Factors Influencing Personal Income in the U.S.

Xiangman Zhao

xiangmaz@andrew.cmu.edu

16 October 2021

Abstract

I address the question of finding appropriate variables to predict personal income per capita in the U.S. I examine data on selected county demographic information for 440 most populous counties in the U.S. I found that personal income is largely affected by the percent of high school graduates, bachelor's degree, below poverty and unemployment, and crime is surprisingly not a significant factor. The level of personal income varies with region due to different economic policies. I finally decided to use land area, population between 18 and 34, doctors, the percent of high school graduates, the percent of bachelor's degree, the percent of people below poverty, the percent of unemployment and region to predict personal income. There is still room for model improvement if we can find additional data on missing counties for model comparison to reduce overfitting noise.

1. Introduction

Personal income per capita 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 personal income per capita?

This question is critical for social scientists to think about 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 we have been asked to build an optimal model to predict the personal income and investigate if there are any additional information needed to better answer the question.

In addition to answer the main question above, we will also answer the following related questions:

- Which variables are related to each other?
- Is there a relationship between crime rate and average income per capita?
- Can we improve the model by including more data on missing counties?

2. Data

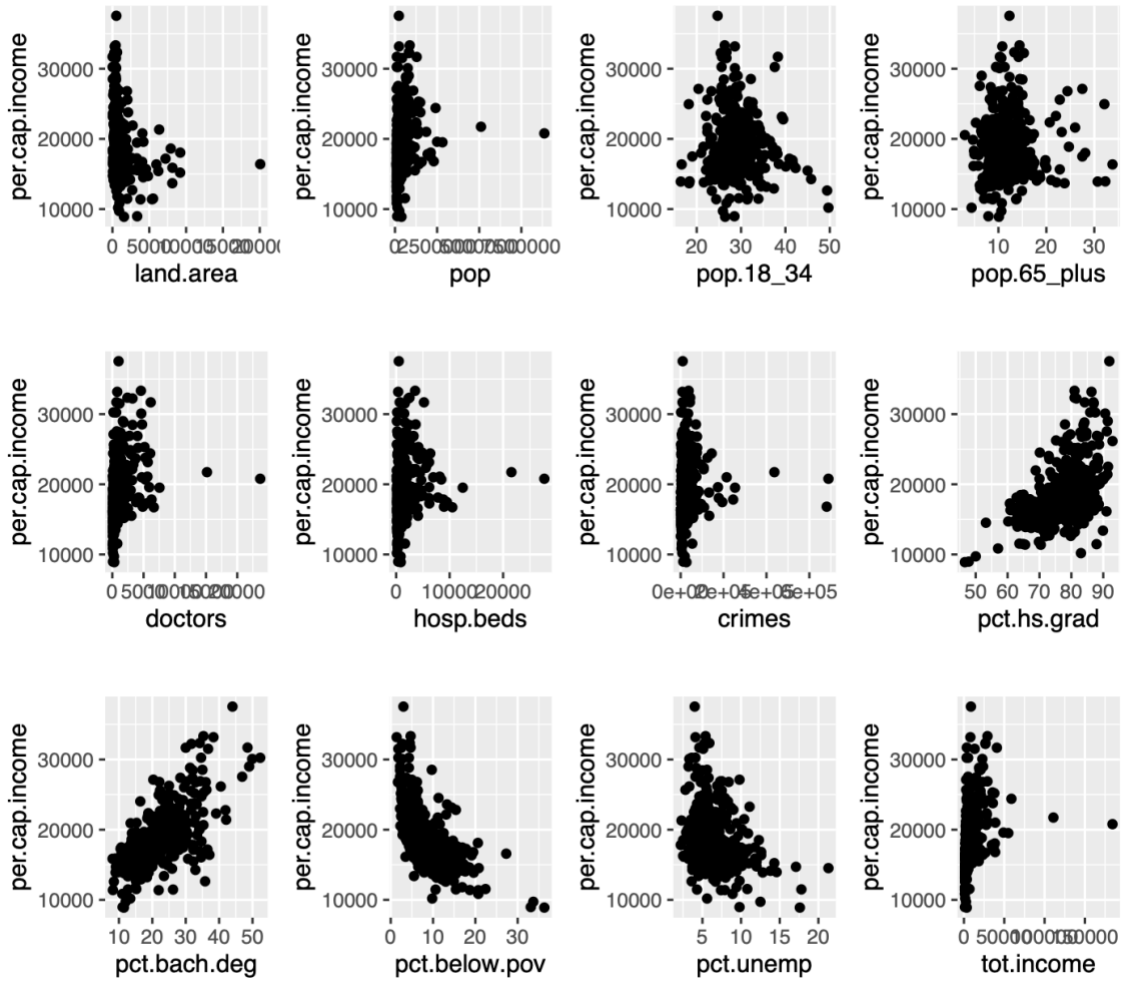
The data for this study provides selected county demographic information for 440 of the most populous counties in the United States. Readers can check Kutner et al. (2005)[1] for more details and information. There are 17 variables in the dataset, and three of them are categorical variables: region,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

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

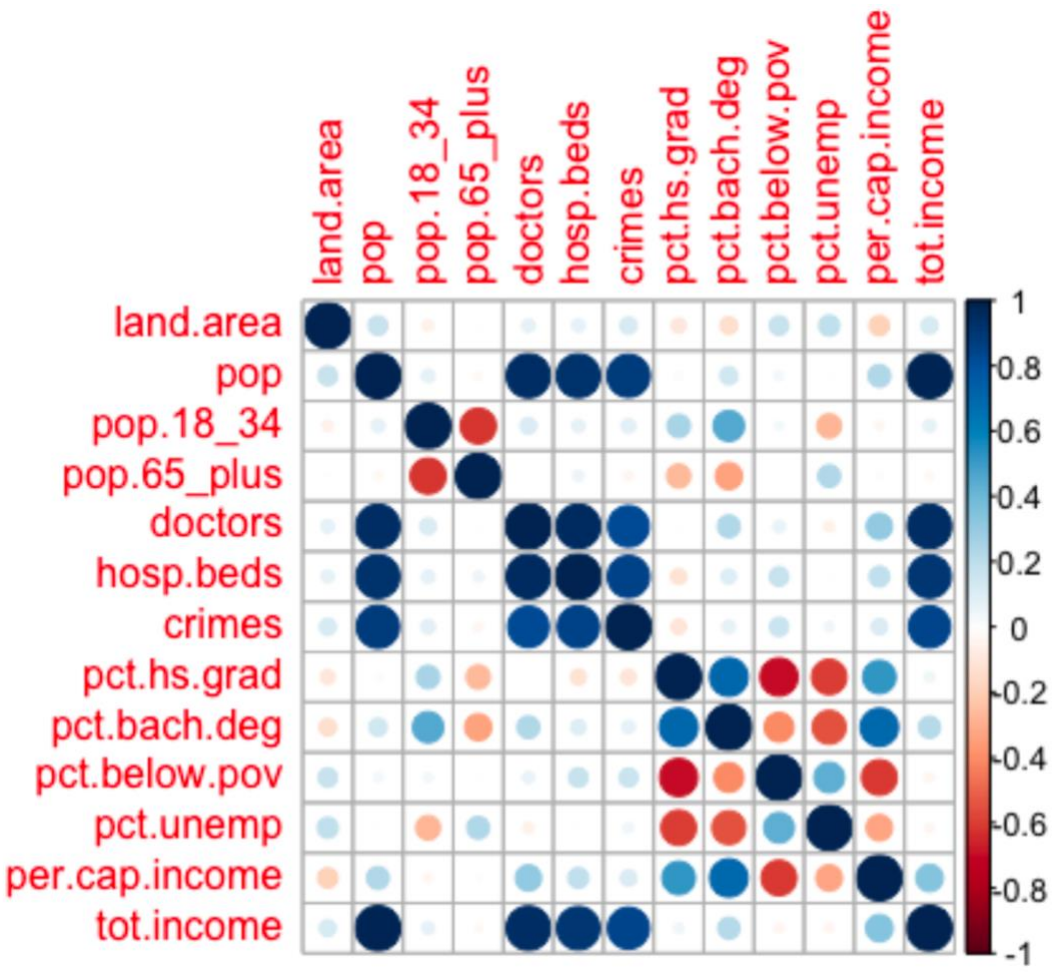
	region of the US), NC (north-central region of the US), S (southern region of the US), and W (Western region of the US)

Except for the response variable average income per capita and identification number for each row, below are 12 scatterplots between average income per capita and quantitative independent variables:



From these scatterplots, I find out that there are five variables that have relatively strong relationship with average income per capita: pop.65_plus, pct.hs.grad, pct.bach.deg, pct.below.pov and pct.unemp.

I also plotted a correlation plot to show the relationship between each variable. The darker the color, the higher correlation between two variables:



Below is a summary table of all the quantitative variables:

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

I also summarize some statistics of three categorical variables:

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

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

Variables county and state have so many unique values that their frequency tables do not provide much useful information. From the summary table for region, I find that NE (Northeastern) has higher baseline salary.

3. Methods

Correlation plot is used to investigate the relationship between each quantitative variable in the dataset. Twelve scatterplots between each independent variable and per.cap.income are plotted to investigate the relationship between these independent variables and the per.cap.income.

I build three linear models using three combinations of crimes, region and the interaction term between crimes and region to predict per.cap.income. I also preform ANOVA test to evaluate the most significant model from these three models.

In order to keep independent variables consistent with the response variable per capita income, we transform variables land.area, doctors, hosp.beds, crimes to land area per capita, doctors per capita, hosp.beds per capita and crimes per capita. We also divide pop.18_34 and pop.65_plus 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, we remove variables pop and tot.income to avoid too much collinearity. We use histograms to check the distribution of each variable and use log transformation to make them normal.

After the transformation, I tried two variables selection methods: stepwise and all subsets to find out most appropriate and significant quantitative variables. A model of interaction terms between region and all other quantitative variables are built to see if any interaction terms will help explain the model. Four diagnostic plots are used to evaluate the

linearity of these models. A five-fold cross validation is used to evaluate the predictability of the model.

4. Results

The main aim of the paper is to build an appropriate model to predict per.cap.income. After variables transformations and variables selection, my final model and variables' coefficients are shown below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.429525e+00	3.077012e-01	30.645073549	9.801801e-109
pct.below.pov	-2.747203e-02	4.238282e-03	-6.481878691	2.568108e-10
pct.unemp	2.463618e-02	5.677918e-03	4.338945826	1.799206e-05
regionNE	1.332848e-01	3.895449e-01	0.342155175	7.324069e-01
regionS	1.980384e-01	3.417511e-01	0.579481283	5.625781e-01
regionW	1.595579e+00	4.387290e-01	3.636820853	3.106677e-04
pct.pop.65_plus	1.034913e+03	1.738661e+02	5.952355263	5.620733e-09
pct.pop.18_34	-6.444283e+02	8.251064e+01	-7.810245023	4.693662e-14
log.land.area.per.capita	-3.341141e-02	6.235304e-03	-5.358424545	1.393375e-07
log.pct.bach.deg	3.161447e-01	2.738703e-02	11.543594833	6.237038e-27
pct.hs.grad	-5.455958e-03	2.924378e-03	-1.865681264	6.278960e-02
log.doctors.per.capita	3.447610e-02	1.746121e-02	1.974439783	4.899274e-02
regionNE:pct.hs.grad	1.887734e-03	3.465985e-03	0.544645682	5.862891e-01
regionS:pct.hs.grad	1.194818e-05	3.037565e-03	0.003933473	9.968634e-01
regionW:pct.hs.grad	-1.180117e-02	4.069307e-03	-2.900044013	3.928905e-03
pct.below.pov:regionNE	-1.619308e-03	5.784708e-03	-0.279929094	7.796710e-01
pct.below.pov:regionS	6.060271e-03	4.700938e-03	1.289161911	1.980582e-01
pct.below.pov:regionW	-1.132848e-02	6.183805e-03	-1.831959958	6.767218e-02
regionNE:log.doctors.per.capita	2.569232e-02	2.663105e-02	0.964750426	3.352303e-01
regionS:log.doctors.per.capita	2.119632e-02	2.196130e-02	0.965166759	3.350220e-01
regionW:log.doctors.per.capita	6.787510e-02	3.060417e-02	2.217838267	2.710539e-02
pct.unemp:regionNE	-1.893370e-02	8.594333e-03	-2.203045262	2.813911e-02
pct.unemp:regionS	-2.557064e-02	7.542125e-03	-3.390375884	7.646562e-04
pct.unemp:regionW	-2.005248e-02	7.864561e-03	-2.549726443	1.113847e-02

The R squared of the model is 0.8213, and the cross-validated R squared is 0.8010. There is not a significant difference between these two values, which shows that there is not a huge problem of overfitting.

To answer the first sub research question, I draw a correlation plot to investigate the relationship between each variable, and I find that tot.income is correlated to doctors, hosp.beds, crimes and pop. Pop.18_34 is negatively related to pop.65_plus. The plot also shows that variables pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp and per.cap.income are correlated to each other. The scatterplots between per.cap.income and independent variables pct.hs.grad, pct.bach.deg, pct.below.pov and pct.unemp further prove that they are correlated. After plotting histograms for all variables, I find that variables per.cap.income, land.area, pop,

doctors, hosp.beds, crimes and tot.income are skewed to the right. Therefore, I use log transformation to make these variables' distributions normal.

For the second sub question on whether variable crimes and region are significant, I used ANOVA test to evaluate these three models and find that either the interaction between crimes and region or the interaction between crimes per capita and region is not significant. After comparing the R squared, I find that the model with only crimes and region has higher R squared and can explain most of the data.

5. Discussions

Average personal income is related to economic factors like percent of unemployment and people below poverty, social factors like education levels and health system and geographical factors like region. In the exploratory analysis, I find 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 collinearity is not an issue in the final model. 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 is an important factor in the 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.

The model does a good job predicting per.cap.income and the prediction error is pretty small from the cross validation process. Diagnostic plots show that the model fulfills all the linearity condition and collinearity is not a problem. 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 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 a little bit complicated, which makes it hard for nontechnical people to understand. Last, 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.

Although overfitting does not seem like a big problem in the model, 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.

6. References

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

R Notebook

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*.

```
library(car)
```

```
## Loading required package: carData
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
```

```
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(ggplot2)
library(dplyr)
library(glmnet)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
## Loaded glmnet 4.1-2
```

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(leaps)
library(MASS)
```

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

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

```
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
```

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

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(grid)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

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

```
library(ggplotify)
library(reshape2)
cdi <- read.table("/Users/zhaoxiangman/Desktop/36-617/cdi.dat", header = TRUE)
head(cdi)
```

```
##   id      county state land.area     pop pop.18_34 pop.65_plus doctors
## 1  1 Los_Angeles    CA      4060 8863164      32.1         9.7   23677
## 2  2        Cook    IL       946 5105067      29.2        12.4   15153
## 3  3      Harris    TX      1729 2818199      31.3         7.1    7553
## 4  4   San_Diego    CA      4205 2498016      33.5        10.9    5905
## 5  5      Orange    CA       790 2410556      32.6         9.2    6062
## 6  6       Kings    NY        71 2300664      28.3        12.4    4861
##   hosp.beds crimes pct.hs.grad pct.bach.deg pct.below.pov pct.unemp
## 1     27700 688936        70.0         22.3          11.6       8.0
```

```
## 2     21550 436936        73.4         22.8          11.1       7.2
## 3     12449 253526        74.9         25.4          12.5       5.7
## 4      6179 173821        81.9         25.3           8.1       6.1
## 5      6369 144524        81.2         27.8           5.2       4.8
## 6      8942 680966        63.7         16.6          19.5       9.5
##   per.cap.income tot.income region
## 1          20786     184230      W
## 2          21729     110928     NC
## 3          19517      55003      S
## 4          19588      48931      W
## 5          24400      58818      W
## 6          16803      38658     NE
```

```
summary(cdi$land.area)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    15.0   451.2   656.5  1041.4   946.8 20062.0
```

```
sd(cdi$land.area)
```

```
## [1] 1549.922
```

```
length(which(!is.na(cdi$land.area)))
```

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

```
summary(cdi$pop)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  100043  139027  217280  393011  436064 8863164
```

```
sd(cdi$pop)
```

```
## [1] 601987
```

```
length(which(!is.na(cdi$pop)))
```

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

```
summary(cdi$pop.18_34)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.40   26.20   28.10   28.57   30.02   49.70
```

```
sd(cdi$pop.18_34)
```

```
## [1] 4.191083
```

```
length(which(!is.na(cdi$pop.18_34)))
```

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

```
summary(cdi$pop.65_plus)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   9.875  11.750  12.170  13.625  33.800
```

```
sd(cdi$pop.65_plus)
```

```
## [1] 3.992666
```

```
length(which(!is.na(cdi$pop.65_plus)))
```

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

```
summary(cdi$doctors)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    39.0   182.8   401.0   988.0  1036.0 23677.0
```

```
sd(cdi$doctors)
```

```
## [1] 1789.75
```

```
length(which(!is.na(cdi$doctors)))
```

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

```
summary(cdi$hosp.beds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    92.0   390.8   755.0  1458.6  1575.8 27700.0
```

```
sd(cdi$hosp.beds)
```

```
## [1] 2289.134
```

```
length(which(!is.na(cdi$hosp.beds)))
```

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

```
summary(cdi$crimes)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     563    6220   11820   27112   26280  688936
```

```
sd(cdi$crimes)
```

```
## [1] 58237.51
```

```
length(which(!is.na(cdi$crimes)))
```

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

```
summary(cdi$pct.hs.grad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   46.60   73.88   77.70   77.56   82.40   92.90
```

```
sd(cdi$pct.hs.grad)
```

```
## [1] 7.015159
```

```
length(which(!is.na(cdi$pct.hs.grad)))
```

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

```
summary(cdi$pct.bach.deg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.10   15.28   19.70   21.08   25.32   52.30
```

```
sd(cdi$pct.bach.deg)
```

```
## [1] 7.654524
```

```
length(which(!is.na(cdi$pct.bach.deg)))
```

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

```
summary(cdi$pct.below.pov)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.400   5.300   7.900   8.721  10.900  36.300
```

```
sd(cdi$pct.below.pov)
```

```
## [1] 4.656737
```

```
length(which(!is.na(cdi$pct.below.pov)))
```

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

```
summary(cdi$pct.unemp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.200   5.100   6.200   6.597   7.500  21.300
```

```
sd(cdi$pct.unemp)
```

```
## [1] 2.337924
```

```
length(which(!is.na(cdi$pct.unemp)))
```

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

```
summary(cdi$per.cap.income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8899   16118   17759   18561   20270   37541
```

```
sd(cdi$per.cap.income)
```

```
## [1] 4059.192
```

```
length(which(!is.na(cdi$per.cap.income)))
```

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

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1141    2311    3857    7869    8654  184230
```

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

```
## [1] 12884.32
```

```
length(which(!is.na(cdi$tot.income)))
```

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

```
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)
```

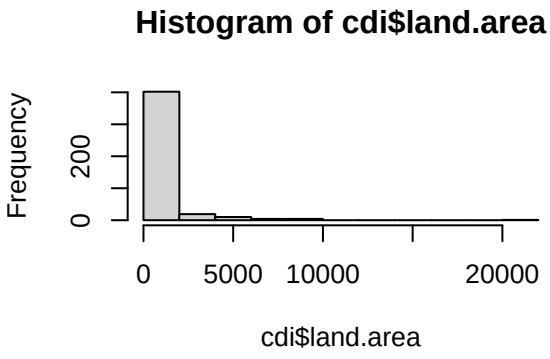
```
## [1] 7
```

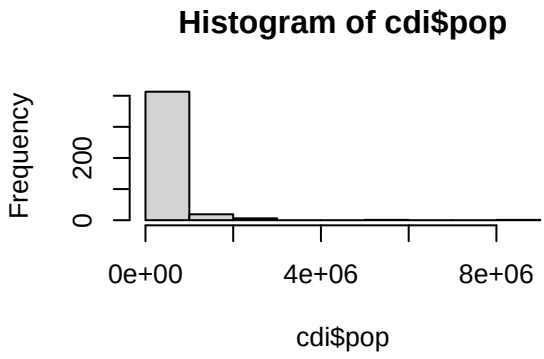
```
which(is.na(cdi))
```

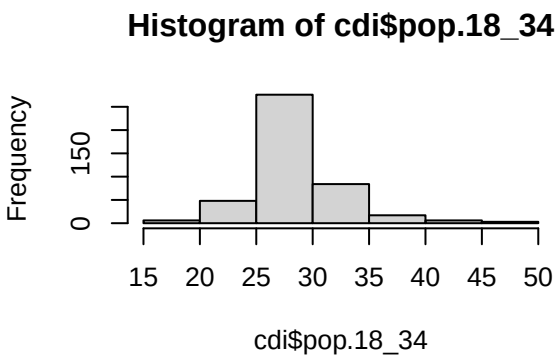
```
## integer(0)
```

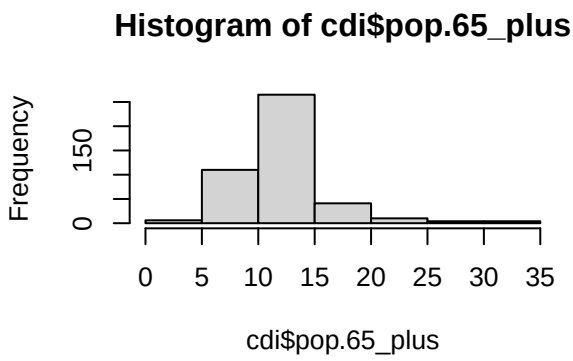
EDA

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

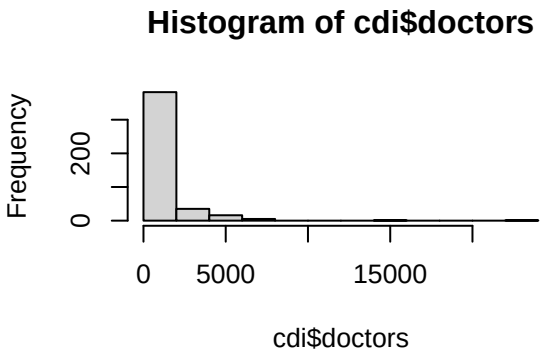


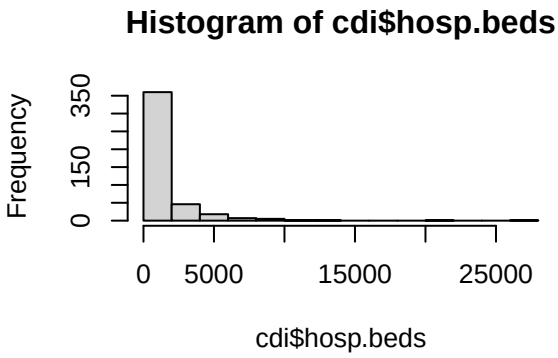


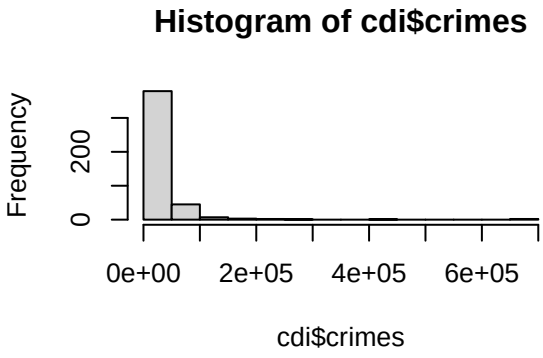


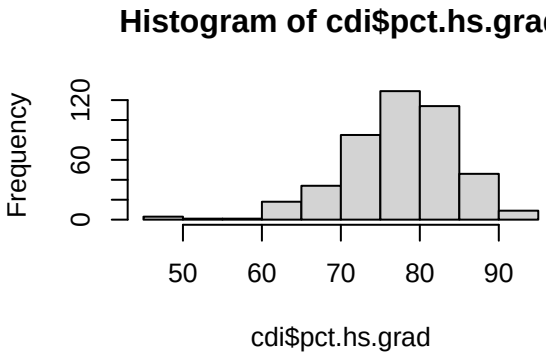



```
hist(cdi$doctors)
hist(cdi$hosp.beds)
hist(cdi$crimes)
hist(cdi$pct.hs.grad)
```

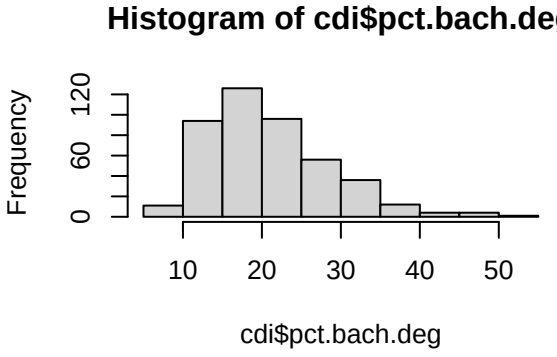


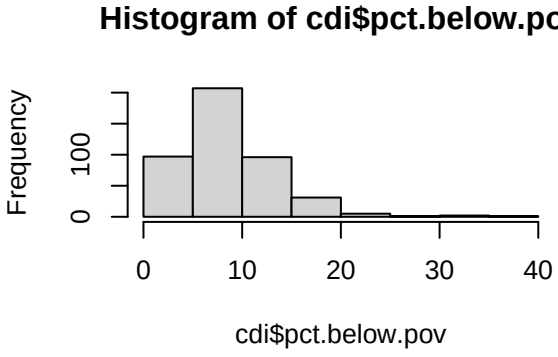


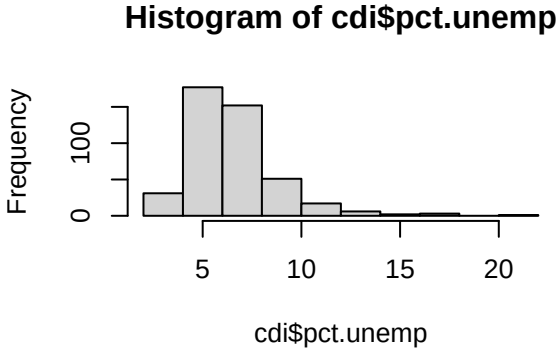


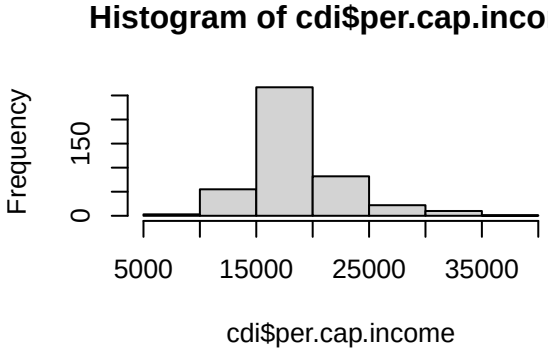



```
hist(cdi$pct.bach.deg)
hist(cdi$pct.below.pov)
hist(cdi$pct.unemp)
hist(cdi$per.cap.income)
```

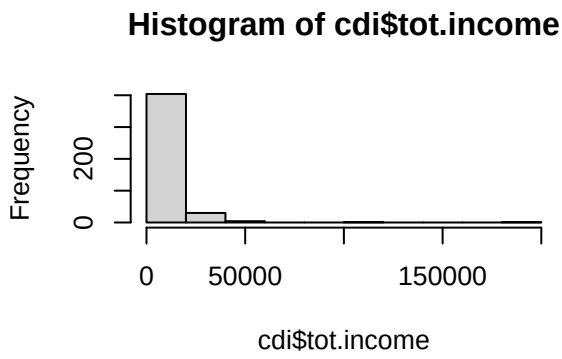




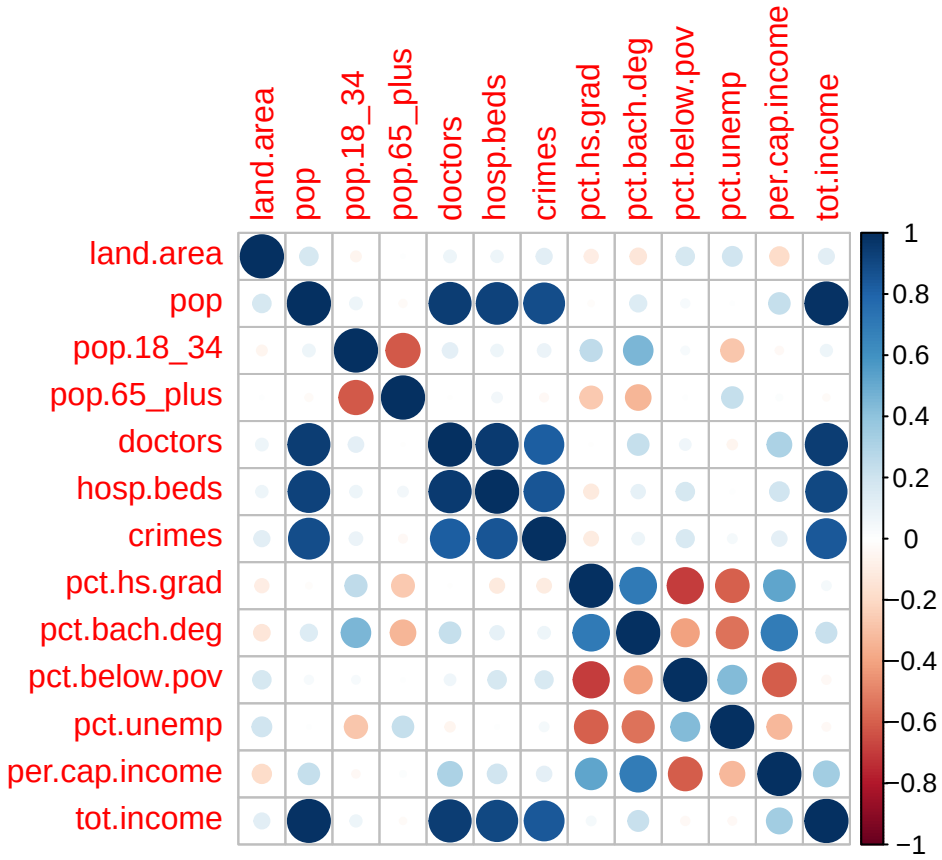




```
hist(cdi$tot.income)
```

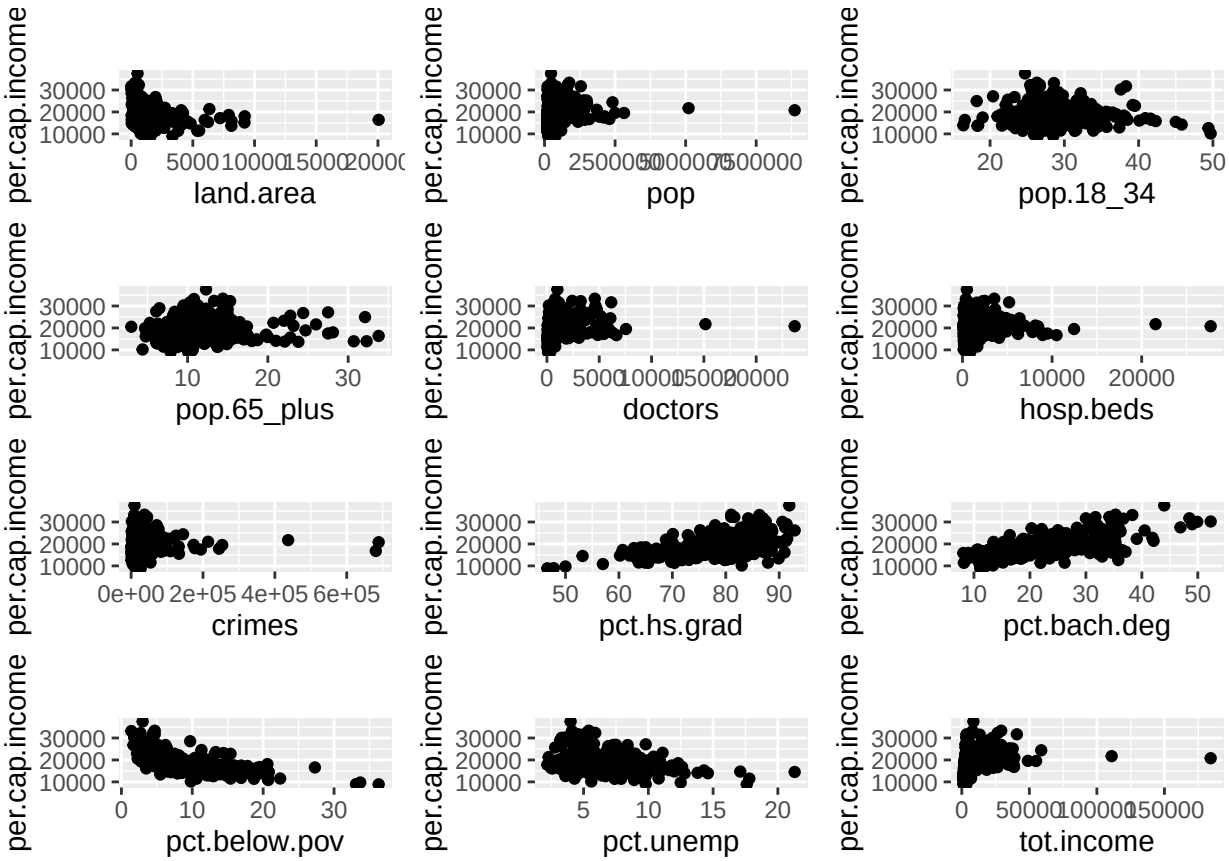



```
cdi_quan <- cdi[4:16]
C <-cor(cdi_quan)
corrplot(C, method="circle")
```

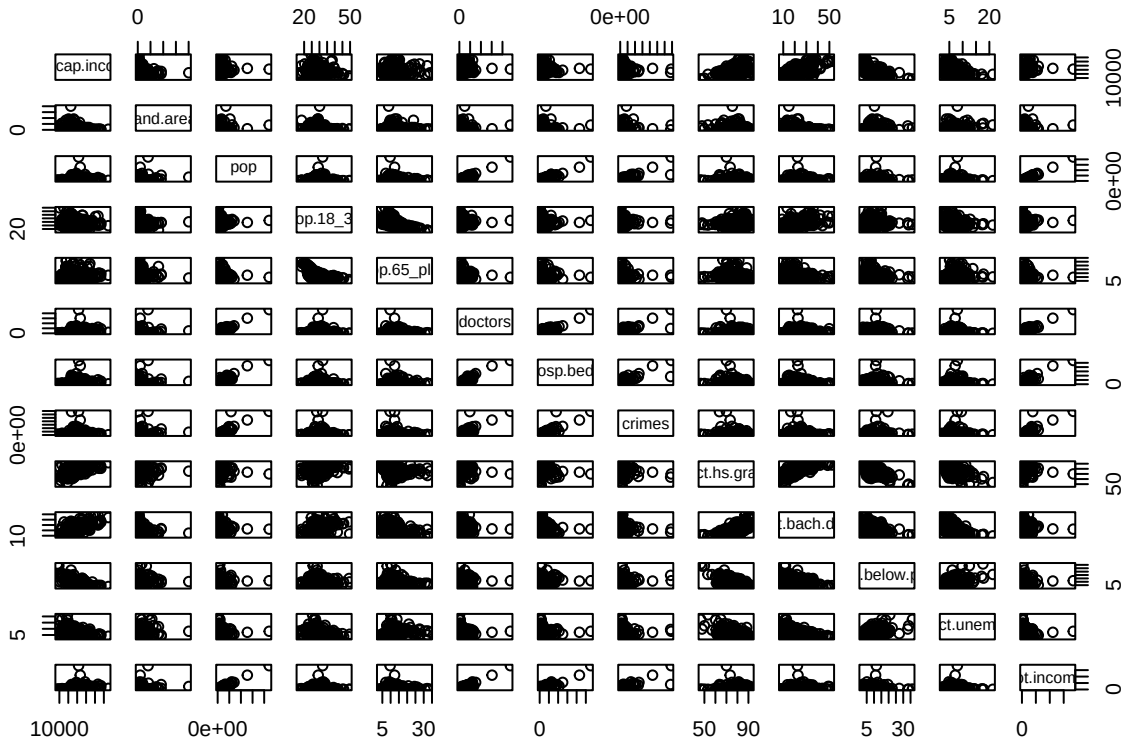



```
scatter.builder <- function(df,yvar="per.cap.income") {
    result <- NULL
    y.index <- grep(yvar,names(df))
    for (xvar in names(df)[-y.index]) {
      d <- data.frame(xx=df[,xvar],yy=df[,y.index])
      if(mode(df[,xvar])=="numeric") {
        p <- ggplot(d,aes(x=xx,y=yy)) + geom_point() +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

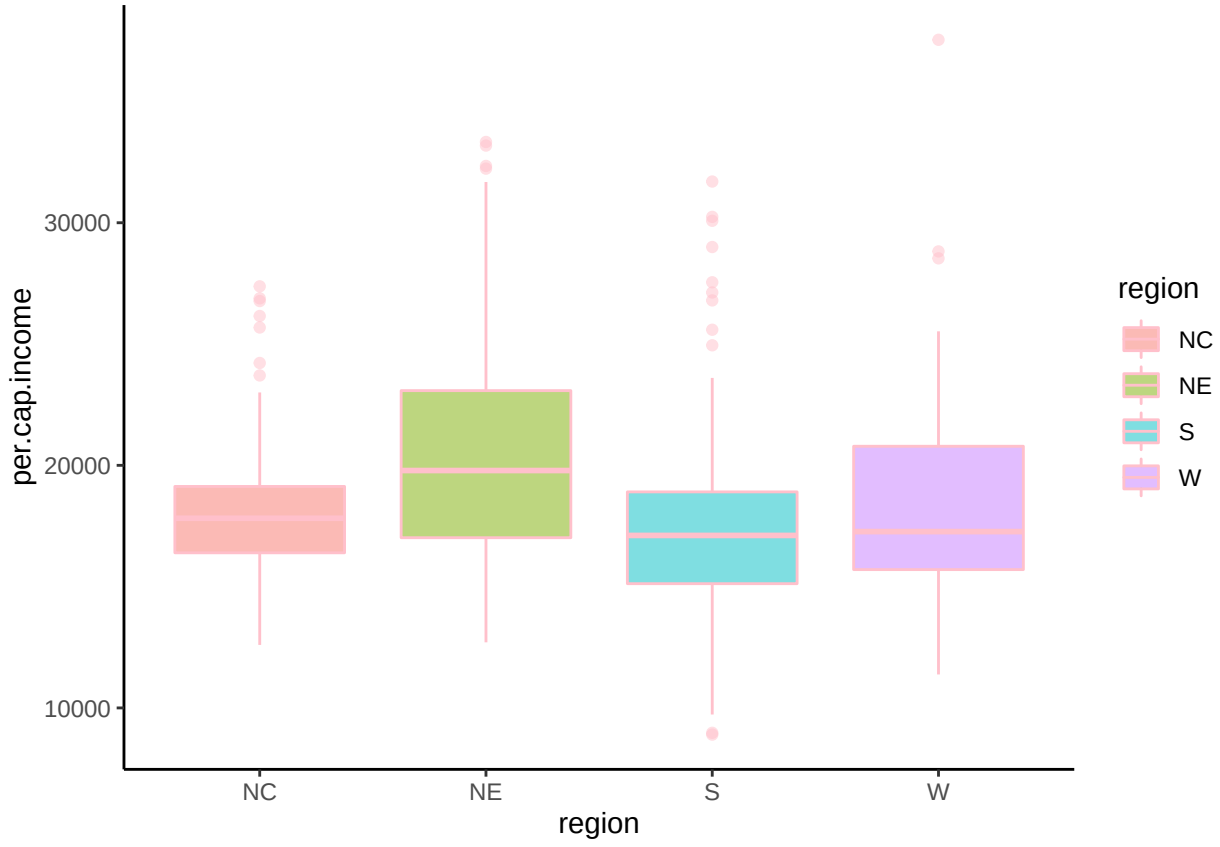
```
          p <- ggplot(d,aes(x=xx,y=yy)) + geom_boxplot(notch=F) +
                 ggtitle("") + xlab(xvar) + ylab(yvar)
      }
    result <- c(result,list(p))
    }
    return(result)
}

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




```
pairs( per.cap.income ~ ., data = cdi_quan)
```

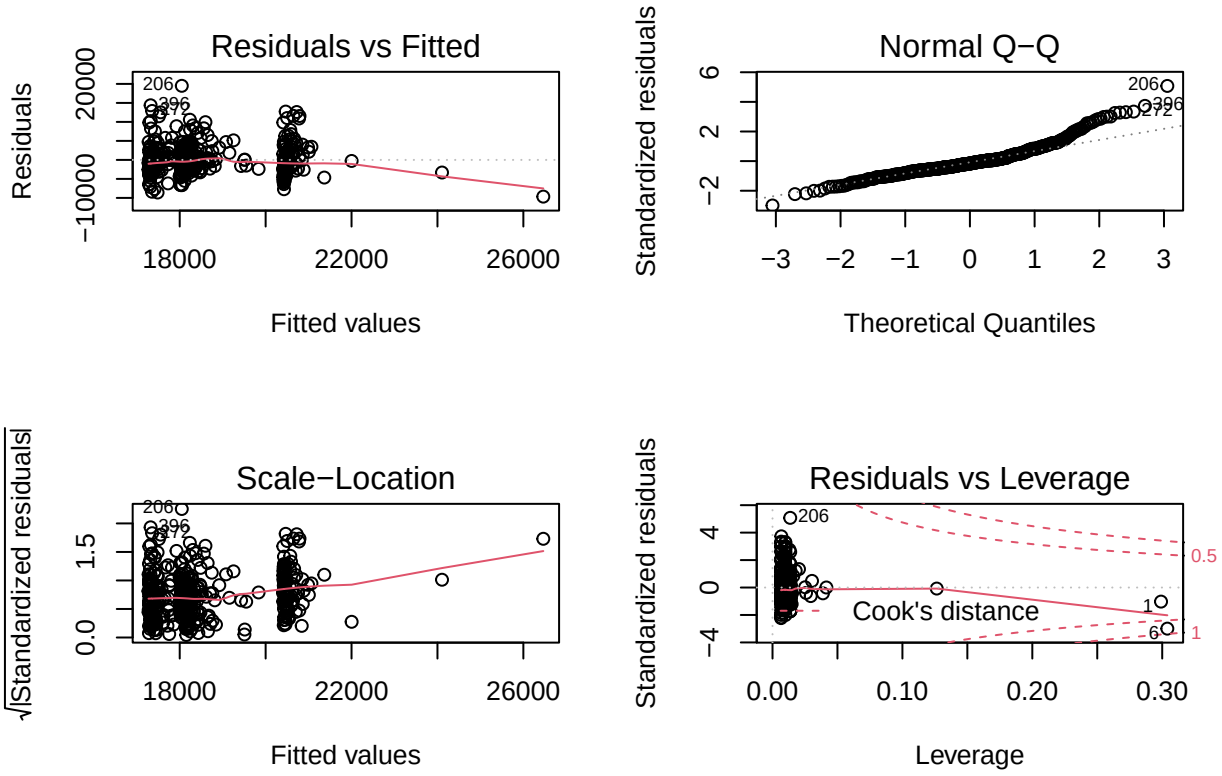



```
ggplot(cdi, aes(x=region, y=per.cap.income, fill=region)) +
    geom_boxplot(color = "pink", alpha=0.5) +
  theme_classic()
```



##(b)

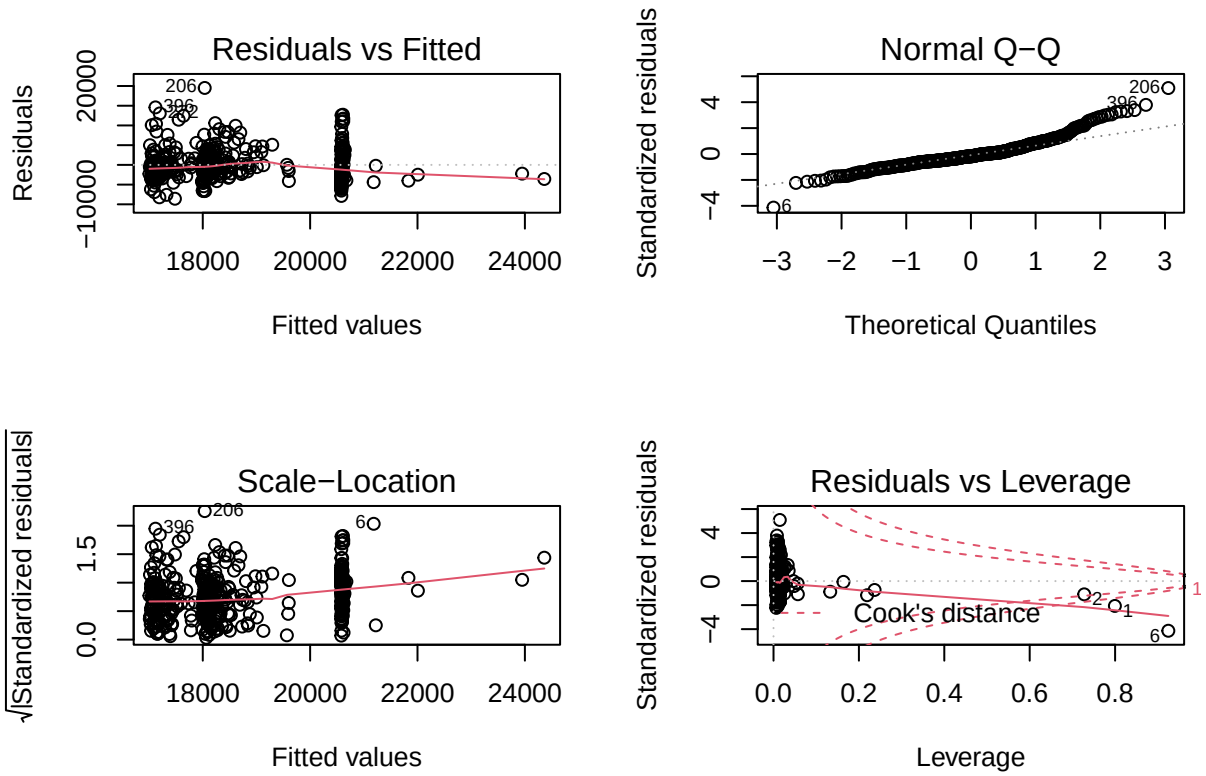
```
mod1 <- lm(per.cap.income~crimes + region, data = cdi)
par(mfrow=c(2,2))
plot(mod1)
```




```
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)
```

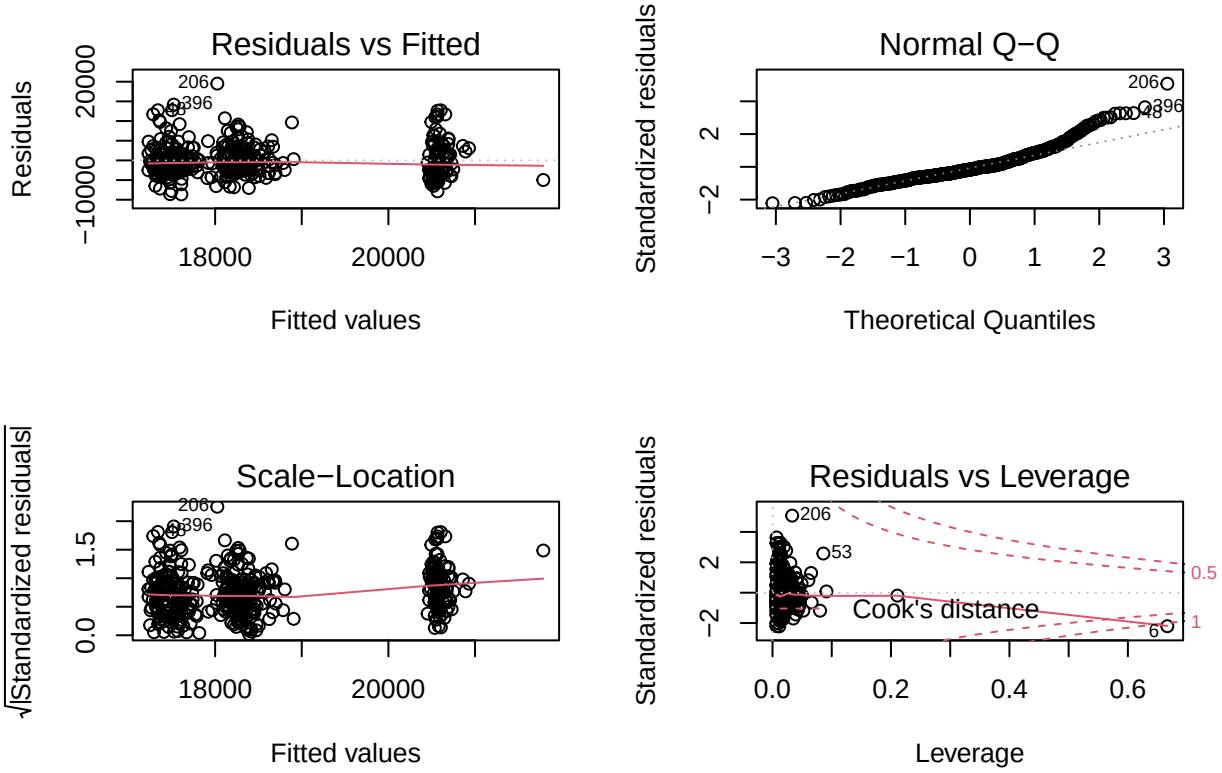
```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region + crimes * region,
##     data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8582.4 -2225.2  -676.2  1563.4 19504.7
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept) 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

```
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)
```

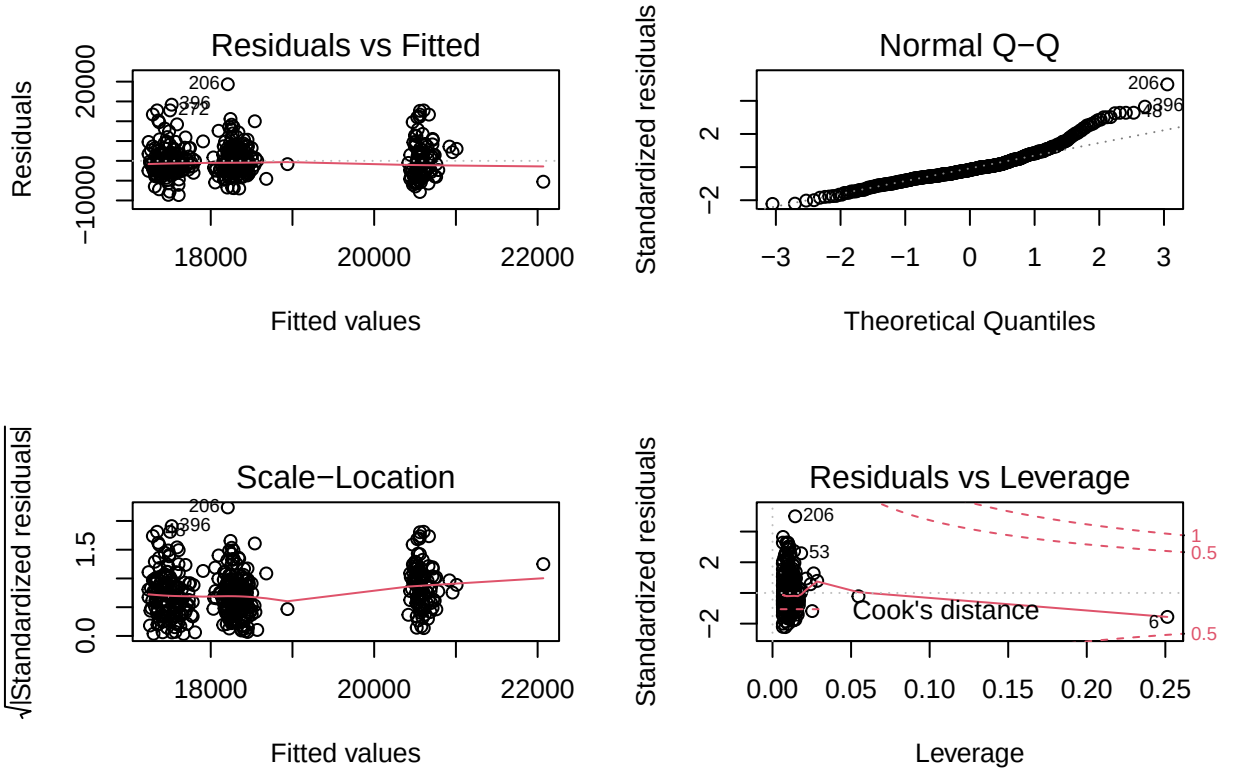



```
summary(mod3)
```

```
##
## 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
```

##(c)

```
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
```

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

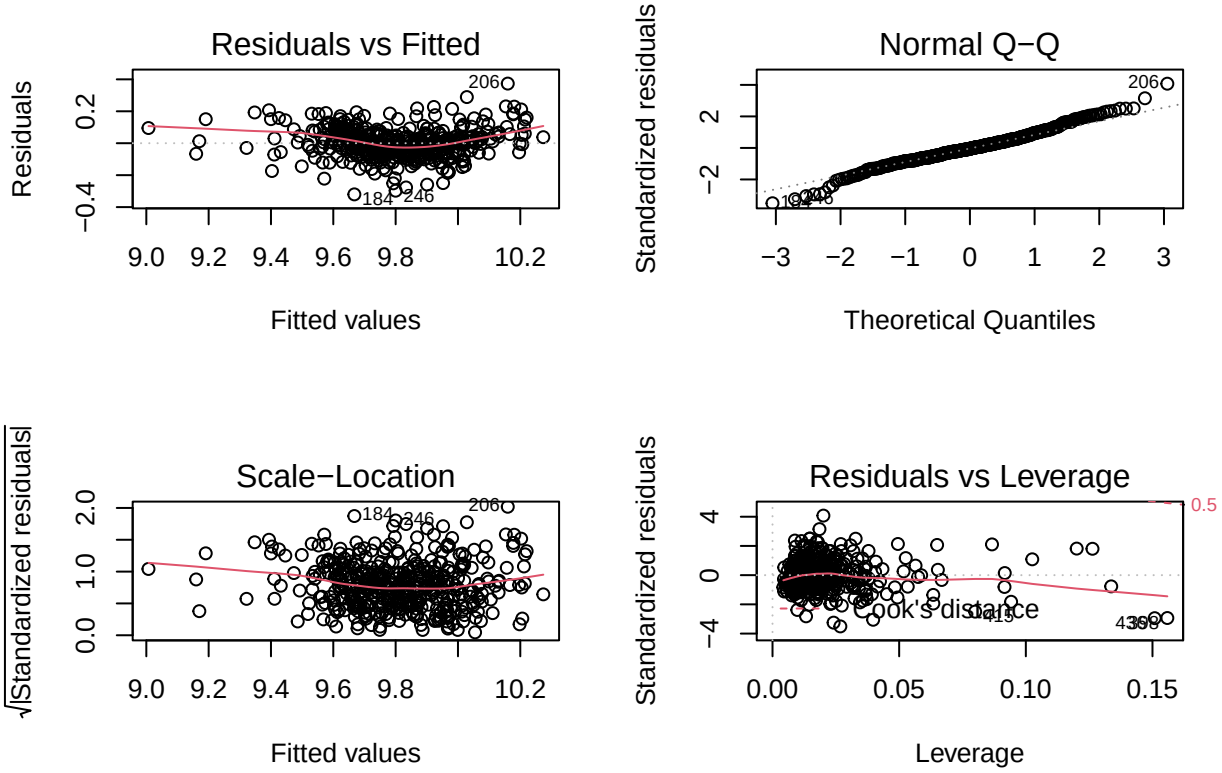
```
##                                Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)                9.735822e+00 1.551815e-01  62.738273 7.504635e-219
## log(land.area.per.capita) -3.199924e-02 5.403586e-03  -5.921852  6.510612e-09
## pct.pop.18_34             -5.815082e+02 8.105052e+01  -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)
```




```
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 ***
```

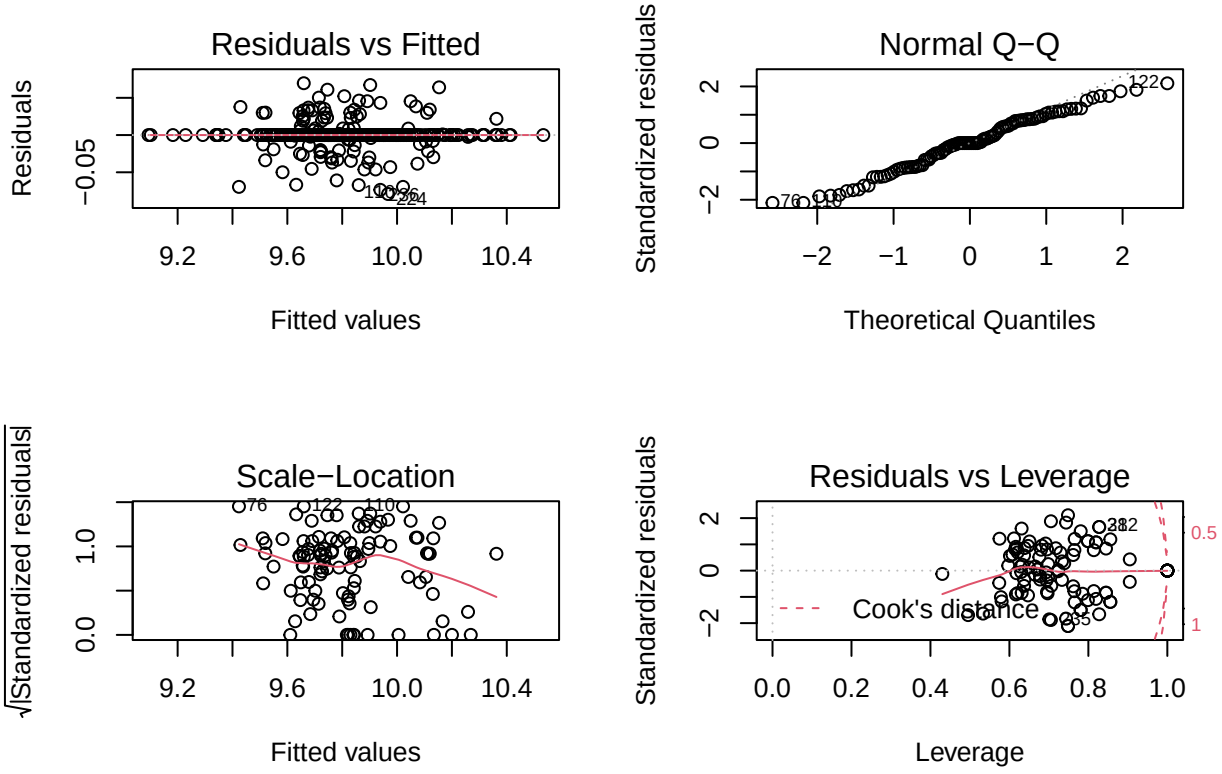
```
## pct.pop.18_34             -5.815e+02  8.105e+01  -7.175 3.19e-12 ***
## pct.pop.65_plus            7.598e+02  1.706e+02   4.453 1.08e-05 ***
## log(doctors_per_capita)    6.326e-02  1.150e-02   5.501 6.48e-08 ***
## pct.hs.grad               -5.330e-03  1.267e-03  -4.207 3.15e-05 ***
## log(pct.bach.deg)          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
```

```
final_mod_test <- lm(log(per.cap.income) ~ log(land.area.per.capita) +
    log(pct.pop.18_34) + log(pct.pop.65_plus) + log(doctors_per_capita) +
    log(pct.bach.deg) + log(pct.below.pov) + log(pct.unemp) +
    log(crimes_percapita)+ state + region + log(crimes_percapita) + county, data = cdi2)
par(mfrow=c(2,2))
plot(final_mod_test)
```

```
## Warning: not plotting observations with leverage one:
##   1, 2, 3, 4, 7, 9, 10, 11, 12, 14, 15, 17, 19, 21, 22, 23, 25, 26, 27, 31, 32, 34, 35, 36, 37, 38, 4
```

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

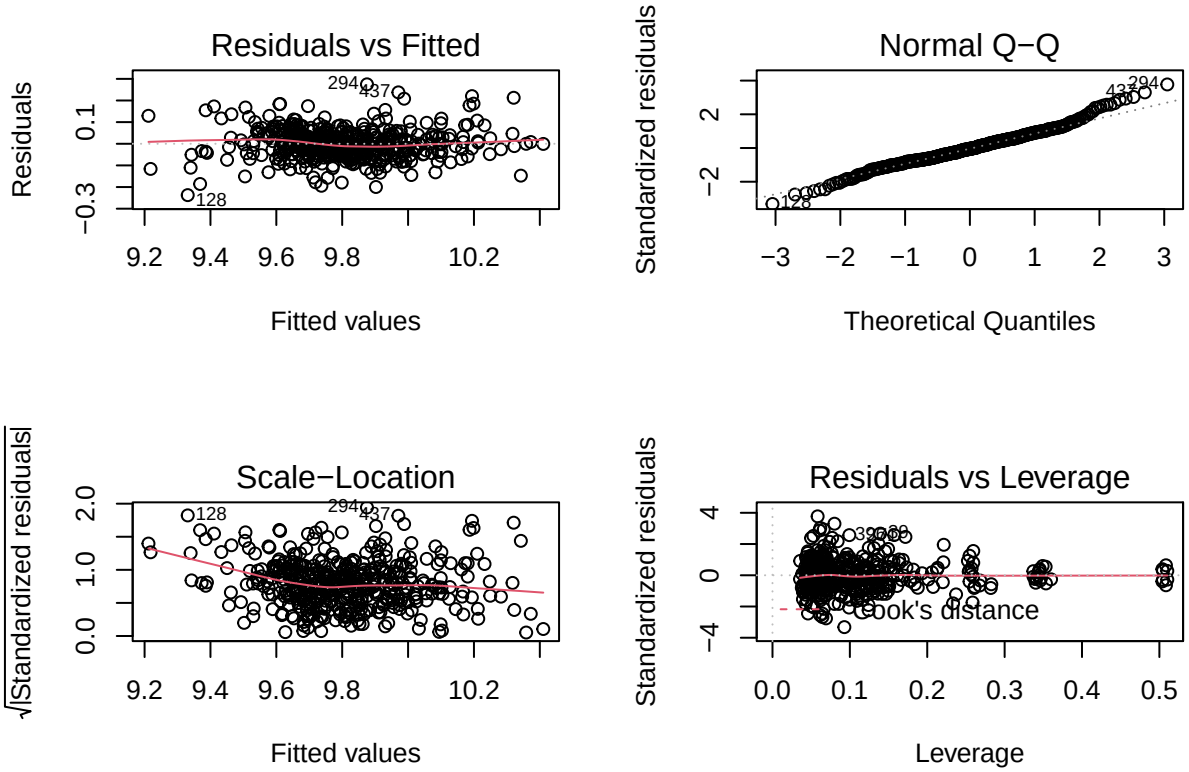
```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```




```
final_mod1 <- lm(log(per.cap.income) ~ log(land.area.per.capita) +
    log(pct.pop.18_34) + log(pct.pop.65_plus) + log(doctors_per_capita) +
    log(pct.bach.deg) + log(pct.below.pov) + log(pct.unemp) +
```

```
    log(crimes_percapita) + state + region, data = cdi2)
par(mfrow=c(2,2))
plot(final_mod1)
```

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




```
summary(final_mod1)
```

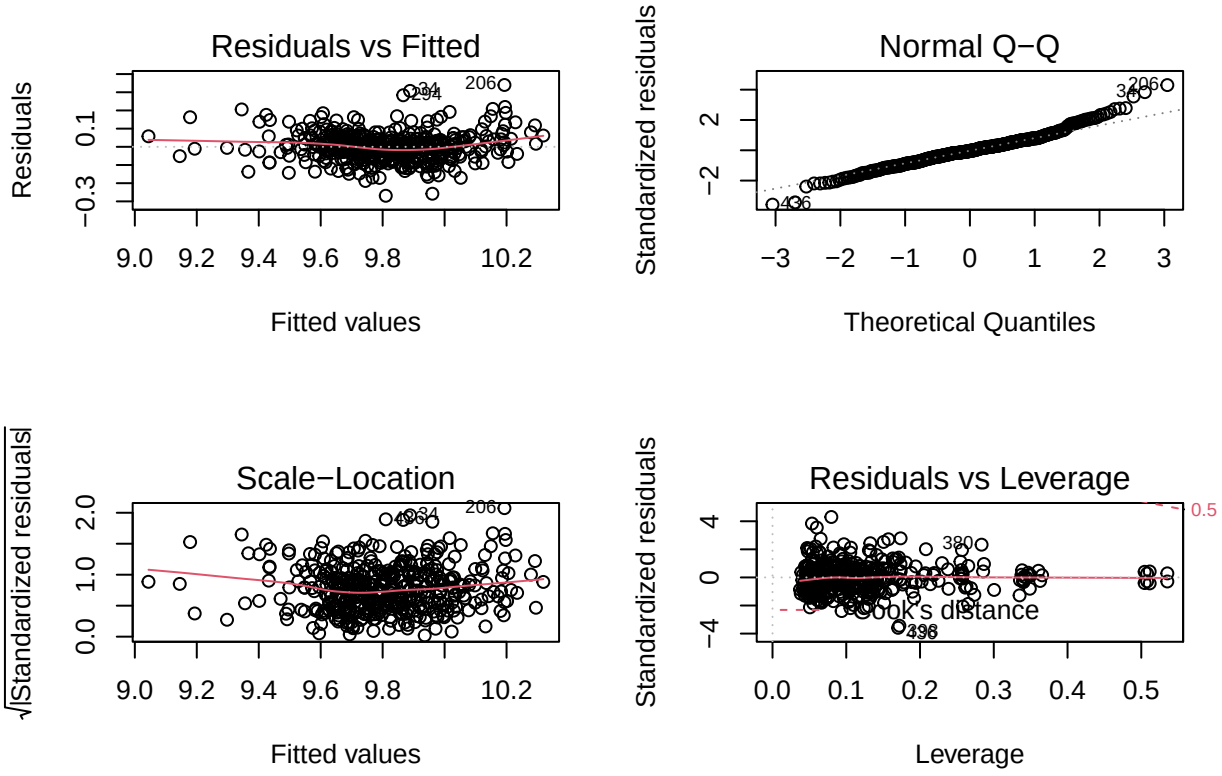
```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area.per.capita) +
##     log(pct.pop.18_34) + log(pct.pop.65_plus) + log(doctors_per_capita) +
##     log(pct.bach.deg) + log(pct.below.pov) + log(pct.unemp) +
##     log(crimes_percapita) + state + region, data = cdi2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.237306 -0.045776 -0.002368  0.039864  0.274632
##
## Coefficients: (3 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 9.7823640  0.1704602  57.388  < 2e-16 ***
## log(land.area.per.capita) -0.0287468  0.0063224  -4.547 7.31e-06 ***
## log(pct.pop.18_34)        -0.1578113  0.0150753 -10.468  < 2e-16 ***
## log(pct.pop.65_plus)       0.1392917  0.0136582  10.198  < 2e-16 ***
## log(doctors_per_capita)    0.0514221  0.0114749   4.481 9.81e-06 ***
## log(pct.bach.deg)          0.2161960  0.0235981   9.162  < 2e-16 ***
## log(pct.below.pov)        -0.2223673  0.0134163 -16.574  < 2e-16 ***
## log(pct.unemp)             0.0271168  0.0237615   1.141  0.25449
```

```
## log(crimes_percapita)      0.0324814  0.0123637   2.627  0.00896 **
## stateAR                   -0.0690588  0.0606258  -1.139  0.25537
## stateAZ                   -0.1119012  0.0460962  -2.428  0.01566 *
## stateCA                    0.0628242  0.0327307   1.919  0.05567 .
## stateCO                    0.0046225  0.0382453   0.121  0.90386
## stateCT                    0.0330740  0.0417848   0.792  0.42912
## stateDC                    0.0727294  0.0829468   0.877  0.38113
## stateDE                   -0.0187782  0.0609234  -0.308  0.75808
## stateFL                   -0.0963482  0.0330917  -2.912  0.00381 **
## stateGA                    0.0357972  0.0390405   0.917  0.35976
## stateHI                    0.0216976  0.0542356   0.400  0.68933
## stateID                   -0.0251768  0.0805299  -0.313  0.75472
## stateIL                    0.0257298  0.0346348   0.743  0.45800
## stateIN                   -0.0315552  0.0357920  -0.882  0.37853
## stateKS                   -0.0373753  0.0480582  -0.778  0.43722
## stateKY                   -0.0117124  0.0525923  -0.223  0.82389
## stateLA                    0.0253947  0.0380256   0.668  0.50464
## stateMA                    0.0007984  0.0409476   0.019  0.98445
## stateMD                   -0.0103516  0.0394484  -0.262  0.79315
## stateME                   -0.0295720  0.0450015  -0.657  0.51149
## stateMI                    0.0391573  0.0350175   1.118  0.26417
## stateMN                   -0.0377144  0.0408420  -0.923  0.35637
## stateMO                   -0.0222381  0.0396489  -0.561  0.57521
## stateMS                   -0.0541300  0.0519253  -1.042  0.29785
## stateMT                    0.0136699  0.0808567   0.169  0.86584
## stateNC                   -0.0338308  0.0344015  -0.983  0.32602
## stateND                   -0.0624228  0.0823718  -0.758  0.44903
## stateNE                   -0.0686782  0.0545918  -1.258  0.20914
## stateNH                   -0.0335185  0.0495419  -0.677  0.49909
## stateNJ                    0.0569001  0.0363777   1.564  0.11861
## stateNM                   -0.1055579  0.0607281  -1.738  0.08298 .
## stateNV                    0.1485466  0.0625724   2.374  0.01809 *
## stateNY                   -0.0153639  0.0337931  -0.455  0.64962
## stateOH                    0.0103876  0.0336007   0.309  0.75738
## stateOK                   -0.0734731  0.0472408  -1.555  0.12070
## stateOR                   -0.0862171  0.0422870  -2.039  0.04215 *
## statePA                   -0.0494212  0.0338484  -1.460  0.14509
## stateRI                   -0.1319924  0.0543371  -2.429  0.01559 *
## stateSC                   -0.0253977  0.0366808  -0.692  0.48911
## stateSD                   -0.0360845  0.0826773  -0.436  0.66276
## stateTN                   -0.0223801  0.0391475  -0.572  0.56787
## stateTX                   -0.0076533  0.0321049  -0.238  0.81171
## stateUT                   -0.2456526  0.0474778  -5.174 3.70e-07 ***
## stateVA                   -0.0053479  0.0413373  -0.129  0.89713
## stateVT                   -0.0520647  0.0820219  -0.635  0.52596
## stateWA                   -0.0426774  0.0374798  -1.139  0.25555
## stateWI                   -0.0302953  0.0376369  -0.805  0.42136
## stateWV                   -0.0114281  0.0806042  -0.142  0.88733
## regionNE                         NA         NA      NA       NA
## regionS                          NA         NA      NA       NA
## regionW                          NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.07502 on 384 degrees of freedom
## Multiple R-squared:  0.8848, Adjusted R-squared:  0.8683
## F-statistic: 53.63 on 55 and 384 DF,  p-value: < 2.2e-16
```

```
final_mod2 <- lm(log(per.cap.income)~log(land.area.per.capita)  + pct.pop.18_34 + pct.pop.65_plus + log
par(mfrow=c(2,2))
plot(final_mod2)
```

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




```
summary(final_mod2)
```

```
##
## 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 * region +
##     pct.unemp + state, data = cdi2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26912 -0.04431 -0.00176  0.04102  0.34000
##
## Coefficients: (3 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  9.681e+00  1.677e-01  57.735  < 2e-16 ***
## log(land.area.per.capita) -3.635e-02  6.677e-03  -5.443 9.40e-08 ***
## pct.pop.18_34             -6.309e+02  8.741e+01  -7.218 2.89e-12 ***
## pct.pop.65_plus            1.093e+03  1.695e+02   6.445 3.50e-10 ***
## log(doctors_per_capita)    5.568e-02  1.170e-02   4.758 2.78e-06 ***
```

```
## pct.hs.grad              -4.105e-03  1.526e-03  -2.690  0.00746 **
## log(pct.bach.deg)         2.639e-01  2.800e-02   9.425  < 2e-16 ***
## pct.below.pov            -2.526e-02  2.962e-03  -8.529 3.53e-16 ***
## regionNE                 -2.878e-02  9.027e-02  -0.319  0.75002
## regionS                  -2.552e-02  4.699e-02  -0.543  0.58740
## regionW                  -5.090e-02  4.667e-02  -1.091  0.27615
## pct.unemp                 5.062e-03  3.472e-03   1.458  0.14571
## stateAR                  -6.501e-02  6.636e-02  -0.980  0.32790
## stateAZ                  -7.866e-02  4.800e-02  -1.639  0.10206
## stateCA                   7.459e-02  3.109e-02   2.399  0.01691 *
## stateCO                   2.415e-02  3.834e-02   0.630  0.52918
## stateCT                   1.194e-01  8.872e-02   1.346  0.17918
## stateDC                   2.296e-02  9.052e-02   0.254  0.79991
## stateDE                   4.379e-02  1.020e-01   0.429  0.66798
## stateFL                  -4.944e-02  3.658e-02  -1.352  0.17724
## stateGA                   2.722e-02  4.245e-02   0.641  0.52178
## stateHI                   5.971e-02  5.606e-02   1.065  0.28750
## stateID                   1.834e-02  8.663e-02   0.212  0.83242
## stateIL                   4.761e-02  3.273e-02   1.455  0.14661
## stateIN                  -2.754e-02  3.383e-02  -0.814  0.41614
## stateKS                   1.036e-02  4.852e-02   0.213  0.83109
## stateKY                  -3.195e-02  5.774e-02  -0.553  0.58032
## stateLA                   5.926e-02  4.274e-02   1.386  0.16645
## stateMA                   6.193e-02  8.822e-02   0.702  0.48314
## stateMD                   5.275e-02  4.287e-02   1.230  0.21929
## stateME                   4.759e-02  9.137e-02   0.521  0.60274
## stateMI                   5.785e-02  3.407e-02   1.698  0.09033 .
## stateMN                  -1.555e-02  4.031e-02  -0.386  0.69992
## stateMO                   4.911e-03  3.910e-02   0.126  0.90011
## stateMS                  -3.458e-02  5.757e-02  -0.601  0.54841
## stateMT                   4.455e-02  8.705e-02   0.512  0.60908
## stateNC                  -4.132e-02  3.757e-02  -1.100  0.27218
## stateND                  -4.195e-02  8.763e-02  -0.479  0.63246
## stateNE                  -5.224e-02  5.490e-02  -0.951  0.34197
## stateNH                   5.319e-02  9.326e-02   0.570  0.56877
## stateNJ                   1.565e-01  8.656e-02   1.808  0.07146 .
## stateNM                  -9.192e-02  6.845e-02  -1.343  0.18014
## stateNV                   1.901e-01  6.550e-02   2.902  0.00392 **
## stateNY                   6.760e-02  8.575e-02   0.788  0.43097
## stateOH                   8.883e-03  3.101e-02   0.286  0.77466
## stateOK                  -6.214e-02  5.276e-02  -1.178  0.23961
## stateOR                  -3.274e-02  4.272e-02  -0.766  0.44397
## statePA                   3.806e-02  8.574e-02   0.444  0.65736
## stateRI                  -5.236e-02  9.681e-02  -0.541  0.58892
## stateSC                  -2.950e-02  4.012e-02  -0.735  0.46269
## stateSD                  -6.996e-03  8.677e-02  -0.081  0.93578
## stateTN                  -3.047e-02  4.304e-02  -0.708  0.47940
## stateTX                   7.995e-03  3.532e-02   0.226  0.82107
## stateUT                  -2.450e-01  4.926e-02  -4.975 9.91e-07 ***
## stateVA                  -2.030e-03  4.549e-02  -0.045  0.96443
## stateVT                         NA         NA      NA       NA
## stateWA                         NA         NA      NA       NA
## stateWI                         NA         NA      NA       NA
## stateWV                   8.207e-03  8.843e-02   0.093  0.92611
```

```
## pct.below.pov:regionNE    -5.114e-03  3.680e-03  -1.390  0.16544
## pct.below.pov:regionS      1.868e-03  3.008e-03   0.621  0.53490
## pct.below.pov:regionW      4.594e-03  4.023e-03   1.142  0.25412
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08237 on 381 degrees of freedom
## Multiple R-squared:  0.8622, Adjusted R-squared:  0.8413
## F-statistic: 41.11 on 58 and 381 DF,  p-value: < 2.2e-16
```

```
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
```

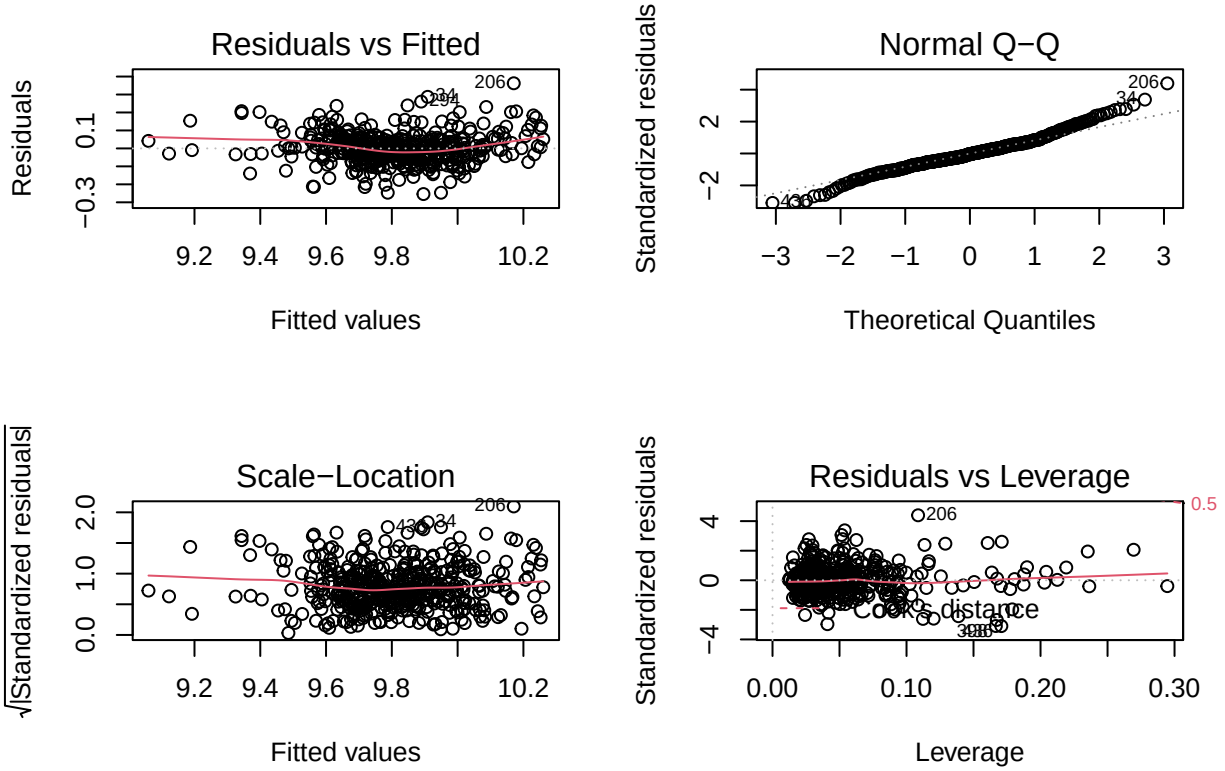
```
## regionW:pct.pop.18_34              8.098e+02  3.217e+02   2.517 0.012222 *
## regionNE:log.land.area.per.capita  8.029e-03  2.218e-02   0.362 0.717585
## regionS:log.land.area.per.capita   2.454e-02  1.948e-02   1.260 0.208575
## regionW:log.land.area.per.capita   3.948e-02  2.085e-02   1.894 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
```

```
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
```

```
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
```

```
##   0.09101312  0.8035181  0.06815723
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

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.