Sport Analytics

Carlo A. Favero,[1]

[1]Bocconi University & CEPR

20630 Lecture 3

The NBA teams database

Our first database is made of 43 seasons (from 1979-1980 to 2022-2023) for all NBA teams.

- the original data for each season are obtained from the following link: https://www.basketball-reference.com/leagues/$NBA\_$"$season$".$html$ where "$season$" identify the season, for example to get the data for 2023_2024 you should specify "$season$" = 2024

- three excel files are obtained for each season with the names team_xx.xlsx, opp_xx.xlsx and misc_xx.xlsx where xx=1...k (k is the (final) year of seasons)

- All the 43x3 files are combined into a .csv file named **teams_overall2023.csv** by the R code **dataset2023.R**

Building the database

Here is the manual procedure to download data from basketball reference

- go to www.basketball-reference.com and select season by season the summary page.
- scroll down the screen to reach the tables called "Team Stats", "Opponent Stats" and "Miscellaneous Stats"
- Save them as Excel files with the names team_xx.xlsx, opp_xx.xlsx and misc_xx.xlsx where xx=1...k (k is the number of seasons)
- check the data format and make sure it is the same across seasons
- the data base can be updated year by year just by adding three files for the new season
- run the R programme **dataset.R** that combines all data and produces a csv file called Teams_overallxxxx.csv, xxxx is the last season

Building the database, webscraping Basketball References

As an alternative to the "manual procedure"

- adapt the code *update_db.R* to select the appropriate webpage in basketball reference
- build three data files with the content of the tables "Team Stats", "Opponent Stats" and "Miscellaneous Stats"
- check the data format and make sure it is the same across seasons
- Save them as Excel files with the names *data_team_xx.xlsx*, *data_opp_xx.xlsx* and *data_misc_xx.xlsx* where xx defines the season
- update the database by running the R programme **dataset2023.R** that combines all data and produces a csv file called *Teams_overallxxxx.csv* where xxxx is the last season in the database

Building the database, accessing Basketball Reference via API

- The most efficient way to access data on the web is via API (Application Programming Interface)
- APIs allow machines to access data programmatically – that is automatically from within a program
- See the intro on accessing API with R on the webpage
- Basketball reference via API: https://github.com/rtelmore/ballr

The relevant dimensions of the data

- There are two relevant dimensions in our data set
  - cross-section (in each year we observed data for all the different teams)
  - time-series (for each team we have 40 seasons of data)
- In general, we shall define $X_{i,t}$ as the statistics observed at time t for team i.
  - the t index captures the time-series dimension
  - the i index captures the cross-sectional dimension

Data Transformation

- After importing the data in the statistical package, the first step in the analysis is data transformation and organization.
- In R data are imported in a data-frame
- We can use the data-frame features to transform the data and organize them (for example, take subsets or sort them)

Descriptive Analysis

- Descriptive analysis can be univariate or multivariate
- Analysis of the marginal distribution of a variable
- Correlation analysis

Graphics

- scatter-plots
- time-series graphics
- multiple graphs
- density estimates (histograms)

QQ-plot

The idea is to plot in a standard Cartesian reference graph:

- the quantiles of the series under consideration, X_t,against the quantiles of any given distribution.If the returns were truly normal, then the graph should look like a straight line with a 45-degree angle.
  - first, sort all (standardized) returns in ascending order, and call the ith sorted value x_i;
  - second, compute the empirical probability of getting a value below the actual as $(i - 0.5)/T$, where T is number of observations available in the sample.
  - Finally, we calculate the quantiles of the benchmark distribution quantiles as $\Phi^{-1}\left((i - 0.5)/T\right)$, where $\Phi^{-1}(\cdot)$ denotes the inverse of the benchmark density.
  - Represent on a scatter plot the (standardized) returns and sort the data on the Y-axis against the standard distribution quantiles on the X-axis.

Matrix Representation of the data

A matrix is a double array of i rows and j columns, whose generic element can be written as a_{ij}, it is a convenient way of collecting simultaneusly information on the time-series and the cross-section of returns:

$$A = \begin{bmatrix} a_{11} & . & . & a_{1j} \\ & & & \\ & & & \\ a_{i1} & & & a_{ij} \end{bmatrix}, 0 = \begin{bmatrix} 0 & . & . & 0 \\ & & & \\ & & & \\ 0 & & & 0 \end{bmatrix}$$

$$I = \begin{bmatrix} 1 & . & . & 0 \\ & & & \\ & & & \\ 0 & & & 1 \end{bmatrix}$$

Matrix Operations

- Transposition $a'_{ij} = a_{ji}$
- Addition: For A and B nxm $(a+b)_{ij} = a_{ij} + b_{ij}$
- Multiplication: For A nxm and B mxp $(ab)_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$
- Inversion for non-singular A nxn, A^{-1} satisfies $A^{-1}A = AA^{-1} = I$

An Illustration with NBA data

- Construct a time-series plot of GSW pace over the seasons
- Did the share of three points/field courts shots taken by Chicago Bulls increase over time ?
- Did the relative efficiency of three points and two points taken by Chicago Bulls increase over time ?