Regression Analysis to Predict Per Capita Income

Ziyan Xia[1]
zxia2@andrew.cmu.edu

Abstract: Some social scientists are interested in looking at historical data to learn how average income per person was related to other variables associated with the county's economic, health and social well-being. To answer relevant question, use a county demographic information (CDI) for 440 of the most populous counties in the United State to do regression analysis. Method including ANOVA, All Subset and Stepwise Regression to select the best model to predict capita income using the variables in the CDI data. The results shows the best model to predict per capita income and the variable selection, transformation all makes sense. However, further research about detecting high-order interactions and whether to include more states and county data should be done

1. Introduction

Some social scientists are interested in looking at historical data to learn how average income per person was related to other variables associated with the county's economic, health and social well-being. There are four questions to answer by this analysis:

1. Looking at the data one pair of variables at a time, which variables seem to be related to which other variables in the data? Which are not? Are all of the relationships what a reasonable person would expect, or are there some surprises? Can you explain these findings in terms of the meanings of the variables?

2. There is a theory that, if we ignore all other variables, per-capita income should be related to crime rate, and that this relationship may be different in different regions of the country (Northeast, Northcentral, South, and West). What do the data say? Does it matter if you use number of crimes, or (number of crimes)/(population), in your analysis?

3. Find the best model predicting per-capita income from the other variables (including possible transformations, interactions, etc.). Here "best" means a good compromise between
   • Best reflects the social science and the meaning of the variables
   • Best satisfies modeling assumptions
   • Is most clearly indicated by the data

[1] Department of Statistics and Data Science,Carnegie Mellon University

• Can be explained to someone who is more interested in social, economic and health factors than in mathematics and statistics.

4. A county is a governmental unit in the United States that is bigger than a city but smaller than a state. There are 50 states in the US, plus the District of Columbia, which is usually coded as a 51st state in data like this. There are 48 states represented in the data. There are approximately 3000 counties in the US, and 373 represented in the data set. Should we be worried about either the missing states or the missing counties? Why or why not?

2. Data

The data is taken from Kutner et al (2005)[i]. It provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The definition is as the table below.

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

Variable Number	Variable Name	Description
1	Identification number	1–440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 or old
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990

9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI population that is unemployed
15	Per capita income	Per-capita income (i.e. average income per person) of 1990 CDI population (in dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification used by the US Bureau of the Census, NE (northeast region of the US), NC (north-central region of the US), S (southern region of the US), and W (Western region of the US)

Table 2: Combine County with Sate

Counties 1-110	Counties 111-220	Counties 221-330	Counties 331-440
Ada ID	Ector TX	Lycoming PA	Rockingham NH
Adams CO	El_Dorado CA	Macomb MI	Rockland NY
Aiken SC	El_Paso CO	Macon IL	Rowan NC
Alachua FL	El_Paso TX	Madison AL	Rutherford TN
Alamance NC	Elkhart IN	Madison IL	Sacramento CA
Alameda CA	Erie NY	Madison IN	Saginaw MI
Albany NY	Erie PA	Mahoning OH	Salt_Lake UT
Alexandria_City VA	Escambia FL	Manatee FL	San_Bernardino CA
Allegheny PA	Essex MA	Marathon WI	San_Diego CA
Allen IN	Essex NJ	Maricopa AZ	San_Francisco CA
Allen OH	Fairfax_County VA	Marin CA	San_Joaquin CA
Anderson SC	Fairfield CT	Marion FL	San_Luis_Obispo CA
Androscoggin ME	Fairfield OH	Marion IN	San_Mateo CA
Anne_Arundel MD	Fayette KY	Marion OR	Sangamon IL
Arapahoe CO	Fayette PA	Martin FL	Santa_Barbara CA
Arlington_County VA	Florence SC	Maui HI	Santa_Clara CA
Atlantic NJ	Forsyth NC	McHenry IL	Santa_Cruz CA
Baltimore MD	Fort_Bend TX	McLean IL	Sarasota FL
Baltimore_City MD	Franklin OH	McLennan TX	Saratoga NY
Barnstable MA	Franklin PA	Mecklenburg NC	Sarpy NE
Bay FL	Frederick MD	Medina OH	Schenectady NY
Bay MI	Fresno CA	Merced CA	Schuylkill PA
Beaver PA	Fulton GA	Mercer NJ	Sedgwick KS
Bell TX	Galveston TX	Mercer PA	Seminole FL
Benton WA	Gaston NC	Merrimack NH	Shasta CA
Bergen NJ	Genesee MI	Middlesex CT	Shawnee KS
Berks PA	Gloucester NJ	Middlesex MA	Sheboygan WI
Berkshire MA	Greene MO	Middlesex NJ	Shelby TN
Bernalillo NM	Greene OH	Midland TX	Smith TX
Berrien MI	Greenville SC	Milwaukee WI	Snohomish WA

Table 3 : The first 5 row of the dataset

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

Table 3: The first 5 row of the dataset

id	pct.hs.grad	pct.bach.deg	pct.below.pov	pct.unemp	per.cap.income	tot.income	region
1	70.0	22.3	11.6	8.0	20786	184230	W
2	73.4	22.8	11.1	7.2	21729	110928	NC
3	74.9	25.4	12.5	5.7	19517	55003	S
4	81.9	25.3	8.1	6.1	19588	48931	W
5	81.2	27.8	5.2	4.8	24400	58818	W
6	63.7	16.6	19.5	9.5	16803	38658	NE

Table 4: Summary Statistics for all variables, including continuous variables and categorical variables

id	county	state	land.area	pop	pop.18_34	pop.65_plus	doctors	hosp.beds
Min. : 1.0	Length:440	Length:440	Min. : 15.0	Min. : 100043	Min. :16.40	Min. : 3.000	Min. : 39.0	Min. : 92.0
1st Qu.:110.8	Class :character	Class :character	1st Qu.: 451.2	1st Qu.: 139027	1st Qu.:26.20	1st Qu.: 9.875	1st Qu.: 182.8	1st Qu.: 390.8
Median :220.5	Mode :character	Mode :character	Median : 656.5	Median : 217280	Median :28.10	Median :11.750	Median : 401.0	Median : 755.0
Mean :220.5	NA	NA	Mean : 1041.4	Mean : 393011	Mean :28.57	Mean :12.170	Mean : 988.0	Mean : 1458.6
3rd Qu.:330.2	NA	NA	3rd Qu.: 946.8	3rd Qu.: 436064	3rd Qu.:30.02	3rd Qu.:13.625	3rd Qu.: 1036.0	3rd Qu.: 1575.8
Max. :440.0	NA	NA	Max. :20062.0	Max. :8863164	Max. :49.70	Max. :33.800	Max. :23677.0	Max. :27700.0

crimes	pct.hs.grad	pct.bach.deg	pct.below.pov	pct.unemp	per.cap.income	tot.income	region
Min. : 563	Min. :46.60	Min. : 8.10	Min. : 1.400	Min. : 2.200	Min. : 8899	Min. : 1141	Length:440
1st Qu.: 6220	1st Qu.:73.88	1st Qu.:15.28	1st Qu.: 5.300	1st Qu.: 5.100	1st Qu.:16118	1st Qu.: 2311	Class :character
Median : 11820	Median :77.70	Median :19.70	Median : 7.900	Median : 6.200	Median :17759	Median : 3857	Mode :character
Mean : 27112	Mean :77.56	Mean :21.08	Mean : 8.721	Mean : 6.597	Mean :18561	Mean : 7869	NA
3rd Qu.: 26280	3rd Qu.:82.40	3rd Qu.:25.32	3rd Qu.:10.900	3rd Qu.: 7.500	3rd Qu.:20270	3rd Qu.: 8654	NA
Max. :688936	Max. :92.90	Max. :52.30	Max. :36.300	Max. :21.300	Max. :37541	Max. :184230	NA

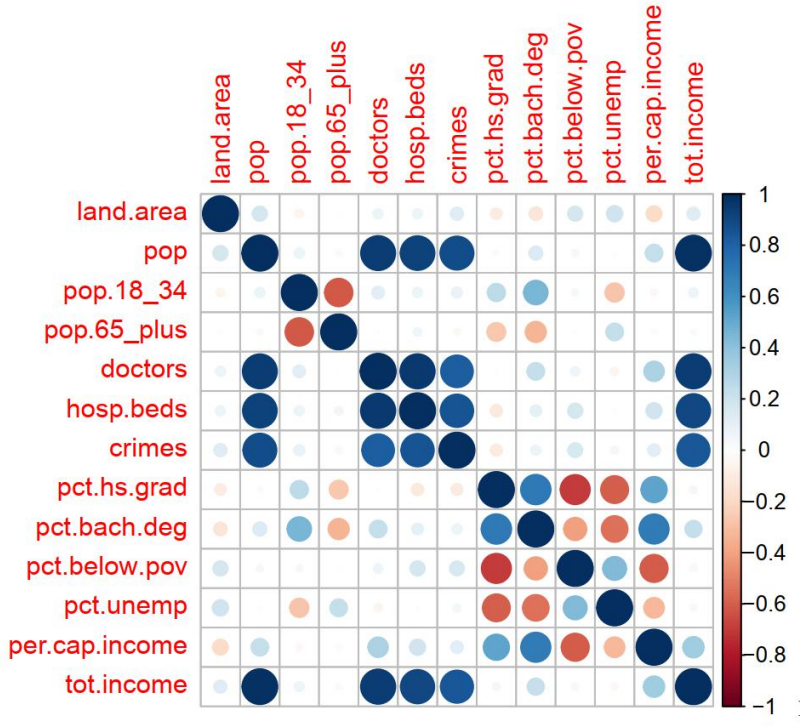
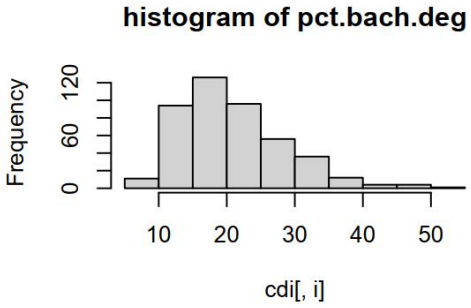
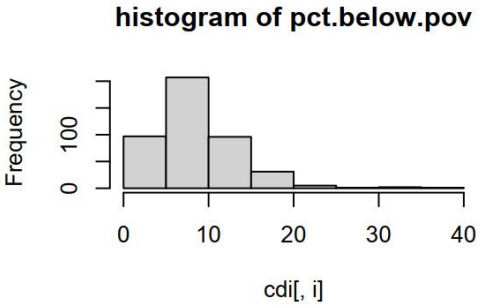
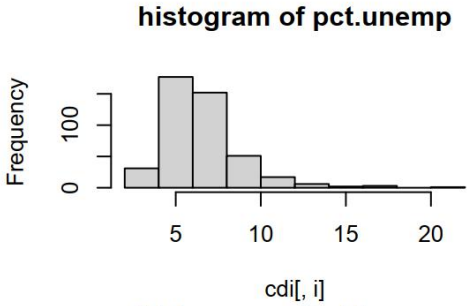


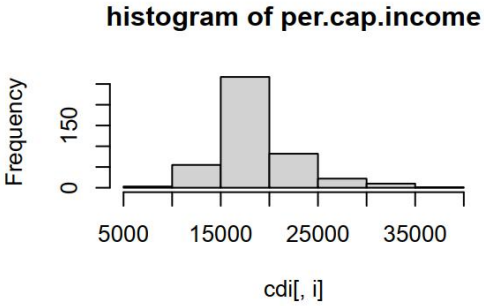
Figure 1: The Correlation Plot of all continuous variables

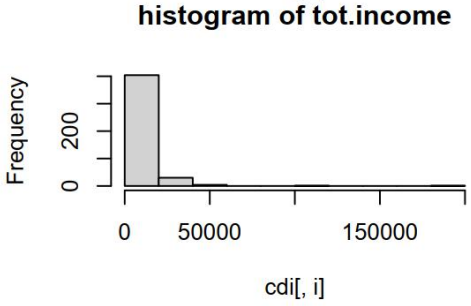
For the correlation plot, there are many variables that are highly correlated with other variables, for example, the correlation between pop and doctors is really high.


histogram of pct.bach.deg
Frequency
0
60
120
10
20
30
40
50
cdi[, i]


histogram of pct.below.pov
Frequency
0
100
0
10
20
30
40
cdi[, i]


histogram of pct.unemp
Frequency
0
100
5
10
15
20
cdi[, i]


histogram of per.cap.income
Frequency
0
150
5000
15000
25000
35000
cdi[, i]


histogram of tot.income
Frequency
0
200
0
50000
150000
cdi[, i]

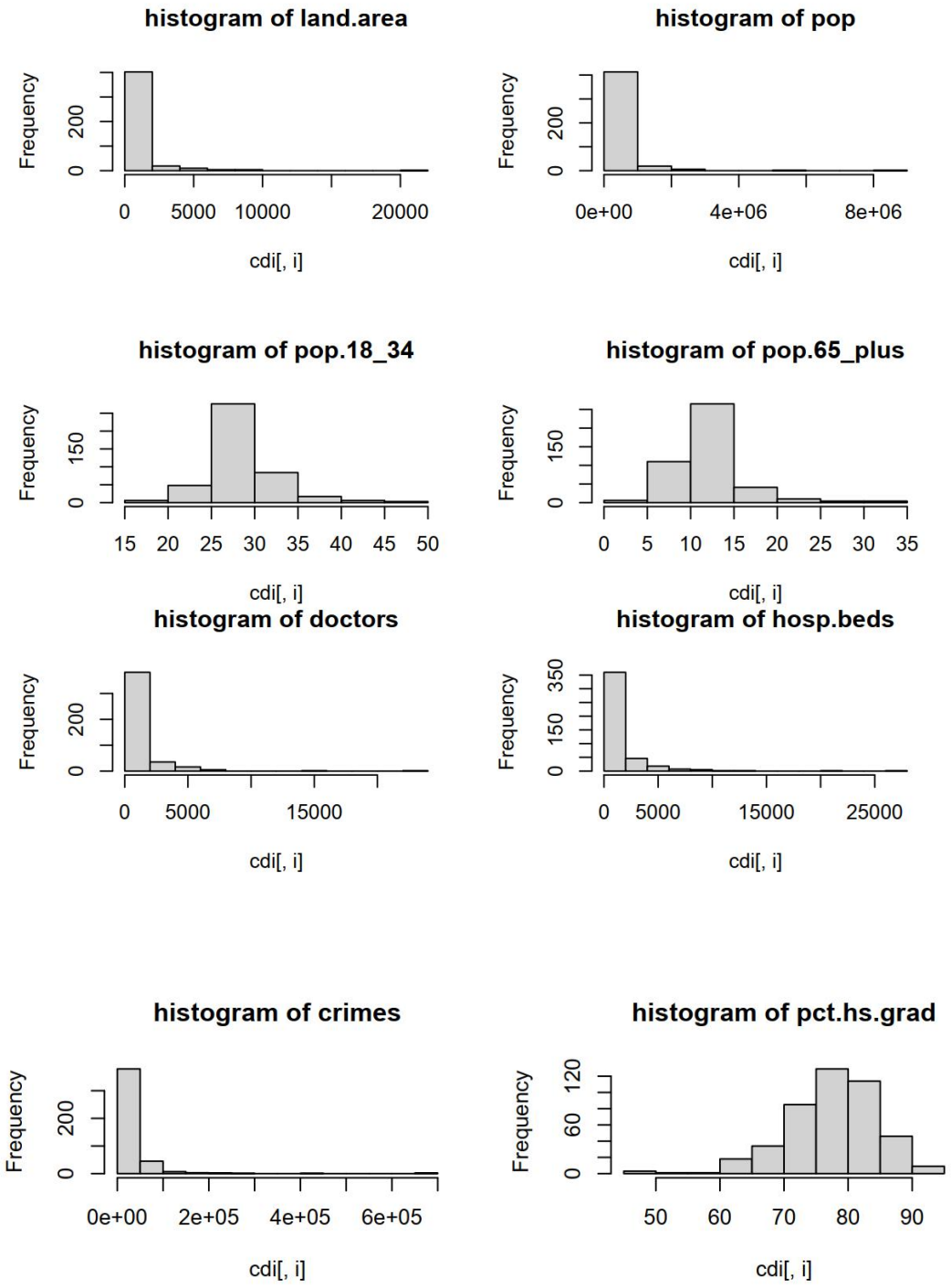


Figure 2: Histograms of all continuous variables

As is shown from those histograms, there are a lot of variables that don't have normal distribution. The distributions of variable land.area, pop, doctors, hosp.beds, crimes, tot.income are all extremely right skewed and the distributions of variable pct.bach.deg and pct.below.pov are quite right skewed.

4. Methods

To answer the first question, a correlational plot of all continuous variables is made to discover the relationship between variables.

To answer the second question, log transformation was applied to variables Per Capita Income, Total serious crimes, Per Capita Income, Total Population are used to fit several linear models. and we use ANOVA method to find the best fitted model to predict Per Capita Income by these variables.

To answer the third question, after taking some variables that's not suitable for regression analysis out, do transformation to variables that have skewed distribution. Then use continuous variables in the transformed data to fit a full model that regress per capita income on all the other variables. After that, use all subsets, stepwise regression and lasso methods to do variable selection and then use ANOVA to test whether to include interaction between the categorical variable and the continuous variable. Finally, use diagnostics plots, add variables plot and marginal plot to test whether the variable selection makes sense and whether this model is a good fit.

To answer the last question, a table combining county with state could be helpful.

5. Results

To answer question 1, as is shown from Figure 1,[2] there are many variables that are highly correlated with other variables, for example, the correlation between pop and doctors is really high. There are several pairs of variables that have really high positive correlation and there are several pairs that have really high negative correlation.

Table 5: correlation of the variables

Variables	obvious positive correlation	obvious negative correlation
pop	**Doctors** **hosp.beds** **crimes** **tot.income**	
pop.18-34	**pop.65_plus**	
hosp.beds	**Pop** **hosp.beds** **crimes** **tot.income**	

[2] See Page 3 Technical Appendix

doctors	**pop** **hosp.beds** **crimes** **tot.income**	
crimes	**Pop** **hosp.beds** **doctors** **tot.income**	
pct.hs.grad	**pct.bach.deg**	**pct.below.pov** **pct.unemp**
pct.below.pov	**pct.hs.grad**	**per.cap.income**

To answer question 2, I created 4 models to address question 2: regress per capita income on total crimes and region (Model 1), regress per capita income on total crimes with interaction between total crimes and region (Model 2), regress per capita income on crime rate and region (Model 3), regress per capita income on crime rate and region with interaction between crime rate and region.

Table 4 shows the adjusted R squared for these 4 models and table 5 shows the F statistics for two ANOVA test, one is to test whether to include the interaction between crimes and region, the other is to test whether to include the interaction between crime rate and region.

Table 6: Adjusted R Squared for 4 models

Model	Adjusted R Squared
per.cap.income~crimes	0.070
per.cap.income~crimes+region	0.093
per.cap.income~crimes+region+crimes*region	0.095
per.cap.income~crime.rate+region	0.078
per.cap.income~crime.rate+region++crime.rate*region	0.072

Table 7: F Statistics for ANOVA test

ANOVA test models	F Statistics
per.cap.income~crimes+region **vs** per.cap.income~crimes+region+crimes*region	0.2396
per.cap.income~crime.rate+region **vs** per.cap.income~crime.rate+region+crime.rate*region	0.9888

per.cap.income~crimes **vs** per.cap.income~crimes+region	1.784e−12
per.cap.income~crime.rate **vs** per.cap.income~crime.rate+region	1.09e−11

From Table 1, as per capita income is actually just total personal income/population, so we cannot use total personal income in our regression analysis, also for this reason, we should exclude total population in the fitted model. Besides, the ID, Country, State will be no helpful for prediction but only added complexity to the model so it will be better to exclude them in the regression analysis.

To answer question 3, I first get a model selected by both All Subsets method and Stepwise regression method because the model selected by both methods is the same. This model regress **Per Capita Income** on **transformed Land area, Percent of population aged 18–34, Number of active physicians, Percent high school graduates, Percent bachelor's degrees, Percent below poverty level, Percent unemployment**. To test whether the variable transformation and selection like this makes sense, I made added-variable plots and marginal model plots for this model.

From Figure 3, All marginal plots show that the non-parametric lines agree with the regression lines and it indicates that the response variable and the predictors don't need to be transformed.

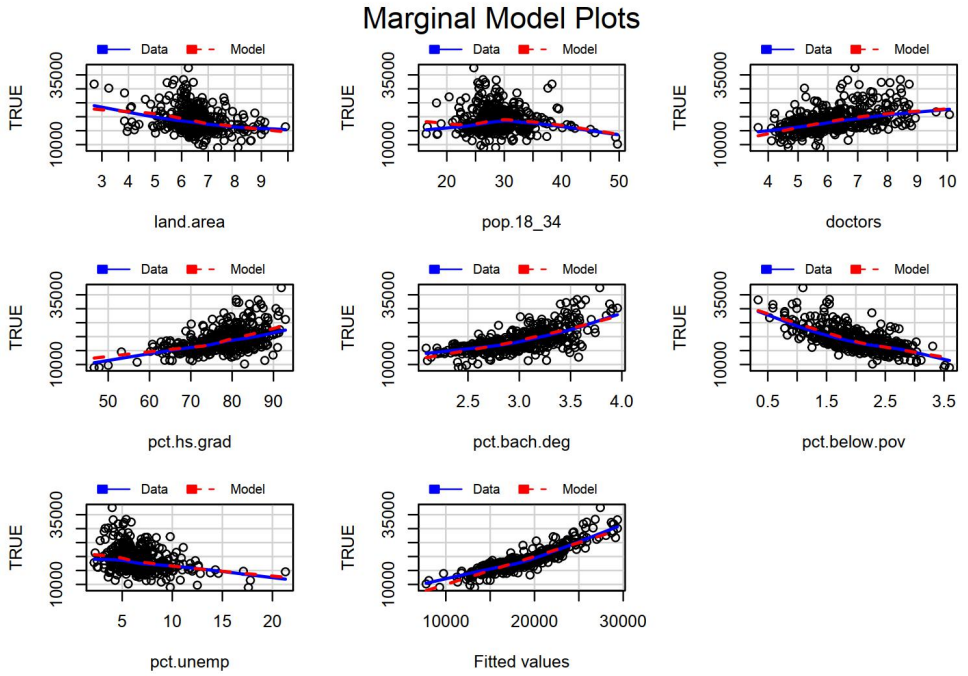


Figure 3: Marginal Model Plot

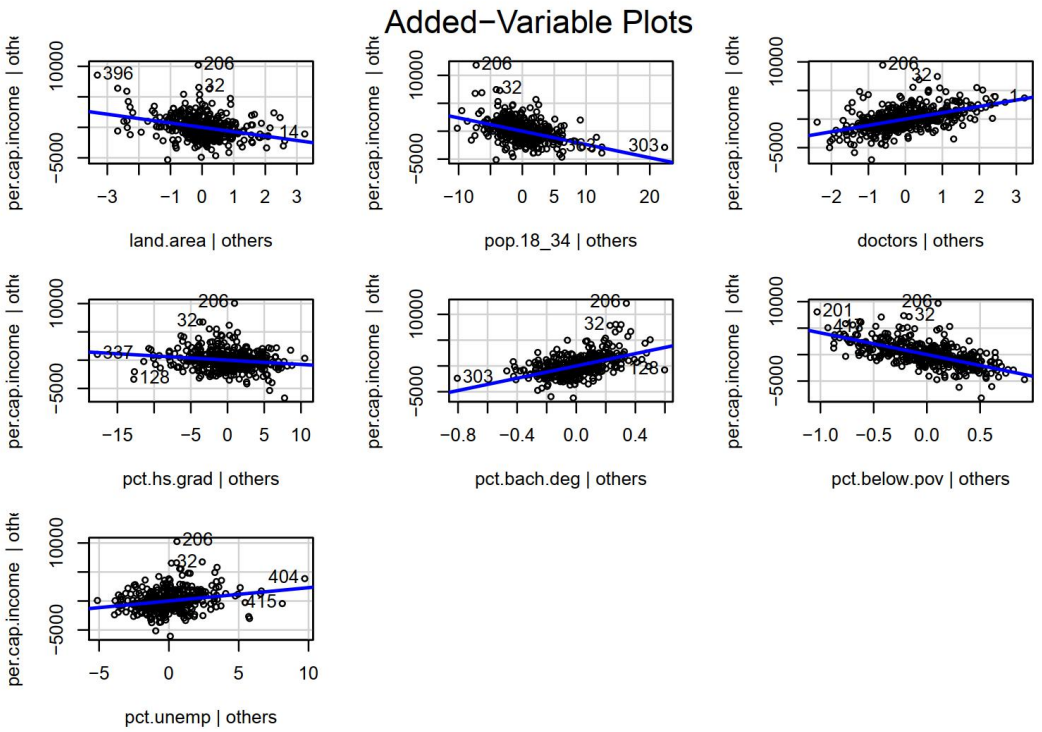


Figure 4: Added-Variable Plot

From Figure 4, Added-Variable plots show that the model won't need any transformation to the variables included in the model.

Table : Multiple AVONA Test Result

Interaction	ANOVA F Statistcs
Land area*region	0.142
Percent of population aged 18–34*region	0.02842
Number of active physicians*region	0.03357
Percent high school graduates *region	0.0002591
Percent bachelor's degrees *region	0.0002673
Percent below poverty level *region	0.006566
Percent unemployment *region	9.305e−05

Based on the F Statistics of ANOVA test, I decided to keep the interaction Percent high school graduates *region, Percent bachelor's degrees *region and Percent below poverty level *region.

Therefore, the final model is to regress **Per Capita Income** on **transformed Land area, Percent of population aged 18–34, Number of active physicians, Percent high school graduates, Percent bachelor's degrees, Percent below poverty level, Percent unemployment** with interaction **Percent high school graduates *region, Percent bachelor's degrees *region** and **Percent below poverty level *region. (**Coefficients' detail is in the technical appendix: summary of final model[3]).
Figure is the diagnostics plot for this final model.

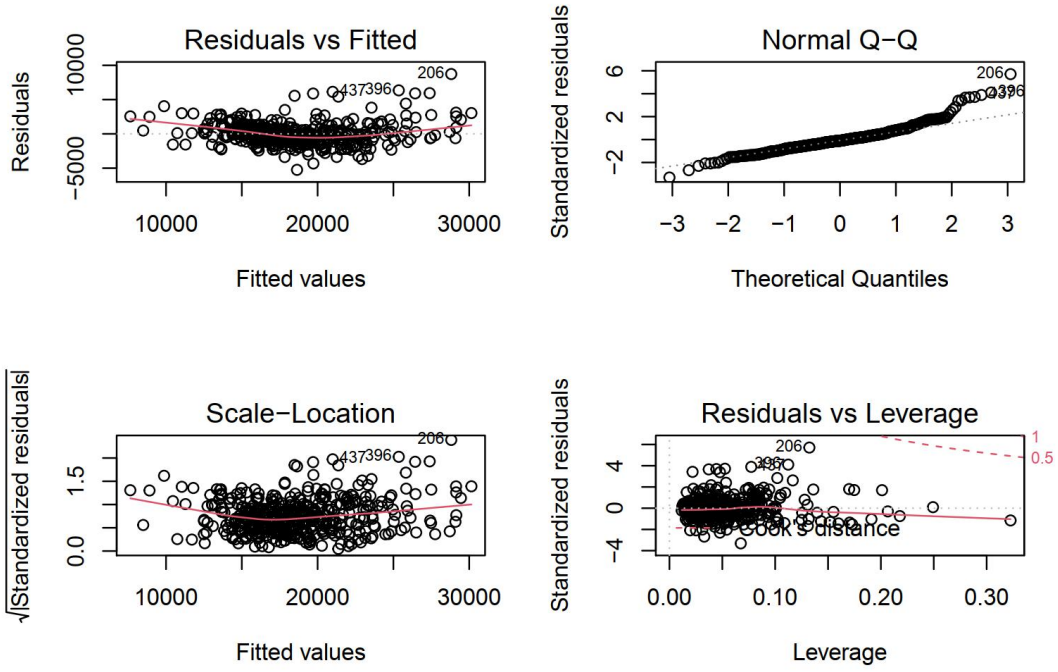


Figure 5: Diagnostics Plot

For the diagnostics plot of our final model, the Residuals vs Fitted plot shows that there is not nonlinear pattern in the model but there are some outliers detected; the normal Q-Q plot shows that the residuals are normally distributed and the assumption that errors are normally distributed holds, but there are some points that deviated from the diagonal line a lot; the scale-location plot shows that the residuals are spread equally
along the ranges of predictors and the residuals have constant variance; the residuals vs leverage plot shows that there are no influential points in the model.

To answer question 4, scanning through table 2 of county & state names, Allen is the first county name that appears in multiple states. And a little further down, both "Baltimore MD" and "Baltimore City MD" are listed.

6. Discussion

[3] See Page 17 on Technical Appendix

Table 5 shows some reasonable relationship between variables. For example, total population is related with number of doctors, total crimes and also total incomes, the higher total population is, the higher number of doctors, total crimes and also total incomes. Also, it makes sense that the higher percentage of high school graduates, the lower the percentage of population below poverty level and the percentage of unemployment. However, it is surprising that crimes is positively related with number of physicians and the number of hospital beds. This is a interesting social situation and maybe worth further investigation to test whether this correlation will vary by different region, state and county.

Table 7 shows that there is no need to include interaction in the model to predict per capita income use crimes/crime rate and region but it's better to include the categorical region in our analysis. Table 6 shows the best model, which in this analysis, means the model with the largest adjusted r squared, which is the model that regresses per capital income on total crimes and region, is the best model here. This answered the second question that if we ignore all other variables, per-capita income should be related to crime rate, and that this relationship may be different in different regions of the country (Northeast, Northcentral, South, and West). Also, It's better to use number of crimes than (number of crimes)/(population) here.

The best model for question 3 is the model that regresses **Per Capita Income** on **logged Land area, Percent of population aged 18–34, logged Number of active physicians, Percent high school graduates, logged Percent bachelor's degrees, logged Percent below poverty level, Percent unemployment** with interaction **Percent high school graduates *region, logged Percent bachelor's degrees *region** and logged **Percent below poverty level *region.** Based on Figure 3,4,5 in results section, the model is overall a good fit. However, for this analysis, I only tested pairwise interactions but didn't test whether there should be higher order interactions. It seems hard to test higher order interactions just use regression analysis so for further research, more methods may be applied to this dataset to do this.

Both "Baltimore MD" and "Baltimore City MD" are listed in table 2, which makes me wonder these two data points are really independent. Therefore, if adding more county data, they may be highly relevant just like these two. However, as I mentioned before, some variables may vary between states and counties, the distribution of these variables may be different in different states and counties. So to answer whether more data should be added still need further research.

[i] Kutner et al. (2005). Original source: Geospatial and Statistical Data Center, University of Virginia.

Technical Appendix

Ziyan Xia

10/18/2021

```
library(arm)      # rescale
```

```
## Loading required package: MASS
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.11-2, built: 2020-7-27)
```

```
## Working directory is /Users/ceciliaxia/Desktop
```

```
library(glmnet)   # for glmnet and cv.glmnet
```

```
## Loaded glmnet 4.1-2
```

```
library(foreign)  # read.dta
library(MASS)     # stepAIC
library(leaps)    # regsubsets
library(car)      # subsets
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:arm':
##
##     logit
```

```
library(boot)     # for cv.glm()$delta[1]
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
## The following object is masked from 'package:arm':
##
##     logit
```

Section Data:

summary table:

```
cdi<-read.table("/Users/ceciliaxia/Desktop/cdi.dat")
summary_cdi<-as.data.frame.matrix(summary(cdi))
```

	id	county	state	land.area	pop	pop.18_34	pop.65_plus	doctors	hosp.beds	crimes	pct.hs.grad	pct.bach.deg	pct.below.pov	pct.unemp	per.cap.income	tot.income	region
X	Min. : 1.0	Length:440	Length:440	Min. : 15.0	Min. : 100043	Min. :16.40	Min. : 3.000	Min. : 39.0	Min. : 92.0	Min. : 563	Min. :46.60	Min. : 8.10	Min. : 1.400	Min. : 2.200	Min. : 8899	Min. : 1141	Length:440
X.1	1st Qu.:110.8	Class :character	Class :character	1st Qu.: 451.2	1st Qu.: 139027	1st Qu.:26.20	1st Qu.: 9.875	1st Qu.: 182.8	1st Qu.: 390.8	1st Qu.: 6220	1st Qu.:73.88	1st Qu.:15.28	1st Qu.: 5.300	1st Qu.: 5.100	1st Qu.:16118	1st Qu.: 2311	Class :character
X.2	Median :220.5	Mode :character	Mode :character	Median : 656.5	Median : 217280	Median :28.10	Median :11.750	Median : 401.0	Median : 755.0	Median : 11820	Median :77.70	Median :19.70	Median : 7.900	Median : 6.200	Median :17759	Median : 3857	Mode :character
X.3	Mean :220.5	NA	NA	Mean : 1041.4	Mean : 393011	Mean :28.57	Mean :12.170	Mean : 988.0	Mean : 1458.6	Mean : 27112	Mean :77.56	Mean :21.08	Mean : 8.721	Mean : 6.597	Mean :18561	Mean : 7869	NA
X.4	3rd Qu.:330.2	NA	NA	3rd Qu.: 946.8	3rd Qu.: 436064	3rd Qu.:30.02	3rd Qu.:13.625	3rd Qu.: 1036.0	3rd Qu.: 1575.8	3rd Qu.: 26280	3rd Qu.:82.40	3rd Qu.:25.32	3rd Qu.:10.900	3rd Qu.: 7.500	3rd Qu.:20270	3rd Qu.: 8654	NA
X.5	Max. :440.0	NA	NA	Max. :20062.0	Max. :8863164	Max. :49.70	Max. :33.800	Max. :23677.0	Max. :27700.0	Max. :688936	Max. :92.90	Max. :52.30	Max. :36.300	Max. :21.300	Max. :37541	Max. :184230	NA

```
library(knitr)
library(dplyr)
```

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

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

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

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

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

```
library(kableExtra)
```

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

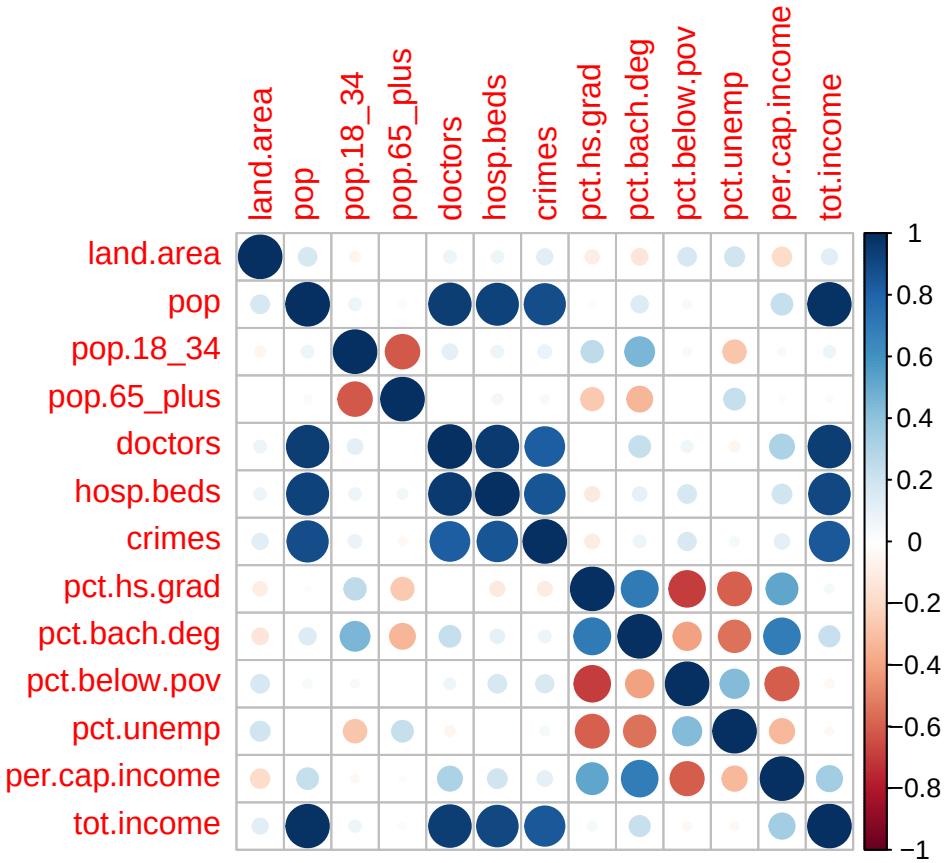
```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
kable(summary_cdi, "latex", booktabs = T) %>%
kable_styling(latex_options = c("striped", "scale_down"))
```

There is missing data in variable Country, state and Geographic Regions. Each of these variables have three missing values because they are all categorical variables whose summary statistics only contain three types.

correlation plot all continuous variables:

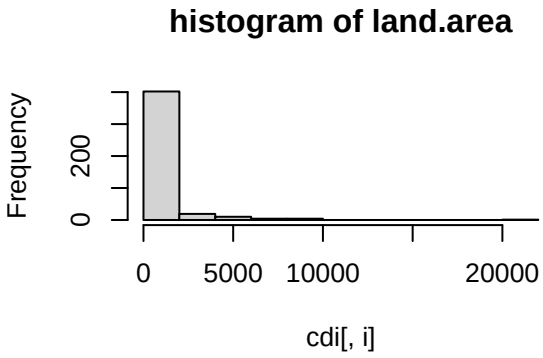
```
corrplot::corrplot(cor(cdi[,c(4:16)]))
```

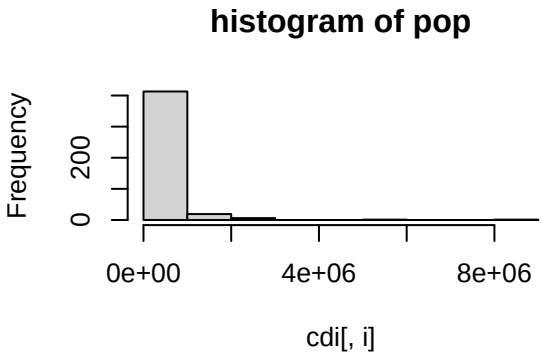


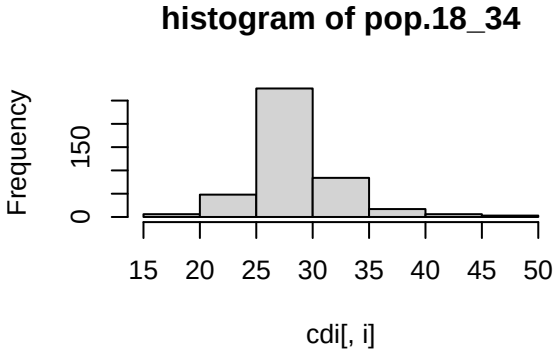
For the correlation plot, there are many variables that are highly correlated with other variables, for example, the correlation between pop and doctors is really high.

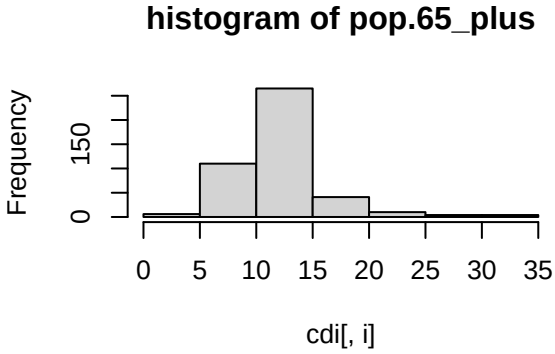
histograms of all continuous variables:

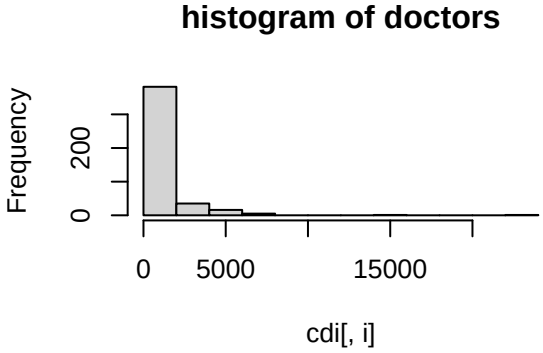
```
par(mfrow=c(2,2))
par(mfrow=c(2,2))
par(mfrow=c(2,2))
par(mfrow=c(2,2))
for (i in c(4:16)){
  hist(cdi[,i],main=paste("histogram of",colnames(cdi)[i]))
}
```

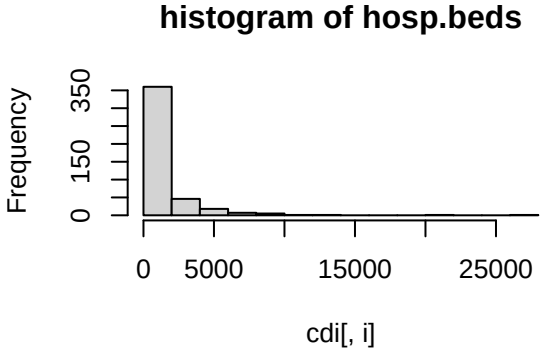

histogram of land.area
Frequency
0
200
0
5000
10000
20000
cdi[, i]

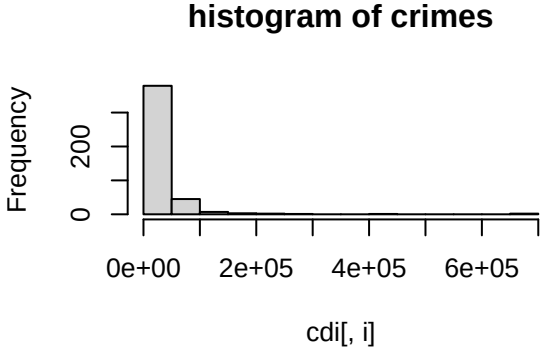

histogram of pop
Frequency
0
200
0e+00
4e+06
8e+06
cdi[, i]

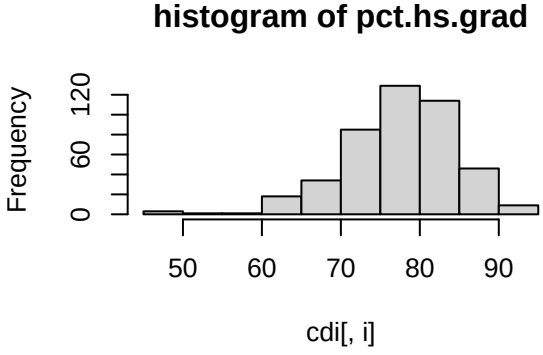

histogram of pop.18_34
Frequency
0
150
15
20
25
30
35
40
45
50
cdi[, i]

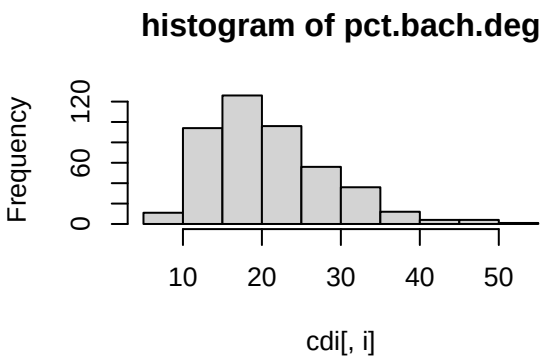

histogram of pop.65_plus
Frequency
0
150
0
5
10
15
20
25
30
35
cdi[, i]

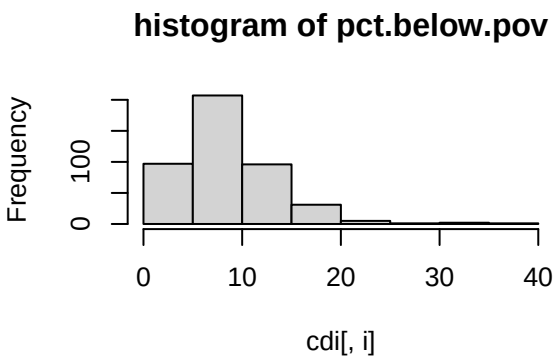

histogram of doctors
Frequency
0
200
0
5000
15000
cdi[, i]

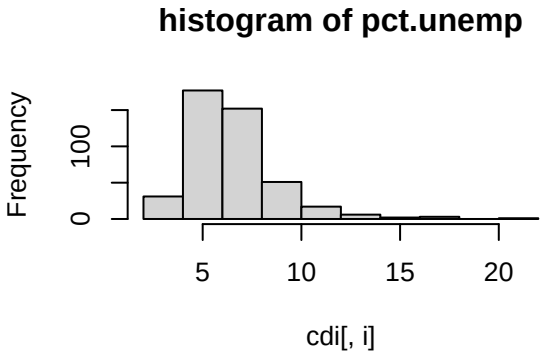

histogram of hosp.beds
Frequency
0
150
350
0
5000
15000
25000
cdi[, i]

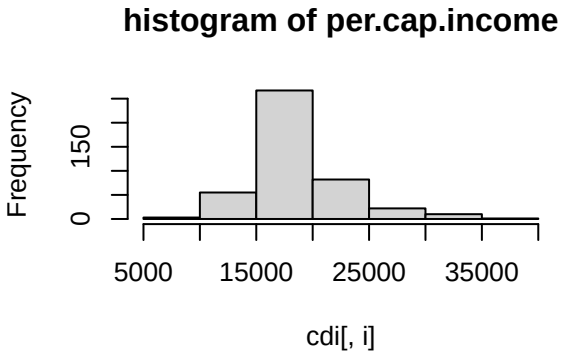

histogram of crimes
Frequency
0
200
0e+00
2e+05
4e+05
6e+05
cdi[, i]


histogram of pct.hs.grad
Frequency
0
60
120
50
60
70
80
90
cdi[, i]









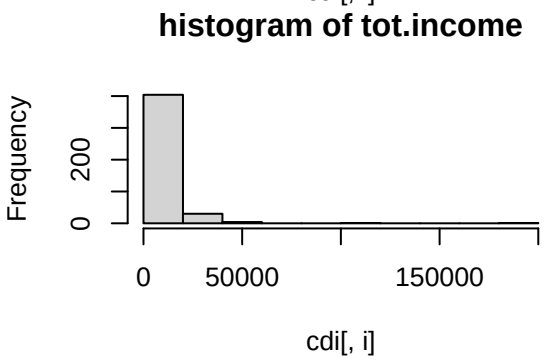


table Combing county and state

```
county.state <- with(cdi,paste(county,state))
tmp <- as.data.frame(matrix(sort(county.state),ncol=4))
names(tmp) <- paste("Counties",c("1-110","111-220","221-330","331-440"))
tmp[1:30,] %>% kbl(booktabs=T,longtable=T,caption=" ") %>% kable_classic(full_width=F)
```

Table 1:

Counties 1-110	Counties 111-220	Counties 221-330	Counties 331-440
Ada ID	Ector TX	Lycoming PA	Rockingham NH
Adams CO	El_Dorado CA	Macomb MI	Rockland NY
Aiken SC	El_Paso CO	Macon IL	Rowan NC
Alachua FL	El_Paso TX	Madison AL	Rutherford TN
Alamance NC	Elkhart IN	Madison IL	Sacramento CA
Alameda CA	Erie NY	Madison IN	Saginaw MI
Albany NY	Erie PA	Mahoning OH	Salt_Lake UT
Alexandria_City VA	Escambia FL	Manatee FL	San_Bernardino CA

Allegheny PA	Essex MA	Marathon WI	San_Diego CA
Allen IN	Essex NJ	Maricopa AZ	San_Francisco CA
Allen OH	Fairfax_County VA	Marin CA	San_Joaquin CA
Anderson SC	Fairfield CT	Marion FL	San_Luis_Obispo CA
Androscoggin ME	Fairfield OH	Marion IN	San_Mateo CA
Anne_Arundel MD	Fayette KY	Marion OR	Sangamon IL
Arapahoe CO	Fayette PA	Martin FL	Santa_Barbara CA
Arlington_County VA	Florence SC	Maui HI	Santa_Clara CA
Atlantic NJ	Forsyth NC	McHenry IL	Santa_Cruz CA
Baltimore MD	Fort_Bend TX	McLean IL	Sarasota FL
Baltimore_City MD	Franklin OH	McLennan TX	Saratoga NY
Barnstable MA	Franklin PA	Mecklenburg NC	Sarpy NE
Bay FL	Frederick MD	Medina OH	Schenectady NY
Bay MI	Fresno CA	Merced CA	Schuylkill PA
Beaver PA	Fulton GA	Mercer NJ	Sedgwick KS
Bell TX	Galveston TX	Mercer PA	Seminole FL
Benton WA	Gaston NC	Merrimack NH	Shasta CA
Bergen NJ	Genesee MI	Middlesex CT	Shawnee KS
Berks PA	Gloucester NJ	Middlesex MA	Sheboygan WI
Berkshire MA	Greene MO	Middlesex NJ	Shelby TN
Bernalillo NM	Greene OH	Midland TX	Smith TX
Berrien MI	Greenville SC	Milwaukee WI	Snohomish WA

As is shown from those histograms, there are a lot of variables that don't have normal distribution. The distributions of variable land.area, pop, doctors, hosp.beds, crimes, tot.income are all extremely right skewed and the distributions of variable pct.bach.deg and pct.below.pov are quite right skewed.

```
fit0<-lm(per.cap.income~crimes,data=cdi)
summary(fit0)
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9659.2 -2432.7  -856.8  1689.4 19124.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.834e+04  2.123e+02  86.400   <2e-16 ***
## crimes      8.193e-03  3.307e-03   2.477   0.0136 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4036 on 438 degrees of freedom
## Multiple R-squared:  0.01382,    Adjusted R-squared:  0.01156
## F-statistic: 6.136 on 1 and 438 DF,  p-value: 0.01362
```

```
fit1<-lm(per.cap.income~crimes+region,data=cdi)
summary(fit1)
```

```
##
```

```
## 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
```

```
fit2<-lm(per.cap.income~crimes+region+crimes*region,data=cdi)
cdi$per.capita.crime<-cdi$crimes/cdi$pop
fit3<-lm(per.cap.income~per.capita.crime+region,data=cdi)
fit4<-lm(per.cap.income~per.capita.crime+region+per.capita.crime*region,data=cdi)
fit5<-lm(per.cap.income~per.capita.crime,data=cdi)
summary(fit5)
```

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

```
anova(fit0,fit1)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ crimes
## Model 2: per.cap.income ~ crimes + region
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)
## 1    438 7133487504
```

```
## 2    435 6501791845  3 631695660 14.088 8.85e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit1,fit2)
```

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

```
anova(fit3,fit4)
```

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

```
anova(fit0,fit1)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ crimes
## Model 2: per.cap.income ~ crimes + region
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)
## 1    438 7133487504
## 2    435 6501791845  3 631695660 14.088 8.85e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit3,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ per.capita.crime + region
## Model 2: per.cap.income ~ per.capita.crime
##   Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
## 1    435 6609753963
## 2    438 7186843542 -3 -577089580 12.66 6.005e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There should not be any interactions in the model and the answer won't change if I use per.capita crime.

To decide whether to include a interaction in the model we can use anova. fit1 and fit2 are two models using only crimes while fit3 and fit4 are two models using per capital crimes. fit2 is fit1 added a interaction and fit3 is fit4 added a interaction.

From ANOVA test comparing fit1 and fit2, the F statistic is small and p value is quite large, which means we can't reject the null hypothesis that interaction is not needed in the model. For ANOVA test comparing fit3 and fit4, the outcomes are the same. Therefore whether I use converted crimes, the interaction should not included in the model.

Then compare the summaries if fit1 and fit3, it is clear the the adjusted R squared of fit1 is larger than that of fit3, which means the model fit3, the one that doesn't convert the crimes to per capita crimes explaines more variation in the per capita income, and therefore better.

hence, my final model is:

per.cap.income ~ crimes + region.

```
summary(fit1)
```

```
##
## 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
```

With every extra nit increase in crimes rate, for region NC, there will be a 8.915e-03 unit increase in per capita income; for region NE, the value will be overall 2.286e+03 unit more than region NC; for region S, the value will be overall 8.606e+02 unit less than region NC; for region W, the value will be overall 1.428e+02 unit less than region NC;

1.c

```
cdi<-read.table("/Users/ceciliaxia/Desktop/cdi.dat")
cdi$land.area<-log(cdi$land.area)
cdi$pop<-log(cdi$pop)
cdi$doctors<-log(cdi$doctors)
cdi$hosp.beds<-log(cdi$hosp.beds)
cdi$crimes<-log(cdi$crimes)
cdi$pct.bach.deg<-log(cdi$pct.bach.deg)
cdi$pct.below.pov<-log(cdi$pct.below.pov)
cdi$tot.income<-log(cdi$tot.income)
new<-cdi[,-c(1,2,3,5,16)]
```

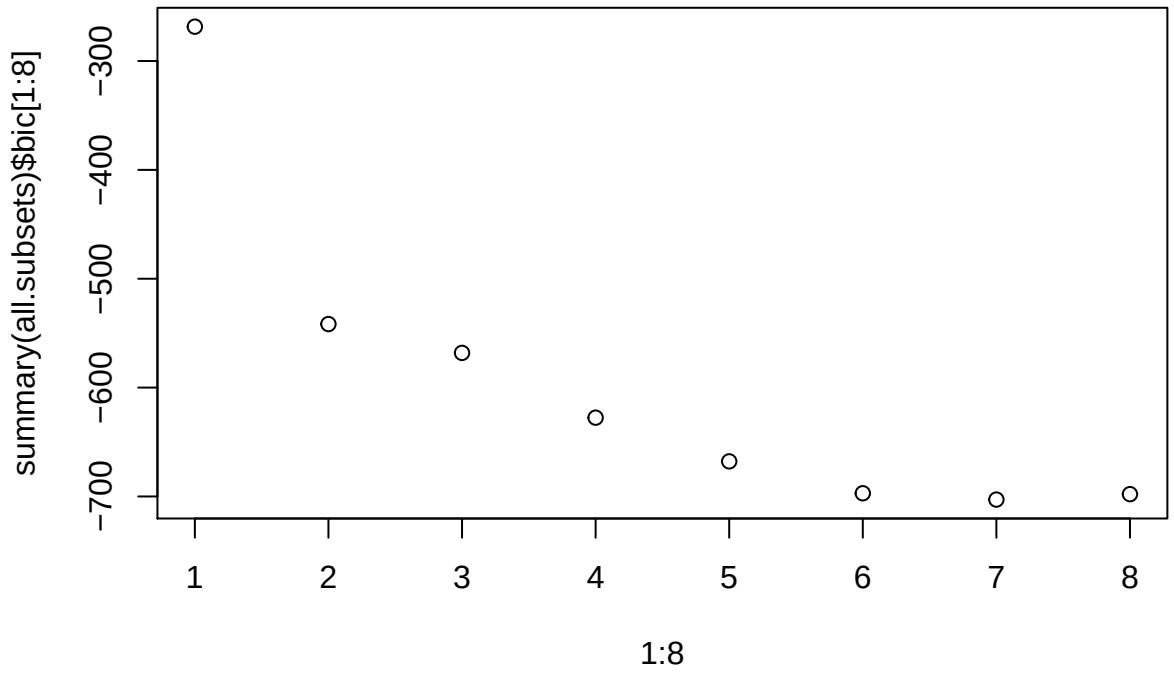
Based on the plots of Q1.a, we still need to do log transformation to the variable land.area, doctors, hosp.beds, crimes, pct.bach.deg and pct.below.pov.

```
all.subsets <- regsubsets(per.cap.income ~.-region,data=new)
coef(all.subsets,1:8)
```

```
## [[1]]
##   (Intercept) pct.below.pov
```

```
##     29237.812     -5254.348
##
## [[2]]
##   (Intercept)       doctors pct.below.pov
##     18350.851      1768.104     -5249.416
##
## [[3]]
##   (Intercept)       doctors  pct.bach.deg pct.below.pov
##     12369.806      1434.751      2211.344     -4545.989
##
## [[4]]
##   (Intercept)     pop.18_34       doctors  pct.bach.deg pct.below.pov
##    12359.3280     -225.1977     1266.7217     4217.8824    -3814.2941
##
## [[5]]
##   (Intercept)     land.area     pop.18_34       doctors  pct.bach.deg
##    17441.6279     -718.6741     -246.5155     1222.1396     4228.8712
## pct.below.pov
##    -3591.8006
##
## [[6]]
##   (Intercept)     land.area     pop.18_34       doctors  pct.bach.deg
##    13779.9659     -789.8828     -238.6530     1168.3924     5155.8231
## pct.below.pov     pct.unemp
##    -3757.5806      273.0216
##
## [[7]]
##   (Intercept)     land.area     pop.18_34       doctors   pct.hs.grad
##   18108.85533    -717.08735    -237.75549    1107.93491     -75.64321
##  pct.bach.deg pct.below.pov     pct.unemp
##    5961.13285   -4105.40240     229.39286
##
## [[8]]
##   (Intercept)     land.area     pop.18_34   pop.65_plus       doctors
##   17213.77020    -705.71566    -221.77597      28.60504    1088.38349
##   pct.hs.grad  pct.bach.deg pct.below.pov     pct.unemp
##     -75.45504    6009.66185   -4114.11839     228.61086
```

```
plot(1:8,summary(all.subsets)$bic[1:8])
```




```
summary(all.subsets)$bic[6:8]
```

```
## [1] -697.0640 -702.8682 -697.8823
```

Then use all subsets to do the variable selection, I chose the model with the smallest BIC, which regress per.capita.income on transformed variables: land.area, pop.18_34, doctors, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp

```
fit<-lm(per.cap.income ~.-region,data=new)
summary(fit)
```

```
##
## Call:
## lm(formula = per.cap.income ~ . - region, data = new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5901.3  -949.6  -206.1   795.0 10099.5
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17731.228   2282.795   7.767 5.96e-14 ***
## land.area      -698.836    103.175  -6.773 4.16e-11 ***
## pop.18_34      -222.780     29.043  -7.671 1.16e-13 ***
## pop.65_plus      23.958     29.043   0.825 0.409885
## doctors        1176.308    267.686   4.394 1.40e-05 ***
## hosp.beds         7.016    255.795   0.027 0.978131
## crimes         -106.524    167.604  -0.636 0.525395
## pct.hs.grad     -75.634     22.052  -3.430 0.000663 ***
## pct.bach.deg   5976.583    508.289  11.758  < 2e-16 ***
## pct.below.pov -4091.242    243.949 -16.771  < 2e-16 ***
## pct.unemp       230.633     47.005   4.907 1.32e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1745 on 429 degrees of freedom
## Multiple R-squared:  0.8194, Adjusted R-squared:  0.8152
## F-statistic: 194.7 on 10 and 429 DF,  p-value: < 2.2e-16
```

```
fit_step<-stepAIC(fit,direction="both",k=2)
```

```
## Start:  AIC=6579.61
## per.cap.income ~ (land.area + pop.18_34 + pop.65_plus + doctors +
##     hosp.beds + crimes + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp + region) - region
##
##                 Df Sum of Sq        RSS    AIC
## - hosp.beds      1      2291 1306259012 6577.6
## - crimes         1   1229984 1307486705 6578.0
## - pop.65_plus    1   2071940 1308328661 6578.3
## <none>                       1306256722 6579.6
## - pct.hs.grad    1  35818081 1342074802 6589.5
## - doctors        1  58798129 1365054850 6597.0
## - pct.unemp      1  73301893 1379558615 6601.6
## - land.area      1 139693176 1445949898 6622.3
## - pop.18_34      1 179160534 1485417255 6634.2
## - pct.bach.deg   1 420974974 1727231696 6700.5
## - pct.below.pov  1 856415001 2162671723 6799.4
##
## Step:  AIC=6577.61
## per.cap.income ~ land.area + pop.18_34 + pop.65_plus + doctors +
##     crimes + pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
##
##                 Df Sum of Sq        RSS    AIC
## - crimes         1   1236733 1307495746 6576.0
## - pop.65_plus    1   2185344 1308444356 6576.3
## <none>                       1306259012 6577.6
## + hosp.beds      1      2291 1306256722 6579.6
## - pct.hs.grad    1  35865100 1342124112 6587.5
## - pct.unemp      1  74062335 1380321348 6599.9
## - doctors        1 140481945 1446740957 6620.6
## - land.area      1 141264745 1447523757 6620.8
## - pop.18_34      1 179307992 1485567004 6632.2
## - pct.bach.deg   1 464358993 1770618005 6709.4
## - pct.below.pov  1 928563708 2234822720 6811.9
##
## Step:  AIC=6576.03
## per.cap.income ~ land.area + pop.18_34 + pop.65_plus + doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
##
##                 Df Sum of Sq        RSS    AIC
## - pop.65_plus    1   3275353 1310771098 6575.1
## <none>                       1307495746 6576.0
## + crimes         1   1236733 1306259012 6577.6
## + hosp.beds      1      9040 1307486705 6578.0
## - pct.hs.grad    1  35725736 1343221482 6585.9
## - pct.unemp      1  73149309 1380645054 6598.0
## - land.area      1 145417926 1452913672 6620.4
## - pop.18_34      1 178234754 1485730500 6630.3
```

```
## - doctors        1 412187735 1719683480 6694.6
## - pct.bach.deg   1 477203458 1784699203 6710.9
## - pct.below.pov  1 966421060 2273916806 6817.5
##
## Step:  AIC=6575.13
## per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
##
##                 Df Sum of Sq        RSS    AIC
## <none>                       1310771098 6575.1
## + pop.65_plus    1   3275353 1307495746 6576.0
## + crimes         1   2326742 1308444356 6576.3
## + hosp.beds      1     47215 1310723884 6577.1
## - pct.hs.grad    1  35906582 1346677680 6585.0
## - pct.unemp      1  73669858 1384440956 6597.2
## - land.area      1 151893139 1462664237 6621.4
## - pop.18_34      1 285501384 1596272482 6659.8
## - doctors        1 445209423 1755980522 6701.8
## - pct.bach.deg   1 474031134 1784802232 6709.0
## - pct.below.pov  1 963606671 2274377770 6815.6
```

```
coef(fit_step)
```

```
##   (Intercept)     land.area     pop.18_34       doctors   pct.hs.grad
##   18108.85533    -717.08735    -237.75549    1107.93491     -75.64321
##  pct.bach.deg pct.below.pov     pct.unemp
##    5961.13285   -4105.40240     229.39286
```

Then use stepwise selection with AIC to do the variable selection, I chose the model with the smallest BIC, which regress per.capita.income on transformed variables: land.area, pop.18_34, pop.65_plus, doctors, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp.

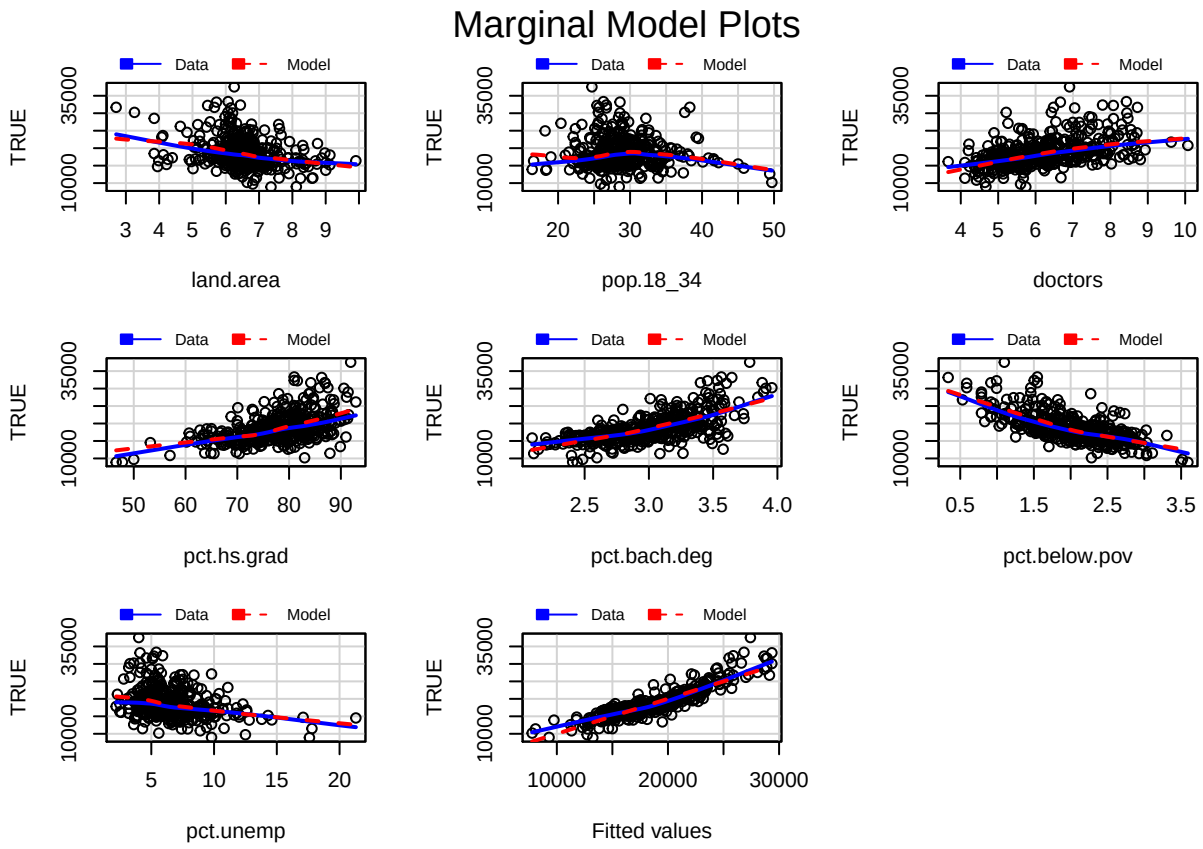
This stepwise model is the same as the all subsets model. Now we fit this model and draw add variable plots and marginal plots for this model.

```
fit_final<-lm(per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad + pct.bach.deg + pct.below
summary(fit_final)
```

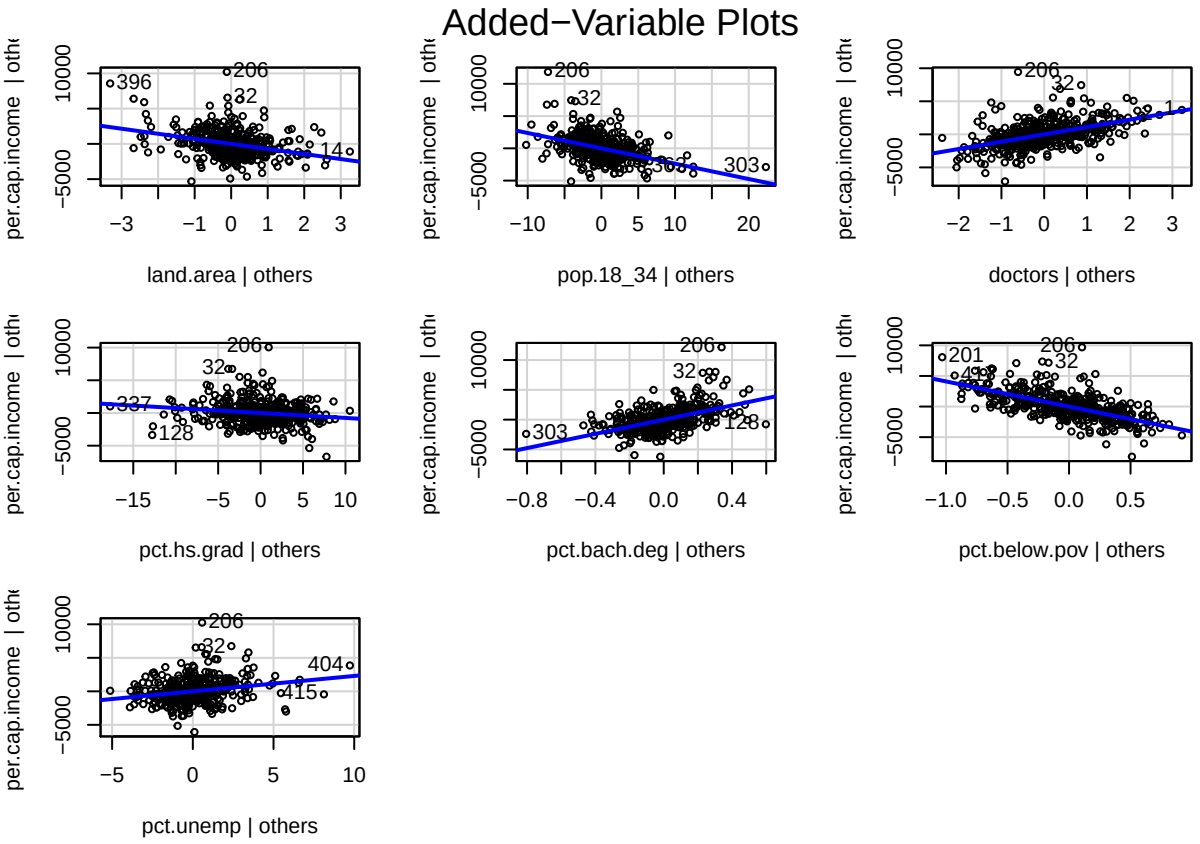
```
##
## Call:
## lm(formula = per.cap.income ~ land.area + pop.18_34 + doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp, data = new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6100.0  -949.7  -213.7   757.6 10121.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18108.86    1867.69   9.696  < 2e-16 ***
## land.area      -717.09     101.35  -7.075 6.05e-12 ***
## pop.18_34      -237.76      24.51  -9.700  < 2e-16 ***
## doctors        1107.93      91.46  12.113  < 2e-16 ***
## pct.hs.grad     -75.64      21.99  -3.440 0.000638 ***
## pct.bach.deg   5961.13     476.92  12.499  < 2e-16 ***
## pct.below.pov -4105.40     230.37 -17.821  < 2e-16 ***
```

```
## pct.unemp       229.39      46.55   4.927 1.19e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1742 on 432 degrees of freedom
## Multiple R-squared:  0.8188, Adjusted R-squared:  0.8159
## F-statistic: 278.9 on 7 and 432 DF,  p-value: < 2.2e-16
```

```
mmps(fit_final)
```




```
avPlots(fit_final)
```



All marginal plots show that the non-parametric lines agree with the regression lines and it indicates that the response variable and the predictors don't need to be transformed.Added-Variable plots show that the model won't need any transformation to the variables included in the model. Overall, the transformation is alright.

```
m1<-lm(per.cap.income ~land.area+pop.18_34+doctors+pct.hs.grad+pct.bach.deg+ pct.below.pov +pct.unemp+la
m2<-lm(per.cap.income ~land.area+pop.18_34+doctors+pct.hs.grad+pct.bach.deg+ pct.below.pov +pct.unemp+pc
m3<-lm(per.cap.income ~land.area+pop.18_34+doctors+pct.hs.grad+pct.bach.deg+ pct.below.pov +pct.unemp+do
m4<-lm(per.cap.income ~land.area+pop.18_34+doctors+pct.hs.grad+pct.bach.deg+ pct.below.pov +pct.unemp+pc
m5<-lm(per.cap.income ~land.area+pop.18_34+doctors+pct.hs.grad+pct.bach.deg+ pct.below.pov +pct.unemp+pc
m6<-lm(per.cap.income ~land.area+pop.18_34+doctors+pct.hs.grad+pct.bach.deg+ pct.below.pov +pct.unemp+pc
m7<-lm(per.cap.income ~land.area+pop.18_34+doctors+pct.hs.grad+pct.bach.deg+ pct.below.pov +pct.unemp+pc
anova(fit_final,m1)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + land.area * region
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    432 1310771098
## 2    426 1281663440  6  29107658 1.6125  0.142
```

```
anova(fit_final,m2)
```

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

```
## Model 2: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + pop.18_34 * region
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1    432 1310771098
## 2    426 1268258641  6  42512457 2.3799 0.02842 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_final,m3)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + doctors * region
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1    432 1310771098
## 2    426 1269571129  6  41199969 2.3041 0.03357 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_final,m4)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + pct.hs.grad *
##     region
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    432 1310771098
## 2    426 1234534538  6  76236560 4.3845 0.0002591 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_final,m5)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + pct.bach.deg *
##     region
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    432 1310771098
## 2    426 1234743417  6  76027681 4.3717 0.0002673 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_final,m6)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
```

```
##     pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + pct.below.pov *
##     region
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)
## 1    432 1310771098
## 2    426 1257169536  6  53601562 3.0272 0.006566 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_final,m7)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + pct.unemp * region
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    432 1310771098
## 2    426 1227697409  6  83073690 4.8043 9.305e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

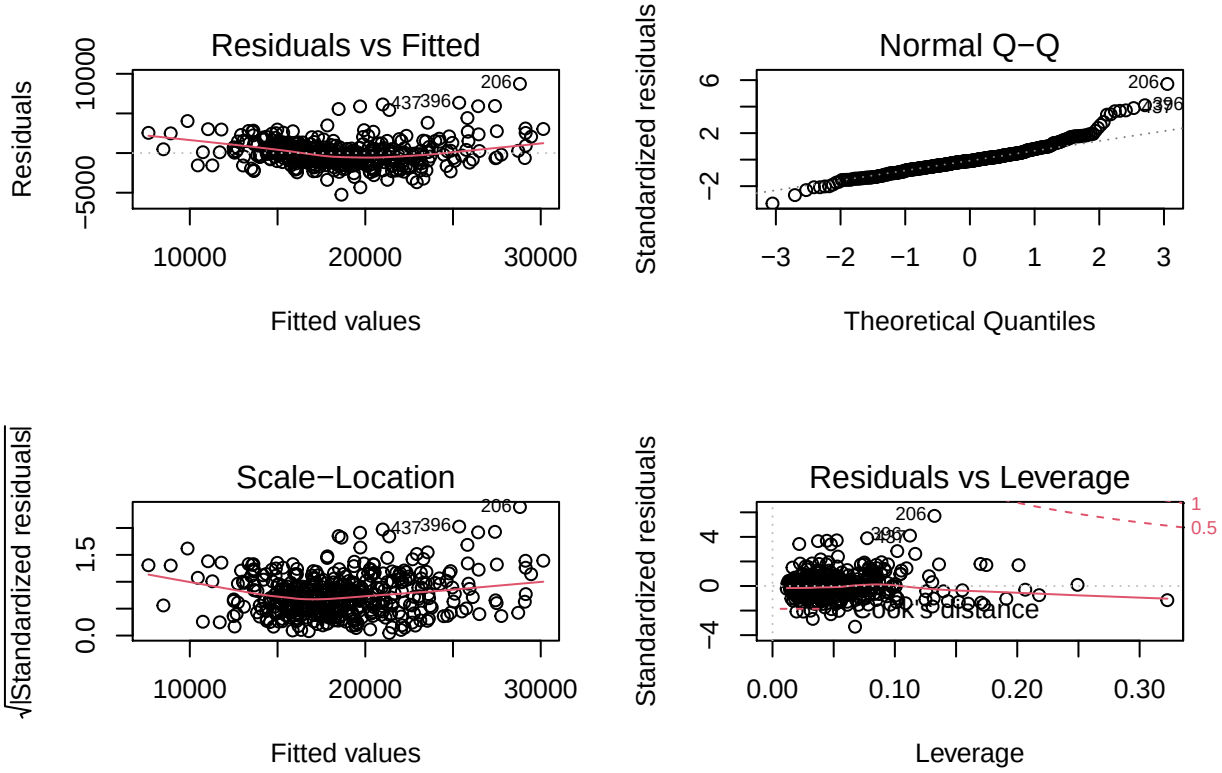
Then I test the interaction between the categorical variable region and all the other continuous variables pair by pair and select every interaction whose coefficient is statistically significant (here I mean the one whose p vale is with more than two stars in the summary)

```
m_final<-lm(per.cap.income ~land.area+pop.18_34+doctors+pct.hs.grad+pct.bach.deg+ pct.below.pov +pct.un
summary(m_final)
```

```
##
## Call:
## lm(formula = per.cap.income ~ land.area + pop.18_34 + doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     pct.hs.grad * region + pct.unemp * region + pct.bach.deg *
##     region + pct.below.pov * region, data = new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5248.2  -926.3  -172.6   689.1  8726.2
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             15099.06    4911.52   3.074  0.00225 **
## land.area                -729.55     116.15  -6.281 8.46e-10 ***
## pop.18_34                -260.76      24.14 -10.804  < 2e-16 ***
## doctors                   978.14      92.56  10.568  < 2e-16 ***
## pct.hs.grad               -21.04      66.76  -0.315  0.75281
## pct.bach.deg             5156.02     977.20   5.276 2.12e-07 ***
## pct.below.pov           -3114.17     503.01  -6.191 1.43e-09 ***
## pct.unemp                 345.40     106.93   3.230  0.00133 **
## regionNE                 2893.77    6256.09   0.463  0.64393
## regionS                  2069.72    5608.57   0.369  0.71229
## regionW                 31802.39    7844.20   4.054 6.00e-05 ***
## pct.hs.grad:regionNE      -87.93      81.24  -1.082  0.27977
```

```
## pct.hs.grad:regionS      -27.70      74.66  -0.371  0.71079
## pct.hs.grad:regionW     -380.61      88.48  -4.302 2.11e-05 ***
## pct.unemp:regionNE      -152.45     160.39  -0.951  0.34239
## pct.unemp:regionS       -363.78     138.98  -2.617  0.00918 **
## pct.unemp:regionW       -366.61     146.54  -2.502  0.01274 *
## pct.bach.deg:regionNE   2327.38    1177.02   1.977  0.04866 *
## pct.bach.deg:regionS     868.05    1080.38   0.803  0.42216
## pct.bach.deg:regionW    2940.05    1249.14   2.354  0.01905 *
## pct.below.pov:regionNE -1050.27     688.23  -1.526  0.12775
## pct.below.pov:regionS   -280.66     595.15  -0.472  0.63747
## pct.below.pov:regionW  -3696.33     886.37  -4.170 3.70e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1641 on 417 degrees of freedom
## Multiple R-squared:  0.8447, Adjusted R-squared:  0.8366
## F-statistic: 103.1 on 22 and 417 DF,  p-value: < 2.2e-16
```

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



The final model is :

per.cap.income~log(land.area)+pop.18_34+log(doctors)+pct.hs.grad+log(pct.bach.deg)+ log(pct.below.pov) +pct.unemp+pct.hs.grad* region+pct.unemp* region+log(pct.bach.deg)* region+log(pct.below.pov)* region

The tradeoff is made in my interactions between the region and the other continous variables in the model, I didn't chose all the interactions that are marked statistically significant under the significance level 0.05. Instead I only chose the one that are more significant (which means with more than two stars). If I only use 0.05 criterion, there will many interactions and may case overfitting so I just exclude that one that is only marked with one star.

For the diagnostics plot of our final model, The Residuals vs Fitted plot shows that there is not nonlinear pattern in the model but there are some outliers detected; the normal Q-Q plot shows that the residuals are normally distributed and the assumption that errors are normally distributed holds, but there are some points that deviated from the diagnal line a lot; the scale-location plot shows that the residuals are spread equally along the ranges of predictors and the residuals have constant variance; the residuals vs leverage plot shows that there are no influential points in the model. Overall, I think it is a good fit.