Data 8 Connector: Sports Analytics

...

3/6/18

Course Stuff

- ➔ Data acquisition for projects: **March 16**
- ➔ Homework coming late this week (~3-4 weeks of time)
- ➔ Rough outline going forward:
  - ◆ Adjustments like park factors and positions
  - ◆ Replacement level and WAR
  - ◆ Inference: the hot hand, uncertainty in WAR, regression modeling

Review

Football: QB Passer Rating

A linear performance metric

$$\text{QB Passer Rating} = 100 \times \frac{1}{6} \times (A + B + C + D),$$

$$A = \left(\frac{\text{COMP}}{\text{ATT}} - .3 \right) \times 5, \quad B = \left(\frac{\text{YDS}}{\text{ATT}} - 3 \right) \times .25,$$

$$C = \left(\frac{\text{TD}}{\text{ATT}} \right) \times 20, \qquad D = 2.375 - \frac{\text{INT}}{\text{ATT}} \times 25$$

Football: QB Passer Rating

Passer Rating misses a lot of things about QB play

- ➔ Rushing
- ➔ Sacks
- ➔ Fumbles
- ➔ Crucial situations (3rd down or 4th quarter, if this matters to you)
- ➔ Quality of teammates (WRs and O-Line)
- ➔ Attribution of performance (interceptions and yards)

Football: DVOA Rating

Defense-adjusted Value Over Average

Recall the question

> One running back runs for three yards.
> Another running back runs for three yards.
> Which is the better run?

Conventional NFL stats measure performance with net yards

- You really want to score points by getting in the endzone
- Not all yards are created equal

Football: DVOA Rating

DVOA tries to distribute credit for scoring points and winning games

- Assigns every play a value based on total yards and yards towards a first down

After adjustments, we football version of RE24

Add up every play by a certain team or player, divide by the total of the various baselines for success in all those situations and you get Value Over Average

Adjust plays to account for defense's ability to defend that play

- Defenses are also rated by the quality of the opposing offenses

Basketball: PER

You can work on intuition that the following matters:

- ➔ Efficient shooting from the floor
- ➔ Not turning the ball over
- ➔ Generating turnovers
- ➔ Rebounding
- ➔ Getting and making free throws

Basketball: PER

So efficient shooting isn't the only part of basketball

- ➔ You can potentially win by just having more possessions/shots than the other team

We saw PER as a way to capture this by using only box score stats

- Obviously box score stats give a very incomplete view of the game but we can start with this

Basketball: PER

Contributions to evaluate a player (via box score stats)

- ➔ Efficient shooting (twos and threes)
- ➔ Assists
- ➔ Free throw makes and misses
- ➔ Turnovers and steals
- ➔ Defensive and offensive rebounds
- ➔ Blocked shots
- ➔ Fouls

Basketball: PER

One key feature of PER is it tries to account for expected outcomes

This includes some key modeling values

- ➔ Assist factor: discounts made fields to balance credit given to assisting player
- ➔ Value of Possession: many contributions based around gaining/losing possession
- ➔ Def Reb Pct: league rebounding rate

Take rebounding: most rebounds are made by the defending team

Any particular defensive rebound is mostly expected

Discount the defensive rebounds based on how likely they are

Basketball: PER

PER also tries to address **bias**

Bias: an observed value is pushed higher or lower due to issues in observing the data

Bias in basketball

- Teams play at different paces so numbers can be inflated for high pace teams
- uPER normalized by minutes played but that still doesn't account for variable number of possessions

PER uses pace adjustment to "debias" the results

More Expected Value Modeling

EV Modeling for Shooting

We know Field Goal Pct misses the different values for shots

We know Effective Field Goal Pct captures that

We know True Shooting Pct captures free throws

But intuitively we know a fair amount about the spatial nature of shooting

EV Modeling for Shooting

Layups and shots around the basket are easy

Three pointers are hard

Distance matters

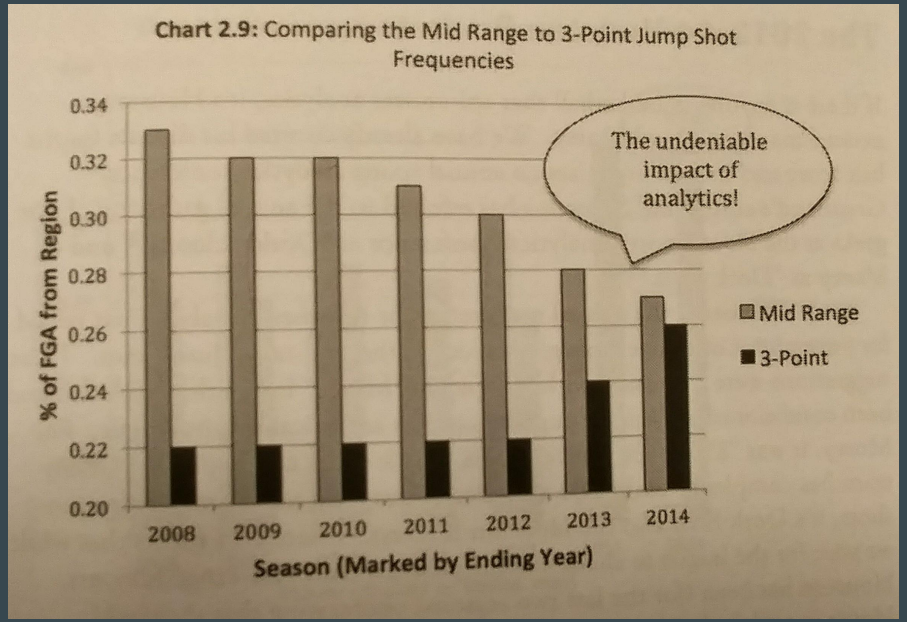
EV Modeling for Shooting

The long two is not a great shot

- 3pt has marginally less % but 1 pt more value

We know distance matters

NBA teams know it too



EV Modeling for Shooting

So we intuitively we know a fair amount about the spatial nature of shooting

Can we get a better sense of expected shooting given spatial information?

...Demo time

Modeling the Game

Modeling the Game

So far we have looked at direct performance measurement

- ➔ Model individual components: run values, eFG, PER
- ➔ Track player stats and build measures
- ➔ Obtain players/build teams around those measures

Modeling the Game

We want to look a bit more broadly at modeling the games (baseball and basketball)

Why?

- Build a better understanding of the game
- Better evaluation of performance
- Transfer analyses to other leagues (Japan, Euroleague, etc)

How? Return to team level performance and try to determine what drives success

Modeling the Game

We'll explore two approaches to modeling baseball and basketball

- ➔ Run Estimation for baseball
- ➔ Dean Oliver's Four Factor Model for basketball

Modeling the Game

Run Estimation

- Using observations (boxscore stats) to estimate team runs
- We actually already have a run estimator: LWTS
- We'll try to improve on it

Dean Oliver's Four Factor Model

- Basketball has a bit more complicated a process: free flowing action
- Will use a common concept, Factor Modeling, to decompose the game into components
- Goal: use the components to explain team success

Run Estimation

Run Estimators

A run estimator is an actual measure of run creation

By definition, it will yield a value that can be used as an estimate of team runs

No need for a regression to get a conversion: Team Runs = $a \cdot$ OPS + b

In other words, it's a model for team runs

Run Estimators: LWTS

We already have a run estimator: LWTS

To get total team runs, add in average runs per game to capture run scoring environment

LWTS produces Pete Palmer's Batting Runs

The varying runs per game is absorbed into value of an out

All other values are fixed for all time

$$BR = .47 \cdot H + .38 \cdot 2B + .55 \cdot 3B + .93 \cdot HR + .33 \cdot (BB + HBP) + .22 \cdot SB - .28 \cdot CS - .1 \cdot (AB - H)$$

Run Estimators: LWTS

We saw LWTS as an estimate of a player's runs produced above/below average

At the team level, we have an estimate of team runs above/below average

- Or total team runs

LWTS is a linear estimator

- Marginal impact of an event is constant

We will return to this question but it's worth asking right now:

Is it okay to use an estimator (like LWTS) for team level production AND player level production?

Run Estimators: LWTS

What happens to LWTS at extreme values?

- ➔ We put together a team (we're terrible compared to MLB)
- ➔ Our stats: no hits, walks, or anything
- ➔ Our predicted total runs (should be 0):

$$5 \text{ Runs per Game} - 0.3 \text{ Runs per Out} \cdot 27 \text{ Outs per Game} = -3.1 \text{ Runs per Game}$$

Why is it negative?

Run Estimators: LWTS

What happens to LWTS at extreme values?

- ➔ Let's play 7 inning softball (super high scoring, ~20 runs per game)
- ➔ Our stats: ~50 batters, 21 out, ~30 get on, and about ⅔ score
- ➔ Is a homerun really worth that much more than a single if you almost always score when getting on base?
- ➔ Now an out is extremely costly (think about a dismissal in Cricket)

Run Estimators: LWTS

LWTS assumes we are operating near league average

- And league average is calibrated to MLB, not our terrible team or our softball league

So we really need to re-estimate if run environment changes

When pitchers are really good (1960s)...

- Outs are not as costly (everyone is getting out anyhow)
- Getting on base not nearly as valuable (you're unlikely to score)
- Power is relatively more valuable (HR value always > 1 while other values will shrink towards 0)

Run Scoring Model

So is the run scoring even linear?

Only approximately when calibrated and used appropriately.

Recall constant marginal impact:
If we have to re-estimate depending on the environment, then it's non-linear

Run Scoring Model

What does it mean to be non-linear?

- ➔ Incremental/marginal value of an event is no longer constant
- ➔ Instead, it changes with itself or according to other values
- ➔ Toy examples of nonlinearity

  Knock-on: singles^2 more singles magnify their own value

  Interaction: $\text{walks} \cdot \text{doubles}$ more walks magnify the effect of doubles

Why non-linearity? The interplay between getting on base and moving runners

Slugging produces more runs when more runners on base

Can also handle the extremes (event values fluctuate via the model instead of re-estimated)

Multiplicative Run Estimators

Instead of a linear estimator, there are multiplicate estimators

- ➔ Bill James' Runs Created
- ➔ David Smyth's BaseRuns

Bill James' Runs Created

Core idea:

$$RC = \frac{\text{On-base Factor} \cdot \text{Advancement Factor}}{\text{Opportunity Factor}}$$

Basic version:

$$RC = \frac{(H + BB) \cdot TB}{AB + BB} = OBP \cdot TB = OBP \cdot SLG \cdot AB$$

"Technical" Version:

$$RC = \frac{(H + BB - CS + HBP - GIDP) \cdot (TB + (.26 \cdot (BB - IBB + HBP)) + (.52 \cdot (SH + SF + SB)))}{AB + BB + HBP + SH + SF}$$

Contribution of OBP and SLG is nonlinear: depends on SLG and OBP, respectively

David Smyth's BaseRuns

Run scoring model:

$$\text{Runs} = \text{Batter reaches base} \times \text{Runner scoring rate} + \text{Batter hits home run}$$

Basic Version of BaseRuns:

$$\text{BaseRuns} = \frac{A \cdot B}{B + C} + D$$

$$A = H + BB - HR$$
$$B = (1.4 \cdot TB - .6 \cdot H - 3 \cdot HR + .1 \cdot BB) \cdot 1.02$$
$$C = AB - H$$
$$D = HR$$

Runs Created vs BaseRuns

Both leverage a non-linear relationship between getting on base and advancing

Both do a good job of predicting team runs (BaseRuns appears to be the best though)

Runs Created is ad-hoc while BaseRuns tries to model run scoring

- Once you get on, you have a success rate of scoring based
- Homeruns automatically score 1 run

A good way to inspect: individual event values

- Take a derivative with respect to an event
- Depends on run environment

Runs Created vs BaseRuns

Recall our reasoning about event values at the extremes

In low scoring environments, values should approach 0 except for the HR

- Getting on base doesn't help: you likely won't score
- Run expectancy overall is so low the change is negligible

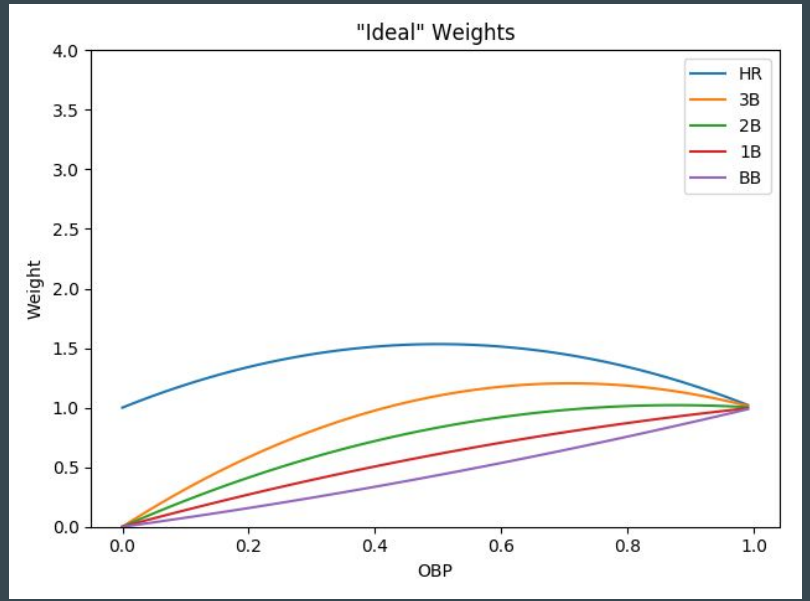
In high scoring environments, all values should converge to 1

- Getting on base is all that matters: you will likely score
- Run Expectancy is so high that all home runs are discounted down to just the 1 run scored

Runs Created vs BaseRuns

Here's a look at how we might think of our "ideal" weights

- ➔ Compare to linear (flat lines)
- ➔ Just conceptual: arbitrary parabolas purely for descriptive purposes
- ➔ Typical MLB team OBP is about .330
- ➔ As runners get on base, event values should rise
- ➔ As runners end up constantly scoring, values revert back to 1.
- ➔ A home run is worth at least 1 run



Runs Created vs BaseRuns

Getting event values from RC and BR

- ➔ Estimate the following values from Lahman Database

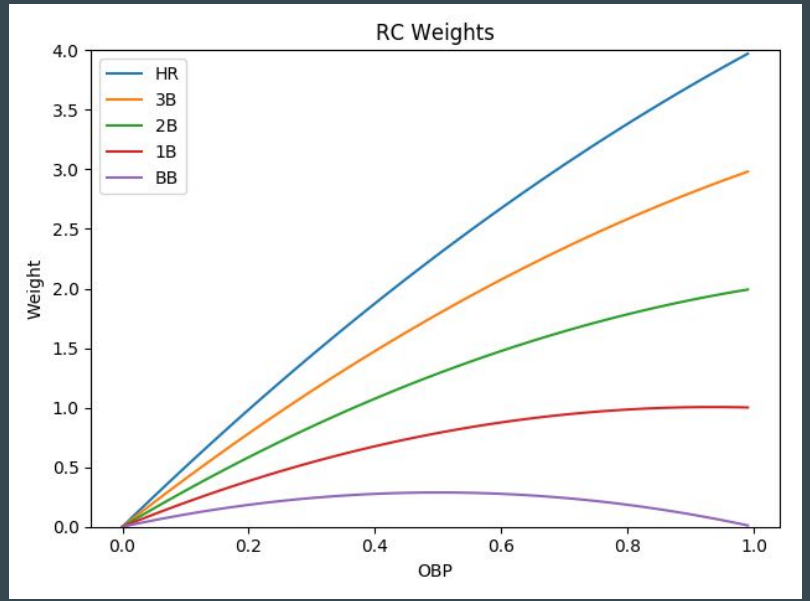
BB/H = 0.367 1B/H = 0.671 2B/H = 0.195 3B/H = 0.021 HR/H = 0.112

- ➔ PA ≈ 162 · 27 / (1 - OBP) (Higher OBP means more PA)
- ➔ Consider OBP varying from 0 to 1
- ➔ This allows us to get a total number of BB, 1B, 2B, 3B, HR depending on OBP
- ➔ Take derivative of formulas and plug in numbers

Runs Created vs BaseRuns

Event weights from Runs Created

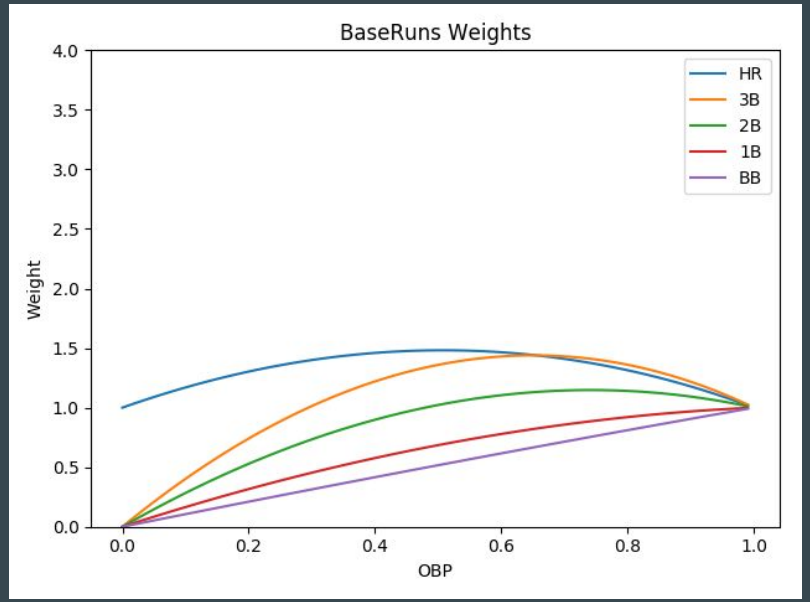
- ➔ Whoa.
- ➔ Home runs not always above 1
- ➔ A home run is eventually worth 4 runs? This is crazy. Those runs were scoring anyhow
- ➔ A walk is worth nothing but a single is 1 run? That runner is scoring regardless.



Runs Created vs BaseRuns

Event weights from BaseRuns

- ➔ This is more like it!
- ➔ The main issue is the value of the triple (worth more than a home run above .500 OBP?)
- ➔ If the triple is over weighted, the double probably is too. But still, this already has some advantages.



Runs Created vs BaseRuns

Moral of the story: non-linearity is good, but actually modeling run scoring is better

- Getting on interacts with scoring rate

What about for individual players?

Player-Level Run Estimators

We started LWTS on individual players and we freely moved between team and players

Why can't we do the same for non-linear estimators?

Actually, we ignored a potentially big problem: the "Ecological Fallacy"

Effects at a group level may not apply at the individual level

Ecological Fallacy in action:

- ➔ 2000 presidential election: strong correlation between states with higher % of AA voters and states voting predominantly for GWB
- ➔ Infer AA voters more likely to vote Republican?
- ➔ In actuality: 90% of AA voted for Al Gore
- ➔ Why? AA voters made up a smaller % of voters, southern states have higher % of white voters who voted GWB

Player-Level Run Estimators

Ecological Fallacy in run estimators

- Definitely causes problems in good players

Runs Created has the term OBP x SLG

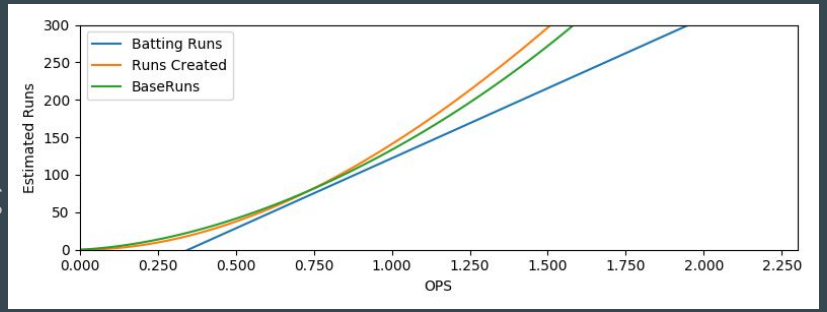
- Barry Bonds does not get to move his on-base self over!

This produces huge effects at the individual level

Huge effects at individual level (OPS vs ... plot)

BJ's flattened RC form

BaseRuns more indicative of run model (getting on x scoring



Player-Level Run Estimators

Ecological Fallacy in run estimators

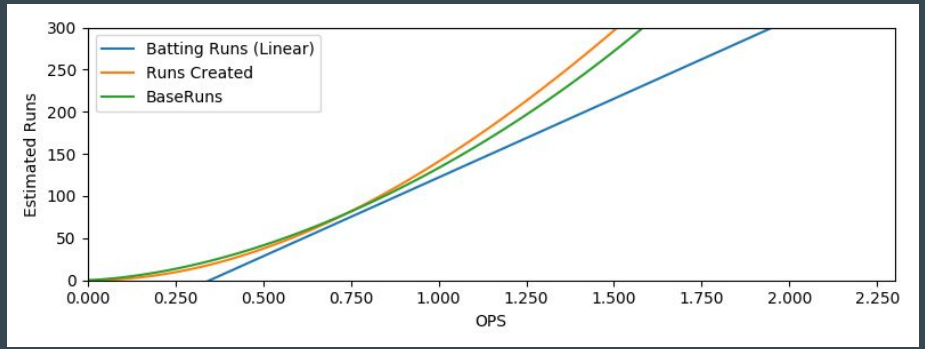
- Definitely causes problems in good players

Runs Created has the term OBP x SLG

- Barry Bonds does not get to move his on-base self over!

This produces huge effects at the individual level

- ➔ Team OPS is consistently around .730
- ➔ Things are okay around there
- ➔ At the extremes, this does not look good!



Player-Level Run Estimators

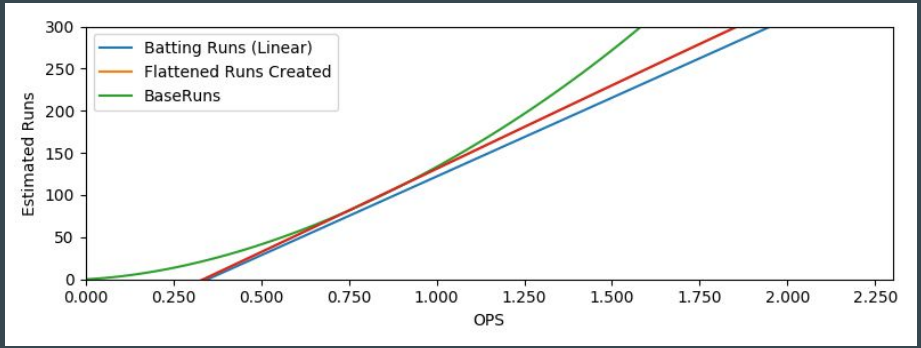
A “Flattened” version of RC

- ➔ Put individual with 8 other “average” players with .300 OBP and .400 SLG
- ➔ A factor: 8 x .300 = 2.4
- ➔ B factor: 8 x .400 = 3
- ➔ C factor: Total plate appearances
- ➔ Final subtraction: remove contribution from 8 phantom players

Does it look any better? What are we even after? It’s basically linear now

$$A = H + BB, \quad B = TB, \quad C = AB + BB,$$

$$RC = \frac{(2.4C + A) \cdot (3C + B)}{9C} - .9C$$



Run Estimators Wrap-Up

What's the point of run estimators and run estimation?

- Always good to remember why you're doing something

Working towards the quantity of interest: runs

- We can take for granted an intuition like BA as a fine measure or FG% in basketball
- Theory and empirics can go beyond intuition

Modeling the run scoring process informs everything

- Ad-hoc models drove classical stats, more informed models drove improved metrics

Run Estimators Wrap-Up

Consensus appears to be non-linear estimators should not be applied at player level

- I can't find player level values anywhere
- For MLB level, linear works fine

Otherwise, BaseRuns is the "gold standard" for run estimators

- Improvements have been made. Research is typically on the scoring rate term (B)

Overall, we worked toward our goal: runs are what matters

- We leveraged our knowledge of baseball to get to a better run scoring model

Up next, a component we've ignored: baserunning

Next Week

The four factor model for basketball