Statistical Inference from NBA Players' Points per Game Regression Model

Chris Doan and Gabe Wallon

April 30, 2023

Introduction

The NBA is the most competitive professional basketball league in the world, and in the game of basketball, scoring points is often considered the most valuable skill to have. In this paper we will build a multiple regression model which predicts the points per game (PPG) statistic for a player based on their other statistics (e.g. assists per game (APG), free throw percentage (FT%) etc.).

We will be using data provided by the official statistics provider for the NBA, SportRadar. Our data comprises information about players who played at least 58 games per season, during the 2020, 2021, and 2022 seasons. In an attempt to avoid a violation in the independence of our observations, we have taken a random sample of the original data. The original data poses an issue with proximity in space of the observations, since a player from one season to the next is likely to have similar statistics.

Our primary goal is to use this model for inferential statistics, to gain insights into how different basketball skills correlate with one another, and how they contribute to predicting PPG. We interpret the slope estimates and p-values of our predictor variables to determine how important of a role they play in predicting the PPG of a player. These slope estimates will give us a fuller picture of the skill palette of high scoring players. We also perform a more thorough analysis on our model's most significant predictor.

If we make the simplifying assumption that a player's value is proportional to their PPG statistic, in finding the variables which are most important in predicting a player's PPG, we implicitly learn which other characteristics are most valuable for a player to have. We ultimately hope to gain an overall deeper understanding of the game.

Exploratory Analysis

Response:

The response variable for this model is Points per Game (PPG), so we would first like to explore and visualize the variable itself to better understand how it is distributed.

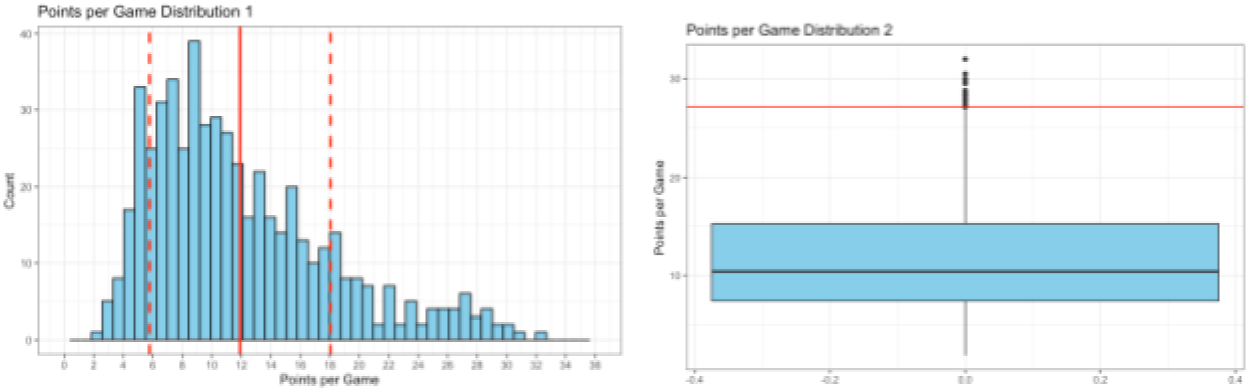


Figure 1: Histogram and Boxplot of Points per Game

The histogram in Figure 1 above shows the distribution of PPG and displays the mean (solid line) and standard deviation (dashed lines) of the variable at 11.9 PPG and 6.1 PPG respectively. The solid line in the above boxplot shows the cutoff for which observations are considered outliers. In Figure 2 below, you can see a sampling distribution of the PPG variable and the associated 90% bootstrapped confidence interval of [11.47, 12.34]. Hence, we are 90% confident that the true mean PPG value for the population is between 11.47 and 12.34.

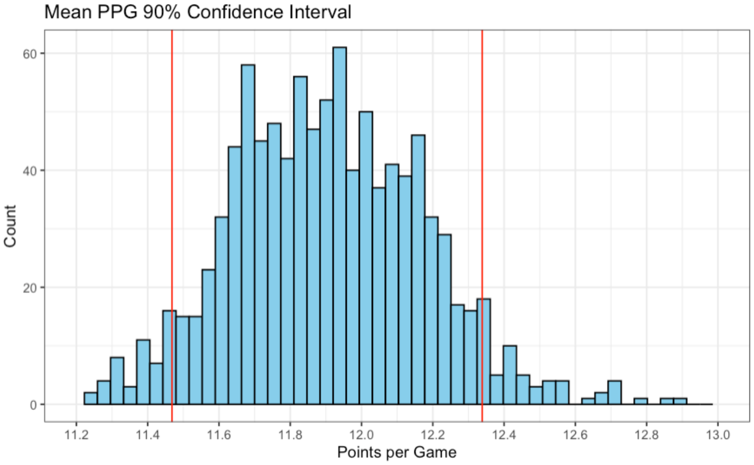


Figure 2: 90% confidence interval of mean PPG

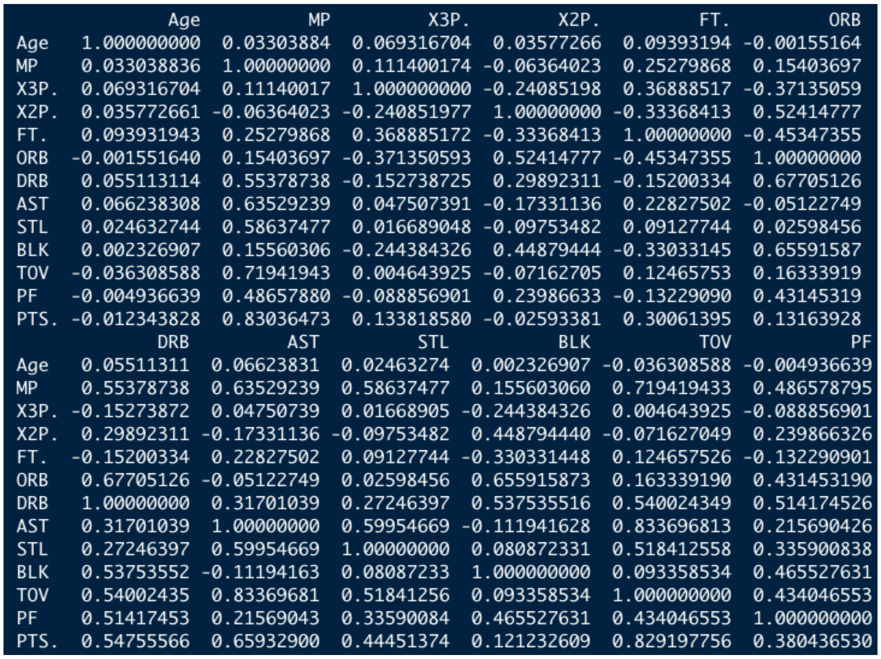
Predictors:

Some of the variables in the original data set we choose to exclude from our model. A major reason for this is obvious collinearity. Variable pairs such as 2-pointers made per game, along with 2-pointers attempted and 2-pointer percentage essentially give us the same information about the player, thus including two of them in the model would be redundant. Collinearity would also become an issue in our model if we were to include the player position variable. A player's position can tell you a lot about which skills they have. For example we would expect Centers to excel in rebounding and Point Guards to excel in assisting. Additionally, we performed an ANOVA test to determine if PPG varied significantly between any of these groups, from which we concluded that only 3 of 10 pairs of positions had a significant difference in average PPG. This leads us to believe that PPG is not strongly influenced by position, which makes sense in the context of the sport. The details of this test can be found in the appendix.

Some variables we will exclude from the model, because including them would overshadow the effect of other variables. For example, including field goals made per game in a model would take away the significance of any other predictors as knowledge of this variable would be vastly more important in making PPG predictions.

With that being said, the predictor variables for this model include the rest of the "basic" box score statistics, which are essentially just any countable statistic in basketball. All of the predictor variables other than Age are considered on a "per game" basis. The predictors that will be considered for the model are as follows:

- Age
- Minutes played (MP)
- 3 Point percentage (3P%)
- 2 Point percentage (2P%)
- Throw percentage (FT%)
- Offensive Rebounds (ORB)
- Defensive Rebounds (DRB)
- Assists (APG)
- Steals (SPG)
- Blocks (BPG)
- Turnovers (TOV)
- Personal Fouls (PF)



```
              Age          MP         X3P.        X2P.         FT.         ORB
Age   1.000000000  0.03303884  0.069316704  0.03577266  0.09393194 -0.00155164
MP    0.033038836  1.00000000  0.111400174 -0.06364023  0.25279868  0.15403697
X3P.  0.069316704  0.11140017  1.000000000 -0.24085198  0.36888517 -0.37135059
X2P.  0.035772661 -0.06364023 -0.240851977  1.00000000 -0.33368413  0.52414777
FT.   0.093931943  0.25279868  0.368885172 -0.33368413  1.00000000 -0.45347355
ORB  -0.001551640  0.15403697 -0.371350593  0.52414777 -0.45347355  1.00000000
DRB   0.055113114  0.55378738 -0.152738725  0.29892311 -0.15200334  0.67705126
AST   0.066238308  0.63529239  0.047507391 -0.17331136  0.22827502 -0.05122749
STL   0.024632744  0.58637477  0.016689048 -0.09753482  0.09127744  0.02598456
BLK   0.002326907  0.15560306 -0.244384326  0.44879444 -0.33033145  0.65591587
TOV  -0.036308588  0.71941943  0.004643925 -0.07162705  0.12465753  0.16333919
PF   -0.004936639  0.48657880 -0.088856901  0.23986633 -0.13229090  0.43145319
PTS. -0.012343828  0.83036473  0.133818580 -0.02593381  0.30061395  0.13163928
             DRB         AST         STL          BLK          TOV           PF
Age   0.05511311  0.06623831  0.02463274  0.002326907 -0.036308588 -0.004936639
MP    0.55378738  0.63529239  0.58637477  0.155603060  0.719419433  0.486578795
X3P. -0.15273872  0.04750739  0.01668905 -0.244384326  0.004643925 -0.088856901
X2P.  0.29892311 -0.17331136 -0.09753482  0.448794440 -0.071627049  0.239866326
FT.  -0.15200334  0.22827502  0.09127744 -0.330331448  0.124657526 -0.132290901
ORB   0.67705126 -0.05122749  0.02598456  0.655915873  0.163339190  0.431453190
DRB   1.00000000  0.31701039  0.27246397  0.537535516  0.540024349  0.514174526
AST   0.31701039  1.00000000  0.59954669 -0.111941628  0.833696813  0.215690426
STL   0.27246397  0.59954669  1.00000000  0.080872331  0.518412558  0.335900838
BLK   0.53753552 -0.11194163  0.08087233  1.000000000  0.093358534  0.465527631
TOV   0.54002435  0.83369681  0.51841256  0.093358534  1.000000000  0.434046553
PF    0.51417453  0.21569043  0.33590084  0.465527631  0.434046553  1.000000000
PTS.  0.54755566  0.65932900  0.44451374  0.121232609  0.829197756  0.380436530
```

Figure 3: Predictor variable correlation matrix

We would like to check for collinearity between these predictors, as multicollinearity would cause fundamental problems in creating an accurate model, and especially since we are primarily interested in inference. A check for collinearity can be done by creating a correlation matrix of all the predictor variables, and looking for any high values of the correlation coefficients. We'll define a value as a "problem" if it has a strong correlation, so a value of 0.7 or greater. By running and examining a correlation matrix as seen in Figure 3, we find that the only predictors with a strong collinearity is between Assists and Turnovers, with a value of 0.834, as well as between Minutes Played and Turnovers, with a value of 0.719. We can also visualize these correlations by plotting the variables against each other:

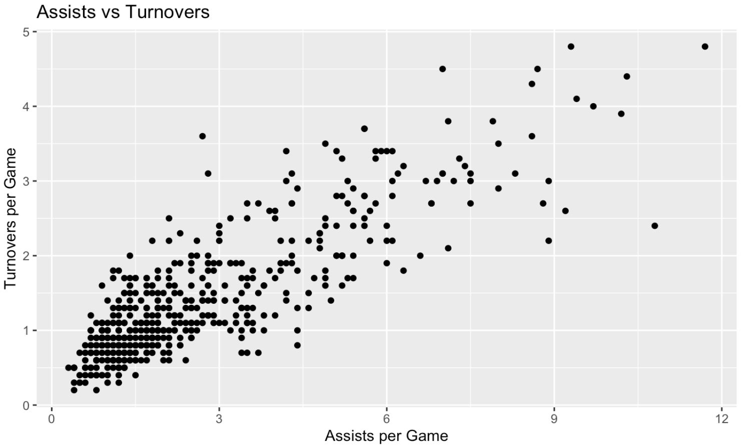


Figure 4: Assists per game plotted against Turnovers per game

As we can see from Figure 4, the plot of APG and TOV has a strong, positive, and linear relationship, which is what is suggested and expected from the correlation coefficient of 0.847. This correlation makes intuitive sense as well, since a high APG statistic implies that this player makes a lot of passes. The more passes you make, the more likely those passes will get stolen by the defense or are thrown away, thus causing more turnovers. The same plot can be made for MP and TOV, and we would find similar results. To solve this issue of multicollinearity we discovered, we will simply remove the TOV variable from consideration as a predictor, since it is the only variable that causes the issue. This leaves us with 11 predictor variables to consider for the model. We will do a further/double check for multicollinearity by computing a variance inflation factor (VIF) among all the predictors once we have our final multiple regression model for PPG.

Model Development

Now that we can assume all the predictor variables are independent of each other after our check for collinearity, we must decide which predictors we actually want to include in our multiple regression model. We will be doing this through a backwards stepwise selection method. This method starts by creating a full regression model (one that includes all the potential predictors) and eliminates one variable at a time from the model until we can't improve it any further. The criterion used to decide which variable to remove at each step is the adjusted R^2 (which describes the strength of a model fit); we will seek to eliminate a variable that leads to the largest improvement in adjusted R^2, and do so until the elimination of a variable doesn't lead to an improvement in the metric.

After completing the backwards stepwise selection process, we find that only the removal of one predictor, Blocks (BPG), leads to the best model. The adjusted R^2 of the model with all the predictors is 0.77309, and removing the Blocks variable leads us to the biggest improvement, which gives us an adjusted R^2 value of 0.77351. The removal of any other variable after this would only cause a decrease in adjusted R^2, so our model selection stops there and we'll go on to build a linear model with every predictor except TOV (which was removed because of multicollinearity) and BPG (which was removed through the backward stepwise selection). It is interesting that to get the best model, only one variable needed to be taken out. This tells us that every other variable is actually meaningful, and each of them can give us insight on and affect PPG.

We create the multiple regression model using statistical software, which gives us slope estimates for each predictor variable, as well as their p-values (among other things). The hypothesis test that corresponds to each of the predictors individually and their p-values is as follows:

$$H_0: \beta_i = 0$$

$$H_A: \beta_i \neq 0$$

where *i* is the *i*-th predictor variable in the model, and β is its slope parameter. The null states that the slope of the *i*-th predictor is 0, or in other words, that predictor has no effect on a player's PPG. The alternate hypothesis states that the slope of the *i*-th predictor has a slope that is not zero, or in other words, that predictor has a statistically significant effect on a player's PPG. Even after our model selection process, we find that not every predictor is statistically significant, and can see the results of each variable's hypothesis test using a 90% significance level (α=0.1):

Predictor Variable	P-value	Result
Age	0.000108	Reject the null
Minutes Played	< 2e-16	Reject the null
3 Point Percentage	0.059229	Reject the null
2 Point Percentage	0.003182	Reject the null
Free Throw Percentage	3.11e-06	Reject the null
Offensive Rebounds	0.283328	Fail to reject the null
Defensive Rebounds	3.69e-08	Reject the null
Assists	< 2e-16	Reject the null
Steals	2.57e-07	Reject the null
Personal Fouls	0.214765	Fail to reject the null

Table 1: P-values for all predictors and their hypothesis test result at 90% significance level

Since we fail to reject the null hypothesis for Offensive Rebounds and Personal Fouls, we can say that they have no effect on a player's PPG when included in a model with the other predictors. So, we can let their slope parameters be zero.

Since ORB and PF are not statistically significant, we can exclude them from the model, leaving us with 8 predictor variables for our final model which can be given by:

$$PPG = -11.988 - 0.117(Age) + 0.577(MP) + 2.641(3P\%) + 6.915(2P\%) + 7.782(FT\%) + 0.688(DRB) + 0.875(APG) - 2.504(SPG)$$

As mentioned earlier, we can do another check for any multicollinearity among the predictors by computing a variance inflation factor (VIF), which measures the total increase in standard error across all predictors when another potentially collinear predictor is included in the model. A large value of VIF is an indication of multicollinearity; We will use the common threshold VIF value of 5 to determine if it is "too large". The predictors and their VIF values can be seen in the table:

Variable:	Age	MP	3P%	2P%	FT%	ORB	DRB	APG	SPG	PF
VIF:	1.026	3.072	1.246	1.475	1.580	3.019	2.986	2.106	1.849	1.603

Table 2: VIF values for all predictors in the model

Since none of VIF's are larger than 5, we can safely assume that multicollinearity is not an issue.

An example of how the slope estimates in the model can be interpreted can be done for MP: for every unit increase in minutes played, a player will score 0.577 more points on average when all other variables are held constant. This makes intuitive sense; the more a player is on the floor and actually playing, the more likely they will score compared to someone who plays less. This has a large impact on PPG because MP will most likely always be the largest value in a box score (a list of player's statistics for a game) other than points itself; this means that the MP variable will also be the largest number in the model equation and thus affect PPG the most. Another example for one of the shooting percentages: for every one unit increase in free throw percentage, a player will score 0.078 more points on average with all other variables held constant. An interpretation of one of the negative slopes (SPG): for every additional steal a player gets, they are expected to score 2.504 less points on average, when all other variables are held constant. The explanation for this relationship, as well as some others will be discussed later on.

For our model to be valid, and in order for inferences made on the slope parameters to be valid (such as the hypothesis tests done mathematically), we must make sure our data meets the four conditions of linear regression: Linearity, independence, normality, and equal variance. We can check these assumptions using diagnostic plots regarding residuals of the model. To check for linearity, we can create a Residuals vs Fitted plot, and look for any pattern (which would indicate a violation in the linearity assumption).

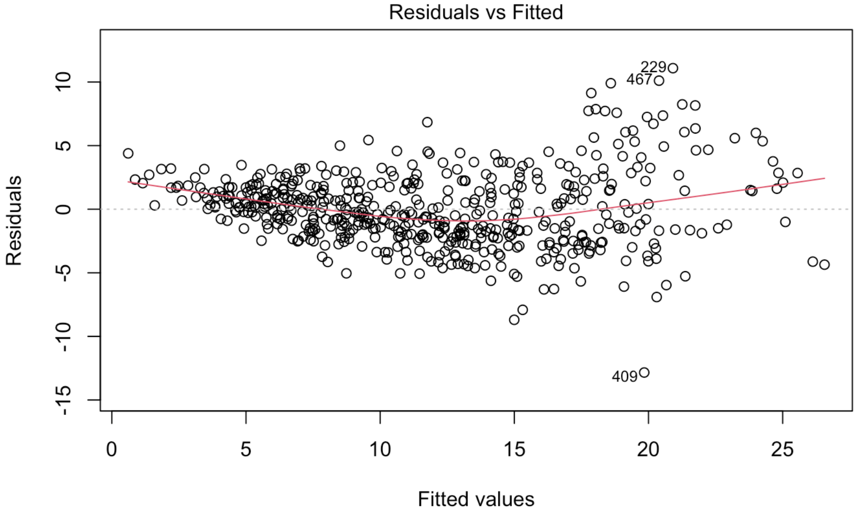


Figure 5: Residuals vs Fitted diagnostic plot

From Figure 5, one could argue that the residuals might be slightly parabolic, but not enough to say that there is an "obvious" nonlinear pattern in the residuals. For the most part, they vary randomly above and below the average value of zero, however we do notice a "cone" shape (this is concerning if we want to assume constant variance of the residuals however we will later see that according to another test, the constant variance assumption holds). Since there is no nonlinear pattern in the residuals, we can assume that the linearity condition holds for the model. The next condition to be checked is independence. We can check this by visualizing the residuals as a function of the individual predictors, and look for any serial correlation, which would mean lack of independence. For example, we can create a scatter plot of the residuals as a function of 2P%:

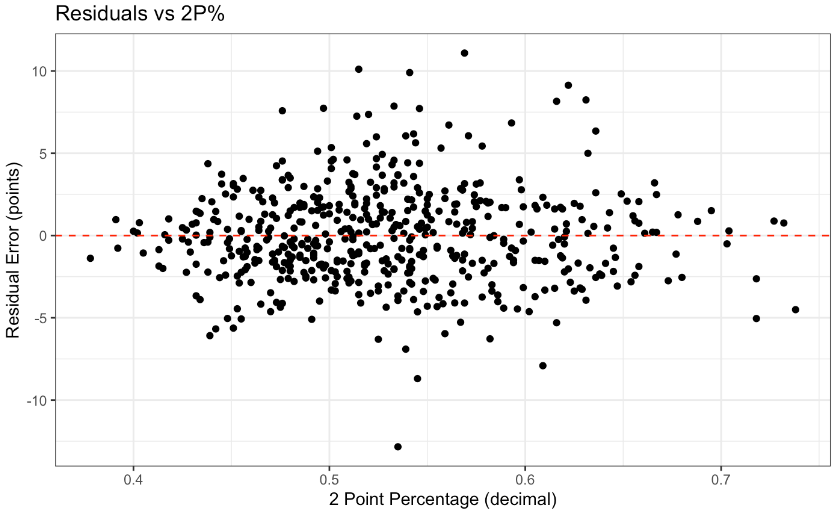


Figure 6: Scatter plot of residuals vs 2P%

Clearly, there is no serial correlation in Figure 6, and no pattern in consecutive residuals (thus does not violate independence). The same is true for every other predictor variable in the model, confirming that the independence assumption holds. The third assumption of linear regression is that the residuals are normally distributed, with a mean of zero. A method to assess normality of residuals is to create a Normal Q-Q plot:

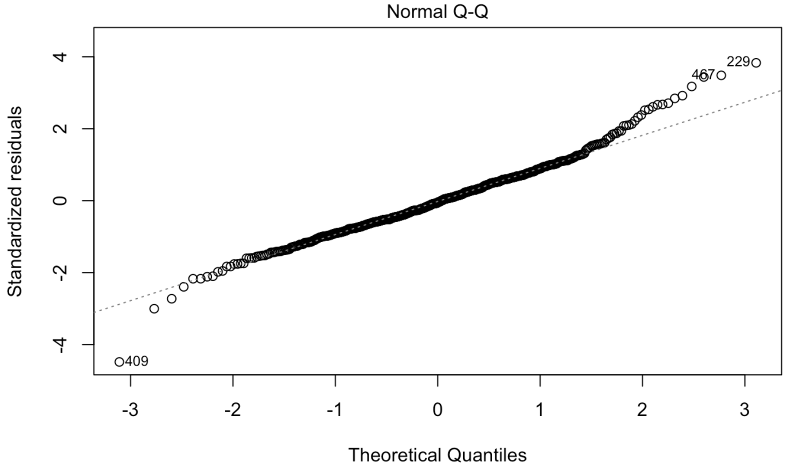


Figure 7: Normal Q-Q plot of model residuals

The Normal Q-Q plot in Figure 7 for the residuals follows the dashed line very well, meaning the residuals are almost perfectly normally distributed, except for the right tail of the distribution. Those observed residuals stray above the line, which indicate larger values than expected. However, since the dataset is so large (534 observations), there are not enough values that stray off the line to put the normality assumption in jeopardy, but it is worth noting the residual errors are slightly right-skewed. Additionally, if we calculate the mean of the residuals, we get 2.1649e-16, which is essentially zero. Lastly, we must see if the residuals all have equal variance, which can be checked using the Breusch-Pagan Test- a hypothesis test whose null is constant variance and alternate hypothesis is non-constant variance. Running this test on our regression model using statistical software gives us a p-value of 0.8022 for the test, so we fail to reject the null and say that the residuals have equal variance. Since all four of the technical conditions for linear regression appear to be satisfied, we can be confident in any inferential analysis we have done and will do with the model.

Model Analysis

Now that we have a model to predict PPG, we can conduct inferential analyses on predictor variables. In Table 1 above, we can see that minutes played (MP) and assists per game (APG) both have equally small p-values. This tells us that these two variables are of the most importance when trying to predict a player's PPG statistic, and thus has the most impact on scoring compared to the other predictors in the model. We choose to focus on APG however because this seems to be a more insightful and unexpected finding. Anyone could have guessed that the more minutes a player gets, the more points they will score. But it is not as obvious that a player who is skilled at making assists will also be skilled at scoring the ball.

Since our data satisfies the technical conditions necessary to assume the sampling distribution of the slope follows a t-distribution, we can construct a confidence interval for the true value of the APG slope parameter in the following way:

$$\beta_i \in \gamma_i \pm t_{n-(k+1),\alpha/2} \cdot SE_{\gamma_i}$$

where β_i is the true slope parameter, γ_i is our slope estimate, $t_{n-(k+1),\alpha/2}$ represents the critical value on the t-distribution with $n - (k + 1)$ degrees of freedom (n is the sample size, k is the number of model features) associated to a significance level of α, and SE_{γ_i} is the standard error associated with the slope sampling distribution. Using this formula, we found the following 90% confidence interval for the true slope parameter of APG: [0.727, 1.023]. Therefore, we can be 90% confident that as a player makes one more assist in a game, the number of points they will score in that game will increase by between 0.727 and 1.023 on average, while keeping the other variables constant.

There are other interesting things we can learn from our model. According to p-value, it appears that the most important shooting accuracy metric for predicting a player's PPG is their free throw percentage. This reveals the importance of free throws in the game of basketball: knowledge of a player's FT% can tell you more about their PPG statistic than their 3 or 2-point accuracy. We think that this variable is so significant because a free throw seems like a good indicator of generally how good a player is at shooting the basketball. It also removes confounding variable effects such as the quality of shot players are getting, since all players shoot a free throw from the same distance without any defense being played on them. Another interesting thing we can observe is the sign of the steals per game (SPG) coefficient. A negative coefficient as this suggests that we can expect players with high SPG statistics to have a lower PPG on average. This leads one to believe that as a player's focus on developing defensive skills increases, skills which likely facilitate a high SPG statistic, their focus on point scoring suffers. This is an interesting result because it reveals how a player's value should not be directly proportional to their PPG statistic, as defensive skills are very important, and steals can lead to momentum swinging plays crucial to a team victory.

Conclusion

We find that after removing variables due to multicollinearity, as well as during the model selection process, that the best model we could create to predict a player's PPG is a multiple linear regression with 8 predictors. Furthermore, the main goal of this regression was to do inference on the slope parameters, and after building the model we have successfully developed a better understanding of how specific basketball statistics impact a player's scoring. We were able to draw insights out of the data which revealed some fascinating things about the game of basketball. Some were not very surprising, but more reinforcing of our intuition, such as the finding that better shooting percentages lead to higher scoring output. Perhaps a more interesting, less expected, result is that both defensive rebounds and assists also have a statistically significant positive relationship with PPG. We also were surprised by the negative associations with PPG: a player's age and steals per game. Overall, it has proved useful to understand the importance of different player characteristics in predicting PPG.

Some ways we could improve this model in the future is by creating confidence intervals for every predictor variable, which would give us a stronger interpretation of the effect each variable has on PPG. Future work could also include reproducing this analysis using data from different seasons, and comparing the models we make. Using data from seasons in the 2010's could reveal interesting insights into how the game is evolving over time.

Appendix

Position Variable ANOVA Test:

First, let's visualize the relationship between a player's position and their PPG (Figure 8).

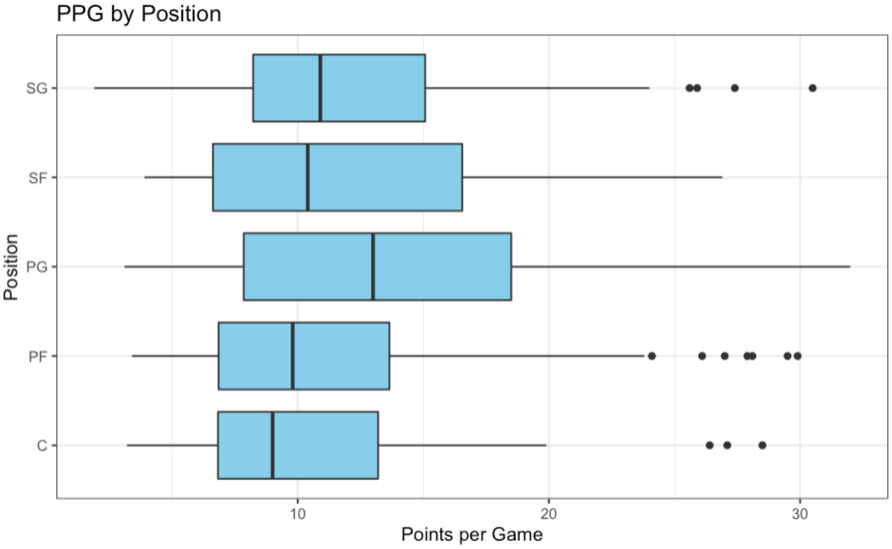
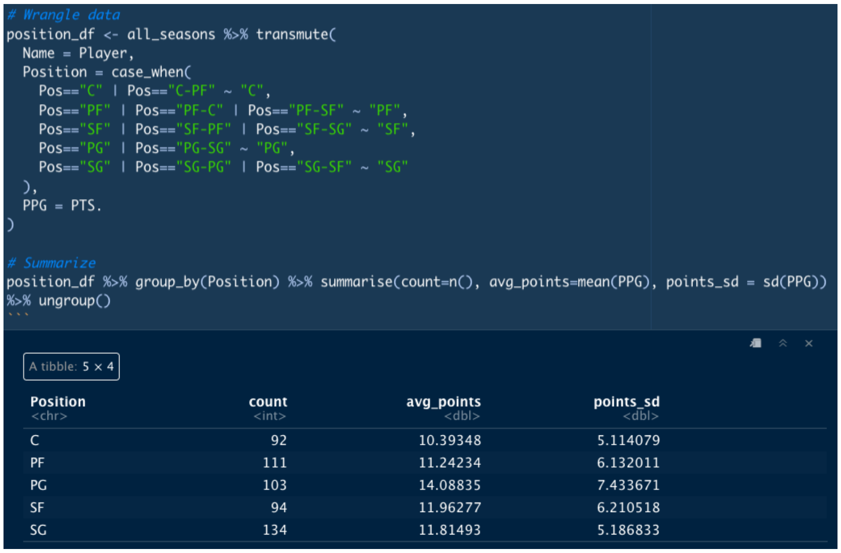


Figure 8: Box plots of PPG sorted by position

Assumptions of an ANOVA Test:

1. Observations are independent both between and within groups (populations):
The PPG statistic of two players who are in the same position will not be dependent on each other unless they are on the same team and one gets more minutes than the other. This should not be pervasive enough to violate the independence assumption of the whole group. Also, the PPG statistic of two players who are in different positions should not depend on each other.
2. Response values are roughly Normally distributed within groups:
Our data does not completely meet this condition, as we know, PPG is generally skewed to the right. However this is not very concerning to us because our sample sizes are not small.
3. Response values have roughly equal variance between groups:

We can see in the summary below (Figure 9) that our groups have roughly equal standard deviations of the PPG statistic.



```
# Wrangle data
position_df <- all_seasons %>% transmute(
  Name = Player,
  Position = case_when(
    Pos=="C" | Pos=="C-PF" ~ "C",
    Pos=="PF" | Pos=="PF-C" | Pos=="PF-SF" ~ "PF",
    Pos=="SF" | Pos=="SF-PF" | Pos=="SF-SG" ~ "SF",
    Pos=="PG" | Pos=="PG-SG" ~ "PG",
    Pos=="SG" | Pos=="SG-PG" | Pos=="SG-SF" ~ "SG"
  ),
  PPG = PTS.
)

# Summarize
position_df %>% group_by(Position) %>% summarise(count=n(), avg_points=mean(PPG), points_sd = sd(PPG))
%>% ungroup()
```
```

```
A tibble: 5 × 4

Position       count     avg_points     points_sd
<chr>          <int>          <dbl>         <dbl>
C                 92       10.39348      5.114079
PF               111       11.24234      6.132011
PG               103       14.08835      7.433671
SF                94       11.96277      6.210518
SG               134       11.81493      5.186833
```

Figure 9: Mean points and standard deviations sorted by positions

With the assumptions fulfilled, we set our level of significance to be 0.05, and complete an ANOVA test. As you can see in the plot to the right (Figure 10), our f-statistic (solid line) was larger than the rejection threshold value (dashed line). So we conclude that at least one of the positions has a statistically significant difference in PPG than the rest. To see which pairs of positions have statistically significant differences in average PPG, we can use a Tukey Honestly Significant Differences (HSD) test.

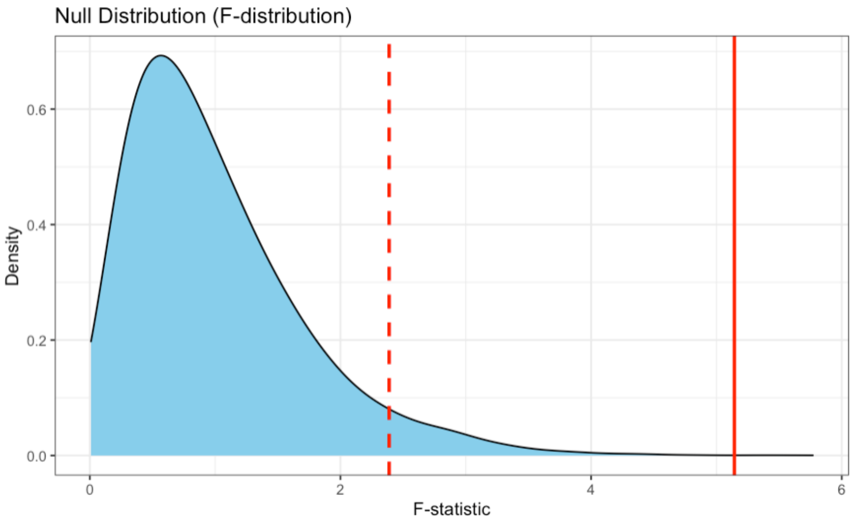


Figure 10: F-distribution with test statistic plotted as solid red line
```

In Figure 11 below, we can visualize the pairwise relationships between positions. We can see that there are only a few pairings that have significantly different average PPG statistics, namely point guard (PG) and center (C), point guard and power forward (PF), and surprisingly shooting guard (SG) and point guard. The interesting takeaway from this test is that point guards, often the smallest players on the court, average a higher PPG statistic than other positions.

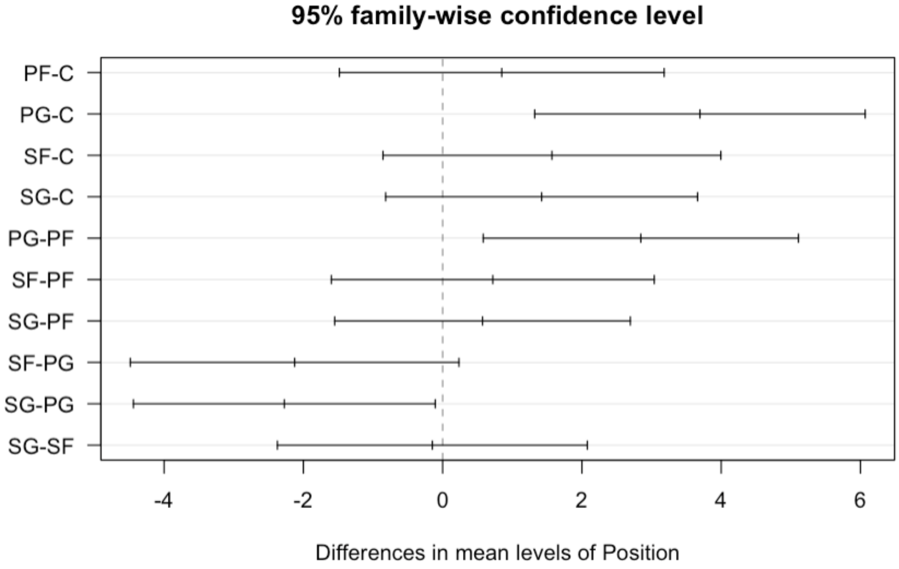


Figure 11: Pairwise relationships between positions regarding mean PPG