Sport Analytics

Carlo A. Favero,[1]

[1]Bocconi University & CEPR

20630 Lecture 1

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 analyze the data and take models to them to answer the relevant questions

The questions

- How do we identify the drivers of teams performance ?
- 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 ?
- Is there an "home advantage" effect in sport events?
- Is there a significant statistical relationship between load management and injury risk ?
- What is the effect of time-outs and are coaches calling them optimally ?

How do we answer ?

- We answer by analyzing the data
- But analyzing the data is not a simple process
- data have many dimensions one needs to select the relevant dimensions
    - build model theory-based and use them
    - select the relevant information with a target variable to predict (supervised machine learning)
    - select the relevant information in absence of a specified target (unsupervised machine learning)
    - use quasi-natural experiments
    - use statistical software to produce user friendly interfaces to analyze the data

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 ?

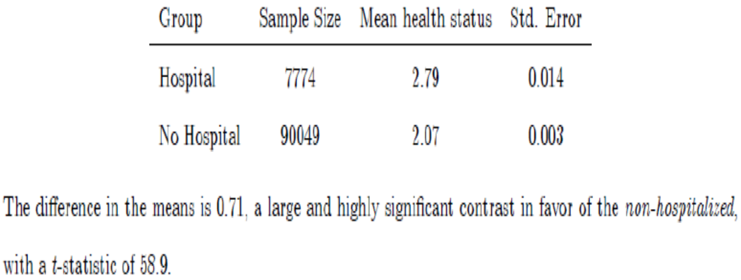
A.Wald choice

- The armor, said Wald, does not go where the bullet holes are. It goes were there are no bullet holes: **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 requires thinking on how it has been selected !!!

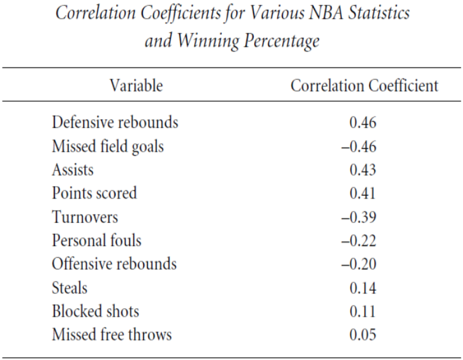
The Effects of Hospitalization



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.

Ok Planes and Hospitals, but what about Sport ? 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 treatments as defined by a binary random variable : $\mathrm{D}_i = [0, 1]$. The outcome of interest is the health status of an individual Y_i.

potential outcome $= \left\{ \begin{matrix} Y_{1i} & if & D_i = 1 \\ Y_{0i} & if & D_i = 0 \end{matrix} \right\}$

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

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"
- Data can be organized in input variables (the information we use to predict) and output variables (the targets of our predictions)
- We use the data by building, estimating and simulating models that relate input variables to target variables
- 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 ”target” $\mathbf{y}_t$ conditional upon ”input” 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.

$$\mathbf{y}_t = f\left(\mathbf{z}_t\right) + \varepsilon_t$$

Modelling Strategy

- Linear Models specify conditional means of the $\mathbf{y}_t$ as linear functions of the $\mathbf{z}_t$.

- modelling implies choosing the relevant set of $\mathbf{z}_t$ (i.e. which variables do we include in the model and which variables do we leave out) and the functional form f.

- Models can be used for predictions (using $\mathbf{z}_t$ to forecast $\mathbf{y}_t$) or simulation (what happens to $\mathbf{y}_t$ when I change some of the $\mathbf{z}_t$

- To use models for prediction and simulation the available data points are divided in two subsample: the training sample, which is used to estimate the f function , and the simulation or prediction sample in which the model is used after estimation and validated by comparing model predcitons with observed data.

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 Simulation and Validation

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 and weight them optimally to pin down the driving factors of players' and teams' performance ?

A simple, naive, approach: the PIR

- possible solution : aggregate statistics with unit weights (Performance Index Rating).

$$\begin{aligned} PIR &= PTS + REB + AST + STL + BLK \\ &\quad - FGMISS - FTMISS - TOV \end{aligned}$$

- The formula combines statistics but the weights are not convincing
    - Why does a Missed FT have the same value with a Missed FG ?
    - are an assist and a missed free throw of equal value ?

The Modelling Process

- 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

- To organize our data 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

We need to predict wins given possessions. We do so by posing a linear relationship between wins and possessions, in particolar we posit:

$$\begin{aligned} W_{it} &= \beta_0 + \beta_1 \left(PTSxEP_{it} - PTSAxAP_{i,t} \right) + u_{it} \\ u_{it} &\sim N.I.D \left(0, \sigma^2 \right) \end{aligned}$$

W_{it} : number of wins by by team i in season t
$PTSxEP_{it}$: points scored per earned possession by team i in season t
$PTSAxAP_{i,t}$: points scored by the opponents per allowed possession by team i in season t

Once we have selected a training sample to estimate $\beta_0, \beta_1, \sigma^2$, we can use the model to predict wins and simulate the impact of statistics on wins. We do so by defining precisely of statistics contribute to the definition of $PTSxEP_{it}$ and $PTSAxAP_{i,t}$

The Modelling Process

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

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 ...

Supervised Machine Learning

- In the example discussed so far ”theory” has been ised to select the ”input” variables.

- Alternatively one can start with a very general model and use statistical regularization methods to reduce the dimensionality of the model.

- Shrinkage methods fit models including many predictors but the estimated coefficients are then shrunken towards zero relative to standard estimates.

- The most widely used shrinkage methods are Ridge Regressions, Lasso and Elastic Nets

- Alternatively to linear regressions, tree-based methods based on stratyfying the predictor space in a number of simple regions can be used.

Ridge Regression

when using OLS estimates are obtained by minimizing the following quantity :

$$RSS \;=\; \sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{k}\beta_j x_{ij}\right)^2$$

in RIDGE regression estimates are obtained by minimizing the following quantity:

$$RSS+\lambda\sum_{j=1}^{k}\beta_j^2$$

Lasso Regression

Ridge regression will always generate a model involving all ten predictors. Increasing the value of λ will tend to reduce the magnitudes of the coefficients, but will not result in exclusion of any of the variables.
In LASSO regression estimates are obtained by minimizing the following quantity:

$$RSS + \lambda \sum_{j=1}^{k} |\beta_j|$$

Elastic Net

Elastic Net generalizes Lasso and Ridge regressions.
In Elastic Net regression estimates are obtained by minimizing the following quantity:

$$RSS + \lambda_1 \sum_{j=1}^{k} \beta_j^2 + \lambda_2 \sum_{j=1}^{k} |\beta_j|$$

Tree-Based Methods

- tree-based methods involve segmenting the predictor space into a number of simple regions. A prediction for a given observation is then the mean or the mode of the training observations in the region to which it belongs.
- the set of splitting rules used to segment the predictor space can be summarized in a tree,
- Tree-based methods are simple but are not competitive with the best supervised learning regularization approaches in terms of prediction accuracy.
- Hence bagging, random forests, and boosting are introduced to produce multiple trees which are then combined to yield a single consensus prediction.

Regression Trees: An Example

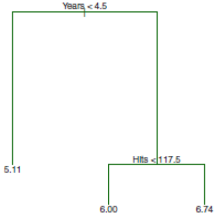


FIGURE 8.1. *For the* Hitters *data, a regression tree for predicting the log salary of a baseball player, based on the number of years that he has played in the major leagues and the number of hits that he made in the previous year. At a given internal node, the label (of the form $X_j < t_k$) indicates the left-hand branch emanating from that split, and the right-hand branch corresponds to $X_j \geq t_k$. For instance, the split at the top of the tree results in two large branches. The left-hand branch corresponds to* Years<4.5*, and the right-hand branch corresponds to* Years>=4.5*. The tree has two internal nodes and three terminal nodes, or leaves. The number in each leaf is the mean of the response for the observations that fall there.*

Regression Trees: An Example

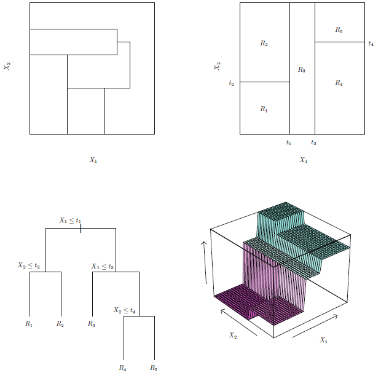


FIGURE 8.3. Top Left: *A partition of two-dimensional feature space that could not result from recursive binary splitting.* Top Right: *The output of recursive binary splitting on a two-dimensional example.* Bottom Left: *A tree corresponding to the partition in the top right panel.* Bottom Right: *A perspective plot of the prediction surface corresponding to that tree.*

Unsupervised Machine Learning

- Unsupervised learning is aimed at extracting information by a given set of features $X_1, X_2, ...X_p$ without relating them to an associated response variable Y.

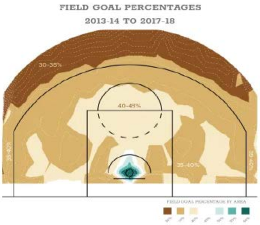
- Think for example of the objective of grouping players according to their observed characteristics

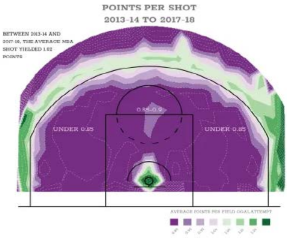
- the most popular unsupervised machine learning tools are
  - *principal component analysis*, common used for data visualization and pre-processing before supervised techniques are applied
  - *clustering*, a class of methods for discovering unknown subgroups in the data. Among clustering methods we distinguish between K-Means Clustering and Hierchical Clustering

Spatial Tracking

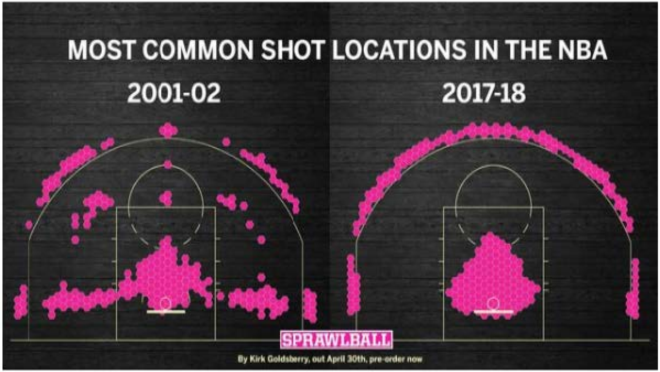
- Spatial Tracking is based on visualization of information based on big data, collected by wearable devices
- Spatial Tracking in R: the SpatialBall Package, the RShiny BallR Applications

Spatial Tracking

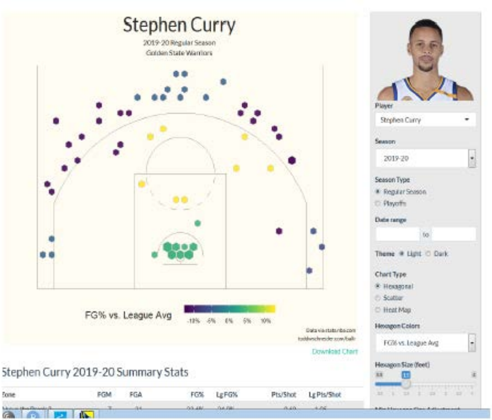




Spatial Tracking



Spatial Tracking



Stephen Curry 2019-20 Summary Stats

Zone	FGM	FGA	FG%	Lg FG%	Pts/Shot	Lg Pts/Shot

Player: Stephen Curry

Season: 2019-20

Season Type
● Regular Season
○ Playoffs

Date range: to

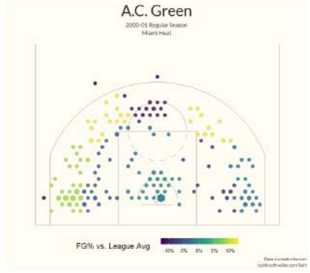
Theme ● Light ○ Dark

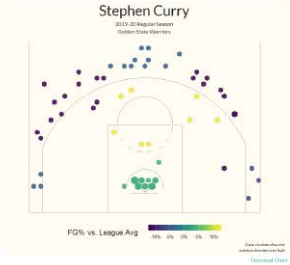
Chart Type
● Hexagonal
○ Scatter
○ Heat Map

Hexagon Colors: FG% vs. League Avg

Hexagon Size (feet)

Spatial Tracking: The Game Changer





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

- Assessment for attending students will depend 50 per cent on group projects and 50 per cent on a final exam
- Assessment for non-attending students will be entirely based on the final exam
- Projects will be offered to students grouped in several groups. Each group will give class presentations.
- 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.

Project 1:Scraping sports data from the web with R

- The objective of this project is to illustrate how data on sports could be efficiently retrieved from the Web.
- Students should feel free to choose their preferred field and application.

Project 2:Creating Web Applications with Rshiny

- The objective of this project is to create a Sport related web application with RShiny.
- An Illustration based on NBA data is provided together with projects produced in 2020.
- Students should feel free to choose their preferred field and application.

Project 3:An Application of Unsupervised Machine Learning to Sport Analytics

- The objective of this project is to apply unsupervised machine learning, and in particular cluster analysis, to finding groups in Sport Analytics data.

Project 4:An Application of Supervised Machine Learning to Sport Analytics

- The objective of this project is to apply supervised machine learning techniques , and in particular techniques to solve the many predictor problem to predict top athletes compensations.
- Students should use as a benchmark the model presented in the lectures and evaluate it against alternatives generated by modern machine learning techniques.

Project 5:Evaluating the Home Advantage Effect from quasi-Natural Experiments

- Following the COVID shock many games in many sport were played without attendance within ”bubbles” in which no team had the ”home advantage effect”. The objective of this project is to use sport data to construct a quasi-natural experiment for the evaluation of the Home Advantage Effect.
- Stock J. and M.Watson (2020) Introduction to Econometrics, 4th edition, Chapter 13

Project 6:Measuring Competitive Balance and its effects

- The objective of this project is to introduce, discuss the concept of Competitive Balance in the Sport Industry.
- Both a discussion of the theory and applications are possible.

Project 7: Load Management and Injury Risk

- A recent report denied the existence of a significant statistical relationship between load management and injury risk in the NBA. The objective of this project is a critical analysis of the report, which will be made available to the groups taking this choice.

- $https://www.espn.com/nba/story/_/id/39288379/nba-report-no-link-load-management-less-injury-risk$

Project 8: The Relevance of Popular Shareholding Contribution to Team Perfomance

- A recent report provided evidence on the popular shareholding contribution to team perfomance in European soccer. The objective of this project is a critical analysis of the report, which will be made available together with the original data to the groups taking this choice

Project 9: Statistical Analysis of time-outs in Basketball

- The practice of coaches calling timeouts to stop runs has been a subject of analysis in basketball discussions.

- Play-by-Play (PBP) datafrom 2019–20 to 2023–24 from the NBA has been utilized to determine whether there is any evidence supporting that some coaches have suboptimal timeout strategies.

- The objective of this project is the extension of the evidence to European Leagues, in particular the Euroleague.

- *https :*
  //medium.com/@ivm9816/analyzing−nba−timeouts−29df987f076a

External Presentation

- Using Video Analysis for a Euroleague Basketball Team, presentation by Mario Fioretti, Assistant Coach, Olimpia Milano

External Presentation

- Using Analytics in the European Soccer Industry, presentation by Mark Nervegna, Head of Strategy and Analytics, Raiola Global