CONTENTS

PART VI METHODS AND MISCELLANEOUS

CHAPTER 1

BASEBALL'S PYTHAGOREAN THEOREM

The more runs that 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 standings and found that the percentage of games won by a baseball team can be well approximated by the formula

$$\frac{\text{runs scored}^2}{\text{runs scored}^2 + \text{runs allowed}^2} = \text{Estimate of percentage of games won.} \quad (1)$$

This formula has several desirable properties:

- 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 length a and b. Recall from high school geometry that the Pythagorean Theorem states that a triangle is a right triangle if and only if $a^2 + b^2 = c^2$ must hold. 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 (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 (see file Mathleticschapter1files.xlsx for all of this chapter's analysis) shows how well (1′) predicts teams' winning percentages for Major League Baseball teams during the 2005–2016 seasons. For example, the 2016 Los Angeles Dodgers scored 725 runs and gave up 638 runs. Their scoring ratio was $R=\frac{725}{638}=1.136$. Their predicted win percentage from Baseball's Pythagorean Theorem was $\frac{1.136^2}{1.136^2+1}=.5636$. The 2016 Dodgers actually won a fraction $\frac{91}{162}=.5618$ of their games. Thus (1′) was off by 0.18% in predicting the percentage of games won by the Dodgers in 2016.

For each team define Error in Win Percentage Prediction to equal Actual Winning Percentage minus Predicted Winning Percentage. For example, for the 2016 Atlanta Braves, Error $=.42-.41=.01$ (or 1.0%), and for the 2016 Colorado Rockies, Error $=.46-.49=-.03$ (or 3%). A positive error means that the team won more games than predicted while a negative error means the team won fewer games than predicted. Column J computes for each team the absolute value of the prediction error. Recall that absolute value of a number is simply the distance of the number from 0. That is, $|5|=|-5|=5$. In cell J1 we average the absolute prediction errors for each team to obtain a measure of how well our predicted win percentages fit the actual team winning percentages. The average of absolute forecasting

	A	B	C	D	E	F	G	H	I	J
1						exp	2.000		MAD:	0.021
2	Year	Team	Wins	Losses	Runs	Opp Runs	Ratio	Pred W–L%	Act W–L%	Error
3	2016	ARI	69	93	752	890	0.845	0.42	0.43	0.009
4	2016	ATL	68	93	649	779	0.833	0.41	0.42	0.010
5	2016	BAL	89	73	744	715	1.041	0.52	0.55	0.030
6	2016	BOS	93	69	878	694	1.265	0.62	0.57	0.041
7	2016	CHC	103	58	808	556	1.453	0.68	0.64	0.043
8	2016	CHW	78	84	686	715	0.959	0.48	0.48	0.002
9	2016	CIN	68	94	716	854	0.838	0.41	0.42	0.007
10	2016	CLE	94	67	777	676	1.149	0.57	0.58	0.011
11	2016	COL	75	87	845	860	0.983	0.49	0.46	0.028
12	2016	DET	86	75	750	721	1.040	0.52	0.53	0.011
13	2016	HOU	84	78	724	701	1.033	0.52	0.52	0.002
14	2016	KCR	81	81	675	712	0.948	0.47	0.50	0.027
15	2016	LAA	74	88	717	727	0.986	0.49	0.46	0.036
16	2016	LAD	91	71	725	638	1.136	0.56	0.56	0.002
17	2016	MIA	79	82	655	682	0.960	0.48	0.49	0.008
18	2016	MIL	73	89	671	733	0.915	0.46	0.45	0.005
19	2016	MIN	59	103	722	889	0.812	0.40	0.36	0.033
20	2016	NYM	87	75	671	617	1.088	0.54	0.54	0.005
21	2016	NYY	84	78	680	702	0.969	0.48	0.52	0.034

FIGURE 1.1 Baseball's Pythagorean Theorem 2005–2016.

errors is called the MAD (mean absolute deviation).[1] We find that for our dataset the predicted winning percentages of the Pythagorean Theorem were off by an average of 2.17% per team.

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

$$\frac{R^{exp}}{R^{exp}+1}. \qquad (2)$$

If we vary exp in (2) we can make (2) better fit the actual dependence of winning percentage on the scoring ratio for different sports.

1. Why didn't we just average the actual errors? 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% less 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.

	N	O
5		MAD
6		0.021
7	1.1	0.02812245
8	1.2	0.02617963
9	1.3	0.02441563
10	1.4	0.02289267
11	1.5	0.02160248
12	1.6	0.02069009
13	1.7	0.02014272
14	1.8	0.0199295
15	1.9	0.0201094
16	2	0.020513
17	2.1	0.02114432
18	2.2	0.02208793
19	2.3	0.02328749
20	2.4	0.02473436
21	2.5	0.02640258
22	2.6	0.02823811
23	2.7	0.03019355
24	2.8	0.03228514
25	2.9	0.03447043
26	3	0.03670606

FIGURE 1.2 Dependence of Pythagorean Theorem Accuracy on Exponent.

For baseball, we will allow exp in (2) (exp is short for exponent) to vary between 1 and 3. Of course exp = 2 reduces to the Pythagorean Theorem.

Figure 1-2 shows how the MAD changes as we vary exp between 1 and 3. This was done using the Data Table feature in Excel.[2] We see that indeed exp = 1.8 yields the smallest MAD (1.99%). An exp value of 2 is almost as good (MAD of 2.05%), so for simplicity we will stick with Bill James's view that exp = 2. Therefore exp = 2 (or 1.8) 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 Pythago-

2. See Chapter 1 Appendix for an explanation of how we used Data Tables to determine how MAD changes as we vary exp between 1 and 3. Additional information available at https://support.office.com/en-us/article/calculate-multiple-results-by-using-a-data-table-e95e2487-6ca6-4413-ad12-77542a5ea50b.

rean Theorem is simple and intuitive, however, and does very well. After all, we are off in predicting team wins by an average of 162 ∗ .0205, which is approximately three wins per team. Therefore, I see no reason to look for a more complicated (albeit slightly more accurate) model.

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. We chose to compare the Pythagorean Theorem's forecast for each Major League Baseball playoff series (2005–2016) 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 46 of 84 playoff series (54.8%) while the "games won" approach correctly predicted the winner of only 55% (44 out of 80) playoff series.[3] The reader is probably disappointed that even the Pythagorean method only correctly forecasts the outcome of under 54% of baseball playoff series. We believe that the regular season is a relatively poor predictor of the playoffs in baseball because a team's regular season record depends a lot on the performance of five starting pitchers. During the playoffs, teams only use three or four starting pitchers, so a lot 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 had extrapolated this win percentage, we would have predicted

3. In four playoff series the opposing teams had identical win-loss records, so the "games won" approach could not make a prediction.

a final record of 99–63. On July 4, 2005, the Nationals' scoring ratio was .991. On July 4, 2005, equation (1) would predict the Nationals to win around half (40) of the remaining 80 games and finish with a 90–72 record. In reality, the Nationals only won 31 of their remaining games and finished at 81–81!

IMPORTANCE OF PYTHAGOREAN THEOREM

The Baseball Pythagorean Theorem is also important because it allows us to determine how many extra wins (or losses) will result from a trade. As an example, suppose a team has scored 850 runs during a season and also given up 800 runs. Suppose we trade an SS (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 $170-150=20$ more runs. 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 have won $\frac{162*1.0875^2}{1+1.0875^2}=87.8$ games. Therefore, we estimate the trade makes our team $87.8-85.9=1.9$ games better. In Chapter 9, we will see how the Pythagorean Theorem can be used to help determine fair salaries for Major League Baseball players.

FOOTBALL AND BASKETBALL "PYTHAGOREAN THEOREMS"

Does the Pythagorean Theorem hold for football and basketball? Daryl Morey, currently the General Manager for the Houston Rockets NBA team, has shown that for the NFL, equation (2) with

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

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1								Exp	2.370	MAD	0.051			
2	Year	Team	Wins	Losses	Ties	PF	PA	Ratio	Pred W–L%	Act W-L%	Error			
3	2015	Arizona Cardinals	13	3	0	489	313	1.56	0.742	0.813	0.071			
4	2015	Atlanta Falcons	8	8	0	339	345	0.98	0.490	0.5	0.010			MAD
5	2015	Baltimore Ravens	5	11	0	328	401	0.82	0.383	0.313	0.070		Exp	0.051130558
6	2015	Buffalo Bills	8	8	0	379	359	1.06	0.532	0.5	0.032		1.5	0.087458019
7	2015	Carolina Panthers	15	1	0	500	308	1.62	0.759	0.938	0.179		1.6	0.083786393
8	2015	Chicago Bears	6	10	0	335	397	0.84	0.401	0.375	0.026		1.7	0.080410576
9	2015	Cincinnati Bengals	12	4	0	419	279	1.50	0.724	0.75	0.026		1.8	0.077291728
10	2015	Cleveland Browns	3	13	0	278	432	0.64	0.260	0.188	0.072		1.9	0.074380834
11	2015	Dallas Cowboys	4	12	0	275	374	0.74	0.325	0.25	0.075		2	0.071698879
12	2015	Denver Broncos	12	4	0	355	296	1.20	0.606	0.75	0.144		2.1	0.069282984
13	2015	Detroit Lions	7	9	0	358	400	0.90	0.435	0.438	0.003		2.2	0.067048672
14	2015	Green Bay Packers	10	6	0	368	323	1.14	0.577	0.625	0.048		2.3	0.065010818
15	2015	Houston Texans	9	7	0	339	313	1.08	0.547	0.563	0.016		2.4	0.063455288
16	2015	Indianapolis Colts	8	8	0	333	408	0.82	0.382	0.5	0.118		2.5	0.062158811
17	2015	Jacksonville Jaguars	5	11	0	376	448	0.84	0.398	0.313	0.085		2.6	0.061279631
18	2015	Kansas City Chiefs	11	5	0	405	287	1.41	0.693	0.688	0.005		2.7	0.060819271
19	2015	Miami Dolphins	6	10	0	310	389	0.80	0.369	0.375	0.006		2.8	0.060758708
20	2015	Minnesota Vikings	11	5	0	365	302	1.21	0.610	0.688	0.078		2.9	0.060941558
21	2015	New England Patriots	12	4	0	465	315	1.48	0.716	0.75	0.034		3	0.061357921
22	2015	New Orleans Saints	7	9	0	408	476	0.86	0.410	0.438	0.028		3.1	0.061891886
23	2015	New York Giants	6	10	0	420	442	0.95	0.470	0.375	0.095		3.2	0.062648637
24	2015	New York Jets	10	6	0	387	314	1.23	0.621	0.625	0.004		3.3	0.063594958
25	2015	Oakland Raiders	7	9	0	359	399	0.90	0.438	0.438	0.000		3.4	0.06474528
26	2015	Philadelphia Eagles	7	9	0	377	430	0.88	0.423	0.438	0.015		3.5	0.065955742
27	2015	Pittsburgh Steelers	10	6	0	423	319	1.33	0.661	0.625	0.036			

FIGURE 1.3 Predicted NFL Winning Percentages: Exp = 2.37.

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 2015 NFL, while Figure 1-4 gives the predicted and actual winning percentages for the 2015–2016 NBA. See the file Sportshw1.xls

For the 2008–2015 NFL seasons we found MAD was minimized by exp = 2.8. Exp = 2.8 yielded a MAD of 6.08%, while Morey's exp = 2.37 yielded a MAD of 6.39%. For the NBA seasons 2008–2016 we found exp = 14.4 best fit actual winning percentages. The MAD for these seasons was 2.84% for exp = 14.4 and 2.87% 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. See file Sportshw1.xls.

Assuming the errors in our forecasts follow a normal random variable (which turns out to be a reasonable assumption) we would

	A	B	C	D	E	F	G	H	I	J	K	L	M
1							Exp	13.910	MAD	0.0287			
2	Year	Team	Wins	Losses	Points	Opp Points	Ratio	Pred W-L%	Act W-L%	Error			
3	2015-16	Atlanta Hawks	48	34	8433	8137	1.04	0.622	0.585	0.037			
4	2015-16	Boston Celtics	48	34	8669	8406	1.03	0.606	0.585	0.021			
5	2015-16	Brooklyn Nets	21	61	8089	8692	0.93	0.269	0.256	0.013		Exp	0.0287
6	2015-16	Charlotte Hornets	48	34	8479	8256	1.03	0.592	0.585	0.007		12	0.0340286
7	2015-16	Chicago Bulls	42	40	8335	8456	0.99	0.450	0.512	0.062		12.2	0.0332135
8	2015-16	Cleveland Cavaliers	57	25	8555	8063	1.06	0.695	0.695	6E-05		12.4	0.0324282
9	2015-16	Dallas Mavericks	42	40	8388	8413	1	0.490	0.512	0.022		12.6	0.0317199
10	2015-16	Denver Nuggets	33	49	8355	8609	0.97	0.397	0.402	0.005		12.8	0.0310445
11	2015-16	Detroit Pistons	44	38	8361	8311	1.01	0.521	0.537	0.016		13	0.0304509
12	2015-16	Golden State Warriors	73	9	9421	8539	1.1	0.797	0.89	0.093		13.2	0.0298964
13	2015-16	Houston Rockets	41	41	8737	8721	1	0.506	0.5	0.006		13.4	0.0294269
14	2015-16	Indiana Pacers	45	37	8377	8237	1.02	0.558	0.549	0.009		13.6	0.0290408
15	2015-16	Los Angeles Clippers	53	29	8569	8218	1.04	0.641	0.646	0.005		13.8	0.0287533
16	2015-16	Los Angeles Lakers	17	65	7982	8766	0.91	0.214	0.207	0.007		14	0.0285995
17	2015-16	Memphis Grizzlies	42	40	8126	8310	0.98	0.423	0.512	0.089		14.2	0.0284997
18	2015-16	Miami Heat	48	34	8204	8069	1.02	0.557	0.585	0.028		14.4	0.0284481
19	2015-16	Milwaukee Bucks	33	49	8122	8465	0.96	0.360	0.402	0.042		14.6	0.0284727
20	2015-16	Minnesota Timberwolves	29	53	8398	8688	0.97	0.384	0.354	0.03		14.8	0.028568
21	2015-16	New Orleans Pelicans	30	52	8423	8734	0.96	0.377	0.366	0.011		15	0.0287573
22	2015-16	New York Knicks	32	50	8065	8289	0.97	0.406	0.39	0.016		15.2	0.0289692
23	2015-16	Oklahoma City Thunder	55	27	9038	8441	1.07	0.721	0.671	0.05		15.4	0.0292675
24	2015-16	Orlando Magic	35	47	8369	8502	0.98	0.445	0.427	0.018		15.6	0.0296178
25	2015-16	Philadelphia 76ers	10	72	7988	8827	0.9	0.200	0.122	0.078		15.8	0.0300081
26	2015-16	Phoenix Suns	23	59	8271	8817	0.94	0.291	0.28	0.011		16	0.0304529
27	2015-16	Portland Trail Blazers	44	38	8622	8554	1.01	0.528	0.537	0.009			

FIGURE 1.4 Predicted NBA Winning Percentages: Exp = 13.91.

expect around 95% of our NBA win forecasts to be accurate within 2.5 * MAD = 7.3%. Over 82 games this is about 6 games. So whenever the Pythagorean forecast for wins is off by more than six games, the Pythagorean prediction is an "outlier." When we spot outliers we try and explain why they occurred. The 2006–2007 Boston Celtics had a scoring ratio of .966, and Pythagoras predicts the Celtics should have won 31 games. They won seven fewer games (24). During that season many people suggested the Celtics "tanked" games to improve their chance of having the #1 pick (Greg Oden and Kevin Durant went 1–2) in the draft lottery. The shortfall in the Celtics' wins does not prove this conjecture, but the evidence is consistent with the Celtics winning substantially fewer games than chance would indicate.

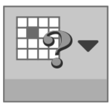
CHAPTER 1 APPENDIX: DATA TABLES

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. In this appendix we show 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 our average error in predicting an MLB's team 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:N26. To enter these values we simply enter 1 in N7 and 1.1 in N8 and select the cell range N7:N8. Now we drag the cross in the lower right-hand corner of N8 down to N26.

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. Then we select the "table range" N6:O26.

Step 3: Now we select Data Table from the What If section of the ribbon's Data tab.



Step 4: We leave the row input cell portion of the dialog box blank but select cell G1 (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 G1 and computed our MAD for each listed value of exp.

INDEX

The letters t or f following a page number indicate a table or figure on that page.