How Diverse Body Shapes in American Football Athletes Address the NFL's 40 Yard Dash

Emile Therrien

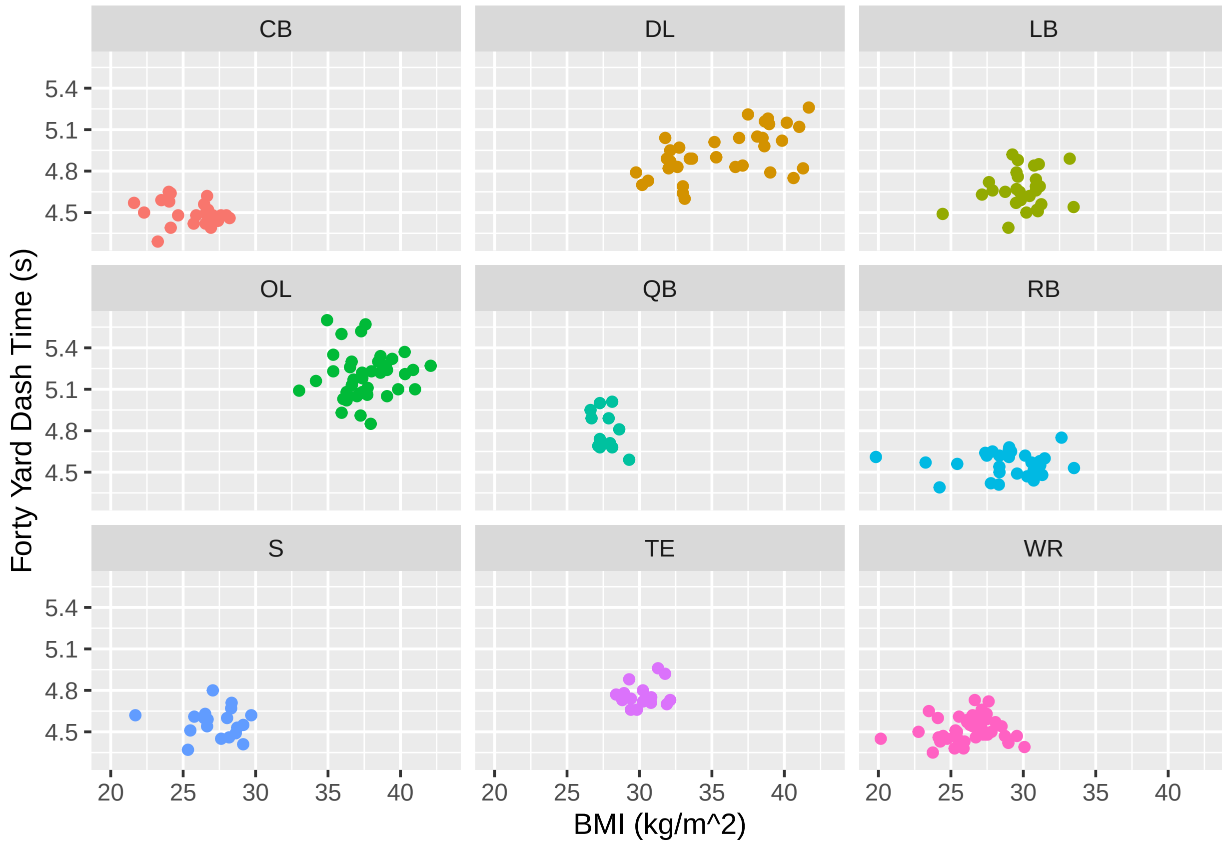
11/08/2020

Introduction

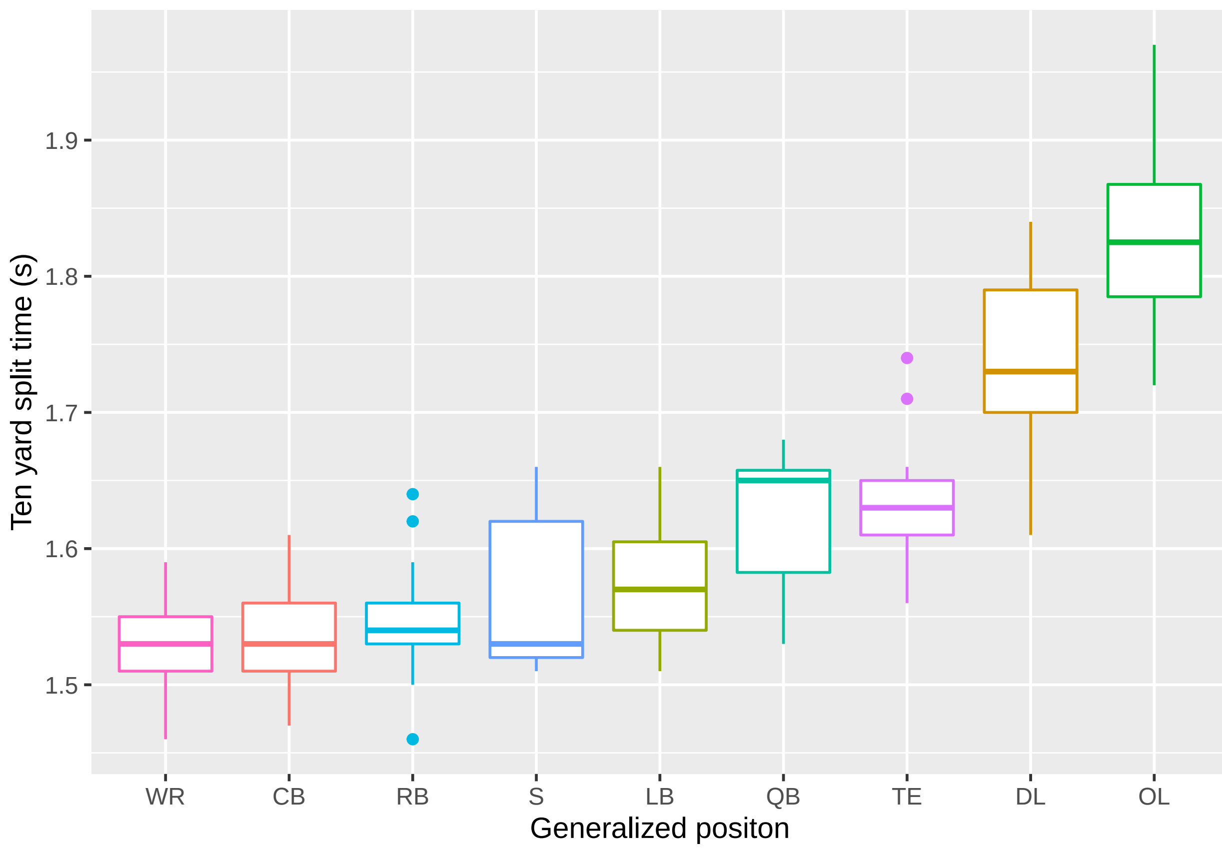
Does body proporitons determine position on the field or is it BMI? * use leg length and BMI: is there is a suggested determinant for how fast the athlete runs the 40? ** Does BMI and/or leg length determine how fast the 40 or first ten yard split is ran? *** Use ovr forty time as a metric the NFL uses to measure athlete speed *** Use ten split time to determine the explosiveness of the athlete

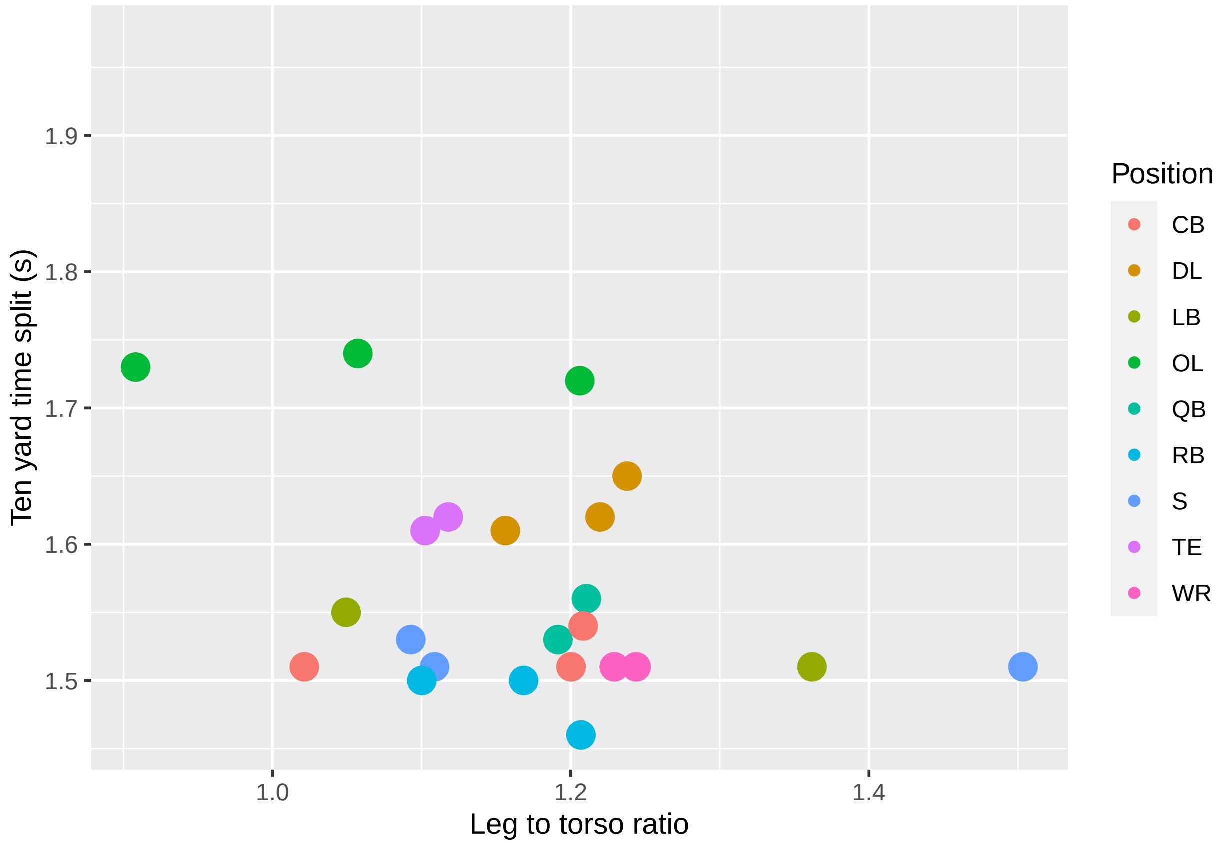
Data Plots

```
## Observations: 308
## Variables: 8
## $ name        <chr> "Zuniga, Jabari", "Young, Chase", "Woodward, David", "W...
## $ POS         <chr> "DE", "DE", "OLB", "TE", "DE", "TE", "OT", "DT", "OT", ...
## $ bmi         <dbl> 32.99737, 31.30548, 29.52993, 31.76957, 30.59400, 28.93...
## $ fortyTime   <dbl> 4.64, NA, 4.79, 4.92, 4.73, 4.78, 4.85, 4.90, 5.32, 4.6...
## $ tenTimeOvr  <dbl> 1.61, NA, 1.62, 1.74, 1.70, 1.63, 1.72, 1.76, 1.87, 1.5...
## $ Height      <dbl> 75, 77, 74, 76, 77, 77, 77, 76, 79, 79, 77, 75, 75, 76,...
## $ legHgtRatio <dbl> 1.156159, NA, NA, NA, NA, NA, 1.206093, NA, NA, NA, NA,...
## $ genPos      <chr> "DL", "DL", "LB", "TE", "DL", "TE", "OL", "DL", "OL", "...
```



CB
DL
LB
OL
QB
RB
S
TE
WR
5.4
5.1
4.8
4.5
20
25
30
35
40
Forty Yard Dash Time (s)
BMI (kg/m^2)


Ten yard split time (s)
1.9
1.8
1.7
1.6
1.5
WR
CB
RB
S
LB
QB
TE
DL
OL
Generalized positon



Statistical Analyses

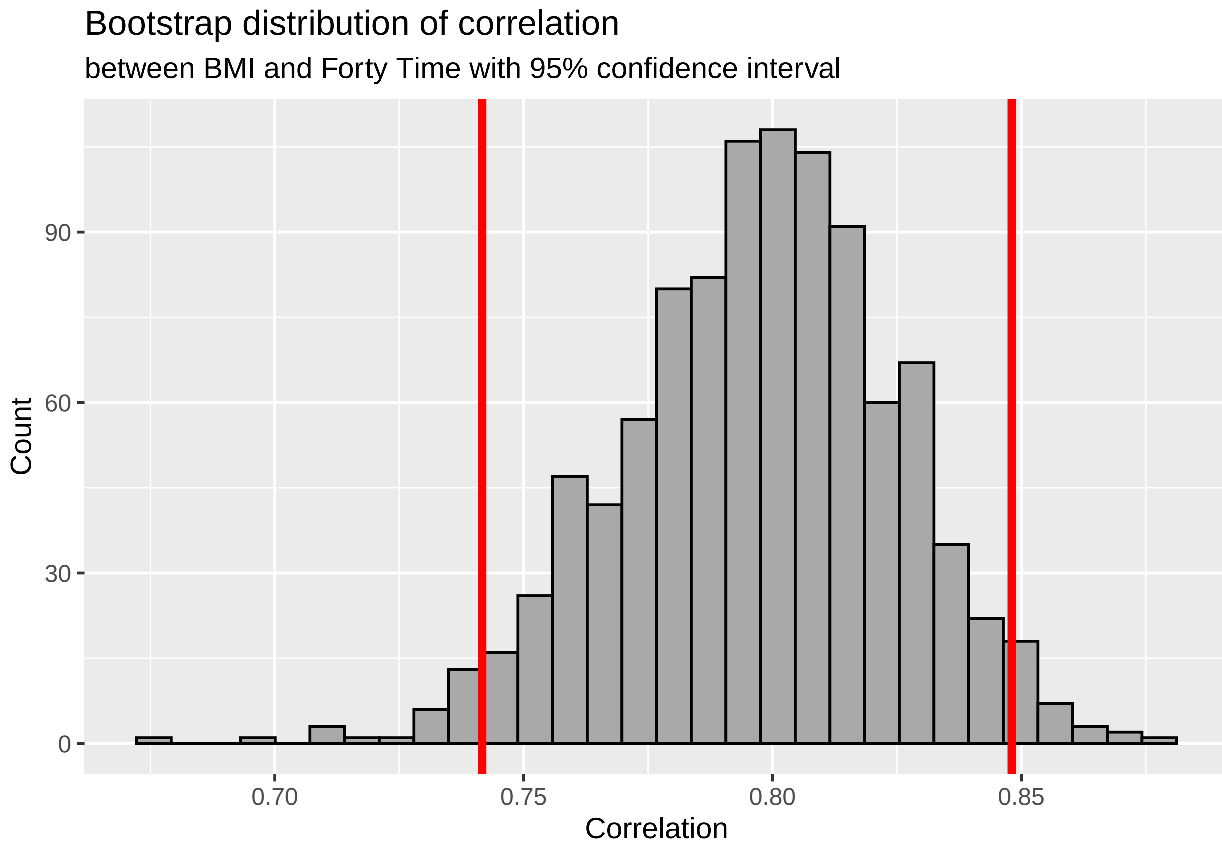
BMI v Forty Time

$\alpha = 0.05$

H_0: BMI and forty time do not have a strong (≤ 0.5), postive correlation.

H_A: BMI and forty time do have a strong (>0.5), postive correlation.

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1 0.742 0.848
```



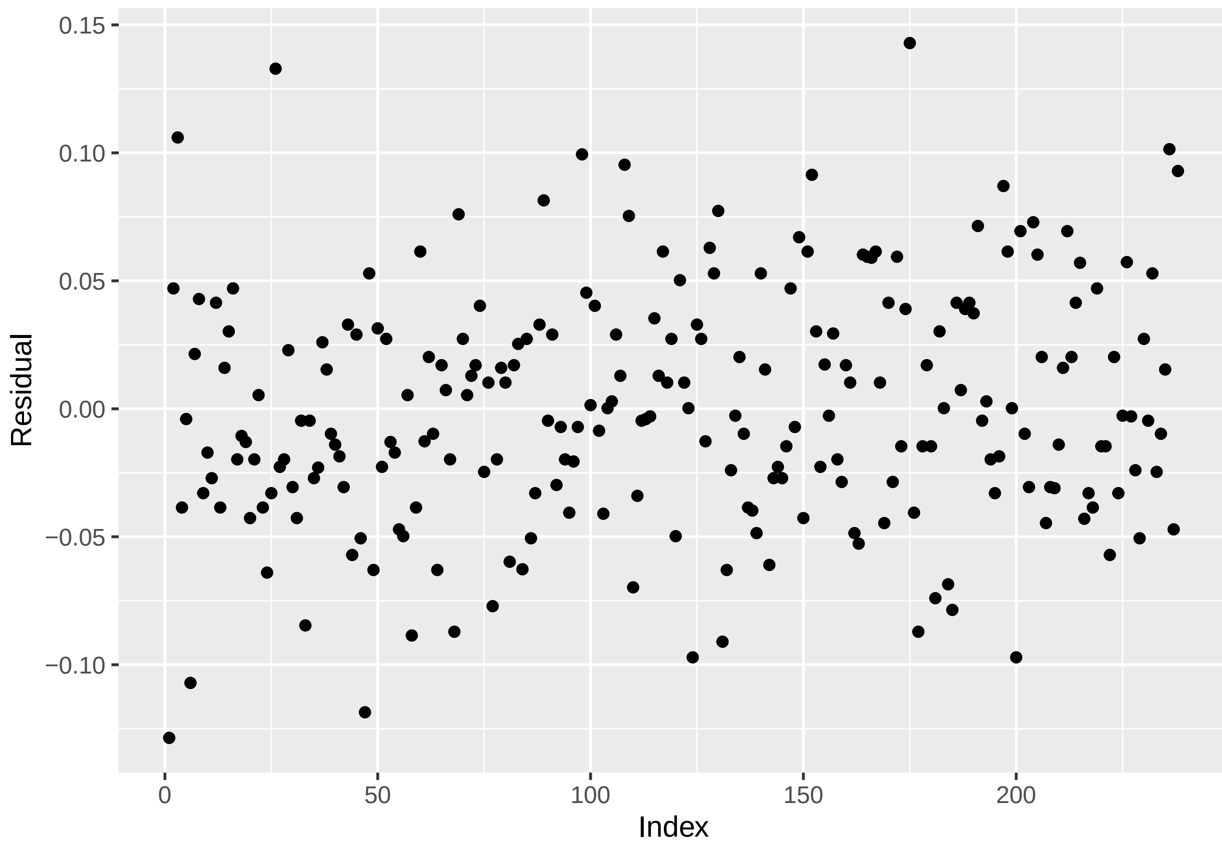
Based on an α level of 0.05, we are 95% confident that the true population coefficient for BMI and forty times is between (0.7485, 0.8436). There is enough evidence to reject the null hypothesis that there is not a strong, positive correlation between BMI and forty time.

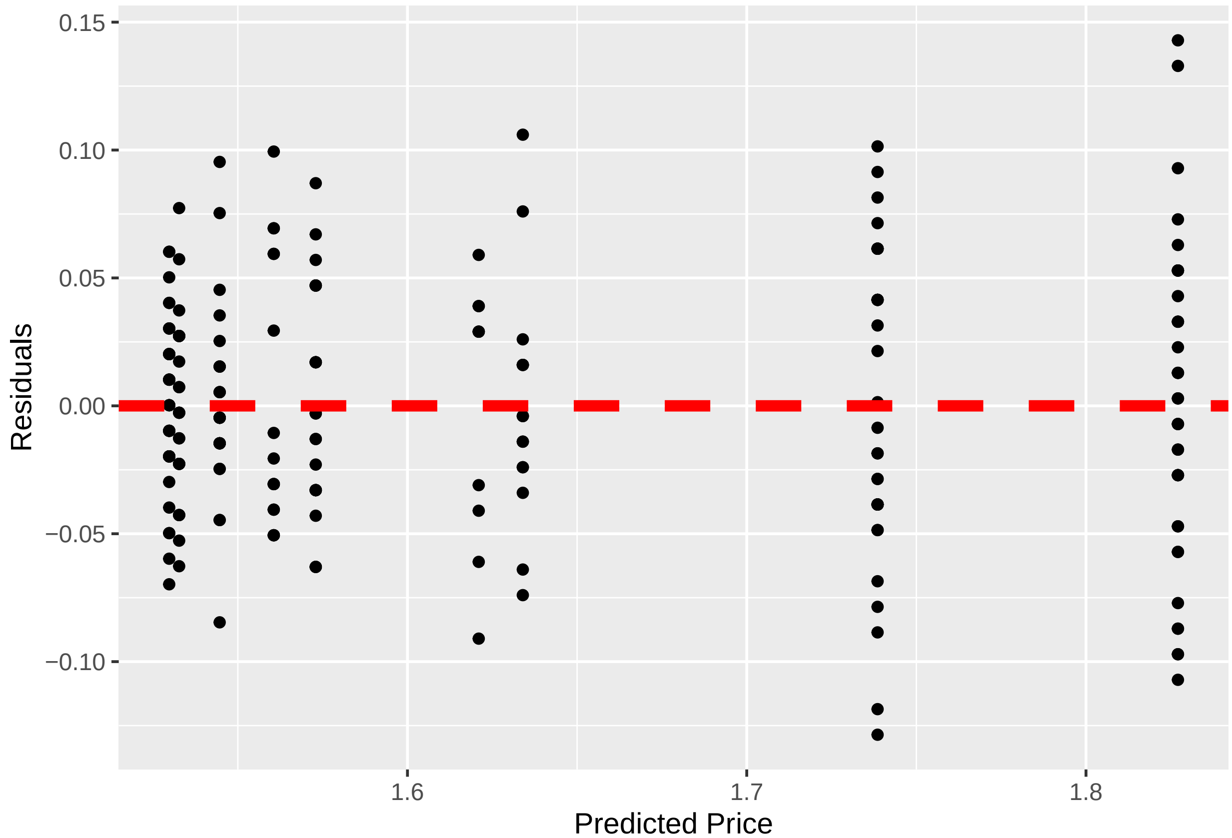
Position v Ten Time

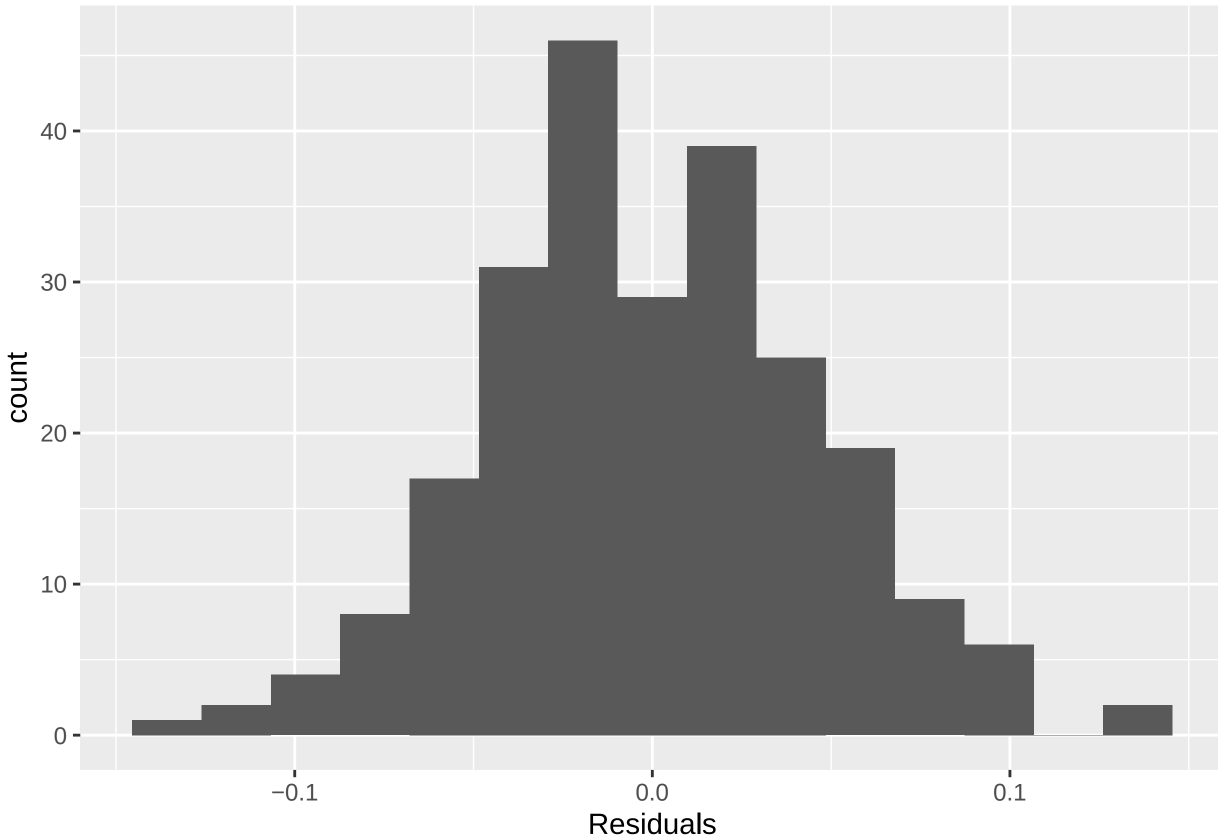
What is the relationship between positions and the frst ten yard splits? All conditions met, can use linear inference to determine relationship.

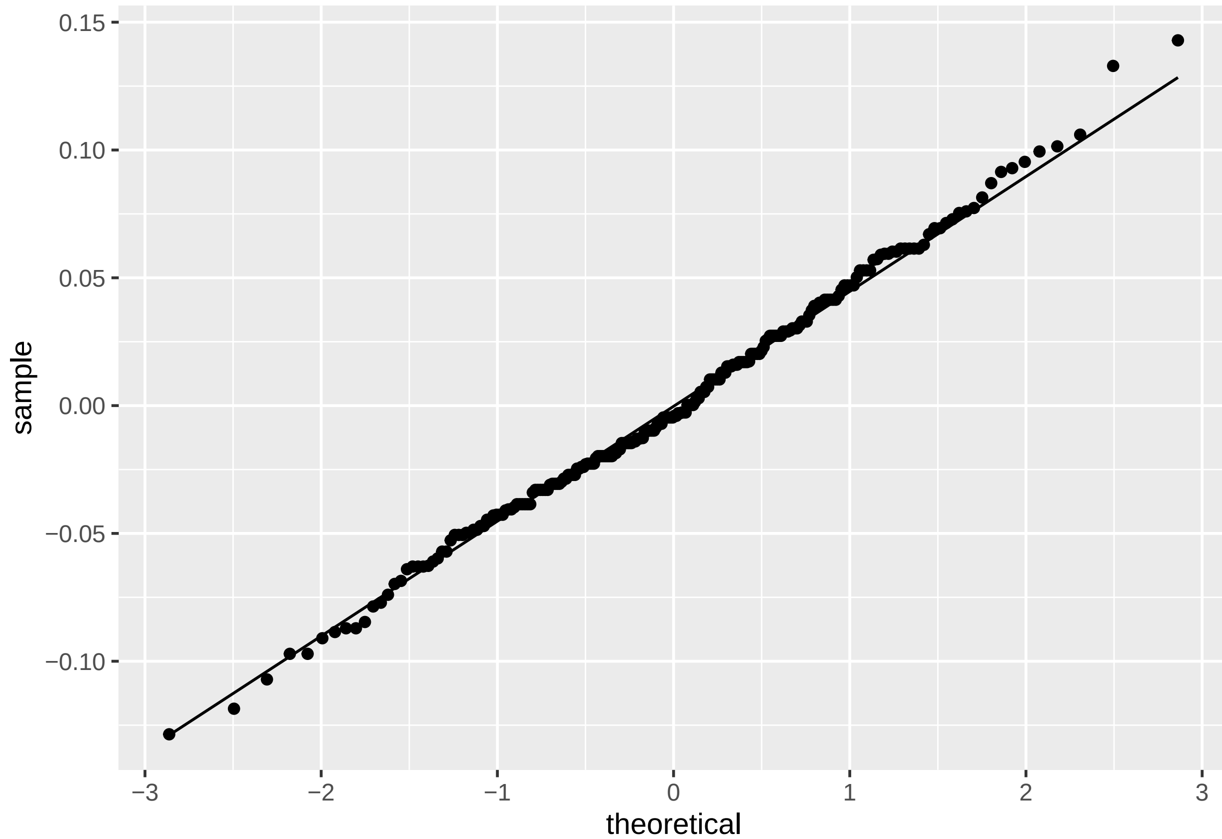
```
## # A tibble: 9 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)  1.53      0.00925   166.    1.85e-240
## 2 genPosDL     0.206     0.0122     16.9   5.09e- 42
## 3 genPosLB     0.0403    0.0130      3.11  2.12e-  3
## 4 genPosOL     0.294     0.0120     24.5   5.16e- 66
```

```
## 5 genPosQB     0.0883    0.0175      5.03  9.78e-  7
## 6 genPosRB     0.0120    0.0128      0.931 3.53e-  1
## 7 genPosS      0.0279    0.0147      1.90  5.91e-  2
## 8 genPosTE     0.101     0.0153      6.63  2.44e- 10
## 9 genPosWR    -0.00293   0.0118     -0.249 8.04e-  1
```




0.15
0.10
0.05
0.00
−0.05
−0.10
Residuals
1.6
1.7
1.8
Predicted Price


40
30
20
10
0
count
−0.1
0.0
0.1
Residuals



$\hat{Split}$ = 1.533 (CB) + 0.206 (DL) + 0.040 (LB) + 0.294 (OL) + 0.088 (QB) + 0.012 (RB) + 0.028 (S) + 0.101 (TE) - 0.003 (WR)

Leg:Torso and Ten Time Split

```
#obs correlation between BMI and forty time
dimension_analysis = combine2 %>%
  summarize(
    sdTenTime = sd(tenTimeOvr, na.rm=TRUE),
    sdRatio = sd(legHgtRatio, na.rm=TRUE),
    covar = cov(tenTimeOvr, legHgtRatio, use ="complete.obs")
  ) %>%
  mutate(
    sample_correlation = (covar/(sdTenTime*sdRatio))
  ) %>%
```

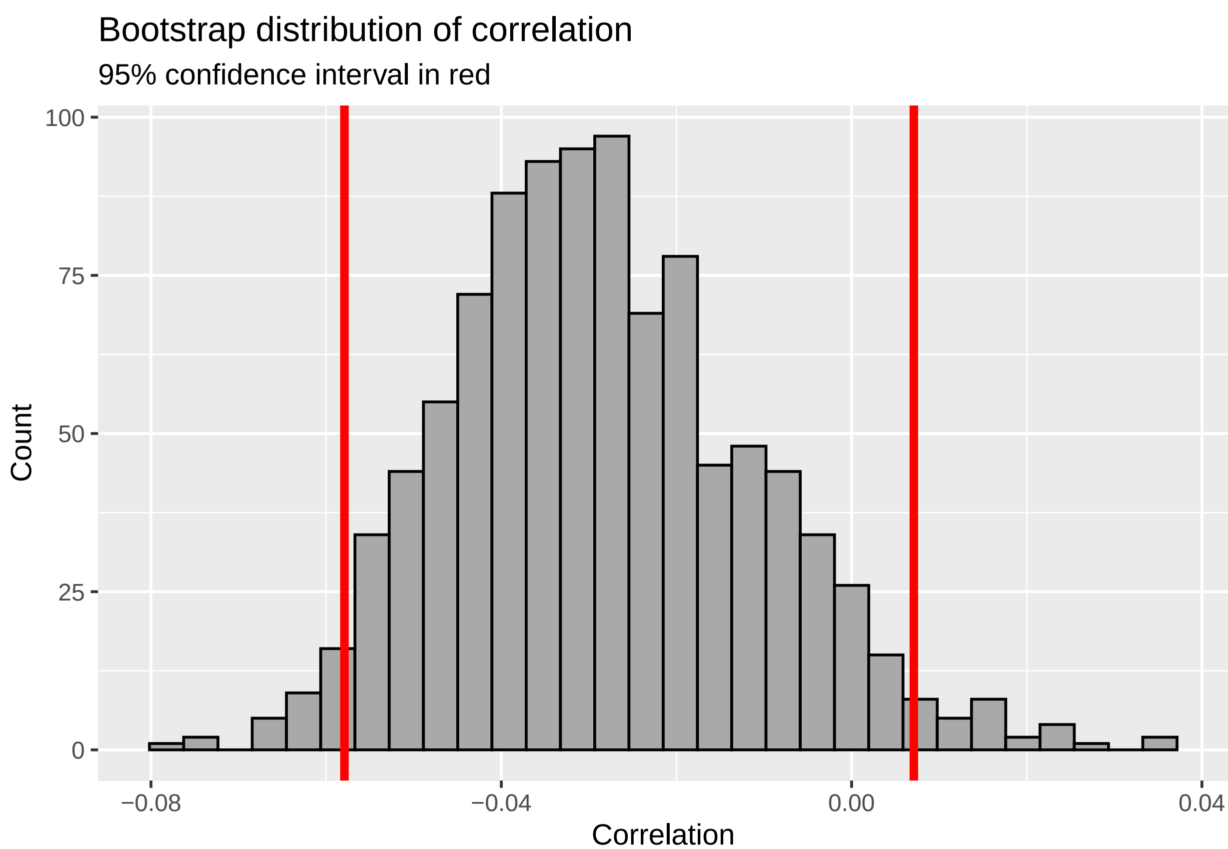
```r
  select(sample_correlation)

#simulation based approach for correlation
set.seed(1)
boot_dist2 = numeric(1000)

for(i in 1:1000){
  indices <- sample(1:nrow(combine2), replace = T)
  boot_ten_time <- combine2 %>%
    slice(indices) %>%
    summarize(boot_sd_tenTime = sd(tenTimeOvr), na.rm=TRUE) %>%
    pull()
  boot_ratio <- combine2 %>%
    slice(indices) %>%
    summarize(boot_sd_ratio = sd(legHgtRatio, na.rm=TRUE)) %>%
    pull()
  boot_covar_ratio <- combine2 %>%
    slice(indices) %>%
    summarize(boot_covar = cov(tenTimeOvr, legHgtRatio,
                               use = "complete.obs")) %>%
    pull()
  boot_dist2[i] <- (boot_covar_ratio/(boot_ratio*boot_ten_time))
}
boot_means2 <- tibble(boot_dist2)

boot_means3 = boot_means2 %>%
  summarize(lower = quantile(boot_dist2, 0.025),
            upper = quantile(boot_dist2, 0.975))
boot_means3
```

```
## # A tibble: 1 x 2
##     lower   upper
##     <dbl>   <dbl>
```

```
## 1 -0.0579 0.00711
```

```r
ggplot(data = boot_means2, aes(x = boot_dist2)) +
  geom_histogram(color = "black",
                 fill = "darkgrey") +
  labs(title = "Bootstrap distribution of correlation",
       subtitle = "95% confidence interval in red",
       x = "Correlation", y = "Count") +
  geom_vline(xintercept = c(boot_means3$lower, boot_means3$upper),
             color = "red", lwd = 1.5)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Based on an α level of 0.05, we are 95% confident that the true population coefficient for leg:torso and ten time splits is between (-0.05468, 0.0006826435).

Can't infer from lm for tentime and leg:torso or fortyTime and leg:torso Not normally distributed

Conclusion: evidence fails to reject the null that there is no correlation between leg:torso and ten time splits