Nicholas Chapman

MAT 8406 Final Project

Dr. Zhang

05/02/2020

Pitching Injuries in Major League Baseball

Abstract: Since 2000, there has been a rise in the number of Tommy John surgeries performed. This rise has been attributed to many factors, including the rise of off-speed pitches, the ever-increasing speed of fastballs, and the volume of pitches required of starting pitchers in Major League Baseball. I gathered data from Baseball Info Solutions Data, John Roegle’s Tommy John database, and the injured list for 2015-2018 from the Spotrac website. In this analysis, I attempted to use Logistic Regression to model the relationship of these qualities and injury in the following season using stepwise logistic regression based on AIC and Likelihood Ratio test to optimize exploratory results. The results found positive associations between the percentage of off-speed pitches and injury the following season, a positive association between a prior injury and injury the next season, and finally an association between pitcher performance and injury the next season.

1. **Introduction**

According to a 2015 article by Forbes, Major League Baseball teams lost an estimated 1.1 billion dollars due to pitchers being on the injured list[[1]](#footnote-1). One of the most pervasive and costly injuries are to the pitcher’s throwing elbow, which often requires a specialized surgery known as the Tommy John Surgery. Tommy John surgery, whose name comes from the first player to receive the surgery, is a reconstructive surgery to replace the torn ulnar collateral ligament in the elbow with tendons taken from the knee or arm. Figure 1 plots the number of documented Tommy John surgeries for Major League players over time[[2]](#footnote-2). Since 2000, there has been a significant increase in surgeries performed and there are numerous explanations proposed for this. One such explanation is that there has been a rise in the surgery as an elective measure to improve performance, a practice which has garnered much criticism, especially in the COVID crisis[[3]](#footnote-3). Other explanations include higher emphasis on off-speed pitches, increased fastball velocities, and sheer volume. Similarly, Will Carroll, a sportswriter who specializes in the coverage of medical issues, makes the argument that a fatigued pitcher is more likely to get injured because of a breakdown in mechanics.

I wanted to evaluate these assertions by applying a logistic regression to these data and see if there was a significant relationship between pitch types and volumes, and an injury to the pitcher. Specifically, I wanted to test three hypotheses:

1. Whether there is a relationship between the percentage of off-speed pitches (sliders and cutters) and injury the next season.

2. Whether there is a relationship between a prior injury and injury the next season.

3. Whether there is a relationship between decreased pitcher performance and injury the next season.

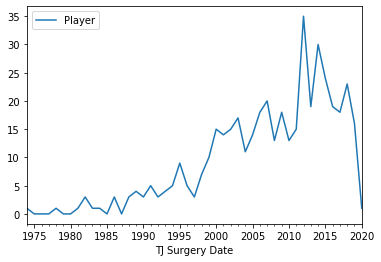


Fig 1: From J. Roegele’s Tommy John Spreadsheet.

1. **Method**

*2.1 Data Gathering & Cleaning*

Baseball Info Solutions Data (BIS), includes aggregate information by season since 2013 for each pitcher, even if they pitched in a single game. This includes information on the number of pitches thrown, innings pitches, games played, and types of pitches thrown. For each type of pitch, this data includes the proportion, average velocity, horizontal and vertical movements, and finally the number of runs above average that pitch type allowed. The Disabled List (DL) from Spotrac includes the player name, season, injury location, and how long the player was inactive. The DL contains a field labelling the location of the injury. In some instances, the player was added to the list because of illness. For purposes of this analysis, I only included those injuries to body parts that are involved in the kinetic chain of pitching[[4]](#footnote-4). This data could only be located for the 2015-2019 Seasons. Lastly, I included the Tommy John Surgery List maintained by John Roegle, which includes the player and year in which the surgery was performed.

The first step in cleaning the data was in generating a field to represent the dependent variable Y. In order to fit a logistic regression, the data must be fit to a binary dependent variable, whose values take on zero or one. Additionally, I needed to fit the dependent variable with the previous year’s data. I created a binary variable, grouped the data by pitcher, then lagged the variable one season. My dependent variable then became: (0: Pitcher does not get injured next season | 1: pitcher gets injured next season).

Not all pitchers are used equally. As baseball has evolved, so have the strategies involving pitches that are used in specialized situations or are pulled up from the minor leagues to cover an injured player. This analysis will cover players and their seasons where there were at least 10 innings pitched. Also, because I create variables based on previous seasons, I will only cover players who have more than one season. This way many of the variables will be populated.

* 1. *Variable Creation*

Because wear on a pitcher cannot be represented in a single season, I wanted to get a sense of the player’s career leading up to the season in question. Using the Python Programming language, I generated accelerations of each of the pitch types and their subcategories. For instance, for the fastball percentage variable, I created a new variable, fastball percentage acceleration, to represent the change over time between the fastball percentage 2 seasons ago, to the fastball percentage in the current season. An example is shown in Table 1. A player who didn’t have seasons available for this calculation will have zero for this value, indicating no change over time. This same process was done for other pitch types, Innings, Pitches, Games, Games Started, wild pitch rate, and walk rate.

|  |  |  |
| --- | --- | --- |
| **Season** | **FA% (pi)** | **FA% (pi)\_accel\_2** |
| 2013 | 0.606 | -0.015667 |
| 2014 | 0.553 | -0.022667 |
| 2015 | 0.538 | -0.022667 |
| 2016 | 0.508 | -0.015 |
| 2017 | 0.466 | -0.024 |
| 2018 | 0.41 | -0.032667 |
| 2019 | 0.438 | -0.009333 |

Table 1: Clayton Kershaw Fastball Percentage 2013-2019

For the first hypothesis, it was necessary to create these different metrics around the pitches to get more information on how they used the off-speed pitches this season, and how that compares to previous seasons. The fastball proportion is negatively correlated with the other pitch types, and it will be shown later that only one of these is necessary to include in the model. For the second hypothesis, it meant lagging the dependent variable and creating a binary variable (1: Injured Prior to this season | 0 : not injured prior to this season). Finally, to test the third hypothesis, I use walk rate, which can be a measure of how well a pitcher is performing because the more he walks, the less strikes he throws. Also, for the fastball pitch type and the off-speed pitch type, there exists a variable measuring the runs above average. Win-Loss rate is also considered here, but by definition is an incomplete variable as a win has to be awarded by decision, a process governed by various rules[[5]](#footnote-5). Additionally, other columns that did not exist in the data were created, but require no deeper explanation. These include: Injury prior to season, walk rate, pitches per innings pitched, and pitchers per batters faced.

After the variable creation step in Python, the data was loaded into R, and some interaction terms were explored. These explorations took on two forms: Domain knowledge and testing for significance from a pool interaction terms using forward selection. The first interaction term that was created was between fastball percentage and fastball velocity. A pitcher’s majority pitch may be mostly fastballs, but it may not be damaging until he is throwing upwards of 100 mph on average. Because the relationship of fastball percentage to injury may be reliant upon fastball velocity, this interaction term is included. Secondly, because a prior injured pitcher may be used in the season differently as it relates to playing time, the injury variables, prior injury and tj (whether the pitcher had received the tommy john surgery), were tested for significance as an interaction term with other variables.

* 1. *Model Initialization*

The variable we are trying to predict is a binary variable. To model these data, a Logistic Regression model was fit in an attempt to test the stated hypotheses. The variables used in this initial model were selected due to a combination of domain knowledge, literature review (reading articles), and for variable testing.

* 1. *Variable Selection*

Multicollinearity in logistic regression can heavily impact the regression coefficients. Before I added any interaction terms, I checked for high collinearity between my predictors. A subset of the correlation matrix is shown in Table 2, which reveals high collinearity between the fastball and off-speed pitches, and a high correlation between innings pitched and pitches. It is obvious that the more fastballs are thrown, the less off-speed pitches are thrown, so I dropped the off-speed proportion variables from consideration. Similarly, innings pitched and pitches thrown are positively correlated, and it is not necessary to have both in the model, and the pitches variable was dropped.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Pitches** | **IP** | **Accel in innings (2 years)** | **Accel in pitches (2 years)** | **FA**  **%** | **Sinkers & Cutters**  **%** |
| **Pitches** | 1 |  |  |  |  |  |
| **IP** | 0.99 | 1 |  |  |  |  |
| **Accel in innings**  **(2 years)**  **%** | 0.41 | 0.42 | 1 |  |  |  |
| **Accel in pitches**  **(2 years)**  **%** | 0.42 | 0.42 | 0.99 | 1 |  |  |
| **FA**  **%** | -0.08 | -0. | 0.04 | 0.04 | 1 |  |
| **Sinkers & Cutters**  **%** | 0.04 | 0.07 | -0.02 | -0.02 | -0.85 | 1 |

Table 2: Subset of the correlation matrix

After removing variables with possible collinearity, the stepwise regression method was used to remove and add variables to the initial model. There were two selection criteria considered: AIC and the Likelihood Ratio Test (LRT). AIC was a good option because the number of variables was high, and the possibility of overfitting greater. Because AIC strikes a balance between Goodness of Fit and model complexity, (penalizing models with more variables), it was a natural candidate. Conversely, the LRT assesses the significance of a difference in fit between two models. At every iteration in the stepwise regression, two models are compared, one with and without the candidate variable. Their log likelihoods are compared and if the difference is significant, the model with more variables is selected. In other words, a candidate variable will not be submitted into the model if it does not provide significant predictive power. An added benefit of these methods is, as candidate variables are evaluated, a natural aversion to multicollinearity may occur. E.g. If A is in the model, and B is collinear with A, B may provide no difference in log likelihood (LRT), or may be penalized with no increased Goodness of Fit (AIC). Variables were admitted or dropped if both metrics agreed. A full breakdown of the stepwise regression can be found in the appendix.

* 1. *Model Diagnostics*

Logistic Regression has fewer assumptions than the General Linear model. The assumptions of Logistic Regression include independence in the observations, and requires that the independent variables are not linear combinations of each other. It is also necessary to check for any influential observations that may have an unwanted effect on the regression coefficients. While determining the independence of the observations may be a difficult task, we can apply some techniques to determine if the model satisfies the other two conditions.

First to determine if the explanatory variables are not linear combinations of themselves, there is a technique, based on multiple linear regression, where the variables that are continuous are fit to a randomly generated continuous variable[[6]](#footnote-6). Then the VIF’s are calculated as if it were a GLM. Moreover, because including interaction terms can sometimes cause multicollinearity, it is necessary to transform the variables to reduce this. Results for the variables where the VIF changed after transformation are displayed in Table 3. Based on a general cutoff of 10, there does not seem to be evidence to suspect that there any of the variables are linear combinations of themselves and, for now, do not reject based on these assumptions of non-linearity.

|  |  |  |
| --- | --- | --- |
|  | VIF | VIF (Transformation) |
| P\_per\_IP:IP\_accel\_2 | 203.7 | 1.1 |
| IP\_accel\_2 | 202.8 | 1.2 |
| walk\_rate:IP\_per\_G | 11.3 | 1.2 |
| IP\_per\_G | 10.7 | 1.2 |
| walk\_rate | 3.2 | 1.9 |
| P\_per\_IP:IP\_accel\_2 | 203.7 | 1.1 |

Table 3: VIF values for model chosen by stepwise regression.

After checking for multicollinearity among the predictor variables, I checked the data for any highly influential observations. Unusual observations could lead to some unwanted effects on our regression coefficients. In this study, four statistics are considered when testing for influential observations: 1. Pearson Residuals, which measures the difference between the observed frequency and the predicted frequency. 2. Deviance residuals which, like raw residuals in OLS, measures the disagreement between the observed and fitted log likelihood functions. 3. Leverage, or what is sometimes referred to as Pregibon leverage. 4. DFBetas, or the change in the jth observation coefficient when the ith observation is held from the fit. Figure 2 shows 5 potentially influential observations. From this plot we can see 5 observations that rank highly due to high Standardized Pearson Residuals, and abnormal leverage.

Similarly, in Figure 3, we can see first that the injured players have a larger effect on the coefficient, but also two players who have notable influence on the beta coefficient. Once again, Matt Harrison’s 2014 season has flagged for one of the influence statistics. Further investigation into these observations shows that Adam Wainwright injured his Achilles in 2015[[7]](#footnote-7) while batting and saw a steep reduction in the innings he pitched and because he isn’t coded as a pitching injury, he has influence on this coefficient. Matt Harrison, in a similar fashion, underwent back surgery in 2013, and represents a data entry error in the disabled list. Sources say he was on the list three years running from 2013-2015[[8]](#footnote-8). While researching one of the outliers in Figure 2, it came to my attention that types of pitchers were not being recognized. This led to the creation of a new variable, Innings pitched per game, which should control for pitchers who start games, and pitchers who just come in for an inning to close it out. Since these latter pitchers are expected to pitch every night, the sheer volume of pitches did not account for this.



Figure 2: Plotting Standardized Pearson Residuals Against Leverage (Red: Injured; Blue: Not Injured)



Figure 3: Plot of DFBetas for P\_per\_IP \* IP\_accel\_2

The DFBetas did reveal a data gap in the plot for the prior injury coefficient. Figure 4a shows the impact of the later seasons on the prior injury variable. While very predictive, the data was missing for the first two seasons of the data, and needed to be filled. To overcome this obstacle, it required some data scraping and regex matching from the transactional data the MLB provides every day during the Season. Figure 4b shows this coefficient after the change.



Figure 4a: DFBetas for Prior Injury. Index is sorted by Season.



Figure 4b: DFBetas for Prior Injury after the fix. Index is sorted by Season.

1. **Results**

The final model can be seen in the appendix. With the final model built and the assumptions checked, the hypotheses could now be tested and their coefficients interpreted. First, in testing where there is a relationship between off-speed pitches thrown and injury the following season. Here, many variables could have been used to represent this relationship, but not many found their way into the model. Here we partially consider this hypothesis with the interaction term between fastball percentage and the acceleration of pitches per inning over a three-year span:

For a 10% increase in fastball percentage, there is an associated decrease in the odds of injury next season by 5.97% while holding all of the other variables constant. This is as I expected, as off-speed pitches can be more harmful to a pitcher’s health as it sometimes requires greater contortion of the arm in the throwing motion.

For the second hypothesis, we want to test whether there is a relationship between a prior injury and injury next season. A prior injury variable, injury in the last two years, was admitted to the model by all three metrics in the form of an interaction term with wSI..pi.\_accel\_2, which measures the acceleration in runs above average for the slider pitch over a three season period. This variable essentially measures the pitcher’s performance over time. We would suspect that there would be a decline in performance immediately prior to injury and after injury, whereas in the two years after injury, the performance might then increase. This interaction term is then interpreted as:

There is an associated increase in the odds of injury next season by 86.6% when there is a prior injury in the two preceding years, and while holding all other variables constant.

Finally, to test the third hypothesis, we want to test the relationship between decreased pitcher performance and likelihood of injury the next season. Here we can use the interaction term wFA..pi.\_accel\_2:wSI..pi.\_accel\_2 which was admitted to the model by all three metrics. As stated above, these two variables measure the acceleration of the number of runs saved with their respective pitch type over the last 3 seasons. It can be interpreted as follows:

For an increase in the acceleration of runs saved with fastballs over a three-year period by 2 runs, there is an associated increase in the odds of injury next season of 5.5%, while holding all other variables constant. For an increase in the acceleration of runs saved by slider pitches by 2 runs for players who were previously injured in the last 2 years, there is an associated increase in the odds by 39%, while holding all other variables constant. For an increase in the acceleration of runs saved with slider pitches by 2 runs for players who don’t have a prior injury in the last 2 years, there is an associated decrease in the odds of injury next season by 4.7%, while holding all other variables constant. Therefore, the likelihood of injury given an increase in runs saved on the slider pitch over three years depends on whether they had a prior injury in the last two years, and contributes little to the likelihood when they didn’t. What is interesting about these results is that it seems an increase in pitcher effectiveness increases the likelihood of injury. This seems to be corroborated by the win record variable which, for every increase in win record by 10%, increases the odds of injury by 7.2%, while holding all other variables constant. This suggests that more successful players are more susceptible to injury.

1. **Discussion**

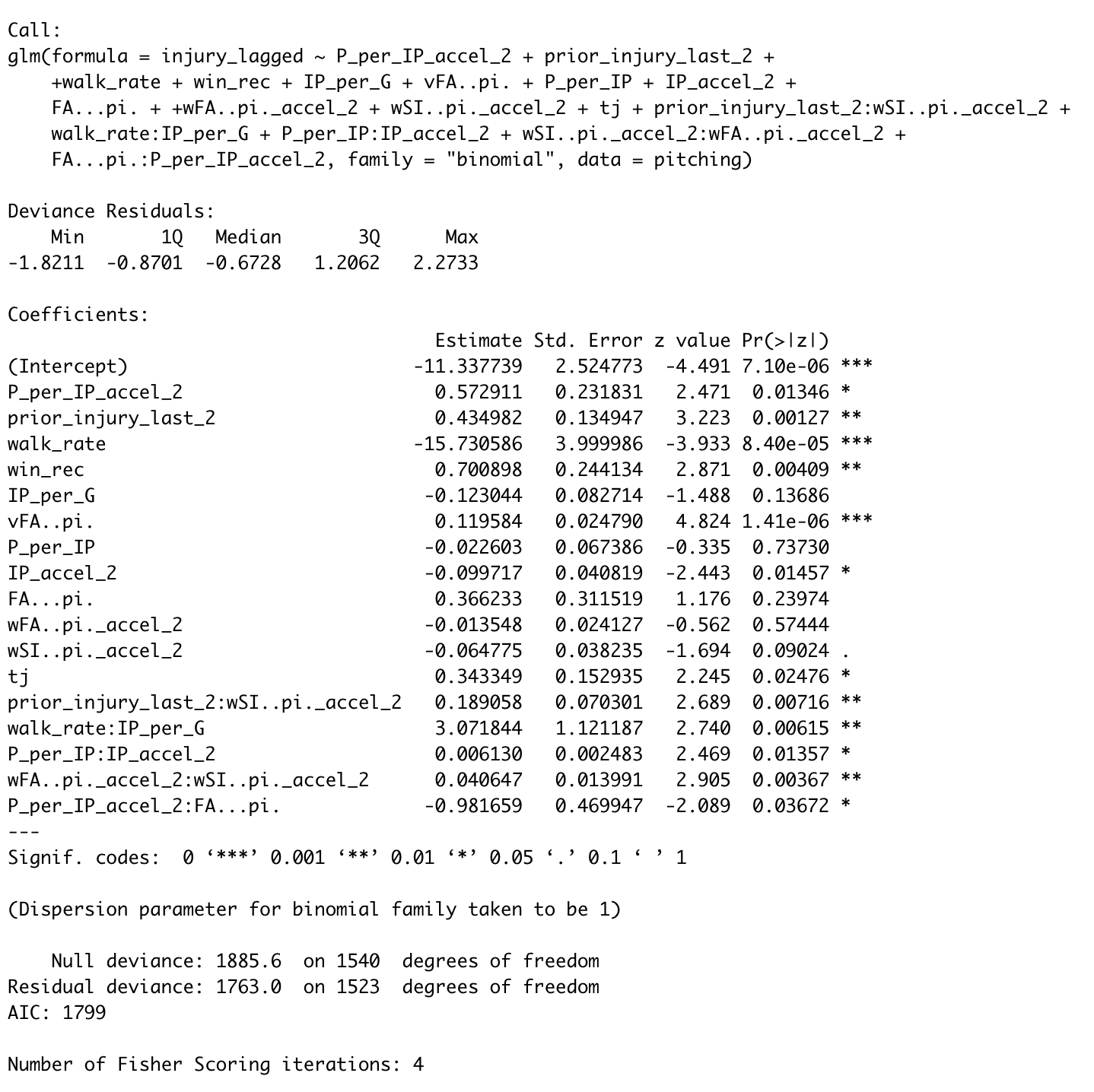
Before fitting the model, multicollinearity was investigated and variables with glaring issues were dropped. The VIF method discussed above cleared any more suspicious about deeply problematic issues related to multicollinearity. Influential observations were identified, and in some cases amended. I opted not to amend any observation unless it was because of a coding error.

Because it can be hard to arrive at a pleasant conclusion as it relates to Goodness-of-Fit in logistic regression, I assume that, in my stepwise regression, I arrive at a more appropriate model every time a variable is accepted or removed from the subsequent model. In the end, I was able to test my hypotheses directly, and in some instances indirectly, as in the case of the third.

For future work, it may be interesting to see how well this data could be used to create a predictive model. There, fitting toward a different metric, such as AUC, might be more appropriate as we would be more concerned with how well the model is separating the two classes, rather than outputting interpretable coefficients.

**Appendix**

**FINAL MODEL**



**STEPWISE REGRESSION**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **model** | **AIC** | **LRT** | **Var Considered** | **direction** |
| reg.2 |  |  |  |  |
| reg.3 | DROP | DROP | FA.Z..pi. | backward |
| reg.4 | DROP | DROP | IP\_accel\_2 | backward |
| reg.5 | DROP | DROP | Pitches | backward |
| reg.6 | DROP | DROP | vFC..pi.\_accel\_2 | backward |
| reg.7 | DROP | DROP | wFA..pi.\_accel\_2 | backward |
| reg.8 | DROP | DROP | Age | backward |
| reg.9 | DROP | DROP | P\_per\_IP | backward |
| reg.10 | DROP | DROP | FA...pi. | backward |
| reg.11 | DROP | DROP | prior\_injury | backward |
| reg.12 | DROP | DROP | wSI..pi.\_accel\_2 | backward |
| reg.13 | DROP | DROP | wild\_pitch\_rate\_accel\_2 | backward |
| reg.14 | DROP | DROP | SI.Z..pi. | backward |
| reg.15 | ADMIT | ADMIT | walk\_rate:IP\_per\_G | forward |
| reg.16 | ADMIT | DENY | SI.Z..pi. | forward |
| reg.17 | ADMIT | DENY | P\_per\_IP\_accel\_2:IP\_per\_G | forward |
| reg.18 | ADMIT | ADMIT | P\_per\_IP:IP\_accel\_2 | forward |
| reg.19 | ADMIT | DENY | IP\_accel\_2:vFA..pi. | forward |
| reg.20 | ADMIT | DENY | P\_per\_IP\_accel\_2:walk\_rate | forward |
| reg.21 | ADMIT | DENY | prior\_injury\_last\_2:IP\_accel\_2 | forward |
| reg.22 | DENY | DENY | FA...pi.:FA.Z..pi. | forward |
| reg.23 | ADMIT | DENY | P\_per\_IP:wFA..pi.\_accel\_2 | forward |
| reg.24 | ADMIT | ADMIT | wSI..pi.\_accel\_2:wFA..pi.\_accel\_2 | forward |
| reg.25 | ADMIT | ADMIT | prior\_injury\_last\_2:wSI..pi.\_accel\_2 | forward |
| reg.26 | ADMIT | ADMIT | FA...pi.:P\_per\_IP\_accel\_2 | forward |

**VIF TABLE**

|  |  |  |
| --- | --- | --- |
|  | VIF Transformed | VIF |
| P\_per\_IP:wFA..pi.\_accel\_2 | 217.6 | 1.1 |
| wFA..pi.\_accel\_2 | 217.4 | 1.1 |
| IP\_accel\_2 | 214.4 | 1.3 |
| P\_per\_IP:IP\_accel\_2 | 213.8 | 1.1 |
| FA...pi.:FA.Z..pi. | 27.2 | 1.2 |
| FA...pi. | 19.4 | 1.3 |
| FA.Z..pi. | 6.4 | 1.3 |
| P\_per\_IP | 2.1 | 2.1 |
| walk\_rate | 1.8 | 1.8 |
| Pitches | 1.4 | 1.4 |
| wild\_pitch\_rate\_accel\_2 | 1.3 | 1.0 |
| vFA..pi. | 1.2 | 1.2 |
| wSI..pi.\_accel\_2 | 1.1 | 1.1 |
| win\_rec | 1.1 | 1.1 |
| wSI..pi.\_accel\_2:wFA..pi.\_accel\_2 | 1.0 | 1.0 |

1. [Forbes.com](https://www.forbes.com/sites/howardcole/2015/03/09/baseball-loses-1-1-billion-to-pitching-injuries-over-five-year-period/#539be72c2972) [↑](#footnote-ref-1)
2. This list is maintained by John Roegle and can be found [here](https://docs.google.com/spreadsheets/d/1gQujXQQGOVNaiuwSN680Hq-FDVsCwvN-3AazykOBON0/edit). [↑](#footnote-ref-2)
3. Nytimes.com [↑](#footnote-ref-3)
4. The kinetic chain of the pitching motion is the functional motion by which a pitcher applies force to the ball to create velocity. It is explored in great detail by Will Carrol in his book, *Saving the Pitcher*. [↑](#footnote-ref-4)
5. A pitcher may not receive a win if: 1. They did not pitch prior to the last half-inning when the winning team took the lead. 2. They did not pitch at least 5 innings. [↑](#footnote-ref-5)
6. This technique is explained in more detail [here](https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-are-the-likelihood-ratio-wald-and-lagrange-multiplier-score-tests-different-andor-similar/). [↑](#footnote-ref-6)
7. <https://www.sbnation.com/mlb/2015/4/26/8499995/adam-wainwright-injury-update-achilles> [↑](#footnote-ref-7)
8. <https://www.wikiwand.com/en/Matt_Harrison_(baseball)> [↑](#footnote-ref-8)