Yale SCHOOL *of* MANAGEMENT

YALE UNIVERSITY
School of Public Health

YALE UNIVERSITY
School of Engineering and
Applied Science

It All Comes Together in Baseball: Linear Weights

Edward H. Kaplan

William N and Marie A Beach Professor of Operations Research
Yale School of Management

Professor of Public Health
Yale School of Public Health

Professor of Engineering
Yale School of Engineering and Applied Science
New Haven, Connecticut

Linear Weights

- "Linear weights" is sabremetrics-speak for regression
- Idea was to come up with an empirical approach to runs created by relating observed runs (at team level) to observable batting/running events (singles, HRs, stolen bases, strikeouts, walks, HBP, etc.)
- Since such batting events are tallied for individual players, use regression coefficients (these are the "linear weights") from team-level regression to model individual runs created

Linear Weights

- So, if y_i is runs scored by team i, and x_{ij} is the frequency of batting/running events of type j by team i, then try estimating

$$y_i = \beta_0 + \sum_j \beta_j x_{ij}$$

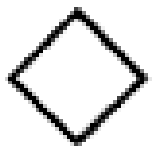
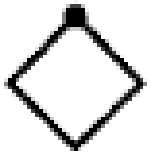
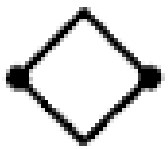
- Easy to do, as discussed in the text

Linear Weights

- But how do we know whether to believe what comes out of this?
- Also, don't we now have a pretty good understanding of how to estimate runs created via the baseball state space?
- WAIT A SECOND -- can we use run value added to figure out what this regression should be doing?
- Can we use run value added to figure out what the β's should represent?
- Let's find out

In A Previous Class…

- Each cell entry below gives expected runs to the end of the inning starting from the specified state

	Bases Situation							
								
0 outs	0.55	0.94	1.17	1.43	1.57	1.85	2.07	2.40
1 out	0.30	0.57	0.70	0.98	0.97	1.22	1.44	1.65
2 outs	0.11	0.24	0.35	0.37	0.49	0.50	0.61	0.81

Recall Run Value Added

- Recall from our state space discussions that at the start of any inning (no outs, nobody on base), the expected number of runs to the end of the inning equals $v_0 \approx 0.55$
- As each baseball play takes place, we update the run-value-added Δv_i for the i^{th} play in sequence according to

$$\Delta v_i = v_i + r_i - v_{i-1}$$

- In above, v_i is the expected number of runs to the end of the inning corresponding to the base/out state that results from the i^{th} play in sequence; r_i is actual runs scored

Evolution of Run Value Added

- If there are n plays in the inning, then the n^{th} play must have led to the final out and thus $v_n = 0$
- Keeping score for the inning is equivalent to summing the value added terms over all plays

$$\sum_{i=1}^{n} \Delta v_i = \sum_{i=1}^{n} (v_i + r_i - v_{i-1})$$

$$= \sum_{i=1}^{n} v_i + \sum_{i=1}^{n} r_i - \sum_{i=0}^{n-1} v_i$$

$$= \sum_{i=1}^{n} r_i + \sum_{i=1}^{n-1} (v_i - v_i) + v_n - v_0$$

$$= \sum_{i=1}^{n} r_i - v_0$$

Evolution of Run Value Added

- So the main result is

$$\sum_{i=1}^{n} \Delta v_i = \sum_{i=1}^{n} r_i - v_0$$

- The value added score for an inning is the difference between the actual runs scored and the number of runs one expects to score from the start of the inning (i.e. v_0)
- What happens if you aggregate to the level of an entire game?

Run Value Added Over A Game

- Aggregating up to an entire game, the sum of the value added by individual plays for a team equals the total runs scored by that team minus the average runs per game!
- Keeping score by run value added is equivalent to keeping track of *runs above average* (which could be a negative number if the team scores fewer runs than expected)

Run Value Added Over A Season

- The sum of the value added by all baseball plays over an entire season is the difference between the actual and expected runs scored over the whole schedule
- So, if y_i is runs scored by team i over the course of a season, and $\bar{y}$ is the average number of runs per team per year, then for team i we have the sum of the run value added for every play during the season equals $y_i - \bar{y}$

The Value of Baseball Events

- Let *n* now represent the total number of plays over the course of a season
- Each value added term observed corresponding to each play can be thought of as stemming from some *type of event*
- Thus, it is sensible to consider the *conditional expectation of the run value added* for a randomly selected play, *given the type of baseball event that produced it*

The Value of Baseball Events

- Let e denote baseball events (singles, doubles, triples, home runs, strike outs etc.)
- Let f_e be the total frequency with which event e occurs over the course of the season (say there are m different kinds of events) for a given team
- Define $\overline{\Delta v_e}$ as the average run value added over all plays corresponding to a baseball event of type e

$$\sum_{i=1}^{n} \Delta v_i = \sum_{e=1}^{m} \overline{\Delta v}_e f_e$$

Deducing Linear Weights

- Hey – we have two different formulas for the sum of run value added over a season:

$$\sum_{i=1}^{n} \Delta v_i = \sum_i r_i - \bar{y} = y_i - \bar{y}$$

$$\sum_{i=1}^{n} \Delta v_i = \sum_{e=1}^{m} \overline{\Delta v}_e f_e$$

- So it must be that

$$y_i - \bar{y} = \sum_{e=1}^{m} \overline{\Delta v}_e f_e$$

Deducing Linear Weights

- Define $\beta_e = \overline{\Delta v_e}$ as the mean run value added for events of type e
- Define $z_i = y_i - \bar{y}$ = runs over average for team i
- This suggests building the regression model

$$z_i = \sum_{e=1}^{m} \beta_e f_{ei}$$

where the values of z_i and f_{ei} are known

- Note – no constant term!

Let's Try This!

- From baseball-reference.com, pulled together all batting events info and runs scored for all teams from 2007-11 seasons (so 150 observations in total)
- Average number of runs per season per team was equal to 736.28; use this as $\bar{y}$
- Let's see what happens

Direct Comparison To Theory

- In *The Book: Playing The Percentages In Baseball* (Tango, Lichtman and Dolphin, 2007), run value added of individual baseball events based on painstaking play-by-play computation over time is presented

Event	Run values
1B	0.475
2B	0.776
3B	1.07
HR	1.397
SB	0.175
CS	-0.467
BB	0.323
SO	-0.301
HBP	0.352
SH	-0.096
IBB	0.179
Other Outs	-0.299

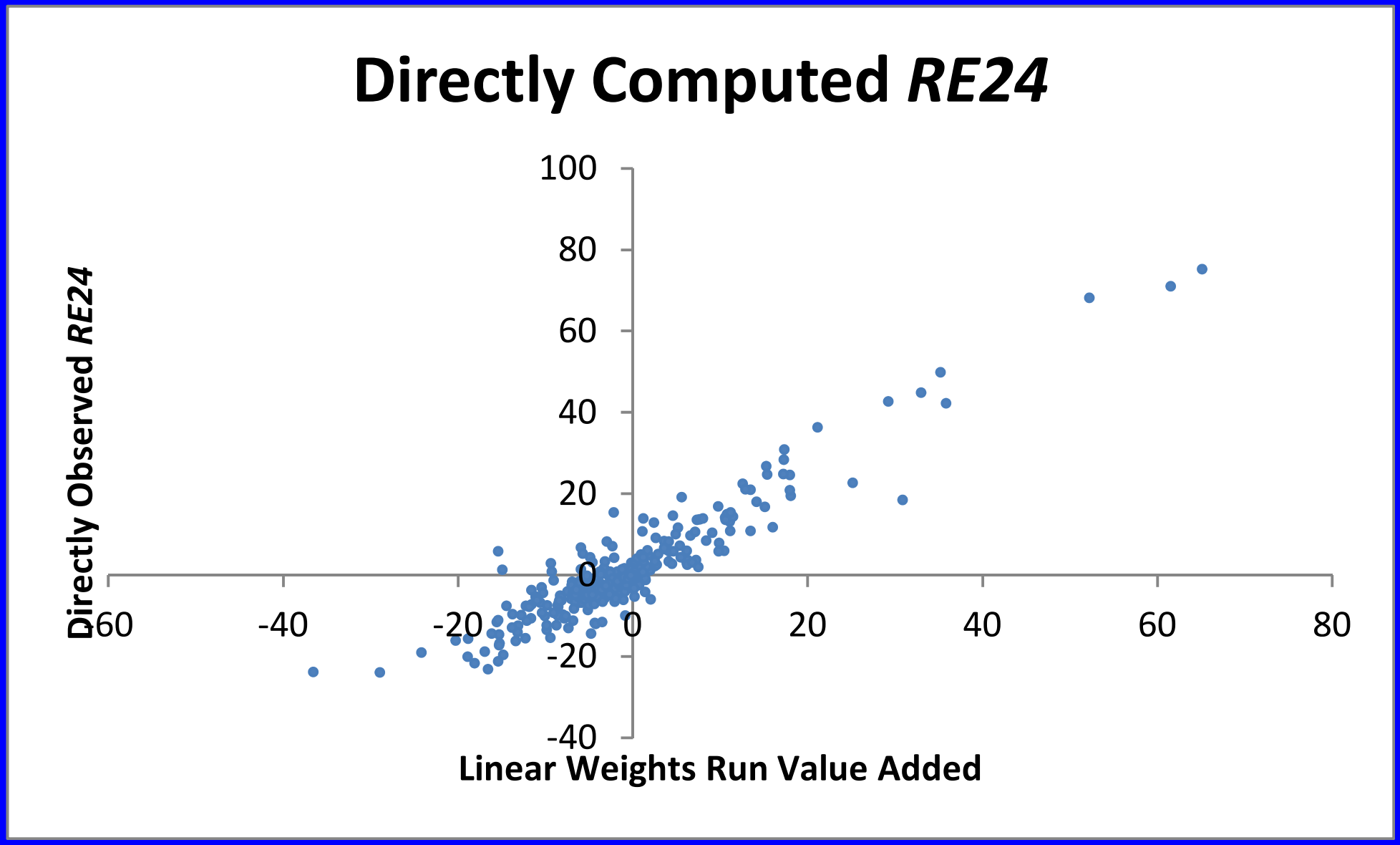
1B	0.489
2B	0.706
3B	1.130
HR	1.344
SB	0.072
CS	-0.156
BB	0.297
SO	-0.278
GDP	-0.700
HBP	0.454
SH	-0.326
SF	0.359
IBB	-0.019
Other Out	-0.250

Regression

But Wait – We Can Directly Observe RE24 For Individual Players!

- For 2013 American League batters
  - Went to baseball-reference.com
  - Recorded all baseball events
  - Recorded RE24
  - Can estimate RE24 using regression estimates
  - Compare directly to individual values
- How do the linear weights perform?
- Let's find out

Individual *RE24* From Linear Weights



Recall *WPA*

- From data, easy to look at individual *WPA* as function of individual *RE24*
- How do you think these will relate?
- Let's take a look…

Predicting *WPA* from *RE24*

- Looks like we need about 11 runs per win
- Not all that different from folklore (10 runs per win)
- Ah – but do you now see a way to evaluate players without having to keep track of all the state space stuff?
- In other words, suppose fangraphs and baseball-reference crashed and all you had were standard baseball event data

It All Comes Together In Baseball

- Using linear weight coefficients, we can translate any player's baseball event frequencies into runs created over average
- Using back-of-envelope result from Pythagoras class, or more recent discovery, can estimate *WPA = RE24 / 10* (OK, 11)
- Thus, can get straight from baseball events to *WPA* via regression without needing all the state space stuff
- But, state space model is why we believe results!