The Effects of Cleveland Guardians' Starting Pitching on Game Result

Hunter Geise
SAL 213: Sport Data Analysis I
Dr. Losak
November 10, 2022

Introduction

The starting pitcher is a key piece for franchises taking crucial steps towards building a championship caliber team. He is called on a given night and expected to throw a minimum of five innings while holding opponents to as few runs possible. These guys usually have an amazing three to four different pitch mix consisting of a four-seam and/or two-seam fastball combined with an off-speed pitch like a curveball, slider, and/or change-up. The dominant pitchers have a fastball that they can paint the strike zone and blow past hitter alongside a reliable off-speed pitch that freeze the hitter in the box or get him to chase the pitch outside the strike zone.

The Cleveland Guardians have had an extremely historic franchise when it comes to their starting pitching. They have had dominant Hall of Famers in the past such as Cy Young, Bob Feller, and Gaylord Perry that carried teams to postseason appearances and even championships. More recently in the 2000's, the likes of CC Sabathia, Cliff Lee, Corey Kluber, and now Shane Bieber have won the Cy Young Award as best pitcher in the league while throwing gems at Progressive Field.

The current Guardians pitching staff is extremely young with an explosive trio of Shane Bieber, Triston McKenzie, and Cal Quantrill. Watching multiple games live in Cleveland over the past two summers has given amazing firsthand views of dominant performances from Quantrill carrying the club to a win. Other games in the 2022 season had McKenzie and Bieber pick up wins for pitching deep into the game while recording double digit strikeouts. These instances inspired the idea of analyzing the correlation between the starting pitcher's performance and the overall outcome of the game.

For the regression model, the dependent "Y" variable is the outcome of the game which is a win or loss. In the case of the Cleveland Guardians, they have been one of the more successful teams in recent MLB seasons, making the postseason in four out of the eight seasons in the data collected from 2014-2022 (did not include 2020 shortened COVID season). Using the filtered data set for the regression, they had an overall win percentage of 58.9%, which included early seasons of playing against division opponents like the Kansas City Royals and Detroit Tigers that were World Series caliber teams.

Model and Variables

The equation used to build the model was basic with the use multiple dummy variables alongside normal quantitative attributes. The equation for the model was

$$\begin{aligned}\text{NumberWinLoss} = {} & \alpha + \beta_1\text{Home} + \beta_2\text{ALEast} + \beta_3\text{ALWest} + \beta_4\text{NLEast} + \beta_5\text{NLCentral} + \\ & \beta_6\text{NLWest} + \beta_7\text{DaysRested} + \beta_8\text{InningsPitched} + \beta_9\text{EarnedRuns} + \\ & \beta_{10}\text{Strikeouts} + \beta_{11}\text{BattersFaced} + \beta_{12}\text{PitchesThrown} + \\ & \beta_{13}\text{AverageLeverageIndex} + \beta_{14}\text{Base-OutRunsSaved} + \varepsilon.\end{aligned}$$

Home was one of the dummy variables used to see if there was a correlation between pitching in Cleveland or on the road affecting game outcome. Away was controlled for and is included in the α coefficient. The other dummy variable was opponent division which was represented by ALEast, ALWest, NLEast, NLCentral, and NLWest. These variables were used to see if the Guardians pitching excelled or regressed against a certain division. Here, AL Central was controlled for and included in the α coefficient.

Transitioning to quantitative attributes, days rested is represented by its variable of days. The relationship is commonly conceived that more rest days causes better pitcher production. InningsPitched, BattersFaced, and PitchesThrown were attributes used to measure how long the pitcher's performance lasted. Innings pitched is represented by a variable of innings, batters faced is represented by a variable of total batters, and pitch count is the variable for pitches thrown. The correlation here would determine if a team succeeded more when the pitcher went further into the game, typically meaning he has more pitches and more batters faced. The attributes EarnedRuns and Strikeouts have variables of earned runs allowed and total strikeouts. These attributes focused on the batter's performance against the pitcher. Looking at the relationships, more earned runs usually means the team loses, but the strikeout relationship is interesting. Certain pitchers specialize in getting batters to ground-out or fly-out while others specialize in getting striking outs, so there is no true relationship. AverageLeverageIndex and Base-OutRunsSaved are measures from baseball-reference.com that focus on pressure that a pitcher faces each game and runs saved each play respectively.

Data

The first step of data collection was going on to baseball-reference.com and finding a page that lists the five most used Guardians pitchers by season. Therefore, none of the eight seasons used would have 162 games because of injuries, player call-ups or reassignments, and the common bullpen game where no true starting pitchers used. The seasons used spanned from 2014 to 2022, but did not include 2020 COVID season data due to the smaller sample size in games and players sitting out. From there, each player's game log for each season was put into Excel for cleaning. Certain pitchers like Cal Quantrill in 2021 came out of the bullpen in the beginning of the season, so games where the player would make a relief appearance would be deleted. That caused measures like ERA and FIP to be excluded because these bullpen appearances influence said rates. On top of that, the opponent division was entered manually for each game. Lastly, all the columns of data that were not in the model were deleted.

After scraping and cleaning the data, summary statistics for the quantitative variables mentioned in the Model and Variables section were calculated and found in Table 1. Most of the minimums and maximums would be as expected as in the cases of the common measures of innings pitched, earned runs, strikeouts, batters faced, and pitches thrown. One thing that stood out was the maximum for days rested which was 99 days. This appeared for the pitcher's first start of the season as long as he had no relief appearances that occurred before then.

Regression Results

The results of running the model from with all the clean data is shown in Table 2. After looking through the results, there were very few statistically significant attributes, and out of those none had a coefficient greater than or equal to one. The adjusted R^2 value is 0.248 which can be interpreted as 25.8% of the variation in the Cleveland Guardians winning a game can be predicted with all of the dummy variables and quantitative variables making up the independent "X" variable.

Significant at the 1% level were the intercept (containing away games and the AL Central) and base-out runs saved. The intercept is significant and positive most likely because a large portion of their schedule come against weak division opponents. When playing the AL Central the past eight seasons, the Guardians faced either a weak White Sox, Royals, or Tigers

Table 1. Summary Statistics

Days Rested		*Innings Pitched*		*Earned Runs*		*Strikeouts*	
Mean	8.449	Mean	5.854	Mean	2.394	Mean	6.054
Standard Error	0.540	Standard Error	0.049	Standard Error	0.057	Standard Error	0.089
Median	4	Median	6	Median	2	Median	6
Standard Deviation	17.401	Standard Deviation	1.583	Standard Deviation	1.840	Standard Deviation	2.860
Sample Variance	302.798	Sample Variance	2.507	Sample Variance	3.386	Sample Variance	8.182
Skewness	4.800	Skewness	-0.713	Skewness	0.602	Skewness	0.410
Minimum	1	Minimum	0	Minimum	0	Minimum	0
Maximum	99	Maximum	9	Maximum	8	Maximum	18
Batters Faced		*Pitches Thrown*		*Average Leverage Index*		*Base-Out Runs Saved*	
Mean	24.586	Mean	94.408	Mean	0.938	Mean	0.524
Standard Error	0.139	Standard Error	0.511	Standard Error	0.008	Standard Error	0.074
Median	25	Median	98	Median	0.96	Median	0.91
Standard Deviation	4.465	Standard Deviation	16.471	Standard Deviation	0.267	Standard Deviation	2.390
Sample Variance	19.934	Sample Variance	271.302	Sample Variance	0.071	Sample Variance	5.714
Skewness	-1.200	Skewness	-1.453	Skewness	-0.150	Skewness	-0.605
Minimum	1	Minimum	2	Minimum	0.12	Minimum	-6.59
Maximum	35	Maximum	127	Maximum	1.81	Maximum	5.17

Table 2. Regression Results

	Dependent variable:
	Num.W.L
(Home.Away)Home	0.052*
	(-0.027)
(Opp.Division)AL East	-0.030
	(-0.035)
(Opp.Division)AL West	-0.026
	(-0.036)
(Opp.Division)NL Central	-0.098*
	(-0.057)
(Opp.Division)NL East	-0.097
	(-0.086)
(Opp.Division)NL West	-0.111
	(-0.069)
Days Rested	0.0004
	(-0.001)
Innings Pitched	-0.019
	(-0.028)
Earned Runs	-0.049**
	(-0.02)
Strikeouts	0.0002
	(-0.006)
Batters Faced	0.007
	(-0.009)
Pitches Thrown	-0.002
	(-0.002)
Average Leverage Index	-0.091*
	(-0.053)
Base-Out Runs Saved	0.075***
	(-0.02)
Intercept	0.829***
	(-0.107)
Observations	1,037
R^2	0.258
Adjusted R^2	0.248
Residual Std. Error	0.427 (df = 1022)
F Statistic	25.447*** (df = 14; 1022)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

roster depending on the season. This resulted in the Guardians having a better shot of winning. Base-Out Runs Saved is significant and positive because saving runs on each play prevents the opponent from scoring, creating a better shot at winning. At the 5% level, earned runs was significant and negative because the less earned runs allowed gives the Guardians a better chance at winning. Lastly, significant at the 10% level were home games, the NL Central, and Average Leverage Index. Based on the filtered data set, the Guardians did not have a large home field advantage considering the larger p-level with a small coefficient. The NL Central was significant and negative because the Guardians played the Cincinnati Reds every season as well as other Central rivals on certain years, but those games the Guardians did not perform well. Average Leverage Index was significant and positive because it is a measure of pressure the pitcher faces in the game. They are experienced professionals, so pressure should not affect them as much.

It was surprising to see the days rested and innings pitched attributes not be significant nor have a large coefficient despite them being thought of impacting the pitcher and thus game outcome. As mentioned earlier, the strikeouts went as expected because of MLB pitchers specializing in different ways of getting batters out. Batters faced and pitch count were unknown depending on the context. A pitcher can have a complete game while allowing five batters reach base (32 batters faced with high number of pitches thrown), or go six innings while allowing 14 batters reach base (32 batters faced with high number of pitches thrown).

Overall, there is a lot more besides pitching that goes into affects the game outcome for the Cleveland Guardians. The attributes used in the regression model did not have a major effect on the team winning since they had weak correlations and/or no significance. If more attributes and variables were used then there could be a slight difference in the results where more significance and stronger coefficients can be found.