CDI ANALYSIS
PART 2

INTRODUCTION:

With rising awareness of socioeconomic inequalities, it is imperative to examine possible causes or symptoms of inequity. This project examines aspects county hospitals and how they correlate with aspects of the socioeconomic status of their counties. The data that is used in this project is the County Demographic Information (CDI) for the 440 most populous counties in the United States. Each line of data is identified by county and state, and has information about 14 variables. In this project, we will compare whether or not each predictors have an effect on the model, how strong those effects are, and how they interact with each other. The main tool used in this project is R. R was used to calculate statistics and create the graphs that follow. In part 1, we will compare and contrast two models: one that includes total population, land area, and total personal income as predictors, and one that includes population density, percent of population greater than 64 years of age, and total personal income as predictors. Then, we will compare each possible two-factor interaction. In part 2, we will be assessing whether or not a third predictor should be added to the combination of total population and total personal income. The predictors tested will be percent of population 65 or older, and number of hospital beds. We will then test which pair of these predictors should be added to the model, given that total population and total personal income are already in the model. Part 3 is a discussion of the findings in this project.

PART 1:
6.28
Here in model 1 (the old model) X1 is Total Population, X2 represent number of Hospital beds, and X3 represents Total personal Income, and Y in model 1 stands for Number of active physicians.
 In Model 2, X1 represents population density, X2 is the Percent of population, 65 or older, and X3 represents Total personal income. And in model 2 Y represents Number of active physicians.

PART A: Stem and Leaf Plots
Model 1: X1
 The decimal point is 6 digit(s) to the right of the |

```
 0 | 11111111111111111111111111111111111111111111111111111111111111111+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
```
Overall X1 in model 1 has a right skewed distribution

Model 1: X2
The decimal point is 3 digit(s) to the right of the |

```
 0 | 11111111122222222222222222222222222222222222222333333333333333333+279
 2 | 00000011222234444455556899990011123333556677889
 4 | 000112356677892458
 6 | 1222498
 8 | 14895
10 | 5
12 | 4
14 |
16 |
18 |
20 | 6
22 |
24 |
26 | 7
```

Overall X2 in model 1 has a right skewed distribution since the right tail seems to be longer

Model 1: X3
The decimal point is 4 digit(s) to the right of the |

```
 0 | 11111111111112222222222222222222222222222222222222222222222222222+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
```
Overall X3 in model 1 has a right skewed distribution since the right tail seems to be longer

Model 2: X1

 The decimal point is 3 digit(s) to the right of the |
```
  0 | 00000000000000001111111111111111111111111111111111111111111111111+321
  2 | 00001112233456700111145
  4 | 05884
  6 | 2464
  8 | 19
 10 | 378
 12 |
 14 | 4
 16 |
 18 |
 20 |
 22 |
 24 |
 26 |
 28 |
 30 |
 32 | 4
```
Overall X1 in model 2 has a right skewed distribution since the right tail seems to be longer

Model 2: X2

 The decimal point is at the |
```
  2 | 0
  4 | 47890389
  6 | 1123455677990134566678899
  8 | 001122222333344445556667777788888999900022223333334444444455555666677
 10 | 00011111122222222222333333344444455555555666666667777777888888888899999+36
 12 | 00000000011111222233333333333444455555555666666677777777778888899900000000+36
 14 | 0000111111112233344444555677889000000011122223455667778
 16 | 12556699901122345
 18 | 06778
 20 | 070
 22 | 018828
 24 | 47
 26 | 055
 28 | 1
 30 | 7
 32 | 138
```
X2 in model 2 has relatively normal distribution

Model 2: X3

 The decimal point is 4 digit(s) to the right of the |

```
  0 | 1111111111111122222222222222222222222222222222222222222222222222222222+263
  1 | 0000000000000111111111122222333334444444555555567788888888999
  2 | 001111233344477788899
  3 | 0255678899
  4 | 19
  5 | 59
  6 |
```
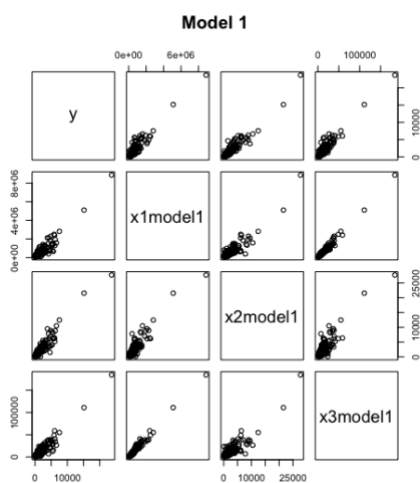
```
 7 |
 8 |
 9 |
10 |
11 | 1
12 |
13 |
14 |
15 |
16 |
17 |
18 | 4
```
Overall X3 in model 2 has a right skewed distribution since the right tail seems to be longer

PART B: Scatter plot and correlation Matrix

Model 1

Model 2

Correlation
Model 1

	Total population	Number of active physicians	Number of hospital beds	Total personal income
Total population	1.0000000	0.9402486	0.9237384	0.9867476
Number of active physicians	0.9402486	1.0000000	0.9504644	0.9481106
Number of hospital beds	0.9237384	0.9504644	1.0000000	0.9020615
Total personal income	0.9867476	0.9481106	0.9020615	1.0000000
```
> |
```
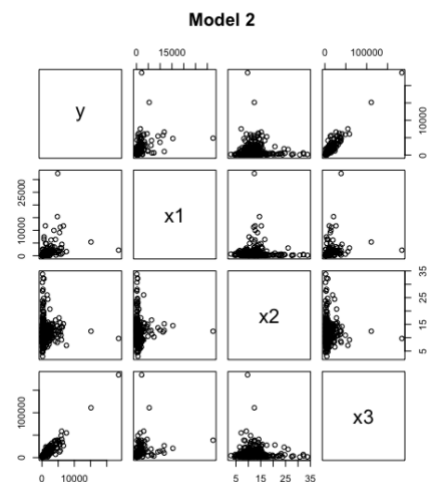
Model 2

	x1	x2	x3	y
x1	1.00000000	0.02918445	0.31620475	0.40643863
x2	0.02918445	1.00000000	-0.02273315	-0.00312863
x3	0.31620475	-0.02273315	1.00000000	0.94811057
y	0.40643863	-0.00312863	0.94811057	1.00000000

These matrices show that the predictors in model 1 generally are more highly correlated with each other than the predictors in model 2.

PART C: First Order Regression Models
Model 1;

$$\hat{Y} = -89.104738 + -0.001833X1 + 0.486604X2 + 0.138200X3$$

Model 2

$$\hat{Y} = -170.57422 + 0.09616X1\ 6.33984X2 + 0.12657X3$$
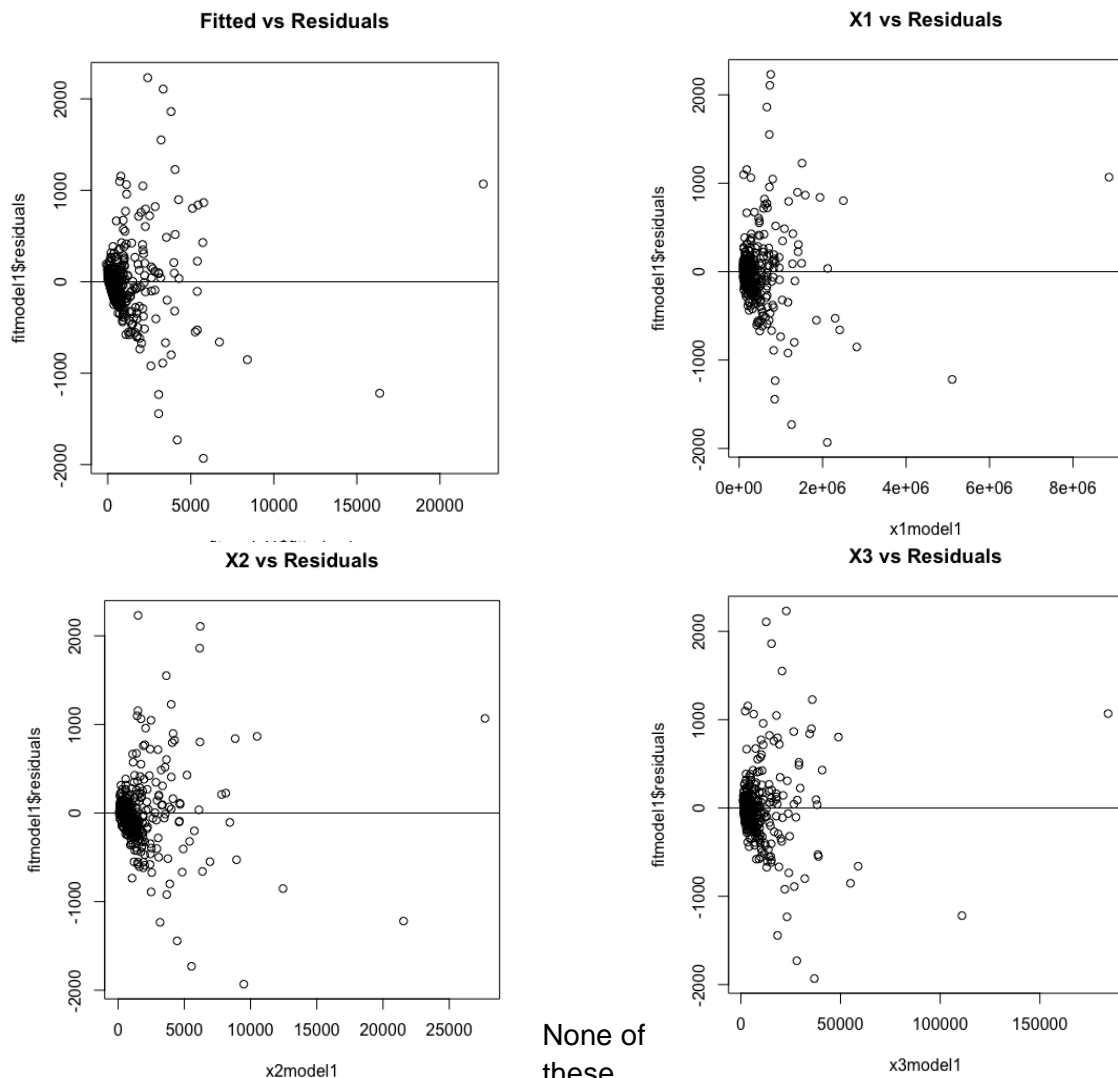
PART D: R^2

The R^2 for model 1 is 0.9553

The R^2 for model 2 is 0.9117

By looking at the R^2 values for each of the models it is clear that model 1 is slightly preferable since it has a higher R^2

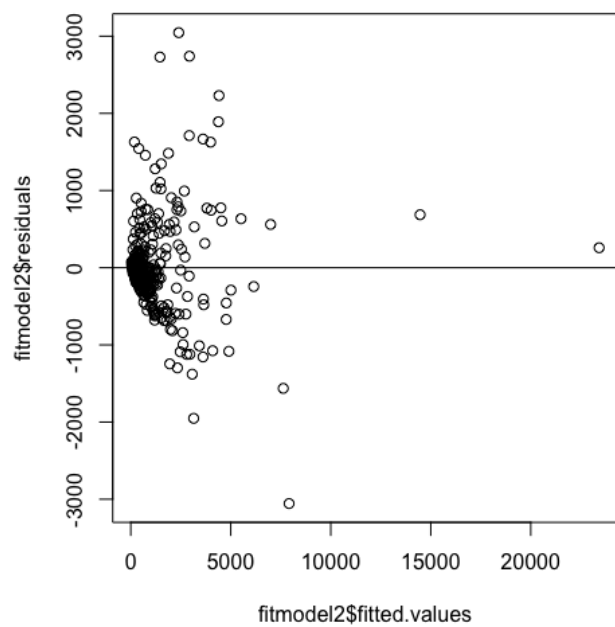PART E: Plots against Resuduals

MODEL 1:

None of these have a
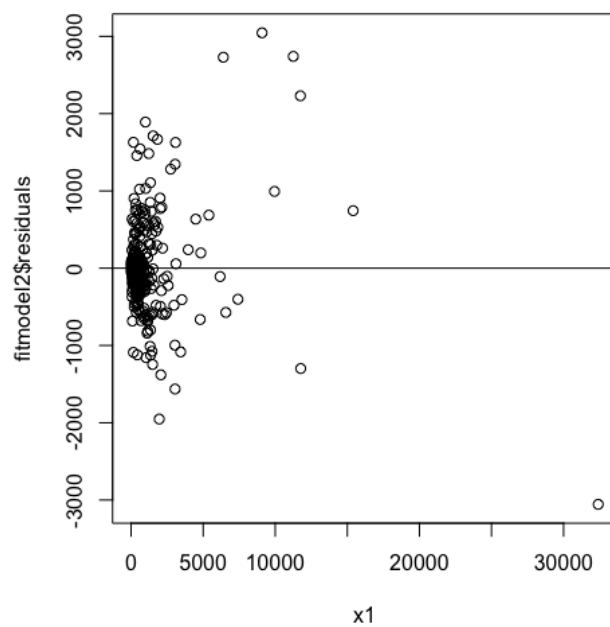
random pattern. There is some trend for x1, x2, and x3 when plotted against the residuals. Based on this, it would appear that x1, x2, and x3 should all be considered in the model.

MODEL 2:

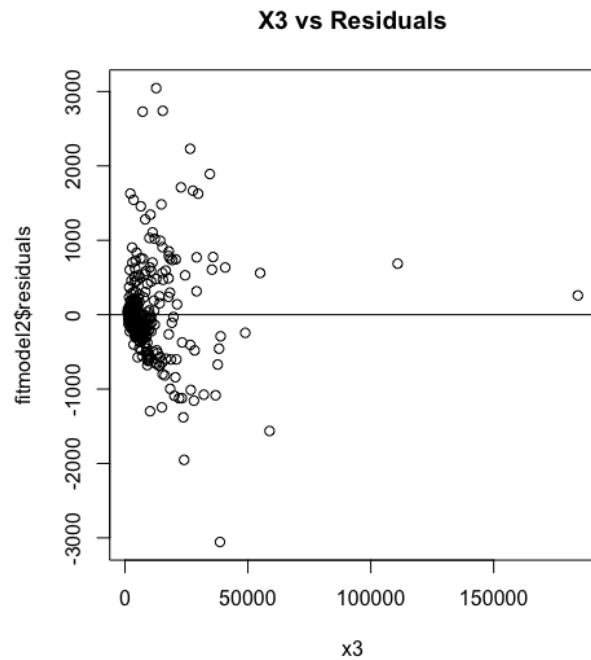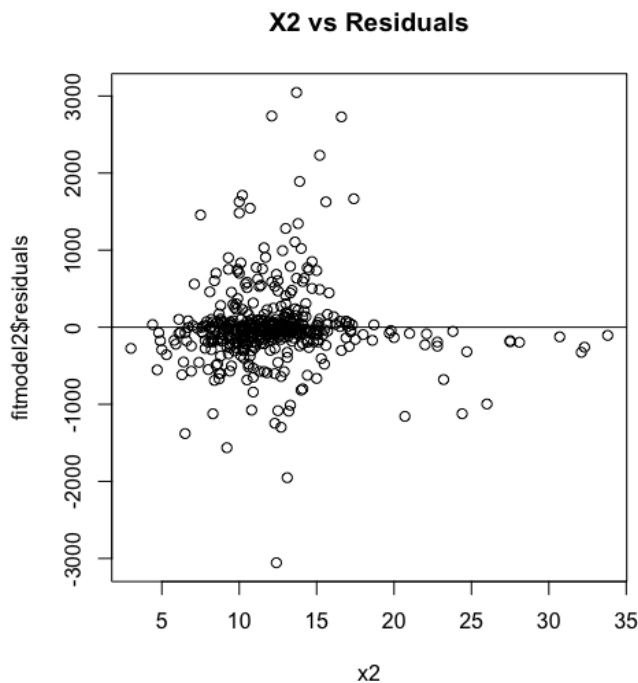**Fitted vs Residuals**

X1 vs Residuals

X2 vs Residuals

X3 vs Residuals

None of these plots appear to have a truly random pattern. There appears to be some sort of trend for x1, x2, and x3 when plotted against the residuals. The models for x1 and x3 against residuals are unbalanced on the x-axis, so this could mean that there is a missing variable that would improve the accuracy of the model. X2 is closest to a random plot. This would indicate that each of the three predictors should be considered in the model.

PART F: Interaction Effects

Model 1:
Regression equation for model 1 with interaction effects is
$\hat{Y}$ = -8.626e+01 +-2.048e-03*X1 + 5.863e-01*X2 +1.311e-01*X3 + -5.463e-08X2*X1*X2 + -1.130e-06*X2*X3 + 1.266e-08*X1*X3

Interaction between X1 and X2 vs Residuals
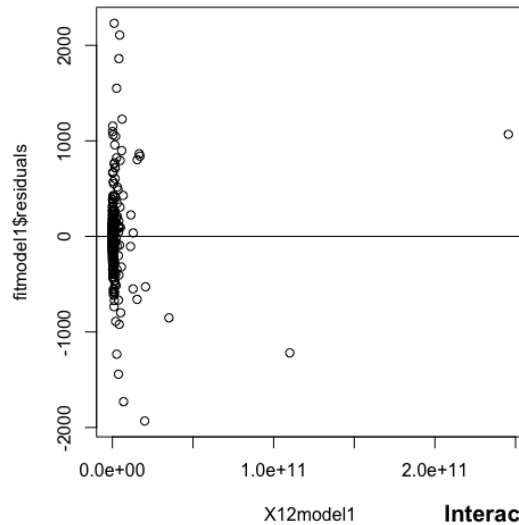
Interaction between X2 and X3 vs Residuals
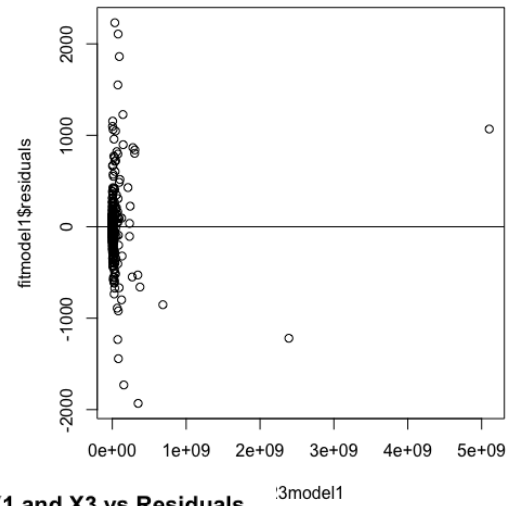
Interaction between X1 and X3 vs Residuals

There is no random pattern in any of these plots. This indicates that each of the two factor interactions appear to be meaningful. That is, the predictors do have some interaction with each other. The models are unbalanced on the x-axis, so this could mean that there is a missing variable that would improve the accuracy of the model.

Model 2: Interaction

$\hat{Y}$ = -9.367e+00 + -4.179e-01*X1 + -1.106e+01*X2 + 1.477e-01* X3 4.652e-02*X1*X2 + -1.289e-03*X2*X3 + -3.276e-06*X1*X3

Interaction between X1 and X2 vs Residuals
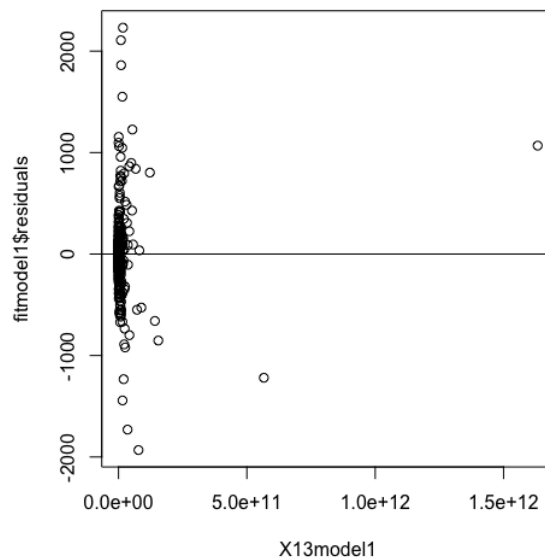

Interaction between X2 and X3 vs Residuals

Interaction between X1 and X3 vs Residuals
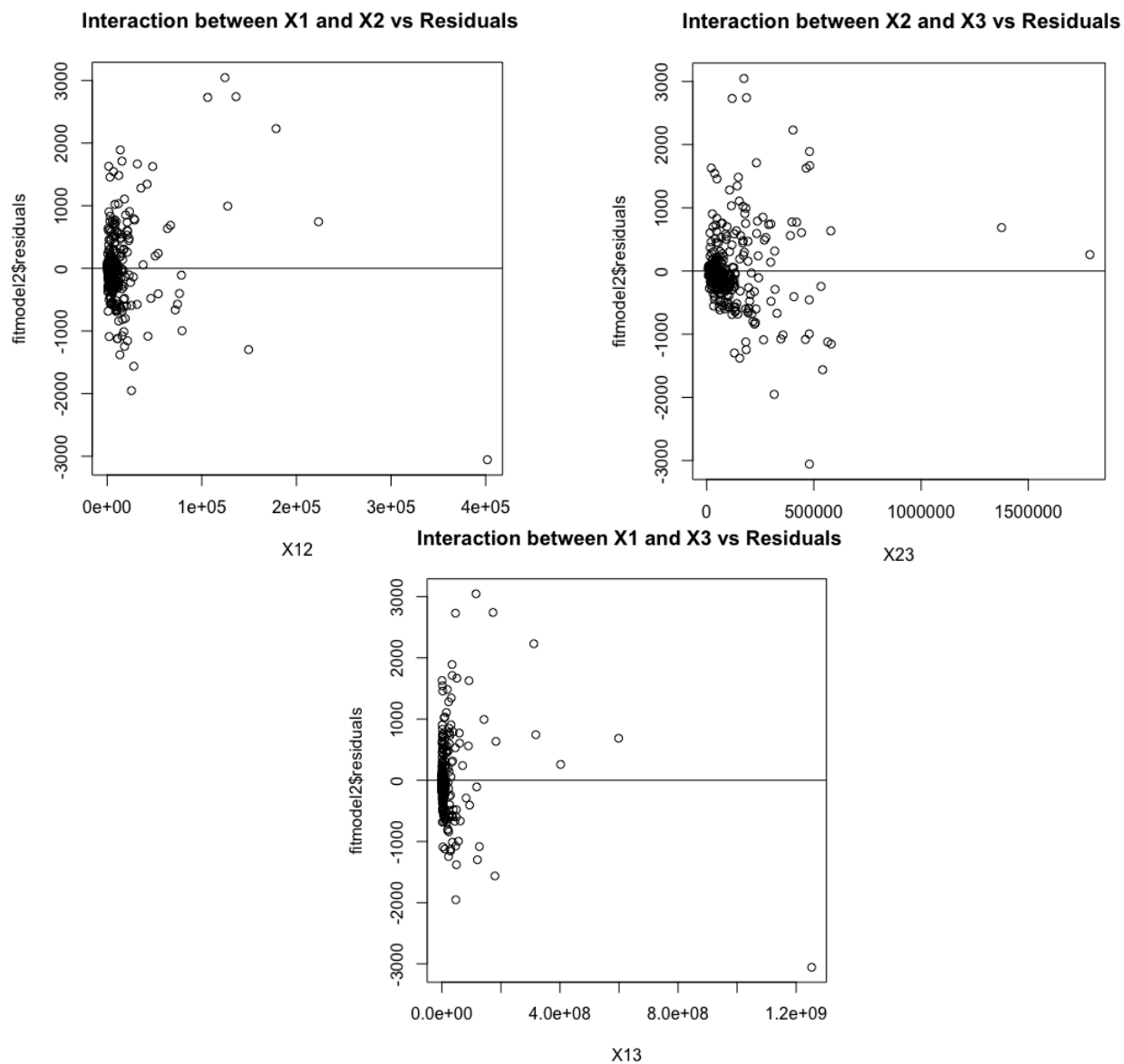
There is no random pattern in any of these plots, although the interaction between x2 and x3 is the closest to a random pattern. This indicates that each of the two factor interactions appear to be meaningful. That is, the predictors do have some interaction with each other. The models are unbalanced on the x-axis, so this could mean that there is a missing variable that would improve the accuracy of the model.

PART 2: 7.37

X1 refers to total population, X2 refers to total personal income, X3 is land area, X4 refers to the percent of population 65 or older, X5 is the number of hospital beds, and Y denotes the number of active physicians.

PART A: Partial Determinations

$R^2_{3|1,2}$ = 4063370/140967081 = 0.02882496

$R^2_{4|1,2}$ = 541647/ 140967081 = 0.003842365

$R^2_{5|1,2}$ = 78070132/140967081 = 0.5538182

These partial determinations indicate that adding x3 to the model results in a reduction in the variation of Y by 0.02882496, adding x4 results in a 0.003842365 reduction, and adding x5 results in a 0.5538182 reduction, given x1 and x2 are in the model.

PART B:

Out of the 3 additional predictor variable X5, the number of hospital beds, is the best is since it has the highest partial determinations given X1 and X2. The extra sum of square for X5 given X1 and X2 is larger compared to the other variables' extra sum of squares. The extra Sum of square for X5 was 78070132 while for X3 it was 4063370 and for X4 it was 541647.

PART C: F* - Test

Ho: $\beta_5 = 0$

Ha: $\beta_5 \neq 0$

F* = (SSE(R) - SSE(F))/(df(R) - df(F))/MSE(F) = 541.1801

Fcrit = F(1,436, 0.99) = 6.693358

F* > Fcrit -> Therefore Reject Ho

We conclude at a 1% significance that β_5 is not equal to 0.

The F* for the other variables would be just as big since the other variables also have large SSE(R) values

PART D: Additional Partial Determinations

$R^2_{3,5|1,2}$ = 0.5558232

Ho: $\beta_3 = \beta_5 = 0$

Ha: $\beta_5 \neq 0 \ or \ \beta_3 \neq 0$

F* = (SSE(R) - SSE(F))/(df(R) - df(F))/MSE(F) = 272.1699

Fcrit = F(2,435, 0.99) = 4.654269

F* > Fcrit

Reject Ho and at the 1% significance we conclude that both X3 and X5 being in the model is significant

$R^2_{4,5|1,2}$ = 0.5642756

Ho: $\beta_4 = \beta_5 = 0$

Ha: $\beta_5 \neq 0 \ and \ or \ \beta_4 \neq 0$

F* = (SSE(R) - SSE(F))/(df(R) - df(F))/MSE(F) = 281.6688
Fcrit = (2,435, 0.99)= 4.654269
F* > Fcrit
Reject Ho and at the 1% significance we conclude that both X4 and X5 being in the model is significant

$R^2_{3,4|1,2}$ = 0.03314181
Ho: $\beta_3 = \beta_4 = 0$
Ha: $\beta_3 \neq 0$ and or $\beta_4 \neq 0$
F* = (SSE(R) - SSE(F))/(df(R) - df(F))/MSE(F) = 7.45543
Fcrit = (2,435, 0.99)= 4.65426
F* > Fcrit
Reject Ho and at the 1% significance we conclude that both X3 and X4 being in the model is significant

Out of the three pairs added to the model, X4 and X5 seem to be the better pair. They're extra sum of squares was larger which led to a higher partial determination and a larger F* value.

PART3: Discussion

The course materials dealing with multicollinearity, multiple linear regression, polynomial regression, and regression variable selection were all helpful in this project. In part 1, we found that model 1 was preferable to model 2. This was due to the slightly higher correlation coefficient for model 1. We also found that all of the predictors interact with each other when included in the model. That is, the predictors are associated with one another. In part 2, we found that the predictor x5, or number of hospital beds, was the best predictor to add to a model already containing total population and total personal income. This means that a model with these three variables give the most accurate prediction of number of physicians. The other two variables considered (land area and percent of population 65 or older) also improve the accuracy of the model, but not as much as number of hospital beds. It would be better to have land area and hospital beds or percent of population 65 and older and hospital beds included in the model, versus having land area and population 65 and older included in the model.

Appendix:

Screenshots of Output
PART 1

The decimal point is 4 digit(s) to the right of the |

```
 0 | 1111111111111222222222222222222222222222222222222222222222222222222+263
 1 | 00000000000011111111112222233333344444445555555567788888888999
 2 | 001111233344477788899
 3 | 0255678899
 4 | 19
 5 | 59
 6 |
 7 |
 8 |
 9 |
10 |
11 | 1
12 |
13 |
14 |
15 |
16 |
17 |
18 | 4
```

```
model_2 = c(x1,x2,x3,y)
mod2 = matrix(model_2, nrow = length(y),byrow = FALSE)
colnames(mod2) = c('x1','x2','x3','y')
pairs(~y+x1+x2+x3, data = df , main = "Model 2")
cor(mod2)
          x1          x2          x3           y
1 1.00000000  0.02918445  0.31620475  0.40643863
2 0.02918445  1.00000000 -0.02273315 -0.00312863
3 0.31620475 -0.02273315  1.00000000  0.94811057
  0.40643863 -0.00312863  0.94811057  1.00000000
pairs(mydata1)
rror in pairs(mydata1) : object 'mydata1' not found
```

```
> plot(X13model1,fitmodel1$residuals, main = " Interaction between X1 and X3 vs Residuals")
> abline(h=0)
>
>
> #ineraction model part 1f
> fitmodel2_int =lm(y~x1+x2+x3+X12 +X23 + X13)
> fitmodel2_int #  -9.367e+00  + -4.179e-01*X1  + -1.106e+01*X2 + 1.477e-01* X3 4.652e-02*X1*X2 + -1.289e-03*X2*X3 + -3.276e
-06*X1*X3

Call:
lm(formula = y ~ x1 + x2 + x3 + X12 + X23 + X13)

Coefficients:
(Intercept)            x1           x2           x3          X12
 -9.367e+00    -4.179e-01   -1.106e+01    1.477e-01    4.652e-02
        X23           X13
 -1.289e-03    -3.276e-06

> summary(fitmodel2_int)

Call:
lm(formula = y ~ x1 + x2 + x3 + X12 + X23 + X13)

Residuals:
     Min       1Q   Median       3Q      Max
-2409.57  -163.91   -12.32   103.25  2721.84

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.367e+00  9.928e+01  -0.094    0.925
x1          -4.179e-01  1.055e-01  -3.960 8.76e-05 ***
x2          -1.106e+01  7.792e+00  -1.419    0.157
x3           1.477e-01  9.739e-03  15.168  < 2e-16 ***
X12          4.652e-02  7.925e-03   5.870 8.67e-09 ***
X23         -1.289e-03  8.743e-04  -1.474    0.141
X13         -3.276e-06  7.439e-07  -4.404 1.34e-05 ***
```

PART 2

```
> ###### PART 2 ########
> x1_pop = df$`Total population`
> x2_inc = df$ `Total personal income`
> x3_land = df$`Land area`
> x4 = df$`Percent of population, 65 or older`
> x5 = df$`Number of hospital beds`
>
> fitfull = lm(y~x1_pop+x2_inc+x3_land+x4+x5)
> fitx1_x2 = lm(y~x1_pop+x2_inc)
> fitx1_x2_x3 = lm(y~x1_pop+x2_inc+x3_land)
> fitx1_x2_x4 = lm(y~x1_pop+x2_inc+x4)
> fitx1_x2_x5 = lm(y~x1_pop+x2_inc+x5)
> #partial determinations
> anova(fitx1_x2)
Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq F value     Pr(>F)
x1_pop      1 1243181164 1243181164 3853.88 < 2.2e-16 ***
x2_inc      1   22058054   22058054   68.38 1.638e-15 ***
Residuals 437  140967081     322579
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(fitx1_x2_x3)
Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq  F value    Pr(>F)
x1_pop      1 1243181164 1243181164 3959.184 < 2.2e-16 ***
x2_inc      1   22058054   22058054   70.249 7.271e-16 ***
x3_land     1    4063370    4063370   12.941 0.0003583 ***
Residuals 436  136903711     313999
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> R2_12_3 = anova(fitx1_x2,fitx1_x2_x3)
```

```
> anova(fitx1_x2_x3)
Analysis of Variance Table

Response: y
            Df      Sum Sq    Mean Sq  F value    Pr(>F)
x1_pop       1 1243181164 1243181164 3959.184 < 2.2e-16 ***
x2_inc       1   22058054   22058054   70.249 7.271e-16 ***
x3_land      1    4063370    4063370   12.941 0.0003583 ***
Residuals  436  136903711     313999
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> R2_12_3 = anova(fitx1_x2,fitx1_x2_x3)
> partial_r123 =4063370/140967081 ; partial_r123
[1] 0.02882496
> R2_12_4 = anova(fitx1_x2, fitx1_x2_x4)
> partial_r124 = 541647/ 140967081 ; partial_r124
[1] 0.003842365
> R2_12_5 = anova(fitx1_x2, fitx1_x2_x5)
> partial_r125 = 78070132/140967081 ; partial_r125
[1] 0.5538182
> 140967081-136903711
[1] 4063370
>
>
>
> #part C
> mse.full = (summary(fitx1_x2_x5)$sigma)^2
> df.full =summary(fitx1_x2_x5)$df[2]
> mse.red= (summary(fitx1_x2)$sigma)^2
> df.red =summary(fitx1_x2)$df[2]
> sse.full = mse.full*df.full
> sse.red = mse.red*df.red
> f = (sse.red - sse.full)/(df.red - df.full)/mse.full ; f
[1] 541.1801
> alpha = 0.01
> df1 =df.red - df.full
```

R-Code
###STA 108 Project 2

```
df = read.table("CDI.txt")
names(df) = c("Identification number", "County", "State","Land area", "Total population","Percent of population aged
18-34", "Percent of population, 65 or older", "Number of active physicians",
        "Number of hospital beds","Total serious crimes" , "Percent high school graduates", "Percent bachelor's
degrees", "Percent below poverty level",
        "Percent unemployment", "Per capita income", "Total personal income", "Geographic region")
#### PART 1####
n = length(df$`Identification number` )
##active_physicans
y= df$`Number of active physicians`
#total_pop density
x1= df$`Total population`/df$`Land area`

#%greater then 64 yrs
x2 = df$`Percent of population, 65 or older`

##personal inc
x3 = df$`Total personal income`

###old model
##active_physicans
```

```r
y= df$`Number of active physicians`
#total_pop
x1model1= df$`Total population`
#hospital_beds
x2model1 = df$`Number of hospital beds`
#total_income
x3model1= df$`Total personal income`

####stem and leaf
stem(x1)
stem(x2)
stem(x3) #new model

stem(x1model1 )
stem(x2model1)
stem(x3model1)
model_2 = c(x1,x2,x3,y)
mod2 = matrix(model_2, nrow = length(y),byrow = FALSE)
colnames(mod2) = c('x1','x2','x3','y')
pairs(~y+x1+x2+x3, data = df , main = "Model 2")
cor(mod2)
pairs(mydata1)

pairs(~y+x1model1+x2model1+x3model1, data = df, main = " Model 1 ")
cor(~x1model1+x2model1+x3model1, data = df)
mydata1 = df[,c(5,8,9,16)]
cor(mydata1)

##1c
fitmodel2 =lm(y~x1+x2+x3)
fitmodel2
summary(fitmodel2)
coef2 = fitmodel2$coefficients
b0hat_m2 = coef2[1]
b1hat_m2 = coef2[2]
b1hat_m2 = coef2[3]
b1hat_m2 = coef2[4] #yhat = -170.57422 + 0.09616x1 6.33984x2+0.12657x3
#r2 = 0.9117

##1c
fitmodel1 =lm(y~x1model1+x2model1+x3model1)
fitmodel1
summary(fitmodel1)
coef1 = fitmodel1$coefficients
b0hat_m1 = coef1[1]
b1hat_m1 = coef1[2]
b1hat_m1 = coef1[3]
b1hat_m1 = coef1[4] # yhat = -89.104738 + -0.001833x1 + 0.486604x2  + 0.138200x3
#r2 model 1 =  0.9553
anova(fitmodel2)

##1e
plot(fitmodel1$fitted.values,fitmodel1$residuals)
abline(h=0)
```

```r
plot(fitmodel2$fitted.values,fitmodel2$residuals)
abline(h=0)


#interaction variables
X12 = as.numeric(x1)*as.numeric(x2)
X23 = as.numeric(x2)*as.numeric(x3)
X13 = as.numeric(x1)*as.numeric(x3)
#model 2
plot(fitmodel2$fitted.values,fitmodel2$residuals, main = "Fitted vs Residuals")
abline(h=0)
plot(x1,fitmodel2$residuals, main = "X1 vs Residuals")
abline(h=0)
plot(x2,fitmodel2$residuals, main = "X2 vs Residuals")
abline(h=0)
plot(x3,fitmodel2$residuals, main = "X3 vs Residuals")
abline(h=0)
plot(X12,fitmodel2$residuals, main = " Interaction between X1 and X2 vs Residuals") #interactions
abline(h=0)
plot(X23,fitmodel2$residuals, main = " Interaction between X2 and X3 vs Residuals")
abline(h=0)
plot(X13,fitmodel2$residuals, main = " Interaction between X1 and X3 vs Residuals")
abline(h=0)
X12model1 = as.numeric(x1model1)*as.numeric(x2model1)
X23model1 = as.numeric(x2model1)*as.numeric(x3model1)
X13model1 = as.numeric(x1model1)*as.numeric(x3model1)
#model 1
plot(fitmodel1$fitted.values,fitmodel1$residuals, main = "Fitted vs Residuals")
abline(h=0)
plot(x1model1,fitmodel1$residuals,  main = "X1 vs Residuals")
abline(h=0)
plot(x2model1,fitmodel1$residuals, main = "X2 vs Residuals")
abline(h=0)
plot(x3model1,fitmodel1$residuals, main = "X3 vs Residuals")
abline(h=0)
plot(X12model1,fitmodel1$residuals, main = " Interaction between X1 and X2 vs Residuals") #interactions
abline(h=0)
plot(X23model1,fitmodel1$residuals, main = " Interaction between X2 and X3 vs Residuals")
abline(h=0)
plot(X13model1,fitmodel1$residuals, main = " Interaction between X1 and X3 vs Residuals")
abline(h=0)
#ineraction model part 1f
fitmodel2_int =lm(y~x1+x2+x3+X12 +X23 + X13)
fitmodel2_int #  -9.367e+00  + -4.179e-01*X1  + -1.106e+01*X2 + 1.477e-01* X3 4.652e-02*X1*X2 + -1.289e-
03*X2*X3 + -3.276e-06*X1*X3
summary(fitmodel2_int)
fitmodel1_int =lm(y~x1model1+x2model1+x3model1+X12model1 +X23model1+ X13model1)
fitmodel1_int # yhat =   -8.626e+01 +-2.048e-03*X1 + 5.863e-01*X2 +1.311e-01*X3 + -5.463e-08X2*X1*X2 + -
1.130e-06*X2*X3 +  1.266e-08*X1*X3
summary(fitmodel1_int)


###### PART 2 ########
x1_pop = df$`Total population`
x2_inc = df$ `Total personal income`
x3_land = df$`Land area`
```

```r
x4 = df$`Percent of population, 65 or older`
x5 = df$`Number of hospital beds`
fitfull = lm(y~x1_pop+x2_inc+x3_land+x4+x5)
fitx1_x2 = lm(y~x1_pop+x2_inc)
fitx1_x2_x3 = lm(y~x1_pop+x2_inc+x3_land)
fitx1_x2_x4 = lm(y~x1_pop+x2_inc+x4)
fitx1_x2_x5 = lm(y~x1_pop+x2_inc+x5)
#partial determinations
anova(fitx1_x2)
anova(fitx1_x2_x3)
R2_12_3 = anova(fitx1_x2,fitx1_x2_x3)
partial_r123 =4063370/140967081 ; partial_r123
R2_12_4 = anova(fitx1_x2, fitx1_x2_x4)
partial_r124 = 541647/ 140967081 ; partial_r124
R2_12_5 = anova(fitx1_x2, fitx1_x2_x5)
partial_r125 = 78070132/140967081 ; partial_r125
140967081-136903711
#part C
mse.full = (summary(fitx1_x2_x5)$sigma)^2
df.full =summary(fitx1_x2_x5)$df[2]
mse.red= (summary(fitx1_x2)$sigma)^2
df.red =summary(fitx1_x2)$df[2]
sse.full = mse.full*df.full
sse.red = mse.red*df.red
f = (sse.red - sse.full)/(df.red - df.full)/mse.full ; f
alpha = 0.01
df1 =df.red - df.full
df2 = df.full
fcrit = qf(1-alpha,df1, df2  )
f > fcrit # Reject H0

#part D
fitx1_x2_x3_x5 = lm(y~x1_pop+x2_inc+x3_land+x5)
fitx1_x2_x4_x5 = lm(y~x1_pop+x2_inc+x4+x5)
aovx12.35 = anova( fitx1_x2,fitx1_x2_x3_x5)
r.partial12.35 = 78352775/140967081 ; r.partial12.35
mse.fulld1 = (summary(fitx1_x2_x3_x5)$sigma)^2
df.fulld1 =summary(fitx1_x2_x3_x5)$df[2]
mse.red= (summary(fitx1_x2)$sigma)^2
df.red =summary(fitx1_x2)$df[2]
sse.fulld1 = mse.fulld1*df.fulld1
sse.red = mse.red*df.red
fd1 = (sse.red - sse.fulld1)/(df.red - df.fulld1)/mse.fulld1 ; fd1
df1r35 =df.red - df.fulld1
df2r35 = df.fulld1
fcrit35 = qf(1-alpha,df1r35, df2r35 )
aovx12.45 = anova( fitx1_x2, fitx1_x2_x4_x5)
mse.fulld2 = (summary(fitx1_x2_x4_x5)$sigma)^2
df.fulld2 =summary( fitx1_x2_x4_x5)$df[2]
mse.red= (summary(fitx1_x2)$sigma)^2
df.red =summary(fitx1_x2)$df[2]
sse.fulld2 = mse.fulld2*df.fulld2
sse.red = mse.red*df.red
fd2 = (sse.red - sse.fulld2)/(df.red - df.fulld2)/mse.fulld2 ; fd2
```

```
df1r45 =df.red - df.fulld2
df2r45 = df.fulld2
fcrit45 = qf(1-alpha,df1r45, df2r45 )
r.partial12.45 = 79544288 /140967081;r.partial12.45
aovx12.34 = anova( fitx1_x2, fitx1_x2_x3_x4)
mse.fulld34 = (summary(fitx1_x2_x3_x4)$sigma)^2
sse.fulld34 = mse.fulld34*df.fulld34
df.fulld34 =summary( fitx1_x2_x3_x4)$df[2]
df1r34 =df.red - df.fulld34
df2r34 = df.fulld34
fcrit34 = qf(1-alpha,df1r34, df2r34 )
fd34 = (sse.red - sse.fulld34)/(df.red - df.fulld34)/mse.fulld34 ; fd34
r.partial12.34 = 4671904/140967081; r.partial12.34
```