The Relationship between Per Capita Income and the County's Economic, Health and Social Well-being

Zixuan Jin

Department of Statistics and Data Science

Carnegie Mellon University

zixuanji@andrew.cmu.edu

Abstract

We are interested in analyzing the association between average income per person and variables related to economic, health, and social well-being. We examined the selected county demographic information data for 440 of the most populous counties in the United States from the years 1990 and 1992. We performed correlation calculations between each pair of variables in our dataset to see the relationship between these variables. We created regression models with possible transformations, interactions based on F-tests, AIC, BIC, and VIF values to predict per-capita income from one or multiple county demographic variables. It appears that the land area, the percent of unemployment, the percent of population aged 18-34, the number of doctors, the percent of with income below poverty level, and geographic region are better in predicting the per capita income. However, it is also surprising to see the negative correlations between per capita income and unemployment or the percent of people aged 18-34. The missing counties or counties are not a primary concern in our analyses. We found the medical level is strongly correlated with the average per-capita income. We need more observation for individual counties and states for future investigations.

1 Introduction

The average income per person measures the economic well-being of the person. How to increase the average income per person has always been one of the significant concerns of society. Learning what is associated with average income per person helps to provide practical suggestions in solving the economic problem. In this report, we will investigate the relationship between average income per person and some variables that indicate economic, health, and social well-being at the county level. Looking at the county level provides more detailed information that could affect the average income per person.

Specifically, we will

- Illustrate the relationship between any two variables in our analyses
- Identify the relationship between per-capita income and crime rate specifically; identify if this relationship would differ in different regions
- Develop a regression model to predict per-capita income from other economic, health, and social well-being related variables
- Discuss how missing states or missing counties in the data might affect our analysis

2 Data

The data provides selected county demographic information (CDI) for 440 of the most populous counties in the United States from the years 1990 and 1992. We took the data from Kutner et al. (2005). The data are given in the file cdi.dat in the files area for our course on Canvas. Each data set line has an identification number with a county name and state abbreviation and provides information on fourteen variables for a single county. Of these fourteen variables, thirteen variables are numeric and one variable describes the county's region. A detailed description of the variables in the dataset are shown in Table 1.

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)

The id variable here is the same as the row number in our data, and we did not include it in our data analyses.

The summary statistics for the quantitative variables in the data are given in Table 2 below.

Table 2: Summary statistics for continuous variables in CDI dataset

	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
per.cap.crime	0.0	0.04	0.05	0.06	0.07	0.3	0.03

In Table 2, there are several variables with mean substantially larger than median (land.area, pop, doctors, hosp.beds, crimes, per.cap.income, and tot.income), indicating possible right-skewing in their distributions. There are no variables with a mean substantially smaller than the median.

Since all categorical variables in the data describe the geographic location, the summary statistics of one categorical variable could at least provide us some geographic information of the data. The summary statistics for the categorical variable region in the data are given in Table 3 below. We see that most counties are in the South (region 'S'), and the least are in the west (region 'W').

Table 3: Summary statistics of geographic region in CDI dataset

	NC	NE	S	W
Freq	108	103	152	77

3 Methods

3.1 Relationship among variables in the dataset

Our analyses consist of three parts. First, we used the correlation matrix to investigate the relationship between any two variables in our data. As the per capita income is the main interest of the response variable in our data, we also checked to see how other predictors in our data are related to the per capita income.

3.2 Per-capita income, crime, and region

Second, to develop a regression model to predict the per capita income from crime rate and region of the county, we considered regression models using untransformed per capita income and crime rate or per capita crime rate with or without interaction with geographic region. We chose our final model based on F-tests, AIC, and BIC values of each regression model, accompanied by a summary of each regression analysis and an examination of residual diagnostic plots.

3.3 Predicting per-capita income from all other variables

Moreover, to develop a regression model to predict the per capita income from the other variables in our data, we started with a complete model using logarithmic, and Box-Cox power transforms of the right-

skewed predictors and power transforms the left-skewed predictors without region. We found all subsets of the whole model. We chose our final model based on the minimum BIC value, accompanied by a summary of each regression analysis and an examination of residual diagnostic plots, VIF values, and marginal model plots. Then, we added the interactions with the region in our last model. We kept the interaction term based on the p-values. At last, we chose our final model based on F-tests, AIC, and BIC values, accompanied by a summary of each regression analysis and an examination of residual diagnostic plots.

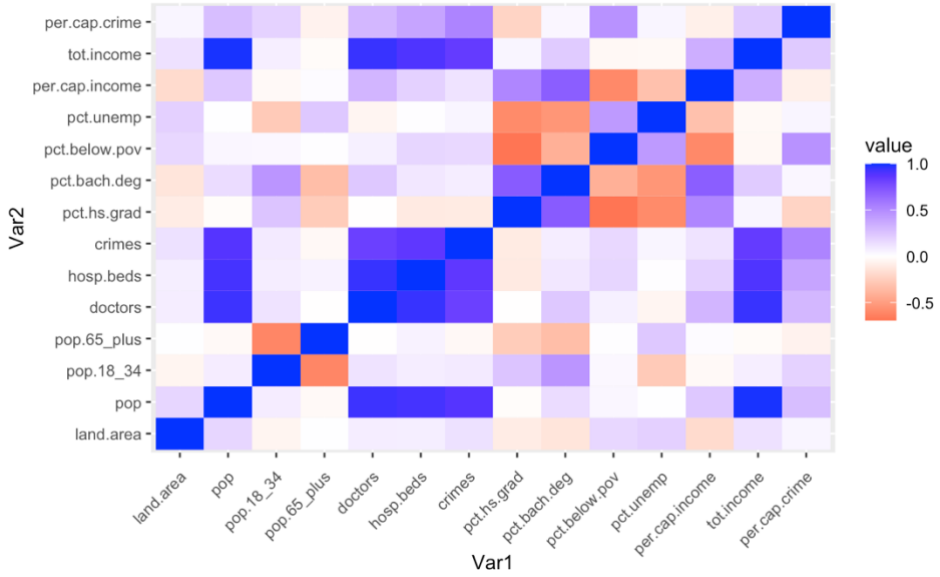
At last, as states are nested in the region, we fitted our final all-subsets regression model with state, and we omitted the region and the interaction with region here. We also chose our final model based on F-tests, AIC, and BIC values, accompanied by a summary of each regression analysis and an examination of residual diagnostic plots. We used the same criteria to compare this final model for the state with the last model for the region.

4 Results

4.1 Relationship Among Variables

We considered a correlation matrix presented in Figure 1 below to capture the correlation coefficient between any two original variables in our data. A value greater than zero indicates there is a positive relationship between the two variables. A value less than zero means there is a negative relationship between the two variables. An absolute correlation coefficient being one indicates a perfect linear correlation between the two variables.

Table 4: Correlation matrix of variables in CDI dataset



In Table 4, the tot.income and pop are highly correlated, which is not surprising as the tot.income is the income of all the population in a county. These two variables are reasonably highly correlated with crimes, hosp.beds, and doctors since we need more doctors, number of beds, cribs, and bassinets for a larger population. The incidents of crimes are more frequent for a larger population. The three variables-crimes, hosp.beds, and doctors-seem strongly positively correlated with one another. The per.cap.income is not highly correlated with any variable. The pct.hs.grad or pct.bach.deg is positively correlated with the per.cap.income since people with higher education tend to more expertise in solving more difficult problems and then earn more. The pct.below.pov or pct.unemp is negatively correlated with the per.cap.income since we have less income in total and less income per capita if more people are in poverty or unemployed. These four variables pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp are moderately correlated with each other. Further exploratory data analyses on the relationship of each variable against per.cap.income are in Appendix A (see page 27 below).

4.2 Predicting Per Capita Income from Crime Rate and Region

We considered one regression model using the original variables per.cap.income, crimes, and region and another using per.cap.income, region, and created per.cap.crime which is the number of crimes over the

population. Then, we investigated whether the associations between per.cap.income and crimes or per.cap.crime differentiate across different regions.

4.2.1 Regression Models without Interaction

The two regression models with crime or per.cap.crime were the following models, shown with estimated regression coefficients:

$$\text{per.cap.income} = 18110 + 0.009 * \text{crime} + 2286*\text{regionNE} - 860.6 * \text{regionS} - 142.8*\text{regionW} \quad (1)$$

$$\text{per.cap.income} = 18006.04 + 5773.20*\text{per.cap.crime} + 2354.70*\text{regionNE} - 927.45*\text{regionS} - 34.92*\text{regionW} \quad (2)$$

Models (1) and (2) had similar low R2 values (0.09288 and 0.07782, respectively). Still, we detected a non-random pattern and highly influential points in the residual diagnostic plots for model (1) (Appendix B, see page 28 below). Model (2) has better residual diagnostic plots, and we prefer model (2), which uses per.cap.crime. Table 5 gives the full table of estimated coefficients and standard errors for model (2).

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	18006.04	537.04	33.528	< 2e-16	***
per.cap.crime	5773.20	7520.41	0.768	0.4431	
regionNE	2354.70	541.97	4.345	1.74e-05	***
regionS	-927.45	512.31	-1.810	0.0709	.
regionW	-34.92	586.03	-0.060	0.9525	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 5: Estimated coefficients and standard errors for model (2)

4.2.2 Regression Models with Interaction

To investigate whether the correlation between per.cap.income and crime differentiate across the different regions, we build on the model (2) with interaction terms between per.cap.crime and region. The model with interaction was shown below with estimated regression coefficients:

$$\text{per.cap.income} = 4379.1*\text{per.cap.crime} + 2329.0*\text{regionNE} - 1010.4*\text{regionS} - 670.0*\text{region} + 288.4*\text{per.cap.income}*\text{regionNE} + 1558.9*\text{per.cap.income}*\text{regionS} + 10655.5*\text{per.cap.income}*\text{regionW} \quad (3)$$

Based on the F-statistics in the ANOVA table in Appendix B (see page 44 below), there is little evidence to support the full model (3). We were happy to adopt the reduced model (2) that contains no interactions between per.cap.crime and region. Table 6 gives the full table of estimated coefficients and standard errors for model (3). In Table 6, the coefficients for the interaction terms between per-capita income and region is not statistically significant at any level. Then, the per-capita income varies with the region in the United States, but the way it is related to crime does not.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	18077.3	895.2	20.193	<2e-16	***
per.cap.crime	4379.1	15893.5	0.276	0.783	
regionNE	2329.0	1101.4	2.115	0.035	*
regionS	-1010.4	1323.8	-0.763	0.446	
regionW	-670.0	1983.9	-0.338	0.736	
per.cap.crime:regionNE	288.4	20184.7	0.014	0.989	
per.cap.crime:regionS	1558.9	20556.1	0.076	0.940	
per.cap.crime:regionW	10655.5	32322.4	0.330	0.742	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Table 6: Estimated coefficients and standard errors for model (3)

Based on Table 5 above, for one unit increase in the number of crime per capita, the per capita income increases 5773.2 dollars on average. The average per capita income in the North Central region is 18006.04 dollars. The average per capita income in the Northeastern region is 20360.74 dollars. The average per capita in the South is 17078.59 dollars. The average per capita income in the West region is 17971.12 dollars.

4.3 Predicting Per capita income from other variables

4.3.1 Logarithmic and Power Transformations

To address right skewness, potential leverage, and influence issues, we took the logarithms of the variables we identified with some or severe right skewness in their univariate distributions (Appendix C, see pages 46 to 59). To address left skewness, potential leverage, and influence issues, we raised the power of the variables we identified with some severe left skewness in their univariate distributions (Appendix C, see page 53).

Before we go through our model selections, we should first mention the inclusion of three categorical variables: county, state, and region in our models. Looking at the unique values for the county (Appendix

A, see pages 17 and 18 below), it has 373 unique values, which is nearly as many as the rows in the CDI dataset (440). A graph describing the combination of county and state showed that we only had one observation per unique county, so the county is not a valid variable to include in our models (Appendix A, see pages 18 to 21).

4.3.2 Variable selection using All Subsets

We started with fitting the following full model

$$\text{per.cap.income} = \log(\text{land.area}) + \log(\text{per.cap.crime}) + \log(\text{pct.unemp}) + \log(\text{pct.below.pov}) + \log(\text{pct.bach.deg}) + (\text{pct.hs.grad})^3 + \log(\text{hosp.beds}) + \log(\text{doctors}) + \log(\text{pop.65\_plus}) + \log(\text{pop.18\_34}) + e \quad (4)$$

Based on the inverse response plot and Box-Cox statistics, we found no need to transform the response variable per.cap.income. A detailed graphical summary of using all possible subsets can be found in Appendix C (see pages 62 to 72 below). We chose our final model with a minimum BIC value (Appendix C, see page 63). The final model is shown below with estimated regression coefficients:

$$\begin{aligned}\text{per.cap.income} = {} & 32700.93 - 681.97 * \log(\text{land.area}) + 1712.85 * \log(\text{pct.unemp}) \\ & - 4165.28 * \log(\text{pct.below.pov}) + 6137.60 * \log(\text{pct.bach.deg}) - 84.97 * (\text{pct.hs.grad})^3 \\ & + 1111.08 * \log(\text{doctors}) - 6863.47 * \log(\text{pop.18\_34}) \end{aligned} \quad (5)$$

Table 7 gives the full table of estimated coefficients and standard errors for model (5).

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       32700.93    2806.52  11.652  < 2e-16 ***
log(land.area)     -681.97     100.57  -6.781 3.93e-11 ***
log(pct.unemp)     1712.85     339.06   5.052 6.47e-07 ***
log(pct.below.pov) -4165.28     229.08 -18.183  < 2e-16 ***
log(pct.bach.deg)  6137.60     485.54  12.641  < 2e-16 ***
pct.hs.grad         -84.97      21.70  -3.916 0.000105 ***
log(doctors)       1111.08      91.42  12.153  < 2e-16 ***
log(pop.18_34)    -6863.47     729.42  -9.409  < 2e-16 ***
```

Table 7: Estimated coefficients and standard errors for model (5)

The fitted model match well with the raw data in the marginal model plots of the model (5) (Appendix C, see page 66 below). The coefficients on percent of unemployment, percent of high school graduates and

percent of population aged 1-34 do not meet what we expected. One possible reason would be that we did not consider geographic effects.

4.3.3 Regression Model with Region

We started with interacting all the variables in the final subset model with region. We kept the whole categorical variable if any indicator for a categorical variable or an interaction with categorical variable is statistically significant coefficient. If none of them is statistically significant, we dropped the variable or the interaction of the variable with the categorical variable. The VIF values are not excessively large and only two of them exceed five (Appendix C, see page 69 below). The F-statistics, AIC, and BIC values in Appendix C suggested that we should choose the model with interaction between some variables and region (see pages 70 and 71 below).

4.3.4 Regression Model with State

Then, we considered model (5) with state and interaction with the state. The ANOVA table, AIC, and BIC values in Appendix D suggest choosing the model with the state but without interaction (see pages 75 and 76 below). A full table of estimated coefficients and standard errors for our model with state is listed in Appendix D (see pages 76 and 77 below).

4.3.5 Final Model

Then we compared the model with some interaction with the region and the model with the state. Looking at the diagnostic plots for the model with state, we saw more deviation in the normal Q-Q plots, and there are some highly influential points in the leverage graph. In addition, the VIF values of the model with state suggest that adding the categorical variable state would result in a severe multicollinearity issue here (Appendix D, see page 80 below). Considering the correlations between the predictor and the response variable across the region, we chose the model with some interaction with the region even though the ANOVA table and AIC values prefer the model with the state (Appendix D, see page 81 below).

The following table gives the full table of estimated coefficients and standard errors for our final model.

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  3.360e+04  4.645e+03   7.235 2.25e-12 ***
log.land.area               -7.042e+02  1.160e+02  -6.070 2.88e-09 ***
log.pop.18_34               -7.613e+03  1.676e+03  -4.543 7.28e-06 ***
log.doctors                  9.303e+02  9.476e+01   9.818  < 2e-16 ***
pct.hs.grad.3               -1.494e-03  3.493e-03  -0.428 0.669130
log.pct.bach.deg             4.676e+03  1.022e+03   4.574 6.33e-06 ***
log.pct.below.pov           -3.094e+03  5.243e+02  -5.901 7.49e-09 ***
log.pct.unemp                1.264e+03  3.599e+02   3.513 0.000491 ***
regionNE                     6.507e+03  6.535e+03   0.996 0.319934
regionS                     -7.750e+03  5.118e+03  -1.514 0.130705
regionW                      8.801e+02  7.303e+03   0.120 0.904145
log.pop.18_34:regionNE      -4.386e+03  2.371e+03  -1.849 0.065105 .
log.pop.18_34:regionS        1.143e+03  1.926e+03   0.594 0.553079
log.pop.18_34:regionW        2.167e+03  2.495e+03   0.869 0.385500
pct.hs.grad.3:regionNE      -6.806e-03  4.468e-03  -1.523 0.128469
pct.hs.grad.3:regionS       -4.696e-03  4.207e-03  -1.116 0.264941
pct.hs.grad.3:regionW       -1.791e-02  4.387e-03  -4.082 5.34e-05 ***
log.pct.bach.deg:regionNE    4.292e+03  1.285e+03   3.340 0.000913 ***
log.pct.bach.deg:regionS     2.482e+03  1.147e+03   2.164 0.030997 *
log.pct.bach.deg:regionW     3.229e+03  1.374e+03   2.350 0.019231 *
log.pct.below.pov:regionNE -7.567e+02  7.117e+02  -1.063 0.288320
log.pct.below.pov:regionS  -8.122e+02  6.451e+02  -1.259 0.208756
log.pct.below.pov:regionW  -3.949e+03  9.132e+02  -4.324 1.92e-05 ***
```

Table 8: Estimated coefficients and standard errors for the final model with region and interaction.

The model explains 83.8% of the variations in the per-capita income around its mean. Looking at the residual diagnostic plots of this model, we found that the assumptions such as the normality of errors, constant variance of errors are followed, and there are no highly influential points in our model. In this case, we believe the final model is an appropriate fit.

Given Table 8, for every 1% increase in a county's land area, the average income per person decreases by about 7.04 dollars on average. For every 1% increase in the percent of the population aged 18-34, the average per capita income decreases by 76.13 dollars on average. We might conjecture that people aged 18-34 are not at peak earning capacity yet, and so perhaps their lower incomes drag down the per-capita average. For every 1% increase in doctors in a county, the average per capita income increases by 9.30 dollars on average. The percent of the population that are high school graduates doesn't have much effect. It might depend on whether college graduates are counted as a subset of high school graduates rather than calculating them separately. For every 1% increase in the percent of the adult population (persons 25 years old or older) with a bachelor's degree, the average income per person increases by about 46.76 dollars on average. For every 1% increase in the population with income below the poverty level, the average per capita income decreases by 30.94 dollars on average. For every 1% increase in the percent of unemployed people, the average per capita income increases by about 12.64 dollars on average. In several

of the interactions for the region, the West shows up as deviating significantly from the North Central part of the United States.

4.4 Missing counties or states in our analyses

Now we will address how the missing counties or the missing states would affect our analyses. The main idea would be whether our data is representative of the overall population. Recall the fact that there are 50 states in the United States, plus the District of Columbia. From Table 9, there are 48 states, and we found Arkansas, Iowa, and Wyoming are missing from the dataset. It would be better to include data from these three states. However, given the VIF values of the model with the state (Appendix D, see page 80 below), it suggests that adding more state values could not provide us with any critical information as different levels are strongly correlated with each other. In this case, the missing states would not result in a significant impact on our analyses.

For the counties, 373 out of 3000 counties in the United States are included in our data given in Table 9. Recall that we only had one observation per unique county. In this case, we failed to capture any associations between per-capita income and counties in our data. Given the summary table of the number of counties in each state in Appendix D (see page 82 below), California has the most number of counties, which is consistent with the population in California from 1970 census data (Wikipedia contributors 2021). The seven states with only one county are also less populous from 1970 census data (Wikipedia contributors 2021). It shows that the data sampled more counties from more populous states and fewer counties from less populated states. Hence, we believe our data is still representative of the overall population. In addition, adding more counties might increase the correlation among the variables in our model. We believe the missing counties would not influence our analyses. But, considering the variations among counties in a state, it would be better to include more counties in our data. To capture the effects of the counties, we need more observations per unique county.

Table 9: The Count of Unique Values for State and County in CDI dataset

Variables	id	state	county
Count	440	48	373

5 Discussion

In this study, we expect to investigate how the average per-capita income was other variables associated with the county's economic, health, and social well-being using historical data. We mainly addressed four research questions with different methods. For the first research question, we did not find any surprising relationship between each two pair of variables in our data. We utilized a regression model to analyze the relationship between crime rate, region, and per-capita income for the second research question. We found a positive relationship between the crime rate and per-capita income, and the per-capita income varies with the regions in the United States. But the association between per-capita income and crime rate does not change with the regions. Our next research question is to predict per-capita income from all other variables. The optimal model that fits per-capita income and other variables best: it includes log-transformed land area, percent of the population aged 18-34, number of doctors, percent of people of bachelor agree, percent of people with income below poverty level, percent of unemployment, power transformed percent of high school graduates, region and their interaction with the region. Finally, we addressed the impacts of the missing states and counties in our analyses. We believe there is no need to worry about the missing values as the data is representative of the overall population.

Given the optimal model in 4.3.5, it is evident that there is a strong positive correlation between the medical level and the average per-capita income. It is surprising to notice a negative correlation between the percent of unemployment and per capita income. It could be the case that we missed some crucial variables related to both unemployment and income. In addition, it could be the case that we failed to capture the county or state effects in the percent of unemployment and per-capita income since we do not include county or state variables in our models. We should require a further investigation into this relationship. A fundamental limitation of this paper is that we cannot refer our model to a more specific place such as a county or a state since we do not have enough observations for a unique county or state. For future analysis, it would be beneficial to perform our study in more extensive data and fit the model using data from another year to see if the relationship among the average per-capita income and other variables still holds.

References

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

Sheather, S. 2009. A Modern Approach to Regression With R. Springer Science & Business Media.

Wikipedia contributors. 2021. "1970 United States Census." *Wikipedia, The Free Encyclopedia*. Retrieved October 28, 2021 (https://en.wikipedia.org/w/index.php?title=1970_United_States_census&oldid=1048003046).

Technical Appendix

Reading Data

```
cdi <- read.table("/Users/rosahhh/Desktop/Fall 2021/36617 Applied Linear Model/project 01/cdi.dat")

attach(cdi)
```

Appendix A

```
pandoc.table(head(cdi[,1:10]), caption="Partial Summary of CDI dataset", style="grid")
```

```
##
##
## +----+-------------+-------+-----------+---------+-----------+
## | id |   county    | state | land.area |   pop   | pop.18_34 |
## +====+=============+=======+===========+=========+===========+
## | 1  | Los_Angeles |  CA   |   4060    | 8863164 |   32.1    |
## +----+-------------+-------+-----------+---------+-----------+
## | 2  |    Cook     |  IL   |    946    | 5105067 |   29.2    |
## +----+-------------+-------+-----------+---------+-----------+
## | 3  |   Harris    |  TX   |   1729    | 2818199 |   31.3    |
## +----+-------------+-------+-----------+---------+-----------+
## | 4  |  San_Diego  |  CA   |   4205    | 2498016 |   33.5    |
## +----+-------------+-------+-----------+---------+-----------+
## | 5  |   Orange    |  CA   |    790    | 2410556 |   32.6    |
## +----+-------------+-------+-----------+---------+-----------+
## | 6  |    Kings    |  NY   |    71     | 2300664 |   28.3    |
## +----+-------------+-------+-----------+---------+-----------+
##
## Table: Partial Summary of CDI dataset (continued below)
##
##
##
## +-------------+---------+-----------+--------+
## | pop.65_plus | doctors | hosp.beds | crimes |
## +=============+=========+===========+========+
## |     9.7     |  23677  |   27700   | 688936 |
## +-------------+---------+-----------+--------+
## |    12.4     |  15153  |   21550   | 436936 |
## +-------------+---------+-----------+--------+
## |     7.1     |  7553   |   12449   | 253526 |
## +-------------+---------+-----------+--------+
## |    10.9     |  5905   |   6179    | 173821 |
## +-------------+---------+-----------+--------+
## |     9.2     |  6062   |   6369    | 144524 |
## +-------------+---------+-----------+--------+
## |    12.4     |  4861   |   8942    | 680966 |
## +-------------+---------+-----------+--------+
```

```
pandoc.table(head(cdi[,c(1,11:17)]), caption="Partial Summary of CDI dataset
(Continued)",
             style="grid")
```

```
##
##
## +----+-------------+--------------+---------------+-----------+
## | id | pct.hs.grad | pct.bach.deg | pct.below.pov | pct.unemp |
## +====+=============+==============+===============+===========+
## | 1  |     70      |     22.3     |     11.6      |     8     |
## +----+-------------+--------------+---------------+-----------+
## | 2  |    73.4     |     22.8     |     11.1      |    7.2    |
## +----+-------------+--------------+---------------+-----------+
## | 3  |    74.9     |     25.4     |     12.5      |    5.7    |
## +----+-------------+--------------+---------------+-----------+
## | 4  |    81.9     |     25.3     |      8.1      |    6.1    |
## +----+-------------+--------------+---------------+-----------+
## | 5  |    81.2     |     27.8     |      5.2      |    4.8    |
## +----+-------------+--------------+---------------+-----------+
## | 6  |    63.7     |     16.6     |     19.5      |    9.5    |
## +----+-------------+--------------+---------------+-----------+
##
## Table: Partial Summary of CDI dataset (Continued) (continued below)
##
##
##
## +----------------+------------+--------+
## | per.cap.income | tot.income | region |
## +================+============+========+
## |     20786      |   184230   |   W    |
## +----------------+------------+--------+
## |     21729      |   110928   |   NC   |
## +----------------+------------+--------+
## |     19517      |   55003    |   S    |
## +----------------+------------+--------+
## |     19588      |   48931    |   W    |
## +----------------+------------+--------+
## |     24400      |   58818    |   W    |
## +----------------+------------+--------+
## |     16803      |   38658    |   NE   |
## +----------------+------------+--------+
```

```
b <- apply(cdi,2,function(x) {length(unique(x))})
pandoc.table(b, caption="The Count of Unique Values for Variables in CDI data
set", style="grid")
```

```
##
##
## +-----+--------+-------+-----------+-----+-----------+-------------+
## | id  | county | state | land.area | pop | pop.18_34 | pop.65_plus |
```

```
## +=====+========+=======+===========+=====+==========+=============+
## | 440 |  373   |  48   |    384    | 440 |   149    |     137     |
## +-----+--------+-------+-----------+-----+----------+-------------+
##
## Table: The Count of Unique Values for Variables in CDI dataset (continued
below)
##
##
##
## +---------+-----------+--------+------------+--------------+
## | doctors | hosp.beds | crimes | pct.hs.grad | pct.bach.deg |
## +=========+===========+========+=============+==============+
## |   360   |    391    |  437   |     223     |     220      |
## +---------+-----------+--------+------------+--------------+
##
## Table: Table continues below
##
##
##
## +---------------+-----------+----------------+------------+--------+
## | pct.below.pov | pct.unemp | per.cap.income | tot.income | region |
## +===============+===========+================+============+========+
## |      155      |    97     |      436       |    428     |   4    |
## +---------------+-----------+----------------+------------+--------+

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"))

pandoc.table(tmp[1:30,], caption="Summary of County and State in CDI", style=
"grid")

##
##
## +--------------------+-------------------+------------------+
## |   Counties 1-110   | Counties 111-220  | Counties 221-330 |
## +====================+===================+==================+
## |       Ada ID       |     Ector TX      |   Lycoming PA    |
## +--------------------+-------------------+------------------+
## |      Adams CO      |   El_Dorado CA    |    Macomb MI     |
## +--------------------+-------------------+------------------+
## |      Aiken SC      |    El_Paso CO     |     Macon IL     |
## +--------------------+-------------------+------------------+
## |     Alachua FL     |    El_Paso TX     |    Madison AL    |
## +--------------------+-------------------+------------------+
## |     Alamance NC    |    Elkhart IN     |    Madison IL    |
## +--------------------+-------------------+------------------+
## |     Alameda CA     |      Erie NY      |    Madison IN    |
## +--------------------+-------------------+------------------+
## |      Albany NY     |      Erie PA      |   Mahoning OH    |
```

```
## +--------------------+-------------------+------------------+
## | Alexandria_City VA |    Escambia FL    |    Manatee FL    |
## +--------------------+-------------------+------------------+
## |    Allegheny PA    |     Essex MA      |   Marathon WI    |
## +--------------------+-------------------+------------------+
## |      Allen IN      |     Essex NJ      |   Maricopa AZ    |
## +--------------------+-------------------+------------------+
## |      Allen OH      | Fairfax_County VA |     Marin CA     |
## +--------------------+-------------------+------------------+
## |     Anderson SC    |   Fairfield CT    |    Marion FL     |
## +--------------------+-------------------+------------------+
## |   Androscoggin ME  |   Fairfield OH    |    Marion IN     |
## +--------------------+-------------------+------------------+
## |   Anne_Arundel MD  |    Fayette KY     |    Marion OR     |
## +--------------------+-------------------+------------------+
## |     Arapahoe CO    |    Fayette PA     |    Martin FL     |
## +--------------------+-------------------+------------------+
## | Arlington_County VA |    Florence SC    |     Maui HI      |
## +--------------------+-------------------+------------------+
## |     Atlantic NJ    |    Forsyth NC     |    McHenry IL    |
## +--------------------+-------------------+------------------+
## |    Baltimore MD    |   Fort_Bend TX    |    McLean IL     |
## +--------------------+-------------------+------------------+
## |  Baltimore_City MD |    Franklin OH    |   McLennan TX    |
## +--------------------+-------------------+------------------+
## |    Barnstable MA   |    Franklin PA    |  Mecklenburg NC  |
## +--------------------+-------------------+------------------+
## |       Bay FL       |   Frederick MD    |    Medina OH     |
## +--------------------+-------------------+------------------+
## |       Bay MI       |     Fresno CA     |    Merced CA     |
## +--------------------+-------------------+------------------+
## |      Beaver PA     |     Fulton GA     |    Mercer NJ     |
## +--------------------+-------------------+------------------+
## |       Bell TX      |   Galveston TX    |    Mercer PA     |
## +--------------------+-------------------+------------------+
## |      Benton WA     |     Gaston NC     |   Merrimack NH   |
## +--------------------+-------------------+------------------+
## |      Bergen NJ     |    Genesee MI     |   Middlesex CT   |
## +--------------------+-------------------+------------------+
## |      Berks PA      |   Gloucester NJ   |   Middlesex MA   |
## +--------------------+-------------------+------------------+
## |    Berkshire MA    |     Greene MO     |   Middlesex NJ   |
## +--------------------+-------------------+------------------+
## |    Bernalillo NM   |     Greene OH     |    Midland TX    |
## +--------------------+-------------------+------------------+
## |     Berrien MI     |   Greenville SC   |   Milwaukee WI   |
## +--------------------+-------------------+------------------+
##
## Table: Summary of County and State in CDI (continued below)
##
```

```
##
##
## +--------------------+
## |  Counties 331-440  |
## +====================+
## |   Rockingham NH    |
## +--------------------+
## |    Rockland NY     |
## +--------------------+
## |      Rowan NC      |
## +--------------------+
## |   Rutherford TN    |
## +--------------------+
## |   Sacramento CA    |
## +--------------------+
## |     Saginaw MI     |
## +--------------------+
## |    Salt_Lake UT    |
## +--------------------+
## | San_Bernardino CA  |
## +--------------------+
## |    San_Diego CA    |
## +--------------------+
## |  San_Francisco CA  |
## +--------------------+
## |   San_Joaquin CA   |
## +--------------------+
## | San_Luis_Obispo CA |
## +--------------------+
## |    San_Mateo CA    |
## +--------------------+
## |    Sangamon IL     |
## +--------------------+
## |  Santa_Barbara CA  |
## +--------------------+
## |   Santa_Clara CA   |
## +--------------------+
## |   Santa_Cruz CA    |
## +--------------------+
## |    Sarasota FL     |
## +--------------------+
## |    Saratoga NY     |
## +--------------------+
## |      Sarpy NE      |
## +--------------------+
## |   Schenectady NY   |
## +--------------------+
## |   Schuylkill PA    |
## +--------------------+
## |    Sedgwick KS     |
```

```
## +--------------------+
## |    Seminole FL     |
## +--------------------+
## |     Shasta CA      |
## +--------------------+
## |     Shawnee KS     |
## +--------------------+
## |    Sheboygan WI    |
## +--------------------+
## |     Shelby TN      |
## +--------------------+
## |      Smith TX      |
## +--------------------+
## |    Snohomish WA    |
## +--------------------+
```

```
cdinumeric <- cdi[,-c(1:3,17)]
a <- apply(cdinumeric,2,function(x) c(summary(x),SD=sd(x))) %>%
  as.data.frame %>%
  t() %>%
  round(digits=2)

  pandoc.table(a, caption="Summary Statistics for Continuous Variables in cdi
.dat dataset ", style="grid")
```

```
##
##
## +--------------------+-------+---------+--------+--------+---------+
## |               | Min.  | 1st Qu. | Median |  Mean  | 3rd Qu. |
## +====================+=======+=========+========+========+=========+
## |   **land.area**    |  15   |  451.2  | 656.5  |  1041  |  946.8  |
## +--------------------+-------+---------+--------+--------+---------+
## |      **pop**       | 1e+05 | 139027  | 217280 | 393011 | 436064  |
## +--------------------+-------+---------+--------+--------+---------+
## |   **pop.18_34**    | 16.4  |  26.2   |  28.1  | 28.57  |  30.02  |
## +--------------------+-------+---------+--------+--------+---------+
## |  **pop.65_plus**   |   3   |  9.88   | 11.75  | 12.17  |  13.62  |
## +--------------------+-------+---------+--------+--------+---------+
## |    **doctors**     |  39   |  182.8  |  401   |  988   |  1036   |
## +--------------------+-------+---------+--------+--------+---------+
## |   **hosp.beds**    |  92   |  390.8  |  755   |  1459  |  1576   |
## +--------------------+-------+---------+--------+--------+---------+
## |     **crimes**     |  563  |  6220   | 11820  | 27112  |  26280  |
## +--------------------+-------+---------+--------+--------+---------+
## |  **pct.hs.grad**   | 46.6  |  73.88  |  77.7  | 77.56  |  82.4   |
## +--------------------+-------+---------+--------+--------+---------+
## |  **pct.bach.deg**  |  8.1  |  15.28  |  19.7  | 21.08  |  25.33  |
## +--------------------+-------+---------+--------+--------+---------+
## | **pct.below.pov**  |  1.4  |   5.3   |  7.9   |  8.72  |  10.9   |
## +--------------------+-------+---------+--------+--------+---------+
```

```
## |   **pct.unemp**    |  2.2  |   5.1   |  6.2   |  6.6   |   7.5   |
## +--------------------+-------+---------+--------+--------+---------+
## | **per.cap.income** | 8899  |  16118  | 17759  | 18561  |  20270  |
## +--------------------+-------+---------+--------+--------+---------+
## |   **tot.income**   | 1141  |  2311   |  3857  |  7869  |  8654   |
## +--------------------+-------+---------+--------+--------+---------+
##
## Table: Summary Statistics for Continuous Variables in cdi.dat dataset  (co
ntinued below)
##
##
##
## +--------------------+---------+--------+
## |               |  Max.   |   SD   |
## +====================+=========+========+
## |   **land.area**    |  20062  |  1550  |
## +--------------------+---------+--------+
## |      **pop**       | 8863164 | 601987 |
## +--------------------+---------+--------+
## |   **pop.18_34**    |  49.7   |  4.19  |
## +--------------------+---------+--------+
## |  **pop.65_plus**   |  33.8   |  3.99  |
## +--------------------+---------+--------+
## |    **doctors**     |  23677  |  1790  |
## +--------------------+---------+--------+
## |   **hosp.beds**    |  27700  |  2289  |
## +--------------------+---------+--------+
## |     **crimes**     | 688936  | 58238  |
## +--------------------+---------+--------+
## |  **pct.hs.grad**   |  92.9   |  7.02  |
## +--------------------+---------+--------+
## |  **pct.bach.deg**  |  52.3   |  7.65  |
## +--------------------+---------+--------+
## | **pct.below.pov**  |  36.3   |  4.66  |
## +--------------------+---------+--------+
## |   **pct.unemp**    |  21.3   |  2.34  |
## +--------------------+---------+--------+
## | **per.cap.income** |  37541  |  4059  |
## +--------------------+---------+--------+
## |   **tot.income**   | 184230  | 12884  |
## +--------------------+---------+--------+

tmp <- rbind(with(cdi,table(region)))
row.names(tmp) <- "Freq"
pandoc.table(tmp,caption="Summary Statistics of Region in CDI", style="grid")

##
##
## +----------+-----+-----+-----+----+
## |     | NC  | NE  |  S  | W  |
```

```
## +==========+=====+=====+=====+====+
## | **Freq** | 108 | 103 | 152 | 77 |
## +----------+-----+-----+-----+----+
##
## Table: Summary Statistics of Region in CDI
```

1. Indicate where (in which variables) there is missing data (NA's), if any, how much there is (in eachvariable) and why it might be there.

```
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
```

```
cdi$county <- as.factor(cdi$county)
cdi$state <- as.factor(cdi$state)
cdi$region <- as.factor(cdi$region)
```

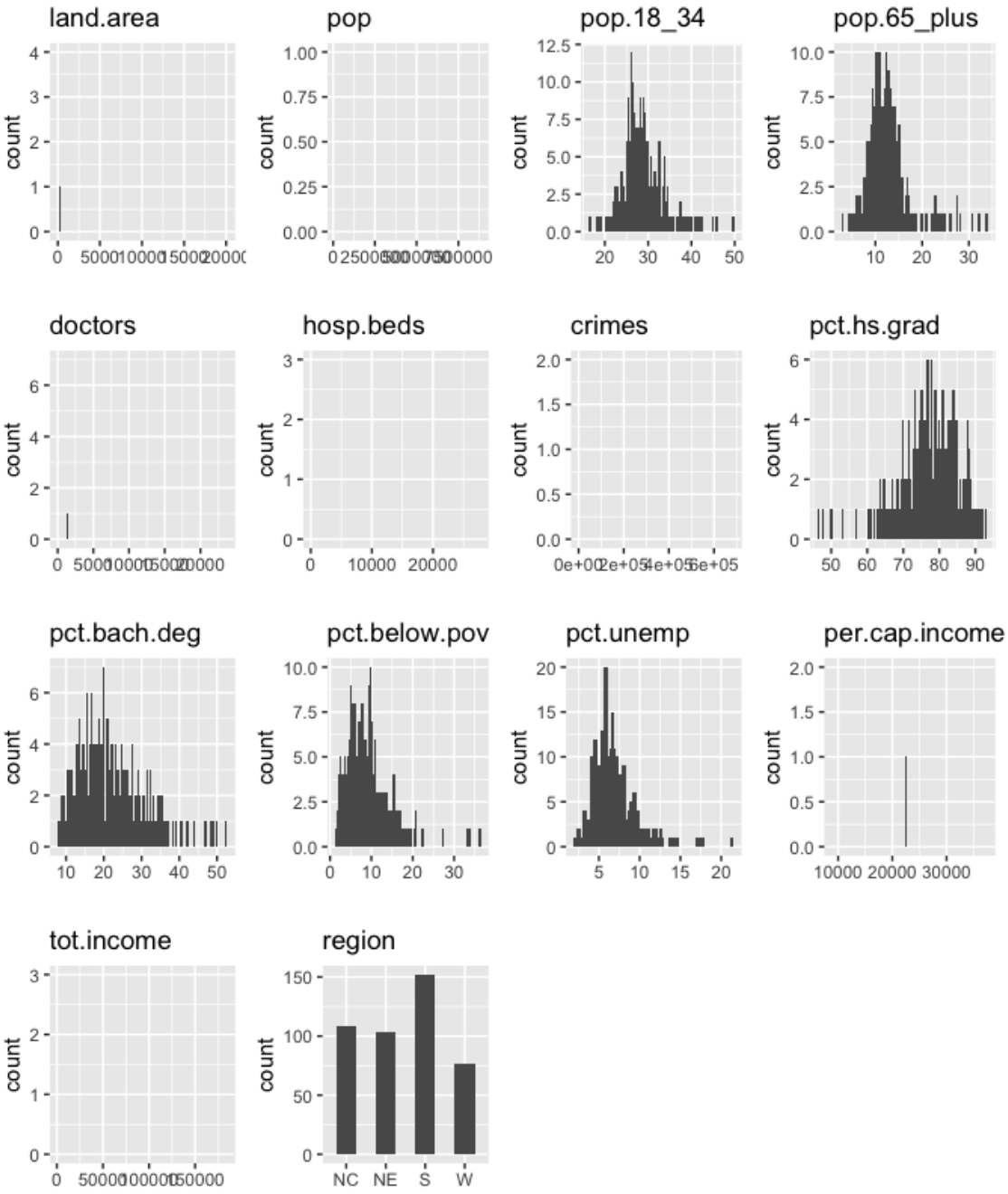
2. Make some appropriate descriptive EDA plots to illustrate any important features of the variables or possible important relationships among them.

```
cdigood <- data.frame(cdinumeric,region=cdi$region)
```

i. Univariate Distributions

```
cdigood$land.area <- as.numeric(cdigood$land.area)
cdigood$pop <- as.numeric(cdigood$pop)
cdigood$doctors <- as.numeric(cdigood$doctors)
cdigood$hosp.beds <- as.numeric(cdigood$hosp.beds)
cdigood$crimes <- as.numeric(cdigood$crimes)
cdigood$per.cap.income <- as.numeric(cdigood$per.cap.income)
cdigood$tot.income <- as.numeric(cdigood$tot.income)

hist.builder <- function(df) {  ## creates a list of graphs
    result <- NULL
    for (var in names(df)) {
      d <- data.frame(dd=df[,var])
      if(mode(df[,var])=="numeric") {
        print(var)
        p <- ggplot(d,aes(x=dd)) + stat_count(width = 0.5) +
              ggtitle(var) + xlab("")
      } else {
```

```
          p <- ggplot(d,aes(x=dd)) + stat_count(width = 0.5) +
                 ggtitle(var) + xlab("")
      }
    result <- c(result,list(p))
    }
    return(result)
}

grid.arrange(grobs=hist.builder(cdigood))
```



Note: There are some errors in putting these graphs in a word doc file. I showed each graph indivually and correctly in Appendix C (See pages 46 to 59).

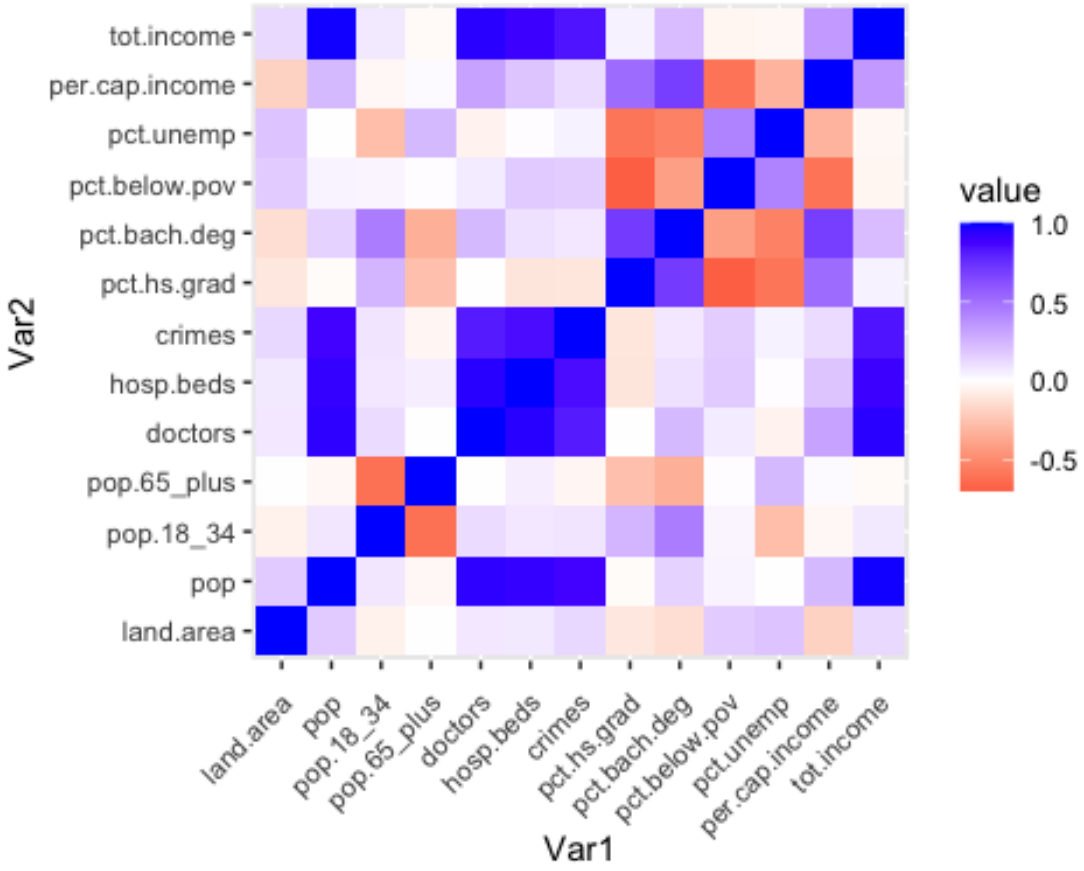
Distributions of Variables

```
## [1] "land.area"
## [1] "pop"
## [1] "pop.18_34"
## [1] "pop.65_plus"
## [1] "doctors"
## [1] "hosp.beds"
```

```
## [1] "crimes"
## [1] "pct.hs.grad"
## [1] "pct.bach.deg"
## [1] "pct.below.pov"
## [1] "pct.unemp"
## [1] "per.cap.income"
## [1] "tot.income"
## [1] "region"
```

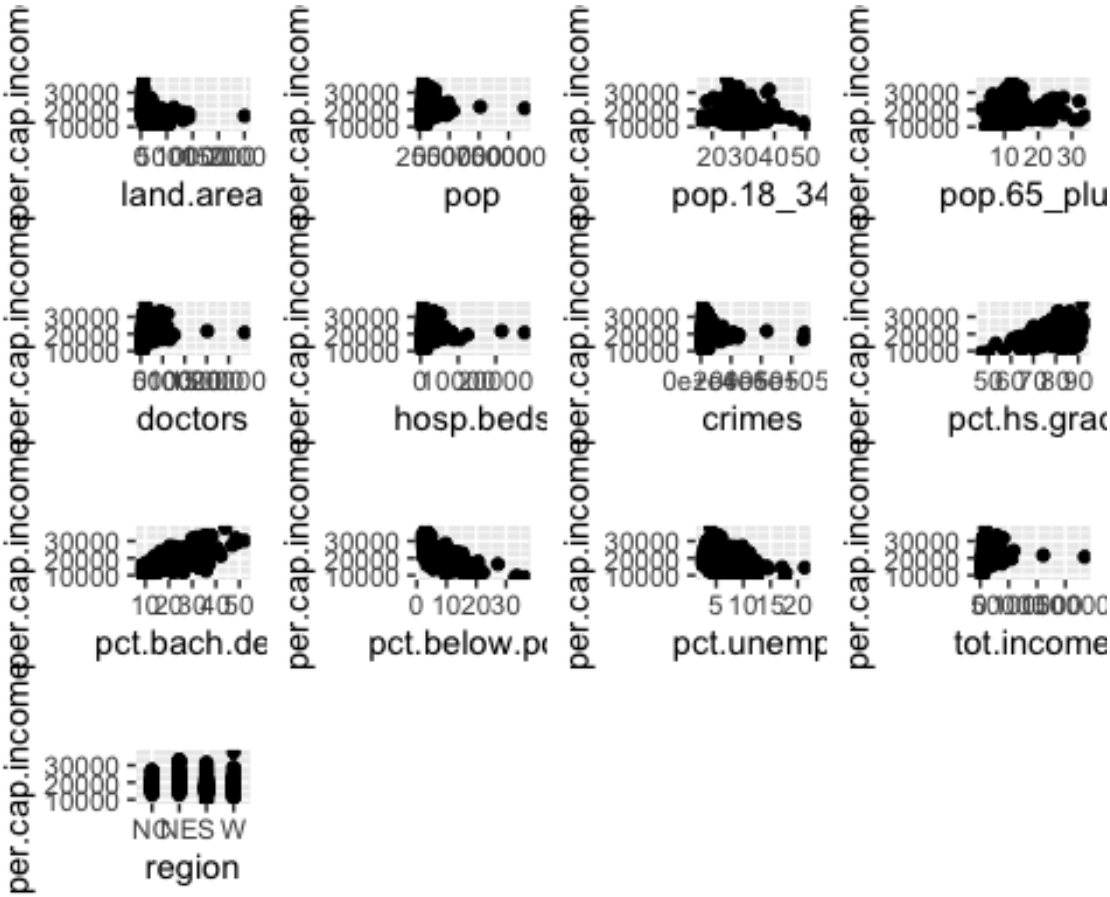
Bivariate Relationship

```
corgraph <-function(df) {
  cormat <- cor(df)
  melted_cormat <- melt(cormat) ## need library(reshape2) for this...
  ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
    geom_tile() +
    theme(axis.text.x = element_text(angle = 45,vjust=0.9,hjust=1)) +
    scale_fill_gradient2(low="red",mid="white",high="blue")
}
corgraph(cdinumeric)
```




```
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)
      } else {
        p <- ggplot(d,aes(x=xx,y=yy)) + geom_boxplot(notch=F) +
          ggtitle("") + xlab(xvar) + ylab(yvar)
      }
    result <- c(result,list(p))
    }
    return(result)
}
grid.arrange(grobs=scatter.builder(cdigood))
```

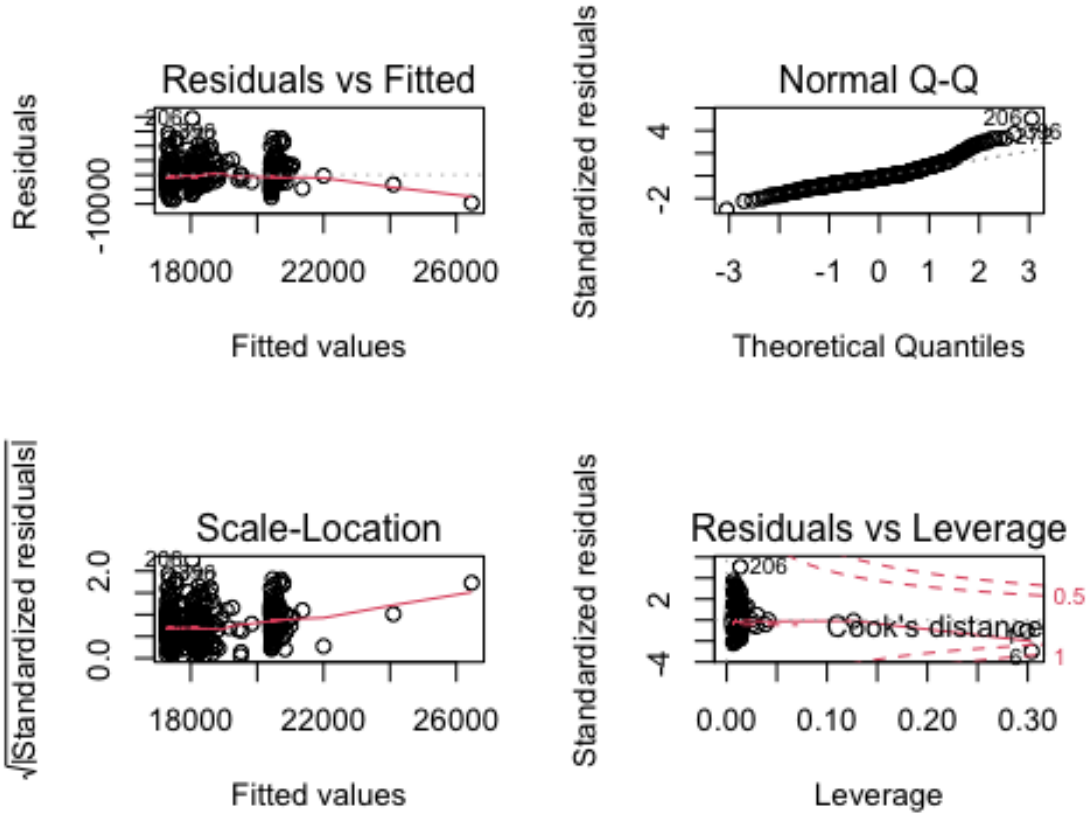


Distributions of Variables against Income per capita

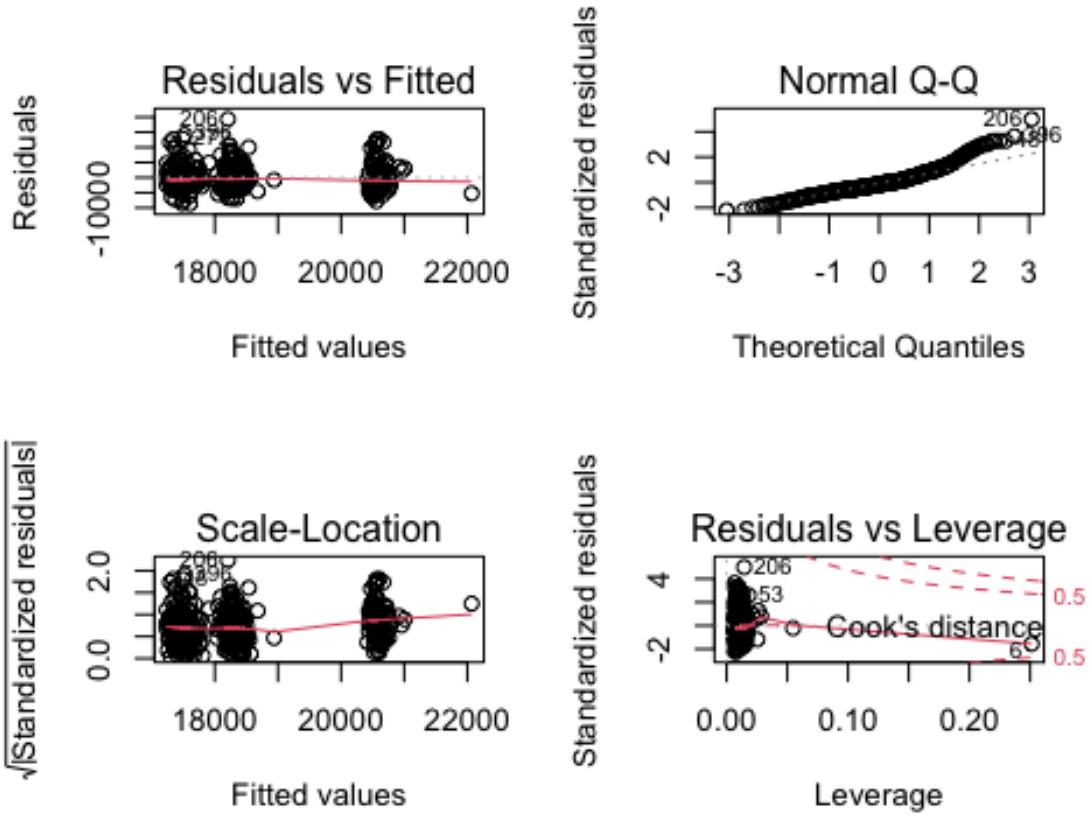
Appendix B

```
lm1 <- lm(per.cap.income ~ crimes + region)

summary(lm1)

##
## Call:
```

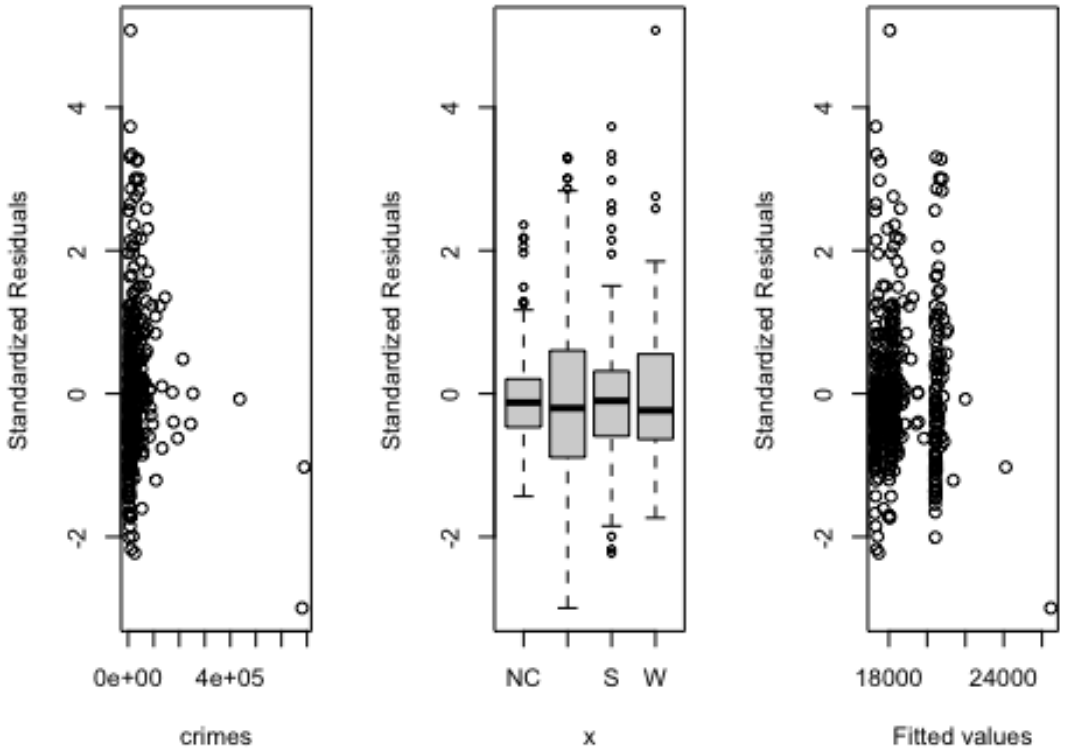
```
## lm(formula = per.cap.income ~ crimes + region)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9661.0 -2260.7  -618.3  1650.0 19492.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.811e+04  3.784e+02  47.846  < 2e-16 ***
## crimes  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

par(mfrow=c(2,2))
plot(lm1)
```



```
cdi$per.cap.crime <- crimes/pop

names(cdi)

##  [1] "id"             "county"         "state"          "land.area"
##  [5] "pop"            "pop.18_34"      "pop.65_plus"    "doctors"
##  [9] "hosp.beds"      "crimes"         "pct.hs.grad"    "pct.bach.deg"
## [13] "pct.below.pov"  "pct.unemp"      "per.cap.income" "tot.income"
## [17] "region"         "per.cap.crime"

detach(cdi)
attach(cdi)

lm2 <- lm(per.cap.income ~ per.cap.crime + region, data=cdi)

summary(lm2)

##
## Call:
## lm(formula = per.cap.income ~ per.cap.crime + region, data = cdi)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8634  -2300   -631   1710  19332
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18006.04     537.04  33.528  < 2e-16 ***
## per.cap.crime  5773.20    7520.41   0.768   0.4431
## regionNE       2354.70     541.97   4.345 1.74e-05 ***
## regionS        -927.45     512.31  -1.810   0.0709 .
## regionW         -34.92     586.03  -0.060   0.9525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3898 on 435 degrees of freedom
## Multiple R-squared:  0.08622,    Adjusted R-squared:  0.07782
## F-statistic: 10.26 on 4 and 435 DF,  p-value: 6.007e-08

par(mfrow=c(2,2))
plot(lm2)
```

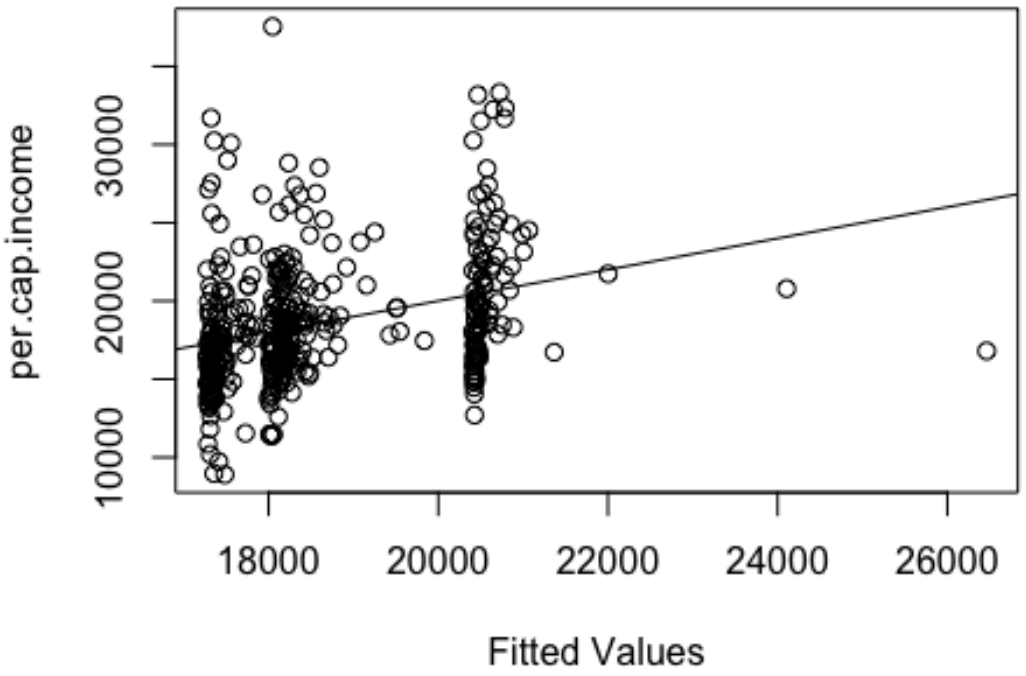



```
StanRes1 <- rstandard(lm1)
par(mfrow=c(1,3))
plot(crimes, StanRes1, ylab="Standardized Residuals")
plot(region, StanRes1, ylab="Standardized Residuals")
plot(lm1$fitted.values,StanRes1, ylab="Standardized Residuals",xlab="Fitted v
alues")
```



*Plots of the standardized residuals from model lm*1

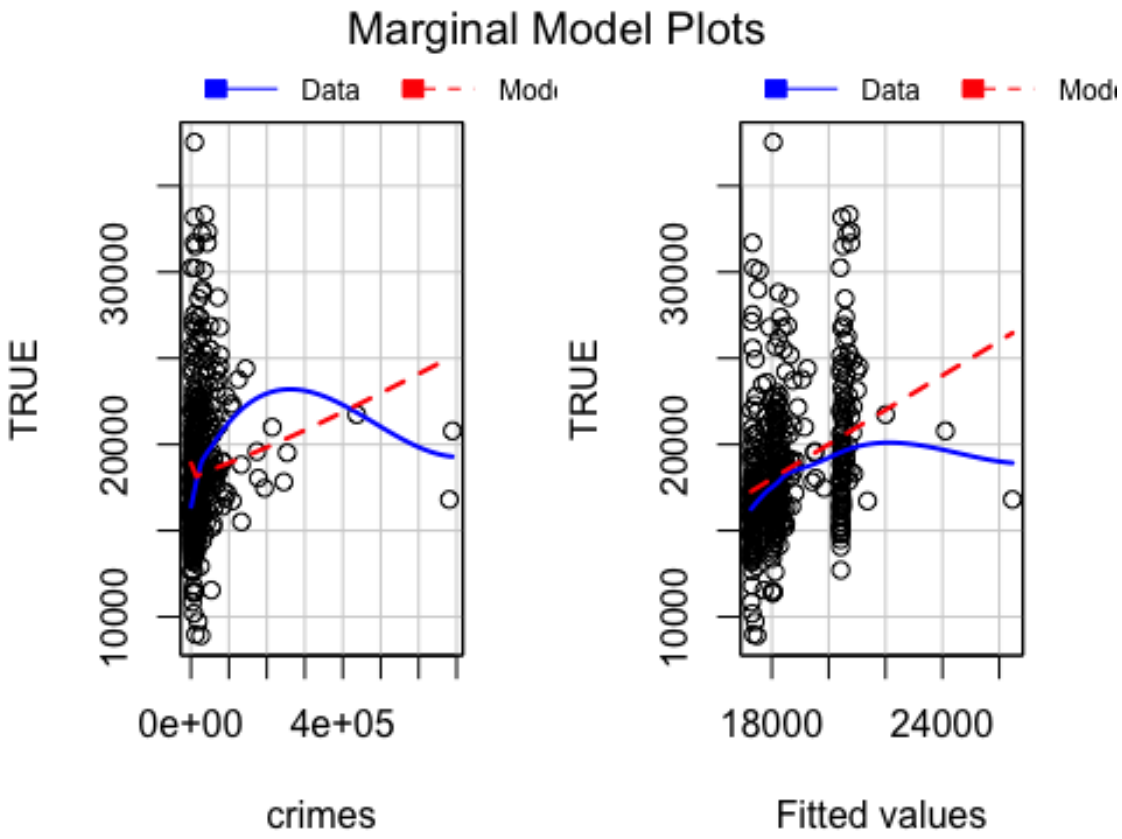
```
par(mfrow=c(1,1))
plot(lm1$fitted.values,per.cap.income,xlab="Fitted Values")
abline(lsfit(lm1$fitted.values,per.cap.income))
```



A plot of per.cap.income against fitted values

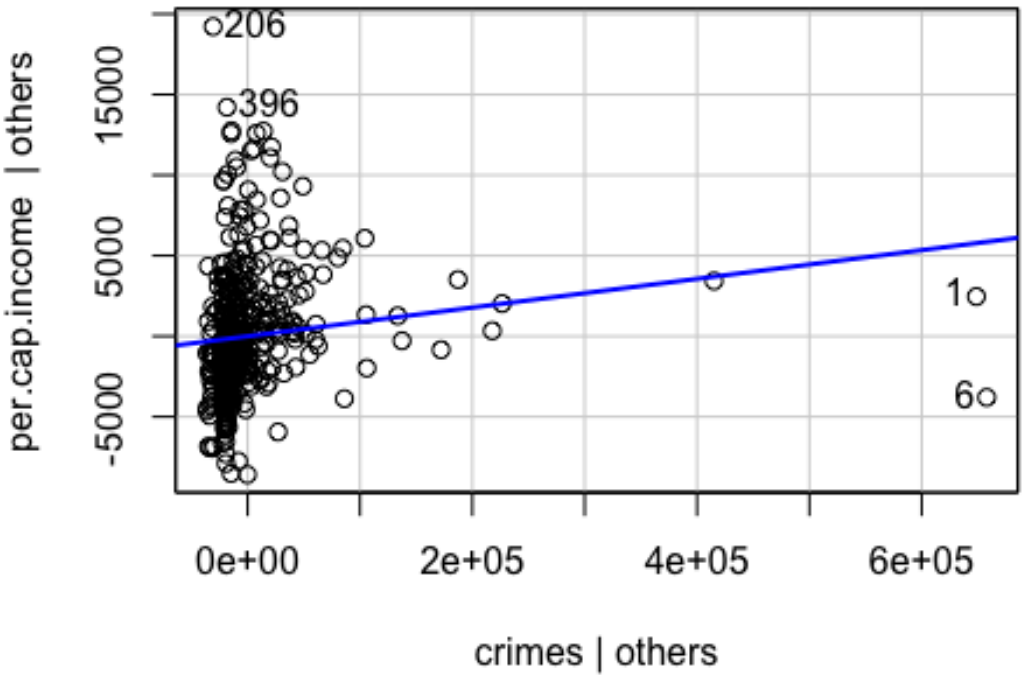
```
mmps(lm1,layout=c(1,2))
```

```
## Warning in mmps(lm1, layout = c(1, 2)): Interactions and/or factors skippe
d
```



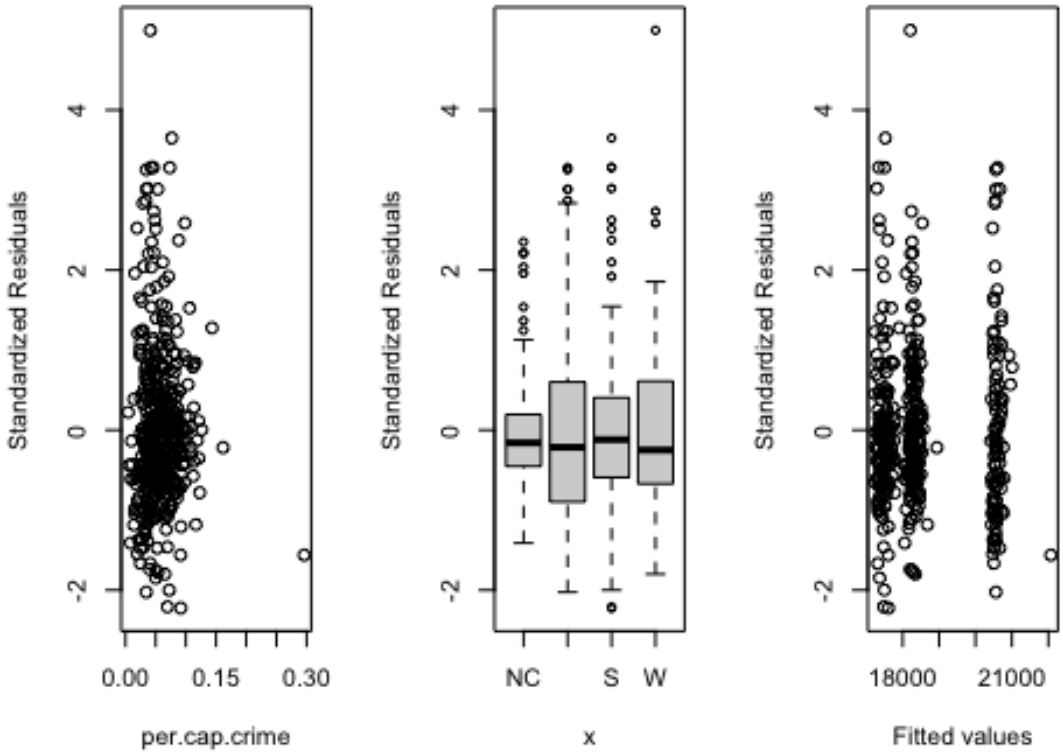
Marginal model plots for model lm1

```
library(car)
par(mfrow=c(1,1))
avPlot(lm1,variable=crimes,ask=FALSE, main="")
```



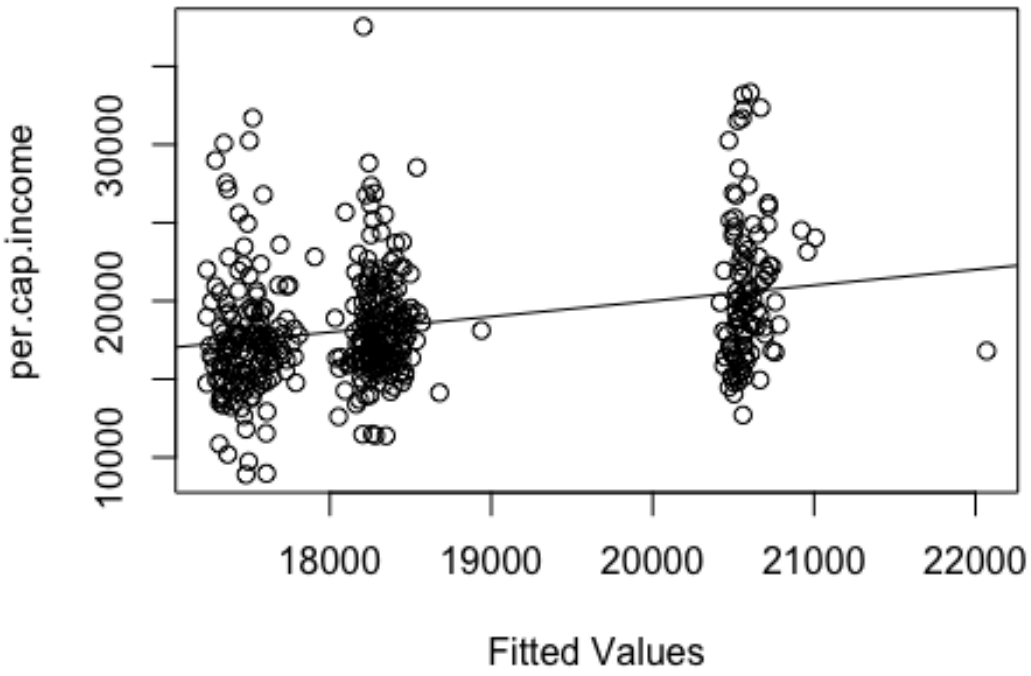
Added-variable plots for model lm1

```
StanRes2 <- rstandard(lm2)
par(mfrow=c(1,3))
plot(per.cap.crime, StanRes2, ylab="Standardized Residuals")
plot(region, StanRes2, ylab="Standardized Residuals")
plot(lm2$fitted.values,StanRes2, ylab="Standardized Residuals",xlab="Fitted v
alues")
```



Plots of the standardized residuals from model lm2

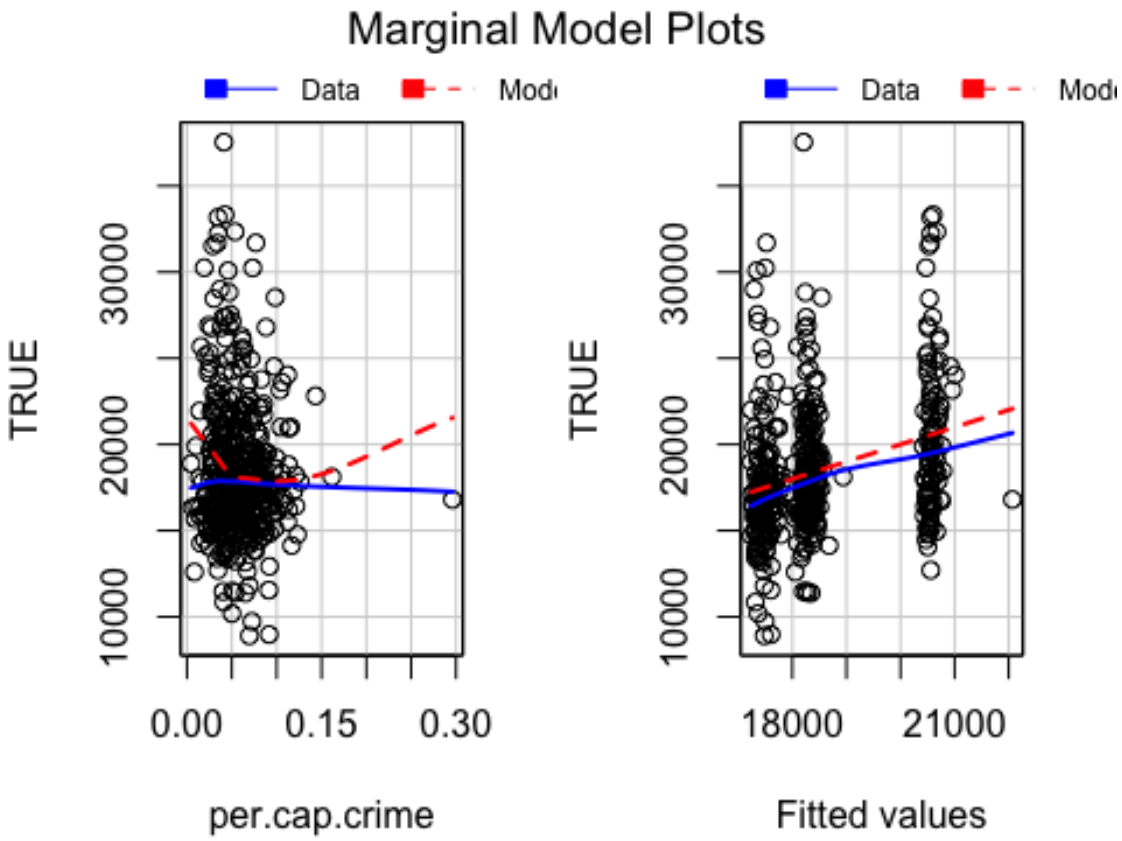
```
par(mfrow=c(1,1))
plot(lm2$fitted.values,per.cap.income,xlab="Fitted Values")
abline(lsfit(lm2$fitted.values,per.cap.income))
```



A plot of per.cap.income against fitted values

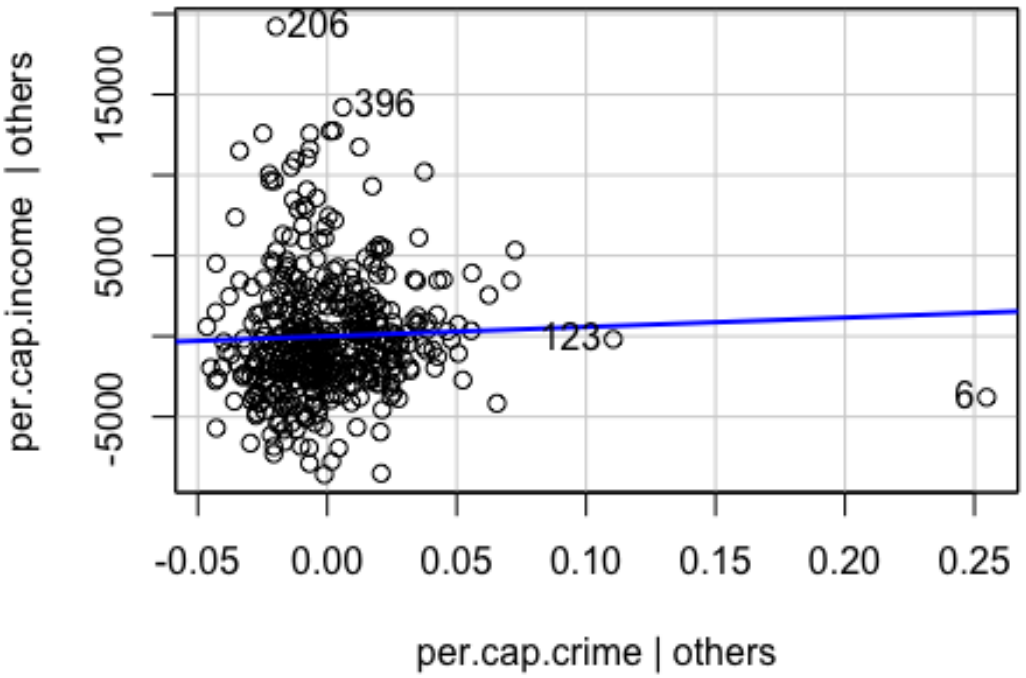
```
mmps(lm2,layout=c(1,2))
```

```
## Warning in mmps(lm2, layout = c(1, 2)): Interactions and/or factors skippe
d
```



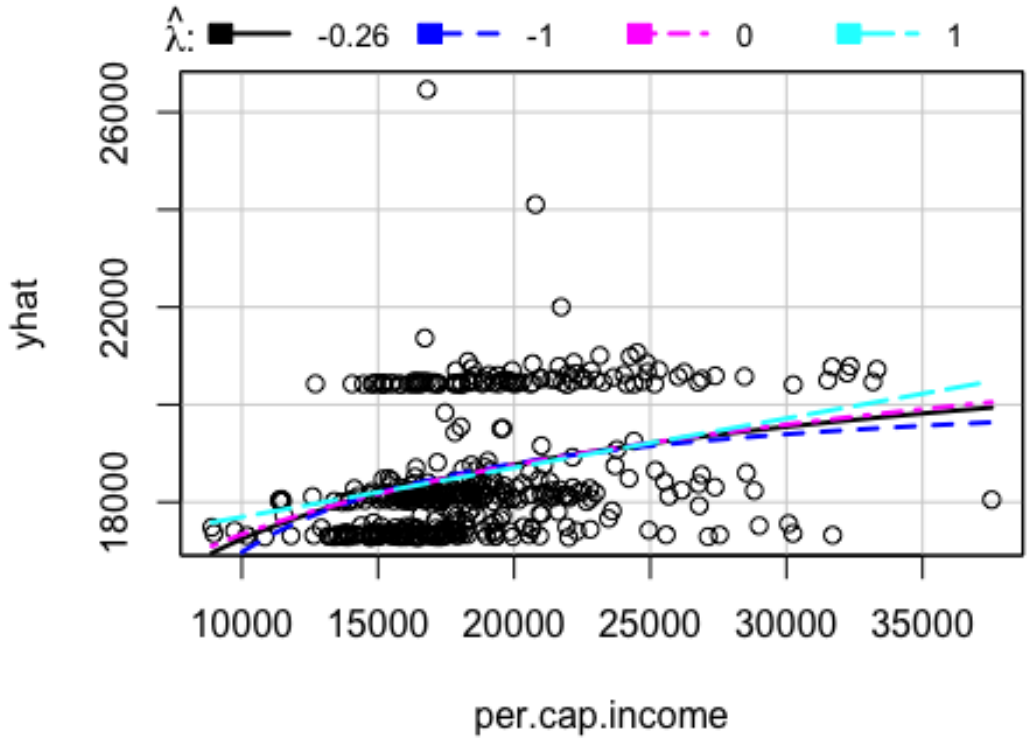
Marginal model plots for model lm2

```
library(car)
par(mfrow=c(1,1))
avPlot(lm2,variable=per.cap.crime,ask=FALSE, main="")
```

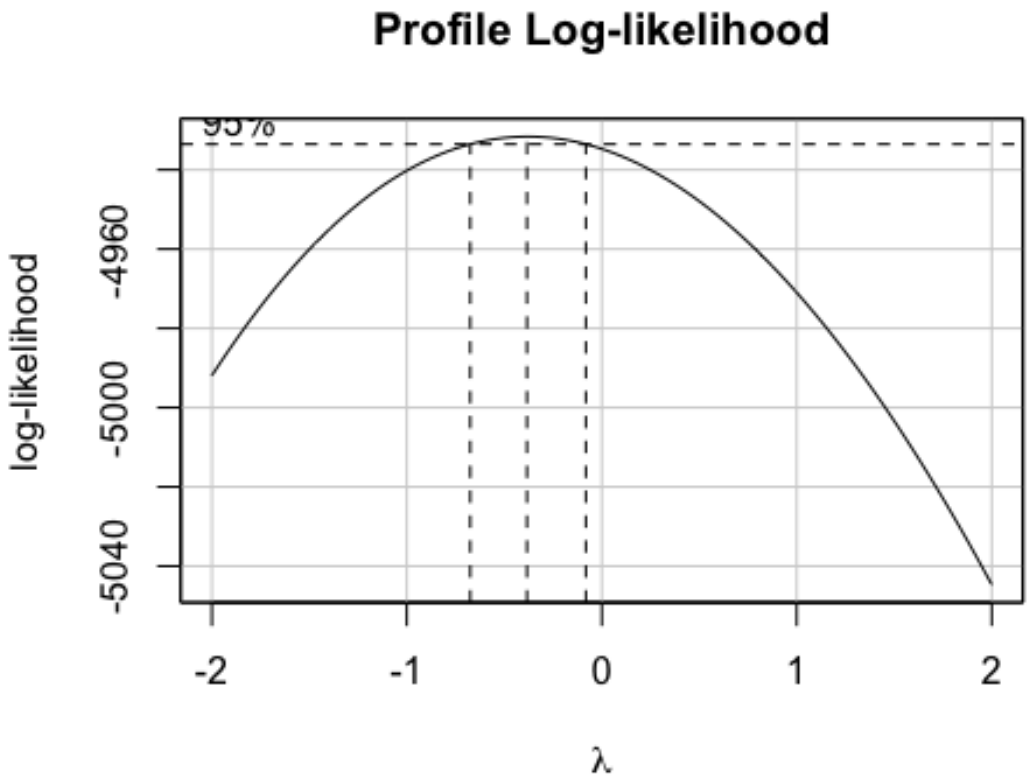


Added-variable plots for model lm2

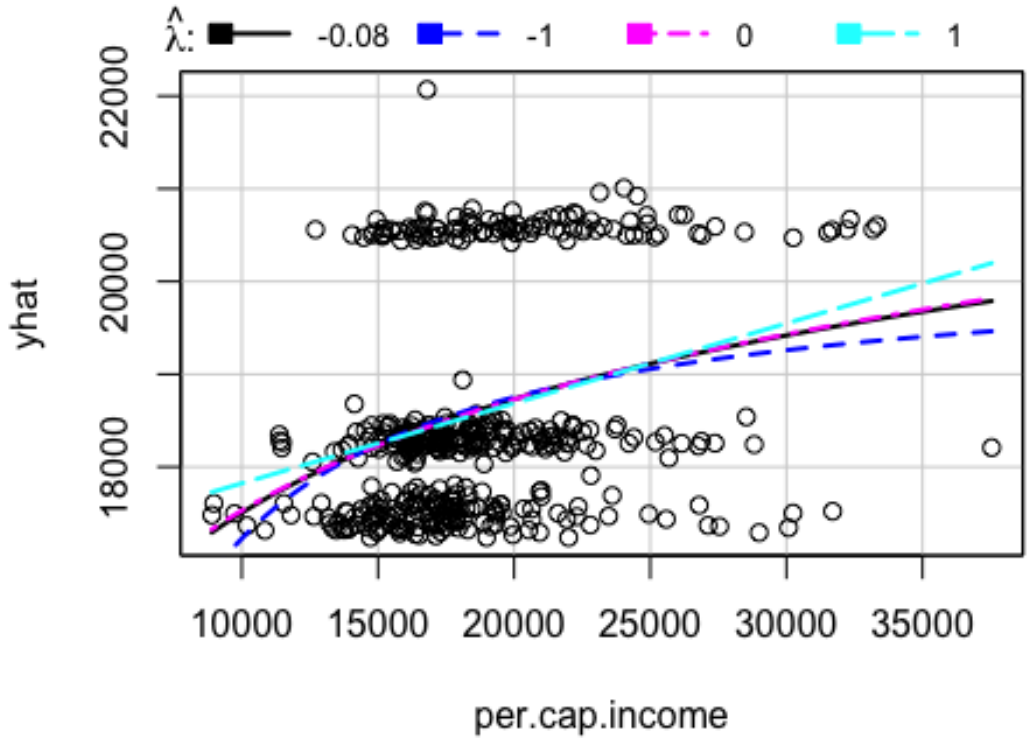
```
inverseResponsePlot(lm1)
```



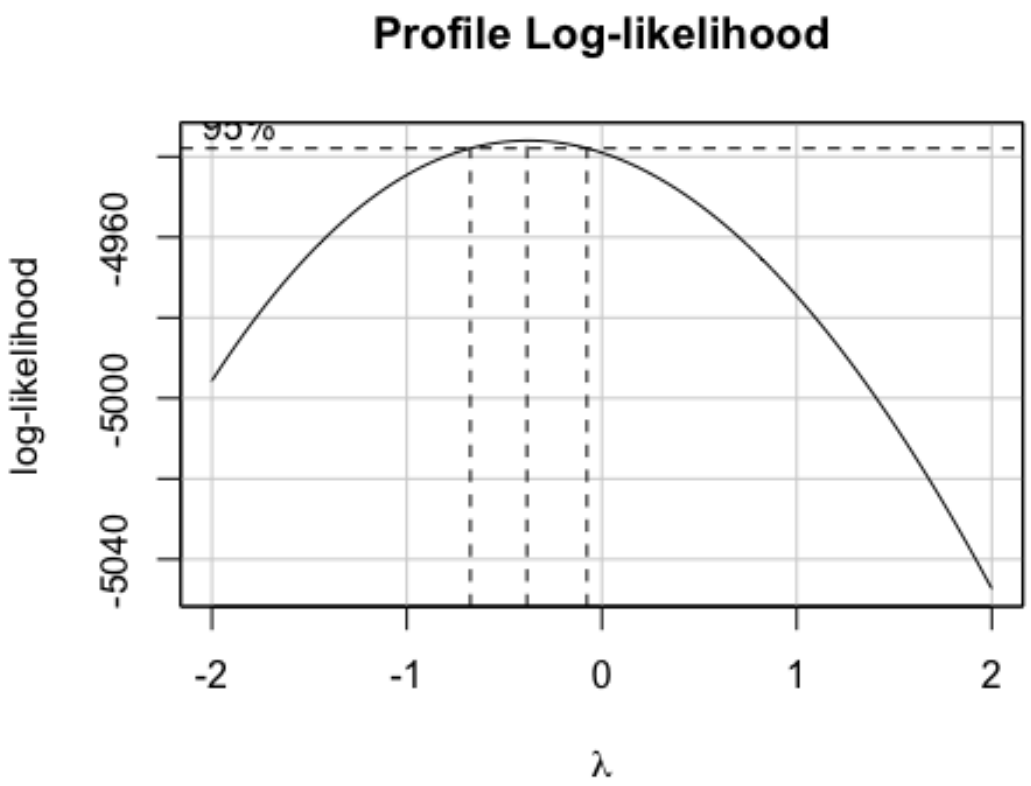

```
##       lambda       RSS
## 1 -0.2581433 652567923
## 2 -1.0000000 654566436
## 3  0.0000000 652800071
## 4  1.0000000 657627602

boxCox(lm1)
```




```
inverseResponsePlot(lm2)
```



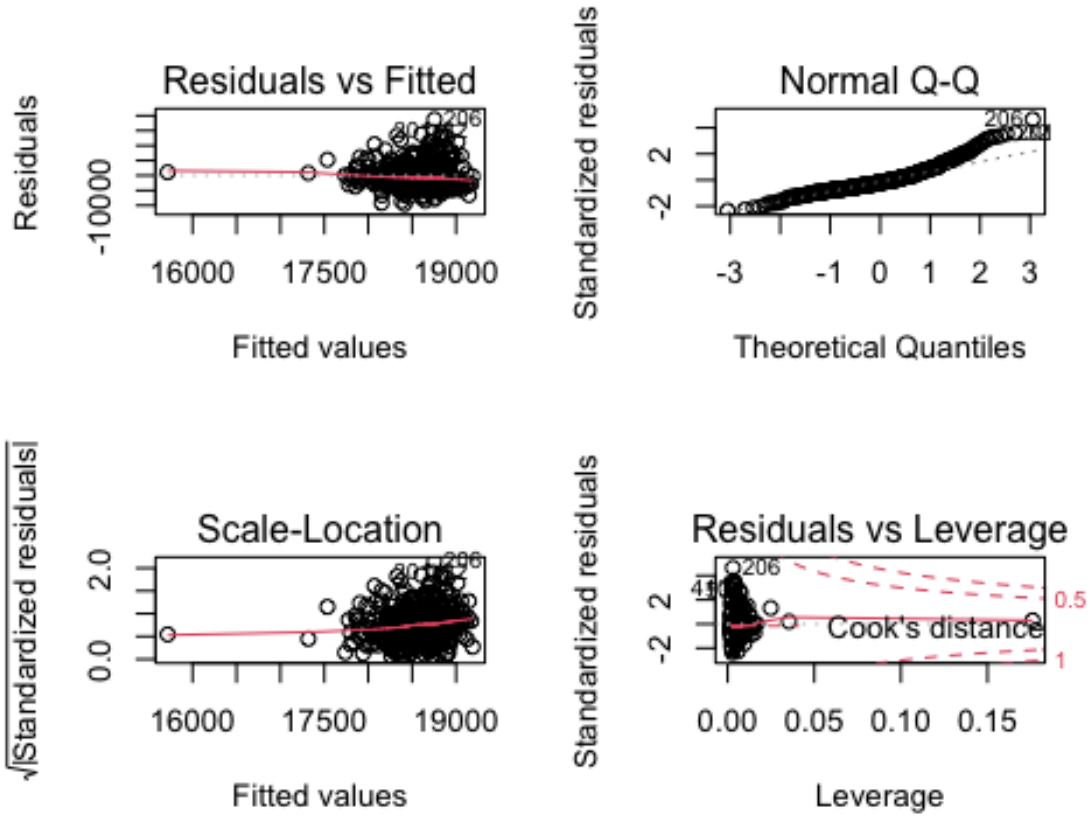

```
##       lambda       RSS
## 1 -0.0774348 567496318
## 2 -1.0000000 569469857
## 3  0.0000000 567509697
## 4  1.0000000 569893885

boxCox(lm2)
```



Interaction

```
lm3 <- lm(per.cap.income ~ per.cap.crime, data=cdi)

summary(lm3)

##
## Call:
## lm(formula = per.cap.income ~ per.cap.crime, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9515.2 -2568.9  -749.9  1574.9 18786.7
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    19244.3      448.9  42.867   <2e-16 ***
## per.cap.crime -11919.3     7074.5  -1.685   0.0927 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4051 on 438 degrees of freedom
## Multiple R-squared:  0.006439,   Adjusted R-squared:  0.004171
## F-statistic: 2.839 on 1 and 438 DF,  p-value: 0.09274
```

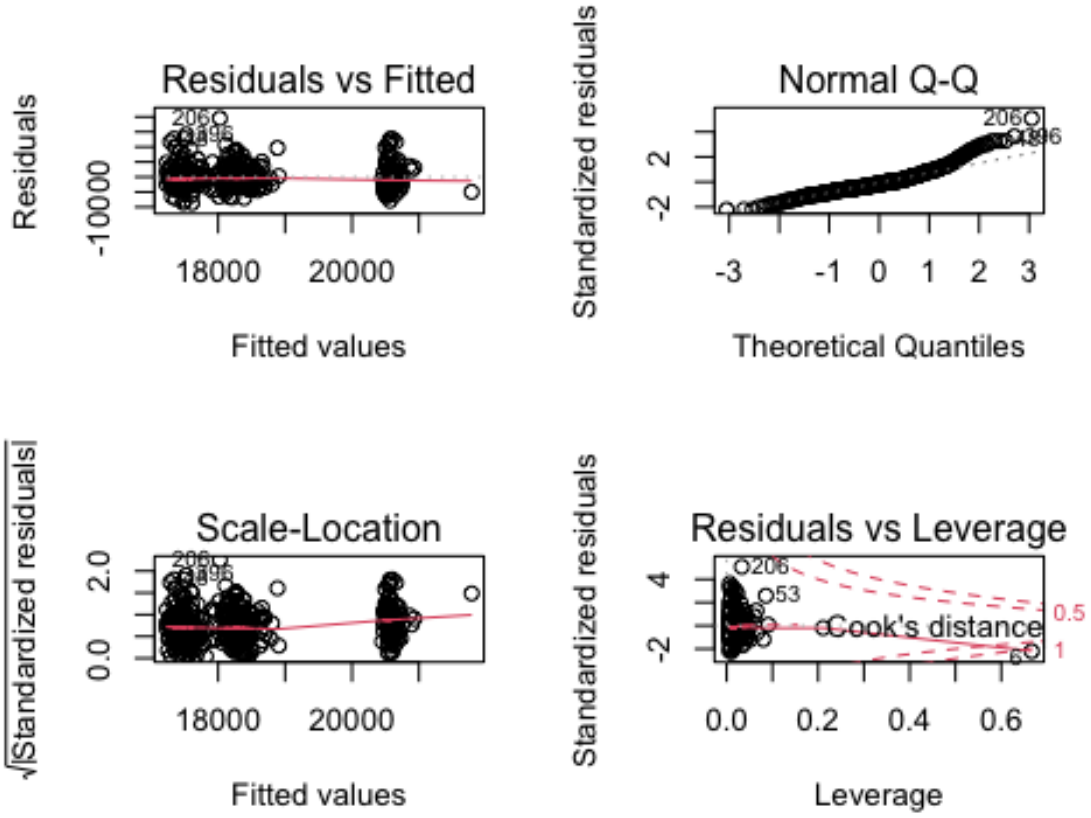
```
par(mfrow=c(2,2))
plot(lm3)
```




```
lm4 <- lm(per.cap.income ~ per.cap.crime*region, data=cdi)

summary(lm4)

##
## Call:
## lm(formula = per.cap.income ~ per.cap.crime * region, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8637.7 -2333.9  -629.5  1759.1 19515.6
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              18077.3      895.2  20.193   <2e-16 ***
## per.cap.crime             4379.1    15893.5   0.276    0.783
## regionNE                  2329.0     1101.4   2.115    0.035 *
## regionS                  -1010.4     1323.8  -0.763    0.446
## regionW                   -670.0     1983.9  -0.338    0.736
## per.cap.crime:regionNE     288.4    20184.7   0.014    0.989
```

```
## per.cap.crime:regionS    1558.9    20556.1   0.076    0.940
## per.cap.crime:regionW   10655.5    32322.4   0.330    0.742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3911 on 432 degrees of freedom
## Multiple R-squared:  0.08648,    Adjusted R-squared:  0.07168
## F-statistic: 5.842 on 7 and 432 DF,  p-value: 1.713e-06
```

```
par(mfrow=c(2,2))
plot(lm4)
```




```
anova(lm3, lm4)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ per.cap.crime
## Model 2: per.cap.income ~ per.cap.crime * region
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    438 7186843542
## 2    432 6607856753  6 578986790 6.3087 2.252e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm2, lm4)
```

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

```
data.frame(AIC=AIC(lm2, lm3, lm4),
           BIC=BIC(lm2, lm3, lm4))[,-3] %>%
  kbl(booktabs=T,col.names=c("df","AIC","BIC")) %>% kable_classic(full_width=
F)
```

df

AIC

BIC

lm2

6

8531.682

8556.203

lm3

3

8562.513

8574.773

lm4

9

8537.556

8574.337

```
formula(lm2)
```

```
## per.cap.income ~ per.cap.crime + region
```

```
round(coef(summary(lm2)),2)
```

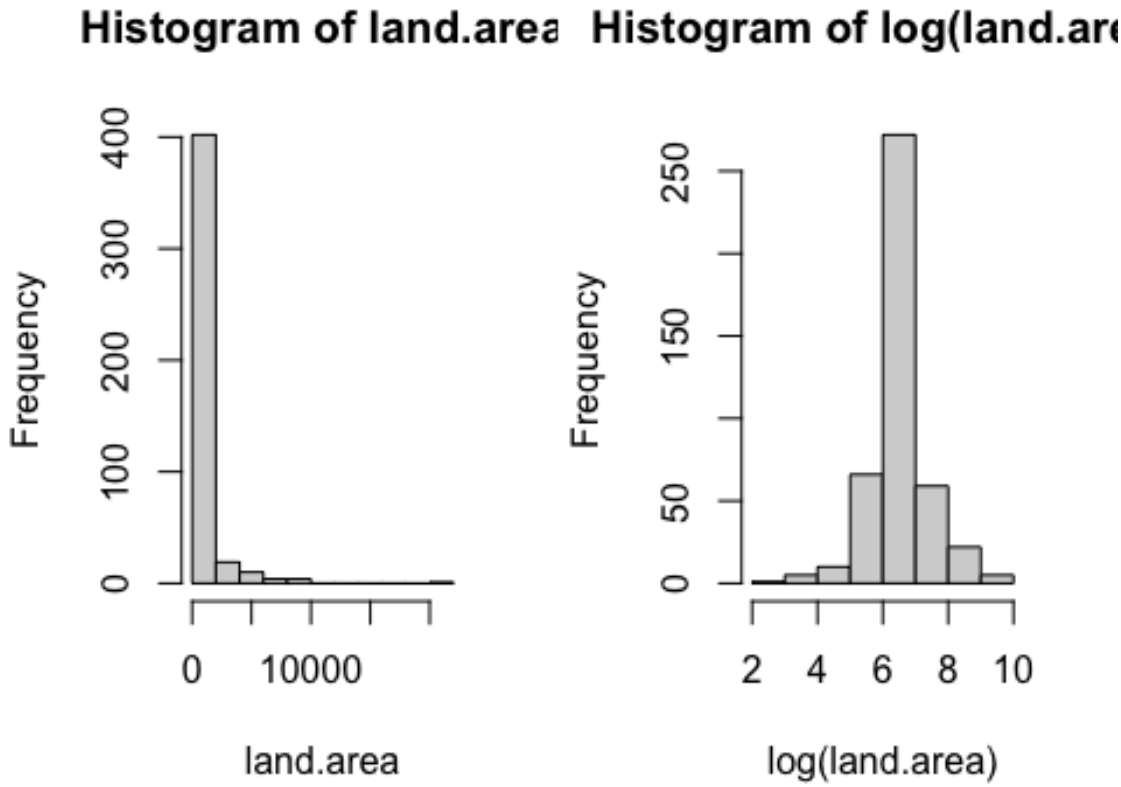
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18006.04     537.04   33.53     0.00
## per.cap.crime  5773.20    7520.41    0.77     0.44
## regionNE       2354.70     541.97    4.34     0.00
```

```
## regionS         -927.45     512.31    -1.81     0.07
## regionW          -34.92     586.03    -0.06     0.95
```

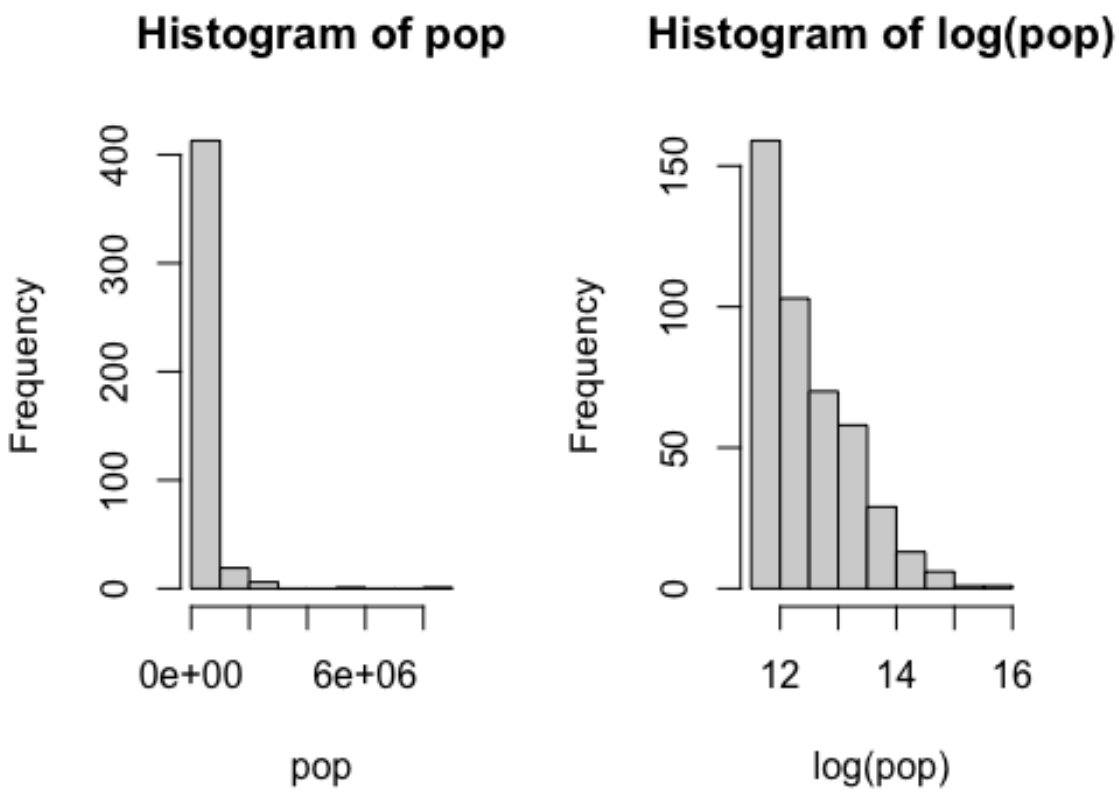
Appendix C

Transformation

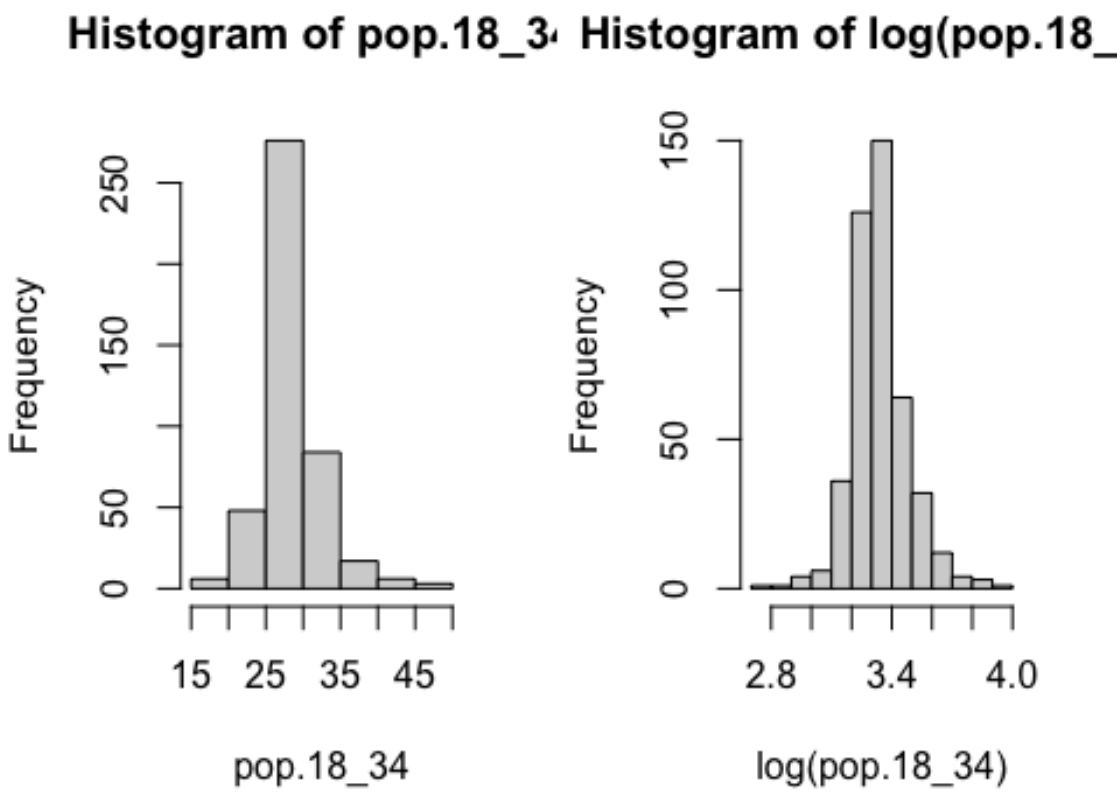
```
par(mfrow=c(1,2))
hist(land.area)
hist(log(land.area))
```



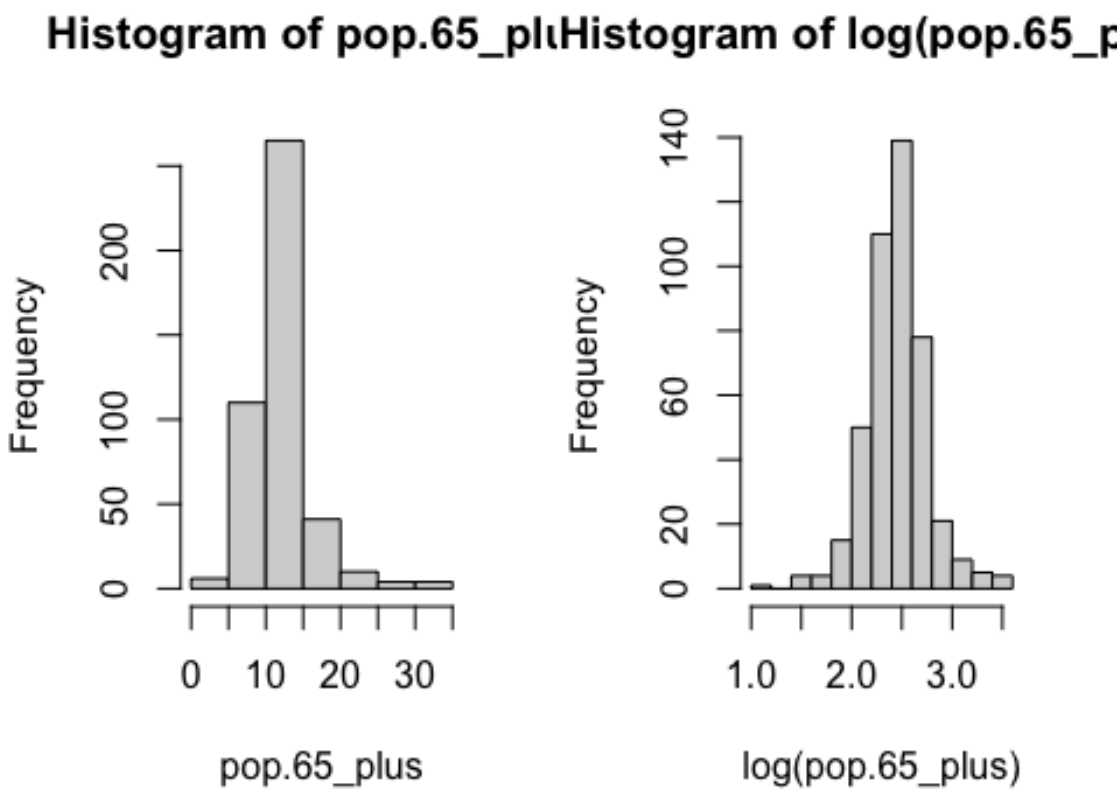

```
par(mfrow=c(1,2))
hist(pop)
hist(log(pop))
```



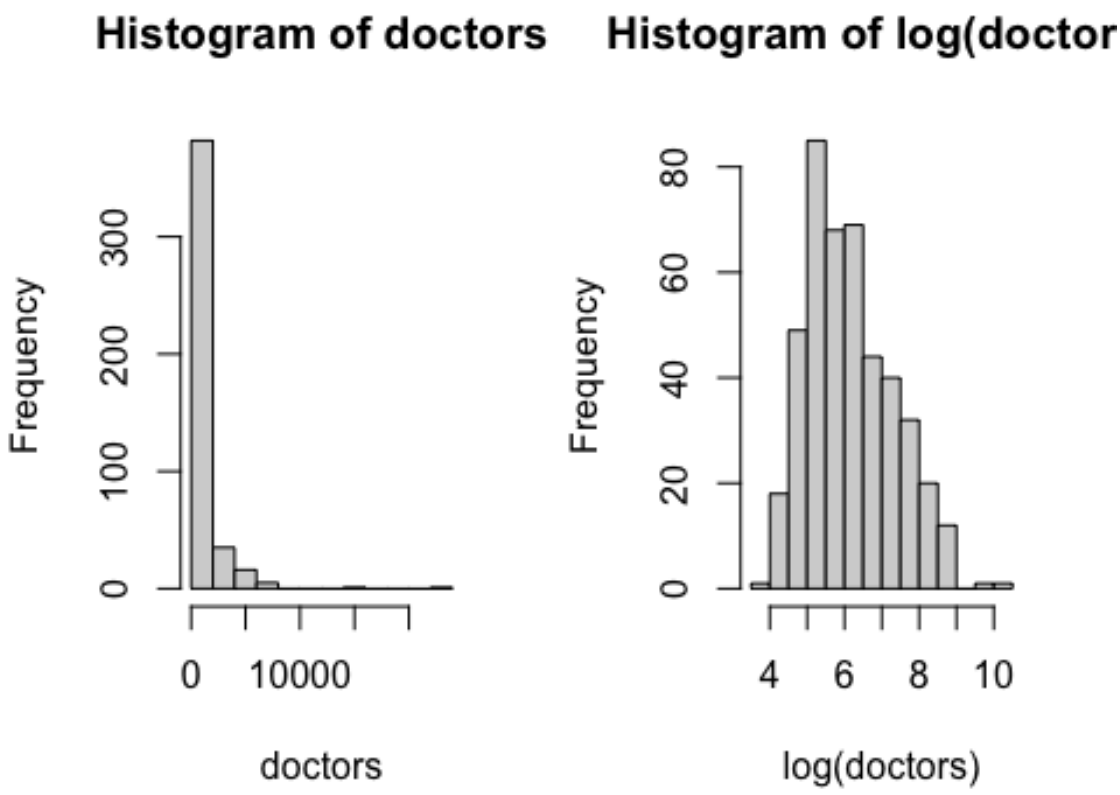

```
par(mfrow=c(1,2))
hist(pop.18_34)
hist(log(pop.18_34))
```



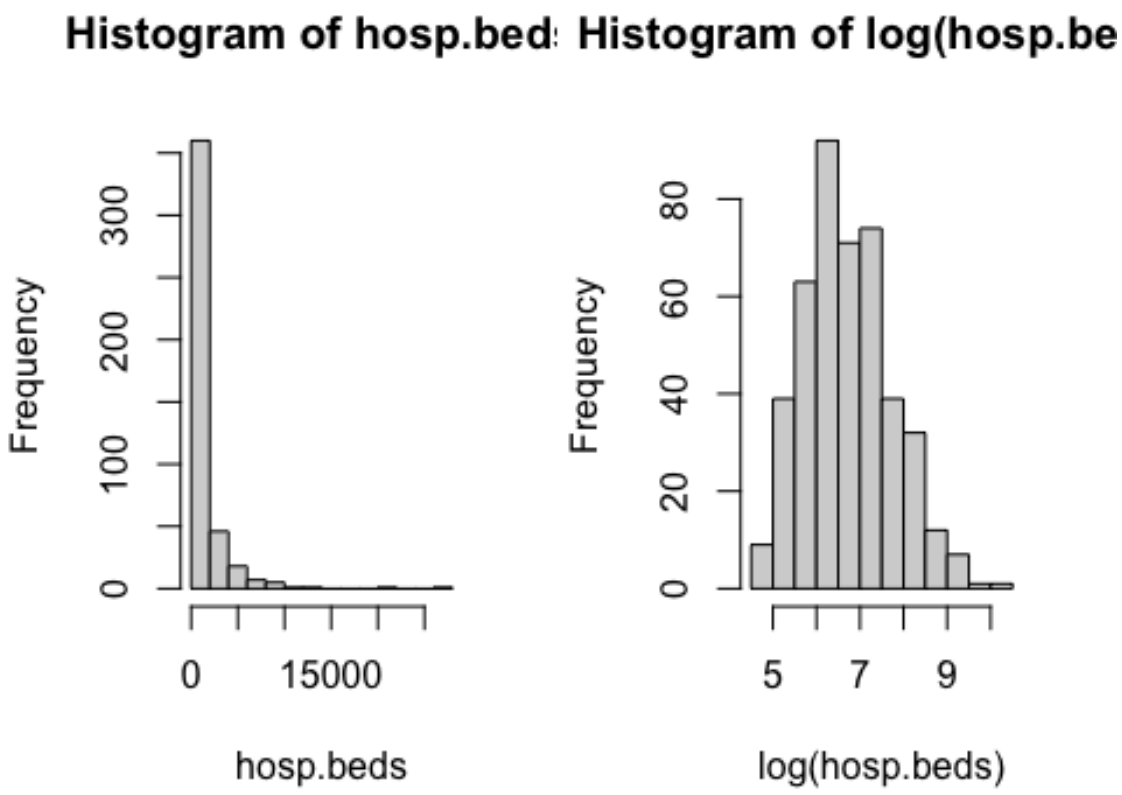

```
par(mfrow=c(1,2))
hist(pop.65_plus)
hist(log(pop.65_plus))
```



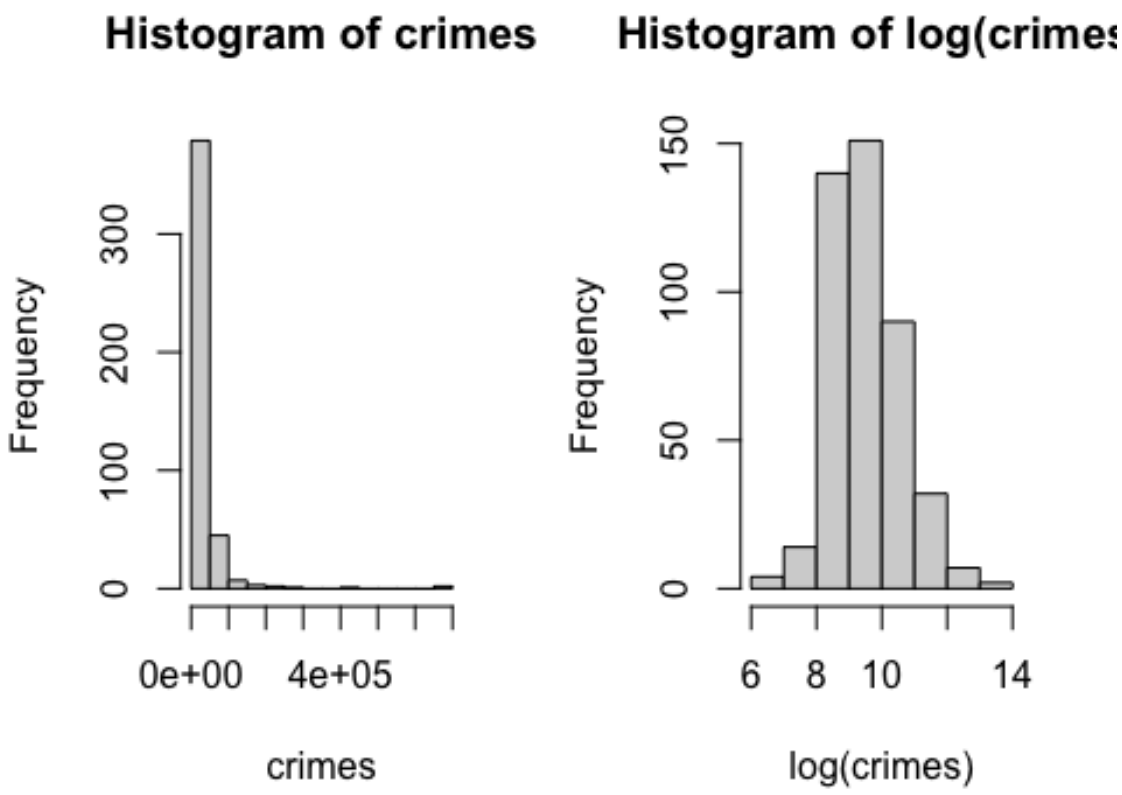

```
par(mfrow=c(1,2))
hist(doctors)
hist(log(doctors))
```



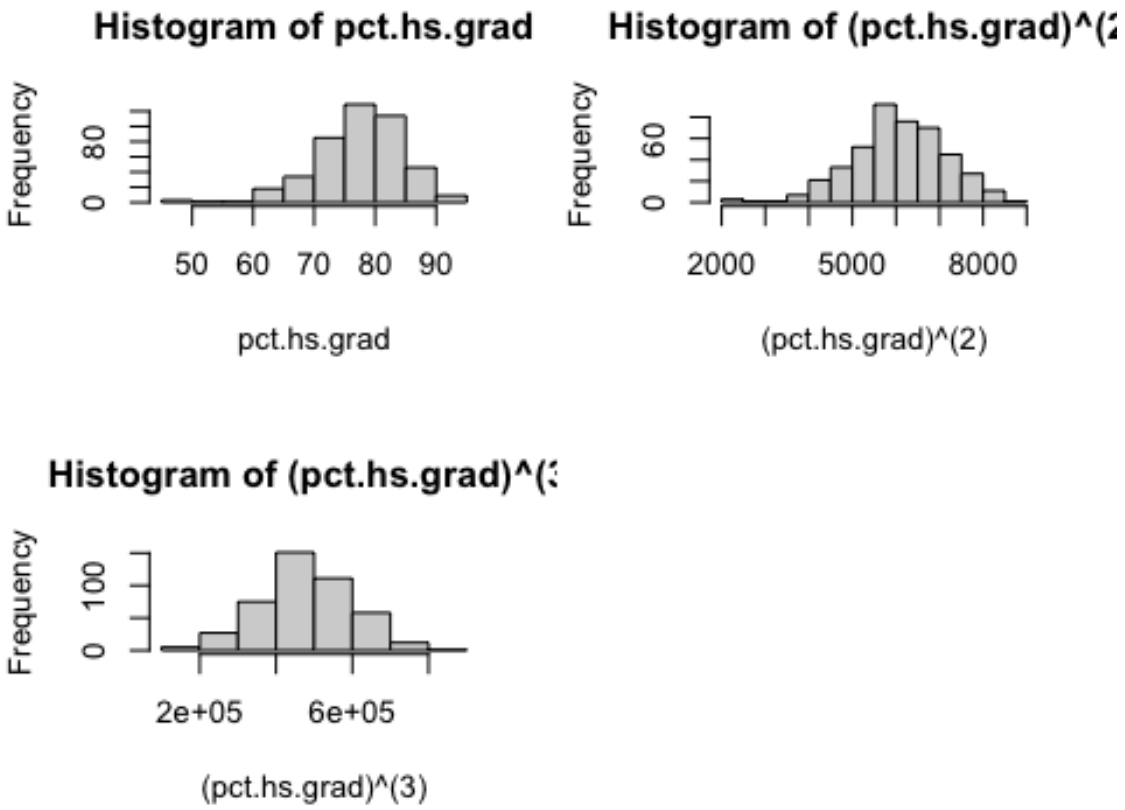

```
par(mfrow=c(1,2))
hist(hosp.beds)
hist(log(hosp.beds))
```



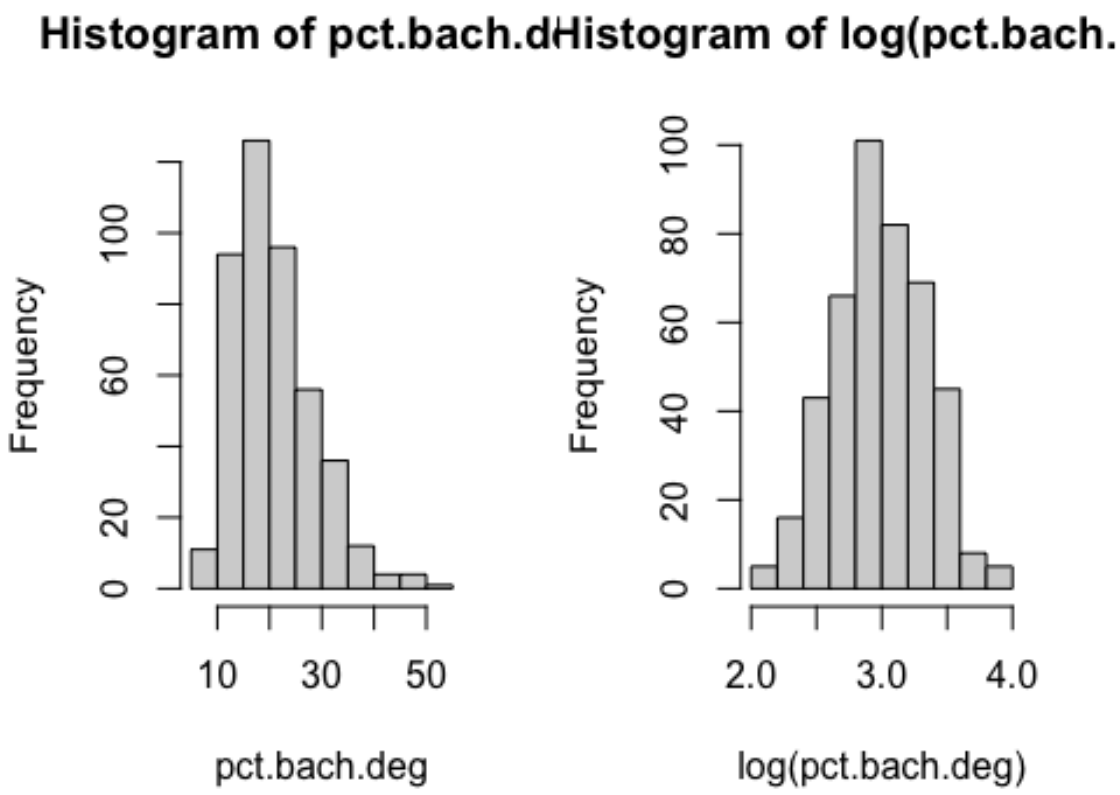

```
par(mfrow=c(1,2))
hist(crimes)
hist(log(crimes))
```



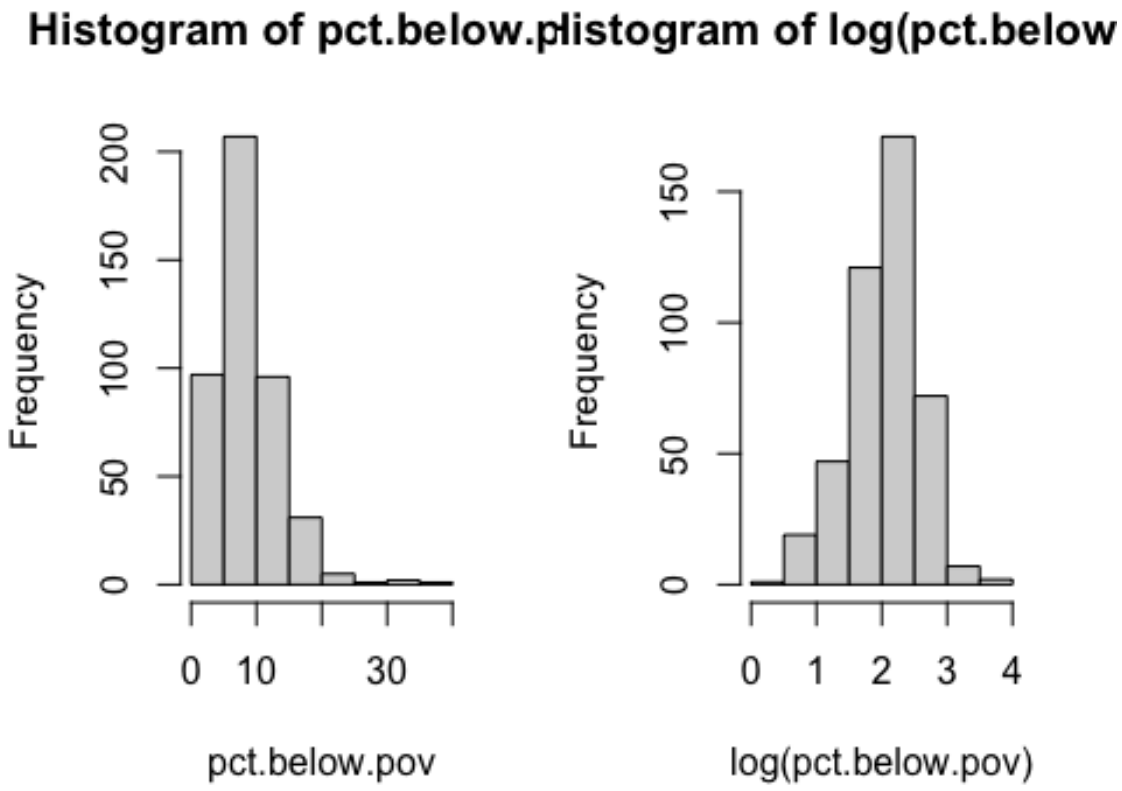

```
par(mfrow=c(2,2))
hist(pct.hs.grad)
hist((pct.hs.grad)^(2))
hist((pct.hs.grad)^(3))
```



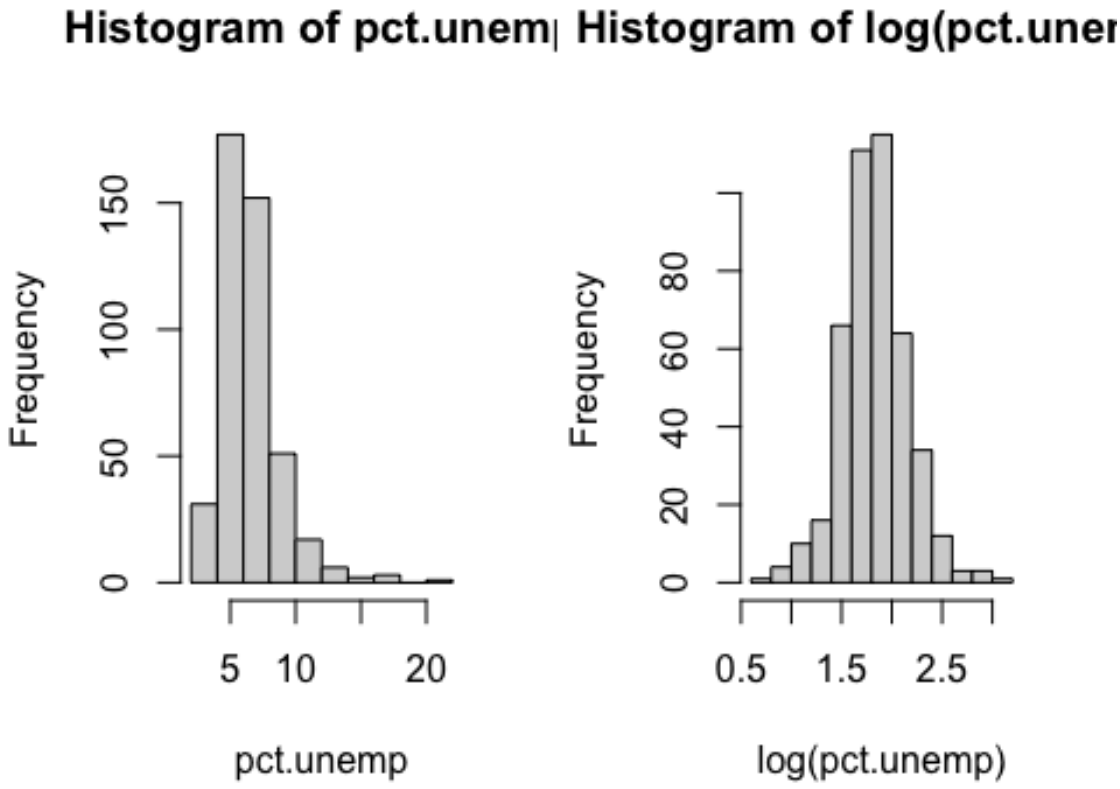

```
par(mfrow=c(1,2))
hist(pct.bach.deg)
hist(log(pct.bach.deg))
```



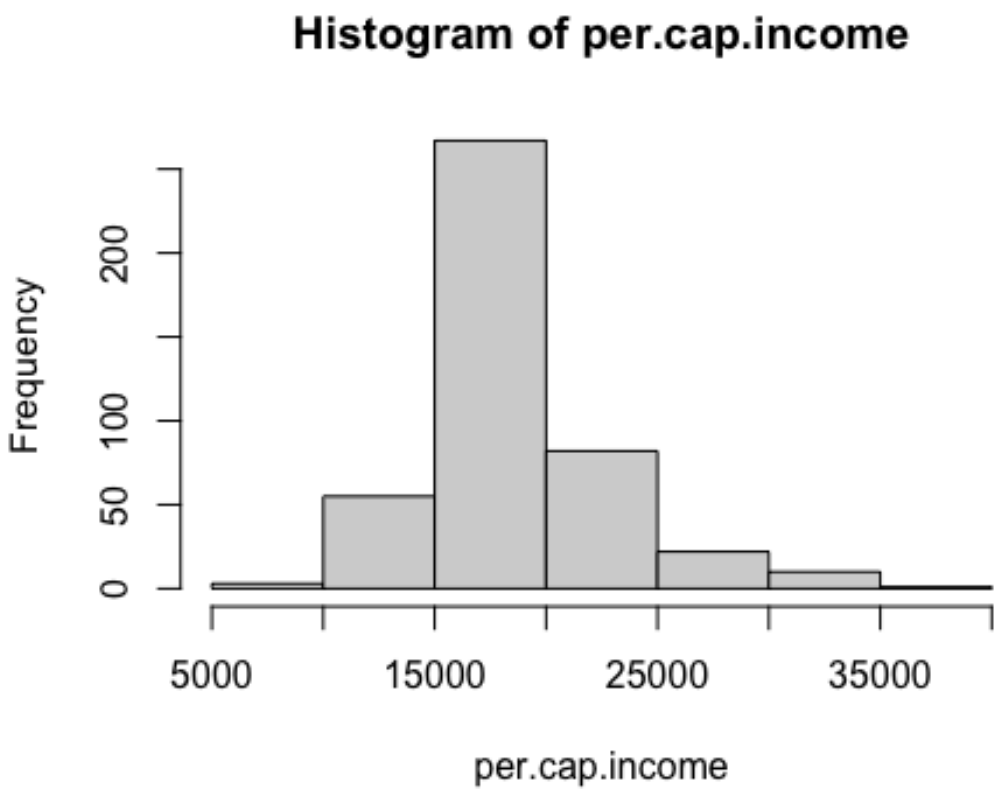

```
par(mfrow=c(1,2))
hist(pct.below.pov)
hist(log(pct.below.pov))
```



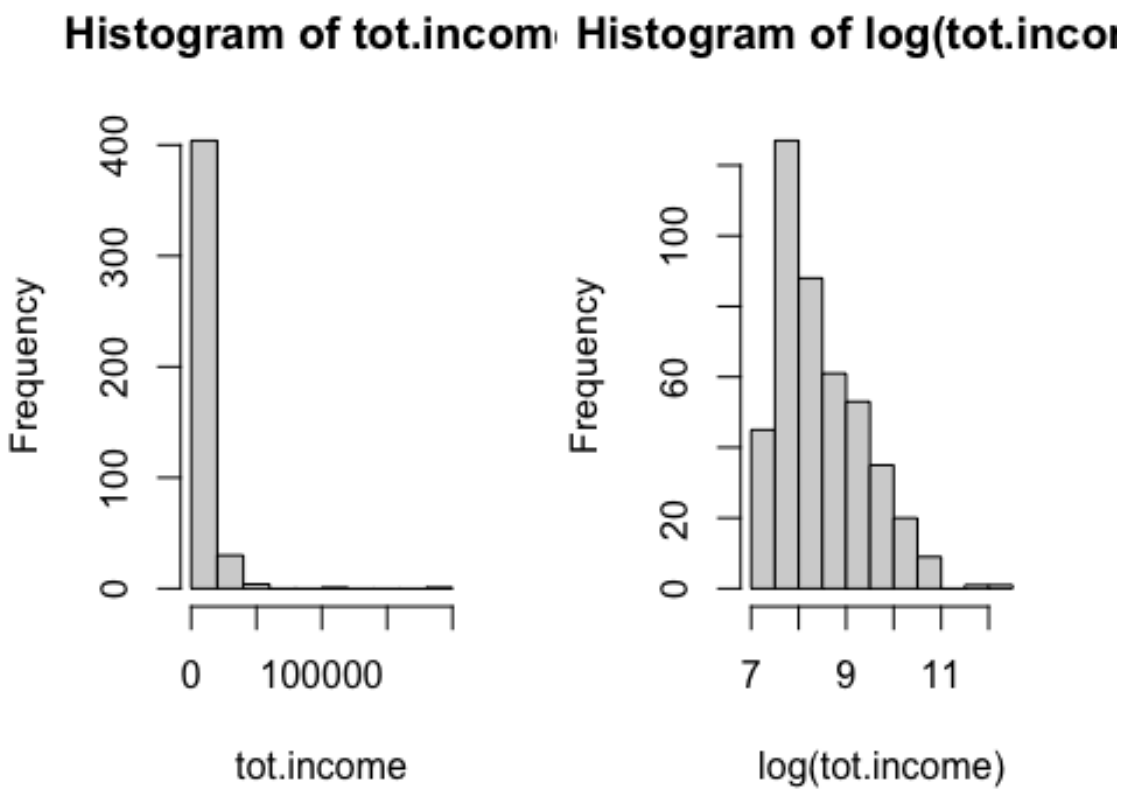

```
par(mfrow=c(1,2))
hist(pct.unemp)
hist(log(pct.unemp))
```



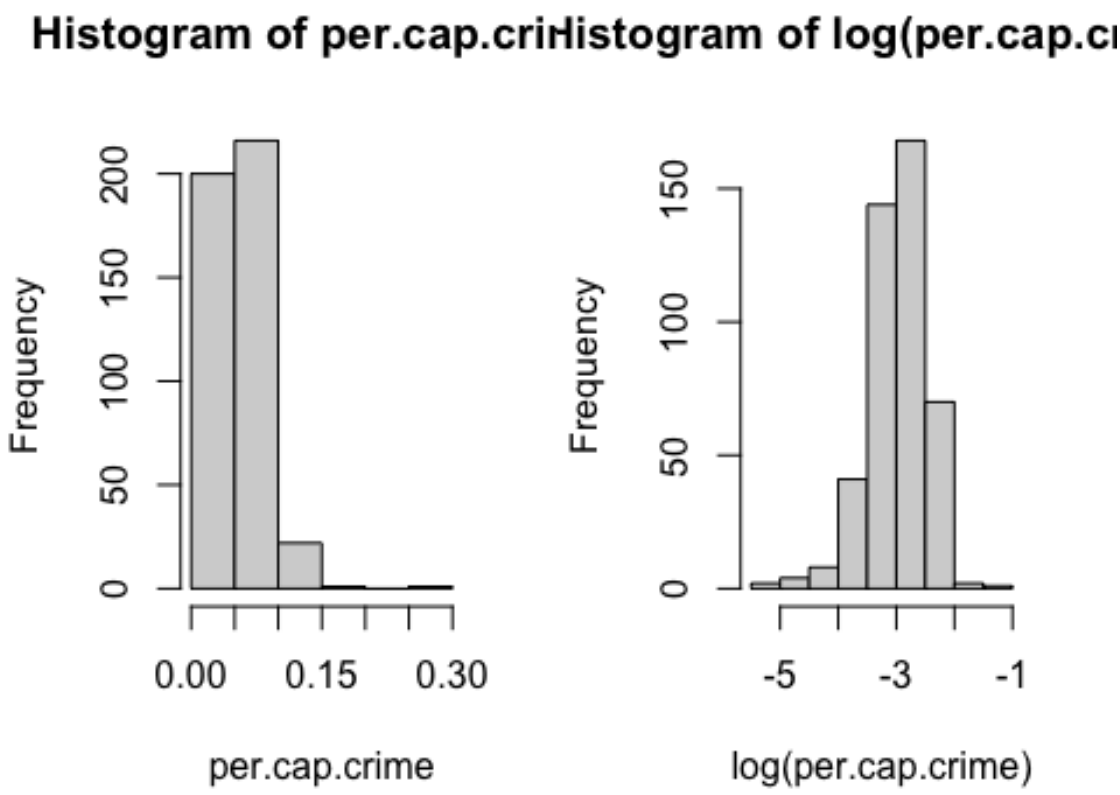

```
hist(per.cap.income)
```




```
par(mfrow=c(1,2))
hist(tot.income)
hist(log(tot.income))
```



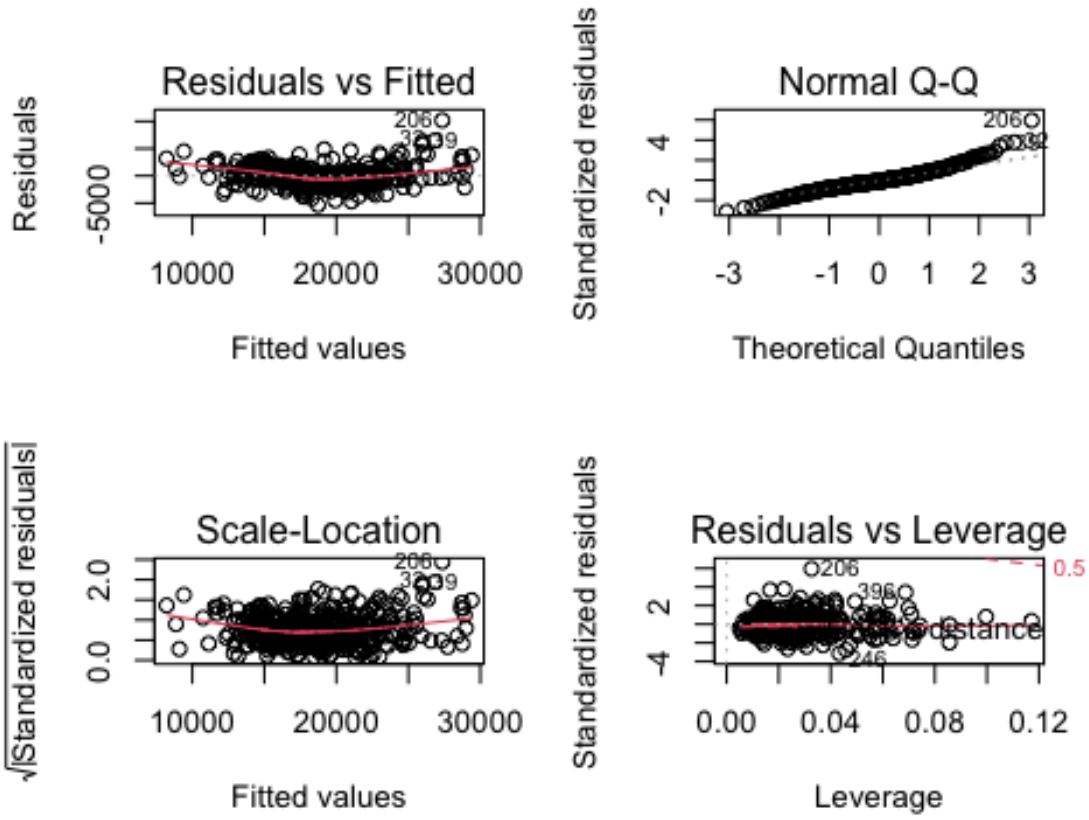

```
par(mfrow=c(1,2))
hist(per.cap.crime)
hist(log(per.cap.crime))
```



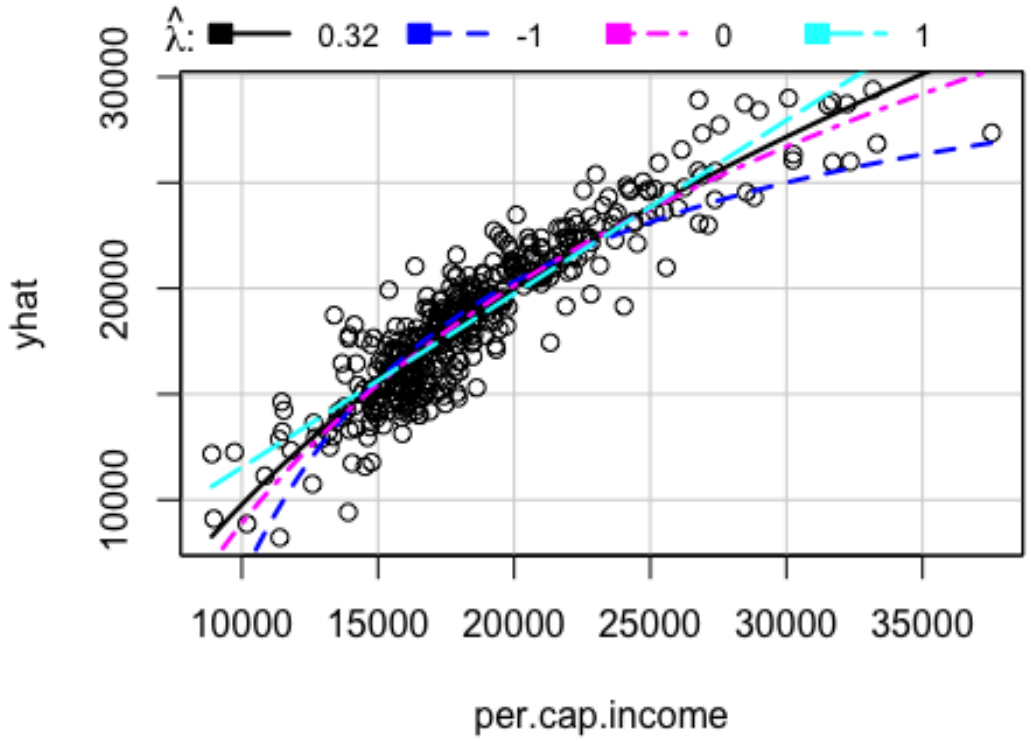
Variable Selection

```
fullmodel <- lm(per.cap.income ~ log(land.area) + log(per.cap.crime)
                + log(pct.unemp) + log(pct.below.pov)
                + log(pct.bach.deg) + (pct.hs.grad)^3
                + log(hosp.beds) + log(doctors) + log(pop.65_plus)
                + log(pop.18_34), data = cdi)

summary(fullmodel)

##
## Call:
## lm(formula = per.cap.income ~ log(land.area) + log(per.cap.crime) +
##     log(pct.unemp) + log(pct.below.pov) + log(pct.bach.deg) +
##     (pct.hs.grad)^3 + log(hosp.beds) + log(doctors) + log(pop.65_plus) +
##     log(pop.18_34), data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5338.8  -996.2  -221.6   822.1 10182.2
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        28683.34    4024.88   7.127 4.39e-12 ***
```

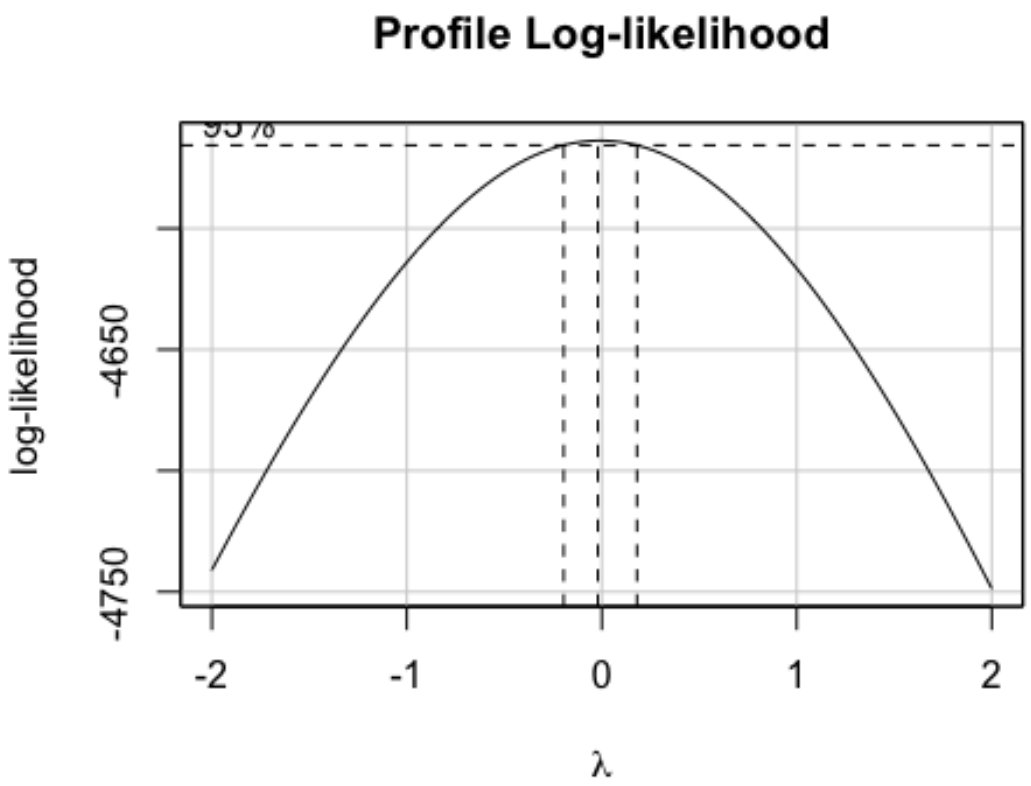
```
## log(land.area)      -653.02     101.85  -6.412 3.80e-10 ***
## log(per.cap.crime)   282.98     217.27   1.302 0.193456
## log(pct.unemp)      1688.07     340.03   4.964 9.96e-07 ***
## log(pct.below.pov) -4308.60     263.22 -16.369  < 2e-16 ***
## log(pct.bach.deg)   6067.05     513.56  11.814  < 2e-16 ***
## pct.hs.grad          -81.56      21.73  -3.754 0.000198 ***
## log(hosp.beds)      -123.18     253.19  -0.486 0.626865
## log(doctors)        1121.19     239.96   4.672 3.99e-06 ***
## log(pop.65_plus)     840.56     406.68   2.067 0.039342 *
## log(pop.18_34)     -5769.21     909.72  -6.342 5.77e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1740 on 429 degrees of freedom
## Multiple R-squared:  0.8205, Adjusted R-squared:  0.8163
## F-statistic: 196.1 on 10 and 429 DF,  p-value: < 2.2e-16

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




```
inverseResponsePlot(fullmodel)
```

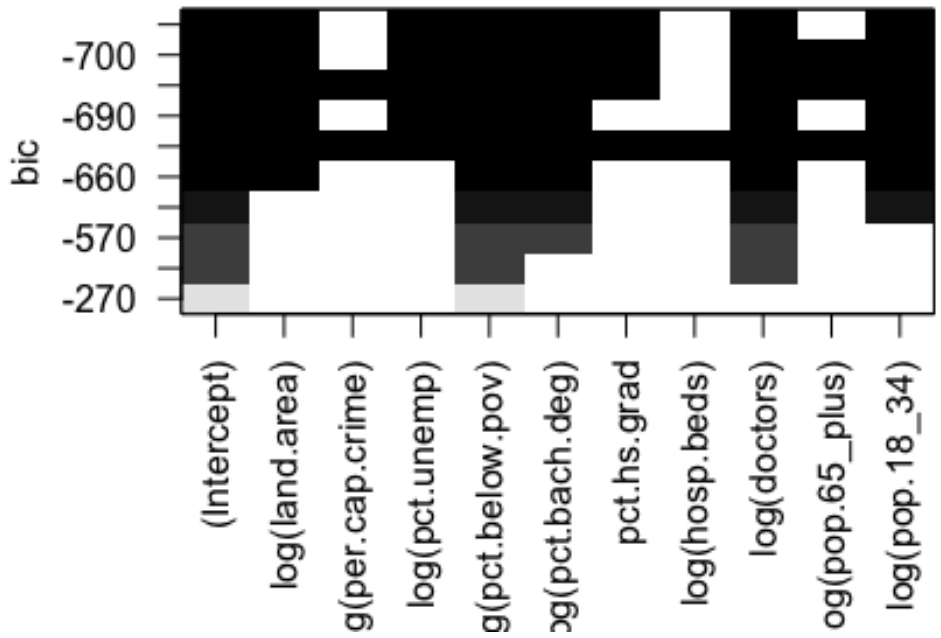



```
##       lambda        RSS
## 1  0.3240419  986682935
## 2 -1.0000000 1320831767
## 3  0.0000000 1006059816
## 4  1.0000000 1065325734

boxCox(fullmodel)
```



All Subsets

```
cdi.subsets <- regsubsets(per.cap.income ~ log(land.area) + log(per.cap.crime
)
                 + log(pct.unemp) + log(pct.below.pov)
                 + log(pct.bach.deg) + (pct.hs.grad)^3
                 + log(hosp.beds) + log(doctors) + log(pop.65_plus)
                 + log(pop.18_34) + log(land.area), data = cdi, nvmax=10)

plot(cdi.subsets)
```

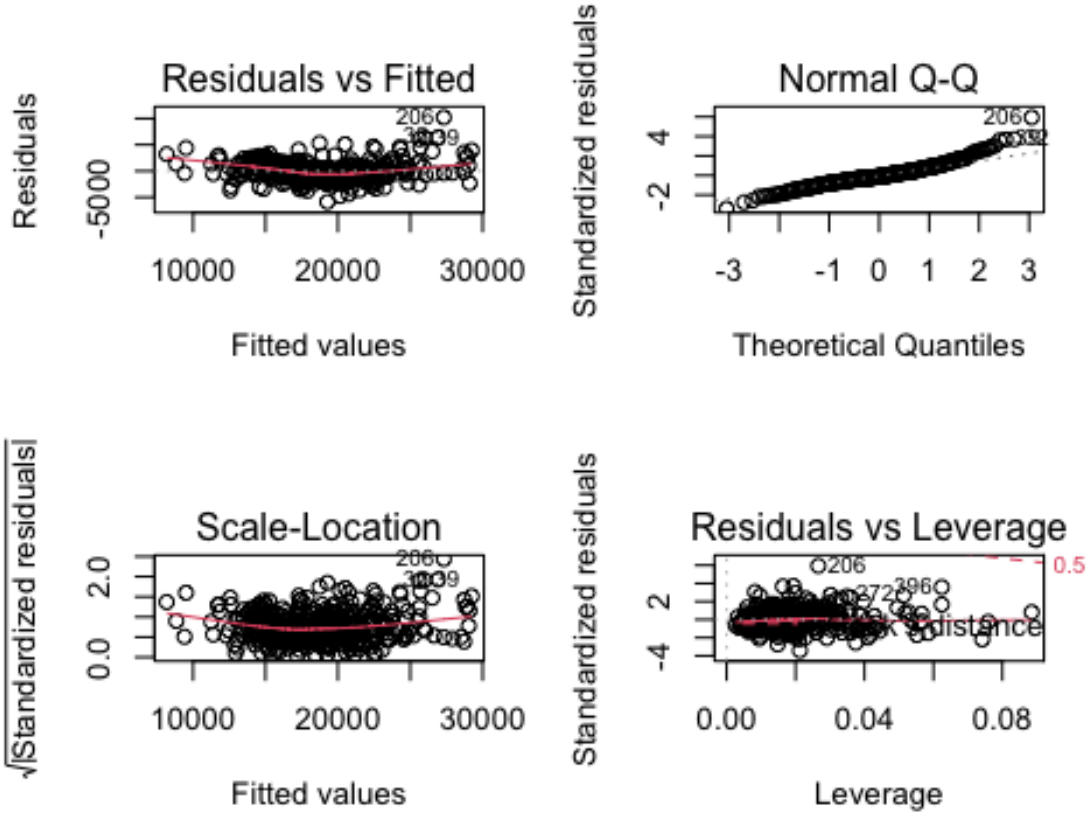



```
cdi.subsets.summary <- summary(cdi.subsets)
names(cdi.subsets.summary)

## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"

cdi.subsets.summary$bic

##  [1] -268.4528 -541.6377 -568.1158 -624.5520 -663.9394 -692.4577 -701.7173
##  [8] -699.0234 -694.6304 -688.7863

min(cdi.subsets.summary$bic)

## [1] -701.7173

print(best.model <- which.min(cdi.subsets.summary$bic))

## [1] 7

coef(cdi.subsets, best.model)

##        (Intercept)     log(land.area)     log(pct.unemp) log(pct.below.pov
)
##        32700.92650         -681.96985         1712.85362        -4165.2773
5
##  log(pct.bach.deg)        pct.hs.grad       log(doctors)     log(pop.18_34
```

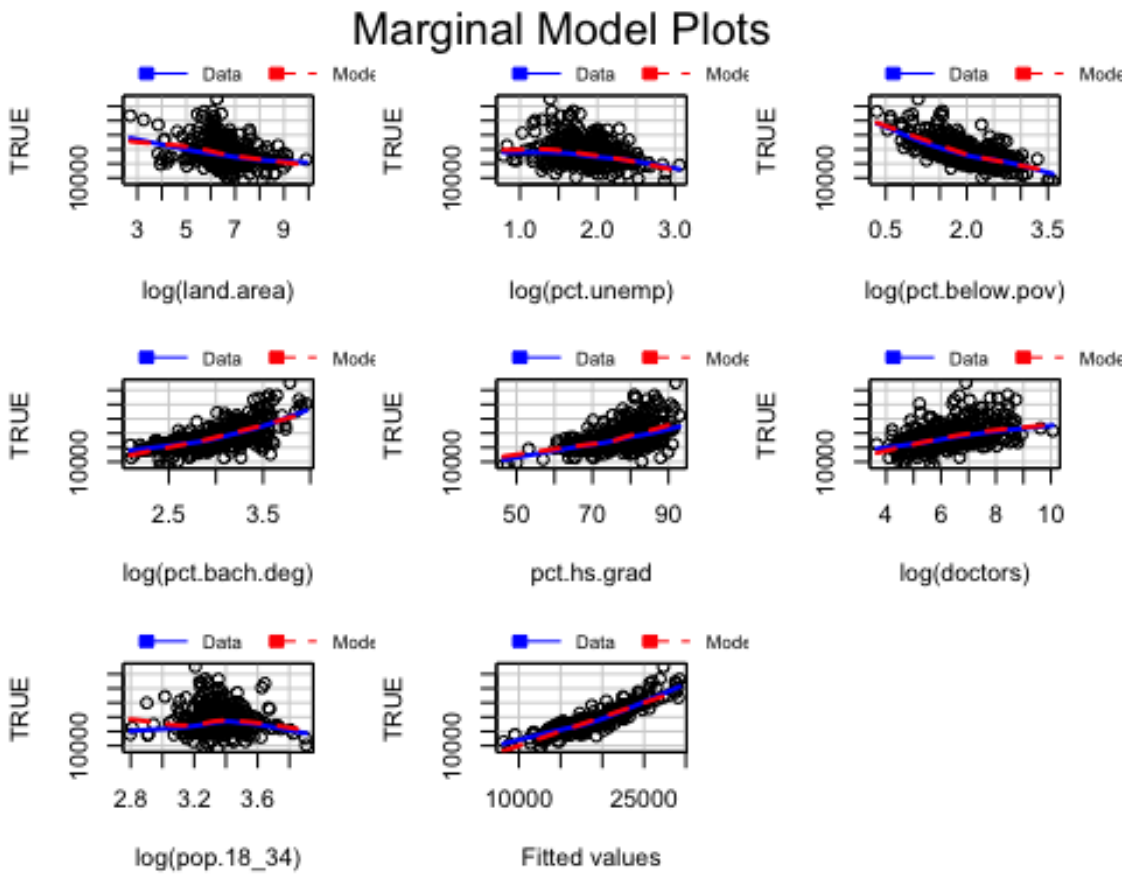
```
)
##         6137.59648          -84.97262         1111.08392        -6863.4662
9

cdi.subsets.final.model <- lm(per.cap.income ~ log(land.area)
                              + log(pct.unemp) + log(pct.below.pov)
                              + log(pct.bach.deg) + (pct.hs.grad)^3
                              + log(doctors) + log(pop.18_34), data = cdi)
summary(cdi.subsets.final.model)

##
## Call:
## lm(formula = per.cap.income ~ log(land.area) + log(pct.unemp) +
##     log(pct.below.pov) + log(pct.bach.deg) + (pct.hs.grad)^3 +
##     log(doctors) + log(pop.18_34), data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5884.1  -986.7  -213.2   809.1 10223.2
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        32700.93    2806.52  11.652  < 2e-16 ***
## log(land.area)      -681.97     100.57  -6.781 3.93e-11 ***
## log(pct.unemp)      1712.85     339.06   5.052 6.47e-07 ***
## log(pct.below.pov) -4165.28     229.08 -18.183  < 2e-16 ***
## log(pct.bach.deg)   6137.60     485.54  12.641  < 2e-16 ***
## pct.hs.grad          -84.97      21.70  -3.916 0.000105 ***
## log(doctors)        1111.08      91.42  12.153  < 2e-16 ***
## log(pop.18_34)     -6863.47     729.42  -9.409  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1744 on 432 degrees of freedom
## Multiple R-squared:  0.8183, Adjusted R-squared:  0.8154
## F-statistic:   278 on 7 and 432 DF,  p-value: < 2.2e-16

summary(cdi.subsets.final.model)$coef

##                       Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)        32700.92650 2806.52362  11.651755 1.802367e-27
## log(land.area)      -681.96985  100.56687  -6.781258 3.929352e-11
## log(pct.unemp)      1712.85362  339.06173   5.051746 6.472157e-07
## log(pct.below.pov) -4165.27735  229.08120 -18.182537 2.831074e-55
## log(pct.bach.deg)   6137.59648  485.53550  12.640881 2.195710e-31
## pct.hs.grad          -84.97262   21.70014  -3.915764 1.047011e-04
## log(doctors)        1111.08392   91.42147  12.153425 1.948148e-29
## log(pop.18_34)     -6863.46629  729.41915  -9.409496 2.981996e-19

vif(cdi.subsets.final.model)
```

```
##     log(land.area)     log(pct.unemp) log(pct.below.pov)  log(pct.bach.deg
)
##           1.109011           1.765859           2.131198           4.27391
9
##        pct.hs.grad       log(doctors)     log(pop.18_34)
##           3.344148           1.578605           1.524308
```

```
par(mfrow=c(2,2))
plot(cdi.subsets.final.model)
```




```
mmps(cdi.subsets.final.model)
```

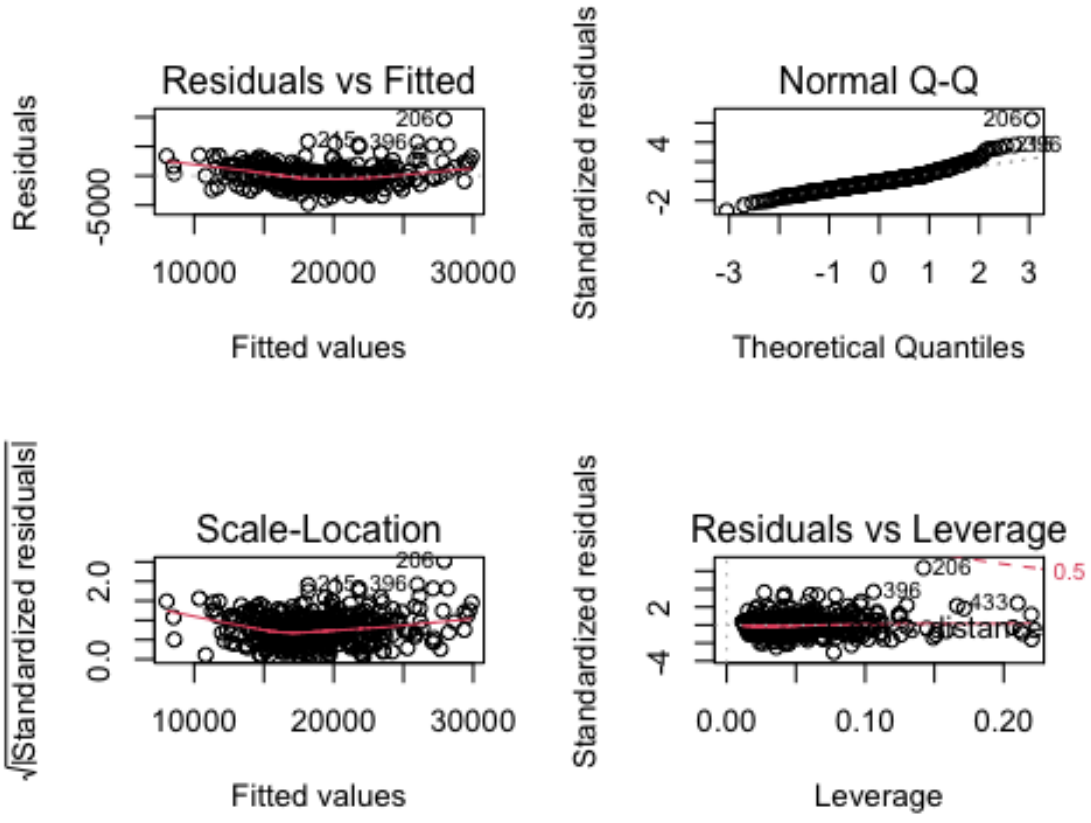



```
cdisub <- cdi[-c(1:3, 5, 10, 16)]
cdi_all_subsets <- cdisub[c(1:2, 4, 6:11)]

cdi_all_subsets <- cdi_all_subsets %>%
  mutate(land.area = log(land.area)) %>%
  dplyr::rename(log.land.area = land.area) %>%
  mutate(pop.18_34 = log(pop.18_34)) %>%
  dplyr::rename(log.pop.18_34 = pop.18_34) %>%
  mutate(pct.unemp = log(pct.unemp)) %>%
  dplyr::rename(log.pct.unemp = pct.unemp) %>%
  mutate(doctors = log(doctors)) %>%
  dplyr::rename(log.doctors = doctors) %>%
  mutate(pct.bach.deg = log(pct.bach.deg)) %>%
  dplyr::rename(log.pct.bach.deg = pct.bach.deg) %>%
  mutate(pct.below.pov = log(pct.below.pov)) %>%
  dplyr::rename(log.pct.below.pov = pct.below.pov) %>%
  mutate(pct.hs.grad = (pct.hs.grad)^3) %>%
  dplyr::rename(pct.hs.grad.3 = pct.hs.grad)

cdi.subsets.final.model.with.region <- lm(per.cap.income ~ .*region,
                                          data=cdi_all_subsets)
summary(cdi.subsets.final.model.with.region)
```

```
##
## Call:
## lm(formula = per.cap.income ~ . * region, data = cdi_all_subsets)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4751.4  -965.6  -130.4   678.9  9679.8
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.844e+04  5.480e+03   5.190 3.32e-07 ***
## log.land.area              -4.389e+02  3.216e+02  -1.365 0.173095
## log.pop.18_34              -7.156e+03  1.792e+03  -3.993 7.75e-05 ***
## log.doctors                 1.017e+03  2.056e+02   4.948 1.10e-06 ***
## pct.hs.grad.3              -6.339e-04  3.889e-03  -0.163 0.870589
## log.pct.bach.deg            4.657e+03  1.357e+03   3.431 0.000664 ***
## log.pct.below.pov          -3.200e+03  5.521e+02  -5.797 1.36e-08 ***
## log.pct.unemp               1.982e+03  6.685e+02   2.965 0.003207 **
## regionNE                    9.776e+03  7.935e+03   1.232 0.218666
## regionS                     1.759e+03  6.890e+03   0.255 0.798592
## regionW                     6.459e+03  8.667e+03   0.745 0.456519
## log.land.area:regionNE     -1.492e+02  4.282e+02  -0.348 0.727646
## log.land.area:regionS      -4.326e+02  3.709e+02  -1.166 0.244139
## log.land.area:regionW      -2.388e+02  3.867e+02  -0.617 0.537271
## log.pop.18_34:regionNE     -4.511e+03  2.602e+03  -1.733 0.083785 .
## log.pop.18_34:regionS       5.709e+01  2.083e+03   0.027 0.978145
## log.pop.18_34:regionW       1.794e+03  2.597e+03   0.691 0.489937
## log.doctors:regionNE        4.744e+01  3.004e+02   0.158 0.874619
## log.doctors:regionS        -1.956e+02  2.586e+02  -0.756 0.449858
## log.doctors:regionW        -1.174e+02  2.907e+02  -0.404 0.686638
## pct.hs.grad.3:regionNE     -7.450e-03  5.315e-03  -1.402 0.161790
## pct.hs.grad.3:regionS      -5.334e-03  4.587e-03  -1.163 0.245531
## pct.hs.grad.3:regionW      -1.936e-02  5.050e-03  -3.833 0.000146 ***
## log.pct.bach.deg:regionNE   3.990e+03  1.983e+03   2.012 0.044842 *
## log.pct.bach.deg:regionS    2.485e+03  1.603e+03   1.550 0.121836
## log.pct.bach.deg:regionW    3.285e+03  1.766e+03   1.860 0.063536 .
## log.pct.below.pov:regionNE -7.876e+02  7.768e+02  -1.014 0.311234
## log.pct.below.pov:regionS  -4.538e+02  6.896e+02  -0.658 0.510818
## log.pct.below.pov:regionW  -3.847e+03  9.403e+02  -4.091 5.18e-05 ***
## log.pct.unemp:regionNE     -5.526e+02  1.108e+03  -0.499 0.618326
## log.pct.unemp:regionS      -1.343e+03  9.540e+02  -1.408 0.159820
## log.pct.unemp:regionW      -9.855e+02  9.924e+02  -0.993 0.321281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1642 on 408 degrees of freedom
## Multiple R-squared:  0.848,  Adjusted R-squared:  0.8364
## F-statistic: 73.41 on 31 and 408 DF,  p-value: < 2.2e-16
```

```
cdi.subsets.final.model.with.some.region <- update(cdi.subsets.final.model.wi
th.region,
                                                   . ~ . - region:log.land.ar
ea -
                                                     region:log.doctors -
                                                     region:log.pct.unemp)
summary(cdi.subsets.final.model.with.some.region)

##
## Call:
## lm(formula = per.cap.income ~ log.land.area + log.pop.18_34 +
##     log.doctors + pct.hs.grad.3 + log.pct.bach.deg + log.pct.below.pov +
##     log.pct.unemp + region + log.pop.18_34:region + pct.hs.grad.3:region +

##     log.pct.bach.deg:region + log.pct.below.pov:region, data = cdi_all_sub
sets)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4787.7  -982.3  -135.6   714.8  9626.3
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.360e+04  4.645e+03   7.235 2.25e-12 ***
## log.land.area               -7.042e+02  1.160e+02  -6.070 2.88e-09 ***
## log.pop.18_34               -7.613e+03  1.676e+03  -4.543 7.28e-06 ***
## log.doctors                  9.303e+02  9.476e+01   9.818  < 2e-16 ***
## pct.hs.grad.3               -1.494e-03  3.493e-03  -0.428 0.669130
## log.pct.bach.deg             4.676e+03  1.022e+03   4.574 6.33e-06 ***
## log.pct.below.pov           -3.094e+03  5.243e+02  -5.901 7.49e-09 ***
## log.pct.unemp                1.264e+03  3.599e+02   3.513 0.000491 ***
## regionNE                     6.507e+03  6.535e+03   0.996 0.319934
## regionS                     -7.750e+03  5.118e+03  -1.514 0.130705
## regionW                      8.801e+02  7.303e+03   0.120 0.904145
## log.pop.18_34:regionNE      -4.386e+03  2.371e+03  -1.849 0.065105 .
## log.pop.18_34:regionS        1.143e+03  1.926e+03   0.594 0.553079
## log.pop.18_34:regionW        2.167e+03  2.495e+03   0.869 0.385500
## pct.hs.grad.3:regionNE      -6.806e-03  4.468e-03  -1.523 0.128469
## pct.hs.grad.3:regionS       -4.696e-03  4.207e-03  -1.116 0.264941
## pct.hs.grad.3:regionW       -1.791e-02  4.387e-03  -4.082 5.34e-05 ***
## log.pct.bach.deg:regionNE    4.292e+03  1.285e+03   3.340 0.000913 ***
## log.pct.bach.deg:regionS     2.482e+03  1.147e+03   2.164 0.030997 *
## log.pct.bach.deg:regionW     3.229e+03  1.374e+03   2.350 0.019231 *
## log.pct.below.pov:regionNE -7.567e+02  7.117e+02  -1.063 0.288320
## log.pct.below.pov:regionS  -8.122e+02  6.451e+02  -1.259 0.208756
## log.pct.below.pov:regionW  -3.949e+03  9.132e+02  -4.324 1.92e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1635 on 417 degrees of freedom
```

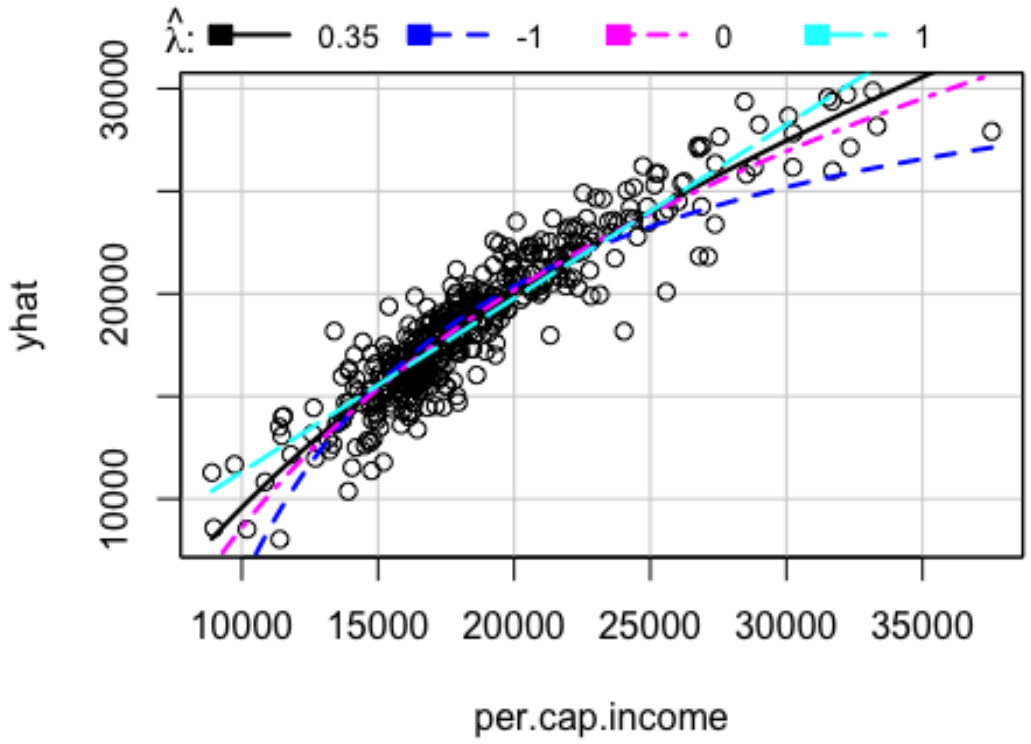
```
## Multiple R-squared:  0.846,  Adjusted R-squared:  0.8378
## F-statistic: 104.1 on 22 and 417 DF,  p-value: < 2.2e-16

vif(cdi.subsets.final.model.with.some.region)

##                                  GVIF Df GVIF^(1/(2*Df))
## log.land.area            1.680366e+00  1        1.296289
## log.pop.18_34            9.161875e+00  1        3.026859
## log.doctors              1.931049e+00  1        1.389622
## pct.hs.grad.3            2.948870e+01  1        5.430350
## log.pct.bach.deg         2.157800e+01  1        4.645212
## log.pct.below.pov        1.271309e+01  1        3.565542
## log.pct.unemp            2.265215e+00  1        1.505063
## region                   4.593495e+08  3       27.777333
## log.pop.18_34:region     1.183970e+09  3       32.525461
## pct.hs.grad.3:region     6.001590e+05  3        9.184265
## log.pct.bach.deg:region  2.276982e+07  3       16.835523
## log.pct.below.pov:region 1.275326e+05  3        7.094747

par(mfrow=c(2,2))
plot(cdi.subsets.final.model.with.some.region)
```



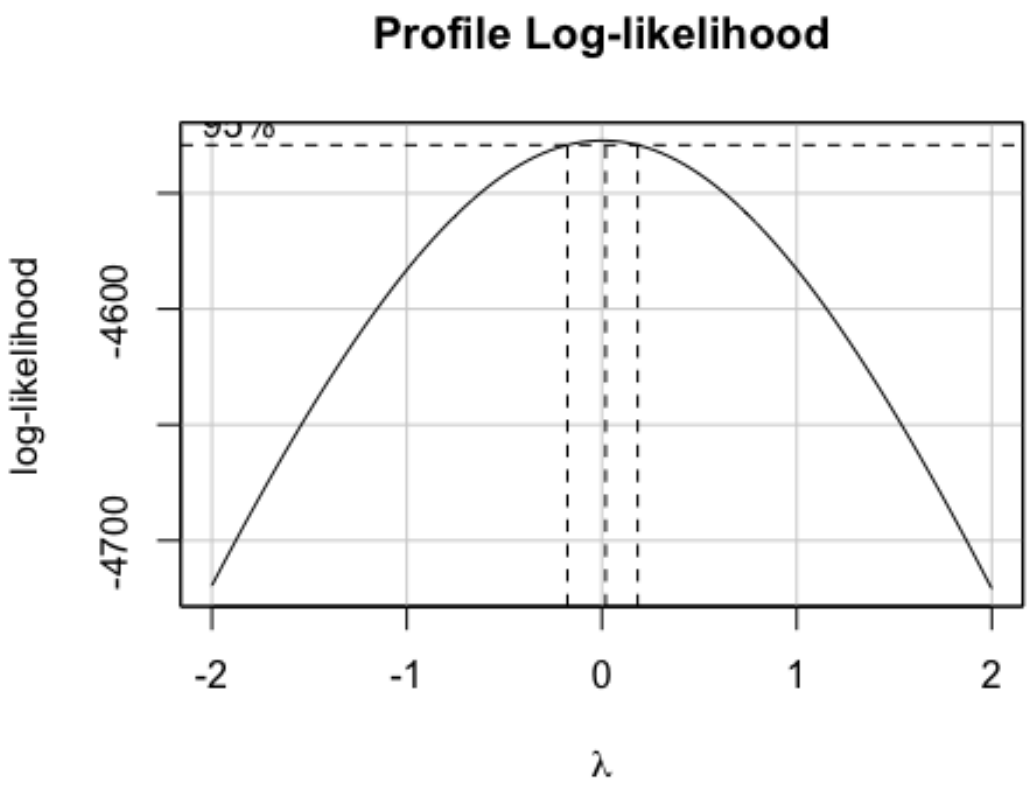

```
anova(cdi.subsets.final.model, cdi.subsets.final.model.with.some.region)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ log(land.area) + log(pct.unemp) + log(pct.below.
pov) +
##     log(pct.bach.deg) + (pct.hs.grad)^3 + log(doctors) + log(pop.18_34)
## Model 2: per.cap.income ~ log.land.area + log.pop.18_34 + log.doctors +
##     pct.hs.grad.3 + log.pct.bach.deg + log.pct.below.pov + log.pct.unemp +

##     region + log.pop.18_34:region + pct.hs.grad.3:region + log.pct.bach.de
g:region +
##     log.pct.below.pov:region
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    432 1314204306
## 2    417 1114159243 15 200045063 4.9914 4.626e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC(cdi.subsets.final.model, cdi.subsets.final.model.with.some.region)

##                                           df      AIC
## cdi.subsets.final.model                    9 7826.944
## cdi.subsets.final.model.with.some.region 24 7784.286

BIC(cdi.subsets.final.model, cdi.subsets.final.model.with.some.region)

##                                           df      BIC
## cdi.subsets.final.model                    9 7863.725
## cdi.subsets.final.model.with.some.region 24 7882.369

formula(cdi.subsets.final.model.with.some.region)

## per.cap.income ~ log.land.area + log.pop.18_34 + log.doctors +
##     pct.hs.grad.3 + log.pct.bach.deg + log.pct.below.pov + log.pct.unemp +

##     region + log.pop.18_34:region + pct.hs.grad.3:region + log.pct.bach.de
g:region +
##     log.pct.below.pov:region

round(summary(cdi.subsets.final.model.with.some.region)$coef,2)

##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                33603.04    4644.53    7.23     0.00
## log.land.area               -704.21     116.01   -6.07     0.00
## log.pop.18_34              -7613.16    1675.90   -4.54     0.00
## log.doctors                  930.31      94.76    9.82     0.00
## pct.hs.grad.3                  0.00       0.00   -0.43     0.67
## log.pct.bach.deg            4676.09    1022.42    4.57     0.00
## log.pct.below.pov          -3094.10     524.35   -5.90     0.00
## log.pct.unemp               1264.35     359.89    3.51     0.00
## regionNE                    6507.45    6535.02    1.00     0.32
## regionS                    -7750.28    5118.03   -1.51     0.13
## regionW                      880.06    7303.42    0.12     0.90
```

```
## log.pop.18_34:regionNE     -4385.78    2371.44   -1.85     0.07
## log.pop.18_34:regionS       1143.33    1925.98    0.59     0.55
## log.pop.18_34:regionW       2167.35    2494.87    0.87     0.39
## pct.hs.grad.3:regionNE        -0.01       0.00   -1.52     0.13
## pct.hs.grad.3:regionS          0.00       0.00   -1.12     0.26
## pct.hs.grad.3:regionW         -0.02       0.00   -4.08     0.00
## log.pct.bach.deg:regionNE   4292.42    1285.09    3.34     0.00
## log.pct.bach.deg:regionS    2481.75    1146.60    2.16     0.03
## log.pct.bach.deg:regionW    3229.20    1374.04    2.35     0.02
## log.pct.below.pov:regionNE  -756.68     711.72   -1.06     0.29
## log.pct.below.pov:regionS   -812.19     645.14   -1.26     0.21
## log.pct.below.pov:regionW  -3948.57     913.22   -4.32     0.00

inverseResponsePlot(cdi.subsets.final.model.with.some.region)
```




```
##       lambda        RSS
## 1  0.3502333  867768343
## 2 -1.0000000 1225709535
## 3  0.0000000  891058518
## 4  1.0000000  942545990

boxCox(cdi.subsets.final.model.with.some.region)
```



Appendix D

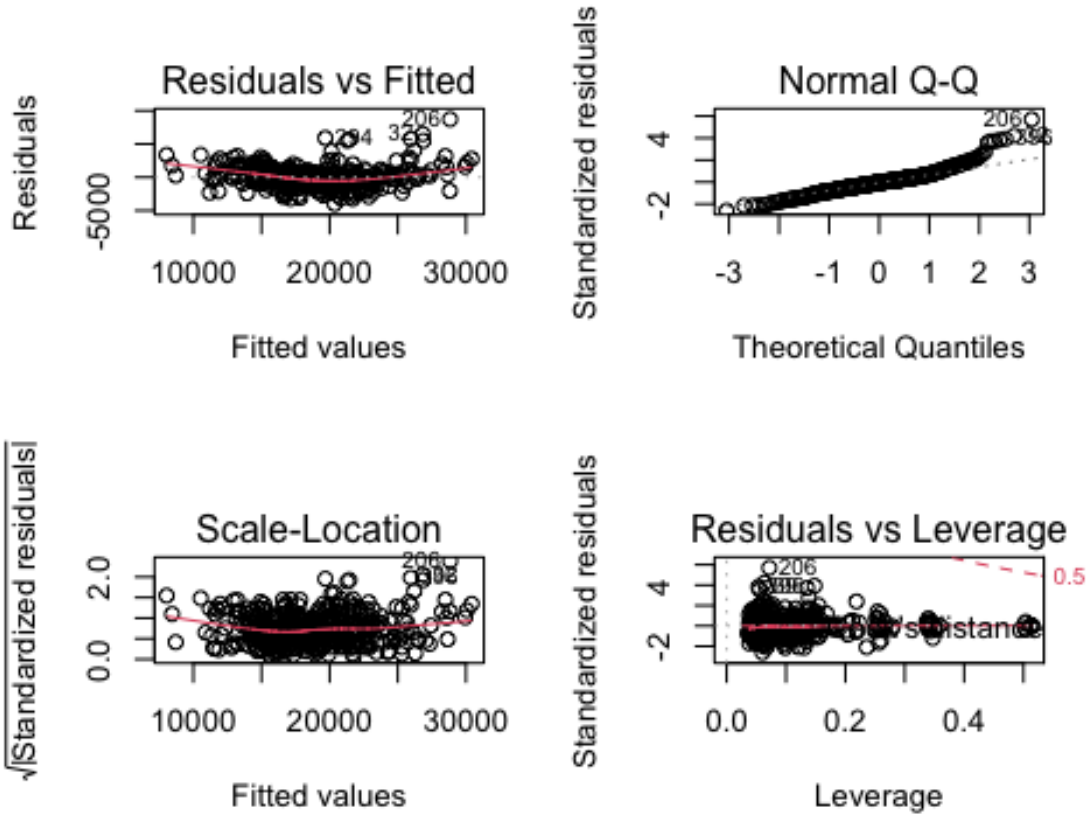
```
cdi.cont <- data.frame(cdi_all_subsets, state=cdi$state)
cdi.cont <- cdi.cont[-c(9)]

final.model.with.state <- lm(per.cap.income ~ ., data=cdi.cont)
summary(final.model.with.state)

##
## Call:
## lm(formula = per.cap.income ~ ., data = cdi.cont)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4005.2  -822.5   -42.7   603.6  8704.6
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3.459e+04  2.834e+03  12.207  < 2e-16 ***
## log.land.area     -7.550e+02  1.290e+02  -5.855 1.02e-08 ***
## log.pop.18_34     -8.270e+03  7.511e+02 -11.010  < 2e-16 ***
## log.doctors        9.530e+02  9.203e+01  10.356  < 2e-16 ***
## pct.hs.grad.3     -4.417e-03  1.672e-03  -2.641 0.008596 **
## log.pct.bach.deg   6.195e+03  5.792e+02  10.696  < 2e-16 ***
```

```
## log.pct.below.pov -3.487e+03  2.912e+02 -11.976  < 2e-16 ***
## log.pct.unemp      5.698e+02  4.983e+02   1.143 0.253543
## stateAR           -7.594e+02  1.276e+03  -0.595 0.552103
## stateAZ           -3.213e+02  9.770e+02  -0.329 0.742457
## stateCA            1.928e+03  7.022e+02   2.746 0.006316 **
## stateCO            4.889e+02  8.326e+02   0.587 0.557410
## stateCT            1.577e+03  8.746e+02   1.803 0.072168 .
## stateDC            1.226e+03  1.753e+03   0.700 0.484632
## stateDE            3.999e+02  1.290e+03   0.310 0.756811
## stateFL           -5.530e+02  6.950e+02  -0.796 0.426667
## stateGA            4.355e+02  8.204e+02   0.531 0.595814
## stateHI            6.876e+02  1.141e+03   0.602 0.547280
## stateID           -1.686e+02  1.722e+03  -0.098 0.922070
## stateIL            9.757e+02  7.451e+02   1.309 0.191192
## stateIN           -4.633e+01  7.606e+02  -0.061 0.951458
## stateKS            3.070e+02  1.032e+03   0.297 0.766311
## stateKY           -3.701e+02  1.110e+03  -0.333 0.739056
## stateLA           -1.874e+02  8.090e+02  -0.232 0.816929
## stateMA            6.244e+02  8.551e+02   0.730 0.465765
## stateMD            2.070e+02  8.325e+02   0.249 0.803753
## stateME            1.815e+02  9.541e+02   0.190 0.849209
## stateMI            1.338e+03  7.681e+02   1.742 0.082269 .
## stateMN           -9.703e+01  8.915e+02  -0.109 0.913388
## stateMO            5.894e+01  8.480e+02   0.070 0.944623
## stateMS           -9.158e+02  1.102e+03  -0.831 0.406349
## stateMT            4.318e+02  1.721e+03   0.251 0.802026
## stateNC            1.708e+02  7.273e+02   0.235 0.814483
## stateND            5.069e+01  1.747e+03   0.029 0.976869
## stateNE           -9.194e+02  1.171e+03  -0.785 0.432875
## stateNH           -6.326e+01  1.050e+03  -0.060 0.952008
## stateNJ            2.155e+03  7.689e+02   2.803 0.005325 **
## stateNM           -1.455e+03  1.284e+03  -1.133 0.257811
## stateNV            4.176e+03  1.327e+03   3.146 0.001782 **
## stateNY            5.421e+02  7.160e+02   0.757 0.449478
## stateOH            3.382e+02  7.148e+02   0.473 0.636382
## stateOK           -7.892e+02  1.015e+03  -0.777 0.437471
## stateOR           -1.088e+03  9.174e+02  -1.186 0.236496
## statePA           -2.777e+02  7.032e+02  -0.395 0.693151
## stateRI           -1.991e+03  1.148e+03  -1.735 0.083480 .
## stateSC           -2.025e+02  7.720e+02  -0.262 0.793217
## stateSD            3.279e+02  1.745e+03   0.188 0.851042
## stateTN           -3.497e+02  8.269e+02  -0.423 0.672590
## stateTX           -2.591e+01  6.735e+02  -0.038 0.969331
## stateUT           -4.061e+03  1.047e+03  -3.877 0.000124 ***
## stateVA            7.491e+02  8.826e+02   0.849 0.396594
## stateVT           -7.505e+02  1.716e+03  -0.437 0.662016
## stateWA           -1.601e+02  8.325e+02  -0.192 0.847561
## stateWI            2.053e+02  8.102e+02   0.253 0.800128
## stateWV           -5.462e+02  1.702e+03  -0.321 0.748382
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1585 on 385 degrees of freedom
## Multiple R-squared:  0.8662, Adjusted R-squared:  0.8475
## F-statistic: 46.16 on 54 and 385 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(final.model.with.state)
```

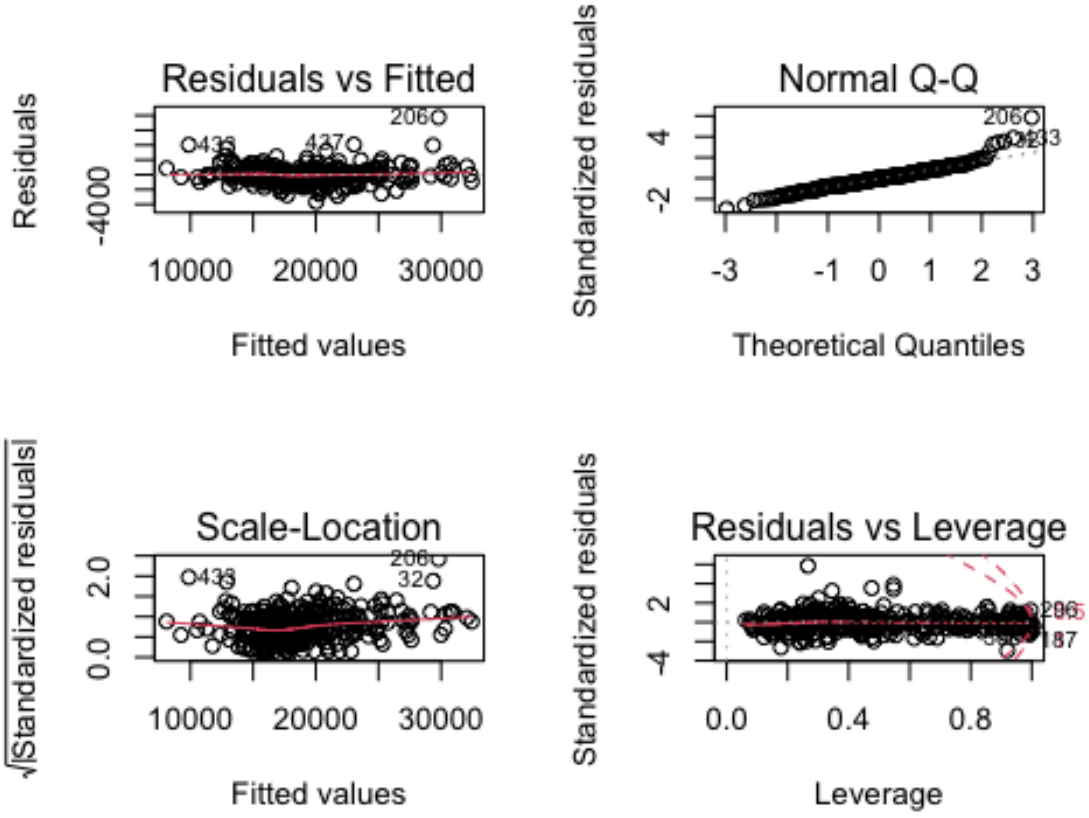
```
## Warning: not plotting observations with leverage one:
##   73, 232, 233, 339, 356, 388, 429
```




```
final.model.with.state.int <- lm(per.cap.income ~ .*state, data=cdi.cont)
par(mfrow=c(2,2))
plot(final.model.with.state.int)
```

```
## Warning: not plotting observations with leverage one:
##   7, 27, 28, 35, 37, 39, 40, 43, 49, 52, 65, 66, 68, 71, 73, 74, 75, 78, 9
0, 92, 97, 99, 109, 118, 120, 123, 130, 141, 143, 148, 149, 156, 167, 169, 17
2, 175, 184, 190, 191, 192, 199, 200, 203, 209, 213, 225, 226, 230, 231, 232,
 233, 235, 246, 267, 268, 271, 276, 280, 282, 285, 286, 288, 291, 298, 309, 3
11, 312, 319, 320, 323, 334, 339, 348, 356, 368, 370, 375, 377, 380, 381, 382
, 385, 388, 389, 392, 395, 397, 402, 412, 415, 418, 421, 425, 428, 429, 431,
432, 438, 439, 440
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```




```
anova(cdi.subsets.final.model, final.model.with.state, final.model.with.state
.int)

## Analysis of Variance Table
##
## Model 1: per.cap.income ~ log(land.area) + log(pct.unemp) + log(pct.below.
pov) +
##     log(pct.bach.deg) + (pct.hs.grad)^3 + log(doctors) + log(pop.18_34)
## Model 2: per.cap.income ~ log.land.area + log.pop.18_34 + log.doctors +
##     pct.hs.grad.3 + log.pct.bach.deg + log.pct.below.pov + log.pct.unemp +

##     state
## Model 3: per.cap.income ~ (log.land.area + log.pop.18_34 + log.doctors +
##     pct.hs.grad.3 + log.pct.bach.deg + log.pct.below.pov + log.pct.unemp +

##     state) * state
##   Res.Df        RSS  Df Sum of Sq      F    Pr(>F)
## 1    432 1314204306
## 2    385  967714409  47 346489897 3.0968 3.526e-08 ***
```

```
## 3    180  428500031 205 539214379 1.1049    0.2464
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(cdi.subsets.final.model, final.model.with.state, final.model.with.state.i
nt)
```

```
##                              df      AIC
## cdi.subsets.final.model       9 7826.944
## final.model.with.state       56 7786.282
## final.model.with.state.int 261 7837.838
```

```
BIC(cdi.subsets.final.model, final.model.with.state, final.model.with.state.i
nt)
```

```
##                              df      BIC
## cdi.subsets.final.model       9 7863.725
## final.model.with.state       56 8015.141
## final.model.with.state.int 261 8904.486
```

```
formula(final.model.with.state)
```

```
## per.cap.income ~ log.land.area + log.pop.18_34 + log.doctors +
##     pct.hs.grad.3 + log.pct.bach.deg + log.pct.below.pov + log.pct.unemp +

##     state
```

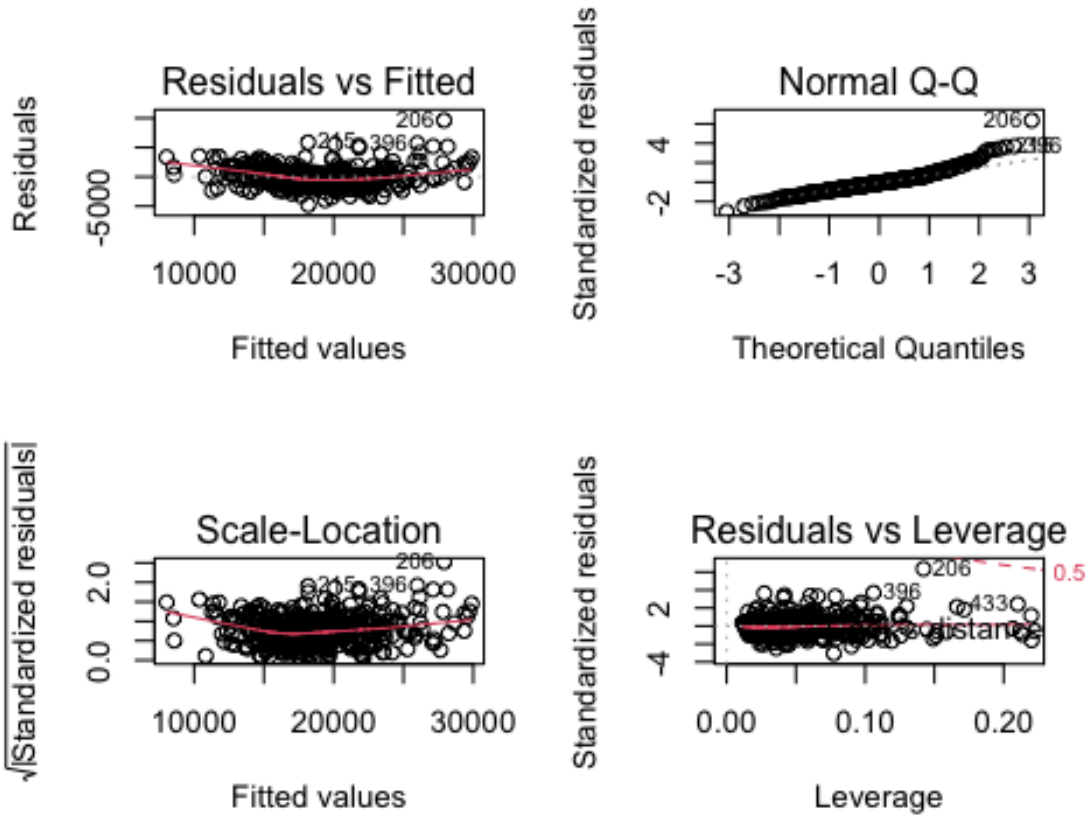
```
round(summary(final.model.with.state)$coef,2)
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       34590.94    2833.74   12.21     0.00
## log.land.area      -755.02     128.95   -5.85     0.00
## log.pop.18_34     -8269.77     751.12  -11.01     0.00
## log.doctors         953.01      92.03   10.36     0.00
## pct.hs.grad.3         0.00       0.00   -2.64     0.01
## log.pct.bach.deg   6195.10     579.20   10.70     0.00
## log.pct.below.pov -3487.25     291.18  -11.98     0.00
## log.pct.unemp       569.76     498.26    1.14     0.25
## stateAR            -759.37    1275.96   -0.60     0.55
## stateAZ            -321.27     976.99   -0.33     0.74
## stateCA            1928.36     702.25    2.75     0.01
## stateCO             488.91     832.60    0.59     0.56
## stateCT            1576.99     874.64    1.80     0.07
## stateDC            1225.97    1752.52    0.70     0.48
## stateDE             399.89    1290.40    0.31     0.76
## stateFL            -553.01     694.96   -0.80     0.43
## stateGA             435.54     820.43    0.53     0.60
## stateHI             687.60    1141.48    0.60     0.55
## stateID            -168.60    1722.38   -0.10     0.92
## stateIL             975.67     745.15    1.31     0.19
## stateIN             -46.33     760.62   -0.06     0.95
## stateKS             306.97    1032.14    0.30     0.77
```

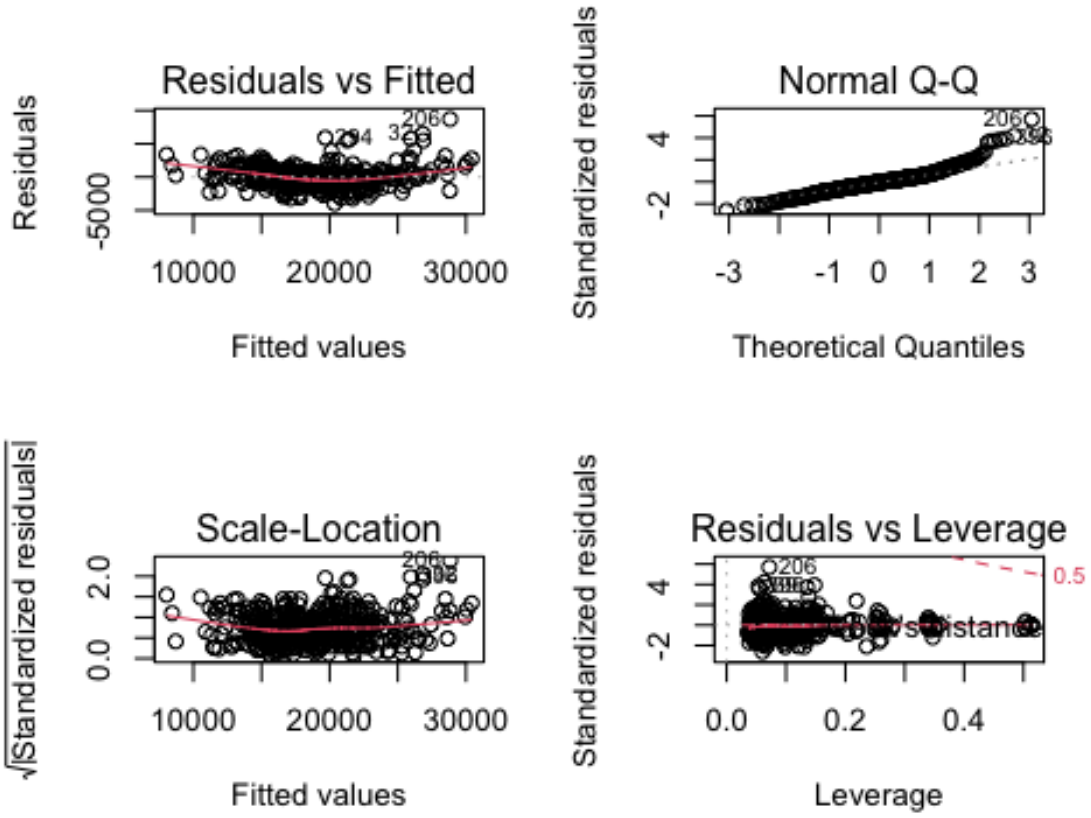
```
## stateKY            -370.07    1110.16   -0.33     0.74
## stateLA            -187.40     808.98   -0.23     0.82
## stateMA             624.35     855.14    0.73     0.47
## stateMD             207.01     832.51    0.25     0.80
## stateME             181.52     954.10    0.19     0.85
## stateMI            1338.13     768.06    1.74     0.08
## stateMN             -97.03     891.51   -0.11     0.91
## stateMO              58.94     847.99    0.07     0.94
## stateMS            -915.79    1101.70   -0.83     0.41
## stateMT             431.82    1721.10    0.25     0.80
## stateNC             170.78     727.32    0.23     0.81
## stateND              50.69    1747.23    0.03     0.98
## stateNE            -919.37    1171.02   -0.79     0.43
## stateNH             -63.26    1050.33   -0.06     0.95
## stateNJ            2154.89     768.88    2.80     0.01
## stateNM           -1455.18    1284.06   -1.13     0.26
## stateNV            4176.17    1327.31    3.15     0.00
## stateNY             542.07     716.02    0.76     0.45
## stateOH             338.22     714.83    0.47     0.64
## stateOK            -789.18    1015.31   -0.78     0.44
## stateOR           -1087.71     917.40   -1.19     0.24
## statePA            -277.69     703.22   -0.39     0.69
## stateRI           -1991.33    1147.52   -1.74     0.08
## stateSC            -202.51     771.99   -0.26     0.79
## stateSD             327.86    1744.72    0.19     0.85
## stateTN            -349.73     826.94   -0.42     0.67
## stateTX             -25.91     673.53   -0.04     0.97
## stateUT           -4060.76    1047.34   -3.88     0.00
## stateVA             749.06     882.63    0.85     0.40
## stateVT            -750.54    1715.64   -0.44     0.66
## stateWA            -160.13     832.45   -0.19     0.85
## stateWI             205.27     810.20    0.25     0.80
## stateWV            -546.25    1701.69   -0.32     0.75

summary(cdi.subsets.final.model.with.some.region)

##
## Call:
## lm(formula = per.cap.income ~ log.land.area + log.pop.18_34 +
##     log.doctors + pct.hs.grad.3 + log.pct.bach.deg + log.pct.below.pov +
##     log.pct.unemp + region + log.pop.18_34:region + pct.hs.grad.3:region +

##     log.pct.bach.deg:region + log.pct.below.pov:region, data = cdi_all_sub
sets)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4787.7  -982.3  -135.6   714.8  9626.3
##
## Coefficients:
```

```
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.360e+04  4.645e+03   7.235 2.25e-12 ***
## log.land.area                -7.042e+02  1.160e+02  -6.070 2.88e-09 ***
## log.pop.18_34                -7.613e+03  1.676e+03  -4.543 7.28e-06 ***
## log.doctors                   9.303e+02  9.476e+01   9.818  < 2e-16 ***
## pct.hs.grad.3                -1.494e-03  3.493e-03  -0.428 0.669130
## log.pct.bach.deg              4.676e+03  1.022e+03   4.574 6.33e-06 ***
## log.pct.below.pov            -3.094e+03  5.243e+02  -5.901 7.49e-09 ***
## log.pct.unemp                 1.264e+03  3.599e+02   3.513 0.000491 ***
## regionNE                      6.507e+03  6.535e+03   0.996 0.319934
## regionS                      -7.750e+03  5.118e+03  -1.514 0.130705
## regionW                       8.801e+02  7.303e+03   0.120 0.904145
## log.pop.18_34:regionNE       -4.386e+03  2.371e+03  -1.849 0.065105 .
## log.pop.18_34:regionS         1.143e+03  1.926e+03   0.594 0.553079
## log.pop.18_34:regionW         2.167e+03  2.495e+03   0.869 0.385500
## pct.hs.grad.3:regionNE       -6.806e-03  4.468e-03  -1.523 0.128469
## pct.hs.grad.3:regionS        -4.696e-03  4.207e-03  -1.116 0.264941
## pct.hs.grad.3:regionW        -1.791e-02  4.387e-03  -4.082 5.34e-05 ***
## log.pct.bach.deg:regionNE     4.292e+03  1.285e+03   3.340 0.000913 ***
## log.pct.bach.deg:regionS      2.482e+03  1.147e+03   2.164 0.030997 *
## log.pct.bach.deg:regionW      3.229e+03  1.374e+03   2.350 0.019231 *
## log.pct.below.pov:regionNE -7.567e+02  7.117e+02  -1.063 0.288320
## log.pct.below.pov:regionS  -8.122e+02  6.451e+02  -1.259 0.208756
## log.pct.below.pov:regionW  -3.949e+03  9.132e+02  -4.324 1.92e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1635 on 417 degrees of freedom
## Multiple R-squared:  0.846,  Adjusted R-squared:  0.8378
## F-statistic: 104.1 on 22 and 417 DF,  p-value: < 2.2e-16

round(summary(cdi.subsets.final.model.with.some.region)$coef,2)

##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                33603.04    4644.53    7.23     0.00
## log.land.area               -704.21     116.01   -6.07     0.00
## log.pop.18_34              -7613.16    1675.90   -4.54     0.00
## log.doctors                  930.31      94.76    9.82     0.00
## pct.hs.grad.3                  0.00       0.00   -0.43     0.67
## log.pct.bach.deg            4676.09    1022.42    4.57     0.00
## log.pct.below.pov          -3094.10     524.35   -5.90     0.00
## log.pct.unemp               1264.35     359.89    3.51     0.00
## regionNE                    6507.45    6535.02    1.00     0.32
## regionS                    -7750.28    5118.03   -1.51     0.13
## regionW                      880.06    7303.42    0.12     0.90
## log.pop.18_34:regionNE     -4385.78    2371.44   -1.85     0.07
## log.pop.18_34:regionS       1143.33    1925.98    0.59     0.55
## log.pop.18_34:regionW       2167.35    2494.87    0.87     0.39
## pct.hs.grad.3:regionNE        -0.01       0.00   -1.52     0.13
## pct.hs.grad.3:regionS          0.00       0.00   -1.12     0.26
```

```
## pct.hs.grad.3:regionW          -0.02       0.00    -4.08     0.00
## log.pct.bach.deg:regionNE    4292.42    1285.09     3.34     0.00
## log.pct.bach.deg:regionS     2481.75    1146.60     2.16     0.03
## log.pct.bach.deg:regionW     3229.20    1374.04     2.35     0.02
## log.pct.below.pov:regionNE   -756.68     711.72    -1.06     0.29
## log.pct.below.pov:regionS    -812.19     645.14    -1.26     0.21
## log.pct.below.pov:regionW   -3948.57     913.22    -4.32     0.00

par(mfrow=c(2,2))
plot(cdi.subsets.final.model.with.some.region)
```




```
plot(final.model.with.state)

## Warning: not plotting observations with leverage one:
##   73, 232, 233, 339, 356, 388, 429
```




```
vif(final.model.with.state)

##                       GVIF Df GVIF^(1/(2*Df))
## log.land.area     2.206895  1        1.485562
## log.pop.18_34     1.956279  1        1.398671
## log.doctors       1.936019  1        1.391409
## pct.hs.grad.3     7.181496  1        2.679831
## log.pct.bach.deg  7.361002  1        2.713117
## log.pct.below.pov 4.167370  1        2.041414
## log.pct.unemp     4.615294  1        2.148323
## state            32.180767 47        1.037620

vif(cdi.subsets.final.model.with.some.region)

##                                 GVIF Df GVIF^(1/(2*Df))
## log.land.area           1.680366e+00  1        1.296289
## log.pop.18_34           9.161875e+00  1        3.026859
## log.doctors             1.931049e+00  1        1.389622
## pct.hs.grad.3           2.948870e+01  1        5.430350
## log.pct.bach.deg        2.157800e+01  1        4.645212
## log.pct.below.pov       1.271309e+01  1        3.565542
## log.pct.unemp           2.265215e+00  1        1.505063
## region                  4.593495e+08  3       27.777333
## log.pop.18_34:region    1.183970e+09  3       32.525461
```

```
## pct.hs.grad.3:region     6.001590e+05  3        9.184265
## log.pct.bach.deg:region  2.276982e+07  3       16.835523
## log.pct.below.pov:region 1.275326e+05  3        7.094747
```

```
anova(cdi.subsets.final.model, cdi.subsets.final.model.with.some.region, fina
l.model.with.state)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ log(land.area) + log(pct.unemp) + log(pct.below.
pov) +
##     log(pct.bach.deg) + (pct.hs.grad)^3 + log(doctors) + log(pop.18_34)
## Model 2: per.cap.income ~ log.land.area + log.pop.18_34 + log.doctors +
##     pct.hs.grad.3 + log.pct.bach.deg + log.pct.below.pov + log.pct.unemp +

##     region + log.pop.18_34:region + pct.hs.grad.3:region + log.pct.bach.de
g:region +
##     log.pct.below.pov:region
## Model 3: per.cap.income ~ log.land.area + log.pop.18_34 + log.doctors +
##     pct.hs.grad.3 + log.pct.bach.deg + log.pct.below.pov + log.pct.unemp +

##     state
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    432 1314204306
## 2    417 1114159243 15 200045063 5.3058 1.037e-09 ***
## 3    385  967714409 32 146444834 1.8207  0.005001 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(cdi.subsets.final.model, cdi.subsets.final.model.with.some.region, final.
model.with.state)
```

```
##                                          df      AIC
## cdi.subsets.final.model                   9 7826.944
## cdi.subsets.final.model.with.some.region 24 7784.286
## final.model.with.state                   56 7786.282
```

```
BIC(cdi.subsets.final.model, cdi.subsets.final.model.with.some.region, final.
model.with.state)
```

```
##                                          df      BIC
## cdi.subsets.final.model                   9 7863.725
## cdi.subsets.final.model.with.some.region 24 7882.369
## final.model.with.state                   56 8015.141
```

```
state_county <- cdi %>%

distinct(state, county) %>%

group_by(state) %>%

summarize("county in state" = n())
```

```
state_county
```

Table: Summary of County in State in CDI

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

```
detach(cdi)
```