Looking Past the Box Score: A Multilevel Modeling Exploration of VORP Variations in NBA Players

Andrew Martinez, Zachary Felix, Ethan Schultz
Stat 414-01
Group 3

Introduction

Basketball is one of the United States' favorite sports to watch each year; it approximately attracts around 1.59 million fans each year that tune in and watch the sport live in person or on TV (PlayToday). Some fans watch casually but others invest a lot of time and interest into the sport and for them it is a part of their identity. Whether it be their favorite team or player, fans will follow them for seasons and seasons on end watching their ups and downs as time goes on they can watch the evolution of their favorite teams or players.

For the people more invested in basketball, statistics actually can play a huge part in their understanding of how well their favorite teams or players are doing, going above and beyond if they are just winning or losing games. There are a plethora of statistics out there that can gauge all sorts of teams and players performance over games, seasons, and even careers. For us basketball is more than a casual sport and we can apply our knowledge of statistics to get a deeper understanding on how some of the players that we have watched all these years are doing. One of these statistics that provide a deeper understanding is a player's VORP. VORP is a measurement of a player's performance on their team when compared to if they were replaced by an "average" player of their same position.

Figure 1: VORP Equation Along With Variable Explanation

$$VORP = [BPM - (-2)] \cdot \frac{5 \cdot MP}{TeMP} \cdot \frac{TeGP}{LgGP}$$

- Box Plus Minus [BPM];
- Minutes played [MP];
- Team minutes played [TeMP];
- Team game played [TeGP];

As can be seen in Figure 1, VORP is a calculation involving a players Box Plus Minus, subtracted by a -2 (which is an arbitrary BPM assigned to a "replacement player"), which is then multiplied by 5 * (Minutes Played / Team Minutes Played) * (Team Games Played / League Games Played). While we do not include any of these variables within our dataframe, we do have variables that can be correlated. For example, as a player's minutes played increases their points per game tends to increase. We hypothesize that we can accurately predict VORP using variables that are not directly included in the VORP calculation, and that the effects these variables have on VORP vary between players. While variables like age, salary, points per game and years spent in college are not directly included in VORP's calculation, we believe that through knowing these variables we can accurately predict a player's VORP. Having the ability to predict a player's VORP can be very beneficial to NBA General Managers for many reasons. For example, general managers can better target players in free agency and trades by evaluating that player's predicted VORP against their current market value. By signing players whose VORP outperform their market value general managers can assemble cost effective rosters.

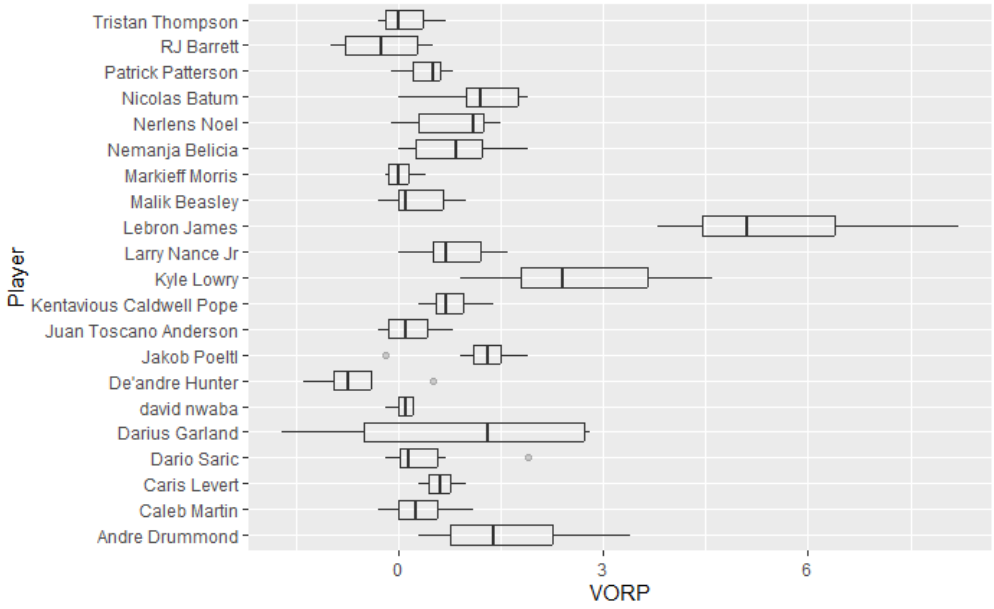
Data source and Methods

All the data that we used came from a site called basketball reference. Different variables were chosen than we thought would be a useful indicator of a players VORP. The level one units that we chose were a players season and the level 2 units are the players careers. Within the level one units certain predictors were chosen which include points per game, their age, and their salary for that particular season. Because the data assembly was largely by hand, 21 players were selected as the data acquisition process was cumbersome. Each of these variables were converted to an average value per season for that specific player which we called cenage, censcalesalary, and cenPTS. We did this to make the interpretations of these variables easier to understand in context when seeing how they affect VORP. The salary variable was also scaled to millions of dollars to also make it easier to interpret(censcalesalary). Then we chose our level two predictors which included how many years they spent in college before being drafted to the NBA, whether or not they were a lottery pick, and the average points across the players career (mean PTS). The quantitative variables of this list which included only a player's average points across their season was centered again for easier interpretation in context. The variable for whether or not the player was a lottery pick was called bindraft and was coded one for if a player is a lottery draft and zero if the player was a non-lottery draft. Often the initially better projected players are drafted in the lottery because they tend to be the more skilled players, but is not a measure of how they actually perform once they are in the NBA. Since we only wanted to look at complete seasons we had to filter out the incomplete data that was on the website for the 2023-2024 NBA season that is currently underway.

Once the data was all cleaned and merged into one file we started to do some analysis of it to see which variables were most useful in seeing how VORP changes from season to season for players. For estimating VORP we used various methods including maximum likelihood (ML) and restricted maximum likelihood (REML). We then created a series of different multilevel models to see which ones were best at modeling VORP. Some of the comparisons between models were ones with and without random intercepts for variables and ones with and without random slopes for variables.

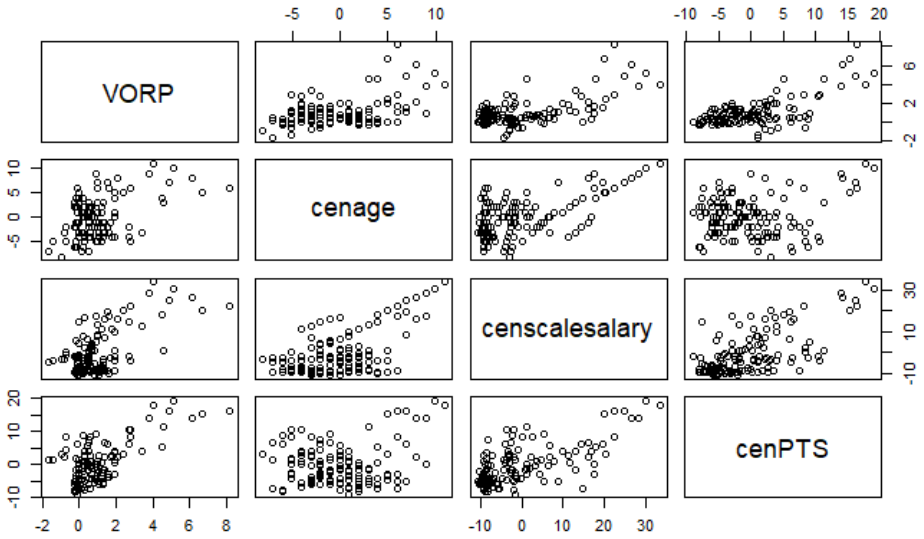
Results

Figure 2: Distribution Of Seasonal VORP By Player



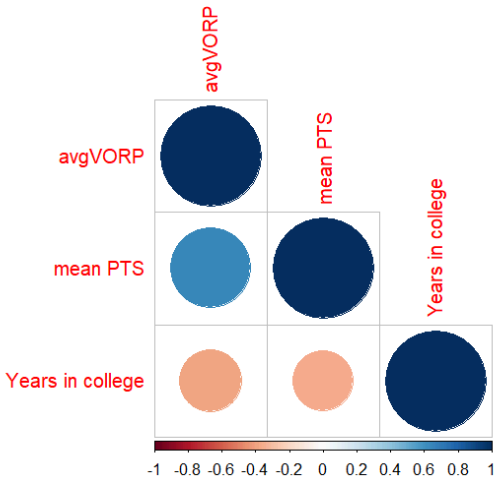
The first step taken in the exploratory data analysis was to analyze the distribution of VORP across the players, shown in figure 2. Players have differing average VORP values for the their, along with differing variations. The differing amount of variation leads us to question the validity of the equal variance assumption for the data. Because of the equal variance assumption being violated, we believe it is worthwhile to investigate random slopes to model the heterogeneity.

Figure 3: Matrix Scatterplot Between Level 1 Variables and VORP



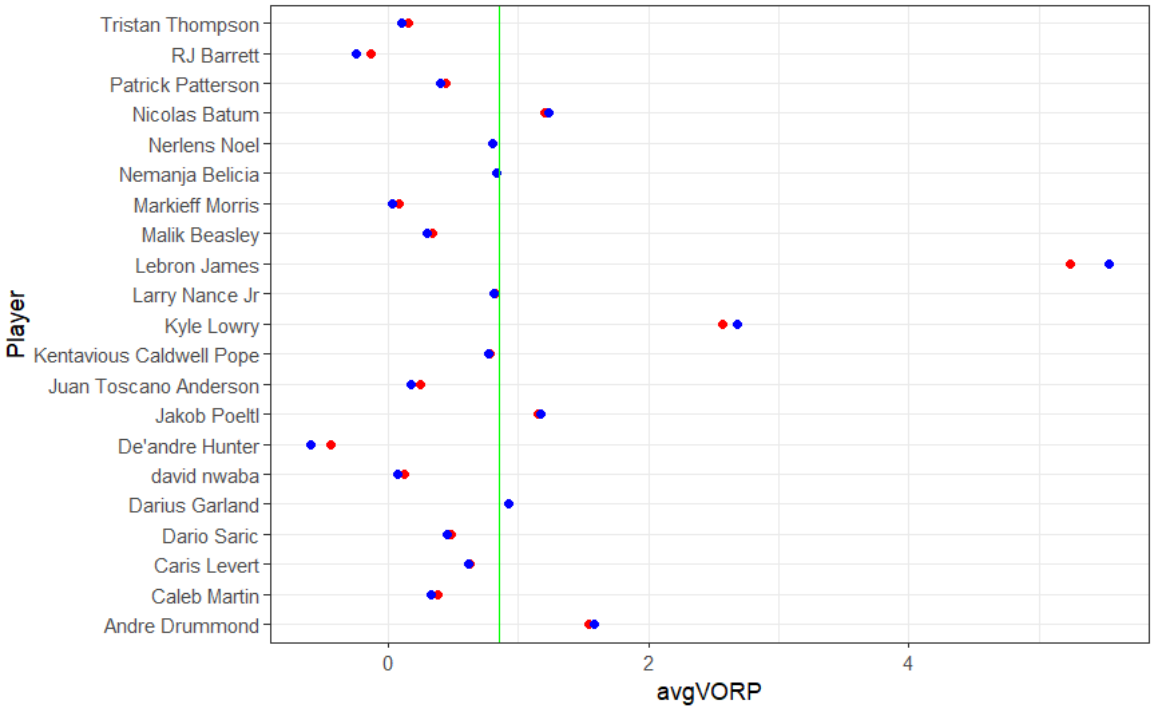
Our next step in our exploratory data analysis was to assess which level 1 variables seem promising to add. Variables that have correlation with VORP would be useful to add to the model. Based on the figure above, cenPTS looks to have the strongest correlation with VORP, compared to cenage and censcalesalary. Again, cenPTS is a centered points per game variable by season, cenage is the centered age of a player within a particular season, and censcalesalary is the centered, scaled (millions), annual salary for a given player in a particular season.

Figure 4: Correlation Matrix Between Vorp & Level 2 Variables



For the level 2 variables, mean PTS is a player's average points per game throughout the 2016-2023 seasons. Years in college is the number of years the player played basketball at the college level, and bindraft is a binary variable where "1" indicates that the player was drafted in the lottery portion of the draft (indicating they had high expectations to perform well). Assessing our level 2 variables using the correlation matrix depicted in figure 4, the strongest correlation is a positive one between mean PTS (a player's career average points per game) and avgVORP (a player's overall average VORP for the 2016-2023 seasons). Additionally, the correlation strengths between VORP and Years in college (how many years in college a player spent) is weaker and negative, leading us to conclude that mean PTS is the most promising level 2 variable. Bindraft is not included because it is a binary categorical variable.

Figure 5: Null Model Shrinkage Effects & Null Model Equation



$$VORP_{ij} = \beta_{0j} + \varepsilon_{ij}$$
$$\beta_{0j} = \beta_{00} + u_{0j}$$

Our null model includes random intercepts for players and no other variables (Output in appendix A). The predicted VORP for an average player in a particular season is 0.8650 (Intercept). The variation in predicted VORP for an average player in a particular season is about 1.464 ($\widehat{\tau_0}$). The variation in predicted VORP within a season is about 0.705 ($\widehat{\sigma}$). 67.49% of the variation in VORP (ICC) is due to different players.

From our null model, we decided to add level 1 variables to see if they would significantly improve the effectiveness of our initial model. The first implemented model included cenage, censcalesalary, and cenPTS and was a significant improvement over the null model (Appendix B) . Of all three of the level 1 variables cenage had the lowest absolute T-value (0.039), so we decided this would be the first variable we remove. Upon removing this variable from the model our resulting model's significant variables did not lose significance therefore the removal of cenage is justified (Appendix C). Furthermore, censcalesalary did not become significance, its T-value was -1.937, so we removed it from the model. Because the T-value of cenage was close to -2, we opted to use a LRT to verify that removing the variable is justified (Appendix D). The resulting p-value of the LRT was .09452 (greater than .05) meaning the removal of cenage is justified since including cenage does not produce a significantly better model. These removals resulted in a model including only cenPTS and random intercepts by player. Comparing the null model to the model with cenPTS added to it, the proportion of additional total variation in VORP explained by the model is 37.83% with the addition of centered points to the null model (see Appendix E).

Moving onto the level 2 variables, we added all three level 2 variables, mean PTS, bindraft, and years in college, to our prior model (Appendix G). To see if the addition of these variables significantly improves the model we performed an LRT between the two models which yielded a p-value of .492

leading to the conclusion that the addition of these variables does not significantly improve the model. After undergoing the process outlined for removing insignificant level 1 variables for level 2 variables, we concluded that the addition of any of the level 2 variables is insignificant. However, we opt to include a cross-level interaction for the sake of creating an interesting model.

As noted in our initial hypothesis, we also believe that the rate at which VORP changes varies between players. To investigate this claim we opted to create a random slopes model that allows players' VORPs between seasons to vary by cenPTS (Appendix H). To see if the inclusion of random slopes by cenPTS significantly improves the model we performed an LRT between our previous model and our new random slopes model which yielded a P-value of 0.001715. This P-value is below .05 leading to the conclusion that the addition of the random slopes significantly improves the performance of our model.

Figure 6: R Output For Final Model & Final Model Equation

```
boundary (singular) fit: see help('isSingular')
Linear mixed model fit by REML ['lmerMod']
Formula: VORP ~ 1 + cenPTS * `Years in college` + (1 + cenPTS | Player)
   Data: cenclean

REML criterion at convergence: 328.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3415 -0.4800 -0.0595  0.5271  3.6757

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Player   (Intercept) 0.445177 0.66722
          cenPTS      0.003479 0.05898  1.00
 Residual             0.535360 0.73168
Number of obs: 126, groups:  Player, 21

Fixed effects:
                           Estimate Std. Error t value
(Intercept)                 1.17331    0.27065   4.335
cenPTS                      0.15303    0.03749   4.081
`Years in college`         -0.18345    0.11984  -1.531
cenPTS:`Years in college` -0.02726    0.01848  -1.475

Correlation of Fixed Effects:
            (Intr) cenPTS `Yicl`
cenPTS       0.589
`Yrsincllg` -0.782 -0.519
cnPTS:`Yic` -0.461 -0.771  0.701
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

$$VORP_{ij} = \beta_{0j} + \beta_{1j}\text{cenPTS}_{ij} + \varepsilon_{ij}$$
$$\beta_{0j} = \beta_{00} + \beta_{01}\text{yearsincollege}_j + u_{0j}$$
$$\beta_{1j} = \beta_{10} + \beta_{11}\text{yearsincollege}_j + u_{1j}$$

The predicted VORP in a season for an average player who scores the average amount of points per game and spent 0 years in college is 1.17331. The predicted VORP in a season for a player who spent 0 years in college is expected to increase by 0.153 for each additional point per game they score for the season. For each additional year spent in college, the predicted VORP for a player who scores the average number of points per game is predicted to decrease by -0.18345. For every additional year spent in college, the effect on VORP of scoring an additional point above the average decreases by .02726. Our $\hat{\tau}_0^2$ (.445) is player to player variability in VORP for players who score the average number of points per

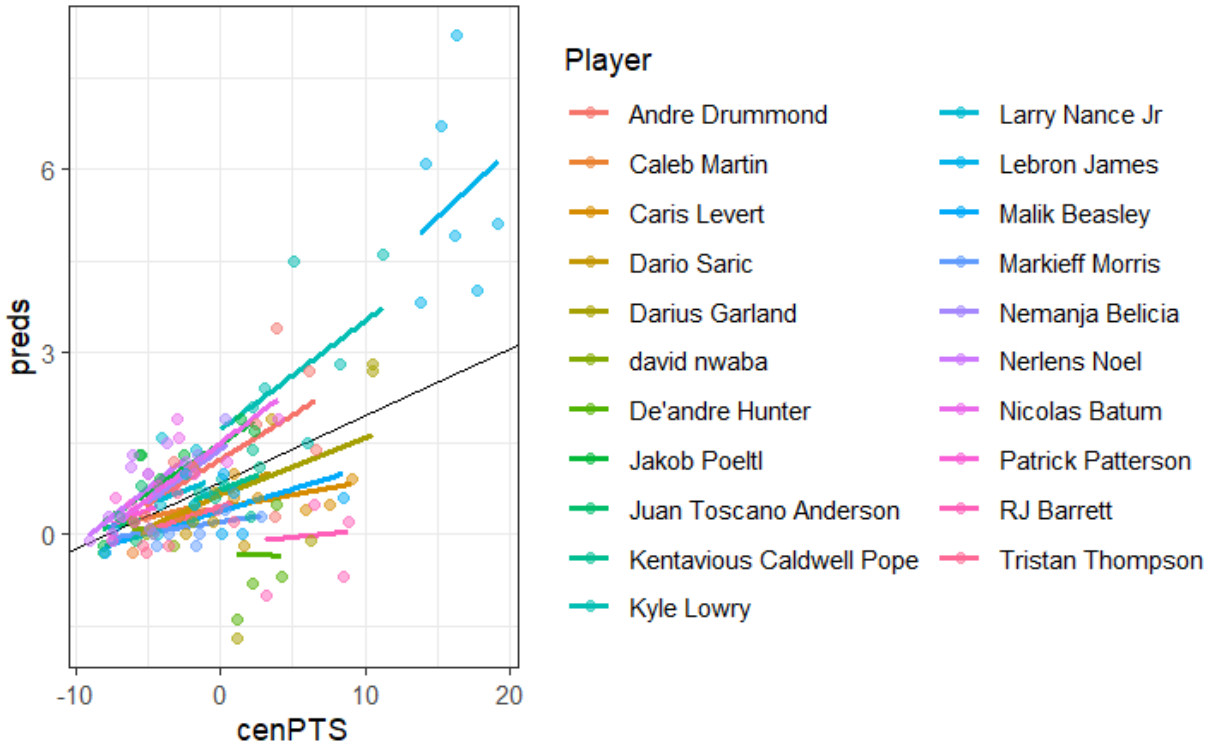
game, and spent 0 years in college. The $\widehat{\tau}_1^2$ (.003479) is between player variation in the effect of centered points per game on VORP.

$\widehat{\tau}_{01}$ = 1*.05898*.66722 = 0.03935. Although τ_{01} is the parameter, the correlation of 1 between random slopes and random intercepts is much easier to interpret. This indicates that we expect a player with the highest predicted VORP in earlier seasons to ALWAYS have the largest season-to-season increase in predicted VORP for each additional point scored per game. Lastly, $\widehat{\sigma}^2$ (.535360) is unexplained within season variation in VORP.

Interval of population slopes (in place of a confidence interval as we found it more interesting): $\widehat{\beta}_1$ +/- $2\widehat{\tau}_1$ = 0.15303 +/- 2*0.003479 = (0.146072, 0.159988). We expect 95% of player slopes for (centered) points per game to be within 0.146072 and 0.159988.

Additionally, 69.59% of Level 2 variation in the null model is explained by the final model, and 24.06% Level 1 variation in the null model is explained by the final model.

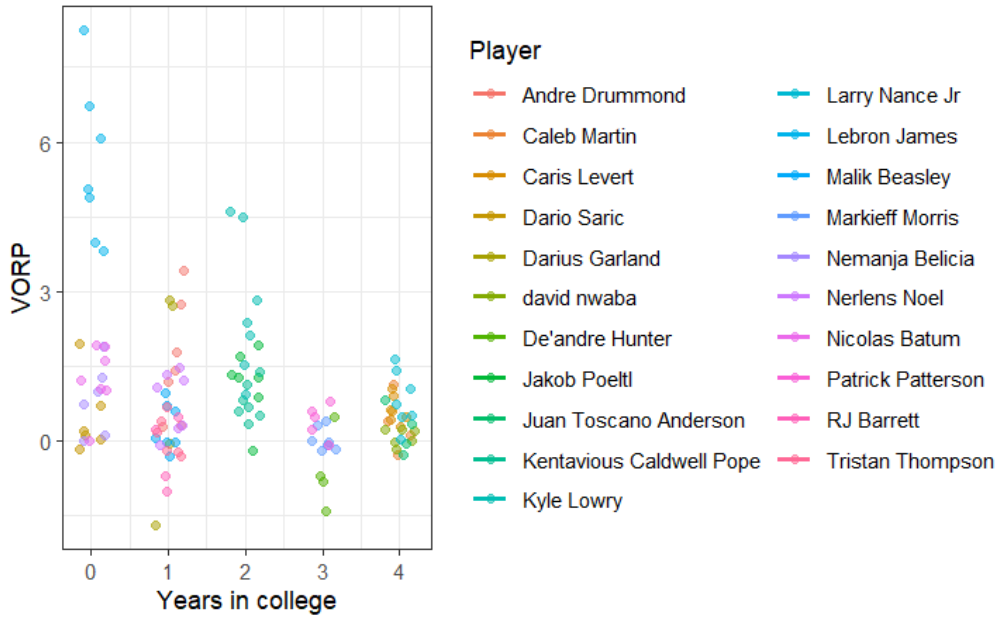
Figure 7: Plotting the Final Model By cenPTS



Each line shows player-based predictions from the model with random slopes and intercepts, and the points are raw data. Notice that Lebron, with the highest intercept, also has the highest slope. Our model coincides with our raw data, as cenPTS increases so does the VORP values. This matches our positive coefficient for cenPTS in our final model. Additionally, note that for players who start out averaging a lot of points per game (high intercept) their slopes (how an additional point per game affects VORP) tend to be more positive which matches our positive Tau-01 estimate. Through the graph it can also be seen that the residuals seem to increase as cenPTS increases. This relationship is likely due to Lebron James being present in the dataset. Upon investigating the data, Lebron's largest outlier season

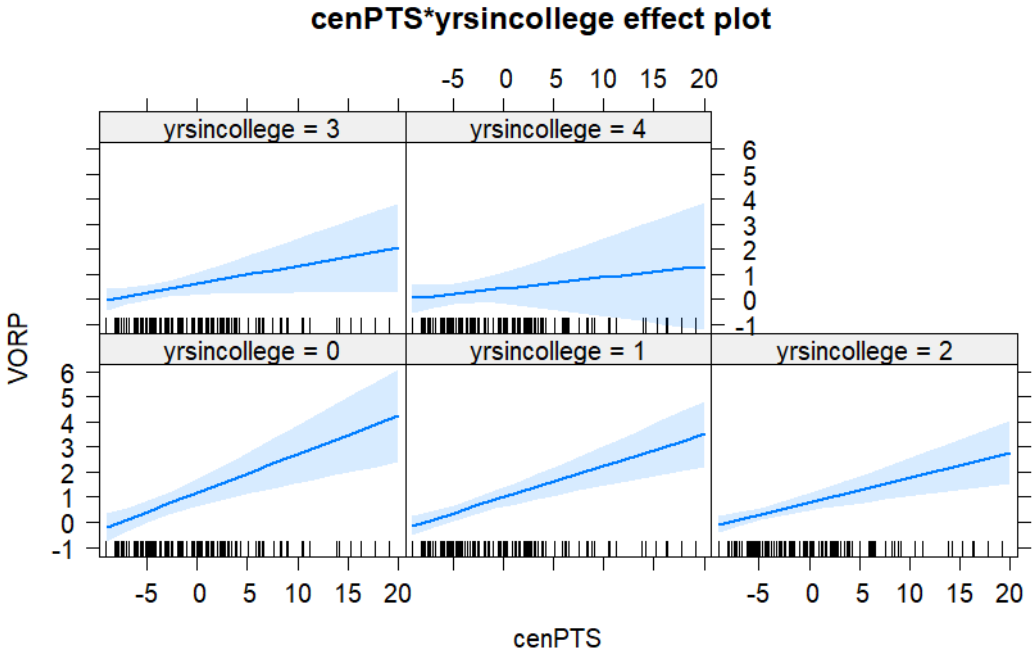
came from the 2017-2018 season and the VORP value was 8.2. Lebron James adds "value" to games through his rebounding, assists, and defense in addition to his points per game, which explains why it may be difficult for the model to accurately predict his VORP values.

Figure 8: Plotting the VORP by Years in College



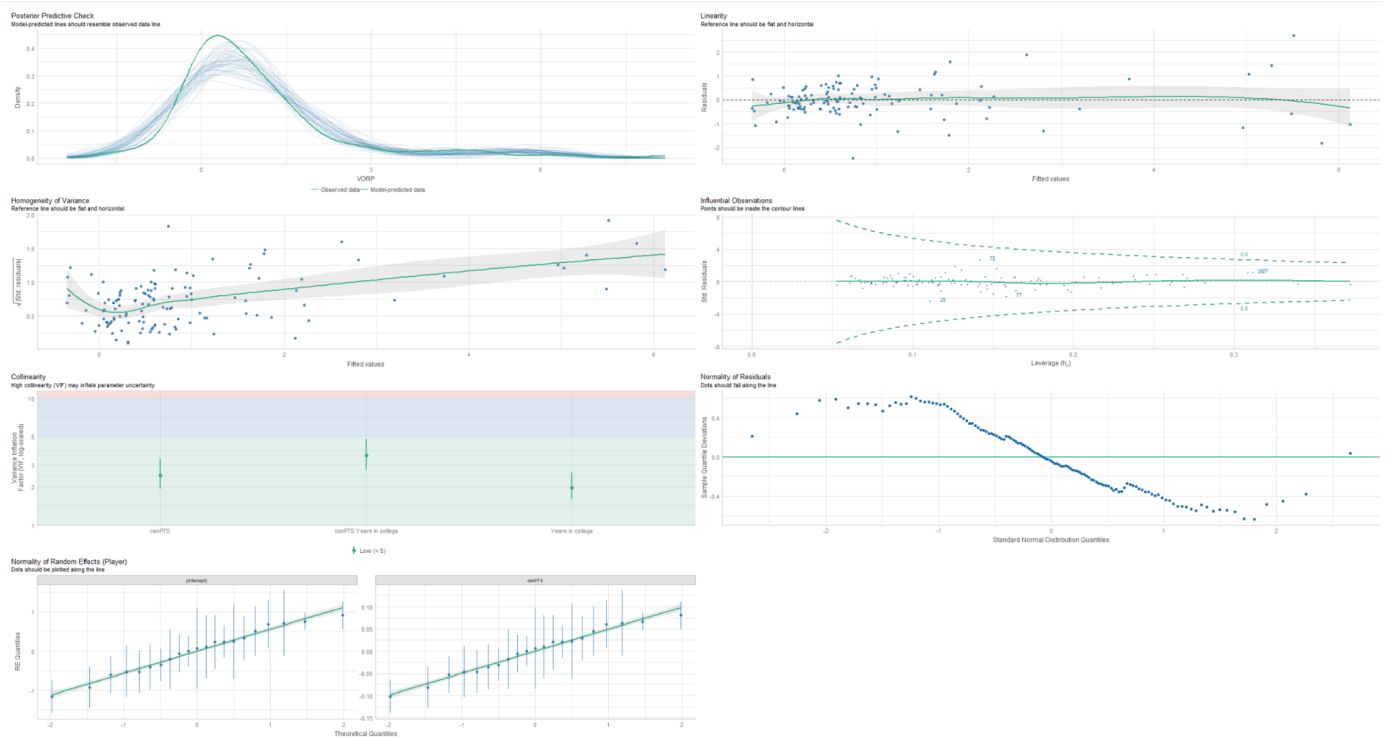
This graph of the raw data shows that players with less years spent playing in college typically have a smaller VORP. Thus, the raw data helps validate the negative coefficient calculated for years in college for the final model.

Figure 9: Interaction Effect Plots for CenPTS and Years in College



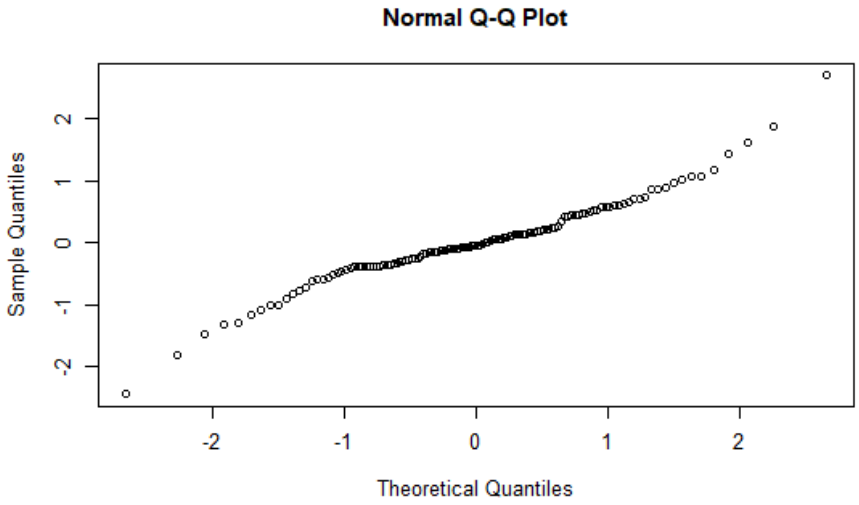
This graph is similar to Figure 8 in that it shows the final model predicts a smaller effect of cenPTS on predicted VORP with each additional year a player spends playing in college. The model predicts that a jump from 15 to 20 cenPTS will have a much larger impact on VORP for players who spent 0 years in college than players who spent 4 years in college. Players with less years in college might typically be more talented and suited for pro play than players who spend longer playing in college.

Figure 10: Performance Final Model Diagnostics Output (LINE Assumptions and Residual Plots)



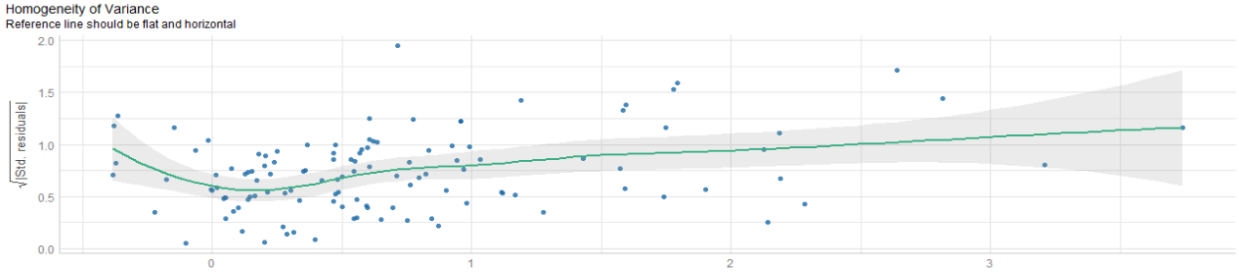
Looking at each of the residual plots, it appears that there are not too many issues with our final model. Our posterior/predictive check is a tad questionable but still reasonable, and our linearity assumption appears to be met (via the horizontal line). The main issue appears to be with the homogeneity of variance assumption, with the graph showing a trend that is not a horizontal line (which can be fixed by using an AR model that we will check out later on). This makes sense because VORP across players do not appear to have equal variance (observing Figure 2). VIFs don't appear to be an issue, and the normality of the random effects also appears reasonable. ADD: without lebron our homogeneity of variance would probably look fine, those rightmost obs are most likely him

Figure 11: Fixed Normality Plot Final Model



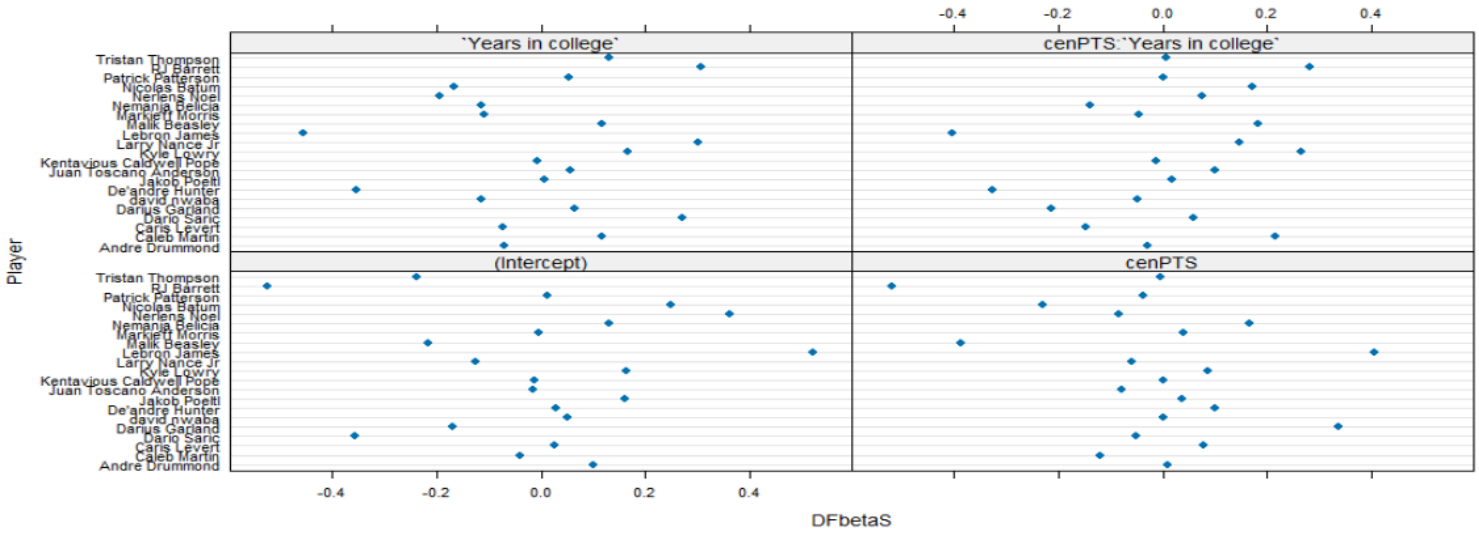
Normal Q-Q plot indicates that our assumption of normality of residuals is met. Although it is worth noting that there does appear to seem outliers at both ends of the graph, which are observations from Darius Garland and Lebron James.

Figure 12: Homogeneity of Variance without Lebron James in Data



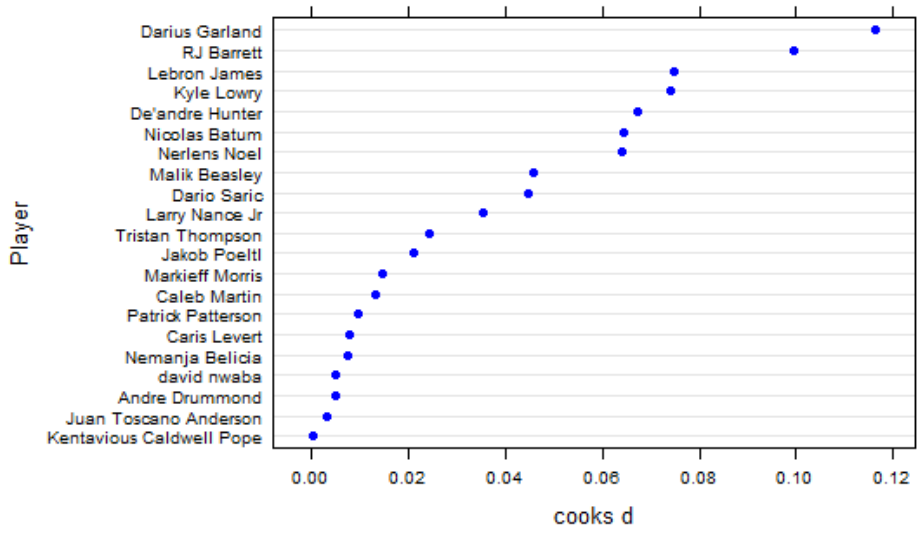
Without Lebron James in the dataset, our only assumption that looked a bit worrisome, homogeneity of variance, looks a lot better. Notably, all the fitted values above 4 are absent, indicating that they were all Lebron's, and they were the main issue that was "dragging" the curve upwards in Figure 10. Because the outliers from Lebron are removed, the model is able to better predict higher fitted values.

Figure 13: DFBetas Graph Final Model (Cutoff: 0.4364)



DFBetas are measured for each observation in the dataset and indicate how much the slope coefficient of a variable changes if that observation is removed. The corresponding cutoff calculated for our data was .4364, and no observations crossed that threshold. Meaning that for all variables, there are not any observations that would significantly alter any slope's coefficient if removed from the data.

Figure 14: Cook's Distance Graph Final Model (Cutoff: 0.2105)



Cook's Distance is measured for each observation and provides an overall measure of how much all of the model predictions change when that observation is removed and is found by considering each observation's leverage and residual. No observation reached the respective cutoff of 0.2105, thus no observations significantly changed the model's predictions when removed.

Discussion

To restate our initial research question, we hypothesize that we can accurately predict VORP using variables that are not directly included in the VORP calculation, and that the effects these variables have on VORP vary between players. Noted in figure 7, our final model seems to accurately predict VORP utilizing cenPTS, years in college, and their interaction as fixed effects, players as random effects along with random slopes allowing the effect of cenPTS on VORP to vary between players. Additionally, because the model with random slopes significantly outperforms the model without random slopes we can conclude that the effects cenPTS has on VORP varies between players.

During our research we did run into some limitations that hindered our final results. One aspect to consider is the relatively small sample size of our population. The study's data only includes 21 players, whereas there were hundreds of NBA players in the league from 2016-2023. This can lead to skewed, inaccurate conclusions if a nonrepresentative sample is randomly drawn. Furthermore, we only considered data from 2016-2023. This is a small window of time in the NBA's existence, so it is difficult to extrapolate our findings to all players within all eras. A better alternative may be to either include more players or expand the scope of the study to a larger time frame. Additionally, some of the players sampled were not in the league for all 7 years. Another limitation is the correlation structure utilized in the final model. Because the data is longitudinal, it may make sense to incorporate an autoregressive(1) correlation structure rather than relying solely on random slopes within the multilevel model to properly model the heteroscedasticity within the data. Possible confounding variables include a player's teammates or team, since certain organizations within the NBA consistently perform better than others. Organizations and teams prove vital in developing a players career, however it can be difficult to quantify an organization's effect on a players development.

Further research that can be done includes exploring alternative time series models to better model the correlation structure. While we did explore this in this study, when attempting to include an AR(1) correlation structure to our final model it was too complicated and could not converge (Appendix J, also for some more AR(1) exploration). However, this model that did not converge did include both random slopes and an AR(1) correlation structure, which is likely not necessary. Additionally, the level 1 time variables included in this data are not the most optimal, if further research was to be done a variable noting how many years in the league a player has by season would be useful. Another possibility is to compare the results found in this study across other 7 year time periods in the NBA to see if the trends found in the research remain true when examining other eras of the NBA. Another relationship that was not explored is between players and organizations. Attempting to quantify how much certain organizations benefit players would be interesting to explore. Additionally, this could be explored in a non-hierarchical multilevel model. For example, data regarding each team's seasons could be included as level 1 units and overall data from each team over the 7 year period could be included as level 2 units.

Appendix A

```r
model1 <- lmer(VORP~1 + (1|Player), data=cenclean, REML="FALSE")
summary(model1)
logLik(model1)
AIC(model1)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: VORP ~ 1 + (1 | Player)
   Data: cenclean

     AIC      BIC   logLik deviance df.resid
   373.7    382.2   -183.8    367.7      123

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1187 -0.4227 -0.1191  0.4191  3.5231

Random effects:
 Groups   Name        Variance Std.Dev.
 Player   (Intercept) 1.464    1.2101
 Residual             0.705    0.8396
Number of obs: 126, groups:  Player, 21

Fixed effects:
            Estimate Std. Error t value
(Intercept)   0.8654     0.2749   3.148
'log Lik.' -183.8448 (df=3)
[1] 373.6897
```

Our null model that only includes random intercepts for different players with the addition of no level 1 or two variables.

Appendix B

```r
model2 <- lmer(VORP~1 + cenage + censcalesalary + cenPTS + (1|Player), data=cenclean, REML="FALSE")
summary(model2)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: VORP ~ 1 + cenage + censcalesalary + cenPTS + (1 | Player)
   Data: cenclean

     AIC      BIC   logLik deviance df.resid
   335.5    352.5   -161.7    323.5      120

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6413 -0.5228 -0.0722  0.5144  3.8343

Random effects:
 Groups   Name        Variance Std.Dev.
 Player   (Intercept) 1.0999   1.0488
 Residual             0.4906   0.7004
Number of obs: 126, groups:  Player, 21

Fixed effects:
               Estimate Std. Error t value
(Intercept)     0.851201   0.237848   3.579
cenage         -0.001316   0.033471  -0.039
censcalesalary -0.025376   0.014455  -1.755
cenPTS          0.153077   0.021862   7.002

Correlation of Fixed Effects:
            (Intr) cenage cnscls
cenage       0.023
censcalslry  0.025 -0.415
cenPTS      -0.008  0.223 -0.460
```

```r
anova(model2, model1)
```

```
Data: cenclean
Models:
model1: VORP ~ 1 + (1 | Player)
model2: VORP ~ 1 + cenage + censcalesalary + cenPTS + (1 | Player)
       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model1    3 373.69 382.20 -183.84   367.69
model2    6 335.49 352.51 -161.75   323.49 44.199  3  1.369e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A new model that inlcudes random intercepts for player and all of our level 1 variables. The anova shows that this new model is significantly better at modeling VORP than the null model.

Appendix C

```r
#removing least sig var
model2d <- lmer(VORP~1 + censcalesalary + cenPTS + (1|Player), data=cenclean, REML="FALSE")
summary(model2d)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: VORP ~ 1 + censcalesalary + cenPTS + (1 | Player)
   Data: cenclean

     AIC      BIC   logLik deviance df.resid
   333.5    347.7   -161.7    323.5      121

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6361 -0.5195 -0.0715  0.5114  3.8360

Random effects:
 Groups   Name        Variance Std.Dev.
 Player   (Intercept) 1.0929   1.0454
 Residual             0.4912   0.7008
Number of obs: 126, groups:  Player, 21

Fixed effects:
             Estimate Std. Error t value
(Intercept)   0.85157    0.23709   3.592
censcalesalary -0.02549    0.01316  -1.937
cenPTS         0.15323    0.02131   7.189

Correlation of Fixed Effects:
            (Intr) cnscls
censcalslry  0.038
cenPTS      -0.013 -0.414
```

Our third model that takes out the least significant level 1 variable which was cenage.

Appendix D

```r
model2e <- lmer(VORP~1 + cenPTS + (1|Player), data=cenclean, REML="FALSE")
summary(model2e)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: VORP ~ 1 + cenPTS + (1 | Player)
   Data: cenclean

     AIC      BIC   logLik deviance df.resid
   334.3    345.6   -163.1    326.3      122

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6571 -0.5589 -0.0535  0.5203  3.8861

Random effects:
 Groups   Name        Variance Std.Dev.
 Player   (Intercept) 0.8173   0.9041
 Residual             0.5312   0.7288
Number of obs: 126, groups:  Player, 21

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.87259    0.20817   4.192
cenPTS       0.13743    0.01944   7.071

Correlation of Fixed Effects:
       (Intr)
cenPTS 0.003
```

```r
anova(model2d, model2e)
```

```
Data: cenclean
Models:
model2e: VORP ~ 1 + cenPTS + (1 | Player)
model2d: VORP ~ 1 + censcalesalary + cenPTS + (1 | Player)
        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model2e    4 334.29 345.63 -163.14   326.29
model2d    5 333.49 347.67 -161.75   323.49 2.7957  1    0.09452 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Our fourth model that takes out the next least significant level 1 variable which was censcalesalary. An anova test that compares models with only cenPTS and random intercepts and a model with cenPTS,censcalesalary, and random intercepts. We see that the model that has both level one is not significantly better than the model with just cenPTS.

Appendix E

```r
model2c <- lmer(VORP~1 + cenPTS + (1|Player), data=cenclean, REML="FALSE")
summary(model2c)

#obtained from the null model
tausq0 <- 1.464
sigmasq0 <- 0.705

tausq1 <- 0.8173
sigmasq1 <- 0.5312

((tausq0 + sigmasq0) - (tausq1 + sigmasq1))/(tausq0 + sigmasq0)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: VORP ~ 1 + cenPTS + (1 | Player)
   Data: cenclean

     AIC      BIC   logLik deviance df.resid
   334.3    345.6   -163.1    326.3      122

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6571 -0.5589 -0.0535  0.5203  3.8861

Random effects:
 Groups   Name        Variance Std.Dev.
 Player   (Intercept) 0.8173   0.9041
 Residual             0.5312   0.7288
Number of obs: 126, groups:  Player, 21

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.87259    0.20817   4.192
cenPTS       0.13743    0.01944   7.071

Correlation of Fixed Effects:
       (Intr)
cenPTS 0.003
[1] 0.3782849
```

Model 2c is our best model containing only level one variables, as shown with the calculations we see the additional amount of level 1 variation that was explained when compared to the null model.

Appendix F

```r
model3 <- lmer(VORP~ 1 + cenPTS + `Years in college` + `mean PTS` + bindraft + (1|Player), data=cenclean, REML="FALSE")
summary(model3)
```

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: VORP ~ 1 + cenPTS + `Years in college` + `mean PTS` + bindraft +      (1 | Player)
   Data: cenclean

     AIC      BIC   logLik deviance df.resid
   337.9    357.7   -161.9    323.9      119

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.7252 -0.5395 -0.0421  0.5385  3.8847

Random effects:
 Groups   Name        Variance Std.Dev.
 Player   (Intercept) 0.7192   0.8481
 Residual             0.5311   0.7288
Number of obs: 126, groups:  Player, 21

Fixed effects:
                  Estimate Std. Error t value
(Intercept)        1.40336    0.72548   1.934
cenPTS             0.13250    0.02246   5.899
`Years in college` -0.19994    0.15051  -1.328
`mean PTS`         0.01163    0.04559   0.255
bindraft          -0.47389    0.42926  -1.104

Correlation of Fixed Effects:
            (Intr) cenPTS `Yicl` `mPTS`
cenPTS       0.345
`Yrsincllg` -0.681  0.000
`mean PTS`  -0.759 -0.493  0.259
bindraft    -0.363  0.000  0.288 -0.130
```

```r
anova(model2e, model3)
```

```
Data: cenclean
Models:
model2e: VORP ~ 1 + cenPTS + (1 | Player)
model3: VORP ~ 1 + cenPTS + `Years in college` + `mean PTS` + bindraft + (1 | Player)
        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model2e    4 334.29 345.63 -163.14   326.29
model3     7 337.88 357.73 -161.94   323.88 2.4087  3      0.492
```

Model 3 now contains all the level 2 variables that we tested in addition to the best level 1 variables that was decided in the outputs above. We see from the anova that when all the level 2 variables are added together there is not a significant amount of variation in VORP explained when compared to only having level 1 variables.

Appendix G

I. Least significant level 2 variable “mean PTS” is removed

```
```{r}
model3d <- lmer(VORP~ 1 + cenPTS + `Years in college` + bindraft  + (1|Player), data=cenclean, REML = "FALSE")
summary(model3d)
```

Linear mixed model fit by maximum likelihood ['lmerMod']
Formula: VORP ~ 1 + cenPTS + `Years in college` + bindraft + (1 | Player)
   Data: cenclean

     AIC   BIC   logLik deviance df.resid
   335.9   353.0   -162.0   323.9   120

Scaled residuals:
    Min   1Q Median   3Q   Max
-2.6915 -0.5466 -0.0410 0.5356 3.8982

Random effects:
 Groups   Name   Variance Std.Dev.
 Player   (Intercept) 0.7225   0.8500
 Residual   0.5310   0.7287
Number of obs: 126, groups: Player, 21

Fixed effects:
                  Estimate Std. Error t value
(Intercept)   1.54383   0.47356   3.260
cenPTS   0.13532   0.01956   6.920
`Years in college` -0.20988   0.14567 -1.441
bindraft   -0.45968   0.42645 -1.078

Correlation of Fixed Effects:
            (Intr) cenPTS `Yicl`
cenPTS   -0.051
`Yrsincllg` -0.770 0.151
bindraft   -0.715 -0.074 0.336

II. Next least significant level 2 variable “bindraft” is removed

```
```{r}
model3e <- lmer(VORP~ 1 + cenPTS + `Years in college`  + (1|Player), data=cenclean)
summary(model3e)
```

Linear mixed model fit by REML ['lmerMod']
Formula: VORP ~ 1 + cenPTS + `Years in college` + (1 | Player)
   Data: cenclean

REML criterion at convergence: 334.5

Scaled residuals:
    Min   1Q Median   3Q   Max
-2.6932 -0.5247 -0.0304 0.5222 3.8331

Random effects:
 Groups   Name   Variance Std.Dev.
 Player   (Intercept) 0.8678   0.9316
 Residual   0.5351   0.7315
Number of obs: 126, groups: Player, 21

Fixed effects:
                  Estimate Std. Error t value
(Intercept)   1.17641   0.35948   3.272
cenPTS   0.13349   0.01994   6.695
`Years in college` -0.15668   0.14869 -1.054

Correlation of Fixed Effects:
            (Intr) cenPTS
cenPTS   -0.141
`Yrsincllg` -0.804 0.178

Appendix H

```r
model4 <- lmer(VORP~ 1 + cenPTS + (1 + cenPTS|Player), data=cenclean, REML="TRUE")
summary(model4)
```

```
boundary (singular) fit: see help('isSingular')
Linear mixed model fit by REML ['lmerMod']
Formula: VORP ~ 1 + cenPTS + (1 + cenPTS | Player)
   Data: cenclean

REML criterion at convergence: 321.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.3772 -0.5044 -0.0974  0.4992  3.7067

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Player   (Intercept) 0.478304 0.6916
          cenPTS      0.004502 0.0671   1.00
 Residual             0.532954 0.7300
Number of obs: 126, groups:  Player, 21

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.86871    0.17143   5.067
cenPTS       0.10958    0.02505   4.374

Correlation of Fixed Effects:
       (Intr)
cenPTS 0.697
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

Model 4 incorporated random slopes for cenPTS which allows the average points to vary from player to player.

Appendix I

```
```{r}
model2eREML <-  lmer(VORP~1 + cenPTS + (1|Player), data=cenclean, REML="TRUE")
summary(model2eREML)

anova(model4, model2eREML)
```
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: VORP ~ 1 + cenPTS + (1 | Player)
   Data: cenclean

REML criterion at convergence: 333.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.6402 -0.5538 -0.0552  0.5147  3.8575

Random effects:
 Groups   Name        Variance Std.Dev.
 Player   (Intercept) 0.8749   0.9354
 Residual             0.5350   0.7314
Number of obs: 126, groups:  Player, 21

Fixed effects:
            Estimate Std. Error t value
(Intercept)  0.87190    0.21473   4.060
cenPTS       0.13719    0.01964   6.987

Correlation of Fixed Effects:
       (Intr)
cenPTS 0.003
refitting model(s) with ML (instead of REML)
Data: cenclean
Models:
model2eREML: VORP ~ 1 + cenPTS + (1 | Player)
model4: VORP ~ 1 + cenPTS + (1 + cenPTS | Player)
            npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model2eREML    4 334.29 345.63 -163.14   326.29
model4         6 325.55 342.57 -156.78   313.55 12.736  2   0.001715 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This output uses anova to compare the model with only our level 1 variables to the model in appendix H with random slopes for cenPTS, and we see that there is a significant improvement between the two models.

Appendix J

Error code with creating an AR model with random slopes:

```r
modelAR = lme(VORP ~ seasonnum + cenPTS, random = ~1 + cenPTS | Player,
correlation=corAR1(), data = cenclean)
summary(modelAR)
```

```
Show Traceback

Error in lme.formula(VORP ~ seasonnum + cenPTS, random = ~1 + cenPTS |
   nlminb problem, convergence error code = 1
   message = iteration limit reached without convergence (10)
```

AR model created without random slopes:

```r
modelAR = lme(VORP ~ seasonnum + cenPTS, random = ~1  | Player, correlation=corAR1(), data = cenclean)
summary(modelAR)
```

```
Linear mixed-effects model fit by REML
  Data: cenclean

Random effects:
 Formula: ~1 | Player
        (Intercept)  Residual
StdDev:   0.8367025 0.8061825

Correlation Structure: AR(1)
 Formula: ~1 | Player
 Parameter estimate(s):
      Phi
0.4298206
Fixed effects:  VORP ~ seasonnum + cenPTS
 Correlation:
          (Intr) sesnnm
seasonnum -0.568
cenPTS    -0.018  0.056
```

Estimates from the AR model:

	Value <chr>	Std.Error <chr>	DF <chr>	t-value <chr>	p-value <chr>
(Intercept)	1.1060890	0.25476416	103	4.341619	0.0000
seasonnum	-0.0754987	0.04532091	103	-1.665869	0.0988
cenPTS	0.1424846	0.02085645	103	6.831681	0.0000

ARModel compared to final model:

```r
texreg::screenreg(list(model5, modelAR), digits = 3, single.row = TRUE, stars = 0, custom.model.names=c("model5", "modelAR"))
```

```
=============================================================
                                model5            modelAR
-------------------------------------------------------------
(Intercept)                      1.173 (0.271)     1.106 (0.255)
cenPTS                           0.153 (0.037)     0.142 (0.021)
`Years in college`              -0.183 (0.120)
cenPTS:`Years in college`       -0.027 (0.018)
seasonnum                                         -0.075 (0.045)
-------------------------------------------------------------
AIC                            344.213           332.459
BIC                            366.903           349.332
Log Likelihood                -164.106          -160.229
Num. obs.                      126               126
Num. groups: Player            21                21
Var: Player (Intercept)          0.445
Var: Player cenPTS               0.003
Cov: Player (Intercept) cenPTS   0.039
Var: Residual                    0.535
=============================================================
* p < 0
```