LS88: Sports Analytics

Data, Observations, & Causality

Outline

- ➔ Experimental vs observation data
- ➔ Causal inference
- ➔ Manual data collection in sports
- ➔ Modern data collection in sports

Observational vs Experimental Data

What is experimental data?

- ➔ Data collected in a controlled setting
- ➔ The collector can assign participants for various “treatments”
  - ◆ +/-: shuffle players around and mix lineups
  - ◆ 4th Down: make the decision for the coach
  - ◆ Baseball: make bullpen choices for the managers
  - ◆ In medicine: giving new drugs to compare to existing treatments
  - ◆ Websites: optimizing websites/campaigns for effectiveness/revenue generation

Observational vs Experimental Data

What is observational data?

- ➔ No control: you get what the data generators (players, coaches, teams) give you from their play
- ➔ The participants select their own treatments
  - ◆ Coaches choose their 4th downs
  - ◆ Players/Coaches/Teams select themselves so no randomization of lineups
  - ◆ British Doctors Study: a longitudinal study of people that provided a key link between smoking and lung cancer

Causal Inference

The gold standard: directly predict outcomes/pathways

➔ Isolate an effect

- Ex: How much a suboptimal 4th down decision costs in wins
- Ex: How much a player is worth in +/- (and not due to other players, coaches, system, etc)

Causal Inference

Proper causal inference takes this statement

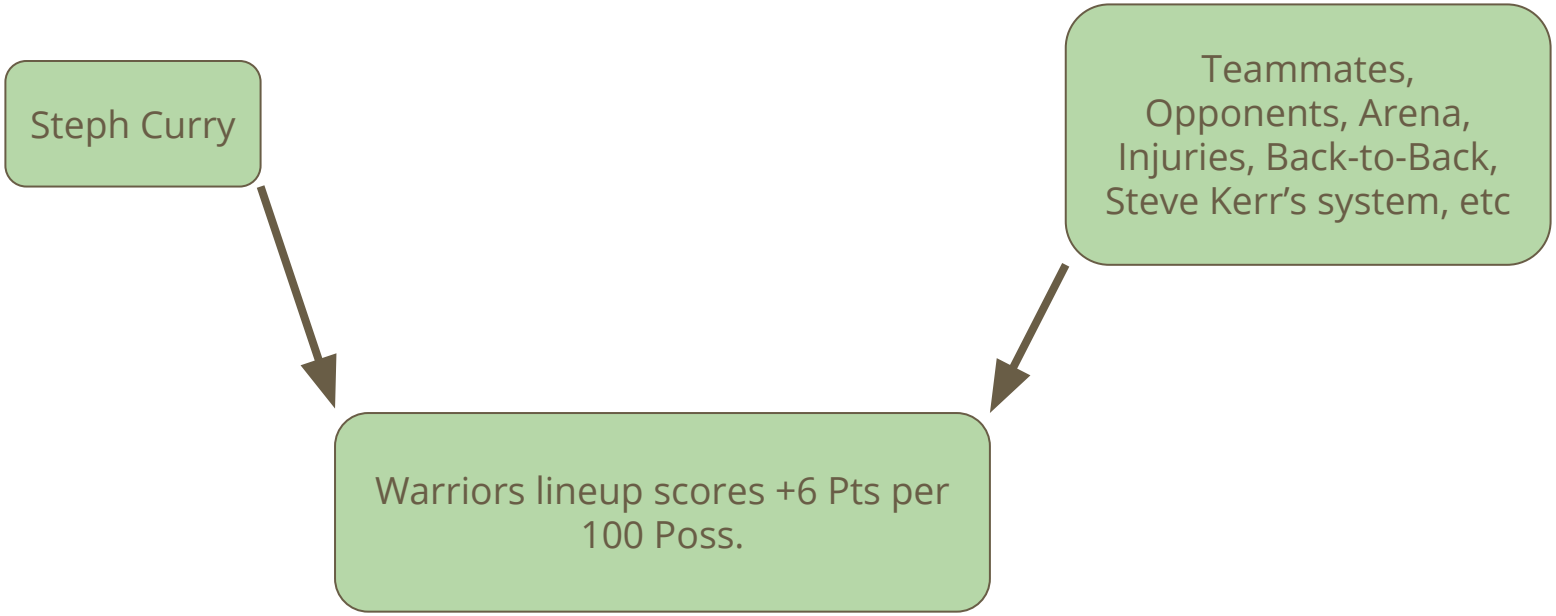
> *Changing player A to player B in a lineup is **associated** with a +4 increase in expected points per 100 possessions*

and turns it into

> *Changing player A to player B in a lineup **causes** a +4 increase in expected points per 100 possessions*

In the first statement, it is implied that there are many other (possibly) hidden factors that could be at work

A Causal Diagram



All About that Bias

What is bias?

- ➔ The difference between the truth and expected output of our model
- ➔ It is *not* the difference between the truth and our estimate: that's error
- ➔ Ex: A player is truly worth +0 points above average per 100 possessions.
  - ◆ Each season our model estimates the player is +4. *The bias is +4.*

Our goal is to remove bias: account for as much as possible to remove biases that will plague our models

All About that Bias

Selection Bias

Some individuals are more likely to be selected for study than others

Estimator Bias

Mathematical bias (mathematically, regression estimates are unbiased but other models can be biased)

Omitted-Variable Bias

Failure to include key variables in a regression

All About that Bias

Attrition Bias

Loss of participants in a study follow-up

Recall Bias

A participant can not reliably recall their own behavior

Observer Bias

The observer subconciously influences the experiment

Aside: Correlation does not imply causation

I presume we've all heard this scolding statement

Fear not our friend correlation, associations are still good and helpful

- ➔ First off, I doubt anyone's proved a causal relationship without an association to start with
- ➔ You can still make predictions:
  - ◆ Run/Point scoring: we haven't formally established a causal relationship, but intuition plus association is good enough
  - ◆ Pairs trading: we don't care *why* two stocks are correlated, we just want to exploit the association to trade

When can we do causal inference?

Do you have experimental data?

…Did you design a good experiment (you properly controlled participants and isolated effects)?

…Have at it.

Do you have observational data?

…Did you do a really good job collecting the data and did you use some important tools from statistical theory to do the heavy lifting?

…Okay, then go ahead.

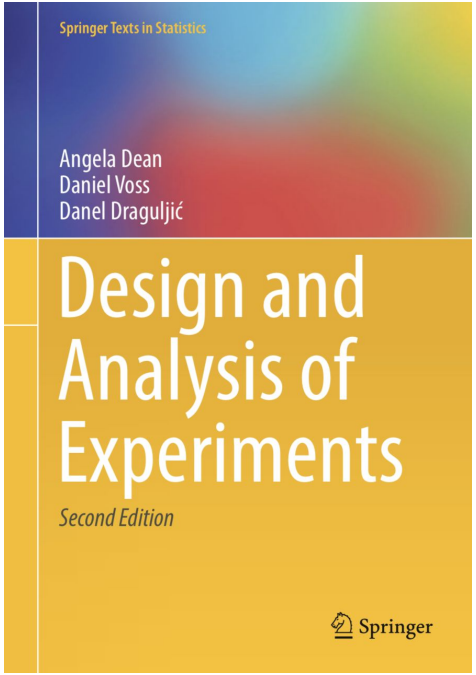
Caveats of Causal Inference

Experiments can be garbage

➔ Experimental design is no joke

- 800+ dense pages devoted to the topic

Just because you run an experiment doesn't mean it's guaranteed to be good and properly tell you about your hypothesis

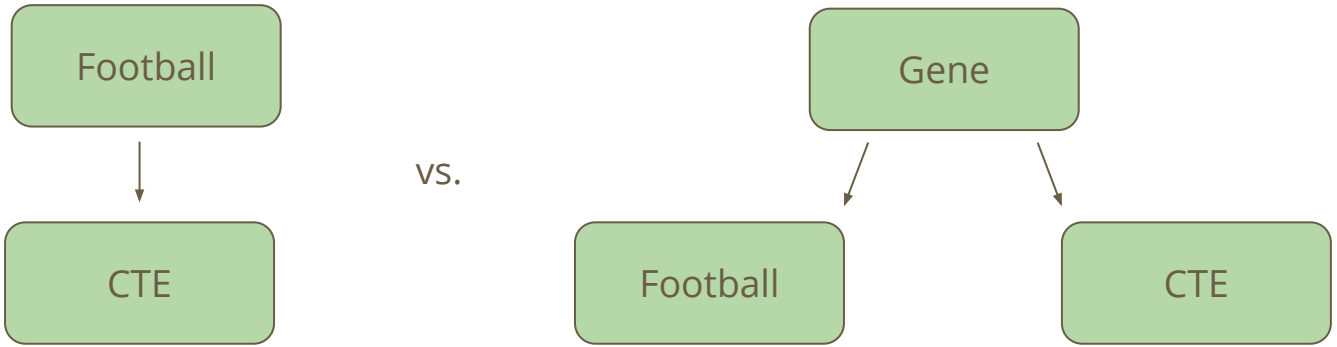


Caveats of Causal Inference

You didn't account for a *confounding* variable

➔ Football is associated CTE, but maybe an unobserved confounder (like a gene) causes CTE *and* causes the person to play football

This is the *exact* alternate hypothesis that was proposed regarding smoking and cancer



Data Collection in Sports

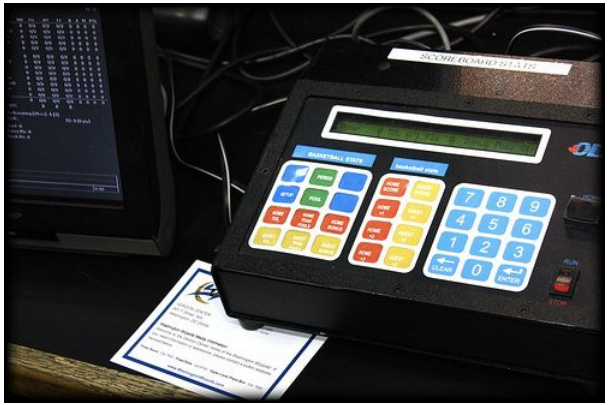
It’s all observational so remember it’s all about bias

The prevailing theme: gather more and more data

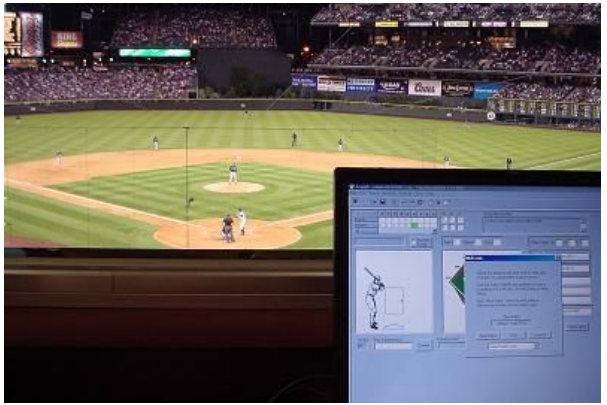
Manual Collection: Stringers

Stringers are people who are paid to collect data

➔ NBA, MLB



NBA Stat Box



MLB Stringer Software

Manual Collection: Stringers

Companies, volunteers, collectives

➔ Opta, Retrosheet, Armchair Analysis

The job is akin to a stenographer

A related tool: Amazon’s Mechanical Turk Service

Modern Data Collection

We need more and more data to continue to reduce bias

- We'll never get in-game experimental data. So we need to collect more and more

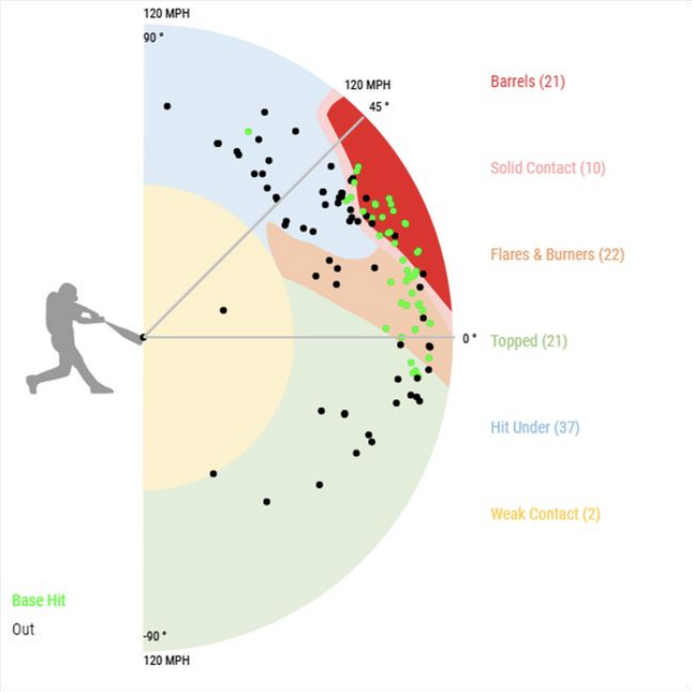
Either have the stringers track more…

- …or use something new

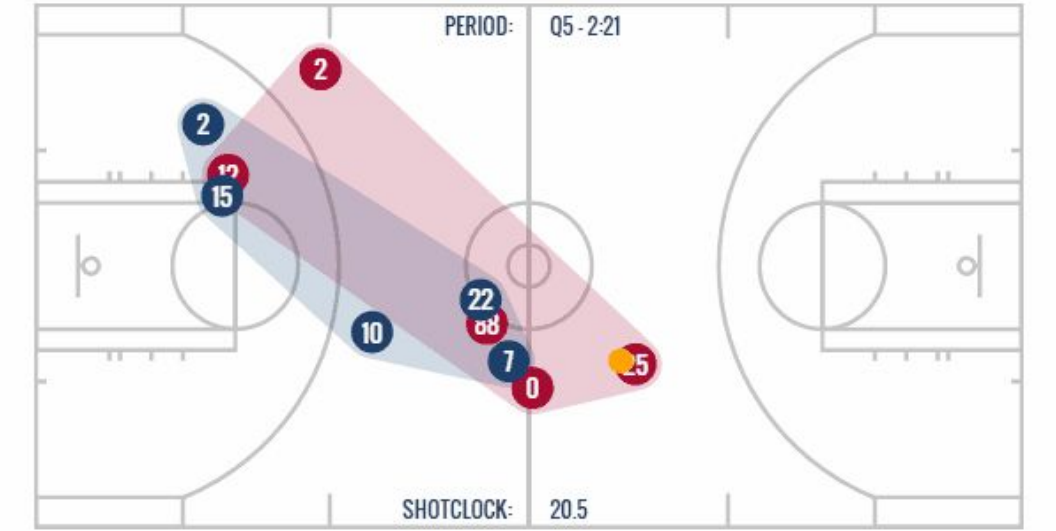
Solution: Radar, Camera, RFID, GPS

- ➔ Statcast: Radar ball tracking
- ➔ SportVu/Second Spectrum: Camera tracking of players in NBA or Soccer
- ➔ NFL: RFID tracking of players

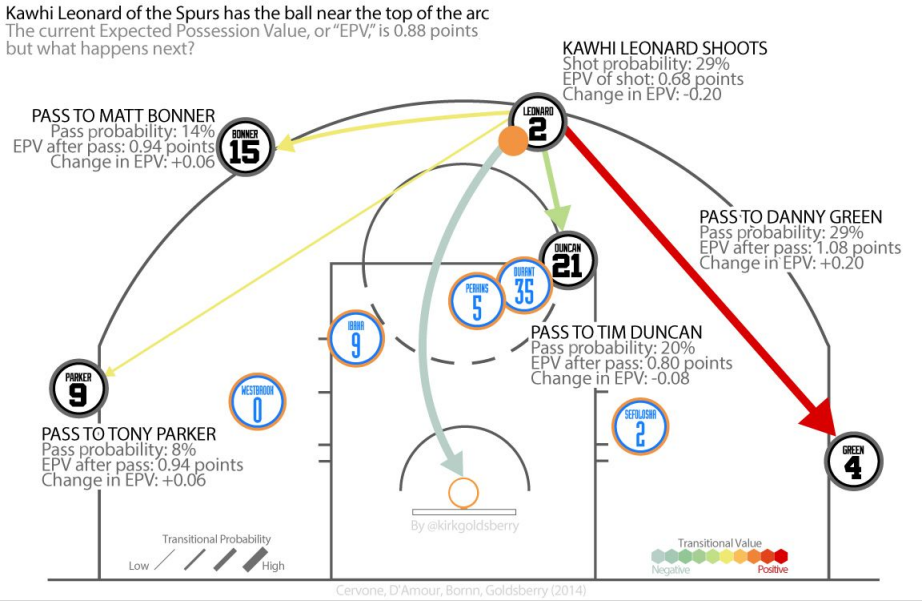
Statcast: Launch Angle



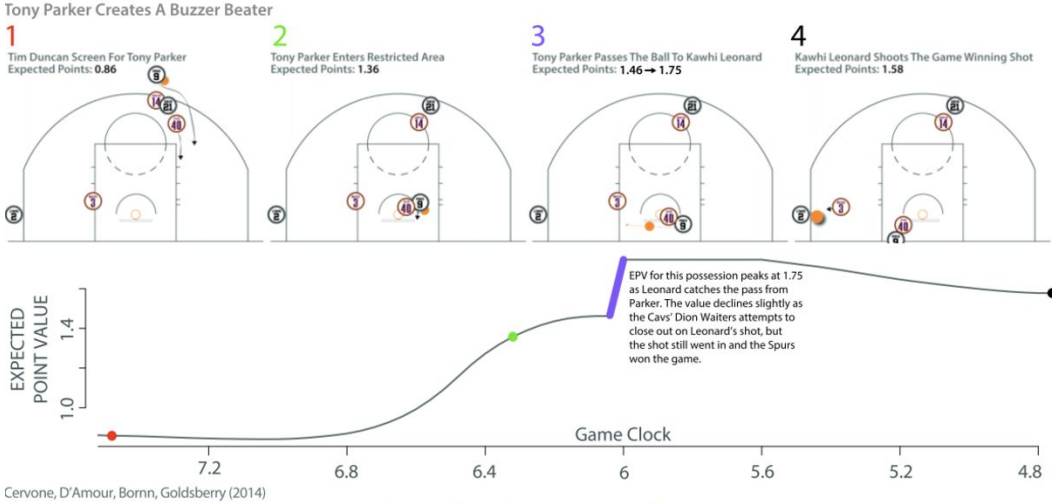
SportVu: Player Tracking



SportVu: Expected Possession Value



SportVu: Expected Possession Value



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

Figure 2. EPV throughout the Spurs' final possession, with annotations of major events.

Summary

- ➔ Causal inference is our objective but it's *very* hard
  - ◆ You can still make predictions without it, but there are caveats
  - ◆ We can predict a lineup performance from +/- ratings, but implicitly we're ignoring all the hidden stuff that actually affects the rating
- ➔ Observational data is a curse but it's all we have
  - ◆ Most of what's out there is observation data anyhow. Exception: websites and A/B tests
  - ◆ You can infer causation if you know what you're doing
  - ◆ Often the observational data blunts statements
    (I've tried to be very careful with my words this semester)
- ➔ Modern data collection in sports aims to solve the problem
  - ◆ Faster collection and richer data