Descriptive and Graphical Analysis of the Data

Carlo Favero

The first database

Our first database is made of 39 seasons (from 1979-1980 to 2017-2018) for all NBA teams, we obtain it at the following link:
https://www.basketball-reference.com/leagues/NBA_2018.html#all_opponent-stats-base
A version of the database is available from the course website and it is named **teams_overall.csv**

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 39 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 $A^{-1}A = AA^{-1} = I$

Running a model with R

- **Problem** : What is the probability that a Great NBA Team (Win% = 85%) Looses Two Consecutive Games at Some Point in the Season ?
- **Solution**: Simulate the binomial distribution