MSDS 6372 Project 3

Team members:
Killion, Kyle and Taylor, Celia

Date: November 28, 2016

Bone Density – Osteoporosis in Women ages 55 and above

<u>Introduction</u>

<u>Basic Statistics</u>

Descriptive Statistics

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.

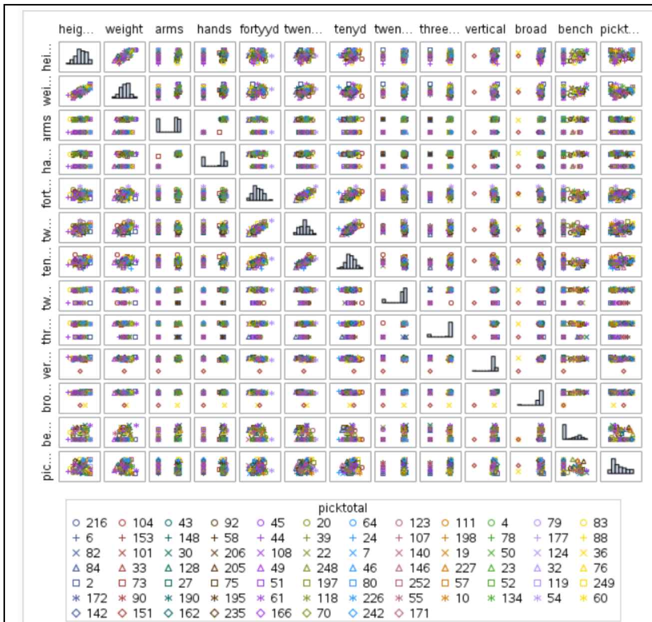

Figure 3a Histogram and Scatter of Drafted Wide Receivers' Variables

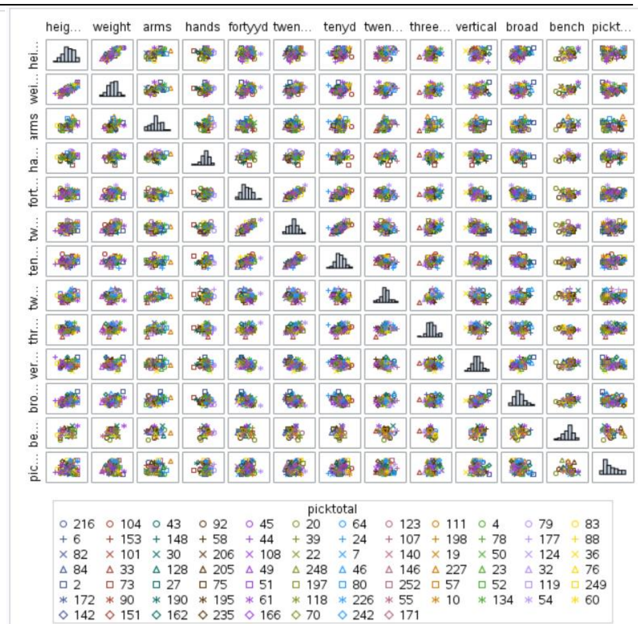

Figure 3b Histogram and Scatter of Drafted Wide Receivers' Variables

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.

Logistic Regression Analysis

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 5^{th} component. This leveling shows a definitive elbow within the Scree Plot at the 4^{th} principal component. "Figure 6 Percent

Variation Accounted for by Principal Components" gives further corroboration of the variance of the principal components.

Figure 6 Percent Variation Accounted for by Principal Components	Figure 7 Variable Analysis within Principal Components

Second Logistic Regression Analysis

.

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

Conclusions

References

Appendix

```
PROC IMPORT OUT= bones

        DATAFILE= '\\Client\C$\Users\hb13316\Documents\Data
Science\Experimental Stats 2\Project 3\glow500.xls'

        DBMS=XLS REPLACE;

   SHEET="GLOW500.TAB.XLS";

   GETNAMES=YES;
```

```
numObs = _N_;

RUN;
```

```
title 'Descriptive Stats - Bone Density';

proc means data=bones;

footnote 'Figure 1';

run;
```

```
title "Bone Density - Scatter and Histogram";

proc sgscatter data=bones;

matrix SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT	BMI
	PREMENO	MOMFRAC	ARMASSIST SMOKE	RATERISK
FRACSCORE /

diagonal=(histogram)

group=FRACTURE;

footnote "Figure 2";

run;
```

```
title "Univariate of Bone Density Variables";

proc univariate data=bones plots normal;

var SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT	BMI
	PREMENO	MOMFRAC	ARMASSIST SMOKE	RATERISK
FRACSCORE FRACTURE;

footnote "Figure 3";
```

```
run;
```

```
title 'Pearson Correlation Analysis';

proc corr PEARSON data=bones;

var SUB_ID SITE_ID PHY_ID PRIORFRAC AGE WEIGHT HEIGHT BMI

PREMENO MOMFRAC ARMASSIST SMOKE RATERISK FRACSCORE FRACTURE;

footnote 'Figure 4';

run;
```

```
/*Reduce variables manually first*/

title 'Regression for Variance Inflation and Lack of Fit Analysis';

proc reg data=bones;

model FRACTURE = PRIORFRAC AGE

HEIGHT    BMI    PREMENO    MOMFRAC    ARMASSIST SMOKE
    RATERISK/ lackfit VIF;

footnote 'Figure 5';

run;

#Options – VIF (Variance Inflation Factor) for multicollinearity (correlation
between predictors) and lackfit for gauging the fitted model
```

```
/*Determine autoselected variables to use for the Logistic Regression*/

title 'Determine autoselected variables to use for the Logistic Regression';

proc glmselect data=bones;

model FRACTURE = PRIORFRAC AGE
```

```
HEIGHT BMI PREMENO MOMFRAC ARMASSIST SMOKE
    RATERISK /

selection = stepwise (choose = cv stop = aic) cvmethod = random(5) stats = (adjrsq
cp bic sbc sl);

footnote 'figure 6';

run;

#Options –

Variable selection criteria = LASSO

Variable stop criteria = AIC

Break up data = Cross Validation / Method = random in 5 parts
```

- *Adjusted R2 = linear correlation*
- *cp = Mallows C(p)*
- *sl = significance level of the F-Stat for entering and exiting effects*
- *sbc = Schwarz Bayesian data measure*
- *bic = Sawa Bayesian data measure*

```
/*

From LAR, LASSO, and STEPWISE ->

All three yielded :

PRIORFRAC AGE RATERISK

These will be utilized for LOGISTIC REGRESSION

*/

title 'Logistic Regression - PRIORFRAC AGE RATERISK';

proc logistic data=bones outest= fracAll;
```

```
model FRACTURE (event='1') =  PRIORFRAC AGE RATERISK /

risklimits lackfit ctable clodds=both;

output out = bonesOut predprobs=I p=predprob resdev=resdev reschi=pearres;

footnote 'Figure 7';

run;
```

```
proc print data=fracAll; run;
```

```
/*Look at residuals to see if anything is out of the norm for any high leverage
points*/

title 'GPlot of output from LR vs Observations';

proc gplot data=bonesOut;

plot resdev * numObs;

plot pearres * numObs;

plot predprob * numObs;

footnote 'Figure 8';

run;

quit;
```

```
/*Check to ensure MOMFRAC and HEIGHT do not impact other variables in the
model*/

proc reg data=bones;

model FRACTURE = PRIORFRAC AGE RATERISK MOMFRAC HEIGHT /

VIF lackfit;
```

```
run;

 #Options – VIF (Variance Inflation Factor) for multicollinearity (correlation
between predictors) and lackfit for gauging the fitted model
```

```
/*Added MOMFRAC and HEIGHT to the model for improvements*/

title 'Logistic Regression 2 - PRIORFRAC AGE RATERISK MOMFRAC HEIGHT';

proc logistic data=bones outest= fracAll;

model FRACTURE (event='1') =  PRIORFRAC AGE RATERISK MOMFRAC HEIGHT /

risklimits lackfit ctable clodds=both;

output out = bonesOut2 predprobs=I p=predprob resdev=resdev reschi=pearres;

footnote 'Figure 9';

run;
```

```
title 'GPlot of output from Logistic Regression 2 vs Observations';

proc gplot data=bonesOut2;

plot resdev * numObs;

plot pearres * numObs;

plot predprob * numObs;

footnote 'Figure 10';

run;

QUIT;
```
