**Analysis of NBA Player Statistics**

Samuel Flohre

February 2024

First, I found the dataset on Kaggle and imported it into R-Studio software for further analysis. Below is the link to the following dataset:

<https://www.kaggle.com/datasets/jacobbaruch/basketball-players-stats-per-season-49-leagues>

There were a few questions that I intended to explore:

1. The Three-Point Shot has revolutionized the sport of basketball. How can I track and visualize the emergence of the three-pointer given the data I have currently?
2. Is there any difference in scoring between NBA Players who are taller vs shorter?
3. How can I predict points scored using multiple factors, like draft rounds and shooting efficiency?

**Question 1: Progression of 3-PT Shot**

Unfortunately, the dataset includes 48 other professional basketball leagues, so I had to single out all the observations pertaining to the NBA. I chose to do this because each basketball league has unique play styles, and my lack of knowledge of most other leagues led me to focus on the NBA, being one of the more well-known leagues across the world.

A graph with a line and a line

Description automatically generated

Here is a line chart that I created using the GGPLOT function in R that highlights how many total three-pointers were attempted in each season, along with a linear fit:

\*Note that for example, the year 2005 represents the 2004-2005 NBA season, and so forth.

Notice that in 2012, there was a lower number of 3PT attempts than expected. This is due to the fact in 2011 there was an NBA lockout, and the regular season was shortened significantly.

On another note, I also wanted to see which teams were historically efficient when shooting 3-pointers, so I programmed R to output each NBA team and their mean 3P% from 1999 to 2020, and the results were the following:

The Phoenix Suns had the highest 3P%, sitting at about 0.325, meaning that from 1999-2020, Suns players made about 32.5% of their three-pointers. Conversely, the Detroit Pistons had the lowest percentage, around 28.5%.

(Technically, the Vancouver Grizzlies had the lowest percentage, but their data only accounted for one season, until they moved to Memphis)

There is no doubt that 3-PT shooting has formulated itself as a vital part of NBA offenses in today’s era of basketball, and based on current data, it is only projected to increase in the coming years.

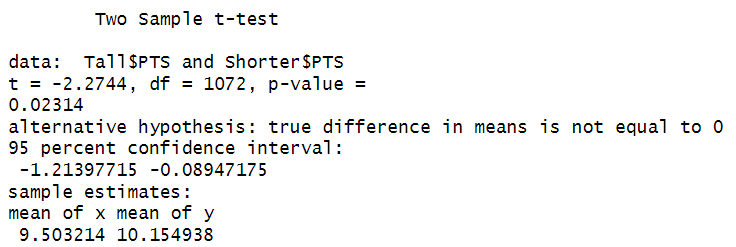
**Question 2: Difference in Scoring between Tall and Short NBA Players**

Compared to the average person, most NBA players are extremely tall. There is an obvious benefit to being tall when it comes to playing basketball, but I wanted to show that even smaller NBA players are just as capable. To do this, I conducted a T-Test to determine if the mean scoring averages among players shorter/taller than the mean height are significantly different.

According to R, the mean height was found to be 80.3 inches, or 6’8”. The two samples for this test included players that were shorter than the mean, as well as those equal to or taller than the mean.

A black text on a white background

Description automatically generatedThe output below is the equal variance test, as well as the two sample T-Test:



Based on the results, it seems that shorter NBA players have a slightly higher mean scoring average than taller players, and the p-value of 0.02314 indicates that this is a significant difference between the two.

Now, we must check and verify the assumptions for the T-Test:

A graph of a distribution of points

Description automatically generatedA graph of a distribution of points

Description automatically generated

A graph of a normal q-q plot

Description automatically generatedA graph with a line

Description automatically generated

A black text on a white background

Description automatically generatedA black text on a white background

Description automatically generated

Both samples appear to not be normally distributed (but instead right-skewed), but by the Central Limit Theorem, due to the large sample size, we can assume both are approximately normal for the sake of our analysis.

The observations are independent of each other, and we have already shown that the variances are equal.

**Question 3: Predicting Points scored based on several factors:**

To predict PTS, the dependent variable must be the number of points scored.

I sorted the NBA data by player and got each player’s average number of points scored per season, along with their mean TS% and games played, as well as which draft round they were selected.

The independent variables to be the following:

* **Games played**: this has an impact on points scored because the more you play, the more your total points increase.
* **Draft Round**: This includes rounds 1 and 2, as well as undrafted players. Normally, players are drafted based on their ability to score in the NBA (as well as other factors). I hypothesized that first-round picks tend to score more points.
* **TS%**: True Shooting Percentage is a somewhat newer metric that is known to be more effective at measuring a player’s shooting efficiency compared to conventional FG%.

The perfect method to conduct this analysis is a linear regression model:

A screenshot of a computer

Description automatically generated

Based on the R Output, all the parameters are statistically significant (shown by the extremely small p-values). However, the R-squared is very small, so only 22% of the variance is explained by the model, which is not ideal. In terms of interpretation, for example, the GP parameter has a value of 15.422, which means that for every 1 game played, the number of points scored is estimated to increase by 15.422 points on average. Interestingly, the intercept has a negative value, which is unrealistic due to the impossibility of having “negative” points. In addition, TS% has a large positive relationship, which makes sense because the higher your TS% is, the more efficient the scorer you are, and teams will most likely trust you with more scoring opportunities.

**Assumptions for Linear Regression:**

1. Independent observations
   1. The observations are independent because each one is a different player and has no effect on the other observations.
2. Linear Relationship between each independent and dependent variable.
   1. This can be verified by plotting the continuous independent variables (GP and TS%) against the dependent (PTS)
   2. A graph of a number of dots

      Description automatically generated with medium confidenceA graph with numbers and dots

      Description automatically generatedHowever, the results of these plots are ambiguous, as there is no clear linear relationship in either of the graphs.
3. Little/No Multicollinearity
   1. For this assumption, we will look at the Variance Inflation Factor (VIF) values of each parameter. Values less than 10 indicate verify this assumption.

A black text on a white background

Description automatically generated

1. Homogeneity of Variance & Normal Distribution of Residuals
   1. Based on the outputs below, we can conclude that the residuals are not normally distributed. More specifically, the QQ plot does not demonstrate a linear trend. The Shapiro-Wilks normality test having a very small p-value indicates a rejection of the null hypothesis of a normal distribution.
   2. Also, the variances are not homogeneous according to the cone-shaped pattern of the residuals vs. fitted values plot on the bottom left. The plot needs to have a random or no pattern for variances to be considered equal.

A black text on a white background

Description automatically generatedA group of graphs and diagrams

Description automatically generated

Not all the assumptions are satisfied, so it is necessary to conduct a transformation of the data. Therefore, I performed a log transformation on PTS below to attempt to validate the regression.

A screenshot of a computer program

Description automatically generatedLog-Transformed Model:

The plots for the assumptions look much better than before. Specifically, you can see a slightly positive linear relationship between GP and TS% vs. PTS in the top two plots.

Furthermore, normality can be assumed based on the QQ Plot mostly following a linear path, only deviating at the extremes, which is common. Also, the Shapiro-Wilks Normality test indicates the residuals are normally distributed at α = 0.01 level of significance.

A group of black dots

Description automatically generatedLastly, the residual vs. fitted plot demonstrates a much more random distribution, so the homogeneity of variance assumption is also satisfied.

A black text on a white background

Description automatically generated