Data 8 Connector: Sports Analytics

•••

3/13/18

Course Stuff

- HW3: 4th Down Bot
  - Mostly written questions, a bit of programming
  - Due April 6, 5pm
- Data acquisition: any issues?

Review

Review: Expected Shot Value

This was our last foray into expected value modeling

The goal to observe the spatial nature of basketball

And try to better capture shooting performance

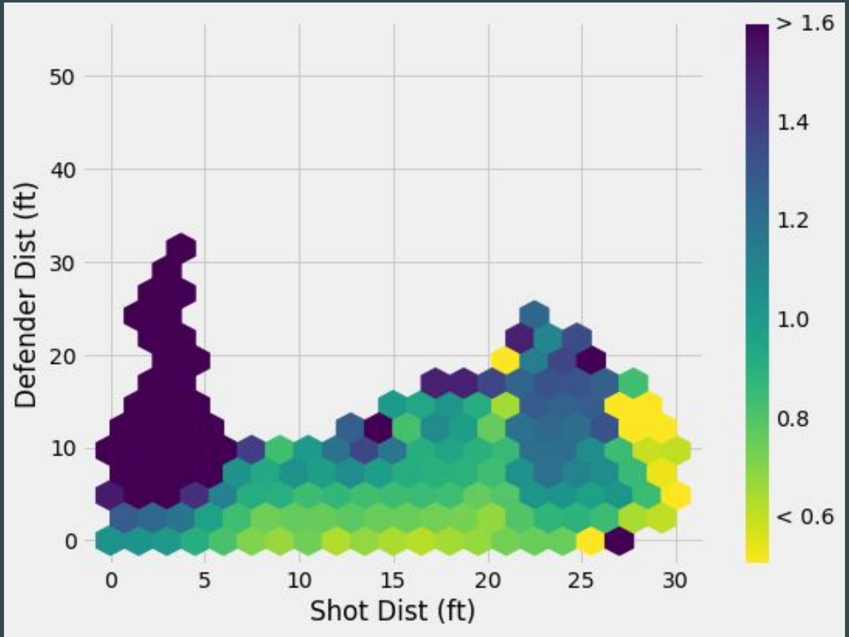
- ➔ Two obvious drivers of performance being shot distance and defender distance

Review: Expected Shot Value

Some things we "learned"

Or at least validated, which is also good

- ➔ Shots further away are harder
- ➔ Shots with a defender nearby are harder
- ➔ Breakaway layups are very high value
- ➔ Open 3s are high value
- ➔ Shots at the basket, regardless of defender are good value
- ➔ Contested long range 2s are not good



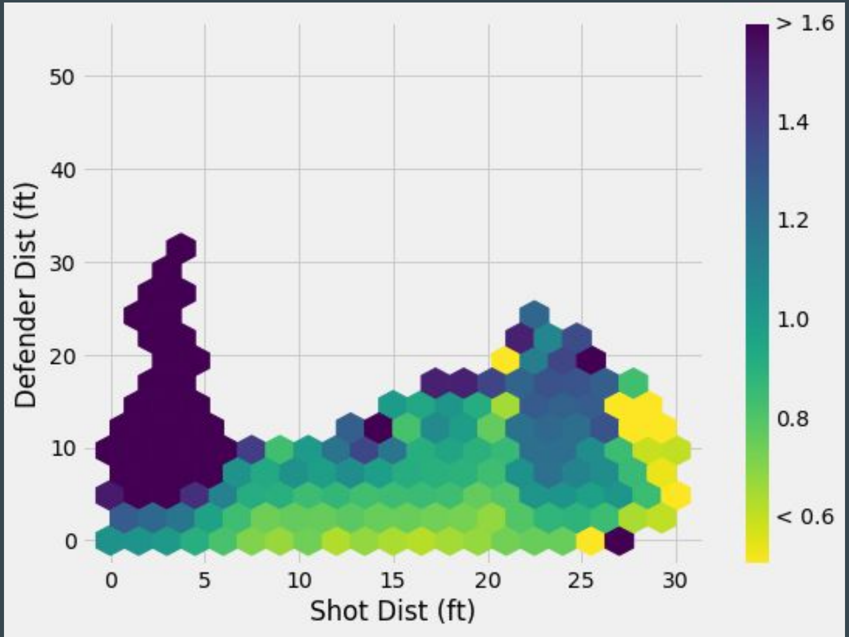
Review: Expected Shot Value

Some things missing

- ➔ We don't know anything about the effect of the shot clock, if there is one
- ➔ We don't know anything about the effect of the style of the shot (catch and shoot, pullup, drive, etc)

This exposes some limits of modeling

- ➔ How can we account for enough that's important?
- ➔ Are you even measuring the important things??



Review: Expected Shot Value

Intuition suggests players like Curry must be guarded closer

- ➔ Evidence says no?

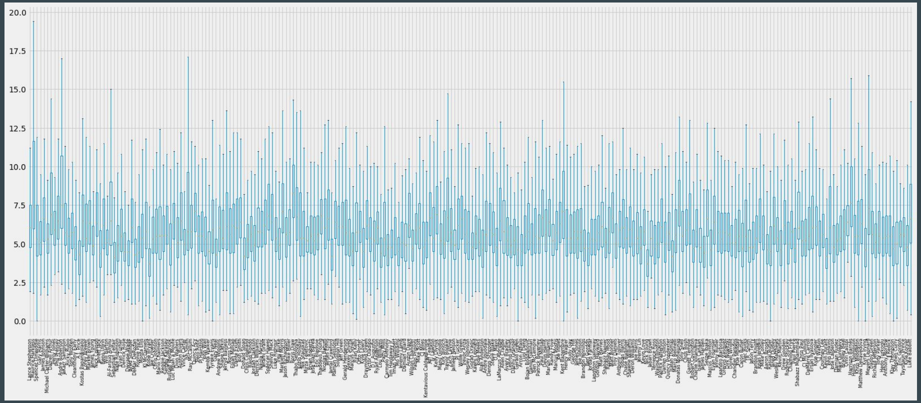
More like...

ceteris paribus, everything else equal

...players like Curry guarded closer

What is everything else?

Things that make up a good player who will get himself open



Expected Possession Value

Use full player tracking data and try to estimate value for any configuration

Player tracking data

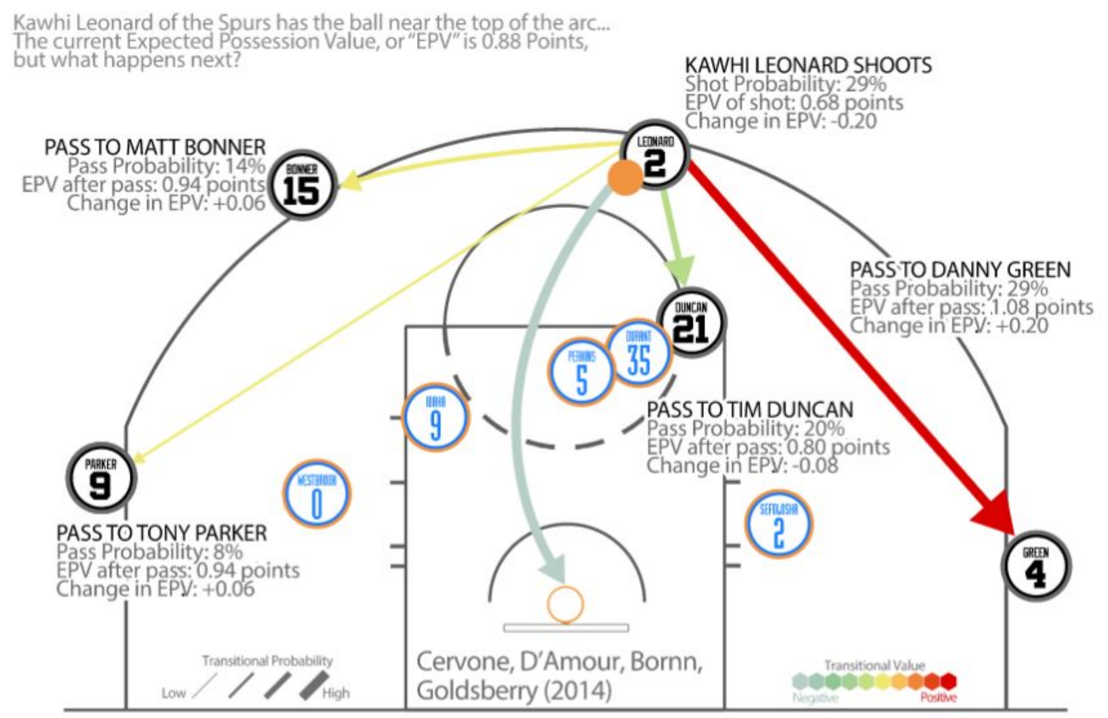
- ➔ X-Y locations of all ten players and the ball sampled 25x per second
- ➔ Augment with things like who has the ball and other macro events

What follows is a high level summary of recent advanced research:

Cervone D, D'Amour A, Bornn L, Goldsberry K. *POINTWISE: Predicting Points and Valuing Decisions in Real Time with NBA Optical Tracking Data.* MIT Sloan Sports Analytics Conference 2014. 2014.

The goal is to get a feel how people are taking advanced approaches to timeless problems

Expected Possession Value



Kawhi Leonard of the Spurs has the ball near the top of the arc...
The current Expected Possession Value, or "EPV" is 0.88 Points,
but what happens next?

KAWHI LEONARD SHOOTS
Shot Probability: 29%
EPV of shot: 0.68 points
Change in EPV: -0.20

PASS TO MATT BONNER
Pass Probability: 14%
EPV after pass: 0.94 points
Change in EPV: +0.06

PASS TO DANNY GREEN
Pass Probability: 29%
EPV after pass: 1.08 points
Change in EPV: +0.20

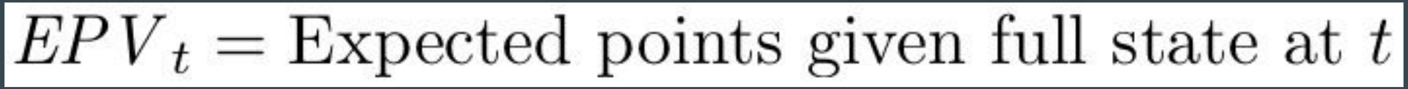
PASS TO TIM DUNCAN
Pass Probability: 20%
EPV after pass: 0.80 points
Change in EPV: -0.08

PASS TO TONY PARKER
Pass Probability: 8%
EPV after pass: 0.94 points
Change in EPV: +0.06

Cervone, D'Amour, Bornn, Goldsberry (2014)

Expected Possession Value

Computed value:



$$EPV_t = \text{Expected points given full state at } t$$

Description

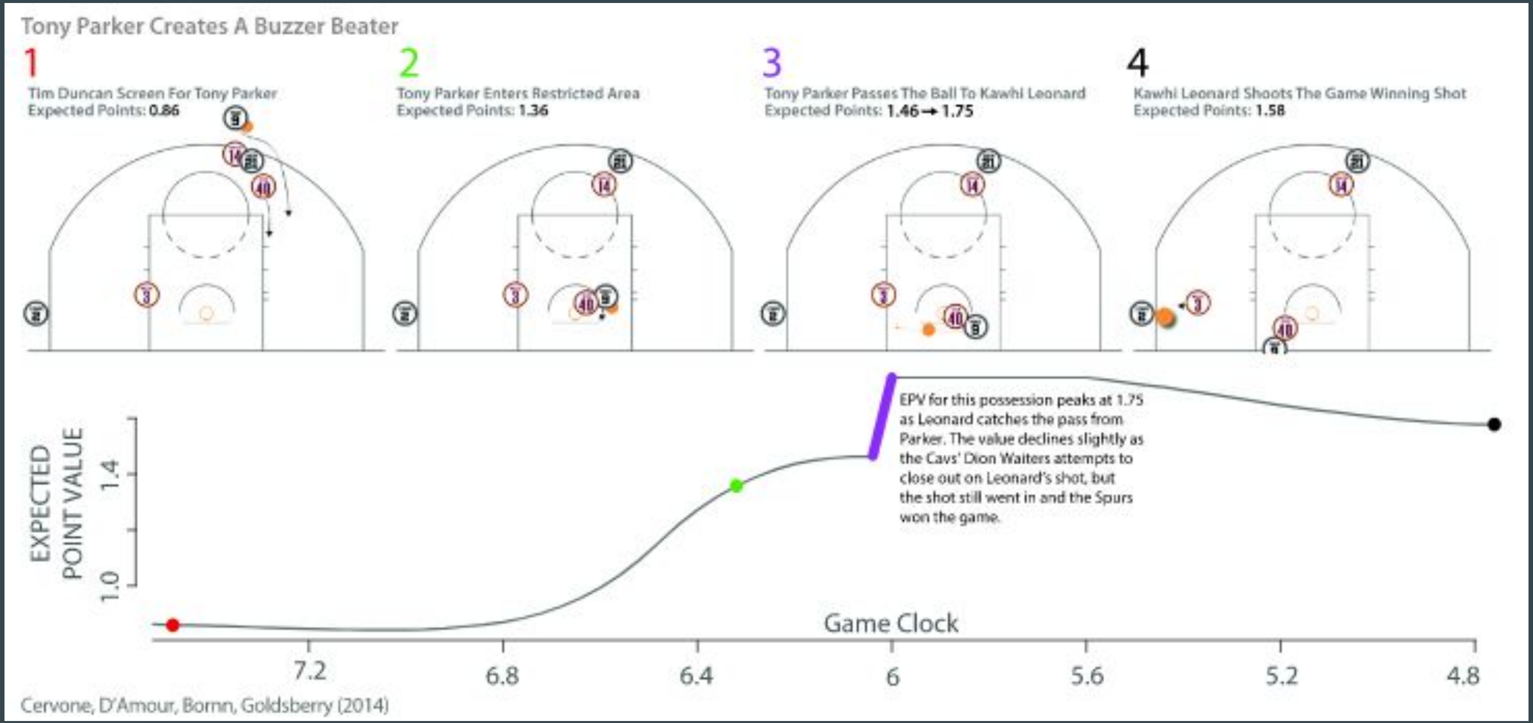
By definition, the current EPV of a possession is the weighted average of the outcomes of all future paths that the possession could take. Calculating this requires a model that defines a probability distribution over what the ballhandler is likely to do next, given the spatial configuration of the players on the court, as we need to understand what future paths the possession can take and how likely they are given the present state.

Expected Possession Value

The model aims to account for...

- ➔ Ballhandler’s shooting tendencies and abilities
- ➔ Ballhandler’s passing tendencies
- ➔ Turnovers and other “basketball” events
- ➔ Defender locations
- ➔ Kinematic motion
- ➔ And probably much more

Expected Possession Value



Tony Parker Creates A Buzzer Beater

1
Tim Duncan Screen For Tony Parker
Expected Points: 0.86

2
Tony Parker Enters Restricted Area
Expected Points: 1.36

3
Tony Parker Passes The Ball To Kawhi Leonard
Expected Points: 1.46 → 1.75

4
Kawhi Leonard Shoots The Game Winning Shot
Expected Points: 1.58

EPV for this possession peaks at 1.75 as Leonard catches the pass from Parker. The value declines slightly as the Cavs' Dion Waiters attempts to close out on Leonard's shot, but the shot still went in and the Spurs won the game.

Cervone, D'Amour, Bornn, Goldsberry (2014)

Expected Possession Value

Deriving metrics:

- ➔ EPV-above-average: Compare EPV added by a player’s decision on all touches to league tendencies in same situations
- ➔ Shot satisfaction: Total value of shots taken relative to passing opportunities not taken
- ➔ Player substitution, pass satisfaction, improved notion of assists, optimal defense

Review: Run Estimation

The premise of run estimation was to work toward a more proper model of baseball

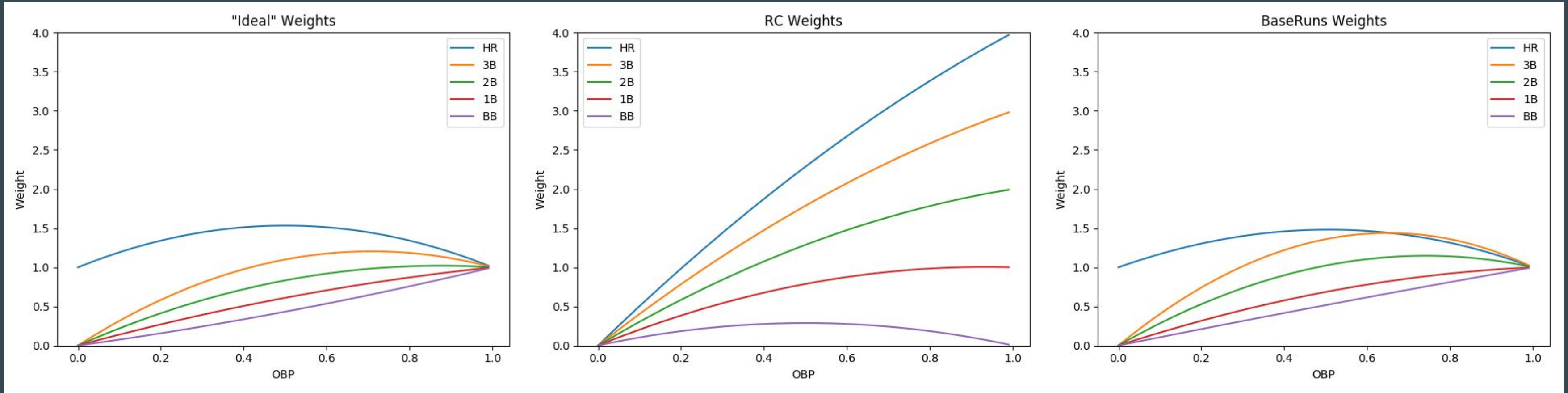
LWTS and RunsCreated are run estimators, but they don’t try to model baseball, just provide a statistically good estimate of run scoring

BaseRuns: directly model the interaction of getting on base and scoring

> AND provide a statistically good estimate of run scoring

Run Estimation

“Ideal” weights vs RC vs BaseRuns as offensive environment changes



Run Estimation

So BaseRuns matches our intuition for the game under more general/extreme settings

Low offense: home runs (ie. power) are extra valuable since runners on base don’t score

High offense: if everyone scores, then a home run is no better than a single or walk

Empirically validated, performs well

Run Estimation

Final comments...

Linear models work well at the player level but miss non-linearity at the team level

Non-linear models are hard to apply at the player level (ecological fallacy)

- A player doesn’t bat in himself

Dean Oliver’s Four Factor Model

Four Factor Model

An empirically driven model that explains basketball performance

It aims to answer “How do teams win basketball games?”

It features four fundamental components of the sport

- ➔ Score efficiently
- ➔ Protect the basketball on offense
- ➔ Grab as many rebounds as possible
- ➔ Get to the foul line as often as possible

Four Factor Model

The outline

- ➔ What is a factor model?
- ➔ How are they used? They're ubiquitous in finance
- ➔ How do we compute the Four Factor Model?
- ➔ What does it tell us about basketball?

Factor Models

First, what is a “factor model”?

Basic idea:

➔ You have have observed outcomes (stock returns, team performance, etc)

➔ There are “factors” or features or just generally quantities that drive performance of your outcomes

➔ Typically you take the factor model as linear

$$\text{Outcome} = \beta_1 \cdot \text{Factor 1} + \cdots + \beta_K \cdot \text{Factor K}$$

➔ Common in Pyschology and Finance (I’ll talk about finance)

Factor Models

Quick aside:

LWTS was a linear model but not a factor model

- We were just estimating the values of events

Dean Oliver’s Four Factor Model

- Factor Models in Finance
  - Use a factor model to explain market phenomena
    - Asset prices/returns, or correlation of returns
  - The most basic factor: the “market”
    - The notion of the market is a bit abstract.
    - Can substitute something for the market though. For example DJIA or S&P500
    - A model for stock returns:

$$R_i = \alpha_i + \beta_i * \text{market return} + \text{idiosyncratic return}$$

    - Market explain/drives returns you observe (and after that, it’s all just noise)
    - Stocks go up/down with the market and everything is correlated

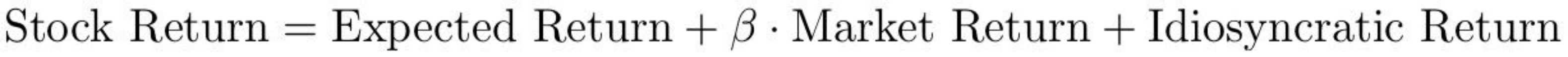
Factor Models in Finance

We use a factor model to explain market phenomena

For example asset prices/returns, or correlation of returns

The most important factor: the “market”

- ➔ The notion of the market is a bit abstract
- ➔ But we can substitute something for the market though. For example DJIA or S&P500 indices
- ➔ A model for stock returns (the famed CAPM)



$$\text{Stock Return} = \text{Expected Return} + \beta \cdot \text{Market Return} + \text{Idiosyncratic Return}$$

Factor Models in Finance

Model for stock returns (the famed CAPM)

$$\text{Stock Return} = \text{Expected Return} + \beta \cdot \text{Market Return} + \text{Idiosyncratic Return}$$

The “market” explain/drives the variation in returns you observe (and after that, it’s all just noise)

Stocks go up/down with the market so every stock is correlated

Factor Models in Finance

Model for stock returns (the famed CAPM)

$$\text{Stock Return} = \text{Expected Return} + \beta \cdot \text{Market Return} + \text{Idiosyncratic Return}$$

Magnitude of idiosyncratic return and beta: fraction of variation due to the market

Beta > 1: stock is extra reactive to market

Beta < 1: stock not as reactive

Beta = 0: stock not affected by market

Beta < 0: stock anti-correlated with market

Factor Models in Finance

Another (potential) factor: momentum

Momentum is recent historical performance. Trending up? Trending down?

What do you do with the factors? Try to build portfolios based on them

➔ Match the market return but minimize risk

➔ Bet on momentum

➔ Anything you want: you look at the “betas” and decide how you want the portfolio to exposed to the risk of the factors

Factor Models in Finance

So what can be a factor?

Potentially anything you want. But you might concoct a bogus factor which doesn’t explain anything about the market (and potentially lose a lot of money!!)

What else?

- Macroeconomic factors: economy level data like employment or interest rates
- “Fundamental” factors: things like value (price/earnings ratio) or momentum
  - Anything related to the company
- Statistical: not firm or directly observable
  - Observed in data but nothing fundamental/macro explains it.

Factor Models in Finance

What about the weights? How do we determine the importance of factors?

Ultimately, we have to figure out on our own how to weight the factors for each stock

- ➔ Market factor is the most important for stocks (not bonds). But how important?

The short answer: regression

Dean Oliver’s Four Factor Model

So what are Dean Oliver’s Four Factors? We know the intuitive ideas from basketball:

- ➔ Efficient shooting
- ➔ Turnovers
- ➔ Rebounding
- ➔ Free Throws

These will be fundamental factors: compute values from team observables

Okay, we know the concepts, but what should we actually compute?

Dean Oliver’s Four Factor Model

Shooting:

$$eFG\% = \frac{FG + .5 \cdot 3FG}{FGA}$$

Rebounding:

$$\text{OREB\%} = \frac{\text{Off Reb}}{\text{Off Reb} + \text{Opposition Def Reb}}$$

Turnovers:

$$\text{Turnover Rate} = \frac{\text{Turnovers}}{\text{Possessions}}$$

Free Throws:

$$\text{FT Rate} = \frac{FT}{FGA}$$

Dean Oliver’s Four Factor Model

Okay, so now what?

- ➔ We want to explain team performance
- ➔ Dean Oliver suggests weighting the factors by importance: 40, 25, 20, 15

Okay... so we just multiply each of those values by those numbers and add them up?

None of those numbers are related at all: different avg level, different variation

Z-Scores

A bit of an important digression for a very useful and great tool for standardizing data

Suppose you have some variables on different levels and different variations

We do. In fact, this happens basically all the time.

The goal is to standardize the variables

Remove average level, inherent variation, and units

That way, you can quantify effects in terms of standard units like *standard deviation*

➔ An example: two variables see a change of .1 and 100, respectively. Two totally different levels of change

➔ But if the standard deviation for those variables are .1 and 100, then it’s really the same level of change

Z-Scores

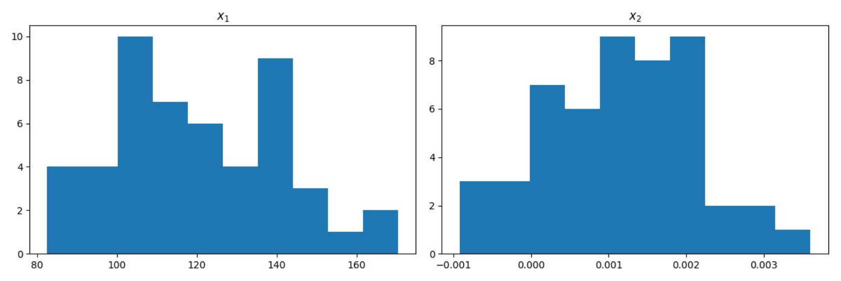
For data observations X_1, ... X_N, the Z-Scores are given by

$$Z_i = \frac{X_i - \text{Avg of } X_1, \ldots, X_N}{\text{Std Dev of } X_1, \ldots, X_N}$$

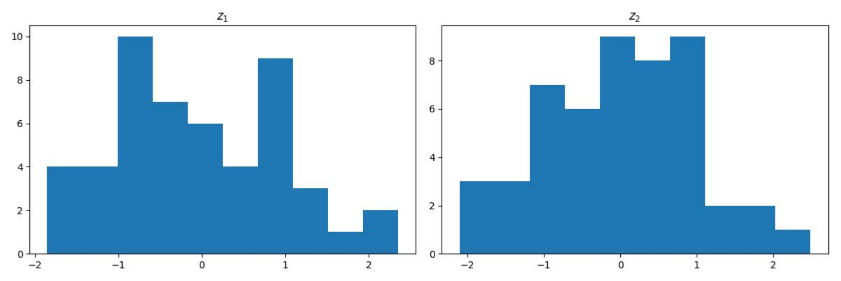
The mechanism is to de-mean the data and then scale it so it has Std Dev = 1

Z-Scores

Non-standardized data:



Standardized data:



Histograms have the same shape: data hasn’t fundamentally change

Just shifted and scaled

Z-Scores

Okay, so now everything is on the same scale

➔ 1.1 means the same thing for each factor

Coefficients have meaning: they now indicate a relative importance

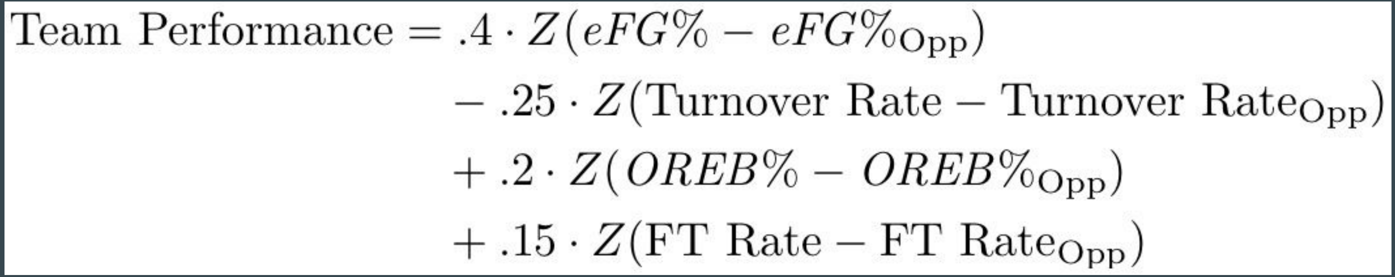
And now we can just use Dean Oliver’s suggested weightings

Dean Oliver's Four Factor Model

But wait, one more issue: there are 8 values for the team and its opposition

We take the raw components, compute the difference, and then compute Z-scores

Dean Oliver's Four Factor Model:



$$\begin{aligned}\text{Team Performance} = {} & .4 \cdot Z(eFG\% - eFG\%_{\text{Opp}}) \\ & - .25 \cdot Z(\text{Turnover Rate} - \text{Turnover Rate}_{\text{Opp}}) \\ & + .2 \cdot Z(OREB\% - OREB\%_{\text{Opp}}) \\ & + .15 \cdot Z(\text{FT Rate} - \text{FT Rate}_{\text{Opp}})\end{aligned}$$

Turnovers are bad hence the negative sign

Dean Oliver’s Four Factor Model

So where did those weightings come from?

- Honestly, don’t know and I haven’t seen anyone offer an explanation

However, I have ve seen evidence that shooting should be weighted more heavily

- We will revisit this when doing statistical inference

Anyhow, let’s look at the data and how this all works