship in comparison to fly balls because it suggests ground balls are more likely to result in a successful hit that

a fly ball on high strikes. As fly balls are often the byproduct of failed homerun attempts, this association supports the idea that aiming to put balls in play rather than prioritizing homeruns may increase game action due to more successful balls in play. Holding all other variables in the model constant, for high strikes resulting in popups the odds of a successful hit are expected to decrease by a factor of 0.193 ([0.120, 0.293]) compared to fly balls. This is not surprising at all as popups are simply fly balls that don't reach the outfield, and with a higher density of infielders than outfielders in their respective areas, they are more likely to be caught.

Interestingly, there was not evidence of statistically significant associations between the odds of a successful batted ball and pitcher related variables like the pitcher handedness, the release speed, or the type of pitch.

Discussion

Given the current state of Major League Baseball and its strikezone, I had hypothesized that hitting low strikes would be associated with reduced odds of a successful batted ball. From the model, it was observed that this general trend was not supported. According to the model, in comparison to high strikes, batted balls on low strikes generally have increased odds of being a successful hit when controlling for other characteristics about the play, pitch, and batted ball, and interactions between vertical pitch location and batted ball type. The model also found evidence to support the intuition that different types of batted balls have different associations with the odds of a batted ball being successful. Interestingly, there was evidence that ground balls on high strikes were more likely to result in a successful batted ball that fly balls hit on high strikes, holding all else constant. As mentioned, because fly balls are often the byproduct of failed homerun attempts, this association supports the idea that aiming to put balls in play rather than prioritizing homeruns may increase game action due to more successful balls in play. Thus, this analysis seems to support both hitting low pitches and avoiding hitting popups and fly balls to increase game action for fans.

While these are interesting observations, it is important to consider that changing the strikezone or batting goals and swing behavior of players is not a simple or easy way to increase game action. For starters, shifting the strikezone would completely change pitching tendencies and pitching form and would require a gradual shift beginning with young players. Secondly, asking batters to change their mentalities and swings is not as simple as it sounds and is likely not a realistic expectations of professional athletes who have trained their whole lives and are likely fixed in their ways.

Limitations and Future Directions

The limitations of this research mainly lie in the scope of my research question and the model. To start, the intense filtering of the data prior to beginning my analysis may have prevented my model's ability to fully capture patterns in the data relevant to high and low strikes. For example, plays with an infield shift in place may have observed different associations or patterns that I missed by choosing to focus on plays with defensive alignments following rules that are soon to be implemented. I aimed to preserve the interpretability of this logistic regression model by limiting the examined scenarios and by only selecting variables that would allow me to do so. Thus, there are other possible covariates that might have improved this analysis had I controlled for them (such as pitch spin rate or launch angle), which I chose not to prioritize in this research.

Another limitation of our model is its reliance on defensive action and skill. Because I chose to only consider balls hit in play, the deciding factor of if it is a successful hit or not depends on if the defensive was successful in playing the ball and getting the batter out. Without quantifying or controlling for the defense's contribution to the play outcome, this analysis may be missing additional patterns in the data.

Also, it is crucial to consider the assumption violations this model makes in regards to independence. As this data includes repeated observations with the same pitcher and repeated observations with the same batter, the observations are not completely independent. However, because I was not interested in the variation within or between these players and their tendencies, I chose to proceed with the model as is. I made this decision in part due to guidance I have received in the past from professional baseball analysts,

considering different at-bats to be at least relatively independent due to situational differences and the inherent randomness of sports. Still, it is necessary to place emphasis on the fact that confidence in my results and interpretations must be hedged due to this independence violation.

Additionally, by filtering for only batted ball events, I exclude the observations where high or low pitches result in a swinging or looking strike. While these instances don't necessarily pertain to my research question, they are relevant to considering the overall associations between high and low strikes and game action. In the future, I would like to also consider the associations between pitch location and strikeout probability as reducing strikeout rates would also be a method of potentially increasing game action.

Separately, it would be interesting to consider the further implications of changing the strikezone such as how pitcher's pitch choice would change to best take advantage of a new strikezone.

References

https://www.mlb.com/glossary/rules/strike-zone

https://www.daktronics.com/en-us/support/kb/DD3312647

https://blogs.fangraphs.com/how-would-we-increase-balls-in-play/

https://blogs.fangraphs.com/what-a-smaller-strike-zone-can-do-for-pace-of-play/

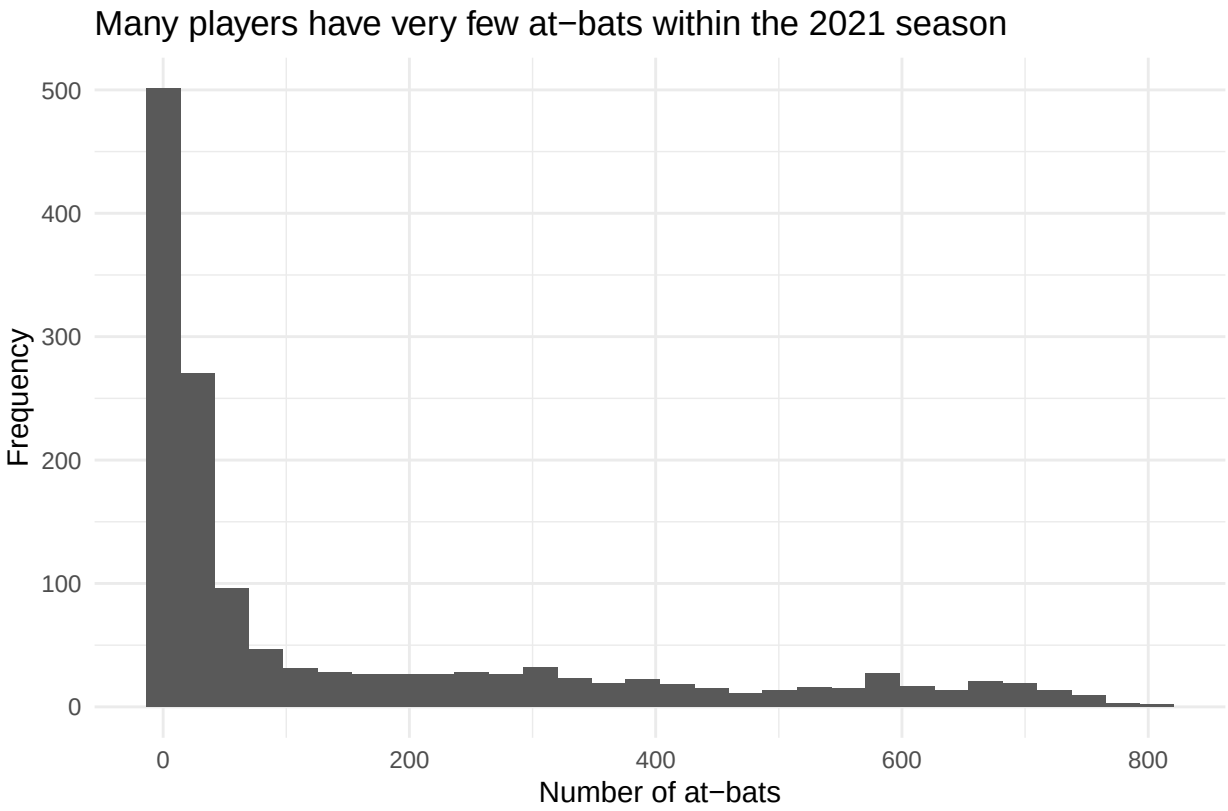
https://blogs.fangraphs.com/a-bigger-strike-zone-is-a-bad-idea/

Appendix

A: At-bat ending events

The following are considered at-bat ending events:

- Triples, fielded outs, doubles, grounding into double plays, strikeout double plays, sacrifice fly double plays, hitting into plays, hit by pitches, bunt foul tips, singles, strikeouts, sacrifice bunts, homeruns, sacrifice flies, catcher interference, other non-specified outs, sacrifice bunt double plays, foul tips, triple plays, fielders choices, double plays, fielders choice outs, field errors, force outs, and walks

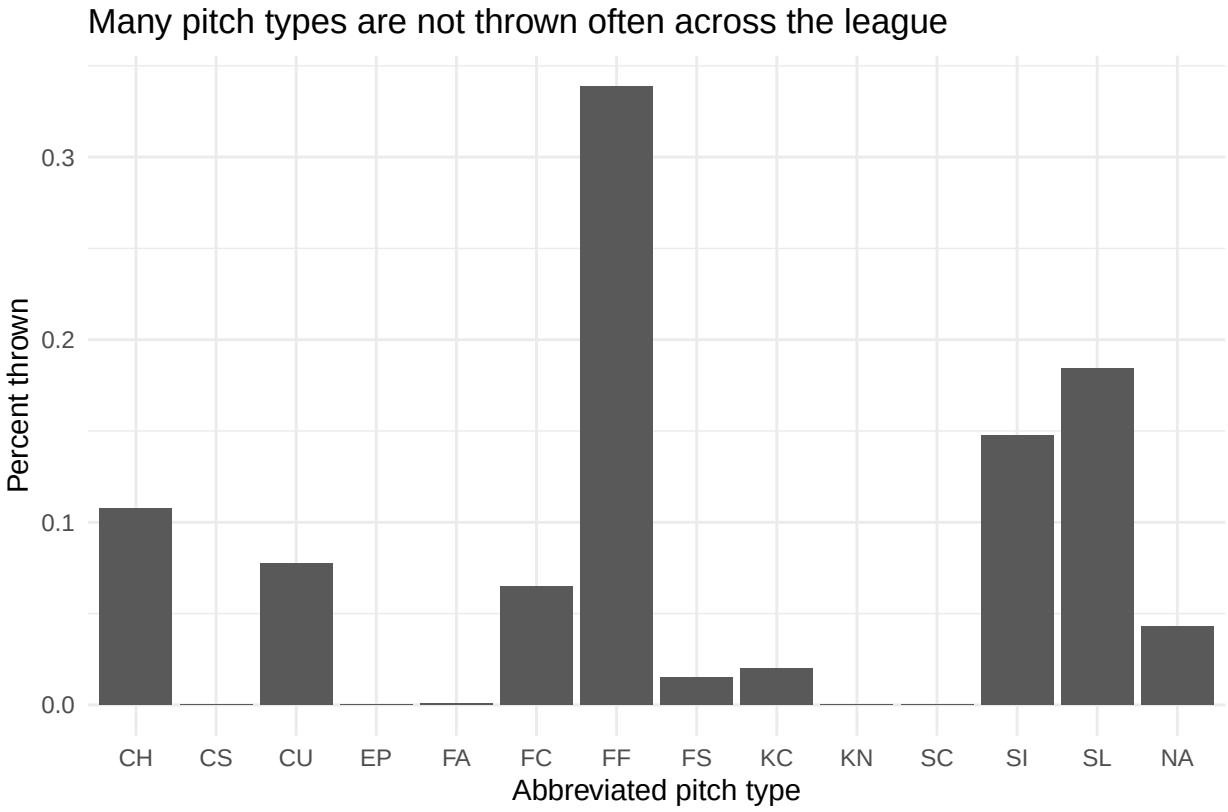


Appendix Plot A: Distribution of number of at−bats players completed

B: Proportions of pitch types

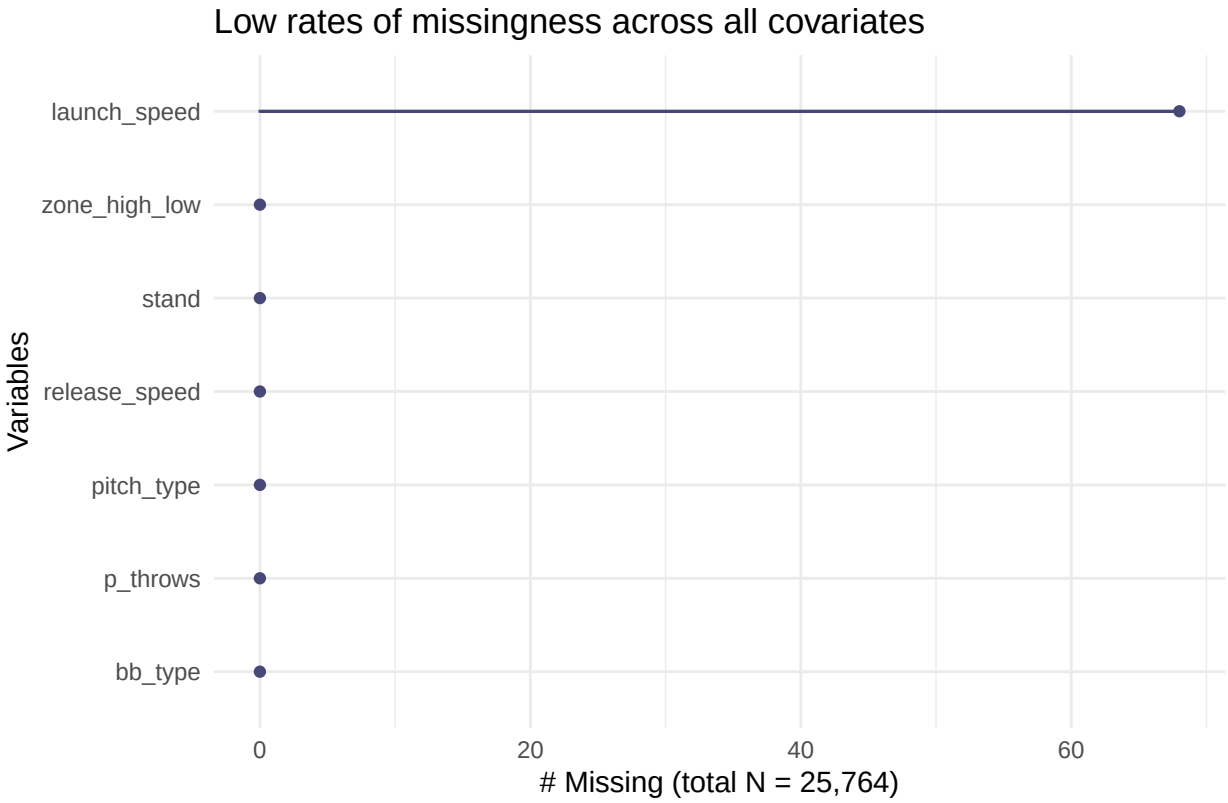
The following are the Pitch Types generated:

- AB Automatic Ball, AS Automatic Strike, CH Change-up, CU Curveball, EP Eephus, FC Cutter, FF Four-Seam Fastball, FO Forkball, FS Splitter, FT Two-Seam Fastball (synonymous with SI), GY Gyroball, IN Intentional Ball, KC Knuckle Curve, KN Knuckleball, NP No Pitch, PO Pitchout, SC Screwball, SI Sinker (synonymous with FT), SL Slider



Appendix Plot B: Proportions of different pitches thrown

C: Missing data



Appendix Plot C: Number of missing observations

D: Correlations between model covariates

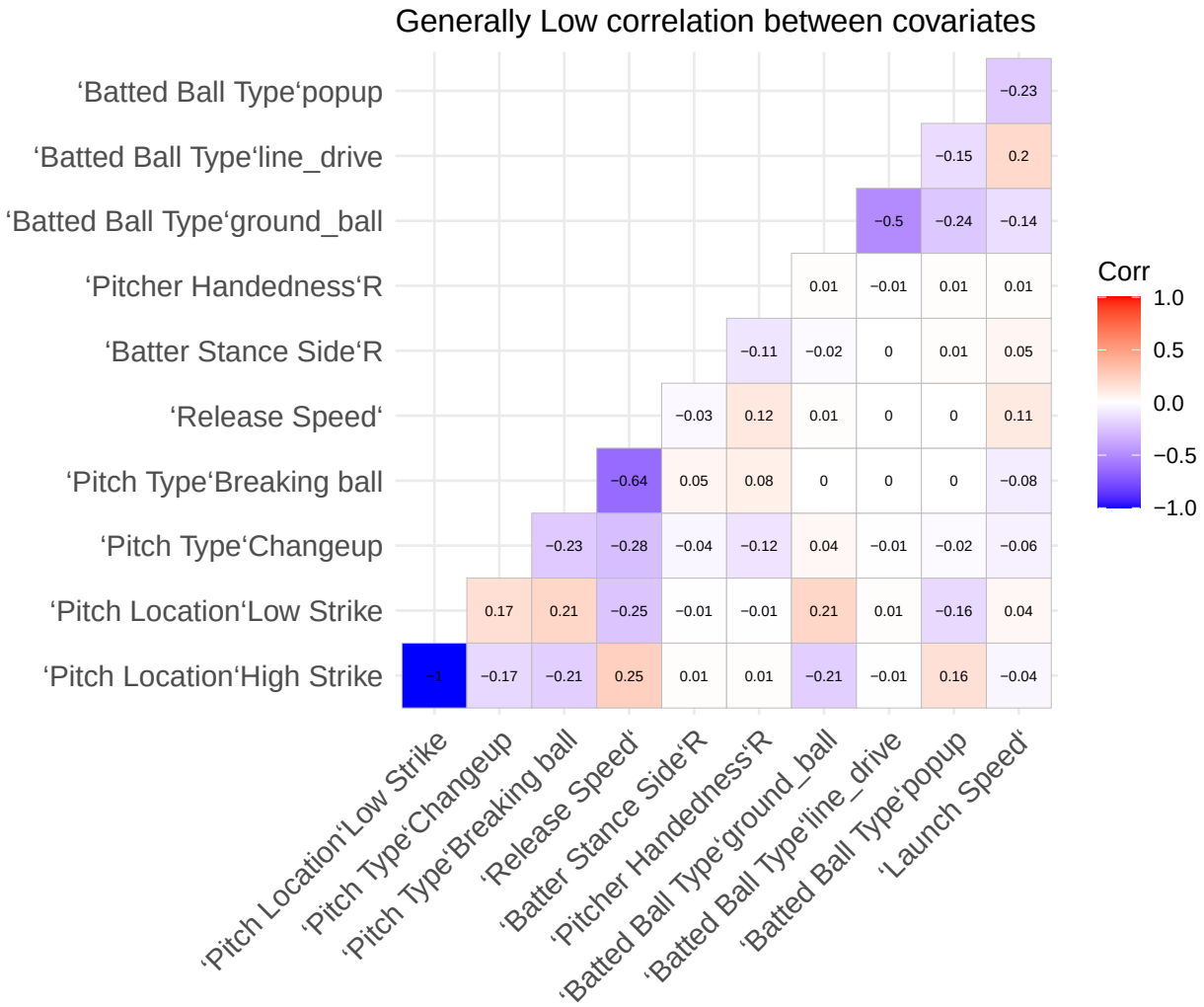
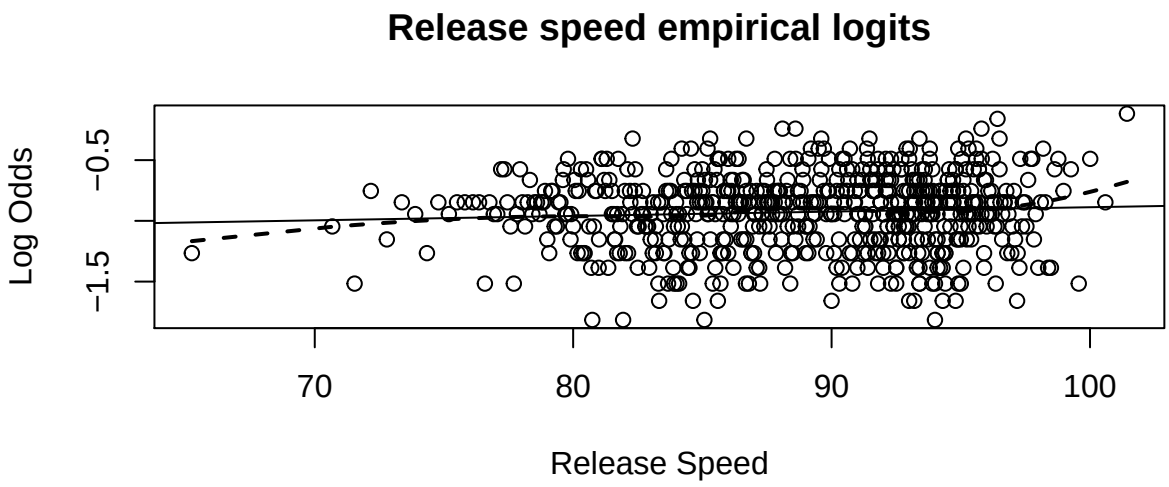


Figure 5: Correlations between model covariates

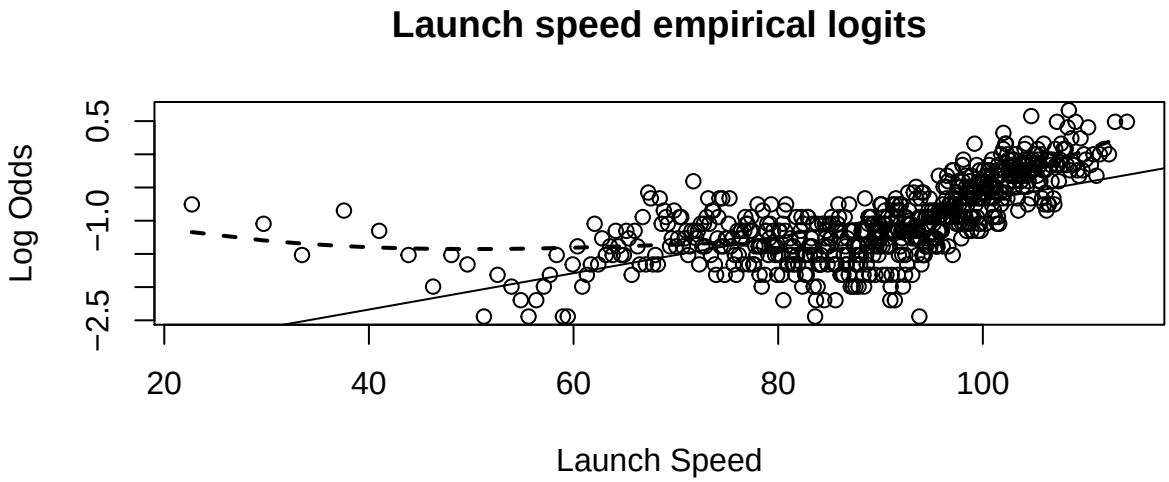
Model Diagnostics

E: Linearity Assumptions

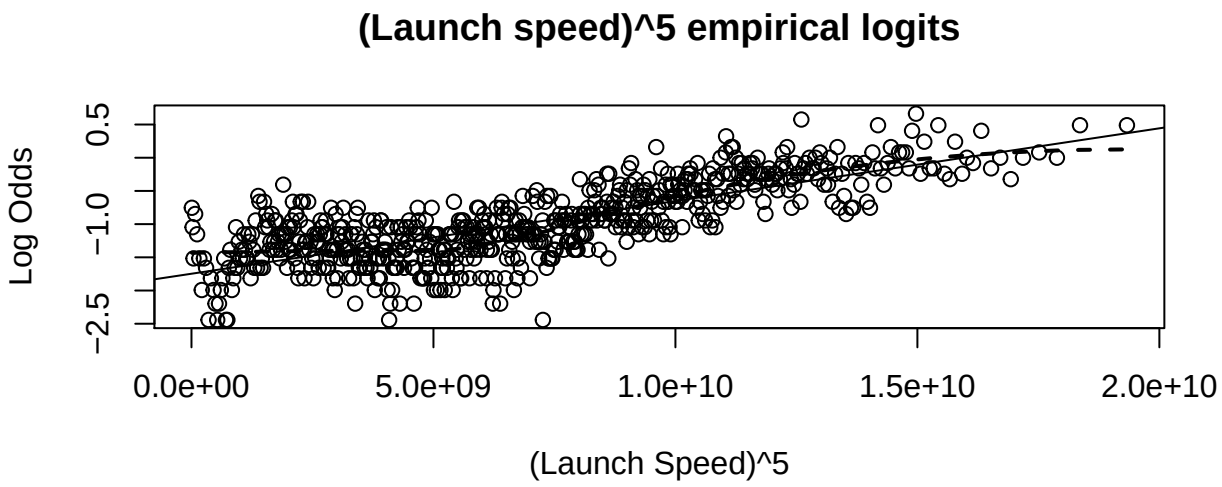
Appendix Plot E.1:



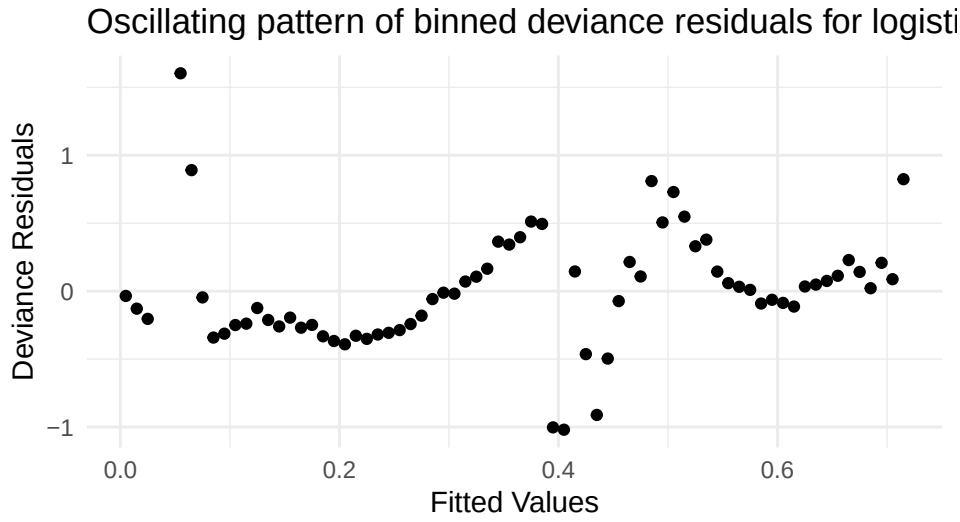
Appendix Plot E.2:



Appendix Plot E.2:



F: Model Fit



Appendix plot F: Assessing residuals of logistic model