Date of submission: June 8, 2022

# Multilevel Modeling to Predict NBA Salaries

Sahib Dhaliwal, Statistics, California Polytechnic State University
Gavin Martinez, Statistics, California Polytechnic University

## 1. Introduction

The National Basketball Association (NBA) is a professional basketball league composed of the best athletes on the planet. During a regular NBA season, millions of people tune in daily to watch some of the greatest players, games, and competitiveness. From its inception in 1946, the NBA has constantly been evolving throughout the years at a rapid rate. Seventy-six years ago, the salary cap for an NBA team to pay all its players was only $55,000. Meaning each player, on average, earned around $4,500 dollars for the season. Today, the NBA's salary cap is at an outlandish $112.4 million, with the average annual salary being around $7.5 million on average. This was not due only to inflation but based around the NBA's popularity and revenue.

Throughout the 2021-2022 season, the NBA generated almost 6.5 billion dollars in revenue. From this revenue, over 50% of it was paid out to NBA players' salaries. This is a record-breaking amount that led to questions regarding how it got to this point, and if the current trajectory is indicative of how salaries for NBA players will continue to grow as the years pass. Therefore, through multilevel modeling, the following research question will be investigated: Is there a significant change in the effect of season on a player's salary based on their position?

## 2. Data Source and Methods

### 2.1 Data Source

Our initial data set "NBA.csv" (https://github.com/erikgregorywebb/datasets/blob/master/nba-salaries.csv) comes from a GitHub repository from an independent user, Erik Gregory. Gregory tracked and recorded the name, position, salary, and year for every player in the NBA from 2000 to 2020. In total, this includes 9,468 observations with over 1,505 different players.

### 2.2 Data Modifications

The initial dataset had 44 observations where the team was classified as "null Unknown". These were considered incomplete observations and were removed from the data set. There were 3 observations under the team "Maccabi Haifa Maccabi Haifa", 3 observations under "Fenerbahce Ulker Fenerbahce Ulker", and 1 observation under "Bilbao Basket Bilbao Basket". These observations were all removed, as we did not recognize them as an NBA franchise. Numerous other observations had outdated team names corresponding with current NBA franchises (e.g. Vancouver Grizzlies and Memphis Grizzlies). These observations were consolidated under the current NBA franchise's team name. NBA was later subset into "nba". For "nba" we took a random sample of 200 NBA players in the original data set, including every observation for each given player with the same variables mentioned above. Our new dataset includes 1,113 observations with 200 players.

### 2.3 Factors

The variables used in our analysis include "salaryM", "name", "season.c", and "position". The variable "salaryM" is derived from NBA's original "salary" and represents an NBA player's annual salary in millions of dollars. The factor, "season.c" is a centered variable derived from the original "season" variable in NBA and represents years before and after the average season in our data set (average season is about 2011). For example, negative values will represent seasons before 2011 and positive values will represent years after 2011. The variable is centered to help with the convergence of the final model. The position variable was included with three levels: centers, guards, and forwards (C, G, F). The variable was included into the model using effect coding..

### 2.4 Modeling Methods

Moving forward, we used the subset of our original dataset to begin building a model to appropriately predict NBA salaries. We started with the unconditional growth model, using time as a fixed effect, and allowing for player-to-player variation in salaries. Then we added in position as fixed effects, comparing likelihood ratio tests along the way, as well as AIC, BIC and deviance. We also explored using random slopes to compare the effect of time on each player, as well as the use of an AR1 model for Level 1 residuals. Model assumptions and influential observations and players were checked for our final model.

### 2.5 Data Analysis

RStudio® Version *2021.09.02* "Ghost Orchid" Release, was used to conduct the data analysis. The multilevel model output was designed and tested in RStudio. In addition, RStudio was used to create all associated graphics, assessment of model fit, and parameter estimates for the model. The main packages used within R were dplyr, ggplot, magittr, and lme4. A significance level of .05 was used for all tests.

## 3. Analysis and Results

### 3.1 Final Model

A Random slopes model was fit to the data. Using the Likelihood Ratio test to assess the model fit, a p-value of <2e-16 provides evidence that this model maximizes the improvement of the fit to the original data compared to an unconditional growth model.

**Figure 3.1.1: Individual Effect of Season on Player**

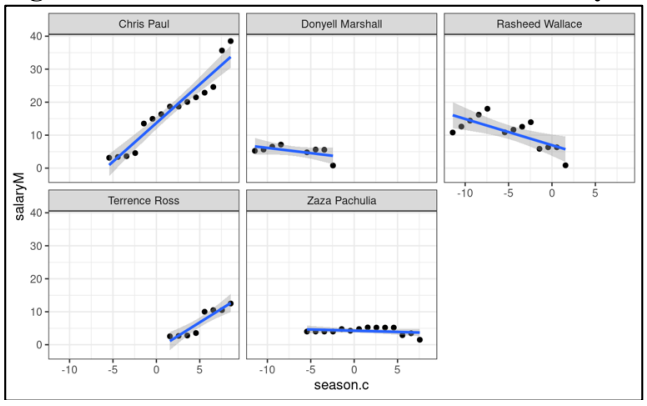



Figure 3.1.1 displays individual effects of a subsample of five players. The between player variation in estimated change in salary is relatively large and matches the estimated between player variation estimated in the final model as shown in Figure 3.1.2.

**Table 3.1.1: Fixed Effects Estimates from R**

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)  2.48316    0.58851   4.219
season.c     0.29618    0.06996   4.233
positionF    0.41300    0.75884   0.544
positionG   -0.12935    0.72320  -0.179
```

Table 3.1.1 shows the parameter estimates for the chosen model fitted in R. Starting in the year 2011 (season.c = 0), for an average player who plays center in the league, the estimated salary is $2.48 million. Moving forward, with each increase of one season after 2011, the salary of an average player who plays center's salary is estimated to increase by $296,180. Looking at the forward position, an average player that is in the forward position is predicted to have their salary increase by $413,000 for each increase one year after the 2011 season compared to an average playing center in the same season. Guards seem to have an opposite effect, with an average player playing the guard position is predicted to have their salary decrease by $129,350 for each one-year increase after 2011, compared to an average player playing center in the same season.

**Table 3.1.2: Random Effects Estimates From R**

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 name     (Intercept) 8.2867   2.8787
          season.c    0.6815   0.8255   -0.15
 Residual             8.2824   2.8779
Number of obs: 1113, groups:  name, 200
```

Table 3.1.2 displays the random effects associated with the final model. There is very similar variation in a player's salary between players as there is between years within a player.

The value of -0.15 is relating to the negative correlation between the random slopes of season and the random intercepts of each player. Players with higher slopes (larger effect of season on average salary) tend to have smaller intercepts (salary when season is at its average year, 2011).

From Table 3.1.2, the level 1 and 2 equations are as follows:

**Level 1:** $salaryM_{ij} = \beta_{0j} + \beta_{1j}(season.c)_{ij} + \varepsilon_{ij}$

**Level 2:** $\beta_{0j} = \beta_{00} + \beta_{01}(position)_j + u_{0j}$

$\beta_{1j} = \beta_{10} + \beta_{11}(position)_j + u_{1j}$

$u_{0j} \sim N(0, \tau_2)$

$u_{1j} \sim N(0, \tau_2)$

$e_{ij} \sim N(0, \sigma_2)$

**Figure 3.1.2: Final Model Random Slopes**


Figure 3.1.2 displays the random slopes of a sample of 9 players from the dataset. Each line in the plot represents the association between salary, in millions, and the centered variable of season for a specific player. The center of the plot represents the average year in the dataset, which is 2011. There is a large amount of variation in intercepts, although less variation around 2011 (season.c = 0). There also appears to be significant variation in slopes.

$$0.29618 \pm 2(0.825) = (-1.35, 1.95)$$

It is estimated that 95% of players experience a change in their year-to-year salary ranging from a loss of $1.35 million to an increase of about $1.95 million. This large interval of estimated change matches the trends shown in Figure 3.1.2. The plot contains a mix of players that have a large positive effect, large negative effect, and small effects. Therefore, the estimated confidence interval is large to be indicative of this wide spread of slopes between players. The plot indicates players with larger intercepts (higher starting salary in the average season), tend to have smaller slopes (less effect of season on salary). If estimated salary is higher than the average, then season is predicted to have has less of an impact. The inclusion of the random slopes variable explains 30.9% of the variation in the predicted salaries of an average NBA player for the average season.

**Figure 3.1.3: Effects Plots for Season.c and Position**

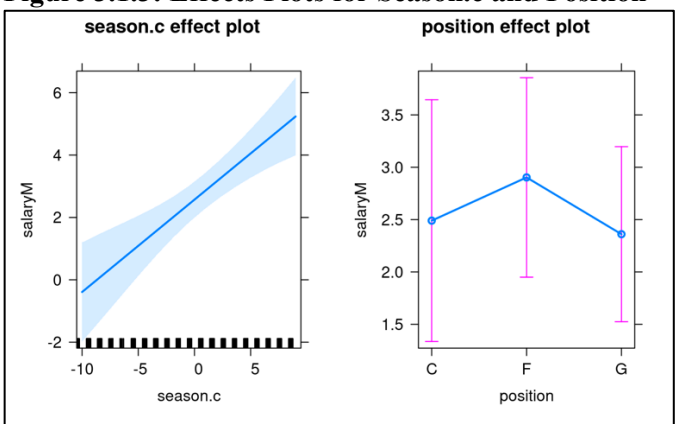



Figure 2.4.1 shows the fixed effects of our final model. Thus, for an average player in the NBA, the effect plot shows an overall increase in salary for every year after 2011. It also shows the predicted salary going from one position to another for the overall model. Here we can see that in 2011, centers and guards tend to see a similar average salary around $2.5 million, while forwards see around $2.9 million.

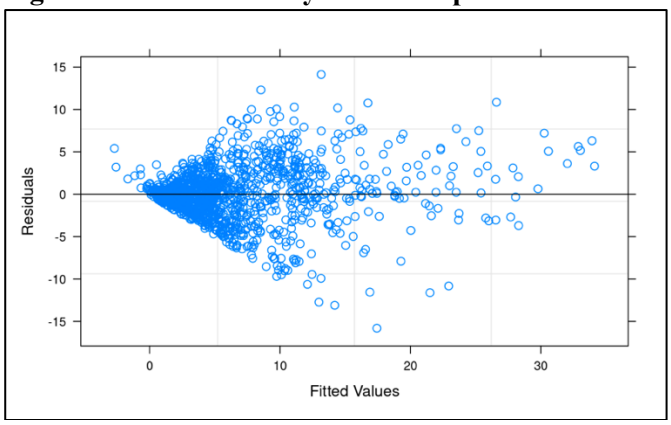
## 3.2: Validity of Assumptions

**Table 3.2.1: Hypothesis Testing for Random Slopes**

```
Data: NBA
Models:
model2: salaryM ~ season.c + position + (1 | name)
model3: salaryM ~ season.s + position + (1 + season.c | name)
       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model2    6 2727.2 2751.6 -1357.6   2715.2
model3    8 2536.7 2569.3 -1260.4   2520.7 194.47  2  < 2.2e-16 ***
```

The output displayed in Table 3.2.1 is testing to see if the differences in random slopes (season.c) is statistically significant. The null hypothesis being tested is if there is between player variation in the data ($H_0: \tau^2 = 0$). The Chi-Square test statistic ($X^2 = 194.47$, DF = 2), along with a small p-value (p-value < 2.2e-16) provides evidence that the inclusion of the random slopes in the model are significantly different and are an improvement on the models fit to the data.

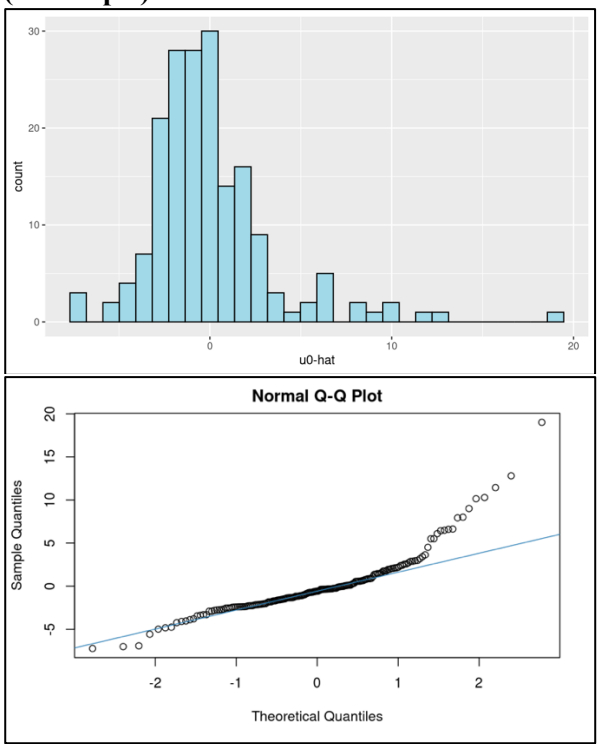
**Figure 3.2.2: Residual by Predicted plot for Model 3**




The Residuals vs. Predicted values plot in Figure 3.2.2 for the final model shows a very clear fanning of the residuals, meaning for larger fitted values we see much more variability about the overall regression line. Although this may seem problematic, our model using random slopes accounts for this heteroscedasticity by allowing each player to have a unique effect for the centered season variable.

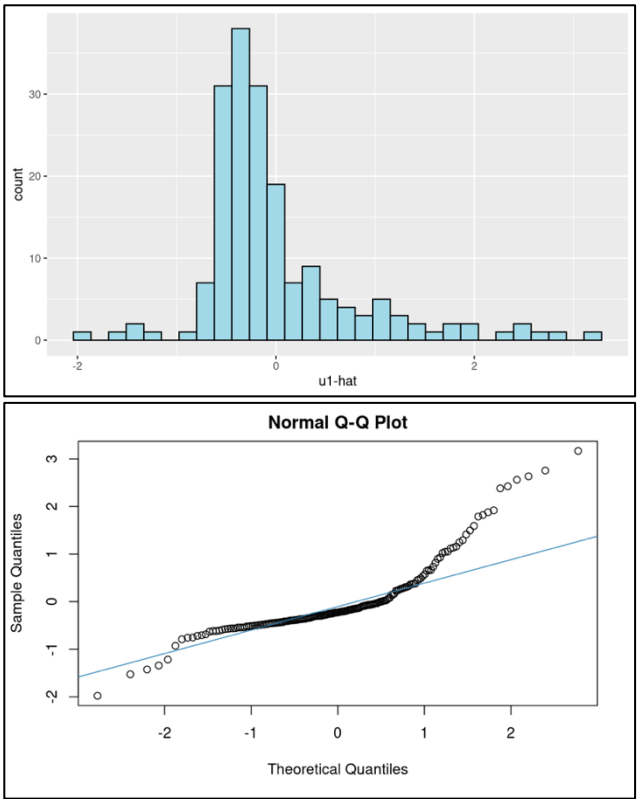
## 3.3 Normality of Random Effects

**Figure 3.3.1 Histogram of Player Random Effects (Intercepts)**




In Figure 3.3.1 we can see a slight right skew in the random intercept's effects, which is where most NBA players lie. Contextually, this makes sense, as most players will see their average salary close to that of the league's average. The skew, however, is a result of the 1 or two superstars on every team that earn the larger contracts. Adding them to the data set brings the average effect size further from 0. From the histogram and QQ-plot, the assumptions made for the level 1 and 2 equations must be taken with a grain of salt.

**Figure 3.3.2 Histogram of Season.c Random Effects (Slopes)**




As shown in Figure 3.3.2, we can see that the random effects for the centered season variable are not approximately normal either, as there is a slight right skew in the histogram and QQ-plot. This means that most players see a smaller predicted change in their salary from year to year, centered close to 0. However, there are a handful of players that see a very large, predicted increase in their salary from year to year and even fewer that see a predicted decrease in salary from year to year. This is potentially due to extrinsic variables such as player's role on a team, along natural regression in physical ability over time.

## 4. Discussion

From the model building process, the ambiguity surrounding the growth of NBA players' salaries has become a bit clearer. When answering the research question regarding the impact position has on player salaries across seasons, there are contradicting narratives between the graphics and model summaries. When looking at Figure 6.4.1, there is evidence of a relationship between position and salary, with centers and guards having steeper slopes across seasons, whereas forwards have a smaller positive effect on salary as seasons progress.

Looking back to Table 3.1.3, the slightly positive coefficient of the centered variable season, is indicative of the trend that the salary for the center position is headed toward. While the coefficient is still positive ($\beta_{season.c} = 0.29$), it appears to be on the downward trend. This is further exemplified through the random effects plots, where centers such as Rasheed Wallace and Zaza Pachulia both have negative slopes. Looking at this in the context of the NBA, the center was a very important position in the game. Therefore, those were the players making the most money. But as the game evolved, it has become a very guard and forward-focused league. In Figure 3.1.2, players such as Wallace and Pachulia, have slopes that show a high intercept (high salary in the year 2000), but a negative trend as the years progress. The year 2011 seems to display a transition period in the league where the shift from the center-dominated play in the earlier years, to a more forward and guard focused attack. Players such as John Wall and Stephen Curry are some prime examples of guards that had intercepts below average, with above average slopes.

These insights shown in the graphs display the relationship between positions and salary as the seasons increase. Although this relationship between position and seasons appears quite apparent in the graphics, when attempting to implement a interaction within the model, it does not appear to improve the fit of the model based off likelihood ratio testing.

This contradiction in the model output and graphics speaks to the number of extraneous factors that go into a player's salary. Subjective factors such as player's impact on the team, and superstar-status are not included in this dataset and speak to the limitations of the dataset.

While our analysis has provided a lot of insight into NBA player salaries over time, there are some limitations that we must acknowledge. As stated earlier, our research question pertains to the relationship between player position and player salary, as well as the relationship between a season, position, and salary (interaction). Given our data set we did our best to find a model that accurately incorporates this. However, our model revealed that none of the coefficients including position proved to be statistically significant (t-value < 2) in predicting player salaries. This prevents us from making any meaningful generalizations about players based on their position. In hindsight, the model with AR(1) may have been the most efficient model, however we were very interested in adding random slopes to the model to further emphasize and embrace player to player variation. At the same time, using a multilevel model with random slopes allowed for us to further address and account for the violated unequal variance assumption seen in Figure 3.2.2.

Going back to the research question, while the analysis done to this point has uncovered interesting insight on the relationship between players and salary, it also uncovered the importance of a versatile dataset. To elaborate, the lack of predictors in the original dataset hinders the amount of variation we are truly able to explain in the response variable.

A main strength of our model may be one that was initially considered a weakness. Figures 3.3.1 and 3.3.2 show a violation of the assumption that the random effects are distributed normally, while this is obviously not ideal, it shows that we have collected enough data so that the population of the NBA is accurately represented, meaning that superstars and role players alike are comprise our sampled data set. This helps us accurately model and predict the salary of a wide range of players: Players that have played for 20 years, players that have played in the early 2000s, players that have played in the 2010s, role players, one-and-dones. They all have a place in our data set and in our model with the use of multi-level

modeling. When considering our interest and research questions, including these players and accounting for the variation of these players was one of our main concerns and for that reason our final model is the most relevant.

Regarding future research, the addition of more predictors in the dataset would be the most important step to take. When conducting the model building and variable selection process, the inclusion of different predictors would significantly improve the ability to predict the relationship between salary and players for each position across seasons. To build off the work done to this point with the data, trying out different interactions along with the addition of team as a random effect may help explain player to player variation.

# 5. Appendix

**Figure 6.1.1 Unconditional Growth Model:** Below is the coding of the unconditional growth model using the lmer function. This includes a fixed effect of the time variable season.c, and allows for random intercepts for player.

```
# unconditional growth model
model1 <- lmer(salaryM ~ season.c + (1|name), data = nba1)
summary(model1)

sample(nba1$name, size = 1, replace = FALSE)
plot(ggpredict(model1, terms = c("season.c", "name [Terrence Ross, Zaza Pachulia, Chris Paul,
Rasheed Wallace, John Wall, Ryan Richards, Donyell Marshall, Stephen Curry, Naz Reid]"), type =
"re"), ci = FALSE)
plot(allEffects(model1))
```

**Figure 6.1.2: Model 1:** Below is an illustration of the unconditional growth model. Each line represents one of the 9 randomly sampled NBA players from our subsetted data set, "nba1". A consistent effect of season.c is predicted for every player in the data set. We can see variation in their predicted salaries at season.c = 0, justifying the use of random intercepts.

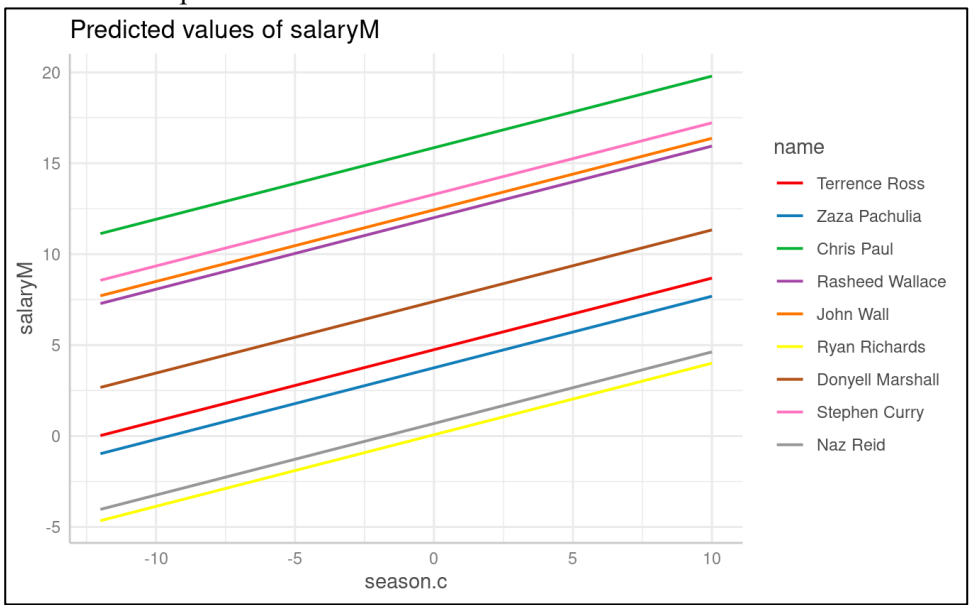



**Figure 6.1.3** Below is the overall effect of season.c for the unconditional growth model. The graph shows a positive effect of our time variable with a generally small standard error, given by the tight confidence bands

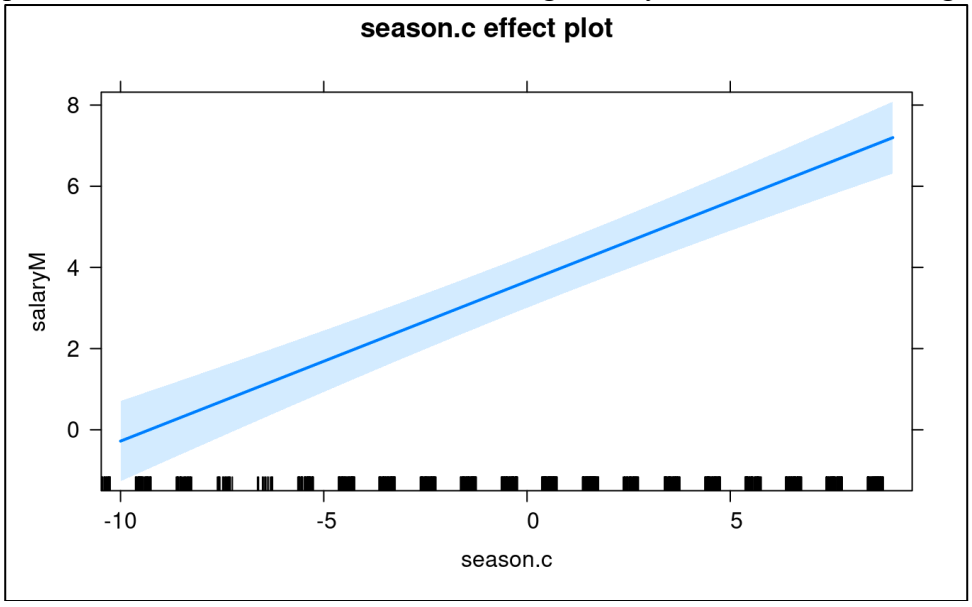


**Figure 6.2.1 Adding position variable to the Unconditional Growth Model:** The model below shows a fixed effect for the season.c and position variable, continuing to use random intercepts for player.

```
# adding position to the original model
model2 <- lmer(salaryM ~ season.c + position + (1|name), data = nba1)
summary(model2)
plot(allEffects(model2))
```

**Figure 6.2.2 Effects Plots for Model 2:** Much like Model 1 we see a positive effect for season.c on a players salary, however this model shows a slightly larger positive effect of season.c after adjusting for a player's position. The effect plot for position shows centers to have a higher average salary than forwards and guards, after adjusting for season.

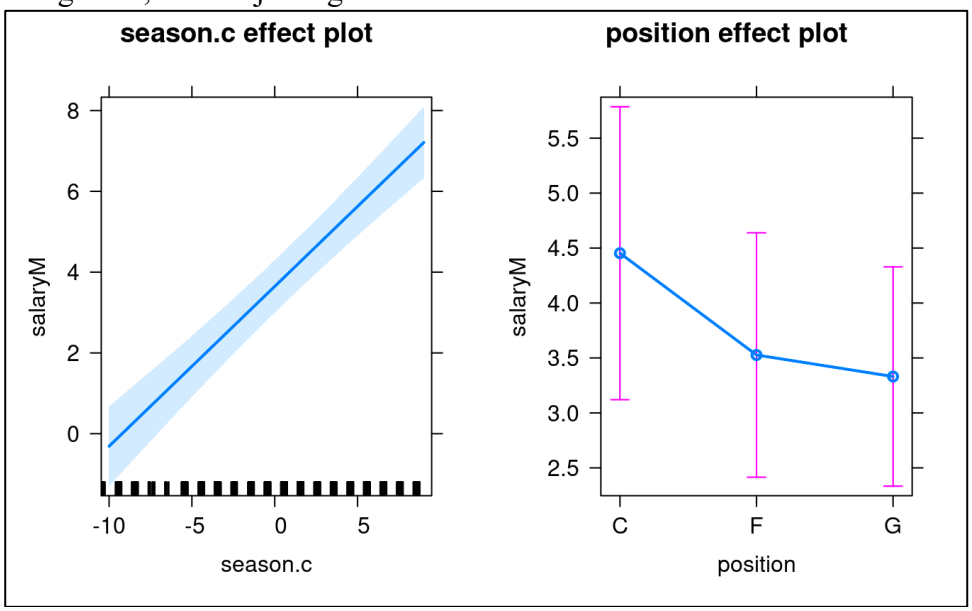



**Figure 6.3.1: Model 3: Code**

Below is the code needed to get the graphs and output in our Results section. Here you can see we are now adding random slopes to the model, while keeping season.c and position as fixed effects.

```
# random slopes/intercepts + 2 fixed
# the between player variation is extremely large....(alot of other factors impact salary /
limitation of ds)
model3 <- lmer(salaryM ~ season.c + position + (1+ season.c | name), data= nba1)
summary(model3)

#sample(nba1$name, size = 4)
# multiple random slopes graphs
  nba1 %>%
  dplyr::filter(name %in% c("Chris Paul","Zaza Pachulia", "Donyell Marshall", "Rasheed Wallace"
, "Terrence Ross" )) %>%
  ggplot(aes(x = season.c, y = salaryM))+
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~ name) +
  theme_bw()


plot(allEffects(model3)) ##effects for season.s and position

plot(ggpredict(model3, terms = c("season.c", "name [Terrence Ross, Zaza Pachulia, Chris Paul,
Rasheed Wallace, John Wall, Ryan Richards, Donyell Marshall, Stephen Curry, Naz Reid]"), type =
"re"), ci = FALSE)
#performance::check_model(model3)
```

**Figure 6.4.1 Facet Grid for Raw data:** Below is a faceted scatter plot of season.c by salaryM, separated by position. The blue lines represent OLS lines fit to the raw data, predicting the relationship between season.c and salaryM for each position. Guards seem to have the larger positive slope, centers seem to have the scond largest positive slope, and forwards appear to have a negative slope. This leads us to believe that an interaction between season.c and position may be something to look for.

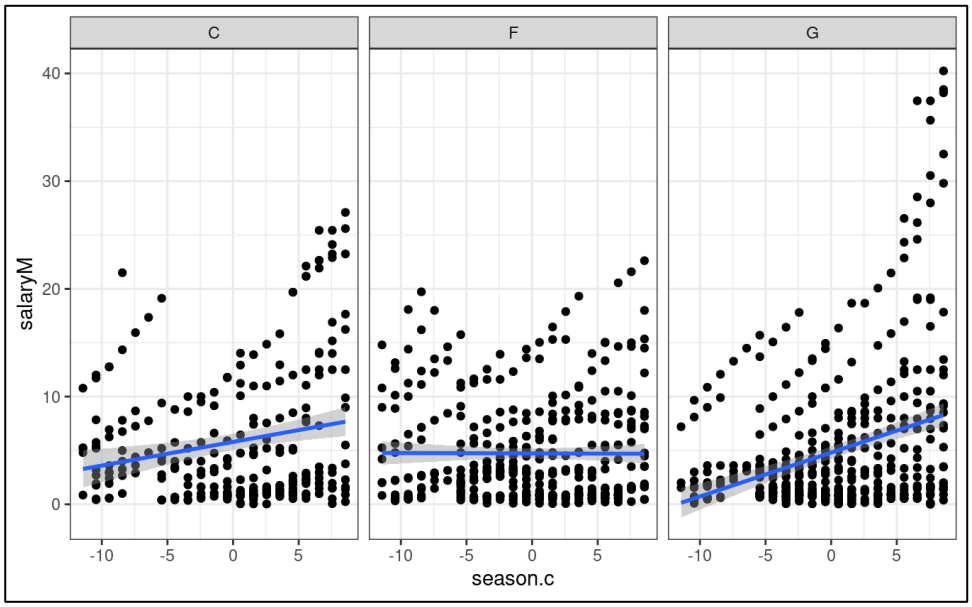



**Figure 6.4.2 Code for Model 4:** Model 4 includes an addition to the interaction between season.c and position, searching for a change in the predicted effect of season.c on a player's salary based on what position a player plays, such change seen in Figure 6.4.1.

```
# adding interaction between season and position
model4 <- lmer(salaryM ~ season.c + position + season.c:position + (1+ season.c | name), data=
nba1)
summary(model4)

plot(ggpredict(model4, terms = c("season.c", "name [Terrence Ross, Zaza Pachulia, Chris Paul,
Rasheed Wallace, John Wall, Ryan Richards, Donyell Marshall, Stephen Curry, Naz Reid]"), type =
"re"), ci = FALSE)
plot(allEffects(model4))
ggplot(data = nba1, aes(x = season.c, y = salaryM))+
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~ position) +
  theme_bw()
```

**Figure 6.4.3: Effects Plot for the interaction between season.c and position:** This plot represents the individual coefficients for the three positions, based on model 4. Although all positive, we see once again guards and centers have the larger slopes, while forward has a slope closer to 0. Despite this change, model 4 shows no significance in. the interaction term (t-values all less than 2)

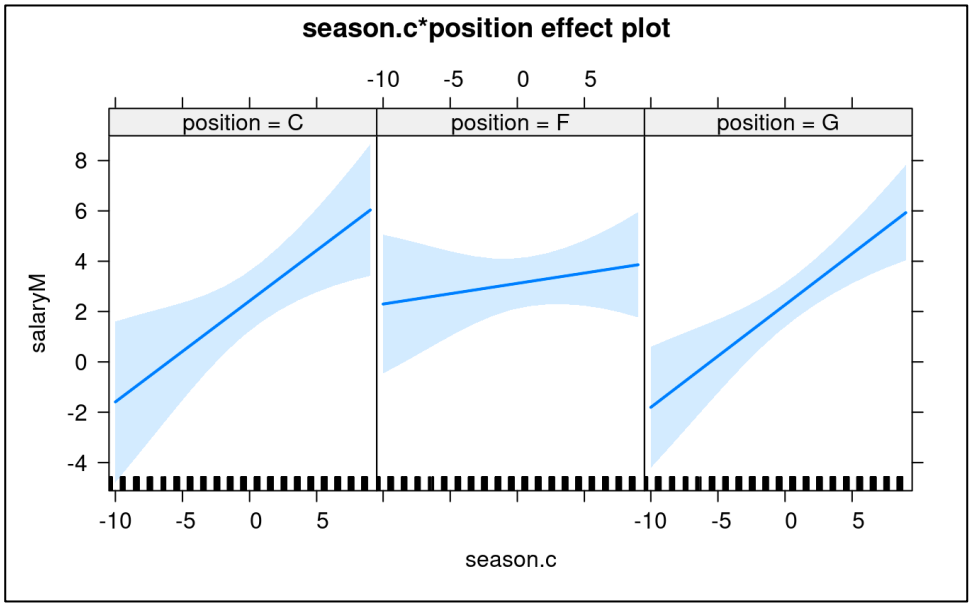



**Figure 6.5.1: Model 5:** In an attempt to find a simpler model, we tried the AR(1) covariance structure, acknowledging the correlation between time points in our data. In model 5 we continue to fit season.c and position as fixed effects, as well as players as random effects. However, the AR(1) covariance structure is specified, using the lme function.

```
# AR(1) covariance error structure with random intercepts only
model5 <- lme(salaryM ~ season.c + position, random = ~1 | name,correlation=corAR1(), data=nba1
)
summary(model5)
plot(allEffects(model5))
```

**Figure 6.5.2: Model 5 Effects Plots:** Compared to previous models, we see a slightly lower effect of season. We also see an even lower effect for the forward position and a higher effect for the guard position. However, the center position's effect on salary stays about the same.

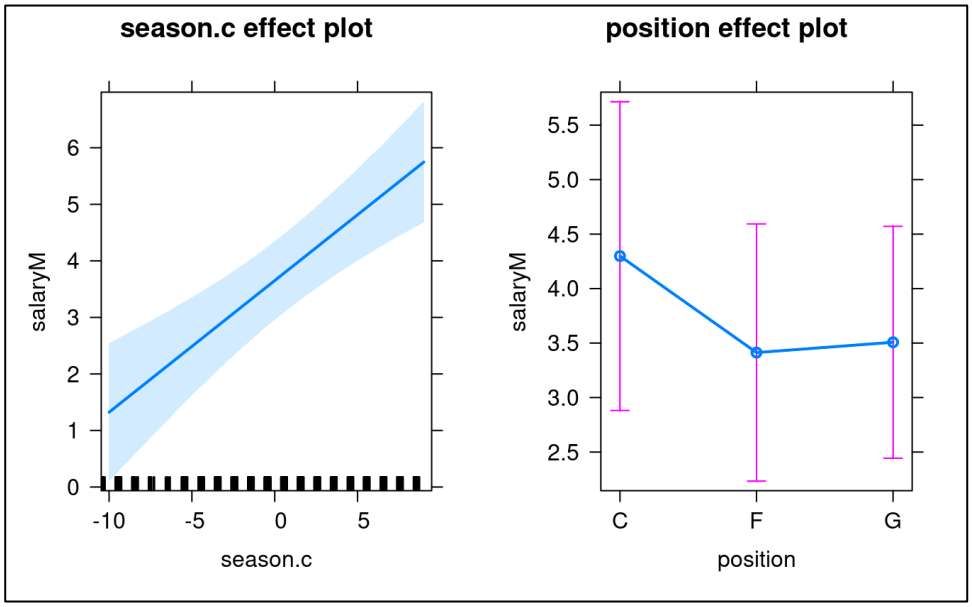



**Figure 6.6.1: ANOVA comparing AIC, BIC, logLik for the 5 models:** The maximum likelihood ratio test was used for the four models that it was applicable to models 1-4 are considered nested models, allowing for the use of the anova() function to perform this test. Since model 5 is not a nested model, the AIC, BIC, and log likelihood were compared with the other models. Of the four nested models, model 3 had to lowest AIC, BIC, and deviance with the highest log-likelihood, an accomplishment that awarded it the most efficient model for predicting player salary. However, model 5 showed significantly lower AIC and BIC with the highest log-likelihood of the 5 models. We know AIC and BIC to be generous to simpler models, so it was no surprise that the addition of just one parameter would favor model 5. Despite these impressive numbers, we aimed for a model that ultimately allowed for random slopes, to capture the relevant level 2 variability in salaries.

```
refitting model(s) with ML (instead of REML)
Data: nba1
Models:
model1: salaryM ~ season.c + (1 | name)
model2: salaryM ~ season.c + position + (1 | name)
model3: salaryM ~ season.c + position + (1 + season.c | name)
model4: salaryM ~ season.c + position + season.c:position + (1 + season.c | name)
       npar    AIC    BIC  logLik deviance    Chisq Df Pr(>Chisq)
model1    4 9084.4 9105.6 -4538.2   9076.4
model2    6 9085.4 9117.1 -4536.7   9073.4   3.0044  2     0.2226
model3    8 8530.3 8572.7 -4257.2   8514.3 559.0255  2     <2e-16 ***
model4   10 8534.1 8587.1 -4257.0   8514.1   0.2387  2     0.8875
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
[1] 7902.44
[1] 7939.5
'log Lik.' -3944.22 (df=7)
```

**Figure 6.7: Code for Checking Assumptions of our Final Model:** Assumptions of the final model include normality of random effects, as well as normality and equal variance of marginal residuals of the model. Our results section used the output from this code.

```
#Normality of random effects
intercepts <-ranef(model2)[[1]][,1]
ggplot()+
  geom_histogram(aes(x = intercepts), fill = "lightblue", col= 'black') +
  labs(x='u0-hat')
qqnorm(intercepts)
qqline(intercepts, col = 'steelblue')

slopes <- ranef(model3)[[1]][,2]
ggplot()+
  geom_histogram(aes(x = slopes), fill = "lightblue", col= 'black') +
  labs(x='u1-hat')

qqnorm(slopes)
qqline(slopes, col = 'steelblue')
```