Social science and economic factors that can affect per-capita income in each US county

Yuqing Xu
yuqingxu@andrew.cmu.edu

Abstract

The project focuses on the relationship between income per capita and other variables associated with the economic, health, and social well-being values in each county in the US. The file cdi.dat is taken from Kutneret al. (2005) and provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. The project uses stepwise variable selection in both directions, and all subsets regression to build potentially best fitting models; diagnostic plots, AIC and BIC values, and some interpretation under social and economic context are used to select the best one as the final fitting model. Final selected variables are pct.unemp:region, pct.hs.grad:region, pct.below.pov:region, region, pop.18_34, pct.unemp, pct.below.pov, pct.bach.deg, log(doctors), log(land.area). For further exploration on this topic, the effect of region/state variable with other variables as interaction terms should be addressed, and some missing data should be collected and added on if possible to eliminate bias.

Introduction

Income is a factor that can reflect people's living quality, and thus the average income can be a factor that shows how well people live in a certain area, which is important in many social problems. The project mainly focuses on how income per capita was related to other variables associated with the county's economic, health, and social well-being like population, crimes, and education. And this research problem may give some perspectives on how to improve in any aspect to improve the income overall. The questions will be addressed related to the topic in this project are:

1. Is there any relation between variables in datasets?
2. How per-capita income was related to crime number, and does different regions of the country matter for this relationship? Is it better or more reasonable to use (number of crimes)/(population)?
3. What is the best model predicting per-capita income from the other variables, which best reflects the social science and the meaning of the variables?
4. Should we be worried about either the missing states or the missing counties?

Data

The file cdi.dat is taken from Kutneret al. (2005). The data 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 variables we are going to use to build models are:

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)

We would not consider id and county as variables, because there is no duplicated value in these two variables, and they are more like identifiers for each row of dataset.
For each numerical variable, we can see the summary in the table below:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
land.area	15.0	451.250	656.50	1.041411e+03	946.750	20062.0
pop	100043.0	139027.250	217280.50	3.930109e+05	436064.500	8863164.0
pop.18_34	16.4	26.200	28.10	2.856841e+01	30.025	49.7
pop.65_plus	3.0	9.875	11.75	1.216977e+01	13.625	33.8
doctors	39.0	182.750	401.00	9.879977e+02	1036.000	23677.0
hosp.beds	92.0	390.750	755.00	1.458627e+03	1575.750	27700.0
crimes	563.0	6219.500	11820.50	2.711162e+04	26279.500	688936.0
pct.hs.grad	46.6	73.875	77.70	7.756068e+01	82.400	92.9
pct.bach.deg	8.1	15.275	19.70	2.108114e+01	25.325	52.3
pct.below.pov	1.4	5.300	7.90	8.720682e+00	10.900	36.3
pct.unemp	2.2	5.100	6.20	6.596591e+00	7.500	21.3
per.cap.income	8899.0	16118.250	17759.00	1.856148e+04	20270.000	37541.0
tot.income	1141.0	2311.000	3857.00	7.869273e+03	8654.250	184230.0

Table 2: numerical variables summary table

For each categorical variable (state and region), we can see the summary in the tables below:

Table 3: state summary

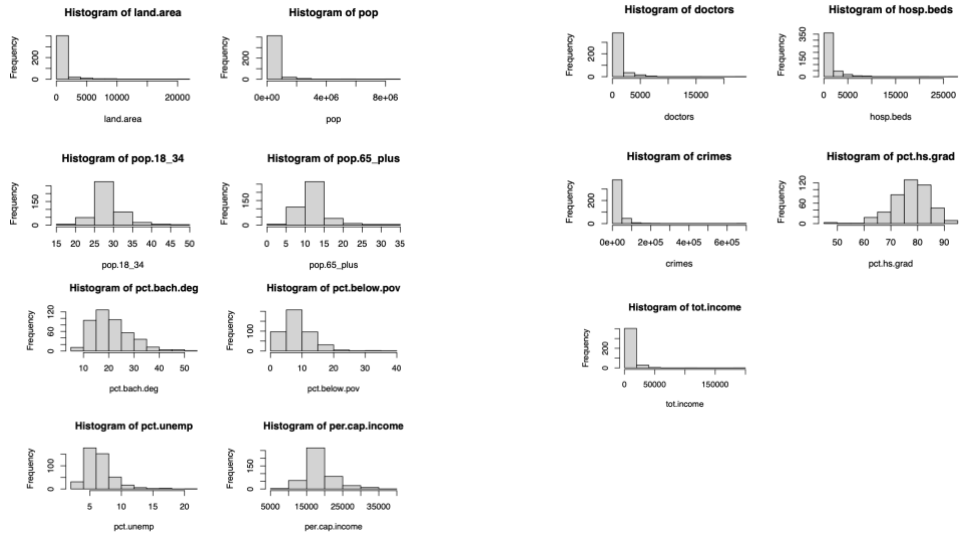
	frequency		
AL	7	NY	22
AR	2	OH	24
AZ	5	OK	4
CA	34	OR	6
CO	9	PA	29
CT	8	RI	3
DC	1	SC	11
DE	2	SD	1
FL	29	TN	8
GA	9	TX	28
HI	3	UT	4
ID	1	VA	9
IL	17	VT	1
IN	14	WA	10
KS	4	WI	11
KY	3	WV	1
LA	9		
MA	11		
MD	10		
ME	5		
MI	18		
MN	7		
MO	8		
MS	3		
MT	1		
NC	18		
ND	1		
NE	3		
NH	4		
NJ	18		
NM	2		
NV	2		

Table 4: region summary

	frequency
NC	108
NE	103
S	152
W	77

Then we can make histogram of each numeric variables to see the distribution:

Plot 1: Distribution of variables



We can see that land.area, pop, doctors, hosp.beds, crimes, and tot.income are skewed to right, thus transformation may need.

Method

At the beginning, we can apply log transformation on the heavily right-skewed variables: land.area, pop, doctors, hosp.beds, crimes, and tot.income. Tranformation is applied to put the tails in and make the distribution of these variables more normal.

Firstly, we want to address the question that if we ignore all other variables, whether per-capita income should be related to crime number, and also that this relationship may be different in different regions of the country, which statistically means that if we will need to add interaction between region and crime number.
Region variable is added as additive term or multiplication term to crimes variable. Then we can apply an ANOVA test to see which one fits the data best to decide if we need the interaction term. After noticing that per-capita income is total income/total population, we want to know if per-capita crime works better than total crime number in terms of using per-capita variable to predict per-capita variable. Thus, we separately add region and interaction term to the per-capita crimes as above. Then ANOVA test is applied to these models. The two best models selected from two sets of models are compared by diagnostic plots, AIC, BIC, and summary to choose the final one, which will indicate which one of crime number and per capita crime number will be used later.

For the further selection of variables and models, total income and population are removed because per-capita income = total income/total population. Region and state are ignored at the beginning because they are hard to address as categorical values. For variable selection, all subsets method, and stepwise selection are used. Then the two models are compared by diagnostic plots, AIC, BIC values. Then it is compared with model with region term multiplicated.

Results

1. Is there any relation between variables in datasets?

Appendix A: page 8
From the correlation plot we can see that tot.income and pop are highly correlated, and they both are reasonably highly correlated with crimes, hosp.beds and doctors, and these three are also strongly correlated with each other.

per.cap.income isn't really highly correlated with anything, but it has some positive correlation with pct.hs.grad, pct.bach.deg and some negative correlation with pct.below.pov, pct.unemp. And pct.hs.grad, pct.bach.deg, pct.below.pov, and pct.unemp. are moderately correlated with each other.

These correlations are expected. total income = per-capita income * population, so it is reasonable that they are correlated, and huge population in some way means that more people are criming, and also more people being doctors, and this county will need more beds in the hospital for this large amount of people. At the same time, more crimes means that more people are hurt, and thus more doctors and hospital beds needed. Also, per-capita income has positive correlation with pct.hs.grad, pct.bach.deg because people with high school or college education are more likely to get high income than those who do not complete at least 12 years of school. And per-capita income has negative correlation with pct.below.pov, and pct.unemp because the increase of people with income below poverty level and people unemployed means that they are getting really low income.

2. How per-capita income was related to crime number, and does different regions of the country matter for this relationship? Is it better or more reasonable to use (number of crimes)/(population)?

Appendix B: page 10

For the model comparison to check the significance of crime number on per-capita income, and the exploration of how region can affect the relationship between per-capita income and crime number, by the ANOVA test, we can know that we cannot reject that the model predicting per-capita income by adding region and crimes is enough comparing to model with variable region times crimes, which means that dummy variable region produces only additive changes in log(per-capita income). For the set of models below, model 2 is selected.

```
## Analysis of Variance Table
##
## Model 1: log(per.cap.income) ~ log(crimes)
## Model 2: log(per.cap.income) ~ log(crimes) + region
## Model 3: log(per.cap.income) ~ log(crimes) + region + (log(crimes):region)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 17.271
## 2    435 14.949  3   2.32194 22.4823 1.523e-13 ***
## 3    432 14.872  3   0.07678  0.7434    0.5266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

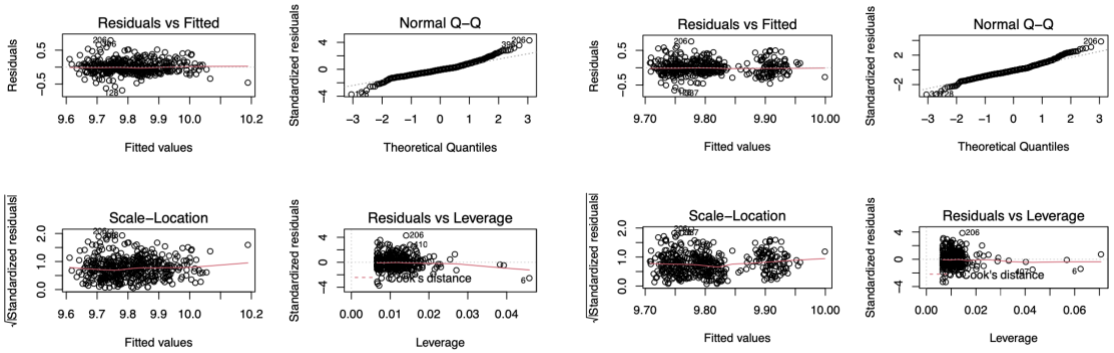
Appendix C: page 15

Then, when per-capita crime is considered in the model, we get the same result that interaction between per-capita crime and region is also unnecessary. For the set of models below, model 2 is selected.

```
## Analysis of Variance Table
##
## Model 1: log(per.cap.income) ~ log(per.capita.crime)
## Model 2: log(per.cap.income) ~ log(per.capita.crime) + region
## Model 3: log(per.cap.income) ~ log(per.capita.crime) + region + (log(per.capita.crime):region)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 18.697
## 2    435 16.952  3   1.74465 14.8407 3.263e-09 ***
## 3    432 16.928  3   0.02408  0.2048     0.893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Appendix D: page 15

Then we come to the choice between total crime number (model 1) and per-capita crime (model 2). For the diagnostic plots, the two sets of plots look quite similar. But model 1 has smaller AIC and BIC values. Overall, I will choose model 1, which means that I would like to use total crimes later.



3. What is the best model predicting per-capita income from the other variables, which best reflects the social science and the meaning of the variables?

Appendix E

Then, we are going to do variable selection on all variables.

Part 1: page 19

Stepwise selection in both directions is used in, which iteratively adding and removing predictors in the predictive model to find the subset of variables in the data set resulting in the model that has lowest prediction error. And at this step we get the model:

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + log(doctors) +
##     pct.bach.deg + pct.below.pov + pct.unemp + pct.hs.grad +
##     pop.18_34 + pop.65_plus, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35756 -0.04551 -0.00543  0.04844  0.27399
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.3159666  0.1025858 100.559  < 2e-16 ***
## log(land.area) -0.0364935  0.0047728  -7.646 1.36e-13 ***
## log(doctors)    0.0626053  0.0041029  15.259  < 2e-16 ***
## pct.bach.deg    0.0152149  0.0009242  16.462  < 2e-16 ***
## pct.below.pov  -0.0246144  0.0012631 -19.488  < 2e-16 ***
## pct.unemp       0.0107688  0.0021696   4.963 9.99e-07 ***
## pct.hs.grad    -0.0046579  0.0010843  -4.296 2.15e-05 ***
## pop.18_34      -0.0153488  0.0012988 -11.818  < 2e-16 ***
## pop.65_plus    -0.0027664  0.0012978  -2.132   0.0336 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08167 on 431 degrees of freedom
## Multiple R-squared:  0.8468, Adjusted R-squared:  0.8439
## F-statistic: 297.7 on 8 and 431 DF,  p-value: < 2.2e-16
```

Part 2: page 20

All subsets method is used, which tests all possible subsets of the set of potential independent variables. For the model that maximizes squared value and minimizes cp and BIC values, we get model with 8 variables, fit_8 :

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + log(doctors) +
##     pct.bach.deg + pct.below.pov + pct.unemp + pct.hs.grad +
##     pop.18_34 + pop.65_plus, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35756 -0.04551 -0.00543  0.04844  0.27399
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.3159666  0.1025858 100.559  < 2e-16 ***
## log(land.area) -0.0364935  0.0047728  -7.646 1.36e-13 ***
## log(doctors)    0.0626053  0.0041029  15.259  < 2e-16 ***
## pct.bach.deg    0.0152149  0.0009242  16.462  < 2e-16 ***
## pct.below.pov  -0.0246144  0.0012631 -19.488  < 2e-16 ***
## pct.unemp       0.0107688  0.0021696   4.963 9.99e-07 ***
## pct.hs.grad    -0.0046579  0.0010843  -4.296 2.15e-05 ***
## pop.18_34      -0.0153488  0.0012988 -11.818  < 2e-16 ***
## pop.65_plus    -0.0027664  0.0012978  -2.132   0.0336 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08167 on 431 degrees of freedom
## Multiple R-squared:  0.8468, Adjusted R-squared:  0.8439
## F-statistic: 297.7 on 8 and 431 DF,  p-value: < 2.2e-16
```

We get same models from this step.
Then, region term is going to be considered.

Part 3: page 21
From the summary table below, we can see that, after removing the insignificant variables, pct.unemp:region, pct.hs.grad:region, pct.below.pov:region, region, pop.18_34, pct.unemp, pct.below.pov, pct.bach.deg, log(doctors), log(land.area) are the variables left significantly for the model with interaction of region.

```
...
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.239497 -0.042518 -0.002899  0.038705  0.315955
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                10.1550994  0.3077758  32.995  < 2e-16 ***
## log(land.area)             -0.0355230  0.0155258  -2.288 0.022654 *
## log(doctors)                0.0548293  0.0094485   5.803 1.32e-08 ***
## pct.bach.deg                0.0140191  0.0029305   4.784 2.41e-06 ***
## pct.below.pov              -0.0233702  0.0038627  -6.050 3.30e-09 ***
## pct.unemp                   0.0176067  0.0051819   3.398 0.000747 ***
## pct.hs.grad                -0.0026649  0.0034861  -0.764 0.445055
## pop.18_34                  -0.0150740  0.0028317  -5.323 1.69e-07 ***
## pop.65_plus                -0.0012483  0.0050165  -0.249 0.803614
## regionNE                    0.4813749  0.3863061   1.246 0.213451
## regionS                    -0.0552517  0.3396107  -0.163 0.870843
## regionW                     1.3969067  0.4575796   3.053 0.002417 **
## log(land.area):regionNE    -0.0050730  0.0204207  -0.248 0.803932
## log(land.area):regionS     -0.0058664  0.0177783  -0.330 0.741589
## log(land.area):regionW      0.0136894  0.0185229   0.739 0.460306
## log(doctors):regionNE       0.0001267  0.0135190   0.009 0.992526
## log(doctors):regionS        0.0042557  0.0116550   0.365 0.715198
## log(doctors):regionW       -0.0046667  0.0132947  -0.351 0.725759
## pct.bach.deg:regionNE       0.0060237  0.0040533   1.486 0.138025
## pct.bach.deg:regionS       -0.0015550  0.0032102  -0.484 0.628384
## pct.bach.deg:regionW        0.0069577  0.0036552   1.903 0.057687 .
## pct.below.pov:regionNE     -0.0009949  0.0052677  -0.189 0.850294
## pct.below.pov:regionS       0.0068718  0.0042992   1.598 0.110736
## pct.below.pov:regionW      -0.0167523  0.0055989  -2.992 0.002941 **
## pct.unemp:regionNE         -0.0063048  0.0075950  -0.830 0.406962
## pct.unemp:regionS          -0.0243492  0.0068439  -3.558 0.000418 ***
## pct.unemp:regionW          -0.0192087  0.0070270  -2.734 0.006541 **
## pct.hs.grad:regionNE       -0.0033331  0.0044706  -0.746 0.456373
## pct.hs.grad:regionS         0.0023152  0.0038518   0.601 0.548134
## pct.hs.grad:regionW        -0.0185423  0.0046646  -3.975 8.33e-05 ***
## pop.18_34:regionNE         -0.0060991  0.0042036  -1.451 0.147582
## pop.18_34:regionS          -0.0008273  0.0034566  -0.239 0.810970
## pop.18_34:regionW           0.0030516  0.0048005   0.636 0.525342
## pop.65_plus:regionNE       -0.0076628  0.0063347  -1.210 0.227119
## pop.65_plus:regionS         0.0009166  0.0052822   0.174 0.862326
## pop.65_plus:regionW         0.0037008  0.0064632   0.573 0.567239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07573 on 404 degrees of freedom
## Multiple R-squared:  0.8765, Adjusted R-squared:  0.8658
## F-statistic: 81.92 on 35 and 404 DF,  p-value: < 2.2e-16
```

Finally we get the new model log(per.cap.income) ~ pct.unemp:region+ pct.hs.grad:region+ pct.below.pov:region+ region+ pop.18_34+ pct.unemp+ pct.below.pov+pct.bach.deg+ log(doctors)+log(land.area)

Part 4: page 24

We will compare the fit_8 model with this new model with region.
ANOVA test:

```
Analysis of Variance Table

Model 1: log(per.cap.income) ~ log(land.area) + log(doctors) + pct.bach.deg +
    pct.below.pov + pct.unemp + pct.hs.grad + pop.18_34 + pop.65_plus
Model 2: log(per.cap.income) ~ pct.unemp:region + pct.hs.grad:region +
    pct.below.pov:region + region + pop.18_34 + pct.unemp + pct.below.pov +
    pct.bach.deg + log(doctors) + log(land.area)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    431 2.8748
2    420 2.4853 11   0.38947 5.9835 4.306e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By the result of test, we can say that the null hypothesis that model without interaction with region is good enough can be rejected. Thus, we think choosing the new model is better.

```
##                df       BIC
## fit_8          10 -904.0206
## fit_region_new 21 -901.1215
```

```
##                df       AIC
## fit_8          10 -944.8883
## fit_region_new 21 -986.9437
```

From BIC and AIC value of the two model, we can see that fit_8 has smaller BIC value but new model has smaller AIC value.
From the two sets of diagnostic plots, we can find it is hard to choose based on those because they are similar.
From the ANOVA test, and value comparisons above, I will choose new model with region as the final model selected. It has lower AIC value, though higher BIC value, but BIC value tends to choose simpler model. Also, from ANOVA test, we can see that the interaction term can make the performance of model better than before to an extent that we cannot accept that the models are quite similar by the p-value far less than 0.05.

For the numeric variables, the positive correlation between y variable and doctors, pct.bach.deg, and the negative correlation between y variable and pct.below.pov can both fit the correlation plot and our expectation at the very beginning, which makes this model reasonable in both statiscal and social aspects.

```
##
## Call:
## lm(formula = log(per.cap.income) ~ pct.unemp:region + pct.hs.grad:region +
##     pct.below.pov:region + region + pop.18_34 + pct.unemp + pct.below.pov +
##     pct.bach.deg + log(doctors) + log(land.area), data = cdi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.294186 -0.043597 -0.001583  0.037667  0.311609
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.2421239  0.2176557  47.057  < 2e-16 ***
## regionNE               -0.0520070  0.2707173  -0.192 0.847750
## regionS                -0.0389718  0.2383516  -0.164 0.870199
## regionW                 1.3910484  0.3408962   4.081 5.38e-05 ***
## pop.18_34              -0.0149347  0.0010897 -13.705  < 2e-16 ***
## pct.unemp               0.0197400  0.0046254   4.268 2.44e-05 ***
## pct.below.pov          -0.0252029  0.0032612  -7.728 8.12e-14 ***
## pct.bach.deg            0.0156310  0.0009715  16.090  < 2e-16 ***
## log(doctors)            0.0572284  0.0040082  14.278  < 2e-16 ***
## log(land.area)         -0.0381738  0.0053996  -7.070 6.51e-12 ***
## pct.unemp:regionNE     -0.0129841  0.0070423  -1.844 0.065929 .
## pct.unemp:regionS      -0.0231138  0.0061365  -3.767 0.000189 ***
## pct.unemp:regionW      -0.0217357  0.0065225  -3.332 0.000937 ***
## regionNC:pct.hs.grad   -0.0043532  0.0024515  -1.776 0.076501 .
## regionNE:pct.hs.grad   -0.0025848  0.0020257  -1.276 0.202657
## regionS:pct.hs.grad    -0.0032007  0.0014122  -2.266 0.023936 *
## regionW:pct.hs.grad    -0.0185005  0.0027800  -6.655 8.88e-11 ***
## regionNE:pct.below.pov -0.0015170  0.0046143  -0.329 0.742493
## regionS:pct.below.pov   0.0070185  0.0035199   1.994 0.046808 *
## regionW:pct.below.pov  -0.0137920  0.0051811  -2.662 0.008066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07692 on 420 degrees of freedom
## Multiple R-squared:  0.8675, Adjusted R-squared:  0.8615
## F-statistic: 144.8 on 19 and 420 DF,  p-value: < 2.2e-16
```

Discussion

1. Is there any relation between variables in datasets?

For the correlation between the variables, from the correlation plots we can see that tot.income and pop are highly correlated, and they both are reasonably highly correlated with crimes, hosp.beds and doctors, and these three are also strongly correlated with each other.

per.cap.income isn't really highly correlated with anything, but it has some positive correlation with pct.hs.grad, pct.bach.deg and some negative correlation with pct.below.pov, pct.unemp. And pct.hs.grad, pct.bach.deg, pct.below.pov, and pct.unemp. are moderately correlated with each other.
These correlations are expected. total income = per-capita income * population, so it is reasonable that they are correlated, and huge population in some way means that more people are criming, and also more people being doctors, and this county will need more beds in the hospital for this large amount of people. At the same time, more crimes means that more people are hurt and thus more doctors and hospital beds needed. Also, per-capita income has positive correlation with pct.hs.grad, pct.bach.deg because people with high school or college education are more likely to get high income than those who do not complete at least 12 years of school. And per-capita income has negative correlation with pct.below.pov, and pct.unemp because the increase of people with income below poverty level and people unemployed means that they are getting really low income.

2. How per-capita income was related to crime number, and does different regions of the country matter for this relationship? Is it better or more reasonable to use (number of crimes)/(population)?

From the two ANOVA test, we can know that total crime number and region are significant variables for per-capita income, which means that per-capita income has relationship with both of them. Also, as the interaction term is not significant, each region results in different additive change on per-capita income. And I think using total crime numbers instead of per-capita crime is better for predicting the per-capita income after comparing the diagnostic plots, summaries, AIC, and BIC values of two models.

3. What is the best model predicting per-capita income from the other variables, which best reflects the social science and the meaning of the variables?

And the final model I select is

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

For the numeric variables, the positive correlation between y variable and doctors, pct.bach.deg, and the negative correlation between y variable and pct.below.pov can both fit the correlation plot and our expectation at the very beginning, which makes this model reasonable in both statiscal and social aspects.

4. Should we be worried about either the missing states or the missing counties?

There is also an important problem is that there are missing states and missing counties in the datasets (48/51 states and 373/3000 counties appear in the dataset). We should be care about it because there are too many missing states such that there might be many other variables that are significant in these states and might be considered as not valued in this project. Also, the missing data can result in some missing information in the state and region variable, which might be biased under what we have now.

Strengths:
Estimated coefficients have the expected sign.
The model is confirmed by stepwise and All subsets procedures.
Variables are either in their original scale, or are transformed by logarithm or power, and final model is concise, which makes explaining the models to people good at social science & economics but not statistics easier.
Weakness:
The residual diagnostic plots are just OK.
Do not have time on exploring deeply on state, and other variables can result in different slopes for variables predicting the per-capita income.

Thus, in the future research on this same topic, how state variable can interactively change per-capita income with other variables should be concerned, and some missing data should be collected and added on if possible.

Reference

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005) Applied Linear Statistical Models. 5th Edition, McGraw-Hill, Irwin, New York.

Code Appendix

Appendix A: Data summary and EDA

```
cdi <- read.table("cdi.dat")
#View(cdi)
```

```
#colnames(cdi)
```

The distribution of each numeric variable and the statistics are shown below.

```
attach(cdi)
table <- matrix(c(summary(land.area),
summary(pop),
summary(pop.18_34),
summary(pop.65_plus),
summary(doctors),
summary(hosp.beds),
summary(crimes),
summary(pct.hs.grad),
summary(pct.bach.deg),
summary(pct.below.pov),
summary(pct.unemp),
summary(per.cap.income),
summary(tot.income)), ncol = 6, byrow = TRUE)
detach(cdi)
```

Make a table or tables showing appropriate summary statistics for each variable in the data set. Note that summary statistics for continuous variables will be different from the summary statistics for categorical variables.

```
rownames(table)<-c("land.area", "pop", "pop.18_34", "pop.65_plus", "doctors", "hosp.beds", "crimes", "p

colnames(table) <- c("Min.", "1st Qu.",  "Median","Mean", "3rd Qu.", "Max.")
table
```

```
##                   Min.    1st Qu.    Median         Mean    3rd Qu.      Max.
## land.area         15.0    451.250    656.50 1.041411e+03    946.750   20062.0
## pop           100043.0 139027.250 217280.50 3.930109e+05 436064.500 8863164.0
## pop.18_34         16.4     26.200     28.10 2.856841e+01     30.025      49.7
## pop.65_plus        3.0      9.875     11.75 1.216977e+01     13.625      33.8
## doctors           39.0    182.750    401.00 9.879977e+02   1036.000   23677.0
## hosp.beds         92.0    390.750    755.00 1.458627e+03   1575.750   27700.0
## crimes           563.0   6219.500  11820.50 2.711162e+04  26279.500  688936.0
## pct.hs.grad       46.6     73.875     77.70 7.756068e+01     82.400      92.9
## pct.bach.deg       8.1     15.275     19.70 2.108114e+01     25.325      52.3
## pct.below.pov      1.4      5.300      7.90 8.720682e+00     10.900      36.3
## pct.unemp          2.2      5.100      6.20 6.596591e+00      7.500      21.3
## per.cap.income  8899.0  16118.250  17759.00 1.856148e+04  20270.000   37541.0
## tot.income      1141.0   2311.000   3857.00 7.869273e+03   8654.250  184230.0
```

The frequency of each unique value of the two categorical variables are shown below.

```
library(dplyr)
```

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

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

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

```
table_1 <- data.frame(summary(as.factor(cdi$state)))
table_1 <- table_1 %>% rename(frequency=summary.as.factor.cdi.state..)
table_1
```

```
##    frequency
## AL         7
## AR         2
## AZ         5
## CA        34
## CO         9
## CT         8
## DC         1
## DE         2
## FL        29
## GA         9
## HI         3
## ID         1
## IL        17
## IN        14
## KS         4
## KY         3
## LA         9
## MA        11
## MD        10
## ME         5
## MI        18
## MN         7
## MO         8
## MS         3
## MT         1
## NC        18
## ND         1
## NE         3
## NH         4
## NJ        18
## NM         2
## NV         2
```

```
## NY        22
## OH        24
## OK         4
## OR         6
## PA        29
## RI         3
## SC        11
## SD         1
## TN         8
## TX        28
## UT         4
## VA         9
## VT         1
## WA        10
## WI        11
## WV         1
```

```
table_2 <- data.frame(summary(as.factor(cdi$region)))
table_2 <- table_2 %>% rename(frequency=summary.as.factor.cdi.region..)
table_2
```

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

id and county are not included because each row has a unique value for both of them, thus useless for the analysis.

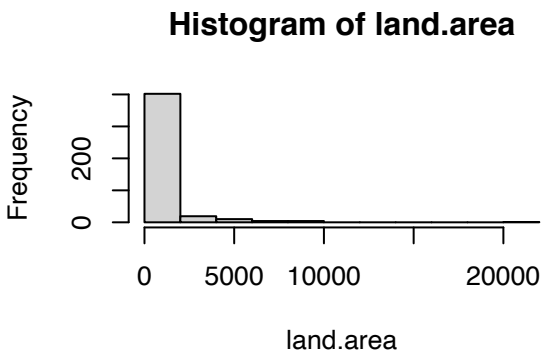
Indicate where (in which variables) there is missing data (NA's), if any, how much there is (in each variable) and why it might be there.

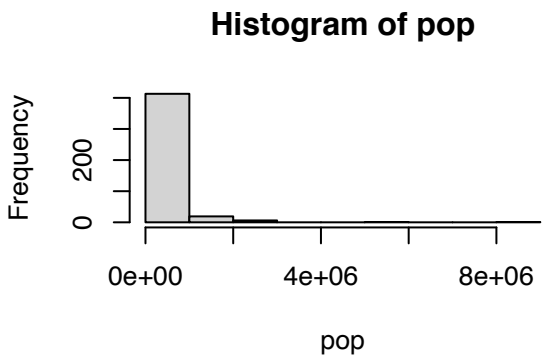
```
#is.na(cdi)
```

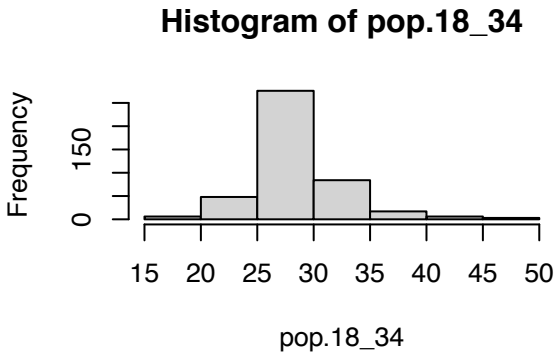
There is no missing data (NA's).

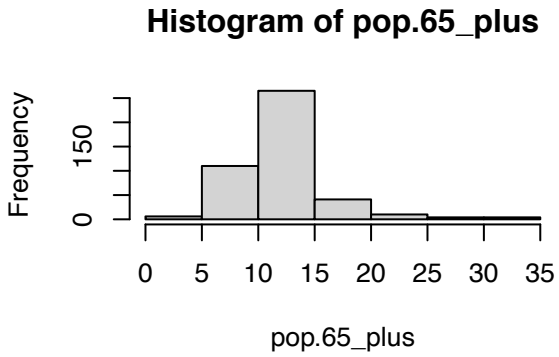
Make some appropriate descriptive EDA plots to illustrate any important features of the variables or possible important relationships among them.

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

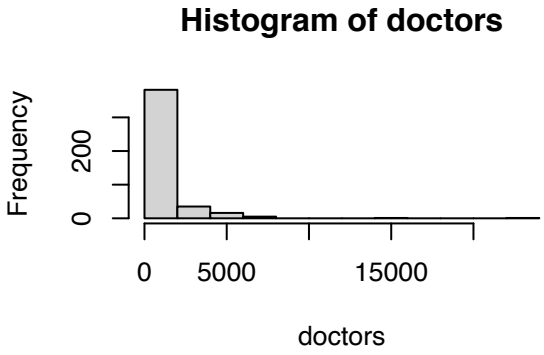


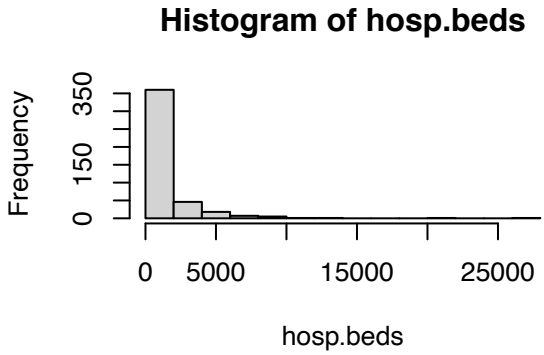


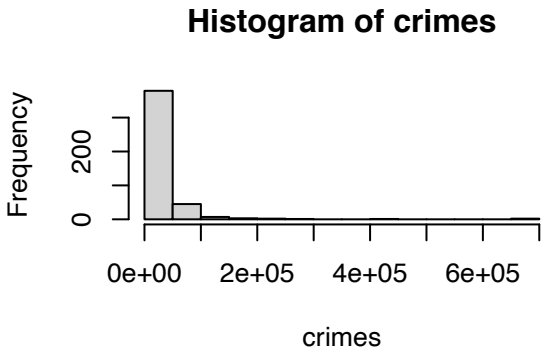


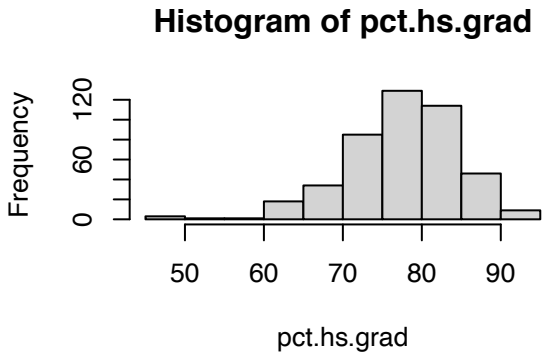



```
hist(doctors)
hist(hosp.beds)
hist(crimes)
hist(pct.hs.grad)
```

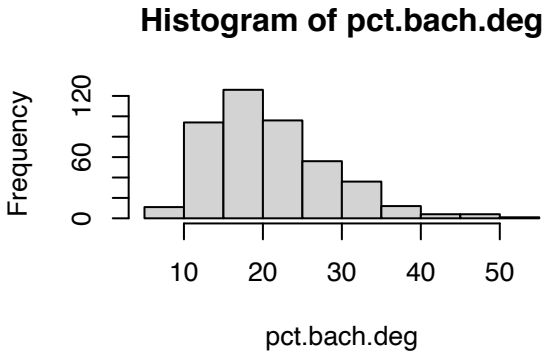


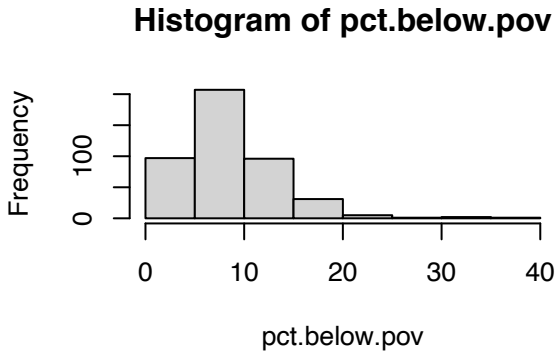


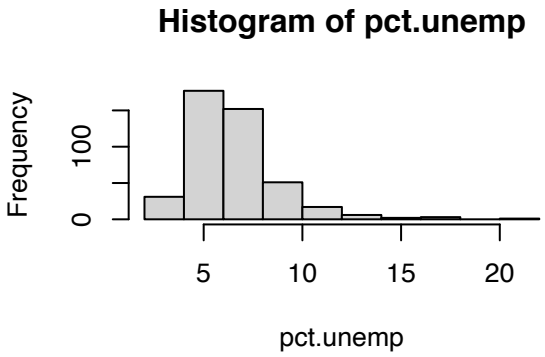


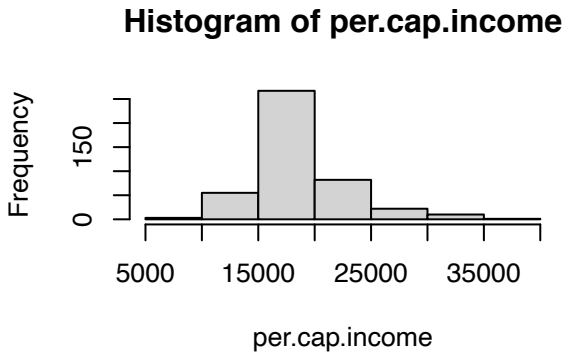



```
hist(pct.bach.deg)
hist(pct.below.pov)
hist(pct.unemp)
hist(per.cap.income)
```

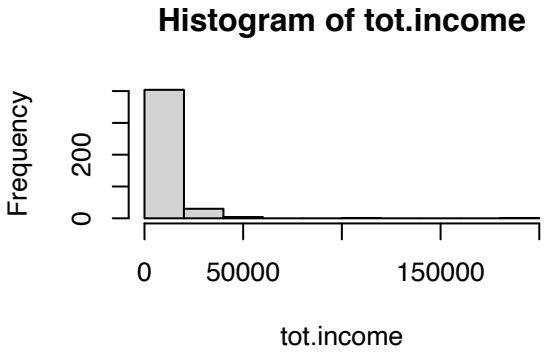









```
hist(tot.income)
detach(cdi)
```

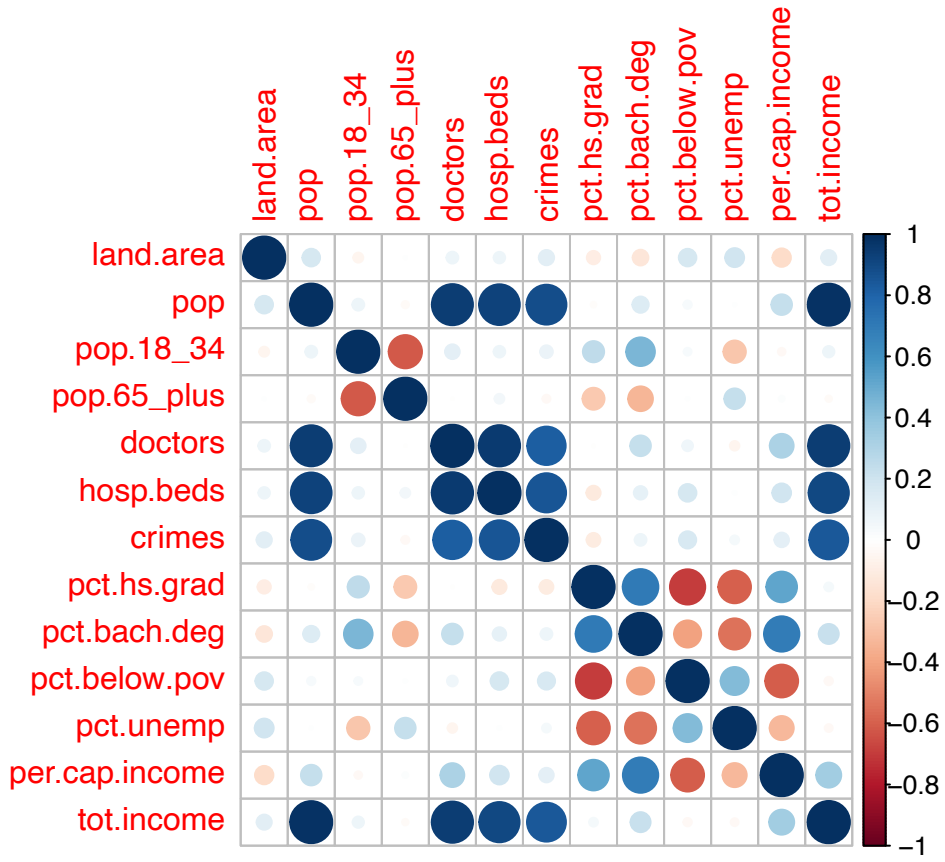


We can see that per.cap.income, land.area, pop, doctors, hosp.beds, crimes, and tot.income are highly skewed to right, thus we may need to apply log transformation to them to put the tails in.

```
library("corrplot")
```

```
## corrplot 0.90 loaded
```

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

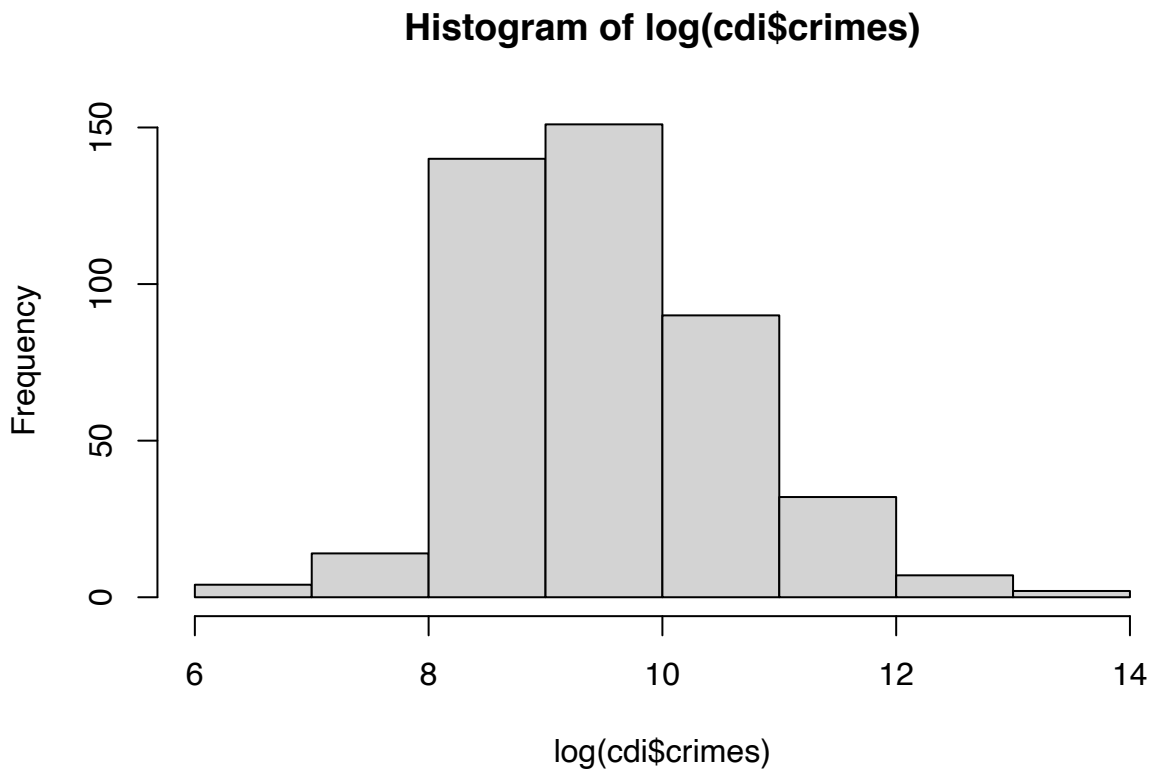


From the correlation plot above, we can see that the dots are dark means that the corresponding two variables are highly correlated, so when the highly correlated variables occur in the same model, we will need to consider if there is any confounding variables or missing variables that may mislead the model. These correlations are expected. total income = per-capita income * population, so it is reasonable that they are correlated, and huge population in some way means that more people are criming, and also more people being doctors, and this county will need more beds in the hospital for this large amount of people. At the same time, more crimes means that more people are hurt and thus more doctors and hospital beds needed. Also, per-capita income has positive correlation with pct.hs.grad, pct.bach.deg because people with high school or college education are more likely to get high income than those who do not complete at least 12 years of school. And per-capita income has negative correlation with pct.below.pov, and pct.unemp because the increase of people with income below poverty level and people unemployed means that they are getting really low income.

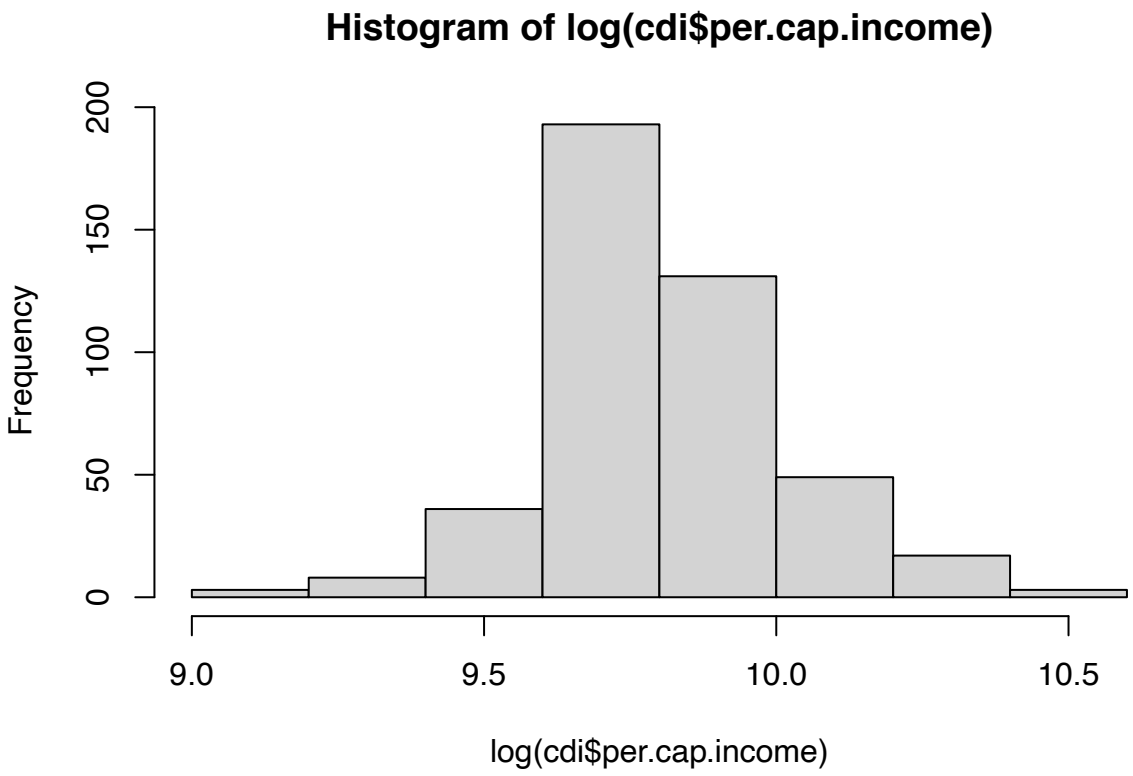
Appendix B: How crimes and region are related to per-capita crime

Build a regression model that predicts per-capita income from crime rate and region of the country. Should there be any interactions in the model? What does your model say about the relationship between per- capita income and crime rate? Do your answers change, depending on whether you use number of crimes, or "per-capita crime" = (number of crimes)/(population) as a crime rate measure? If so, which one best answers the question? Why? Show the fitted model results and explain your answer to these questions in terms of those results.

```
hist(log(cdi$crimes))
```




```
hist(log(cdi$per.cap.income))
```



Log transformation is applied to crimes and per.cap.income.

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

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(crimes), data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75042 -0.11569 -0.02976  0.09597  0.74498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.295146   0.083764  110.97  < 2e-16 ***
## log(crimes) 0.053858   0.008758    6.15 1.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1986 on 438 degrees of freedom
## Multiple R-squared:  0.07948,    Adjusted R-squared:  0.07738
## F-statistic: 37.82 on 1 and 438 DF,  p-value: 1.752e-09
```

```r
summary(fit1)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(crimes) + region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68757 -0.10557 -0.01422  0.08905  0.78946
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.188431   0.079812 115.125  < 2e-16 ***
## log(crimes)  0.066695   0.008421   7.920 2.00e-14 ***
## regionNE     0.104458   0.025531   4.091 5.11e-05 ***
## regionS     -0.086983   0.023618  -3.683  0.00026 ***
## regionW     -0.055280   0.028167  -1.963  0.05033 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1854 on 435 degrees of freedom
## Multiple R-squared:  0.2032, Adjusted R-squared:  0.1959
## F-statistic: 27.74 on 4 and 435 DF,  p-value: < 2.2e-16
```

For this original model, we can see that log(crimes) is a significant variable for per.capita.income, and each percent increase of crime leads to about 5 percent increase of per.capita.income. Then we check if the model becomes better with interaction term in the model.

```r
fit2<-lm(log(per.cap.income)~log(crimes)+region+(log(crimes):region), data = cdi)
summary(fit2)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(crimes) + region + (log(crimes):region),
##     data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68552 -0.10418 -0.01444  0.08302  0.79755
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.33677    0.14579  64.044  < 2e-16 ***
## log(crimes)           0.05064    0.01566   3.233  0.00132 **
## regionNE             -0.18407    0.21515  -0.856  0.39272
## regionS              -0.19717    0.21211  -0.930  0.35312
## regionW              -0.31439    0.24465  -1.285  0.19947
## log(crimes):regionNE  0.03122    0.02311   1.351  0.17749
## log(crimes):regionS   0.01211    0.02228   0.544  0.58696
## log(crimes):regionW   0.02727    0.02523   1.081  0.28028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1855 on 432 degrees of freedom
## Multiple R-squared:  0.2073, Adjusted R-squared:  0.1945
## F-statistic: 16.14 on 7 and 432 DF,  p-value: < 2.2e-16
```

It seems that r squared value does not change a lot with interaction term. Then we use anova test to check H0: fit1(no interaction term) is enough or Ha: reject H0 so that we will need an interaction term.

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

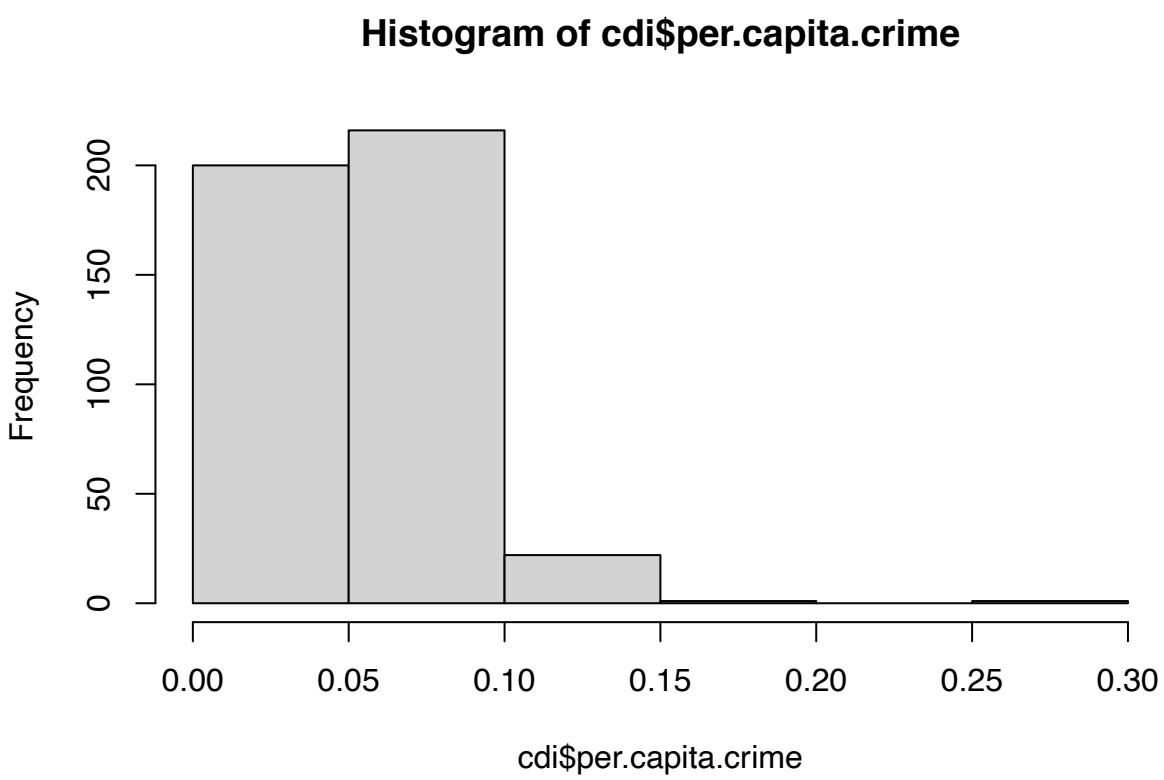
```
## Analysis of Variance Table
##
## Model 1: log(per.cap.income) ~ log(crimes)
## Model 2: log(per.cap.income) ~ log(crimes) + region
## Model 3: log(per.cap.income) ~ log(crimes) + region + (log(crimes):region)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 17.271
## 2    435 14.949  3   2.32194 22.4823 1.523e-13 ***
## 3    432 14.872  3   0.07678  0.7434    0.5266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Thus, we can know that model with region as an interaction term is not better than region as a additive term. And we will choose model 2 in this case.

Appendix C: How per-capita crime and region are related to per-capita crime

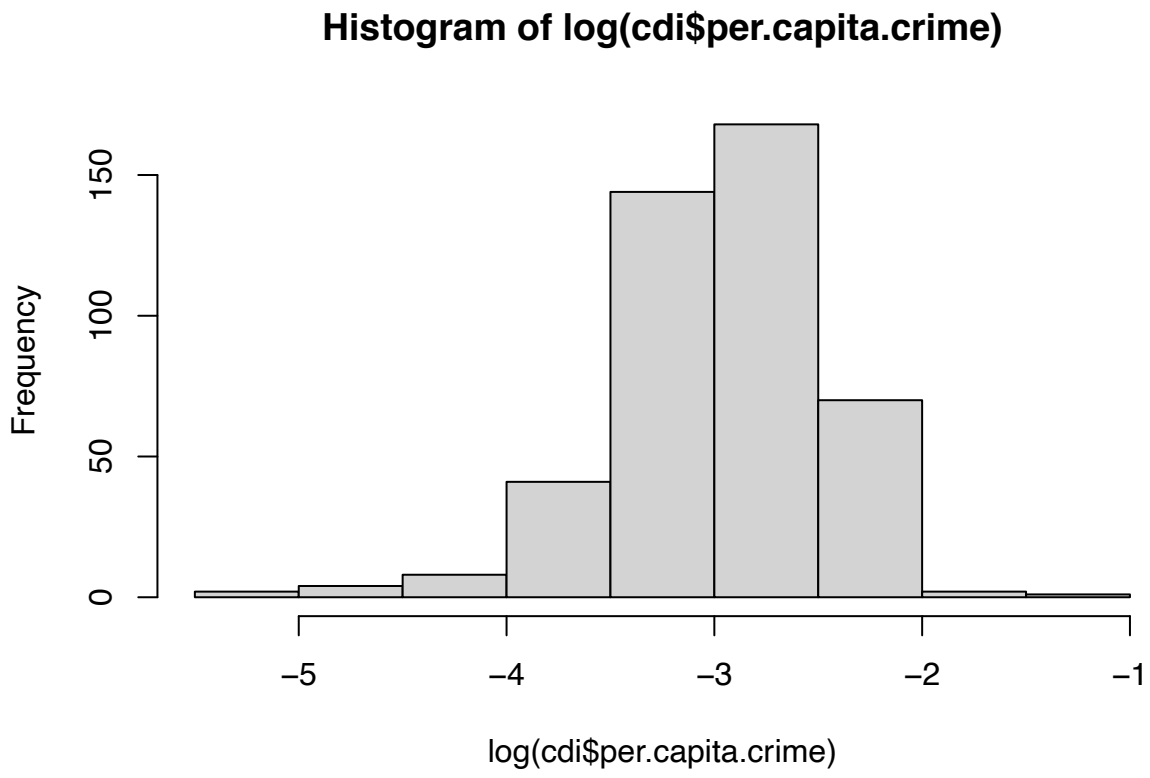
```
cdi["per.capita.crime"] = cdi$crimes/cdi$pop
```

```
hist(cdi$per.capita.crime)
```



We can see that the per.capita.crime is highle right skewed, so we may apply a log transformation on it, and it looks better now.

```
hist(log(cdi$per.capita.crime))
```




```
fit <- lm(log(per.cap.income)~log(per.capita.crime), data = cdi)
fit3 <- lm(log(per.cap.income)~log(per.capita.crime)+region, data = cdi)
summary(fit3)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(per.capita.crime) + region,
##     data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65832 -0.11431 -0.01548  0.10838  0.75657
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.93628    0.06934 143.303  < 2e-16 ***
## log(per.capita.crime)  0.04243    0.02148   1.975  0.04885 *
## regionNE               0.11457    0.02760   4.151 3.99e-05 ***
## regionS               -0.07456    0.02624  -2.841  0.00471 **
## regionW               -0.02426    0.03002  -0.808  0.41952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1974 on 435 degrees of freedom
## Multiple R-squared:  0.09645,    Adjusted R-squared:  0.08814
## F-statistic: 11.61 on 4 and 435 DF,  p-value: 5.776e-09
```

```r
fit4 <- lm(log(per.cap.income)~log(per.capita.crime)+region+(log(per.capita.crime):region), data = cdi)
summary(fit4)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(per.capita.crime) + region +
##     (log(per.capita.crime):region), data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65410 -0.11829 -0.01708  0.10399  0.76628
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     9.91177    0.10503  94.367   <2e-16 ***
## log(per.capita.crime)           0.03454    0.03327   1.038    0.300
## regionNE                        0.21007    0.17165   1.224    0.222
## regionS                        -0.10137    0.16072  -0.631    0.529
## regionW                         0.07689    0.26753   0.287    0.774
## log(per.capita.crime):regionNE  0.02924    0.05232   0.559    0.577
## log(per.capita.crime):regionS  -0.01104    0.05554  -0.199    0.843
## log(per.capita.crime):regionW   0.03495    0.09268   0.377    0.706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.198 on 432 degrees of freedom
## Multiple R-squared:  0.09773,    Adjusted R-squared:  0.08311
## F-statistic: 6.685 on 7 and 432 DF,  p-value: 1.575e-07
```

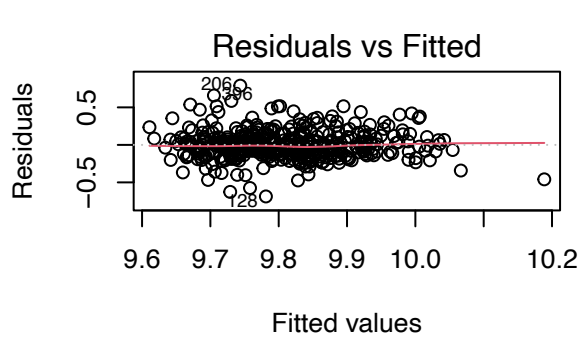
```r
anova(fit, fit3, fit4)
```

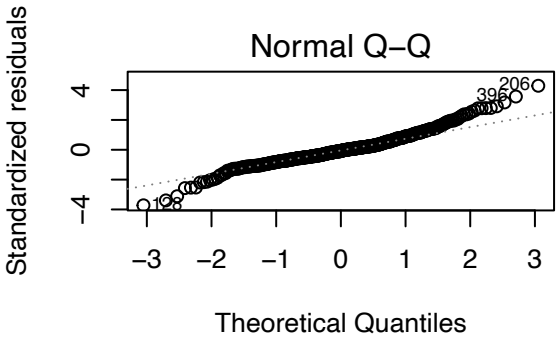
```
## Analysis of Variance Table
##
## Model 1: log(per.cap.income) ~ log(per.capita.crime)
## Model 2: log(per.cap.income) ~ log(per.capita.crime) + region
## Model 3: log(per.cap.income) ~ log(per.capita.crime) + region + (log(per.capita.crime):region)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 18.697
## 2    435 16.952  3   1.74465 14.8407 3.263e-09 ***
## 3    432 16.928  3   0.02408  0.2048     0.893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

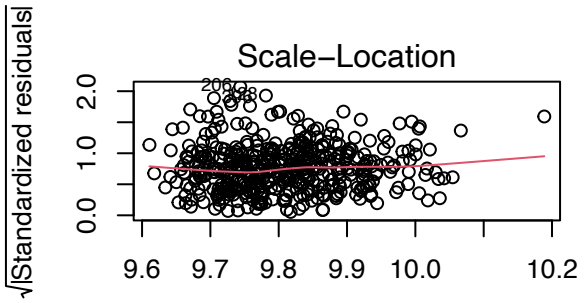
We can see that for model per.cap.income ~ log(per.capita.crime) + region, we also do not need to add the interaction term as the p-value is $0.89 > 0.05$.

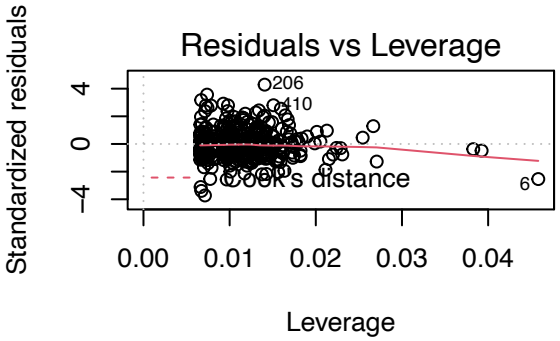
Appendix D: Choose between total crime number and per-capita crime

```r
par(mfrow = c(2,2))
plot(fit1)
```

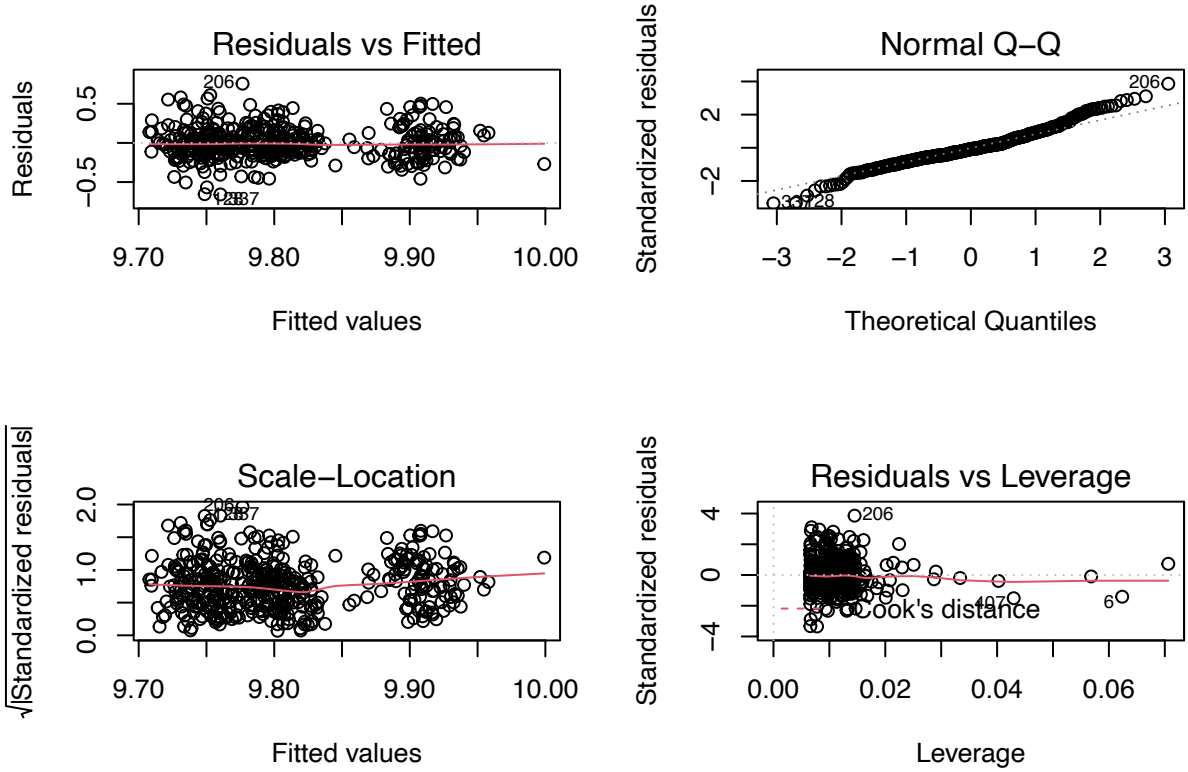









```
par(mfrow = c(2,2))
plot(fit3)
```




```
AIC(fit1, fit3)
```

```
##      df       AIC
## fit1  6 -227.4746
## fit3  6 -172.1347
```

```
BIC(fit1, fit3)
```

```
##      df       BIC
## fit1  6 -202.9539
## fit3  6 -147.6140
```

To decide which one of fit 1 and fit 3, apart from the summary of two models, we will also need to check the diagnostic plots for them. Based on the summary, we can see that the r squared value is larger for model 1, per.cap.income ~ log(crimes) + region, than model 3, per.cap.income ~ log(per.capita.crime) + region, and also the log(per.capita.crime) is not a significant variable in model 3. For the diagnostic plots, residuals vs.fitted plot for model 1 looks good as there is no pattern and the mean is at about 0; there are some points off the line at the sides; there is a slightly upward pattern in the scale-location plot; there is no points with both high residuals and high leverage. For model 3, the plots are quite similar with those of model 1, except that points on residuals and scale-location plots are roughly clustered into three groups, so that maybe regions affect the model more for this one. Also, when comparing their AIC values and BIC values, we can see that fit 1 has both lower AIC value and BIC value, which makes it better than fit 3. Overall, I will choose model 1 for its lower AIC and BIC values and higher r squared value.

Appendix E: Variable selection process

Use methods we have discussed in class and/or methods from Sheather Chapters 5, 6 & 7 (including, as needed: transformations, interactions, variable selection, residual analysis, fit indices, etc.) to find the multiple regression model predicting per-capita income from the other variables.

(1) Stepwise selection

We can see that per.cap.income, land.area, pop, doctors, hosp.beds, crimes, pct.bach.deg, pct.below.pov, pct.unemp, and tot.income are highly skewed to right, thus we may need to apply log transformation to them to put the tails in. Also, pct.hs.grad is somehow skewed to the left, so we apply a power transformation of degree 2 to it. Also, as per capita income = total income/population, and we also can see the high correlation among them in the correlation plot. Thus, I decide to remove these two features. Here I also exclude the variable state and region, and I will process with these categorical variables later in the process. Now I jsut want to know how these numerical variables perform in the model.

```
fit_all <- lm(log(per.cap.income) ~ (log(land.area)+log(doctors)+log(hosp.beds)+log(crimes)+pct.bach.deg
```

```
summary(fit_all)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ (log(land.area) + log(doctors) +
##     log(hosp.beds) + log(crimes) + pct.bach.deg + pct.below.pov +
##     pct.unemp + pct.hs.grad + pop.18_34 + pop.65_plus), data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35561 -0.04712 -0.00846  0.04522  0.27681
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.313312   0.108644  94.927  < 2e-16 ***
## log(land.area) -0.035641   0.004825  -7.386 7.95e-13 ***
## log(doctors)    0.052055   0.012502   4.164 3.79e-05 ***
## log(hosp.beds)  0.016215   0.012008   1.350   0.1776
## log(crimes)    -0.004066   0.007831  -0.519   0.6039
## pct.bach.deg    0.015712   0.001027  15.305  < 2e-16 ***
## pct.below.pov  -0.024945   0.001303 -19.138  < 2e-16 ***
## pct.unemp       0.011130   0.002186   5.091 5.34e-07 ***
## pct.hs.grad    -0.004738   0.001086  -4.363 1.61e-05 ***
## pop.18_34      -0.015542   0.001306 -11.897  < 2e-16 ***
## pop.65_plus    -0.003309   0.001371  -2.413   0.0162 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08168 on 429 degrees of freedom
## Multiple R-squared:  0.8475, Adjusted R-squared:  0.8439
## F-statistic: 238.4 on 10 and 429 DF,  p-value: < 2.2e-16
```

```r
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

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

Here we start doing a variable selection by stepwise selection and all subsets selection and make a comparison.

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

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

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

```r
step_model <- stepAIC(fit_all, direction = "both", trace = FALSE)
summary(step_model)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + log(doctors) +
##     pct.bach.deg + pct.below.pov + pct.unemp + pct.hs.grad +
##     pop.18_34 + pop.65_plus, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35756 -0.04551 -0.00543  0.04844  0.27399
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.3159666  0.1025858 100.559  < 2e-16 ***
## log(land.area) -0.0364935  0.0047728  -7.646 1.36e-13 ***
## log(doctors)    0.0626053  0.0041029  15.259  < 2e-16 ***
## pct.bach.deg    0.0152149  0.0009242  16.462  < 2e-16 ***
## pct.below.pov  -0.0246144  0.0012631 -19.488  < 2e-16 ***
## pct.unemp       0.0107688  0.0021696   4.963 9.99e-07 ***
## pct.hs.grad    -0.0046579  0.0010843  -4.296 2.15e-05 ***
## pop.18_34      -0.0153488  0.0012988 -11.818  < 2e-16 ***
## pop.65_plus    -0.0027664  0.0012978  -2.132   0.0336 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08167 on 431 degrees of freedom
## Multiple R-squared:  0.8468, Adjusted R-squared:  0.8439
## F-statistic: 297.7 on 8 and 431 DF,  p-value: < 2.2e-16
```

From the stepwise selction, model we can is log(per.cap.income) ~ log(land.area) + log(doctors) + log(pct.bach.deg) + log(pct.below.pov) + log(pct.unemp) + pop.18_34 + pct.hs.grad + pop.65_plus.

(2) All subsets selection

```
all_subsets <- regsubsets(log(per.cap.income) ~ log(land.area)+log(doctors)+log(hosp.beds)+log(crimes)+
```

```
cdi_sum <- summary(all_subsets)
data.frame(
  adj_r2 = which.max(cdi_sum$adjr2),
  cp = which.min(cdi_sum$cp),
  bic = which.min(cdi_sum$bic)
)
```

```
##   adj_r2 cp bic
## 1      8  8   7
```

```
coef(all_subsets, 1:8)
```

```
## [[1]]
##  (Intercept) pct.bach.deg
##   9.42153388   0.01828273
##
## [[2]]
##   (Intercept)  log(doctors) pct.below.pov
##    9.48447189    0.09381497   -0.02919999
##
## [[3]]
##   (Intercept)  log(doctors)  pct.bach.deg pct.below.pov
##   9.422105059   0.070869806   0.007533288  -0.024073152
##
## [[4]]
##   (Intercept)  log(doctors)  pct.bach.deg pct.below.pov     pop.18_34
##    9.73156229    0.06434105    0.01245298   -0.02031627   -0.01420342
##
## [[5]]
##    (Intercept) log(land.area)   log(doctors)   pct.bach.deg  pct.below.pov
##    10.00506798    -0.03733574     0.06324794     0.01197955    -0.01957811
##      pop.18_34
##    -0.01490014
##
## [[6]]
##    (Intercept) log(land.area)   log(doctors)   pct.bach.deg  pct.below.pov
##     9.90343798    -0.04021183     0.06286862     0.01341559    -0.02138922
##      pct.unemp      pop.18_34
##     0.01290540    -0.01409166
##
## [[7]]
##    (Intercept) log(land.area)   log(doctors)   pct.bach.deg  pct.below.pov
##   10.222495041   -0.035674062    0.060676872    0.015385301   -0.024278371
##      pct.unemp    pct.hs.grad      pop.18_34
##    0.010603691   -0.004406396   -0.013900201
```

```
##
## [[8]]
##    (Intercept) log(land.area)   log(doctors)   pct.bach.deg  pct.below.pov
##   10.315966592   -0.036493494    0.062605267    0.015214937   -0.024614405
##      pct.unemp    pct.hs.grad      pop.18_34    pop.65_plus
##    0.010768825   -0.004657948   -0.015348817   -0.002766377
```

For the reason that bic tends to select simple model, we choose the model with eight variables.

```
fit_8 <- lm(log(per.cap.income) ~ log(land.area) + log(doctors) + pct.bach.deg + pct.below.pov +pct.une
summary(fit_8)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(land.area) + log(doctors) +
##     pct.bach.deg + pct.below.pov + pct.unemp + pct.hs.grad +
##     pop.18_34 + pop.65_plus, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35756 -0.04551 -0.00543  0.04844  0.27399
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.3159666  0.1025858 100.559  < 2e-16 ***
## log(land.area) -0.0364935  0.0047728  -7.646 1.36e-13 ***
## log(doctors)    0.0626053  0.0041029  15.259  < 2e-16 ***
## pct.bach.deg    0.0152149  0.0009242  16.462  < 2e-16 ***
## pct.below.pov  -0.0246144  0.0012631 -19.488  < 2e-16 ***
## pct.unemp       0.0107688  0.0021696   4.963 9.99e-07 ***
## pct.hs.grad    -0.0046579  0.0010843  -4.296 2.15e-05 ***
## pop.18_34      -0.0153488  0.0012988 -11.818  < 2e-16 ***
## pop.65_plus    -0.0027664  0.0012978  -2.132   0.0336 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08167 on 431 degrees of freedom
## Multiple R-squared:  0.8468, Adjusted R-squared:  0.8439
## F-statistic: 297.7 on 8 and 431 DF,  p-value: < 2.2e-16
```

From the selection of both methods, we can see that we get same variables from them. Then, we will need to consider the how can the variable indicating the regions can affect the model and prediction.

(3) Interaction with region

```
fit_region <- lm(log(per.cap.income) ~ (log(land.area) + log(doctors) + pct.bach.deg + pct.below.pov +p
summary(fit_region)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ (log(land.area) + log(doctors) +
##     pct.bach.deg + pct.below.pov + pct.unemp + pct.hs.grad +
```

```
##     pop.18_34 + pop.65_plus) * region, data = cdi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.239497 -0.042518 -0.002899  0.038705  0.315955
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              10.1550994  0.3077758  32.995  < 2e-16 ***
## log(land.area)           -0.0355230  0.0155258  -2.288 0.022654 *
## log(doctors)              0.0548293  0.0094485   5.803 1.32e-08 ***
## pct.bach.deg              0.0140191  0.0029305   4.784 2.41e-06 ***
## pct.below.pov            -0.0233702  0.0038627  -6.050 3.30e-09 ***
## pct.unemp                 0.0176067  0.0051819   3.398 0.000747 ***
## pct.hs.grad              -0.0026649  0.0034861  -0.764 0.445055
## pop.18_34                -0.0150740  0.0028317  -5.323 1.69e-07 ***
## pop.65_plus              -0.0012483  0.0050165  -0.249 0.803614
## regionNE                  0.4813749  0.3863061   1.246 0.213451
## regionS                  -0.0552517  0.3396107  -0.163 0.870843
## regionW                   1.3969067  0.4575796   3.053 0.002417 **
## log(land.area):regionNE -0.0050730  0.0204207  -0.248 0.803932
## log(land.area):regionS  -0.0058664  0.0177783  -0.330 0.741589
## log(land.area):regionW   0.0136894  0.0185229   0.739 0.460306
## log(doctors):regionNE    0.0001267  0.0135190   0.009 0.992526
## log(doctors):regionS     0.0042557  0.0116550   0.365 0.715198
## log(doctors):regionW    -0.0046667  0.0132947  -0.351 0.725759
## pct.bach.deg:regionNE    0.0060237  0.0040533   1.486 0.138025
## pct.bach.deg:regionS    -0.0015550  0.0032102  -0.484 0.628384
## pct.bach.deg:regionW     0.0069577  0.0036552   1.903 0.057687 .
## pct.below.pov:regionNE  -0.0009949  0.0052677  -0.189 0.850294
## pct.below.pov:regionS    0.0068718  0.0042992   1.598 0.110736
## pct.below.pov:regionW   -0.0167523  0.0055989  -2.992 0.002941 **
## pct.unemp:regionNE      -0.0063048  0.0075950  -0.830 0.406962
## pct.unemp:regionS       -0.0243492  0.0068439  -3.558 0.000418 ***
## pct.unemp:regionW       -0.0192087  0.0070270  -2.734 0.006541 **
## pct.hs.grad:regionNE    -0.0033331  0.0044706  -0.746 0.456373
## pct.hs.grad:regionS      0.0023152  0.0038518   0.601 0.548134
## pct.hs.grad:regionW     -0.0185423  0.0046646  -3.975 8.33e-05 ***
## pop.18_34:regionNE      -0.0060991  0.0042036  -1.451 0.147582
## pop.18_34:regionS       -0.0008273  0.0034566  -0.239 0.810970
## pop.18_34:regionW        0.0030516  0.0048005   0.636 0.525342
## pop.65_plus:regionNE    -0.0076628  0.0063347  -1.210 0.227119
## pop.65_plus:regionS      0.0009166  0.0052822   0.174 0.862326
## pop.65_plus:regionW      0.0037008  0.0064632   0.573 0.567239
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07573 on 404 degrees of freedom
## Multiple R-squared:  0.8765, Adjusted R-squared:  0.8658
## F-statistic: 81.92 on 35 and 404 DF,  p-value: < 2.2e-16
```

From the summary table above, we can see that, after removing the insignificnat variables, pct.unemp:region, pct.hs.grad:region, pct.below.pov:region, region, pop.18_34, pct.unemp, pct.below.pov, pct.bach.deg, log(doctors), log(land.area) are the variables left significant for the new model with interaction of region. Here

we get the new model log(per.cap.income) ~ pct.unemp:region+ pct.hs.grad:region+ pct.below.pov:region+ region+ pop.18_34+ pct.unemp+ pct.below.pov+pct.bach.deg+ log(doctors)+log(land.area)

```
fit_region_new <- lm(log(per.cap.income) ~ pct.unemp:region+pct.hs.grad:region+ pct.below.pov:region+ r
summary(fit_region_new)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ pct.unemp:region + pct.hs.grad:region +
##     pct.below.pov:region + region + pop.18_34 + pct.unemp + pct.below.pov +
##     pct.bach.deg + log(doctors) + log(land.area), data = cdi)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.294186 -0.043597 -0.001583  0.037667  0.311609
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             10.2421239  0.2176557  47.057  < 2e-16 ***
## regionNE                -0.0520070  0.2707173  -0.192 0.847750
## regionS                 -0.0389718  0.2383516  -0.164 0.870199
## regionW                  1.3910484  0.3408962   4.081 5.38e-05 ***
## pop.18_34               -0.0149347  0.0010897 -13.705  < 2e-16 ***
## pct.unemp                0.0197400  0.0046254   4.268 2.44e-05 ***
## pct.below.pov           -0.0252029  0.0032612  -7.728 8.12e-14 ***
## pct.bach.deg             0.0156310  0.0009715  16.090  < 2e-16 ***
## log(doctors)             0.0572284  0.0040082  14.278  < 2e-16 ***
## log(land.area)          -0.0381738  0.0053996  -7.070 6.51e-12 ***
## pct.unemp:regionNE      -0.0129841  0.0070423  -1.844 0.065929 .
## pct.unemp:regionS       -0.0231138  0.0061365  -3.767 0.000189 ***
## pct.unemp:regionW       -0.0217357  0.0065225  -3.332 0.000937 ***
## regionNC:pct.hs.grad    -0.0043532  0.0024515  -1.776 0.076501 .
## regionNE:pct.hs.grad    -0.0025848  0.0020257  -1.276 0.202657
## regionS:pct.hs.grad     -0.0032007  0.0014122  -2.266 0.023936 *
## regionW:pct.hs.grad     -0.0185005  0.0027800  -6.655 8.88e-11 ***
## regionNE:pct.below.pov -0.0015170  0.0046143  -0.329 0.742493
## regionS:pct.below.pov    0.0070185  0.0035199   1.994 0.046808 *
## regionW:pct.below.pov   -0.0137920  0.0051811  -2.662 0.008066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07692 on 420 degrees of freedom
## Multiple R-squared:  0.8675, Adjusted R-squared:  0.8615
## F-statistic: 144.8 on 19 and 420 DF,  p-value: < 2.2e-16
```

(4) Model selection

```
anova(fit_8, fit_region_new)
```

```
## Analysis of Variance Table
##
## Model 1: log(per.cap.income) ~ log(land.area) + log(doctors) + pct.bach.deg +
```

```
##     pct.below.pov + pct.unemp + pct.hs.grad + pop.18_34 + pop.65_plus
## Model 2: log(per.cap.income) ~ pct.unemp:region + pct.hs.grad:region +
##     pct.below.pov:region + region + pop.18_34 + pct.unemp + pct.below.pov +
##     pct.bach.deg + log(doctors) + log(land.area)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    431 2.8748
## 2    420 2.4853 11   0.38947 5.9835 4.306e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA test is made bwteen the model selected by all subsets and the model with interaction of region. By the result of test, we can say that the null hypothesis that model without interaction with region is good enough can be rejected. Thus, we think choosing the fit_region+new model is better. Several other comparisons are also made between the two models to find the best one.

```
BIC(fit_8, fit_region_new)
```

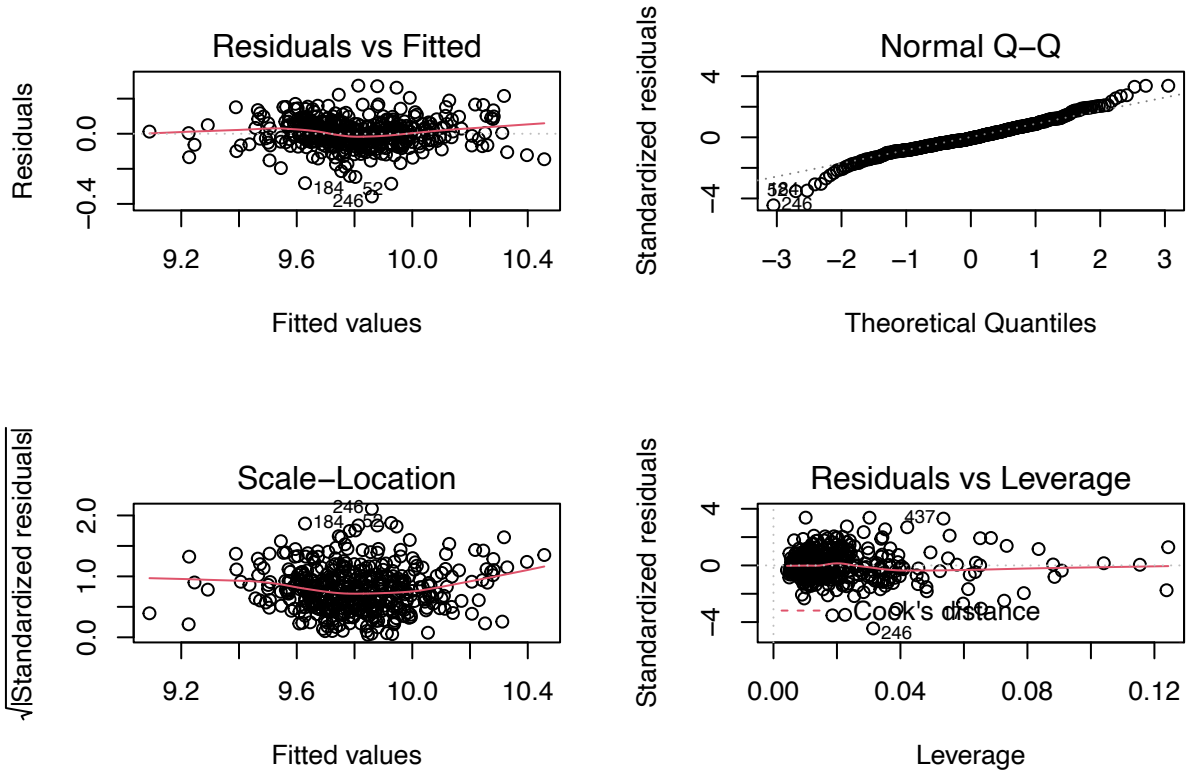
```
##                df       BIC
## fit_8          10 -904.0206
## fit_region_new 21 -901.1215
```

```
AIC(fit_8, fit_region_new)
```

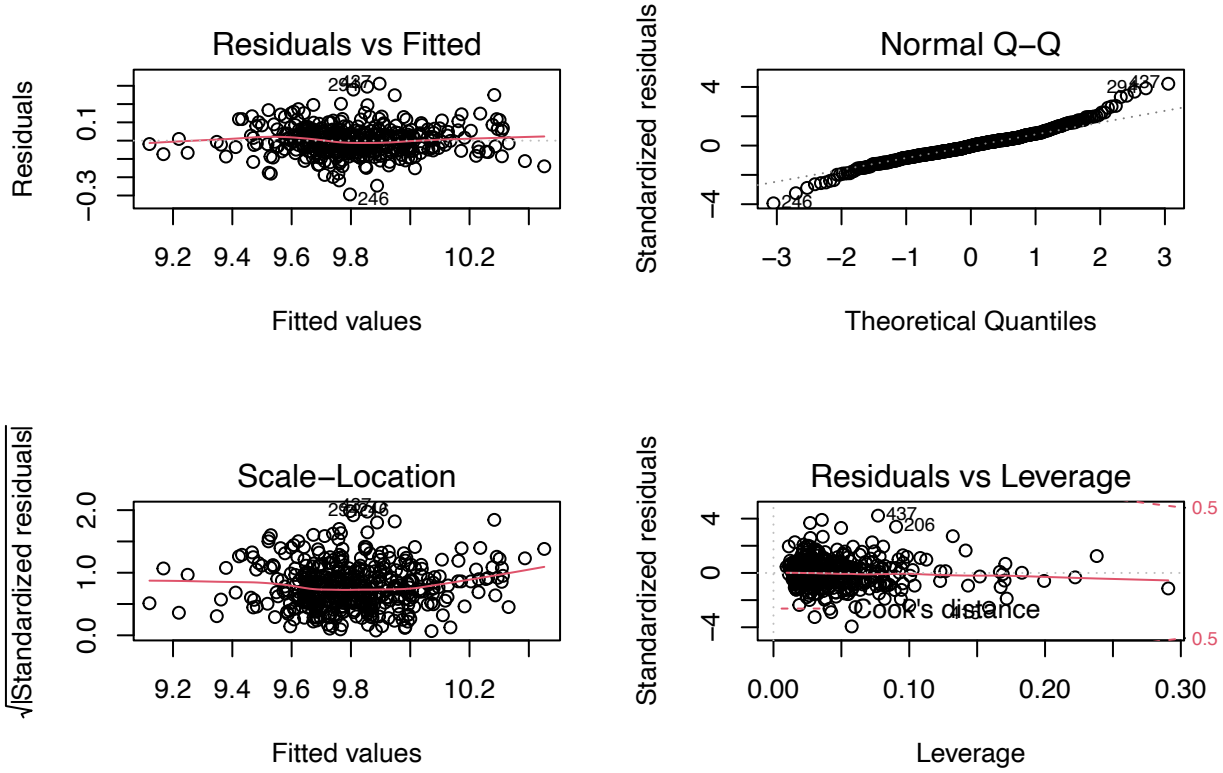
```
##                df       AIC
## fit_8          10 -944.8883
## fit_region_new 21 -986.9437
```

From BIC and AIC value of the two model, we can see that fit_8 has smaller BIC value but fit-region_new has smaller AIC value.

```
par(mfrow = c(2,2))
plot(fit_8)
```




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



From the two sets of diagnostic plots, we can find it is hard to choose based on those. Both resduals vs.fitted plots and scale-location plots has a slight upward concavity patterns shown. And from both normal q-q plots, we can see that they all have some points off the line at the very right side, which also corresponds to the residuals vs.leverage plots that they both have no points with either high residuals or high leverage, but no point with both high residuals and high leverage. They have quite similar r squared values and it is also hard to choose based on this single value. from the ANOVA test, and value comparisons above, I will choose fit_region_new as the final model selected. It has lower AIC value, though higher BIC value, but BIC value tends to choose simpler model. Also, from ANOVA test, we can see that the interaction term can make the performance of model better than before to an extent that we cannot accept that the models are quite similar by the p-value far less than 0.05.

(5) Model with state

```
all_subsets_state <- regsubsets(log(per.cap.income) ~ log(land.area)+log(doctors)+log(hosp.beds)+log(cr
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 3 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
cdi_sum <- summary(all_subsets_state)
data.frame(
  adj_r2 = which.max(cdi_sum$adjr2),
  cp = which.min(cdi_sum$cp),
```

```
  bic = which.min(cdi_sum$bic)
)
```

```
##   adj_r2 cp bic
## 1      9  9   9
```

```
coef(all_subsets_state, 1:9)
```

```
## [[1]]
##  (Intercept) pct.bach.deg
##   9.42153388   0.01828273
##
## [[2]]
##   (Intercept)  log(doctors) pct.below.pov
##    9.48447189    0.09381497   -0.02919999
##
## [[3]]
##   (Intercept)  log(doctors)  pct.bach.deg pct.below.pov
##   9.422105059   0.070869806   0.007533288  -0.024073152
##
## [[4]]
##   (Intercept)  log(doctors)  pct.bach.deg pct.below.pov     pop.18_34
##    9.73156229    0.06434105    0.01245298   -0.02031627   -0.01420342
##
## [[5]]
##    (Intercept) log(land.area)   log(doctors)   pct.bach.deg  pct.below.pov
##    10.00506798    -0.03733574     0.06324794     0.01197955    -0.01957811
##      pop.18_34
##    -0.01490014
##
## [[6]]
##    (Intercept) log(land.area)   log(doctors)   pct.bach.deg  pct.below.pov
##    10.00988950    -0.03700469     0.06250060     0.01215494    -0.01956914
##      pop.18_34        stateUT
##    -0.01501832    -0.30594763
##
## [[7]]
##    (Intercept) log(land.area)   log(doctors)   pct.bach.deg  pct.below.pov
##     9.96456492    -0.03222361     0.06117849     0.01208768    -0.01900949
##      pop.18_34        stateNJ        stateUT
##    -0.01452636     0.11574866    -0.30098690
##
## [[8]]
##    (Intercept) log(land.area)   log(doctors)   pct.bach.deg  pct.below.pov
##    10.04080958    -0.04091209     0.05799269     0.01205856    -0.01896206
##      pop.18_34        stateCA        stateNJ        stateUT
##    -0.01475241     0.08467020     0.11660717    -0.29429494
##
## [[9]]
##    (Intercept) log(land.area)   log(doctors)   pct.bach.deg  pct.below.pov
##    10.03919588    -0.04102795     0.05756553     0.01201238    -0.01857332
##      pop.18_34        stateCA        stateCT        stateNJ        stateUT
##    -0.01474387     0.08713006     0.10785995     0.12053000    -0.29161335
```

From this selection that linear model with state and region considered, we can find some state with specific meaning or are special to this model, like CA, CT, NJ, UT.