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 | 11111111111111111111111111111111111111111111111111111111111111111111+254
##   0 | 55555555555555555555555556666666666666777777777777777777778888888888
##   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 | 00001111111111112222222222222222222233333333333333333333333333444444+252
##    1 | 00000000000000011111111111111222222222223333334444555666677778889999
##    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 | 11111111111112222222222222222222222222222222222222222222222222222222+263
##    1 | 000000000000111111111222223333344444445555555567788888888999
##    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 | 00000000000000001111111111111111111111111111111111111111111111111111+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 $(X_{2,2})$

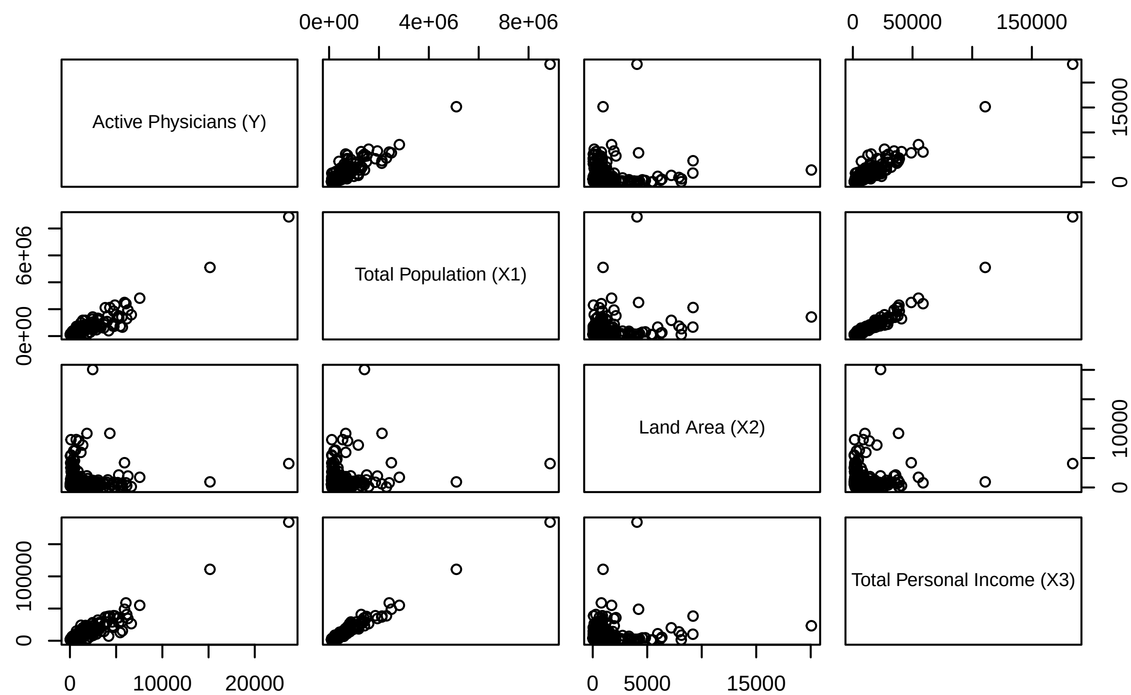
```
##
##   The decimal point is at the |
##
##    2 | 0
##    4 | 47890389
```

```
##     6 | 1123455677990134566678899
##     8 | 0011222223333444445556667777788888999900022222333333444444445555666677
##    10 | 000111111222222222223333334444445555555666666667777777888888888899999+36
##    12 | 000000001111122222333333333344444555555566666667777777778888999000000+36
##    14 | 00001111111223333444445556778890000001111222223455667778
##    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

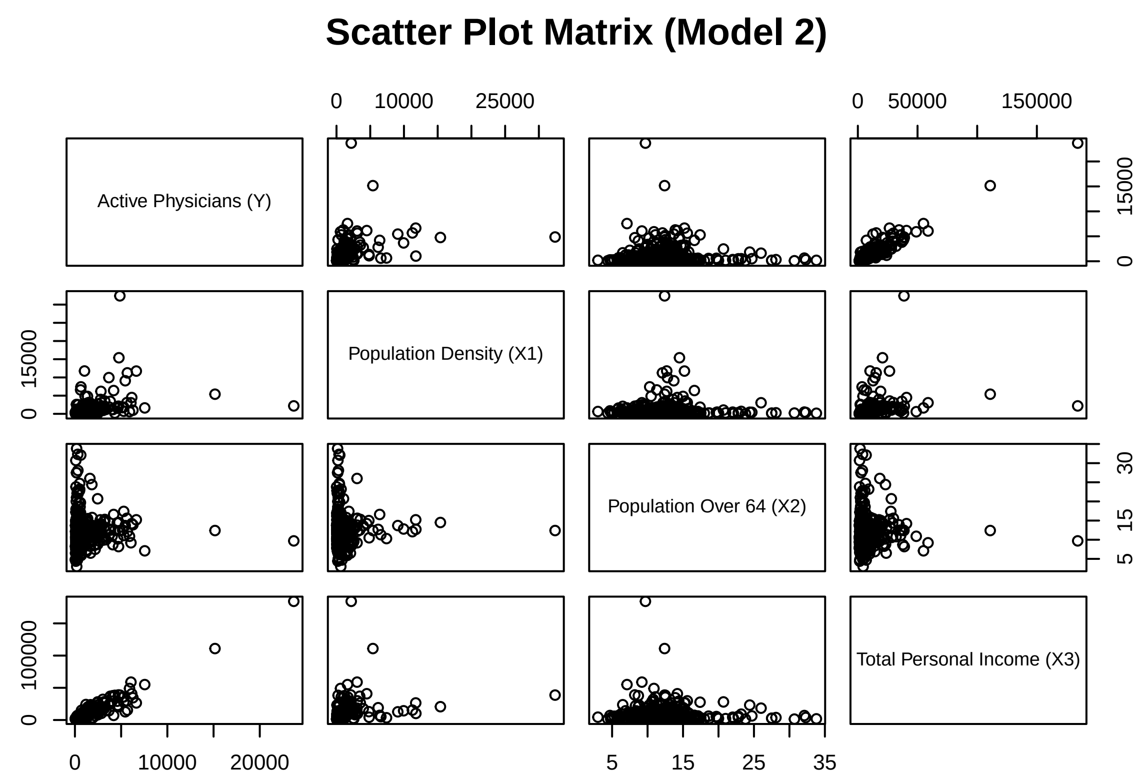
Scatter Plot Matrix (Model 1)




```
##                             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.




```
##                              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

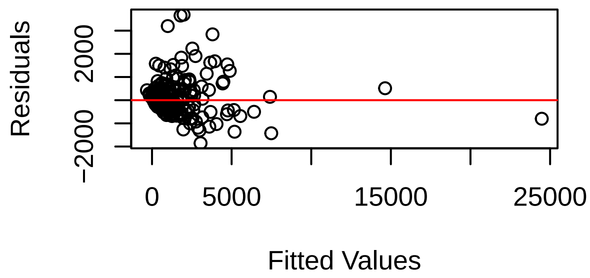
$$\begin{aligned} \text{Model 1 } R^2 : &\quad 0.9026432 \\ \text{Model 2 } R^2 : &\quad 0.9117491 \end{aligned}$$

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

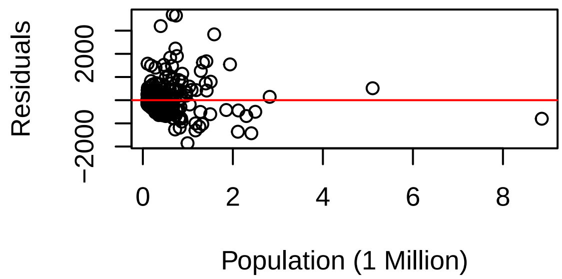
e

Model 1

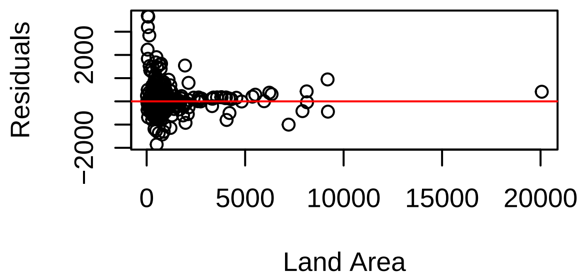
Model 1: Residuals ~ Fitted values



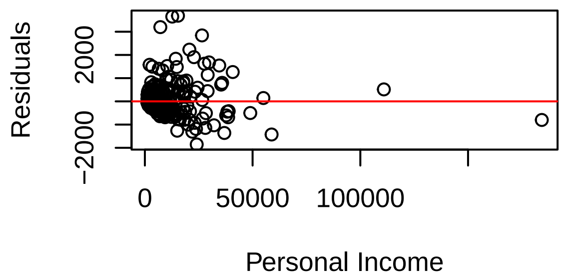
Model 1: Residuals ~ Population



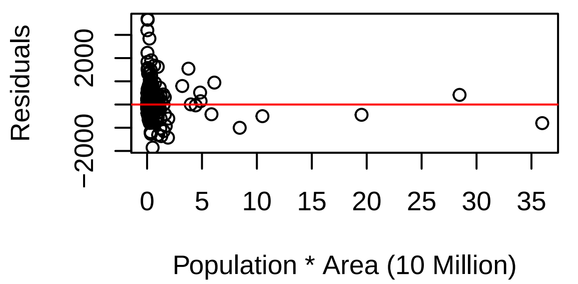
Model 1: Residuals ~ Area



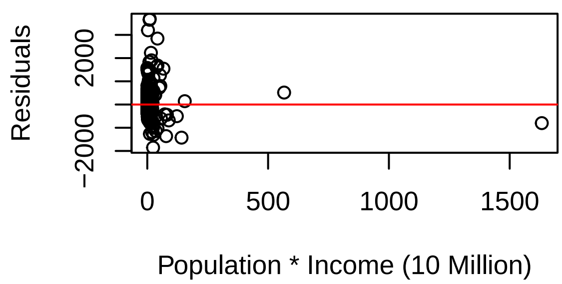
Model 1: Residuals ~ Income



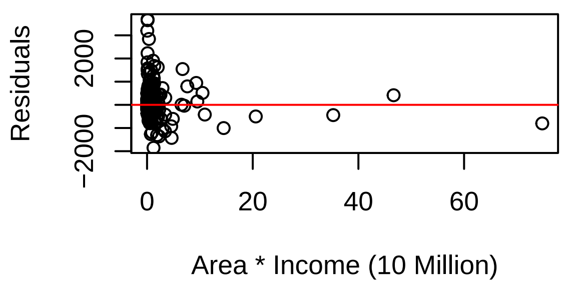
Model 1: Residuals ~ Population * Are



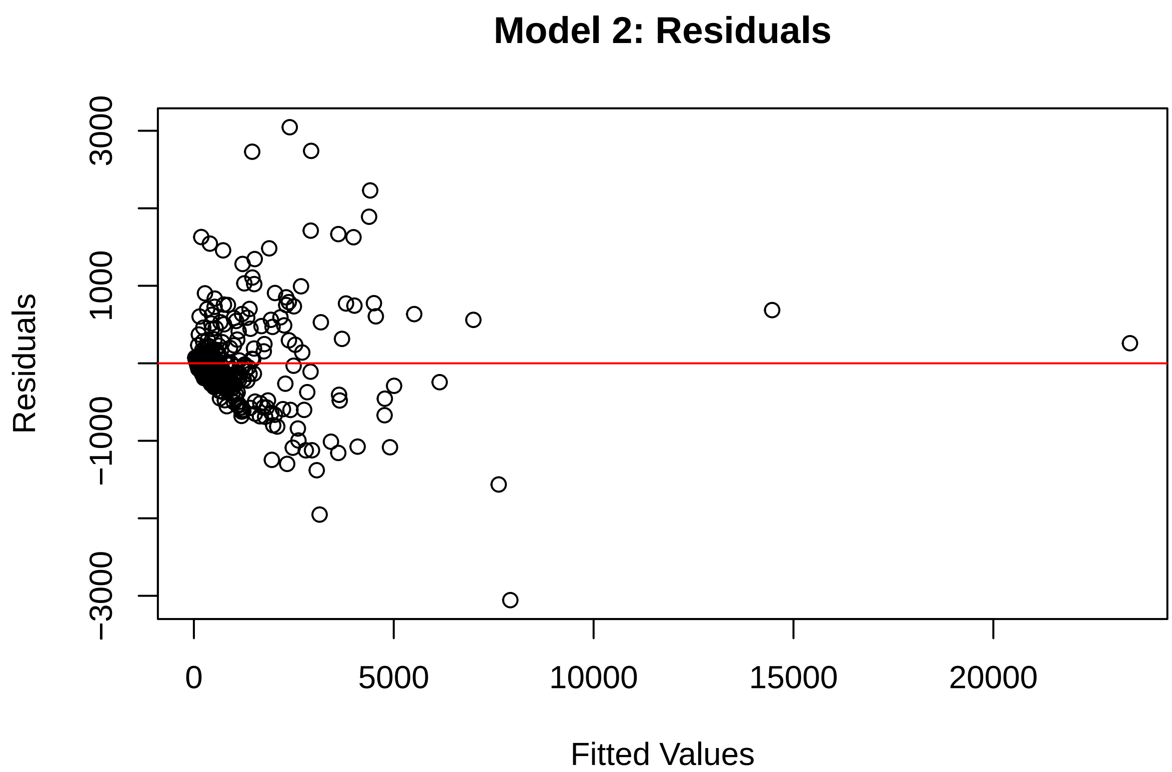
Model 1: Residuals ~ Population * Incor



Model 1: Residuals ~ Area * Income



Model 2



Model 2: Residuals
Residuals
Fitted Values
3000
1000
−1000
−3000
0
5000
10000
15000
20000

f

Part 2

Part 3

Appendix

```
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
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)", "To
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 = f1s$coefficients
betas2 = f2s$coefficients
model1.R2 = f1s$r.squared
model2.R2 = f2s$r.squared
# Get resids
m1.resid = f1s$residuals
m2.resid = f2s$residuals

m1.yhat = fit1$fitted.values
m2.yhat = fit2$fitted.values

par(mfrow = c(2,2))

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

```r
# 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 = "R
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 = "Residual
abline(0,0,col = "red")


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