Alvaro Hu

Springboard

August 28, 2018

Inferential Statistics

In this portion of my project, I just wanted to get some certainty on some of the stats that I obtained through my initial exploratory data analysis. Some of these are the means of some per-game stats, the pearson correlation coefficients of some of the stats, the differences between means of two different groups in the data (like teams or positions), and using principal component analysis to check how many dimension I can reduce my problem to.

To start off, I wanted to see how sound the correlations between different stats and the player's salary were. To do this, I took 10% of the data and shuffled it, and then obtained the pearson correlation coefficient from this mostly unshuffled data. I reran this 10,000 times and checked to see how many times I achieved a correlation coefficient as strong as the one I got from the data. For points per game, win shares, rebound per game, and minutes per game, I deduced that, at a 95% confidence interval, the correlations were those that I achieved from the empirical data. I achieved p-values of 0.004, 0.015, and 0.0328 for points per game, win shares, and rebound per game, respectively. To affirm my findings, I did a Mann Whitney U test for each, since my data is non-normal. These tests confirmed that my correlation coefficients were statistically significant.

The next thing I did was confirm whether or not several means were actually different from one another, and not just because of a random sample. My first test was between the two teams with the highest average salaries; The Golden State Warriors and the Oklahoma City Thunder. I calculated the mean difference between the two and stored the value. I then wrote a function that would take the two teams, shuffle the players up, and then redistribute the players

onto the teams, and find the mean difference between these shuffled teams. I repeated this 10,000 times to see the probability of getting a mean difference as extreme as my empirical value. I got a p-value of 0.466, which means I can not rule out that there is no statistical difference between the average salaries of the two teams. I repeated this process for the highest and lowest average salaried teams; The Warriors and the Kings. This time, there was a much larger difference, and I got a p-value of 0.0265. Meaning that I *can* rule out that there is no difference between the average salaries of these two teams. Finally, I did the test for the Warriors and the rest of the NBA, yielding a p-value of 0.13. I confirmed these numbers by using a t-test module from scipy.stats and got relatively the same numbers.

Another thing I wanted to check was whether or not means between two different positions differed. I took the highest salaried position, Centers, and the lowest salaried position, Shooting Guards, and ran this test with them. For this test, I for a p-value of 0.3, which is not nearly low enough to rule out that there is a difference between average salary and position.

The next thing I did was PCA and found that the intrinsic dimensionality of my problem is only 1, which makes sense since game statistics are all intercorrelated based on how good a player is. If a player has high “basketball IQ”, then chances are he scores more, rebounds more, gets more assists, and plays more games/minutes. I ran an OLS regression with 2 variables, points per game and Games played, since the rest were too correlated with one another, and obtained an r-squared value of 0.413 and p-values of 0.00 and 0.032 for each variable respectively. I ran the OLS again, slowly subtracting a feature with a high p-value each time until all the p-values were relatively low, and ended up with the variables G, AST, PS/G, WS, and AST%. The r-squared for this regression is 0.504, which is slightly higher than the first.

From my inferential statistical analysis, I have concluded that Games Played, Points per Game and Win Shares are indeed good indicators of how much a player’s salary is, but Total

Rebounds is not. I also concluded with a 95% certainty that there is no difference between the salaries of 2 positions, and also that there is a difference between the salaries between two teams, but only on opposite sides of the spectrum, and not between every team.