Shifting in Major League Baseball

Introduction

"It's not hard to be romantic about baseball," is an often-quoted line from 2011 movie *Moneyball*, and it is a statement that is simply true. There is something in the air at a ball field, from the most highly funded stadium to the overgrown outfield of a small town's field, something that makes everything feel a little simpler, a little more like childhood and happiness. In the rush of a game, there is nothing that matters but the ball and the base, and perhaps the scoreboard. But baseball is also much more than that, it is a complex game with rules that require both instinct and intellect, and an ability to adapt to any situation that arises. With this idea comes the ever-changing strategies of the game, from the beginnings of batting statistics in 1917 to the development of statistical recruiting in 2003 (Lee, 2018). One of the most recent developments in fielding, however, is the concept of "shifting," or the movement of the majority infielders to one side of the field based on where each individual batter is most probable to hit.

Shifting has only become popular in the last decade, first being used to remarkable success by the Tampa Bay Rays in their 2010 season. Much like the statistical hirings developed in 2003 by the Oakland A's, shifting was a strategy built because of budget. The Tampa Bay Rays in 2010 needed a way to win games on a budget much smaller than many of their competitors, and from this need came the idea of moving fielders to only the spots where a batter was likely to hit – why use a valuable player in a place where they are improbable to be useful when there is so much more to gain in having them placed somewhere else (Heyen, 2020). And it worked – according to Bill Heyen of *Sporting News*, between 2009 and 2011, opponents of the Rays had hits 1.2% less than the league average. Since the Rays first success using this strategy, the form has only grown more popular, with every team in Major League Baseball (MLB) using it in the 2021 season and has become a controversial topic amongst baseball fans due to this.

The initial goal of this paper was to investigate certain aspects of the claims made against shifting, and demonstrating the statistical realities of the effect of shifting on teams based on win/loss proportion and funding, as well as it's effect on individual batters. However, when progressing in the paper, I have found other aspects which lead me to investigate the best model of MLB ranking based on win/loss proportion vs the shifting proportion of teams in 2019. To do this, I plan to run various kinds of regressions on the total data of the 30 MLB teams, looking specifically at their ranking and at the proportion of plate appearances they shifted on. To collect the data for this modeling exploration, I will use the information collected by *Baseball Savant*, which is a database specifically made by the MLB and which stores vast amounts of data about the MLB teams. Too, I will be sourcing my data from the 2019 season, as at the time of writing it is the most recently completed full season, and therefore the most truly accurate and relevant data available. In total, my goal is to better understand the relationship between the proportion of shifting and the win/loss rates of the various major league teams.

Data Collection and Initial Findings

As previously stated, to get data for this project, I used sites such as *Baseball Savant* and *Baseball Reference* to collect information on teams ranking and shift rate. Information on the payroll – the amount of money the team paid to its players over the course of the season – of specific teams was collected from *SpoTrac* and used to operationalize the funding of teams. All this data has been compiled in a variety of tables, which are displayed below.

Table 1 shows the top ten and bottom ten ranked teams of the MLB in 2019 and the percentage of plays they shifted on out of the plate appearances in the season.

TEAM NAME	RANKING	SHIFT PER PLAY (%)
ASTROS	1	49.5
DODGERS	2	50.6
YANKEES	3	36
TWINS	4	35.5
ATHLETICS	5	19.2
BRAVES	6	14.9
RAYS	7	37.2
INDIANS	8	14
NATIONALS	9	14.3
CARDINALS	10	15.8
ANGELS	21	16.8
ROCKIES	22	18.7
PADRES	23	16.7
PIRATES	24	30.2
MARINERS	25	19.1
BLUE JAYS	26	28.5
ROYALS	27	17.9
MARLINS	28	36.4
ORIOLES	29	42.8
TIGERS	30	29.3

Table 1

We can see from this table that the top two teams in 2019, the Astros and the Dodgers, have the highest shift per play (SpP) percentage of these twenty 2019 teams at 49.5% and 50.6%. Meanwhile, in the bottom half of the ranking, teams like the Orioles managed to place in the top 10 highest shift rates while maintaining a spot in the bottom 10 in terms of win/loss rate. The overall mixed use of shifting leads to no apparent correlations between the shift and ranking just by looking at the data. To better display this data however, I went through the process of placing the data into graphs.

Due to the generally decreasing SpP rates in the top ten teams, I suspected that the graph would follow a linear trend. To be able to prove linearity, we must calculate *r* using the formula:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

with

$$\bar{x} = \frac{1+2+3+4+5+6+7+8+9+10}{10} = \frac{11}{2}$$

$$s_x = \sqrt{\frac{\left(1-\frac{31}{2}\right)^2+\left(2-\frac{31}{2}\right)^2+\left(3-\frac{31}{2}\right)^2+\left(4-\frac{31}{2}\right)^2+\left(5-\frac{31}{2}\right)^2+\left(6-\frac{31}{2}\right)^2+\left(7-\frac{31}{2}\right)^2+\left(8-\frac{31}{2}\right)^2+\left(9-\frac{31}{2}\right)^2+\left(10-\frac{31}{2}\right)^2}{30}} = 6.01$$

$$\bar{y} = \frac{49.5+50.6+36+35.5+19.2+14.9+37.2+14+14.3+15.8}{10} = 28.7$$

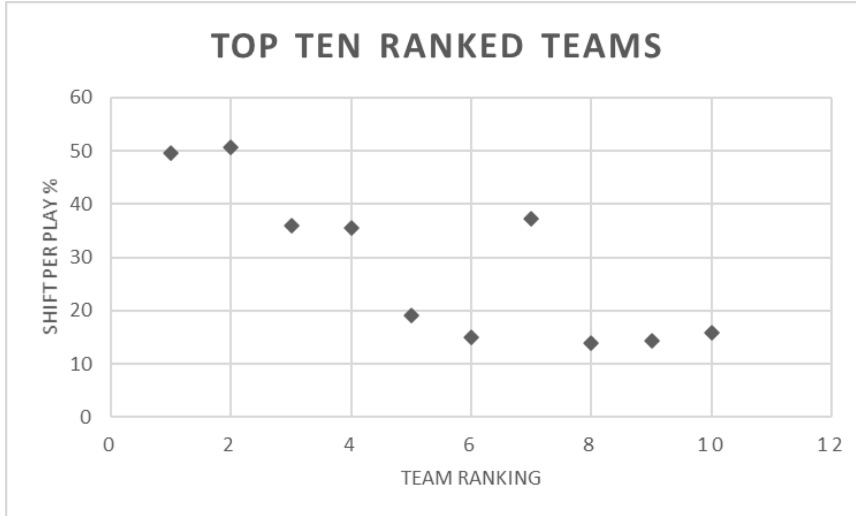
$$s_y = \sqrt{\frac{(49.5-25.6)^2+(50.6-25.6)^2+(36-25.6)^2+(35.5-25.6)^2+(19.2-25.6)^2+(14.9-25.6)^2+(37.2-25.6)^2+(14-25.6)^2+(14.3-25.6)^2+(15.8-25.6)^2}{30}}$$

$$8.26$$

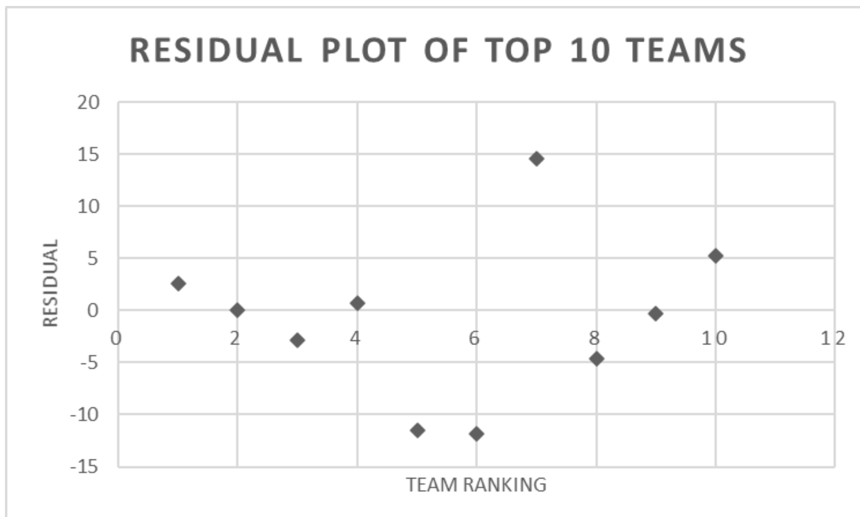
Resulting in a final equation of:

$$r = \frac{1}{10-1}\left(\left(\left(\frac{1-5.5}{6.01}\right)\left(\frac{49.5-28.7}{8.26}\right)\right) + \left(\left(\frac{2-5.5}{6.01}\right)\left(\frac{50.6-28.7}{8.26}\right)\right) + \left(\left(\frac{3-5.5}{6.01}\right)\left(\frac{36-28.7}{8.26}\right)\right) + \left(\left(\frac{4-5.5}{6.01}\right)\left(\frac{35.5-28.7}{8.26}\right)\right) + \left(\left(\frac{5-5.5}{6.01}\right)\left(\frac{19.2-28.7}{8.26}\right)\right) + \left(\left(\frac{6-5.5}{6.01}\right)\left(\frac{14.9-28.7}{8.26}\right)\right) + \left(\left(\frac{7-5.5}{6.01}\right)\left(\frac{37.2-28.7}{8.26}\right)\right) + \left(\left(\frac{8-5.5}{6.01}\right)\left(\frac{14-28.7}{8.26}\right)\right) + \left(\left(\frac{9-5.5}{6.01}\right)\left(\frac{14.3-28.7}{8.26}\right)\right) + \left(\left(\frac{10-5.5}{6.01}\right)\left(\frac{15.8-28.7}{8.26}\right)\right)\right) = -0.746$$

With this r value being above |0.7| there is a strong negative correlation between rank and shifting rate in the top ten MLB teams. Because of this, I put the data into a scatterplot in Graph 1, which further revealed a negative correlation trend, and when looking at the graph we can see what looks to be a linear trend. With this hypothesis of linearity, I moved onto a residuals plot, displayed in Graph 2, where we can see an approximately random scatter, which further implicates the linearity of the data correlation. Because of all these factors, running a linear regression is viable, and results in the equation of $y = 50.92 - 4.04x$. This equation results in an r^2 value of 0.68, which means that the least-squared regression line (LSRL) explains 68% of variation in y or SpP.



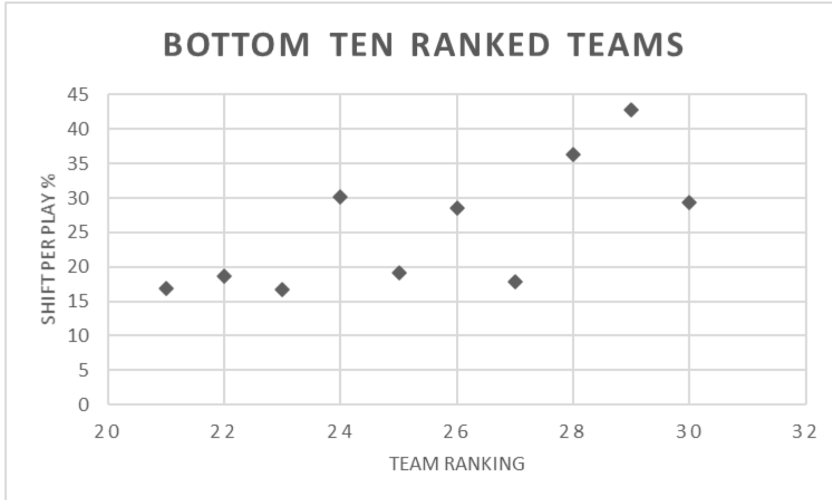
Graph 2



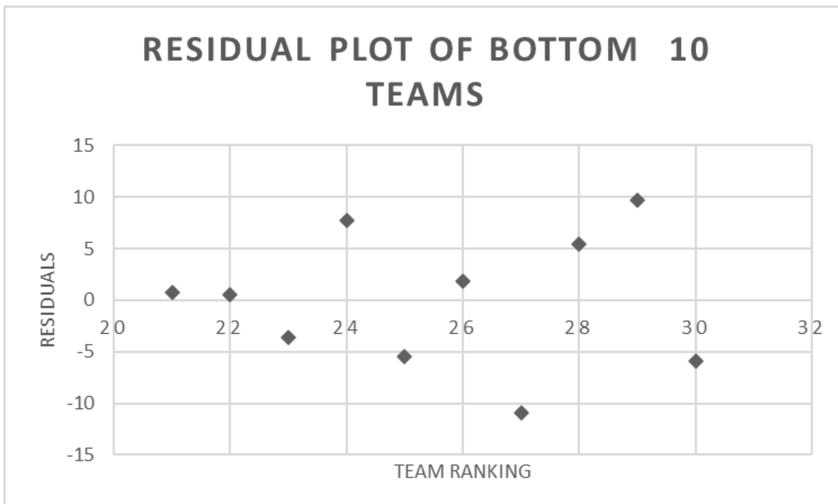
Graph 1

A similar process was gone through with the bottom ten teams in the MLB, with the r value calculations being largely similar, though the workings aren't shown in full here. With this, we see an r value of 0.7, equating to a strong positive correlation. I again then put the data into a scatterplot in Graph 3, where there is what appears to be a positive linear trend. When this is transferred to a residual plot in Graph 4, the trend maintains its relevance, as the plot is randomly scattered, and therefore a linear regression is possible.

When the regression is run, the result is $y = -28.7 + 2.13x$ and have a r^2 value of 0.49 and is therefore only accounting for 49% of variation.



Graph 4



Graph 3

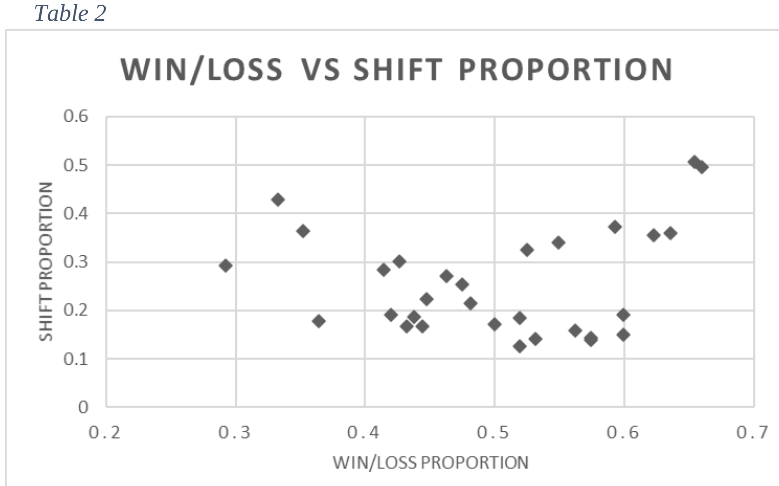
These two sets of data interestingly are opposite to one in terms of correlation direction, despite coming from the same overall population, though on different ends. I also noticed that the 10th and 21st ranked teams have SpP rates which are only separated by 1%, and if graphed on the same graph, would be the same distance to the x-axis from an axis of symmetry for a quadratic function. Noticing this phenomenon and having access to the intermediate ten data points which would make this a complete data set, I chose to expand my initial plan to create a model to understand the total relationship between shifting and the win/loss rate of teams in the 2019 season.

Analysis & Mathematics

The full set of data, displayed in Table 2, shows the team's name, their win/loss proportion, and the teams shifting proportion. To model this data, I first plotted a scatterplot in order to look at the initial shape of the data without having run any regressions or residuals in order to make an initial prediction or interpretation. This is displayed in Graph 5, where we can see what appears to be an approximately curved line. At the very least the association is not linear, though it's unclear what the model may end up being.

TEAM NAME	WIN/LOSS RATE	SHIFT PROPORTION
ASTROS	0.66	0.495
DODGERS	0.654	0.506
YANKEES	0.636	0.36
TWINS	0.623	0.355
ATHLETICS	0.599	0.192
BRAVES	0.599	0.149
RAYS	0.593	0.372
INDIANS	0.574	0.14
NATIONALS	0.574	0.143
CARDINALS	0.562	0.158
BREWERS	0.549	0.341
METS	0.531	0.141
DIAMONDBACKS	0.525	0.325
RED SOX	0.519	0.184
CUBS	0.519	0.127
PHILLIES	0.5	0.171
RANGERS	0.481	0.214
GIANTS	0.475	0.254
REDS	0.463	0.27
WHITE SOX	0.447	0.223
ANGELS	0.444	0.168
ROCKIES	0.438	0.187
PADRES	0.432	0.167
PIRATES	0.426	0.302
MARINERS	0.42	0.191
BLUE JAYS	0.414	0.285
ROYALS	0.364	0.179
MARLINS	0.352	0.364
ORIOLES	0.333	0.428
TIGERS	0.292	0.293

Table 2



Graph 5

To look at what potential model it could be, I began to run regressions of various kinds, looking for the model with the highest *r* value. These various regressions, their modeled formula, *r* values, and r^2 values are all displayed in Table 3. As one can see demonstrated in the table, the *r* value is highest for the power model graphs, with the highest of those being the quartic model at an *r* value of 0.744. This means that using the quartic model, there would be an estimated 74.4% correlation, and because of the r^2 value of 0.553, the model accounts for 55.3% of variation.

REGRESSION NAME	EQUATION	r	r^2
LINEAR	$y = 0.172 + 0.169x$	0.153	0.023
QUADRATIC	$y = 6.64x^2 - 6.34x + 1.7$	0.652	0.426
CUBIC	$y = 35.8x^3 - 44.9x^2 + 17.7x - 1.91$	0.738	0.545
QUARTIC	$y = 95.2x^4 - 148x^3 + 84.6x^2 - 22x + 2.52$	0.744	0.553
POWER	$y = 0.238x^{0.0171}$	0.00598	0.0000358
EXPONENTIAL	$y = 0.205(1.33)^x$	0.0687	0.00472
LOGARITHMIC	$y = 0.288 + 0.0441\,(\log x)$	0.0848	0.0072

Table 3

Because the quartic model only has an *r* value of 0.738, which, while being a strong correlation, is a fairly low correlation in comparison to the possible *r* values if we continue to increase the polynomial order. To attempt to find the regression models, as the graphing calculator used with the initial regressions does not go past a quartic power model, I first attempted to use systems of equations and matrices and then calculate the *r* and r^2.

To test what powers might work best, I can work with systems of equations and matrices in order to create models and then calculate the r and r^2 value of the model. The first system of equations for the model $y = ax^5 + bx^4 + cx^3 + dx^2 + ex + f$ is displayed below.

$$0.495 = a(.66)^5 + b(.66)^4 + c(.66)^3 + d(.66)^2 + e(.66) + f$$

$$0.506 = a(.654)^5 + b(.654)^4 + c(.654)^3 + d(.654)^2 + e(.654) + f$$

$$0.325 = a(.525)^5 + b(.525)^4 + c(.525)^3 + d(.525)^2 + e(.525) + f$$

$$0.127 = a(.519)^5 + b(.519)^4 + c(.519)^3 + d(.525)^2 + e(.525) + f$$

$$0.428 = a(.333)^5 + b(.333)^4 + c(.333)^3 + d(.333)^2 + e(.333) + f$$

$$0.364 = a(.352)^5 + b(.352)^4 + c(.352)^3 + d(.352)^2 + e(.352) + f$$

I then reduced the equations, which are also displayed below.

$$0.495 = 0.125a + 0.19b + 0.287c + 0.436d + 0.66e + f$$

$$0.506 = 0.12a + 0.183b + 0.28c + 0.428d + 0.654e + f$$

$$0.325 = 0.0399a + 0.76b + 0.145c + 0.276d + 0.525e + f$$

$$0.127 = 0.0377a + 0.0726b + 0.14c + 0.269d + 0.519e + f$$

$$0.428 = 0.0041a + 0.0122b + 0.0369c + 0.111d + 0.333e + f$$

$$0.364 = 0.0054a + 0.0154b + 0.0436c + 0.124d + 0.352e + f$$

From here, the new coefficients must be inserted into a matrix, where the coefficients of each system are placed in order, then multiplied by a second matrix of the variables, which are equal the solutions to the systems. This is all displayed in the matrix below.

$$\begin{matrix} 0.125 & 0.19 & 0.287 & 0.436 & 0.66 & 1 \\ 0.12 & 0.183 & 0.28 & 0.428 & 0.654 & 1 \\ 0.0399 & 0.76 & 0.145 & 0.27 & 0.525 & 1 \\ 0.0377 & 0.0726 & 0.14 & 0.269 & 0.519 & 1 \\ 0.0041 & 0.0122 & 0.0369 & 0.111 & 0.333 & 1 \\ 0.0054 & 0.0154 & 0.0436 & 0.124 & 0.352 & 1 \end{matrix} \times \begin{matrix} a \\ b \\ c \\ d \\ e \\ f \end{matrix} = \begin{matrix} 0.495 \\ 0.506 \\ 0.325 \\ 0.127 \\ 0.428 \\ 0.364 \end{matrix}$$

This is all then calculated using the matrix function in a graphing calculator, coming to the following results for each variables posited solution; $a = -648, b = 626, c = 1040, d = -1501, e = 595, f = -74.7$. Thus, the regression line for a power model with the highest power of 5 would be:

$$y = -58.6x^5 + 30.3x^4 + 91.2x^3 - 89.6x^2 + 24.7x - 1.33$$

To finish out this problem, I must calculate the *r* and r^2 values of the equation. To calculate *r* the following equation must be employed:

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

with

$$\bar{x} = \frac{1+2+3+4+5+6+7+8+9+10+11+12+13+14+15+16+17+18+19+20+21+22+23+24+25+26+27+28+29+30}{30} = 15.5$$

$$s_x =$$

$$\sqrt{\frac{\begin{gathered}\left(1-\frac{31}{2}\right)^2+\left(2-\frac{31}{2}\right)^2+\left(3-\frac{31}{2}\right)^2+\left(4-\frac{31}{2}\right)^2+\left(5-\frac{31}{2}\right)^2+\left(6-\frac{31}{2}\right)^2+\left(7-\frac{31}{2}\right)^2+\left(8-\frac{31}{2}\right)^2+\left(9-\frac{31}{2}\right)^2+\left(10-\frac{31}{2}\right)^2+\left(11-\frac{31}{2}\right)^2+\left(12-\frac{31}{2}\right)^2+\left(13-\frac{31}{2}\right)^2+ \\ \left(14-\frac{31}{2}\right)^2+\left(15-\frac{31}{2}\right)^2+\left(16-\frac{31}{2}\right)^2+\left(17-\frac{31}{2}\right)^2+\left(18-\frac{31}{2}\right)^2+\left(19-\frac{31}{2}\right)^2+\left(20-\frac{31}{2}\right)^2+\left(21-\frac{31}{2}\right)^2+\left(22-\frac{31}{2}\right)^2+\left(23-\frac{31}{2}\right)^2+\left(24-\frac{31}{2}\right)^2+\left(25-\frac{31}{2}\right)^2+ \\ \left(26-\frac{31}{2}\right)^2+\left(27-\frac{31}{2}\right)^2+\left(28-\frac{31}{2}\right)^2+\left(29-\frac{31}{2}\right)^2+\left(30-\frac{31}{2}\right)^2\end{gathered}}{30}}$$

$$=8.66$$

$$\bar{y} = \frac{\begin{gathered}49.5+50.6+36+35.5+19.2+14.9+37.2+14+14.3+15.8+34.1+14.1+32.5+18.4+12.7+ \\ 17.1+21.4+25.4+27+22.3+16.8+18.7+16.7+30.2+19.1+28.5+17.9+36.4+42.8+29.3\end{gathered}}{30} = 25.6$$

$$s_y =$$

$$\sqrt{\frac{\begin{array}{c}(49.5-25.6)^2+(50.6-25.6)^2+(36-25.6)^2+(35.5-25.6)^2+(19.2-25.6)^2+(14.9-25.6)^2+(37.2-25.6)^2+(14-25.6)^2+(14.3-25.6)^2+\\(15.8-25.6)^2+(34.1-25.6)^2+(14.1-25.6)^2+(32.5-25.6)^2+(18.4-25.6)^2+(12.7-25.6)^2+(17.1-25.6)^2+(21.4-25.6)^2+(25.4-25.6)^2+\\(27-25.6)^2+(22.3-25.6)^2+(16.8-25.6)^2+(18.7-25.6)^2+(16.7-25.6)^2+(30.2-25.6)^2+(19.1-25.6)^2+(28.5-25.6)^2+(17.9-25.6)^2+\\(36.4-25.6)^2+(42.8-25.6)^2+(29.3-25.6)^2\end{array}}{30}}$$

$$= 10.6$$

Which gives us a final equation of:

$$r = \frac{1}{30-1}\left(\left(\left(\frac{1-15.5}{8.66}\right)\left(\frac{49.5-25.6}{10.6}\right)\right) + \left(\left(\frac{2-15.5}{8.66}\right)\left(\frac{50.6-25.6}{10.6}\right)\right) + \left(\left(\frac{3-15.5}{8.66}\right)\left(\frac{36-25.6}{10.6}\right)\right) +\right.$$

$$\left(\left(\frac{4-15.5}{8.66}\right)\left(\frac{35.5-25.6}{10.6}\right)\right) + \left(\left(\frac{5-15.5}{8.66}\right)\left(\frac{19.2-25.6}{10.6}\right)\right) + \left(\left(\frac{6-15.5}{8.66}\right)\left(\frac{14.9-25.6}{10.6}\right)\right) + \left(\left(\frac{7-15.5}{8.66}\right)\left(\frac{37.2-25.6}{10.6}\right)\right) +$$

$$\left(\left(\frac{8-15.5}{8.66}\right)\left(\frac{14-25.6}{10.6}\right)\right) + \left(\left(\frac{9-15.5}{8.66}\right)\left(\frac{14.3-25.6}{10.6}\right)\right) + \left(\left(\frac{10-15.5}{8.66}\right)\left(\frac{15.8-25.6}{10.6}\right)\right) + \left(\left(\frac{11-15.5}{8.66}\right)\left(\frac{34.1-25.6}{10.6}\right)\right) +$$

$$\left(\left(\frac{12-15.5}{8.66}\right)\left(\frac{14.1-15.5}{10.6}\right)\right) + \left(\left(\frac{13-15.5}{8.66}\right)\left(\frac{32.5-25.6}{10.6}\right)\right) + \left(\left(\frac{14-15.5}{8.66}\right)\left(\frac{18.4-25.6}{10.6}\right)\right) + \left(\left(\frac{15-15.5}{8.66}\right)\left(\frac{12.7-25.6}{10.6}\right)\right) +$$

$$\left(\left(\frac{16-15.5}{8.66}\right)\left(\frac{17.1-25.6}{10.6}\right)\right) + \left(\left(\frac{17-15.5}{8.66}\right)\left(\frac{21.4-25.6}{10.6}\right)\right) + \left(\left(\frac{18-15.5}{8.66}\right)\left(\frac{25.4-25.6}{10.6}\right)\right) + \left(\left(\frac{19-15.5}{8.66}\right)\left(\frac{27-25.6}{10.6}\right)\right) +$$

$$\left(\left(\frac{20-15.5}{8.66}\right)\left(\frac{22.3-25.6}{10.6}\right)\right) + \left(\left(\frac{21-15.5}{8.66}\right)\left(\frac{16.8-25.6}{10.6}\right)\right) + \left(\left(\frac{22-15.5}{8.66}\right)\left(\frac{18.7-25.6}{10.6}\right)\right) + \left(\left(\frac{23-15.5}{8.66}\right)\left(\frac{16.7-25.6}{10.6}\right)\right) +$$

$$\left(\left(\frac{24-15.5}{8.66}\right)\left(\frac{30.2-25.6}{10.6}\right)\right) + \left(\left(\frac{25-15.5}{8.66}\right)\left(\frac{19.1-25.6}{10.6}\right)\right) + \left(\left(\frac{26-15.5}{8.66}\right)\left(\frac{28.5-25.6}{10.6}\right)\right) + \left(\left(\frac{27-15.5}{8.66}\right)\left(\frac{17.9-25.6}{10.6}\right)\right) +$$

$$\left.\left(\left(\frac{28-15.5}{8.66}\right)\left(\frac{36.6-25.6}{10.6}\right)\right) + \left(\left(\frac{29-15.5}{8.66}\right)\left(\frac{42.8-25.6}{10.6}\right)\right) + \left(\left(\frac{30-15.5}{8.66}\right)\left(\frac{29.3-25.6}{10.6}\right)\right)\right)$$

$$r = -0.193$$

As you can see, the *r* value for the model generated using this method is incredibly low, and in no frame of mind would be seen as a valid way to model our data. This at first, confused me as I had checked over my work countless times. However, upon closer observation, this model and methodologies flaws become clear. To create this model, and any model like it, you must create six-variable systems of equations, which requires you to input six points of data to form said equation. The issue lies here, when dealing with a sample of 30 – much larger than the available six points of data. Because of this, it is impossible to get a true accurate representation of the full sample being modeled after; my model used data points from both the highest, middling, and lowest ranking teams, which should theoretically offer the widest range of data, and still it resulted in an *r* value of only $|0.193|$. Clearly, this systems of equations and matrices method is not viable for a data set as large as this.

Due to this, I chose to instead use an online calculator, which both uses the entirety of the data set, and removes chances of deviation through rounding and other factors which are prevalent with hand-done math. The results of these regressions - from a newly calculated power 5 up to power 11 – are displayed in Table 4 below. As we look at the table, we can see that up to x^{10} the *r* value is steadily increasing, though it is beginning to level out as we approach x^{10} with only a 0.001 difference between the *r* values of x^8 and x^9, and a 0.003 difference between x^9 and x^{10}. However, after the 10th power, the *r* value drops significantly, with over 0.021 of a difference between the *r* value of x^{10} and x^{11}.

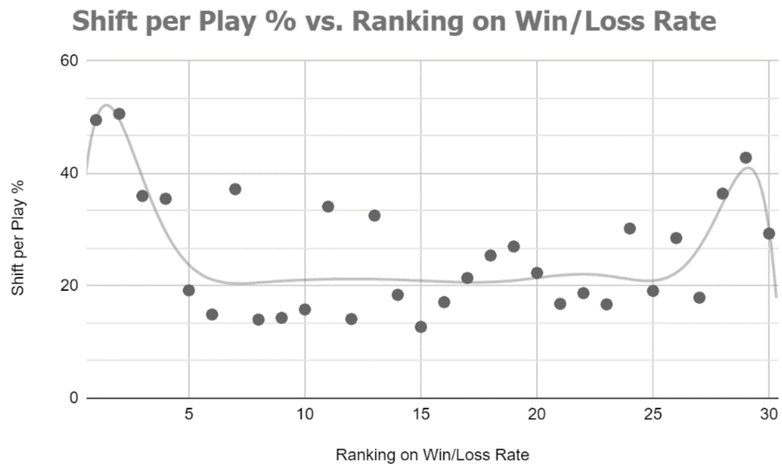
POWER	EQUATION	r	r^2
5	$y = -0.0000317x^5 + 0.003180x^4 - 0.117x^3 + 2x^2 - 16x + 68.7$	0.746	0.558
6	$y = -0.00000606x^6 + 0.000532x^5 - 0.0168x^4 + 0.221x^3 - 0.734x^2 - 6.62x - 59.5$	0.756	0.572
7	$y = -0.000000146x^7 + 0.00000973x^6 - 0.000191x^4 + 0.0465x^3 + 0.305x^2 - 9.37x - 61.7$	0.757	0.573
8	$y = -0.000000206x^8 + 0.0000254x^7 - 0.00129x^6 + 0.539x^5 - 0.539x^4 + 4.72x^3 - 21.4x^2 + 36.9x - 30.2$	0.786	0.618
9	$y = -0.0000000043x^9 + 0.000000393x^8 - 0.00000975x^7 - 0.000164x^6 + 0.0134x^5 - 0.291x^4 + 3.03x^3 - 15.02x^2 + 25.6x + 36.9$	0.787	0.619
10	$y = -0.0000000015x^{10} + 0.000000230x^9 - 0.0000152x^8 + 0.000570x^7 - 0.0134x^6 + 0.205x^5 - 2.04x^4 + 12.8x^3 - 45.7x^2 + 72.6x + 11.7$	0.79	0.625
11	$y = -0.0000000002x^{11} + 0.00000000307x^{10} - 0.00000216x^9 + 0.0000858x^8 - 0.00212x^7 + 0.0335x^6 - 0.334x^5 + 1.99x^4 - 6.04x^3 + 5.05x^2 + 4.09x + 45.2$	0.769	0.591

Table 4

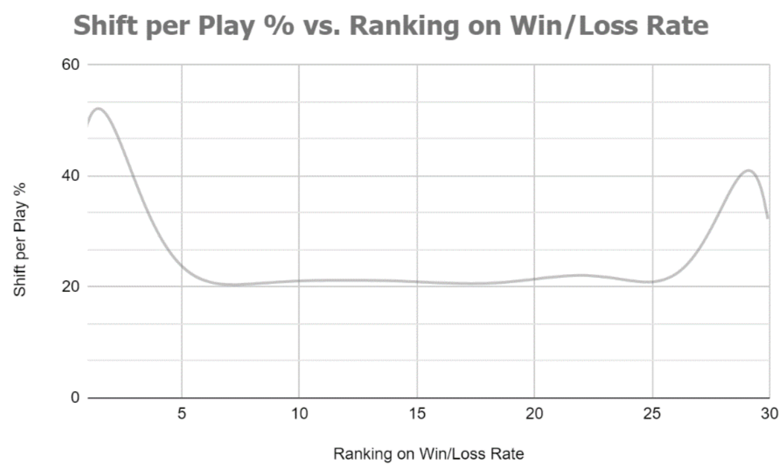
Through this, we can conclude that the best fitting model for the overall relationship between the ranking of win/loss and the SpP of plate appearances shifted on for all 30 MLB teams in the 2019 season is the 10th power parabolic model, specifically:

$$y = -0.0000000015x^{10} + 0.000000230x^9 - 0.0000152x^8 + 0.000570x^7 - 0.0134x^6 + 0.205x^5 - 2.04x^4 + 12.8x^3 - 45.7x^2 + 72.6x + 11.7$$

To look more closely at this, I have graphed the model over the scatterplot in Graph 6 and removed the points on Graph 7. Overall, the model appears to fit the data quite well, with Graph 6 demonstrating that it largely follows the trends of the data, only seeming to have a few outliers which are more removed from the trendline in the top 15 teams. This model also shows, for the most part, that there is a power relationship between the two variables with an absolute maximum at the highest ranked teams and a clear relative maximum at the lowest ranked teams. We can also see that in the higher rankings, there is a sharp decline in shifting as the rankings go down, however at approximately $x = 6$, the SpP rate begins to level off until $x = 26$, middling around 20% SpP for those teams ranked 6^{th} through 26^{th}. We do see a mild relative maximum around $x = 22$, immediately followed by the final rise to the second, more distinct, relative maximum at the lowest ranked teams. The graph also shows that while both the highest and lowest ranked teams tend to have SpP rates above the middling teams, the highest ranked teams are expected to shift a little over 10% more often than the lowest ranking teams. It also shows that it is likely only the top and bottom 5 teams who have SpP rates higher than middling teams, which goes against my initial hypothesis which posed that it would be the top and bottom 10 teams.



Graph 6



Graph 7

Findings and Reflections

After analysis of the models created for the relationship between the ranking of MLB teams based off of win/loss rate and the SpP percentage, I have reached several conclusions about the nature of the shift's use in the MLB in the 2019 season and the most effective ways to model this phenomenon. First, I have learned much about the approaches used to create models, specifically methods involving systems of equations and matrices, which, while an interesting aspect of math and perhaps a good tool for small-sample-models when you lack a calculator, they failed to accurately portray data which is significantly larger than the variable slots. However, when transferring to a far more accurate calculation technique using a digital calculator, I found that the power model of x^{10} was the model with the highest *r* value with a 79% correlation to the raw data. The model revealed that, similar to my hypothesis, there was a correlation between having a more "extreme" ranking, with the relative maximums completing their rise or decent from the 20% SpP rate mark within the top and bottom 5 ranked teams. This could have several implications – perhaps poor playing teams are using it at the wrong times due to ineffective statistical analysis, or they are using it to compensate for other aspects of their team. Better playing teams may simply being using it more effectively and with less error – either due to the superiority of their analysis or simply due to having better players. Meanwhile, middling teams seem to opt out and reap neither the reward nor the risk of shifting, as demonstrated by their placement in the middle of the rankings.

My data, of course, has flaws; I used a limited sample in comparison to the vast amount of data available for the MLB, and specifically only sourced from one season, which offers a limited idea of how shifting is continually being used across seasons. Too, the expansion of data points may allow for a more relevant model; while my data had an *r* value of 0.79 and therefore is strongly correlated, a more comprehensive data set could lead to more powerful models. Topics such as this, which explore the more specific cases of shifting and who uses it, could be even more expansive in looking at how this strategy may be changing the game, something I hope to address in the future.

Bibliography

Heyen, B. (2020, October 21). *Why do baseball teams use the shift? How Dodgers, Rays deployed MLB strategy to reach 2020 World Series*. https://www.sportingnews.com/us/mlb/news/shift-baseball-dodgers-rays-world-series/c3z6i15o1z7h1hh7mpj2bca2b

Lee, C. (2018, March 7). *History of Baseball Statistics*. https://medium.com/@190654/history-of-baseball-statistics-6f2b13f5de20

Tango, T. M., Lichtman, M. G., & Dolphin, A. E. (2014). *The book: Playing the percentages in baseball*.

Appendix A

League Average				vs RHH				vs LHH			
Year	PA	Shifts	Shift %	PA	Shifts	%	wOBA	PA	Shifts	%	wOBA
2019	184392	47178	25.6	108892	15548	14.3	.350	75500	31630	41.9	.330

						vs RHH				vs LHH			
Rk.	Year	Team	PA	Total Shifts	%	PA	Shifts	%	wOBA	PA	Shifts	%	wOBA
1	2019	Dodgers	5884	2975	50.6	3538	1493	42.2	.302	2346	1482	63.2	.293
2	2019	Astros	5922	2934	49.5	3243	862	26.6	.308	2679	2072	77.3	.277
3	2019	Orioles	6381	2732	42.8	3754	1095	29.2	.381	2627	1637	62.3	.339
4	2019	Rays	6059	2255	37.2	3777	1248	33.0	.323	2282	1007	44.1	.282
5	2019	Marlins	6146	2238	36.4	3472	760	21.9	.362	2674	1478	55.3	.352
6	2019	Yankees	6027	2168	36.0	3737	766	20.5	.378	2290	1402	61.2	.310
7	2019	Twins	6236	2216	35.5	3686	1287	34.9	.323	2550	929	36.4	.326
8	2019	Brewers	6223	2124	34.1	3621	692	19.1	.379	2602	1432	55.0	.338
9	2019	D-backs	6164	2002	32.5	3477	698	20.1	.350	2687	1304	48.5	.320
10	2019	Pirates	6325	1910	30.2	3692	718	19.4	.405	2633	1192	45.3	.377
11	2019	Tigers	6237	1830	29.3	3967	539	13.6	.384	2270	1291	56.9	.354
12	2019	Blue Jays	6287	1789	28.5	3471	422	12.2	.364	2816	1367	48.5	.333
13	2019	Reds	5944	1606	27.0	3156	229	7.3	.272	2788	1377	49.4	.317
14	2019	Giants	6230	1581	25.4	3849	636	16.5	.353	2381	945	39.7	.316
15	2019	White Sox	6096	1357	22.3	3558	291	8.2	.391	2538	1066	42.0	.323
16	2019	Rangers	6343	1359	21.4	3802	429	11.3	.382	2541	930	36.6	.324
17	2019	Athletics	6008	1156	19.2	3202	133	4.2	.315	2806	1023	36.5	.296
18	2019	Mariners	6153	1175	19.1	4010	226	5.6	.353	2143	949	44.3	.339
19	2019	Rockies	6386	1197	18.7	3568	290	8.1	.356	2818	907	32.2	.369
20	2019	Red Sox	6198	1141	18.4	3953	38	1.0	.478	2245	1103	49.1	.338
21	2019	Royals	6163	1101	17.9	3463	364	10.5	.315	2700	737	27.3	.369
22	2019	Phillies	6197	1057	17.1	3700	274	7.4	.430	2497	783	31.4	.379
23	2019	Angels	6145	1033	16.8	3721	560	15.0	.405	2424	473	19.5	.317
24	2019	Padres	6141	1023	16.7	3665	319	8.7	.292	2476	704	28.4	.343
25	2019	Cardinals	5915	934	15.8	3267	115	3.5	.452	2648	819	30.9	.320
26	2019	Braves	6209	924	14.9	3936	339	8.6	.368	2273	585	25.7	.354
27	2019	Nationals	6093	872	14.3	3570	116	3.2	.375	2523	756	30.0	.320
28	2019	Mets	6153	870	14.1	3770	221	5.9	.309	2383	649	27.2	.345
29	2019	Indians	5989	840	14.0	3345	96	2.9	.276	2644	744	28.1	.311
30	2019	Cubs	6138	779	12.7	3922	292	7.4	.370	2216	487	22.0	.322

Key:

Red: Above Average Shift per Play Percentage

Blue: Below Average Shift Per Play Percentage