Predicting Per Capita Income from County Demographic Information

Maxine Graves
Department of Statistics, Carnegie Mellon University
mgraves@andrew.cmu.edu

Abstract
Four research questions pertaining to the prediction of per capita income are asked (1. Are there any relationships between variables included in dataset? Are these relationships expected? 2. Do crime, per capita crime, and region have any influence on per capita income? 3. Of variables included, what is the best model to predict per capita income? 4. Is it a cause for concern that not all counties and states are included in dataset?). Data used comes from Kutner et al. (2005) and includes information on 17 different variables for the 440 counties in the United States with the greatest populations. Methods used to answer research questions include exploratory data analysis, simple linear regression, all subsets regression, and stepwise regression. Results found that 1. Total income vs number of hospital beds and total income vs crimes were correlated unexpectedly, 2. Although there appears to be a (questionable) relationship between per capita income, log(crime), and region, this model cannot be used to draw reliable conclusions between variables of interest, 3. Per capita crime predicted by log(land area), percent of population between 18 and 34, log(doctors), percent of those 25+ with 12 or more years of education, percent of those 25+ with a bachelor's degree, percent of those below the poverty level, and percent unemployment was the "best" model, 4. Depending on perspective, missing states and counties may be cause for concern. Overall, main limitations included criteria prioritization used to determine "best" model for predicting per capita income and dataset used.

Introduction
Given that the way in which the United States is set-up socially, politically, and economically, it is no surprise that personal income and its predictors are a topic of interest among many disparate spheres. As a result, it is also no surprise that there is a large body of research addressing this facet of American life and livelihood. By way of adding to this already broad area of study, the present paper seeks to answer four questions pertaining to per capita income in the United States. These questions are presented below.

1. What relationships are found between variables included in the present dataset and are these relationships expected?
2. How do crime, per capita crime, and geographic region (exclusively) relate to per capita income? Further, does it matter whether total crime in a county or per capita crime is used to predict per capita income?
3. Based on the variables in the dataset, what is the best way to model predictions of per capita income?
4. Are counties and states missing from the dataset a cause for concern?

Data
Data used is known as county demographic information. This particular dataset tracks 17 different variables for the 440 United States counties with the largest populations. Data originally comes from Kutner et al. (2005). A breakdown of the variables can be found in Table 1.

Variable	Description
id	County ID number 1-440
county	Name of county
state	Abbreviation of state name
land.area	Land area (square miles)
pop	Estimated population (1990)
pop.18_34	Estimated percent of population between 18 and 34 years old (1990)
pop.65_plus	Estimated percent of population 65 years or older (1990)
doctors	Number of nonfederal practicing physicians (1990)
hosp.beds	Includes hospital beds, bassinets, and cribs (1990)
crimes	Total number of serious crimes (i.e. murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft) in 1990
pct.hs.grad	Percent of population 25 years or older with 12 or more years of schooling
pct.bach.deg	Percent of population 25 years or older with a bachelor's degree
pct.below.pov	Estimated percent of population below poverty level based on income (1990)
pct.unemp	Estimated percent of population that is unemployed (1990)
per.cap.income	Estimated per capita income of population in dollars (1990)
tot.income	Estimated total income of population in millions of dollars (1990)
region	Geographic region (NE: northeast, NC: north-central, S: southern, W: western)

Table 1 Breakdown of variables and definitions. References Kutner et al. (2005). *Original Source*: Geospatial and Statistical Data Center, University of Virginia.

By way of a brief overview of data, Figure 1 shows numeric summaries of each variable included. For a more detailed overview, including frequency plots for state and region, and histograms of all other variables, see Section 1 of the code appendix.

```
       id              county        state       land.area            pop             pop.18_34       pop.65_plus        doctors
 Min.   :  1.0   Jefferson :  7   CA     : 34   Min.   :   15.0   Min.   : 100043   Min.   :16.40   Min.   : 3.000   Min.   :   39.0
 1st Qu.:110.8   Montgomery:  6   FL     : 29   1st Qu.:  451.2   1st Qu.: 139027   1st Qu.:26.20   1st Qu.: 9.875   1st Qu.:  182.8
 Median :220.5   Washington:  5   PA     : 29   Median :  656.5   Median : 217280   Median :28.10   Median :11.750   Median :  401.0
 Mean   :220.5   Cumberland:  4   TX     : 28   Mean   : 1041.4   Mean   : 393011   Mean   :28.57   Mean   :12.170   Mean   :  988.0
 3rd Qu.:330.2   Jackson   :  4   OH     : 24   3rd Qu.:  946.8   3rd Qu.: 436064   3rd Qu.:30.02   3rd Qu.:13.625   3rd Qu.: 1036.0
 Max.   :440.0   Lake      :  4   NY     : 22   Max.   :20062.0   Max.   :8863164   Max.   :49.70   Max.   :33.800   Max.   :23677.0
                 (Other)   :410   (Other):274
   hosp.beds           crimes         pct.hs.grad     pct.bach.deg   pct.below.pov       pct.unemp       per.cap.income    tot.income
 Min.   :   92.0   Min.   :   563   Min.   :46.60   Min.   : 8.10   Min.   : 1.400   Min.   : 2.200   Min.   : 8899   Min.   :  1141
 1st Qu.:  390.8   1st Qu.:  6220   1st Qu.:73.88   1st Qu.:15.28   1st Qu.: 5.300   1st Qu.: 5.100   1st Qu.:16118   1st Qu.:  2311
 Median :  755.0   Median : 11820   Median :77.70   Median :19.70   Median : 7.900   Median : 6.200   Median :17759   Median :  3857
 Mean   : 1458.6   Mean   : 27112   Mean   :77.56   Mean   :21.08   Mean   : 8.721   Mean   : 6.597   Mean   :18561   Mean   :  7869
 3rd Qu.: 1575.8   3rd Qu.: 26280   3rd Qu.:82.40   3rd Qu.:25.32   3rd Qu.:10.900   3rd Qu.: 7.500   3rd Qu.:20270   3rd Qu.:  8654
 Max.   :27700.0   Max.   :688936   Max.   :92.90   Max.   :52.30   Max.   :36.300   Max.   :21.300   Max.   :37541   Max.   :184230

 region
 NC:108
 NE:103
 S :152
 W : 77
```

Figure 1 Summaries of variables included in main, untransformed dataset.

Further, of notable interest are the histograms of land area, population, doctors, hospital beds, crimes, and total income. All six of these variables seem to have a significant right skew, as can be seen in Figure 2.

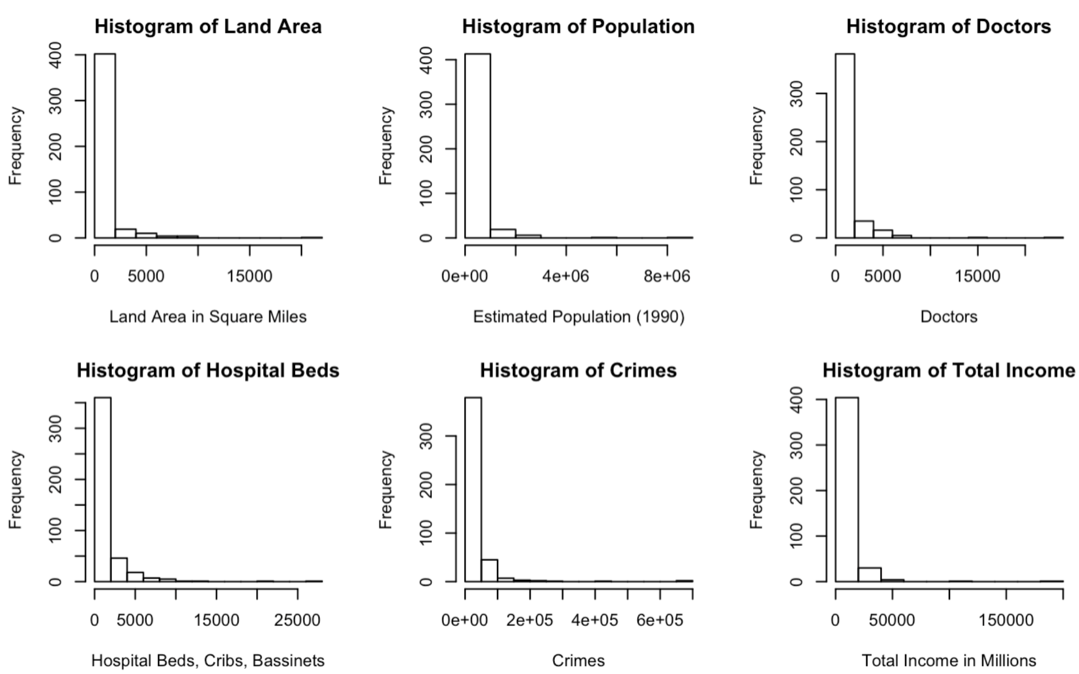


Figure 2 Variables with visually significant right skew.

In addition to the above, pairs plots of all variables were also created to check correlation (see Section 1 of code appendix).

Methods

This section will be broken down by methods used to answer each of the research questions specified above.

1. **What relationships are found between variables included in the present dataset and are these relationships expected?**
   A pairs plot of all variables included in the data set was visually inspected for possible correlation.
   Variables used: all
2. **How do crime, per capita crime, and geographic region (exclusively) relate to per capita income? Further, does it matter whether total crime in a county or per capita crime is used to predict per capita income?**
   Six linear regression models were fit. The summaries of each model, along with their diagnostic plots were assessed to determine whether there is evidence to suggest a relationship between per capita income and (per capita) crime and region.
   Variables used: per.cap.income, per.cap.crime, crime, log(crime), and region
3. **Based on the variables in the dataset, what is the best way to model predictions of per capita income?**
   Two linear models were chosen via all-subsets and stepwise regression (on transformed data): a simple model and a model that included interaction terms. These models were compared based on Bayes Information Criterion (BIC) value, R squared value, diagnostic plots, and ease of understanding.
   Variables used:
   - Simple Model (all-subsets): per.cap.income, log(land.area), pop.18-34, log(doctors), pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp
   - Interaction Model (stepwise): per.cap.income, log(land.area), pop.18-34, log(doctors), pct.hs.grad, pct.bach.deg, pct.below.pov, region, pct.hs.grad:region, pct.bach.deg:region, pct.below.pov:region, pct.bach.deg:pct.below.pov
4. **Are counties and states missing from the dataset a cause for concern?**
   Simple EDA and critical thinking based on understanding of relevant concepts.
   Variables used: region, state

Results

Beginning with the first question posed in the introduction (What relationships are found between variables included in the present dataset and are these relationships expected?), there does appear to be relationships between some variables included in the dataset. While a full breakdown of scatterplots showing the correlation between variables can be found in Figure 3, of plots that showed a possible relationship between the variables graphed, total income vs hospital beds and total income vs crimes seemed the most unexpected. As there is no immediate connection between an increase in total income and an increase in hospital beds and crimes, it seems possible there is a confounding variable that affects the correlation between the former variable and the two latter variables. More information can be found in Section 1 of the code appendix.

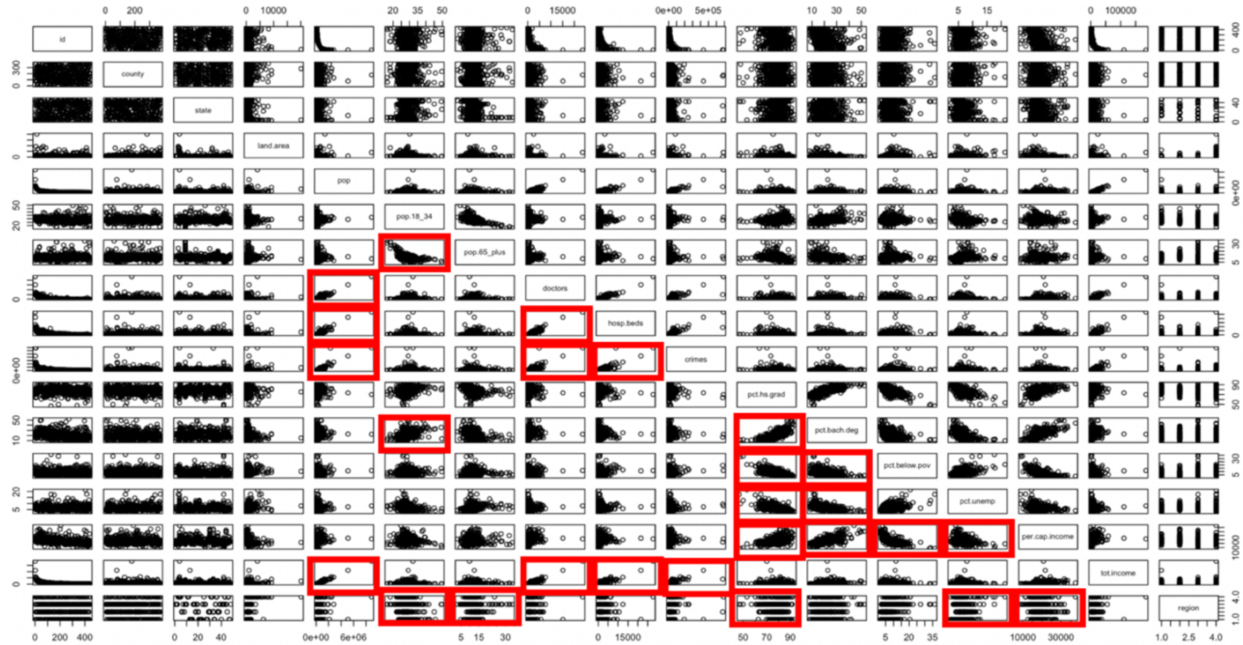


Figure 3 Pairs plots of all variables included in data set. Plots that denote a relationship between variables are highlighted in red.

Turning next to the second question presented in the introduction (How do crime, per capita crime, and geographic region (exclusively) relate to per capita income? Further, does it matter whether total crime in a county or per capita crime is used to predict per capita income?), it seems pertinent to first address the differences between two of the possible predictor variables, crime and per capita crime. Given that the response variable in question is per capita income, it makes intuitive sense that one would choose per capita crime as a predictor of crime, as opposed to total crime. This being said, two models were fit using per capita crime and region to predict per capita income. The first of these models took only per capita crime and region as predictors of per capita income, while the second model included an interaction term between per capita crime and region. Neither of the models fit showed per capita crime as a significant predictor of per capita income. Similarly, both models showed inconclusive results regarding region as an explanatory variable of per capita income, with only one region (northeast) being significant in either model. Looking more specifically at the second model, the interaction term between the two explanatory variables included in this model does not appear significant in predicting per capita income. Put more succinctly, models fit do not provide any strong evidence that per capita crime, region, or an interaction term between the two, are significant in predicting per capita income.

Note, a simple model and interaction model predicting per capita income via crime and region were also fit using log(crime) to address right skew in crime data. It was found that the simple model fit using this transformed data produced significant results for log(crime) and three of the four regions. Although this model produced significant results, based on the above discussion of total crime vs per capita crime, the results obtained do not change the final analysis regarding whether a meaningful significant relationship exists between per capita income, crime, and region. More information can be found in Section 2 of the code appendix.

Continuing to the third question of interest (Based on the variables in the dataset, what is the best way to model predictions of per capita income?), as noted in the methods section, there were two main contenders in the search for the best model for predicting per capita income. It is important to note that both models were fit on a modified version of the original dataset. This modified version did not include the following variables from the original dataset: id, county, state, pop, tot.income, or crimes. Further the modified version log transformed land.area, doctors, and hosp.beds and used these transformed variables in place of their untransformed counterparts. See Section 3 of the code appendix for more information on modified dataset. Taking the above into account, the first model used an all-subsets regression that took as its starting point all variables in the modified dataset. The equation for this model is output below.

$$
\begin{aligned}
&expected\ per.cap.income \\
&\quad = 28748.6 - 683.89 * land.area - 300.39 * pop.18\_34 + 1000.9 * doctors - 116.8039 \\
&\quad * pct.hs.grad + 371.01 * pct.bach.\mathrm{deg} - 427.27 * pct.below.pov + 251.44 * pct.unemp
\end{aligned}
$$

Similarly, the second model also took the modified dataset as its starting point, however included interaction terms between all variables and region, interaction terms between pct.hs.grad and pct.below.pov and pct.bach.deg and pct.below.pov, and was fit using stepwise regression (based on Bayes Information Criterion). The ensuing model equation is shown below.

$$
\begin{aligned}
&expected\ per.cap.income \\
&\quad = 30493.95 - 653.54 * land.area - 291.41 * pop.18\_34 + 979.17 * doctors - 106.827 \\
&\quad * pct.hs.grad + 321.73 \\
&\quad * pct.bach.\mathrm{deg} - 260.61 * pct.below.pov + 707.388 * regionNE - 9533.428 * regionS \\
&\quad + 20392.519 * regionW - 54.776 * pct.hs.grad{:}regionNE + 92.87 \\
&\quad * pct.hs.grad{:}regionS - 283.267 * pct.hs.grad{:}regionW + 187.64 \\
&\quad * pct.bach.deg{:}regionNE + 26.89 * pct.bach.deg{:}regionS + 201.07 \\
&\quad * pct.bach.deg{:}regionW - 29.57 * pct.below.pov{:}regionNE + 161.097 \\
&\quad * pct.below.pov{:}regionS - 219.63 * pct.below.pov{:}regionW - 9.59 \\
&\quad * pct.bach.deg{:}pct.below.pov
\end{aligned}
$$

While the model including interactions had a significantly better BIC score, it was ultimately decided that the simpler model that did not include any interaction terms was best. This decision was based on two main findings, namely that the more complex model only increased R squared by 2.85% and that the simpler model was more elegant and easily interpretable. Given the former finding, it seemed unwise to add unnecessary complexity to the final model, in turn leading to the importance of the latter finding (see Section 4 of the code appendix for more information on the above, as well as diagnostic plots for both models). One can find a breakdown of how to interpret the results of the chosen model in Table 2.

Variable	**Interpretation**
Intercept	Given a county with 0% land area, 0% population between the ages of 18 and 34, 0% doctors, 0% people 25 years or older with more than 12 years of education, 0% people 25 years or older with a bachelor's degree, 0% of people under the poverty level, and 0%

	unemployment, per capita income is expected to be $28,748.60.
land.area	For every 1% increase in land area, one would expect a $6.84 decrease in per capita income.
pop.18_34	For every 1% increase in estimated county population between 18 and 34 years of age, one would expect a $300.39 decrease in per capita income.
doctors	For every 1% increase in doctors, one would expect a $10.01 increase in per capita income.
pct.hs.grad	For every 1% increase in those 25 years or older with 12 or more years of education (but no bachelor's degree), one would expect a $116.80 decrease in per capita income.
pct.bach.deg	For every 1% increase in those 25 years or older with a bachelor's degree, one would expect a $371.01 increase in per capita income.
pct.below.pov	For every 1% increase in those below the poverty level, one would expect a $427.27 decrease in per capita income.
pct.unemp	For every 1% increase in unemployment, one would expect a $251.44 increase in per capita income.

Table 2 Interpretations of predictors included in best model.

Finally, looking at the fourth question posed in the introduction (Are counties and states missing from the dataset a cause for concern?), the author would argue that missing states and counties may be a cause for a concern. On the one hand, one could take the stance that looking at this data from the state or county level is too granular, and that aggregated forms of this data like the region variable included in the present dataset are better equipped to act as explanatory variables. From this point of view, missing states and counties would not be a cause for concern given that each of the four regions included in the dataset have relatively large sample sizes (see Section 5 of the code appendix for more information), with which regressions could be confidently computed. On the other hand, it is possible that simply looking at the 440 most populous counties fails to account for underlying relationships that may exist between predictors if more counties and states were included in the dataset.

Discussion

By way of a quick summary of the above, answers to the four questions presented in the introduction are included below, along with the reasoning behind the approach to each answer.

1. **What relationships are found between variables included in the present dataset and are these relationships expected?**
   While a full list of all relationships found between variables included in the dataset can be seen in Figure 3, the apparent correlation between total income vs hospital beds and total income vs crimes are of special note. In comparison to the other correlations

noticed, the relationships between these variables seemed especially unexpected. The relationship noted between these variables may point to possible confounding variables. Scatterplots graphing all variables against each other were used to determine the above since this type of graph allows for easy visual assessment of correlation between variables.

2. **How do crime, per capita crime, and geographic region (exclusively) relate to per capita income? Further, does it matter whether total crime in a county or per capita crime is used to predict per capita income?**
   Via simple linear regression it was determined that there does not seem to be a significant relationship between per capita income and (exclusively) per capita crime and region; even when an interaction term is included between the two. Although a significant relationship appeared to exist when a simple model was fit with log(crime) and region, this model did not make logical sense and was therefore discarded in favor of the above. Simple linear regression was used to answer this question because this statistical approach is relatively intuitive and easy to interpret.
3. **Based on the variables in the dataset, what is the best way to model predictions of per capita income?**
   It was found through all subsets and stepwise regression that the best model for predicting per capita income included log(land area), percent of population between 18 and 34 years old, log(doctors), percent of population 25 years or older with twelve or more years of education, percent of population 25 years or older with a bachelor's degree, percent of population below poverty level, and percent of population unemployed as explanatory variables. These forms of regression were used to show the above, as they lessen computational strain and provide reliable results that can be tested and interpreted easily.
4. **Are counties and states missing from the dataset a cause for concern?**
   Depending on perspective, it is possible that missing states and counties are cause for concern. Simple EDA was used to answer this question to augment ease of interpretability and to allow for a more conceptual approach.

Some possible limitations to present research may include final model selection regarding the third research question and data used to conduct research. As observed above, the model chosen as "best" in regards to the third research question did not have the lowest BIC value of models fit. In turn, although the model picked is easily interpretable, it does lack some of the predictive ability of a more complex model. Additional research may look into fitting and interpreting more complex models to overcome this issue. Further, as previously noted, data used to conduct research only included information on the 440 most populous counties. Including data for only these counties may create an unrealistic picture of which variables can effectively predict per capita income. Future research would benefit from a larger dataset that includes information on more counties.

Bibliography

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

R Documentation. "Merge Two Data Frames." Retrieved October 18, 2021 (https://stat.ethz.ch/R-manual/R-devel/library/base/html/merge.html).

Tidyr. "Pivot data from long to wide." Retrieved October 18, 2021 (https://tidyr.tidyverse.org/reference/pivot_wider.html).

note: class materials and office hours were also used as resources

36617 Project 1 Code Appendix

Maxine Graves

```
sources:
2. https://stat.ethz.ch/R-manual/R-devel/library/base/html/merge.html
3. https://tidyr.tidyverse.org/reference/pivot_wider.html
```

```r
library(Hmisc)
```

```
## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:Hmisc':
##
##     src, summarize

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

######################## **SECTION 1** ########################

```r
cdi = read.csv("cdi.dat", header=TRUE, sep=" ")
```

```r
lapply(cdi, summary)
```

```
## $id
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0   110.8   220.5   220.5   330.2   440.0
##
## $county
##       Jefferson       Montgomery       Washington       Cumberland
```

```
##                 7                 6                 5                 4
##           Jackson              Lake             Clark          Hamilton
##                 4                 4                 3                 3
##              Kent           Madison            Marion         Middlesex
##                 3                 3                 3                 3
##            Monroe            Orange             Wayne              York
##                 3                 3                 3                 3
##             Allen               Bay            Butler           Calhoun
##                 2                 2                 2                 2
##              Clay          Davidson          Delaware           El_Paso
##                 2                 2                 2                 2
##              Erie             Essex         Fairfield           Fayette
##                 2                 2                 2                 2
##          Franklin            Greene      Hillsborough             Kings
##                 2                 2                 2                 2
##         Lancaster            Mercer          Richland         St._Clair
##                 2                 2                 2                 2
##         St._Louis           Suffolk         Winnebago               Ada
##                 2                 2                 2                 1
##             Adams             Aiken           Alachua          Alamance
##                 1                 1                 1                 1
##           Alameda            Albany  Alexandria_City         Allegheny
##                 1                 1                 1                 1
##          Anderson      Androscoggin      Anne_Arundel          Arapahoe
##                 1                 1                 1                 1
## Arlington_County          Atlantic         Baltimore    Baltimore_City
##                 1                 1                 1                 1
##        Barnstable            Beaver              Bell            Benton
##                 1                 1                 1                 1
##            Bergen             Berks         Berkshire        Bernalillo
##                 1                 1                 1                 1
##           Berrien             Bexar              Bibb             Blair
##                 1                 1                 1                 1
##             Boone           Boulder          Brazoria            Brazos
##                 1                 1                 1                 1
##           Brevard           Bristol            Broome           Broward
##                 1                 1                 1                 1
##             Brown             Bucks          Buncombe        Burlington
##                 1                 1                 1                 1
##             Butte             Caddo         Calcasieu           Cambria
##                 1                 1                 1                 1
##            Camden           Cameron           Carroll              Cass
##                 1                 1                 1                 1
##           Catawba            Centre         Champaign           Charles
##                 1                 1                 1                 1
##        Charleston         Charlotte           Chatham        Chautauqua
##                 1                 1                 1                 1
##  Chesapeake_City           Chester        Chittenden           (Other)
##                 1                 1                 1               274
##
## $state
## AL AR AZ CA CO CT DC DE FL GA HI ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC
##  7  2  5 34  9  8  1  2 29  9  3  1 17 14  4  3  9 11 10  5 18  7  8  3  1 18
## ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV
```

```
##  1  3  4 18  2  2 22 24  4  6 29  3 11  1  8 28  4  9  1 10 11  1
##
## $land.area
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0   451.2   656.5  1041.4   946.8 20062.0
##
## $pop
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  100043  139027  217280  393011  436064 8863164
##
## $pop.18_34
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.40   26.20   28.10   28.57   30.02   49.70
##
## $pop.65_plus
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   9.875  11.750  12.170  13.625  33.800
##
## $doctors
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    39.0   182.8   401.0   988.0  1036.0 23677.0
##
## $hosp.beds
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    92.0   390.8   755.0  1458.6  1575.8 27700.0
##
## $crimes
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     563    6220   11820   27112   26280  688936
##
## $pct.hs.grad
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   46.60   73.88   77.70   77.56   82.40   92.90
##
## $pct.bach.deg
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.10   15.28   19.70   21.08   25.32   52.30
##
## $pct.below.pov
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   5.300   7.900   8.721  10.900  36.300
##
## $pct.unemp
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.200   5.100   6.200   6.597   7.500  21.300
##
## $per.cap.income
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8899   16118   17759   18561   20270   37541
##
## $tot.income
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1141    2311    3857    7869    8654  184230
##
```

```
## $region
##  NC  NE   S   W
## 108 103 152  77
```

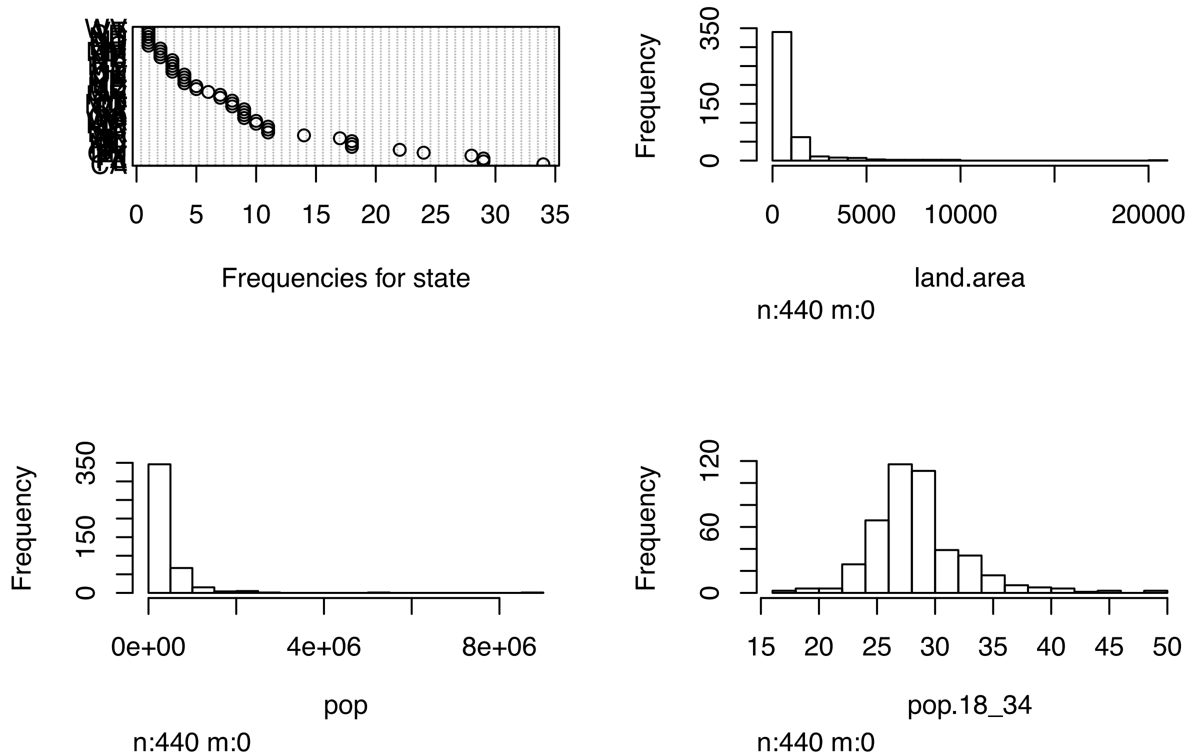
```r
which(is.na(cdi)==TRUE)
```

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

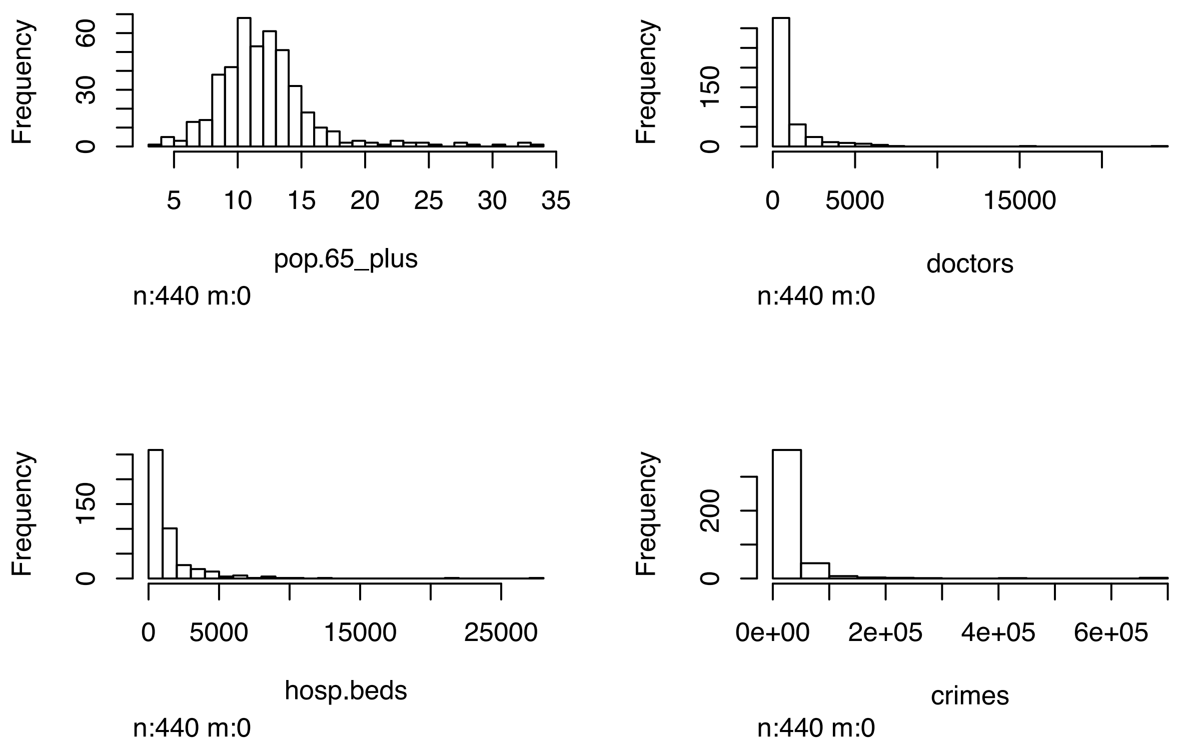
```r
#there doesn't seem to be any missing data

summary(cdi)
```

```
##        id                county        state        land.area
##  Min.   :  1.0   Jefferson :  7   CA     : 34   Min.   :   15.0
##  1st Qu.:110.8   Montgomery:  6   FL     : 29   1st Qu.:  451.2
##  Median :220.5   Washington:  5   PA     : 29   Median :  656.5
##  Mean   :220.5   Cumberland:  4   TX     : 28   Mean   : 1041.4
##  3rd Qu.:330.2   Jackson   :  4   OH     : 24   3rd Qu.:  946.8
##  Max.   :440.0   Lake      :  4   NY     : 22   Max.   :20062.0
##                  (Other)   :410   (Other):274
##       pop              pop.18_34      pop.65_plus        doctors
##  Min.   : 100043   Min.   :16.40   Min.   : 3.000   Min.   :   39.0
##  1st Qu.: 139027   1st Qu.:26.20   1st Qu.: 9.875   1st Qu.:  182.8
##  Median : 217280   Median :28.10   Median :11.750   Median :  401.0
##  Mean   : 393011   Mean   :28.57   Mean   :12.170   Mean   :  988.0
##  3rd Qu.: 436064   3rd Qu.:30.02   3rd Qu.:13.625   3rd Qu.: 1036.0
##  Max.   :8863164   Max.   :49.70   Max.   :33.800   Max.   :23677.0
##
##    hosp.beds           crimes         pct.hs.grad     pct.bach.deg
##  Min.   :   92.0   Min.   :   563   Min.   :46.60   Min.   : 8.10
##  1st Qu.:  390.8   1st Qu.:  6220   1st Qu.:73.88   1st Qu.:15.28
##  Median :  755.0   Median : 11820   Median :77.70   Median :19.70
##  Mean   : 1458.6   Mean   : 27112   Mean   :77.56   Mean   :21.08
##  3rd Qu.: 1575.8   3rd Qu.: 26280   3rd Qu.:82.40   3rd Qu.:25.32
##  Max.   :27700.0   Max.   :688936   Max.   :92.90   Max.   :52.30
##
##  pct.below.pov      pct.unemp      per.cap.income    tot.income     region
##  Min.   : 1.400   Min.   : 2.200   Min.   : 8899   Min.   :  1141   NC:108
##  1st Qu.: 5.300   1st Qu.: 5.100   1st Qu.:16118   1st Qu.:  2311   NE:103
##  Median : 7.900   Median : 6.200   Median :17759   Median :  3857   S :152
##  Mean   : 8.721   Mean   : 6.597   Mean   :18561   Mean   :  7869   W : 77
##  3rd Qu.:10.900   3rd Qu.: 7.500   3rd Qu.:20270   3rd Qu.:  8654
##  Max.   :36.300   Max.   :21.300   Max.   :37541   Max.   :184230
##
```

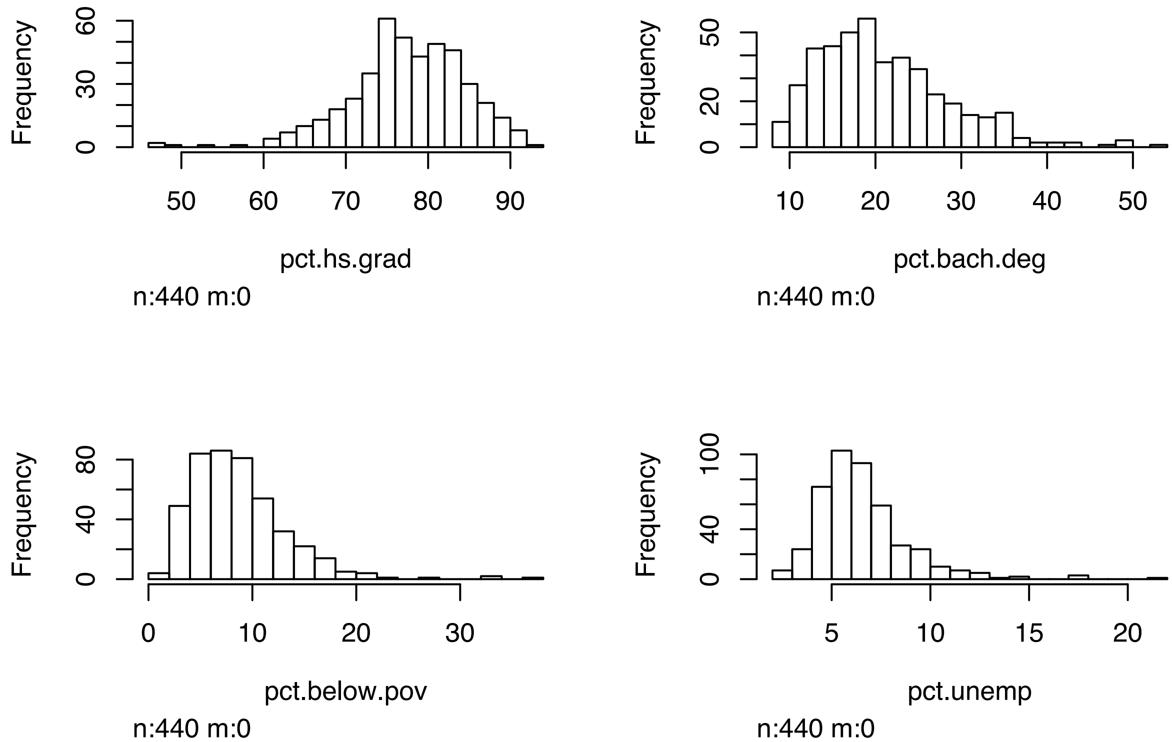
```r
hist.data.frame(cdi[,3:6])
```



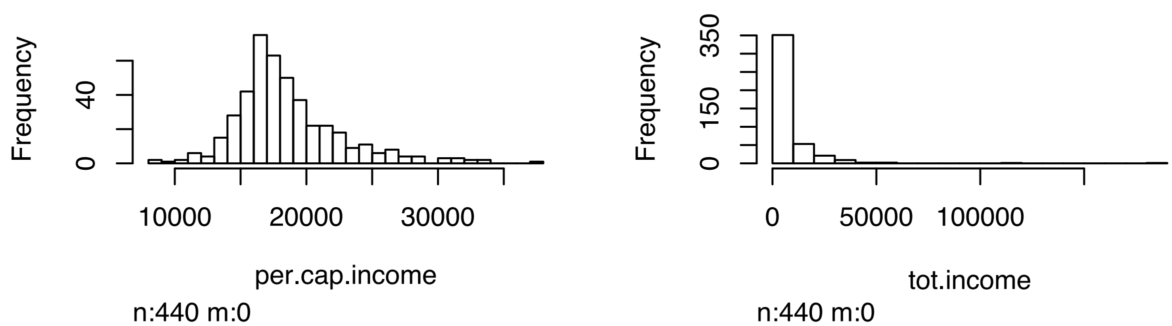

```
hist.data.frame(cdi[,7:10])
```

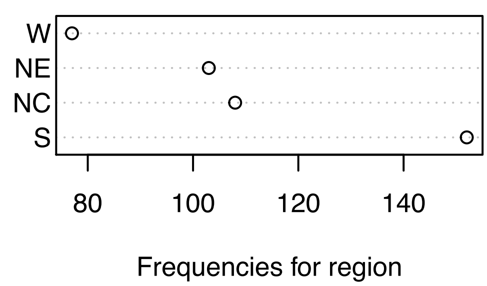



```
hist.data.frame(cdi[,11:14])
```



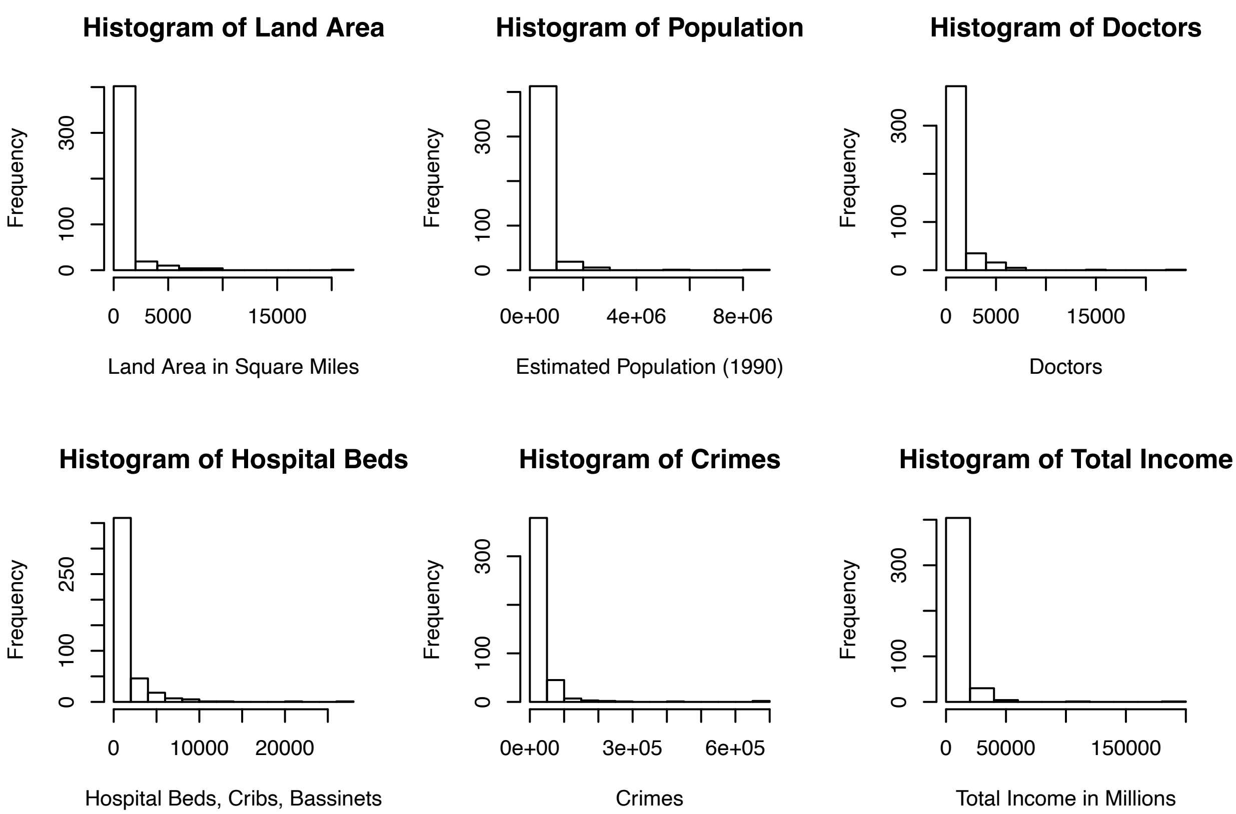

```
hist.data.frame(cdi[,15:17])
```



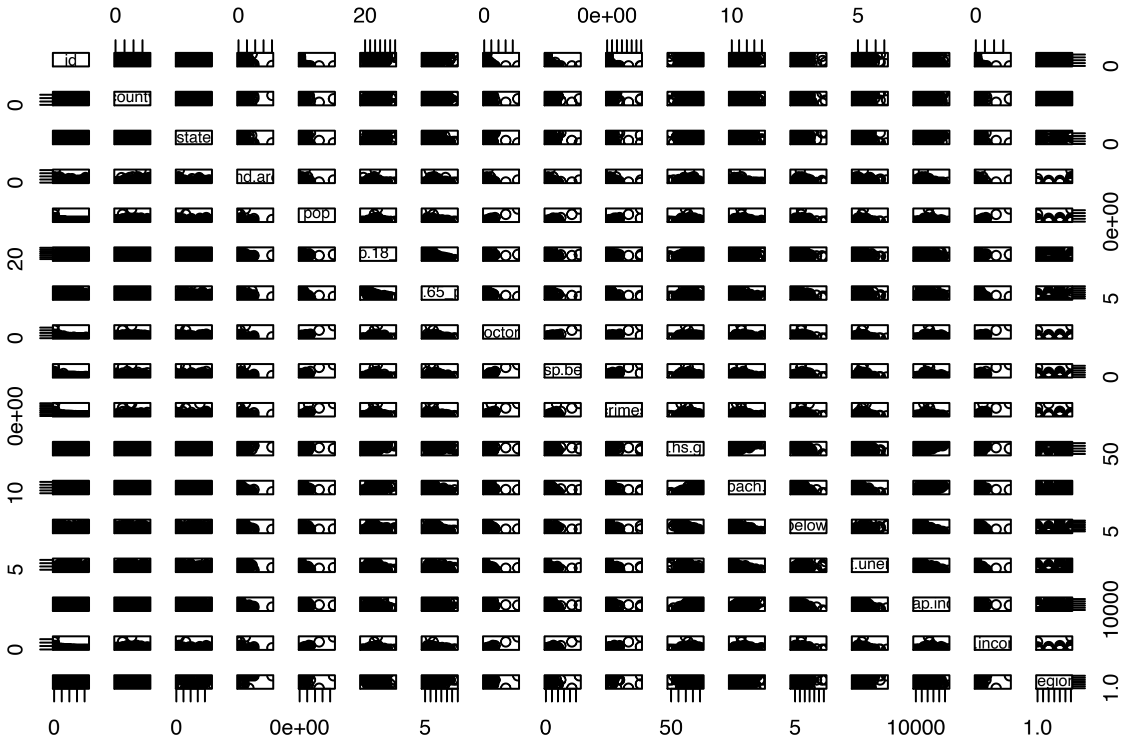



```
par(mfrow=c(2,3))
hist(cdi$land.area,
     xlab="Land Area in Square Miles",
     main="Histogram of Land Area")
hist(cdi$pop,
```

```
      xlab="Estimated Population (1990)",
      main="Histogram of Population")
hist(cdi$doctors,
      xlab="Doctors",
      main="Histogram of Doctors")
hist(cdi$hosp.beds,
      xlab="Hospital Beds, Cribs, Bassinets",
      main="Histogram of Hospital Beds")
hist(cdi$crimes,
      xlab="Crimes",
      main="Histogram of Crimes")
hist(cdi$tot.income,
      xlab="Total Income in Millions",
      main="Histogram of Total Income")
```




```
pairs(cdi)
```




```
cor(cdi$tot.income, cdi$hosp.beds)
```

```
## [1] 0.9020615
```

```
cor(cdi$tot.income, cdi$crimes)
```

```
## [1] 0.843098
```

SECTION 2

```
#untransformed data
simple_mod = lm(per.cap.income~crimes+region, data=cdi)
interaction_mod = lm(per.cap.income~crimes+region+crimes:region, data=cdi)
summary(simple_mod)
```

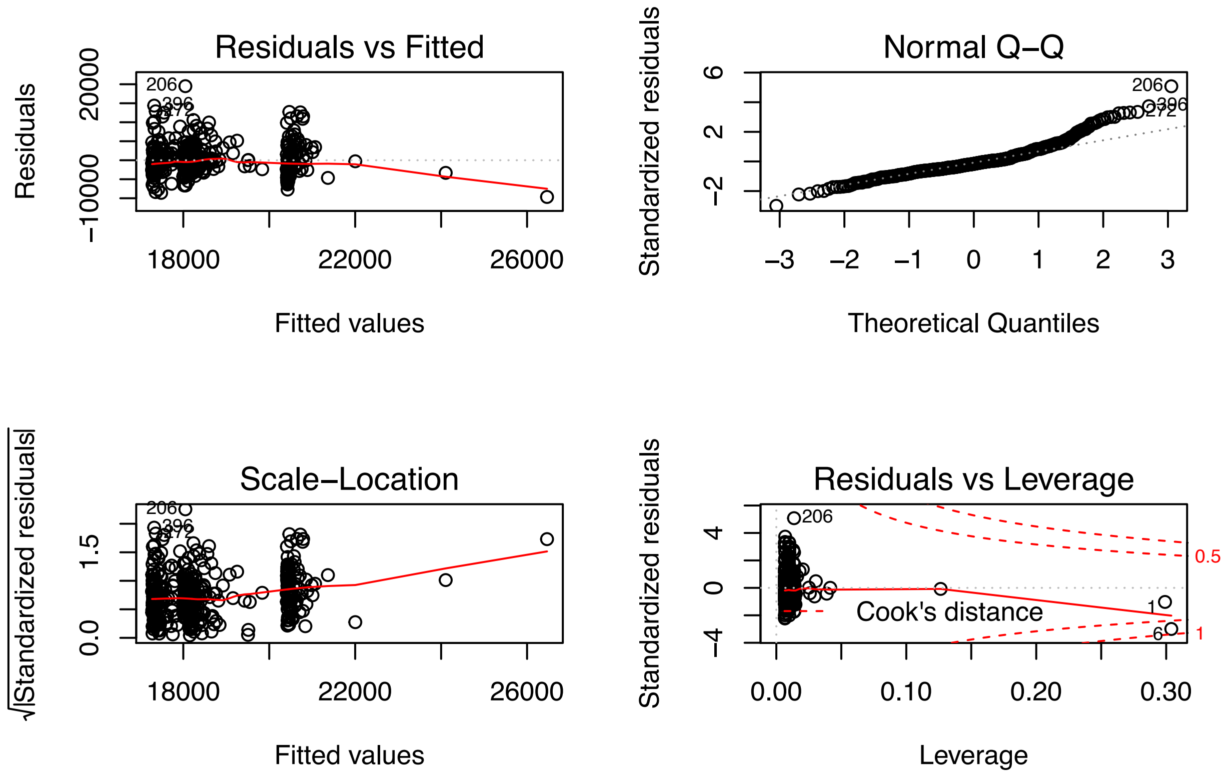
```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9661.0 -2260.7  -618.3  1650.0 19492.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.811e+04  3.784e+02  47.846  < 2e-16 ***
## crimes       8.915e-03  3.188e-03   2.797  0.00539 **
## regionNE     2.286e+03  5.325e+02   4.293 2.17e-05 ***
## regionS     -8.606e+02  4.868e+02  -1.768  0.07782 .
## regionW     -1.428e+02  5.796e+02  -0.246  0.80548
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3866 on 435 degrees of freedom
## Multiple R-squared:  0.1011, Adjusted R-squared:  0.09288
## F-statistic: 12.24 on 4 and 435 DF,  p-value: 1.946e-09
```

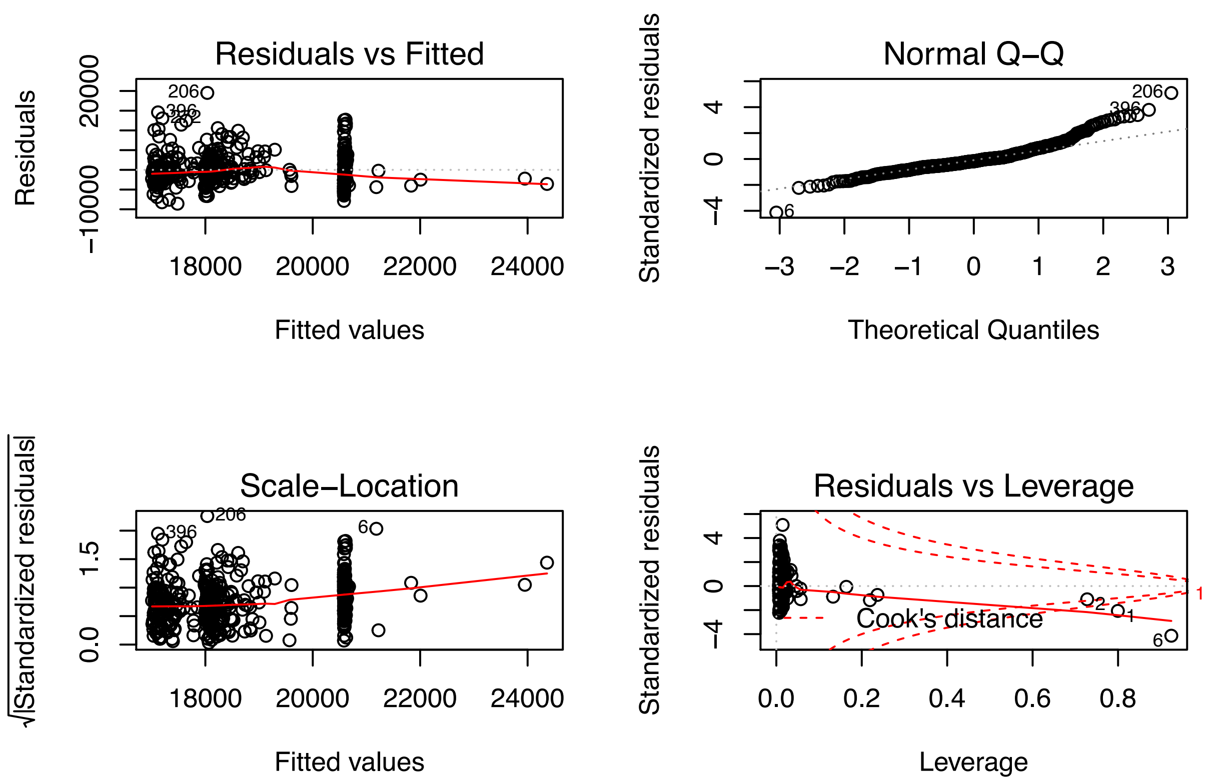
```
summary(interaction_mod) #interaction term does not seem to be significant
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region + crimes:region,
##     data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8582.4 -2225.2  -676.2  1563.4 19504.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.800e+04  4.092e+02  43.995  < 2e-16 ***
## crimes           1.361e-02  7.882e-03   1.726   0.0851 .
## regionNE         2.573e+03  5.736e+02   4.487 9.28e-06 ***
## regionS         -1.056e+03  5.606e+02  -1.884   0.0602 .
## regionW         -5.654e+01  6.372e+02  -0.089   0.9293
## crimes:regionNE -1.272e-02  9.677e-03  -1.314   0.1895
## crimes:regionS   6.348e-03  1.136e-02   0.559   0.5765
## crimes:regionW  -4.295e-03  9.486e-03  -0.453   0.6509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3861 on 432 degrees of freedom
## Multiple R-squared:  0.1099, Adjusted R-squared:  0.09543
## F-statistic: 7.616 on 7 and 432 DF,  p-value: 1.122e-08
```

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




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




```r
cdi$per.cap.crime = cdi$crimes/cdi$pop
hist(cdi$per.cap.crime)
```

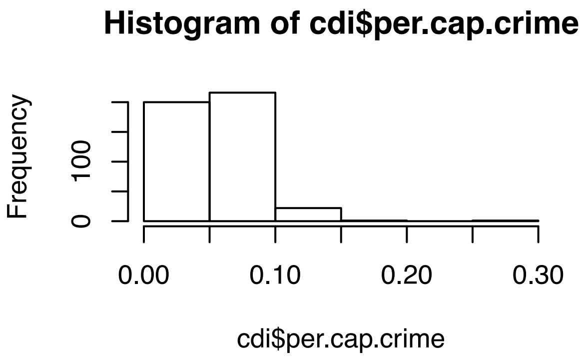
```r
pcc_mod = lm(per.cap.income~per.cap.crime+region, data=cdi)
pcc_interaction_mod = lm(per.cap.income~per.cap.crime+region+per.cap.crime:region, data=cdi)
summary(pcc_mod)
```

```
##
## 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
```

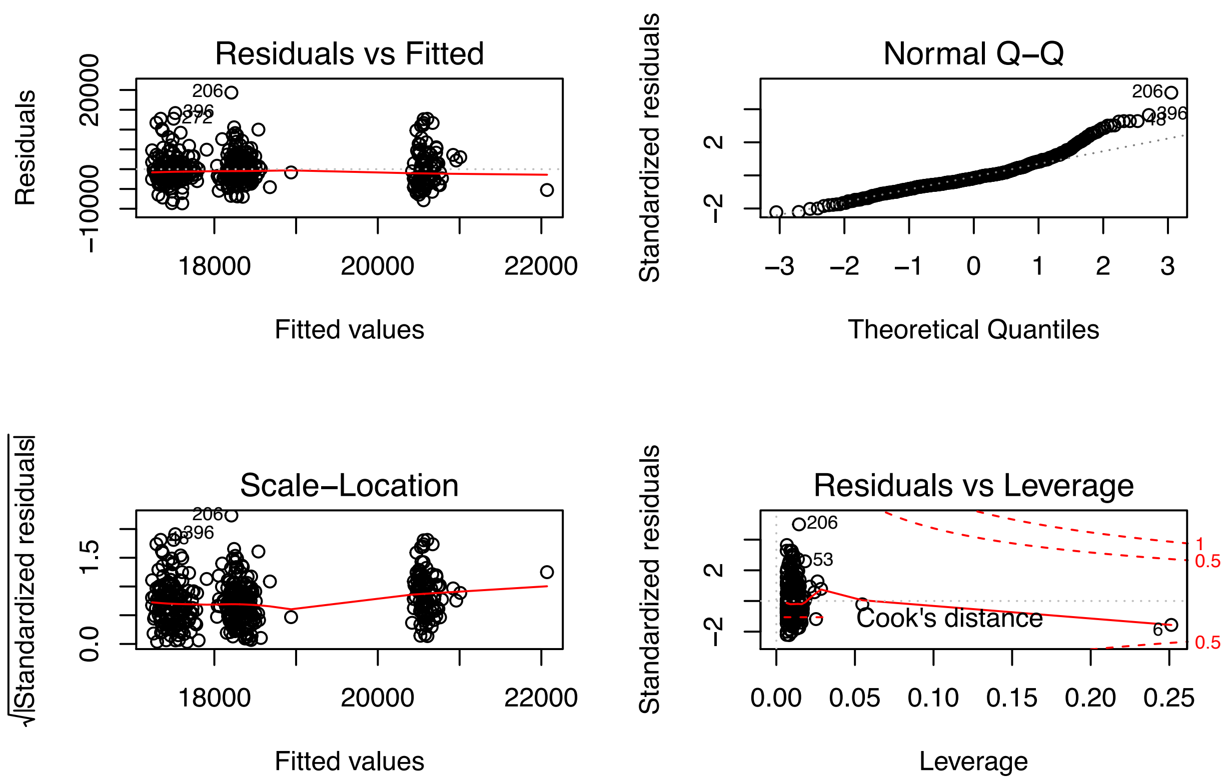
```r
summary(pcc_interaction_mod)
```

```
##
## Call:
## lm(formula = per.cap.income ~ per.cap.crime + region + 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
```

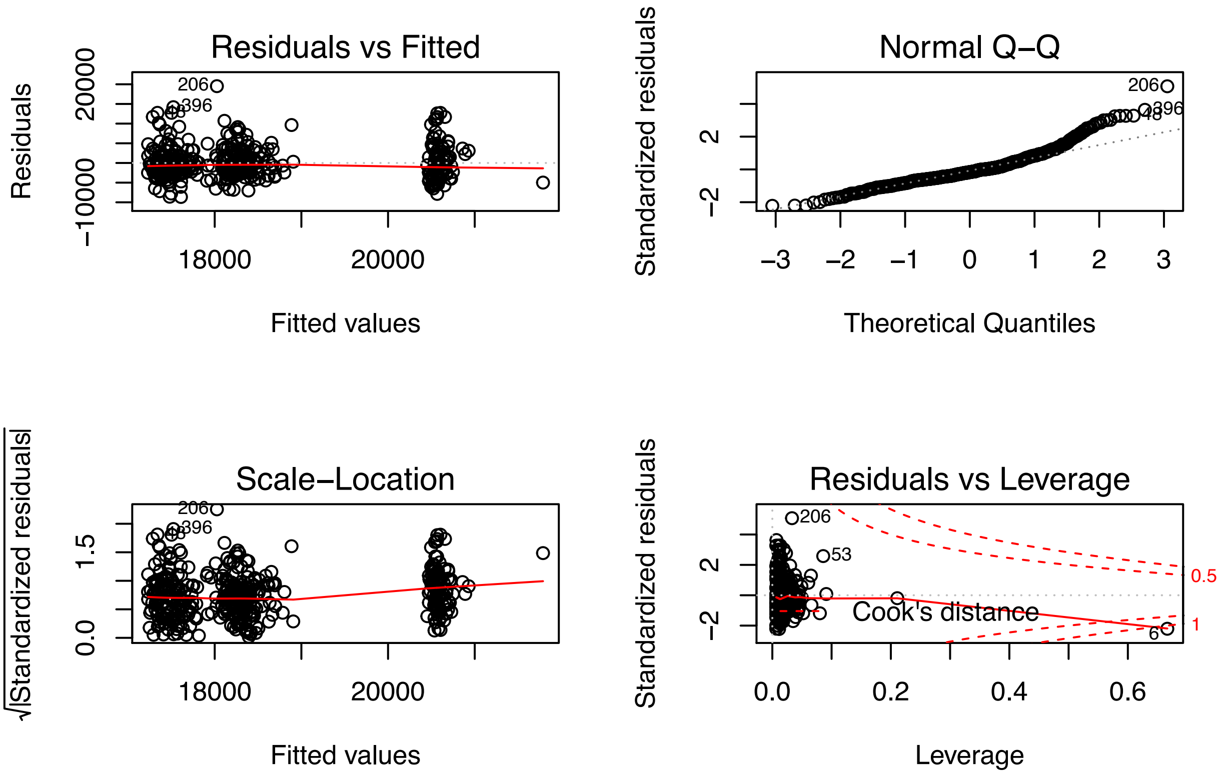
```
par(mfrow=c(2,2))
```




```
plot(pcc_mod)
```




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




```
#transformed data
cdi_transformed = cdi %>%
  mutate(land.area = log(land.area),
         pop = log(pop),
         doctors = log(doctors),
         hosp.beds = log(hosp.beds),
         crimes = log(crimes),
         tot.income = log(tot.income))

simple_mod = lm(per.cap.income~crimes+region, data=cdi_transformed)
interaction_mod = lm(per.cap.income~crimes+region+crimes:region, data=cdi_transformed)
summary(simple_mod)
```

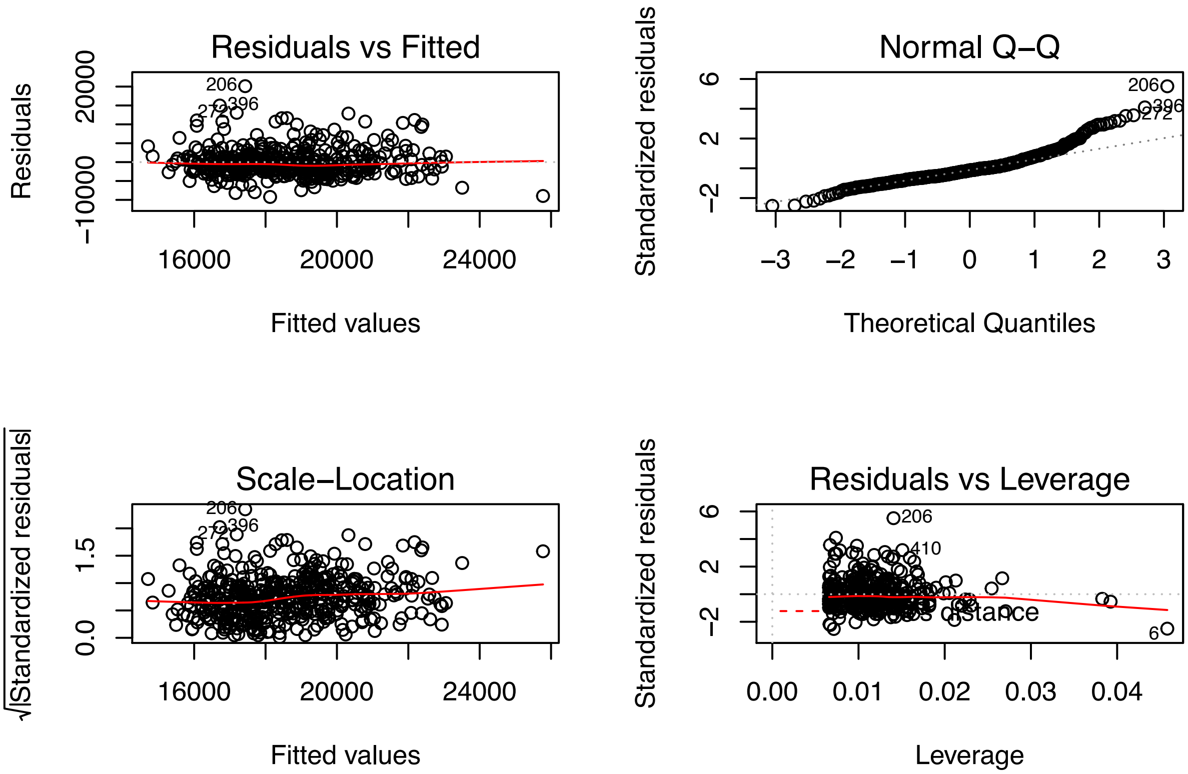
```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region, data = cdi_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9229.2 -2183.6  -502.4  1339.3 20110.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6870.8     1582.5   4.342 1.76e-05 ***
## crimes        1237.3      167.0   7.411 6.61e-13 ***
## regionNE      2284.9      506.2   4.514 8.21e-06 ***
## regionS      -1354.4      468.3  -2.892  0.00402 **
## regionW       -768.2      558.5  -1.376  0.16968
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3676 on 435 degrees of freedom
## Multiple R-squared:  0.1875, Adjusted R-squared:  0.1801
## F-statistic:  25.1 on 4 and 435 DF,  p-value: < 2.2e-16
```

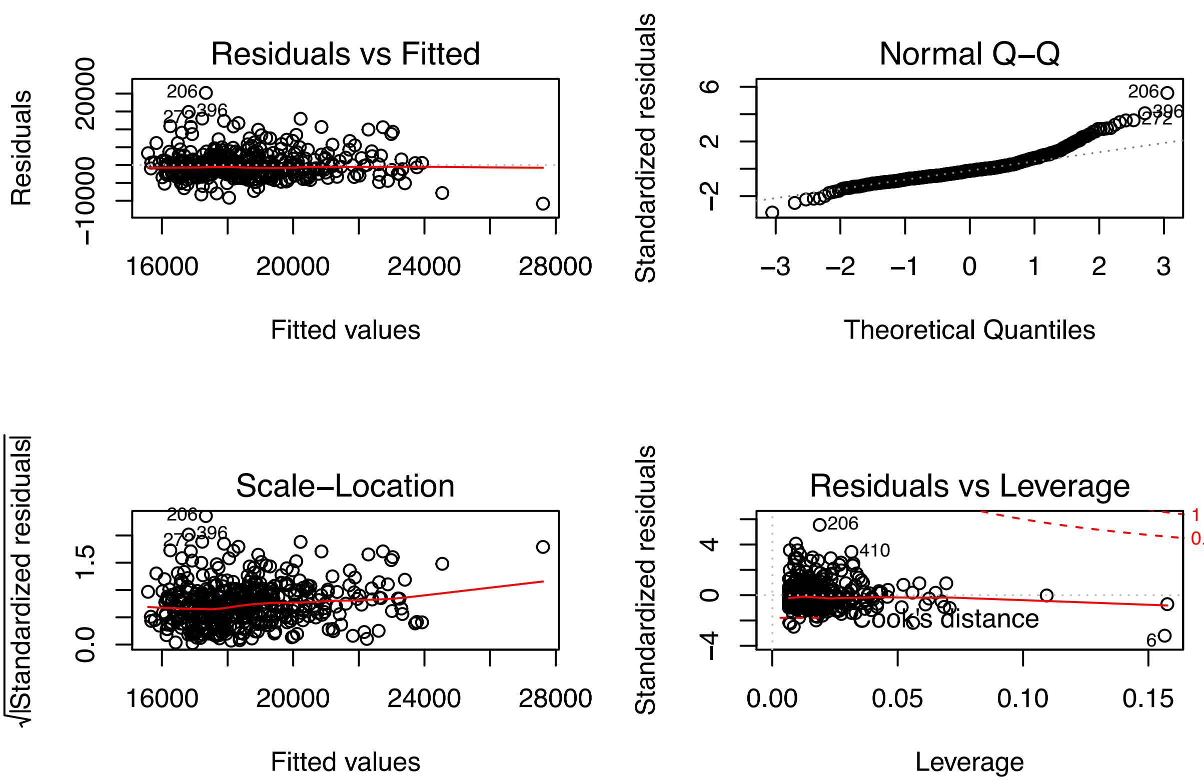
```
summary(interaction_mod) #interaction term does not seem to be significant
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region + crimes:region,
##     data = cdi_transformed)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10810  -2127   -533   1187  20202
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9634.8     2888.0   3.336 0.000923 ***
## crimes              938.1      310.3   3.024 0.002648 **
## regionNE          -4544.8     4262.1  -1.066 0.286870
## regionS           -2595.4     4201.9  -0.618 0.537117
## regionW           -4784.6     4846.6  -0.987 0.324093
## crimes:regionNE     738.8      457.8   1.614 0.107313
## crimes:regionS      141.8      441.4   0.321 0.748223
## crimes:regionW      426.0      499.8   0.852 0.394467
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3676 on 432 degrees of freedom
## Multiple R-squared:  0.1931, Adjusted R-squared:  0.1801
## F-statistic: 14.77 on 7 and 432 DF,  p-value: < 2.2e-16
```

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




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



Untransformed Data:/ Looking at the simple linear model of per capita income vs crime and region, it appears that for every additional crime committed, one would expect a 8.915e-03 dollar increase in per capita income. When the same regression is run using per capita crime in place of crime, the per capita crime coefficient is not counted as significant. Further, including interactions between the two predictors included in each model does not seem to be significant. Since the response variable is per capita, it would follow that one would want the predictor variable to also be per capita. Therefore, of the two basic models fit (predicting per capita income by crime or per capita crime), it seems that the model using per capita crime more aptly captures the true relationship (or lack thereof) between per capita income and crime rate. This being said, looking at the diagnostic plots of all four models fit above, there seems to be other underlying relationships that the models do not account for (variance is not constant in residual and standardized residual plots). / Transformed Data:/ Of models fit on the transformed data, the simple linear model predicting per capita income by log of total crime and region seems to produce the most significant results, with all of the predictors (barring regionW) having p-values $< .05$. Of the six models fit (on untransformed and transformed data), this model is the only one that shows a significant relationship between (almost all of) the predictors of interest and per capita income. Note, diagnostic plots of models fit on transformed data seem to adhere to model assumptions relatively well.

SECTION 3

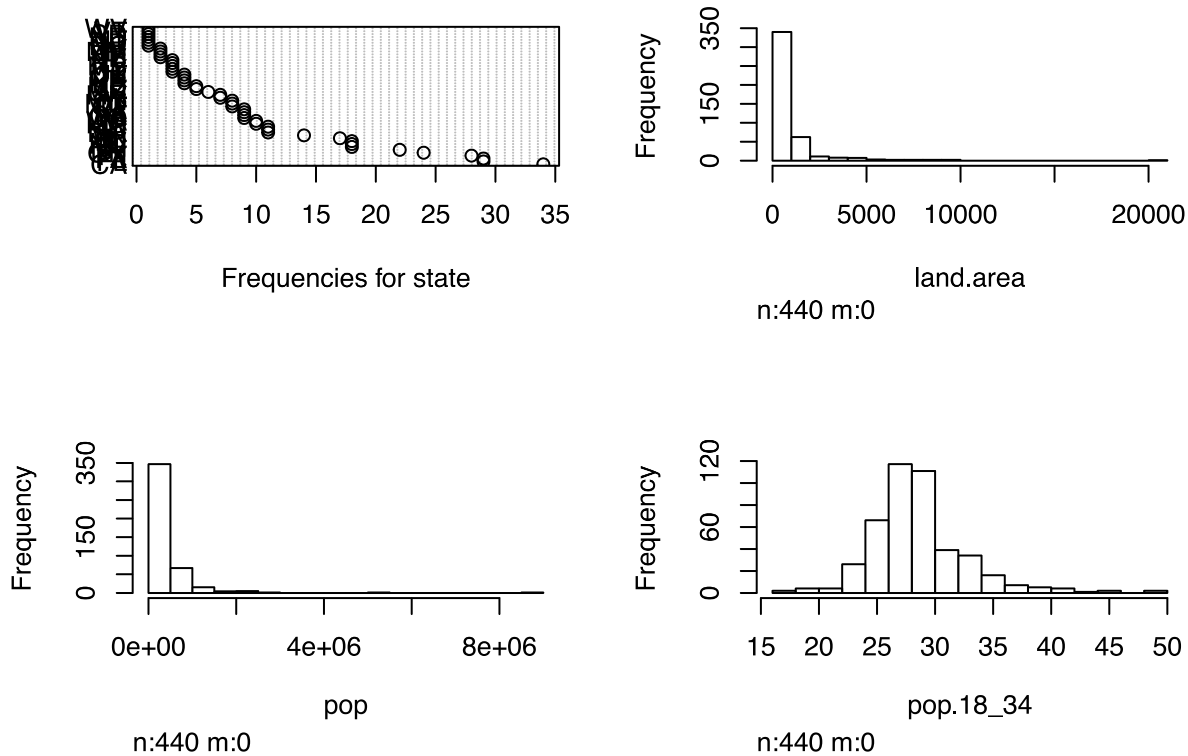
```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

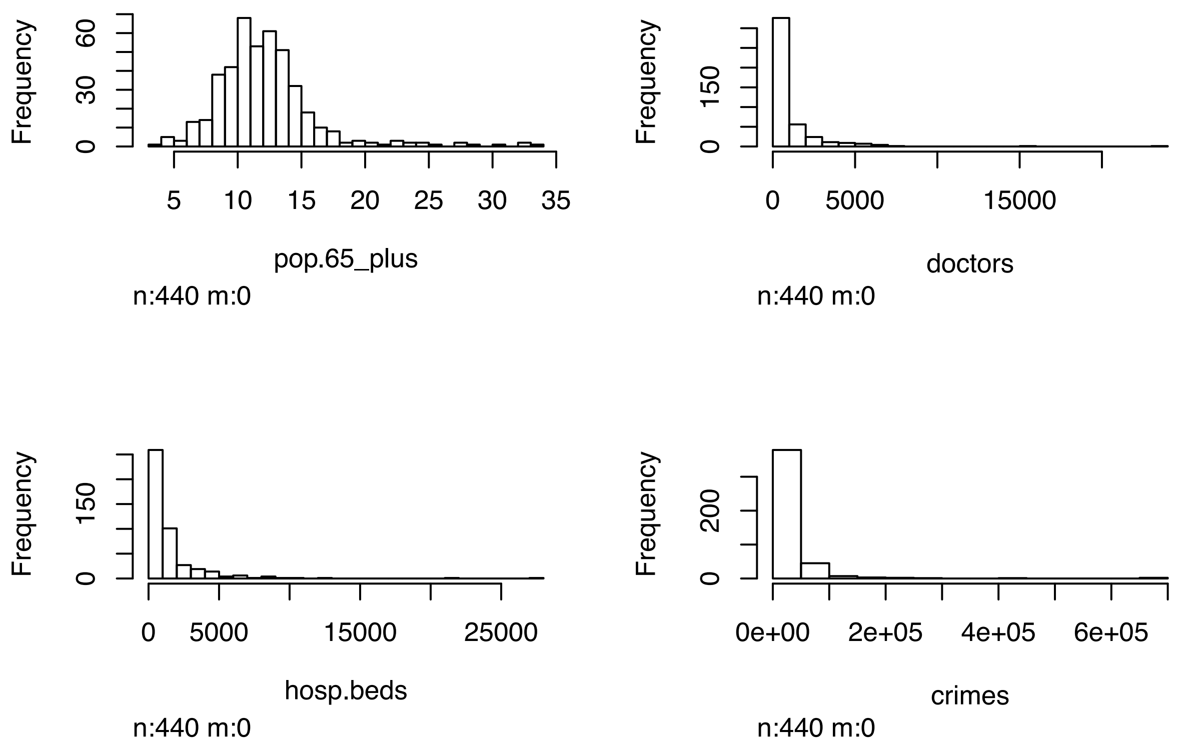
```
library(dplyr)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

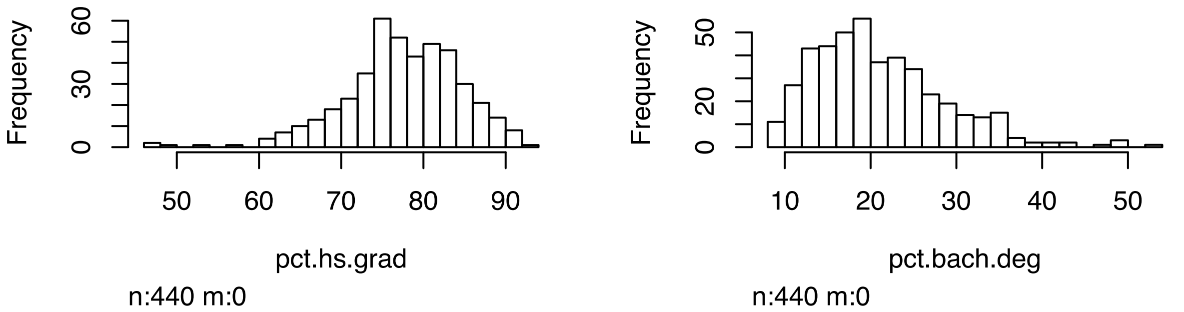
```
library(leaps)
hist.data.frame(cdi[,3:6])
```



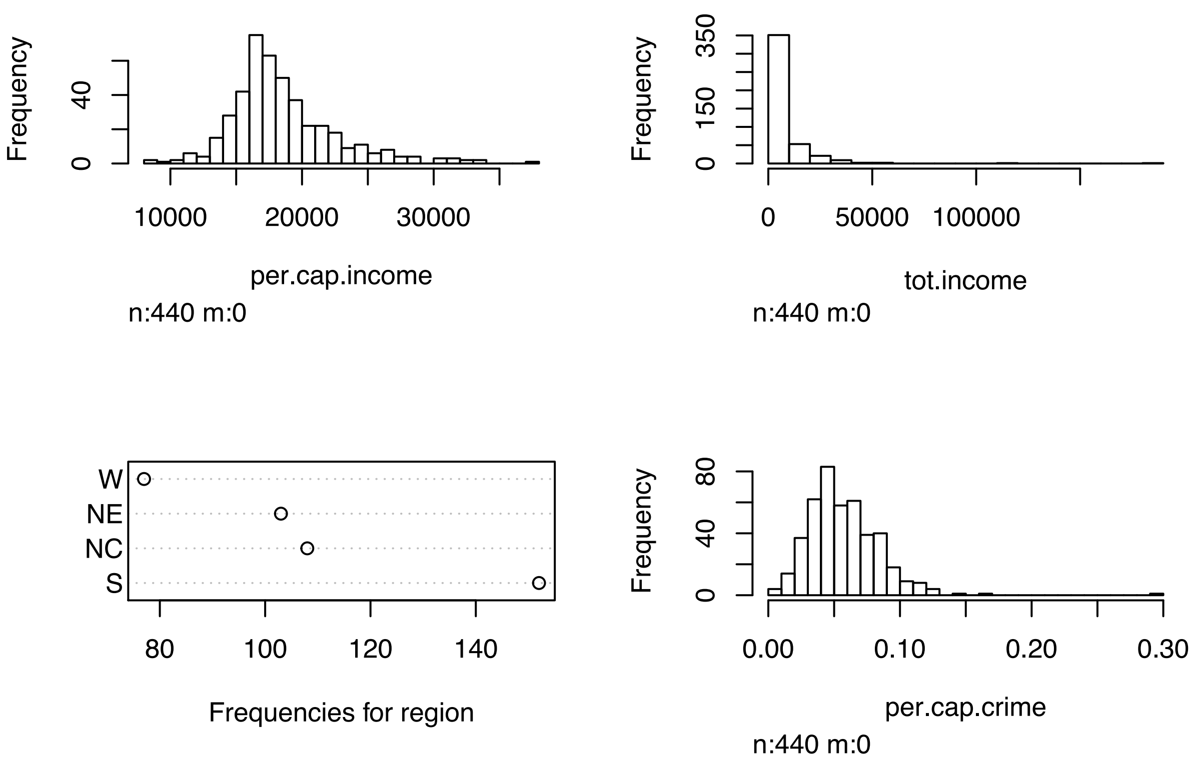

```
hist.data.frame(cdi[,7:10])
```




```
hist.data.frame(cdi[,11:14])
```

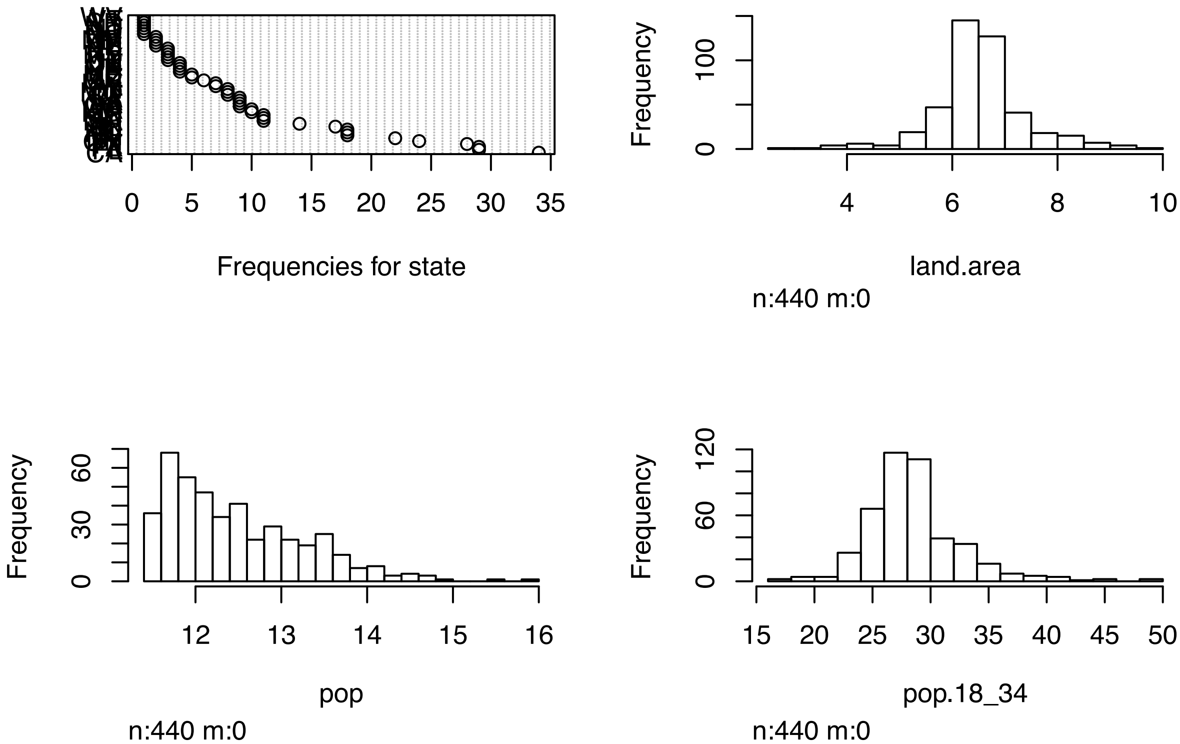



```
hist.data.frame(cdi[,15:18])
```

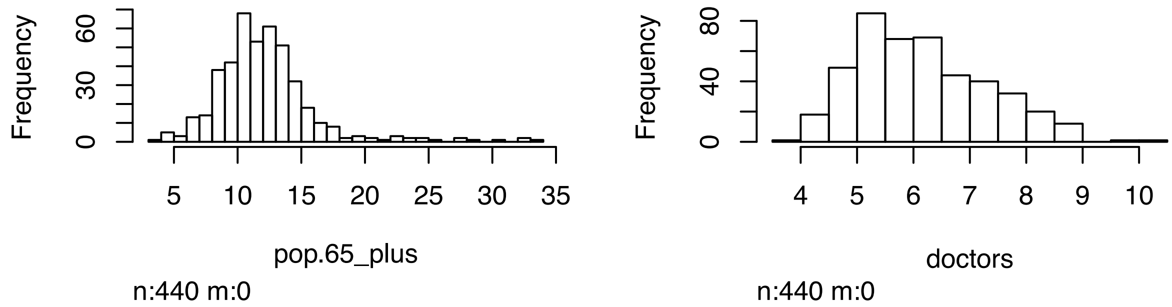



```
cdi_transformed = cdi %>%
  mutate(land.area = log(land.area),
         pop = log(pop),
         doctors = log(doctors),
```

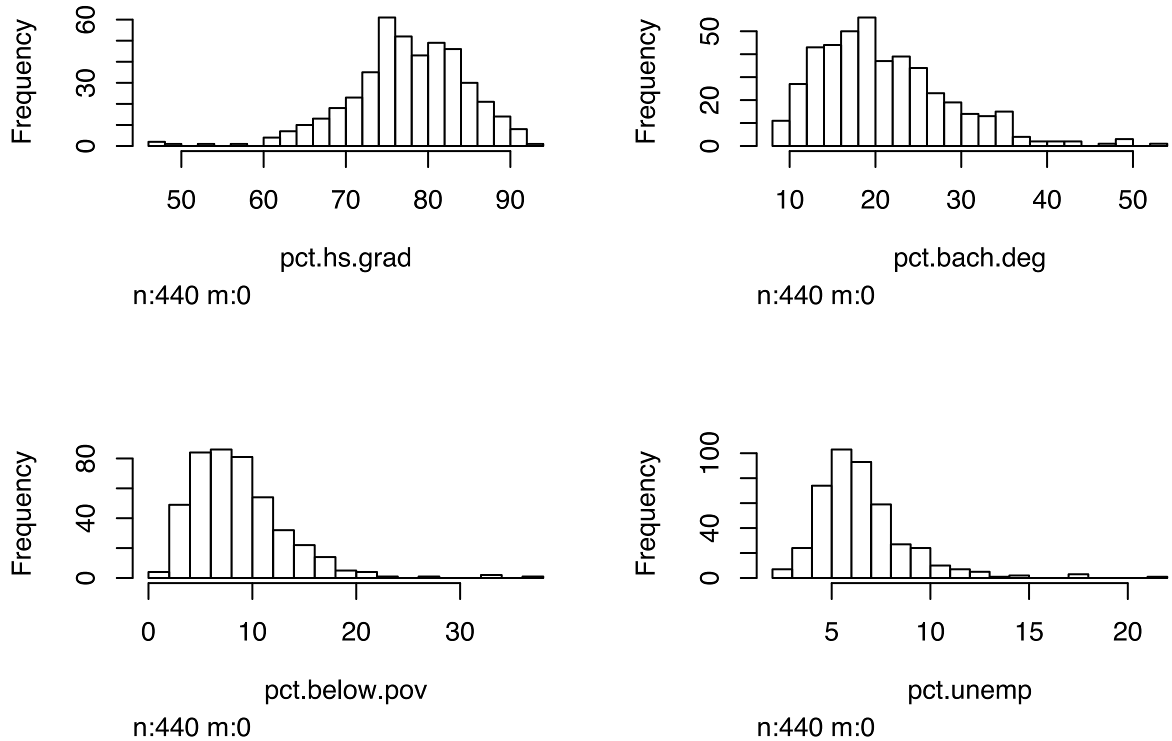
```r
         hosp.beds = log(hosp.beds),
         crimes = log(crimes),
         tot.income = log(tot.income))

hist.data.frame(cdi_transformed[,3:6])
```



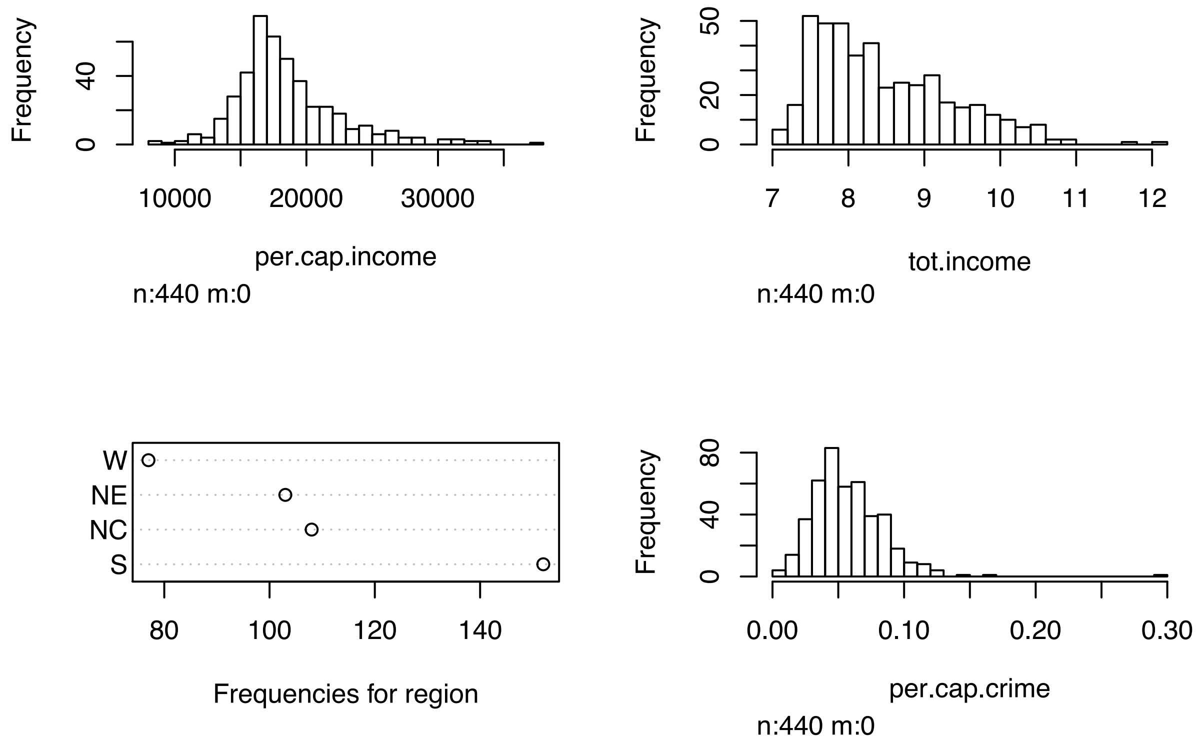

```r
hist.data.frame(cdi_transformed[,7:10])
```




```
hist.data.frame(cdi_transformed[,11:14])
```




```
hist.data.frame(cdi_transformed[,15:18])
```




```
cor(cdi_transformed$pct.hs.grad, cdi_transformed$pct.bach.deg)
```

```
## [1] 0.7077867
```

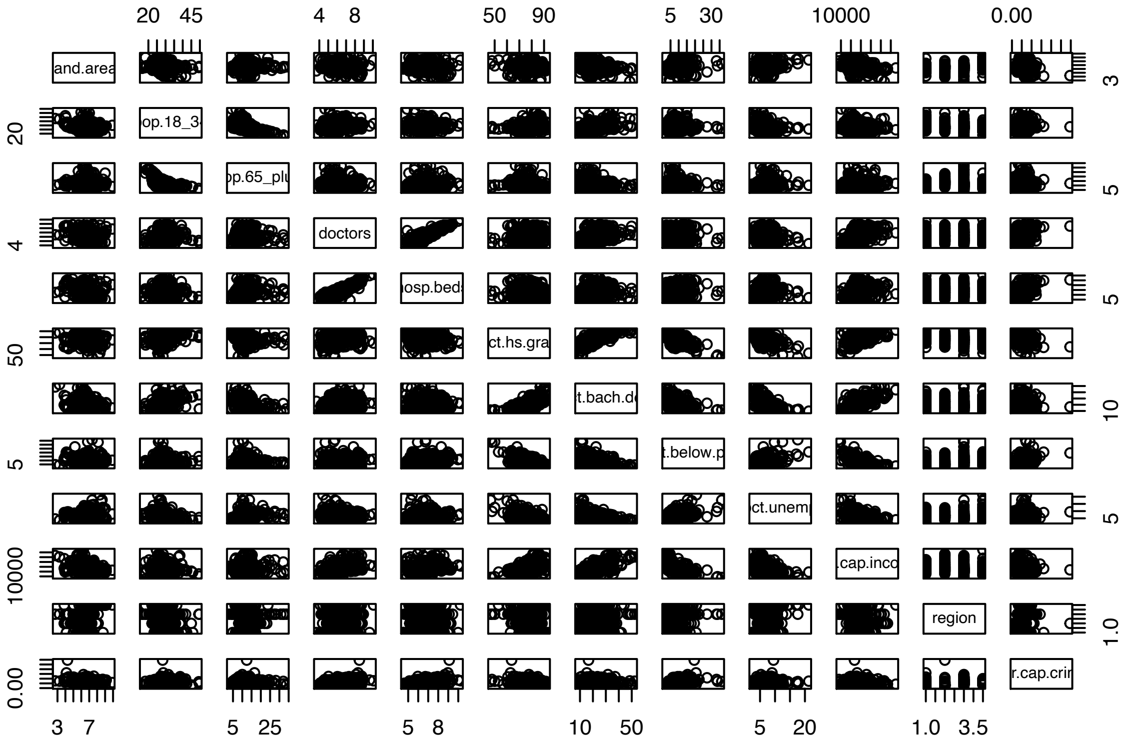
```
summary(cdi$state)
```

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

```
cdi_transformed = cdi_transformed %>%
  mutate(id = NULL, #removed since does not add to analytical ability of dataset
         county = NULL, #removed since does not add to analytical ability of dataset
         state = NULL, #removed since some states had very few or no observations
         pop = NULL, #removed since per capita income = total income/population
         tot.income = NULL, #removed since per capita income = total income/population
         crimes=NULL) #removed since per capita crime was added to dataset

(apply(cdi_transformed, 2, function(x) {which(is.infinite(x))}))
```

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

```
pairs(cdi_transformed)
```



```r
colnames(cdi_transformed[ ,c(which(vif(lm(per.cap.income~.,
       data=cdi_transformed))[,1]>5))])
```

```
## [1] "doctors"   "hosp.beds"
```

```r
#signs of multicollinearity

cor(cdi_transformed$pct.bach.deg, cdi_transformed$pct.hs.grad)
```

```
## [1] 0.7077867
```

Note, while the correlation between pct.bach.deg and pct.hs.grad is quite high (.7078), both variables were left in the transformed dataset. This decision was made under the assumption that pct.bach.deg acts as a subset of pct.hs.grad, allowing the two variables to behave similar to factors of education.

######################### SECTION 4 ##########################

```r
#simple model
cdi_subsets_mod = regsubsets(per.cap.income~.-region,
                             data=cdi_transformed,
                             nvmax = 12)

coef(cdi_subsets_mod, which.min(summary(cdi_subsets_mod)$bic))
```

```
##   (Intercept)     land.area     pop.18_34       doctors   pct.hs.grad
##    28748.6035     -683.8873     -300.3892     1000.9013     -116.8039
##  pct.bach.deg pct.below.pov    pct.unemp
##      371.0053     -427.2673      251.4416
```

```r
#checking BIC value for simple model
BIC(lm(per.cap.income~land.area+
         pop.18_34+
         doctors+
         pct.hs.grad+
```

```
          pct.bach.deg+
          pct.below.pov+
          pct.unemp,
        data=cdi_transformed))
```

```
## [1] 7847.685
```

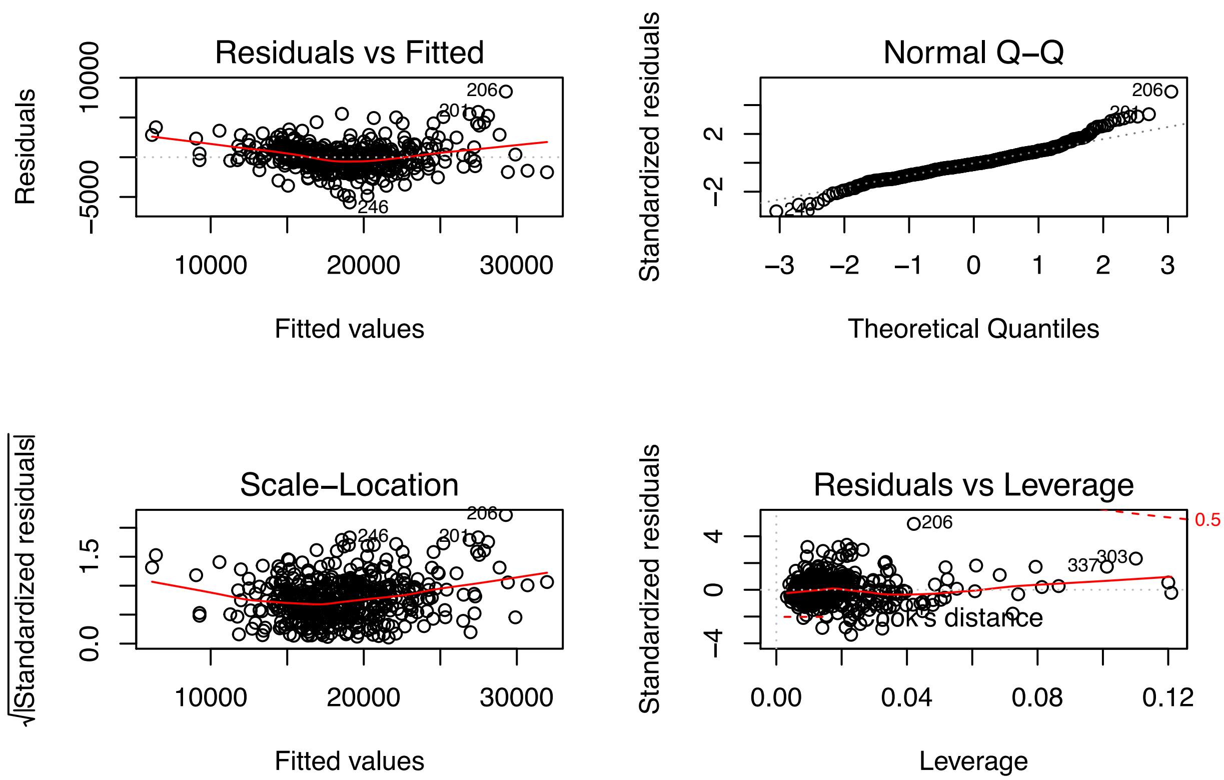
```
#checking for collinearity in simple model
which(vif(lm(per.cap.income~., data=cdi_transformed))>5)
```

```
## [1] 4 5
```

```
subsets_mod = lm(per.cap.income~land.area+
          pop.18_34+
          doctors+
          pct.hs.grad+
          pct.bach.deg+
          pct.below.pov+
          pct.unemp,
        data=cdi_transformed)

which(vif(subsets_mod)>5)
```

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

```
#checking model assumptions for simple model
par(mfrow=c(2,2))
plot(subsets_mod)
```




```
#summary of simple model
summary(subsets_mod)
```

```
##
## Call:
## lm(formula = per.cap.income ~ land.area + pop.18_34 + doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp, data = cdi_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5688.4 -1015.1  -123.4   892.2  8260.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   28748.60    1944.84  14.782  < 2e-16 ***
## land.area      -683.89      99.76  -6.855 2.47e-11 ***
## pop.18_34      -300.39      23.21 -12.942  < 2e-16 ***
## doctors        1000.90      83.92  11.926  < 2e-16 ***
## pct.hs.grad    -116.80      22.60  -5.168 3.63e-07 ***
## pct.bach.deg    371.01      19.31  19.214  < 2e-16 ***
## pct.below.pov  -427.27      26.28 -16.258  < 2e-16 ***
## pct.unemp       251.44      45.47   5.530 5.56e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1713 on 432 degrees of freedom
## Multiple R-squared:  0.8248, Adjusted R-squared:  0.822
## F-statistic: 290.6 on 7 and 432 DF,  p-value: < 2.2e-16
```

Based on the above diagnostic plots, one can see that model assumptions are met for the most part by the simple (non-interaction) model. Residuals are centered around 0 with relatively constant variance above and below the x-axis. The normal q-q plot shows a pretty straight line, with slight deviation around tails. Standardized residuals seem somewhat centered around 1 and all except one point are in the acceptable [-2,2] range. The leverage plot does not denote any points with exceptionally high Cook's distance values (all points have values $< .5$).

```r
#interaction model
cdi_sw_t_mod = stepAIC(lm(per.cap.income~.*region+
                              pct.hs.grad:pct.below.pov+
                           pct.bach.deg:pct.below.pov,
                      data = cdi_transformed),
                      direction="both",
                      k=log(nrow(cdi_transformed)),
                      trace=0)

cdi_sw_t_mod
```

```
##
## Call:
## lm(formula = per.cap.income ~ land.area + pop.18_34 + doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + region + pct.hs.grad:region +
##     pct.bach.deg:region + pct.below.pov:region + pct.bach.deg:pct.below.pov,
##     data = cdi_transformed)
##
## Coefficients:
##                (Intercept)                   land.area
##                  30493.952                    -653.641
##                  pop.18_34                     doctors
```

```
##                    -291.409                    979.173
##                 pct.hs.grad               pct.bach.deg
##                    -106.827                    321.733
##               pct.below.pov                   regionNE
##                    -260.607                    707.388
##                     regionS                    regionW
##                   -9533.428                  20392.519
##        pct.hs.grad:regionNE        pct.hs.grad:regionS
##                     -54.776                     92.873
##         pct.hs.grad:regionW      pct.bach.deg:regionNE
##                    -283.267                    187.640
##        pct.bach.deg:regionS       pct.bach.deg:regionW
##                      26.891                    201.069
##      pct.below.pov:regionNE      pct.below.pov:regionS
##                     -29.571                    161.097
##       pct.below.pov:regionW  pct.bach.deg:pct.below.pov
##                    -219.626                     -9.588
```

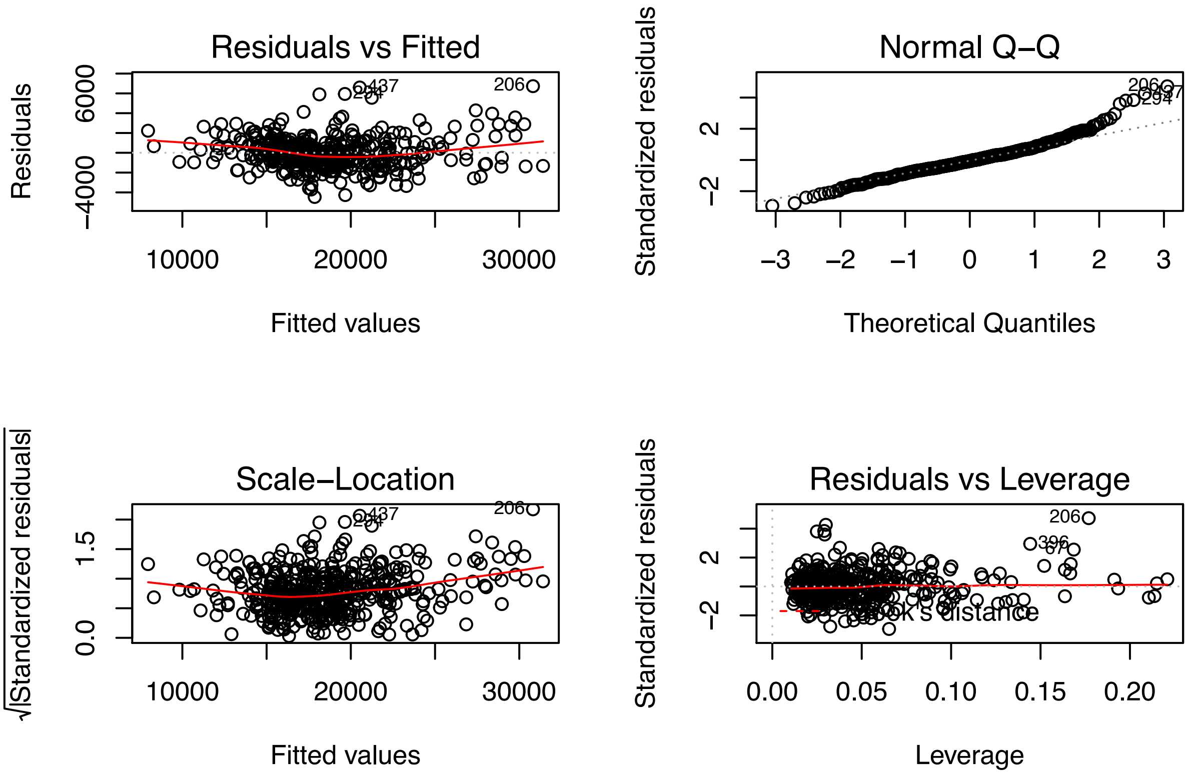
```
#checking BIC value for interaction model
BIC(cdi_sw_t_mod)
```

```
## [1] 7831.622
```

```
#checking for collinearity in interaction model
which(vif(cdi_sw_t_mod)[,1]>5)
```

```
##                 pct.hs.grad               pct.bach.deg
##                           4                          5
##               pct.below.pov                     region
##                           6                          7
##          pct.hs.grad:region        pct.bach.deg:region
##                           8                          9
##        pct.below.pov:region pct.bach.deg:pct.below.pov
##                          10                         11
```

```
#checking model assumptions for interaction model
par(mfrow=c(2,2))
plot(cdi_sw_t_mod)
```




```r
#summary of interaction model
summary(cdi_sw_t_mod)
```

```
##
## Call:
## lm(formula = per.cap.income ~ land.area + pop.18_34 + doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + region + pct.hs.grad:region +
##     pct.bach.deg:region + pct.below.pov:region + pct.bach.deg:pct.below.pov,
##     data = cdi_transformed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4460.0  -919.5   -95.3   759.2  6727.8
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 30493.952   5098.626   5.981 4.75e-09 ***
## land.area                    -653.641    111.633  -5.855 9.62e-09 ***
## pop.18_34                    -291.409     23.441 -12.432  < 2e-16 ***
## doctors                       979.173     86.659  11.299  < 2e-16 ***
## pct.hs.grad                  -106.827     65.250  -1.637 0.102337
## pct.bach.deg                  321.733     42.879   7.503 3.75e-13 ***
## pct.below.pov                -260.607     80.260  -3.247 0.001260 **
## regionNE                      707.388   6082.532   0.116 0.907472
## regionS                     -9533.428   5546.773  -1.719 0.086400 .
## regionW                     20392.519   6464.088   3.155 0.001722 **
## pct.hs.grad:regionNE          -54.776     80.428  -0.681 0.496210
## pct.hs.grad:regionS            92.873     73.408   1.265 0.206516
## pct.hs.grad:regionW          -283.267     80.297  -3.528 0.000465 ***
```

```
## pct.bach.deg:regionNE        187.640     51.710   3.629 0.000320 ***
## pct.bach.deg:regionS          26.891     45.461   0.592 0.554486
## pct.bach.deg:regionW         201.069     52.591   3.823 0.000152 ***
## pct.below.pov:regionNE       -29.571     94.345  -0.313 0.754108
## pct.below.pov:regionS        161.097     75.687   2.128 0.033880 *
## pct.below.pov:regionW       -219.626    108.255  -2.029 0.043111 *
## pct.bach.deg:pct.below.pov    -9.588      2.548  -3.762 0.000192 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1570 on 420 degrees of freedom
## Multiple R-squared:  0.8569, Adjusted R-squared:  0.8505
## F-statistic: 132.4 on 19 and 420 DF,  p-value: < 2.2e-16
```

Looking at the diagnostic plots above, it appears that assumptions are relatively well met by the more complex interaction model. Residuals seem centered around 0 with relatively constant variance around the x-axis. The normal q-q plot shows a pretty straight line with variance around the tails (especially on the upper tail, suggesting data may be skewed right). Standardized residuals center loosely around 1 and most points fall within the acceptable [-2,2] range (visually, there seems to be two points that fall outside of this range). The leverage plot does not show any points with exceptionally high leverage, with all points having Cook's distance values < .5.

SECTION 5

```
table(cdi_transformed$region)
```

```
##
##  NC  NE   S   W
## 108 103 152  77
```

```
table(cdi$state)
```

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

```
table(cdi$state, cdi_transformed$region)
```

```
##
##      NC NE  S  W
##   AL  0  0  7  0
##   AR  0  0  2  0
##   AZ  0  0  0  5
##   CA  0  0  0 34
##   CO  0  0  0  9
##   CT  0  8  0  0
##   DC  0  0  1  0
##   DE  0  2  0  0
##   FL  0  0 29  0
##   GA  0  0  9  0
##   HI  0  0  0  3
##   ID  0  0  0  1
##   IL 17  0  0  0
##   IN 14  0  0  0
##   KS  4  0  0  0
```

```
##    KY  0  0  3  0
##    LA  0  0  9  0
##    MA  0 11  0  0
##    MD  0  0 10  0
##    ME  0  5  0  0
##    MI 18  0  0  0
##    MN  7  0  0  0
##    MO  8  0  0  0
##    MS  0  0  3  0
##    MT  0  0  0  1
##    NC  0  0 18  0
##    ND  1  0  0  0
##    NE  3  0  0  0
##    NH  0  4  0  0
##    NJ  0 18  0  0
##    NM  0  0  0  2
##    NV  0  0  0  2
##    NY  0 22  0  0
##    OH 24  0  0  0
##    OK  0  0  4  0
##    OR  0  0  0  6
##    PA  0 29  0  0
##    RI  0  3  0  0
##    SC  0  0 11  0
##    SD  1  0  0  0
##    TN  0  0  8  0
##    TX  0  0 28  0
##    UT  0  0  0  4
##    VA  0  0  9  0
##    VT  0  1  0  0
##    WA  0  0  0 10
##    WI 11  0  0  0
##    WV  0  0  1  0
```