proj2

Dylan M Ang

3/5/2022

Background

Part 1

a

Total Population $(X_{1,1})$

```
##
##   The decimal point is 6 digit(s) to the right of the |
##
##    0 | 1111111111111111111111111111111111111111111111111111111111111111+254
##    0 | 5555555555555555555555555666666666666666777777777777777777778888888888
##    1 | 000000122233333444
##    1 | 55699
##    2 | 1134
##    2 | 58
##    3 |
##    3 |
##    4 |
##    4 |
##    5 | 1
##    5 |
##    6 |
##    6 |
##    7 |
##    7 |
##    8 |
##    8 | 9
```

Land Area $(X_{1,2})$

```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##     0 | 000011111111111112222222222222222222233333333333333333333333333444444+252
##     1 | 00000000000000011111111111111112222222222233333344445555666677778889999
##     2 | 0001111466778
##     3 | 3344688
##     4 | 00122368
##     5 | 45
##     6 | 023
##     7 | 29
##     8 | 11
##     9 | 22
##    10 |
##    11 |
##    12 |
##    13 |
##    14 |
##    15 |
##    16 |
##    17 |
##    18 |
##    19 |
##    20 | 1
```

Total Personal Income $(X_{1,3})$ and $(X_{2,3})$

```
##
##   The decimal point is 4 digit(s) to the right of the |
##
##     0 | 111111111111122222222222222222222222222222222222222222222222222222+263
##     1 | 0000000000000111111111122222333334444445555555567788888888999
##     2 | 001111233344477788899
##     3 | 0255678899
##     4 | 19
##     5 | 59
##     6 |
##     7 |
##     8 |
##     9 |
##    10 |
##    11 | 1
##    12 |
##    13 |
##    14 |
##    15 |
##    16 |
##    17 |
##    18 | 4
```

Population Density $(X_{2,1})$

```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##     0 | 000000000000000001111111111111111111111111111111111111111111111111111+321
##     2 | 00001112233456700111145
##     4 | 05884
##     6 | 2464
##     8 | 19
##    10 | 378
##    12 |
##    14 | 4
##    16 |
##    18 |
##    20 |
##    22 |
##    24 |
##    26 |
##    28 |
##    30 |
##    32 | 4
```

Total Population over 64 $(X_{2,2})$

```
##
##   The decimal point is at the |
##
##     2 | 0
##     4 | 47890389
```

```
##      6 | 112345567799013456678899
##      8 | 001122222333344445556667777788888999900022223333334444444555566677
##     10 | 00011111122222222233333334444445555555566666666677777778888888888899999+36
##     12 | 00000000011111222233333333334444455555555666666777777777888899900000000+36
##     14 | 000011111112233344444555677889000000011122223455667778
##     16 | 12556699901122345
##     18 | 06778
##     20 | 070
##     22 | 018828
##     24 | 47
##     26 | 055
##     28 | 1
##     30 | 7
##     32 | 138
```

Stem and Leaf plots tell us about the distribution of the data. In this case we can see that all of the
predictor variable data is right-skewed. We see that Total Population, Land Area, Total Personal Income,
and Population Density all have outliers.

b

Scatter Plot Matrix (Model 1)


```
##                             Active Physicians (Y) Total Population (X1)
## Active Physicians (Y)                 1.00000000            0.9402486
## Total Population (X1)                 0.94024859            1.0000000
## Land Area (X2)                        0.07807466            0.1730834
## Total Personal Income (X3)            0.94811057            0.9867476
##                             Land Area (X2) Total Personal Income (X3)
## Active Physicians (Y)           0.07807466                  0.9481106
## Total Population (X1)           0.17308335                  0.9867476
## Land Area (X2)                 1.00000000                  0.1270743
## Total Personal Income (X3)     0.12707426                  1.0000000
```

Total Population (X1) and Total Personal Income (X3) are highly correlated to The Number of Active Physicians (Y) with correlation coefficients of 0.940 and 0.948, respectively. This means there is a strong linear relationship between the predictor variables and number of active physicians. Land Area (X2) has a weak correlation with Active Physicians (Y) with a coefficient of 0.078. This means there is a weak linear relationship between land area and the number of active physicians. Based on the scatter plot, Total Population and Total Personal Income are positively correlated with Active Physicians, meaning that when they increase, the number of Active Physicians increases as well.

Scatter Plot Matrix (Model 2)
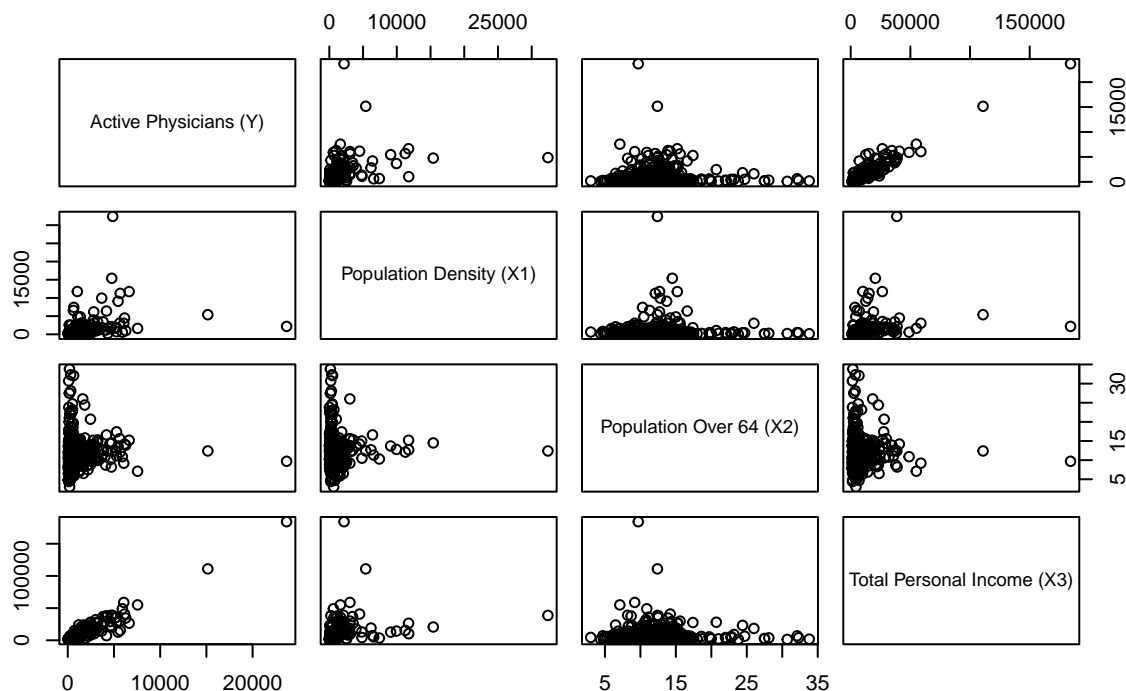

```
##                              Active Physicians (Y) Population Density (X1)
## Active Physicians (Y)                  1.00000000             0.40643863
## Population Density (X1)                0.40643863             1.00000000
## Population Over 64 (X2)               -0.00312863             0.02918445
## Total Personal Income (X3)            0.94811057             0.31620475
##                              Population Over 64 (X2) Total Personal Income (X3)
## Active Physicians (Y)                  -0.00312863                 0.94811057
## Population Density (X1)                 0.02918445                 0.31620475
## Population Over 64 (X2)                 1.00000000                -0.02273315
## Total Personal Income (X3)            -0.02273315                 1.00000000
```

Of the three predictor variables, Total Personal Income (X3) has the strongest linear relationship with the number of Active Physicians (Y) with a correlation coefficient of 0.948. Then, Population Density (X1) has a weaker linear relationship with Active Physicians (Y) with a correlation coefficient of 0.406. Finally, Population Over 64 (X2) has the weakest linear relationship with Active Physicians (Y) with a correlation coefficient of -0.00312. X2 also has the only negative correlation coefficient. Based on the scatter plot, Population Density and Total Personal Income are positively correlated with Active Physicians (Y). Therefore, as those predictors increase, the number of active physicians will also increase.

c

Model 1

$$Y = -13.316 + 0.001X_1 + -0.066X_2 + 0.094X_3$$

Model 2

$$Y = -170.574 + 0.096X_1 + 6.34X_2 + 0.127X_3$$

d

$$\text{Model 1 } R^2 : \quad 0.9026432$$
$$\text{Model 2 } R^2 : \quad 0.9117491$$

Based on the R^2 value, Model 2 is a slightly better model, but they are close.
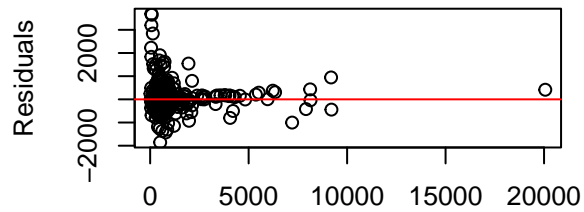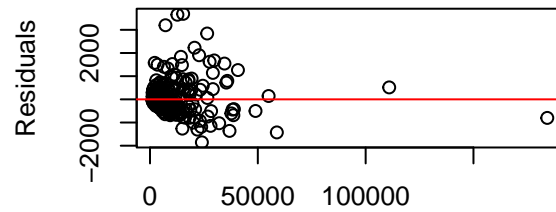
e: Model 1

Model 1: Residuals ~ Fitted values
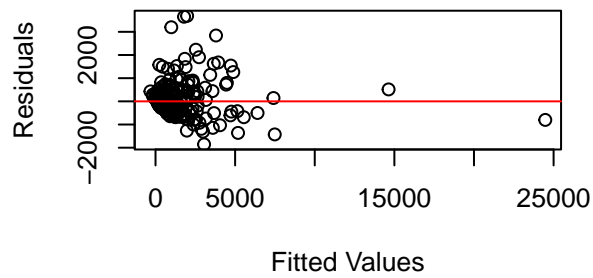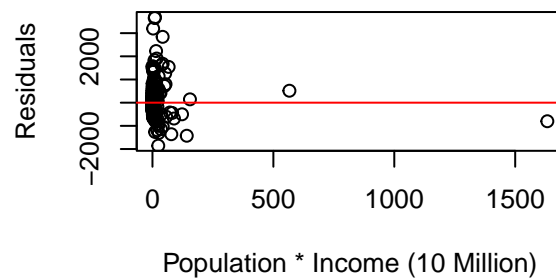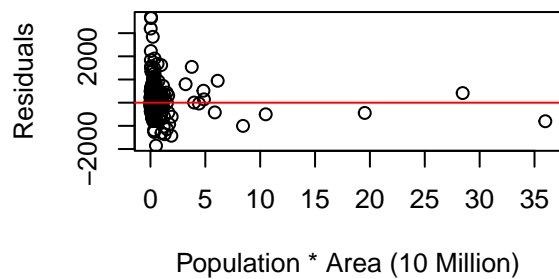

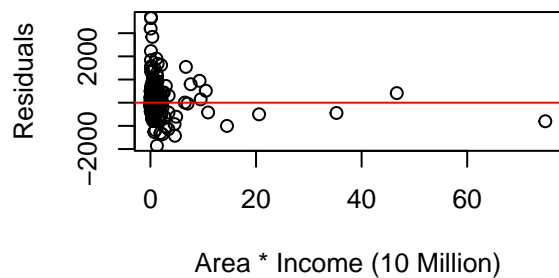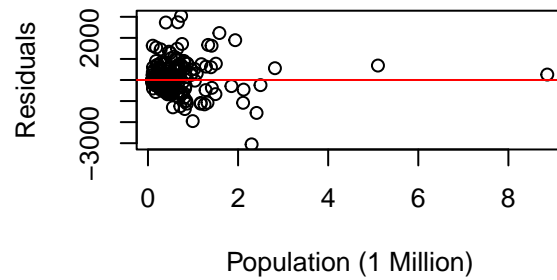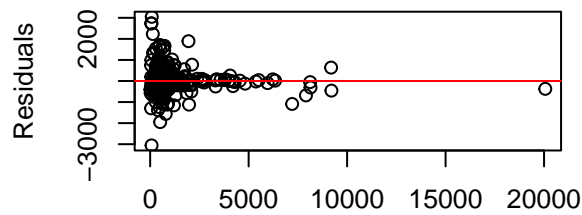**Model 1: Residuals ~ Population**
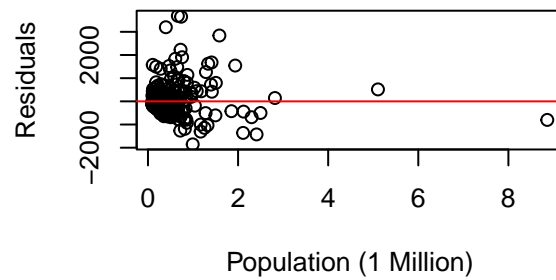
Model 1: Residuals ~ Area

Model 1: Residuals ~ Income

Model 1: Residuals ~ Population * Are

Model 1: Residuals ~ Population * Incor

Model 1: Residuals ~ Area * Income

Model 2
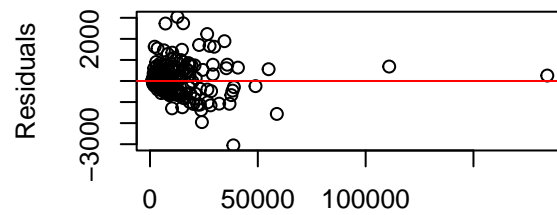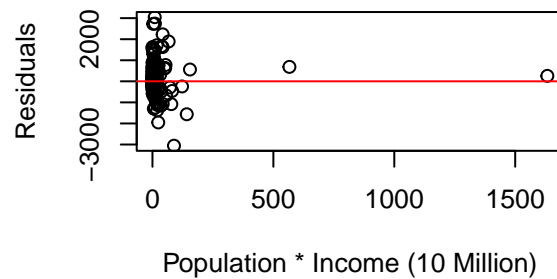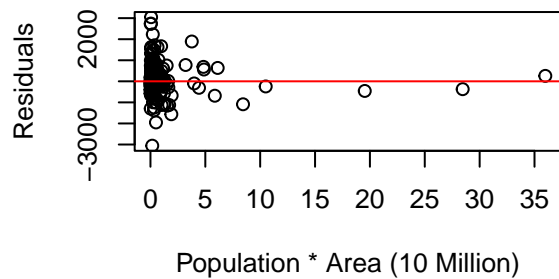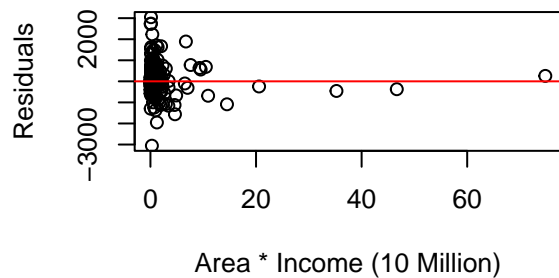
Model 2: Residuals ~ Fitted values

Fitted Values

Model 2: Residuals ~ Population

Population (1 Million)

Model 2: Residuals ~ Area

Land Area

Model 2: Residuals ~ Income

Personal Income

Model 2: Residuals ~ Population * Area

Population * Area (10 Million)

Model 2: Residuals ~ Population * Income

Population * Income (10 Million)

Model 1: Residuals ~ Area * Income

Area * Income (10 Million)

f

Model 1

$$Y_{1,2} = -84.914 + 0.001x_1 + -0.021x_2 + 0.089x_3 + -0.00000005x_1x_2 \tag{1}$$
$$Y_{1,3} = -58.222 + 0.001x_1 + 0.093x_2 + -0.069x_3 + -0.000000001x_1x_3 \tag{2}$$
$$Y_{2,3} = -94.964 + -0.021x_1 + 0.09x_2 + 0.001x_3 + -0.00000325x_2x_3 \tag{3}$$

Model 2

$$Y_{1,2} = 133.388 + -0.476x_1 + -17.763x_2 + 0.128x_3 + 0.04428047x_1x_2 \tag{4}$$
$$Y_{1,3} = -252.911 + 0.192x_1 + 0.134x_2 + 6.371x_3 + -0.0000038x_1x_3 \tag{5}$$
$$Y_{2,3} = -84.499 + -1.454x_1 + 0.113x_2 + 0.092x_3 + 0.00127165x_2x_3 \tag{6}$$

R^2 values of two factor interaction terms in Model 1

```
## (R2 of X1*X2) (R2 of X1*X3) (R2 of X2*X3)
##     0.9039154      0.9036285      0.9043730
```

R^2 values of two factor interaction terms in Model 2

```
## (R2 of X1*X2) (R2 of X1*X3) (R2 of X2*X3)
##     0.9190910      0.9164615      0.9122407
```

For Model 1, the model with the $x_2 * x_3$ interaction term had the highest R^2 value of the other possible interaction models. Therefore, this is the preferable model, since a higher proportion of the variability in our response variable (number of active physicians) is explained by our predictors.

For Model 2, the model with the $x_1 * x_2$ interaction term had the highest R^2 value of the other possible interaction models. Therefore, this is the preferable model, since a higher proportion of the variability in our response variable (number of active physicians) is explained by our predictors.

Part 2

a

```
## [1] 0.028824954 0.003842367 0.553818175
```

$$R^2_{Y,3|1,2} = 0.0288$$
$$R^2_{Y,4|1,2} = 0.0038$$
$$R^2_{Y,5|1,2} = 0.5538$$

Part 3

Appendix

```r
knitr::opts_chunk$set(echo = FALSE)
cdi_data = read.table("./CDI.txt")
phy = cdi_data$V8 # Number of Active Physicians
pop = cdi_data$V5 # Total Population
are = cdi_data$V4 # Land Area
inc = cdi_data$V16 # total personal income
den = pop / are # Population density
sen = cdi_data$V7
# par(mfrow = c(1,2))
# stem(pop, width = 20)
# stem(are, width = 20)
# stem(inc, width = 20)
# stem(den, width = 20)
# stem(sen, width = 20)
stem(pop)
stem(are)
stem(inc)
stem(den)
stem(sen)
# Create Dataframes
model1 = data.frame(Y = phy, X1 = pop, X2 = are, X3 = inc)
colnames(model1) = c("Active Physicians (Y)", "Total Population (X1)", "Land Area (X2)", "Total Personal

model2 = data.frame(Y = phy, X1 = den, X2 = sen, X3 = inc)
colnames(model2) = c("Active Physicians (Y)", "Population Density (X1)", "Population Over 64 (X2)", "Total
pairs(model1, main = "Scatter Plot Matrix (Model 1)")
cor(model1)
pairs(model2, main = "Scatter Plot Matrix (Model 2)")
cor(model2)
# Get models
fit1 = lm(model1)
fit2 = lm(model2)

f1s = summary(fit1)
f2s = summary(fit2)

betas1 = fit1$coefficients
betas2 = fit2$coefficients
slope = betas1[2] + betas1[3] + betas1[4]
int = betas1[1]
model1.R2 = f1s$r.squared
model2.R2 = f2s$r.squared
# Get residuals
m1.resid = f1s$residuals

m1.yhat = fit1$fitted.values

par(mfrow = c(2,2))
```

```r
# Residual Plots
plot(m1.resid ~ m1.yhat,
     main = "Model 1: Residuals ~ Fitted values", xlab = "Fitted Values", ylab = "Residuals")
abline(0,0,col = "red")

# Predictors
plot(x = (pop)/1000000,y = m1.resid,
     main = "Model 1: Residuals ~ Population", xlab = "Population (1 Million)", ylab = "Residuals")
abline(0,0,col = "red")

plot(m1.resid ~ are,
     main = "Model 1: Residuals ~ Area", xlab = "Land Area", ylab = "Residuals")
abline(0,0,col = "red")

plot(m1.resid ~ inc,
     main = "Model 1: Residuals ~ Income", xlab = "Personal Income", ylab = "Residuals")
abline(0,0,col = "red")

par(mfrow = c(2,2))

# Two Factor
plot(x = (pop/100000) * (are/10000), y = m1.resid,
     main = "Model 1: Residuals ~ Population * Area", xlab = "Population * Area (10 Million)", ylab = "
abline(0,0,col = "red")

plot(x = (pop/100000) * (inc/10000), y = m1.resid,
     main = "Model 1: Residuals ~ Population * Income", xlab = "Population * Income (10 Million)", ylab
abline(0,0,col = "red")

plot(x = (are * inc)/10000000, y = m1.resid,
     main = "Model 1: Residuals ~ Area * Income", xlab = "Area * Income (10 Million)", ylab = "Residuals
abline(0,0,col = "red")

# Get residuals
m2.resid = f2s$residuals

m2.yhat = fit2$fitted.values

par(mfrow = c(2,2))

# Residual Plots
plot(m2.resid ~ m2.yhat,
     main = "Model 2: Residuals ~ Fitted values", xlab = "Fitted Values", ylab = "Residuals")
abline(0,0,col = "red")

# Predictors
plot(x = (pop)/1000000,y = m2.resid,
     main = "Model 2: Residuals ~ Population", xlab = "Population (1 Million)", ylab = "Residuals")
abline(0,0,col = "red")

plot(m2.resid ~ are,
     main = "Model 2: Residuals ~ Area", xlab = "Land Area", ylab = "Residuals")
abline(0,0,col = "red")
```

```r
plot(m2.resid ~ inc,
     main = "Model 2: Residuals ~ Income", xlab = "Personal Income", ylab = "Residuals")
abline(0,0,col = "red")

par(mfrow = c(2,2))

# Two Factor
plot(x = (pop/100000) * (are/10000), y = m2.resid,
     main = "Model 2: Residuals ~ Population * Area", xlab = "Population * Area (10 Million)", ylab = "
abline(0,0,col = "red")

plot(x = (pop/100000) * (inc/10000), y = m2.resid,
     main = "Model 2: Residuals ~ Population * Income", xlab = "Population * Income (10 Million)", ylab
abline(0,0,col = "red")

plot(x = (are * inc)/10000000, y = m2.resid,
     main = "Model 1: Residuals ~ Area * Income", xlab = "Area * Income (10 Million)", ylab = "Residuals
abline(0,0,col = "red")

tf1_12 <- lm(phy ~ pop*are+inc)
tf1_13 <- lm(phy ~ pop*inc+are)
tf1_23 <- lm(phy ~ are*inc+pop)

r2_12 <- summary(tf1_12)$r.squared
r2_13 <- summary(tf1_13)$r.squared
r2_23 <- summary(tf1_23)$r.squared

res1 <- c(a = r2_12, b = r2_13, c = r2_23)
names(res1) <- c("(R2 of X1*X2)", "(R2 of X1*X3)", "(R2 of X2*X3)")

tf2_12 <- lm(phy ~ den*sen+inc)
tf2_13 <- lm(phy ~ den*inc+sen)
tf2_23 <- lm(phy ~ sen*inc+den)

r2_12 <- summary(tf2_12)$r.squared
r2_13 <- summary(tf2_13)$r.squared
r2_23 <- summary(tf2_23)$r.squared

res2 <- c(a = r2_12, b = r2_13, c = r2_23)
names(res2) <- c("(R2 of X1*X2)", "(R2 of X1*X3)", "(R2 of X2*X3)")
tf1_b12 <- tf1_12$coefficients
tf1_b13 <- tf1_13$coefficients
tf1_b23 <- tf1_23$coefficients
tf2_b12 <- tf2_12$coefficients
tf2_b13 <- tf2_13$coefficients
tf2_b23 <- tf2_23$coefficients
res1
res2
# already have other predictors
bed <- cdi_data$V9
fit  <- lm(phy ~ pop+inc)
fit3 <- lm(phy ~ pop+inc+are)
fit4 <- lm(phy ~ pop+inc+sen)
```

```
fit5 <- lm(phy ~ pop+inc+bed)

SSE <- function(fit) {
  tail(anova(fit)$`Sum Sq`, n = 1)
}

r2_3 = 1 - SSE(fit3)/SSE(fit)
r2_4 = 1 - SSE(fit4)/SSE(fit)
r2_5 = 1 - SSE(fit5)/SSE(fit)

c(r2_3, r2_4, r2_5)
```