STA 108 Project II

County Demographic Information

A Report Presented to

Professor Jiming Jiang of the Department of Statistics

University of California, Davis

In Partial Fulfillment

of the Requirements for the

Completion of STA 108

---

by
Ian Xu
Jeffrey Ugochukwu
Yiheng Lu

March 9th, 2020

Introduction

In this project we use a data set containing county demographic information for 440 of the most populated counties in the United States. Generally the data pertains to the years 1990-1992. Some counties with incomplete data were deleted from the data set.

We are interested in the relationship between the number of active physicians and total population, hospital beds, and total personal income respectively. We also want to discuss the relationship between per capita income and the percentage of individuals in a county having at least a bachelor degree.

We will be analysing the Fitting regression model to describe the relationship between predictors and response. ANOVA will be used as our approach to determine the relationship.

I. Multiple linear regression I

Part a. Stem and leaf plots

Model I:

Total Population:

```
 0 | 1111111111111111111111111111111111111111111111111111111111111111111111111111+254
 0 | 555555555555555555555556666666666666677777777777777777777778888888888
 1 | 000000122233333444
 1 | 55699
 2 | 1134
 2 | 58
 3 |
 3 |
 4 |
 4 |
 5 | 1
 5 |
 6 |
 6 |
 7 |
 7 |
 8 |
 8 | 9
```

Land Area:

```
 0 | 000011111111111122222222222222222222333333333333333333333333333444444+252
 1 | 000000000000000111111111111112222222222233333344445556666777778889999
 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:

```
 0 | 11111111111112222222222222222222222222222222222222222222222222222222222+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
```

Model II:

Population Density:

```
 0 | 00000000000000001111111111111111111111111111111111111111111111111111111+321
 2 | 00001112233456700111145
 4 | 05884
 6 | 2464
 8 | 19
10 | 378
12 |
14 | 4
16 |
18 |
20 |
22 |
24 |
26 |
28 |
30 |
32 | 4
```

Percent of Population 65 or Older:

```
 2 | 0
 4 | 47890389
 6 | 1123455677990134566678899
 8 | 00112222233334444555666777778888899990002222333333444444445555666677
10 | 000111111222222222233333344444455555556666666677777778888888888899999+36
12 | 000000001111122223333333333444455555556666667777777778888999000000000+36
14 | 00001111111223334444455567788900000011112222345566778
16 | 12556699901122345
18 | 06778
20 | 070
22 | 018828
24 | 47
26 | 055
28 | 1
30 | 7
32 | 138
```

Total Personal Income:

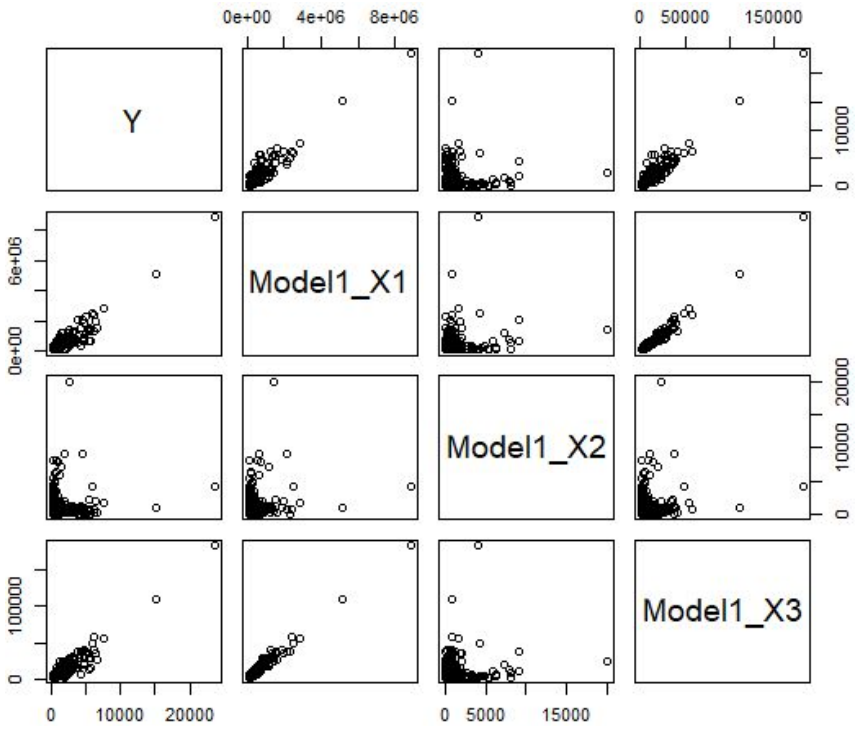
```
 0 | 1111111111111222222222222222222222222222222222222222222222222222222222+263
 1 | 000000000001111111112222233333444444455555555567788888888999
 2 | 001111233344477788899
 3 | 0255678899
 4 | 19
 5 | 59
 6 |
 7 |
 8 |
 9 |
10 |
11 | 1
12 |
13 |
14 |
15 |
16 |
17 |
18 | 4
```

Analysis: After analyzing our stem and leaf plots, we see that the data is skewed towards including smaller numbers compared to larger numbers. This would make the majority of the predictor variables for each model would be skewed towards the right since the unimodal peaks tend to be in the lower values of the columns for the stem and leaf plots and the rest of the values trail off towards the right. The

only exception would be the Model II's second predictor variable (X2: Percent of Population 65 or Older) in which the distribution is closer to being normally distributed (albeit, still slightly skewed to the right)because there's a bimodal peak of values in the middle range of values in the stem and leaf plot (10-12).

Part b.

Model I's Scatterplot Matrix



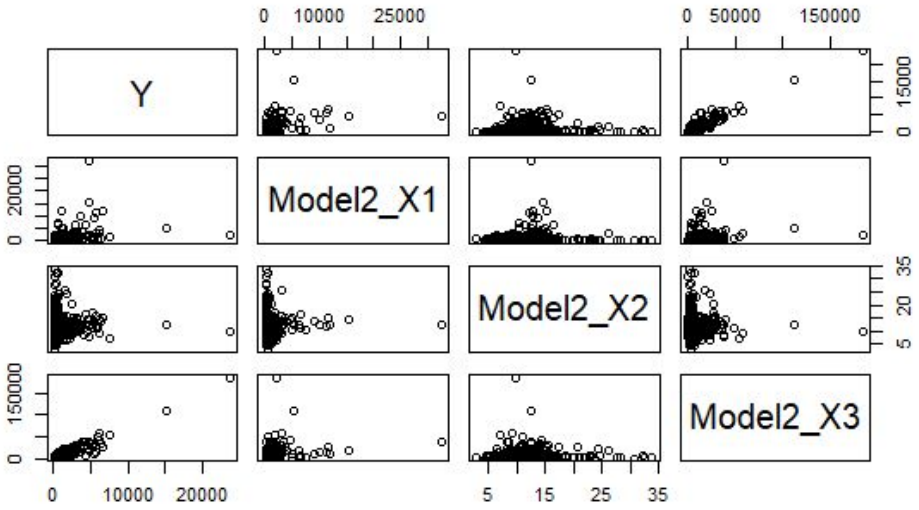
Model I's Correlation Matrix

	Y	Model1_X1	Model1_X2	Model1_X3
Y	1.00000000	0.9402486	0.07807466	0.9481106
Model1_X1	0.94024859	1.0000000	0.17308335	0.9867476
Model1_X2	0.07807466	0.1730834	1.00000000	0.1270743
Model1_X3	0.94811057	0.9867476	0.12707426	1.0000000

Analysis: After analyzing the scatter plots and correlation matrix for Model 1, we see that some variables are strongly correlated with each other. The observed variable, Number of Active Physicians, is strongly correlated with X1, Total population, and X3, Total Personal Income. The correlation coefficient for Y and X1 and X3 is about 0.94

and 0.948 respectively. X1 and X3 are also strongly correlated with each other, with a correlation coefficient of about 0.987, which indicates the possible multicollinearity. Variables that are strongly correlated can also be seen in the scatterplots because they form a straight line when plotted against each other.

Model II's Scatterplot Matrix



Model II's Correlation Matrix

	Y	Model2_X1	Model2_X2	Model2_X3
Y	1.00000000	0.40643863	-0.00312863	0.94811057
Model2_X1	0.40643863	1.00000000	0.02918445	0.31620475
Model2_X2	-0.00312863	0.02918445	1.00000000	-0.02273315
Model2_X3	0.94811057	0.31620475	-0.02273315	1.00000000

Analysis: After analyzing the scatter plots and correlation matrix for Model 2, we see that only the observed variable, Number of Active Physicians, and X3, Total Personal Income are strongly correlated (the correlation coefficient is about 0.948). Some variables that have low to moderate correlation are the observed variable and X1, Population Density (total population/land area), and X1 and X3. The correlation coefficients are about 0.406 and 0.3162 respectively. The other variables have little correlation with each other.

Part c.

Model I: $\widehat{Y} = -13.32 + 0.0008366X_1 - 0.06552X_2 + 0.09413X_3$

Model II: $\widehat{Y} = -170.57422325 + 0.09615889X_1 + 6.33984064X_2 + 0.12656649X_3$

Part d.

Model I's R-squared: 0.9026432

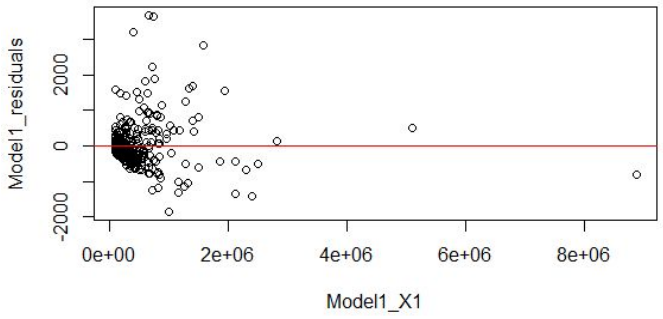
Model II's R-squared: 0.9117

Analysis: Based on the values of R squared for each model, Model II would be the preferable model since its R-squared is the closest to 1. This means the predictor variables for Model II account for more variation in our observed variable compared to the predictor variables for Model I.

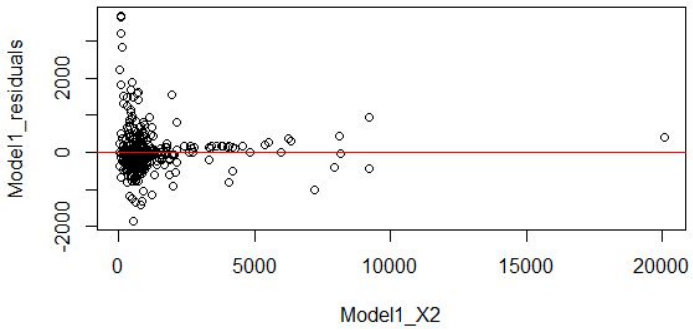
Part e.

Model I Residual Plots:

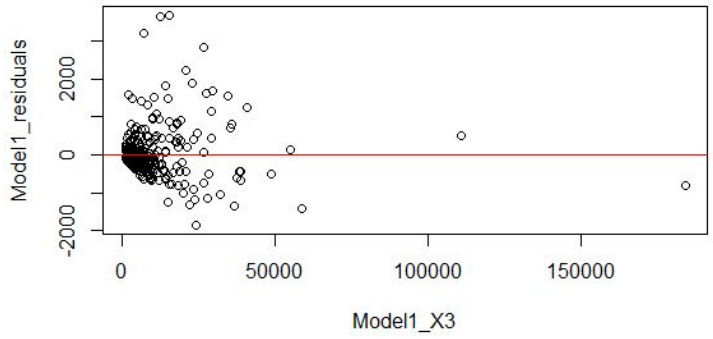
Total Population (X1)



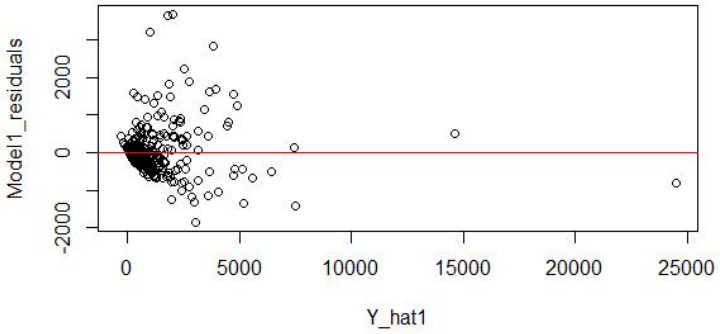
Land Area (X2)



Total Personal Income (X3)

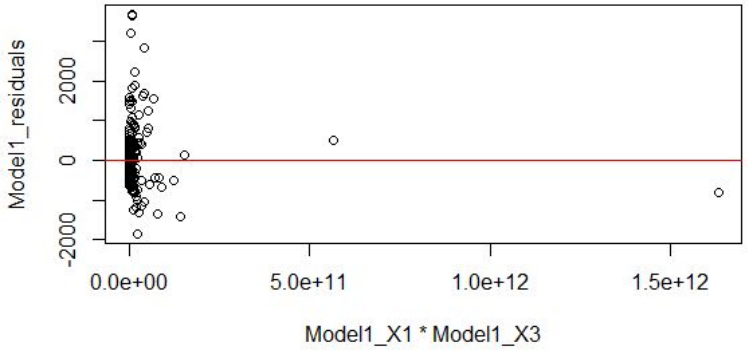


Model I's fitted values $(\widehat{Y})$

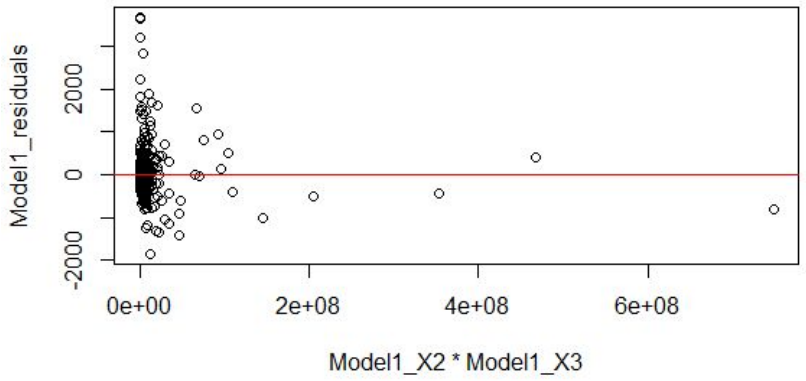


Total Population(X1)*Land Area(X2)

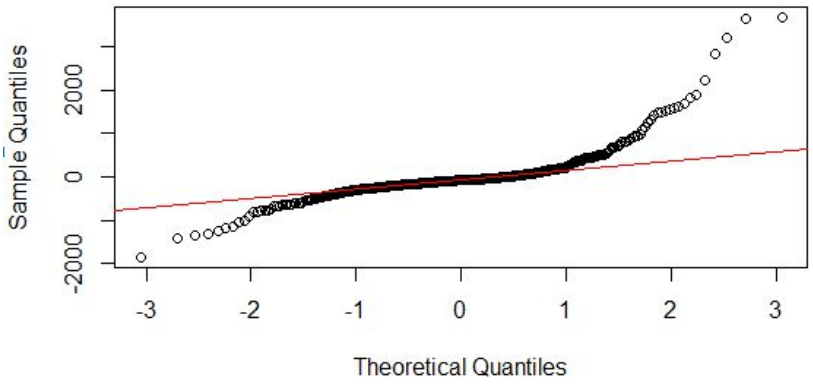
Total Population(X1)*Total Personal Income(X3)



Land Area(X2)*Total Income(X3)

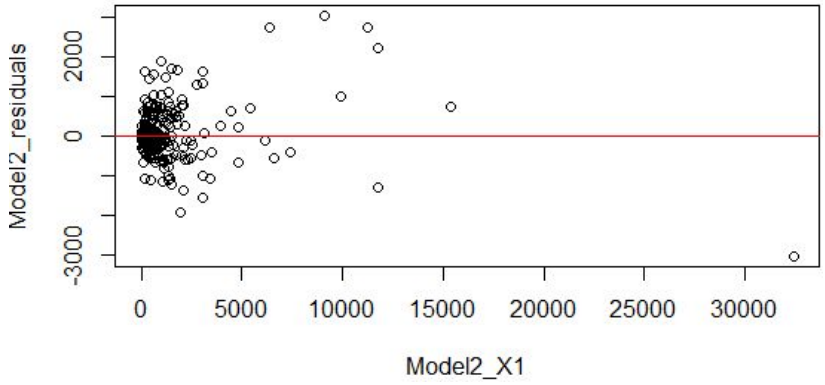


Normal Q-Q Plot

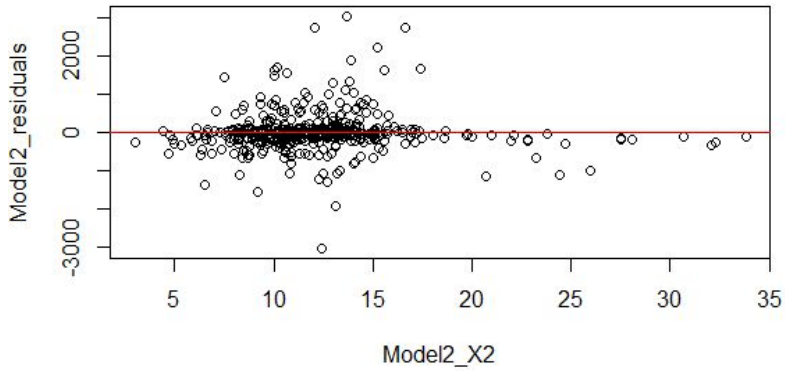


Model II Residual Plots:

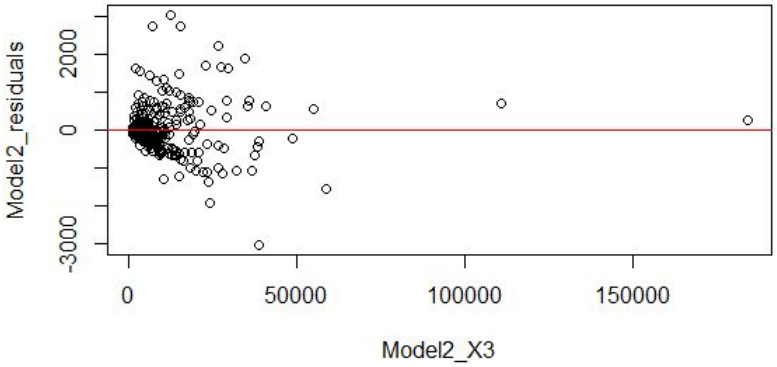
Population Density(X1)



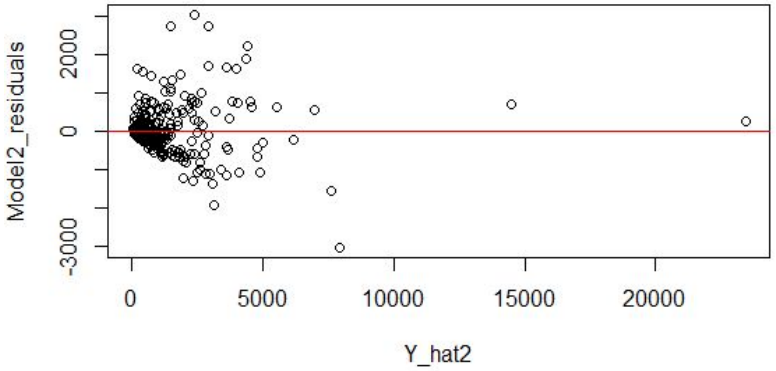
Percent of Population that are older than 64 years old(X2)



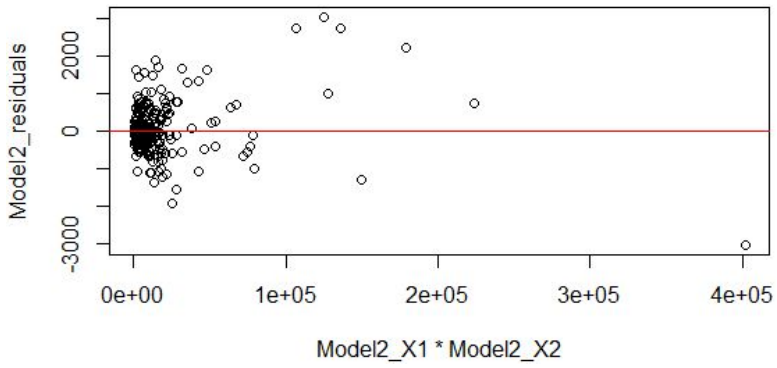
Total Personal Income(X3)



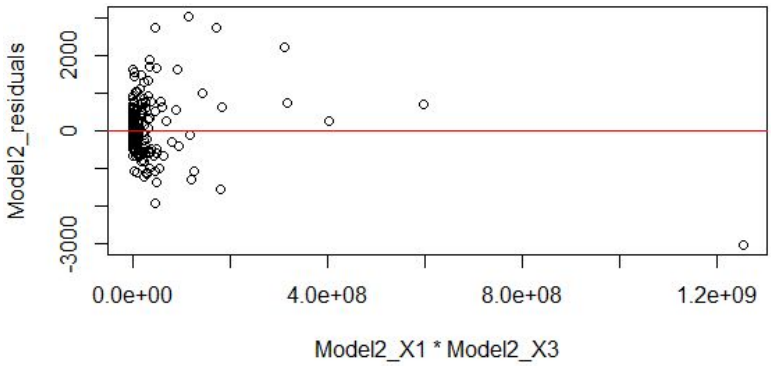
Model II's fitted values $(\widehat{Y})$



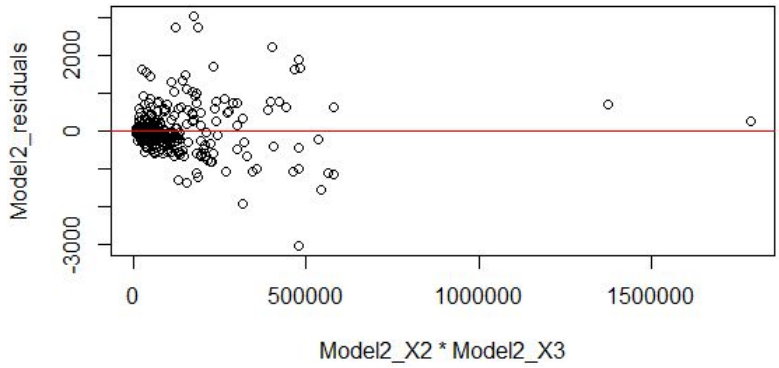
Population Density(X1) * Percent of population that's older than 64 years old(X2)



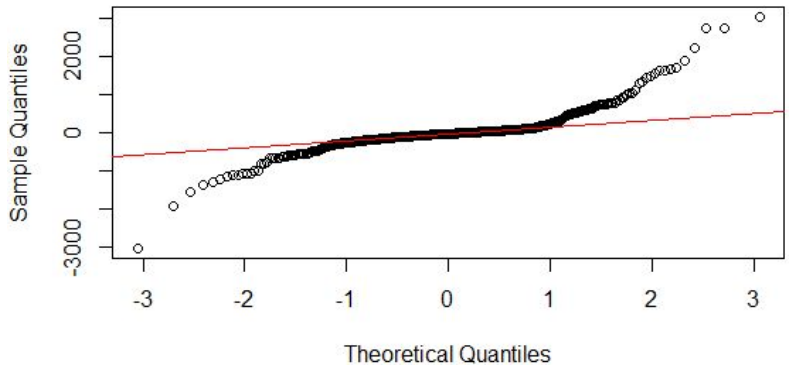
Population Density(X1) * Total Personal Income(X2)



Percent of Population 65 or Older(X2) * Total Personal Income(X3)



Normal Q-Q Plot



Analysis: When comparing Model 1 from Model 2, in terms of their residual plots for their predictor variables, most of their residuals' data points are more scattered towards the lower end of the x-axis and are fairly distributed against the horizontal line of zero; however, Model 2's X2's residuals are very evenly distributed throughout the x-axis whereas Model 1's X2's residuals still mainly towards the lower end. This would give model 2 a greater advantage in terms of being more appropriate. In terms of the residual plot for each Model's Y hat, they're both evenly distributed against the horizontal line of zero and are scattered towards the lower end of

the x-axis, so they both seem to share a level of being appropriate in that regard. In terms of the residuals in each model's interaction plots, Model 1's interaction plots tend to make a pattern of a vertical line on the lower end of the x-axis whereas Model 2's interaction plots are evenly scattered throughout the entire residual plot, so Model 2 would be the most appropriate out of the 2 in that regard. In terms of the normal probability plots for each model, both models are very close to the line of best fit, so there really isn't much of a difference regarding linear relationship.Overall, Model 2 is the most appropriate mode since it's interaction plots are the most evenly scattered for residuals and their 2nd predictor variable's residuals are evenly scattered across the x-axis.

Part f.

Model I Combinations for Interactions (R Squared Value):
0.9058047

Model II Combinations for Interactions (R Squared Value):
0.9195752

Analysis: Based off of the values of R squared for each interaction combination per model, Model 2 would still be the preferable model since it's R squared is the closest to 1, meaning that most of its values are closer to the regression line.

Part II. Multiple linear regression II

Part a.

$R^2_{Y\ 3|12}=0.02882495$

$R^2_{Y\ 4|12}=0.003842367$

$R^2_{Y\ 5|12}=0.5538182$

$R^2_{Y\ 5|12} > R^2_{Y\ 3|12} > R^2_{Y\ 4|12}$

$SSR_{(3|12)}=4063370$

$SSR_{(4|12)}$=541647.3
$SSR_{(5|12)}$=78070132

$SSR_{(5|12)} > SSR_{(3|12)} > SSR_{(4|12)}$

Part b.

X5(Number of hospital beds) is the best since $R^2_{Y\ 5|12}$ is the largest. Also, $SSR_{(5|12)}$ is larger than the extra sum of squares associated with the other two variables.

Part c.

H_0:β_5=0 H_a:$\beta_5 \neq 0$

Decision rule: Reject H_0 if $F^* > F_{(1-\alpha,\ ,1,n-1)}$

F^*=541.18 $F_{(1-\alpha,\ ,1,n-1)}=F_{(0.99,1,436)}$=6.69336 $F^*>F_{(0.99,1,436)}$

Conclusion:Reject H_0, X5(Number of hospital beds) is helpful in the regression model when X1 and X2 are included in the model

The F* values for X3 and X4 might not be as large as the one here for X5, since their partial R^2 values and extra sum of squares are not as high

Part d.

$R^2_{Y\ 34|12}$=0.03314181
$R^2_{Y\ 35|12}$=0.5558232
$R^2_{Y\ 45|12}$=0.5642756

$R^2_{Y\ 45|12} > R^2_{Y\ 35|12} > R^2_{Y\ 34|12}$

$SSR_{(34|12)}$=4671904
$SSR_{(35|12)}$=78352775
$SSR_{(45|12)}$=79544288

$SSR_{(45|12)} > SSR_{(35|12)} > SSR_{(34|12)}$

X4X5(V7/Percent of population 65 or older)(V9/Number of hospital beds) is relatively more important than other pairs since $R^2_{Y\ 45|12}$ is the largest. Also, $SSR_{(45|12)}$ is larger than the extra sum of squares associated with other two pairs

$H_0: \beta_4=\beta_5=0$ $H_a: \beta_4 \neq 0$ and/or $\beta_5 \neq 0$

F*=281.67 P-value=2.2e-16

Reject H_0 at any standard significant level, adding X4 and X5 as a pair to the model is helpful when X1 and X2 are included in the model.

Part III.Discussion

Part 1: The results of the linear regression for model 1 indicate all three of our predictor variables—Total Population, Land Area, and Total Personal Income—are strongly correlated with our observed variable—Number of Active Physicians. However, based on the correlation matrix, we see that Total Population

and Total Personal Income are very strongly correlated, which is evidence that multicollinearity may be present in our line model.

The results of the linear regression for Model II indicate Population Density and Total Personal Income are strongly correlated with the Number of Active Physicians, but Percent 65 or Older was not significantly correlated with the Number of Active Physicians. It should be noted that since Population Density's p-value is lower than Total Personal Income's p-value, Population Density is more strongly correlated with Number of Active Physicians than Total Personal Income.

Both Model I and Model II have very high R-squared values, but Model II's R-squared was slightly higher than Model I's R-squared (even when comparing adjusted R-squared). Both models seem to have similar problems with the residuals—they suffer from heteroskedasticity and are not normal. Since the problems are similar, but Model II's R-squared value is slightly higher, Model II is preferred.

Part 2: Given Total Population and Total Personal Income are already included in our model, the Number of Hospital Beds is

the strongest predictor for the Number of Active Physicians. The other predictor variables, Land Area and Percent 65 or Older, are not nearly as strong predictors as Number of Hospital Beds, given that Total Population and Total Personal Income are included in our model.

R Appendix

Part I Code

```
Model1_X1 = CDI$`Total population
stem(CDI$`Total population`)
Model1_X2 = CDI$`Land area`
stem(CDI$`Land area`)
Model1_X3 = CDI$`Total personal income`
stem(CDI$`Total personal income`)
Model2_X1 = CDI$`Population Density`
stem(CDI$`Population Density`)
Model2_X2 = CDI$`Percent of population 65 or older`
stem(CDI$`Percent of population 65 or older`)
Model2_X3 = CDI$`Total personal income`
stem(CDI$`Total personal income`)
data1 = CDI[,c(8,5,4,16)]
colnames(data1) = c("Y","Model1_X1", "Model1_X2", "Model1_X3")
pairs(data1)
cor(data1)
data2 = CDI[,c(8,18,7,16)]
colnames(data2) = c("Y", "Model2_X1", "Model2_X2", "Model2_X3")
pairs(data2)
cor(data2)
Model1_fit = lm(Y~Model1_X1+Model1_X2+Model1_X3, data = CDI)
Model2_fit = lm(Y~Model2_X1+Model2_X2+Model2_X3, data = CDI)
summary(Model1_fit)
summary(Model2_fit)$coefficients
Model1_residuals = Model1_fit$residuals
plot(Model1_X1, Model1_residuals)
abline(h=0, col = 'red')
plot(Model1_X2, Model1_residuals)
abline(h=0, col = 'red')
plot(Model1_X3, Model1_residuals)
abline(h=0, col = 'red')
Y_hat1 = Model1_fit$fitted.values
plot(Y_hat1, Model1_residuals)
abline(h=0, col = 'red')
plot(Model1_X1*Model1_X2, Model1_residuals)
abline(h=0, col = 'red')
plot(Model1_X1*Model1_X3, Model1_residuals)
abline(h=0, col = 'red')
plot(Model1_X2*Model1_X3, Model1_residuals)
abline(h=0, col = 'red')
qqnorm(Model1_residuals)
qqline(Model1_residuals, col = "red")
Model2_residuals = Model2_fit$residuals
plot(Model2_X1, Model2_residuals)
abline(h=0, col = 'red')
plot(Model2_X2, Model2_residuals)
abline(h=0, col = 'red')
plot(Model2_X3, Model2_residuals)
abline(h=0, col = 'red')
Y_hat2 = Model2_fit$fitted.values
plot(Y_hat2, Model2_residuals)
```

```
abline(h=0, col = 'red')
plot(Model2_X1*Model2_X2, Model2_residuals)
abline(h=0, col = 'red')
plot(Model2_X1*Model2_X3, Model2_residuals)
abline(h=0, col = 'red')
plot(Model2_X2*Model2_X3, Model2_residuals)
abline(h=0, col = 'red')
qqnorm(Model2_residuals)
qqline(Model2_residuals, col = "red")
Model1_fit_int =
lm(Y~Model1_X1+Model1_X2+Model1_X3+Model1_X1*Model1_X2+Model1_X2*Model1_X3+Model1_X2*M
odel1_X3, data = CDI)
summary(Model1_fit_int)$r.squared
Model2_fit_int =
lm(Y~Model2_X1+Model2_X2+Model2_X3+Model2_X1*Model2_X2+Model2_X2*Model2_X3+Model2_X2*M
odel2_X3, data = CDI)
summary(Model2_fit_int)$r.squared
```

Part II Code

```
fit = lm(V8 ~ V5+V16, data=CDI) #v8=Y=Number of active physicians V5=X1=Total
population V16=X2=Total personal income
anova(fit)
model.before = lm(V8 ~ V5+V16, data=CDI)
model.afterX3 = lm(V8 ~ V5+V16+V4, data=CDI)
SSE.before = sum(model.before$residuals^2)
SSE.afterX3 = sum(model.afterX3$residuals^2)
X3extra.SS = SSE.before - SSE.afterX3
X3extra.SS
X3partial.R2 = (SSE.before - SSE.afterX3)/(SSE.before)
X3partial.R2
model.before = lm(V8 ~ V5+V16, data=CDI)
model.afterX4 = lm(V8 ~ V5+V16+V7, data=CDI)
SSE.before = sum(model.before$residuals^2)
SSE.afterX4 = sum(model.afterX4$residuals^2)
X4extra.SS = SSE.before - SSE.afterX4
X4extra.SS
X4partial.R2 = (SSE.before - SSE.afterX4)/(SSE.before)
X4partial.R2

model.before = lm(V8 ~ V5+V16, data=CDI)
model.afterX5 = lm(V8 ~ V5+V16+V9, data=CDI)
SSE.before = sum(model.before$residuals^2)
SSE.afterX5 = sum(model.afterX5$residuals^2)
X5extra.SS = SSE.before - SSE.afterX5
X5extra.SS
X5partial.R2 = (SSE.before - SSE.afterX5)/(SSE.before)
X5partial.R2
partialR2.X3X4X5=c(X3partial.R2,X4partial.R2,X5partial.R2)
rank(partialR2.X3X4X5)
```

```
extraSS.X3X4X5=c(X3extra.SS,X4extra.SS,X5extra.SS)
rank(extraSS.X3X4X5)
anova(model.before, model.afterX5)
model.before = lm(V8 ~ V5+V16, data=CDI)
model.afterX3X4 = lm(V8 ~ V5+V16+V4+V7, data=CDI)
SSE.before = sum(model.before$residuals^2)
SSE.afterX3X4 = sum(model.afterX3X4$residuals^2)
X3X4extra.SS = SSE.before - SSE.afterX3X4
X3X4extra.SS
X3X4partial.R2 = (SSE.before - SSE.afterX3X4)/(SSE.before)
X3X4partial.R2
model.before = lm(V8 ~ V5+V16, data=CDI)
model.afterX3X5 = lm(V8 ~ V5+V16+V4+V9, data=CDI)
SSE.before = sum(model.before$residuals^2)
SSE.afterX3X5 = sum(model.afterX3X5$residuals^2)
X3X5extra.SS = SSE.before - SSE.afterX3X5
X3X5extra.SS
X3X5partial.R2 = (SSE.before - SSE.afterX3X5)/(SSE.before)
X3X5partial.R2
model.before = lm(V8 ~ V5+V16, data=CDI)
model.afterX4X5 = lm(V8 ~ V5+V16+V7+V9, data=CDI)
SSE.before = sum(model.before$residuals^2)
SSE.afterX4X5 = sum(model.afterX4X5$residuals^2)
X4X5extra.SS = SSE.before - SSE.afterX4X5
X4X5extra.SS
X4X5partial.R2 = (SSE.before - SSE.afterX4X5)/(SSE.before)
X4X5partial.R2
partialR2.X3X4.X3X5.X4X5=c(X3X4partial.R2,X3X5partial.R2,X4X5partial.R2)
rank(partialR2.X3X4.X3X5.X4X5)
extraSS.X3X4.X3X5.X4X5=c(X3X4extra.SS,X3X5extra.SS,X4X5extra.SS)
rank(extraSS.X3X4.X3X5.X4X5)
anova(model.before, model.afterX4X5)
```