Final Project - 2016 NFL DRAFT Pick Analysis

Tyler Millan

6/12/2018

## Introduction

Every year after the end of the NFL season, all sports news coverage seems to shift towards covering the Draft. It is a time for teams to either rebuild or strengthen their respective rosters with college talent. Many teams scout players years in advance in hopes of landing the next big star. In this statistical report I analyze the 2016 NFL draft which occurred between April 28th to April 30th, 2016.The basic format of the NFL draft is each team is given a position in the drafting order in reverse order relative to its record in the previous year, which means that the last place team is positioned first. From this position, the team can either select a player or trade their position to another team for other draft positions, a player or players, or any combination thereof. In the 2016 draft specifically there were 7 rounds of 32 picks which is 253 total picks. The dataset I am working is a combination of data that was taken during the days of the draft, as well as data which is current up to when the 2018-2019 NFL season starts in August. The data was collected by the National Football League and pulled from the pro football reference website. The specific variables I am using to analyza drafter players are as follows:

**Quantitative Variables:**

* *GamesPlayed*: The number of regular season NFL games the player has played in since being drafted in 2016 which at can range from 0-32.
* *Age*: A player’s age in years at the time of being drafted in 2016.

**Categorical Variables:**

* *RoundChosen*: the round the player was chosen in the draft out of the 7.
* *PickChosen*: the ranked pick the player was chosen, out of the 253 total picks.
* *College/Univ*: Indicates which college a player attended prior to being drafted.
* *Team:* Indicates which NFL team the player was drafted by

The motivation behind this analysis is to gain a deeper understanding of how players have performed after being drafted based on measurable characteristics as well as what teams will look for on draft days. Given these variables to work with, the specific question’s I look to answer will be the following:

*What does the age distribution of NFL players in the 2016 NFL Draft look like?*

*Does a player’s age at the time of being drafted have any correlation/effect on the amount of playing time they get?*

*Does the university/college that a player attended have any relationship with the NFL team that they are drafted by?*

*Is there a difference in the mean number of games played for players drafted in the first round to the mean number of games played in the seventh round?*



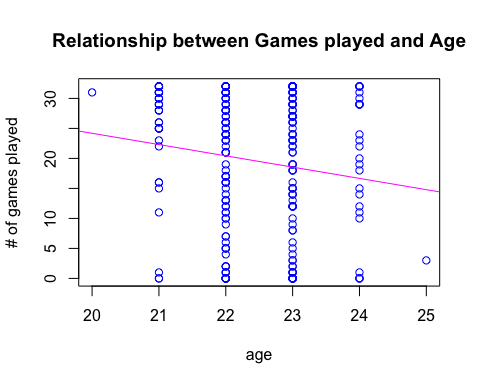
**What does the age distribution of NFL players in the 2016 NFL Draft look like?**

I first want to see what the age breakdown of players that were drafted. Since players need to be in at least three years of college, we can expect the age of players drafted to be around 21 years old assuming most people graduate highschool when they are 18. After running some analysis on the age data of all players in the 2016 NFL Draft class, I find that the mean age of players in this class is 22.42. The standard deviation of age for the players is 0.89, which means we would expect a majority of player ages to lie within about a year or two of our mean of 22.42. After running a 5-number summary of our data we find that the youngest player out of all of the observed players in the draft class is 20 years old while the oldest player is 25 years old, which confirms our previous expectation based on our previously computed mean and standard deviation.

We also find that the median age of players is 22 years old which is very close to our mean, indicating that outliers are most likely not affecting the mean which is more susceptible to being affected by outliers. We also see that our 1st quartile age resides on 22 years of age which is equal to our mean. This can be attributed to ages in our dataset not being spread. We can also infer from this that 22 is a very common age out of the 253 total observations and will later be able to confirm this by looking at the histogram of age frequencies.

We also observe that the third quartile or 75th percentile age is 23 years old. Another way to find any outliers is by taking the interquartile range which we would find by subtracting our 3rd quartile observation of 23 and subtracting it by our 1st quartile observation of 22 and multiplying that, which in this case is 1, by 1.5 and anything greater than quartile 3 plus 1.5 which would be 24.5 in this case would be considered an outlier. We would also perform the inverse of this, so quartile 1 minus 1.5 which is equal 20.5, and any observation less than 20.5 is considered an outlier. Realistically this means that any observation of 20 years old or 25 years old in the data is an outlier. We find that there are 2 outliers in our data that fit these criteria, one which is a single observation that happens to also be to minimum age of 20 years old and the other which also happens to be our maximum age of 25. After graphing the frequency histogram of player ages, we see that the shape of our data is approximately uniform and symmetric. We confirm our prior inference that 22 is the most common age out of all of our observations. We also confirm that 20 and 25 looks to the be outliers in the data.

##   
## Call:  
## lm(formula = gamesplayed ~ age)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -22.306 -7.425 3.457 9.457 15.338   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 61.8096 17.4934 3.533 0.000488 \*\*\*  
## age -1.8811 0.7797 -2.413 0.016553 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 11.02 on 251 degrees of freedom  
## Multiple R-squared: 0.02267, Adjusted R-squared: 0.01877   
## F-statistic: 5.821 on 1 and 251 DF, p-value: 0.01655



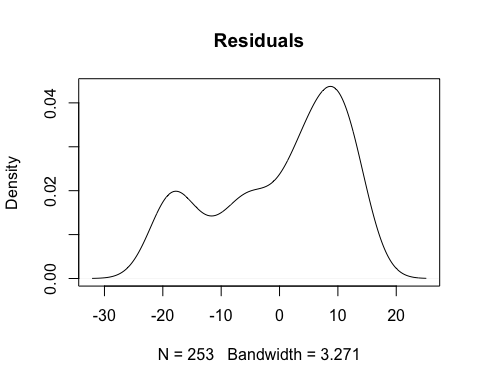
**Does a player’s age at the time of being drafted have any correlation/effect on the amount of playing time they get?**

Another question I would like to further explore is if a players age has any effect on their playing time since being drafted. The motivation behind this is to possibly confirm the commonly held belief that younger players are less game ready than older, more expierenced players. After running a linear regression, I find that the relationship between the number of games played since being drafted is negative in relation to the age of a player at the time of being drafted.

We see that the coefficent on the age is -1.881, which we can interpret that for every one-year increase in a players age at the time they get drafted decreases the amount of games played since being drafted by 1.881 games. This is the inverse of the commonly held belief that older players will be played more than their younger counterparts. After checking the correlation among the two variables, we see the correlation is -0.150552 which implies low negative correlation amongst the two variables.

The scatterplot graph’s appearence above can be attributed to age being a discrete number in addition to the spread of ages being small. We see that all of our observations fall on their respective age intervals. And each observation with more games played end up on the upper part of the graph.

The residuals for the given scatterplot are graphed below:



After observing the residual graph above, we can see a somewhat abnormal distribution of negative residual values. There also seems to be a disproportionate amount of positive residual values. Based on these two observations, there may be bias in our best fit line above of the relationship between age and games played.

To gain a better understanding of the possible bias in my regession line above, I want to compute a 95% confidence interval about the slope of the regression line I previously computed. We recieved a slope of -1.584 in the instance above, but I would like to see how large the range of the true slope of the regression line falls.

We find this by taking our coefficient on age or the slope which is -1.584 and adding/subtracting the standard error multiplied by 2(since 95% of our observations lie within two standard errors of the mean).

We observe a confidence interval of:

**(-3.4405, -0.3217)**

We notice from our 95% confidence interval of the slope of our regression line that 0 is not contained in the range of values implying that we can be 95% confident that there is an effect of age on games played in our data set. Though we can also observe that the range of our values is quite large due to a comparatively high standard error, which means our regression line slope may be slightly inaccurate.

***Does the university/college that a player attended have any relationship with the NFL team that they are drafted by?***

Next, I want to observe if there is a relationship between NFL teams drafting from specific college programs in the 2016 NFL Draft, which would be possible for a number of reasons. Some include the fact that most NFL coaches were previously college football coaches and may prefer players that either they personally coached or were in a system that the coach still uses with there NFL team. A notable scenario of this happening came with Chip Kelly, a former head coach for the University of Oregon that, after becoming the head coach of the Philadelphia Eagles, drafted/traded for Oregon alumnus.

To test this, I run a Chi-Squared Test of Independence between my two categorical variables of NFL teams that drafted players with the college’s/universities that those players attended prior to the draft. The null and alternative hypothesis of my proposed test are as follows:

**Null Hypothesis (H0)**: NFL team’s draft pick decisions are independent of a specific players college.

**Alternative Hypothesis (HA)**: NFL team’s draft pick decisions and the college a player attended are related.

The p-vlaue from my Chi-squared test yielded a p-value of .5744. This means that I fail to reject the null hypothesis at any meaningful significance level heavily implying that in the 2016 NFL Draft there seems to be no identifiable relationship between a specific NFL team and a specific college program.

## Warning in chisq.test(crosstab): Chi-squared approximation may be incorrect

##   
## Pearson’s Chi-squared test  
##   
## data: crosstab  
## X-squared = 3146.4, df = 3162, p-value = 0.5744

***Is there a difference in the mean number of games played for players drafted in the first round to the mean number of games played in the seventh round?***

We would expect players drafted in earlier rounds of the draft to be much more desirable for teams than players at the end of the draft. We we would also assume players more valuable would recieve more playing time that those that are not. I am interested in seeing if there is a signficant enough difference between playing time of players in the first and last round.

**Null Hypothesis (H0)**: The mean number of games played for players drafted in the seventh round is more than or equal to the amount of games played in the first round. (m1-m2=0: where m1 is the mean number of games played for first round players)

**Alternative Hypothesis (HA)**: The mean number of games played by players drafted in the first round is greater than the amount of games played by players drafted in the seventh round. (m1-m2>0)

**T-Test Results**

##   
## Two Sample t-test  
##   
## data: firstrnd$X\_\_11 and lastrnd$X\_\_11  
## t = 5.462, df = 61, p-value = 4.582e-07  
## alternative hypothesis: true difference in means is greater than 0  
## 95 percent confidence interval:  
## 9.382382 Inf  
## sample estimates:  
## mean of x mean of y   
## 24.48387 10.96875

After running a two-sample t-test on our data we see that we get a t-statistic of 5.462 and a p-value of 4.582e-07 which we can round to being 0. This means I am able to reject the null hypothesis that there is no difference in the amount of games played by first round players and seventh round players. We can reject this null hypothesis at any significance level, which means that there is clear evidence to support the alternative hypothesis. We can conclude that players drafted in the first round are significantly more impactful early in their careers than those drafted in the last round. After running this test, we also see that mean number of games for players in the first round is 24.48387 while the mean number of games played in the seventh round is 10.96875.

We can use a small simulation to test the above hypothesis test by randomly sampling 1 player from the last round of the draft and compare their amount of games played to a randomly sampled player from the first round of the draft. We will take the proportion of every success (instance where the first-round player had more games played then the last round player) to the total amount of simulations.

**Small Simulation results (10 iterations): 0.8**

**Large Simulation results (10,000 iterations): 0.7796**

We see that about 80% in our test, one random sample or player chosen out of all the first-round players has more games played than the randomly chosen last round player which confirms our above p-value of zero, since a large majority of the time in our simulation we see the first-round player receiving more playing time.

## Conclusion

After running multiple tests on my data, I have come to a number of conclusions. I noticed a relationship that was the inverse of the belief that older players are more mature and therefore more game ready. After running a linear regression, I observed a negative relationship between age and games played meaning young players ending up recieving more playing time than older players. I also found that there was clear evidence that players drafted in the first round on average have recieved much more playing time than those drafted in the last round which would be expected. To further confirm this, I ran a simulation comparing a single randomly chosen player from the first round and comparing their amount of games played with a randomly chosen seventh-round player. Lastly, I looked to test the independence between each NFL team and the college’s that players attended to try to confirm historic occurrences of NFL teams purposely drafting from specific colleges. After running a chi-squared independence test, I came to find this not being true in the 2016 class.

I think it should be noted that there are plenty instances of last round picks becoming more successful later in their careers as well as first round player’s declining after only a few years. If I were to expand on this report, I would include data from earlier draft classes, so I could observe trends and compare players careers after all players in the draft retired. I would also be able to better observe patterns with a larger sample size of players which may change some of the patterns observed in this report.