**Ordinary Regression Model building Code**

Sree Sarada Alapaty

HAP 719

**Question 1.**

library(readr)

#To read CSV files

df <- read.csv("C:/Users/sarad/OneDrive - George Mason University - O365 Production/Documents/Nomissingfinal\_030123.csv")

names(df)

sum(is.na(df))

sapply(df,class)

##A)Initially explain variation in complete Series Vaccination rates by demographics(including age, race and gender)

# Fit the model with rescaled variables

model\_xa <- lm(Series\_Complete\_Pop\_Pct\_x~

Census2019\_5PlusPop\_x+Census2019\_5to17Pop\_x+Census2019\_12PlusPop\_x+

Census2019\_18PlusPop\_x+Census2019\_65PlusPop\_x+

below18yearsofage2019\_x+older65over2019\_x+below18yearsofage2020+

older65over2020 + below18yearsofageApril2022+

older65overApril2022 + Asian2019\_x +Asian2020+AsianApril2022 +

AmericanIndian\_AlaskaNative2019+AmericanIndian\_AlaskaNative2020+

AmericanIndian\_AlaskaNativeApri + Females2019\_x + Females2020+

FemalesApril2022,

data = df)

#Print summary of the model

summary(model\_xa)

#Calculate the percent of variation explained

total\_variation <- var(df$Series\_Complete\_Pop\_Pct\_x)

explained\_variation <- 1 - (var(model\_xa$residuals) / total\_variation)

#Print the percent of variation explained

cat("Percent of variation explained by demographics:",

round(explained\_variation \* 100, 2), "%\n")

#Percent of variation explained by demographics: 26.12 %

##B)Explain variation in complete Series Vaccination rates by demographics (age, race, gender),

# and social determinants (including high school completion rate, percent nor

# proficient inenglish, percent employed, percent of children in poverty,

# and media household income). Report the percent of variation explained.

names(df)

#Fit the Linear Regression Model

model\_xb <- lm(Series\_Complete\_Pop\_Pct\_x ~

Census2019\_5PlusPop\_x+Census2019\_5to17Pop\_x+Census2019\_12PlusPop\_x+

Census2019\_18PlusPop\_x+Census2019\_65PlusPop\_x+

below18yearsofage2019\_x+older65over2019\_x+below18yearsofage2020+

older65over2020 + below18yearsofageApril2022+older65overApril2022 +

Asian2019\_x+Asian2020 + AsianApril2022 +

AmericanIndian\_AlaskaNative2019+AmericanIndian\_AlaskaNative2020+

AmericanIndian\_AlaskaNativeApri + Females2019\_x + Females2020+

FemalesApril2022 +Highschoolcompletion2019\_x +Highschoolcompletion2020+

HighschoolcompletionApril2022+notproficientinEnglish2019\_x+notproficientinEnglish2020+

notproficientinEnglishApril2022+Unemployment2019\_x+Unemployment2020+

UnemploymentApril2022+Trafficvolume2019\_x+Trafficvolume2020+

TrafficvolumeApril2022+Childreninpoverty2019\_x+Childreninpoverty2020+

ChildreninpovertyApril2022+Childreninsingleparenthousehold+

Childreninsingleparenthouseho\_1+Childreninsingleparenthouseho\_2+

Medianhouseholdincome2019\_x+Medianhouseholdincome2020+MedianhouseholdincomeApril2022,

data = df)

#Fit the Model

summary(model\_xb)

#Include interaction items

model\_xb\_interaction<- lm(Series\_Complete\_Pop\_Pct\_x~

Census2019\_5PlusPop\_x+Census2019\_5to17Pop\_x+Census2019\_12PlusPop\_x+

Census2019\_18PlusPop\_x+Census2019\_65PlusPop\_x+

below18yearsofage2019\_x+older65over2019\_x+below18yearsofage2020+

older65over2020 + below18yearsofageApril2022+

older65overApril2022 + Asian2019\_x +Asian2020 + AsianApril2022 +

AmericanIndian\_AlaskaNative2019+AmericanIndian\_AlaskaNative2020+

AmericanIndian\_AlaskaNativeApri + Females2019\_x + Females2020+

FemalesApril2022 +Highschoolcompletion2019\_x +Highschoolcompletion2020+

HighschoolcompletionApril2022+notproficientinEnglish2019\_x+notproficientinEnglish2020+

notproficientinEnglishApril2022+Unemployment2019\_x+Unemployment2020+

UnemploymentApril2022+Trafficvolume2019\_x+Trafficvolume2020+

TrafficvolumeApril2022+Childreninpoverty2019\_x+Childreninpoverty2020+

ChildreninpovertyApril2022+Childreninsingleparenthousehold+

Childreninsingleparenthouseho\_1+Childreninsingleparenthouseho\_2+

Medianhouseholdincome2019\_x+Medianhouseholdincome2020+MedianhouseholdincomeApril2022+

Unemployment2019\_x\*Trafficvolume2019\_x+

Unemployment2020\*Trafficvolume2020+

UnemploymentApril2022\*TrafficvolumeApril2022+

Childreninpoverty2019\_x\*Childreninsingleparenthousehold+

Childreninpoverty2020\*Childreninsingleparenthouseho\_1+

ChildreninpovertyApril2022\*Childreninsingleparenthouseho\_2+

Childreninpoverty2019\_x\*Medianhouseholdincome2019\_x+

Childreninpoverty2020\*Medianhouseholdincome2020+

ChildreninpovertyApril2022\*MedianhouseholdincomeApril2022,

data = df)

#Fit the Model

summary(model\_xb\_interaction)

#Calculate the percent of variation explained

total\_variation <- var(df$Series\_Complete\_Pop\_Pct\_x)

explained\_variation <- 1 - (var(model\_xb$residuals) / total\_variation)

#Print the percent of variation explained

cat("Percent of variation explained by demographics and social determinates:",

round(explained\_variation \* 100, 2), "%\n")

#Percent of variation explained by demographics and social determinates: 46.2 %

#Calculate the percent of variation explained

total\_variation <- var(df$Series\_Complete\_Pop\_Pct\_x)

explained\_variation <- 1 - (var(model\_xb\_interaction$residuals) / total\_variation)

#Print the percent of variation explained

cat("Percent of variation explained by demographics and scoial determinants

including interactionitems:", round(explained\_variation \* 100, 2), "%\n")

#Percent of variation explained by demographics and scoial determinants

#including interactionitems: 48.33 %

##C)Explain variation in Complete Series Vaccination rates by demographics (age, race, gender),

#social deteminants(including high school completion rate, percent not proficient in English,

#percent employed, percnet of children in poverty, median household income) and health of residents

#(including percent population disabled, life expectancy, percent population having premature morbidity).

#Report the percent of variation explained.

names(df)

#Fit the Linear Regression Model without Alias

model\_xc <- lm(Series\_Complete\_Pop\_Pct\_x~

Census2019\_5PlusPop\_x+Census2019\_5to17Pop\_x+Census2019\_12PlusPop\_x+

Census2019\_18PlusPop\_x+Census2019\_65PlusPop\_x+

below18yearsofage2019\_x+older65over2019\_x+below18yearsofage2020+

older65over2020 + below18yearsofageApril2022+

older65overApril2022 + Asian2019\_x +Asian2020+AsianApril2022 +

AmericanIndian\_AlaskaNative2019+AmericanIndian\_AlaskaNative2020+

AmericanIndian\_AlaskaNativeApri + Females2019\_x + Females2020+

FemalesApril2022 +Highschoolcompletion2019\_x +Highschoolcompletion2020+

HighschoolcompletionApril2022+notproficientinEnglish2019\_x+notproficientinEnglish2020+

notproficientinEnglishApril2022+Unemployment2019\_x+Unemployment2020+

UnemploymentApril2022+Trafficvolume2019\_x+Trafficvolume2020+

TrafficvolumeApril2022+Childreninpoverty2019\_x+Childreninpoverty2020+

ChildreninpovertyApril2022+Childreninsingleparenthousehold+

Childreninsingleparenthouseho\_1+Childreninsingleparenthouseho\_2+

Medianhouseholdincome2019\_x+Medianhouseholdincome2020+MedianhouseholdincomeApril2022+

Disability2019\_x+Disability2022+DisabilityApril2022+

Lifeexpectancy2019\_x+Lifeexpectancy2020+LifeexpectancyApril2022+

Prematureageadjustedmortality20+Prematureageadjustedmortality\_1+

PrematureageadjustedmortalityAp,

data = df)

#Fit the Model

summary(model\_xc)

#Calculate the percent of variation explained

total\_variation <- var(df$Series\_Complete\_Pop\_Pct\_x)

explained\_variation <- 1 - (var(model\_xc$residuals) / total\_variation)

#Print the percent of variation explained

cat("Percentof variation explained by demographics, social determinants and

health of residents:", round(explained\_variation \* 100, 2), "%\n")

#Percentof variation explained by demographics, social determinants and

#health of residents: 47.17 %

##D)Explain variation in Complete Series Vaccination rates by demographic (age, race, gender),

#social deteminants(including high school completion rate, percent not proficient in English,

#percent employed, percnet of children in poverty, median household income) and health of residents

#(including percent population disabled, life expectancy, percent population having premature morbidity),

#and political leaning of the population (including republican leaning, democrat leaning).

#Report the percent of variation explained.

names(df)

#Fit the Linear Regression Model without Alias

model\_xd <- lm(Series\_Complete\_Pop\_Pct\_x~

Census2019\_5PlusPop\_x+Census2019\_5to17Pop\_x+Census2019\_12PlusPop\_x+

Census2019\_18PlusPop\_x+Census2019\_65PlusPop\_x+

below18yearsofage2019\_x+older65over2019\_x+below18yearsofage2020+

older65over2020 + below18yearsofageApril2022+

older65overApril2022 + Asian2019\_x +Asian2020+AsianApril2022 +

AmericanIndian\_AlaskaNative2019+AmericanIndian\_AlaskaNative2020+

AmericanIndian\_AlaskaNativeApri + Females2019\_x + Females2020+

FemalesApril2022 +Highschoolcompletion2019\_x +Highschoolcompletion2020+

HighschoolcompletionApril2022+notproficientinEnglish2019\_x+notproficientinEnglish2020+

notproficientinEnglishApril2022+Unemployment2019\_x+Unemployment2020+

UnemploymentApril2022+Trafficvolume2019\_x+Trafficvolume2020+

TrafficvolumeApril2022+Childreninpoverty2019\_x+Childreninpoverty2020+

ChildreninpovertyApril2022+Childreninsingleparenthousehold+

Childreninsingleparenthouseho\_1+Childreninsingleparenthouseho\_2+

Medianhouseholdincome2019\_x+Medianhouseholdincome2020+MedianhouseholdincomeApril2022+

Disability2019\_x+Disability2022+DisabilityApril2022+

Lifeexpectancy2019\_x+Lifeexpectancy2020+LifeexpectancyApril2022+

Prematureageadjustedmortality20+Prematureageadjustedmortality\_1+

PrematureageadjustedmortalityAp+Republicanpercent+DemocraticPercent+

Homeownership2019\_x+Homeownership2020+HomeownershipApril2022+

Povertyrat\_2019+Povertyrat2020+PovertyratApril2022+

Severehousingcostburden2019\_x+Severehousingcostburden2020+SeverehousingcostburdenApril202+

ResidentialsegregationBlackWhit+ResidentialsegregationBlackWh\_1+

ResidentialsegregationBlackWh\_2+Rural2010\_x+Rural2010\_1+RuralApril2022,

data = df)

#Fit the Model

summary(model\_xd)

#Include interaction items

model\_xd\_interaction<- lm(Series\_Complete\_Pop\_Pct\_x~

Census2019\_5PlusPop\_x+Census2019\_5to17Pop\_x+Census2019\_12PlusPop\_x+

Census2019\_18PlusPop\_x+Census2019\_65PlusPop\_x+

below18yearsofage2019\_x+older65over2019\_x+below18yearsofage2020+

older65over2020 + below18yearsofageApril2022+

older65overApril2022 + Asian2019\_x +Asian2020 + AsianApril2022 +

AmericanIndian\_AlaskaNative2019+AmericanIndian\_AlaskaNative2020+

AmericanIndian\_AlaskaNativeApri + Females2019\_x + Females2020+FemalesApril2022+

Highschoolcompletion2019\_x +Highschoolcompletion2020+HighschoolcompletionApril2022+

notproficientinEnglish2019\_x+notproficientinEnglish2020+notproficientinEnglishApril2022+

Unemployment2019\_x+Unemployment2020+UnemploymentApril2022+

Trafficvolume2019\_x+Trafficvolume2020+TrafficvolumeApril2022+

Childreninpoverty2019\_x+Childreninpoverty2020+ChildreninpovertyApril2022+

Childreninsingleparenthousehold+Childreninsingleparenthouseho\_1+Childreninsingleparenthouseho\_2+

Medianhouseholdincome2019\_x+Medianhouseholdincome2020+MedianhouseholdincomeApril2022+

Disability2019\_x+Disability2022+DisabilityApril2022+

Lifeexpectancy2019\_x+Lifeexpectancy2020+LifeexpectancyApril2022+

Prematureageadjustedmortality20+Prematureageadjustedmortality\_1+

PrematureageadjustedmortalityAp+Republicanpercent+DemocraticPercent+

Homeownership2019\_x+Homeownership2020+HomeownershipApril2022+

Povertyrat\_2019+Povertyrat2020+PovertyratApril2022+

Severehousingcostburden2019\_x+Severehousingcostburden2020+SeverehousingcostburdenApril202+

ResidentialsegregationBlackWhit+ResidentialsegregationBlackWh\_1+

ResidentialsegregationBlackWh\_2+Rural2010\_x+Rural2010\_1+RuralApril2022+

Unemployment2019\_x\*Trafficvolume2019\_x+

Unemployment2020\*Trafficvolume2020+

UnemploymentApril2022\*TrafficvolumeApril2022+

Childreninpoverty2019\_x\*Childreninsingleparenthousehold+

Childreninpoverty2020\*Childreninsingleparenthouseho\_1+

ChildreninpovertyApril2022\*Childreninsingleparenthouseho\_2+

Childreninpoverty2019\_x\*Medianhouseholdincome2019\_x+

Childreninpoverty2020\*Medianhouseholdincome2020+

ChildreninpovertyApril2022\*MedianhouseholdincomeApril2022+

Republicanpercent\*Trafficvolume2019\_x+

Republicanpercent\*Trafficvolume2020+

Republicanpercent\*TrafficvolumeApril2022+

Republicanpercent\*Homeownership2019\_x+

Republicanpercent\*Homeownership2020+

Republicanpercent\*HomeownershipApril2022+

Republicanpercent\*Severehousingcostburden2019\_x+

Republicanpercent\*Severehousingcostburden2020+

Republicanpercent\*SeverehousingcostburdenApril202+

Republicanpercent\*older65over2019\_x+

Republicanpercent\*older65over2020+

Republicanpercent\*older65overApril2022+

Republicanpercent\*below18yearsofage2019\_x+

Republicanpercent\*below18yearsofage2020+

Republicanpercent\*below18yearsofageApril2022+

Republicanpercent\*ResidentialsegregationBlackWhit+

Republicanpercent\*ResidentialsegregationBlackWh\_1+

Republicanpercent\*ResidentialsegregationBlackWh\_2+

Republicanpercent\*Rural2010\_x+

Republicanpercent\*Rural2010\_1+

Republicanpercent\*RuralApril2022,

data = df)

#Fit the Model

summary(model\_xd\_interaction)

#Calculate the percent of variation explained

total\_variation <- var(df$Series\_Complete\_Pop\_Pct\_x)

explained\_variation <- 1 - (var(model\_xd$residuals) / total\_variation)

#print the percent of variation explained

cat("Percent of variation explained by demographics, social determinants,

health of residents and political leaning of the population:",

round(explained\_variation \* 100, 2), "%\n")

#Percent of variation explained by demographics, social determinants,

#health of residents and political leaning of the population: 50.9 %

#Calculate the percent of variation explained

total\_variation <- var(df$Series\_Complete\_Pop\_Pct\_x)

explained\_variation <- 1 - (var(model\_xd\_interaction$residuals) / total\_variation)

#Print the percent of variation explained

cat("Percent of variation explained by demographics, social determinants,

health of residents and political leaning of the population including interaction items:",

round(explained\_variation \* 100, 2), "%\n")

#Percent of variation explained by demographics, social determinants,

#health of residents and political leaning of the population including interaction items: 52.49 %

######

#Republicanpercent: Estimate = 57.80, t value = 1.440, Pr(>|t|) = 0.150

#Democraticpercent: Estimate = 18.24, t value = 2.485, Pr(>|t|) = 0.013

#Given the p-values, DemocraticPercent appears to be statistically significant (p < 0.05)

#Suggesting that there is evidence to reject the null hypothesis that the coefficient

#for DemocraticPercent is zero.

#In other words, there is evidence to suggest that a county's Democratic leaning is associated

#with changes in the vaccination rate.

#On the other hand, the p-value for RepublicanPercent is 0.150, which is greater than 0.05.

#While the coefficient is positive, suggesting a positive associaiton,

#the evidence for the association is not strong enough to reject the null hypothesis.

##E)Does a county's political leaning affect vaccination rates.

#In Conclusion, based on the results of the regression, there is some evidence to

#suggest that a county's political leaning, specifically leaning towards Democratic,

#is associated with changes in the COVID-19 vaccination rate.

#However, the association with Republican leaning is not statistically significant

#at the conventional Significance level (p > 0.05).

**Question 2.**

library(dplyr)

data = read.csv("C:\\Users\\sarad\\OneDrive - George Mason University - O365 Production\\Documents\\DATA\_2017 (Diabetes).csv")

#Checking the column names

column\_names <- names(data)

#Creating a new dataframe considering variables from 2015

data\_2015<- data%>% select(American\_Indian\_and\_Alaskan\_Native\_2015, obesity\_2015,Limited\_access\_to\_healthy\_foods\_2015,

Diabetes\_2015, Poor\_or\_fair\_health\_2015, Frequent\_physical\_distress\_2015, Median\_household\_income\_2015,

Hispanic\_2015, Children\_in\_poverty\_2015, Adult\_smoking\_2015, Inactivity\_2015)

#Multiple regression model to determine the percent of variation explained

model <- lm(Diabetes\_2015 ~., data = data\_2015)

summary(model)

#Creating a new dataframe considering variables from 2016

data\_2016<- data%>%

select (American\_Indian\_and\_Alaskan\_Native\_2016, obesity\_2016, Diabetes\_2016, Poor\_or\_fair\_health\_2016, Inactivity\_2016)

#Multiple regression model to determine the percent of variation explained

model1 <- lm(Diabetes\_2016 ~., data = data\_2016)

summary(model1)

#Combining the data from two years using the required variables

combined\_data <- cbind(data\_2015, data\_2016)

#Multiple regression model to determine the percent of variation explained

model3 <- lm(Diabetes\_2016 ~., data = combined\_data)

summary(model3)

#Diagnostic plots

par(mfrow = c(2, 2))

plot(model)