Analysis of NFL Combine Data with Linear Regression

Applied Statistics: Hans Larson (‘20)

ABSTRACT

This data comes from the NFL Combine between the years 2012 and 2017. It contains data from the physical tests that NFL prospects took to catch the eye of possible NFL draft recruiters. My primary goal is to see what predictor variables have the highest correlation with a player's 40-yard-dash time. Quarterbacks were excluded from the dataset because of their inconsistent data, and I deleted any players that had N/As for any of the predictor variables. There are 627 observations in this data set.

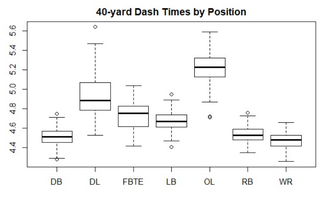
DATA CHARACTERISTICS

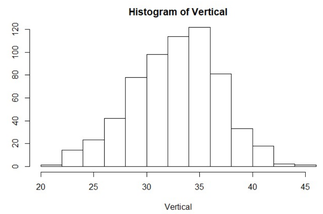
Predictor variables used in the initial analysis:

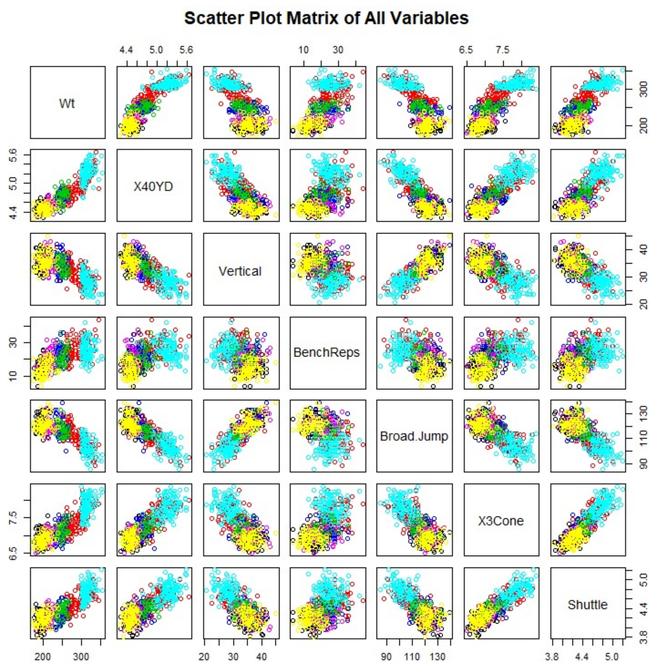
Variable Name	Description
Weight	(lbs.)
Vertical	Standing Vertical Jump Height (in.)
BenchReps	Number of bench reps at 225 lbs.
Broad.Jump	Horizontal distance jumped (in.)
X3Cone	Three-cone drill time (s)
Shuttle	Shuttle drill time (s)
Pos.cat	Category of position (DB, OL, etc.)

Predictor variables used in the initial analysis:

Variable Name	Description
X40YD	40-yard dash time (s)







METHODOLOGY

Multiple linear regression is a form of predictive regression analysis used to explain the relationship between two or more continuous or categorical predictor variables and one continuous response variable. The general equation for a multiple linear regression model is:

$$y_i = B_0 + B_1x_{i1} + B_2x_{i2} + \ldots + B_px_{ip} + E \text{ where } i = 1,2, \ldots, n$$

INITIAL MODEL FIT

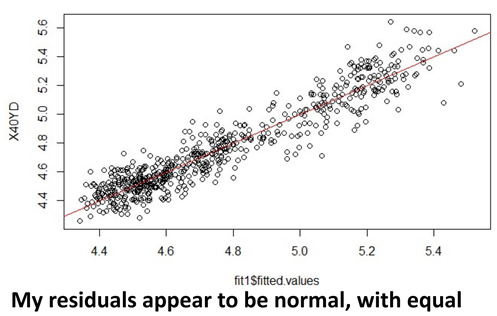
A first-order model of 40-yard dash vs. weight, vertical, bench reps, broad jump, 3-cone drill, shuttle drill, and position category gave the following results:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.2460549  0.1922023  22.092  < 2e-16 ***
Wt           0.0035313  0.0003461  10.203  < 2e-16 ***
Vertical    -0.0013759  0.0018658  -0.737 0.461144
BenchReps   -0.0041684  0.0008972  -4.646 4.14e-06 ***
Broad.Jump  -0.0083942  0.0008988  -9.340  < 2e-16 ***
X3Cone       0.0898675  0.0238364   3.770 0.000179 ***
Shuttle      0.0191337  0.0355028   0.539 0.590127
pos.catDL    0.0032358  0.0281569   0.115 0.908546
pos.catFBTE -0.0008956  0.0234698  -0.038 0.969573
pos.catLB   -0.0094577  0.0196528  -0.481 0.630517
pos.catOL    0.0802511  0.0345552   2.322 0.020537 *
pos.catRB   -0.0339966  0.0180168  -1.887 0.059641 .
pos.catWR   -0.0502122  0.0146411  -3.430 0.000645 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From this fit, it appears that an increase in vertical, bench reps, and broad jump results in a faster 40-yard dash time, while an increase in weight and a slower 3-cone drill and shuttle time result in a slower 40-yard dash time.

We can also see that, when compared to defensive backs, positions like running backs and wide receivers tend to be faster, and positions like defensive line and offensive line tend to be slower players.

RESIDUAL ANALYSIS OF INITIAL MODEL



My residuals appear to be normal, with equal variance and no obvious outliers

There also appears to be two groups of data that have faster and slower 40-yard dash times, with more data points being in the faster cluster. This is due to the fact that we have more fast players in this dataset than slow.

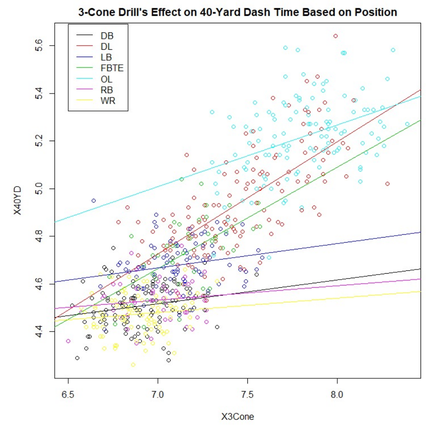
MODEL SELECTION

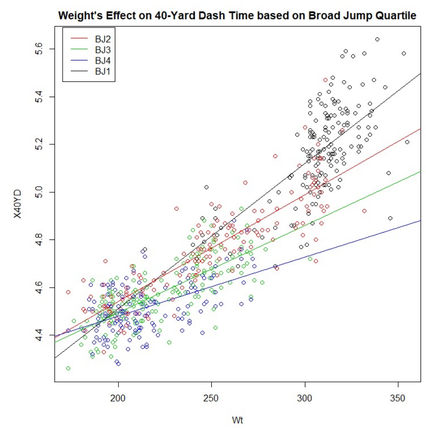
A stepwise model selection procedure was used, where I used both backward and forward selection. starting with the full interaction model.

The initial model assumed there are no interactions between the predictor variables. As a part of selecting the final model, I checked for possible interactions between variables. The variables selected by stepwise regression returned:

Predictor Variables	Interaction Effects
Weight	Weight:Broad Jump
Bench Reps	Broad Jump:3-cone drill
Broad Jump	3-cone drill:Position Category
3-cone drill	
Position Category	

Interaction Effects Plots from the final model:

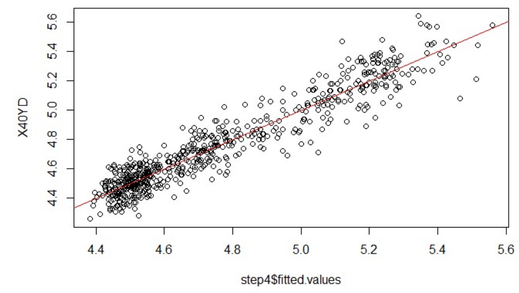




In the interaction plots, I discovered that the faster, more agile positions' 40-yard dash times are less affected by their three-cone drill times. We also discovered that the longer the broad jump, the less a player's 40-yard dash time is affected by weight.

MODEL DIAGNOSTICS

A plot of the residuals vs. fitted values reveals no serious problems. The residuals still appear to be normal, with equal variance and no obvious outliers.



Final Model Parameter Estimates

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.667e+00  1.933e+00   2.932 0.003497 **
Wt                   1.302e-02  2.624e-03   4.960 9.16e-07 ***
BenchReps           -4.362e-03  8.627e-04  -5.057 5.66e-07 ***
Broad.Jump          -1.635e-02  1.558e-02  -1.049 0.294382
X3Cone              -4.467e-01  3.355e-01  -1.331 0.183544
pos.catDL           -7.809e-01  4.331e-01  -1.803 0.071868 .
pos.catFBTE         -2.036e+00  6.389e-01  -3.187 0.001512 **
pos.catLB            2.324e-01  4.927e-01   0.472 0.637334
pos.catOL           -4.205e-01  5.505e-01  -0.764 0.445315
pos.catRB            5.724e-01  6.266e-01   0.913 0.361355
pos.catWR            4.143e-01  5.471e-01   0.757 0.449176
Wt:Broad.Jump       -8.757e-05  2.300e-05  -3.808 0.000154 ***
Broad.Jump:X3Cone    4.186e-03  2.748e-03   1.523 0.128207
X3Cone:pos.catDL     1.198e-01  6.188e-02   1.936 0.053301 .
X3Cone:pos.catFBTE   2.952e-01  9.024e-02   3.271 0.001130 **
X3Cone:pos.catLB    -2.473e-02  7.012e-02  -0.353 0.724428
X3Cone:pos.catOL     8.117e-02  7.538e-02   1.077 0.281992
X3Cone:pos.catRB    -8.284e-02  8.990e-02  -0.921 0.357186
X3Cone:pos.catWR    -6.690e-02  7.931e-02  -0.843 0.399308
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

CONCLUSIONS

The final model, after many methods of predictor variable elimination, boiled down to these most signicant predictors: weight, bench reps, broad jump, three-cone drill, and position category., with interaction effects weight x broad jump, broad jump x three-cone drill, and three-cone drill x position category.

- **I learned that average 40-yard dash time is faster with faster three-cone and shuttle drills, higher vertical, longer broad jump, and higher bench reps. 40-yard dash time becomes slower with higher weight.**
- **Adding interaction effects increased our r-squared the most (increased from originally .8904 to .8975) and gave us some very interesting interaction plots.**
- **Example prediction intervals predicted 95% correct, and appeared to become more accurate as the 40-yard dash time got slower.**