Regression Analysis on Average Income Per Person Associated with County's Economic, Health, and Social Variables

Sifeng Li
sifengl@andrew.cmu.edu
29 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 counties 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, number of hospital beds, and crimes. A simple linear regression analysis shows that our best linear regression model should use "number of crimes" 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 and conclude that the best model would be the one including variables land area, percent of population aged 18-34, number of active physicians, percent below poverty level, percent bachelor's degrees, state with specification on California, New Jersey, Nevada, and Utah. The better model can be improved by obtaining additional data on the missing counties and doing further improvement on the models' case-wise diagnostic plots.

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, we want to further investigate how the average income per person is 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 need to consider their interaction term? Does the level of salary vary from region to region? Which expression of the variable performs better in the model, using number of crimes or number of crimes divided by population as the variable measurement?
- How to find the best model predicting per-capita income from the other variables by having both statistical and practical meaning?
- Are there any concerns on the fact that the dataset has missing states and missing counties? Why or why not?

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. Also, we checked on the missing data of each variable, and no missing data appeared. 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:

Variable Number	Variable Name	Description
1	Identification number	1–440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 or old
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI population that is unemployed
15	Per capita income	Per-capita income (i.e. average income per person) of 1990 CDI population (in dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic 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)

Table 1: Variable Definitions for CDI Data from Kutner et al. (2005)

In Table 2, Table 3, and Table 4, 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	16.40	28.10	28.57	49.70	4.19
pop.65_plus	440	3.00	11.75	12.17	33.80	3.99
doctors	440	39.00	401.00	988.00	23677.00	1789.75
hosp.beds	440	92.00	755.00	1458.60	27700.00	2289.13
crimes	440	563.00	11820.00	27112.00	688936.00	58237.51
pct.hs.grad	440	46.60	77.70	77.56	92.90	7.02
pct.bach.deg	440	8.10	19.70	21.08	52.3	7.65
pct.below.pov	440	1.40	7.90	8.70	36.30	4.66
pct.unemp	440	2.20	6.20	7.50	21.30	2.34
per.cap.income	440	8899.00	17759.00	18561.00	37541.00	4059.19
tot.income	440	1141.00	3857.00	7869.00	184230.00	12884.32

Table 2: Summary Statistics for Continuous Variables of CDI Dataset

Region	Frequency
NC	108
NE	103
S	152
W	77

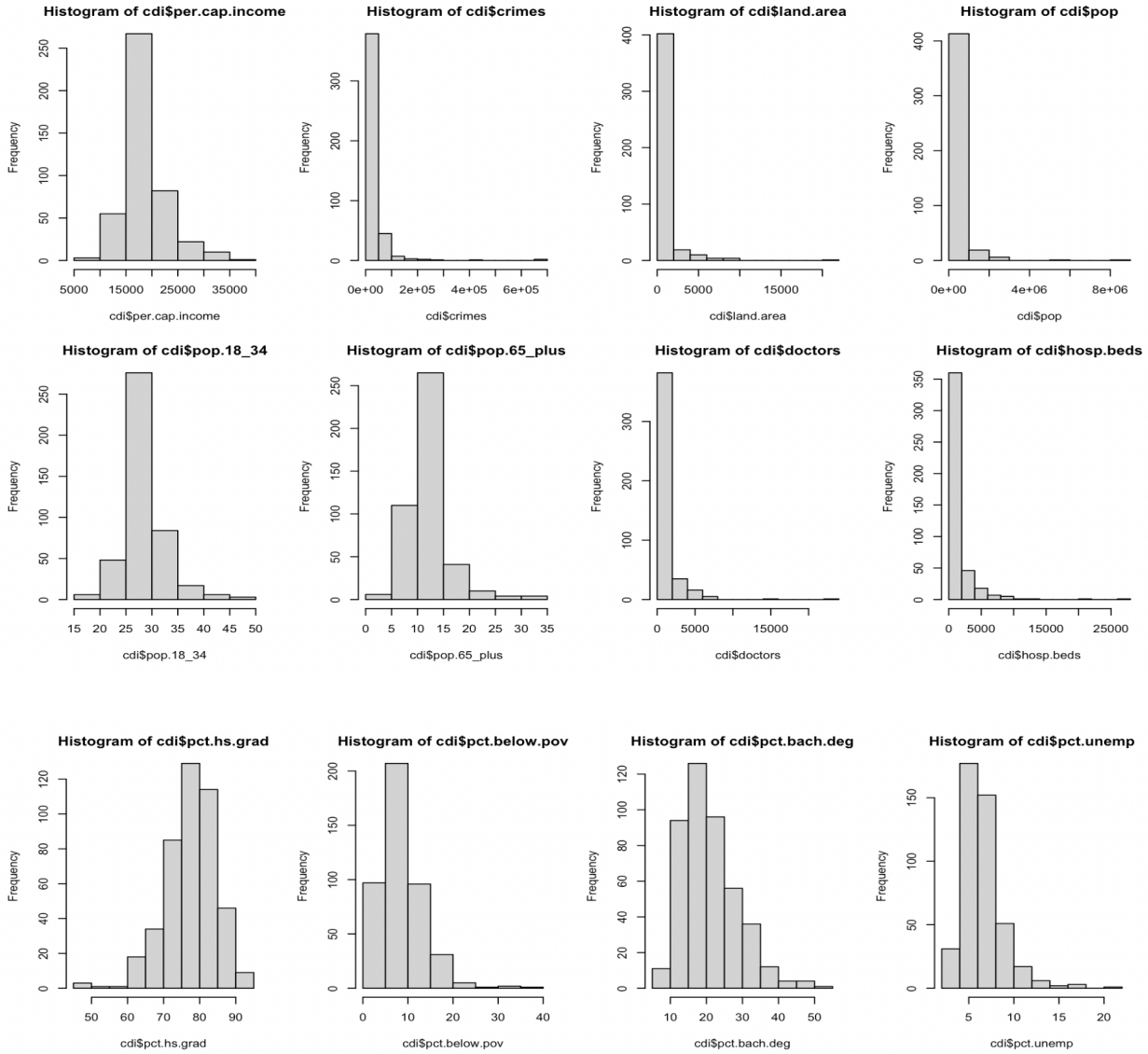
Table 3: Summary Statistics for Categorical Variables Region of CDI Dataset

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

Table 4: Summary Statistics for Categorical Variables County and State of CDI Dataset

From Table 2, we can observe that for variables land area, total population, number of active physicians/doctors, number of hospital beds, crimes, per capita income, and total income, their mean is substantially larger than their median, with the possibility of their distribution being right-skewed. There are no variables that have their mean substantially smaller than their median. From Table 3, we can observe that most counties are located in the South (region “S”) and the least are in the West (region “W”). The potential interpretation of fewer numbers in the West can be a lack of sampling in the West and the potential interpretation of larger numbers in the South can be a result of over-sampling in this region.

In Figure 1 we show all histograms of all the continuous variables. From the histograms, we find that for variables total income, total population, per capita income, crimes, land area, number of active physicians/doctors, and number of hospital beds, we need to do data transformation on each of them.



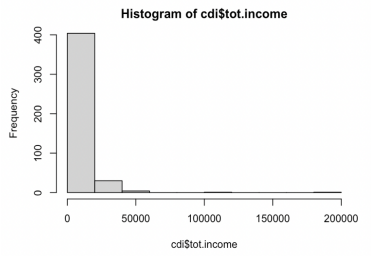


Figure 1: Histogram of all Continuous Variables for CDI Dataset

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. From the correlation plot, we can observe that Total Income and Total Population are highly correlated. Furthermore, variables Number of Doctors, Number of Hospital Beds, and Crimes are highly correlated to both variable Total Population and variable Total Income. Besides, variables Number of Doctors, Number of Hospital Beds, and Crimes have strong correlation with one another. However, the variable Per Capita Income does not have a strong correlation with any variable, the great correlation relationship that we can make is its correlation with variable percent high school graduates, percent bachelor's degrees, percent below poverty level, and percent unemployment, respectively. Not surprisingly, these four variables have a moderate correlation relationship with one another.

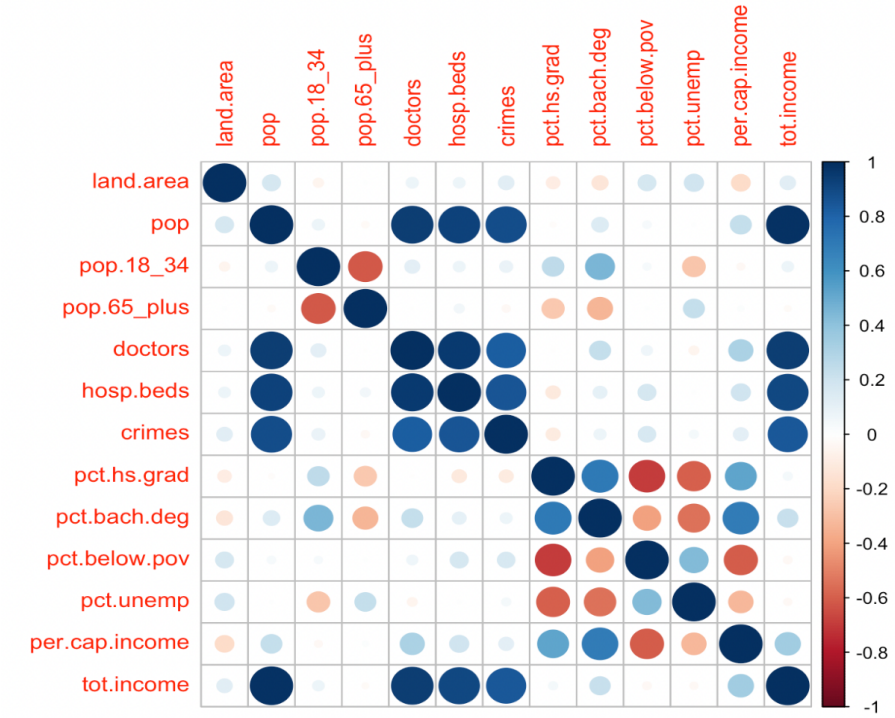


Figure 2: Correlation Plot of all Continuous Variables for CDI Dataset

In Figure 3 we show the boxplot for categorical variable region by plotting each region's boxplot and showing how per.cap.income varies across the four regions of the country. There is

a lot of overlap in the boxplots, but the Northeast and the West seem to do a little better than the North Central and the South. We can observe that for region “S,” it has the most number 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 are evenly distributed but for the region “S” and “W,” data points have more dispersions compared to data points in the region “NE.”

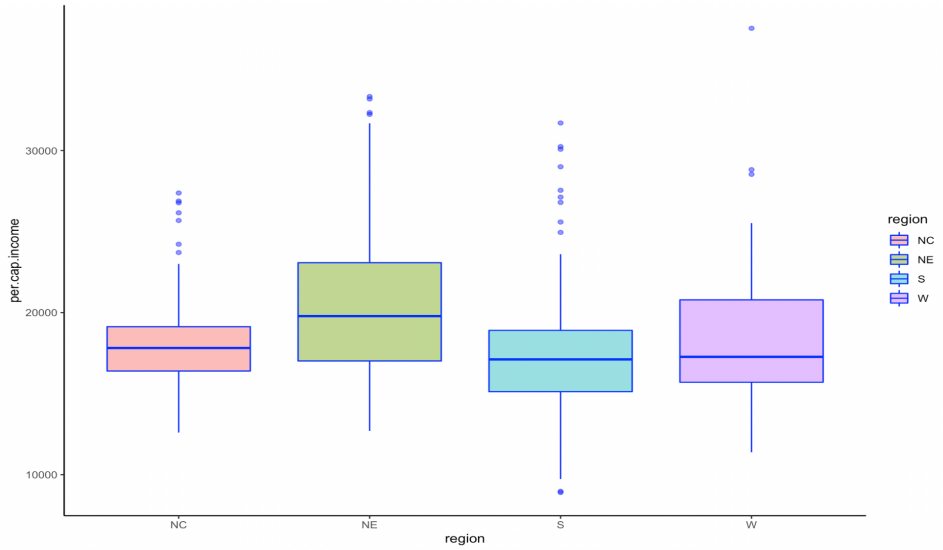


Figure 3: Boxplot of all Continuous Variable Region for CDI Dataset

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

3 Methods

We will address the methods used for each research question defined in the Introduction section.

3.1 Relationship Between Each Individual Pair of Variables

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 divided 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. With the immediate visualization of raw data, we perform histogram on transformed data again to illustrate the data distribution looks better. Detailed R analyses can be found in Appendix 1.

3.2 Examining How Variables Crimes and Region Affect Average Income per Person

For this 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 the criteria of choosing the smallest AIC/BIC value as well as putting them in real-world settings and choosing the one which has more practical meanings. Details of these analyses in R can be found in the Appendix 2.

3.3 Finding the Best Model to Predict Average Income per Person

For this 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 Appendix 3.

3.4 Researching on Whether the Missing States and Missing Counties Matter

For this part, we did some preliminary research on the missing states and missing counties as well as critical thinking based on the understanding of relevant concepts.

For this paper, all analyses were carried out in R and RStudio (RStudio Team, 2020).

4 Results

4.1 Relationship Between Each Individual Pair of Variables

After discussing from the previous exploratory analysis histograms, we consider doing log-transformation on variables total income, total population, land area, number of active physicians, and number of hospital beds. In Figure 4, we show the histograms of all continuous variables with data transformation needed. We can observe that except for variable log(pop), all other variables have nearly normal distribution compared to their previous corresponding histograms.

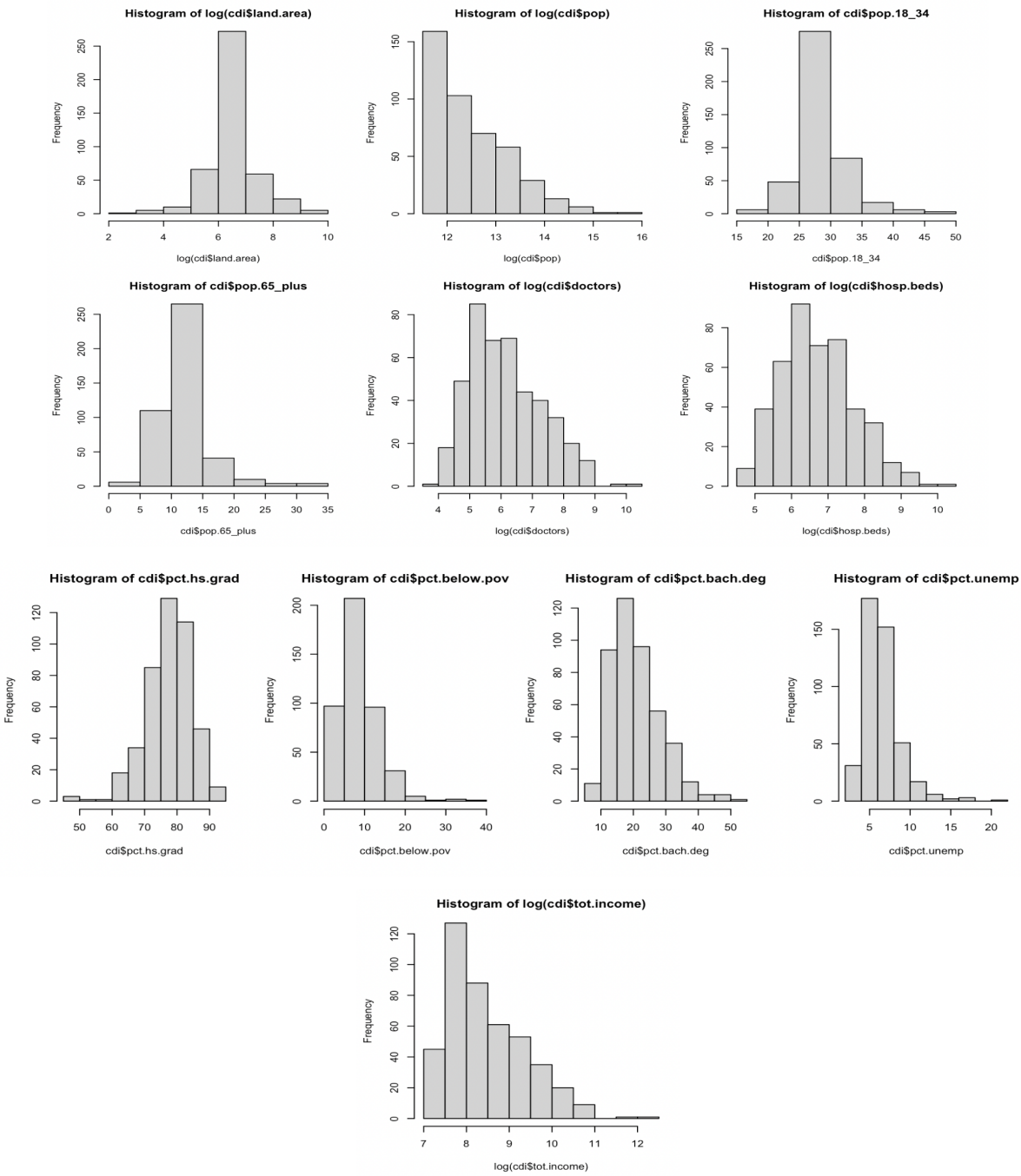


Figure 4: Histogram of all Transformed Continuous Variables for CDI Dataset

The correlation matrix on page 11 of the Technical Appendix suggests that:

- Total Income and Total Population are highly correlated. This is expected because the relationship between total.income and population can be defined as:

  per.cap.income = tot.income/pop

- Number of Doctors, Number of Hospital Beds, and Crimes are highly correlated to both variable Total Population and variable Total Income. Besides, variables Number of Doctors, Number of Hospital Beds, and Crimes have strong correlation with one another.

- Percent of High School Graduates and Percent of Bachelor's Degree have moderately high correlation. This is expected because a person can get a bachelor's degree only if he/she graduates from high school in most cases.
- Percent high school graduates and percent bachelor's degrees have a slight negative correlation with percent unemployment. Not surprising results since people who graduate from high school and university are more likely to find a job compared to people who have different educational backgrounds.
- However, the variable Per Capita Income does not have a strong correlation with any variable. This is surprising since Per Capita Income can be defined as total income divided by total population.

As for relating exploratory data analysis plots into the real-world setting, with the boxplots, we can know that people who live in the region "NE" (Northeast) tend to have a very evenly distributed per capita income; however, people who live in the region "S"(South) tend to have the most extreme and not well dispersed per capita income.

4.2 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 regression equation log(per.cap.income) ~ log(crimes) + region. 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 log transformation on data to solve the non-linearity, non-normality, and non-constant variance problems.

We considered fitting the model again after using the log-transformed variables log(per.cap.income) and log(crimes), as well as the additive and interaction terms with the region variable (details are in page 15-19 in Technical Appendix).

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 coefficient estimates presented in page 16-17 in Technical Appendix. The regression model involving log(per.cap.income), log(crimes), and region variable has the following estimated regression coefficients:

log(per.cap.income) = 0.067*log(crimes) + 0.1*regionNE - 0.09*regionS - 0.06*regionW+9.19

Page 16 and Page 17 of the technical appendix shows that all of the coefficient estimates are statistically significant with p-value less than 0.05, and adjusted R-Squared is 0.1959, meaning that 19.59% of the variance can be explained by the model. The interpretation of the regression model can be a unit percent increase in total crimes can lead to a 0.067 percent increase in average income per person.

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. We considered fitting the model again by replacing log(crimes) with per-capita crime measure, as the formula presented per-capita crime = number of crimes/population (Details can be found on page 19-22 in Technical Appendix).

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 in page 19-20 in Technical Appendix. The regression model involving log(per.cap.income), log(crimes), and region variable has the following estimated regression coefficients:

log(per.cap.income)=0.04*log(crimes/pop)+0.11*regionNE-0.07*regionS-0.02*regionW+9.94

Page 19 and Page 20 of the technical appendix shows that all of the coefficient estimators except for regionW are statistically significant with p-value less than 0.05, and adjusted R-Squared is 0.08814, meaning that only 8.81% of the variance can be explained by the model. The interpretation of the regression model can be an unit percent increase in per-capita crimes can lead to a 0.04 percent increase in average income per person.

From the above analysis, we picked one model respectively for each measurement using the number of crimes or using the “per-capita income”. Then, we wanted to identify which model performs better for the CDI dataset. As for choosing the better model to answer the question, we considered the value of comparison on AIC and BIC between these two models as well as the real-world setting situation. For comparing AIC and BIC of two candidate models, we have the following table 5 results presented:

Model	df	AIC	BIC
log(per.cap.income)~log(crimes)+region	6	-227.4746	-202.9539
log(per.cap.income)~log(crime/pop)+region	6	-172.1347	-147.6140

Table 5: Summary Table of AIC and BIC Values for Two Candidate Models

We prefer using the model with the “number of crimes” as the crime rate measure with two following reasons. First, the model with “number of crimes” as the measurement has both the smaller AIC and BIC value. Since we know that the smaller AIC and BIC value, the better the model, then the model with “number of crimes” performs better compared to the one with measurement “number of crimes/population”. Second, for the real-world setting, if we introduce the concept of “number of crimes” to social scientists or people outside the world of statistics, it would be easier for them to understand. Later, if they want to use the dataset to do further analysis with more updated data, then the data form can keep consistent without any calculation needed.

In summary, the formula of the model is: log(per.cap.income)~log(crimes)+region, with summary table of coefficient estimators presented in Table 6:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.19	0.08	115.13	0.00
log(crimes)	0.07	0.01	7.92	0.00
regionNE	0.10	0.03	4.09	0.00
regionS	-0.09	0.02	-3.68	0.00
regionW	-0.06	0.03	-1.96	0.05

Table 6: Summary Table of Coefficient Estimators for Final Model

From the above summary table, we can interpret the results as following:

- Among the entire United States, for every 1% increase in crimes, we are expecting to observe a 0.07% increase in per-capita income, on average.
- Different regions of the country have different baseline per-capita income. In the NC region, the baseline salary is exp(9.19)=$9,798.65. Similarly, in the NE region, the baseline salary is $10,829.18; in the S region, the baseline salary is $8,955.29; and in the W region, the baseline salary is $9,228.02. Therefore, the level of salary varies from region to region.

4.3 Multiple Regression Analysis

Before beginning the analysis, we need to take two variables: tot.income and pop out of consideration since per.cap.income is a deterministic function of them.

With the above presented data transformation on each numerical variable, we fit the model using all subsets regression and stepwise regression and show that we should choose the model with 9 predictor variables to be our best model (details of the analysis can be found on page 23-29 in Technical Appendix). The best model would be:

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

with specifications on state CA, stateNJ, stateNV, and stateUT.

The meaning of the model would be: if we want to predict the per-capita income, the best model would be the model include variables land area, percent of population aged 18-34, number of active physicians, percent below poverty level, percent bachelor's degrees, and state with main focus on CA, NJ, NV, and UT.

On page 31 of the Technical Appendix, we could see that the model is valid, with all four diagnostic plots performing well. On page 32 of the Technical Appendix, we could see that none of the chosen 9 variables have excessively large VIF, meaning that multicollinearity is no longer an issue. On page 32 of the Technical Appendix, we could observe that for all variables' marginal plots, the real model line matches with the blue data line pretty well. Therefore, we conclude that the chosen predictors are appropriate and the model is valid.

Summary table of coefficient estimators presented in Table 7:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.07	0.05	216.30	0
log(land.area)	-0.04	0.00	-9.61	0
pop.18_34	-0.02	0.00	-14.83	0
log(doctors)	0.06	0.00	15.37	0
pct.below.pov	-0.02	0.00	-20.34	0
pct.bach.deg	0.01	0.00	17.78	0
stateCA	0.09	0.01	6.24	0
stateNJ	0.12	0.02	6.20	0
stateNV	0.20	0.05	3.72	0
stateUT	-0.29	0.04	-7.79	0

Table 7: Summary Table of Coefficient Estimators for Best Model

From the above summary table, we can interpret the results as following:

- For every one percent increase in a country's land area, one would expect a 0.04 percent decrease in per-capita income.
- For every one percent increase in the percent of population aged 18-34, one would expect a 0.02 percent decrease in per-capita income.
- For every one percent increase in the number of active physicians, one would expect a 0.06 percent increase in per-capita income.
- For every one percent increase in the percent below poverty level, one would expect a 0.02 percent decrease in per-capita income.
- For every one percent increase in the percent bachelor's degrees, one would expect a 0.01 percent increase in per-capita income.

As for the variable state included in the best model, we find that state California (CA), state New Jersey (NJ), state Nevada (NV), and state Utah (UT) have coefficients significantly different from zero. We indicate Alabama to be the reference category state since it has the median population among all states' population (World Population Review 2021). To interpret the results as following:

- For state California, it has 0.09 percent higher per capita income than Alabama. It's reasonable since many high-tech companies are located in California and those companies produce more working opportunities with decent salaries.
- For state New Jersey, it has 0.12 percent higher per capita income than Alabama. It's reasonable since there are many prestigious institutions located in New Jersey. Also, with the fact that New Jersey is close to New York City, students who graduate from universities in New Jersey are more likely to find a job in NYC and make a great amount of earnings.
- For state Nevada, it has 0.20 percent higher per capita income than Alabama. It's reasonable since Las Vegas, located in the heart of Nevada, has countless casinos. Then, people can make significant profits through running those casino businesses thus improving the per-capita income.

- For state Utah, it has 0.29 percent lower per capita income than Alabama. It's reasonable as Utah is famous for its natural landscapes (mountains and deserts). Then, it is easy to imagine the working opportunities and population density would not be as much available as previously mentioned states.

4.4 Researching on Whether the Missing States and Missing Counties Matter

As from the dataset, we know that there are three missing states in the dataset: Alaska, Iowa, and Wyoming. Also, there are approximately 3000 counties in the United States, but only 373 of them are presented in the data set. After researching on information related to missing states and countries, we consider that missing states and counties would not be a cause for concern given that states Alaska and Wyoming are on the list of top five least populated states in the country and Iowa has less populated states compared to many other states in the country (World Population Review 2021). From this point of view, we believe that less populated states will not make a huge difference on the computation and analysis of our final chosen model.

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.

5.1 Relationship Between Each Individual Pair of Variables

In our correlation plot, we observe that total income and total population are highly correlated. Variable number of doctors, number of hospital beds, and crimes are highly correlated to both variable total population and variable total income. Besides, variables number of doctors, number of hospital beds, and crimes have strong correlation with one another. Moreover, with the boxplots, we can know that the Northeast and the West seem to do a little better than the North Central and the South.

5.2 Simple Linear Regression Analysis

With the question of investigating the relationship between the per-capita income and crime rate, we performed models including the additive term region as well as the interaction terms with region and concluded that the model without interaction term to be our final model, and the regression model should be:

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

5.3 Multiple Regression Analysis

We use multiple regression analysis including subsets regression and stepwise regression to output the best model for the dataset. We decide to choose our final model as presented:

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

with specifications on state CA, stateNJ, stateNV, and stateUT.

5.4 Researching on Whether the Missing States and Missing Counties Matter

The question of whether the missing states and missing counties matter depends on different perspectives. For the perspective of taking population influence on the best model into consideration, we grasp that it would not be the cause of concern.

5.5 Limitations and Future Works

There are some limitations that we would like to discuss regarding our data analysis. The first scope is that some models that we explored in the data analysis do not have perfectly case-wise diagnostic plots. The approximate horizontal but slight curve on the plot of Residuals vs. Fitted Value and the slight right tails of the Normal Q-Q plot show that we can still make further improvements of the model within the next step. One possible improvement can be made is to research more on secondary resources and include interaction terms involving the region categorical variable.

As previously noted, another limitation can be the dataset only includes information on the 440 most populous countries in the United States. Further research can be focused on having a larger dataset that includes information on more countries and make sure whether those counties will make significant influence on predicting per-capita income.

6 References

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.

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.

World Population Review. 2021. “US States - Ranked by Population 2021.” Retrieved Oct. 28, 2021 (https://worldpopulationreview.com/states)

Appendices: Regression Analysis on Average Income Per Person Associate with Country's Economic, Health, and Social Vairables

Sifeng Li

10/29/2021

Contents

Appendix 1. Initial Data Import & Exploration

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

```
## 'data.frame':    440 obs. of  17 variables:
##  $ id            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ county        : chr  "Los_Angeles" "Cook" "Harris" "San_Diego" ...
##  $ state         : chr  "CA" "IL" "TX" "CA" ...
##  $ land.area     : int  4060 946 1729 4205 790 71 9204 614 1945 880 ...
##  $ pop           : int  8863164 5105067 2818199 2498016 2410556 2300664 2122101 2111687 1937094 1852
##  $ pop.18_34     : num  32.1 29.2 31.3 33.5 32.6 28.3 29.2 27.4 27.1 32.6 ...
##  $ pop.65_plus   : num  9.7 12.4 7.1 10.9 9.2 12.4 12.5 12.5 13.9 8.2 ...
##  $ doctors       : int  23677 15153 7553 5905 6062 4861 4320 3823 6274 4718 ...
##  $ hosp.beds     : int  27700 21550 12449 6179 6369 8942 6104 9490 8840 6934 ...
##  $ crimes        : int  688936 436936 253526 173821 144524 680966 177593 193978 244725 214258 ...
##  $ pct.hs.grad   : num  70 73.4 74.9 81.9 81.2 63.7 81.5 70 65 77.1 ...
##  $ pct.bach.deg  : num  22.3 22.8 25.4 25.3 27.8 16.6 22.1 13.7 18.8 26.3 ...
##  $ pct.below.pov : num  11.6 11.1 12.5 8.1 5.2 19.5 8.8 16.9 14.2 10.4 ...
##  $ pct.unemp     : num  8 7.2 5.7 6.1 4.8 9.5 4.9 10 8.7 6.1 ...
##  $ per.cap.income: int  20786 21729 19517 19588 24400 16803 18042 17461 17823 21001 ...
##  $ tot.income    : int  184230 110928 55003 48931 58818 38658 38287 36872 34525 38911 ...
##  $ region        : chr  "W" "NC" "S" "W" ...
```

```
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
```

We use the above statistics to make summary table on continuous variables and categorical variables.

Moreover, we double check on missing values to make sure that there is no trouble with missing value here:

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

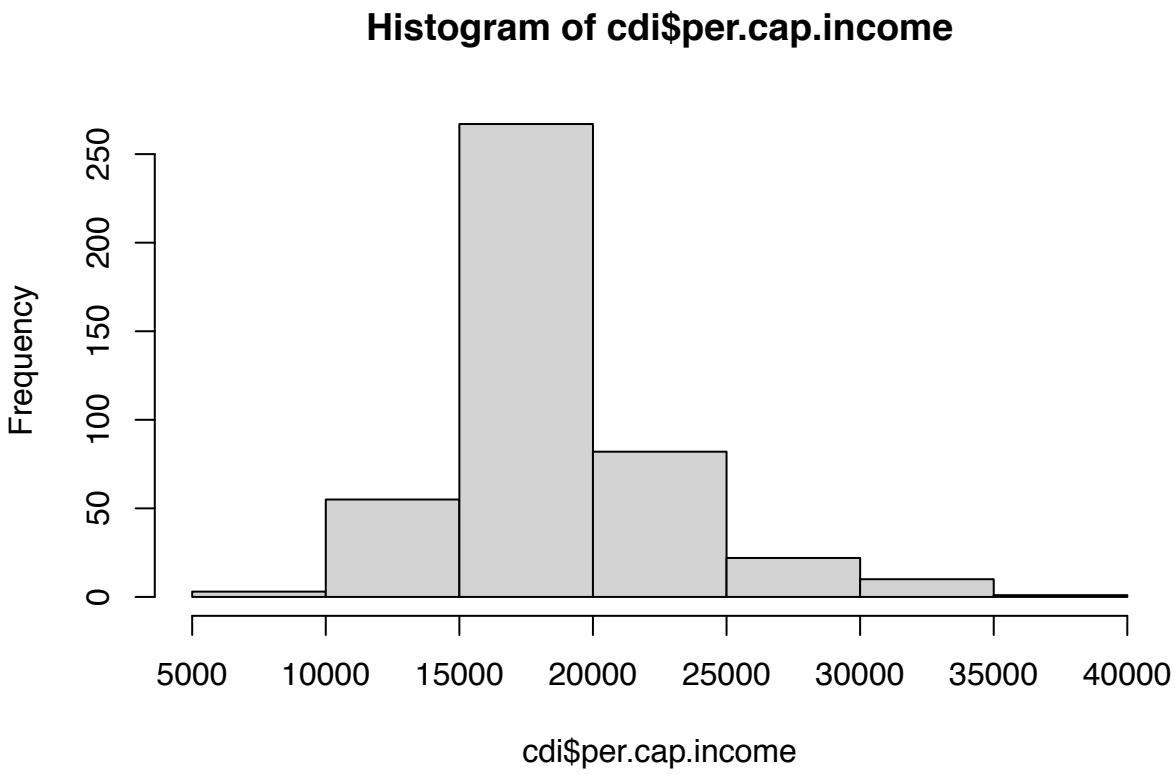
```
## integer(0)
```

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

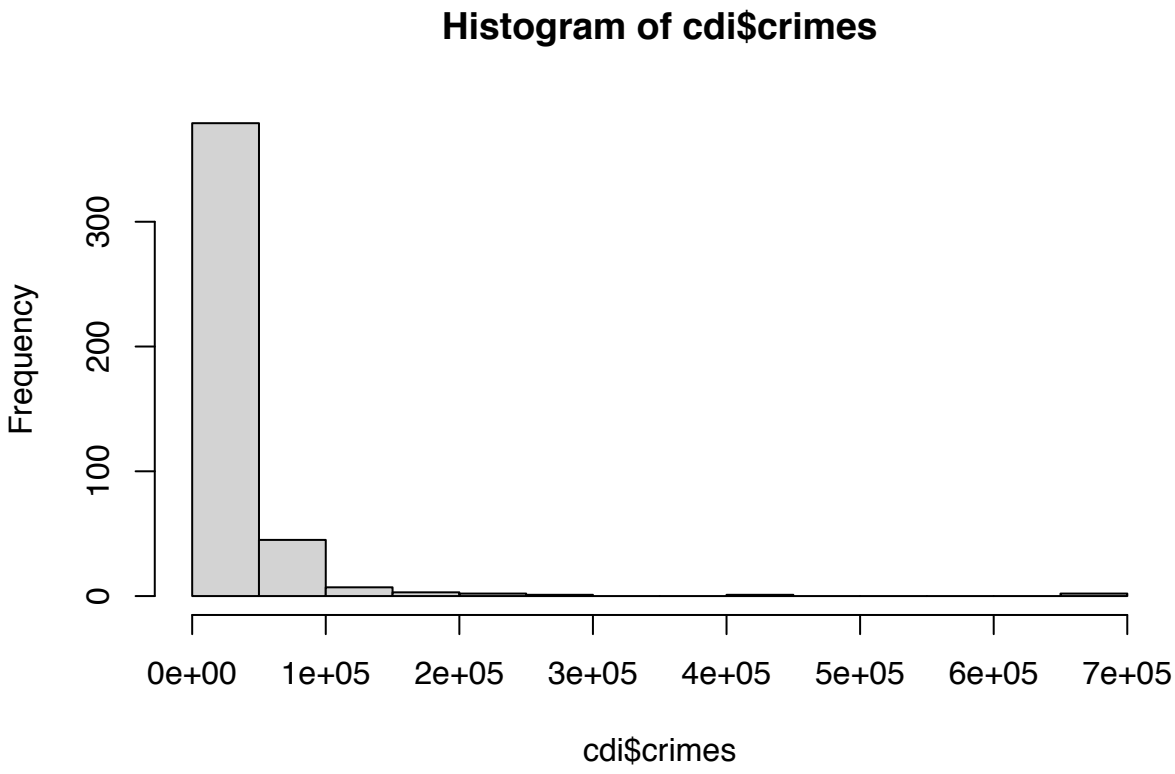
Next, we make some appropriate descriptive EDA plots as the following presented:

```
library(psych)
```

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

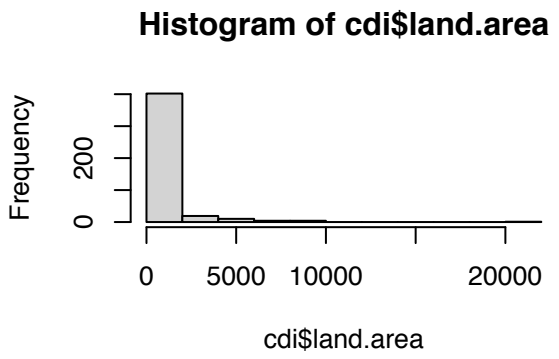


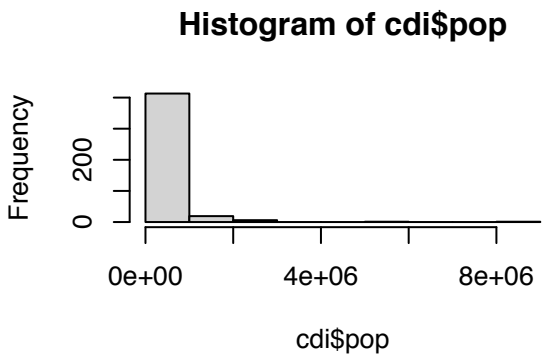

```
hist(cdi$crimes)
```

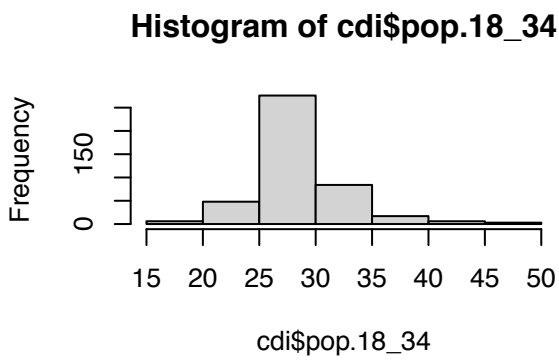


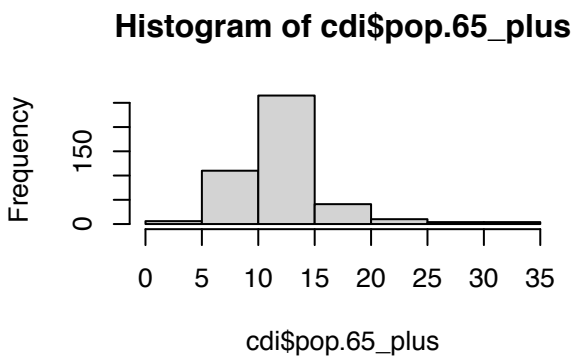

```
par(mfrow=c(2,2))
hist(cdi$land.area)
hist(cdi$pop)
```

```
hist(cdi$pop.18_34)
hist(cdi$pop.65_plus)
```

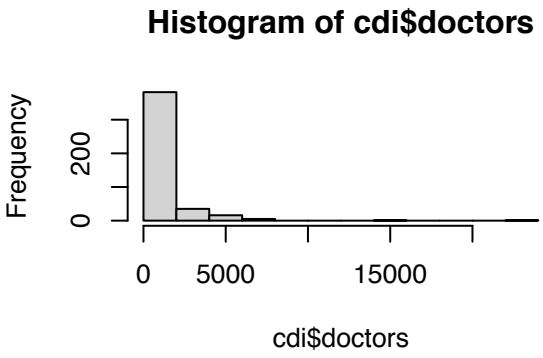


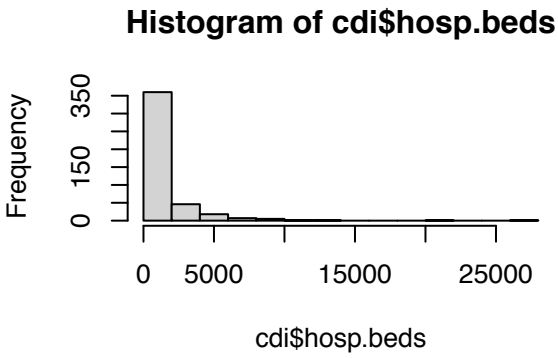


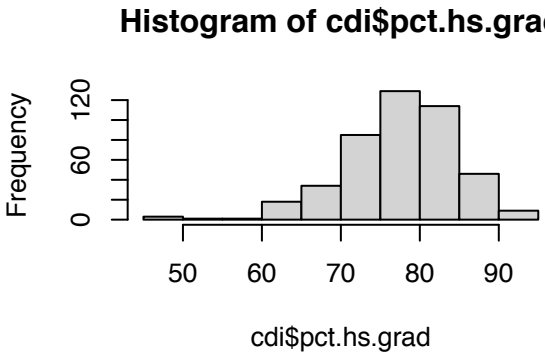


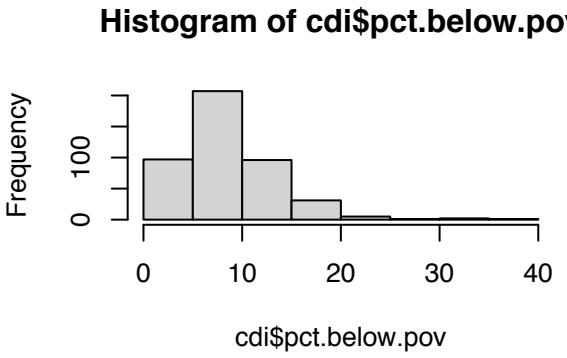



```
hist(cdi$doctors)
hist(cdi$hosp.beds)
hist(cdi$pct.hs.grad)
hist(cdi$pct.below.pov)
```

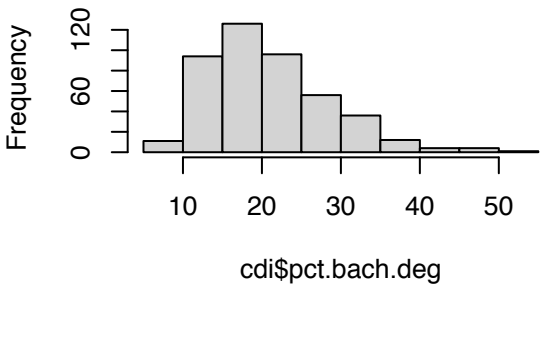


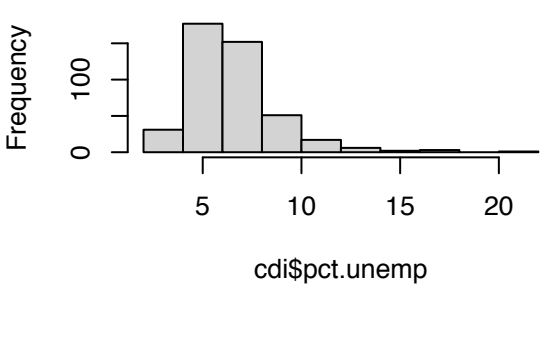


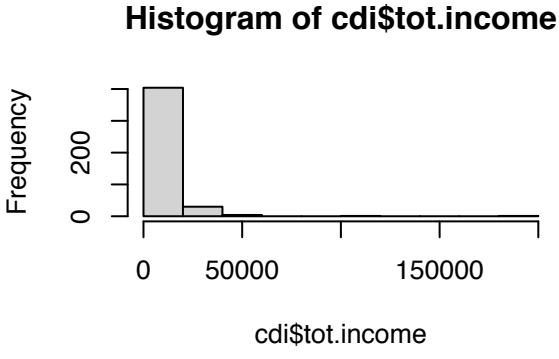





```
hist(cdi$pct.bach.deg)
hist(cdi$pct.unemp)
hist(cdi$tot.income)
```

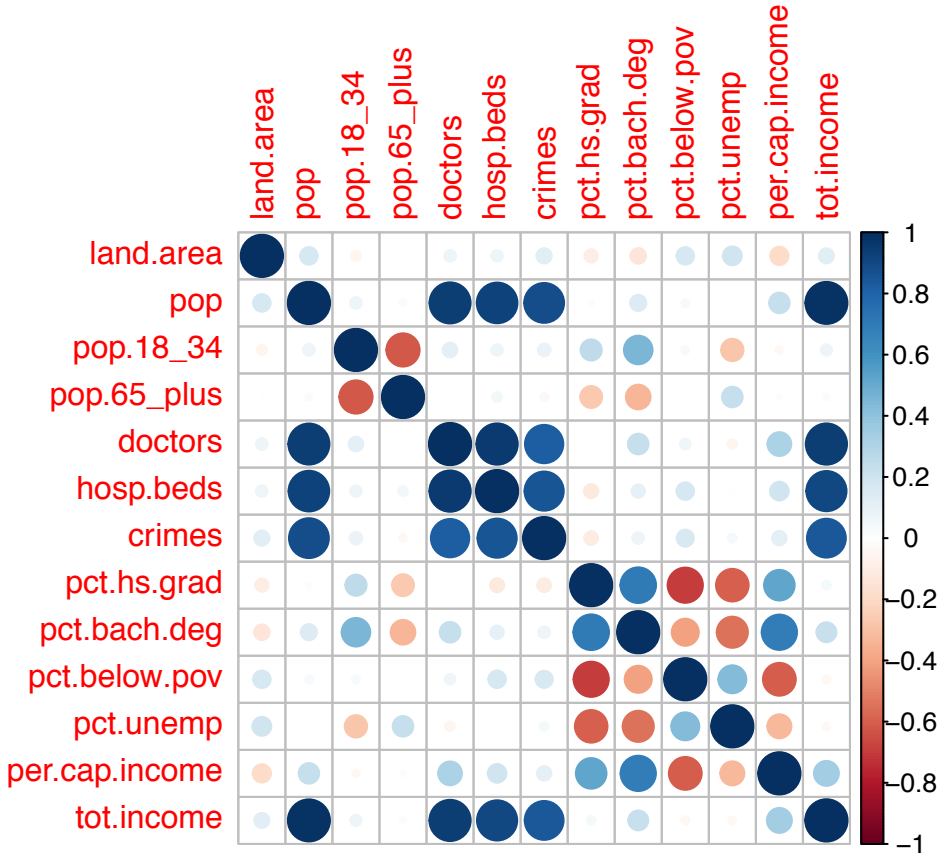






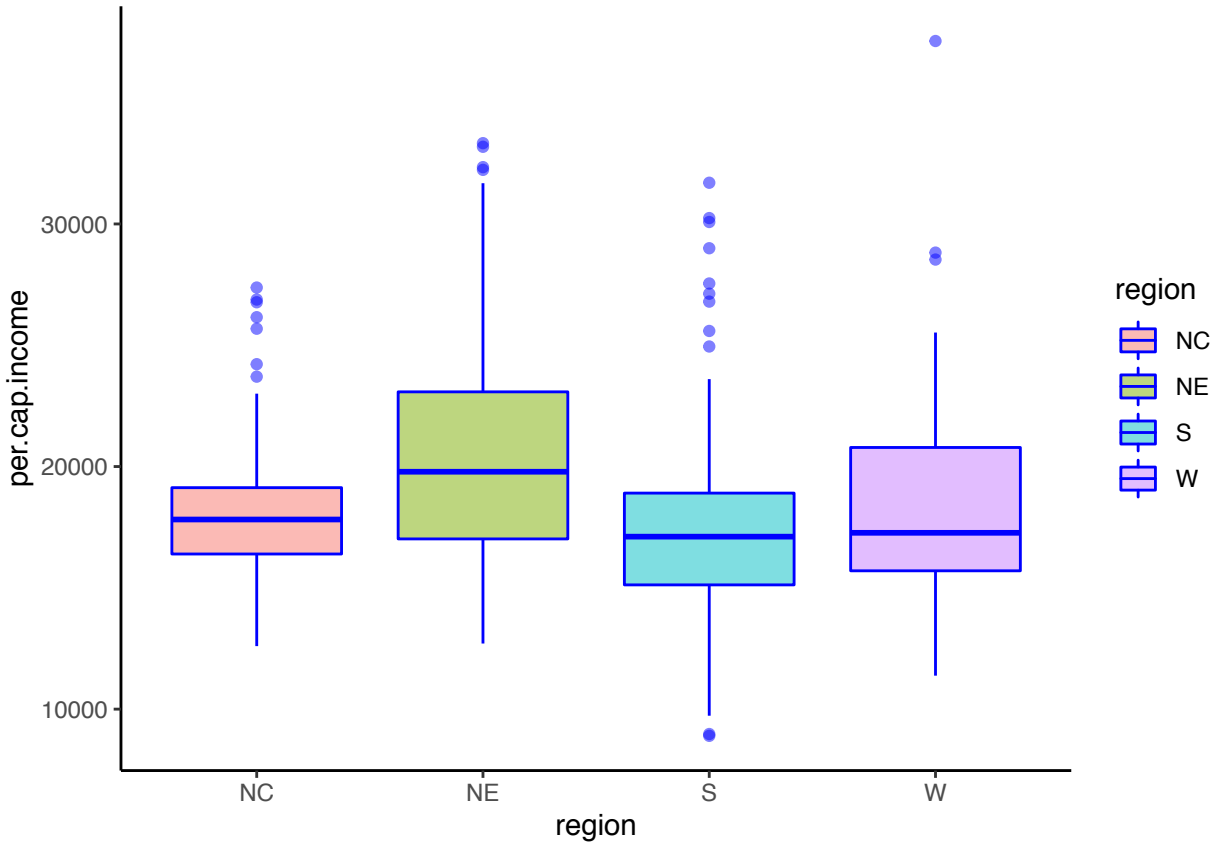
From the histograms, we find that for variables total income, pop, per.cap.income, crimes, land.area, doctors, and hosp.beds, we need to do data transformation on each of them.

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



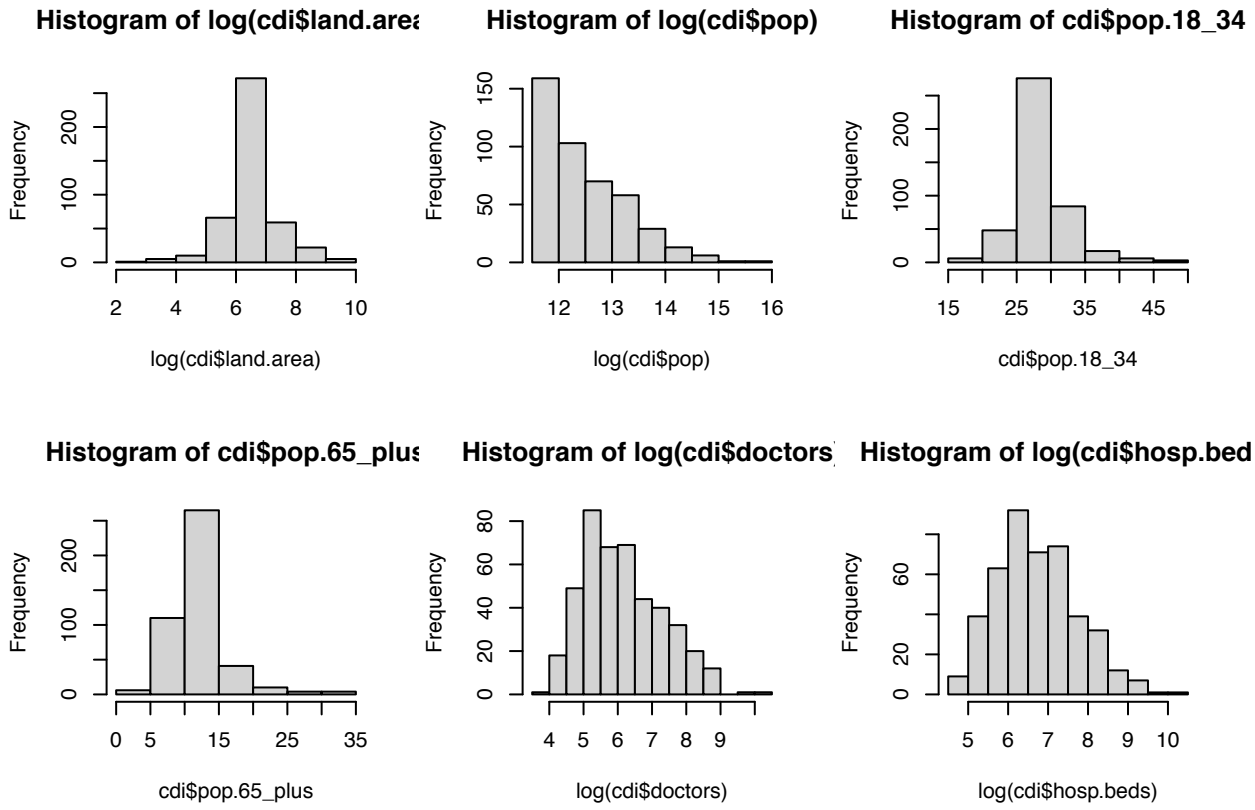
From the correlation plot, we can notice that the darker colors and bigger size the circle is, the more connected relationship, i.e. the bigger correlation, the two variables have. We can notice that the darker colors and bigger size the circle is, the more connected the relationship, i.e. the bigger correlation, that two variables have. From the correlation plot, we can observe that tot.income and pop are highly correlated. Furthermore, Variables doctors, hosp.beds, and crimes are highly correlated to both variable pop and variable tot.income. Besides, variables doctors, hosp.beds, and crimes have strong correlation with one another. However, variable per.cap.income does not have a strong correlation with any variable, the great correlation relationship that we can make is its correlation with variable pct.hs.grad, pct.bach.deg, pct.below.pov, and pct.unemp, respectively. Not surprisingly, these four variables have a moderate correlation relationship with one another.

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

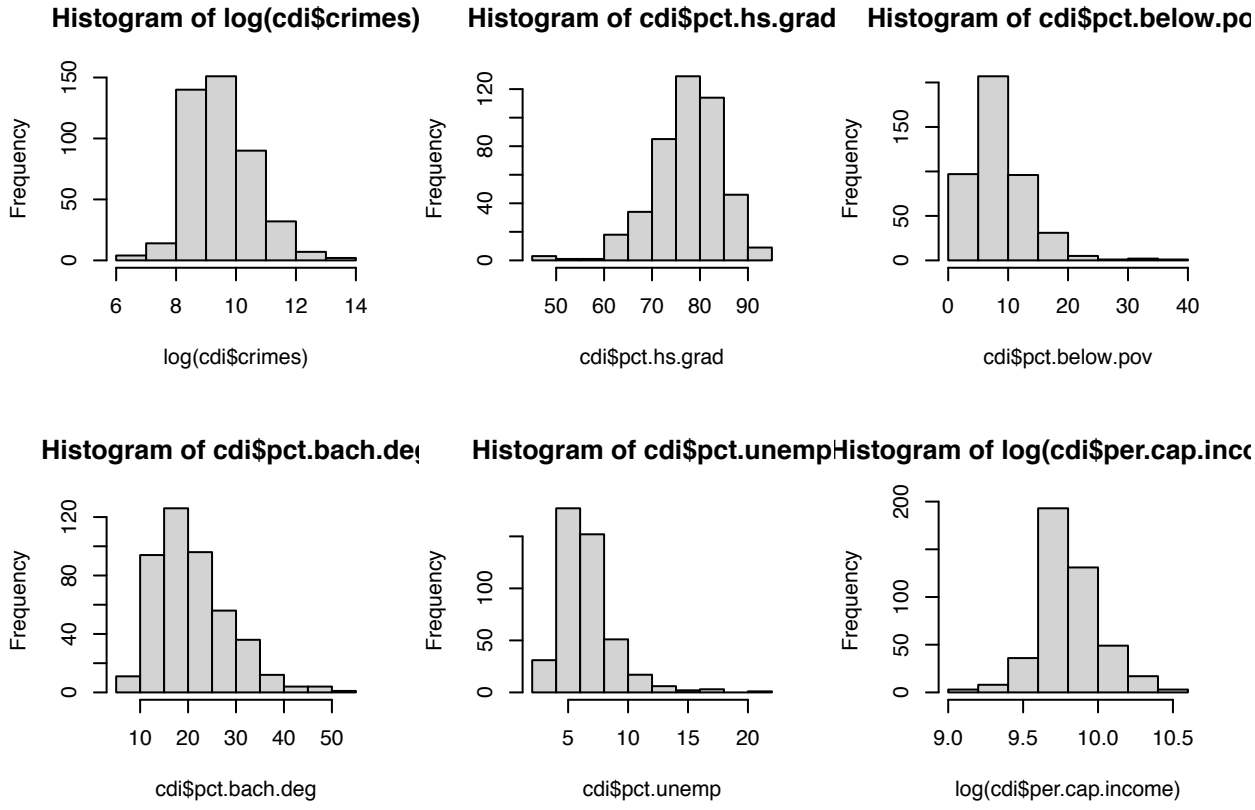


We show the boxplot for categorical variable region by plotting each region's boxplot and showing how per.cap.income varies across the four regions of the country. There is a lot of overlap in the boxplots, but the Northeast and the West seem to do a little better than the North Central and the South. We can observe that for region "S," it has the most number 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 are evenly distributed but for the region "S" and "W," data points have more dispersions compared to data points in the region "NE."

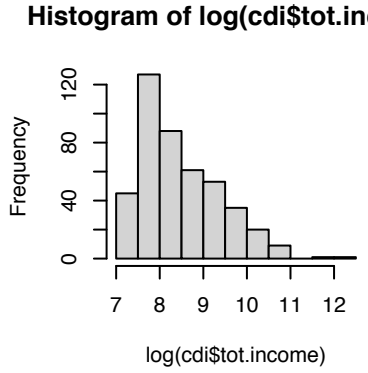
```
# histograms for transformed data
par(mfrow=c(2,3))
hist(log(cdi$land.area))
hist(log(cdi$pop))
hist(cdi$pop.18_34)
hist(cdi$pop.65_plus)
hist(log(cdi$doctors))
hist(log(cdi$hosp.beds))
```




```
hist(log(cdi$crimes))
hist(cdi$pct.hs.grad)
hist(cdi$pct.below.pov)
hist(cdi$pct.bach.deg)
hist(cdi$pct.unemp)
hist(log(cdi$per.cap.income))
```




```
hist(log(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", "Percent Below Poverty Level", "Percent Bachelor's Degrees", and "Percent Unemployment": keep them simple for later explanation to social scientists, no need to do data transformation
For variable "Total Income": right-skewed, need to do log transformation

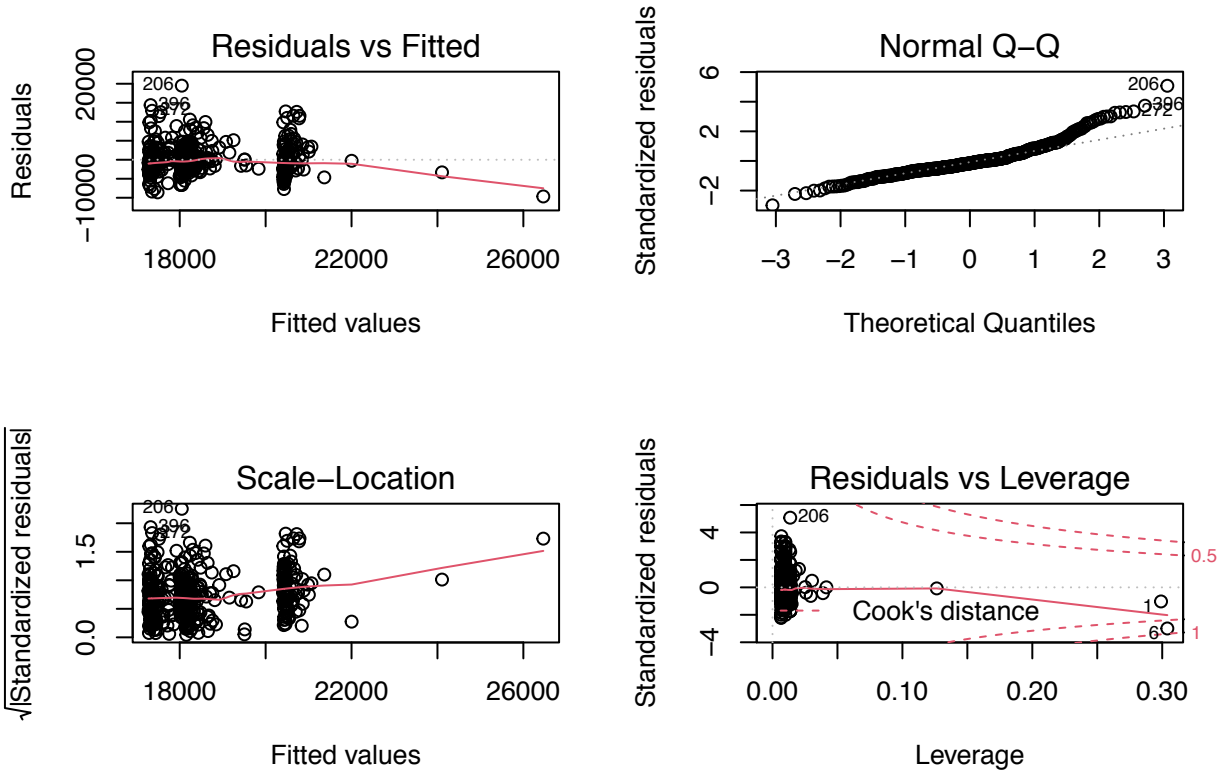
Since we know that the response variable per.cap.income is mathematically calculated by tot.income divided by pop, then we remove two variables: tot.income and pop when fitting the multiple regression model.

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)
```



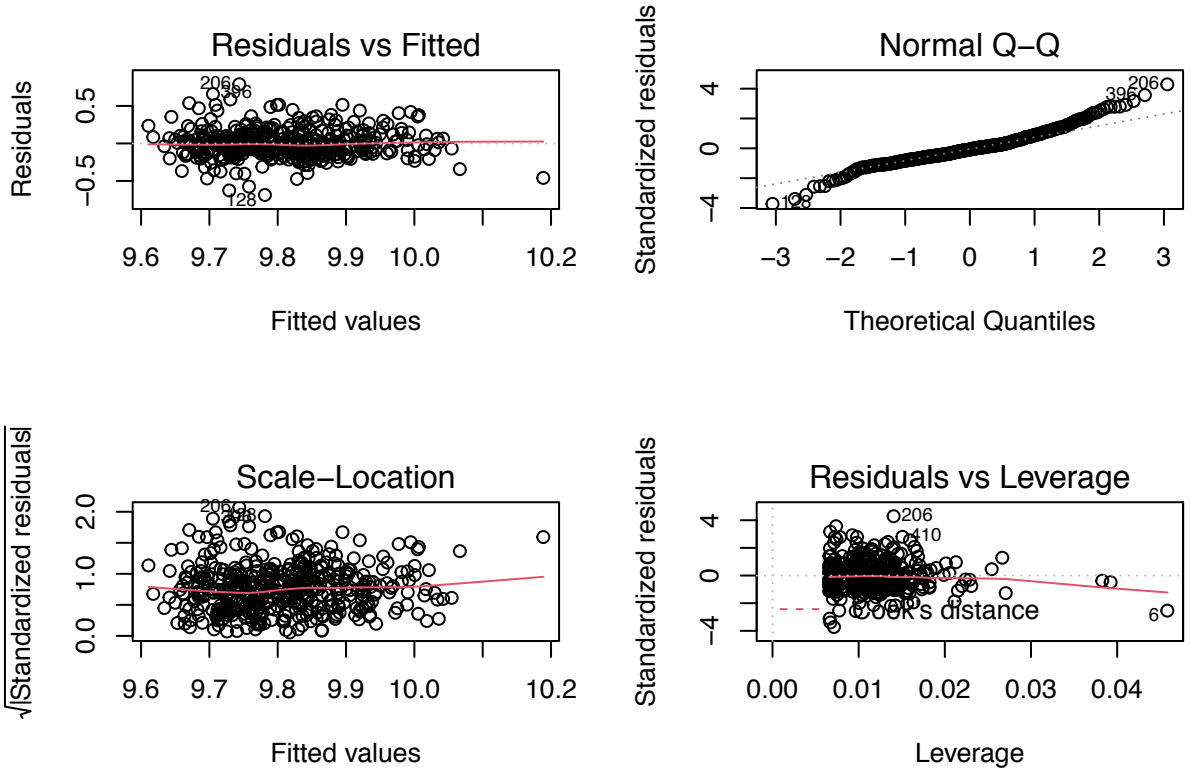
From the summary of data with only including crime rate and region of the country, we can observe that the adjusted R-squared is 0.09288, meaning that only 9.29% of its variability can be explained. Also, looking back to the four diagnostic plots show that the model performs not well with validity problems. From the Residuals vs. Fitted plot, we can clearly that the linearity condition is satisfied since there is not a horizontal line around 0, and the line has a downward-curve pattern. From the Normal Q-Q plot, we can clearly observe that the normality assumption is not satisfied because there is apparent outliers for example point 206 and point 396. From the Scale-Location plot, we can clearly observe that there is a non-constant variance problem with the non-horizontal line around 1 with upward-curve pattern. 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 points appeared like point 1 and point 6.

```
# 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)
```



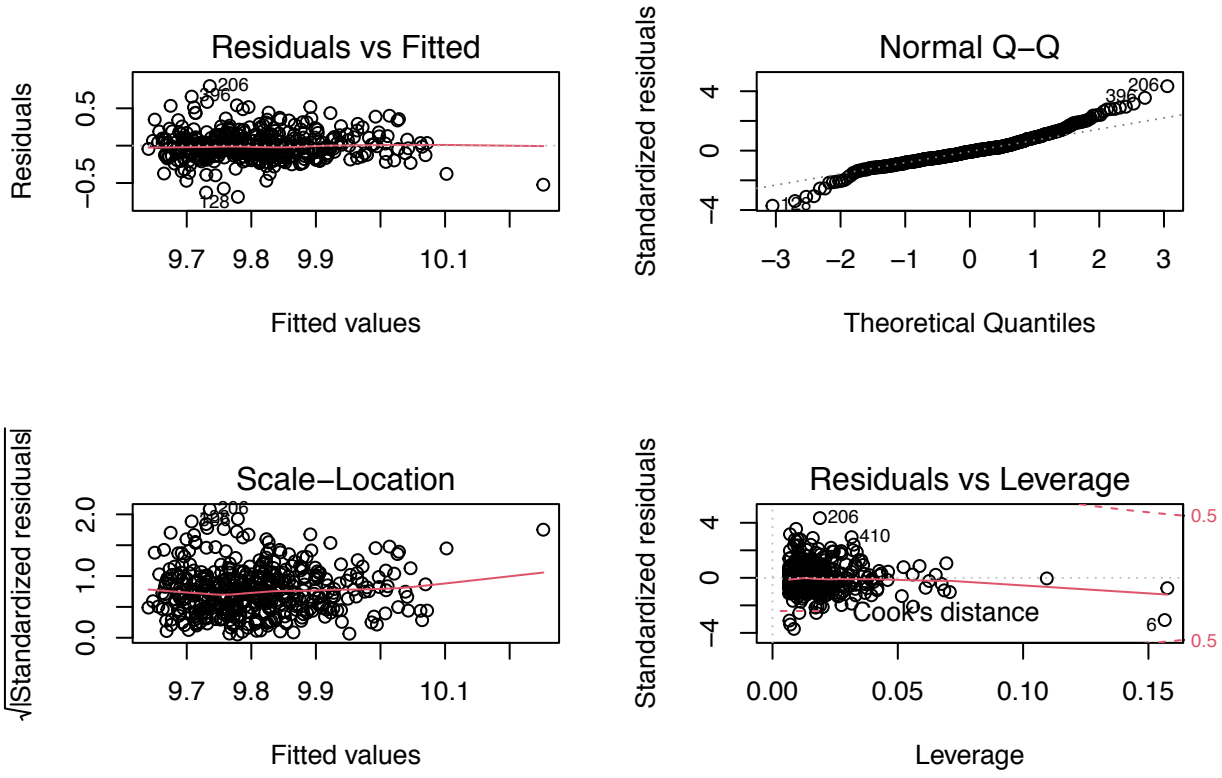
From the summary of data with including log-transformation on both per-capita income and crime rates, we can observe that the adjusted R-squared is 0.1959, meaning that only 19.59% of its variability can be explained, and the adjusted R-squared value improves a lot comparing to preivous model without any data transformation. Also, looking back to the four diagnostic plots, they show that the model performs better comparing to the previous model. From the Residuals vs. Fitted plot, we can clearly 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 apparent outlier for example point 206. From the Scale-Location plot, we can clearly observe that there is a slightly non-constant variance problem with the nearly horizontal line around 1 but just with slight fluctuating pattern. From the Residuals vs. Leverage plot, we can clearly observe that there are several outliers for example point 206 and point 410 (greater or lower than the absolute value of 2) and high leverage point appeared like point 6.

Later, we want 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.

```
# 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)
```



```r
# 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
```

From the above ANOVA table, we can observe that the with the F-statistics=0.7434, there is no 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 be not included in the model.

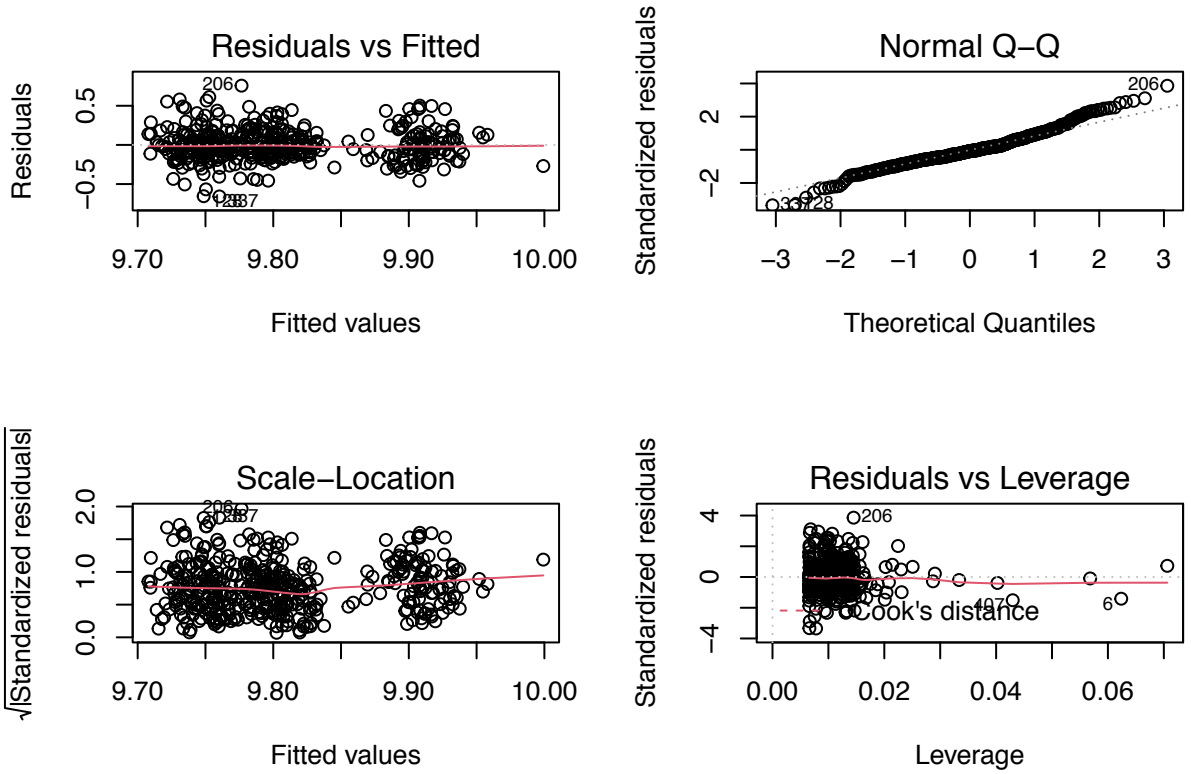
Later, we would like to investigate whether using the number of crimes or using per-capita crime (which is defined as number of crimes/population) will make any difference on choosing the best model.

First, we fit the model with transformed data by replacing log(crimes) with per-capita crime measure:

```r
# 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
## -0.65832 -0.11431 -0.01548  0.10838  0.75657
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.93628    0.06934 143.303  < 2e-16 ***
## log(crimes/pop)  0.04243    0.02148   1.975  0.04885 *
## regionNE         0.11457    0.02760   4.151 3.99e-05 ***
## regionS         -0.07456    0.02624  -2.841  0.00471 **
## regionW         -0.02426    0.03002  -0.808  0.41952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1974 on 435 degrees of freedom
## Multiple R-squared:  0.09645,    Adjusted R-squared:  0.08814
## F-statistic: 11.61 on 4 and 435 DF,  p-value: 5.776e-09
```

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



From the summary of data with using per-capita crime, we can observe that the adjusted R-squared is 0.08841, meaning that 8.841% of its variability can be explained, and the adjusted R-squared value is much smaller 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 comparing to the previous model with using the number of crimes as the variable. From the Residuals vs. Fitted plot, we can clearly 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 apparent outlier for example point 206. From the Scale-Location plot, we can clearly observe that there is a slightly non-constant variance problem with the nearly horizontal line around 1 but just with slight fluctuating pattern. 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.

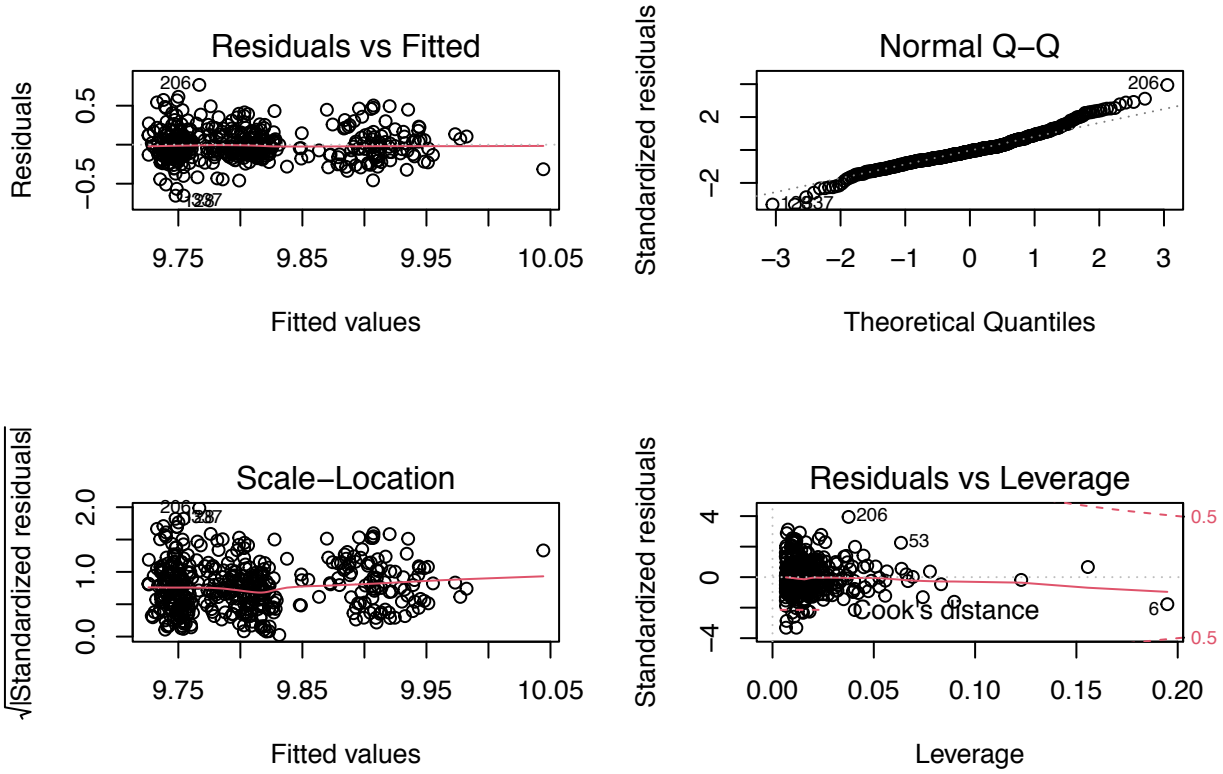
Furthermore, we want to check whether it is necessary to include the interaction term in the model using "per-capita crime" as the predictor variable. We create the ANOVA table and compare the model with transformed data added interaction term between "per-capita crime" and regions to the model with only transformed data.

```
# 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
```

```
## -0.65410 -0.11829 -0.01708  0.10399  0.76628
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 9.91177    0.10503  94.367   <2e-16 ***
## log(crimes/pop)             0.03454    0.03327   1.038    0.300
## regionNE                    0.21007    0.17165   1.224    0.222
## regionS                    -0.10137    0.16072  -0.631    0.529
## regionW                     0.07689    0.26753   0.287    0.774
## log(crimes/pop):regionNE    0.02924    0.05232   0.559    0.577
## log(crimes/pop):regionS    -0.01104    0.05554  -0.199    0.843
## log(crimes/pop):regionW     0.03495    0.09268   0.377    0.706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.198 on 432 degrees of freedom
## Multiple R-squared:  0.09773,    Adjusted R-squared:  0.08311
## F-statistic: 6.685 on 7 and 432 DF,  p-value: 1.575e-07
```

```
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    435 16.952
## 2    432 16.928  3   0.02408 0.2048  0.893
```

From the above ANOVA table, we can observe that the with the F-statistics=0.2048, there is no enough evidence against the reduced model in favor of the full model. Also, given the p-value 0.893, 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 be not included in the model.

If we want to compare these two winners (ancova.02 vs. ancova.05), we need to use AIC or BIC, because the two winners are not nested models (you can't get one from the other by imposing one or more linear constraints).

```
# compare the final two candidate models - revise
AIC(cdi_fit2,cdi_fit4)
```

```
##          df       AIC
## cdi_fit2  6 -227.4746
## cdi_fit4  6 -172.1347
```

```
BIC(cdi_fit2,cdi_fit4)
```

```
##          df       BIC
## cdi_fit2  6 -202.9539
## cdi_fit4  6 -147.6140
```

I prefer using the model with the "number of crimes" as the crime rate measure with two following reasons. First, we can observe from the AIC and BIC comparison results on these two candidate models, the model with "number of crimes" as measurement has both the smaller AIC and BIC value. Since we know that the smaller AIC and BIC value, the better the model, then the model with "number of crimes" performs better compared to the one with measurement "number of crimes/population". Second, for the real-world setting, if we introduce the concept of "number of crimes" to social scientists, it would be easier for them to understand and later if they want to use the dataset to do further analysis with more updated data, then the data form can keep consistent without any calculation needed.

Here, we make summary on the final model's formula and summary table of coefficient estimators:

```
formula(cdi_fit2)
```

```
## log(per.cap.income) ~ log(crimes) + region
```

```
round(coef(summary(cdi_fit2)),2)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.19       0.08  115.13     0.00
## log(crimes)     0.07       0.01    7.92     0.00
## regionNE        0.10       0.03    4.09     0.00
## regionS        -0.09       0.02   -3.68     0.00
## regionW        -0.06       0.03   -1.96     0.05
```

From the above summary table, we can interpret the results as following:
1) Among the entire United States, for every 1% increase in crimes, we are expecting to observe a 0.07% increase in per-capita income, on average.
2) Different regions of the country have different baseline per-capita income. In the NC region, the baseline salary is exp(9.19)= 9,798.65 dollars. Similarly, in the NE region, the baseline salary is 10,829.18 dollars; in the S region, the baseline salary is 8,955.29 dollars; and in the W region, the baseline salary is 9,228.02 dollars. Therefore, the level of salary varies from region to region.

Also, I keep the other candidate model's summary here just for reference:

```
# summary of the model using per-capita crime as measurement
formula(cdi_fit4)
```

```
## log(per.cap.income) ~ log(crimes/pop) + region
```

```
round(coef(summary(cdi_fit4)),2)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)         9.94       0.07  143.30     0.00
## log(crimes/pop)     0.04       0.02    1.98     0.05
## regionNE            0.11       0.03    4.15     0.00
## regionS            -0.07       0.03   -2.84     0.00
## regionW            -0.02       0.03   -0.81     0.42
```

From the above summary table, we can interpret the results as following:
1) Among the entire United States, for every 1% increase in crimes, we are expecting to observe a 0.04% increase in per-capita income, on average.
2) Different regions of the country have different baseline per-capita income. In the NC region, the baseline salary is exp(9.94)= 20,743.74 dollars. Similarly, in the NE region, the baseline salary is 23,155.79 dollars; in the S region, the baseline salary is 19,341.34 dollars; and in the W region, the baseline salary is 20,332.99 dollars. Therefore, the level of salary varies from region to region.

Appendix 3. Multiple Regression Analysis

Before beginning, I need to take log.pop and log.tot.income out of consideration, since per.cap.income is a deterministic function of them.

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

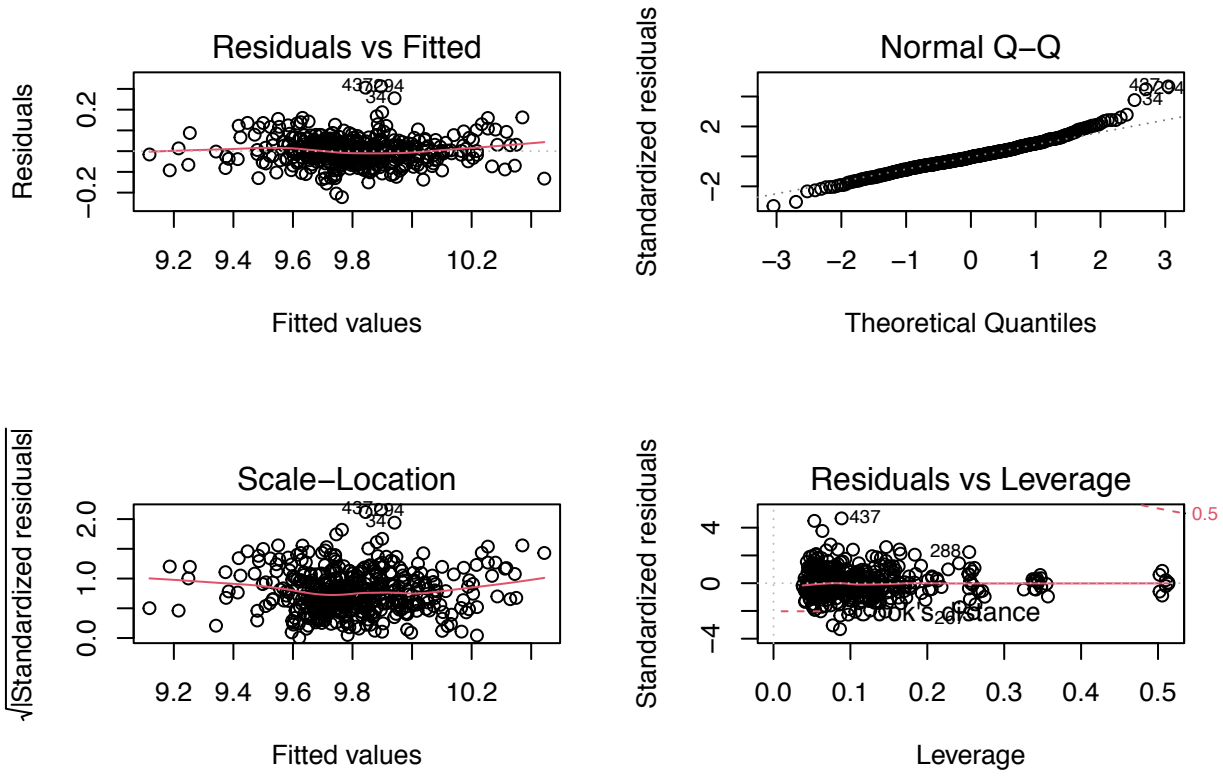
```
mulreg_fit1<-lm(log(per.cap.income)~log(land.area)+pop.18_34+pop.65_plus
                +log(doctors)+log(hosp.beds)+log(crimes)
                +pct.hs.grad+pct.below.pov+pct.bach.deg
                +pct.unemp+region+state,data=cdi)
summary(mulreg_fit1)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + pop.18_34 +
##     pop.65_plus + log(doctors) + log(hosp.beds) + log(crimes) +
##     pct.hs.grad + pct.below.pov + pct.bach.deg + pct.unemp +
##     region + state, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22173 -0.03797 -0.00185  0.03334  0.31229
##
## Coefficients: (3 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.1822860  0.1241179  82.037  < 2e-16 ***
## log(land.area) -0.0340120  0.0057658  -5.899 8.06e-09 ***
## pop.18_34      -0.0154677  0.0012431 -12.443  < 2e-16 ***
## pop.65_plus    -0.0001396  0.0015022  -0.093 0.925984
## log(doctors)    0.0429905  0.0128155   3.355 0.000874 ***
## log(hosp.beds)  0.0074359  0.0126342   0.589 0.556510
## log(crimes)     0.0059008  0.0087784   0.672 0.501866
```

```
## pct.hs.grad    -0.0026516  0.0012170  -2.179 0.029957 *
## pct.below.pov  -0.0203284  0.0015125 -13.440  < 2e-16 ***
## pct.bach.deg    0.0142266  0.0009906  14.361  < 2e-16 ***
## pct.unemp       0.0012396  0.0027903   0.444 0.657106
## regionNE       -0.0491818  0.0743131  -0.662 0.508486
## regionS        -0.0175857  0.0353292  -0.498 0.618935
## regionW        -0.0378182  0.0321571  -1.176 0.240308
## stateAR        -0.0512794  0.0562663  -0.911 0.362675
## stateAZ        -0.0629723  0.0403582  -1.560 0.119509
## stateCA         0.0992255  0.0266825   3.719 0.000230 ***
## stateCO         0.0160398  0.0326781   0.491 0.623820
## stateCT         0.1370406  0.0749204   1.829 0.068157 .
## stateDC         0.0662625  0.0773568   0.857 0.392212
## stateDE         0.0514564  0.0867537   0.593 0.553443
## stateFL        -0.0408069  0.0326580  -1.250 0.212239
## stateGA         0.0345818  0.0362400   0.954 0.340565
## stateHI         0.0546152  0.0476508   1.146 0.252448
## stateID         0.0043993  0.0737425   0.060 0.952460
## stateIL         0.0383543  0.0278136   1.379 0.168709
## stateIN        -0.0247281  0.0286778  -0.862 0.389079
## stateKS        -0.0354295  0.0413187  -0.857 0.391723
## stateKY        -0.0142186  0.0491681  -0.289 0.772599
## stateLA         0.0355765  0.0363969   0.977 0.328959
## stateMA         0.0805178  0.0743617   1.083 0.279587
## stateMD         0.0363191  0.0371440   0.978 0.328797
## stateME         0.0416228  0.0775607   0.537 0.591823
## stateMI         0.0427306  0.0281412   1.518 0.129732
## stateMN        -0.0276522  0.0342913  -0.806 0.420519
## stateMO        -0.0046421  0.0331309  -0.140 0.888644
## stateMS        -0.0524006  0.0490281  -1.069 0.285840
## stateMT         0.0353103  0.0743390   0.475 0.635066
## stateNC        -0.0205307  0.0322535  -0.637 0.524804
## stateND        -0.0449402  0.0744223  -0.604 0.546299
## stateNE        -0.0760466  0.0468574  -1.623 0.105428
## stateNH         0.0823994  0.0789629   1.044 0.297367
## stateNJ         0.1458632  0.0732375   1.992 0.047121 *
## stateNM        -0.0642114  0.0554348  -1.158 0.247456
## stateNV         0.2105018  0.0556949   3.780 0.000182 ***
## stateNY         0.0529419  0.0724781   0.730 0.465561
## stateOH        -0.0093760  0.0257145  -0.365 0.715599
## stateOK        -0.0565242  0.0448730  -1.260 0.208565
## stateOR        -0.0409777  0.0363544  -1.127 0.260377
## statePA         0.0151572  0.0722563   0.210 0.833959
## stateRI        -0.0153824  0.0820802  -0.187 0.851441
## stateSC        -0.0200973  0.0344451  -0.583 0.559929
## stateSD         0.0008370  0.0737597   0.011 0.990952
## stateTN        -0.0187489  0.0365972  -0.512 0.608734
## stateTX         0.0004159  0.0303082   0.014 0.989058
## stateUT        -0.2514920  0.0421830  -5.962 5.68e-09 ***
## stateVA         0.0054090  0.0394628   0.137 0.891052
## stateVT                NA         NA      NA       NA
## stateWA                NA         NA      NA       NA
## stateWI                NA         NA      NA       NA
## stateWV        -0.0075967  0.0753022  -0.101 0.919696
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07004 on 382 degrees of freedom
## Multiple R-squared:  0.9001, Adjusted R-squared:  0.8852
## F-statistic: 60.39 on 57 and 382 DF,  p-value: < 2.2e-16
```

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



From the summary of table, we can observe 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)+pct.hs.grad+pct.below.pov
                    +pct.bach.deg+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) +
##     pct.hs.grad + pct.below.pov + pct.bach.deg + pct.unemp +
##     region + state, data = cdi, really.big = T)
## 60 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
## pct.below.pov      FALSE      FALSE
## pct.bach.deg       FALSE      FALSE
## 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
## 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 pct.below.pov pct.bach.deg pct.unemp regionNE
## 1  ( 1 ) " "         " "         " "           "*"          " "       " "
## 2  ( 1 ) " "         " "         "*"           " "          " "       " "
## 3  ( 1 ) " "         " "         "*"           "*"          " "       " "
## 4  ( 1 ) " "         " "         "*"           "*"          " "       " "
## 5  ( 1 ) " "         " "         "*"           "*"          " "       " "
## 6  ( 1 ) " "         " "         "*"           "*"          " "       " "
## 7  ( 1 ) " "         " "         "*"           "*"          " "       " "
## 8  ( 1 ) " "         " "         "*"           "*"          " "       " "
## 9  ( 1 ) " "         " "         "*"           "*"          " "       " "
##          regionS regionW stateAR stateAZ stateCA stateCO stateCT stateDC
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 7  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 8  ( 1 ) " "     " "     " "     " "     "*"     " "     " "     " "
## 9  ( 1 ) " "     " "     " "     " "     "*"     " "     "*"     " "
##          stateDE stateFL stateGA stateHI stateID stateIL stateIN stateKS
## 1  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 2  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 3  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 4  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 5  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
## 6  ( 1 ) " "     " "     " "     " "     " "     " "     " "     " "
```

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

```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
```

```r
best.model <- which.min(cdi_sum$bic)
# ADD revise - coeff of the subsets regression best model
coef(mulreg_fit2,best.model)
```

```
##    (Intercept) log(land.area)      pop.18_34   log(doctors)  pct.below.pov
##    10.03919588    -0.04102795    -0.01474387     0.05756553    -0.01857332
##   pct.bach.deg        stateCA        stateCT        stateNJ        stateUT
##     0.01201238     0.08713006     0.10785995     0.12053000    -0.29161335
```

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)
+pct.below.pov+pct.bach.deg+state, with state being specifically in CA, NJ, NV, and UT.

Later, we try to include region as the interaction term with other variables to see whether it helps anyways:

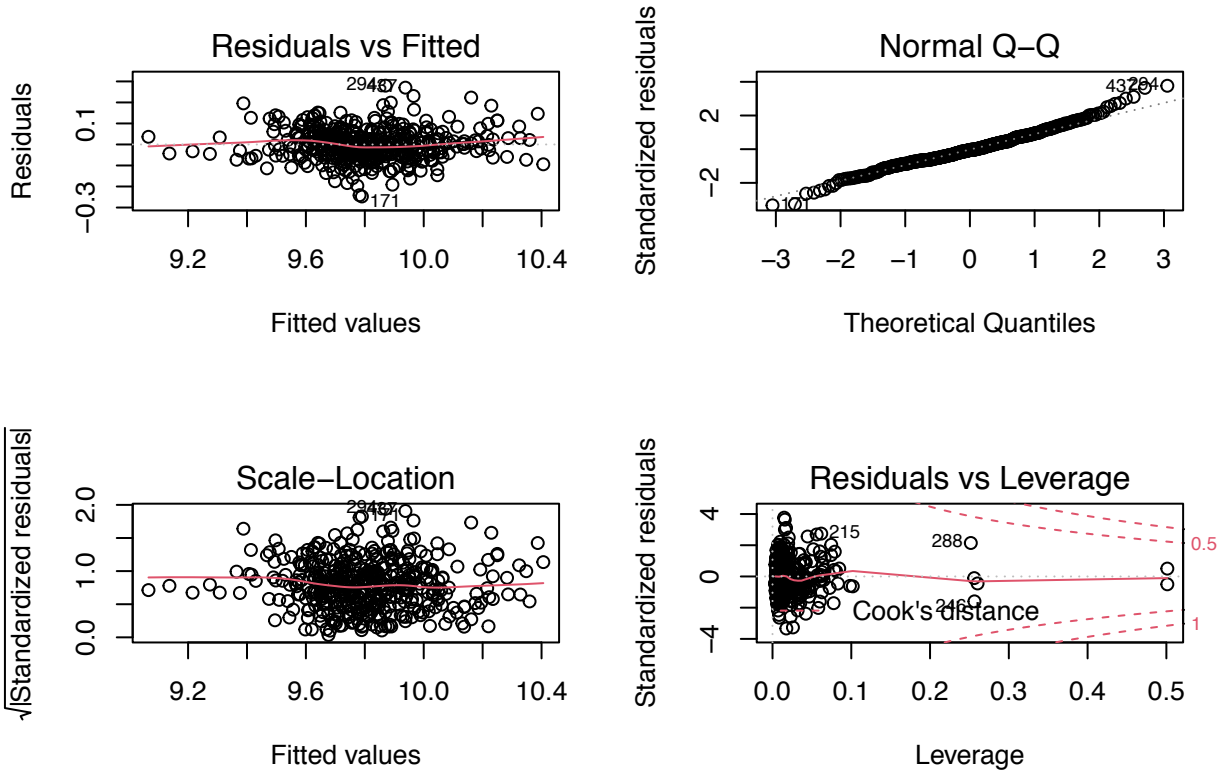
```r
# include the interaction term for model 1
subsets_with_region <-lm(log(per.cap.income)~log(land.area)+pop.18_34
                     +log(doctors)+pct.bach.deg+pct.below.pov
                   +log(land.area)*region+pop.18_34*region
                   +log(doctors)*region+pct.bach.deg*region
                   +pct.below.pov*region,data=cdi)
summary(subsets_with_region)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + pop.18_34 +
##     log(doctors) + pct.bach.deg + pct.below.pov + log(land.area) *
##     region + pop.18_34 * region + log(doctors) * region + pct.bach.deg *
##     region + pct.below.pov * region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36231 -0.04825 -0.00362  0.04544  0.30180
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             10.0704991  0.1245747  80.839  < 2e-16 ***
## log(land.area)          -0.0319298  0.0161166  -1.981   0.0482 *
## pop.18_34               -0.0155107  0.0028049  -5.530 5.67e-08 ***
## log(doctors)             0.0575045  0.0097083   5.923 6.62e-09 ***
## pct.bach.deg             0.0098008  0.0021747   4.507 8.57e-06 ***
## pct.below.pov           -0.0198948  0.0035522  -5.601 3.89e-08 ***
```

```
## regionNE                 0.0728614  0.1838947   0.396   0.6922
## regionS                 -0.0654637  0.1536426  -0.426   0.6703
## regionW                 -0.3762732  0.1789846  -2.102   0.0361 *
## log(land.area):regionNE -0.0216170  0.0208017  -1.039   0.2993
## log(land.area):regionS  -0.0095513  0.0186916  -0.511   0.6096
## log(land.area):regionW   0.0066567  0.0195926   0.340   0.7342
## pop.18_34:regionNE      -0.0005554  0.0039893  -0.139   0.8893
## pop.18_34:regionS        0.0004853  0.0032486   0.149   0.8813
## pop.18_34:regionW        0.0058919  0.0044526   1.323   0.1865
## log(doctors):regionNE   -0.0026384  0.0137913  -0.191   0.8484
## log(doctors):regionS     0.0015352  0.0121565   0.126   0.8996
## log(doctors):regionW     0.0125259  0.0137510   0.911   0.3629
## pct.bach.deg:regionNE    0.0060710  0.0030237   2.008   0.0453 *
## pct.bach.deg:regionS     0.0029394  0.0024771   1.187   0.2360
## pct.bach.deg:regionW     0.0015277  0.0029735   0.514   0.6077
## pct.below.pov:regionNE  -0.0018406  0.0051342  -0.358   0.7202
## pct.below.pov:regionS    0.0026630  0.0038148   0.698   0.4855
## pct.below.pov:regionW    0.0047389  0.0047519   0.997   0.3192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08394 on 416 degrees of freedom
## Multiple R-squared:  0.8438, Adjusted R-squared:  0.8351
## F-statistic: 97.69 on 23 and 416 DF,  p-value: < 2.2e-16
```

From the above summary table, we can observe that all the interaction terms that involve region have p-value greater than 0.05, which is not statistically significant. Therefore, we conclude that we do not choose the model with any interaction term on region.

```
#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)+pct.below.pov+pct.bach.deg
               +stateCA+stateNJ+stateNV+stateUT,data=cdi)
par(mfrow=c(2,2))
plot(can_model1)
```




```
summary(can_model1)
```

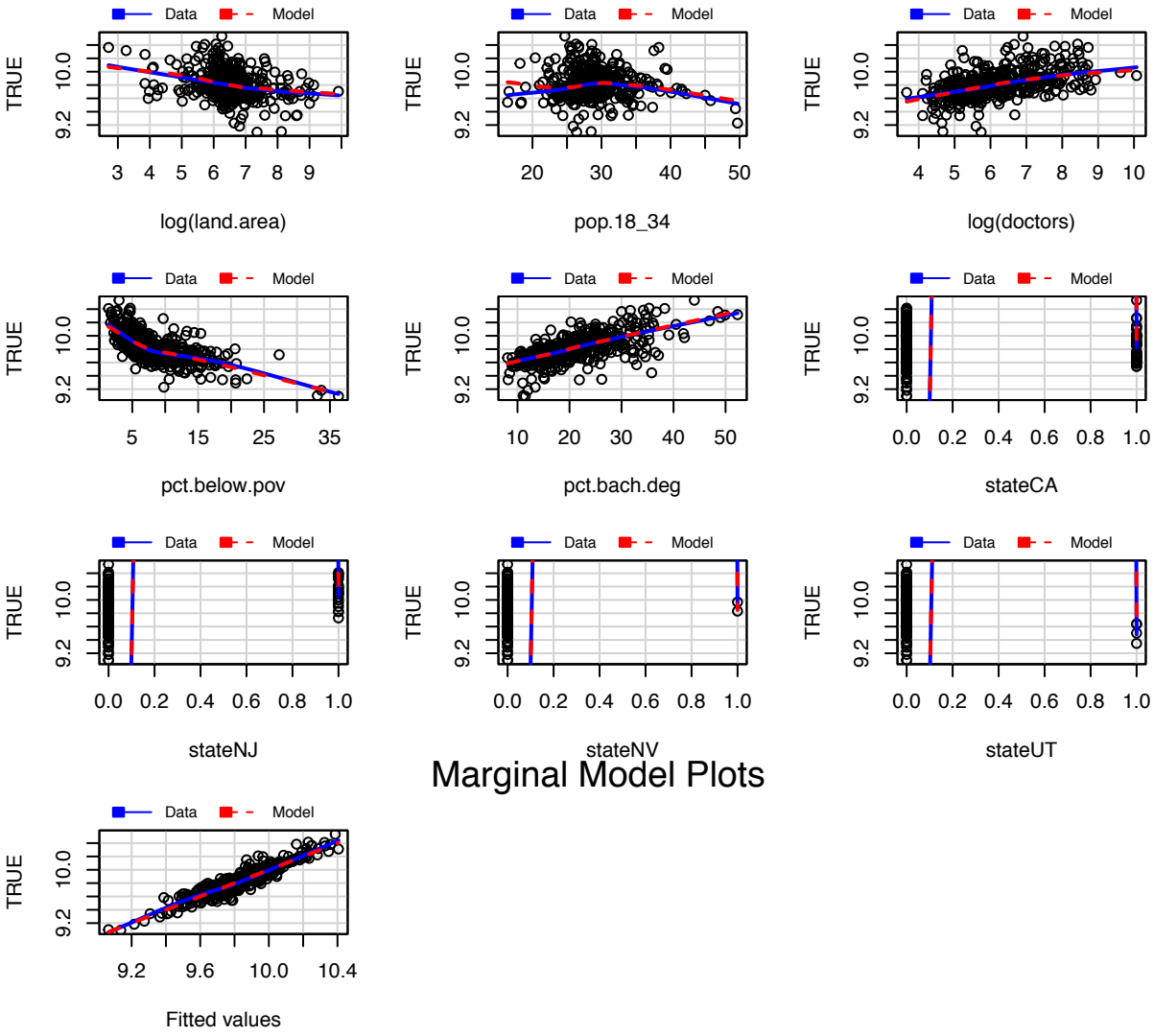
```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + pop.18_34 +
##     log(doctors) + pct.below.pov + pct.bach.deg + stateCA + stateNJ +
##     stateNV + stateUT, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24732 -0.04806 -0.00380  0.04463  0.27972
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.0715855  0.0465624 216.303  < 2e-16 ***
## log(land.area) -0.0445245  0.0046354  -9.605  < 2e-16 ***
## pop.18_34      -0.0150253  0.0010130 -14.833  < 2e-16 ***
## log(doctors)    0.0567865  0.0036957  15.365  < 2e-16 ***
## pct.below.pov  -0.0186611  0.0009176 -20.337  < 2e-16 ***
## pct.bach.deg    0.0122523  0.0006892  17.779  < 2e-16 ***
## stateCA         0.0897060  0.0143763   6.240 1.05e-09 ***
## stateNJ         0.1154899  0.0186336   6.198 1.34e-09 ***
## stateNV         0.2030343  0.0545426   3.722 0.000223 ***
## stateUT        -0.2931336  0.0376534  -7.785 5.25e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07486 on 430 degrees of freedom
## Multiple R-squared:  0.8716, Adjusted R-squared:  0.8689
```

```
## F-statistic: 324.3 on 9 and 430 DF,  p-value: < 2.2e-16
```

```
# vif of candidate model1
vif(can_model1)
```

```
## log(land.area)      pop.18_34   log(doctors)  pct.below.pov   pct.bach.deg
##       1.279192       1.412043       1.400598       1.430453       2.180185
##        stateCA        stateNJ        stateNV        stateUT
##       1.157187       1.069748       1.057007       1.002899
```

```
# mmps of candidate model1
mmps(can_model1)
```



As we explore more on the model with variable state included but without interaction term on state, we can observe that in the summary table of VIF, none of the variables seem to have an excessively large value. Moreover, from the marginal plots, we can observe that for all variables' plots, the red model line matches with the blue data line very well. Therefore, we think that this model is valid.

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+pct.below.pov
                    +pct.bach.deg+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 +
##     log(doctors) + log(crimes/pop) + pct.hs.grad + pct.below.pov +
##     pct.bach.deg + state, data = cdi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.215557 -0.037591 -0.002943  0.032779  0.309925
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     10.3337515  0.1078747  95.794  < 2e-16 ***
## log(land.area)  -0.0306871  0.0057463  -5.340 1.59e-07 ***
## pop.18_34       -0.0157165  0.0010368 -15.159  < 2e-16 ***
## log(doctors)     0.0486410  0.0042871  11.346  < 2e-16 ***
## log(crimes/pop)  0.0281387  0.0111018   2.535 0.011652 *
## pct.hs.grad     -0.0028766  0.0011599  -2.480 0.013564 *
## pct.below.pov   -0.0208524  0.0014524 -14.358  < 2e-16 ***
## pct.bach.deg     0.0141053  0.0008935  15.786  < 2e-16 ***
## stateAR         -0.0601365  0.0557679  -1.078 0.281558
## stateAZ         -0.0902805  0.0425846  -2.120 0.034643 *
## stateCA          0.0796579  0.0293439   2.715 0.006934 **
## stateCO         -0.0108905  0.0358686  -0.304 0.761581
## stateCT          0.1168350  0.0372323   3.138 0.001832 **
## stateDC          0.0713433  0.0758554   0.941 0.347542
## stateDE          0.0176536  0.0560511   0.315 0.752966
## stateFL         -0.0455794  0.0299239  -1.523 0.128535
## stateGA          0.0299551  0.0356478   0.840 0.401258
## stateHI          0.0204598  0.0483687   0.423 0.672534
## stateID         -0.0178748  0.0748672  -0.239 0.811424
## stateIL          0.0619231  0.0317602   1.950 0.051936 .
## stateIN          0.0013978  0.0331543   0.042 0.966392
## stateKS         -0.0255016  0.0443662  -0.575 0.565763
## stateKY         -0.0099289  0.0485621  -0.204 0.838105
## stateLA          0.0372188  0.0358874   1.037 0.300340
## stateMA          0.0673322  0.0349176   1.928 0.054551 .
## stateMD          0.0349912  0.0351388   0.996 0.319973
## stateME          0.0139983  0.0411385   0.340 0.733836
## stateMI          0.0641118  0.0316433   2.026 0.043446 *
## stateMN         -0.0073216  0.0382036  -0.192 0.848120
## stateMO          0.0176265  0.0364924   0.483 0.629356
## stateMS         -0.0517426  0.0483535  -1.070 0.285250
## stateMT          0.0164568  0.0750786   0.219 0.826614
## stateNC         -0.0254100  0.0315640  -0.805 0.421299
## stateND         -0.0253177  0.0750439  -0.337 0.736020
## stateNE         -0.0561666  0.0491984  -1.142 0.254316
## stateNH          0.0603136  0.0443388   1.360 0.174534
```

```
## stateNJ          0.1214027  0.0323415   3.754 0.000201 ***
## stateNM         -0.0927136  0.0563457  -1.645 0.100695
## stateNV          0.1793033  0.0576783   3.109 0.002019 **
## stateNY          0.0292735  0.0308607   0.949 0.343434
## stateOH          0.0209051  0.0312844   0.668 0.504389
## stateOK         -0.0586660  0.0441857  -1.328 0.185059
## stateOR         -0.0673362  0.0397221  -1.695 0.090849 .
## statePA         -0.0010710  0.0309196  -0.035 0.972387
## stateRI         -0.0423503  0.0489006  -0.866 0.387003
## stateSC         -0.0293659  0.0338706  -0.867 0.386480
## stateSD          0.0152107  0.0746465   0.204 0.838641
## stateTN         -0.0202202  0.0361493  -0.559 0.576247
## stateTX         -0.0029967  0.0295618  -0.101 0.919310
## stateUT         -0.2699189  0.0449612  -6.003 4.47e-09 ***
## stateVA          0.0057092  0.0376180   0.152 0.879450
## stateVT         -0.0283266  0.0748568  -0.378 0.705334
## stateWA         -0.0238256  0.0355036  -0.671 0.502575
## stateWI          0.0187875  0.0346471   0.542 0.587959
## stateWV         -0.0044904  0.0743231  -0.060 0.951855
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06932 on 385 degrees of freedom
## Multiple R-squared:  0.9014, Adjusted R-squared:  0.8876
## F-statistic: 65.17 on 54 and 385 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) + pct.below.pov + pct.bach.deg + state;

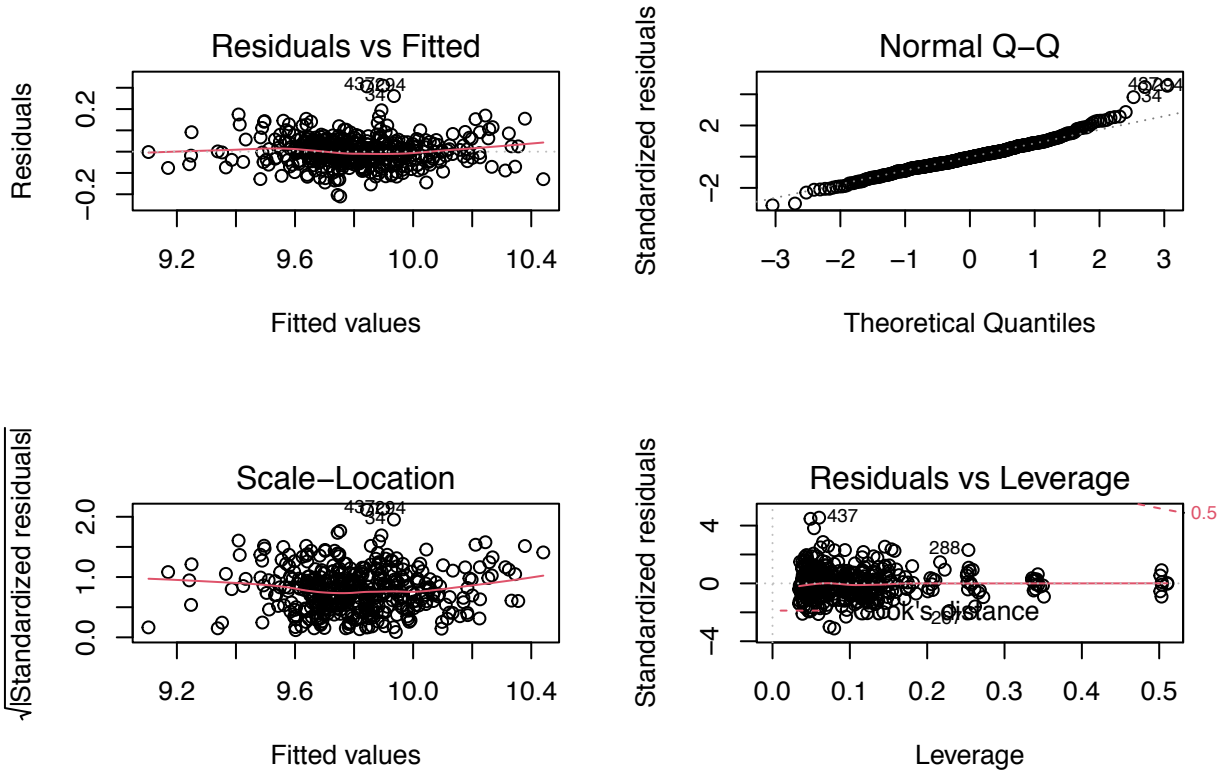
```
# include the interaction term for model 2
stepwise_with_region <-lm(log(per.cap.income)~log(land.area)+pop.18_34+pop.65_plus
                          +log(doctors)+pct.bach.deg+pct.below.pov
                    +log(land.area)*region+pop.18_34*region
                    +log(doctors)*region+pct.bach.deg*region
                    +pct.below.pov*region,data=cdi)
summary(stepwise_with_region)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + pop.18_34 +
##     pop.65_plus + log(doctors) + pct.bach.deg + pct.below.pov +
##     log(land.area) * region + pop.18_34 * region + log(doctors) *
##     region + pct.bach.deg * region + pct.below.pov * region,
##     data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36440 -0.04772 -0.00468  0.04534  0.30326
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             10.0869558  0.1256405  80.284  < 2e-16 ***
## log(land.area)          -0.0312783  0.0161293  -1.939   0.0532 .
```

```
## pop.18_34               -0.0157845  0.0028180  -5.601 3.88e-08 ***
## pop.65_plus             -0.0014996  0.0014895  -1.007   0.3146
## log(doctors)             0.0581858  0.0097317   5.979 4.85e-09 ***
## pct.bach.deg             0.0096614  0.0021791   4.434 1.19e-05 ***
## pct.below.pov           -0.0194579  0.0035785  -5.437 9.25e-08 ***
## regionNE                 0.0926692  0.1849413   0.501   0.6166
## regionS                 -0.0279912  0.1580845  -0.177   0.8595
## regionW                 -0.3633323  0.1794427  -2.025   0.0435 *
## log(land.area):regionNE -0.0222686  0.0208115  -1.070   0.2852
## log(land.area):regionS  -0.0111330  0.0187572  -0.594   0.5531
## log(land.area):regionW   0.0063222  0.0195951   0.323   0.7471
## pop.18_34:regionNE      -0.0009411  0.0040076  -0.235   0.8145
## pop.18_34:regionS       -0.0003458  0.0033518  -0.103   0.9179
## pop.18_34:regionW        0.0053464  0.0044854   1.192   0.2340
## log(doctors):regionNE   -0.0025398  0.0137914  -0.184   0.8540
## log(doctors):regionS     0.0023354  0.0121823   0.192   0.8481
## log(doctors):regionW     0.0124941  0.0137508   0.909   0.3641
## pct.bach.deg:regionNE    0.0060530  0.0030237   2.002   0.0459 *
## pct.bach.deg:regionS     0.0028651  0.0024781   1.156   0.2483
## pct.bach.deg:regionW     0.0017605  0.0029824   0.590   0.5553
## pct.below.pov:regionNE  -0.0021058  0.0051409  -0.410   0.6823
## pct.below.pov:regionS    0.0020103  0.0038695   0.520   0.6037
## pct.below.pov:regionW    0.0045315  0.0047563   0.953   0.3413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08394 on 415 degrees of freedom
## Multiple R-squared:  0.8442, Adjusted R-squared:  0.8351
## F-statistic: 93.66 on 24 and 415 DF,  p-value: < 2.2e-16
```

From the above summary table, we can observe that all the interaction terms that involve region have p-value greater than 0.05, which is not statistically significant. Therefore, we conclude that we do not choose the model with any interaction term on region.

```
#compare candidate model2
can_model2<-lm(log(per.cap.income) ~ log(land.area) + pop.18_34
               + pop.65_plus + log(doctors) + pct.below.pov
               + 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) + pct.below.pov + pct.bach.deg +
##     state, data = cdi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.210328 -0.039617 -0.001907  0.036728  0.309850
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.0313019  0.0686766 146.066  < 2e-16 ***
## log(land.area) -0.0348435  0.0057425  -6.068 3.10e-09 ***
## pop.18_34      -0.0157836  0.0012144 -12.997  < 2e-16 ***
## pop.65_plus    -0.0001723  0.0014106  -0.122 0.902852
## log(doctors)    0.0554740  0.0036939  15.018  < 2e-16 ***
## pct.below.pov  -0.0178011  0.0010935 -16.279  < 2e-16 ***
## pct.bach.deg    0.0125070  0.0007307  17.115  < 2e-16 ***
## stateAR        -0.0547513  0.0564686  -0.970 0.332859
## stateAZ        -0.0926747  0.0432099  -2.145 0.032596 *
## stateCA         0.0768789  0.0297389   2.585 0.010100 *
## stateCO        -0.0203631  0.0357424  -0.570 0.569201
## stateCT         0.1059931  0.0373126   2.841 0.004740 **
## stateDC         0.0784831  0.0768780   1.021 0.307952
## stateDE         0.0194527  0.0568603   0.342 0.732451
## stateFL        -0.0428086  0.0314333  -1.362 0.174027
```

```
## stateGA         0.0284604  0.0361215   0.788 0.431234
## stateHI         0.0221152  0.0489009   0.452 0.651347
## stateID        -0.0391106  0.0754157  -0.519 0.604337
## stateIL         0.0509156  0.0320136   1.590 0.112555
## stateIN        -0.0154675  0.0332743  -0.465 0.642303
## stateKS        -0.0339707  0.0444695  -0.764 0.445387
## stateKY        -0.0209132  0.0491572  -0.425 0.670756
## stateLA         0.0196264  0.0360290   0.545 0.586247
## stateMA         0.0467257  0.0348883   1.339 0.181263
## stateMD         0.0345232  0.0357146   0.967 0.334329
## stateME         0.0014253  0.0416049   0.034 0.972689
## stateMI         0.0526146  0.0317048   1.660 0.097824 .
## stateMN        -0.0293679  0.0380516  -0.772 0.440710
## stateMO         0.0091964  0.0369105   0.249 0.803374
## stateMS        -0.0703663  0.0486966  -1.445 0.149271
## stateMT        -0.0114021  0.0755765  -0.151 0.880159
## stateNC        -0.0174468  0.0319212  -0.547 0.584999
## stateND        -0.0540949  0.0756636  -0.715 0.475078
## stateNE        -0.0772458  0.0491487  -1.572 0.116845
## stateNH         0.0478295  0.0448244   1.067 0.286620
## stateNJ         0.1216821  0.0326429   3.728 0.000222 ***
## stateNM        -0.0991374  0.0569192  -1.742 0.082354 .
## stateNV         0.1798472  0.0581954   3.090 0.002144 **
## stateNY         0.0171161  0.0311144   0.550 0.582568
## stateOH        -0.0038858  0.0309010  -0.126 0.899994
## stateOK        -0.0712550  0.0442120  -1.612 0.107852
## stateOR        -0.0812459  0.0395956  -2.052 0.040854 *
## statePA        -0.0213913  0.0306356  -0.698 0.485440
## stateRI        -0.0382753  0.0495975  -0.772 0.440753
## stateSC        -0.0151501  0.0341159  -0.444 0.657236
## stateSD         0.0004943  0.0755569   0.007 0.994784
## stateTN        -0.0196143  0.0366752  -0.535 0.593089
## stateTX        -0.0022855  0.0299401  -0.076 0.939193
## stateUT        -0.2981919  0.0445173  -6.698 7.47e-11 ***
## stateVA        -0.0015872  0.0381144  -0.042 0.966805
## stateVT        -0.0445434  0.0757751  -0.588 0.556985
## stateWA        -0.0400277  0.0350376  -1.142 0.253987
## stateWI         0.0048751  0.0348451   0.140 0.888806
## stateWV        -0.0151237  0.0753667  -0.201 0.841064
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07033 on 386 degrees of freedom
## Multiple R-squared:  0.8982, Adjusted R-squared:  0.8843
## F-statistic: 64.29 on 53 and 386 DF,  p-value: < 2.2e-16
```

```
vif(income_stepmod)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## log(land.area)   2.292199  1        1.514001
## pop.18_34        1.724818  1        1.313323
## log(doctors)     2.197605  1        1.482432
## log(crimes/pop)  2.851405  1        1.688610
## pct.hs.grad      6.048592  1        2.459389
## pct.below.pov    4.178754  1        2.044200
```

```
## pct.bach.deg     4.273605  1        2.067270
## state           16.726798 47        1.030422
```

```
formula(income_stepmod)
```

```
## log(per.cap.income) ~ log(land.area) + pop.18_34 + log(doctors) +
##     log(crimes/pop) + pct.hs.grad + pct.below.pov + pct.bach.deg +
##     state
```

```
coef(income_stepmod)
```

```
##     (Intercept)  log(land.area)       pop.18_34    log(doctors) log(crimes/pop)
##    10.333751521    -0.030687121    -0.015716547     0.048640957     0.028138716
##     pct.hs.grad   pct.below.pov    pct.bach.deg         stateAR         stateAZ
##    -0.002876599    -0.020852358     0.014105307    -0.060136549    -0.090280459
##         stateCA         stateCO         stateCT         stateDC         stateDE
##     0.079657931    -0.010890462     0.116835009     0.071343340     0.017653588
##         stateFL         stateGA         stateHI         stateID         stateIL
##    -0.045579408     0.029955122     0.020459751    -0.017874792     0.061923132
##         stateIN         stateKS         stateKY         stateLA         stateMA
##     0.001397818    -0.025501625    -0.009928850     0.037218829     0.067332157
##         stateMD         stateME         stateMI         stateMN         stateMO
##     0.034991223     0.013998338     0.064111848    -0.007321606     0.017626530
##         stateMS         stateMT         stateNC         stateND         stateNE
##    -0.051742608     0.016456839    -0.025409963    -0.025317724    -0.056166596
##         stateNH         stateNJ         stateNM         stateNV         stateNY
##     0.060313551     0.121402680    -0.092713612     0.179303276     0.029273510
##         stateOH         stateOK         stateOR         statePA         stateRI
##     0.020905103    -0.058665969    -0.067336192    -0.001070982    -0.042350341
##         stateSC         stateSD         stateTN         stateTX         stateUT
##    -0.029365935     0.015210717    -0.020220170    -0.002996652    -0.269918853
##         stateVA         stateVT         stateWA         stateWI         stateWV
##     0.005709179    -0.028326580    -0.023825605     0.018787456    -0.004490356
```

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)
+pct.below.pov+pct.bach.deg+state, with state being specifically in CA, NJ, NV, and UT.

```
# write out the formula for our final chosen model
formula(can_model1)
```

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

```
# write out the estimator coefficients for our final chosen model
round(summary(can_model1)$coef,2)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       10.07       0.05  216.30        0
## log(land.area)    -0.04       0.00   -9.61        0
```

```
## pop.18_34         -0.02       0.00  -14.83        0
## log(doctors)       0.06       0.00   15.37        0
## pct.below.pov     -0.02       0.00  -20.34        0
## pct.bach.deg       0.01       0.00   17.78        0
## stateCA            0.09       0.01    6.24        0
## stateNJ            0.12       0.02    6.20        0
## stateNV            0.20       0.05    3.72        0
## stateUT           -0.29       0.04   -7.79        0
```

```
# for research question #4, checking states
table(cdi$state)
```

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