Base-Advance Average

Gary M. Hardegree
Department of Philosophy
University of Massachusetts
Amherst, MA 01003

1. Introduction

In what follows, I propose a new performance measure ("stat") for baseball, which I call *base-advance average*, which I compare with conventional stats, including batting average, on-base percentage, slugging percentage, and OPS.

My comparisons are based on play-by-play records of all the regular season games currently available at Retrosheet, which total 74,255 games, and include the years 1960-1992 and 2000-2004 (alas, the years 1993-1999 are still unavailable to the public).

In order to facilitate the comparison, I further propose a simple way to evaluate stats called *win-tracking rate*. As it turns out, among the conventional stats, the best is OPS, which tracks wins at a rate of 84%, and the worst is hits, which tracks wins at a rate of 71%. But the conventional stats are all eclipsed by base-advance average, which tracks wins at a rate of 95%.

Along the way, by way of illustration, I evaluate various players and teams with respect to base-advance average. The evaluation of teams leads to further statistical comparisons. Here, the results are equally decisive in favor of base-advance average. For example, over the period 2000-2004, base-advance average for-and-against predicts wins just as well as runs for-and-against. This is because, over this period, base-advance average correlates with runs at a "rate" of .988!

2. Evaluating Stats

Baseball comes with a vast array of performance measures ("stats"), some of which are enshrined, such as batting-average and earned-run-average, and others of which are less prestigious, but which are nevertheless useful in managing and appreciating the game.

Stats are used to evaluate and compare players and teams, but how do we evaluate and compare stats? For this purpose, I propose that we measure all stats against the collective purpose of the team, which (presumably) is winning. In particular, in order to evaluate various stats, I propose a very simple measure, which I call *win-tracking rate*, which is defined as follows.

> Let S be a stat. Then the *win-tracking-rate* (WTR) of S is, by definition, the frequency at which the winning team out-performs the losing team with respect to S.

For example, the win-tracking-rate of *hits* is how frequently the winning team out-*hits* the losing team.[1]

Besides being fundamentally easy to understand, *win-tracking-rate* can be applied indifferently to *count-stats*, such as hits, and *rate-stats*, such as batting average.[2]

---

[1] Note that, by the win-tracking criterion, the best stat of all is, of course, *runs*. The winning team out-performs the losing team 100% of the time with respect to *runs*, by definition. Unfortunately, *runs* is not a very good stat for evaluating individual players because scoring runs generally involves teamwork. Accordingly, baseball analysts seek other stats that apply meaningfully to individual players but manage to track runs and hence wins.

[2] A count-stat is one that results from totaling a kind of events – e.g., hits – whereas a rate-stat is one that results from dividing one total by another – e.g., batting average – which is total hits divided by total at-bats.

3. The Data

The data I examine come from play-by-play records of all the regular-season games currently available from Retrosheet,[3] which include the years 1960-1992 and 2000-2004, and comprise 74,225 games. The following table summarizes the findings for selected conventional stats.[4]

stat	WTR	maximum	minimum	variance
OPS	84.0%	85.0%	81.5%	0.99%
weighted batting average[5]	81.7%	83.7%	78.6%	1.60%
slugging percentage	80.8%	83.1%	78.9%	1.07%
on-base percentage	79.8%	82.3%	76.6%	1.79%
total bases + times on base	79.7%	81.0%	76.0%	1.47%
batting average	78.6%	80.8%	76.0%	1.41%
total bases	75.0%	78.8%	73.9%	1.33%
times on base	74.5%	76.9%	70.9%	2.01%
hits	70.5%	73.2%	67.8%	1.90%

As one can see in the above chart, the winner is OPS, and the loser is *hits*. Over the course of 74,225 games, the winning team out-performed the losing team 84% of the time with respect to OPS, whereas the winning team out-hit the losing team only 71% of the time. That it tracks winning the best among conventional stats supports its increased attention over the last few years.

Nevertheless, there remains considerable room for improvement.

4. A New Way to Measure Offensive Performance

By way of improving upon the above results, in what follows, I propose a new way of measuring offensive performance. The new measure is not constructed from existing measures, but is rather created from scratch from play-by-play data. Of course, this means that the proposed stat cannot be calculated from the usual baseball summaries (e.g., box-scores). Nevertheless, it can be readily calculated from widely available play-by-play accounts of games, and it can be calculated on the spot by anyone who keeps score during the game.[6]

1. The Basic Idea – Advancing Base Runners

It has often been said that the goal of the offense in baseball is to get men on base, move them along, and get them home.[7] We can simplify this account by treating every offensive player as a base-runner, in which case we can say that the goal of the offense is to advance base-runners ultimately home.

---

[3] Baseball researchers are all profoundly grateful to Retrosheet and its many dedicated volunteers over the years, who have made this data available, and whose only requirement in return is that we include the following official disclaimer.

> The information used here was obtained free of charge from and is copyrighted by Retrosheet. Interested parties may contact Retrosheet at 20 Sunset Rd., Newark, DE 19711.

[4] In this chart, WTR is win-tracking rate, whereas *maximum*, *minimum*, and *variance* refer to year-to-year results. For example, in its best year, OPS tracked wins at a rate of 85.0%, and in its worst year OPS tracked wins at a rate of 81.5%.

[5] Weighted batting average is similar to slugging percentage, except that it includes walks and it also weights the various events according to well-known ideas according to how much each is worth. The weights I employ, which come from Albert and Bennett, *Curveball*, are as follows. Walk – .36; Single – .52; Double – .67; Triple – 1.18; Homerun – 1.50.

[6] That is, provided one includes a few more details in the account. See Section 4.3.

[7] Bear in mind that, in this context, a "man" may very well be a woman or a child. For the sake of simplifying my description, and grammar, I simply pretend that all baseball players are men.

Accordingly, the success of an offensive player is measured by *how many* base-runners he advances *how far*.[8] The unit of measurement is *base-runners-bases-advanced*, or more simply *bases-advanced*, or even more simply *bases*. For example, if the batter moves a runner from first-base to second-base, he is credited with one base; and if he moves a runner from first-base to third-base, he is credited with two bases; etc. In the meantime, if the batter advances himself to first-base, he is credited with one base, and if he advances himself to second-base, he is credited with two bases, etc. A player is also credited with a base-advance for each base he steals.

Note that the exact manner in which the base-advances are achieved is irrelevant to score-keeping. For example, with the bases empty, reaching first-base is scored as one base, irrespective of whether it achieved *via* a hit, walk, or error. Similarly, a lead-off walk followed by a stolen base is tallied as two bases, which is the same as a lead-off double. By the same token, a lead-off batter who hits *but not safely* is credited with zero bases, just as if he had struck out.

One can also achieve *negative* bases-advanced, which happens when base-runners are erased. In particular, erasing a base-runner on first / second / third is scored as –1 / –2 / –3 bases, respectively. For example, suppose the lead-off batter reaches first on a single, but is subsequently caught stealing. The first event counts as +1, and the second event counts as –1, so that the net result is 0. An extreme example involves a batter hitting into a triple-play (say 5-4-3). Since the man on second is erased, this counts as –2, and since the man on first is erased, this counts as –1, so the net result is –3. Other examples involve force-outs and fielder's choices. For example, a force-out at second-base produces a net result of 0, since the batter reaches first (+1), but the man on first is out (–1).[9]

2. Base-Advance Average

The next obvious question is what denominator do we use to produce a base-advance *average*? One's initial inclination is to divide bases-advanced by plate-appearances, but this does not properly count *opportunities*, which can be greater or smaller, and which can be converted or squandered. For example, it is generally agreed that striking out with the bases loaded is considerably worse than striking out with the bases empty; yet both involve zero bases-advanced over one plate-appearance (0-for-1).

In order to take these important differences into account, I propose a considerably more useful denominator, called *base-advance opportunities*, or simply *opportunities*. For each plate appearance, for each base-runner (including the batter), there is a maximum number of bases that the base-runner can be advanced; for example, the batter can be advanced a maximum of 4 bases, and a man on first can be advanced a maximum of 3 bases. The number of opportunities is then the total of all of these individual maxima. So, for example, if the bases are empty there are 4 opportunities, whereas if the bases are loaded there are 10 (4+3+2+1) opportunities.

---

[8] Since beginning my investigations, I have learned that this measure was proposed over 90 years ago. In Alan Schwarz's fabulous book *The Numbers Game*, he reports on page 37 that, in a letter to *Baseball Magazine* in 1913, a fan J.H. Hamel proposes measuring base-advances in exactly this manner. Also, I recently learned that, for the past few years, Steve Winters has promoted this idea on his extensive website http://www.basesproduced.com/. There are differences in our approaches. I count base-losses as negative base-advances, whereas Winters counts them separately, and Hamel does not mention them. Furthermore, in the matter of constructing an associated base-advance average, Hamel proposes using plate-appearances as the denominator, and Winters proposes using base-runners (including the batter) as the denominator, whereas I propose base-advance opportunities as the denominator (see Section 4.2).

[9] As with many stats, such as hits versus errors and wild pitches versus passed balls, score-keeping requires interpretation. For example, suppose a runner is sent home by the third-base coach, but is thrown out at home. Who do we charge the –3 bases to? The batter? The base-runner? The third-base coach? The team? In live score-keeping, this will be a matter of on-the-spot evaluation, since we have the play right in front of us. For example, in Game 3 of the 2004 World Series, the audience witnessed a colossal base-running error, which clearly should be charged to the base-runner. On the other hand, in retrospective score-keeping, we do not have this luxury; we only have the before-states and the after-states. For retrospective score-keeping, I propose that we charge –3 bases to the batter, treating such a play as similar to a fielder's choice. This is not entirely fair to the batter, but the batter gets breaks on other plays in which he is credited for base-advances that perhaps should be credited to the swiftness of a base-runner ahead of him.

Base-advance average is computed by dividing the total bases-advanced by the total number of base-advance opportunities.[10] So, a strike-out with the bases empty counts as 0-for-4, whereas a strike-out with the bases loaded counts as 0-for-10. On the other hand, a walk with the bases empty counts as 1-for-4, whereas a walk with the bases loaded counts 4-for-10, and a grand-slam counts as 10-for-10.

3. A Sample Inning

By way of a illustration, I present the following sample half-inning, which recreates the top of the third inning of the fourth game of the 2004 World Series, between the Boston Red Sox and the St. Louis Cardinals.

player	play	base-runner advances				bases adv'ed	base advance opp's
		0	1	2	3		
Cabrera	F7	0				0	4
Ramirez	1b 7	1				1	4
Ortiz	2b 9	2	2			4	7
Varitek	FC 4-2	1		1	–3	–1	7
Mueller	BB	1	1		0	2	8
Nixon	2b 8	2	2	2	1	7	10
Bellhorn	BB	1		0	0	1	7
Lowe	K	0	0	0	0	0	10
Total						14	57

Fundamental to score-keeping are the base-runner advance entries, which are highlighted. For example, Varitek bats with men on second and third, so the number of base-advance opportunities for him is 7. He grounds into a fielder's choice, which moves him to first (+1), and moves the man on second to third (+1), but erases the man on third (–3), and which accordingly tallies as –1 for 7. All told, the inning produces 14 base-advances over 57 opportunities, for a base-advance average of .245.[11]

5. How Good a Stat is Base-Advance Average?

Irrespective of its intuitive appeal, the "cash value" of a stat is how well it tracks winning. Here, our findings demonstrate that total bases-advanced is a very good stat, and base-advance average is an excellent stat, which may be seen in the following summary table.

stat	WTR	maximum	minimum	variance
base-advance average	95.5%	96.4%	94.7%	0.16%
total bases-advanced	88.5%	90.0%	86.3%	0.68%

What is remarkable, I believe, is that not only does base-advance-average track winning at an astonishing rate of 95.5% over the course of 74,225 games, it also does so with a rock-steady consistency from year to year, as indicated by the exceptionally small variance.

[10] Note carefully that non-batting plays (e.g., stolen bases) add to the total number of bases advanced, but do not add to the total number of opportunities. Accordingly, a lead-off walk followed by a stolen base nets the player +2 base-advances over 4 opportunities. The calculations produce occasional oddities; for example, if a pinch-runner steals a base, he is credited with +1 base over 0 opportunities for a base-advance average of "infinity". The latter, of course, is similar to those occasions in which an unfortunate pitcher gives up runs but records no outs, and whose ERA is accordingly infinite.

[11] The fraction .245 is not very good as a batting average, let alone a slugging average, but it is a very good number for base-advance average, whose average value from 2000 to 2004 was 150. See Section 6.

6. How do the Players Stack Up?

The following table presents the base-advance-average champions for the available years, as well as the league averages for those years.[12]

year	player		plate appearances	bases advanced as a batter	bases advanced as a runner	base-advance opportunities	base-advance average	league average
1960	Williams	Ted	390	455	5	2256	203.9	140.9
1961	Mantle	Mickey	646	758	22	3543	220.2	143.8
1962	Mantle	Mickey	502	613	19	2839	222.6	142.7
1963	Aaron	Hank	654	653	38	3518	196.4	133.7
1964	Mantle	Mickey	567	615	10	3170	197.2	136.7
1965	Mays	Willie	638	649	26	3445	195.9	133.1
1966	Allen	Dick	599	636	8	3307	194.7	133.6
1967	Yastrzemski	Carl	679	735	25	3662	207.5	130.2
1968	Yastrzemski	Carl	664	616	39	3623	180.8	126.8
1969	McCovey	Willie	583	684	10	3319	209.1	136.4
1970	McCovey	Willie	638	758	10	3714	206.8	144.3
1971	Aaron	Hank	546	598	5	3011	200.3	136.1
1972	Williams	Billy	650	717	13	3533	206.6	132.1
1973	Stargell	Willie	609	657	14	3307	202.9	139.3
1974	Morgan	Joe	641	632	66	3580	195.0	138.2
1975	Morgan	Joe	639	701	72	3636	212.6	140.9
1976	Morgan	Joe	599	708	65	3397	227.6	136.6
1977	Carew	Rod	694	746	27	3765	205.3	145.1
1978	Parker	Dave	642	676	38	3417	209.0	140.1
1979	Lynn	Fred	622	682	10	3335	207.5	145.6
1980	Brett	George	515	629	20	2849	227.8	140.0
1981	Schmidt	Mike	434	495	17	2456	208.5	135.2
1982	McRae	Hal	676	704	5	3727	190.2	139.7
1983	Murphy	Dale	687	663	38	3705	189.2	140.0
1984	Davis	Alvin	678	695	8	3740	188.0	139.6
1985	Brett	George	665	678	15	3497	198.2	143.9
1986	Davis	Eric	487	440	91	2613	203.2	146.0
1987	Davis	Eric	562	577	66	3059	210.2	149.5
1988	Canseco	Jose	705	705	40	3837	194.2	140.0
1989	Clark	Will	675	701	24	3631	199.7	140.2
1990	Bonds	Barry	621	610	42	3350	194.6	142.0
1991	Bonds	Barry	634	657	45	3626	193.6	141.3
1992	Bonds	Barry	612	653	38	3382	204.3	140.9
2000	Helton	Todd	697	903	13	3971	230.7	155.6
2001	Bonds	Barry	664	864	13	3604	243.3	150.6
2002	Bonds	Barry	612	783	17	3331	240.2	148.2
2003	Bonds	Barry	550	670	17	3062	224.4	149.9
2004	Bonds	Barry	617	835	21	3449	248.2	151.3

[12] Note that the units for base-advance average we employ are bases *per thousand* opportunities. So, for example, in 1960 Ted Williams produced 203.9 bases per thousand opportunities.

On the other hand, the following are the top ten *cumulative* performances for each decade except the 90's (for which we only have three years of data).

decade	player		plate appearances	bases advanced as a batter	bases advanced as a runner	base-advance opportunities	base-advance average
60-69	Mantle	Mickey	4499	4495	94	24619	186.4
	Robinson	Frank	6172	6091	205	33950	185.4
	Mays	Willie	6181	5979	198	33652	183.6
	Aaron	Hank	6175	5960	241	33895	182.9
	McCovey	Willie	4961	4875	90	27279	182.0
	Killebrew	Harmon	6000	5858	80	32663	181.8
	Allen	Dick	3664	3452	72	19932	176.8
	Kaline	Al	5450	5037	172	29909	174.2
	Cash	Norm	5688	5365	117	31594	173.5
	Yastrzemski	Carl	5942	5444	164	32561	172.2
70-79	Morgan	Joe	6273	5784	556	34060	186.1
	Stargell	Willie	5083	5069	70	27971	183.7
	Lynn	Fred	3035	2968	50	16636	181.4
	Parker	Dave	3607	3417	113	19459	181.4
	Williams	Billy	4189	3997	91	23103	176.9
	Jackson	Reggie	5913	5513	180	32528	175.0
	Rice	Jim	3456	3259	71	19053	174.8
	Schmidt	Mike	4506	4117	154	24654	173.2
	Aaron	Hank	3363	3203	49	18811	172.9
	Carew	Rod	5916	5287	288	32272	172.8
80-89	Brett	George	5381	5148	150	29454	179.9
	Strawberry	Darryl	3928	3746	164	21897	178.6
	Henderson	Rickey	6206	4818	762	31571	176.7
	Mattingly	Don	4423	4223	83	24490	175.8
	Schmidt	Mike	5556	5294	90	30648	175.7
	Gibson	Kirk	4557	4037	240	24814	172.4
	Raines Sr	Tim	5621	4294	629	28663	171.8
	Murray	Eddie	6437	6052	135	36058	171.6
	Hrbek	Kent	4767	4441	87	26390	171.6
	Guerrero	Pedro	4858	4423	120	26565	171.0
00-04	Bonds	Barry	3050	3828	87	16732	234.0
	Helton	Todd	3448	3935	81	19157	209.6
	Pujols	Albert	2728	2967	43	15180	198.3
	Giambi	Jason	3036	3310	60	17021	198.0
	Ramirez	Manny	3012	3304	43	16915	197.9
	Berkman	Lance	3142	3389	63	17596	196.2
	Walker	Larry	2406	2481	62	13046	194.9
	Edmonds	Jim	2970	3149	64	16546	194.2
	Rodriguez	Alex	3542	3618	125	19484	192.1
	Delgado	Carlos	3299	3470	53	18387	191.6

7. How do the Teams Stack Up?

The following table presents the cumulative team-data for 2000-2004, including wins and runs, and is sorted by wins.[13]

	FOR							AGAINST						
	wins	runs	BAA	ba	ob%	slg%	OPS	wins	runs	BAA	ba	ob%	slg%	OPS
Yankees	488	4346	158.2	272	350	459	809	320	3748	146.9	263	319	415	734
Athletics	483	4192	153.9	261	337	434	772	326	3497	142.1	256	321	399	720
Braves	482	3957	151.8	267	333	438	772	327	3330	139.2	254	317	395	712
Cardinals	475	4219	156.4	273	342	452	795	335	3558	144.0	258	324	425	749
Giants	473	4112	154.3	266	342	445	786	336	3519	142.0	258	325	406	731
Mariners	456	4141	154.0	276	346	429	775	354	3566	141.9	253	319	410	729
Red Sox	453	4333	157.2	282	352	470	823	356	3732	147.7	255	319	408	727
Dodgers	442	3604	143.6	263	326	425	751	368	3356	139.9	247	316	398	714
Twins	430	3868	150.7	272	335	441	776	379	3831	147.8	270	324	437	761
White Sox	428	4288	156.6	270	335	456	791	382	3978	150.4	265	332	433	764
Astros	428	4142	154.1	264	335	435	770	382	3783	148.2	262	329	434	763
Angels	425	3978	151.5	271	333	425	758	385	3720	146.5	261	328	423	751
D'backs	410	3761	147.4	263	330	424	755	400	3689	145.7	255	319	416	735
Marlins	408	3641	145.8	268	333	431	764	401	3696	146.5	260	332	420	752
Indians	403	4143	154.0	266	334	434	768	407	4109	153.3	270	340	433	772
Phillies	403	3795	149.1	258	333	426	759	406	3751	147.9	259	327	431	758
Cubs	397	3760	147.8	263	328	443	772	413	3712	146.3	252	325	410	735
Blue Jays	394	4054	152.6	264	331	430	761	415	4138	154.4	276	342	445	787
Mets	388	3465	142.0	258	322	402	724	420	3639	146.4	260	326	419	745
Rangers	376	4267	156.0	269	334	457	792	434	4587	161.8	284	356	472	828
Reds	375	3713	146.3	255	324	419	743	436	4182	155.7	273	340	459	798
Padres	372	3649	144.2	261	329	408	737	438	3978	151.4	266	332	442	773
Rockies	370	4355	159.4	276	341	455	796	440	4516	161.7	283	354	479	832
Expos	368	3489	142.7	257	321	408	729	442	3917	150.7	270	335	439	773
Orioles	354	3733	146.3	264	326	419	745	457	4165	154.5	270	342	440	782
Pirates	350	3524	142.7	258	323	407	730	458	4021	152.4	271	341	433	774
Royals	345	3901	149.4	266	326	422	749	465	4451	160.0	282	350	469	818
Brewers	332	3455	141.7	261	327	422	748	478	4083	153.9	268	342	442	784
Devil Rays	319	3507	143.3	257	318	404	722	488	4341	158.0	272	345	454	800
Tigers	315	3540	143.2	258	318	416	733	494	4339	158.3	283	344	460	804

8. Correlation Results

With the above data in hand, we are immediately led to ask how the various stats correlate with wins. First of all, one *should not* expect any offensive measure to correlate *too* closely with wins, since winning depends upon both offense and defense. This is born out by the following correlation values.[14]

	runs	BAA	ba	ob%	slg%	OPS
correlation with wins	.502	.523	.381	.606	.467	.539

On the other hand, one *should* expect winning to correlate with *differential* offensive production, which is born out by the following much stronger correlation values.[15]

[13] Here BAA is base-advance-average, ba is batting average, ob% is on-base-percentage, slg% is slugging percentage.
[14] The correlation measure I employ is R^2, which best delineates strong from weak correlation.
[15] For example, differential-runs is simply runs-for minus runs-against.

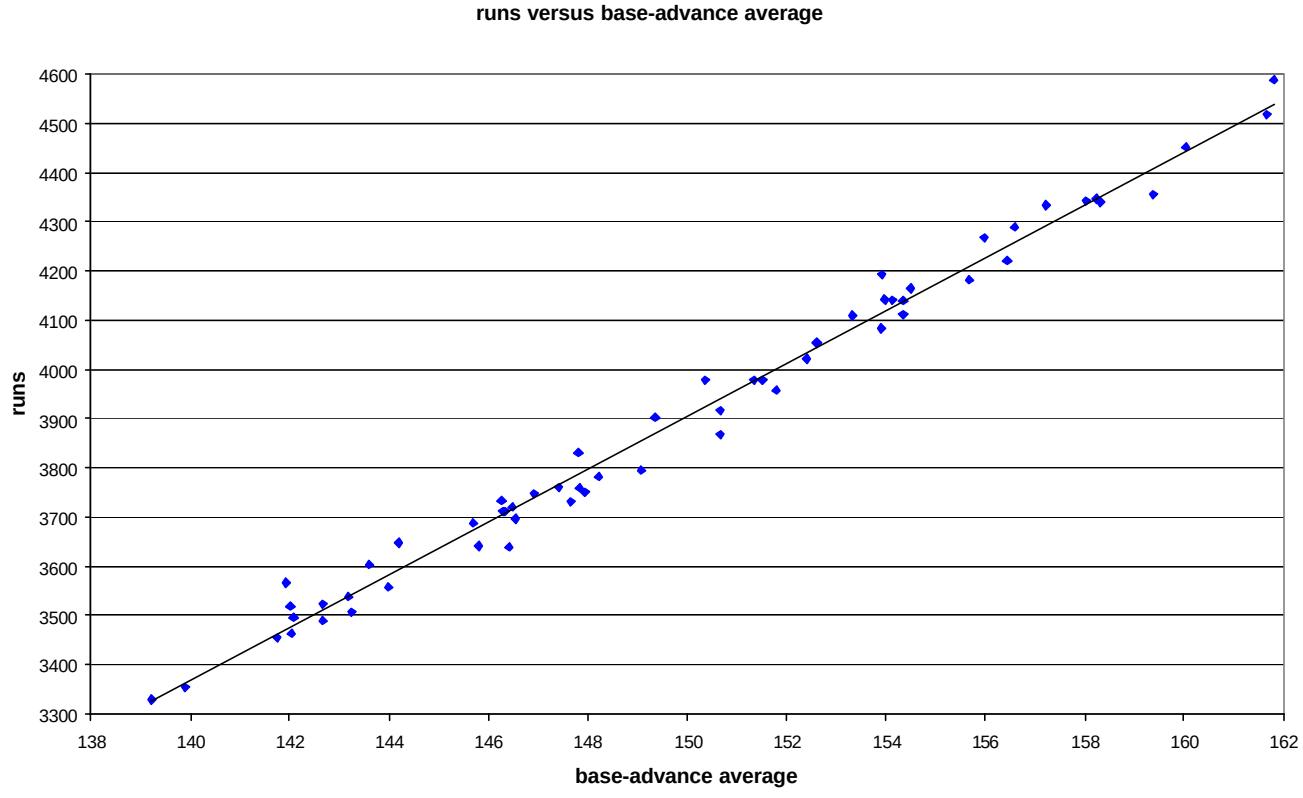
	differential					
	runs	**BAA**	**ba**	**ob%**	**slg%**	**OPS**
correlation with wins	.958	.956	.667	.850	.807	.861

What is quite surprising is that base-advance average performs almost as well as *runs*. The latter stat, of course, is linked by definition to winning. Nevertheless, *runs* is only very slightly better than base-advance average in predicting the number of wins a team achieves over the five year period under consideration.

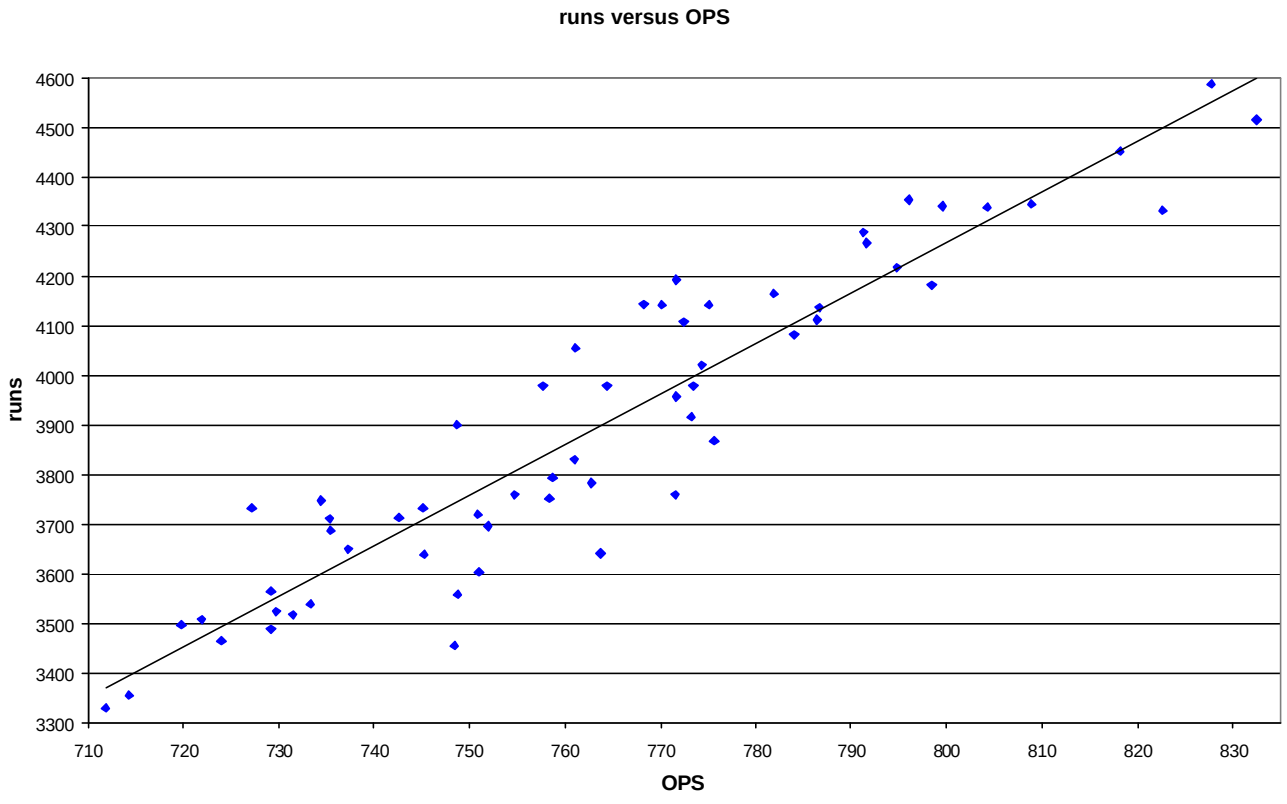
The reason that base-advance average does as well as *runs* is that they are very strongly correlated with each other, which is presented in the following table of correlation values.[16]

	BAA	**ba**	**ob%**	**slg%**	**OPS**
correlation with runs	.988	.739	.823	.831	.874

With a correlation (R^2) of .874, OPS is definitely a very good stat, but with a correlation of .988, base-advance average is "ridiculous" (as they might say on ESPN's *Sports Center*). This point is further emphasized when we examine the associated linear-regression scatter-plots for base-advance average and OPS, which are given as follows.



[16] Technically, one should use runs per game, or runs per out, or some other rate-stat. But, on the whole, the teams played the same number of games, so I simply use runs. Also, note that the correlation values involve 60 points, rather than 30, since they include both runs-for and runs-against for each of the 30 teams.



9. Conclusion

When Bill James began his pioneering efforts to shed light on baseball using baseball statistics, he complained that much of the needed data was unjustly withheld from the general public. By way of correcting the latter problem, he proposed the formation of Project Scoresheet, which put baseball data collection in the hands of the people, and which eventually gave rise to Retrosheet.

The present work is one of the many beneficiaries of these efforts. Using play-by-play data available from Retrosheet, I have compared a proposed stat – base-advance average – with numerous conventional stats, in regard to how well they model win-production and run-production.

In particular, whereas the best available conventional stat – OPS – has a win-tracking rate of 84%, base-advance average has a win-tracking rate of 95%, which constitutes a dramatic improvement.[17]

Questions remain concerning missing play-by-play data – for games prior to 1960, and for games between 1993 and 1999. From a purely statistical point of view, the existing data set of 74,255 games is vastly more than sufficient to make our point. On the other hand, from a historical point of view, the play-by-play data from the missing games are sorely missed. We can examine box scores to ascertain the conventional stats for Babe Ruth, or Lou Gehrig, or Ty Cobb, but we need the play-by-play data to ascertain their base-advance averages, which we would love to know.

Likewise, we have a tantalizing glimpse of Ted Williams' career based on his farewell year, 1960, when he was the base-advance average champion, with a base-advance average of .204, which is comparable to our contemporary slugging goliaths. We would love to know Ted Williams' base-advance average during his prime years.

[17] For example, a stat with a win-tracking rate of 84% makes over *three times* as many prediction/tracking errors as a stat with a win-tracking rate of 95%.