By the Numbers

Volume 18, Number 2 The Newsletter of the SABR Statistical Analysis Committee May, 2008

Review

Review: The Baseball Economist

Charlie Pavitt

The author reviews "The Baseball Economist," a recent baseball book with an economics slant.

The Baseball Economist: The Real Game Exposed, By J. C. Bradbury, Dutton, 2007, 336 pages, $24.95, ISBN 9780525949930

This is another example of the recent outpouring of books whose authors claim unique insights into baseball through their specialized training; in this case, economics. In fact, Bradbury is so confident in the power of economic theory that he has coined the term "sabernomics" as a reference to his work. I have no problem with the general concept of applying one's discipline to baseball when it is done well (i.e., Curve Ball), but that is only sporadically the case here.

Some of Bradbury's attempts to illuminate baseball through economic theory are developed via verbal argument alone, and most of the relevant chapters are pedestrian. Case in point: an attempt to place the decision of whether or not to use performance enhancing drugs into the context of the prisoner's dilemma game. It works, but is trite. Here's the whole thing in a nutshell: "Gee, that guy's no more talented than me but he's doing better than me cause he's juicing, so I better juice too." The problem with material like this is that it makes no contribution to the further scientific understanding of the game while, in reflecting only the idiosyncrasies of the author's interests, provides a skewed view of the area to neophytes who would be better served by, say, *The Hidden Game of Baseball*.

Other issues are presented through statistical research, including a summary of the very good work he and Douglas Drinen have performed on hit-by-pitches, earlier published in the *Journal of Sports Economics* (Vol. 7 No 3., pages 319-329) and *Economic Inquiry* (Vol. 45 No. 1, pages 131-144); I reviewed the latter in BTN Vol. 17 No. 1. There is a good analysis of DIPS (also published in the *Journal of Sports Economics*, Vol. 8 No. 6, pages 616-632) but it reflects my pet peeve; authors presenting as new knowledge something that is in actuality a replication because they are too lazy to seriously review past research. Two statistical analyses that, to my knowledge, make new contributions consist of a nice study of the (absence of) impact of supposed "batter protection" (published with Douglas Drinen in *Journal of Sports Economics*, Vol. 9 No. 2) and a second on managerial influence on umpire ball-strike judgments (to the best of my knowledge not published elsewhere). There is also an insightful examination of the relative amount of hitting versus pitching talent across the twentieth century. Finally, the last three chapters develop a very interesting verbal-only argument that baseball's monopoly power is benign and should be ignored by potential trust-busters.

Nonetheless, I personally don't think these chapters to be worth the price of the book. For readers of this esteemed newsletter, I would suggest limiting their attention to the published work cited above instead.

Charlie Pavitt, chazzq@udel.edu ♦

In this issue

Informal Peer Review

The following committee members have volunteered to be contacted by other members for informal peer review of articles.

Please contact any of our volunteers on an as-needed basis – that is, if you want someone to look over your manuscript in advance, these people are willing. Of course, I'll be doing a bit of that too, but, as much as I'd like to, I don't have time to contact every contributor with detailed comments on their work. (I will get back to you on more serious issues, like if I don't understand part of your method or results.)

If you'd like to be added to the list, send your name, e-mail address, and areas of expertise (don't worry if you don't have any – I certainly don't), and you'll see your name in print next issue.

Expertise in "Statistics" below means "real" statistics, as opposed to baseball statistics: confidence intervals, testing, sampling, and so on.

Member	E-mail	Expertise
Shelly Appleton	slappleton@sbcglobal.net	Statistics
Ben Baumer	bbaumer@nymets.com	Statistics
Chris Beauchamp	cbeaucha@asinc.ca	Statistics
Jim Box	jim.box@duke.edu	Statistics
Keith Carlson	kcsqrd@charter.net	General
Dan Evans	devans@seattlemariners.com	General
Rob Fabrizzio	rfabrizzio@bigfoot.com	Statistics
Larry Grasso	l.grasso@juno.com	Statistics
Tom Hanrahan	Han60Man@aol.com	Statistics
John Heer	jheer@walterhav.com	Proofreading
Dan Heisman	danheisman@comcast.net	General
Bill Johnson	firebee02@hotmail.com	Statistics
Mark E. Johnson	maejohns@yahoo.com	General
David Kaplan	dkaplan@education.wisc.edu	Statistics (regression)
Keith Karcher	karcherk@earthlink.net	Statistics
Chris Leach	chrisleach@yahoo.com	General
Chris Long	clong@padres.com	Statistics
John Matthew IV	john.matthew@rogers.com	Apostrophes
Nicholas Miceli	nsmiceli@yahoo.com	Statistics
John Stryker	john.stryker@gmail.com	General
Tom Thress	TomThress@aol.com	Statistics (regression)
Joel Tscherne	Joel@tscherne.org	General
Dick Unruh	runruhjr@iw.net	Proofreading
Steve Wang	scwang@fas.harvard.edu	Statistics

Changes in League Averages Over the Course of the Season

Brian Morrow

It has been previously observed that run scoring is higher in the middle of the season than at the beginning and end. Here, the author investigates other baseball measures to see if they, too, change as the season progresses.

League averages often increase or decrease in predictable ways as the baseball season progresses. This paper uses data from the last 20 years to track and display these changes in graphical form.

Data Sources

This study uses data from the 1988-2007 major league baseball regular seasons. All data elements can be extracted from the current version of commonly-formatted on-line boxscores. The data was acquired in a variety of ways. Data from 1988-1991 was originally typed in from the Sporting News annual books of boxscores (as much fun as it sounds). Data from 1992-2007 was either hand-entered, or electronically extracted, on a daily basis from on-line boxscores from various websites. Data from 1988-1999 was supplemented by data licensed from Gary Gillette and Pete Palmer. All of the data used in this study is also available from Retrosheet.

Methods

The following league averages were evaluated:

- runs per inning
- hits per plate appearance
- homeruns per plate appearance
- walks plus hit batters per plate appearance
- strikeouts per plate appearance
- hits per balls in play, where hits is defined as (hits minus homeruns), and balls in play is defined as (plate appearances minus (homeruns + walks + hit batters + strikeouts))

Computations were done separately for each season. 1988-1996 computations were done separately for the American and National Leagues, while, due to interleague play, 1997-2007 computations combine both leagues, with league averages and data for each game adjusted for the presence or absence of the designated hitter.

For each season:

- the cumulative league average was computed as a baseline. Since the database includes the strike-shortened years of 1994 and 1995, the study used the cumulative league average as of 112 games into the season as the baseline, so that every year could be included.

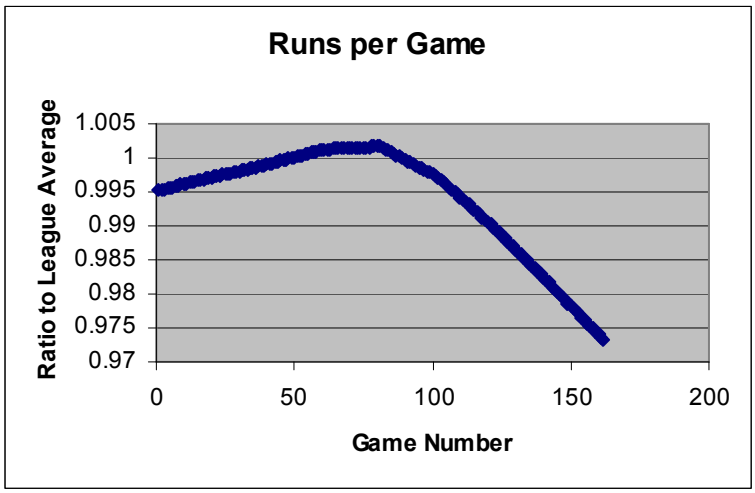
- for each game of the season, game 1 to game 162, the league average of that game number was computed by summing the results from all of the teams for that game number, and dividing by the number of teams.

- for each game of the season, the league average for that game number was divided by the (constant) cumulative league average as of 112 games, to obtain a relative ratio. Obtaining these ratios separately for every season allows the ratios for relatively low-scoring years to be fairly compared to those for relatively high-scoring years.

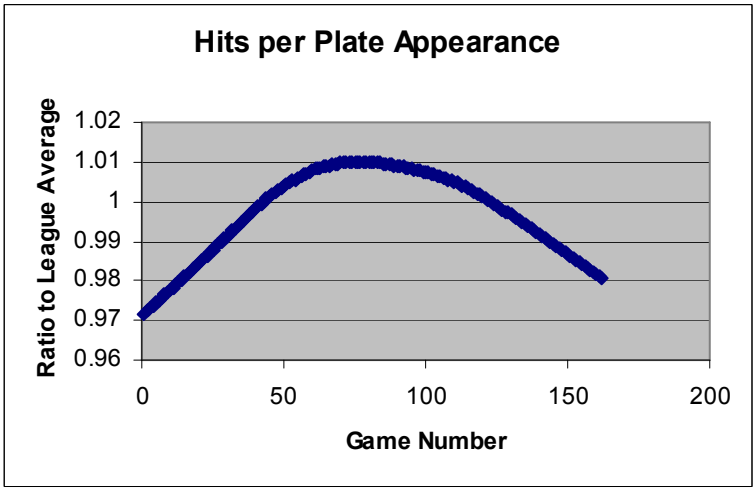
- Once this process is done for each of the 20 seasons, the ratios for each of the 162 games are averaged over the 20 seasons.

- the 162 averaged ratios are smoothed, using PROC LOESS in the SAS software, and graphed. The graphical results obtained from this procedure are very similar to those one would obtain by grouping together sets of games (say, games 1-10, games 11-20, etc.) and then connecting the dots.

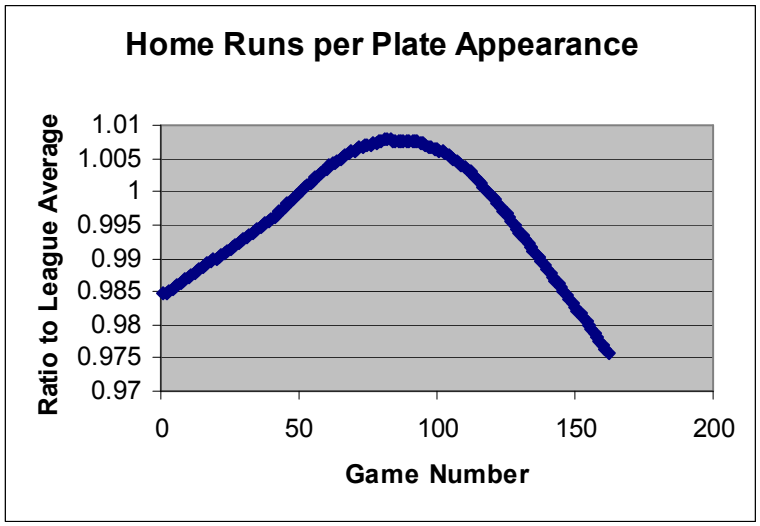
Results



At the start of the season, runs scored are only slightly lower than they are during the peak run-scoring period during the middle of the season – less than a 1 percentage point difference. In fact, the smoothing method used masks the fact that over Games 1-10 of the season, run-scoring is actually higher than the cumulative season average after 112 games have been played. Run-scoring undergoes a steady descent beginning at approximately the middle of the season, finishing at about 3 percentage points below the peak period.

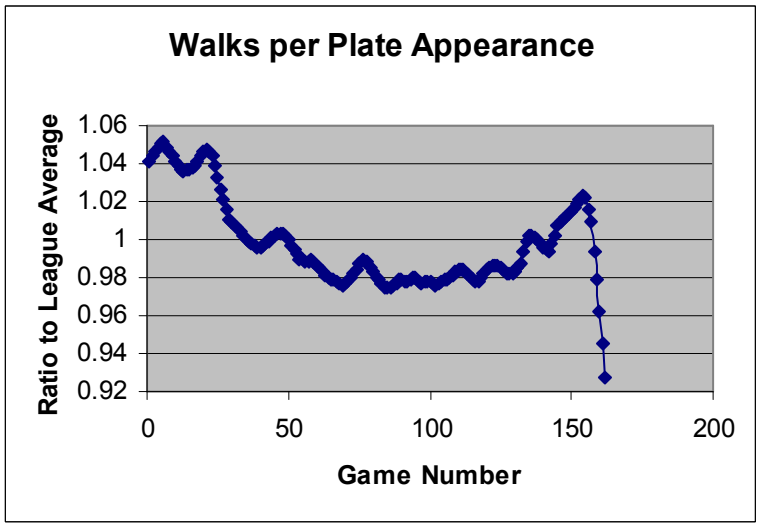


Hits per plate appearance takes on a shape more typical of weather and temperature patterns. There is approximately a 4 percentage point difference between the start of the season and peak period.

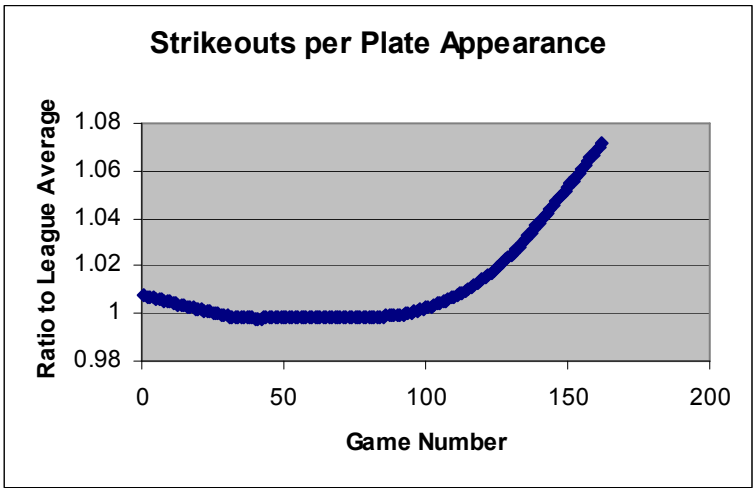


Homeruns per plate appearance has a shape similar to that of hits per plate appearance, although the minimum rate occurs at the end of the season, rather than at the beginning. Again, the difference between the maximum and minimum homerun rate is about 3.5 percentage points.

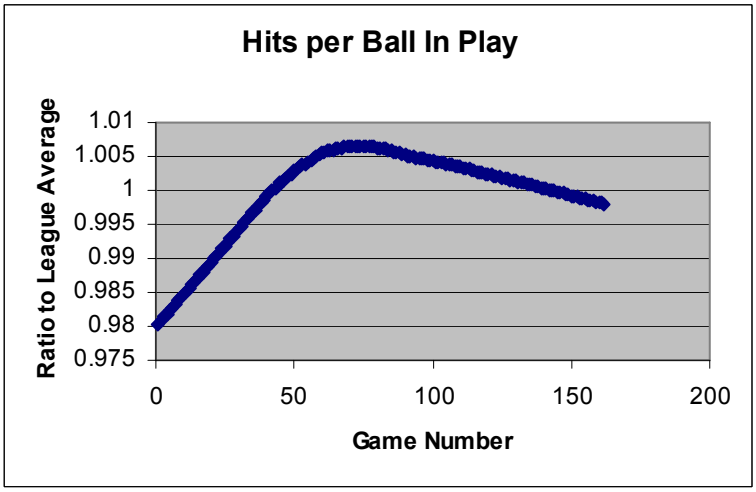
Since hits and homeruns are both down at the start of the season, then why are runs only trivially so? The answer lies in another component of run-scoring: walks and hit batters.



Check it out! Pitchers are walking the stadium at the beginning of the season, at a rate 7 percentage points above that at mid-season. Even more signficantly, the walk rate falls off the table in late September. In the last two weeks or so of the season, batters are swinging at everything. There is a 12 (!) percentage point difference in the rate of walks between the beginning and end of the season.



Strikeouts occur at a fairly constant rate through the first two-thirds of the season, before steadily increasing through year's end. By the end of the season, batters are striking out at a rate 7 percentage points higher than they do mid-season.

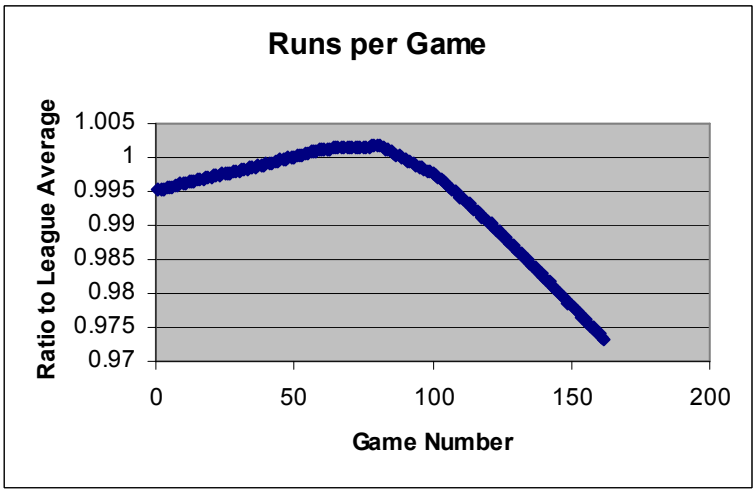


Hits per ball in play follow a similar shape to hits and homeruns per plate appearance. The rate increases about 2.5 percentage points from Opening Day through the mid-season peak.

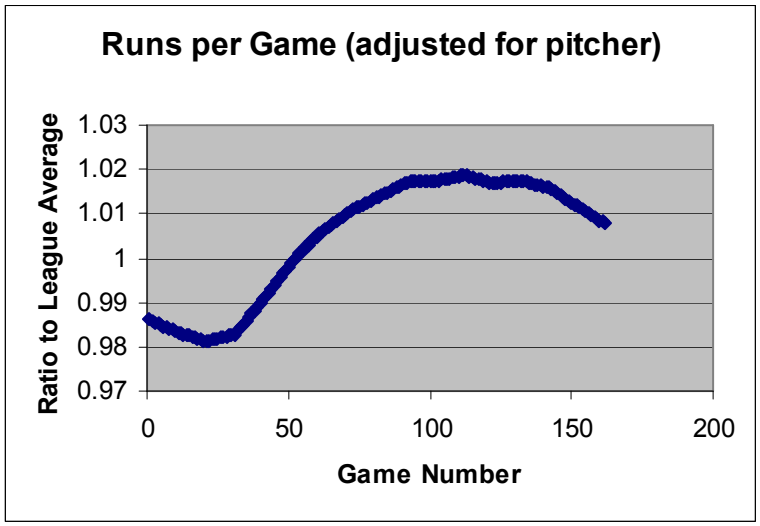
This would seem to complete the picture of how rates of events vary within a season, but, as always in baseball, where everything is connected to everything else, there are other factors in play. In this case, a complicating factor is that the quality of overall player performance is demonstrably not constant throughout the season.

Specifically, if (for example) runs per inning for each game is adjusted for the season-long performance of the starting pitcher, the shape of the runs per inning graph changes significantly.

Below is the runs per inning graph, reproduced from above, not adjusted for starting pitcher performance.



Compare that to the same graph, except with runs per inning is adjusted for the season performance of the starting pitcher:



As one can see, these graphs are not at all alike. The reason is that, as a group, the set of pitchers who take the mound in April and May are not as good as the group of June-July pitchers, who in turn are of poorer quality than the August-September pitchers. Pitchers who perform poorly are replaced throughout the year, often by better pitchers. The early-season unadjusted runs per game (top graph) ratio is near 1.0, but once the fact that this performance is achieved against a set of below-average pitchers is accounted for, the adjusted (bottom) graph shows early-season performance is as low as 0.98. Similarly, the unadjusted graph shows runs taking a nosedive towards the end of the season, but when the above-average group of pitchers during that period is considered, the adjusted graph displays a ratio of above 1.0.

Is the same process happening with batters? Probably. The boxscore data extracted for this study does not easily lend itself to examination of the change in overall batter quality as the season progresses. If the magnitude in the change in batter quality is equal to that of the pitchers, the changes would cancel each other out, and we would be left with the unadjusted (top) graph as a description of how run scoring varies with the time of the season. However, it is not a given that the batter and pitcher changes are close to equal.

The Role of Temperature

Constancio[1] sampled 2000 major league games from 2005-2006 in order to examine the relationship between temperature and strikeouts, walks, homeruns, and hits per balls in play. He found that cooler weather is associated with higher strikeout rates, higher walk rates, lower home run rates, and lower hits per balls in play rates. Comparing his results to those in this paper, the home run rates and hits per balls in play rates can be mostly explained by temperature effects, as both graphs show the inverted-U shape, reflecting cooler temperatures at the beginning and end of the season. However, the strikeout and walk graphs show that these rates are quite different at the start compared to the finish, suggesting weather conditions play only a partial role in determining the rate changes.

Conclusions

Average rates of common events in baseball games can vary significantly as the season progresses. Although game-time temperature is certainly a contributing factor, the different shapes of the above graphs indicate that there are other forces at work. Roster changes, as poorly-performing pitchers are replaced, also have an effect on the rates.

These results can be used to predict future individual game rates and future league cumulative averages from any point in the season. They are also useful for evaluating individual and team performance over time. The pitcher who allows 4 runs a game in April and 4 runs a game in May is improving, while his opponent, allowing 4 runs a game in August and 4 runs a game in September, is declining.

The game you are watching in July is not being played in quite the same way as the one you are watching in April or September.

Further Study

To further isolate the effects of seasonal progression on league rates, gametime temperature could be controlled for. Also, it would be interesting to control for the set of batter season performances for each individual game number, in the same manner that starting pitchers were accounted for, in order to compare the two effects.

Brian Morrow, bmorrow@cdc.gov ♦

[1] Constancio, Chris. Temperature Effects. 2006, www.hardballtimes.com/main/article/temperature-effects