Modeling the relationship between personal income and a community's characteristics

Clare Cruz

Department of Statistics and Data Science

Carnegie Mellon University

clarecru@andrew.cmu.edu

Modeling the relationship between personal income and a community's characteristics

While the relationship between someone's income and their locality's qualities has critical social impacts, the connection is mainly unknown. Thus, this study seeks to understand the association between average income per person and the variables associated with a county's economic, health, and social well-being. The data consists of 14 county characteristics for 440 individual counties in the U.S based on information from 1990 to 1992. Exploratory data analysis is performed to determine the pairwise relationships between the given variables and address concerns about missing values. Additionally, regression modeling is utilized to model per capita income and confirm existing theories for the relationship between income, crime, and region. The exploratory data analysis found unsurprising relationships between the variables and that the missing values were not an issue for the study. Then, the regression modeling found that a county's geographical region does not affect the wealth gap since the region does not affect the positive association between per capita income and crime, even when crime per capita is used. Additionally, the final model for per capita income found that uneducated and rural areas with fewer job opportunities are at risk of low incomes. The analysis confirms several existing theories for income and suggests potential areas to close the wealth gap.

Introduction

As the wage gap increases in the United States, it is becoming increasingly important to determine what factors contribute to the economic disparity to address the issues. One way to tackle this problem is to investigate the relationship between a person's income and the quality of their surroundings, such as their county. Looking at average income in a county is essential because it provides more specific variables that could potentially affect the wealth gap. Therefore, this study aims to learn how average income per person is related to other variables associated with the county's economic, health, and social well-being. Since this goal requires a robust analysis, four sets of questions are addressed.

1.) **Economic Relationships** – Looking at the data one pair of variables at a time, which variables seem to be related to other variables in the data? Which are not?

2.) **Income, Crime, and Region** – There is a theory that, if all other variables are ignored, per-capita income should be related to the crime rate. This relationship may differ in different country regions (Northeast, Northcentral, South, and West). Does the data support this theory? Does the relationship change if the number of crimes, or (number of crimes)/(population), is included in the analysis?

3.) **Modeling Per-capita Income** – Find the best model predicting per-capita income from the other variables.

4.) **Missingness** – Should we be worried about either the missing states or the missing counties? Why or why not?

Data

The data set for this study includes selected county demographic information (CDI) for 440 of the most populous counties in the United States from Kutner et al. (2005). The original data is from the Geospatial and Statistical Data Center at the University of Virginia. Each county has an identification number, along with the county's name and state abbreviation. For each of the counties, there is information on fourteen

variables from the years 1990 to 1992. Of the fourteen variables, thirteen contain numeric data, and the remaining variable includes the county's region. A complete list of the variables and their definitions is in Table 1. The counties included in this dataset come from 48 states with 337 unique county names. The states that are excluded in the data are Arkansas, Iowa, and Wyoming. Some county names, like Jefferson, occur in multiple states. However, when the county and state are considered together, there are 440 unique combinations. Most counties are in the south, and the least are in the west for the region variable. There are no missing values in the dataset since any county with missing data was excluded before the analysis.

Table 1: Variable definitions for CDI data from Kutner et al. (2005). Original source: Geospatial and Statistical Data Center, University of Virginia.

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

Table 2: Descriptive statistics for the continuous variables.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
land.area	15.0	451.25	656.50	1041.41	946.75	20062.0	1549.92
pop	100043.0	139027.25	217280.50	393010.92	436064.50	8863164.0	601987.02
pop.18_34	16.4	26.20	28.10	28.57	30.02	49.7	4.19
pop.65_plus	3.0	9.88	11.75	12.17	13.62	33.8	3.99
doctors	39.0	182.75	401.00	988.00	1036.00	23677.0	1789.75
hosp.beds	92.0	390.75	755.00	1458.63	1575.75	27700.0	2289.13
crimes	563.0	6219.50	11820.50	27111.62	26279.50	688936.0	58237.51
pct.hs.grad	46.6	73.88	77.70	77.56	82.40	92.9	7.02
pct.bach.deg	8.1	15.28	19.70	21.08	25.33	52.3	7.65
pct.below.pov	1.4	5.30	7.90	8.72	10.90	36.3	4.66
pct.unemp	2.2	5.10	6.20	6.60	7.50	21.3	2.34
per.cap.income	8899.0	16118.25	17759.00	18561.48	20270.00	37541.0	4059.19
tot.income	1141.0	2311.00	3857.00	7869.27	8654.25	184230.0	12884.32

The descriptive statistics for the continuous variables outlined in Table 2 show no apparent abnormalities in the data. However, many continuous variables such as land area, population, doctors, hospital beds, crimes, total income, and per capita income have a substantially higher mean than the median. The plot of histograms in Figure 1 supports this idea by showing how land area, population, doctors, hospital beds, crimes, total income, and per capita income have right-skewed distributions. The remaining variables also have skewness to them but are not as extreme as the variables already mentioned.

Methods

Recall that this study seeks to answer four sets of questions to analyze the relationship between average income and county characteristics. The methods for each set of questions are outlined here:

1. Economic Relationships

The first question asks for the relationships between the variables in the dataset. For this question, pairwise correlations and scatterplots with the response variable were calculated to explore the connections for all the predictor variables. The categorical variables (state, county, and region) are excluded from this analysis because there are too many categories to make a reasonable conclusion.

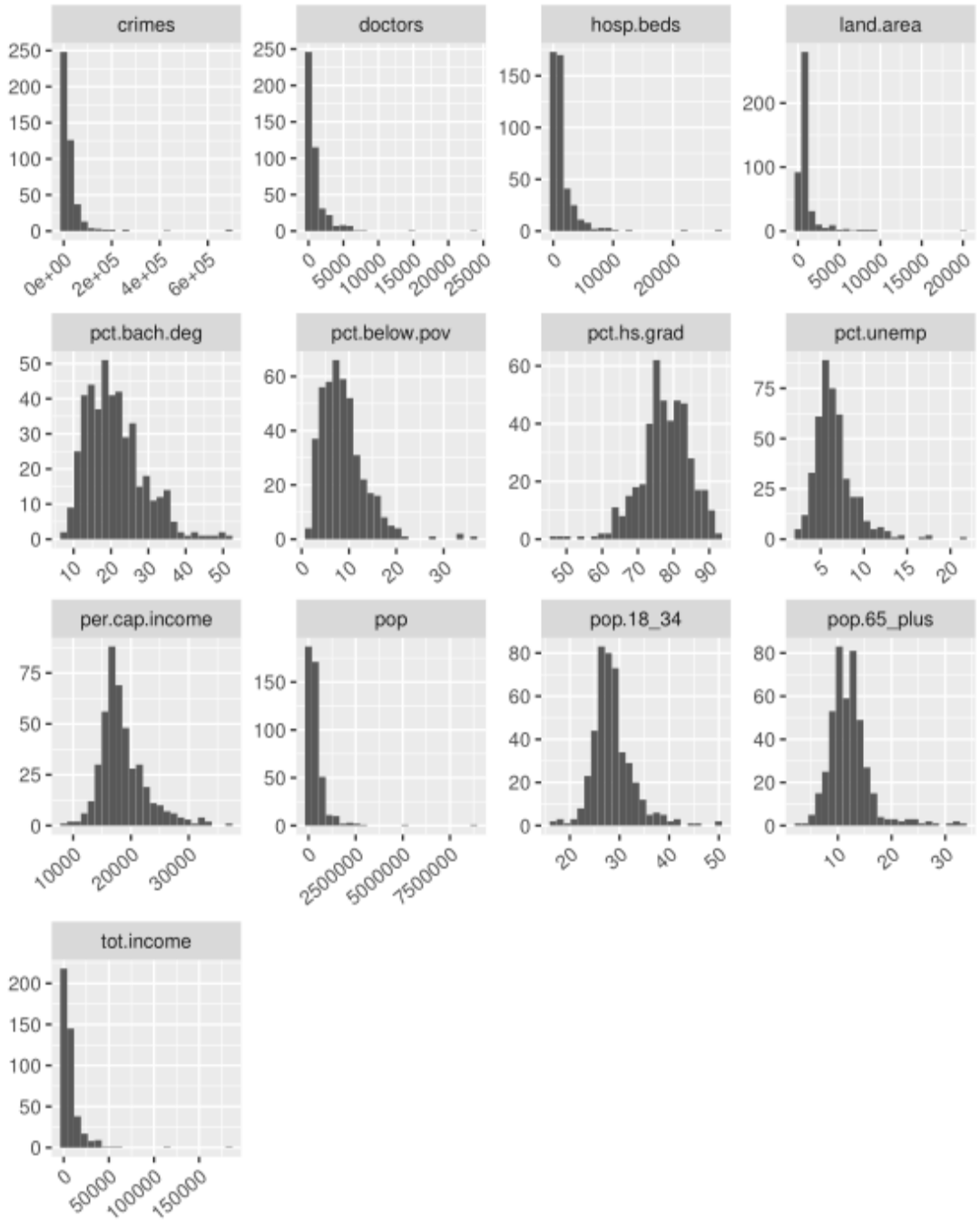


Figure 1: Histograms of the continuous variables.

2. Crime, Income, and Region

The next batch of questions pertains to the relationship between crime, income, and region. The first question asks if the relationship between crime and income differs based on the area. The best method to answer this question is to build a regression model since it directly translates with interaction terms. Recall that the data indicates that transformations are necessary to meet the linear model assumptions. Therefore, preliminary shifts using the log function were applied to the data to make the models valid.

Three regression models will be built to determine the relationship between income, crime, and region. The first model acts as a baseline with per capita income as the response and crime as the sole predictor. Next, the county's region is added to the previous model to determine if it is significant to predict per capita income. Then, a third model is created with the same parameters as the second model, plus their interaction terms. Finally, an ANOVA test compares the three models and determines which model best predicts per capita income. If the model with the interaction terms is the best, then there is evidence that the relationship between crime and income is dependent on the region.

The second goal is to see if the relationship between crime and income changes if crime per capita is used instead of crime. The analysis framework outlined in the previous paragraph will be performed again with the crime per capita variable for this task. Afterward, there will be two final models, the best model with the crime per capita variable and the optimal model with the given crime variable. Finally, a comparison is made between the two models by their coefficients, diagnostic plots, and summary regression statistics. If these two models are significantly different, then there is evidence that the relationship between crime and income depends on the crime metric used in the models.

3. Modeling Per-capita Income

This part of the study aims to build the best model that predicts per capita income using the variables available in the dataset. For this study, 'best model' means a compromise between meeting the statistical

assumptions, reflecting the social science behind the variables, relaying the data, and finding a model that can be easily explained to a non-statistician.

The first step in the modeling process is to find optimal transformations for the predictors if a shift is necessary. The Box-Cox method calculates the initial benchmark transformations, but the final changes will be simple, consistent, and interpretable functions that make each predictor as close to normal as possible. After the best transformations are identified, the variance inflation factors will be checked to see if any underlying relationships between the variables need to be excluded from the model building process. Next, three variable selection methods will be used to determine potential candidates for the final model: stepwise, best subset, and LASSO. In all three methods, the Bayes information criterion (BIC) will be used as the selection criterion since the focus of this study is to find the 'true' model for income per capita (Sheather 2009). For the lasso method, the choice of lambda will be determined by cross-validation.

The region variable is excluded from the model building process in all three methods since its categorical attribute can skew the variable selection methods. Nevertheless, the first-order interaction terms for the region variable will be considered after identifying a model to see if they improve the model fit. Higher-order interaction terms are not considered in the model since they complicate the model interpretation. Once the best model is found for each selection method, the three final models will be compared using diagnostic plots, summary regression statistics, and context to determine the overall best model.

4. Missingness

Finally, this section of the study will address any concern over the lost data present in the dataset. Specifically, the missing values will be analyzed by comparing the sample against the population to see if any important features are excluded from the data.

Results

1. **Economic Relationships** – Looking at the data one pair of variables at a time, which variables seem to be related to other variables in the data? Which are not?

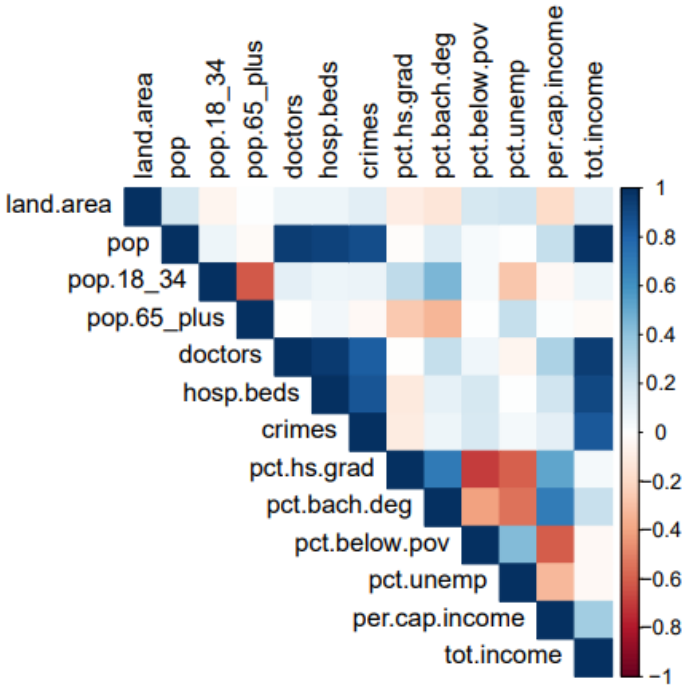


Figure 2: Correlation plot of the continuous variables.

The correlation plot in Figure 2 shows that there are several linear relationships among the variables. To start, total income and population have a high positive correlation, and these two variables are also highly correlated with crime, hospital beds, and doctors. Additionally, crime, hospital beds, and doctors have strong positive correlations with each other. These results are unsurprising since they all are related to an area's population. In contrast, only a couple of variables have negative associations between them. For example, the percentage of people below poverty and the rate of high school graduates have a negative relationship. This result makes sense because people who graduate from high school tend to have more job opportunities and avoid extreme poverty. While these relationships were foreseeable, the presence of high correlations between the predictor variables demonstrates that there may be some issues with multicollinearity during the model fitting process.

Interestingly, the response variable, per capita income, is not strongly correlated with any other variable. This result is supported by the scatterplots presented in Figure 3. None of the scatterplots show a linear relationship between the response and the predictors, which indicates that some additional work is needed for the data to meet the modeling assumptions. The skewed distribution also confirms a wealth gap since a small fraction of the counties has a significantly higher income than the majority.

2. Crime, Income, and Region

After transforming the skewed data using the log function, three regression models were built to evaluate the relationship between income, region, and crime. Each model contained income per capita as the response variable and crime as one of the predictors. The variation in the model came from the region variable, with one model containing the interaction terms for both region and crime. The ANOVA test concluded that the relationship between income and crime does not change based on the region because the ANOVA test returned a p-value that is well over 0.05 for the model with the interaction terms. The best model contained income per capita, crime rates, and region. More specifically, the model indicated that for every 1% increase in U.S. crimes, we expect a 0.07% increase in per-capita income, on average. Additionally, the four regions have different baseline per-capita incomes (Technical Appendix Part E, pg 21). Therefore, the best model indicates that the magnitude of salary varies with a region in the U.S., but the positive association is the same.

The results were very similar when the above modeling process was applied with the crime per capita variables. Once again, the ANOVA test concluded that the best model did not include the interaction terms since the p-value for the model with the interaction terms was well over 0.05. The best model also had the income per capita, crime rates, and region variables, but the coefficients differed slightly. More specifically, for every 1% increase in U.S. per-capita crime, there is an associated 0.04% increase in per-capita income. Again, the salary level varied with region, but the connection between crime and income remained the same. Thus, both models agree that crime and income have a relationship that does not depend on the region.

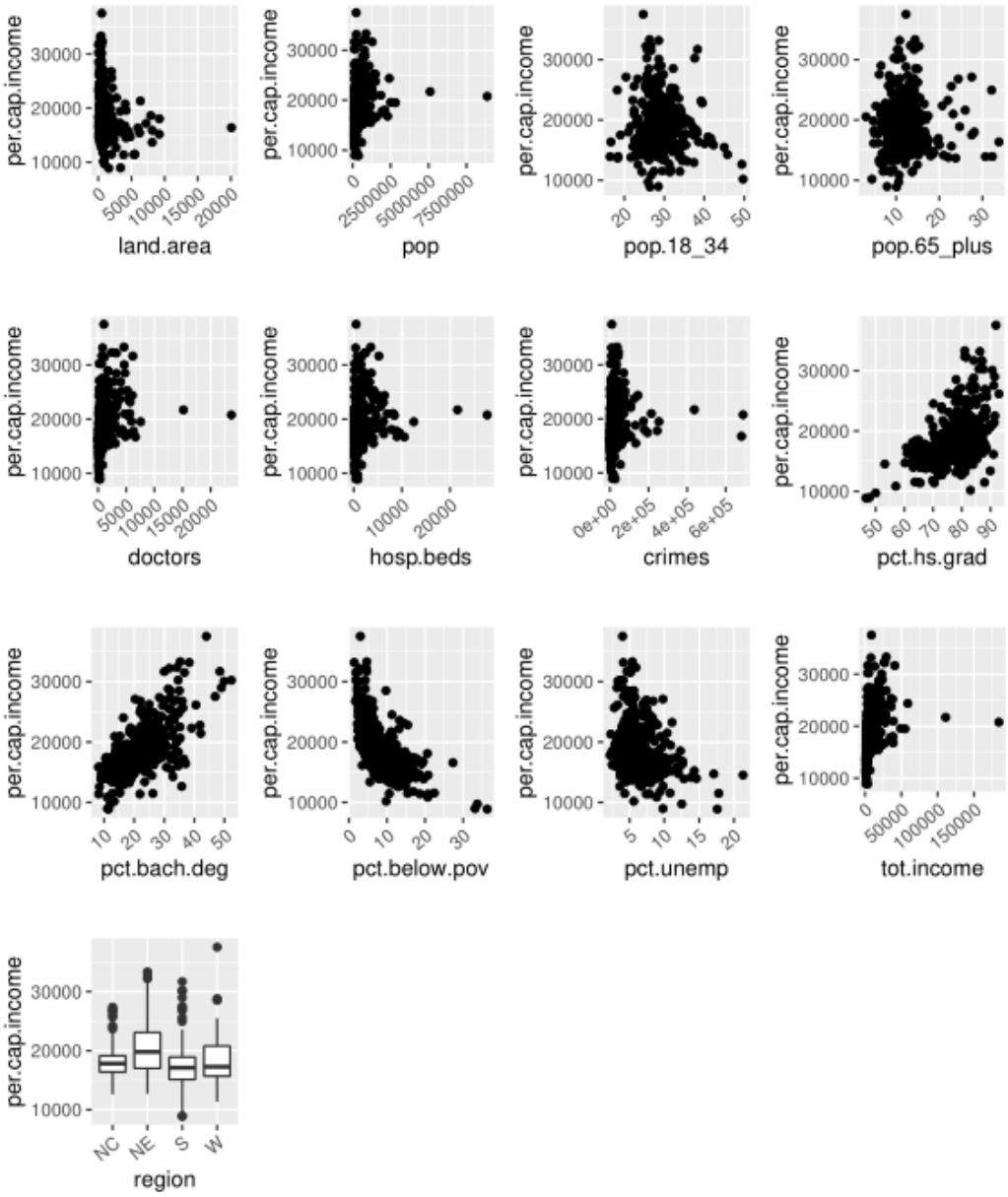


Figure 3: Scatterplots of the predictor variables against the response variable per capita income.

The final two models both included income per capita, region, and crime/per capita crime. The diagnostic plots and summary regression statistics are similar between the model with the original crime variable and the model with crime per capita (Technical Appendix Part E, pg 24). Therefore, the relationship between crime and income does not depend on the crime metric used in the models. This result suggests that

location is not contributing to the wealth since wealthier areas will experience more crime than poorer locations in every region in the country.

3. Modeling Per-capita Income

The focus of the model selection process was on finding the best model to explain income per capita from the county's qualities. At the beginning of the model-building process, the skewed data needed to be transformed to meet the linear model assumptions. The Box-Cox method indicated that many variables had ideal transformations that were unreasonable to interpret (Technical Appendix Part F, pg 27). For example, the optimal power for the land area variable was around -0.05, which is not easy to interpret, even for statisticians. Additionally, some variables could have been transformed in multiple ways to reach the normal distribution. Overall, the transformation that was the best balance between statistics and context was the log transformation applied to the land area, population, doctors, hospital beds, crimes, total income, and per capita income variables (Technical Appendix Part F, pg 28). Then, two variables had a high variance inflation factor: population and total income. This collinearity was an intuitive result because both are a function of the response variable. Consequently, both variables were removed from model consideration since multicollinearity interferes with the interpretation of the model.

In the variable selection process, the three different selection methods returned the same model with seven variables. However, their coefficients were relatively small, and some variables appeared to have the wrong sign (Technical Appendix Part F, pg 30). None of the diagnostic plots or regression statistics indicated any particular variable was causing the problem (Technical Appendix Part F, pg 31). Thus, the region variable with corresponding interaction terms was added to the model to improve the coefficients. After fitting the interaction terms to the model, the signs of the variables did not change, the diagnostic plots were the same, and the regression statistics were comparable to the model without the interaction terms (Technical Appendix Part F, 34). An ANOVA test was performed as a last resort, and the BIC and AIC values were also calculated. The ANOVA test concluded that the interaction terms should be included in the model since the p-value for the model with the interaction terms is significantly less than

0.05. This result aligns with the AIC value, which was lower for the model with the interaction terms. However, the BIC value favored the model without the interaction terms, and the model with the interaction terms still did not change the signs for the two variables of concern. The interaction terms were excluded from the final model since they complicate its interpretation and do not add much value. The final model contains seven variables which are listed with their summary statistics in Table 3.

Even though the selection methods returned the same model, the final model's diagnostic plots and summary regression statistics were evaluated to see how well the model fit the data. The diagnostic plots showed the residuals were random, normally distributed, consistently varied, and had no outliers or high leverage points. All of which indicates that the model is valid and a good fit for the data. Additionally, the summary regression statistics demonstrate that the final model has predictive power. The r-squared value was high ($R^2 = 0.85$), the standard error was low in context ($SE = 0.082$), all the coefficients were significant to the model, and the BIC metric was low ($BIC = -905.45$). Therefore, the best model is statistically valid and can effectively predict per capita income.

Table 3: Summary statistics for the final model that predicts per capita income.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.22	0.09	109.78	0
log.land.area	-0.04	0.00	-7.47	0
pop.18_34	-0.01	0.00	-12.51	0
log.doctors	0.06	0.00	15.10	0
pct.hs.grad	0.00	0.00	-4.07	0
pct.bach.deg	0.02	0.00	16.64	0
pct.below.pov	-0.02	0.00	-19.29	0
pct.unemp	0.01	0.00	4.87	0

The variable's coefficients in the final model have the following interpretations:

- **Land Area** – For every 1% increase in a county's land area, there is a 0.04% decrease in expected per-capita income. This association makes sense because larger areas are less populated, which leads to less income overall.
- **Percent Population 18-34** – For every 1% increase in the percentage of people aged 18 to 34, there is a 1% decrease in the expected per-capita income. This result aligns with the expectations that younger adults make less as they start their careers.
- **Doctors** – For every 1% increase in the number of doctors in a county, the expected per-capita income increases by about 0.06%. This result makes sense because doctors and medical personnel are high-paying jobs.
- **Percent High School Graduate** – For every 1% increase in the percentage of people with high school degrees, the expected per capita income increases by less than 1%. This result aligns with the expectations that people with high school degrees have more job opportunities and more income.
- **Percent Bachelor Degree** – For every 1% increase in the percentage of people with a bachelor's degree, the expected per capita income increases by 2%. This result aligns with the expectations that people with college degrees have more job opportunities and, therefore, more income.
- **Percent Below Poverty** – For every 1% increase in the percentage of people in poverty, there is a 2% drop in the expected per capita income. This result is intuitive since people who are in poverty have low incomes.
- **Percent Unemployment –** For every 1% increase in the unemployment rate, there is a 1% increase in the expected per capita income. This result is surprising since we would think that unemployed people have less income than employed people.

The final model suggests several interesting relationships with per capita income. For one, the presence of both education variables suggests that more resources should be devoted to schools since these variables had positive associations with income per capita. The number of doctors is also positively connected to

the response, implying that areas that house higher-paying jobs increase the overall income. However, the model also concluded that a higher unemployment rate increased the overall per capita income. While this seems incorrect, the generous unemployment benefits may be skewing this variable. Most unemployed people may receive government compensation that provides them with a consistent and reasonable income. A deeper investigation beyond the scope of this study is needed to confirm this theory. Additionally, the size of the area has a negative association with per capita income which suggests that rural areas are at an economic disadvantage. Finally, areas with a large percentage of younger adults tend to have less income. This result suggests that younger adults have less stable income are more vulnerable to the effects of economic disparity.

4. Missingness

Finally, the missing counties and states in the study are evaluated to see if any patterns could cause an issue to the analysis. The concern stems from the sampling method being unknown to the study, so a separate analysis is necessary to address the concern. Recall that 48 states and 440 individual counties are included in the data set. There are 50 states in the U.S., plus the District of Columbia. From this list, three states are missing from the dataset: Arkansas, Iowa, and Wyoming. These states are relatively unpopulated, with Wyoming being the least populist state in the country (Worldpopulationreview.com 2021). It would be best if three states were included in the dataset, but there was no glaringly unique pattern to these states that could cause a significant impact on analysis.

For the counties, the data includes 440 of the 3000 counties in the U.S. However, the counties that are included seem to be representative of the overall population. More specifically, California has the most counties, which aligns with the state's large population. There are also seven states with only one county, but they are all unpopulous. These observations suggest that the counties have been purposively selected because it ensures that the sample represents the population. Therefore, the missing values are not a concern since the data is representative of the population, and there are no apparent patterns in the missing data.

Table 4: Descriptive statistics for the state and county variables.

Statistic	Value
Total States	48
State with the Most Counties	CA with 34 counties.
Number of States with 1 County	7
Excluded States	AK, IA, WY

Discussion

The purpose of this study was to investigate the relationships between per capita income and county characteristics. The investigation was broken down into four sections with individual methods. In the first section, an exploratory data analysis found several unsurprising relationships between the variables. However, per capita income had many skewed relationships with the other variables. This result confirms the presence of a wealth gap and informed potential multicollinearity issues for the model. The next part of the analysis utilized regression models to determine the relationship between crime, region, and per capita income. The final model suggests that crime and income have a positive association when all other variables are constant. It also implies that location is not contributing to the wealth gap because the connection between crime and income does not change based on the region, even when the crime variable is transformed into crime per capita. Next, an optimal model was built to predict per capita income. After testing three selection methods, the best model included seven variables listed in Table 3. Finally, the missing values were analyzed to see if their exclusion could negatively influence the study. Descriptive analysis showed that the states and counties included in the dataset represented an accurate population sample. Therefore, no concerns are surrounding the missing values in the data.

Overall, the analysis has confirmed several existing theories relating to income and suggested several possible explanations for the country's wealth gap at a county level. This evidence can be helpful to those trying to implement solutions to the wealth gap in specific areas of the country.

The study has several strengths, including an interpretable model agreed upon by three modeling selection techniques. Also, the model's explicit validity and substantial predictive power support the implications behind the study. However, there are several areas for further analysis and improvement. While it was hypothesized that the positive association between unemployment and per capita income could be explained by government aid, further investigation should confirm the relationship. Additionally, there were not enough resources to fully consider the state variable. Individual state policy could explain the model's variability, so it would be valuable to see this variable utilized. Finally, it would be constructive to perform the analysis on a larger dataset. The analysis in this study implemented valid methods for variable selection. Nevertheless, the model may overfit the data since it was the only data used to create it. More data would allow for cross-validation, which would help create a generalizable model.

References & Citations

Kutner, Michael H., Christopher J. Nachtsheim, John Neter, and William Wasserman. 2005. *Applied Linear Statistical Models, Fifth Edition*. NY: McGraw-Hill/Irwin

Sheather, Simon. 2009. *A modern approach to regression with R.* Springer Science & Business Media.

Worldpopulationreview.com. 2021. U.S. States - Ranked by Population 2021. [online] Available at: <https://worldpopulationreview.com/states> [Accessed 18 October 2021]

Technical Appendix

Part A - Data Processing

```
head(data)
```

```
##   id      county state land.area     pop pop.18_34 pop.65_plus doctors
## 1  1 Los_Angeles    CA      4060 8863164      32.1         9.7   23677
## 2  2        Cook    IL       946 5105067      29.2        12.4   15153
## 3  3      Harris    TX      1729 2818199      31.3         7.1    7553
## 4  4   San_Diego    CA      4205 2498016      33.5        10.9    5905
## 5  5      Orange    CA       790 2410556      32.6         9.2    6062
## 6  6       Kings    NY        71 2300664      28.3        12.4    4861
##   hosp.beds crimes pct.hs.grad pct.bach.deg pct.below.pov pct.unemp
## 1     27700 688936        70.0         22.3          11.6       8.0
## 2     21550 436936        73.4         22.8          11.1       7.2
## 3     12449 253526        74.9         25.4          12.5       5.7
## 4      6179 173821        81.9         25.3           8.1       6.1
## 5      6369 144524        81.2         27.8           5.2       4.8
## 6      8942 680966        63.7         16.6          19.5       9.5
##   per.cap.income tot.income region
## 1          20786     184230      W
## 2          21729     110928     NC
## 3          19517      55003      S
## 4          19588      48931      W
## 5          24400      58818      W
## 6          16803      38658     NE
```

```
dim(data)
```

```
## [1] 440  17
```

```
str(data)
```

```
## 'data.frame':    440 obs. of  17 variables:
##  $ id            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ county        : chr  "Los_Angeles" "Cook" "Harris" "San_Diego" ...
##  $ state         : chr  "CA" "IL" "TX" "CA" ...
##  $ land.area     : int  4060 946 1729 4205 790 71 9204 614 1945 880 ...
##  $ pop           : int  8863164 5105067 2818199 2498016 2410556 2300664 2122101 2111687 1937094 1852
##  $ pop.18_34     : num  32.1 29.2 31.3 33.5 32.6 28.3 29.2 27.4 27.1 32.6 ...
##  $ pop.65_plus   : num  9.7 12.4 7.1 10.9 9.2 12.4 12.5 12.5 13.9 8.2 ...
##  $ doctors       : int  23677 15153 7553 5905 6062 4861 4320 3823 6274 4718 ...
##  $ hosp.beds     : int  27700 21550 12449 6179 6369 8942 6104 9490 8840 6934 ...
```

```
##  $ crimes        : int  688936 436936 253526 173821 144524 680966 177593 193978 244725 214258 ...
##  $ pct.hs.grad   : num  70 73.4 74.9 81.9 81.2 63.7 81.5 70 65 77.1 ...
##  $ pct.bach.deg  : num  22.3 22.8 25.4 25.3 27.8 16.6 22.1 13.7 18.8 26.3 ...
##  $ pct.below.pov : num  11.6 11.1 12.5 8.1 5.2 19.5 8.8 16.9 14.2 10.4 ...
##  $ pct.unemp     : num  8 7.2 5.7 6.1 4.8 9.5 4.9 10 8.7 6.1 ...
##  $ per.cap.income: int  20786 21729 19517 19588 24400 16803 18042 17461 17823 21001 ...
##  $ tot.income    : int  184230 110928 55003 48931 58818 38658 38287 36872 34525 38911 ...
##  $ region        : chr  "W" "NC" "S" "W" ...
```

```
# Change id to character for ease
data$id <- as.character(data$id)
```

ID is a numeric variable by default. This doesn't make sense for our analysis so it is changed to a character variable.

Part B - Data Description and Tables

```
cont_var <- unlist(lapply(data, is.numeric))
cont_table <- as.data.frame(apply(data[,cont_var], 2, summary))

# Splitting the continuous and integer variables since they all can't fit on one table
percent_var <- c('pop.18_34','pop.65_plus','pct.hs.grad','pct.bach.deg',
                 'pct.below.pov','pct.unemp')
integer_var <- c('pop','doctors','hosp.beds','crimes','tot.income')

percent_table <- cont_table[,which((names(cont_table) %in% percent_var)==TRUE)]
percent_table <- percent_table %>% mutate_if(is.numeric, ~round(., 1))

integer_table <- cont_table[,which((names(cont_table) %in% percent_var)==FALSE)]
integer_table <- integer_table %>% mutate_if(is.numeric, ~round(., 1))
```

```
# Region
# Only extract the values with the sum so that we can say 'total'
cross <- table(data$region)
region_table <- as.vector(addmargins(cross))
region_df <- data.frame(c('NC','NE','S','W','Total'),region_table)
colnames(region_df) <- c('Region','Freq')
```

```
# State and county
unique_state <- length(unique(data$state))
popular_state <- data %>%
 group_by(state) %>%
 summarise(n = n()) %>%
 top_n(1,n)

popular_state_stat <- paste(popular_state$state,'with',popular_state$n,'counties.',sep = " ")

unpopular_state <- data %>%
 group_by(state) %>%
 summarise(n = n()) %>%
 filter(n == 1)
```

Table 1:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
land.area	15.0	451.25	656.50	1041.41	946.75	20062.0	1549.92
pop	100043.0	139027.25	217280.50	393010.92	436064.50	8863164.0	601987.02
pop.18_34	16.4	26.20	28.10	28.57	30.02	49.7	4.19
pop.65_plus	3.0	9.88	11.75	12.17	13.62	33.8	3.99
doctors	39.0	182.75	401.00	988.00	1036.00	23677.0	1789.75
hosp.beds	92.0	390.75	755.00	1458.63	1575.75	27700.0	2289.13
crimes	563.0	6219.50	11820.50	27111.62	26279.50	688936.0	58237.51
pct.hs.grad	46.6	73.88	77.70	77.56	82.40	92.9	7.02
pct.bach.deg	8.1	15.28	19.70	21.08	25.33	52.3	7.65
pct.below.pov	1.4	5.30	7.90	8.72	10.90	36.3	4.66
pct.unemp	2.2	5.10	6.20	6.60	7.50	21.3	2.34
per.cap.income	8899.0	16118.25	17759.00	18561.48	20270.00	37541.0	4059.19
tot.income	1141.0	2311.00	3857.00	7869.27	8654.25	184230.0	12884.32

```
excluded_states <- toString(setdiff(state.abb,data$state))

state_df <- data.frame(c('Total States', 'State with the Most Counties','Number of States with 1 County
colnames(state_df) <- c('Statistic', 'Value')
```

```
popular_county <- data %>%
 group_by(county) %>%
 summarise(n = n()) %>%
 group_by(n) %>%
 summarize(count = n())

most_popular_county <- data %>%
 group_by(county) %>%
 summarise(n = n()) %>%
 group_by(n) %>%
 filter(n > 3)

unique_county <- length(unique(data$county))
sample_county <- c('...', '...' , '...' , 'Cumberland, Jackson, Lake', 'Washington', 'Montgomery', 'Jef

county_df <- data.frame(popular_county,sample_county)
colnames(county_df) <- c('Number of States','Total Counties','County')

# Add a total bar so unique counties are shown
county_df <- county_df %>% adorn_totals("row")
```

```
cdinumeric <- data[,-c(1,2,3,17)] ## get rid of id, county, state and (for now) region
apply(cdinumeric,2,function(x) c(summary(x),SD=sd(x))) %>% as.data.frame %>% t() %>%
  round(digits=2) %>% kbl(booktabs=T,caption=" ") %>% kable_classic()
```

```
# kableExtra::kbl(percent_table, caption = "Descriptive statistics for the continuous variables that ar
#   kableExtra::kable_styling(latex_options = "HOLD_position") %>%
#   kableExtra::kable_classic() %>%
#   kableExtra::row_spec(6, hline_after = TRUE)
#
# kableExtra::kbl(integer_table, caption = "Descriptive statistics for the continuous variables that ar
#   kableExtra::kable_styling(latex_options = "HOLD_position") %>%
#   kableExtra::kable_classic() %>%
#   kableExtra::row_spec(6, hline_after = TRUE)
```

```
kableExtra::kbl(region_df, caption = "Descriptive statistics for the region variable.", booktabs = T, li
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic() %>%
  kableExtra::row_spec(5, bold=T)
```

Table 2: Descriptive statistics for the region variable.

Region	Freq
NC	108
NE	103
S	152
W	77
Total	**440**

```
kableExtra::kbl(state_df, caption = "Descriptive statistics for the state variable.", booktabs = T, lin
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

Table 3: Descriptive statistics for the state variable.

Statistic	Value
Total States	48
State with the Most Counties	CA with 34 counties.
Number of States with 1 County	7
Excluded States	AK, IA, WY

```
kableExtra::kbl(county_df, caption = "Descriptive statistics for the county variable.", booktabs = T, li
  kableExtra::kable_styling(latex_options = "HOLD_position") %>%
  kableExtra::kable_classic()
```

Table 4: Descriptive statistics for the county variable.

Number of States	Total Counties	County
1	334	...
2	23	...
3	10	...
4	3	Cumberland, Jackson, Lake
5	1	Washington
6	1	Montgomery
7	1	Jefferson
Total	373	-

There are 48 states; Arkansas, Idaho, and Wyoming are the only excluded states. 48 categories is too much for it to be useful to the dataset. Therefore, this variable will be excluded from the analysis.

There are 373 unique counties, with some county names repeated several times in different states. If you combine state and county together, you will get 440 unique values which corresponds to the number of rows. Again, 373 categories is too much for it to be useful in the dataset. Therefore, this variable will be excluded from the analysis.

Region only has four unique values and it's fairly evenly distributed among the four regions. Therefore, it will be considered in the data analysis.

Part C - Missing Values

```
unlist(lapply(data, function(x){sum(is.na(x))}))
```

```
##             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
```

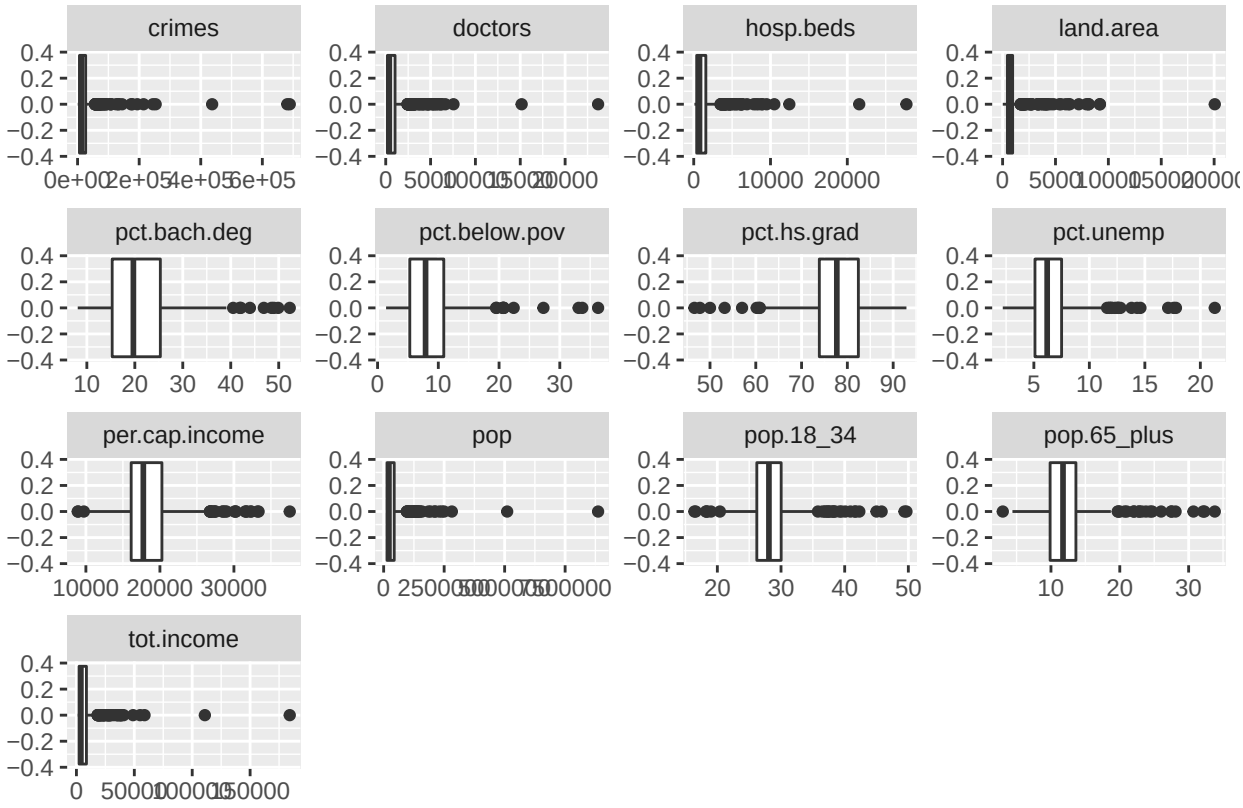
There are no missing values in the data since counties with missing values were removed prior to analysis.

Part D - Descriptive EDA

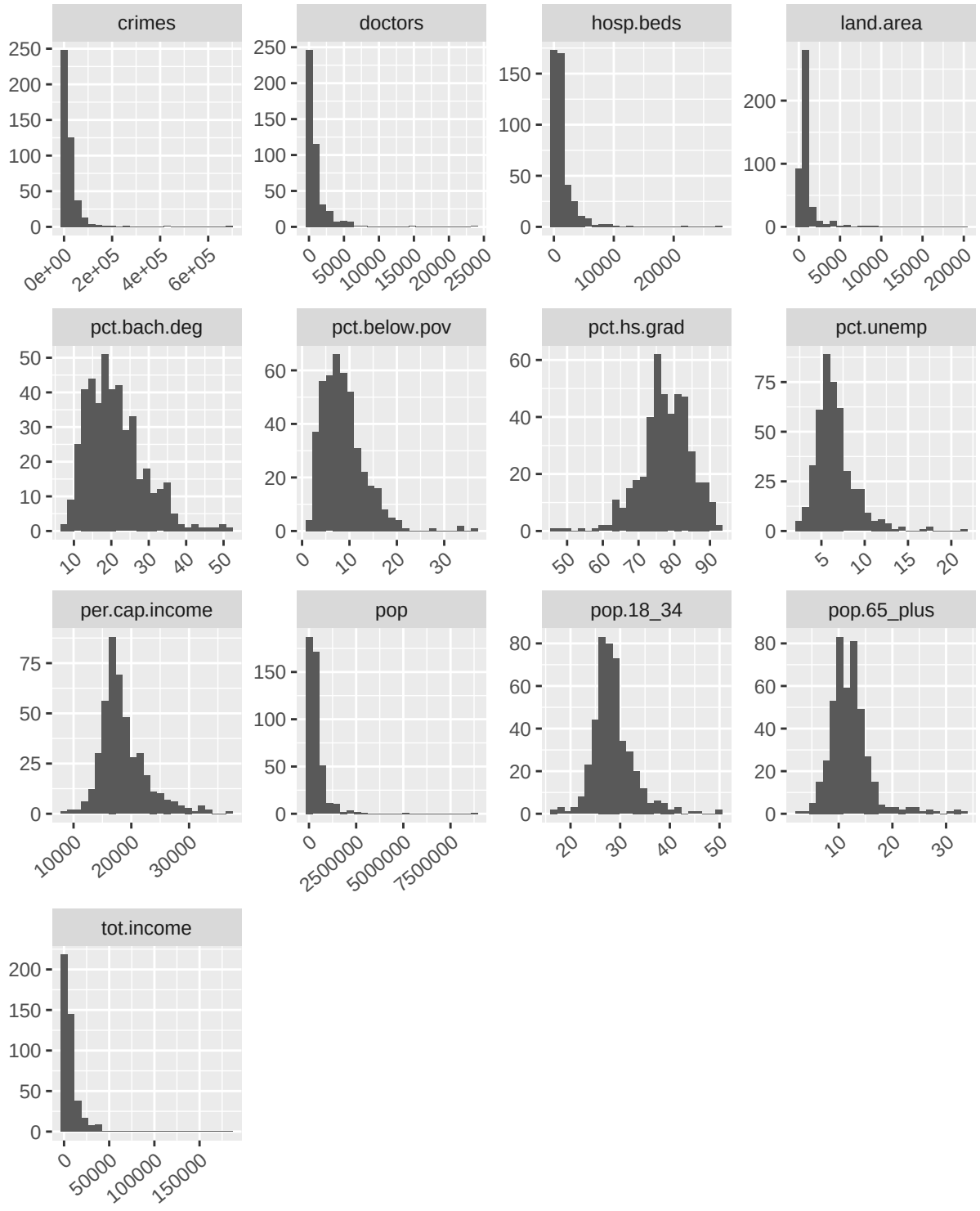
Since we have determined which variables will be useful to the model - we will condense the data to the variables were are interested in.

```
viz_data <- data[,-c(1,2,3)] # get rid of id, state, and count for the visualizations
```

```
ggplot(gather(viz_data[,-c(14)]), aes(value)) +
 geom_boxplot()+
 facet_wrap(~key, scales = 'free')
```




```
ggplot(gather(viz_data[,-c(14)]), aes(value)) +
  geom_histogram(bins = 25) +
 facet_wrap(~key, scales = "free") +
  theme(axis.text.x = element_text(angle = 40,  hjust=1))+
  xlab("")+
  ylab("")
```

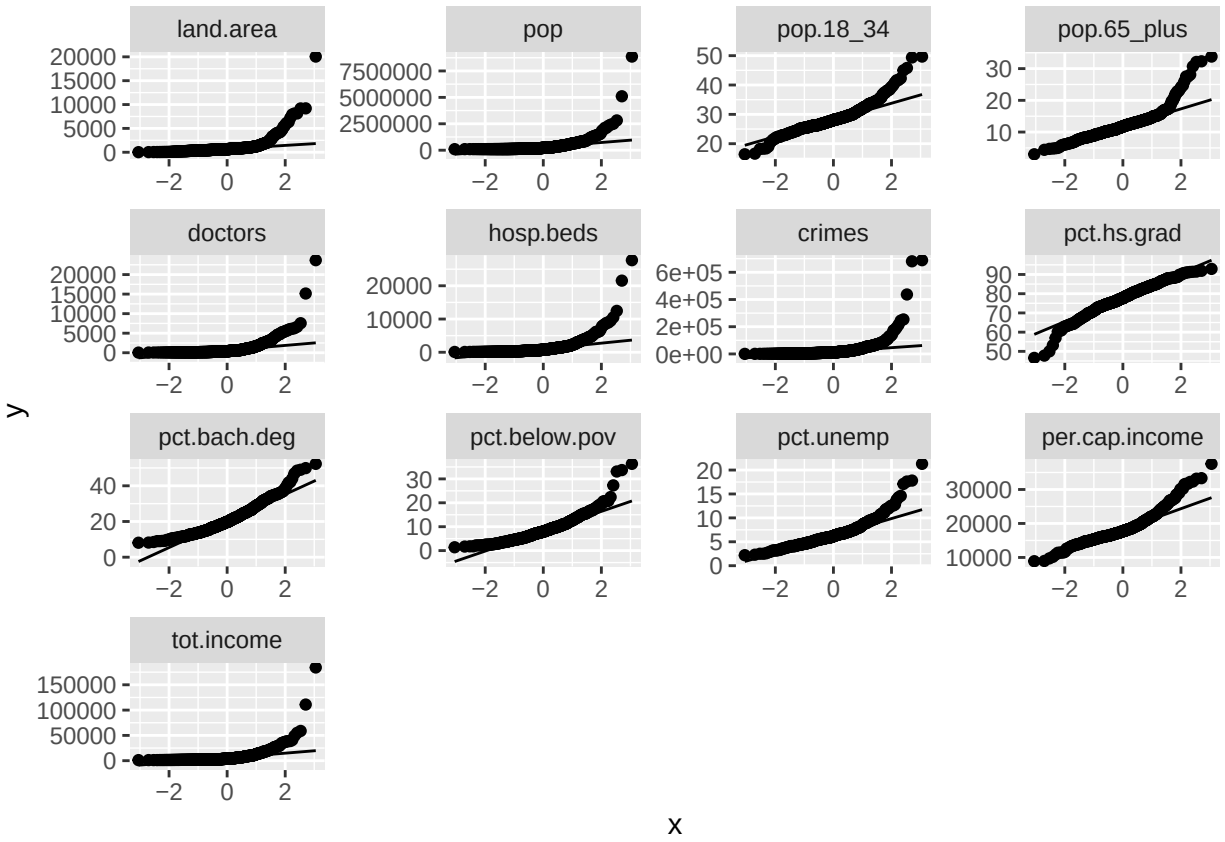


It looks from the histograms like the variables that will really need attention (because they are severely right-skewed) are `land.area`, `pop`, `doctors`, `hosp.beds`, `crimes`, and `tot.income`, and maybe `per.cap.income`.

We can look at QQ plots to determine if the other variables would be worth transforming.

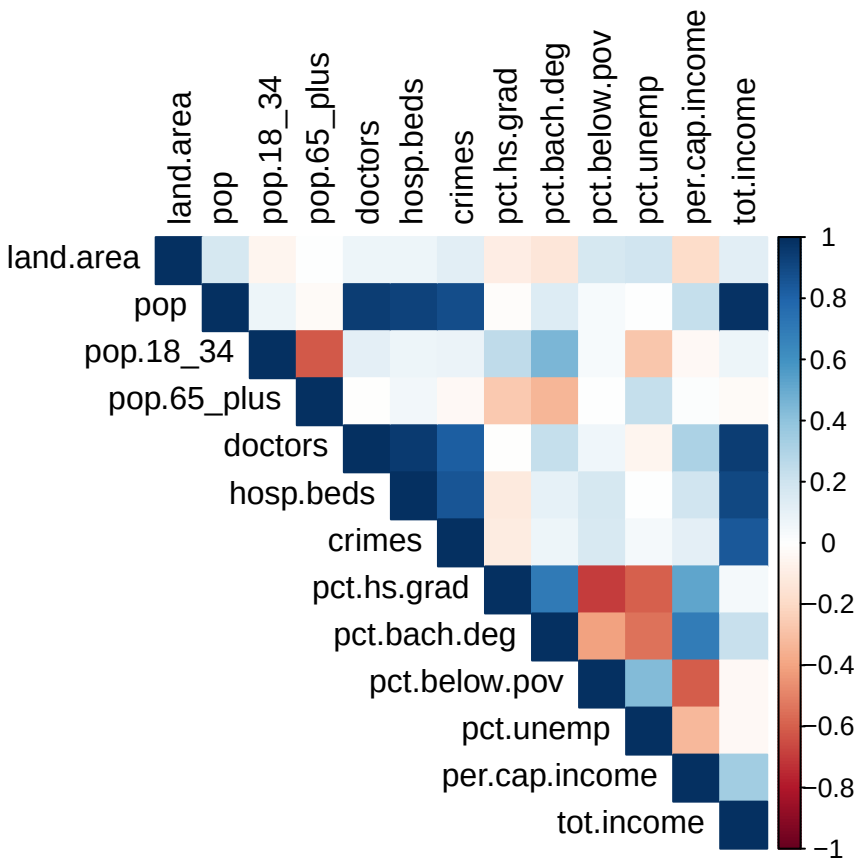
```
X2 <- stack(data[,cont_var])

ggplot(X2, aes(sample = values)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap( ~ ind, scales = 'free')
```



Then we can look at a correlation plot to visualize the relationships between the variables

```
m <- cor(data[,cont_var])
corrplot(m, method="color", type="upper",tl.col="black")
```



We can make the following conclusions from the correlation matrix:

- `tot.income` and `pop` are highly correlated (no surprise there)
- both are reasonably highly correlated with `crimes`, `hosp.beds` and `doctors`
- the three variables `crimes`, `hosp.beds` and `doctors` seem strongly correlated with one another
- `per.cap.income` isn't really highly correlated with anything, but the best possibilities seem to be `pct.hs.grad`, `pct.bach.deg` (postively correlated with `per.cap.income`) and `pct.below.pov`, `pct.unemp` (negatively correlated with `per.cap.income`); all four of these variables are moderately highly correlated with one another

These observations suggest that we may run into multi-collinearity problems when we start fitting models, but there is also some hope that we can make a good model for `per.cap.income`.

Now we turn to scatter plots, but we are just going to concentrate on relationships with `per.cap.income`:

```
## I'm not sure how I'd do this directly with ggplot...

scatter.builder <- function(df,yvar="per.cap.income") {
    result <- NULL
    y.index <- grep(yvar,names(df))
    for (xvar in names(df)[-y.index]) {
      d <- data.frame(xx=df[,xvar],yy=df[,y.index])
      if(mode(df[,xvar])=="numeric") {
        p <- ggplot(d,aes(x=xx,y=yy)) + geom_point() +
               ggtitle("") + xlab(xvar) + ylab(yvar) +
```

```
                theme(axis.text.x = element_text(angle = 40,  hjust=1))
      } else {
          p <- ggplot(d,aes(x=xx,y=yy)) + geom_boxplot(notch=F) +
                ggtitle("") + xlab(xvar) + ylab(yvar)  +
                theme(axis.text.x = element_text(angle = 40,  hjust=1))
      }
    result <- c(result,list(p))
    }
    return(result)
}

grid.arrange(grobs=scatter.builder(viz_data))
```

The best possibilities for predicting `per.cap.income` are the same variables we identified from the correlation matrix: `pct.hs.grad`, `pct.bach.deg`,`pct.below.pov`, and `pct.unemp`. The last plot shows how `per.cap.income` varies across the four regions of the country. There is a lot of overlap in the boxplots, but the Northeast and the West seem to be doing a little better than the North Central and South regions.

Part E - Crime, Income, and Region

Build a regression model that predicts per-capita income from crime rate and region of the country. Should there be any interactions in the model? What does your model say about the relationship between per-capita income and crime rate? Do your answers change, depending on whether you use number of crimes, or "per-capita crime" = (number of crimes)/(population) as a crime rate measure? If so, which one best answers the question? Why? Show the fitted model results and explain your answer to these questions in terms of those results

```
raw_crime_model <- lm(log(per.cap.income) ~ log(crimes), data = data)
nointer_model <- lm(log(per.cap.income) ~ log(crimes) + region, data = data)
inter_model <- lm(log(per.cap.income) ~ log(crimes)*region, data = data)

anova(raw_crime_model, nointer_model, inter_model)
```

```
## Analysis of Variance Table
##
## Model 1: log(per.cap.income) ~ log(crimes)
## Model 2: log(per.cap.income) ~ log(crimes) + region
## Model 3: log(per.cap.income) ~ 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
```

> There should not be an interaction term in the model because the ANOVA test returned a p-value that is well over 0.05 for the interactions terms which means that they should not be added to the model. Additionally, the coefficients for the interaction terms were insignificant when they were fitted to the model which also indicates that they should not be included in the model. Also, the model that include crime and region is the best at predicting income per capita.

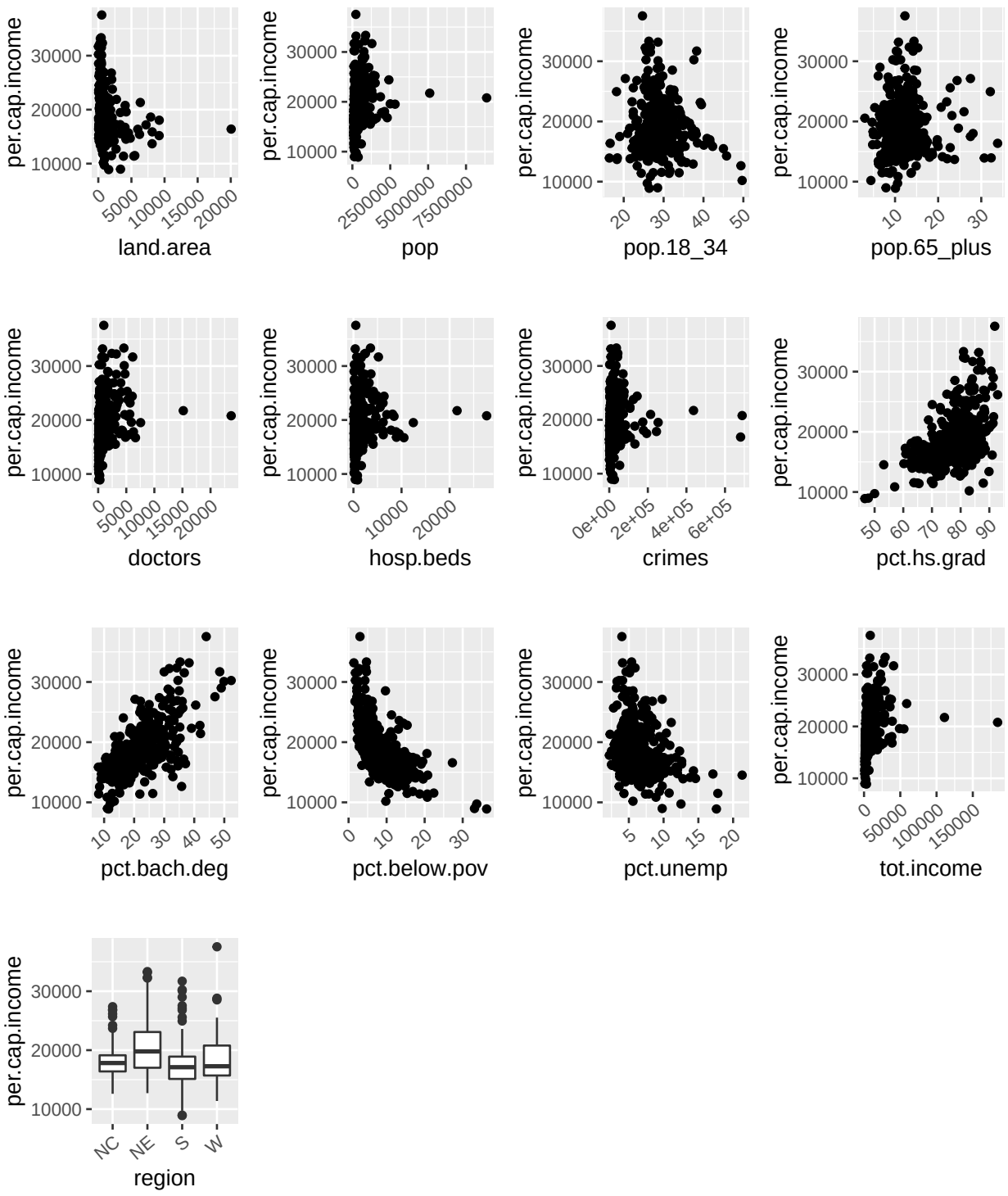
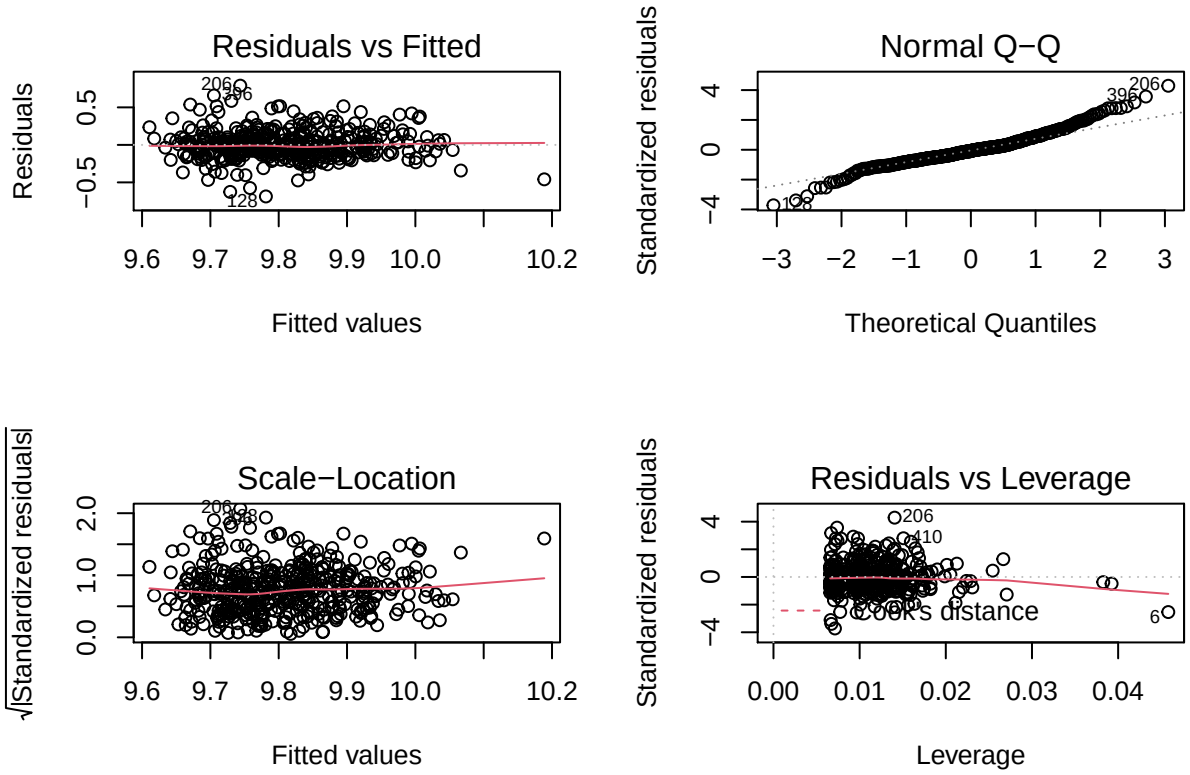


Figure 1: Scatter Plots with y = per.cap.income

```r
summary(nointer_model)
```

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

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



The model says that there is a positive linear relationship between income per capita income and crime rates when the region is constant. More specifically, the crime rate variable's coefficient is significant with a p-value well below 0.05. This result indicates that crime rate has a linear relationship with income per capita. Then, the positive coefficient value for crime rate shows that the linear relationship between income per capita and crime is positive. More specifically for every 1% increase in US crimes, we expect a 0.07% increase in per-capita income, on average.

Different regions of the country have different baseline per-capita incomes however: In the NC region, the baseline salary is , and in the W it is All of these region baselines are, according to the model, significantly different from the NC baseline.

Therefore, according to the model, the *level* of salary varies with region in the US, but the *way it is related to crime* does not. .

Create the new crime per capita variable

```
data_prob1 <- data
data_prob1$per.capita.crime <- data_prob1$crime/data_prob1$pop
```

```
crime_cap_model <- lm(log(per.cap.income) ~ log(per.capita.crime), data = data_prob1)
nointer_model_pcc <- lm(log(per.cap.income) ~ log(per.capita.crime) + region, data = data_prob1)
inter_model_pcc <- lm(log(per.cap.income) ~ log(per.capita.crime)*region, data = data_prob1)

anova(crime_cap_model, nointer_model_pcc, inter_model_pcc)
```

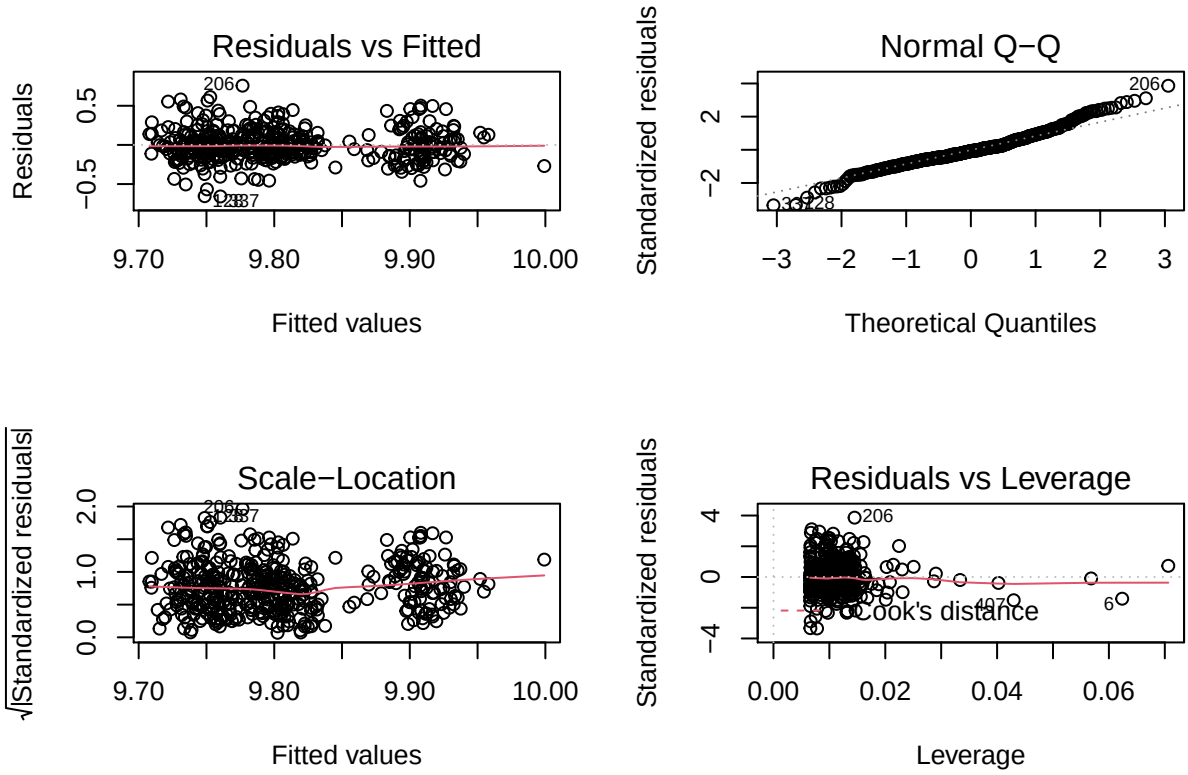
```
## Analysis of Variance Table
##
## Model 1: log(per.cap.income) ~ log(per.capita.crime)
## Model 2: log(per.cap.income) ~ log(per.capita.crime) + region
## Model 3: log(per.cap.income) ~ log(per.capita.crime) * region
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 18.697
## 2    435 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
```

Once again, there should not be an interaction term in the model because the ANOVA test returned a p-value that is well over 0.05 for the interactions terms which means that they should not be added to the model. Also, the model that include crime per capita and region is the best at predicting income per capita.

```
summary(nointer_model_pcc)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(per.capita.crime) + region,
##     data = data_prob1)
##
## 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(per.capita.crime)  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
```

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



If we want to statistically compare the best models in both scenarios (raw crime and crime per capita) we need to use AIC or BIC because the two best models are not nested models.

```
AIC(nointer_model,nointer_model_pcc)
```

```
##                    df       AIC
## nointer_model       6 -227.4746
## nointer_model_pcc   6 -172.1347
```

```
BIC(nointer_model,nointer_model_pcc)
```

```
##                    df       BIC
## nointer_model       6 -202.9539
## nointer_model_pcc   6 -147.6140
```

The relationship between crime and income per capita is similar when the transformed crime rate is fit to income. Again, the crime rate variable's coefficient is significant with a p-value well below 0.05. This result indicates that crime rate has a linear relationship with income per capita. Then, the positive coefficient value for crime rate shows that the linear relationship between income per capita and crime is positive. More specifically, for every 1% increase in US per-capita crime, there is an associated 0.04% increase in per-capita income. Again, the level of salary varies with region, but not the way it varies with crime, according to the model.

From a statistical perspective, the original crime variable is the best variable to answer the question because it produces a valid model to predict per capita income. More specifically, the

diagnostic plots and summary regression statistics are better in the model with the original crime variable than the model with crime per capita. In the diagnostic plots, the model with crime per capita has distinct groups in the residuals and several potential high leverage values. This result indicates that the model does not meet the assumptions for linear modeling. In contrast, the residuals in the model with the original crime variable have no pattern that are centered at 0, approximately normally distributed, constant variance, and only a handful of potential high leverage values. This result indicates that the model does meet the model assumptions for linear modeling and therefore is a better model than the model with crime per capita. The summary regression statistics also agree with this conclusion since the r-squared value and variable significance's are better for the model with the original crime variable. Lastlty, BIC and AIC agree that the model with region and crime is the best model since they have the lowest values.

However, this conclusion does not make sense contextually since the crime per capita variable has the same standardization as income per capita. In other words, crime per capita eliminates the possible confounding effect that population can have on crime and describes crime relative to the area's population (more dense places have more crime since there are more people). Therefore, the crime per capita variable is the best variable to answer the question.

Part F - Transformations and VIF

Use methods we have discussed in class and/or methods from Sheather Chapters 5, 6 & 7 (including, as needed: transformations, interactions, variable selection, residual analysis, fit indices, etc.) to find the multiple regression model predicting per-capita income from the other variables, that makes the "best" tradeoff between the following criteria:

• Reflects the social science and the meaning of the variables • Satisfies modeling assumptions • 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.

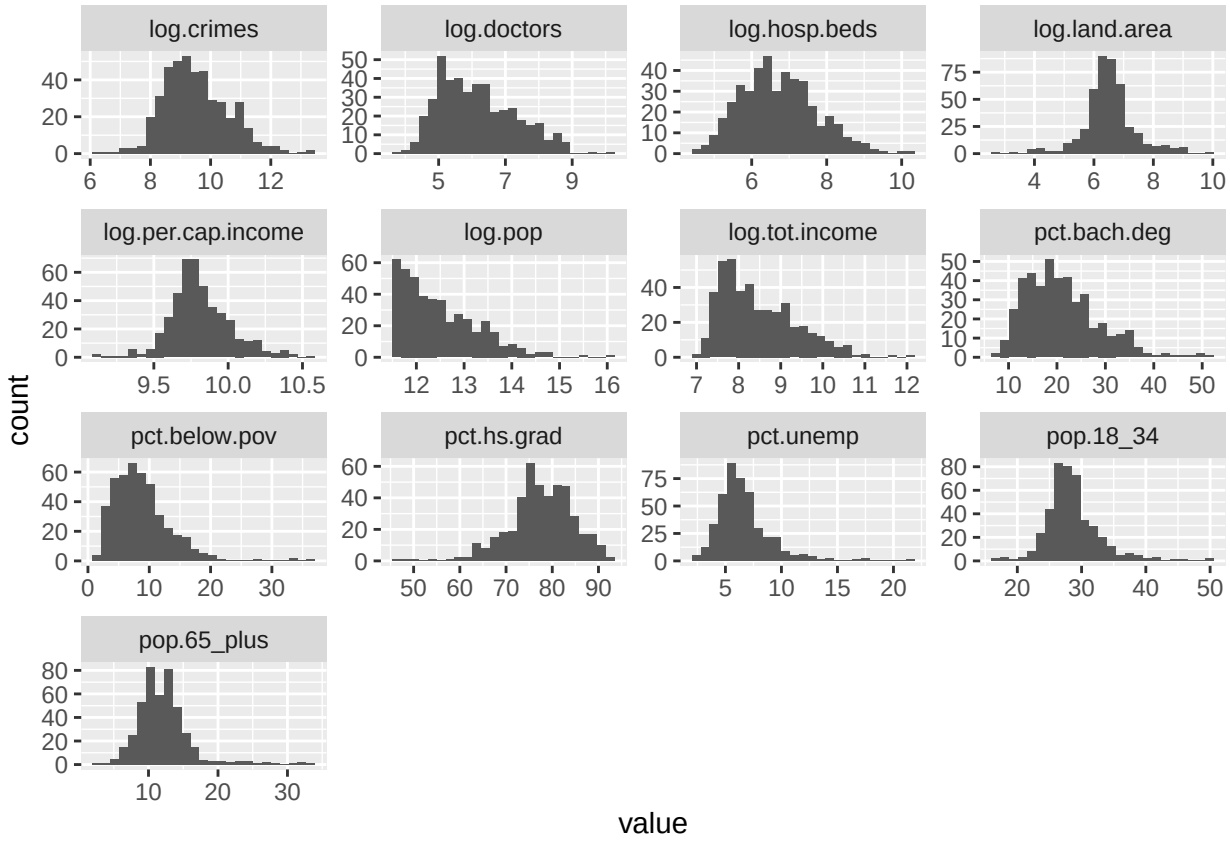
```
# Power transformations suggested by BoxCox
unlist(lapply(data[,cont_var], function(x){powerTransform(x)$roundlam}))
```

```
##      land.area.x            pop.x      pop.18_34.x    pop.65_plus.x
##        0.0000000       -0.5000000        0.0000000        0.0000000
##        doctors.x      hosp.beds.x         crimes.x    pct.hs.grad.x
##       -0.2174773       -0.1541052       -0.1307109        3.0719249
##   pct.bach.deg.x  pct.below.pov.x      pct.unemp.x per.cap.income.x
##        0.0000000        0.1817562        0.0000000       -0.5000000
##     tot.income.x
##       -0.5000000
```

```
# Applying log transformations to the data
data_trans <- viz_data
log_trans <- c('crimes','hosp.beds','doctors','land.area','pop','tot.income','per.cap.income')

for (tmp in log_trans) {
  loc <- grep(paste("^",tmp,"$",sep=""),names(data_trans))
  data_trans[,loc] <- log(data_trans[,loc])
  names(data_trans)[loc] <- paste("log.",names(data_trans)[loc],sep="")
}
#data_trans$pct.hs.grad <- data_trans$pct.hs.grad^3 # only left skewed variable
```

```
ggplot(gather(data_trans[,-c(14)]), aes(value)) +
 geom_histogram(bins = 25)+
 facet_wrap(~key, scales = 'free')
```




```
vif(lm(log.per.cap.income ~., data = data_trans[,c(-1,-2,-3)]))
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## pop.65_plus     1.529781  1        1.236843
## log.doctors    18.068402  1        4.250694
## log.hosp.beds  11.278397  1        3.358332
## log.crimes      9.557573  1        3.091533
## pct.hs.grad     4.482986  1        2.117306
## pct.bach.deg    4.033704  1        2.008408
## pct.below.pov   4.136517  1        2.033843
## pct.unemp       2.143984  1        1.464235
## log.tot.income 13.788017  1        3.713222
## region          2.889704  3        1.193463
```

We ran the initial VIFs before variable selection to ensure that there was no underlying relationship in the data. From the VIFs, we found that log.pop and log.tot.income need to be excluded from consideration, since per.cap.income is a deterministic function of them (so if they are included, no other predictors can possibly matter, and so I won't learn anything about what is associated with per.cap.income).

```
# Disregard state, county, tot.income (because of VIF), and region since it is categorical
#data_region <- data_trans[,c(-14)] # need region for the interactions model

# Disregard region since it messes with the variable selection methods and pop and income
log_data_cont <- data_trans[,c(-2,-13,-14)]
```

Part G - All Subsets

```
# Including region produces more variables in the final model
all.subsets <- regsubsets(log.per.cap.income ~., data = log_data_cont, nvmax = 10)
#subsets(all.subsets)
reg_summary <- summary(all.subsets)

reg_summary$bic
```

```
##  [1] -257.5260 -502.4302 -572.5538 -682.8532 -732.1894 -761.5908 -772.0715
##  [8] -770.5990 -766.2235 -760.4131
```

```
min(reg_summary$bic)
```

```
## [1] -772.0715
```

```
print(best.model <- which.min(reg_summary$bic))
```

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

```
coef(all.subsets,best.model)
```

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

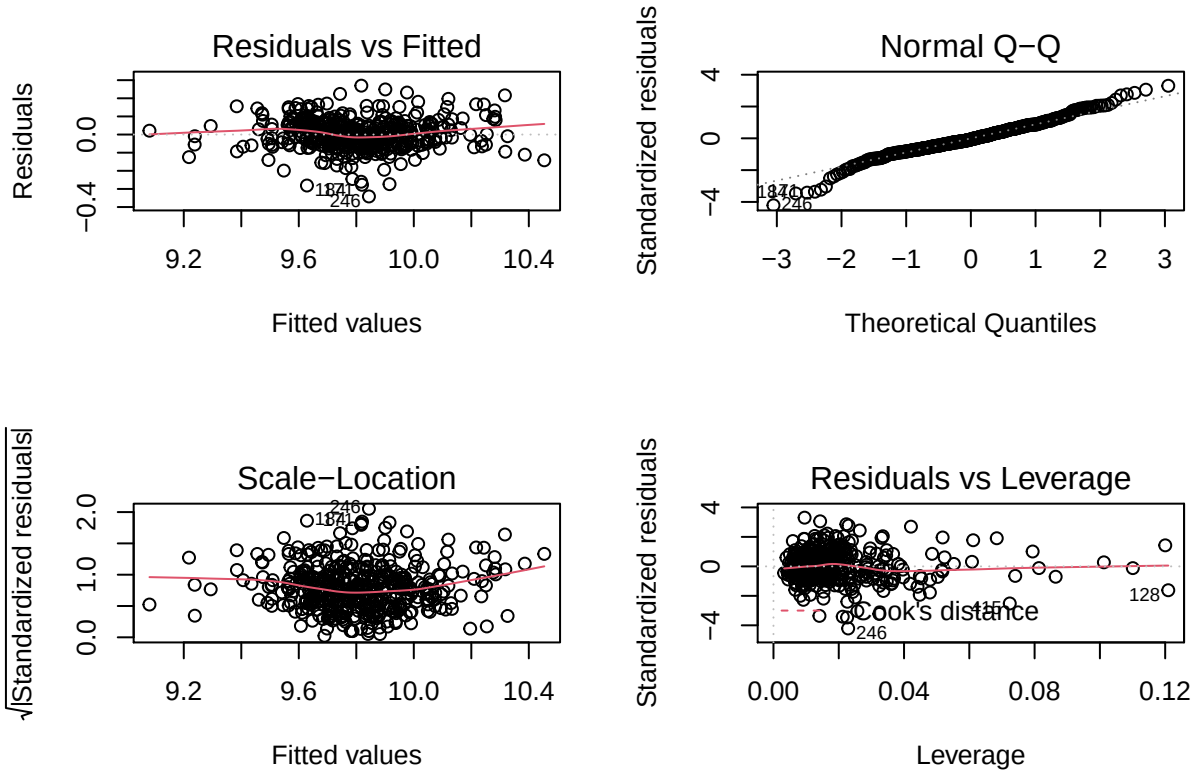
```
reg_summary$which[best.model,]
```

```
##   (Intercept) log.land.area     pop.18_34   pop.65_plus   log.doctors
##          TRUE          TRUE          TRUE         FALSE          TRUE
## log.hosp.beds    log.crimes   pct.hs.grad  pct.bach.deg pct.below.pov
##         FALSE         FALSE          TRUE          TRUE          TRUE
##     pct.unemp
##          TRUE
```

```
tmp <- log_data_cont[,reg_summary$which[best.model,][-1]]
best_subset_model <- lm(log.per.cap.income ~ .,data=tmp)
summary(best_subset_model)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ ., data = tmp)
##
## 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 ***
## log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
## pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
## log.doctors    0.0606769  0.0040183  15.100  < 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 ***
## ---
## 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
```

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



All the predictors have coefficients significantly different from zero and the diagnostic plots look really good. However, most of the coefficients are small, and some seem to have the wrong sign (e.g. `pct.hs.grad` and `pct.unemp`). We need to rule out what is causing this weird affect.

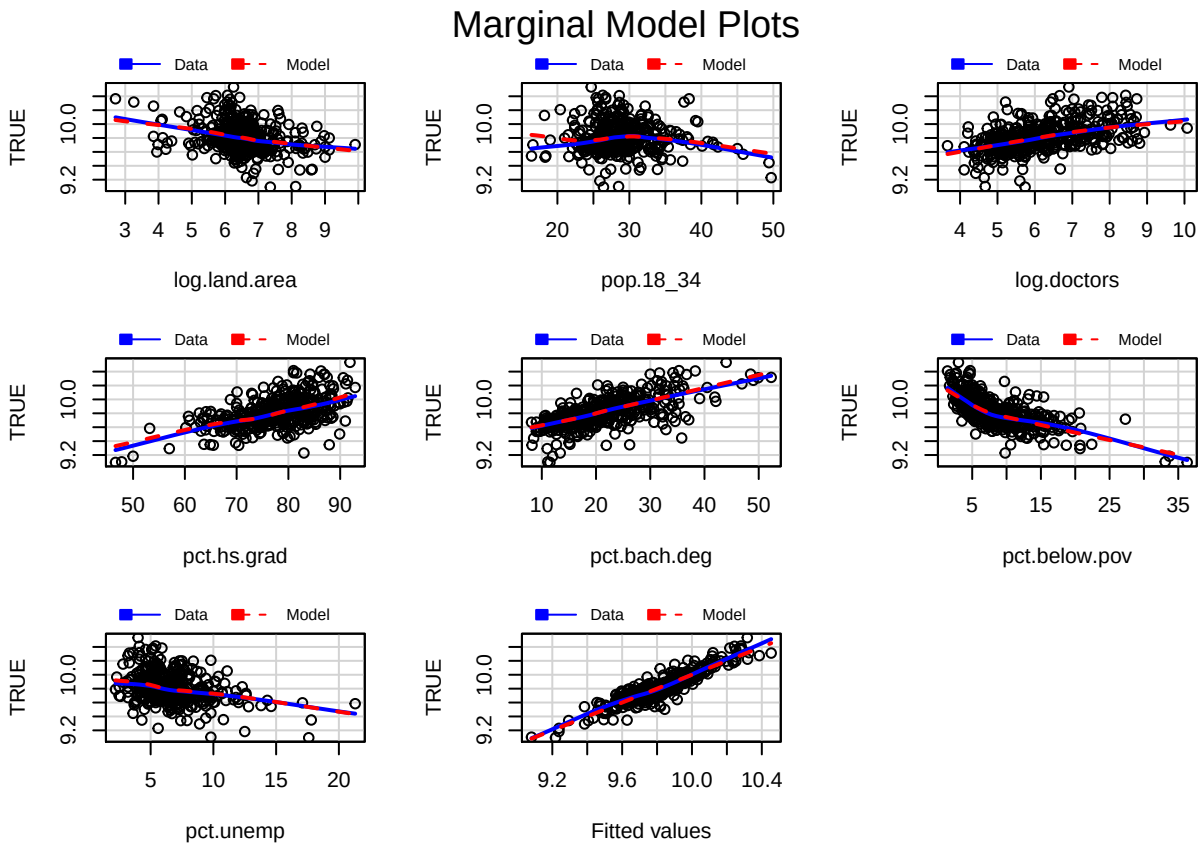
```
vif(best_subset_model)
```

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

Variance inflation is good since none of the variance inflation factors are above 5.

Maybe the marginal plots will tell us if a specific variable is causing the problem?

```
mmps(best_subset_model)
```



No - all the marginal plots looks good. We don't seem to be missing any important transformations, interactions, etc.

One last thing to try is to see if interaction with `region` helps in any way.

```
tmp <- cbind(tmp,region=data_trans$region)
subset_region <- lm(log.per.cap.income ~ .*region,data=tmp)
summary(subset_region)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ . * region, data = tmp)
##
## 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 ***
## log.land.area           -0.0364187  0.0151355  -2.406 0.016564 *
## pop.18_34               -0.0147940  0.0026043  -5.681 2.55e-08 ***
## log.doctors              0.0544169  0.0093221   5.837 1.08e-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.land.area:regionNE -0.0037179  0.0201435  -0.185 0.853656
## log.land.area:regionS  -0.0047582  0.0174155  -0.273 0.784825
## log.land.area:regionW   0.0151234  0.0181871   0.832 0.406154
## 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
## log.doctors:regionNE   -0.0046251  0.0132571  -0.349 0.727359
## log.doctors:regionS     0.0043337  0.0114401   0.379 0.705019
## log.doctors:regionW    -0.0034863  0.0131576  -0.265 0.791173
## 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 **
## ---
## 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
```

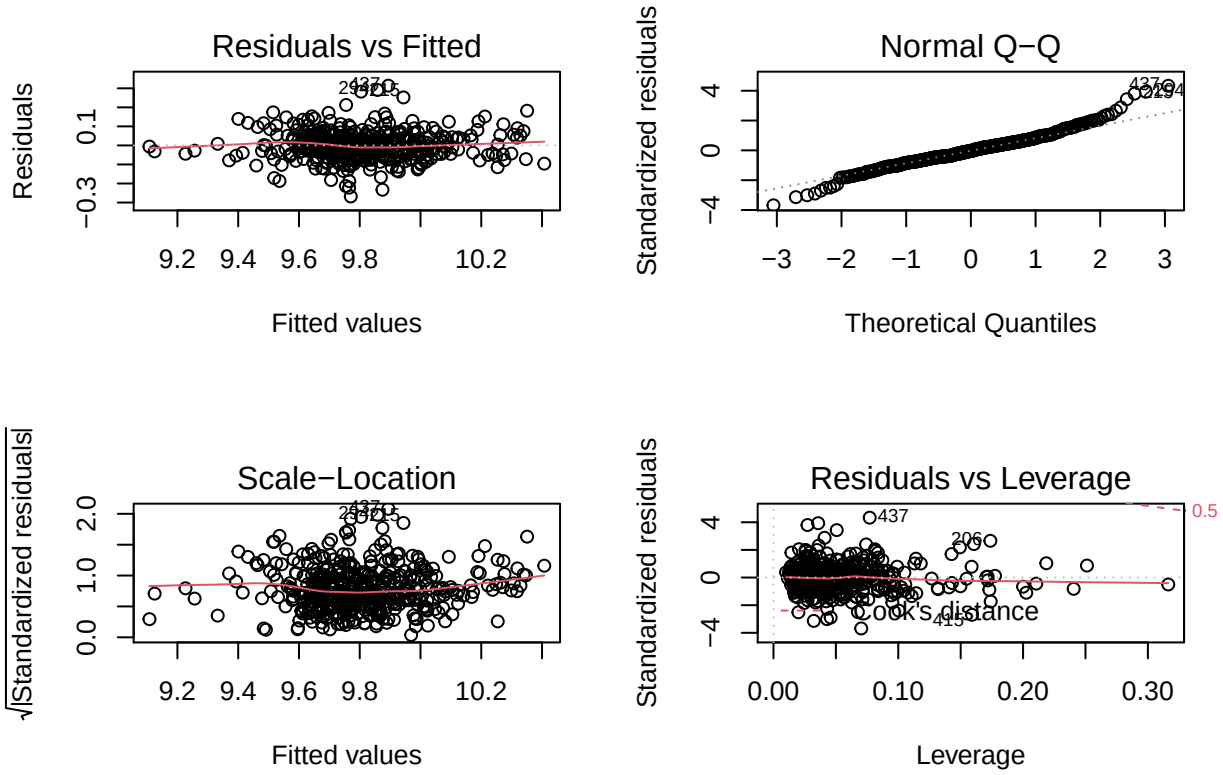
A handful of the interaction terms seem to be statistically significant. Therefore, the interaction terms that have at least one significant variable between the three regions.

```
best_subset_region_inter <- update(subset_region,. ~ . - region:log.land.area - region:pop.18_34 - regi
summary(best_subset_region_inter)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp + region + pct.hs.grad:region + pct.bach.deg:region +
##     pct.below.pov:region + pct.unemp:region, data = tmp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.268015 -0.043459 -0.002511  0.039967  0.313939
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.125260   0.251582  40.246  < 2e-16 ***
## log.land.area          -0.034569   0.005376  -6.430 3.50e-10 ***
```

```
## pop.18_34              -0.015404   0.001087 -14.170  < 2e-16 ***
## log.doctors             0.055342   0.004034  13.720  < 2e-16 ***
## pct.hs.grad            -0.002503   0.003151  -0.794 0.427456
## pct.bach.deg            0.014208   0.002108   6.741 5.24e-11 ***
## pct.below.pov          -0.023634   0.003351  -7.054 7.30e-12 ***
## pct.unemp               0.017787   0.004783   3.719 0.000228 ***
## regionNE                0.219429   0.302526   0.725 0.468661
## regionS                -0.062648   0.276125  -0.227 0.820627
## regionW                 1.629351   0.357633   4.556 6.86e-06 ***
## pct.hs.grad:regionNE   -0.003640   0.003876  -0.939 0.348271
## pct.hs.grad:regionS     0.002014   0.003539   0.569 0.569690
## pct.hs.grad:regionW    -0.018916   0.004204  -4.499 8.85e-06 ***
## pct.bach.deg:regionNE   0.005905   0.002618   2.256 0.024611 *
## pct.bach.deg:regionS   -0.001298   0.002321  -0.559 0.576352
## pct.bach.deg:regionW    0.006326   0.002620   2.415 0.016183 *
## pct.below.pov:regionNE -0.002435   0.004647  -0.524 0.600488
## pct.below.pov:regionS   0.007137   0.003686   1.937 0.053482 .
## pct.below.pov:regionW  -0.015224   0.005169  -2.945 0.003407 **
## pct.unemp:regionNE     -0.007967   0.007255  -1.098 0.272761
## pct.unemp:regionS      -0.024668   0.006377  -3.868 0.000127 ***
## pct.unemp:regionW      -0.019757   0.006603  -2.992 0.002935 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07545 on 417 degrees of freedom
## Multiple R-squared:  0.8735, Adjusted R-squared:  0.8668
## F-statistic: 130.9 on 22 and 417 DF,  p-value: < 2.2e-16
```

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



The signs of the variables we were worried about have not changed. The diagnostic plots and regression statistics are also comparable. Let's use ANOVA to see if they should be included in the model and look at BIC and AIC.

```
anova(best_subset_model,subset_region)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ (log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     region) * region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    432 2.9051
## 2    408 2.3502 24   0.55491 4.0139 2.307e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(best_subset_model,subset_region)
```

```
##                    df      AIC
## best_subset_model   9 -942.274
## subset_region      33 -987.542
```

```
BIC(best_subset_model,subset_region)
```

```
##                    df       BIC
## best_subset_model   9 -905.4931
## subset_region      33 -852.6784
```

The ANOVA test concludes that the interaction terms should be included in the model since the p-value for the model with the interaction terms is significantly less than 0.05. This result aligns with the AIC value which is lower for the model with the interaction terms. However, the BIC value favored the model without the interaction terms and the model with the interaction terms still didn't change the signs for the two variables that we were concerned with. Plus, the diagnostic plots and regression statistics are so similar to the model without the interaction terms. Since interaction terms complicate the interpretation of the model, it's better to exclude them since they don't provide enough value for their inclusion to be worth it.

```
formula(best_subset_model)
```

```
## log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
```

```
round(summary(best_subset_model)$coef,2)
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      10.22       0.09  109.78        0
## log.land.area    -0.04       0.00   -7.47        0
## pop.18_34        -0.01       0.00  -12.51        0
## log.doctors       0.06       0.00   15.10        0
## pct.hs.grad       0.00       0.00   -4.07        0
## pct.bach.deg      0.02       0.00   16.64        0
## pct.below.pov    -0.02       0.00  -19.29        0
## pct.unemp         0.01       0.00    4.87        0
```

Therefore, the final best subset model contains 7 variables which have the following intepretation.

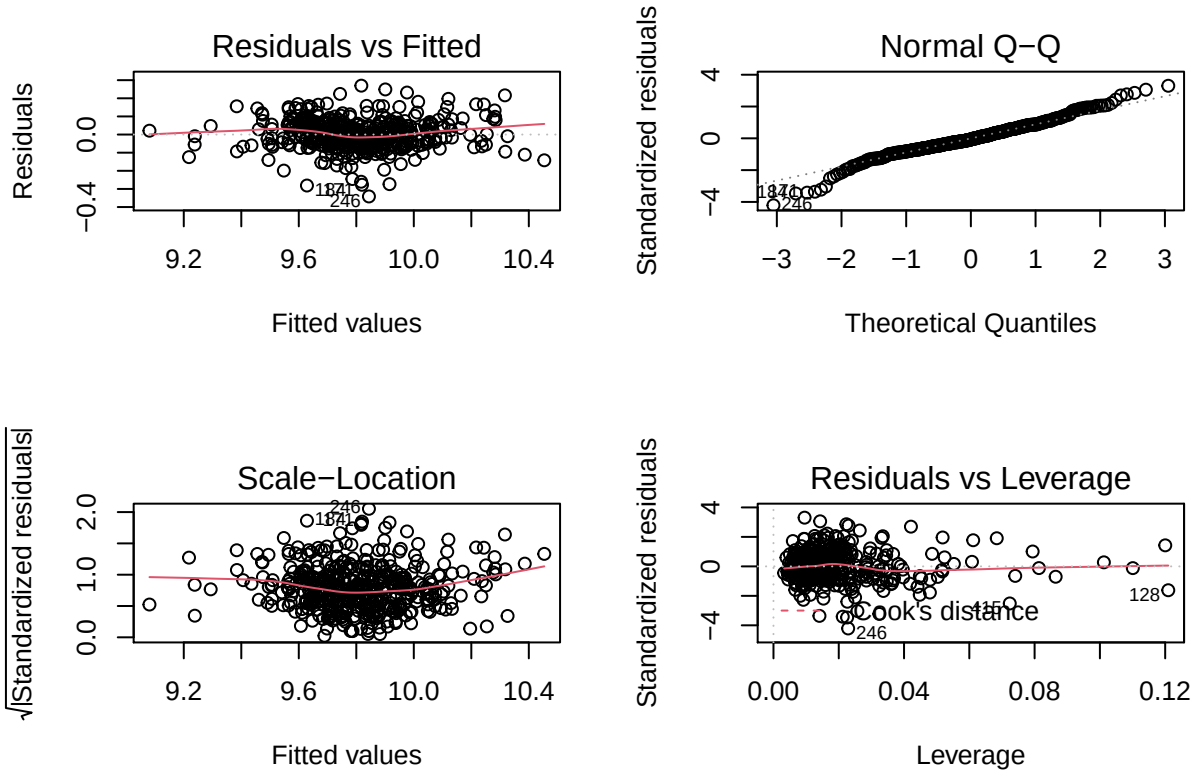
- For every 1% increase in a county's land area, there is a 0.04% decrease in expected per-capita income. (We might conjecture that this could be due to an urban-rural contrast: rural counties tend to be bigger than urban ones).
- For every 1% increase in the number of doctors in a county, the expected per-capita income increases by about 0.06%. That makes sense; doctors are well-paid and could be big contributors to the per-capita average income.
- Percent of the population that are high school graduates doesn't have much effect, except in the South, where a one percentage point increase in hs graduates induces an expected 2% decrease in per-capita income.

Part H - Stepwise

```
stepwise_model <- stepAIC(lm(log.per.cap.income ~., data = log_data_cont), direction = 'both', k = log(
summary(stepwise_model)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp, data = log_data_cont)
##
## 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 ***
## log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
## pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
## log.doctors    0.0606769  0.0040183  15.100  < 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 ***
## ---
## 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
```

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



Yay the stepwise method returns the same model as the best subset method! We will make the same conclusions about the interaction terms.

Part I - Lasso

```
# Lasso removes more variables
x <- as.matrix(dplyr::select(log_data_cont, !c(log.per.cap.income))) # remove region since it's a facto
result <- cv.glmnet(x,log_data_cont$log.per.cap.income, alpha = 1)
# plot(result)
c(lambda.1se=result$lambda.1se,lambda.min=result$lambda.min)
```

```
##   lambda.1se   lambda.min
## 0.0064883132 0.0005262871
```

```
cbind(coef(result),coef(result,s=result$lambda.1se), coef(result,s=result$lambda.min))
```

```
## 11 x 3 sparse Matrix of class "dgCMatrix"
##                            1            1            1
## (Intercept)    9.878369962  9.878369962 10.249164856
## log.land.area -0.032063002 -0.032063002 -0.035735118
## pop.18_34     -0.011810866 -0.011810866 -0.015066793
## pop.65_plus    .            .           -0.002580217
## log.doctors    0.059230219  0.059230219  0.051704309
```

Table 5:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.222	0.093	109.776	0
log.land.area	-0.036	0.005	-7.468	0
pop.18_34	-0.014	0.001	-12.508	0
log.doctors	0.061	0.004	15.100	0
pct.hs.grad	-0.004	0.001	-4.071	0
pct.bach.deg	0.015	0.001	16.641	0
pct.below.pov	-0.024	0.001	-19.294	0
pct.unemp	0.011	0.002	4.871	0

```
## log.hosp.beds   .            .            0.012082909
## log.crimes      .            .            .
## pct.hs.grad     .            .           -0.004214225
## pct.bach.deg    0.011645778  0.011645778  0.015328825
## pct.below.pov -0.019928341 -0.019928341 -0.024446716
## pct.unemp       0.005894554  0.005894554  0.010632241
```

```
# Also tested with region
# lasso_model <- lm(log.per.cap.income ~ log.land.area + log.doctors + pct.bach.deg + pct.below.pov + p
# summary(lasso_model)
#
# par(mfrow = c(2,2))
# plot(lasso_model)
```

Yay the Lasso method also agrees with the best subset model! We will make the same conclusions about the interaction terms.

Part J - Final Model

```
formula(best_subset_model)
```

```
## log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
```

```
round(summary(best_subset_model)$coef,3) %>%
  round(digits=4) %>% kbl(booktabs=T,caption=" ") %>% kable_classic()
```

Part K - Tradeoffs

No matter what you do, you are likely to be unhappy with some or all of these criteria; the better you make one criterion, the worse another is likely to get. So you will have to find a compromise or tradeoff between these criteria. Explain how you decided to make the tradeoff(s) you made.

The focus of the model selection process was on finding a model that could be used for inference and interpretation since that is the focus of the research questions. This goal resulted in many tradeoffs that favored simplicity over accuracy.

To start, there were many variables that had an ideal transformation that was an unreasonable power to interpret. For example, the optimal power for land area was found to be around -0.05, which is not easy to interpret to business collaborators. In order to still try and meet the model assumptions, the best common transformation was performed on the variables. Additionally, some variables can be transformed multiple ways to reach the normal distribution. But to keep the interpretation simple, the same transformations were chosen so that only two types of transformations were applied to the data.

Then, there were two variables that had a high variance inflation factor: population and total income. It makes sense that these two variables were accounting for the same variance since they both account for the population of the area. Consequently, one of the variables was removed from model consideration since the multicollinearity interferes with the interpretation of the model. Since the response is already a measure of income, the total income variable was removed.

Next, interactions terms were tested to see if region should be included in the final model. Even though they complicate the model, the weren't too many to make the interpretation too difficult.

Also, the state and county variables were not considered in the model since there are so many possible categories between them it was not useful to the model. It might be useful to see if they improve the model at all in the future.

Finally, in the variable selection process a tradeoff was made while choosing the selection criterion. BIC was chosen as the selection criterion since the research question is geared towards selecting the 'true' model over prediction which is better for AIC.