Regression Analysis on Average Income Per Person Associate with Other Variables in CDI

Sifeng Li
sifengl@andrew.cmu.edu
18 October 2021

Abstract

We address the question of how to identify the factors that are associated with the average income per person in the United States. We examine data on selected county demographic information (CDI) for 440 of the most populous countries in the United States collected by Kutner et al. (2005), using exploratory data analyses to make preliminary findings. From exploratory data analysis, it appears that both variable population and total income have association with doctors, hosp.beds, and crimes. A simple linear regression analysis shows that our best linear regression model should use “per-capita crime” as the measurement of our predictor variable and the model should be performed without interaction term. In multiple regression analysis, we perform the analysis through both subsets regression and stepwise regression. By comparing the adjusted R-Squared Values, AIC, BIC, and case-wise diagnostic plots, we conclude that the most suitable model would be the one selected from stepwise regression and average income per person has association with land area, percent of population aged 18-34, number of active physicians, percent below poverty level, percent high school graduates, state of California, state of New Jersey, state of Nevada, state of Utah.

1 Introduction

The average income per person can be considered as an important factor of identifying the economic and social aspects of a country. With the common understanding that the average income per person is difficult to be measured based on only a single variable, how should the average income per person be related to various kinds of variables on the country’s economic, health, and social aspects.

This question is especially critical in the research area where social scientists would like to gain first-hand information on the relationship between the average income per person and other potential factors, and thus helping them understand the current situation of a well being’s income status in the United States, but at the same time determine further directions on understanding how the average income per person in the United States can reflect social and economic problems.

In addition to answering the main question posed above, we will address the following questions:

- Among all the variables that we’re considering from the dataset, which variables seem to be related to which other variables closely in the data? Is there any practical meaning with regards to findings on these variables?

- If we ignore all other variables, is per-capita income related to crime rate? If so, does the relationship vary from region to region in the United States? Which expression of the variable performs better in the model, using number of crimes or number of crimes divided by population as the variable?
- How to find the best model predicting per-capita income from the other variables by having both statistical and practical meaning?

2 Data

The data for this study come from Kutner et al. (2005) with providing selected county demographic information (CDI) for 440 of the most populous counties in the United States. The information generally pertains to the years 1990 and 1992. Counties with missing data were deleted from the data set. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county.

In all, 440 observations are represented in the data available to us, and the following variables were measured on each:

id = Identification number = 1-40
county = County = County Name
state = State = Two-letter state abbreviation
land.area = Land area = Land area (square miles)
pop = Total population = Estimated 1990 population
pop.18_34 = Percent of 1990 CDI population aged 18–34
pop.65_plus = Percent of 1990 CDI population aged 65 or older
doctors = Number of professionally active nonfederal physicians during 1990
hosp.beds =Total number of beds, cribs, and bassinets during 1990
crimes = Total number of serious crimes in 1990
pct.hs.grad = Percent of adult who are 25 years old or older who are high school graduates
pct.bach.deg =Percent of adult who are 25 years old or older who have bachelor's degree
pct.below.pov = Percent of 1990 CDI population with income below poverty level
pct.unemp = Percent of 1990 CDI population that is unemployed
per.cap.income = Average income per person of 1990 CDI population (in dollars)
tot.income = Total personal income of 1990 CDI population (in millions of dollars)
region = Geographic region classification used by the US Bureau of the Census, NE (northeast region of the US), NC (north-central region of the US), S (southern region of the US), and W (Western region of the US)

In Table 1 and Table 2, we show the summary statistics for continuous variables and categorical variables, respectively.

Variables	Obs	Minimum	Median	Mean	Maximum	S.D.
land.area	440	15.0	656.5	1041.4	20062.0	1549.9

pop	440	100043	217280	393011	8863164	601987
pop.18_34	440	16.4	28.1	28.57	49.7	4.19
pop.65_plus	440	3	11.75	12.17	33.8	3.99
doctors	440	39	401	988	23677	1789.75
hosp.beds	440	92	755	1458.6	27700	2289.13
crimes	440	563	11820	27112	688936	58237.51
pct.hs.grad	440	46.6	77.7	77.56	92.9	7.02
pct.bach.deg	440	8.1	19.7	21.08	52.3	7.65
pct.below.pov	440	1.4	7.9	8.7	36.3	4.66
pct.unemp	440	2.2	6.2	7.5	21.3	2.34
per.cap.income	440	8899	17759	18561	37541	4059.19
tot.income	440	1141	3857	7869	184230	12884.32

Table 1: Summary Statistics for Continuous Variables of CDI Dataset

Region	Frequency
NC	108
NE	103
S	152
W	77

Table 2: Summary Statistics for Categorical Variables Region of CDI Dataset

	Maximum Frequency	Median Frequency	Minimum Frequency
County	Jefferson 7	1	1
State	CA 34	7	1

Table 3: Summary Statistics for Categorical Variables County and State of CDI Dataset

In Figure 1 we show all histograms of all the continuous variables, except for variables pop.18_34 and pop.65_plus, we need to do data transformation on each variable.

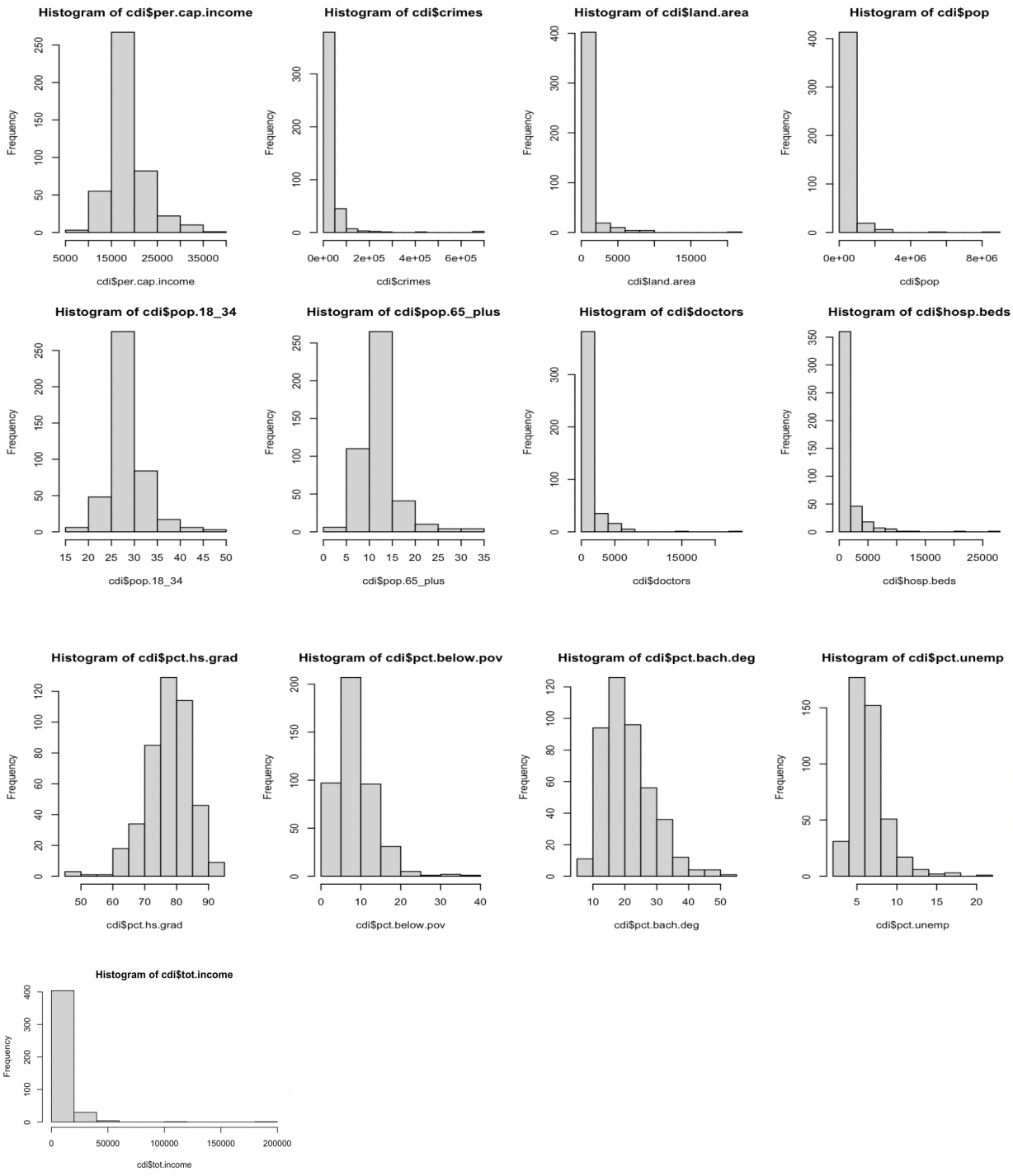


Figure 1: Histogram of all Continuous Variables for CDI Dataset

In Figure 2 we show the correlation plot of all the continuous variables. We can notice that the darker colors and bigger size the circle is, the more connected the relationship, i.e. the bigger

correlation, that two variables have. Variable doctors, hosp.beds, and crimes are highly correlated to both variable pop and variable tot.income.

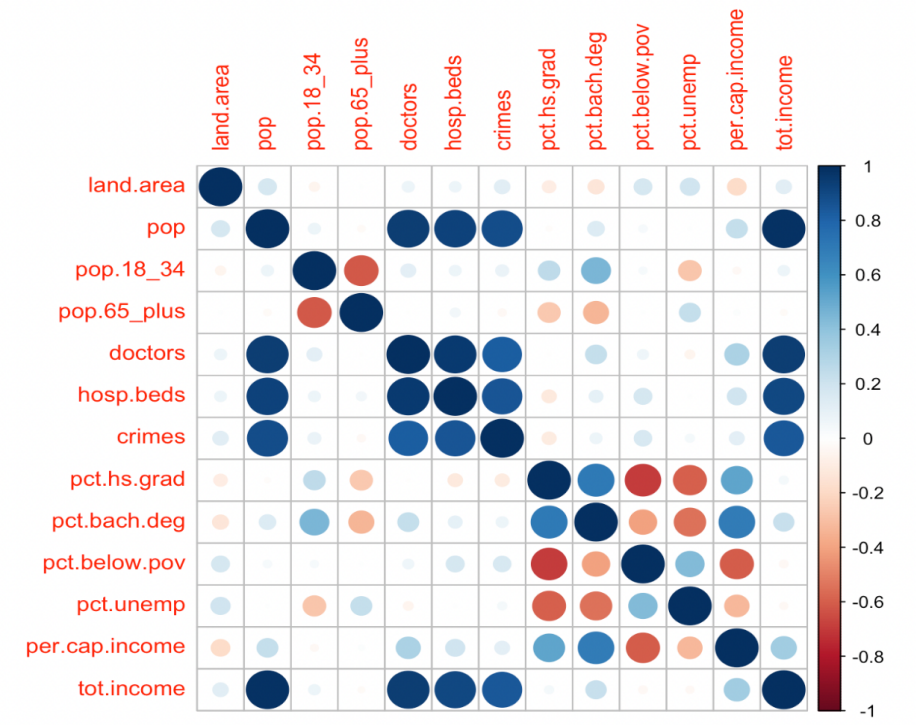


Figure 2: Correlation Plot of all Continuous Variables for CDI Dataset

In Figure 3 we show the boxplot for categorical variable region by plotting each region's boxplot. For the region is "S," there are more outliers compared to other regions.

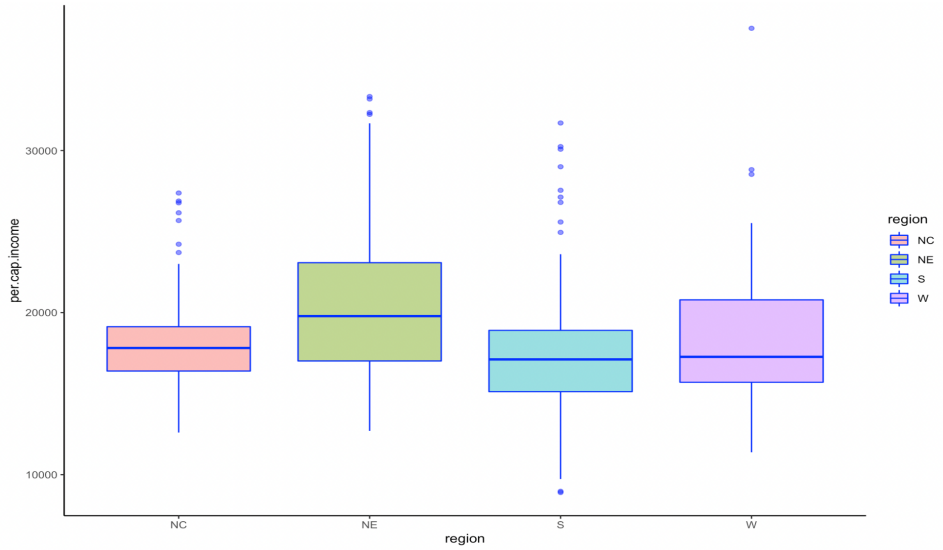


Figure 3: Boxplot of all Continuous Variable Region for CDI Dataset

More details from an Exploratory Data Analysis (EDA) can be found in Appendix 1.

3 Methods

Our analysis consists of three parts. For the first part, in order to identify the different functions that continuous variables and categorical variables could have on affecting the response variable average income per person, we divide all variables into continuous and categorical variables and examined raw data in cdi.dat. Then, we relied on visual observation of the exploratory correlation plot to further investigate the closely related relationship between groups of variables. This analysis can tell us how variables work in combination to affect the average income per person. Detailed R analyses can be found in Appendix 1 and Appendix 2.

For the second part, we considered two simple linear regression models, also in R, predicting average income per person from the variable crimes. The difference of these two linear regression models is the expression use of the variable crimes. For one model, we used the number of crimes as the predictor variable, but for the other model, we used the crime rate, defined as the number of crimes divided by total population, to be the predictor variable. We took the interaction term into consideration and examined the summary table of the linear regression model and four case-wise residual diagnostic plots, including Residuals vs. Fitted plot, Normal Q-Q plot, Scale-Location plot, and Residuals vs. Leverage plot, respectively, to select the best model using each expression of crimes. Then, we compared those two candidate models by putting them in real-world settings and choosing the one which has more practical meanings.

For the third part, we considered two multiple regression models, also in R, by using subsets regression and stepwise regression. They are able to help predict the per-capita income from each of the potential variables in the dataset, since multiple regression can tell us about the effect of each individual predictor variable, after controlling for all other predictor variables. For using the subsets regression, we considered the criteria of picking the model with maximum adjusted R-Squared, minimum Cp value, and minimum BIC value, respectively to choose the best model for the method of subsets regression. For using the stepwise regression, we considered the criteria of using forward selection on the minimum AIC value to choose the best model fitting the prediction relationship. Then, we compare the two candidate models by examining their summary table of statistics coefficients and case-wise diagnostic plots. Details of these analyses in R can be found in the Appendices 3,4, and 5.

Analyses were carried out in R and RStudio (RStudio Team, 2020).

4 Results

4.1 Visual Observation of Exploratory Plots

We investigate the closely related relationship between groups of variables by plotting the correlation plot between different variables and compare them one-to-one. From the correlation plot, we can notice that the darker colors and bigger size the circle is, the more connected the relationship, i.e. the bigger correlation, the two variables have. There are many groups of correlation relationships in the dataset between two variables. To be more specific, the most significant variables that are closely related to variable population are variable doctors, hosp.beds, and crimes; the most significant variables that are closely related to variables

doctors are variable hosp.beds, crimes, and tot.income; the most significant variables that are closely related to tot.income are variable pop, doctors, hosp.beds, and crimes. Later, we show the boxplot for categorical variable region by plotting each region's boxplot. From four boxplots in the overall box plot graphs, we can observe that for region "S," it has the most amount numbers of outliers and for the region "NE," it has the biggest value of median and for the region "S," it has the smallest value of median. Moreover, we can observe that for the region "NE," data points is evenly distributed but for the region "S" and "W," data points have more dispersions compared to data points in the region "NE."

As for relating the above two plots into the real-world setting, we can grasp that both variable population and total income have association with doctors, hosp.beds, and crimes before doing any model fitting on different groups of variables to select the optimal model. With this fact in mind, we can have a preliminary direction of what variables should be included in the model at the first glance. Moreover, with the boxplots, we can know that people who live in the region "NE" tend to have very evenly distributed per capita income; however, people who live in the region "S" tend to have the most extreme and not well dispersed per capita income.

4.2 Regression Analysis

4.2.1 Simple Linear Regression Analysis

With the question of investigating the relationship between the per-capita income and crime rate and region of the country, we first do a linear regression model without adding any interaction term on the original data. With the problem of having extremely low adjusted R-squared value 0.09288, the violation on both linearity and normality assumptions, and the appearance of non-constant variance problem, we decide to make possible data transformation to solve the non-linearity, non-normality, and non-constant variance problems.

Following that, we first observe the histogram of those variables to decide whether they need any kind of transformation. From the histogram of per-capita income and crime rate respectively, we can observe that the histogram of per-capita shows slightly right-skewed distribution, so we decide to do log-transformation on the variable per-capita income. Also, from the histogram of crime rates, we can clearly observe that it shows right-skewed distribution, so we decide to also do log-transformation on crime rates as well. After doing the model of data transformation, we fit the model again and get the result that the adjusted R-squared value improves a lot and the linearity condition is satisfied. However, there are still problems regarding to violate the normality assumption and cause non-constant variance problems. Then, we decide to check whether it is necessary to include the interaction term in the model. We create the ANOVA table and compare the model with transformed data added interaction term between crime rates and regions to the model with only transformed data. From the ANOVA table, we can observe that with the F-statistics=0.7434, there is not enough evidence against the reduced model in favor of the full model. Also, given the p-value 0.5266, which is greater than 0.05, we conclude that we cannot tell whether there is an association between the crime rates and the region of the country, therefore we cannot reject the hypothesis that there is no difference on including the interaction term. In other words, the interaction term can not be included in the model. Therefore, as for using the log(crimes) as our predictor variables, we choose the model without

interaction term to be our final model, with the diagnostic plots and estimator coefficients presented:

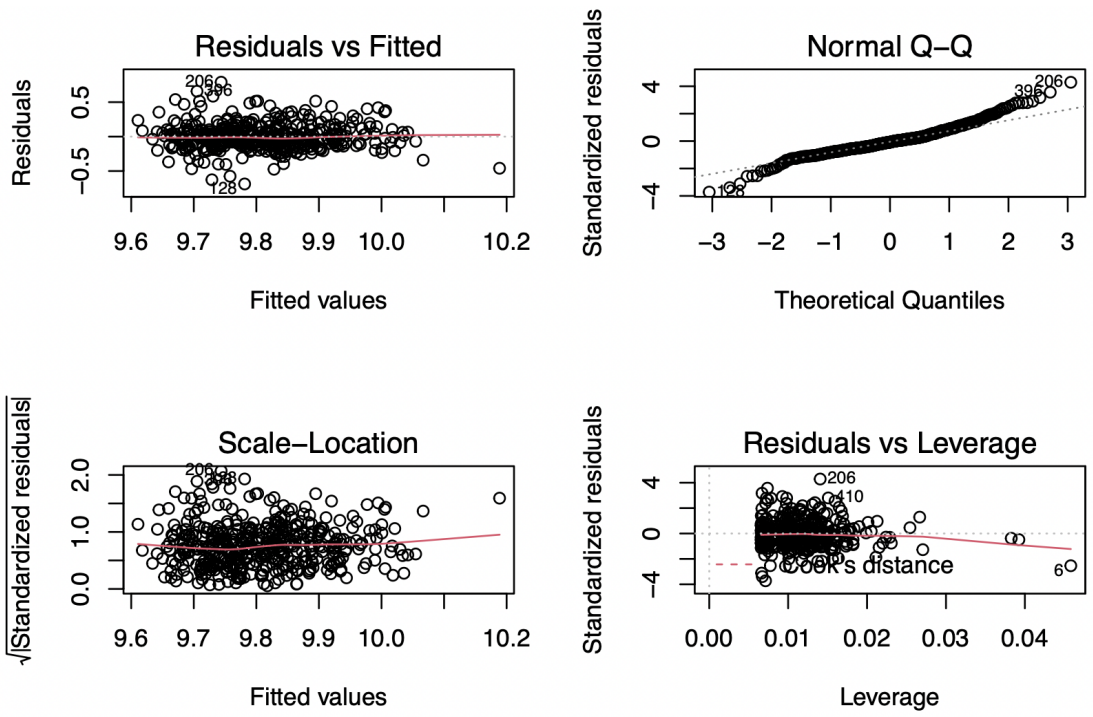


Figure 4: Four Diagnostic Plots of Simple Linear Regression Chosen Model

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.68757 -0.10557 -0.01422  0.08905  0.78946

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.188431   0.079812 115.125  < 2e-16 ***
log(crimes)  0.066695   0.008421   7.920 2.00e-14 ***
regionNE     0.104458   0.025531   4.091 5.11e-05 ***
regionS     -0.086983   0.023618  -3.683  0.00026 ***
regionW     -0.055280   0.028167  -1.963  0.05033 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1854 on 435 degrees of freedom
Multiple R-squared:  0.2032,    Adjusted R-squared:  0.1959
F-statistic: 27.74 on 4 and 435 DF,  p-value: < 2.2e-16
```

Table 4: Summary of Coefficient Estimators of Linear Regression Chosen Model

To explain the coefficient estimator, we can state that for every one unit increase in the number of crime cases, the per-capita income can increase 1.07 dollars.

Later, we would like to investigate whether using the number of crimes or using per-capita crime (which is defined as number of crimes/population) will make any difference on choosing the best model. First, we do model fitting with transformed data by replacing log(crimes) with per-capita crime measure. From the summary of data with using per-capita crime, we can observe that the adjusted R-squared is 0.1941, meaning that 19.41% of its variability can be explained, and the adjusted R-squared value does not change a lot comparing to previous model with using the number of crimes as the variable. Also, looking back to the four diagnostic plots, they show that the model has a very similar performance compared to the previous model with using the number of crimes as the variable. From the Residuals vs. Fitted plot, we can clearly see that the linearity condition is satisfied since there is a horizontal line around 0. From the Normal Q-Q plot, we can clearly observe that the normality assumption is not completely satisfied because there is still an apparent outlier for example point 206. From the Scale-Location plot, we can clearly observe that there is not a non-constant variance problem with the nearly horizontal line around 1. From the Residuals vs. Leverage plot, we can clearly observe that there are several outliers for example point 206 (greater or lower than the absolute value of 2) and high leverage point appeared like point 1. Furthermore, we want to check whether it is necessary to include the interaction term in the model using “per-capita crime” as the predictor variable. We create the ANOVA table and compare the model with transformed data added interaction term between “per-capita crime” and regions to the model with only transformed data. From the above ANOVA table, we can observe that with the F-statistics=0.8816, there is not enough evidence against the reduced model in favor of the full model. Also, given the p-value 0.5082, which is greater than 0.05, we conclude that we cannot tell whether there is an association between “per-capita crime” and the region of the country, therefore we cannot reject the hypothesis that there is no difference on including the interaction term. In other words, the interaction term can not be included in the model. Therefore, as for using the log(crimes)/pop as our predictor variables, we choose the model without interaction term to be our final model, and the diagnostic plots and estimator coefficients presented as:

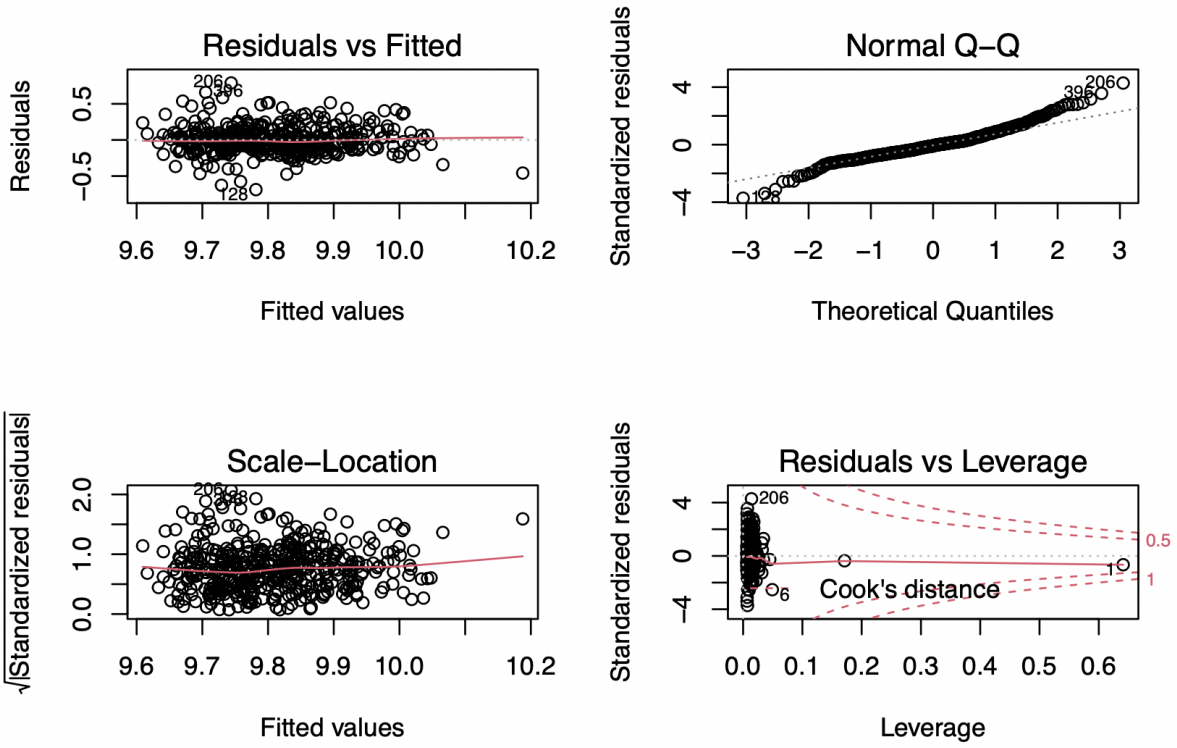


Figure 5: Four Diagnostic Plots of Simple Linear Regression Chosen Model (per person)

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.68786 -0.10559 -0.01417  0.08906  0.78930

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.183e+00  9.995e-02  91.869  < 2e-16 ***
log(crimes)      6.739e-02  1.102e-02   6.118 2.12e-09 ***
regionNE         1.045e-01  2.557e-02   4.088 5.19e-05 ***
regionS         -8.731e-02  2.388e-02  -3.656 0.000287 ***
regionW         -5.529e-02  2.820e-02  -1.961 0.050557 .
log(crimes):pop -1.466e-10  1.503e-09  -0.098 0.922344
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1856 on 434 degrees of freedom
Multiple R-squared:  0.2033,    Adjusted R-squared:  0.1941
F-statistic: 22.14 on 5 and 434 DF,  p-value: < 2.2e-16
```

Table 5: Summary of Coefficient Estimators of Linear Regression Chosen Model (per person)

To explain the coefficient estimator, we can state that for every one unit increase in the per-capita crime cases, the per-capita income can increase 1.07 dollars.

To be concluded, from the above analysis, we can observe that whether using the number of crimes or using the “per-capita income” as the predictor variable does not influence my final picked model that much since the adjusted R-Squared, the four diagnostic plots, ANOVA table, and F-statistics are pretty close comparing the model with “the number of crimes” and “per-capita crime”. As for choosing the better model to answer the question, we need to consider the real-world setting environment. Here, we want to investigate the model to predict the per-capita income from crime rate and region of the country. Then, I prefer using the model with the “per-capita crime” = number of crimes/population as the crime rate measure with two following reasons. First, if we introduce the concept of “per-capita crime,” then we make an accurate definition for per-capita crime and it is statistically cautious. Second, as we would like to predict per-capita income, it would be better to use the concept of “per-capita crime” to have a consistent unit as the analysis further goes.

Therefore, we prefer choosing the model using “per-capita crime” as the measure and without interaction term to be our final model.

4.2.2 Multiple Regression Analysis

The first step of doing the multiple regression analysis is to identify the distribution of variables and decide whether to make data transformation on all potential variables. So here, we plot the histogram on each potential predictor variable again to make sure whether there is a need for data transformation on each of them. The result turns out that for variables “Land Area,” “Total Population,” “Number of Active Physician,” “Number of Hospital Beds,” “Percent Below Poverty Level,” “Percent Bachelor’s Degrees,” “Percent Unemployment,” and “Total Income,” their histograms look right-skewed and need log-transformation on the data. For variables “Percent of Population Aged 18-34” and “Percent of Population Aged 65 or Older,” the distribution looks normal and there is no need for data transformation. For variable “Percent High School Graduates,” the histogram looks left-skewed and needs squared-transformation on the data.

Since we know that the response variable per.cap.income is mathematically calculated by tot.income divided by pop, then we remove two variables: tot.income and pop when fitting the multiple regression model.

With all the above presented data transformation on each numerical variable, we fit the multiple regression model and have both the adjusted R-Squared and the residual standard error perform pretty well. With the adjusted R-Squared value of 0.87, meaning that 87% of its variability can be explained. Also, the residual standard error is 0.07455, which means that the model fits well for the dataset.

Then, we perform subsets regression analysis to observe the suitable multiple regression model for the dataset and consider different criteria including choosing the model with maximum adjusted R-Squared value, choosing the model with minimum Cp value, or choosing the model with minimum BIC value in order to select the best model. From the summary table below, we can observe that no matter whether we choose to use the criteria of adjusted R-Squared, Cp, or BIC, we all should choose the model with 9 predictor variables to be our best model.

```r
# with different criteria to select the best model
cdi_sum<-summary(mulreg_fit2)
data.frame(
  Adj_R2 = which.max(cdi_sum$adjr2),
  CP = which.min(cdi_sum$cp),
  BIC = which.min(cdi_sum$bic)
)
```

```
##   Adj_R2 CP BIC
## 1      9  9   9
```

Table 6: Summary Table of Choosing Model in Subsets Regression

Therefore, the best model would be:

log(per.cap.income)~log(land.area)+pop.18_34+log(doctors)
+log(pct.below.pov)+log(pct.bach.deg)+stateCA+stateNJ+stateNV+stateUT

Next, we would like to perform our model selection by using stepwise regression and create the summary table. From the summary table, we can observe that if we choose to use the criteria of AIC, then we should choose the model with smallest AIC values, which the best model would be:

log(per.cap.income) ~ log(land.area) + pop.18_34 + pop.65_plus + log(doctors) +
log(pct.below.pov) + log(pct.bach.deg) + state;

With the fact that we have one best model both from the subset regression and the stepwise regression, we tend to compare those two models by plotting their corresponding residual diagnostic plots and summary table. By comparing the candidate models from subsets regression and stepwise regression, we decide to choose the model selecting from subsets regression as our final model. As for the value of adjusted R-Squared and Residual Standard Error, both of the two models have pretty much the same performance. However, looking into the variables, we believe that the model selecting from subsets regression has more specific preference on influential states in doing the prediction. With the consideration to explain our model to someone who is more interested in economic factors, the model with specific states can be more convincing. Therefore, our preferred final model would be:

log(per.cap.income)~log(land.area)+pop.18_34+log(doctors)+log(pct.below.pov)+log(pct.bach.deg)+stateCA+stateNJ+stateNV+stateUT

5 Discussion

The study aims to help social scientists gain first-hand information on the relationship between the average income per person and other potential factors, and thus helping them understand the current situation of a well being's income status in the United States and determine further directions on understanding how the average income per person in the United States can reflect social and economic problems.

In our correlation plot, both variable population and total income have association with doctors, hosp.beds, and crimes. Moreover, with the boxplots, we can know that people who live in the region "NE" tend to have very evenly distributed per capita income; however, people who live in the region "S" tend to have the most extreme and not well dispersed per capita income.

With the question of investigating the relationship between the per-capita income and crime rate and region of the country, we plot histograms and notice that there is a need to do data transformation on both the per-capita income and crime rate. Later, we perform linear regression model fitting and ANOVA table to show that as for using the log(crimes) as our predictor variables, we choose the model without interaction term to be our final model, and the mathematical association should be:

$$\log(\text{per.cap.income}) \sim \log(\text{crimes}) + \text{region}$$

Later, we investigate whether using the number of crimes or using per-capita crime (which is defined as number of crimes/population) will make any difference on choosing the best model. By changing the measure to per-capita crime and do linear regression model fitting, we find that the model without interaction term to be our final model, and the mathematical association should be:

$$\log(\text{per.cap.income}) \sim (\log(\text{crimes})/\text{pop}) + \text{region}$$

Comparing the above two models in real-world settings, we decide to choose using per-capita crime as our measure and the final model is:

$$\log(\text{per.cap.income}) \sim (\log(\text{crimes})/\text{pop}) + \text{region}$$

Besides, we also use multiple regression analysis including subsets regression and stepwise regression to output the best model for the dataset. Considering the summary table of the model and the rule of choosing model with fewer variables, we decide to choose as our final model as presented:

$$\log(\text{per.cap.income}) \sim \log(\text{land.area}) + \text{pop.18\_34} + \log(\text{doctors}) + \log(\text{pct.below.pov}) + \log(\text{pct.bach.deg}) + \text{stateCA} + \text{stateNJ} + \text{stateNV} + \text{stateUT}$$

If say, social scientists really want to explore more in the relationship between average income per person and other potential variables that are associated with the country's economic, health, and social well-being, the first thing that social scientists should be aware of is to better use the concept of per-capita income, which is defined as the total income divided by total population as a measurement of the term "average income per person." The second thing

is that as we go through deeply in the final model that we've selected from the multiple regression model, those variables make sense in various aspects. At the beginning of the EDA, we plot the correlation between different continuous variables in the dataset and notice that variables total income and population are highly related to doctors, crimes, and hosp.beds. As we've discussed and proved that there is an association between per-capita income and crime rate, we focus on developing and proving that there is an association between per-capita income and doctors. As for considering other variables included in our final model, they can be well-explained by putting them in real-world context. For variable "pop.18_34," they can be considered as the most significant group of the population that can make huge contributions on the average income per person since normally people in this age are in the development/peak of their career. For variables "percent high school graduates" and "percent bachelor's degrees," we can classify them as individuals' education background information, and it can be related to average income per person since the fact that people with higher education background are more inclined to enter well-known and high level businesses, and thus earning more money. For variables "land.area" and "stateCA, stateNJ, stateNV, and stateUT," we can classify them as having geographical location impact on average income per person. With many high-tech companies located and plenty of landmass, state California can be an influential state factor. With the fact that Las Vegas, located in the heart of Nevada, has countless casinos and can make profitable money through running the casino business, so it can be explained as an influential state factor as well. With lots of job opportunities and prestigious institutions located in New Jersey, students who graduate from there might have jobs with an acceptable salary and therefore, state New Jersey can be counted as an influential factor as well.

There is scope for establishing the final model to be in real-world settings; indeed, the dataset is the 1990 CDI population. Therefore, if the social scientist want to use our model to predict the current situation of the relationship between average income per person and other variables, there might be bias and concerns since the variables "total population," "percent of population aged 18-34," "percent of population 65 or older," haven't been updated for nearly 20 years. One recommendation on solving this problem might be to first update or compare the real-time data with the 1990 CDI population data. By doing time series analysis, we can investigate whether there is a linear trend on the growth of the population and each sub-category population, then we can do further research on our research question.

Our study was also limited by the fact that in the dataset, we only have 48 out of 51 total states represented as the state level data and 373 out of 3000 total counties represented as the county level data. It can be a concern since it is not known that for the countries that we are missing, which state do those countries belong to since counties located in different states can be considered as a potential influential factor on the average income per person. One recommendation for this limitation is to collect more useful information and comprehensive real-time data on more counties in the United States in order to provide an unbiased characterization of the relationship between average income per person and other variables in the United States now.

In summary, keeping the caveats of the last two paragraphs in mind, there is scope to establish the relationship that there is an association between per-capita income and land area,

percent of population aged 18-34, number of active physicians, percent below poverty level and percent bachelor;s degrees and state of California, New Jersey, Nevada, and Utah.

6 References

R Core Team (2017), *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2020), *R Studio: Integrated Development Environment for R.* RStudio, PBC, Boston MA. URL http://www.rstudio.com.

Kutner, M.H., Nachsheim, C.J., Neter, J. & Li, W. (2005), *Applied Linear Statistical Models, Fifth Edition*. NY: McGraw-Hill/Irwin.

Kutner et al. (2005). Original source:Geospatial and Statistical Data Center, University of Virginia.

Appendices: Regression Analysis on Average Income Per Person Associate with Other Variables in CDI

Sifeng Li

10/18/2021

Contents

Appendix 1. Initial Data Import & Exploration

```
cdi<-read.table('/Users/sifengli/Desktop/CMU/Fall 2021/Applied Linear Models/cdi.dat')
#cdi
```

```
#install.packages('plyr')
library(plyr)
```

```
summary(cdi)
```

```
##        id           county             state              land.area
##  Min.   :  1.0   Length:440         Length:440         Min.   :   15.0
##  1st Qu.:110.8   Class :character   Class :character   1st Qu.:  451.2
##  Median :220.5   Mode  :character   Mode  :character   Median :  656.5
##  Mean   :220.5                                         Mean   : 1041.4
##  3rd Qu.:330.2                                         3rd Qu.:  946.8
##  Max.   :440.0                                         Max.   :20062.0
##       pop             pop.18_34      pop.65_plus        doctors
##  Min.   : 100043   Min.   :16.40   Min.   : 3.000   Min.   :   39.0
##  1st Qu.: 139027   1st Qu.:26.20   1st Qu.: 9.875   1st Qu.:  182.8
##  Median : 217280   Median :28.10   Median :11.750   Median :  401.0
##  Mean   : 393011   Mean   :28.57   Mean   :12.170   Mean   :  988.0
##  3rd Qu.: 436064   3rd Qu.:30.02   3rd Qu.:13.625   3rd Qu.: 1036.0
##  Max.   :8863164   Max.   :49.70   Max.   :33.800   Max.   :23677.0
##    hosp.beds           crimes        pct.hs.grad     pct.bach.deg
##  Min.   :   92.0   Min.   :   563   Min.   :46.60   Min.   : 8.10
##  1st Qu.:  390.8   1st Qu.:  6220   1st Qu.:73.88   1st Qu.:15.28
##  Median :  755.0   Median : 11820   Median :77.70   Median :19.70
##  Mean   : 1458.6   Mean   : 27112   Mean   :77.56   Mean   :21.08
##  3rd Qu.: 1575.8   3rd Qu.: 26280   3rd Qu.:82.40   3rd Qu.:25.32
##  Max.   :27700.0   Max.   :688936   Max.   :92.90   Max.   :52.30
##  pct.below.pov      pct.unemp      per.cap.income    tot.income
##  Min.   : 1.400   Min.   : 2.200   Min.   : 8899   Min.   :  1141
##  1st Qu.: 5.300   1st Qu.: 5.100   1st Qu.:16118   1st Qu.:  2311
##  Median : 7.900   Median : 6.200   Median :17759   Median :  3857
```

```
##  Mean   : 8.721   Mean   : 6.597   Mean   :18561   Mean   :  7869
##  3rd Qu.:10.900   3rd Qu.: 7.500   3rd Qu.:20270   3rd Qu.:  8654
##  Max.   :36.300   Max.   :21.300   Max.   :37541   Max.   :184230
##     region
##  Length:440
##  Class :character
##  Mode  :character
##
##
##
```

```
sd(cdi$land.area)
```

```
## [1] 1549.922
```

```
sd(cdi$pop)
```

```
## [1] 601987
```

```
sd(cdi$pop.18_34)
```

```
## [1] 4.191083
```

```
sd(cdi$pop.65_plus)
```

```
## [1] 3.992666
```

```
sd(cdi$doctors)
```

```
## [1] 1789.75
```

```
sd(cdi$hosp.beds)
```

```
## [1] 2289.134
```

```
sd(cdi$crimes)
```

```
## [1] 58237.51
```

```
sd(cdi$pct.hs.grad)
```

```
## [1] 7.015159
```

```
sd(cdi$pct.bach.deg)
```

```
## [1] 7.654524
```

```
sd(cdi$pct.below.pov)
```

```
## [1] 4.656737
```

```
sd(cdi$pct.unemp)
```

```
## [1] 2.337924
```

```
sd(cdi$per.cap.income)
```

```
## [1] 4059.192
```

```
sd(cdi$tot.income)
```

```
## [1] 12884.32
```

```
count(cdi,'region')
```

```
##   region freq
## 1     NC  108
## 2     NE  103
## 3      S  152
## 4      W   77
```

```
count(cdi,'state')
```

```
##    state freq
## 1     AL    7
## 2     AR    2
## 3     AZ    5
## 4     CA   34
## 5     CO    9
## 6     CT    8
## 7     DC    1
## 8     DE    2
## 9     FL   29
## 10    GA    9
## 11    HI    3
## 12    ID    1
## 13    IL   17
## 14    IN   14
## 15    KS    4
## 16    KY    3
## 17    LA    9
## 18    MA   11
## 19    MD   10
## 20    ME    5
## 21    MI   18
## 22    MN    7
## 23    MO    8
## 24    MS    3
## 25    MT    1
## 26    NC   18
## 27    ND    1
## 28    NE    3
## 29    NH    4
## 30    NJ   18
## 31    NM    2
## 32    NV    2
## 33    NY   22
## 34    OH   24
## 35    OK    4
## 36    OR    6
## 37    PA   29
## 38    RI    3
## 39    SC   11
## 40    SD    1
## 41    TN    8
## 42    TX   28
## 43    UT    4
```

```
## 44    VA    9
## 45    VT    1
## 46    WA   10
## 47    WI   11
## 48    WV    1
```

```
county_freq<-data.frame(summary(as.factor(cdi$county)))
transform(county_freq,County_Frequency=ave(seq(nrow(county_freq)),cdi$county,FUN=length))
```

```
##                  summary.as.factor.cdi.county.. County_Frequency
## Jefferson                                     7                 1
## Montgomery                                    6                 1
## Washington                                    5                 1
## Cumberland                                    4                 1
## Jackson                                       4                 2
## Lake                                          4                 1
## Clark                                         3                 1
## Hamilton                                      3                 1
## Kent                                          3                 1
## Madison                                       3                 1
## Marion                                        3                 1
## Middlesex                                     3                 1
## Monroe                                        3                 1
## Orange                                        3                 1
## Wayne                                         3                 1
## York                                          3                 2
## Allen                                         2                 1
## Bay                                           2                 2
## Butler                                        2                 1
## Calhoun                                       2                 1
## Clay                                          2                 1
## Davidson                                      2                 1
## Delaware                                      2                 1
## El_Paso                                       2                 1
## Erie                                          2                 1
## Essex                                         2                 1
## Fairfield                                     2                 1
## Fayette                                       2                 1
## Franklin                                      2                 1
## Greene                                        2                 1
## Hillsborough                                  2                 1
## Kings                                         2                 1
## Lancaster                                     2                 1
## Mercer                                        2                 1
## Richland                                      2                 1
## St._Clair                                     2                 1
## St._Louis                                     2                 1
## Suffolk                                       2                 1
## Winnebago                                     2                 1
## Ada                                           1                 1
## Adams                                         1                 1
## Aiken                                         1                 1
## Alachua                                       1                 1
## Alamance                                      1                 1
## Alameda                                       1                 1
```

```
## Albany                    1               1
## Alexandria_City           1               2
## Allegheny                 1               3
## Anderson                  1               1
## Androscoggin              1               1
## Anne_Arundel              1               1
## Arapahoe                  1               1
## Arlington_County          1               1
## Atlantic                  1               1
## Baltimore                 1               1
## Baltimore_City            1               1
## Barnstable                1               1
## Beaver                    1               3
## Bell                      1               2
## Benton                    1               1
## Bergen                    1               2
## Berks                     1               2
## Berkshire                 1               1
## Bernalillo                1               1
## Berrien                   1               1
## Bexar                     1               2
## Bibb                      1               2
## Blair                     1               2
## Boone                     1               1
## Boulder                   1               1
## Brazoria                  1               1
## Brazos                    1               1
## Brevard                   1               1
## Bristol                   1               1
## Broome                    1               1
## Broward                   1               1
## Brown                     1               1
## Bucks                     1               1
## Buncombe                  1               1
## Burlington                1               3
## Butte                     1               1
## Caddo                     1               1
## Calcasieu                 1               1
## Cambria                   1               1
## Camden                    1               1
## Cameron                   1               1
## Carroll                   1               2
## Cass                      1               1
## Catawba                   1               1
## Centre                    1               1
## Champaign                 1               1
## Charles                   1               1
## Charleston                1               1
## Charlotte                 1               1
## Chatham                   1               1
## Chautauqua                1               1
## Chesapeake_City           1               1
## Chester                   1               1
## Chittenden                1               1
```

```
## (Other)                                   274                 2
```

```
median(county_freq$summary.as.factor.cdi.county..)
```

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

```
state_freq<-data.frame(summary(as.factor(cdi$state)))
transform(state_freq,State_Frequency=ave(seq(nrow(state_freq)),cdi$state,FUN=length))
```

```
##    summary.as.factor.cdi.state.. State_Frequency
## AL                              7                9
## AR                              2                2
## AZ                              5                4
## CA                             34                9
## CO                              9                9
## CT                              8                5
## DC                              1                1
## DE                              2                2
## FL                             29                5
## GA                              9                4
## HI                              3                2
## ID                              1                1
## IL                             17                9
## IN                             14                9
## KS                              4                3
## KY                              3                1
## LA                              9                2
## MA                             11                5
## MD                             10                5
## ME                              5                9
## MI                             18                5
## MN                              7                4
## MO                              8                9
## MS                              3                4
## MT                              1                2
## NC                             18                9
## ND                              1                1
## NE                              3                1
## NH                              4                5
## NJ                             18                3
## NM                              2                1
## NV                              2                5
## NY                             22                3
## OH                             24                5
## OK                              4                3
## OR                              6                5
## PA                             29                1
## RI                              3                5
## SC                             11                3
## SD                              1                1
## TN                              8                2
## TX                             28                1
## UT                              4                3
## VA                              9                9
## VT                              1                1
```

```
## WA                                10               2
## WI                                11               2
## WV                                 1               1
```

```
median(state_freq$summary.as.factor.cdi.state..)
```

```
## [1] 7
```

For the Data Description, please refer to the end of the homework document (last page) Sorry for any inconvenience!

```
which(is.na(cdi))
```

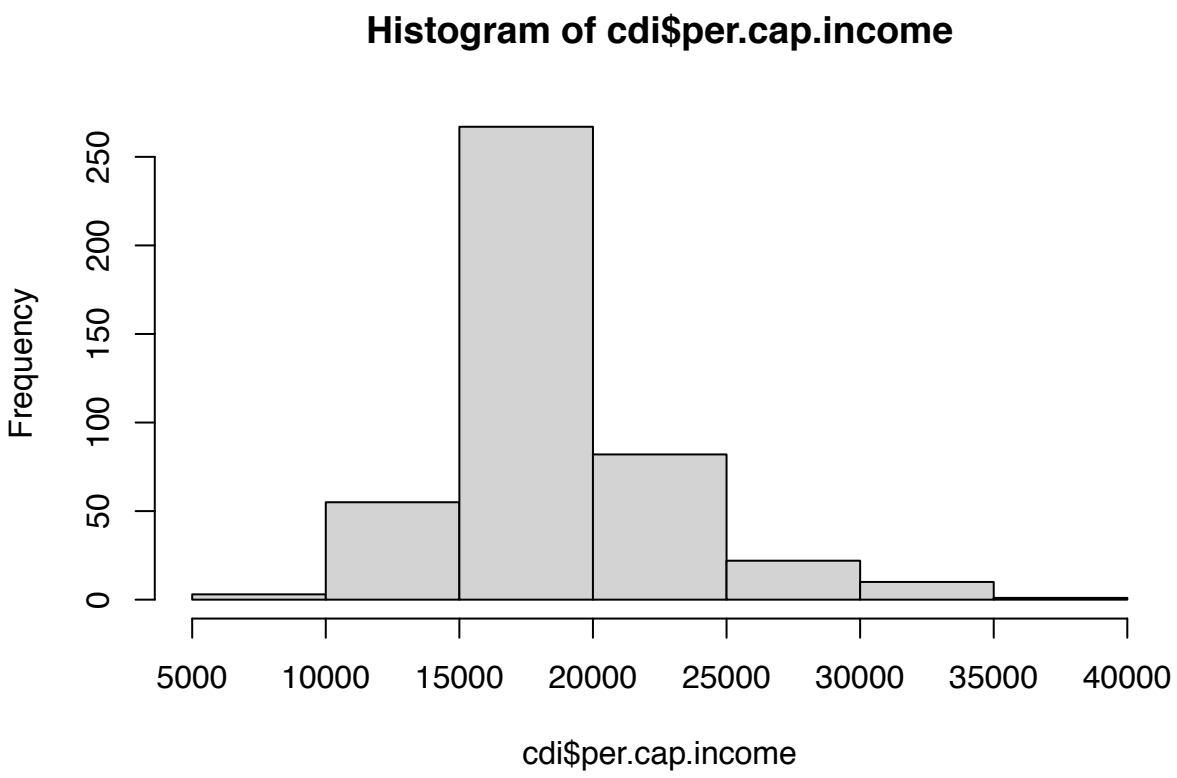
```
## integer(0)
```

As for checking missing values before processing further analysis, we've noticed that there is no missing data in this dataset.

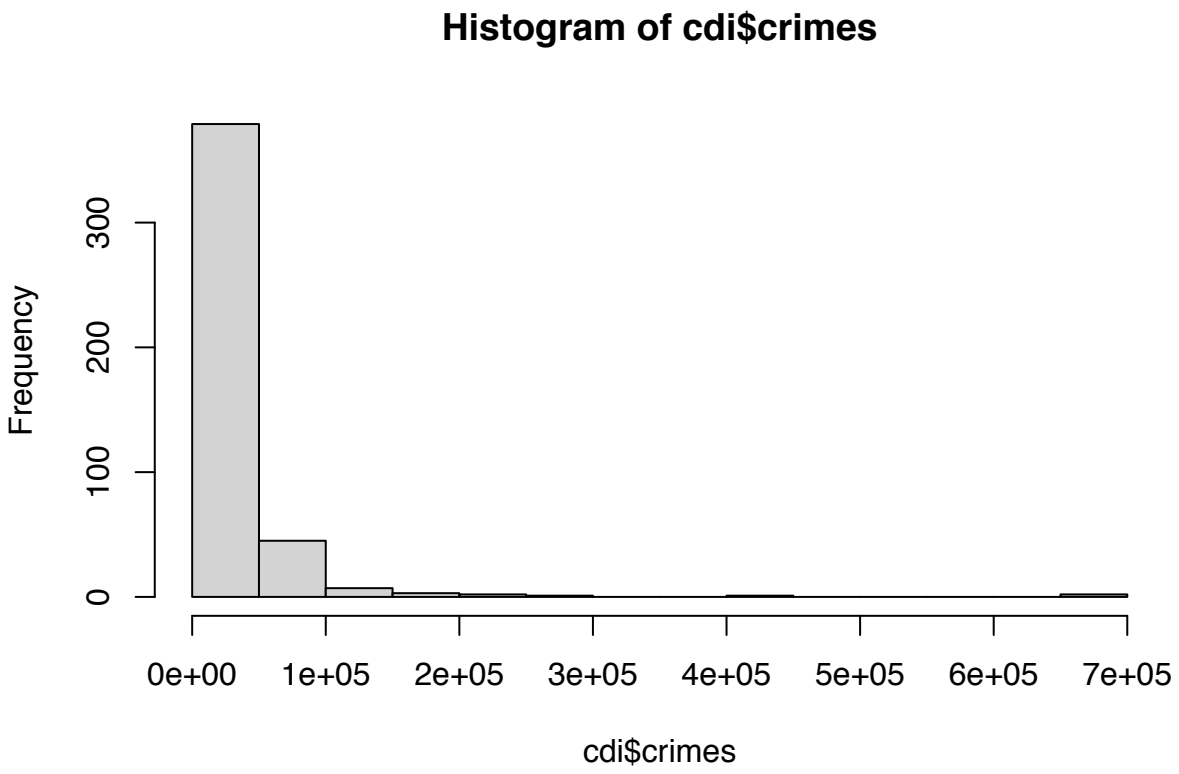
Next, we make some appropriate descriptive EDA plots as the following presented:

```
library(psych)
```

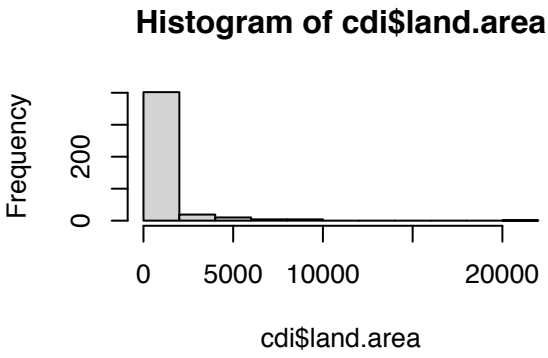
```
hist(cdi$per.cap.income)
```

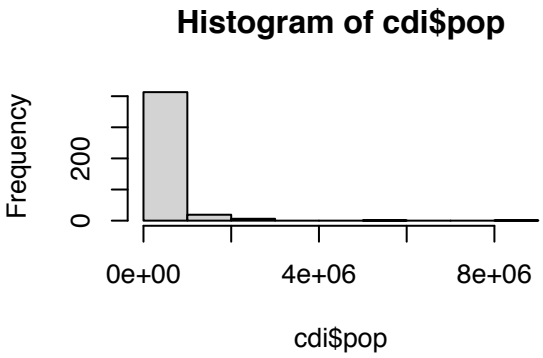


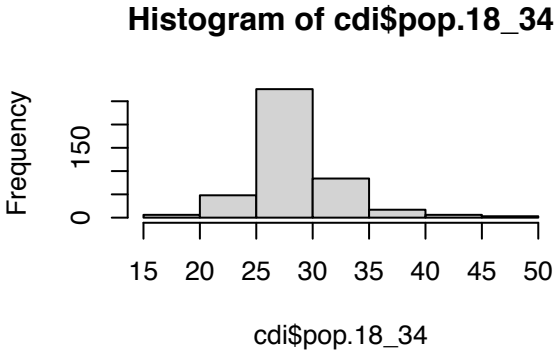

```
hist(cdi$crimes)
```

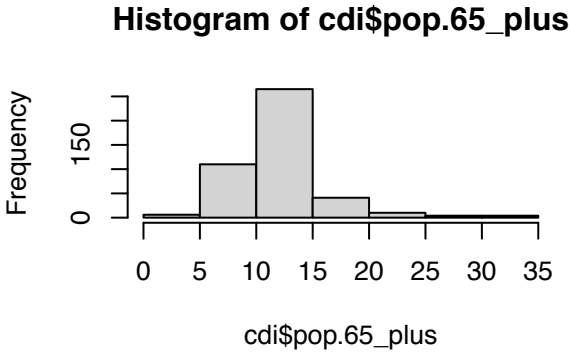



```
par(mfrow=c(2,2))
hist(cdi$land.area)
hist(cdi$pop)
hist(cdi$pop.18_34)
hist(cdi$pop.65_plus)
```

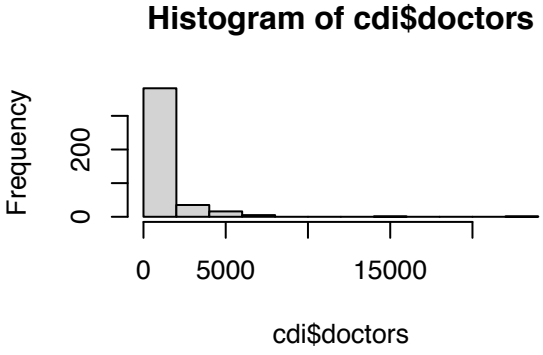


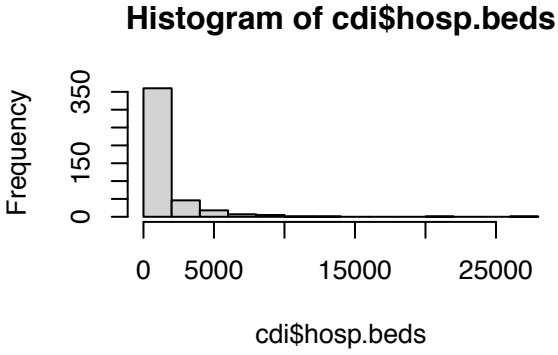


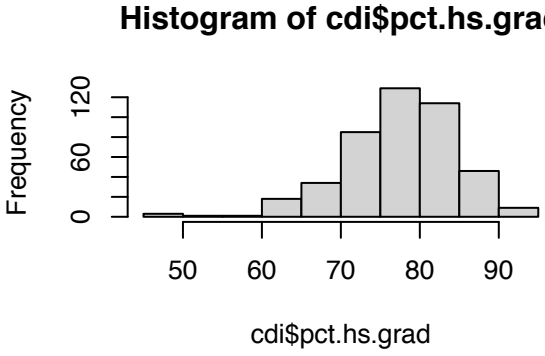


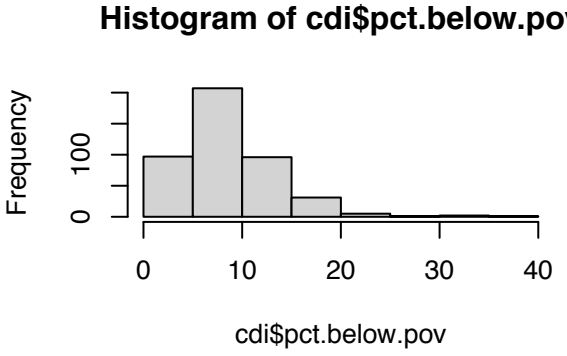



```
hist(cdi$doctors)
hist(cdi$hosp.beds)
hist(cdi$pct.hs.grad)
hist(cdi$pct.below.pov)
```

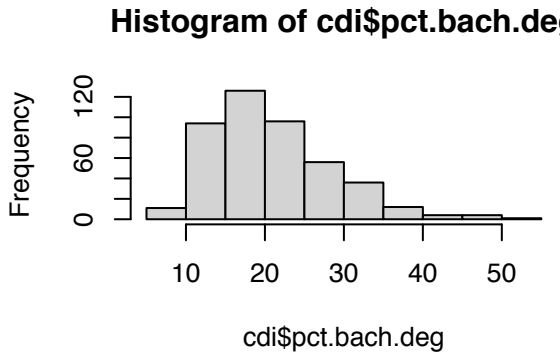


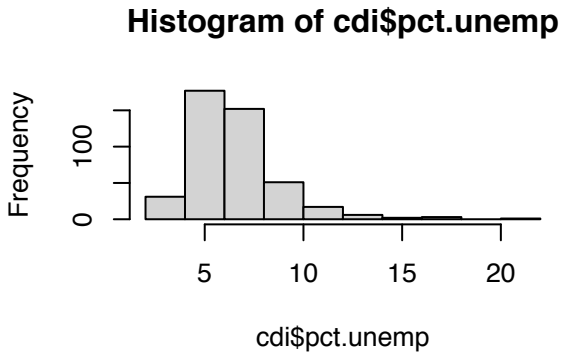


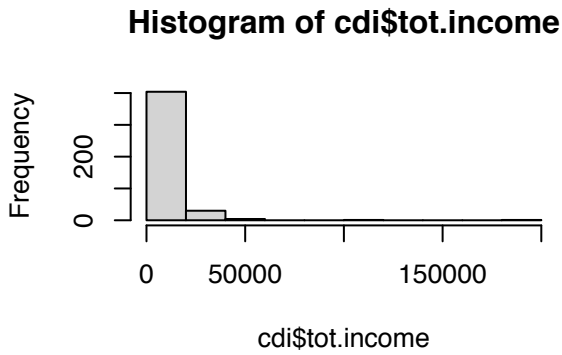




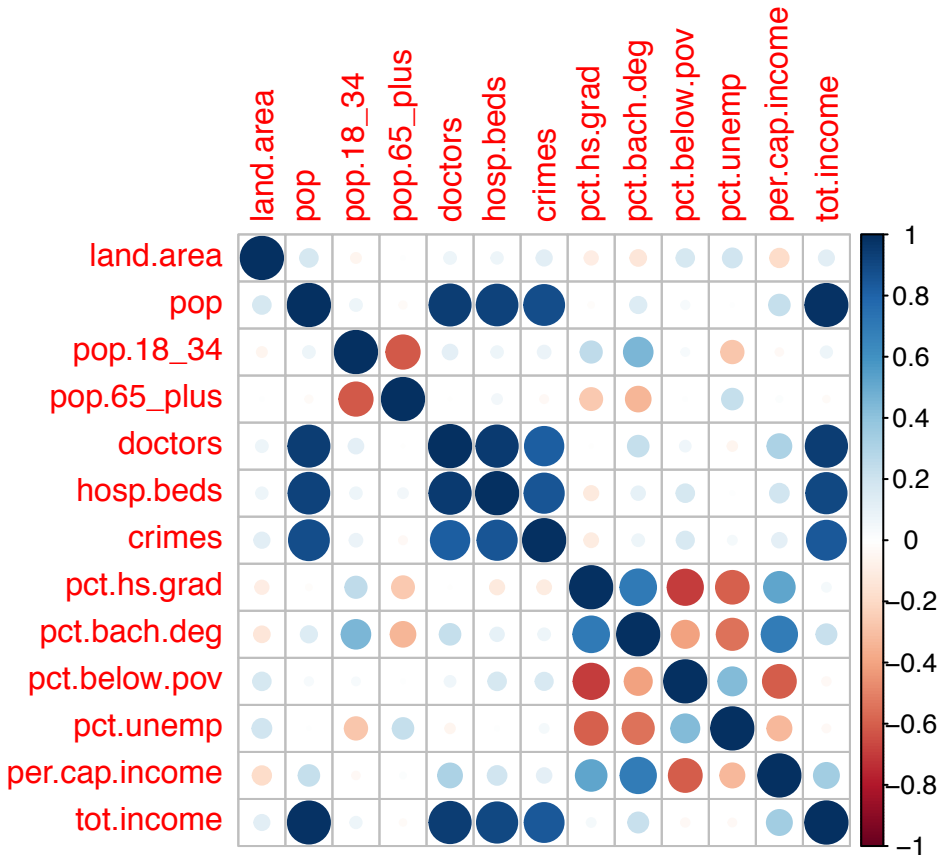
```
hist(cdi$pct.bach.deg)
hist(cdi$pct.unemp)
hist(cdi$tot.income)
```





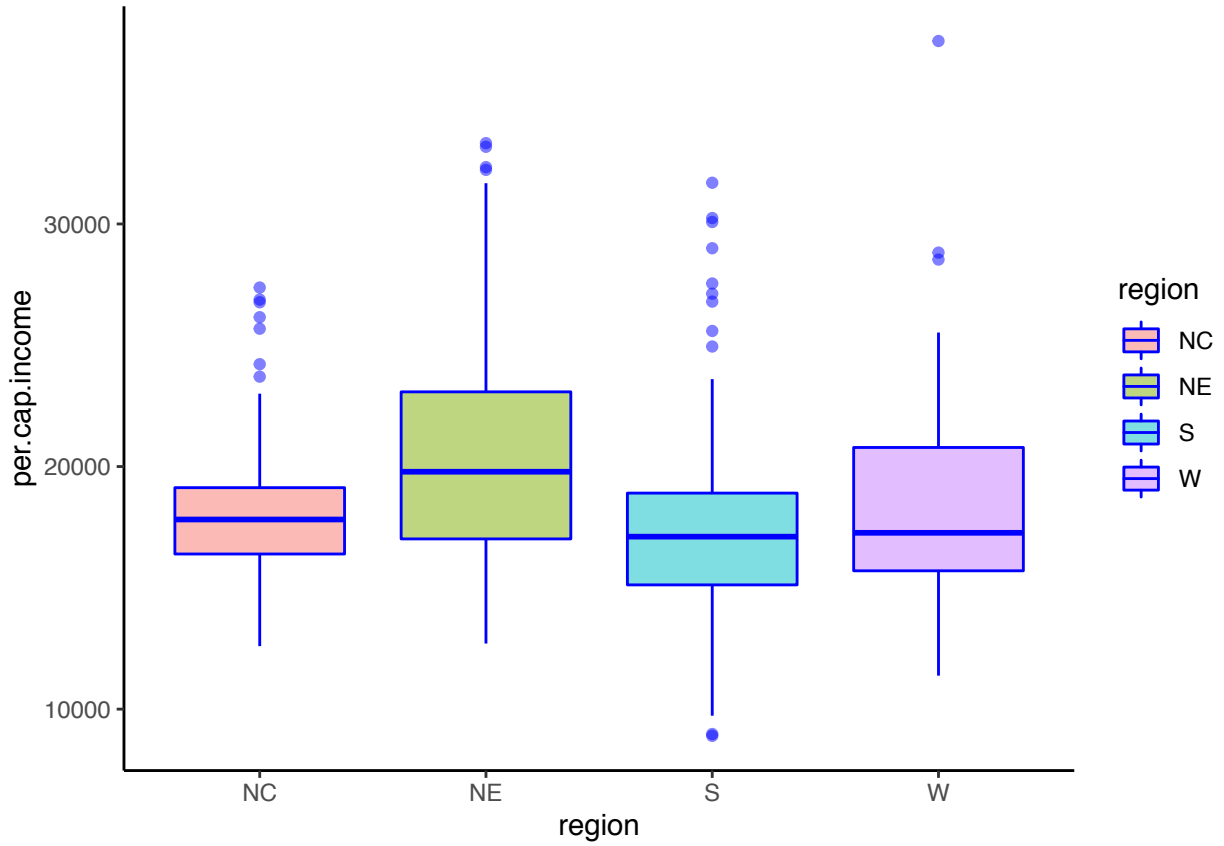



```
library(corrplot)
cdi_corr<-cdi[,-c(1:3,17)]
C <- cor(cdi_corr)
corrplot(C,method="circle")
```

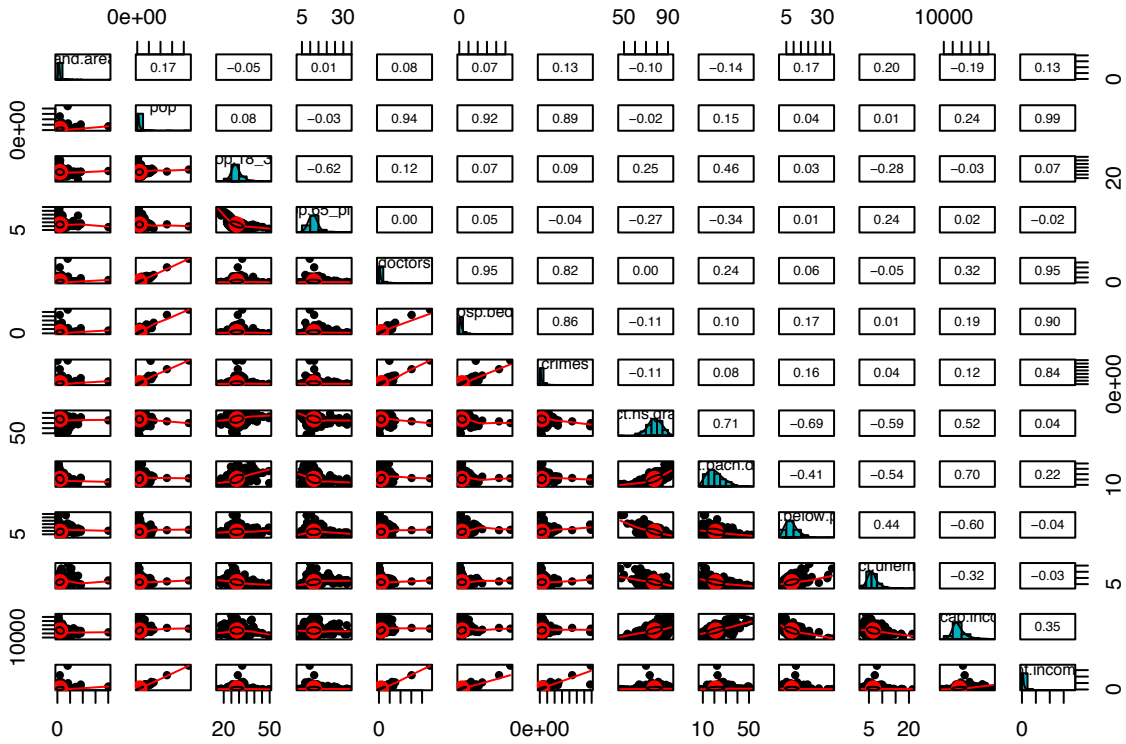


From the correlation plot, we can notice that the darker colors and bigger size the circle is, the more connected relationship, i.e. the bigger correlation, the two variables have.

```
library(tidyverse)
ggplot(cdi,aes(x=region,y=per.cap.income,fill=region)) + geom_boxplot(color="blue",alpha=0.5) + theme_cl
```




```
pairs.panels(cdi[,-c(1:3,17)],
             method="pearson",
             hist.col="#00AFBB",
             density=TRUE,
             ellipses=TRUE)
```

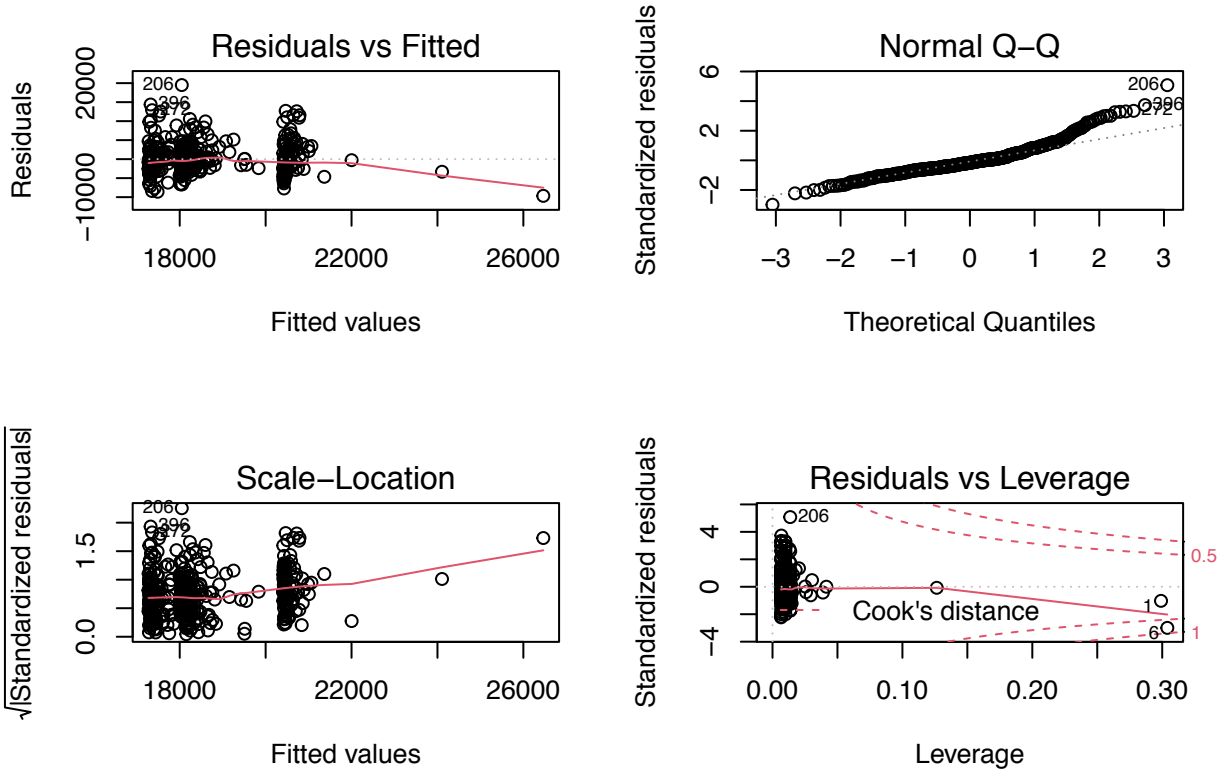


Appendix 2. Simple Linear Regression Analysis

```
# linear regression model with no interaction term on the original data
region<-as.factor(cdi$region)
cdi_fit1<-lm(per.cap.income~crimes+region,data=cdi)
summary(cdi_fit1)
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9661.0 -2260.7  -618.3  1650.0 19492.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.811e+04  3.784e+02  47.846  < 2e-16 ***
## crimes       8.915e-03  3.188e-03   2.797  0.00539 **
## regionNE     2.286e+03  5.325e+02   4.293 2.17e-05 ***
## regionS     -8.606e+02  4.868e+02  -1.768  0.07782 .
## regionW     -1.428e+02  5.796e+02  -0.246  0.80548
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3866 on 435 degrees of freedom
## Multiple R-squared:  0.1011, Adjusted R-squared:  0.09288
## F-statistic: 12.24 on 4 and 435 DF,  p-value: 1.946e-09
```

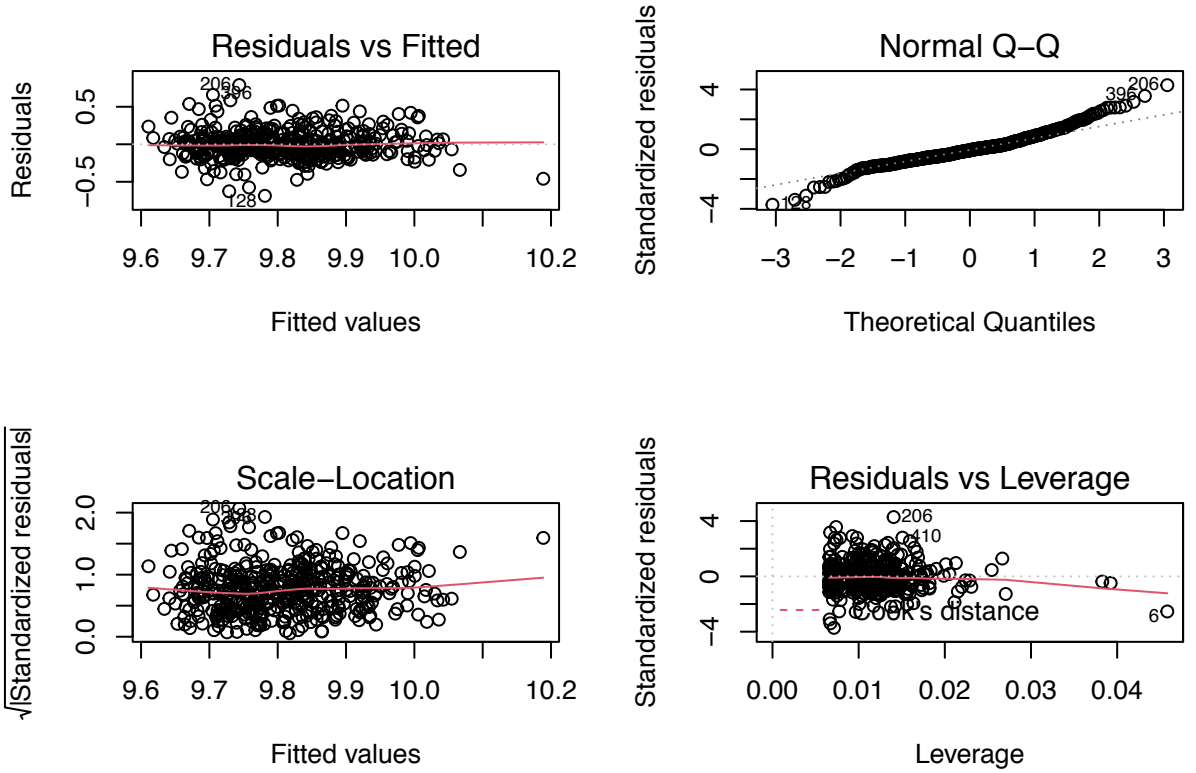
```
par(mfrow=c(2,2))
plot(cdi_fit1)
```




```
# linear regression model with no interaction term on the transformed data
region<-as.factor(cdi$region)
cdi_fit2<-lm(log(per.cap.income)~log(crimes)+region,data=cdi)
summary(cdi_fit2)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(crimes) + region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68757 -0.10557 -0.01422  0.08905  0.78946
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.188431   0.079812 115.125  < 2e-16 ***
## log(crimes)  0.066695   0.008421   7.920 2.00e-14 ***
## regionNE     0.104458   0.025531   4.091 5.11e-05 ***
## regionS     -0.086983   0.023618  -3.683  0.00026 ***
## regionW     -0.055280   0.028167  -1.963  0.05033 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1854 on 435 degrees of freedom
## Multiple R-squared:  0.2032, Adjusted R-squared:  0.1959
## F-statistic: 27.74 on 4 and 435 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(cdi_fit2)
```

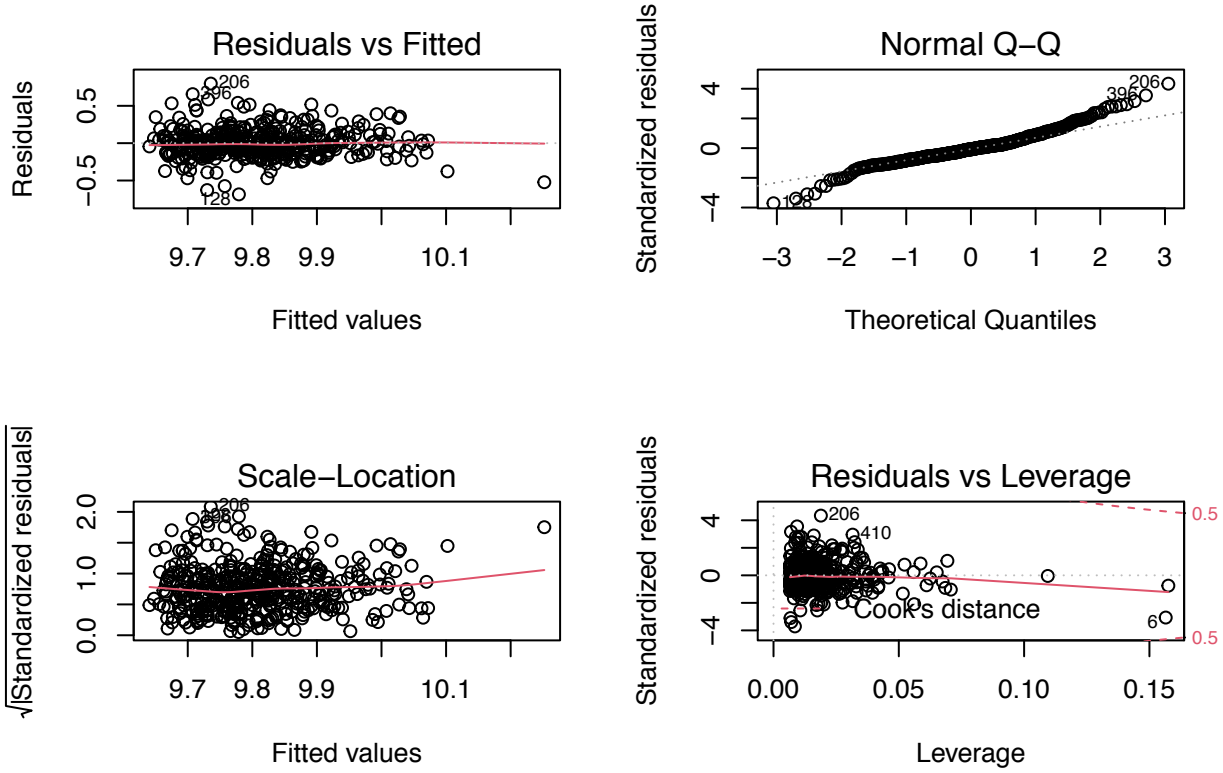



```
# linear regression model with the interaction term on the transformed data
region<-as.factor(cdi$region)
cdi_fit3<-lm(log(per.cap.income)~log(crimes)+region+log(crimes)*region,data=cdi)
summary(cdi_fit3)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(crimes) + region + log(crimes) *
##     region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68552 -0.10418 -0.01444  0.08302  0.79755
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.33677    0.14579  64.044  < 2e-16 ***
## log(crimes)            0.05064    0.01566   3.233  0.00132 **
## regionNE              -0.18407    0.21515  -0.856  0.39272
## regionS               -0.19717    0.21211  -0.930  0.35312
## regionW               -0.31439    0.24465  -1.285  0.19947
## log(crimes):regionNE   0.03122    0.02311   1.351  0.17749
## log(crimes):regionS    0.01211    0.02228   0.544  0.58696
## log(crimes):regionW    0.02727    0.02523   1.081  0.28028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1855 on 432 degrees of freedom
## Multiple R-squared:  0.2073, Adjusted R-squared:  0.1945
## F-statistic: 16.14 on 7 and 432 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(cdi_fit3)
```




```
# compare whether is the need to interaction term on the transformed data
anova(cdi_fit2,cdi_fit3)
```

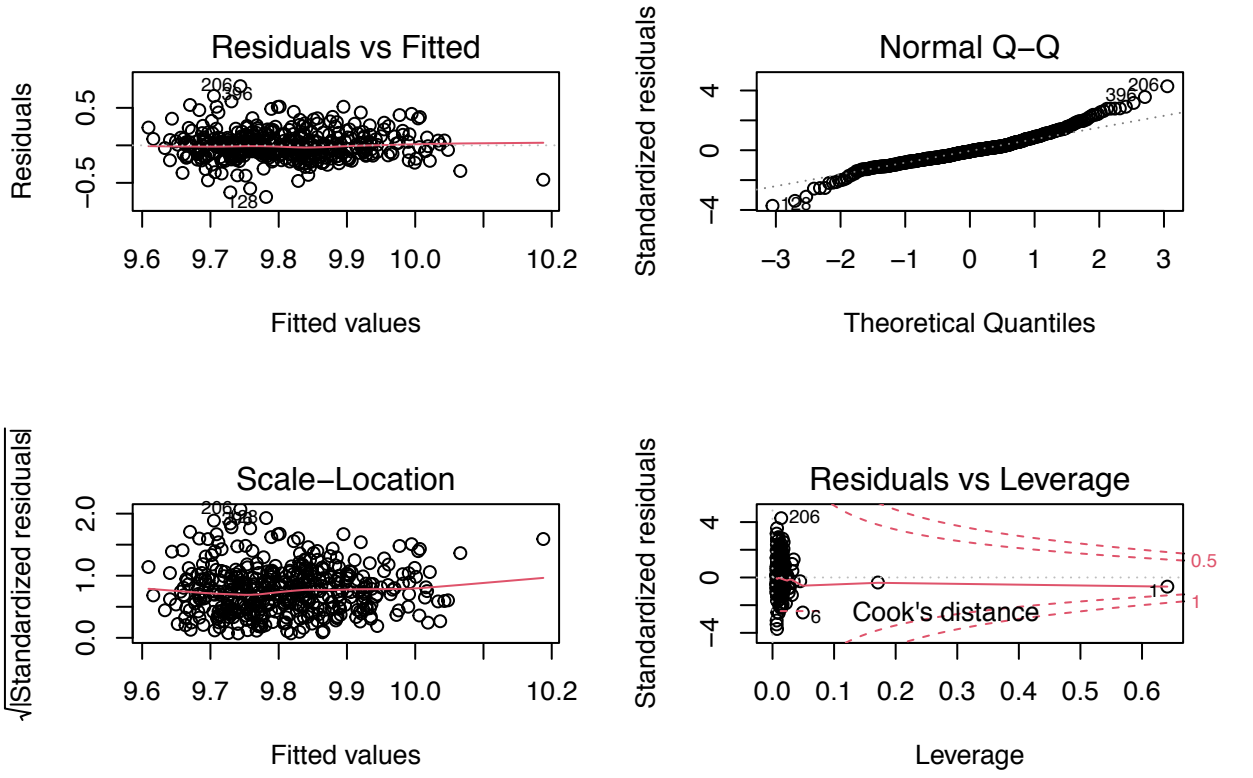
```
## Analysis of Variance Table
##
## Model 1: log(per.cap.income) ~ log(crimes) + region
## Model 2: log(per.cap.income) ~ log(crimes) + region + log(crimes) * region
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    435 14.949
## 2    432 14.872  3  0.076778 0.7434 0.5266
```

```
# model with per-capita crime measure with no interaction term on transformed data
region<-as.factor(cdi$region)
cdi_fit4<-lm(log(per.cap.income)~(log(crimes)/pop)+region,data=cdi)
summary(cdi_fit4)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ (log(crimes)/pop) + region,
##     data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.68786 -0.10559 -0.01417  0.08906  0.78930
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.183e+00  9.995e-02  91.869  < 2e-16 ***
## log(crimes)       6.739e-02  1.102e-02   6.118 2.12e-09 ***
## regionNE          1.045e-01  2.557e-02   4.088 5.19e-05 ***
## regionS          -8.731e-02  2.388e-02  -3.656 0.000287 ***
## regionW          -5.529e-02  2.820e-02  -1.961 0.050557 .
## log(crimes):pop -1.466e-10  1.503e-09  -0.098 0.922344
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1856 on 434 degrees of freedom
## Multiple R-squared:  0.2033, Adjusted R-squared:  0.1941
## F-statistic: 22.14 on 5 and 434 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(cdi_fit4)
```

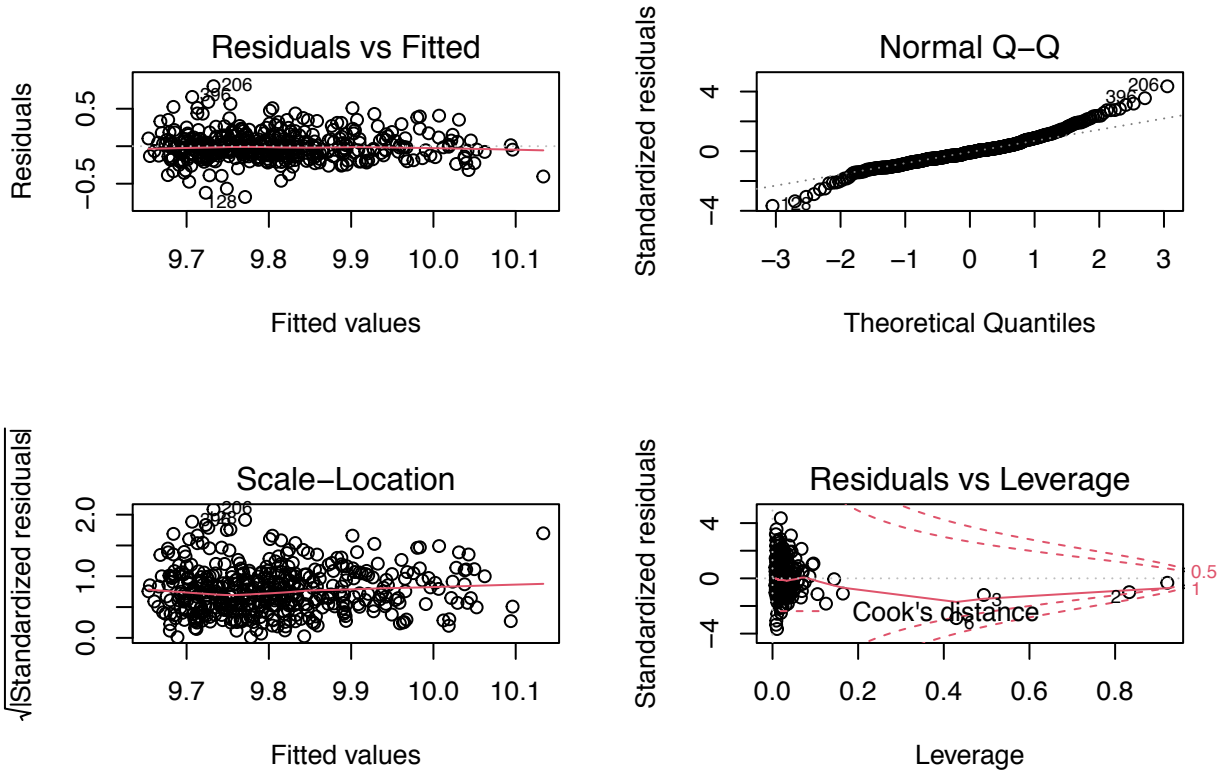



```
# model with per-capita crime measure with an interaction term on transformed data
region<-as.factor(cdi$region)
cdi_fit5<-lm(log(per.cap.income)~(log(crimes)/pop)+region+(log(crimes)/pop)*region,data=cdi)
summary(cdi_fit5)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ (log(crimes)/pop) + region +
##     (log(crimes)/pop) * region, data = cdi)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.67743 -0.10319 -0.01547  0.08266  0.80046
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 9.384e+00  1.824e-01  51.451   <2e-16 ***
## log(crimes)                 4.499e-02  2.045e-02   2.200   0.0284 *
## regionNE                   -5.089e-01  3.317e-01  -1.534   0.1257
## regionS                    -6.599e-02  3.020e-01  -0.218   0.8272
## regionW                    -4.766e-01  3.169e-01  -1.504   0.1333
## log(crimes):pop             1.421e-09  3.306e-09   0.430   0.6675
## log(crimes):regionNE        7.111e-02  3.864e-02   1.840   0.0664 .
## log(crimes):regionS        -2.627e-03  3.343e-02  -0.079   0.9374
## log(crimes):regionW         4.551e-02  3.391e-02   1.342   0.1803
## log(crimes):pop:regionNE -1.115e-08  8.623e-09  -1.293   0.1965
## log(crimes):pop:regionS     4.083e-09  6.599e-09   0.619   0.5364
## log(crimes):pop:regionW    -2.805e-09  3.878e-09  -0.723   0.4698
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1857 on 428 degrees of freedom
## Multiple R-squared:  0.213,  Adjusted R-squared:  0.1928
## F-statistic: 10.53 on 11 and 428 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(cdi_fit5)
```

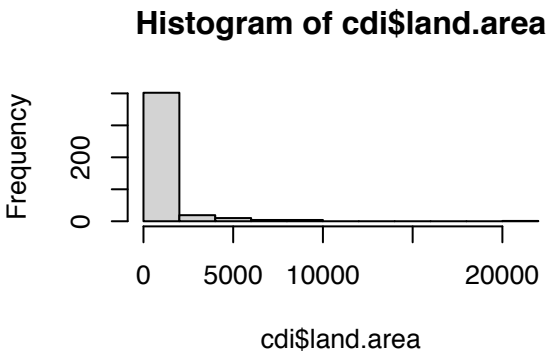


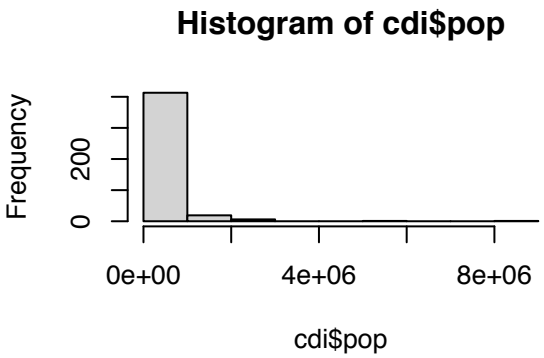

```
# compare whether is the need to interaction term on the transformed data
anova(cdi_fit4,cdi_fit5)
```

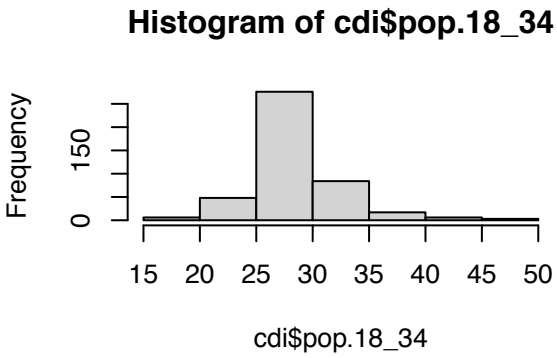
```
## Analysis of Variance Table
##
## Model 1: log(per.cap.income) ~ (log(crimes)/pop) + region
## Model 2: log(per.cap.income) ~ (log(crimes)/pop) + region + (log(crimes)/pop) *
##     region
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    434 14.949
## 2    428 14.766  6    0.1825 0.8816 0.5082
```

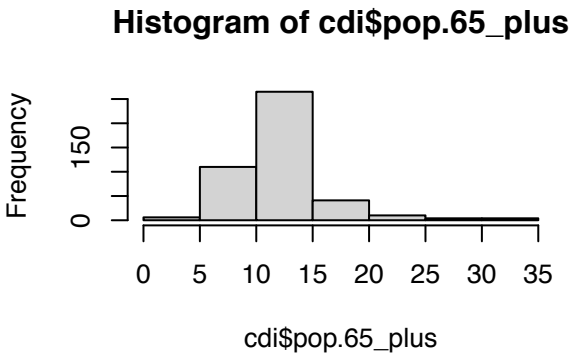
Appendix 3. Multiple Regression Analysis

```
par(mfrow=c(2,2))
hist(cdi$land.area)
hist(cdi$pop)
hist(cdi$pop.18_34)
hist(cdi$pop.65_plus)
```

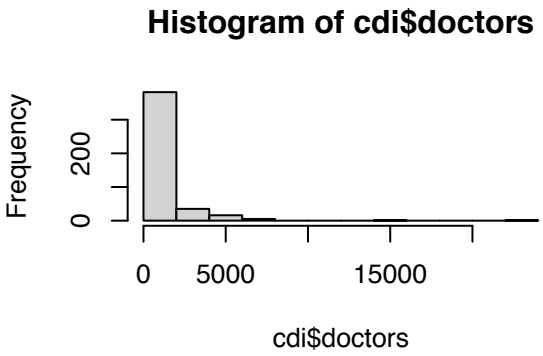


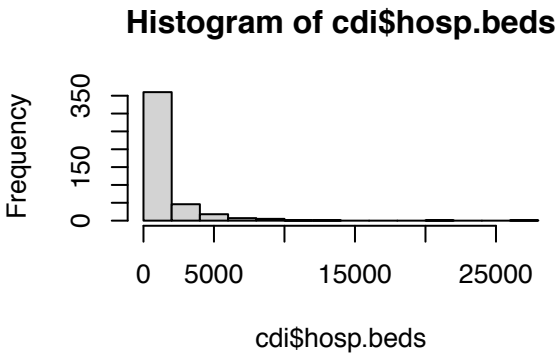


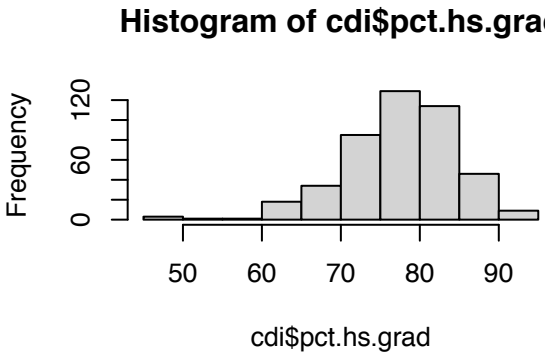


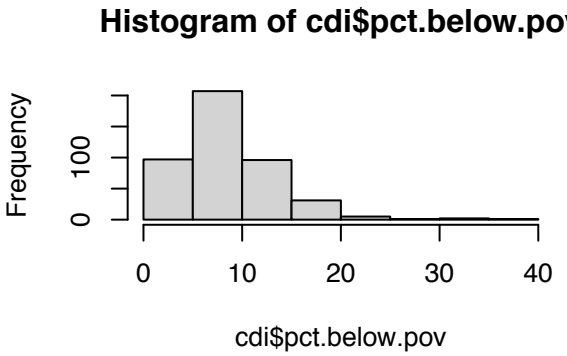



```
hist(cdi$doctors)
hist(cdi$hosp.beds)
hist(cdi$pct.hs.grad)
hist(cdi$pct.below.pov)
```

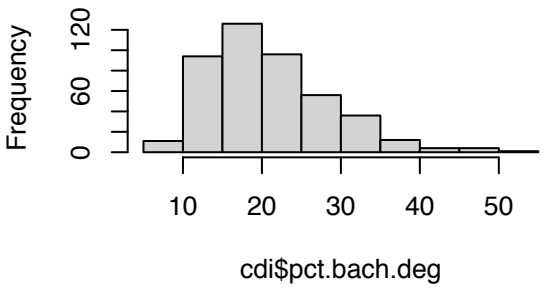


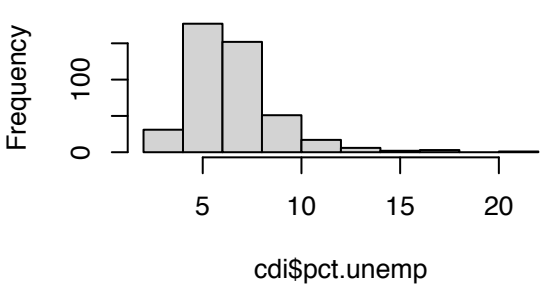


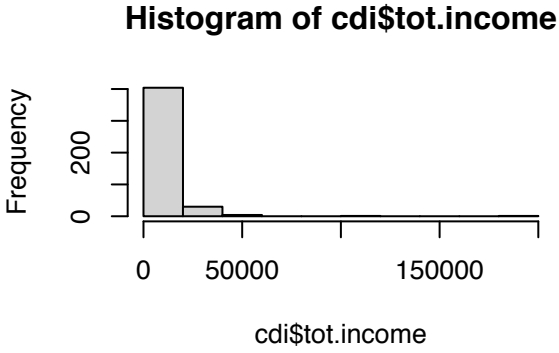





```
hist(cdi$pct.bach.deg)
hist(cdi$pct.unemp)
hist(cdi$tot.income)
```







The first step of doing the multiple regression model is to identify the distribution of variables and decide whether

to make data transformation on them.
For variable "Land Area": right-skewed, need to do log transformation
For variable "Total Population": right-skewed, need to do log transformation
For variable "Percent of Population Aged 18-34" and "Percent of Population Aged 65 or Older", the distribution looks normal and there is no need on data transformation
For variable "Number of Active Physician": right-skewed, need to do log transformation
For variable "Number of Hopsital Beds": right-skewed, need to do log transformation
For variable "Percent High School Graduates": left-skewed, need to do squared transformation
For variable "Percent Below Poverty Level": right-skewed, need to do log transformation
For variable "Percent Bachelor's Degrees": right-skewed, need to do log transformation
For variable "Percent Unemployment": right-skewed, need to do log transformation
For variable "Total Income": right-skewed, need to do log transformation

Since we know that the response variable per.cap.income is mathematically calculated by tot.income divided by pop, then we remove two variables: tot.income and pop when fitting the multiple regression model.

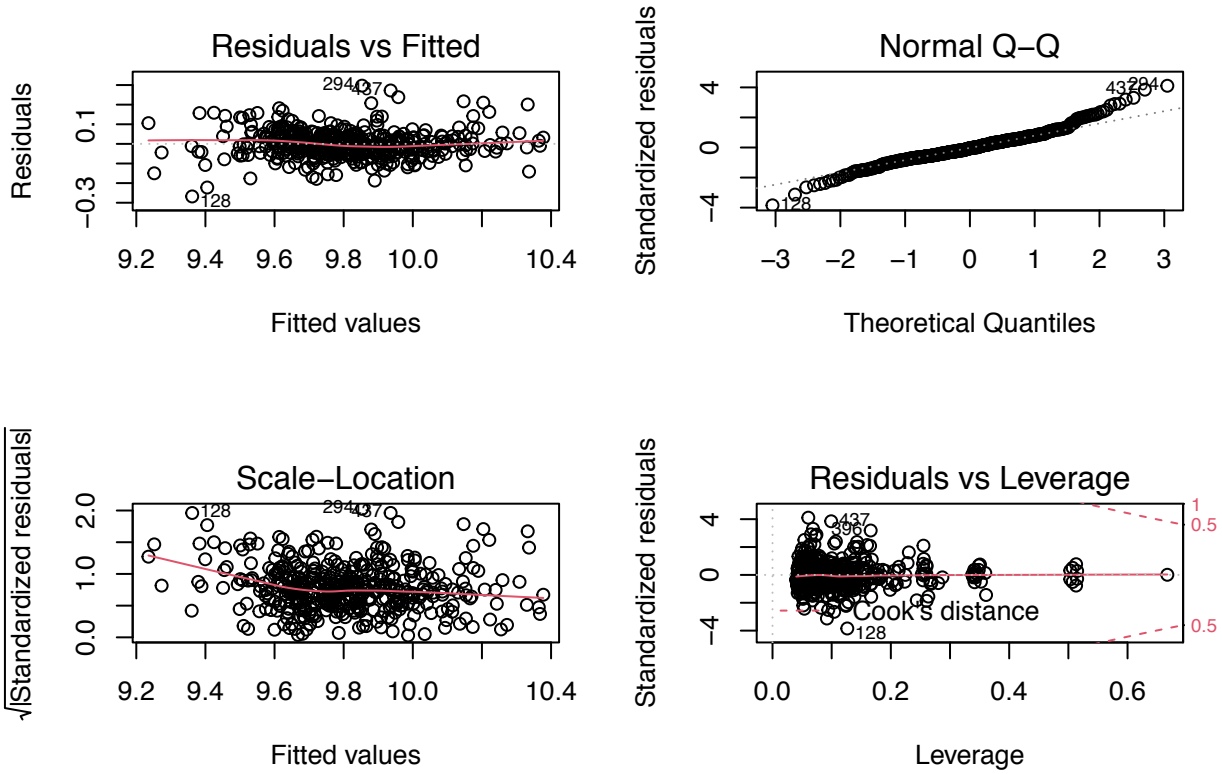
With all the above presented data transformation on each numerical variable, we fit the multiple regression model as the following equation:

```
mulreg_fit1<-lm(log(per.cap.income)~log(land.area)+pop.18_34+pop.65_plus
                +log(doctors)+log(hosp.beds)+log(crimes)/pop
                +pct.hs.grad**2+log(pct.below.pov)+log(pct.bach.deg)
                +log(pct.unemp)+region+state,data=cdi)
summary(mulreg_fit1)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + pop.18_34 +
##     pop.65_plus + log(doctors) + log(hosp.beds) + log(crimes)/pop +
##     pct.hs.grad^2 + log(pct.below.pov) + log(pct.bach.deg) +
##     log(pct.unemp) + region + state, data = cdi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.267982 -0.040635 -0.002401  0.037440  0.296768
##
## Coefficients: (3 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.590e+00  1.386e-01  69.179  < 2e-16 ***
## log(land.area)      -3.753e-02  6.160e-03  -6.093 2.72e-09 ***
## pop.18_34           -1.246e-02  1.435e-03  -8.681  < 2e-16 ***
## pop.65_plus          3.047e-03  1.629e-03   1.871 0.062173 .
## log(doctors)         3.523e-02  1.377e-02   2.559 0.010890 *
## log(hosp.beds)       1.550e-02  1.368e-02   1.132 0.258145
## log(crimes)          5.139e-03  9.446e-03   0.544 0.586713
## pct.hs.grad         -2.581e-04  1.220e-03  -0.212 0.832540
## log(pct.below.pov) -1.720e-01  1.464e-02 -11.744  < 2e-16 ***
## log(pct.bach.deg)    2.586e-01  2.671e-02   9.683  < 2e-16 ***
## log(pct.unemp)       1.216e-03  2.430e-02   0.050 0.960107
## regionNE            -1.932e-02  7.917e-02  -0.244 0.807301
## regionS             -1.114e-03  3.806e-02  -0.029 0.976668
## regionW             -2.353e-02  3.484e-02  -0.675 0.499889
## stateAR             -5.267e-02  6.001e-02  -0.878 0.380680
## stateAZ             -5.151e-02  4.301e-02  -1.198 0.231726
## stateCA              1.169e-01  2.829e-02   4.132 4.42e-05 ***
## stateCO              3.227e-02  3.483e-02   0.926 0.354913
```

```
## stateCT             9.813e-02  8.002e-02   1.226 0.220816
## stateDC             9.361e-02  8.260e-02   1.133 0.257817
## stateDE             3.862e-02  9.232e-02   0.418 0.675955
## stateFL            -4.595e-02  3.487e-02  -1.318 0.188386
## stateGA             2.271e-02  3.881e-02   0.585 0.558802
## stateHI             5.659e-02  5.270e-02   1.074 0.283615
## stateID            -2.250e-03  7.871e-02  -0.029 0.977204
## stateIL             4.525e-02  3.000e-02   1.509 0.132242
## stateIN            -1.393e-02  3.055e-02  -0.456 0.648713
## stateKS            -1.785e-02  4.397e-02  -0.406 0.685057
## stateKY            -8.963e-03  5.233e-02  -0.171 0.864092
## stateLA            -5.920e-03  3.826e-02  -0.155 0.877108
## stateMA             6.424e-02  7.947e-02   0.808 0.419398
## stateMD             1.655e-02  4.014e-02   0.412 0.680441
## stateME             2.486e-02  8.257e-02   0.301 0.763495
## stateMI             6.470e-02  3.046e-02   2.124 0.034332 *
## stateMN            -2.710e-02  3.656e-02  -0.741 0.458991
## stateMO            -4.211e-03  3.533e-02  -0.119 0.905176
## stateMS            -7.129e-02  5.209e-02  -1.369 0.171920
## stateMT             3.451e-02  7.919e-02   0.436 0.663226
## stateNC            -5.224e-03  3.463e-02  -0.151 0.880157
## stateND            -4.535e-02  8.010e-02  -0.566 0.571620
## stateNE            -9.244e-02  5.136e-02  -1.800 0.072680 .
## stateNH             2.922e-02  8.439e-02   0.346 0.729359
## stateNJ             1.077e-01  7.821e-02   1.377 0.169320
## stateNM            -8.753e-02  5.896e-02  -1.485 0.138502
## stateNV             2.186e-01  5.968e-02   3.662 0.000286 ***
## stateNY             3.113e-02  7.716e-02   0.403 0.686888
## stateOH             1.143e-02  2.760e-02   0.414 0.679083
## stateOK            -6.836e-02  4.769e-02  -1.433 0.152570
## stateOR            -4.067e-02  3.874e-02  -1.050 0.294504
## statePA            -2.086e-03  7.694e-02  -0.027 0.978388
## stateRI            -5.157e-02  8.768e-02  -0.588 0.556787
## stateSC            -1.054e-02  3.667e-02  -0.287 0.774058
## stateSD            -1.675e-02  7.970e-02  -0.210 0.833674
## stateTN            -1.491e-02  3.894e-02  -0.383 0.702046
## stateTX            -1.753e-02  3.209e-02  -0.546 0.585233
## stateUT            -2.543e-01  4.510e-02  -5.639 3.33e-08 ***
## stateVA             1.287e-02  4.242e-02   0.303 0.761798
## stateVT                    NA         NA      NA       NA
## stateWA                    NA         NA      NA       NA
## stateWI                    NA         NA      NA       NA
## stateWV            -7.792e-03  8.016e-02  -0.097 0.922616
## log(crimes):pop    -6.685e-11  6.447e-10  -0.104 0.917469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07455 on 381 degrees of freedom
## Multiple R-squared:  0.8872, Adjusted R-squared:   0.87
## F-statistic: 51.64 on 58 and 381 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(mulreg_fit1)
```



From the summary of table, we can obserce that the adjusted R-Squared performs pretty well, with the value of 0.87 meaning that 87% of its variability can be explained. Also, the residual standard error is 0.07455, which is pretty small.

From the previous understanding, we perform subsets regression analysis to observe the suitable multiple regression model for the dataset:

```
library(leaps)
library(car)
library(MASS)
library(glmnet)
```

```
# variable selection - subsets regression
mulreg_fit2<-regsubsets(log(per.cap.income)~log(land.area)+pop.18_34
                    +pop.65_plus+log(doctors)+log(hosp.beds)
                    +(log(crimes)/pop)+pct.hs.grad**2+log(pct.below.pov)
                    +log(pct.bach.deg)+log(pct.unemp)+region
                    +state,data=cdi,really.big=T)
```

```
## Reordering variables and trying again:
summary(mulreg_fit2)
```

```
## Subset selection object
## Call: regsubsets.formula(log(per.cap.income) ~ log(land.area) + pop.18_34 +
##     pop.65_plus + log(doctors) + log(hosp.beds) + (log(crimes)/pop) +
##     pct.hs.grad^2 + log(pct.below.pov) + log(pct.bach.deg) +
##     log(pct.unemp) + region + state, data = cdi, really.big = T)
## 61 Variables  (and intercept)
##                     Forced in Forced out
## log(land.area)         FALSE      FALSE
```

```
## pop.18_34              FALSE      FALSE
## pop.65_plus            FALSE      FALSE
## log(doctors)           FALSE      FALSE
## log(hosp.beds)         FALSE      FALSE
## log(crimes)            FALSE      FALSE
## pct.hs.grad            FALSE      FALSE
## log(pct.below.pov)     FALSE      FALSE
## log(pct.bach.deg)      FALSE      FALSE
## log(pct.unemp)         FALSE      FALSE
## regionNE               FALSE      FALSE
## regionS                FALSE      FALSE
## regionW                FALSE      FALSE
## stateAR                FALSE      FALSE
## stateAZ                FALSE      FALSE
## stateCA                FALSE      FALSE
## stateCO                FALSE      FALSE
## stateCT                FALSE      FALSE
## stateDC                FALSE      FALSE
## stateDE                FALSE      FALSE
## stateFL                FALSE      FALSE
## stateGA                FALSE      FALSE
## stateHI                FALSE      FALSE
## stateID                FALSE      FALSE
## stateIL                FALSE      FALSE
## stateIN                FALSE      FALSE
## stateKS                FALSE      FALSE
## stateKY                FALSE      FALSE
## stateLA                FALSE      FALSE
## stateMA                FALSE      FALSE
## stateMD                FALSE      FALSE
## stateME                FALSE      FALSE
## stateMI                FALSE      FALSE
## stateMN                FALSE      FALSE
## stateMO                FALSE      FALSE
## stateMS                FALSE      FALSE
## stateMT                FALSE      FALSE
## stateNC                FALSE      FALSE
## stateND                FALSE      FALSE
## stateNE                FALSE      FALSE
## stateNH                FALSE      FALSE
## stateNJ                FALSE      FALSE
## stateNM                FALSE      FALSE
## stateNV                FALSE      FALSE
## stateNY                FALSE      FALSE
## stateOH                FALSE      FALSE
## stateOK                FALSE      FALSE
## stateOR                FALSE      FALSE
## statePA                FALSE      FALSE
## stateRI                FALSE      FALSE
## stateSC                FALSE      FALSE
## stateSD                FALSE      FALSE
## stateTN                FALSE      FALSE
## stateTX                FALSE      FALSE
## stateUT                FALSE      FALSE
```

```
## stateVA                FALSE      FALSE
## stateWV                FALSE      FALSE
## log(crimes):pop        FALSE      FALSE
## stateVT                FALSE      FALSE
## stateWA                FALSE      FALSE
## stateWI                FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: exhaustive
##          log(land.area) pop.18_34 pop.65_plus log(doctors) log(hosp.beds)
## 1  ( 1 ) " "            " "       " "          " "           " "
## 2  ( 1 ) " "            " "       " "          "*"           " "
## 3  ( 1 ) "*"            " "       " "          "*"           " "
## 4  ( 1 ) " "            "*"       " "          "*"           " "
## 5  ( 1 ) "*"            "*"       " "          "*"           " "
## 6  ( 1 ) "*"            "*"       " "          "*"           " "
## 7  ( 1 ) "*"            "*"       " "          "*"           " "
## 8  ( 1 ) "*"            "*"       " "          "*"           " "
## 9  ( 1 ) "*"            "*"       " "          "*"           " "
##          log(crimes) pct.hs.grad log(pct.below.pov) log(pct.bach.deg)
## 1  ( 1 ) " "         " "         "*"                " "
## 2  ( 1 ) " "         " "         "*"                " "
## 3  ( 1 ) " "         " "         "*"                " "
## 4  ( 1 ) " "         " "         "*"                "*"
## 5  ( 1 ) " "         " "         "*"                "*"
## 6  ( 1 ) " "         " "         "*"                "*"
## 7  ( 1 ) " "         " "         "*"                "*"
## 8  ( 1 ) " "         " "         "*"                "*"
## 9  ( 1 ) " "         " "         "*"                "*"
##          log(pct.unemp) regionNE regionS regionW stateAR stateAZ stateCA
## 1  ( 1 ) " "            " "      " "     " "     " "     " "     " "
## 2  ( 1 ) " "            " "      " "     " "     " "     " "     " "
## 3  ( 1 ) " "            " "      " "     " "     " "     " "     " "
## 4  ( 1 ) " "            " "      " "     " "     " "     " "     " "
## 5  ( 1 ) " "            " "      " "     " "     " "     " "     " "
## 6  ( 1 ) " "            " "      " "     " "     " "     " "     " "
## 7  ( 1 ) " "            " "      " "     " "     " "     " "     "*"
## 8  ( 1 ) " "            " "      " "     " "     " "     " "     "*"
## 9  ( 1 ) " "            " "      " "     " "     " "     " "     "*"
##          stateCO stateCT stateDC stateDE stateFL stateGA stateHI stateID
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 7  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 8  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 9  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
##          stateIL stateIN stateKS stateKY stateLA stateMA stateMD stateME
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
```

```
## 6  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 7  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 8  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 9  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
##          stateMI stateMN stateMO stateMS stateMT stateNC stateND stateNE
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 7  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 8  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 9  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
##          stateNH stateNJ stateNM stateNV stateNY stateOH stateOK stateOR
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 7  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 8  ( 1 ) " "     "*"     " "     " "     " "     " "     " "     " "
## 9  ( 1 ) " "     "*"     " "     "*"     " "     " "     " "     " "
##          statePA stateRI stateSC stateSD stateTN stateTX stateUT stateVA
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     " "     " "     " "     " "     "*"     " "
## 7  ( 1 ) " "     " "     " "     " "     " "     " "     "*"     " "
## 8  ( 1 ) " "     " "     " "     " "     " "     " "     "*"     " "
## 9  ( 1 ) " "     " "     " "     " "     " "     " "     "*"     " "
##          stateVT stateWA stateWI stateWV log(crimes):pop
## 1  ( 1 ) " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     " "     " "     " "
## 7  ( 1 ) " "     " "     " "     " "     " "
## 8  ( 1 ) " "     " "     " "     " "     " "
## 9  ( 1 ) " "     " "     " "     " "     " "
```

```
# with different criteria to select the best model
cdi_sum<-summary(mulreg_fit2)
data.frame(
  Adj_R2 = which.max(cdi_sum$adjr2),
  CP = which.min(cdi_sum$cp),
  BIC = which.min(cdi_sum$bic)
)
```

```
##   Adj_R2 CP BIC
## 1      9  9   9
```

From the above summary table, we can observe that no matter we choose to use the criteria of adjusted R-Squared, C_P, or BIC, we all should choose the model with 9 predictor variables to be our best model.

Therefore, the best model would be:

log(per.cap.income)~log(land.area)+pop.18_34+log(doctors)
+log(pct.below.pov)+log(pct.bach.deg)+stateCA+stateNJ+stateNV+stateUT

Next, we would like to perform our model selection by using stepwise regression:

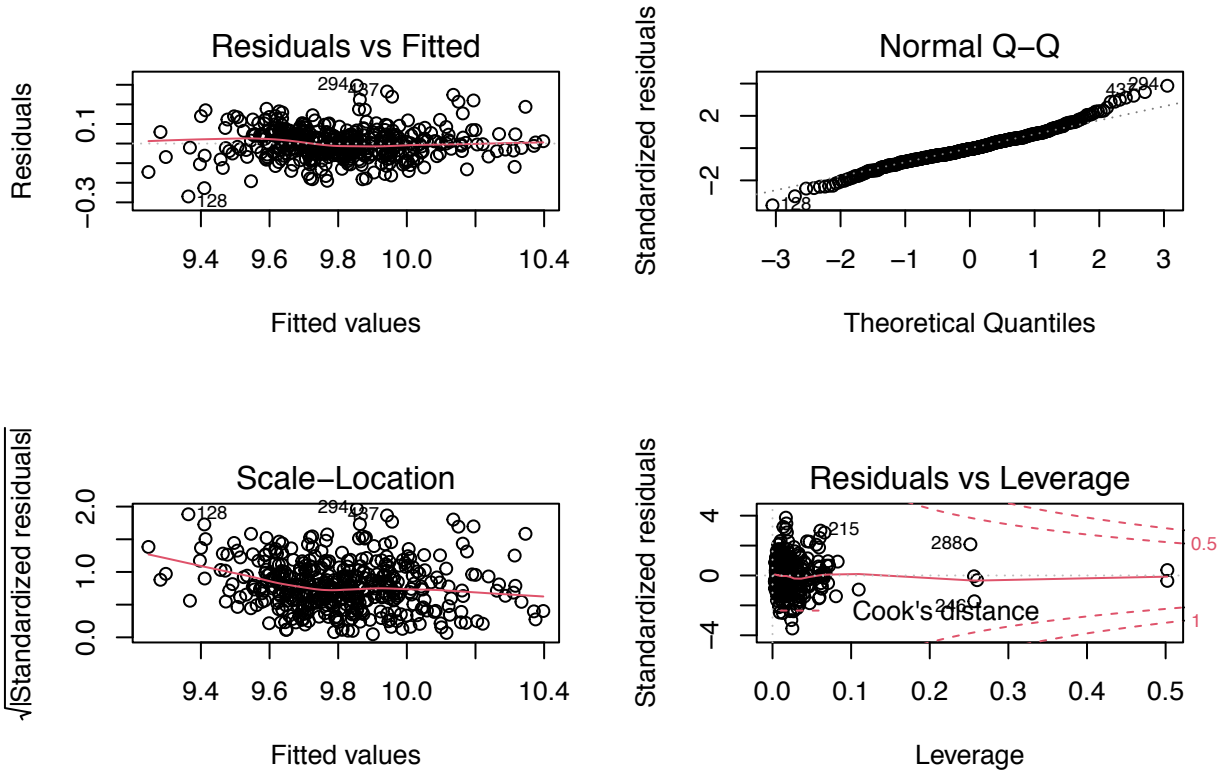
```
# variable selection - stepwise regression
income_stepmod<-stepAIC(lm(log(per.cap.income)~log(land.area)+pop.18_34
                    +pop.65_plus+log(doctors)+log(hosp.beds)
                    +log(crimes)/pop+pct.hs.grad**2+log(pct.below.pov)
                    +log(pct.bach.deg)+log(pct.unemp)+region
                    +state,data=cdi),direction="both",trace=FALSE)
summary(income_stepmod)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + pop.18_34 +
##     pop.65_plus + log(doctors) + log(pct.below.pov) + log(pct.bach.deg) +
##     state, data = cdi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.267293 -0.039773 -0.002899  0.037594  0.293303
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.646673   0.087940 109.696  < 2e-16 ***
## log(land.area)      -0.037470   0.006051  -6.193 1.51e-09 ***
## pop.18_34           -0.012447   0.001402  -8.880  < 2e-16 ***
## pop.65_plus          0.003106   0.001530   2.030  0.04304 *
## log(doctors)         0.053373   0.004180  12.769  < 2e-16 ***
## log(pct.below.pov) -0.166261   0.012019 -13.834  < 2e-16 ***
## log(pct.bach.deg)    0.244320   0.019657  12.429  < 2e-16 ***
## stateAR             -0.052290   0.059598  -0.877  0.38082
## stateAZ             -0.082270   0.045604  -1.804  0.07201 .
## stateCA              0.086381   0.031541   2.739  0.00646 **
## stateCO              0.001609   0.037688   0.043  0.96598
## stateCT              0.071273   0.040130   1.776  0.07651 .
## stateDC              0.090587   0.081144   1.116  0.26496
## stateDE              0.014682   0.060159   0.244  0.80732
## stateFL             -0.047778   0.033505  -1.426  0.15468
## stateGA              0.021186   0.038222   0.554  0.57971
## stateHI              0.024836   0.051613   0.481  0.63065
## stateID             -0.031016   0.079600  -0.390  0.69701
## stateIL              0.047108   0.034062   1.383  0.16746
## stateIN             -0.016526   0.035392  -0.467  0.64081
## stateKS             -0.023147   0.047144  -0.491  0.62372
## stateKY             -0.013390   0.051899  -0.258  0.79655
## stateLA             -0.009165   0.037895  -0.242  0.80901
## stateMA              0.037652   0.037123   1.014  0.31109
## stateMD              0.006623   0.038723   0.171  0.86428
## stateME              0.001426   0.044013   0.032  0.97417
```

```
## stateMI             0.063008   0.033814   1.863  0.06317 .
## stateMN            -0.033165   0.040313  -0.823  0.41119
## stateMO            -0.001327   0.039345  -0.034  0.97311
## stateMS            -0.078030   0.051394  -1.518  0.12977
## stateMT             0.003968   0.079819   0.050  0.96038
## stateNC            -0.010597   0.034057  -0.311  0.75585
## stateND            -0.053140   0.079898  -0.665  0.50639
## stateNE            -0.088663   0.051960  -1.706  0.08874 .
## stateNH             0.008170   0.047859   0.171  0.86454
## stateNJ             0.088985   0.035351   2.517  0.01223 *
## stateNM            -0.119155   0.060086  -1.983  0.04807 *
## stateNV             0.192661   0.061625   3.126  0.00190 **
## stateNY             0.008028   0.033150   0.242  0.80878
## stateOH             0.006344   0.032786   0.194  0.84666
## stateOK            -0.069977   0.046661  -1.500  0.13451
## stateOR            -0.073557   0.041770  -1.761  0.07903 .
## statePA            -0.026585   0.033076  -0.804  0.42203
## stateRI            -0.074835   0.052832  -1.416  0.15744
## stateSC            -0.013347   0.036033  -0.370  0.71128
## stateSD            -0.018598   0.079897  -0.233  0.81606
## stateTN            -0.017699   0.038710  -0.457  0.64778
## stateTX            -0.018621   0.031521  -0.591  0.55503
## stateUT            -0.283266   0.046930  -6.036 3.71e-09 ***
## stateVA             0.002370   0.040768   0.058  0.95368
## stateVT            -0.033379   0.080124  -0.417  0.67721
## stateWA            -0.031058   0.036956  -0.840  0.40121
## stateWI            -0.004460   0.037281  -0.120  0.90483
## stateWV            -0.012768   0.079580  -0.160  0.87262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07426 on 386 degrees of freedom
## Multiple R-squared:  0.8865, Adjusted R-squared:  0.871
## F-statistic: 56.91 on 53 and 386 DF,  p-value: < 2.2e-16
```

From the above summary table, we can observe that if we choose to use the criteria of AIC, then we should choose the model with smallest AIC values, which the best model would be:
log(per.cap.income) ~ log(land.area) + pop.18_34 + pop.65_plus + log(doctors) + log(pct.below.pov) + log(pct.bach.deg) + state;

```
#compare candidate model1
cdi["stateCA"]<-ifelse(cdi$state=="CA",1,0)
cdi["stateNJ"]<-ifelse(cdi$state=="NJ",1,0)
cdi["stateNV"]<-ifelse(cdi$state=="NV",1,0)
cdi["stateUT"]<-ifelse(cdi$state=="UT",1,0)
can_model1<-lm(log(per.cap.income)~log(land.area)+pop.18_34
               +log(doctors)+log(pct.below.pov)+log(pct.bach.deg)
               +stateCA+stateNJ+stateNV+stateUT,data=cdi)
par(mfrow=c(2,2))
plot(can_model1)
```

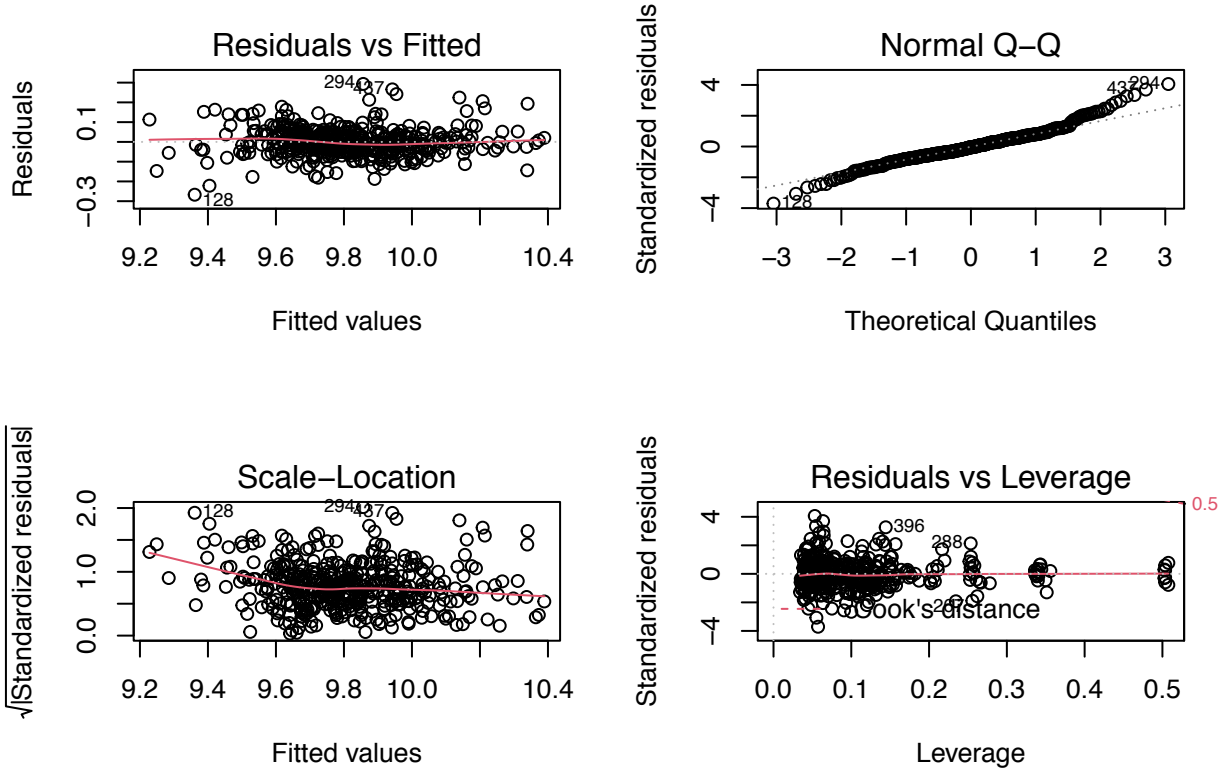



```
summary(can_model1)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + pop.18_34 +
##     log(doctors) + log(pct.below.pov) + log(pct.bach.deg) + stateCA +
##     stateNJ + stateNV + stateUT, data = cdi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.269816 -0.045495 -0.004464  0.044270  0.295412
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.826000   0.057502 170.880  < 2e-16 ***
## log(land.area)      -0.046640   0.004781  -9.756  < 2e-16 ***
## pop.18_34           -0.013034   0.001091 -11.951  < 2e-16 ***
## log(doctors)         0.058378   0.004041  14.446  < 2e-16 ***
## log(pct.below.pov) -0.179521   0.009267 -19.373  < 2e-16 ***
## log(pct.bach.deg)    0.218856   0.017251  12.686  < 2e-16 ***
## stateCA              0.099027   0.014814   6.685 7.21e-11 ***
## stateNJ              0.084771   0.019385   4.373 1.54e-05 ***
## stateNV              0.208425   0.056236   3.706 0.000238 ***
## stateUT             -0.283781   0.038851  -7.304 1.36e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0771 on 430 degrees of freedom
## Multiple R-squared:  0.8637, Adjusted R-squared:  0.8609
```

```
## F-statistic: 302.9 on 9 and 430 DF,  p-value: < 2.2e-16
```

```
#compare candidate model2
can_model2<-lm(log(per.cap.income) ~ log(land.area) + pop.18_34
               + pop.65_plus + log(doctors) + log(pct.below.pov)
               + log(pct.bach.deg) + state,data=cdi)
par(mfrow=c(2,2))
plot(can_model2)
```




```
summary(can_model2)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + pop.18_34 +
##     pop.65_plus + log(doctors) + log(pct.below.pov) + log(pct.bach.deg) +
##     state, data = cdi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.267293 -0.039773 -0.002899  0.037594  0.293303
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.646673   0.087940 109.696  < 2e-16 ***
## log(land.area)      -0.037470   0.006051  -6.193 1.51e-09 ***
## pop.18_34           -0.012447   0.001402  -8.880  < 2e-16 ***
## pop.65_plus          0.003106   0.001530   2.030  0.04304 *
## log(doctors)         0.053373   0.004180  12.769  < 2e-16 ***
## log(pct.below.pov) -0.166261   0.012019 -13.834  < 2e-16 ***
## log(pct.bach.deg)    0.244320   0.019657  12.429  < 2e-16 ***
```

```
## stateAR           -0.052290   0.059598  -0.877  0.38082
## stateAZ           -0.082270   0.045604  -1.804  0.07201 .
## stateCA            0.086381   0.031541   2.739  0.00646 **
## stateCO            0.001609   0.037688   0.043  0.96598
## stateCT            0.071273   0.040130   1.776  0.07651 .
## stateDC            0.090587   0.081144   1.116  0.26496
## stateDE            0.014682   0.060159   0.244  0.80732
## stateFL           -0.047778   0.033505  -1.426  0.15468
## stateGA            0.021186   0.038222   0.554  0.57971
## stateHI            0.024836   0.051613   0.481  0.63065
## stateID           -0.031016   0.079600  -0.390  0.69701
## stateIL            0.047108   0.034062   1.383  0.16746
## stateIN           -0.016526   0.035392  -0.467  0.64081
## stateKS           -0.023147   0.047144  -0.491  0.62372
## stateKY           -0.013390   0.051899  -0.258  0.79655
## stateLA           -0.009165   0.037895  -0.242  0.80901
## stateMA            0.037652   0.037123   1.014  0.31109
## stateMD            0.006623   0.038723   0.171  0.86428
## stateME            0.001426   0.044013   0.032  0.97417
## stateMI            0.063008   0.033814   1.863  0.06317 .
## stateMN           -0.033165   0.040313  -0.823  0.41119
## stateMO           -0.001327   0.039345  -0.034  0.97311
## stateMS           -0.078030   0.051394  -1.518  0.12977
## stateMT            0.003968   0.079819   0.050  0.96038
## stateNC           -0.010597   0.034057  -0.311  0.75585
## stateND           -0.053140   0.079898  -0.665  0.50639
## stateNE           -0.088663   0.051960  -1.706  0.08874 .
## stateNH            0.008170   0.047859   0.171  0.86454
## stateNJ            0.088985   0.035351   2.517  0.01223 *
## stateNM           -0.119155   0.060086  -1.983  0.04807 *
## stateNV            0.192661   0.061625   3.126  0.00190 **
## stateNY            0.008028   0.033150   0.242  0.80878
## stateOH            0.006344   0.032786   0.194  0.84666
## stateOK           -0.069977   0.046661  -1.500  0.13451
## stateOR           -0.073557   0.041770  -1.761  0.07903 .
## statePA           -0.026585   0.033076  -0.804  0.42203
## stateRI           -0.074835   0.052832  -1.416  0.15744
## stateSC           -0.013347   0.036033  -0.370  0.71128
## stateSD           -0.018598   0.079897  -0.233  0.81606
## stateTN           -0.017699   0.038710  -0.457  0.64778
## stateTX           -0.018621   0.031521  -0.591  0.55503
## stateUT           -0.283266   0.046930  -6.036 3.71e-09 ***
## stateVA            0.002370   0.040768   0.058  0.95368
## stateVT           -0.033379   0.080124  -0.417  0.67721
## stateWA           -0.031058   0.036956  -0.840  0.40121
## stateWI           -0.004460   0.037281  -0.120  0.90483
## stateWV           -0.012768   0.079580  -0.160  0.87262
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07426 on 386 degrees of freedom
## Multiple R-squared:  0.8865, Adjusted R-squared:  0.871
## F-statistic: 56.91 on 53 and 386 DF,  p-value: < 2.2e-16
```

By comparing the candidate models from subsets regression and stepwise regression, we decide to choose the model selecting from subsets regression as our final model. As for the value of adjusted R-Squared and Residual Standard Error, both of two models have pretty much the same performance. However, looking into the variables, we believe that the model selecting from subsets regression has more specific preference on influential states in doing the prediction. With the consideration to explain our model to someone who is more interested in economic factors, the model with specific states can be more convincing. Therefore, our preferred final model would be:

log(per.cap.income)~log(land.area)+pop.18_34+log(doctors)
+log(pct.below.pov)+log(pct.bach.deg)+stateCA+stateNJ+stateNV+stateUT