1

BASEBALL'S PYTHAGOREAN THEOREM

The more runs a baseball team scores, the more games the team should win. Conversely, the fewer runs a team gives up, the more games the team should win. Bill James, probably the most celebrated advocate of applying mathematics to analysis of Major League Baseball (often called sabermetrics), studied many years of Major League Baseball (MLB) standings and found that the percentage of games won by a baseball team can be well approximated by the formula

$$\frac{\text{runs scored}^2}{\text{runs scored}^2 + \text{runs allowed}^2} = \begin{array}{l}\text{estimate of percentage} \\ \text{of games won.}\end{array} \tag{1}$$

This formula has several desirable properties.

- The predicted win percentage is always between 0 and 1.
- An increase in runs scored increases predicted win percentage.
- A decrease in runs allowed increases predicted win percentage.

Consider a right triangle with a hypotenuse (the longest side) of length c and two other sides of lengths a and b. Recall from high school geometry that the Pythagorean Theorem states that a triangle is a right triangle if and only if $a^2 + b^2 = c^2$. For example, a triangle with sides of lengths 3, 4, and 5 is a right triangle because $3^2 + 4^2 = 5^2$. The fact that equation (1) adds up the squares of two numbers led Bill James to call the relationship described in (1) Baseball's Pythagorean Theorem.

Let's define $R = \frac{\text{runs scored}}{\text{runs allowed}}$ as a team's scoring ratio. If we divide the numerator and denominator of (1) by $(\text{runs allowed})^2$, then the value of the fraction remains unchanged and we may rewrite (1) as equation (1)′.

$$\frac{R^2}{R^2+1} = \text{estimate of percentage of games won.} \qquad (1)'$$

Figure 1.1 shows how well (1)′ predicts MLB teams' winning percentages for the 1980–2006 seasons.

For example, the 2006 Detroit Tigers (DET) scored 822 runs and gave up 675 runs. Their scoring ratio was $R = \frac{822}{675} = 1.218$. Their predicted win percentage from Baseball's Pythagorean Theorem was $\frac{1.218^2}{(1.218)^2 + 1} = .597$. The 2006 Tigers actually won a fraction of their games, or $\frac{95}{162} = .586$. Thus (1)′ was off by 1.1% in predicting the percentage of games won by the Tigers in 2006.

For each team define error in winning percentage prediction as actual winning percentage minus predicted winning percentage. For example, for the 2006 Arizona Diamondbacks (ARI), error = .469 − .490 = −.021 and for the 2006 Boston Red Sox (BOS), error = .531 − .497 = .034. A positive

	A	B	C	D	E	F	G	H	I	J
1									MAD = 0.020	
2										
3	Year	Team	Wins	Losses	Runs scored	Runs allowed	Scoring ratio	Predicted winning %	Actual Winning %	Absolute Error
4	2006	Diamondbacks	76	86	773	788	0.981	0.490	0.469	0.021
5	2006	Braves	79	83	849	805	1.055	0.527	0.488	0.039
6	2006	Orioles	70	92	768	899	0.854	0.422	0.432	0.010
7	2006	Red Sox	86	76	820	825	0.994	0.497	0.531	0.034
8	2006	White Sox	90	72	868	794	1.093	0.544	0.556	0.011
9	2006	Cubs	66	96	716	834	0.859	0.424	0.407	0.017
10	2006	Reds	80	82	749	801	0.935	0.466	0.494	0.027
11	2006	Indians	78	84	870	782	1.113	0.553	0.481	0.072
12	2006	Rockies	76	86	813	812	1.001	0.501	0.469	0.031
13	2006	Tigers	95	67	822	675	1.218	0.597	0.586	0.011
14	2006	Marlins	78	84	758	772	0.982	0.491	0.481	0.009
15	2006	Astros	82	80	735	719	1.022	0.511	0.506	0.005
16	2006	Royals	62	100	757	971	0.780	0.378	0.383	0.005
17	2006	Angels	89	73	766	732	1.046	0.523	0.549	0.027
18	2006	Dodgers	88	74	820	751	1.092	0.544	0.543	0.001
19	2006	Brewers	75	87	730	833	0.876	0.434	0.463	0.029
20	2006	Twins	96	66	801	683	1.173	0.579	0.593	0.014
21	2006	Yankees	97	65	930	767	1.213	0.595	0.599	0.004

Figure 1.1. Baseball's Pythagorean Theorem, 1980–2006. See file Standings.xls.

error means that the team won more games than predicted while a negative error means the team won fewer games than predicted. Column J in figure 1.1 computes the absolute value of the prediction error for each team. Recall that the absolute value of a number is simply the distance of the number from 0. That is, $|5| = |-5| = 5$. The absolute prediction errors for each team were averaged to obtain a measure of how well the predicted win percentages fit the actual team winning percentages. The average of absolute forecasting errors is called the MAD (Mean Absolute Deviation).[1] For this data set, the predicted winning percentages of the Pythagorean Theorem were off by an average of 2% per team (cell J1).

Instead of blindly assuming winning percentage can be approximated by using the square of the scoring ratio, perhaps we should try a formula to predict winning percentage, such as

$$\frac{R^{\exp}}{R^{\exp}+1}. \tag{2}$$

If we vary exp (exponent) in (2) we can make (2) better fit the actual dependence of winning percentage on scoring ratio for different sports. For baseball, we will allow exp in (2) to vary between 1 and 3. Of course, exp = 2 reduces to the Pythagorean Theorem.

Figure 1.2 shows how MAD changes as we vary exp between 1 and 3.[2] We see that indeed exp = 1.9 yields the smallest MAD (1.96%). An exp value of 2 is almost as good (MAD of 1.97%), so for simplicity we will stick with Bill James's view that exp = 2. Therefore, exp = 2 (or 1.9) yields the best forecasts if we use an equation of form (2). Of course, there might be another equation that predicts winning percentage better than the Pythagorean Theorem from runs scored and allowed. The Pythagorean Theorem is simple and intuitive, however, and works very well. After all, we are off in predicting team wins by an average of $162 \times .02$, which is approximately three wins per team. Therefore, I see no reason to look for a more complicated (albeit slightly more accurate) model.

[1] The actual errors were not simply averaged because averaging positive and negative errors would result in positive and negative errors canceling out. For example, if one team wins 5% more games than (1)′ predicts and another team wins 5% fewer games than (1)′ predicts, the average of the errors is 0 but the average of the absolute errors is 5%. Of course, in this simple situation estimating the average error as 5% is correct while estimating the average error as 0% is nonsensical.

[2] See the chapter appendix for an explanation of how Excel's great Data Table feature was used to determine how MAD changes as exp varied between 1 and 3.

	N	O
2		**EXP**
3		2
4	**Variation of MAD as Exp changes**	
5		**MAD**
6	**Exp**	0.0197
7	1.0	0.0318
8	1.1	0.0297
9	1.2	0.0277
10	1.3	0.0259
11	1.4	0.0243
12	1.5	0.0228
13	1.6	0.0216
14	1.7	0.0206
15	1.8	0.0200
16	1.9	0.0196
17	2.0	0.0197
18	2.1	0.0200
19	2.2	0.0207
20	2.3	0.0216
21	2.4	0.0228
22	2.5	0.0243
23	2.6	0.0260
24	2.7	0.0278
25	2.8	0.0298
26	2.9	0.0318
27	3.0	0.0339

Figure 1.2. Dependence of Pythagorean Theorem accuracy on exponent. See file Standings.xls.

How Well Does the Pythagorean Theorem Forecast?

To test the utility of the Pythagorean Theorem (or any prediction model), we should check how well it forecasts the future. I compared the Pythagorean Theorem's forecast for each MLB playoff series (1980–2007) against a prediction based just on games won. For each playoff series the Pythagorean method would predict the winner to be the team with the higher scoring ratio, while the "games won" approach simply predicts the winner of a playoff series to be the team that won more games. We found that the Pythagorean approach correctly predicted 57 of 106 playoff series (53.8%) while the "games won" approach correctly predicted the winner of only 50% (50 out of 100) of playoff series.[3] The reader is prob-

[3] In six playoff series the opposing teams had identical win-loss records so the "Games Won" approach could not make a prediction.

ably disappointed that even the Pythagorean method only correctly forecasts the outcome of less than 54% of baseball playoff series. I believe that the regular season is a relatively poor predictor of the playoffs in baseball because a team's regular season record depends greatly on the performance of five starting pitchers. During the playoffs teams only use three or four starting pitchers, so much of the regular season data (games involving the fourth and fifth starting pitchers) are not relevant for predicting the outcome of the playoffs.

For anecdotal evidence of how the Pythagorean Theorem forecasts the future performance of a team better than a team's win-loss record, consider the case of the 2005 Washington Nationals. On July 4, 2005, the Nationals were in first place with a record of 50–32. If we extrapolate this winning percentage we would have predicted a final record of 99–63. On July 4, 2005, the Nationals scoring ratio was .991. On July 4, 2005, (1)′ would have predicted a final record of 80–82. Sure enough, the poor Nationals finished 81–81.

The Importance of the Pythagorean Theorem

Baseball's Pythagorean Theorem is also important because it allows us to determine how many extra wins (or losses) will result from a trade. Suppose a team has scored 850 runs during a season and has given up 800 runs. Suppose we trade a shortstop (Joe) who "created"[4] 150 runs for a shortstop (Greg) who created 170 runs in the same number of plate appearances. This trade will cause the team (all other things being equal) to score 20 more runs $(170-150=20)$. Before the trade, $R = \frac{850}{800} = 1.0625$, and we would predict the team to have won $\frac{162(1.0625)^2}{1+(1.0625)^2} = 85.9$ games. After the trade, $R=\frac{870}{800}=1.0875$, and we would predict the team to win $\frac{162(1.0875)^2}{1+(1.0875)^2}=87.8$ games. Therefore, we estimate the trade makes our team 1.9 games better $(87.8-85.9=1.9)$. In chapter 9, we will see how the Pythagorean Theorem can be used to help determine fair salaries for MLB players.

[4] In chapters 2–4 we will explain in detail how to determine how many runs a hitter creates.

Football and Basketball "Pythagorean Theorems"

Does the Pythagorean Theorem hold for football and basketball? Daryl Morey, the general manager for the Houston Rockets, has shown that for the NFL, equation (2) with exp = 2.37 gives the most accurate predictions for winning percentage while for the NBA, equation (2) with exp = 13.91 gives the most accurate predictions for winning percentage. Figure 1.3 gives the predicted and actual winning percentages for the NFL for the 2006 season, while figure 1.4 gives the predicted and actual winning percentages for the NBA for the 2006–7 season.

For the 2005–7 NFL seasons, MAD was minimized by exp = 2.7. Exp = 2.7 yielded a MAD of 5.9%, while Morey's exp = 2.37 yielded a MAD of 6.1%. For the 2004–7 NBA seasons, exp = 15.4 best fit actual winning percentages. MAD for these seasons was 3.36% for exp = 15.4 and 3.40% for exp = 13.91. Since Morey's values of exp are very close in accuracy to the values we found from recent seasons we will stick with Morey's values of exp.

These predicted winning percentages are based on regular season data. Therefore, we could look at teams that performed much better than expected during the regular season and predict that "luck would catch up

	B	C	D	E	F	G	H	I	J	K	L	M	N
3			exp = 2.4						MAD = 0.061497				
4													
5	Year	Team	Wins	Losses	Points for	Points against	Ratio	Predicted winning %	Annual winning %	abserr		exp	MAD
6	2007	N.E. Patriots	16	0	589	274	2.149635	0.859815262	1	0.140185			0.061497
7	2007	B. Bills	7	9	252	354	0.711864	0.308853076	0.4375	0.128647		1.5	0.08419
8	2007	N.Y. Jets	4	12	268	355	0.75493	0.339330307	0.25	0.08933		1.6	0.080449
9	2007	M.Dolphins	1	15	267	437	0.610984	0.237277785	0.625	0.174778		1.7	0.077006
10	2007	C. Browns	10	6	402	382	1.052356	0.530199349	0.625	0.094801		1.8	0.073795
11	2007	P. Steelers	10	6	393	269	1.460967	0.710633507	0.625	0.085634		1.9	0.070675
12	2007	C. Bengals	7	9	380	385	0.987013	0.492255411	0.4375	0.054755		2	0.068155
13	2007	B. Ravens	5	11	275	384	0.716146	0.311894893	0.3125	0.000605		2.1	0.06588
14	2007	I. Colts	13	3	450	262	1.717557	0.782779877	0.8125	0.02972		2.2	0.064002
15	2007	J. Jaguars	11	5	411	304	1.351974	0.67144112	0.6875	0.016059		2.3	0.062394
16	2007	T. Titans	10	6	301	297	1.013468	0.507925876	0.625	0.117074		2.4	0.061216
17	2007	H. Texans	8	8	379	384	0.986979	0.492235113	0.5	0.007765		2.5	0.060312
18	2007	S.D. Chargers	11	5	412	284	1.450704	0.707186057	0.6875	0.019686		2.6	0.059554
19	2007	D. Broncos	7	9	320	409	0.782396	0.35856816	0.4375	0.078932	best!	2.7	0.059456
20	2007	O. Raiders	4	12	283	398	0.711055	0.308278013	0.25	0.058278		2.8	0.059828
21	2007	K.C. Chiefs	4	12	226	335	0.674627	0.282352662	0.25	0.032353		2.9	0.060934
22	2007	D. Cowboys	13	3	455	325	1.4	0.689426435	0.8125	0.123074		3	0.062411
23	2007	N.Y. Giants	10	6	373	351	1.062678	0.535957197	0.625	0.089043		3.4	0.063891

Figure 1.3. Predicted NFL winning percentages. Exp = 2.4. See file Sportshw1.xls.

	E	F	G	H	I	J	K
37	2006–2007 NBA						**MAD = 0.05**
38							
39	**Team**	**PF**	**PA**	**Ratio**	**Predicted Win %**	**Actual Win %**	**Abs. Error**
40	Phoenix Suns	110.2	102.9	1.07	0.722	0.744	0.022
41	Golden State Warriors	106.5	106.9	1.00	0.487	0.512	0.025
42	Denver Nuggets	105.4	103.7	1.02	0.556	0.549	0.008
43	Washington Wizards	104.3	104.9	0.99	0.480	0.500	0.020
44	L.A. Lakers	103.3	103.4	1.00	0.497	0.512	0.016
45	Memphis Grizzlies	101.6	106.7	0.95	0.336	0.268	0.068
46	Utah Jazz	101.5	98.6	1.03	0.599	0.622	0.022
47	Sacramento Kings	101.3	103.1	0.98	0.439	0.395	0.044
48	Dallas Mavericks	100	92.8	1.08	0.739	0.817	0.078
49	Milwaukee Bucks	99.7	104	0.96	0.357	0.341	0.016
50	Toronto Raptors	99.5	98.5	1.01	0.535	0.573	0.038
51	Seattle Supersonics	99.1	102	0.97	0.401	0.378	0.023
52	Chicago Bulls	98.8	93.8	1.05	0.673	0.598	0.076
53	San Antonio Spurs	98.5	90.1	1.09	0.776	0.707	0.068
54	New Jersey Nets	97.6	98.3	0.99	0.475	0.500	0.025
55	New York Knicks	97.5	100.3	0.97	0.403	0.402	0.000
56	Houston Rockets	97	92.1	1.05	0.673	0.634	0.039
57	Charlotte Bobcats	96.9	100.6	0.96	0.373	0.402	0.030
58	Cleveland Cavaliers	96.8	92.9	1.04	0.639	0.610	0.029
59	Minnesota Timberwolves	96.1	99.7	0.96	0.375	0.395	0.020
60	Detroit Pistons	96	91.8	1.05	0.651	0.646	0.004
61	Boston Celtics	95.8	99.2	0.97	0.381	0.293	0.088
62	Indiana Pacers	95.6	98	0.98	0.415	0.427	0.012
63	L.A. Clippers	95.6	96.1	0.99	0.482	0.952	0.471
64	New Orleans Hornets	95.5	97.1	0.98	0.442	0.476	0.033
65	Philadelphia 76ers	94.9	98	0.97	0.390	0.427	0.037
66	Orlando Magic	94.8	94	1.01	0.529	0.488	0.042
67	Miami Heat	94.6	95.5	0.99	0.467	0.537	0.069
68	Portland Trail Blazers	94.1	98.4	0.96	0.349	0.390	0.041
69	Atlanta Hawks	93.7	98.4	0.95	0.336	0.366	0.030

Figure 1.4. Predicted NBA winning percentages. Exp = 13.91. See file Footballbasketballpythagoras.xls.

with them." This train of thought would lead us to believe that these teams would perform worse during the playoffs. Note that the Miami Heat and Dallas Mavericks both won about 8% more games than expected during the regular season. Therefore, we would have predicted Miami and Dallas to perform worse during the playoffs than their actual win-loss record indicated. Sure enough, both Dallas and Miami suffered unexpected first-round defeats. Conversely, during the regular season the San Antonio Spurs and Chicago Bulls won around 8% fewer games than the Pythagorean Theorem predicts, indicating that these teams would perform better than expected in the playoffs. Sure enough, the Bulls upset the Heat and gave the Detroit Pistons a tough time. Of course, the Spurs won the 2007 NBA title. In addition, the Pythagorean Theorem had the Spurs as by far the league's best team (78% predicted winning percentage). Note the team that underachieved the most was the Boston Celtics, who won nearly 9% fewer (or 7) games than predicted. Many people suggested the Celtics "tanked" games during the regular season to improve their chances of obtaining potential future superstars such as Greg Oden and Kevin Durant in the 2007 draft lottery. The fact that the Celtics won seven fewer games than expected does not prove this conjecture, but it is certainly consistent with the view that Celtics did not go all out to win every close game.

APPENDIX

Data Tables

The Excel Data Table feature enables us to see how a formula changes as the values of one or two cells in a spreadsheet are modified. This appendix shows how to use a One Way Data Table to determine how the accuracy of (2) for predicting team winning percentage depends on the value of exp. To illustrate, let's show how to use a One Way Data Table to determine how varying exp from 1 to 3 changes the average error in predicting a MLB team's winning percentage (see figure 1.2).

Step 1. We begin by entering the possible values of exp (1, 1.1, . . . 3) in the cell range N7:N27. To enter these values, simply enter 1 in N7, 1.1 in N8, and select the cell range N8. Now drag the cross in the lower right-hand corner of N8 down to N27.

Step 2. In cell O6 we enter the formula we want to loop through and calculate for different values of exp by entering the formula = J1.

Step 3. In Excel 2003 or earlier, select Table from the Data Menu. In Excel 2007 select Data Table from the What If portion of the ribbon's Data tab (figure 1-a).

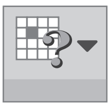

Figure 1-a. What If icon for Excel 2007.

Step 4. Do not select a row input cell but select cell L2 (which contains the value of exp) as the column input cell. After selecting OK we see the results shown in figure 1.2. In effect Excel has placed the values 1, 1.1, . . . 3 into cell M2 and computed our MAD for each listed value of exp.