LS88: Sports Analytics

Fall 2018 Connector Course

What is this course?

Data Science + Sports = Sports Analytics

Why this course?

Data Scientist + Sports Enthusiast = Teach Sports Analytics

Why this course?

Data Science Student + Sports Enthusiast = Learn Sports Analytics

What is this course about?

Demystifying sports analytics

Lots of talk, lots of jargon, lots of sites and blogs.

Theory/methods/motivations/history/data/statistics

All as it applies to sports

What is this course about?

Demystifying sports analytics

- Usually quite simple:
  - Knowing to look, ie. not assuming from the eye test, and collecting data that wasn't there before followed by basic summarizations
  - Ex: Pick and roll pairs efficiency. The challenge is *knowing* and *getting* the data
- There's still a hefty chunk of quite interesting work that isn't so simple.

What is this course about?

The what, the how, and the why

I could plow through a lot of concepts in two weeks and give a coarse overview.
Ex: What is a park factor? Not how it's computed or how to use it or why.

The point is to get a feel for how this kind of work is done and why it's done the way it is.

Also a great way to connect with the DS/Stats/Math field at-large through sports and data

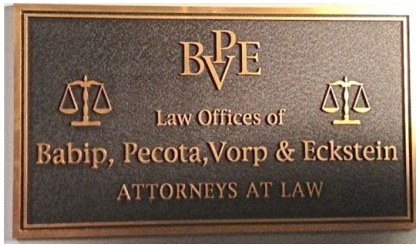
What is this course about?

Hands-on with data and Python

We’re going to try to be as hands on as possible.

You don’t want to hear me just talk. And it’s pretty fun developing this stuff.

Analytical/Scientific Thinking



Bill James of Sabermetrics: “the search for objective knowledge about baseball.”

Developing hypotheses, using logic and reasoning, and grounding it in data

Ex: Is a sacrifice bunt a worthwhile strategy?

(Go to the data and see if a bunt increases your chances to score)

Analytical/Scientific Thinking

> When big league scouts road-tested a group of elite amateur prospects, foot speed was the first item they checked off their lists…
>
> *-Moneyball*

Five tools: Run, throw, field, hit, and hit with power

But do these actually matter? Which is the most important?

Cognitive Bias

- ➔ A lot of economic models are built on rational decision making
- ➔ Behavioral econ/finance and psychology have demonstrated humans fail at rational decision making

Cognitive bias is the mind failing at interpretation and decision making

Premise of Moneyball A's:

> Exploit irrational decisions by MLB clubs to build a good team, cheaper

Outcome vs Process THinking

- ➔ Perfect predictions in certain fields aren't possible
- ➔ End goal: a decision making process that is as unimpeachable as possible
- ➔ Deandre Jordan and Joey Dorsey
  - ◆ Forget the outcome (Dorsey a big bust, DJ a sleeper success)
  - ◆ Can we investigate the process? Yes
- ➔ SF Giants, Brian Sabean, 3 WS in 5 years

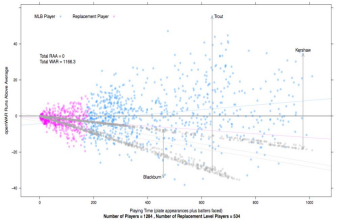
More than just talking

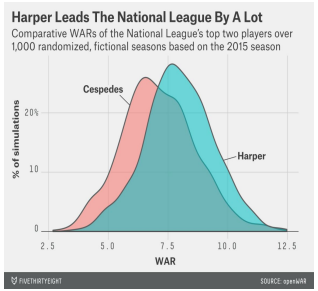
This semester, we're going to try to do it the best that we can

- ➔ Breaking down a sport and understanding how it works
- ➔ Measurement, statistics, and how to interpret given all the assumptions
- ➔ Modeling and decisions

What do we want to explore?

Measuring performance



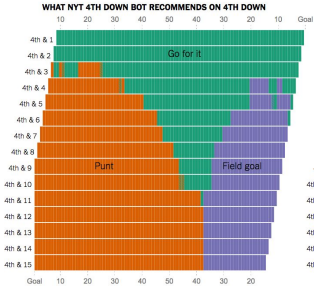


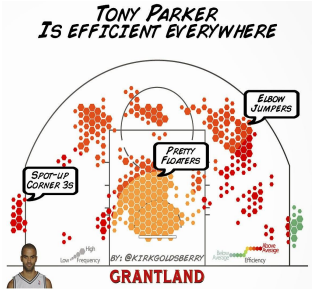
Advanced Metrics

Decision Analysis

Regression modeling

Inference





Assorted Topics

What we hope to learn

What goes into producing advanced metrics

The universality of it all can serve you well elsewhere

Read and discuss data oriented analysis (and ignore the taeks)

You get to appreciate the games at a higher level

Begin your own path of getting into the world of sports analytics

Maybe you are just curious, maybe you play DFS, maybe you want a job?

Course Details

Who are we

- ➔ Your Instructor: Alex Papanicolaou
  - Formerly a Data Scientist in FX trading
  - Researcher in risk for finance
  - *Not* a professor

- ➔ Your Course Assistant: Alex Almond
  - Junior in Applied Math
  - Took the course Spring 18

Who are you

- ➔ Data 8 students
  You like the connector? Maybe it was the most interesting? (I hope not)
- ➔ Data 100 students
  Good opportunity to apply things you've learned?
- ➔ Something else?
  Stats/CS/Math/Psych/Econ/Biz
  Background somewhere at the level of Data 8

Who are you

We're a mix: Data 8 students and non-Data 8 students

➔ Goal: A more open class (no prerequisite for Data 8)

If you're a Data 8 student:

- You'll be tossed into the deep end (a bit)
- But we're not here to drown you

If you're a more advanced student:

- Be a mentor: in the group project, on Piazza, in class
- Push your limits: go beyond the bare minimum answer

What we expect

Try

What we expect

- ➔ Seriously, just try.
- ➔ You do not need to know every sport
  You will need to learn the rules though or else little will make sense
- ➔ You need a curious mind and a willingness to be wrong
  Throw out any idea/though/potential answer
  Wrong answers are great too! Say “I think/know this is wrong but…”

The worst thing you can do is stay silent and assume everyone else knows what’s going on.

Pro tip: they don’t.

Course Structure

- ➔ Still a work in progress: you're round 2 guinea pigs
- ➔ Less lecture this time: ~1hr, then lab
- ➔ No textbook but I'll point to some sources
- ➔ Homework: bit of coding, bit of reading, bit of writing (~every 2 weeks)
- ➔ Quizzes: just 5-10 min, recapping stuff
- ➔ Final project: groups (mix of Data 8 and non-Data 8 students)

Specifics to come... (syllabus on Piazza, OH schedule, etc)

Final Project

- ➔ Groups of about 3-4
- ➔ Final written report at the end (3-5 pages). Presentations last week.
- ➔ More details on Piazza

Topic Outline

In no particular order:

- ➔ Measuring performance
  - ◆ Baseball: classical stats, OBP, OPS, and run expectancy
- ➔ Expected value modeling
  - ◆ PER, DVOA, and more
- ➔ Decision analysis
- ➔ Testing hypotheses
- ➔ Regressing modeling
- ➔ Forecasting/projection
- ➔ Other topics (rankings, etc)

“Textbooks”

Nothing required

- ➔ The Book (Tom Tango and Mitchel Lichtman)
- ➔ Analyzing Baseball Data with R (Jim Albert and Max Marchi)
- ➔ Mathletics (Wayne L. Winston)
- ➔ Basketball on Paper (Dean Oliver)

“Textbooks”

Blogs are good

- ➔ The Hardball Times (FanGraphs Blog)
- ➔ 538
- ➔ Beyond the Box Score
- ➔ Nylon Calculus

“Textbooks”

Other interesting books

- ➔ Moneyball
- ➔ The Undoing Project
- ➔ Big Data Baseball

There’s no shortage of resources and these lists are by no means complete

“Prerequisites”

Some things you need to have a bit of an idea about…

- ➔ Central Tendency
- ➔ Dispersion/variation
- ➔ Correlation/Line-of-best-fit/Regression line

Central Tendency

Sample averages and expected values

➔ We flipped a coin 100 times, it came up heads 51 times.
The *sample* of 100 flips *averaged* .51 heads per coin flip

➔ The coin wasn't special (not weighted or biased)
The *likelihood* of heads is .5 (aka 50-50, aka 50%)
We *expected* the coin to come up heads 50 times or .5 heads per coin flip
In the next 100 flips, we still *expect* the coin to come up heads 50 times

Central Tendency

Sample averages and expected values

- ➔ *Expected Value (aka mean)*
  The value we expect *samples* or *observations* to be centered around
- ➔ *Sample average (aka sample mean)*
  The actual average value of the *sample*
- ➔ The sample average will tend to be close to the expected value
  Collect more samples and it should get even closer
- ➔ We don't always know the expected value
  We use the sample mean to estimate it*

*Statistics! Using data to estimate or infer about something!

Central Tendency

- ➔ In 2014, Giancarlo Stanton got on base 39.5% of the time
  He *averaged* about .395 times on base per plate appearance
  In his career, he's *averaged* about .356 times on base per PA
- ➔ Next season, we probably should *expect* Giancarlo Stanton to get on base about 35-36% of the time

Dispersion/Variation

- ➔ I go to Vegas with $1000 and start playing Blackjack for 2 hours
- ➔ The house “edge” means I expect to lose money.
  Let’s say I expect to finish at $800 after 2 hours (-20% return)
- ➔ This is insane, yet I still play. Other than being some kind of degenerate, why would I play this game?
- ➔ If I play 100 hands in 2 hours, I may lose everything, I may win $2k, I may be right at $800
- ➔ There is *variance/variation/dispersion* in the outcomes

Dispersion/Variation

- ➔ Casinos make money because lots of people plays lots of hands.
  Sample average gets closer to the expected value with more samples = Casino makes tons of money
- ➔ I play the game because the variation in the outcome makes it possible to win money
- ➔ How much variation is there? Is there more variation in Craps or Blackjack?

Dispersion/Variation

- ➔ How much variation is there? Is there more variation in Craps or Blackjack?
- ➔ We could answer this one of two ways
  With mathematics: we know the probabilities., so compute the variation in the outcome
  Sample: give a bunch of people a bunch of money and see how much variation there is in what they come back with
  Sample (the cheap way): simulate with a computer

The specific definitions/formulas aren't important now, the key is the concept of variation in outcome

Dispersion/Variation

- ➔ Variations can be large, they can be small
- ➔ Smaller variation is good when we estimate quantities: tighter, more reliable estimate
- ➔ But a smaller variation in a casino game means…

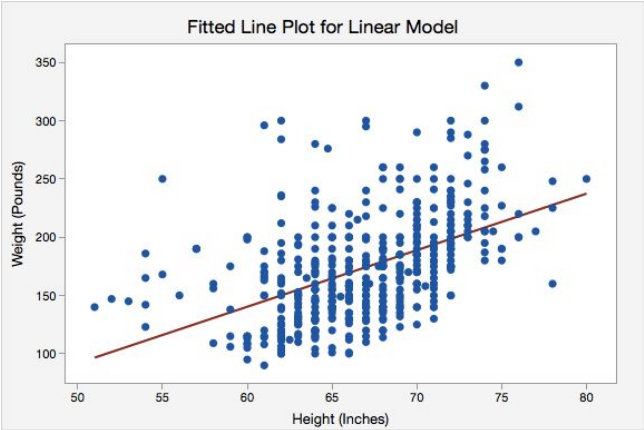
…the average outcome of my play (my sample) will tend to be closer to the expected outcome

…I am more likely to lose money

If there's zero variation, it's a sure thing and I would just be handing money to the casino

Correlation/Regression

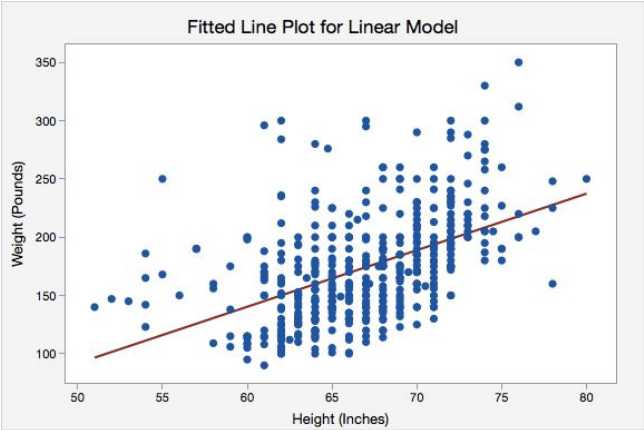
- ➔ We poll 100 students their height and weight
- ➔ We gather the data and plot the result*



*Taken from a PSU course

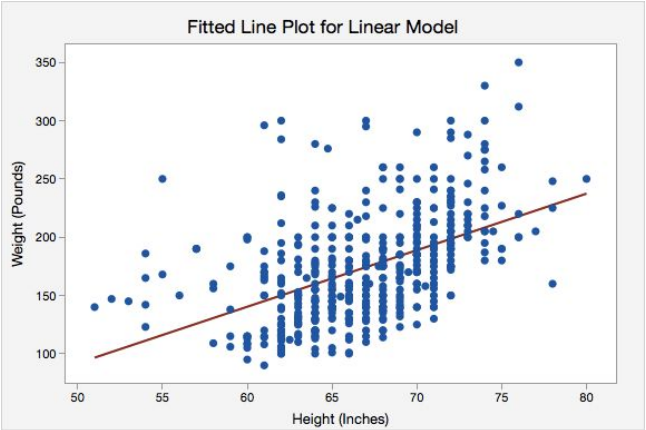
Correlation/Regression

- ➔ Height and weight are *correlated*: taller people tend to be heavier
- ➔ Correlation can be positive or negative
- ➔ Sometimes represented by the letter *R*



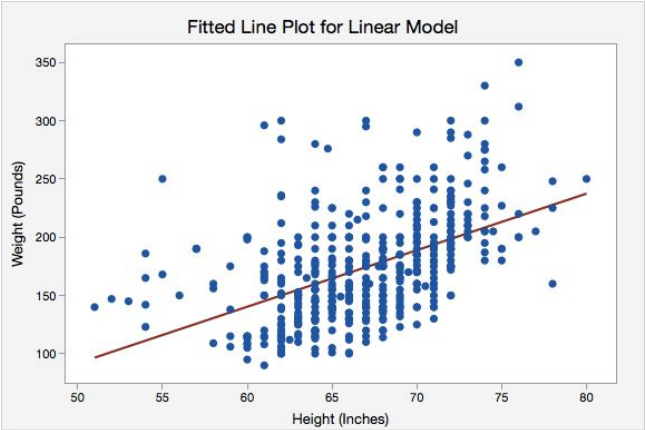
Correlation/Regression

- ➔ The *regression line* describes the general relationship
- ➔ It models the relationship between height and weight



Correlation/Regression

- ➔ We predict a person 65” to weigh about 155lbs
- ➔ For every 1” inch increase in height, the predicted weight increases by 4.854lbs.

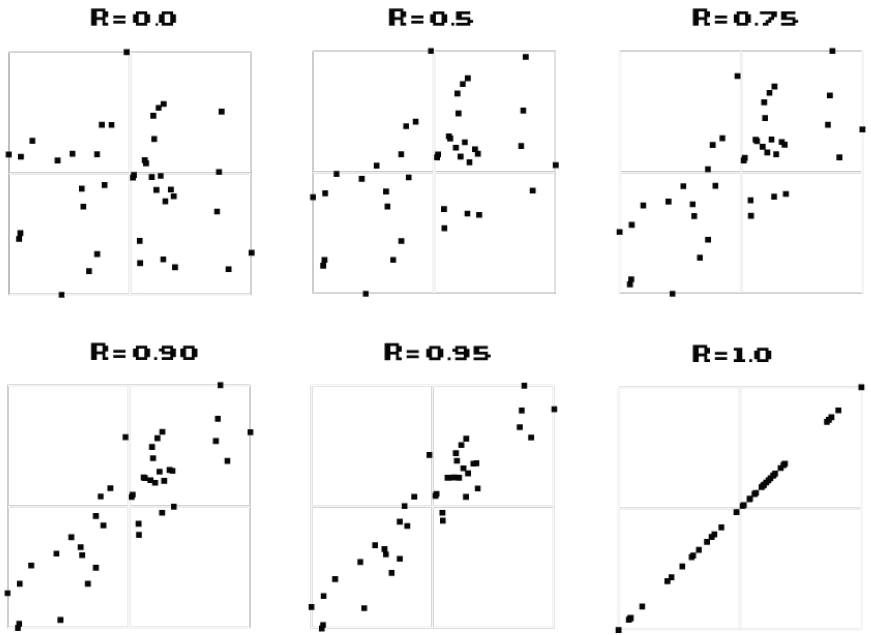


Correlation/Regression

Play *Guess the Correlation*

http://guessthecorrelation.com/

(Hope it still works)



Data

Let’s cover a bit of the data that’s out there to help us answer questions.

Data

What type of datasets are out there?

- ➔ Raw: game box scores; play-by-play; player tracking
- ➔ Extracted Events: hits, runs, points, rebounds, assists, etc
- ➔ Stats: batting avg, total bases, RBI, shooting %, etc

Data

Where can we find this data?

- Websites:
  - Leagues: MLB.com, NBA.com
  - General: ESPN, Baseball/Basketball/Football Reference, FanGraphs
- API/Published
  - PitchF/X, Statcast
  - NBA Stats
- Curated (not necessarily free)
  - Lahman Database, Retrosheet, armchairanalysis.com (cheap with .edu email)
- Other
  - API tools and scrapers published on GitHub (LOTS of repos out there)
  - Data Collectives: Kaggle, data.world

Data

Our focus is on MLB/NBA

Other data is not as prominent or collected as well: e.g. NFL