Building a Winning Formula: Key Factors for World Series Success

Lissandro Alvarado, Jack Krasulak, Daniel Kloner, Ryan Blatney

Project Overview

Roster Depth and strength

- Star player impact: MVP-caliber performances
- Depth of bullpen and starting rotation

In-Season Performance Trends

- Win-loss record against teams
- Late-season momentum and Energy

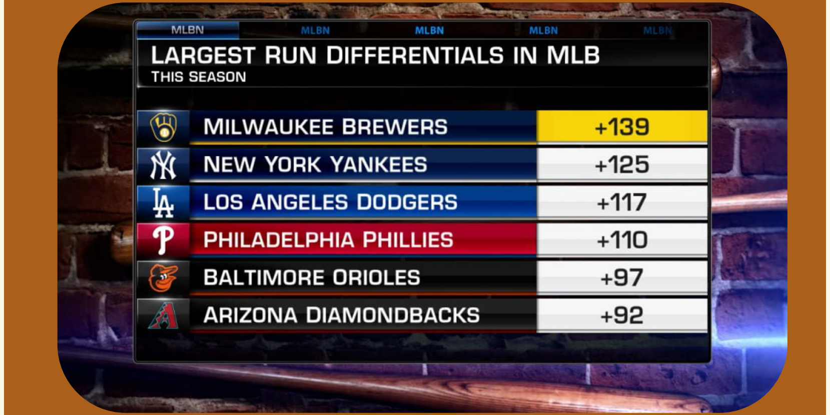
Data Alignment and Discrepancies

- Different models emphasize by offensive or defensive performance
- Metrics vary in predictive accuracy

The Goal

To provide insights on a team's chances of reaching the World Series while understanding the variability in predictive models and the variability of live sport

Subproblem One


MLBN
MLBN
MLBN
MLBN
MLBN

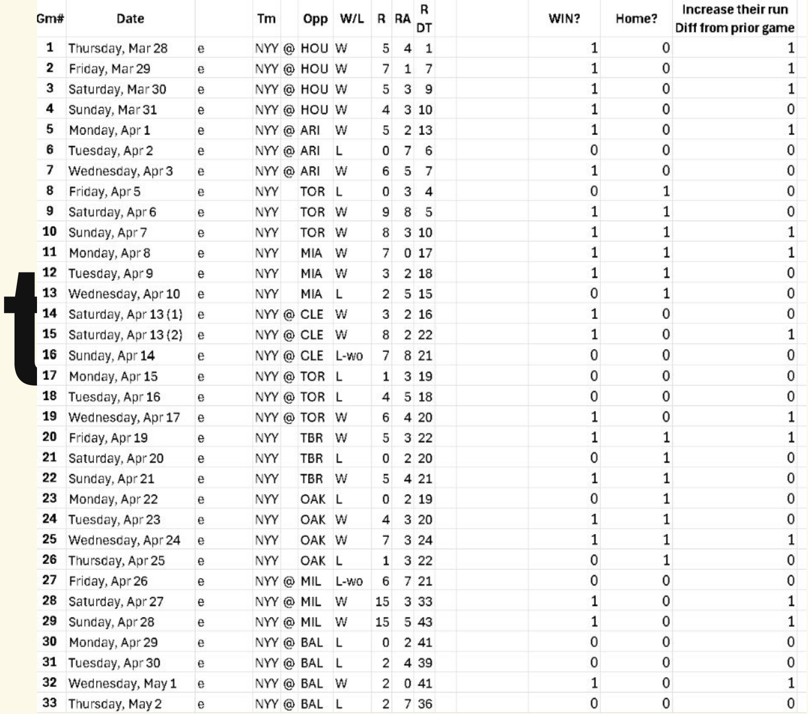
LARGEST RUN DIFFERENTIALS IN MLB

THIS SEASON

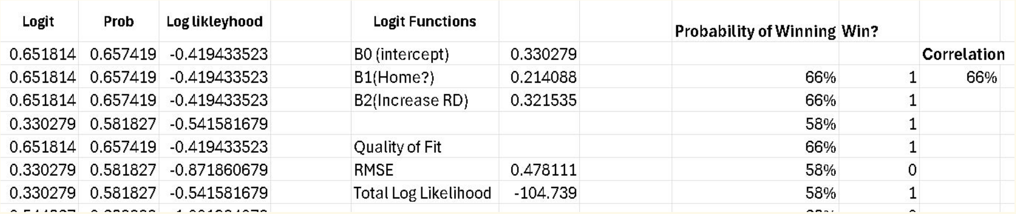
MILWAUKEE BREWERS	+139
NEW YORK YANKEES	+125
LOS ANGELES DODGERS	+117
PHILADELPHIA PHILLIES	+110
BALTIMORE ORIOLES	+97
ARIZONA DIAMONDBACKS	+92

- **Definition of Run Differential:** The difference between runs scored and runs allowed over the season.
- **Central Question:** Does a higher average run differential correlate with better initial postseason success?
- **Why Run Differential:** Run differential as a key performance indicator for team dominance and consistency.

Data



Gm#	Date		Tm		Opp	W/L	R	RA	R DT	WIN?	Home?	Increase their run Diff from prior game
1	Thursday, Mar 28	e	NYY	@	HOU	W	5	4	1	1	0	1
2	Friday, Mar 29	e	NYY	@	HOU	W	7	1	7	1	0	1
3	Saturday, Mar 30	e	NYY	@	HOU	W	5	3	9	1	0	1
4	Sunday, Mar 31	e	NYY	@	HOU	W	4	3	10	1	0	0
5	Monday, Apr 1	e	NYY	@	ARI	W	5	2	13	1	0	1
6	Tuesday, Apr 2	e	NYY	@	ARI	L	0	7	6	0	0	0
7	Wednesday, Apr 3	e	NYY	@	ARI	W	6	5	7	1	0	0
8	Friday, Apr 5	e	NYY		TOR	L	0	3	4	0	1	0
9	Saturday, Apr 6	e	NYY		TOR	W	9	8	5	1	1	0
10	Sunday, Apr 7	e	NYY		TOR	W	8	3	10	1	1	1
11	Monday, Apr 8	e	NYY		MIA	W	7	0	17	1	1	1
12	Tuesday, Apr 9	e	NYY		MIA	W	3	2	18	1	1	0
13	Wednesday, Apr 10	e	NYY		MIA	L	2	5	15	0	1	0
14	Saturday, Apr 13 (1)	e	NYY	@	CLE	W	3	2	16	1	0	0
15	Saturday, Apr 13 (2)	e	NYY	@	CLE	W	8	2	22	1	0	1
16	Sunday, Apr 14	e	NYY	@	CLE	L-wo	7	8	21	0	0	0
17	Monday, Apr 15	e	NYY	@	TOR	L	1	3	19	0	0	0
18	Tuesday, Apr 16	e	NYY	@	TOR	L	4	5	18	0	0	0
19	Wednesday, Apr 17	e	NYY	@	TOR	W	6	4	20	1	0	1
20	Friday, Apr 19	e	NYY		TBR	W	5	3	22	1	1	1
21	Saturday, Apr 20	e	NYY		TBR	L	0	2	20	0	1	0
22	Sunday, Apr 21	e	NYY		TBR	W	5	4	21	1	1	0
23	Monday, Apr 22	e	NYY		OAK	L	0	2	19	0	1	0
24	Tuesday, Apr 23	e	NYY		OAK	W	4	3	20	1	1	0
25	Wednesday, Apr 24	e	NYY		OAK	W	7	3	24	1	1	1
26	Thursday, Apr 25	e	NYY		OAK	L	1	3	22	0	1	0
27	Friday, Apr 26	e	NYY	@	MIL	L-wo	6	7	21	0	0	0
28	Saturday, Apr 27	e	NYY	@	MIL	W	15	3	33	1	0	1
29	Sunday, Apr 28	e	NYY	@	MIL	W	15	5	43	1	0	1
30	Monday, Apr 29	e	NYY	@	BAL	L	0	2	41	0	0	0
31	Tuesday, Apr 30	e	NYY	@	BAL	L	2	4	39	0	0	0
32	Wednesday, May 1	e	NYY	@	BAL	W	2	0	41	1	0	1
33	Thursday, May 2	e	NYY	@	BAL	L	2	7	36	0	0	0



Logit	Prob	Log likleyhood	Logit Functions		Probability of Winning	Win?	
0.651814	0.657419	-0.419433523	B0 (intercept)	0.330279			Correlation
0.651814	0.657419	-0.419433523	B1(Home?)	0.214088	66%	1	66%
0.651814	0.657419	-0.419433523	B2(Increase RD)	0.321535	66%	1	
0.330279	0.581827	-0.541581679			58%	1	
0.651814	0.657419	-0.419433523	Quality of Fit		66%	1	
0.330279	0.581827	-0.871860679	RMSE	0.478111	58%	0	
0.330279	0.581827	-0.541581679	Total Log Likelihood	-104.739	58%	1	

KC	80%
NY	66%
HOUSTON H ASTROS	86%
A	83%
SD	76%
D	66%

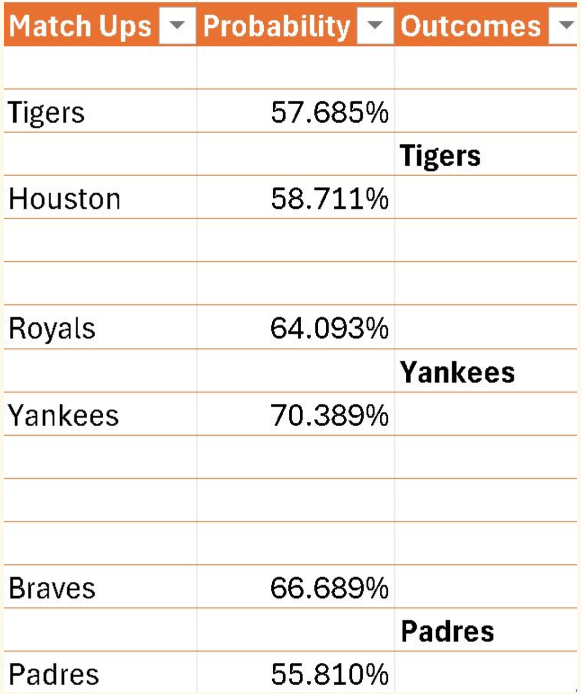
Correlation between Reality and probability of winning given by the model

Applying the Methodology

1. How much does Momentum mean going into the playoffs

2. Why The difference between Regular and postseason

3. What needs to be added



Match Ups	Probability	Outcomes
Tigers	57.685%	
		Tigers
Houston	58.711%	
Royals	64.093%	
		Yankees
Yankees	70.389%	
Braves	66.689%	
		Padres
Padres	55.810%	

Subproblem Two

- **Central Question:** Do teams with higher payrolls have better chances to get into the world series?
- **Why Payroll:** Payroll is a tangible, quantifiable variable that MLB teams can control to some extent through resource allocation.
- **Method:** Logistic Regression and Monte Carlo Simulation

Logistic Regression

Y variable - Whether team made it to the world series or not (binary)

X variable - Payroll per MLB team per year in millions

B0	B1	Exponentiated B1
-4.56	0.012	1.0122

Monte Carlo Simulation

Teams with higher payroll are more likely to make it to the World Series

Average Success in Making it to the World Series By Payroll Level	
High Payroll	14%
Medium Payroll	5%
Low Payroll	2%

Subproblem Three

- **Central Question:** How do different defensive metrics help to predict how a team performs in a season?
- **Methodology:** Used a Logistic Regression to look at specific team defensive stats and determine which were most important.
- **Why Defensive Metrics?:** These stats are crucial in impacting run prevention which is just as important as scoring runs.

The Data

Predicting Runs Allowed per Game

1. Using Solver to Minimize RMSE

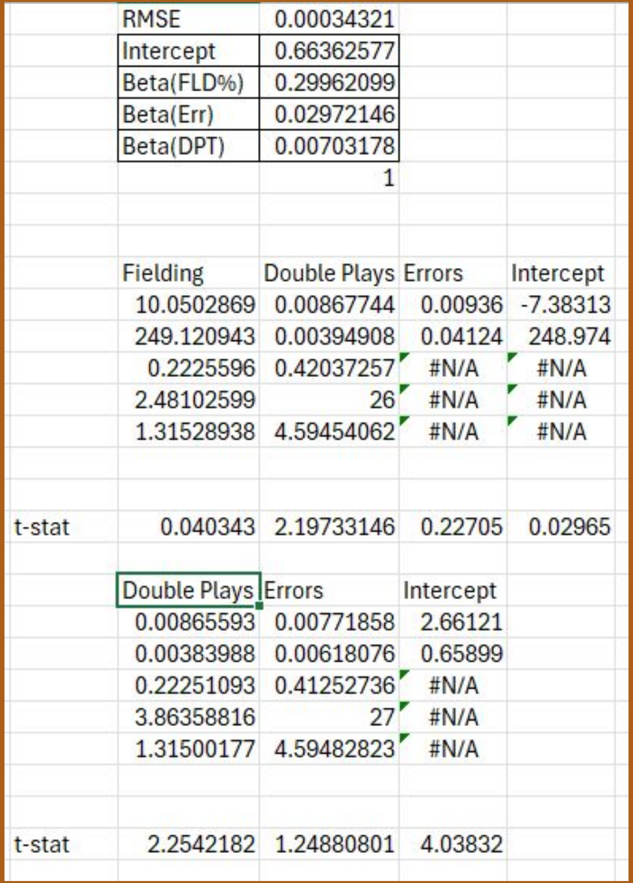
Fielding %, Errors, and Defensive Plays Turned were the three metrics used. Model provided an “ok” prediction in regards to team performance.

2. Using Linest Function

Using the same metrics tried using “Linest” to check for any variability. Model was different and was a little bit better.

3. Using Linest Function Minus Fielding %

In an effort to increase the t-stat figures I removed Fielding % as it was insignificant in helping to predict team performance. Even with the change the difference was minimal compared to the original Linest Model

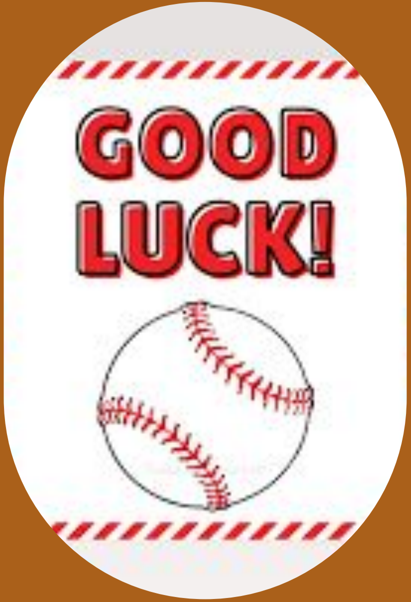


RMSE	0.00034321
Intercept	0.66362577
Beta(FLD%)	0.29962099
Beta(Err)	0.02972146
Beta(DPT)	0.00703178
	1

	Fielding	Double Plays	Errors	Intercept
	10.0502869	0.00867744	0.00936	-7.38313
	249.120943	0.00394908	0.04124	248.974
	0.2225596	0.42037257	#N/A	#N/A
	2.48102599	26	#N/A	#N/A
	1.31528938	4.59454062	#N/A	#N/A
t-stat	0.040343	2.19733146	0.22705	0.02965

	Double Plays	Errors	Intercept
	0.00865593	0.00771858	2.66121
	0.00383988	0.00618076	0.65899
	0.22251093	0.41252736	#N/A
	3.86358816	27	#N/A
	1.31500177	4.59482823	#N/A
t-stat	2.2542182	1.24880801	4.03832

Subproblem Four



- **Central Question:** Do teams win more games than expected, and therefore have higher chances of winning world series if they have higher team stats in playoffs?
- **Why "Luck":** Important to see if despite many different factors in building a team, does some of winning just come down to a team getting hot at the right time
- **Method:** Logistic Regression Linest test to see if betas are zero or not, taking data since 2015

The Data

Y variable- wins in playoffs- expected wins in playoffs

“Luck” variable 1

Batting average difference

“Luck” variable 2

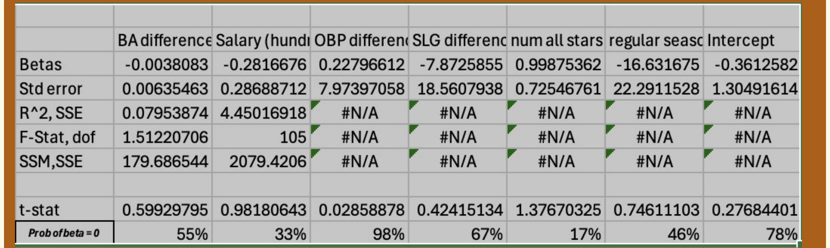
On base percentage difference

“Luck” variable 3

Slugging percentage difference

Other variables involved

All stars per team, run differential, Salary



	BA difference	Salary (hundı	OBP differen	SLG differenc	num all stars	regular seasc	Intercept
Betas	-0.0038083	-0.2816676	0.22796612	-7.8725855	0.99875362	-16.631675	-0.3612582
Std error	0.00635463	0.28688712	7.97397058	18.5607938	0.72546761	22.2911528	1.30491614
R^2, SSE	0.07953874	4.45016918	#N/A	#N/A	#N/A	#N/A	#N/A
F-Stat, dof	1.51220706	105	#N/A	#N/A	#N/A	#N/A	#N/A
SSM,SSE	179.686544	2079.4206	#N/A	#N/A	#N/A	#N/A	#N/A
t-stat	0.59929795	0.98180643	0.02858878	0.42415134	1.37670325	0.74611103	0.27684401
Prob of beta = 0	55%	33%	98%	67%	17%	46%	78%

- No variables are significant
- OBP the only positive beta out of the luck variables
- Number of all stars seems to be the only one close to being important

Conclusion

- More Time and More Data
    - While individually there were no strong correlations with the individual metrics
    - We think combining them would produce better outputs.

Questions?