IEOR E4650 Business Analytics

Session 14: Sports Analytics

Spring 2018

Prof. Adam Elmachtoub

Sports Analytics Opportunities

Professional sports teams now use analytics to drive decisions about nearly every aspect of the game

- Performance evaluation: players and teams
  - Building a team
  - Contract negotiations
  - Fantasy sports and sports betting
  - Injury prevention
  - Identifying at-risk players in pro drafts
- Training and practice
  - Identify strengths and weaknesses
  - Identify best drills and practice techniques
- Strategy
  - Game film analysis
  - Lineups, match-ups, play calling
  - Defensive alignment in baseball; pitch selection

Outline

- Bean bag toss competition
- Identifying skill versus luck
- Shrinkage estimators
- Predicting sports outcomes with optimization

Bean Bag Toss: Rules

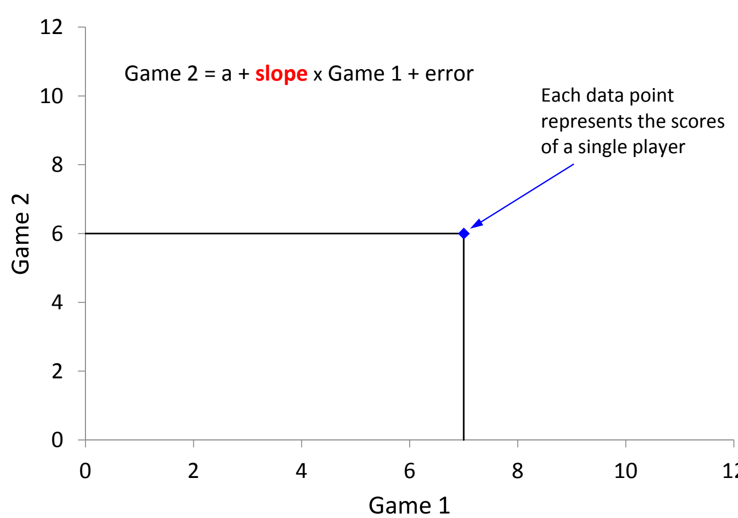
CORNHOLE PLAYERS ASSOCIATION
"FOR THE PLAYERS, BY THE PLAYERS"
WWW.CORNHOLEPLAYERSASSOCIATION.COM

- A game is 4 tosses
- Scoring of tosses
  - Miss board: score = 0
  - Land on board: score = 1
  - Through the hole: score = 3
  - Mercy Rule: You get 1 point free if you did not make any good tosses
- **Winner**: team with the highest score (average score per player) after 2 games

Bean Bag Toss: Instructions

- Teams
  - There are four teams: Target, Nomis, Pandora and Tahoe
  - Teams are formed based on alphabetical order (last name): see board
- We will proceed in two rounds
  - round 1
  - round 2
- During each round, organize tosses as follows:
  - Each team member makes four consecutive tosses, alternating with opposite team
  - Captain (first name in team) records scores on sheet
  - Turn in sheet to IT Czar

Predicting Bean Bag Toss Performance



Class poll:

The "game 2 versus game 1" regression line has a slope

(A) ≥ 1.4

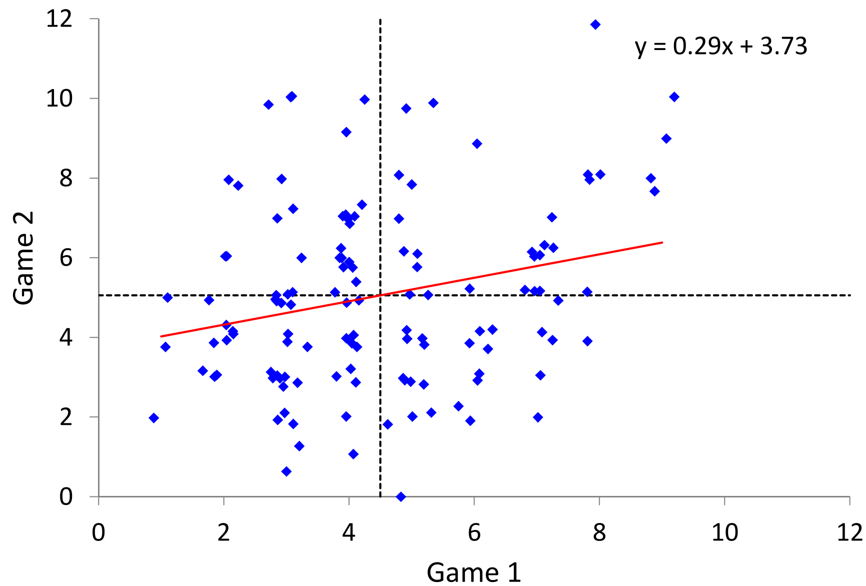
(B) 1.2

(C) 1

(D) 0.8

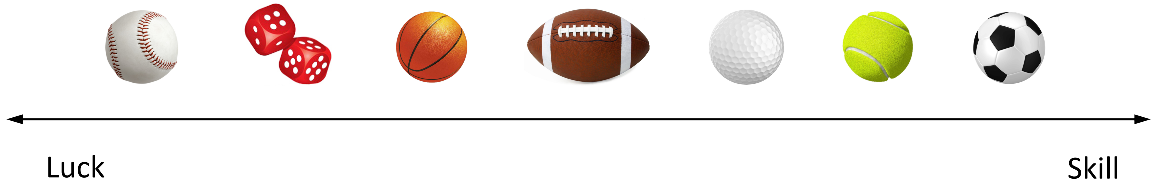
(E) ≤ 0.6

Game 2 Versus Game 1 Performance



Skill Versus Luck

Luck-Skill Continuum



Where would you place these games on the luck-skill continuum?

More skill means more predictability

- Past performance is a good predictor of future performance
- A better player/team is more likely to beat a worse player/team
- Past better-than-average performance predicts future better-than-average performance

Skill Versus Luck Equation

$$\text{performance} \;=\; \text{skill} + \text{luck}$$

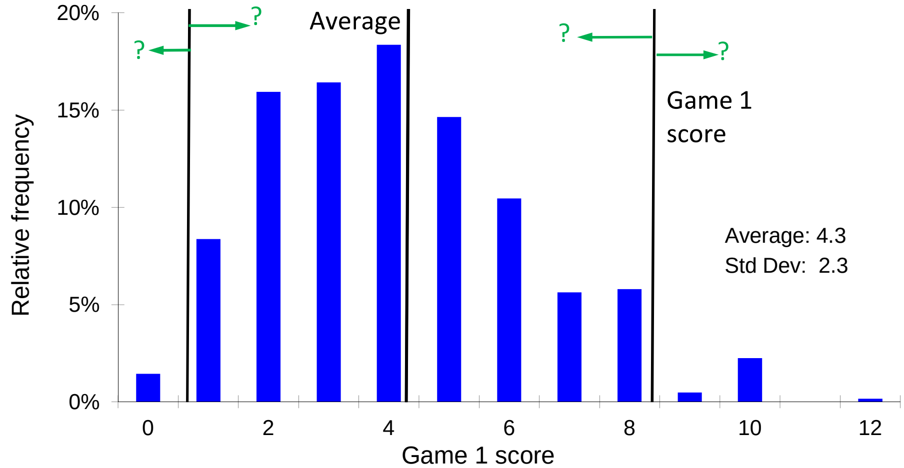
Simple idea: skill and luck contribute to performance (success)

Observations:

- Skill is expected to be stable over appropriate time frames
- Luck is unpredictable

Consequence?

Predicting Game 2 Performance from Game 1 Performance



Mean reversion: why should predictions of game 2 scores shrink toward the mean?

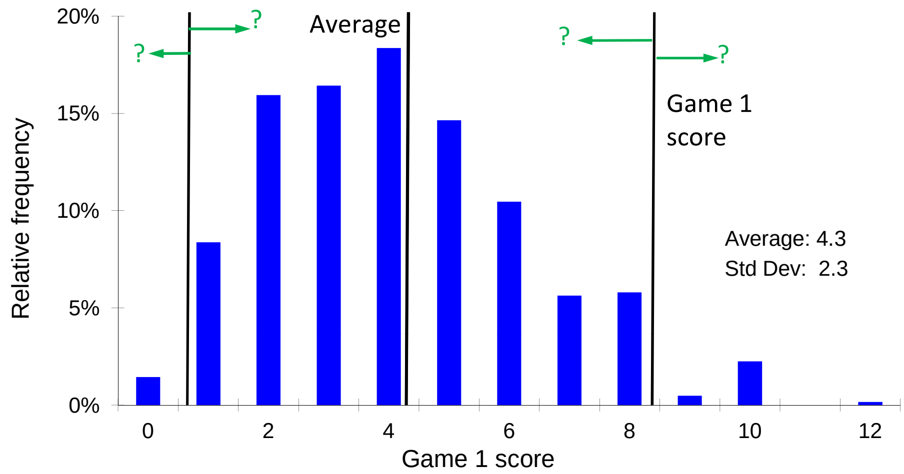
$$\text{performance (score)} = \text{skill} + \text{luck}$$

A game 1 score of 8 could happen with

- (a) Skill $= 10$, luck $= -2$, or (b) skill $= 6$, luck $= +2$
- (b) is more likely than (a) $\implies$ mean reversion!

Shrinkage Estimators

Shrinkage Estimators and Mean Reversion



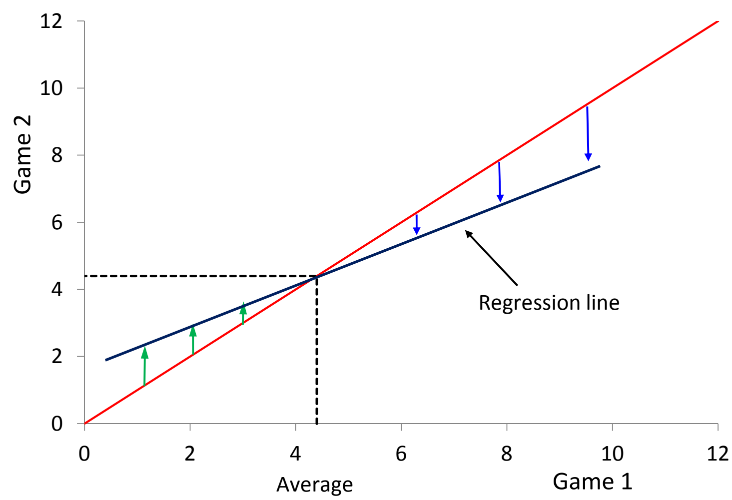
Shrinkage estimator of game 2:

$$\text{Game 2 prediction} = c \times (\text{Game 1 score}) + (1 - c) \times (\text{Game 1 average})$$

Shrinkage coefficient c

- Weight on the past outcome in the prediction
- The prediction shrinks from the past outcome to the population average

Connecting Shrinkage to Mean Reversion



Mean reversion

- Above average game 1 scores tend to decrease; below average tend to increase
- Below average game 1 score tend to increase
- Regression slope will be less than 1
  - Slope close to zero: mostly luck
  - Slope close to one: mostly skill
- Fact: regression slope $\approx$ optimal shrinkage coefficient c^*
  since both measure mean reversion (see the appendix for details)

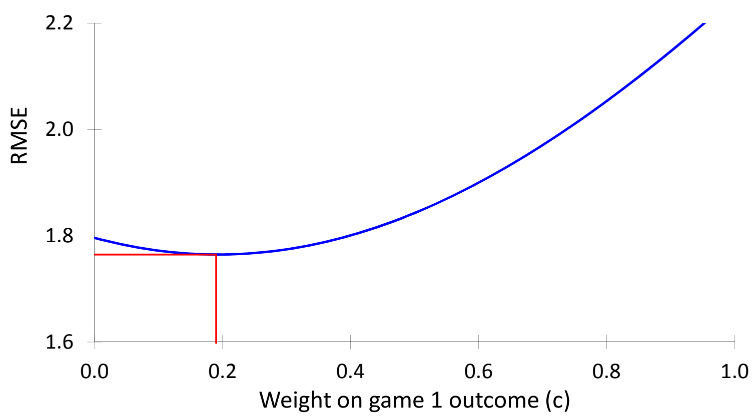
Performance of the Shrinkage Estimator: RMSE

Game 1 avg	c	RMSE
4.0	0.5	2.18

Player	Game 1	Game 2	Shrinkage estimator	Prediction error
1	5	7	4.5	−2.5
2	10	6	7.0	1.0
⋮	⋮	⋮	⋮	⋮
n	1	4	2.5	−1.5

- Game 1 avg: `=AVERAGE(Game 1 column)`
- Shrinkage coefficient (weight on game 1 score): c
- Shrinkage estimator: `=`c`*(Game 1 score) + `$(1-c)$`*(Game 1 avg)`
- Prediction error: `(Shrinkage estimator) - (Game 2 score)`
- Root-mean-squared-error (RMSE):
  `=SQRT(SUMSQ(Pred error column) / COUNT(Pred error column))`

Optimal Shrinkage Coefficient c Minimizes RMSE



$$\text{Game 2 prediction} = c \times (\text{Game 1 score}) + (1-c) \times (\text{Game 1 average})$$

Interpreting the optimal shrinkage coefficient c^*

- c^* close to one
  - Mostly skill: game 1 outcomes are good predictors of game 2 outcomes
  - Little mean reversion of scores
- c^* close to zero
  - Mostly luck: game 1 outcomes are poor predictors of game 2 outcomes
  - Significant mean reversion of scores

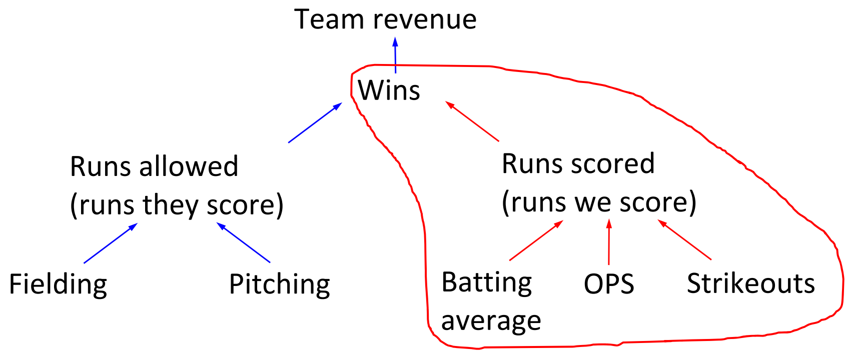
Baseball Analytics: From Shrinkage Estimators to Moneyball

Sports Analytics Opportunities

Professional sports teams now use analytics to drive decisions about nearly every aspect of the game

Value of a Baseball Player

Miguel Cabrera

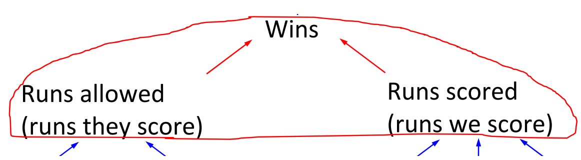


Value of a player

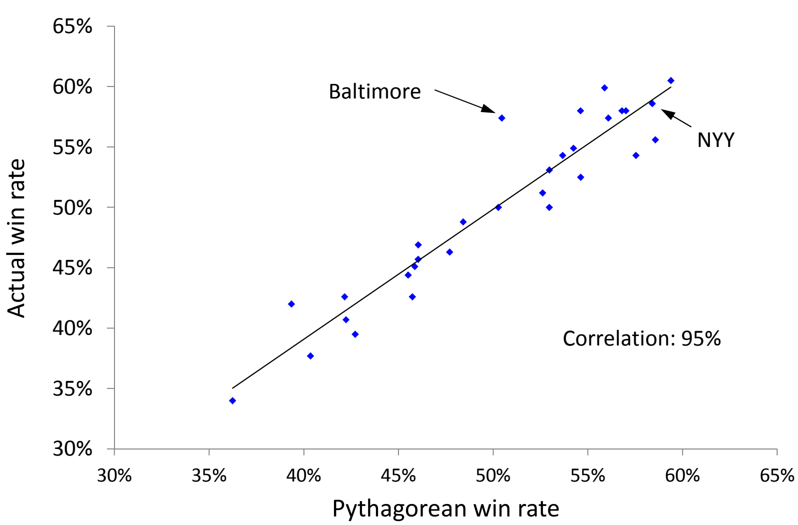
- Better hitters help teams score more runs
- Teams that score more runs win more games

Connecting Runs to Wins: Pythagorean Win Rate

$$\text{Pythagorean win rate} = \frac{\text{Runs scored}^{1.83}}{(\text{Runs scored}^{1.83} + \text{Runs allowed}^{1.83})}$$



Bill James' original formula had an exponent of 2. He later refined it to 1.83.

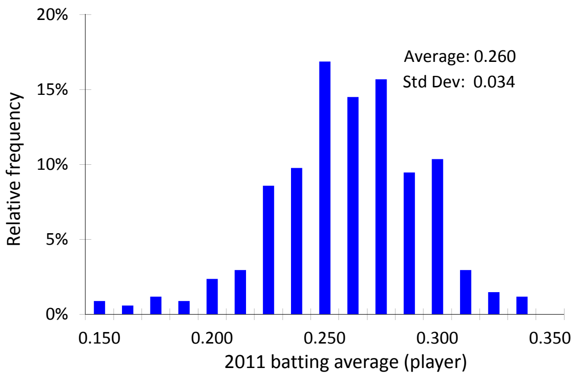


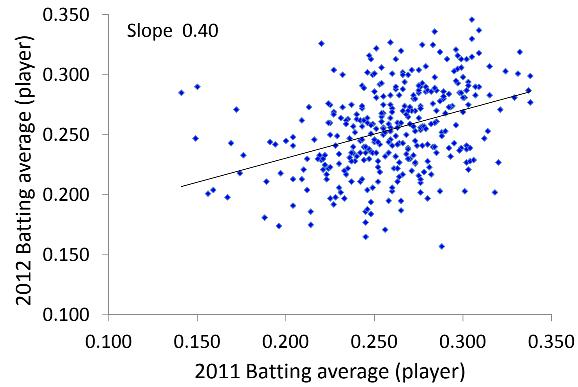
Example: 2012 NY Yankees

RS = 804, RA = 668,
Pythagorean win rate = 0.584,
actual win rate = 0.586

Pythagorean wins:
$162 \times 0.584 = 94.6$ wins;
actual wins: 95

Predicting Batting Average: Using 2011 to Predict 2012

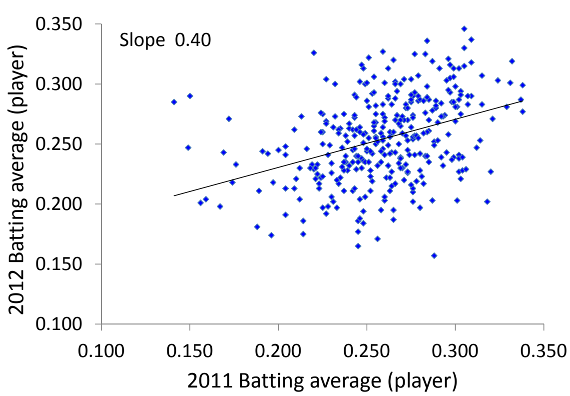


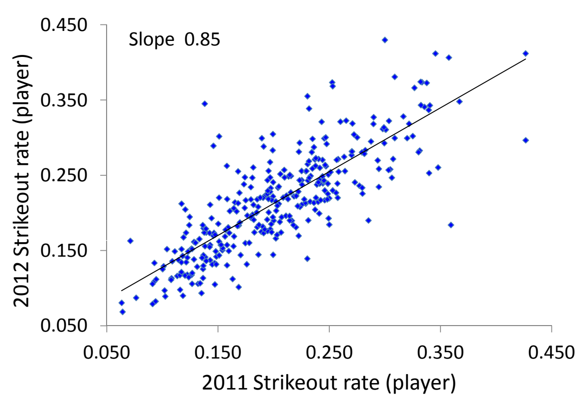


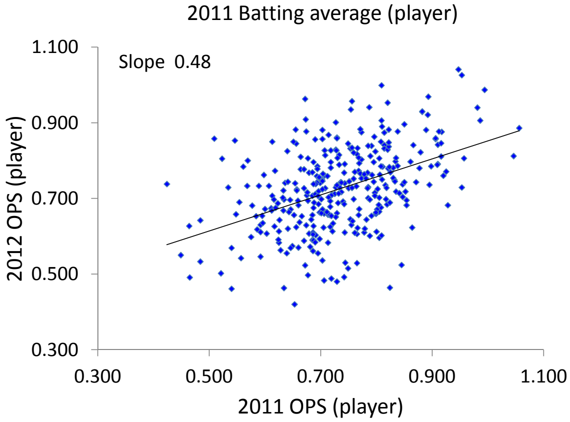
$$\text{Predicted 2012 BA} = c \times (\text{Observed 2011 BA}) + (1-c) \times (\text{2011 Average BA})$$

$c^* =$ 0.4 minimizes RMSE prediction error across players

Skill: Year-to-Year Persistence





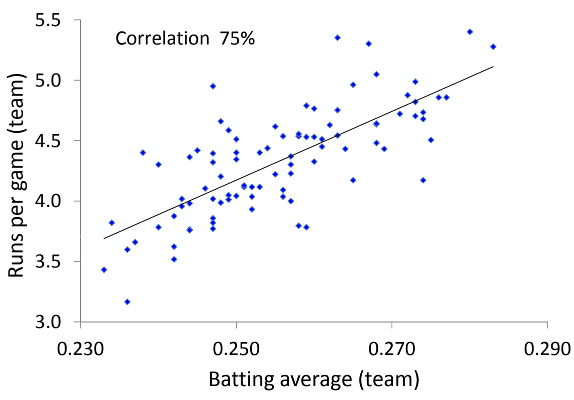


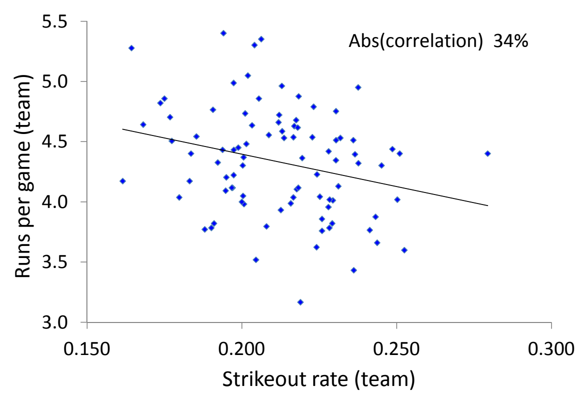
Which stat indicates the most persistent?

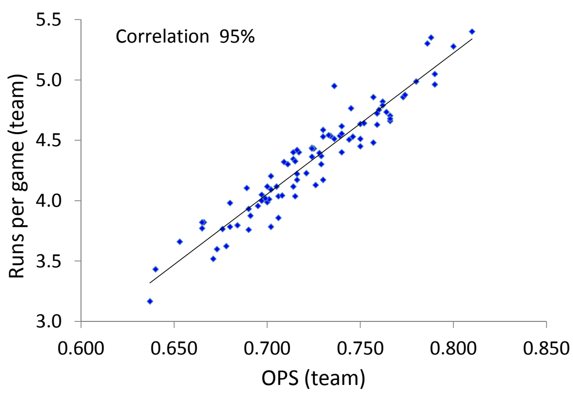
- BA: batting average (hits / at-bats)
- OPS: on-base plus slugging
    - On-base: (hits + walks) / (plate appear)
    - Slugging: total bases / at-bats
- SO: strikeout rate (strikeouts / AB)

Performance Measures: What is a Good Stat?

Predictive: Correlated with Runs



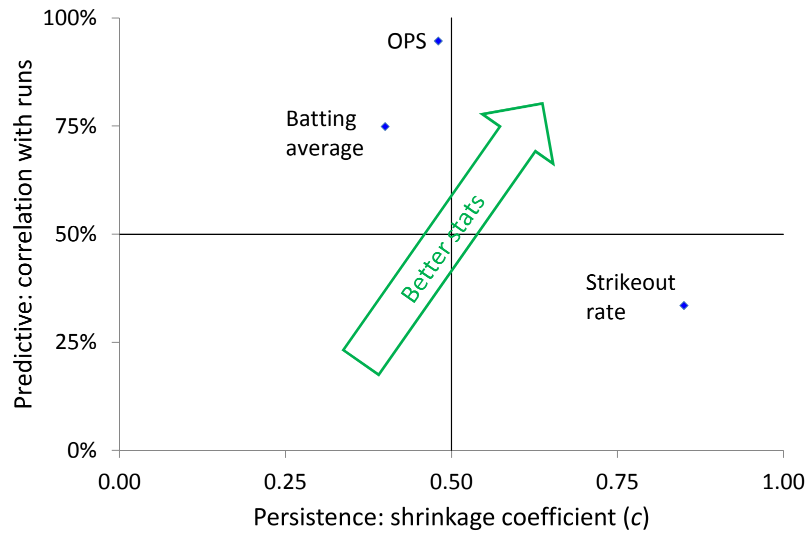




Hitting $\Longrightarrow$ Runs $\Longrightarrow$ Wins

Which is a better stat: BA, OPS, or SO?

Best Stats: Persistent and Predictive



Moneyball, p.128: OPS "was a much better indicator than any other offensive statistic of the number of runs a team would score ... The one attribute most critical to the success of a baseball team was an attribute they could afford to buy."

See: "The Sabermetric Revolution" (posted)

Moneyball: Bill James, Billy Beane and Brad Pitt

Bill James

- Father of modern baseball analytics (Sabermetrics)
- With Red Sox since 2003: Boston won World Series in 2004, 2007 and 2013
- 60 Minutes video: `http://cbsn.ws/wGuOBb`

Billy Beane

- General manager, Oakland Athletics
- 2002, Oakland payroll: $41M; Texas payroll: $107M
- Oakland: 103 wins (64%); Texas: 72 wins (44%)
- Billy Beane interview: `http://bit.ly/1biBahq`

Brad Pitt

- Played Billy Beane in the movie *Moneyball*
- Moneyball video: `http://bit.ly/y1dQ13`

Other Applications of Shrinkage Estimators:

Predicting Future Stock β

Predicting Winning Margins in Football Games

Week	Day	Date		Winner/tie		Loser/tie	PtsW	PtsL
11	Thu	November 14	boxscore	**Indianapolis Colts**	@	Tennessee Titans	**30**	27
11	Sun	November 17	boxscore	**Seattle Seahawks**		Minnesota Vikings	**41**	20
11	Sun	November 17	boxscore	**Denver Broncos**		Kansas City Chiefs	**27**	17
11	Sun	November 17	boxscore	**Tampa Bay Buccaneers**		Atlanta Falcons	**41**	28
11	Sun	November 17	boxscore	**Oakland Raiders**	@	Houston Texans	**28**	23
11	Sun	November 17	boxscore	**New Orleans Saints**		San Francisco 49ers	**23**	20
11	Sun	November 17	boxscore	**New York Giants**		Green Bay Packers	**27**	13
11	Sun	November 17	boxscore	**Pittsburgh Steelers**		Detroit Lions	**37**	27
11	Sun	November 17	boxscore	**Arizona Cardinals**	@	Jacksonville Jaguars	**27**	14
11	Sun	November 17	boxscore	**Miami Dolphins**		San Diego Chargers	**20**	16
11	Sun	November 17	boxscore	**Chicago Bears**		Baltimore Ravens	**23**	20
11	Sun	November 17	boxscore	**Cincinnati Bengals**		Cleveland Browns	**41**	20
11	Sun	November 17	boxscore	**Buffalo Bills**		New York Jets	**37**	14

Goal

- Predict margin of victory in football games
- Objective: minimize RMSE of prediction error

Data

- Winning margin in past games
  - Game 1: Indianapolis beat Tennessee by 3 points
  - Game 2: Seattle beat Minnesota by 21 points
  - Source: Pro-football-reference.com (http://bit.ly/17pgCWz)

Prediction Model

Game	Home team number	Away team number	Home team score	Away team score	Margin (Home − Away)	Prediction	Error (Pred − actual)
						RMSE	10.2
1	1	3	17	7	10	4	−6
2	2	6	10	24	−14	−12	2
3	4	5	12	10	2	3	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

$$\text{MARGIN} = \text{HOME\_TEAM\_RATING} - \text{AWAY\_TEAM\_RATING} + \text{error}$$

Main idea: assign each team a rating

- Teams with higher ratings score more points
- Predicted margin: difference in the ratings of the two teams
- Use optimization to find ratings that minimize prediction error
- Decision variables: team ratings
- Objective: minimize RMSE prediction error

Team number	Team rating
1	10
2	−4
3	6
4	−5
5	−8
6	8
⋮	⋮

Excel Solver and 2014 Results

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	nfl_prediction.xlsm													
2														
3				Number of games in training set	256			RMSE	Average					
4				Number of games in test set	11			13.3	72%					
5				Total number of games	267								Sum	0.00
6	Training set													
7		home	away	home	away				correct				Team	
8	game #	team	team	score	score	margin	predict	error	pred?			Team name	number	Rating
9	1	21	20	36	16	20	1.2	-18.8	1			Dallas Cowboys	1	5.4
10	2	8	22	37	34	3	-1.0	-4.0	0			New England Patriots	2	10.9
11	3	24	18	10	26	-16	17.5	33.5	0			Denver Broncos	3	9.6
12	4	19	30	30	27	3	6.1	3.1	1			San Francisco 49ers	4	-1.0
13	5	13	32	19	14	5	4.0	-1.0	1			Arizona Cardinals	5	2.0
14	6	25	2	33	20	13	-8.4	-21.4	0			Washington Redskins	6	-8.7
15	7	26	12	6	34	-28	0.8	28.8	0			Chicago Bears	7	-6.7
16	8	3	23	31	24	7	5.1	-1.9	1			Atlanta Falcons	8	-3.8
17	9	11	27	14	20	-6	-6.7	-0.7	1			Houston Texans	9	1.7
18	10	14	28	34	17	17	14.4	-2.6	1			Detroit Lions	10	2.1

Prediction results for 2014 (see `nfl_prediction_2014.xlsm`)

- Decision variables: ratings in N7:N38 (one for each of 32 teams)
- Constraint: N3 (sum of ratings) equal to zero, i.e., an "average" team will have a zero rating
- Objective: minimize RMSE in cell H3
- Result: RMSE of 13.3 (average prediction error of 13.3 points)
- Win-loss predicted correctly in 72% of games

2014 Results Through Week 17 (before playoffs)

Team name	Rating
New England Patriots	10.9
Denver Broncos	9.6
Seattle Seahawks	9.5
Green Bay Packers	8.3
Kansas City Chiefs	5.6
Dallas Cowboys	5.4
Buffalo Bills	4.9
Baltimore Ravens	4.6
Indianapolis Colts	4.4
Philadelphia Eagles	3.9
Miami Dolphins	2.6
Pittsburgh Steelers	2.2
Detroit Lions	2.1
Arizona Cardinals	2.0
San Diego Chargers	1.9
Houston Texans	1.7

Team name	Rating
Cincinnati Bengals	0.7
St. Louis Rams	-0.8
San Francisco 49ers	-1.0
Minnesota Vikings	-1.7
New York Giants	-1.7
New Orleans Saints	-2.9
Carolina Panthers	-3.1
Atlanta Falcons	-3.8
Cleveland Browns	-3.9
New York Jets	-5.0
Chicago Bears	-6.7
Washington Redskins	-8.7
Oakland Raiders	-9.0
Tampa Bay Buccaneers	-9.8
Jacksonville Jaguars	-10.5
Tennessee Titans	-11.8

(see `nfl_prediction_2014.xlsm`)

2014 Playoff Predictions

RMSE	Average
12.8	82%

Test set

game #	home team	away team	home score	away score	margin	predict	error	correct pred?
257	Pittsburgh Steelers	Baltimore Ravens	17	30	-13	-2.4	10.6	1
258	Carolina Panthers	Arizona Cardinals	27	16	11	-5.0	-16.0	0
259	Dallas Cowboys	Detroit Lions	24	20	4	3.3	-0.7	1
260	Indianapolis Colts	Cincinnati Bengals	26	10	16	3.7	-12.3	1
261	New England Patriots	Baltimore Ravens	35	31	4	6.3	2.3	1
262	Seattle Seahawks	Carolina Panthers	31	17	14	12.6	-1.4	1
263	Green Bay Packers	Dallas Cowboys	26	21	5	2.9	-2.1	1
264	Denver Broncos	Indianapolis Colts	13	24	-11	5.1	16.1	0
265	Seattle Seahawks	Green Bay Packers	28	22	6	1.2	-4.8	1
266	New England Patriots	Indianapolis Colts	45	7	38	6.5	-31.5	1
267	New England Patriots	Seattle Seahawks	28	24	4	1.4	-2.6	1

Model Extensions

- More sophisticated models might consider effect of additional variables, e.g.,

$$\begin{aligned}\text{MARGIN} = {} & \text{CONSTANT} \\ & + \text{HOME\_TEAM\_RATING} - \text{AWAY\_TEAM\_RATING} \\ & + \text{error}\end{aligned}$$

  accounts for home field advantage.
- More sophisticated models might consider offense & defense separately
  (offense rating, defense rating)
- General technique can be used in many score-based sports
  (e.g., basketball, soccer, etc.)

Wrap-Up

Challenge in practice: separate skill from luck

- What value will a player bring to a team?
- What bonus to give to an employee given her yearly performance?
- What fund to invest in given the past returns?

Want to reward (or invest) based on true performance (skill)
not random performance (luck)

Shrinkage

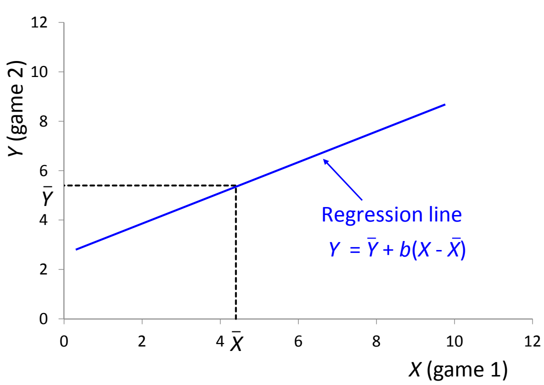
- General principle of regression to the mean
- Almost always happens because of imperfect correlation between performances in different periods
- Key is to understand to what extent is "past performance an indicator of future performance"

Predicting outcomes via optimization

- Come up with simple scoring rules
- Define meaningful constraints and objective

Appendix: Regression Slope, Shrinkage and Correlation

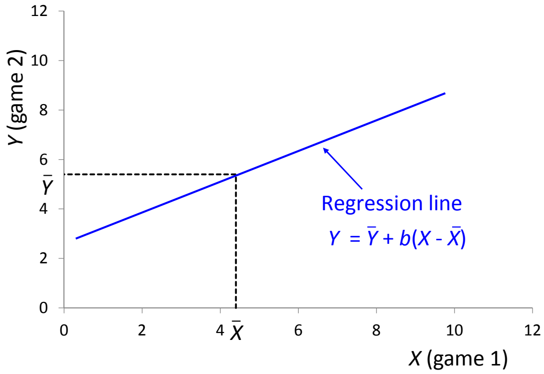
Regression Background



- Regression equation: $Y = a + bX$
- Fact: the regression line passes through $(\bar{X}, \bar{Y})$, i.e., $\bar{Y} = a + b\bar{X}$
- Subtracting and rearranging gives: $Y = \bar{Y} + b(X - \bar{X})$

- Fact: the regression slope is $b = \text{Cov}(X, Y)/\text{Var}(X) = \rho(X, Y)\sigma(Y)/\sigma(X)$
- When there is no change in volatility of performance ($\sigma(Y) = \sigma(X)$): $b = \rho(X, Y)$, i.e., the slope of the regression line is the correlation of performance in the before and after periods

Regression Slope, Shrinkage and Correlation



- Regression equation: $Y = \bar{Y} + b(X - \bar{X})$
- When there is no change in average performance ($\bar{Y} = \bar{X}$), re-arranging gives

$$Y = bX + (1 - b)\bar{X},$$

i.e., the slope of the regression line is the best shrinkage coefficient c^*

- When $\bar{Y} = \bar{X}$ and $\sigma(Y) = \sigma(X)$:

$$b = c^* = \rho$$

i.e., the regression slope (b), the optimal shrinkage coefficient (c^*) and the correlation of before/after performance (ρ) are all identical!