MSDS 6372 Project 2

Team members:
Killion, Kyle and Taylor, Celia

Date: November 14, 2016

NFL Combine – Analyzing Drafted Wide Receivers 1999-2013

Introduction

The National Football League (NFL) Combine is the yearly gathering of approximately 300 invited top football athletes applying to play for one of the 32 NFL professional football teams. Preceding and during the Combine, coaches and recruiters measure and test each athlete to access their physical attributes and skills. These physical attributes include height, weight, arm length, and hand size. Physical skills include running speed (ten yards, twenty yards, and forty yards), agility (twentyss and threecone), jumping (broad and vertical), and weightlifting (bench). Different physical attributes and skills are recognized as having more importance to one or more of the 22 different football player positions in the NFL draft. Because of the diversity in the original data set, the focus of this analysis is on the wide receiver position to determine which physical attributes and skills in a published combine data set correlate to the actual drafted wide receiver athletes from the 1999 – 2015 combine years.

The original data set contains close to 5000 observations for the 22 football player positions for the 1999-2015 combine years. This data set is statistically a challenge to assess because the data is missing for different measurements, for different years, and for different player positions. A zero or null in almost all cases in this data set means that the coaches did not need or record the value, like the hand measurement, and not that the player could not jump ten inches. A nonzero value in our response variable, picktotal, means that the player was drafted. The lowest nonzero picktotal values indicate the more highly recruited players.

Also, the value and significance of some measurements is quite different for one player position versus another player position. For instance, wide receivers, who run to catch passes, and the cornerbacks, who defend against those passes, need to be very fast. In contrast, linemen, who protect the quarterback for that pass, need to be heavy and strong. Linemen are almost a second slower in the 40 yard dash than wide receivers or cornerbacks. Both are world class athletes, but no wide receiver will ever get drafted into the NFL with a lineman's 40 yard dash time.

Basic Statistics and PCA

Descriptive Statistics

As we ran different groupings of observations through exploratory data analysis and principal components analysis, the zeros, null values, and different measurements for player positions would cancel each other out from showing statistical significance to the NFL draft variable, picktotal, our chosen response variable. These trials resulted in the understanding that the master Combine data set is very complex with many layers or dimensions of information. In order to maintain the most uniformity, homogeneity, and independence possible in this project, a data set of 95 drafted wide receiver player observations over several Combine years from the original master Combine data set was collected. These 95 observations are the most complete wide receiver observations, but these still had some fields with missing values. In data set "a", the same 95 observations' missing values were replaced with zeros. In data set "b", the same 95 observations' missing values are represented with a null. Figure 1a and 1b show the

descriptive statistics of the 95 drafted wide receivers' data set "a" with the zeros versus the 95 drafted wide receivers' data set "b" with the null values.

Variable	N	Mean	Minimum	Maximum	Std Dev	Variance
id	95	2517.86	777.0000000	4893.00	956.4618044	914819.18
year	95	2007.03	1999.00	2013.00	3.4129061	11.6479283
heightinchestotal	95	73.4750000	68.0000000	77.0000000	1.9790794	3.9167553
weight	95	204.3368421	169.0000000	241.0000000	14.0293086	196.8215006
arms	95	21.8039474	0	35.8750000	15.2922185	233.8519457
hands	95	6.3894737	0	10.6250000	4.3880453	19.2549412
fortyyd	95	4.4520000	4.2800000	4.7200000	0.0832019	0.0069226
twentyyd	95	2.5896842	2.4500000	2.7700000	0.0674529	0.0045499
tenyd	95	1.5434737	1.4000000	1.6900000	0.0548454	0.0030080
twentyss	95	3.7316842	0	4.6000000	1.3640995	1.8607673
threecone	95	5.9742105	0	7.3900000	2.3983565	5.7521140
vertical	95	36.5368421	0	42.5000000	4.3480505	18.9055431
broad	95	120.6105263	0	139.0000000	18.5307650	343.3892497
bench	95	6.5263158	0	23.0000000	7.8290812	61.2945129
picktotal	95	98.7473684	2.0000000	252.0000000	69.8818103	4883.47

Figure 1a Descriptive Statistics of Drafted Wide Receivers' Variables

Variable	N	Mean	Minimum	Maximum	Std Dev	Variance
id	95	2517.86	777.0000000	4893.00	956.4618044	914819.18
year	95	2007.03	1999.00	2013.00	3.4129061	11.6479283
heightinchestotal	95	73.4750000	68.0000000	77.0000000	1.9790794	3.9167553
weight	95	204.3368421	169.0000000	241.0000000	14.0293086	196.8215006
arms	64	32.3652344	30.0000000	35.8750000	1.2952724	1.6777305
hands	65	9.3384615	7.5000000	10.6250000	0.5580217	0.3113882
fortyyd	95	4.4520000	4.2800000	4.7200000	0.0832019	0.0069226
twentyyd	95	2.5896842	2.4500000	2.7700000	0.0674529	0.0045499
tenyd	95	1.5434737	1.4000000	1.6900000	0.0548454	0.0030080
twentyss	84	4.2203571	3.9100000	4.6000000	0.1419609	0.0201529
threecone	82	6.9213415	6.3000000	7.3900000	0.1973451	0.0389451
vertical	94	36.9255319	31.0000000	42.5000000	2.1452090	4.6019218
broad	93	123.2043011	114.0000000	139.0000000	5.2741353	27.8165030
bench	42	14.7619048	4.0000000	23.0000000	3.9988384	15.9907085
picktotal	95	98.7473684	2.0000000	252.0000000	69.8818103	4883.47

Figure 1b Descriptive Statistics of Drafted Wide Receivers' Variables

The minimum in "Figure 1a Descriptive Statistics of Drafted Wide Receivers' Variables" is zero for several variables. The minimum is not zero in "Figure 1b Descriptive Statistics of Drafted Wide Receivers' Variables". The "N" in "Figure 1b" represents the total population and shows that the variable "bench" only has 42 of the 95 total possible nonzero or non-null values. Variables, "arms", "hands", "twentyss", and "threecone", as well as "bench" are not fully populated to the possible 95 values.

Figure 2 shows the initial normality of the selected drafted wide receiver data set with zero values (data set "a") versus the null values (data set "b").

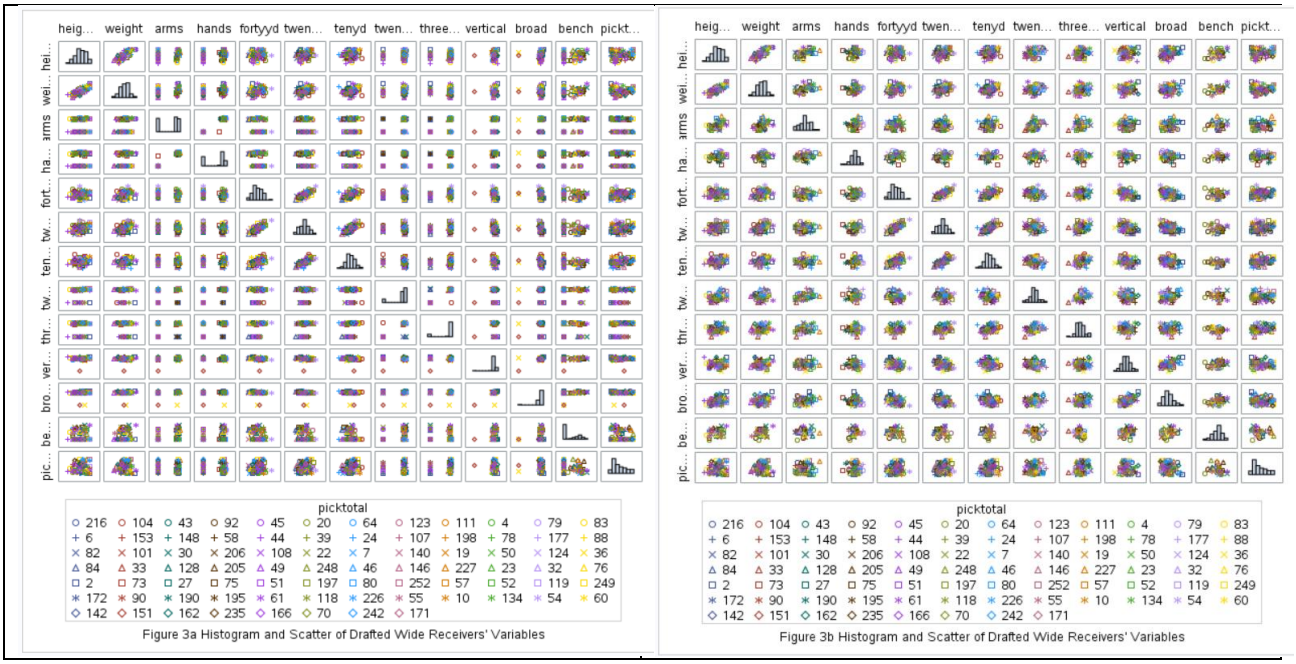
Original variables with zeros (data set a)	Shapiro-Wilk (data set a)	Skewness (data set a)	Original variables with nulls (data set b)	Shapiro-Wilk (data set b)	Skewness (data set b)
heightinchestotal	0.95	- 0.32	heightinchestotal	0.95	-0.32
weight	0.65	-0.02	weight	0.99	-0.02
arms	0.66	-0.74	arms	0.98	0.19
hands	0.98	-0.77	hands	0.97	-0.37
fortyyd	0.98	0.25	fortyyd	0.98	0.26
twentyyd	0.99	0.13	twentyyd	0.98	0.13
tenyd	0.46	0.04	tenyd	0.99	0.04
twentyss	0.48	-2.39	twentyss	0.98	0.18
threecone	0.50	-2.12	threecone	0.98	0.09
vertical	0.50	-6.35	vertical	0.98	0.29
broad	0.36	-5.94	broad	0.97	0.55
bench	0.75	0.54	bench	0.97	-0.36
picktotal	0.92	0.66	picktotal	0.92	0.66

Figure 2 Initial Data Set Normality Test Results

Shapiro-Wilk preferred values for normality range from 0.95 to 1.00. These normality tests highlight the difficulty of using the data set with the zero values. The zero values in data set "a" are not a test result of "zero", but represent a "player did not participate" or "coach did not require" or "coach did not record" value. The corresponding null values in data set "b" are really "unrecorded" or "unneeded" values from the NFL Combine business viewpoint.

"Figure 3a Histogram and Scatter of Drafted Wide Receivers' Variables" shows a graphical representation of the effect of zeros on normality tests. Contrast this to "Figure 3b Histogram

and Scatter of Drafted Wide Receivers' Variables" of the data set "b" with null values to see the major effect the zeros have on the overall data normality.



Processing the data set "a" with the zero values to a "statistical normal" required much effort with many different extreme manual transforms (like taking variables to the 9th power) and unsuccessful "Box Cox" transform trials. Because statistical software interprets the zeros as a zero value test result and not a "player did not participate" value, the remaining part of this paper will be executed with the null value data set "b". As previously seen in Figure 2, the actual recorded data in data set "b" is "statistically normal" or requires very minor transformation. The resulting transformed data set "b" is shown in all subsequent procedures. All variables in the transformed data set "b" besides the response variable, picktotal, have a Shapiro-Wilk value of > 0.95. Some of these procedures eliminate any observations that have any null values. There are 33 observations that have no null values.

Within the NFL Combine, there are certain variables that correspond to one another as indicated by the Pearson Correlation Coefficients in "Figure 4 – Drafted Wide Receivers' Correlation of Transformed Variables". As one would expect, the variables for height, weight, arms, and hands show strong relationships with each other due to the nature of the human body. The forty yard dash with its interim split times, the ten yard split and twenty yard split, also show high correlation. This make perfect sense because the faster a player is in the beginning of the forty yard dash due to faster ten and twenty yard splits then the more likely the overall forty yard dash time will be faster. The vertical and broad jump carry a strong, positive correlation between each other as they are a test of an athlete's explosive jumping abilities. The variables for the remaining two tests, three cone drill and twenty yard short shuttle, have a positive correlation, especially among wide receivers. These drills showcase a wide receiver's ability to run crisp and effective routes and to synchronize the timing between Quarterbacks and Wide Receivers.

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	heightinchestotal	weight	tfarms	tfhands	fortyyd	twentyyd	tenyd	tftwentyss	tfthreecone	tfvertical	tfbroad	tfbench	picktotal
heightinchestotal	1.00000 95	0.78716 <.0001 95	-0.58737 <.0001 64	0.19298 0.1235 65	0.15334 0.1379 95	0.25903 0.0113 95	0.12050 0.2447 95	-0.30987 0.0041 84	-0.15636 0.1607 82	-0.00747 0.9430 94	-0.25454 0.0138 93	0.28650 0.0658 42	0.11250 0.2777 95
weight	0.78716 <.0001 95	1.00000 95	-0.50286 <.0001 64	0.21230 0.0895 65	0.28568 0.0050 95	0.35974 0.0003 95	0.15649 0.1299 95	-0.23925 0.0284 84	-0.18056 0.1045 82	-0.01152 0.9123 94	-0.17500 0.0934 93	0.35612 0.0206 42	0.03152 0.7617 95
tfarms	-0.58737 <.0001 64	-0.50286 <.0001 64	1.00000 64	-0.28113 0.0244 64	-0.07068 0.5789 64	-0.07996 0.5299 64	-0.06875 0.5894 64	0.43130 0.0006 60	0.26698 0.0428 58	0.03050 0.8109 64	0.21608 0.0890 63	-0.07748 0.6485 37	-0.11511 0.3651 64
tfhands	0.19298 0.1235 65	0.21230 0.0895 65	-0.28113 0.0244 64	1.00000 65	0.13843 0.2715 65	0.11129 0.3775 65	-0.02490 0.8439 65	-0.09195 0.4810 61	0.00810 0.9514 59	0.12177 0.3339 65	0.05077 0.6903 64	0.17226 0.3080 37	0.00828 0.9478 65
fortyyd	0.15334 0.1379 95	0.28568 0.0050 95	-0.07068 0.5789 64	0.13843 0.2715 65	1.00000 95	0.75303 <.0001 95	0.51764 <.0001 95	0.03668 0.7405 84	-0.23113 0.0367 82	0.40467 <.0001 94	0.38904 0.0001 93	-0.12135 0.4440 42	0.18606 0.0710 95
twentyyd	0.25903 0.0113 95	0.35974 0.0003 95	-0.07996 0.5299 64	0.11129 0.3775 65	0.75303 <.0001 95	1.00000 95	0.69505 <.0001 95	0.10088 0.3612 84	-0.27950 0.0110 82	0.33082 0.0011 94	0.34145 0.0008 93	-0.11405 0.4720 42	0.21827 0.0336 95
tenyd	0.12050 0.2447 95	0.15649 0.1299 95	-0.06875 0.5894 64	-0.02490 0.8439 65	0.51764 <.0001 95	0.69505 <.0001 95	1.00000 95	0.14827 0.1783 84	-0.27158 0.0136 82	0.20890 0.0433 94	0.26687 0.0097 93	-0.04162 0.7935 42	0.25620 0.0122 95
tftwentyss	-0.30987 0.0041 84	-0.23925 0.0284 84	0.43130 0.0006 60	-0.09195 0.4810 61	0.03668 0.7405 84	0.10088 0.3612 84	0.14827 0.1783 84	1.00000 84	0.37704 0.0005 81	-0.19720 0.0722 84	0.10032 0.3669 83	0.11645 0.4802 39	0.00357 0.9743 84
tfthreecone	-0.15636 0.1607 82	-0.18056 0.1045 82	0.26698 0.0428 58	0.00810 0.9514 59	-0.23113 0.0367 82	-0.27950 0.0110 82	-0.27158 0.0136 82	0.37704 0.0005 81	1.00000 82	-0.04545 0.6851 82	-0.03159 0.7795 81	0.23393 0.1697 36	-0.27452 0.0126 82
tfvertical	-0.00747 0.9430 94	-0.01152 0.9123 94	0.03050 0.8109 64	0.12177 0.3339 65	0.40467 <.0001 94	0.33082 0.0011 94	0.20890 0.0433 94	-0.19720 0.0722 84	-0.04545 0.6851 82	1.00000 94	0.51059 <.0001 93	-0.22792 0.1466 42	0.12791 0.2192 94
tfbroad	-0.25454 0.0138 93	-0.17500 0.0934 93	0.21608 0.0890 63	0.05077 0.6903 64	0.38904 0.0001 93	0.34145 0.0008 93	0.26687 0.0097 93	0.10032 0.3669 83	-0.03159 0.7795 81	0.51059 <.0001 93	1.00000 93	-0.31642 0.0412 42	0.18175 0.0812 93
tfbench	0.28650 0.0658 42	0.35612 0.0206 42	-0.07748 0.6485 37	0.17226 0.3080 37	-0.12135 0.4440 42	-0.11405 0.4720 42	-0.04162 0.7935 42	0.11645 0.4802 39	0.23393 0.1697 36	-0.22792 0.1466 42	-0.31642 0.0412 42	1.00000 42	-0.21209 0.1775 42
picktotal	0.11250 0.2777 95	0.03152 0.7617 95	-0.11511 0.3651 64	0.00828 0.9478 65	0.18606 0.0710 95	0.21827 0.0336 95	0.25620 0.0122 95	0.00357 0.9743 84	-0.27452 0.0126 82	0.12791 0.2192 94	0.18175 0.0812 93	-0.21209 0.1775 42	1.00000 95

Figure 4 - Drafted Wide Receivers' Correlation of Transformed Variables

Principal Component Analysis

Because the NFL Combine produces so many metrics and resulting variables of physical attributes and skills, a Principal Component Analysis (PCA) is a prime procedure to determine which variables are most important for an upcoming athlete entering the Combine. The focus of this variable analysis is on wide receivers to discern which variables are the most important to be drafted into the NFL.

The goal of the PCA is to filter the list of 13 variables to find the significance between the variables and hopefully reduce the number of variables into a smaller number of principal components. The procedures, PRINCOMP, PLS, and REG, process the data set to determine the component selection to build the regression model.

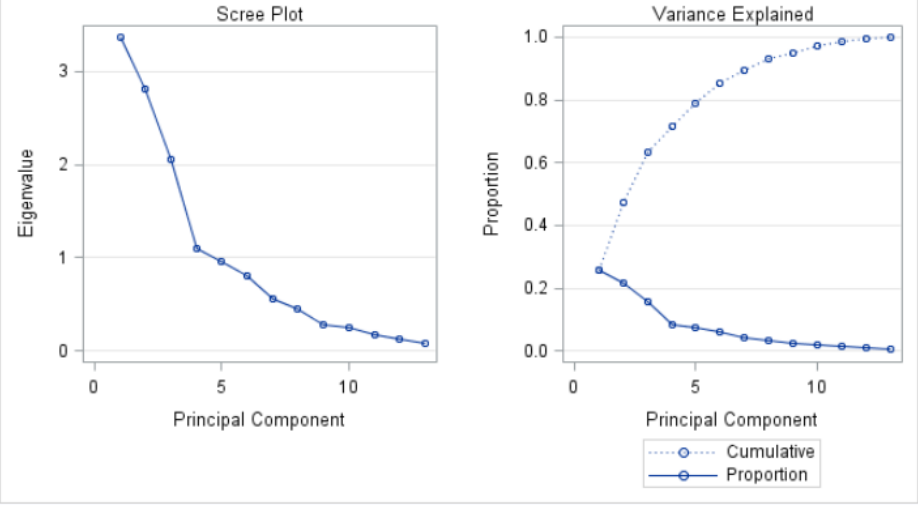


Figure 5 Principal Component vs Eigenvalue Scree Plot and Principal Component vs Proportion of Variance Explained

"Figure 5 Principal Component vs Eigenvalue Scree Plot and Principal Component vs Proportion of Variance Explained" shows the relationships between the principal components and the Eigenvalues and variance. As the graphic depicts, the principal components clearly account for substantial variance which levels off right at the 5th component. This leveling shows a definitive elbow within the Scree Plot at the 4th principal component. "Figure 6 Percent Variation Accounted for by Principal Components" gives further corroboration of the variance of the principal components.

Percent Variation Accounted for by Principal Components

Number of Extracted Factors	Model Effects		Dependent Variables	
	Current	Total	Current	Total
1	27.6053	27.6053	3.2010	3.2010
2	22.5915	50.1968	6.6033	9.8044
3	16.9998	67.1966	0.7419	10.5463
4	7.9880	75.1846	0.3953	10.9416
5	7.7418	82.9264	2.7457	13.6873
6	4.9280	87.8544	2.7043	16.3916

Figure 6 Percent Variation Accounted for by Principal Components

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
heightinchestotal	-.223183	0.477461	0.157624	-.197315	0.028029	-.043926
weight	-.146782	0.455656	0.241647	-.233146	-.201328	-.040712
tfarms	0.173317	-.384915	0.138703	0.209874	-.416531	0.111070
tfhands	0.002735	0.302613	0.174340	0.533891	0.355549	0.281249
fortyyd	0.404021	0.038016	0.319726	-.253250	-.011303	-.247091
twentyyd	0.386500	0.154291	0.326897	-.139378	0.120841	-.223278
tenyd	0.380644	0.099889	0.213360	0.062178	-.428336	0.079212
tftwentyss	-.029193	-.343849	0.486432	0.088261	0.094038	0.017412
tfthreecone	-.145238	-.245545	0.436076	0.012027	0.505070	-.105407
tfvertical	0.382739	0.221177	-.123039	0.055423	0.215980	0.328962
tfbroad	0.440650	-.013233	-.146327	-.120810	0.225500	0.391382
tfbench	-.232523	0.133343	0.375397	0.209044	-.309466	0.459620
picktotal	0.156266	0.209872	-.090930	0.648764	-.057972	-.551120

Figure 7 Variable Analysis within Principal Components

"Figure 7 Variable Analysis within Principal Components" itemizes which variables each principal component and associated Eigenvector emphasizes. For Prin1, the variables within its Eigenvector present a combination of higher relationships with the forty yard dash and the ten yard and twenty yard split times along with the explosive nature of the vertical and broad jump

tests for the athletes. The three running speed variables in conjunction with the vertical and broad jump seems to explain a good portion of the variance. Because the main factors in Prin1 are the speed and jumping variables, these are the most influential in a lower picktotal. The business world of the NFL Combine heavily emphasized these variables for a wide receiver.

Prin2 is the second most important component in explaining the variance in the data set. Prin2 seems to possess a combination of the body and the route running ability of the athlete. Within the Combine business context, the physical nature of the body has positive height and weight correlations. A tall, big body, and big hands are important and play a key role in success for these athletes as they compete for passes thrown in the air. The faster times of the twenty yard short shuttle and three cone drills also contribute to earlier draft pick selection.

Prin3 is clearly taking into account the twenty yard short shuttle and three cone drills as the third most important variance dimension. This Prin3 does also have emphasis on some of the same variables as Prin1 and Prin2. These twenty yard short shuttle and three cone drill variables have increased times in contrast to a slightly negative correlation with picktotal. In this principal component, the balancing factors are the speed and strength variables.

Prin4 delineates the relationship of the wide receiver's hands and a faster forty time. The prominence of these variables highlights what coaches really desire in a wide receiver athlete. This fourth Eigenvector has the most variance in regards to the hands and picktotal. Both variables have a positive variance.

Prin5 revealing another combination of arms, ten yard split, and three cone. This seems somewhat offsetting with a large positive correlation in threecone having increasing times yet suggesting earlier picks. This possibly can be explained with the positive broad and vertical jumps integrated here along with a strong negative ten yard split in the forty suggesting a lower time. Overall, Prin5 could be favoring explosiveness in an athlete's agility.

Prin6 again focuses on the explosive nature within Wide Receivers with positive correlations of vertical and broad jumps. Joining the emphasis on the jumping ability is more repetitions within the 225 bench test. This strength test allows coaches to judge how well wide receivers will separate off from cornerbacks in the initial jams at the line of scrimmage.

Multiple Regression

Through principal components analysis, the list of 13 variables was cut down to 6 principal components and the response variable, picktotal. Proc Reg processed the six principal components into a partial least squares regression . Figure 8 and Figure 9 show the results.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	115822	19304	56.83	<.0001
Error	26	8831.31624	339.66601		
Corrected Total	32	124653			

Root MSE	18.43003	R-Square	0.9292
Dependent Mean	92.66667	Adj R-Sq	0.9128
Coeff Var	19.88852		

Figure 8 The Analysis of Variance Table for the Multiple Regression of the Principal Components and R^2 Table.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	92.66667	3.20826	28.88	<.0001
Prin1	1	9.75305	1.77621	5.49	<.0001
Prin2	1	13.09879	1.94607	6.73	<.0001
Prin3	1	-5.67525	2.27480	-2.49	0.0193
Prin4	1	40.49146	3.11401	13.00	<.0001
Prin5	1	-3.61825	3.33436	-1.09	0.2878
Prin6	1	-34.39720	3.64500	-9.44	<.0001

Figure 9 Parameter Estimates of the Principal Components in Regression Model

The model equation is as follows:

$$picktotal = \beta_0 + \beta_1\, Prin1 + \beta_2\, Prin2 - \beta_3\, Prin3 + \beta_4\, Prin - \beta_5\, Prin5 + \beta_6\, Prin6$$

$$picktotal = 92.67 + 9.75\, Prin1 + 13.10\, Prin2 - 5.68\, Prin3 + 40.49\, Prin - 3.62\, Prin5 - 34.40\, Prin6$$

The intercept 92.67 is representative of the mean picktotal for wide receivers within the dataset, which is then calculated with our principal components and their coefficients. The Analysis of Variance contains an F-value of 56.83 and p-value < .0001. This conveys that the model does indeed rationalize and explain variance among these athletes.

Not all of our Principal Components carried a statistically significant p-value < 0.05. Prin5's p-value = 0.29 was not statistically significant. Going back to Prin5 within the Eigenvectors, we see that this combination of variables suggested most of the variance in threecone, ten yard, and arms. Both Prin4 (40.49) and Prin6 (-34.40) inherited substantial coefficients in the regression model.

The model ended up with a R^2 value of an astonishing 0.93 which is a substantial amount of variance explained within the model. The difference between R^2 and adjusted R^2 (0.91) is 0.0164 which translates that the model also is highly effective with each component contributing to correlating with picktotal.

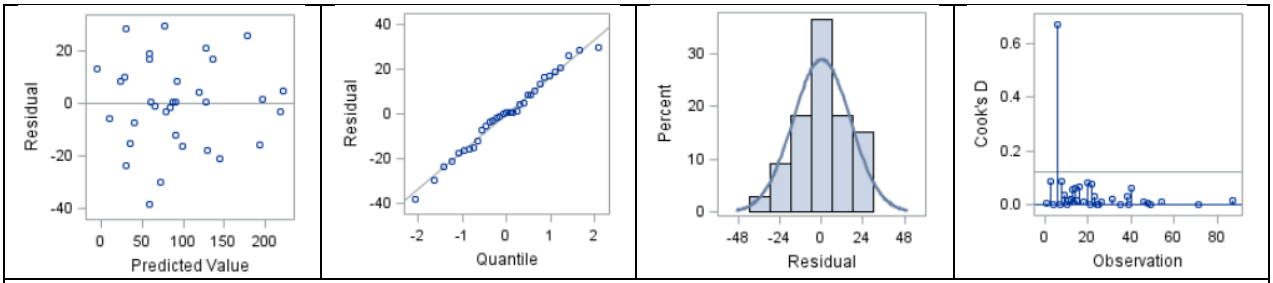


Figure 10 Principal Components Regression Model Residuals and Cook's D Analysis Graph

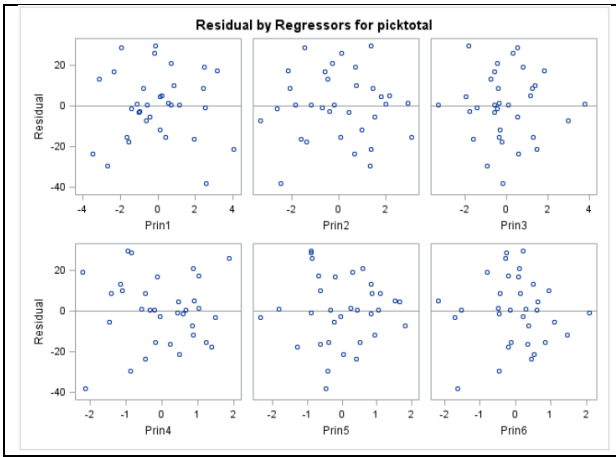


Figure 11 Principal Components' Residual by Regressors for picktotal.

Within the Regression model and its residuals, as noted in Figure10 and Figure11, we observe how the 33 athletes are analyzed employing the six Principle Components. The residuals retain a healthy balanced cluster around zero and within a range approximately (+/-) 40 in regarding picktotal. The QQ plot and histogram indicate much the same assimilating into a balance of normality. We did encounter an abnormality within the Cook's D early in the observations, but could tolerate and considered non-influential maintaining a value under 0.12 with 33 athletes.

Conclusions

The analysis conducted is full of fruitful information which is inherited by these athletes playing throughout the years. They work furiously to better themselves for opportunities to perform at the highest level of competition that exists today. Athletes know what translates to the field of competition and focus on these areas.

We felt as though we peered into a Coach's mind with each different Principle Component dimension. The sluggish forty yard times and unimpressive vertical and broad jumps seem to speak magnitudes for considering a wide receiver in the NFL. The following Prin2, which was orthogonal to Prin1, then speaks to the wide receiver's tests over their threecone drill and twenty yard short shuttle. These tests translated into the abilities to run crisp and effective routes, which is vital in the NFL. The last of the top three Principle Components again explained an ample amount of variance from mostly a positive scope over the variables. This Prin3 conducted a negative coefficient within the regression projecting the player to go in earlier picks. The final regression component is worth noting as the largest of the negative coefficients suggesting that focusing on performing very well on vertical, broad, and bench press to increase your success in getting drafted in earlier picks.

In conclusion, the athlete's overall performances within the components are valuable but some more than others. An athlete competing for a wide receiver position in the draft needs to focus mainly on his forty yard dash, threecone, twenty yard short shuttle, vertical, and broad jump to thoroughly impress scouts among professional teams within the NFL.

References

1. https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_princomp_sect021.htm
2. *http://www.nflsavant.com/index.php*

Appendix

```
data NFLDataData2WR;
/*
Citrix access
infile '\\Client\C$\Users\hb13316\Documents\Data Science\Experimental Stats 2\Project
2\NFL_PreliminarySet.csv' firstobs=2 dlm = ',' DSD;
input ID year
heightinchestotal weight arms hands fortyyd twentyyd tenyd twentyss threecone
vertical broad bench picktotal;
run;
WRDataData2.csv input has zeros in all absent values.
NFL_PreliminarySet.csv has null in all absent values.
*/
infile '/folders/myfolders/SASDATA/NFL_PreliminarySet.csv' firstobs=2 dlm = ',' DSD;
input id year heightinchestotal weight arms hands fortyyd twentyyd tenyd twentyss threecone
vertical broad bench picktotal;
run;
title "Show all of the Clean 95 Drafted Wide Receiver players";
proc print data=NFLDataData2WR;
footnote "Display - Show all of the Clean 95 Drafted Wide Receiver players";
run;

proc means data=NFLDataData2WR n mean min max std var;
Footnote "Figure 1b Descriptive Statistics of Drafted Wide Receivers' Variables";
run;

title "NFL Combine - Scatter and Histogram";
proc sgscatter data=NFLDataData2WR;
matrix heightinchestotal weight arms hands fortyyd twentyyd tenyd twentyss
    threecone vertical broad bench picktotal / diagonal=(histogram) group=picktotal;
Footnote "Figure 3b Histogram and Scatter of Drafted Wide Receivers' Variables";
run;

title "Univariate Picktotal Clean 95 Wide Receivers' Variables";
proc univariate data=NFLDataData2WR plots normal;
var heightinchestotal weight arms hands fortyyd twentyyd tenyd twentyss
```

```
   threecone vertical broad bench picktotal;
footnote "Display - Univariate Picktotal Clean 95 Wide Receivers' Variables";

proc corr data=NFLDataData2WR nosimple PEARSON PLOTS = (MATRIX SCATTER);
var heightinchestotal weight arms hands fortyyd twentyyd tenyd twentyss
   threecone vertical broad bench picktotal;
footnote "Display - Drafted Wide Receivers' Correlation of Variables and Scatter Matrix";
run;

proc corr data=NFLDataData2WR nosimple PLOTS = SCATTER;
var heightinchestotal weight arms hands    fortyyd    twentyyd tenyd twentyss
   threecone vertical broad bench picktotal;
footnote "Display - Correlation picktotal - Drafted Wide Receiver PLOTS = SCATTER";
run;
/*
Manual Transforms for WRDataData2.csv
data NFLDataData2WR;
set NFLDataData2WR;
tfarms = arms*arms*arms*arms*arms*arms*arms;
tfhands = hands*hands*hands*hands*hands*hands;
tftwentyss = twentyss*twentyss*twentyss*twentyss*twentyss*twentyss*
          twentyss*twentyss*twentyss*twentyss;
tfthreecone = threecone*threecone*threecone*threecone*threecone*threecone*
           threecone*threecone*threecone*threecone;
tfvertical = vertical*vertical*vertical*vertical*vertical*vertical;
tfbroad = broad*broad*broad*broad*broad*broad;
tfbench = bench;
run;
*/
/*
Manual transforms for NFL_PreliminarySet.csv
*/
data NFLDataData2WR;
set NFLDataData2WR;
tfarms = arms**-1/arms;
tfhands = log(hands);
tftwentyss = twentyss**-1/twentyss**2;
tfthreecone = threecone**-1/threecone;
tfvertical = vertical**-1/vertical;
tfbroad = broad**-1/broad;
tfbench = bench**1/6;
footnote " Display - Manual Transforms";
run;

title  "Display - Proc Univariate - Original and Manual Transformed Variables";
proc univariate data=NFLDataData2WR plots normal;
var heightinchestotal weight tfarms tfhands fortyyd twentyyd tenyd tftwentyss
      tfthreecone tfvertical tfbroad tfbench picktotal;
```

```
footnote "Display - Proc Univariate - Original and Manual Transformed Variables";
run;

title "Descriptive Statistics Drafted Wide Receivers";
proc means data=NFLDataData2WR n mean min max std var;
FOOTNOTE "Display - WR Stats with Transformation";
run;

title "NFL Final Combine Drafted Wide Receivers Histogram and Scatter";
proc sgscatter data=NFLDataData2WR;
matrix heightinchestotal weight tfarms tfhands fortyyd twentyyd tenyd tftwentyss
    tfthreecone tfvertical tfbroad tfbench picktotal  / diagonal=(histogram) group=picktotal;
FOOTNOTE "Display - NFL Final Combine Drafted Wide Receivers Histogram and Scatter";
run;

proc corr data=NFLDataData2WR nosimple PEARSON PLOTS = (MATRIX SCATTER);
var heightinchestotal weight tfarms tfhands fortyyd twentyyd tenyd tftwentyss
    tfthreecone tfvertical tfbroad tfbench picktotal;
footnote "Figure 4 - Drafted Wide Receivers' Correlation of Transformed Variables";
run;

title "PCA with NFL Final Combine Drafted Wide Receivers";
proc princomp data=NFLDataData2WR out=NFLDataData2WRP2
plots(ncomp=6)=all n=6;
var heightinchestotal weight tfarms tfhands fortyyd twentyyd tenyd tftwentyss
tfthreecone tfvertical tfbroad tfbench picktotal;
FOOTNOTE "Display - Principal Component Analysis";
run;
/*
proc print data=NFLDataData2WRP2;
run;
*/
title "NFL Final Combine Drafted Wide Receivers: PCR";
proc pls data=NFLDataData2WR method=PCR cv=one cvtest (stat=PRESS);
model picktotal = heightinchestotal weight tfarms tfhands fortyyd twentyyd tenyd tftwentyss
tfthreecone tfvertical tfbroad tfbench;
FOOTNOTE "Display - Principal Component Regression";
run;

title "NFL Final Combine Drafted Wide Receivers: PCR with 5";
proc pls data=NFLDataData2WR method=PCR nfac=6;
model picktotal = heightinchestotal weight tfarms tfhands fortyyd twentyyd tenyd tftwentyss
tfthreecone tfvertical tfbroad tfbench;
FOOTNOTE "Display - Principal Component Regression 6 nfac utilized";
run;

title "NFL Final Combine Drafted Wide Receiver Regression Analysis with PCA results";
proc reg data= NFLDataData2WRP2;
```

```
model picktotal = Prin1 Prin2 Prin3 Prin4 Prin5 Prin6 / CLI CLM CLB;
FOOTNOTE "Display - Prin1 - Prin6";
run;

title "Proc Reg data=NFLDataData2WRP2 corr plots(label)=(RSTUDENTBYLEVERAGE cooksd)";
proc reg data=NFLDataData2WRP2 corr plots(label)=(RSTUDENTBYLEVERAGE cooksd);
model picktotal = Prin1 Prin2 Prin3 Prin4 Prin5 Prin6;
FOOTNOTE "Display - Regression with the 6 Principals utilized";
run;

title "Correlations Analysis: NFL Final Combine Drafted Wide Receivers";
proc corr data=NFLDataData2WR PEARSON nosimple PLOTS = (MATRIX SCATTER);
var heightinchestotal weight tfarms tfhands fortyyd twentyyd tenyd tftwentyss
tfthreecone tfvertical tfbroad tfbench picktotal;
FOOTNOTE "Display - Correlation Analysis";
run;
quit;
```