Part III: Modeling Results

Libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(lme4)
```

```
## Loading required package: Matrix
```

```
library(ggplot2)
```

Data

```
NBA <- read.csv('https://raw.githubusercontent.com/erikgregorywebb/datasets/master/nba-salaries.csv')
NBA$season.c <- NBA$season - mean(NBA$season)
NBA$salaryM <- NBA$salary/1000000

table(NBA$team)
```

```
##
##                    Atlanta Hawks       Bilbao Basket Bilbao Basket
##                              281                                 1
##                   Boston Celtics                      Brooklyn Nets
##                              267                               171
##               Charlotte Bobcats                 Charlotte Hornets
##                              162                               150
##                   Chicago Bulls               Cleveland Cavaliers
##                              285                               283
##                Dallas Mavericks                    Denver Nuggets
##                              297                               285
##                 Detroit Pistons Fenerbahce Ulker Fenerbahce Ulker
##                              283                                 3
```

```
##            Golden State Warriors                        Houston Rockets
##                              311                                    299
##                   Indiana Pacers                            LA Clippers
##                              303                                    294
##             Los Angeles Clippers                     Los Angeles Lakers
##                               17                                    308
##       Maccabi Haifa Maccabi Haifa                    Madrid Real Madrid
##                                3                                      7
##                Memphis Grizzlies                             Miami Heat
##                              354                                    308
##                  Milwaukee Bucks                 Minnesota Timberwolves
##                              313                                    311
##                  New Jersey Nets                    New Orleans Hornets
##                              167                                    102
##             New Orleans Pelicans                        New York Knicks
##                              138                                    346
##        NO/Oklahoma City\n Hornets             NO/Oklahoma City Hornets
##                               11                                     15
##                     null Unknown                  Oklahoma City Thunder
##                               44                                    238
##                    Orlando Magic                     Philadelphia 76ers
##                              297                                    351
##                     Phoenix Suns                 Portland Trail Blazers
##                              332                                    325
##                 Sacramento Kings                      San Antonio Spurs
##                              336                                    324
##              Seattle SuperSonics                        Toronto Raptors
##                              106                                    337
##                        Utah Jazz                    Vancouver Grizzlies
##                              314                                     20
##               Washington Wizards
##                              357
```

```r
NBA[NBA$team == 'null Unknown',]
```

```
##      rank             name position         team salary season   season.c
## 2404  441      Mike Gansey       NA null Unknown 412718   2007 -4.4380288
## 2410  447      Pat Carroll       NA null Unknown 412718   2007 -4.4380288
## 2866  433       Kevin Lyde       NA null Unknown 427163   2008 -3.4380288
## 2878  445     Larry Turner       NA null Unknown 427163   2008 -3.4380288
## 2879  446      Elton Brown       NA null Unknown 427163   2008 -3.4380288
## 2884  451      Sammy Mejia       NA null Unknown 427163   2008 -3.4380288
## 2887  454     Jared Newson       NA null Unknown 427163   2008 -3.4380288
## 2893  460     Jackie Manuel       NA null Unknown 427163   2008 -3.4380288
## 3379  455    Taj McCullough       NA null Unknown 442114   2009 -2.4380288
## 3389  465   Jason Richards       NA null Unknown 442114   2009 -2.4380288
## 3390  466     David Padgett       NA null Unknown 442114   2009 -2.4380288
## 3391  467       C.J. Giles       NA null Unknown 442114   2009 -2.4380288
## 3392  468  Dwayne Mitchell       NA null Unknown 442114   2009 -2.4380288
## 3394  470      Dion Dowell       NA null Unknown 442114   2009 -2.4380288
## 3395  471  Richard Hendrix       NA null Unknown 442114   2009 -2.4380288
## 3401  477     Jamaal Tatum       NA null Unknown 442114   2009 -2.4380288
## 3861  452 Deron Washington       NA null Unknown 457588   2010 -1.4380288
## 3875  466  Kenny Hasbrouck       NA null Unknown  75476   2010 -1.4380288
```

```
## 3878  469  Curtis Jerrells       NA null Unknown  59217   2010 -1.4380288
## 4339  448  Kenny Hasbrouck       NA null Unknown 762195   2011 -0.4380288
## 4343  452  Curtis Jerrells       NA null Unknown 762195   2011 -0.4380288
## 4380  489 Stanley Robinson       NA null Unknown 473604   2011 -0.4380288
## 4381  490     Brian Zoubek       NA null Unknown 473604   2011 -0.4380288
## 4383  492      Tiny Gallon       NA null Unknown 473604   2011 -0.4380288
## 4400  509 Deron Washington       NA null Unknown 457588   2011 -0.4380288
## 4419  528   Da'Sean Butler       NA null Unknown 125000   2011 -0.4380288
## 4433  542     Magnum Rolle       NA null Unknown   8358   2011 -0.4380288
## 4917  484   Da'Sean Butler       NA null Unknown 788872   2012  0.5619712
## 4925  492     Magnum Rolle       NA null Unknown 788872   2012  0.5619712
## 4938  505  Kenny Hasbrouck       NA null Unknown 762195   2012  0.5619712
## 4947  514  Curtis Jerrells       NA null Unknown 762195   2012  0.5619712
## 4982  549 Stanley Robinson       NA null Unknown 473604   2012  0.5619712
## 4986  553     Brian Zoubek       NA null Unknown 473604   2012  0.5619712
## 4989  556      Tiny Gallon       NA null Unknown 473604   2012  0.5619712
## 5004  571 Deron Washington       NA null Unknown 457588   2012  0.5619712
## 5578  526   Da'Sean Butler       NA null Unknown 788872   2013  1.5619712
## 5591  539     Magnum Rolle       NA null Unknown 788872   2013  1.5619712
## 5613  561  Kenny Hasbrouck       NA null Unknown 762195   2013  1.5619712
## 5625  573  Curtis Jerrells       NA null Unknown 762195   2013  1.5619712
## 5664  612 Stanley Robinson       NA null Unknown 473604   2013  1.5619712
## 5672  620     Brian Zoubek       NA null Unknown 473604   2013  1.5619712
## 5677  625      Tiny Gallon       NA null Unknown 473604   2013  1.5619712
## 5686  634      Tu Holloway       NA null Unknown 473604   2013  1.5619712
## 5700  648 Deron Washington       NA null Unknown 457588   2013  1.5619712
##       salaryM
## 2404 0.412718
## 2410 0.412718
## 2866 0.427163
## 2878 0.427163
## 2879 0.427163
## 2884 0.427163
## 2887 0.427163
## 2893 0.427163
## 3379 0.442114
## 3389 0.442114
## 3390 0.442114
## 3391 0.442114
## 3392 0.442114
## 3394 0.442114
## 3395 0.442114
## 3401 0.442114
## 3861 0.457588
## 3875 0.075476
## 3878 0.059217
## 4339 0.762195
## 4343 0.762195
## 4380 0.473604
## 4381 0.473604
## 4383 0.473604
## 4400 0.457588
## 4419 0.125000
## 4433 0.008358
```

```
## 4917 0.788872
## 4925 0.788872
## 4938 0.762195
## 4947 0.762195
## 4982 0.473604
## 4986 0.473604
## 4989 0.473604
## 5004 0.457588
## 5578 0.788872
## 5591 0.788872
## 5613 0.762195
## 5625 0.762195
## 5664 0.473604
## 5672 0.473604
## 5677 0.473604
## 5686 0.473604
## 5700 0.457588
```

```
NBA <- NBA[-c(4555,4672,5303,5622,6621,8195,8330,2762,2877,4352,4957,5157,5336,6465,2404,2410,2866,2878

NBA %<>%
  mutate(team = ifelse(team %in% c('Atlanta Hawks'), 'Hawks',
                     ifelse(team %in% c('Brooklyn Nets', 'New Jersey Nets'), 'Nets',
                            ifelse(team %in% c('Charlotte Bobcats', 'Charlotte Hornets'), 'Hornets',
                                   ifelse(team == "Boston Celtics", 'Celtics',
                                          ifelse(team %in% c('Chicago Bulls'), 'Bulls',
                                                 ifelse(team == 'Cleveland Cavaliers', 'Cavs',
                                                        ifelse(team == 'Dallas Mavericks', 'Mavericks
                                                               ifelse(team == 'Denver Nuggets', 'Nugg
                                                                      ifelse(team == 'Detroit Pistons
                                                                             ifelse(team == 'Golden S
                                                                                    ifelse(team == 'H
                                                                                           ifelse(tea
                                                                                                  ifel
```

```
NBA %<>%
  mutate(position = ifelse(position %in% c(' PF', ' SF', ' GF'), 'F',
                           ifelse(position %in% c(' SG', " PG"), 'G',
                                  ifelse(position == ' C', 'C', position))))
NBA <- NBA[NBA$position %in% c('G', 'F', 'C'),]
```

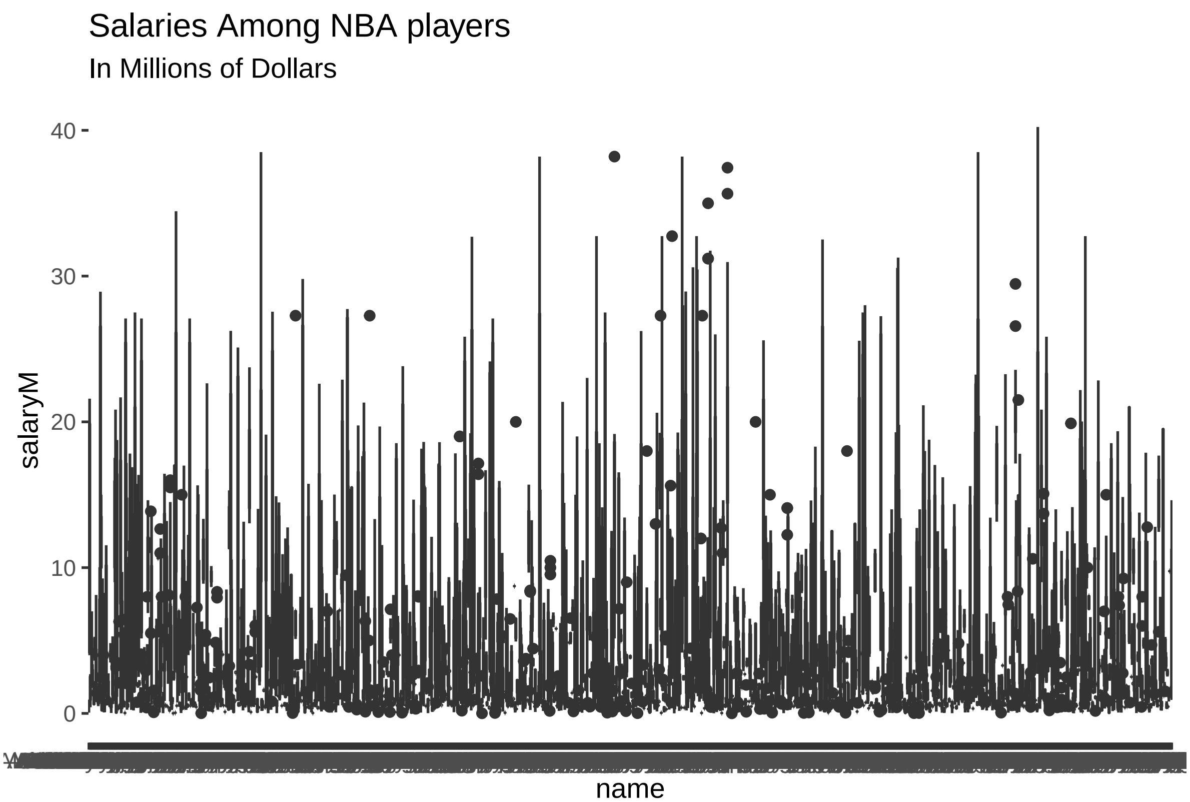
1. ggplot(NBA, aes(x = name , y = salaryM) +

   geom_boxplot()+

   theme_bw())

2. Include a graph exploring the variability in the response variable across the Level-2 units. Fit an ANOVA using OLS for your response variable and the Level-2 group variable. Does the group effect appear to be statistically significant?

```
library(ggplot2)
ggplot(data = NBA, aes(x = name, y = salaryM))+
  geom_boxplot() +
  labs(title = 'Salaries Among NBA players',
       subtitle = 'In Millions of Dollars',
       xlab = NULL) +
  theme(axis.text.x = NULL)
```



The bocplot plot shows varying distributions of salaries. The ANOVA using OLS also indicates that there is significant player to player variation in mean salaries (1504 and 6608 DF, F-value = 4.702, p-value < 2e-16).

2. Fit the "intercepts only" model. Interpret each of the estimated parameters in context. Interpret the intraclass correlation coefficient in context. Does the value of the ICC seem "substantial" to you? Report the likelihood, deviance, and AIC values for later comparison.

```
model0 = lmer(salaryM ~ 1 + (1 | name), data = NBA)
summary(model0)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: salaryM ~ 1 + (1 | name)
##    Data: NBA
##
## REML criterion at convergence: 48187.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7207 -0.3802 -0.1541  0.2916  6.2576
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  name     (Intercept) 10.20    3.194
##  Residual             17.57    4.191
## Number of obs: 8113, groups:  name, 1505
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   3.6731     0.1002   36.67
```

```
AIC(model0)
```

```
## [1] 48193.42
```

```
logLik(model0)
```

```
## 'log Lik.' -24093.71 (df=3)
```

```
deviance(model0, REML = F)
```

```
## [1] 48184.66
```

Overrall intercept: The predicted average salary for an average NBA team is about 3.6731 million.

τ^2 **= The estimated variation in average salaries among NBA players is about 10.20**

σ^2 **= The estimated variation in salaries between salary observations from the same player is about 17.57**

ICC = $\tau^2/\tau^2 + \sigma^2$ **—> How correlated two salary observations for the same player.**

= 10.20/10.20+17.57

= .3673

The ICC is fairly small but does seem to be substantial.

AIC = 48193.42

Deviance = 48184.66

logLik = -24093.71

3. Add 1-3 Level 1 variables. Carry out a likelihood ratio test to compare this model to the model in step 2 (using ML, clearly explain how you find the chi-square value and df). Include details. Also report/compare the AIC values to the intercepts only model. Calculate a "proportion of variation explained" for each variable (and what variation) and interpret the results in context. Did the Level 2 variance decrease? What does the tell you? Remove (one at a time) any insignificant variables.

)

```
model1 <- lmer(salaryM ~ season.c + (1|name), data = NBA)
summary(model1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: salaryM ~ season.c + (1 | name)
##    Data: NBA
##
## REML criterion at convergence: 47802.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.2979 -0.3946 -0.1077  0.3151  6.1001
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  name     (Intercept) 12.00    3.464
##  Residual             16.22    4.028
## Number of obs: 8113, groups:  name, 1505
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  3.52067    0.10531   33.43
## season.c     0.25308    0.01227   20.62
##
## Correlation of Fixed Effects:
##          (Intr)
## season.c -0.042
```

```
summary(model0)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: salaryM ~ 1 + (1 | name)
##    Data: NBA
##
## REML criterion at convergence: 48187.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7207 -0.3802 -0.1541  0.2916  6.2576
##
## Random effects:
```

```
##  Groups   Name        Variance Std.Dev.
##  name     (Intercept) 10.20    3.194
##  Residual             17.57    4.191
## Number of obs: 8113, groups:  name, 1505
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   3.6731     0.1002   36.67
```

```
anova(model0, model1) #X^2 = 392.12, DF = 1, p-value <2.2e-16
```

```
## refitting model(s) with ML (instead of REML)

## Data: NBA
## Models:
## model0: salaryM ~ 1 + (1 | name)
## model1: salaryM ~ season.c + (1 | name)
##        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## model0    3 48191 48212 -24092    48185
## model1    4 47801 47829 -23896    47793 392.12  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the likelihood ratio tests we can see that model 1 is significantly better than model 0 (reasoning is commented next to code). However, it does slightly increase Level 2 variance, which means that the despite the positive association between salary and season, after adjusting for player, we might see a different relationship between year and salary. Going from Model0 to Model1, the AIC is smaller by about 400, BIC is smaller by about 200 and logLik is larger by about 200 as well. The deviance is also smaller by about 400.

Change in Level 1 variance : 17.57-16.22/17.57 = 7.68% decrease

Change in Level 2 variance: 10.20-12/10.20 = 17.64% increase

4. Add 1-3 Level 2 variables. Carry out a likelihood ratio test to compare the models (using ML). Include details. Also report/compare the AIC values. Calculate a "proportion of variation explained" for each level and interpret the results in context. Remove (one at a time) any insignificant variables.

```
model2 <- lmer(salaryM ~ season.c + position + (1|name), data = NBA)
summary(model2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: salaryM ~ season.c + position + (1 | name)
##    Data: NBA
##
## REML criterion at convergence: 47788.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.3193 -0.3915 -0.1080  0.3166  6.1103
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  name     (Intercept) 11.93    3.454
```

```
##  Residual             16.21    4.026   
## Number of obs: 8113, groups:  name, 1505
## 
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  4.26249    0.22683  18.792
## season.c     0.25578    0.01228  20.820
## positionF   -0.79062    0.28253  -2.798
## positionG   -1.09122    0.28112  -3.882
## 
## Correlation of Fixed Effects:
##           (Intr) sesn.c postnF
## season.c   0.035              
## positionF -0.804 -0.048       
## positionG -0.806 -0.064  0.649
```

```
anova(model1, model2) #X^2 = 15.194, DF = 2, p-value = .0005
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: NBA
## Models:
## model1: salaryM ~ season.c + (1 | name)
## model2: salaryM ~ season.c + position + (1 | name)
##        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)    
## model1    4 47801 47829 -23896    47793                         
## model2    6 47789 47831 -23889    47777 15.194  2  0.0005019 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Via the likelihood ratio test, we can see that the model with the position (Level 2 variable) is better at predicting salaries (reasoning commented in code). The AIC and BIC decrease by more than 10, the logLik increases by about 7, and the deviance decreases by about 10 as well.

Change in Level 2 variance: 17.57-16.21/17.57 = 7.74% decrease

Change in Level 1 variance: 10.20-11.93/ 10.20 = 16.96% increase

5. Consider random slopes for one Level 1 variable. (This could be one of the variables that was removed earlier…) Include a graph illustrating variability in the estimated random slopes and discuss what you learn in context. Interpret the amount of group-to-group variation in these slopes in context. Once you have a model with at least one set of random slopes, compare this model to the model in step 4, is adding random slopes a significant improvement (REML, be clear how you are determining degrees of freedom)?

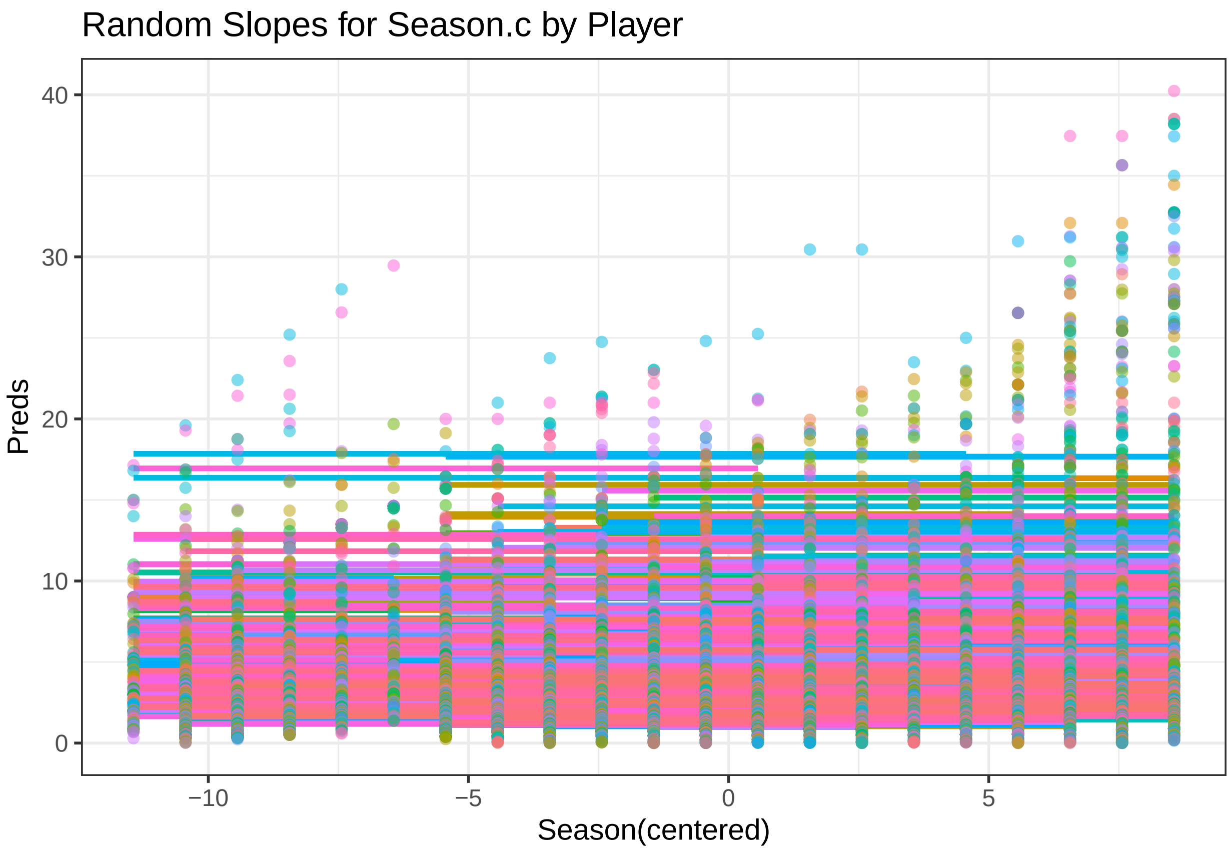
```
model3 <- lmer(salaryM ~ season.c + position + (1+season.c|name), data= NBA)
summary(model3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: salaryM ~ season.c + position + (1 + season.c | name)
##    Data: NBA
## 
## REML criterion at convergence: 45116.5
## 
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.3047 -0.3337 -0.0767  0.2721  5.5473
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  name     (Intercept) 8.1716   2.8586
##           season.c    0.4768   0.6905   -0.10
##  Residual             8.9544   2.9924
## Number of obs: 8113, groups:  name, 1505
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  3.15886    0.22491  14.045
## season.c     0.23738    0.02216  10.713
## positionF   -0.46611    0.27830  -1.675
## positionG   -0.78955    0.27664  -2.854
##
## Correlation of Fixed Effects:
##           (Intr) sesn.c postnF
## season.c  -0.036
## positionF -0.806 -0.041
## positionG -0.808 -0.047  0.657
```

```
preds = predict(model0, newdata = NBA)
ggplot(NBA, aes(x = season.c , y = preds , group = name, color = name)) +
geom_smooth(method = "lm", alpha = .5, se = FALSE) +
geom_point(data = NBA, aes(y = salaryM, color=name), alpha = .5) +
  theme_bw() +
  theme(legend.position = 'none') +
  labs(title = 'Random Slopes for Season.c by Player',
       x = 'Season(centered)',
       y = 'Preds')
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
anova(model2, model3)
```

```
## refitting model(s) with ML (instead of REML)

## Data: NBA
## Models:
## model2: salaryM ~ season.c + position + (1 | name)
## model3: salaryM ~ season.c + position + (1 + season.c | name)
##        npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
## model2    6 47789 47831 -23889    47777
## model3    8 45122 45178 -22553    45106 2671.4  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The variability in the effect of season on slope from player to player is given by $\tau_1^2 = .4768$. We can see from the graph that players with lower salaries in 2011 (the average season in the dataset) see a much larger effect of season.c on their salaries, on average.

The loglikelihood ratio test shows that the model with random slopes is significantly better at predicting salaries than the model without (chi-squared = 2671.4, DF = 2, p-value = 2.2e-16)

6. Add and interpret a cross-level interaction (you may have to use insignificant variables, focus on interpreting the interaction). Are you able to explain much of the slope variation you found in step 5? Is this a significantly better model?

```r
model4 <- lmer(salaryM ~ season.c + position + season.c*position + (1+ season.c | name), data = NBA)
summary(model4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: salaryM ~ season.c + position + season.c * position + (1 + season.c |
##     name)
##    Data: NBA
##
## REML criterion at convergence: 45123.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.3006 -0.3340 -0.0771  0.2722  5.5387
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  name     (Intercept) 8.172    2.8586
##           season.c    0.477    0.6907   -0.10
##  Residual             8.955    2.9925
## Number of obs: 8113, groups:  name, 1505
##
## Fixed effects:
##                    Estimate Std. Error t value
## (Intercept)         3.15759    0.22545  14.006
## season.c            0.24123    0.04702   5.131
## positionF          -0.43926    0.28038  -1.567
## positionG          -0.81318    0.27895  -2.915
## season.c:positionF -0.03307    0.05909  -0.560
## season.c:positionG  0.02227    0.05877   0.379
##
## Correlation of Fixed Effects:
##             (Intr) sesn.c postnF postnG ssn.:F
## season.c    -0.077
## positionF   -0.804  0.062
## positionG   -0.806  0.062  0.649
## ssn.c:pstnF  0.061 -0.796 -0.117 -0.049
## ssn.c:pstnG  0.062 -0.800 -0.050 -0.122  0.636
```

The predicted decrease in season.c's effect on salary for an average forward in the NBA is 0.03307.

The predicted increase in the season.c's effect on salary for an average guard in the NBA is .02227.

Change in random slopes variance coefficient :

.4768-.4777/.4768 = .0018 increase

We see less than a 1% change in the variability between slopes from player to player.

```r
anova(model3,model4)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: NBA
## Models:
## 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)
## model3    8 45122 45178 -22553    45106
## model4   10 45125 45195 -22552    45105 1.2263  2     0.5416
```

Adding the interaction term does not prove to be significantly better than the model without an interaction term (chi-squared = 1.2263, DF = 2, p-value = .5416).