Project2

Arya Gupta

3/9/2022

Introduction

- This project consists of 3 parts: the first two deal with multiple linear regression and the final part is a discussion on the project. We are given a data set, called "CDI.txt", which consists of county demographic information for 440 of the most populous counties in the United States. The data set provides 14 different variables for a single county. In this project, I worked with 7 of these variables: Number of Active Physicians, Total Population, Land Area, Total Personal Income, Population Density (which is found by dividing Total Population and Land Area), Percent of Population 65 or Older, and Number of Hospital Beds.
- For Part 1, the task is to create two different models with different regression functions, and answer various questions through that model. In Model 1, the predictor variables are Total Population (X_1), Land Area (X_2), and Total Personal Income (X_3). In Model 2, the predictor variables are Population Density (X_1), Percent of Population 65 or Older (X_2), and Total Personal Income (X_3). The response variable (Y_i) is Number of Active Physicians for both models.
- For Part 2, the predictor variables are Total Population (X_1) and Total Personal Income (X_2). In this part, we have to test which additional variable, between Land Area(X_3), Percent of Population 65 or Older (X_4), and Number of Hospital Beds (X_5), would be the most helpful to the regression model.
- There were a few tools I used throughout this project. For example, I used the $pairs()$ function to create scatter plot matrices. I used the $lm()$ function numerous times to fit the regression model into a variable through which I was able to find coefficients of the regression function and ANOVA terms such as SSR and SSE. I used the "qf" function to find the critical value of the F-distribution. Finally, I used the "plot" function for creating graphs, and the "qqplot" function to create normal probability plots.
- In this report, I have detailed the 3 different parts of the project, ranging from part I to part III. I used the "echo = FALSE" command to hide the code from the project so it only shows the results. I created an Appendix at the end of the project where I have attached the code and screen-shots of the outputs of the code that is not shown in the project. In the Appendix, I used the "results = 'hide' " command to hide the results.

Part I: Multiple linear regression I

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

```
##
##   The decimal point is 6 digit(s) to the right of the |
##
##   0 | 11111111111111111111111111111111111111111111111111111111111111111+254
##   0 | 55555555555555555555556666666666666667777777777777777778888888888
##   1 | 000000122233333444
##   1 | 55699
##   2 | 1134
##   2 | 58
##   3 |
##   3 |
##   4 |
##   4 |
##   5 | 1
##   5 |
##   6 |
##   6 |
##   7 |
##   7 |
##   8 |
##   8 | 9
```

- This stem-and-leaf plot of the total population is the first plot of Model 1. It is heavily right-skewed. We can see that the majority of data points are 6-digits, with a few 7-digit numbers. The highest value can easily be seen in this stem-and-leaf plot which is rounded to 890,0000. This shows that the total population in each county skews towards the lower end.

```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 | 000011111111111122222222222222222223333333333333333333333333444444+252
##    1 | 00000000000000011111111111112222222222233333344445556666777788899999
##    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 stem-and-leaf plot of the land area is the second plot of Model 1. It is right-skewed, but not as much as the stem-and-leaf plot of the total population. This is because we can clearly see more numbers towards the center of the plot while there were relatively no data points towards the center in the plot of the total population. The highest value can easily be seen in this stem-and-leaf plot which is rounded to 20,100. This plot shows that land area of each county skews towards the lower end.

```
##
##   The decimal point is 4 digit(s) to the right of the |
##
##    0 | 11111111111112222222222222222222222222222222222222222222222222222+263
##    1 | 00000000000011111111122222333334444444555555556778888888999
##    2 | 001111233344477788899
##    3 | 0255678899
##    4 | 19
##    5 | 59
##    6 |
##    7 |
##    8 |
##    9 |
##   10 |
##   11 | 1
##   12 |
##   13 |
##   14 |
##   15 |
##   16 |
##   17 |
##   18 | 4
```

- This stem-and-leaf plot of total personal income is the third plot of Model 1. It is very right-skewed - more right-skewed than the stem-and-leaf plot of land area but not as much as the stem-and-leaf plot of the total population. The highest value can easily be seen in this stem-and-leaf plot which is rounded to 184,000. This plot shows that the total personal income is skewed towards the lower end.

```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 | 000000000000000011111111111111111111111111111111111111111111111111111+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 stem-and-leaf plot of population density is the first plot of Model 2. It is fairly right-skewed. This plot shows that the population density is fairly low across all counties with the highest population density at 32,400.

```
##
##   The decimal point is at the |
##
##    2 | 0
##    4 | 47890389
##    6 | 112345567799013456667899
##    8 | 00112222233334444555666777778888899990002222333333444444445555666677
##   10 | 000111111222222222233333344444455555556666666677777778888888888999999+36
##   12 | 000000001111122223333333333444455555556666667777777778888999000000000+36
##   14 | 000011111112233344444555677889000000111122223455667778
##   16 | 12556699901122345
##   18 | 06778
##   20 | 070
##   22 | 018828
##   24 | 47
##   26 | 055
##   28 | 1
##   30 | 7
##   32 | 138
```

- This stem-and-leaf plot of the percent of population 65 or older is the second plot of Model 2. It is fairly symmetrical, and is easily the most symmetrical plot out of all the predictor variables so far. There is a slight right-skew. It's clear through this plot that the percentage of the population 65 or older is much more normally distributed throughout all counties and isn't heavily skewed towards one side. The highest number this plot shows us is 32.8%. It doesn't show the highest number to be 33.8% (which is the actual number) as the axis of this plot goes up by 2. The average percentage from this stem-and-leaf plot seems to be between 12% and 13%.

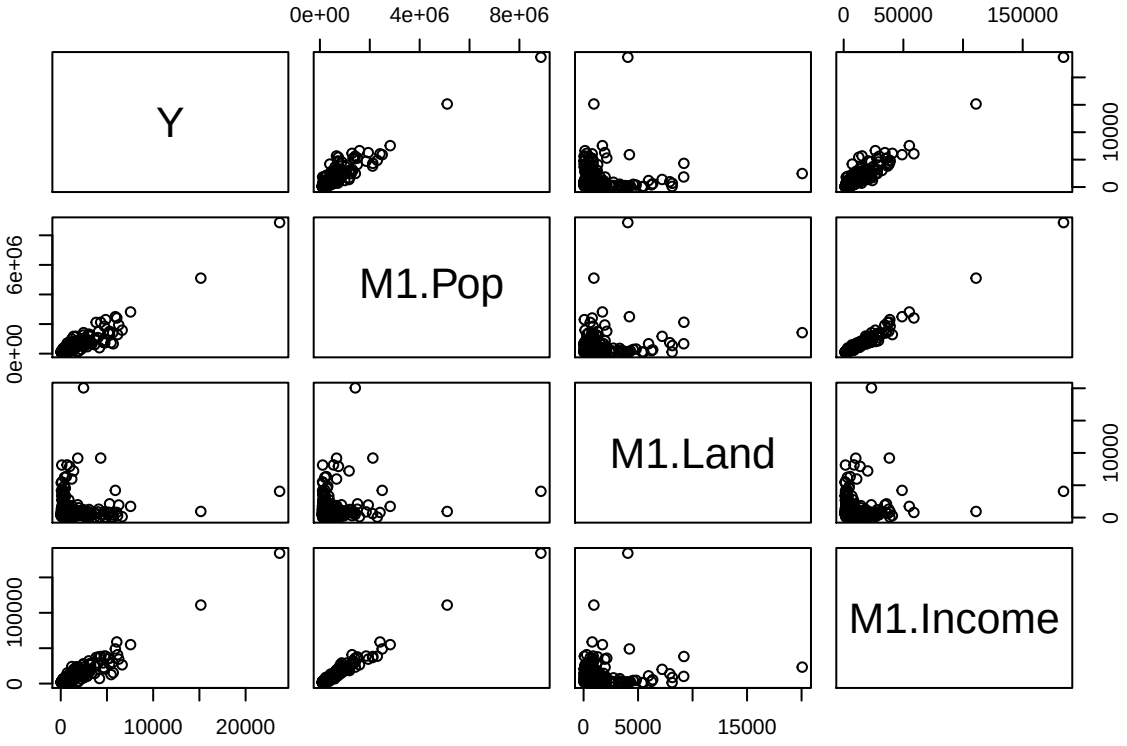
```
##
##   The decimal point is 4 digit(s) to the right of the |
##
##    0 | 11111111111112222222222222222222222222222222222222222222222222222+263
##    1 | 000000000000111111111222223333344444455555555567788888888999
##    2 | 001111233344477788899
##    3 | 0255678899
##    4 | 19
##    5 | 59
##    6 |
##    7 |
##    8 |
##    9 |
##   10 |
##   11 | 1
```

```
##    12 |
##    13 |
##    14 |
##    15 |
##    16 |
##    17 |
##    18 | 4
```

- This stem-and-leaf plot of total personal income is the third plot of Model 2. It is the same as the stem-and-leaf plot of the previous total personal income plot from Model 1. It's very right-skewed compared to the other two plots in Model 2. The highest value can easily be seen in this stem-and-leaf plot which is rounded to 184,000. This plot shows that the total personal income is skewed towards the lower end.

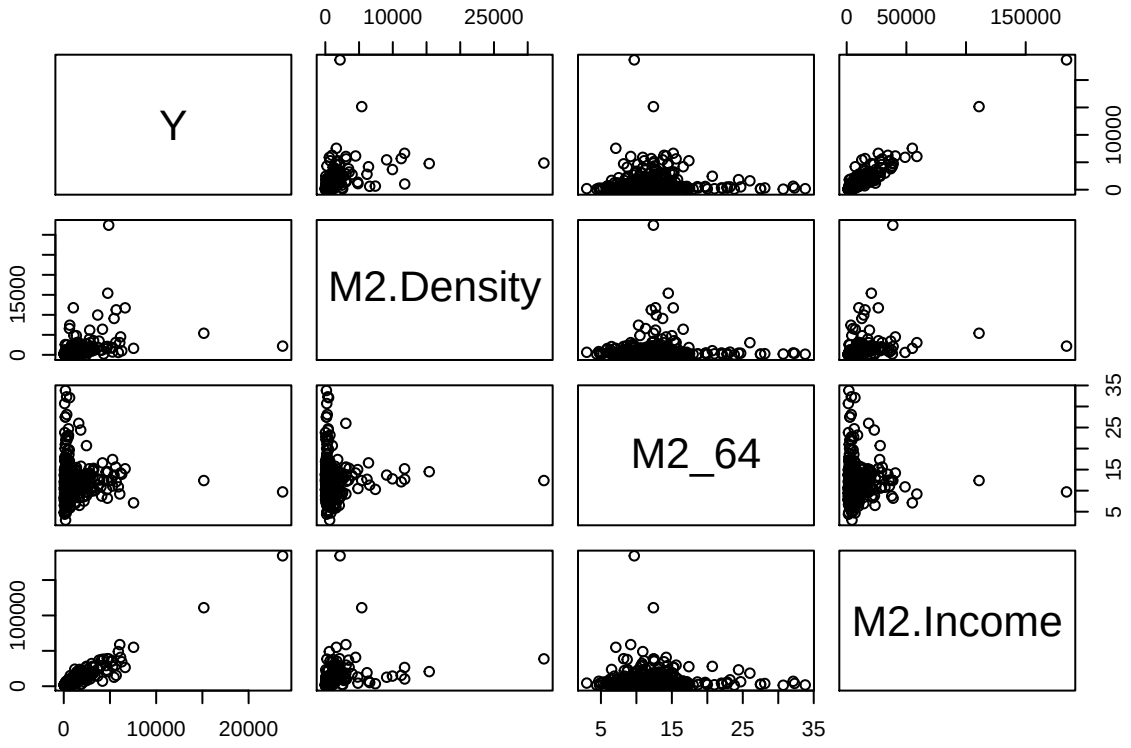
b) Obtain the scatter plot matrix and the correlation matrix for each proposed model. Summarize the information provided.

Model 1 Scatter Plot Matrix



- From this scatterplot matrix, it seems as if there is a correlation between "total population" and "number of active physicians". There also seems to be a correlation between "total personal income" and "number of active physicians". This is because both plots look like a straight line. There does not appear to be a correlation between "land area" and "number of active physicians". Although this scatterplot is useful for observations, further tests must occur to test correlation.

Model 2 Scatter Plot Matrix



- From this scatterplot matrix, it seems as if there is a solid correlation between "total personal income" and "number of active physicians". There does not appear to be any correlation between "population density" and "number of active physicians" nor "percent of population 65 or older" and "number of active physicians". Although this scatterplot is useful for observations, further tests must occur to test correlation.

Model 1 Correlation Matrix

```
##                    Y    M1.Pop    M1.Land M1.Income
## Y          1.00000000 0.9402486 0.07807466 0.9481106
## M1.Pop     0.94024859 1.0000000 0.17308335 0.9867476
## M1.Land    0.07807466 0.1730834 1.00000000 0.1270743
## M1.Income  0.94811057 0.9867476 0.12707426 1.0000000
```

- For the correlation matrix of Model 1, it shows that "total personal income" has the largest correlation with the number of active physicians. It also shows that "land area" has the lowest correlation with the number of active physicians.

Model 2 Correlation Matrix

```
##                     Y M2.Density       M2_64   M2.Income
## Y           1.00000000 0.40643863 -0.00312863  0.94811057
## M2.Density  0.40643863 1.00000000  0.02918445  0.31620475
## M2_64      -0.00312863 0.02918445  1.00000000 -0.02273315
## M2.Income   0.94811057 0.31620475 -0.02273315  1.00000000
```

- For the correlation matrix of Model 2, it shows that "total personal income" has the largest correlation with the number of active physicians. It also shows that "percentage of population 65 or older" has the lowest correlation with the number of active physicians.

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

Model 1: $\hat{Y}$ = -13.31615 + 0.0008366178(X_1) - 0.06552296(X_2) + 0.09413199(X_3)

Model 2: $\hat{Y}$ = -170.57422325 + 0.09615889(X_1) + 6.33984064(X_2) + 0.12656649(X_3)

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

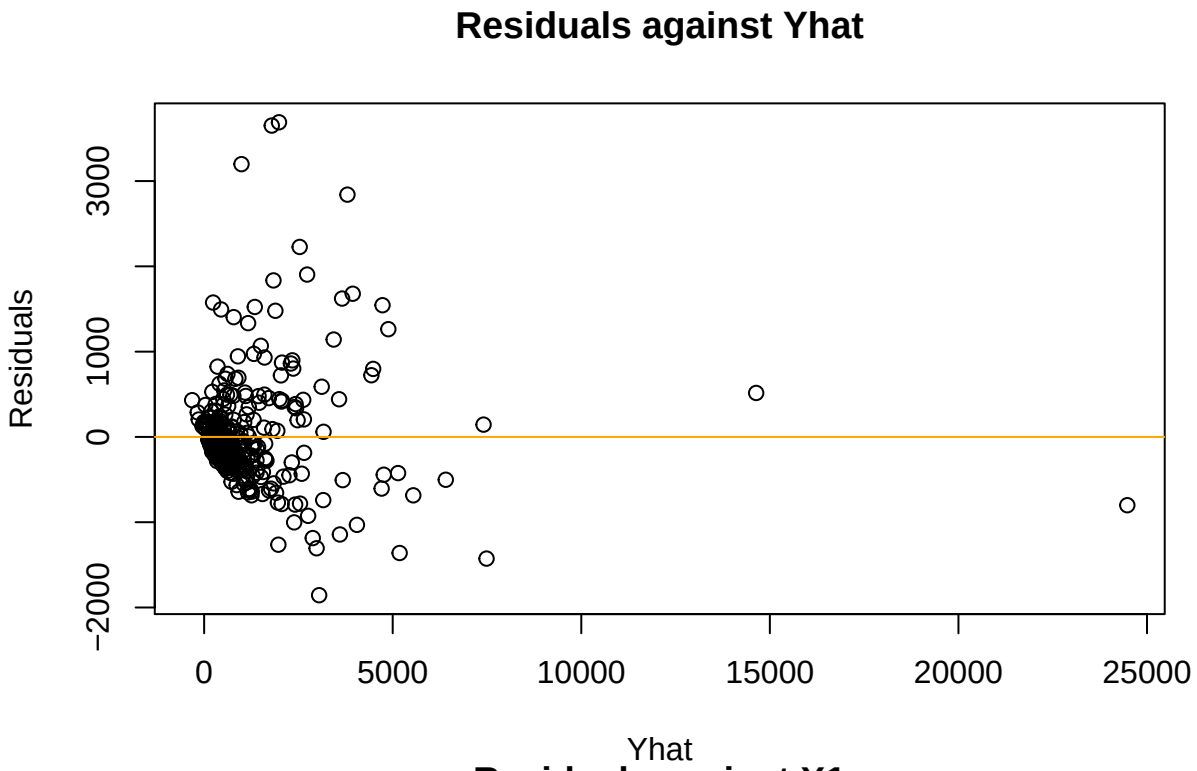
R^2 value for Model 1 is 0.9026432

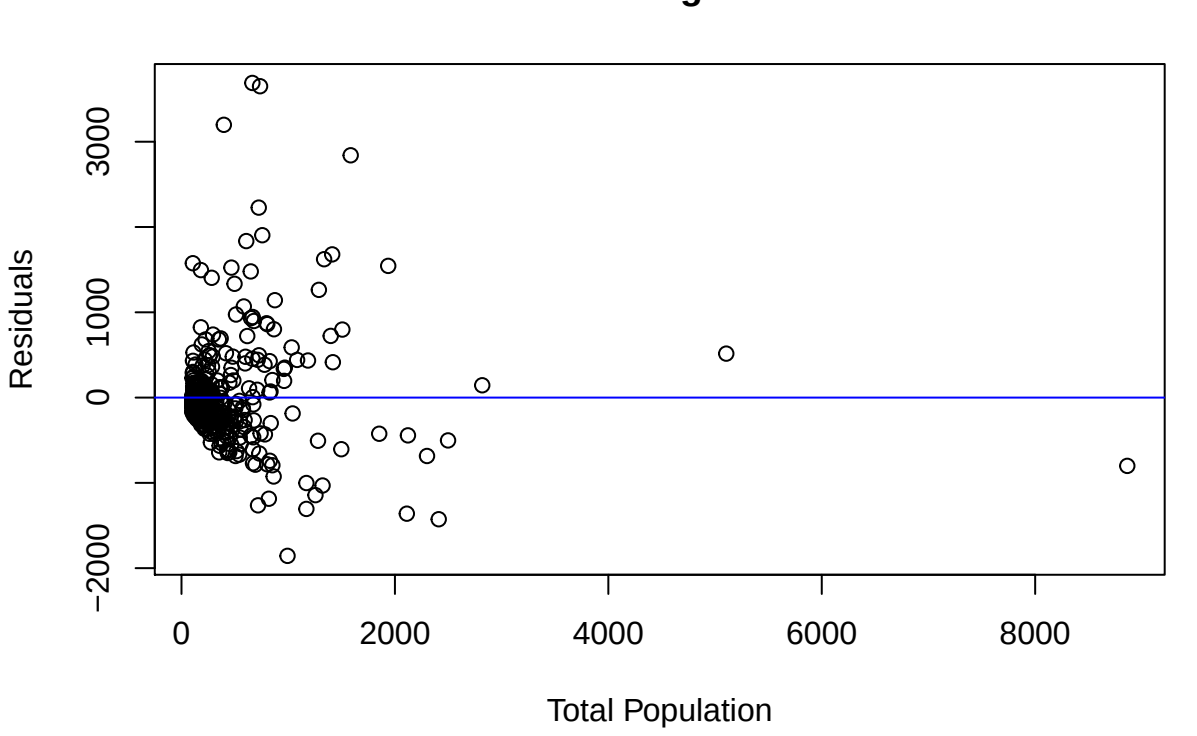
R^2 value for Model 2 is 0.9117491

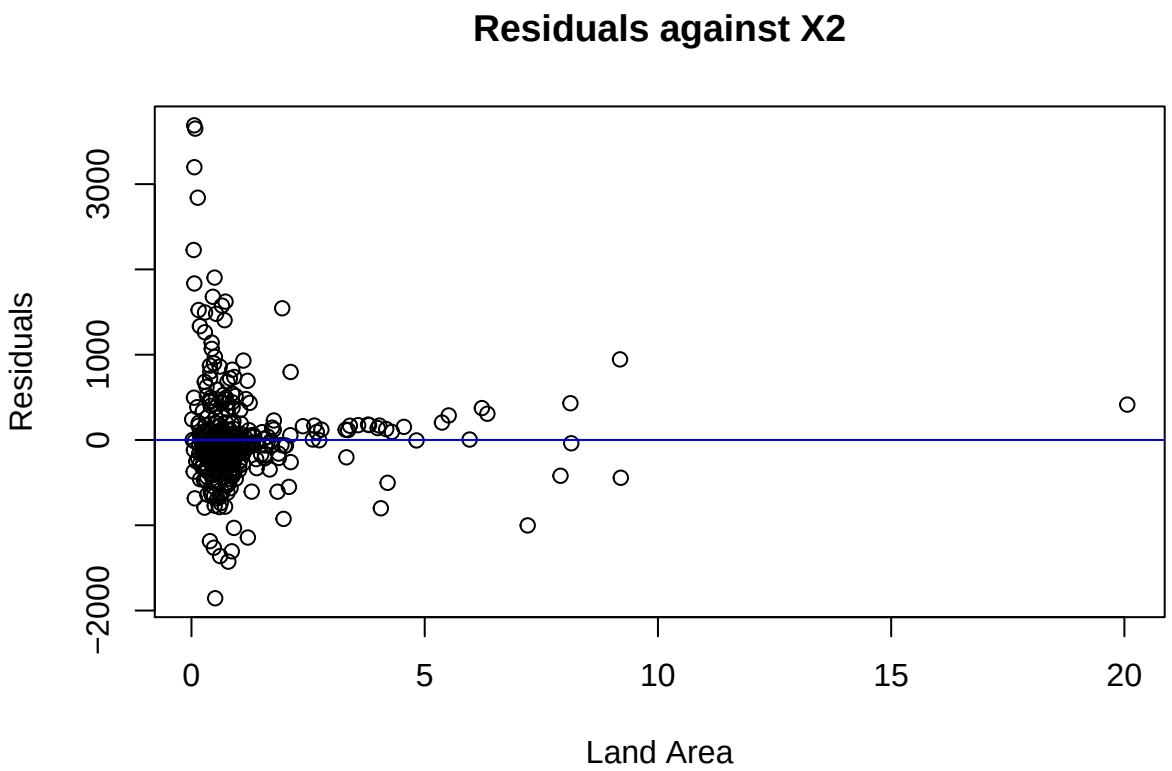
Model 2 is clearly preferable as evident by the higher R^2 value.

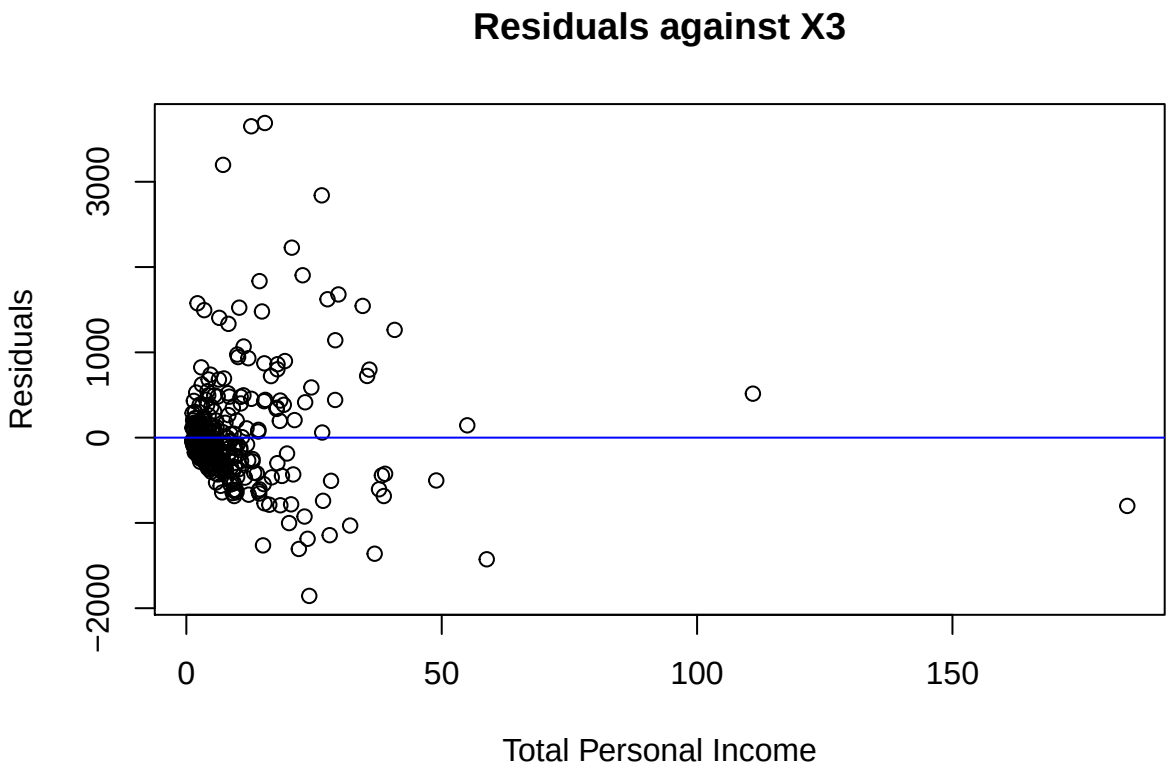
e) For each model, obtain the residuals and plot them against $\hat{Y}$, each of the three predictor variables, and each of the two-factor interaction terms. Also prepare a normal probability plot for each of the two fitted models. Interpret your plots and state your findings. Is one model clearly preferable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

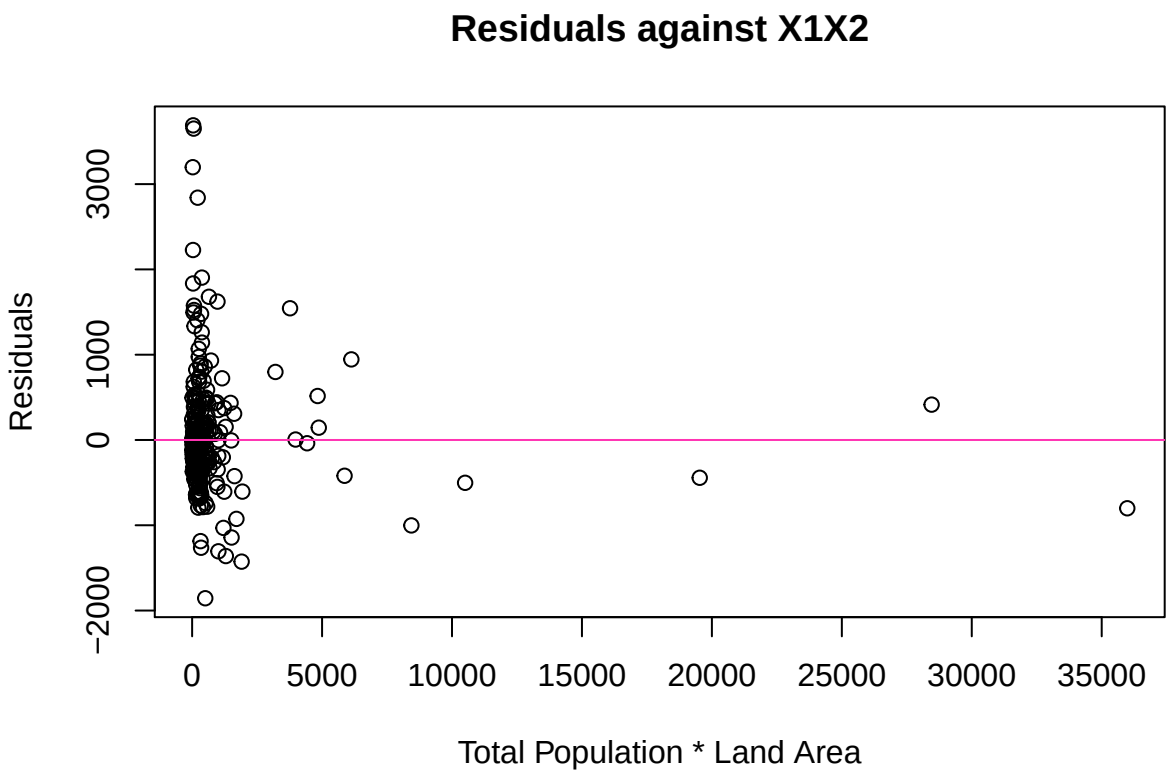
Model 1 Plots

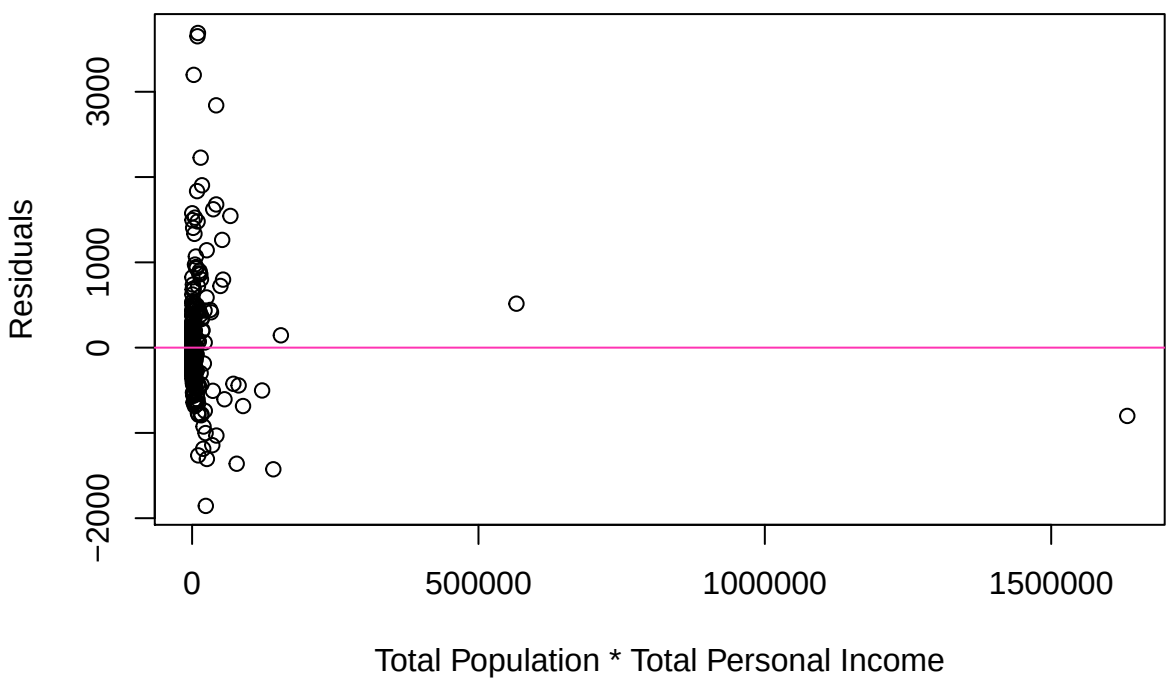


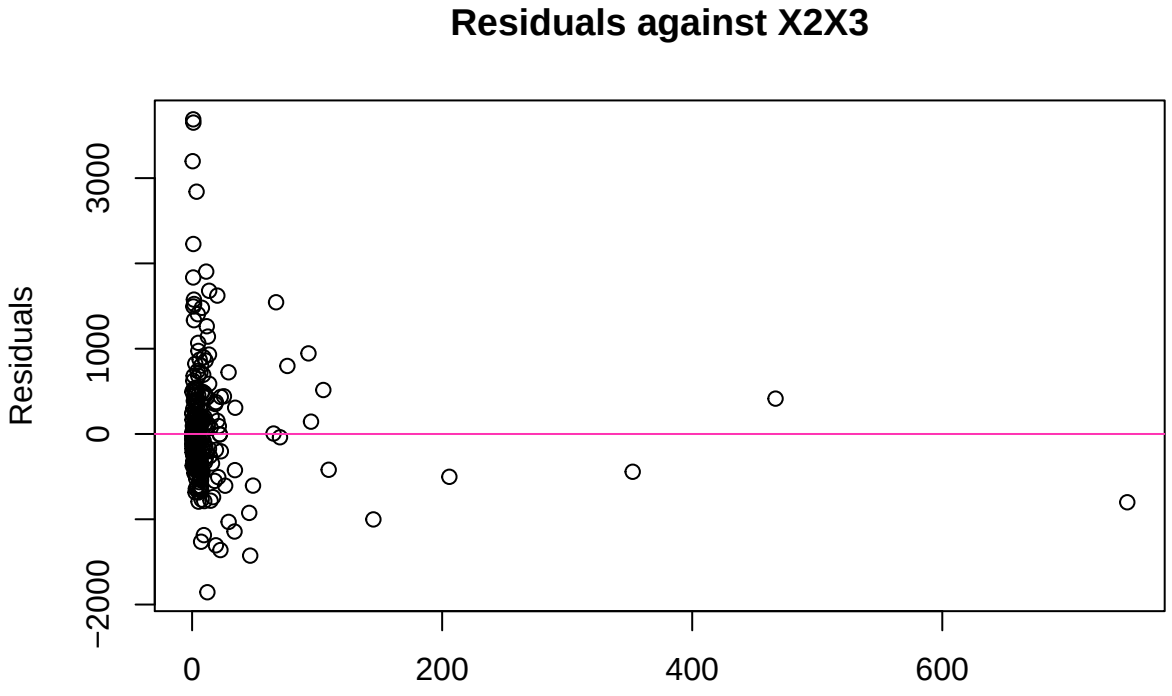


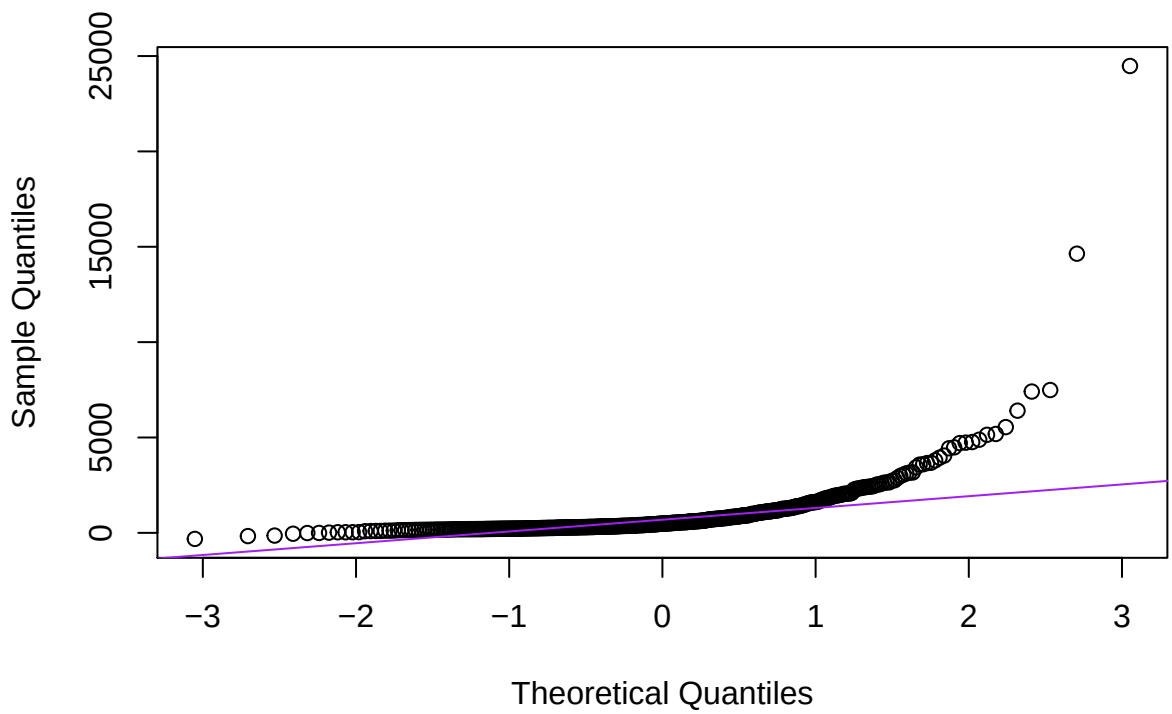

Residuals against X2
Residuals
3000
1000
0
−2000
0
5
10
15
20
Land Area


Residuals against X3
Residuals
3000
1000
0
−2000
0
50
100
150
Total Personal Income


Residuals against X1X2
Residuals
3000
1000
0
−2000
0
5000
10000
15000
20000
25000
30000
35000
Total Population * Land Area


Residuals against X1X3
Residuals
3000
1000
0
−2000
0
500000
1000000
1500000
Total Population * Total Personal Income



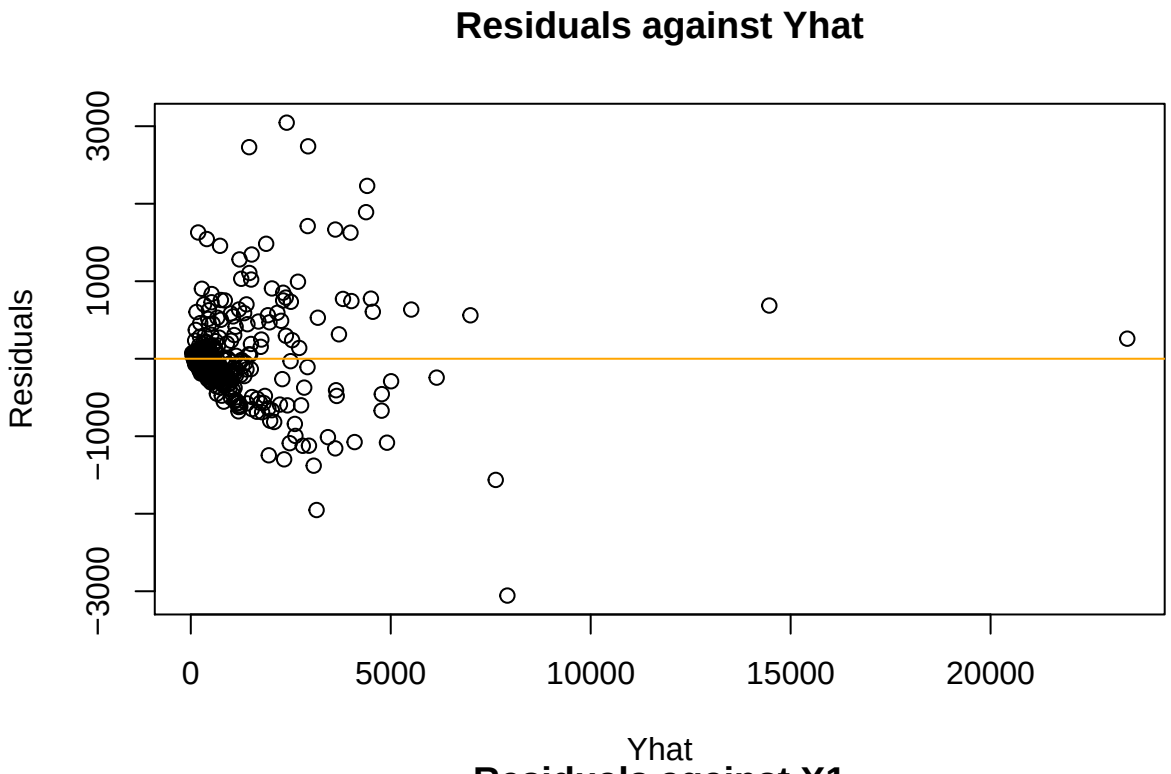


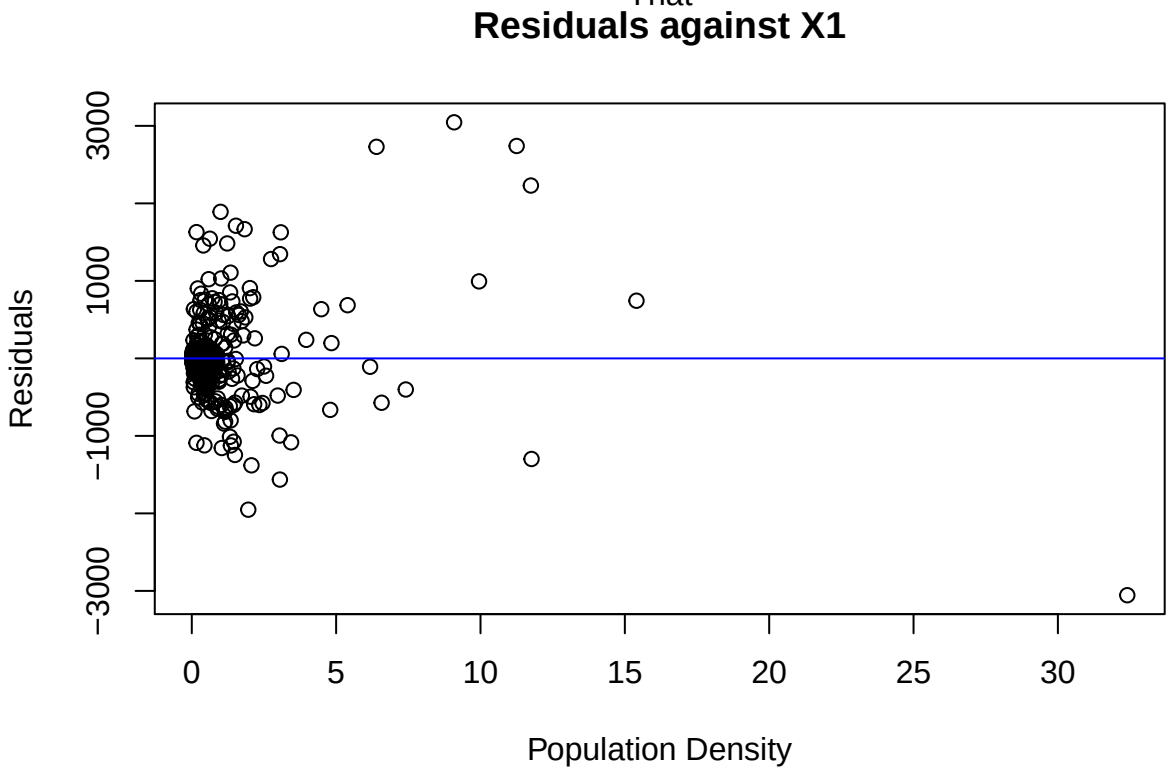
- The residual plot of the residuals against $\hat{Y}$ looks normal with constant variance and randomness.
- The residual plot of the residuals against the three predictor variables also look normal with constant variance and randomness, however the residual plot against X_2 looks much different that the other two. The residual plot of X_2 is more clustered and less scattered than either X_1 or X_3.
- The residual plot of the residuals against the three interaction terms look much more different than

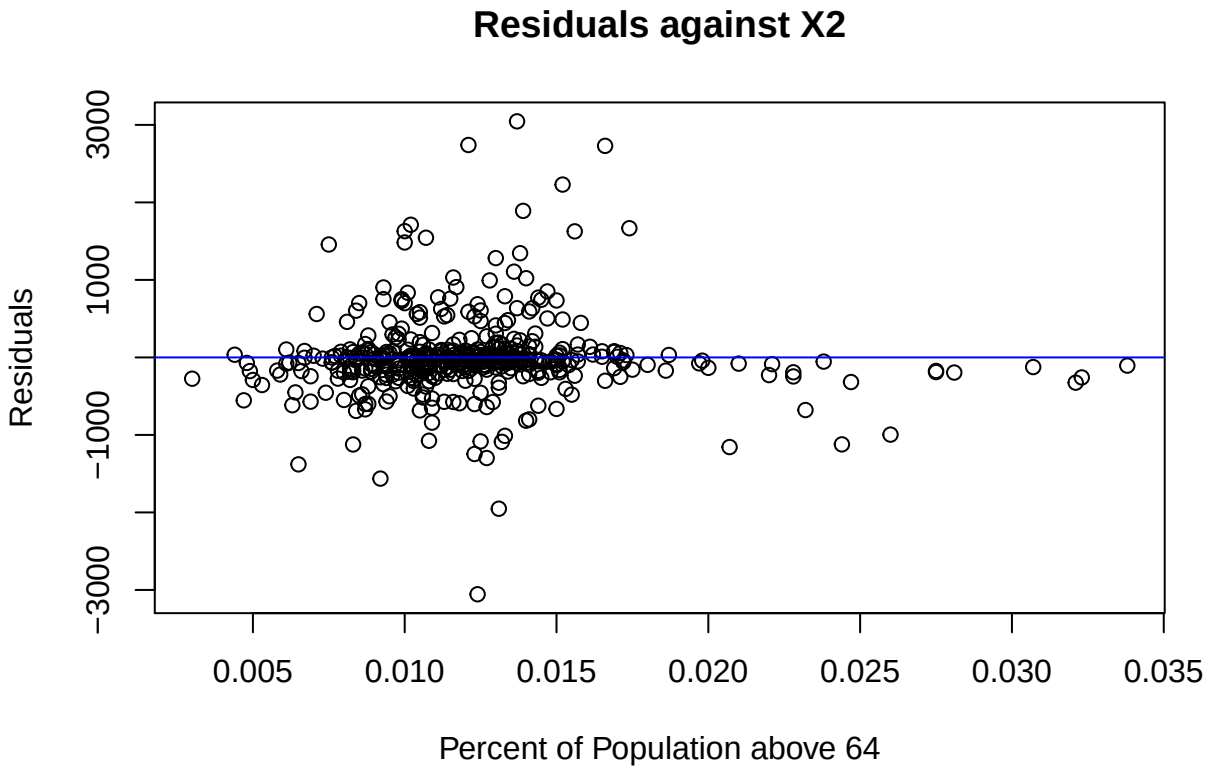
the four previous plots. It's more clustered than the previous plots, and the interaction term of X_1X_2 seems to have much more outliers. However, the three plots do seem to pass the test of constant variance and randomness.

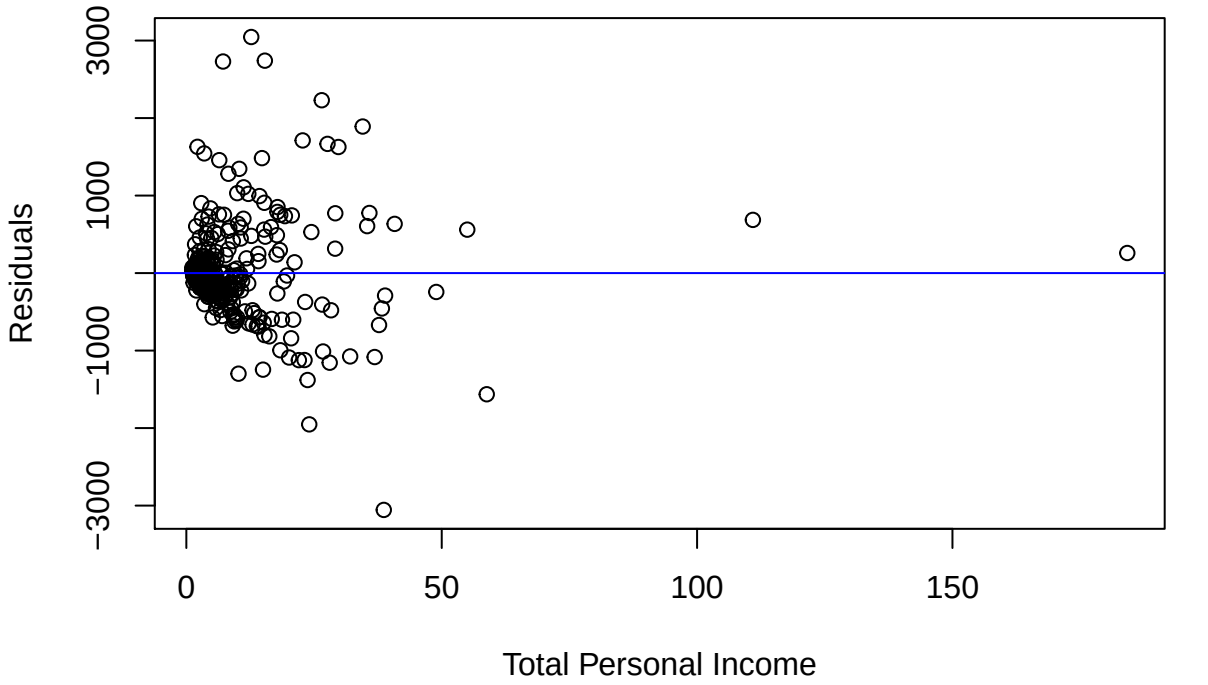
- The normal probability plot of the fitted regression line for model 1 skews to the right but appears to be normal.

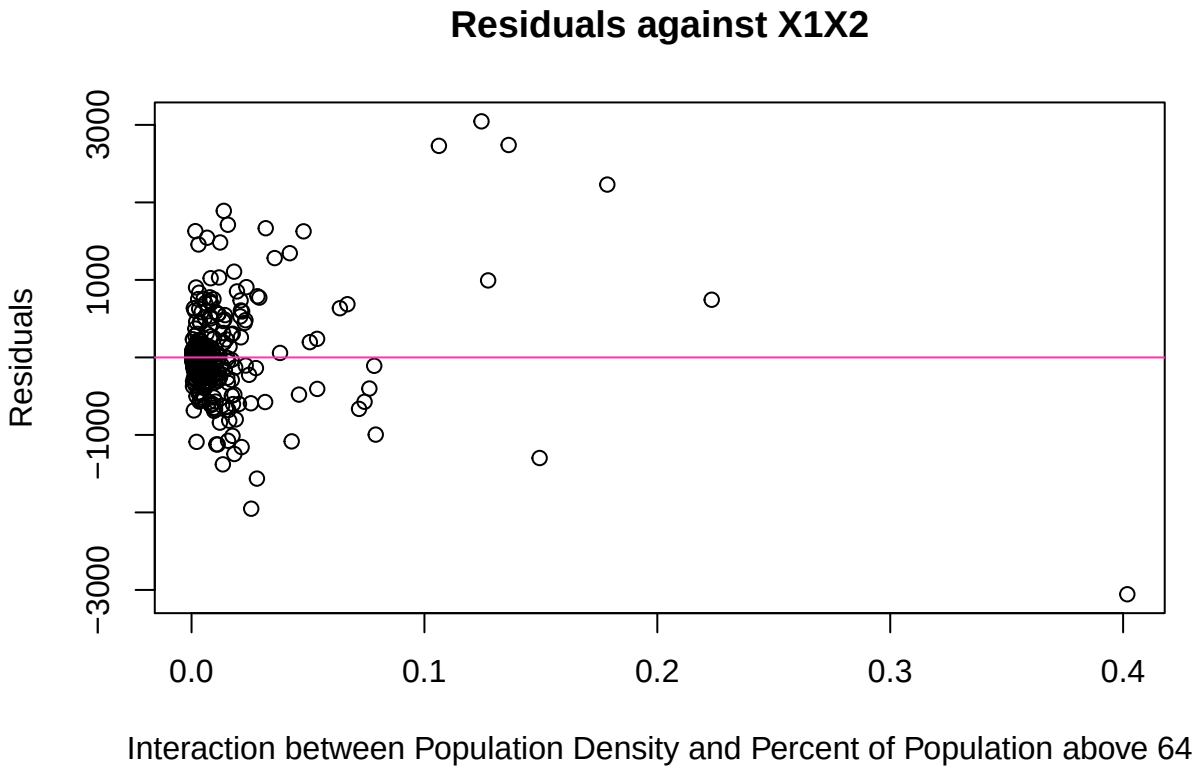
Model 2 Plots

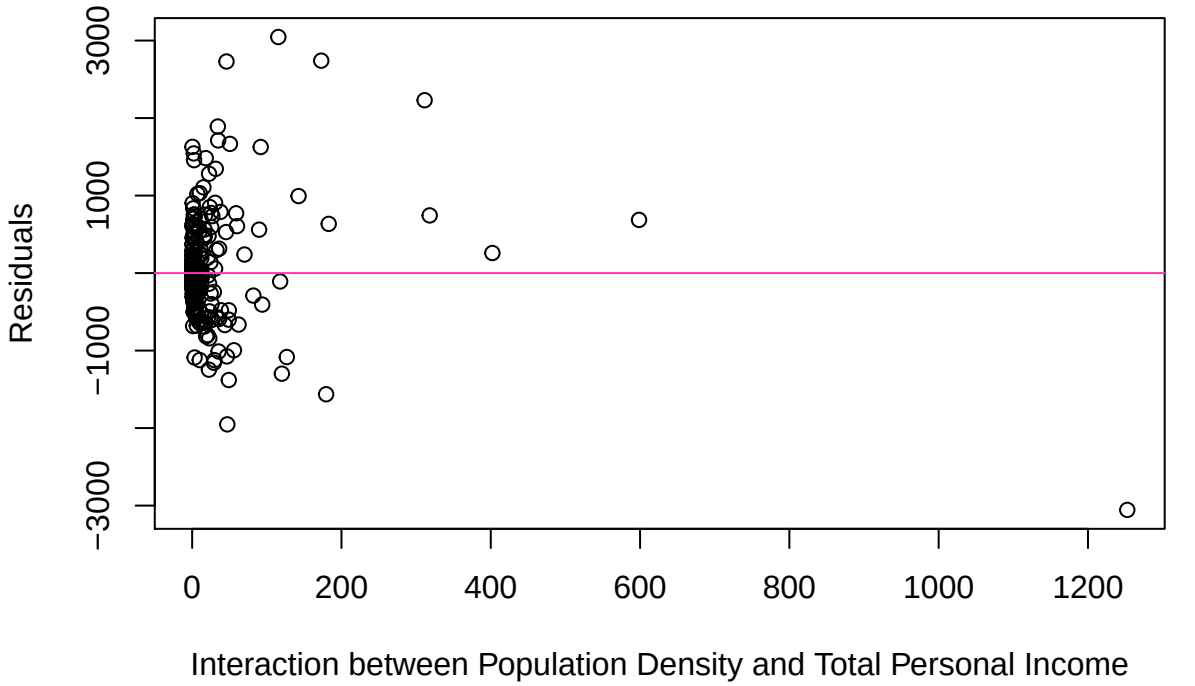


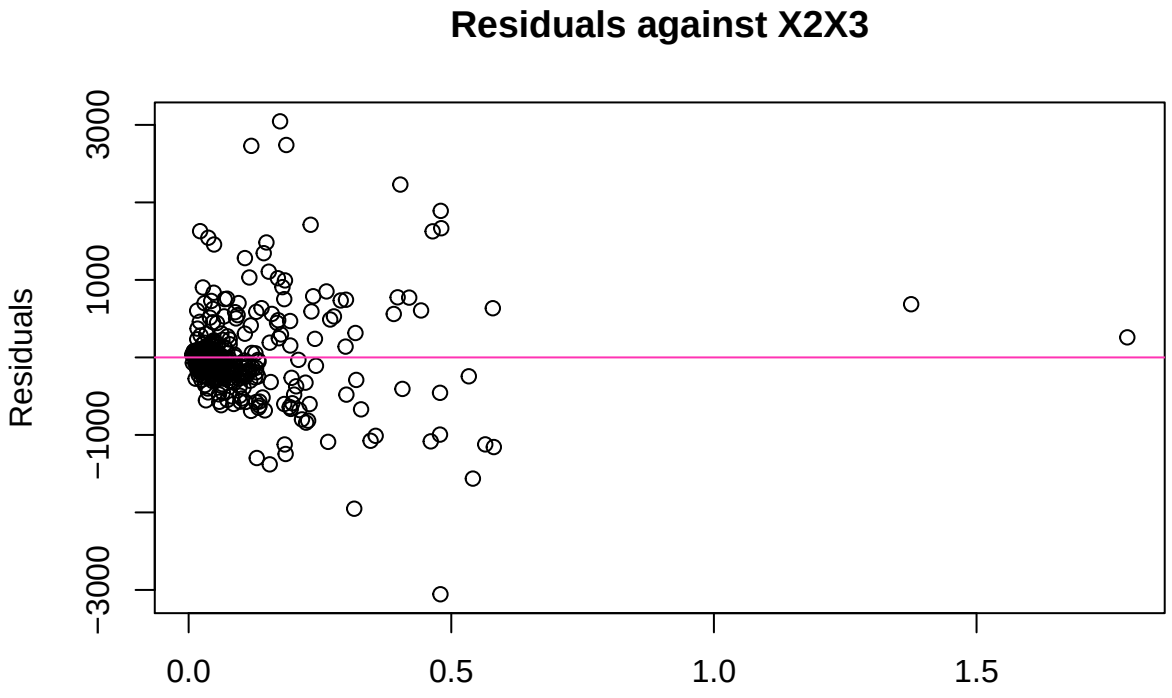


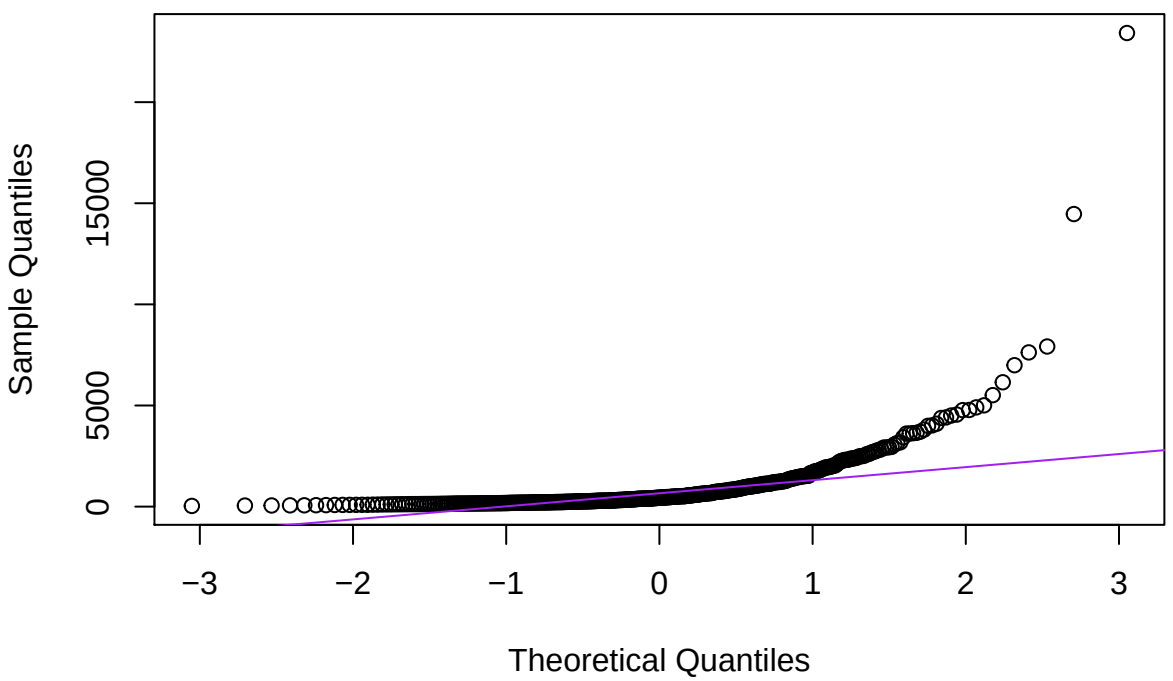

Residuals against X2
Residuals
3000
1000
−1000
−3000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
Percent of Population above 64


Residuals against X3
Residuals
3000
1000
−1000
−3000
0
50
100
150
Total Personal Income


Residuals against X1X2
Residuals
3000
1000
−1000
−3000
0.0
0.1
0.2
0.3
0.4
Interaction between Population Density and Percent of Population above 64


Residuals against X1X3
Residuals
3000
1000
−1000
−3000
0
200
400
600
800
1000
1200
Interaction between Population Density and Total Personal Income





- The residual plot of the residuals against Model 2's $\hat{Y}$ looks normal with constant variance and randomness.
- The residual plot of Model 2's residuals against the three predictor variables also look normal with constant variance and randomness. It's noteworthy that the residual plot against X_2 is far more

symmetrical than the other two predictor variables. Unlike the plots of Model 1, no residual plot against the three predictor variable looks overly clustered on one side or heavily skewed.

- The residual plot of the residuals against the three interaction terms look different than the four previous plots. The residual plot against X_1X_3 is far more clustered, than the other plots. The three plots do seem to pass the test of constant variance and randomness.
- The normal probability plot of the fitted regression line for Model 2 skews to the right but appears to be normal.
- The plots between Model 1 and Model 2 appear similar with a few differences. One difference is that the residual plots against the predictor variables in Model 2 seems to be slightly more symmetrical than those from Model 1. A much bigger difference is in the interaction terms, where Model 2's residual plots against interaction terms are far more symmetrical than those of Model 1. Finally, the normal probability plots of both models seem similar, with Model 2 being slightly more linear.
- Between the two plots, Model 2 is clearly preferable in terms of appropriateness. This can also be proven by calculating the correlation test for normality. The correlation coefficient between Model 1 and number of active physicians (Y) is equal to 0.8690221. The correlation coefficient between Model 2 and number of active physicians (Y) is equal to 0.8940804. With a higher correlation coefficient, Model 2 is more preferable to Model 1.

f) Now expand both models proposed above by adding all possible two-factor interactions. Repeat part d for the two expanded models.

Model 1: $\hat{Y}$ = -70.35217 + 0.0005108667(X_1) + 0.0172478(X_2) + 0.09797127(X_3) - 0.0000001603799(X_4) + 0.00000003954438(X_5) - 0.0000002169211(X_6)

Model 2: $\hat{Y}$ = -9.367002 - 0.4179492(X_1) - 11.05857(X_2) + 0.147717(X_3) + 0.04652245(X_4) - 0.000003276354(X_5) - 0.001288565(X_6)

R^2 value for Model 1 is 0.9063789

R^2 value for Model 2 is 0.9230238

Model 2 is clearly preferable as evident by the higher R^2 value.

Part II: Multiple linear regression II

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), and number of hospital beds (X_5).

$R^2_{X_3|X_1,X_2} = 0.02882495$

$R^2_{X_4|X_1,X_2} = 0.003842367$

$R^2_{X_5|X_1,X_2} = 0.5538182$

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

Extra SS of X_3 = 4,063,370

Extra SS of X_4 = 541,647.3

Extra SS of X_5 = 78,070,132

Based on our results in part A, it's clear that the best additional predictor variable is X_5 (number of hospital beds) as it has the highest coefficient of partial determination. Yes, the extra sum of squares associated with variable X_5 is larger than those for the other two variables.

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 α = .0l. State the alternatives, decision rule, and conclusion. Would the F* test statistics for the other two potential predictor variables be as large as the one here?

H0: B5 = 0

HA: B5 != 0

F* statistic = 541.1801

Critical Value = 6.693358

Since the F* statistic is larger than the critical value, we reject the null hypothesis and conclude that the variable X_5 is helpful in the regression model when X_1 and X_2 are included in the model.

As the partial coefficient of determination was the largest for X_5, it's obvious that the F* statistic when adding X_5 would be larger than either X_3 or X_4. This can also be proved by solving the F* statistic of the other two predictor variables. The F* statistic of adding X_3 into the model is 12.94069 and the F* statistic of adding X_4 into the model is 1.681734. Therefore, the F* test statistics for the other two potential predictor variables will not be as large as the one with X_5.

d) Compute three additional coefficients of partial determination: $R^2_{Y,X_3,X_4|X_1,X_2}$, $R^2_{Y,X_3,X_5|X_1,X_2}$, and $R^2_{Y,X_4,X_5|X_1,X_2}$. 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

$R^2_{Y,X_3,X_4|X_1,X_2}$ = 0.03314181

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

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

The pair of X_4, X_5 is relatively more important than the other pairs as it has the highest R^2 value.

H0: B4 = B5 = 0

HA: B4 != B5 != 0

F* statistic = 281.6688

Critical Value = 3.016458

Since the F* statistic is larger than the critical value, we reject the null hypothesis and conclude that adding the pair of X_4,X_5 to the model is helpful given that X_1,X_2 are already in the model.

Part III: Discussion

- For Part 1, I created stem-and-leaf plots for each predictor variable and created a scatter plot matrix and correlation matrix for both models. After calculating the R^2 value (coefficient of determination) for each model, I found that the R^2 value for Model 2 (0.9026432) was higher than that of Model 1

(0.9117491). Next, I obtained residual plots for each model which were then plotted against numerous variables. Finally, I calculated the R^2 value for the models when interaction terms were added to the regression function and found that Model 2 still had a higher R^2 than Model 1. After completing this part, it showed me that Model 2 was the superior model. The most challenging piece of Part 1 was R's inability to work with very large numbers. When calculating the interaction terms, many terms showed up as "NA". To solve this problem, I divided each predictor variable by 1000 so the output would have smaller numbers, yet still be proportional.

- For Part 2, to find the answer to the question "which additional variable would be the most helpful to the regression model", I calculated the partial coefficient of determination of each predictor variable when X_1 and X_2 are present in the model. I found that the partial coefficient of determination was highest when X_5 was included in the model, and lowest when X_3 was included. I then chose X_5 as the best additional predictor variable and conducted an F-test to test whether or not including it in the regression model is helpful. Through the F-test, I concluded that it is helpful in the model. Finally, I calculated the partial coefficient of determination for pairs of additional predictor variables when X_1 and X_2 are present in the model. The pair which had the highest coefficient of partial determination was the pair of (X_4,X_5) and the lowest was the pair of (X_3,X_4). Once finding the best pair, I conducted an F-test to test whether or not the pair of variables I chose is helpful in the regression model. The F-test showed me that the pair of (X_4,X_5) was indeed helpful.

- The entire course material was useful to me for completing this project. The lecture summaries were vital as I was able to write R code using the formulas of extra sums of squares and the coefficient of partial determination. Discussions were very helpful as it taught me how to use the $lm()$ function, find coefficients, plot residual & normal probability plots, and find scatterplot matrices. All in all, the lectures for multiple linear regression helped me become confident with my calculations, and discussion sections helped with becoming confident in my R coding.

- Finally, to improve the linear regression models even more, I can test more variables to see which additional predictor would help the regression model the most. I can also create additional models with different predictors to compare with models 1 and 2.

Appendix

Project 6.28 code

a)

```
CDI = read.table("CDI.txt")
Y = CDI$V8
M1.Pop = CDI$V5
M1.Land= CDI$V4
M1.Income = CDI$V16
n = length(Y)

M2.Density = CDI$V5/CDI$V4
M2_64 = CDI$V7
M2.Income = CDI$V16

stem(M1.Pop)
stem(M1.Land)
stem(M1.Income)
stem(M2.Density)
```

```
stem(M2_64)
stem(M2.Income)
```

b)

```
CDImodel1 = data.frame(M1.Pop, M1.Land, M1.Income)
CDImodel2 = data.frame(M2.Density, M2_64, M2.Income)
pairs(CDImodel1)
```

```
pairs(CDImodel2)
```

```
cor(CDImodel1)
cor(CDImodel2)
```

c)

```
fit.model1 = lm(Y~M1.Pop + M1.Land + M1.Income, data =CDI)
fit.model2 = lm(Y~M2.Density + M2_64 + M2.Income, data=CDI)
beta.model1 = fit.model1$coefficients
beta.model2 = fit.model2$coefficients

Yhat.model1 = beta.model1[1] + beta.model1[2]*M1.Pop + beta.model1[3]*M1.Land + beta.model1[4]*M1.Income

Yhat.model2 = beta.model2[1] + beta.model2[2]*M2.Density + beta.model2[3]* M2_64 + beta.model2[4]*M2.In

beta.model2
beta.model1
```

d)

```
SSE_model1 = sum((Y - Yhat.model1)^2)
SSR_model1 = sum((Yhat.model1 - mean(Y))^2)
SSTO_model1 = SSE_model1 + SSR_model1
RSquared_model1 = SSR_model1/SSTO_model1
RSquared_model1

SSE_model2 = sum((Y - Yhat.model2)^2)
SSR_model2 = sum((Yhat.model2 - mean(Y))^2)
SSTO_model2 = SSE_model2 + SSR_model2
RSquared_model2 = SSR_model2/SSTO_model2
RSquared_model2
```

e)

```
res_model1 = fit.model1$residuals
res_model2 = fit.model2$residuals

M1.Pop = M1.Pop/1000
M1.Land = M1.Land/1000
M1.Income = M1.Income/1000

M2.Density = M2.Density/1000
M2_64 = M2_64/1000
M2.Income = M2.Income/1000

model1.x1x2 = (M1.Pop *M1.Land)
model1.x1x3 = (M1.Pop * M1.Income)
model1.x2x3 = (M1.Land * M1.Income)

model2.x1x2 = (M2.Density * M2_64)
model2.x1x3 = (M2.Density * M2.Income)
model2.x2x3 = (M2_64 * M2.Income)

res_m1 <- qqnorm(res_model1, plot.it = FALSE)
r_m1 = cor(res_m1$x, res_m1$y)
r_m1
res_m2 <- qqnorm(res_model2, plot.it = FALSE)
r_m2 = cor(res_m2$x, res_m2$y)
r_m2

plot(Yhat.model1, res_model1)
abline(h=0, col='red')
```

```
plot(M1.Pop, res_model1)
abline(h=0, col='red')
```

```
plot(M1.Land, res_model1)
abline(h=0, col='red')
```

```
plot(M1.Income, res_model1)
abline(h=0, col='red')
```

```
plot(model1.x1x2, res_model1)
abline(h=0, col='red')
```

```
plot(model1.x1x3, res_model1)
abline(h=0, col='red')
```

```
plot(model1.x2x3, res_model1)
abline(h=0, col='red')
```

```
qqnorm(Yhat.model1)
qqline(Yhat.model1, col='red')
```

```
plot(Yhat.model2, res_model2)
abline(h=0, col='red')
```

```
plot(M2.Density, res_model2)
abline(h=0, col='red')
```

```
plot(M2_64, res_model2)
abline(h=0, col='red')
```

```
plot(M2.Income, res_model2)
abline(h=0, col='red')
```

```
plot(model2.x1x2, res_model2)
abline(h=0, col='red')
```

```
plot(model2.x1x3, res_model2)
abline(h=0, col='red')
```

```
plot(model2.x2x3, res_model2)
abline(h=0, col='red')
```

```
qqnorm(Yhat.model2)
qqline(Yhat.model2, col='red')
```

f)

```
newfit.model1 = lm(Y ~ M1.Pop + M1.Land + M1.Income + model1.x1x2 + model1.x1x3 + model1.x2x3, data =CD

newfit.model2 = lm(Y~M2.Density + M2_64 + M2.Income + model2.x1x2 + model2.x1x3 + model2.x2x3, data=CDI

newbeta.model1 = newfit.model1$coefficients
newbeta.model2 = newfit.model2$coefficients

NewYhat.model1 = newbeta.model1[1] + newbeta.model1[2]*M1.Pop + newbeta.model1[3]*M1.Land + newbeta.mod

NewYhat.model2 = newbeta.model2[1] + newbeta.model2[2]*M2.Density + newbeta.model2[3]* M2_64 + newbeta.

NewSSE_model1 = sum((Y - NewYhat.model1)^2)
NewSSR_model1 = sum((NewYhat.model1 - mean(Y))^2)
NewSSTO_model1 = NewSSE_model1 + NewSSR_model1
NewRSquared_model1 = NewSSR_model1/NewSSTO_model1
NewRSquared_model1

NewSSE_model2 = sum((Y - NewYhat.model2)^2)
NewSSR_model2 = sum((NewYhat.model2 - mean(Y))^2)
NewSSTO_model2 = NewSSE_model2 + NewSSR_model2
NewRSquared_model2 = NewSSR_model2/NewSSTO_model2
NewRSquared_model2
```

Project 7.37 code

a)

```
X1 = CDI$V5
X2 = CDI$V16
X3 = CDI$V4
X4 = CDI$V7
X5 = CDI$V9
Y = CDI$V8

typeX1X2 = lm(Y~X1+X2, data=CDI)
typeX1X2X3 = lm(Y~X1+X2+X3, data=CDI)
typeX1X2X4 = lm(Y~X1+X2+X4, data=CDI)
typeX1X2X5 = lm(Y~X1+X2+X5, data=CDI)
typeX1X2X3X4 = lm(Y~X1+X2+X3+X4, data=CDI)
typeX1X2X3X5 = lm(Y~X1+X2+X3+X5, data=CDI)
typeX1X2X4X5 = lm(Y~X1+X2+X4+X5, data=CDI)

coef.x1x2 = typeX1X2$coefficients
coef.x1x2x3 = typeX1X2X3$coefficients
coef.x1x2x4 = typeX1X2X4$coefficients
coef.x1x2x5 = typeX1X2X5$coefficients
coef.x1x2x3x4 = typeX1X2X3X4$coefficients
coef.x1x2x3x5 = typeX1X2X3X5$coefficients
coef.x1x2x4x5 = typeX1X2X4X5$coefficients

Y_typeX1X2 = coef.x1x2[1] + coef.x1x2[2]*X1 + coef.x1x2[3]*X2
Y_typeX1X2X3 = coef.x1x2x3[1] + coef.x1x2x3[2]*X1 + coef.x1x2x3[3]*X2 + coef.x1x2x3[4]*X3
Y_typeX1X2X4 = coef.x1x2x4[1] + coef.x1x2x4[2]*X1 + coef.x1x2x4[3]*X2 + coef.x1x2x4[4]*X4
Y_typeX1X2X5 = coef.x1x2x5[1] + coef.x1x2x5[2]*X1 + coef.x1x2x5[3]*X2 + coef.x1x2x5[4]*X5
Y_typeX1X2X3X4 = coef.x1x2x3x4[1] + coef.x1x2x3x4[2]*X1 + coef.x1x2x3x4[3]*X2 + coef.x1x2x3x4[4]*X3 + c
Y_typeX1X2X3X5 = coef.x1x2x3x5[1] + coef.x1x2x3x5[2]*X1 + coef.x1x2x3x5[3]*X2 + coef.x1x2x3x5[4]*X3 + c
Y_typeX1X2X4X5 = coef.x1x2x4x5[1] + coef.x1x2x4x5[2]*X1 + coef.x1x2x4x5[3]*X2 + coef.x1x2x4x5[4]*X4 + c

SSE.X1X2 = sum((Y - Y_typeX1X2)^2)
SSR.X1X2 = sum((Y_typeX1X2 - mean(Y))^2)

SSE.X1X2X3 =  sum((Y - Y_typeX1X2X3)^2)
SSR.X1X2X3 = sum((Y_typeX1X2X3 - mean(Y))^2)

SSE.X1X2X4 =  sum((Y - Y_typeX1X2X4)^2)
SSR.X1X2X4 = sum((Y_typeX1X2X4 - mean(Y))^2)

SSE.X1X2X5 =  sum((Y - Y_typeX1X2X5)^2)
SSR.X1X2X5 = sum((Y_typeX1X2X5 - mean(Y))^2)

SSE.X1X2X3X4 =  sum((Y - Y_typeX1X2X3X4)^2)
SSR.X1X2X3X4 = sum((Y_typeX1X2X3X4 - mean(Y))^2)

SSE.X1X2X3X5 =  sum((Y - Y_typeX1X2X3X5)^2)
SSR.X1X2X3X5 = sum((Y_typeX1X2X3X5 - mean(Y))^2)
```

```
SSE.X1X2X4X5 =  sum((Y - Y_typeX1X2X4X5)^2)
SSR.X1X2X4X5 = sum((Y_typeX1X2X4X5 - mean(Y))^2)

partialX3 = (SSE.X1X2 - SSE.X1X2X3)/SSE.X1X2
partialX3

partialX4 = (SSE.X1X2 - SSE.X1X2X4)/SSE.X1X2
partialX4

partialX5 = (SSE.X1X2 - SSE.X1X2X5)/SSE.X1X2
partialX5

extraSSX3 = (SSE.X1X2 - SSE.X1X2X3)
extraSSX4 = (SSE.X1X2 - SSE.X1X2X4)
extraSSX5 = (SSE.X1X2 - SSE.X1X2X5)

extraSSX3
extraSSX4
extraSSX5
```

c)

```
F_value = ((SSE.X1X2 - SSE.X1X2X5)/((SSE.X1X2X5)/(n-4)))
F_value
qf(0.99,1,n-4)
F_valuex3 = ((SSE.X1X2 - SSE.X1X2X3)/((SSE.X1X2X3)/(n-4)))
F_valuex4 = ((SSE.X1X2 - SSE.X1X2X4)/((SSE.X1X2X4)/(n-4)))
F_valuex3
F_valuex4
```

d)

```
partial.x1x2x3x4 = (SSE.X1X2 - SSE.X1X2X3X4)/SSE.X1X2
partial.x1x2x3x4

partial.x1x2x3x5 = (SSE.X1X2 - SSE.X1X2X3X5)/SSE.X1X2
partial.x1x2x3x5

partial.x1x2x4x5 = (SSE.X1X2 - SSE.X1X2X4X5)/SSE.X1X2
partial.x1x2x4x5

Fv = anova(typeX1X2,typeX1X2X4X5)$'F'[2]
Fv
qf(1-0.05, 2, n-5)
```

Screenshots of Output for Part 1

c)

```
beta.model2
beta.model1
```

```
          (Intercept)     Model2_PopDensity          Model2_Pop64 Model2_PersonalIncome
        -1.705742e+02          9.615889e+19          6.339841e+21          1.265665e+20
          (Intercept)       Model1_TotalPop       Model1_LandArea Model1_PersonalIncome
        -1.331615e+01          8.366178e+17         -6.552296e+19          9.413199e+19
```

Figure 1: 6.28c output

d)

```
RSquared_model1
RSquared_model2
```

```
[1] 0.9026432
[1] 0.9117491
```

Figure 2: 6.28d output

f)

```
NewRSquared_model1
NewRSquared_model2
```

```
[1] 0.9063789
[1] 0.9230238
```

Figure 3: 6.28f output

Screenshots of Output for Part 2

a)

```
partialX3
partialX4
partialX5

extraSSX3
extraSSX4
extraSSX5
```

```
[1] 0.02882495
[1] 0.003842367
[1] 0.5538182
[1] 4063370
[1] 541647.3
[1] 78070132
```

Figure 4: 7.37a output

c)

```
```{r, echo=FALSE, results='hide'}
F_value = ((SSE.X1X2 - SSE.X1X2X5)/((SSE.X1X2X5)/(n-4)))
F_value
qf(0.99,1,n-4)
F_valuex3 = ((SSE.X1X2 - SSE.X1X2X3)/((SSE.X1X2X3)/(n-4)))
F_valuex4 = ((SSE.X1X2 - SSE.X1X2X4)/((SSE.X1X2X4)/(n-4)))
F_valuex3
F_valuex4
```
```

```
[1] 541.1801
[1] 6.693358
[1] 12.94069
[1] 1.681734
```

Figure 5: 7.37c output

d)

```
partial.x1x2x3x4
partial.x1x2x3x5
partial.x1x2x4x5
Fv
qf(1-0.05, 2, n-5)
```

```
[1] 0.03314181
[1] 0.5558232
[1] 0.5642756
[1] 281.6688
[1] 3.016458
```

Figure 6: 7.37d output