What statistic (traditional or advanced sabermetric) can best predict how many runs will be scored by a baseball team?

Word Count:

Design

Research Question: What statistic (traditional or advanced SABERMETRIC) can best predict how many runs will be scored by a player/team?

Aim: To determine how one can best predict how many runs will be scored by a baseball team over the course of a season, based on other facilitating stats.

Hypothesis: A sabermetric stat will best predict how many runs will be scored by a team because they were designed to perform that exact task better than and traditional stat could.

Background: Baseball started as a simple enough game, throw and hit the ball, score more runs to win. But even since the very beginning of baseball, fans, managers, and owners have wanted to track their players and teams. Henry Chadwick is credited with the creation of the very first baseball statistics creating what we now call Batting Average, Runs, and Runs allowed (by a pitcher). This provided the foundation for the development of more statistics such as hits, runs batted in (RBI), etc., but in the early 1970s, a statistical revolution started to take place. The Society for American Baseball Research (SABR) was founded with the purpose of recording the history of baseball, and their members, specifically the famed Bill James, are credited for developing the advanced statistics known as sabermetrics. He was able to dig beneath the surface of the seemingly simple game of baseball and develop stats that not only track player and team performance, but also factor in things such as good/bad luck, poor defense for pitchers, and opponent quality. This greatly increased the quality of and ability to analyze performance in a baseball game.

Personal Engagement: I have always been interested in baseball, and recently have explored advanced statistics. The way advanced statistics dig deeper into the surface of what otherwise seems to be such a simple game creates a whole new way to enjoy the game. Investigating how different stats are related to each other is very interesting and will let me, and others, better understand how projection models work.

Data Collection: BASEBALL-REFERENCE.COM is a website that provides hundreds of baseball statistics that covers nearly every game from baseball's inception to today. Of use in this investigation, it contains a database of yearly total team stats, organized nicely in a table for comparison. For this experiment the complete team-by-team stats for the year 2019 were downloaded to be used for analysis.

Data: The full table of data is available in the appendix.

Math

Equations for sabermetrics stats:

BRA: Batter's Runs Average

A way to measure a players ability to score (OBP) and drive in (SLG) runs.

$$BRA = OBP * SLG$$

TA: Total Average

A way to measure how many bases a player touches, thus possibly advancing any runners ahead of the batter, making run scoring possible.

$$TA = \frac{TB + BB}{AB - H}$$

RC: Runs Created

A way to estimate the number of runs a hitter contributed to his team, but in this case it is applied to total team stats.

$$RC = \frac{TB * (H + BB)}{AB + BB}$$

BR: Base Runs (Simple)

Is a way to estimate how many runs a team "should have scored" given their offensive statistics.

$$BR = \frac{(H + BB - HR)((1.4 * TB - .6 * H - 3 * HR + .1 * BB) * 1.02)}{((1.4 * TB - .6 * H - 3 * HR + .1 * BB) * 1.02)(AB - H)} + HR$$

BR: Base Runs (Advanced)

This calculated the same as above, but considers more statistics that may impact run scoring.

$$BR = \frac{(H + BB - HR)((1.4 * TB - .6 * H - 3 * HR + .1 * (BB + HBP - IBB) + .9 * (SB - CS - GIDP}{((1.4 * TB - .6 * H - 3 * HR + .1 * (BB + HBP - IBB) + .9 * (SB - CS - GIDP)) * 1.1) + (AB - H} + HR$$

Overview of Analysis: In order to determine which statistic best predicts the number of runs that will be scored by a team, there are a couple steps that will be performed in order to do this. First, correlation analysis is performed between every variable given in the table obtained from BASEBALL-REFERENCE.COM, in order to determine which stats may be a good indicator of how many runs will be scored. Once all the correlation coefficients are calculated, a value close to one is ideal. This is a good indicator that there is a good relationship between the given statistic and runs, but not always. For example, just because the number of batters used by a team

and the number of runs they score have an interestingly high correlation coefficient, it doesn't mean that they have any influence over each other, because they really do not. Next, regression analysis is performed to model and calculate the relationship between the two variables with a high correlation coefficient. With this we'll also be able to produce a residual plot with a trend line, which we'll use to determine which statistic might best predict how many runs are scored.

Sample Math:

Correlation Coefficient:

The equation to calculate the r-value is: $r_{xy} = \frac{n(\Sigma xy)-(\Sigma x)(\Sigma y)}{\sqrt{(n\Sigma x^2-(\Sigma x^2))(n\Sigma y^2-(\Sigma y^2)}}$, and an example of working it out is as follows.

This calculation will determine the correlation between the RBI statistics and the number of runs a team scored, first a table is set up (example below) with the teams labeled, and the two statistics that are being tested are put in their respective columns (labeled both by their name and x and y respectively). Next, for each team multiply their RBI and Run values together, then square each individually and put them in their own columns. Next, sum up the totals in every column (except the team) and put in the final row of the table.

Team	Runs (x)	RBI (y)	xy	x^2	y^2
1	813	778	632514	660969	605284
2	855	824	704520	731025	678976
3	729	698	508842	531441	487204
4	901	857	772157	811801	734449
5	814	783	637362	662596	613089
6	708	676	478608	501264	456976

7	701	679	475979	491401	461041
8	769	731	562139	591361	534361
9	835	803	670505	697225	644809
10	582	556	323592	338724	309136
11	920	891	819720	846400	793881
12	691	655	452605	477481	429025
13	769	734	564446	591361	538756
14	886	861	762846	784996	741321
15	615	593	364695	378225	351649
16	769	744	572136	591361	553536
17	939	906	850734	881721	820836
18	791	767	606697	625681	588289
19	943	904	852472	889249	817216
20	845	800	676000	714025	640000
21	774	742	574308	599076	550564
22	758	722	547276	574564	521284
23	682	652	444664	465124	425104
24	758	730	553340	574564	532900
25	678	655	444090	459684	429025
26	764	714	545496	583696	509796
27	769	730	561370	591361	532900
28	810	765	619650	656100	585225
29	726	697	506022	527076	485809
30	873	824	719352	762129	678976

Totals (Σ)	23467	22471	17804137	18591681	17051417

Plugging these values into our equation (where $n = 30$) gives us:

$$r = \frac{30(17804137) - (23467 * 22471)}{\sqrt{(30(18591681) - (23467)^2)(30(17051417) - (22471)^2)}}$$

$$r = 0.9966889372$$

Next, we must calculate r^2which is "the proportion of the variance in the dependent variable that is predictable from the independent variable" and will allow us to determine how well RBI can predict how many runs are scored. Calculating $r^2$is simply in its name, we will square the r value obtained from the previous equation.

$$r^2 = 0.9966889372^2 = 0.99338883753$$

With these results, we can determine that the number of RBI a team has and the number of runs scored have a very high correlation coefficient, and are very likely to influence each other, and the r^2value indicates that the residual plot and regression model (generated by Excel) might be a good indication for predicted values.

Results

Results

After calculating the R-Value between all possible statistics and the number of runs scored, I decided to further examine the statistics, number of bats (r-value = 0.999351035), plate appearances(0.999820513), RBI (0.9999984), runs created (0.999986406), and base runs (advanced) (0.999987485). The full table of r-values is available in the appendix. I chose all of

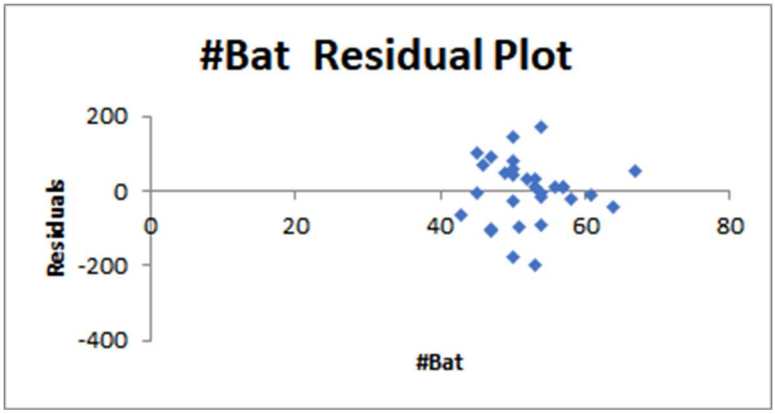
these because of their very high r-values, but the "number of bats" statistic specifically interested me, and will be discussed further later.

Interpreting Graphs

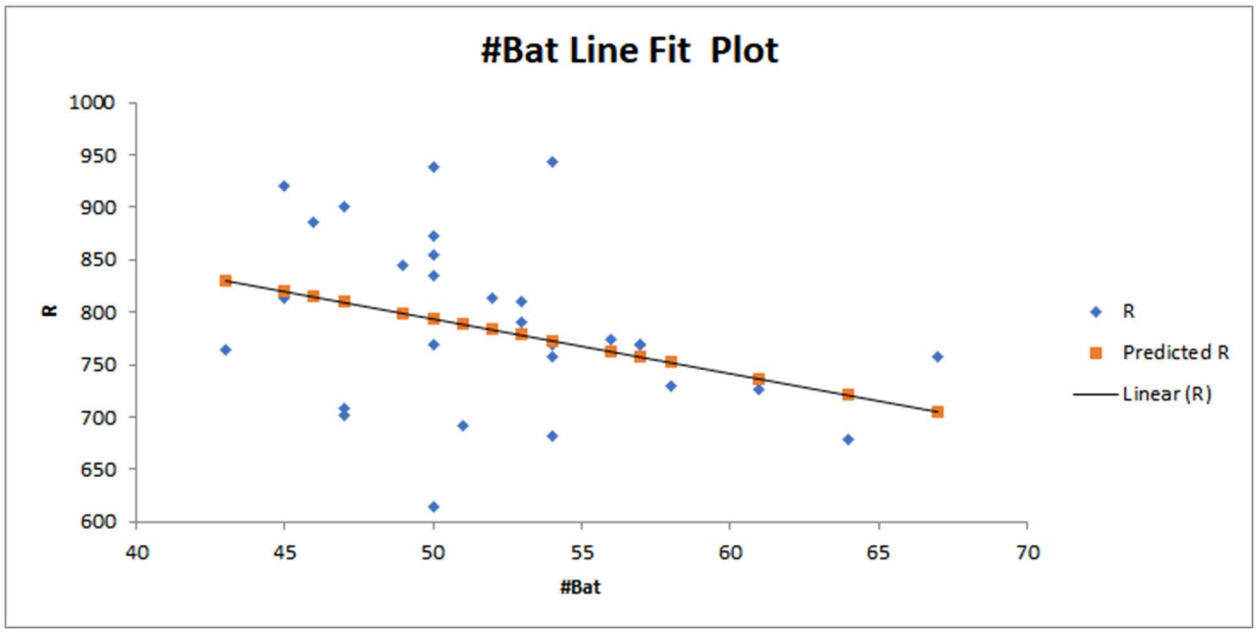
When interpreting the graphs generated by excel there are certain things that must be considered. For residual plots, the plots should look fairly "random", as in the plots don't seem to form a shape that might indicate an ability to predict the next value. For the line fit plot that displays the predicted run values versus the actual values. Looking at these graphs and the distance between the predicted values and the fitted line, we can determine how well the run values were predicted.

Number of Bats

# of bats produced an r-value of 0.999351035 and an r^2value of 0.99870249115, which made it a likely candidate to best predict the number of runs scored. This statistic interested me because a team that uses more batters might indicate that either a trade happened to bring in a better batter, thus increasing the chance that more runs will be scored. When modeled on a residual plot it produced:



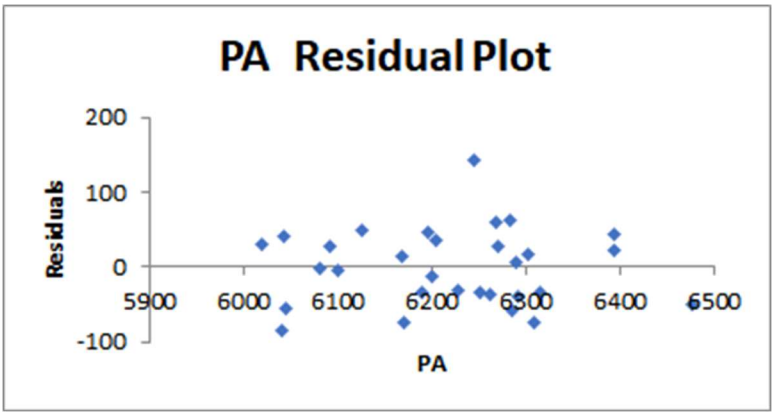
And when plotted on a regression analysis plot showing both the true amount of runs scored by a team, as well as the predicted runs amount based on the # of bats value it produced:



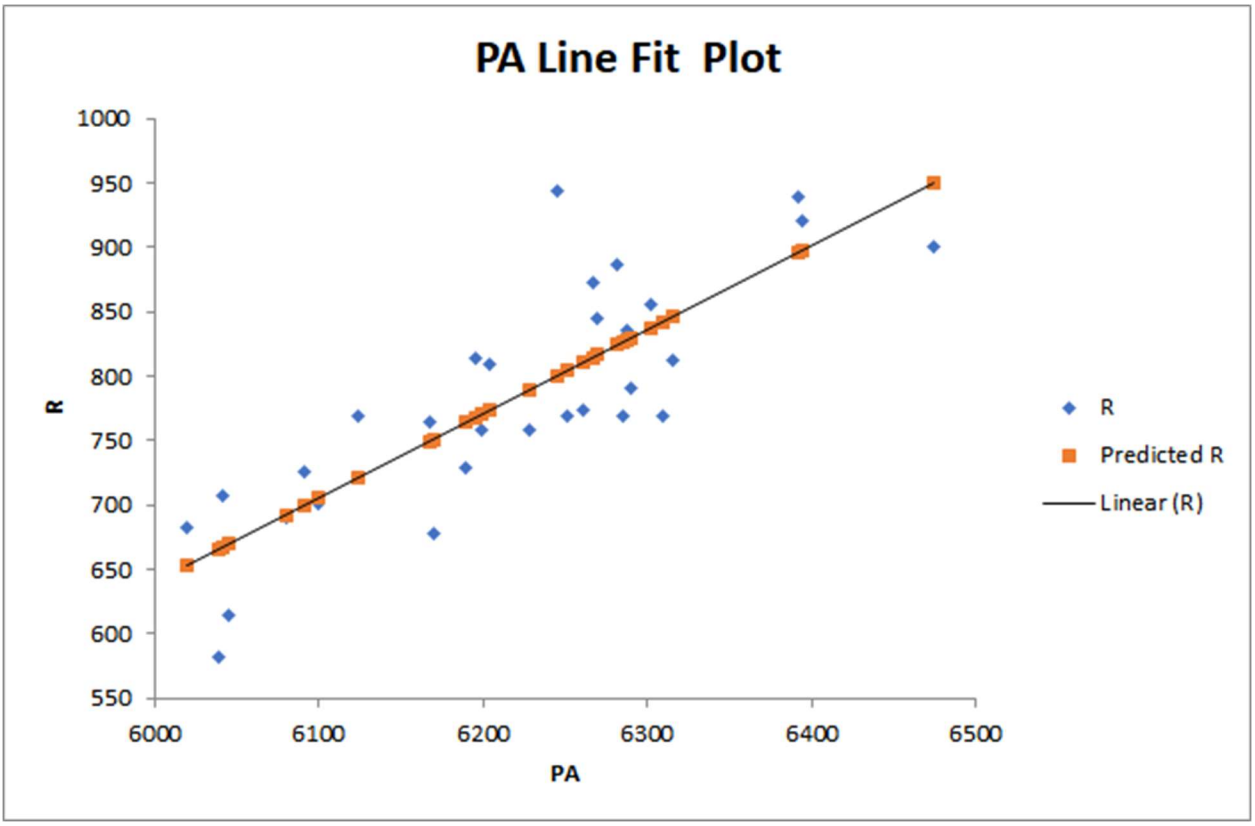
The results for this statistic show that my specific interest in it did not result in any significant results.

Plate Appearances

Plate Appearances produced an r-value of 0.999820513 and an r^2value of 0.99964105821. This statistic interested me because the more plate appearances a team has, the more chances they have to drive in runs. When modeled on a residual plot it produced:



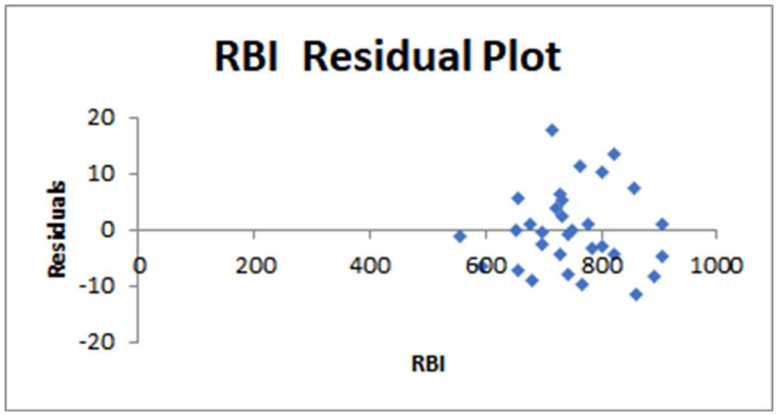
And when plotted on a regression analysis plot showing both the true amount of runs scored by a team, as well as the predicted runs amount based on the Plate Appearances value it produced:



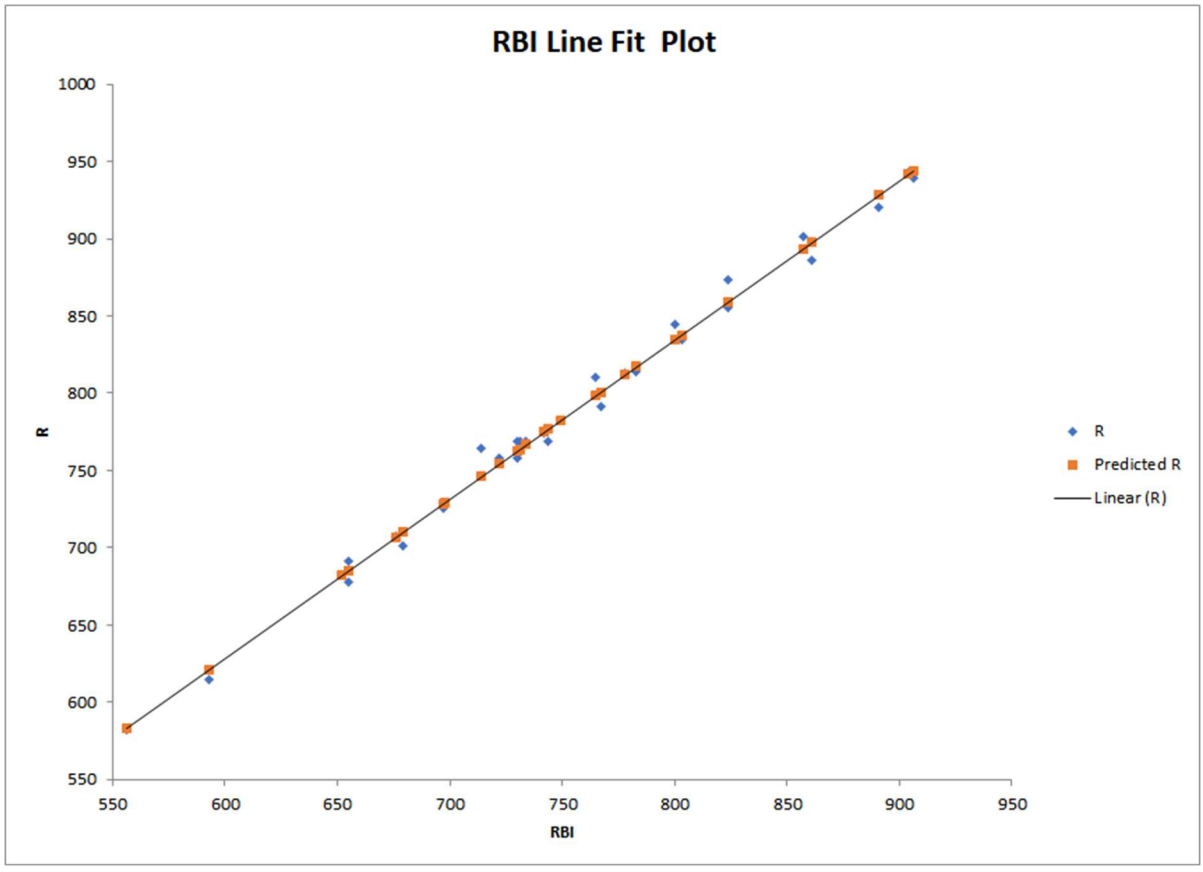
RBI

RBI produced an r-value of 0.9999984, and an r^2value of 0.9999968. This statistic interested me because it is the defacto statistic to consider when looking at how many runs will be scored.

When modeled on a residual plot it produced:



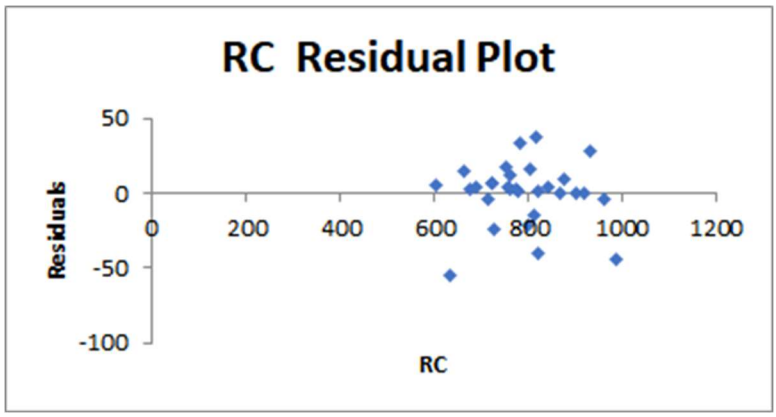
And when plotted on a regression analysis plot showing both the true amount of runs scored by a team, as well as the predicted runs amount based on the RBI value it produced:



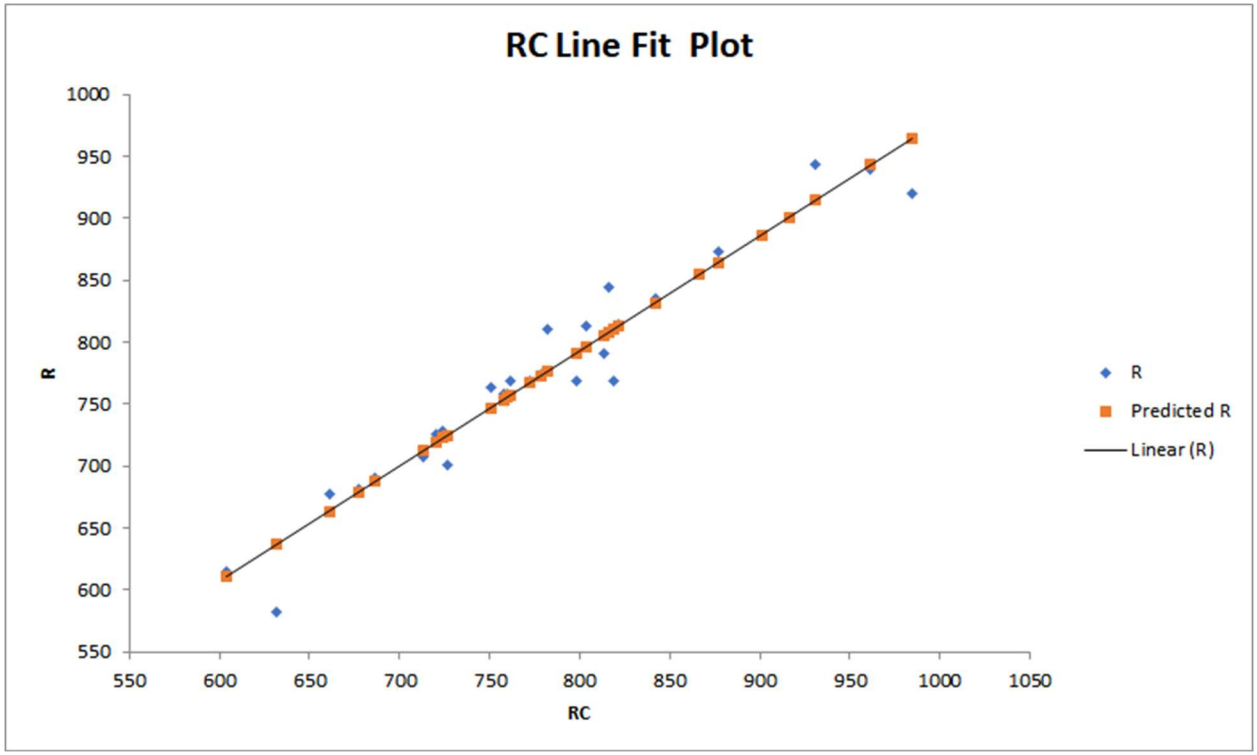
Runs Created

Runs Created produced an r-value of 0.999986406, and an r^2value of 0.99997281218. This statistic (along with the other sabermetric stat Base Runs) interested me because they were

developed with the purpose of predicting how many runs should be scored. When modeled on a residual plot it produced:



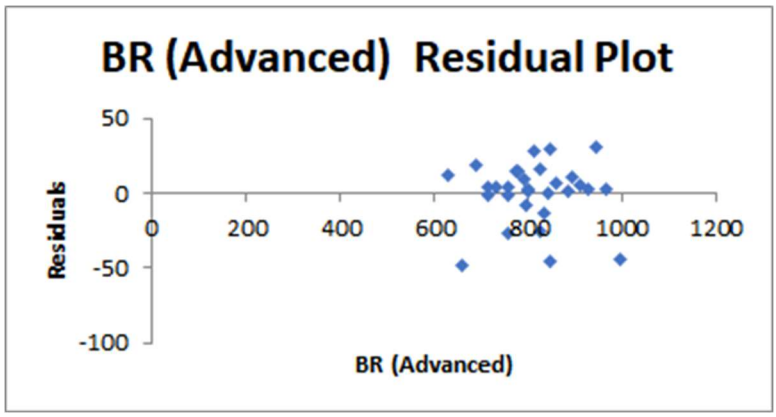
And when plotted on a regression analysis plot showing both the true amount of runs scored by a team, as well as the predicted runs amount based on the Runs Created value it produced:



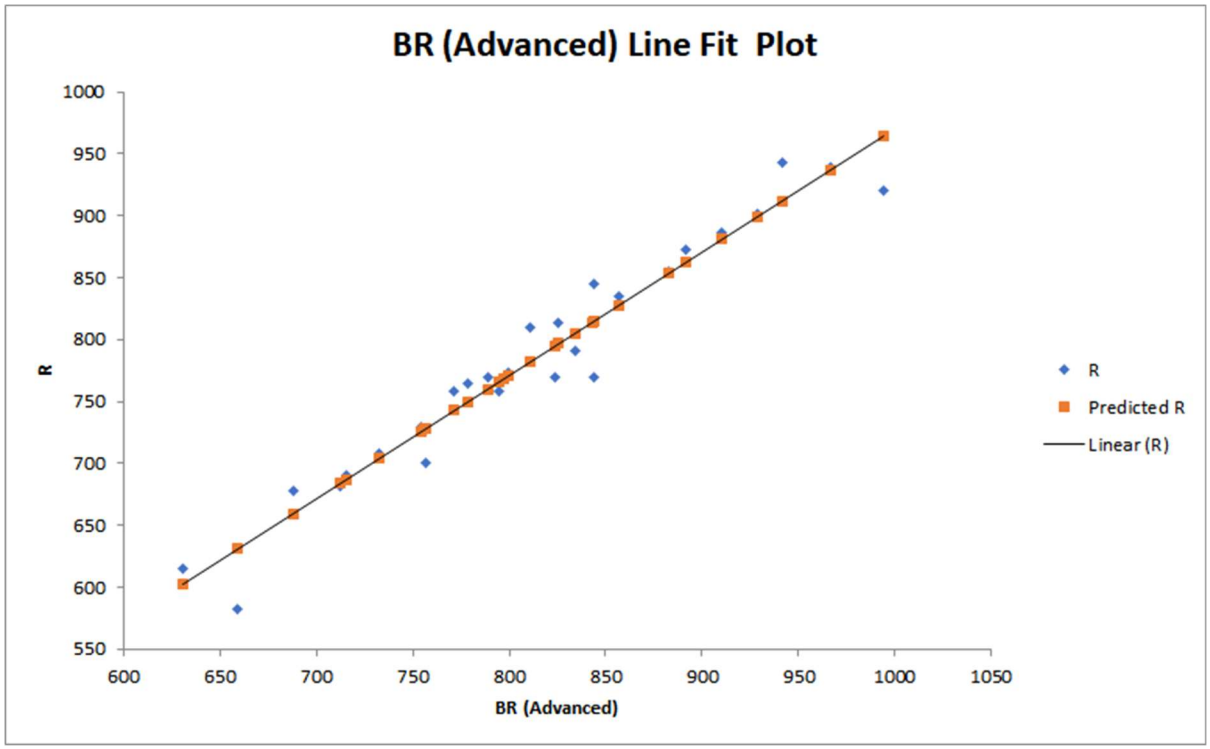
Base Runs (Advanced)

Base Runs (Advanced) produced an r-value of 0.999987485 and an r^2 value of 0.99997497015. This statistic (along with the other sabermetric stat Runs Created) interested me because they

were developed with the purpose of predicting how many runs should be scored. When modeled on a residual plot it produced:



And when plotted on a regression analysis plot showing both the true amount of runs scored by a team, as well as the predicted runs amount based on the Base Runs (Advanced) value it produced:



Review of Results

Based on the results three main candidates appear: RBI, runs created, and base runs (advanced). These fit the criteria of having a very good r-value and r^2value, but also excelled in their predictions for runs scored in their regression analysis tests. They were able to very closely predict the number of runs scored in all cases. The question of which is best cannot be determined solely on that however, a more in depth analysis of the stats must be done. RBI only accounts for plays where the batter directly drives in a run (e.g. a hit that allows the runner on 3rd base to score), but as described earlier, Runs Created and Base Runs (Advanced) account for many factors including opportunities to drive in runs, the number of bases touched that may advance runners in front of the batter, as well as hits that can directly drive runs in to score. Because of this, Base Runs (Advanced) is the best statistic to predict how many runs will be scored. This is due to both it's very good accuracy in predicting, as well as the variables that are taken into account which take into consideration all factors that come into play when batting in baseball.

Reflection

On Results

Going into this investigation I expected that one of the sabermetric statistics would be the best at predicting the number of runs scored, however I was unsure of which one. I thought this because they were specifically designed to dig deeper into the game of baseball than the traditional more surface level stats by considering all possible factors and variables that affect the game. This gave me the confidence to think that it would predict better than other stats. My conclusion ended up lining up with my hypothesis.

On Validity

The results are valid, however they have their limitations. I only used data from the 2019 season, which is very significant because the run changing environment changes between seasons due to changes in rules, ability of players, and changes in equipment (specifically the ball). This means that the results of this investigation won't necessarily remain true between seasons, but because we determined a sabermetric stat to be the best (which as said multiple times before takes in multiple important variables that other statistics do not), it is a possibility that the Base Runs (Advanced) stat will be a good predictor of runs scored in past and future seasons.

Bibliography

Sports-Reference. “Baseball-Reference Bullpen Main Page.” *BR Bullpen*, 2020, www.baseball-reference.com/bullpen/.

Sports-Reference. “2019 Major League Baseball Season Summary.” *Baseball*, 2019, www.baseball-reference.com/leagues/MLB/2019.shtml.

Appendix

2019 Batting Data Table:

Tm	#Bat	BatAge	R/G	G	PA	AB	R	H
ARI	45	28.7	5.02	162	6315	5633	813	1419
ATL	50	28	5.28	162	6302	5560	855	1432
BAL	58	26.5	4.5	162	6189	5596	729	1379
BOS	47	27.3	5.56	162	6475	5770	901	1554
CHC	52	27.7	5.02	162	6195	5461	814	1378
CHW	47	27.6	4.4	161	6042	5529	708	1443
CIN	47	27.8	4.33	162	6100	5450	701	1328
CLE	54	27.7	4.75	162	6124	5425	769	1354
COL	50	28.2	5.15	162	6288	5660	835	1502
DET	53	27.6	3.61	161	6039	5549	582	1333
HOU	45	29	5.68	162	6394	5613	920	1538
KCR	51	27.6	4.27	162	6080	5496	691	1356
LAA	57	28.8	4.75	162	6251	5542	769	1368
LAD	46	27.9	5.47	162	6282	5493	886	1414
MIA	50	28.4	3.8	162	6045	5512	615	1326
MIL	50	28.9	4.75	162	6309	5542	769	1366
MIN	50	27.8	5.8	162	6392	5732	939	1547

NYM	53	27.9	4.88	162	6290	5624	791	1445
NYY	54	28.3	5.82	162	6245	5583	943	1493
OAK	49	27.8	5.22	162	6270	5561	845	1384
PHI	56	27.7	4.78	162	6261	5571	774	1369
PIT	54	27.5	4.68	162	6228	5657	758	1497
SDP	54	26.2	4.21	162	6019	5391	682	1281
SEA	67	27.8	4.68	162	6199	5500	758	1305
SFG	64	29.9	4.19	162	6170	5579	678	1332
STL	43	28.8	4.72	162	6167	5449	764	1336
TBR	57	27.2	4.75	162	6285	5628	769	1427
TEX	53	28.8	5	162	6204	5540	810	1374
TOR	61	25.9	4.48	162	6091	5493	726	1299
WSN	50	28.8	5.39	162	6267	5512	873	1460
LgAvg	47	27.9	4.83	162	6217	5555	782	1401
Totals	1410	27.9	4.83	4858	186518	166651	23467	42039

2B	3B	HR	RBI	SB	CS	BB	SO	BA
288	40	220	778	88	14	540	1360	0.252
277	29	249	824	89	28	619	1467	0.258
252	25	213	698	84	30	462	1435	0.246
345	27	245	857	68	30	590	1382	0.269
270	26	256	783	45	24	581	1460	0.252
260	20	182	676	63	28	378	1549	0.261
235	27	227	679	80	38	492	1436	0.244

286	18	223	731	103	35	563	1332	0.25
323	41	224	803	71	31	489	1503	0.265
292	41	149	556	57	20	391	1595	0.24
323	28	288	891	67	27	645	1166	0.274
281	40	162	655	117	39	456	1405	0.247
268	21	220	734	65	20	586	1276	0.247
302	20	279	861	57	10	607	1356	0.257
265	18	146	593	55	30	395	1469	0.241
279	17	250	744	101	25	629	1563	0.246
318	23	307	906	28	21	525	1334	0.27
280	17	242	767	56	27	516	1384	0.257
290	17	306	904	55	22	569	1437	0.267
292	23	257	800	49	21	578	1338	0.249
311	26	215	742	78	18	562	1453	0.246
315	38	163	722	64	29	425	1213	0.265
224	24	219	652	70	37	504	1581	0.238
254	28	239	730	115	47	588	1581	0.237
300	26	167	655	47	28	475	1435	0.239
246	24	210	714	116	29	561	1420	0.245
291	29	217	730	94	37	542	1493	0.254
296	24	223	765	131	38	534	1578	0.248
270	21	247	697	51	20	509	1514	0.236
298	27	231	824	116	29	584	1308	0.265
284	26	226	749	76	28	530	1427	0.252
8531	785	6776	22471	2280	832	15895	42823	0.252

OBP	SLG	OPS	OPS+	TB	GDP	HBP	SH	SF
0.323	0.434	0.757	94	2447	120	70	31	40
0.336	0.452	0.789	98	2514	104	60	25	35
0.31	0.415	0.725	90	2320	111	71	22	37
0.34	0.466	0.806	107	2688	127	49	20	44
0.331	0.452	0.783	102	2468	127	83	30	39
0.314	0.414	0.728	93	2289	114	66	36	32
0.315	0.422	0.736	86	2298	111	89	30	33
0.323	0.432	0.756	95	2345	110	50	40	46
0.326	0.456	0.782	88	2579	111	43	51	43
0.294	0.388	0.682	78	2154	108	48	9	42
0.352	0.495	0.848	118	2781	146	66	10	57
0.309	0.401	0.71	86	2203	113	59	24	42
0.324	0.422	0.746	97	2338	143	67	4	42
0.338	0.472	0.81	112	2593	100	81	55	45
0.298	0.375	0.673	79	2065	139	73	31	33
0.329	0.438	0.767	97	2429	120	72	20	38
0.338	0.494	0.832	119	2832	101	81	10	41
0.328	0.442	0.77	106	2485	129	95	28	27
0.339	0.49	0.829	118	2735	113	49	10	33
0.327	0.448	0.776	108	2493	140	87	7	36
0.319	0.427	0.746	91	2377	97	57	34	34
0.321	0.42	0.741	94	2377	119	63	47	34
0.308	0.41	0.718	90	2210	120	55	37	31
0.316	0.424	0.74	100	2332	83	58	14	37
0.302	0.392	0.694	83	2185	111	50	24	42

0.322	0.415	0.737	93	2260	110	76	40	39
0.325	0.431	0.757	102	2427	114	73	8	34
0.319	0.431	0.75	88	2387	98	67	17	44
0.305	0.428	0.733	94	2352	107	45	14	28
0.342	0.454	0.796	102	2505	117	81	48	42
0.323	0.435	0.758	97	2416	115	66	26	38
0.323	0.435	0.758	97	72468	3463	1984	776	1150

IBB	LOB	BRA	TA	RC	BR (Advanced)	BR (Simple)
36	1119	0.140182	0.723321	804.0154	825.7404817	787.4399
39	1138	0.151872	0.770423	866.4621	882.8845605	829.4906
8	1063	0.12865	0.673933	724.3298	754.3025339	685.1403
36	1170	0.15844	0.776355	916.3793	928.4157234	935.6208
33	1071	0.149612	0.750354	821.6538	842.9756759	768.0455
13	1071	0.129996	0.661306	713.246	732.2042584	717.0479
25	1073	0.13293	0.692812	726.9083	757.0528332	668.2512
30	1072	0.139536	0.726044	772.0175	796.9308706	747.6079
25	1075	0.148656	0.74	841.7444	856.8625676	858.1063
14	1069	0.114072	0.610037	631.7544	658.92846	665.3818

					31	
17	1168	0.17424	0.837806	984.6041	994.0975018	961.8426
17	1056	0.123909	0.660531	686.7061	715.3174613	715.9987
29	1125	0.136728	0.704635	761.5269	788.78465	745.0675
47	1124	0.159536	0.796849	901.1096	910.2148125	817.6807
16	1034	0.11175	0.594259	603.4332	630.7019312	629.2867
42	1180	0.144102	0.747744	818.4041	843.8983021	761.6251
21	1115	0.166972	0.804736	961.6748	966.5489682	868.9293
34	1128	0.144976	0.727105	813.3873	834.0805323	768.4572
18	1039	0.16611	0.806627	931.0239	941.4198516	844.2629
17	1081	0.146496	0.739281	815.9509	844.1479744	763.4939
47	1129	0.136213	0.712069	778.6628	799.9709922	759.8854
41	1103	0.13482	0.679898	757.879	771.6425857	823.0484
19	1008	0.12628	0.665339	677.3837	712.0408663	634.205
7	1080	0.133984	0.715145	759.4484	794.5838767	693.8564
26	1069	0.118384	0.628591	660.887	687.9129289	684.5443
15	1107	0.13363	0.708608	750.5924	778.4610755	711.9

20	1130	0.140075	0.720588	798.0471	823.4750432	786.6536
18	1066	0.137489	0.725012	782.337	810.8417943	744.7308
10	1003	0.13054	0.684332	720.5037	756.3789256	653.7029
33	1114	0.155268	0.782754	876.7472	891.5775169	861.9805
25	1093	0.140505	0.718641	786.7225	810.4893481	761.0746
753	32780	0.140505	0.718557	23599.23	24311.14409	22833.65

Correlation Coefficient Table

Statistic	Correlation Coefficient with R(uns)
#Bat	0.999351035
BatAge	-0.003351979
R/G	0.021279561
G	0.999766988
PA	0.999820513
AB	0.999789042
R	1
H	0.999878671
2B	0.99981476
3B	0.998207309
HR	0.999799614
RBI	0.9999984

SB	0.997583768
CS	0.997928255
BB	0.999832489
SO	0.999510514
BA	0.014057518
OBP	0.027578845
SLG	0.022744476
OPS	0.02418011
OPS+	0.020029431
TB	0.999958058
GDP	0.999513689
HBP	0.999099381
SH	0.994366496
SF	0.999532526
IBB	0.996643491
LOB	0.999817249
BRA	0.020852145
TA	0.019860142
RC	0.999986406
BR (Advanced)	0.999987485
BR (Simple)	0.999940973