**Bayesian hierarchical modelling for performance forecasting in NBA basketball, with Markov chain Monte Carlo sampling**

Arjun Sahota

Contents

[Introduction 3](#_Toc196157810)

[Data 4](#_Toc196157811)

[Exploratory Data Analysis 4](#_Toc196157812)

[Population histograms for BPM 4](#_Toc196157813)

[Population scatterplot BPM against Age 5](#_Toc196157814)

[Boxplot of team BPMs 5](#_Toc196157815)

[Within season team BPMS 5](#_Toc196157816)

[Plotting how BPM changes with age for specific players 6](#_Toc196157817)

[Constructing the Model 7](#_Toc196157818)

[Linear Single Level Model 7](#_Toc196157819)

[Linear Hierarchical Model 7](#_Toc196157820)

[Bayesian Hierarchical Model 8](#_Toc196157821)

[Priorsimulationplot 9](#_Toc196157822)

[Quantile**-**Quantileplot 9](#_Toc196157823)

[Modellingforplayers 9](#_Toc196157824)

[Forecasting 10](#_Toc196157825)

[Conclusion 10](#_Toc196157826)

# Introduction

At the elite level of sport, recordings of every measurable attribute of performance are being constantly made, and massive databases constantly expanded. The American National Basketball Association (NBA) is no different and is perhaps one of the sports where coverage is most extensive. Every possession is short, with only a few possible outcomes which can end each, and begin a possession for the opposing team. Therefore, every possession will have a few attached statistics, which are carefully analysed to understand the value of each player action and each team composition. Basketball-reference collates data on each point, assist, rebound, block, and steal, then further manipulates the data to create advanced statistics, which provide further insight into player performance.

Advanced statistics are a key point of contention in the league, as fans and analysts argue over which most effectively encapsulates the performance of a player. Two of the most popular metrics are Win Shares per 48 (WS/48) and Box Plus/Minus (BPM). WS is calculated by taking the points created and denied by a player, scaling by how often they play, comparing to the rest of league and estimating how many wins this performance would result in for a regular season (WS/48 is per 48 minutes). BPM aims to show how many points a player contributes above league average per 100 possessions through box scores. The methodology for calculating BPM is as follows:

1. **Estimate player position by running box scores through a regression formula**
2. Positions are generally increasing in size and decreasing in assists as their number increases (1 is a point guard, 5 a centre). For example, % of team rebounding has a coefficient of 8.7.
3. **Based on role, generate coefficients based on player statistics (points, reb, etc.)**
4. **Calculate team-average points per shot, adjust player points accordingly**
5. Increase if points per shot is higher than the baseline used in BPM calculation
6. **Calculate Raw BPM, adjust so that player BPMs sum to the team’s net rating**
7. Net rating is (points scored – allowed) / 100 possessions. This final adjustment means that player BPMs (net points) sum to that of the team.

For this analysis, we used BPM, as it better represents the individual’s contribution. While Win Shares is an effective metric, it can be meaningfully influenced by team performance, we are focussing exclusively on the individual. However, BPM has its own limitations, as the statistic is disproportionately weighted on offensive production. Indeed, a player’s defensive impact can be difficult to measure, but one might contend that teams rely on strong offensive performances to win in today’s NBA. With the three-point revolution in full swing, defences are increasingly more spaced out, so a good offense will often find a way through.

Regarding our approach to modelling, there are three key points to the model: hierarchy, Bayesian coefficients, and Markov chain Monte Carlo (MCMC) sampling. We now proceed with a discussion of each point. For a more comprehensive introduction to these concepts, the reader is referred to a useful tutorial article by Veenman, Stefan & Haaf (2023).

Hierarchical modelling structures relations between predictors by nesting them within each other. Consequently, one predictors effect may be shifted by the value of another. Examples include modelling student scores while considering which school they attend. Here, the population of students across all schools will have some distribution of scores, and students within each school will have an additional effect from their school.

Bayesian statistics introduces the idea of prior knowledge, which, in combination with new observations, brings us to a final (posterior) distribution of data. A classic example is a comparison of a musical expert and a drunk man. If the expert gets the composer of a piece right 9 times out of 10 (between two choices) and the drunk gets the outcome of a coin flip right 9 times out of 10, a frequentist says they are as skilled as each other. In contrast, a Bayesian will assign prior distributions such that the drunk’s chance to guess a single flip correctly is around 50%, while that of the expert is higher. After observing the outcome of the 10 trials, the likelihood for each outcome will be updated (higher), but the expert’s chance to be correct in a trial will still be higher.

Finally, MCMC algorithms draw samples from likelihood informed posterior distributions attempting to converge to a target distribution. Often the target is not easily modelled, so we formulate it piecewise – looking at how probability behaves in small sections of the distribution. A Markov chain determines a sample based on the previous sample. Thus, MCMC begins from priors and iterates based on projected (stationary) posteriors to map a distribution. Hamiltonian MCMC (used by Stan) utilises information from the gradient to find proposals more efficiently.

Data

As mentioned, data scraped from basketball-reference forms the basis of this analysis. We consider 20 years of NBA history, from 2004-05 to the most recent 2024-25 season. Prior to the 2004-05 season, the NBA implemented rule changes, banning hand-checking in certain situations and introducing the three-second rule. As a result, players had fewer defensive options and greater freedom on offense. We analyse, then, every season after this change, for continuity in the league’s rules. An alternative timeframe would be 2014-2025, as 3-point shots became much more popular throughout this period.

Our data consists of player-team-season observations (e.g. LeBron James on the Miami Heat in the 2011). We focus on age as a predictor for BPM, a hierarchical model allows us to vary the effect of age between players: a player who is injury prone will not age as well as a fit player who practices frequently, for instance. Further, a Bayesian approach allows us to incorporate our beliefs/knowledge of the data through prior distributions. MCMC helps us model our final distribution, as it has a complex distribution.

To ignore players with little data, we removed all observations where a player played less than 400 minutes for a team (an average of 5 minutes per game). Season averages for players who were on multiple teams were not considered. Additionally, observations with team names that no longer exist were changed (e.g. Bobcats -> Hornets). Finally, age was centred to be mean 0.

Exploratory Data Analysis

A graph of a number of colored bars

AI-generated content may be incorrect.A graph of a red pyramid

AI-generated content may be incorrect.Population histograms for BPM

Figure 2

Figure 1

To begin this section, two crucial histograms. On the left is the observed density of BPM scores for data used in this analysis (figure 1). The right-hand plot uses data from 1974-2019, provided by basketball-reference (figure 2). They define the mean BPM to be 0, however, our empirical result was -0.487. The distributions are similar, with a peak between -1 and 0, and a slight positive skew. From this, we can already begin to consider viable priors for intercept and variance, as well as the overall distribution, which looks like a skewed normal.

## Population scatterplot BPM against Age

A graph with lines and dots

AI-generated content may be incorrect.Another plot which is essential to our understanding of modelling BPM is this scatterplot (figure 3). It displays that, from 2004-2025, the average 27-year-old player had a BPM around 2.5 higher than an average 19-year-old. This growth seems logarithmic on inspection. Looking at the next period, average BPM slowly declines before settling at -0.25 for 35-year-old players. Finally, performance increases slightly until the average player in their early forties has a BPM of 0. Clearly, this final section is less certain than the other trends. To look further into the distribution of BPM within age groups, see the histograms in the attached Rmd. file.

Figure 3

## A graph with lines and dots AI-generated content may be incorrect.Boxplot of team BPMs

Another useful predictor will be what team a player is on. Over the last 20 seasons, the average Wizards player (WAS) had a BPM of -2.2, while an average Celtics player (BOS) had a BPM of 0. This plot also demonstrates the number of star players that a team has had in the last 20 years, with the Lakers (LAL) and Warrior (GSW) leading (of course!). The Lakers have had over 15 player-seasons with BPM above the team’s normal values (think Kobe, LeBron), while the Knicks (NYK) have had only two. Finally, looking at the Celtics again, we see that they have had a large range of talent, resulting in an interquartile range of ~4.6; so, while they only have one anomalous player-season, this can be attributed to high IQR. If Boston had the same quartiles as the Lakers, we would say that eight of their player-seasons were well above average. All teams have relatively few abnormally bad entries, with their boxplot’s lower tails ranging from -7.5 to -5.

Figure 4

A graph with lines and letters

AI-generated content may be incorrect.A graph with lines and dots

AI-generated content may be incorrect.Within season team BPM

Figure 6

Figure 5

While there is little variance between the distribution of BPM league-wide across seasons, if we look within each season, we can see that teams vary quite significantly, when compared to others in the same season, and themselves in other seasons. Observe how Cleveland (CLE) went from a bottom two team by BPM (median player BPM of -2) to top three (median 1) in only five years. For the 2019-20 season, team median BPM ranged from around -2.7 to +1.4, from 2004-2025, the medians ranged from -1.4 to 0.

## Plotting how BPM changes with age for specific players

**A graph with a line

AI-generated content may be incorrect.A graph with a line graph and numbers

AI-generated content may be incorrect.**Over the page, our final plot demonstrates that, while the general trend is to improve performance until peaking peaking at age 27, slowly decline and then flatten out around 35, most players do not exhibit this pattern. There are huge variations in the paths of different players. Shown above, LeBron James has been well above average for 20 years, despite a slow decline; Tyson Chandler struggled in his early career before a Renaissance around 30; Kawhi Leonard saw massive improvement from 20 to 25, then declined just as quickly; Alex Caruso started playing better and better every year after 27. Additionally, the points around the intercept for each player (around a1 = 0, age = 26 or 27) have very different scores: Kawhi’s BPM at age 27 was 7.2, while Harden’s was 8.7, even with them experiencing similar effects of age on their BPM.

A graph of different age groups

AI-generated content may be incorrect.

Figure 7

# Constructing the Model

Having thoroughly examined the data, we list the following takeaways:

1. The effect of age on BPM seems logarithmic, ranging from -2.5 at 19, to approximately 0 27-40. However, the age effect is significantly different between individuals. It increases at first, then decreases.
2. Season has a negligible effect on BPM. Within each season, though, each team has a very different distribution of BPM scores.
3. Compared to each other, teams have different distributions of average BPMs across 2004-2025. Despite these average effects, one team’s distribution can vary between seasons.

## Linear Single Level Model

Now, to create a model based on these findings, we start with a simple linear model, incorporating the three takeaways. Because age is centred, the lowest possible value in the NBA (given mean ~26.6) would be -9 for an 18-year-old athlete. Hence, to avoid taking the logarithm of a negative number, we add 9 before creating the logarithmic relationship (all logs are base e). We add an age squared term to let the function decrease with age at some point. Season and team are interacted so that each team can have a different mean based on the season:

BPM ~ log( age + 9 ) + log( age ^ 2 ) + season \* team

For our linear models, the key plot will be the Q-Q plot, displaying how residuals behave based on the theoretical quantile of fitted data. For example, the quarter of data with the highest fitted values will be graphed between 2 and 4 on the x-axis, with 4 being the highest value. This plot shows that, as predictions increase, standardised residuals (standardised difference between A graph with a line

AI-generated content may be incorrect.fitted value and true value) increase disproportionately. Clearly, the model not a good fit for players who are well above average, they will have BPM estimates which are far too low, as shown by the residual plot.

Figure 8

## 

## Linear Hierarchical Model

Using a similar structure, we now allow the age effect to vary for each player. Within the model, this means that we decompose the effect of age into a population term and a player term (Griffin et al., 2022). Also, or interaction term is now separated into a random team effect (1 | team) and a random interaction term season (1 | team / season):

BPM ~ log( age + 9 ) + log( age^2 ) + ( log( age + 9 ) + log(age^2) | player) + (1 | team / season)

A graph with a line

AI-generated content may be incorrect.Here, the Q-Q plot (figure 9) handles the top end of athletes much better, though it still struggles for the most extreme observations. However, the lower end of fitted values has also become a problem. The residual plot is much better than the single level model (range of -6 to 6 compared to -7 to 14), but there are still improvements to be made.

Figure 9

## Bayesian Hierarchical Model

In fitting a Bayesian model, prior selection is instrumental in determining the reliability and usefulness of results. To this end, much time was invested in finding reasonable priors for the distribution of each coefficient in the final model. The posterior model was derived using MCMC.

Only one change was made to the formula from the linear hierarchical model: to maintain a level of parsimony, the interaction between team and season was reduced to one random effect for each team, across all seasons:

BPMi​ = β0​ + β1​ \* log( a1,i​ + 9 ) + β2 \* ​log( a2,i ​) + b0j[i]​ + b1j[i] \* ​log( a1,i​ + 9 )

+ b2j[i] \* ​log( a2,i​ ) + uk[i]​ + εi

Where: βm are coefficients we determined priors for ; a1,i  is the age of the i-th player-team-season observation ; a2,i is a1,i^2 ; bmj[i] are random effects of player j applied on their observation i ; uk[i] is the random effect of team k applied on their observation i ; εi is random noise.

A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

AI-generated content may be incorrect.To the left is a comparison between the formula used to form the basis of our model’s fixed effect priors, and the true average effect of age on BPM (figure 10). Multiplying the log(age) term by 0.8, the log of squared age by -0.2 and determining intercept as -1.8 gives a function which follows the true average value closely. It starts at -2.5, rises to around 0 by the average age, then flattens out just below 0. It is possible to emulate the final increase in average BPM by using a cubic term, but, again, for a parsimonious model, we forgo this option. We select reasonably small variance terms, as our constructed formula follows the mean.

Figure 10

Thus, our priors: β0 ~ N(-1.8, 1) ; β1 ~ N(0.8, 0.1) ; β2 ~ N(-0.2, 0.03) ; εi  ~ | N(1, 0.5) |

β0 is the population intercept: with 0 centred-age (a1) and no player-age or team effect, BPM is, in an average sample, -1.8 + 0.8 \* log(9) ~ 0. Crucially, a1 is never 0: at age 26, a1 = -0.55, at age 27, a1 = 0.45. So, BPM of the average 27-year-old player, with no random effects, is 0.32.

Next, we decide on a regularisation factor for the player level effects of age. We set this constant at 1.2, meaning that unexpected observations are strongly shifted towards the average, unless there is clear evidence that a player experiences a different effect of age. This is done by shrinking the variance and correlations of player level effects - the model is kept from overfitting but maintains a level of flexibility in modelling the best and worst players.

### **A graph of a normal distribution AI-generated content may be incorrect.Prior simulation plot**

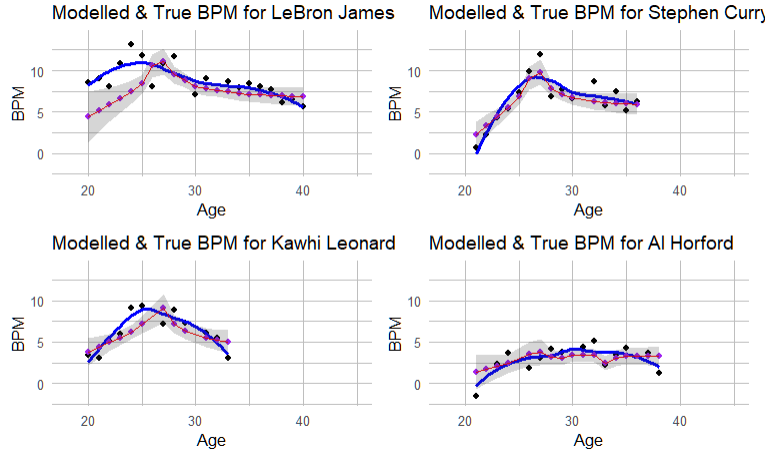
This plot uses our priors to simulate data and compares it to the true distribution (y). We see that the shape of a sample set of simulated data is generally acceptable (figure 11). There are some trials with variance too high, and some with peaks in their density that are too high, nonetheless, clearly the priors work effectively with our data.

Figure 11

### **A graph of a graph AI-generated content may be incorrect.Quantile-Quantile plot**

The Bayesian Model Q-Q has similar characteristics to the linear hierarchical model, but it performs slightly better, staying closer to a normal distribution, at the top and bottom ends (figure 12).

Figure 12

**Modelling for players**

Here we show selected results from our model (red) with a 95% confidence interval (grey) overlayed with the true values (blue) (figure 13). We see that the model is flexible enough for a variety of shapes, from the (relatively) orthodox Steph Curry to the more abnormal Kawhi and Horford. Also demonstrated is one of the key limitations of the model – it struggles with high values before the intercept (~ age 27). This is really an issue with the formula, because, while a1 is less than 0 (equivalently age < 27), there is almost guaranteed to be a positive gradient for BPM w.r.t. a1, thanks to a combination of the coefficient of log(a1 + 9) being positive, log(a2 ^ 2) shrinking, and β2 being less than 0. Lebron is already in his third year by the time he’s 20 and has already gone through much of his developmental phase, but the model predicts as though he is a rookie. A similar problem arises for players whose first entries have above average BPM for their age and who then fall into line with the population. As their player effects are minimal, the model will completely ignore the abnormal values, so that the predictions for later years are a better fit.

Figure 13

### **Forecasting**

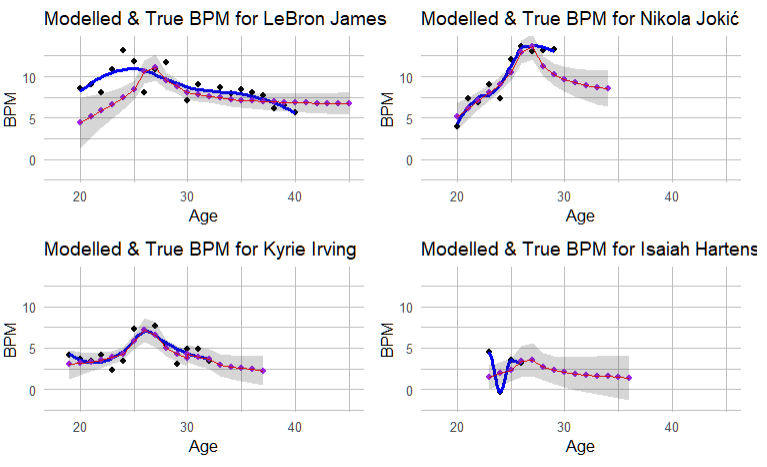
Forecasting using our model yields promising results: the fits for players like Jokić and Kyrie follow reasonable patterns of a slow decline with age, matched to observed data (figure 14). We see that Jokic has a forecasted BPM of ~8.5 when he is 34, compared to the observed value of 13.3 today. This represents a decrease of production akin to almost 5 points per 100 possessions. Certain caveats must be mentioned, however. This model should not be used to predict performance for players past 40, as there is too little data to accurately fit the model; also, with our current formula, a player’s predicted BPM won’t change must past their fortieth birthday, as can be seen by LeBron’s plot. Also, it is fairly argued that our regularisation should not have been so tight: Hartenstein’s plot is typical of any young player, since there isn’t yet enough evidence to determine whether their age effect is unique. Even so, Hartenstein is predicted a much higher intercept (3.5) than an average player (~0), so there is certainly useful description of his specific performance.

Figure 14

# Conclusion

In this article, we discussed the use of statistics to model the performance of basketball players. We determined BPM as the best single metric for performance and conducted an analysis of data from NBA players from the 2004-05 to 2024-25 seasons. Using a variety of plots, we explored the relationship between BPM and our three key predictors: age, team and season. Finally, this knowledge informed the fitting of three regression models: linear single level, linear hierarchical and, most importantly, Bayesian hierarchical.

Through MCMC, we developed a solid Bayesian hierarchical model for how NBA players’ performance varies with age. With sensible priors derived from careful examination of the data, our model gives a reasonable fit to all players tested. It is not without room for improvement, however, as issues persist for modelling athletes with high first year BPM, and of predicting into a player’s forties. Even so, the model can robustly generate predictions for most players in the NBA.

References:

basketball-reference. *Basketball database*. <https://www.basketball-reference.com/about/bpm2.html>

Griffin, Jim E., Hinoveanu, Laurenţiu C. and Hopker, James G.. "Bayesian modelling of elite sporting performance with large databases" *Journal of Quantitative Analysis in Sports*, vol. 18, no. 4, 2022, pp. 253-268. <https://doi.org/10.1515/jqas-2021-0112>

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. Psychonomic Bulletin & Review, 12(4), 573–604. <https://doi.org/10.3758/BF03196750>

Veenman, M., Stefan, AM., & Haaf, J.M. (2023). *Bayesian hierarchical modeling: an introduction and reassessment*. *Behav Res* **56**, 4600–4631 (2024). <https://doi.org/10.3758/s13428-023-02204-3>