Analysis on the Average Income per Person in US

Zhuoheng Han
Department of Statistics and Data Science, Carnegie Mellon University
zhouhenh@andrew.cmu.edu

Abstract

In this paper, we focus on the factors influence the per capita income for different counties in US. We use the county demographic information (CDI) dataset to help us analyze the per capita income. Methods such as correlation heatmap, transformation, regression model, variable selection, and analysis of variance are used to find the influencing factors. We find that there are high correlated pairs such as the number of the doctors and the total number of beds, discover that positive relationship between crimes and per capita income, build a regression model predicting the per capita income, and discuss the shortcoming of missing values. In order to improve our analysis, we need to research on other counties since the dataset only contains 1/9 total counties in US.

Introduction

Social scientists are interested in determining per capita income to evaluate the life quality of the population. In this paper, we are discussing how average income per person was related to other variables associated with the county's economic, health and social well-being from county demographic information (CDI) dataset. We address four research questions:

- Which variables seem to be related to other variables in the dataset? Which are not? Are these relationships reasonable?
- Prove or Disprove a theory that per capita income should be related to crime rate, and that relationship may be different in different regions of the country. Does it matter if you use number of crimes or (number of crimes)/(population) in your analysis?
- What is the best model predicting per capita income?
- There are 51 states and around 3000 counties in US, but 48 states and 440 counties are represented in the dataset. Should we be worried about either the missing states or the missing counties? Why or why not?

Data

The county demographic information (CDI) dataset is taken from Kutner et al. (2005), which provides 440 most populous counties in the United States. There are total 17 columns and 440 rows in this dataset. Each line of the dataset provides information for a single county. There are no missing values in this dataset. The definition of each variable is given below:

1. id: Identification number 1–440
2. county: County name
3. state: Two-letter state abbreviation
4. land.area: Land area (square miles)
5. pop: Estimated 1990 CDI total population
6. pop.18_34: Percent of 1990 CDI population aged 18–34
7. pop.65_plus: Percent of 1990 CDI population aged 65 or old
8. doctors: Number of professionally active non-federal doctors during 1990
9. hosp.beds: Total number of beds, cribs, and bassinets during 1990
10. 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. pct.hs.grad: Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12. pct.bach.deg: Percent of adult population (persons 25 years old or older) with bachelor's degree
13. pct.below.pov: Percent of 1990 CDI population with income below poverty level
14. pct.unemp: Percent of 1990 CDI population that is unemployed
15. per.cap.income: Per-capita income (i.e. average income per person) of 1990 CDI population (in dollars)
16. tot.income: Total personal income of 1990 CDI population (in millions of dollars)
17. 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)

Variables id and county are unique for each row, which means we can ignore these two variables when doing data analysis. Below are the summary tables for two category variables state and region:

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

Table 1. Summary Table of State

NC	NE	S	W
108	103	152	77

Table 2. Summary Table of Geographic Region

We can find that most counties in this dataset are in CA, FL, MI, NJ, NY, OH, PA and TX. The largest number of counties are in Southern US, and the smallest number of counties are in Western US.

Then, the summary of numerical variables are shown below and the histograms of those numerical variables are shown in Appendix (see Appendix **Figure 1 – Figure 3**).

	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 3. Summary Table of Numerical Variables

From the summary statistics, we can find that there are huge difference between minimal value and maximal value in land area, population, number of doctors, number of hospital beds, number of crimes and total income. From histograms of all numerical variables, we can find that variables land area, population, number of doctors, number of hospital beds, number of crimes, per capital income, and tot income are skewed to the right.

Methods

In order to find the relationship between each variable, we plot the correlation heatmap based on the whole dataset. Since we care more about the per captia income as the response variable, we make scatter plots between per capita income and other numerical variables.

We build three regression models on per-capita income, region, and number of crimes as well as three regression models on per-capita income, region, and (number of crimes)/(population). Then we use ANOVA to get the best two models from each three models, and apply AIC to find the best model between two winners to inspect whether per-capita income is related to crime rate and region.

To find the best model predicting per-capita income, we build all-subsets regression, step-wise AIC regression, and step-wise BIC regression. We first consider the models without regions and then add regions as interaction. After we get all those models, we use AIC, BIC, and adjusted R^2 to compare the model performance and get the best model.

Finally, we make boxplot of per-capita income and population grouped by region and perform exploratory data analysis on those plots. We compare our findings to facts that follow our understanding.

Results

Relationship Between Variables

Since our response variable is average income, we look at the scatter plots (see Appendix **Figure 4.** – **Figure 5.**) of per capita income with other numerical variables. From the scatter plots, land area, population, number of doctors, number of hospital beds, crime, and total incomes look like have no relationship with per capita income. In addition, percentage of population with high school graduation and percentage of population with bachelor degree have a positive linear relationship with per capita income and percentage of population with income below poverty has a negative linear relationship with per capita income.

In order to have a better visualization on the relationship between variables, we make a heatmap for numerical variables. From the correlation heatmap (see Appendix **Figure 6**), we can find that population is highly correlated with total income, number of doctors, number of hospital beds, and number of crimes. That is no surprise since more population result in more total incomes; more population result in more people choosing to be doctors; more hospital beds are needed for more population; and more crimes might occur due to the more population. Also, three variables doctors, number of hospital beds, and total crimes are strongly correlated with one another, which is reasonable because more hospital beds are needed if there exist more crimes and result in more doctors to take care. Per capita

income is positively correlated with percentage of high school graduation, and percentage of population with bachelor degree; is negatively correlated with and percentage of below poverty, and percentage of unemployment. All four of these variables are highly correlated with one another. This is also reasonable since people with higher degree have more chance to be employed and always earn more.

Analysis on Average Income and Crime in Different Region

Before building models, we need to transform the skewed data first. Since logarithms clean up a lot of the skewing in the data, we use log-transform on variables land.area, pop, doctors, hosp.beds, crimes, per.cap.income, and tot.income variables. Then there are three models to think about.

$$\begin{aligned}
\text{log.per.cap.income} = &\ \beta_0 + \beta_1 \text{log.crimes} \\
\text{log.per.cap.income} = &\ \beta_0 + \beta_1 \text{log.crimes} + \beta_2 \text{regionNE} + \beta_3 \text{regionS} + \beta_4 \text{regionW} \\
\text{log.per.cap.income} = &\ \beta_0 + \beta_1 \text{log.crimes} + \beta_2 \text{regionNE} + \beta_3 \text{regionS} + \beta_4 \text{regionW} + \\
&\ \beta_5 \text{log.crimes*regionNE} + \beta_6 \text{log.crimes*regionS} + \beta_7 \text{log.crimes*regionW}
\end{aligned}$$

```
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
  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
```

Table 4. ANOVA on three models

From the ANOVA result, we can find that model

$$\text{log.per.cap.income} = \beta_0 + \beta_1 \text{log.crimes} + \beta_2 \text{regionNE} + \beta_3 \text{regionS} + \beta_4 \text{regionW}$$

is the best among those three models since its p-value $= 1.523e - 13 < 0.05$.

In order to compare this with a model involving per-capita crime, we construct a new variable log.per.cap.crimes, which is equal to log.crimes - log.pop. Once again, there are three models to think about.

$$\begin{aligned}
\text{log.per.cap.income} = &\ \beta_0 + \beta_1 \text{log.per.cap.crimes} \\
\text{log.per.cap.income} = &\ \beta_0 + \beta_1 \text{log.per.cap.crimes} + \beta_2 \text{regionNE} + \beta_3 \text{regionS} + \beta_4 \text{regionW} \\
\text{log.per.cap.income} = &\ \beta_0 + \beta_1 \text{log.per.cap.crimes} + \beta_2 \text{regionNE} + \beta_3 \text{regionS} + \beta_4 \text{regionW} + \\
&\ \beta_5 \text{log.per.cap.crimes*regionNE} + \beta_6 \text{log.per.cap.crimes*regionS} + \\
&\ \beta_7 \text{log.per.cap.crimes*regionW}
\end{aligned}$$

```
Analysis of Variance Table

Model 1: log.per.cap.income ~ log.per.cap.crimes
Model 2: log.per.cap.income ~ log.per.cap.crimes + region
Model 3: log.per.cap.income ~ log.per.cap.crimes * 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
```

Table 5. ANOVA on three models

From the ANOVA result, we can find that model

$$\text{log.per.cap.income} = \beta_0 + \beta_1 \text{log.per.cap.crimes} + \beta_2 \text{regionNE} + \beta_3 \text{regionS} + \beta_4 \text{regionW}$$

is the best among those three models since its p-value $= 3.263e - 09 < 0.05$.

To compare two winners, we use AIC because the two winners are not nested models.

	df <dbl>	AIC <dbl>
q2model2	6	-227.4746
q2model5	6	-172.1347

Table 6. AIC between winner models

From the AIC result, it shows that

$$\text{log.per.cap.income} = \beta_0 + \beta_1 \text{log.crimes} + \beta_2 \text{regionNE} + \beta_3 \text{regionS} + \beta_4 \text{regionW}$$

is the best model since AIC value of this model is smaller. Then let's see the summary of this model to check the feasibility.

```
Call:
lm(formula = log.per.cap.income ~ log.crimes + region, data = cdi_transform)

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
```

Table 7. Summary of model

From the summary statistics of the best model, for every 1% increase in crimes, we expect a 0.067% increase in per capita income, which is a slightly small effect. Different regions of the country have different baseline per-capita incomes. In the north-central region, the baseline salary is $e^{9.188} = 9779.073$; In the northeastern region, it is $e^{9.188+0.104} = 10850.864$; In the southern region, it is $e^{9.188-0.087} = 8964.252$; In the western region, it is $e^{9.188-0.055} = 9255.748$. Based on the model, the level of income varies with region in the US, but is not related to crime.

Best Model Predicting Income per Person

From the Data section, id and county variables are not useful so we decide to drop these two variables. Also, we take variables population and total income out of consideration, since log.per.cap.income = log.tot.income - log.pop, which is a deterministic function of those two variables. Lastly, state and region are two category variables for the location, and region contains states geographically so we decide to drop off state to avoid duplicate information. In addition, since region is a categorical variable, variable selection functions may not deal with that kind of variables well. Then we first analyze the model without the region variable, and try to include the region variable as an interaction term to see the different result.

First, we start with all-subsets regression. Based on the all-subsets plot (Appendix **Figure 7.**), we can find that the best model is with variables percentage of population age from 18 to 34, percentage of population with high school graduation, percentage of population with bachelor degree, percentage of population with income below poverty, percentage of unemployment, log of land area, and log of the number of doctors. Then, let us see the summary statistics first.

```
Call:
lm(formula = log.per.cap.income ~ ., data = tmp)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34147 -0.04886 -0.00538  0.04818  0.26969

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.2224950  0.0931210 109.776  < 2e-16 ***
pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
pct.hs.grad   -0.0044064  0.0010823  -4.071 5.56e-05 ***
pct.bach.deg   0.0153853  0.0009246  16.641  < 2e-16 ***
pct.below.pov -0.0242784  0.0012583 -19.294  < 2e-16 ***
pct.unemp      0.0106037  0.0021771   4.871 1.56e-06 ***
log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
log.doctors    0.0606769  0.0040183  15.100  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.082 on 432 degrees of freedom
Multiple R-squared:  0.8452,    Adjusted R-squared:  0.8427
F-statistic: 336.9 on 7 and 432 DF,  p-value: < 2.2e-16
```

Table 8. Summary of all-subsets regression model without region

From summary statistics of the final model using all.subsets regression method, we can find variables are statistically significant even though the estimated values are kind of small.

Then We apply VIF method to check the multicollinearity (see Appendix **Table 7.**) and plot the diagnostic plots to check the assumption (see Appendix **Figure 8.**). We can find that none of the VIF values seem excessively large, i.e., there is no multicollinearity issue that need to be addressed. From Residuals vs Fitted plot, residuals are randomly distributed around 0. From Q-Q plot, it suggests both the left and the right tails are a bit longer than expected for the normal distribution. From the Scale-Location plot, it has constant variance. From Residuals vs Leverage plot, there is no high influential points that might influence the model performance. To sum up, this model works well.

Next thing is to try to see whether adding interaction term region helps in any way. Based on the rule of thumb: if any indicator for a categorical variable seems important (e.g. a statistically significant coefficient), then keep the whole categorical variable. The summary of the full model with region below suggests that we should keep region, interaction between region and percentage of population with high school graduation, interaction between region and percentage of population with income below poverty, and interaction between region and percentage of unemployment if we choose $\alpha = 0.05$.

```
Call:
lm(formula = log.per.cap.income ~ . * region, data = tmp)

Residuals:
      Min        1Q    Median        3Q       Max
-0.250782 -0.042332 -0.002298  0.040559  0.313570

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           10.1244260  0.2826240  35.823  < 2e-16 ***
pop.18_34             -0.0147940  0.0026043  -5.681 2.55e-08 ***
pct.hs.grad           -0.0024773  0.0034110  -0.726 0.468088
pct.bach.deg           0.0140833  0.0029254   4.814 2.09e-06 ***
pct.below.pov         -0.0237085  0.0036234  -6.543 1.81e-10 ***
pct.unemp              0.0180393  0.0048923   3.687 0.000257 ***
log.land.area         -0.0364187  0.0151355  -2.406 0.016564 *
log.doctors            0.0544169  0.0093221   5.837 1.08e-08 ***
regionNE               0.3243992  0.3577081   0.907 0.365004
regionS               -0.0345856  0.3131668  -0.110 0.912116
regionW                1.5043946  0.4226868   3.559 0.000416 ***
pop.18_34:regionNE    -0.0024780  0.0036873  -0.672 0.501939
pop.18_34:regionS     -0.0008777  0.0030680  -0.286 0.774970
pop.18_34:regionW      0.0014122  0.0040925   0.345 0.730220
pct.hs.grad:regionNE  -0.0037529  0.0044150  -0.850 0.395813
pct.hs.grad:regionS    0.0021198  0.0037853   0.560 0.575790
pct.hs.grad:regionW   -0.0190188  0.0045881  -4.145 4.13e-05 ***
pct.bach.deg:regionNE  0.0069429  0.0040312   1.722 0.085776 .
pct.bach.deg:regionS  -0.0015774  0.0032000  -0.493 0.622328
pct.bach.deg:regionW   0.0071026  0.0036374   1.953 0.051541 .
pct.below.pov:regionNE -0.0014134  0.0050896  -0.278 0.781381
pct.below.pov:regionS  0.0072764  0.0040739   1.786 0.074827 .
pct.below.pov:regionW -0.0161639  0.0054271  -2.978 0.003071 **
pct.unemp:regionNE    -0.0083596  0.0073758  -1.133 0.257720
pct.unemp:regionS     -0.0249396  0.0065867  -3.786 0.000176 ***
pct.unemp:regionW     -0.0201466  0.0067713  -2.975 0.003101 **
log.land.area:regionNE -0.0037179  0.0201435  -0.185 0.853656
log.land.area:regionS -0.0047582  0.0174155  -0.273 0.784825
log.land.area:regionW  0.0151234  0.0181871   0.832 0.406154
log.doctors:regionNE  -0.0046251  0.0132571  -0.349 0.727359
log.doctors:regionS    0.0043337  0.0114401   0.379 0.705019
log.doctors:regionW   -0.0034863  0.0131576  -0.265 0.791173
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0759 on 408 degrees of freedom
Multiple R-squared:  0.8747,    Adjusted R-squared:  0.8652
F-statistic: 91.91 on 31 and 408 DF,  p-value: < 2.2e-16
```

Table 9. Summary of all-subsets regression model with region

After dropping other insignificant variables, we can find all variables are statistically significant except variable percentage of population with high school graduation from summary statistics of the final model considering region as shown below.

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

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 ***
pop.18_34               -0.0149347  0.0010897 -13.705  < 2e-16 ***
pct.hs.grad             -0.0043532  0.0024515  -1.776 0.076501 .
pct.bach.deg             0.0156310  0.0009715  16.090  < 2e-16 ***
pct.below.pov           -0.0252029  0.0032612  -7.728 8.12e-14 ***
pct.unemp                0.0197400  0.0046254   4.268 2.44e-05 ***
log.land.area           -0.0381738  0.0053996  -7.070 6.51e-12 ***
log.doctors              0.0572284  0.0040082  14.278  < 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 ***
pct.hs.grad:regionNE     0.0017684  0.0029293   0.604 0.546374
pct.hs.grad:regionS      0.0011525  0.0025618   0.450 0.653024
pct.hs.grad:regionW     -0.0141473  0.0035826  -3.949 9.20e-05 ***
pct.below.pov:regionNE -0.0015170  0.0046143  -0.329 0.742493
pct.below.pov:regionS    0.0070185  0.0035199   1.994 0.046808 *
pct.below.pov:regionW   -0.0137920  0.0051811  -2.662 0.008066 **
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 ***
---
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
```

Table 10. Summary of all-subsets regression model with some regions

Again, we apply VIF method (see Appendix **Table 8.**) and plot the diagnostic plots (see Appendix **Figure 9.**). We can find that VIF values of percentage of population with high school graduation, percentage of population with below poverty, region, interaction between percentage of population with high school graduation and region, interaction between percentage of population below poverty and region, interaction between percentage of unemployment and region are excessively large, i.e., there is multicollinearity issue that need to be addressed. From Residuals vs Fitted plot, residuals are randomly distributed around 0. From Q-Q plot, it suggests both the left and the right tails are a bit longer than expected for the normal distribution. From the Scale-Location plot, it has constant variance. From Residuals vs Leverage plot, there is no high influential points that might influence the model performance negatively.

Since when we take region as interaction term, it creates collinearities, which may not result in a good prediction. The interpretation of the coefficient is that it provides an estimate of one unit change in an independent variable, holding the other variables constant. If one independent variable is highly correlated with another independent variable, we would

have an imprecise estimate when that independent variable changes. Because of this reason, we choose the model

$$\begin{aligned}\text{log.per.cap.income} = \; & \beta_0 + \beta_1 \text{pop.18\_34} + \beta_2 \text{pct.hs.grad} + \beta_3 \text{pct.bach.deg} + \\ & \beta_4 \text{pct.below.pov} + \beta_5 \text{pct.unemp} + \beta_6 \text{log.land.area} + \beta_7 \text{log.doctors}\end{aligned}$$

Then, we consider stepwise regression using BIC criterion. We can find that the best model is with variables percentage of population age from 18 to 34, percentage of population with high school graduation, percentage of population with bachelor degree, percentage of population with income below poverty, percentage of unemployment, log of land area, and log of the number of doctors. Below is the summary of stepwise BIC model

```
Call:
lm(formula = log.per.cap.income ~ ., data = tmp)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34147 -0.04886 -0.00538  0.04818  0.26969

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.2224950  0.0931210 109.776  < 2e-16 ***
pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
pct.hs.grad   -0.0044064  0.0010823  -4.071 5.56e-05 ***
pct.bach.deg   0.0153853  0.0009246  16.641  < 2e-16 ***
pct.below.pov -0.0242784  0.0012583 -19.294  < 2e-16 ***
pct.unemp      0.0106037  0.0021771   4.871 1.56e-06 ***
log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
log.doctors    0.0606769  0.0040183  15.100  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.082 on 432 degrees of freedom
Multiple R-squared:  0.8452,    Adjusted R-squared:  0.8427
F-statistic: 336.9 on 7 and 432 DF,  p-value: < 2.2e-16
```

Table 11. Summary of stepwise BIC model

We can see that stepwise regression using the BIC criterion actually find exactly same as the all-subsets regression model, so we skip the VIF and diagnostic plots which will be the same as the all subsets regression model (we still do the code analysis (see Appendix **Table 9.** and **Figure 10.**)). Also, if we add the interaction term region, the result will also be the same as the all-subsets regression model with region. By using BIC stepwise regression model, we still get the final model

$$\begin{aligned}\text{log.per.cap.income} = \; & \beta_0 + \beta_1 \text{pop.18\_34} + \beta_2 \text{pct.hs.grad} + \beta_3 \text{pct.bach.deg} + \\ & \beta_4 \text{pct.below.pov} + \beta_5 \text{pct.unemp} + \beta_6 \text{log.land.area} + \beta_7 \text{log.doctors}\end{aligned}$$

Then, we use the AIC criterion stepwise regression. We can find that the best model is with variables percentage of population age from 18 to 34, percentage of population age equal or above 65, percentage of population with high school graduation, percentage of population with bachelor degree, percentage of population below the poverty, percentage of unemployment, log of land area, and log of the number of doctors. Below are the summary statistics of the final model using AIC stepwise.

```
Call:
lm(formula = log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
    pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
    log.doctors, data = cdi_nonregion)

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 ***
pop.18_34     -0.0153488  0.0012988 -11.818  < 2e-16 ***
pop.65_plus   -0.0027664  0.0012978  -2.132   0.0336 *
pct.hs.grad   -0.0046579  0.0010843  -4.296 2.15e-05 ***
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 ***
log.land.area -0.0364935  0.0047728  -7.646 1.36e-13 ***
log.doctors    0.0626053  0.0041029  15.259  < 2e-16 ***
---
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
```

Table 12. Summary of stepwise AIC model without region

From the summary statistics, We can find variables are statistically significant. In order to further decide whether it is a good model, we apply VIF method (see Appendix **Table 10**) to check the multicollinearity and plot the diagnostic plots (see Appendix **Figure 11**) to check the assumptions. None of the VIF values seem excessively large, i.e., there is no multicollinearity issue that need to be addressed. From Residuals vs Fitted plot, residuals are randomly distributed around 0. From Q-Q plot, it suggests both the left tails and right tails are a bit longer than expected for the normal distribution. From the Scale-Location plot, it has constant variance. From Residuals vs Leverage plot, there is no high influential points that might influence the model performance.

Same as the procedure as in the all-subsets regression method, next thing is to see if interaction with region helps in any way. Based on the rule of thumb, the summary of the full model with interaction terms below suggests that we should keep region, interaction between percentage of population with high school graduation and region, percentage of population with income below poverty and region, and percentage of unemployment and region if we choose $\alpha = 0.05$.

```
Call:
lm(formula = log.per.cap.income ~ . * region, data = tmp)

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 ***
pop.18_34               -0.0150740  0.0028317  -5.323 1.69e-07 ***
pop.65_plus             -0.0012483  0.0050165  -0.249 0.803614
pct.hs.grad             -0.0026649  0.0034861  -0.764 0.445055
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 ***
log.land.area           -0.0355230  0.0155258  -2.288 0.022654 *
log.doctors              0.0548293  0.0094485   5.803 1.32e-08 ***
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 **
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
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 ***
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 **
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
---
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
```

Table 13. Summary of stepwise AIC model with region

After dropping other insignificant variables, from summary of the final model below considering region using AIC stepwise regression method, we can find all variables are statistically significant except variable percentage of population with high school graduation.

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

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 ***
pop.18_34              -0.0149347  0.0010897 -13.705  < 2e-16 ***
pct.hs.grad            -0.0043532  0.0024515  -1.776 0.076501 .
pct.bach.deg            0.0156310  0.0009715  16.090  < 2e-16 ***
pct.below.pov          -0.0252029  0.0032612  -7.728 8.12e-14 ***
pct.unemp               0.0197400  0.0046254   4.268 2.44e-05 ***
log.land.area          -0.0381738  0.0053996  -7.070 6.51e-12 ***
log.doctors             0.0572284  0.0040082  14.278  < 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 ***
pct.hs.grad:regionNE    0.0017684  0.0029293   0.604 0.546374
pct.hs.grad:regionS     0.0011525  0.0025618   0.450 0.653024
pct.hs.grad:regionW    -0.0141473  0.0035826  -3.949 9.20e-05 ***
pct.below.pov:regionNE -0.0015170  0.0046143  -0.329 0.742493
pct.below.pov:regionS   0.0070185  0.0035199   1.994 0.046808 *
pct.below.pov:regionW  -0.0137920  0.0051811  -2.662 0.008066 **
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 ***
---
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
```

Table 14. Summary of stepwise AIC model with some regions

We again apply VIF method to check the multicollinearity (see Appendix **Table 11.**) and plot the diagnostic plots (see Appendix **Figure 12.**). We can find that VIF values of percentage of population with high school graduation, percentage of population with income below poverty, region, interaction between percentage of population with high school graduation and region, interaction between percentage of population with income below poverty and region, interaction between percentage of unemployment and region are excessively large, i.e., there is multicollinearity issue that need to be addressed. From Residuals vs Fitted plot, residuals are randomly distributed around 0. From Q-Q plot, it suggests both the left and the right tails are a bit longer than expected for the normal distribution. From the Scale-Location plot, it has constant variance. From Residuals vs Leverage plot, there is no high influential points that might influence the model performance negatively.

Because of the collinearity issue as mentioned in all-subsets regression model, we decide our final model without region using AIC criterion

$$\begin{aligned}\text{log.per.cap.income} = \; & \beta_0 + \beta_1 \text{pop.18\_34} + \beta_2 \text{pop.65\_plus} + \beta_3 \text{pct.hs.grad} + \\ & \beta_4 \text{pct.bach.deg} + \beta_5 \text{pct.below.pov} + \beta_6 \text{pct.unemp} + \\ & \beta_7 \text{log.land.area} + \beta_8 \text{log.doctors}\end{aligned}$$

In order to find the best model using these three methods, we compare AIC, BIC, and adjusted R^2. From the comparison table below,

	AIC <dbl>	BIC <dbl>	adjusted_R2 <dbl>
all.subsets.model	-942.2740	-905.4931	0.8426532
BIC.model	-942.2740	-905.4931	0.8426532
AIC.model	-944.8883	-904.0206	0.8439334

Table 15. Comparison

AIC stepwise model has lower AIC and higher adjusted R^2, so we decide our final model as

$$\begin{aligned}\text{log.per.cap.income} = \; & 10.316 - 0.015\text{pop.18\_34} - 0.002\text{pop.65\_plus} - 0.005\text{pct.hs.grad} + \\ & 0.015\text{pct.bach.deg} - 0.025\text{pct.below.pov} + 0.011\text{pct.unemp} - \\ & 0.036\text{log.land.area} + 0.063\text{log.doctors}\end{aligned}$$

For every 1% increase in percentage of population age from 18 to 34, there is a $e^{0.015}\%$ decrease in per capita income. For every 1% increase in percentage of population age 65 or above, there is a $e^{0.002}\%$ decrease in per capita income. For every 1% increase in percentage of population with high school graduation, there is a $e^{0.005}\%$ decrease in per capita income. For every 1% increase in percentage of population with bachelor degree, there is a $e^{0.015}\%$ increase in per capita income. For every 1% increase in percentage of population with income below poverty, there is a $e^{0.025}\%$ decrease in per capita income. For every 1% increase in percentage of unemployment, there is a $e^{0.011}\%$ increase in per capita income. For every 1 unit increase in land area, there is a 0.036 unit decrease in per capita income. For every 1 unit increase in number of doctors, there is a 0.063 unit increase in per capita income.

Analysis on Missing Counties

Based on these 440 counties, if these counties can represent the around 3000 counties in United States, then we do not need to worry about the missing counties. However, since the counties are 440 of the most populous counties in the United States, it is not randomly sampled, which might cause the bias. In order to check the feasibility, we plot two boxplots (see Appendix **Figure 13.** and **Figure 14.**), which are per capital income in different region and population in different regions (remove outliers population > 4000000 to have a better visualization). We can find that the median of per capital income in northeastern region is the highest, and the medians of other three regions are close. For each region's per capita income, there are some outliers. AS for the population, medians of northeastern region and western region are close and are higher than medians of north-central region and southern region. From the summary table (see Appendix **Table 13.**), the median of per capita income is reasonable since northeastern US are economically developed area. However, the sum of population in northeastern region is higher than that in north-central region. According to Population Change and Distribution by Marc J. Perry and Paul J. Mackun (2001), we can find that Midwestern US (Prior to June 1984, the Midwest Region was designated as the North Central Region) has the second largest population, , which conflicts the data that northeastern region has the lowest population. Thus, we might worry about the missing counties.

Discussion

According to the results, per capita income is not highly correlated with other variables except total income and population. If we only keep per capita income, total crimes, and region, we can find that there is no strong relationship between per capita income and crimes, but per capita income varies in different regions. Our final model predicting per capita income contains variables percentage of population age from 18 to 34, percentage of population age equal or above 65, percentage of population with high school degree, percentage of population with bachelor degree, percentage of population below poverty percentage of unemployment, land area, and number of doctors.

Even though the performance of model is well, we still can make some improvements. If we have no time limitation, we are going to use LASSO and Ridge regression with cross-validation when analyzing the dataset. In this way, we might be able to distinguish which model is the best model better, at least in terms of prediction error. Also, one more weakness is that there are only 440 counties in our dataset, which are 1/9 of all counties in US. It might be biased since we use this dataset to build a model predicting the per capital income in US. It would be an improvement if we spend more time researching on other counties. Taking other counties into consideration will help us address the issues much better, especially the last research question.

Reference

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

Marc, J.P. & Paul J.M. (2001) Population Change and Distribution, U.S. Department of Commerce.

Appendix

```
library(leaps)
library(car)
library(MASS)
library(dplyr)
library(corrplot)
```

Data

```
cdi <- read.table("~/Desktop/CMU/36-617_Applied_Linear_Models/Project1/cdi.dat")
cdi_num <- cdi[,-c(1:3,17)]
cdi_cat <- cdi[,c(1:3,17)]
```

First, we check whether there are missing values in this dataset.

```
check_na = function(i){
  n = sum(cdi[,i] == "NA")
  n
}
sapply(1:length(cdi), check_na)
```

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Then, we see the summary statistics of numerical variables and categorical variables, as well as make histograms of numerical variables to visualize the data.

```
summary_statistics <- apply(cdi_num, 2, summary)
t(as.data.frame(summary_statistics))
```

```
##                   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 1. Summary statistics of numerical variables

```
apply(cdi_cat[,-c(1,2)], 2, table)
```

```
## $state
##
## AL AR AZ CA CO CT DC DE FL GA HI ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC
##  7  2  5 34  9  8  1  2 29  9  3  1 17 14  4  3  9 11 10  5 18  7  8  3  1 18
## ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV
##  1  3  4 18  2  2 22 24  4  6 29  3 11  1  8 28  4  9  1 10 11  1
##
## $region
##
##  NC  NE   S   W
## 108 103 152  77
```

Table 2. Summary statistics of categorical variables

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

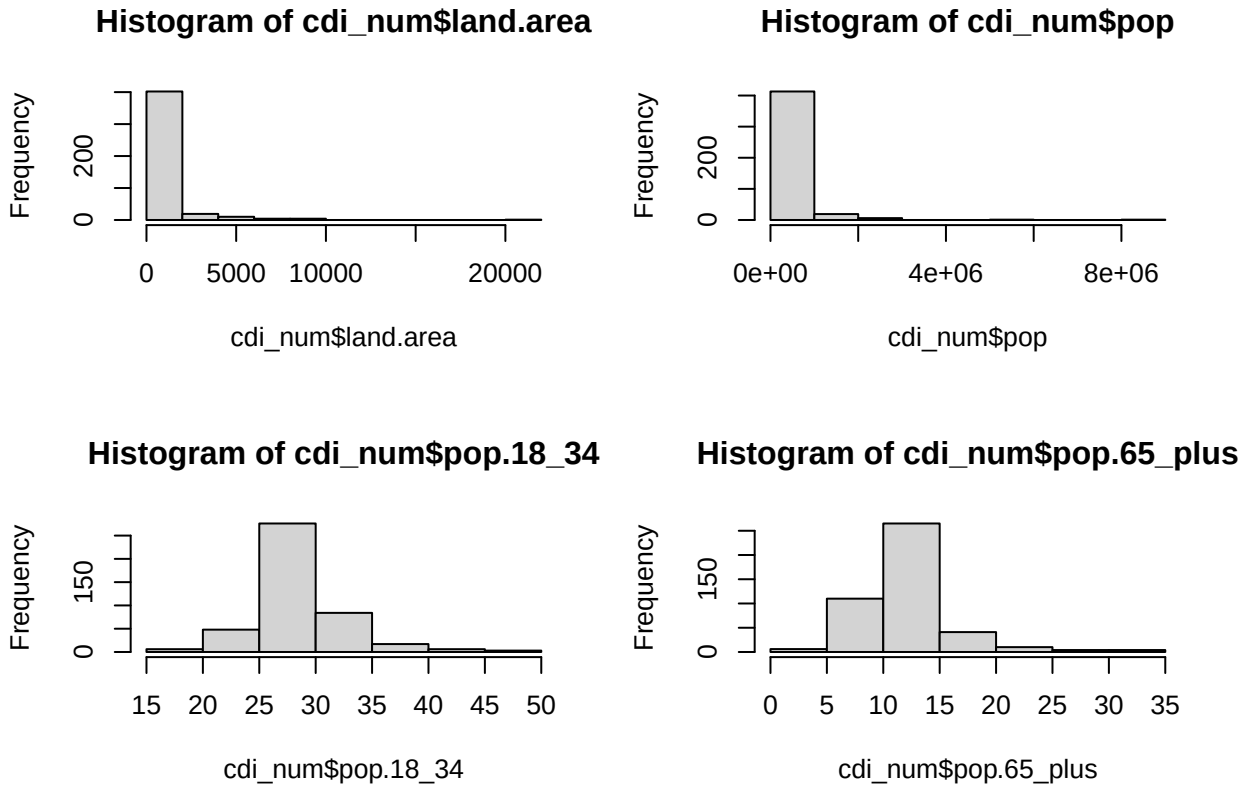

Figure 1. Histograms of numerical variables

```
par(mfrow = c(2,2))
hist(cdi_num$doctors)
hist(cdi_num$hosp.beds)
hist(cdi_num$crimes)
hist(cdi_num$pct.hs.grad)
```

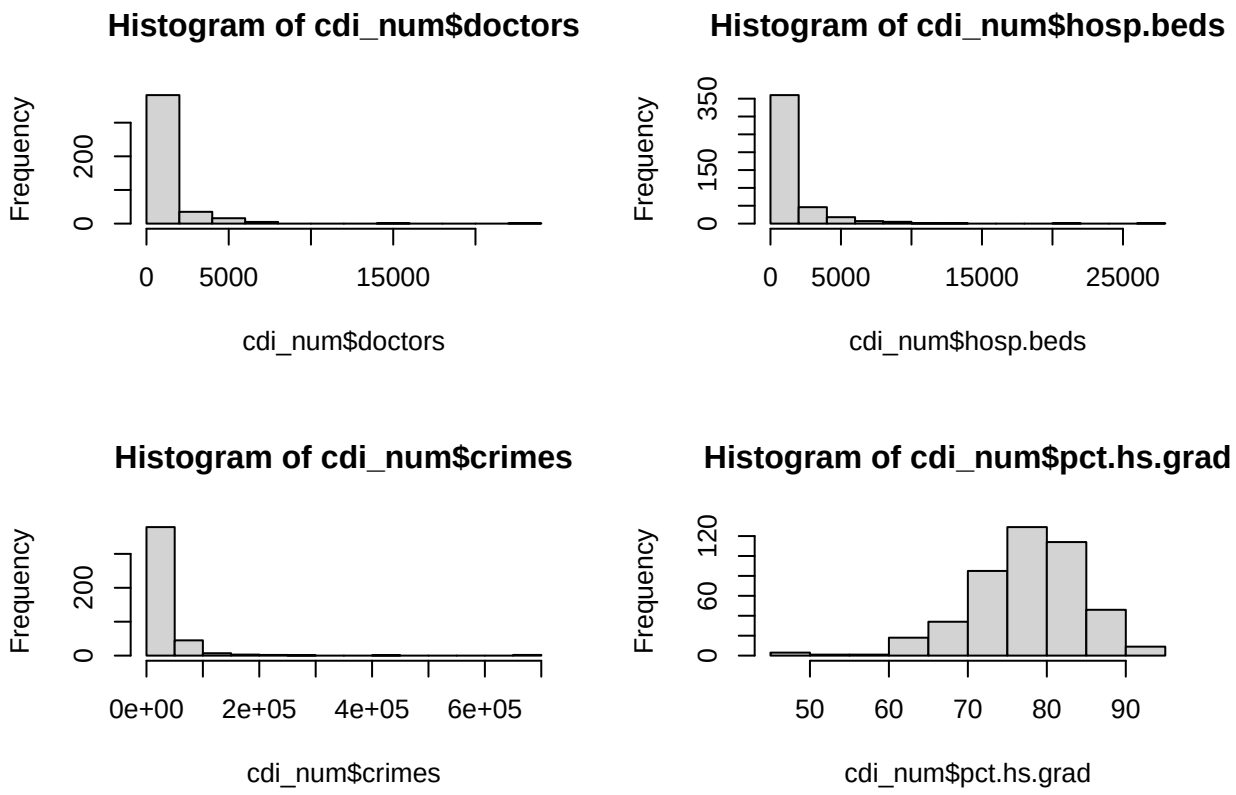

Figure 2. Histograms of numerical variables

```
par(mfrow = c(2,3))
hist(cdi_num$pct.bach.deg)
hist(cdi_num$pct.below.pov)
hist(cdi_num$pct.unemp)
hist(cdi_num$per.cap.income)
hist(cdi_num$tot.income)
```

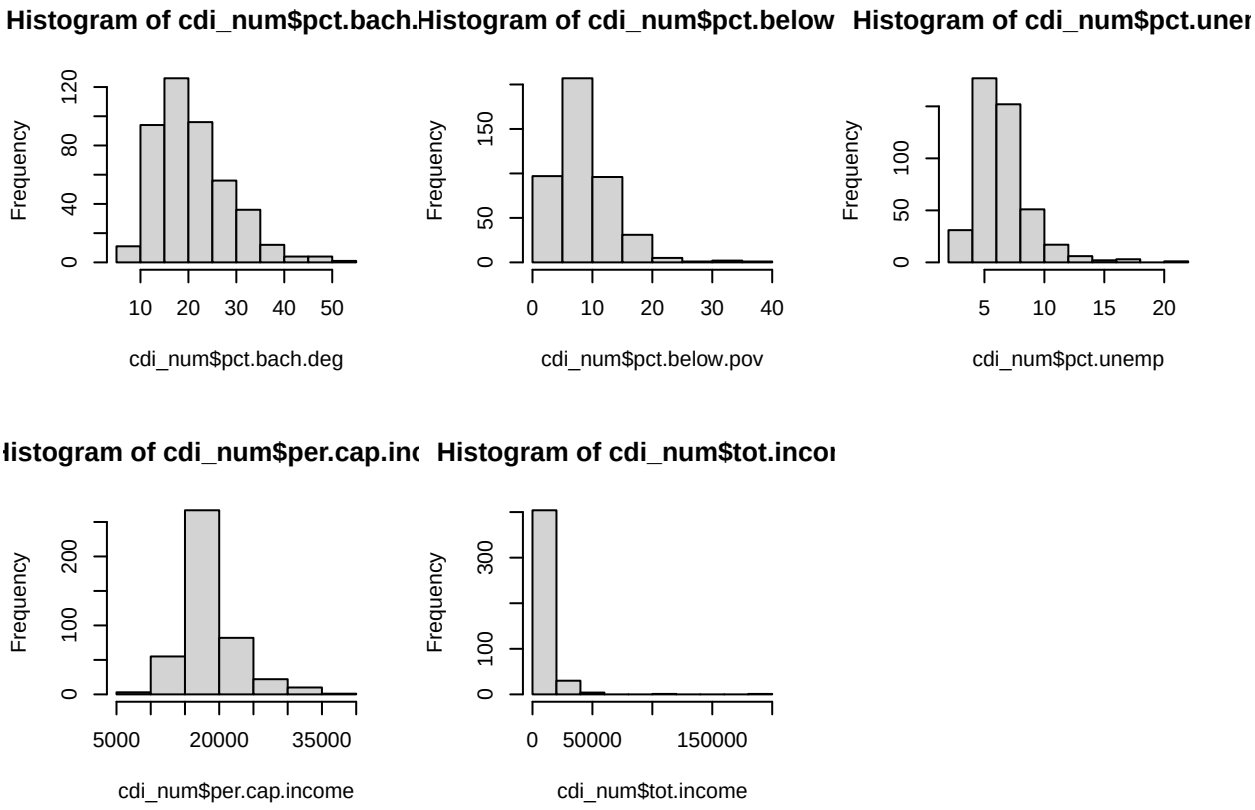

Figure 3. Histograms of numerical variables

Results

Question 1

Below are the scatter plots of per capita income and other numerical variables.

```
par(mfrow = c(2,3))
plot(cdi_num$land.area, cdi_num$per.cap.income)
plot(cdi_num$pop, cdi_num$per.cap.income)
plot(cdi_num$pop.18_34, cdi_num$per.cap.income)
plot(cdi_num$pop.65_plus, cdi_num$per.cap.income)
plot(cdi_num$doctors, cdi_num$per.cap.income)
plot(cdi_num$hosp.beds, cdi_num$per.cap.income)
```

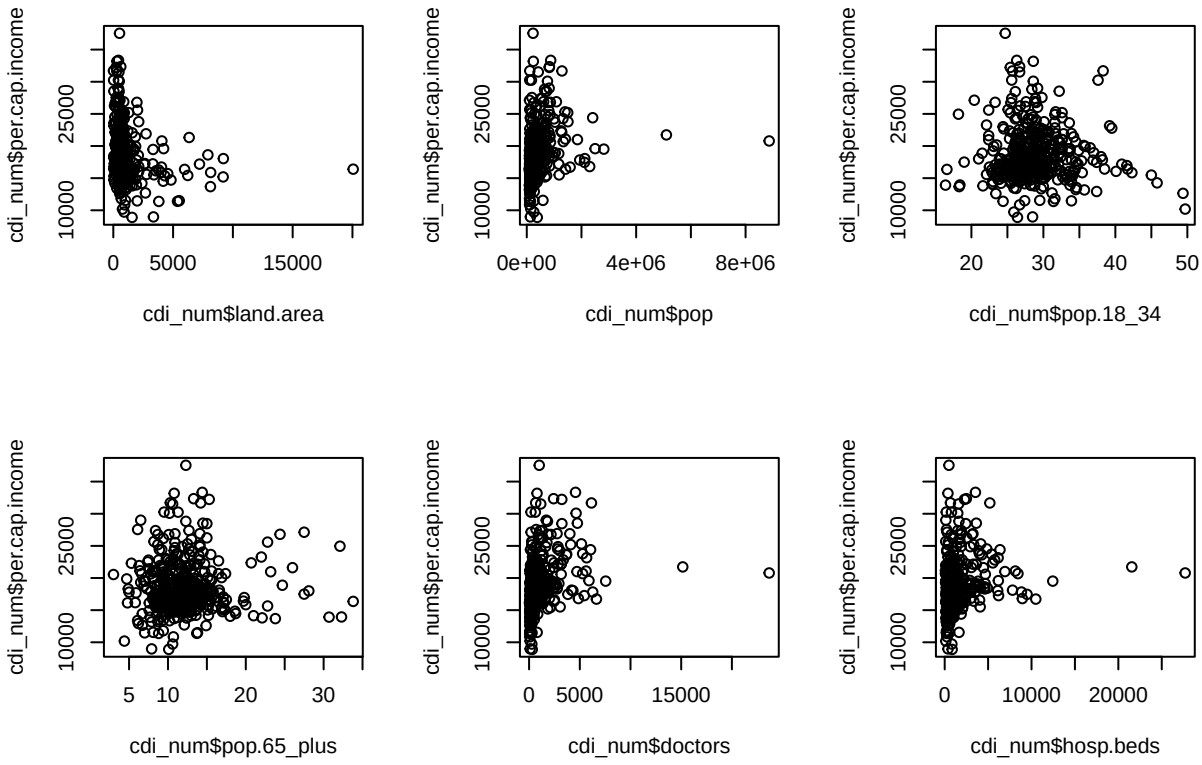

Figure 4. Scatter plots between per capita income and other numerical varibles

```
par(mfrow = c(2,3))
plot(cdi_num$crimes, cdi_num$per.cap.income)
plot(cdi_num$pct.hs.grad, cdi_num$per.cap.income)
plot(cdi_num$pct.bach.deg, cdi_num$per.cap.income)
plot(cdi_num$pct.below.pov, cdi_num$per.cap.income)
plot(cdi_num$pct.unemp, cdi_num$per.cap.income)
plot(cdi_num$tot.income, cdi_num$per.cap.income)
```

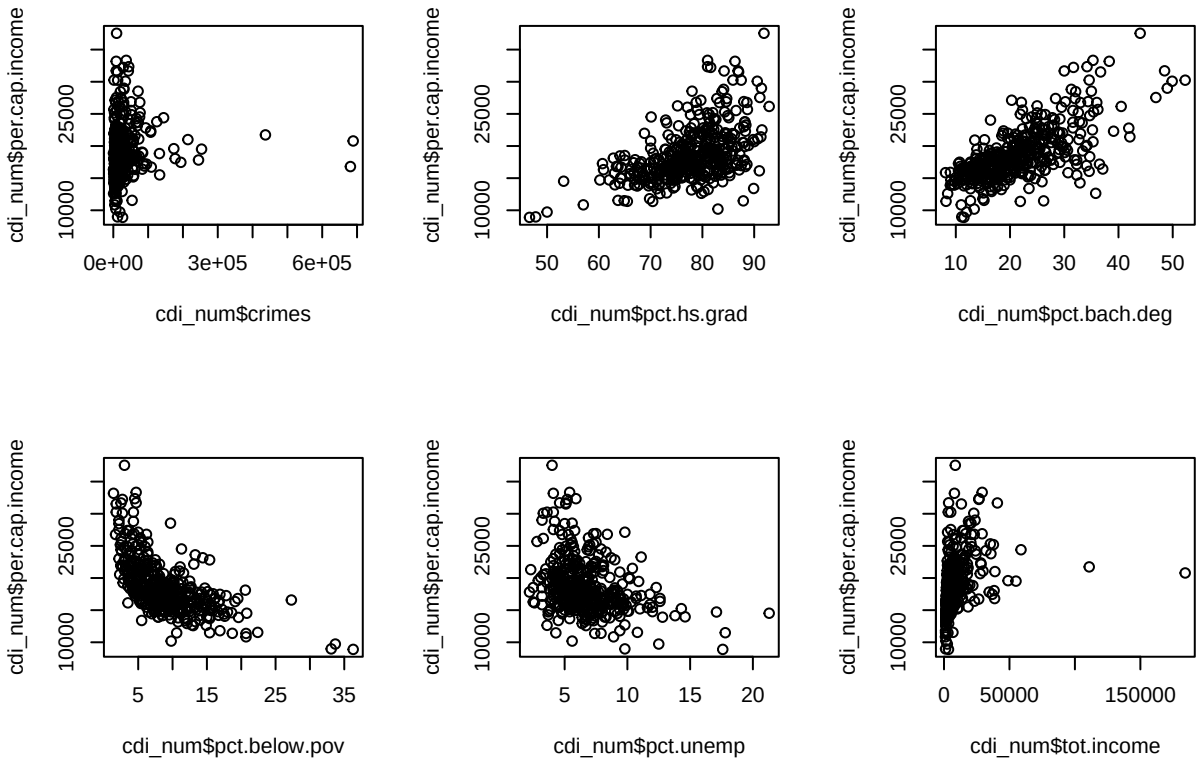

Figure 5. Scatter plots between per capita income and other numerical varibles

Here is the plot the correlation matrix.

```
corrplot(cor(cdi_num), method = "color", tl.col="black")
```

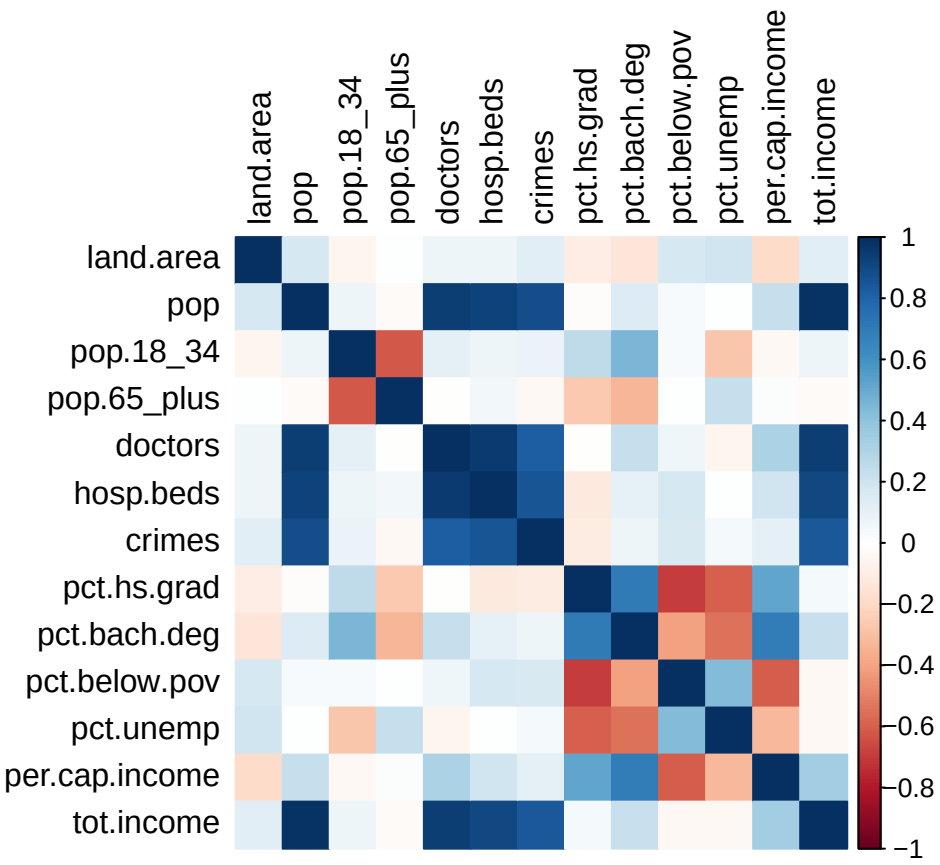

Figure 6. Correlation heatmap

Question 2

From **Figure 1-3.**, we can find that several variables land.area, pop, doctors, hosp.beds, crimes, per.cap.income, and tot.income variables are skewed to the right, so we take log transform on those variables and rename the column name.

```
cdi_transform <- cdi
skewed.vars <- c(4,5,8,9,10,15,16)

for (i in skewed.vars){
  cdi_transform[,i] <- log(cdi_transform[,i])
}
```

```
newname = paste("log.", names(cdi_transform[skewed.vars]), sep = "")
cdi_transform[newname] = cdi_transform[,skewed.vars]
cdi_transform = cdi_transform[,-skewed.vars]
```

Then we build three models predicting the log per capital income on total crime number and region and three models predicting the log per capital income on per capital crime (total crime number/population) and region.

```r
q2model1 <- lm(log.per.cap.income ~ log.crimes, data = cdi_transform)
q2model2 <- lm(log.per.cap.income ~ log.crimes + region, data = cdi_transform)
q2model3 <- lm(log.per.cap.income ~ log.crimes * region, data = cdi_transform)
anova(q2model1, q2model2, q2model3)
```

```
## 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
##   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
```

Table 3. ANOVA of models with total crimes and region

```r
cdi_transform["log.per.cap.crimes"] <- cdi_transform$log.crimes - cdi_transform$log.pop
q2model4 <- lm(log.per.cap.income ~ log.per.cap.crimes, data = cdi_transform)
q2model5 <- lm(log.per.cap.income ~ log.per.cap.crimes + region, data = cdi_transform)
q2model6 <- lm(log.per.cap.income ~ log.per.cap.crimes * region, data = cdi_transform)
anova(q2model4, q2model5, q2model6)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.per.cap.crimes
## Model 2: log.per.cap.income ~ log.per.cap.crimes + region
## Model 3: log.per.cap.income ~ log.per.cap.crimes * 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
```

Table 4. ANOVA of models with per capita crimes and region

```r
AIC(q2model2, q2model5)
```

```
##          df       AIC
## q2model2  6 -227.4746
## q2model5  6 -172.1347
```

Table 5. AIC of two models

Base on the ANOVA result, we find two winners and apply AIC to find that log.per.cap.income $= \beta_0 + \beta_1$log.crimes $+ \beta_2$regionNE $+\beta_3$regionS $+ \beta_4$regionW is the best model. Here is the summary of this model.

```
summary(q2model2)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.crimes + region, data = cdi_transform)
##
## 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
```

Summary 1. Summary of model

Question 3

In order to find the best model, we first drop useless variables id, county, log.tot.income, log.pop, and state. We first analyze the model without the region variable

```
cdi_nonregion <- cdi_transform[, -which(names(cdi_transform) %in% c("id", "county", "state", "log.pop",
cdi_region <- cdi_transform[, -which(names(cdi_transform) %in% c("id", "county", "state", "log.pop", "l
```

Apply all-subsets regression

```
all.subsets <- regsubsets(log.per.cap.income ~ ., cdi_nonregion, nvmax = 10)
plot(all.subsets)
```

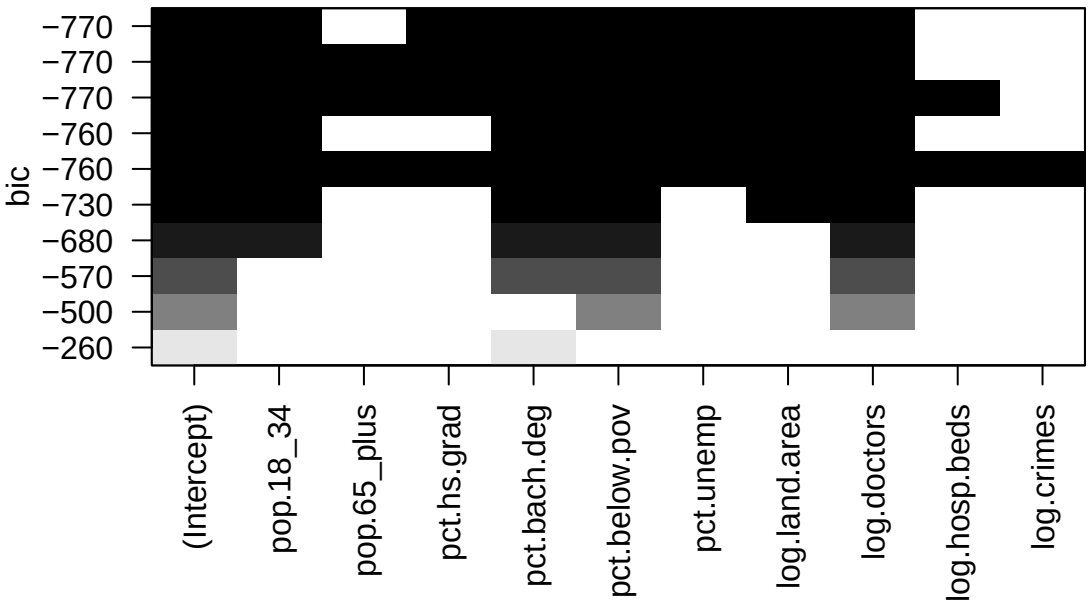

Figure 7. All-subsets plots

Let's see the coefficients of the remaining variables.

```
all.subsets.summary <- summary(all.subsets)
best.model <- which.min(all.subsets.summary$bic)
coef(all.subsets, best.model)
```

```
##   (Intercept)     pop.18_34   pct.hs.grad  pct.bach.deg pct.below.pov
##  10.222495041  -0.013900201  -0.004406396   0.015385301  -0.024278371
##     pct.unemp log.land.area   log.doctors
##   0.010603691  -0.035674062   0.060676872
```

Table 6. Coefficients

```
tmp <- cdi_nonregion[,all.subsets.summary$which[best.model,][-1]]
tmp["log.per.cap.income"] <- cdi_nonregion["log.per.cap.income"]
all.subsets.model <- lm(log.per.cap.income ~ ., data=tmp)
summary(all.subsets.model)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ ., data = tmp)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34147 -0.04886 -0.00538  0.04818  0.26969
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.2224950  0.0931210 109.776  < 2e-16 ***
## pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
## pct.hs.grad   -0.0044064  0.0010823  -4.071 5.56e-05 ***
## pct.bach.deg   0.0153853  0.0009246  16.641  < 2e-16 ***
## pct.below.pov -0.0242784  0.0012583 -19.294  < 2e-16 ***
## pct.unemp      0.0106037  0.0021771   4.871 1.56e-06 ***
## log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
## log.doctors    0.0606769  0.0040183  15.100  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.082 on 432 degrees of freedom
## Multiple R-squared:  0.8452, Adjusted R-squared:  0.8427
## F-statistic: 336.9 on 7 and 432 DF,  p-value: < 2.2e-16
```

Summary 2. Summary of all-subsets model without region

```
vif(all.subsets.model)
```

```
##     pop.18_34   pct.hs.grad  pct.bach.deg pct.below.pov     pct.unemp
##      1.416145      3.763103      3.269565      2.241555      1.691280
## log.land.area   log.doctors
##      1.131867      1.379671
```

Table 7. VIF of all-subsets model without region

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

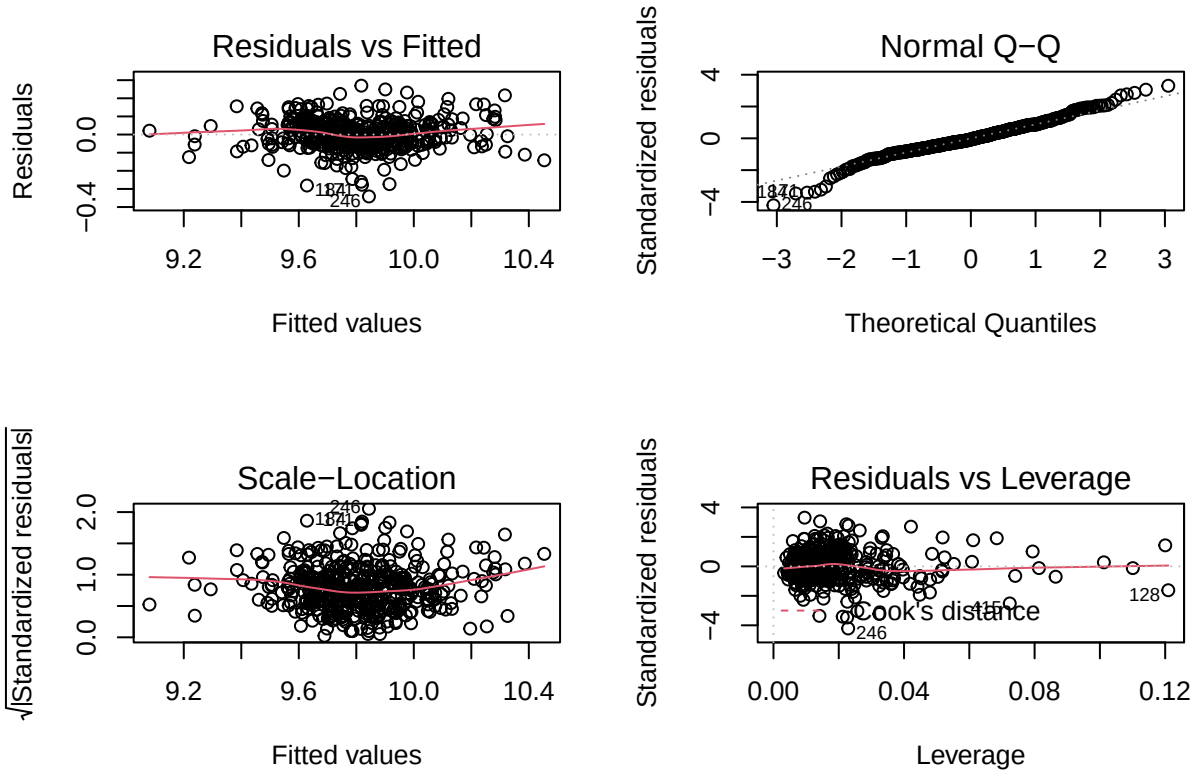

Figure 8. Diagnostic plots of all-subsets model without region

To see if interaction with region helps in any way.

```
tmp <- cbind(tmp, region=cdi_transform$region)
all.subsets.model.with.region <- lm(log.per.cap.income ~ .*region,data=tmp)
summary(all.subsets.model.with.region)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ . * region, data = tmp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.250782 -0.042332 -0.002298  0.040559  0.313570
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.1244260  0.2826240  35.823  < 2e-16 ***
## pop.18_34              -0.0147940  0.0026043  -5.681 2.55e-08 ***
## pct.hs.grad            -0.0024773  0.0034110  -0.726 0.468088
## pct.bach.deg            0.0140833  0.0029254   4.814 2.09e-06 ***
## pct.below.pov          -0.0237085  0.0036234  -6.543 1.81e-10 ***
## pct.unemp               0.0180393  0.0048923   3.687 0.000257 ***
## log.land.area          -0.0364187  0.0151355  -2.406 0.016564 *
```

```
## log.doctors             0.0544169  0.0093221   5.837 1.08e-08 ***
## regionNE                0.3243992  0.3577081   0.907 0.365004
## regionS                -0.0345856  0.3131668  -0.110 0.912116
## regionW                 1.5043946  0.4226868   3.559 0.000416 ***
## pop.18_34:regionNE     -0.0024780  0.0036873  -0.672 0.501939
## pop.18_34:regionS      -0.0008777  0.0030680  -0.286 0.774970
## pop.18_34:regionW       0.0014122  0.0040925   0.345 0.730220
## pct.hs.grad:regionNE   -0.0037529  0.0044150  -0.850 0.395813
## pct.hs.grad:regionS     0.0021198  0.0037853   0.560 0.575790
## pct.hs.grad:regionW    -0.0190188  0.0045881  -4.145 4.13e-05 ***
## pct.bach.deg:regionNE   0.0069429  0.0040312   1.722 0.085776 .
## pct.bach.deg:regionS   -0.0015774  0.0032000  -0.493 0.622328
## pct.bach.deg:regionW    0.0071026  0.0036374   1.953 0.051541 .
## pct.below.pov:regionNE -0.0014134  0.0050896  -0.278 0.781381
## pct.below.pov:regionS   0.0072764  0.0040739   1.786 0.074827 .
## pct.below.pov:regionW  -0.0161639  0.0054271  -2.978 0.003071 **
## pct.unemp:regionNE     -0.0083596  0.0073758  -1.133 0.257720
## pct.unemp:regionS      -0.0249396  0.0065867  -3.786 0.000176 ***
## pct.unemp:regionW      -0.0201466  0.0067713  -2.975 0.003101 **
## log.land.area:regionNE -0.0037179  0.0201435  -0.185 0.853656
## log.land.area:regionS  -0.0047582  0.0174155  -0.273 0.784825
## log.land.area:regionW   0.0151234  0.0181871   0.832 0.406154
## log.doctors:regionNE   -0.0046251  0.0132571  -0.349 0.727359
## log.doctors:regionS     0.0043337  0.0114401   0.379 0.705019
## log.doctors:regionW    -0.0034863  0.0131576  -0.265 0.791173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0759 on 408 degrees of freedom
## Multiple R-squared:  0.8747, Adjusted R-squared:  0.8652
## F-statistic: 91.91 on 31 and 408 DF,  p-value: < 2.2e-16
```

Summary 3. Summary of all-subsets model with region

Based on the rule of thumb: if any indicator for a categorical variable seems important (e.g. a statistically significant coefficient), then keep the whole categorical variable.

```
all.subsets.model.with.some.region <- update(all.subsets.model.with.region,
. ~ . - region:log.land.area - region:pop.18_34 - region:log.doctors - region:pct.bach.deg)
summary(all.subsets.model.with.some.region)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log.land.area + log.doctors +
##     region + pct.hs.grad:region + pct.below.pov:region + pct.unemp:region,
##     data = tmp)
##
## 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 ***
## pop.18_34                -0.0149347  0.0010897 -13.705  < 2e-16 ***
## pct.hs.grad              -0.0043532  0.0024515  -1.776 0.076501 .
## pct.bach.deg              0.0156310  0.0009715  16.090  < 2e-16 ***
## pct.below.pov            -0.0252029  0.0032612  -7.728 8.12e-14 ***
## pct.unemp                 0.0197400  0.0046254   4.268 2.44e-05 ***
## log.land.area            -0.0381738  0.0053996  -7.070 6.51e-12 ***
## log.doctors               0.0572284  0.0040082  14.278  < 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 ***
## pct.hs.grad:regionNE      0.0017684  0.0029293   0.604 0.546374
## pct.hs.grad:regionS       0.0011525  0.0025618   0.450 0.653024
## pct.hs.grad:regionW      -0.0141473  0.0035826  -3.949 9.20e-05 ***
## pct.below.pov:regionNE -0.0015170  0.0046143  -0.329 0.742493
## pct.below.pov:regionS     0.0070185  0.0035199   1.994 0.046808 *
## pct.below.pov:regionW    -0.0137920  0.0051811  -2.662 0.008066 **
## 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 ***
## ---
## 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
```

Summary 4. Summary of all-subsets model with some regions

```
vif(all.subsets.model.with.some.region)
```

```
##                               GVIF Df GVIF^(1/(2*Df))
## pop.18_34             1.547481e+00  1        1.243978
## pct.hs.grad           2.194177e+01  1        4.684205
## pct.bach.deg          4.102307e+00  1        2.025415
## pct.below.pov         1.710982e+01  1        4.136402
## pct.unemp             8.675528e+00  1        2.945425
## log.land.area         1.643605e+00  1        1.282032
## log.doctors           1.559981e+00  1        1.248992
## region                2.454546e+08  3       25.022374
## pct.hs.grad:region    8.506975e+07  3       20.971486
## pct.below.pov:region  5.278685e+03  3        4.172736
## pct.unemp:region      1.108865e+04  3        4.722222
```

Table 8. VIF of all-subsets model with some regions

```
par(mfrow=c(2,2))
plot(all.subsets.model.with.some.region)
```

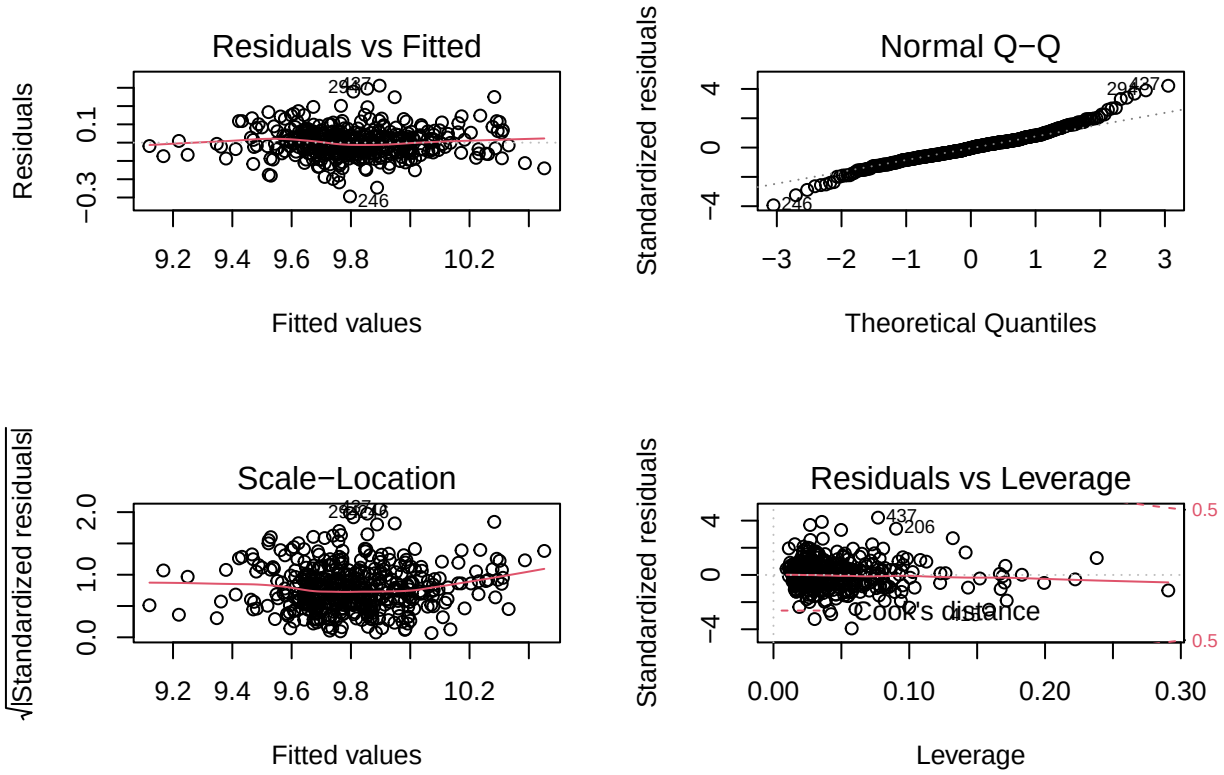

Figure 9. Diagnostic plots of all-subsets model with some region

Apply stepwise BIC regression

```
BIC.model <- stepAIC(lm(log.per.cap.income ~ .,data = cdi_nonregion), direction = "both", k = log(dim(cd
```

```
## Start:  AIC=-2148.59
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##     log.doctors + log.hosp.beds + log.crimes
##
##                 Df Sum of Sq    RSS     AIC
## - log.crimes     1   0.00180 2.8636 -2154.4
## - log.hosp.beds  1   0.01216 2.8740 -2152.8
## - pop.65_plus    1   0.03884 2.9006 -2148.7
## <none>                       2.8618 -2148.6
## - log.doctors    1   0.11565 2.9775 -2137.2
## - pct.hs.grad    1   0.12699 2.9888 -2135.6
## - pct.unemp      1   0.17289 3.0347 -2128.9
## - log.land.area  1   0.36392 3.2257 -2102.0
## - pop.18_34      1   0.94423 3.8060 -2029.2
## - pct.bach.deg   1   1.56251 4.4243 -1963.0
## - pct.below.pov  1   2.44318 5.3050 -1883.1
```

```
##
## Step:  AIC=-2154.4
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##     log.doctors + log.hosp.beds
##
##                 Df Sum of Sq    RSS     AIC
## - log.hosp.beds  1   0.01116 2.8748 -2158.8
## - pop.65_plus    1   0.03709 2.9007 -2154.8
## <none>                       2.8636 -2154.4
## + log.crimes     1   0.00180 2.8618 -2148.6
## - pct.hs.grad    1   0.12662 2.9902 -2141.4
## - log.doctors    1   0.12889 2.9925 -2141.1
## - pct.unemp      1   0.17123 3.0348 -2134.9
## - log.land.area  1   0.37492 3.2385 -2106.3
## - pop.18_34      1   0.94270 3.8063 -2035.3
## - pct.bach.deg   1   1.59514 4.4587 -1965.7
## - pct.below.pov  1   2.47345 5.3371 -1886.5
##
## Step:  AIC=-2158.77
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##     log.doctors
##
##                 Df Sum of Sq    RSS     AIC
## - pop.65_plus    1   0.03031 2.9051 -2160.2
## <none>                       2.8748 -2158.8
## + log.hosp.beds  1   0.01116 2.8636 -2154.4
## + log.crimes     1   0.00079 2.8740 -2152.8
## - pct.hs.grad    1   0.12309 2.9978 -2146.4
## - pct.unemp      1   0.16432 3.0391 -2140.4
## - log.land.area  1   0.38995 3.2647 -2108.9
## - pop.18_34      1   0.93157 3.8063 -2041.3
## - log.doctors    1   1.55295 4.4277 -1974.8
## - pct.bach.deg   1   1.80755 4.6823 -1950.2
## - pct.below.pov  1   2.53302 5.4078 -1886.8
##
## Step:  AIC=-2160.25
## log.per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log.land.area + log.doctors
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       2.9051 -2160.2
## + pop.65_plus    1   0.03031 2.8748 -2158.8
## + log.hosp.beds  1   0.00438 2.9007 -2154.8
## + log.crimes     1   0.00014 2.9049 -2154.2
## - pct.hs.grad    1   0.11147 3.0165 -2149.8
## - pct.unemp      1   0.15952 3.0646 -2142.8
## - log.land.area  1   0.37507 3.2801 -2112.9
## - pop.18_34      1   1.05209 3.9572 -2030.3
## - log.doctors    1   1.53330 4.4384 -1979.8
## - pct.bach.deg   1   1.86219 4.7673 -1948.4
## - pct.below.pov  1   2.50333 5.4084 -1892.9
```

```
summary(BIC.model)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log.land.area + log.doctors,
##     data = cdi_nonregion)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34147 -0.04886 -0.00538  0.04818  0.26969
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.2224950  0.0931210 109.776  < 2e-16 ***
## pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
## pct.hs.grad   -0.0044064  0.0010823  -4.071 5.56e-05 ***
## pct.bach.deg   0.0153853  0.0009246  16.641  < 2e-16 ***
## pct.below.pov -0.0242784  0.0012583 -19.294  < 2e-16 ***
## pct.unemp      0.0106037  0.0021771   4.871 1.56e-06 ***
## log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
## log.doctors    0.0606769  0.0040183  15.100  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.082 on 432 degrees of freedom
## Multiple R-squared:  0.8452, Adjusted R-squared:  0.8427
## F-statistic: 336.9 on 7 and 432 DF,  p-value: < 2.2e-16
```

Summary 5. Summary of BIC model without region

```
vif(BIC.model)
```

```
##     pop.18_34   pct.hs.grad  pct.bach.deg pct.below.pov     pct.unemp
##      1.416145      3.763103      3.269565      2.241555      1.691280
## log.land.area   log.doctors
##      1.131867      1.379671
```

Table 9. VIF of BIC model without region

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

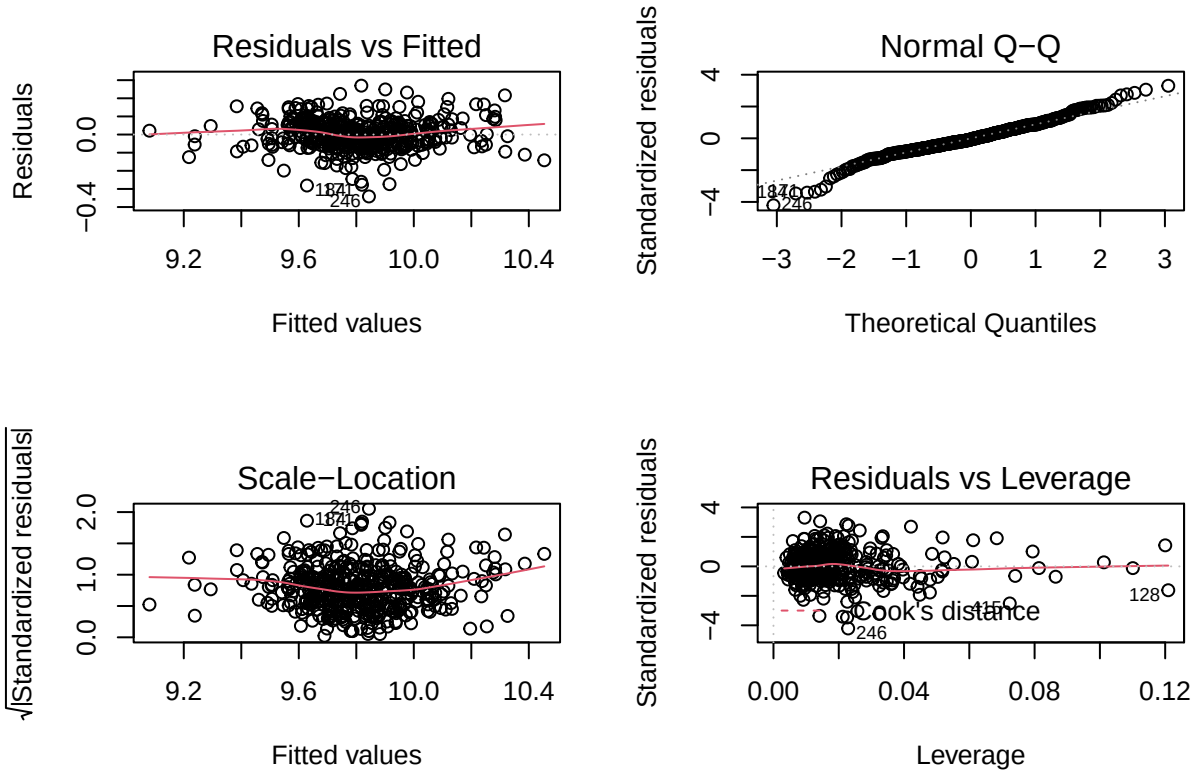

Figure 10. Diagnostic plots of BIC model without region

```
tmp <- cdi_region[c("log.per.cap.income", "pop.18_34", "pct.hs.grad", "pct.bach.deg", "pct.below.pov",
BIC.model.with.region <- lm(log.per.cap.income ~ .*region, data=tmp)
summary(BIC.model.with.region)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ . * region, data = tmp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.250782 -0.042332 -0.002298  0.040559  0.313570
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.1244260  0.2826240  35.823  < 2e-16 ***
## pop.18_34              -0.0147940  0.0026043  -5.681 2.55e-08 ***
## pct.hs.grad            -0.0024773  0.0034110  -0.726 0.468088
## pct.bach.deg            0.0140833  0.0029254   4.814 2.09e-06 ***
## pct.below.pov          -0.0237085  0.0036234  -6.543 1.81e-10 ***
## pct.unemp               0.0180393  0.0048923   3.687 0.000257 ***
## log.land.area          -0.0364187  0.0151355  -2.406 0.016564 *
## log.doctors             0.0544169  0.0093221   5.837 1.08e-08 ***
## regionNE                0.3243992  0.3577081   0.907 0.365004
```

```
## regionS                -0.0345856  0.3131668  -0.110 0.912116
## regionW                 1.5043946  0.4226868   3.559 0.000416 ***
## pop.18_34:regionNE     -0.0024780  0.0036873  -0.672 0.501939
## pop.18_34:regionS      -0.0008777  0.0030680  -0.286 0.774970
## pop.18_34:regionW       0.0014122  0.0040925   0.345 0.730220
## pct.hs.grad:regionNE   -0.0037529  0.0044150  -0.850 0.395813
## pct.hs.grad:regionS     0.0021198  0.0037853   0.560 0.575790
## pct.hs.grad:regionW    -0.0190188  0.0045881  -4.145 4.13e-05 ***
## pct.bach.deg:regionNE   0.0069429  0.0040312   1.722 0.085776 .
## pct.bach.deg:regionS   -0.0015774  0.0032000  -0.493 0.622328
## pct.bach.deg:regionW    0.0071026  0.0036374   1.953 0.051541 .
## pct.below.pov:regionNE -0.0014134  0.0050896  -0.278 0.781381
## pct.below.pov:regionS   0.0072764  0.0040739   1.786 0.074827 .
## pct.below.pov:regionW  -0.0161639  0.0054271  -2.978 0.003071 **
## pct.unemp:regionNE     -0.0083596  0.0073758  -1.133 0.257720
## pct.unemp:regionS      -0.0249396  0.0065867  -3.786 0.000176 ***
## pct.unemp:regionW      -0.0201466  0.0067713  -2.975 0.003101 **
## log.land.area:regionNE -0.0037179  0.0201435  -0.185 0.853656
## log.land.area:regionS  -0.0047582  0.0174155  -0.273 0.784825
## log.land.area:regionW   0.0151234  0.0181871   0.832 0.406154
## log.doctors:regionNE   -0.0046251  0.0132571  -0.349 0.727359
## log.doctors:regionS     0.0043337  0.0114401   0.379 0.705019
## log.doctors:regionW    -0.0034863  0.0131576  -0.265 0.791173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0759 on 408 degrees of freedom
## Multiple R-squared:  0.8747, Adjusted R-squared:  0.8652
## F-statistic: 91.91 on 31 and 408 DF,  p-value: < 2.2e-16
```

Summary 6. Summary of BIC model with region

We can see that stepwise regression using the BIC criterion actually is exactly the same as all-subsets model no matter we consider region or not, so we skip the part when we consider the region.

Apply stepwise AIC regression

```
AIC.model <- stepAIC(lm(log.per.cap.income ~ .,data = cdi_nonregion), direction = "both", k = 2)
```

```
## Start:  AIC=-2193.54
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##     log.doctors + log.hosp.beds + log.crimes
##
##                  Df Sum of Sq    RSS     AIC
## - log.crimes      1   0.00180 2.8636 -2195.3
## - log.hosp.beds   1   0.01216 2.8740 -2193.7
## <none>                        2.8618 -2193.5
## - pop.65_plus     1   0.03884 2.9006 -2189.6
## - log.doctors     1   0.11565 2.9775 -2178.1
## - pct.hs.grad     1   0.12699 2.9888 -2176.4
```

```
## - pct.unemp      1   0.17289 3.0347 -2169.7
## - log.land.area  1   0.36392 3.2257 -2142.9
## - pop.18_34      1   0.94423 3.8060 -2070.1
## - pct.bach.deg   1   1.56251 4.4243 -2003.8
## - pct.below.pov  1   2.44318 5.3050 -1924.0
##
## Step:  AIC=-2195.27
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##     log.doctors + log.hosp.beds
##
##                 Df Sum of Sq    RSS     AIC
## - log.hosp.beds  1   0.01116 2.8748 -2195.6
## <none>                       2.8636 -2195.3
## + log.crimes     1   0.00180 2.8618 -2193.5
## - pop.65_plus    1   0.03709 2.9007 -2191.6
## - pct.hs.grad    1   0.12662 2.9902 -2178.2
## - log.doctors    1   0.12889 2.9925 -2177.9
## - pct.unemp      1   0.17123 3.0348 -2171.7
## - log.land.area  1   0.37492 3.2385 -2143.1
## - pop.18_34      1   0.94270 3.8063 -2072.1
## - pct.bach.deg   1   1.59514 4.4587 -2002.4
## - pct.below.pov  1   2.47345 5.3371 -1923.3
##
## Step:  AIC=-2195.55
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##     log.doctors
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       2.8748 -2195.6
## + log.hosp.beds  1   0.01116 2.8636 -2195.3
## + log.crimes     1   0.00079 2.8740 -2193.7
## - pop.65_plus    1   0.03031 2.9051 -2192.9
## - pct.hs.grad    1   0.12309 2.9978 -2179.1
## - pct.unemp      1   0.16432 3.0391 -2173.1
## - log.land.area  1   0.38995 3.2647 -2141.6
## - pop.18_34      1   0.93157 3.8063 -2074.1
## - log.doctors    1   1.55295 4.4277 -2007.5
## - pct.bach.deg   1   1.80755 4.6823 -1982.9
## - pct.below.pov  1   2.53302 5.4078 -1919.5
```

```
summary(AIC.model)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##     log.doctors, data = cdi_nonregion)
##
## 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 ***
## pop.18_34     -0.0153488  0.0012988 -11.818  < 2e-16 ***
## pop.65_plus   -0.0027664  0.0012978  -2.132   0.0336 *
## pct.hs.grad   -0.0046579  0.0010843  -4.296 2.15e-05 ***
## 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 ***
## log.land.area -0.0364935  0.0047728  -7.646 1.36e-13 ***
## log.doctors    0.0626053  0.0041029  15.259  < 2e-16 ***
## ---
## 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
```

Summary 7.Summary of AIC model without region

```
vif(AIC.model)
```

```
##     pop.18_34   pop.65_plus   pct.hs.grad  pct.bach.deg pct.below.pov
##      1.950084      1.767181      3.808211      3.294199      2.277025
##     pct.unemp log.land.area   log.doctors
##      1.693439      1.139258      1.450175
```

Table 10. VIF of AIC model without region

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

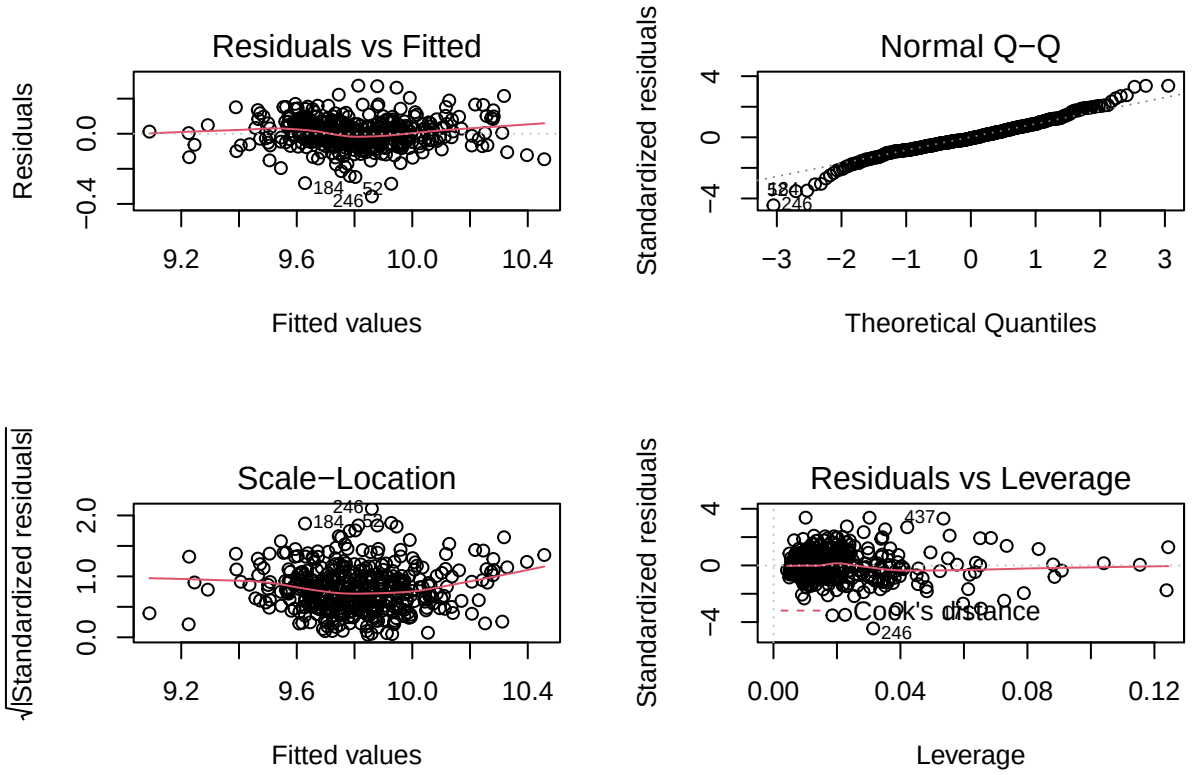

Figure 11. Diagnostic plots of AIC model without region

```
tmp <- cdi_region[c("log.per.cap.income", "pop.18_34", "pop.65_plus", "pct.hs.grad", "pct.bach.deg", "p
AIC.model.with.region <- lm(log.per.cap.income ~ .*region, data=tmp)
summary(AIC.model.with.region)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ . * region, data = tmp)
##
## 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 ***
## pop.18_34              -0.0150740  0.0028317  -5.323 1.69e-07 ***
## pop.65_plus            -0.0012483  0.0050165  -0.249 0.803614
## pct.hs.grad            -0.0026649  0.0034861  -0.764 0.445055
## 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 ***
## log.land.area          -0.0355230  0.0155258  -2.288 0.022654 *
## log.doctors             0.0548293  0.0094485   5.803 1.32e-08 ***
```

```
## 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 **
## 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
## 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 ***
## 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 **
## 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
## ---
## 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
```

Summary 8. Summary of AIC model with region

```
AIC.model.with.some.region <- update(AIC.model.with.region,
. ~ . - region:log.land.area - region:pop.18_34 - pop.65_plus:region - region:log.doctors - region:pct.
summary(all.subsets.model.with.some.region)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log.land.area + log.doctors +
##     region + pct.hs.grad:region + pct.below.pov:region + pct.unemp:region,
##     data = tmp)
##
## 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 ***
## pop.18_34              -0.0149347  0.0010897 -13.705  < 2e-16 ***
## pct.hs.grad            -0.0043532  0.0024515  -1.776 0.076501 .
## pct.bach.deg            0.0156310  0.0009715  16.090  < 2e-16 ***
## pct.below.pov          -0.0252029  0.0032612  -7.728 8.12e-14 ***
## pct.unemp               0.0197400  0.0046254   4.268 2.44e-05 ***
## log.land.area          -0.0381738  0.0053996  -7.070 6.51e-12 ***
## log.doctors             0.0572284  0.0040082  14.278  < 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 ***
## pct.hs.grad:regionNE    0.0017684  0.0029293   0.604 0.546374
## pct.hs.grad:regionS     0.0011525  0.0025618   0.450 0.653024
## pct.hs.grad:regionW    -0.0141473  0.0035826  -3.949 9.20e-05 ***
## pct.below.pov:regionNE -0.0015170  0.0046143  -0.329 0.742493
## pct.below.pov:regionS   0.0070185  0.0035199   1.994 0.046808 *
## pct.below.pov:regionW  -0.0137920  0.0051811  -2.662 0.008066 **
## 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 ***
## ---
## 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
```

Summary 9. Summary of AIC model with some region

```
vif(AIC.model.with.some.region)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## pop.18_34            2.147957e+00  1        1.465591
## pop.65_plus          2.088636e+00  1        1.445212
## pct.hs.grad          2.194977e+01  1        4.685058
## pct.bach.deg         4.122860e+00  1        2.030483
## pct.below.pov        1.748503e+01  1        4.181510
## pct.unemp            8.817304e+00  1        2.969395
## log.land.area        1.643607e+00  1        1.282032
## log.doctors          1.619490e+00  1        1.272592
## region               2.457552e+08  3       25.027480
## pct.hs.grad:region   8.520213e+07  3       20.976921
## pct.below.pov:region 5.700842e+03  3        4.226587
## pct.unemp:region     1.179976e+04  3        4.771397
```

Table 11. VIF of AIC model with some region

```
par(mfrow=c(2,2))
plot(AIC.model.with.some.region)
```

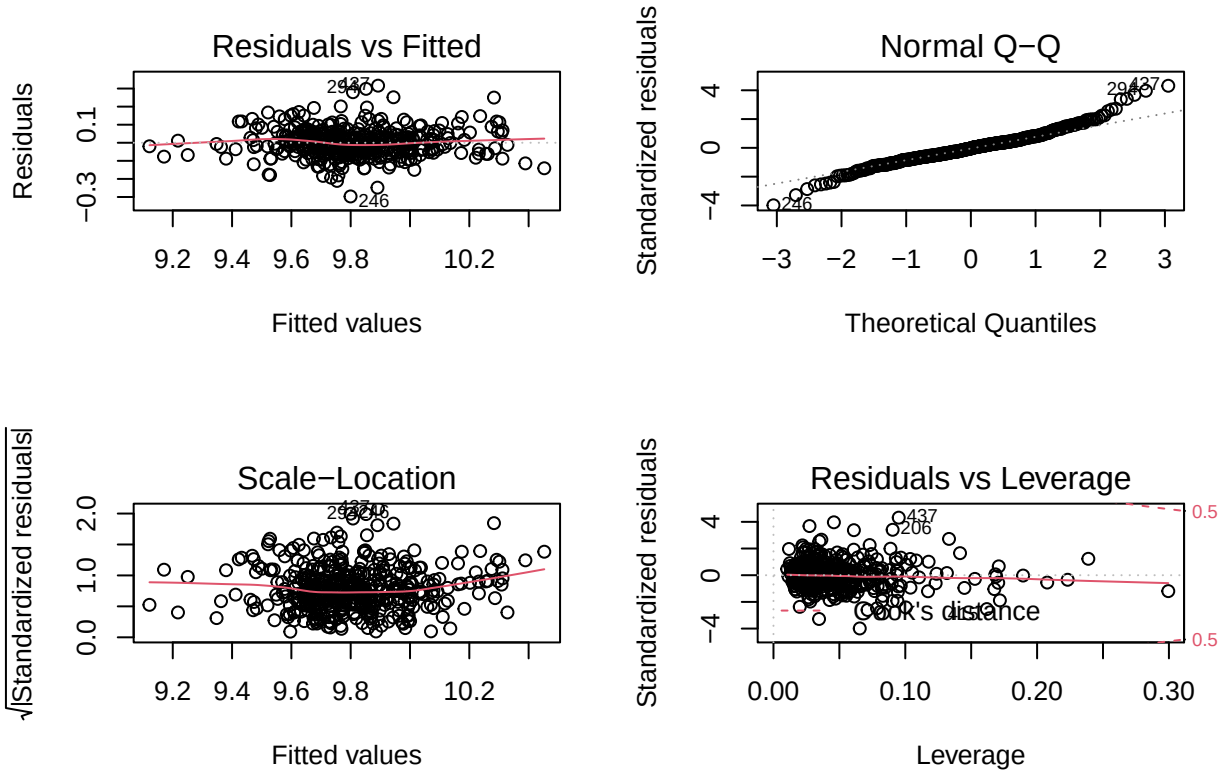

Figure 12. Diagnostic plots of AIC model with some region

```
adjusted_R2 = c(summary(all.subsets.model)$adj.r.squared, summary(BIC.model)$adj.r.squared, summary(AIC
compare <- cbind(AIC(all.subsets.model, BIC.model, AIC.model), BIC(all.subsets.model, BIC.model, AIC.mo

compare <- compare[,-c(1,3)]
compare
```

```
##                          AIC       BIC adjusted_R2
## all.subsets.model -942.2740 -905.4931   0.8426532
## BIC.model         -942.2740 -905.4931   0.8426532
## AIC.model         -944.8883 -904.0206   0.8439334
```

Table 12. Comparison of three models

Analysis on missing counties

```
boxplot(per.cap.income ~ region, data = cdi)
```

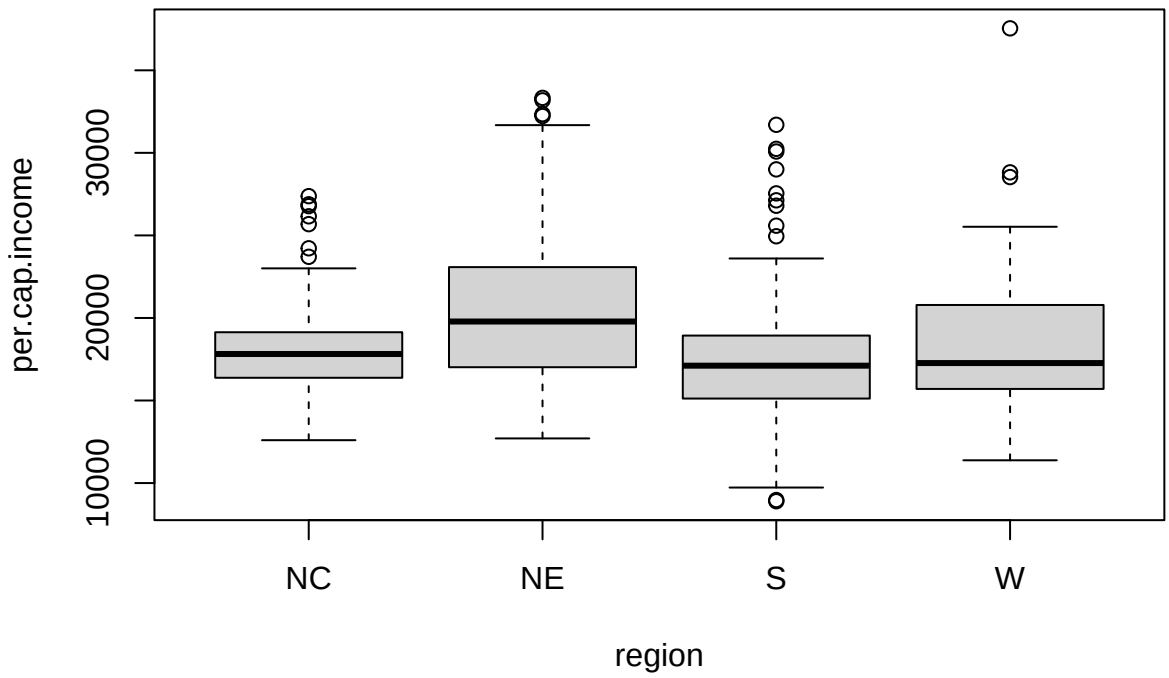

Figure 13. Boxplot of per capita income by region

```
cdi_remove <- cdi %>%
  filter(pop < 4000000)

boxplot(pop ~ region, data = cdi_remove)
```

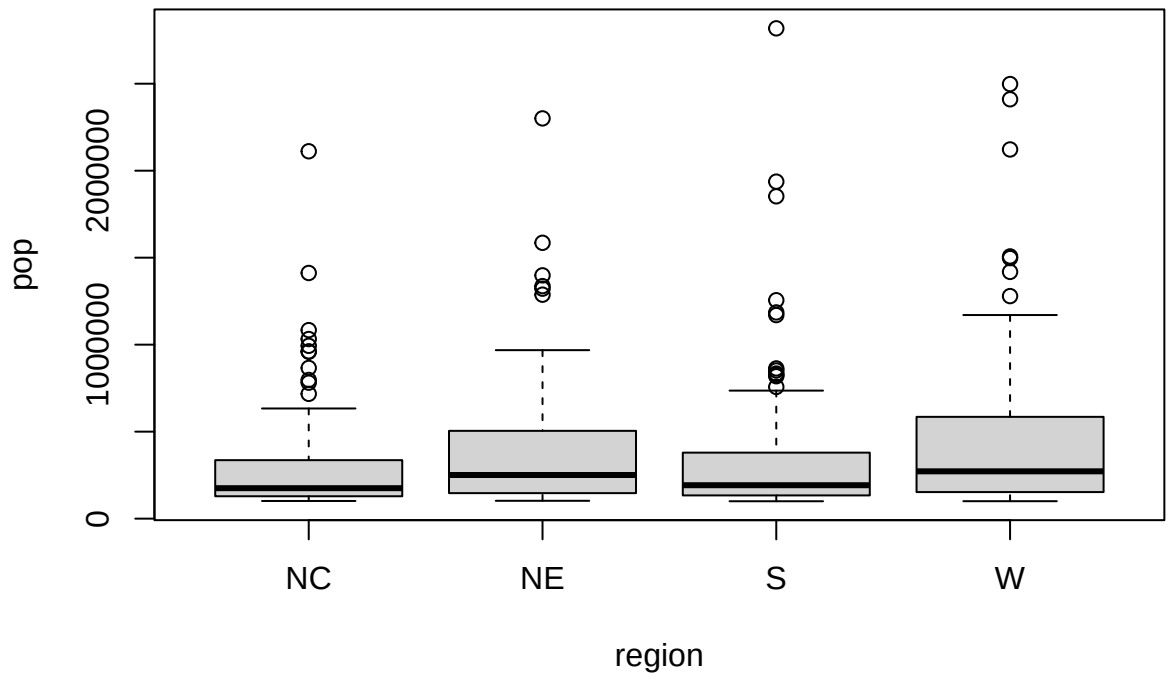

Figure 14. Boxplot of population by region

```
cdi %>%
  group_by(region) %>%
  summarise(median_per_cap_income = median(per.cap.income),
            mean_per_cap_income = mean(per.cap.income),
            sum_pop = sum(pop))
```

```
## # A tibble: 4 x 4
##   region median_per_cap_income mean_per_cap_income  sum_pop
##   <chr>                  <dbl>               <dbl>    <int>
## 1 NC                     17817              18301. 37386529
## 2 NE                     19785              20599. 40770956
## 3 S                      17110              17487. 50008592
## 4 W                      17268              18323. 44758728
```

Table 13. Median of per capita income and sum of population by region