Sport Analytics. An Introduction

Carlo Favero

Course 20630 Year 2018/19

Sport Analytics

- Sport Analytics is the statistical analysis of economics data.
- How to use data-driven decision making processes to:
  - evaluate the drivers of team performance
  - evaluate players talent
  - assess market efficiency (players'compensation vs players value. Michael Lewis's 2003 *Moneyball)*

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

- 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

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

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

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

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

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

Our Case

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

To do list

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

Why a model can be wrong ?

There are many ways in which the model can go wrong:

- other factors beyond those explicitly considered are relevant in determining WINS
- the model is non-linear
- the residuals are non-normal and their variance is not constant
- Teams are different and Seasons are different

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.