Title: An investigation of the relationship between county demographic information and per capita income in the US during 1990 and 1992

Author: Daniel Nason
Department of Statistics and Data Science, Carnegie Mellon University
dnason@andrew.cmu.edu

Abstract

In this paper, we address questions regarding the relationships between data containing county-level demographic variables in the sample and develop models to predict per capita income (PCI) while noting limitations of the data in the context of the research questions. We utilize County Demographic Information (CDI) data collected by Kutner et al. (2005), which includes information and measures on income, population, economic activity, as well as education, age and crime demographics across the most populous counties in the USA from 1990 and 1992. To examine our research questions of interest and build models, we employed exploratory data analysis (EDA), multiple linear regression, and variable selection to capture the relationships between the variables in the study. From the analysis we choose the best set of predictor variables with variable selection techniques to predict PCI while still being reasonably explainable to social scientists, although some relationships, such as a positive association between income and crime and unemployment rate, are surprising. The limitations of the selected model and the data in the study are also detailed and future directions to the analysis are considered to improve the study and validly generalize the results in a contemporary setting.

1 Introduction

The ability to predict the well-being of its citizens is often the aim of policymakers and social scientists as they develop theories and enact programs related to human interactions. Attempts at improving quality of life are typically targeted at improving average income per person (PCI), as classical economic reasoning dictates that more income leads to greater happiness for the individual and, by extension, the community. Developing models that capture the relationship between PCI and other variables to provide accurate predictions would greatly assist both academics and policymakers, allowing them to tailor their actions to improving metrics that will have a positive impact on the community.

However, predicting these metrics with reasonable accuracy has proven surprising challenging due to the unpredictability of human behavior and the interrelated complexities of economies even at the county level. Using data from Kutner et al. (2005), this paper attempts to provide a solution by utilizing multiple linear regression analysis to capture the relationship between PCI and numerous other variables to build a prediction model for the outcome of interest. Through this, we hope to detail important relationships between variables that are common among counties in the US and, as a result, clarify the impacts of these relationships on PCI so social scientists and policymakers can target these metrics when considering new initiatives to improve the general welfare of their citizens.

In attempting to build this model, we will address the following research questions:

1. *Relationships between the variables*: Which variables are related to each other, and which are not? Do these relationships align with our expectations, or are they surprising?
2. *Crime and crime rate*: How is crime related to PCI in the US? Does this relationship depend on the region of the country? Is the relationship in crime data better captured by looking at the crime levels or the crime rate?
3. *What is the best model for predicting PCI, accounting for the following criteria*:
    - Best reflects the social science and the meaning of the variables?
    - Best satisfies modeling assumptions?
    - Is most clearly indicated by the data?
    - Can be explained to someone who is more interested in social, economic and health factors than in mathematics and statistics?
4. *Limitations of the Sample*: Since the data sampled only represents a subset of all counties in the US (373 of 3000) and does not include certain states or locations, should we be worried about either the missing states or the missing counties? Why or why not?

2 Data

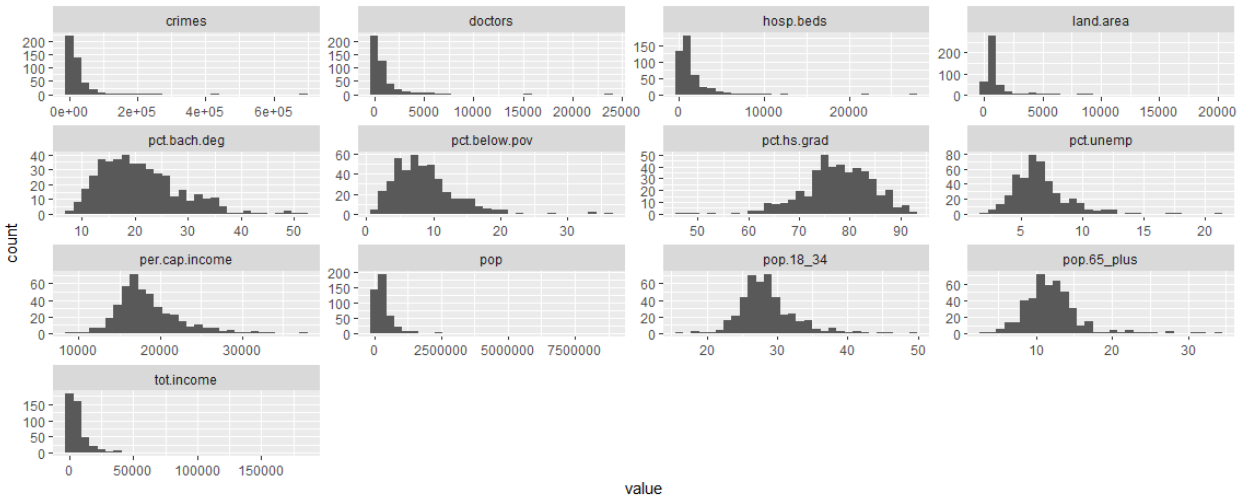
The data for this paper come from Kutner et al. (2005), which was originally collected by the Geospatial and Statistical Data Center at the University of Virginia. It provides county demographic information (CDI) for 440 of the most populous counties in the US, as counties with missing data were removed from the data set. The information generally pertains to the years 1990 and 1992, and for each county the following variables are measured:

Variable definitions for CDI data from Kutner et al. (2005)		
Number	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	Percentage of population aged 18-34	Percentage of 1990 CDI population aged 18-34
7	Percentage of population 65 or older	Percentage of 1990 CID population 65 or older
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 (PCI)	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)
Original source: Geospatial and Statistical Data Center, University of Virginia.		

Table 1: Summary statistics for quantitative variables

	n	mean	sd	min	max	range	q1	median	q3
per.cap.income	440	18561.48	4059.19	8899.0	37541.0	28642.0	16118.25	17759.00	20270.00
land.area	440	1041.41	1549.92	15.0	20062.0	20047.0	451.25	656.50	946.75
pop	440	393010.92	601987.02	100043.0	8863164.0	8763121.0	139027.25	217280.50	436064.50
pop.18_34	440	28.57	4.19	16.4	49.7	33.3	26.20	28.10	30.02
pop.65_plus	440	12.17	3.99	3.0	33.8	30.8	9.88	11.75	13.62
doctors	440	988.00	1789.75	39.0	23677.0	23638.0	182.75	401.00	1036.00
hosp.beds	440	1458.63	2289.13	92.0	27700.0	27608.0	390.75	755.00	1575.75
crimes	440	27111.62	58237.51	563.0	688936.0	688373.0	6219.50	11820.50	26279.50
pct.hs.grad	440	77.56	7.02	46.6	92.9	46.3	73.88	77.70	82.40
pct.bach.deg	440	21.08	7.65	8.1	52.3	44.2	15.28	19.70	25.33
pct.below.pov	440	8.72	4.66	1.4	36.3	34.9	5.30	7.90	10.90
pct.unemp	440	6.60	2.34	2.2	21.3	19.1	5.10	6.20	7.50
tot.income	440	7869.27	12884.32	1141.0	184230.0	183089.0	2311.00	3857.00	8654.25

Figure 1: Histograms of continuous variables for CDI data



We examine the data for both qualitative and quantitative variables. Table 1 (Appendix, page 17) and Figure 1 (Appendix, page 18) provide information on the distributions for the quantitative variables, illustrating that the following variables are noticeable right-skewed: Total Serious Crimes, Number of Active Physicians, Number of Hospital Beds, Land Area, Total Population, Per capita income, and Total personal income.

The remaining quantitative variables are approximately normally distributed and will remain untransformed. Tables 2 and 3 (Appendix, pages 18-19) show that the categorical variables ID, County, and State have a considerable number of unique values associated with them, and since these levels add little to the analysis, they are ignored. Table 4 (Appendix, page 20) and Figure 2 (Appendix, page 21) show the information for the categorical variable Region. The most observations are in the South region (152), while the fewest observations are within the West region (77). We also see 47 states and the District of Columbia have at least one observation in the data set, but three states (Alaska, Wyoming, and Iowa) do not have a single county with a large enough population to be included in the data set when the study was conducted.

3 Methods

To investigate the research questions of interest, we outline the approach of how each question will be addressed. We use methods outlined in the Sheather (2009) textbook for exploratory data analysis, regression modeling, and variable selection.

Question 1: Relationships between the variables

Before examining the relationships between variables, we addressed our findings in the Data section about the distributions of some quantitative variables in the data. Specifically, since some of the quantitative variables (Total serious Crimes, Number of active physicians, Number of hospital beds, Land area, Total population, PCI, and Total personal income) are right-skewed, we applied transformations to them to reduce skew and more closely resemble normal distributions. These transformations have the advantage of not compromising the interpretability of the data, so any transformations would need to be interpreted as percentages rather than as levels. To investigate the relationship between the predictor variables and PCI, we generated scatter plot and correlation matrices to examine both the

visual and numerical strength of the associations. We also evaluated the relationship between Region (categorical) and PCI using a side-by-side boxplot and the corresponding five number summary. These graphics and tables allowed us to determine if there are any unexpected relationships in the data compared to our intuition. The results also guided our approach to the other research questions and for modeling the data appropriately to make predictions.

Question 2: Crime and crime rate

To determine if there is a relationship between PCI and crime, we created a linear regression model to predict PCI using the transformed variables obtained from Question 1. We then included dummy variables for region in the model in two steps: first, only including the dummy variables with no interactions; second, including the dummy variables with interactions between the region and crime. These results illustrated whether the relationship between crime and PCI depends on the region of the country.

Similarly, we repeated this process for crime rate, which was formulated from the ratio of Number of Crimes to Total Population, as well as the transformed crime rate, based on the distribution of the variable. This permitted us to determine the impact of using crime versus crime rate, as well as which variable to use when building the model to best predict PCI.

Question 3: Best model for predicting PCI

Using the results from Questions 1 and 2, we built a regression model to predict PCI. We also excluded the Total Population and Total Income variables due to their functional relationship with PCI, and temporarily excluded Region. This term was later accounted for to see its impact on the model, whether as a main effects or interaction effects term. After examining the full additive linear model, we employed all subsets regression, stepwise regression, and LASSO as variable selection techniques on the data. With these we found the optimal subset of predictor variables to keep in the model, and utilized residual diagnostics, summary outputs, and information criteria (adjusted R^2, AIC, and BIC) to determine the "best" model. We considered the criteria detailed in the research question and accounted for it as we evaluated more complicated transformations such as including interactions to improve the predictive power of the model. Interactions between the continuous variables in the data were included as well to see if they added any noteworthy predictive power to the model.

Question 4: Limitations of the Sample

To answer this question, we assess the data itself in the context of the research problem. How the data was collected and what it represents was evaluated, including its shortcomings for building a model and generalizing the results to other counties and to a more contemporary setting. We also considered how the categorical variables are defined and related to each other.

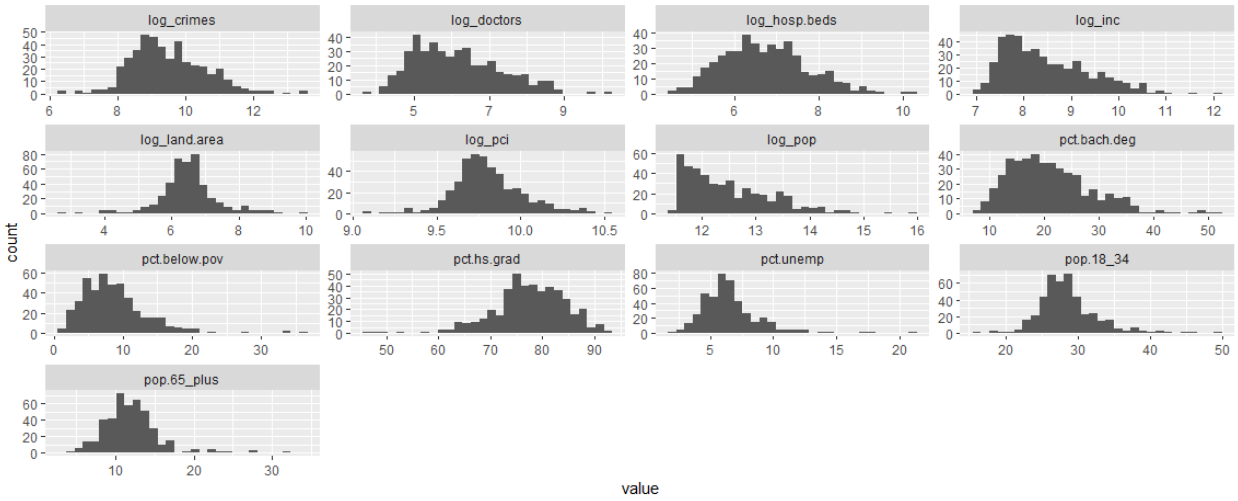
4 Results

Question 1: Relationships between the variables

The scatterplot and correlation coefficient matrices in Figures 3 and 4, respectively (Appendix, pages 22-23) illustrate that there is evidence of some non-linear relationships between the predictor and response variables, specifically for the predictors Land area, Total population, Percentage of population aged 18-34, Percentage of population 65 or older, Number of physicians, Number of hospital beds, Total

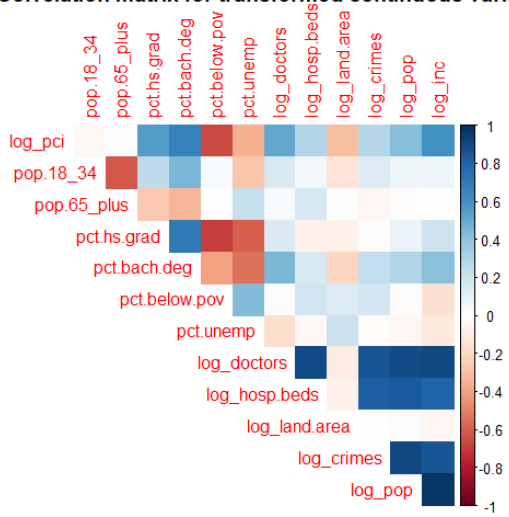
serious crimes, Percentage of population below poverty line, and Total personal income. However, since we previously identified that a few of the quantitative variables are right-skewed (Total serious Crimes, Number of active physicians, Number of hospital beds, Land area, Total population, PCI, and Total personal income), log transformations are applied to each of the variables.

Figure 5: Histograms of transformed continuous variables



As illustrated in Figure 5 (Appendix, page 24), these variables appear to resemble the normal distribution more closely after the log transformations are applied. We utilize the transformed versions of these variables as we proceed with answering both this question and the other research questions. We also examine how the relationships between the continuous random variables have changed.

Figure 7: Correlation matrix for transformed continuous variables



Figures 6 and 7 (Appendix, pages 25-26) display the scatter plot and correlation coefficient matrices for the continuous variables. After applying these transformations, we see that the relationships between PCI and the remainder of the variables more closely resemble linear associations. The correlation matrix also identifies strong linear relationships between the transformed predictor variables, specifically: Number of doctors and Number of hospital beds, Total serious crimes, Total population, and Total personal income; Number of hospital beds and Total serious crimes, Total population, and Total

personal income; Total serious crimes and Total population and Total income, and Total population and Total income. Each relationship has a correlation coefficient of at least 0.7, so these relationships will need to be accounted for when building a prediction model for PCI to answer Question 3 due to potential collinearity between the predictors.

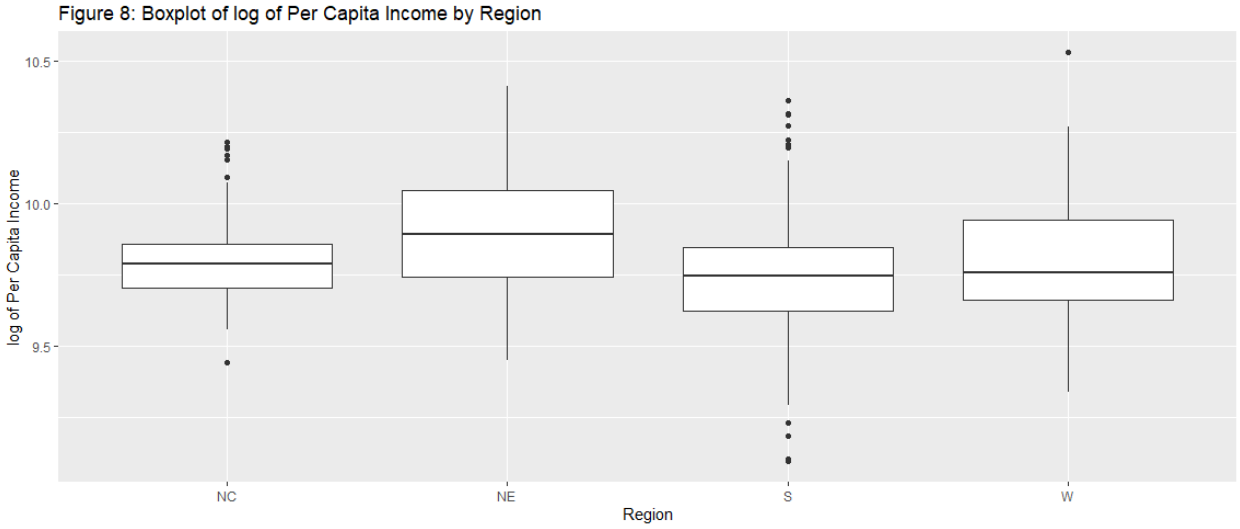


Table 5 and Figure 8 (Appendix, page 27) illustrate the relationship between the log transformed PCI and Region. While overlapping IQRs suggest that PCI is similar across regions, further investigation is conducted with regards to Question 3 to see if this relationship becomes useful for prediction after account for the other variables in the analysis.

Question 2: Crime and crime rate

We see that after applying log transformations to the noticeably skewed continuous variables in the data, their pairwise relationships between the transformed variables are also affected. There also does not appear to be a noteworthy difference in PCI after accounting for region; however, this warrants further investigation to determine whether accounting for other variables impacts this relationship. Specifically, we examine this relationship in the context of the association between Total serious crimes and PCI.

$$Model\ B: \log(PCI) = \beta_0 + \beta_1 crimes + \beta_2 region + \varepsilon$$

$$Model\ E: \log(PCI) = \beta_0 + \beta_1\, crimes/population + \beta_2 region + \varepsilon$$

Table 6: Comparing models including Crime and Crime Rate

	df	AIC	BIC	R2 adj.
model_b	6	-227.4746	-202.9539	0.1959087
model_e	6	-172.1347	-147.6140	0.0881411

Three models are fit to investigate how crime relates to PCI. The first model regresses PCI on crimes, while the next two models include a dummy variable for region: one model examining only the additive effects, and the other including both the additive and interaction effects between region and crime. The results of the nested F-test (Appendix, page 28) illustrate that the model that includes the additive

effects is most appropriate for modeling PCI. This suggests that both crime and region of the US are related to PCI, but the relationship between income and crime does not depend on region. It should also be noted there is a positive linear relationship between crime and PCI and that this relationship is statistically significant in the best model selected. We see that a 1 percent increase in the number of Total serious crimes is associated with a 0.07 percent increase in PCI, and an associated t-value of 7.92 (Appendix, page 29). There are statistically significant disparities across regions as well; PCI is on average 10 percent higher in Northeast, and 6 and 9 percent lower in the West and South, respectively compared to North Central (Appendix, page 29). However, only the coefficient for West is not statistically significant at the 5% since it has a p-value of 0.0503.

We also examine the relationship between PCI and crime rate, after applying the log transformation to crime rate (Appendix, page 30). The nested F-test yields similar results when the crime rate variable is included in the model (Appendix, page 30) in that the additive effects is most appropriate for modeling PCI. As shown in Table 7 (Appendix, page 34), the model also finds a positive, statistically significant linear relationship between PCI and crime rate, although the effect is smaller and less significant in comparison to the model using the crimes variable. It predicts that a 1 percent increase in the crime rate is associated with a 0.04 percent increase in PCI, and an associated t-value of 1.98 (Appendix, page 34). There are statistically significant disparities across regions as well; PCI is on average 11 percent higher in Northeast, and 2 and 7 percent lower in the West and South, respectively, compared to the North Central (Appendix, page 34). While the differences are statistically significant for Northeast and South, the coefficient estimate for West loses its statistical significance when crime rate is included in the model instead of crime (p-value of 0.4195).

$$Fitted\ Model\ E: \log(PCI) = 9.94 + 0.04\log(crime\ rate) + 0.11 regionNE - 0.07 regionS - 0.02 regionW$$

Table 7: Model E Coefficient and Standard Error Estimates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.9363	0.0693	143.3029	0.0000
log_crime_rate	0.0424	0.0215	1.9754	0.0489
regionNE	0.1146	0.0276	4.1505	0.0000
regionS	-0.0746	0.0262	-2.8410	0.0047
regionW	-0.0243	0.0300	-0.8080	0.4195

Since the results of the selected models are similar for either of the additive models, we examine the residual diagnostic plots, AIC, BIC, and related outputs for these models. The results are presented in Figures 10 and 11 (Appendix, pages 32-33) and summarized in Table 6 (Appendix, page 34). While the model analyses suggest that the crime model slightly outperforms the crime rate model for predicting PCI in these categories, since neither model performs exceptionally well, the crime rate variable is preferred due its interpretability in the context of the problem. This variable will remain in the analysis when modeling PCI for Question 3, though the model will be supplemented with the other variables to improve predictive power.

Question 3: Best model for predicting PCI

Using the results obtained from Questions 1 and 2, we include the transformed crime rate variable in lieu of Total serious crimes our analysis but exclude the Region variable when initially building the

regression model. This will later be included after the model selection procedures have been applied. Since Total population and Total income can be used to deterministically model PCI, these are dropped from the model building process.

Naively using the remaining continuous variables to model PCI (Appendix, pages 35-36), we see that the coefficient estimates for Number of hospital beds and crime rate are not statistically significant, and Figure 12 (Appendix, page 37) illustrates that the regression model assumptions are not exactly satisfied. Multicollinearity is present among some of the predictors based on their VIF values as well (Appendix, page 36). Since these issues are present in the model, we utilize variable selection techniques to determine the optimal subset of the predictor variables needed to predict PCI. Using the following table, we explore regression models selected by the variable selection techniques:

Variable definitions for model selection		
Symbol	Variable Name	Variable Description Notes
y	Per capita income	Log transformation is applied
x1	Percentage of population aged 18-34	N/A
x2	Percentage of population 65 or older	N/A
X3	Percent high school graduates	N/A
X4	Percent bachelor's degrees	N/A
X5	Percent below poverty level	N/A
X6	Percent unemployment	N/A
X7	Number of active physicians	Log transformation is applied
X8	Number of hospital beds	Log transformation is applied
X9	Land Area	Log transformation is applied
X10	Crime rate = Total serious crimes / Total population	Log transformation is applied

Utilizing all subsets regression, stepwise regression, and LASSO as variable selection techniques, we obtain the following two models. The predictors obtained in Model 1 were selected by all subsets regression, LASSO, and stepwise regression when specified for BIC, while the predictors for Model 2 were selected by stepwise regression when specified for AIC (Appendix, pages 37-42).

$$Model\ 1: \log(PCI) = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 \log(x_6) + \beta_7 \log(x_7) + \beta_8 \log(x_8) + \varepsilon$$

$$Model\ 2: \log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 \log(x_6) + \beta_7 \log(x_7) + \beta_8 \log(x_8) + \varepsilon$$

Comparing the two models, we see that they are similar except that Model 2 includes the variable Percentage of population 65 or older while Model 1 does not. Examining the model analyses for each model, we find that the results are extremely similar for each model's goodness of fit measures, statistical significance of beta coefficient estimates, diagnostic plots, VIF values, and marginal model plots (Appendix, pages 42-46). Therefore, we consider the effect of including interaction terms for the Region variable in both models before determining which model to use.

Table 8: Comparison Table for Models 1 and 2 (including interactions)

	df	AIC	BIC	R2 adj.
model1	9	-942.2740	-905.4931	0.8427
model2	10	-944.8883	-904.0206	0.8439
model1_region_interactions	21	-986.9437	-901.1215	0.8615
model2_region_interactions	25	-983.6060	-881.4366	0.8617

The results of including the interaction terms are outlined in the Appendix on pages 47-55, and the information for model comparisons are summarized in Table 8 (Appendix, page 54). We see in Table 8 that despite the considerable number of terms added to both models when accounting for interactions across each of the levels of the Regions variable, there is little predictive power added to the model for BIC and adjusted R^2. Similarly, the residual diagnostic plots displayed in Figure 19 (Appendix, page 50) and Figure 20 (Appendix, page 53) show no discernable improvement in comparison with the residual plots for Models 1 and 2. These results suggest that including the terms does not help to better satisfy the regression model assumptions and do not improve the predictive power of the model tremendously despite the number of terms in each model more than doubling. Therefore, we ignore these models to avoid potentially overfitting the data and select between the models without the interaction terms.

There is little difference in the more technical aspects of the Models 1 and 2 such as goodness of fit, satisfaction of the regression model assumptions, multicollinearity measures, and inclusion of the properly specified terms in the model. Therefore, based on the criteria outlined in Question 1, Model 1 is selected since it is more parsimonious and provides similar predictive power compared to Model 2.

$$\begin{aligned} \textit{Fitted Model } 1\text{: } &\log(y) \\ &= 10.222 - 0.014x_1 - 0.004x_3 + 0.015x_4 - 0.024x_5 + 0.011\log(x_6) \\ &+ 0.061\log(x_7) - 0.036\log(x_8) \end{aligned}$$

Table 9: Model 1 Coefficient and Standard Error Estimates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.2225	0.0931	109.7765	0e+00
pop.18_34	-0.0139	0.0011	-12.5080	0e+00
pct.hs.grad	-0.0044	0.0011	-4.0714	1e-04
pct.bach.deg	0.0154	0.0009	16.6408	0e+00
pct.below.pov	-0.0243	0.0013	-19.2940	0e+00
pct.unemp	0.0106	0.0022	4.8705	0e+00
log_doctors	0.0607	0.0040	15.1000	0e+00
log_land.area	-0.0357	0.0048	-7.4683	0e+00

The simplicity of the model as well as the transformations applied to the variables permit the coefficients to be easily interpreted in the context of the data by social scientists. For example, Table 9 (Appendix, page 55) shows that the coefficient on Percent bachelor's degree suggests, holding other variables in the model constant, a 1-unit increase in the percentage of people with bachelor's degrees in the county is associated with a 0.015 percent increase in PCI. Interestingly, the coefficient for percent of high school graduates does not have the same sign, although the coefficient suggests that there is little practice effect; holding everything else constant, a 1-unit increase in the percentage of people with high school degrees is associated with a 0.004 percent decrease in PCI. Other noteworthy coefficient

interpretations include the estimate for percentage unemployment, which suggests that, holding everything else constant, a 1-unit increase in the percentage of people unemployed is associated with a 0.011 percentage increase in PCI. The contextual meaning and significances of these estimates are further explored in the Discussion.

Question 4: Limitations of the Sample

From Tables 2 and 3 (Appendix, pages 18-19), we see that there are many unique levels for the categorical variables County (373) and State (48). Table 4 (Appendix, page 20) shows which state is categorized in which region. Within the State variable, 47 states and the District of Columbia are represented in the data set. Alaska, Iowa, and Wyoming are omitted from the data since at the time of the sample, they did not contain any of the 440 most populous counties in the US. These states could have been omitted from the data set because they have lower relative population sizes or the population within their state is more evenly distributed across all counties. While this idea will be further explored in the Discussion, it is important to note that the limitations of the data may reduce the validity of generalizing the model beyond the sample for making predictions.

5 Discussion

Question 1: Relationships between the variables

We first focus on the pairwise relationships in the untransformed data between PCI and the other continuous random variables illustrated in Figures 3 and 4. The strong positive linear association between PCI and education aligns with our expectation of their relationship, as increased education is positively associated on average with more earnings, especially in a predominately service-based economy such as the US. Similarly, the negative relationship between PCI and the indicators of lower economic activity (Percent below poverty level and Percent unemployed) is also consistent with our expectation; as increases in unemployment imply less people are working and income is reduced, and therefore poverty is more likely to increase.

Interestingly, there appears to be a lack of a pairwise relationship between PCI and some of the other continuous variables. Land area would expect to be more negatively correlated with PCI since rural and suburban areas tend to be more spacious and have lower incomes than urban areas due to the relatively lower cost of living. We would also expect a negative relationship between PCI and Percentage of population aged 18-34, since younger people have had less time to work or develop their skills and therefore usually have lower incomes compared to the rest of the population. Percentage of population aged 65 and older having no relationship is also surprising, since we'd expect that incomes would be lower for predominately senior citizens since many people of them are retired and living on fixed incomes or government assistance (or both). Number of physicians and hospital beds is noteworthy since we'd expect wealthier areas to have more doctors and medical resources since they are able to afford better healthcare compared to their less-wealthy counterparts. Finally, Total serious crimes does not appear to have a clear relationship with PCI when we would expect this relationship to be negative since wealthy areas also tend to have fewer crimes. However, this is not always the case as some high-income urban areas tend to have more crimes committed in comparison to their suburban and rural counterparts.

Examining the pairwise relationships outside of PCI, we see a few noteworthy associations. Specifically, population has a strong positive linear association with doctors, hospital beds, and crimes. This aligns with our intuition as larger populations require more medical personnel and resources, and a large population also provides more opportunities from individuals to commit crimes. The negative relationship between the population variables (18-34 versus 65 and older) is reasonable since they are mutually exclusive subsets of the total population and are separated by at least one generation. There is also a strong relationship between educational attainment and economic indicators in the directions we would expect. Specifically, the percentage of high school graduates and bachelor's degree holders is highly correlated since the former is required for the latter. Both measures are negatively correlated with unemployment and poverty rates, which aligns with our results for PCI; more educational attainment leads to higher average income and more employment opportunities.

After applying log transformations to the right skewed variables as illustrated in Figures 6 and 7, we see that these transformations more closely align to our expectations for the relationships between PCI and the other continuous random variables. This is also true for relationships between the continuous variables outside of PCI. Interestingly, the relationship between PCI and the population subset variables is unchanged from the transformation, and crimes appears to be weakly positively associated with PCI. Figure 8 also illustrates while PCI is slightly higher in the Northeast and West regions of the US, there is still some overlap with the IQRs and therefore the relationship would not be statistically significant. This result is consistent with our expectations given the higher cost of living and population density for the coastal regions of the US coupled with more densely populated metropolitan areas in comparison with the non-coastal areas of the country.

While our intuition is more consistent for the transformed variables, there are some limitations to looking at just the pairwise relationships between variables. We saw in some instances that the relationships are likely due to not addressing confounding variables, such as in the case between crime and income or crime and medical resources. These variables are very likely both related to population, and without controlling for this we could make inappropriate conclusions about the data. Controlling for these omitted variables could explain any non-sensical relationships in the data as we answer the remaining research questions.

Question 2: Crime and crime rate

$$Model\ E: \log(PCI) = \beta_0 + \beta_1\, crimes/population + \beta_2 region + \varepsilon$$

Our analysis between crime and PCI yields some interesting results. We find a positive, statistically significant relationship between these variables, and this association persists after transforming the crime rate variable. Since both models are not exceptionally useful, the crime rate model (Model E) is used due to its interpretability in the context of the data; similar units are present for both PCI and crime rate (i.e. crime per capita).

The relationship between crime and PCI runs counter to our intuition, as we would not expect a positive relationship between the two variables. The relationship, if any, would be anticipated to be negative since wealthy areas also tend to have fewer crimes. However, there are exceptions to this rule, as some high-income urban areas tend to have more crimes committed in comparison to non-urban areas. Since this model is not exceptionally useful in predicting PCI and shows a relationship that is the opposite of our expectation, confounding variables are likely driving this result. Specifically, we see from the

correlation matrix in Figure 6 that population is strongly correlated with both the predictor and response variables in the model we selected. It would be reasonable to assume that a larger population provides more opportunities from crime but also generates more income and wealth due to increased economic activity.

Question 3: Best model for predicting PCI

$$\begin{aligned} Final\ Model\ (Model\ 1): \log(PCI) \\ = \beta_0 + \beta_1(percent\ of\ populate\ 18-34) + \beta_{23}(percent\ high\ school\ graduates) \\ + \beta_3(percent\ bachelor's\ degree) + \beta_4(percent\ below\ poverty\ level) \\ + \beta_5 \log(percent\ unemployment) + \beta_6 \log(number\ of\ active\ physicians) \\ + \beta_7 \log(number\ of\ hospital\ beds) + \varepsilon \end{aligned}$$

Table 9: Model 1 Coefficient and Standard Error Estimates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.2225	0.0931	109.7765	0e+00
pop.18_34	-0.0139	0.0011	-12.5080	0e+00
pct.hs.grad	-0.0044	0.0011	-4.0714	1e-04
pct.bach.deg	0.0154	0.0009	16.6408	0e+00
pct.below.pov	-0.0243	0.0013	-19.2940	0e+00
pct.unemp	0.0106	0.0022	4.8705	0e+00
log_doctors	0.0607	0.0040	15.1000	0e+00
log_land.area	-0.0357	0.0048	-7.4683	0e+00

We find that Model 1 is the most appropriate prediction model for PCI because it best satisfies the criteria outlined in Question 3. It is the most parsimonious of the models identified from the variable selection procedure while still providing similar predictive power compared to more complicated models, such as Model 2 or the hybrid of Models 1 and 2 that include interaction terms for region. However, we must also examine the more practical features of the model to determine its utility in the context of the social sciences.

In addition to reasonably satisfying the more technical features of a good model (Appendix, pages 42-46), the model features also bolster the argument that it is the most appropriate for the data. Since the modeling assumptions are roughly satisfied for the model (with the exception of the Normal QQ plot in Figure 15 on page 43 in the Appendix), inference can also be conducted with reasonable certainty. Additionally, because it is a first-order additive linear model, the coefficient estimates are easily interpretable; this is even true for the transformed variables (doctors and land area) since log transformations were applied to reduce skew. It therefore is useful for inference and interpretation in the context of the social sciences.

We examine the coefficients of the model in Table 9 to see how they align with our intuition. While they are mostly consistent with our expectations, some of the signs of the estimates are surprising. The signs of the estimates for Percentage of population aged 18-34, Number of physicians, and Percentage of Bachelor's degrees align with our expectations. A younger workforce has not had as much time to obtain the skills needed to earn higher wages and therefore would have lower incomes, while doctors are highly skilled positions that require extensive education and typically have higher incomes. This aligns with the story about how an educated workforce tends to earn higher incomes and contributes

positively to PCI. Surprisingly, this relationship is negative (but small) for Percent of high school graduates. A possible partial explanation for this could be the increase education requirements of the workforce in the US's predominantly service economy, as it is more difficult to obtain such a position with just a high school diploma. It is also surprising to see that while estimates for Percentage below the poverty line and Land area are negative related, Percentage unemployment is positively associated PCI. One possible explanation for this is rural migration to urban and suburban counties since the lack of opportunity in rural areas would force people to move into cities and therefore contribute to the unemployment rate in those areas. The relatively higher incomes and cost of living with these urban areas would offset any increases in PCI due to having a larger unemployed workforce.

Model 1 makes the best tradeoff between reflecting the social science and meaning of the variables, satisfying the model assumptions, modeling the variation in the response variable, and simplicity in explaining the results of the model to a social scientist instead of someone focused on the more technical aspects of the model. It is not without flaws, however, as lack of normality in the residuals suggests that the model may not generate valid prediction intervals. It also sacrifices complexity for practicality and interpretability, as more complicated interactions were omitted from the model to avoid confusion in interpretation of the coefficient estimates. We are also not sure about the predictive capabilities of the model since all the data was utilized to train the model. Additional evaluation would be necessary to determine whether the model is useful or potentially overfitting the data; this could be done on similar test data or using cross-validation. A noteworthy consideration is that the data are from approximately 30 years ago and may provide little similarities to more modern CDI data due to the rapid pace of economic development and technological innovation over the last three decades. Including more practically useful variables to help predict PCI, such as whether the county is urban or not, the type of workforce in the county (STEM or otherwise), and budget resources for the local government, could improve the model. Training and testing the model on more updated data may help to improve its practical utility and predictive power as well.

Question 4: Limitations of the Sample

To determine whether our analysis is generalizable to the omitted states and counties from the CDI data set, we take a closer look at the data. The advantages of the data are that they are sufficiently large (440 observations) such that it can be assumed that the Central Limit Theorem applies for our analysis. Additionally, since the sample is stratified across the 47 states and Washington D.C., we can be sure that we are not excluding those states as we analyze the subset of the population. These facts give us confidence in the model and any associated inferences that are made from the data.

However, there are some important weaknesses that must be considered before considering applying the model elsewhere. We are not aware of the study design employed by the authors to avoid sampling biases. Also, the data do not include observations from Alaska, Iowa, or Wyoming. This provides information about the states based on when the sampling was performed in 1990 and 1992. Their omission could reflect the geographic distribution of the population across these states as well as the relatively small population size in the state. For example, while Alaska or Wyoming account for a large portion of the geographic area of the US, they are sparsely populated and predominantly rural states. Additionally, while Iowa might have more population density than Alaska or Wyoming, its population could be uniformly distributed across each county and lack the density from a major urban center that would have caused it to be included in the data set. The fact that these states are not included in the

data set suggests that CDI information for these states are fundamentally different from data in the sample.

Another weakness to note is that the data was only collected from the most populous counties in the country roughly three decades prior to today. The relationships that appear between the variables in these counties may not validly translate to other smaller counties. Specifically, it is more likely that urban areas are captured in this data since cities usually have higher population density in comparison with smaller suburban or rural counties. Applying the model generated from urban data would probably not be exceptionally useful for these counties. Also, it is unlikely that if we resampled the 440 most populous counties today the same counties would appear in the data due to changes over the last thirty years, especially in the workforce. As one of the largest working generations in American history approaches retirement age, these individuals may relocate to other parts of the country and thus alter the demographics. These weaknesses must be considered when attempting to apply this model in a contemporary setting.

Addressing the shortcomings of the data and the model built using them is critical to providing an appropriate statistical tool for predictions. Future work to handle these limitations includes collecting more variables from updated samples of the county demographic information and expanding the dataset to include a larger cross section of the counties in the country as well as from the states omitted. This would avoid omitting potentially important relationships between variables by not collecting data in the less populated counties. If this is impractical, some investigation would be needed to determine how representative the sample is relative to the counties omitted from the data. EDA would yield insights of how representative the data are for the remainder of the country. Repeating the analysis of the data with updated information and training and testing the data would greatly improve the quality of the model and account for shifts in the relationships between the variables over time. Such an analysis could inform a discussion on how craft appropriate policies in a modern setting to improve PCI and strengthen the validity of the predictions generated by the model in the context of this critical social science issue.

6 References

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

Sheather, S.J. (2009), A Modern Approach to Regression with R. New York: Springer Science + Business Media LLC.

7 Technical Appendix

```
library(tidyverse)
library(arm)
library(car)
library(leaps)
library(kableExtra)
library(glmnet)
library(MASS)
library(psych)
library(corrplot)
library(reshape2)
setwd("C:/Users/Owner/CMU MSP Program/Fall 2021/36-617 - Applied Linear Models/Midterm Project")
cdi <- read.table("cdi.dat")
cdi_dat <- cdi[, -which(colnames(cdi) == "id")] %>%
  mutate(region = as.factor(region)) # categorical variables
```

Data

Quantitative variables EDA

```
# checking if there are any NAs in the data
colSums(is.na(cdi))
```

```
##             id         county          state      land.area            pop
##              0              0              0              0              0
##      pop.18_34    pop.65_plus        doctors      hosp.beds         crimes
##              0              0              0              0              0
##    pct.hs.grad   pct.bach.deg  pct.below.pov      pct.unemp per.cap.income
##              0              0              0              0              0
##     tot.income         region
##              0              0
```

```
# making dataframes for ease of analysis
cdi_cont <- cdi_dat[, !names(cdi_dat) %in% c("county", "state", "region")] %>%
  relocate(per.cap.income)

q1 <- rep(0, ncol(cdi_cont))
q3 <- rep(0, ncol(cdi_cont))

for (i in seq(ncol(cdi_cont))) {
  q1[i] <- quantile(cdi_cont[,i], 0.25)
  q3[i] <- quantile(cdi_cont[,i], 0.75)
}
```

Table 1: Summary statistics for quantitative variables

	n	mean	sd	min	max	range	q1	median	q3
per.cap.income	440	18561.48	4059.19	8899.0	37541.0	28642.0	16118.25	17759.00	20270.00
land.area	440	1041.41	1549.92	15.0	20062.0	20047.0	451.25	656.50	946.75
pop	440	393010.92	601987.02	100043.0	8863164.0	8763121.0	139027.25	217280.50	436064.50
pop.18_34	440	28.57	4.19	16.4	49.7	33.3	26.20	28.10	30.02
pop.65_plus	440	12.17	3.99	3.0	33.8	30.8	9.88	11.75	13.62
doctors	440	988.00	1789.75	39.0	23677.0	23638.0	182.75	401.00	1036.00
hosp.beds	440	1458.63	2289.13	92.0	27700.0	27608.0	390.75	755.00	1575.75
crimes	440	27111.62	58237.51	563.0	688936.0	688373.0	6219.50	11820.50	26279.50
pct.hs.grad	440	77.56	7.02	46.6	92.9	46.3	73.88	77.70	82.40
pct.bach.deg	440	21.08	7.65	8.1	52.3	44.2	15.28	19.70	25.33
pct.below.pov	440	8.72	4.66	1.4	36.3	34.9	5.30	7.90	10.90
pct.unemp	440	6.60	2.34	2.2	21.3	19.1	5.10	6.20	7.50
tot.income	440	7869.27	12884.32	1141.0	184230.0	183089.0	2311.00	3857.00	8654.25

```
tab <- as.data.frame(describe(cdi_cont, skew = F))
tab <- tab[ ,-c(1, ncol(tab))]
tab$q1 <- q1
tab$median <- apply(cdi_cont, 2, median)
tab$q3 <- q3

round(tab,2) %>%
  kbl(booktabs=T, caption = "Summary statistics for quantitative variables") %>%
  kable_classic()
```

```
ggplot(gather(cdi_cont), aes(x = value)) +
  geom_histogram() +
  facet_wrap(~key, scales = 'free') +
  labs(title = "Figure 1: Histograms of continuous variables for CDI data")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Figure 1: Histograms of continuous variables for CDI data

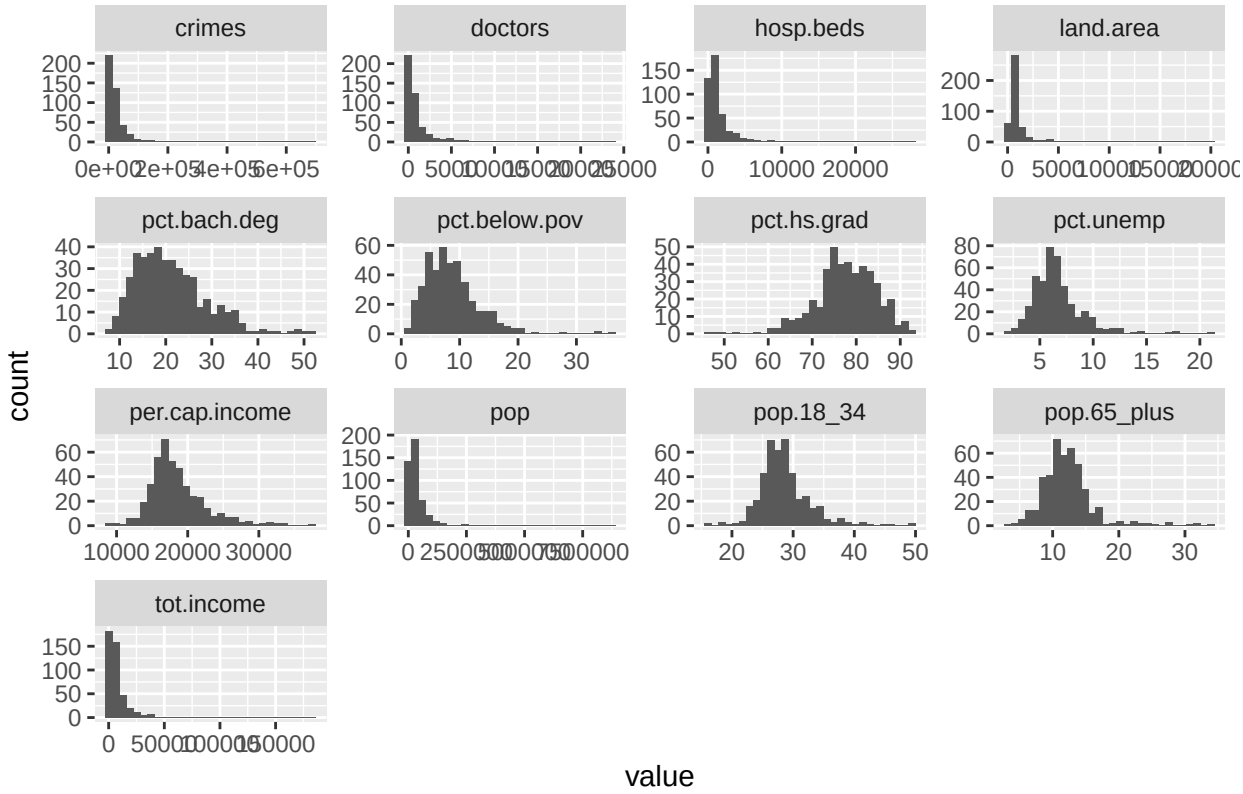


Table 1 presents the summary statistics for the continuous variables, and the histograms displayed in Figure 1 show that some variables have a noteworthy right-skew (crimes, doctors, hosp.beds, land.area, pop, tot.income, per.cap.income), and transformations should be considered on the data. We also verify that there are no NAs (missing data) in the data frame.

Categorical Variables EDA

```
county.state <- with(cdi, paste(county, state))
tmp <- as.data.frame(matrix(sort(county.state),ncol=4))
names(tmp) <- paste("Counties",c("1-110","111-220","221-330","331-440"))
tmp[1:10,] %>%
  kbl(booktabs=T,longtable=T,caption="Unique counties with states") %>%
  kable_classic(full_width=F)
```

Table 2: Unique counties with states

Counties 1-110	Counties 111-220	Counties 221-330	Counties 331-440
Ada ID	Ector TX	Lycoming PA	Rockingham NH
Adams CO	El_Dorado CA	Macomb MI	Rockland NY
Aiken SC	El_Paso CO	Macon IL	Rowan NC
Alachua FL	El_Paso TX	Madison AL	Rutherford TN
Alamance NC	Elkhart IN	Madison IL	Sacramento CA
Alameda CA	Erie NY	Madison IN	Saginaw MI

Table 3: Unique values in CDI data

	unique values
id	440
county	373
state	48
land.area	384
pop	440
pop.18_34	149
pop.65_plus	137
doctors	360
hosp.beds	391
crimes	437
pct.hs.grad	223
pct.bach.deg	220
pct.below.pov	155
pct.unemp	97
per.cap.income	436
tot.income	428
region	4

Albany NY	Erie PA	Mahoning OH	Salt_Lake UT
Alexandria_City VA	Escambia FL	Manatee FL	San_Bernardino CA
Allegheny PA	Essex MA	Marathon WI	San_Diego CA
Allen IN	Essex NJ	Maricopa AZ	San_Francisco CA

```
apply(cdi,2,function(x) {length(unique(x))}) %>%
  kbl(booktabs=T,col.names="unique values",caption="Unique values in CDI data") %>%
  kable_classic(full_width=F)
```

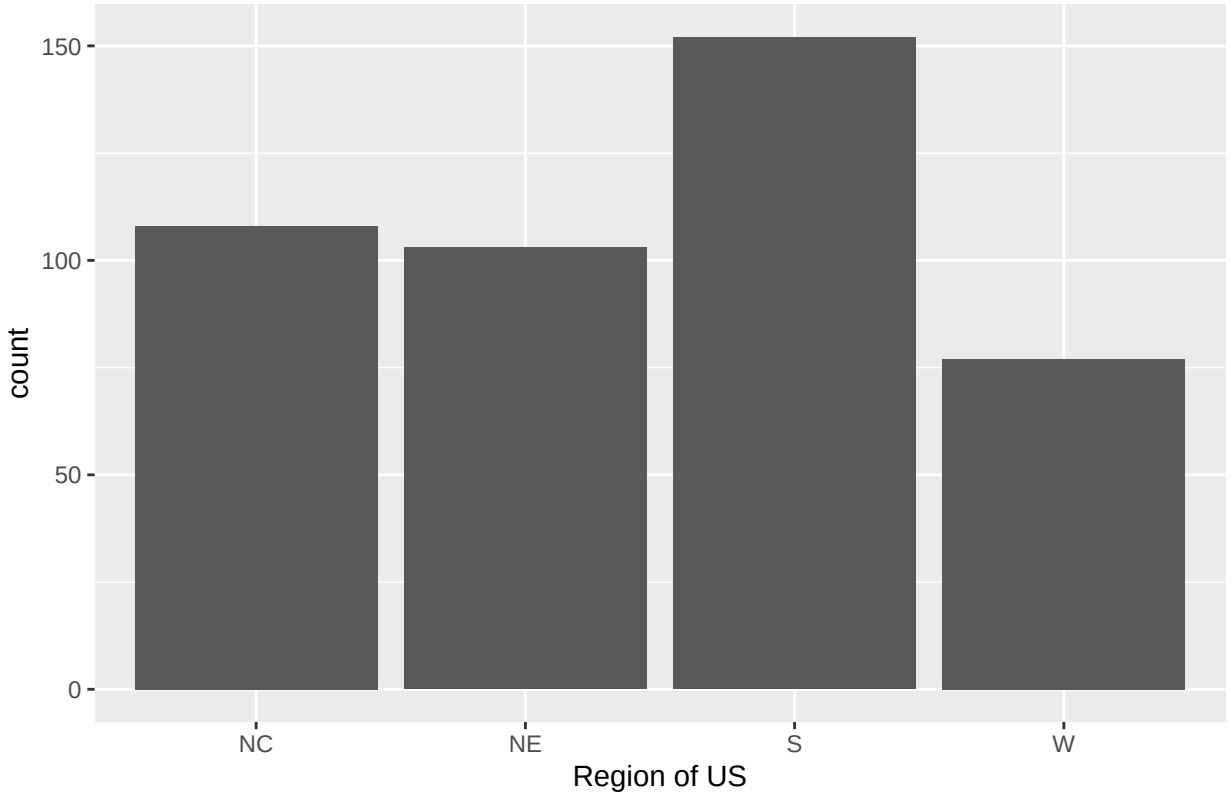
```
margin_table <- addmargins(table(cdi_dat$state, cdi_dat$region))
margin_table %>%
  kbl(booktabs=T, caption = "Frequency table of observations of states across each region") %>%
  kable_classic()
```

```
ggplot(data = cdi_dat, mapping = aes(region)) +
  geom_bar() +
  labs(title = "Figure 2: Where the most populous US counties are located",
       x = "Region of US")
```

Table 4: Frequency table of observations of states across each region

	NC	NE	S	W	Sum
AL	0	0	7	0	7
AR	0	0	2	0	2
AZ	0	0	0	5	5
CA	0	0	0	34	34
CO	0	0	0	9	9
CT	0	8	0	0	8
DC	0	0	1	0	1
DE	0	2	0	0	2
FL	0	0	29	0	29
GA	0	0	9	0	9
HI	0	0	0	3	3
ID	0	0	0	1	1
IL	17	0	0	0	17
IN	14	0	0	0	14
KS	4	0	0	0	4
KY	0	0	3	0	3
LA	0	0	9	0	9
MA	0	11	0	0	11
MD	0	0	10	0	10
ME	0	5	0	0	5
MI	18	0	0	0	18
MN	7	0	0	0	7
MO	8	0	0	0	8
MS	0	0	3	0	3
MT	0	0	0	1	1
NC	0	0	18	0	18
ND	1	0	0	0	1
NE	3	0	0	0	3
NH	0	4	0	0	4
NJ	0	18	0	0	18
NM	0	0	0	2	2
NV	0	0	0	2	2
NY	0	22	0	0	22
OH	24	0	0	0	24
OK	0	0	4	0	4
OR	0	0	0	6	6
PA	0	29	0	0	29
RI	0	3	0	0	3
SC	0	0	11	0	11
SD	1	0	0	0	1
TN	0	0	8	0	8
TX	0	0	28	0	28
UT	0	0	0	4	4
VA	0	0	9	0	9
VT	0	1	0	0	1
WA	0	0	0	10	10
WI	11	0	0	0	11
WV	0	0	1	0	1
Sum	108	103	152	77	440

Figure 2: Where the most populous US counties are located



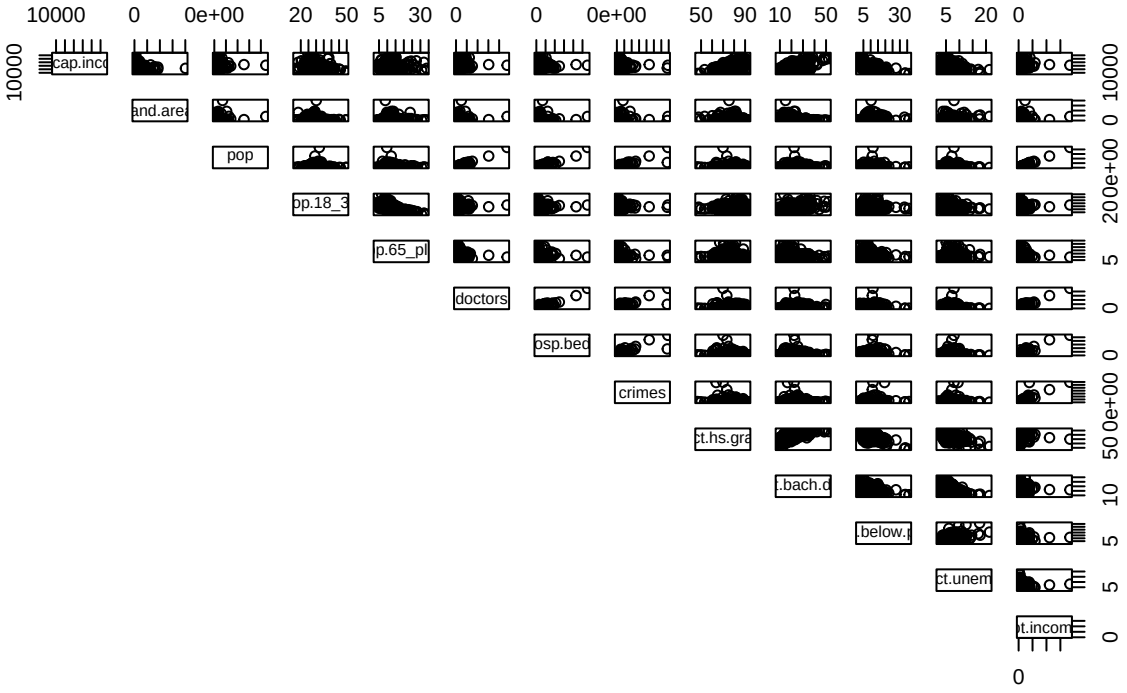
We see in Table 3 that id is just an index variable for the number of rows, and therefore it is dropped from the analysis. Similarly, we see in Table 2 that there are numerous unique values for state (48) and county (373). Therefore, they are not included as categorical variables since they have too many different levels to justify remaining in the analysis. Table 4 displays which states are classified in which region, while Figure 2 shows the distribution of the region variable.

Results

Research Question 1

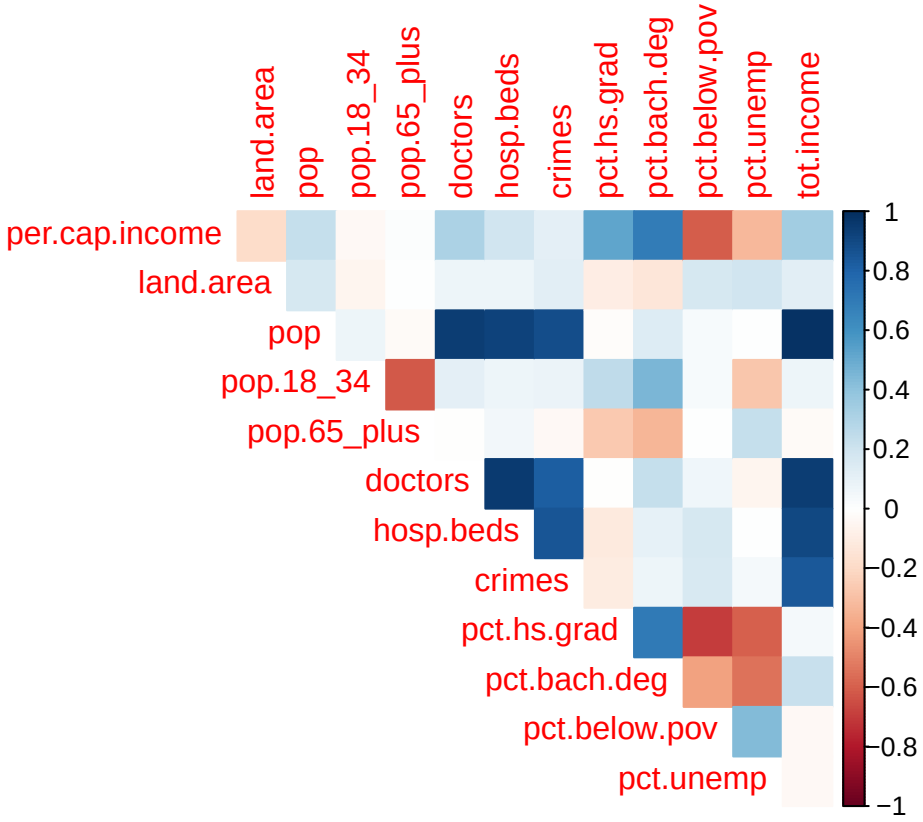
```
# results section - relationship between variables
pairs(cdi_cont, lower.panel = NULL,
      main = "Figure 3: Scatterplot matrix for continuous random variables")
```

Figure 3: Scatterplot matrix for continuous random variables



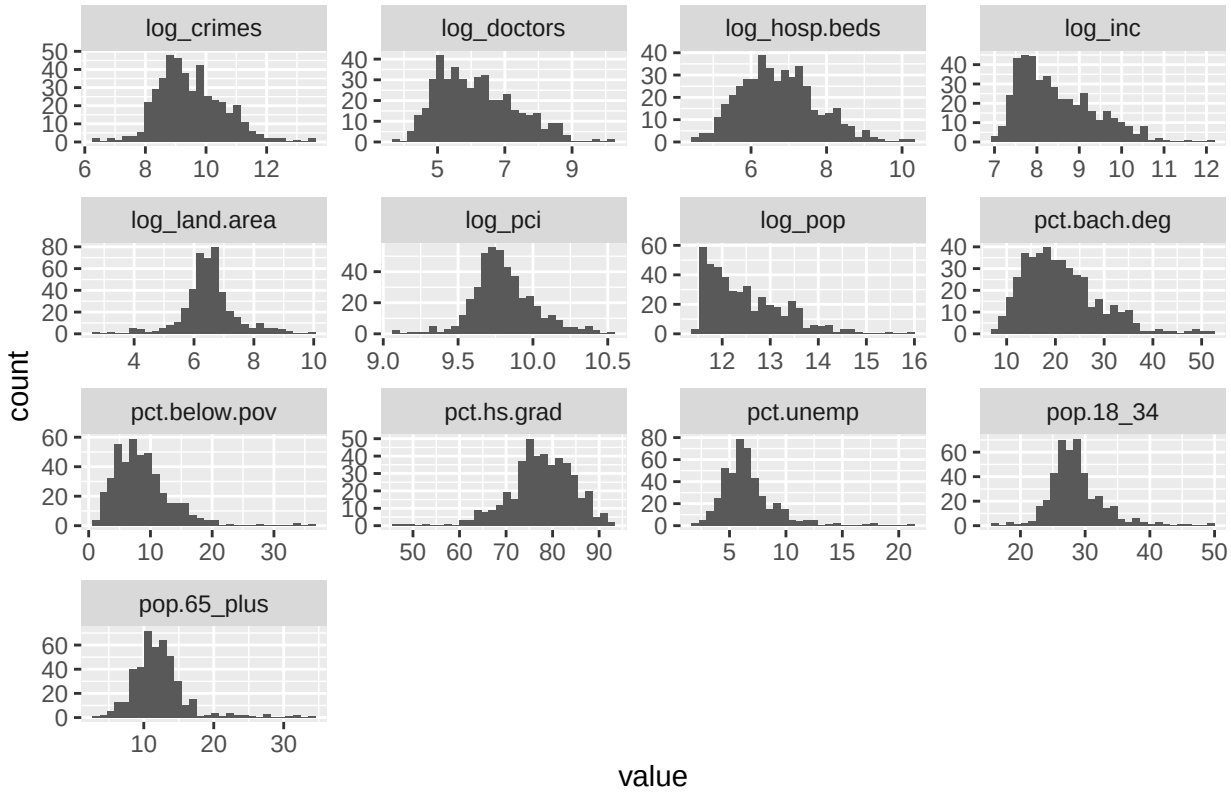

```
corrplot(cor(cdi_cont), color = T, type = "upper",
         title = "Figure 4: Correlation matrix for continuous random variables",
         mar=c(0,0,1,0), diag = F, method = "color")
```

Figure 4: Correlation matrix for continuous random variables



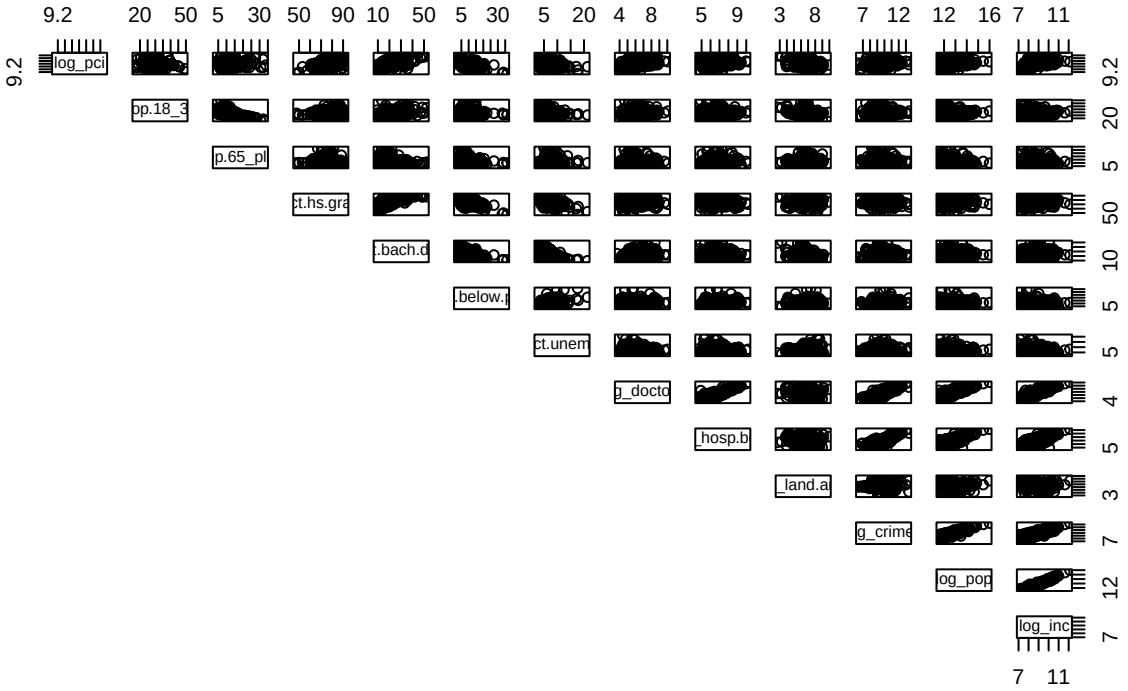

```
# updated cdi data set for analysis
cdi_cont2 <- cdi_cont %>%
  mutate(log_doctors = log(doctors),
         log_hosp.beds = log(hosp.beds),
         log_land.area = log(land.area),
         log_crimes = log(crimes),
         log_pop = log(pop),
         log_inc = log(tot.income),
         log_pci = log(per.cap.income)
  ) %>%
  relocate(log_pci)
idx1 <- c("doctors", "hosp.beds", "land.area", "crimes", "pop",
          "tot.income", "per.cap.income")
cdi_cont2 <- cdi_cont2[,!names(cdi_cont2) %in% idx1]

ggplot(gather(cdi_cont2), aes(x = value)) +
  geom_histogram() +
  facet_wrap(~key, scales = 'free') +
  labs(title = "Figure 5: Histograms of transformed continuous variables")
```

Figure 5: Histograms of transformed continuous variables



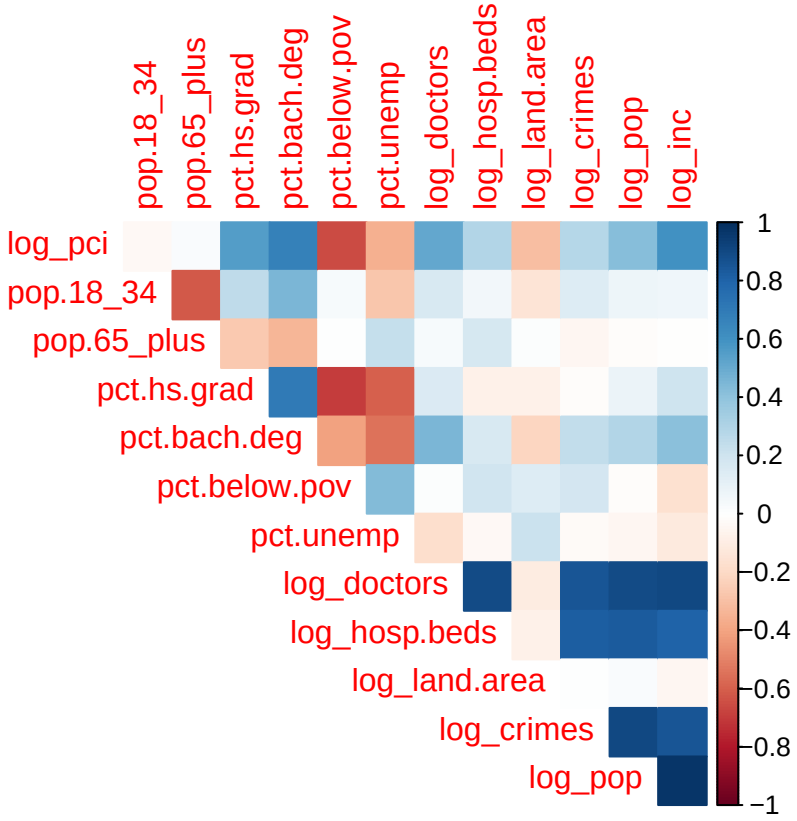

```
pairs(cdi_cont2, lower.panel = NULL,
      main = "Figure 6: Scatterplot matrix for transformed continuous variables")
```

Figure 6: Scatterplot matrix for transformed continuous variables




```
corrplot(cor(cdi_cont2), color = T, type = "upper",
         title = "Figure 7: Correlation matrix for transformed continuous variables",
         mar=c(0,0,1,0), diag = F, method = "color")
```

Figure 7: Correlation matrix for transformed continuous variables

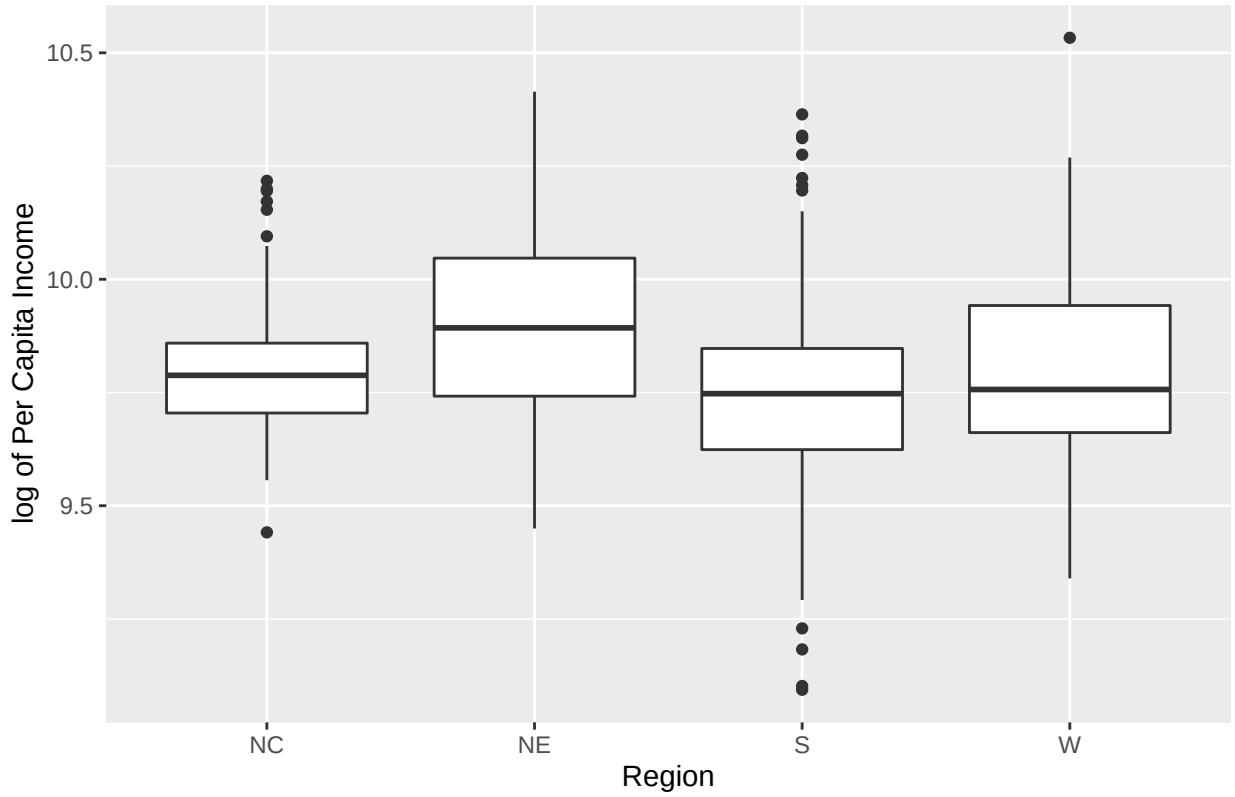



```
ggplot(data = cdi_dat, mapping = aes(x = region, y = log(per.cap.income))) +
  geom_boxplot() +
  labs(title = "Figure 8: Boxplot of log of Per Capita Income by Region",
       x = "Region", y = "log of Per Capita Income")
```

Table 5: Five number summary of Per Capita Income by Region

region	n	min	Q1	median	Q3	max
NC	108	12597	16377.0	17817	19135.5	27378
NE	103	12704	17016.0	19785	23079.0	33330
S	152	8899	15118.5	17110	18933.5	31699
W	77	11379	15701.0	17268	20786.0	37541

Figure 8: Boxplot of log of Per Capita Income by Region




```
cdi_dat %>%
  group_by(region) %>%
  summarise(n = n(),
            min = fivenum(per.cap.income)[1],
            Q1 = fivenum(per.cap.income)[2],
            median = fivenum(per.cap.income)[3],
            Q3 = fivenum(per.cap.income)[4],
            max = fivenum(per.cap.income)[5]) %>%
  kbl(booktabs=T, caption="Five number summary of Per Capita Income by Region") %>%
  kable_classic(full_width=F)
```

Figures 3 shows that there is some evidence of non-linear relationships between PCI and the other variables, specifically for land area, population, pop.18_34, pop.65_plus, doctors, hosp.beds, crimes, pct.below.pov, and tot.income. There is also some evidence of linearity among these variables as seen in Figures 3 and 4. After applying transformations to the skewed distributions, we see that the transformed variables now more closely resemble a Normal distribution as illustrated in Figure 5. Figure 6 shows that the relationships

between PCI and the other variables more closely resemble linear relationships after the transformations are applied. It should be noted that after the transformation is applied, strong linear relationships also appear between pairs of the continuous variables, as illustrated in Figure 7. This should be kept in mind when generating models to predict per capita income.

Figure 8 and Table 5 show that the medians of per capita income by region are relatively similar, but it is worth investigating whether region is useful in predicting the response variable, as well as its relationship with the other continuous variables.

Research Question 2

Crime Model

```
# building a model to predict per-capita income from crime/region
cdi_analysis_1 <- cdi_dat[, !names(cdi_dat) %in% c("county", "state")] %>%
  mutate(
    log_doctors = log(doctors),
    log_hosp.beds = log(hosp.beds),
    log_land.area = log(land.area),
    log_crimes = log(crimes),
    log_pop = log(pop),
    log_inc = log(tot.income),
    log_pci = log(per.cap.income)
  ) %>%
  relocate(log_pci)

lm.q2a <- lm(log_pci ~ log_crimes, data = cdi_analysis_1)
model_b <- lm(log_pci ~ log_crimes + region, data = cdi_analysis_1)
lm.q2c <- lm(log_pci ~ log_crimes * region, data = cdi_analysis_1)
anova(lm.q2a, model_b, lm.q2c)
```

```
## Analysis of Variance Table
##
## Model 1: log_pci ~ log_crimes
## Model 2: log_pci ~ log_crimes + region
## Model 3: log_pci ~ log_crimes * region
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 17.271
## 2    435 14.949  3   2.32194 22.4823 1.523e-13 ***
## 3    432 14.872  3   0.07678  0.7434    0.5266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model_b)
```

```
##
## Call:
## lm(formula = log_pci ~ log_crimes + region, data = cdi_analysis_1)
##
## 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
```

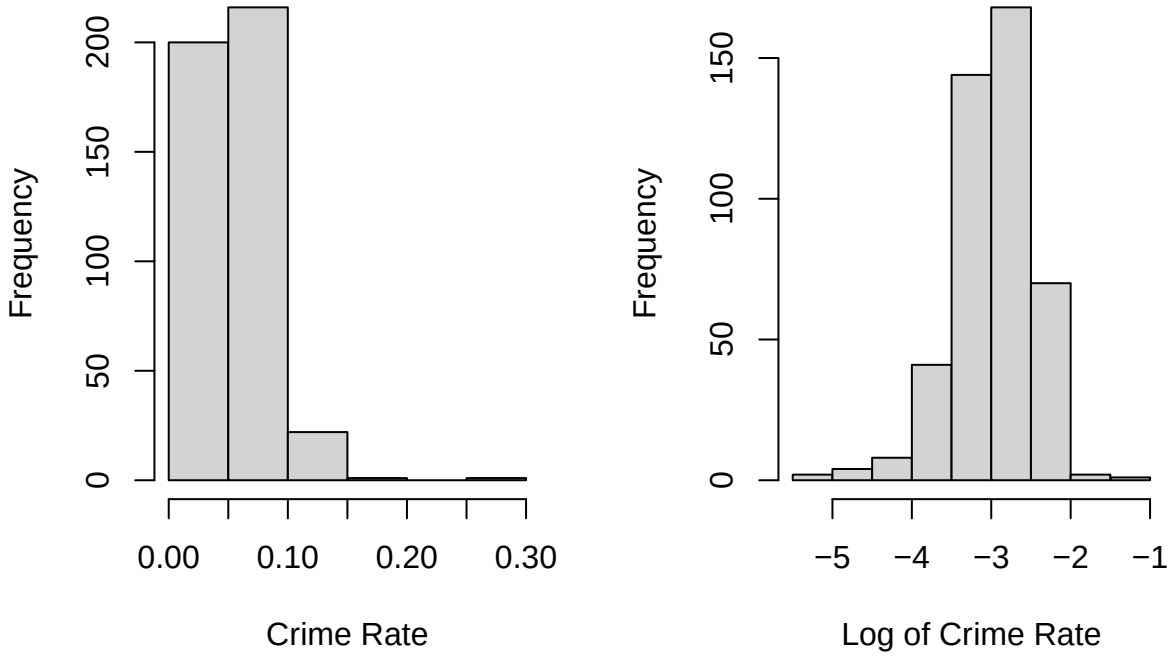
Since two of the models are nested versions of each other, we apply the nested F-test to determine whether the relationship between per capita income and crime rate depends on different regions of the country.

The results of the nested F-test illustrate that including region in the additive ANCOVA model is appropriate based on the p-value being less than 0.05, but the interaction terms are unnecessary due to the relatively large p-value (greater than 0.05). The model also displays a statistically significant relationship between the transformed crimes and per capita income variables, suggesting that there is a positive linear relationship between the variables. This result is surprising, however, since it would be expected that higher crime rates are not associated with wealthier areas. The result may be driven by omitted variables such as population, which is likely correlated with both crime and income. It can be argued that population density in a county implies more workers and higher income, but can also be associated with higher crime rates.

Crime Rate Model

```
cdi_analysis_2 <- cdi_analysis_1 %>%
  mutate(
    crime_rate = crimes / pop,
    log_crime_rate = log(crime_rate)
  )

par(mfrow = c(1,2))
hist(cdi_analysis_2$crime_rate, main = "", xlab = "Crime Rate")
hist(cdi_analysis_2$log_crime_rate, main = "", xlab = "Log of Crime Rate")
mtext("Figure 9: Histograms of Crime Rate and Log of Crime Rate",
      side=3, adj = 1, cex=1.2)
```

Figure 9: Histograms of Crime Rate and Log of Crime Rate




```
par(mfrow = c(1,1))

lm.q2d <- lm(log_pci ~ log_crime_rate, data = cdi_analysis_2)
model_e <- lm(log_pci ~ log_crime_rate + region, data = cdi_analysis_2)
lm.q2f <- lm(log_pci ~ log_crime_rate * region, data = cdi_analysis_2)
anova(lm.q2d, model_e, lm.q2f)
```

```
## Analysis of Variance Table
##
## Model 1: log_pci ~ log_crime_rate
## Model 2: log_pci ~ log_crime_rate + region
## Model 3: log_pci ~ log_crime_rate * region
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 18.697
## 2    435 16.952  3   1.74465 14.8407 3.263e-09 ***
## 3    432 16.928  3   0.02408  0.2048     0.893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model_e)
```

```
##
## Call:
## lm(formula = log_pci ~ log_crime_rate + region, data = cdi_analysis_2)
```

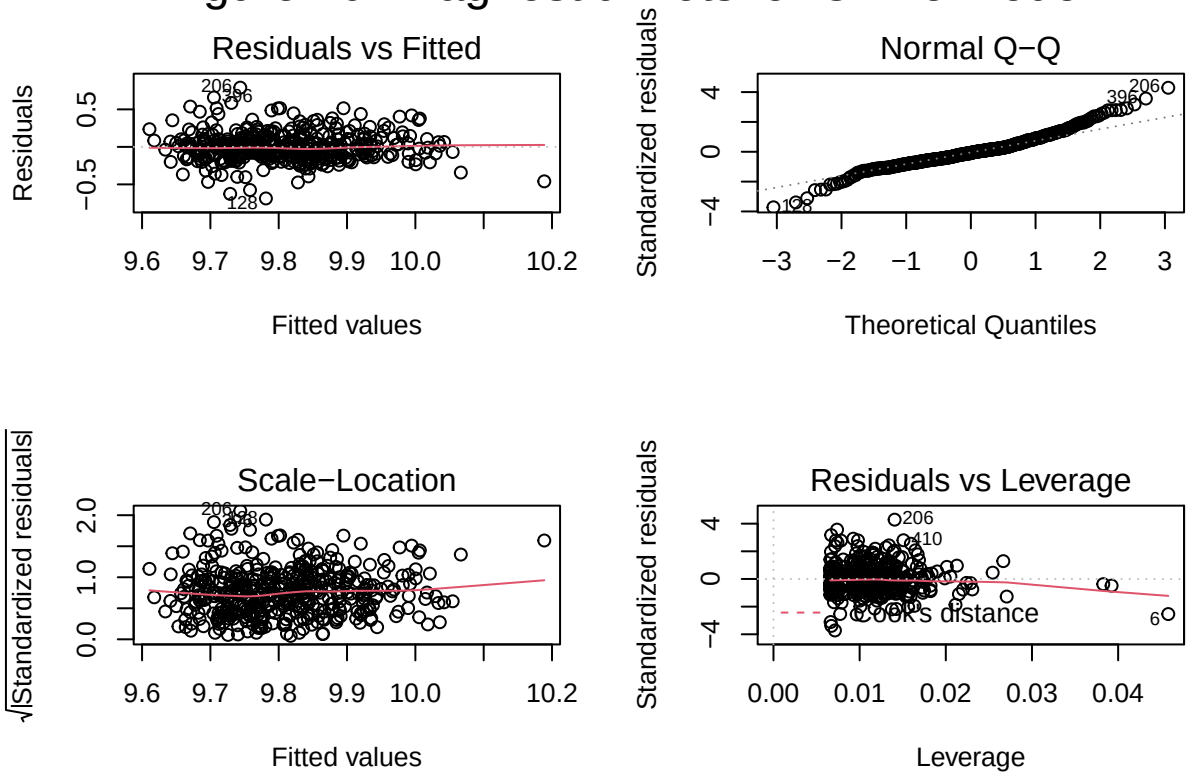
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65832 -0.11431 -0.01548  0.10838  0.75657
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.93628    0.06934 143.303  < 2e-16 ***
## log_crime_rate  0.04243    0.02148   1.975  0.04885 *
## 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
```

We also examine the crime rate variable created by taking the ratio of Number of Crimes to Total Population. Figure 9 illustrates that Crime Rate is right skewed, but becomes approximately normal after the log transformation is applied. Therefore, the log transformation of Crime Rate is used when building the models.

Similar to the Crime variable case, since two of the models are nested versions of each other, we apply the nested F-test to determine whether the relationship between per capita income and crime rate depends on different regions of the country. The results of the nested F-test illustrate that including region in the additive ANCOVA model is appropriate based on the p-value being less than 0.05, but the interaction terms are unnecessary due to the relatively large p-value (greater than 0.05). The model also displays a statistically significant relationship between the transformed crimes and per capita income variables, suggesting that there is a positive linear relationship between the variables.

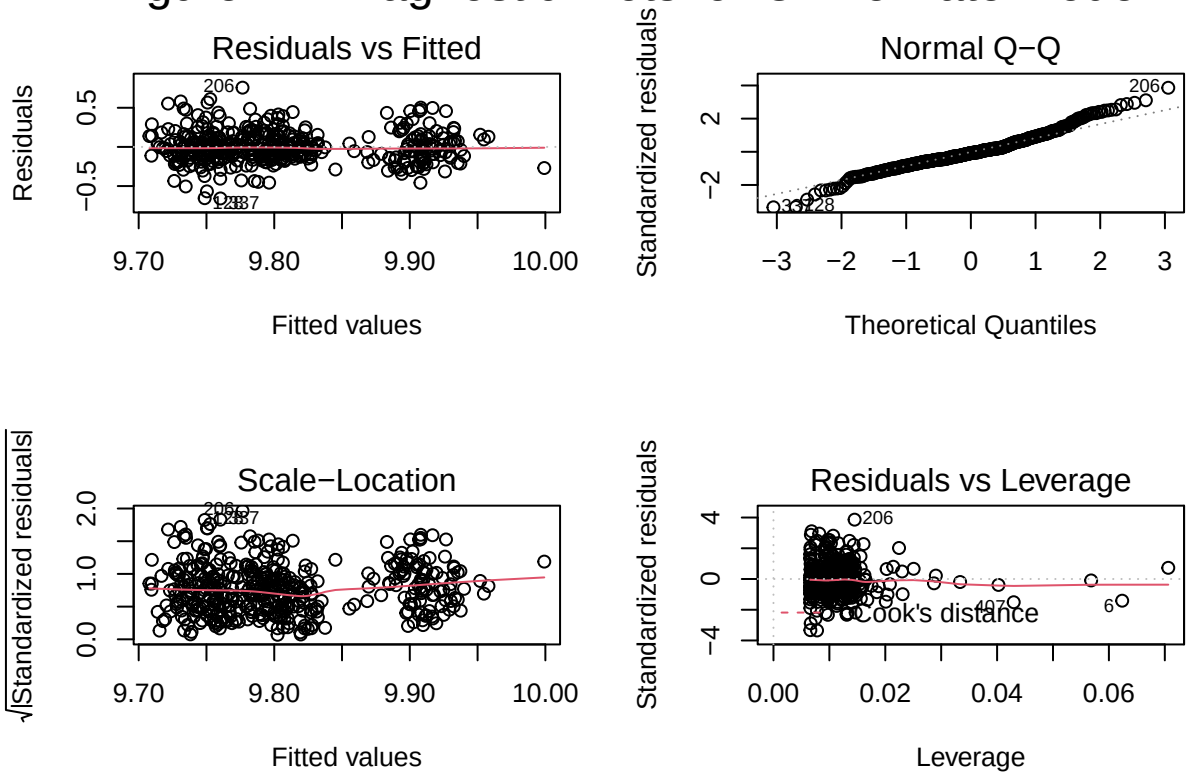
```
par(mfrow=c(2,2))
plot(model_b)
mtext("Figure 10: Diagnostic Plots for Crime model",
      side = 3, line = -2, outer = TRUE, cex = 1.5)
```

Figure 10: Diagnostic Plots for Crime model




```
plot(model_e)
mtext("Figure 11: Diagnostic Plots for Crime Rate model",
      side = 3, line = -2, outer = TRUE, cex = 1.5)
```

Figure 11: Diagnostic Plots for Crime Rate model




```
par(mfrow=c(1,1))

formula(model_b)
```

```
## log_pci ~ log_crimes + region
```

```
round(coef(summary(model_b)),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
```

```
formula(model_e)
```

```
## log_pci ~ log_crime_rate + region
```

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

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.94       0.07  143.30     0.00
```

Table 6: Comparing models including Crime and Crime Rate

	df	AIC	BIC	R2 adj.
model_b	6	-227.4746	-202.9539	0.1959087
model_e	6	-172.1347	-147.6140	0.0881411

Table 7: Model E Coefficient and Standard Error Estimates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.9363	0.0693	143.3029	0.0000
log_crime_rate	0.0424	0.0215	1.9754	0.0489
regionNE	0.1146	0.0276	4.1505	0.0000
regionS	-0.0746	0.0262	-2.8410	0.0047
regionW	-0.0243	0.0300	-0.8080	0.4195

```
## log_crime_rate     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
```

```
data.frame(AIC=AIC(model_b,model_e),
           BIC=BIC(model_b,model_e),
           R2 = c(summary(model_b)$adj.r.squared, summary(model_e)$adj.r.squared))[,-3] %>%
  kbl(booktabs=T, caption = "Comparing models including Crime and Crime Rate",
      col.names=c("df", "AIC","BIC","R2 adj.")) %>%
  kable_classic(full_width=F)
```

```
q2_final <- round(summary(model_e)$coefficients,4)
q2_final %>%
  kbl(booktabs=T,
      caption = "Model E Coefficient and Standard Error Estimates") %>%
  kable_classic()
```

Since we find that the additive model that includes Region is the best model when either the log transformation is applied to either Crimes or Crime Rate, we compare these two models by examining the diagnostic plots, AIC, BIC, and regression outputs for each of these models.

Diagnostics:

For the Crime variable model, we investigate the residual diagnostics displayed in Figure 10. We see that the residuals vs fitted value plot does not display a major vertical trend for the majority of the fitted values and the data are centered at 0., although there are some points that deviate from the pattern of the data. The Normal QQ plot suggests that the normality in error terms is violated due to the deviation of the points from the linear relationship illustrated by the qqline (standardized residuals and theoretical quantiles). It also identifies numerous potential outliers based on the values of the standardized residuals, such as observations 128, 206, and 396. The Scale Location plot illustrates evidence of many outliers since their square rooted absolute value standardized residuals are greater than 1.5, although the spread majority of the data is relatively constant and centered between 0.5 and 1, suggesting that the constant error variance assumption is not violated for the model. The residuals vs leverage plot does not identify any influential point based on having a Cook's distance value greater than 0.5, though there are some observations that are

high leverage (i.e. observation 6) or have a large standardized residual value (observation 206) that merit further investigation to determine if they should remain in the analysis.

For the model with the crime rate variable included, we examine the residual diagnostics displayed in Figure 11. We see that the residuals are roughly centered at 0 and the variance is relatively constant for all values, suggesting that the constant variance assumption is roughly satisfied; there are some points that deviate from the pattern of the data. The Normal QQ plot suggests that the normality in error terms approximately satisfied for the majority of the data, but there is some deviation in the tails as illustrated by the deviation of the points from the linear relationship illustrated by the qqline (standardized residuals and theoretical quantiles). It also identifies some potential outliers based on the values of the standardized residuals, such as observations 128, 206, and 337. The scale location plot does not show any major vertical trends and that the data is centered around 1, which confirms that the constant variance assumption is satisfied. However, there are multiple observations with square rooted absolute value standardized residuals that are greater than 1.5 that could be classified as outliers. The residuals vs. leverage plot shows no influential points based on Cook's distance, but a few observations are either highly leveraged (observation 6) or can be classified as outliers (observation 206) based on its standardized residual value. These points should be investigated to see if they should remain in the model.

We see from Table 6 that neither model explains more than 20% of the variation in the response variable, although the Crime model explains roughly 10% more of the variation in the response variable and has better measures for AIC and BIC. Additionally, the coefficients for both crime rate and crimes have signs that are the opposite of what is expected. This is likely due to omitted variable bias from not controlling for variables like population. However, the coefficient for the crime rate variable is only slightly statistically significant (unlike for the crime variable), which more aligns with our intuition since it seems unreasonable that counties with higher per capita income would also have higher crime rates. We would expect either no relationship or a negative relationship between these variables.

In summary, both models provide a similar (but relatively weak) fit for the response variable, and have similar diagnostic plots approximately showing that the regression model assumptions are satisfied. However, the coefficient estimate for crime rate better aligns with our intuition about the relationship between crime and response variable, and crime rate is on a similar scale as the response variable (per capita income). Therefore, since neither variable explains the response variable exceptionally well after accounting for region, the crime rate variable is the more appropriate variable to include in the analysis due to its interpretability. This variable will be included in the model to predict per capita income, although other variables will also need to be included to improve the predictive power of the model.

Research Question 3

Naive Full Model with continuous predictors

```
# creating data frames to be used to select predictors for the final model
idx2 <- c("doctors", "hosp.beds", "land.area", "pop", "tot.income", "crimes",
          "crime_rate", "log_inc", "log_pop", "per.cap.income", "log_crimes")

cdi_df1 <- cdi_analysis_2[,!names(cdi_analysis_2) %in% idx2]
cdi_df2 <- cdi_df1[,-which(colnames(cdi_df1) == "region")]

lm.q3a <- lm(log_pci ~ ., data = cdi_df2)
summary(lm.q3a)
```

```
##
## Call:
## lm(formula = log_pci ~ ., data = cdi_df2)
```

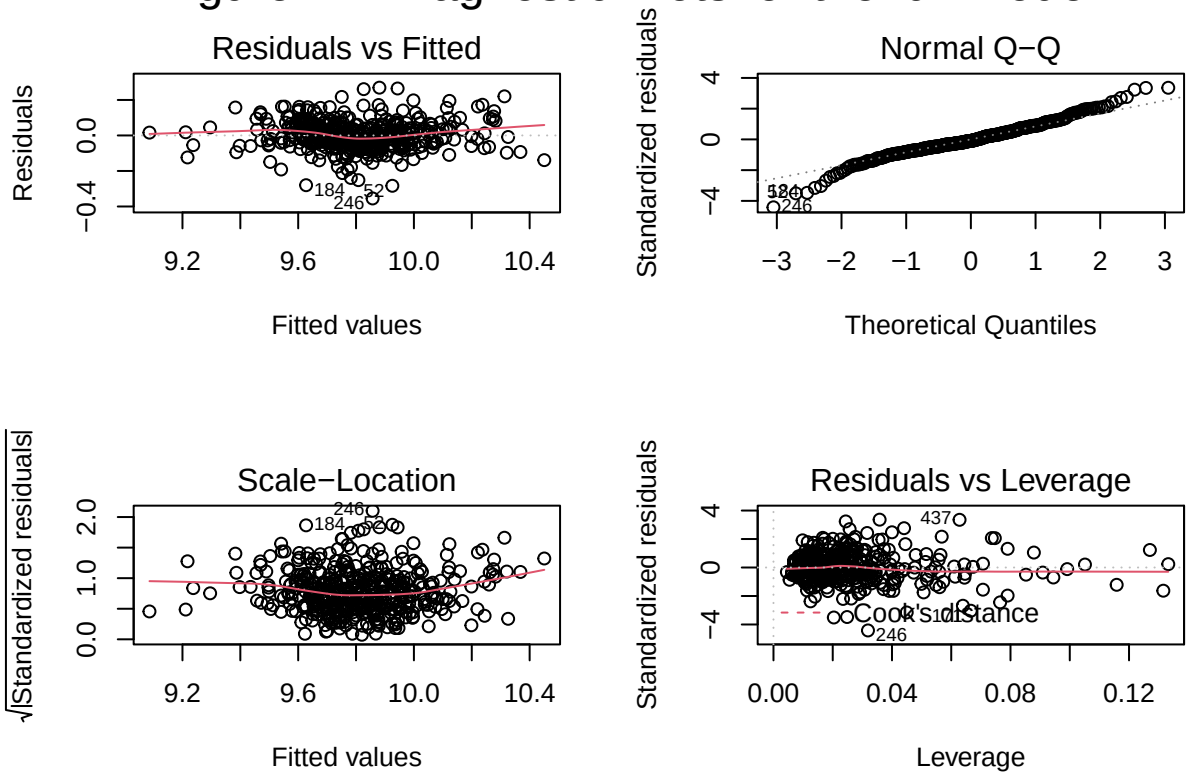
```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35474 -0.04577 -0.00794  0.04585  0.26911
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.343656   0.113345  91.258  < 2e-16 ***
## pop.18_34      -0.015613   0.001308 -11.940  < 2e-16 ***
## pop.65_plus    -0.003080   0.001328  -2.319   0.0209 *
## pct.hs.grad    -0.004755   0.001085  -4.382 1.48e-05 ***
## pct.bach.deg    0.015793   0.001019  15.495  < 2e-16 ***
## pct.below.pov  -0.025487   0.001380 -18.467  < 2e-16 ***
## pct.unemp       0.011229   0.002186   5.138 4.23e-07 ***
## log_doctors     0.047859   0.011243   4.257 2.55e-05 ***
## log_hosp.beds   0.014801   0.011908   1.243   0.2146
## log_land.area  -0.035783   0.004791  -7.469 4.55e-13 ***
## log_crime_rate  0.010047   0.009792   1.026   0.3055
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0816 on 429 degrees of freedom
## Multiple R-squared:  0.8477, Adjusted R-squared:  0.8442
## F-statistic: 238.9 on 10 and 429 DF,  p-value: < 2.2e-16
```

```
vif(lm.q3a)
```

```
##      pop.18_34    pop.65_plus    pct.hs.grad   pct.bach.deg  pct.below.pov
##       1.979952       1.853750       3.820223       4.013215       2.723069
##      pct.unemp    log_doctors  log_hosp.beds  log_land.area log_crime_rate
##       1.721429      10.906872       9.410985       1.149748       1.600834
```

```
par(mfrow=c(2,2))
plot(lm.q3a)
mtext("Figure 12: Diagnostic Plots for the full model",
      side = 3, line = -2, outer = TRUE, cex = 1.5)
```

Figure 12: Diagnostic Plots for the full model




```
par(mfrow=c(1,1))
```

Based on the results from research questions 1 and 2, we apply transformations to the Per Capital Income, Doctors, Hospital Beds, Land Area, and Crime Rate variables. We also remove the Population and Total Income variables due to their deterministic functional relationship with the response variable. The Region variable is temporarily removed for variable selection purposes, and will be added back once a model has been chosen using variable selection techniques.

Fitting all the variables in the multiple linear regression model, we see that some of the coefficients for the predictor variables are not statistically significant, and their VIFs show that there is multicollinearity present among the predictors. Figure 12 shows that while the model roughly satisfies the constant variance assumption, there is evidence of deviation from the normal distribution in the tails based on the Normal QQ plot. High leverage points and outliers are also present in the data.

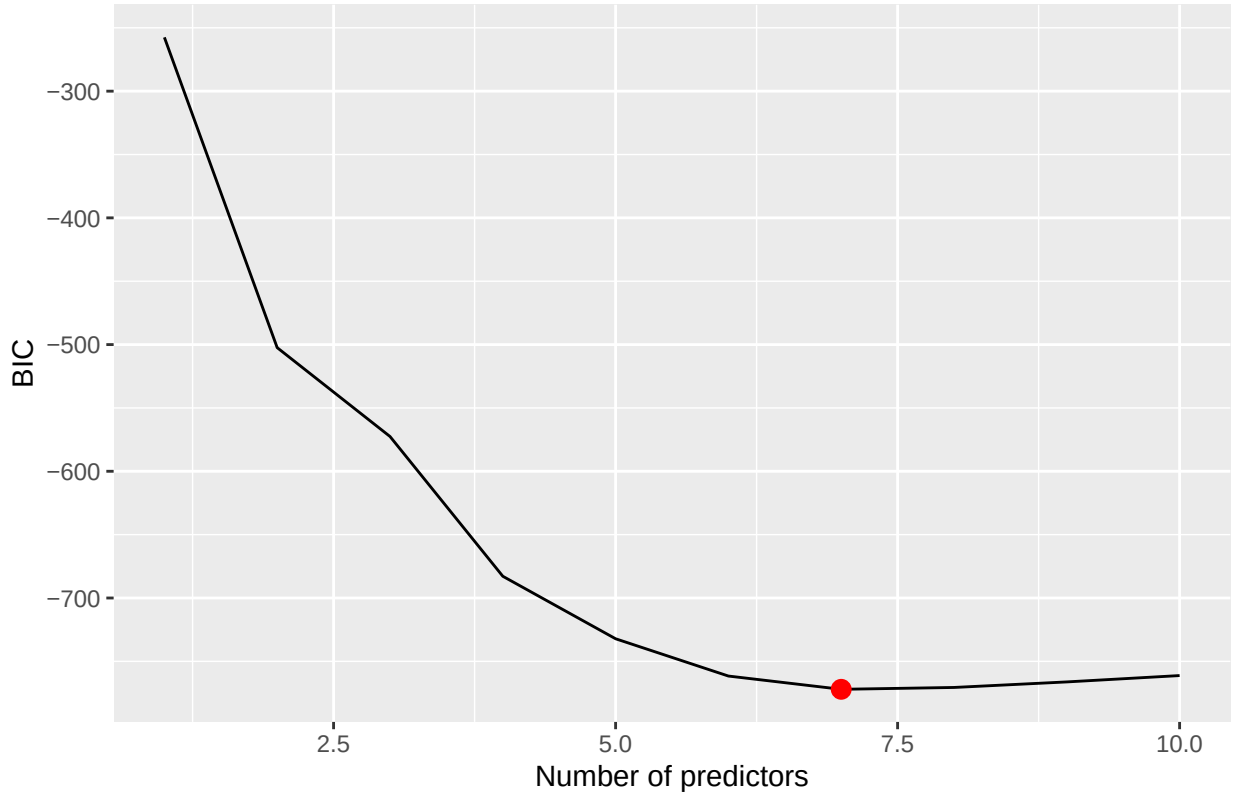
The results suggests that the model roughly satisfies the regression model assumptions, but we will further investigate subsets of predictors using variable selection to find the best subset that does not suffer from multicollinearity and still satisfies the regression model assumptions.

Variable Selection Technique: All Subsets Regression

```
# variable selection
lm.q3a <- leaps::regsubsets(log_pci ~ ., data = cdi_df2, nvmax = 10)
tibble(x = 1:10, y = summary(lm.q3a)$bic) %>%
  ggplot(aes(x = x, y = y)) +
  geom_line() +
```

```
  labs(x = "Number of predictors", y = "BIC",
       title = "Figure 13: BIC values for All Subsets selection method") +
  annotate("point", y = min(summary(lm.q3a)$bic), x = which.min(summary(lm.q3a)$bic), colour = "red", si
```

Figure 13: BIC values for All Subsets selection method




```
# generating the best model
summary(lm.q3a)$which[which.min(summary(lm.q3a)$bic), ]
```

```
##    (Intercept)      pop.18_34    pop.65_plus    pct.hs.grad   pct.bach.deg
##           TRUE           TRUE          FALSE           TRUE           TRUE
##  pct.below.pov      pct.unemp    log_doctors  log_hosp.beds  log_land.area
##           TRUE           TRUE           TRUE          FALSE           TRUE
## log_crime_rate
##          FALSE
```

```
coef(lm.q3a, which.min(summary(lm.q3a)$bic))
```

```
##   (Intercept)     pop.18_34   pct.hs.grad  pct.bach.deg pct.below.pov
##  10.222495041  -0.013900201  -0.004406396   0.015385301  -0.024278371
##     pct.unemp   log_doctors log_land.area
##   0.010603691   0.060676872  -0.035674062
```

```
# refitting the model to get the minimum standard errors
lm.q3a_fit <- lm(log_pci ~ . - pop.65_plus - log_hosp.beds - log_crime_rate, data = cdi_df2)
summary(lm.q3a_fit)$coef
```

```
##                   Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)   10.222495041 0.0931210074 109.776465 1.127483e-317
## pop.18_34     -0.013900201 0.0011113007 -12.508046  7.514862e-31
## pct.hs.grad   -0.004406396 0.0010822796  -4.071403  5.558448e-05
## pct.bach.deg   0.015385301 0.0009245509  16.640838  2.100590e-48
## pct.below.pov -0.024278371 0.0012583372 -19.294011  2.812246e-60
## pct.unemp      0.010603691 0.0021771148   4.870525  1.564524e-06
## log_doctors    0.060676872 0.0040183327  15.100012  1.133432e-41
## log_land.area -0.035674062 0.0047767371  -7.468291  4.533156e-13
```

From the All Subsets Regression variable selection technique, we see that following variables are selected: pop.18_34, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, log_doctors, and log_land.area.

Variable Selection Technique: Stepwise Regression (AIC and BIC)

```
lm.q3_base <- lm(log_pci ~ ., data = cdi_df2)
lm.q3b <- stepAIC(lm.q3_base, direction = "both", k = 2)
```

```
anova(lm.q3_base, lm.q3b)
```

```
## Analysis of Variance Table
##
## Model 1: log_pci ~ pop.18_34 + pop.65_plus + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log_doctors + log_hosp.beds +
##     log_land.area + log_crime_rate
## Model 2: log_pci ~ pop.18_34 + pop.65_plus + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log_doctors + log_land.area
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    429 2.8566
## 2    431 2.8748 -2 -0.018169 1.3643 0.2567
```

```
names(coef(lm.q3b))
```

```
## [1] "(Intercept)"   "pop.18_34"     "pop.65_plus"   "pct.hs.grad"
## [5] "pct.bach.deg"  "pct.below.pov" "pct.unemp"     "log_doctors"
## [9] "log_land.area"
```

```
summary(lm.q3b)$coef
```

```
##                   Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)   10.315966592 0.1025857893 100.559411 2.359405e-301
## pop.18_34     -0.015348817 0.0012987646 -11.818014  4.136902e-28
## pop.65_plus   -0.002766377 0.0012977992  -2.131591  3.360555e-02
## pct.hs.grad   -0.004657948 0.0010843088  -4.295776  2.153275e-05
## pct.bach.deg   0.015214937 0.0009242442  16.462032  1.361311e-47
## pct.below.pov -0.024614405 0.0012630840 -19.487544  4.083797e-61
## pct.unemp      0.010768825 0.0021696234   4.963454  9.990989e-07
## log_doctors    0.062605267 0.0041029328  15.258663  2.438771e-42
## log_land.area -0.036493494 0.0047727720  -7.646184  1.360706e-13
```

From the Stepwise Regression AIC variable selection technique, we see that following variables are selected: pop.18_34, pop.65_plus, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, log_doctors, and log_land.area.

```
lm.q3c <- stepAIC(lm.q3_base, direction = "both", k = log(dim(cdi_df2)[1]))
```

```
anova(lm.q3_base, lm.q3c)
```

```
## Analysis of Variance Table
##
## Model 1: log_pci ~ pop.18_34 + pop.65_plus + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log_doctors + log_hosp.beds +
##     log_land.area + log_crime_rate
## Model 2: log_pci ~ pop.18_34 + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp + log_doctors + log_land.area
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    429 2.8566
## 2    432 2.9051 -3 -0.048475 2.4267  0.065 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
names(coef(lm.q3c))
```

```
## [1] "(Intercept)"   "pop.18_34"     "pct.hs.grad"   "pct.bach.deg"
## [5] "pct.below.pov" "pct.unemp"     "log_doctors"   "log_land.area"
```

```
summary(lm.q3c)$coef
```

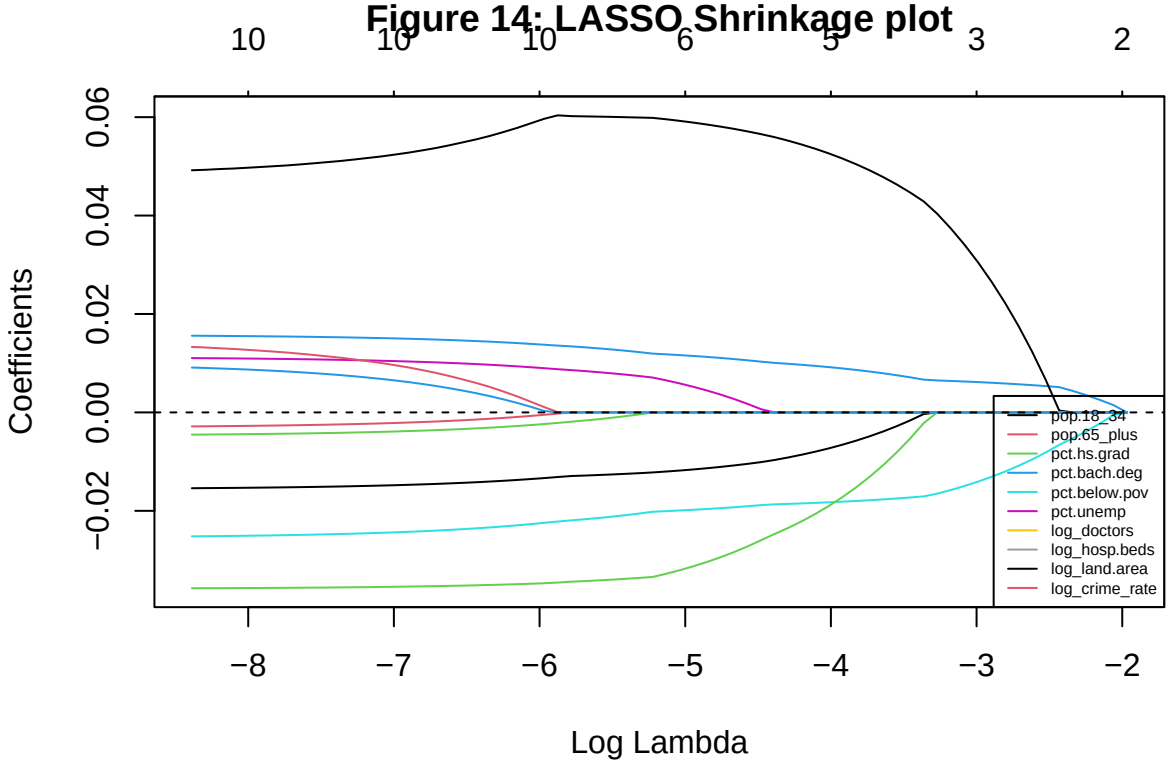
```
##                   Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)   10.222495041 0.0931210074 109.776465 1.127483e-317
## pop.18_34     -0.013900201 0.0011113007 -12.508046  7.514862e-31
## pct.hs.grad   -0.004406396 0.0010822796  -4.071403  5.558448e-05
## pct.bach.deg   0.015385301 0.0009245509  16.640838  2.100590e-48
## pct.below.pov -0.024278371 0.0012583372 -19.294011  2.812246e-60
## pct.unemp      0.010603691 0.0021771148   4.870525  1.564524e-06
## log_doctors    0.060676872 0.0040183327  15.100012  1.133432e-41
## log_land.area -0.035674062 0.0047767371  -7.468291  4.533156e-13
```

From the Stepwise Regression BIC variable selection technique, we see that following variables are selected: pop.18_34, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, log_doctors, and log_land.area.

Variable Selection Technique: LASSO

```
cdi_mat <- as.matrix(cdi_df2[,-1])

# LASSO without cross-validation
lasso <- glmnet(cdi_mat, cdi_df2[,1], alpha=1)
Xnames <- dimnames(cdi_mat)[[2]]

plot(lasso,xvar="lambda", main = "Figure 14: LASSO Shrinkage plot")
abline(h=0,lty=2)
legend('bottomright',lty=1,col=1:length(Xnames),legend=Xnames,cex=0.5)
```

Figure 14: LASSO Shrinkage plot




```
# LASSO with cross-validation
set.seed(20)
lasso_cv <- glmnet::cv.glmnet(cdi_mat, cdi_df2[,1],alpha=1)

c(lambda.1se=lasso_cv$lambda.1se,lambda.min=lasso_cv$lambda.min)
```

```
##   lambda.1se   lambda.min
## 0.0044721426 0.0002278171
```

```
lasso_mat <- cbind(coef(lasso_cv,s=lasso_cv$lambda.min), coef(lasso_cv,s=lasso_cv$lambda.1se))
dimnames(lasso_mat)[[2]] <- c("lambda(minMSE)","lambda(minMSE+1se)")

lasso_mat
```

```
## 11 x 2 sparse Matrix of class "dgCMatrix"
##                lambda(minMSE) lambda(minMSE+1se)
## (Intercept)      10.317582891       9.9409568402
## pop.18_34        -0.015401734      -0.0124830611
## pop.65_plus      -0.002842193       .
## pct.hs.grad      -0.004512574      -0.0007592586
## pct.bach.deg      0.015585598       0.0125356027
## pct.below.pov    -0.025186168      -0.0208825185
## pct.unemp         0.011038990       0.0076826679
## log_doctors       0.049207126       0.0599889248
## log_hosp.beds     0.013309126       .
```

```
## log_land.area    -0.035726978      -0.0338113848
## log_crime_rate    0.009115750       .
```

Figure 14 does not display an obvious place to cut off the shrinkage plot and select predictor variables, so we utilize cross validation to select the appropriate lambda value for LASSO regression. From the cross-validation LASSO regression (utilizing the minimum lambda value plus 1 standard error to avoid capitalization on chance), we see that following variables are selected: pop.18_34, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, log_doctors, and log_land.area.

Comparing Models from variable selection

```
# assigning objects of interest to new names for consistency with the written analysis
model1 <- lm.q3a_fit
model2 <- lm.q3b

summary(model1)
```

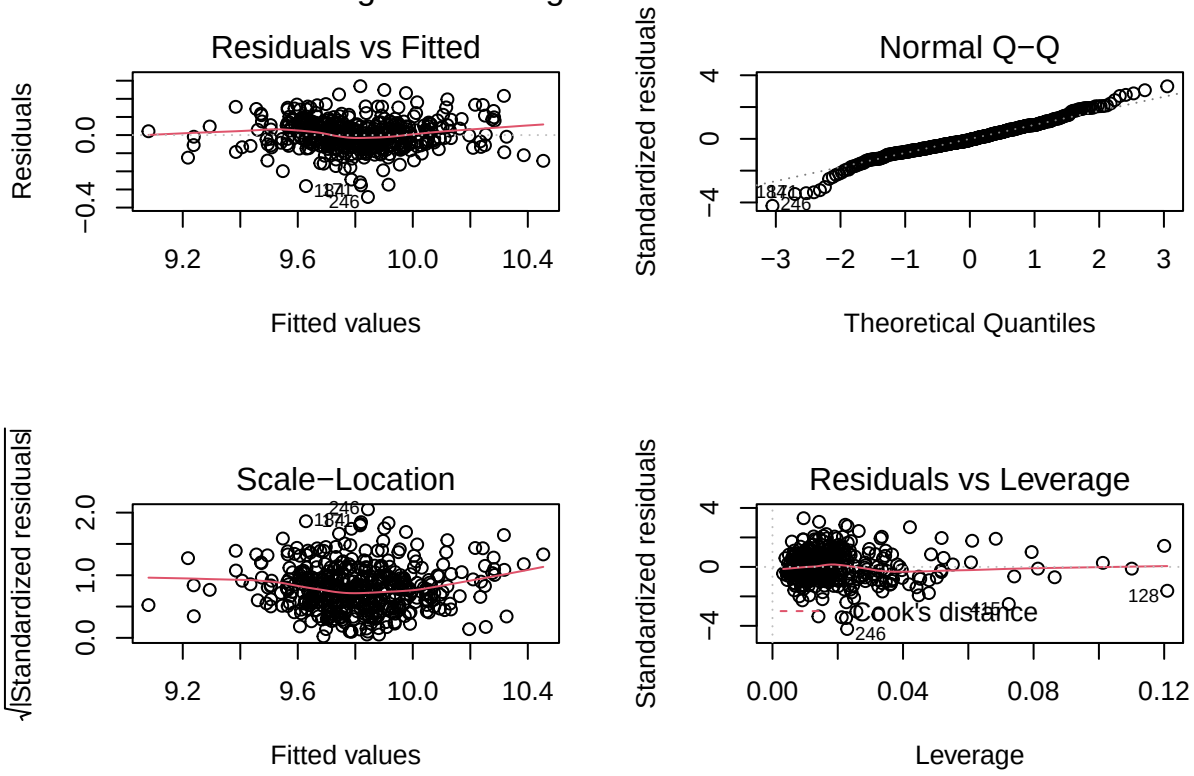
```
##
## Call:
## lm(formula = log_pci ~ . - pop.65_plus - log_hosp.beds - log_crime_rate,
##     data = cdi_df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34147 -0.04886 -0.00538  0.04818  0.26969
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.2224950  0.0931210 109.776  < 2e-16 ***
## pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
## pct.hs.grad   -0.0044064  0.0010823  -4.071 5.56e-05 ***
## pct.bach.deg   0.0153853  0.0009246  16.641  < 2e-16 ***
## pct.below.pov -0.0242784  0.0012583 -19.294  < 2e-16 ***
## pct.unemp      0.0106037  0.0021771   4.871 1.56e-06 ***
## log_doctors    0.0606769  0.0040183  15.100  < 2e-16 ***
## log_land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.082 on 432 degrees of freedom
## Multiple R-squared:  0.8452, Adjusted R-squared:  0.8427
## F-statistic: 336.9 on 7 and 432 DF,  p-value: < 2.2e-16
```

```
vif(model1)
```

```
##     pop.18_34   pct.hs.grad  pct.bach.deg pct.below.pov     pct.unemp
##      1.416145      3.763103      3.269565      2.241555      1.691280
##   log_doctors log_land.area
##      1.379671      1.131867
```

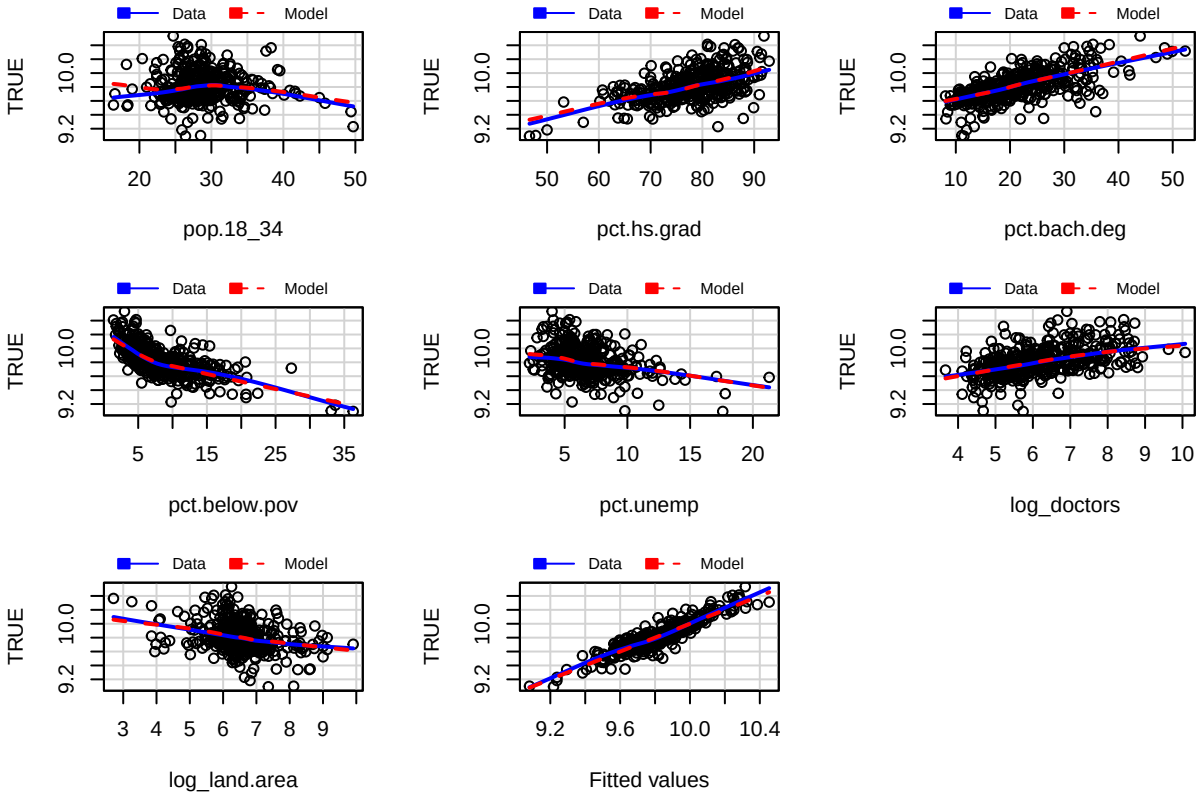
```
par(mfrow=c(2,2))
plot(lm.q3a_fit)
mtext("Figure 15: Diagnostic Plots for Model 1", side = 3, line = -2, outer = TRUE, cex = 1.1)
```

Figure 15: Diagnostic Plots for Model 1




```
par(mfrow=c(1,1))
mmps(model1, main = "Figure 16: Marginal model plots for Model 1")
```

Figure 16: Marginal model plots for Model 1




```
summary(model2)
```

```
##
## Call:
## lm(formula = log_pci ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + log_doctors +
##     log_land.area, data = cdi_df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35756 -0.04551 -0.00543  0.04844  0.27399
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.3159666  0.1025858 100.559  < 2e-16 ***
## pop.18_34     -0.0153488  0.0012988 -11.818  < 2e-16 ***
## pop.65_plus   -0.0027664  0.0012978  -2.132   0.0336 *
## pct.hs.grad   -0.0046579  0.0010843  -4.296 2.15e-05 ***
## pct.bach.deg   0.0152149  0.0009242  16.462  < 2e-16 ***
## pct.below.pov -0.0246144  0.0012631 -19.488  < 2e-16 ***
## pct.unemp      0.0107688  0.0021696   4.963 9.99e-07 ***
## log_doctors    0.0626053  0.0041029  15.259  < 2e-16 ***
## log_land.area -0.0364935  0.0047728  -7.646 1.36e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

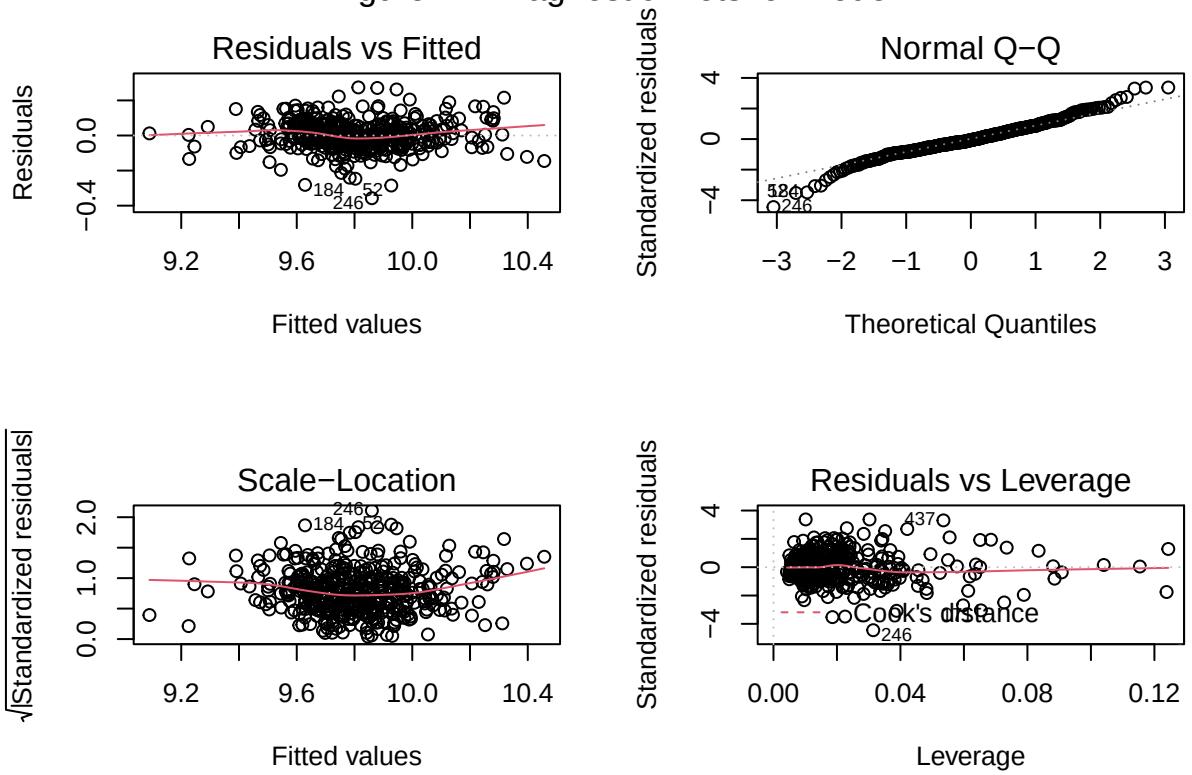
```
##
## Residual standard error: 0.08167 on 431 degrees of freedom
## Multiple R-squared:  0.8468, Adjusted R-squared:  0.8439
## F-statistic: 297.7 on 8 and 431 DF,  p-value: < 2.2e-16
```

```
vif(model2)
```

```
##     pop.18_34   pop.65_plus   pct.hs.grad  pct.bach.deg pct.below.pov
##      1.950084      1.767181      3.808211      3.294199      2.277025
##     pct.unemp   log_doctors log_land.area
##      1.693439      1.450175      1.139258
```

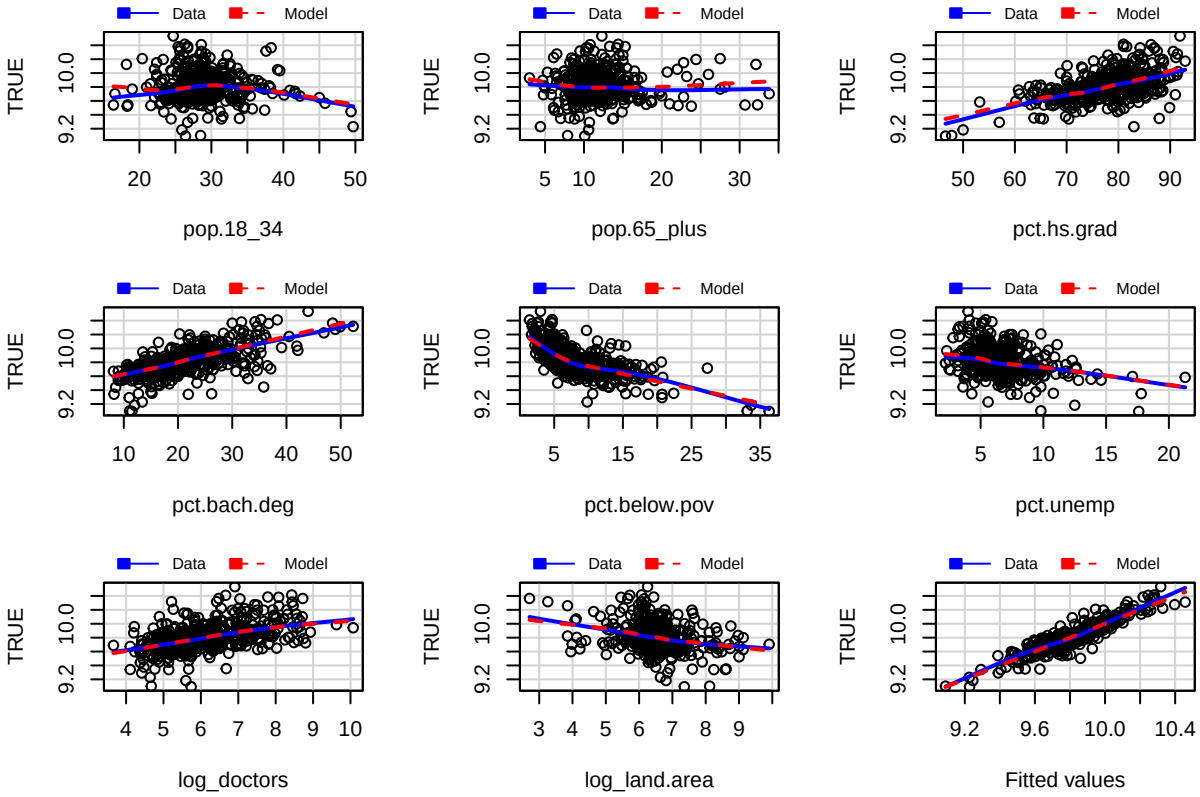
```
par(mfrow=c(2,2))
plot(model2)
mtext("Figure 17: Diagnostic Plots for Model 2", side = 3, line = -2, outer = TRUE, cex = 1.1)
```

Figure 17: Diagnostic Plots for Model 2




```
par(mfrow=c(1,1))
mmps(model2, main = "Figure 18: Marginal model plots for Model 1")
```

Figure 18: Marginal model plots for Model 1



Results of Variable Selection Methods Examining the output, we see that the All Subsets, Stepwise BIC, and LASSO (using the model with lambda that is 1 standard error larger than the minimum lambda value found) regression techniques select the same model using the predictor variables: pop.18_34, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, log_doctors, and log_land.area. This will be referred to as Model 1. All of the coefficients in the model are statistically significant, the adjusted R^2 is approximately 84%, and the VIFs of the coefficients are all less than 5, implying multicollinearity is not present within the predictors. Additionally, the diagnostic plots in Figure 15 are similar to the full model in that the regression model assumptions are approximately satisfied with the exception of Normal QQ plot, since the deviation from the QQ line implies the tails are slightly longer than those of the Normal Distribution. The marginal model plots in Figure 16 also show that the appropriate form of the predictor variables are included since the non-parametric data line and model line trend closely together for each of the predictor variables and the fitted values.

We see similar results when we examining the selected Stepwise AIC model, finding that it includes identical predictors from Model 1 along with the Percentage of population 65 or older. This will be referred to as Model 2. Similar to model 1, all of the coefficients in the model are statistically significant, the adjusted R^2 is approximately 84%, and the VIFs of the coefficients are all less than 5, implying multicollinearity is not present within the predictors. The diagnostic plots in Figure 17 are consistent with the full model in that the regression model assumptions are approximately satisfied with the exception of Normal QQ plot, since the deviation from the QQ line implies the tails are slightly longer than those of the Normal Distribution. The marginal model from Figure 18 plots illustrate that the appropriate form of the predictor variables are specified in the model since since the non-parametric data line and model line trend closely together for each of the predictor variables and the fitted values.

Since these models are close to identical and therefore have similar interpretations for the beta coefficients, we consider interaction terms for both models before making our determination about which model to use.

Adding interaction terms for region variables

```
# adding interaction terms for models 1 and 2
idx3 <- c("log_pci", "pop.18_34", "pct.hs.grad", "pct.bach.deg", "pct.below.pov",
          "pct.unemp", "log_doctors", "log_land.area", "region")
cdi_region1 <- cdi_df1[, names(cdi_df1) %in% idx3]
model1_region <- lm(log_pci ~ .*region, data = cdi_region1)
summary(model1_region)
```

```
##
## Call:
## lm(formula = log_pci ~ . * region, data = cdi_region1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.250782 -0.042332 -0.002298  0.040559  0.313570
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             10.1244260  0.2826240  35.823  < 2e-16 ***
## pop.18_34               -0.0147940  0.0026043  -5.681 2.55e-08 ***
## pct.hs.grad             -0.0024773  0.0034110  -0.726 0.468088
## pct.bach.deg             0.0140833  0.0029254   4.814 2.09e-06 ***
## pct.below.pov           -0.0237085  0.0036234  -6.543 1.81e-10 ***
## pct.unemp                0.0180393  0.0048923   3.687 0.000257 ***
## regionNE                 0.3243992  0.3577081   0.907 0.365004
## regionS                 -0.0345856  0.3131668  -0.110 0.912116
## regionW                  1.5043946  0.4226868   3.559 0.000416 ***
## log_doctors              0.0544169  0.0093221   5.837 1.08e-08 ***
## log_land.area           -0.0364187  0.0151355  -2.406 0.016564 *
## pop.18_34:regionNE      -0.0024780  0.0036873  -0.672 0.501939
## pop.18_34:regionS       -0.0008777  0.0030680  -0.286 0.774970
## pop.18_34:regionW        0.0014122  0.0040925   0.345 0.730220
## pct.hs.grad:regionNE    -0.0037529  0.0044150  -0.850 0.395813
## pct.hs.grad:regionS      0.0021198  0.0037853   0.560 0.575790
## pct.hs.grad:regionW     -0.0190188  0.0045881  -4.145 4.13e-05 ***
## pct.bach.deg:regionNE    0.0069429  0.0040312   1.722 0.085776 .
## pct.bach.deg:regionS    -0.0015774  0.0032000  -0.493 0.622328
## pct.bach.deg:regionW     0.0071026  0.0036374   1.953 0.051541 .
## pct.below.pov:regionNE -0.0014134  0.0050896  -0.278 0.781381
## pct.below.pov:regionS    0.0072764  0.0040739   1.786 0.074827 .
## pct.below.pov:regionW   -0.0161639  0.0054271  -2.978 0.003071 **
## pct.unemp:regionNE      -0.0083596  0.0073758  -1.133 0.257720
## pct.unemp:regionS       -0.0249396  0.0065867  -3.786 0.000176 ***
## pct.unemp:regionW       -0.0201466  0.0067713  -2.975 0.003101 **
## regionNE:log_doctors    -0.0046251  0.0132571  -0.349 0.727359
## regionS:log_doctors      0.0043337  0.0114401   0.379 0.705019
## regionW:log_doctors     -0.0034863  0.0131576  -0.265 0.791173
## regionNE:log_land.area -0.0037179  0.0201435  -0.185 0.853656
## regionS:log_land.area  -0.0047582  0.0174155  -0.273 0.784825
## regionW:log_land.area   0.0151234  0.0181871   0.832 0.406154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.0759 on 408 degrees of freedom
## Multiple R-squared:  0.8747, Adjusted R-squared:  0.8652
## F-statistic: 91.91 on 31 and 408 DF,  p-value: < 2.2e-16
```

```
# include interactions for pct.hs.grad, pct.below.pov, pct.unemp
model1_region_interactions <- update(model1_region, . ~ . -
                                region:log_land.area - region:pop.18_34 -
                                region:log_doctors - region:pct.bach.deg)
summary(model1_region_interactions)
```

```
##
## Call:
## lm(formula = log_pci ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + region + log_doctors + log_land.area +
##     pct.hs.grad:region + pct.below.pov:region + pct.unemp:region,
##     data = cdi_region1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.294186 -0.043597 -0.001583  0.037667  0.311609
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.2421239  0.2176557  47.057  < 2e-16 ***
## pop.18_34              -0.0149347  0.0010897 -13.705  < 2e-16 ***
## pct.hs.grad            -0.0043532  0.0024515  -1.776 0.076501 .
## pct.bach.deg            0.0156310  0.0009715  16.090  < 2e-16 ***
## pct.below.pov          -0.0252029  0.0032612  -7.728 8.12e-14 ***
## pct.unemp               0.0197400  0.0046254   4.268 2.44e-05 ***
## regionNE               -0.0520070  0.2707173  -0.192 0.847750
## regionS                -0.0389718  0.2383516  -0.164 0.870199
## regionW                 1.3910484  0.3408962   4.081 5.38e-05 ***
## log_doctors             0.0572284  0.0040082  14.278  < 2e-16 ***
## log_land.area          -0.0381738  0.0053996  -7.070 6.51e-12 ***
## pct.hs.grad:regionNE    0.0017684  0.0029293   0.604 0.546374
## pct.hs.grad:regionS     0.0011525  0.0025618   0.450 0.653024
## pct.hs.grad:regionW    -0.0141473  0.0035826  -3.949 9.20e-05 ***
## pct.below.pov:regionNE -0.0015170  0.0046143  -0.329 0.742493
## pct.below.pov:regionS   0.0070185  0.0035199   1.994 0.046808 *
## pct.below.pov:regionW  -0.0137920  0.0051811  -2.662 0.008066 **
## pct.unemp:regionNE     -0.0129841  0.0070423  -1.844 0.065929 .
## pct.unemp:regionS      -0.0231138  0.0061365  -3.767 0.000189 ***
## pct.unemp:regionW      -0.0217357  0.0065225  -3.332 0.000937 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07692 on 420 degrees of freedom
## Multiple R-squared:  0.8675, Adjusted R-squared:  0.8615
## F-statistic: 144.8 on 19 and 420 DF,  p-value: < 2.2e-16
```

```
anova(model1, model1_region_interactions)
```

```
## Analysis of Variance Table
```

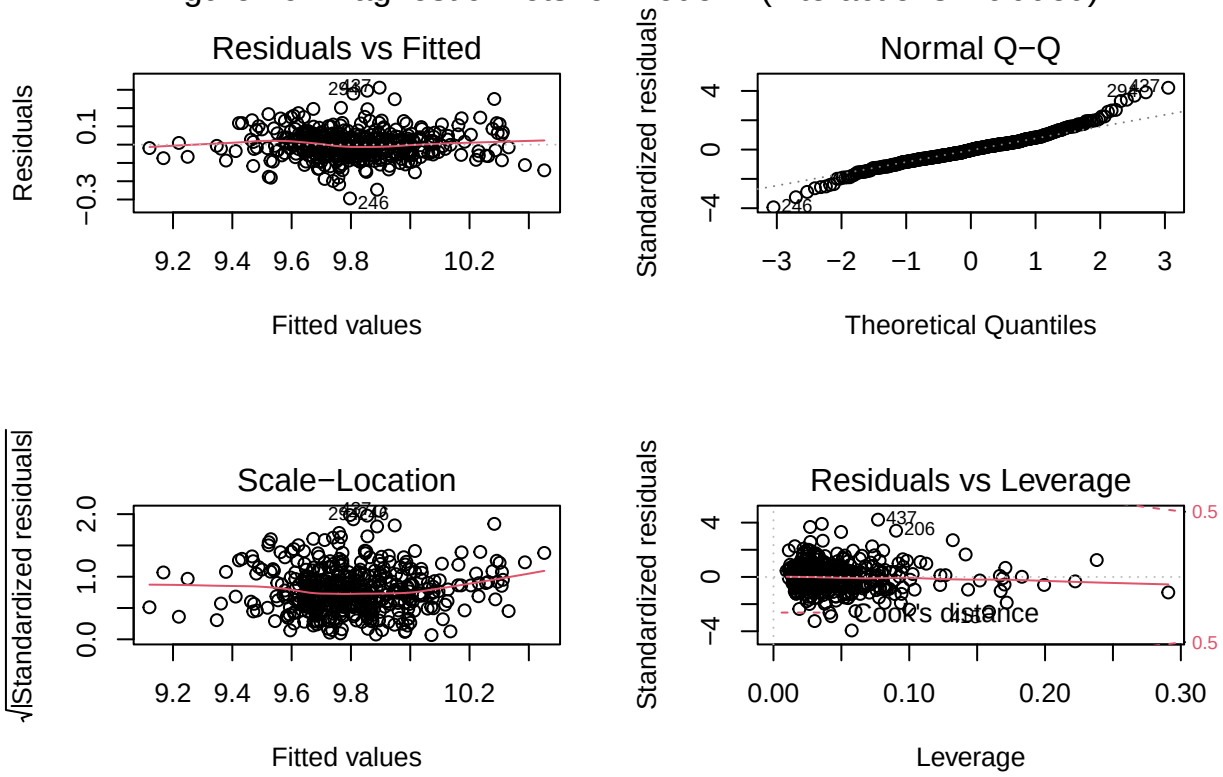
```
##
## Model 1: log_pci ~ (pop.18_34 + pop.65_plus + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log_doctors + log_hosp.beds +
##     log_land.area + log_crime_rate) - pop.65_plus - log_hosp.beds -
##     log_crime_rate
## Model 2: log_pci ~ pop.18_34 + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp + region + log_doctors + log_land.area + pct.hs.grad:region +
##     pct.below.pov:region + pct.unemp:region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    432 2.9051
## 2    420 2.4853 12   0.41978 5.9117 1.555e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vif(model1_region_interactions)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## pop.18_34            1.547481e+00  1        1.243978
## pct.hs.grad          2.194177e+01  1        4.684205
## pct.bach.deg         4.102307e+00  1        2.025415
## pct.below.pov        1.710982e+01  1        4.136402
## pct.unemp            8.675528e+00  1        2.945425
## region               2.454546e+08  3       25.022374
## log_doctors          1.559981e+00  1        1.248992
## log_land.area        1.643605e+00  1        1.282032
## pct.hs.grad:region   8.506975e+07  3       20.971486
## pct.below.pov:region 5.278685e+03  3        4.172736
## pct.unemp:region     1.108865e+04  3        4.722222
```

```
par(mfrow=c(2,2))
plot(model1_region_interactions)
mtext("Figure 19: Diagnostic Plots for Model 1 (interactions included)",
      side = 3, line = -2, outer = TRUE, cex = 1.1)
```

Figure 19: Diagnostic Plots for Model 1 (interactions included)




```
par(mfrow=c(1,1))


idx4 <- c("log_pci", "pop.18_34", "pop.65_plus","pct.hs.grad", "pct.bach.deg",
          "pct.below.pov", "pct.unemp", "log_doctors", "log_land.area", "region")
cdi_region2 <- cdi_df1[, names(cdi_df1) %in% idx4]
model2_region <- lm(log_pci ~ .*region, data = cdi_region2)
summary(model2_region)
```

```
##
## Call:
## lm(formula = log_pci ~ . * region, data = cdi_region2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.239497 -0.042518 -0.002899  0.038705  0.315955
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.1550994  0.3077758  32.995  < 2e-16 ***
## pop.18_34              -0.0150740  0.0028317  -5.323 1.69e-07 ***
## pop.65_plus            -0.0012483  0.0050165  -0.249 0.803614
## pct.hs.grad            -0.0026649  0.0034861  -0.764 0.445055
## pct.bach.deg            0.0140191  0.0029305   4.784 2.41e-06 ***
## pct.below.pov          -0.0233702  0.0038627  -6.050 3.30e-09 ***
```

```
## pct.unemp               0.0176067  0.0051819   3.398 0.000747 ***
## regionNE                0.4813749  0.3863061   1.246 0.213451
## regionS                -0.0552517  0.3396107  -0.163 0.870843
## regionW                 1.3969067  0.4575796   3.053 0.002417 **
## log_doctors             0.0548293  0.0094485   5.803 1.32e-08 ***
## log_land.area          -0.0355230  0.0155258  -2.288 0.022654 *
## pop.18_34:regionNE     -0.0060991  0.0042036  -1.451 0.147582
## pop.18_34:regionS      -0.0008273  0.0034566  -0.239 0.810970
## pop.18_34:regionW       0.0030516  0.0048005   0.636 0.525342
## pop.65_plus:regionNE   -0.0076628  0.0063347  -1.210 0.227119
## pop.65_plus:regionS     0.0009166  0.0052822   0.174 0.862326
## pop.65_plus:regionW     0.0037008  0.0064632   0.573 0.567239
## pct.hs.grad:regionNE   -0.0033331  0.0044706  -0.746 0.456373
## pct.hs.grad:regionS     0.0023152  0.0038518   0.601 0.548134
## pct.hs.grad:regionW    -0.0185423  0.0046646  -3.975 8.33e-05 ***
## pct.bach.deg:regionNE   0.0060237  0.0040533   1.486 0.138025
## pct.bach.deg:regionS   -0.0015550  0.0032102  -0.484 0.628384
## pct.bach.deg:regionW    0.0069577  0.0036552   1.903 0.057687 .
## pct.below.pov:regionNE -0.0009949  0.0052677  -0.189 0.850294
## pct.below.pov:regionS   0.0068718  0.0042992   1.598 0.110736
## pct.below.pov:regionW  -0.0167523  0.0055989  -2.992 0.002941 **
## pct.unemp:regionNE     -0.0063048  0.0075950  -0.830 0.406962
## pct.unemp:regionS      -0.0243492  0.0068439  -3.558 0.000418 ***
## pct.unemp:regionW      -0.0192087  0.0070270  -2.734 0.006541 **
## regionNE:log_doctors    0.0001267  0.0135190   0.009 0.992526
## regionS:log_doctors     0.0042557  0.0116550   0.365 0.715198
## regionW:log_doctors    -0.0046667  0.0132947  -0.351 0.725759
## regionNE:log_land.area -0.0050730  0.0204207  -0.248 0.803932
## regionS:log_land.area  -0.0058664  0.0177783  -0.330 0.741589
## regionW:log_land.area   0.0136894  0.0185229   0.739 0.460306
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07573 on 404 degrees of freedom
## Multiple R-squared:  0.8765, Adjusted R-squared:  0.8658
## F-statistic: 81.92 on 35 and 404 DF,  p-value: < 2.2e-16
```

```
# include interactions for pct.hs.grad, pct.below.pov, pct.unemp
model2_region_interactions <- update(model2_region, . ~ . -
                                region:log_land.area - region:pop.18_34 -
                                region:log_doctors - region:pct.bach.deg)
summary(model2_region_interactions)
```

```
##
## Call:
## lm(formula = log_pci ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + region + log_doctors +
##     log_land.area + pop.65_plus:region + pct.hs.grad:region +
##     pct.below.pov:region + pct.unemp:region, data = cdi_region2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.296590 -0.043466 -0.002885  0.037861  0.306999
##
```

```
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             10.2648646  0.2554925  40.177  < 2e-16 ***
## pop.18_34               -0.0154990  0.0012989 -11.932  < 2e-16 ***
## pop.65_plus             -0.0010923  0.0044591  -0.245 0.806616
## pct.hs.grad             -0.0042827  0.0026464  -1.618 0.106348
## pct.bach.deg             0.0154781  0.0009776  15.832  < 2e-16 ***
## pct.below.pov           -0.0247649  0.0033948  -7.295 1.52e-12 ***
## pct.unemp                0.0191862  0.0047300   4.056 5.96e-05 ***
## regionNE                 0.0688649  0.3090596   0.223 0.823784
## regionS                 -0.0757821  0.2761089  -0.274 0.783864
## regionW                  1.3795407  0.3711718   3.717 0.000229 ***
## log_doctors              0.0579753  0.0041166  14.083  < 2e-16 ***
## log_land.area           -0.0383115  0.0054021  -7.092 5.72e-12 ***
## pop.65_plus:regionNE    -0.0053560  0.0052010  -1.030 0.303700
## pop.65_plus:regionS      0.0013909  0.0045027   0.309 0.757554
## pop.65_plus:regionW      0.0008047  0.0054499   0.148 0.882686
## pct.hs.grad:regionNE     0.0010360  0.0031422   0.330 0.741792
## pct.hs.grad:regionS      0.0014490  0.0027996   0.518 0.605020
## pct.hs.grad:regionW     -0.0141288  0.0037424  -3.775 0.000183 ***
## pct.below.pov:regionNE -0.0016883  0.0046938  -0.360 0.719255
## pct.below.pov:regionS    0.0070531  0.0036916   1.911 0.056748 .
## pct.below.pov:regionW   -0.0141419  0.0052507  -2.693 0.007360 **
## pct.unemp:regionNE      -0.0111462  0.0071677  -1.555 0.120696
## pct.unemp:regionS       -0.0235386  0.0063571  -3.703 0.000242 ***
## pct.unemp:regionW       -0.0212297  0.0066069  -3.213 0.001414 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07688 on 416 degrees of freedom
## Multiple R-squared:  0.8689, Adjusted R-squared:  0.8617
## F-statistic: 119.9 on 23 and 416 DF,  p-value: < 2.2e-16
```

```
anova(model2, model2_region_interactions)
```

```
## Analysis of Variance Table
##
## Model 1: log_pci ~ pop.18_34 + pop.65_plus + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log_doctors + log_land.area
## Model 2: log_pci ~ pop.18_34 + pop.65_plus + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + region + log_doctors + log_land.area +
##     pop.65_plus:region + pct.hs.grad:region + pct.below.pov:region +
##     pct.unemp:region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    431 2.8748
## 2    416 2.4591 15   0.41567 4.6878 2.311e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

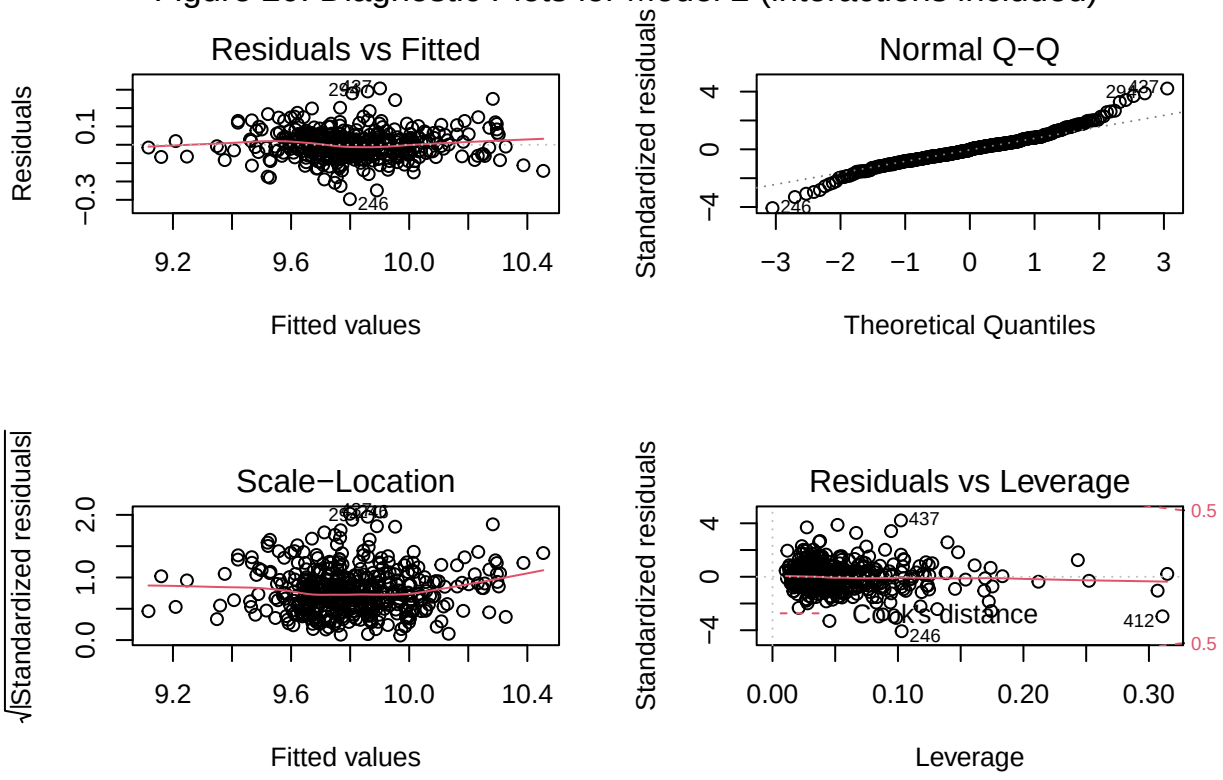
```
vif(model2_region_interactions)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## pop.18_34            2.200948e+00  1        1.483559
```

```
## pop.65_plus           2.353980e+01  1        4.851783
## pct.hs.grad           2.559501e+01  1        5.059151
## pct.bach.deg          4.158740e+00  1        2.039299
## pct.below.pov         1.855937e+01  1        4.308059
## pct.unemp             9.081772e+00  1        3.013598
## region                4.083716e+08  3       27.238044
## log_doctors           1.647251e+00  1        1.283453
## log_land.area         1.646810e+00  1        1.283281
## pop.65_plus:region    2.326687e+04  3        5.343055
## pct.hs.grad:region    1.096598e+08  3       21.878013
## pct.below.pov:region 6.011332e+03  3        4.264111
## pct.unemp:region      1.255322e+04  3        4.820875
```

```
par(mfrow=c(2,2))
plot(model2_region_interactions)
mtext("Figure 20: Diagnostic Plots for Model 2 (interactions included)",
      side = 3, line = -2, outer = TRUE, cex = 1.1)
```

Figure 20: Diagnostic Plots for Model 2 (interactions included)




```
par(mfrow=c(1,1))
```

Initially, we consider interaction terms for each of the continuous random variables for both Models 1 and 2. In both instances, the interactions with the region variable are kept in the model only if the interaction term is statistically significant (at the 5% level) and useful for predicting per capita income. The results illustrate that in both models, the interaction terms are significant for the following variables: Percent high school graduates, Percent below poverty level, and Percent unemployment. We therefore only look at the models that include these interactions.

Table 8: Comparison Table for Models 1 and 2 (including interactions)

	df	AIC	BIC	R2 adj.
model1	9	-942.2740	-905.4931	0.8427
model2	10	-944.8883	-904.0206	0.8439
model1_region_interactions	21	-986.9437	-901.1215	0.8615
model2_region_interactions	25	-983.6060	-881.4366	0.8617

For Model 1 with the interaction term included, we see that the nested F-test is highly significant, suggesting that the interactions should remain in the model. The diagnostic plots in Figure 19 display that the regression model assumptions are similarly satisfied for the diagnostic plots in comparison with the models that do not include the interactions. While the VIFs for some of the coefficients are elevated, this is expected since the interaction terms introduce some collinearity into the model due to the nature of the relationship of interaction terms. This is acceptable, however, since the collinearities do not appear to cause noteworthy changes in the t-statistics and p-values based on the summary output.

We see similar results in Model 2 with the interaction term included, as the nested F-test is highly significant, suggesting that the interactions should remain in the model. While the VIFs for some of the coefficients are elevated, this is expected since the interaction terms introduce some collinearity into the model due to the nature of the relationship of interaction terms. This is acceptable, however, since the collinearities do not appear to cause noteworthy changes in the t-statistics and p-values based on the summary output. The diagnostic plots in Figure 20 also display that the regression model assumptions are similarly satisfied for the diagnostic plots in comparison with the models that do not include the interactions.

```
# comparison table for Models 1 and 2
comparison <- cbind(
  AIC = AIC(model1, model2, model1_region_interactions, model2_region_interactions),
  BIC = BIC(model1, model2, model1_region_interactions, model2_region_interactions),
  R2_adj = c(round(summary(model1)$adj.r.squared,4), round(summary(model2)$adj.r.squared,4), round(summa
comparison <- comparison[,-3]
names(comparison) <- c("df","AIC","BIC", "R2 adj.")
comparison %>%
  kbl(booktabs=T,
      caption = "Comparison Table for Models 1 and 2 (including interactions)") %>%
  kable_classic()
```

```
q3_final <- round(summary(model1)$coefficients,4)
q3_final %>%
  kbl(booktabs=T,
      caption = "Model 1 Coefficient and Standard Error Estimates") %>%
  kable_classic()
```

Selecting the model Table 8 displays the resulting adjusted R^2, AIC, and BIC values for Models 1 and 2, both with and without including the statistically significant regional interaction terms. While the AIC values improve for both models, the steeper penalty for adding coefficients from BIC illustrates that not much information is added when we include the interaction terms for region. Additionally, we see that there is only a marginal increase in the adjusted R^2 (less than 2% for both models) despite the degrees for freedom

Table 9: Model 1 Coefficient and Standard Error Estimates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.2225	0.0931	109.7765	0e+00
pop.18_34	-0.0139	0.0011	-12.5080	0e+00
pct.hs.grad	-0.0044	0.0011	-4.0714	1e-04
pct.bach.deg	0.0154	0.0009	16.6408	0e+00
pct.below.pov	-0.0243	0.0013	-19.2940	0e+00
pct.unemp	0.0106	0.0022	4.8705	0e+00
log_doctors	0.0607	0.0040	15.1000	0e+00
log_land.area	-0.0357	0.0048	-7.4683	0e+00

more than doubling for both Models 1 and 2. Given the criteria stated in the research question, we ignore these models to avoid the risk of overfitting the model to the data and decide between the models that do not include the interaction terms.

As previously stated, Models 1 and 2 are similar in many key aspects. The AIC, BIC, and adjusted R^2 are extremely similar; each of the coefficients present in the model is statisitcally significant and properly specified per the marginal model plots; the VIF values of the coefficients are low enough to suggest multicollinearity is not present in the model; and the diagnostic plots suggest that the regression model assumptions are approximately satisfied (with the slight deviation in the Normal QQ plot). Selecting the appropriate prediction model is therefore a decision made based on the more practical aspects of the model. Accounting for the criteria in the specified in the research question, Model 1 is selected since it is the more parsimonious model and provides virtually identical prediction power for the response variable. Table 9 provides the coefficient and standard error estimates for Model 1.