STA302 Assigment 1

Zikun Chen
1001117882

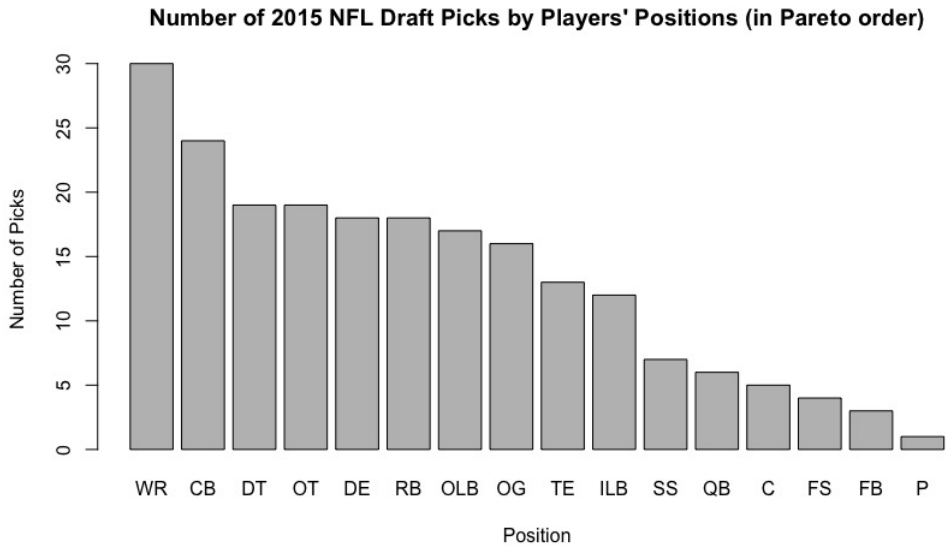
September 7, 2018

A Initial Data Analysis

2. (a) Cleveland Browns had the most draft picks this year. They made 11 picks.

   (b) Florida State had the most players drafted, and 11 were drafted.

   (c) Bar-plot in Pareto order:

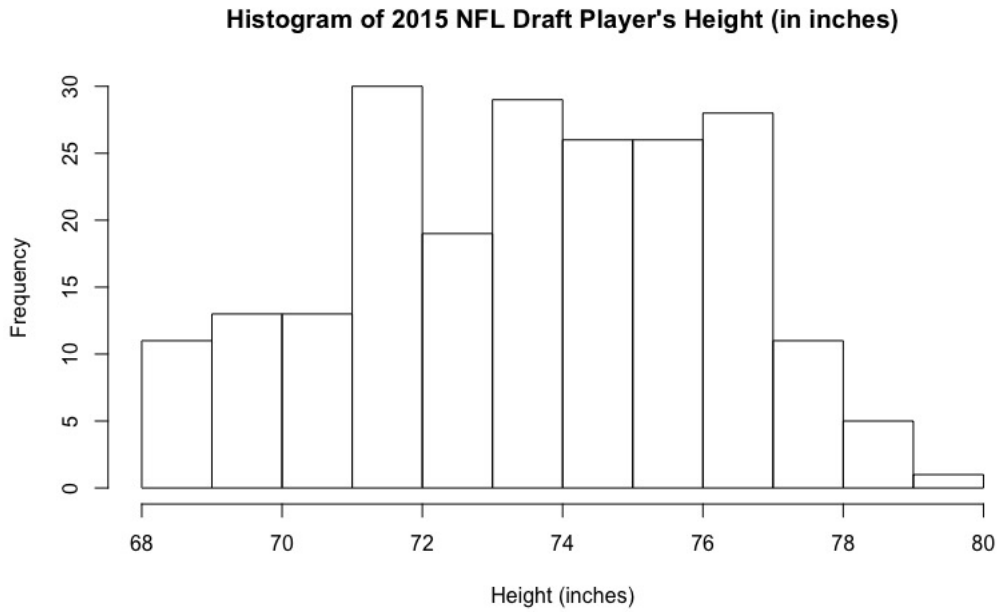


   (d) The following table presents the 5-number summary of players' heights in this years draft:

Minimum	1st Quantile	Median	3rd Quantile	Maximum
68	72	74	76	80

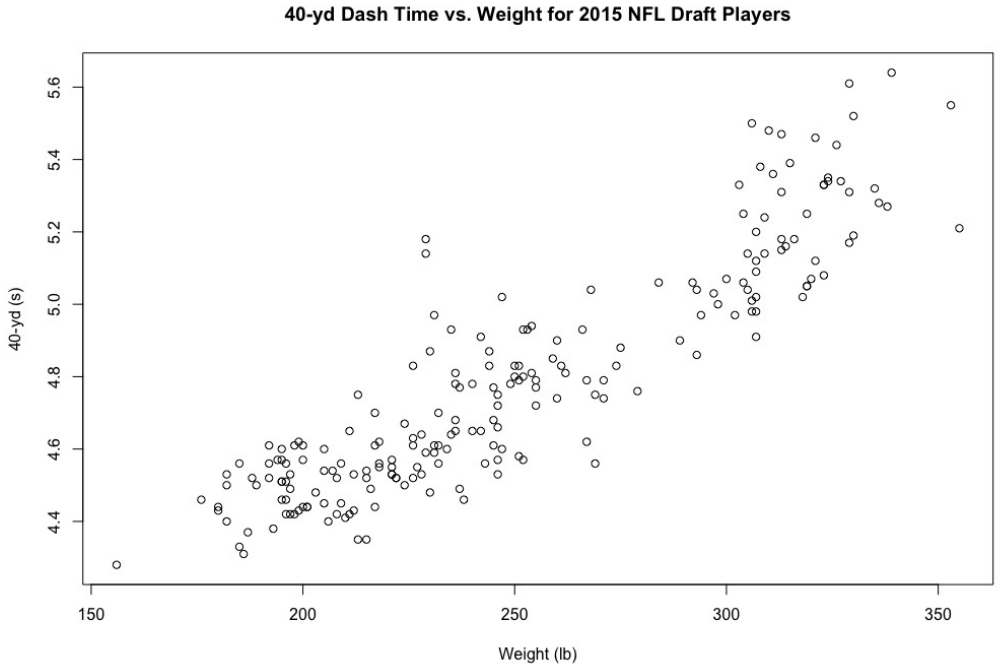
   The average height for an NFL draftee is 6 foot 2 (74.04245 inches).

   (e) Histogram of player heights:



(f) Jamison Crowder, Josh Robinson and Mario Alford are the shortest players in the draft.

(g) Plot of a players 40-yd dash time (Y) vs. their Weight (X):

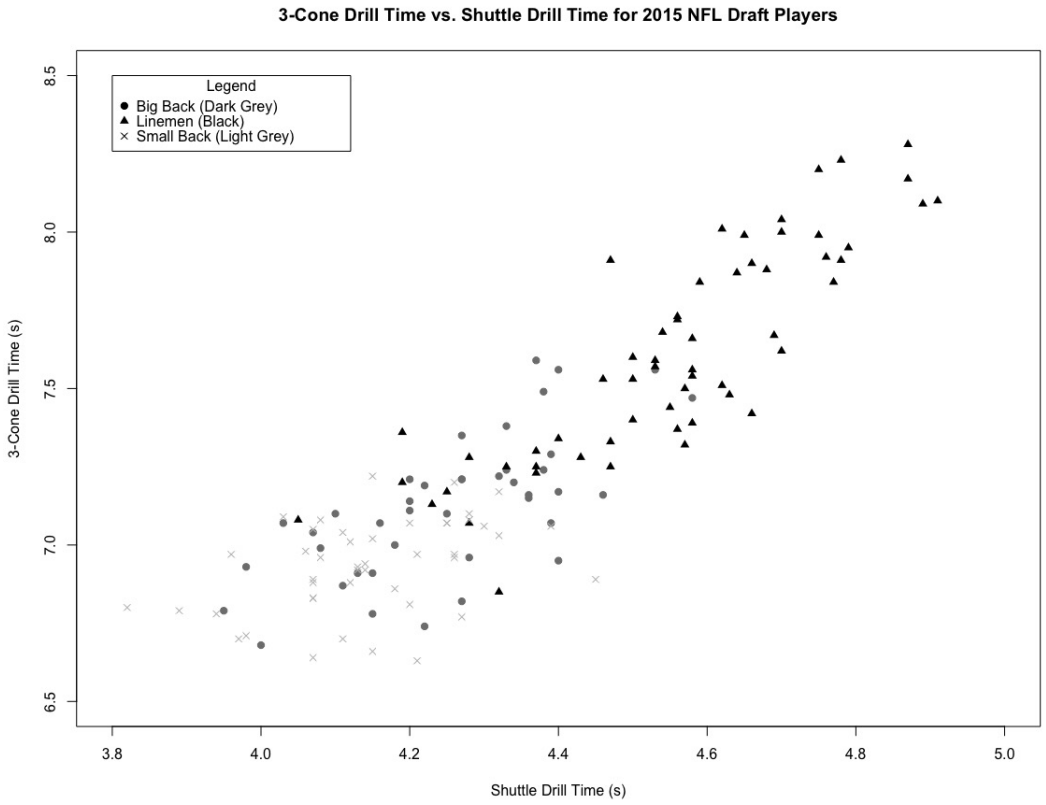


The relationship does look linear.

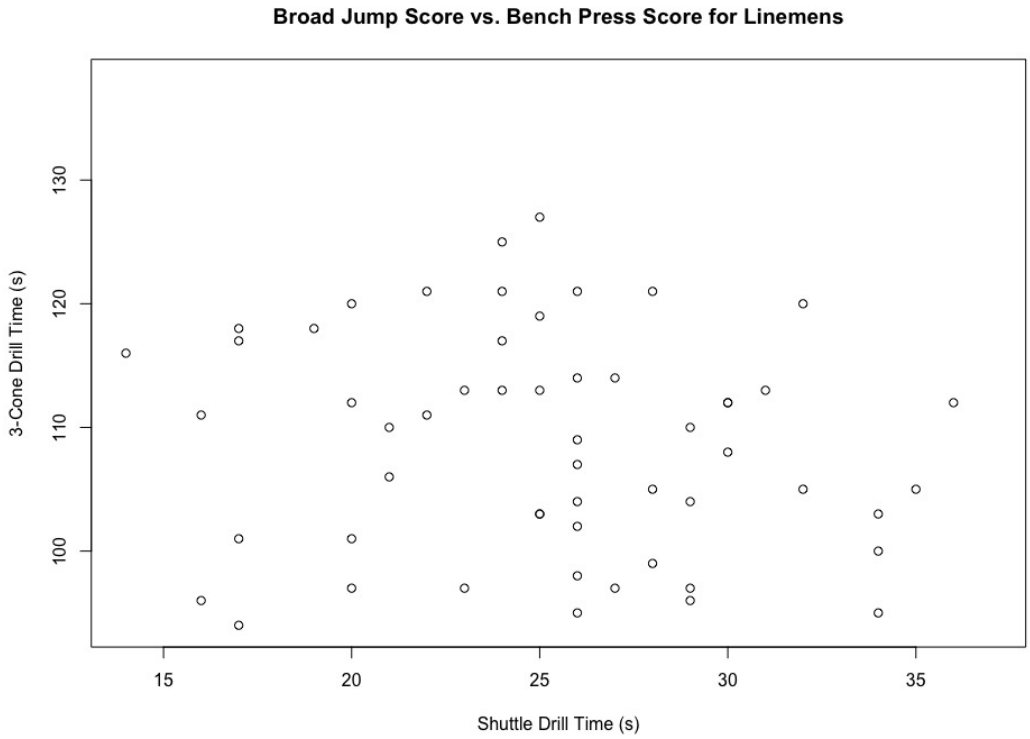
There are some outliers visible. There are two players who took relatively longer to complete the 40-yd dash test than players from the same weight range. In addition, the player who weighs the most could complete the 40-yd dash test with a very impressive score.

(h) It does look like there is a linear association present between 3-Cone drill time and the shuttle drill time.

(i) Plot of 3-cone drill time vs. shuttle drill time grouped by the three positional groups (Big Back, Linemen, Small Back):



(j) Plot of the Broad jump score vs. Bench press score, for linemen only:



There doesn't seem to be a relationship between these two test scores.

(k) Austin Shepherd had the shortest broad jump (with lighter weight). He got drafted in the 7th round.

(l) Byron Jones had the longest broad jump. There is no record of his bench press reps.

B First Linear Model

1. (a) Estimated regression equation is:
3-Cone Drill Time $= 0.9509769 + 1.4530307\times$ (Shuttle Drill Time)

(b) The slope is the expected increase in time spent on completing a 3-Cone Drill in seconds for a player when he takes 1 second longer to complete the Shuttle Drill.
The intercept is meaningless since it is not humanly possible to complete the shuttle drill in no time.
The relationship is statistically significant. The p-value is 2.2×10^{-6}.

(c) 92% Confidence Interval for the slope:

$$[1.3417969, 1.564264]$$

92% Confidence Interval for the intercept:

$$[0.4672872, 1.434667]$$

(d) The expected 3-Cone drill time is 7.489615s for players with a shuttle time of 4.5s.
95% Confidence Interval for this estimate:

$$[7.453684, 7.525546]$$

(e) I predict this new player's 3-Cone drill time to be 7.780221s
95% Confidence Interval for this estimate:

$$[7.415916, 8.144527]$$

2. I do not think it is reasonable to model the overall draft rank as a function of the 40-yard dash test score.
Because when I tried to fit an SLR model with the draft rank as the response and 40-yard dash result as the predictor and did inference on the slope, the p-value was 0.244 for the 2-sided test where the slope is claimed to be 0 (no linear relationship). We do not have enough evidence the claim that the slope is 0 at an significance level of 0.1. In other words, 0 lies in the 90% confidence interval for the slope (we cannot be 90% certain that 0 is not the true slope).
Therefore, the slope is not statistically significantly different from 0. So we cannot conclude that overall draft rank has a linear relationship with 40-yard dash score.

The broad jump is a better test score. The p-value of 2-sided test for the slope being zero for the SLR model fitting overall draft rank and broad score was 0.008371 (< 0.01). This is considered highly statistically significant. We can be 99% sure that 0 is not the true slope of the SLR model. So consequently, we are 99% sure that there is a linear relationship between overall draft rank and broad jump score. Thus, it is a good idea to model them in a linear model.

Appendix R Code

```
rm(list=ls())

# Part A
#1(a) reading
df <- read.csv("NFLdraft.csv", head=T, strip.white=T, stringsAsFactors=F)

#1(b) change to factor
df$Pos <- ifelse(df$Pos == "LS", "C", df$Pos)
df$Pos <- as.factor(df$Pos)

#1(c) new factor
new_Pos=as.character(df$Pos)
new_Pos[new_Pos %in% c("C", "OG","OT", "TE", "DT", "DE")] <- "L"
new_Pos[new_Pos %in% c("CB", "WR","FS")] <- "SB"
new_Pos[new_Pos != "SB" & new_Pos != "L"] <- "BB"
new_factor=as.factor(new_Pos)
rm(new_Pos)

#1(d) changing heights
df$Ht <- gsub("-", "", df$Ht)
df$Ht <- sapply(df$Ht, function(x) as.numeric(substr(x, 1, 1)) * 12
                + as.numeric(substr(x, 2, nchar(x))))

#1(e) breaking last column and storing
dftlist <- lapply(df$Drafted, function(x) unlist(strsplit(x, " / ")))

tmlist <- sapply(dftlist, function(x) x[1])
rdlist <- sapply(dftlist, function(x) x[2])
pklist <- sapply(dftlist, function(x) x[3])
yrlist <- sapply(dftlist, function(x) x[4])

df <- within(df, {
  Tm <-tmlist
})
df <- within(df, {
  Rd <-rdlist
})
df <- within(df, {
  Pk <-pklist
})
df <- within(df, {
  Yr <-yrlist
})
rm(dftlist, tmlist, rdlist, pklist, yrlist)

#1(f) Overall draft pick as number
```

```
df$Pk <- as.numeric(gsub("[ a-zA-Z]*", "", df$Pk))

#2(a)
team <- as.factor(df$Tm)
barplot(table(factor(team, levels = levels(team)[order(-table(team))])))
max(table(team))
rm(team)

#2(b)
cteam <- as.factor(df$CollegeTeam)
barplot(table(factor(cteam, levels = levels(cteam)[order(-table(cteam))])))
max(table(cteam))
rm(cteam)

#2(c)
position <- as.factor(df$Pos)
barplot(table(factor(position, levels = levels(position)[order(-table(position))])),
        main="Number of 2015 NFL Draft Picks by Players' Positions (in Pareto order)",
        xlab="Position", ylab="Number of Picks")
max(table(position))
rm(position)

#2(d)
summary(df$Ht, digits=7)
mean(df$Ht)

#2(e)
hist(df$Ht, main="Histogram of 2015 NFL Draft Player's Height (in inches)",
     xlab="Height (inches)")

#2(f)
subset(df, Ht==min(df$Ht))$Name

#2(g)
with(df, plot(Wt, Yd40,
     main="40-yd Dash Time vs. Weight for 2015 NFL Draft Players",
     xlab="Weight (lb)", ylab="40-yd (s)"))

#2(h)
df1 <- df
df1$Pos <- new_factor

#2(i)
with(subset(df1, Pos=="BB"), plot(Shuttle, Cone3,
     main="3-Cone Drill Time vs. Shuttle Drill Time for 2015 NFL Draft Players",
     xlab="Shuttle Drill Time (s)", ylab="3-Cone Drill Time (s)",
     xlim=c(3.8, 5), ylim=c(6.5, 8.5),pch=19, col="grey50"))
with(subset(df1, Pos=="L"), points(Shuttle, Cone3, pch=17, col="black"))
```

```
with(subset(df1, Pos=="SB"), points(Shuttle, Cone3, pch=4, col="grey75"))
legend(title="Legend", x=3.8, y=8.5, pch=c(19,17,4), text.width = 0.28,
       legend=c("Big Back (Dark Grey)", "Linemen (Black)", "Small Back (Light Grey)"))

#2(j)
with(subset(df1, df1$Pos=="L"), plot(Bench, Broad,
     main="Broad Jump Score vs. Bench Press Score for Linemens",
     xlab="Shuttle Drill Time (s)", ylab="3-Cone Drill Time (s)", col=1))
rm(df1)

#2(k)
nona <- subset(df, !is.na(Broad))
short <- subset(nona, Broad == min(nona$Broad))
target <- subset(short, Wt == min(short$Wt))
target$Name
target$Rd
rm(short, target)

#2(l)
subset(nona, Broad == max(nona$Broad))$Bench
rm(nona)

# Part B
# 1(a)
d <- data.frame(pred=df$Shuttle, resp=df$Cone3)
fit <- lm(resp ~ pred, data=d)
fit$coefficients
# 1(b)
summary(fit)
# 1(c)
confint(fit, level= 0.92)
# 1(d)
newData1 <- data.frame(pred=4.5)
predict(fit, newData1)
predict(fit, newData1, interval = "confidence", level = 0.95)
rm(newData1)
#1(e)
newData2 <- data.frame(pred=4.7)
predict(fit, newData2)
predict(fit, newData2, interval = "prediction", level = 0.95)
rm(newData2)
rm(fit, d)

#2
d <- data.frame(pred=df$Yd40, resp=df$Pk)
fit <- lm(resp ~ pred, data=d)
summary(fit)
confint(fit)
```

```
rm(fit, d)

d <- data.frame(pred=df$Broad, resp=df$Pk)
fit <- lm(resp ~ pred, data=d)
summary(fit)
confint(fit)
rm(fit, d)
rm(new_factor, df)
```