Final Research Project - Predictive Analysis of Pitchers In The MLB Using Spin-Rate

Tyler Scott

12/01/2022

Introduction

Baseball is a sport that has been around for almost two-hundred years and just like anything that has been around for an extensive amount of time, it has changed drastically over the years. With advancements in technology, physical education, and everything else that has improved since the berth of baseball, the players have increasingly gotten better as well. Stories of players like Babe Ruth eating hot dogs and drinking beers before the game simply do not exist with players in the game today. Year after year players are throwing harder and harder with no stop in sight. This increase in skill level has also transpired at the youth levels as one can turn on the Little League World Series and see a twelve year old kid touching eighty miles an hour. With that being said, while skill level has routinely increased over time in the sport, injury has grown along with it.

Injury throughout all levels of baseball has increased drastically through the years. It seems there are more and more high school kids that undergo Tommy John Surgery every single year. Let that sink in, a surgery for an injury once regarded as career altering and possibly career ending has trickled to a high school level.

This rise in injury across all levels can be attributed to many different aspects. As stated before, the velocities at which players propel objects or themselves as increased at a drastic weight. It is reasonable to assume that the force required to produce such velocities puts high levels of stress that people's extremities are not trained to handle. For example, it has been observed that the UCL (Ulnar Collateral Ligament) in a person's elbow can handle a load of force up to on average 32Nm. Throwing a baseball at hide speeds as pitchers do in the MLB puts a load of stress that has potential to be higher than this threshold. Now, this stress is dispersed throughout the arm and not just the UCL which is why pitchers don't experience UCL tears on every pitch. That being said if a pitcher has relatively weak body composition there is more stress placed on the UCL putting a pitcher at risk. Not only do weaker body compositions put pitchers at higher risk but, mechanical inefficiencies also have an effect as well. A pitcher who throws a ball at high velocities without being mechanically efficient is also at high risk for injury. This is due to pitchers that are able to optimize body movement are able to reduce stress on the UCL relatively to a person who don't use their body in a correct manner to manage the load of force.

With advances in technology we are able to not only use this for metrics to raise player performance, we can use this new information as an attempt to limit injury in the sport as well. Take the company Trackman for example, this company has developed a monitoring system to where they can measure any metric about a pitch one could ask for such as velocity, spin-rate, horizontal break, height at release, and so much more. Using advanced tracking information like this we can use metrics as an tool to spot opportunities for injury development before they reach a stage at which serious medical attention is needed.

This now ties us into the purpose of this research as it will be broken up into two parts. In the first part, we will use available spin-rate data from MLB pitchers to predict expected batting average against which will be referred to from now on as xBA. The point of investigating this relationship is to see if the spin-rate has a genuine effect on the performance and success of this pitch. By finding whether spin-rate does have an important effect on the success can allow for better measuring of a pitcher's ability to perform.

In Part 2 of the research we will again use this spin-rate data from MLB pitchers to predict which players may need Tommy John Surgery based off the difference in spin of their pitches. As an example for further understanding, take player A who has a fastball of 2,000 RPM (Revolutions Per Minute) with a slider of 2,400 RPM and player B who has a fastball of 2,300 RPM with a slider of 2,200 RPM. Say player A remains healthy and player B ends up needing Tommy John surgery. We are going to use these differences in RPM for player B in order to attempt to classify players that will get injured and use player A's differences in RPM to classify players that will remain healthy.

The reason for trying to predict the need for Tommy John surgery is due to the nature of the injury. Many types of injuries are spontaneously caused by a specific traumatic event like getting hit with a pitch that fractures the hand. This type of injury is due to the specific event and most likely not because of a build up of trauma over a longer period of time. Predicting this type of injury is seemingly impossible. A UCL tear, which is the type of injury fixed by Tommy John, is a different type of injury. This injury is caused

by a slow build up of stress on the UCL that starts with micro-tears in the ligament that progressively grow with repeated force. Due to the nature of this injury occurring over time due to repeatable action (pitching) we are able to form predictions on the likelihood of this injury happening.

The basic idea of using RPM to predict injury relates back to our idea presented earlier about mechanical efficiency. We are attempting to find a relationship between the difference of RPM of pitches in order to spot out mechanical inefficiencies in a pitcher's pitching motion by using spin-rate. The measure of spin-rate is a relatively new metric that we are able to measure and could possibly hold valuable information about a pitcher's delivery as spin-rate is the result of the output of the entire motion of a pitcher to throw the ball.

Data

The data being used for this research was retrieved from the website Baseball Savant. The website allows you to customize pitcher data based off many different types of inputs. For our data, we are using average spin-rates of individual pitches from the years 2015-2022. For example, Walker Buehler will have one entry for each year between 2018-2022. The website link is listed below:

https://baseballsavant.mlb.com/statcast_search

Due to the amount of pitches that a player is able to throw we have categorized pitches based off the nature of the type of pitch. For example, a traditional curve ball and knuckle curve are similar pitches with similar actions in terms of bodily motion and flight of the baseball. Due to these pitches being similar in nature they will be categorized under the same category. Below is a table of the pitches that are represented in each category:

	Pitches Included
Category: Fastball	4-Seam Fastball | 2-Seam Fastball | Sinker | Cutter
Category: Offspeed	Changup | Split-finger
Category Breaking Ball	Curveball | Knuckle Curve | Slow Curve
Category: Slider	Slider

As you can see, some more uncommon pitches, like a knuckle ball, were left off. This is because of the rarity of their usage and the potential for a drastic effect on the model.

The next piece of data that will be used is categorical data denoting whether a player needed Tommy John that season. This is a little more complicated as we are able to load in a data set that denotes the player, the date they got the surgery, and a few more supporting facts as well. The link to this data set is below:

https://docs.google.com/spreadsheets/d/1gQujXQQGOVNaiuwSN680Hq-FDVsCwvN-3AazykOBON0/edit#gid=0

Using these two data sets and some data cleaning maneuvers we are able to retrieve our final data frame which we will conduct our research. Below is a portion of our final data frame as well as a data dictionary denoting an explanation for each variable included:

Variables	Type	Explanation
player_name	Character	Name of Pitcher
fast_spin	Integer	Average Fastball RPM
off_spin	Integer	Average Offspeed RPM
slid_spin	Integer	Average Slider RPM
break_spin	Integer	Average Breaking Ball RPM
year	Integer	Year
xba	Numeric	Expected Batting Average Against
tj	Integer (1 or 0)	TJ Needed | 1-Yes, 0-No

	player_name	fast_spin	off_spin	slid_spin	break_spin	year	xba	tj
1	Miller, Shelby	2216	1789	0	2178	2015	0.226	0
2	Lackey, John	2137	1956	0	2082	2015	0.244	0
4	Lynn, Lance	2349	1903	2304	2032	2015	0.234	1
5	Lester, Jon	2213	1653	0	2360	2015	0.225	0
6	Haren, Dan	2162	1383	0	2212	2015	0.249	0
7	Cueto, Johnny	2264	1558	2101	1993	2015	0.243	0

Looking at our above table, there are a few spin rate entries that have zeros. If a pitcher doesn't throw a certain pitch then the RPM for that pitch is set to zero.Naturally, all pitchers are different with their pitch arsenal and don't have a pitch in every of our categories. To account for this we will break our observations up into three groups: all, fast-off-slid, and fast-off-break. All other pitchers that don't fit into our category will be thrown out. Having a zero in a category and using a prediction including that zero would be misleading and throw off our accuracy.

Part 1 - Predicting xBA With Spin-Rate

Methodology

As touched on in the introduction, this part will be using spin-rate to predict xBA. It will be done by using the expected batting averages for each entry in our four different pitch categories explained in the introduction. Each pitch category will have it's own model fitted. The reason for this is that different pitchers have different strengths when it comes to their pitch arsenal. Due to this, it would be rather difficult to fit a significant model predicting xBA for each pitcher. We are able to learn more about each individual pitch by creating different models and evaluating the relationship from each individual model.

The models were creating by dividing the data up into a 70%/30% testing/training split and fit using SLR. The data is not high-dimensional for this application as we are fitting four different models for our pitch categories so their is only one predictor variable in each relationship. Due to this, SLR was the chosen way to measure xBA from spin-rate.

Results

Looking below, we are able to see our four estimate tables for the four different models fitted.

Fastball Model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4463839	0.0154342	28.92174	0
spin_rate	-0.0000763	0.0000069	-11.05664	0

Off-Speed Model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2649260	0.0152901	17.3266451	0.0000000
spin_rate	-0.0000017	0.0000088	-0.1990032	0.8422727

Breaking-Ball Model:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3337636	0.0194969	17.118789	0.0e+00
spin_rate	-0.0000392	0.0000080	-4.881651	1.1e-06

Slider Model:

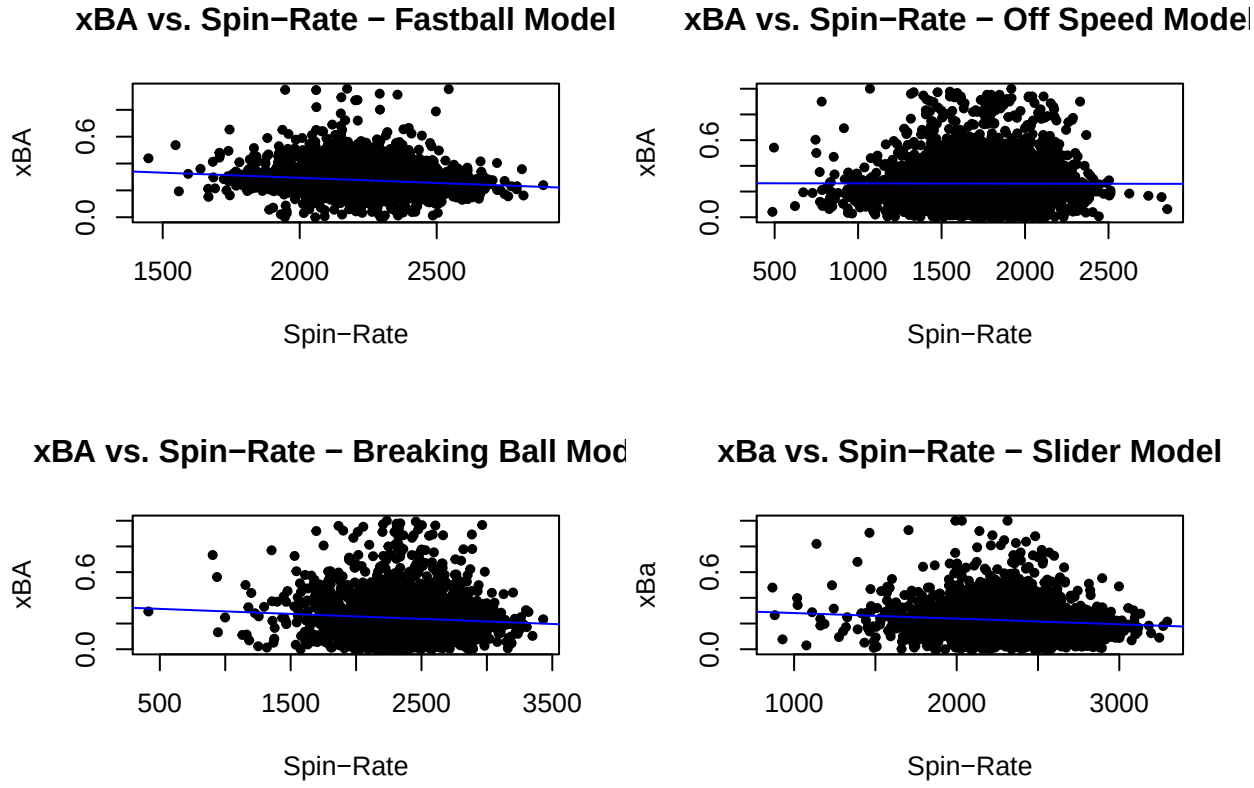
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.3253155	0.0162008	20.080270	0
spin_rate	-0.0000437	0.0000069	-6.325118	0

Looking at these estimate tables, we are able to see that three of four of our predictor spin-rates have a significant relationship with xBA, the three being: fastball, breaking ball, and slider. For all three of these pitches as we increase the spin-rate of the pitch our models show us that we can expect xBA for these pitches to decrease. Naturally, with a decrease in xBA, the pitcher can expect to have higher performance.

It is interesting to see these results as they agree with the physics of the pitch. For a fastball, with a higher spin-rate the pitch gives off more of an illusion that it is rising out of the pitchers hand as it approaches the batter. Higher spin-rates also generally lead to a faster pitch in terms of a fastball which also would correspond with a lower xBA. The results from our other two significant models also concur with the action of both sliders and breaking balls as these pitches often have tighter movement with higher spin-rates making these pitchers harder to hit. Lastly, the off-speed category was the only model that didn't show lower expected batting average with higher spin. This can be explained by the nature of the pitch. The off-speed pitch isn't reliant upon movement as much as it is supposed to throw off a hitter's timing. Due to the pitch's effectiveness not being entirely reliant upon movement but rather difference in speed, it is not surprising that there isn't a significant relationship between spin-rate on off-speed pitches and xBA.

The testing data was then fed into our models that we generated which produced the following visualizations as well as the MSE table to examine accuracy for each model. These visualizations will help represent the trends discussed above and can be seen below:

Fastball Model	Off Speed Model	Breaking Ball Model	Slider Model
0.0053542	0.0190941	0.0177976	0.0106821



Examining the above plots, we naturally are able to see agreement with the assumptions generated from our initial estimates. The three significant models generated a noticeable linear relationship with our

testing data while looking at the only non-significant predictor of xBA, off-speed, this line effectively has no slope. The MSE table agrees with our results as well as we can see the off-speed pitch category has the highest of all four models. This was expected as we have the most uncertainty of our xBA from the spin-rate of the off-speed pitch as spin-rate is less important than the other pitches for reasons explained earlier.

The import part of observing and obtaining these results is how we can apply these to the baseball world. This answer can be broken down into two parts. First, it serves as a metric where we can predict player performance based off these spin-rate tendencies. For example, if a GM is deciding on the type of contract to give a guy these models could be used as a predictive tool to measure what type of season the pitcher will have with their past spin-rates. This can also be applied to pitchers during the off-season who have developed a new pitch or different spin-rates on their pitches. With different spin-rates this pitcher is effectively different and we can use these models to predict how their performance will be effected.

Secondly, players can use this information, as a tool to develop pitches that meet an optimal spin-rate. If they are looking to better their performance with a certain pitch to a reasonable extent, these models can provide guidance when bettering their pitches. These models would also be useful in the R&D stage of a pitcher's off season when looking to develop a specific pitch. Tracking the spin-rate of their new pitch and using the information from the corresponding model can give them a reasonable idea of how that pitch will perform in game.

Spin-rate is able to tell us a lot about a pitch and pitcher just by simply tracking the revolutions of the ball. This idea will be investigated further in the next part as we try and take the applications of spin-rate a step further.

Part 2 - Attempting To Identify "At-Risk" Pitchers For Tommy John

Methodology

In comparison to the previous part, rather than using raw spin-rate, the spin-rate differences between a specific pitcher's pitches will be used. Pitchers are all unique and have different pitch arsenals which led to the decision of breaking them up into three categories: All Pitch Categories, Fastball | Off-Speed | Breaking Ball, and Fastball | Off-Speed | Slider which will be referenced later in the paper as All, F-O-B, and F-O-S respectively. These are the most common combinations for pitchers in the MLB which is why they were chosen. All other combinations were thrown out of the study. Possible further research would be to fit models for these other types of pitchers however, due to there uniqueness and the already present uniqueness of the surgery(which will be revealed later) it wasn't reasonable to fit a model for these other pitch combinations. An example of the pitch difference data frame used can be seen below:

fastoff	fastbreak	fastslid	tj
446	317	45	1
706	271	163	0
251	144	385	0
434	403	-17	0
446	-311	229	0
1135	390	302	0

This part of the analysis is focused on predicting pitchers who will need to get Tommy John Surgery based off these spin-rate differences discussed above. For this part of the predictive analysis, logistic regression was used to fit our three different models. This method was chosen due to the idea of attempting to "classify" pitchers who would possibly need Tommy John Surgery.

The results turn out to not be cut out to dry as hoped and further manipulation to our methods when trying to fit a model is carried out however, it is more appropriate to include these methods in the results section as they will be better understood in this context.

Results

Using the spin-rate difference tables generated for each pitch category, the following three logistic regression models were fitted. The estimates for each model can be seen below:

All Model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.2908894	0.3910571	-10.9725393	0.0000000
fastoff	0.0005201	0.0006091	0.8538625	0.3931812
fastbreak	0.0000793	0.0006903	0.1149327	0.9084985
fastslid	-0.0001273	0.0008291	-0.1535533	0.8779620

Looking at the model estimates for the All Model, we are able to decipher a few different observations made. First, it seems that as the difference of spin-rate between the fastball and off-speed increases, the pitcher becomes more likely to have a UCL injury. This notion is also the same for the difference of spin-rate between the fastball and breaking ball as the likelihood of an injury increases as well according to the estimates. Lastly, the pitcher seems to be more likely to have a UCL injury as the difference in spin-rate between the fastball and slider decreases meaning the spin of the two pitches becomes similar. The actual significance of these predictors will be touched on together after all three models have been discussed.

F-O-B Model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.6928912	0.3306090	-14.1946852	0.0000000
fastoff	0.0011483	0.0004849	2.3679398	0.0178874
fastbreak	-0.0001542	0.0004961	-0.3107167	0.7560160

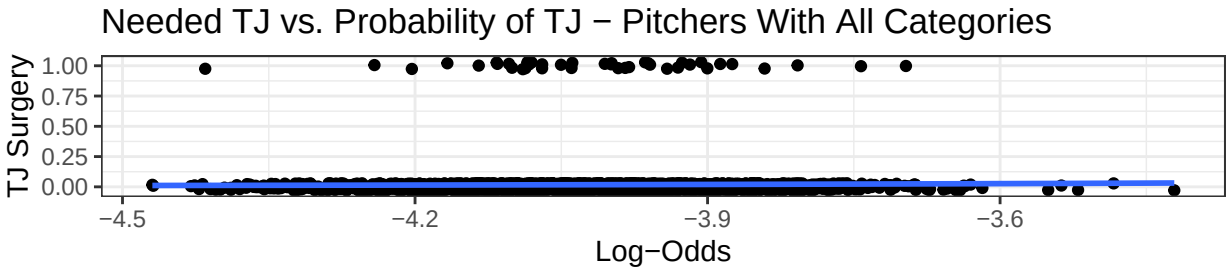
The F-O-B Model estimates are shown in the above table. Similar to the All Model, we can see that as the difference of spin-rate between fastball and off-speed increases also seems to lead to a higher probability that a pitcher will need Tommy John Surgery. In contrast to the All Model, the difference between fastball and breaking ball serves as a different relationship. For this particular model, it shows that the pitcher becomes more likely to need Tommy John Surgery as the spin-rate difference of these two pitches approach zero. Again, the significance of these predictors will be touched on later however, most likely hold an explanation on why there are different relationships between two pitches for the two different models examined so far.

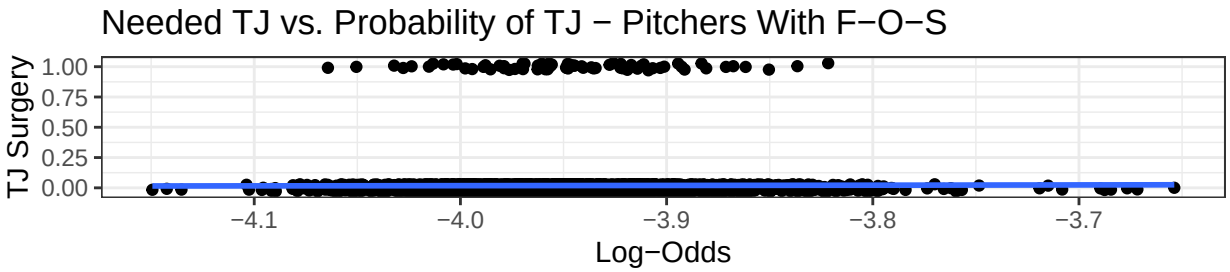
F-O-S Model:

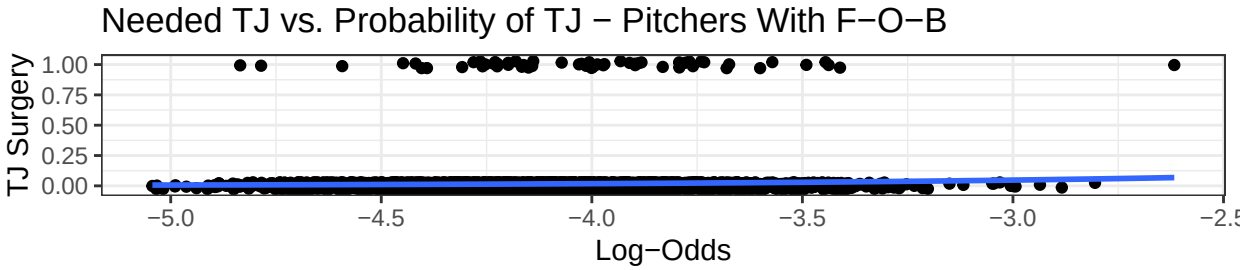
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.9944509	0.2484190	-16.0794916	0.0000000
fastoff	0.0001155	0.0004096	0.2820497	0.7779054
fastslid	0.0001422	0.0004772	0.2980136	0.7656928

Lastly, the F-O-S Model estimates are shown above. Again, like the previous two models, the estimates suggests that as the difference in spin of the fastball and off-speed increase, so does the probability for needing Tommy John Surgery. Looking at the next predictor, there is again an interesting trend as the estimate for the difference of spin between the fastball and slider is different than what was shown in the All Model. The F-O-S Model seems to show that as the difference between spin of these pitches increases so does the probability for Tommy John Surgery which was the opposite of the estimte found for the All Model. This again is most likely due to the significance of these predictors.

As foreshadowed in each model estimate discussion, there are drawbacks to the findings of each model. None of the three models reflected a significant relationship when attempting to predict Tommy John Surgery. Only one predictor variable out of all the variables included in the three models besides the intercepts was shown to be statistically significant. This is a problem as it would not be statistically correct to use these results when trying to predict whether a player is at risk for this type of injury. It is easy to see the problem with these three models after examining the graphical visualizations of the models which can be seen below:







These visualizations support the concerns generated from the estimate tables and it is clear that these models are not particularly relevant. The plots make it easy to see that despite the increase of occurrences in Tommy John Surgery, the proportion of pitchers who need this operation is still very small when compared to the rest of the pitchers who do not. In an effort to account for this, the amount of players who did not need the surgery was reduced. 10% of pitchers that did not need the surgery were randomly selected to keep in the data while all pitchers who needed the surgery were still included. The idea behind this is that our estimates won't be weighed down by the large amount of players not needing the surgery and will allow for focusing on characteristics of players who did need the surgery.

Using this new strategy for each pitch category, new models were fitted and the estimates can be seen below:

All Model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.9297772	0.4319360	-4.4677388	0.0000079
fastoff	0.0004025	0.0006686	0.6020060	0.5471701
fastbreak	0.0003822	0.0008024	0.4763769	0.6338059
fastslid	-0.0004110	0.0008754	-0.4694239	0.6387667

The estimates generated from the All Model of the mutated data can be observed above. These estimates are consistent in the relationships between the difference of each pitch however, the exact estimates changed slightly than what was found in the first fitting of the All Model. The different estimates was to be expected as the data itself fitting these estimates is different than initially used.

F-O-B Model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.4404508	0.3538281	-6.8972788	0.0000000
fastoff	0.0010909	0.0005233	2.0848548	0.0370825
fastbreak	-0.0003827	0.0005552	-0.6891967	0.4906995

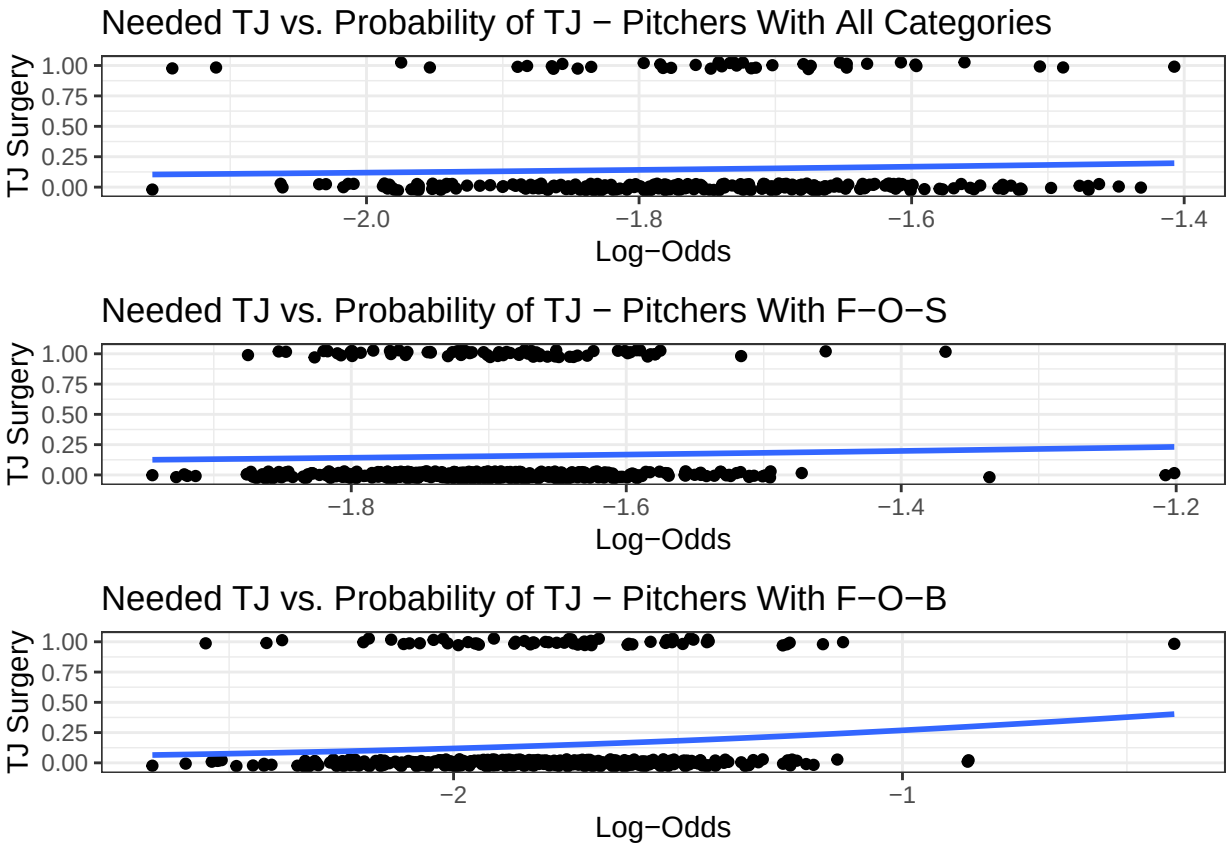
The new estimates generated for the F-O-B Model are displayed in the above table. The new estimates for this model as well follow the same trend as the new All Model as the estimates generated from the relationship between the variables is the same in terms of the sign of the slope while the estimates are slightly different. The significance of these new predictors will be touched on after all models have been examined.

F-O-S Model:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.6315124	0.2786290	-5.8555009	0.0000000
fastoff	-0.0000610	0.0004590	-0.1329153	0.8942604
fastslid	0.0003914	0.0005155	0.7593901	0.4476192

While the estimates were generally the same for the two new models discussed as their predecessors the F-O-S Model is different from the previous fit. The first new estimate seems to reflect the idea that a pitcher becomes more likely for Tommy John Surgery as the spin of their fastball and off-speed approach each other. The next estimate insinuates that as the pitcher is more likely to get Tommy John Surgery as the difference in spin of their fastball and slider increases. These difference in models and why they occur will be touched on in the paragraph following the visualizations.

To better understand the new model effects the following plots show our new data with the new predicted probability of Tommy John below:



Examining the new plots it does seem that we get a better idea of characteristics that lead to a UCL injury by using our data less populated by pitchers who didn't get injured. Despite this, the predictor

variables still generally all lacked a significant relationship with our response variable and all three models lacked a significant relationship as a whole. Even if we had found a significant relationship using this method of data manipulation is problematic because despite the non-injury pitchers being randomly selected, they are still being excluded knowingly while allowing all players who did get injured. Therefore, all applications of using these models should proceed with caution.

Using the results from the initial attempt to predict Tommy John Surgery and the revised attempt both point to the same idea that we can conclude that there isn't a significant relationship when predicting the need for Tommy John Surgery using only spin-rate difference between pitches. This doesn't mean that spin-rate shouldn't be used at all but just that we may need to examine it a different light. This will be touched on in the conclusion.

Conclusion

Spin-rate is a relatively new metric that holds a lot of potential in development and measuring of future players. This large amount of potential was the point of focusing this research on it in the first place. Despite the use of spin-rate not being able to effectively predict the potential for a UCL injury as shown in Part 2, the use of spin-rate should still be included in future studies for injury prevention. For example, an interesting possible extension of this research could be to measure the different levels of force put on the pitcher's elbow and the corresponding spin-rate of the pitch. This could potentially be more effective than just using spin-rate differences to identify at-risk pitchers for injury as it allows us to see the actual stress put on the UCL.

At the end of the day, while Part 2 proved to be relatively ineffective, the results in Part 1 proved to be beneficial as three significant models were generated from the research. For the three pitch categories of fastball, slider, and breaking ball, we learned that as we increase the spin-rate we can expect xBA to decrease. Therefore, as we increase the spin-rate we can expect the pitch to perform better. This part of the research would also benefit from further research such as potentially investigating the corresponding movement of the baseball on certain pitches with a particular spin-rate. This would provide further guidelines for pitchers to develop pitches to meet their specific desired movement amounts.

In conclusion, not every theory is going to lead to a direct breakthrough which can be seen in this paper itself. In Part 1 important relationships were identified while, in Part 2, the research failed to find any significant relationships. This is the nature of the scientific process and was understood when embarking on this project that there may not be any relevant relationships uncovered. These shortcomings of the results shouldn't discourage future research and should encourage future developments as there is so much to be learned when using spin-rate in order to predict certain outcomes in baseball.

References

Biomechanics: Ulnar collateral ligament. Biomechanics: Ulnar Collateral Ligament - Dec 18, 2008 - Blog - TexasLeaguers.com. (n.d.). Retrieved December 5, 2022, from https://texasleaguers.com/blog/2008/12/18/biomechanics-ulnar-collateral-ligament

Google. (n.d.). Tommy John Surgery List (@mlbplayeranalys). Google Sheets. Retrieved December 5, 2022, from https://docs.google.com/spreadsheets/d/1gQujXQQGOVNaiuwSN680Hq-FDVsCwvN-3AazykOBON0/edit#gid=0

Statcast Search. baseballsavant.com. (n.d.). Retrieved December 5, 2022, from https://baseballsavant.mlb.com/statcast_search

Appendix

Introduction

```
# Table Showing Pitch Category
fast <- c("4-Seam Fastball | 2-Seam Fastball | Sinker | Cutter")
off <- c("Changup | Split-finger")
breaki <- c("Curveball | Knuckle Curve | Slow Curve")
slid <- c("Slider")
cats <- data.frame(c(fast, off, breaki, slid))
colnames(cats) <- c("Pitches Included")
rownames(cats) <- c("Category: Fastball", "Category: Offspeed",
                    "Category Breaking Ball", "Category: Slider")
kable(cats)
```

Data

```
# Loading In Data
fast <- read.csv("C:/Users/tscot/Downloads/SA_Final_Proj/fastballspin.csv")
off <- read.csv("C:/Users/tscot/Downloads/SA_Final_Proj/offspeedspin.csv")
slid <- read.csv("C:/Users/tscot/Downloads/SA_Final_Proj/sliderspin.csv")
breaki <- read.csv("C:/Users/tscot/Downloads/SA_Final_Proj/breakingspin.csv")
xba <- read.csv("C:/Users/tscot/Downloads/SA_Final_Proj/xba.csv")
tj <- read_excel("C:/Users/tscot/Downloads/SA_Final_Proj/tj.xlsx")


# Data Cleaning/Combining
tj <- tj[tj$Year >= 2015,]

years <- c(2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022)
for (j in 1:length(years)) {
  year <- fast[fast$year == years[j],]
  for (i in 1:nrow(year)) {

  o <- off[off$player_name == year$player_name[i] & off$year == years[j],]
  s <- slid[slid$player_name == year$player_name[i] & slid$year == years[j],]
  b <- breaki[breaki$player_name == year$player_name[i] & breaki$year == years[j],]
  x <- xba[xba$player_name == year$player_name[i] & xba$year == years[j],]

    if (nrow(o) == 0) {
      o <- data.frame(spin_rate = 0)
    }
    if (nrow(s) == 0) {
      s <- data.frame(spin_rate = 0)
    }
    if (nrow(b) == 0) {
      b <- data.frame(spin_rate = 0)
    }
    if (nrow(x) == 0) {
      x <- data.frame(xba = 0)
```

```
    }
    if (i == 1 & j == 1) {
      final <- data.frame(year$player_name[i], year$spin_rate[i], o$spin_rate,
                          s$spin_rate, b$spin_rate, years[j], x$xba)
    }
    else {
      entry <- c(year$player_name[i], year$spin_rate[i], o$spin_rate, s$spin_rate,
                 b$spin_rate, years[j], x$xba)
      final <- rbind(final, entry)
    }
  }
}


# Data Dictionary
dic.vars <- c("player_name", "fast_spin", "off_spin", "slid_spin", "break_spin",
              "year", "xba", "tj")
dic.type <- c("Character", "Integer", "Integer", "Integer", "Integer", "Integer",
              "Numeric", "Integer (1 or 0)")
dic.expl <- c("Name of Pitcher", "Average Fastball RPM", "Average Offspeed RPM",
              "Average Slider RPM", "Average Breaking Ball RPM", "Year",
              "Expected Batting Average Against", "TJ Needed | 1-Yes, 0-No")
dic.frame <- data.frame(dic.vars, dic.type, dic.expl)
colnames(dic.frame) <- c("Variables", "Type", "Explanation")
kable(dic.frame)

ftj <- c()
# Head of Final Frame
for (i in 1:nrow(final)) {
  name <- strsplit(final$year.player_name.i.[i], split = ", ")
  name <- paste(name[[1]][2], name[[1]][1])
  tommy <- 0
  for (a in 1:nrow(tj)) {
    if (tj$Player[a] == name & tj$Year[a] == final$years.j.[i]) {
      tommy <- 1
    }
  }
  ftj <- c(ftj, tommy)
}
final$tj <- ftj

# Frame Manipulations
colnames(final) <- dic.vars
final$fast_spin <- as.integer(final$fast_spin)
final$off_spin <- as.integer(final$off_spin)
final$slid_spin <- as.integer(final$slid_spin)
final$break_spin <- as.integer(final$break_spin)
final$xba <- as.numeric(final$xba)
final$tj <- as.integer(final$tj)

final[is.na(final$off_spin),] <- 0
final[is.na(final$slid_spin),] <- 0
final[is.na(final$break_spin),] <- 0
```

```
final <- final[-3,]
kable(head(final))
```

Part 1

Results - Data Cleaning and Training Code

```
# xBA SLR Training

## Data Splitting
fast <- fast[!is.na(fast$xba),]
set.seed(0)
sample <- sample(c(T, F), nrow(fast), replace=TRUE, prob=c(0.7,0.3))
fast.train <- fast[sample,]
fast.test <- fast[-sample,]

## Model Fitting - Fast
fastmod <- lm(xba ~ spin_rate, data = fast.train)


## Data Splitting
off <- off[!is.na(off$xba) & !is.na(off$spin_rate),]
set.seed(0)
sample <- sample(c(T, F), nrow(off), replace=TRUE, prob=c(0.7,0.3))
off.train <- off[sample,]
off.test <- off[-sample,]

## Model Fitting - Off
offmod <- lm(xba ~ spin_rate, data = off.train)

## Data Splitting
breaki <- breaki[!is.na(breaki$xba) & !is.na(breaki$spin_rate),]
set.seed(0)
sample <- sample(c(T, F), nrow(breaki), replace=TRUE, prob=c(0.7,0.3))
breaki.train <- breaki[sample,]
breaki.test <- breaki[-sample,]

## Model Fitting - Breaking
breakmod <- lm(xba ~ spin_rate, data = breaki.train)

## Data Splitting
slid <- slid[!is.na(slid$xba) & !is.na(slid$spin_rate),]
set.seed(0)
sample <- sample(c(T, F), nrow(fast), replace=TRUE, prob=c(0.7,0.3))
slid.train <- slid[sample,]
slid.test <- slid[-sample,]
## Model Fitting - Slider
slidmod <- lm(xba ~ spin_rate, data = slid.train)

# Estimate Table Output
kable(summary(fastmod)$coefficients)
```

```
kable(summary(offmod)$coefficients)
kable(summary(breakmod)$coefficients)
kable(summary(slidmod)$coefficients)
```

Results - Graph COde

```
# xBA SLR Testing/Graph
## MSE Calculations
fast.mse <- mean((predict(fastmod, newdata = fast.test) - fast.test$xba)**2)
off.mse <- mean((predict(offmod, newdata = off.test) - off.test$xba)**2)
breaki.mse <- mean((predict(breakmod, newdata = breaki.test) - breaki.test$xba)**2)
slid.mse <- mean((predict(slidmod, newdata = slid.test) - slid.test$xba)**2)

mse.frame <- data.frame(fast.mse, off.mse, breaki.mse, slid.mse)
colnames(mse.frame) <- c("Fastball Model", "Off Speed Model", "Breaking Ball Model",
                         "Slider Model")
kable(mse.frame)

## Plots
par(mfrow = c(2,2))
### Fast
plot(xba ~ spin_rate, data = fast.test, pch = 20, xlab = "Spin-Rate",
     ylab = "xBA", main = "xBA vs. Spin-Rate - Fastball Model")
abline(fastmod$coefficients[1], fastmod$coefficients[2], col = 'blue')

### Off
plot(xba ~ spin_rate, data = off.test, pch = 20, xlab = "Spin-Rate",
     ylab = "xBA", main = "xBA vs. Spin-Rate - Off Speed Model")
abline(offmod$coefficients[1], offmod$coefficients[2], col = "blue")

### Break
plot(xba ~ spin_rate, data = breaki.test, pch = 20, xlab = "Spin-Rate",
     ylab = "xBA", main = "xBA vs. Spin-Rate - Breaking Ball Model")
abline(breakmod$coefficients[1], breakmod$coefficients[2], col = "blue")

### Slid
plot(xba ~ spin_rate, data = slid.test, pch = 20, xlab = "Spin-Rate",
     ylab = "xBa", main = "xBa vs. Spin-Rate - Slider Model")
abline(slidmod$coefficients[1], slidmod$coefficients[2], col = "blue")
```

Part 2

Results - Initial Model Code

```
# TJ Prediction

# All
## Manipulating Frame to All
final.all <- final[which(final$fast_spin > 0),]
```

```r
final.all <- final.all[which(final.all$off_spin > 0),]
final.all <- final.all[which(final.all$break_spin > 0),]
final.all <- final.all[which(final.all$slid_spin > 0),]

## Creating Frame of Differences
final.all <- data.frame(final.all$fast_spin-final.all$off_spin,
                        final.all$fast_spin-final.all$break_spin,
                        final.all$fast_spin-final.all$slid_spin,  final.all$tj)
colnames(final.all) <- c("fastoff", "fastbreak", "fastslid", "tj")


# Model Fitting With Plot
logmod.all <- glm(tj ~ ., data = final.all, family = binomial)
all <- ggplot(final.all, aes(x = predict(logmod.all, type = "link"), y = tj)) +
  geom_point(position = position_jitter(height = 0.03, width = 0)) +
  geom_smooth(method = "glm", method.args = list(family = binomial), se = F) +
  theme_bw() +
  labs(y = "TJ Surgery",
       title = "Needed TJ vs. Probability of TJ - Pitchers With All Categories",
       x = "Log-Odds")

# Fast-Off-Slid
## Manipulating Frame To Fast-Off-Slid
final.fos <- final[which(final$fast_spin > 0),]
final.fos <- final.fos[which(final.fos$off_spin > 0),]
final.fos <- final.fos[which(final.fos$slid_spin > 0),]

## Creating Frame of Differences
final.fos <- data.frame(final.fos$fast_spin-final.fos$off_spin,
                        final.fos$fast_spin-final.fos$slid_spin,
                        final.fos$tj)
colnames(final.fos) <- c("fastoff", "fastslid", "tj")

# Model Fitting and Plot
logmod.fos <- glm(tj ~ ., data = final.fos, family = binomial)
fos <- ggplot(final.fos, aes(x = predict(logmod.fos, type = "link"), y = tj)) +
  geom_point(position = position_jitter(height = 0.03, width = 0)) +
  geom_smooth(method = "glm", method.args = list(family = binomial), se = F) +
  theme_bw() +
  labs(y = "TJ Surgery",
       title = "Needed TJ vs. Probability of TJ - Pitchers With F-O-S",
       x = "Log-Odds")

# Fast-Off-Break
## Manipulation Frame To Fast-Off-Break
final.fob <- final[which(final$fast_spin > 0),]
final.fob <- final.fob[which(final.fob$off_spin > 0),]
final.fob <- final.fob[which(final.fob$break_spin > 0),]

## Creating Frame of Differences
final.fob <- data.frame(final.fob$fast_spin-final.fob$off_spin,
                        final.fob$fast_spin-final.fob$break_spin,
                        final.fob$tj)
```

```
colnames(final.fob) <- c("fastoff", "fastbreak", "tj")


# Model Fitting and Plot
logmod.fob <- glm(tj ~ ., data = final.fob, family = binomial)
fob <- ggplot(final.fob, aes(x = predict(logmod.fob, type = "link"), y = tj)) +
  geom_point(position = position_jitter(height = 0.03, width = 0)) +
  geom_smooth(method = "glm", method.args = list(family = binomial), se = F) +
  theme_bw() +
  labs(y = "TJ Surgery",
       title = "Needed TJ vs. Probability of TJ - Pitchers With F-O-B",
       x = "Log-Odds")

# Example Table Output
kable(head(final.all))

# Estimate Tables Output
kable(summary(logmod.all)$coefficients)
kable(summary(logmod.fob)$coefficients)
kable(summary(logmod.fos)$coefficients)

# Plotting Data
grid.arrange(all, fos, fob)
```

Results - Revised Model Code

```
# Frame Reduction All
tjy <- which(final.all$tj == 1)
tjn <- which(final.all$tj == 0)
set.seed(0)
sample <- sample(c(T, F), length(tjn), replace=TRUE, prob=c(0.1,0.9))
tjn <- tjn[sample]
final.all <- final.all[c(tjy,tjn),]



# Frame Reduction F-O-B
tjy <- which(final.fob$tj == 1)
tjn <- which(final.fob$tj == 0)
set.seed(0)
sample <- sample(c(T, F), length(tjn), replace=TRUE, prob=c(0.1,0.9))
tjn <- tjn[sample]
final.fob <- final.fob[c(tjy,tjn),]

# Frame Reduction F-O-S
tjy <- which(final.fos$tj == 1)
tjn <- which(final.fos$tj == 0)
set.seed(0)
sample <- sample(c(T, F), length(tjn), replace=TRUE, prob=c(0.1,0.9))
tjn <- tjn[sample]
final.fos <- final.fos[c(tjy,tjn),]
```

```r
# All
logmod.all <- glm(tj ~ ., data = final.all, family = binomial)
all <- ggplot(final.all, aes(x = predict(logmod.all, type = "link"), y = tj)) +
  geom_point(position = position_jitter(height = 0.03, width = 0)) +
  geom_smooth(method = "glm", method.args = list(family = binomial), se = F) +
  theme_bw() +
  labs(y = "TJ Surgery",
       title = "Needed TJ vs. Probability of TJ - Pitchers With All Categories",
       x = "Log-Odds")

# Fast-Off-Slid
logmod.fos <- glm(tj ~ ., data = final.fos, family = binomial)
fos <- ggplot(final.fos, aes(x = predict(logmod.fos, type = "link"), y = tj)) +
  geom_point(position = position_jitter(height = 0.03, width = 0)) +
  geom_smooth(method = "glm", method.args = list(family = binomial), se = F) +
  theme_bw() +
  labs(y = "TJ Surgery",
       title = "Needed TJ vs. Probability of TJ - Pitchers With F-O-S",
       x = "Log-Odds")

# Fast-Off-Break
logmod.fob <- glm(tj ~ ., data = final.fob, family = binomial)
fob <- ggplot(final.fob, aes(x = predict(logmod.fob, type = "link"), y = tj)) +
  geom_point(position = position_jitter(height = 0.03, width = 0)) +
  geom_smooth(method = "glm", method.args = list(family = binomial), se = F) +
  theme_bw() +
  labs(y = "TJ Surgery",
       title = "Needed TJ vs. Probability of TJ - Pitchers With F-O-B",
       x = "Log-Odds")

# Estimate Table Output
kable(summary(logmod.all)$coefficients)
kable(summary(logmod.fob)$coefficients)
kable(summary(logmod.fos)$coefficients)


# Plotting Output
grid.arrange(all, fos, fob)
```