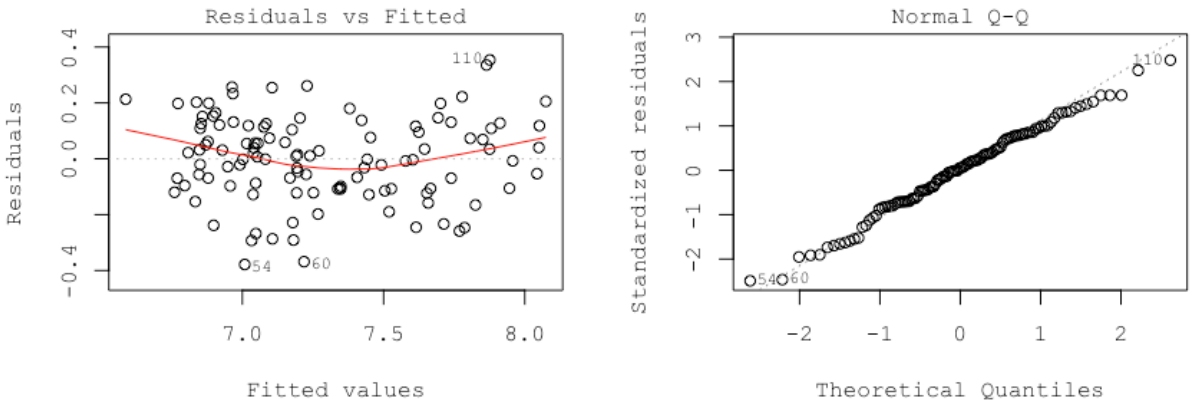
Name: Omika DHARAMDASANI
Student Number: 1000984483
Lecture: 5101
Date: June 15, 2016

STA302 ASSIGNMENT 3

PART A:

1.



Looking at the residual plot, we can see that the residuals seem to be evenly distributed across the fitted values, indicating constant variance. The linearity assumption is not met because the residuals do not add up to zero, as indicated by the curved red line. The normality assumption seems to be met since the normal Q-Q plot only shows some deviation at both ends. Hence, I would conclude that we could probably trust inference on this model, but we can also do better than this.

2. Our reduced model has 3-Cone Times as the response and 40-yard dash times and Shuttle times as predictors.

```
Call:
lm(formula = nfl$Cone3 ~ nfl$Yd40 + nfl$Shuttle)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4031 -0.1022  0.0013  0.1228  0.3402

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.74934    0.26951   2.780   0.0064 **
nfl$Yd40     0.55241    0.08377   6.594 1.58e-09 ***
```

```
nfl$Shuttle  0.89171    0.10878   8.198 5.22e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.157 on 109 degrees of freedom
Multiple R-squared:  0.851,Adjusted R-squared:  0.8483
F-statistic: 311.3 on 2 and 109 DF,  p-value: < 2.2e-16
```

3. F-statistic = 1.0733
   Degrees of freedom = 5
   p-value = 0.3795

   Since the p-value is greater than our significance level of 0.05, there is no evidence against our null hypothesis, which states that both models are equivalent. Hence, we fail to reject it. This means that we can now use our reduced model to explain the data.

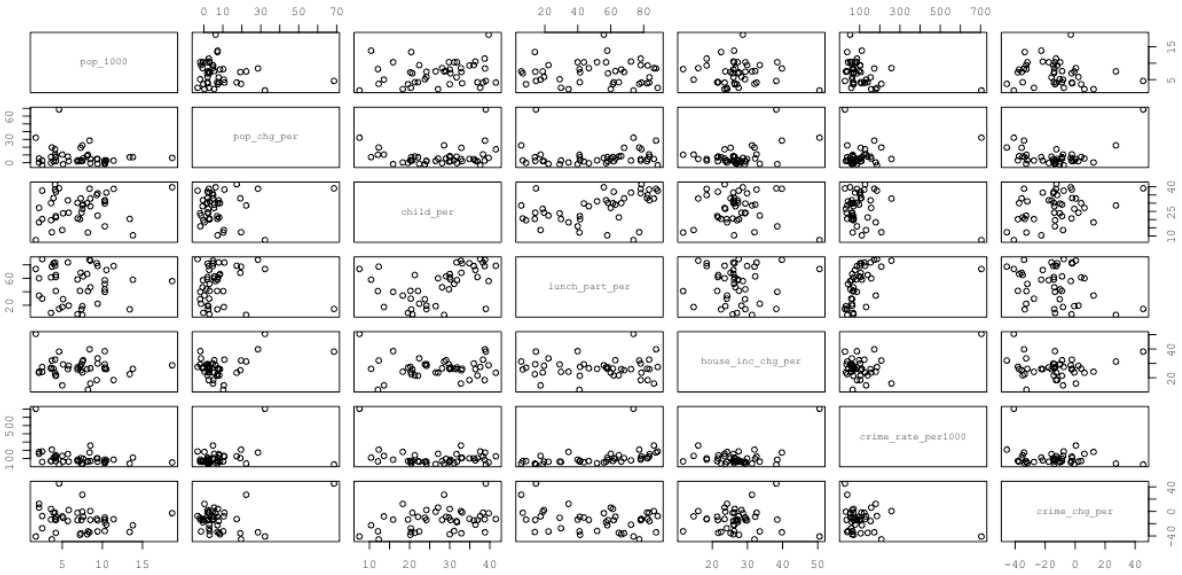
4. We are looking for the coefficient of partial determination. This can be found using the following formula:
   ((SSE(Reduced) – SSE(full))/SSE(reduced))*100
   Therefore, the answer is ((2.6857 – 2.5539)/2.6857)*100 = **4.9075%**

5. Adding weight is useful in predicting 3-cone time on top of the 40-yard times and Shuttle times.

```
Call:
lm(formula = origdf$Cone3 ~ origdf$Wt + origdf$Yd40 +
origdf$Shuttle)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40851 -0.10164  0.00574  0.12255  0.32185

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.5911663  0.3965814   4.012 9.75e-05 ***
origdf$Wt      0.0016749  0.0006748   2.482  0.01424 *
origdf$Yd40    0.3253517  0.1038521   3.133  0.00211 **
origdf$Shuttle 0.8526824  0.0955407   8.925 2.24e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1545 on 140 degrees of freedom
(68 observations deleted due to missingness)
Multiple R-squared:  0.8504,     Adjusted R-squared:  0.8472
F-statistic: 265.4 on 3 and 140 DF,  p-value: < 2.2e-16
```

PART B:

1.



The communities with the high crime rate and large change in population do indeed stand out. There is another observation that stands out as an outlier with seemingly large population per 1000, which looks like it might have high leverage.

2.

	pop_1000	pop_chg_per	child_per	lunch_part_per	house_inc_chg_per	crime_rate_per1000
pop_1000	1.000	-0.031	0.120	-0.020	0.092	-0.368
pop_chg_per	-0.031	1.000	0.100	0.233	0.105	0.267
child_per	0.120	0.100	1.000	0.600	0.121	0.089
lunch_part_per	-0.020	0.233	0.600	1.000	0.007	0.606
house_inc_chg_per	0.092	0.105	0.121	0.007	1.000	-0.091
crime_rate_per1000	-0.368	0.267	0.089	0.606	-0.091	1.000
crime_chg_per	-0.193	-0.168	0.207	-0.184	-0.038	0.021

	crime_chg_per
pop_1000	-0.193
pop_chg_per	-0.168
child_per	0.207
lunch_part_per	-0.184
house_inc_chg_per	-0.038
crime_rate_per1000	0.021
crime_chg_per	1.000

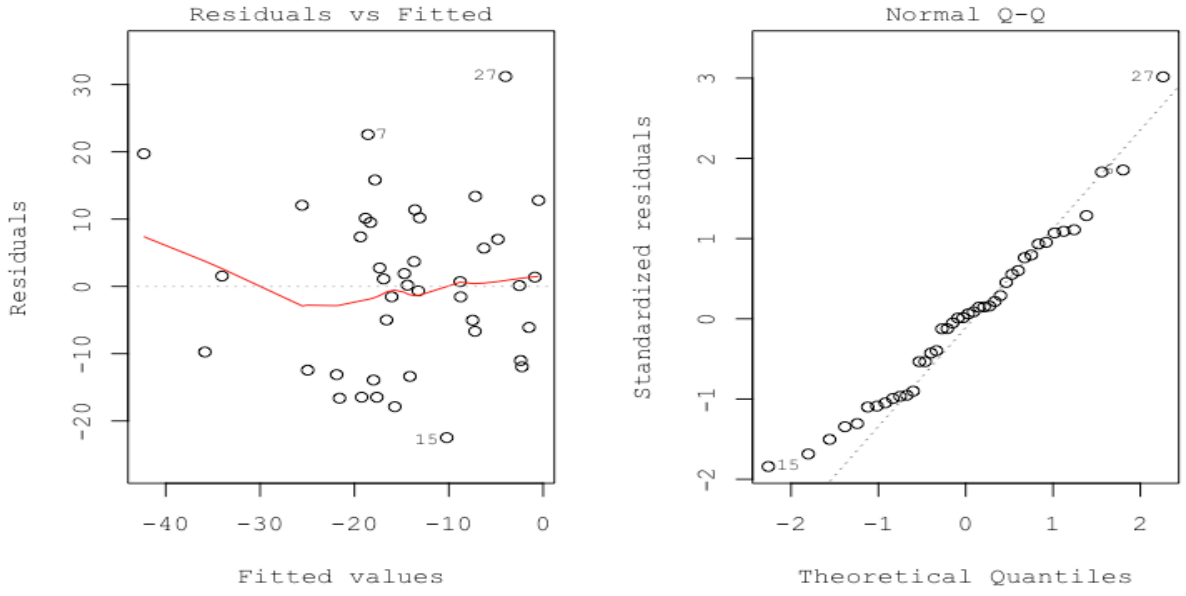
It looks like the percentage of children under 18 is highly correlated with the lunch program participation percentage (0.6). Crime rate per 1000 also seems to be highly correlated with lunch participation program percentage (0.606). This could present the problem of multicollinearity, which could potentially make our coefficient estimates very sensitive to minor changes in the model.

3.

	Estimate	Std. Error	t value	Pr(>|t|)	vifs
(Intercept)	-28.6174912	13.65274488	-2.0960980	0.0433720391	1.257794
denver$pop_1000	-0.5294447	0.65413039	-0.8093871	0.4237610032	1.107021
denver$pop_chg_per	-0.3643101	0.31336186	-1.1625860	0.2528635308	1.944417
denver$child_per	1.3362415	0.34344669	3.8906809	0.0004278115	3.179425
denver$lunch_part_per	-0.5238328	0.14271429	-3.6705001	0.0008003091	1.044703
denver$house_inc_chg_per	-0.1311643	0.37771820	-0.3472543	0.7304797713	2.438403
denver$crime_rate_per1000	0.1450314	0.06293124	2.3046012	0.0272399832	NA

Since VIFs are all greater than 1, we do not have to worry about linear independence. Also, since no VIF is greater than 10, we do not need to worry about multicollinearity.

4.



The residual plot seems to show that the residuals do not sum to zero. The variance looks constant because the residual plot shows no fanning. The normal QQ-plot deviates significantly,

which implies that the normality assumption is not met. Hence, some of the assumptions for linear regression are not met.

5. Test-statistic = 0.7962
   Df = 3
   p-value = 0.5043

Since the p-value is greater than our significance level of 0.05, there is no evidence against our null hypothesis that both models are equivalent. Hence, we fail to reject it. This means that we can now use our reduced model to explain the data, which is what we prefer.

```
Call:
lm(formula = denver$crime_chg_per ~ denver$child_per +
denver$lunch_part_per +
    denver$crime_rate_per1000)

Residuals:
    Min      1Q  Median      3Q     Max
-26.740  -8.151   1.362   8.516  24.542

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               -37.07406    8.28513  -4.475 6.74e-05 ***
denver$child_per            1.32442    0.33851   3.913 0.000366 ***
denver$lunch_part_per      -0.55765    0.13778  -4.047 0.000245 ***
denver$crime_rate_per1000   0.15713    0.05572   2.820 0.007588 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.03 on 38 degrees of freedom
Multiple R-squared:  0.3311,    Adjusted R-squared:  0.2783
F-statistic:  6.27 on 3 and 38 DF,  p-value: 0.001454
```

6. There is an increase in crime rate percentage of 1.32442 on average, as the children under 18 increase by 1 percent, given that lunch participation and crime rate per 1000 are fixed.

On average, there is a decrease in crime rate percentage of 0.55765 as the lunch participation percentage increases by 1 percent, given that children under 18 and crime rate per 1000 are fixed.

7. Test-statistic = 1.5555
   Df = 4
   p-value = 0.2085

Since the p-value is less than our significance level of 0.05, there is evidence against our null hypothesis that both models are

equivalent. Hence, we reject it and conclude that the two models are not equivalent.

8. I will use the interaction model, since it has a higher adjusted R-squared.

	fit	lwr	upr
1	-15.36544	-42.19293	11.46206

The prediction interval for the new community is from -42.19293 to 11.46206.

APPENDIX: R PROGRAM FILE FOR ASSIGNMENT 3

```
#A1 data prep code taken from Portal

## Data prep ##
nfl.raw <- read.csv("NFLdraft.csv", head= T, strip.white= T ,
stringsAsFactors = F)
# head(nfl.raw)
# tail(nfl.raw)
# str(nfl.raw)

nfl.raw$Pos[nfl.raw$Pos == "LS"] <- "C"

nfl <- within(nfl.raw, {
  Pos <- factor(Pos)
  PosGroup <- factor(ifelse(Pos %in% c("C", "DE", "DT", "OG",
"OT", "TE"), "Linemen",
                            ifelse(Pos %in% c("CB", "WR", "FS"),
"Small Backs", "Big Backs")))

  ProTeam <- factor(matrix(unlist(strsplit(Drafted, " / ")),
ncol= 4, byrow = T)[,1])
  Round   <- factor(matrix(unlist(strsplit(Drafted, " / ")),
ncol= 4, byrow = T)[,2])
  Overall <- factor(matrix(unlist(strsplit(Drafted, " / ")),
ncol= 4, byrow = T)[,3])
  Year    <- factor(matrix(unlist(strsplit(Drafted, " / ")),
ncol= 4, byrow = T)[,4])

  Overall <- as.numeric(gsub("[^0-9]", "", Overall))

  HtFt <- as.numeric(sub("-.*", "", Ht))
  HtIn <- as.numeric(sub("*.-", "", Ht))
  Ht   <- 12*HtFt + HtIn
})
#nfl <- subset(nfl, select= -c(Link, Drafted, Year, HtFt, HtIn))
# Remove unused columns
head(nfl) # Check
str(nfl)
View(nfl)

#storing the original data frame in another variable for later
use
origdf <- nfl

### Part A ###
```

```
#1. Removing all NAs
nfl <- nfl[!is.na(nfl$Vertical),]
nfl <- nfl[!is.na(nfl$Bench),]
nfl <- nfl[!is.na(nfl$Broad),]
nfl <- nfl[!is.na(nfl$Cone3),]
nfl <- nfl[!is.na(nfl$Shuttle),]
View(nfl)

par(family="Courier New")
#1. Fitting MLR model
fitFull <- lm(nfl$Cone3 ~ nfl$Ht + nfl$Wt + nfl$Yd40 +
nfl$Vertical + nfl$Bench + nfl$Broad + nfl$Shuttle)
par(mfrow=c(1,2))
plot(fitFull,1); plot(fitFull, 2)

par(mfrow=c(1,1))

#2.
summary(fitFull)
#Removing height
fitRed1 <- lm(nfl$Cone3 ~ nfl$Wt + nfl$Yd40 + nfl$Vertical +
nfl$Bench + nfl$Broad + nfl$Shuttle)
summary(fitRed1)
#Removing bench press
fitRed2 <- lm(nfl$Cone3 ~ nfl$Wt + nfl$Yd40 + nfl$Vertical +
nfl$Broad + nfl$Shuttle)
summary(fitRed2)
#Removing broad jump
fitRed3 <- lm(nfl$Cone3 ~ nfl$Wt + nfl$Yd40 + nfl$Vertical +
nfl$Shuttle)
summary(fitRed3)
#Removing vertical
fitRed4 <- lm(nfl$Cone3 ~ nfl$Wt + nfl$Yd40+ nfl$Shuttle)
summary(fitRed4)
#Removing weight
fitRed5 <- lm(nfl$Cone3 ~ nfl$Yd40+ nfl$Shuttle)
summary(fitRed5)
#fitRed5 is our reduced model

#3. Partial F-test
anova(fitRed5, fitFull)

#4.
anova(fitRed5)
anova(fitFull)

#5.
```

```
fitReduced <- lm(origdf$Cone3 ~ origdf$Yd40 + origdf$Shuttle)
summary(fitReduced)
#Adding broad jump
fitReduced1 <- lm(origdf$Cone3 ~ origdf$Broad + origdf$Yd40 +
origdf$Shuttle)
summary(fitReduced1)
#Adding Bench Press
fitReduced2 <- lm(origdf$Cone3 ~ origdf$Bench + origdf$Yd40 +
origdf$Shuttle)
summary(fitReduced2)
#Adding Vertical
fitReduced3 <- lm(origdf$Cone3 ~ origdf$Vertical + origdf$Yd40 +
origdf$Shuttle)
summary(fitReduced3)
#Adding Wt
fitReduced4 <- lm(origdf$Cone3 ~ origdf$Wt + origdf$Yd40 +
origdf$Shuttle)
summary(fitReduced4)
#all the predictors are significant
#Adding Ht
fitReduced5 <- lm(origdf$Cone3 ~ origdf$Ht + origdf$Yd40 +
origdf$Shuttle)
summary(fitReduced5)


### PART B ###
#Reading in the Denver file
denver <- read.csv("Denver.csv", head= T, strip.white= T ,
stringsAsFactors = F)
head(denver)
tail(denver)
str(denver)
View(denver)

#1. Pairwise scatterplot matrix for all variables
pairs(denver)
a <- which(denver$pop_chg_per == max(denver$pop_chg_per))
b <- which(denver$crime_rate_per1000 ==
max(denver$crime_rate_per1000))
denver <- denver[-a, ]
denver <- denver[-b, ]
View(denver)

#2. Correlation matrix rounded to 3 dp
round(cor(denver), 3)

#3. Fitting an MLR model with crime rate change as response and
```

```
the other six as predictor variables
fit.full <- lm(crime_chg_per ~ pop_1000 + pop_chg_per +
child_per + lunch_part_per + house_inc_chg_per +
crime_rate_per1000, data=denver)

install.packages("car")
library(car)

vifs <- vif(fit.full)
vifs
vifs[7] = NA
tab <- coefficients(summary(fit.full))
cbind(tab, vifs)

#4. Diagnostic plots
par(mfrow=c(1,2))
plot(fit.full,1); plot(fit.full, 2)
par(mfrow=c(1,1))

#5. Reducing the model
summary(fit.full)
#Removing pop_1000, pop_chg_per and house_inc_chg_per
fit.red <- lm(crime_chg_per ~ child_per + lunch_part_per +
crime_rate_per1000, data=denver)
anova(fit.red, fit.full)
summary(fit.red)

#7. Fitting an interaction model and comparing it with the
reduced model from 5
fit.int <- lm(crime_chg_per ~
child_per*lunch_part_per*crime_rate_per1000, data=denver)
anova(fit.red, fit.int)

#8. Predicting the change in crime rate for a community with
population = 7, children = 25%, lunch program = 55%
crime <- mean(denver$crime_rate_per1000)
newData <- data.frame(child_per = 25, lunch_part_per = 55,
crime_rate_per1000 = crime)
predict(fit.int, newData, interval="prediction") # PI
```