Project 2

Tina Lagerblad, Keon Sadeghi

#Introduction

. . . . . .

In this report, we will perform regression analysis using data from the CDI data set which includes demographic information of the 440 most populous counties in the United States. The information in the data set is mostly from 1990-1992 and there are 17 variables.

In the previous project, we performed tasks such as stating the estimated regression functions relating to the data and calculated the mean squared error for the three predictor variables. We also found the predictor variable that accounted for the biggest reduction in the variability in the number of active physicians. We created a confidence interval and ANOVA table using the relationship between regress per capita income and the percentage of people that have at least a bachelor's degree for every geographic region. Lastly, we performed regression diagnostics to help us determine whether the linear regression model is better suited for one case than in others.

For this project, we are once again using the CDI data set to find different things. In Part I, we are constructing stem-and-leaf plots for all of the predictor variables in order to make observations on our data and variables and note any outliers. We are also obtaining the scatter plot matrix and correlation matrix for each model and calculating R_2 to observe the nature and strength of any bivariate relationships between variables and then to determine if either model provides a better fit. Additionally, we will obtain the residuals for each model and plot them against things such as Y-hat, the predictor variables and the two-factor interaction terms. We will also create a normal probability plot and determine whether or not one model is more appropriate than the other.

In Part II, we will calculate the coefficients of partial determination and use our findings to determine which of the four additional predictor variables is best and whether the extra sum of squares associated with this variable is larger than that of the other three variables. We will perform analysis using the F* test statistic and determine whether the other 3 predictor variables' F* test statistics will be as large. Also, we will use the F test to discover if adding the best pair to the model is helpful if X1,X2 are already included. Finally, in part III we will discuss our results and potential ways to improve the linear regression models. The main software that will be used as a a tool to conduct our analysis and research in this report is RStudio.

. . . . . .

#Part I: Multiple Linear Regression I

##Project 6.28

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

For total population:

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

```
##   0 | 11111111111111111111111111111111111111111111111111111111111111111111111111+254
##   0 | 55555555555555555555555566666666666666677777777777777777778888888888
##   1 | 000000122233333444
##   1 | 55699
##   2 | 1134
##   2 | 58
##   3 |
##   3 |
##   4 |
##   4 |
##   5 | 1
##   5 |
##   6 |
##   6 |
##   7 |
##   7 |
##   8 |
##   8 | 9
```

For land area:

```
##
##   The decimal point is 3 digit(s) to the right of the |
##
##    0 | 000011111111111122222222222222222223333333333333333333333333333444444+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
```

For population density:

```
##
##   The decimal point is 2 digit(s) to the left of the |
##
##   0 | 00000000000000000000000000000000000000011111111111111111111111111111+221
##   0 | 55555555555555555555555555555555555555555566666666666666666666666777+17
```

```
##   1 | 00011111223344444
##   1 | 555567799
##   2 | 3334
##   2 | 5568
##   3 | 0013
##   3 | 8
##   4 |
##   4 | 6
##   5 | 2
##   5 |
##   6 |
##   6 |
##   7 |
##   7 | 5
```

For percent of population older than 64:

```
##
##   The decimal point is at the |
##
##    2 | 0
##    4 | 47890389
##    6 | 112345567799013456667899
##    8 | 00112222233334444555666777778888899990002222333333444444445555666677
##   10 | 000111111222222222233333344444455555556666666677777778888888888999999+36
##   12 | 000000001111122223333333333444455555556666666777777777888899900000000+36
##   14 | 000011111112233344444555677889000000111122223455667778
##   16 | 12556699901122345
##   18 | 06778
##   20 | 070
##   22 | 018828
##   24 | 47
##   26 | 055
##   28 | 1
##   30 | 7
##   32 | 138
```

For total personal income:

```
##
##   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 |
```

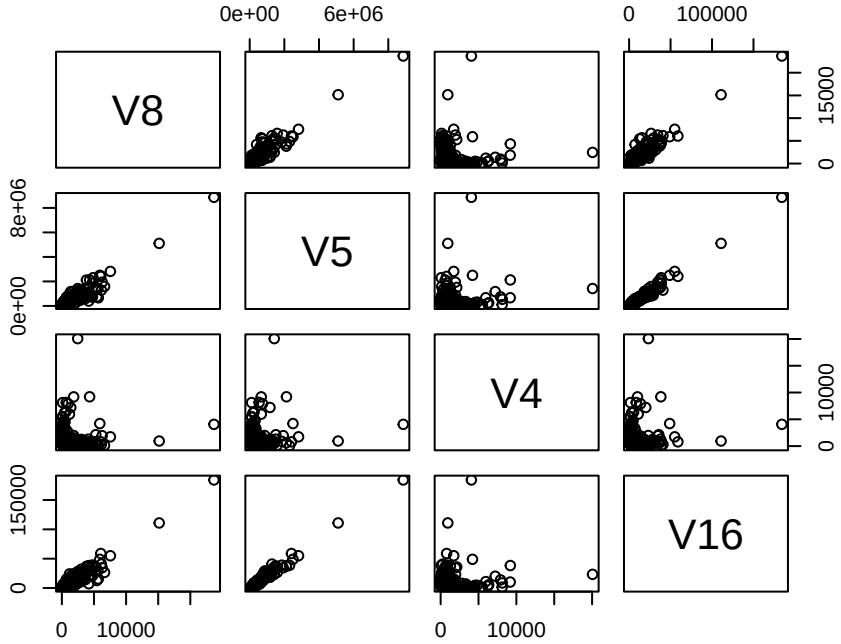
```
##   11 | 1
##   12 |
##   13 |
##   14 |
##   15 |
##   16 |
##   17 |
##   18 | 4
```

There are a few outliers for each plot, however nothing extreme. There are far more people with lower incomes than higher. The data for the percentage of old people seems pretty normally distributed with the mean at the lower end of the range, appearing to be around 10 or 12 percent, meaning lower percentage of people over 64 is more common. Lower population density is much more frequent than high. Smaller land and population are more common than larger land size and population. In general for the graphs, the data seems to be consistently at the top of the graphs, showing that smaller values for each variable are more frequent.

. . . . . .

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

Given that V8 is the variable for the number of active physicians, V5 represents total population, V4 represents land area here, V16 represents total personal income. Scatter plot matrix for model 1:

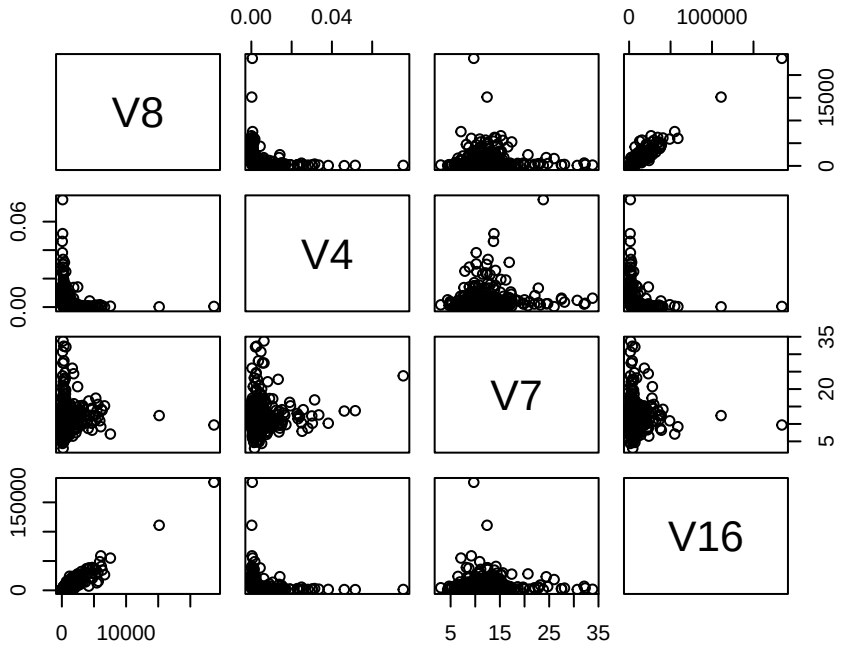


From this scatterplot matrix we can guess that there is the least correlation between land area (V4) and the other variables since the data for the plots involving V4 is the least linear. The plots that have the best linear relationship would be those between total population (V5) and total personal income (V16), meaning

that it is likely that there is a pretty strong correlation between these variables. Each other relationship not mentioned also seems to be pretty linear, indicating correlation between those variables.

. . . . . .

Given that V8 is the variable for the number of active physicians, V4 represents population density here, V7 represents the percent of population over 64, V16 represents total personal income. Scatter plot matrix for model 2:



As we can see there does not appear to be any notably linear relationship between the variables except for between active physicians (V8) and total personal income (V16). This tells us that there is likely a relationship between these variables but there does not appear to be any strong correllation with any other notable pairs.

. . . . . .

Given that V8 is the variable for the number of active physicians, V5 represents total population, V4 represents land area here, V16 represents total personal income. Correlation matrix for model 1:

```
##             V8        V5         V4       V16
## V8  1.00000000 0.9402486 0.07807466 0.9481106
## V5  0.94024859 1.0000000 0.17308335 0.9867476
## V4  0.07807466 0.1730834 1.00000000 0.1270743
## V16 0.94811057 0.9867476 0.12707426 1.0000000
```

This correlation matrix supports our assumptions made about bivariate relationships made from the scatterplot matrix. As we see here, the relationships involving V4 show the weakest relationships between those variables, since those correlation coefficents are the furthest from zero. Additionally, we see that the correlation coefficients closest to 1, indicating the strongest linear correlation, are those between V5 and

V16, confirming that these are the variables with the strongest relationship. The rest of the variables not mentioned all seem to have pretty strong relationships as well, since they are also close to 1.

. . . . . .

Given that V8 is the variable for the number of active physicians, V4 represents population density here, V7 represents the percent of population over 64, V16 represents total personal income. Correlation matrix for model 2:

```
##              V8          V4          V7         V16
## V8   1.00000000 -0.22232620 -0.00312863  0.94811057
## V4  -0.22232620  1.00000000  0.07944933 -0.21541496
## V7  -0.00312863  0.07944933  1.00000000 -0.02273315
## V16  0.94811057 -0.21541496 -0.02273315  1.00000000
```

This matrix supports our predictions made from the scatterplot matrix for this model. as we can see there is pretty close to a perfectly positive linear correlation between the variables V8 and V19 since 0.94811057 is close to 1, which is a perfectly positive linear relationship. Additionally, we can also confirm that there appears to be little correlation between any of the other variables, as the rest of the correlation coefficients are much farther from 1.

Note: While the variable V4 is used for the matrices for both models, it has different meanings for each model.

. . . . . .

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

- The fitted first-order regression functions for the first proposed model (Y = Number of Active Physicians, X_{11} = Total Population, X_{12} = Land Area & X_{13} = Total Personal Income):

$$Y_1 = -13.3161522 + (8.3661782 \times 10^{-4})X_{11} + (-0.065523)X_{12} + (0.094132)X_{13} + \varepsilon_1$$

- The fitted first-order regression functions for the second proposed model (Y = Number of Active Physicians, X_{21} = population density, X_{22} = percent of population greater than 64 years old & X_{23} = total personal income):

$$Y_2 = -170.5742233 + (0.0961589)X_{21} + (6.3398406)X_{22} + (0.1265665)X_{23} + \varepsilon_2$$

. . . . . .

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

Model 1 R^2:

```
## [1] 0.9026432
```

Model 2 R^2:

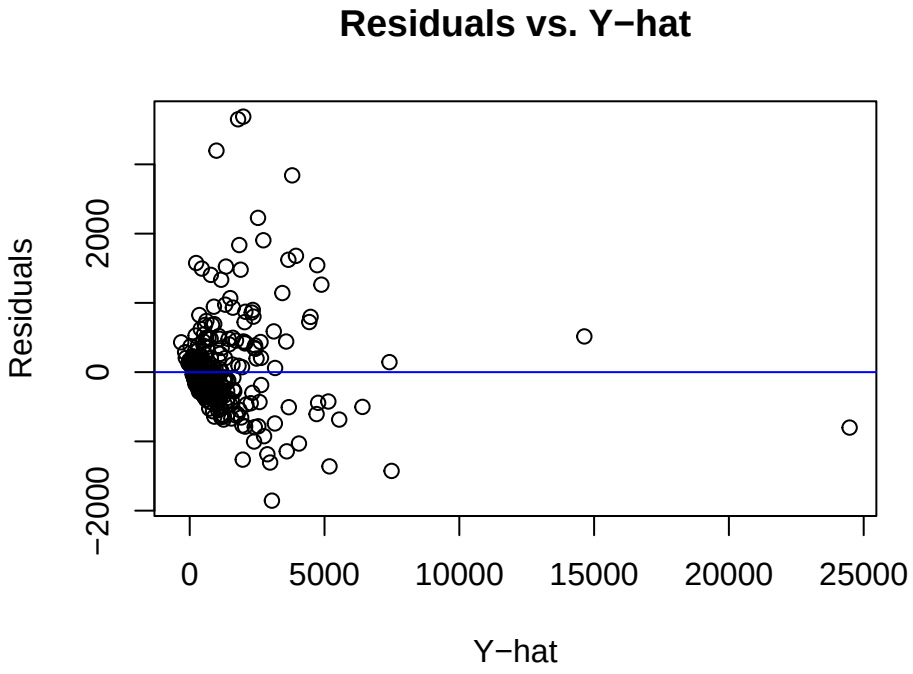
```
## [1] 0.9117491
```

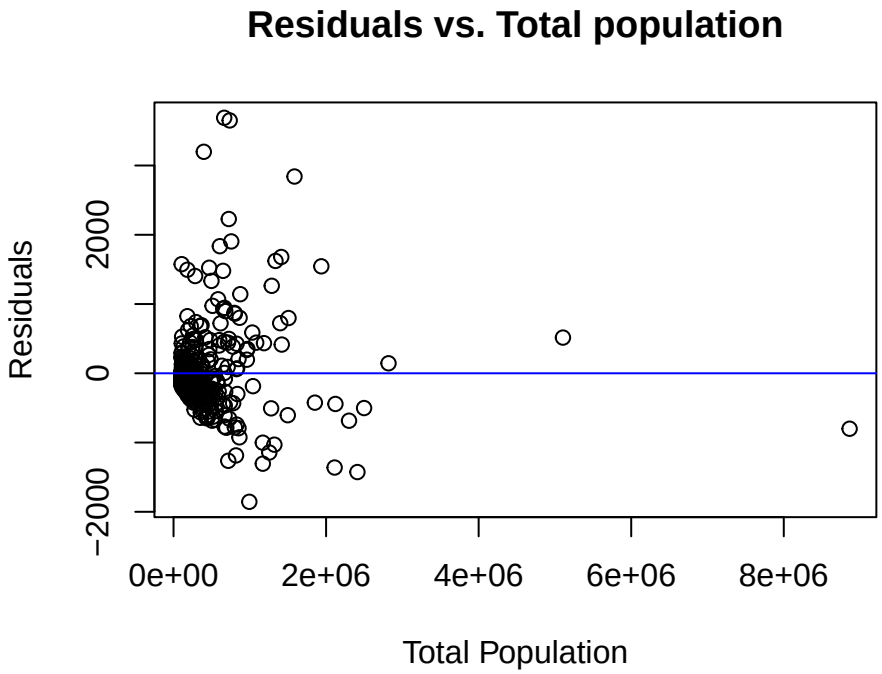
Since the R^2 values are so close for the models, we can not provide a confidently definite answer as to which would be preferable. We need to do further testing in order to be able to say if one is clearly superior.

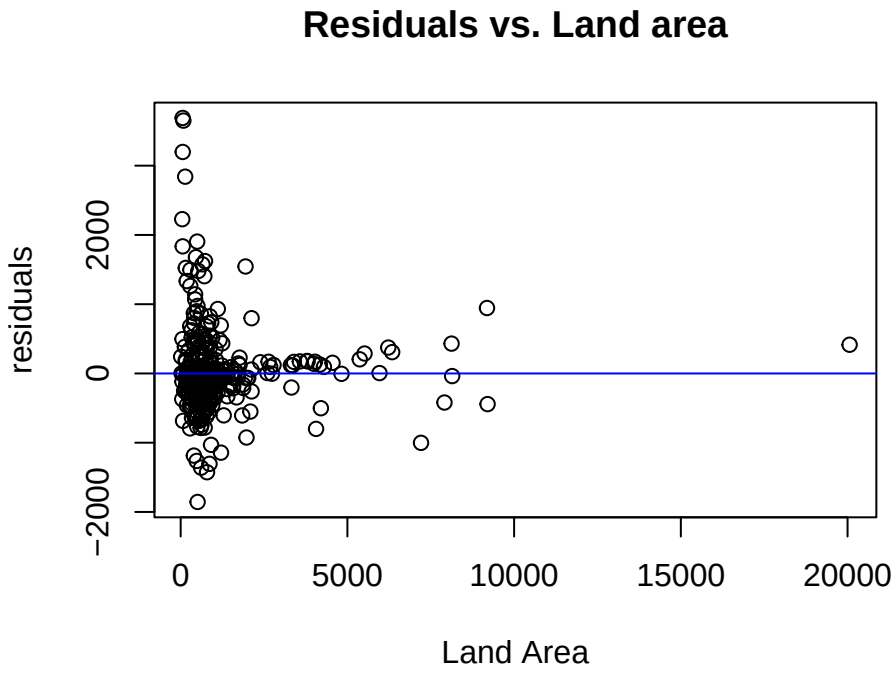
. . . . . .

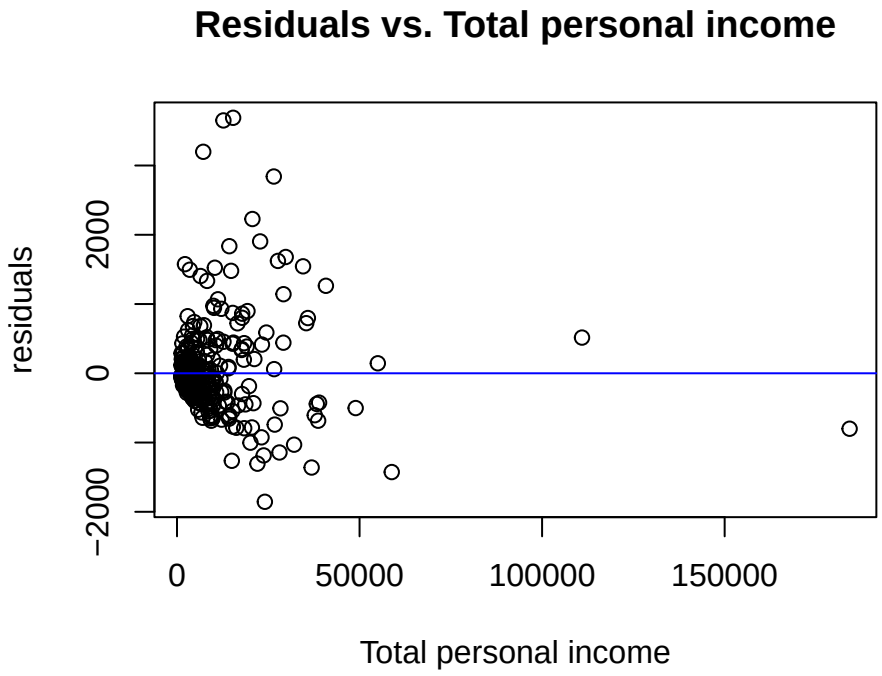
(e) For each model, obtain the residuals and plot them against Y, each of the three predictor variables, and each of the two-factor interaction terms. Also prepare a normal probability plot for each of the two fitted models. Interpret your plots and state your findings. Is one model clearly preferable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

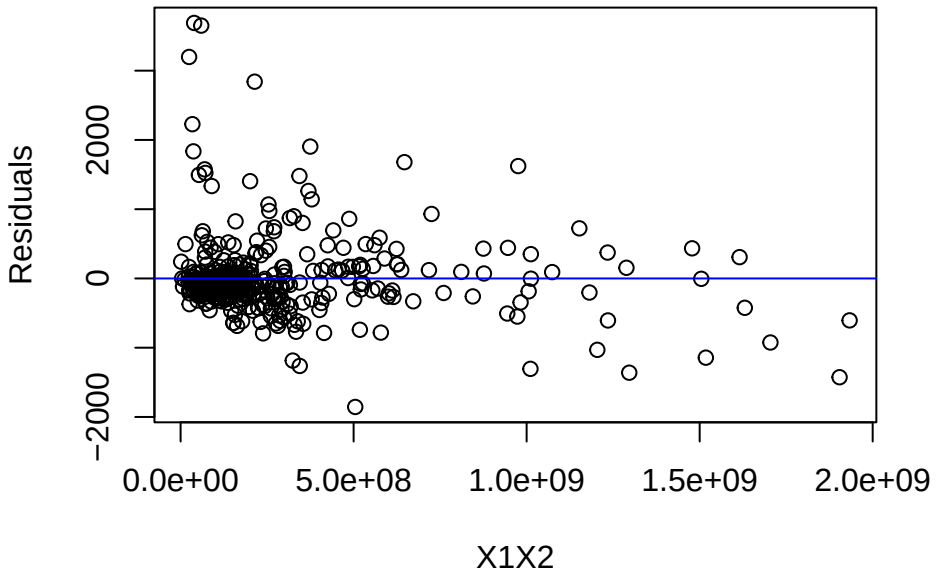
Plots for model 1:

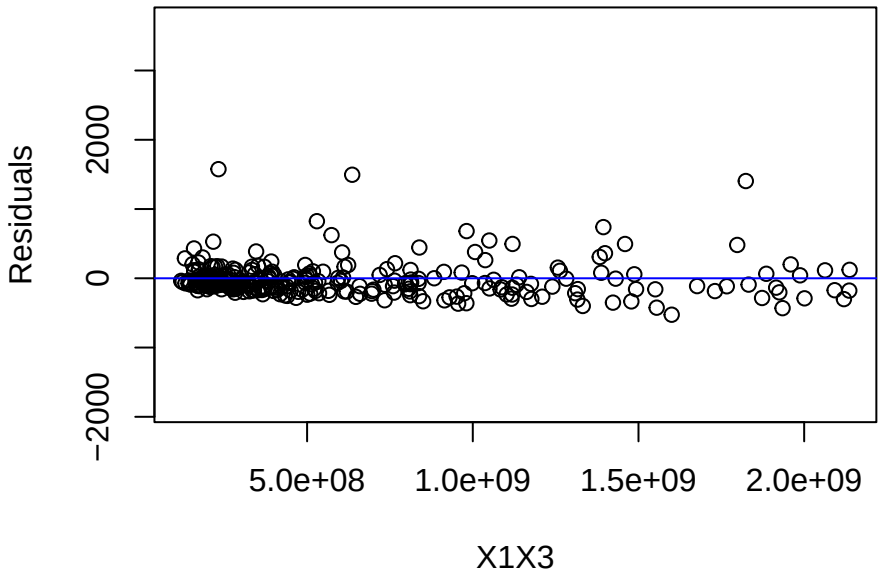


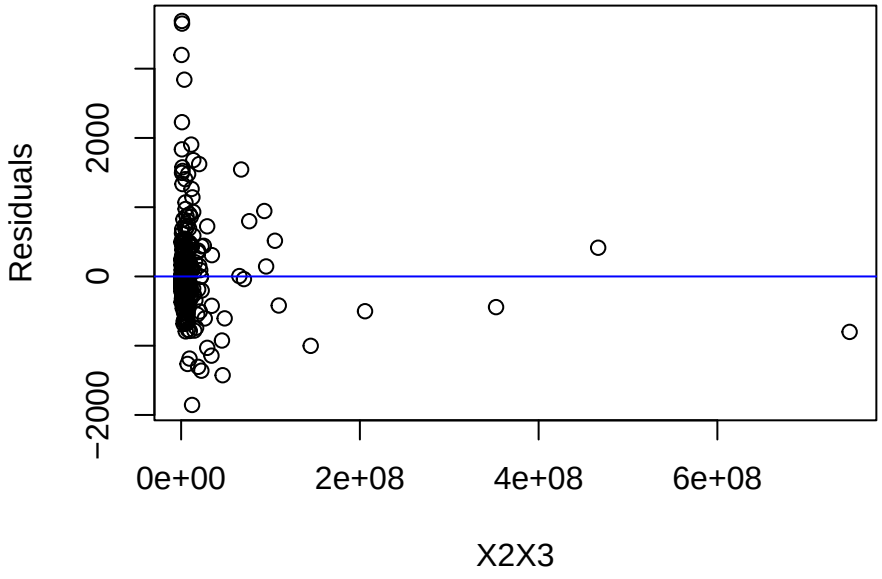

Residuals vs. Total population
Residuals
2000
0
−2000
0e+00
2e+06
4e+06
6e+06
8e+06
Total Population

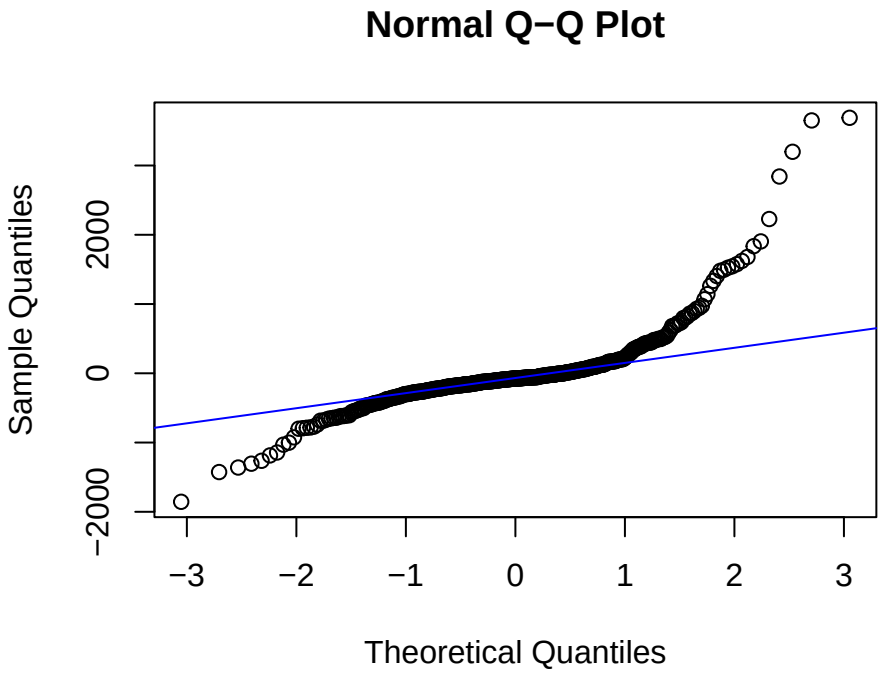

Residuals vs. Land area
residuals
2000
0
−2000
0
5000
10000
15000
20000
Land Area


Residuals vs. Total personal income
residuals
2000
0
−2000
0
50000
100000
150000
Total personal income

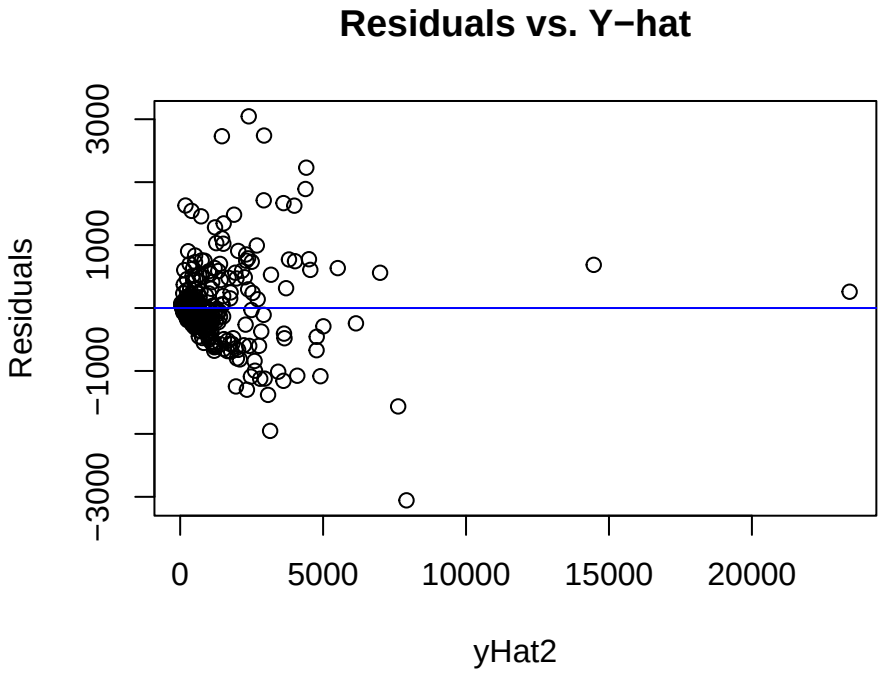

Residuals
2000
0
−2000
0.0e+00
5.0e+08
1.0e+09
1.5e+09
2.0e+09
X1X2

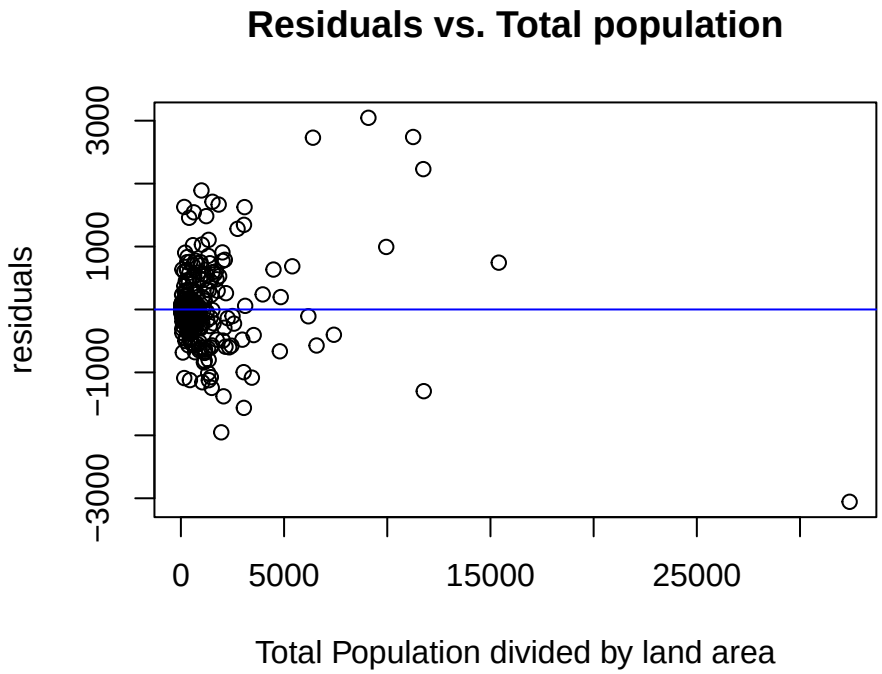

Residuals
2000
0
−2000
5.0e+08
1.0e+09
1.5e+09
2.0e+09
X1X3

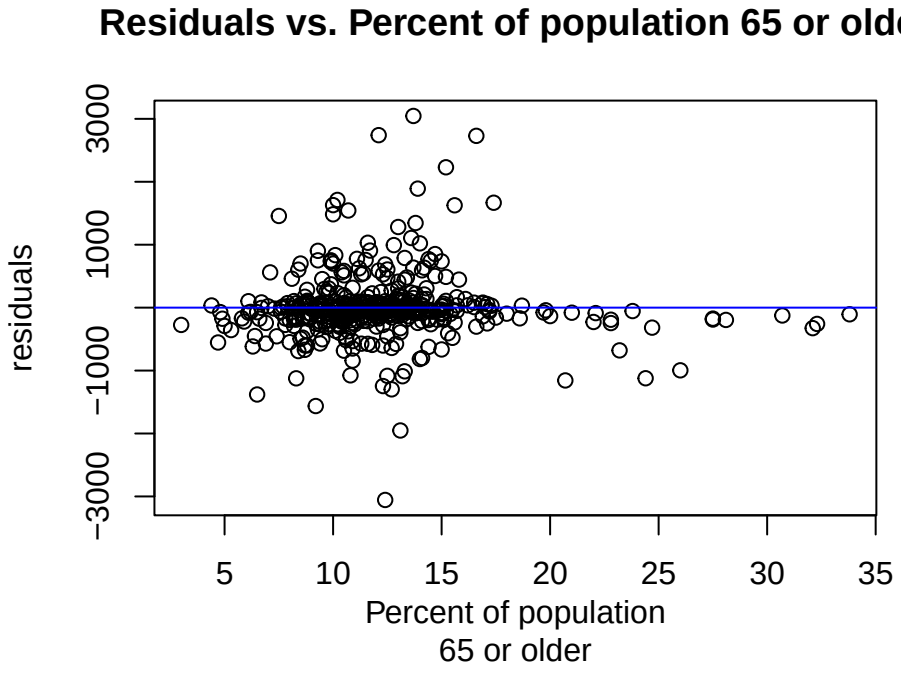

Residuals
2000
0
−2000
0e+00
2e+08
4e+08
6e+08
X2X3

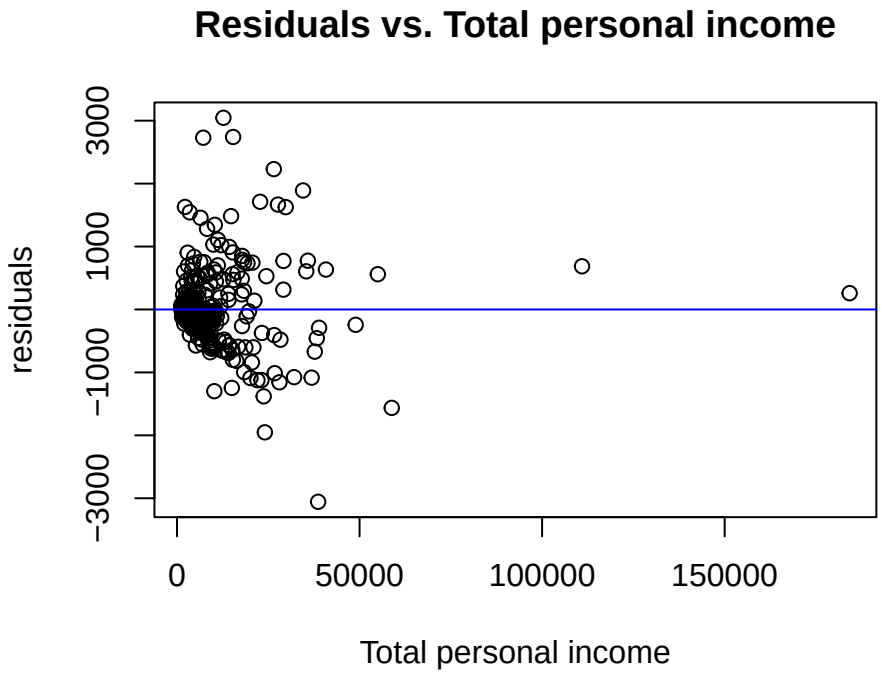


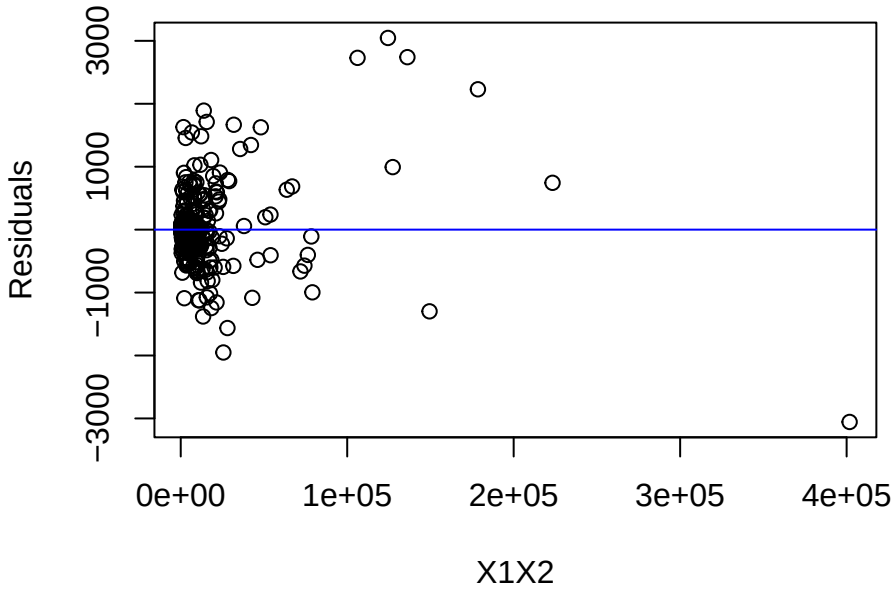
Plots for model 2:

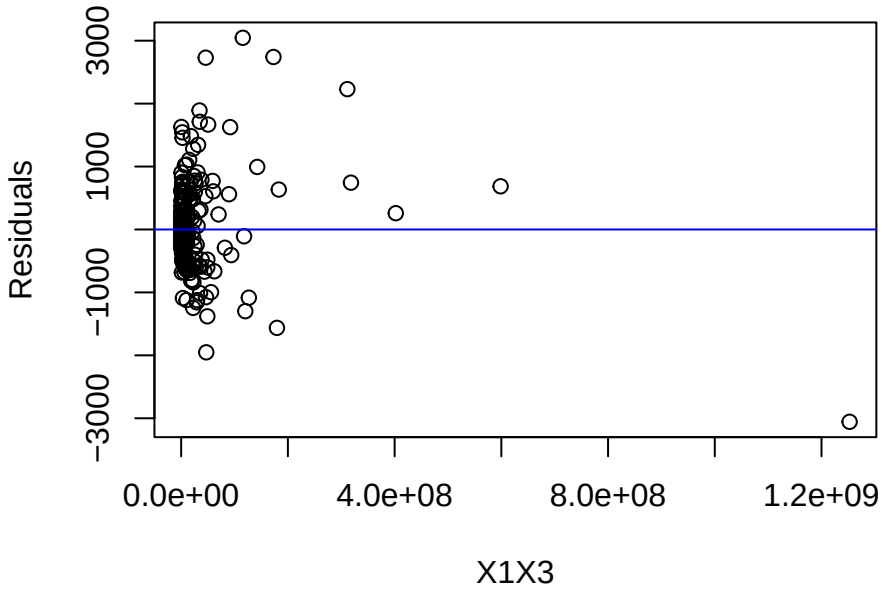


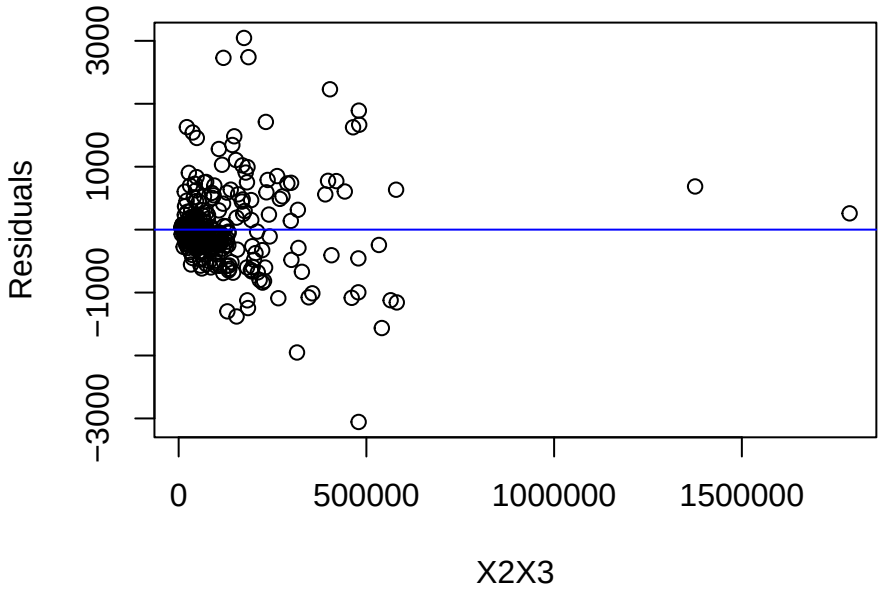

Residuals vs. Total population
residuals
3000
1000
−1000
−3000
0
5000
15000
25000
Total Population divided by land area

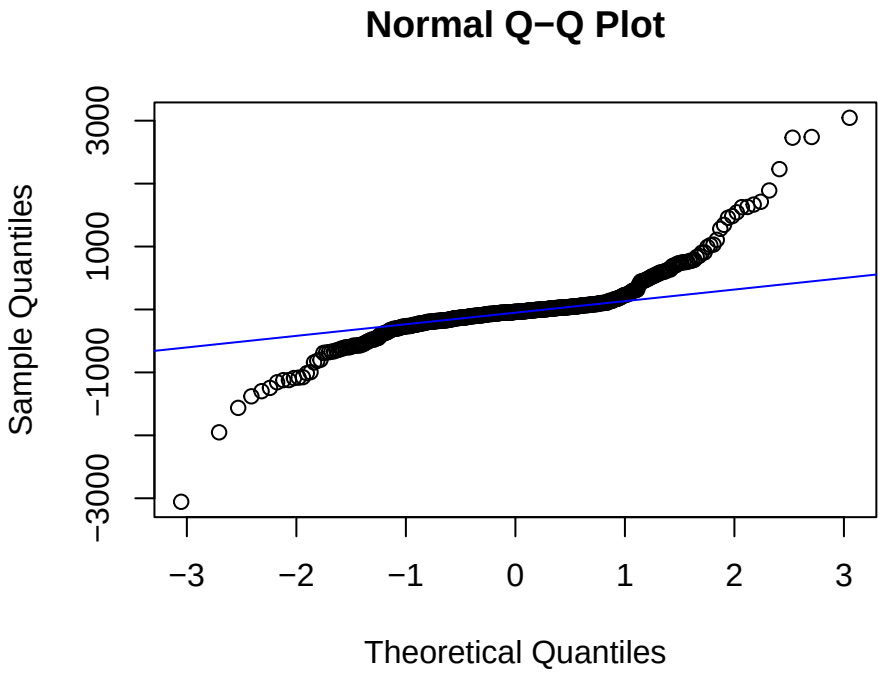

Residuals vs. Percent of population 65 or older
residuals
3000
1000
−1000
−3000
5
10
15
20
25
30
35
Percent of population
65 or older


Residuals vs. Total personal income
residuals
3000
1000
−1000
−3000
0
50000
100000
150000
Total personal income


Residuals
3000
1000
−1000
−3000
0e+00
1e+05
2e+05
3e+05
4e+05
X1X2


Residuals
3000
1000
−1000
−3000
0.0e+00
4.0e+08
8.0e+08
1.2e+09
X1X3


Residuals
3000
1000
−1000
−3000
0
500000
1000000
1500000
X2X3



The residual plots appear roughly normal and evenly distributed about 0. The residual graphs also contain some outliers. When looking at the normal probability plots for model 1 and 2 the plot for model 2 appears to show the most normal distribution making it more preferable in terms of appropriateness.

. . . . . .

(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-factor interactions are X_1X_2, X_1X_3, X_2X_3. Repeat part (d) for the two expanded models.

```
## [1] 0.357187
```

```
## [1] 0.9230238
```

. . . . . .

#Part II: Multiple Linear Regression II

##Project 7.37

(a) For each of the following variables, calculate the coefficient of partial determination given that XI and X2 are included in the model: land area (X3), percent of population 65 or older (X4), number of hospital beds (X5).

Partial R^2 for land area:

```
## [1] 0.02882495
```

Partial R^2 for percent of population 65 or older:

```
## [1] 0.003842367
```

Partial R^2 for number of hospital beds:

```
## [1] 0.5538182
```

.....

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

Since the extra sum of squares value for the number of hospital beds (X_5) is the largest compared to that of the other tested variables, we assume that the variable X_5 has the strongest relationship with variables X_1 and X_2. Therefore, we conclude that X_5 is the best additional predictor variable.

.....

(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 = 0.01. State the alternatives, decision rule, and conclusion. Would the F* test statistics for the other potential predictor variables be as large as the one here? Discuss.

```
## [1] 541.1801
```

```
## [1] 6.693358
```

H_0: β_5=0

H_a: β_5 does not equal 0

Reject the null hypothesis if the F* Value is greater than the critical value (f(0.99;1,436)).

Since our F* Value (541.1801) is greater than our F critical value (6.693358), we reject the null hypothesis and conclude the alternative hypothesis. The variable determined in part a will be helpful in the model when X_1 and X_2 are included. The F* value would not be as large for the other variables, since we observed that their coefficients of partian determination were smaller.

.....

(d) Compute three additional coefficients of partial determination: $R^2_{(Y,X3,X4|X1,X2)}$, $R^2_{(Y,X3,X5|X1,X2)}$, and $R^2_{(Y,X4,X5|X1,X2)}$. 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.

$R^2_{(Y,X3,X4|X1,X2)}$:

```
## [1] 0.03314181
```

$R^2_{(Y,X3,X5|X1,X2)}$:

```
## [1] 0.5558232
```

$R^2_{(Y,X4,X5|X1,X2)}$:

```
## [1] 0.5642756
```

From these coefficients of partial determination we conclude that X^4 and X^5 seem to be the most useful to include, as the coefficient of partial determination associated with this pair is the largest compared to the others, meaning they are able to cover the most variation.

F Test:

```
## [1] 4.654156
```

```
## [1] 281.6688
```

H_0: β_4=β_5=0

H_a: Not all β_4 and β_5=0

Reject the null hypothesis if F*>F(0.99;2,436).

Since our F value: 281.6688, is greater than the critical value: 4.654156, we reject the null hypothesis and conclude that the pair is statistically significant and should be included.

. . . . . .

#Part III: Discussion

. . . . . .

After analyzing several parts of the CDI dataset we were able to come to many conclusions about how to best fit linear regression models to this data. From part I we were able to learn about the nature of our chosen variables and the nature of the relationships between them. We used R^2 and residual, interaction and normality plots for each proposed model in order to try to conclude which of the two would be more appropriate. We found both to be very similar in fit, but decided that the data's normality distribution of model two may be marginally better. In part II we found the variable for the number of hospital beds (X_5) to be the most effective additional variable to include as compared to the others and that the pair predictor X_4 and X_5 was the most significant amongst the tested pairs.

There were several parts of course material that was relevant to our analysis in this project. In particular, information on how to calculate things such as R^2 and the coefficient of partial determination as well as knowledge of creating plots such as the stem-and-leaf plot and scatter plot matrix. Also, we used knowledge of how to obtain residuals and plot them and knowledge of F tests. Finally, we used our knowledge of F* test statistics to help us determine alternatives, decision rules and the conclusion. All these different topics that we learned throughout the quarter helped us throughout the project.

To improve linear regression models, we could consider more predictor variables and test to see if including them in our model would provide a better fit. We could also try to gather more data through surveys or any other type of data collecting method for the variables we found to be most efficient in order to widen the scope of our data which may improve our model. Another way we could improve them is by finding out whether or not the outliers were a result of human error, and remove any erroneous values them from our dataset in order to get a more accurate fit based on the data. Using our knowledge of the course, we were able to perform analysis and learn more about the CDI dataset.

. . . . . .

#Appendix

```
knitr::opts_chunk$set(
    error = FALSE,
    message = FALSE,
    warning = FALSE,
    echo = FALSE, # hide all R codes!!
    fig.width=5, fig.height=4,#set figure size
    fig.align='center',#center plot
    options(knitr.kable.NA = ''), #do not print NA in knitr table
    tidy = FALSE #add line breaks in R codes
)
data = read.table("CDI.txt", header = FALSE)
income = data$V16
ap = data$V8
pd = data$V4 / data$V5
tp = data$V5
po = data$V7
la = data$V4

pd2 = data$V5/data$V4
po2 = data$V7
income2 = data$V16
ap2 = data$V8

secondMod = lm(ap2 ~ pd2 + po2 + income2, data = data)

firstMod = lm(ap ~ tp + la + income)
#secondMod = lm(ap ~ pd + po + income)
mod1.co= firstMod$coefficients
mod2.co= secondMod$coefficients
stem(tp)
stem(la)
stem(pd)
stem(po)
stem(income)
modDat1 = data.frame(c(data[8:8], data[5:5], data[4:4], data[16:16]))
modDat2 = data.frame(c(data[8:8], data[4:4]/data[5:5], data[7:7], data[16:16]))
pairs(modDat1)
pairs(modDat2)
cor(modDat1)
cor(modDat2)
mod2.coef0 = mod2.co[1]
mod2.coef1 = mod2.co[2]
mod2.coef2 = mod2.co[3]
mod2.coef3 = mod2.co[4]
mod1.coef0 = mod1.co[1]
mod1.coef1 = mod1.co[2]
mod1.coef2 = mod1.co[3]
mod1.coef3 = mod1.co[4]
R_1 = summary(firstMod)$r.squared
R_1
R_2 = summary(secondMod)$r.squared
R_2
data1X1 = data$V5
```

```
data1X2 = data$V4
data1X3 = data$V16
data1Y = data$V8
data1both = data1X1*data1X2
data1both2 = data1X1*data1X3
data1both3 = data1X2*data1X3
data2X1 = data$V5/data$V4
data2X2 = data$V7
data2X3 = data$V16
data2Y = data$V8
data2X1 = data$V5/data$V4
data2X2 = data$V7
data2X3 = data$V16
data2Y = data$V8
data2both = data2X1*data2X2
data2both2 = data2X1*data2X3
data2both3 = data2X2*data2X3

fit1 = lm(data1Y ~ data1X1 + data1X2 + data1X3, data = data)
res1 = fit1$residuals
yHat1 = fit1$fitted.values
plot(x=yHat1, y=res1, main = "Residuals vs. Y-hat", ylab = 'Residuals', xlab = 'Y-hat')
abline(h = 0, col = 'blue')

plot(x=data1X1, y=res1, main = "Residuals vs. Total population", xlab = 'Total Population', ylab = 'Resi
abline(h = 0, col = 'blue')

plot(x=data1X2, y=res1, main = "Residuals vs. Land area", xlab = 'Land Area', ylab = 'residuals')
abline(h = 0, col = 'blue')

plot(x=data1X3, y=res1, main = "Residuals vs. Total personal income", xlab = 'Total personal income', yl
abline(h = 0, col = 'blue')

plot(x=data1X1*data1X2, y=res1, ylab="Residuals", xlab="X1X2")
abline(h = 0, col = 'blue')

plot(x=data1X1*data1X3, y=res1, ylab="Residuals", xlab="X1X3")
abline(h = 0, col = 'blue')

plot(x=data1X2*data1X3, y=res1, ylab="Residuals", xlab="X2X3")
abline(h = 0, col = 'blue')

qqnorm(res1)
qqline(res1, col = 'blue')
fit2 = lm(data2Y ~ data2X1 + data2X2 + data2X3, data = data)
res2 = fit2$residuals
yHat2 = fit2$fitted.values
plot(x=yHat2, y=res2, main = "Residuals vs. Y-hat", ylab = 'Residuals')
abline(h = 0, col = 'blue')

plot(x=data2X1, y=res2, main = "Residuals vs. Total population", xlab = 'Total Population divided by lar
abline(h = 0, col = 'blue')
```

```
plot(x=data2X2, y=res2, main = "Residuals vs. Percent of population 65 or older", xlab = 'Percent of pop
65 or older', ylab = 'residuals')
abline(h = 0, col = 'blue')

plot(x=data2X3, y=res2, main = "Residuals vs. Total personal income", xlab = 'Total personal income', yl
abline(h = 0, col = 'blue')

plot(x=data2X1*data2X2, y=res2, ylab="Residuals", xlab="X1X2")
abline(h = 0, col = 'blue')

plot(x=data2X1*data2X3, y=res2, ylab="Residuals", xlab="X1X3")
abline(h = 0, col = 'blue')

plot(x=data2X2*data2X3, y=res2, ylab="Residuals", xlab="X2X3")
abline(h = 0, col = 'blue')

qqnorm(res2)
qqline(res2, col = 'blue')
model1Ex = lm(data1Y ~ data1X1 + data1X2 + data1X3 + data1both + data1both2 + data1both3, data = data)
RS1Ex = summary(model1Ex)$r.squared
RS1Ex

model2Ex = lm(data1Y ~ data2X1 + data2X2 + data2X3 + data2both + data2both2 + data2both3, data = data)
R2Ex = summary(model2Ex)$r.squared
R2Ex
hb=data$V9
fit21=lm(ap~tp+income, data=data)
sumsq1=(summary(fit21)$sigma^2)*fit21$df.residual

fit22=lm(ap~tp+income+la, data=data)
sumsq2=(summary(fit22)$sigma^2)*fit22$df.residual
rsqx1x2x3=1-sumsq2/sumsq1
rsqx1x2x3
fit23=lm(ap~tp+income+po, data=data)
fit24=(summary(fit23)$sigma^2)*fit23$df.residual
rsqx1x2x4=1-fit24/sumsq1
rsqx1x2x4
fit24=lm(ap~tp+income+hb, data=data)
sumsq3=(summary(fit24)$sigma^2)*fit24$df.residual
rsq1x2x5x=1-sumsq3/sumsq1
rsq1x2x5x
df125=fit24$df.residual
#df125

Part_df=fit21$df.residual
#Part_df
Fstat=((sumsq1-sumsq3)/(Part_df-df125))/(sumsq3/df125)
Fstat
qf(0.99,1,df125)
fit221=lm(ap~tp+income+la+po, data=data)
sumsq21=(summary(fit221)$sigma^2)*fit221$df.residual
rsq1=1-sumsq21/sumsq1
rsq1
```

```
fit222=lm(ap~tp+income+la+hb, data=data)
sumsq22=(summary(fit222)$sigma^2)*fit222$df.residual
rsq2=1-sumsq22/sumsq1
rsq2
fit223=lm(ap~tp+income+po+hb, data=data)
sumsq23=(summary(fit223)$sigma^2)*fit223$df.residual
rsq3=1-sumsq23/sumsq1
rsq3
Tot_df=fit223$df.residual
#Tot_df
f_stat = qf(0.99,2,436)
f_stat
f_star=((sumsq1-sumsq23)/(Part_df-Tot_df))/(sumsq23/Tot_df)
f_star
```