Data Analysis of
County Demographic Information
(CDI)

Youngchan Park	914955115
Chan Woong Joo	913053322

Jiming Jiang STA 108

March 9th, 2020

Project 2

Introduction

The United States Census Bureau is an agency responsible for producing population data across the United States. They not only produce population information but also gather information about so many variables to make models to analyze the county and country's economy.

Throughout our analysis, we will be using the County Demographic Information (CDI) data provided by the United States Census Bureau. The data have information of 440 counties with 14 different variables for a single county such as land area, total population, percent of population in a specific age group, number of active physicians, number of hospital beds, percent bachelor's degree, per capita income, total personal income, etc. Furthermore, counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992.

In continuation to our Project 1, we will be conducting our analysis assuming the linear regression model is appropriate. We will further analyze data with multiple regression and investigate how each variable is significant. We object to finding the correlation of certain variables and how one variable could explain the other variable to which extent. We will formulate a linear regression model to estimate an independent variable and, with statistical analysis, determine how well the variable is explained by the model.

Throughout our research, we will be using Rstudio, a widely used statistical and graphing utility by coding a programming language. Furthermore, we will be referencing the textbook "Applied Linear Statistical Models, Fifth Edition" by Kutner, et.al. The outline of our project is as follow:

- Part I: *Multiple linear regression I*
- Part II: *Multiple linear regression II*
- Part III: *Discussion*
- Appendix

Part I: *Multiple linear regression I*

6.28

a.

According to the data of Country Demographic Information (CDI) and assuming that first-order multiple regression model is appropriate, we could construct a regression model for two Models with number of active physicians as dependent variable: Model 1 including total population (X1), land area (X2) and total personal income (X3) as predictor variables, and Model 2 including population density (X1), which can be derived as dividing total population by land area, percent of population greater than 64 years old (X2) and total personal income as predictor variables. However, before regressing the variables, we first prepared stem-and-leaf plots for each of the predictor variables to obtain some noteworthy aspects of each predictors.

Stem-and-leaf plot of Total Population

```
The decimal point is 6 digit(s) to the right of the |

0 | 111111111111111111111111111111111111111111111111111111111111111+254
0 | 55555555555555555555555566666666666666777777777777777777778888888888
1 | 000000122233333444
1 | 55699
2 | 1134
2 | 58
3 |
3 |
4 |
4 |
5 | 1
5 |
6 |
6 |
7 |
7 |
8 |
8 | 9
```

Stem-and-leaf plot of Land Area

```
The decimal point is 3 digit(s) to the right of the |

  0 | 000011111111111222222222222222222222223333333333333333333333333333334444444+252
  1 | 000000000000000111111111111122222222222333333444455566667777888999 9
  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
```

Stem-and-leaf plot of Total Income

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

   0 | 1111111111111222222222222222222222222222222222222222222222222222222+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
```

Stem-and-leaf plot of Population Density

```
The decimal point is 3 digit(s) to the right of the |

   0 | 00000000000000001111111111111111111111111111111111111111111111111+321
   2 | 0000111223345670011114 5
   4 | 05884
   6 | 2464
   8 | 19
  10 | 378
  12 |
  14 | 4
  16 |
  18 |
  20 |
  22 |
  24 |
  26 |
  28 |
  30 |
  32 | 4
```

Stem-and-leaf of Elderly Percentage

```
The decimal point is at the |

   2 | 0
   4 | 47890389
   6 | 1123455677990134566678899
   8 | 0011222223333444455566677777888889999000222233333344444444555666677
  10 | 000111111222222222233333344444455555556666666667777777888888888899999+36
  12 | 000000001111122223333333333444455555556666666777777777888899900000000+36
  14 | 00001111111223334444455567788900000011112222345566777 8
  16 | 12556699901122345
  18 | 06778
  20 | 070
  22 | 018828
  24 | 47
  26 | 055
  28 | 1
```

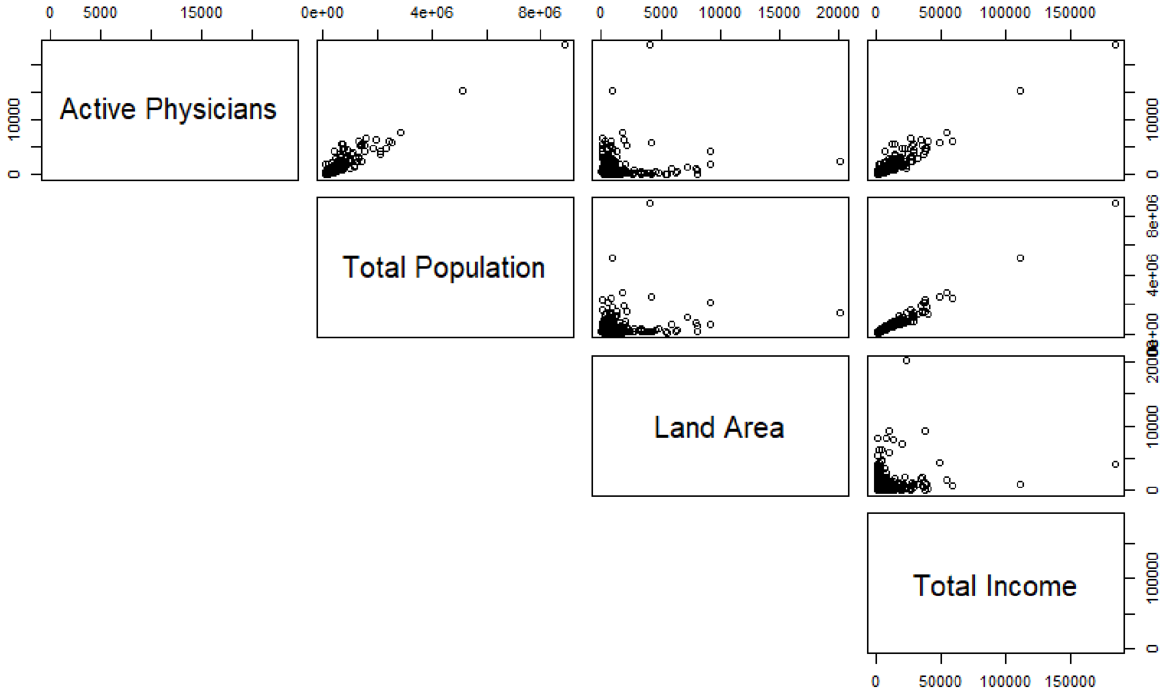
```
30 | 7
32 | 138
```

Stem-and-leaf plots provide information of where most of the data points are located. We have investigated 5 predictor variables: total population, land area, total income, population density, and elderly percentage. According to the stem-and-leaf plot of total population, it seems that most of the data points have population of approximately 100,000 people with some outliers. Similarly, according to the stem-and-leaf plot of land area, most of the data points have land area approximately between 100 and 1000 with some outliers. Likewise, most of the counties have total personal income of approximately 1000 millions of dollars with some outliers. With the same interpretation, most of the counties have a population density of 1000 people per square mile. Interestingly, the stem-and-leaf plot of elderly percentage provides information that most of the data points are distributed around 8 to 14 percentage.

With the stem-and-leaf plots of specific predictor variables, we were able to have a glimpse of the data points distribution that we could further analyze in first-order multiple regression model.

b.

Model 1:



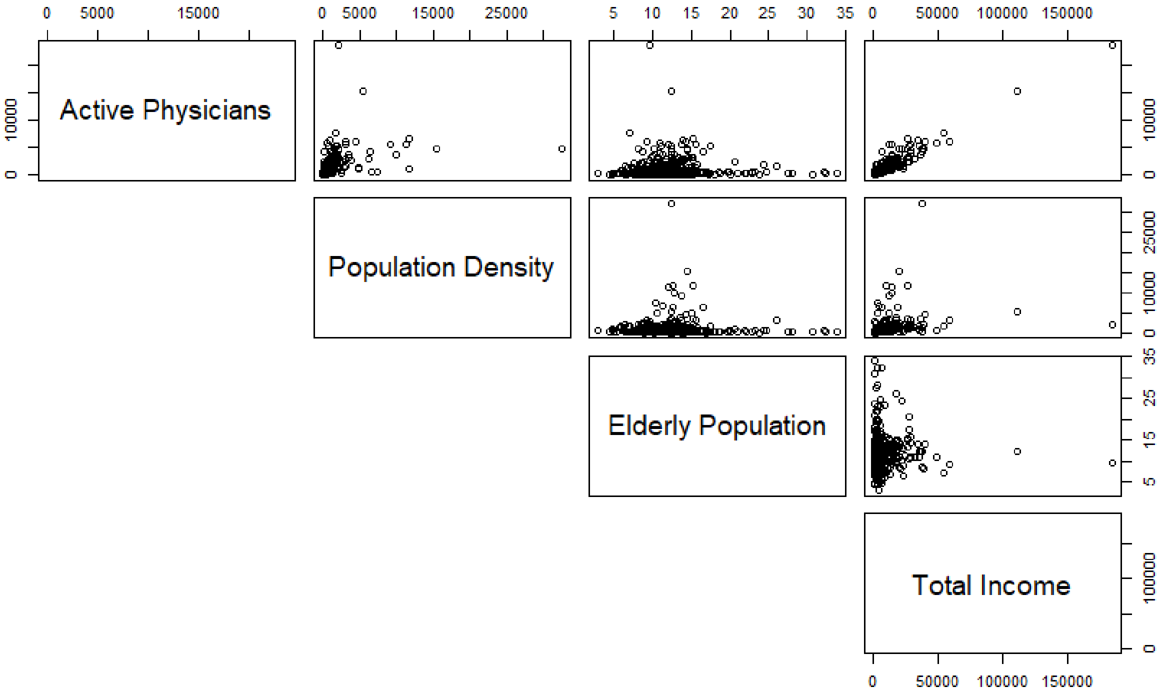
	Active Physicians	Total Population	Land Area	Total Income
Active Physicians	1.00000000	0.9402486	0.07807466	0.9481106
Total Population		1.0000000	0.17308335	0.9867476
Land Area			1.00000000	0.1270743
Total Income				1.0000000

Correlation matrix provides information showing correlation coefficient between sets of variables. The diagonal of the table is always one because the correlation between one variable and itself is always 1.

From the scatter plot matrix Model 1 above, we found out that the predictor variables of total income and total population is highly correlated with the number of active physicians. In other words, total income and total population are highly correlated.

From the correlation matrix, Active physicians and total population and total income have correlation coefficients of 0.9402486 and 0.9481106. The total population and total income have correlation coefficient of 0.9867476, implying that one unit change in either predictor variable will cause 0.9867476, this would apply the same for the other coefficients.

Model 2:



	Active Physicians	Population Density	Elderly Population	Total Income
Active Physicians	1.00000000	0.40643863	-0.00312863	0.94811057
Population Density		1.00000000	0.02918445	0.31620475
Elderly Population			1.00000000	-0.02273315
Total Income				1.00000000

From the scatterplot matrix Model 2 above, we found out that total income and the number of active physicians are highly correlated, with the coefficient value of 0.94811067. This implies that one unit change in either variable will cause unit change by the amount of coefficient. Population density and active physicians has weak but somewhat significant correlation coefficient with 0.40643863. Population density and total income has a weak but somewhat significant correlation coefficient with 0.31620475. We found out that the predictor variable of total income is mostly correlated with the dependent variable number of active physicians.

c.

Having obtained the scatter plot matrix and correlation matrix, we could then derive the multiple linear regression model with three predictor variables for each model. For model 1, the three predictor variables are: total population (X_1), land area (X_2), and total personal income(X_3). For model 2, the three predictor variables are: population density (X_1), percent of population greater than 64 years old (X_2), and total personal income (X_3). The first-older multiple linear regression models are as follows:

Model 1:

$$\hat{Y} = -13.32 + 0.0008366X_1 - 0.06552X_2 + 0.09413X_3$$

Model 2:

$$\hat{Y} = -170.6 + 0.09616X_1 + 6.34X_2 + 0.1266X_3$$

With the regression models, we could estimate dependent variable, number of active physicians, with each predictor variable.

d.

Multiple R^2 values provide information on how the dependent variable and predictors are correlated, R^2 close to 0 implying insignificant correlation and R^2 close to 1 implying significant correlation.

Model 1: $R^2 = 0.9026$

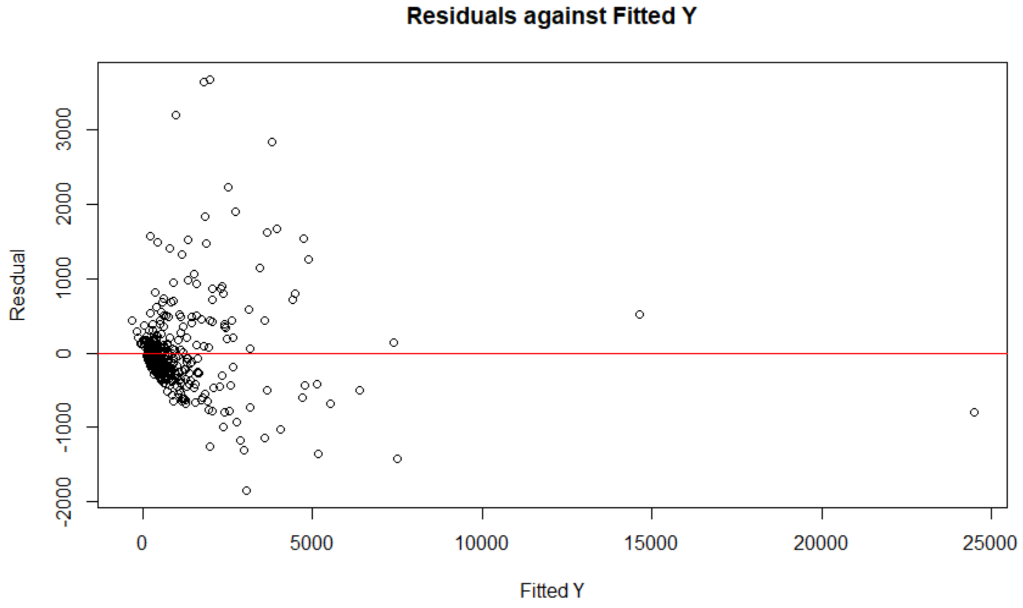
Model 2: $R^2 = 0.9117$

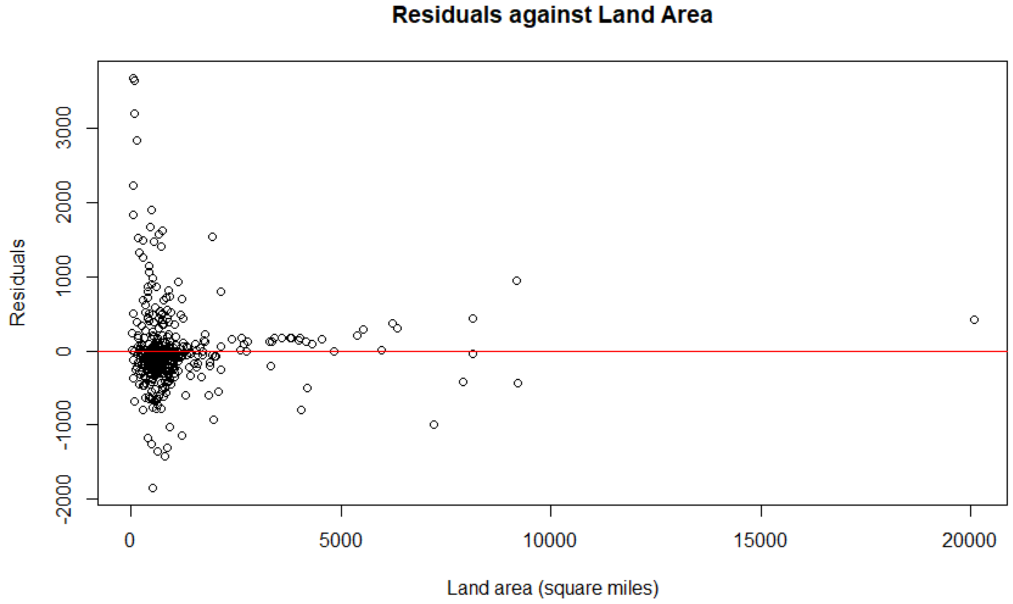
Model 1 and 2 both provide noticeable R^2 values; with R^2 values of 0.9026 and 0.9117 correspondingly, they imply that the number of active physicians and each predictor variable are highly correlated. Although similar, since R^2 of Model 2 is closer to 1, it is clear that Model 2 is more preferable than Model 1 in terms of this measure.

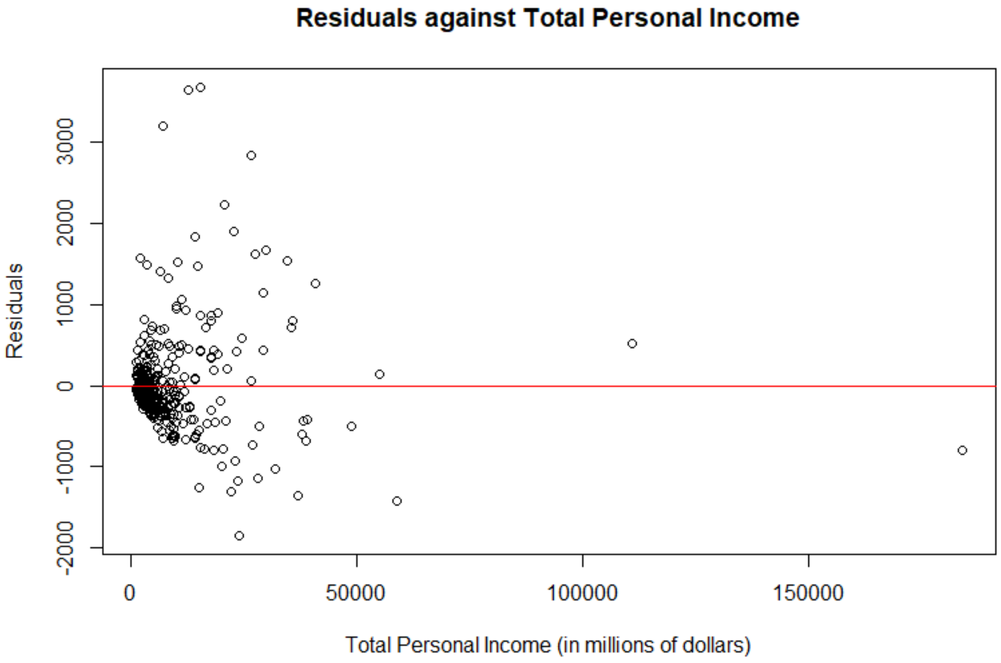
e.

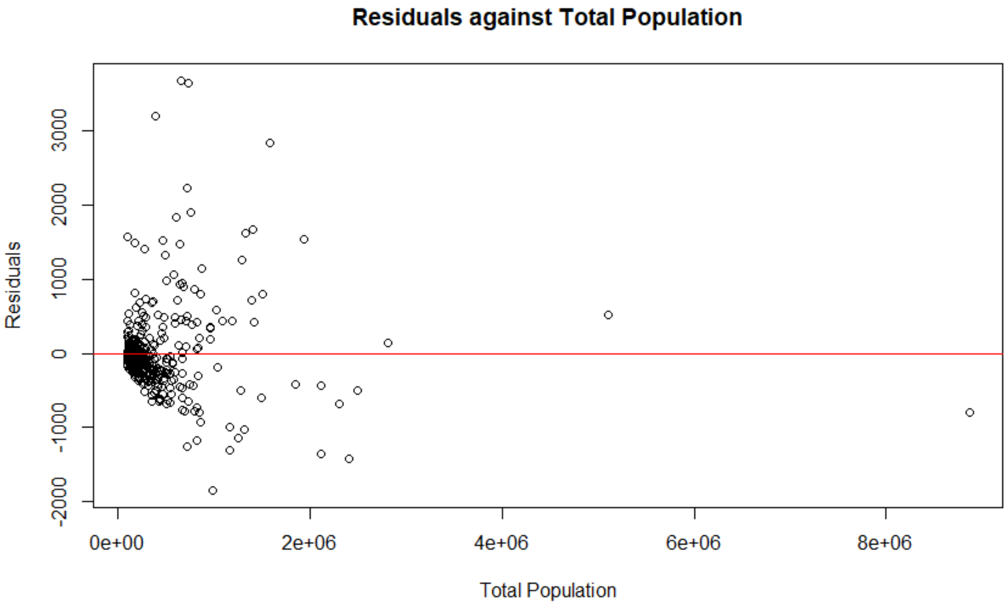
Residual plot is a measure of how the data points are regressing from the estimator. If the residuals are evenly distributed across the zero-line, we could obtain information that the regression model is a good estimator. We have prepared the residuals plot against fitted Y, Ŷ, and each predictor variable to see the quality of our regression model.

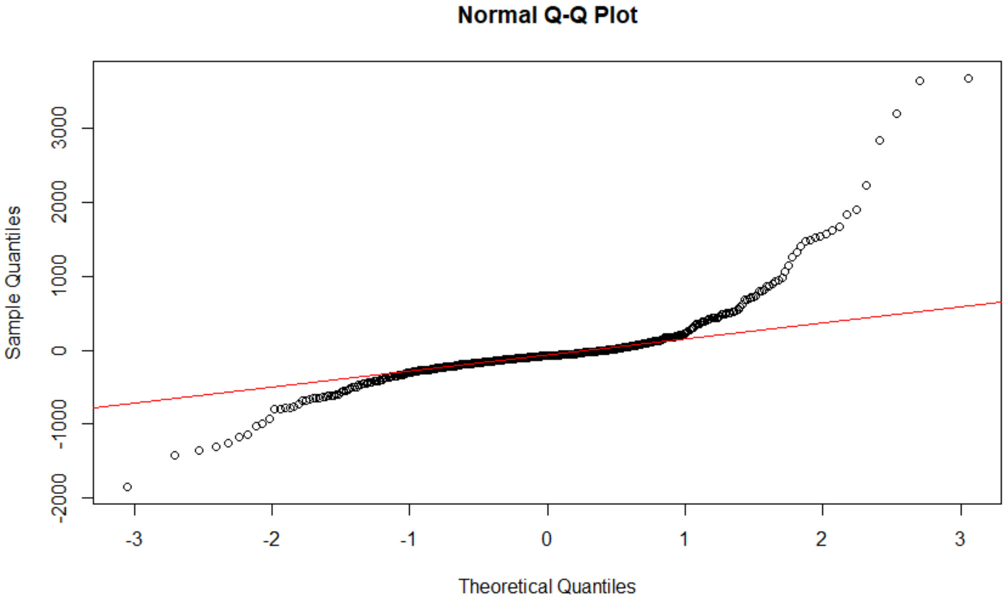
Model 1:



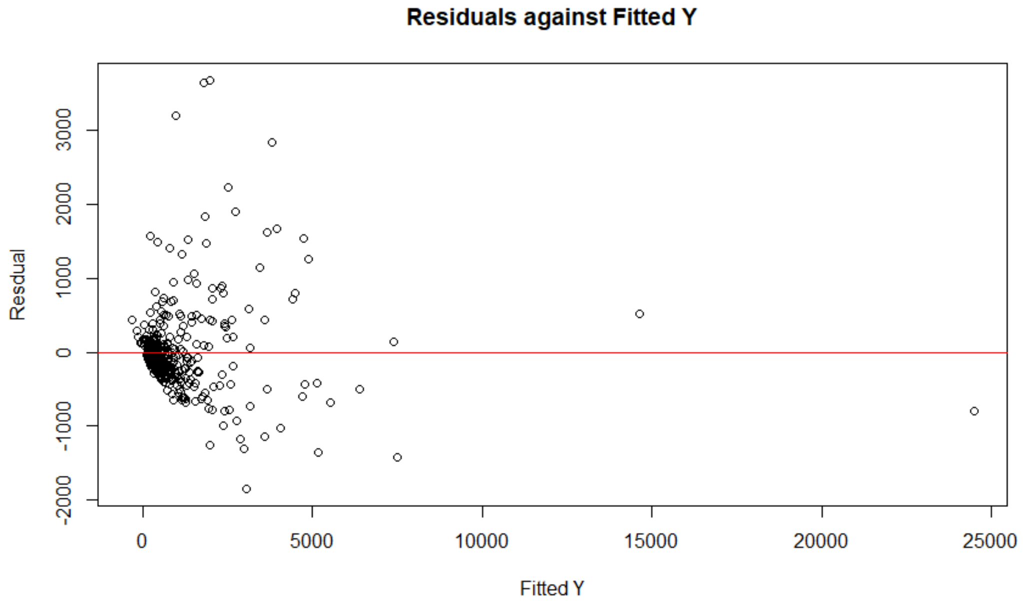



Residuals against Total Personal Income

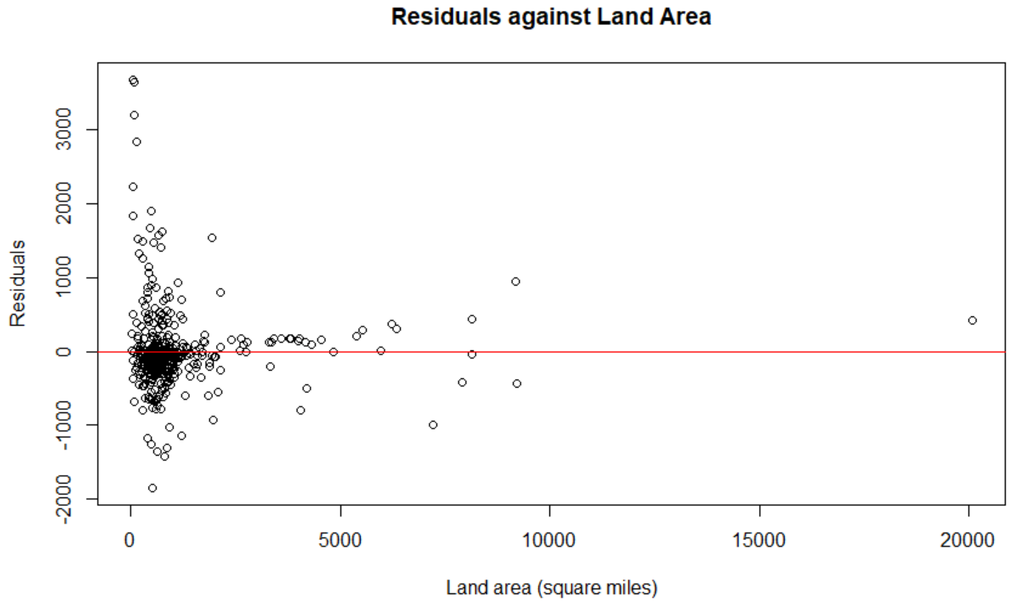

Residuals against Total Population


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
3000
2000
1000
0
-1000
-2000
-3
-2
-1
0
1
2
3

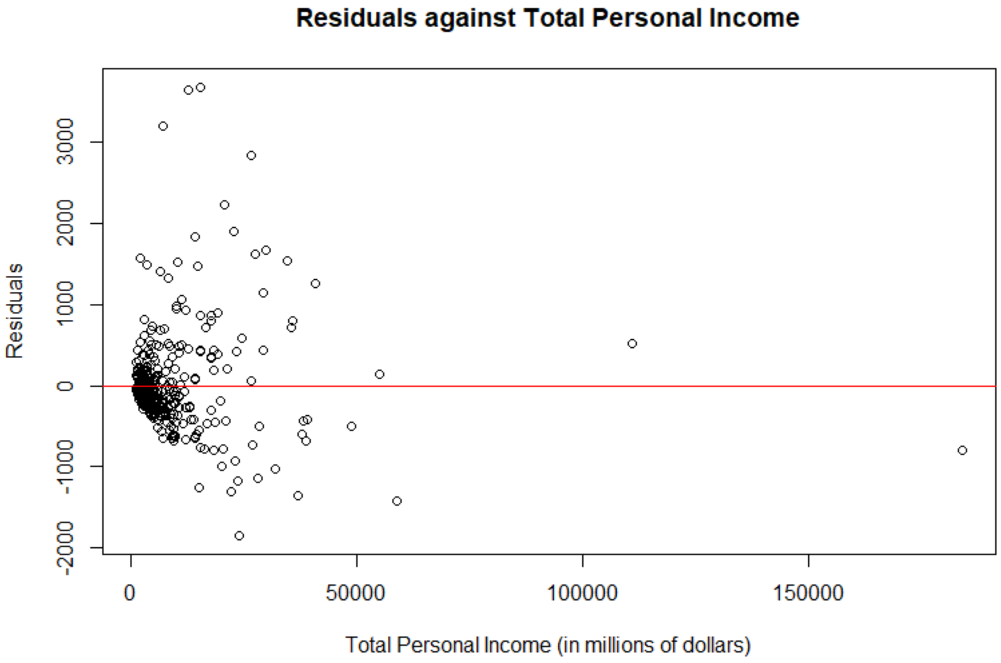
Model 2:

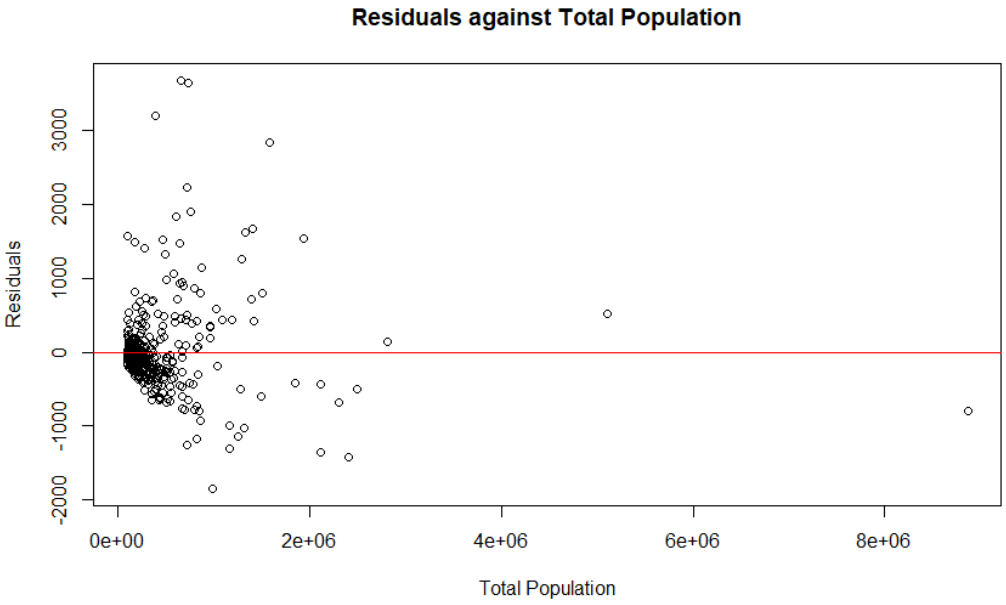


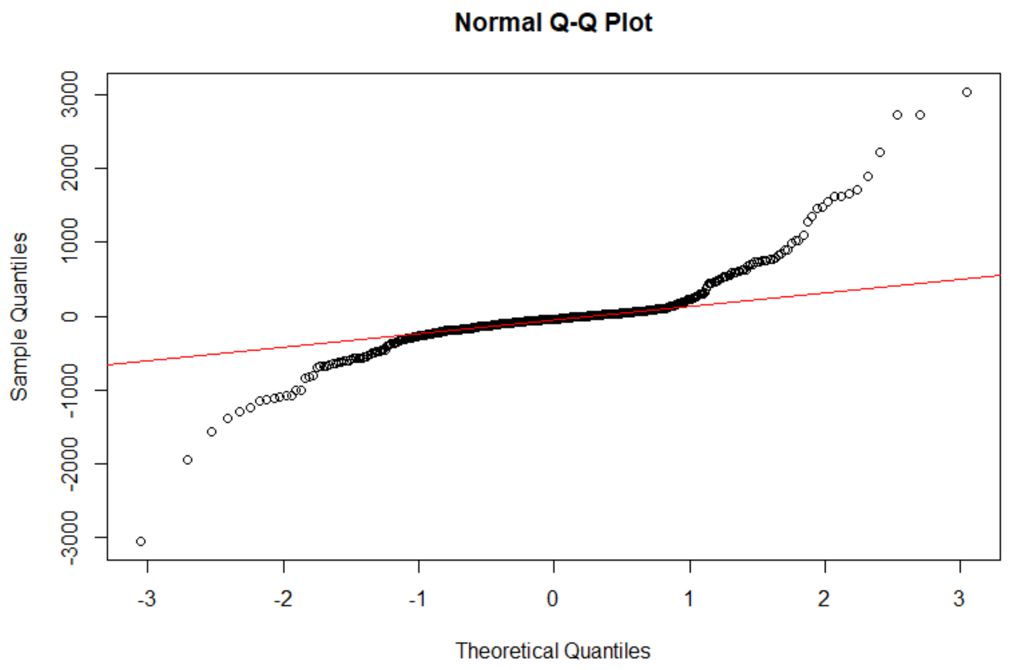
Residuals against Fitted Y
Resdual
3000
2000
1000
0
-1000
-2000
0
5000
10000
15000
20000
25000
Fitted Y



Residuals against Land Area
Residuals
3000
2000
1000
0
-1000
-2000
0
5000
10000
15000
20000
Land area (square miles)


Residuals against Total Personal Income
Residuals
-2000
-1000
0
1000
2000
3000
0
50000
100000
150000
Total Personal Income (in millions of dollars)


Residuals against Total Population
Residuals
-2000
-1000
0
1000
2000
3000
0e+00
2e+06
4e+06
6e+06
8e+06
Total Population



From the residual plots for Model 1 and Model 2 above, it seems that residuals against fitted value are normally dispersed, not being concentrated to a certain value. At the same time, it seems that there are many extreme values within the data. Most of the samples are located near zero residual.

Furthermore, our normal-QQ plots do not provide a straight line, a line that implies the residuals are normally distributed. The graphs clearly suggest that most of the samples fit in the regression line. However, the QQ plots we achieved all pertain tails at the end. Therefore, the plots we obtained provide information that our data points have more outliers, or extreme values, than estimated assuming they came from a normal distribution. The relationship between the sample quantile and theoretical quantile is not linear, implying that errors are not normally distributed.

f.

To improve the multiple regression model as an estimator, it is a great approach to add two-factor interaction terms to the model. When regressing three or more predictor variables, it is possible that an interaction is evident in which the change of one variable is affected by another uncertain predictor variable. Therefore, it is possible that, if there is a interaction between predictor variables, adding the interaction terms into the model may provide a better estimating model. For example, considering full model to be:

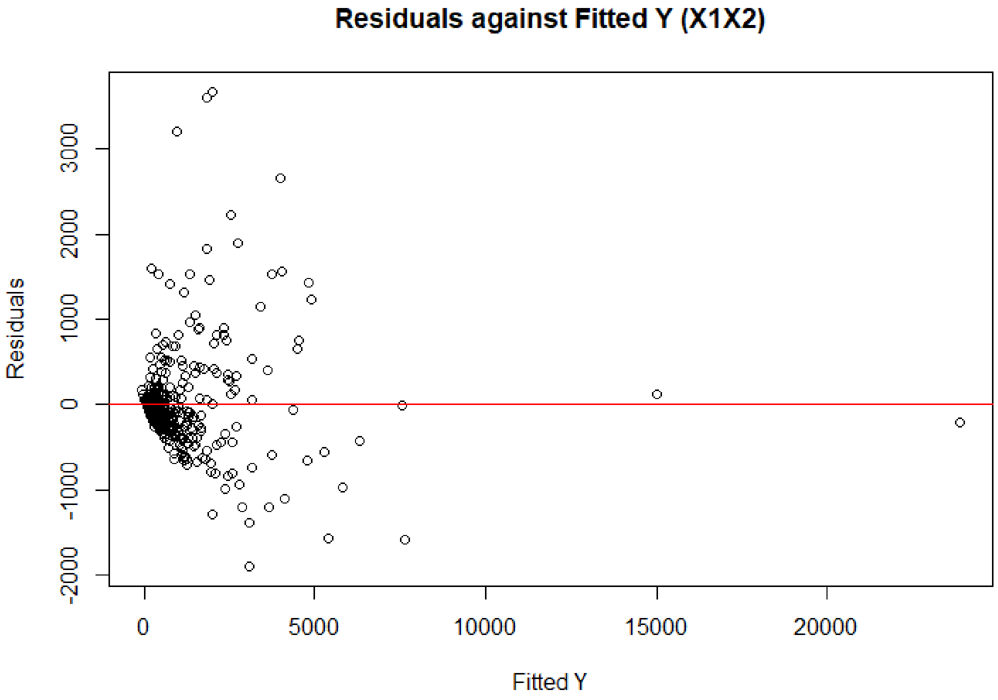
$$\hat{Y} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3,$$

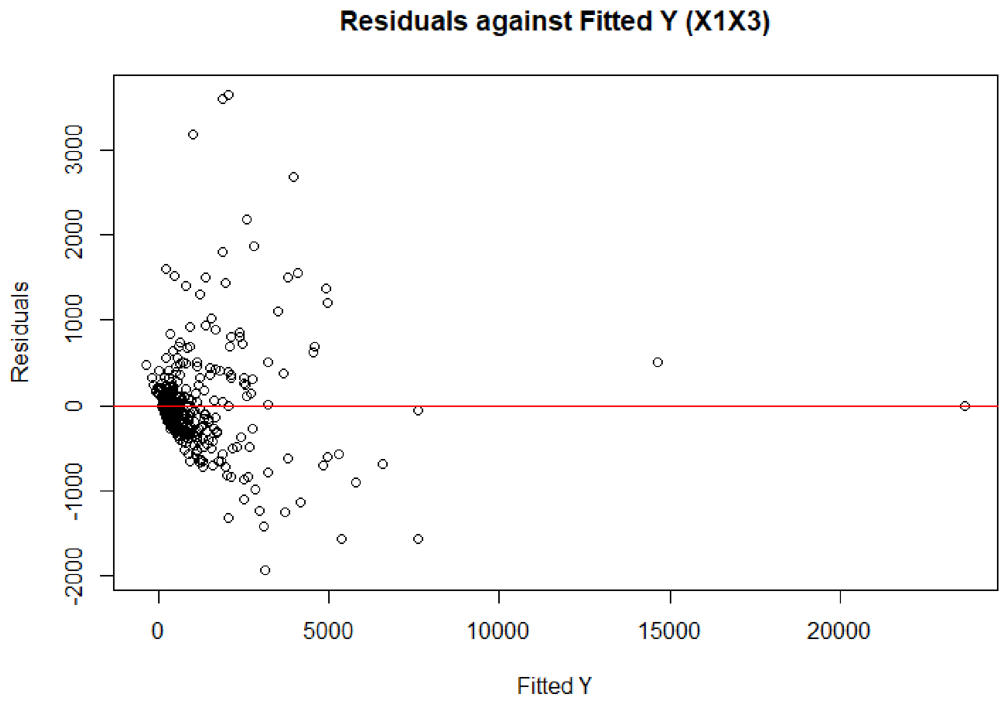
we can add a two-factor interaction term to the model as

$$\hat{Y} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_1X_2 .$$

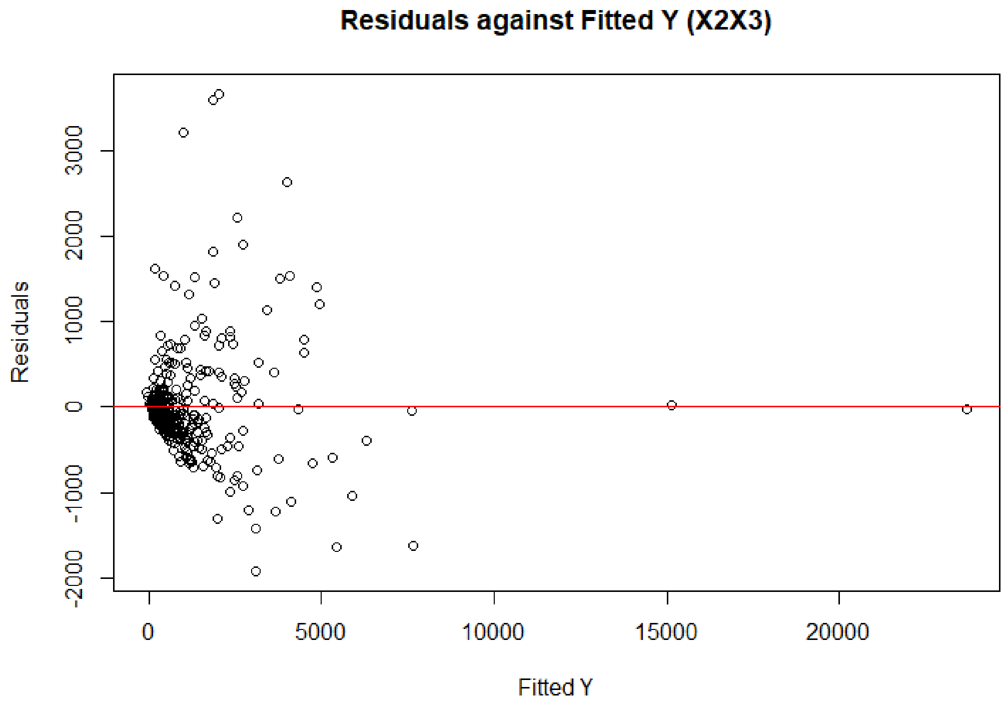
To analyze if adding interaction terms into the model provides a better estimator, it is also important to calculate coefficient of correlation of each model with additional interaction predictors. In order to figure out if the predictor variables are statistically significant, we need to obtain residuals plot against two-factor interaction terms:

Model 1:



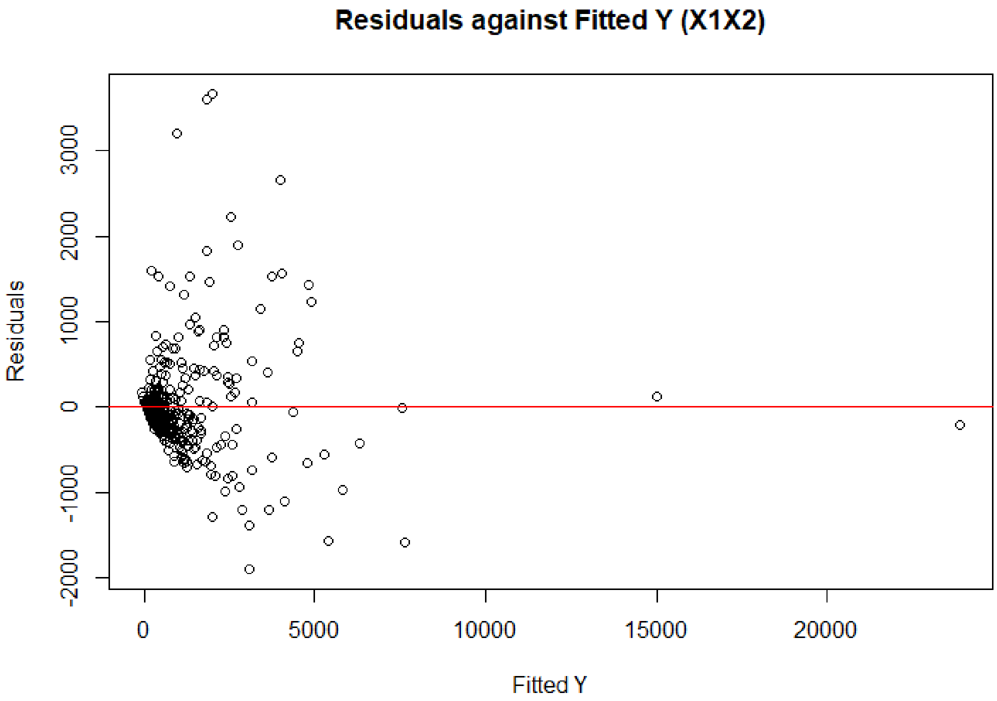


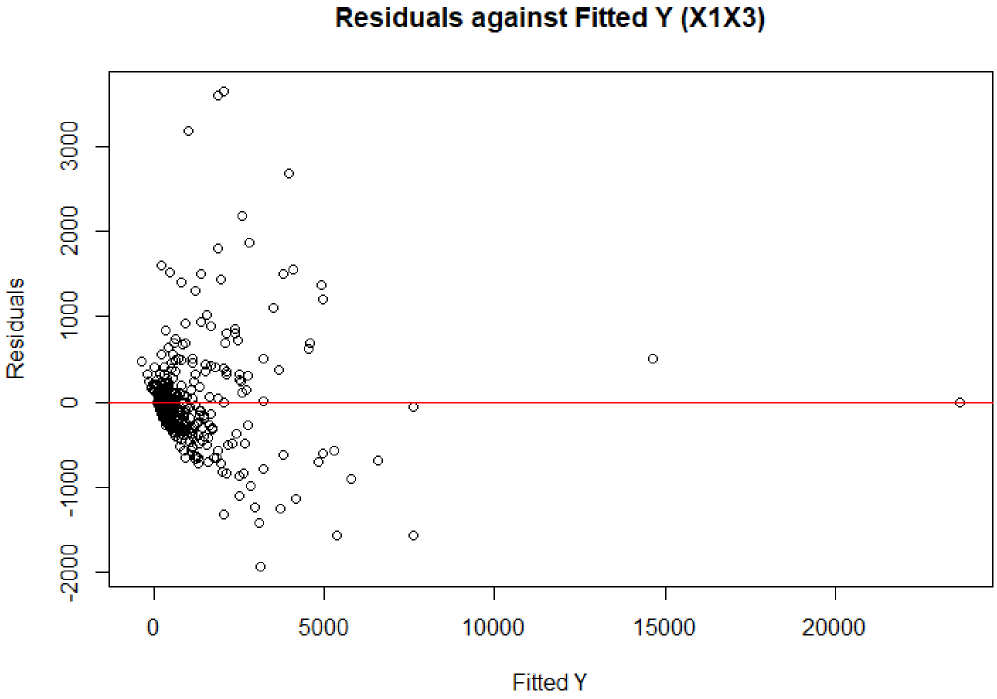
Residuals against Fitted Y (X1X3)

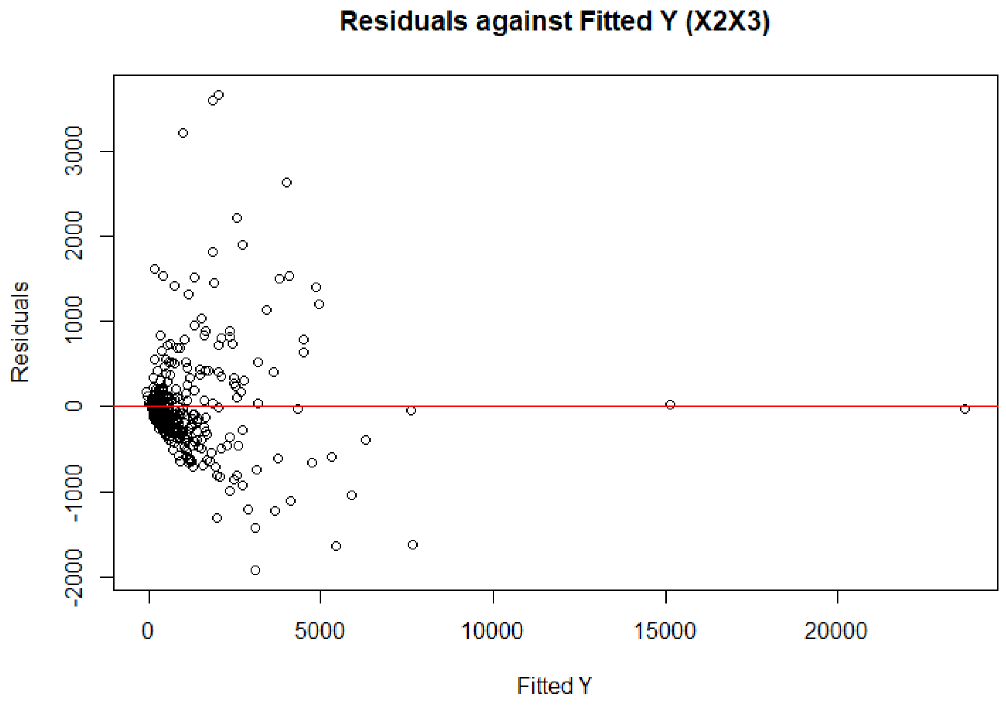


Residuals against Fitted Y (X2X3)

Model 2:


Residuals against Fitted Y (X1X2)
Residuals
-2000
-1000
0
1000
2000
3000
0
5000
10000
15000
20000
Fitted Y


Residuals against Fitted Y (X1X3)
Residuals
-2000
-1000
0
1000
2000
3000
0
5000
10000
15000
20000
Fitted Y



Similar to residual plots before adding two-factor interaction terms, the model with additional two-factor interaction terms also provide equally distributed residuals across the zero-line. For this reason, it is difficult to analyze if the new model with additional interaction term provide a better estimating model, and therefore, it is important to analyze using multiple R^2 of each model with additional two-factor interaction term

Model 1:

R^2 with X_1X_2 added to the full model = 0.9039

R^2 with X_1X_3 added to the full model = 0.9036

R^2 with X_2X_3 added to the full model = 0.9044

Model 2:

R^2 with X_1X_2 added to the full model = 0.9191

R^2 with X_1X_3 added to the full model = 0.9165

R^2 with X_2X_3 added to the full model = 0.9122

For Model 1, since R^2 with X_2X_3 added to the full model is the largest among the three, adding X_2X_3 (land area and total personal income) into the model provides a better estimating regression model.

Similarly for Model 2, since R^2 with X_1X_2 added to the full model is the largest among the three, adding X_1X_2 (population density and percent of population greater than 64 years old) into the model provides a better estimating regression model.

Again, for Model 1, R^2 value of full model before adding two-factor interaction term was 0.9026 and for Model 2, R^2 value of full model before adding two-factor interaction term was 0.9117. By adding the product of two predictors into the model, we could achieve an increase in R^2 value, meaning that the regression model could be improved in estimating the dependent variable.

Part II: *Multiple linear regression II*

In order to predict if an additional predictor variable is effective, it is necessary to investigate the coefficient of partial determination. Coefficient of partial determination is a measure of correlation between variables given certain predictor variables. Assuming a first-order multiple regression is appropriate, we could calculate if an additional variable is helpful in estimating dependent variable.

a.

Given that X_1 (total population) and X_2 (total personal income) are already in the model, we could derive coefficient of partial determination of X_3 (land area), $R^2_{x3|x1x2}$, X_4 (percent of population 65 or older), $R^2_{x4|x1x2}$, and X_5 (number of hospital beds), $R^2_{x5|x1x2}$. The formula for deriving coefficient of partial determination is:

$$\text{Partial } R^2 = \frac{SSEreduced - SSEfull}{SSEfull}$$

Using the formula above, we could then derive the coefficient of partial determination:

$R^2_{x3|x1x2} = 0.02882495$

$R^2_{x4|x1x2} = 0.003842367$

$R^2_{x5|x1x2} = 0.5538182$

b.

On the basis of the result from part(a), we could conclude that an additional predictor variable of X_5 (number of hospital beds) provides a better regression model. Among the three Partial R^2 values, $R^2_{x5|x1x2}$ of 0.5538182 is the best additional predictor. However, to verify the result, we could look at the extra sum of squares of each additional predictor variable, and whichever predictor with large extra sum of squares is the proof that the additional predictor is appropriate.The values are given as follows:

Extra SSE X_3 = 4063370

Extra SSE X_4 = 541647.3

Extra SSE X_5 = 78070132

Since the extra sum of squares of X_5 (number of hospital beds) is the largest, it is clear that the additional predictor of X_5 (number of hospital beds) is the best compared to X_3 (land area) and X_4 (percent of population 65 or older).

c.

On the basis of the result from part(b), we could conclude that an additional predictor variable of X_5 (number of hospital beds) provides a better regression model. To verify using F* test statistic, we should test whether X_5 is helpful in the regression model when X_1 and X_2 are already included in the model. We define a full model which includes X_1, X_2 and X_5 and reduced model including X_1 and X_2.

To test if an additional predictor variable is significant, we establish hypothesis as follows:

$$H_0 : \beta_5 = 0$$

$$H_a : \beta_5 \neq 0$$

Using the F statistics with α = 0.01, we could test our hypothesis with the following decision rules:

Decision rule: if F* > F(.99,1, 436) conclude H_a
Decision rule: if F(.99,1, 436) > F* conclude H_0

F* could be derived by dividing MSR of X_5 given that X_1 and X_2 are already in the model and MSE of the full model. The F* could be calculated as follows:

$$F^* = \frac{MSR(X5|\, X1, X2)}{MSE(\, X1, X2, X5)} = \frac{78070132}{144259.1} = 541.1801$$

Since F(.99,1, 436) = 6.69336, and 541.1801 is way larger than 6.69336, we should reject our null hypothesis and conclude that an additional predictor variable of X_5 (number of hospital beds) provides a better regression model.

Furthermore, F statistic for the other predictors X_3, X_4 should not be large as X_5, because MSR $(X3|\, X1,\ X2)$ and MSR$(X4|X1, X2)$ are less than $MSR\ (X5|\ X1,\ X2)$ Since MSR of X_3 given X_1 and X_2 and MSR X_4 given X_1 and X_2 are smaller than MSR of X_5 given X_1 and X_2 it is impossible to have larger F* test statistics for the other predictor variables.

d.

Given that X_1 (total population) and X_2 (total personal income) are already in the model, we could derive coefficient of partial determination of X_3 (land area) and X_4 (percent of population 65 or older), $R^2_{x3,\ x4|x1x2}$, partial determination of X_3 (land area) and X_5 (number of hospital beds) $R^2_{x3,\ x5|x1x2}$, X_4 (percent of population 65 or older) and X_5 (number of hospital beds), $R^2_{x4,\ x5|x1x2}$. The formula for deriving coefficient of partial determination is:

$$\text{Partial } R^2 = \frac{SSEreduced - SSEfull}{SSEfull}$$

Using the formula above, we could then derive the coefficient of partial determination as:

$R^2_{x3,\ x4|x1x2} = 0.03314181$

$R^2_{x3,\ x5|x1x2} = 0.5558232$

$R^2_{x4,\ x5|x1x2} = 0.5642756$

From the partial coefficient, the pair of two variables x4 and x5 has more correlation than other pairs x3,x4 and x3, x5. This implies that the pair of x4 and x5 is relatively more important than other pairs.

To test if a pair of two predictor variables are significant, we establish hypothesis as follows:

$$H_0 : \beta_4, \beta_5 = 0,$$
$$H_a : \text{not all } \beta_4, \beta_5 = 0$$

Using the F statistics with $\alpha = 0.01$, we could test our hypothesis with the following decision rules:

Decision rule: if $F^* > F(.99,2, 435)$ conclude H_a
Decision rule: if $F(.99,2, 435) > F^*$ conclude H_0

F* could be derived by dividing MSR of X_4, X_5 given that X_1 and X_2 are already in the model and MSE of the full model. The F* could be calculated as follows:

$$F^* = \frac{MSR(X4,X5|\ X1, X2)}{MSE(\ X1, X2, X4, X5)} = 281.67$$

Since F(.99,2, 435) =4.65426919, and 281.67 is way larger than 4.65426919, we should reject our null hypothesis and conclude that an additional predictor variables of X_4(percent of population 65 or older), X_5 (number of hospital beds) provides a better regression model.

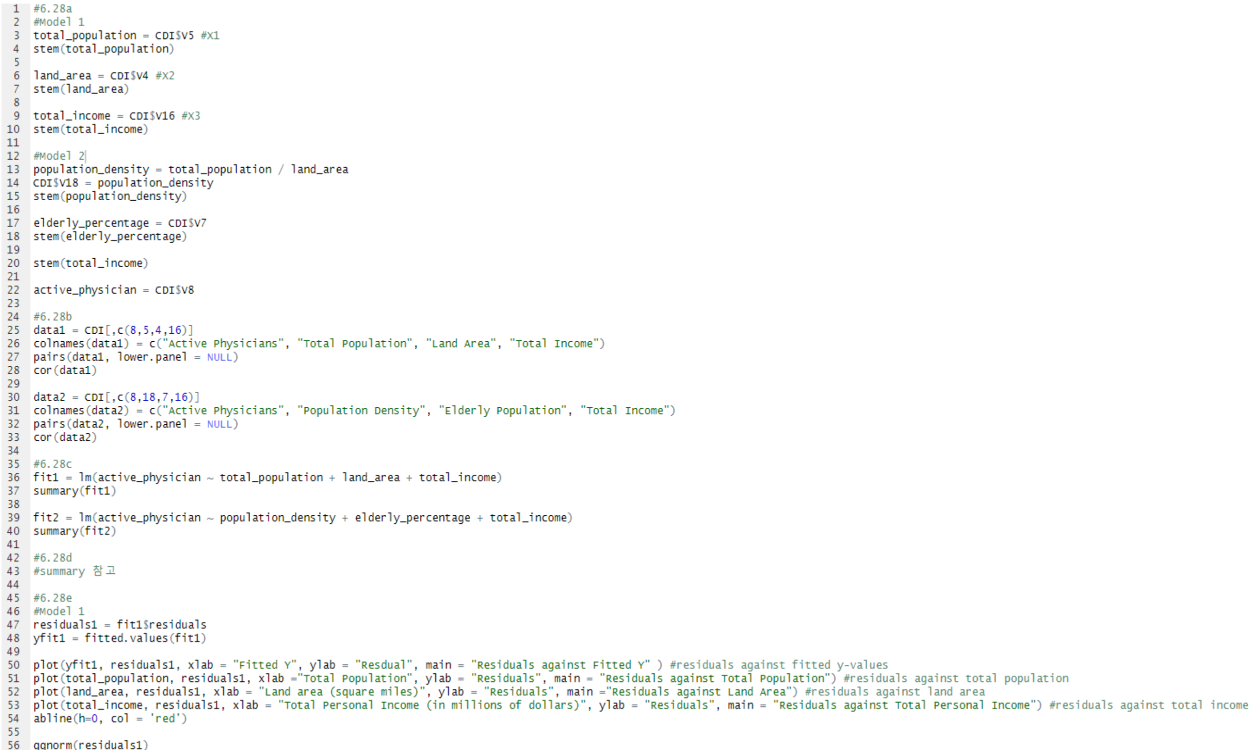
Part III: Discussion

Throughout analyzing the County Demographic Information (CDI) data, we further expanded our multiple linear regression analysis. We have investigated how each predictor variable could be an estimator to predict a certain dependent variable. We have found out that the coefficients of correlation provide important information in multiple linear regression models, giving the most helpful evidence to CDI data analysis.

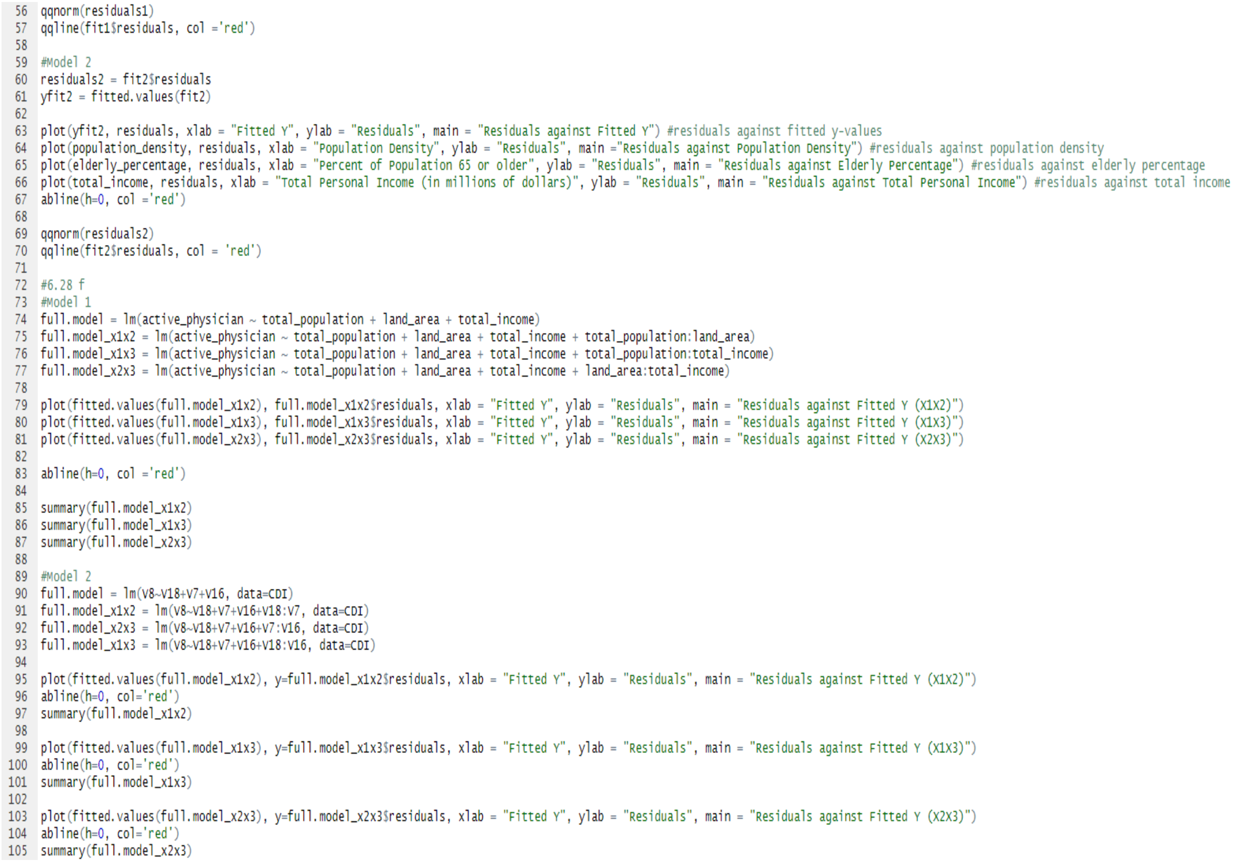
Obtaining various kinds of plots such as stem-and-leaf plot, residuals plot, QQ plot, we have found out how the data points are distributed across the regression line, providing insight for the data. We further expanded our analysis by deriving a coefficient of correlation and it directly gave us information of which predictor variables affect the dependent variable we desired the most. R^2 value gave us the criteria to determine which model is more useful to predict data, and coefficient of partial determination gave us the criteria to which predictors are correlated in the multiple regression analysis.

In order to enhance our regression analysis, the model should be sufficient enough to solve more complexity. In other words, to improve our regression model, it is necessary to further add a predictor variable to the linear regression model to best estimate the desired expected dependent variable. However, we understand adding variables is sensitive, and therefore, it is necessary to analyze and select variables that would only help us improve the model. Since adding extra variables into the model tends to make the model somewhat redundant and complex, it is important to maintain a certain level of efficiency and simplicity.

Appendix Part I

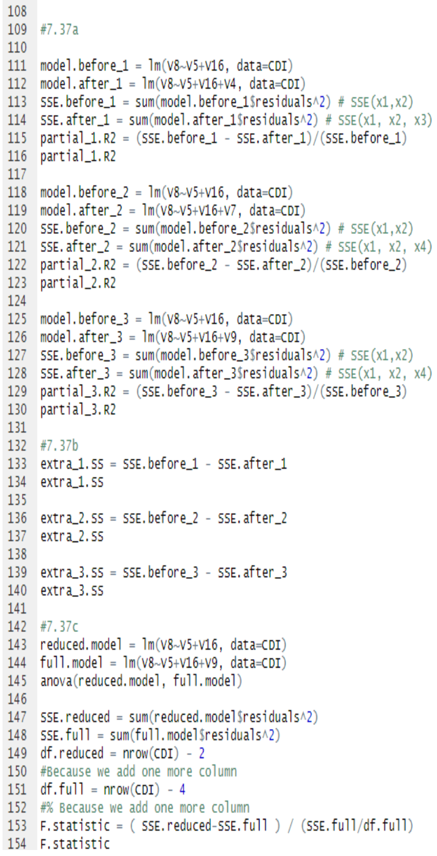


```
#6.28a
#Model 1
total_population = CDI$V5 #X1
stem(total_population)

land_area = CDI$V4 #X2
stem(land_area)

total_income = CDI$V16 #X3
stem(total_income)

#Model 2
population_density = total_population / land_area
CDI$V18 = population_density
stem(population_density)

elderly_percentage = CDI$V7
stem(elderly_percentage)

stem(total_income)

active_physician = CDI$V8

#6.28b
data1 = CDI[,c(8,5,4,16)]
colnames(data1) = c("Active Physicians", "Total Population", "Land Area", "Total Income")
pairs(data1, lower.panel = NULL)
cor(data1)

data2 = CDI[,c(8,18,7,16)]
colnames(data2) = c("Active Physicians", "Population Density", "Elderly Population", "Total Income")
pairs(data2, lower.panel = NULL)
cor(data2)

#6.28c
fit1 = lm(active_physician ~ total_population + land_area + total_income)
summary(fit1)

fit2 = lm(active_physician ~ population_density + elderly_percentage + total_income)
summary(fit2)

#6.28d
#summary 참고

#6.28e
#Model 1
residuals1 = fit1$residuals
yfit1 = fitted.values(fit1)

plot(yfit1, residuals1, xlab = "Fitted Y", ylab = "Resdual", main = "Residuals against Fitted Y" ) #residuals against fitted y-values
plot(total_population, residuals1, xlab ="Total Population", ylab = "Residuals", main = "Residuals against Total Population") #residuals against total population
plot(land_area, residuals1, xlab = "Land area (square miles)", ylab = "Residuals", main ="Residuals against Land Area") #residuals against land area
plot(total_income, residuals1, xlab = "Total Personal Income (in millions of dollars)", ylab = "Residuals", main = "Residuals against Total Personal Income") #residuals against total income
abline(h=0, col = 'red')

qqnorm(residuals1)
```

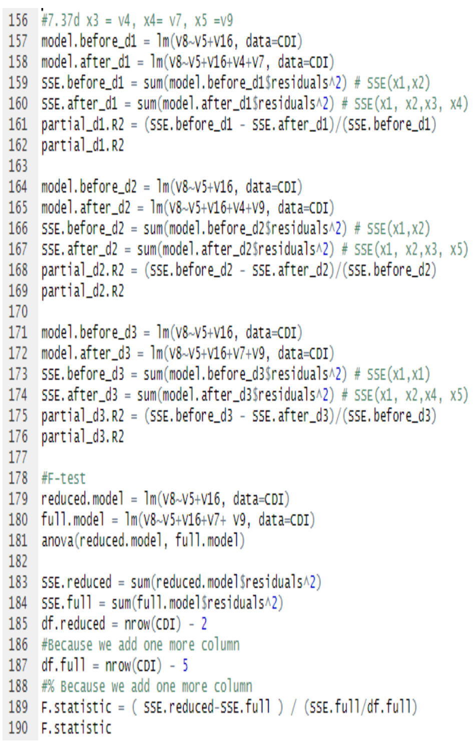


```
qqnorm(residuals1)
qqline(fit1$residuals, col ='red')

#Model 2
residuals2 = fit2$residuals
yfit2 = fitted.values(fit2)

plot(yfit2, residuals, xlab = "Fitted Y", ylab = "Residuals", main = "Residuals against Fitted Y") #residuals against fitted y-values
plot(population_density, residuals, xlab = "Population Density", ylab = "Residuals", main ="Residuals against Population Density") #residuals against population density
plot(elderly_percentage, residuals, xlab = "Percent of Population 65 or older", ylab = "Residuals", main = "Residuals against Elderly Percentage") #residuals against elderly percentage
plot(total_income, residuals, xlab = "Total Personal Income (in millions of dollars)", ylab = "Residuals", main = "Residuals against Total Personal Income") #residuals against total income
abline(h=0, col ='red')

qqnorm(residuals2)
qqline(fit2$residuals, col = 'red')

#6.28 f
#Model 1
full.model = lm(active_physician ~ total_population + land_area + total_income)
full.model_x1x2 = lm(active_physician ~ total_population + land_area + total_income + total_population:land_area)
full.model_x1x3 = lm(active_physician ~ total_population + land_area + total_income + total_population:total_income)
full.model_x2x3 = lm(active_physician ~ total_population + land_area + total_income + land_area:total_income)

plot(fitted.values(full.model_x1x2), full.model_x1x2$residuals, xlab = "Fitted Y", ylab = "Residuals", main = "Residuals against Fitted Y (X1X2)")
plot(fitted.values(full.model_x1x3), full.model_x1x3$residuals, xlab = "Fitted Y", ylab = "Residuals", main = "Residuals against Fitted Y (X1X3)")
plot(fitted.values(full.model_x2x3), full.model_x2x3$residuals, xlab = "Fitted Y", ylab = "Residuals", main = "Residuals against Fitted Y (X2X3)")

abline(h=0, col ='red')

summary(full.model_x1x2)
summary(full.model_x1x3)
summary(full.model_x2x3)

#Model 2
full.model = lm(V8~V18+V7+V16, data=CDI)
full.model_x1x2 = lm(V8~V18+V7+V16+V18:V7, data=CDI)
full.model_x2x3 = lm(V8~V18+V7+V16+V7:V16, data=CDI)
full.model_x1x3 = lm(V8~V18+V7+V16+V18:V16, data=CDI)

plot(fitted.values(full.model_x1x2), y=full.model_x1x2$residuals, xlab = "Fitted Y", ylab = "Residuals", main = "Residuals against Fitted Y (X1X2)")
abline(h=0, col='red')
summary(full.model_x1x2)

plot(fitted.values(full.model_x1x3), y=full.model_x1x3$residuals, xlab = "Fitted Y", ylab = "Residuals", main = "Residuals against Fitted Y (X1X3)")
abline(h=0, col='red')
summary(full.model_x1x3)

plot(fitted.values(full.model_x2x3), y=full.model_x2x3$residuals, xlab = "Fitted Y", ylab = "Residuals", main = "Residuals against Fitted Y (X2X3)")
abline(h=0, col='red')
summary(full.model_x2x3)
```

Appendix Part II



```
#7.37a

model.before_1 = lm(V8~V5+V16, data=CDI)
model.after_1 = lm(V8~V5+V16+V4, data=CDI)
SSE.before_1 = sum(model.before_1$residuals^2) # SSE(x1,x2)
SSE.after_1 = sum(model.after_1$residuals^2) # SSE(x1, x2, x3)
partial_1.R2 = (SSE.before_1 - SSE.after_1)/(SSE.before_1)
partial_1.R2

model.before_2 = lm(V8~V5+V16, data=CDI)
model.after_2 = lm(V8~V5+V16+V7, data=CDI)
SSE.before_2 = sum(model.before_2$residuals^2) # SSE(x1,x2)
SSE.after_2 = sum(model.after_2$residuals^2) # SSE(x1, x2, x4)
partial_2.R2 = (SSE.before_2 - SSE.after_2)/(SSE.before_2)
partial_2.R2

model.before_3 = lm(V8~V5+V16, data=CDI)
model.after_3 = lm(V8~V5+V16+V9, data=CDI)
SSE.before_3 = sum(model.before_3$residuals^2) # SSE(x1,x2)
SSE.after_3 = sum(model.after_3$residuals^2) # SSE(x1, x2, x4)
partial_3.R2 = (SSE.before_3 - SSE.after_3)/(SSE.before_3)
partial_3.R2

#7.37b
extra_1.SS = SSE.before_1 - SSE.after_1
extra_1.SS

extra_2.SS = SSE.before_2 - SSE.after_2
extra_2.SS

extra_3.SS = SSE.before_3 - SSE.after_3
extra_3.SS

#7.37c
reduced.model = lm(V8~V5+V16, data=CDI)
full.model = lm(V8~V5+V16+V9, data=CDI)
anova(reduced.model, full.model)

SSE.reduced = sum(reduced.model$residuals^2)
SSE.full = sum(full.model$residuals^2)
df.reduced = nrow(CDI) - 2
#Because we add one more column
df.full = nrow(CDI) - 4
#% Because we add one more column
F.statistic = ( SSE.reduced-SSE.full ) / (SSE.full/df.full)
F.statistic
```



```
#7.37d x3 = v4, x4= v7, x5 =v9
model.before_d1 = lm(V8~V5+V16, data=CDI)
model.after_d1 = lm(V8~V5+V16+V4+V7, data=CDI)
SSE.before_d1 = sum(model.before_d1$residuals^2) # SSE(x1,x2)
SSE.after_d1 = sum(model.after_d1$residuals^2) # SSE(x1, x2,x3, x4)
partial_d1.R2 = (SSE.before_d1 - SSE.after_d1)/(SSE.before_d1)
partial_d1.R2

model.before_d2 = lm(V8~V5+V16, data=CDI)
model.after_d2 = lm(V8~V5+V16+V4+V9, data=CDI)
SSE.before_d2 = sum(model.before_d2$residuals^2) # SSE(x1,x2)
SSE.after_d2 = sum(model.after_d2$residuals^2) # SSE(x1, x2,x3, x5)
partial_d2.R2 = (SSE.before_d2 - SSE.after_d2)/(SSE.before_d2)
partial_d2.R2

model.before_d3 = lm(V8~V5+V16, data=CDI)
model.after_d3 = lm(V8~V5+V16+V7+V9, data=CDI)
SSE.before_d3 = sum(model.before_d3$residuals^2) # SSE(x1,x1)
SSE.after_d3 = sum(model.after_d3$residuals^2) # SSE(x1, x2,x4, x5)
partial_d3.R2 = (SSE.before_d3 - SSE.after_d3)/(SSE.before_d3)
partial_d3.R2

#F-test
reduced.model = lm(V8~V5+V16, data=CDI)
full.model = lm(V8~V5+V16+V7+ V9, data=CDI)
anova(reduced.model, full.model)

SSE.reduced = sum(reduced.model$residuals^2)
SSE.full = sum(full.model$residuals^2)
df.reduced = nrow(CDI) - 2
#Because we add one more column
df.full = nrow(CDI) - 5
#% Because we add one more column
F.statistic = ( SSE.reduced-SSE.full ) / (SSE.full/df.full)
F.statistic
```