BAYESIAN BASKETBALL

A PREPRINT

Ben Brennan
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109
brennben@umich.edu

Jieru Shi
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109
herashi@umich.edu

December 9, 2019

ABSTRACT

The recent rise and availability of big data in sports has changed the way owners and mangers can - and should - operate their franchises. In this project, we take the data of 256 NBA basketball players from the 2013-2014 season to the 2016-2017 season and use that data to make informed decisions regarding player section. First, we develop a loss function that we posit provides managers the opportunity to make informed decisions in the light of relevant information. We then fit a Bayesian Linear regression as a way to provide the necessary information to make this decision. Lastly, we briefly describe Gaussian Processes and provide an example of how they can be used to help predict future performance and how we can use them with respect to our loss function. Through these methods and the use of our loss function, we present two separate possible starting line ups that are deemed ideal.

1 Introduction

Every year, NBA franchises are offered the opportunity to buy, sell and trade players in order to improve their team. Whilst there are other factors to consider (money, personality fit, etc.), putting together a team of highly skilled players will always be important. The process we hope to describe will allow NBA team to make informed decisions when presented with this opportunity, although likely for a smaller subset of available players. For instance, the methods could be applied to a subset of players who are free agents or are at the end of their contracts in order to decide who is the best choice to pursue at a given position. In our project, we are lucky enough to have data for all players in the league. The data contains the demographic physical and historical performance attributes of 264 national basketball association players, spanning from the 2013-2014 season to the 2016-2017 season. We chose Points Per Game (PPG), Free Throw Percentage (FT%) and DRAYMOND [1] score as outcomes of interest as we felt they were a good proxy for overall performance.

2 Methods

2.1 Loss Function

The idea behind the loss function we present is as follows: we want a well-rounded team, but we do not want to compromise when it comes to not selecting the best players. This is in line with current thought on how to select players. The best players, while they may have some shortcomings, are pursued to a much greater extent than players with more average performance stats who may very well be more well-rounded players. To introduce our loss function, we must introduce a bit of notation. Let y_i^* (where $i \in \{1, \dots, 256\}$) be the predicted value of a certain metric for one player. Then, each player has a predicted rank $p_i^* \geq 1$ ($p_i^* \in \mathbb{N}$) for that outcome, associated with the magnitude of y_i^*. The ideal predicted rank is 1, so an ideal player has three predicted ranks of 1. Therefore, the ideal $\mathbf{p}^*$ is $< 1, 1, 1 >= \mathbf{1}$.

A natural idea for the loss function is then

$$\mathcal{L} = ||\mathbf{p}^* - \mathbf{1}||_2$$

however, this weights the difference in loss between selecting say, the best free throw shooter over the second best free throw shooter equally to the loss between selecting the 100th best free throw shooter over the 101st best. Intuitively, we would like to say that the difference in loss between selecting the first setting is higher than the second. So, if we add a monotone increasing, concave penalty term to the loss function, we can achieve that goal. We chose to use a loss function of the form

$$\mathcal{L} = ||\mathbf{p}^* - \mathbf{1}||_2 + \lambda|| \log(\mathbf{p}^*)||_1$$

where λ is a tune-able parameter that reflects the level that one favors players with better single ranks. Any function $f(x)$ such that $f'(x) > 0$ and $f''(x) < 0$ will achieve this goal, along with any norm. We chose the Manhattan norm so that each parameter's penalty was considered equally, and $f(x) = \log(x)$ for simplicity.

2.2 Bayesian Linear Regression

Now, we fit a simple Bayesian Linear Regression model [2] of the form:

$$Y_{i,j} = \alpha_i + \beta_i t_j + \epsilon_{i,j}$$

where $\epsilon \sim \mathcal{N}(0, \tau^{-1}), \tau \sim Gamma(0.01, 0.01)$ and α and β are outcome specific. Specifically,

- $\alpha \sim \mathcal{N}(0, 4), \beta \sim \mathcal{N}(0, 4)$ for DRAYMOND
- $\alpha \sim \mathcal{N}(12, 4), \beta \sim \mathcal{N}(0, 10)$ for PPG
- $\alpha \sim \mathcal{N}(0.7, 0.01), \beta \sim \mathcal{N}(0, 0.05)$ for FT%

These are subjective priors, chosen to reflect our prior knowledge of the metric and how variable it tends to be amongst players. Above, t is a measure of the season, coded as 14,15,16, and 17 for the 2013-2014, 2014-2015, 2015-2016 and 2016-2017 season, respectively.

We used rstan [3] for this analysis, obtaining 4,000 draws from the posterior of each player's α and each player's β. Then, setting t equal to 18, we obtained an approximation of

$$\pi(y^*|X, x, y_1, \ldots y_n)$$

which is the posterior predictive distribution of next season's performance in the given outcomes. Figure 1 shows these estimated predictive distributions for three well-known players - Jimmy Butler, Kevin Durant, and LeBron James.

While we are aware of the limitations of modeling FT% and PPG using linear regression as FT% $\in [0, 1]$ and PPG $\in [0, \infty)$, we decided to use this method anyways as we were not attempting to predict the correct metric next season, but rather the correct ranking of players with respect to that metric. If one wanted to predict the metric accurately, we believe that modeling PPG on the log scale and FT% on the log-negative-log scale - and then transforming back - will yield accurate results.

Let $\alpha_{i,j,k}$ and $\beta_{i,j,k}$ be the j-th draw ($j \in \{1, \ldots 4000\}$) from the posterior generated by rstan for a certain individual $i \in \{1, \ldots, 256\}$ associated with one of the three metrics $k \in \{1, 2, 3\}$ of interest. Then, for each draw, we can obtain $p^*_{j,i,k}$, the rank of a player at that draw by $\alpha_{i,j,k}$ + $18\beta_{i,j,k}$ and then comparing all players for metric k. We compute the average posterior rank of a player over the 4000 draws as

$$\sum_{j=1}^{4000} \frac{p^*_{j,i,k}}{4000}$$

for all players i and all outcomes of interest k. Using a player's average posterior rank and the loss function presented above, we select the following players given our posterior predictive distributions for DRAYMOND, PPG and FT% :

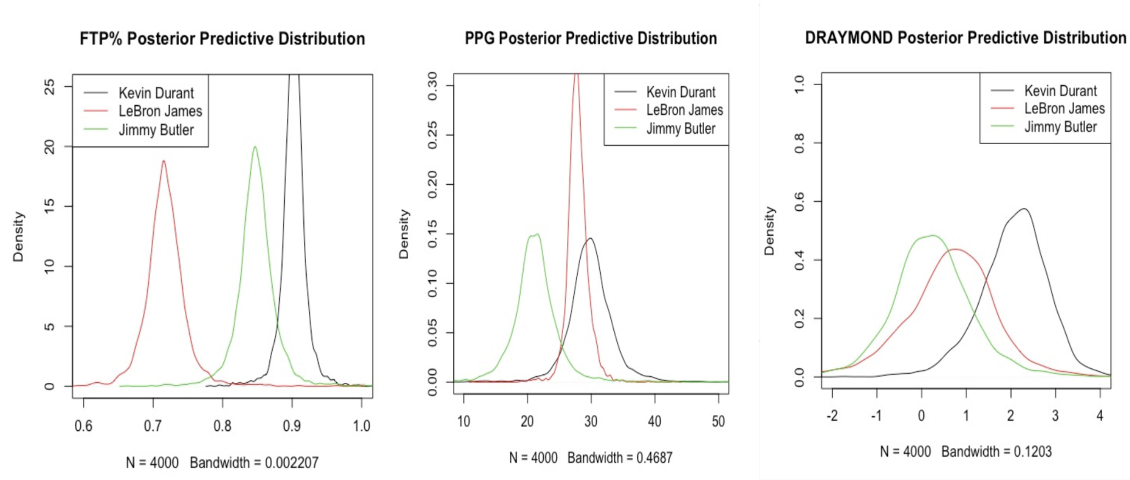

Figure 1: Estimated Posterior Predictive Distributions

- Brook Lopez
- Anthony Davis
- Kevin Durant
- Klay Thompson
- Steph Curry

which, from a subjective point of view, is a highly accurate and efficient group of players.

2.3 Gaussian Process Regression

2.3.1 Gaussian Process

Despite the fact that the Bayesian linear regression gave good predictions and picked an ideal team, it limits the model to a linear case. In this part we will extend it to a non-parametric model and use physical characteristics as inputs (instead of just time) to diversify the analysis. [4]

Instead of defining prior distributions on α and β, we will treat $f(x)$ as a whole and define a prior distribution over possible forms of $f(x)$.[5] Since Gaussian processes let us describe probability distributions over functions we can use Bayes' rule to update our distribution of functions by adding new observations.

2.3.2 Assumptions

Y represents the actual DRAYMOND scores from training set, and

$$Y = f(X) + \epsilon, \;\; f(X) \sim GP(m(X), \mathcal{K}) \;\; \epsilon \sim (0, \sigma^2) \tag{1}$$

where

$$m(X) = (1 \;\; X)\beta \tag{2}$$

$$\mathcal{K} = \begin{pmatrix} k(x_1, x_1) & k(x_1, x_1) & \cdots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \cdots & k(x_2, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n, x_1) & k(x_n, x_2) & \cdots & k(x_n, x_n) \end{pmatrix} \tag{3}$$

and each entry of the kernel matrix is defined as:

$$k(x_i, x_j) = \exp(-\frac{(x_i - x_j)^2}{2\gamma^2}) \tag{4}$$

A time continuous stochastic process $\{X_t; t \in T\}$ is Gaussian if and only if for every finite set of indices $t_1, \ldots, t_k$ in the index set T

$$\mathbf{X}_{t_1,\ldots,t_k} = (X_{t_1}, \ldots, X_{t_k})$$

is a multivariate Gaussian random variable. Therefore, we can write the joint distribution of observed target value $y(\mathbf{X})$ and function value at test point x^* as:

$$\begin{pmatrix} y(\mathbf{X}) \\ f(x^*) \end{pmatrix} \sim N\left(0, \begin{bmatrix} \mathcal{K} + \sigma^2 I & k(x^*, \mathbf{x}) \\ k(x^*, \mathbf{x})^T & k(x^*, x^*) \end{bmatrix}\right) \tag{5}$$

Which leads to the conditional distribution of $f(x^*)$ [6]:

$$f(x^*)|y(\mathbf{X}) \sim N(k(x^*, \mathbf{X})^T(\mathcal{K} + \sigma^2 I)^{-1} y(\mathbf{X}), \sigma^2 + k(x^*, x^*) + k(x^*, \mathbf{x})^T(\mathcal{K} + \sigma^2 I)^{-1} k(x^*, \mathbf{X})) \tag{6}$$

2.3.3 Predictions

By using the first three seasons as training data, and the last season as testing data, we obtained 4000 draws (1000 burn-in) from the posterior distribution of each test point [7], and here we present the prediction results as follows:

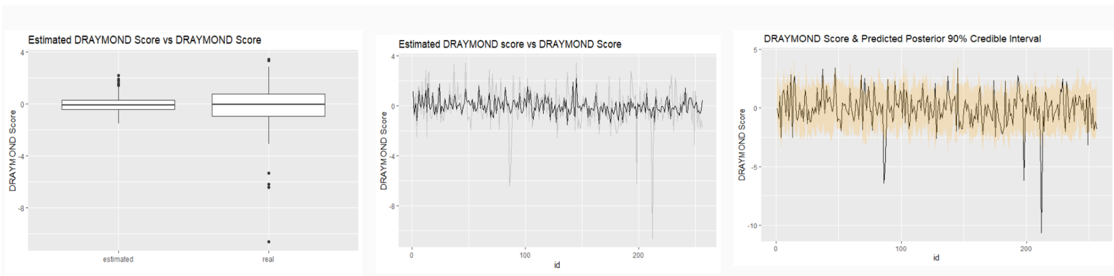

Figure 2: Box plot, estimated value and 90% CI for posterior mean of DRAYMOND Score

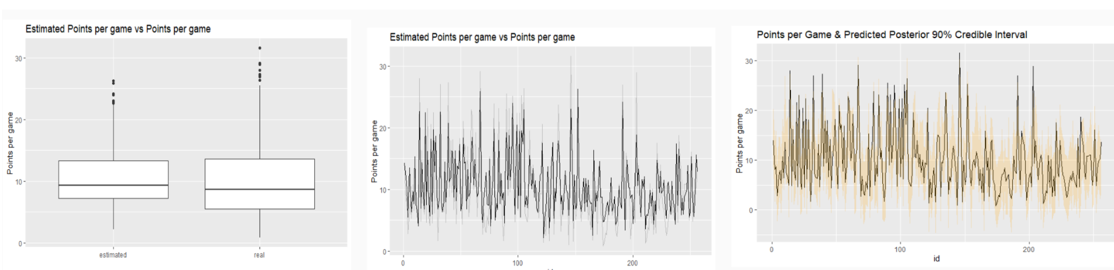

Figure 3: Box plot, estimated value and 90% CI for posterior mean of PPG

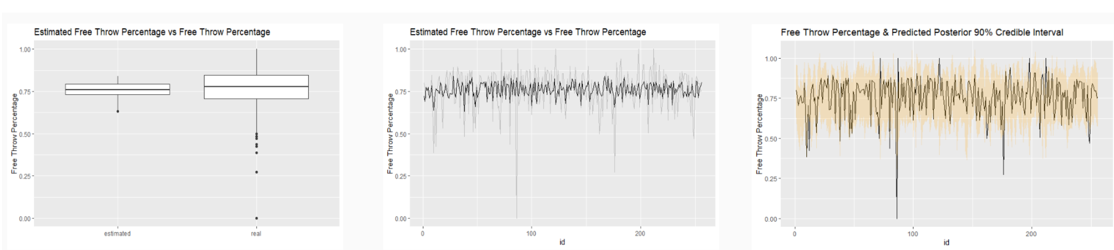

Figure 4: Box plot, estimated value and 90% CI for posterior mean of FT%

The first box plot shows the distribution of last season's predicted DRAYMOND Scores, PPG and FT% compared to the real records. From the plots we can see in general our predictions did a great job estimating the locations of the distributions but performed relatively conservative in terms of capturing the variances.

The black line in the second plot shows the predictive posterior mean of the last season's DRAYMOND Scores, PPG and FT% and the grey line indicates the real records. Worth noticing, the estimated PPG aligns with the real records fairly well, and the other two also successfully captured the overall patterns of DRAYMOND Scores and FT%.

The last plot shows the 90% credible interval of predictive posterior mean and the real records. It's easy to see apart from some extreme values, our predicted credible intervals basically included all observations, which favors the prediction accuracy of the method we proposed.

For all players i and all outcomes of interest k. Using a player's average posterior rank and the loss function presented in **2.1**, we select the following players given our posterior predictive distributions for DRAYMOND, PPG and FT% :

- Nikola Vucevic
- Dirk Nowitzki
- Steph Curry
- Paul George
- Will Barton

which is also a highly accurate and efficient group of players.

3 Discussion

As a simple, beginning analysis, we feel the results are very promising. Despite many limitations, the algorithms are efficient computationally (as there are not many inputs) and tend to select a good group of players. Both approaches not only give us predictions, but associated uncertainty with that prediction. For instance, from the Frequentist point of view, we could achieve a prediction, but it would be a point estimate completely decided but the data, with no associated uncertainty. Using Bayesian Linear Regression, we obtain a point estimate along with uncertainty, which can be useful to managers when considering players in the context of salary. Maybe a certain player has the highest expected rank, but maybe they want to know how certain they can be that the player will achieve this rank before they pay more for this player than another one. These methods are also flexible, adding for the possibility to include other covariates in the regression and to consider different kernels for the Gaussian Process, essentially allowing for the possibility of many different types of functions to fit the data.

Despite these pros, there are significant limitations that can be improved upon in future research. First, we can see in Figures 5 - 10 (Appendix) that the real time trend, as seen by the spaghetti plots produced by the average values at the seasons, seems to be non-linear. We assumed linearity would work as none of the lines were flat, but adding in a non-linear time trend seems like it would most likely improve the efficiency of the Bayesian Linear Regression. Furthermore, the algorithms are unreliable in predicting the actual future values of the metrics, as seen by the high variance in predictions in the Gaussian Process Regression and the domain limitations in the Linear Regression setting. To deal with this, future work may want to use the log scale and the log-negative-log scale (as mentioned above) to model outcomes based on their domain. To improve the prediction accuracy of the Gaussian Process, it would be useful to have physical characteristics that vary more substantially over time - such as mobile health data - or to use a subset of players (training set) to predict another subset of player's (testing set) outcomes. This was not feasible in our case as we needed to predict the future values, meaning we needed values that we knew beforehand.

Overall, we believe that these methods are a strong starting point for future work. They provide NBA franchises with a concrete set of algorithms that can be used to make objective decisions, and to be able to consider uncertainty in those decisions.

References

[1] Nate Silver. A better way to evaluate nba defense.

[2] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[3] Ben Goodrich [cre aut] Daniel Lee [ctb] Krzysztof Sakrejda [ctb] Modrak Martin [ctb] Trustees of Columbia University [cph] Oleg Sklyar [cph] (R/cxxfunplus.R) The R Core Team [cph] (R/pairs.R R/dynGet.R) Jens Oehlschlaegel-Akiyoshi [cph] (R/pairs.R) Jiqiang Guo [aut], Jonah Gabry [aut]. rstan: R interface to stan.

[4] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[5] Oscar Knagg. An intuitive guide to gaussian processes.

[6] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[7] Robert B. Gramacy and Matt A. Taddy. tgp: Bayesian treed gaussian process models.

4 Appendix

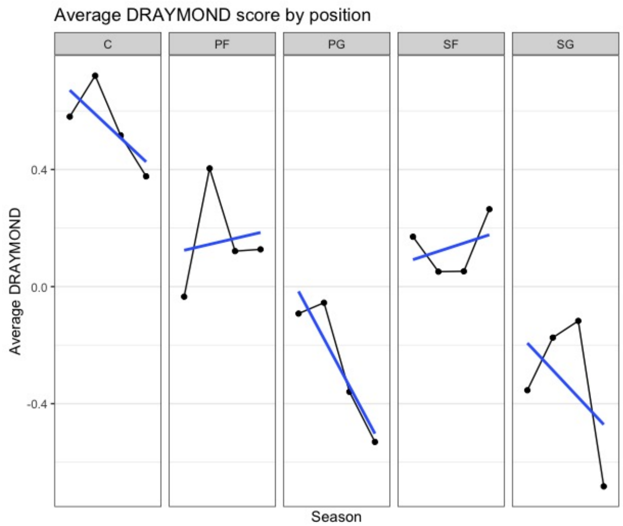

Figure 5:

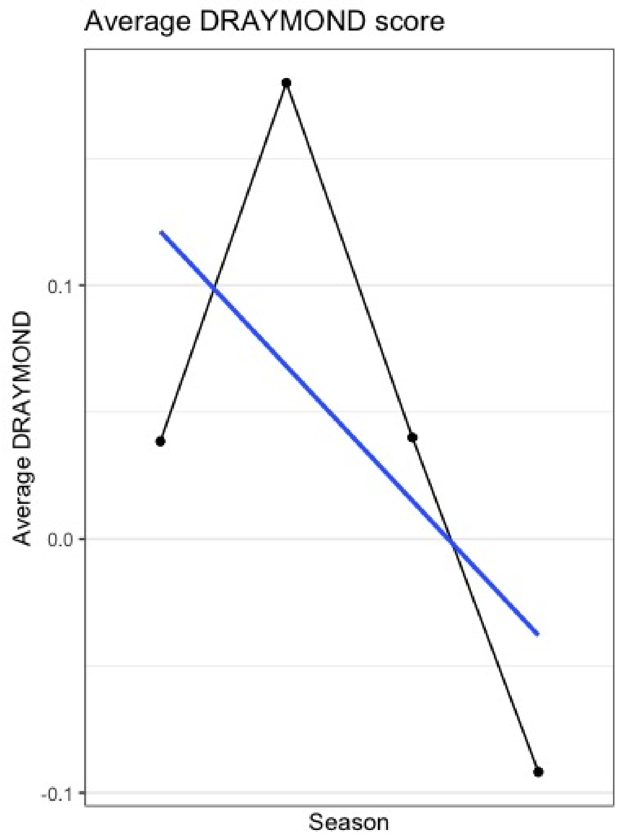

Figure 6:

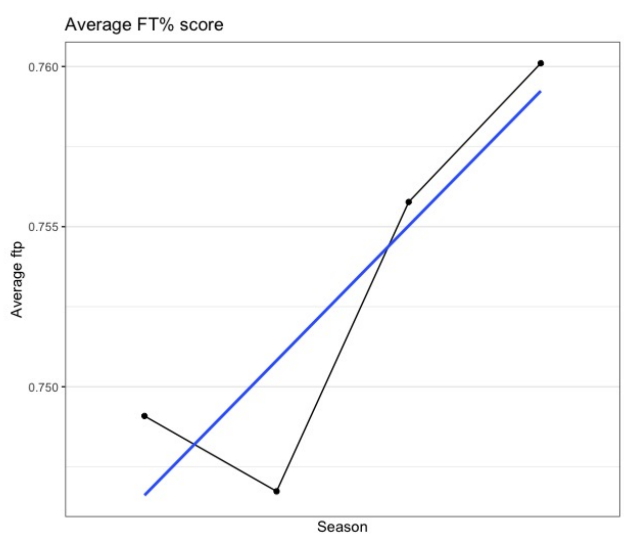

Figure 7:

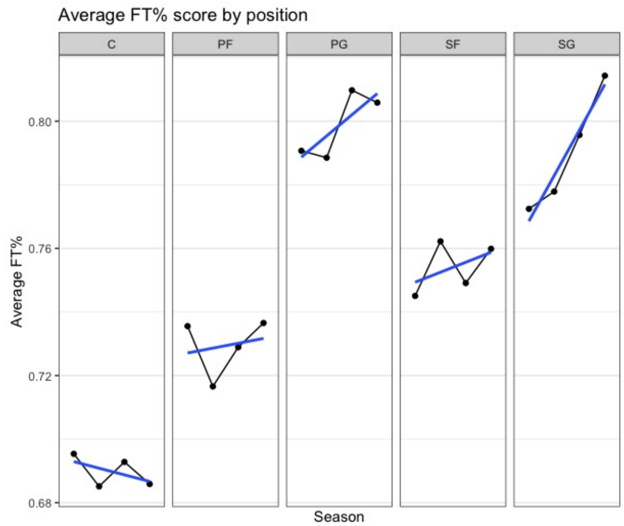

Figure 8:

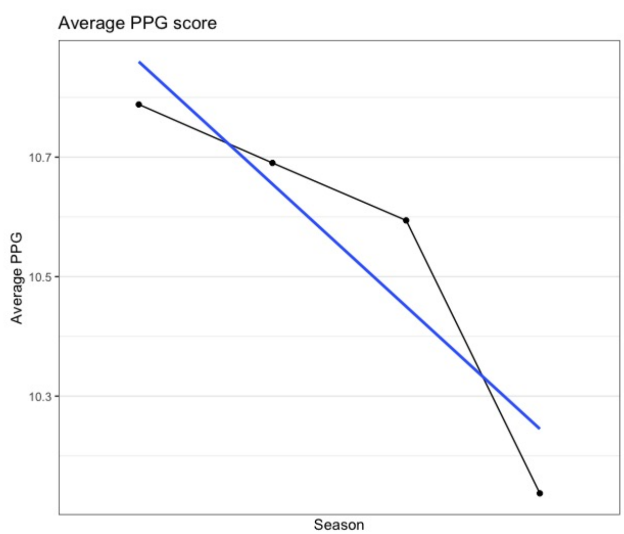

Figure 9:

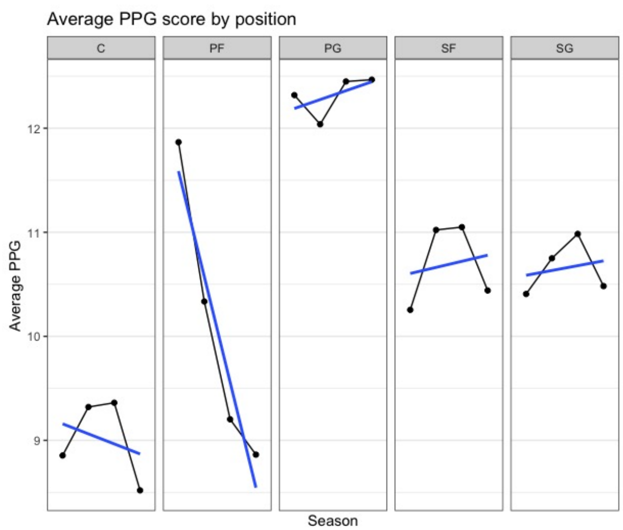
Average PPG score by position
C
PF
PG
SF
SG
12
11
10
9
Average PPG
Season

Figure 10: