"AN EMPIRICAL TEST OF BILL JAMES'S PYTHAGOREAN FORMULA"

by

Paul M. Sommers
David U. Cha
And
Daniel P. Glatt

March 2010

MIDDLEBURY COLLEGE ECONOMICS DISCUSSION PAPER NO. 10-06

DEPARTMENT OF ECONOMICS
MIDDLEBURY COLLEGE
MIDDLEBURY, VERMONT 05753

http://www.middlebury.edu/~econ

AN EMPIRICAL TEST OF BILL JAMES'S PYTHAGOREAN FORMULA

by

David U. Cha
Daniel P. Glatt
Paul M. Sommers

Department of Economics
Middlebury College
Middlebury, Vermont 05753
psommers@middlebury.edu

AN EMPIRICAL TEST OF BILL JAMES'S PYTHAGOREAN FORMULA

> The Giants do not usually need to score many runs.
> All that we must do is score more than the other fellow.
>
> — Bill Terry, manager of the
> 1932 New York Giants
> [1, p.136]

Bill James, baseball writer and statistician, in 1980 developed a formula that related a team's won-lost percentage to the number of runs they scored and allowed, as follows:

(1) $$\text{Won-Lost Percentage} = \frac{(RunsScored)^2}{(RunsScored)^2 + (RunsAllowed)^2}$$

Since the Won-Lost Percentage is the ratio of games won to the total number of games played (games won plus games lost), equation (1) can be re-written as follows:

(2) $$\frac{Wins}{Losses} = \frac{(RunsScored)^2}{(RunsAllowed)^2} = \left(\frac{RunsScored}{RunsAllowed}\right)^2$$

If, for example, the Boston Red Sox score 867 runs and allow 657 runs (as they did in 2007), Bill James's "Pythagorean" method [so dubbed because of the presence of three squared terms in equation (1)][1] projects that the team would have a won-lost percentage of $(867)^2/[(867)^2 + (657)^2]$ or .635 (and hence win about .635 × 162 or 103 games). In fact, the Red Sox (world champions in 2007) won 96 regular season games or about 6.8 percent fewer games than the Pythagorean method would predict. In this instance, the exponent of "2" on the right-hand side of equation (2)

is too high. Or, one could argue that "2" is accurate, but the Boston Red Sox should have won more regular season games in 2007 than they actually did.

The purpose of this brief note is to empirically test Bill James's Pythagorean method for all teams in both leagues, by decade, from 1950 to 2007. Does the method work as well since 1980 as it did before 1980? Does the method work better for one league (American or National) than the other? Has the exponent in equation (2) changed in recent decades?

The Models

Equation (2) can be written in log-linear form as follows:

$$\ln\left(\frac{Wins}{Losses}\right) = 2\ln\left(\frac{RS}{RA}\right) \tag{3}$$

where "ln" is the natural logarithm, *RS* denotes runs scored, and *RA* denotes runs allowed. That is, if one first takes logs of both sides of equation (2) and then if we define $y_{i,t} = \ln\left(\frac{Wins}{Losses}\right)_{i,t}$ and $x_{i,t} = \ln\left(\frac{RS}{RA}\right)_{i,t}$ for each team *i* in year *t*, we can estimate the coefficients β_0 and β_1 by applying least squares to *y* and *x* in the following regression:

$$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \varepsilon_{i,t} \tag{4}$$

where $\varepsilon_{i,t}$ is a disturbance term. According to Bill James, β_0 should be indistinguishable from zero and β_1 should be close to "2". To test the null hypothesis H_0: $\beta_1 = 2$, we employ a *t*-test. The test statistic is $t_{calc} = \frac{b_1 - 2}{SE(b_1)}$, where b_1 is the estimated slope coefficient and $SE(b_1)$ is the

standard error of the estimated slope coefficient.[2] Hereafter, equation (4) where $y = \ln\left(\frac{Wins}{Losses}\right)$ and $x = \ln\left(\frac{RS}{RA}\right)$ will be called Model (1).

Model (1) assumes that one more run scored has the same impact on a team's win percentage as does one less run allowed. But what if scoring runs was more (or less) important to winning games than allowing runs? Model (1) might then be revised as follows:

$$\ln\left(\frac{Wins}{Losses}\right)_{i,t} = \beta_0 + \beta_1 \ln(RS)_{i,t} + \beta_2 \ln(RA)_{i,t} + \varepsilon_{i,t} \quad (5)$$

If we relax the assumption that the exponent on the ratio $\left(\frac{RS}{RA}\right)$ is the same (and, according to James, equal to "2"), then the revised model would be described by equation (5), hereafter, Model (2).

The Data

Data on regular season wins, losses, runs scored, and runs allowed for all teams were gleaned from two primary sources: *Total Baseball* [3] for the years 1950 through 2003 and http://sports.espn.go.com/mlb/standings for the years 2004 through 2007.

The Results

Table 1 shows the regression results for each league (as well as for both leagues combined) for each decade since the 1950s. The estimated intercept (b_0) in *all* regressions is not discernible from zero, as Bill James would expect. Since the year 2000, the exponent in the ratio of runs scored to runs allowed in James's Pythagorean formula has been indistinguishable from "2". But, in decades before the turn of the millennium the exponent was not equal to "2". And, in all cases when we could reject H_0: $\beta_1 = 2$ (in favor the alternative hypothesis H_A: $\beta_1 \neq 2$),

our estimate b_1 was invariably less than "2". A comparison of the 30-year period 1950-1979 to the 28-year period 1980-2007 shows that b_1 was in most cases (with the exception of the American League (AL) from 1980 to 2007) significantly less than "2". The impact of *RS/RA* on winning is marginally higher now (1980-2007) than it was in the earlier period (1950-1979). Compare the value of b_1 (1.9202) estimated for both leagues combined in the period 1980-2007 to the corresponding estimate for b_1 (1.8099) in the period 1950-1979. Moreover, it is worth noting that the average number of runs scored is also higher in the National (American) League in the period 1980-2007 than it was in the period 1950-1979 [$\overline{RS}_{1980-2007,NL} = 699$, $\overline{RS}_{1950-1979,NL} = 667.3$, *p*-value on the difference between means is less than .001; $\overline{RS}_{1980-2007,AL} = 747$, $\overline{RS}_{1950-1979,AL} = 669.6$, *p*-value on the difference is again less than .001].

Figures 1 and 2 show scatterplots of *ln(W/L)* against *ln(RS/RA)* for each subperiod (1950-1979 and 1980-2007, respectively) for each league. Each point represents an observation on one team in one year. The points more closely fall on a straight line for the National League, 1950-1979 than they do for the National League, 1980-2007 (compare $R^2 = .878$ for 1950-1979 with $R^2 = .847$ for 1980-2007 in Table 1). Still, the differences between the two periods by league are admittedly very small.

Table 2 shows the regression results for Model (2), which isolates the impact of runs scored from the impact of runs allowed on the win-loss ratio. The right-hand column reports the coefficient of determination (R^2) for each regression each decade, by league. A look down this column and the corresponding column in Table 1 clearly shows that the explanatory power (that is, how well the regressors as a group explain variation in the dependent variable, namely, *ln(Wins/Losses)*) of Model (2) is not an improvement over Model (1). In other words, runs scored and runs allowed seemingly have an equal (and opposite) effect on the win-loss ratio.

Concluding Remarks

Early in the 1980s, Bill James developed a formula in response to the question: Can you tell how many games a team will win, based on its runs scored and runs allowed? A regression analysis of data on regular season runs scored, runs allowed, and wins (and losses) for each team each season in Major League Baseball since 1950 reveals that Bill James's Pythagorean formula has stood the test of time very well indeed. Runs scored and runs allowed have equal (and opposite) effects on team winning, in both leagues and in years before and since 1980. If any modification should be made to the formula, the exponent on runs scored and runs allowed should be reduced to a power slightly below "2" ["1.92" for both leagues since the year 1980].

Table 1. Regression Results for Model (1)
$\ln(WINS/LOSSES) = b_0 + b_1 \ln(RS/RA)$

	Intercept b_0	Slope coefficient on ln(*RS/RA*) b_1	R^2
1950-1959			
AL	-.0059 [.0131][a]	**1.7543** [.0598]	.917
NL	.00003 [.0122]	1.8758 [.0737]	.893
Both leagues	-.0030 [.0089]	**1.7985** [.0461]	.906
1960-1969			
AL	-.0017 [.0094]	*1.8757* [.0593]	.911
NL	-.0013 [.0111]	1.9323 [.0655]	.901
Both leagues	-.0016 [.0072]	*1.9055* [.0441]	.905
1970-1979			
AL	.00001 [.0091]	**1.8139** [.0560]	.894
NL	.0012 [.0101]	**1.6576** [.0642]	.850
Both leagues	.0006 [.0068]	**1.7398** [.0425]	.873
1980-1989			
AL	.0005 [.0078]	*1.8849* [.0577]	.885
NL	-.0017 [.0100]	2.0195 [.0848]	.828
Both leagues	-.0005 [.0063]	1.9381 [.0489]	.859
1990-1999			
AL	.00003 [.0078]	1.9324 [.0599]	.883
NL	-.0021 [.0090]	*1.8370* [.0645]	.856
Both leagues	-.0012 [.0060]	**1.8814** [.0441]	.869
2000-2007			
AL	-.0055 [.0099]	2.0026 [.0624]	.904
NL	-.0004 [.0089]	1.8720 [.0682]	.857
Both leagues	-.0023 [.0066]	1.9445 [.0458]	.883
1950-1979			
AL	-.0021 [.0059]	**1.8062** [.0333]	.907
NL	.00005 [.0064]	**1.8146** [.0393]	.878
Both leagues	-.0011 [.0044]	**1.8099** [.0255]	.893
1980-2007			
AL	-.0012 [.0048]	1.9415 [.0344]	.891
NL	-.0014 [.0054]	*1.8951* [.0411]	.847
Both leagues	-.0013 [.0036]	**1.9202** [.0266]	.870

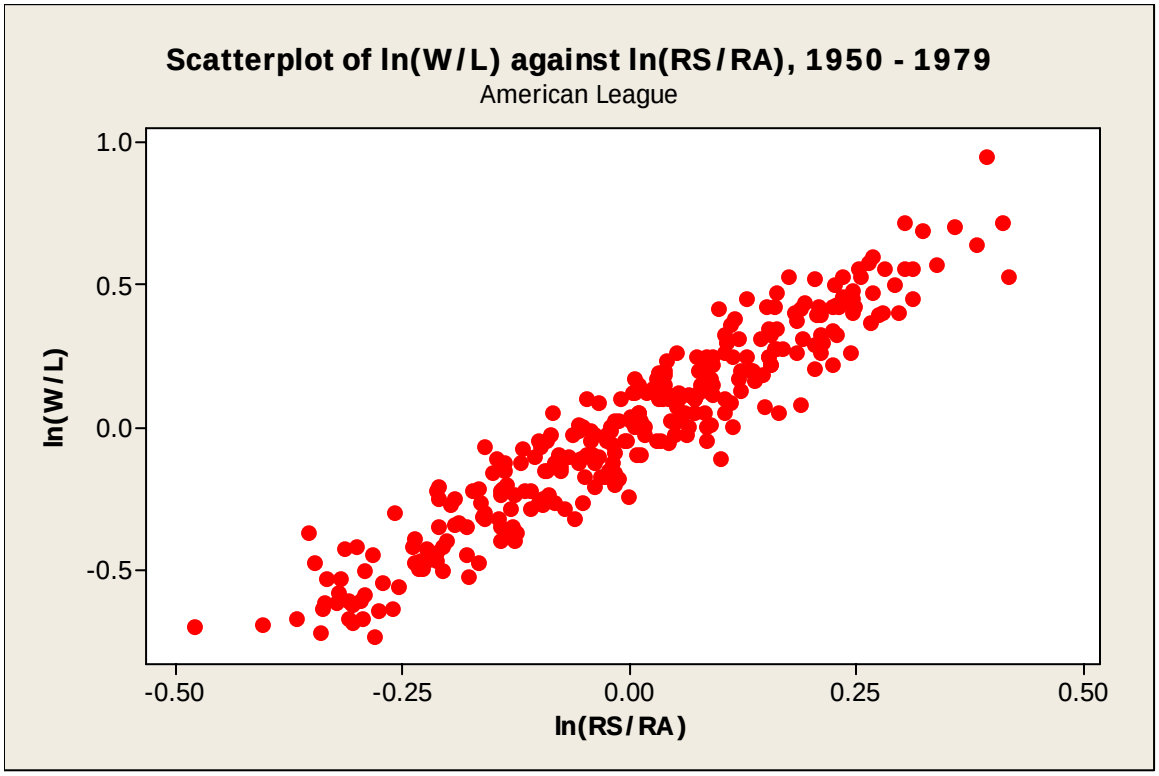
[a]Numbers in brackets are standard errors and numbers in boldface (italics) are significant at better than the .01 (.05) level. The null hypothesis for the intercept is H_0:$\beta_0 = 0$ and the null hypothesis for the slope coefficient on ln(*RS/RA*) is H_0: $\beta_1 = 2$. In both cases, the alternative hypothesis is two-tailed.

Table 2. Regression Results for Model (2)
$\ln(WINS/LOSSES) = b_0 + b_1 \ln(RS) + b_2 \ln(RA)$

	Intercept b_0	Slope coefficient on: $\ln(RS)$ b_1	$\ln(RA)$ b_2	R^2
1950-1959				
AL	.3888 [.9790]	**1.7224** [.0995]	*-1.7829* [.0930]	.917
NL	-1.0380 [1.1370]	1.9526 [.1119]	-1.7937 [.1164]	.894
Both leagues	-.2284 [.7351]	*1.8162* [.0739]	**-1.7816** [.0718]	.906
1960-1969				
AL	.3171 [.5948]	1.8494 [.0771]	-1.8986 [.0733]	.911
NL	.8684 [.7091]	1.8708 [.0823]	-2.0052 [.0883]	.902
Both leagues	.5648 [.4578]	*1.8624* [.0562]	-1.9499 [.0567]	.906
1970-1979				
AL	-.3023 [.5620]	*1.8365* [.0702]	**-1.7900** [.0715]	.895
NL	-.6484 [.7772]	**1.7046** [.0854]	**-1.6046** [.0903]	.850
Both leagues	-.4204 [.4613]	**1.7060** [.0545]	**-1.7060** [.0564]	.873
1980-1989				
AL	.0077 [.3052]	1.8844 [.0619]	-1.8855 [.0630]	.885
NL	-.3260 [.4232]	2.0462 [.0919]	-1.9959 [.0904]	.829
Both leagues	-.1265 [.2429]	1.9477 [.0524]	-1.9283 [.0525]	.859
1990-1999				
AL	.0794 [.4242]	1.9265 [.0680]	-1.9385 [.0683]	.883
NL	-.1865 [.4482]	1.8535 [.0762]	*-1.8253* [.0707]	.856
Both leagues	-.0895 [.2951]	*1.8887* [.0505]	*-1.8753* [.0486]	.869
2000-2007				
AL	-.8735 [.9428]	2.0721 [.0980]	-1.9421 [.0907]	.904
NL	*1.6195* [.7686]	**1.7343** [.0938]	-1.9788 [.0842]	.862
Both leagues	.5502 [.5714]	1.8998 [.0652]	-1.9829 [.0666]	.884

[a]Numbers in brackets are standard errors and numbers in boldface (italics) are significant at better than the .01 (.05) level. The null hypothesis for the intercept is H_0:$\beta_0 = 0$ and the null hypotheses for the slope coefficients on $\ln(RS)$ and $\ln(RA)$ are H_0: $\beta_1 = 2$ and H_0: $\beta_2 = -2$, respectively. In all three cases, the alternative hypothesis is two-tailed.

Figure 1


Scatterplot of ln(W/L) against ln(RS/RA), 1950 - 1979
American League
ln(W/L)
1.0
0.5
0.0
-0.5
-0.50
-0.25
0.00
0.25
0.50
ln(RS/RA)

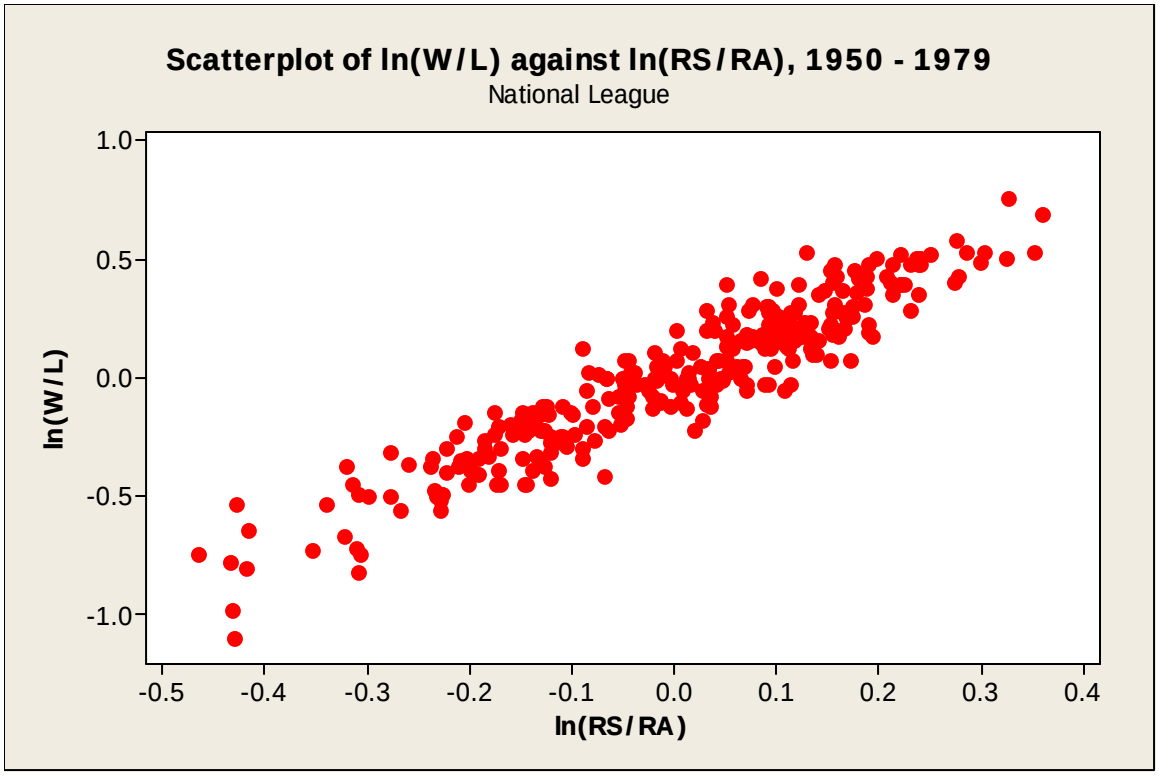
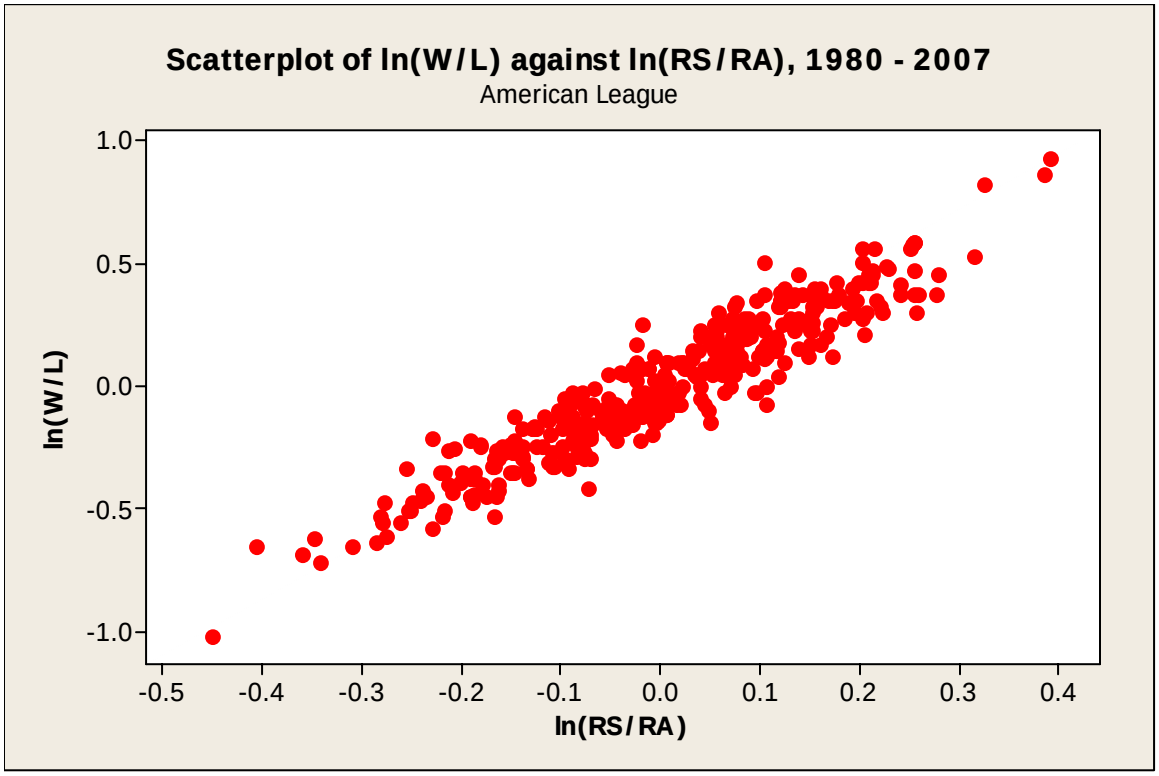
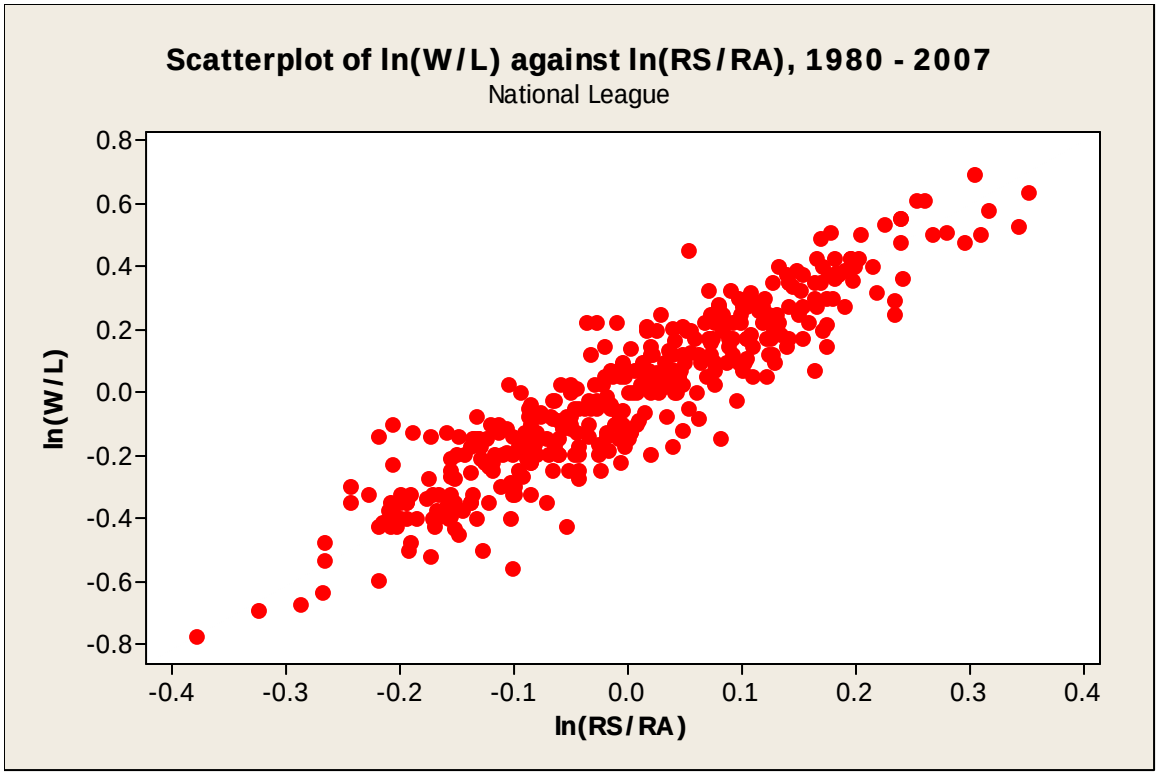

Scatterplot of ln(W/L) against ln(RS/RA), 1950 - 1979
National League
ln(W/L)
1.0
0.5
0.0
-0.5
-1.0
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
ln(RS/RA)

Figure 2


Scatterplot of ln(W/L) against ln(RS/RA), 1980 - 2007
American League
ln(W/L)
1.0
0.5
0.0
-0.5
-1.0
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
ln(RS/RA)


Scatterplot of ln(W/L) against ln(RS/RA), 1980 - 2007
National League
ln(W/L)
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
ln(RS/RA)

References

1. P. Williams, *When the Giants Were Giants*, Chapel Hill, NC: Algonquin Books, 1994.

2. B. James, *The Bill James Baseball Abstract 1983*, New York: Ballantine Books, 1983.

3. *Total Baseball: The Ultimate Baseball Encyclopedia* (edited by H. Thorn, P. Birnbaum, and B. Deane), Wilmington, DE: Sports Media Publishing, 2004.

Footnotes

1. See, for example, the reference to "The Pythagorean Formula" in [2, p. 10].

2. The b_1 estimate also interprets as an elasticity of (*Wins/Losses*) to (*RS/RA*), where (in general) the elasticity of *Y* with respect to *X* is defined as $\frac{X}{Y} \cdot \frac{dY}{dX}$. In other words, a 1 percent increase in (*RS/RA*) will lead to a "b_1" percent increase in (*Wins/Losses*). Moreover, since "*Wins* + *Losses*" is equal to a constant (162 games since the year 1962 and 154 games in years before 1962), it should also be noted that a given percentage change in *Wins* is equal to the percentage change in the winning percentage, [*Wins*/(*Wins* + *Losses*)].