Pitcher Injury Analysis

STATS 141XP Final Project

Kyle Biagan, Donggyu Kim, Min Jung Kim, Cherry Li, Jun Ryu

Statistics and Data Science
University of California, Los Angeles (UCLA)

9th July 2024

Contents

Abstract

In Major League Baseball (MLB), pitcher injuries significantly impact team performance. This study investigates whether a pitcher's spin rate predicts injury risk. Spin rate, while crucial for pitch effectiveness, may also increase injury risk due to the physical demands it places on pitchers. Using multiple linear regression and logistic regression models, our analysis found that a lower spin rate can put the pitcher at a higher risk of injury.

1. Introduction

Injuries were and remain a prominent part of sports. As players of the game tend to exert a lot of physical force within short periods, an injury can be a defining factor for individual health and overall team success. Especially in the case of baseball, a pitcher's heavy use of arm strength and the resulting pressure applied can be highly indicative of a pitcher's risk of injury.

One potential factor contributing to these injuries is the spin rate of the pitches. Here, spin rate is a measure of how much a baseball spins after the ball has been thrown, measured in revolutions per minute. Spin rate has gained attention for its relation to pitch effectiveness. However, the increased emphasis on maximizing spin rate may come at a cost: potentially leading to higher injury rates among pitchers. This study aims to analyze whether or not spin rate portends injury.

2. Background

2.1 About the Data

The data is provided by Stan Conte, a well-known MLB Physical Therapist and Athletic Trainer with 23 years of experience with major-league teams. There are two different data frames in the `SpinRateVsInjuryKit.RData` provided. The first is the `mx_out_DL_2021c` dataset, which has 221,407 observations and 16 variables. This data frame contains all the recorded transactions, which include a pitcher being put on the injury list. The second is the `mx_pitchFx_tall_2021c` dataset, which has 5,214,064 observations and 8 variables. This data frame contains all the pitching data tracking individual pitches.

2.2 Data Cleaning

We refined our dataset to facilitate easier analysis. To extract our variable of interest, "injury," we created a new column called `injury_duration`. This was done using the **grepl** function on the `description` variable in the `mx_out_DL_2021c` dataset. We filtered out rows with "injury" or "disabled" in the description. This allowed us to narrow down our analysis to determine if the spin rate portends injury. The `injury_duration` includes

information about whether the player's injury was included in the 7-day, 10-day, 15-day, 60-day, or full-season injury list.

For the `mx_pitchFx_tall_2021c` data, we focused on pitch type 2 (Fastball) instead of a mixture of pitch type 2 (Fastball) and 1 (off-speed ball). The rationale for this decision is that the spin rates between fastball and off-speed balls differ significantly and we were unable to distinctly identify the various types of off-speed balls within the pitches that were all categorized as pitch type 1 (off-speed ball).

A few more transformations to the dataset involved taking the sine and cosine of the spin direction to split the direction into two separate components: y-direction using sine and x-direction using cosine. We also transformed the `z0` column, which measures the vertical release point of the pitch. Since every pitcher's height may vary, we standardized this column by taking each observation and subtracting the corresponding pitcher's height. The pitcher's height was collected through the MLB API.[1]

Each row of the `mx_pitchFx_tall_2021c` represents the information of every single pitch thrown by a pitcher. Therefore, if a pitcher throws 30 pitches in an appearance, there will be 30 different rows of pitching information. Our goal was to group all the pitching data for each appearance of pitchers and determine whether the pitcher was injured during that appearance. To achieve this result, we grouped rows with the same dates and pitchers and calculated the average values for the rest of the columns. Instead of taking only the last pitch thrown by a pitcher before an injury is recorded, we considered that all pitches thrown during an appearance collectively contribute to the risk of injury. Thus, we calculated the average values for all pitches thrown by a pitcher on the same day.

Now with the dataset that contains the average values, we wanted to combine this with the `injury_duration` extracted earlier. When observing the data, we noticed that if a player is injured during or immediately after a game, in most cases, the injury effective date begins the next day of the game appearance.[2] Therefore, we added one day to the game appearance date to match it with the injury effective date. Then, we were able to merge the two in order to align each injury to a game, if applicable. As a result, each row in the combined data represents the average values of all pitches thrown by one pitcher during an appearance, with the last column containing the `injury_duration` values.

The combined dataset consists of 103,323 rows, with only 1,029 rows indicating an injury. This means that a pitcher has approximately a 1% chance of getting injured during an appearance. With a total of 4,535 injury records for pitchers, we found that 1,029 of these injuries, or 22%, occurred during or immediately after a game is played. In the combined dataset, we were unable to include full-season injuries because none of the full-season injuries matched the game-played dates.

However, we could not model using this dataset as is. Valid modeling requires a balanced comparison between injured and non-injured data. If a player has never been

[1] 'MLB Data API', n.d.
[2] '10-Day Injured List', n.d.

injured, their data does not contribute to understanding the factors that lead to injury. Including players with no injury history would introduce significant imbalance and reduce the effectiveness of our models. Therefore, we filtered out players who never had an injury. This step was crucial for our analysis, ensuring that our dataset is more balanced and suitable for modeling, which would then allow us to understand the variables that contribute to injury risk.

Below is a description of our variables of interest after data cleaning:

Variable	**Type**	**Data Type**	**Description**
injury_duration	Outcome	Categorical	Duration of the injured list (IL)
spin_rate	Predictor	Numerical	Rate of spin on a pitch (RPM)
spinDir_x	Predictor	Numerical	Direction of spin in the x-direction
spinDir_y	Predictor	Numerical	Direction of spin in the y-direction
startSpeed	Predictor	Numerical	Velocity of pitch (mph)
nEvents	Predictor	Numerical	Number of events for pitcher-date-game
x0	Predictor	Numerical	Lateral release point
fastball_ratio	Predictor	Numerical	Proportion of fastballs thrown
z0_new	Predictor	Numerical	Vertical release point

Table 1: Variable List

After data cleaning, our final dataset consisted of 50,770 observations and 12 variables.

2.3 Explanatory Data Analysis

To get a better idea about our data, we conducted some explanatory data analysis.

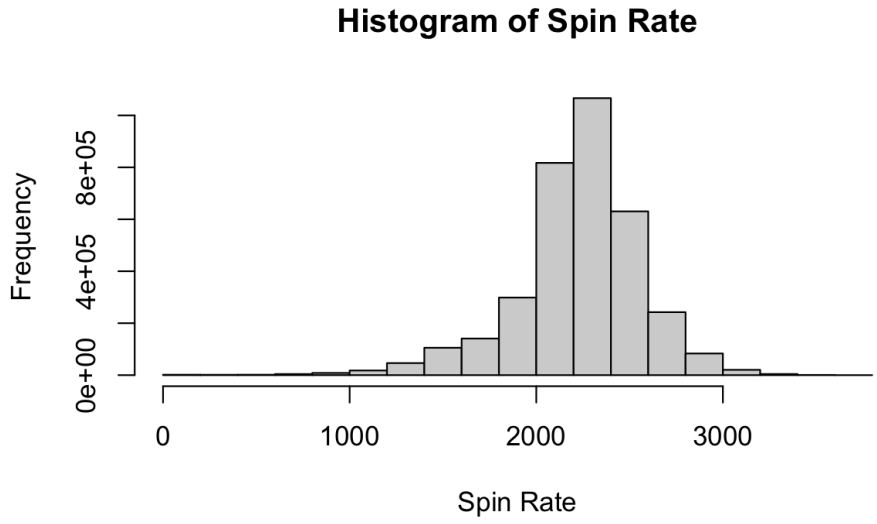

Figure 1: Histogram of Spin Rate

Figure 1 represents the distribution of spin rates. It shows a range of spin rates,

with most values concentrated between 2000 and 2500. The distribution appears to be normally distributed.

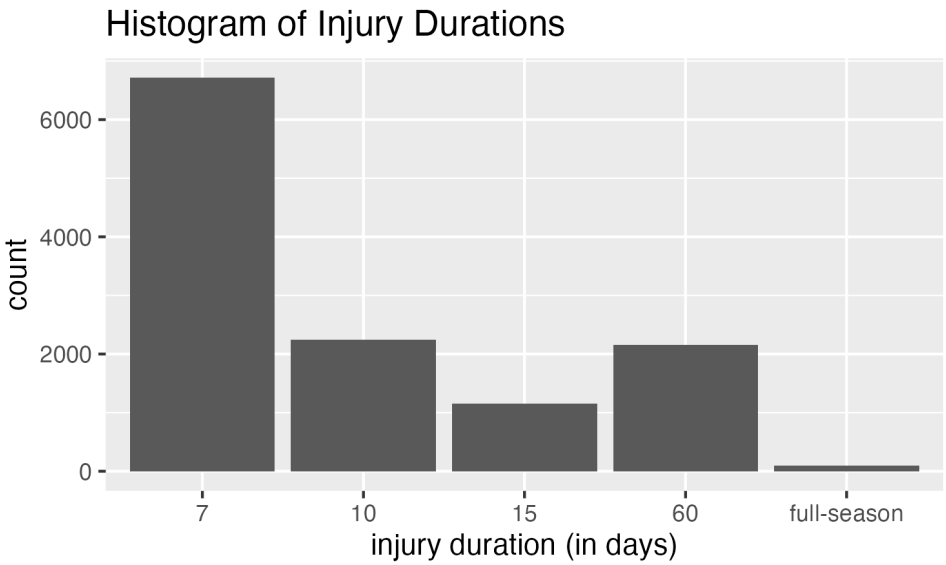

Figure 2: Histogram of Total Injury Durations

Figure 2 illustrates the frequency distribution of various injury durations. The 7-day injury duration has the highest frequency, indicating it is the most common injury duration. Additionally, the 10-day and 60-day injury durations exhibit similar frequencies, suggesting comparable occurrence rates for these durations.

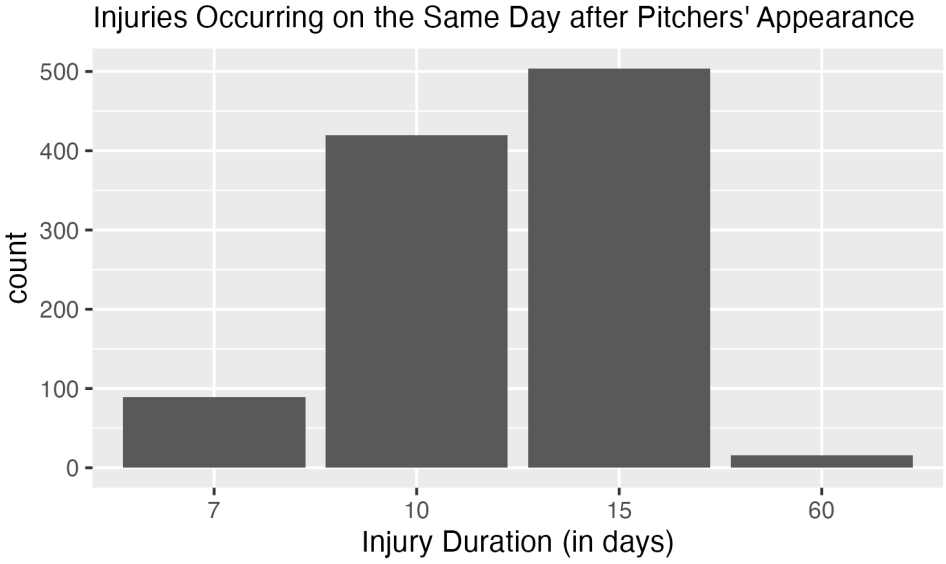

Figure 3: Histogram of In-game Injury Durations

Figure 3 shows the injuries sustained by pitchers either during the game or immediately after. Unlike Figure 2, where the 7-day injury duration had the highest frequency, this plot shows that the 15-day injury duration is the most common.

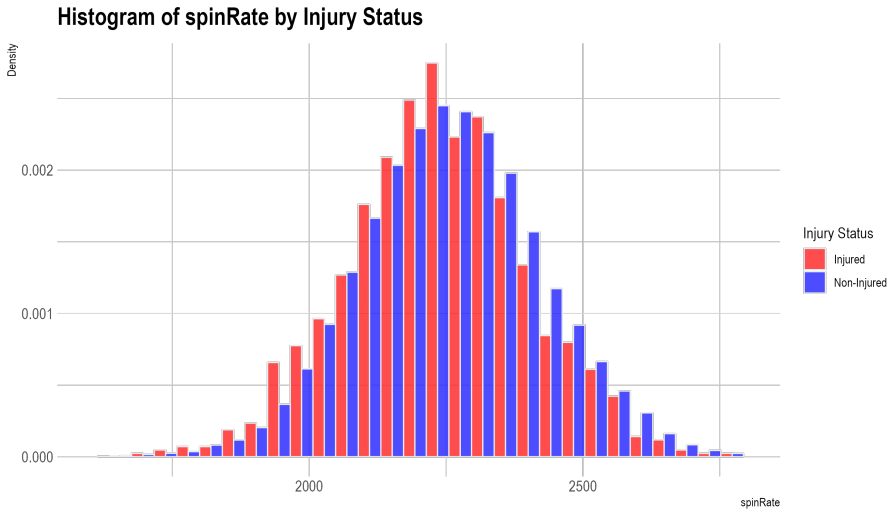

Figure 4: Histogram of Spin Rate and Injury Status

Figure 4 illustrates the distribution of spin rate among pitchers based on injury status. It aggregates all pitches from a pitcher's game day and calculates the mean, resulting in a distribution that resembles a normal curve. From the plot, we observe that the red bars representing injured players are slightly denser on the left side, while the blue bars representing non-injured players are more distributed to the right. This pattern suggests a potential association between lower spin rates and an increased risk of injury. To explore whether this relationship holds statistical significance and to examine the influence of other variables on injury likelihood, we will utilize common statistical models for further analysis.

3. Methodology and Results

To analyze whether spin rate portends injury, we used the multiple linear regression and logistic regression models.

3.1 Model 1: Multiple Linear Regression

Our multiple linear regression model includes the `injury_duration` against the following predictors: `startSpeed`, `x0`, `spinRate`, `fastball_ratio`, `spinDir_x`, `spinDir_y`, `z0_new`, `nEvents`.

	Estimate	Std. Error	t value	Pr(|t|)
(Intercept)	2.123e+00	3.445e-01	6.162	7.23e-10
startSpeed	-1.709e-02	3.540e-03	-4.827	1.39e-06
x0	-6.888e-03	1.216e-02	-0.567	0.5710
spinRate	-1.418e-04	5.942e-05	-2.386	0.0171
fastball_ratio	5.836e-02	5.039e-02	1.158	0.2468
spinDir_x	-3.584e-02	7.423e-02	-0.483	0.6292
spinDir_y	-2.960e-02	3.807e-02	-0.777	0.4369
z0_new	4.724e-02	2.584e-02	1.828	0.0676
nEvents	2.485e-04	3.050e-04	0.815	0.4152

Table 2: Summary of Model 1

	Value
Residual standard error	2.052
Multiple R-squared	0.0009124
Adjusted R-squared	0.000755
F-statistic	5.795
p-value	2.049e-07

Table 3: Summary Statistics of Model 1

As can be seen in Table 2, the significant predictors are start speed and spin rate. The coefficient for start speed is 0.01709. This means that for each unit increase in start speed, the injury duration decreases approximately by 0.01709 units, holding all variables constant.

The coefficient of spin rate is -0.0001418. This means that for each unit increase in spin rate, the injury duration decreases by approximately 0.0001418 units, holding all variables constant. The p-value for spin rate is 0.0171, which is less than 0.05, indicating that the effect of spin rate on injury duration is statistically significant at the 5% level. Thus, there is sufficient evidence to conclude that spin rate affects injury duration. It is important to note that while this effect is statistically significant, the magnitude of the effect is very small.

Furthermore, `z0_new` has a p-value of 0.0676, which is slightly above the 0.05 threshold, suggesting that the evidence against the null hypothesis is not strong enough to declare statistical significance at the 5% level. However, it is close to the threshold, indicating that there may be a trend or weak evidence suggesting that `z0_new` is associated with injury duration. The coefficient can be interpreted as: for each unit increase in `z0_new`, the injury duration increases by approximately 0.04724 units, holding all other variables constant.

As noted in Table 3, the model's R-squared value is low (0.0009124), indicating that the model explains less than 0.1% of the variability in injury duration. This suggests that the predictors in the model, including spin rate, do not explain much of the variation in injury duration. While the spin rate is statistically significant, its practical significance

may be limited due to the small effect size and low explanatory power of the model. However, it is important to note that most of the events of `injury_duration` are 0, meaning no injury has occurred.

To learn more about our multiple linear regression model, we plotted a diagnostic plot to test for linear model assumptions.

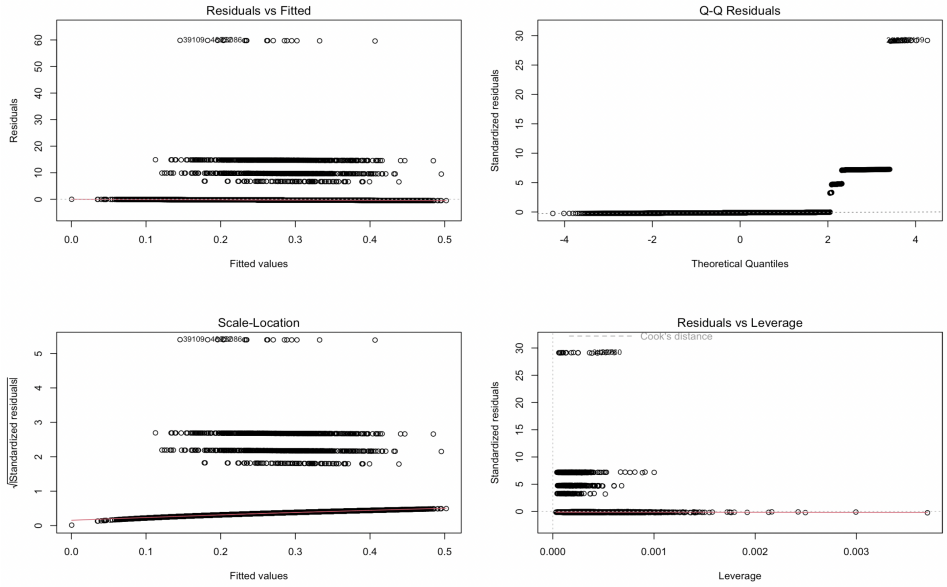

Figure 5: Multiple Linear Regression Diagnostic Plots

The diagnostic plot above presents several issues of non-linearity, non-normality of residuals, heteroscedasticity, and influential observations with high leverage; however, since the response variable is only among the values {0, 7, 10, 15, 60} (which is technically a categorical variable), traditional methods (i.e. diagnostic plots as in Figure 5) for checking normality and linearity is not directly applicable. Therefore, we have applied the logistic regression model for more accurate analysis.

3.2 Model 2: Logistic Regression

Our logistic regression model incorporates the same predictors as the multiple linear regression model with one minor difference: we added another feature called `spinRate_sd`, which takes into account the standard deviation of the spin rates in a single game-pitcher instance. Note here that since logistic regression is mostly a binary classification algorithm, we labeled any instance of injury as label 1 and any instance of non-injury as label 0.

Due to the unbalanced dataset we have, it required us to figure out a threshold value other than the default value of 0.5 for the model. With the default threshold of 0.5, the model had no better approach but to classify everything as 0 or non-injured since that is the majority class in our dataset.

To tune the threshold for the logistic model, we first needed to see a level of imbalance between the two classes; specifically, the proportion of observations with label 1 over the whole set. This proportion came out to be about 0.02, which again shows how unbalanced the dataset is. The next step was running the logistic model and extracting all the $\hat{p}$ values for each observation, which represent the estimated probability that the dependent variable will be labeled 1. Then, out of all these $\hat{p}$ values, we took the 98th percentile value and set that as a threshold. The rationale for this is that we want approximately 2% of the predicted values to be of label 1 so that it would match the proportion derived from the dataset.

The following classification report corresponds to a logistic regression model fitted with a regularization parameter C of 0.5:

	Precision	Recall	f1-score	Support
0	0.98	0.98	0.98	9575
1	0.04	0.03	0.03	210
accuracy			0.96	9785
macro avg	0.51	0.51	0.51	9785
weighted avg	0.96	0.96	0.96	9785

Table 4: Summary of Model 2

From Table 4, we see that the macro average score is 0.51, while the weighted average score is 0.96, which paints two contrasting narratives. Here, the more accurate measure is the macro average score, which indicates a low success in predicting injury status. The reason why the macro average score makes more sense is that from the above table, we see that the precision score for label 1 (indicating injury) is 0.04, which is extremely low. Since the model cannot predict injury well, the average precision score cannot be high. Thus, we will visualize the results of the model using a confusion matrix.

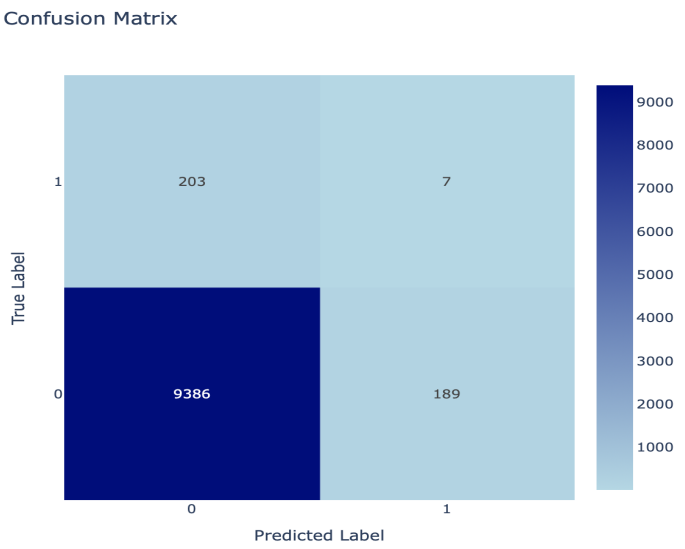

Figure 6: Confusion Matrix using Logistic Regression

The confusion matrix in Figure 6 confirms that the accuracy of the model is quite low. Although there are 9,386 correct classified instances of non-injury, there are also a nontrivial amount of misclassified instances (as noted by the numbers in the top left and bottom right squares). Specifically, for classifying label 1 (injury), only 7 out of 210 instances were labeled correctly, revealing the weakness of the results.

The weakness shown in the logistic model could be due to many factors; however, we suspected that some of the predictors involved in the model were not useful or significant. With this in mind, we also wanted to plot the coefficients from the logistic regression for each feature to see if we could extract some of the more significant features. The following is the plot of the feature coefficients:

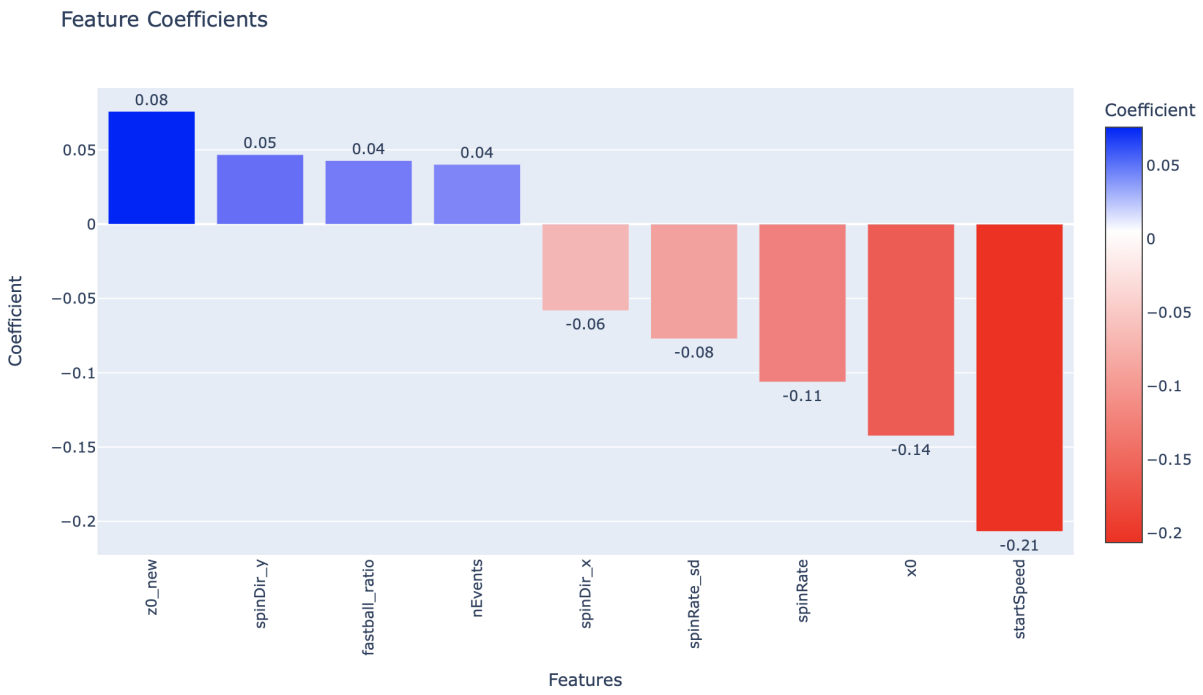

Figure 7: Feature Coefficients

In the above plot, we see the coefficients for each feature, indicating the change in the log odds of the response variable. What that means is that every time we increase one specific feature by one unit, the log odds of the response variable will change by whatever the coefficient value is for that specific feature. This is a comparative measure that we can use to assess feature importance as each feature has been standardized through `StandardScaler` implemented in the model. Thus, the positive coefficients indicate that for every unit increase of the predictor variable, the odds of success (or in this case, the chance of injury as that is our label 1) will increase. If the coefficient is negative, then the odds of success will decrease, meaning that the chance of injury will decrease.

Here, we assess some of the more extreme values such as `startSpeed`, `x0`, `spinRate`, and `z0_new`. These extreme values indicate, regardless of their sign, that their impact on the odds of success is large. Thus, these features can be seen as the more significant attributes in predicting injury status.

We also wanted to confirm the findings from the coefficient plot by extracting the p-values of each feature from the logistic model.

Coefficients	P-value	
(Intercept)	4.45e-0.5	***
startSpeed	7.94e-11	***
x0	0.29267	
spinRate	0.00133	**
fastball_ratio	0.15403	
spinDir_x	0.26820	
spinDir_y	0.52464	
z0_new	0.03002	*
spinRate_sd	0.06723	.

Table 5: P-Values for Features

From Table 5, we do find some features significant in terms of p-values, such as `startSpeed`, `spinRate`, and `z0_new`. These features seem to match what was discovered through Figure 7, which confirms the significance of these features. To analyze some of these features, increasing the `startSpeed` and `spinRate` by a standardized unit of one will result in lower odds of success or lower risk of injury. On the other hand, increasing `z0_new` by a single unit will increase the risk of injury.

4. Conclusions

The multiple linear regression model indicated that spin rate and start speed are significant predictors of injury duration. Our analysis shows that as the spin rate increases, the injury duration slightly decreases. The model also indicated that as the start speed increases, the injury duration decreases. `z0_new` has a p-value of 0.0676, which is close to the threshold of 0.05, indicating there may be weak evidence suggesting that `z0_new` is associated with injury duration. `z0_new` has a positive coefficient, meaning pitchers with higher release points have a higher risk of injury. However, the magnitude of both of the individual effects is extremely small. The R-squared value of the model is also very low, indicating that predictors in the model do not explain much of the variation in injury duration.

When running logistic regression as a full model with 9 different predictors, the results were not optimal and did not present a significant finding. The macro average precision value was so low only performed marginally better than the "benchmark" approach, which would've been where we just classified everything to be of a single label. The "benchmark" approach would have given us a macro average value of 0.5, and our logistic regression gave us 0.52, showing the weakness of using the full model.

When we proceeded to determine feature importance using coefficients in the model, we found that a couple of features were certainly more significant than others. These

features namely were `startSpeed`, `x0`, `spinRate`, and `z0_new`, and we found that all of these features except `x0` were significant in terms of their p-values.

In conclusion, the spin rate is a significant predictor of injury risk, with a negative coefficient in the logistic model. This suggests that pitchers who show a lower spin rate on average are at a higher risk of injury. Specifically, if a pitcher's spin rate and start speed during their appearance is lower than usual, the risk of getting injured increases as we also found start speed to be a significant predictor with a negative coefficient. Additionally, with the p-values of `spinRate_sd` and `z0_new` nearly below 0.05, pitchers who show less variation in their spin rate and have a higher release point than usual are potentially at a higher risk of injury.

5. Limitations

One significant limitation is our lack of pitch information (i.e. spin rate) corresponding to a specific injury. As previously mentioned, only 22% of the injury occurrences in our dataset could be traced back to a specific game. One reason for this is that injuries can be recorded retroactively, up to three days after they occur. This introduces potential inaccuracies in the effective dates of injuries, which can affect the precision of our analysis. Another major reason is that most injuries occur during Spring Training, and our dataset only includes pitch information for game appearances. Thus, we are unable to connect Spring Training injuries to the pitcher's spin rate. These discrepancies need to be considered when interpreting the relationship between spin rates and injury occurrences.

As a result of these discrepancies, our dataset is unbalanced, in the sense that the mass majority of our data corresponds to no injury. Our final dataset consisted of 50,770 observations with only 1,029 of those observations indicating an injury. With this lack of balance, it is more difficult to pick up on the patterns in spin rate that factor into a higher risk of injury.

Another limitation is that we only used fastball data. For a more precise analysis, it is essential to understand the characteristics of various off-speed balls and include their spin rate in the analysis.

One last important limitation to keep in mind is that injury list data is not entirely reliable. It can be utilized to cycle players—in this case, pitchers—on and off the active roster. Thus, in comparison to official medical records, injured lists may entail some form of data manipulation.

For future research, it would be interesting to investigate the effect that a past injury may have on the association between a pitcher's spin rate and injury status. For instance, a past injury resulting in lingering health issues may be associated with a lower spin rate and thus a higher risk of another injury.

6. Bibliography & References

10-day injured list. (n.d.). https://www.mlb.com/glossary/injuries/10-day-injured-list

Baseballr r package. (January 16, 2024). https://cran.r-project.org/package=baseballr

Logistic regression. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression

Mlb data api. (n.d.). https://appac.github.io/mlb-data-api-docs/