Modelling player performance in basketball through mixed models

Martí Casals [1,2,3] and Jose A. Martinez[4]

[1] CIBER de Epidemiología y Salud Pública (CIBERESP), Spain.

[2] Biostatistics Unit, Public Health Department, University of Barcelona, Barcelona, Spain

[3] Departament de Ciencies Basiques, Universitat Internacional de Catalunya, Barcelona, Spain.

[4] Departamento de Economía de la Empresa. Universidad Politécnica de Cartagena. Spain

Abstract

The aims of this study were to identify variables which may potentially influence player performance, and to implement a statistical model to study their relative contribution in order to explain two outcomes: points and win score. We used all the possible variables affecting player performance creating a comprehensive database from two sources of statistical information about the NBA 2007 regular season: www.basketball-reference.com and www.nbastuffer.com. The data employed for the analysis were composed of 2187 cases (27 players * 81 games), having followed a filtering process. We dealt with a balanced study design with repeated measurements given that each player was observed the same number of games, and therefore the player was considered as a random effect. We carried out mixed models to quantify the variability in points and win score among players. Minutes played, the usage percentage and the difference of quality between teams were the main factors for variations in points made and win score. The interaction between player position and age was important in win score. We encourage managers and coaches of sports teams to choose appropriate methods according to their aims. Future research should take into consideration the use of models with random effects on players' characteristics.

Keywords: players, performance, mixed models, random effects, basketball

1. Introduction

To know which variables influence performance of a player together with the relative contribution of these variables is especially important for coaches and managers of sports teams. In basketball, although the academic and professional attention to quantitative analysis of data has exponentially grown since the appearance of the "Moneyball" phenomenon (Lewis, 2003), there is no study about modelling player performance in a game by game scenario. Therefore, none of the academic works such as Berri (1999), Berri and Bradbury (2010), Esteller-Moré and Eres-García (2002), Kubatko et al. (2007), Piette et al. (2010); seminal books as Berri et al. (2006), Berri and Scmidht (2010), Oliver (2004), Winston (2009); prospective books such as Doolittle and Pelton (2009), Hollinger, (2005); and a plethora of specialized websites (e.g. www.apbrmetrics.com, www.hoopsstats.com, www.nbastuffer.com, www.basketball-reference.com, www.82games.com, www.basketballvalue.com), speak about any specific model to determine variation in player performance through quantifiable data. Consequently, there is no tool for understanding why players vary their performance from one game to another, nor to know about which variables are actually affecting performance.

The aim of this research was to fill this gap, by identifying variables which may potentially influence player performance, and by implementing a statistical model to study their relative contribution to explain two outcomes: points and win score. The former metric reflects only the raw scoring production of players, and it has been linked to revenues and rewards for players in the NBA (Berri et al (2007)). The latter metric reflects the overall quantifiable contribution of players to team success, and it has been linked to team wins (Berri & Schmidt, 2010). In the NBA there are so many games and only a few players who end up playing in all of them during the regular season. This select group includes only those players who do not get injured or suspended (as well as being disciplined and in excellent shape) and have the complete trust of the coaching staff and fellow players throughout the season. This is one of the reasons why we focused this study on the performance of players with these characteristics. At this time, performance research into this type of player is unknown. We analyse data on the game by game performance of 27 NBA players, who played the entire length of the regular season. As we had repeated measurements for each player, we estimated a mixed model (also called multilevel model). Therefore, as an alternative of developing a particular model for each player, a comprehensive multilevel model was constructed, allowing parameters to vary from one individual to another, to take into account the heterogeneity between players. This method has been used in ecology, epidemiology, genetics and other aplications but it is gaining attention in sports sciences (e.g. Avalos et al. (2003); Bullock & Hopkins, 2009; Sampaio et al.(2010)).

Therefore, the contribution of this paper to sports sciences is two-fold: First of all, we make an extensive work of identifying and recording variables which may affect the performance of a basketball player from one game to another. These variables are put together in a comprehensive statistical model in order to control their covariance; and secondly, we study the significance of these variables through mixed models, showing the conditional effects on performance. Consequently, this research is the first attempt in the basketball literature to try to understand variation in player performance in a game by game context using such extensive amount of information available from quantifiable statistics.

2. Methods

2.1 Data

We purchased two data bases of statistical information about the NBA 2007 regular season from www.basketball-reference.com and www.nbastuffer.com. The first one had basic statistics of all players in a game by game format. The second one also had basic statistics of teams for each game of that season. We worked roughly joining both databases in a single one, in order to assign to each game played by any player the information about his team and opponents. At this stage, we wanted to identify all the possible variables affecting player performance. We had to complete the available variables with data from other sources and to create new variables from the existing ones. In the following lines we show the chosen variables and the rationale of their selection.

Dependent variables

- *Points:* This variable represents the points made by each player in each game. As Berri, Brook and Schmidt (2007) explain, points scored traditionally have dominated the evaluation of player productivity in the NBA. Player evaluation in the NBA seems overly focused upon scoring in terms of salary and coaches evaluation of player talent. Obviously, this variable only reflects a portion of the overall performance of a player. Therefore, other metrics are needed in order to quantify other aspects of performance beyond scoring. And this is the reason why we have added a second dependent variable: win score.
- *Win score:* Among the hundreds of methods proposed to evaluate player performance form quantifiable statistics, the works of Berri and colleagues are among the best recognized by the academic literature. The “Wins Produced” and its simplification “Win Score” are two of the most accepted ways to rate players by the academic community. The works of Berri (1999; 2008; 2012), Berri et al. (2007) or Berri and Bradbury (2010) explain the foundation of this method, the important limitations of other famous metrics such as “Efficiency”, “PER”, “Plus-Minus” and “Adjusted Plus-Minus”, and why Wins Produced overcomes such metrics. Berri (1999; 2008) links box-score statistics with team wins through regression analysis. Win Produced correlation with win score is 0.99, when both variables are normalized per minute (Berri & Schmidt, 2010). Therefore, the major contribution of Berri (1999; 2008) is to propose an easily understandable metric, which may be manually computed for any person, and which exceptionally correlates with team wins. To summarize: Win Score = Points + Rebounds + Steals + 0.5 Assists + 0.5 Blocks – Turnovers – Field Goals Attempted – 0.5 Fouls – 0.5 Free Throws Attempted. This metric can be easily computed from the box-score of each game and it is accessible to analysts and fans.

Independent variables

- *Individual, Team, Division and Conference:* These are the "level variables" of the multilevel model. We could expect individual differences in performance linked to personal factors of players beyond what we measure with observable statistics. Individuals (players) are nested within teams (30 teams), teams are nested within divisions (6 divisions), and division are nested within conferences (2 conferences). A priori, these are different levels of analysis, and repeated measures of player performance are sequentially nested within these levels. Again, teams, divisions and conferences have their particular characteristics. For example, in the NBA, the schedule of teams is different depending on the division and conference where the teams play. Therefore, there is no homogenous calendar for all teams, because some teams play against more powerful teams than others, and some teams play more times against a specific team in the same season than others.
- *Season period:* Following Sampaio et al. (2010) we hypothesize that performance could vary depending on the season period. For example Clark et al. (2008) identified performance changes in physiological variables in mid-season, although a manifest decline in some indicators towards the end of the season. This could be related with a decrease of performance beyond a threshold of games played. In addition, it is well known in sports the periodization of physical training in order to get performance peaks in some key moments of the competition. Metaxas et al. (2006) split the season into four different periods in order to analyze changes in performance. We used the same criteria, so that we divided the 82 games of the NBA regular season into four periods (21, 20, 20 and 21 games, respectively).
- *Home advantage:* Home advantage is a pervasive phenomenon in sport (Koning, 2010). In basketball, the home team wins 60% of the games (Arkes & Martínez, 2011), and there is a margin of 3.2 points in the final result for the home team (Winston, 2009). Therefore, it is plausible to expect that players perform better at home than at road.
- *Difference of team quality:* Some teams are better than others, and this is reflected in the winning percentage obtained at the end of the season. As Arkes and Martínez (2011) showed, the difference of quality between teams influences results, so we hypothesize this could also influence player performance. Therefore, we followed the same procedure Arkes and Martinez (2011) implemented: We made two partitions of the regular season (the first 41 games and the second 41 games), and we computed the winning percentage of each team for the two partitions. Then we also considered the home winning percentage and the road winning percentage of each team at each partition, in order to take into account that some teams are better at home or at road than others. Consequently, this variable is computed from the winning percentage of the home team at home minus the winning percentage of the road team at road, both referred to the partition they pertain in such game. Winning percentages as taken as continuous variables ranged in a [0,1] interval. For a deep explanation of the advantages of this procedure, see Arkes and Martínez (2011).
- *Quality factor of a game:* Martínez (2012) proposes that the difference of team quality could not reflect a factor related with the "quality" of the game. For example, if the difference of quality between teams is 0.5, this could be due to the

winning percentages of both teams are 1 and 0.5, or they are 0.6 and 0.1. For both cases, the difference of quality is 0.5, but Martínez (2012) suggests that, in the former case, the quality of this game is superior to the quality of the latter game, because globally, the first two teams are better than the second two teams, and this fact could influence results. Martínez (2012) uses an exponential parameter λ to address this fact, thus quality factor of a game=|difference of quality between teams|$^{1/\lambda}$. Therefore, this variable increases to the extent that the winning percentage of the two teams increases, and it is also a continuous variable ranged in a in a [0,1] interval.

- *Rest days:* Reed and O'Donohue (2005) suggested counting with this variable to study factors affecting performance in sports. However, Arkes and Martínez (2011) did not find an important effect of this variable on the probability of winning a game. We have considered this variable using four categories. The first three categories reflect one, two and three rest days between games, and the fourth category includes the situation when players rest four or more days from one game to another.
- *Game started:* Players in the starting line-up are usually the best players of the team, and they play many more minutes than the remaining players. For the 2007 season in the NBA, starting players played a mean of 32.2 minutes against 15.2 of the non-starting players. Started players also use to play the clutch minutes of the game. For the vast majority of cases, started players are the most trustworthy players for their coaches, and this could influence their own confidence for playing. Therefore, it is a variable susceptible to be analyzed in order to explore its effect on performance.
- *Player momentum:* Arkes and Martínez (2011) showed there was evidence for a momentum effect in the NBA teams, i.e. positive or negative trends in results could positively or negatively influence the outcome of a subsequent game for that team, other things being equal. This is the reason why we measured the performance of players in the prior five games (points and win score) using the median of these five games[1].
- *Player wage relative to team salary:* This variable is usually to be linked to the quality or value of a player. However, as Berri, Brook and Schmidt (2007) or Berri and Schmidt (2010) show, decision makers in the NBA are eminently irrational,

[1] We labeled these five games as G-1, G-2, G-3, G-4 and G-5. In addition, we also considered other alternative ways of forming this single variable, using linear weights and non-linear weights, in order to give more importance to the last games. For the case of linear weights, the last five games weighted 1, 0.9, 0.8, 0.7 and 0.6, respectively. For the case of non-linear weights, the last five games weighted 1, 0.5, 0.33, 0.25 and 0.2. Therefore, and taking the last game as the reference point, for the case of non-linear weights, G-5 weighted 1/5 times G-1, G-4 weighted 1/4 times G-1, and so on. Finally, we considered another potential influencing variable: the revenge factor. This variable refers to the performance of a player against the same team in the last game both teams played, and it has been suggested as relevant for forecasting basketball results (Mallios, 2011). It is suggested that if a player achieve a poor performance against a specific team, it could be expected to be highly motivated to "revenge", and to perform better the next time both teams play. We incorporated the revenge factor to by forming the arithmetic mean from the variable which reflects the performance of the last five games played and from the variable which reflects the performance of the last game played against the seam team.

because they mainly value players by their scoring ability instead of other criteria as individual wins produced. The salary paid to players by each team only explains 6% of the variance of teams wins. As teams possessing more economic resources also spend much more money than others (Eldridge, 2010), because of the exceptions of the salary cap and the different size of the markets of the cities where they are located, then salary of players should be normalized considering the total budget of their respective teams. In other words, as there are teams spending almost twice other teams in player salaries, then the wage paid to a specific player could largely vary in function of the team he is playing for, regardless of the quality of that player. Therefore, we use salary as an indicator of the quality of a player, but we divided the salary perceived by the total salary paid to the roster he belongs to.

- *Teams fighting for the playoffs:* We distinguished between players pertaining to teams classified for playoffs or to teams which have been fighting until the end of the season to enter to the playoffs, from players pertaining to the remaining teams. We hypothesize that motivation of players pertaining to teams with low winning percentage could decrease when their teams lose any chance to be classified for the playoffs. And this fact could influence performance.
- *Player position:* As Berri and Schmidt (2010) explain, players should be evaluated in function of their position. Therefore, performance could be affected by the position played. In fact, win score rewards tall players, because obviously they generally grab many more rebounds than smaller players. We divided positions into three categories considering the characteristics of each player: Guard or Point Guard (1), Forward (2), and Power Forward and Center (3). This is a traditional classification that distinguished playmakers (1) from other exterior players (2), and all these players from big players who use to play in the paint (3).
- *Age:* Berri et al. (2006) showed age influenced the career performance of a player. We included age in our model instead of years of experience in the NBA because some players can be incorporated to the NBA directly from high-school (about 18 years old) and others when they finish college (about 22 years old). In addition, some international players usually play several seasons in highly competitive leagues before playing in the NBA.
- *Contract condition:* Contract conditions for players are very diverse, but we used a simple classification taken from www.storytellerscontracts.com in order to categorize players into two groups: players with salary guaranteed for at least the successive year (0), and players whose contracts finalize at the end of the season (1). We hypothesized that players pertaining to this latter group could play with an extra motivation in order to obtain a good contract for the next year.
- *Minutes played:* There is some body of work reflecting weak evidence of a lineal effect between points made and minutes played (Martínez & Martínez, 2010). However, it would be unrealistic to expect players to always perform equally (the same points made per minute) regardless of minutes played. Fatigue and a learning effect of defense could make players to obtain diminishing returns of points made to the extent that minutes played increase. This is akin to non-linear decreasing returns modeled using logarithmic transformations of variables (see Wooldridge, 2002). This reasoning is extendable to win score.

- *Usage percentage:* Usage percentage is an estimate of the percentage of team plays used by a player while he was on the floor. The formula is 100 * ((Field goals attempted + 0.44 * Free throw attempted + Turnovers) * (Team minutes played/ 5)) / (Player minutes played * (Team Field goals attempted + 0.44 * Team Free throw attempted + Team Turnovers)). It is obvious to suppose points and win score increase with usage percentage. However, diminishing returns again could be expectable to the extent that usage percentage increases, because of fatigue and defense learning effect.

2.2 Filtering process

The raw data base was composed of 458 players and 25806 games. Therefore, the mean of games played by players was 56. There were only 3 players who participated in only 1 game and 38 players who played all the 82 games of the season. Therefore, the heterogeneity of the data was high, and we decided to employ an exhaustive filtering process. We used listwise deletion, so when a specific case (game) did not fulfil the requirements, we eliminated that player, and consequently all the cases (games) pertaining to that player.

First of all, we dropped players who had played less than 5 minutes in some of their games. We considered that playing less than 5 minutes didn't allow players to correctly develop their skills. This cut-off criterion has been also used by Martínez and Martínez (2010) and Sampaio et al. (2010). Secondly, we removed players who had been traded that season, because they changed their teams, and hence they significantly changed their context for playing (teammates, coaches, city, etc.). And thirdly, we only considered players who had played the entire season. The rationale of this decision is to exclude the possible influence of injuries or sanctions, among others, which can make players to miss games. When this fact occurs, the effect of other variables on players' performance could be seriously biased when these players return. For example, rest days or player momentum are variables which would be highly distorted by player injuries. Consequently, we only considered players who played 82 games.

Finally, only 27 players passed this filtering process. In addition, the first game of each player was also deleted, in order to correctly use both variables: the momentum effect variable and the rest days. It is obvious that in the first game of the season there is no possibility to compute any data reflecting momentum neither rest days between games. For the second, third, fourth, and fifth game of each player, momentum was computed using the 1, 2, 3 and 4 latter games. Therefore, from the game 6 to the game 82 of each player, momentum reflected the performance of the five last games. The final data base was hence composed of 2187 cases (27 players * 81 games), with no missing values in any of the aforementioned variables, and a completely balanced design. Table1 summarizes all the variables employed, with the succinct distinction between time variant covariates (also called inner variables), and time invariant covariates (also called outer variables). The former variables vary within each level but the latter vary only in the cross-sectional part of the data, i.e., outer variables are constant for each player but different between players. In addition, minutes played and usage percentage are endogenous covariates because it is also plausible to think that they are also determined by points and win score. For example, if a player is scoring a lot of points, then his coach could have him playing more minutes than expected. In addition, if a player feels "on fire" (playing very well and making points), then he could try to use more plays (increasing his usage percentage) because of the effect of self-confidence.

Table 1. List of variables and descriptive statistics.

Variables	Type	Categories/ Range	Mean	St. Dev.
Dependent				
Points (P)	Continuous		12.77	7.89
Win Score (WS)	Continuous		5.29	6.04
Level				
*Player**	Nominal	27		
Team	Nominal	22		
Division	Nominal	6		
Conference	Nominal	2		
Covariates (inner)				
Season period	Nominal	4		
Home/road game	Nominal	2		
Difference of team quality	Continuous	[0,1]	0.0094	0.294
Quality factor of a game	Continuous	[0,1]	0.246	0.175
Rest days	Nominal	4		
Game started	Nominal	2		
Player momentum				
- *Median (P)*	Continuous		0.405	0.146
- *Median (WS)*	Continuous		0.159	0.125
Covariates (outer)				
Player wage relative to team salary	Continuous	[0,1]	0.076	0.056
Team fighting for the play-off	Nominal	2		
*Player position**	Nominal	3		
Age	Continuous		26.37	3.955
Contract conditions	Nominal	2		
Covariates (endogenous)**				
Minutes played	Continuous		30.17	8.98
Usage percentage	Continuous		20.55	6.94

* PLAYERS: Alston, Rafer (Houston); Battier, Shane (Houston); Bell, Charlie (Milwaukee); Bibby, Mike (Sacramento); Biedrins, Andris (Golden State), Blount, Mark (Minnesota); Carter, Vince (New Jersey); Collison, Nick (Seattle); Dalembert, Samuel (Philadelphia); Deng, Luol (Chicago); Finley, Michael (San Antonio); Fisher, Derek (Utah); Foye, Randy (Minnesota); Gordon, Ben (Chicago); Granger, Danny (Indiana); Howard, Dwight (Orlando); James, Mike (Minnesota); McDyess, Antonio (Detroit); Pargo, Jannero (New Orleans); Parker, Smush (L.A. Lakers); Prince, Tayshaun (Detroit); Snow, Eric (Cleveland); Stevenson, DeShawn (Washington); Stoudemire, Amare (Phoenix); Warrick, Hakim (Memphis); Webster, Martell (Portland); Wilcox, Chris (Seattle).

* * PERCENTAGES IN THE SAMPLE: Guard or Point Guard (52%); Forward (26%); Power Forward and Center (22%).

*** Endogenous covariates were log-transformed in order to count with diminishing returns.

3. Model building and Statistical analyses

We carried out two mixed models to quantify the variability in points and win score among players. For win score as a response variable, we fitted a linear mixed model, and for points we fitted a generalized linear mixed model assuming Poisson error structure and log link function.

Linear mixed models (LMMs) and generalized linear mixed models are an extension of linear models (LMs) and generalized linear models (GLM) adding random effects in the linear predictor term to the regression setting. They allow us to model the dependence structure among dependent variables for longitudinal or repeated measures data.

Let Y_{ij} of individual i (i=1,…..,N players) and the j-th game be the dependent variable. Y_{ij} is modeled by a linear mixed model when it comes from the multivariate gaussian parametric density family. On the one hand, the LMM is commonly written:

$$Y_{ij} = X_{ij}\beta + u_i + \epsilon_{ij} \quad u_i \sim \mathrm{N}(0, \sigma_u^2) \quad \epsilon_{ij} \sim \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2\, \mathbf{I}) \qquad (1)$$

This model assumes that the variances of the errors are normally distributed. In the LMM, there are two sources of variability in the model, an individual-to-individual variability in the level of the dependent variable and the residual. The σ_u^2 is the variance of random effect and σ_ϵ^2 is the variance of residual error. Both $X_{ij}\,\beta$ and u_i are lineal predictors with known design matrices. The vector β contains the fixed effects (covariates of interest) that change within a player and u_i (players) is the vector of random effects.

On the other hand, the GLMM does not model a normally distributed response variable, but rather counts or proportions. The notation of GLMM is similar to LMM , but it uses a logit link and the response variable Y_{ij}~Poisson(μ_i), which μ_i is the expected value of Y_{ij}. The Poisson mixed model indicates the expected number of counts as

$$\log(\mu_i) = X_{ij}\beta + u_i \quad u_i \sim \mathrm{N}(0, \sigma_u^2) \qquad (2)$$

These models incorporated fixed effects that were treated as categorical and continuous variables, in function of each variable type. We dealt with a balanced study design with repeated measures given that each player is observed the same number of games, and therefore player is considered as a random effect.

As a first stage of the model building procedure, we considered the possibility of adding a new random effect in the analysis using nested or crossed random effects designs. We dropped this possibility because the variability of the other covariate as a random effect was not important. The division and the conference variables were correlated, so we only used the first variable because it collected more information, but again the variability of the division as a random effect was not relevant. Consequently, we only had a random intercept (the player effect), and it was taken as the random component of the model. Before fitting the mixed models instead of the classical models (LM or GLM) we checked likelihood ratio tests for evaluating whether including a random effects parameter. We fitted a model with and without the variance component and compared the quality of the fits. The null hypothesis of this test was H_0: σ_u^2 =0 versus the alternative H_A: σ_u^2 >0 proposed (Verbeke & Molenberghs, 2000). The likelihood ratio with a corrected chi-squared is a reasonable statistic test for the comparison (Baayen, Davidson & Bates, 2008). The presence of random effect (player) was clearly necessary.

As a second stage in the LMM we followed the strategy proposed by Pinheiro and Bates (2000) to select the final model to be estimated, using a forward stepwise approach. This model building approach (to be used for incorporating covariates in the model) consists of starting with an LMM without covariates to explain inter-individual variation. We used plots of the estimated random effects versus the candidate covariates to identify interesting patterns. A pattern in each step would indicate that the covariate will be included in the model. This procedure was applied sequentially until no further interesting patterns were found. Before deciding on the most appropriate models we used the Akaike information criterion (AIC; Akaike, 1973) that determines the maximum likelihood of a candidate model. The model with the lowest AIC represents the best model. When we compared models we refitted all models using the maximum likelihood criterion (ML) as a estimation criterion. For the final model we refitted it again using the restricted maximum likelihood criterion (REML), and we used this model for inference. On the other hand, as a second stage in the GLMM model simplification was performed by backward selection of variables from full model as suggested by Crawley (2007) and Bolker (2009), and models were compared using likelihood ratio tests (LR) until a minimal adequate model was obtained. Model selection was based on the AIC. There are several ways to approximate the likelihood to estimate GLMM parameters. In our case, we used the Gauss-Hermite quadrature (GHQ) which presents better advantages of estimation than others because we have only one random effect (Bolker, 2009).

In both models, the significance of the fixed-effects associated with a covariate included in the model was assessed using the Wald tests. A *p*-value below 0.05 was considered significant. We also checked the correlation and main possible interaction of covariates.

After fitting the models, they were validated. Residuals of the final models were explored for normality, homogeneity (except in the case of the GLMM) and independence assumptions. Normality assumption of the residuals was checked via q-q normal graphs of residuals. In addition, a correction for heteroscedasticity was applied in the LMM by modeling variance as a power of the fitted values (Pinheiro & Bates 2000). The constant plus power of variance covariance function was used to model the variance structure (varConstPower(form=~fitted(.)) as it had the lowest AIC. In the GLMM, we checked for overdispersion in Poisson regression model using the Pearson dispersion parameter, which was < 1.5. Hence, it was unnecessary to use quasipoisson errors and quasi-AIC (QAIC) to correct for overdispersion (Bolker et al. 2009).

All statistical analyses were performed using the statistical package R (The R Foundation for Statistical Computing, Vienna, Austria), version 2.13.1. We used the main packages for mixed models called *nlme* (Pinheiro & Bates, 2011) and *lme4* (Bates, Maechler & Bolker, 2011).

4. Results

Regarding win score, the final set of covariates selected after applying the model building procedure were: player, difference of team quality, age, player positions, the interaction between player position and age, minutes played and usage percentage. Table 2 shows the estimates of the LMM. The importance of the interaction meant that in the older centers the win score decreased. This is the way age influenced win score, because it had no significant effect by itself. The difference of quality between teams had a strong influence on wins produced, together with the minutes played and usage percentage. The Hausman-Taylor estimate (Hausman & Taylor, 1981) yielded a similar result, which indicated that endogeneity was not an important issue for such data.

Table 2: Linear mixed models with the win score as dependent variable.

	Linear mixed model with random intercept				
Coefficients	Estimate	SE	*df*	*t*	*p*-value
Intercept	-19.85	2.74	2157	-7.28	<0.001
Log Minutes played	6.47	0.31	2157	20.71	<0.001
Log Usage percentage	0.67	0.28	2157	2.34	<0.001
Age	-0.01	0.08	21	-0.13	0.894
Difference of team quality	2.42	0.36	2157	6.65	<0.001
Forward	5.91	6.18	21	0.95	0.349
Center	12.86	3.87	21	3.32	0.003
Forward x Age	-0.14	0.24	21	-0.60	0.549
Center x Age	-0.31	0.14	21	-2.17	0.041
σ_u		1.16			
σ_ϵ		4.85*			

* It is the value of residual standard deviation when the model is homocedastic. However, with the variance function, the residual standard deviation would have the following expression: 0.50*(5.96+abs(x)^0.77), which is denote the variance function evaluated at x.

Regarding points, the final set of covariates selected after applying the model building procedure were: player, difference of team quality, game started, age, player position, minutes played and usage percentage. Table 3 shows the estimates of the GLMM. As it is shown in the results (Table 3), the Western Conference is not significant. However, we must acknowledge that there was weak evidence this variable is important, because the AIC of the model was better than the simpler model without the conference variable. As expected, the average points of the big players (forwards and centers) were greater than those expected for guard players.

The game started variable was significant, with a negative effect. This result reflects the complex relationship between minutes played and points made. This finding does not mean that non-started line-up players would make more points than game started players, because if they would play more minutes they would be game started players. This finding means that they score relatively more than game started players for the minutes they play, but it they would play more minutes, then diminishing returns probably would make they would score the same as game started players.

As expected, minutes played and usage percentage were highly associated with points, together with the difference of quality between teams.

Table 3: Generalized linear mixed models with points as dependent variable.

	Generalized Linear mixed model with random intercept			
Coefficients	Estimate	SE	z	p-value
Intercept	-4.96	0.15	-32.85	<0.001
Log Minutes played	1.31	0.03	47.59	<0.001
Log Usage percentage	1.05	0.02	46.76	<0.001
Age	-0.006	0.003	-1.84	0.06
Difference of team quality	0.12	0.02	5.49	<0.001
Forward	0.06	0.03	2.08	0.03
Center	0.07	0.03	2.22	0.02
Western Conference	0.02	0.01	1.57	0.11
Game started	-0.09	0.02	-4.33	<0.001
σ_u	0.05			

Both models allowed us to obtain the coefficients of each player when we added the random effects (Figure 1 and Figure 2) to the intercept of the model (average population). We could see the archetype player (value 0) in the middle of the plot and the best players' performance to the right.

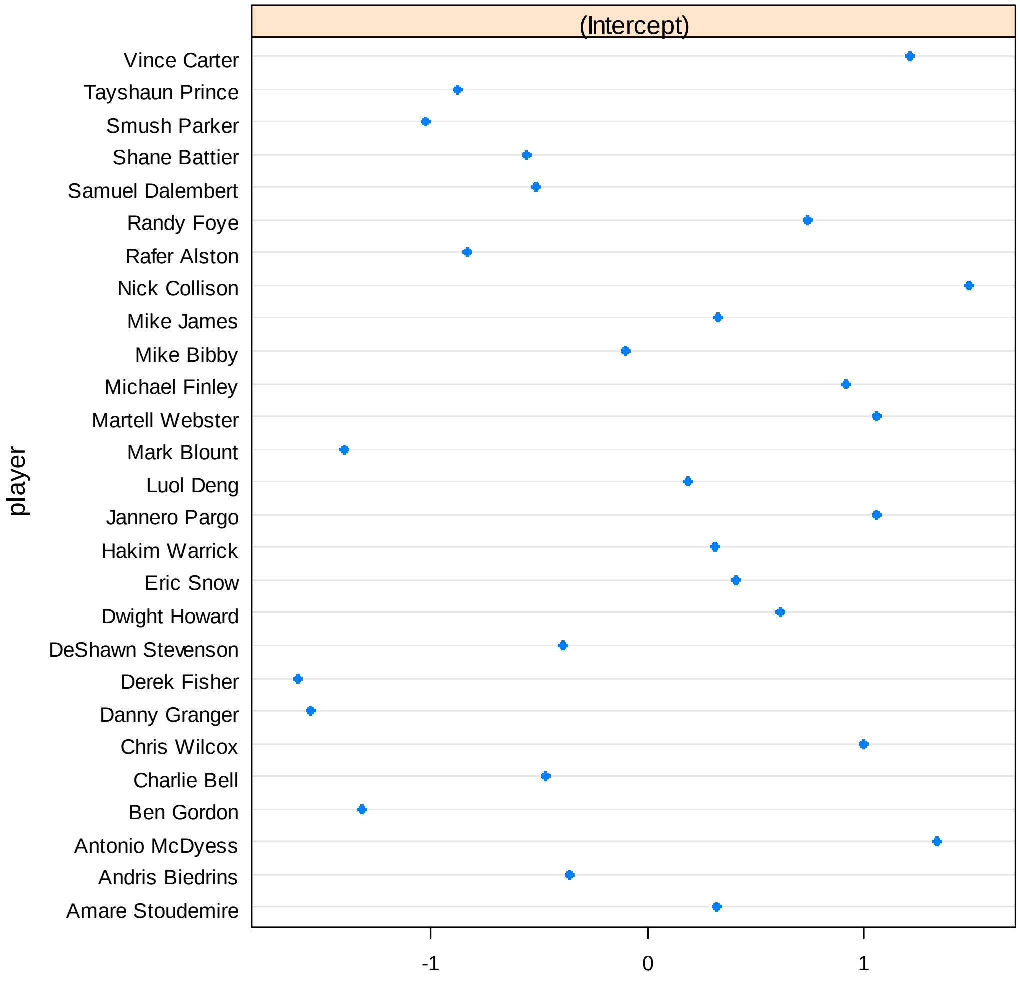

Figure 1: Coefficients for the random effects intercept in the win score model

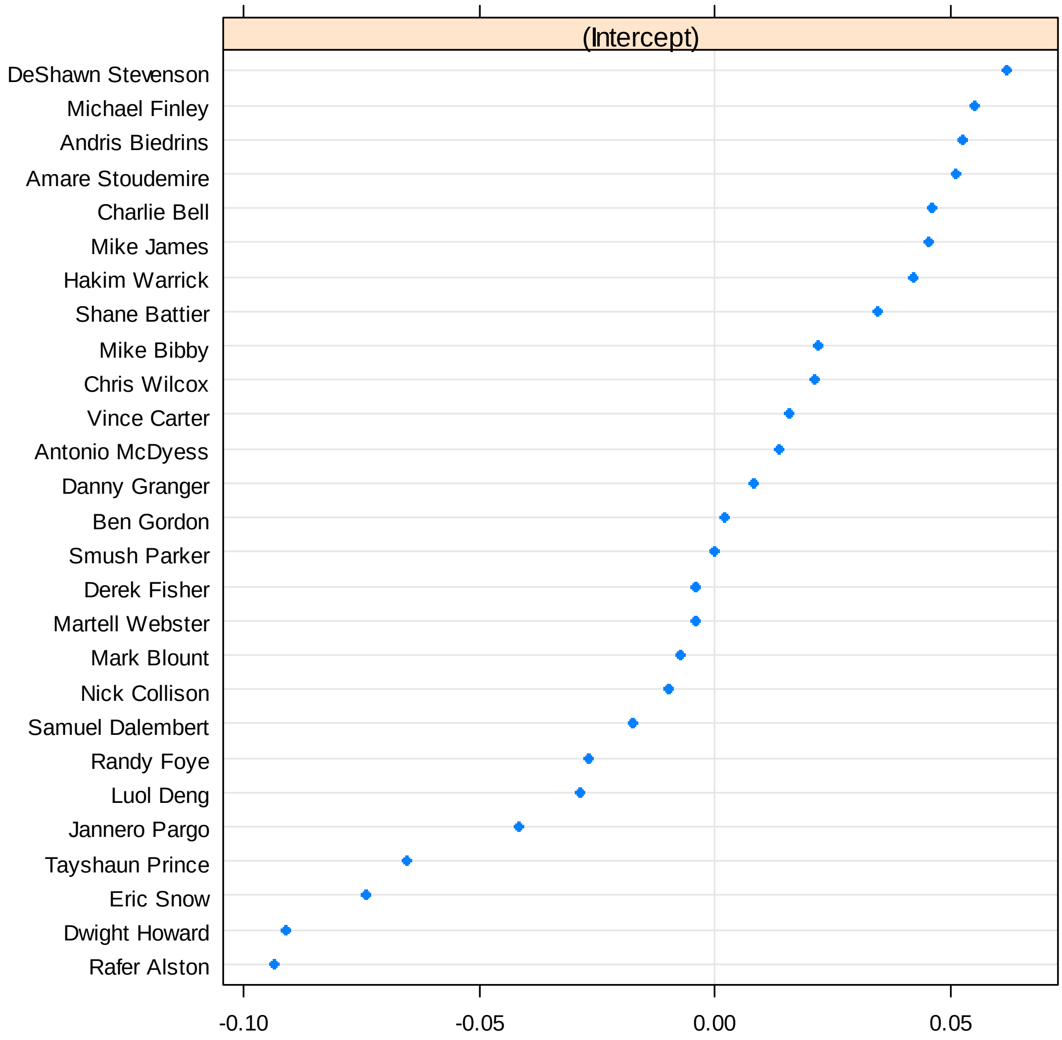

Figure 2: Coefficients for the random effects intercept in the points model

5. Discussion and implications

This research has modelled player performance in basketball using a mixed models approach. We considered the use of covariates to model random effects variability, taking into account the strategies described to give an adequate and parsimonious model. Therefore, it is possible to know which variables are potentially useful in explaining random effects variation and which random effects may have their variability explained by covariates. It is also noticeable that in the LMM this is possible by analyzing plots of random effects estimates versus the covariates, and looking for trends and patterns.

Results derived from the outcome of the model building procedure showed that minutes played and usage percentage had a great impact on points and win score. The importance of these variables to determine performance was not surprising. Therefore, the most relevant result of our analysis was related to the difference of quality between teams. This is the unique time-variant (inner) covariable affecting significantly player performance, with the exception of the game started variable for the points made. The remaining inner covariables were not significant, so variables such as player momentum (in all of its different manifestations), rest days, home/road games or the quality of a game were not relevant. Consequently, variables affecting the result of a game from a team perspective such as home advantage or team momentum (see Arkes & Martínez, 2011), are not relevant from a player perspective. This apparent contradictory finding is easily understandable if we think that, for example, the margin of points derived from the home advantage: 3.2 points (Winston, 2009) has to be redistributed among the players of the team playing at home. Therefore, it is plausible to think that each individual player probably performs a little better at home, but this effect size is so small, that we would need a sample of several ten thousands of players to find it significant. The same reasoning can be made with momentum.

There is not evidence that season period influenced points and win score. Considering that the recent study of Sampaio, Drikwater and Leite (2010) did not find evidence for the effect of season period on game related statistics, we acknowledge that more support to our result would be desirable. It is true that the method and variables used by Sampaio, Drikwater and Leite (2010) were different, but again further research should deepen into this issue in order to obtain clearer evidence.

The effect of time-invariant covariates (outer) should be also taken with caution. It is true that Rabe-Hesketh and Skrondal (2012) suggest using a minimum sample from 20 clusters to estimate a random effect model when such clusters were sampled from a larger population. However, although we had 27 different players (players are the cluster or level variable), and with a heterogeneous/representative profile, we must acknowledge that this could be insufficient to guarantee accurate inferences about the role of outer variables to determine performance over a population of more than 400 NBA players. As we explained, our filtering process was considered necessary to obtain a balanced design of data with similar conditions for the inner variables.

The results obtained can only be applied to players that fit these characteristics and cannot be applied to all NBA players. Possible future research may focus on trying to use these statistical analysis models to study all the players that participate in the NBA season, or even multiple seasons. With observational studies we could deal with other complex situations such as unbalanced designs (Schafer & Yucel, 2002).This would distort a little the measure of variables such as player momentum or rest days, but it would gain power to detect the effect of outer variables on performance.

Anyway, the main value of this research is on the implications derived from inner covariates. Researchers, managers and coaches probably already know that best paid players are very good scorers (Berri et al., 2007), but this is a minor issue to predict how a particular player will vary his performance from one game to another.

Nevertheless, some of our results regarding outer variables should be studied in future research. For example, the interaction between age and position (big men players), and its negative coefficient associated with win score was interesting. Young centers are preferred for many teams, and the history of the first numbers of the NBA draft selection is full of examples of this fact. Again further studies must ascertain with more evidence whether this is a rationale decision. In addition, the effect of age on performance would deserve major attention. Age was slightly negatively associated to points made, and this could contradict some studies regarding aging curves and performance. For example, Bradbury (2009) found that baseball players peak much later than assumed, at age 29 or 30. Meanwhile, Berri et al. (2006) found different results in basketball --a peak somewhat younger than 27, instead as early as 24 or 25. These studies were commented by Pelton (2010), who proposed an interesting form of analyzing aging curves for NBA players, by focusing in the relative increase or reduction of performance from one year to another. Pelton (2010) found a slightly negative linear association between age and performance, which it would be similar to the small negative linear trend we have found.

Globally, all results show that performance of players with these characteristics was very stable once controlling for minutes played and usage. In order to predict changes from one game to another only the difference of team quality between the player team and the opposite team would yield important variations. Beside the implications of this result for coaching and managing teams, it is also important to other contexts, such as for example, fantasy games. This industry has vastly increased in the last years, where players spend more than $1.65 billion per year only in the United States (Ballard, 2004). Players of this type of games could improve their scores focusing on NBA players whose team plays against a low quality team, because points made and win score would significantly increase. Predictions could be made on the basis of current information about quality of teams, which it would be somewhat different from the retrospective information we have used to form this variable. Therefore, predictions would be more accurate to the extent that season advances, because the difference of teams quality variable would be more realistic.

In conclusion, this research has applied the mixed models analysis to understand performance of basketball players in a game by game context. Minutes played, the usage percentage and the difference of quality between teams are the main factors for variations in points made and win score. From the game's point of view, win score and points are different variables when looking at the players' performance. This is also reflected from a statistical point of view, given that we have to take into account that the distribution of the two variables is different, as it is shown in the use of a linear mixed model or Poisson mixed model. Points made was much more explained than win score by a subset of covariates, which indicates that the overall production of a player would be more difficult to predict, because a player could produce the same contribution to the team by focusing on divergent aspects of the game. The statistical methods commonly used by sports researchers focus on the mean tendency in the data through a linear model (e.g., linear regression, analysis of variance) they might ignore random effects altogether (thus committing pseudoreplication) or treat them as fixed factors. This idea clashes with the importance of variability in basketball and sports research in general. The currently increasing interest in mixed models will surely continue in the future. Mixed models will be very useful in disciplines such as sports research. Sports coaches and managers have much to gain from LMMs and GLMMs. By incorporating random effects, the mixed models also allow sport researchers to generalize their conclusions to other leagues and players. We encourage managers and coaches of sports team to analyze it in more detail according to their aims. Future research should take into consideration the use of models with random effects on players' characteristics. Ignoring the correlation of observations between players may lead to an underestimation of the standard error.

6. References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. **Second International Symposium on Information Theory**, 267-281.

Arkes, J. & Martínez, J. A. (2011). Finally, evidence for a momentum effect in the NBA. **Journal of Quantitative Analysis in Sports** 7 (3), Article 13.

Avalos, M., Hellard, P. & Chatard, J. C. (2003). Modeling the training-performance relationship using a mixed model in elite swimmers. **Medicine & Science in Sports & Exercise** 35 (5), 838-846.

Baayen, R. H., Davidson, D. J. & Bates, D.M. (2008) Mixed-effects modeling with crossed random effects for subjects and items. **Journal of Memory and Language** 59, 390-412.

Ballard, C. (2004, 21 June). Fantasy World. **Sports Illustrated**

Berri, D. J. (1999). Who is 'most valuable'? Measuring the player's production of wins in the National Basketball Association. **Managerial and Decision Economics** 20, 411-427.

Berri, D. J. (2008). A simple measure of worker productivity in the National Basketball Association. in **The Business of Sport,** eds. Brad Humphreys and Dennis Howard, editors, 3 volumes, Westport, Conn.

Berri, D.J. (2012). Measuring performance in the National Basketball Association. In Stephen Shmanske, S. and Kahane, L. (Eds): **The Handbook of Sports Economics.** Oxford University

Berri, D. J., & Bradbury, J. C. (2010). Working in the land of metricians. **Journal of Sports Economics**. 11 (1), 29-47.

Berri, D. J., Brook, S. L., & Schmidt, M. B. (2007). Does One Simply Need to Score to Score? **International Journal of Sport Finance** 2 (4), 190-205

Berri, D. J., & Schmidt, M. B. (2010). **Stumbling on wins: Two economists expose the pitfalls on the road to victory in professional sports**. FT Press.

Berri, D. J., Schmidt, M. B., & Brook, S.L. (2006). **The wages of wins: Taking measure of the many myths in modern sports.** Stanford, California: Stanford University

Bradbury, J. C. (2009). Peak athletic performance and ageing: Evidence from baseball. **Journal of Sport Sciences** 27 (6), 599-610.

Bates, D., Maechler, M., Bolker, B. (2011). lme4: linear mixed-effects models using S4 classes. R package version 0.999375-39.http://CRAN.R-project.org/package=lme4

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H.,and White, J. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. **Trends in Ecology and Evolution** 24 (3):127–135. doi:10.1016/j.tree.2008.10.008.

Bullock, N. & Hopkins, W (2009). Methods for tracking athletes' competitive performance in skeleton. **Journal of Sports Sciences** 27 (9), 937-940

Clark, N., Edwards, A., Morton, R., & Butterly, R. (2008). Season-to-season variations of physiological fitness within a squad of professional male soccer players. **Journal of Sports Science and Medicine** 7, 157-165.

Crawley, M. J. (2007). **The R book.** Chichester: J.Wiley.

Doolittle, B. & Pelton, K. (2009). **Pro Basketball Prospectus 2009-10**. Prospectus Entertainment Ventures LLC.

Eldirdge, R. (2010). **Measuring Efficiency in the National Basketball Association: A "Moneyball" Approach.** Thesis. New York University Leonard N. Stern School of Business

Esteller-Moré, A., & Eres-García, M. (2002). A note on consistent players' valuation. **Journal of Sports Economics,** 3 (4), 354-360.

Hausman, J. A. & Taylor, W. E. (1981). Panel data and unobservable individual effects. **Econometrica** 49 (6), 1377-1398.

Hollinger, J. (2005). **Pro Basketball Forecast**. Washington, D.C.: Potomac, Inc.

Koning, R. (2010). Home advantage in professional tennis. **Journal of Sports Sciences** 29 (1), 19-27.

Kubatko, J., Oliver, D., Pelton, K, & Rosenbaum, D. T. (2007). A starting point for analyzing basketball statistics. **Journal of Quantitative Analysis in Sports** 3 (3), Article 1.

Lewis, M. M. (2003) **Moneyball: The art of winning an unfair game**. W.W. Norton & Company Inc.

Mallios, W. S. (2011). **Forecasting in Financial and Sports Gambling Markets: Adaptive Drift Modeling.** Wiley John Wiley & Sons, Inc.

Martínez, J. A. & Martínez, L. (2010). El uso de indicadores de desempeño normalizados para la valoración de jugadores: El caso de las estadísticas por minuto en baloncesto. **Motricidad. European Journal of Human Movement** 24, 39-62.

Martínez, J. A. (2012). Entrenador nuevo, ¿victoria segura? Evidencia en baloncesto /Chaning a coach, guarantee the win? **Revista Internacional de Medicina y Ciencias de la Actividad Física y el Deporte** 12 (48), 663-679.

Metaxas, T., Sendelides, T., Koutlianos, N., & Mandroukas, K. (2006). Seasonal variation of aerobic performance in soccer players according to positional role. **Journal of Sports Medicine and Physical Fitness** 46, 520-525.

Oliver, D. (2004). **Basketball on paper. Rules and tools for performance analysis**. Washington, D. C.: Brassey's, INC.

Pelton, K. (2010), Rethinking NBA aging. Retrieved from http://basketballprospectus.com/article.php?articleid=896

Piette, J., Sathyanarayan, A. & Kai, Z. (2010). Scoring and shooting abilities of NBA players. **Journal of Quantitative Analysis in Sports** 6 (1), Article 1.

Pinheiro, J, C., & Bates, D. M. (2000). **Mixed – effects models in S and S-Plus**, Springer, New York.

Rabe-Hesketh, S., & Skrondal, A. (2012). **Multilevel and Longitudinal Modeling using Stata (Third Edition)**. College Station, TX: Stata Press.

Reed, D. & O'Donoghue, P. G. (2005), Development and application of computer-based prediction methods. **International Journal of Performance Analysis of Sport (e)** 5 (3), 12-28

Sampaio, J., Drikwater, E. J. & Leite, N. M. (2010). Effects of season period, team quality, and playing time on basketball players' game-related statistics. **European Journal of Sport Science** 10 (2), 141-149.

Schafer, J. L., &. Yucel, R. M. (2002). Computational Strategies for Multivariate Linear Mixed-Effects Models with Missing Values. **Journal of Computational and Graphical Statistics** 11 (2), 437–457.

Verbecke, G., & Molenberghs, G. (2000). **Linear mixed models for longitudinal data.** New York: Springer.

Winston, W. L. (2009). **Mathletics**. New Yersey: Princeton University Press

Wooldridge. J. M. (2002). **Introductory econometrics: A modern approach**. Thomspon

Acknowledgements

Jose A. Martínez has been supported by the following research grant: ECO2012-36032-C03-03 from the Ministerio de Economía y Competitividad del Reino de España.

We are in debt to Ángel López, M. Luz Mate, Josep Anton Sánchez and Mariona Salvador for their helpful suggestion in previous versions of this manuscript.

Correspondence

Martí Casals.
Address: Josep Trueta s/n – 08195 Sant Cugat del Vallès (Barcelona), Spain.
E-mail: marticasals@gmail.com