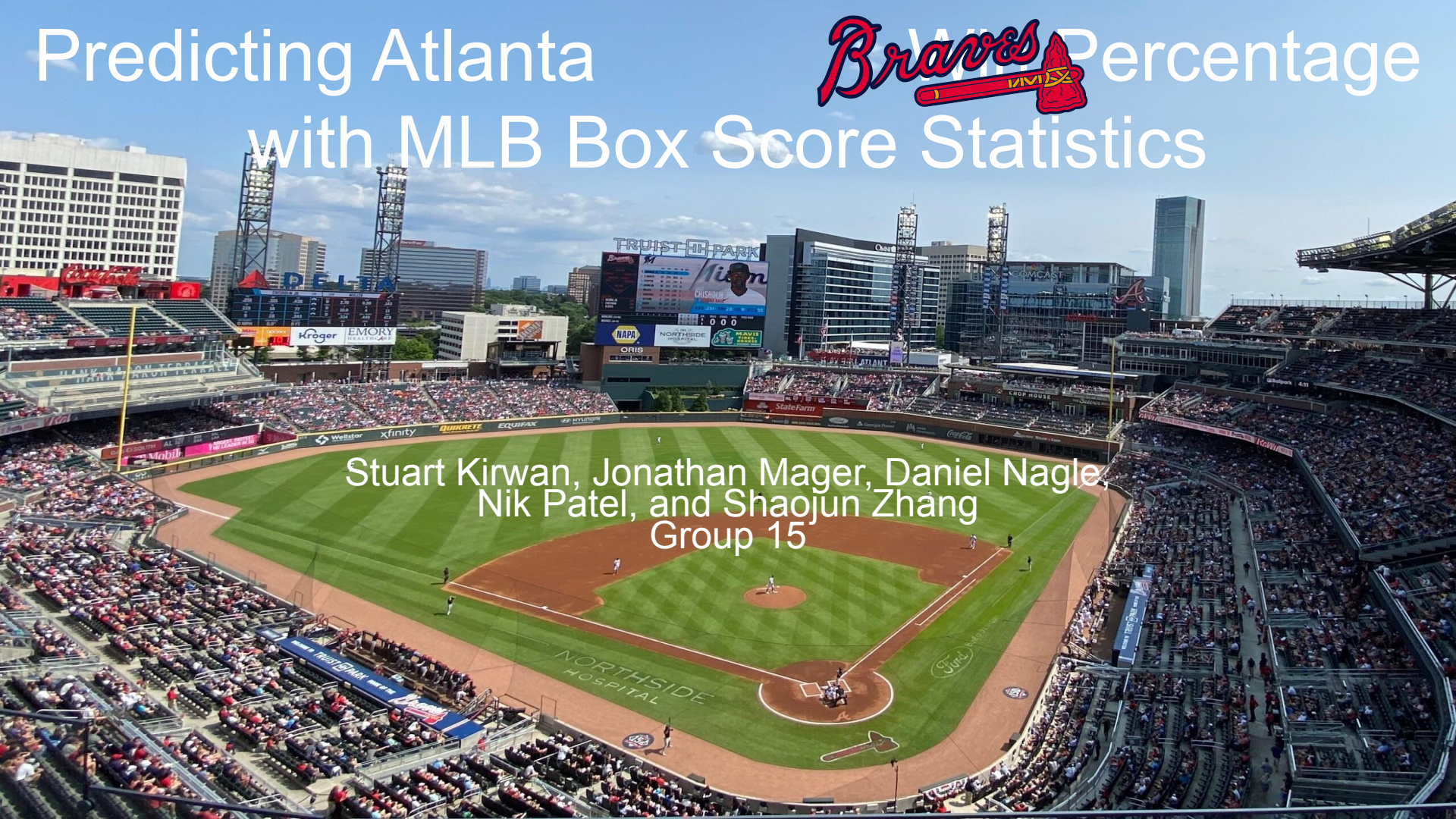
Predicting Atlanta Braves Win Percentage with MLB Box Score Statistics



Stuart Kirwan, Jonathan Mager, Daniel Nagle,
Nik Patel, and Shaojun Zhang
Group 15

Introduction

- In baseball, teams are trying to win as many games as possible which is approximately determined by:

$$WP \approx \frac{RS^2}{(RS^2 + RA^2)}$$

WP = Win Percentage
RS = Runs Scored
RA = Runs Allowed
RD = Run Differential

- RD = RS - RA, heavily influences the above equation
- It is extremely useful for teams to determine what factors leads to scoring more runs and preventing other teams from scoring

Problem Statement

- Which box score metrics have the most influence in predicting run differential for a game?
- Can we predict win probability for any given game based on these data points?
- Can we predict win percentage in a season based on run differential?

> “All models are wrong, but some are useful.” - George Box

Data Description (collection/cleaning/preparation)

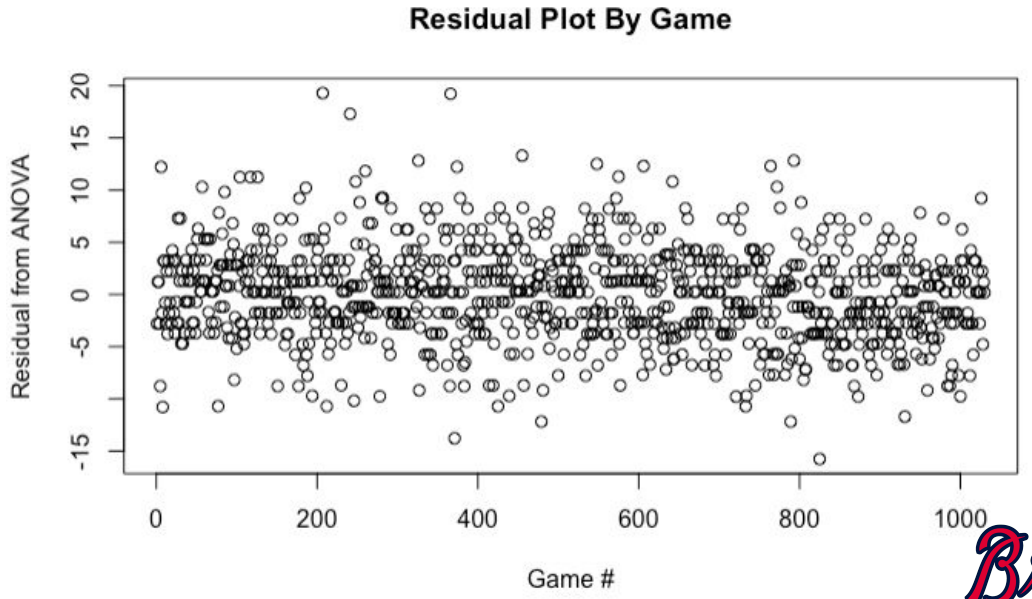
- Collected pitching and batting logs from 2016 through 2022 season
  - Typical season has 162 games
  - Only 60 games played in 2020 due to COVID
  - 161 games played in 2021 due to MLB Players Lockout
- Interested in individual game box score stats like Home Runs, Strikeouts, Earned Run Average
- 1,030 games/data points

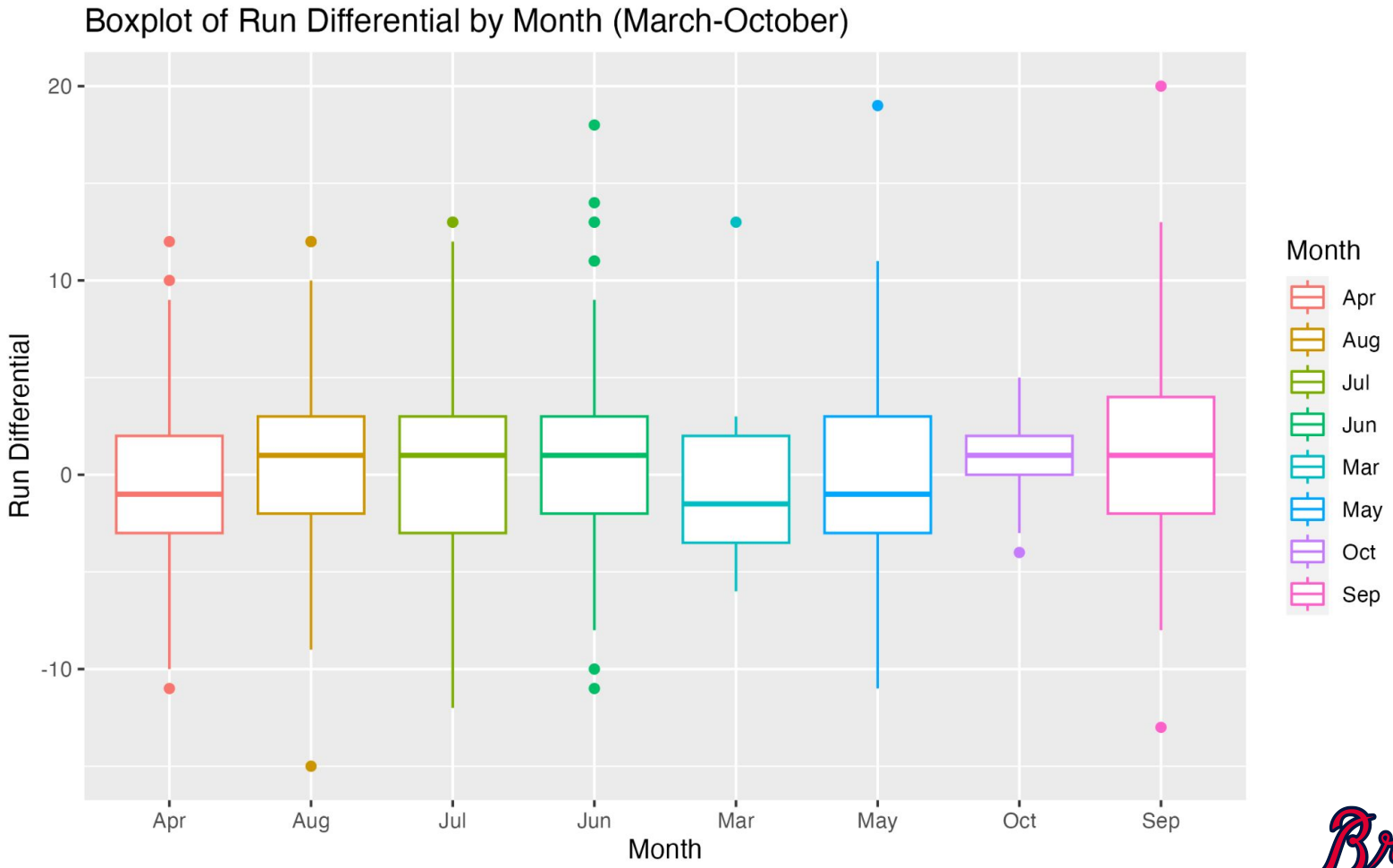
Models

- Linear Regression Model: Run Differential
  - Simple Model: Uses only quantitative metrics
  - Advanced Model: Removed variables that had collinearity and added dummy variables for our categorical columns (Month, Opponent)
- Logistic Model: Binomial Prediction for Win or Loss
  - Instead of predicting a continuous metric such as Run Differential, we can classify our games as a Win with 0 or 1
- Poisson Model: Split separately
  - Pitching model contains only pitching stats to predict Runs Allowed
  - Batting model contains only batting stats to predict Runs Scored
- Regularized Regression:
  - LASSO: Uses L1 regularization to force some coefficients to 0

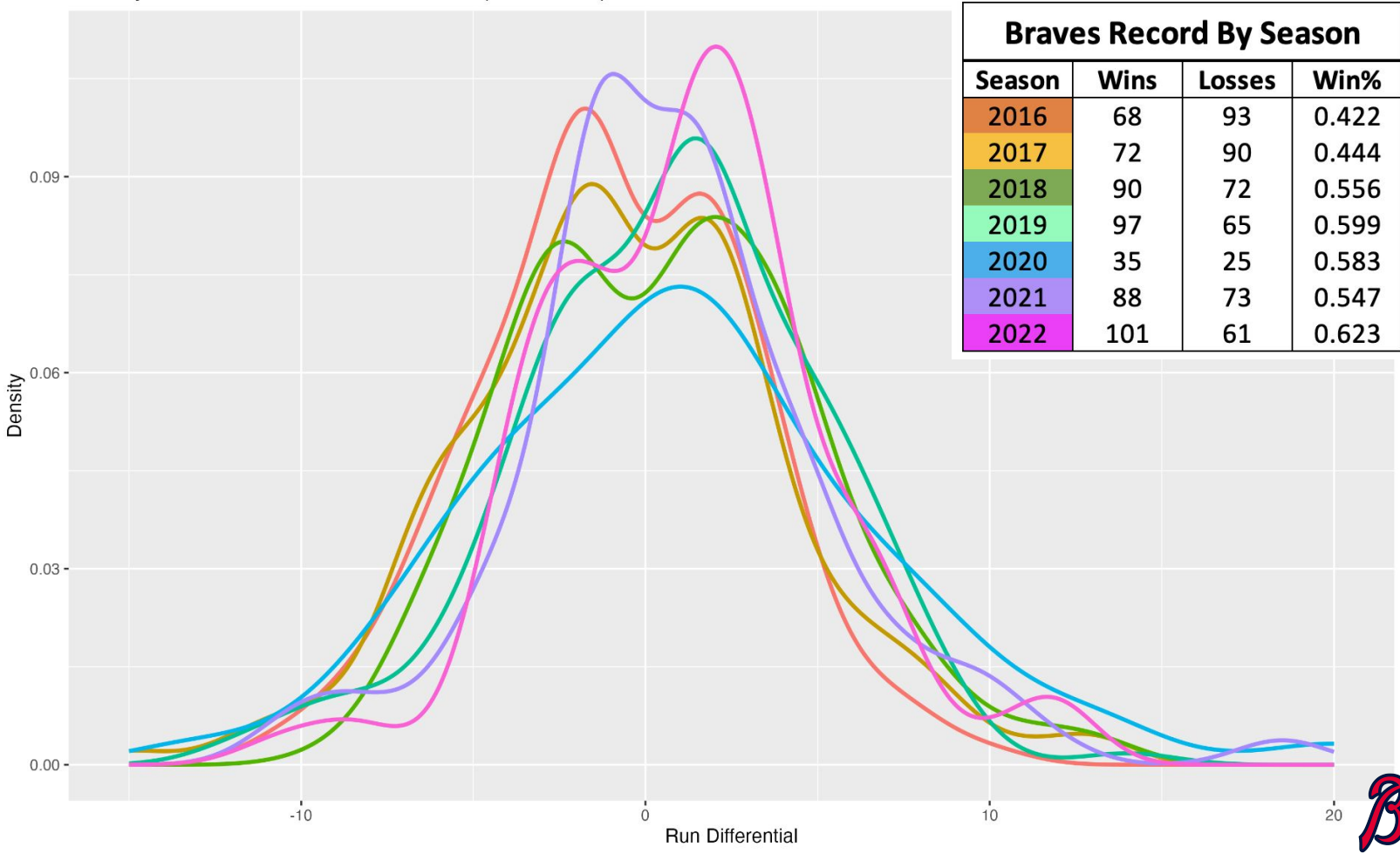
ANOVA of Run Differential by Game

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Month	7	208	29.65	1.486	0.168
Residuals	1022	20395	19.96		




Boxplot of Run Differential by Month (March-October)
Run Differential
20
10
0
-10
Apr
Aug
Jul
Jun
Mar
May
Oct
Sep
Month
Month
Apr
Aug
Jul
Jun
Mar
May
Oct
Sep

Density Plot of Braves Run Differentials (2016-2022)



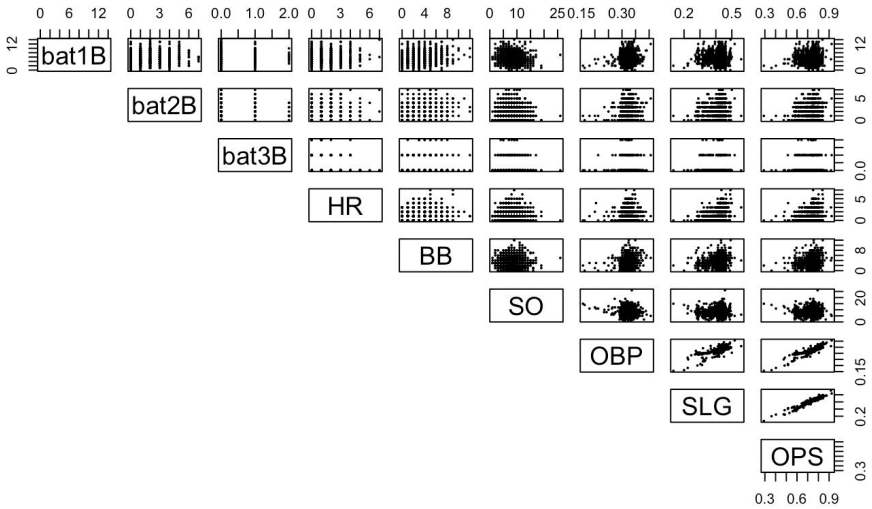
Braves Record By Season

Season	Wins	Losses	Win%
2016	68	93	0.422
2017	72	90	0.444
2018	90	72	0.556
2019	97	65	0.599
2020	35	25	0.583
2021	88	73	0.547
2022	101	61	0.623

Analysis: Basic MLR

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.56725    1.20809  -0.470   0.6388
pitchERA     -0.35838    0.13165  -2.722   0.0066 **
pitchBB      -0.39090    0.03317 -11.786  < 2e-16 ***
pitchSO       0.01865    0.02422   0.770   0.4415
pitch1B      -0.48629    0.02695 -18.044  < 2e-16 ***
pitch2B      -0.72983    0.04995 -14.613  < 2e-16 ***
pitch3B      -1.10936    0.15484  -7.165 1.50e-12 ***
pitchHR      -1.35119    0.06215 -21.742  < 2e-16 ***
bat1B         0.47881    0.02827  16.936  < 2e-16 ***
bat2B         0.78231    0.04919  15.906  < 2e-16 ***
bat3B         0.88289    0.18349   4.812 1.73e-06 ***
HR            1.60608    0.06151  26.112  < 2e-16 ***
BB            0.32117    0.03471   9.251  < 2e-16 ***
SO           -0.05533    0.02373  -2.332   0.0199 *
OBP         -66.04848  139.81417  -0.472   0.6367
SLG         -76.24369  140.06130  -0.544   0.5863
OPS          74.90012  139.97544   0.535   0.5927
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

Residual standard error: 2.186 on 1013 degrees of freedom
Multiple R-squared:  0.765,    Adjusted R-squared:  0.7613
F-statistic: 206.1 on 16 and 1013 DF,  p-value: < 2.2e-16
```

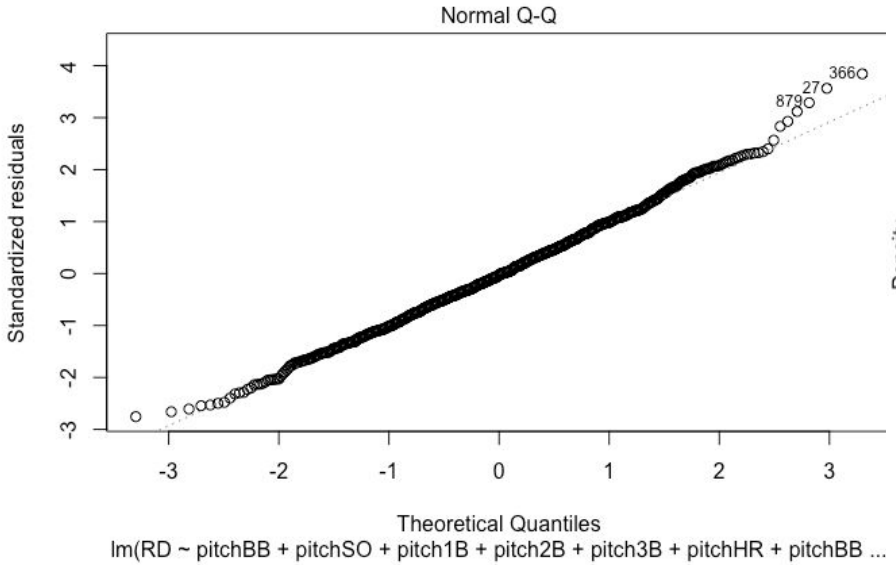
```
pitchERA    pitchBB    pitchSO    pitch1B    pitch2B    pitch3B    pitchHR     bat1B
1.379551   1.055173   1.058285   1.052799   1.042651   1.022690   1.045788  1.093867
   bat2B      bat3B         HR         BB         SO        OBP        SLG       OPS
1.058678   1.018522   1.139255   1.119767   1.099220 1724.885524 8915.826862 16236.844777
```

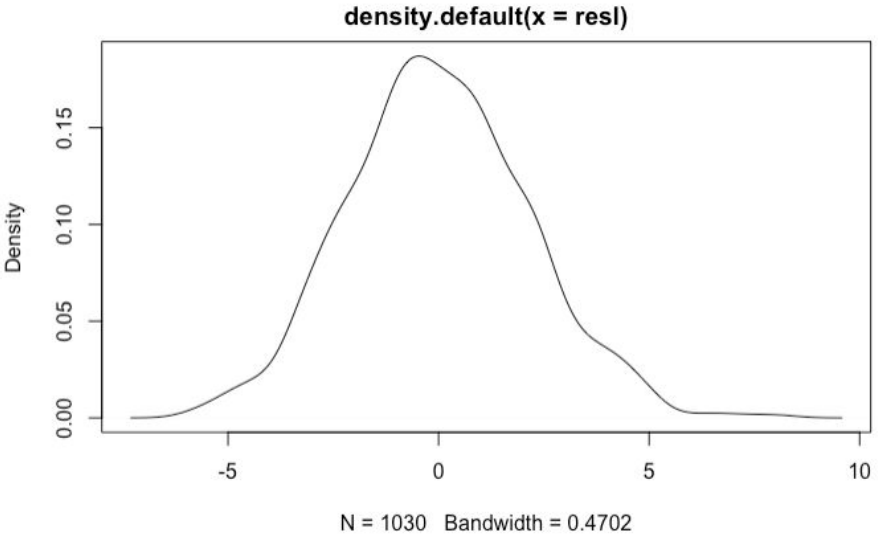


Advanced MLR

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.1644338  0.5510752   0.298   0.7655
pitchBB         -0.3832987  0.0342373 -11.195  < 2e-16 ***
pitchSO          0.0219828  0.0248304   0.885   0.3762
pitch1B         -0.5010635  0.0279902 -17.901  < 2e-16 ***
pitch2B         -0.7307052  0.0514774 -14.195  < 2e-16 ***
pitch3B         -1.1112856  0.1586206  -7.006 4.55e-12 ***
pitchHR         -1.3565004  0.0628660 -21.578  < 2e-16 ***
bat1B            0.4861490  0.0282744  17.194  < 2e-16 ***
bat2B            0.8000501  0.0502669  15.916  < 2e-16 ***
bat3B            0.9236318  0.1892403   4.881 1.23e-06 ***
HR               1.6216109  0.0595613  27.226  < 2e-16 ***
BB               0.3239862  0.0355162   9.122  < 2e-16 ***
SO              -0.0536458  0.0239693  -2.238   0.0254 *
as.factor(Opp)BAL -1.5157185  0.8196242  -1.849   0.0647 .
as.factor(Opp)BOS -0.8936003  0.5720668  -1.562   0.1186
as.factor(Opp)CHC -0.4945878  0.5051351  -0.979   0.3278
as.factor(Opp)CHW -0.9908377  0.9741290  -1.017   0.3093
as.factor(Opp)CIN -1.1170041  0.5010399  -2.229   0.0260 *
as.factor(Opp)CLE -0.6820989  0.9759311  -0.699   0.4848
as.factor(Opp)COL -0.8702265  0.4970939  -1.751   0.0803 .
as.factor(Opp)DET  1.6939164  1.0004296   1.693   0.0907 .
as.factor(Opp)HOU -1.1949211  0.9054741  -1.320   0.1873
as.factor(Opp)KCR -0.7139778  0.9086873  -0.786   0.4322
Home               0.1833173  0.1401919   1.308   0.1913
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.191 on 981 degrees of freedom
Multiple R-squared:  0.7714,    Adjusted R-squared:  0.7602
F-statistic: 68.97 on 48 and 981 DF,  p-value: < 2.2e-16
```

MLR continued





Logistic

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.75008	1.21366	0.618	0.536552	
pitchERA	-0.16120	0.20434	-0.789	0.430191	
pitchBB	-0.33333	0.05539	-6.018	1.76e-09	***
pitchSO	0.07672	0.03700	2.073	0.038135	*
pitch1B	-0.46457	0.04801	-9.677	< 2e-16	***
pitch2B	-0.65392	0.08576	-7.625	2.45e-14	***
pitch3B	-1.45614	0.24909	-5.846	5.04e-09	***
pitchHR	-1.24439	0.12787	-9.732	< 2e-16	***
bat1B	0.40195	0.04938	8.141	3.93e-16	***
bat2B	0.62645	0.08593	7.290	3.10e-13	***
bat3B	1.01983	0.27709	3.680	0.000233	***
HR	1.51054	0.13186	11.455	< 2e-16	***
BB	0.32302	0.05647	5.720	1.06e-08	***
SO	-0.13014	0.03702	-3.515	0.000440	***
as.factor(Opp)BAL	-0.03025	1.11864	-0.027	0.978427	
as.factor(Opp)BOS	-1.18383	0.84872	-1.395	0.163065	
as.factor(Opp)CHC	-0.25259	0.74871	-0.337	0.735841	
as.factor(Opp)CHW	-0.58476	1.33618	-0.438	0.661647	
as.factor(Opp)CIN	-0.90165	0.72629	-1.241	0.214444	
as.factor(Opp)CLE	-0.24663	1.34707	-0.183	0.854733	
as.factor(Opp)COL	-0.86374	0.75060	-1.151	0.249839	
as.factor(Opp)DET	3.28128	1.65424	1.984	0.047305	*
as.factor(Opp)HOU	-1.33424	1.34070	-0.995	0.319647	
as.factor(Opp)KCR	-2.07315	1.32044	-1.570	0.116405	
as.factor(Opp)LAA	0.43039	1.34065	0.321	0.748189	
as.factor(Opp)LAD	-0.38164	0.79058	-0.483	0.629290	
as.factor(Opp)MIA	0.83950	0.61064	1.375	0.169196	
as.factor(Opp)MIL	-0.34459	0.73813	-0.467	0.640616	
as.factor(Opp)MIN	-0.33209	1.29101	-0.257	0.797002	
as.factor(Opp)NYM	0.32043	0.59525	0.538	0.590366	
as.factor(Opp)NYY	-0.37879	0.94388	-0.401	0.688190	
as.factor(Opp)OAK	15.63473	701.16029	0.022	0.982210	
as.factor(Opp)PHI	-0.14303	0.60356	-0.237	0.812679	
as.factor(Opp)PIT	-0.19369	0.75516	-0.256	0.797576	
as.factor(Opp)SDP	0.84102	0.73831	1.139	0.254657	
as.factor(Opp)SEA	-2.16587	1.40112	-1.546	0.122150	
as.factor(Opp)SFG	0.30714	0.72464	0.424	0.671675	
as.factor(Opp)STL	-0.08929	0.75593	-0.118	0.905972	
as.factor(Opp)TBR	-0.08639	0.94326	-0.092	0.927027	
as.factor(Opp)TEX	1.09541	1.50061	0.730	0.465402	
as.factor(Opp)TOR	-0.31085	0.83706	-0.371	0.710372	
as.factor(Opp)WSN	0.20243	0.60530	0.334	0.738056	
as.factor(Month)Aug	0.67618	0.37660	1.795	0.072581	.
as.factor(Month)Jul	0.66190	0.39367	1.681	0.092696	.
as.factor(Month)Jun	0.62137	0.38804	1.601	0.109311	
as.factor(Month)Mar	-0.44709	2.27050	-0.197	0.843897	
as.factor(Month)May	0.60352	0.39841	1.515	0.129816	
as.factor(Month)Oct	1.35780	1.06969	1.269	0.204321	
as.factor(Month)Sep	0.35261	0.37589	0.938	0.348207	
Home	0.10565	0.20529	0.515	0.606784	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1422.85 on 1029 degrees of freedom
Residual deviance: 643.82 on 980 degrees of freedom
AIC: 743.82

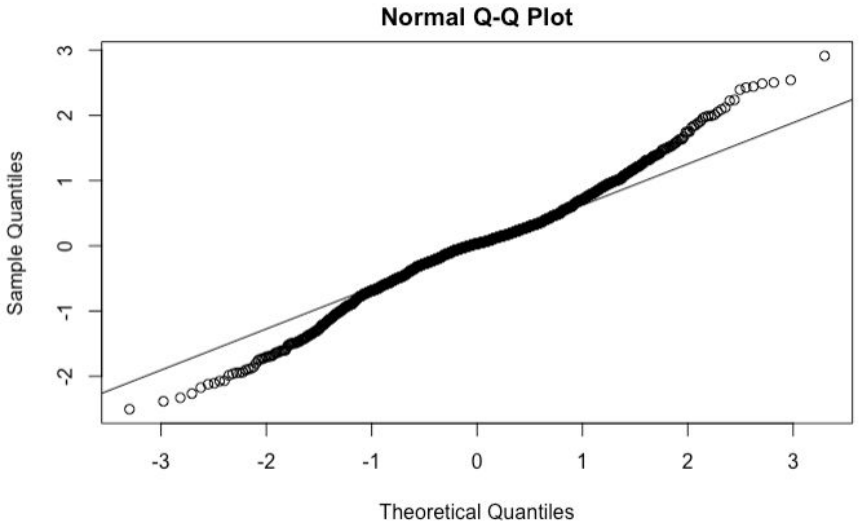
Number of Fisher Scoring iterations: 15

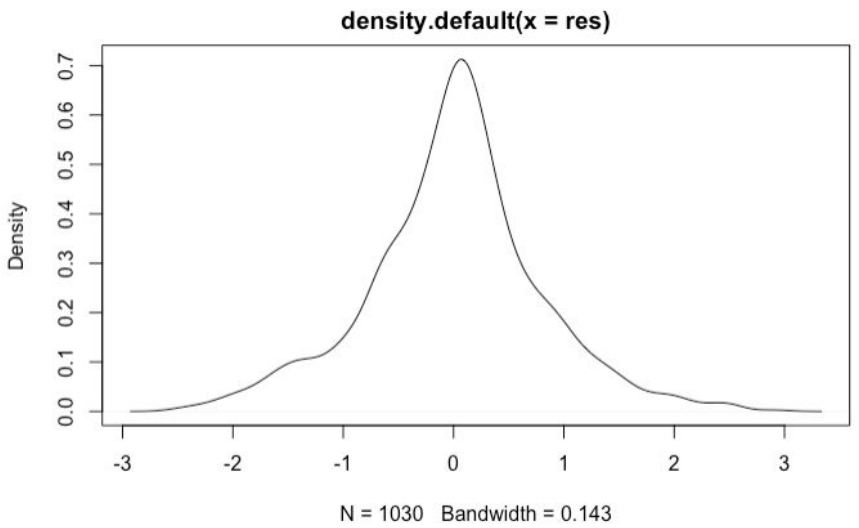
```{r}
1-pchisq((1422.85-640.69),(1029-974))

```

[1] 0

Logistic continued





Poisson - Pitching

- Modeling the rate of runs allowed per game using various measures of pitching performance

```
Call:
glm(formula = pitchR ~ pitchBB + pitchSO + pitch1B + pitch2B +
    pitch3B, family = "poisson", data = ds)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4361  -0.8832  -0.1453   0.6520   3.6557

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.316031   0.063120   5.007 5.53e-07 ***
pitchBB      0.090441   0.006719  13.461  < 2e-16 ***
pitchSO     -0.007401   0.005277  -1.403    0.161
pitch1B      0.100038   0.005144  19.446  < 2e-16 ***
pitch2B      0.147895   0.009267  15.959  < 2e-16 ***
pitch3B      0.246131   0.027987   8.795  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2377.7  on 1029  degrees of freedom
Residual deviance: 1343.8  on 1024  degrees of freedom
AIC: 4463.9
```

Rate Ratios:

```
(Intercept)     pitchBB     pitchSO     pitch1B     pitch2B     pitch3B
  1.3716734   1.0946570   0.9926262   1.1052132   1.1593908   1.2790665
```

Overall Significance:

```
```{r}
# test for overall significance
1-pchisq((2377.7-1343.8),(1029-1024))
```
```

```
[1] 0
```

Poisson - Batting

- Modeling the rate of runs scored per game using various measures of batting performance

```
Call:
glm(formula = R ~ X1B + X2B + X3B + HR + BB + SO, family = "poisson",
    data = ds)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.07185  -0.62689  -0.08552   0.50525   2.37209

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.264315   0.061880   4.271 1.94e-05 ***
X1B          0.084287   0.005373  15.688  < 2e-16 ***
X2B          0.134272   0.009087  14.776  < 2e-16 ***
X3B          0.203957   0.034199   5.964 2.47e-09 ***
HR           0.237992   0.009791  24.306  < 2e-16 ***
BB           0.065952   0.006625   9.955  < 2e-16 ***
SO          -0.010383   0.004950  -2.098   0.0359 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

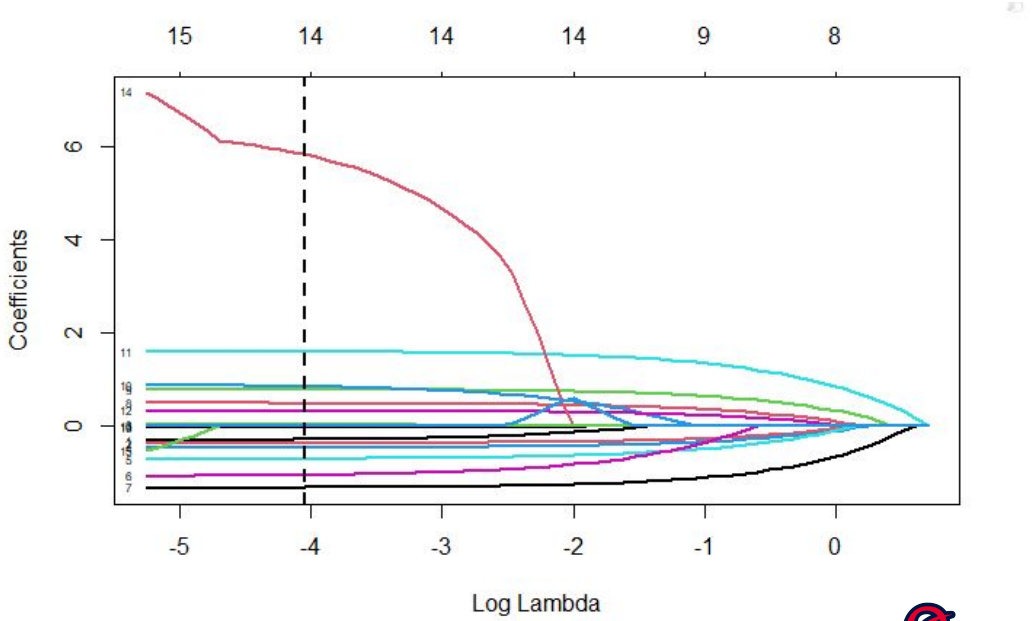
    Null deviance: 2314.37  on 1029  degrees of freedom
Residual deviance:  782.78  on 1023  degrees of freedom
AIC: 4026.5
```

Rate Ratios:

(Intercept)	X1B	X2B	X3B	HR	BB	SO
1.3025380	1.0879412	1.1437043	1.2262450	1.2686995	1.0681755	0.9896706

Overall Significance:

```
```{r}
# test for overall significance
1-pchisq((2314.37-782.78),(1029-1023))
```

[1] 0
```

LASSO Regression

- LASSO took the coefficients of SLG and OPS to 0

(Intercept)	-0.35821324
pitchERA	-0.30458933
pitchBB	-0.38534947
pitchSO	0.01265111
pitch1B	-0.48077774
pitch2B	-0.72019758
pitch3B	-1.07099294
pitchHR	-1.34100448
bat1B	0.47533232
bat2B	0.77622644
bat3B	0.83799207
HR	1.58628829
BB	0.31604371
SO	-0.05252379
OBP	5.83704184
SLG	.
OPS	.



2023 Game by Game Predictions

Wrong Prediction

Game#	Month	Home	Opp	Win	RD	Basic MLR (RD)	Advanced MLR (RD)	Lasso (RD)	Logistic (W/L)
1	Mar	0	WSN	1	5	4	2	4	1
2	Apr	0	WSN	1	6	7	6	7	1
3	Apr	0	WSN	0	-3	1	-1	-1	0
4	Apr	0	STL	1	4	3	3	3	1
5	Apr	0	STL	1	3	4	3	4	1
6	Apr	0	STL	1	3	2	1	2	1
7	Apr	1	SDP	1	1	4	3	4	1
8	Apr	1	SDP	0	-1	1	1	1	1
9	Apr	1	SDP	0	-3	-3	-3	-3	0
10	Apr	1	SDP	0	-8	-5	-5	-5	0
11	Apr	1	CIN	1	1	1	-1	1	0
12	Apr	1	CIN	1	1	3	2	3	1
13	Apr	1	CIN	1	1	3	2	3	1
14	Apr	0	KCR	1	7	8	7	7	1
15	Apr	0	KCR	1	6	3	2	3	1
			Record	**11 - 4**		**13 - 2**	**11 - 4**	**10 - 5**	**11 - 4**
			Accuracy			**0.8667**	**0.8667**	**0.9333**	**0.8667**

Sources

- Bill James Pythagorean Theorem of Baseball https://www.baseball-reference.com/bullpen/Pythagorean_Theorem_of_Baseball#:~:text=The%20Pythagorean%20Theorem%20of%20Baseball,a%20team's%20actual%20winning%20percentage.
- Batting Game Logs: https://www.baseball-reference.com/teams/tgl.cgi?team=ATL&t=b&year=2017
- Pitching Game Logs: https://www.baseball-reference.com/teams/tgl.cgi?team=ATL&t=p&year=2022