Calculating SPARQ and Projecting 2016 NFL Edge Rushing Prospects

Adam Wright

May 10, 2016

1 - Introduction

The purpose of this project is to develop a model to predict the four year sack totals of NFL edge rushing prospects. Sacks are chosen as the outcome variable because they represent the simplest direct measurement of a NFL edge rusher's primary function: disrupting the opponents passing game by harassing the opposing quarterback. Four years are chosen as the time horizon because that is the number of guranteed years for NFL rookie contracts.

My model will include three predictor variables:

1. **A measure of the player's qualitatively scouted ability** - either the round he was drafted in for current/former NFL players or a consesnsus projection of where the player will be drafted for prospects.
2. **A measure of the player's college production** - his production ratio which is defined as:

$$ProductionRatio = (Sacks + TacklesforLoss)/GamesPlayed$$

This ratio will only be calculated for the last two years of a player's eligibility to reflect the fact that many players, even stars, do not play very much as underclassmen.

3. **A measure of the player's athletic ability** - his SPARQ score.

Player's college stats, draft position/draft projection, and NFL statistics are freely available on the web. Therefore, collecting the data for my outcome variable (NFL sacks in first five years), and two of my predictor variables (draft position/projection and college production) will be a rather extensive but straightforward exercise in web scraping and data munging. SPARQ scores, on the other hand, are not publically available and to incorporate them into the model I will have to create my own means of calculating them.

2 - SPARQ Scores

2.1 - SPARQ Scores - Introduction

SPARQ scores - which stand for Speed, Power, Agility, Reactions, and Quickness scores - are a metric that was developed by Nike in conjunction with the staff of the Seattle Seahawks to provide a single quantitative measurement of a player's football specific athleticism. The scores are used extensively in the football scouting community from the time players are high school prospects and their SPARQ scores are measured at regional Nike camps up through their testing for the NFL draft where several teams (Seattle Seahawks, Dallas Cowboys, Cleveland Browns) are known to incorporate them into their desicions about whom to draft.

While the explicit formula for calculating SPARQ scores has never been made publically available the five components that go into the final score are known:

1. **Weight** - used to normalize results so that 300 lb. lineman can be compared to 180 lb. wide receivers on the same scale.
2. **Forty-yard dash time** - a measure of a prospect's speed and reaction time
3. **Short shuttle time** - a measure of a prospect's quickness and reaction time
4. **Vertical jump** - a measure of a prospect's explosive, lower body power
5. **Powerball throw** - a measure of a prospect's strength and upper body power

This component data (or an easily translatable analogue in the case of powerball throw) ***is publically available*** for a great majority of former and current NFL prospects in the form of NFL combine athletic testing data. I was also able to find over 18,000 SPARQ/5-component pairs for 2012 highschool and 2014 NFL athletes. These pairs will allow me to model the SPARQ score and create my own explicit formula for its calculation. This secured I will have all of the necessary components for my model.

2.2 - SPARQ Scores - Data Munging

I begin by loading the packages I will need for this analysis:

```
setwd('C:/Users/Adam/Udacity/Data_Analyst_Nanodegree/Project_4')
library(moments) # to calculate skewness
library(tidyr)
library(dplyr)
library(plyr)
library(ggplot2)
```

Next, I load my SPARQ data for 2012 highschool athletes and 2014 NFL prospects, clean the tables up, and join them into a single dataframe. The 2012 highschool data was scraped from ESPN and the 2014 NFL data was scraped from Zach Whitman's 3sigmaathlete.com.

As part of this process it is neccessary to translate high school players powerball throws into bench press scores. The bench press is the test used at the NFL combine to measure a player's upper body power - specifically how many repetitions a player can do at 225 pounds. Because not many high school players can do many (or any!) repetitions at that weight they do an alternate drill - the powerball throw - to test their strength and upper body power. Fortunately, it is straightforward to convert between the two using z-scores, which I did for all of my highschool athletes following Zach Whitman's table from the "Converting Bench Press to Powerball Throw" section of this page.

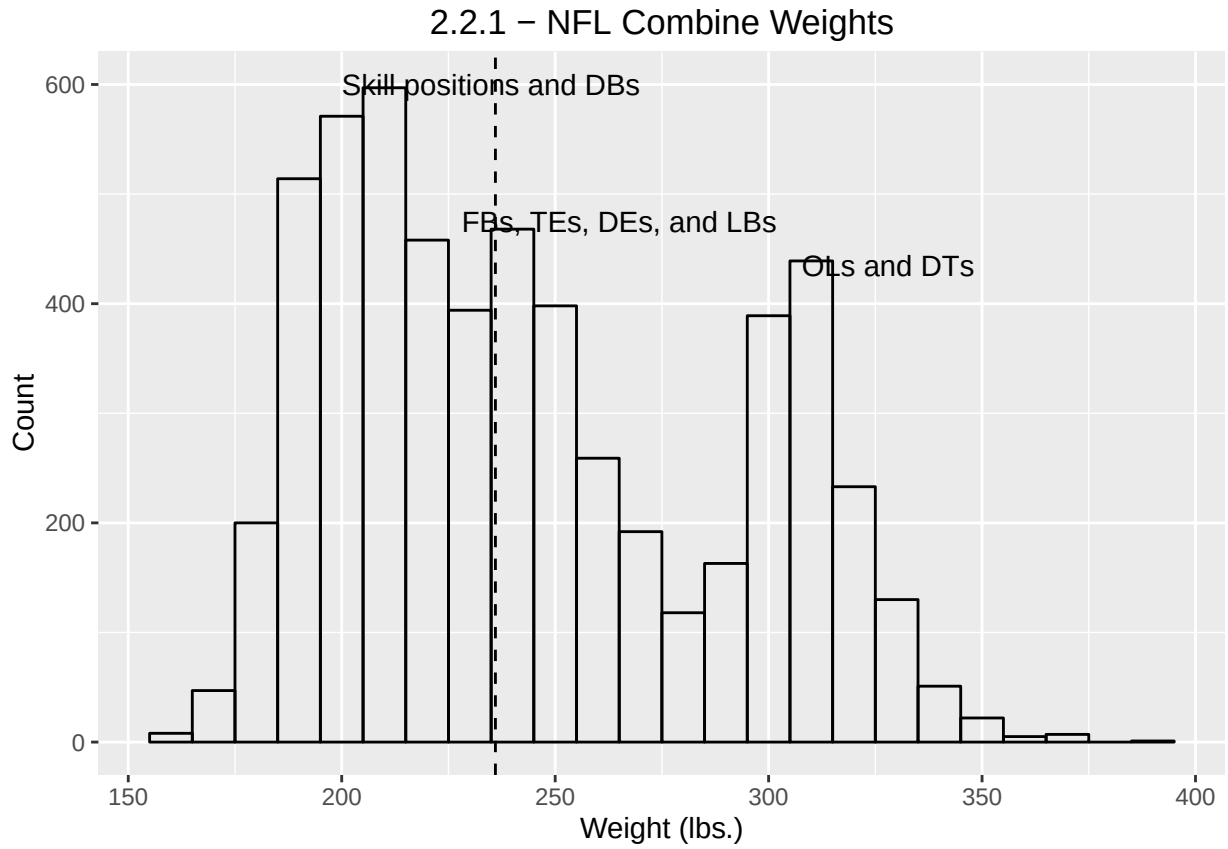
Note: the conversion from powerball throw to bench press was done in Excel since I could not find a good way to accomplish Excel's vlookup functionality in R!

```
# read in data
highschool_sparq_2012 <- read.csv("highschool_sparq_2012.csv")
nfl_sparq_2014 <- read.csv("nfl_sparq_2014.csv")

# clean and combine data
highschool_2012 <- highschool_sparq_2012 %>% select(-height)
highschool_2012$name <- as.character(highschool_2012$name)
highschool_2012$pos <- as.character(highschool_2012$pos)
nfl_sparq_2014_clean <- nfl_sparq_2014 %>% select(-c(Pick, Team, Class, School,
                                                     HT..ft., Arm.Length..in.,
                                                     Hand.Size..in., X10.yd..s.,
                                                     X3Cone..s.,
```

```
                                                    Broad.Jump..ft.,
                                                    pSPARQ, rSPARQ.z.score,
                                                    pSPARQ.z.score))
colnames(nfl_sparq_2014_clean) <- c('name', 'pos', 'weight', 'forty_yard',
                                    'shuttle', 'bench_press','vertical_jump',
                                    'sparq')
nfl_sparq_2014_clean$name <- as.character(nfl_sparq_2014_clean$name)
nfl_sparq_2014_clean$pos <- as.character(nfl_sparq_2014_clean$pos)
nfl_sparq_2014_clean <- nfl_sparq_2014_clean[, c(1:5, 7, 6, 8)]
nfl_sparq_2014_clean <- nfl_sparq_2014_clean[-1704,]
# remove case with ridiculous value (shuttle > 40)
sparq_calc_data <- union(nfl_sparq_2014_clean, highschool_2012)
sparq_calc_data <- sparq_calc_data[-c(17882,2940),]
# remove cases with ridiculous values (e.g. shuttle > 40)
```

In order to give readers unfamiliar with NFL Combine testing data a sense of the variation for each of the five SPARQ predictor variables what follows are histograms and summary statistics for each:

2.2.1 - NFL Combine Weights

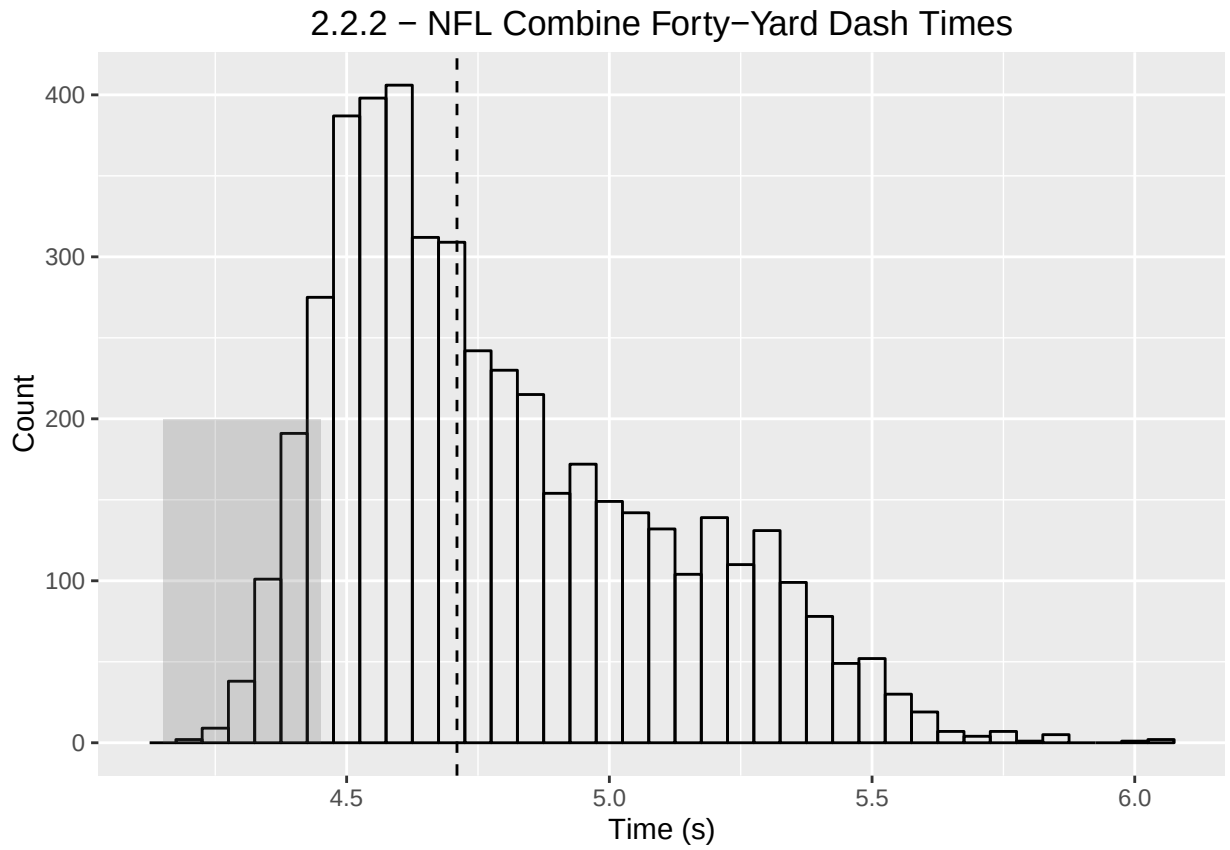



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   155.0   207.0   236.0   245.1   288.0   386.0
```

The distribution of weights in the NFL is multimodal, with 3 distinct peaks for different position groups as shown above. One of the things that makes football so fascintating is the tremendous physical variation in

players as is exemplified by the range of weights observed at the NFL combine (>200 lbs!). Finally, NFL players are enormous compared to the average human being, with a median weight of 236 lbs., which is 41 lbs. greater (21%!) than the average American male at 195 lbs.

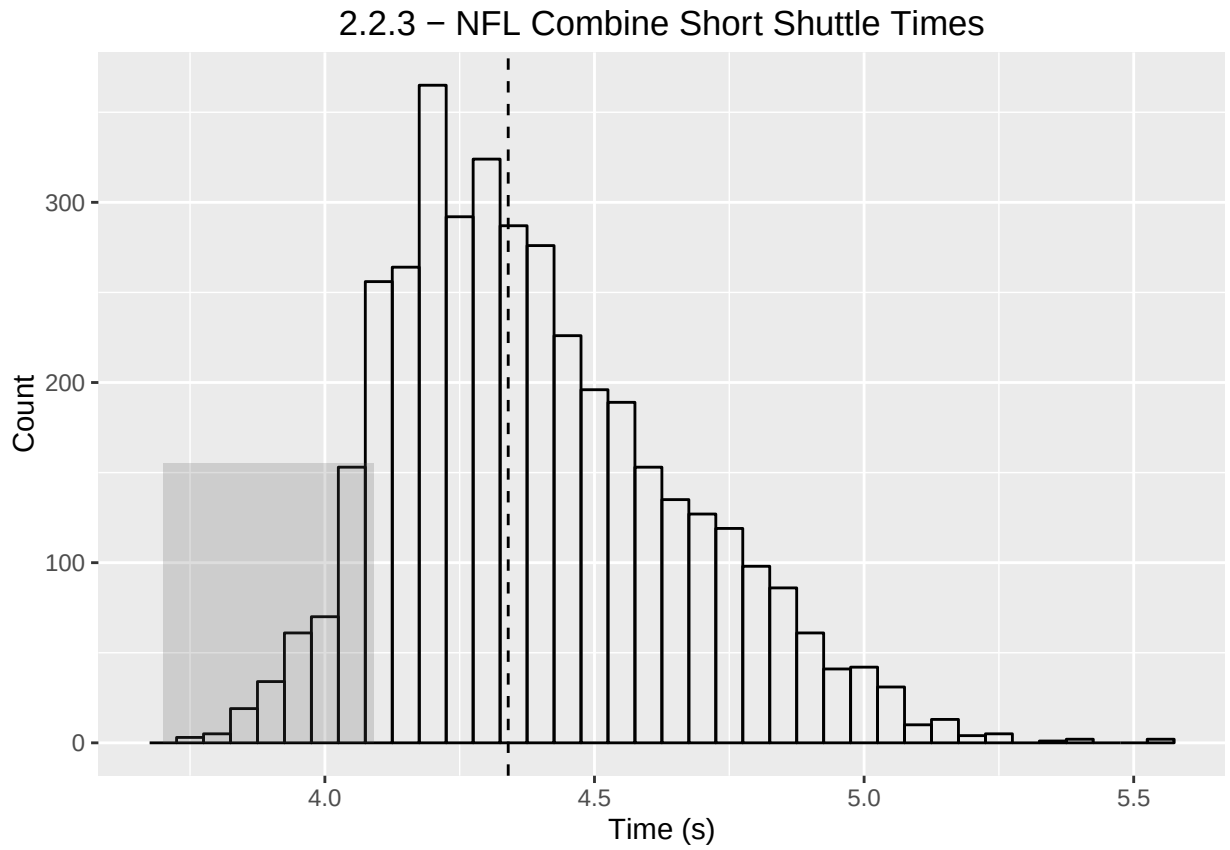
2.2.2 - NFL Combine 40-yard Dash Times




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   4.210   4.550   4.710   4.793   5.000   6.050     962
```

Forty-yard times at the NFL Combine have a decided right skew - most players meet a minimum speed standard though there is a long right tail composed of many QBs and huge linemen. The elite speed players - those with a 40 time of 4.45 or lower - make up just 10% of the population of NFL prospects. Most succesful WRs and CBs fall in this group (highlighted by rectangle, above).

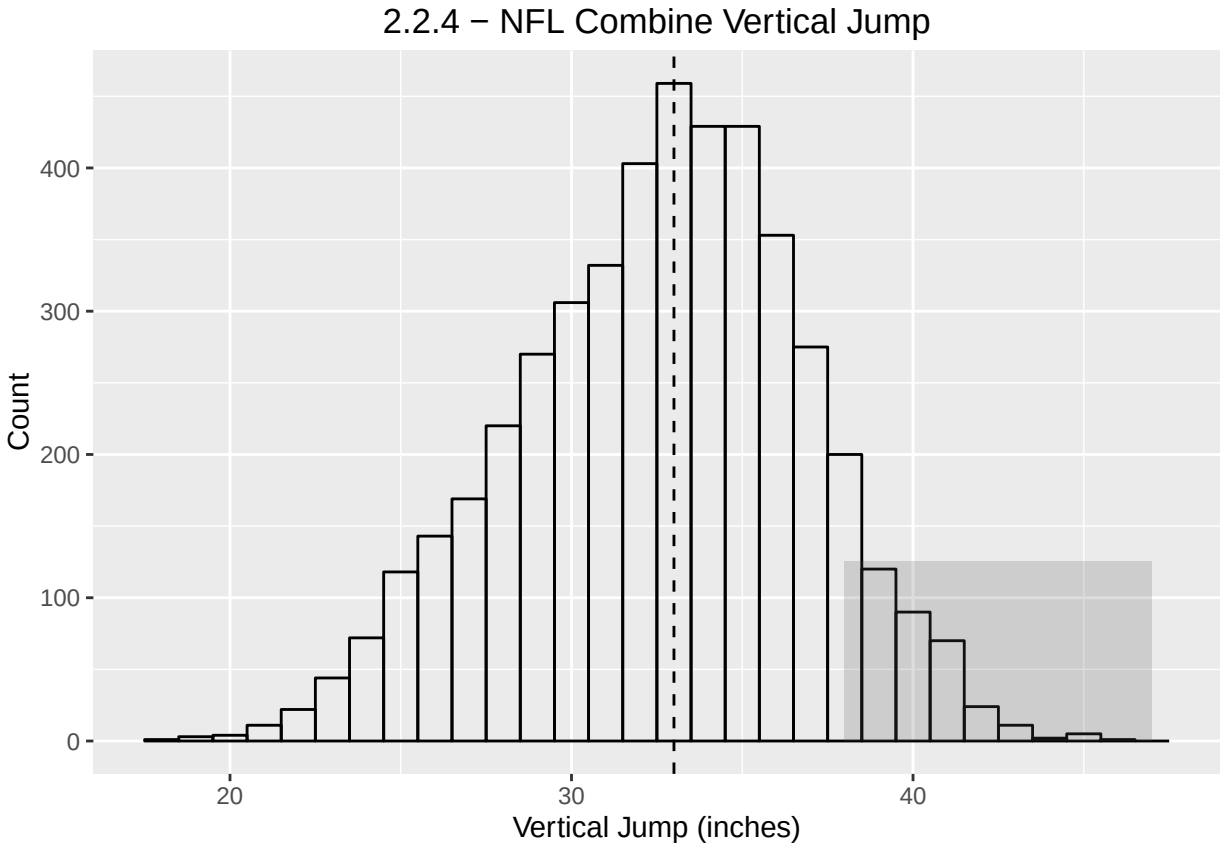
2.2.3 - NFL Combine Short Shuttle Times




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    3.73    4.19    4.34    4.39    4.56    5.56    1713
```

The distribution of shuttle times is similar to that of forty-yard times, though not quite as right skewed. An elite shuttle time is 4.1 seconds or below (highlighted by rectangle, above).

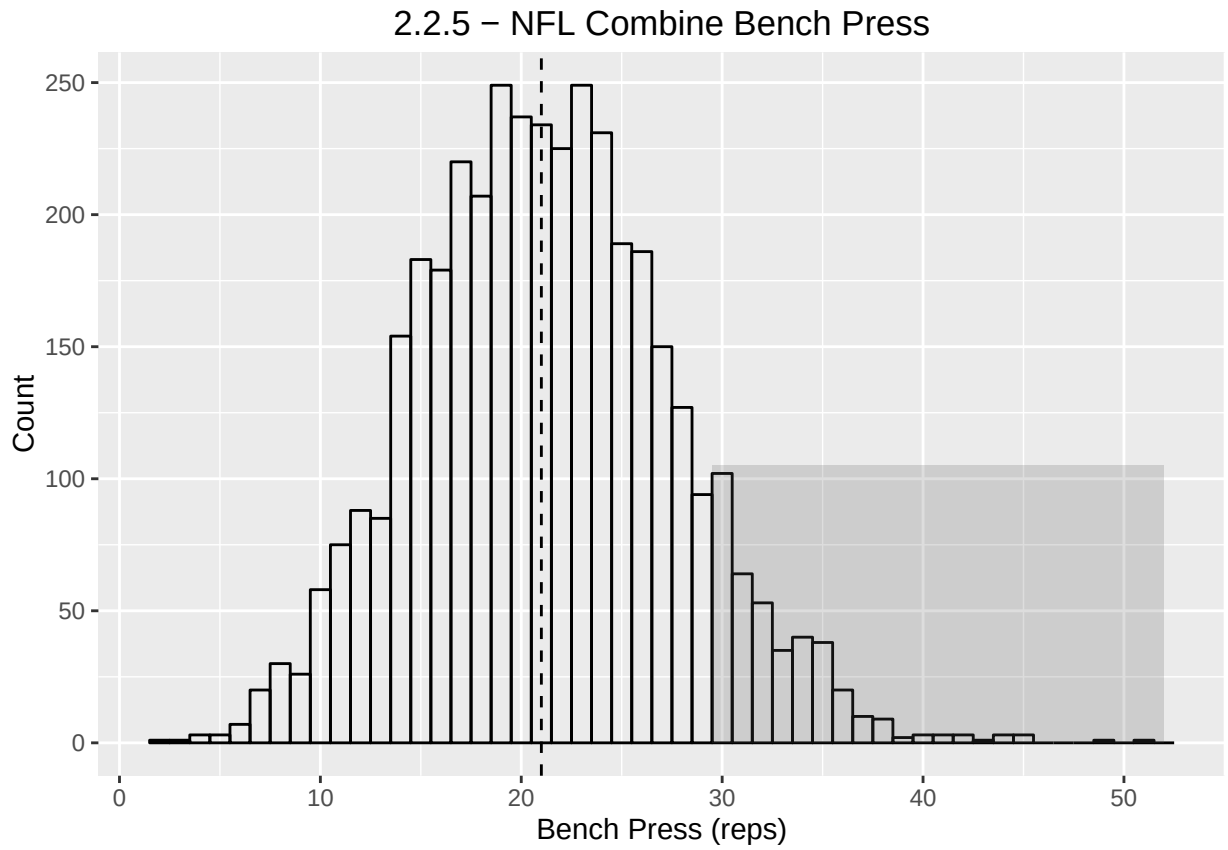
2.2.4 - NFL Combine Vertical Jump




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   17.50   30.00   33.00   32.82   36.00   46.00    1077
```

Vertical jump scores display a slight left skew, but are relatively symmetric compared to 40-yard and shuttle times. An elite vertical jump is 38 inches or higher (highlighted by rectangle, above).

2.2.5 - NFL Combine Bench Press

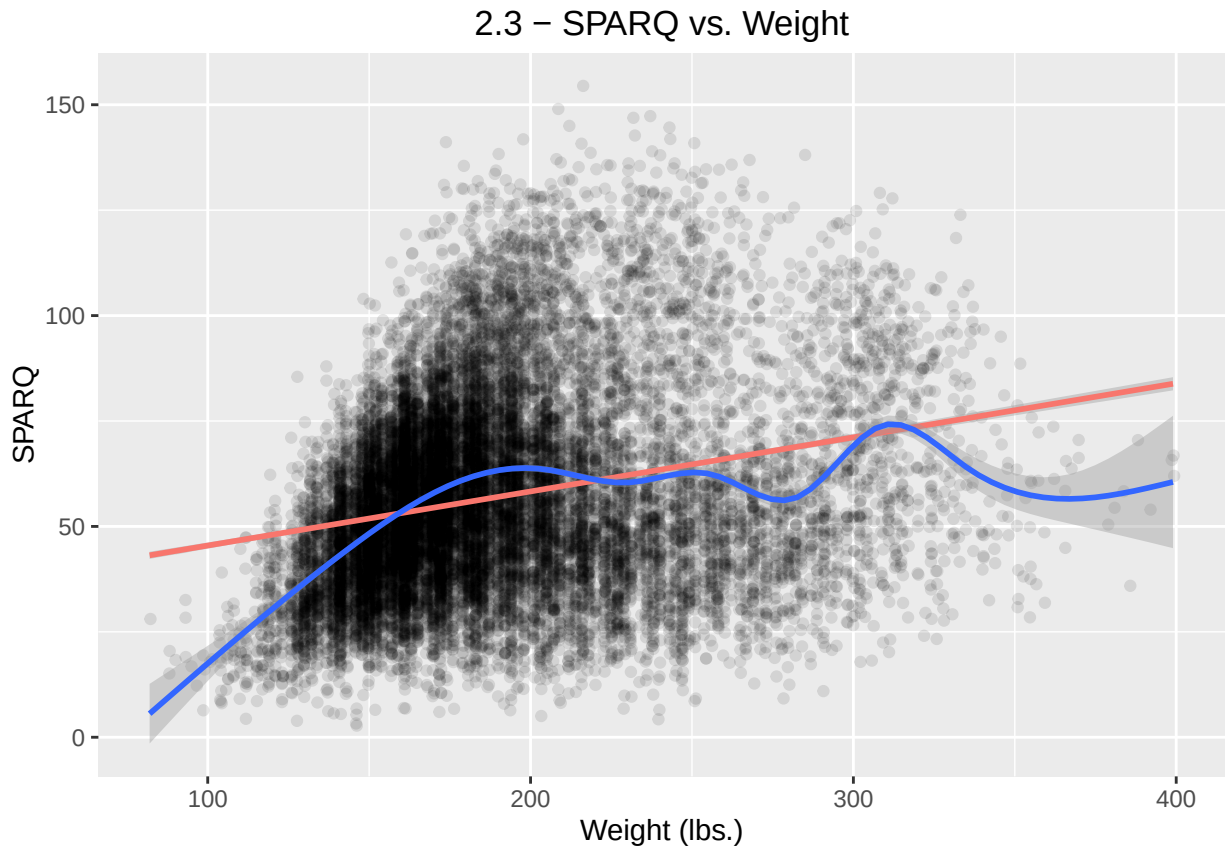



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.00   17.00   21.00   21.23   25.00   51.00    1761
```

The bench press distribution has a slight right skew, with a relatively long but sparsely populated tail of freakishly strong players. An elite bench press is anything at or above 30 reps (highlighted by rectangle, above).

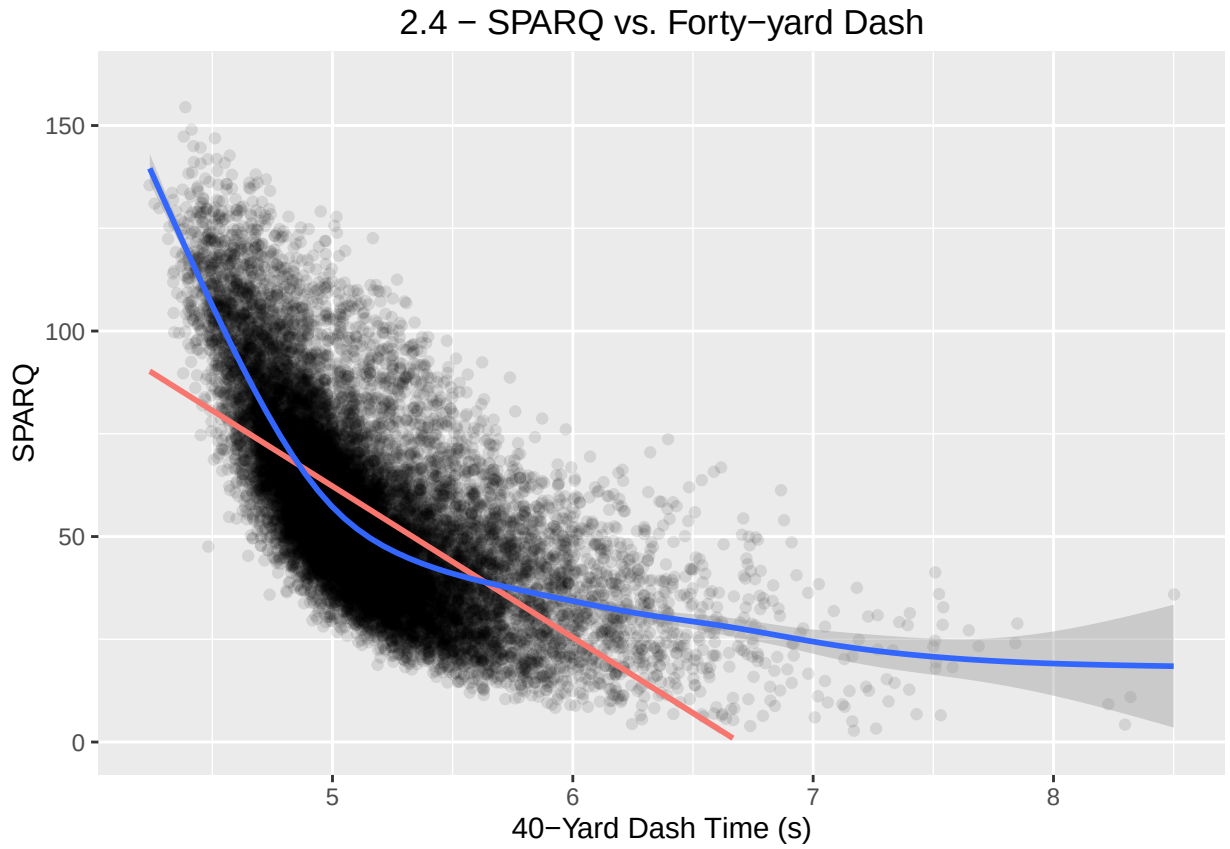
2.3 - SPARQ Scores vs. Weight

With the data combined into a single data frame, I proceed to examine the univariate relationships between each of the SPARQ components and the overall score, beginning with weight.



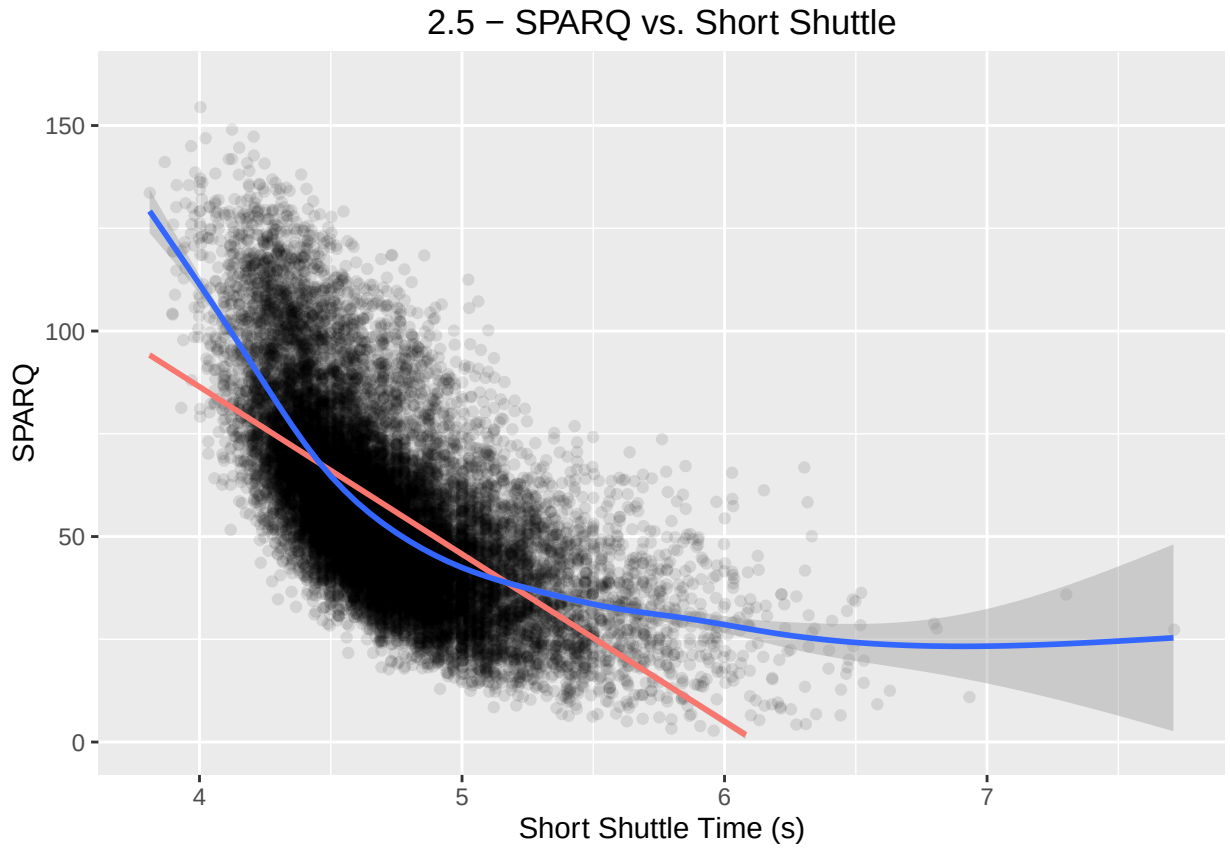
As expected, there is no strong correlation between weight and SPARQ score ($R^2 = 0.06$), as the primary purpose of incorporating a player's weight is to normalize results so that small, fast players can be compared on the same scale to big, strong players. It is worth noting that what relationship there is positive which makes sense - all else equal a bigger football player is a better football player since he can generate more force with the same acceleration.

2.4 - SPARQ Scores vs. Forty-yard Dash



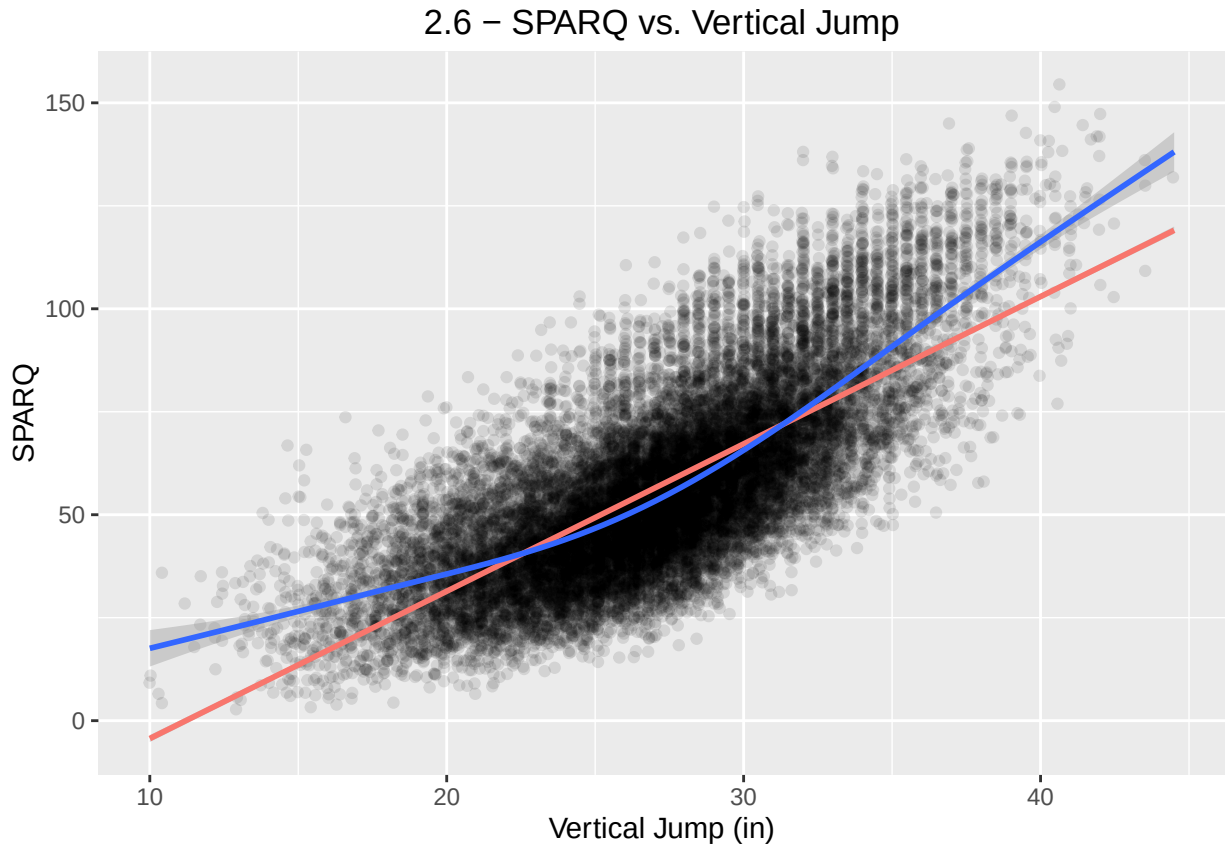
The correlation between forty-yard dash and SQARQ score ($R^2 = 0.42$) is significantly stronger than between weight and SPARQ. The important thing to note about the above graph is that the relationship is clearly not linear - rather it is a clear example of exponential decay. That is, the rewards of going from a 4.6 to a 4.5 forty are much higher than is the penalty from going to a 5.0 to a 5.1 forty. Speed kills in football and it appears that Nike incorporated this into thier SPARQ score by modelling it as an exponential function.

2.5 - SPARQ Scores vs. Short Shuttle



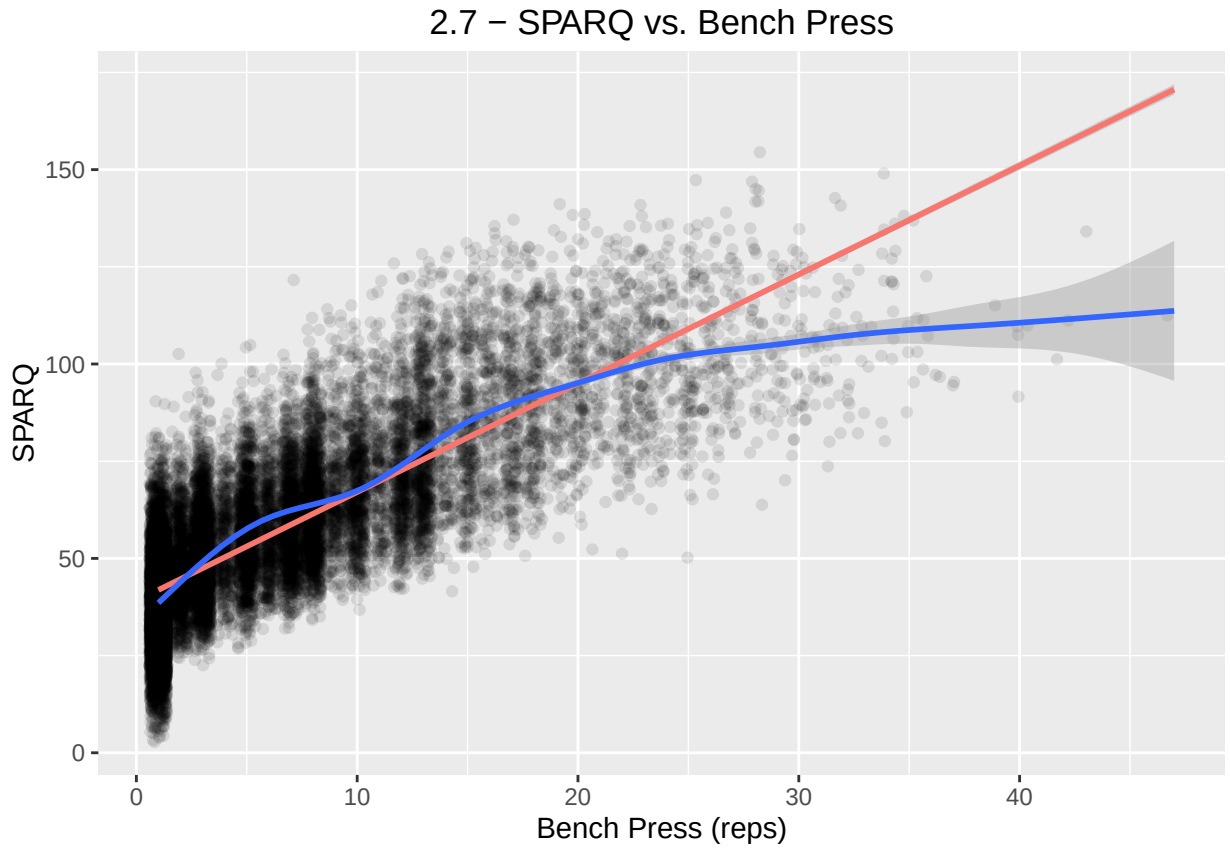
Unsurprisingly, a player's quickness is also relatively highly correlated with his SPARQ score ($R^2 = 0.40$). And like forty-yard dash, the relationship between short shuttle and SPARQ is clearly not linear but once again an example of exponential decay.

2.6 - SPARQ Scores vs. Vertical Jump



The correlation between a player's vertical jump and his SPARQ score ($R^2 = 0.54$) is actually higher than both forty-yard dash and short shuttle. I was initially surprised by this, but clearly Nike puts a lot of emphasis on a player's lower body explosion. This actually makes a lot of sense, especially for power players playing close to the line of scrimmage (i.e. offensive and defensive linemen). The relationship between vertical jump and SPARQ appears nearly linear, but perhaps could best be modelled as a low degree polynomial.

2.7 - SPARQ Scores vs. Bench Press



Bench press is the component most highly correlated with the overall SPARQ score ($R^2 = 0.62$). The relationship looks like the most strictly linear of the five components, though the possibility of a logarithmic growth function with a maximum possible benefit (that is doing, extra bench press reps would have no effect on the SPARQ score beyond a certain point) being a better fit exists.

2.8 - Fitting the SPARQ Model

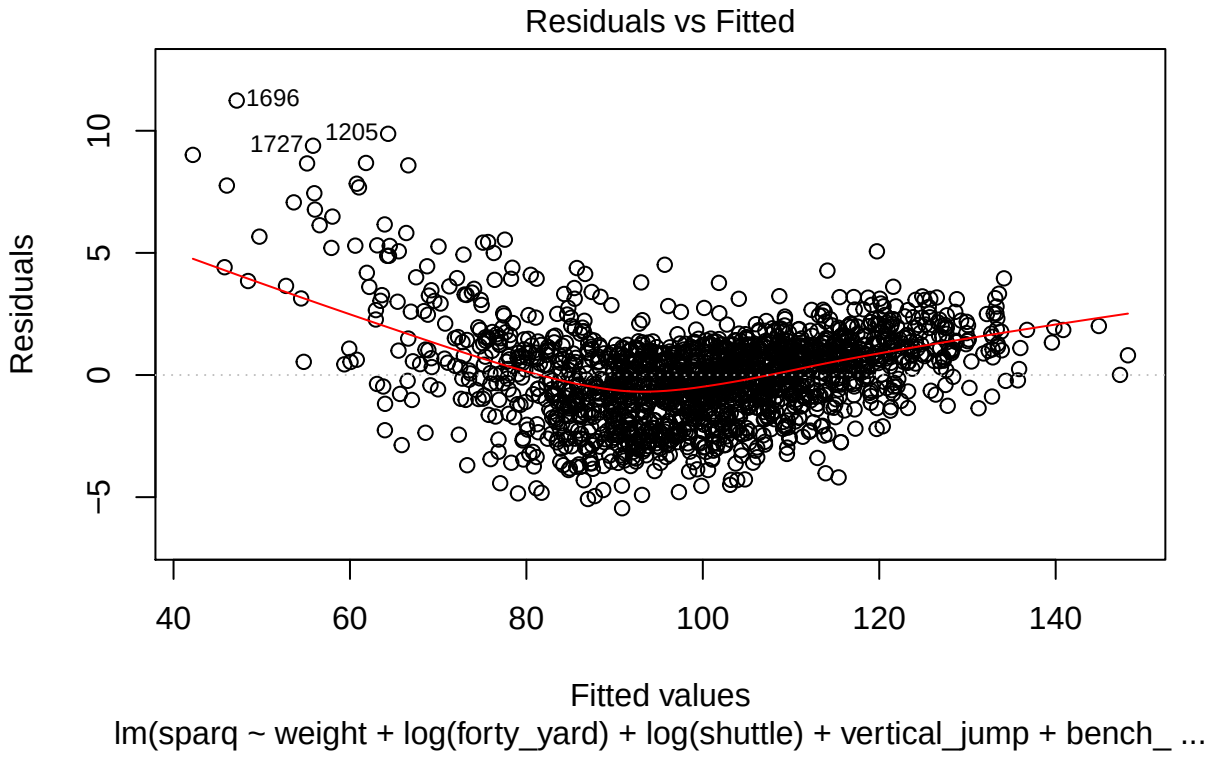
Now that the relationships between individual variables and the SPARQ score are understood, it is time to model the SPARQ score. I will attempt to fit the SPARQ score to linear model consisting of the five SPARQ factors: weight, forty-yard dash time, short shuttle time, vertical jump, and bench press repetitions. Based upon the relationships observed above, the forty-yard dash and short shuttle times will be modelled as log functions of the value to reflect their exponential relationship with the SPARQ score. The other three factors will be modelled as simple, linear contributors to the SPARQ score.

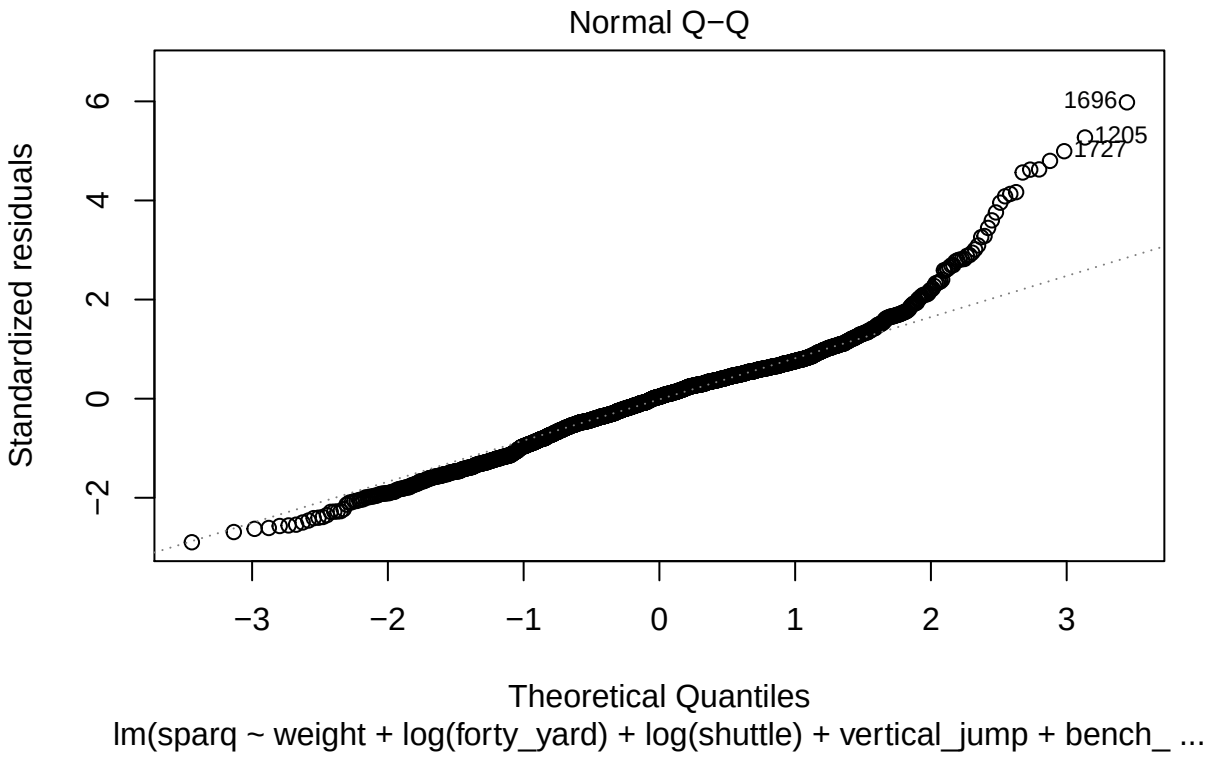
As a final note, I will only try to fit my model to the NFL players as the high school data contains many cases of players either much too small and/or not nearly athletic enough to be considered NFL prospects. Trying to fit those extreme cases will only introduce noise for the SPARQ scores that I actually care about: those of individuals meriting consideration as NFL prospects.

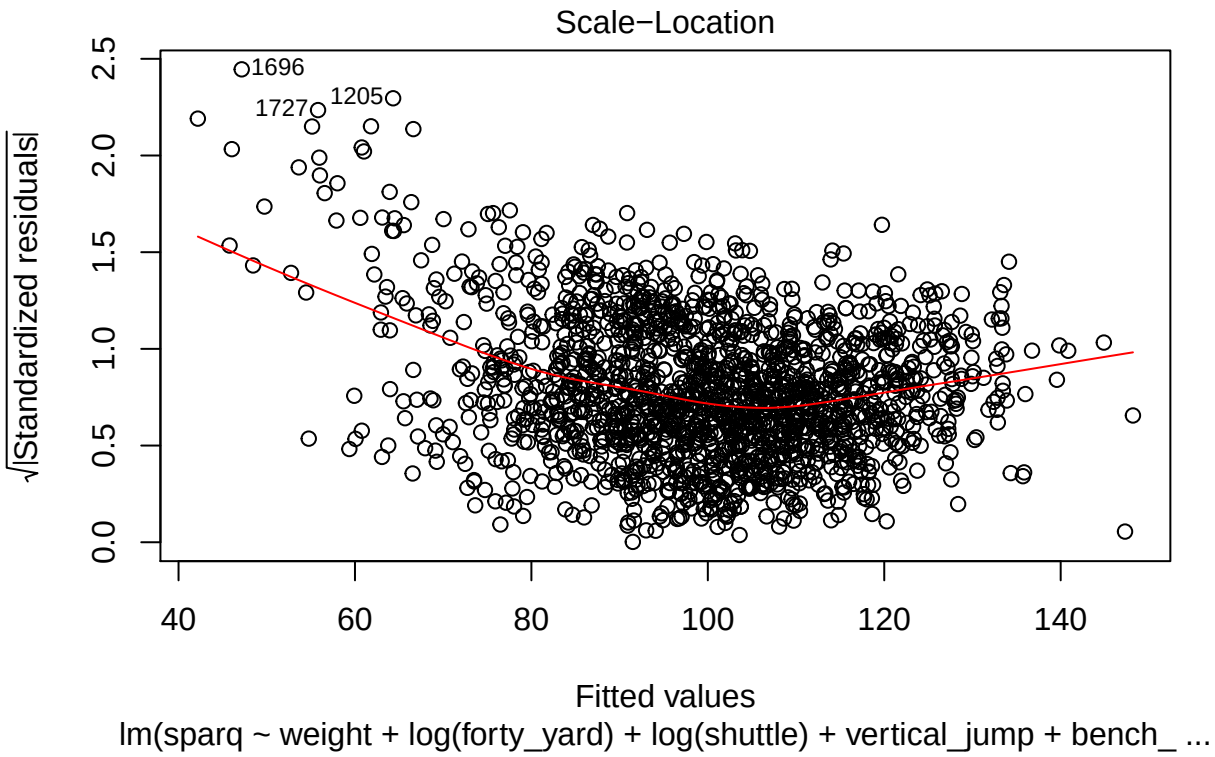
```
sparq_fit = lm(sparq ~ weight + log(forty_yard) + log(shuttle) +
                 vertical_jump + bench_press, data = nfl_sparq_2014_clean)

summary(sparq_fit)
```

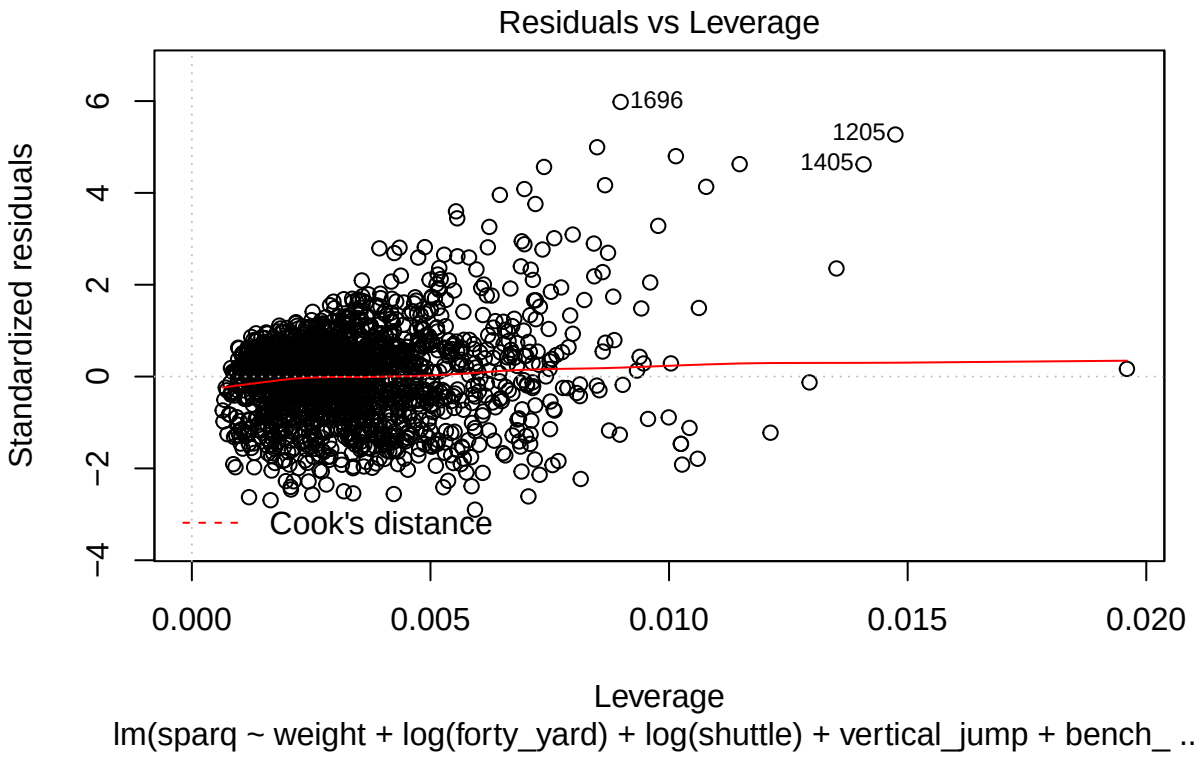
```
##
## Call:
## lm(formula = sparq ~ weight + log(forty_yard) + log(shuttle) +
##     vertical_jump + bench_press, data = nfl_sparq_2014_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4527 -1.0834  0.0718  1.0259 11.2364
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.583e+02  2.887e+00  158.76   <2e-16 ***
## weight           2.416e-01  2.353e-03  102.71   <2e-16 ***
## log(forty_yard) -1.809e+02  1.752e+00 -103.22   <2e-16 ***
## log(shuttle)    -1.350e+02  1.373e+00  -98.28   <2e-16 ***
## vertical_jump    1.652e+00  1.717e-02   96.20   <2e-16 ***
## bench_press      9.399e-01  9.600e-03   97.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.887 on 1739 degrees of freedom
## Multiple R-squared:  0.9858, Adjusted R-squared:  0.9858
## F-statistic: 2.421e+04 on 5 and 1739 DF,  p-value: < 2.2e-16
```

```
plot(sparq_fit)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sparq ~ weight + log(forty_yard) + log(shuttle) + vertical_jump + bench_ ...
1696
1205
1727
−2
0
2
4
6
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
1696
1727
1205
40
60
80
100
120
140
Fitted values
lm(sparq ~ weight + log(forty_yard) + log(shuttle) + vertical_jump + bench_ ...



The results of the model are displayed in the summary table and plots, above. Happily, this model produces an extremely strong fit for the NFL data with an ($R^2 = 0.986$). The residual standard error is also a relativley low 1.89 - meaning that about once in every 100 cases the model will be off by ~5.67 SPARQ points (i.e. three standard deviations). I am quite pleased with this model and the explicit equation that I will use to calculate SPARQ scores from here on out is below:

$$SPARQ = 458.3 + 0.2416(weight) - 180.9(log(forty)) - 135.0(log(shuttle)) + 1.652(verticaljump) + 0.9399(bench)$$

3 - Data Collection and Data Munging

With a means for calculating SPARQ scores in hand, it is time to collect and prepare the data to fit my final model.

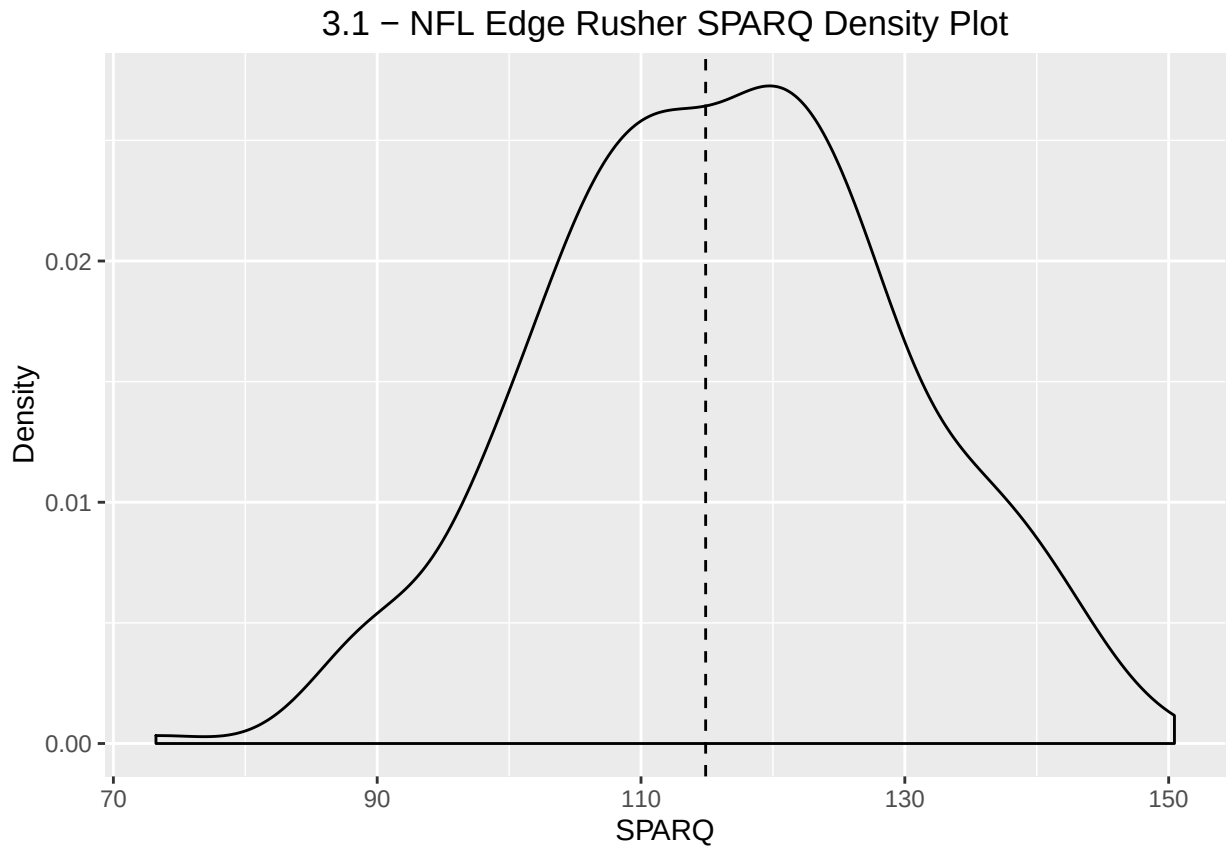
3.1 - NFL Combine Data

First, I read in the NFL combine data - scraped from the invaluable NFL Combine Results - pare it down to what I need to calculate SPARQ scores, calculate the SPARQ scores, and calculate z-scores and percent ranks for those SPARQ scores.

```
# read in data
nfl_combine <- read.csv('nfl_combine.csv')

# pare data down to what is required, eliminate non-complete cases,
```

```r
# and munge variable into proper formats
nfl_combine_edge <- nfl_combine %>% select(-college, -height, -wonderlic,
                                           -broad_jump, -three_cone) %>%
  # remove non-edge rushers and players with less than 4 years of NFL experience
  filter(pos %in% c('DE', 'OLB') &
           !year %in% c('2016', '2015', '2014', '2013')) %>%
  # removes off-line of scrimmage linebackers (i.e. 4-3)
  filter(!(pos == 'OLB' & weight < 240))
nfl_combine_edge$forty_yard <- as.character(nfl_combine_edge$forty_yard)
nfl_combine_edge$forty_yard <- as.numeric(nfl_combine_edge$forty_yard)
colnames(nfl_combine_edge)[7] <- 'vertical_jump'
nfl_combine_edge <- nfl_combine_edge[complete.cases(nfl_combine_edge), ]
nfl_combine_edge <- separate(nfl_combine_edge, name,
                             sep = " ", c('first_name', 'last_name'))
write.csv(nfl_combine_edge, 'nfl_combine_edge.csv')

# calculate sparq, sparq z-score, and sparq percent rank
nfl_combine_edge$sparq <- round(predict(sparq_fit,
                                        newdata = nfl_combine_edge), 2)
nfl_combine_edge$sparq_zscore <- round(scale(nfl_combine_edge$sparq,
                                             center = TRUE, scale = TRUE), 2)
nfl_combine_edge$sparq_percent_rank <- round(percent_rank
                                             (nfl_combine_edge$sparq), 3)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   73.23  106.70  116.60  116.40  125.70  150.40
```



I am left with 253 edge rushers with SPARQ scores. The mean and median SPARQ score for this group is 116.36 with a standard deviation of 13.5. As can be seen in plot 3.1, above, the distribution is normal.

3.2 - Production Ratio and NFL Draft Round

With the SPARQ scores calculated, it is not time to gather my other two model parameters: the round the player was drafted in and their college production ratio. Fortunately, I can scrape both from the player pages at Sports Reference's College football page.

The scraping for the college stats was a bit problematic for three reasons:

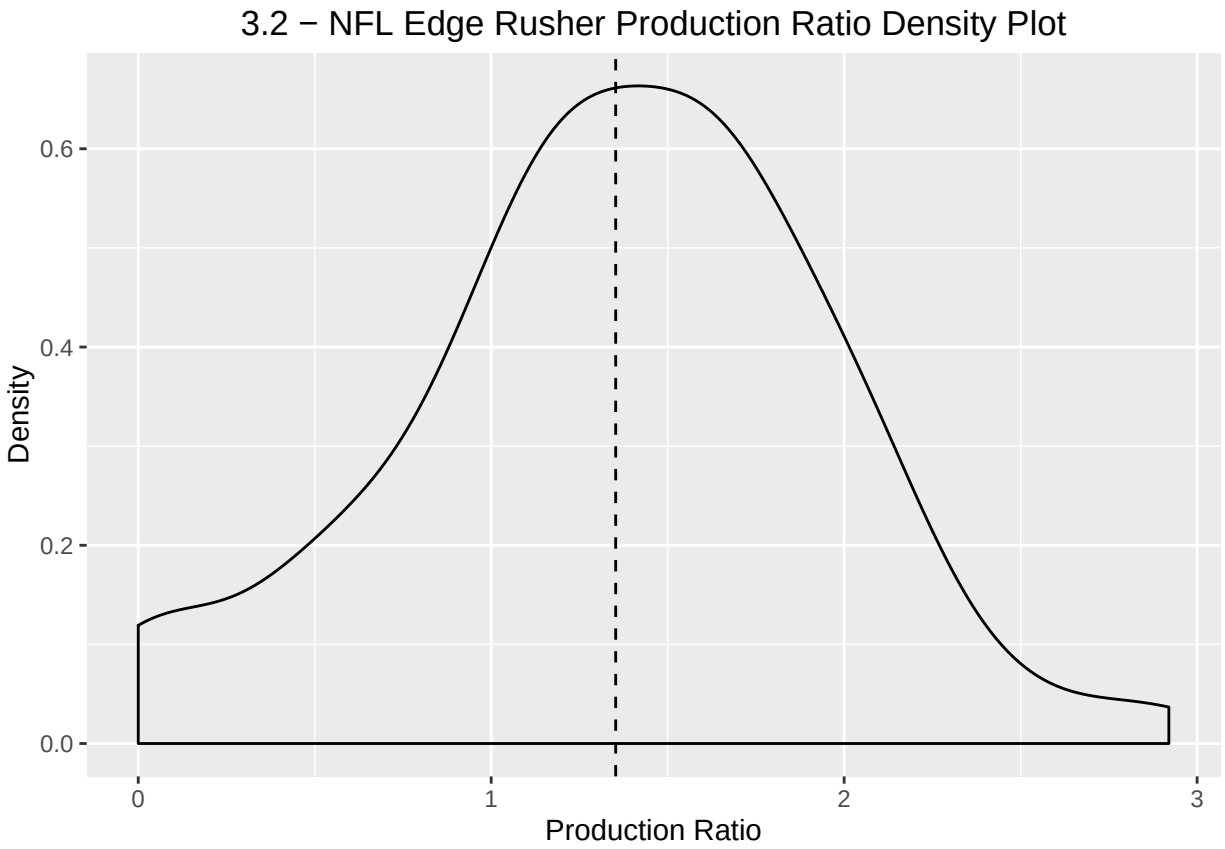
1. Many players - particularly players pre 2010 - did not have a player page
2. Of those players with pages - many of the players I was interested in shared names with other college players and there was no simple method to automate choosing the correct player.
3. Even the players with complete pages often had incomplete stats.

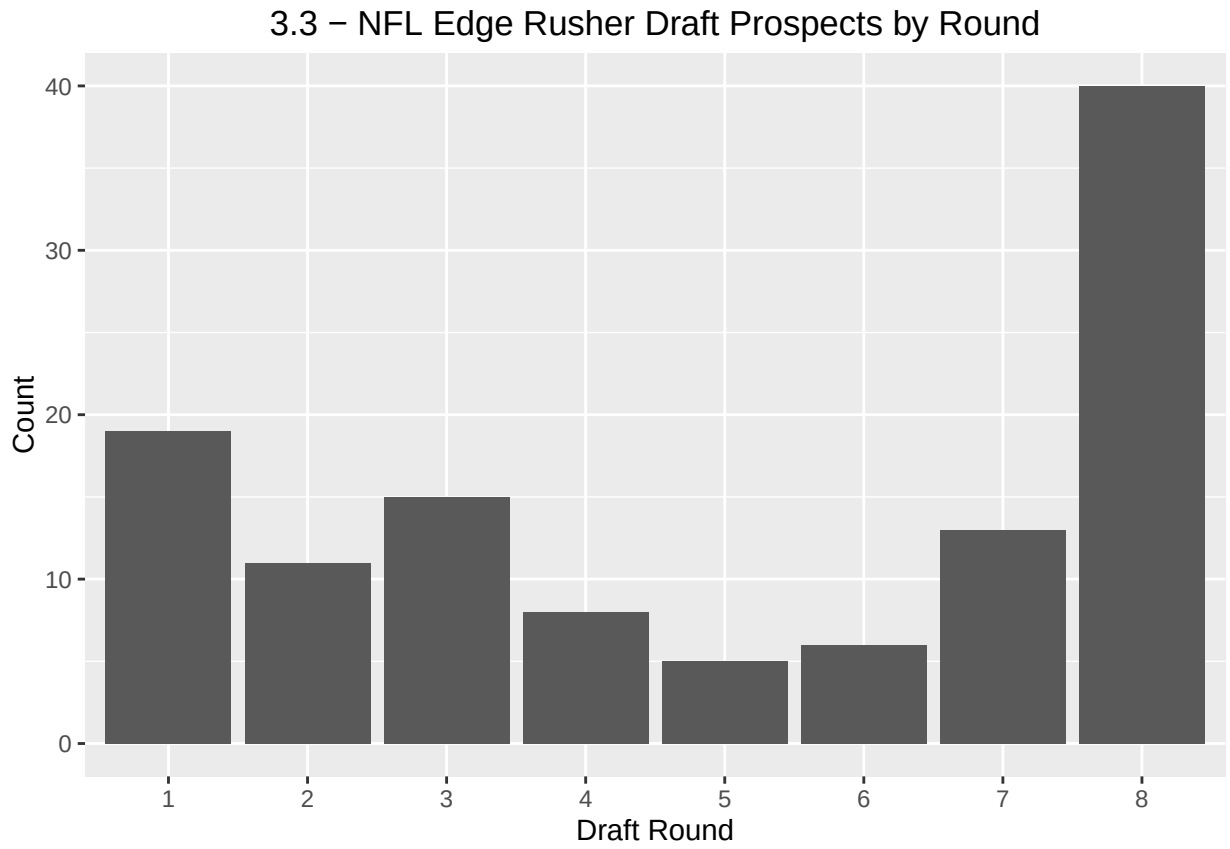
Because of this - even after some significant manual intervention - my sample of edge rushers dwindled to 143 players. This relatively small sample may prove problematic when it comes time to fit the model.

```
# read in and clean data
edge_college_stats <- read.csv('edge_college_stats.csv')
edge_college_stats$first_name <- as.character(edge_college_stats$first_name)
edge_college_stats$last_name <- as.character(edge_college_stats$last_name)
edge_college_stats$production_ratio <-
  as.numeric(edge_college_stats$production_ratio)
```

```
edge_college_stats$draft <- as.numeric(edge_college_stats$draft)
edge_college_stats <- filter(edge_college_stats, production_ratio < 3.0)
# filters out instances where an edge rusher does not have a page but a player
# from another position does and a ridiculous number results

#add players
Hardy <- c('Hardy', 'Greg', 6, 0, 0, 0, 1.64)
edge_college_stats <- rbind(edge_college_stats, Hardy)
Hughes <- c('Hughes', 'Jerry', 1, 0, 0, 0, 2.4)
edge_college_stats <- rbind(edge_college_stats, Hughes)
Irvin <- c('Irvin', 'Bruce', 1, 0, 0, 0, 1.98)
edge_college_stats <- rbind(edge_college_stats, Irvin)
Jackson <- c('Jackson', 'Malik', 5, 0, 0, 0, 1.18)
edge_college_stats <- rbind(edge_college_stats, Jackson)
Kerrigan <- c('Kerrigan', 'Ryan', 1, 0, 0, 0, 2.92)
edge_college_stats <- rbind(edge_college_stats, Kerrigan)
edge_college_stats$production_ratio <-
  as.numeric(edge_college_stats$production_ratio)
edge_college_stats$draft <- factor(edge_college_stats$draft)

# calculate z-scores and percent ranks
edge_college_stats$production_ratio_zscore <-
  round(scale(edge_college_stats$production_ratio,
              center = TRUE, scale = TRUE), 2)
edge_college_stats$production_ratio_percent_rank <-
  round(percent_rank(edge_college_stats$production_ratio), 2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.350   1.353   1.770   2.920
```


3.2 − NFL Edge Rusher Production Ratio Density Plot
Density
0.0
0.2
0.4
0.6
0
1
2
3
Production Ratio



The mean production ratio for my sample of edge rushers was 1.35 while the median was 1.35. The standard deviation was 0.6 and the distribution, shown in plot 3.2 above, of the ratios was normal.

I was hoping for a roughly uniform distribution of draft rounds, but alas that was not to be. Instead, as plot 3.3 above shows, I am left with a two tailed distribution with a lot of players from the first two rounds of the draft and a lot of undrafted players. This is primarily because players drafted in the first two rounds of the draft tend to be stars in college and are thus more likely to have player pages on the Sports Reference site. Meanwhile, the sheer number of undrafted players in the original combine database make them more likely to filter through. This non-uniform sampling is another thing to keep in mind when my final model is fitted.

3.3 - NFL Statistics and Final Data

Next, I combine the college stats/draft round and SPARQ scores into a single data frame.

```
# clean and join data
nfl_combine_edge_unite <- unite(nfl_combine_edge, name, c(first_name, last_name),
                                sep = ' ')
nfl_combine_edge_unite$sparq_zscore <-
  as.numeric(nfl_combine_edge_unite$sparq_zscore) # convert matrix to numeric
edge_college_stats_unite <- unite(edge_college_stats, name,
                                  c(first_name, last_name), sep = ' ')
# convert matrix to numeric
edge_college_stats_unite$production_ratio_zscore <-
  as.numeric(edge_college_stats_unite$production_ratio_zscore)
edge_combine_college_all <- left_join(nfl_combine_edge_unite,
                                      edge_college_stats_unite, by = 'name')
```

```
edge_combine_college <- inner_join(nfl_combine_edge_unite,
                                   edge_college_stats_unite, by = 'name')
edge_combine_college <- edge_combine_college %>%
  separate(name, c('first_name', 'last_name'), sep = ' ')

# check players without college stats
no_stats <- filter(edge_combine_college_all, is.na(production_ratio))
```

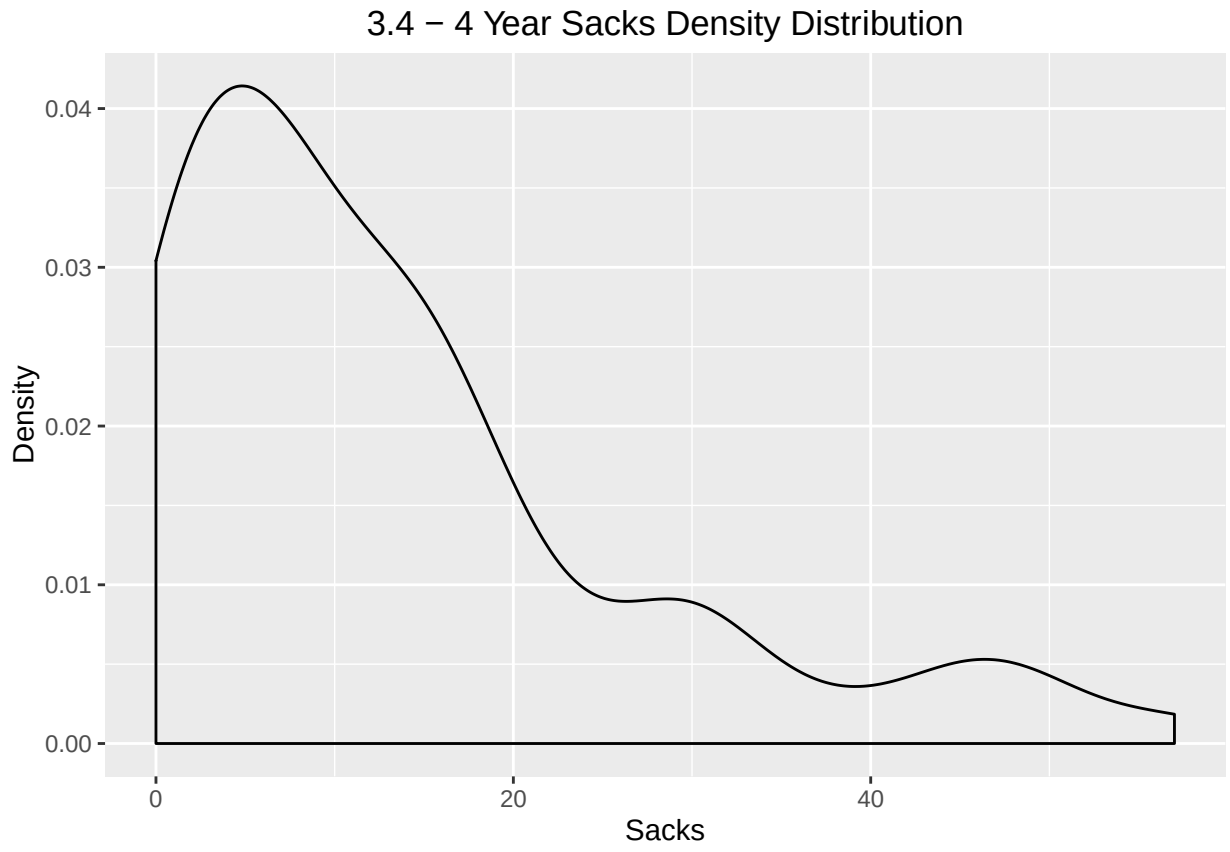
Finally, I combine my predictor variables and my outcome variable - sacks in first four NFL seasons - into a single dataframe from which I will fit my model. NFL stats were scraped from Pro Football Reference player pages.

```
# read in NFL sack data
edge_nfl_stats <- read.csv('edge_nfl_stats.csv')

# join NFL sack data and predictor variables and export for manual manipulation
edge_nfl_stats_unite <- unite(edge_nfl_stats, name, c(first_name, last_name),
                              sep = ' ')
edge_combine_college_unite <- unite(edge_combine_college, name,
                                    c(first_name, last_name), sep = ' ')
draft_production <-
  select(edge_combine_college_unite, c(year, name, draft, production_ratio,
                                       production_ratio_zscore,
                                       production_ratio_percent_rank, sparq,
                                       sparq_zscore, sparq_percent_rank))
edge_model_all <- left_join(edge_nfl_stats_unite,
                            draft_production, by = 'name')
write.csv(edge_model_all, 'edge_model_all.csv')
```

As I looked over this final dataset, I realized that - despite my best efforts - a number of off scrimmage linebackers (i.e. players who rarely rush the passer and accrue sacks) filtered through to this final dataset. In addition, there were a couple of players who gained weight and became defensive tackles prior to starting their NFL careers and even two players who switched to the offensive side of the ball! After dropping these players I had 67 complete cases (sacks, production ratio, SPARQ) with which to build my model. First, I read this manually manipulated data back into R and coded draft position as a factor variable (i.e. dummy variable) in preperation for its inclusion in my final regression model.

```
# read in complete cases and code draft position as factor variable
edge_model_manual <- read.csv('edge_model_manualv2.csv')
edge_model_manual$draft <- factor(edge_model_manual$draft)
```

Next, I examine the probability distribution for sacks.



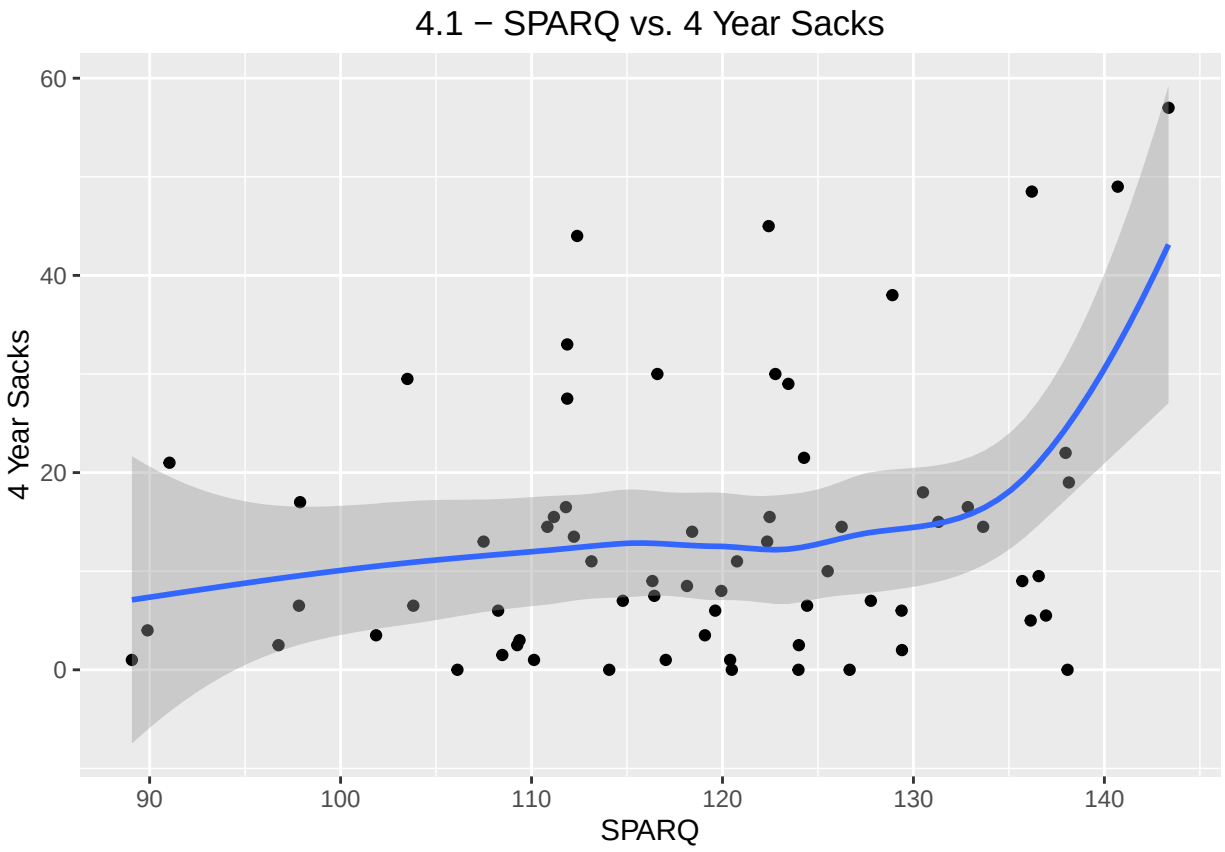

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    3.50    9.25   13.67   17.25   57.00
```

Unsurprisingly, sacks were not normally distributed - they showed a strong positive skew of ~1.39 indicating that many players had few sacks while a few very succesful ones had a lot of sacks. This is the reality of the NFL draft - most prospects do not succeed. The mean number of sacks for a prospect in my database was 13.67 and the median was 9.25.

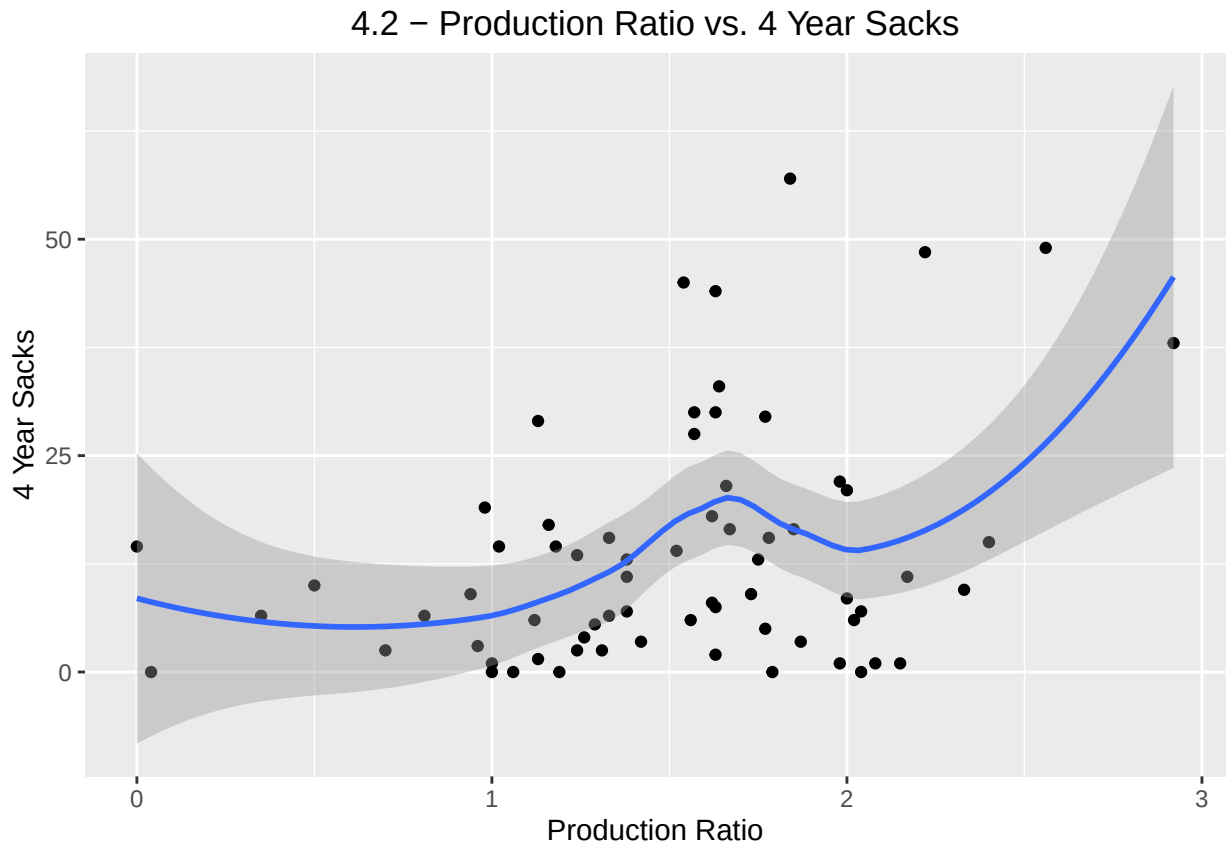
4 - 4 Year Sack Model

4.1 - Predictor Variables vs. Sacks

It is now time to build my final model. As a preliminary I will examine the relationship between each of my predictor variables and my outcome variable in turn. I begin with the relationship between SPARQ score and first four year sacks.

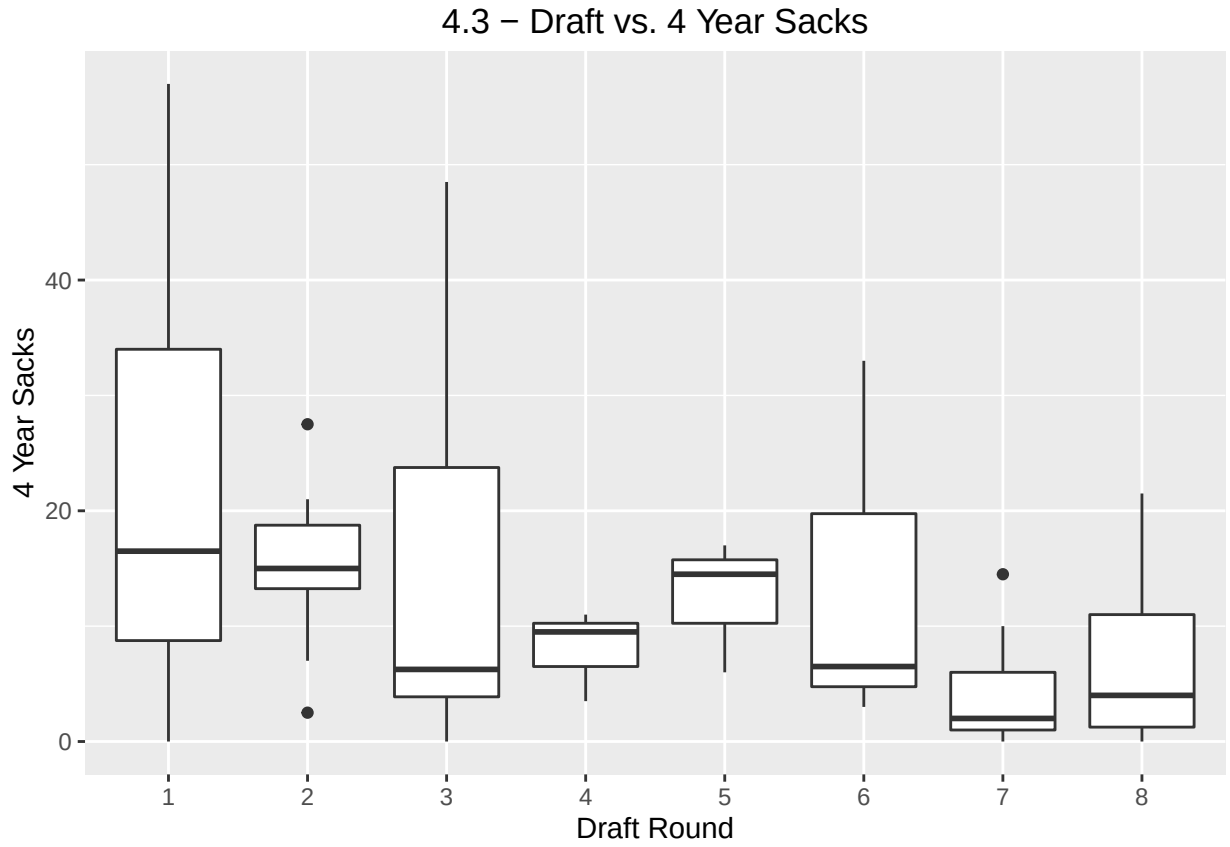


Given the amount of time that I spent back calculating a formula for the SPARQ score, to say that the relative lack of a correlation between SPARQ score and sacks exhibited in plot 4.1, above, is a dissapointment is a serious understatement. Modelling sacks as a simple function of SPARQ results in a $R^2 = 0.076$ and a non-significant p-value for the regrsssion of $p = 0.09$. However, sacks do not appear to be homoscedastic with respect to SPARQ; as you get to the higher reaches of SPARQ scores (i.e. the truly elite athletes) there appears to be an increasingly positive association between SPARQ and sacks. This can best be modelled as an exponential relationship and doing so markedly increases the predictive power of the regression: $R^2 = 0.173$ with a highly significant $p = 0.0004$. While there are certainly exceptions (lookng at you Vernon Gholston) it makes sense that in a league composed of elite athletes only the truly exceptional derive a measurable benefit from pure athleticism.



The relationship between production ratio and sacks - though not particularly strong - is somewhat clearer than that of SPARQ and sacks. Very few players with a production ratio < 1 have any success, while most players with production ratios > 2 have at least moderate (10+ sacks) success as pros. Unfortunately, production ratio does not appear to be particualrly predictive in the 1-2 range where most players fall. Nevertheless, it's ability to highlight players on the tails as either likely failures or successes makes it a worthy metric to include in the final model.

In looking at the line of best fit, a low degree polynomial is clearly the best description of the relationship. In particular, modelling sacks as a function production ratio as a third degree polynomial results in a $R^2 = 0.132$ at a high signficance level $p = 0.028$.



The final variable that I am going to include in my sack model is the round in which the player was drafted. I intend for this primarily to serve to as a stand-in for a player's qualitatively scouted ability, though of course NFL teams take into account a players physical attributes and college production when drafting so there will be some unavoidable overlap.

As a categorical variable, draft will be included in my model as a dummy variable, such that being drafted in the first round will be considered the base case and being drafted in subsequent rounds will be modelled as a deprature from that base case.

As the boxplots in plot 3.3, above, indicate draft position generally follows the expected pattern with a higher draft round resulting in higher average sack totals. Draft round loses some of its predictive power in the mid-to late rounds though this may be an artifact of a relatively small sample size of players from these rounds (these "fringe" prospects were much less likely to have college stats available).

Overall, draft position has the strongest correaltion with 4 year sack totals of my three predictor variables: $R^2 = 0.241$ highly significant at $p = 0.016$. This is not unexpected as, in addition into the millions of dollars that NFL teams pour into scouting, draft position also reflects a player's athletic chracteristics and college production.

The correlation between draft position and future production provides an excellent baseline against which to judge my model. If the model does not do appreciably better than the "scouting consensus" - information that can be gleaned by reading any number of NFL draft websites/publications - then it is not worth the (digital) paper that it is printed on. However, if it does add appreciable predictive power then it is an indication that this or a similar modelling approach (using slightly different predictors, for example) could be useful evidence in deciding between NFL edge rushing prospects.

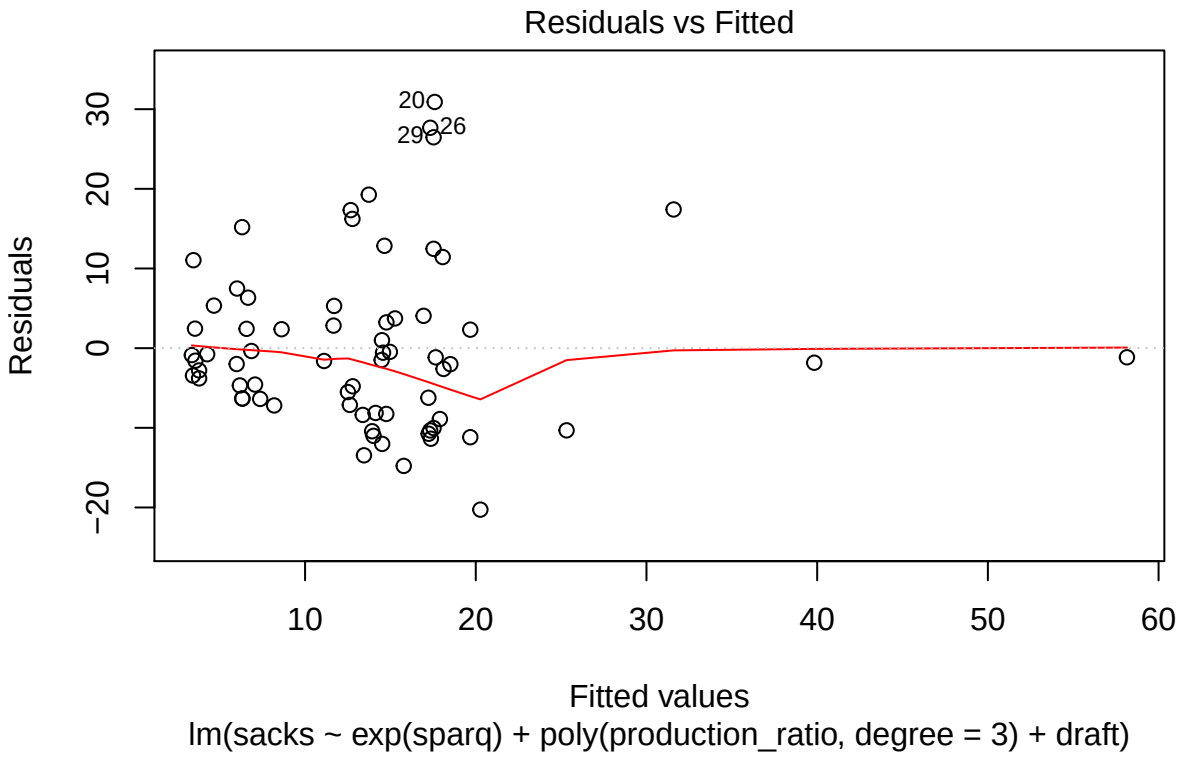
4.2 - 4 Year Sack Model Summary and Discussion

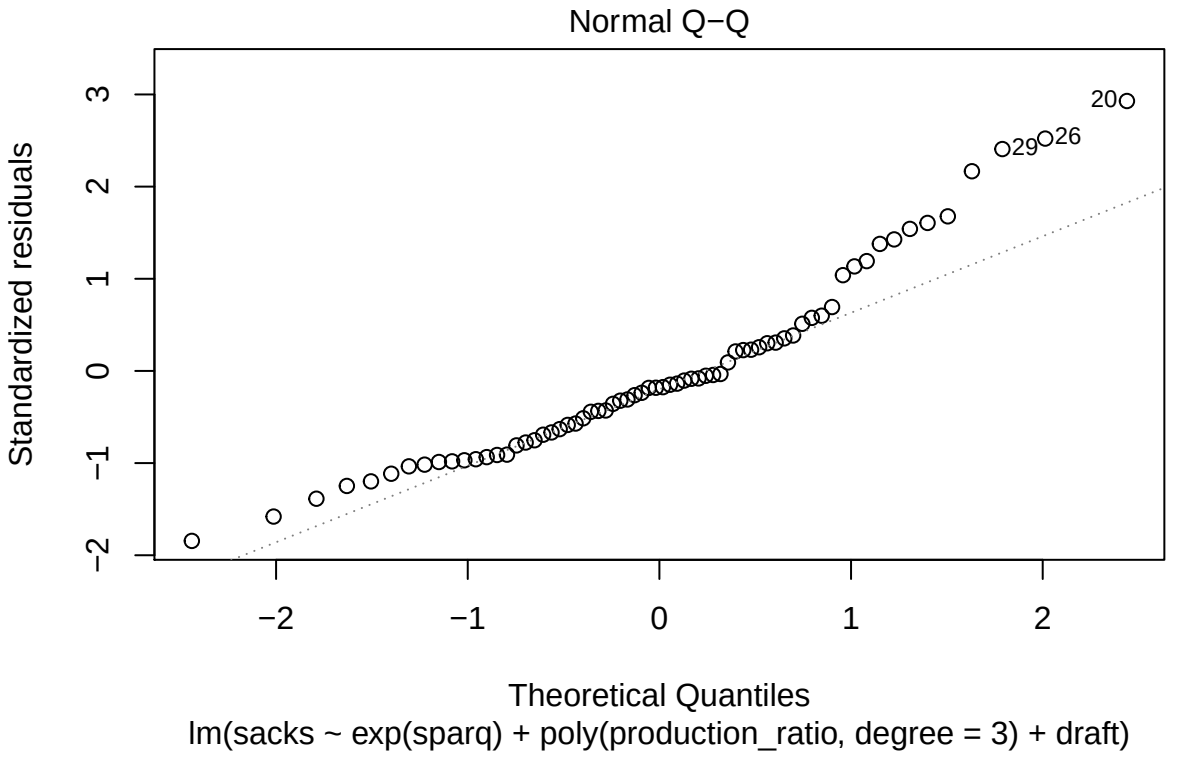
Having examined the relationships between my three predictor variables and my outcome variable, it is finally time to model 4 year sack totals as a function of SPARQ scores, college production ratio, and NFL draft position.

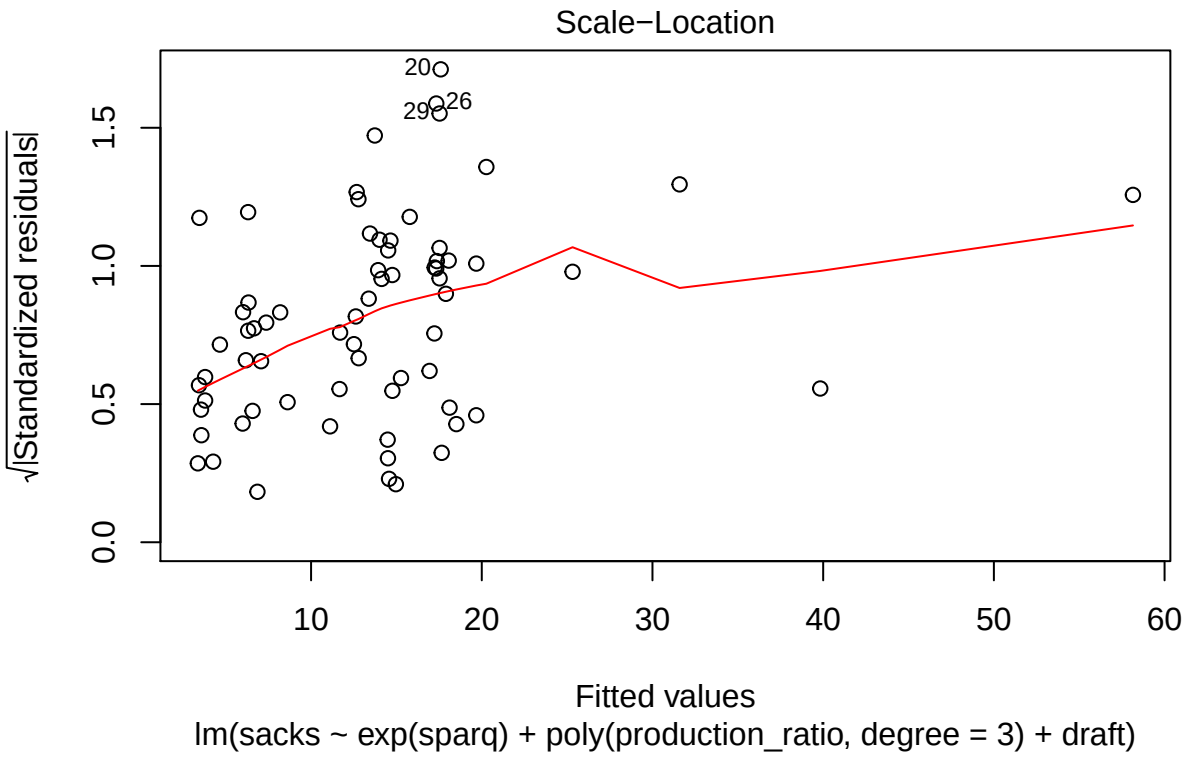
```
# build and summarize final sack model
sack_model <- lm(sacks ~ exp(sparq) + poly(production_ratio, degree = 3) +
                   draft, data = edge_model_manual)
summary(sack_model)
```

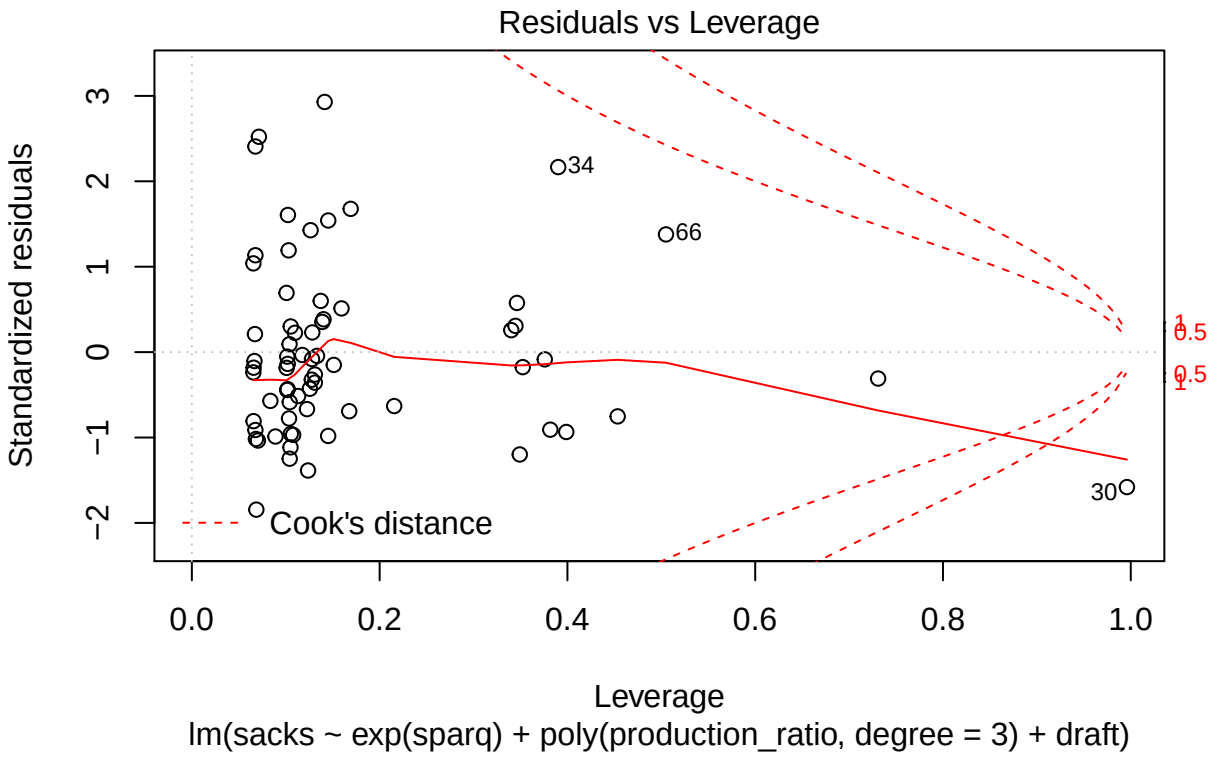
```
##
## Call:
## lm(formula = sacks ~ exp(sparq) + poly(production_ratio, degree = 3) +
##     draft, data = edge_model_manual)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.269  -7.138  -1.593   3.813  30.905
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         1.880e+01  2.921e+00   6.436 2.95e-08
## exp(sparq)                          2.179e-61  6.457e-62   3.375  0.00135
## poly(production_ratio, degree = 3)1  1.619e+01  1.426e+01   1.135  0.26122
## poly(production_ratio, degree = 3)2  1.842e+01  1.178e+01   1.564  0.12352
## poly(production_ratio, degree = 3)3  1.113e+01  1.163e+01   0.957  0.34247
## draft2                              -2.737e+00  4.711e+00  -0.581  0.56353
## draft3                              -4.714e+00  4.573e+00  -1.031  0.30706
## draft4                              -1.296e+01  7.134e+00  -1.817  0.07456
## draft5                              -5.737e+00  7.255e+00  -0.791  0.43242
## draft6                              -3.825e+00  7.666e+00  -0.499  0.61981
## draft7                              -1.395e+01  5.275e+00  -2.646  0.01056
## draft8                              -1.131e+01  4.943e+00  -2.287  0.02597
##
## (Intercept)                         ***
## exp(sparq)                          **
## poly(production_ratio, degree = 3)1
## poly(production_ratio, degree = 3)2
## poly(production_ratio, degree = 3)3
## draft2
## draft3
## draft4                              .
## draft5
## draft6
## draft7                              *
## draft8                              *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.39 on 56 degrees of freedom
## Multiple R-squared:  0.4081, Adjusted R-squared:  0.2918
## F-statistic:  3.51 on 11 and 56 DF,  p-value: 0.0008894
```

```r
plot(sack_model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(sacks ~ exp(sparq) + poly(production_ratio, degree = 3) + draft)
20
29
26


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sacks ~ exp(sparq) + poly(production_ratio, degree = 3) + draft)
20
29
26


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
26
29
10
20
30
40
50
60
Fitted values
lm(sacks ~ exp(sparq) + poly(production_ratio, degree = 3) + draft)



The table and plots above summarize my final regression model. The $R^2 = 0.408$ is a marked improvement over looking at a player's draft projection alone and is highly signficant $p < 0.001$. For a simple model with three inputs, explaining almost 41% of the variance in player's sack production is quite an accomplishment.

However, the model is not without faults. In particular, the residual standard error is quite high at 11.39. This means that the 95% confidence interval for any given prediction will be on the order of +/- 22 sacks which is very high considering a star player will accrue ~32+ sacks (i.e. 8+ sacks a season) in 4 seasons. Clearly, this model can and should be improved on.

I think my general framework for the model - to include predictor variables for athleticism, production, and qualitatively scouted ability - is conceptually sound and potentially a fruitful one. However, I certainly think there is room to improve the predictions derived from within each of these buckets. In particular I would suggest the following four follow-ups to improve upon this model:

1. Get more complete cases. Lots of players had to be dropped because of a lack of college stats and/or combine numbers. This will be less of a problem going forwards as this information is becoming more readily available online by the day but in the interim probably the only real fix for this issue is manual effort to fill out required variables for the (many) problem players.

2. Focus more on the athletic traits that make for succesful pass rushers versus the overall athleticism measured by the SPARQ score. In particular, an attempt to measure an athlete's explosive power as is done in Football Outsiders Sackseer model would be more relevant and could prove more predictive for edge rushers.

3. Examine college statistical profiles more granually for more predictive indicators. Production ratio is probably not the best measure of a player's college production - as I demonstrated above it lacks predictive value in the 1-2 range where most players ultimately fall. A more thorough examination of which stats actually correlate with NFL success would potentially improve the model.

4. Determine whether or not a categorical variable to indicate level of competition should be included. There is a big difference in the level of competition between divisions of college football, big differences between Power 5 competition and the rest of college football, and even big differences between the Power 5 conferences. Attempting to credit players who played the toughest competition in some manner may improve predictive power.

5 - 2016 Edge Rusher 4 Year Sack Projections

What's the fun of having a shiny new (if imperfect) car if you don't take it for a spin? What follows is my projection for top 5 edge rushing prospects in the 2016 draft class.

First, I select only 2016 NFL edge rushing prospects from the NFL combine data, manually add missing combine data (from Pro-Days), and calculate SPARQ scores, z-scores, and percent ranks.

```
# prepare 2016 edge rusher combine and pro-day data
# filter 2016 DE
prospects_2016 <- nfl_combine %>% filter(pos == 'DE' & year == 2016)
prospects_2016 <- rbind(prospects_2016, nfl_combine[95, ]) # add Leonard Floyd
prospects_2016 <- select(prospects_2016, -c(college, pos, height, wonderlic,
                                            broad_jump, three_cone))
prospects_2016[5, 5] <- 21 # DeForest Buckner Bench Press
prospects_2016[10, 5:7] <- c(22, 30.5, 4.44) # Kevin Dodd Pro-Day
prospects_2016[15, 5] <- 20 # Shaq Lawson Bench Press
prospects_2016[33, 5:7] <- c(16, 39.5, 4.34) # Leonard Floyd Pro-Day
prospects_2016[28, 7] <- 4.50 # Charles Tapper Shuttle
# eliminate incomplete cases
prospects_2016 <- prospects_2016[complete.cases(prospects_2016), ]
colnames(prospects_2016) <- c('year', 'name', 'weight', 'forty_yard',
                              'bench_press', 'vertical_jump', 'shuttle')

# fix forty_yard class
prospects_2016$forty_yard <- as.character(prospects_2016$forty_yard)
prospects_2016$forty_yard <- as.numeric(prospects_2016$forty_yard)

# join with all DE for centering z-score, percent rank
DE_complete <- nfl_combine %>% filter(pos == 'DE') %>%
  select(-c(college, pos, height, wonderlic, broad_jump, three_cone))
colnames(DE_complete) <- c('year', 'name', 'weight',
                           'forty_yard', 'bench_press', 'vertical_jump',
                           'shuttle')
DE_complete$bench_press <- as.numeric(DE_complete$bench_press)
DE_complete$forty_yard <- as.character(DE_complete$forty_yard)
DE_complete$forty_yard <- as.numeric(DE_complete$forty_yard)
DE_complete <- DE_complete[complete.cases(DE_complete), ]
prospects_2016 <- union(DE_complete, prospects_2016)

# calculate sparq, sparq z-score, and sparq percent rank
prospects_2016$sparq <- round(predict(sparq_fit, newdata = prospects_2016), 2)
prospects_2016$sparq_zscore <- round(scale(prospects_2016$sparq, center = TRUE,
                                           scale = TRUE), 2)
prospects_2016$sparq_percent_rank <- round(percent_rank(prospects_2016$sparq),
                                           3)
prospects_2016$sparq_z <- as.numeric(prospects_2016$sparq_zscore)
```

```
prospects_2016 <- prospects_2016 %>% select(-sparq_zscore) %>%
  filter(year == 2016)
```

With the SPARQ scores in hand, it is time to gather college statistics and draft round.

```
# prepare csv for Python scraping
prospects_2016 <- separate(prospects_2016, name, c('first_name', 'last_name'),
                           sep = ' ')
write.csv(prospects_2016, 'prospects_2016.csv')

#load college stats
prospects_2016_college <- read.csv('prospects_2016_college.csv')
prospects_2016_college <- unite(prospects_2016_college, 'name',
                                c(first_name, last_name), sep = ' ')
prospects_2016_college <- select(prospects_2016_college,
                                 c(name, draft, production_ratio))
prospects_2016 <- unite(prospects_2016, 'name', c(first_name, last_name),
                        sep = ' ')
prospects_2016_sparq <- select(prospects_2016,
                               c(name, sparq, sparq_z, sparq_percent_rank))
```

Finally, with the three predictor variables gathered, it is time to merge the data into a single data frame and predict the prospects 4 year sack production.

```
# join and munge data
prospects_2016_final <- left_join(prospects_2016_college, prospects_2016_sparq,
                                  by = 'name')
prospects_2016_final$draft <- factor(prospects_2016_final$draft)

# apply model and sort by sacks
prospects_2016_final$sacks <- round(predict(sack_model,
                                            newdata = prospects_2016_final), 1)
prospects_2016_final <- arrange(prospects_2016_final, -sacks)
```

Top 5 2016 Edge Rushing Prospects

1. Carl Nassib (Cleveland Browns, Pick 65) - *24.3 sacks*

SPARQ Percent Rank: 31.4%

Production Ratio: 2.57

Carl Nassib is a fascinating case. Despite incredible production in a major conference (Big 10) he lacked both the pedigree (walk-on at Penn State) and athleticism of an elite prospect. The new regime in Cleveland - by far the most analytically sophisticated in the NFL - did not overlook his extreme production and took him perhaps a round earlier then he was expected to go.

2. Joey Bosa (San Diego Chargers, Pick 3) - *20.4 sacks*

SPARQ Percent Rank: 64.3%

Production Ratio: 2.07

The consensus top edge rushing prospect lacks elite athleticism and saw his production sharply decline in a somewhat dissapointing junior season (early entry to the draft). He is a player that scouts love much more than the numbers do.

3. Shaq Lawson (Buffalo Bills, Pick 19) - *19.1 sacks*

SPARQ Percent Rank: 76.0%

Production Ratio: 1.93

Shaq Lawson is another interesting case. He combines well above average athleticism for the position with incredible productivity in his one year as a starter. He is punished by the model for his sophmore season when he was a backup to former top-10 pick Vic Beasley of the Atlanta Falcons. But because the model cannot tell that he was a backup, it dings him for his sophmore production. This is a limitation of the model and not the player - I think he is probably the best prospect from the 2016 class.

4. Leonard Floyd (Chicago Bears, Pick 9) - *17.5 sacks*

SPARQ Percent Rank: 77.0%

Production Ratio: 1.15

Floyd was drafted as 3-4 OLB in coordinator Vic Fangio's defense despite never being a particularly productive pass rusher in college. At least he exhibits well above athleticism.

5. DeForest Buckner (San Francisco 49ers, Pick 7) - *17.4 sacks*

SPARQ Percent Rank: 26.9%

Production Ratio: 1.59

Buckner is not - and was not drafted to be - a classic edge rusher. Rather, at 6'7" 290 lbs., he is the quintessential 5-Technique, two-gapping DE in 3-4 defense. In a sense his responsibilities are more akin to a DT and his SPARQ percent rank would not be so terrible if compared to that group. Still, his 21 bench reps are a major red flag for a player of his size. High risk for a top 10 pick.

Summary

As a whole this group is decidedly underwhelming. The Browns will probably be quite happy if they get 24+ sacks from 3rd rounder Nassib, but all of the top 10 picks project as middling NFL pass rushers. An NFL team hopes to get a Pro Bowler with a top-10 pick but no edge rusher in the 2016 class profiles as such. That these players were probably overdrafted is a testament to the premium that NFL teams apply to pass rushers.