NFL Draft

Linear Regression Analysis

Samuel Haley

Applied Multivariate Data Analysis

Rowan University

Glassboro, NJ 08028

[haleys57@students.rowan.edu](mailto:haleys57@students.rowan.edu)

# **Abstract**

The pick at which a player is selected in the NFL draft depends upon a number of factors. Data on running backs and wide receivers drafted in the 2000 – 2018 NFL drafts is consolidated and analyzed using linear regression techniques. An assessment of the regression coefficients shows that college rushing attempts per game is the most important factor in determing draft position. For wide receivers the most important factor is 40 yard dash time.

# **Introduction**

Every spring 256 players are selected in the NFL draft. In the lead up to the draft teams expend significant resources scouting and analyzing players to attempt to identify the draft players that will bring the most value to their organizations.

The vast majority of the players selected played division I college football. A player's college football career offers NFL teams a significant amount of information which they can attempt to make predictions about how that player will fair in the NFL.

In addition to the information offered by player's experience playing college football the NFL draft scouting combine, held every February, offers teams additional data with which to differentiate potential draft selections. Players will perform a number of drills the most famous of which is the 40 yard dash.

Prior to and throughout the draft process there is a significant amount of information available both to teams as well as the public about the players eligible for that year's draft. The goal of this paper is to identify the variables that contribute to most significantly to where running backs and wide receivers are selected in the NFL draft.

# **Data**

There are a number of factors that can affect a player's draft selection. In order to get insightful analysis results it was important to collect as many variables as possible for players selected in the NFL draft. The following categories of data where utilized to analyze and answer the overarching research question.

### *Datasets*

* "NFL Draft" Dataset – This dataset was compiled and publicized by Sean J. Taylor. The observations are all of the draft selections from the 2000 through 2018 NFL Draft. It includes NFL scouting combine metrics and individual college statistics aggregated over the player's college carreer.
* "College Football Rankings" Dataset – This dataset was produced by pulling the college football rankings for each year from 1999 – 2017.

### *NFL Draft Results*

The "NFL Draft" dataset included draft results for each player selected. This included the draft pick number (1 – 256) for each player selected in the draft.

### *NFL Scouting Combine Results*

The "NFL Draft" dataset included results from the NFL scouting combine. The data collected at the NFL scouting combine primarily identifies a player's physical traits and capabilities. The following variables were used within this analysis.

* Height – A player's height in inches
* Weight – A player's weight in pounds
* Forty – A player's 40 yard dash time in seconds
* Vertical – A player's vertical leap distance in inches
* Bench – A player's number of reps on the bench press at 225 lbs
* Broad – A player's broad jump distance in inches
* Threecone – A player's threecone drill time in seconds
* Shuttle – A player's shuttle drill time in seconds

### *College Individual Statistics*

The "NFL Draft" dataset included each player's college statistics aggregated over the course of there careers. The following variables related to individual college statistics were used within this analysis.

* Rushing Games – The total number of games played in a rushing position
* Rushing Seasons – The total number of seasons played in a rushing position
* Rushing Plays – The total number of plays played in a rushing position
* Rushing Attempts – The total number of rushing attempts from a rushing position
* Rushing Rush Yards – The total number of yards gained on rushing attempts
* Rushing Rush Touchdowns – The total number of touchdowns scored on rushing attempts
* Rushing Receptions – The total number of receptions from a rushing position
* Rushing Receiving Yards – The total number of yards gained on rushing receptions
* Rushing Receiving Touchdowns – The total number of touchdowns scored on rushing receptions
* Rushing Scrimmage Yards – The total number of yards gained from a rushing position
* Rushing Scrimmage Touchdowns – The total number of touchdowns scored from a rushing position
* Receiving Games – The total number of games played in a receiving position
* Receiving Seasons – The total number of seasons played in a receiving position
* Receiving Plays – The total number of plays played in a receiving position
* Receiving Receptions – The total number of receptions from a receiving position
* Receiving Receiving Yards – The total number of yards gained on receiving receptions
* Receiving Receiving Touchdowns – The total number of touchdowns scored on receiving receptions
* Receiving Scrimmage Yards – The total number of yards gained from a receiving position
* Receiving Scrimmage Touchdowns – The total number of touchdowns scored from a receiving position

### *College Team Statistics*

The "College Football Standings" dataset offered important information about the teams that majority of the players drafted played on. The following variables related to college team statistics were used within this analysis.

* Wins – Total number of wins in the season
* Losses – Ttal number of losses in the season
* Win Percentage – Total wins / total games played for the season
* Conference – The conference the team played in during the season
* Conference Wins – Total number of wins against conference opponents
* Conference Losses – Total number of losses to conference opponents
* Conference Win Percentage – Total wins / total games played against conference opponents
* AP Preseason Rank – The rank of the team in preseason according to the Associated Press
* AP Highest Rank – The highest rank achieved throughout the course of the season according to the Associated Press
* AP Final Rank – The rank of the team at the end of the season according to the Associated Press
* Simple Rating System Score – The simple rating system score for the team. A SRS score is given to all division I teams with higher score indicating better teams. SRS has a mean of zero.
* Strength of Schedule Score – The strength of schedule score for the teams. A SOS score is given to all division I teams with higher scores indicating more difficult schedules. SOS has a mean of zero.

# **Methdology**

### *Overview*

To answer the overarching research question it was decided that a multivariate linear regression was appropriate. Using pick as the dependent variable would produce regression coefficients and statistical significances for each of the independent variables. These coefficients could be compared assessed and compared to one another to provide insight into the relationship between pick and all of the dependent variables.

The following steps were taken to combine, transform and analyze the data and produce the results within this paper.

1. Combine “NFL Draft” dataset with “College Football Standings” dataset from the year prior
2. Transform rank variables and encode categorical variables
3. Group players by position (RBs, WRs)
4. Standardize variables within each position group
5. Perform principal component analysis per position to inform number of factors to utilize in subsequent factor analysis
6. Perform factor analysis per position to identify correlated variables and identify underlying factors
7. Reduce the number of variables
8. Perform linear regression per position to identify variables that contribute most significantly to draft position

### *Issues with Raw Data*

In order to fit a linear regression model and produce regression coefficients and p-values for the independent variable that could be compared to one another it was important not to violate any of the 4 assumptions of linear regression (linearity, homoscedacity, independence and normality). The issues found within the the raw data are listed below.

* Variables had different means and standard deviations (makes comparison of regression coefiicients invalid)
* Rank variables (pick and AP ranks) were uniformily distrubuted (violating the normality assumption)
* Categorical variables (conference) could not be used as independent variables
* High correlation between variables (violating the independence assumption)

### *Data Transformation*

To absolve the linear regression model of these issues it was necessary to transform the raw data in various ways. The following sections discuss some of the important transformations that were performed to alleviate the issues with the suitability of the raw data for linear regression.

**Transforming Ranks**

Both the pick and AP rank variables had the issue of thaving uniform distributions within their values. Since pick was the dependent variable it was especially important that this variable was transformed to have a normal distribution.

The methodology utilized to transform the variables was to calculate the area to the right of the curve and compute the inverse CDF (PPF) of that value using the standard normal distrbution. The equation is shown below:

It should also be noted that this had the effect giving low picks higher values.

**Encoding Categorical Variables**

College conference was provided simply as a string value within the "College Football Standings" dataset. To avoid removing this variable from the analysis performed, and therefore losing important information about the relationship between coference and draft pick, it needed to be transformed into a numerical value.

The methodology utilize to transform conference was "One-Hot Encoding". Each conference that had 50 or more players drafted in the years analyzed was one hot encoded as variable in addition to "Other" for players from small and non division I conferences.

**Transforming Individual Statistics**

Assessing the correlation between the independent variables showed there was significant correlation between some of the individual statistics. Rushing attempts and rushing yards, for example, were highly correlated. High multicolinearity between variables would violate the assumptions of linear regression and throw off the coefficients and p-values of the linear regression model. The desire was to reduce the correlation between the individual statistics without losing all of the information that those variables provided.

I order to reduce correlation without losing the information additional statistics were computed by combining variables. Rushing attempts per game, yards per attempt, touchdowns per attempt, receptions per game, yards per reception and touchdowns per reception variables were created in order to reduce correlation between variables while still capturing efficiency information within the model.

# **Results**

### *Principal Component Analysis*

After the data had been transformed, separated into groups by position, and standardized the variables could be analyzed. The first step in this analysis was to perform a principal component analysis to determine the dimensionality of the variance within all of the independent variables available. The goal of this was simply to inform the number of factors that should be utilized in subsequent factor analysis. The explained variance ratios for each of the components in both the running back and wide receiver position groups is shown below.

|  |  |  |
| --- | --- | --- |
| Component | RB EVR Cumulative | WR EVR Cumulative |
| c1 | 0.24857 | 0.246936 |
| c2 | 0.393652 | 0.430913 |
| c3 | 0.501476 | 0.533977 |
| c4 | 0.585316 | 0.60974 |
| c5 | 0.64412 | 0.677979 |
| c6 | 0.696157 | 0.727188 |
| c7 | 0.738346 | 0.766527 |
| c8 | 0.769736 | 0.803012 |
| c9 | 0.796587 | 0.83163 |
| c10 | 0.822348 | 0.858572 |
| c11 | 0.84688 | 0.882502 |
| c12 | 0.86816 | 0.902146 |
| c13 | 0.887504 |  |
| c14 | 0.9045 |  |

As a result of principal component analysis it was decided that 9 factors should be utilized for both the running back and wide receiver position groups.

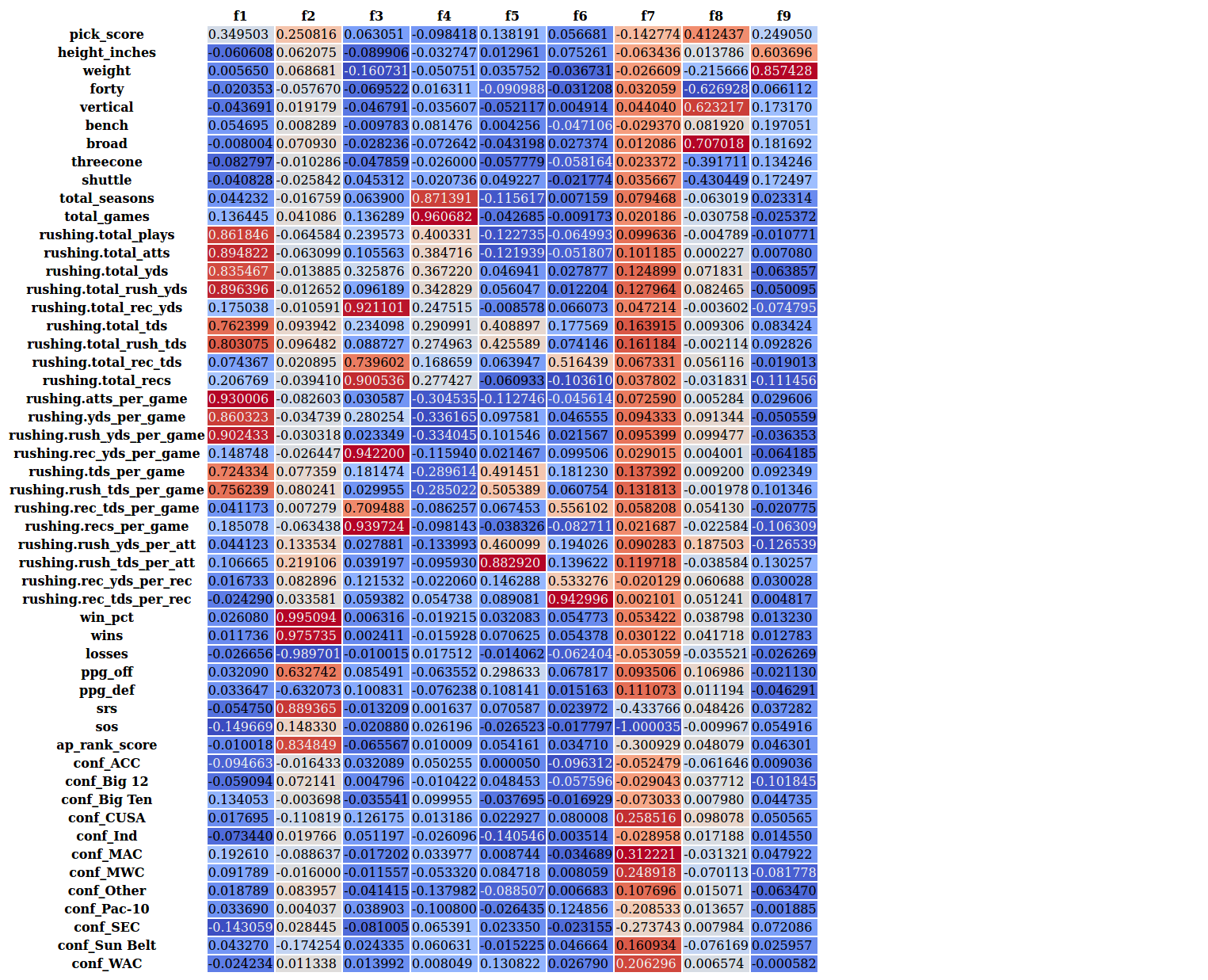
### *Factor Analysis*

For both running backs and wide receivers a factor analysis was performed. The goal of the factor analysis was to:

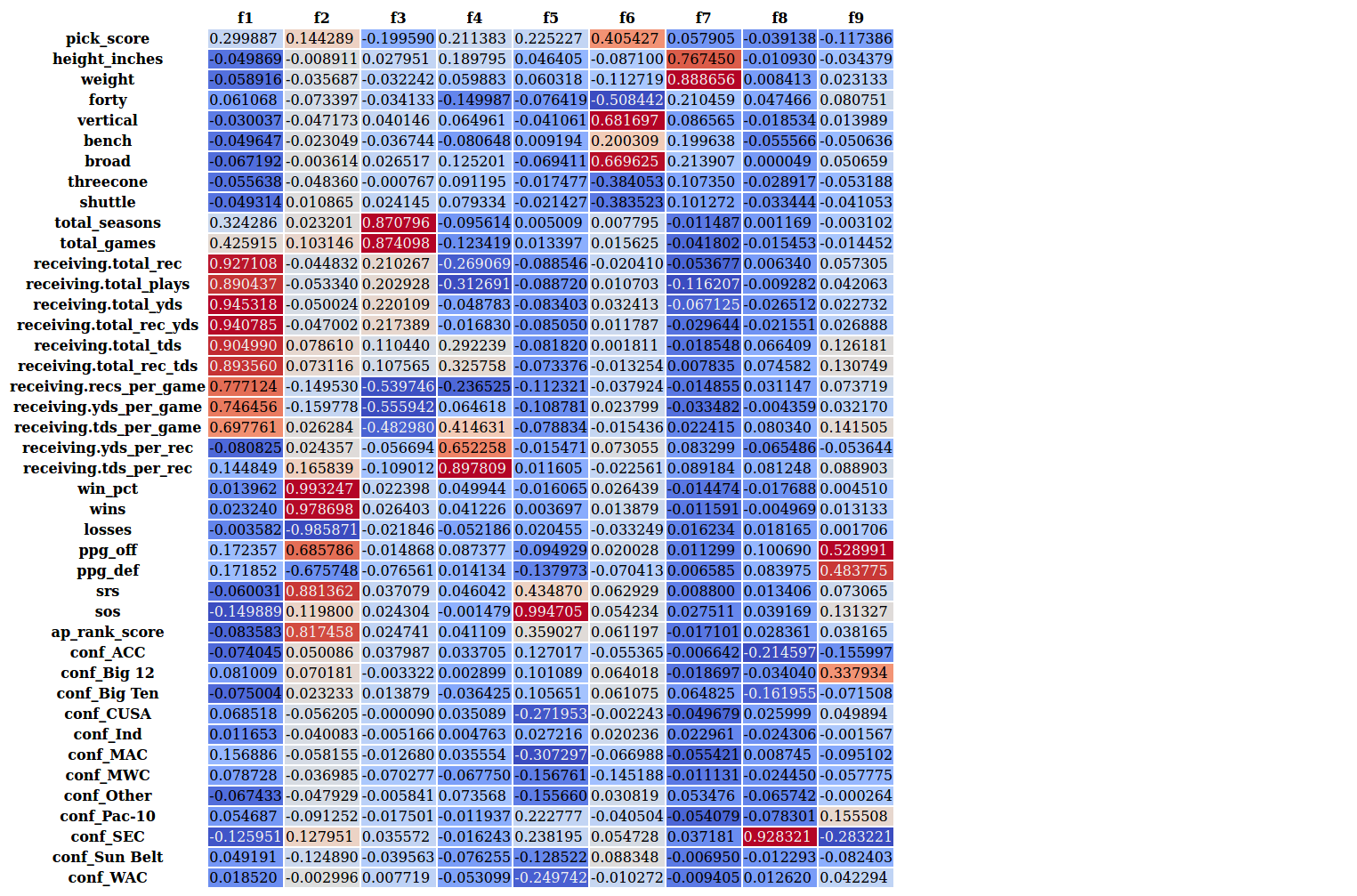
1. Identify and interpret the underlying factors of the independent variables
2. Identify variables that had high correlations with one another

In order to interpret the factors a factor analysis model with 9 factors was fit using all of the variables available. The loadings for each of the variables was extracted from the model and styled using a background gradient to indicate the value of the loading (dark red = high positive loading, dark blue = high negative loading). The results of factor analysis for bothe running backs and wide receivers is shown below.

**Running Back Factor Loadings**



**Wide Receiver Factor Loadings**



**Interpretation of Factor Analysis Results**

Based upon the factor analysis results above each factor was interpreted in terms of what it represents. In addition it was determined whether or not the number of variables within that factor needed to be reduced. The table below summarizes that information.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Running Backs | | Wide Receivers | |
| Factor | Interpretation | Reduce? | Interpretation | Reduce? |
| 1 | College Rushing Production | Y | College Receiving Production | Y |
| 2 | College Team Success | Y | College Team Success | Y |
| 3 | College Receiving Production | Y | College Experience | Y |
| 4 | College Experience | Y | College Receiving Efficiency | N |
| 5 | College Rushing Efficiency | N | Stronger Conferences | N |
| 6 | College Receiving Efficiency | N | Speed / Power | N |
| 7 | Lesser Conferences | N | Size | N |
| 8 | Speed / Power | N | From SEC | N |
| 9 | Size | N | College Team Scoring | N |

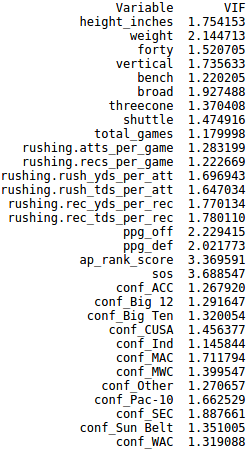
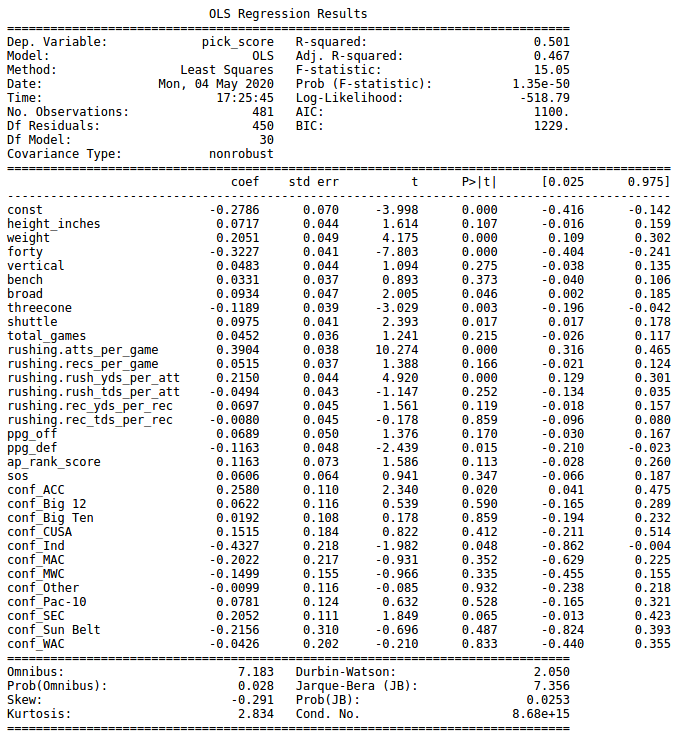
### *Linear Regression Analysis*

After completing the factor analysis more was understood about the correlation of the independent variables. In order to avoid undesirable effects of multicolinearity it was necessary to reduce the number of variables within some of the factors. This was especially true of the "college production" variables and the "team success" variables.

To aide in the determination of what variables should be removed an iterative process was utilized. A set of independent variables would be selected and subsequently analyzed to generate a correlation matrix and linear regression summary which included the computation of variance inflation factors for each variable. The set of independent variables was modified independently for each position group until the best performing linear regression model (based upon R2 value) with all correlation values below 0.8 and variance inflation factors below 5.0 was reached.

Shown below are the results of the linear regression analysis for both running backs and wide receivers.

**Running Backs Linear Regression Results**

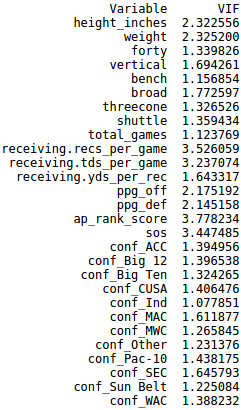
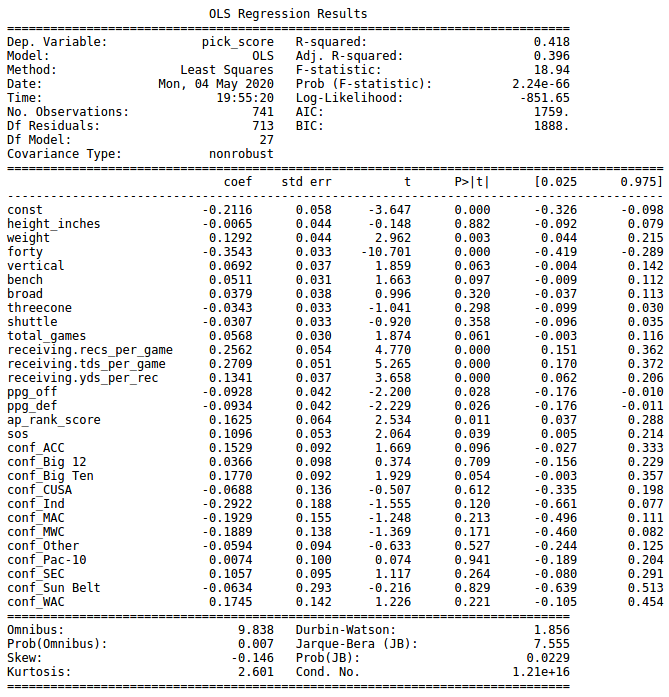


**Findings from Running Back Linear Regression Results**

Rushing attempts per game is the variable with the highest regression coefficient and significance at the 95% level or above which would indicate its the best predictor of draft selection. While this may seem counter intuitive at first its important to remeber from the factor analysis that there is a strong correlation between attempts per game and yards per game. In addition thinking about it from a more qualitative view point running backs who have a lot of attempts in a game typically will typically be dominating that game, running for a lot of yards and picking up a lot of first downs other wise they wouldn't have the oppurtunities.

One of the more surprising results is that receiving production (receptions per game) is not significant. The conventional knowledge is that NFL teams value running backs who are efficient receivers. These results do not seem to indicate that at least at a level thats significant. The other somewhat surprising result is the lack of relationship between rushing touchdowns per attempt and pick. My initial thinking would be that players that score touchdowns at a higher clip are more appealing to NFL teams.

**Wide Receivers Linear Regression Results**



**Findings from Wide Receivers Linear Regression Results**

Forty yard dash time is the variable with the largest regression coefficient and significance at the 95% level of above. This isn't necessarily surprising given the value of speed at the receiver position. Compared to the running back position there seems to be a greater emphasis on touchdowns at the receiver position as evidenced by touchdowns per game being the third highest coefficient.

One of the more surprising results is the lack of relationship between height and pick. Larger receivers would seemingly be more desirable, however the results don't seem to indicate that. Another surprising result is offensive points per game having negative regression coefficient and significance at the 95% level. That would seem to indicate that receivers on teams that score more points get drafted lower. It may point to NFL teams justifying their success as the result of their teammates most notably quarterbacks.

# **Conclusion**

The primary goal of this analysis was to determine what factors contribute most significantly to where running backs and wide receivers are selected in the NFL draft. Based upon the above analysis the answer for running backs is college rushing attampts per game and 40 yard dash time. For wide receivers the answer is 40 yard dash time and touchdowns per game.

In addition to answering the main research question other interesting relationships were discovered:

* College production (attempts / receptions per game) and efficiency (yards per attempt / yards per reception) are important factors for draft selection; college experience (total games) is not
* Team success is an important factor for draft selection although team offensive points per game is not necessarily
* Players from the ACC, SEC and Big Ten conferences are drafted higher than players from other conferences

# **Data Sources**

"NFL Draft" dataset: <https://docs.google.com/spreadsheets/d/1sFcj1W38gR_Bg3BQz8lZSuFrRQr0BDk8K066ctrxrOA/edit#gid=1586587221>

"College Football Standings" dataset:

<https://www.sports-reference.com/cfb/years/2017-standings.html>

# **References**

"Learning the NFL Draft" <https://seanjtaylor.github.io/learning-the-draft>.

"Assumptions of Multiple Linear Regression" <https://www.statisticssolutions.com/assumptions-of-multiple-linear-regression>

"The Race to Make the NFL Draft an Exact Science" <https://www.theringer.com/nfl/2019/4/22/18510099/nfl-draft-analytics-player-tracking>

"Normalization vs. Standardization" <https://towardsdatascience.com/normalization-vs-standardization-quantitative-analysis-a91e8a79cebf>