STA 108 Project 2

Abstract:

In project 1, we first used regression model (1.1) as a test of good fit for the relationships between the number of active physicians and the 3 predictor variables. However, upon further inspection through regression diagnostics on residuals, it can be seen that the linear regression model (2.1) would be a better fit for 3 relationships. We came to this conclusion because despite the residuals showing randomness, did not seem to follow a Normal Distribution in the Normal Q-Q plot. We found that all 3 predictor variables were not independent of Y. We then ran further tests and found that in each of the 4 regions, per capita income was significantly joint and dependent on the percentage of those with at least a bachelor's degree.

The CDI data set has been collected from 440 different areas split into 4 regions, with each area having independent data surrounding these variables. In project 2, we will firstly analyze using the first-order regression model (6.5) to determine out of 2 separate models, which would be the best to predict the number of active physicians. The two models being: Proposed model I includes as predictor variables total population (X_1), land area (X_2), and total personal income (X_3). Proposed model II includes as predictor variables population density (X_1, total population divided by land area), percent of population greater 64 years old (X_2), and total personal income (X_3). Through ANOVA, F-tests, coefficient of determination and residual analysis we will conclude which of the two models is the more favourable one.

The second part will be to conclude whether adding certain predictor variables to an established model will be helpful to the model or not. The model is to predict the number of active physicians, and already contains the variables total population (X_1) and total personal income (X_2). Having assumed that a first-order multiple regression model is appropriate, we will determine which predictor and/or pairs of predictor variables will be most helpful to add to the model. We will come to this conclusion through analysis based on the coefficient of partial determination, ANOVA and F-tests.

<u>Part I:</u> Multiple Linear Regression I (Project 6.28)

a) Stem and Leaf Plots

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

0 | 11111111111111111111111111111111111111111111111111111111111111111+254
0 | 55555555555555555555555666666666666667777777777777777777778888888888
1 | 000000122233333444
1 | 55699
2 | 1134
2 | 58
3 |
3 |
4 |
4 |
5 | 1
5 |
6 |
6 |
7 |
7 |
8 |
8 | 9
```

This is the plot for the total population. The left-Hand side is in millions. The majority of the data is in areas with a population of less than a million. The data for this predictor variable is heavily right-skewed.

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

 0 | 000011111111111122222222222222222223333333333333333333333333444444+252
 1 | 0000000000000001111111111111122222222222333333444455566667777888999 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
```

This plot is for the area of land for each place. The left-hand side is in thousands of square miles. The majority of the data is from areas under 1000 square miles. The data for this predictor variable is also right-skewed.

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

 0 | 1111111111111222222222222222222222222222222222222222222222222222222222+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
```

This plot is for the total personal income. The left-hand side is in 10000's. The majority of the data has income under $10,000. The data for this predictor variable is right-skewed.

```
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
```

This plot is for population density. A predictor variable was made from the total population/total land area for each area. The left-hand side is in thousands of people per square mile. The majority of the data has under 2000 people per square mile. The data for this predictor variable is right-skewed.

```
The decimal point is at the |

 2 | 0
 4 | 47890389
 6 | 112345567799O134566678899
 8 | 00112222233334444555666777778888899990002222333334444444455556666777
10 | 000111111222222222233333344444455555556666666677777778888888888999999+36
12 | 000000001111122223333333333444455555556666667777777778888999000000000+36
14 | 00001111111223334444455567788900000011112222345566777 8
16 | 12556699901122345
18 | 06778
20 | 070
22 | 018828
24 | 47
26 | 055
28 | 1
30 | 7
32 | 138
```

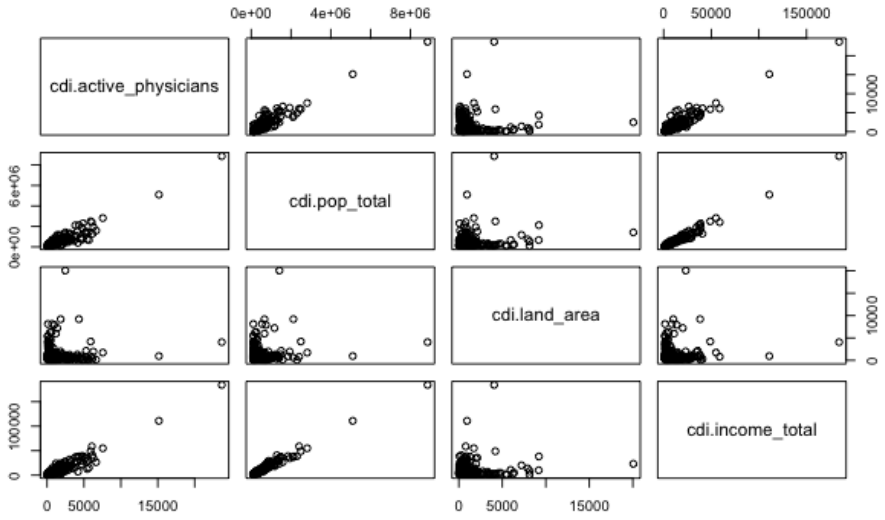
The plot is for the percentage of the population greater than 64 years old. The left-hand side is in percentages, with the right-hand side being the number on the other side of the decimal point. This data set is the most symmetrical of all the plots but is still right-skewed. It has the most bell-shaped data but the majority of the data is still on the lower side.

Overall, the first 4 plots at least have one huge outlier affecting the data set. All the data sets are right-skewed meaning that they have a mode smaller than the mean.

b) Scatter Plot and Correlation Matrix

Below are the Scatter Plot and Correlation Matrices for Model 1. Regression Model 1 is that of the total number of active physicians against the Total Population, Total Land Area and Total Personal Income.

	cdi.active_physicians	cdi.pop_total	cdi.land_area	cdi.income_total
cdi.active_physicians	1.00000000	0.9402486	0.07807466	0.9481106
cdi.pop_total	0.94024859	1.0000000	0.17308335	0.9867476
cdi.land_area	0.07807466	0.1730834	1.00000000	0.1270743
cdi.income_total	0.94811057	0.9867476	0.12707426	1.0000000

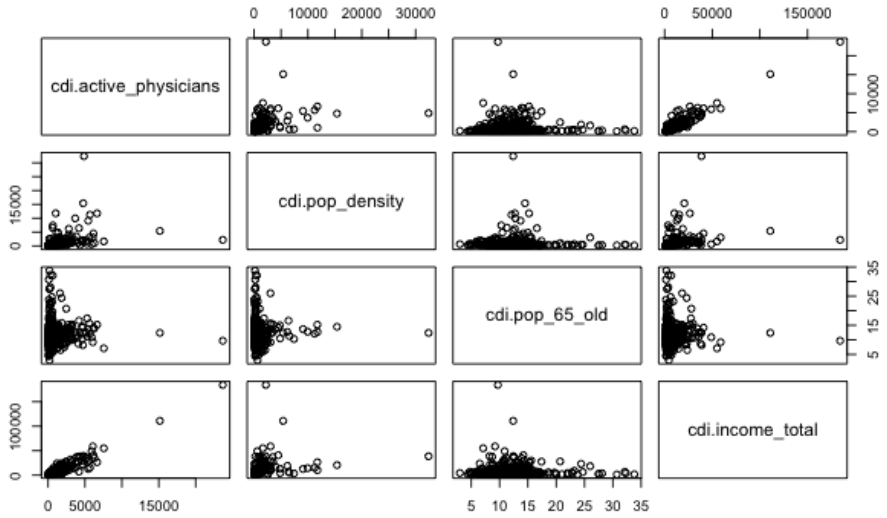


From these matrices, there is a very strong positive correlation (roughly 0.94) between the total number of active physicians and the population total as well as between the total number of active physicians and the total personal income. In regards to the correlation between the predictor variables, total population and total income have a very strong positive correlation of nearly 0.99. Total Land Area, on the other hand, has a low correlation (below 0.2) with the other two predictor variables and an even lower correlation (below 0.1) with the total number of active physicians.

This means that the two predictor variables of the total population and total income are strongly related both to the total number of active physicians and to each other. Whereas the variable of Total Land Area is hardly related to any of the 3.

Below are the Scatter Plot and Correlation Matrices for Model 2. Regression Model 2 is that of the Total number of Active Physicians against the Population Density, Percentage of Population over 64 years old and Total Personal Income.

	cdi.active_physicians	cdi.pop_density	cdi.pop_65_old	cdi.income_total
cdi.active_physicians	1.00000000	0.40643863	-0.00312863	0.94811057
cdi.pop_density	0.40643863	1.00000000	0.02918445	0.31620475
cdi.pop_65_old	-0.00312863	0.02918445	1.00000000	-0.02273315
cdi.income_total	0.94811057	0.31620475	-0.02273315	1.00000000



From these matrices, there is a very strong positive correlation (roughly 0.94) between the total number of active physicians and the total personal income. There is a relatively low correlation (around 0.4) between population density and the total number of active physicians. There is a really small negative correlation (around 0) between the total number of active physicians and the Percentage of the Population over 64 years old. In regards to the correlation between the predictor variables, all three have relatively low correlations with each other. Population Density and Total Personal Income have a 0.31 & Population Density and the Percentage of the Population over 64 years old have a 0.03. Total Personal Income and the Percentage of the Population over 64 years old also have a small negative correlation of -.0.03.

This means there is a strong relation between the total number of active physicians and total personal income. Whereas there is a range from low to no relation between the rest of the variable predictors and the total number of active physicians. There also is little to no relation between the three predictor variables. A negative correlation, despite it being small, means that as one variable increases the other decreases.

c) Fitting the first-order Regression Model (6.5)

Having assumed that the first-order Regression Model(6.5) is appropriate, we fitted the following models.

Model 1:

Active Physicians (Y) regressed against the Total Population (X_1), Total Land Area (X_2) and Total Personal Income(X_3).

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

Model 2:

Active Physicians (Y) regressed against the Population Density (X_1), the Percentage of the Population over 64 years old (X_2) and Total Personal Income (X_3).

$\hat{Y} = -170.57 + 0.09616X_1 + 6.33984X_2 + 0.12657X_3$

d) Coefficient of determination (R^2)

R^2 is the Coefficient of Determination. It measures how well a statistical model predicts an outcome. It is the proportion of the variance for a dependent variable that's explained by an independent variable in a regression model. To calculate the R^2 criterion, the summary function in RStudio was used.

Our results for the Coefficient of Determination of the two models are as follows:

Model 1: 0.9026432

Model 2: 0.9117491

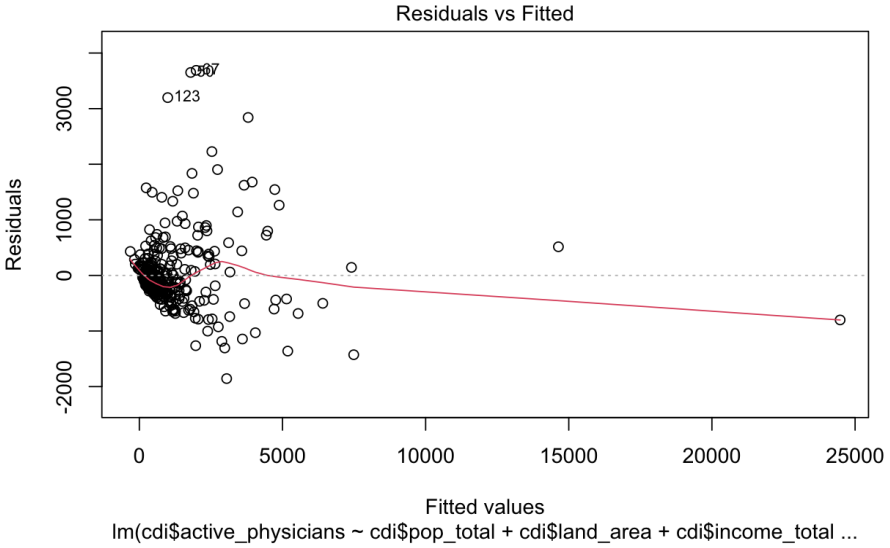
The higher the value of R^2, the better the model. The two R^2 values are very close however that doesn't mean that one isn't better than the other. In this case, the results show that Model 2 accounts for the largest reduction in the variability in the number of active physicians.

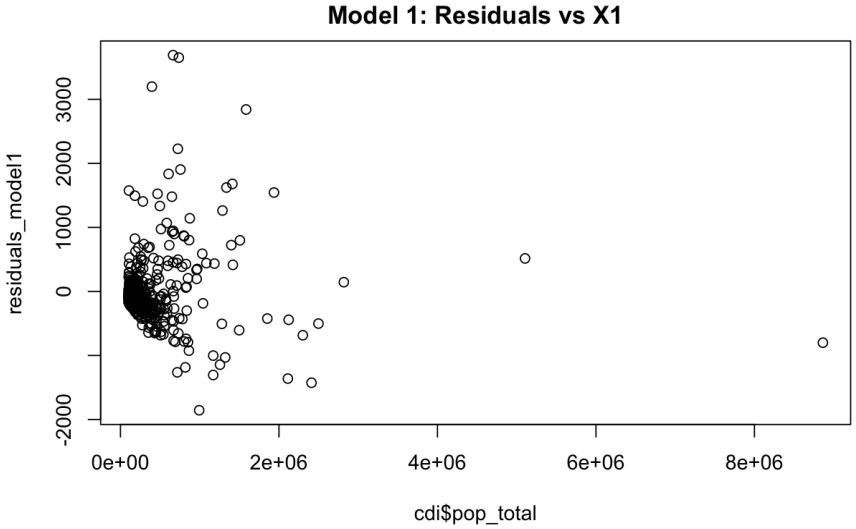
e) Residual Plots and Regression Diagnostics

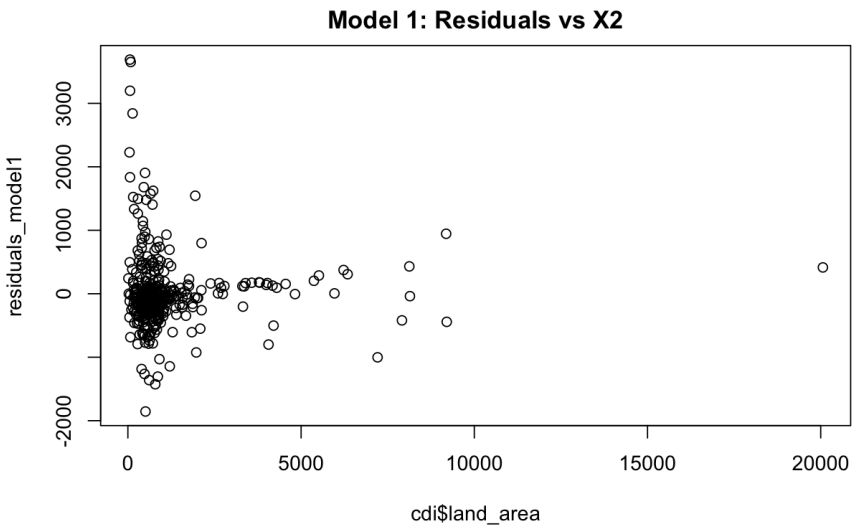
For both regression models, we can plot the residuals against the fitted values, each of the three predictor variables and each of the two-factor interaction terms in a series of scatter plots. Then we can create a normal probability plot for both models to determine if the assumptions surrounding the regression models are true or not. This will allow us to determine which model is preferable in terms of appropriateness.

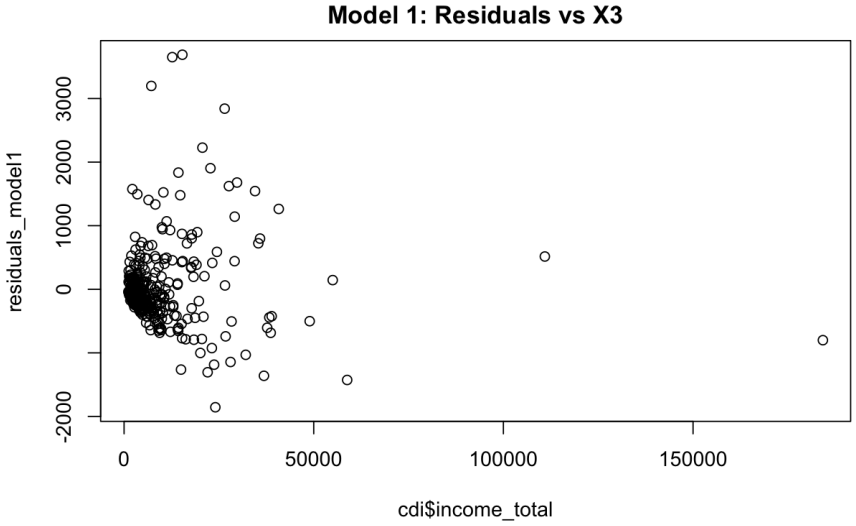
We use these scatterplots to determine whether there is a distinct randomness to the residuals when compared to each predictor and interaction term. This is to check whether they follow the assumption of independence.

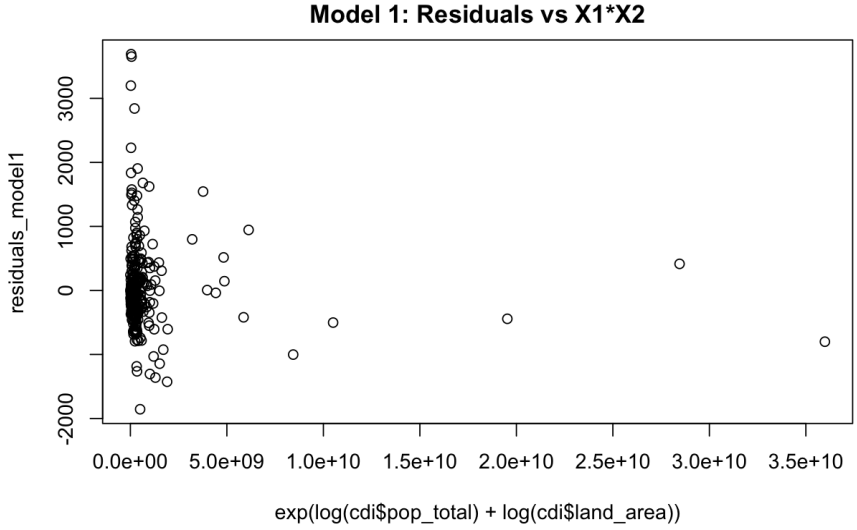
The following are the residual plots of Model 1:

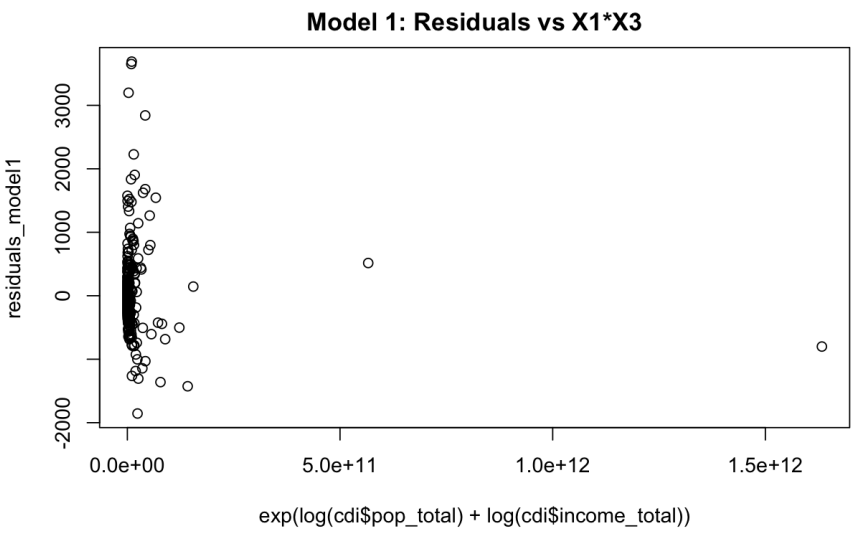


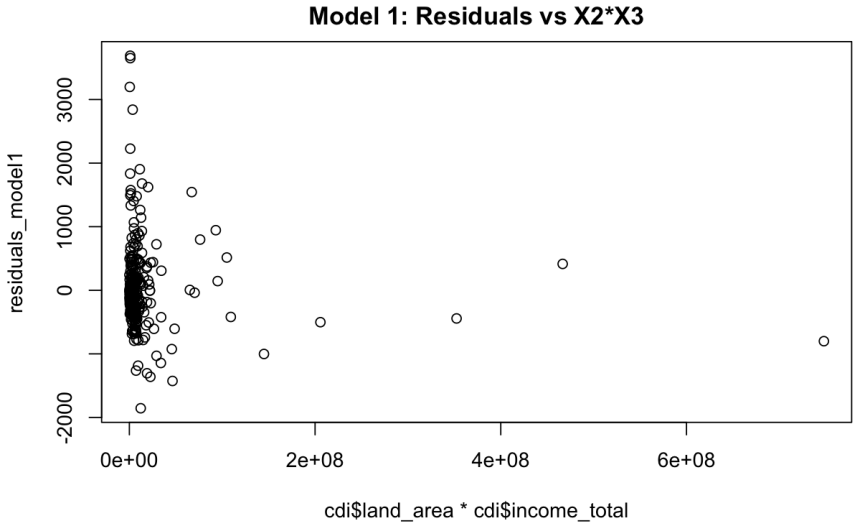



Model 1: Residuals vs X2
residuals_model1
-2000
0
1000
2000
3000
0
5000
10000
15000
20000
cdi$land_area


Model 1: Residuals vs X3
residuals_model1
-2000
0
1000
2000
3000
0
50000
100000
150000
cdi$income_total

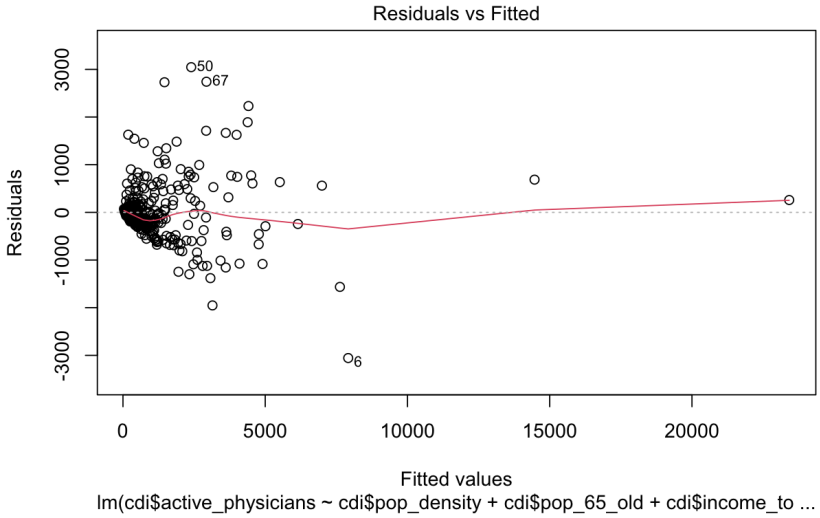

Model 1: Residuals vs X1*X2
residuals_model1
-2000
0
1000
2000
3000
0.0e+00
5.0e+09
1.0e+10
1.5e+10
2.0e+10
2.5e+10
3.0e+10
3.5e+10
exp(log(cdi$pop_total) + log(cdi$land_area))

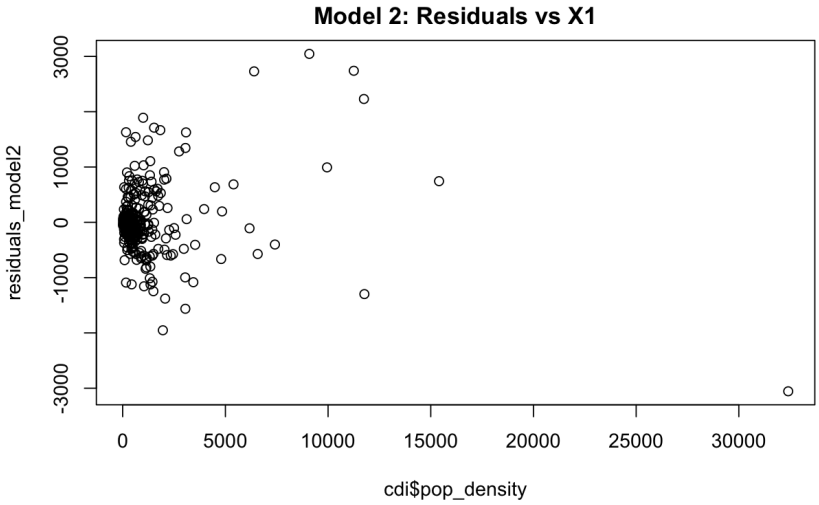


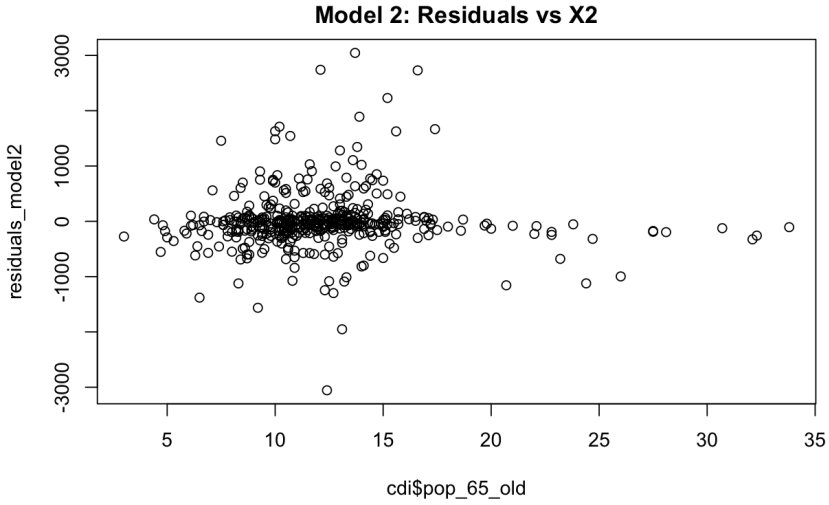


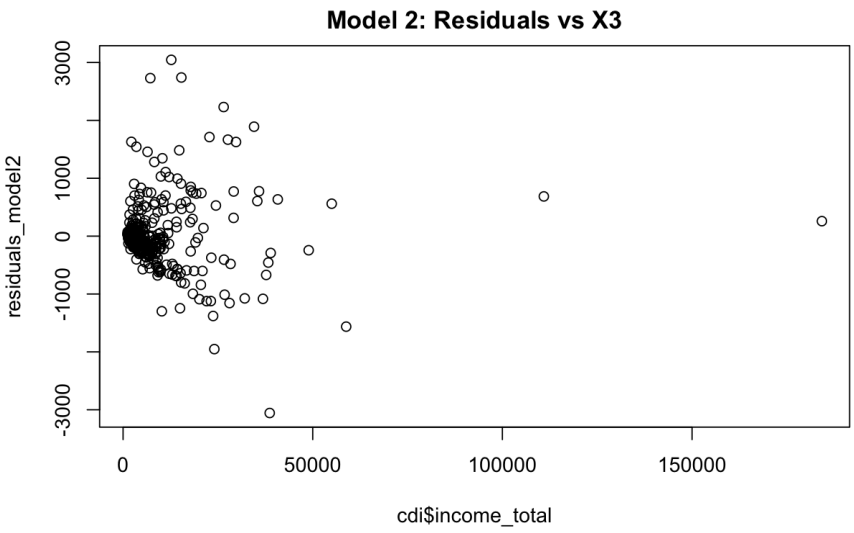
Upon reviewing this first set of plots, there is a clear aspect of randomness. There does not seem to be a pattern in any of them, therefore confirming the individuality assumption and that they do not affect each other. When looking at the residuals vs fitted plot, it can be seen that there is an unequal variance meaning that there is heteroscedasticity, which can affect the Type I error rate and also lead to incorrect conclusions in future F-tests.

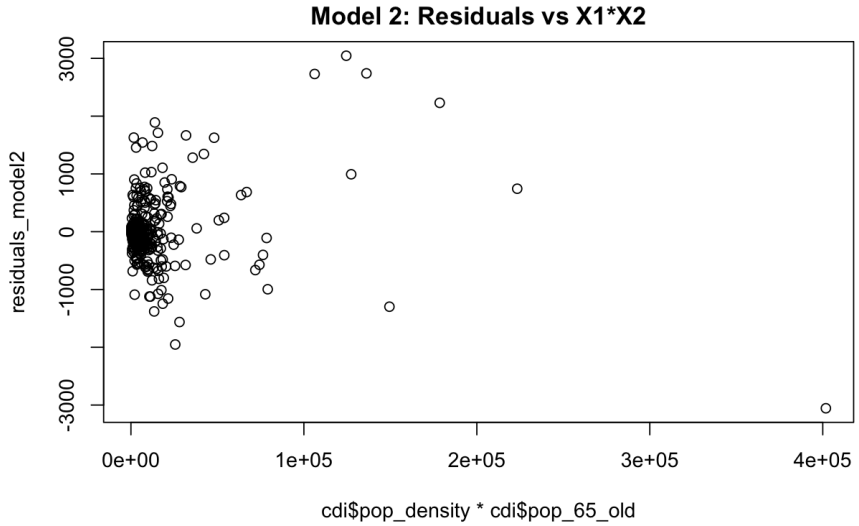
The following are the residual plots of Model 2:

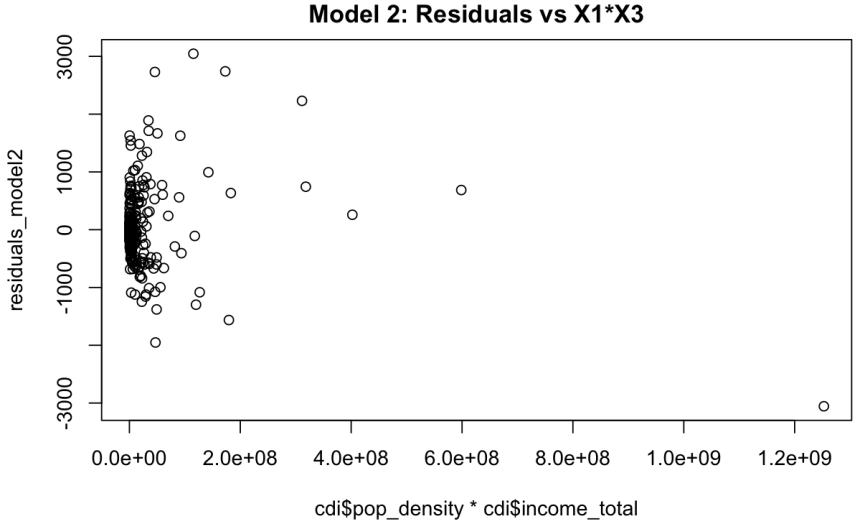


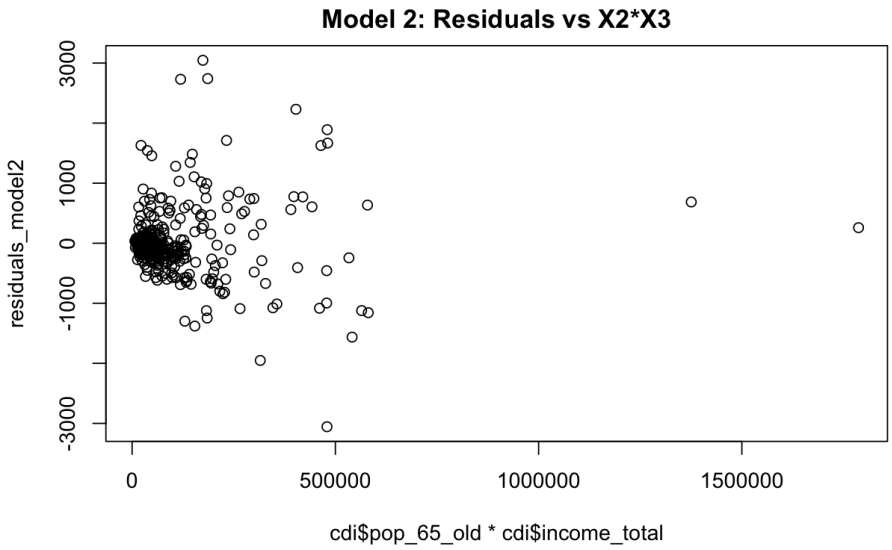





Model 2: Residuals vs X3
residuals_model2
3000
1000
0
-1000
-3000
0
50000
100000
150000
cdi$income_total


Model 2: Residuals vs X1*X2
residuals_model2
3000
1000
0
-1000
-3000
0e+00
1e+05
2e+05
3e+05
4e+05
cdi$pop_density * cdi$pop_65_old

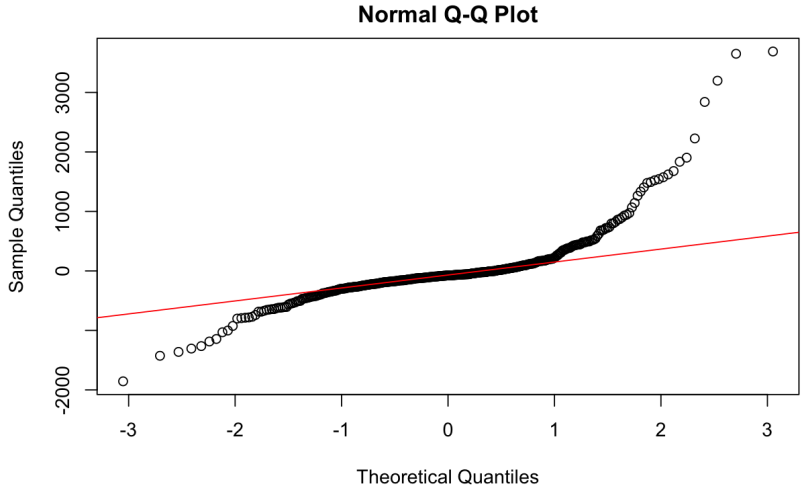

Model 2: Residuals vs X1*X3
residuals_model2
3000
1000
0
-1000
-3000
0.0e+00
2.0e+08
4.0e+08
6.0e+08
8.0e+08
1.0e+09
1.2e+09
cdi$pop_density * cdi$income_total



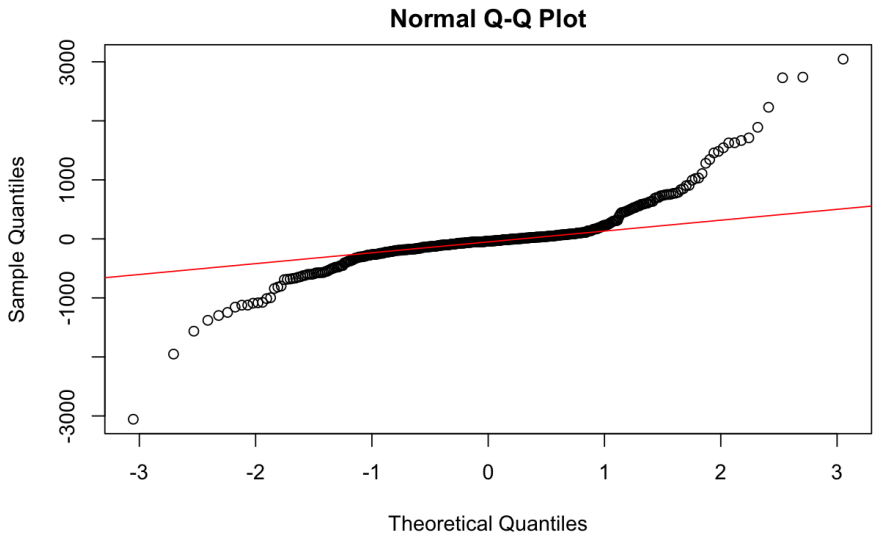
When reviewing this set of plots, most of the plots do display a strong degree of randomness. The case can be made that the plot of residuals vs X2 does not display randomness and possibly could be seen as a pattern, however upon further investigation, the residuals are still relatively random. In regards to the residuals vs fitted values plot, it can be seen that despite it also displaying heteroscedasticity, the variances are closer for the second model.

Keeping these findings in mind, we now observe the Normal Probability plots of the two models. The objective is to assess whether or not the residuals follow the assumption that they follow a normal distribution.

Model 1 QQ plot:



Model 2 QQ plot:



In these plots, we see that both residual normal probability plot does not actually follow a continually straight line. The main issue that can be deducted from all 3 plots is the fact that there are far outliers that are completely off the expected line. When comparing the two, despite the outliers in the plot for model 2 being more extreme, it can actually be seen that it is the more 'normal' plot. This is because the points as a whole are closer to the line than in the plot for model 1. Both of these, therefore, do not follow a normal distribution meaning a different linear regression model could be used.

Based on these findings, it can be concluded that Model 2 is preferable in terms of appropriateness. This is because the residuals in both models break the univariance and normality assumptions, but the regression errors in Model 2 are closer to following these assumptions than those in Model 1.

f) Expansion of Models

We are now adding all possible two-factor interactions to both models.
These being X_1X_2 , X_1X_3, X_2X_3. After doing this, we will then examine the R^2
(Coefficient of Determination) of the two new models and see which model is preferable in terms of this measure.

Expanded Model 1:

Active Physicians (Y) regressed against the Total Population (X_1), Total Land Area (X_2), Total Personal Income (X_3), Total Population * Total Land Area (X_1X_2),
Total Population*Total Personal Income (X_1X_3) and Total Land Area*Total Personal Income(X_2X_3)

$$\hat{Y} = -58.26 + 0.0007252X_1 - 0.06421X_2 + 0.1087X_3 + 0.0000006173X_1X_2 + 1.696x10^{-9}X_1X_3 + 0.00003706X_2X_3$$

Expanded Model 2:

Active Physicians (Y) regressed against the Population Density (X_1), the Percentage of the Population over 64 years old (X_2), Total Personal Income (X_3), Population Density * Percentage of the Population over 64 years old (X_1X_2), Population Density * Total Personal Income (X_1X_3) and Percentage of the Population over 64 years old * Total Personal Income (X_2X_3).

$$\hat{Y} = -9.367 - 0.4179X_1 - 11.06X_2 + 0.1477X_3 + 0.04652X_1X_2 - 0.000003276X_1X_3 - 0.001289X_2X_3$$

Using the summary(model)$r.squared function in r, we now find the Coefficients of Determination for each model.

Expanded Model 1: 0.9063789

Expanded Model 2: 0.9230238

Therefore based on this measure, Expanded Model 2 has the larger Coefficient of Determination, meaning that it is the better model. Expanded Model 2 accounts for the largest reduction in the variability in the number of active physicians. Meaning that before and after expansion, model 2 is the better model in relation to the measure of the Coefficient of Determination.

<u>Part II:</u> Multiple Linear Regression II (Project 7.37)

For this part: X_1 is total population, X_2 is total personal income, X_3 is land area, X_4 is the percentage of the population 65 or older and X_5 is the number of hospital beds.

a) Coefficient of Partial Determination

Formula:

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{SSE(X_1, X_2)}$$

A coefficient of partial determination, in contrast, measures the marginal contribution of one X variable when all others are already included in the model.

We are adding a predictor variable to the regression model with the number of active physicians against the total population and total personal income. We are trying to determine which third predictor variable would be the best predictor would be the most helpful. The three predictor variables we are testing to choose are land area, percentage of the population older than 64 and the number of hospital beds. This is all assuming that a first-order multiple regression model is appropriate.

Coefficient of partial determination:

X_3 = 4063370/140967081 = 0.02882496

X_4 = 541647.3/140967081 = 0.003842367

X_5 = 78070132/140967081 = 0.5538182

b) Inferences about ANOVA

Based on part a), it can be seen that X_5 has the largest coefficient of partial determination. The larger the coefficient, the better the fit that predictor variable is for the model. That means that the number of hospital beds is the best predictor variable to add to the model.

Using the ANOVA function in RStudio, we then are able to check the Extra Sum of Squares associated with each predictor variable.

Formula: ESS = Residual sum of squares (reduced) – Residual Sum of Squares (full). For example, your model contains one predictor variable, X_1. If you add a second predictor, X_2 to the model, ESS explains the additional variation explained by X_2.

ESS for all 3 predictor variables:

Land Area (X_3) = 4063370

Population older than 64 (X_4) = 541647.3

Hospital Beds (X_5) = 78070132

From these results, we can see that the total number of hospital beds has the largest ESS. The extra sum of squares values for the other variables are far smaller than the value of the number of Hospital beds.

c) Statistical Inference

We will now complete an F test to determine whether or not the predictor variable, the number of Hospital Beds, is helpful in the regression model when X_1 and X_2 are included in the model. We want to see if X_3 is needed in this F-test.

$H_0 : \beta_3 = 0$

$H_a : \beta_3 \neq 0$

Using the alpha = 0.01, we can calculate the critical F value that we compare our F* test statistic. We compute F values in R through the function qf(1-α, df_R - df_F, df_F).

The statistic is

$$F = \frac{MSR(X_3, X_4|X_1, X_2)}{MSE(X_1, X_2, X_3, X_4)} = \frac{SSR(X_3, X_4|X_1, X_2)/2}{SSE(X_1, X_2, X_3, X_4)/435} = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F}$$

Or can be found with F = MSR/MSE

For this F test our F value is qf(1-0.01, 437-436, 436) = 6.693

The F* test statistic that we get is 78070132/144259 = 541.1801

Since 541.1801>6.693 (F*stat > F critical), we, therefore, can reject the null hypothesis (H_0) and can conclude that the predictor variable, X_3, (the number of Hospital Beds) is helpful in the regression model when X_1 and X_2 are included in the model.

The F* test statistics for the other two potential predictor variables would not be as large because the MSR values for the other models are significantly smaller and simultaneously their MSE values are also larger. This in turn would lead to far smaller F* values.

d) Inference on Pairs of Predictors

We will now compute more coefficients of partial determination and follow up with F tests to find which pairs of predictors are the best and whether they should be added to the model. The three pairs of Predictors are $R^2_{Y,X_3,X_4 \mid X_1,X_2}$, $R^2_{Y,X_3,X_5 \mid X_1,X_2}$, and $R^2_{Y,X_4,X_5 \mid X_1,X_2}$.

The coefficients of partial determination are as follows:

$R^2_{Y,X_3,X_4 \mid X_1,X_2} = 0.0331418$

$R^2_{Y,X_3,X_5 \mid X_1,X_2} = 0.5558232$

$R^2_{Y,X_4,X_5 \mid X_1,X_2} = 0.5642756$

From this, that all of these coefficients are not high meaning that there is not a huge amount of contribution that these pairs of predictors would add to the model. The pair that has the highest coefficient is that of the percentage of the population 65 or older and the number of hospital beds.

Now we run further F-tests to find out whether adding the best pair to the model is helpful given that X_1, X_2 are already included at a 0.01 alpha value.

$H_0 : \beta_3\beta_4 = 0$

$H_a : \beta_3\beta_4 \neq 0$

The F critical value for this test is: F(1-0.01, 2, 435) = 4.654269
Using the R Anova function between reduced and full models, we were able to find that the F* statistical value = 281.67

281.67 > 4.654, (F*stat > F critical) meaning that we can reject the null hypothesis and can conclude that the pair of predictor variables, X_3X_4 (percentage of population 65 or older and the number of Hospital Beds) is helpful in a regression model that X_1 and X_2 are included in.

Part III: Discussion

In the review of the data, we found that most of the data in the CDI dataset are right-skewed with only one major outlier. In the two models, the total number of active physicians had the strongest correlation with the population total and the total personal income. This means that these were the two predictor variables that we checked that were the most important in the models. We then tried fitting three predictor variables into the first-order regression model (6.5) for each proposed model and based on the R^2 values, we found that model 2 was better than model 1. Upon inspecting the residuals, they were all consistent in their randomness but did not seem to follow a Normal Distribution in the Normal Q-Q plot.

In the second part of the project, we were trying to find the best predictor variable to fit into a model which already had two other variables in it. Through multiple layers of testing and inference, we found that the number of hospital beds was the best one due to it having the largest coefficient of partial determination as well as passing the F-test which confirmed its significance if it was added to the model. We then tested which pair of predictors would be the best to add to the model, in which we found that it was the percentage of the population 65 or older and the number of Hospital Beds. This also passed the F-test for significance if it was added to the model.

The most relevant parts of the course in relation to analysis for the project would be the comparison of the reduced vs full model. This allows you to be able to see if adding a new predictor variable would be beneficial and helpful for the old model through an F test. Another relevant part would be the Coefficient of Partial and normal Determination which literally are coefficients which measure how much a new predictor will contribute to a model and how well the data fits the model respectively. These literally tell us how well the model we created fits and if another variable should be added.

In order to improve the linear regression models, a possible way to improve them would be to increase the degrees of freedom. In general, this would be through increasing the sample size and in turn mean there is more power to reject a false null hypothesis and find a significant result. To create normality in the data, we possibly could perform power transformations such as taking the square root and the logarithm of the observations. Using the logarithm of observations could also help with the issue of heteroscedasticity. I think a variable that we could've checked would be the percentage of people with a bachelor's degree.

Appendix:

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
library(dplyr)
getwd()
setwd("/Users/owen/Downloads/")
cdi <- read.table("CDI.txt")
colnames(cdi) <- c("id_num", "county",
                   "state", "land_area",
                   "pop_total", "pop_18_34",
                   "pop_65_old", "active_physicians",
                   "hospital_beds", "serious_crimes",
                   "pct_hsgrad", "pct_bachelors",
                   "pct_poverty", "pct_unemp",
                   "income_percap", "income_total",
                   "region")
cdi <- cdi %>%
  group_by(pop_total, land_area) %>%
  mutate(pop_density = pop_total/land_area)
  head(cdi)



#part a
stem(cdi$pop_total)
stem(cdi$land_area)
stem(cdi$income_total)
stem(cdi$pop_density)
stem(cdi$pop_65_old)


#part b
cdi_model1 <- data.frame(cdi$active_physicians,cdi$pop_total, cdi$land_area,cdi$income_total)
cdi_model2 <-data.frame(cdi$active_physicians, cdi$pop_density, cdi$pop_65_old, cdi$income_total)
pairs(cdi_model1)
cor(cdi_model1)
pairs(cdi_model2)
cor(cdi_model2)
#part c

model1 <- lm(cdi$active_physicians ~ cdi$pop_total + cdi$land_area + cdi$income_total, data = cdi)
model2 <- lm(cdi$active_physicians ~ cdi$pop_density + cdi$pop_65_old + cdi$income_total, data = cdi)
#part d
```

```r
summary(model1)$r.squared
summary(model2)$r.squared
#part e
residuals_model1 <- residuals(model1)
residuals_model2 <- residuals(model2)
y_hat = model1$fitted.values
plot(x=y_hat, y=residuals_model1, xlab='fitted values', ylab='residuals')
abline(h=0, col='red')
y_hat2 = model2$fitted.values
plot(x=y_hat2, y=residuals_model2, xlab='fitted values', ylab='residuals')
abline(h=0, col='red')


#rm1
plot(cdi$pop_total,residuals_model1,
     main = "Model 1: Residuals vs X1")
plot(cdi$land_area,residuals_model1,
     main = "Model 1: Residuals vs X2")
plot(cdi$income_total,residuals_model1,
     main = "Model 1: Residuals vs X3")

#rm2
plot(cdi$pop_density,residuals_model2,
     main = "Model 2: Residuals vs X1")
plot(cdi$pop_65_old,residuals_model2,
     main = "Model 2: Residuals vs X2")
plot(cdi$income_total,residuals_model2,
     main = "Model 2: Residuals vs X3")

#rm1
residuals_model1 <- residuals(model1)
residuals_model2 <- residuals(model2)
plot(exp(log(cdi$pop_total)+log(cdi$land_area)), residuals_model1,
main = "Model 1: Residuals vs X1*X2")
plot(exp(log(cdi$pop_total)+log(cdi$income_total)), residuals_model1,
main = "Model 1: Residuals vs X1*X3")
plot(cdi$land_area*cdi$income_total, residuals_model1,
     main = "Model 1: Residuals vs X2*X3")

#rm2
plot(cdi$pop_density*cdi$pop_65_old, residuals_model2,
     main = "Model 2: Residuals vs X1*X2")
plot(cdi$pop_density*cdi$income_total, residuals_model2,
      main = "Model 2: Residuals vs X1*X3")
plot(cdi$pop_65_old*cdi$income_total, residuals_model2,
      main = "Model 2: Residuals vs X2*X3")

#rm1
qqnorm(residuals(model1))
qqline(residuals(model1), col = "red")

#rm2
qqnorm(residuals(model2))
```

```r
qqline(residuals(model2), col = "red")
#part f

model3 <- lm(active_physicians ~ pop_total + land_area + income_total + pop_total*land_area + pop_tota
model4 <- lm(active_physicians ~ pop_density + pop_65_old + income_total + pop_density*pop_65_old + po

summary(model3)$r.squared
summary(model4)$r.squared
#part a
reduced=lm(active_physicians~pop_total+income_total, data=cdi)
full= lm(active_physicians~pop_total+income_total+land_area, data=cdi)
#SSR of land area= 4063370
r2landarea= 4063370/140967081

knitr::kable(anova(reduced))
knitr::kable(anova(reduced,full))

full1= lm(active_physicians~pop_total+income_total+pop_65_old, data=cdi)
knitr::kable(anova(reduced,full1))
#SSR of pop_65_old = 541647.3
r2pop65=  541647.3/140967081

full2= lm(active_physicians~pop_total+income_total+hospital_beds, data=cdi)
knitr::kable(anova(reduced,full2))
#SSR of hospital beds = 78070132
r2hospitalbeds =78070132/140967081

r2landarea
r2pop65
r2hospitalbeds
#part b
anova(reduced)
anova(full)
anova(full1)
anova(full2)

#part c
a= 78070132/144259
a
Fstat=541.1801
qf(1-0.01,1,436)
#part d
reduced=lm(active_physicians~pop_total+income_total, data=cdi)
full = lm(active_physicians~pop_total+income_total + land_area +pop_65_old, data=cdi)
full1 = lm(active_physicians~pop_total+income_total + land_area + hospital_beds , data=cdi)
full2 = lm(active_physicians~pop_total+income_total + pop_65_old + hospital_beds , data=cdi)
anova(reduced,full)
anova(reduced,full1)
anova(reduced,full2)
r2_x3x4=4671904/140967081
r2_x3x5= 78352775/140967081
r2_x4x5 = 79544288/140967081
r2_x3x4
```

```r
r2_x3x5
r2_x4x5

f_x3x4=7.4554
f_x3x5=272.17
f_x4x5=281.67
fcrit_1= qf(1-0.01,2,435)
fcrit_1
```