Who is DRAYMOND? A Longitudinal Analysis of Defensive Performance in the NBA

Ben Brennan, Debraj Bose, Soumik Purkayastha, Keejeong Ryu, Jieru Shi

University of Michigan, Biostatistics

December 21, 2019

1 Introduction

1.1 Overview

The recent rise and availability of big data in sports has made understanding the performance of athletes more accessible. In this report, we analyze 239 NBA players over the time period of the 2013-2014 season to the 2016-2017 season in hopes of understanding the factors that affect a player's defensive performance. Defensive performance that is notoriously undervalued and, thus, under-studied. Unsurprisingly, people have been interested in modeling offensive performance longitudinally [1] [2] but, as far we know, there is no study yet to do the same with defensive performance. To do this, we collected an umbrella metric for defense - each player's DRAYMOND score [3] over the course of the seasons. First, we fit a simple mean model to the data to understand what the major differences in players were across career status, position and time in order to see if any of those factors interacted with one-another. After deciding on a mean model, we fit a linear mixed model (LMM) to the data and assess the need for random effects. We also fit a marginal model, using GEE, to understand the trend over time in the population. Third, we use the model to make predictions on future defensive performance. Finally, we try to explain the leftover variance by modeling the residuals, stratified by position, as a function of physical characteristics.

1.2 Objective

The goal of our study is to better understand what makes a player a good defender. Does a player tend to get worse or better, on average, over time? Is his defensive performance related to the amount of time he has been in the league? Do certain positions objectively play better defense than others? What else can explain the variation between players that is not explained by these factors and individual differences? Through this study, we hope to provide insights into these questions and to the general idea of how to model defensive performance.

A secondary goal we have is to see what our model can predict in regards to the future. In particular, we want to know: Is Draymond Green the most complete defensive player in the game?

2 Methods

2.1 Mean Model

First, to get a start on modeling, we fit a mean model to the data. The definitions below will be used through the rest of the report.

- Rookie (time in NBA is less than 5 years in 2017)
- Mid-career (time in NBA is between 5 and 10 years in 2017)
- Late-career (time in NBA is greater than 10 years in 2017)

- Pos (Position)

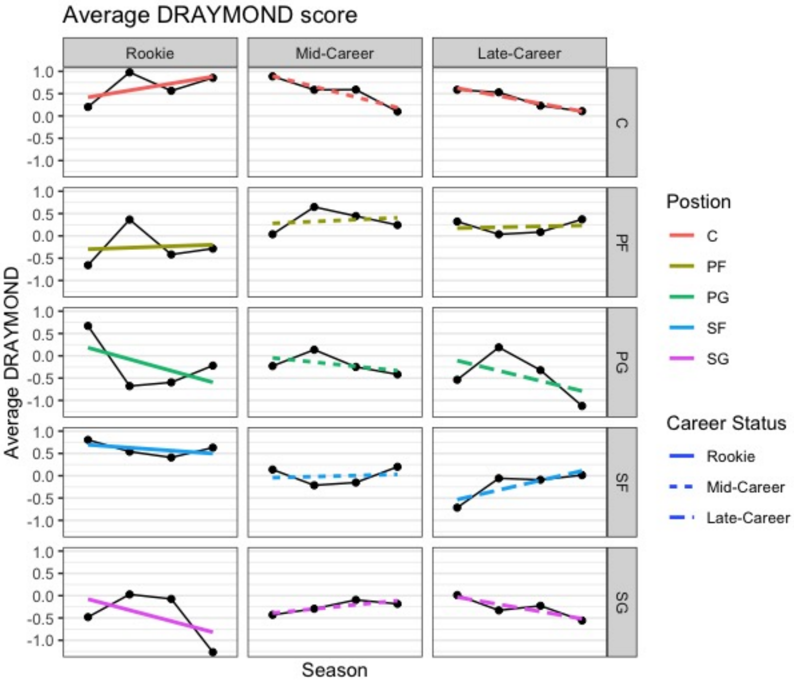

Figure 1: DRAYMOND Score by Position and Time

Figure 1 shows that DRAYMOND score over time tends to change differently with respect to position and career status. Including a position, career status interaction would add significant complexity to our model so, instead, we posit the following mean model to describe the data.

$$\text{DRAYMOND} = \beta_0 + \beta_1 \times \text{season} + \textstyle\sum_{i=2}^{6} \beta_i \times \text{Pos} + \beta_7 \times \text{rookie} + \beta_8 \times \text{mid-career} +$$

$$\textstyle\sum_{i=9}^{13} \beta_i \times \text{Pos} \times \text{season} + \beta_{14} \times \text{rookie} \times \text{season} + \beta_{15} \times \text{mid-career} \times \text{season}$$

This model has a good interpretation and allows for us to model the mean efficiently. However, to make it more parsimonious, we tested the interaction terms and the main effects terms to see if they were necessary. We used a One-Way ANOVA at baseline to check group differences (Table 1) and concluded that there was a significant difference in DRAYMOND scores between positions in 2013-2014, but no significant difference between players with different career status.

main effect	**p value**
Position	0.00417
Career Status	0.57

Table 1: ANOVA for Group Differences in 2013-2014

Therefore, we decided to leave out the main effect for career status. When testing the interaction terms, we get a LRT test statistics of 2.94 testing the model with a career status and time interaction, and a LRT test statistic of 0.42 when testing the need for a position and time interaction. While neither are significant, we wanted to control for the affect of age and experience on change - so we decided to leave the career status and time interaction. We justified this by the fact that the test statistic was higher, as well as through a subjective interest in the interaction. Moving forward, we considered the following mean model:

$$\text{DRAYMOND} = \beta_0 + \beta_1 \times \text{season} + \sum_{i=2}^{5} \beta_i \times \text{Pos} + \beta_6 \times \text{rookie} \times \text{season} + \beta_7 \times \text{mid-career} \times \text{season}$$

To assess these models, we used linear regression (i.e. an independent covariance structure) as it was a mean model. Further along, we modify the covariance structure to appropriately model the variation between individuals (**2.3**, **2.4**).

2.2 Linear Mixed Model

Using the mean model as a population average, we fit a linear mixed model with a random intercept and a random slope.[4] A random slope was included to model the idea that some players are objectively better at defense than others, and that should be taken into account. While, on average, we considered players at the same career status to begin at the same point in the mean model, we know that there is most likely substantial variation in where that exact point is. Furthermore, some players respond to practice more than others, so we included a random slope to test for differences in change over time. That is, we would expect some players to improve at different rates than others. Using this linear model, we compare the model with a random intercept to our mean model, and then our model with a random slope and random intercept to the mixed model with only a random intercept. To do this, we used a 50-50 χ^2 mixture distribution to conduct the likelihood ratio test, as well as compared the models with respect to AIC and BIC (Table 2).

model	**LRT χ^2**	**AIC**	**BIC**
RANDOM INTERCEPT + SLOPE	6.27	3322.71	3381.063
RANDOM INTERCEPT	46.98	3324.98	3373.608
MEAN MODEL	–	3369.97	3413.730

Table 2: Comparison of Models

2.3 Generalized Estimating Equations(GEE)

GEE is specified by a mean model and a correlation model. We fitted GEE using marginal mean model, considering various specifications for the 'working' correlation structure(Independence, Exchangeable, Auto-regressive, Unstructured). We aimed to make inferences about the population, accounting for the within-subject correlation.

GEE is based on the quasilikelihood theory, and also it has no assumption with distribution of response observations. Therefore, AIC, a widely used method for model selection in GLM, is not applicable to GEE directly. Thus, we calculated QIC for selecting the working correlation structure.

$$\mu_{ij} = \beta_0 + \beta_1 \times \text{season}_{ij} + \sum_{i=2}^{5} \beta_i \times \text{Pos}_i + \beta_6 \times \text{rookie}_i \times \text{season}_{ij} + \beta_7 \times \text{mid-career}_i \times \text{season}_{ij}$$

2.4 Residual Analysis

An important question we would like to address is which of the defensive covariates are most influential while calculation of DRAYMOND scores. Does this vary between positions? The residuals from the above GEE model can be interpreted as the DRAYMOND score of all the players with the effects of 'experience' of each player removed. We first fit a population-level model using GEE to figure out the effect of these defensive covariates.

$$\begin{aligned}\text{Resid} = \beta_0 + \beta_1 \times \text{season} + \beta_2 \times \text{Pos} + \beta_3 \times \text{DRB} + \beta_4 \times \text{BLK} + \beta_5 \times \text{STL} + \beta_6 \times \text{Possession}\\ + \beta_7 \times \text{DRB} \times \text{Pos} + \beta_8 \times \text{BLK} \times \text{Pos} + \beta_9 \times \text{STL} \times \text{Pos} + \beta_{10} \times \text{Possession} \times \text{Pos}\end{aligned}$$

We would expect to find significant interactions between position and defensive covariates so we would split our data into five groups based on the five positions and fit five separate position-specific models using GEE. Now, we want to figure out which covariates are more important for each position as far as prediction of DRAYMOND score goes.

$$\text{Resid} = \beta_0 + \beta_1 \times \text{season} + \beta_2 \times \text{DRB} + \beta_3 \times \text{BLK} + \beta_4 \times \text{STL} + \beta_5 \times \text{Possession}$$

3 Results

3.1 Mean Model

Presented here are the results of the mean model presented in **2.1**. As seen in the Table 3, all covariates except for season × rookie, are statistically significant(p-value < 0.05), and regardless of career status, DRAYMOND scores for all players decrease as time goes by, shown by the time slope (reference for late-career players) and the interactions between the other two career statuses. From the data, it appears DRAYMOND scores tend to decrease more slowly on average among the players in their early career compared to those in their late career. Furthermore, it appears that late-career players are the ones getting worse at, on average, and other players, while getting worse, are mostly staying the same in terms of defensive performance. We also see a significant difference in terms of where players begin depending on the position they play - as we should given the preliminary analysis in **2.1**. Centers seem to play better defense overall, with shooting guards playing the worst defense - a result that subjectively makes sense.

covariates	**estimation**	**standard error**	**p value**
(Intercept)	0.60941	0.11830	**3.14e-07**
season	-0.13081	0.05212	**0.0122**
PosPF	-0.30414	0.14188	**0.0323**
PosPG	-0.76923	0.14197	**7.63e-08**
PosSF	-0.36310	0.15249	**0.0175**
PosSG	-0.86093	0.14049	**1.30e-09**
season:rookie	0.10933	0.06956	0.1163
season:midcareer	0.10489	0.05433	**0.0583**

Table 3: Coefficients estimation of Mean Model

3.2 Linear Mixed Model

3.2.1 Coefficients Results

First, we showed the necessity of random effects through both standard hypothesis testing methods such as the Likelihood Ratio Test (LRT) and through the lens of information criteria. Referring to the results in Table 2, we see that, when comparing the mean model with no random effects to the model with a random intercept, the LRT gives us a test statistic of 46.98 and the random intercept model is objectively better than the mean model with respect to both AIC and BIC. When comparing the random intercept model to the model that includes a random slope, the random intercept is better with respect to BIC but the random slope and random intercept model is better in terms of AIC. To make the decision, we consult the LRT and obtain a test statistic of 6.27 which, while not high, is greater than 5.14. Therefore, we conclude that the model that includes a random intercept and a random slope is the best fit for our data.

The fitted fixed effect coefficients are shown below in Table 5 [5] and the random effect variance matrix is shown in Table 4. We can see in Table 5 that our original mean model inference basically remains unchanged - centers play the best defense, and late career players are getting worse overall. However, it appears the interaction terms are no longer as significant as they originally were once we include the random effects. Looking at Table 4, we can see that only around 34% of the variation in DRAYMOND scores is explained by the variance attributed to between unit differences in this model. For this reason, we explore the unexplained variance (the residuals from this model) in **3.4**.

Groups	**Name**	**Variance**	**Std.Dev.**	**Corr**
player	(Intercept)	0.61306	0.7830	
	time	0.07937	0.2817	-0.39
Residual		1.34441	1.1595	

Table 4: Variance of Random Effects

covariates	**estimation**	**standard error**	**t value**
(INTERCEPT)	0.61706	0.14485	4.260
TIME	-0.11553	0.05416	-2.133
POSPF	-0.31286	0.18826	-1.662
POSPG	-0.77136	0.18805	-4.102
POSSF	-0.38177	0.20232	-1.887
POSSG	-0.87196	0.18643	-4.677
TIME:ROOKIE	0.07626	0.08160	0.935
TIME:MIDCAREER	0.08505	0.06383	1.332

Table 5: Coefficients estimation of Linear Mixed Model

3.2.2 Model Diagnostics

The Residuals versus Fits plot is used to detect non-linearity, unequal error variances, and outliers. As for our LMM model, first we applied Cholesky decomposition on the residuals, and plot them versus their corresponding fitted values.

The other 3 plots listed are the measurements of influence. Cook's Distance directly summarizes how much all of the fitted fixed effects change when the ith observation is deleted. A data point with a large Cook's distance indicates that the data point strongly influences the fitted values. The Covariance Ratio (covratio) measures the change in the covariance matrix of the fixed effects based on the deletion of each player. By comparing the ratio of the determinants, this measurement quantifies the influence on precision of fixed effects. The bottom right plot shows the relative variance change to assess the influence of each player on the variance components.

We can see a set of influential points are somewhat consistent from all three perspectives. Further work may include removing the influential points and refitting the LMM again.

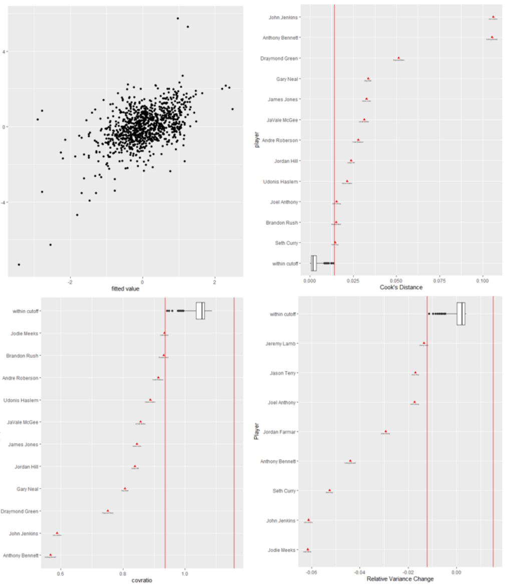

Figure 2: DRAYMOND Score by Position and Time

3.2.3 Prediction

The predictors we used in the linear mixed model are all well defined for future seasons. Using BLUP estimates of the random effects and REML estimates of the fixed effects using the lme4 [5] package in R, we can get estimates for future DRAYMOND scores. We did just that. Presented in Table 7 are the Top 10 defensive players of the 2017 - 2018 season, as predicted by our model.

Player	DRAYMOND Score
Draymond Green	2.623
Rudy Gobert	1.804
Dewayne Dedmon	1.635
Andre Roberson	1.568
Aron Baynes	1.373
Timofey Mozgov	1.349
Jodie Meeks	1.346
Miles Plumlee	1.313
James Johnson	1.264
Michael Kidd-Gilchrist	1.217

Table 6: DRAYMOND Scores for 2017-2018 season

3.3 GEE

We fit our mean model using GEE to understand the population level effects that we saw in the mean model, while accounting for different possible covariance structures. Parameter estimates obtained are consistent irrespective of the underlying true correlation structure, but may be inefficient when there is a misspecification of the correlation structure. However, GEE is fairly robust against a misspecified correlation structure (particularly with large sample size), so we expect to see consistent estimates. We can see this is true in Table 8, where parameters are relatively equal despite different covariance structures.

As you can see in the Table 9, we also computed QIC for model selection. Here, the exchangeable correlation model has the smallest QIC value and thus is the preferred correlation structure. However, the differences of QIC between the models is subtle.

corr	$\hat{\beta}_{time}$	$\hat{\beta}_{PF}$	$\hat{\beta}_{PG}$	$\hat{\beta}_{SF}$	$\hat{\beta}_{SG}$	$\hat{\beta}_{t \text{ x } rookie}$	$\hat{\beta}_{t \text{ x } midcareer}$
IN	-0.131(0.052)	-0.304(0.142)	-0.769(0.142)	-0.363(0.152)	-0.861(0.140)	0.109(0.070)	0.105(0.054)
UN	-0.111(0.051)	-0.301(0.187)	-0.777(0.187)	-0.356(0.201)	-0.865(0.185)	0.071(0.076)	0.083(0.060)
EX	-0.113(0.050)	-0.308(0.187)	-0.777(0.188)	-0.368(0.201)	-0.864(0.185)	0.070(0.076)	0.081(0.059)
AR-1	-0.116(0.056)	-0.294(0.172)	-0.770(0.172)	-0.341(0.185)	-0.876(0.171)	0.083(0.079)	0.091(0.062)

Table 7: GEE Parameter Estimates with Sandwich Variance Estimates

Corr	QIC
EX	973.3
UN	973.5
AR	974.3
IN	974.6

Table 8: QIC for the working correlation

3.4 Residual Analysis

From the analysis of the residuals at the population level, we found significant interactions between position and all four defensive covariates considered. The main effects of blocks and number of possessions also came out to be significant. The results can be found in Table 10.

After splitting the data based on the five positions, the analysis of the residuals gave us Table 11-15 (in Appendix). In all of these tables, the coefficients in bold are the ones which are significant. We can see that for shooting guards, blocks and steals are important; for centers, blocks are the driving factor; for power forwards, defensive rebounds and blocks seem to be most influential; for small forwards, steals count more; and for point guards, blocks and steals get you further as far as DRAYMOND score is concerned.

Covariates	Df	χ^2	pvalue
season	1	0.00	0.9554
Pos	4	0.14	0.9976
drb	1	1.03	0.3100
blk	1	25.38	**4.7e-07**
p	1	10.63	**0.0011**
stl	1	0.06	0.8102
Pos:drb	4	14.97	**0.0048**
Pos:blk	4	15.67	**0.0035**
Pos:p	4	13.77	**0.0081**
Pos:stl	4	14.77	**0.0052**

Table 9: Analysis of Wald Statistics

4 Discussion

The biggest conclusion that we could draw from our analysis is that the NBA, as a whole, is getting worse defensively - with late career players getting worse at a faster rate in general. Even though this may break the hearts of those basketball fans who love watching low-scoring games with solid defensive performances, it is not as ominous as it sounds. This only means that the offensive players are finding better ways to successfully evade the defenders and take their shots or in the worst case, successfully draw fouls on themselves. In recent times, we have seen how the game has evolved into a quick and intricate passing game (the Warriors' triumphs in 2014-15 and 2015-16 is the perfect example) with lots of stress on 3-point shooting. If teams can shoot $\sim 40\%$ successfully from 3-point range (Miami Heat has a 3-point success rate of 38.2% and they are currently second in the Eastern Conference), there is only so much a defender can do. If players like Stephen Curry can shoot from 40 feet away and still make a basket, your defensive stats (and your morale if you're on the opposing team) are bound to take a huge hit. From our analysis, this shift in playing style becomes pretty evident. It also begs the question - how can a player play better defense?

To look into this question, we have seen that different covariates are important for improving a defender's DRAYMOND score for different positions. For example, for a shooting guard, blocks and steals turned out to be statistically significant. This is a nice and interpretable result as a shooting guard is not expected to be near the board and get defensive rebounds. Neither will he have a lot of possession. His job is to try and block 3-pointers, get the occasional steal if he can and kill the other team on transition.

Finally, it would appear from our analysis that, in fact, when considering our mixed model, Draymond Green is the most complete defensive player in the league. He consistently has the highest DRAYMOND score and our predictions for the next season also supports our claim. Of course, we don't need to perform an analysis to arrive at this conclusion because - even if you know little about the NBA - you know who Draymond Green is!

References

[1] Marti Casals and Jose Martinez. Modelling player performance in basketball through mixed models. *International Journal of Performance Analysis in Sport*, page 64–82, 2013.

[2] Jamie Sampaio, Eric Drinkwater, and Nuno Leite. Effects of season period, team quality, and playing time on basketball players' game-related statistics. *European Journal of Sport Science*, 10:141–149, 2010.

[3] Nate Silver. A better way to evaluate nba defense. https://fivethirtyeight.com/features/a-better-way-to-evaluate-nba-defense/, note = Accessed: 2019-11-27.

[4] Geert Verbeke and Geert Molenberghs. *Linear mixed models for longitudinal data*. Springer Science & Business Media, 2009.

[5] Ben Bolker ORCID iD [aut cre] Steven Walker ORCID iD [aut] Rune Haubo Bojesen Christensen ORCID iD [ctb] Henrik Singmann ORCID iD [ctb] Bin Dai [ctb] Fabian Scheipl ORCID iD [ctb] Gabor Grothendieck [ctb] Peter Green ORCID iD [ctb] John Fox [ctb] Douglas Bates ORCID iD [aut], Martin Maechler ORCID iD [aut]. lme4: Linear mixed-effects models using 'eigen' and s4.

5 Appendix

Covariates	Coef
(Intercept)	-88.051
season	0.043
DRB	0.008
BLK	**0.267**
STL	**0.158**
Possession	**8.983x10^{-5}**

Table 10: Shooting Guard

Covariates	Coef
(Intercept)	11.032
season	-0.006
DRB	-0.004
BLK	**0.117**
STL	-0.076
Possession	**1.15x10^{-4}**

Table 11: Center

Covariates	Coef
(Intercept)	-112.313
season	0.056
DRB	**-0.029**
BLK	**0.188**
STL	-0.045
Possession	**1.13 x 10^{-4}**

Table 12: Power Forward

Covariates	Coef
(Intercept)	-126.310
season	0.062
DRB	0.043
BLK	0.010
STL	**0.182**
Possession	-6.75 $\times 10^{-5}$

Table 13: Small Forward

Covariates	Coef
(Intercept)	168.448
season	-0.084
DRB	-0.016
BLK	**0.498**
STL	**-0.147**
Possession	1.17 $\times 10^{-5}$

Table 14: Point Guard

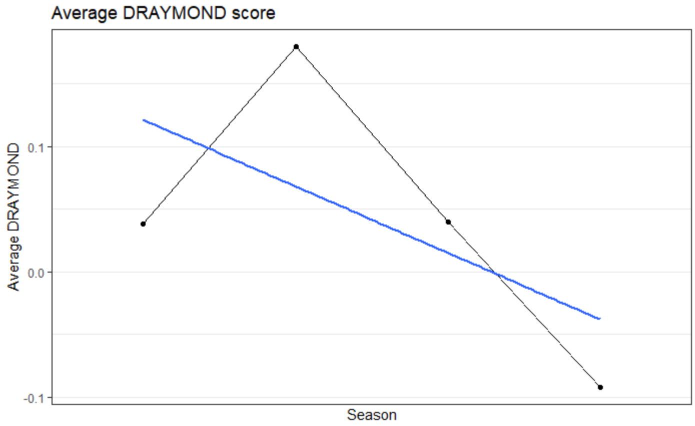

Figure 3: Average DRAYMOND Score by Time

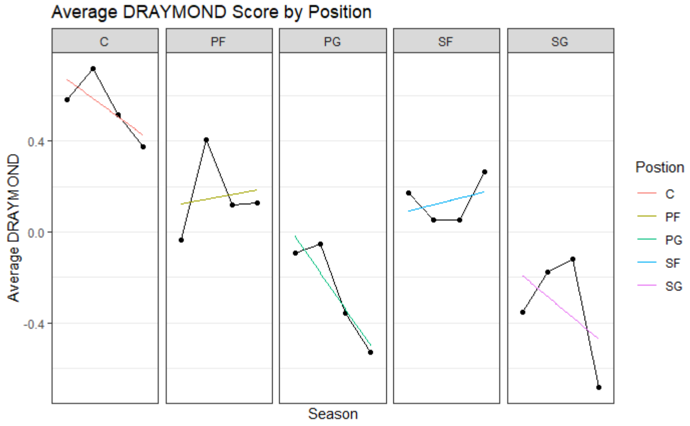

Figure 4: Average DRAYMOND Score stratified by Pos

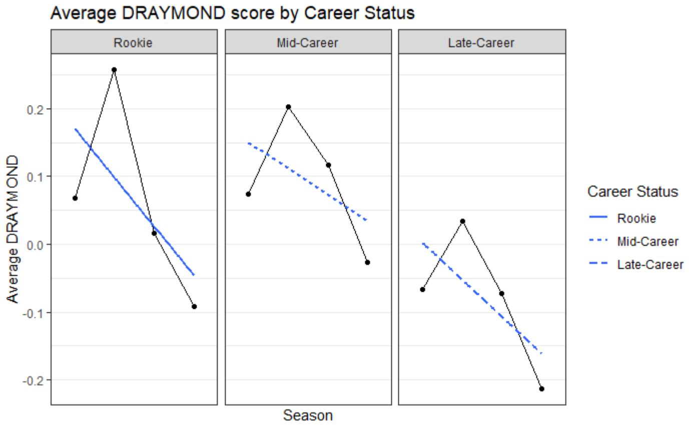

Figure 5: Average DRAYMOND Score by Career Status