Project 2

Tomer Fidelman, Meghana Vemula, Quilvio Hernandez

5/29/2019

Background and Introduction

The CDI is a data set that provides county demographic information for 440 of the most populous counties in the US. The primary objective of the project was to identify if there was a linear relationship between several significant attributes of US continental counties. Initially, the project team was interested in determining if there were linear relationships present between the number of active physicians in a county and 3 specific predictor variables: total population, number of hospital beds, and total personal income. For this part, the team formed linear regression models for each predictor variable. Later, the R2 goodness of fit values were compared between the three predictor values to determine which factor accounted for the largest variability reduction of the number of physicians. Next, the research team analyzed per capita income of the counties against the predictor variable containing the percentage of bachelor degree holding county residents. The researchers analyzed regression models between the four continental regions of the US for similarity. They cross compared error values for variance and regression coefficients. To further analyze the effects of bachelor degree holding residents of counties on per capita income, confidence intervals and F-tests were utilized to identify similarity and presence of linear relationships between these variables. Lastly, the researchers used residual and normal probability plots as regression diagnostics to determine if a linear model was an appropriate fit for the original 3 predictor variables (total population, number of hospital beds, and total personal income) when related to the number of active physicians. Overall, the research team used advanced linear regression analysis tools such as confidence intervals, normal probability plots, goodness of fit statistics, residual plots, and regression modeling to determine if there were linear relationships between several unique variables of interest.

In this project, the project team was interested in learning if any of the variables could be used as predictor variables. Specifically, we wanted to see what predictor variables would best predict the number of active physicians. To do this, we used scatterplot and correlation matrices to evaluate 7 of the variables. We then used the highest R^2 value to determine the best predictor variables for number of active physicians. After this, we wanted to see if any additional predictor variables could be added to better predict the number of active physicians, aside from total population and personal income. To do this, we used the lowest coefficient of partial determination and the highest extra sum of squares. We then performed F tests to see if adding this extra predictor variable would make the model more predictable. Finally, we tested if a pair of predictor variables should be added to the model that already contained total population and personal income. Overall, the research team used tools like stem-and-leaf plots, scatterplot matrix, correlation matrix, coefficient of partial determination, extra sum of squares, and F tests to determine if these variables would be good predictors of the number of active physicians.

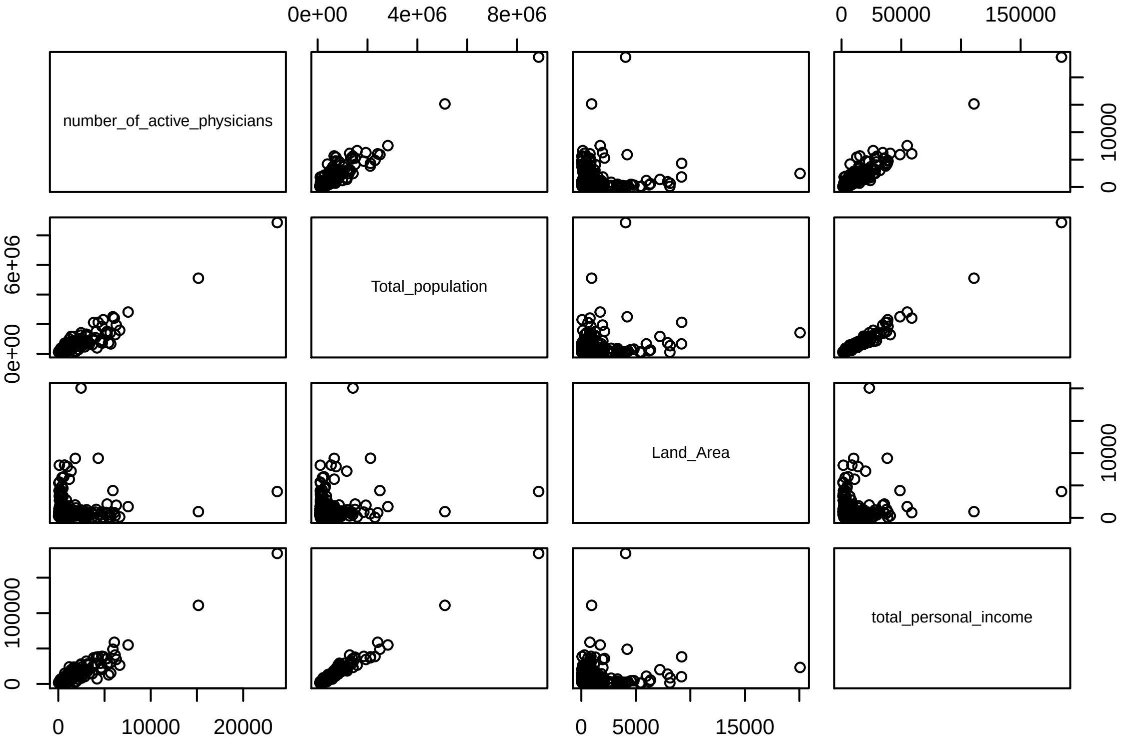
Part I: Multiple Linear Regression I

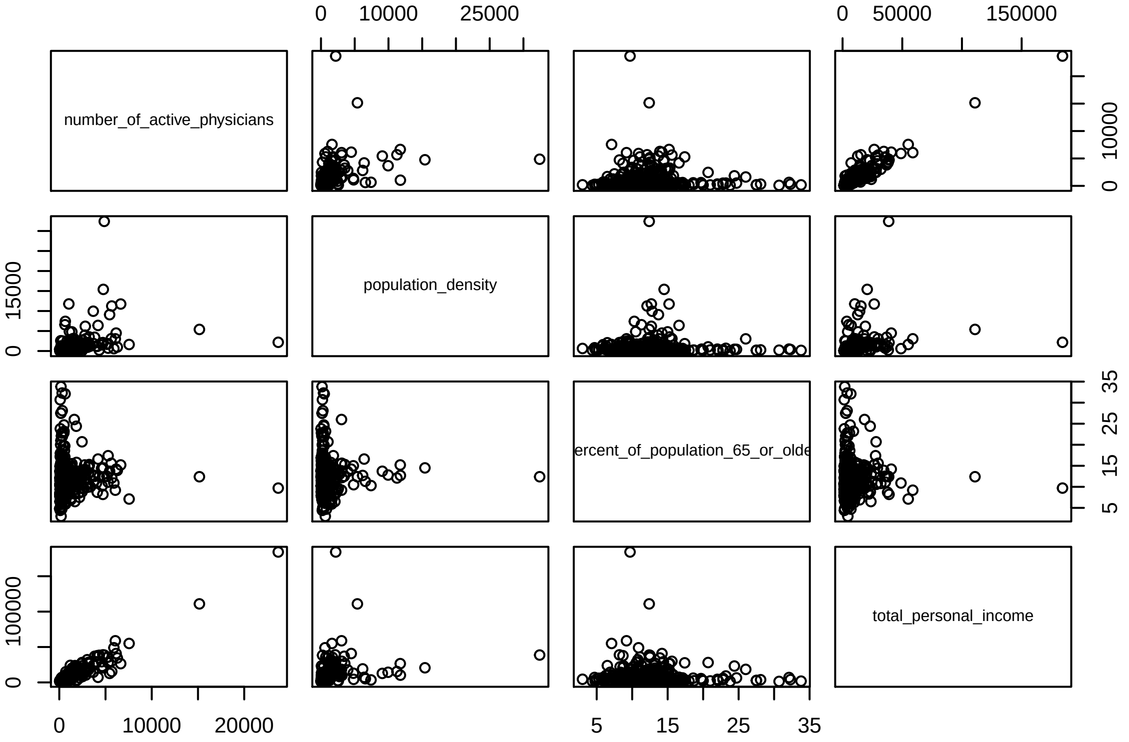
6.28. Refer to the **CDI** data set in Appendix C.2. You have been asked to evaluate two alternative models for predicting the number of active physicians (Y) in CDI. Proposed model I includes as predictor variables total population (X_1), land area (X_2), and total personal income (X_3). Proposed model II includes as predictor variables density (X_1, total population divided by land area), percent of population greater than 64 years old (X_2), and total personal income (X_3)

a. Prepare a stem-and-leaf plot for each of the predictor variables. What noteworthy information is provided by your plots?

Noteworthy information obtained from the plots: For all the stem and leaf plots, except for the Percent of Population 65 or Older, the data is heavily skewed to the left. There seem to be about 2 outliers for each of the extremely skewed plots. The presence of these outliers may assist in making these plots look extremely skewed. The Percent of Population 65 or Older plot has a more unimodal, normal looking plot. This may be because there doesn't seem to be any outliers and most of the data is clustered around the middle.

b. Obtain the scatter plot matrix and the correlation matrix for each proposed model. Summarize the information provided.




0
10000
25000
0
50000
150000
number_of_active_physicians
0
10000
0
15000
population_density
ercent_of_population_65_or_olde
35
25
15
5
0
100000
total_personal_income
0
10000
20000
5
15
25
35

The results of the correlation and scatterplot matrix are Model 1: When "Land Area" acts as either the X or Y variable in the plot, the points on the scatterplot show a U-shaped pattern instead of 1 diagonal line across the plot. This may suggest that Land Area is not good predictor of this data. In addition, the Rˆ2 values of Land Area show values extremely close to 0 (0.127, 0.078, 0.173) which also support the theory that Land Area is not linearly correlated to the any of the other predictor variables. For the rest of the scatterplots, the presence of a couple of outliers seem to skew the data, but the Rˆ2 values show a strong linear correlation between predictor variable, with the exception of Land Area.

Model 2: The scatterplots show the only linear correlation between predictor variables seem to be between "Total Personal Income" and "Number of Active Physicians." The correlation matrix supports this as well, because these predictors have the only strong Rˆ2 value (close to 1). This means the variation in Total Personal Income accounts for more of the variability in Number of Active Physicians. The other predictors don't seem to be linearly correlated, as the scatterplots show patterns that deviate from a diagonal line from the bottom left corner of the plot to the top right corner.

c. For each proposed model, fit the first-order regression model (6.5) with three predictor variables.

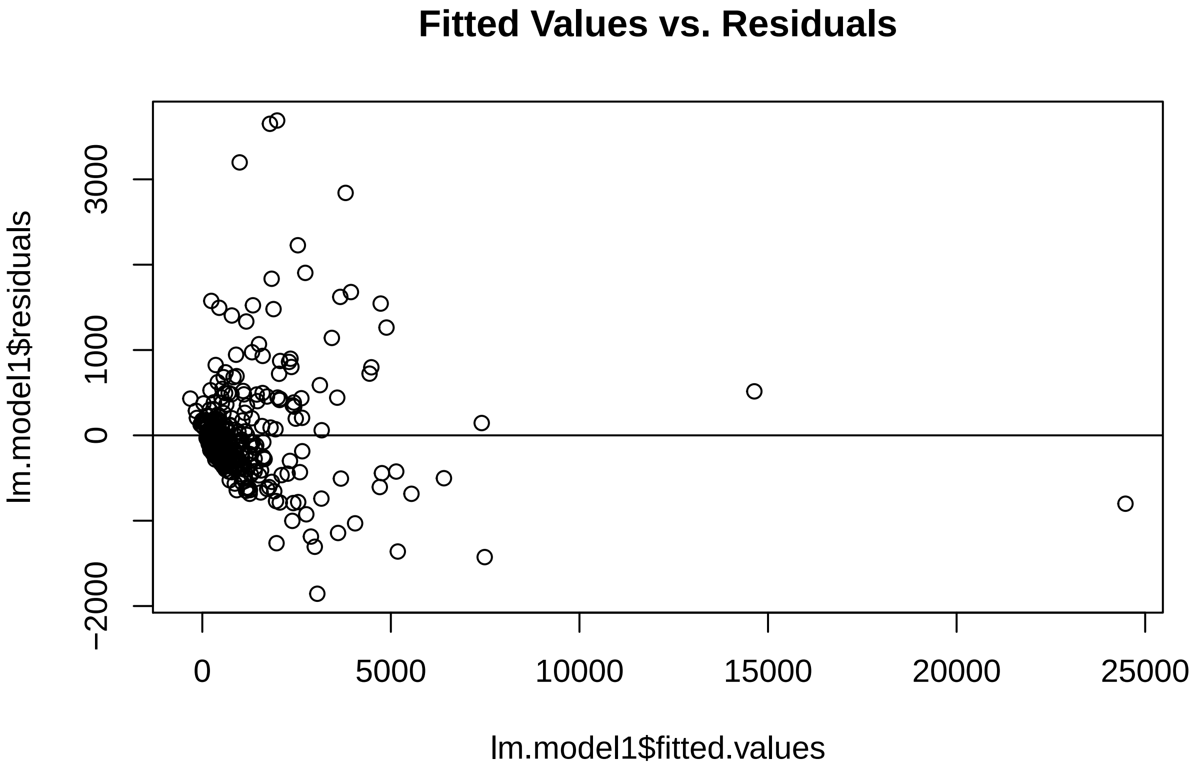
The first-order linear regression model for model 1 is $\hat{Y} = -13.3161522 + 8.3661782 \times 10^{-4} X_1 + -0.065523 X_2 + 0.094132 X_3$

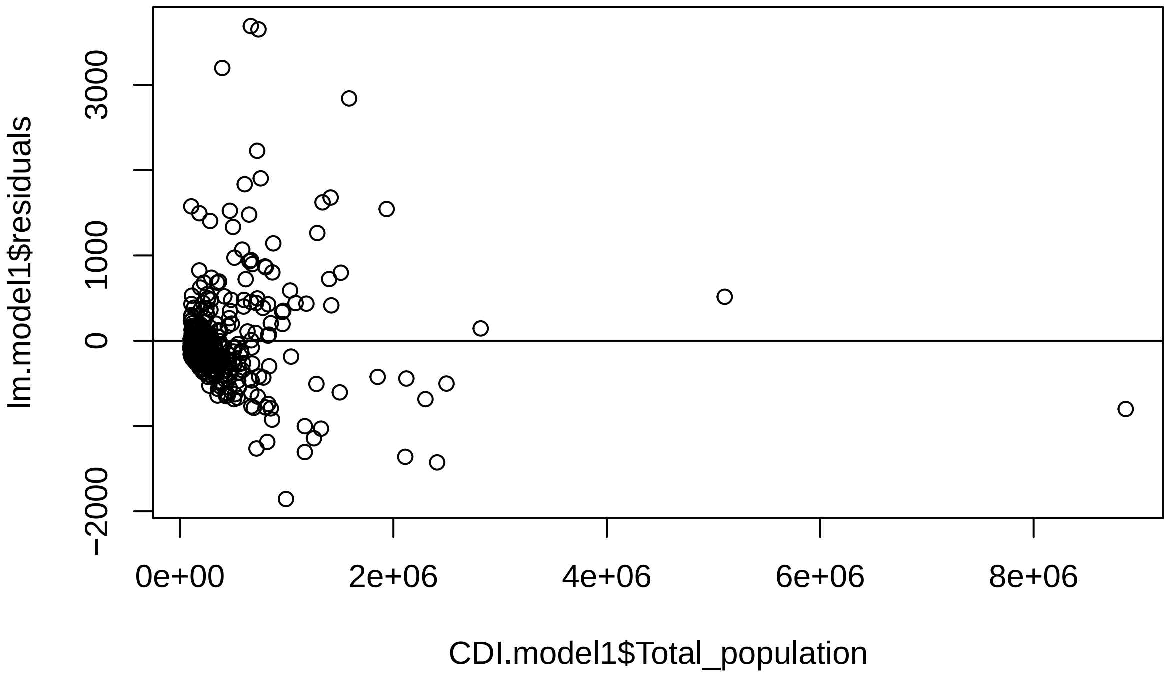
The first-order linear regression model for model 2 is $\hat{Y} = -170.5742233 + 0.0961589 X_1 + 6.3398406 X_2 + 0.1265665 X_3$

d. Calculate R^2 for each model. Is one model clearly preferable in terms of this measure?

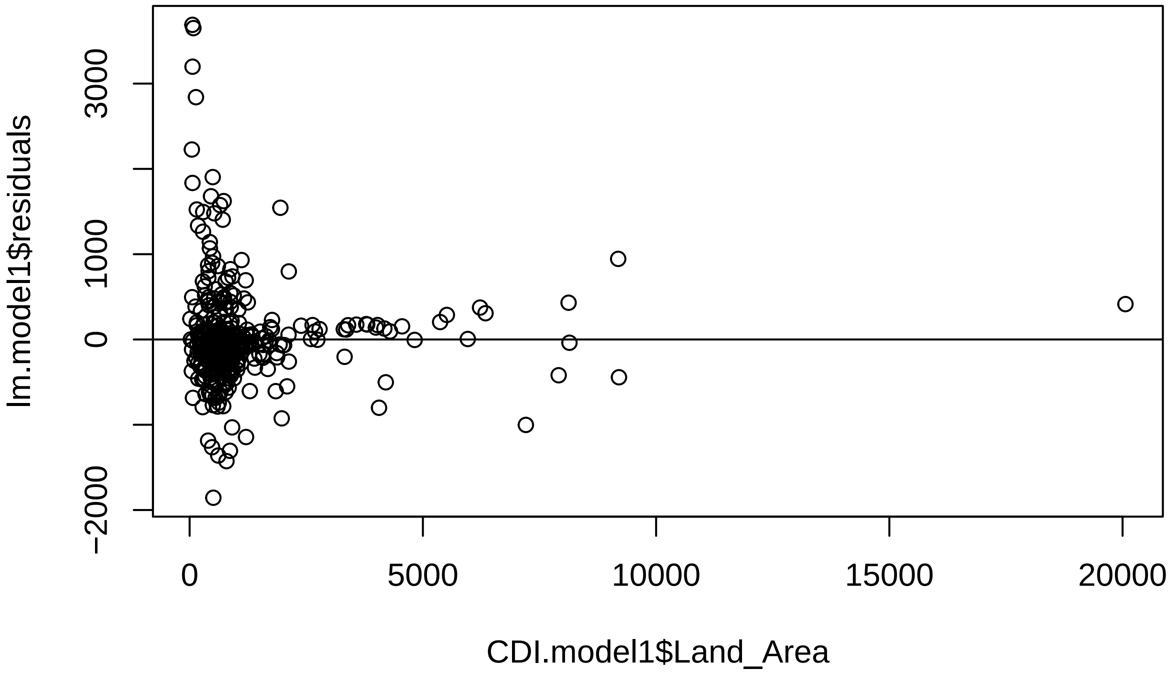
R^2 for model 1 is 0.9026432 and R^2 for model 2 is 0.9117491. In terms of R^2 model 2 is better since it has a higher R^2 meaning it accounts for more of the variability in the number of active physicians, however the two are very close and we cannot conclude that one is *clearly* better than the other.

e. For each model, obtain the residuals and plot them agaisnt $\hat{Y}$, each of the three predictor variables, and each of the two-factor interaction terms. Also prepare a normal probabilty plot for each of the two fitted models. Interpret your plots and state your findings. Is one model clearly preferable in terms of appropriateness?

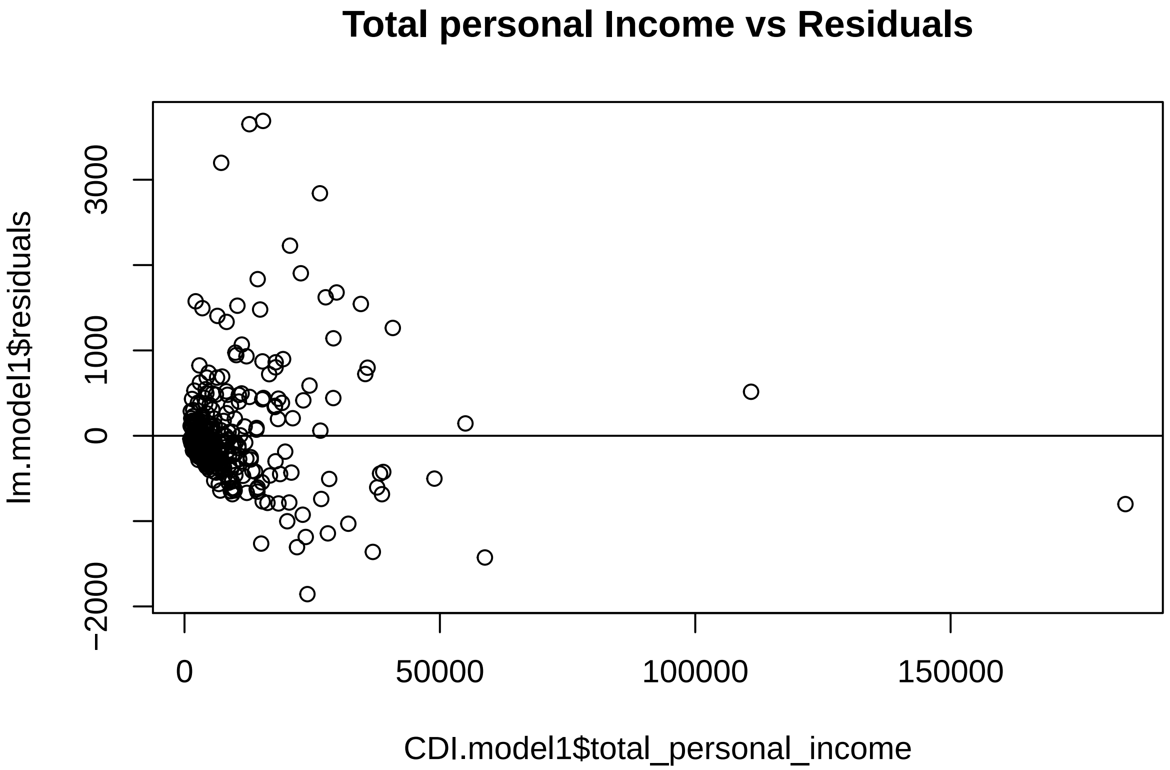


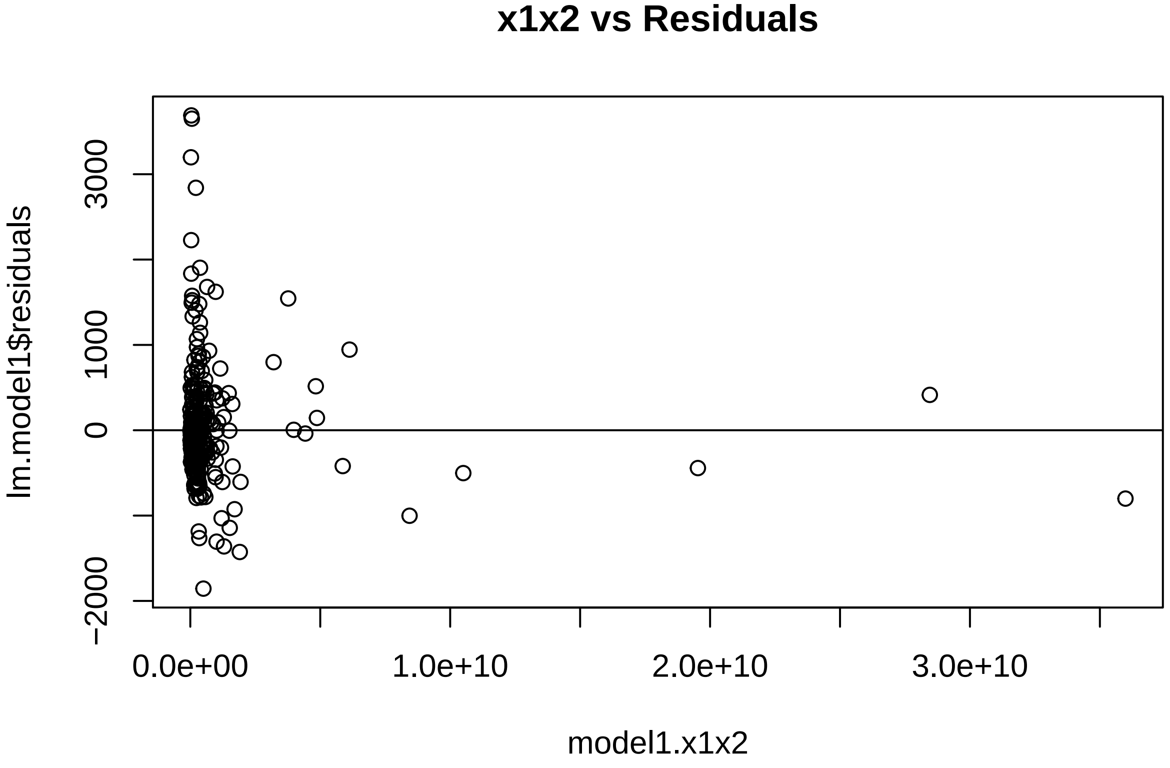


Total Population vs Residuals

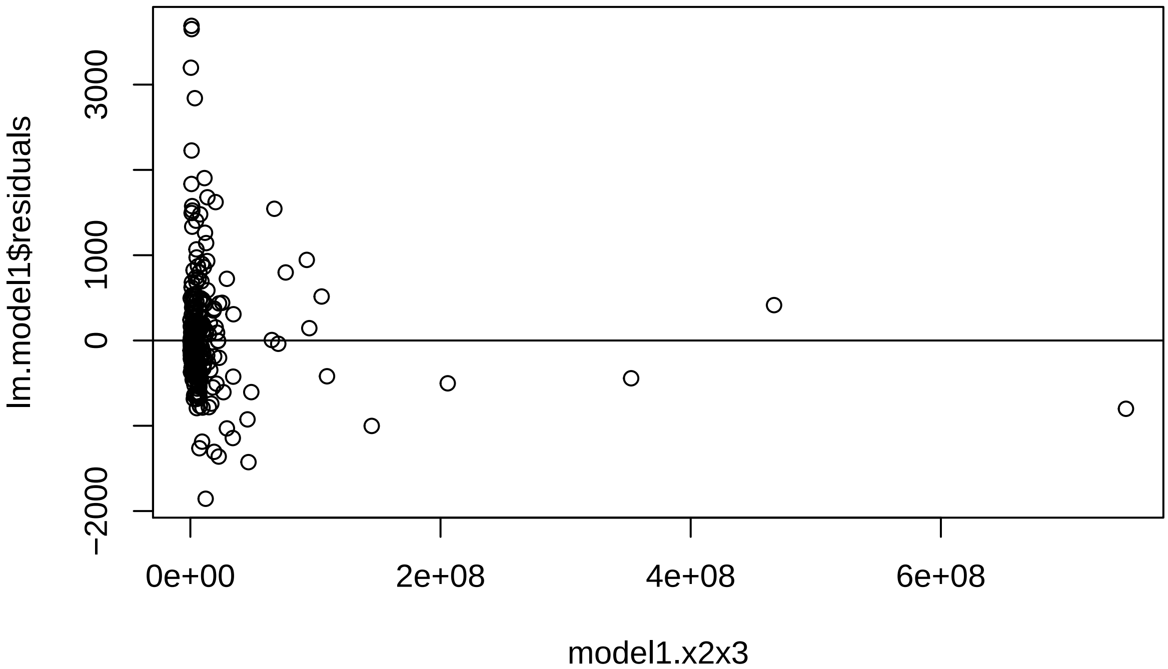


Land Area vs Residuals

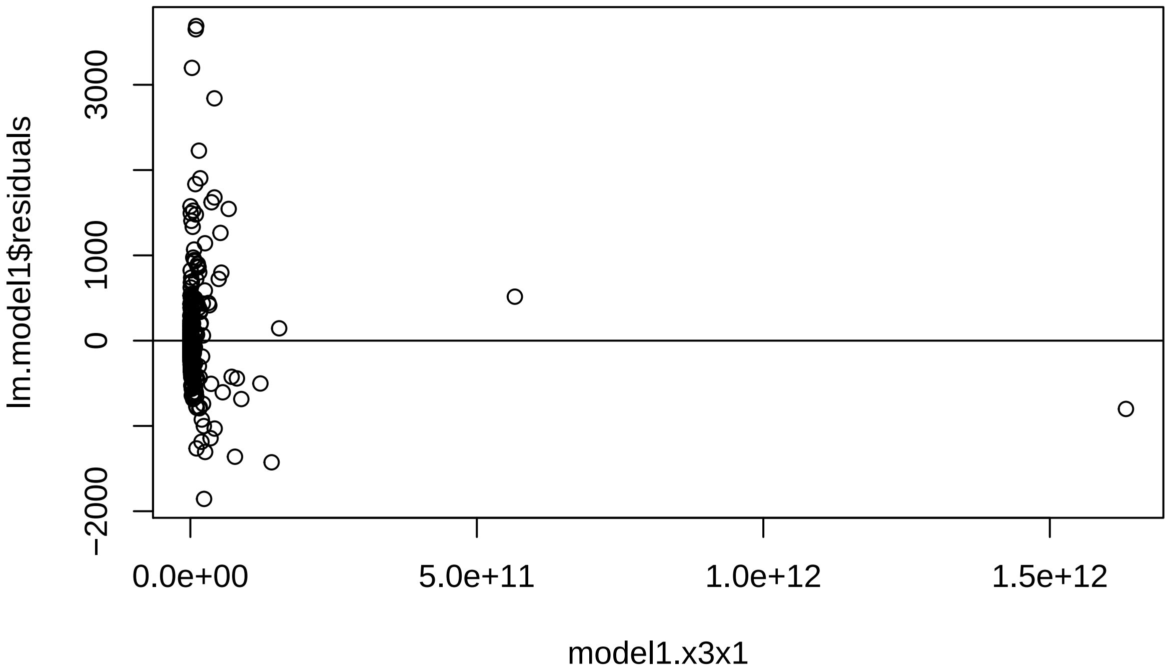

Total personal Income vs Residuals
lm.model1$residuals
−2000
0
1000
3000
0
50000
100000
150000
CDI.model1$total_personal_income

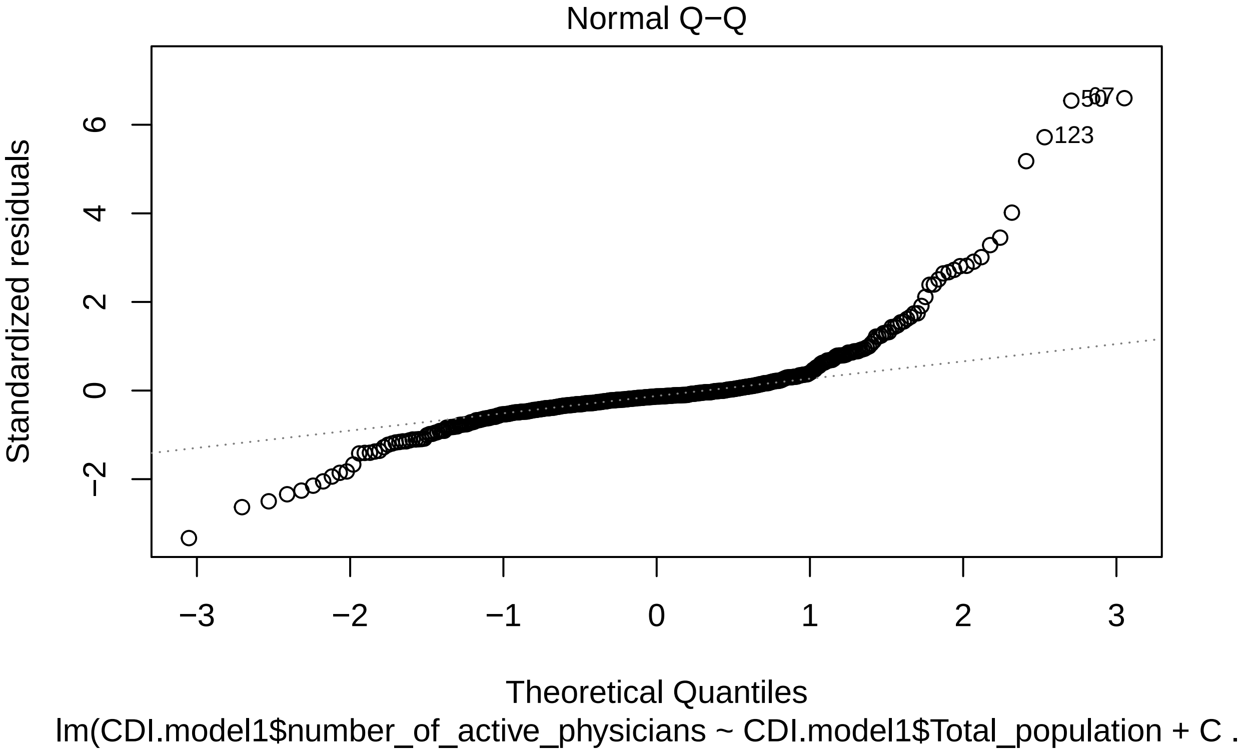

x1x2 vs Residuals
lm.model1$residuals
−2000
0
1000
3000
0.0e+00
1.0e+10
2.0e+10
3.0e+10
model1.x1x2

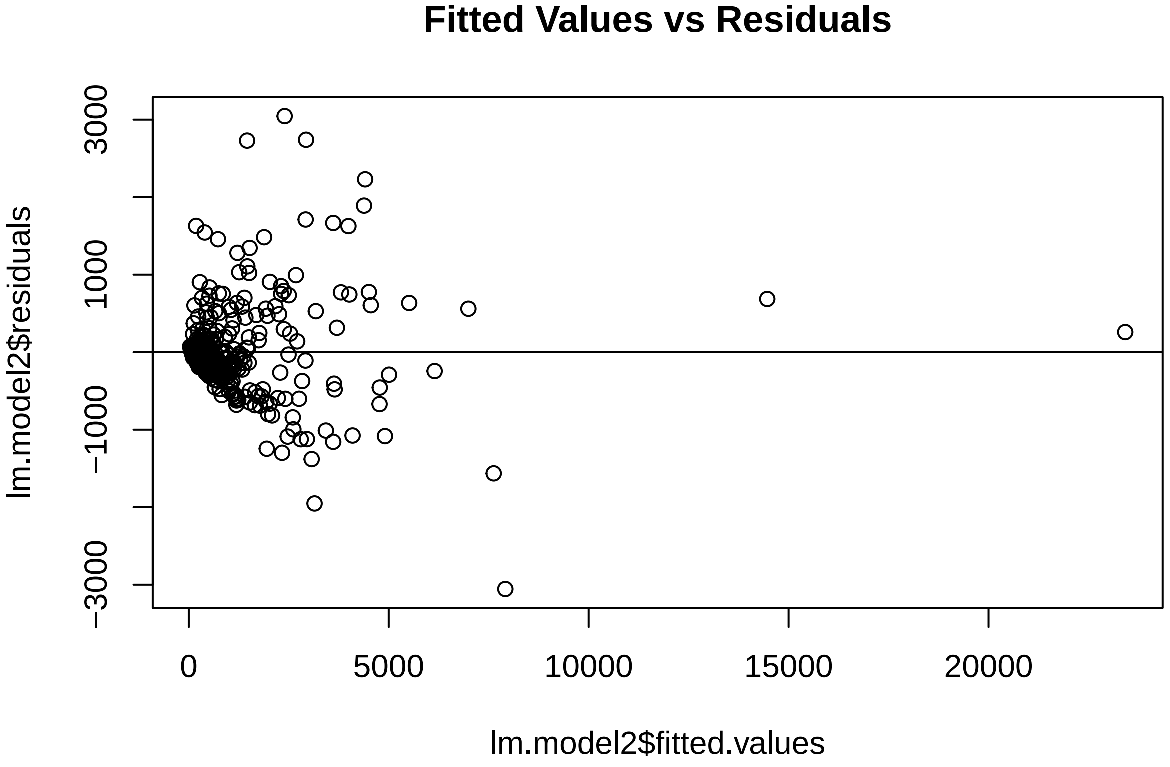
x2x3 vs Residuals

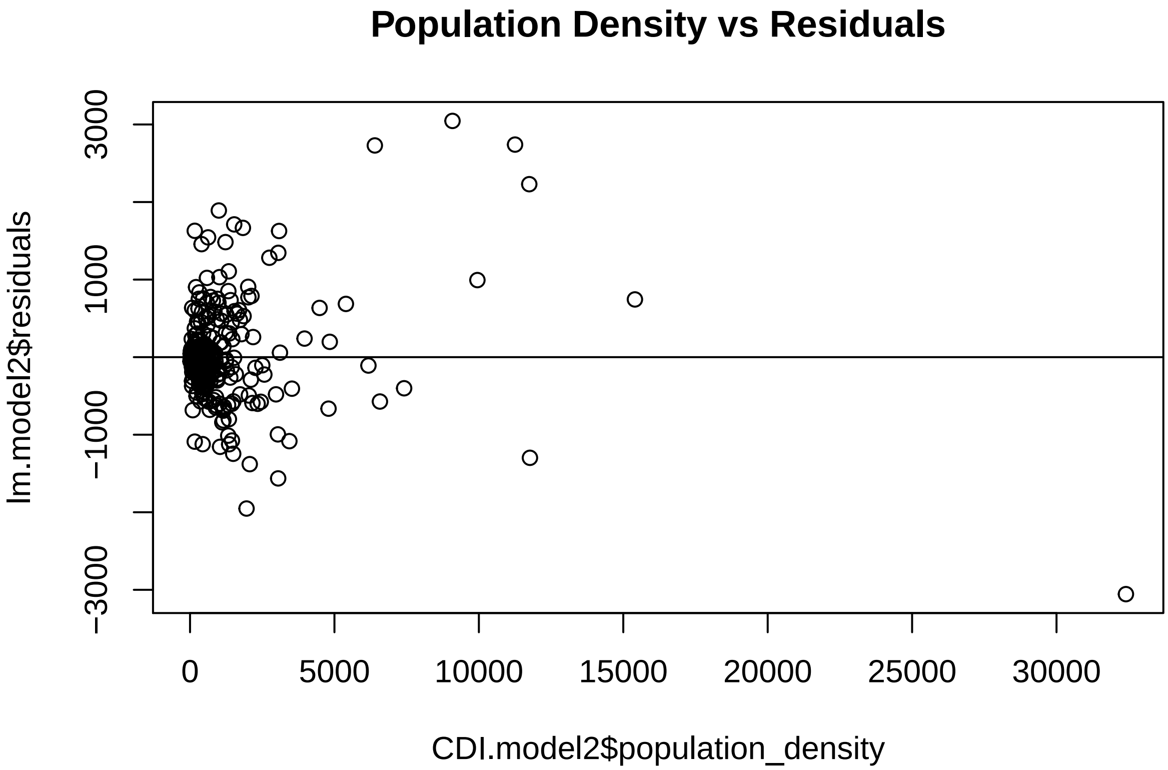


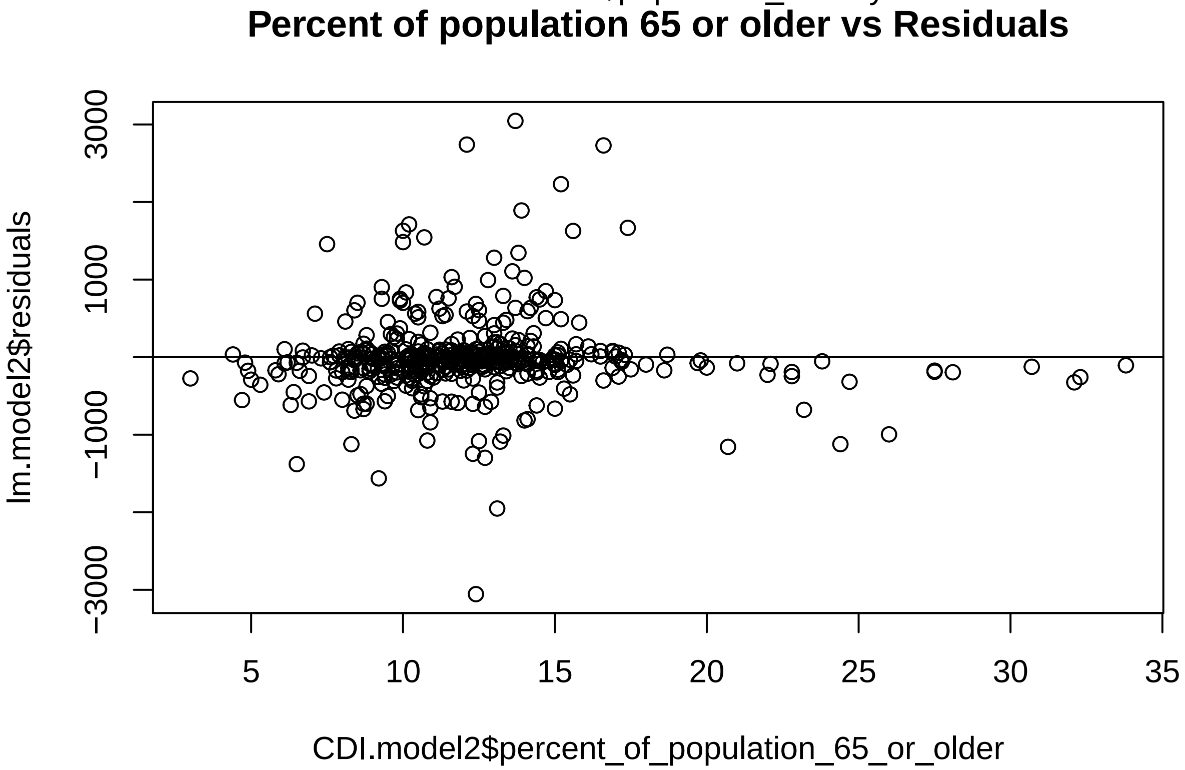
x1x3 vs Residuals

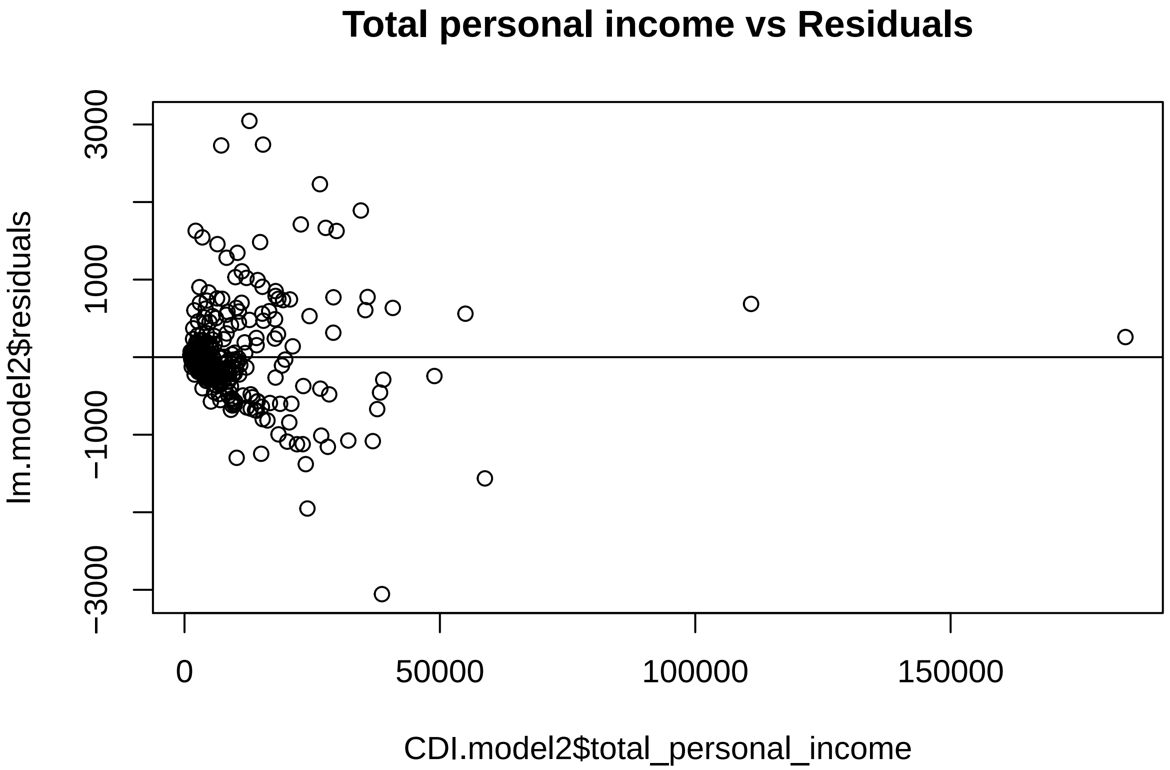


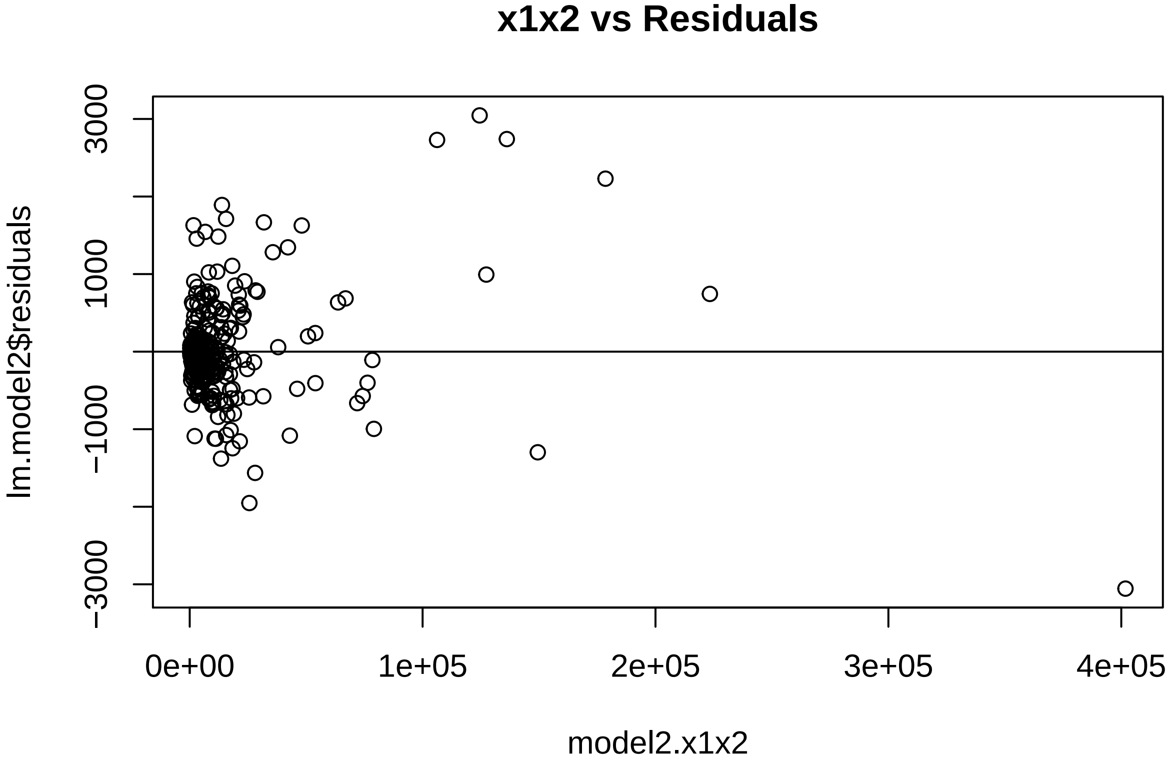

Normal Q−Q
Standardized residuals
−2
0
2
4
6
50
67
123
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(CDI.model1$number_of_active_physicians ~ CDI.model1$Total_population + C .

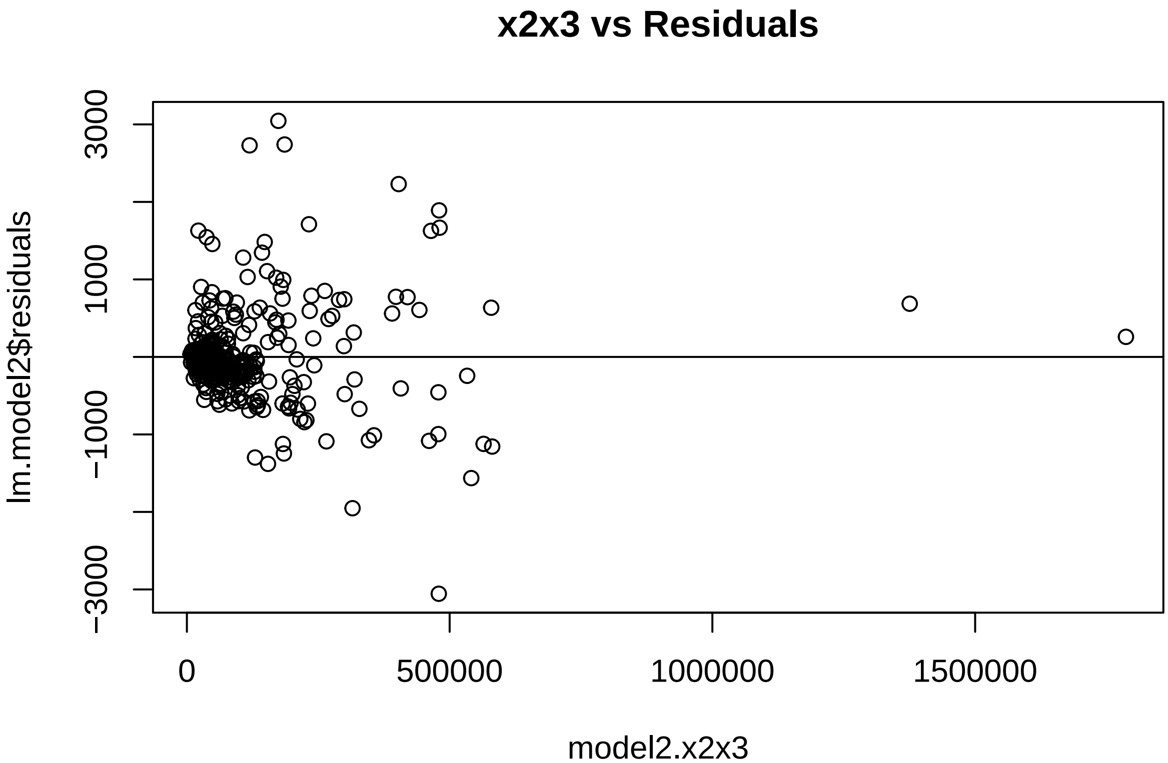

Fitted Values vs Residuals
lm.model2$residuals
−3000
−1000
1000
3000
0
5000
10000
15000
20000
lm.model2$fitted.values

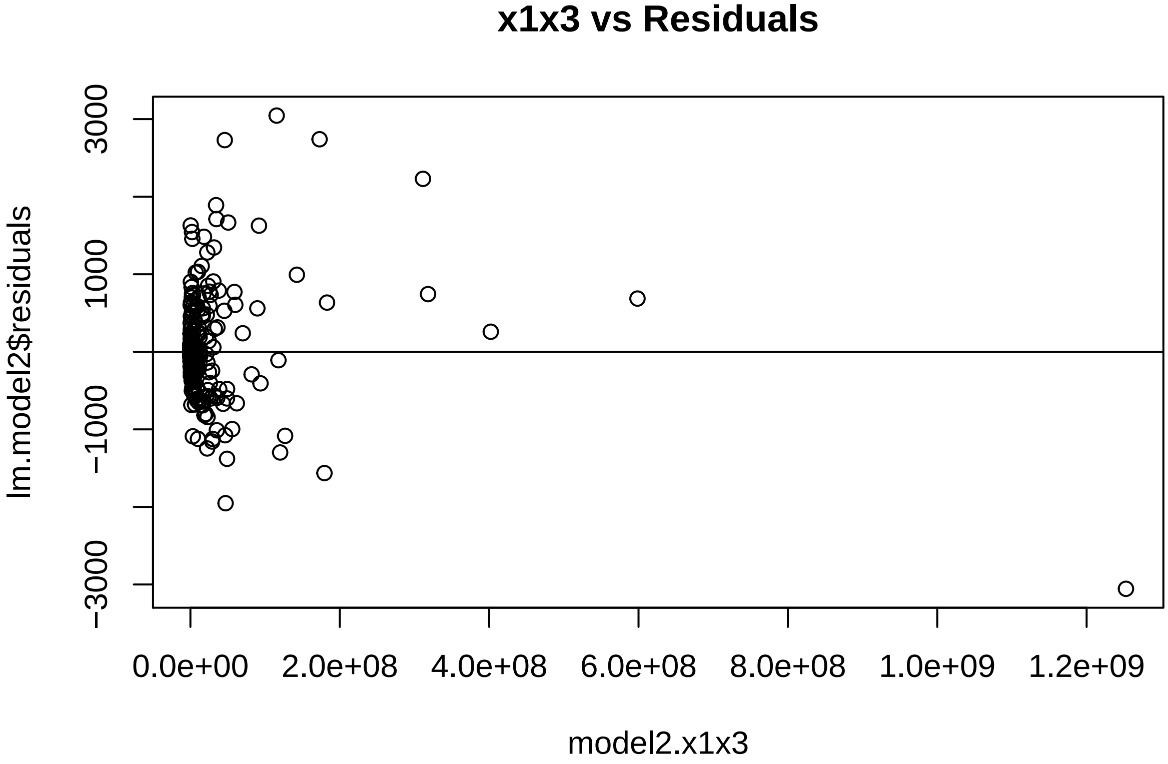

Population Density vs Residuals
lm.model2$residuals
3000
1000
−1000
−3000
0
5000
10000
15000
20000
25000
30000
CDI.model2$population_density

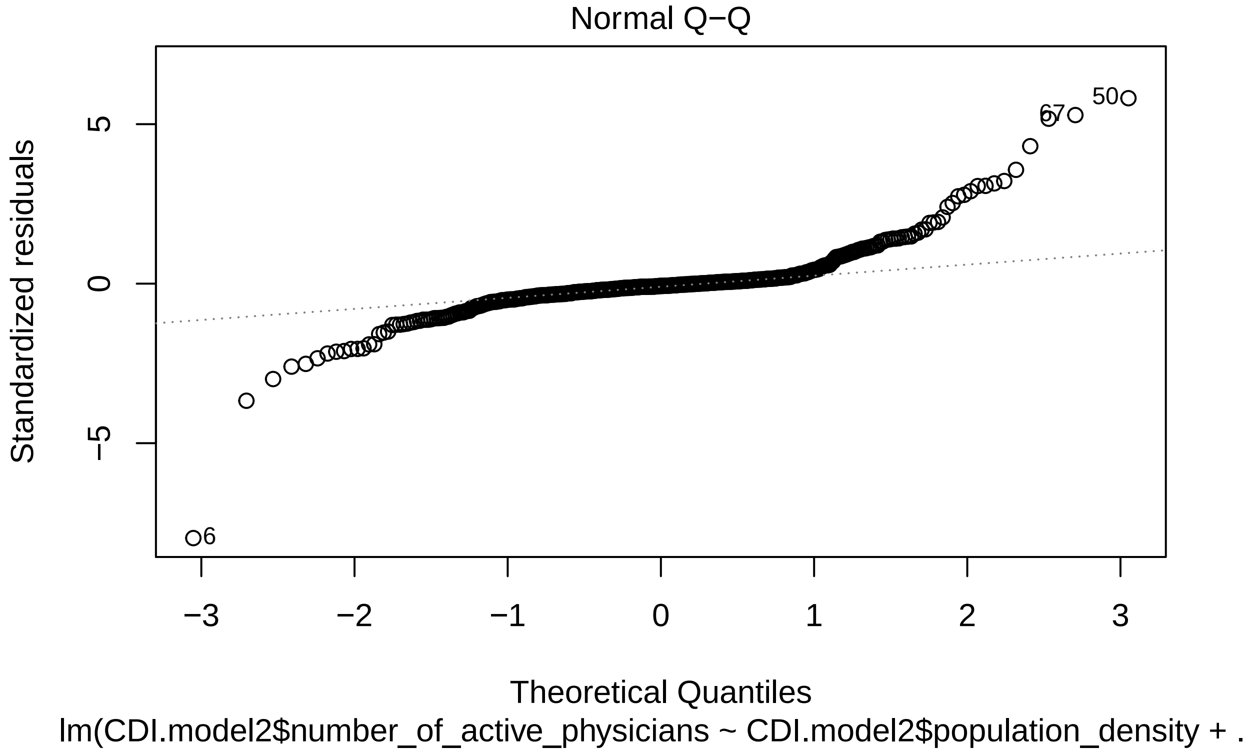

Percent of population 65 or older vs Residuals
lm.model2$residuals
3000
1000
−1000
−3000
5
10
15
20
25
30
35
CDI.model2$percent_of_population_65_or_older


Total personal income vs Residuals
lm.model2$residuals
3000
1000
−1000
−3000
0
50000
100000
150000
CDI.model2$total_personal_income


x1x2 vs Residuals
lm.model2$residuals
3000
1000
−1000
−3000
0e+00
1e+05
2e+05
3e+05
4e+05
model2.x1x2


x2x3 vs Residuals
lm.model2$residuals
3000
1000
−1000
−3000
0
500000
1000000
1500000
model2.x2x3


x1x3 vs Residuals
lm.model2$residuals
3000
1000
−1000
−3000
0.0e+00
2.0e+08
4.0e+08
6.0e+08
8.0e+08
1.0e+09
1.2e+09
model2.x1x3



Neither model seems to be normally distributed as the normal probability plots stray away at both ends. In fact for Residuals vs Fitted Values, Total Population and totla personal income appear to have increasing variance for model 1, while Residuals vs Land Area has decreasing variance. A great number of two way interactions for model 1 is centered at x = 0. For model 2, Population density and total personal income vs Resiudals appears to have incrasing varince. Percent of population older than 64 appears to have constant variance. Again the two way plots are mostly around x = 0.

f. Now expand both models proposed above by adding all possible two-factor interactions. Note that, for a model with X_1, X_2, X_3 as the predictors, the two-factorinteractions are X_1X_2, X_1X_3, X_2X_3. Repeat part d for the two expanded models.

R^2 for model 1 is 0.9063789 and R^2 for model 2 is 0.9230238. In terms of R^2 model 2 is better since it has a higher R^2 meaning it accounts for more of the variability in the number of active physicians, however the two are close and we cannot conclude that one is *clearly* better than the other.

7.37

a) For each of the following variables, calculate the coefficient of partial determination given that X_1 and X_2 are included in the model: land area(X_3), percent of population 65 or older (X_4), number of hospital beds X_5, and total serious crimes (X_6)

coefficient of partial determination for land area: 0.028825

coefficient of partial determination for percent of population 65 or older: 0.0038424

coefficient of partial determination for number of hospital beds: 0.5538182

b) On the basis of results in part (a), which of the four additional predictor variables is best? Is the extra sum of square associated with this variable larger than those for the other three variables?

Since x5 (number of hospital beds) has the greatest coefficient of partial determination, it is the best predictor variable for the response variable 'number of active physicians'

Yes, the extra SS is largest for x5 (number of hospital beds, 7.8070132×10^7). The extra SS for land area and percent of population 65 or older are: 4.0633696×10^6, 5.416473×10^5, respectively.

c) Using the F* test statistic, test whether or not the variable determined to be best in part (b) is helpful in the regression model when X_1 and X_2 are included in the model; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. Would the F* test statistics for the other three potential variables be as large as the one here? Discuss

H0 : B5 = 0

HA : B5 $\neq$ 0

Since Fstar > F (541.1800993 > 6.6933576), we reject the null hypothesis at a 99% confidence level and conclude that the number of hospital beds is an important regressor in the model

The Fstar predictors of the other two variables would not be as large as this one because those variables (land area and percent of population over 65) are not as important and contributional to the regression model as the number of hospital beds

d) Compute three additional coefficients of partial determination: $R^2_{Y,X_3,X_4|X_1,X_2}$, $R^2_{Y,X_3,X_5|X_1,X_2}$ $R^2_{Y,X_4,X_5|X_1,X_2}$. Which pair of predictors is relatively more important than other pairs? Use the F test to find out whether adding the best pair to the model is helpful given that X_1, X_2 are already included.

The pair of variables land area (x4) and number of hospital beds (x5) is the most important predictor pair to forecast the active number of physicians with a coefficient of partial determination of 0.5642756 compared to 0.0331418 for x3x4 and 0.5558232 for x3x5.

H0 : B4 = B5 = 0

HA : B4 $\neq$ 0 or B5 $\neq$ 0

Since Fstar > F (281.6687674 > 4.6542692), we reject the null hypothesis with 99% confidence and conclude that adding the predictor pair percent of population 65 or older (x4) and number of hospital beds (x5) is important in explaining the model given that X1 and X2 are already included.

Discussion

For the first part of the project, we were able to use stem-and-leaf plots, scatterplot matrices, and correlation matrices to gather that although most of the predictor variables are heavily skewed, this is mostly due to the presence of 1 or 2 outliers. In particular, Number of Active Physicians, Total Population, and Total Personal Income are strongly, linearly correlated with high Rˆ2 values. Considering the second part of the project, this suggests Total Population and Personal Income are good predictors of Number of Active Physicians. If we were to add another variable to assist in predicting Number of Active Physicians, it would have to be Number of Hospital Beds. Since it has the smallest coefficient of partial determination and the largest estra sum of squares, Number of Hospital Beds would make a better additional predictor than Land Area, Percent of Population 65 or Older, or Total Serious Crimes.

The most relevant parts of the course materials for this project were the correlation matrix and the extra sum of squares. For the first part, the correlation matrix really showed which of the predictor variables were most correlated, aiding in figuring out which were the best predictor variables for the Number of Active Physicians. It was obvious from the plots that only Total Population and Personal Income had a strong linear correlation to Number of Active Physicians. For the second part, the extra sum of squares allowed us to confirm our suspicion that Number of Hospital Beds would be the additional variable to better predict the Number of Active Physicians. The largest value of the extra SS was very obviously the Number of Hospital Beds. The F test only confirmed that adding Number of Hospital Beds would be useful, not that Number of Hospital Beds was the next best predictor.

To improve the linear regression models, it might be helpful to test all the predictor variables, apart from the ones we tested in this project. Other predictor variables may aid Total Population and Total Personal Income or even be better than them at predicting the Number of Active Physicians. To do this, we can first create a scatterplot matrix of the other variables along with Number of Active Physicians. If any seem to result in a higher correlation, we can go on to use extra SS and F tests to see if they would be helpful in aiding to predict Number of Active Physicians.

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE, results = 'hide')
#import the data
CDI <- read.table("~/Desktop/Spring 2019/STA 108 /Projects/Project 2/CDI.txt", quote="\"", comment.char
#give proper variable names to the data
names(CDI) <- c("ID", "City", "State", "Land_Area", "Total_population", "percent_of_population_aged_18-3
CDI$population_density <- CDI$Total_population/CDI$Land_Area
#6.28 part a
#creating stem and leaf plots
#model I
stem(CDI$Total_population)
stem(CDI$Land_Area)
stem(CDI$total_personal_income)

#model II
stem(CDI$population_density)
stem(CDI$percent_of_population_65_or_older)
stem(CDI$total_personal_income)
#scatterplot matrix and correlation matrix for model 1
model1 <- c("number_of_active_physicians", "Total_population", "Land_Area", "total_personal_income")
CDI.model1 <- CDI[model1]
pairs(CDI.model1)
cor(CDI.model1)
#scatterplot matrix and correlation matrix for model 2
model2 <- c("number_of_active_physicians", "population_density", "percent_of_population_65_or_older", "
CDI.model2 <- CDI[model2]
pairs(CDI.model2)
cor(CDI.model2)
#linear regression for model 1
lm.model1 <- lm(CDI.model1$number_of_active_physicians ~ CDI.model1$Total_population + CDI.model1$Land_
model1.beta_0 = lm.model1$coefficients[[1]]
model1.beta_1 = lm.model1$coefficients[[2]]
model1.beta_2 = lm.model1$coefficients[[3]]
model1.beta_3 = lm.model1$coefficients[[4]]

#linear regression for model 2
lm.model2 <- lm(CDI.model2$number_of_active_physicians ~ CDI.model2$population_density + CDI.model2$perc
model2.beta_0 = lm.model2$coefficients[[1]]
model2.beta_1 = lm.model2$coefficients[[2]]
model2.beta_2 = lm.model2$coefficients[[3]]
model2.beta_3 = lm.model2$coefficients[[4]]
#obtaining R-squared for each model
model1.r2 = summary(lm.model1)$r.squared
model2.r2 = summary(lm.model2)$r.squared
#residual plot agaisnt Y, each of the three predictor variables, and each two way interations
{plot(lm.model1$fitted.values, lm.model1$residuals, main = "Fitted Values vs. Residuals")
abline(0,0)}
{plot(CDI.model1$Total_population, lm.model1$residuals, main = "Total Population vs Residuals")
```

```
abline(0,0)}
{plot(CDI.model1$Land_Area, lm.model1$residuals, main = "Land Area vs Residuals")
abline(0,0)}
{plot(CDI.model1$total_personal_income, lm.model1$residuals, main = "Total personal Income vs Residuals
abline(0,0)}

model1.x1x2 = as.numeric(CDI.model1$Total_population)*as.numeric(CDI.model1$Land_Area)
{plot(model1.x1x2, lm.model1$residuals, main = "x1x2 vs Residuals")
abline(0,0)}
model1.x2x3 = CDI.model1$Land_Area*CDI.model1$total_personal_income
{plot(model1.x2x3, lm.model1$residuals, main = "x2x3 vs Residuals")
abline(0,0)}
model1.x3x1 = as.numeric(CDI.model1$total_personal_income)*as.numeric(CDI.model1$Total_population)
{plot(model1.x3x1, lm.model1$residuals, main = "x1x3 vs Residuals")
abline(0,0)}

#normal probability plot for model 1
plot(lm.model1, which = 2)

{plot(lm.model2$fitted.values, lm.model2$residuals, main = "Fitted Values vs Residuals")
abline(0,0)}
{plot(CDI.model2$population_density, lm.model2$residuals, main = "Population Density vs Residuals")
abline(0,0)}
{plot(CDI.model2$percent_of_population_65_or_older, lm.model2$residuals, main = "Percent of population 
abline(0,0)}
{plot(CDI.model2$total_personal_income, lm.model2$residuals, main = "Total personal income vs Residuals
abline(0,0)}
model2.x1x2 = CDI.model2$population_density*CDI.model2$percent_of_population_65_or_older
{plot(model2.x1x2, lm.model2$residuals, main = "x1x2 vs Residuals")
abline(0,0)}
model2.x2x3 = CDI.model2$percent_of_population_65_or_older*CDI.model2$total_personal_income
{plot(model2.x2x3, lm.model2$residuals, main = "x2x3 vs Residuals")
abline(0,0)}
model2.x1x3 = CDI.model2$population_density*CDI.model2$total_personal_income
{plot(model2.x1x3, lm.model2$residuals, main = "x1x3 vs Residuals")
abline(0,0)}
#normal probability plot for model 2
plot(lm.model2, which = 2)
expanded1 = lm(number_of_active_physicians ~ .^2, data = CDI.model1)
expanded2 = lm(number_of_active_physicians ~ .^2, data = CDI.model2)

expanded1.r2 = summary(expanded1)$r.squared
expanded2.r2 = summary(expanded2)$r.squared
#a
lm.modelbase <- lm(CDI$number_of_active_physicians ~ CDI$Total_population + CDI$total_personal_income)
lm.modelx3 <- lm(CDI$number_of_active_physicians ~ CDI$Total_population + CDI$total_personal_income + C
lm.modelx4 <- lm(CDI$number_of_active_physicians ~ CDI$Total_population + CDI$total_personal_income + C
lm.modelx5 <- lm(CDI$number_of_active_physicians ~ CDI$Total_population + CDI$total_personal_income + C

parR2 <- function(m.full, m.reduced)
  {
  a.full <- anova(m.full)
  a.reduced <- anova(m.reduced)
  sse.full <- tail(a.full$"Sum Sq", 1)
  sse.red <- tail(a.reduced$"Sum Sq", 1)

  pR2 <- (sse.red - sse.full) / sse.red
  return(pR2)}

parR2.land = parR2(lm.modelx3,lm.modelbase)
# 0.02882495
parR2.65 = parR2(lm.modelx4, lm.modelbase)
# 0.003842367
parR2.beds = parR2(lm.modelx5, lm.modelbase)
# 0.5538182
#b
extrass.x3 = anova(lm.modelx3)$"Sum Sq"[3]
# 4063369.6
extrass.x4 = anova(lm.modelx4)$"Sum Sq"[3]
# 541647.3
extrass.x5 = anova(lm.modelx5)$"Sum Sq"[3]
# 78070131.6

extrass = c(extrass.x3, extrass.x4, extrass.x5)
extrass
max(extrass)
# 78070131.6
#c
sse.fftest = anova(lm.modelx5)$"Sum Sq"[4]
sse.rftest = anova(lm.modelbase)$"Sum Sq"[3]
n=440
alpha = 0.01
df.r = n-3
df.f = n-4

Fstar = ((sse.rftest-sse.fftest)/(df.r-df.f)/(sse.fftest/df.f))
# 541.1801
F = qf(1-alpha, df.r-df.f, df.f)
# 6.693358

Fstar > F
#d
lm.modelx3x4 = lm(CDI$number_of_active_physicians ~ CDI$Total_population + CDI$total_personal_income + C
lm.modelx3x5 = lm(CDI$number_of_active_physicians ~ CDI$Total_population + CDI$total_personal_income + C
lm.modelx4x5 = lm(CDI$number_of_active_physicians ~ CDI$Total_population + CDI$total_personal_income + C

x3x4 = parR2(lm.modelx3x4, lm.modelbase)
# 0.03314181
x3x5 = parR2(lm.modelx3x5, lm.modelbase)
# 0.5558232
x4x5 = parR2(lm.modelx4x5, lm.modelbase)
# 0.3042514

coef.parpairs = c(x3x4, x3x5, x4x5)
max(coef.parpairs)
# 0.5642756
```

```
sse.fftestpair = anova(lm.modelx4x5)$"Sum Sq"[5]
sse.rftestpair = anova(lm.modelbase)$"Sum Sq"[3]
n=440
alpha = 0.01
df.rpair = n-3
df.fpair = n-5

Fstar.pair = ((sse.rftestpair-sse.fftestpair)/(df.rpair-df.fpair)/(sse.fftestpair/df.fpair))
# 281.6688
F.pair = qf(1-alpha, df.rpair-df.fpair, df.fpair)
# 4.654269
Fstar.pair > F.pair
```