Offense, Defense and Causality in Baseball

Charles F. Hofacker
University of California, Los Angeles

Abstract

While baseball fans are known for their penchant for statistics, the application of rational statistical methods to the results of baseball games is apparently a new phenomenon (James, 1982). The lack of sophisticated analyses is all the more surprising when it is considered that almost every aspect of the sport has been religiously tabulated for over 100 years.

SAS was used to enter and analyze published data for the American and National baseball leagues. Various econometric, time series, and simple descriptive techniques were used to look at a number of questions of interest to baseball fans, and perhaps those interested in the development and structure of competitive systems. The first topic considered is the relationship between offense and defense. Is it indeed true, as Casey Stengel has said, that "good pitching beats good hitting and vice versa?" Other topics looked at include hitting, fielding, speed, pitching, and the relationship between the two major leagues.

Introduction

Recently, several reference works have been published that contain a myriad of baseball statistics. In particular, Macmillan Publishing Company's Baseball Encyclopedia contains information on every major league team since 1876. In one of the most useful sections, the editors tabulated data for each team for the entire season. They included the following variables pertaining to team offensive performance: runs, doubles, triples, home runs, batting average, slugging average, and stolen bases. Defensively, opponent's runs, the team's fielding average, errors, bases on balls, strikeouts, shutouts, earned run average, complete games, and saves were included. Additionally, each team's won and lost record is recorded. These data are fertile hunting grounds for researchers with access to SAS and a fondness for the grand old game.

All the current research involved the following two subsets of data from the Baseball Encyclopedia. The first dataset consists of individual team data from 1962 until 1975. Each of the variables mentioned above was entered (with SAS/FSP) for each team in both major leagues for the above years. Differences between successful teams and unsuccessful teams might be of use to see what factors lead to winning seasons. We will call this dataset the team dataset.

The second dataset consists of yearly data for both leagues from 1911 to 1975, aggregated to the league level. The variables are the same as those in the team database, but without the variables for won and lost records. Thus there are two observations per year in the second dataset: one for the American League and one for the National League. It is interesting to reflect on the vast scope of this second dataset. The dataset represents the efforts of the thousands of ballplayers playing nearly 91,000 games under relatively similar conditions. For the 65 years recorded in the dataset, the same basic test of skill was played out over 4 1/2 million times: a batter walked to the plate and tried to create runs. Throughout the paper, this dataset is called the time dataset.

Both sets of data were used to explore some issues in the following topics: offense and defense, hitting, fielding, speed, pitching, and differences between the leagues.

The Relationship between Offense and Defense

The Casey Stengel quote reflects, I think, a certain belief in the symmetry between offense and defense. In other words, a run your pitchers prevent the other team from scoring is exactly as valuable as a run that you score off the other team's pitchers. The symmetry principle seems intuitively necessary, but do the data show this? We can empirically explore the question with the team dataset. What is required is a regression equation where the number of wins for the season for a team is modeled as a function of the number of runs scored and the number of runs allowed. The symmetry idea proposes that the regression coefficients for runs should be the same as that for opponent's runs, but of the opposite sign.

In order to perform the analysis, the functional form of the model must be determined. James (1982) proposed the following equation:

$$P(\text{Win}) = \frac{\beta \text{RUNS}}{\beta \text{RUNS} + \beta \text{OR}}$$

OR refers to the opponents' runs. James (1982) suggests 1.83 as a value for $\beta$. Some algebra reveals that the above equation is equivalent to the following logistic regression equation:

$$P(\text{Win}) = \frac{1}{1 + \exp[\ln(\text{RUNS})\beta - \ln(\text{OR})\beta]}$$

The logistic form would seem reasonable on the basis that $p(\text{Win})$ should follow a binomial distribution. When the logistic model (PROC LOGIST) is applied to the team dataset, a value
of 1.74 results for 8, close to James' (1982) value. The coefficient for offense agrees with that for defense to the second decimal place. A chi square fit for the model of 291.325, p > .99, indicates that the functional form is reasonable, although one must bear in mind that the observations are not independent. There doesn't seem to be much reason a priori to take logs of the runs variables. In fact, when raw runs are used, the model fits better.

There is another sense in which the relationship between offense and defense may be asymmetric. Historically, it may have been the case that innovation in pitching led to temporary advantages for pitchers until batters adjusted. Batting has not profited from improved technology while pitching has seen the slider and other pitches introduced into the defensive arsenal.

Haugh and Box (1977) have described a method by which a researcher can compare one time series to another in order to see if one is causally antecedent to the other. Both series are pre-whitened with their own appropriate ARIMA model. In our case, both series needed only to be differenced. The cross-correlations for a number of leads and lags between the two stationary series are of special use. If series x causes series y, then when x is lagged, there should be significant cross correlations. If y causes x, then these correlations will be zero, and instead, the correlations will be significant when y is lagged. In the time dataset, earned run average positively correlated with batting average, but only contemporaneously. It appears then that some common factor influences both to move up or down.

Finally, on the topic of offense and defense, one can ask if teams that produce a lot of runs also tend to prevent other teams from scoring. Such would be the case if successful teams could afford both good hitting and good pitching, or had good farm systems. Conversely, teams that buy excellent pitching may do so to the neglect of hitting. The latter hypothesis predicts a positive correlation between the runs a team produces and the runs it gives up. The team by team dataset was used to calculate this correlation, and it was not significantly different than zero. r = -.10272. The correlation is in the direction predicted by the first hypothesis, however. The correlation was calculated after both variables had been regressed on time, so it is really a partial correlation.

Hitting

What is more important for winning a pennant: players that hit for average or players that hit home runs? These two skills are not unrelated. For both leagues pooled, the team dataset shows a correlation of .28 (p = .0001) between these two hitting statistics. As before, this correlation was calculated on variables corrected for the effect of time using time dummies, which is to say it is a partial correlation.

Time series and cross sections are pooled in the team dataset, and so PROC TSCSREG is useful in determining the efficacy of various offensive statistics for predicting team runs. No doubt some variance in the dataset pertains to the effect of teams, and a contemporaneous term for years is perhaps needed. Additionally, a given team changes personnel slowly from year to year, so there is an autoregressive component to runs. Parks (1967) proposal for error structure models these three aspects, and has been programmed into PROC TSCSREG. Each of the offensive statistics - batting average, slugging average, doubles, triples, home runs, and stolen bases, contributed to predicting a team's runs.

Fielding

Critics have attacked the fielding average statistic for years. Many players that have high fielding averages are said to have little range. That fielding average is not totally irrelevant can be seen by noting that the correlation between opponent's runs and fielding average is strongly negative: r = -.48, p = .0001. The correlation comes from the team dataset with the effect of time partialed out. A GLS model for opponent's runs revealed the team fielding average statistic to be a good predictor.

Surprisingly, there is a zero-order correlation between ERA and fielding average. ERA is supposed to reflect only the pitcher, not the quality of the fielders behind him. However, when time and a team's won-loss record are partialed out, the correlation vanishes.

Of all the defensive statistics in the team dataset, only one does not appear to relate to the number of runs a team will yield by the end of the season. That one statistic is the number of stolen bases since 1911. American League data are pictured in triangles and the solid line, while the National League is presented with squares and the dotted line. The ordinate shows the average number of stolen bases per game for an average team in the league. Recent times have seen a upswing in the number of stolen bases.

Speed

It has been said that the introduction of artificial surfaces and large multi-purpose stadia has turned baseball into a game of speed. Turning to the time dataset for an moment, Figure 1 shows the historical trend for stolen bases since 1911. American League data are pictured in triangles and the solid line, while the National League is presented with squares and the dotted line. The ordinate shows the average number of stolen bases per game for an average team in the league. Recent times have seen a upswing in the number of stolen bases.

Using the team dataset, the correlation between stolen bases for a team and the number of triples that it gets is r = .23, p = .0001.
Triples would appear to be a measure of team speed, not team power. Note that triples and home runs are negatively correlated, \( r = -0.20 \), \( p = 0.004 \). Thus, teams that hit a lot of home runs do not hit a lot of triples. Perhaps power hitters are slow, as one's stereotype might indicate. Both of the above correlations have been corrected for time. No correlation exists between triples and either stolen bases or home runs at any lag in the time dataset.

Pitching

Billy Martin has something of a reputation for getting a lot out of starting pitchers and not going to the bullpen very often. Is this strategy a good idea? The statistical answer to this question is fraught with the following dangers. First, teams with a lot of complete games might be teams that have poor bullpens. Conversely, teams that have a lot of saves might be teams that have poor starters. In any event, both complete games and saves predict opponent's runs in the team dataset using the GLS method of Parks (1967). The correlation between complete games and saves is \( r = -0.41 \), \( p = 0.001 \), after time has been partialed out.

Turning to the time dataset, we find that years associated with a large number of complete games are associated with a small number of saves. For the American League the correlation is \( r = -0.60 \), and for the National, \( r = -0.38 \). Both of these correlations were calculated after all four series were detrended by first-order differencing.

The league saves totals are one of the few series in the time dataset that are not white noise processes after first differencing. Saves and the total number of shutouts for the year follow ARIMA \((1,1,0)\) models, perhaps because both are dependent on the performance of a small number of individuals.

Both strikeouts and bases on balls reliably predict the number of wins per season. The two measures slightly correlate with each other in the team dataset, \( r = 0.13 \), \( p = 0.028 \), where the correlation has been corrected for time. A positive correlation here is counterintuitive. As a pitcher increases his skill, bases on balls should decrease and strikeouts should increase. Instead, it appears that a team's strikeout total is expanded at the cost of more walks. Thus the distinction between power pitching and finesse pitching is meaningful. There is no relationship between bases on balls and strikeouts in the time dataset.

The baseball powers-that-be tamper with the product only infrequently. But in 1968, the strike zone was made smaller in order to give an advantage to the offense. In Figure 2, you can see the average number of walks given up per team per game from 1911 to 1975. Box and Tiao (1975) have proposed a method for assessing the impact of interventions over time. Using their method, it appears that the number of bases on balls went up as a result of the rule change. The following model was arrived at for bases on balls:

\[
BB(t) = uZ(t) + \frac{a(t)}{1-B},
\]

or equivalently,

\[
(1-B)BB(t) = u(1-B)Z(t) + a(t).
\]

Here, \( B \) is the backshift operator, and \( Z \) is an indicator variable whose value is one after 1968 and zero for 1968 and before. The value \( a(t) \) is the \( t \)-th white noise input to the series, and \( BB(t) \) is the number of bases on balls given up for the league in the \( t \)-th year. The value of \( u \) is 0.043 for the American League with \( t(63) = 2.86, p < 0.05 \). For the National League, the value of \( u \) is 0.6517, \( t(63) = 3.30, p < 0.05 \).

Differences between the Leagues

Writers sometimes claim that the American League is a curveball league, and the National League is a fastball league. Some differences in pitching, or perhaps in umpiring, can be clearly seen in Figure 2. The number of stolen bases in the National League has historically been higher in the American League. The reason for this might well be that the curveball is harder to get over the plate and American League pitchers throw more curve balls.

There are a number of variables that show contemporaneous correlation in the time dataset between the leagues. The number of runs produced in the American League correlates with the number of runs produced in the National League across years. The same is true of home runs, earned run average, and strikeouts.

There is a hint that the number of stolen bases in the National League predicts stolen bases in the American League at a three year lag. Also, American League saves seem to be a leading indicator for National League saves at a three year lag.

Acknowledgements

Paul Hoffman assisted with the figures, and the Office of Academic Computing providing computing time for this paper. For further information write to Dr. Charles Hofacker, Office of Academic Computing, University of California, Los Angeles, California, 90024.

References


Haugh, L. D., & Box, G. E. P. Identification of dynamic regression (distributed lag)

