Is the MLB Pitching "Opener" Strategy More Effective than Traditional Pitching Strategies?

Authors: Hank Hopkins, Shixuan Song, Saisharan Chimbili

Part 1: Question

The goal is to determine whether or not the "Opener" strategy is a statistically significant strategy to affect pitching ERA. If the strategy is deemed significant, there are likely some measurable factors that play into this difference; our next goal, then, is to investigate the potential factors. Lastly, we want to investigate whether the "Opener" strategy is more effective than traditional pitching strategy.

Part 2: Motivation

Normally, in a baseball game, there is a starting pitcher who plays the first 5-7 innings out of 9. After the starting pitcher is removed, the manager summons relief pitchers to finish the remaining innings. These starting pitchers can be forced to pitch close to 100 pitches and only rest for 4 days to a week; this combination can deteriorate a pitcher's arm and lead to injury. In 2018, the Tampa Bay Rays made the MLB universe question the effectiveness of this traditional pitching strategy.

Near the ⅔ mark of the season, the Major League Baseball team, Tampa Bay Rays, had none of their original quality starting pitchers on their roster, with 2 months of the season to play. Usually, this lack of starting pitchers would indicate imminent disaster; however, the Rays still had many of their relief pitchers/closers on the roster. Traditionally, relief pitchers are effective pitchers, but are not known for having much stamina; usually, a relief pitcher will typically only pitch 1-4 innings in a game. Kevin Kash, the manager of the Rays, decided to utilize the plethora of relief pitchers by using many pitchers per game, rotating new pitchers in, often. Thus, the "opener" strategy was implemented. The strategy is centered around inserting a relief pitcher to start the game, rather than how they are typically used, to end the game. When the Rays used this “opener” strategy for a few games, their ERA was 3.97, compared to the league average, 4.15. After seeing the results of the “opener” strategy, teams like the Los Angeles Dodgers, Minnesota Twins, Oakland Athletics, Texas Rangers, and Milwaukee Brewers implemented the concept in at least one of their games in 2018.

Of course, the goal of baseball is to win games. We may find that the “opener” strategy is not much more effective in reducing opponents’ run scores. On the contrary, *if* the “opener” strategy has similar efficacy to traditional strategy, then the “opener” strategy may be highly preferable for a number of other potential factors. For example, many teams spend much of their payroll towards their starting pitchers; if teams can mitigate their payout to pitchers, they can utilize their money on other aspects of the team. Lastly, many pitchers face injuries when pitching too much and can ultimately be forced to have Tommy John surgery which can detriment their career. If teams can successfully implement the opener strategy, they could save their pitchers from facing these injuries in their careers. In summary, if the “opener” strategy is deemed as similarly effective (or more effective) to traditional pitching strategy, then this could change the MLB pitching landscape in terms of many factors, mainly payroll and pitchers’ MLB lifespan.

*If you have any further questions about baseball in general or you may not be familiar with the game. Please watch this five minute YouTube Video to give you a generic overview of baseball.

Topic source for our project: https://www.si.com/mlb/2018/08/23/tampa-bay-rays-bullpen

Part 3: Dataset

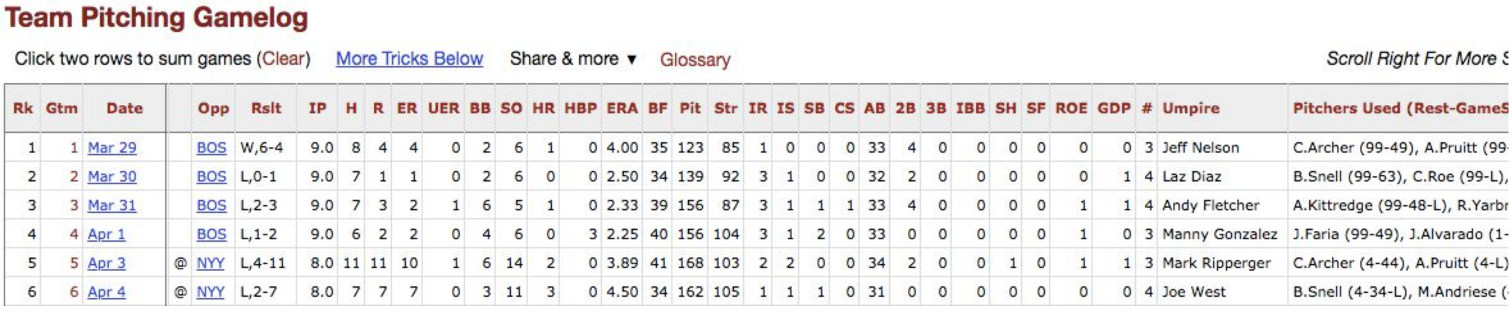
The dataset we want to use stems from baseball-reference.com. Baseball-reference.com is a reliable and complete baseball history dataset used by many national media members. We will particularly be analyzing pitching data from the Tampa Bay Rays 2018 season while using 2018 MLB Team Batting data, as a blocking factor.

Here are the Database Sources:

- *Link for 2018 Tampa Bay Rays game by game pitching log*:
    - https://www.baseball-reference.com/teams/tgl.cgi?team=TBR&t=p&year=2018
- *Link for 2018 MLB Team Statistics for Team Standard Batting*:
    - https://www.baseball-reference.com/leagues/MLB/2018.shtml

Data Cleaning:

- We will be using Tampa Bay Rays game by game pitching log.
- The unit of analysis is one game's worth of pitching data.
- Per the length of the MLB season, there are 162 data points.
- This is a sample of what our pitching data looks like:

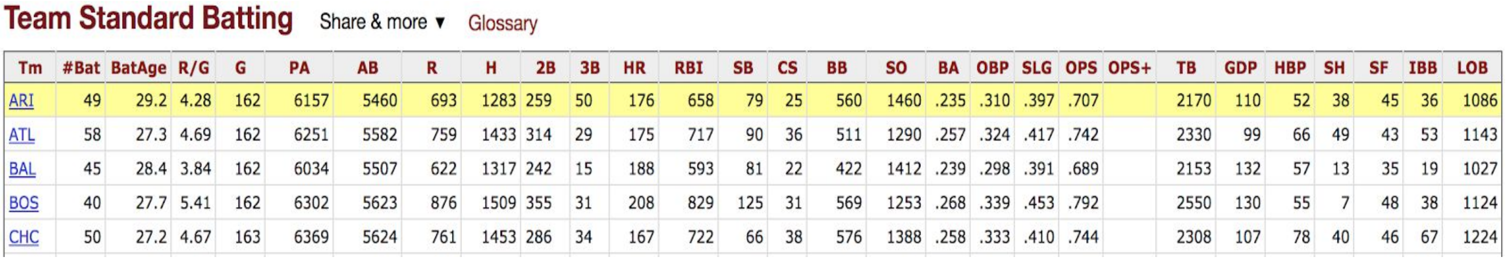


Team Pitching Gamelog

Click two rows to sum games (Clear) More Tricks Below Share & more ▾ Glossary Scroll Right For More S

Rk	Gtm	Date		Opp	Rslt	IP	H	R	ER	UER	BB	SO	HR	HBP	ERA	BF	Pit	Str	IR	IS	SB	CS	AB	2B	3B	IBB	SH	SF	ROE	GDP	#	Umpire	Pitchers Used (Rest-GameS
1	1	Mar 29		BOS	W,6-4	9.0	8	4	4	0	2	6	1	0	4.00	35	123	85	1	0	0	0	33	4	0	0	0	0	0	0	3	Jeff Nelson	C.Archer (99-49), A.Pruitt (99-
2	2	Mar 30		BOS	L,0-1	9.0	7	1	1	0	2	6	0	0	2.50	34	139	92	3	1	0	0	32	2	0	0	0	0	0	1	4	Laz Diaz	B.Snell (99-63), C.Roe (99-L),
3	3	Mar 31		BOS	L,2-3	9.0	7	3	2	1	6	5	1	0	2.33	39	156	87	3	1	1	1	33	4	0	0	0	0	1	1	4	Andy Fletcher	A.Kittredge (99-48-L), R.Yarbr
4	4	Apr 1		BOS	L,1-2	9.0	6	2	2	0	4	6	0	3	2.25	40	156	104	3	1	2	0	33	0	0	0	0	0	1	0	3	Manny Gonzalez	J.Faria (99-49), J.Alvarado (1-
5	5	Apr 3	@	NYY	L,4-11	8.0	11	11	10	1	6	14	2	0	3.89	41	168	103	2	2	0	0	34	2	0	0	1	0	1	1	3	Mark Ripperger	C.Archer (4-44), A.Pruitt (4-L)
6	6	Apr 4	@	NYY	L,2-7	8.0	7	7	7	0	3	11	3	0	4.50	34	162	105	1	1	1	0	31	0	0	0	0	0	0	0	4	Joe West	B.Snell (4-34-L), M.Andriese (

- We will also be using the MLB Team Statistics for 2018 Season Team Standard Batting.
- We will match each Tampa Bay Rays' opponent with their respective On-Base Percentage + Slugging (OPS).

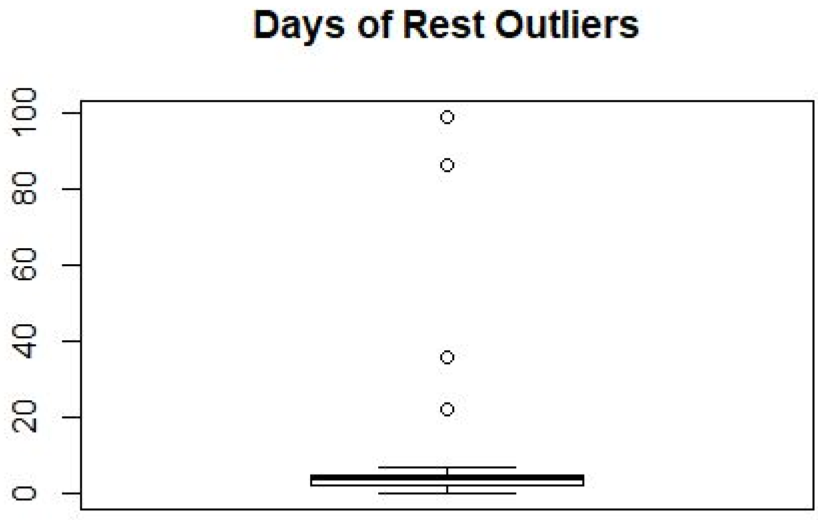


Team Standard Batting Share & more ▾ Glossary

Tm	#Bat	BatAge	R/G	G	PA	AB	R	H	2B	3B	HR	RBI	SB	CS	BB	SO	BA	OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF	IBB	LOB
ARI	49	29.2	4.28	162	6157	5460	693	1283	259	50	176	658	79	25	560	1460	.235	.310	.397	.707		2170	110	52	38	45	36	1086
ATL	58	27.3	4.69	162	6251	5582	759	1433	314	29	175	717	90	36	511	1290	.257	.324	.417	.742		2330	99	66	49	43	53	1143
BAL	45	28.4	3.84	162	6034	5507	622	1317	242	15	188	593	81	22	422	1412	.239	.298	.391	.689		2153	132	57	13	35	19	1027
BOS	40	27.7	5.41	162	6302	5623	876	1509	355	31	208	829	125	31	569	1253	.268	.339	.453	.792		2550	130	55	7	48	38	1124
CHC	50	27.2	4.67	163	6369	5624	761	1453	286	34	167	722	66	38	576	1388	.258	.333	.410	.744		2308	107	78	40	46	67	1224

Create columns in our data frame with:

1. Starting/Opening Pitcher Name
    - The starting pitcher for a game is the first pitcher listed in the “Pitchers Used” column.
    - {1 if opener; 0 if starter}
    - We used the Sports Illustrated article to define our Starters/Openers.
2. Days of Rest.
    - The days of rest is the first number associated with the starting pitcher, from step 1.
    - We found that there were 153 data points from [1,7] and 9 data points with days rest >7.
        - We decided to minimize all values above 7 to the value 7. (e.g. 99 days of rest becomes 7 days of rest). We wanted to preserve as much data as possible, while also minimizing outlier affects.
        - This technique of shifting outlier values is known as Winsorization
    - As you can see in the Boxplot below, there are multiple outlier values. We thought there would be little difference between 7 days of rest and any amount above 7.



*Note: There are multiple data points at each open circle, and all were converted to 7.

3. Home vs Away

- We created a new column with a binary variable. {1 = “home” and 0-away}

4. Opponent On-Base Slugging Percentage
    - Corresponding to opponent for a game.
5. ERA - Binary Dependent variable
    - We will convert the continuous ERA column to a Binary variable
    - Using the two pitching strategies, Starter vs “Opener,” we will use the mean Starter Strategy game ERA as our boundary condition for success/failure.
    - Success = One game’s “Opener” ERA better than mean “Starter” ERA
    - Failure = One game’s “Opener” ERA worse than mean “Starter” ERA
    - The Rays starting pitchers’ ERA was 3.77 for the 2018 season.
    - ERA = {1 if “Opener” game ERA <3.77; 0 if “Opener” game ERA >3.77}

Part 4: Methodology

Model:

ERA ~ Opener_Starting + Num_of_Pitches+ Days_of_Rest + Home + Opp_OnBase_SluggPct

Dependent:

-*ERA for one game*

-ERA = (earned runs x 9{standard innings per game}) / (innings pitched)

-ERA is how many runs a pitcher allows on average per 9 innings pitched.

- We will convert ERA to a binary variable.

- ERA = {1 if game ERA < 3.77 ; 0 if game ERA >3.77}

- Reason for choosing factor: ERA allows us to determine if the pitcher is performing well. The goal of the pitcher is to minimize the amount of runs the opponent scores, and the ERA accounts for the rate of amount of runs the pitcher gives up to the opposing team. We use 3.77 as a baseline, because this is the average ERA for the starting pitcher strategy. By doing this, we can compare the “opener” strategy directly to the starting pitcher strategy to see if the “opener” strategy is statistically significant and if it is better or not.

Explanatory:

-*Opener*

- This is a categorical/binary variable.

- {1 if opener starts ; 0 when opener doesn’t start}

- Reason for choosing factor: The strategy is in effect when a opener is starting, otherwise the strategy is not being used when an opener doesn’t start. This allows us to assess the effectiveness of the opener strategy on the ERA for a game.

-*Pit*

- This is a continuous variable that shows the amount of pitches thrown in the game.
- Reason for choosing factor: We want to see whether the number of pitches thrown in a game is a significant factor in affecting the log odds of achieving a better ERA than the mean Starting pitcher ERA. We suspect that more pitches means increased pitcher fatigue, which would cause increased (worse) ERA.

-*DaysRest*

-This is a continuous variable that measures the amount of days of rest for the starting pitcher since last pitched

- Reason for choosing factor: We believe that the more rest pitchers are able to have, the better their ERA will be.

- *Home*
    - This is a categorical/binary variable
        -{1 = Rays playing at Home, 0=Rays playing at opponent's field}
    - Reason for choosing factor: There are studies done that show pitchers perform better at their home stadium compared to their opponent's field.

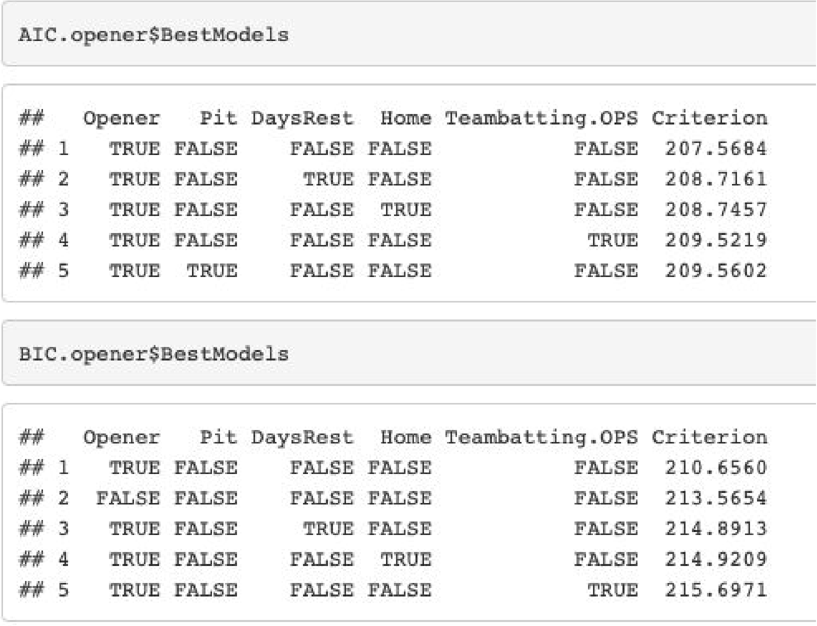
- *Teambatting.OPS*
    - This is a continuous variable depicting Tampa Bay opponent's batting success.
    - Reason for choosing factor: We believe this is the best batting statistic to Include.

OPS - (On Base Percentage + Slugging percentage)

is a commonly used statistical measure for batting in the MLB. It takes into account 2 major aspects of batting success: reaching base successfully (On-base percentage), and successfully reaching base, but accounting for larger magnitudes of success (for example, a home run, triple, or double is more valuable than a single). Thus, Slugging percentage also accounts for the batter's tendency to have extra-base hits.

We will be using R. We will use the glm function to model a logistic regression. We will also use dplyr, mcprofile and bestglm to help us clean the data and select the proper model.

Part 5: Results/Interpretation



```
AIC.opener$BestModels
```

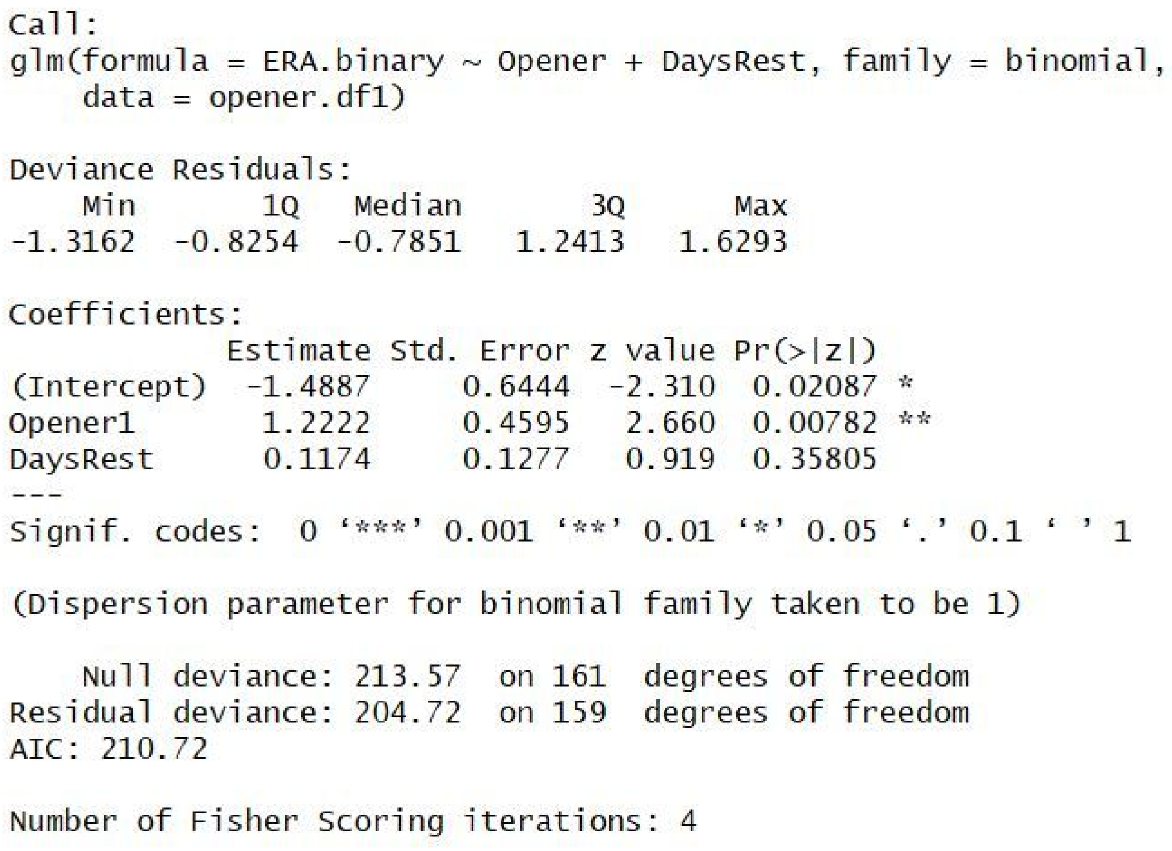
```
##    Opener   Pit DaysRest  Home Teambatting.OPS Criterion
## 1   TRUE FALSE    FALSE FALSE           FALSE  207.5684
## 2   TRUE FALSE     TRUE FALSE           FALSE  208.7161
## 3   TRUE FALSE    FALSE  TRUE           FALSE  208.7457
## 4   TRUE FALSE    FALSE FALSE            TRUE  209.5219
## 5   TRUE  TRUE    FALSE FALSE           FALSE  209.5602
```

```
BIC.opener$BestModels
```

```
##    Opener   Pit DaysRest  Home Teambatting.OPS Criterion
## 1   TRUE FALSE    FALSE FALSE           FALSE  210.6560
## 2  FALSE FALSE    FALSE FALSE           FALSE  213.5654
## 3   TRUE FALSE     TRUE FALSE           FALSE  214.8913
## 4   TRUE FALSE    FALSE  TRUE           FALSE  214.9209
## 5   TRUE FALSE    FALSE FALSE            TRUE  215.6971
```

Considering the low number of variables that we ought to include in our model, we suspect that using the AIC's results is potentially more beneficial, because we do not have to worry about the possibility of the AIC's results leading to overfitting, due to the low order of dimensionality of the number of variables to reasonably suggest.

We are also going to use the model that includes both Opener and DaysRest, because they are close in Criterion values and including more variables helps account for the explained variability in our dependent variable as best as possible.



```
Call:
glm(formula = ERA.binary ~ Opener + DaysRest, family = binomial,
    data = opener.df1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3162  -0.8254  -0.7851   1.2413   1.6293

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4887     0.6444  -2.310  0.02087 *
Opener1       1.2222     0.4595   2.660  0.00782 **
DaysRest      0.1174     0.1277   0.919  0.35805
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 213.57  on 161  degrees of freedom
Residual deviance: 204.72  on 159  degrees of freedom
AIC: 210.72

Number of Fisher Scoring iterations: 4
```

The regression output suggest that Opener is the most significant variable. Although, the AIC model informed us to use just the *Opener* explanatory variable, we decided to also include the *DaysRest* variable as well since our book states that we are allowed to use the other four models given as long as they fall within two Criterion units of the best model from the AIC. Our regression shows that the *Opener* variable is very significant at a p-value of 0.00782. The coefficient for *Opener* (1.2222) represents the change in log odds of success when an opener pitcher starts the game. However, the DaysRest variable does not seem to be significant (due to the relatively high p-value: 0.35805). The coefficient for *DaysRest* represents the change in log odds of success for one more day of rest the pitcher (either opener or starting) receives.

Additionally, we can look at the Odds Ratio between Opener pitchers and Starting pitchers. The interpretation of our result would be that the odds of getting game ERA < 3.77 (starting pitcher average) with opener strategy is 3.3939 times as large as without using the opener strategy. Furthermore, we can look at the 95% Confidence Interval for the Odds Ratio, which is (1.3793, 8.3547). Since 1 is not in the interval, the Odds Ratio is statistically significant.

Work

Interpretation of Coefficients:

- *Opener:* 1.2222 represents the change in log odds of success when an opener pitcher starts the game
- *DaysRest*: 0.1174 represents the change in log odds of success when a pitcher has one more day of rest before playing.

  Note*: We used average days of rest, 3.685185, in the calculations below. We realize that the DaysRest value might affect the model with different values, so we keep it in our model. But we want to mainly focus on how the opener strategy would affect the game ERA rather than the Days Rest, so we simply use the average days of rest for this generalized linear model to calculate the probability of success, odds of success, and the odds ratio with its confidence interval.

Probabilities

Probability of success: the probability that the opener pitcher strategy records a lower 1-game ERA compared with the mean ERA of starters.

With opener strategy:

- pi.opener=P(Success| Opener=1, daysrest=3.685185)

  = $\exp(\beta_0+\beta_1+\beta_2(3.685185)) / [1+\exp(\beta_0+\beta_1+\beta_2*3.685185)]$

  = 0.5414
- odds of success with opener strategy = pi.opener/(1-pi.opener) = 1.1807

Without opener strategy:

- pi.without.opener = P(Success| Opener = 0, daysrest = 3.685185)

  $=\exp(\beta_0+\beta_2*3.685185) / [1+\exp(\beta_0+\beta_2*3.685185)]$

  = .2581
- odds of success without opener strategy= pi.without.opener/(1-pi.without.opener) = $\exp(\beta_0+\beta_2*3.685185)$

  = 0.3478

Odds Ratio

Odds of with opener strategy vs. odds without opener strategy:

$\exp(\beta_0+\beta_1+\beta_2(3.685185))/\exp(\beta_0+\beta_2(3.685185)) = \exp(1.222)= 3.3939$

Interpretation: The odds of getting game ERA < 3.77 (starting pitcher average) with opener strategy is 3.3939 times as large as without using the opener strategy.

95% confidence interval of the odds ratio:

$\exp(\beta_1 \pm 1.96*\sqrt{Var.hat(\beta1.hat)}) = \exp(1.2222 \pm 1.96*0.4595) = (1.3793, 8.3547)$

Interpretation:

With 95% confidence, we can conclude that the odds of getting game ERA < 3.77 (starting pitcher average) with opener strategy are between 1.3793 and 8.3547 times as large than without using the opener strategy. Since the interval does not include 1, we can conclude, with 95% certainty that the odds of success for Opener Strategy are greater than odds of success for Starter Strategy.

Part 6: Conclusion and Reflection

After observing our output, we are able to determine that the opener strategy is very significant with a p-value of 0.00782. Furthermore, if we use the odds ratio to compare the opener strategy to the starting pitcher strategy, the odds of having an ERA < 3.77 for one game is 3.3939 times as large as without using the opener strategy. Furthermore, we calculated a 95% Confidence Interval of the odds ratio which was 1.3793 to 8.3547. Since 1 is not in the interval, we can see that the opener strategy is significant and further supports our very significant p-value.

Although our analysis may show that the opener strategy is significantly better than the starting pitcher strategy, there could be limitations with our data that may affect our analysis.

One issue is the incorporation of only one team. Our results can only be shown to be effective for the Tampa Bay Rays and their roster of pitchers for the 2018 season. There is a chance that the relief pitchers (i.e. non-starters) for the Tampa Bay Rays are significantly better than the TB Rays' starters as well as pitchers from other MLB teams. In this case, our results would indicate the effectiveness of these specific pitchers rather than the actual effectiveness of the Opener Strategy.

We also only had 162 data points to work with. If other teams incorporated the strategy over the next few years, we would have many more data points to verify our results.

There is also the problem of independence, which seems to naturally arise often. In terms of independence of ERA, particularly, within one game, there could be an exponentiation factor behind ERA in a game. For example, let us lay out two scenarios:

1. The pitchers have given up 6 runs over the first 3 innings.
2. The pitchers have given up 0 runs over the first 3 innings.

There is a possibility that scenario 1 will lend its way to many more runs being scored over the remaining 6 innings compared to scenario 2. This would indicate that inning-by-inning pitching performance is not independent. The effect described could lead to skewed ERA results.

Furthermore, the opener strategy may not be independent from the starting pitching strategy. We found that the Rays were switching between strategies which ultimately gave the starting pitcher’s more rest before they played. This may be a reason why starting pitchers may have performed better than their own strategy due to the opener strategy benefiting them.

Overall, we were very excited to see our results and to use the material we’ve learned throughout the semester to a real topic. One of the most difficult parts of the project was to make sure we considered the right explanatory variables and to parse our data with R. If we had more knowledge and data, we would consider using interactions between explanatory variable and other explanatory variables which were not provided to us in order to reduce the significance of the intercept of our model. Lastly, our group spent about 40 hours together and went to Office Hours three times with Professor Holt.