Can you Model the Change? An Analysis into Modeling NBA Scoring By: Pierre Aucoin

2024-04-9

Throughout the last few years, I have observed young players under 25 in the National Basketball Association (NBA) scoring an extreme amount of points in games. This scoring brought up an interesting question. Are current young NBA players scoring more than young players have in the past? Certainly, there is ample data to show that the number of points scored in NBA games has increased in recent years, but is that because of the young players scoring more, or is it because of the veterans (CleaningtheGlass)? When considering young players, there is another question that I have always been curious about: Does draft order really matter in their scoring productivity? Of course, I believe that in a perfect world, a player who gets drafted with the 1st pick should be the best player in the draft. However, this is not always the case. Throughout my lifetime, I have followed many top NBA draft picks that did not work out and performed poorly while also seeing many later picks that have outperformed players picked before them, being key pieces for championship-winning teams and getting enshrined into the Basketball Hall of Fame. I could not find any studies that model this information, so I decided to try this myself. Is it possible to model how draft position and draft year affect a player's scoring capabilities?

Hypothesis

My hypothesis is that younger players today score more points than players from earlier years will be correct. As I have been following the NBA for a long time, I have seen more young players score 40,50 and even 60 points in a single NBA game, which would support my hypothesis. However, this does not necessarily mean the players must score more on average. Regarding the question regarding the NBA draft, we will end up with the following: As the pick number increases, the scoring prowess will decrease. This is because, through recent memory, I have seen more success with earlier picks than failures.

Models, Assumptions, and Expected Results

For this project, I will be looking at two different models. These models observe how the year a player was drafted affects their Points per Game in their fifth season, and how their draft position (1st, 2nd, etc.) affected their Points per Game in their fifth season. If my hypothesis is correct, then the slope of the model regarding the year drafted and points per game should be positive, as it means that as the year increases, players are scoring more in their 5th season in the NBA. On the other hand, if my hypothesis is correct regarding how draft position affects scoring, then the slope of the second model should be negative, as that would mean that as draft position increases, the scoring would decrease. To help create this model, I have found and used the data from the top 15 picks from the NBA drafts from 2000 to 2017 and their points per game in their fifth season playing.

I chose the 5th season of playing as that is often considered the beginning of the prime of a player's career. According to some analyses, players often peak around 27-31 years old (Salameh 1). Players must be at least 19 years old to be drafted, their peak would be around their seventh in the NBA (NBAPA 296). However, if I used seven seasons, this could have issues, as when looking at players' statistics, many of them have less than seven years of experience in the NBA. I had to find a middle ground where many players played in the

NBA for a long time while also giving the players time to develop and adapt to the NBA. Therefore, I chose their fifth season. It also worked out well because, for many of the players whose statistics I observed, their fifth season was one of their best seasons from a scoring perspective. The reason I chose to focus only on the top 15 picks is these picks are the top half of the first round of the NBA draft. Therefore, in theory, these players are the most impactful and can help create a model that could be extended later.

Another assumption made is that there are no rule changes between seasons that affect the number of points scored. This assumption allows me to simplify the model as I do not have to worry about how outside forces affect scoring. This also allows me to use the fifth season a player played in the NBA, allowing me to ignore seasons where a player did not play at all due to a severe injury.

The reason I am using points per game as my stat of interest instead of just using points scored is that players may not play the same number of games in a season. For example, a player may miss four games due to an injury. Points per game allow me to look at the points without worrying that one player played more games than another.

Model Creation and Analysis

Firstly, I got all my data from an accurate basketball statistics website called Basketball Reference.com. The website had statistics for every season a player has played in the NBA. I went through all 270 players (15 players for 18 different years of the NBA draft) and put their points per game of their fifth active season in the NBA into a CSV so I could work with it in R. Here is a small sample of the data used.

```
##    Year Pick  PPG
## 1  2000    1 15.5
## 2  2000    2 10.1
## 3  2000    3 12.6
## 4  2000    4  6.2
## 5  2000    5 13.4
## 6  2000    6  6.1
## 7  2000    7  9.8
## 8  2000    8 17.7
## 9  2000    9  6.4
## 10 2000   10  5.2
## 11 2000   11  1.7
## 12 2000   12  4.7
## 13 2000   13  0.0
## 14 2000   14  0.9
## 15 2000   15  5.7
```

After getting the data into R, I plotted the players' points per game against both their draft year and the position they were picked in.

Although there were too many data points crowded together to create an accurate model, we see that there is a general shape that as the draft year increases, the points per game also increase. As the position picked increases, there are more players that have very low points per game. I do notice that for the graph involving points per game against the draft position, there are many top picks that scored many points per game (top left corner) compared to top picks that did not score a lot (bottom left corner of a graph). The inverse of this can be seen on the right side of the graph, where lower-drafted players do not tend to score a lot compared to other picks.

Something similar can be seen when plotting points per game against draft year. It seems that more recent players scored a lot in their fifth season (upper right corner in the graph) compared to those in earlier draft years (upper left corner of the graph).

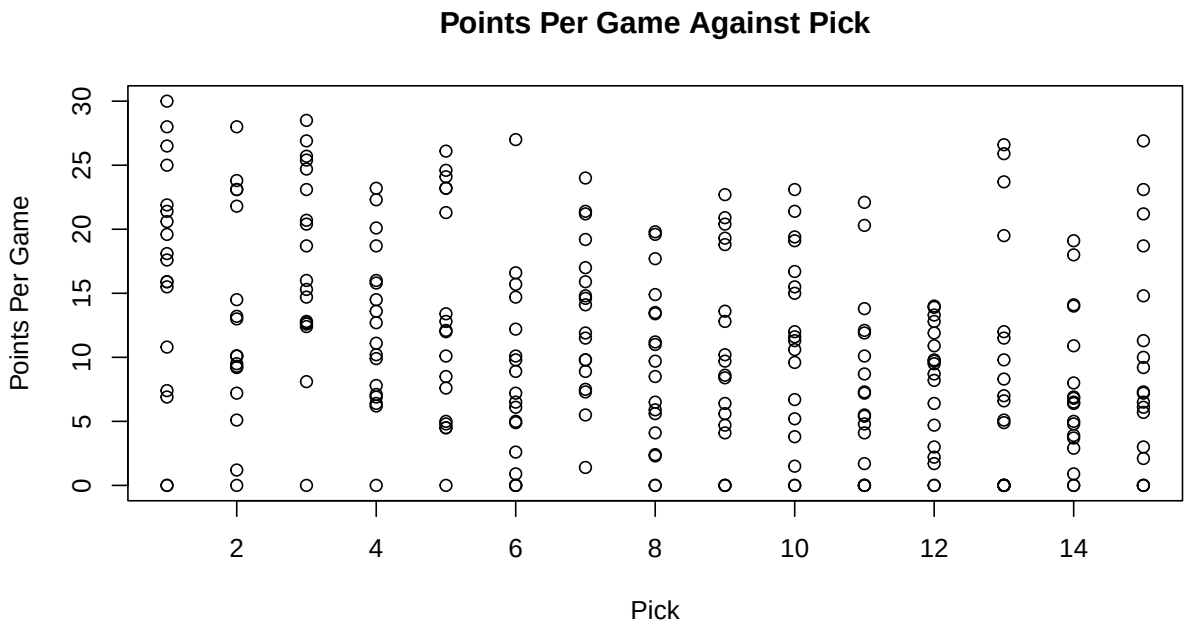

Figure 1: Figure 1, Graph of Points per Game in 5th season against Pick

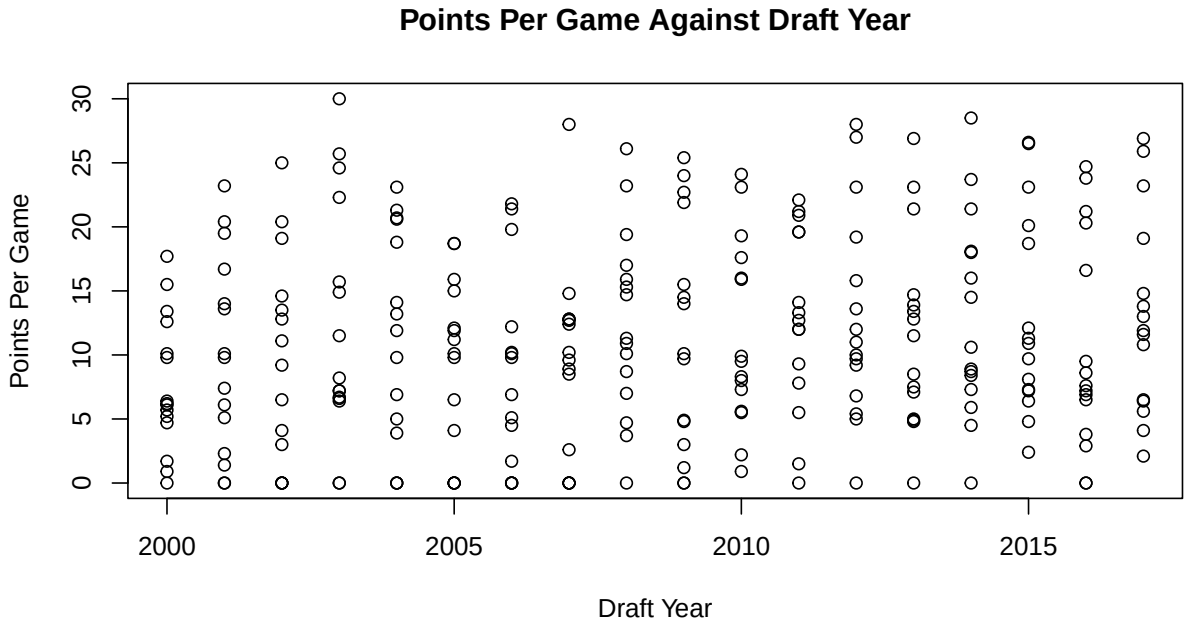

Figure 2: Figure 2, Graph of Points per Game in 5th season against Draft Year

Modelling Points Per Game Against Draft Year

Firstly, I decided to look at points per game against the year drafted to see whether players recently drafted are better scorers than players who came before them. What I decided to do was divide the data by pick and try to model each pick individually to see whether the model works for each pick (i.e. build a model for all the 1st overall picks, then do the same for the 2nd overall picks, etc.). If the model works for each of the different picks, then it would work for all the picks at once. By analyzing each of the graphs, I noticed the line of best fit must be linear so that y can be written in the form:

$$y = a + bx$$

Where y is the points per game for a player in their fifth season, b is the rate of change between years (the slope), x is the year the player was drafted, and a is the y-intercept. I used least square regression through R to verify that the models fit well and get the values of a and b for each set. Below are graph for 3 models for pick 1,5 and 11. I did calculate the values for a and b as well as the p-values for all of the picks.

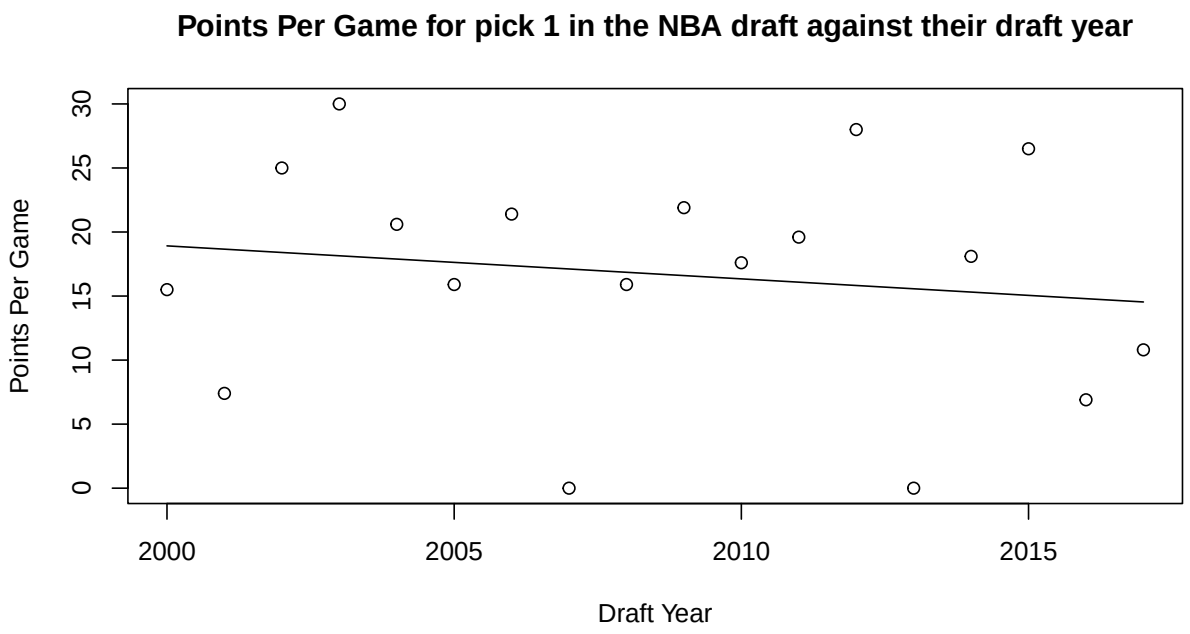

Figure 3: Figure 3,4,5, Graphs of Points per Game in 5th season against draft year for the 1st, 5th and 11th picks respectively

Here is a table of all of the YIntercepts(a), revisedYIntercepts(a'), slope(b) and p-values for each model:

```
##    DraftPick YIntercept revisedYIntercept      Slope     PValue
## 1          1   534.2948         19.175817 -0.2576883 0.53661524
## 2          2 -1368.1714          6.373203  0.6876161 0.05634619
## 3          3  -683.9222         14.386928  0.3493292 0.31712409
## 4          4   427.9486         14.326797 -0.2069143 0.48311265
## 5          5   708.8278         16.501307 -0.3463364 0.38228413
## 6          6  -327.1385          6.647059  0.1669763 0.62042910
## 7          7  -803.9802          9.235294  0.4068111 0.14367098
## 8          8   615.0942         12.093464 -0.3016512 0.31141383
```

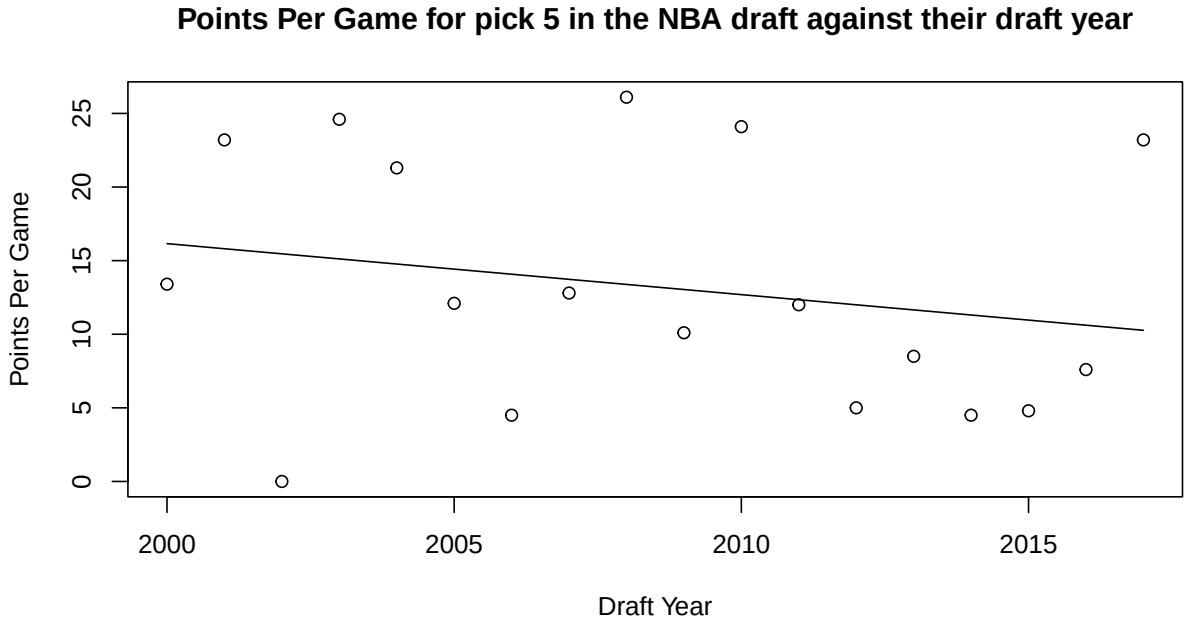

Figure 4: Figure 3,4,5, Graphs of Points per Game in 5th season against draft year for the 1st, 5th and 11th picks respectively

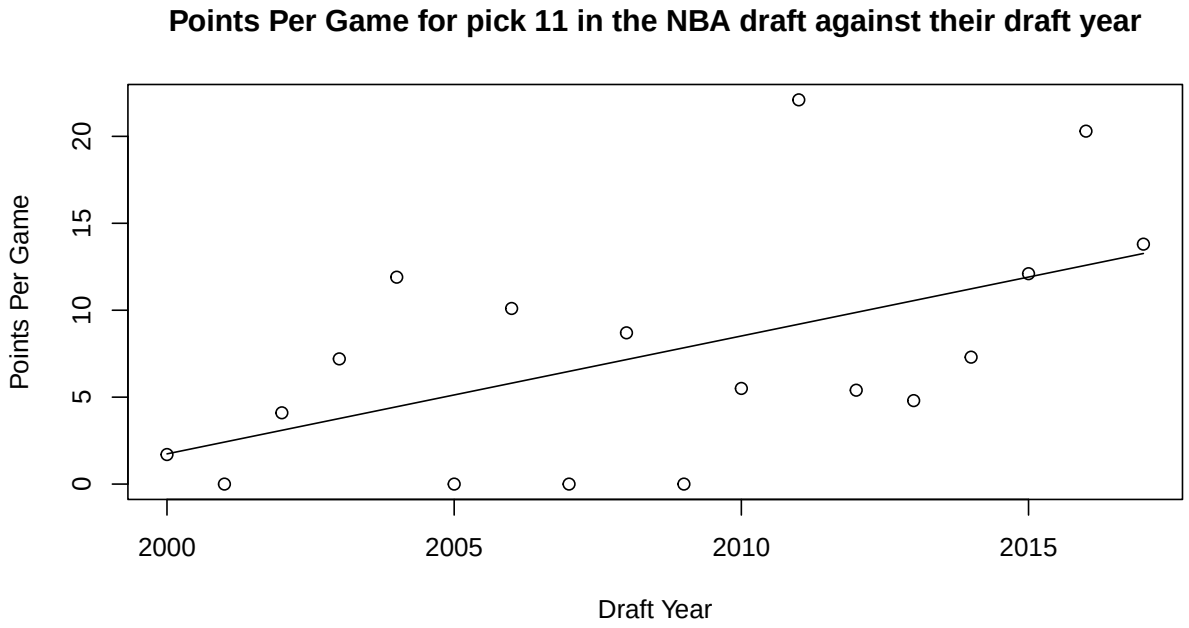

Figure 5: Figure 3,4,5, Graphs of Points per Game in 5th season against draft year for the 1st, 5th and 11th picks respectively

```
## 9          9  -468.4621          8.079739  0.2383901 0.50411140
## 10        10  -139.0248         10.539216  0.0748194 0.82854418
## 11        11 -1354.3003          1.058824  0.6780186 0.01949036
## 12        12  -800.3507          3.994118  0.4023736 0.06330993
## 13        13 -1554.5405          1.543791  0.7784314 0.06255879
## 14        14  -917.7430          2.952288  0.4605779 0.07700822
## 15        15  -431.2584          7.531373  0.2195046 0.57337526
```

For nearly all the picks, the model did not fit well as especially in the three example graphs, there were outliers. Something interesting was that some graphs (picks 1,4,5, and 8) all have negative slopes, meaning b<0, while all the other graphs had a positive slope (b>0). Also, the values for a initially did not make much sense. Some picks had negative values, while the others had very high positive numbers that would not make sense for a player. This is because when looking at this data set, the first x value we are looking at is 2000, meaning the large positive and negative values for a is at x=0, which would explain why they are so extreme. I tried to model it again, but this time in the form:

$$y = a' + b(x - 1999)$$

This form of the equation worked slightly better, as the b values made a lot more sense, and the value of b, as well as the p-value (which will be explained later in the paper), was the same between the two different equations. The changed value of a is the revisedYIntercept column in the table above.

The other reason I used linear least square regression through R is that it provides a p-value for every model. In summary, the p-value represents the probability that the linear model explains as much or more variance if the x and y values are unrelated (W3schools.com). The typical rule is that if the p-value is less than 0.05, there is a significant relationship between the x and y variables (W3schools.com). In this situation, with the different picks, only the model for the 11th overall pick had a p-value less than 0.05 (0.01949036). This means that for the 11th overall pick, it seems to be that as the years have gone on, young players are scoring more in their fifth season. However, for all the other models, the p-values were greater than 0.05, meaning I cannot conclude that there is a significant relationship between year drafted and scoring.

However, I was unsatisfied with that answer, so I tried another model with this relationship. Instead of doing it one pick at a time, I decided to take the average points per game for each of the top 15 in each draft and use that as my y-value. Here is the graph and the table of values for the equation.

```
##             YIntercept revisedYIntercept     Slope      PValue
## (Intercept)  -437.5151          8.962614 0.2233505 0.003089872
```

The graph created from this idea looked significantly better than the other models. Thankfully, this model had a p-value less than 0.05 (0.003089872). Because of this, I can conclude that there is a significant relationship between the year of a draft and the average points per game of the top 15 players in their fifth season in the NBA.

Modelling Points Per Game Against Draft Position

Next, I worked to see if it was true that top picks score more per game than later picks. Similar to the previous set of models, I started by separating the data and seeing if the model fit for each individual draft year. This time, instead of a linear model, I found that the points scored followed an exponential model described by:

$$y = ae^{bx}$$

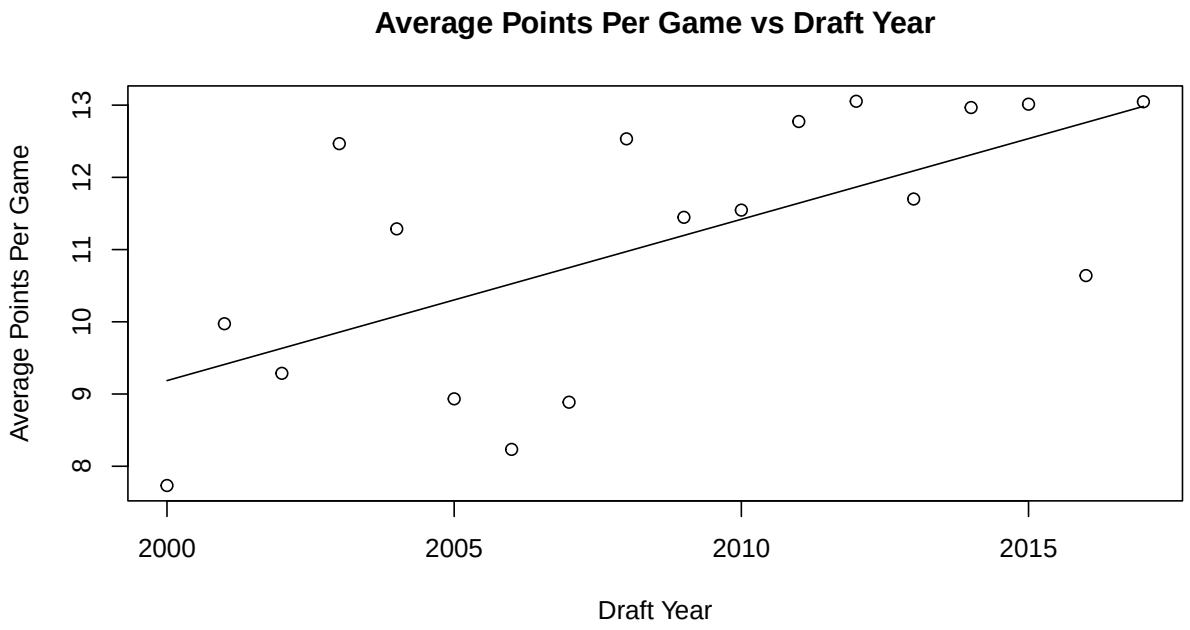

Figure 6: Figure 6, Graph of Average Points Per Game against Draft Year

Where Y is the points per game, a is the y-intercept, x is the position the player was drafted, and b is the exponential rate of change. As d approaches 0, the line created by this equation becomes more and more linear. This was certainly seen throughout the 18 different models. To get the values of a and b, I created a method that used the least-square method taught in this class. However, it should be noted that when working on these values, some scored 0 points per game, and as ln(0) equals infinity, I used ln(y+1) to ensure the model functioned correctly. Below are three example graphs from the 2000, 2007 and 2011 NBA drafts. I did get the values for a and b of all of the draft years.

Below is a table of all of the vlaues of a and b for each year.

```
##    DraftYear         a           b
## 1       2000 19.427155 -0.130554329
## 2       2001  9.119915 -0.023930033
## 3       2002  6.279712 -0.002879463
## 4       2003 31.556539 -0.150510327
## 5       2004 15.124083 -0.079743937
## 6       2005 30.194104 -0.195763753
## 7       2006 13.084759 -0.107876451
## 8       2007 10.672031 -0.069363123
## 9       2008 17.922493 -0.062451970
## 10      2009 12.806574 -0.056029384
## 11      2010 14.918168 -0.047431654
## 12      2011  8.342867  0.033109386
## 13      2012 24.756947 -0.099201941
## 14      2013  6.056382  0.064030425
## 15      2014 20.839156 -0.078126421
## 16      2015 12.524132 -0.006083201
## 17      2016 13.682836 -0.065647562
## 18      2017 14.288574 -0.023060532
```

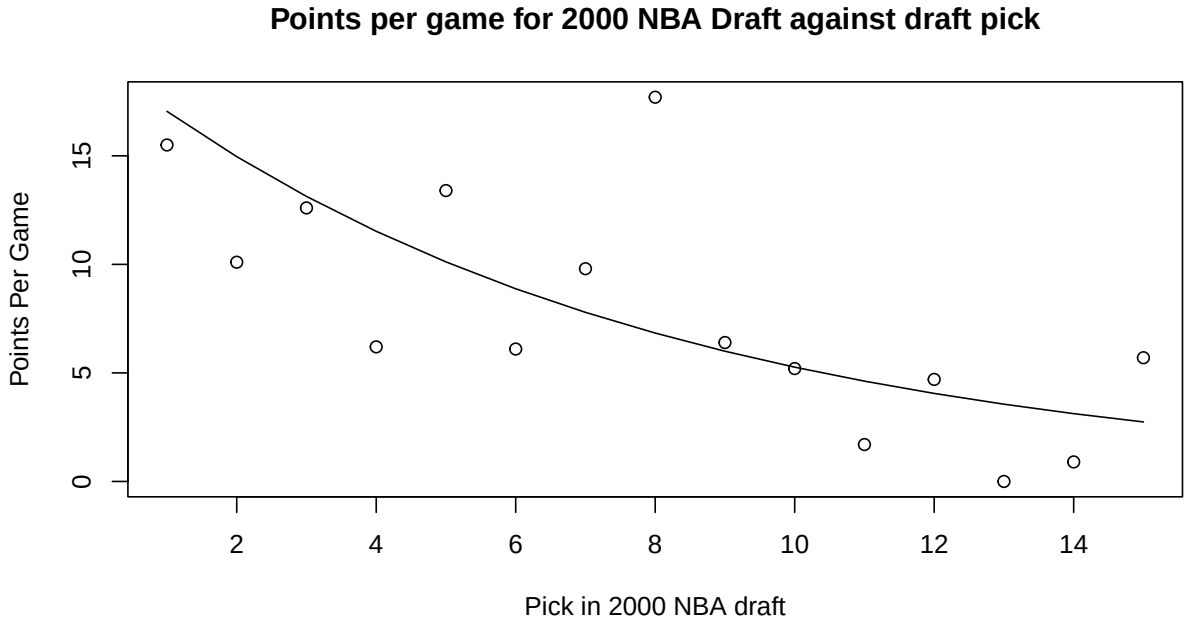

Figure 7: Figure 7,8,9, Graphs of Points Per Game against Draft Position for the 2000, 2007 and 2011 NBA Drafts respectively

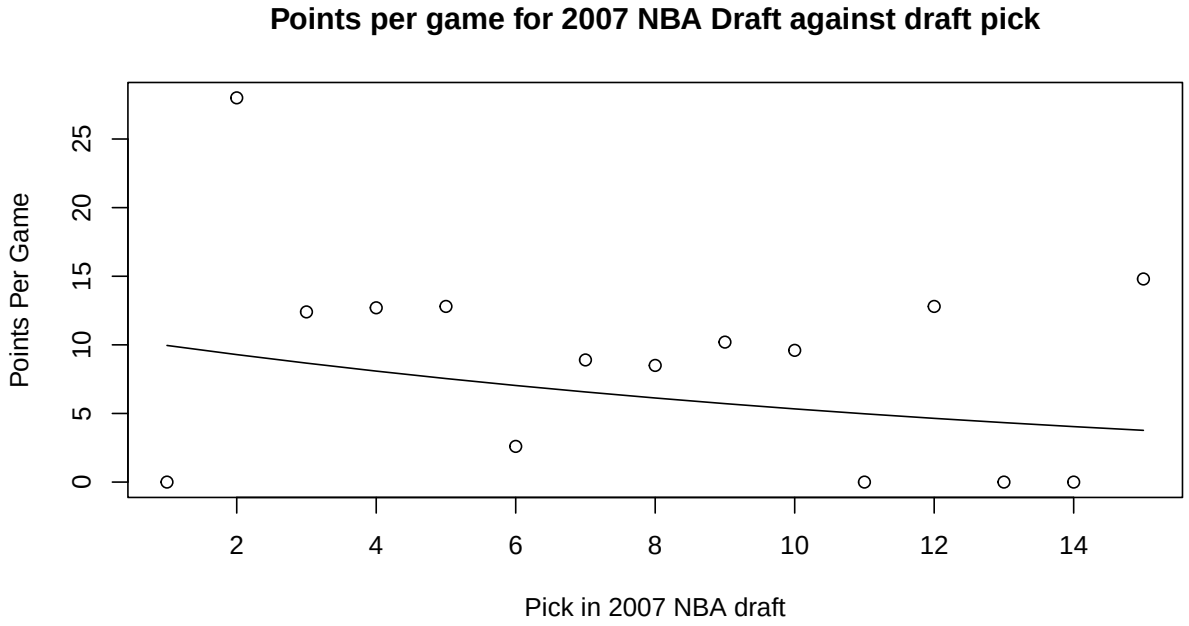

Figure 8: Figure 7,8,9, Graphs of Points Per Game against Draft Position for the 2000, 2007 and 2011 NBA Drafts respectively

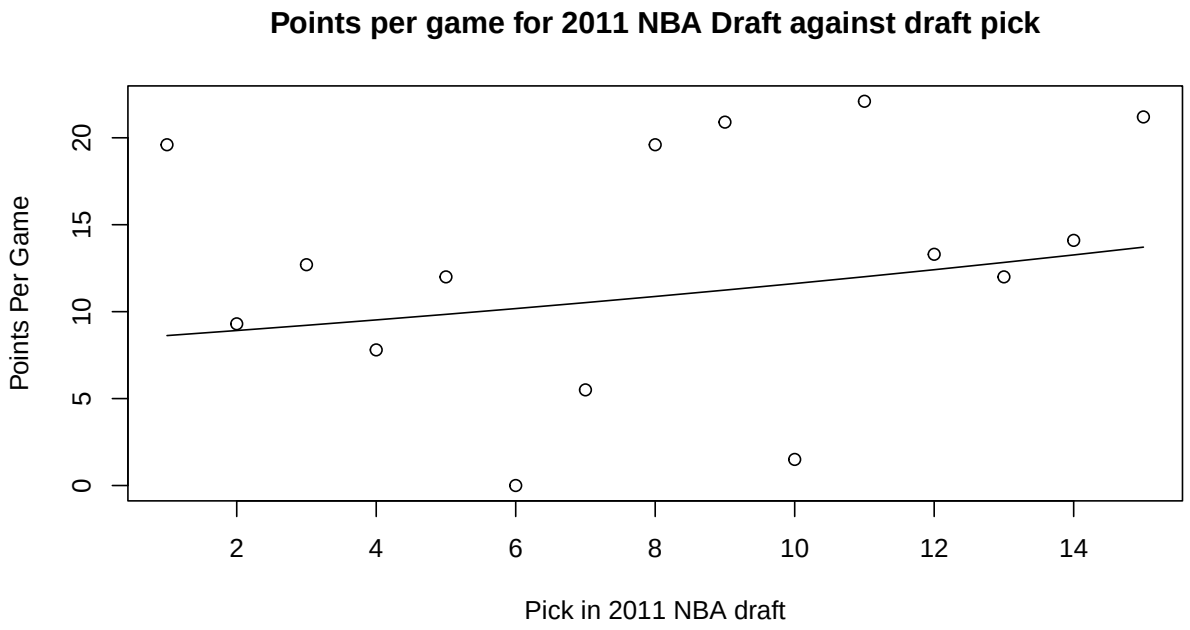

Figure 9: Figure 7,8,9, Graphs of Points Per Game against Draft Position for the 2000, 2007 and 2011 NBA Drafts respectively

Some of them, such as the 2000 NBA Draft, look like a traditional exponential model where b<0. Others, such as the 2007 model, have a line that looks very linear, as the b value is less than 0 but is very close to 0. Finally, we have years such as 2011, where the model looks linearly positive, meaning that the b-value is greater than 0, but still very close to 0. I also noticed that compared to the models whose lines of best fit looked more linear, the years that looked more exponential seemed to fit the data better. Unfortunately, I could not find how to directly determine whether there is a relationship between draft position and scoring, meaning there is not a clear-cut way to determine whether the model works in this form.

However, I can convert this model to a linear equation. By taking the logarithm of both sides of the equation, I get:

$$ln(y) = ln(a) + bx$$

In this equation, ln(a) is the y-intercept and b is the slope. Below are the linear versions of the three models created earlier. Similar to previous methods, I did get all of the vlaues for all of the other draft years.

Below is a table of ln(a), b and the p-values for each draft year and also verifying that ln(a) found in the linear model is equal to ln of the a value found in the exponential model and that the two b values match.

```
##    DraftYear      lna lnofaIslna            b bvaluesMatch      pValue
## 1       2000 2.966672       TRUE -0.130554329         TRUE 0.003122229
## 2       2001 2.210460       TRUE -0.023930033         TRUE 0.719046390
## 3       2002 1.837324       TRUE -0.002879463         TRUE 0.970654996
## 4       2003 3.451781       TRUE -0.150510327         TRUE 0.009433032
## 5       2004 2.716288       TRUE -0.079743937         TRUE 0.272604416
## 6       2005 3.407647       TRUE -0.195763753         TRUE 0.002025735
## 7       2006 2.571448       TRUE -0.107876451         TRUE 0.138441202
## 8       2007 2.367626       TRUE -0.069363123         TRUE 0.355255945
```

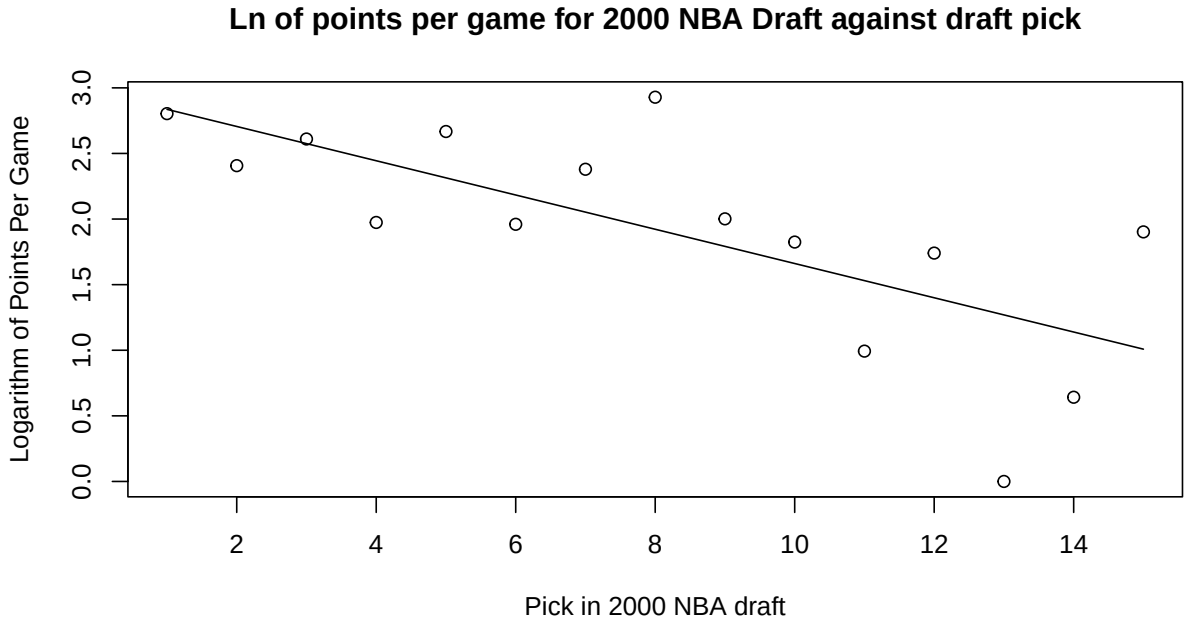

Figure 10: Figure 10,11,12, Graphs of the logarithm of Points Per Game against Draft Position for the 2000, 2007 and 2011 NBA Drafts respectively

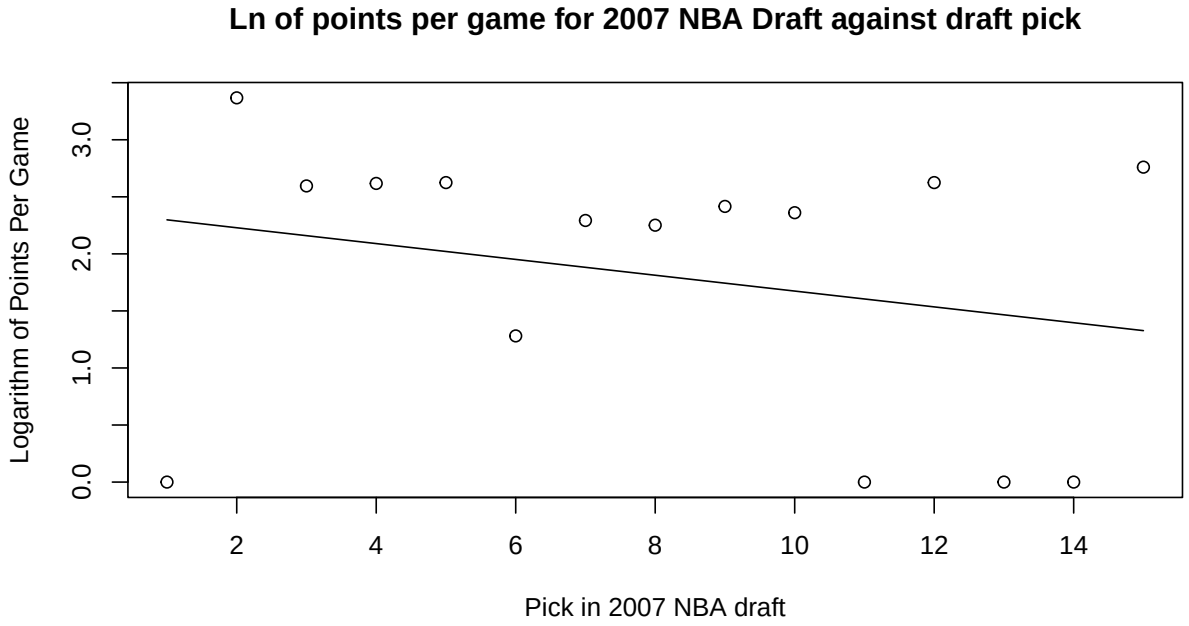

Figure 11: Figure 10,11,12, Graphs of the logarithm of Points Per Game against Draft Position for the 2000, 2007 and 2011 NBA Drafts respectively

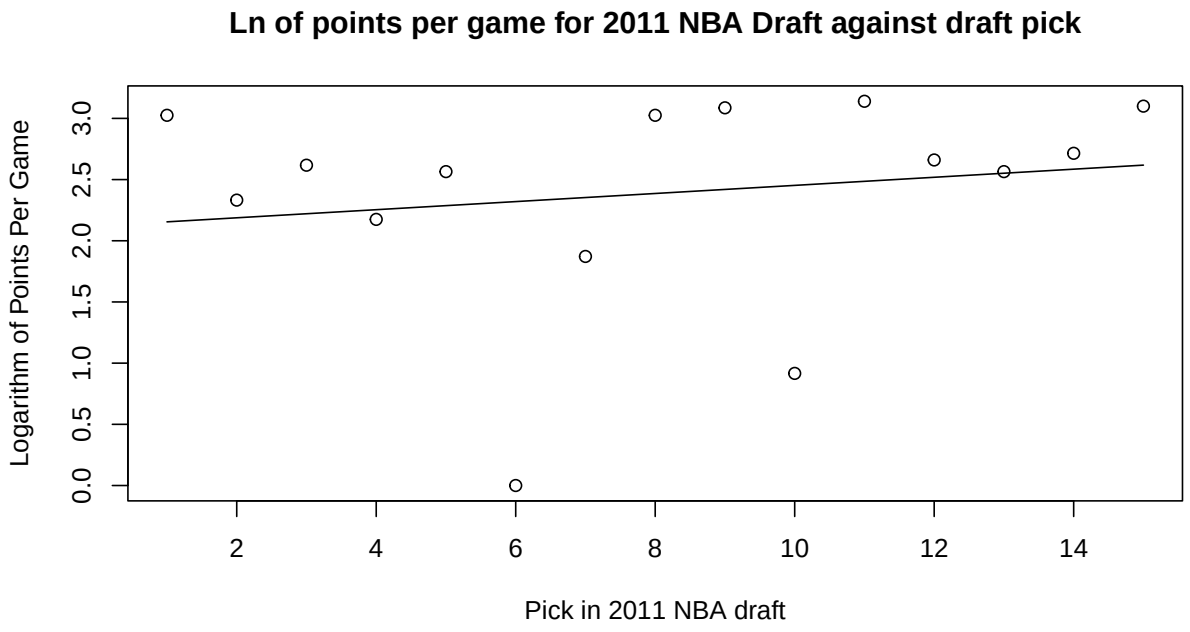

Figure 12: Figure 10,11,12, Graphs of the logarithm of Points Per Game against Draft Position for the 2000, 2007 and 2011 NBA Drafts respectively

```
## 9       2008 2.886057       TRUE -0.062451970        TRUE 0.217997044
## 10      2009 2.549959       TRUE -0.056029384        TRUE 0.422545057
## 11      2010 2.702580       TRUE -0.047431654        TRUE 0.295859965
## 12      2011 2.121407       TRUE  0.033109386        TRUE 0.549077723
## 13      2012 3.209106       TRUE -0.099201941        TRUE 0.039830520
## 14      2013 1.801113       TRUE  0.064030425        TRUE 0.200029126
## 15      2014 3.036834       TRUE -0.078126421        TRUE 0.117686282
## 16      2015 2.527657       TRUE -0.006083201        TRUE 0.874041254
## 17      2016 2.616142       TRUE -0.065647562        TRUE 0.302331190
## 18      2017 2.659460       TRUE -0.023060532        TRUE 0.563145236
```

Using the linear modelling method in R, I used earlier; I confirmed that my previous calculations for a and b worked as values of b matched while the intercept is ln(a) by modelling the line of best fit between x and ln(y). This also provided p-values I could use to determine whether there is a relationship between the points per game and draft position. Some of these graphs did seem to have outliers that were far away from the line, so I was doubtful that this model would be effective for every year.

Observing the table of all the p-values, one can see that there were only four years where the p-values were less than 0.05. These years were 2000, 2003, 2005 and 2012. Because all these years had negative slopes, I could conclude for those years that as draft position increased, points per game in a player's fifth season decreased. However, I could not conclude something similar for any of the other years, as all of them had a p-value greater than 0.05, meaning that for these years, there is not a significant relationship between draft position and points per game.

```
##             a     loga           b       pValue
## value 16.8239 2.822801 -0.04529878 0.0004766719
```

Finally, just like the previous model, I decided to try to model it using the averages of each pick as the

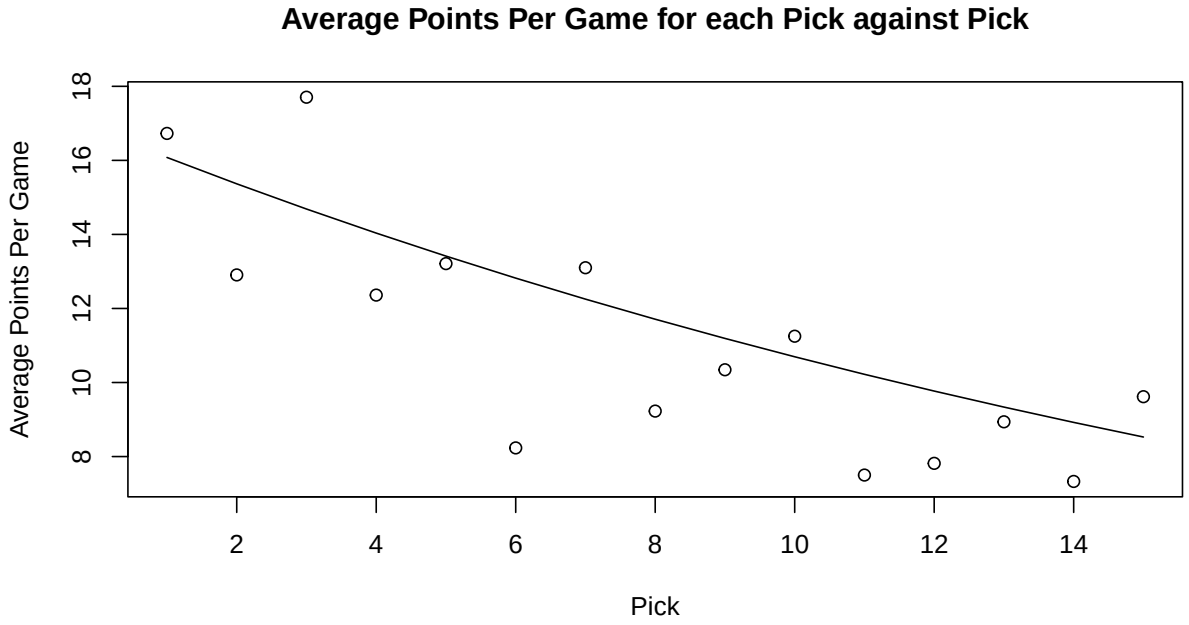

Figure 13: Figure 13,14, Graphs of Average Points Per Game against Draft Position, the logarithm of Average Points Per Game against Draft Position and a table providing the values of both models

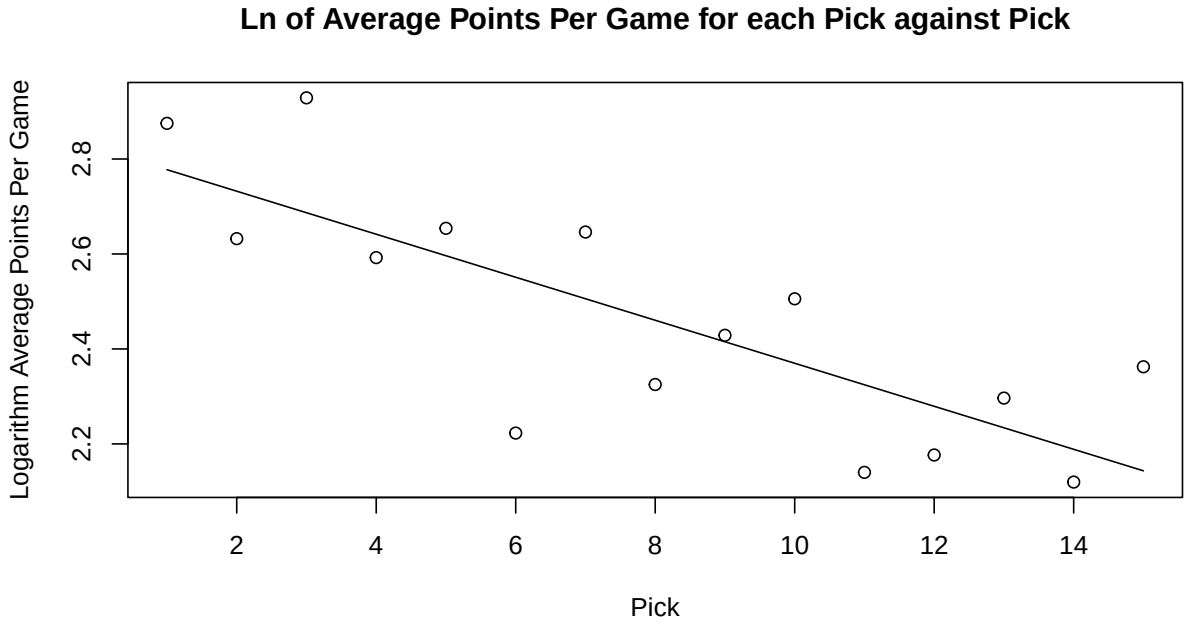

Figure 14: Figure 13,14, Graphs of Average Points Per Game against Draft Position, the logarithm of Average Points Per Game against Draft Position and a table providing the values of both models

y-values. The exponential graph looked similar to the years with a p-value less than 0.05. When looking at the linear graph of x and ln(y), we see that the line fits almost perfectly, which is reflected in the p-value (0.0004766719). Because the slope is negative, I can conclude that as draft position increases, the average points per game decreases.

Conclusion

Based on the results, it seems that my hypothesis is correct. Players drafted at a higher position tend to score more than players drafted later, and young players score more than those who came before them. However, it does not look like the models work on a year-by-year or a pick-by-pick basis, as many of the p-values for individual models were greater than 0.05, meaning there was no significant relationship between points scored and draft year/ draft pick. This makes some sense, as it seems that there is still a bit of luck when it comes to drafting players to play in the NBA. Every year ends up slightly different, but by averaging scoring, it cancels out extreme outliers.

Possible Improvements

Regarding possible improvements, one could be to expand the timeframe of the data used. I would be very interested to see if the model still holds true when using players from the 1980's and 1990's. Another possible improvement could be calculating their points per 48 minutes of play (which is the total points scored divided by minutes played and then multiplied by 48, which is the length of an NBA game) (BasketballReference.com). Using this statistic would remove the assumption that players would have played the same number of minutes per game, as the scoring statistics would have taken that into account. Another possible improvement would be to add an additional variable for rule changes. There have been slight rule changes throughout the history of the NBA. Some of these rule changes could have influenced scoring, and I believe they should be accounted for when creating another model. (Zsolt) Finally, I would like to investigate whether these models work for other important basketball statistics, such as rebounds or assists. Even though the winning team is the one who scores more points than their opponent, assists and rebounds are very important aspects of the game of basketball, and I would be very interested to see if similar models to what I created in this project could work for these statistics as well.

Bibliography

"Data Science- Regression Table: P-Value". W3schools.com. 2024. https://www.w3schools.com/datascience/ds_linear_regression_pvalue.asp

"Four Factors", Cleaning The Glass. https://cleaningtheglass.com/stats/league/fourfactors?season=2023&seasontype=regseason&start=10/1/2023&end=10/15/2024

"Glossary". BasketballReference.com. https://www.basketball-reference.com/draft/

NBAPA, "COLLECTIVE BARGAINING AGREEMENT", July 2023, https://nbpa.com/cba

Salameh, Tony. "An Empirical Analysis of Prime Performing Age of NBA Players; When Do They Reach Their Prime?". Bryant Univeristy. https://digitalcommons.bryant.edu/cgi/viewcontent.cgi?article=1223&context=eeb#:

Zsolt, Hartyàni. "History of Basketball", Basketref.com. https://www.basketref.com/en/index.php/rules/rules-history

Data

"NBA Draft Index". BasketballReference.com. https://www.basketball-reference.com/draft/

Note about the data: The link I used is a hub that allows me to get access to every single draft class. Then, I went to each individual draft and got the points per game from each of the player's fifth seasons.

Note about coding: All of the code was created by me using techniques learned in other classes.