

INTRODUCTION TO SPORT ANALYTICS (20630)

Exam Mock 2026

Allowed time: 75 minutes

Q1 The file `Teams_overall12026.csv` contains NBA data as in the lectures.

1. Clear the environment, set the correct default directory, import the data in R, install and run all the relevant packages for the code.
2. Check the data and their type. Indicate precisely the size of cross-sectional and the time-series dimensions of the available data in panel format. Remove all NA from the dataset.

Q2

1. Build a subset of the database that contains data for all the NBA teams from season 2010 to 2024, excluding seasons 2019-2020 and 2020-2021.
2. Provide a scatterplot of ORB on teams' win percentage and assess the correlation among these two variables. Comment the results.

Q3

1. Describe the four factor model
2. Construct the four factors using the data in your database.
3. Estimate the four factor model on data from season 2023-2024, and use data transformation so that you can get from your estimated coefficients an estimate of the number of wins of the average team in the league.
4. Comment the regression coefficients
5. Which statistical tools have you seen in class to perform hypothesis testing? Test the hypothesis that the coefficients on F1 and F2 are equal in magnitude but have opposite sign.
6. Indicate the partial R2 associated to F1.
7. Indicate the teams with the minimum and the maximum residual from the regression.

Q4 The file `Players04-05_ws.csv` contains NBA players' data as in the lectures.

1. Compute a measure of player efficiency based on the WINSORE model.
2. Compare and make a scatterplot of the actual wins of a team against the fitted number of wins obtained summing players' contributions to wins.
3. Use a regression model to project salaries on the efficiency measure computed in the point above. Comment the results.