

RUN PRODUCTION IN THE D.H. ERA

by Jefferson Glapski

I. INTRODUCTION

Runs, as Bill James¹ observes, generate wins. Proficient pitching, heavy hitting and fabled fielding contribute to wins. Yet these feats can be measured by a common denominator: runs. A pitcher can be measured by how many runs he gives up, a hitter by how many he produces, a fielder indirectly by how many he saves. Our concern is with offense, specifically with how hitting generates runs. Is a walk as good as a hit? Is a home run the equivalent of four singles? We shall attempt to answer these questions econometrically.

Previous studies of runs begin with Lindsey (1963), who measured the run value of various offensive events, such as singles or doubles. Lindsey based his estimates upon recorded play-by-play data and basic probability theory. Cook (1964) assigned run values on the same basis, except he only calculated the direct effect of offensive events. Cook didn't consider the indirect effect of bringing additional batters to the plate. James (1980) develops a more comprehensive explanation of runs with his runs created formula, where runs created are a function of: i) getting on base; ii) advancing runners on base; and, iii) the amount of plate appearances. By including the amount of plate appearances, James indirectly included the opportunity cost of not getting on base or advancing a runner into the formula. Palmer (1978) simulates all games in Major League history, and uses the collected frequencies to deduce assigned values for various offensive events. This linear weights (LWTS) formula, while determining the value of comprehensive statistics, values the events in terms of runs above average. Our study differs from Palmer's as we intend to measure runs in absolute terms by means of linear regression. This difference, though somewhat preferential, allows average performance to be recognized as valuable in the context of team performance. As Bill James notes:

[Palmer] used an analytic structure that considers an average player to be of no value and a below-average player to have negative value. [Brooks] Robinson spent a couple of years at the start of his career and several at the end as a below-average player, which (in Pete's [Palmer's] system) reduces his career value. This is irrational, since it assumes that the player is being used by a team for several seasons although he has a negative value, not to mention that Robby [Brooks Robinson] was being used for several seasons by a team that was winning 95 games a year. If a team is willing to play a player, then by definition that player must have some value.²

These findings are becoming increasingly important, as baseball fans show a growing interest in the statistical nature of baseball. Player agents are more likely to use statisticians in arguing on their clients' behalf. Indeed, Bill James, a sabermetrician,³ was used in the Doug Drabek arbitration case in 1990, by providing empirical evidence of Drabek's worth as a pitcher. As well, teams seeking to economize on salaries, and get the greatest runs per dollar could use

this research profitably. In other words, this research can serve as a methodology which baseball teams can more efficiently allocate assets. The rest of this paper is organized as follows: section II discusses the data and variables. Section III presents descriptive statistics, the models tested and the results of the linear regression estimates. Section IV concludes the study.

II. DATA AND VARIABLES

The data consist of 460 observations, which represent a very large sample. The cross-section data cover the years 1974-1990, which coincide with the use of the designated hitter in American League play. During these 17 seasons, the data incorporate two unique events. The 1981 players' strike decreased the length of the season by roughly 40%, reducing the values of the collected variables by roughly the same amount. This should not affect the O.L.S. estimators.

Additionally, the American League expanded in 1977 by welcoming the Seattle Mariners and the Toronto Blue Jays, which could possibly affect the run value of events as a higher amount of new players and pitchers entered the league. While it is possible that expansion could affect the run value of offensive events, it is unlikely, as expansion is characterized not only by an influx of "green" pitchers, but also with swells of inexperienced players.

There are a plethora of variables under study, but due to certain shortcomings, certain variables, out of necessity, have been omitted. Variables omitted from the study, but which may pertain to the study include hit-by-pitcher, ground into double-plays, intentional bases on balls, sacrifice hits and sacrifice flies. Our data source for the years 1974-1988, Total Baseball, does not have these statistics. Our source for 1989-1990, Bill Mazerowski's Annual Baseball Preview, contains these variables, but has only been published in the past few years. While this can be considered to be a shortcoming in the data, it is a negligible problem, as these variables do not occur as frequently as the other variables we have employed. While these may be significant events, the amount of runs accounted for by these events is probably very small. Another shortcoming in the data is to imperfections in data collection. Ideally, it would be beneficial to collect opposition errors, since errors create unearned runs, which account for as much as 10% of runs scored in some years.

The variables collected are team statistics for one season, rather than individual statistics. This choice was made, since individual statistics like runs scored and R.B.I.s are largely a function of a player's place in the batting order. Trades from one team to another within a season may also tend to skew individual batting statistics. Additionally, individual statistics are not entirely indicative of the quantitative relationship they hold with runs. For example, a player drawing a walk in the first inning with two out has produced an event that may create a run, by becoming a baserunner, advancing another baserunner from first to second base, increasing that player's likelihood of scoring. Additionally, by extending the inning by an out, that batter has allowed a batter in another inning to make an additional plate appearance. This contribution is captured by using team statistics. The variables retrieved include:

R: runs scored, the dependent variable;
H: hits;
D: doubles;
T: triples;
HR: home runs;
BB: bases on balls;
SO: strikeouts;
SB: stolen bases;
CS: caught stealing; and,
L: league, the dummy variable (1=National League, 0 if American League).

Certain transformations were required to avoid double counting of particular variables. Hits, for example, subsume doubles, triples and home runs. We generated singles by subtracting the extra-base hits from hits. To generate outs, we simply subtracted hits from at-bats. Again to avoid double counting, we subtracted strikeouts from outs, to create the variable outs in play. In summary, the transformed variables we will use are:

S: singles (equal to $H - D - T - HR$); and,
O: outs in play (equal to $AB - H - SO$).

By including outs, we have included variables that will exact the opportunity cost of grounding, flying or striking out.

Conceptually, we expect the signs of all variables, except for outs, outs in play, caught stealing and strikeouts, to be positively signed, indicating a positive contribution to runs. We expect that home runs has a larger coefficient than triples, which has a larger coefficient than doubles, which has a larger coefficient than singles, which has a larger coefficient than bases on balls. This indicates a decreasingly positive contribution to run production. A single, for example, can't be more valuable than a double, because a double possesses higher run-producing potential. In getting on base, the runner is on second rather than first, increasing the probabilities of scoring. As well, a double will usually advance runners already on base further than a single, again increasing the likelihood of runs being scored. With regard to our prior expectations of the specific magnitudes of variables, we expect a home run to be greater than one, since at least one run is scored when a home run is hit. The coefficient for stolen bases should be approximately the difference between the coefficient for a triple and a double, or the coefficient for a double and a single, as a stolen base represents the difference between being on second base rather than first; again this would increase the likelihood of runs being scored. The magnitude on the coefficients for various "out" is less certain. A priori, we have no indication whether a strikeout is more detrimental to the production of runs than outs in play. A strikeout negates the possibility of double plays (usually), taking away less runs than other outs. However, outs in play can advance baserunners, which is a component in producing runs. Caught stealing, however, is likely to be of greater

magnitude than other types of outs, because the potential to score runs with a runner on base exceeds that with the bases empty.⁴

III. ANALYSIS OF DATA

A. DESCRIPTIVE STATISTICS

The players' strike of 1981, while not affecting the O.L.S. estimators, will alter the descriptive statistics. The means of all variables will be less than normal, and the standard errors will be larger than normal (see Appendix for output). While the strike will affect the covariances to some extent, the correlations among variables should be unchanged, as the variance (or the covariance between and variable and itself) is affected by the strike to the same extent.

Table I: Correlation Matrix of Variables - 460 Observations

[INSERT TABLE I HERE]

Table II: Covariance Matrix of Variables - 460 Observations

[INSERT TABLE II HERE]

It is interesting to note that all variables except for the dummy variable L covary positively with the dependent variable runs. In retrospect, it becomes fairly obvious that teams use up a great amount of outs to generate runs. The correlation matrix reveals a few surprises contrasting with conventional baseball wisdom. Bases on balls correlate almost as strongly with runs as do singles, contrasting the infamous words of Garry Templeton, "They don't pay me to walk." Interestingly, the strongest correlation is between outs and singles, 0.84766. This is a puzzling result, since all teams use 27 outs in a game, 24 if they are victorious at home, excluding extra-inning affairs. Perhaps singles-hitting corresponds to some degree to extra-inning games, which makes some sense, as singles-hitting lineups are perceived to be apt at "playing for one run," a common sense strategy in close games. Another revelation is the weak correlation of 0.32709 between triples and stolen bases, confounding conventional wisdom that speed is the common denominator which binds these events.

Other results confirm traditional baseball wisdom. Home runs are negatively correlated with stolen bases and caught stealing, substantiating the claim that a running game need not advance runners if a lineup has power. The dummy variable correlates positively with strikeouts, corroborating with the belief that the National League is the fastball league, while the A.L. is the curveball league.⁵

The dummy variable correlates negatively with home runs, yet positively with stolen bases and caught stealing, enforcing a further characterization of the leagues: that the N.L. relies on the running game to advance baserunners, while the American League waits for "three-run homers."

B. MODEL AND REGRESSION RESULTS

All the variables were included in the model, consistent with prior beliefs. The model is of the form:

$$(1) R = \beta_1S + \beta_2D + \beta_3T + \beta_4HR + \beta_5BB + \beta_6SO + \beta_7SB + \beta_8CS + \beta_9O + \beta_{10}L + e.$$

Notice that we omit the constant term in this regression; this is due to the fact that there is no theoretical reason for runs to be scored without plate appearances. Though unearned runs can be scored without plate appearances (a player is not credited with plate appearance under official baseball scoring rules), unearned runs still depend upon the presence of baserunners, who were credited with at bats, or the following hitters to register plate appearances.⁶ We ran an O.L.S. estimate for (1) and gathered the following results:

$$(2) R = 0.505S + 0.716D + 1.000T + 1.423HR + 0.321BB - 0.087SO + 0.203SB - 0.154CS - 0.099O - 8.114L$$

R² = 0.947
s.e. = 21.838
df = 450
F = 45096.235

We find that (2) is significant at 0.05, since $F = 45096.235 > F^* = 1.83$. All variables are significant at 5%, since they exceed the critical value of t , 1.960.⁷ The high R² indicates that there is a "good fit." All coefficients are signed in accordance with our prior beliefs. As well, our prior expectations concerning the magnitude of coefficients are achieved

Table III: Confidence Interval and t-tests

[INSERT TABLE III HERE]

Note: confidence intervals for variables based on t-distribution with 450 d.f.; t-value = 1.96. T-tests' critical values are based on 450 d.f..

as well. Home runs were greater than unity, and coefficients on triples, doubles, singles and bases on balls decreased as we expected. We also theorized that the coefficient on stolen bases should not exceed the difference between triples and doubles, or doubles and singles. This hypothesis has been verified empirically, and it is interesting to note that the difference in coefficients is greater for T - D, 0.28377, than it is for D - S, 0.21123. This suggests that the primary advantage in the value of a double, compared to a single, is not in advancing baserunners, but in being closer to home plate, because the value of a stolen base, 0.20291, is negligibly different from this result. The dummy variable denoting league of play, L, has a negative coefficient indicating that less runs are scored in the National League. This may be due to the designated hitter rule, as the presence of an "easy out" in the N.L., with the pitcher batting, may take away some of a lineup's continuity in scoring runs. Or, the negative coefficient may reflect the fact that stadia in the A.L. tend to be better hitters' parks. While these may not be the reasons for the negative coefficient, we can

successfully postulate that there are significant⁸ differences between run production in the National and American Leagues.

Problems in Estimation

We tested for the possibility of heteroscedasticity in the model, using the Breusch-Pagan test for heteroscedasticity. We discovered heteroscedastic disturbances with respect to stolen bases, caught stealing and triples. Perhaps this is due to the contextual nature of the running game, entirely dependent upon a manager's decision, though usually foreseeable under the circumstances. We ran the Breusch-Pagan test on individual variables first, before proceeding to test groups of variables. The Breusch-Pagan test identified stolen bases and caught stealing as separately rejecting the null of homoscedasticity. While triples did not reject null at 0.05 significance, it did at 0.10 significance. Further tests revealed that stolen bases and caught stealing together rejected the null of homoscedasticity at 0.05. Finally, the Breusch-Pagan test revealed that stolen bases, caught stealing and triples together rejected the null of homoscedasticity at 0.05 significance. None of the tests were significant at 0.01 significance, but heteroscedasticity may be present in our model.

However, the Breusch-Pagan test "is conservative in that it rejects the null hypothesis (of homoscedasticity) in instances where the hypothesis should be accepted."⁹ As well, the Breusch-Pagan test uses a critical value which follows a Chi-squared distribution, where the degrees of freedom are equal to the number of explanatory variables in the Breusch-Pagan regression. The test statistic, theta, is formed by halving the value of the SSR in the regression of the standardized O.L.S. residuals on the suspect variables. Our sample size is 460, so even in regressions where R² values are below 0.10, we will tend to get large SSR values, which may cause the Breusch-Pagan test to incorrectly identify heteroscedasticity.

Table IV: Tests for Heteroscedasticity

[INSERT TABLE IV HERE]

Note: * indicates significance at 0.05, rejecting the null hypothesis of homoscedasticity. 200 d.f. were used to approximate the 198 d.f. for the critical values of the Goldfeld-Quandt test.

We also used the Goldfeld-Quandt test to test the presence of heteroscedasticity with regard to stolen bases and caught stealing separately. A weakness of the Goldfeld-Quandt test is that it cannot test whether variables are jointly heteroscedastic. However, the Goldfeld-Quandt test does adjust the test statistic according to sample size. We did not reject the null hypothesis of homoscedasticity at 0.05 for either variable. As the are conflicting results concerning the presence of heteroscedasticity, we will opt for the conclusion that the model is homoscedastic, cognizant that correcting heteroscedasticity in a model that is homoscedastic can introduce heteroscedasticity.

IV. CONCLUSION

Baseball has unrivaled popularity as a sport in North America, drawing 350,770,108 in Major League attendance during the 1980s.¹⁰ Fan interest has never been higher, and the results presented here provide an intelligent manner in which to judge players, teams and strategies employed by managers.

With this popularity, the revenues generated by baseball have increased dramatically. Recently having signed a multi-billion dollar television contract, baseball can be characterized more as an industry than ever before. Because of this, it is becoming increasingly important for baseball management to quantify and verify decisions: millions of dollars are at stake. The model presented here can provide baseball management with a means to reduce payroll costs (in comparison with other teams)¹¹ by valuing players differently than present conventional wisdom.

This simple model, estimated by O.L.S., provides a rudimentary means to calculate a team's runs, based upon baseball statistics of offensive events. By estimating the determinants of run-production, the value of individual players can be estimated. This research complements the Palmer (1978) study, by estimating the run-value of offensive events on an absolute basis. Future investigations could include studies that include a greater amount of variables or consider other data sources, such as pitching statistics.

FOOTNOTES

1. Bill James, well known baseball commentator and sabermetrician (sabermetrics, broadly, is the statistical study of baseball), is the author of Bill James' Baseball Abstract, The Bill James Historical Abstract and The Baseball Book.

2. Bill James, The Bill James Historical Baseball Abstract. New York: Villard Books, 1988. pp. 369-70.

3. Sabermetrics refers to the statistical study of baseball. Its name derives from the S.A.B.R., the Society of American Baseball Research.

4. Caught stealing implies a situation involving baserunners, where the potential to score runs is higher than otherwise. A strikeout or an out in play, while occurring with runners on base, can also occur with nobody on base. One would expect more strikeouts and outs in play to occur with nobody on base than with somebody on base.

5. Strikeout pitchers usually are fastball pitchers, while curve ball pitchers rely on groundouts.

6. It is entirely possible that an unearned may result from two or more cumulative errors. However, the lowest fielding average usually belongs to a shortstop, whose fielding average is close to 0.950. His chances of making two consecutive errors are then $0.052 = .0025$ (assuming independence among errors). As well, two consecutive errors may not result in runs being scored, as most infield errors are single base errors. The probability of unearned runs being scored without some official statistic being registered is quite small.

7. 1.645 is the one-tailed critical value f_0 , t with 450 degrees of freedom. 1.960 is the two-tailed critical value of t with 450 degrees of freedom. Caught stealing is the only variable not significant using the two-tailed critical value of t , but we have strong prior beliefs that the coefficient on caught stealing is negative. Caught stealing is significant using the one-tailed t test, hence, all variables are significant.

8. Significance is used in the statistical sense. 10 runs above average contribute to a win over the course of an average season. 8 runs a year is not that much.

9. Johnson, Aaron C., Marvin B. Johnson and Rueben C. Buse, Econometrics Basic and Applied. New York: MacMillan Publishing Company, 1987. p. 306.

10. James, Bill, op. cit., p. 267.

11. If all teams were to use the model, presumably no cost savings would be realized. Cost cutting would be possible if a player were relatively unappreciated, where management could sign that player without having to bid progressively higher. Of course, if the amount of bidders increase, this opportunity would disappear.

REFERENCES

Cook, Earnshaw, Percentage Baseball. Baltimore: Waverly Press, 1964.

Gujarti, Damodar N., Basic Econometrics (2nd ed.). New York: McGraw-Hill, 1988.

James, Bill, The Baseball Abstract. New York: Ballantine Books, 1985, 1986, 1987.

James, Bill, The Bill James Historical Abstract (2nd ed.). New York: Villard Books, 1988.

James, Bill, with Jack Etkin, Mike Kopf and Rob Neyer, The Baseball Book 1991. New York: Villard Books, 1991.

Johnson, Aaron C., Marvin B. Johnson and Rueben C. Buse, Econometrics Basic and Applied. New York: MacMillan 1987.

Lindsey, George R., "An Investigation of Strategies in Baseball," Operations Research, 11:4, July-August, 1963.

Mazeroski, Bill (ed.), Bill Mazeroski's Baseball Preview. New York: Preview Publishing, 1990, 1991.

Reichler, Joseph L. (ed.), The Baseball Encyclopedia (5th ed.). New York: Macmillan, 1982.

Thorn, John and Pete Palmer, The Hidden Game of Baseball. Garden City, NY: Doubleday-Dolphin, 1985.

Thorn, John and Pete Palmer (ed.), Total Baseball. New York: Warner Books, 1989.

White, Kenneth J. and Linda M. Bui, Basic Econometrics: A Computer Handbook Using Shazam. New York: McGraw-Hill, 1988.