Sport Analytics. An Introduction

Carlo Favero

Course 20630 Year 2019/20

Sport Analytics

- Sport Analytics is the statistical analysis of economics data.
- Numbers allow us to see what your eyes cannot follow
- Sport Analytics uses the tools of Mathematics, Economics and Statistics to address questions relevant for sports.
- Programming in R is the way in which we take models to the data to answer the relevant questions

The questions

- What are the drivers of team performance ? Can you buy the fans' love ?
- How can you evaluate players' talent ?
- Is the market for athletes efficient ? (Does players'compensation reflect their value? see Michael Lewis 2003 Moneyball)
- Is competitive balance relevant to determine a league success ?
- Should leagues pursue "competitive balance" in their policies ?
- Does a team need "stars" to attract the fans ?
- What is the importance of a Recreation Center for a university ?

How do we answer ?

- We answer by analyzing the data
- But analyzing the data is not a simple process

J.Ellenberg(2015) "How Not to be Wrong. The Hidden Maths in Everyday Life"

- A. Wald and the Statistical Research Group (SRG) where faced with a problem
  - to prevent planes from being shot down by enemy fighters your armor them.
  - but armour makes the plane heavier ...
  - armoring the planes too much is a problem and armoring them too little is a problem
- which is the optimum level of armoring ?

The Data

When American planes came back from engagements over Europe, they were covered with bullet holes. Here are the data

Section of the Plane	Bullet Holes per sq. f.
Engine	1.11
Fuselage	1.73
Fuel System	1.55
Rest of the plane	1.8

Data Analytics is about using the data to make decisions. So in the light of the data where do you put the armoring ?

A.Wald choice

- The armor, said Wald, does not go where the bullet holes are. It goes were the bullet holes are not: THE ENGINE
- Why? The sample is selected: it is made only **by planes which came back from engagement.** The inference is that planes with bullet holes in the engine did not make it back**.**
- Using a sample reuires thinking on how it has been selected !!!

The Effects of Hospitalization

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003

The difference in the means is 0.71, a large and highly significant contrast in favor of the *non-hospitalized*, with a t-statistic of 58.9.

Offensive Rebounds and Wins

Correlation Coefficients for Various NBA Statistics and Winning Percentage

Variable	Correlation Coefficient
Defensive rebounds	0.46
Missed field goals	–0.46
Assists	0.43
Points scored	0.41
Turnovers	–0.39
Personal fouls	–0.22
Offensive rebounds	–0.20
Steals	0.14
Blocked shots	0.11
Missed free throws	0.05

Selection Bias

Think of Hospital treatmes as defined by a binary random variable : $D_i = [0, 1]$. The outcome of interest is the health status of an indivdual Y_i.

$$\text{potential outcome} = \begin{cases} Y_{1i} & if \quad D_i = 1 \\ Y_{0i} & if \quad D_i = 0 \end{cases}$$

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) D_i$$

Selection Bias

$$\underbrace{E[Y_i|D_i=1]-E[Y_i|D_i=0]}_{\text{Observed difference in average health}} = \underbrace{E[Y_{1i}|D_i=1]-E[Y_{0i}|D_i=1]}_{\text{average treatment effect on the treated}} + \underbrace{E[Y_{0i}|D_i=1]-E[Y_{0i}|D_i=0]}_{\text{selection bias}}$$

The selection bias is the difference in the average status of health between those who were hospitalized and those who were not. The classical solution to the selection bias is to introduce treatment randomly via experiments.

Experiments

- In an experiment you divide your available sample in two groups(the **treatment group** to whom the treatment is administered and the **control group** to whom the "placebo" is administered).
- random allocation of individuals in the two subgroups avoids the selection bias

The data

- In Sports data are not generated by experiments, we have only "observational data"
- We use the data by building, estimating and simulating models
- models need to be validated to minimize the risk of using a "wrong" model
- alternatively there is the possibility of focussing on quasi-natural experiments

Modelling Strategy

- Empirical models specify the distribution of a vector of some ("endogenous") variables to "be explained" $\mathbf{y}_t$ conditional upon "explanatory" variables $\mathbf{z}_t$ that do not depend on them (i.e. are "exogenous).
- The mapping between $\mathbf{y}_t$ and $\mathbf{z}_t$ is determined by some functional relation and some unknown parameters. The unconditional density of $\mathbf{z}_t$ might or might not be specified.
- All the relevant variables are stochastic and they are therefore characterized by a density function.
- Linear Models specify conditional means of the $\mathbf{y}_t$ as linear functions of the $\mathbf{z}_t$.

Modelling Process

- the data

$$D\left(\mathbf{y}_t, \mathbf{z}_t, \mathbf{w}_t \mid \dot{}_{t-1,}\boldsymbol{\theta}\right)$$

- a general multivariate model

$$D\left(\mathbf{y}_t, \mathbf{z}_t \mid \mathbf{I}_{t-1,}, \boldsymbol{\beta}\right)$$

- decomposing a multivariate into conditional and marginal

$$D\left(\mathbf{y}_t \mid \mathbf{z}_t, \mathbf{I}_{t-1,}, \boldsymbol{\beta}_1\right) D\left(\mathbf{z}_t \mid \mathbf{I}_{t-1,}, \boldsymbol{\beta}_2\right)$$

- a general linear univariate conditional model

$$\begin{aligned} y_t &= \boldsymbol{\beta}_1' \mathbf{z}_t + u_{1t} \\ \mathbf{z}_t &= \boldsymbol{\gamma}_1' \mathbf{x}_t + \mathbf{u}_{2t} \end{aligned}$$

Assessing Model Validity

- For a model to be valid to assess the effect of a given variable z_t on another variable y_t we need that the absence of correlation between z_t and the residuals of the models used to predict y_t
- for a model to deliver precisely the effect of z_t on y_t what is left out needs to be uncorrelated to what is included in the model
- There are other ways in which the model can go wrong:
  - the model is non-linear
  - the residuals are non-normal and their variance is not constant

The Steps in the Modelling Process

- Sport Analytics uses the "available data" to predict the distribution of variables of interest. This process involves several steps:
  - Data collection and transformation
  - Graphical and descriptive data analysis
  - Model Specification
  - Model Estimation
  - Model Validation
  - Model Simulation

Selection Bias in Regression

$$
\begin{aligned}
Y_i &= Y_{0i} + (Y_{1i} - Y_{0i})\, D_i \\
Y_{0i} &= \alpha + \eta_i, (Y_{1i} - Y_{0i}) = \beta \\
Y_i &= \alpha + \beta D_i + \eta_i
\end{aligned}
$$

where D_i is correlated with η_i

find a set of instruments X_i, such that

$$\begin{aligned} E(\eta_i \mid W_i, X_i) &= E(\eta_i \mid X_i) \\ E(\eta_i \mid X_i) &= \gamma X_i \\ Y_i &= \alpha + \beta D_i + \gamma X_i + e_i \end{aligned}$$

and in the last model there is no correlation among regressors and residuals

An Illustration

- For each team and their opponents NBA box scores track the following info:
  - 1P,2P and 3P made and missed
  - offensive and defensive rebounds
  - turnovers and steals
  - blocked shots, fouls and assists

How can we use the data to pin down the driving factors of team performance ?

The Modelling Process

- Modelling takes the quantities being analyzed as random variables. An model then is a joint probability distributions for the variables of interest which is taken to be as a valid approximation to their true joint probability distribution.
- Suppose we want to build a model for the determinants of a basketball team perfomance.
- We use the number of WINS in a regular season as the measurable counterpart of performance
- We theorize that the key concept to determine peformance is how efficiently teams use **possession**
- A possession starts when one team gains control of the ball and ends when that team gives it up (in other words, an offensive rebound would start a new play, not a new possession). Possession totals are guaranteed to be approximately the same for the two teams in a game.

The Modelling Process

$$
\begin{aligned}
EP_{i,t} &= FGA_{i,t} + 0.45 * FTA_{i,t} + TOV_{i,t} - ORB_{i,t} \\
AP_{i,t} &= OTOV_{i,t} + DRB_{i,t} + TR_{i,t} + OFG_{i,t} + 0.45 * OFT_{i,t} \\
PTS_{i,t} &= 1 * FT_{i,t} + 2 * 2PFG_{i,t} + 3 * 3PFG_{i,t} \\
PTSA_{i,t} &= 1 * OFT_{i,t} + 2 * O2PFG_{i,t} + 3 * O3PFG_{i,t} \\
PTSxEP_{it} &= \frac{PTS_{i,t}}{EP_{i,t}} \\
PTSAxAP_{i,t} &= \frac{PTSA_{i,t}}{AP_{i,t}} \\
W_{it} &= \beta_0 + \beta_1 \left(PTSxEP_{it} - PTSAxAP_{i,t} \right) + u_{it} \\
u_{it} &\sim N.I.D \left(0, \sigma^2 \right)
\end{aligned}
$$

To do list

- estimate $\beta_0, \beta_1, \sigma^2$ from the data
- validate the model
- simulate the model to predict the impact on WINS, of shots, rebounds, turnovers etc ...

We are interested in the relation between Competitive Balance in a league and Attendance.
There are many determinants of attendance

$$ATT_i = \beta_0 + \beta_1 CB_i + \boldsymbol{\gamma}' X_i + u_i$$

if one runs a regression between ATT_i and CB_i only, the omitted variable problems can cause carrelation between residuals and the "treatment of interest" i.e. competitive balance

Possible solution (Szymanski 2001) find two sets of data in which competitive balance is different but all other relevant factors are the same.

$$\begin{aligned} ATT_{1,i} &= \beta_0 + \beta_1 CB_{1,i} + \gamma' X_i + u_{1i} \\ ATT_{2,i} &= \beta_0 + \beta_1 CB_{2,i} + \gamma' X_i + u_{2i} \\ ATT_{1,i} - ATT_{2,i} &= \beta_1 \left(CB_{1,i} - CB_{2,i} \right) + e_i \end{aligned}$$

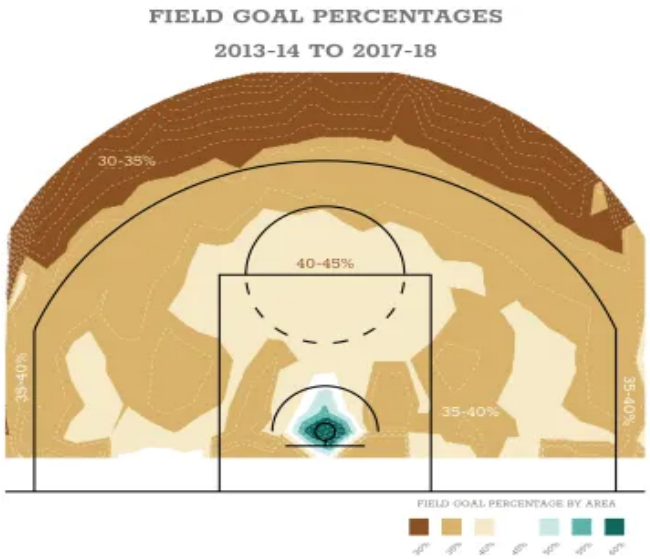
A quasi natural experiment

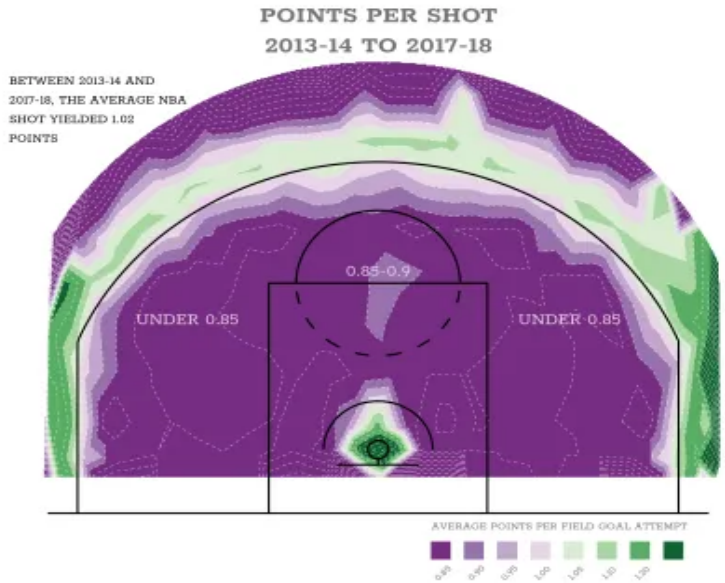
- Compare matches between the same teams in the FA cup and in the league. Under the null that the FA feature less competitive balance than the league (by its nature) and that this gap has been increasing over time
- Consider one thousand same division FA-CUP matches

Spatial Tracking

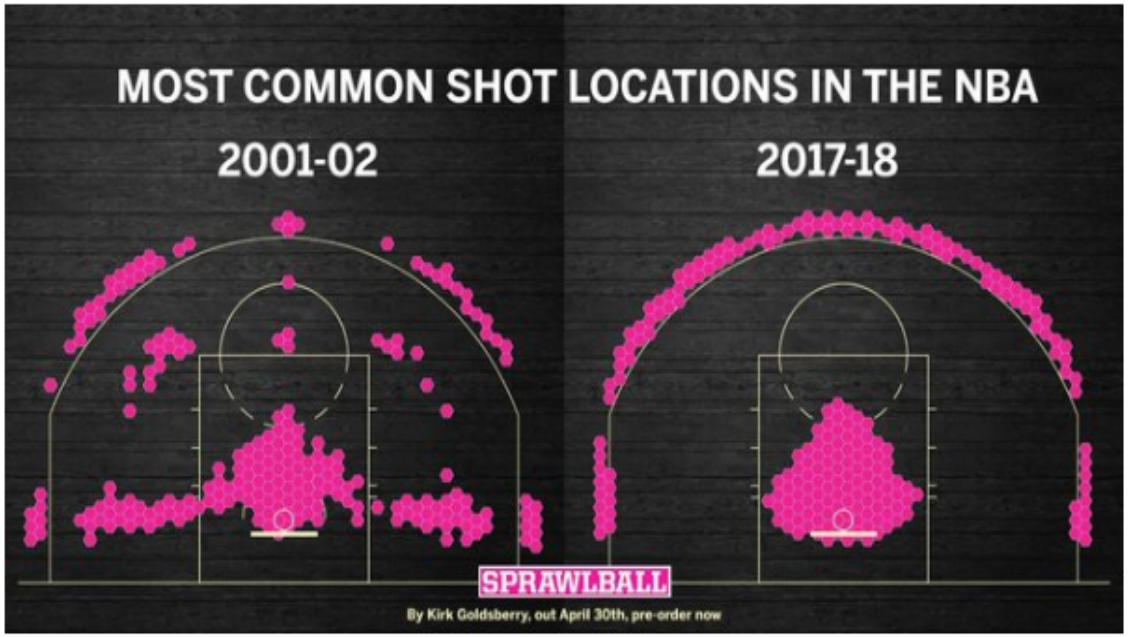
- Spatial Tracking is based on visualization of information based on big data, collected by wearable devices
- Spatial Tracking in R: the SpatialBall Package, the RShiny BallR Applications

Spatial Tracking

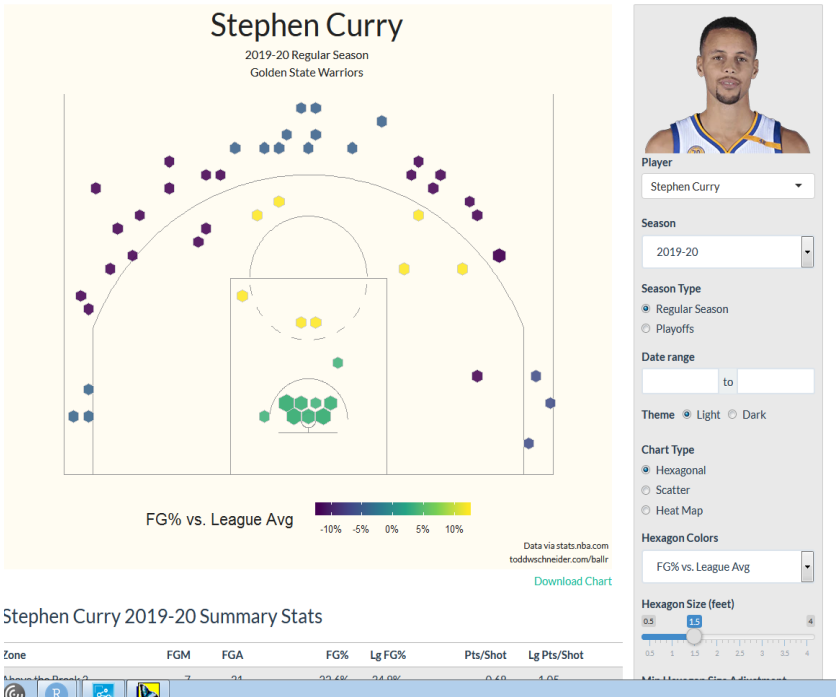




Spatial Tracking



Spatial Tracking



This course

The objective of this course is to lead students to learn the Sport Analytics by developing skills along different, but highly interrelated, dimensions:

- knowledge of the relevant data;
- knowledge of the relevant statistical methods;
- capability of implementing empirical applications (coding).

Assessment

- Students assessment will depend 50 per cent on class exercises and 50 per cent on a final exam
- Solutions to class exercises must be handed in the day before the class, on a rotation basis all students will be in charge of presenting their solution the day of the class, a general discussion will follow.
- The objective of the exam will be to evaluate the individual capability of students of using the inputs given to build the relevant output
- During the exam students will be required to modify the R codes that they have built during the course to generate answers to the questions posed in the exercises.
- Working on the exercises step by step and using all the inputs given is the best preparation strategy for the exam.
- The exams will be open books.