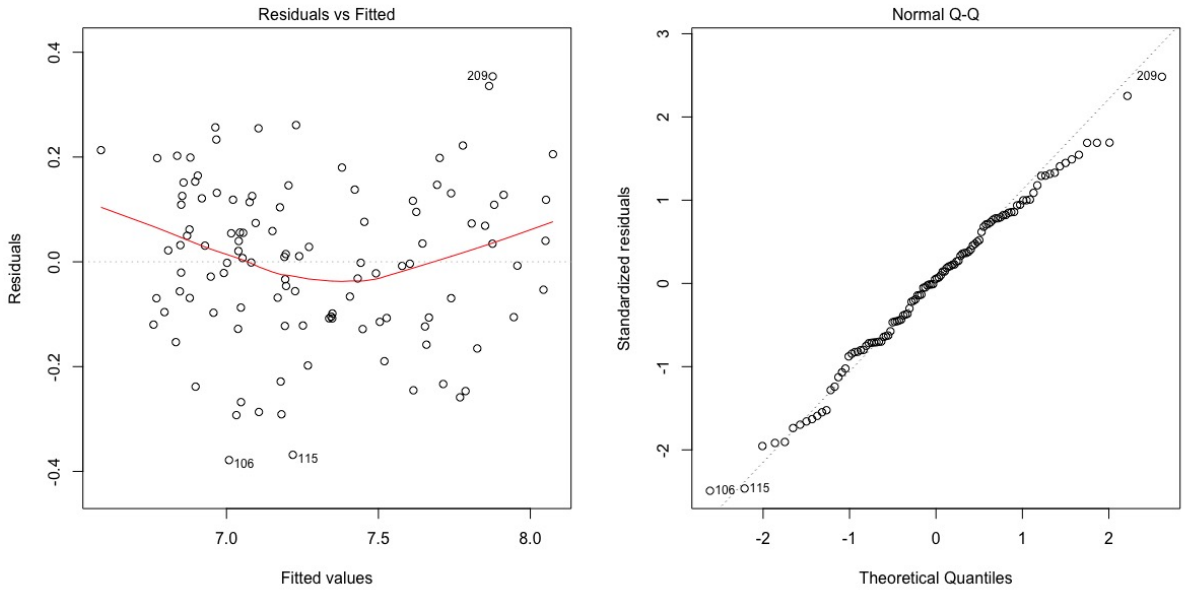
STA302 Assigment 3

Zikun CHEN
1001117882

September 7, 2018

A NFL Draft

1. Residual and QQ plots for MLR:



The residual plot shows the presence of a couple outliers present in our data. There is no sign of violation of normality or constant variance assumptions from the plots.
However, we cannot trust inference in this case because the p-values in the summary table are for type III tests, examining if each of the parameter are useful given we already have the other ones in the model. There might be extra predictors that we do not need. Another problem might occur when our predictors are correlated with each other. We need to run further tests and diagnostics to select a precise and appropriate model.

2. Summary table for the reduced model:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.74934    0.26951   2.780   0.0064
Yd40         0.55241    0.08377   6.594 1.58e-09
Shuttle      0.89171    0.10878   8.198 5.22e-13
```

3.

Test Statistic	Degrees of Freedom	p-value
1.07328	5 and 104	0.3795427

Conclusion: The p-value for partial F-test for additional predictors (height, weight, vertical jump, bench press, and broad jump) for the reduced model (with predictors 40-yard time and shuttle time) was greater than 0.1. We do not have enough evidence against the claim that additional predictors are all statistically insignificant. Thus, the inclusion of these predictors to the reduced multiple linear regression model is unnecessary. The reduced model and the full model are equivalent.

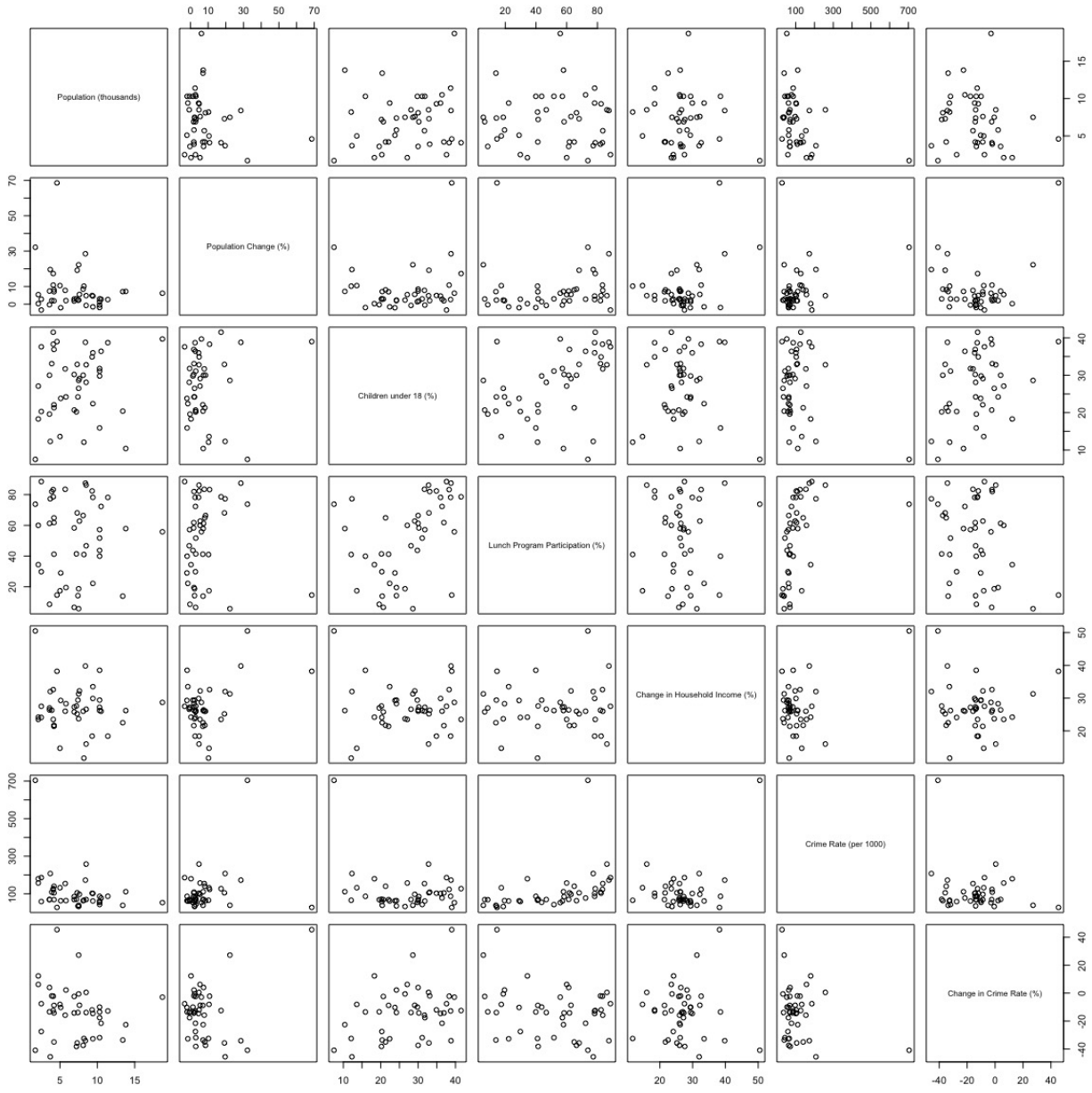
4. The coefficient of partial determination is 4.906807%, which is the percentage of the unexplained variation in the reduced model is explained by using the full model instead.

5. Weight is also useful in predicting 3-cone time on top of 40-yd time and shuttle time.
Summary table for MLR model fitting all the useful predictors:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.6583386  0.4063855   4.081 7.56e-05
Wt          0.0018057  0.0006898   2.618  0.00984
Yd40        0.3227538  0.1041097   3.100  0.00235
Shuttle     0.8328283  0.0988449   8.426 4.19e-14
```

B Denver

1. Pairwise scatterplot matrix for all variables:



From the scatterplot matrix, we can see that there is a community with a really high crime rate and one with a really large change in population. We will omit them from our following regression analysis.

2. Correlation matrix for the seven variables:

Predictors	Population	Population Change	Percentage of Children	Lunch Program Participation	Change in Household Income	Crime Rate	Change in Crime Rate
Population	1.000	-0.031	0.120	-0.020	0.092	-0.368	-0.193
Population Change	-0.031	1.000	0.100	0.233	0.105	0.267	-0.168
Percentage of Children	0.120	0.100	1.000	0.600	0.121	0.089	0.207
Lunch Program Participation	-0.020	0.233	0.600	1.000	0.007	0.606	-0.184
Change in Household Income	0.092	0.105	0.121	0.007	1.000	-0.091	-0.038
Crime Rate	-0.368	0.267	0.089	0.606	-0.091	1.000	0.021
Change in Crime Rate	-0.193	-0.168	0.207	-0.184	-0.038	0.021	1.000

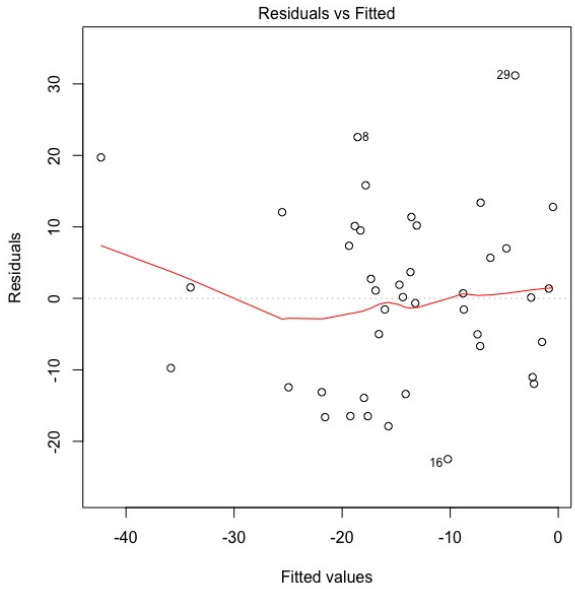
From looking at the above correlation matrix, we see that change in crime rate has a weak correlation with some of the other predictors like crime rate and percentage change in household income, which may not be significant enough to be included in the MLR model.
In addition, some of the intended predictor variables are correlated. For example, lunch program participation rate is relatively highly correlated with crime rate (with correlation coefficient of 0.606) and percentage of children (with correlation coefficient of 0.6) in the community. So there is a problem of multicollinearity when we fit a MLR including these variables as predictors.

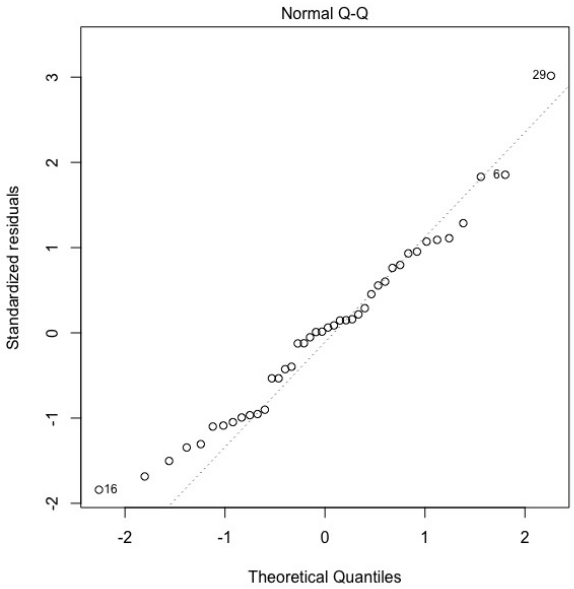
3. Summary table with VIFs:

	Estimate	Std. Error	t value	Pr(>|t|)	VIF
(Intercept)	-28.6174912	13.65274488	-2.0960980	0.0433720391	NA
Population	-0.5294447	0.65413039	-0.8093871	0.4237610032	1.257794
PopulationChange	-0.3643101	0.31336186	-1.1625860	0.2528635308	1.107021
Children	1.3362415	0.34344669	3.8906809	0.0004278115	1.944417
Lunch	-0.5238328	0.14271429	-3.6705001	0.0008003091	3.179425
IncomeChange	-0.1311643	0.37771820	-0.3472543	0.7304797713	1.044703
CrimeRate	0.1450314	0.06293124	2.3046012	0.0272399832	2.438403

Most of the predictors have a VIF slightly larger than 1. None of the predictors have a VIF of more than 10. Thus, we should not be worried about multicollinearity here.

4. Residual and normal QQ plots for the full MLR model:





Residual plot does not suggest violation of constant variance. There are some outliers present in the residual plot that might be influential.
The normal QQ plot suggests that the residuals follow a right-tailed distribution.

5.

Test Statistic	Degrees of Freedom	p-value
0.796155	3 and 35	0.5043457

Conclusion: The p-value for partial F-test for additional predictors (population, population change, change in household income) for the reduced model (with predictors percentage of children, lunch participation rate and crime rate) was greater than 0.1. We do not have enough evidence against the claim that all additional predictors are statistically insignificant. Thus, the inclusion of these predictors to the reduced multiple linear regression model is unnecessary. The reduced model and the full model are equivalent.
I prefer the simpler reduced model without the redundant predictors since the partial F-tests suggest that the full model and the interaction model are both essentially equivalent to the reduced model.
Summary table for the reduced model:

```
                Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -37.0740562 8.28513072 -4.474770 6.744452e-05
Children      1.3244224 0.33850655  3.912546 3.657212e-04
Lunch        -0.5576526 0.13777772 -4.047480 2.452195e-04
CrimeRate     0.1571348 0.05572025  2.820067 7.588123e-03
```

6. Interpretation of the coefficient of children under 18 precentage:
The change in crime rate will increase, on average by about 1.32442% if there is 1% increase

in the percentage of children under 18 in the community when lunch program participation rate and crime rate stays the same.
Interpretation of the coefficient of lunch program participation:
The change in crime rate will decrease, on average by about 0.55765% if there is 1% increase in lunch program participation rate in the community when percentage of children under 18 and crime rate stays the same.

7.

Test Statistic	Degrees of Freedom	p-value
1.555512	4 and 34	0.2085115

Conclusion: The p-value for partial F-test for additional interaction regressors for the reduced model (with predictors percentage of children, lunch participation Rate and crime rate) was greater than 0.1. We do not have enough evidence against the claim that all additional interaction regressors are statistically insignificant. Thus, the inclusion of these regressors to the reduced multiple linear regression model is unnecessary. The reduced model and the interaction model are equivalent.

8. Based on the preferred reduced model (with predictors percentage of children under 18, lunch participation rate and crime rate), we predict the new community to have a change in crime rate of -19.48809%. A 95% prediction interval for it is $[-46.28844, 7.312266]$.

Appendix R Code

```
## Part A ##
options(show.signif.stars = F)
#1
full <- nfl[complete.cases(nfl),]
fitf <- lm(Cone3 ~ Ht + Wt + Yd40 + Vertical + Bench + Broad + Shuttle, data = full)
hist(fitf$residuals, breaks=10)
par(mfrow=c(1, 2))
plot(fitf, 1)
plot(fitf, 2)
par(mfrow=c(1, 1))

#2 reduce model
summary(fitf)
fit <- fitf
max(as.data.frame(summary(fit)$coefficients)$"Pr(>|t|)")
# height has the highest p value, refitting MLR with height removed
fit <- lm(Cone3 ~ Wt + Yd40 + Vertical + Bench + Broad + Shuttle, data = full)
summary(fit)
max(as.data.frame(summary(fit)$coefficients)$"Pr(>|t|)")
# remove bench
fit <- lm(Cone3 ~ Wt + Yd40 + Vertical + Broad + Shuttle, data = full)
summary(fit)
max(as.data.frame(summary(fit)$coefficients)$"Pr(>|t|)")2
#remove broad
fit <- lm(Cone3 ~ Wt + Yd40 + Vertical + Shuttle, data = full)
summary(fit)
max(as.data.frame(summary(fit)$coefficients)$"Pr(>|t|)")
#remove vertical
fit <- lm(Cone3 ~ Wt + Yd40 + Shuttle, data = full)
summary(fit)
max(as.data.frame(summary(fit)$coefficients)$"Pr(>|t|)")
# remove weight
fit <- lm(Cone3 ~ Yd40 + Shuttle, data = full)
summary(fit)
as.data.frame(summary(fit)$coefficients)
rm(fit)

#3
fitr <- lm(Cone3 ~ Yd40 + Shuttle, data = full)
p=nrow(as.data.frame(summary(fitf)$coefficients))
q=nrow(as.data.frame(summary(fitr)$coefficients))
#df
p-q
nrow(full)-p
# f-statistics
anova(fitr, fitf)$F
```

```
# p-value
anova(fitr, fitf)$"Pr(>F)"

#4 finding coeff of partial determination
sser=tail(anova(fitr)$"Sum Sq", n=1)
ssef=tail(anova(fitf)$"Sum Sq", n=1)
(sser-ssef)/sser

#5 adding back rows with bench press missing
full5 <- subset(nfl, (!is.na(Ht)) & (!is.na(Wt)) & (!is.na(Yd40)) & (!is.na(Vertical))
                & (!is.na(Broad)) & (!is.na(Cone3)) & (!is.na(Shuttle)))
# full model without bench press score
fit5 <- lm(Cone3 ~ Ht + Wt + Yd40 + Vertical + Broad + Shuttle, data = full5)
summary(fit5)
max(as.data.frame(summary(fit5)$coefficients)$"Pr(>|t|)")
# remove height
fit5 <- lm(Cone3 ~ Wt + Yd40 + Vertical + Broad + Shuttle, data = full5)
summary(fit5)
max(as.data.frame(summary(fit5)$coefficients)$"Pr(>|t|)")
# remove broad
fit5 <- lm(Cone3 ~ Wt + Yd40 + Vertical + Shuttle, data = full5)
summary(fit5)
max(as.data.frame(summary(fit5)$coefficients)$"Pr(>|t|)")

# remove vertical
fit5 <- lm(Cone3 ~ Wt + Yd40 + Shuttle, data = full5)
summary(fit5)
max(as.data.frame(summary(fit5)$coefficients)$"Pr(>|t|)")
rm(fit5)

# Partial F-test
fitr<-lm(Cone3 ~ Wt + Yd40 + Shuttle, data = full5)
fitf <- lm(Cone3 ~ Ht + Wt + Yd40 + Vertical + Broad + Shuttle, data = full5)
# p-value for partial F-test
anova(fitr, fitf)$"Pr(>F)"


## Part B ##
rm(list=ls())
#1 scatterplot matrix
dv <- read.csv("Denver.csv", head= T, strip.white= T , stringsAsFactors = F)
names(dv) <- c("Population", "PopulationChange", "Children",
               "Lunch", "IncomeChange", "CrimeRate", "CrimeRateChange")
pairs(dv, labels = c("Population (thousands)", "Population Change (%)",
                     "Children under 18 (%)", "Lunch Program Participation (%)",
                     "Change in Household Income (%)", "Crime Rate (per 1000)",
                     "Change in Crime Rate (%)"))
# remove community with high crime rate
```

```
dv <- subset(dv, CrimeRate < max(dv$CrimeRate))
# remove community with large change in population
dv <- subset(dv, PopulationChange < max(dv$PopulationChange))

#2 correlation matrix
cormat <- as.data.frame(round(as.matrix(cor(dv)), digits=3))
rownames(cormat) <- c("Population (thousands)", "Population Change (%)",
                      "Children under 18 (%)", "Lunch Program Participation (%)",
                      "Change in Household Income (%)", "Crime Rate (per 1000)",
                      "Change in Crime Rate (%)")
colnames(cormat) <- c("Population (thousands)", "Population Change (%)",
                      "Children under 18 (%)", "Lunch Program Participation (%)",
                      "Change in Household Income (%)", "Crime Rate (per 1000)",
                      "Change in Crime Rate (%)")
cormat

#3 adding VIF
library(car)
fitb <- lm(CrimeRateChange ~ Population+PopulationChange+Children
           +Lunch+IncomeChange+CrimeRate, data = dv)
summary(fitb)
table<-as.data.frame(summary(fitb)$coefficients)
table$VIF <- c(NA ,as.data.frame(vif(fitb))$"vif(fitb)")
table

#4 Residual and Normal QQ plot
par(mfrow=c(1, 2))
plot(fitb, 1)
plot(fitb, 2)
par(mfrow=c(1, 1))

#5 reduced model
df <- subset(as.data.frame(summary(fitb)$coefficients))
names(df) <-  c(c(names(df)[1: length(names(df))-1]),"pvalue")
rownames(subset(df, pvalue < 0.05))
rm(df)

fitr <-lm(CrimeRateChange ~ Children + Lunch + CrimeRate, data = dv)
fitf <- fitb
p=nrow(as.data.frame(summary(fitf)$coefficients))
q=nrow(as.data.frame(summary(fitr)$coefficients))
#df
p-q
nrow(dv)-p
# f-statistics
anova(fitr, fitf)$F
# p-value
anova(fitr, fitf)$"Pr(>F)"
```

```
summary(fitr)$coefficients

#7 interaction model
fitrint <-lm(CrimeRateChange ~ Children * Lunch * CrimeRate, data = dv)
p=nrow(as.data.frame(summary(fitrint)$coefficients))
q=nrow(as.data.frame(summary(fitr)$coefficients))
#df
p-q
nrow(dv)-p
# f-statistics
anova(fitr, fitrint)$F
# p-value
anova(fitr, fitrint)$"Pr(>F)"

#8 prediction
newData =  data.frame(Children=25, Lunch=55, CrimeRate=mean(dv$CrimeRate))
predict(fitr, newData, interval="prediction")
```