Question 3:

In this part we aim at using our data about player performances during the past ten years to better understand the factors that determine players' salaries. As classical linear regression techniques cannot necessarily take advantage of time-dependent variables measured on different entities, we use here a panel regression which is more adapted to this problem.

We recall the hypothesis of panel regression :

* The predictors of the dependent variables have to be independent, or at least not correlated.
* The panel regression can only find the role of time-dependent predictors in the value observed for the dependent variable X
* The random model can find the role of time-independent predictors but then unobserved time-independent variable has to be uncorrelated with other predictors

It is quite difficult to know in real life if a many variables are correlated or not, and if they are, how important will be the bias on the regression result. Therefore, our first idea is to keep only the following information about players for each season : age, games played, points scored, 3-points average, number of assists, number of steal, 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 performer during the whole championship.

All these variables are time-invariant. Indeed, as we saw in question 2, height and weight have little correlation with players' salaries, therefore we decide not to include them in this panel data. Moreover, fixed effect model couldn't take advantage of these variables. The dummy variables we introduced aims 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.

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 this criteria, we have still 202 players for 1907 observations.

The result of the panel regression are shown here :

*# Panel regression*

*formula = Salary ~ Age+G+PTS+X3P+AST+STL+WL\_percentage+Champion+RunnerUp+TopPerformerPoints+TopPerformerAssists+TopPerformerRebounds+TopPerformerWinShares*

*fixed <- plm(formula, data = data, index=c("Season", "PlayerID"),effect = "time", model = "within")*

*summary(fixed)*

|  |
| --- |
| *Coefficients :*  *Estimate Std. Error t-value Pr(>|t|)*  *Age 403289.9 30074.0 13.4099 < 2.2e-16 \*\*\**  *G -33281.5 6445.8 -5.1633 2.713e-07 \*\*\**  *PTS 3040.3 484.9 6.2700 4.572e-10 \*\*\**  *X3P 4311.4 3248.8 1.3271 0.1846603*  *AST 3763.4 1502.4 2.5049 0.0123404 \**  *STL -10254.9 6334.8 -1.6188 0.1056711*  *WL\_percentage -1349672.0 592780.8 -2.2768 0.0229197 \**  *Champion -230657.8 465308.1 -0.4957 0.6201636*  *RunnerUp -7398.1 415514.6 -0.0178 0.9857968*  *TopPerformerPoints 915521.6 1191645.0 0.7683 0.4424258*  *TopPerformerAssists -1649869.2 1261086.4 -1.3083 0.1909520*  *TopPerformerRebounds -1265388.4 1246707.4 -1.0150 0.3102585*  *TopPerformerWinShares 5011845.4 1461137.0 3.4301 0.0006179 \*\*\** |

We see here some unwanted results as : the more a player plays, the less he is paid. We also see that the more the team of a player is performing , 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 :

*formula = Salary ~ Age+PTS+X3P+AST+TopPerformerWinShares*

|  |
| --- |
| *Coefficients :*  *Estimate Std. Error t-value Pr(>|t|)*  *Age 427683.56 30020.92 14.2462 < 2.2e-16 \*\*\**  *PTS 1457.77 400.55 3.6394 0.0002815 \*\*\**  *X3P 2281.71 3274.58 0.6968 0.4860268*  *AST 2167.56 1438.25 1.5071 0.1319751*  *TopPerformerWinShares 4200212.53 1312704.25 3.1997 0.0014013 \*\** |

We find in this new regression the same predictors as before. Thus, we can feel disturbed to have remove the "number of games played" variable. Thus, we observe that even if we put this variables in the list of predictors, the regression will also conclude that TopPerformerWinShares, Age and PTS are good (i.e. Pr(>|t|) < 0.001) 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 are 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 considerate that one year of experience in a player is worth 400k$. If a player score more points than usual, he could also give bonus up to 5000$ per point. Finally, owner of the best teams, should reward their top performer with a bonus of 4M$.

n our project, we had the following information about players and teams for each season:

* Player's position in team
* Player's personal information : height, weight, age, experience
* Player's global data: games played, minutes played, points scored
* Detailed number of points : field goals scored/attempted/average, 3-points scored/ attempted/average, 2-points scored/attempted/average
* Behavior during games : assists, steal, block, turnover, personal foul
* Information about team's result : Win, Losses, Win-Loss percentage, Finish, Playoffs, Coaches, Top Win-Share player
* Information about winners, runner-up and best points/rebounds/assists/win-shares performer