National University of Singapore

Report Assignment 2

NBA Data Analysis

Clustering, Linear Regression, Panel Data

Emery Ong A0136591B
Emile Brès A0132365L
Simon Helmlinger A0134470M
Juan Manuel Muñoz Perez A0134739X

Monday 30th March, 2015

Clustering. Which players are similar? (15 points)

Introduction Using the stats library in R, the purpose of this part is to determine the 'closest' players thanks to the kmeans algorithm. We should not forget the main idea of the NBA analysis started in the assignment 1: *explain the factors that influence the player's salary.* After defining 'closest', we will explain the approach used in order to conduct the analysis.

In the context of NBA players, two players are close if their statistics (weight, height, age, experience and its games statistics) are similar. We aim to cluster the current NBA active players in order to understand the characteristics of the differents groups and how it influence their salary.

Data The first step to conduct the analysis is to build the dataset. To do that, we used the data scraped during the assignment 1 as follows:

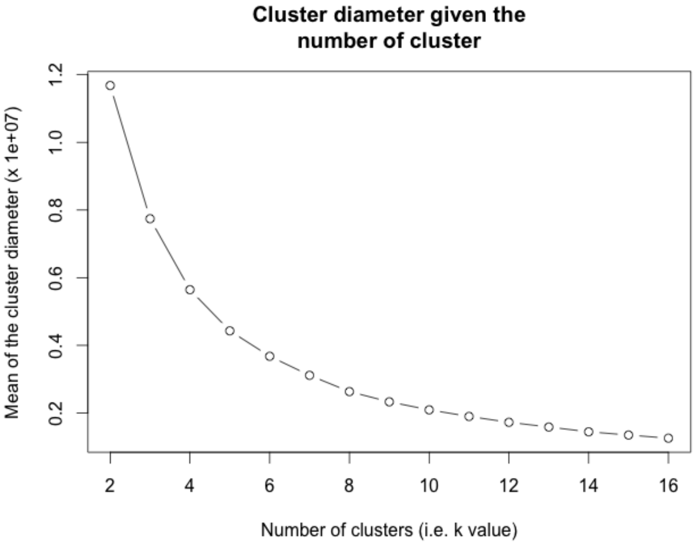
- Filtering
  1. extract the profile of the active players into the active_player_profile dataframe (attributes: PlayerID, name, shoots, weight, height, dob, birth_city, birth_state, experience and age)
  2. extract the most recent salary recorded for the active players into the active_salaries dataframe (attributes: PlayerID, Season, Team, FranchiseID and Salary)
  3. extract the totals statistics for the current active players into the active_totals_final dataframe (attributes: PlayerID, Season, Age, FranchiseID, Lg, Pos, G, GS, MP, FG, FGA, FG%, X3P, X3PA, X3P%, X2P, X2PA, X2P%, eFG%, FT, FTA, FT%, ORB, DRB, TRB, AST, STL, BLK, TOV, PF and PTS)
- Merging
  1. merge active_player_profile with active_salaries into the player_information_inter dataframe
  2. merge player_information_inter with active_totals_final into the player_information dataframe

The player_information dataframe contains 600 active players but at the end the kmeans algorithm is applied to only 531 active players. Indeed, the dataset is build so that for each active player we keep its last salary recorded - note that the salary of the current season can be missing - with its corresponding totals statistics for the same team and season - some players change team during the season and so have two records in salaries and teams for the same season.

Attributes chosen to explain the player's salary The salary is influenced directly by the experience of the player and by all the statistical attributes that gather data about the player's performance (see active_totals_final for the list).

kmeans algorithm The main issue in dealing with the kmeans algorithm is the difficulty in finding the optimal number of centrois (k). In order to find the better parameter we use the cluster-diameter mean analysis. The figure below show the result for our dataset.

From that figure we cannot get directly the optimal k without being sure of the chosen k. Then, we proceed to a deeper analysis. The idea is to compute the divergence of the slope from the kth to the k+1th clusters. The table below shows the results computed.



k	diameter	slope	variation (%)	divergence (%)	diff_div (%)
2	11680914	0.00	0.000000	0.00000	0.00000000
3	7742893	-3938021.36	-Inf	0.00000	0.00000000
4	5644795	-2098098.18	46.722021	46.72202	46.72202131
5	4429310	-1215484.24	42.067333	69.13465	22.41262459
6	3677566	-751744.45	38.152678	80.91061	11.77595914
7	3113996	-563570.26	25.031670	85.68900	4.77839436
8	2633222	-480773.84	14.691410	87.79149	2.10248776
9	2330790	-302431.24	37.094905	92.32022	4.52873621
10	2093675	-237115.06	21.597036	93.97883	1.65860413
11	1899240	-194435.25	17.999620	95.06262	1.08378818

The `variation` is the % of variation between the `k`$^{\text{th}}$ and the `k-1`$^{\text{th}}$ slopes. The `divergence` is the variance of each slope from the first one ($k = 3$). Finally, the `diff_div` is the variation between the `k`$^{\text{th}}$ and the `k-1`$^{\text{th}}$ of the `divergence`.

Theses numbers conduct us to choose $k = 6$. Indeed, for that value, the slope is up to 81% different from the first slope. For $k = 7$, the difference is of 85.7%. Thus, the variance of the `divergence` start becoming insignificant (about 4.7% compare to previous which are about > 12%).

Cluster centroids With $k = 6$ and the previous attributes chosen, the kmeans algorithm output the same centers (it is 'stable'). The data below shows the difference from the mean.

```
> km_final$centers
       Salary     weight      height experience        age         G
1 -1796384.28 -0.3256596 -2.92232958  0.1824454 -0.1335164  1.199887
2 -3013671.09  0.1906397 -0.06971589 -1.9161103 -0.9032700 -7.110760
3  2944608.94  0.1305320 -1.79908192  2.2301966  1.5640596  8.234934
4  7231951.43 -1.7428654  6.30508475  2.4661741  1.2319281  4.138780
5 13826170.47  0.4130320  8.31841808  3.1420716  1.3221846  7.864934
6    67988.91  0.5091240 -0.13169686  0.7296578  0.2513800  5.433210
          GS         MP         FG         FGA        X3P       X3PA
1 -1.6590457  -17.75074  -8.767042   -17.67844  -2.952233  -5.719221
2 -7.9715261 -311.13325 -57.616928  -123.64942 -11.010633 -30.181503
3  8.1495115  367.00992  64.626942   142.05535  21.107639  54.038518
```

```
4 16.5513545  370.29958  75.619731   150.50847   0.991453   4.316964
5 24.8426365  643.31804 172.798192   359.58847  18.648889  51.924143
6  0.6371193  148.42425  18.100721    43.48549   8.750958  22.930580
         X2P      X2PA         FT        FTA         ORB          PF
1  -5.814809 -11.95922  -7.560660  -8.029586   0.0530475   0.9898798
2 -46.606295 -93.46792 -28.379378 -35.958352 -13.8642266 -22.2423341
3  43.519303  88.01683  30.527395  35.792844  10.9232286  23.7870468
4  74.628278 146.19151  36.672027  51.408228  25.2361292  23.0502680
5 154.149303 307.66433 119.181770 148.632844  32.4751036  32.0564218
6   9.349763  20.55491   4.330736   4.172154   4.5606208  15.1722839
         DRB        TRB       AST         STL      BLK       TOV        PTS
1  -4.401007  -4.347959  -3.01709  -0.8676804 -1.08407  -1.97136  -28.04698
2 -45.046135 -58.910361 -33.95444 -10.1036953 -5.92591 -19.98455 -154.62387
3  42.724723  53.647952  30.13609   9.9164901  4.22345  17.47066  180.88892
4  72.749964  97.986093  60.06678  14.7362017 15.70983  29.32362  188.90294
5 121.165348 153.640452 108.44422  21.9577401 14.18470  62.22003  483.42704
6  14.767647  19.328268   5.02169   5.1287746  1.34378   6.55704   49.28314

km_final$size
[1]  107  209  64  39  25  87
```

Interpretation By analyzing the centers coordinates, we can derive the caracteristics of the different player clusters. We can assume the clusters are as follows: **rookies** (cluster 2), **intermediate experienced** (cluster 1), **advanced experienced** (cluster 6), **seniors 2P** (cluster 4), **seniors 3P** (cluster 3) and **all-stars** (cluster 5). The experience is the attribute that influence the most the player salary. The more experienced a player is, the higher salary he earns - senior players are above 3 millions. At the opposite, the rookies are under 3 millions the mean salary. Note also that the age is strongly correlated with experience. The more experienced a player is, the higher the probability to be older and finally the higher salary he earns. Let's understand deeper the differences between clusters.

Rookies - 209 players Since rookies have played lesser games than other players, their statistics are lower - the centroid's coordinates are all below the mean. But be careful, that doesn't mean the players are bad. It just translates a lack of experience compared to experienced and seniors players. Players is this category can be very promising.

Intermediate experienced - 109 players This category corresponds to players with some experience in NBA (equals to the mean). This is by no doubt the category of the worst players since it concerns the players with experience but with statistics below the mean. One important thing is the `height` which is the lower - of $3cm$ from the mean - between the clusters. Since basket-ball uses to be a sport with tall players, we can induce that this attribute may influence the salary (< 1.7 million). We should be aware of this result. Indeed, Tony Parker is a 'small' player but still earns more than 12.5 million USD. This cluster groups the worst players (lower statistics).

Advanced experienced - 87 players This category group the players with some experience in NBA (a little higher than the mean) but who do not separate from the crowd. Their statistics shows they are close to the mean - including the salary.

Seniors 2P and Seniors 3P - 39 and 64 players These two categories are quite complementary. What distinguish the most these clusters are the difference in `height` of the players (about $8cm$) and the salary. That difference influence also their statistics. Indeed, seniors 2P have higher statistics in X2P, X2PA, DRB TRB, AST, STL and TOV than seniors 3P. Since seniors 2P are taller, we can assume they tend to be positionned under the basket, then are prone to score more 2 points. In the opposite, seniors 3P, smaller, are prone to score more 3 points and their position requires less defensive than seniors 2P.

All-stars - 25 players This last category contains the best players of the current NBA season. As expected, they have the higher statistics in all the attributes making them the most experienced and talented NBA players. They obviously are the best paid players (13.8 million above the mean).

Linear regression models. What factors predict a player's salary in the 2011-2012 season? (20 points)

Predictor variables We build 3 models for the linear regression:

- one with the usual in-game statistics: points per game PTS, assists per game AST, rebound per game TRB, turnovers TOV, blocks BLK, stealsSTL;

```
lm(Salary ~ PTS+AST+TRB+BLK+TOV+STL, data = per_game_salaries)
```

- one with the statistics regarding the usage of players: number of games played G, games started GS, minutes played per game MP;

```
lm(Salary ~ G+GS+MP, data = per_game_salaries)
```

- one with the statistics regarding the shooting accuracy of players: field goal percentage FG%, 3-points-percentage 3PA%, free throw percentage FT%, multiplied by the number of minutes played to account for the importance of players in a team.

```
lm(Salary ~ X3P.*MP+X2P.*MP+FT.*MP+FG.*MP, data = per_game_salaries)
```

Preditor explanation We tried to evaluate the impact of a player on the field as, excluding off-field factors, this determines the value of a player and hence his salary. To do that, we chose variables that are commonly used to assess the performance of a player by the coaches and the players themselves as well as by the sports commentators. Let's analyse the results of the linear regression.

Model 1 The points, assists and rebounds per game are significant variables, which are positively correlated with the salary. There is also a significant negative offset.

```
lm(formula = Salary ~ PTS+AST+TRB+BLK+TOV+STL, data = per_game_salaries)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -697236     318395  -2.190  0.02899 *
PTS           328918      51412   6.398 3.59e-10 ***
AST           486670     164995   2.950  0.00333 **
TRB           563932     112552   5.010 7.51e-07 ***
BLK           281477     458148   0.614  0.53924
TOV             6553     472058   0.014  0.98893
STL          -900142     528883  -1.702  0.08937 .
Multiple R-squared:  0.4623,        Adjusted R-squared:  0.4559
```

Model 2 The number of games started, games played and minutes per game are all significant variables, positively correlated for the games started and minutes played and negatively for the number of games played.

```
lm(formula = Salary ~ G + GS + MP, data = per_game_salaries)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -787854     485370  -1.623  0.10516
G             -44112       9994  -4.414 1.24e-05 ***
GS             30220      11165   2.707  0.00702 **
MP            308776      27763  11.122  < 2e-16 ***
Multiple R-squared:  0.4181,        Adjusted R-squared:  0.4147
```

Model 3 The field goal percentage as well as the field goal percentage times the minutes played are significant variables.

```
lm(formula = Salary ~ X3P. * MP + X2P. * MP + FT. * MP + FG. *
    MP, data = per_game_salaries)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   7296398    3662965   1.992 0.047042 *
X3P.          1150100    3585118   0.321 0.748527
MP            -281849     203865  -1.383 0.167565
X2P.         17803703    9854024   1.807 0.071537 .
FT.          -5118364    3106271  -1.648 0.100172
FG.         -34587121   10796146  -3.204 0.001463 **
X3P.:MP        -71879     173444  -0.414 0.678781
MP:X2P.      -1288671     602168  -2.140 0.032942 *
MP:FT.         265472     171688   1.546 0.122818
MP:FG.        2379415     625467   3.804 0.000164 ***
Multiple R-squared:  0.4216,        Adjusted R-squared:  0.4089
```

We can see that the model with the highest adjusted R-squared value is the model 1.

Interpretation Based on the results for each model, we can make a few interpretations. First the number of points, assists and rebounds per game are the best predictors of the salaries. Teams are ready to pay higher salaries to players who can score more points, give more assists and collect more rebounds. These results were expected. We also see that the number of games started and minutes played are positively correlated with the salaries. Surprisingly, the number of games played is negatively correlated with the salaries. To explain this, we can say that players with very high salaries play more games and thus are more prone to injuries. Since their salaries are guaranteed, there are cases of highly-paid players who do not play in a whole season because of grave injuries such as torn ligaments in the knee. Using model 3, we can see that field goal percentage times minutes played is positively correlated with salary (which we expected) but that field goal percentage is negatively correlated. To interpret that, we say that role players, who typically play very few minutes and have low salaries, have a field goal percentage comparable to star players (in part because they play mostly when games are already decided and against low-class competition). You are much more valuable to your team if you can sustain a good field-goal percentage for long minutes.

Advice from the team coach A coach can use these results to tell its players on what area of their games they should focus on. The total salary of a player can indeed be divided between his performance in points, rebounds and assists. For example, 10% of the salary of a player could be because of his points performance, 40% because of his assists and 50% because of its rebounds. If this player neglects his rebounding game to try to score more points, the coach could tell him to focus on the rebounds, because that is what he is paid for.

Pertinence of the results I think the results are valid but that the models are too simple to effectively reflect the performance of the players. For example, to obtain a finer analysis, we could have used the clusters found in question 1 and make a linear regression on each of those clusters. We could also have taken into account the position of each player. Here we make the linear regression on very different groups of players (stars and roles players for example) which can prove to be problematic to analyse the results. There are also off-field factors who influence a player salary. For example, the contract structure of the NBA limits the salaries of new players in the league, which is one of the reasons we observe a strong positive correlation between salary and experience in the league.

Panel data. What predicts the players' salaries in the past 10 years? (20 points)

In this question, we used our data on player performances during the past ten years to better understand the factors that determine players' salaries. Classical linear regression techniques cannot necessarily take advantage of time-dependent variables measured on different entities. Therefore, a panel regression is more suited for this problem.

Time-varying predictors It is quite difficult to know in real life exactly how many variables are correlated, and if they are, how important will be their bias on the regression result. Therefore, our first idea is to keep only the following information about players for each season: age, position, games played, points scored, 3-points average, number of assists, number of steals, total rebounds, Win-Loss percentage of team during season; and some dummy variables to distinguish players that played exceptionally well during each season: championship winner member, championship runner-up member and best points, rebounds, assists, win-shares performance during the whole championship.

Time-invariant predictors All these variables are time-invariant. Indeed, as we saw in question 2, height and weight have little correlation with players' salaries; therefore we decided not to include them in this panel data. Moreover, fixed effect model couldn't take advantage of these variables. The dummy variables we introduced aim at separating exceptional players from others, since we think that salaries do not necessarily depend linearly on performances but also on player's fame, which can be less rational.

Regression The data we use consists in players that have been active in the past 10 seasons (i.e. from 2004-2005 to 2014-2015) for at least 8 seasons. This limitation aims at reducing the bias resulting from unbalanced data. After removing players who do not match these criteria, we have still 202 players for 1907 observations. The result of the panel regression is shown here:

```
# Panel regression
Coefficients :
                             Estimate  Std. Error t-value  Pr(>|t|)
Age                     409870.87         30376.82 13.4929 < 2.2e-16 ***
G                           -33248.53            6448.52 -5.1560  2.82e-07 ***
PTS                           3003.38             486.98  6.1674  8.67e-10 ***
X3P                           4709.47           3256.09  1.4464 0.1482633
```

```
AST                           3642.69           1526.60  2.3861 0.0171358 *
STL                          -9675.03           6328.84 -1.5287 0.1265214
PosC                    863717.09  2109942.06  0.4094 0.6823305
PosPF                   784927.66  2074082.68  0.3784 0.7051471
PosPG                  1506901.02  2027284.45  0.7433 0.4573975
PosSF                   138135.30  2009546.42  0.0687 0.9452051
PosSG                  1390703.09  2002512.41  0.6945 0.4874774
WL_percentage         -1391715.89   592015.18 -2.3508 0.0188471 *
Champion               -194477.12   465851.92 -0.4175 0.6763911
RunnerUp                     11731.55    415405.61  0.0282 0.9774731
TopPerformerPoints      831576.87  1193764.94  0.6966 0.4861490
TopPerformerAssists   -1695243.10  1259194.51 -1.3463 0.1783893
TopPerformerRebounds  -1306341.45  1245226.75 -1.0491 0.2942919
TopPerformerWinShares  4941261.14  1459454.98  3.3857 0.0007263 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Total Sum of Squares:        2.1218e+16
Residual Sum of Squares: 1.784e+16
R-Squared      :  0.15918
         Adj. R-Squared :  0.14081
F-statistic: 17.7426 on 18 and 1687 DF, p-value: < 2.22e-16
```

We understand that this is an unbalanced panel due to players having their debut seasons on different seasons in the past decade. Nevertheless, we see some unwanted results here: the more a player plays (`G` coefficient), the less he is paid. We also see that the better a player's team performs (`WL_percentage` coefficient), the less this player earns. We cannot be satisfied with this regression. Thus, we remove variables that produce these errors and see if the variables that were detected as good predictors are still predictors after our model's transformation:

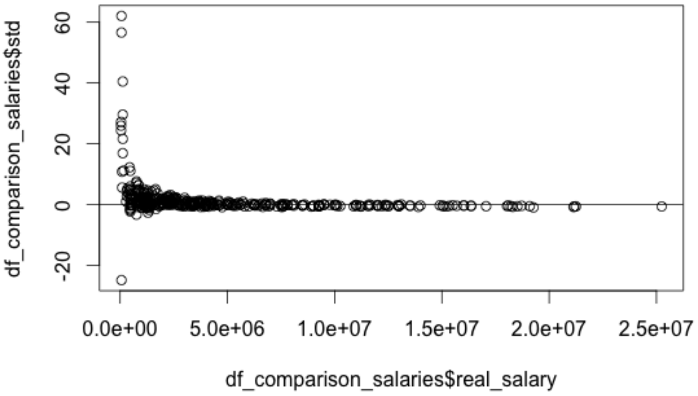
```
Coefficients :
                        Estimate Std. Error t-value  Pr(>|t|)
Age                    434450.61   30331.37 14.3235 < 2.2e-16 ***
PTS                          1456.62         404.42  3.6018 0.0003252 ***
X3P                          2769.00        3283.94  0.8432 0.3992387
AST                          2044.82        1462.27  1.3984 0.1621788
PosC                   152097.29 2134818.36  0.0712 0.9432104
PosPF                   69280.69 2098642.23  0.0330 0.9736688
PosPG                  919394.29 2050932.03  0.4483 0.6540076
PosSF                 -589410.59 2032567.80 -0.2900 0.7718646
PosSG                  700798.70 2025538.47  0.3460 0.7293996
TopPerformerPoints     239969.53 1164789.50  0.2060 0.8368003
TopPerformerWinShares 4044273.65 1332347.65  3.0354 0.0024384 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Total Sum of Squares:        2.1218e+16
Residual Sum of Squares: 1.8418e+16
R-Squared      :  0.13194
          Adj. R-Squared :  0.1172
F-statistic: 23.4073 on 11 and 1694 DF, p-value: < 2.22e-16
```

Interpretation We find in this new regression the same predictors as before. Thus, we can feel disturbed to have removed the "number of games played" variable. Thus, we observe that even if we put this variable in the list of predictors, the regression will also conclude that `TopPerformerWinShares`, `Age`, and `PTS` have significant influence on our outcome variable – salary (i.e. Pr(¿|) ¡ 0.05) – and are probably good predictors of players' salary. We also tried to many other combination of predictors, trying to replace variables by other variables with which they could be correlated (e.g. 2-points scored with total points scored) but we always found that these three variables are the best predictor of players' salaries. We also tried a random model regression by including the time-invariant variables height and weight but the Hausman test always gave us that the fixed effect model was better.

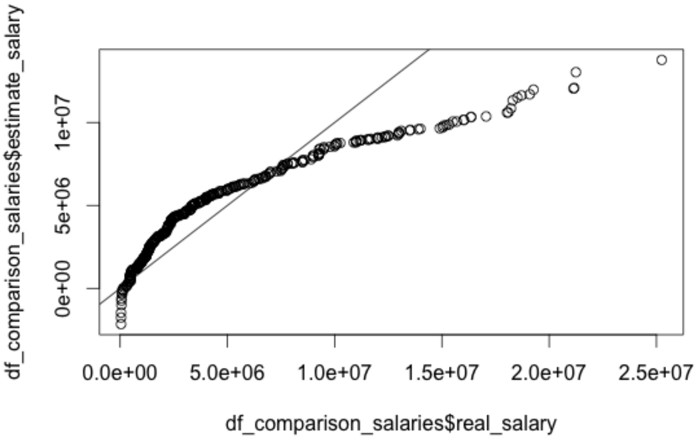
In conclusion, we suggest that the most important factor in players' salaries is age of players (which corresponds to their experience) and the number of points they score in one season. As we initially suggested, there is also an exception in player's salaries: players that have the best statistics for win shares in the championship are usually paid 4M$ than others. In concrete terms, team owner should consider that one year of experience in a player is worth 400k$. If a player score more points than usual, he could also give bonus up to 2000$ per point. Finally, owner of the best teams should reward their top performer with a bonus of 4M$.

Question 4

Upper-bound or lower-bound coefficients We use the best model we have, i.e. the model with the points, rebounds and assists. We first plot the standard error between the estimate salary and the real salary, with the real salary on the x axis. We can see that our prediction is not very good for low salaries but rather good for high salaries (even if it is normal that the standard error diminishes when the salary increases).



We then plot the estimate salaries with the real salaries on the x axis. We can see that the coefficients we predicted are upper-bound for low salaries and lower-bound for high salaries.


df_comparison_salaries$estimate_salary
0e+00
5e+06
1e+07
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
2.5e+07
df_comparison_salaries$real_salary