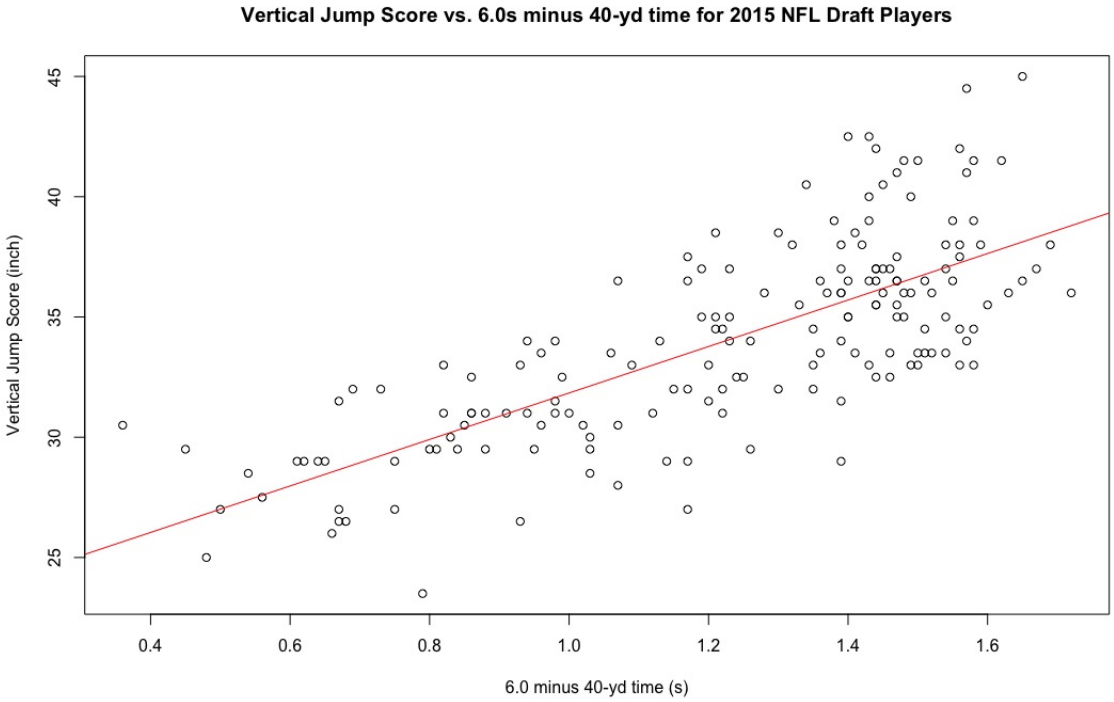
STA302 Assigment 2

Zikun CHEN
1001117882

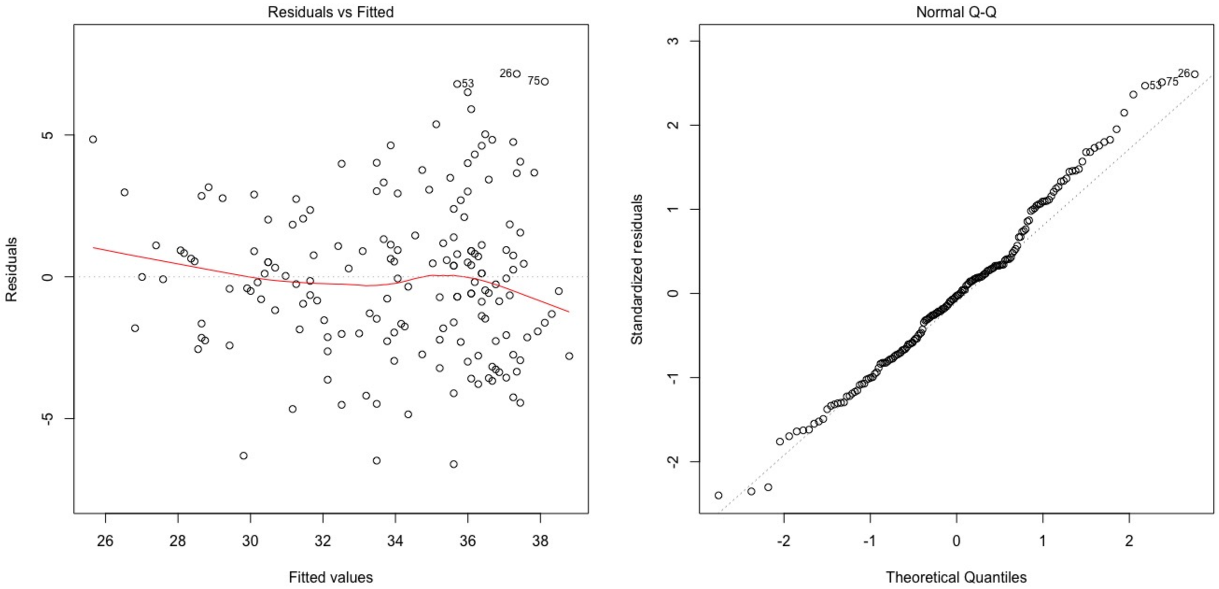
September 7, 2018

A NFL Draft

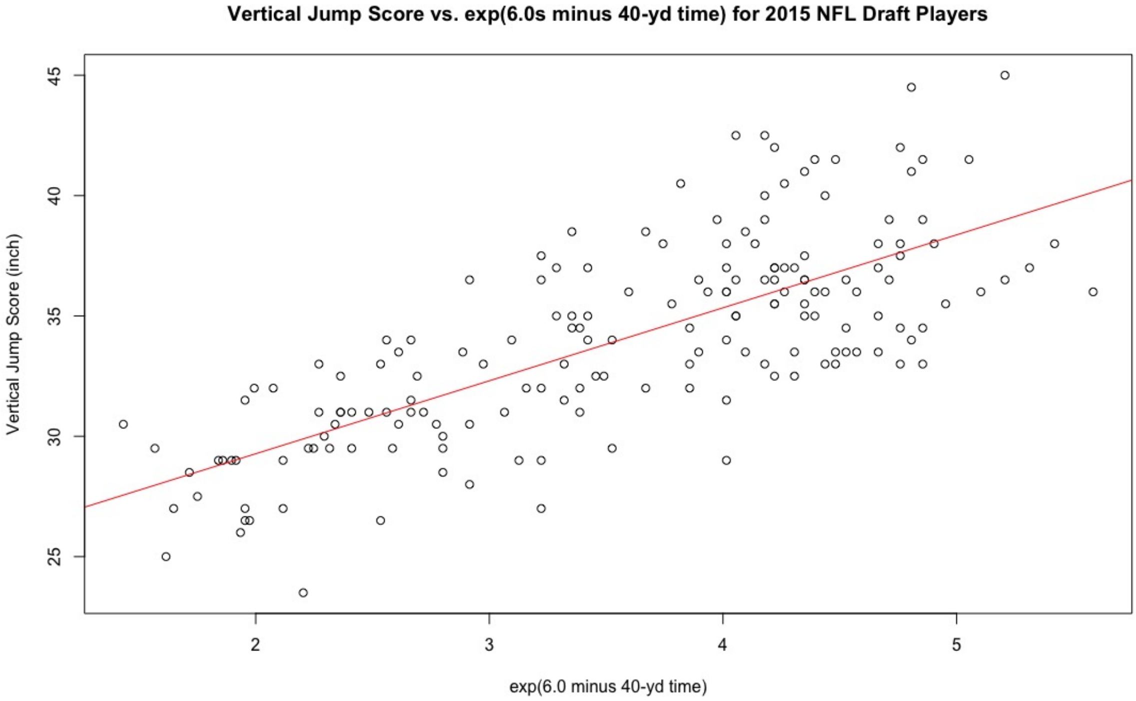
1. & 3. Plot of the vertical jump against 6.0s minus the 40-Yd time with LS trend-line:



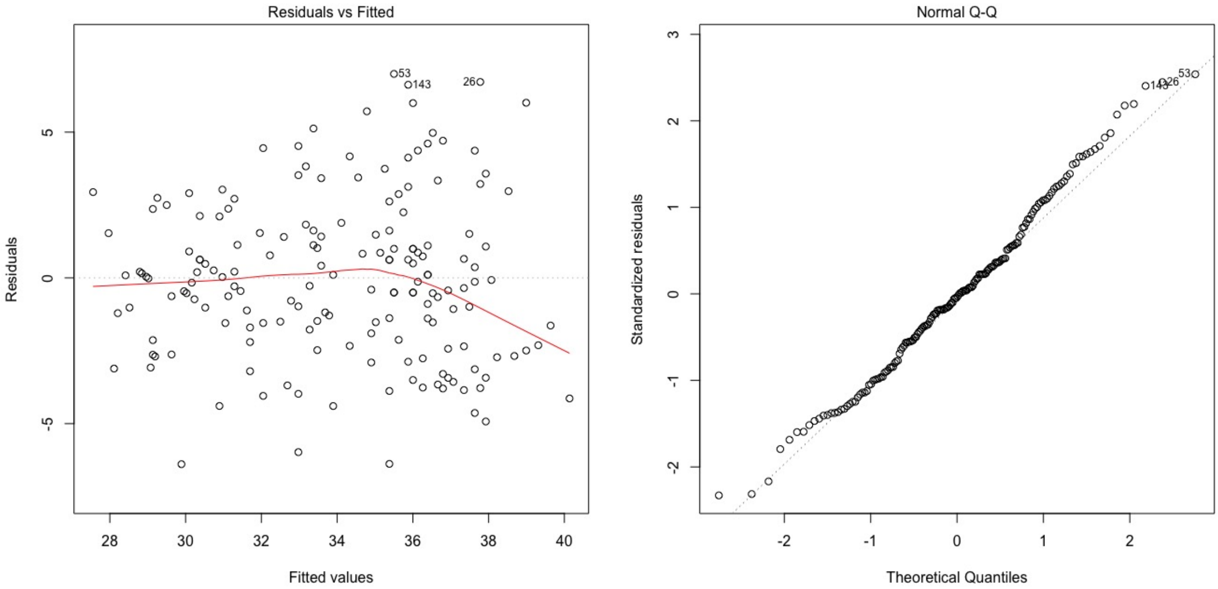
4. Residual plot and the Normal QQ:



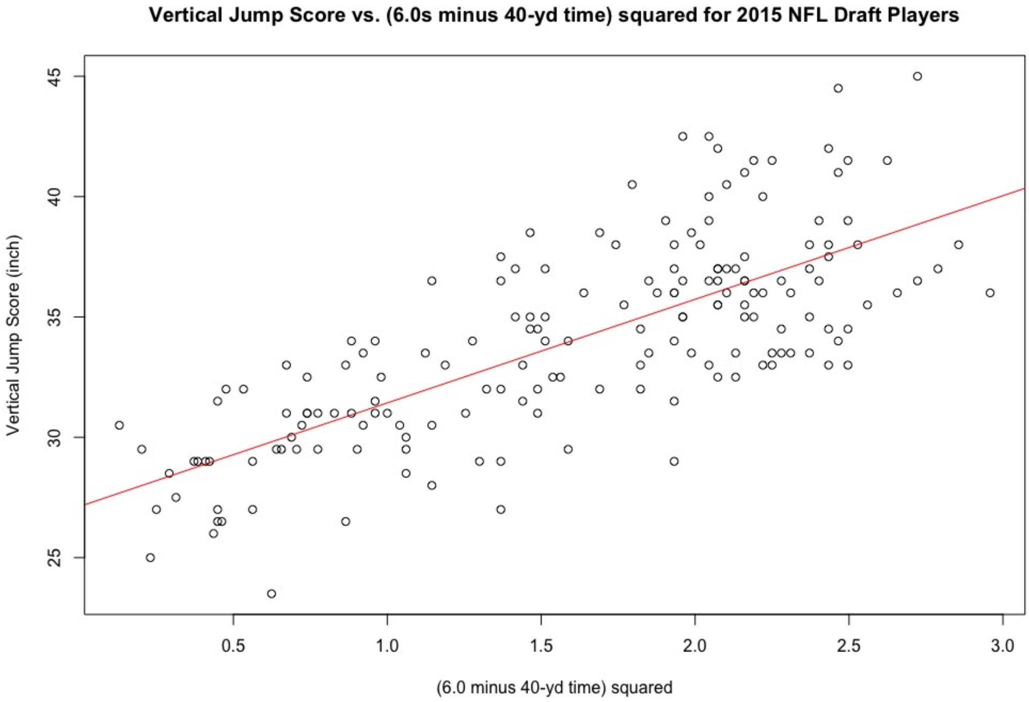
5. Plot of new fit with exponential transformation on (6 minus 40-Yd time):



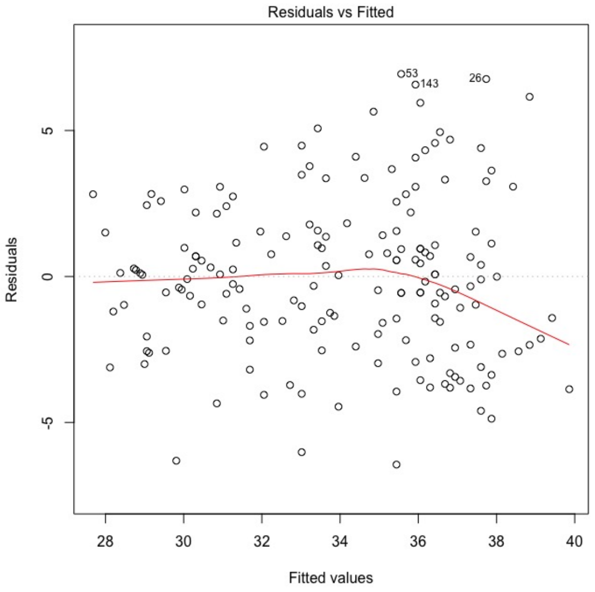
Residual plot and the Normal QQ of the exponential model:

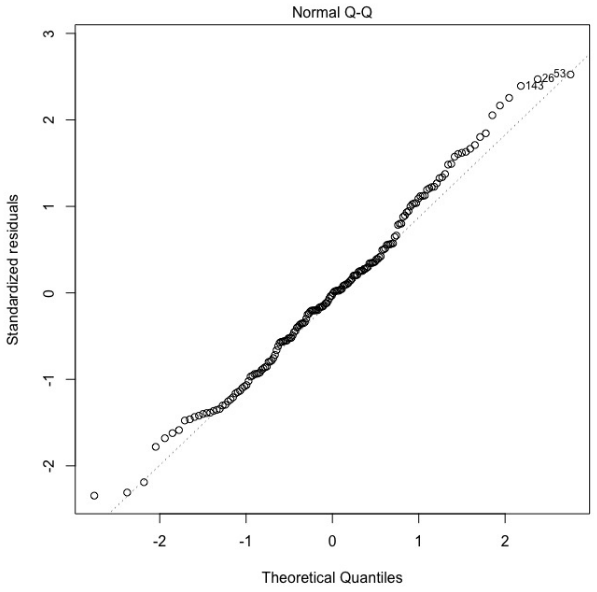


6. Plot of new fit with squared transformation on (6 minus 40-Yd time):



Residual plot and the Normal QQ of the squared model:

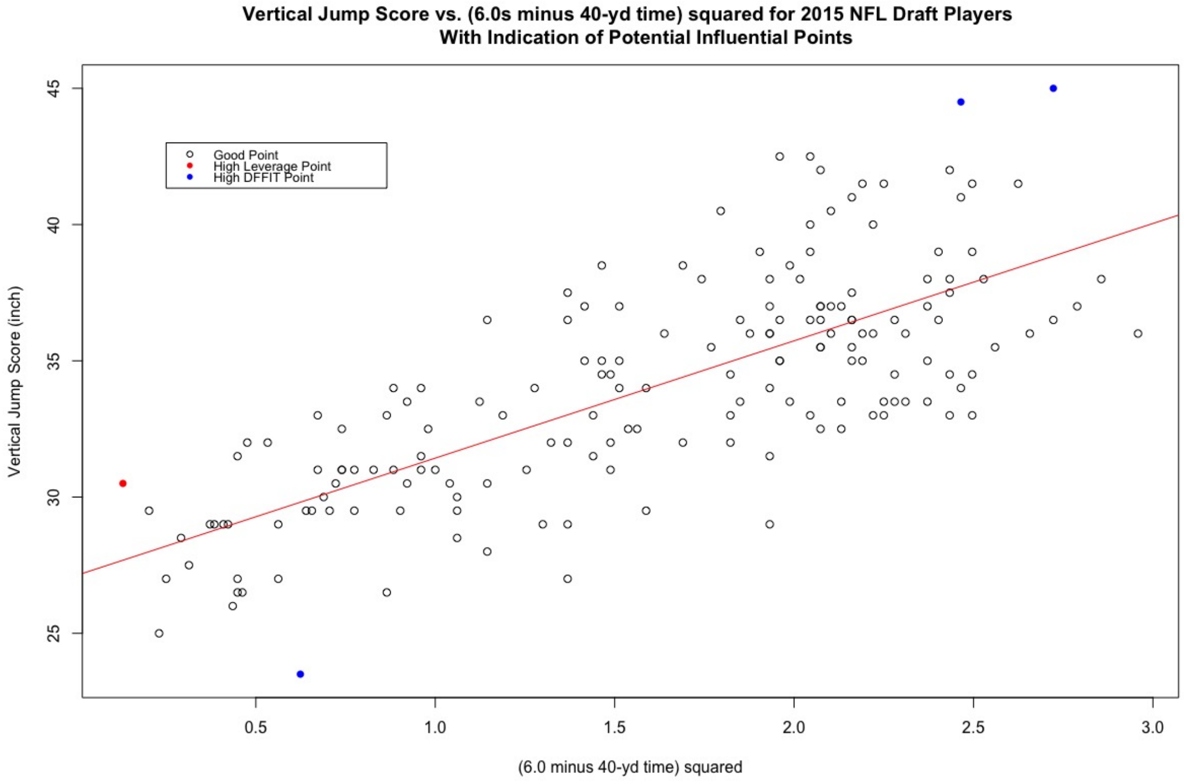




7. Here is a table contrasting the three models:

Model	R-squared Value	SSE	t-statistic	Slope Estimate
Linear	0.5501375	1298.136	14.41850	9.663182
Exponential	0.5489415	1301.587	14.38371	3.030209
Squared	0.5519850	1292.805	14.47244	4.305439

8. Here is a scatterplot of the squared model (highest R-squared and lowest SSE) with potential influential points (high leverage or high DFFIT) indicated (with red or blue respectively):



Danny Shelton has a high leverage.
Byron Jones, Chris Conley and Trenton Brown have high DFFITs.
These players are distant from the sample mean for (6 minus 40-yd time) squared. In particular, Danny Shelton took the longest time to complete 40-yd. In addition, the vertical jump scores of players who have high DFFITs are relatively further away (either higher or lower) from their expected vertical jump performance.

B NFL Season

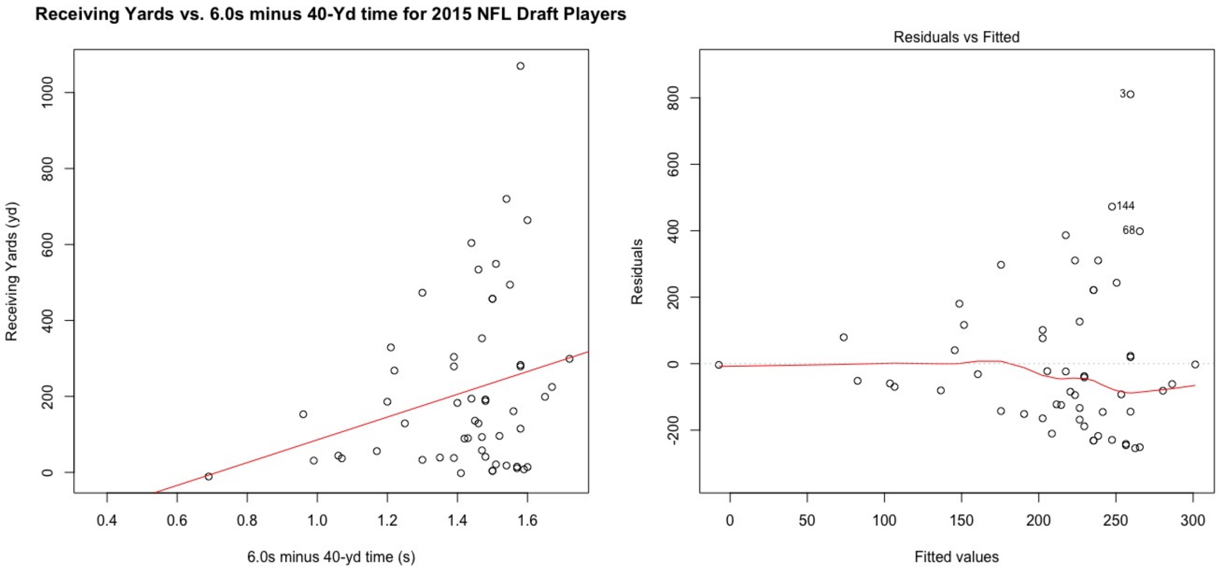
4. Summary table form R output of linear model with the player's receiving yards against 6 minus his 40-yard dash time:

```
Call:
lm(formula = REC.YDS ~ I(6 - Yd40), data = nflrr)

Residuals:
    Min      1Q  Median      3Q     Max
-254.46 -145.22  -55.60   95.97  810.54

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   -214.1       220.5  -0.971   0.3359
I(6 - Yd40)    299.7       153.8   1.949   0.0566 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 219 on 52 degrees of freedom
  (161 observations deleted due to missingness)
Multiple R-squared:  0.0681,Adjusted R-squared:  0.05018
F-statistic:   3.8 on 1 and 52 DF,  p-value: 0.05665
```

5. Interpretation of slope: The player's receiving yards will increase, on average by about 299.7 yards if he took 1 second less to finish the 40-yd test.
   The relationship is not statistically significant since the p-value for testing whether the slope is zero was 0.05665, which is bigger than 0.05.

6. Scatter plot and Residual plot of the linear model with the player's receiving yards against 6 minus his 40-yard dash time:



   The assumption that errors have constant variance seems to be violated. The scatterplot and residual plot shows the existence of nonconstant variance (heteroscedasticity). The variance of residuals grows as receiving yard distance increases. They don't spread out evenly across. The scatterplot also suggests that errors do not follow a normal distribution as in the assumption. They appear to have a right-tailed distribution.

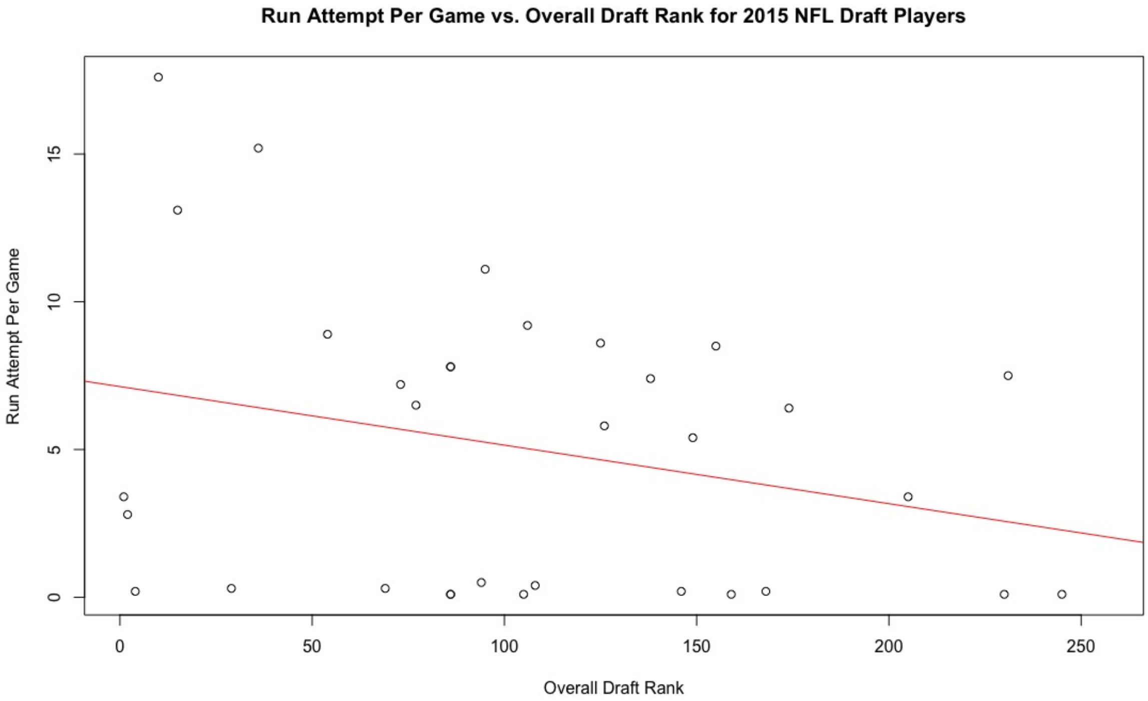
7. Summary table form R outpout of linear model with the player's run attempt per game against his overall draft rank:

```
Call:
```

```
lm(formula = ATTG ~ Overall, data = nflrr)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.8501 -4.2898  0.8997  3.0112 10.6691 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.12955    1.55571   4.583 7.07e-05 ***
Overall     -0.01986    0.01248  -1.592    0.122    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.792 on 31 degrees of freedom
  (182 observations deleted due to missingness)
Multiple R-squared:  0.07555,Adjusted R-squared:  0.04573 
F-statistic: 2.533 on 1 and 31 DF,  p-value: 0.1216
```

Interpretation of slope: The player's run attempts per game will decrease, on average by about 0.01986 times if his overall rank in the draft dropped by 1.
The relationship is not statistically significant with a p-value greater than 0.5 on testing whether the slope is zero. So there is no evidence that a players overall draft selection rank is linearly related with the run attempts per game. Thus, a players overall draft selection rank is not a good predictor of the run attempts per game.

8. Scatterplot of the linear model with the player's run attempt per game against his overall draft rank:



Based on the scatterplot, the error variance seems to grow as run attempts per game grow (not constant), and the errors do not look like they follow a normal distribution, their distribution seems to be left-tailed (or heavy-tailed, need to look at normal QQ plot).

Appendix R Code

```
## Part A ##
#1 plotting vertical jump vs 6-time
with(nfl, plot(I(6 - Yd40), Vertical,
        main="Vertical Jump Score vs. 6.0s minus 40-yd time for 2015 NFL Draft Players",
        xlab="6.0 minus 40-yd time (s)", ylab="Vertical Jump Score (inch)"))

#2 fitting linear model between vertical jump against 6-time
fit <- lm(Vertical ~ I(6 - Yd40), data= nfl)

#3 putting a red LS trendline
abline(coef(fit), col="red")

#4 residual plot & Normal QQ plot (1x2 grid)
par(mfrow=c(1, 2))
plot(fit, 1)
plot(fit, 2)
par(mfrow=c(1, 1))

#5 exp transformation
fitexp <- lm(Vertical ~ I(exp(1) ^ (6-Yd40)), data= nfl)
with(nfl, plot(exp(1)^(6-Yd40), Vertical,
    main="Vertical Jump Score vs. exp(6.0s minus 40-yd time) for 2015 NFL Draft Players",
    xlab="exp(6.0 minus 40-yd time)", ylab="Vertical Jump Score (inch)"))
abline(coef(fitexp), col="red")
par(mfrow=c(1, 2))
plot(fitexp, 1)
plot(fitexp, 2)
par(mfrow=c(1, 1))

#6 squared transformation
fitsq <- lm(Vertical ~ I((6-Yd40) ^ 2), data= nfl)
with(nfl, plot((6-Yd40)^2, Vertical,
    main="Vertical Jump Score vs. (6.0s minus 40-yd time) squared for 2015 NFL Draft Players",
    xlab="(6.0 minus 40-yd time) squared", ylab="Vertical Jump Score (inch)"))
abline(coef(fitsq), col="red")
par(mfrow=c(1, 2))
plot(fitsq, 1)
plot(fitsq, 2)
par(mfrow=c(1, 1))

#7 comparison table of 3 models
#(1) linear fit
r.sq1 = summary(fit)$r.squared
sse1 = anova(fit)$"Sum Sq"[2]
tstat1 = summary(fit)$coefficients[2,3]
slope1 = summary(fit)$coefficients[2,1]
```

```
#(2) exponential fit
r.sq2 = summary(fitexp)$r.squared
sse2 = anova(fitexp)$"Sum Sq"[2]
tstat2 = summary(fitexp)$coefficients[2,3]
slope2 = summary(fitexp)$coefficients[2,1]


#(3) squared fit
r.sq3 = summary(fitsq)$r.squared
sse3 = anova(fitsq)$"Sum Sq"[2]
tstat3 = summary(fitsq)$coefficients[2,3]
slope3 = summary(fitsq)$coefficients[2,1]

data.frame(Model=c("linear", "exp", "square"),
           R.square = c(r.sq1, r.sq2, r.sq3), SSE=c(sse1, sse2, sse3),
           tstat=c(tstat1, tstat2, tstat3), slope=c(slope1, slope2, slope3))
rm(r.sq1, r.sq2, r.sq3, sse1, sse2, sse3, tstat1, tstat2, tstat3, slope1, slope2, slope3)

#8 DFFIT
avglvg=2/(summary(fitsq)$df[2]+2)
dflvg=subset(as.data.frame(influence.measures(fitsq)$infmat), hat > 2.5*avglvg)
dfdifft=subset(as.data.frame(influence.measures(fitsq)$infmat), abs(dffit) > 0.3)
badrows=as.numeric(rownames(subset(as.data.frame(influence.measures(fitsq)$infmat),
                                   hat > 2.5*avglvg | abs(dffit) > 0.3)))
with(nfl[-badrows,], plot((6-Yd40)^2, Vertical,
      xlim= c((6-max(nfl$Yd40, na.rm= T))^2, (6-min(nfl$Yd40, na.rm= T))^2),
      ylim= c(min(nfl$Vertical, na.rm= T),   max(nfl$Vertical, na.rm= T)),
      main="Vertical Jump Score vs. (6.0s minus 40-yd time) squared for 2015 NFL Draft Players
      With Indication for Potential Influential Points",
      xlab="(6.0 minus 40-yd time) squared", ylab="Vertical Jump Score (inch)"))
with(nfl[as.numeric(rownames(dflvg)),], points((6-Yd40)^2, Vertical, col="red", pch=16))
with(nfl[as.numeric(rownames(dfdifft)),], points((6-Yd40)^2, Vertical, col="blue", pch=16))
abline(coef(fitsq), col="red")
with(nfl, legend(x=0.25, y=43, cex = 0.80, text.width = 0.45,
                 legend=c("Good Point", "High Leverage Point", "High DFFIT Point"),
                 col= c("black", "red", "blue") ,pch=c(1,16,16)))
#return players' name
highlvg <- subset(nfl, rownames(nfl) %in% rownames(subset(
  as.data.frame(influence.measures(fitsq)$infmat), hat > 2.5*avglvg)))
highdffit <- subset(nfl, rownames(nfl) %in% rownames(subset(
  as.data.frame(influence.measures(fitsq)$infmat), abs(dffit) > 0.3)))
highlvg$Name
highdffit$Name
rm(avglvg, dfdifft, dflvg, badrows, highlvg, highdffit, fitsq, fitexp)

## Part B ##
require(plyr)
```

```
#1 Reading and cleaning data
rec.raw <- read.csv("Receiving.csv", head= T, strip.white= T , stringsAsFactors = F)
run.raw <- read.csv("Rushing.csv", head= T, strip.white= T , stringsAsFactors = F)
rec <- subset(rec.raw, select=-c(Pos, Team))
run <- subset(run.raw, select=-c(Pos, Team))
rm(rec.raw, run.raw)

#2 changing colnames
samecol <- intersect(colnames(rec)[2:length(colnames(rec))],
                     colnames(run)[2:length(colnames(run))])
colnames(rec) <- ifelse(colnames(rec) %in% samecol,
                        paste("REC.", colnames(rec), sep = ""), colnames(rec))
colnames(run) <- ifelse(colnames(run) %in% samecol,
                        paste("RUN.", colnames(run), sep = ""), colnames(run))
rm(samecol)

#3 merging data
colnames(rec)[1] <- "Name"
colnames(run)[1] <- "Name"
rookies <- nfl$Name
nflrr <- join(join(nfl, rec, by = "Name", type="left"),
              run, by = "Name", type="left")
rm(rookies)

#4 linear model between receiving yards against 6 minus 40-Yd
fit4 <- lm(REC.YDS ~ I(6 - Yd40), data= nflrr)
summary(fit4)

#6 scatterplot and residual plot
par(mfrow=c(1, 2))
with(nflrr, plot(6 - Yd40, REC.YDS,
     main="Receiving Yards vs. 6.0s minus 40-Yd time for 2015 NFL Draft Players",
     xlab="6.0s minus 40-yd time (s)", ylab="Receiving Yards (yd)"))
abline(coef(fit4), col="red")
plot(fit4, 1)
par(mfrow=c(1, 1))
rm(fit4)

#7 linear model between run attempts per game against overall draft rank
fit7 <- lm(ATTG ~ Overall, data= nflrr)
summary(fit7)

#8 scatterplot
with(nflrr, plot(Overall, ATTG,
     main="Run Attempt Per Game vs. Overall Draft Rank for 2015 NFL Draft Players",
     xlab="Overall Draft Rank", ylab="Run Attempt Per Game"))
abline(coef(fit7), col="red")
rm(fit7)
```