Keely Sweet

Predicting Major League Baseball Attendance with Game Statistics

Using Major League Baseball (MLB) game statistics for the St. Louis Cardinals over the past five years, I attempt to predict game attendance. Previous studies have found a wide variety of variables that impact attendance. Some have shown that significant factors impacting attendance, for example, have changed over time (Lee 2018). Lee finds that from 1904 to 1957, the number of significant variables were much fewer than the number of variables from 1958 to 2012. In recent years, tight competition and high offensive output have been driving attendance up. He also finds that per capita gross domestic product was significant throughout the whole sample. In contrast, Davis (2007) finds that win percentage is the focal point in explaining differences in attendance. Although each team is affected by a differing degree, interleague play also boosts the number of fans in the stands. Accordingly, I will also examine the impact of playing a division rival on Cardinal's attendance. Competitive balance has the power to make a game more or less exciting and sway attendance. Meehan (2007) finds that the effects of a change in competitive balance increase as a team falls further and further behind the division leader. Competitive balance is not as important a factor as the season progresses, and if the team is playing a statistically inferior opponent. People respond strongly to the uncertainty of outcomes. Given these results, I am going to analyze attendance for the Cardinals from 2015 to 2019, focusing on win percentage, team batting average, games back from the division leader and whether or not they are playing another member of the National League Central (NLC). Thus, measurements of overall performance, offensive performance and competition are all included.

I use data from baseballreference.com on Cardinals home games during the regular season from 2015 to 2019. There are a total of 405 observations, with 81 games played each year. Looking at Table I, over the five years, attendance averaged 42,710 with a minimum at 33,448 and maximum of 48,555. Busch Stadium has a capacity of 48,261 as of 2019. The maximum is slightly over the official capacity which may be a result of imprecise calculation of attendance or the nature of the event. Since the game with the highest attendance was played on Mother's Day in 2019, there may have been additional seating or club tickets sold in an attempt to fit more fans. Although capacity has changed from 2015 to 2019, there is less than a 200 seat difference throughout the five years which will not affect the data in significant ways. The day and division indicators show that approximately 32 percent of games were played during the day, and that 47 percent of them were against division opponents. Ranking can take a range of values from one to five, as there are five teams in the National League Central, the St. Louis Cardinals, the Milwaukee Brewers, the Pittsburgh Pirates, the Chicago Cubs, and the Cincinnati Reds. Over the five year time period, the Cardinals were, on average, two and a half games back. However, this variable ranges quite a bit from a ninge game lead to nineteen games back. Statistics that were not as variable, however, include the winning percentage and batting average. For each of these, I take the monthly average and create a two week lag. For example, the win percentage in May was applied to the second half of May and the beginning half of June to help predict attendance more accurately. This lag time allows for Cardinal fans to consider the last two weeks' of performance by the Cardinals when deciding whether to attend a game. The win percentage seems to have a large range with values from .333 to .786. However, some months do not have many games played in them so the extreme lows and highs are due to a small number of observations. Overall, the Cardinals won a little more than 55 percent of the games they played throughout the five year period. Batting average also remained relatively constant with a mean of .251 and standard deviation of .014. Around 10.8 percent of home games were played on Mondays. We can also find that 16.7 percent of games were played on Sundays by adding all the weekday indicator means and subtracting that value from 1. Along the same lines, April is the omitted month, an account for 16.2 percent of the games. May is the month with the most games, with around 18.5 percent of total home games being played then. On the other hand, only a few games are played in October, making up less than 1 percent of the total games.

Table II presents the results of the following regression:

$$\ln(\text{Attendance}_{i,t})= \text{Bo} + B_1{*}\text{DivOpp}_{i,t} + B_2{*}\text{Day}_{i,t} + B_3{*}\text{Rank}_{i,t} + B_4{*}\text{GB}_{i,t} +B_5{*}\text{WLPerc}_{i,t} + B_6{*}\text{BA}_{i,t} + \text{B7}{*}\text{Mon}_{i,t} + B_8{*}\text{Tues}_{i,t} + B_9{*}\text{Wed}_{i,t} + B_{10}{*}\text{Thurs}_{i,t} + B_{11}{*}\text{Fri}_{i,t} + B_{12}{*}\text{Sat}_{i,t} + B_{13}{*}\text{May}_{i,t} + B_{14}{*}\text{Jun}_{i,t} + B_{15}{*}\text{Jul}_{i,t} + B_{16}{*}\text{Aug}_{i,t} + B_{17}{*}\text{Sep}_{i,t} + B_{18}{*}\text{Oct}_{i,t} + B_{19}{*}2016_{i,t} + B_{20}{*}2017_{i,t} + B_{21}{*}2018_{i,t} + B_{22}{*}2019_{i,t} + u_i$$

The results show that interleague play is statistically significant, with a t-statistic of 2.54. There are approximately 650 more fans at games where the Cardinals are playing a division opponent. With average 2019 Cardinals' ticket prices at $34.20, playing an NLC team leads to an average increase in revenue of $22,230. With an operating income of $65 million in 2019, each game should be expected to bring in approximately $800,000. An additional 650 fans, or $22,230, contributes around 2.8 percent to the total revenue expected to come in that day. Although division play is statistically significant, the practical change is noticeable, but small. The number of games back from the division leader is also statistically significant with a t-statistic of -2.75. Compared to a Cardinals team that is two games ahead, a Cardinals team that is six games back would have approximately 2.5 percent fewer fans. Again, this losing Cardinals team would generate a revenue that was $35,191.80 less than the winning Cardinals team. This variable will most likely be practically significant in the long run as well. A team that is eight games back will most likely stay behind for a while and continue to lose potential revenue. It is possible for a losing team to become a winning team, but it takes time to climb up the ranks. A decrease in revenue of $35,191.80 for one game might not be significant in the long run, but a decrease in revenue of $35,191.80 over 40 games would be practically significant. Due to the nature of the games back statistics, there is a higher chance of it significantly affecting revenue in the long run compared to interleague play that does not change based on how a team is playing. Just because the Cardinals play a division opponent today does not mean they will play a division opponent tomorrow. However, if the Cardinals are eight games back today, they will still be at least seven games back tomorrow. Over time, games back can accumulate a large loss, or gain, in revenue. With a t-statistic of -3.65, win-loss percentage proves to be statistically significant; however, a team that has a win-loss percentage of .400 should only expect around 1,630 fewer fans than a team with a monthly win-loss percentage of .650. Considering the large difference between a team with a .650 win-loss percentage and .400 win-loss percentage, the total impact of a 3.8 percent difference in attendance seems practically small. However, win-loss percentage is also a statistic that changes slowly and in the long run, the team may see a significant change in revenue if caught at the extreme ends of the win-loss percentage range for an extended period of time. An offensive metric, batting average, is both statistically and practically insignificant with a t-statistic of 1.65. However, the factor that most influences attendance is the day of the week. Compared to the average Sunday, the average Tuesday had 9.3 percent fewer fans, or $55,746 less in revenue. Not only is Tuesday significantly different than Sunday, it is also practically significant. Saturdays and Sundays are much more comparable when it comes to attendance, as the t-statistic for the Saturday indicator is only 1.11. Some days of the week produce similar attendance, but weekend games pull more fans. When it comes to the year indicators, there is no year that stands out as being statistically or practically significant. Therefore, the specific year is not necessarily important when trying to explain attendance in this five year interval.

When it comes to the St. Louis Cardinals in the past five years, performance metrics affect attendance but only by 1 to 3 percent per game. However, total revenue could be significantly impacted if the team remains at an extreme for an extended period of time. Previous authors have found similar results, as Meehan also finds that interleague competition affects attendance, with an additional 2,337 fans on average for all MLB teams. Although I only find that there are approximately 650 more fans when playing a division opponent, I only use data on the St. Louis Cardinals. Davis also looks at team specific responses and finds that interleague play leaves the Cardinals with around 520 more fans which is much closer to my estimate. On average, other MLB teams in Davis' regression have a much stronger response to playing a division opponent. Each team's fan base responds differently to each variable. Looking at just one team may help to pinpoint the factors needed to increase fan attendance specific to them. While other teams in the MLB respond strongly to playing a division opponent, perhaps the Cardinals respond more strongly to having a winning team. Using this knowledge to their advantage, the Cardinals could allocate their resources more efficiently in order to maximize their profit.

Table I: Descriptive Statistics
St. Louis Cardinals 2015-2019

Variable	Mean	Std. Dev.	Min.	Max.
Attendance	42710.36	3024.51	33448	48555
Division Opponent Indicator	.46666667	.499047	0	1
Day Indicator	.320197	.4671278	0	1
Division Rank	2.192593	5.5328	1	5
Games Back from Division Leader	2.511111	5.5328	-9	19
Win-Loss Percentage	.5575827	.0955174	.333	.786
Batting Average	.2512346	.014256	.223	.280
Monday Indicator	.1083744	.3112362	0	1
Tuesday Indicator	.1453202	.3528582	0	1
Wednesday Indicator	.1477833	.3553229	0	1
Thursday Indicator	.1108374	.3143181	0	1
Friday Indicator	.1576355	.3648485	0	1
Saturday Indicator	.1625616	.3694204	0	1
May Indicator	.1847291	.3885564	0	1
June Indicator	.1576355	.3648485	0	1
July Indicator	.1724138	.3782058	0	1
August Indicator	.1502463	.3577535	0	1
September Indicator	.1699507	.3760529	0	1
October Indicator	.0073892	.0857477	0	1
Year	2017	1.415963	2015	2019

Table II: Regression Results

Number of obs = 405
R-squared = 0.4865
Root MSE = 2228.9

Attendance	Estimated Coefficient	t	P>|t|
Division Opponent Dummy	.0145494 (.0057279)	2.54	0.011
Day Dummy	.0054227 (.0070566)	0.77	0.443
Division Rank	.0016242 (.0042873)	0.38	0.705
Games Back from Division Leader	-.00317382 (.0011555)	-2.75	0.006
Win-Loss Percentage	-.1564987 (.0428501)	-3.65	0.000
Batting Average	.4853433 (.2949347)	1.65	0.101
Monday Dummy	-.076577875 (.0113085)	-6.877	0.000
Tuesday Dummy	-.0934157 (.0108143)	-8.64	0.000
Wednesday Dummy	-.0801018 (.0102400)	-7.82	0.000
Thursday Dummy	-.0715375 (.01077962)	-6.64	0.000
Friday Dummy	-.0067205 (.0107664)	-0.62	0.533
Saturday Dummy	.0104943 (.0094908)	1.11	0.270
May Dummy	.0152035 (.0100033)	1.52	0.129
June Dummy	.0435845 6 (.0109737)	3.97	0.000
July Dummy	.0258496 (.0102029)	2.53	0.012
August Dummy	-.0004095 (.0105273)	-0.04	0.969
September Dummy	.0181616 (.0118766)	1.53	0.127
October Dummy	.0306505 (.0344315)	0.89	0.374
2016 Dummy	.0047543 (.0176013)	0.27	0.787
2017 Dummy	-.0189719 (.0127097)	-1.49	0.136
2018 Dummy	-.0225603 (.0126961)	-1.78	0.076
2019 Dummy	-.0057755 (.0105174)	-0.55	0.583
Constant	10.653 (.0705401)	151.02	0.000

Note: Values in parentheses show standard errors

Works Cited

Davis, Michael C. “Analyzing the Relationship Between Team Success and MLB Attendance With GARCH Effects.” *Journal of Sports Economics*, vol. 10, no. 1, Feb. 2009, pp. 44–58, doi:10.1177/1527002508327387.

Lee, Young H. “Common Factors in Major League Baseball Game Attendance.” *Journal of Sports Economics*, vol. 19, no. 4, May 2018, pp. 583–598, doi:10.1177/1527002516672061.

Meehan, James W., et al. “Competitive Balance and Game Attendance in Major League Baseball.” *Journal of Sports Economics*, vol. 8, no. 6, Dec. 2007, pp. 563–580, doi:10.1177/1527002506296545.