Exploring the factors related to average income per person in the United States

Yanlin Li, yanlinli@andrew.cmu.edu

Abstract

In this study, we focused on various factors influencing per-capita income for different counties in the United States, which is an interesting topic for social scientists. We used the county demographic information data set from Geospatial and Statistical Data Center including several aspects of social well-beings to do our analysis. Methods such as Multiple Linear Regression, Box-Cox Transformation, Analysis of Variance, Variance Inflation Factor, and Bayesian Information Criterion were used in our analysis. We discovered several pairwise relationships of the social factors, discussed the relationship between crime and per-capita income, developed an optimal linear model for per-capita income, and illustrated the influence of missing values. Finally, all the relationships found were interpreted with real life circumstances. Our analysis is still limited by missing data, lack of other models, and further references, which should be tackled during next step.

Key words: social science, per-capita income, multiple linear regression

Introduction

Personal income is always something that can attract attention from all parts of society. People use their income for basic living, entertainment, and investments. From a social science perspective, scientists are especially interested in how average income per person is related to other variables representing economic, health, and social well-being. In this study, we will use county (a governmental unit in the United States that is larger than a city but smaller than a state) as a basic unit of calculating per-capita income, together with other variables of the county to address various social science questions by data analysis. The four questions below are our main focuses:

1. Are there any variables related to each other?
2. Is per-capita income related to crime rate, and is the relationship influenced by region?
3. What is the best model to predict per-capita income?
4. Is there any problem caused by missing states?

The first two questions explore pairwise relationships between different social factors. These can help us better address question three, which is the most important one for social scientists. That is, how can all the listed factors in our data work together to influence a county's per-capita income? The last question serves as a diagnostic of our whole analysis. We will analyze and try to solve these problems in the following sections.

Data

The data set we will use is from Geospatial and Statistical Data Center, University of Virginia (Kutner et al. 2005). This data set includes some demographic information (CDI) from the 440 counties with the largest population in the United States in the year 1990. Counties with missing data were deleted from the data

set. There are 17 columns in this data, including three basic identification of the county, and 14 variables to consider. Here are the definitions of each column:

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)

We first investigate the response variable: per-capita income.

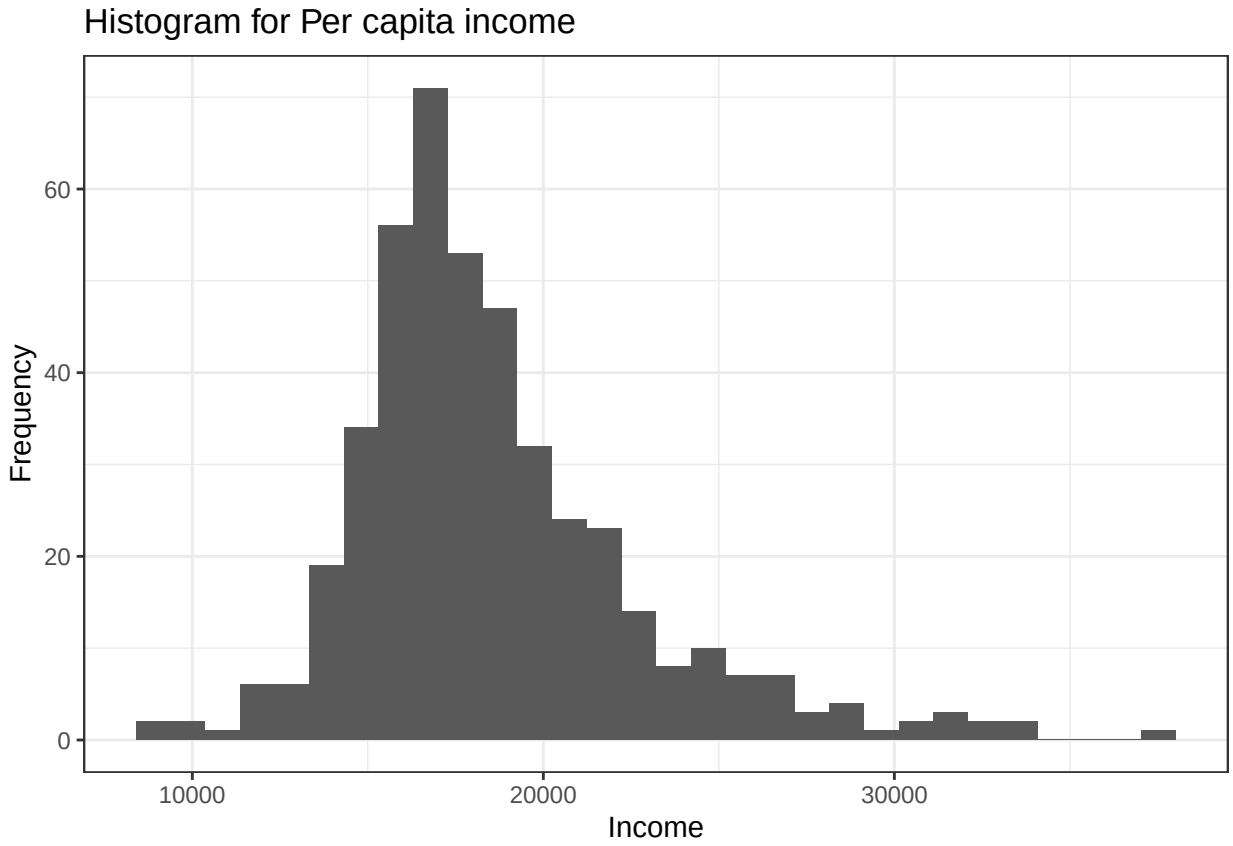


Figure 1: Histogram for Per capita income

As is shown in Figure 1, the distribution of per-capita income in different counties is right skewed and peaking at around 17,000 dollars. Most counties have per capital income within the range of 14,000 and 23,000 dollars. There is one county which has a very high per-capita income.

Here is a basic summary for the numerical variables. (Table 1)

Table 1: Summary table: Continuous variables

Variable	Min	First.Qu	Median	Mean	SD	Third.Qu	Max
land.area	15.0	451.25	656.50	1041.41	1549.92	946.75	20062.0
pop	100043.0	139027.25	217280.50	393010.92	601987.02	436064.50	8863164.0
pop.18_34	16.4	26.20	28.10	28.57	4.19	30.02	49.7
pop.65_plus	3.0	9.88	11.75	12.17	3.99	13.62	33.8
doctors	39.0	182.75	401.00	988.00	1789.75	1036.00	23677.0
hosp.beds	92.0	390.75	755.00	1458.63	2289.13	1575.75	27700.0
crimes	563.0	6219.50	11820.50	27111.62	58237.51	26279.50	688936.0
pct.hs.grad	46.6	73.88	77.70	77.56	7.02	82.40	92.9
pct.bach.deg	8.1	15.28	19.70	21.08	7.65	25.33	52.3
pct.below.pov	1.4	5.30	7.90	8.72	4.66	10.90	36.3
pct.unemp	2.2	5.10	6.20	6.60	2.34	7.50	21.3
per.cap.income	8899.0	16118.25	17759.00	18561.48	4059.19	20270.00	37541.0
tot.income	1141.0	2311.00	3857.00	7869.27	12884.32	8654.25	184230.0

Per-capita income has a mean of 18,561.48 and standard deviation of 4,059.19. The minimum is 8,899 and the maximum is 37,541, which represents a huge difference.

Other interesting findings from the summary table above (Table 1) includes the statistics for crimes. The minimum number of crimes is only 563, and the maximum is 688,936, about 1,224 times higher than the minimum. The maximum unemployment rate is 21.3%, which means that over one fifth of the population do not have a job. The medical condition of the county also varies a lot. The standard deviation of number of active physicians and number of hospital beds are 1,789.75 and 2289.13 respectively.

Here is another summary for the categorical variable: region. (Table 2)

Table 2: Summary table: Geographic region

Region	Count
NC	108
NE	103
S	152
W	77

There are highest number of counties in the southern region, and the number is the lowest in western region.

Methods

To address the questions listed in the introduction section, I did the following analysis of the data.

Question 1

Are there any variables related to each other?

To answer the first question, I did some graphing (boxplots, correlation plot, scatterplots) using R to analyze the pairwise relationships between some variables. The variable per-capita income was especially emphasized in this part because this is the variable of interest. Confounding factor (factor correlated to two or more variables, making these variables correlated) was detected and analyzed. (Appendix (a), Page 15)

Question 2

Is per-capita income related to crime rate, and is the relationship influenced by region?

For the second question, a model was first built for predicting per-capita income using region and crime rate. Another model including the income grouped by regions was also included for comparison. An ANOVA (Analysis of Variance) table and BIC (Bayesian Information Criterion) were used to assess the necessity for the added interaction. We picked the better model of the two choices and changed the factor crime rate into the total number of crimes. Finally, we compared the changed model to the original one for the best interpretation of the relationship between per-capita income and crimes and region. Statisical and social scientific interpretation were both took into considertation. (Appendix (b), Page 19)

Question 3

What is the best model to predict per-capita income?

For the third question, our goal was to find the best model within the given variables to predict the per-capita income of the counties. Here are the steps towards it.

First, we changed all the factors influenced by confounding variable population (mentioned in Question 1) by dividing them by population. The purpose was to eliminate the collinearity problem brought by confounding variable. This can be a better approach than simply deleting variables, because we are interested in these factors. (Appendix (c) Confounding variables, Page 21)

Second, we explored some transformations for each variable to make them follow a normal distribution using histogram and Box-Cox methods. The reason for that is to satisfy the normality assumption of linear regression. (Appendix (c) Transformations, Page 22)

Third, we did some variable selection. We first deleted the total income variable, which is the one that can calculate per-capita income directly when divided by population. Then, VIF (Variance Inflation Factor) was used to check for multicollinearity conditions. We expect all the variables to be independent of each other to satisfy the multiple linear regression assumptions. Scatterplots were used to detect correlations for factors with high VIF values (over the benchmark VIF of 5). BIC (Bayesian Information Criterion) was also used after deleting variables for collinearity, to develop a simple model. (Appendix (c) Variable selection, Page 25)

Fourth, we checked for any possible improvement when grouping one of the variables in the model by region. ANOVA (Analysis of Variance) table were used in this process and we extracted all the p-values of the F-tests. The null hypothesis of the test is that the interaction between the region and the variable does not help improve the model. When the p-value is below the threshold of 0.05, we can reject that null hypothesis and include the interaction term. We compared the cases of single interaction term and multiple ones using BIC (Bayesian Information Criterion) values. The model with the lowest BIC value was chosen. (Appendix (c) Interaction, Page 26)

Finally, we got the optimal model, and did some model diagnostics. (Appendix (c) Final model discussion, Page 27)

Question 4

Is there any problem caused by missing states?

To answer the fourth question, we basically examined two type of problems.

The first one comes from the extreme values of per-capita income in the missing data. We looked over the per-capita income of the three missing states. For counties, we examined 10 counties with highest per-capita income and another 10 counties with lowest per-capita income to see whether our data included all these counties with extreme values.

The second problem is the common feature of the missing data. We looked into that feature and see whether it can affect our result.

Results

Now we can look at the results of the above analysis.

Question 1 (See Appendix (a), Page 15, for details)

Are there any variables related to each other?

We first consider the effect of the only categorical variable region to per-capita income:

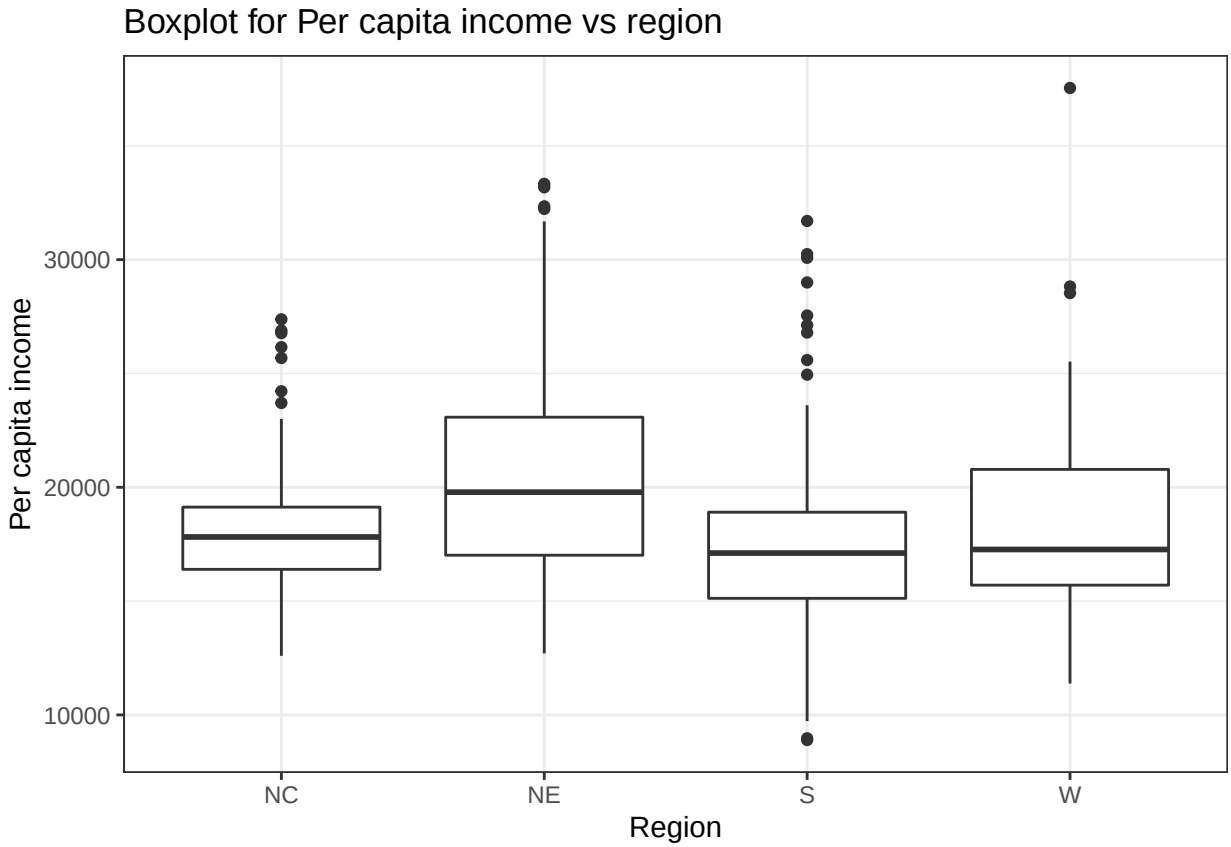


Figure 2: Boxplot for Per capita income vs region

From the boxplot above (Figure 2), we can see that the median per-capita income is the highest in northeast region, and the lowest in southern region. There are some outliers for the west region with very high values.

We then explore any relationship between numerical variables. Here is a correlation plot for all those variables (Figure 3). We will look into the variables with highly positive (red) or negative (blue) relationships. Any relationship between the factors we are interested in will also be included.

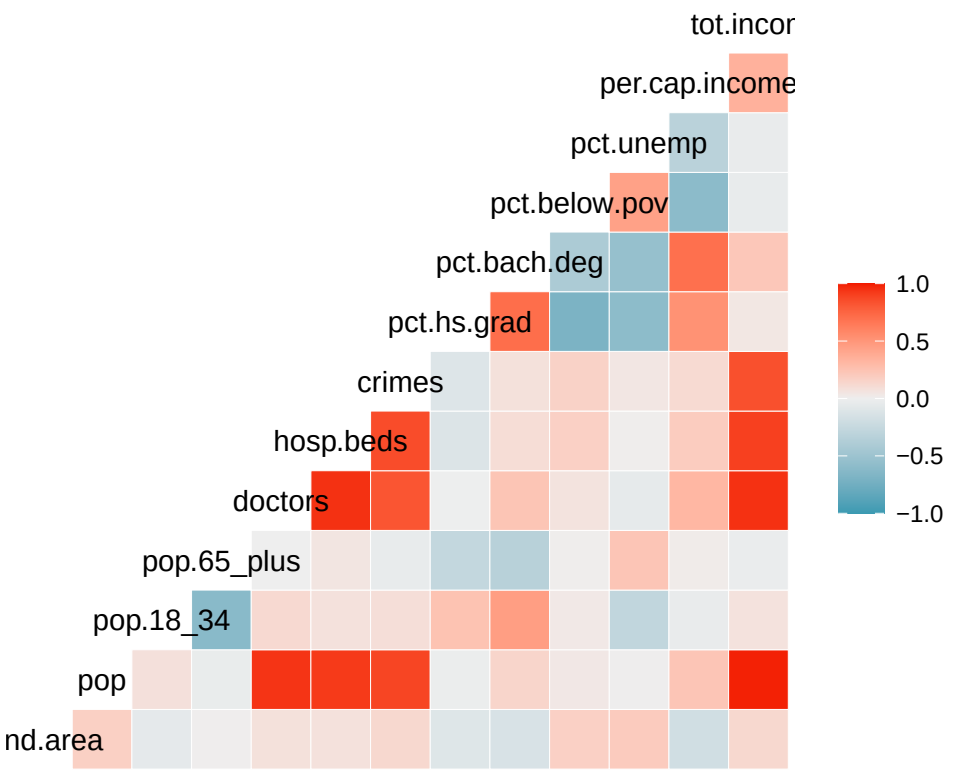


Figure 3: Correlation plot of numerical variables

We first look at the factors related to per-capita income. There are some scatterplots for selected factors. (Either high in correlation, or interesting)

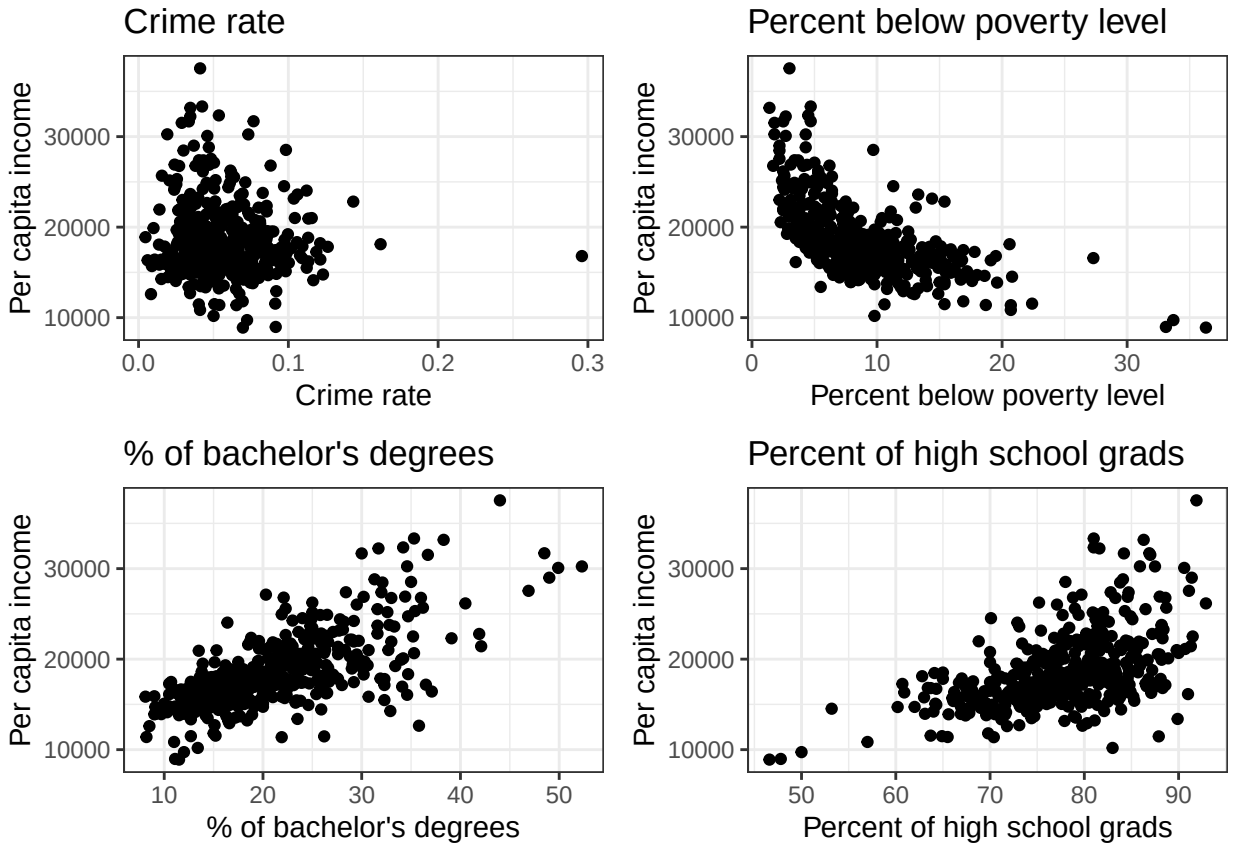


Figure 4: Per capita income vs Crime rate, Percent below poverty level, Percent of bachelor's degrees, Percent of high school grads

By Figure 4, we can see that the counties having high crime rate usually have medium per capital income (around 20000). There is a negative relationship between per capital income and percent below poverty level. Also, we can see increasing trend in the graphs for percent of bachelor's degrees and percent of high school grads. For the latter, there also exists an increasing variance of per-capita income with the growing percent of high school grads. No relationship can be observed between per-capita income and percent of elderly population, and number of active physicians. Given that the data for both of the factors are right-skewed, maybe we can so some transformation for further study. (We will include that in the third question)

When we look at other relationships between variables, there is an interesting finding. From Figure 2, we can see a significant positive relationship between the number of active physicians and the number of hospital beds. There is also a positive relationship between number of active physicians and total serious crimes. This result brought us to think about whether there exists any confounding variables.

We then detected that the factor population is highly correlated with four variables: Number of active physicians, Number of hospital beds, Total serious crimes, Total personal income. The five variables including population are also pairwise correlated. Given that the four variables are all some kind of "total" values of a certain population, we conclude that population is a confounding factor.

Question 2 (See Appendix (b), Page 19, for details)

Is per-capita income related to crime rate, and is the relationship influenced by region?

We first built two models, one predicts per-capita income with crime rate and region (Model 1), and the other added the interaction of the two variables into the model (Model 2). We tested the null hypothesis that we prefer Model 1 to Model 2. The p-value is $0.99 > 0.05$, so Model 1 makes more sense. We should not include the interaction between region and crime rate. Below are the summary tables for both Model 1

and Model 2. The tables include the model estimates for the coefficients, the standard error, t statistics and p-values for the coefficient significance tests (To test whether the coefficient can be 0, which means excluding the corresponding factor).

Table 3: Coefficients for Model 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18006.04469	537.0395	33.5283439	0.0000000
regionNE	2354.69663	541.9715	4.3446875	0.0000174
regionS	-927.44668	512.3059	-1.8103378	0.0709333
regionW	-34.92294	586.0281	-0.0595926	0.9525075
crime.rate	5773.20230	7520.4126	0.7676710	0.4430992

Table 4: Coefficients for Model 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18077.2939	895.2077	20.1934068	0.0000000
regionNE	2329.0367	1101.3942	2.1146260	0.0350340
regionS	-1010.3526	1323.8024	-0.7632201	0.4457487
regionW	-669.9909	1983.8920	-0.3377154	0.7357417
crime.rate	4379.0699	15893.5069	0.2755257	0.7830441
regionNE:crime.rate	288.3868	20184.6607	0.0142874	0.9886073
regionS:crime.rate	1558.9186	20556.1122	0.0758372	0.9395837
regionW:crime.rate	10655.5422	32322.4079	0.3296642	0.7418135

From Table 3, we see that there is a positive relationship between total serious crimes and per capital income, because the coefficient for crime rate in the model is positive. However, this relationship is not significant, because the p-value for it is $0.44 > 0.05$.

Here is a summary for the model with total crimes, instead of crime rate. (Model 3)

Table 5: Coefficients for Model 3

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18106.9099510	378.4380302	47.8464333	0.0000000
regionNE	2286.0373120	532.4709260	4.2932622	0.0000217
regionS	-860.5567325	486.8305023	-1.7676722	0.0778167
regionW	-142.8266313	579.6196624	-0.2464144	0.8054777
crimes	0.0089153	0.0031877	2.7968322	0.0053895

The coefficient is now strictly above 0, indicating a positive relationship between total crimes and per-capita income. (See Appendix (b) for details) (See Discussion section for final choice of predictor)

Question 3

What is the best model to predict per-capita income?

Confounding Variables (See Appendix (c) Confounding Variables, Page 21, for details)

We made changes to three variables: number of active physicians, number of hospital beds, and total serious crimes. All of them were divided by population, so we got three new variables: percentage of active physicians,

average hospital beds per person, and crime rate.

Transformations (See Appendix (c) Transformations, Page 22, for details)

Here are the transformations we used for modeling:

1. Log transform $\log(X)$ for: Per capita income, Land area, Percent of population aged 18–34, Percent of population 65 or older, Percentage of active physicians, Average hospital beds per person, Crime rate, Percent bachelor's degrees, Percent below poverty level, Percent unemployment
2. One over square root $X^{-\frac{1}{2}}$ for: Total population, Total personal income
3. Cube X^3 for: Percent high school graduates

The transformation suggests an inverse change of relationship (i.e. a positive relationship will switch to a negative one) for only total population and total personal income.

Variable selection (See Appendix (c) Variable selection, Page 25, for details)

We first deleted the variable total income of the county, because this is actually another form of evaluating income. In our data set, there is another factor called population. Our response variable per-capita income is calculated by: per-capita income = total income / population. Including both total income and population will resulting a meaningless model and exclude all the other variables of interest. Thus, we chose to delete total income

For multicollinearity condition, using VIF table and scatterplots, we decided to delete Percent bachelor's degrees. This variable is correlated with percent below poverty, unemployment rate and percent of high school graduates.

Using BIC, we finally chose the model with seven variables: land area, population, percent of population aged 18-34, population below poverty level, unemployment rate, percent of active physicians, and crime rate. This model can explain 79.19% of the total variability of per-capita income, which is acceptable for such a simple model.

Interaction (See Appendix (c) Interaction, Page 26, for details)

In our exploration of interaction, we found adding an interaction term to unemployment rate and percent of active physicians can improve the model significantly. We then examined three possible models under this observation:

Model 1: Only include region interaction with unemployment rate
Model 2: Only include region interaction with percentage of active physicians
Model 3: Including interaction with both of them

The BIC (Bayesian information criterion) value for Model 2 is the lowest, so we chose Model 2 as our final model.

Final model (See Appendix (c) Final model discussion, Page 27, for details)

Here is the summary table for our final model. (Table 6) The table includes the model estimates for the coefficients, the standard error, t statistics and p-values for the coefficient significance tests (To test whether the coefficient can be 0, which means excluding the corresponding factor).

Table 6: Final Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.7958118	0.1660919	71.019802	0.0000000
land.area	-0.0254894	0.0064618	-3.944633	0.0000934
pop	-55.2253736	8.0467534	-6.863063	0.0000000

	Estimate	Std. Error	t value	Pr(>|t|)
pop.18_34	-0.1447567	0.0345886	-4.185094	0.0000346
pct.below.pov	-0.2875406	0.0117311	-24.510922	0.0000000
pct.unemp	0.0367914	0.0186342	1.974400	0.0489818
pct.doctors	0.0914359	0.0154388	5.922482	0.0000000
regionNE	0.3959113	0.1532993	2.582603	0.0101384
regionS	0.4388434	0.1237713	3.545599	0.0004352
regionW	0.6894023	0.1815267	3.797801	0.0001672
crime.rate	0.0442536	0.0132308	3.344732	0.0008965
pct.doctors:regionNE	0.0613486	0.0241245	2.543001	0.0113432
pct.doctors:regionS	0.0711746	0.0194159	3.665781	0.0002777
pct.doctors:regionW	0.1053775	0.0286868	3.673383	0.0002698

We can see that all the p-values are below 0.05, which means that we have statistical evidence that all the factors are correlated to per-capita income. The relationships are:

- Positive for: total population, unemployment rate, percentage of active physicians, region (northeastern, southern, western), crime rate, percentage of active physicians in northeastern, southern, and western regions.
- Negative for: percent of population aged 18-34, percent below poverty.

The relationships are highly significant for total population, percent below poverty and percentage of active physicians. In contrast, this relationship is the weakest for unemployment rate.

According to the model diagnostics done in Technical Appendix Page 28, this model can be appropriate for prediction.

Question 4

Is there any problem caused by missing states?

We first look at the extreme value problem.

The missing states are: Alaska, Iowa, Wyoming. The state per-capita income for all the three states are between $31,557 and $38,915 (United States Census Bureau, 2019), which is not extremely high or low. We do not expect any values from these states to influence our model.

We then examined the county rankings of per-capita income and examined whether the top 10 and bottom 10 counties are in our data set. (Wikipedia, 2021) The top 10 counties are: Marin, New York, City of Falls Church, Pitkin, Fairfield, Teton, Somerset, Arlington, City of Alexandria, Morris. The Bottom 10 counties are: Wake, Sussex, Union, Denton, Newport, City and Borough of Juneau, Alameda, Rockingham, Suffolk, Middlesex. We found that there are several missing high income counties: New York, City of Falls Church, Pitkin, Teton. And also, one missing low income county: City and Borough of Juneau. These missing data may cause bias in our study.

The population for the three missing states are 626,932 (Alaska), 2,926,324 (Iowa), and 493,782 (Wyoming) (The States of the USA on a Map, 2021). The values are relatively small, especially for Alaska and Wyoming. According to the data background, the reason of not including a county is either including missing values, or small in population (majority). So relatively small population can be a common factor in all the missing data, including missing counties and missing states.

The figure below (Figure 5) shows the relationship between per-capita income and population. The slope for this graph is negative. Considering the effect of our transformation mentioned in Question 3 Transformation section, we can conclude that population and per-capita income has a slightly positive relationship. The per-capita income can be higher in more populous counties. Consequently, the missing counties with relatively small population can result a bias in our analysis.

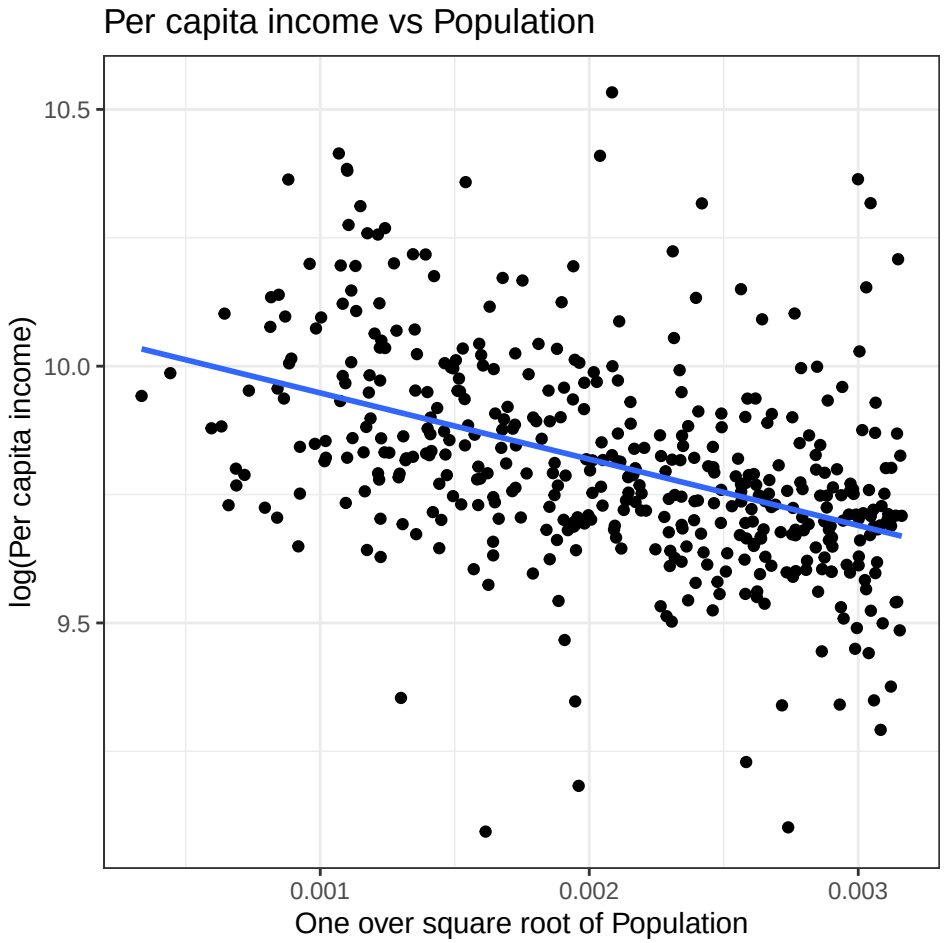


Figure 5: Total population vs per-capita income

Discussion

In this section, we will dig into the results we got from the data and discuss the research questions in the introduction.

Question 1

Are there any variables related to each other?

From the graphs related to per-capita income, there are four main findings:

1. The median per-capita income is the highest in northeast region, and the lowest in southern region. This is consistent to the reality that the density of big cities is the highest in northeast region and the lowest in the southern region. The rich zones in the western region resulted in the outliers.
2. Crime rates are usually higher in areas with medium per-capita income. This can be an interesting result that the crime rate does not significantly decrease with higher income. In reality, almost everyone is having low income in counties with very low per-capita income. One will not get much money from others even if they commit crimes like robbery. In contrast, the income disparity can be larger in middle-income regions, which encourages the criminals to commit crimes.
3. The negative relationship between per-capita income and percent below poverty level is within expectation, because the poverty means low income.

4. Education is an important factor in determining per-capita income. High school education is a basis of higher per-capita income. It will not definitely result in an increase in income, but a low percent of high school grads will certainly result in a low income. In contrast, percent of adults with a bachelor degree can be a good indicator of per-capita income. The higher the percentage, the higher the per-capita income will be.

We detected in Result section that the number of active physicians and the number of hospital beds are significantly correlated. This is possible because both of them can be treated as indicators of a county's medical resource. However, the number of active physicians is surprisingly correlated with total crimes. This cannot be explained by any real life situations, because increasing number of crimes are not likely to make more people become physicians. Thus, we switched to explore some factors that acts as a "medium" to stick these unrelated factors together (i.e. confounding factor). From the red blocks (highly positive correlation) related to population in the correlation graph (Figure 3), we convinced that population is related to all the variables with aggregation: number of active physicians, number of hospital beds, total serious crimes, and total personal income.

Consequently, the answer of the question is definitely yes, but some of the relationships are influenced by confounding factors.

Question 2

Is per-capita income related to crime rate, and is the relationship influenced by region?

From the result of the model summary, we can see a slightly positive relationship between per-capita income and crime rate. However, we cannot be sure that the relationship is significant and correct. Thus, we conclude that per-capita income is not necessarily related to crime rate.

When we group crime rate with region, we can see positive relationship between per-capita income and crime rate for all the four regions, indicating that the relationship is not influenced by region. Our test also show that grouping crime rate with regions is unnecessary.

Furthermore, we replaced crime rate with total crimes to see the output. We can infer from the model that total crimes has a significantly positive relationship with per-capita income.

Statisticians may think that the significance of coefficient is important, because it at least confirm a certain relationship. However, in social science, total crime itself is not a property that can measure the stability of a certain region. Actually, more populous counties will certainly have more cases of serious crimes. In contrast, crime rate can serve as a social stability factor, and is more meaningful in social science perspective. Consequently, we still choose crime rate in our analysis.

Question 3

What is the best model to predict per-capita income?

The best model contains nine factors: land area, population, percent of population aged 18-34, percent below poverty level, unemployment rate, percentage of active physicians, region, crime rate, and percentage of active physicians influenced by region.

From the optimal model, there are some findings from the highly significant factors:

1. Land area is negatively correlated with per-capita income. The larger the county, the lower per-capita income will be. It is interesting with the fact that people with highest income are usually in large cities, which are usually small in land area but dense in capital.
2. Population is positively related to per-capita income. This is an understandable result, because the average income in populous large cities are usually higher than rural areas with small population.
3. Proportion of people with age 18 to 34 is negatively correlated with per-capita income. This can be surprising that the population with many young adults usually have less income. There can be several possible reasons for this fact. The first one is that young adults are usually new in the industry and may

still be students without income. The second one is that high income families are less willing to give birth to children, causing aging societies in rich zones. Both of them need further evidence to support.

4. Per-capita income is negatively related to percent below poverty. This is a straightforward result, since poverty means low income.

5. An increase in percentage of active physicians is related to an increase in per-capita income. This is understandable, since high-income area usually have more medical resources. Also, physicians usually have high income, which can contribute to the higher per-capita income of the county. This increasing effect is more obvious in southern and western region, which means that medical strength is more important to income in these regions. The reason is still unclear at this point.

6. Per-capita income in western region is significantly higher than other regions. Many big and high-tech companies have there main sites in western region. Employees and owners of these companies occupy a large percentage of working population there. They usually have higher income than average, which is largely related to the high per-capita income.

Question 4

Is there any problem caused by missing states?

We have explored two types of bias that can be caused by missing states and counties. The first one is missing extreme values in our variable of interest, per-capita income. Although no extreme value can be found in missing states, some counties with extremely high or low per-capita income are not in our data. We thus conclude that extreme value problem exists. The second bias is from the common properties of missing values, which are different from those in the data. Because we included only 440 most populous counties, this common property can be small population. From the scatterplot of population and per-capita income, we can observe a positive relationship between population and per-capita income. In that case, our study may overestimate the average per-capita income.

Consequently, the answer to Question 4 is yes. Missing value can cause bias.

Weaknesses & Next Step

1. Limited by dataset, we can only consider the data for top 440 counties in population. According to our analysis in Question 4, missing data can cause bias in our study. In the future, we can include all the counties in our data.

2. We used only multiple linear regression in this analysis, which cannot reflect all the possible relationships in reality. In the future, we can try other models such as Generalized Additive Model and Linear Mixed Model.

3. In this analysis, we cannot dig into all the social scientific inferences we made. That is, the reason why the relationship is negative or positive, and why the two factors are correlated with each other. We made some guesses in our discussion, but none of them can be proved by real life researches and data. For the next step, we can do more researches to find certain reasons and make recommendations for social well-beings.

References

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

2. Englisch-hilfen.de. (2021). *The States of the USA on a Map.* [online] Available at: https://www.englisch-hilfen.de/en/texte/states.htm [Accessed 18 October 2021].

3. United States Census Bureau. (2019). *Per capita income in past 12 months (in 2019 dollars), 2015-2019.* [online] Available at: https://www.census.gov/quickfacts/fact/map/US/INC910219 [Accessed 18 October 2021].

4. En.wikipedia.org. (2021). *List of highest-income counties in the United States - Wikipedia.* [online] Available at: https://en.wikipedia.org/wiki/List_of_highest-income_counties_in_the_United_States [Accessed 30 October 2021].

Technical Appendix

(a)

Summary statistics

Categorial variables

Table 7: Summary table: Geographic region

Region	Count
NC	108
NE	103
S	152
W	77

Continuous variables

Table 8: Summary table: Continuous variables

Variable	Min	First.Qu	Median	Mean	Third.Qu	Max
land.area	15.0	451.250	656.50	1.041411e+03	946.750	20062.0
pop	100043.0	139027.250	217280.50	3.930109e+05	436064.500	8863164.0
pop.18_34	16.4	26.200	28.10	2.856841e+01	30.025	49.7
pop.65_plus	3.0	9.875	11.75	1.216977e+01	13.625	33.8
doctors	39.0	182.750	401.00	9.879977e+02	1036.000	23677.0
hosp.beds	92.0	390.750	755.00	1.458627e+03	1575.750	27700.0
crimes	563.0	6219.500	11820.50	2.711162e+04	26279.500	688936.0
pct.hs.grad	46.6	73.875	77.70	7.756068e+01	82.400	92.9
pct.bach.deg	8.1	15.275	19.70	2.108114e+01	25.325	52.3
pct.below.pov	1.4	5.300	7.90	8.720682e+00	10.900	36.3
pct.unemp	2.2	5.100	6.20	6.596591e+00	7.500	21.3
per.cap.income	8899.0	16118.250	17759.00	1.856148e+04	20270.000	37541.0
tot.income	1141.0	2311.000	3857.00	7.869273e+03	8654.250	184230.0

Missing values

There is no missing data in this dataset.

Variable features EDA

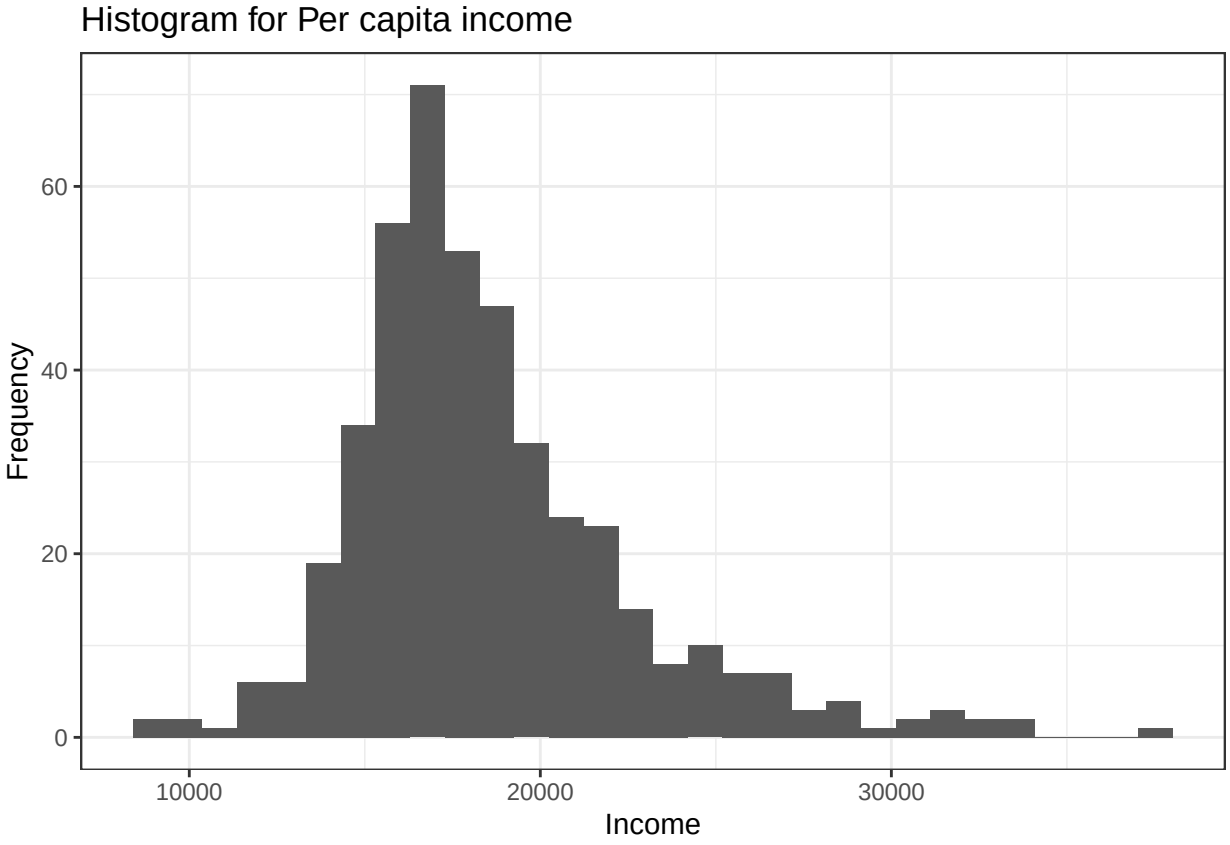


Figure 6: Histogram for Per capita income

As is shown in Figure 6, the distribution of per capital income in different counties is right skewed and peaking around 17000 dollars. Most counties have per capital income within the range of 14000 and 23000 dollars. There is one county which has a very high per capital income.

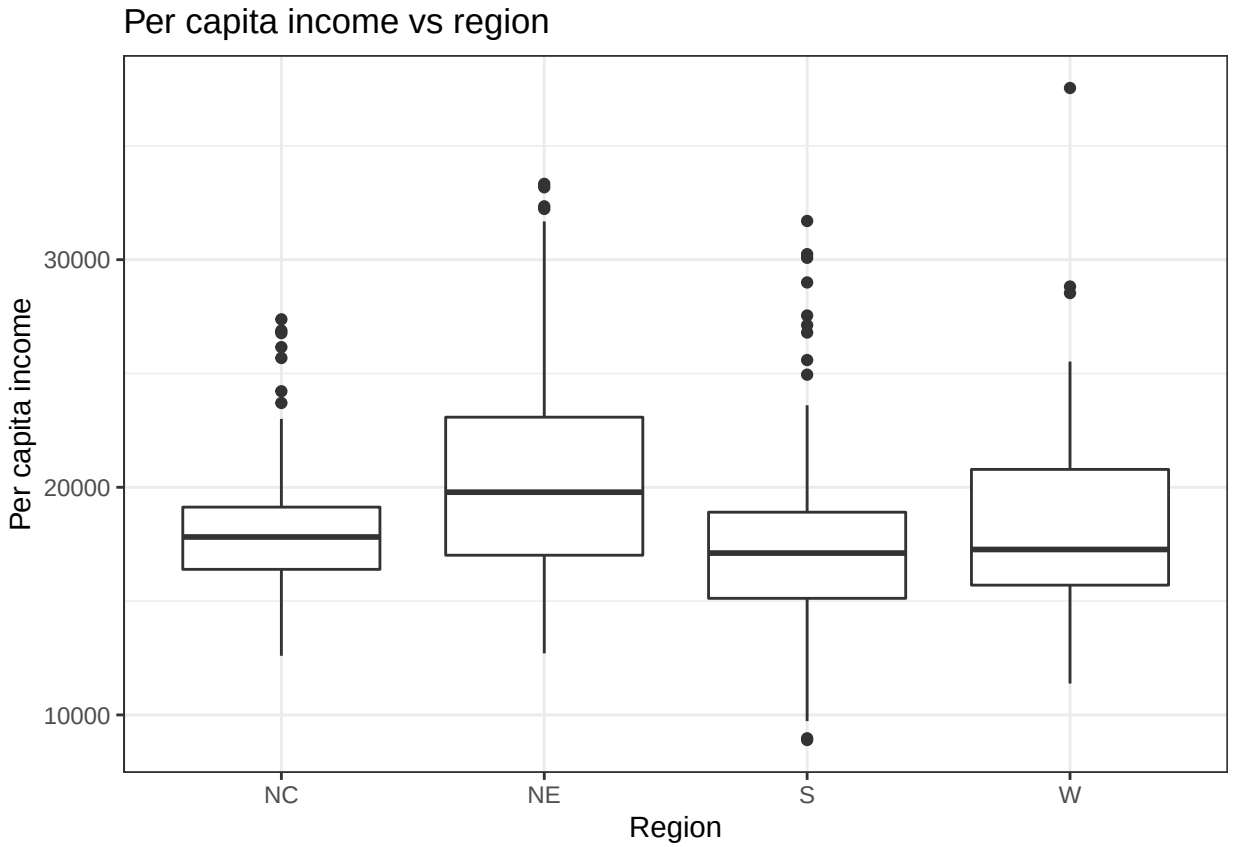


Figure 7: Boxplot for Per capita income vs region

The median per capital is the highest in northeast region, and the lowest in southern region. This is consistent to the reality that the density of big cities is the highest in northeast region and the lowest in the southern region. Per capital income is usually higher in large cities than other areas. Besides, there are some outliers for the west region with very high values. This is because of there is some rich zone in the western region.

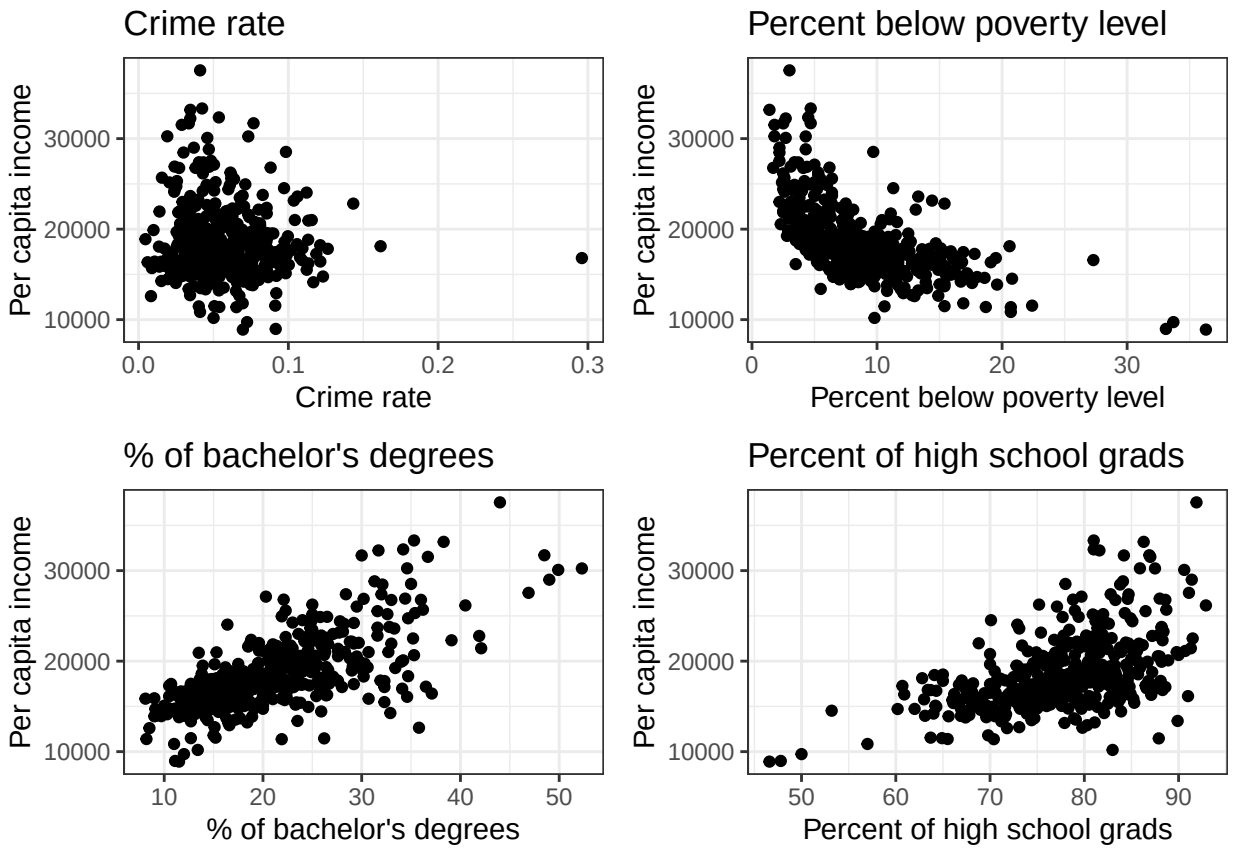


Figure 8: Per capita income vs Crime rate, Percent below poverty level, Percent of bachelor's degrees, Percent of high school grads

By Figure 8:

1. The counties having high crime rate usually have medium per capital income (around 20000).
2. There is a negative relationship between per capital income and percent below poverty level.
3. Per capital income for a county is significantly correlated to the percent of adult population with a bachelor's degree in that county.
4. We can also see an increasing trend in the graph for percent of high school grads, together with an increasing variance with the growing percent of high school grads. We can infer that the higher percent of high school grads will not definitely result in an increase in income, but a low percent of high school grads will certainly result in a low income.

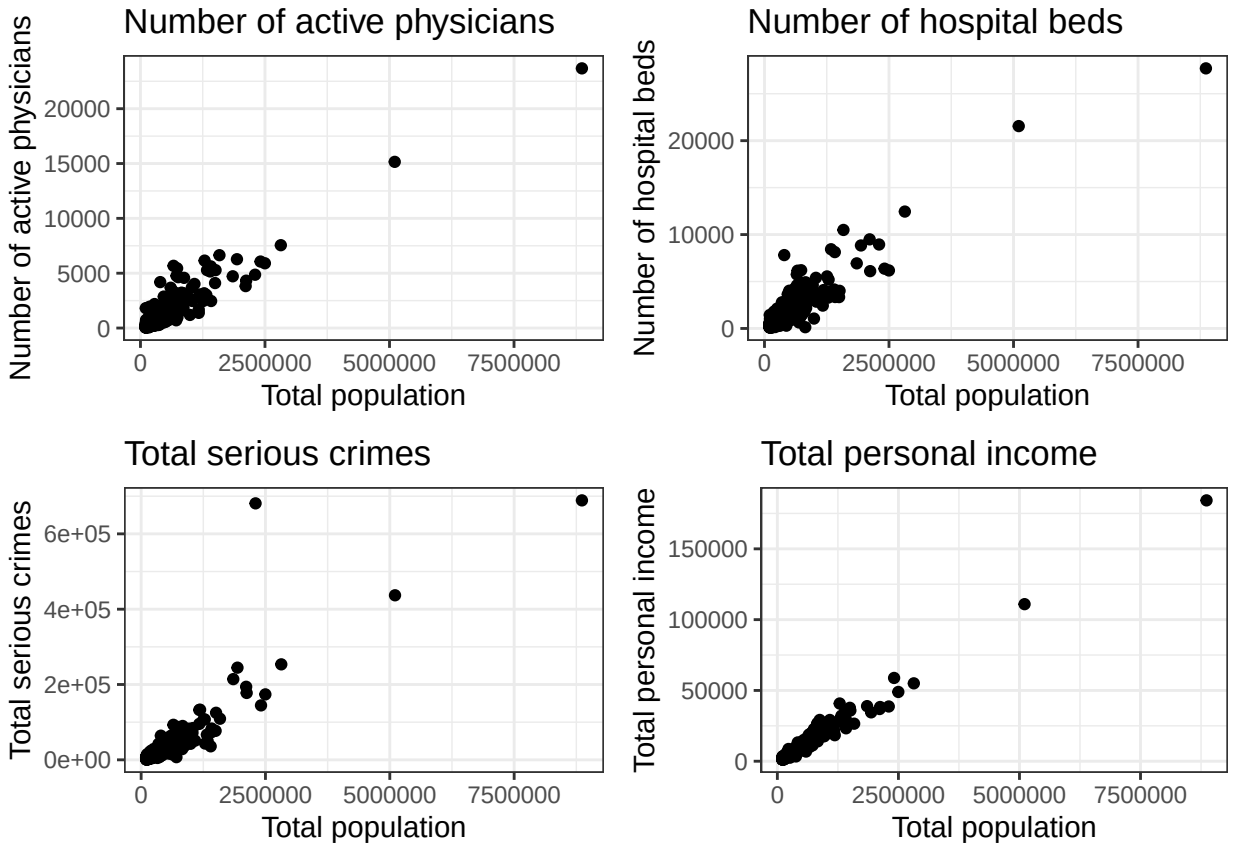


Figure 9: Total population vs Number of active physicians, Number of hospital beds, Total serious crimes, Total personal income

We can see that the factor population is correlated with all the four variables. So it is a confounding factor to make these variable correlated with each other.

(b)

Model 1 predicts per-capita income using region and crime rate. Model 2 added the interaction term of region and crime.

We will first compare the models with and without the interaction terms. We use a partial F test to compare the models. The null hypothesis is that the coefficient of the interaction terms is zero.

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ region + crime.rate
## Model 2: per.cap.income ~ region * crime.rate
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    435 6609753963
## 2    432 6607856753  3   1897210 0.0413 0.9888
```

Table 9: BIC comparison of models

Model	BIC
Model 1	8556.203
Model 2	8574.337

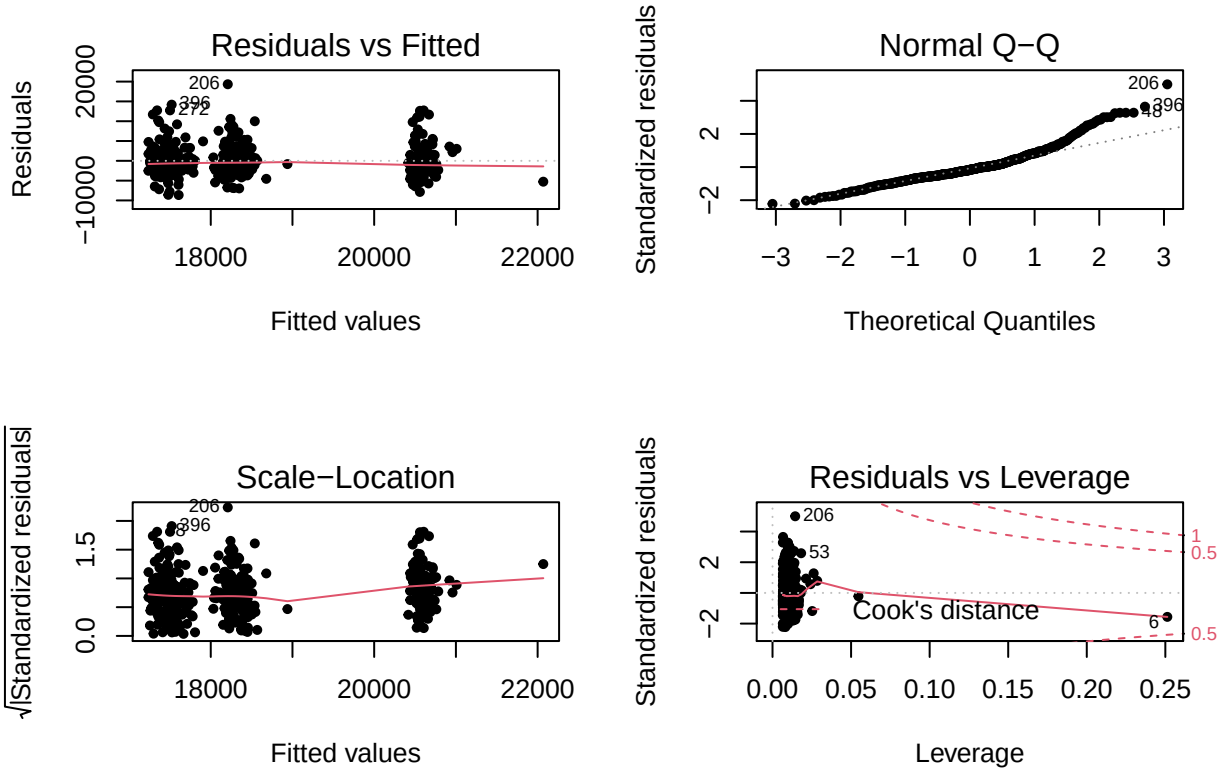
By the above ANOVA table, we can see that the p-value for the partial F test is $0.98 > 0.05$. So we fail to reject the null hypothesis and confirm that there should not be any interaction terms in the model. Besides, the BIC value is also lower for model 1.

The final model is:

$$PerCapitalIncome = \beta_0 + \beta_1 Region + \beta_2 Crimes + \epsilon$$

Table 10: Coefficients for Model 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18006.04469	537.0395	33.5283439	0.0000000
regionNE	2354.69663	541.9715	4.3446875	0.0000174
regionS	-927.44668	512.3059	-1.8103378	0.0709333
regionW	-34.92294	586.0281	-0.0595926	0.9525075
crime.rate	5773.20230	7520.4126	0.7676710	0.4430992



The diagnostic plot above for Model 1 reveals that the only problem for this model is that fitted value is not normal.

By Table 10, there is a positive relationship between total serious crimes and per capital income, because the coefficient for crime rate in the model is positive. However, this relationship is not significant, because the p-value for it is $0.44 > 0.05$.

Now we fit the model with the total number of crimes. (Model 3)

Table 11: Coefficients for Model 3

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18106.9099510	378.4380302	47.8464333	0.0000000
regionNE	2286.0373120	532.4709260	4.2932622	0.0000217
regionS	-860.5567325	486.8305023	-1.7676722	0.0778167
regionW	-142.8266313	579.6196624	-0.2464144	0.8054777
crimes	0.0089153	0.0031877	2.7968322	0.0053895

By Table 11, we can see that the coefficient for number of crimes is also positive, meaning that there is a slightly positive relationship between total crimes and per-capita income. The coefficient is significant (the p-value is less than 0.05), so we can confirm this positive relationship.

Statisticians may think that the significance of coefficient is important, because it at least confirm a certain relationship. Here, we can say that total crime is positively related to per-capita income. However, for social scientists, total crime itself is not a property that can measure the stability of a certain region. Actually, more populous counties will certainly have more cases of serious crimes. In contrast, crime rate can serve as a social stability factor, and is more meaningful in social science perspective. Consequently, we still choose crime rate in our analysis.

(c)

Confounding Variable

In order to get rid of the problem brought by confounding variable population, we decided to divide all the related variables by population. We think all the values got from dividing population is a property inside the society. In our data, crime rate is not related to population, and can serve as a factor measuring the stability of the society. Percent of doctors and average hospital beds per person are measurements of a county's medical strength. In that case, we no longer have to exclude population in our model to avoid collinearity.

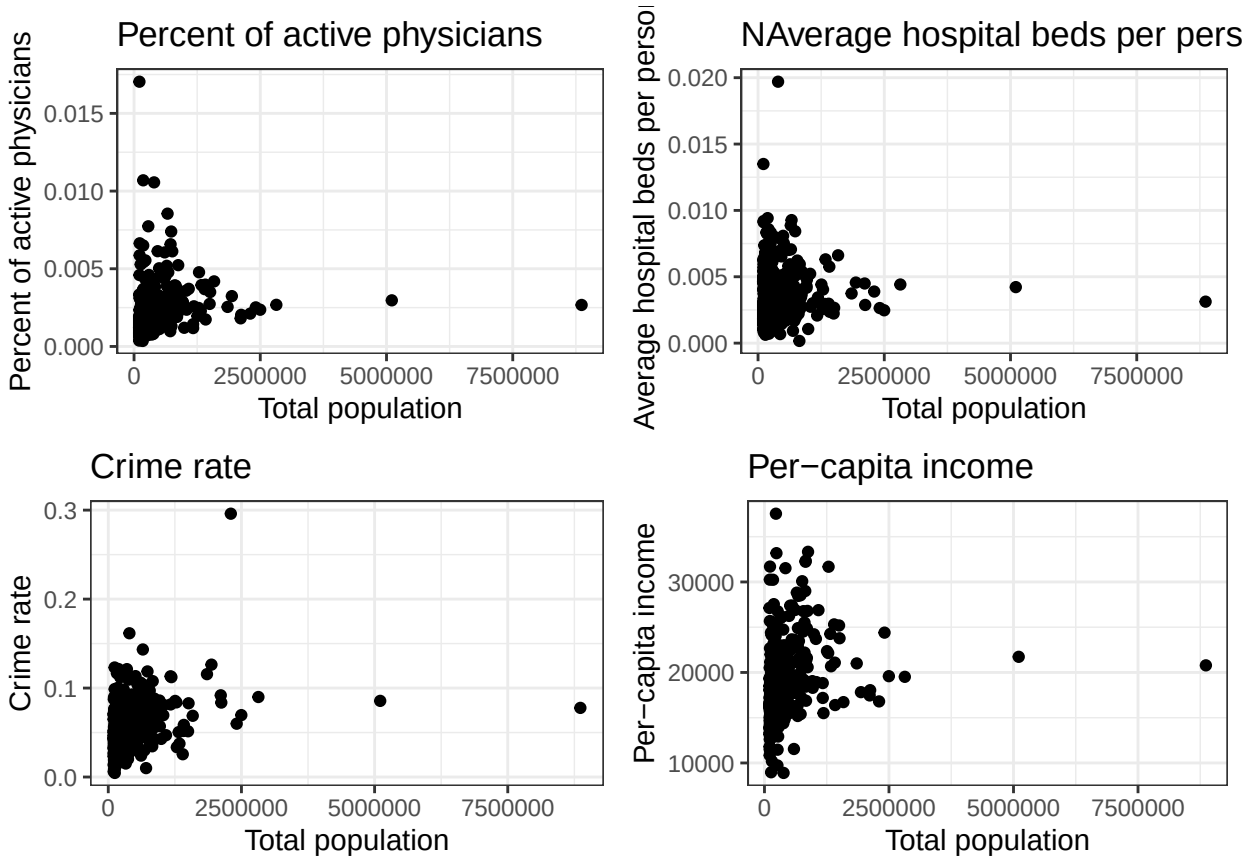


Figure 10: Total population vs Percent of active physicians, Average hospital beds per person, Crime rate, Per-capita income

Now, we can observe no extreme pattern from these scatterplots.

Transformations

In this part, we will explore transformations for the numeric variables to make them look normal.

For the response variable `per.cap.income`, here is its histogram.

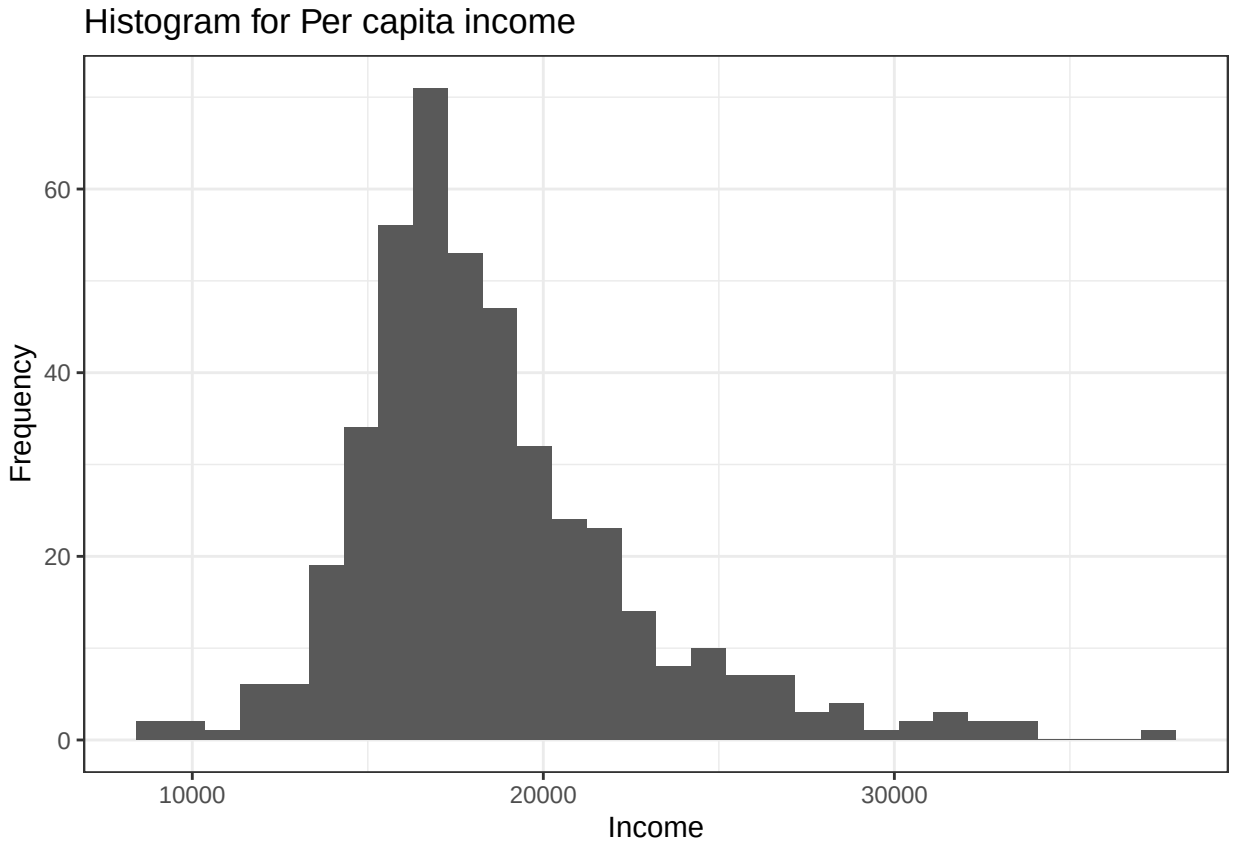


Figure 11: Histogram for Per capital income

We can see that it is right-skewed, so we need a power transform X^λ with $\lambda < 1$. From the Box-Cox method for power transform below, we pick the nearby value $\lambda = 0$. This means we choose the log transform.

```
## [1] "Power Transform Lambda: -0.368336534678286"
```

For the other variables, here is a histogram for all the variables.

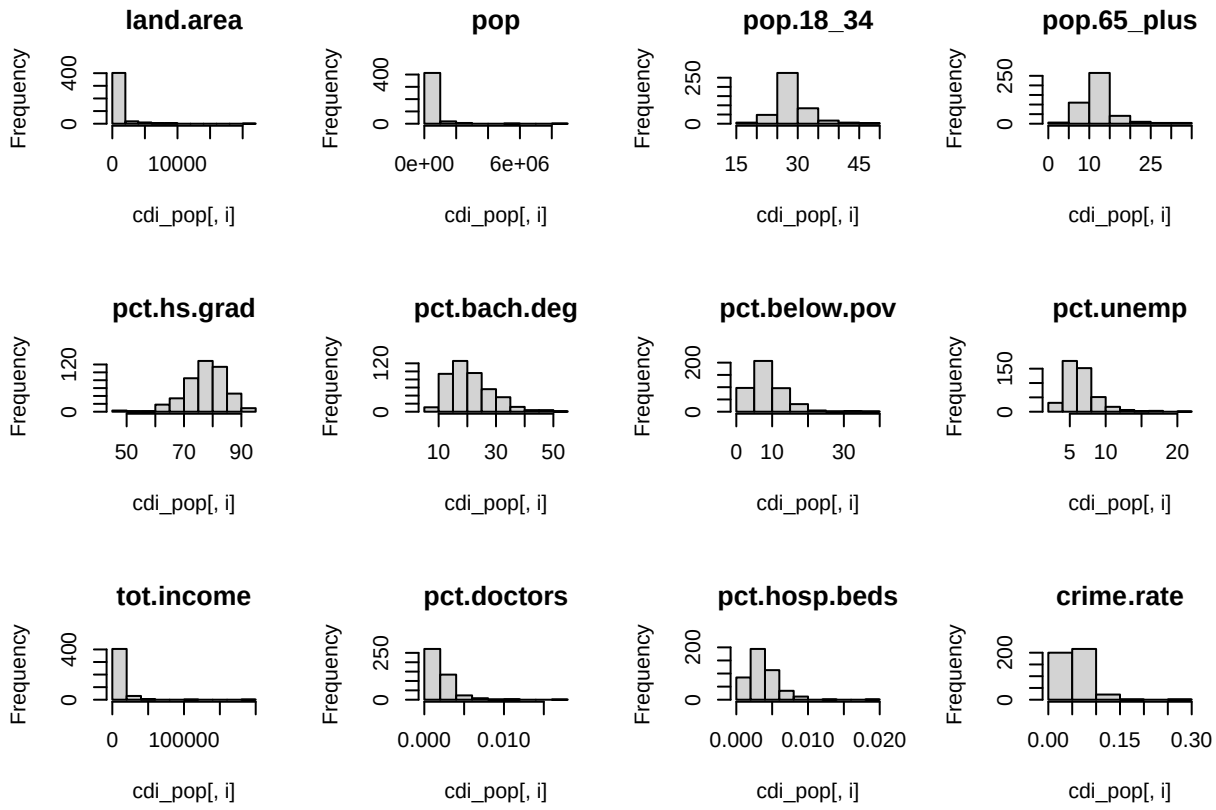


Figure 12: Histograms for variables

We can see from Figure 12 that percent of high school graduates is left-skewed, and all the other variables are right-skewed. We will use Box-Cox method to find proper power transformations.

Table 12: Box-Cox Power transforms

Variable	Lambda
land.area	0.0023048
pop	-0.5795787
pop.18_34	-0.3857010
pop.65_plus	-0.0075542
pct.hs.grad	3.0719249
pct.bach.deg	-0.0317479
pct.below.pov	0.1817562
pct.unemp	-0.1130851
tot.income	-0.4379530
pct.doctors	-0.2302523
pct.hosp.beds	0.2349813
crime.rate	0.3776893

In Table 12, we can see that we can use:

1. Log transform $\log(X)$ for: `land.area`, `pop.18_34`, `pop.65_plus`, `pct.doctors`, `pct.hosp.beds`, `crime.rate`, `pct.bach.deg`, `pct.below.pov`, `pct.unemp`
2. One over square root $X^{-\frac{1}{2}}$ for: `pop`, `tot.income`
3. Cube X^3 for: `pct.hs.grad`

Variable selection

Pre-selection We will delete the variable `tot.income`, because total income divided by population (`pop`) is actually the response variable per capital income.

Collinearity In this part, we will check the collinearity condition for all the numeric variables.

Table 13: Coefficients statistics

	Estimate	Std. Error	t value	Pr(>|t|)	VIF
land.area	-0.03	0.01	-5.21	0.00	1.23
pop	-49.51	7.37	-6.72	0.00	1.60
pop.18_34	-0.28	0.05	-6.18	0.00	2.46
pop.65_plus	0.03	0.02	1.18	0.24	2.78
pct.hs.grad	0.00	0.00	-3.39	0.00	3.79
pct.bach.deg	0.27	0.03	10.17	0.00	5.43
pct.below.pov	-0.25	0.01	-17.91	0.00	3.24
pct.unemp	0.08	0.02	4.73	0.00	1.90
pct.doctors	0.07	0.01	4.78	0.00	3.98
pct.hosp.beds	0.01	0.01	1.07	0.29	2.96
crime.rate	0.02	0.01	2.32	0.02	1.76

We set 5 as the benchmark VIF value, and there is only one predictor with high VIF value: `pct.bach.deg`.

We will use some scatterplots to check which variable it is correlated with. There are several candidates: `pct.below.pov`, `pct.unemp`, `pct.hs.grad`

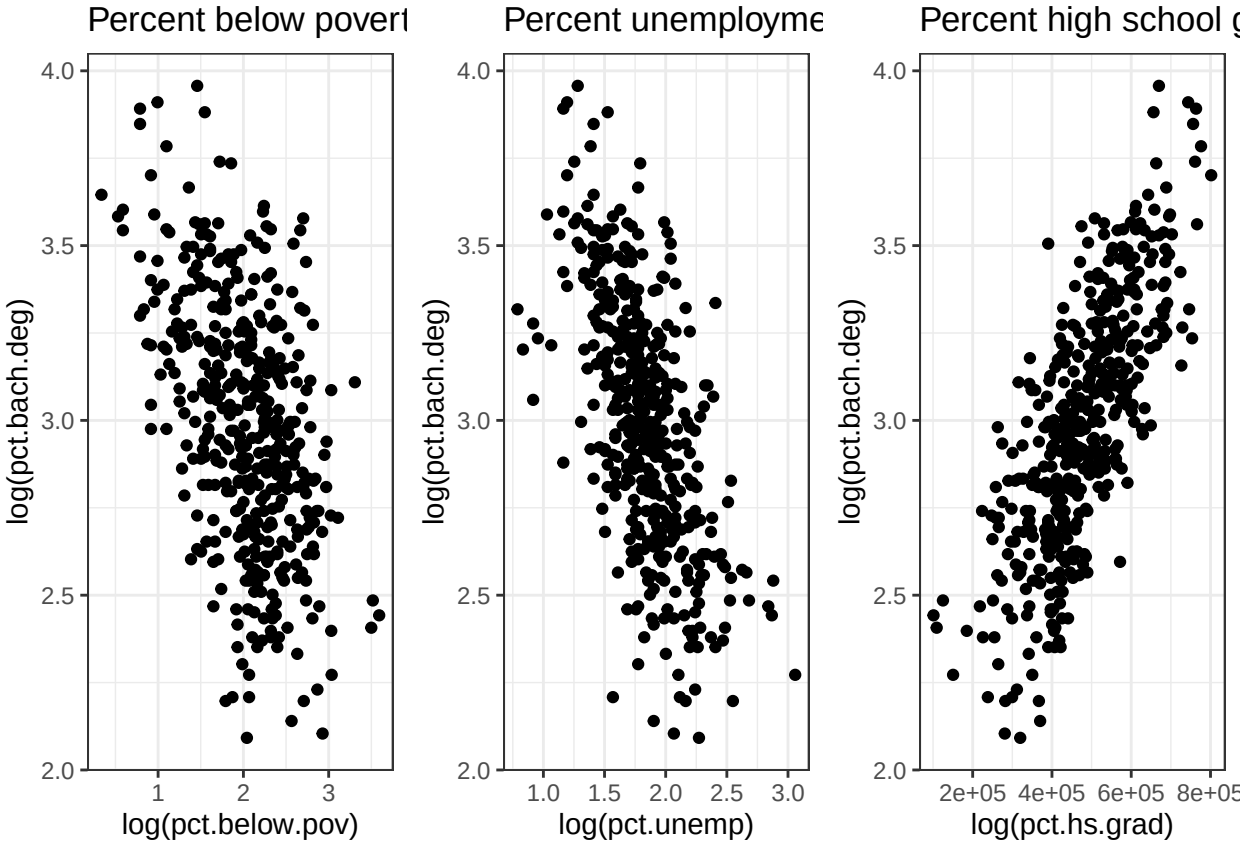


Figure 13: Collinearity check for percent bachelor's degrees.

We can see a correlation with all the three variables, so we choose to delete `pct.bach.deg`

BIC We can use BIC to do model selection. We chose BIC because we would like a simple model

```
##
## Call:
## lm(formula = per.cap.income ~ land.area + pop + pop.18_34 + pct.below.pov +
##     pct.unemp + pct.doctors + crime.rate, data = cdi_adj)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28871 -0.05409 -0.00729  0.05959  0.32874
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    12.09677    0.14890  81.239  < 2e-16 ***
## land.area      -0.02509    0.00558  -4.497 8.86e-06 ***
## pop           -61.79178    8.01821  -7.706 8.96e-14 ***
## pop.18_34      -0.14534    0.03526  -4.122 4.51e-05 ***
## pct.below.pov  -0.29128    0.01152 -25.287  < 2e-16 ***
## pct.unemp       0.04269    0.01740   2.454   0.0145 *
## pct.doctors     0.14148    0.01029  13.747  < 2e-16 ***
## crime.rate      0.03458    0.01181   2.927   0.0036 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09506 on 432 degrees of freedom
## Multiple R-squared:  0.7919, Adjusted R-squared:  0.7886
## F-statistic: 234.9 on 7 and 432 DF,  p-value: < 2.2e-16
```

From the summary of BIC result above, we finally chose the model with seven variables: `land.area`,`pop`, `pop.18_34`, `pop.below.pov`, `pct.unemp`, `pct.doctors`, `crime.rate`. The R-squared value is 0.7919, which is an acceptable value for a model with only 7 parameters.

Interaction

We would like to explore the interaction of `region` with other terms. We test the significance of the interaction term with all the variables using ANOVA table. The null hypothesis is including the interaction term does not make an improvement in the fit. We extract their p-values into a single table below.

Table 14: Interaction analysis

Variables	p_values
land.area	0.2514920
pop	0.2008832
pop.18_34	0.1595394
pct.below.pov	0.1640199
pct.unemp	0.0004508
pct.doctors	0.0002572
crime.rate	0.0609544

The p-values are significant for `pct.unemp` and `pct.doctors`, meaning that we only reject the null hypothesis

for these two variables. Because the test before only considered adding one interaction term, we also need to evaluate whether adding both of them also improve the model.

There are only three models possible:

Model 1: Interaction with `pct.unemp`

Model 2: Interaction with `pct.doctors`

Model 3: Interaction with both of them

We can evaluate the BIC values for the three models and choose from them.

Table 15: Comparing BIC for three models

Model	BIC
Model 1	-763.9879
Model 2	-765.3415
Model 3	-765.1857

We can see that the BIC value is the lowest for Model 2. So we add the interaction term for `pct.doctors` only.

Final model discussion

Final model

Here are the coefficients for the final model.

Table 16: Final Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.7958118	0.1660919	71.019802	0.0000000
land.area	-0.0254894	0.0064618	-3.944633	0.0000934
pop	-55.2253736	8.0467534	-6.863063	0.0000000
pop.18_34	-0.1447567	0.0345886	-4.185094	0.0000346
pct.below.pov	-0.2875406	0.0117311	-24.510922	0.0000000
pct.unemp	0.0367914	0.0186342	1.974400	0.0489818
pct.doctors	0.0914359	0.0154388	5.922482	0.0000000
regionNE	0.3959113	0.1532993	2.582603	0.0101384
regionS	0.4388434	0.1237713	3.545599	0.0004352
regionW	0.6894023	0.1815267	3.797801	0.0001672
crime.rate	0.0442536	0.0132308	3.344732	0.0008965
pct.doctors:regionNE	0.0613486	0.0241245	2.543001	0.0113432
pct.doctors:regionS	0.0711746	0.0194159	3.665781	0.0002777
pct.doctors:regionW	0.1053775	0.0286868	3.673383	0.0002698

Model diagnostic check

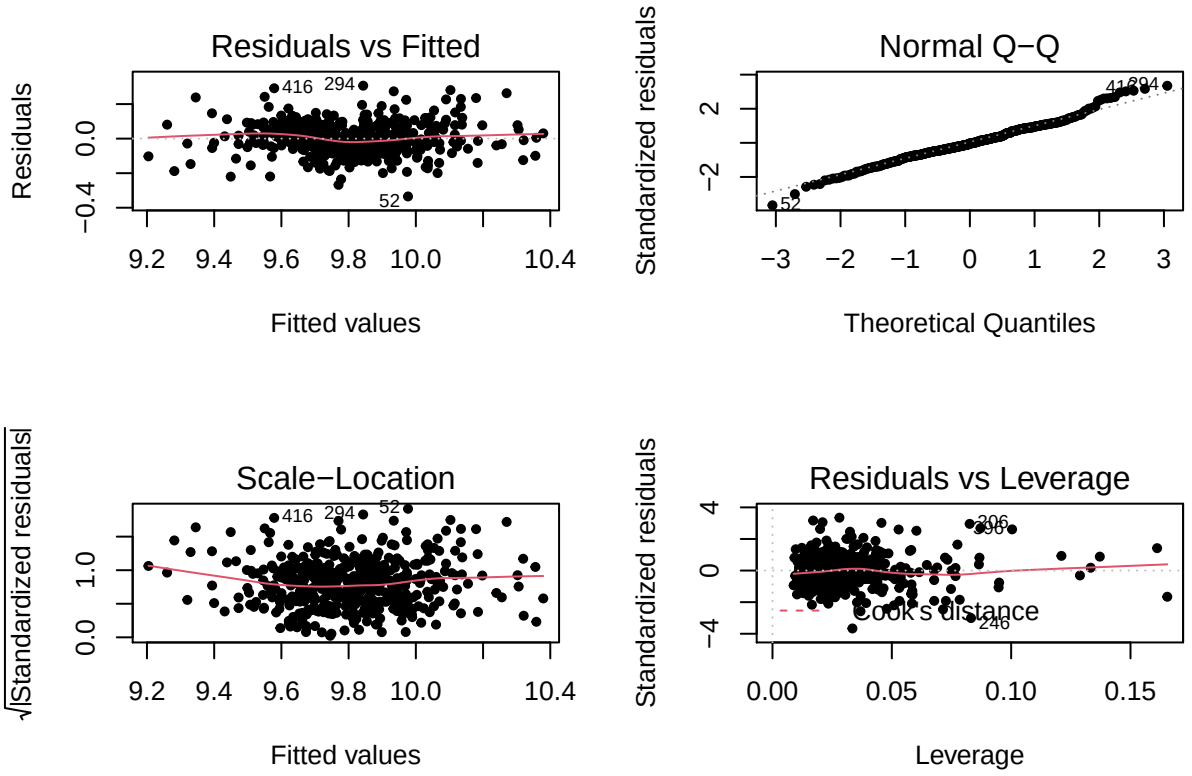


Figure 14: Model diagnostic plots

1. Residuals vs Fitted: The residuals are distributed evenly with a mean 0, with no pattern detectable.
2. Normal Q-Q: The fitted value is slightly skewed and not completely normal.
3. Scale-Location: There is no observable pattern, so the variance looks constant.
4. Residuals vs Leverage: There are no observable leverage points or outliers.

Tradeoff

1. In order to satisfy the normality condition for the variables, we transformed all the variables, which can make the model less interpretable.
2. The transformed data is still not normal, because we chose a nearby value of power for the simplicity of the model.
3. A simple model was chosen using BIC. Some significant factors, such as number of crimes, and unemployment rate were deleted. Other factors, such as population and number of hospital beds were also deleted due to collinearity conditions. This may cause a decrease in R-squared.

Codes

```
## (a)

### Summary statistics

#### Categorial variables

smry_state = data.frame(
  Region = names(table(cdi$region)),
  Count = as.numeric(table(cdi$region))
```

```
)
kable(smry_state, caption = "Summary table: Geographic region")

smry_state = data.frame(
  Region = names(table(cdi$region)),
  Count = as.numeric(table(cdi$region))
)
kable(smry_state, caption = "Summary table: Geographic region")

#### Continuous variables

smry_cts = tibble(
  Variable = c(),
  Min = c(),
  First.Qu = c(),
  Median = c(),
  Mean = c(),
  Third.Qu = c(),
  Max = c()
)

for (i in 4:16){
  values = as.numeric(summary(cdi[,i]))
  smry_cts = add_row(smry_cts,
    Variable = names(cdi)[i],
    Min = values[1],
    First.Qu = values[2],
    Median = values[3],
    Mean = values[4],
    Third.Qu = values[5],
    Max = values[6]
  )
}

kable(smry_cts, caption = "Summary table: Continuous variables")

### Variable features EDA

cdi %>% ggplot()+
  geom_histogram(aes(per.cap.income)) +
  labs(title = "Histogram for Per capita income",
       x = "Income",
       y = "Frequency") +
  theme_bw()

cdi %>% ggplot(aes(x = region, y = per.cap.income)) +
  geom_boxplot()+
  labs(title = "Per capita income vs region",
       x = "Region",
       y = "Per capita income") +
  theme_bw()

p1 = cdi %>% ggplot(aes(x = crimes/pop, y = per.cap.income))+
```

```
  geom_point() +
  labs(title = "Crime rate",
       x = "Crime rate",
       y = "Per capita income") +
  theme_bw()

p2 = cdi %>% ggplot(aes(x = pct.below.pov, y = per.cap.income))+
  geom_point() +
  labs(title = "Percent below poverty level",
       x = "Percent below poverty level",
       y = "Per capita income") +
  theme_bw()

p3 = cdi %>% ggplot(aes(x = pct.bach.deg, y = per.cap.income))+
  geom_point() +
  labs(title = "% of bachelor's degrees",
       x = "% of bachelor's degrees",
       y = "Per capita income") +
  theme_bw()

p4 = cdi %>% ggplot(aes(x = pct.hs.grad, y = per.cap.income))+
  geom_point() +
  labs(title = "Percent of high school grads",
       x = "Percent of high school grads",
       y = "Per capita income") +
  theme_bw()


grid.arrange(p1,p2,p3,p4, nrow = 2, ncol = 2)

p1 = cdi %>% ggplot(aes(x = pop, y = doctors))+
  geom_point() +
  labs(title = "Number of active physicians",
       x = "Total population",
       y = "Number of active physicians") +
  theme_bw()

p2 = cdi %>% ggplot(aes(x = pop, y = hosp.beds))+
  geom_point() +
  labs(title = "Number of hospital beds",
       x = "Total population",
       y = "Number of hospital beds") +
  theme_bw()

p3 = cdi %>% ggplot(aes(x = pop, y = crimes))+
  geom_point() +
  labs(title = "Total serious crimes",
       x = "Total population",
       y = "Total serious crimes") +
  theme_bw()

p4 = cdi %>% ggplot(aes(x = pop, y = tot.income))+
  geom_point() +
```

```
  labs(title = "Total personal income",
       x = "Total population",
       y = "Total personal income") +
  theme_bw()

grid.arrange(p1,p2,p3,p4, nrow = 2, ncol = 2)

## (b)

cdi_more = cdi
cdi_more$crime.rate = cdi$crimes/cdi$pop
lm_1b1 = lm(per.cap.income ~ region + crime.rate, data = cdi_more)
lm_1b2 = lm(per.cap.income ~ region * crime.rate, data = cdi_more)
anova(lm_1b1, lm_1b2)

biccrime = tibble(
  Model = c('Model 1', 'Model 2'),
  BIC = c(BIC(lm_1b1), BIC(lm_1b2))
)
kable(biccrime, caption = "BIC comparison of models")

kable(summary(lm_1b1)$coef, caption = "Coefficients for Model 1")

par(mfrow = c(2,2))
plot(lm_1b1, pch = 20)

lm_1b3 = lm(per.cap.income ~ region + crimes, data = cdi)
kable(summary(lm_1b3)$coef, caption = "Coefficients for Model 3")

## (c)

### Confounding Variable

cdi_pop = cdi %>% mutate(
  pct.doctors = doctors / pop,
  pct.hosp.beds = hosp.beds / pop,
  crime.rate = crimes / pop
)
cdi_pop = cdi_pop[,-c(8:10)]

p1 = cdi_pop %>% ggplot(aes(x = pop, y = pct.doctors))+
  geom_point() +
  labs(title = "Percent of active physicians",
       x = "Total population",
       y = "Percent of active physicians") +
  theme_bw()

p2 = cdi_pop %>% ggplot(aes(x = pop, y = pct.hosp.beds))+
  geom_point() +
  labs(title = "NAverage hospital beds per person",
       x = "Total population",
       y = "Average hospital beds per person") +
  theme_bw()
```

```r
p3 = cdi_pop %>% ggplot(aes(x = pop, y = crime.rate))+
  geom_point() +
  labs(title = "Crime rate",
       x = "Total population",
       y = "Crime rate") +
  theme_bw()

p4 = cdi_pop %>% ggplot(aes(x = pop, y = per.cap.income))+
  geom_point() +
  labs(title = "Per-capita income",
       x = "Total population",
       y = "Per-capita income") +
  theme_bw()

grid.arrange(p1,p2,p3,p4, nrow = 2, ncol = 2)

### Transformations

cdi_pop %>% ggplot()+
  geom_histogram(aes(per.cap.income)) +
  labs(title = "Histogram for Per capita income",
       x = "Income",
       y = "Frequency") +
  theme_bw()

paste("Power Transform Lambda:",powerTransform(cdi_pop$per.cap.income ~ 1)$lambda)

par(mfrow = c(3,4))
for (i in c(4:11,13,15:17)){
  hist(cdi_pop[,i], main = names(cdi_pop)[i])
}

box_cdi = tibble(
  Variable = c(),
  Lambda = c()
)

for (i in c(4:11,13,15:17)){
  box_cdi = add_row(box_cdi, Variable = names(cdi_pop)[i],
                    Lambda = powerTransform(cdi_pop[,i] ~ 1)$lambda)
}

kable(box_cdi, caption = "Box-Cox Power transforms")

cdi_adj = cdi_pop[,-c(1:3)] %>% mutate(
  per.cap.income = log(per.cap.income),
  land.area = log(land.area),
  pop.18_34 = log(pop.18_34),
  pop.65_plus = log(pop.65_plus),
  pct.doctors = log(pct.doctors),
  pct.hosp.beds = log(pct.hosp.beds),
  crime.rate = log(crime.rate),
  pct.bach.deg = log(pct.bach.deg),
```

```
  pct.below.pov = log(pct.below.pov),
  pct.unemp = log(pct.unemp),
  pop = 1/sqrt(pop),
  tot.income = 1/sqrt(tot.income),
  pct.hs.grad = (pct.hs.grad)^3
)


### Variable selection

#### Pre-selection

cdi_adj = cdi_adj[,-10]

#### Collinearity

lm_cdi_num = lm(per.cap.income ~ ., data = cdi_adj[,-10])
VIF = vif(lm_cdi_num)
cdi_vif_table = cbind(summary(lm_cdi_num)$coef[-1,], VIF)
kable(round(cdi_vif_table,2), caption = "Coefficients statistics")

p1 = cdi_adj %>% ggplot(aes(x = pct.below.pov, y = pct.bach.deg))+
  geom_point() +
  labs(title = "Percent below poverty level",
       x = "log(pct.below.pov)",
       y = "log(pct.bach.deg)") +
  theme_bw()

p2 = cdi_adj %>% ggplot(aes(x = pct.unemp, y = pct.bach.deg))+
  geom_point() +
  labs(title = "Percent unemployment",
       x = "log(pct.unemp)",
       y = "log(pct.bach.deg)") +
  theme_bw()

p3 = cdi_adj %>% ggplot(aes(x = pct.hs.grad, y = pct.bach.deg))+
  geom_point() +
  labs(title = "Percent high school graduates",
       x = "log(pct.hs.grad)",
       y = "log(pct.bach.deg)") +
  theme_bw()

grid.arrange(p1,p2,p3,nrow = 1, ncol = 3)

#### BIC

cdi_adj = cdi_adj[,-6]
lm_cdi_full = lm(per.cap.income ~ ., data = cdi_adj)
bic_cdi = stepAIC(lm_cdi_full, direction = 'both', k = log(nrow(cdi_adj)))

summary(bic_cdi)

### Interaction
```

```
lm_cdi_r0 = lm(per.cap.income ~ land.area + pop + pop.18_34 + pct.below.pov +
                 pct.unemp + pct.doctors + crime.rate, data = cdi_adj)
lm_cdi_r1 = lm(per.cap.income ~ land.area*region + pop + pop.18_34 +
                 pct.below.pov + pct.unemp + pct.doctors + crime.rate, data = cdi_adj)
lm_cdi_r2 = lm(per.cap.income ~ land.area + pop*region + pop.18_34 +
                 pct.below.pov + pct.unemp + pct.doctors + crime.rate, data = cdi_adj)
lm_cdi_r3 = lm(per.cap.income ~ land.area + pop + pop.18_34*region +
                 pct.below.pov + pct.unemp + pct.doctors + crime.rate, data = cdi_adj)
lm_cdi_r4 = lm(per.cap.income ~ land.area + pop + pop.18_34 + pct.below.pov*region +
                 pct.unemp + pct.doctors + crime.rate, data = cdi_adj)
lm_cdi_r5 = lm(per.cap.income ~ land.area + pop + pop.18_34 + pct.below.pov +
                 pct.unemp*region + pct.doctors + crime.rate, data = cdi_adj)
lm_cdi_r6 = lm(per.cap.income ~ land.area + pop + pop.18_34 + pct.below.pov +
                 pct.unemp + pct.doctors*region + crime.rate, data = cdi_adj)
lm_cdi_r7 = lm(per.cap.income ~ land.area + pop + pop.18_34 + pct.below.pov +
                 pct.unemp + pct.doctors + crime.rate*region, data = cdi_adj)
inter_table = tibble(
  Variables = c("land.area", "pop", "pop.18_34", "pct.below.pov",
                "pct.unemp", "pct.doctors", "crime.rate"),
  p_values = c(anova(lm_cdi_r1,lm_cdi_r0)[2,6], anova(lm_cdi_r2,lm_cdi_r0)[2,6],
               anova(lm_cdi_r3,lm_cdi_r0)[2,6], anova(lm_cdi_r4,lm_cdi_r0)[2,6],
               anova(lm_cdi_r5,lm_cdi_r0)[2,6], anova(lm_cdi_r6,lm_cdi_r0)[2,6],
               anova(lm_cdi_r7,lm_cdi_r0)[2,6])
)

kable(inter_table, caption = "Interaction analysis")

lm_cdi_rx = lm(per.cap.income ~ land.area + pop + pop.18_34 + pct.below.pov +
                 pct.unemp*region + pct.doctors*region + crime.rate, data = cdi_adj)
bic3_table = tibble(
  Model = c("Model 1", "Model 2", "Model 3"),
  BIC = c(BIC(lm_cdi_r5), BIC(lm_cdi_r6), BIC(lm_cdi_rx))
)
kable(bic3_table, caption = "Comparing BIC for three models")

### Final model discussion

lm_cdi = lm_cdi_r6
kable(summary(lm_cdi)$coef, caption = "Final Model")

par(mfrow = c(2,2))
plot(lm_cdi, pch = 20)
```