Per Capita Income Prediction from County Demographic Information

Xiangman Zhao
xiangmaz@andrew.cmu.edu
October 2021

Abstract

Four related questions about finding appropriate variables to predict personal income per capita in the United States are examined in this paper. The data used comes from Kutner et al. (2005) and includes selected county demographic information for 440 most populous counties in the United States. The model shows that personal income is largely affected by pct.hs.grad, pct.bach.deg, pct.below.pov and pct.unemp. Crimes is not a significant factor, but region is related to per.cap.income. All subsets regression and stepwise regression are used to select significant variables. ANOVA test and diagnostic plots are used to evaluate the linearity of the model. The final model includes some important predictors like doctors, pct.bach.deg, pct.unemp as well as some added interaction terms with region. There is still room for model improvement if additional data on missing counties can be added to the existing dataset to reduce overfitting noise.

1. Introduction

Personal income is an important metric to evaluate local wealth and prosperity. The local government can also use it as a way to evaluate the standard of living and quality of life of a population. Many social scientists are trying to figure out what factors affect per capita income.

This question is critical for social scientists to solve to find the relationship between person income and other variables associated with the country's economic, social well beings and health. Solving the problem also helps to improve people's living, and the main aim of the paper is to build an optimal model to predict the personal income and investigate if there are any additional information needed to better answer the question.

Four questions related to the per capita income are presented below:

1. Which variables are related to each other?
2. How do crime rate, per capita crime and region relate to per capita income?
3. What is the best model to predict per capita income?
4. Can we improve the model by including more data on missing counties?

2. Data

The data for this study provides selected county demographic information for 440 of the most populous counties in the United States. Readers can check Kutner et al. (2005) for more details and information. There are 17 variables in the dataset, and three of them are categorical variables: region, county and state. Table 1 includes definitions for all the variables in the dataset:

Variable names	Descriptions
Identification number	1 – 440
County	County name
State	Two-letter state abbreviation
Land area	Land area (square miles)
Total population	Estimated 1990 population
Percent of population aged 18-34	Percent of 1990 CDI population aged 18–34
Percent of population 65 or older	Percent of 1990 CDI population aged 60 or old
Number of active physicians	Number of professionally active nonfederal physicians during 1990
Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
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
Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
Percent Bachelor's degree	Percent of adult population (persons 25 years old or older) with bachelor's degree
Percent below poverty level	Percent of 1990 CDI population with income below poverty level
Percent unemployment	Percent of 1990 CDI population that is unemployed
Per capita income	Per-capita income (i.e. average income per person) of 1990 CDI population (in dollars)
Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
Geographic region	Geographic region classification used by the US Bureau of the Census, NE (northeast region of the US), NC (north-central region of the US), S (southern region of the US), and W (Western region of the US)

Table 1: Variables Definitions

Table 2 is a summary table of all the quantitative variables in the dataset except for variable identification number:

	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

Table 2: Summary Table for Quantitative Variables

Table 3 and Table 4 summarize three categorical variables region, county and state in the dataset. There are 4 unique regions, 373 unique counties and 48 unique states:

Region	Frequency	Baseline salary
W	77	20332.99
NC	108	20743.74
S	152	19341.34
NE	103	23155.79

Table 3: Summary Table for Region

	Max frequency	Median frequency	Min frequency	Count
County	Jefferson 7	1	1	373
State	CA 34	7	1	48

Table 4: Summary Table for County and State

Variables county and state have so many unique values that their frequency tables do not provide much useful information. Table 3 shows that NE (Northeastern) has higher baseline salary.

Based on Table 2, some variables seem to have very different means and medians, which shows that they have very skewed distributions. Histograms are plotted for these variables for further analysis in Figure 1:

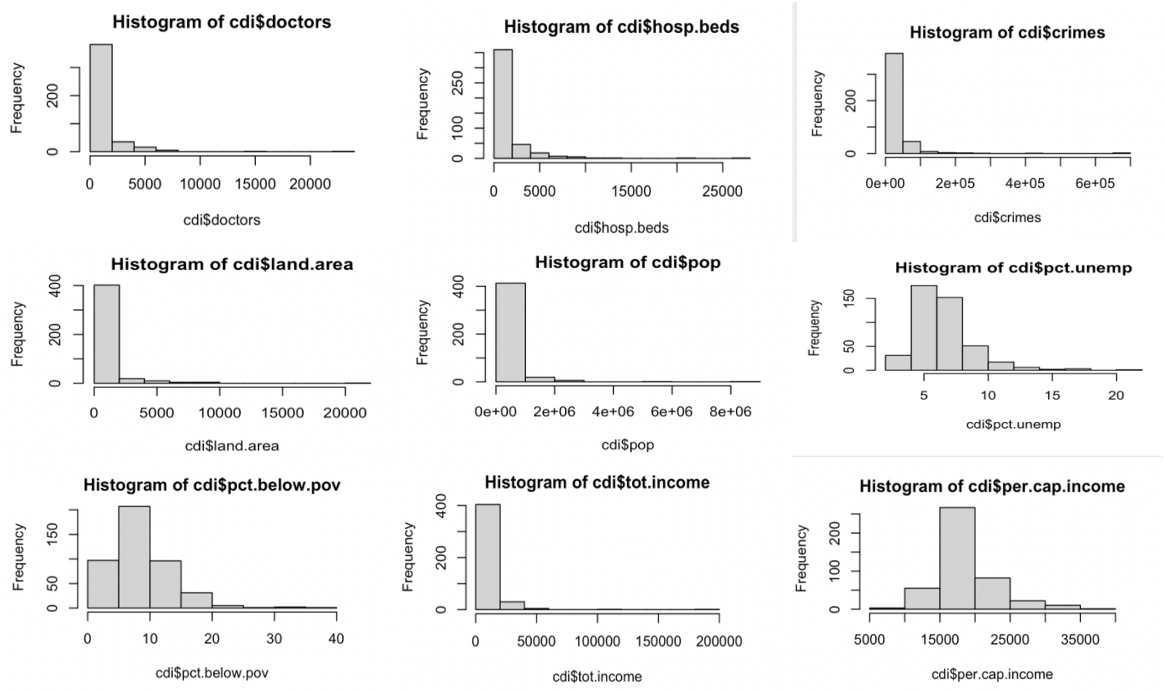


Figure 1: Histograms of Skewed Variables

3. Methods

3.1 Relationship between each pair of variables

For the first research question, correlation plot is used to investigate the relationship between each quantitative variable in the dataset. 12 scatterplots between each independent variable and per.cap.income are plotted to investigate the relationship between these independent variables and the per.cap.income.

3.2 How crimes and region are related to per.cap.income

For the second research question, linear models based on combinations of (per capita) crimes, region and the interaction term between (per capita) crimes and region are built to predict per.cap.income. ANOVA test is used to find the most significant model between these models.

3.3 Finding the best model to predict per.cap.income

For the third research question, in order to keep independent variables consistent with the response variable per capita income, variables land.area, doctors, hosp.beds, crimes are transformed to land area per capita, doctors per capita, hosp.beds per capita and crimes per capita. Variables pop.18_34 and pop.65_plus are divided by population to get percent of pop.18_34 and percent of pop.65_plus. Since per.cap.income is calculated by dividing population by total income, variables pop and tot.income are removed to avoid too much

collinearity. Log transformation is used to make those skewed variables normal. After the transformation, two variables selection methods: stepwise and all subsets are applied to find the most appropriate and significant quantitative variables in the dataset. The VIF is calculated for each of the predictors to assess the severity of multicollinearity when all significant quantitative variables are included. A model containing all interaction terms between region and all other quantitative variables is built to see if any interaction terms will help explain the model. Four diagnostic plots are used to evaluate the linearity of these models. In addition to these key modeling assumptions, the final model also needs to be interpretable in the context of economics and social well-beings.

3.4 Investigating the need for additional data

For the fourth research question, a five-fold cross validation is used to evaluate the predictability of the model. A model containing all interaction terms between state and all other quantitative variables is built to see if state can provide more information than region.

4. Results

4.1 Relationship between each pair of variables

To answer the first research question, a correlation plot in Figure 2 and scatterplots between per.cap.income and quantitative variables in Figure 3 are used to investigate the relationship between each variable, and it is found that:

- Tot.income is correlated to doctors, hosp.beds, crimes and pop. It is expected as total income can be largely affected by crime rate and health system. Per.cap.income is calculated by tot.income/pop, which is not surprising to see that tot.income and pop are highly correlated.
- Pop.18_34 is negatively related to pop.65_plus.
- Variables pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp and per.cap.income are correlated to each other. It is expected as per capita income is related to education levels and unemployment rate.
- Doctors and hosp.beds are highly correlated, which makes sense as the number of doctors increases as the number of hospital beds increases.

More information can be found in EDA section (page 1 to 6) of the code appendix.

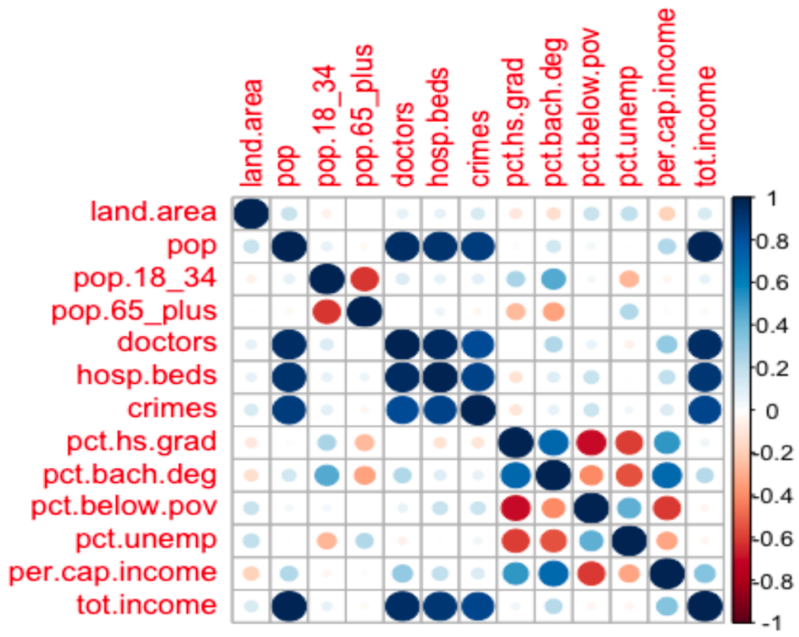


Figure 2: Correlation plot

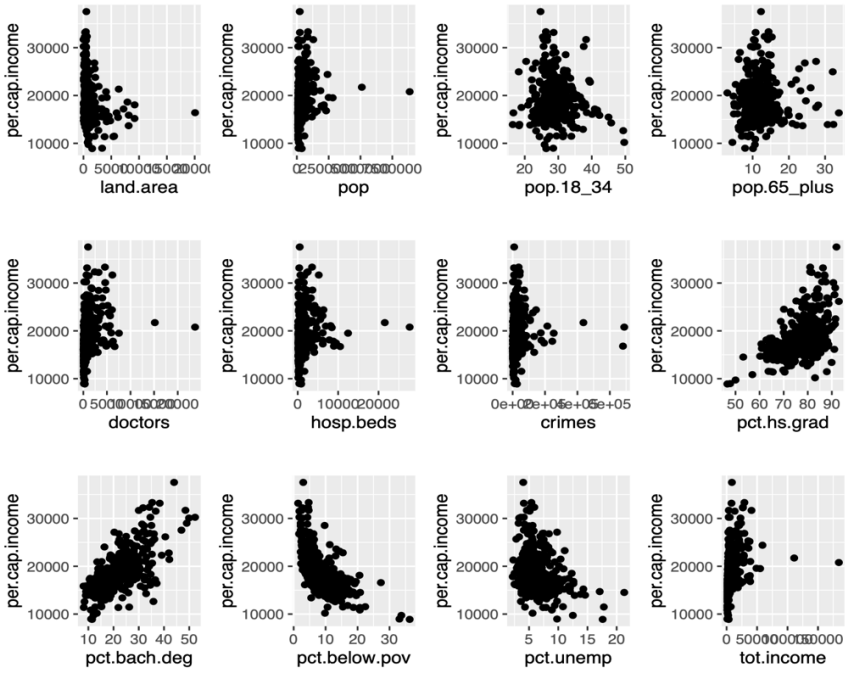


Figure 3: Scatterplots between Quantitative Variables and Per.cap income

4.2 How crimes and region are related to per.cap.income

For the second research question on whether variable (per capita) crimes and region are significant, 4 modes are built for comparison.

1. Per.cap.income predicted by crimes and region
2. Per.cap.income predicted by crimes, region and the interaction term between crimes and region
3. Per.cap.income predicted by crimes per capita and region
4. Per.cap.income predicted by crimes per capita, region and the interaction term between crimes per capita and region

ANOVA tests between model 1 and model 2, model 3 and model 4 are used to compare these three models and find that neither the interaction between crimes and region nor the interaction between crimes per capita and region is significant. These models' summary tables show that only region is significant at 5% significance level, especially for Northeast which is significant in all models. Therefore, neither crimes nor crimes per capita is significant in predicting per.cap.income. Region is a significant predictor and needs to be included in the model. More details can be found in crimes section (page 6 to 11) of the code appendix.

4.3 Finding the best model to predict per.cap.income

The third research question is trying to build the most appropriate model to predict per.cap.income. After variables transformations stated in the method section, 2 variables selections methods, all subsets regression and stepwise AIC regression, are applied to find significant quantitative predictors from 10 quantitative variables. Both variables selection methods choose the same 8 variables as below:

Log(per.cap.income) = 9.736 – 0.03199 * log(land.area.per.capita) – 581.51 * *pct.pop.18_34* + 759.34 * *pct.pop.65_plus* + 0.063 * *log(doctors_per_capita)* – 0.00533 * *pct.hs.grad* + 0.294 * *log(pct.bach.deg)* – 0.0272 * *pct.below.pov* + 0.0135 * *pct.unemp*

Although categorical variable region is not included in the variables selection process, the previous part finds that region is important and will be included in the model. The next step is to investigate which interaction terms with region should be added to the model. The summary table of the model containing all interaction terms between region and all other quantitative variables shows 4 significant interaction terms: *pct.hs.grad * region*, *pct.below.pov * region*, *log(doctors.per.capita) * region* and *pct.unemp * region*. These 4 interaction terms are added to the model above. Four diagnostic plots in Figure 4 show that the model satisfies all the linearity conditions: independence, normality, constant variance and no bad leverage points or outliers. R squared is 0.8269, which shows that 82.69% of the data can be explained by the model. The final model is shown as below:

Log(per.cap.income) = 9.082 – 0.0334 * log(land.area.per.capita) – 644.43 * *pct.pop.18_34* + 1034.91 * *pct.pop.65_plus* + 0.0345 * *log(doctors.per.capita)* – 0.00546 * *pct.hs.grad* + 0.316 * *log(pct.bach.deg)* – 0.0275 * *pct.below.pov* + 0.0246 * *pct.unemp* + 0.00189 * *regionNE:pct.hs.grad* + 0.0000119 * *regionS:pct.hs.grad* – 0.0118 * *regionW:pct.hs.grad* – 0.00162 * *pct.below.pov:regionNE* + 0.00606 * *pct.below.pov:regionS* – 0.0113 *

pct.below.pov:regionW + 0.0257 * *regionNE:log(doctors.per.capita)* + 0.0212 * *regionS:log(doctors.per.capita)* + 0.0679 * *regionW:log(doctors.per.capita)* – 0.0189 * *pct.unemp:regionNE* – 0.0256 * *pct.unemp:regionS* – 0.02 * *pct.unemp:regionW*

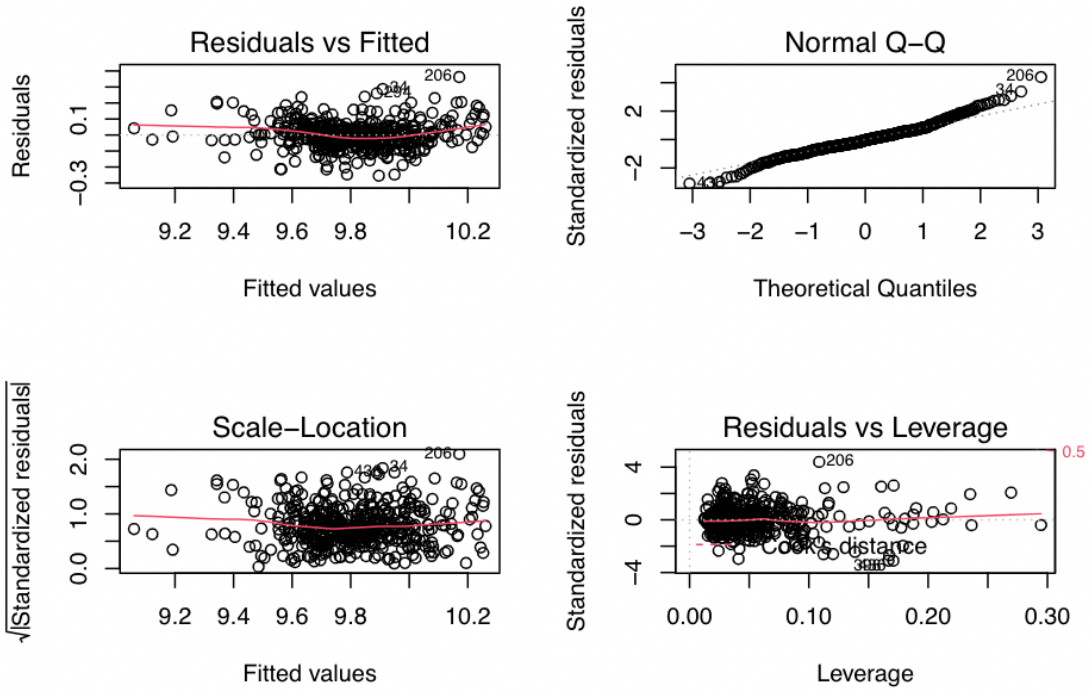


Figure 4: Diagnostic plots of the final model

Table 5 is the interpretation of the final model:

Predictor	Interpretation
Log(land.area.per.capita)	For every 1% increase in per capita land area, per capita income will increase by 0.0334%.
Pct.pop.18_34	For every 1% increase in the percent of population between age 18 and 34, per capita income will decrease by 644.43%.
Pct.pop.65_plus	For every 1% increase in the percent of population over age 65, per capita income will increase by 1034.91%.
Log(doctors.per.capita)	For every 1% increase in doctors per capita, per capita income will increase by 0.0345%.
Pct.hs.grad	For every 1% increase in the percent of high school graduates, per capita income will decrease by 0.00546%.
Log(pct.bach.deg)	For every 1% increase in the percent of bachelor's degree, per capita income will increase by 0.316%.
Pct.below.pov	For every 1% increase in the percent of people below poverty, per capita income will decrease by 0.0275%.

Pct.unemp	For every 1% increase in the percent of unemployment, per capita income will increase by 0.0246%.
Interaction terms with region	North central part is not significant in any of the interaction terms with percent of high school graduates, percent of people below poverty, doctors per capita and percent of unemployment. Among interaction terms, percent of unemployment seems to have a negative effect on per capita income for all three regions. Doctors per capita has a positive effect on per capita income for all three regions. Percent of high school graduates has a negative effect on per capita income in the west, but positive effect on per capita income in southern and northeastern parts. Percent of people below poverty seems to positively affect per capita income in the south while negatively affect per capita income in the west and northeast.

Table 5: Interpretations of the final model

More information can be found in model fitting section (page 11 to 18) of the code appendix.

4.4 Investigating the need for additional data

For the fourth research question, the five-fold cross validation shows that the cross-validated R squared is 0.8010. Compared with the model's R squared of 0.8269, there is not a significant difference between these two values, which shows that there is not a huge problem of overfitting. The summary table of the model containing all interaction terms between state and all other quantitative variables indicates that none of the interaction term is significant. There are also many NAs for some variables' p-values, because there are 48 unique states and 373 unique counties in the dataset, but the dataset only contains 440 rows of data. However, it does not mean that variable region is enough to account for the geographical influence on per.cap.income. If more data can be added to the existing dataset, variables state and county can be added to the model for future improvements. More information can be found in the cross-validation section (page 18 to 27) of the code appendix.

5. Discussions

5.1 Relationship between each pair of variables

For the first research question, per.cap.income is related to economic factors such as pct.unemp and pct.below.pov, social factors like pct.hs.grad and hosp.beds and geographical factors like region. In the exploratory analysis, it is found that pct.hs.grad, pct.bach.deg and pct.below.pov are highly related to per.cap.income. There are also some relationships existing between pop.18_34 and pop.65_plus, but the VIF indicates that collinearity is not an issue in the final model. Interaction terms between region and quantitative variables pct.unemp, pct.below.pov and log.doctors.per.capita are also crucial to explain the change in per.cap.income.

5.2 How crimes and region are related to per.cap.income

For the second research question, it is surprising to find out that the interaction term between crimes and region is not a significant factor to predict per.cap.income, but region alone is an important factor in the model. ANOVA tests between different linear regression models effectively show that neither crimes nor crimes per capita is significant factor. It is intuitive to conclude that only region is significant to predict per.cap.income.

5.3 Finding the best model to predict per.cap.income

For the third research question, the model does a good job predicting per.cap.income, and the prediction error is pretty small from the cross validation. Diagnostic plots show that the model fulfills all the linearity condition, and collinearity is not a problem. The final model for predicting per.cap.income includes log(land.area.per.capita), pct.below.pov, pct.unemp, region, pct.pop.65_plus, pct.pop.18_34, log(pct.bach.deg), pct.hs.grad, log(doctors.per.capita), regionNE:pct.hs.grad, regionS:pct.hs.grad, regionW:pct.hs.grad, pct.below.pov:regionNE, pct.below.pov:regionS, pct.below.pov:regionW, regionNE:log(doctors.per.capita), regionS:log(doctors.per.capita), regionW:log(doctors.per.capita), pct.unemp:regionNE, pct.unemp:regionS and pct.unemp:region.

5.4 Investigating the need for additional data

For the last question, the five-fold cross validation indicates that overfitting is not a big problem in the model, but it is still useful to find more data on state and add to the existing dataset. There are 48 unique values in variable state, but there are only 440 rows of data. If more data can be combined into the original dataset, the model will provide more information on how locations influence per.cap.income and probably be more accurate.

5.5 Limitations and future improvements

Some strengths of the final model is that all the independent variable and per.cap.income have consistent measurement scales, and log transformations on both dependent variable and independent variables make the model fairly easy to interpret. There are still some limitations with the final model. First, the coefficients of variables pct.unemp and log.land.area.per.capita don’t make sense, since it is common to expect that pct.unemp is negatively related to per.cap.income and log.land.area.per.capita should be positively related to per.cap.income. Second, the model is still complicated with 8 independent variables and 4 interaction terms, which makes it hard for nontechnical people to understand. Lastly, the variable region only has four values, which is not sufficient to explain the geographical influence on per.cap.income. If I want to include some interaction terms between state and other quantitative variables, there will be many NA values. Therefore, more data on states and counties is needed to expand the existing dataset to provide a more accurate prediction of per.cap.income.

6. References

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

[2]Sheather, S. J. (2009). A modern approach to regression with R. (Springer eBooks.)

Technical Appendix

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Cmd+Shift+Enter*.

EDA

The code below tries to make a summary table of all the quantitaitive variables.

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

code below tries to create histograms for each quantitative variables to check distributions.

```
par(mfrow=c(2,2))
hist(cdi$land.area)
hist(cdi$pop)
hist(cdi$pop.18_34)
hist(cdi$pop.65_plus)
```

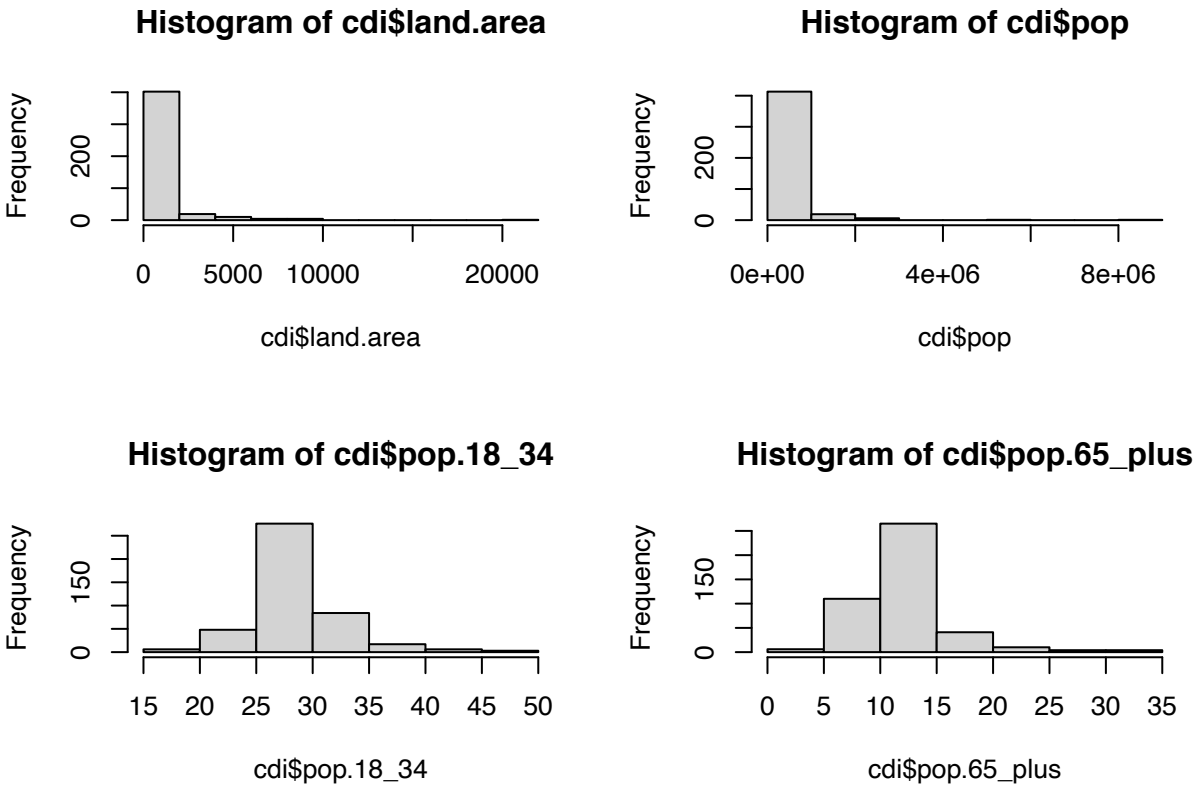
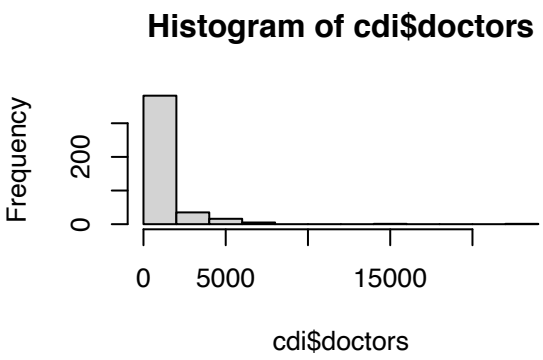
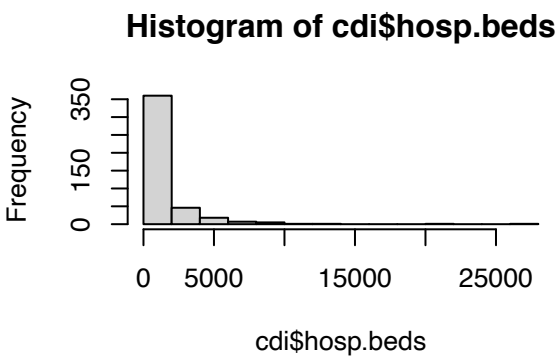


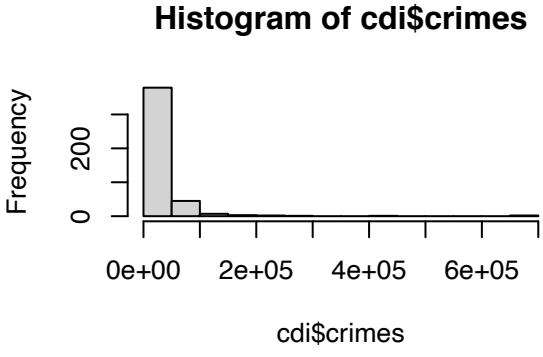
Table 1:

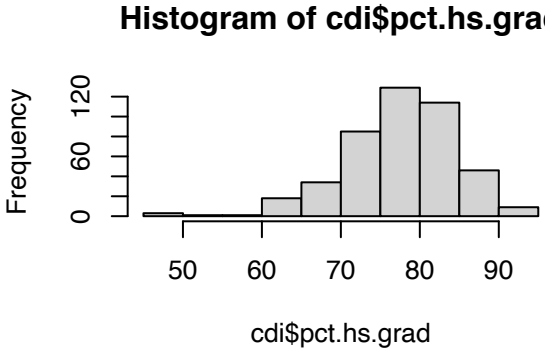
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
land.area	15.0	451.25	656.50	1041.41	946.75	20062.0	1549.92
pop	100043.0	139027.25	217280.50	393010.92	436064.50	8863164.0	601987.02
pop.18_34	16.4	26.20	28.10	28.57	30.02	49.7	4.19
pop.65_plus	3.0	9.88	11.75	12.17	13.62	33.8	3.99
doctors	39.0	182.75	401.00	988.00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

```
hist(cdi$doctors)
hist(cdi$hosp.beds)
hist(cdi$crimes)
hist(cdi$pct.hs.grad)
```

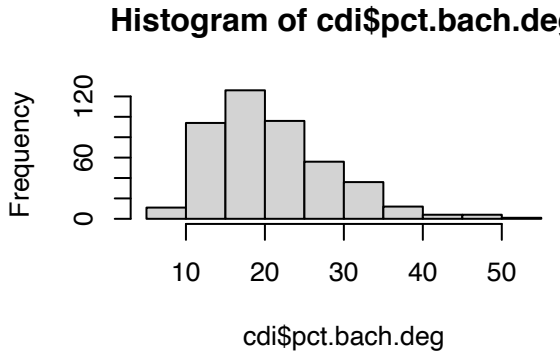


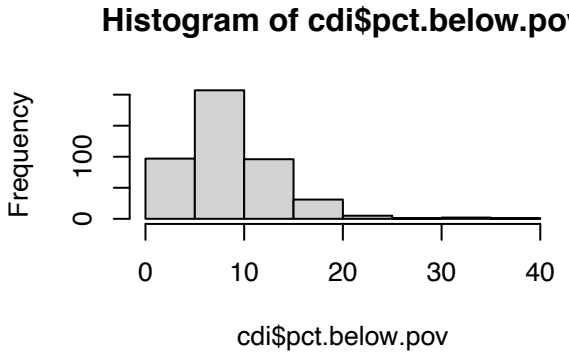


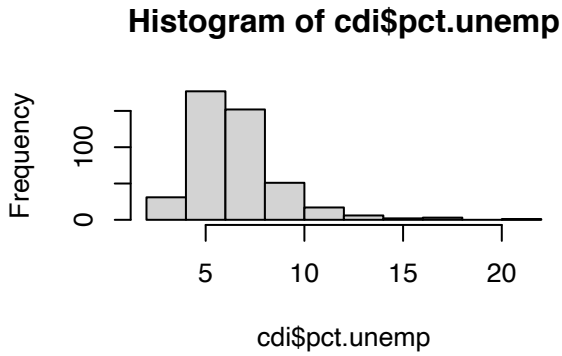


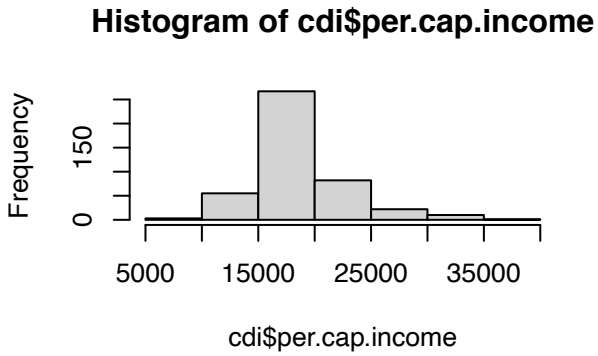



```
hist(cdi$pct.bach.deg)
hist(cdi$pct.below.pov)
hist(cdi$pct.unemp)
hist(cdi$per.cap.income)
```

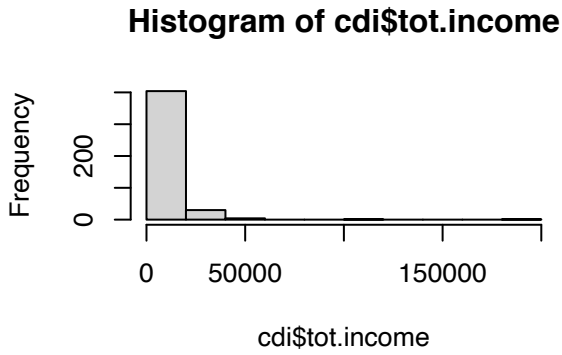






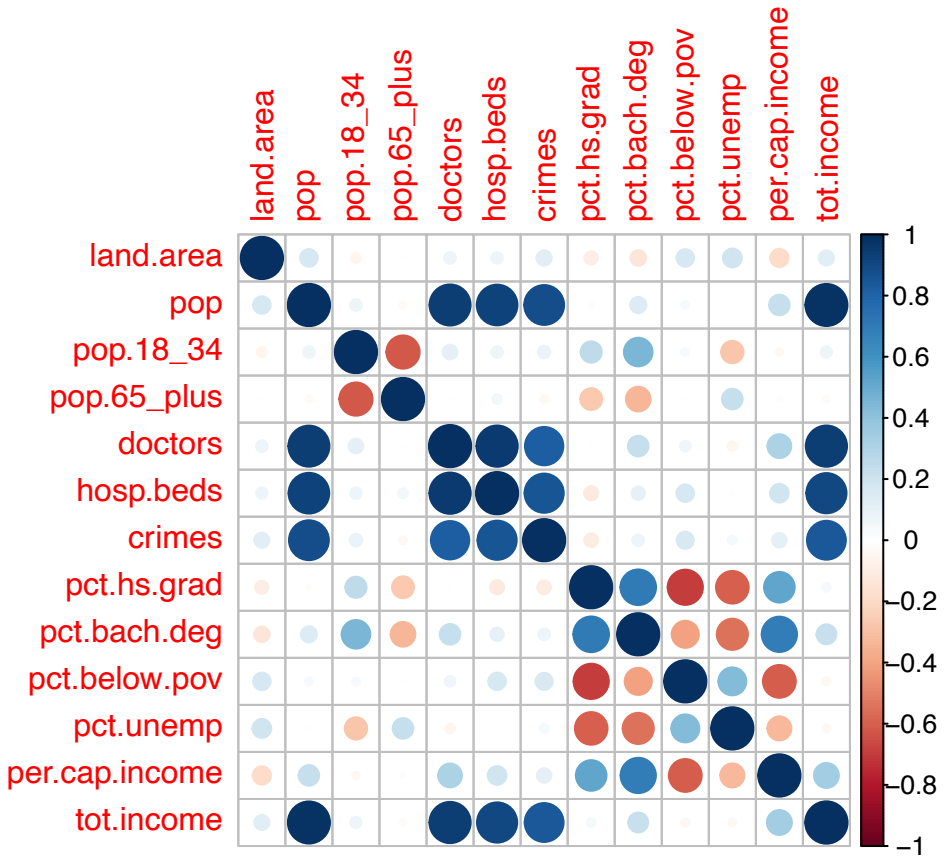



```
hist(cdi$tot.income)
```



correlation plot is used to analyze the relationship between each variables

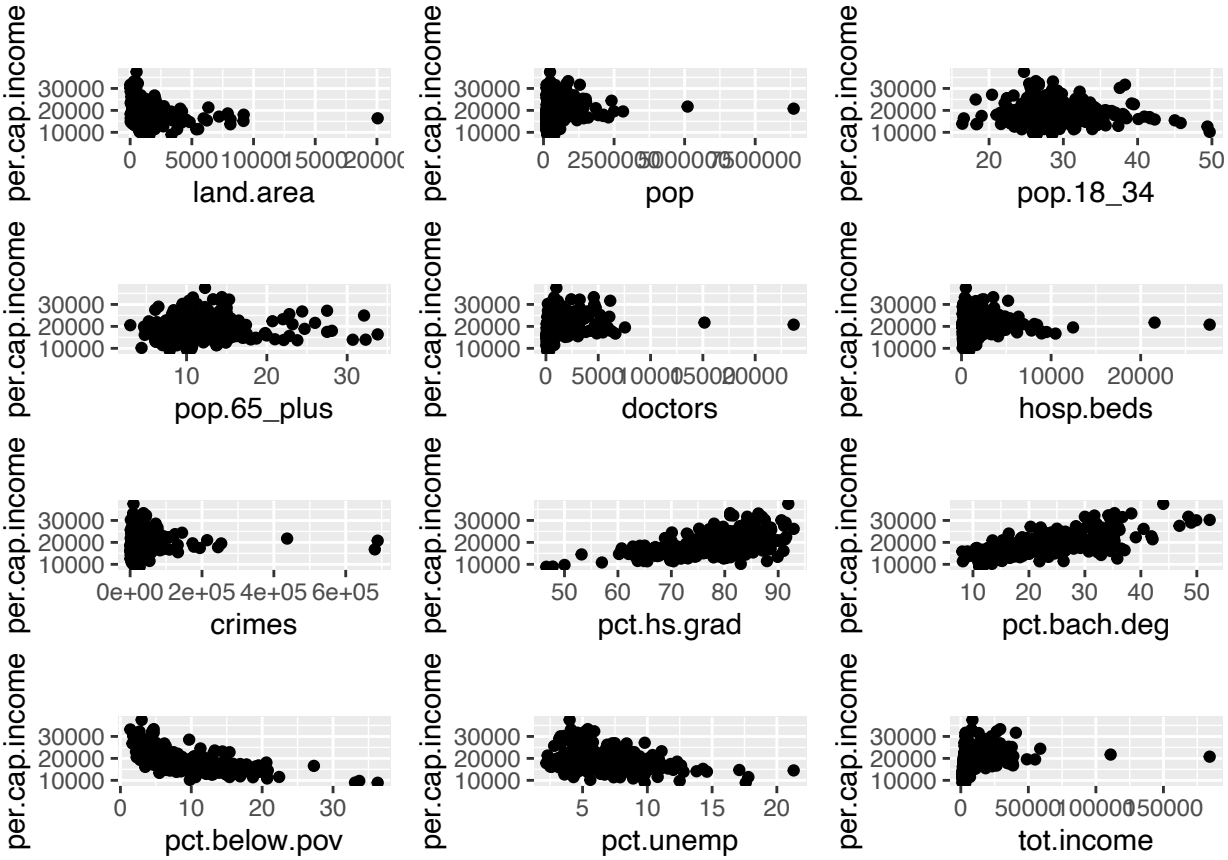
```
cdi_quan <- cdi[4:16]
C <-cor(cdi_quan)
corrplot(C, method="circle")
```



12 scatterplots are used to evaluate the relationship between each numeric variables and the response variable.

```
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(cdi_quan))
```



chuncks below are summary table for the categorcical variables.

```
region_frequency <- cdi%>% dplyr ::select(region)
t1 <- transform(region_frequency,region_Frequency=ave(seq(nrow(region_frequency)),region,FUN=length)) %
t1
```

```
##   region region_Frequency
## 1      W               77
## 2     NC              108
## 3      S              152
## 6     NE              103
```

```
county_frequency <- cdi%>% dplyr :: select(county)
t2 <- transform(county_frequency,county_Frequency=ave(seq(nrow(county_frequency)),county,FUN=length)) %
median(t2$county_Frequency)
```

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

```
state_frequency <- cdi%>% dplyr ::select(state)
t3 <- transform(state_frequency,state_Frequency=ave(seq(nrow(state_frequency)),state,FUN=length)) %>% u
median(t3$state_Frequency)
```

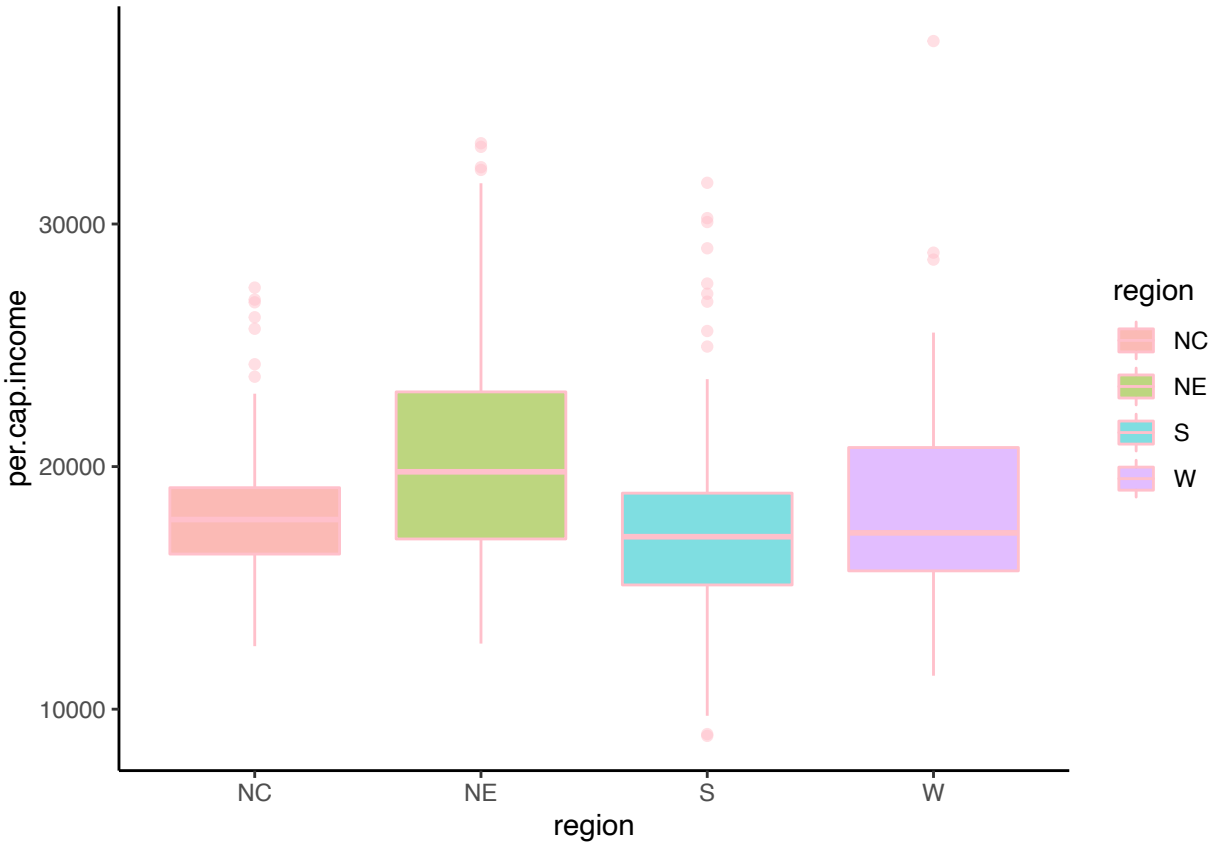
```
## [1] 7
```

```
which(is.na(cdi))
```

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

```
ggplot(cdi, aes(x=region, y=per.cap.income, fill=region)) +
    geom_boxplot(color = "pink", alpha=0.5) +
```

```
  theme_classic()
```

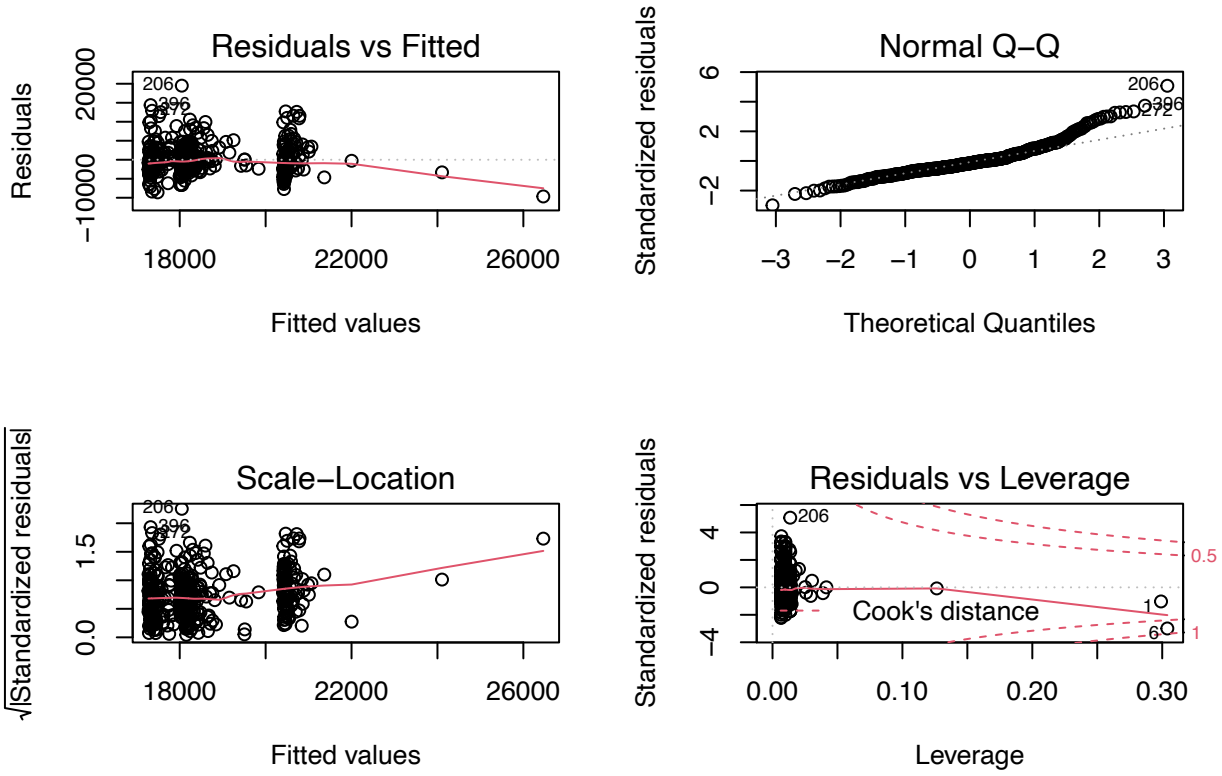


There is no NA value in the dataset. I plotted histograms for each quantitative variables and a correlation plot for all the quantitative variables. I also did a boxplot for the categorical variable region, and there are so many unique values in variables county and state, which makes it hard to plot boxplots. From the histograms, I find that land.area, pop, doctors, hosp.beds, crimes and tot.income are skewed to the right, which require further transformations. From the correlation plot, some variables are highly correlated, such as that tot.income with doctors, hosp.beds and crimes. Also, the response variable per.cap.income is calculated by dividing tot.income by pop, therefore, I will remove tot.income and pop when fitting the model.

crimes

crimes, crimes per capita and region are turned into four different combinations to predict per.cap.income.

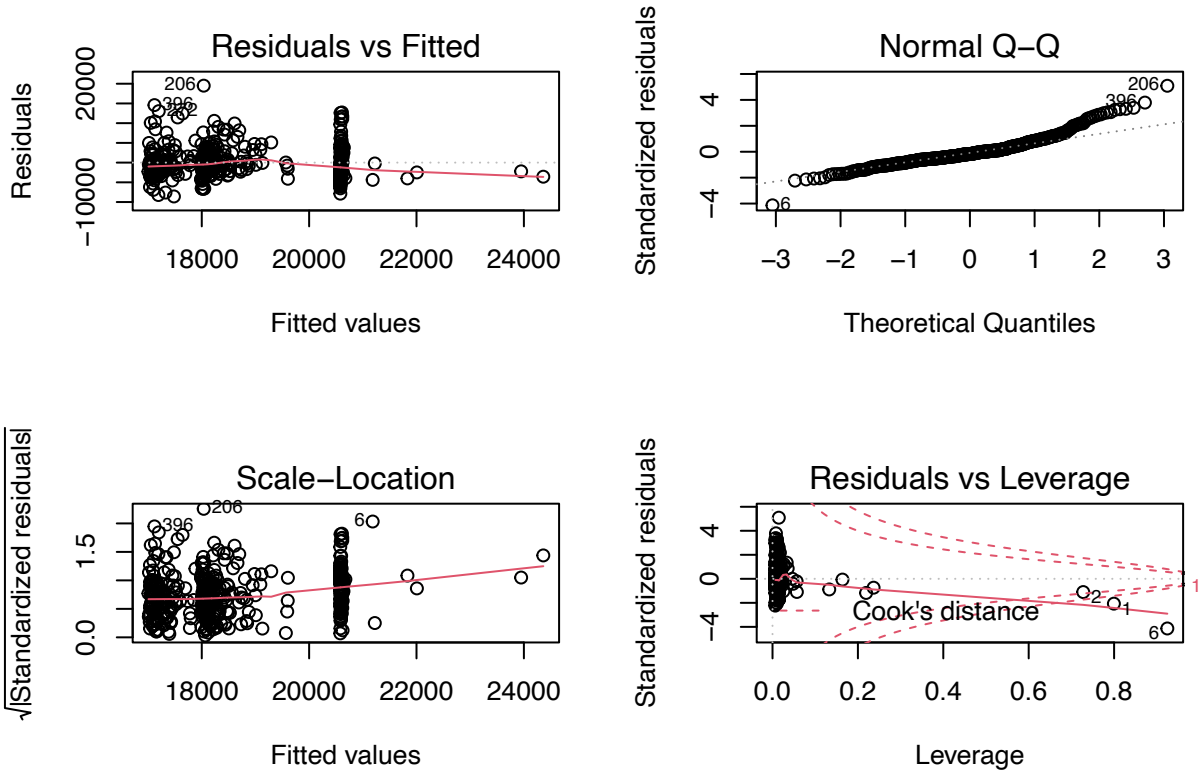
```
mod1 <- lm(per.cap.income~crimes + region, data = cdi)
par(mfrow=c(2,2))
plot(mod1)
```




```
summary(mod1)
```

```
##
## 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
```

```
mod2 <- lm(per.cap.income~crimes + region + crimes*region, data = cdi)
par(mfrow=c(2,2))
plot(mod2)
```




```
summary(mod2)
```

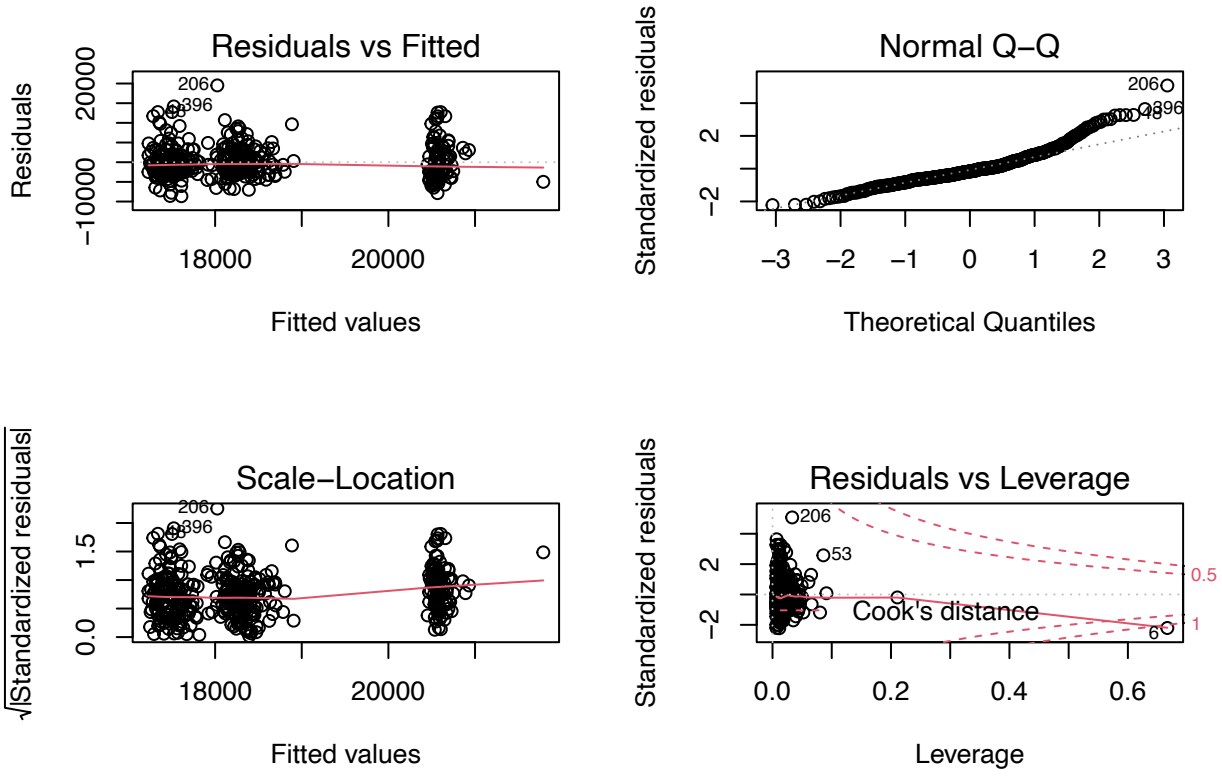
```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region + crimes * region,
##     data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8582.4 -2225.2  -676.2  1563.4 19504.7
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       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
```

```
anova(mod2, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ crimes + region + crimes * region
## Model 2: per.cap.income ~ crimes + region
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    432 6438799739
## 2    435 6501791845 -3 -62992106 1.4088 0.2396
```

After running an anova test between mod1 and mod2, the p-value is 0.2396, which is larger than 0.05, which cannot reject the null hypothesis that there is no difference between these two models. The test further shows that the interaction term between crimes and region is not significant.

```
cdi1 <- cdi %>% mutate(crimes_percapita = crimes/pop)
mod3 <- lm(per.cap.income~crimes_percapita + region + crimes_percapita*region, data = cdi1)
mod4<- lm(per.cap.income~crimes_percapita + region, data=cdi1)
par(mfrow=c(2,2))
plot(mod3)
```

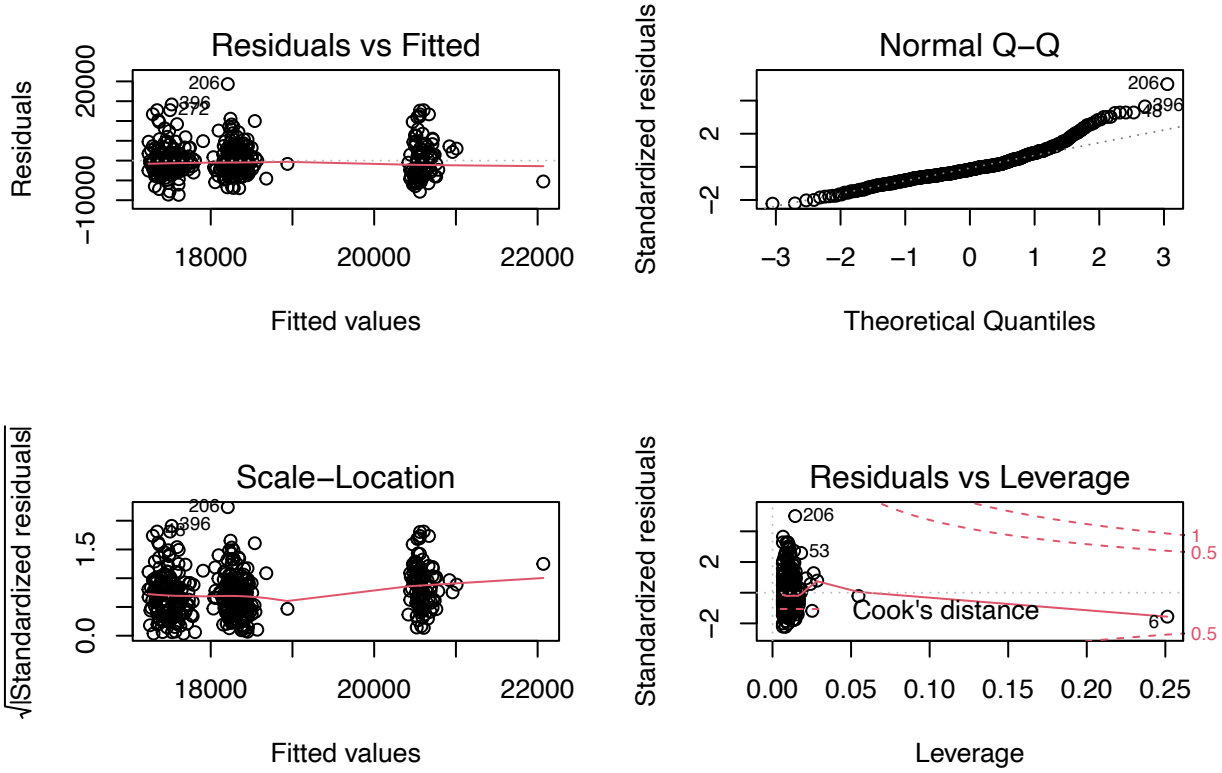



```
summary(mod3)
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes_percapita + region + crimes_percapita *
##     region, data = cdi1)
##
## 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 ***
## crimes_percapita              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
## crimes_percapita:regionNE      288.4    20184.7   0.014    0.989
## crimes_percapita:regionS      1558.9    20556.1   0.076    0.940
## crimes_percapita: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(mod4)
```




```
summary(mod4)
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes_percapita + region, data = cdi1)
##
## 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 ***
## crimes_percapita  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
```

```
anova(mod4, mod3)
```

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

For per-capita crime models, the anova test between the reduced model and the full model has a very high p-value of 0.9888, which cannot reject the null hypothesis that there is no difference between these two models. Similarly, the interaction term crimes_percapita * region is not significant.

From both anova tests, we find that the interaction term between crimes and region or crimes per capita and region is not significant. From the diagnostic plots for two models (one without interaction term and the other with the interaction term from cirmes per capita model), the model without the interaction term between crimes and region is better. These two models have very similar diagnostic plots, the redline is nearly horizontal and kind of random, which satisfies the independence and constant variance conditions. Two models have similar QQ plots as the upper tail is not following the diagonal line, which shows that the normality condition is not completely fulfilled. These two models also have similart scale vs location plots and Residuals vs Leverage plot.

Either crimes or crimes per capita is not significant predictor in either model from the summary table. The coefficient of variable crimes is a very small positive number, so the variable crimes will not really affect the per.cap.income. When I add another variable crimes_percapita to the model, there is a positive relationship between crimes_percapita and per.cap.income as the coefficient is as high as 5773.2. Therefore, it is enough to include only variables region and crimes per capita or crimes without the interaction term. When comparing two models (region + crimes and region + crimes per capita), I found that the model with region and crimes has a higher R squared of 0.09288, which is higher than 0.07782 of the model of region + crimes per capita. Therefore, I will use the model with predictors crimes and region, but crimes is still not very useful.

model fitting

```
cdi2 <- cdi1 %>% dplyr::select(-crimes, -id)
cdi2$land.area.per.capita <- cdi2$land.area/cdi2$pop
cdi2$pct.pop.65_plus <- cdi2$pop.65_plus/cdi2$pop
```

```
cdi2$pct.pop.18_34 <- cdi2$pop.18_34/cdi2$pop
cdi2$hosp.beds.capita <- cdi2$hosp.beds/cdi2$pop
cdi2$doctors_per_capita <- cdi2$doctors/cdi2$pop
cdi2 <- cdi2 %>% dplyr::select(-pop, -tot.income, -pop.65_plus, -pop.18_34, -hosp.beds, -doctors, -land
```

As I mentioned before, I will remove variables pop and tot.income from the model fitting. I will log all the skewed variables including per.cap.income. Also, since the response variables is income per capita, I will keep the indipendent variables to be per capita to keep the unit consistent. For example, I change the variable crimes, land.area, doctors, hosp.beds into crimes_per_capita, land.area per capita, doctors per capita, hosp.beds per capita since I thing it makes more sense to use per_capita variables in the model. I also transform pop.65_plus and pop.18_34 into percentage.

```
allsubset <- regsubsets(log(per.cap.income)~log(land.area.per.capita)  + pct.pop.18_34 + pct.pop.65_plus
summary<-summary(allsubset)
summary
```

```
## Subset selection object
## Call: regsubsets.formula(log(per.cap.income) ~ log(land.area.per.capita) +
##     pct.pop.18_34 + pct.pop.65_plus + log(doctors_per_capita) +
##     hosp.beds.capita + pct.hs.grad + log(pct.bach.deg) + pct.below.pov +
##     pct.unemp + log(crimes_percapita), data = cdi2)
## 10 Variables  (and intercept)
##                           Forced in Forced out
## log(land.area.per.capita)     FALSE      FALSE
## pct.pop.18_34                 FALSE      FALSE
## pct.pop.65_plus               FALSE      FALSE
## log(doctors_per_capita)       FALSE      FALSE
## hosp.beds.capita              FALSE      FALSE
## pct.hs.grad                   FALSE      FALSE
## log(pct.bach.deg)             FALSE      FALSE
## pct.below.pov                 FALSE      FALSE
## pct.unemp                     FALSE      FALSE
## log(crimes_percapita)         FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          log(land.area.per.capita) pct.pop.18_34 pct.pop.65_plus
## 1  ( 1 ) " "                       " "           " "
## 2  ( 1 ) " "                       " "           " "
## 3  ( 1 ) "*"                       " "           " "
## 4  ( 1 ) " "                       "*"           " "
## 5  ( 1 ) "*"                       " "           " "
## 6  ( 1 ) "*"                       "*"           " "
## 7  ( 1 ) "*"                       "*"           "*"
## 8  ( 1 ) "*"                       "*"           "*"
##          log(doctors_per_capita) hosp.beds.capita pct.hs.grad log(pct.bach.deg)
## 1  ( 1 ) " "                     " "               " "         "*"
## 2  ( 1 ) "*"                     " "               " "         " "
## 3  ( 1 ) " "                     " "               " "         "*"
## 4  ( 1 ) "*"                     " "               " "         "*"
## 5  ( 1 ) "*"                     " "               " "         "*"
## 6  ( 1 ) "*"                     " "               " "         "*"
## 7  ( 1 ) "*"                     " "               " "         "*"
## 8  ( 1 ) "*"                     " "               "*"         "*"
```

```
##          pct.below.pov pct.unemp log(crimes_percapita)
## 1  ( 1 ) " "           " "       " "
## 2  ( 1 ) "*"           " "       " "
## 3  ( 1 ) "*"           " "       " "
## 4  ( 1 ) "*"           " "       " "
## 5  ( 1 ) "*"           "*"       " "
## 6  ( 1 ) "*"           "*"       " "
## 7  ( 1 ) "*"           "*"       " "
## 8  ( 1 ) "*"           "*"       " "
```

```
data_frame (R2 = which.max(summary$adjr2), CP = which.min(summary$cp), BIC = which.min(summary$bic))
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## Please use `tibble()` instead.
```

```
## # A tibble: 1 x 3
##      R2    CP   BIC
##   <int> <int> <int>
## 1     8     8     8
```

All subsets regression chooses 8 variables to predict the per.cap.income.

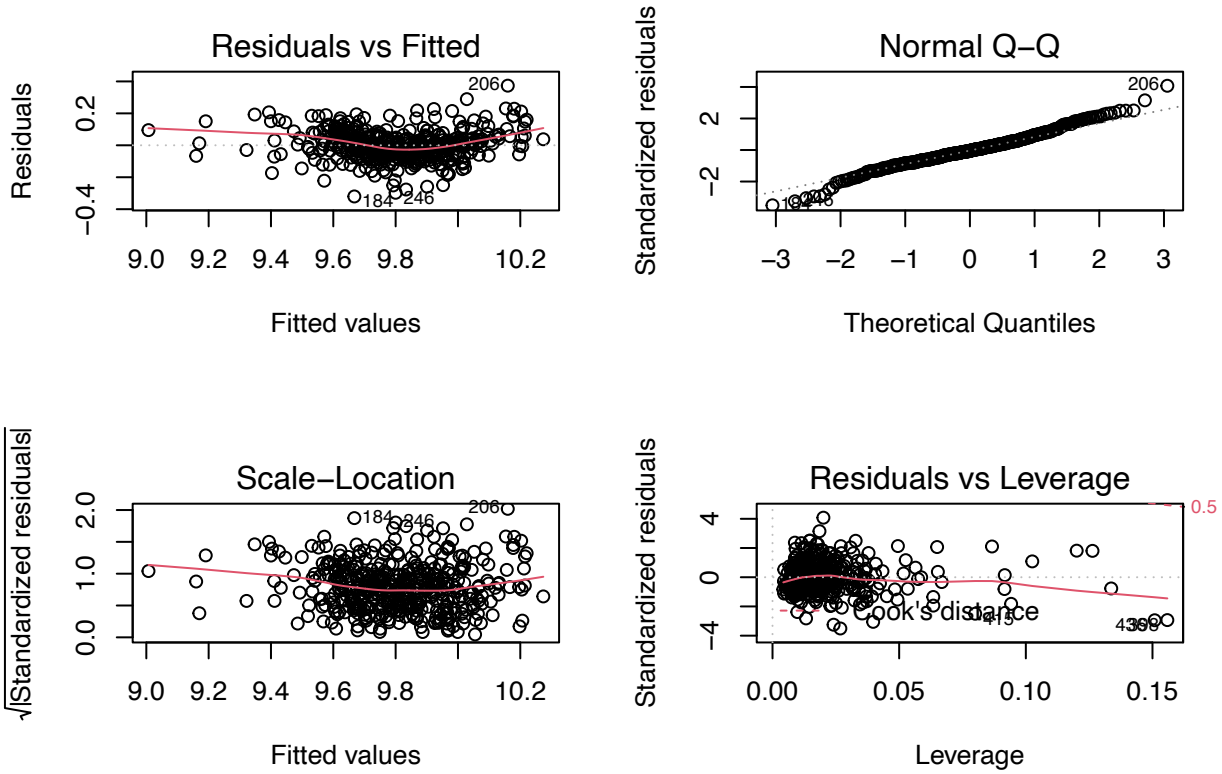
```
all.final1 <- lm(log(per.cap.income)~log(land.area.per.capita)  + pct.pop.18_34 + pct.pop.65_plus + log
summary(all.final1)$coef
```

```
##                                Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)                9.735822e+00 1.551815e-01  62.738273 7.504635e-219
## log(land.area.per.capita) -3.199924e-02 5.403586e-03  -5.921852  6.510612e-09
## pct.pop.18_34             -5.815082e+02 8.105052e+01  -7.174638  3.186427e-12
## pct.pop.65_plus            7.598345e+02 1.706300e+02   4.453112  1.079594e-05
## log(doctors_per_capita)    6.325965e-02 1.149944e-02   5.501106  6.483299e-08
## pct.hs.grad               -5.329863e-03 1.266842e-03  -4.207203  3.147335e-05
## log(pct.bach.deg)          2.939542e-01 2.518297e-02  11.672736  1.520685e-27
## pct.below.pov             -2.723836e-02 1.473078e-03 -18.490778  1.254607e-56
## pct.unemp                  1.349877e-02 2.527943e-03   5.339824  1.508595e-07
```

```
vif(all.final1)
```

```
## log(land.area.per.capita)            pct.pop.18_34           pct.pop.65_plus
##                  2.016549                 2.401883                 2.683723
##   log(doctors_per_capita)               pct.hs.grad        log(pct.bach.deg)
##                  2.096849                 4.048137                 4.083645
##             pct.below.pov                pct.unemp
##                  2.411847                 1.790318
```

```
par(mfrow=c(2,2))
plot(all.final1)
```



No collinearity found in the model. All the diagnostic plots show that the model is a good fit. Except that there is a slight curve in the residual plot and QQ plot has a lower tail. There is no bad leverage point on the residuals vs leverage plot.

```
fullmod <- lm(log(per.cap.income)~log(land.area.per.capita)  + pct.pop.18_34 + pct.pop.65_plus + log(do

step_model <- stepAIC(fullmod, direction = "both",
                      trace = FALSE)

summary(step_model)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area.per.capita) +
##     pct.pop.18_34 + pct.pop.65_plus + log(doctors_per_capita) +
##     pct.hs.grad + log(pct.bach.deg) + pct.below.pov + pct.unemp,
##     data = cdi2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32030 -0.05650 -0.00460  0.05037  0.37335
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 9.736e+00  1.552e-01  62.738  < 2e-16 ***
## log(land.area.per.capita) -3.200e-02  5.404e-03  -5.922 6.51e-09 ***
## pct.pop.18_34             -5.815e+02  8.105e+01  -7.175 3.19e-12 ***
## pct.pop.65_plus            7.598e+02  1.706e+02   4.453 1.08e-05 ***
## log(doctors_per_capita)    6.326e-02  1.150e-02   5.501 6.48e-08 ***
## pct.hs.grad               -5.330e-03  1.267e-03  -4.207 3.15e-05 ***
```

```
## log(pct.bach.deg)          2.940e-01  2.518e-02  11.673  < 2e-16 ***
## pct.below.pov             -2.724e-02  1.473e-03 -18.491  < 2e-16 ***
## pct.unemp                  1.350e-02  2.528e-03   5.340 1.51e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09255 on 431 degrees of freedom
## Multiple R-squared:  0.8032, Adjusted R-squared:  0.7996
## F-statistic: 219.9 on 8 and 431 DF,  p-value: < 2.2e-16
```

stepwise regression chooses the same 8 variables, which is the same model as the all subsets model.

```
cdi2$log.per.cap.income <- log(cdi2$per.cap.income)
cdi2$log.land.area.per.capita <- log(cdi2$land.area.per.capita)
cdi2$log.doctors.per.capita <- log(cdi2$doctors_per_capita)
cdi2$log.pct.bach.deg <- log(cdi2$pct.bach.deg)
cdi3 <- cdi2 %>% dplyr::select(-c(per.cap.income, land.area.per.capita, doctors_per_capita, pct.bach.deg
```

```
all.region <- lm(log.per.cap.income ~ .*region,data=cdi3)
summary(all.region)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ . * region, data = cdi3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23813 -0.04574 -0.00719  0.04413  0.33462
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      9.082e+00  3.882e-01  23.392  < 2e-16 ***
## pct.hs.grad                      3.322e-04  3.920e-03   0.085 0.932517
## pct.below.pov                   -2.619e-02  4.256e-03  -6.153 1.83e-09 ***
## pct.unemp                        2.176e-02  6.173e-03   3.525 0.000472 ***
## regionNE                        -6.081e-02  5.216e-01  -0.117 0.907242
## regionS                          6.093e-01  4.477e-01   1.361 0.174272
## regionW                          2.518e+00  5.564e-01   4.525 7.94e-06 ***
## pct.pop.65_plus                  1.796e+03  5.711e+02   3.145 0.001786 **
## pct.pop.18_34                   -8.546e+02  2.152e+02  -3.971 8.46e-05 ***
## log.land.area.per.capita        -5.240e-02  1.671e-02  -3.135 0.001843 **
## log.doctors.per.capita           3.766e-02  2.091e-02   1.801 0.072478 .
## log.pct.bach.deg                 2.430e-01  6.980e-02   3.481 0.000553 ***
## pct.hs.grad:regionNE            -3.449e-03  5.107e-03  -0.675 0.499939
## pct.hs.grad:regionS             -6.920e-03  4.402e-03  -1.572 0.116708
## pct.hs.grad:regionW             -1.848e-02  5.260e-03  -3.514 0.000492 ***
## pct.below.pov:regionNE          -9.971e-04  5.936e-03  -0.168 0.866695
## pct.below.pov:regionS            3.372e-03  4.828e-03   0.699 0.485258
## pct.below.pov:regionW           -1.698e-02  6.298e-03  -2.696 0.007312 **
## pct.unemp:regionNE              -1.732e-02  8.910e-03  -1.943 0.052653 .
## pct.unemp:regionS               -1.936e-02  8.236e-03  -2.351 0.019228 *
## pct.unemp:regionW               -1.761e-02  8.178e-03  -2.153 0.031904 *
## regionNE:pct.pop.65_plus        -9.544e+01  7.835e+02  -0.122 0.903108
## regionS:pct.pop.65_plus         -8.725e+02  6.106e+02  -1.429 0.153813
## regionW:pct.pop.65_plus         -1.737e+03  7.277e+02  -2.387 0.017444 *
```

```
## regionNE:pct.pop.18_34            -1.926e+02  3.201e+02  -0.602 0.547791
## regionS:pct.pop.18_34              3.149e+02  2.429e+02   1.296 0.195596
## regionW:pct.pop.18_34              8.098e+02  3.217e+02   2.517 0.012222 *
## regionNE:log.land.area.per.capita  8.029e-03  2.218e-02   0.362 0.717585
## regionS:log.land.area.per.capita   2.454e-02  1.948e-02   1.260 0.208575
## regionW:log.land.area.per.capita   3.948e-02  2.085e-02   1.894 0.058973 .
## regionNE:log.doctors.per.capita   -2.345e-02  3.552e-02  -0.660 0.509603
## regionS:log.doctors.per.capita     2.483e-02  2.766e-02   0.898 0.369972
## regionW:log.doctors.per.capita     1.163e-01  4.023e-02   2.891 0.004048 **
## regionNE:log.pct.bach.deg          1.255e-01  1.019e-01   1.231 0.218973
## regionS:log.pct.bach.deg           1.014e-01  8.081e-02   1.255 0.210249
## regionW:log.pct.bach.deg           5.818e-02  9.198e-02   0.633 0.527361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08602 on 404 degrees of freedom
## Multiple R-squared:  0.8407, Adjusted R-squared:  0.8269
## F-statistic: 60.91 on 35 and 404 DF,  p-value: < 2.2e-16
```

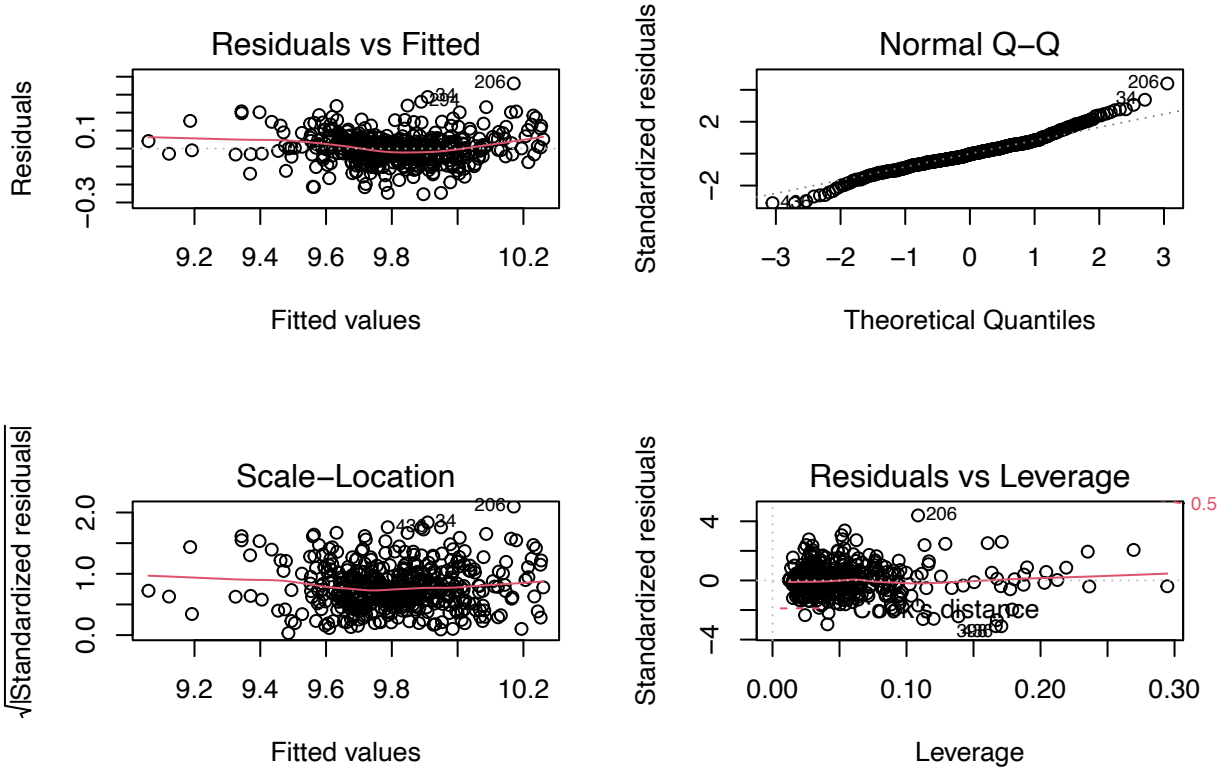
A model with all interaction terms between region and all quantitative variables is built to find significant interaction terms. Four interaction terms pct.hs.grad and region, pct.below.pov and region, log.doctors.per.capita and region + pct.unemp and region are significant. Therefore, I decided to add them in the model.

```
final.model <- lm(log.per.cap.income~pct.below.pov + pct.unemp + region + pct.pop.65_plus + pct.pop.18_3
summary(final.model)$coeff
```

```
##                                     Estimate   Std. Error      t value
## (Intercept)                     9.429525e+00 3.077012e-01 30.645073549
## pct.below.pov                  -2.747203e-02 4.238282e-03 -6.481878691
## pct.unemp                       2.463618e-02 5.677918e-03  4.338945826
## regionNE                        1.332848e-01 3.895449e-01  0.342155175
## regionS                         1.980384e-01 3.417511e-01  0.579481283
## regionW                         1.595579e+00 4.387290e-01  3.636820853
## pct.pop.65_plus                 1.034913e+03 1.738661e+02  5.952355263
## pct.pop.18_34                  -6.444283e+02 8.251064e+01 -7.810245023
## log.land.area.per.capita       -3.341141e-02 6.235304e-03 -5.358424545
## log.pct.bach.deg                3.161447e-01 2.738703e-02 11.543594833
## pct.hs.grad                    -5.455958e-03 2.924378e-03 -1.865681264
## log.doctors.per.capita          3.447610e-02 1.746121e-02  1.974439783
## regionNE:pct.hs.grad            1.887734e-03 3.465985e-03  0.544645682
## regionS:pct.hs.grad             1.194818e-05 3.037565e-03  0.003933473
## regionW:pct.hs.grad            -1.180117e-02 4.069307e-03 -2.900044013
## pct.below.pov:regionNE         -1.619308e-03 5.784708e-03 -0.279929094
## pct.below.pov:regionS           6.060271e-03 4.700938e-03  1.289161911
## pct.below.pov:regionW          -1.132848e-02 6.183805e-03 -1.831959958
## regionNE:log.doctors.per.capita 2.569232e-02 2.663105e-02  0.964750426
## regionS:log.doctors.per.capita  2.119632e-02 2.196130e-02  0.965166759
## regionW:log.doctors.per.capita  6.787510e-02 3.060417e-02  2.217838267
## pct.unemp:regionNE             -1.893370e-02 8.594333e-03 -2.203045262
## pct.unemp:regionS              -2.557064e-02 7.542125e-03 -3.390375884
## pct.unemp:regionW              -2.005248e-02 7.864561e-03 -2.549726443
##                                     Pr(>|t|)
## (Intercept)                    9.801801e-109
## pct.below.pov                   2.568108e-10
```

```
## pct.unemp                          1.799206e-05
## regionNE                           7.324069e-01
## regionS                            5.625781e-01
## regionW                            3.106677e-04
## pct.pop.65_plus                    5.620733e-09
## pct.pop.18_34                      4.693662e-14
## log.land.area.per.capita           1.393375e-07
## log.pct.bach.deg                   6.237038e-27
## pct.hs.grad                        6.278960e-02
## log.doctors.per.capita             4.899274e-02
## regionNE:pct.hs.grad               5.862891e-01
## regionS:pct.hs.grad                9.968634e-01
## regionW:pct.hs.grad                3.928905e-03
## pct.below.pov:regionNE             7.796710e-01
## pct.below.pov:regionS              1.980582e-01
## pct.below.pov:regionW              6.767218e-02
## regionNE:log.doctors.per.capita    3.352303e-01
## regionS:log.doctors.per.capita     3.350220e-01
## regionW:log.doctors.per.capita     2.710539e-02
## pct.unemp:regionNE                 2.813911e-02
## pct.unemp:regionS                  7.646562e-04
## pct.unemp:regionW                  1.113847e-02
```

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




```
summary(final.model)
```

```
##
## Call:
```

```
## lm(formula = log.per.cap.income ~ pct.below.pov + pct.unemp +
##     region + pct.pop.65_plus + pct.pop.18_34 + log.land.area.per.capita +
##     log.pct.bach.deg + pct.hs.grad * region + pct.below.pov *
##     region + log.doctors.per.capita * region + pct.unemp * region,
##     data = cdi3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25410 -0.04740 -0.00391  0.04693  0.36256
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     9.430e+00  3.077e-01  30.645  < 2e-16 ***
## pct.below.pov                  -2.747e-02  4.238e-03  -6.482 2.57e-10 ***
## pct.unemp                       2.464e-02  5.678e-03   4.339 1.80e-05 ***
## regionNE                        1.333e-01  3.895e-01   0.342 0.732407
## regionS                         1.980e-01  3.418e-01   0.579 0.562578
## regionW                         1.596e+00  4.387e-01   3.637 0.000311 ***
## pct.pop.65_plus                 1.035e+03  1.739e+02   5.952 5.62e-09 ***
## pct.pop.18_34                  -6.444e+02  8.251e+01  -7.810 4.69e-14 ***
## log.land.area.per.capita       -3.341e-02  6.235e-03  -5.358 1.39e-07 ***
## log.pct.bach.deg                3.161e-01  2.739e-02  11.544  < 2e-16 ***
## pct.hs.grad                    -5.456e-03  2.924e-03  -1.866 0.062790 .
## log.doctors.per.capita          3.448e-02  1.746e-02   1.974 0.048993 *
## regionNE:pct.hs.grad            1.888e-03  3.466e-03   0.545 0.586289
## regionS:pct.hs.grad             1.195e-05  3.038e-03   0.004 0.996863
## regionW:pct.hs.grad            -1.180e-02  4.069e-03  -2.900 0.003929 **
## pct.below.pov:regionNE         -1.619e-03  5.785e-03  -0.280 0.779671
## pct.below.pov:regionS           6.060e-03  4.701e-03   1.289 0.198058
## pct.below.pov:regionW          -1.133e-02  6.184e-03  -1.832 0.067672 .
## regionNE:log.doctors.per.capita  2.569e-02  2.663e-02   0.965 0.335230
## regionS:log.doctors.per.capita   2.120e-02  2.196e-02   0.965 0.335022
## regionW:log.doctors.per.capita   6.788e-02  3.060e-02   2.218 0.027105 *
## pct.unemp:regionNE             -1.893e-02  8.594e-03  -2.203 0.028139 *
## pct.unemp:regionS              -2.557e-02  7.542e-03  -3.390 0.000765 ***
## pct.unemp:regionW              -2.005e-02  7.865e-03  -2.550 0.011138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0874 on 416 degrees of freedom
## Multiple R-squared:  0.8306, Adjusted R-squared:  0.8213
## F-statistic: 88.71 on 23 and 416 DF,  p-value: < 2.2e-16
```

After adding these four interaction terms, I find that there is less curve in the residual plot and the lower tail of QQ plot is also not obivious any more. The red line in the scale-location plot is horizontal, which shows that the constant variance condition is satisfied. The residual vs Leverage plot also does not show any bad influential point. The R sqaured is 0.8213, which is higher than 0.7996 of the original model with all quantitative variables.

cross validation

```
data_cdi <-trainControl(method = "cv", number = 5)
model_caret <- train(log.per.cap.income~pct.below.pov + pct.unemp + region + pct.pop.65_plus + pct.pop.
                     trControl = data_cdi,              # folds
```

```
                      method = "lm",                      # specifying regression model
                      na.action = na.pass)
```

```
print(model_caret)
```

```
## Linear Regression
##
## 440 samples
##   9 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 352, 352, 352, 352, 352
## Resampling results:
##
##   RMSE        Rsquared   MAE
##   0.09150005  0.8041748  0.06801731
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

A five-fold cross validation is performed and the R squared is 0.7971, which is not very different from 0.8213, which shows that overfitting is not a great problm. RMSE is only 0.09413, which is also very close to the standard error in the final model, which is 0.0874. All these prove that the model does not have an overfitting problem and the prediction error is pretty small.

```
cdi4 <- cdi2 %>% dplyr::select(-c(per.cap.income, land.area.per.capita, doctors_per_capita, pct.bach.deg
state_interaction <- lm(log.per.cap.income ~ .*state,data=cdi4)
summary(state_interaction)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ . * state, data = cdi4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19966 -0.01225  0.00000  0.01065  0.21148
##
## Coefficients: (152 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.218e+01  3.637e+00   3.348  0.00101 **
## stateAR                        -2.153e+00  3.400e+00  -0.633  0.52753
## stateAZ                         9.333e+00  1.018e+01   0.917  0.36065
## stateCA                        -1.868e+00  3.735e+00  -0.500  0.61763
## stateCO                        -1.143e+01  1.082e+01  -1.057  0.29225
## stateCT                         1.527e+00  1.759e+01   0.087  0.93093
## stateDC                         1.708e+00  1.340e+00   1.274  0.20434
## stateDE                         1.813e+00  2.819e+00   0.643  0.52113
## stateFL                        -2.010e-01  3.728e+00  -0.054  0.95707
## stateGA                         2.914e+00  3.099e+01   0.094  0.92520
## stateHI                        -6.482e+00  1.042e+01  -0.622  0.53459
## stateID                        -4.672e-01  4.322e-01  -1.081  0.28137
## stateIL                        -3.378e+00  3.969e+00  -0.851  0.39607
## stateIN                        -2.238e+00  4.401e+00  -0.508  0.61185
## stateKS                         3.331e+01  3.233e+01   1.030  0.30443
## stateKY                        -1.910e+00  3.365e+00  -0.568  0.57103
```

```
## stateLA                          2.732e+00  6.213e+00   0.440  0.66072
## stateMA                         -2.167e+00  3.824e+00  -0.566  0.57185
## stateMD                          1.278e+01  9.862e+00   1.296  0.19691
## stateME                         -1.102e+02  9.020e+01  -1.222  0.22362
## stateMI                         -2.833e+00  3.894e+00  -0.728  0.46790
## stateMN                         -7.229e+00  9.792e+00  -0.738  0.46145
## stateMO                         -5.247e+00  8.487e+00  -0.618  0.53728
## stateMS                          6.917e+01  1.275e+02   0.542  0.58833
## stateMT                         -8.588e-01  8.449e-01  -1.016  0.31094
## stateNC                         -2.772e+00  3.748e+00  -0.739  0.46074
## stateND                          1.838e-01  7.558e-01   0.243  0.80821
## stateNE                          1.312e+02  1.066e+02   1.231  0.22022
## stateNH                         -3.277e+01  3.118e+01  -1.051  0.29481
## stateNJ                         -3.041e+00  3.957e+00  -0.768  0.44340
## stateNM                          7.059e-01  1.839e+00   0.384  0.70164
## stateNV                         -1.299e+00  4.084e+00  -0.318  0.75077
## stateNY                         -1.774e+00  4.018e+00  -0.441  0.65948
## stateOH                         -6.307e-01  4.020e+00  -0.157  0.87553
## stateOK                         -1.433e+01  1.514e+01  -0.947  0.34507
## stateOR                         -4.313e+01  4.423e+01  -0.975  0.33104
## statePA                         -3.079e+00  3.690e+00  -0.835  0.40521
## stateRI                         -2.643e+00  4.733e+00  -0.558  0.57730
## stateSC                         -2.588e+00  4.834e+00  -0.535  0.59310
## stateSD                         -3.039e-01  4.790e-01  -0.635  0.52662
## stateTN                         -1.085e+00  6.054e+00  -0.179  0.85794
## stateTX                         -1.656e+00  3.730e+00  -0.444  0.65777
## stateUT                          5.141e+00  6.592e+00   0.780  0.43659
## stateVA                         -2.968e+00  4.843e+00  -0.613  0.54087
## stateVT                          9.756e-01  9.124e-01   1.069  0.28659
## stateWA                         -3.906e+00  5.056e+00  -0.772  0.44101
## stateWI                         -4.130e+00  5.012e+00  -0.824  0.41114
## stateWV                         -3.862e-01  4.064e-01  -0.950  0.34338
## pct.hs.grad                     -6.471e-03  4.010e-02  -0.161  0.87200
## pct.below.pov                   -4.252e-02  4.052e-02  -1.049  0.29560
## pct.unemp                       -1.092e-01  8.183e-02  -1.335  0.18377
## pct.pop.65_plus                  1.639e+04  1.049e+04   1.561  0.12041
## pct.pop.18_34                   -1.020e+04  7.238e+03  -1.409  0.16071
## log.land.area.per.capita         3.924e-01  4.204e-01   0.934  0.35196
## log.doctors.per.capita          -1.692e-01  2.268e-01  -0.746  0.45678
## log.pct.bach.deg                 2.946e-01  4.706e-01   0.626  0.53226
## stateAR:pct.hs.grad              2.812e-02  4.440e-02   0.633  0.52750
## stateAZ:pct.hs.grad             -1.188e-01  1.147e-01  -1.036  0.30183
## stateCA:pct.hs.grad              1.634e-03  4.059e-02   0.040  0.96794
## stateCO:pct.hs.grad              5.364e-02  7.791e-02   0.688  0.49214
## stateCT:pct.hs.grad             -6.516e-02  1.351e-01  -0.482  0.63019
## stateDC:pct.hs.grad                     NA         NA      NA       NA
## stateDE:pct.hs.grad             -1.900e-02  3.466e-02  -0.548  0.58440
## stateFL:pct.hs.grad             -7.593e-03  4.112e-02  -0.185  0.85374
## stateGA:pct.hs.grad             -9.783e-02  3.615e-01  -0.271  0.78702
## stateHI:pct.hs.grad              8.989e-02  1.387e-01   0.648  0.51792
## stateID:pct.hs.grad                     NA         NA      NA       NA
## stateIL:pct.hs.grad             -6.901e-03  4.271e-02  -0.162  0.87182
## stateIN:pct.hs.grad              2.006e-02  4.273e-02   0.469  0.63937
## stateKS:pct.hs.grad             -3.473e-01  3.300e-01  -1.053  0.29414
```

```
## stateKY:pct.hs.grad              2.970e-02  5.602e-02   0.530  0.59666
## stateLA:pct.hs.grad             -2.464e-02  5.777e-02  -0.426  0.67039
## stateMA:pct.hs.grad              1.344e-02  4.120e-02   0.326  0.74465
## stateMD:pct.hs.grad             -4.540e-02  5.115e-02  -0.888  0.37609
## stateME:pct.hs.grad              1.093e+00  8.986e-01   1.217  0.22557
## stateMI:pct.hs.grad              1.646e-02  4.271e-02   0.385  0.70055
## stateMN:pct.hs.grad              9.087e-02  1.161e-01   0.783  0.43499
## stateMO:pct.hs.grad              5.882e-03  5.135e-02   0.115  0.90896
## stateMS:pct.hs.grad             -9.785e-01  1.804e+00  -0.542  0.58838
## stateMT:pct.hs.grad                     NA         NA      NA       NA
## stateNC:pct.hs.grad             -7.847e-03  4.052e-02  -0.194  0.84669
## stateND:pct.hs.grad                     NA         NA      NA       NA
## stateNE:pct.hs.grad             -1.352e+00  1.095e+00  -1.235  0.21880
## stateNH:pct.hs.grad              3.302e-01  3.443e-01   0.959  0.33894
## stateNJ:pct.hs.grad              2.071e-03  4.286e-02   0.048  0.96153
## stateNM:pct.hs.grad             -1.044e-02  2.344e-02  -0.445  0.65661
## stateNV:pct.hs.grad              5.967e-03  5.394e-02   0.111  0.91205
## stateNY:pct.hs.grad             -1.297e-02  4.221e-02  -0.307  0.75904
## stateOH:pct.hs.grad             -2.517e-03  4.139e-02  -0.061  0.95158
## stateOK:pct.hs.grad              1.751e-01  1.950e-01   0.898  0.37049
## stateOR:pct.hs.grad              5.051e-01  5.202e-01   0.971  0.33310
## statePA:pct.hs.grad              5.497e-03  4.064e-02   0.135  0.89259
## stateRI:pct.hs.grad              3.057e-02  5.694e-02   0.537  0.59211
## stateSC:pct.hs.grad             -8.396e-03  4.391e-02  -0.191  0.84860
## stateSD:pct.hs.grad                     NA         NA      NA       NA
## stateTN:pct.hs.grad             -6.773e-02  6.558e-02  -1.033  0.30329
## stateTX:pct.hs.grad             -6.529e-03  4.111e-02  -0.159  0.87402
## stateUT:pct.hs.grad             -4.226e-02  5.625e-02  -0.751  0.45354
## stateVA:pct.hs.grad             -2.310e-02  5.434e-02  -0.425  0.67134
## stateVT:pct.hs.grad                     NA         NA      NA       NA
## stateWA:pct.hs.grad             -1.156e-02  5.157e-02  -0.224  0.82290
## stateWI:pct.hs.grad              5.948e-03  3.732e-02   0.159  0.87357
## stateWV:pct.hs.grad                     NA         NA      NA       NA
## stateAR:pct.below.pov                   NA         NA      NA       NA
## stateAZ:pct.below.pov           -8.777e-02  1.121e-01  -0.783  0.43465
## stateCA:pct.below.pov            1.105e-02  4.135e-02   0.267  0.78969
## stateCO:pct.below.pov            2.829e-01  3.002e-01   0.942  0.34753
## stateCT:pct.below.pov           -7.778e-02  1.648e-01  -0.472  0.63760
## stateDC:pct.below.pov                   NA         NA      NA       NA
## stateDE:pct.below.pov                   NA         NA      NA       NA
## stateFL:pct.below.pov           -2.079e-02  4.191e-02  -0.496  0.62065
## stateGA:pct.below.pov           -2.674e-02  1.746e-01  -0.153  0.87844
## stateHI:pct.below.pov           -1.819e-01  1.556e-01  -1.169  0.24424
## stateID:pct.below.pov                   NA         NA      NA       NA
## stateIL:pct.below.pov            2.186e-02  4.285e-02   0.510  0.61059
## stateIN:pct.below.pov            1.250e-02  4.473e-02   0.280  0.78017
## stateKS:pct.below.pov           -7.097e-01  6.065e-01  -1.170  0.24372
## stateKY:pct.below.pov            1.775e-02  8.540e-02   0.208  0.83563
## stateLA:pct.below.pov            3.877e-02  4.426e-02   0.876  0.38237
## stateMA:pct.below.pov            1.778e-02  4.526e-02   0.393  0.69500
## stateMD:pct.below.pov           -1.397e-01  1.089e-01  -1.282  0.20156
## stateME:pct.below.pov           -5.739e-01  4.980e-01  -1.152  0.25092
## stateMI:pct.below.pov            1.830e-02  4.136e-02   0.442  0.65873
## stateMN:pct.below.pov            3.974e-01  3.842e-01   1.034  0.30261
```

```
## stateMO:pct.below.pov           7.330e-02  1.647e-01   0.445  0.65683
## stateMS:pct.below.pov           2.765e-01  5.026e-01   0.550  0.58304
## stateMT:pct.below.pov                  NA         NA      NA       NA
## stateNC:pct.below.pov           1.859e-02  4.189e-02   0.444  0.65778
## stateND:pct.below.pov                  NA         NA      NA       NA
## stateNE:pct.below.pov          -2.082e+00  1.695e+00  -1.228  0.22120
## stateNH:pct.below.pov           1.475e+00  1.334e+00   1.106  0.27058
## stateNJ:pct.below.pov           4.560e-03  4.591e-02   0.099  0.92099
## stateNM:pct.below.pov                  NA         NA      NA       NA
## stateNV:pct.below.pov                  NA         NA      NA       NA
## stateNY:pct.below.pov          -1.359e-02  4.328e-02  -0.314  0.75386
## stateOH:pct.below.pov           7.716e-03  4.269e-02   0.181  0.85681
## stateOK:pct.below.pov           2.810e-01  3.283e-01   0.856  0.39339
## stateOR:pct.below.pov           1.273e+00  1.306e+00   0.975  0.33126
## statePA:pct.below.pov           2.370e-02  4.099e-02   0.578  0.56403
## stateRI:pct.below.pov           1.242e-01  1.154e-01   1.077  0.28329
## stateSC:pct.below.pov           4.197e-02  4.890e-02   0.858  0.39197
## stateSD:pct.below.pov                  NA         NA      NA       NA
## stateTN:pct.below.pov          -3.099e-03  4.962e-02  -0.062  0.95027
## stateTX:pct.below.pov           9.496e-03  4.199e-02   0.226  0.82137
## stateUT:pct.below.pov          -1.372e-01  1.214e-01  -1.131  0.25989
## stateVA:pct.below.pov           4.956e-02  5.479e-02   0.905  0.36705
## stateVT:pct.below.pov                  NA         NA      NA       NA
## stateWA:pct.below.pov           2.692e-02  5.593e-02   0.481  0.63104
## stateWI:pct.below.pov           2.582e-02  5.726e-02   0.451  0.65260
## stateWV:pct.below.pov                  NA         NA      NA       NA
## stateAR:pct.unemp                      NA         NA      NA       NA
## stateAZ:pct.unemp               1.026e-01  1.142e-01   0.899  0.37013
## stateCA:pct.unemp               1.331e-01  8.292e-02   1.605  0.11041
## stateCO:pct.unemp              -3.187e-01  4.898e-01  -0.651  0.51623
## stateCT:pct.unemp              -5.850e-02  3.052e-01  -0.192  0.84823
## stateDC:pct.unemp                      NA         NA      NA       NA
## stateDE:pct.unemp                      NA         NA      NA       NA
## stateFL:pct.unemp               1.220e-01  8.251e-02   1.478  0.14135
## stateGA:pct.unemp               2.609e-01  4.779e-01   0.546  0.58586
## stateHI:pct.unemp                      NA         NA      NA       NA
## stateID:pct.unemp                      NA         NA      NA       NA
## stateIL:pct.unemp               9.662e-02  8.931e-02   1.082  0.28092
## stateIN:pct.unemp               1.609e-01  1.028e-01   1.566  0.11944
## stateKS:pct.unemp               2.092e-01  4.871e-01   0.429  0.66823
## stateKY:pct.unemp                      NA         NA      NA       NA
## stateLA:pct.unemp              -4.462e-02  2.212e-01  -0.202  0.84038
## stateMA:pct.unemp               7.376e-02  9.275e-02   0.795  0.42762
## stateMD:pct.unemp               1.191e-01  1.033e-01   1.153  0.25057
## stateME:pct.unemp               2.958e+00  2.357e+00   1.255  0.21126
## stateMI:pct.unemp               1.168e-01  8.442e-02   1.383  0.16851
## stateMN:pct.unemp              -4.892e-01  5.019e-01  -0.975  0.33123
## stateMO:pct.unemp               6.196e-02  1.507e-01   0.411  0.68155
## stateMS:pct.unemp                      NA         NA      NA       NA
## stateMT:pct.unemp                      NA         NA      NA       NA
## stateNC:pct.unemp               1.061e-01  8.807e-02   1.205  0.22987
## stateND:pct.unemp                      NA         NA      NA       NA
## stateNE:pct.unemp                      NA         NA      NA       NA
## stateNH:pct.unemp              -2.620e-02  5.697e-01  -0.046  0.96338
```

```
## stateNJ:pct.unemp                1.818e-01  9.233e-02   1.970  0.05062 .
## stateNM:pct.unemp                       NA         NA      NA       NA
## stateNV:pct.unemp                       NA         NA      NA       NA
## stateNY:pct.unemp                1.686e-01  8.718e-02   1.934  0.05490 .
## stateOH:pct.unemp                9.995e-02  8.436e-02   1.185  0.23785
## stateOK:pct.unemp               -4.213e-01  6.709e-01  -0.628  0.53091
## stateOR:pct.unemp               -2.133e+00  2.251e+00  -0.948  0.34464
## statePA:pct.unemp                1.024e-01  8.399e-02   1.219  0.22479
## stateRI:pct.unemp                       NA         NA      NA       NA
## stateSC:pct.unemp                7.701e-02  9.273e-02   0.830  0.40754
## stateSD:pct.unemp                       NA         NA      NA       NA
## stateTN:pct.unemp                1.837e-01  1.210e-01   1.517  0.13113
## stateTX:pct.unemp                1.025e-01  8.291e-02   1.236  0.21836
## stateUT:pct.unemp               -1.470e-01  3.156e-01  -0.466  0.64203
## stateVA:pct.unemp                2.052e-01  2.934e-01   0.699  0.48530
## stateVT:pct.unemp                       NA         NA      NA       NA
## stateWA:pct.unemp                1.223e-01  1.632e-01   0.749  0.45470
## stateWI:pct.unemp                1.075e-01  7.389e-02   1.454  0.14779
## stateWV:pct.unemp                       NA         NA      NA       NA
## stateAR:pct.pop.65_plus                 NA         NA      NA       NA
## stateAZ:pct.pop.65_plus         -1.658e+04  1.117e+04  -1.484  0.13974
## stateCA:pct.pop.65_plus         -1.686e+04  1.054e+04  -1.600  0.11156
## stateCO:pct.pop.65_plus         -2.075e+04  1.284e+04  -1.616  0.10796
## stateCT:pct.pop.65_plus         -1.620e+04  1.568e+04  -1.033  0.30309
## stateDC:pct.pop.65_plus                 NA         NA      NA       NA
## stateDE:pct.pop.65_plus                 NA         NA      NA       NA
## stateFL:pct.pop.65_plus         -1.625e+04  1.050e+04  -1.548  0.12360
## stateGA:pct.pop.65_plus         -1.612e+04  3.867e+04  -0.417  0.67728
## stateHI:pct.pop.65_plus                 NA         NA      NA       NA
## stateID:pct.pop.65_plus                 NA         NA      NA       NA
## stateIL:pct.pop.65_plus         -1.269e+04  1.067e+04  -1.189  0.23611
## stateIN:pct.pop.65_plus         -1.493e+04  1.063e+04  -1.405  0.16208
## stateKS:pct.pop.65_plus                 NA         NA      NA       NA
## stateKY:pct.pop.65_plus                 NA         NA      NA       NA
## stateLA:pct.pop.65_plus         -2.443e+04  1.340e+04  -1.824  0.07007 .
## stateMA:pct.pop.65_plus         -1.367e+04  1.063e+04  -1.285  0.20055
## stateMD:pct.pop.65_plus         -3.207e+04  1.508e+04  -2.126  0.03506 *
## stateME:pct.pop.65_plus          6.557e+04  5.705e+04   1.149  0.25211
## stateMI:pct.pop.65_plus         -1.551e+04  1.070e+04  -1.449  0.14916
## stateMN:pct.pop.65_plus         -4.495e+04  3.325e+04  -1.352  0.17834
## stateMO:pct.pop.65_plus         -1.976e+04  1.440e+04  -1.372  0.17209
## stateMS:pct.pop.65_plus                 NA         NA      NA       NA
## stateMT:pct.pop.65_plus                 NA         NA      NA       NA
## stateNC:pct.pop.65_plus         -1.547e+04  1.054e+04  -1.468  0.14419
## stateND:pct.pop.65_plus                 NA         NA      NA       NA
## stateNE:pct.pop.65_plus                 NA         NA      NA       NA
## stateNH:pct.pop.65_plus                 NA         NA      NA       NA
## stateNJ:pct.pop.65_plus         -1.212e+04  1.103e+04  -1.099  0.27363
## stateNM:pct.pop.65_plus                 NA         NA      NA       NA
## stateNV:pct.pop.65_plus                 NA         NA      NA       NA
## stateNY:pct.pop.65_plus         -1.506e+04  1.065e+04  -1.414  0.15921
## stateOH:pct.pop.65_plus         -1.658e+04  1.070e+04  -1.549  0.12327
## stateOK:pct.pop.65_plus                 NA         NA      NA       NA
## stateOR:pct.pop.65_plus          4.846e+04  5.096e+04   0.951  0.34307
```

```
## statePA:pct.pop.65_plus          -1.475e+04  1.055e+04  -1.398  0.16401
## stateRI:pct.pop.65_plus                  NA         NA      NA       NA
## stateSC:pct.pop.65_plus          -1.625e+04  1.199e+04  -1.355  0.17723
## stateSD:pct.pop.65_plus                  NA         NA      NA       NA
## stateTN:pct.pop.65_plus          -2.337e+04  1.376e+04  -1.698  0.09146 .
## stateTX:pct.pop.65_plus          -1.325e+04  1.054e+04  -1.257  0.21047
## stateUT:pct.pop.65_plus                  NA         NA      NA       NA
## stateVA:pct.pop.65_plus          -2.115e+04  1.325e+04  -1.596  0.11235
## stateVT:pct.pop.65_plus                  NA         NA      NA       NA
## stateWA:pct.pop.65_plus          -2.040e+04  1.482e+04  -1.376  0.17064
## stateWI:pct.pop.65_plus          -1.490e+04  1.077e+04  -1.383  0.16849
## stateWV:pct.pop.65_plus                  NA         NA      NA       NA
## stateAR:pct.pop.18_34                    NA         NA      NA       NA
## stateAZ:pct.pop.18_34                    NA         NA      NA       NA
## stateCA:pct.pop.18_34             1.052e+04  7.248e+03   1.452  0.14856
## stateCO:pct.pop.18_34             8.259e+03  8.553e+03   0.966  0.33570
## stateCT:pct.pop.18_34             8.459e+03  7.742e+03   1.093  0.27621
## stateDC:pct.pop.18_34                    NA         NA      NA       NA
## stateDE:pct.pop.18_34                    NA         NA      NA       NA
## stateFL:pct.pop.18_34             9.509e+03  7.249e+03   1.312  0.19148
## stateGA:pct.pop.18_34             5.274e+03  1.350e+04   0.391  0.69645
## stateHI:pct.pop.18_34                    NA         NA      NA       NA
## stateID:pct.pop.18_34                    NA         NA      NA       NA
## stateIL:pct.pop.18_34             8.252e+03  7.356e+03   1.122  0.26362
## stateIN:pct.pop.18_34             9.407e+03  7.287e+03   1.291  0.19860
## stateKS:pct.pop.18_34                    NA         NA      NA       NA
## stateKY:pct.pop.18_34                    NA         NA      NA       NA
## stateLA:pct.pop.18_34             8.522e+03  8.319e+03   1.024  0.30719
## stateMA:pct.pop.18_34             8.966e+03  7.291e+03   1.230  0.22063
## stateMD:pct.pop.18_34             1.366e+04  7.600e+03   1.798  0.07411 .
## stateME:pct.pop.18_34                    NA         NA      NA       NA
## stateMI:pct.pop.18_34             9.854e+03  7.317e+03   1.347  0.18000
## stateMN:pct.pop.18_34             1.576e+04  1.123e+04   1.404  0.16236
## stateMO:pct.pop.18_34             1.082e+04  7.525e+03   1.437  0.15257
## stateMS:pct.pop.18_34                    NA         NA      NA       NA
## stateMT:pct.pop.18_34                    NA         NA      NA       NA
## stateNC:pct.pop.18_34             9.745e+03  7.254e+03   1.343  0.18104
## stateND:pct.pop.18_34                    NA         NA      NA       NA
## stateNE:pct.pop.18_34                    NA         NA      NA       NA
## stateNH:pct.pop.18_34                    NA         NA      NA       NA
## stateNJ:pct.pop.18_34             8.868e+03  7.376e+03   1.202  0.23102
## stateNM:pct.pop.18_34                    NA         NA      NA       NA
## stateNV:pct.pop.18_34                    NA         NA      NA       NA
## stateNY:pct.pop.18_34             9.382e+03  7.314e+03   1.283  0.20140
## stateOH:pct.pop.18_34             1.041e+04  7.292e+03   1.428  0.15529
## stateOK:pct.pop.18_34                    NA         NA      NA       NA
## stateOR:pct.pop.18_34            -1.454e+03  6.973e+03  -0.209  0.83504
## statePA:pct.pop.18_34             8.812e+03  7.252e+03   1.215  0.22615
## stateRI:pct.pop.18_34                    NA         NA      NA       NA
## stateSC:pct.pop.18_34             8.544e+03  7.519e+03   1.136  0.25753
## stateSD:pct.pop.18_34                    NA         NA      NA       NA
## stateTN:pct.pop.18_34             1.234e+04  8.626e+03   1.430  0.15465
## stateTX:pct.pop.18_34             9.084e+03  7.250e+03   1.253  0.21207
## stateUT:pct.pop.18_34                    NA         NA      NA       NA
```

```
## stateVA:pct.pop.18_34             1.128e+04  7.377e+03   1.529  0.12835
## stateVT:pct.pop.18_34                    NA         NA      NA       NA
## stateWA:pct.pop.18_34             1.116e+04  8.829e+03   1.264  0.20820
## stateWI:pct.pop.18_34             8.653e+03  7.294e+03   1.186  0.23723
## stateWV:pct.pop.18_34                    NA         NA      NA       NA
## stateAR:log.land.area.per.capita         NA         NA      NA       NA
## stateAZ:log.land.area.per.capita         NA         NA      NA       NA
## stateCA:log.land.area.per.capita -4.084e-01  4.208e-01  -0.971  0.33320
## stateCO:log.land.area.per.capita -5.161e-01  4.626e-01  -1.116  0.26620
## stateCT:log.land.area.per.capita -5.270e-01  5.924e-01  -0.890  0.37499
## stateDC:log.land.area.per.capita         NA         NA      NA       NA
## stateDE:log.land.area.per.capita         NA         NA      NA       NA
## stateFL:log.land.area.per.capita -3.466e-01  4.215e-01  -0.822  0.41210
## stateGA:log.land.area.per.capita -3.030e-01  4.380e-01  -0.692  0.49008
## stateHI:log.land.area.per.capita         NA         NA      NA       NA
## stateID:log.land.area.per.capita         NA         NA      NA       NA
## stateIL:log.land.area.per.capita -4.208e-01  4.253e-01  -0.990  0.32387
## stateIN:log.land.area.per.capita -3.708e-01  4.477e-01  -0.828  0.40881
## stateKS:log.land.area.per.capita         NA         NA      NA       NA
## stateKY:log.land.area.per.capita         NA         NA      NA       NA
## stateLA:log.land.area.per.capita -1.229e-01  4.963e-01  -0.248  0.80476
## stateMA:log.land.area.per.capita -5.015e-01  4.279e-01  -1.172  0.24297
## stateMD:log.land.area.per.capita -1.716e-01  4.707e-01  -0.364  0.71597
## stateME:log.land.area.per.capita         NA         NA      NA       NA
## stateMI:log.land.area.per.capita -4.690e-01  4.254e-01  -1.102  0.27198
## stateMN:log.land.area.per.capita  1.182e-01  5.733e-01   0.206  0.83686
## stateMO:log.land.area.per.capita -5.011e-01  4.191e-01  -1.196  0.23354
## stateMS:log.land.area.per.capita         NA         NA      NA       NA
## stateMT:log.land.area.per.capita         NA         NA      NA       NA
## stateNC:log.land.area.per.capita -4.624e-01  4.263e-01  -1.085  0.27963
## stateND:log.land.area.per.capita         NA         NA      NA       NA
## stateNE:log.land.area.per.capita         NA         NA      NA       NA
## stateNH:log.land.area.per.capita         NA         NA      NA       NA
## stateNJ:log.land.area.per.capita -3.730e-01  4.249e-01  -0.878  0.38134
## stateNM:log.land.area.per.capita         NA         NA      NA       NA
## stateNV:log.land.area.per.capita         NA         NA      NA       NA
## stateNY:log.land.area.per.capita -4.392e-01  4.214e-01  -1.042  0.29896
## stateOH:log.land.area.per.capita -4.177e-01  4.293e-01  -0.973  0.33204
## stateOK:log.land.area.per.capita         NA         NA      NA       NA
## stateOR:log.land.area.per.capita         NA         NA      NA       NA
## statePA:log.land.area.per.capita -4.169e-01  4.216e-01  -0.989  0.32424
## stateRI:log.land.area.per.capita         NA         NA      NA       NA
## stateSC:log.land.area.per.capita -1.813e-01  5.077e-01  -0.357  0.72155
## stateSD:log.land.area.per.capita         NA         NA      NA       NA
## stateTN:log.land.area.per.capita -8.343e-01  6.819e-01  -1.223  0.22295
## stateTX:log.land.area.per.capita -4.271e-01  4.217e-01  -1.013  0.31274
## stateUT:log.land.area.per.capita         NA         NA      NA       NA
## stateVA:log.land.area.per.capita -2.346e-01  4.508e-01  -0.520  0.60348
## stateVT:log.land.area.per.capita         NA         NA      NA       NA
## stateWA:log.land.area.per.capita -4.335e-01  4.255e-01  -1.019  0.30990
## stateWI:log.land.area.per.capita -4.171e-01  4.149e-01  -1.005  0.31625
## stateWV:log.land.area.per.capita         NA         NA      NA       NA
## stateAR:log.doctors.per.capita           NA         NA      NA       NA
## stateAZ:log.doctors.per.capita           NA         NA      NA       NA
```

```
## stateCA:log.doctors.per.capita    3.544e-01  2.416e-01   1.467  0.14434
## stateCO:log.doctors.per.capita   -5.905e-01  8.811e-01  -0.670  0.50368
## stateCT:log.doctors.per.capita   -1.839e-01  5.415e-01  -0.340  0.73454
## stateDC:log.doctors.per.capita           NA         NA      NA       NA
## stateDE:log.doctors.per.capita           NA         NA      NA       NA
## stateFL:log.doctors.per.capita    4.000e-01  2.395e-01   1.670  0.09680 .
## stateGA:log.doctors.per.capita   -1.954e-02  5.461e-01  -0.036  0.97149
## stateHI:log.doctors.per.capita           NA         NA      NA       NA
## stateID:log.doctors.per.capita           NA         NA      NA       NA
## stateIL:log.doctors.per.capita    7.544e-02  2.488e-01   0.303  0.76215
## stateIN:log.doctors.per.capita    3.820e-01  2.756e-01   1.386  0.16766
## stateKS:log.doctors.per.capita           NA         NA      NA       NA
## stateKY:log.doctors.per.capita           NA         NA      NA       NA
## stateLA:log.doctors.per.capita    1.957e-01  3.282e-01   0.596  0.55170
## stateMA:log.doctors.per.capita    1.983e-01  2.428e-01   0.817  0.41519
## stateMD:log.doctors.per.capita    1.126e+00  6.090e-01   1.849  0.06624 .
## stateME:log.doctors.per.capita           NA         NA      NA       NA
## stateMI:log.doctors.per.capita    2.524e-01  2.401e-01   1.051  0.29479
## stateMN:log.doctors.per.capita           NA         NA      NA       NA
## stateMO:log.doctors.per.capita   -6.039e-02  5.572e-01  -0.108  0.91384
## stateMS:log.doctors.per.capita           NA         NA      NA       NA
## stateMT:log.doctors.per.capita           NA         NA      NA       NA
## stateNC:log.doctors.per.capita    1.793e-01  2.353e-01   0.762  0.44706
## stateND:log.doctors.per.capita           NA         NA      NA       NA
## stateNE:log.doctors.per.capita           NA         NA      NA       NA
## stateNH:log.doctors.per.capita           NA         NA      NA       NA
## stateNJ:log.doctors.per.capita    3.355e-01  2.505e-01   1.340  0.18225
## stateNM:log.doctors.per.capita           NA         NA      NA       NA
## stateNV:log.doctors.per.capita           NA         NA      NA       NA
## stateNY:log.doctors.per.capita    2.821e-01  2.423e-01   1.164  0.24604
## stateOH:log.doctors.per.capita    3.298e-01  2.443e-01   1.350  0.17893
## stateOK:log.doctors.per.capita           NA         NA      NA       NA
## stateOR:log.doctors.per.capita           NA         NA      NA       NA
## statePA:log.doctors.per.capita    1.579e-01  2.325e-01   0.679  0.49792
## stateRI:log.doctors.per.capita           NA         NA      NA       NA
## stateSC:log.doctors.per.capita    4.976e-02  2.931e-01   0.170  0.86542
## stateSD:log.doctors.per.capita           NA         NA      NA       NA
## stateTN:log.doctors.per.capita    6.605e-02  3.039e-01   0.217  0.82825
## stateTX:log.doctors.per.capita    2.052e-01  2.312e-01   0.887  0.37614
## stateUT:log.doctors.per.capita           NA         NA      NA       NA
## stateVA:log.doctors.per.capita    2.564e-01  2.661e-01   0.963  0.33688
## stateVT:log.doctors.per.capita           NA         NA      NA       NA
## stateWA:log.doctors.per.capita    5.342e-02  3.887e-01   0.137  0.89087
## stateWI:log.doctors.per.capita           NA         NA      NA       NA
## stateWV:log.doctors.per.capita           NA         NA      NA       NA
## stateAR:log.pct.bach.deg                 NA         NA      NA       NA
## stateAZ:log.pct.bach.deg                 NA         NA      NA       NA
## stateCA:log.pct.bach.deg          6.448e-02  4.811e-01   0.134  0.89355
## stateCO:log.pct.bach.deg         -2.698e-01  5.723e-01  -0.471  0.63794
## stateCT:log.pct.bach.deg                 NA         NA      NA       NA
## stateDC:log.pct.bach.deg                 NA         NA      NA       NA
## stateDE:log.pct.bach.deg                 NA         NA      NA       NA
## stateFL:log.pct.bach.deg          9.254e-02  4.930e-01   0.188  0.85135
## stateGA:log.pct.bach.deg          6.329e-01  8.273e-01   0.765  0.44537
```

```
## stateHI:log.pct.bach.deg                 NA         NA      NA       NA
## stateID:log.pct.bach.deg                 NA         NA      NA       NA
## stateIL:log.pct.bach.deg          2.563e-01  5.892e-01   0.435  0.66419
## stateIN:log.pct.bach.deg         -2.087e-01  5.283e-01  -0.395  0.69341
## stateKS:log.pct.bach.deg                 NA         NA      NA       NA
## stateKY:log.pct.bach.deg                 NA         NA      NA       NA
## stateLA:log.pct.bach.deg         -1.182e-01  6.540e-01  -0.181  0.85681
## stateMA:log.pct.bach.deg         -4.989e-01  6.335e-01  -0.788  0.43215
## stateMD:log.pct.bach.deg         -1.039e+00  9.609e-01  -1.081  0.28121
## stateME:log.pct.bach.deg                 NA         NA      NA       NA
## stateMI:log.pct.bach.deg         -3.251e-01  5.349e-01  -0.608  0.54428
## stateMN:log.pct.bach.deg                 NA         NA      NA       NA
## stateMO:log.pct.bach.deg                 NA         NA      NA       NA
## stateMS:log.pct.bach.deg                 NA         NA      NA       NA
## stateMT:log.pct.bach.deg                 NA         NA      NA       NA
## stateNC:log.pct.bach.deg          1.424e-01  4.987e-01   0.286  0.77551
## stateND:log.pct.bach.deg                 NA         NA      NA       NA
## stateNE:log.pct.bach.deg                 NA         NA      NA       NA
## stateNH:log.pct.bach.deg                 NA         NA      NA       NA
## stateNJ:log.pct.bach.deg          3.820e-01  5.786e-01   0.660  0.51000
## stateNM:log.pct.bach.deg                 NA         NA      NA       NA
## stateNV:log.pct.bach.deg                 NA         NA      NA       NA
## stateNY:log.pct.bach.deg          1.787e-01  5.288e-01   0.338  0.73580
## stateOH:log.pct.bach.deg         -2.399e-01  5.173e-01  -0.464  0.64343
## stateOK:log.pct.bach.deg                 NA         NA      NA       NA
## stateOR:log.pct.bach.deg                 NA         NA      NA       NA
## statePA:log.pct.bach.deg         -1.145e-02  4.996e-01  -0.023  0.98173
## stateRI:log.pct.bach.deg                 NA         NA      NA       NA
## stateSC:log.pct.bach.deg          3.913e-01  7.636e-01   0.512  0.60907
## stateSD:log.pct.bach.deg                 NA         NA      NA       NA
## stateTN:log.pct.bach.deg                 NA         NA      NA       NA
## stateTX:log.pct.bach.deg         -7.342e-02  4.815e-01  -0.152  0.87900
## stateUT:log.pct.bach.deg                 NA         NA      NA       NA
## stateVA:log.pct.bach.deg          1.039e+00  1.426e+00   0.729  0.46715
## stateVT:log.pct.bach.deg                 NA         NA      NA       NA
## stateWA:log.pct.bach.deg          4.202e-01  6.520e-01   0.645  0.52015
## stateWI:log.pct.bach.deg                 NA         NA      NA       NA
## stateWV:log.pct.bach.deg                 NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07731 on 160 degrees of freedom
## Multiple R-squared:  0.949,  Adjusted R-squared:  0.8601
## F-statistic: 10.68 on 279 and 160 DF,  p-value: < 2.2e-16
```

> None of the interaction terms between state and other quantitative variables is significant. We can only use variable region to represent the geographic factor. If we want to use variables state or county, we need to include more data to expand the existing dataset for future improvements.

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Cmd+Option+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Cmd+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor

is displayed.