Yale SCHOOL *of* MANAGEMENT

YALE UNIVERSITY
School of Public Health

YALE UNIVERSITY
School of Engineering and
Applied Science

Decomposing Pythagoras

Edward H. Kaplan

William N and Marie A Beach Professor of Operations Research

Professor of Public Health

Professor of Engineering

Yale University

Candler Rich

Applied Mathematics Program

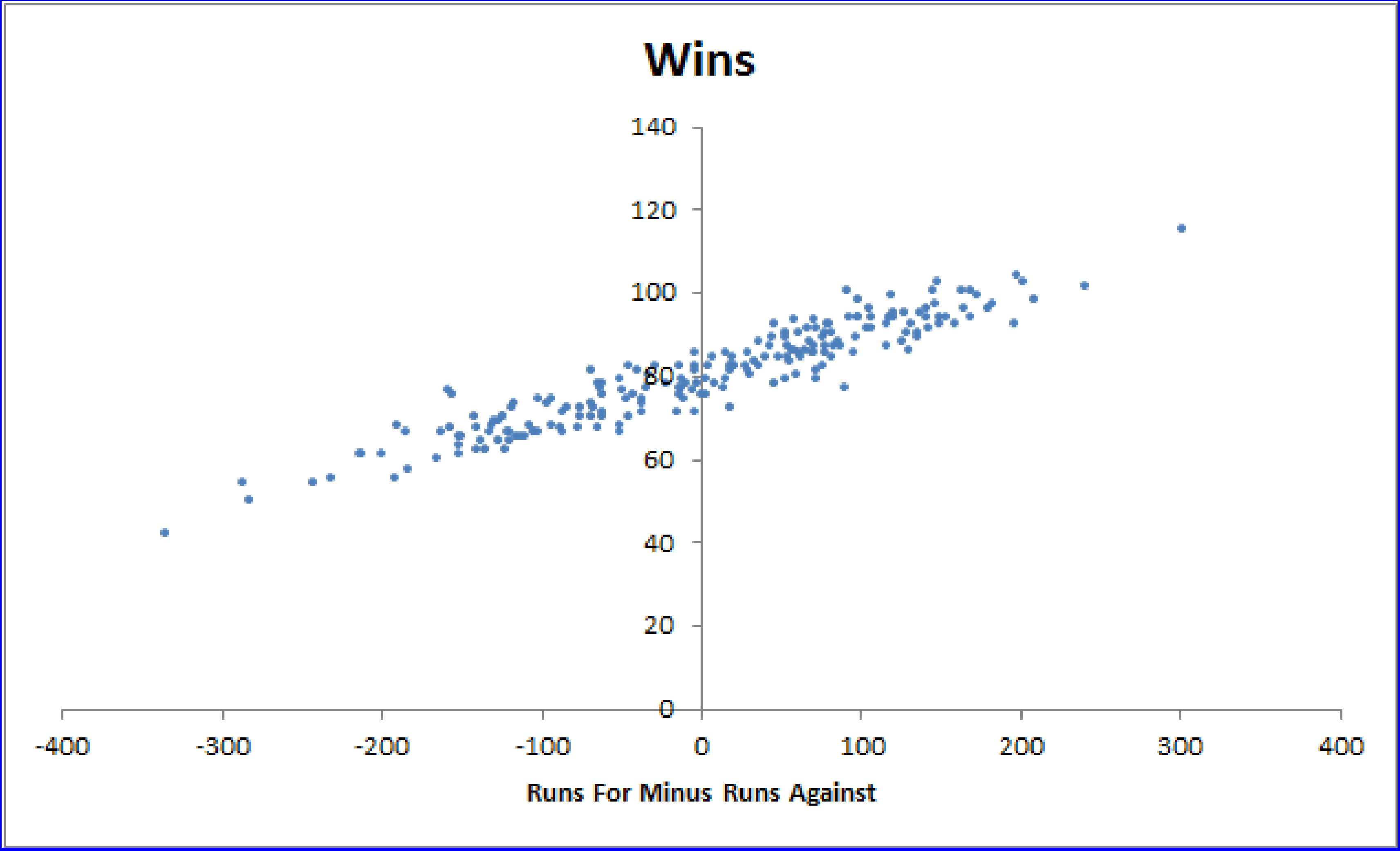
Yale College

Let's Start With Baseball

- Problem: given data for every MLB team over several seasons reporting the total number of wins, and the seasonal difference between runs for and against (*run differential*), predict the number of wins from the run differential

Let's Look At The Data

- For 2000-2006 MLB seasons:



Pose Question To MBA Students…

- They all want to commit an act of regression!!
- So, before punching buttons in Excel…
- Tell me: what is the intercept, and what is the slope?
  – Huh? That's why we run the regression Prof!
- I know, I know, but humor me. What is the intercept, and what is the slope?

Predicting Wins From Run Diff

- How many MLB games per season?
  - 162
- Great. So what is the average number of wins per team per season?
  - Uhhh…81!
- So if Wins = a + b * Run Diff, what is a?
  - tick…tick…tick… 81!
- Mazel Tov
- (Why? What is average Run Diff? Zero!)

Predicting Wins From Run Diff

- Now, on average, how many runs per team per season?
  - check data… 775.5
- And how many runs against on average?
  - don't check data… 775.5
- What is ratio of average number of wins per runs for (runs against)?
  - 81/775.5 = 0.10
- So, if Wins = 81 + b * Run Diff, what's b?
  - 0.10 (!!)

Back-of-Envelope Result…

- Wins = 81 + 0.10 * Run Diff
- OK, *now* run a regression

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.94
R Square	0.89
Adjusted R Square	0.89
Standard Error	4.04
Observations	210

	Coefficients	*Standard Error*	*t Stat*
Intercept	80.94	0.28	289.99
Run Diff	0.10	0.00	41.05

- Estimated intercept: 80.94 (vs 81)
- Estimated slope: 0.10 (vs 0.10)

Wow! Let’s Try Basketball!

- Try to predict seasonal wins from difference in points for and against

2010-11 NBA Season Stats	
Games per team per season?	82
Average wins per team?	41
Points per team per season	8163.373
Average points per win?	199.107
Average wins per point?	0.005
Back of Envelope suggests	
Wins = 41 + .005*Point Differential	

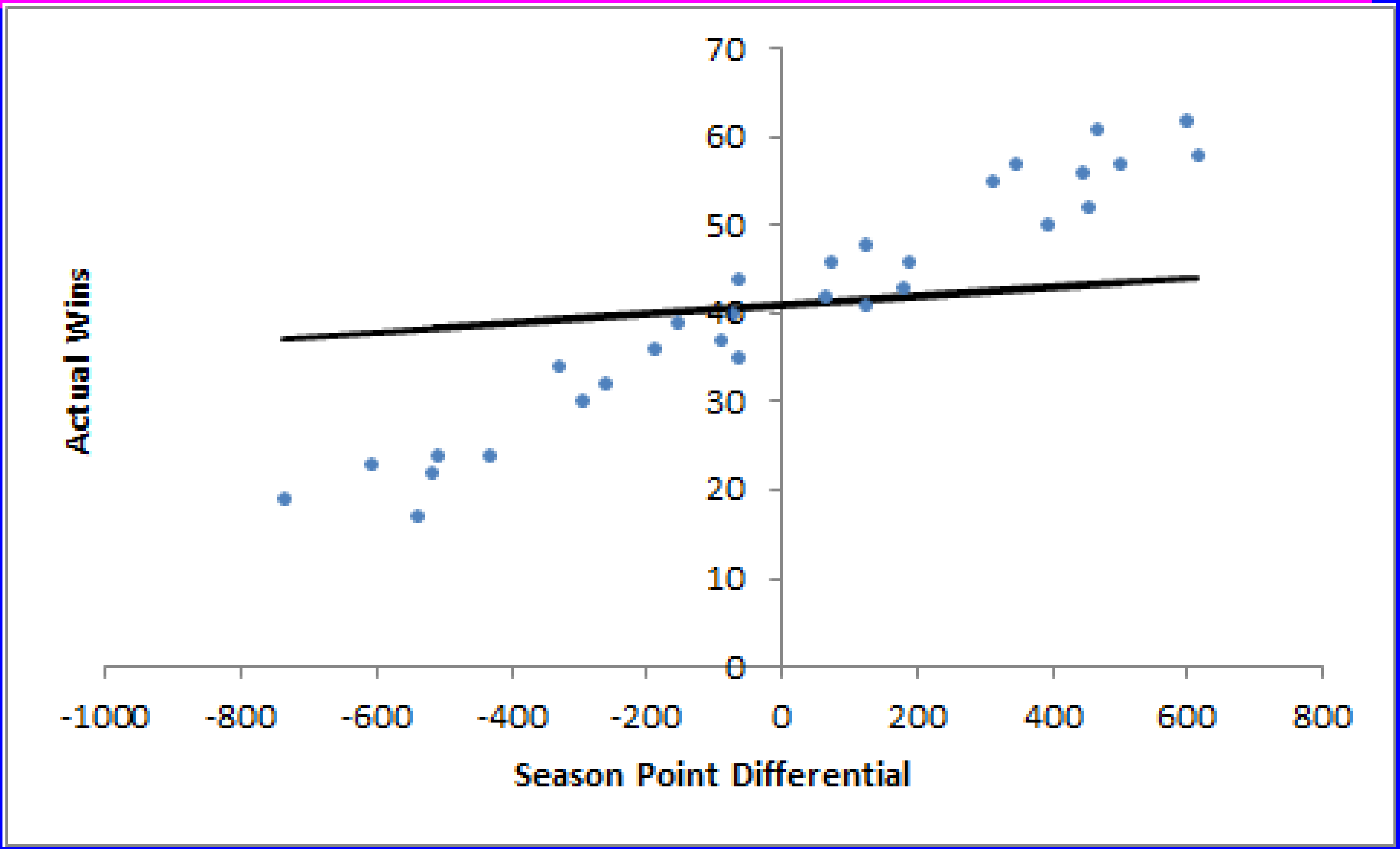
Basketball Regression

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.97
R Square	0.94
Adjusted R Squar	0.94
Standard Error	3.17
Observations	30

	Coefficients	*Standard Error*	*t Stat*
Intercept	40.991	0.578	70.915
Point Differential	0.033	0.002	21.777

- Back of envelope: Wins = 41 + 0.005 * Pt Diff
- What happened?

What Went Wrong?



- Worked for baseball, why not basketball?
- We'll get back to this, but first…

Bill James

- Published highly influential *Baseball Abstract* from 1977-1988
- First widely-known baseball “Sabermetrician”
- One of the inspirations for Michael Lewis’s book *Moneyball*

James's Pythagorean Model

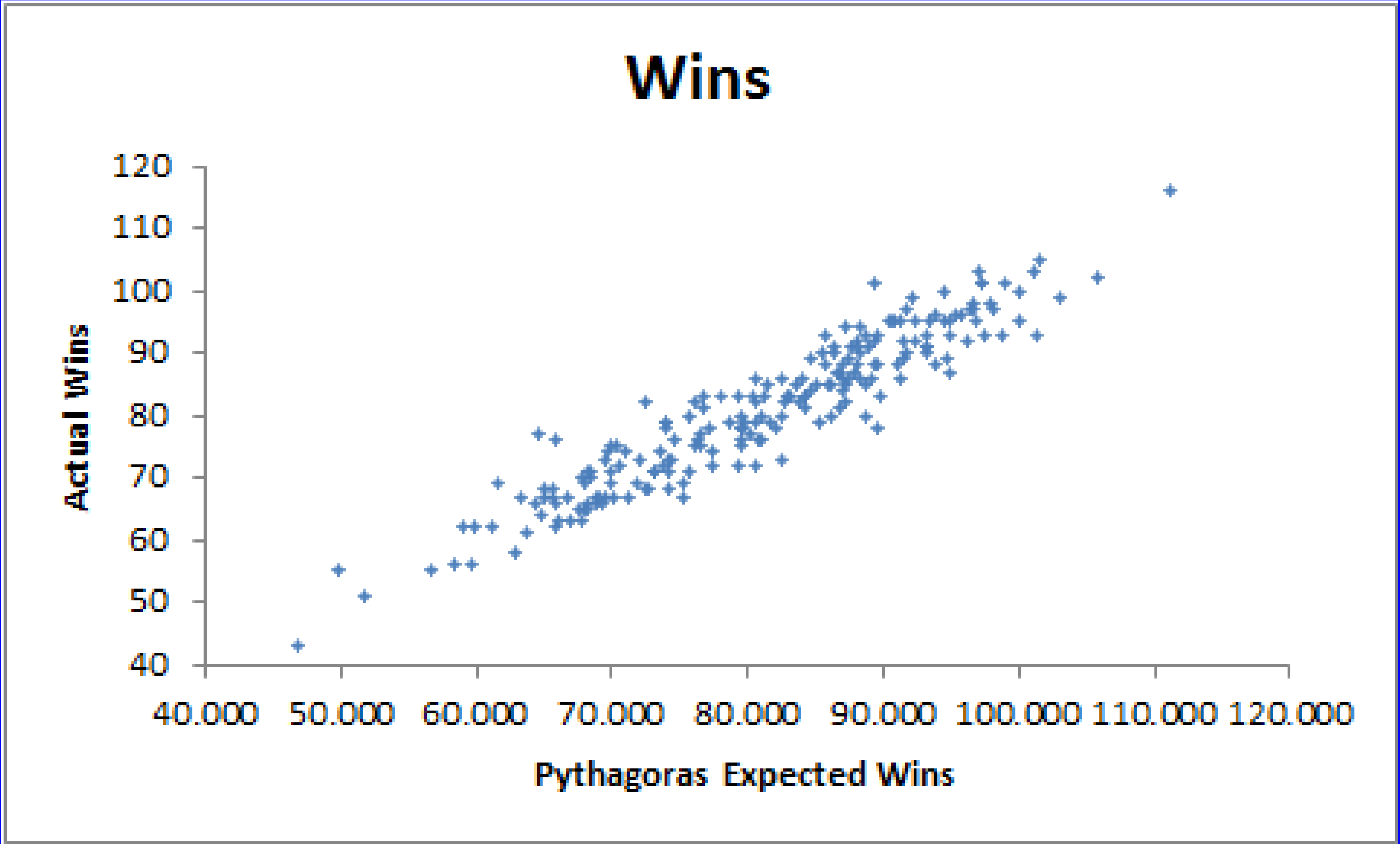
- In 1980, Bill James presented his Pythagorean model for baseball relating seasonal win percentage (*WP*), runs scored (*RS*), and runs against (*RA*)

$$WP = \frac{RS^2}{RS^2 + RA^2}.$$

- He called it the Pythagorean model because the denominator has a sum of squares…

Pythagorean Model

- James was the Johannes Kepler of baseball
- Works pretty well!



Pythagorean Model

- Why "2" in the exponent? Why not γ?

$$WP = \frac{RS^{\gamma}}{RS^{\gamma} + RA^{\gamma}}$$

- Lots of papers in sports modeling literature estimating γ for different sports, e.g.
  - Baseball: $\gamma = 2$
  - Basketball: $\gamma = 14$

Aside: Reverse Engineering Pythagoras

- Miller (*Chance*, 2007): if number of points scored by two teams playing each other are independent and Weibull distributed with same shape parameter but different scale parameters, probability of winning follows the Pythagorean formula

How Does This Relate To…

- …the baseball vs basketball "back of envelope" models?
- Dayaratna and Miller (2012) showed 1st-order Taylor expansion of Pythagoras is

$$WP = \frac{RS^{\gamma}}{RS^{\gamma} + RA^{\gamma}} \approx \frac{1}{2} + \frac{\gamma}{4 \times R_{\text{total}}} \times (RS - RA).$$

- To get baseball model, set $\gamma = 2$, and multiply both sides by 162. Result is that intercept = 81, slope = 81/average runs/team. Perfect!

How Does This Relate To…

$$WP = \frac{RS^{\gamma}}{RS^{\gamma} + RA^{\gamma}} \approx \frac{1}{2} + \frac{\gamma}{4 \times R_{\text{total}}} \times (RS - RA).$$

- To get basketball model, set $\gamma = 14$, and multiply both sides by 82. Result is that intercept = 41, slope = 41*7/average points/team. So need to multiply back of envelope basketball slope by 7!
- Back of envelope basketball slope = 0.005; multiply by 7 to get 0.035
- Basketball regression gave slope = 0.033 (!!)

Worded Differently…

- If n games per season, R_{total} seasonal points on avg
- *Average* wins/point = $\frac{n/2}{R_{total}}$
- *Marginal* wins/point = $\frac{n/2}{R_{total}} \times \frac{\gamma}{2}$

  $= \frac{\gamma}{2} \times$ *Average* wins/point

- For baseball, $\gamma = 2$ and Marginal = Average

- For basketball, $\gamma = 14$ and Marginal = 7 × Average

Still Unsatisfying…

- Pythagoras with Taylor series explains why back of envelope works for baseball, but not for basketball
- But that is just because $\gamma = 2$ for baseball, and 14 for basketball
- Does the Pythagorean γ tell us anything about baseball versus basketball beyond "that's what the data say?"
- What is it about the sports of baseball vs basketball that give rise to the different γ's?

Aside: Quick Way To Estimate γ

$$WP = \frac{RS^{\gamma}}{RS^{\gamma} + RA^{\gamma}} \approx \frac{1}{2} + \frac{\gamma}{4 \times R_{\text{total}}} \times (RS - RA).$$

- Consider running regression

$$WP = \alpha + \beta(RS - RA) + \varepsilon$$

- Then can estimate γ as

$$\widehat{\gamma} \approx 4 \times R_{\text{total}} \times \widehat{\beta},$$

- We'll use this later

An Exact Win Probability Model!

- Let X denote the spread (i.e. point difference) between a team and its opponent
- Normalize RS (RA) as average points for (against) team *per game* (so just divide by 162 for baseball, 82 for basketball, etc.)
- In a randomly chosen game, we have

$$E(X) \approx RS - RA$$

An Exact Win Probability Model!

- Now for any team, define:

$$\Pr\{\text{Win}\} = \Pr\{X > 0\}$$

- We can estimate this by *WP*, the team's observed winning percentage

Now, define a team's expected *margin of victory* by

$$MOV = E(X|X > 0), \tag{9}$$

and similarly define a team's expected *margin of defeat* by

$$MOD = -E(X|X < 0). \tag{10}$$

An Exact Win Probability Model!

- Invoke law of total expectation and write

$$
\begin{aligned}
E(X) &= E(X|X>0)\,\Pr\{X>0\} + E(X|X<0)\,\Pr\{X<0\} \\
&= MOV \times \Pr\{\text{Win}\} - MOD \times (1 - \Pr\{\text{Win}\}) \\
&= (MOD + MOV) \times \Pr\{\text{Win}\} - MOD \qquad (11)
\end{aligned}
$$

- Rearrange terms to arrive at

$$
\Pr\{\text{Win}\} = \frac{MOD}{MOD + MOV} + \frac{1}{MOD + MOV} \times E(X).
$$

An Exact Win Probability Model!

- This win probability equation is also exact at the data level for each team over a season

$$WP_i = \frac{mod_i}{mod_i + mov_i} + \frac{1}{mod_i + mov_i} \times (RS_i - RA_i) \quad (13)$$

where WP_i, RS_i and RA_i are the observed win percentage, runs scored and runs against while mod_i and mov_i are the observed average margins of victory and defeat respectively for the ith team, $i = 1, 2, \ldots, n$.

Exact Model For 2016 MLB

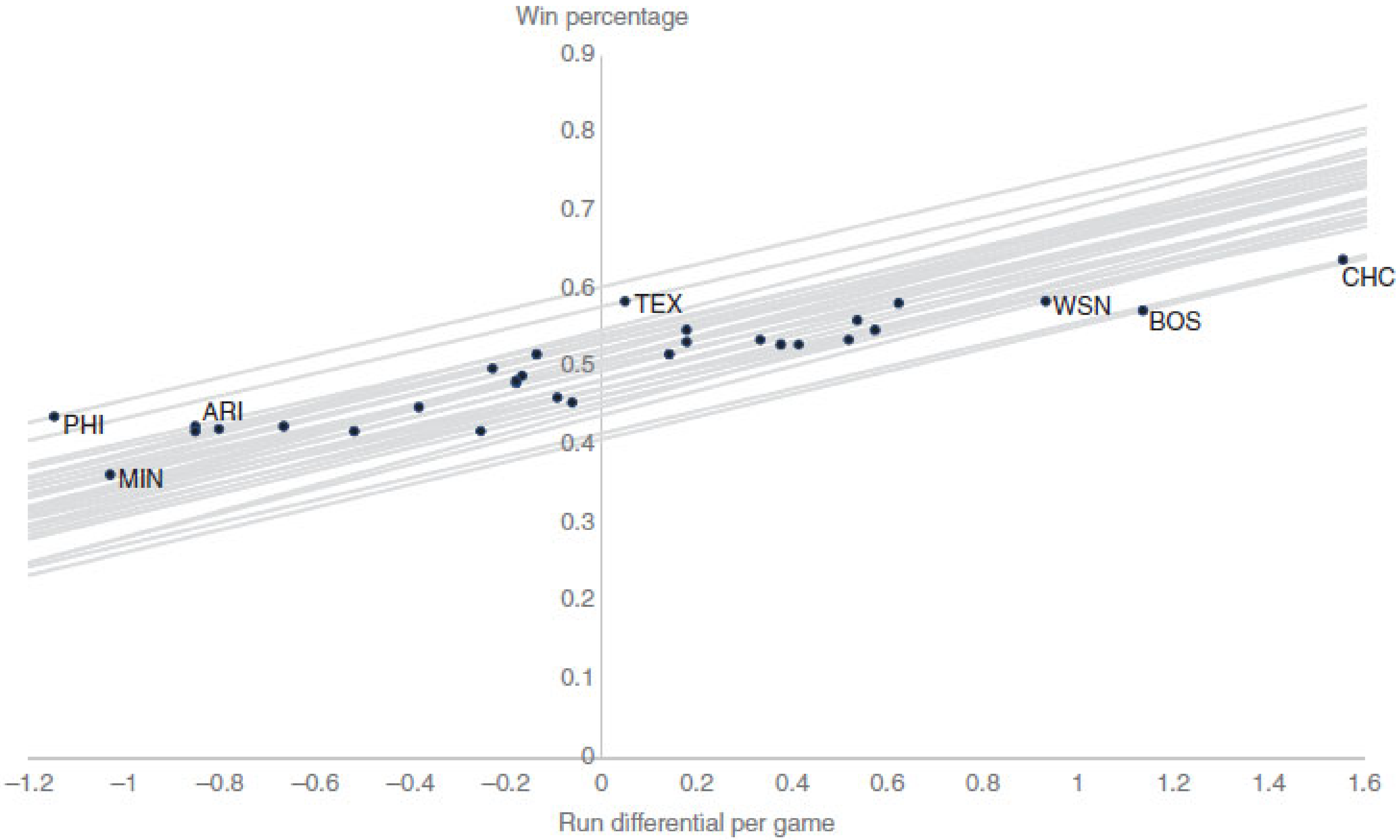


Figure 1: Exact win percentage for the 2016 major league baseball season.

An Exact Win Probability Model!

- Note in figure: one line per team, and observed win percentage for each team is single point on each team-specific line

The intercept a_i and slope b_i for the ith team are given by

$$a_i = \frac{mod_i}{mod_i + mov_i}$$

and

$$b_i = \frac{1}{mod_i + mov_i}.$$

Back To Pythagoras

- Tempting to compare exact model to Pythagoras Taylor expansion:

$$\frac{1}{2} + \frac{\gamma}{4 \times R_{\text{total}}} \times (RS_i - RA_i)$$
$$\approx \frac{mod_i}{mod_i + mov_i} + \frac{1}{mod_i + mov_i} \times (RS_i - RA_i)$$

which in turn suggests that

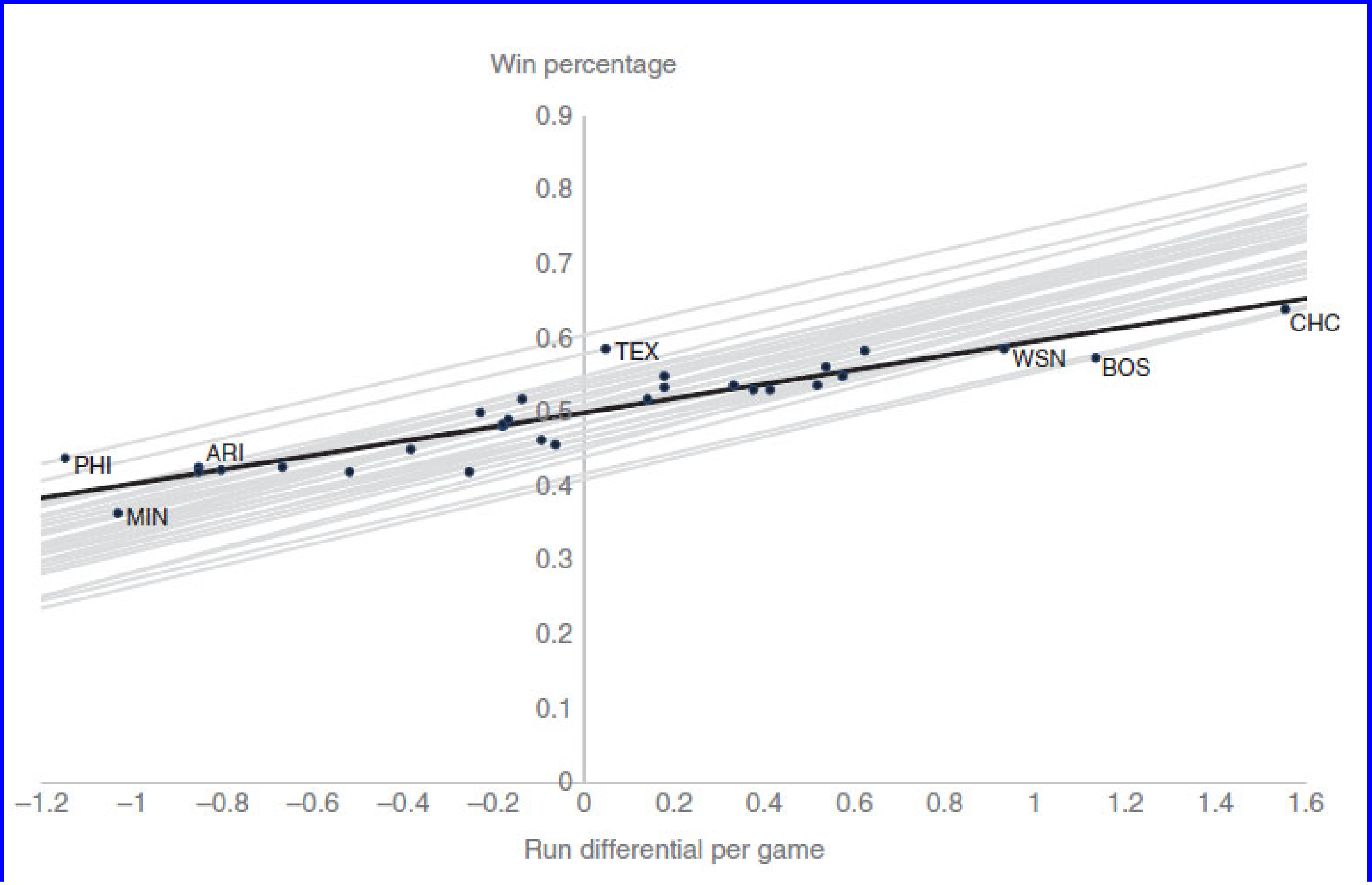
$$\frac{mod_i}{mod_i + mov_i} \approx \frac{1}{2}$$

and

$$\frac{\gamma}{4 \times R_{\text{total}}} \approx \frac{1}{mod_i + mov_i}$$

But That's Not Right!

- Exact model is correct on a team-by-team basis
- Pythagoras model is cross sectional



Decomposing Pythagoras

To simplify notation, let: $x_i = RS_i - RA_i$ = observed average run differential per game over the course of a season for the ith team; $y_i = WP_i$ = seasonal win percentage for the ith team; and recall the definitions of a_i and b_i

$$a_i = \frac{mod_i}{mod_i + mov_i}$$

$$b_i = \frac{1}{mod_i + mov_i}.$$

Decomposing Pythagoras

Now, as is well known, the estimated regression slope $\widehat{\beta}$ in the model $E(Y) = \alpha + \beta X$ is given by

$$\widehat{\beta} = \frac{s_{xy}}{s_x^2}$$

where

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

However, owing to the exact model, for any team i we have

$$y_i = a_i + b_i x_i \text{ for } i = 1, 2, \ldots, n$$

Decomposing Pythagoras

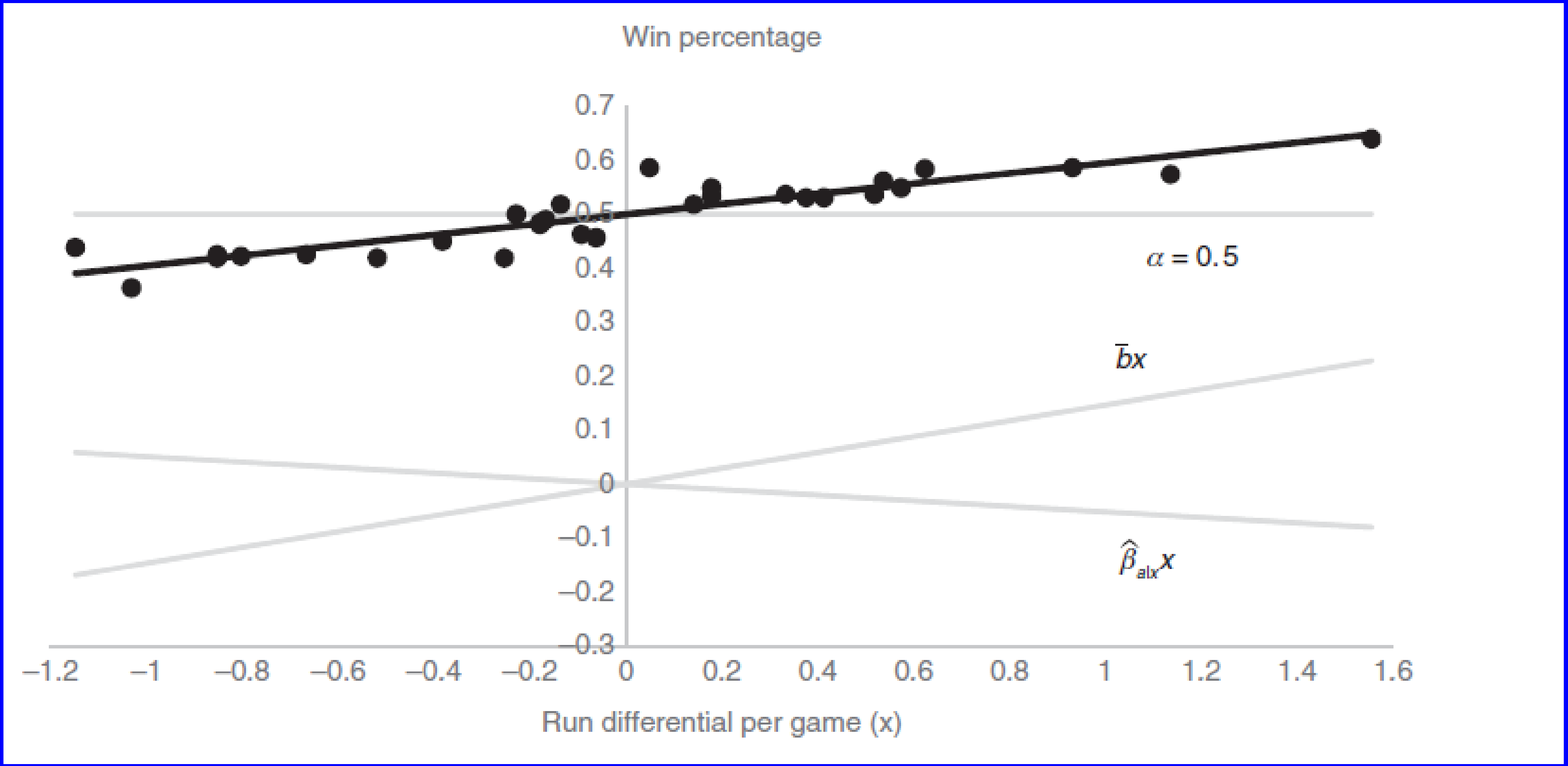
- Substituting the exact model we obtain the cross-team regression slope estimate as

$$
\begin{aligned}
\widehat{\beta} &= \frac{s_{x,a+bx}}{s_x^2} \\
&= \frac{s_{ax} + \overline{b} s_x^2 + s_{b,x^2}}{s_x^2} \\
&= \overline{b} + \hat{\beta}_{a|x} + \hat{\beta}_{b|x^2} \frac{s_{x^2}^2}{s_x^2}
\end{aligned}
$$

where

$$\overline{b} = \frac{1}{n} \sum_{i=1}^{n} b_i$$

Decomposing Pythagoras



- Adding the three gray lines together yields the solid black line

Decomposing Pythagoras

- Finally we can approximate γ as

$$\hat{\gamma} \approx 4 \times R_{\text{total}} \times \left(\overline{b} + \hat{\beta}_{a|x} + \hat{\beta}_{b|x^2} \frac{s^2_{x^2}}{s^2_x} \right)$$

$$a_i = \frac{mod_i}{mod_i + mov_i}$$

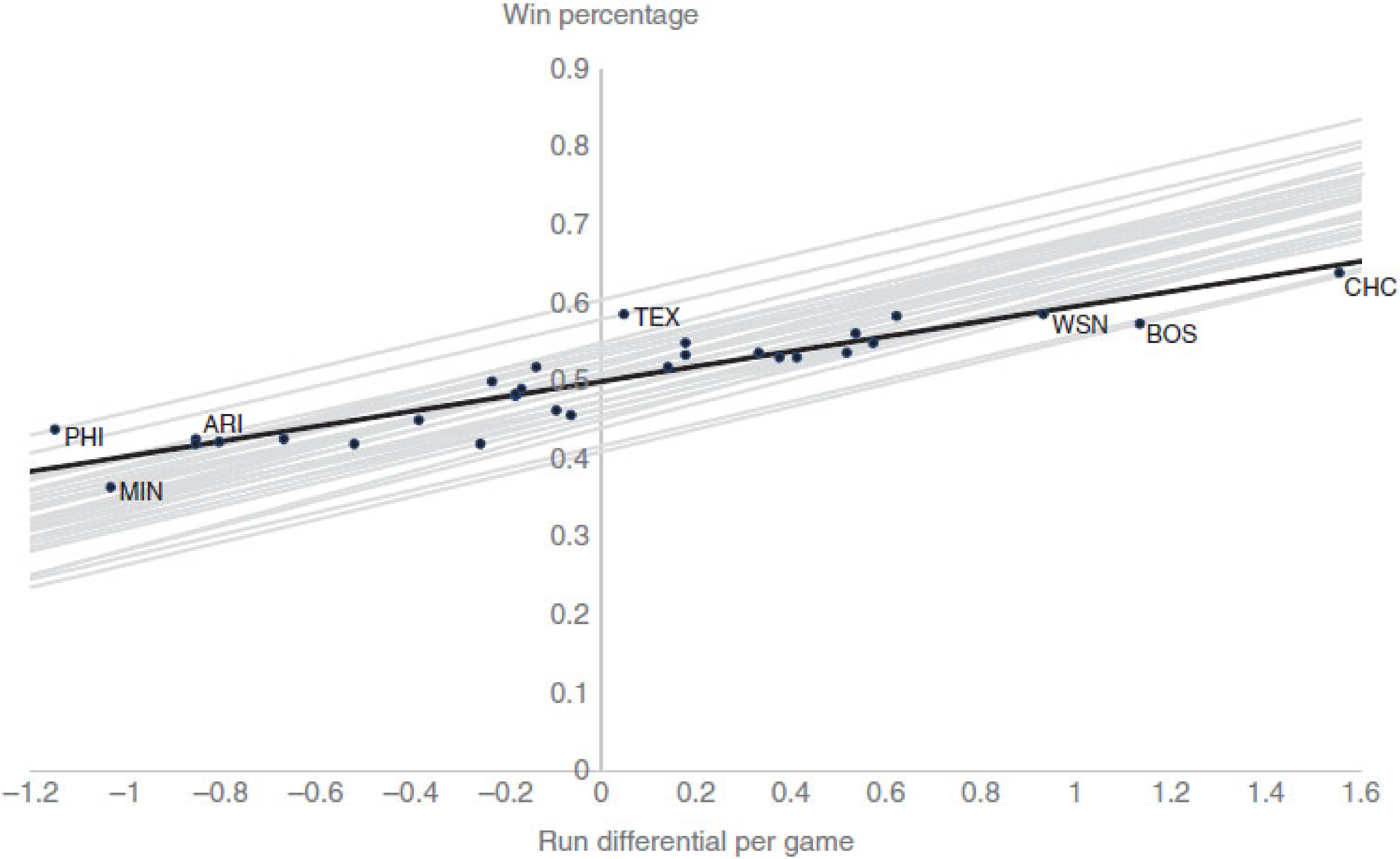
$$b_i = \frac{1}{mod_i + mov_i}.$$

Decomposing Pythagoras

Other things being equal, we see that not only does $\widehat{\gamma}$ increase with the average number of points scored per game (R_{total}), it also increases with $\overline{b}$, which itself declines with scoring margin. The term $\widehat{\beta}_{a|x}$ is the rate with which the ratio of mod_i to $mod_i + mov_i$ changes with score differential x_i across teams

teams with higher average net scores (higher values of $x_i = RS_i - RA_i$) have *lower* values of $mod_i/(mod_i + mov_i)$. This implies that $\widehat{\beta}_{a|x} < 0$. Simply stated, better teams have higher average net scores, larger margins of victory (so when they win they win by more), and smaller margins of defeat (so when they lose they lose by less). Finally,

the term $\widehat{\beta}_{b|x^2}$ essentially equals zero.



Win percentage
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
Run differential per game
PHI
MIN
ARI
TEX
WSN
BOS
CHC

Example: NHL Hockey

$$\hat{\gamma} \approx 4 \times R_{\text{total}} \times \left(\bar{b} + \hat{\beta}_{a|x} + \hat{\beta}_{b|x^2} \frac{s^2_{x^2}}{s^2_x} \right)$$

Table 4: Pythagorean Decomposition Results for NHL Hockey.

Year	$\hat{\beta}$	se	$\bar{b}$	se	$\hat{\beta}_{a\|x}$	se	$\hat{\beta}_{b\|x^2}\frac{s^2_{x^2}}{s^2_x}$	se	R_{total}	$\hat{\gamma}$	se
2007	0.1755	0.0141	0.2449	0.0037	−0.0689	0.0141	−0.0005	0.0054	2.78	1.95	0.16
2008	0.1830	0.0123	0.2497	0.0026	−0.0701	0.0123	0.0034	0.0033	2.91	2.13	0.14
2009	0.1852	0.0124	0.2497	0.0028	−0.0642	0.0128	−0.0004	0.0038	2.84	2.10	0.14
2010	0.1727	0.0135	0.2448	0.0024	−0.0673	0.0137	−0.0048	0.0026	2.79	1.93	0.15
2011	0.1697	0.0146	0.2501	0.0030	−0.0758	0.0146	−0.0046	0.0030	2.73	1.86	0.16
2012	0.1783	0.0133	0.2546	0.0035	−0.0724	0.0139	−0.0039	0.0059	2.73	1.95	0.15
2013	0.1865	0.0095	0.2562	0.0036	−0.0682	0.0094	−0.0015	0.0046	2.74	2.04	0.10
2014	0.1857	0.0116	0.2536	0.0035	−0.0627	0.0122	−0.0052	0.0058	2.73	2.03	0.13
2015	0.2092	0.0114	0.2481	0.0029	−0.0369	0.0116	−0.0020	0.0026	2.71	2.27	0.12
2016	0.1810	0.0075	0.2502	0.0033	−0.0707	0.0078	0.0016	0.0049	2.77	2.00	0.08

Decomposing Pythagoras

- Recall

$$\widehat{\gamma} \approx 4 \times R_{\text{total}} \times \left(\overline{b} + \hat{\beta}_{a|x} + \hat{\beta}_{b|x^2} \frac{s_{x^2}^2}{s_x^2} \right)$$

- Empirically, for baseball and basketball over many seasons

$\hat{\beta}_{a|x}$ is about 30% of $\overline{b}$, which suggests a further approximation

$$\widehat{\gamma} \approx 4 \times R_{\text{total}} \times 0.7 \times \overline{b}.$$

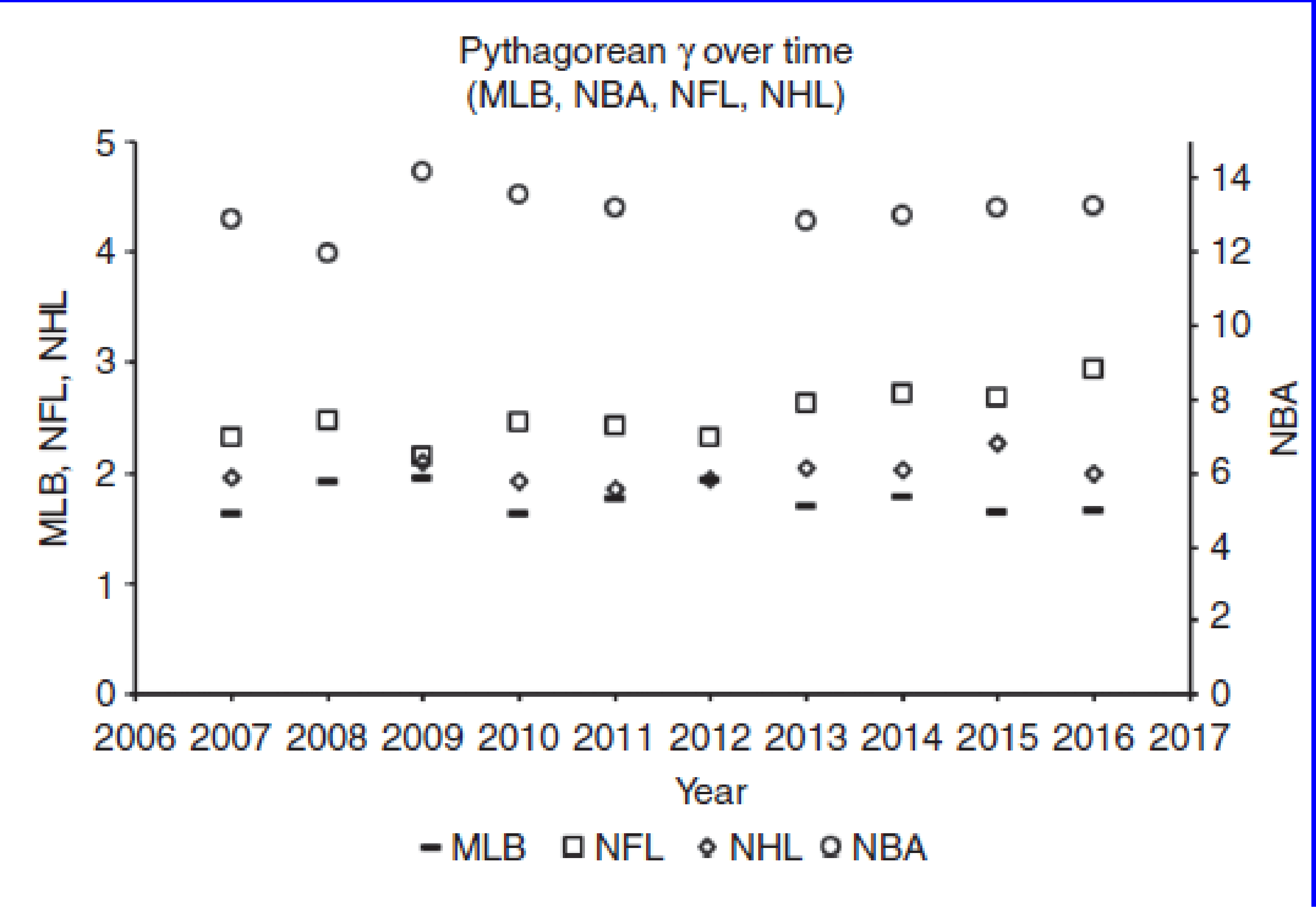
We Have A Story!

This means that the Pythagorean gammas should roughly be in proportion to the ratio of scoring to scoring margin for baseball and basketball.

- Note: baseball is a game with low scores, and low margins of victory/defeat (scores of 4-2, or 5-4 are common)
- Note: basketball is a game with high scores, but very low margins of victory/defeat (scores of 100-95 are common; scores of 100-50 are not)

Football? Hockey?

- Decomposition also works; resulting Pythagorean coefficients show no trend over time



To Sum Up…

- Relationship between scoring and winning can be captured in exact win probability model on a team by team basis
- Pythagoras model yields cross-sectional summary of this relationship across teams
- Decomposing Pythagoras reveals importance of both scale of scoring (average points) and margins of victory/defeat
- Ratios of Pythagorean coefficients across sports interpretable as ratios of scoring to scoring margin

One More Thing…

- Bill James deduced the Pythagorean model based solely on observing patterns in data
- Daryl Morey was first to apply it to basketball
- When James heard about Morey's result, he said "I would never have guessed that you could adapt the Pythagorean to basketball. Basketball has very small margins, relative to the score."
- It appears James really got it!