Effects of Economic and Social Factors on Average Personal Income in the United States

Olivia Wang
ziyanw2@andrew.cmu.edu

13 October 2021

Abstract

This paper will investigate the effects of economic and social factors on average personal income. We examine CDI data obtained from Geospatial and Statistical Data Center, University of Virginia (2005)[1], using exploratory data analyses. We use exploratory data analysis and deploy all subsets method, stepwise regression as well as LASSO method to choose the best regression model. It appears that the per-capita income is highly influenced by land area, percent of population aged 18-34, percent of population 65 or order, number of serious crimes, percent high school graduates, percent bachelor's degrees, percent below poverty level, and percent unemployment; we found the level of average personal income also varies with region due to different economic policies. This conclusion, however, is not applicable to every case because of limits in our outdate data and model shortcomings. Ideally, we would need additional up to date dataset to perform a comprehensive analysis.

1. Introduction

Per-capita income is also called as average income per person is a measure of the amount of money earned per person in a nation or geographic region. Per-capita income can be used to determine the average per-person income for an area and to evaluate the standard of living and quality of life of the population. Per-capita income can be largely different from one county to another and average income per person was associated with many various factors especially economics and social wellness. Thus, some social scientists are interested in looking at this historical data, to learn the effects of economic and social factors on per-capita income in the United States. In this report, we investigate how average income per person was related to other variables associated with the county's economic, health and social well-being.

In particular we will:

- Investigate the data one pair of variables at a time and find variables which seem to be related to other variables.
- Build a regression model that predicts per-capita income of 1990 CDI population (in dollars) from crime rate and region of the country.
- Develop the best multiple regression model to predict per-capita income of 1990 CDI population (in dollars) from other variables associated with the county's economic, health and social well-being.

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

- Illustrate the reasons about if we should be worried about either the missing states or the missing counties.

2. Data

The CDI data for this study are from Kutner et al. (2005)[2]. *Original source:* Geospatial and Statistical Data Center, University of Virginia. The data were obtained from http://www.worldcat.org and are given in the file cdi.dat which is available on the website of University of Virginia Library Geospatial and Statistical Data Center.

The variables in the data set are shown in Table 1.

Variable Number	Variable	Definition & Comments
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 (northcentral region of the US), S (southern region of the US), and W (Western region of the US)

Table 1: Variable Definitions for the cdi.dat data set.

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

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

Table 2: Summary Statistics for the cdi.dat data set.

	NC	NE	S	W
Freq	108	103	152	77

Table 3: Sample size of region

There are total of 440 observations of 17 variables collected in cdi.dat data set. There do not appear to be any missing values in the data set.

One-dimensional summary statistics for the 13 continuous variables (except for id) are given in Table 2. Table 3 shows the sample size of the observations collected within each of the four geographic regions. Further EDA including histogram and scatterplot of continuous variables (Appendix 1, p. 4: Histogram; p. 6: Scatterplot) show that several variables are substantially skewed right. Besides, these observations suggest that we may run into multi-collinearity problems.

3. Methods

3.1 Investigate the data one pair of variables at a time and find variables which seem to be related to other variables

Our analysis contains of three parts:

- First, we relied on visual comparison of histograms and boxplots to capture univariate distributions.
- Then, we investigated two-variable relationships by using scatter plots and correlation matrix, we especially used scatter plots to concentrate on relationships with per-capita income.
- As needed, we took logarithms of seven continuous variables to address heavy skewing and potential leverage and influence issues (detailed R analysis can be found in Appendix 1, p. 6-9).

3.2 Build a regression model that predicts per-capita income of 1990 CDI population from crime rate and region of the country

We considered to build a regression model that predicts per-capita income of 1990 CDI population (in dollars) from crime rate and region of the country.

- In this case, for each version of the income and crime variables (total serious crimes and per-capita serious crimes), we considered two models using log-transformed variables. The first model was not included with interaction term, and we added interaction term with region in the second model.
- We examined residual diagnostic plots, ANOVA F-tests and we used AIC, BIC and R^2_{adj} to decide the best way to measure crime rate and the best model to predict per-capita income from crime rate and region of the country (details of these analyses in R can be found in Appendix 2, p. 9-19).

3.3 Develop the best multiple regression model to predict per-capita income of 1990 CDI population from other variables associated with the county's economic, health and social well-being

To develop the best multiple regression model to predict per-capita income (in dollars) from other variables associated with the county's economic, health and social well-being, we considered two parts.

- The first part is predictor selection:

We started by analyzing predictor variables which are correlated with per-capita income and also meaningful to be included in model. Next, we created two new predictor variables per-capita doctor and per-capita hospital beds with logarithmic transform which are comparable to per-capita income. Then, we kept total of eleven predictor variables including a categorical variable region and ten continuous variables.

- The second part is model fit:

We considered to choose the best regression model by using all subsets method, stepwise regression as well as LASSO method accordingly of ten continuous variables. Then, we added region to see if interaction with region in each of the method helps with model fit. We choosed our final model not only by examining summary of regression analysis, residual diagnostic plots, VIF test, marginal model plots, and ANOVA F-tests, but also comparing model performance based on AIC, BIC and R^2_{adj} (details of this R analysis can be found in Appendix 3, p. 20-47).

3.4 Illustrate the reasons about if we should be worried about either the missing states or the missing counties

Lastly, we illustrated the reasons about if we should be worried about either the missing states or the missing counties. In this part, we used some external information to comment.

4. Results

4.1 Visual Comparison of Exploratory Plots

Univariate Distributions

First, we relied on visual comparison of histograms and boxplots to capture univariate distributions. Based on the Figure 1, we can see there are seven continuous variables are extremely right skewed, in order to solve right skewed problem, we took logarithms of seven variables to address heavy skewing and potential leverage and influence issues. They are Per-capita income, Total personal income, Total serious crimes, Total population, Land area, Number of hospital beds and Number of active physicians.

Two-Variable Relationships

We investigated two-variable relationships by using scatter plots and correlation matrix. Based on the Figure 2, we can see the correlation matrix of numeric variables, it is obviously that

- Total personal income and Total population are highly correlated.
- Total serious crimes, Number of hospital beds and Number of active physicians are also reasonably highly correlated with each other.
- Percent high school graduates and Percent bachelor's degree are both positively correlated with Per-capita income.
- Percent below poverty level, Percent unemployment are both negatively correlated with Per-capita income.
- Percent high school graduates and Percent bachelor's degree, Percent below poverty level, Percent unemployment are moderately highly correlated with each other.

Relationships with Per-Capita Income

Then, we especially used scatter plots to concentrate on relationships with per-capita income from Figure 3. We get the same four variables mentioned above to predict per-capita income. The last plot shows how it varies across the four regions of the country.

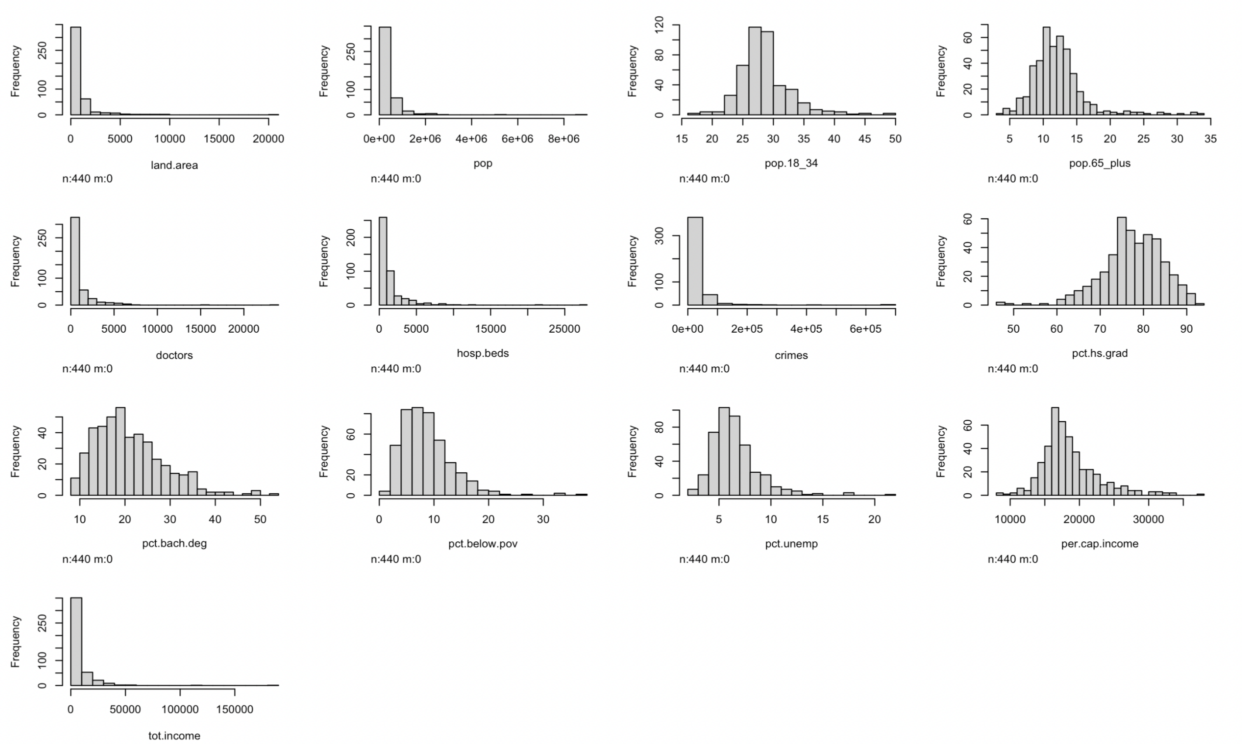


Figure 1: Distributions of variables.

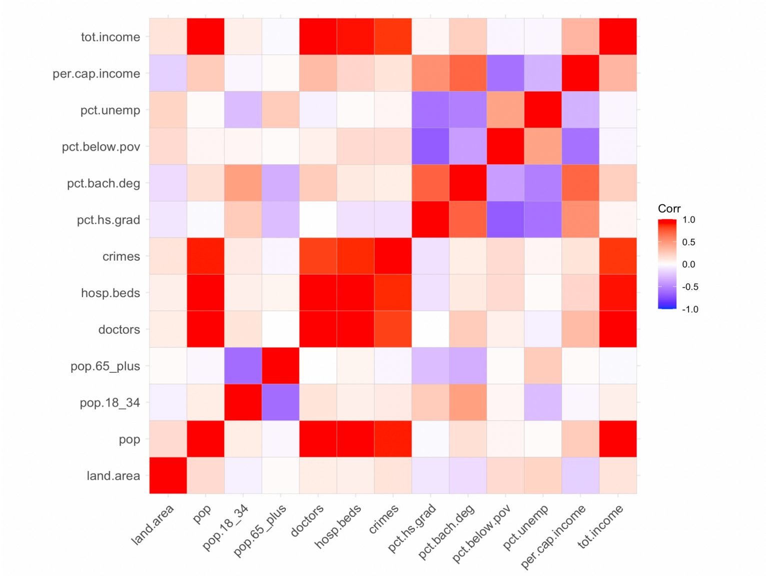


Figure 2: Correlation matrix of numeric variables.

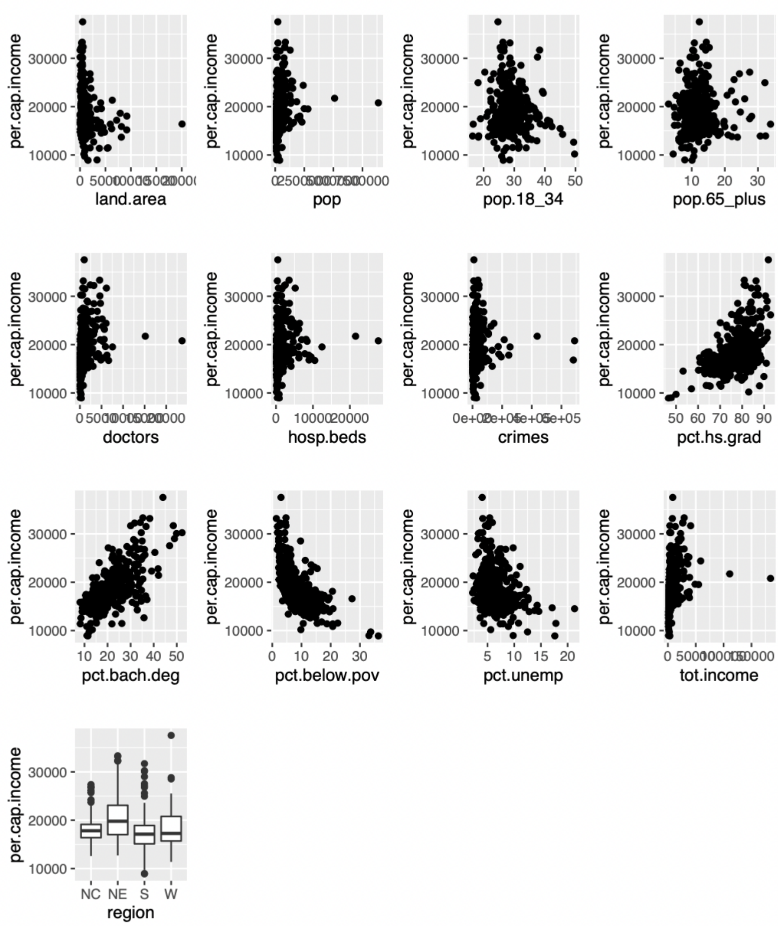


Figure 3: Scatter plots of numeric variables with per.cap.income.

4.2 Regression Analysis I

Total Serious Crimes

We considered to build a regression model that predicts per-capita income of 1990 CDI population (in dollars) from crime rate and region of the country. Firstly, we used total serious crimes as a crime rate measure, we considered these two regression models:

- Model A:

$$\text{log.per.cap.income} = 9.19 + 0.07 \times \text{log.crimes} + 0.10 \times \text{regionNE} - 0.09 \times \text{regionS} + \varepsilon$$

- Model B:

$$\begin{aligned}\text{log.per.cap.income} &\\ &= 9.34 + 0.05 \times \text{log.crimes} - 0.18 \times \text{regionNE} - 0.20 \times \text{regionS} - 0.31 \times \text{regionW} \\ &+ 0.03 \times \text{log.crimes} * \text{regionNE} + 0.01 \times \text{log.crimes} * \text{regionS} + 0.03 \times \text{log.crimes} \\ &* \text{regionW} + \varepsilon\end{aligned}$$

From residual plots (Appendix 2, p. 11: Model A; p. 13: Model B), none of them are awful, so we could trust F-test to compare these models:

```
Analysis of Variance Table

Model 1: log.per.cap.income ~ log.crimes + region
Model 2: log.per.cap.income ~ log.crimes + region + log.crimes * region
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    435 14.949
2    432 14.872  3  0.076778 0.7434 0.5266
```

Table 4: Analysis of variance table 1.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.18843110	0.079812437	115.125305	0.000000e+00
log.crimes	0.06669491	0.008421114	7.919963	2.002771e-14
regionNE	0.10445836	0.025531314	4.091382	5.110827e-05
regionS	-0.08698350	0.023617956	-3.682939	2.595887e-04
regionW	-0.05527965	0.028167096	-1.962561	5.033416e-02

Table 5: Estimated coefficients and standard errors for Model A.

Table 4 provides Analysis of variance table, besides, Model A had larger R^2_{adj} than Model B and better residual diagnostic plots. Table 5 gives the full table of estimated coefficients and standard errors for Model A. Based on the output, we can get Model A, shown with estimated regression coefficients:

$$\text{log.per.cap.income} = 9.19 + 0.07 \times \text{log.crimes} + 0.10 \times \text{regionNE} - 0.09 \times \text{regionS} + \varepsilon$$

Per-Capita Serious Crimes

Similar to the first section, but in this time, we used per-capita serious crimes as a crime rate measure, we considered these two regression models:

- Model C:

$$\begin{aligned}\text{log. per. cap. income} \\ &= 9.94 + 0.04 \times \text{log. per. cap. crimes} + 0.11 \times \text{regionNE} - 0.07 \times \text{regionS} + \varepsilon\end{aligned}$$

- Model D:

$$\begin{aligned}\text{log. per. cap. income} \\ &= 9.91 + 0.03 \times \text{log. per. cap. crimes} + 0.21 \times \text{regionNE} - 0.10 \times \text{regionS} \\ &+ 0.08 \times \text{regionW} + 0.03 \times \text{log. per. cap. crimes} * \text{regionNE} \\ &- 0.01 \times \text{log. per. cap. crimes} * \text{regionS} + 0.03 \times \text{log. per. cap. crimes} * \text{regionW} + \varepsilon\end{aligned}$$

Again, residual plots (Appendix 2, p. 16: Model C; p. 18: Model D) look ok. So we could trust F-test to compare these models:

```
Analysis of Variance Table

Model 1: log.per.cap.income ~ log.per.cap.crimes + region
Model 2: log.per.cap.income ~ log.per.cap.crimes + region + log.per.cap.crimes *
    region
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    435 16.952
2    432 16.928  3   0.02408 0.2048  0.893
```

Table 6: Analysis of variance table 2.

```
                      Estimate Std. Error     t value     Pr(>|t|)
(Intercept)         9.93628386 0.06933762 143.3029302 0.000000e+00
log.per.cap.crimes  0.04242970 0.02147872   1.9754301 4.885103e-02
regionNE            0.11456811 0.02760334   4.1505159 3.992234e-05
regionS            -0.07455751 0.02624295  -2.8410493 4.707758e-03
regionW            -0.02425540 0.03001832  -0.8080196 4.195209e-01
```

Table 7: Estimated coefficients and standard errors for Model C.

Table 6 provides Analysis of variance table, besides, Model C had higher R^2_{adj} than Model D with and better residual diagnostic plots. Table 7 gives the full table of estimated coefficients and standard errors for Model C. Based on the output, we can get Model C, shown with estimated regression coefficients:

$$\begin{aligned}\text{log. per. cap. income} \\ &= 9.94 + 0.04 \times \text{log. per. cap. crimes} + 0.11 \times \text{regionNE} - 0.07 \times \text{regionS} + \varepsilon\end{aligned}$$

Conclusion

We compared these two winners (Model A vs. Model C), by using AIC or BIC and R^2_{adj}, and the output (Appendix 2, p. 19) shows that Model A is the best model, by either AIC or BIC, R^2_{adj}.

Thus, the best way to measure crime rate is total serious crimes and the best model to predict per-capita income of 1990 CDI population (in dollars) from crime rate and region of the country is:

$$\text{log. per. cap. income} = 9.19 + 0.07 \times \text{log. crimes} + 0.10 \times \text{regionNE} - 0.09 \times \text{regionS} + \varepsilon$$

Here are somethings to interpret:

- For every 1% increase in number of crimes, there is 0.07% increase in expected per-capita income.
- In the main effects for region, the Northeastern part of the U.S. is positively correlated with per-capita income, while South part of the U.S. is negatively correlated with per-capita income.

4.3 Regression Analysis II

Predictor Selection

To develop the best multiple regression model to predict per-capita income (in dollars) from other variables associated with the county's economic, health and social well-being. The first part is predictor selection: we kept total of eleven predictor variables including a categorical variable region and ten continuous variables.

Considering predictor variables which are correlated with per.cap.income and also meaningful to be included in model, we removed total population and total serious crimes, since per-capita income is a deterministic function of them; we also removed id and county, since they are not useful to include in models; with regard to state, we considered the large sample size of state which is 48 and the collinearity between state and region, hence we decided to remove it as well.

The reason to consider per-capita active physicians and per-capita hospital beds is that per-capita crime is more comparable to, or at least the same on the same scale as per-capita income, besides other predictor variables are displayed as percentage such as "percent unemployment", "percent of population aged 18-34" etc.. Considering that, and we equally got logarithmic transformation on them**.** Thus, these two variables should be helpful to interpret.

Multiple Regression Model

We considered to choose the best regression model by using all subsets method, stepwise regression as well as LASSO method.

- **All Subsets Method:**

We started variable selection by all subsets method, and ended with two models (the second model with region included), shown here with estimated regression coefficients:

Model (1):

$$\begin{aligned}\mathrm{log.per.cap.income} &= 10.59 - 0.03 \times \mathrm{log.land.area} - 0.02 \times \mathrm{pop.18\_34} - 0.003 \times \mathrm{pop.65\_plus} \\ &+ 0.05 \times \mathrm{log.crimes} - 0.005 \times \mathrm{pct.hs.grad} + 0.02 \times \mathrm{pct.bath.deg} \\ &- 0.03 \times \mathrm{pct.below.pov} + 0.01 \times \mathrm{pct.uemp} + 0.06 \times \mathrm{log.per.cap.hosp.beds} + \varepsilon\end{aligned}$$

Model (2):

$$\begin{aligned}\mathrm{log.per.cap.income} &= 10.30 - 0.04 \times \mathrm{log.land.area} - 0.02 \times \mathrm{pop.18\_34} - 0.0009 \times \mathrm{pop.65\_plus} \\ &+ 0.05 \times \mathrm{log.crimes} - 0.005 \times \mathrm{pct.hs.grad} + 0.018 \times \mathrm{pct.bath.deg} \\ &- 0.02 \times \mathrm{pct.below.pov} + 0.01 \times \mathrm{pct.uemp} + 0.05 \times \mathrm{log.per.cap.hosp.beds} \\ &+ 0.71 \times \mathrm{regionW} + 0.09 \times \mathrm{regionS} - 0.01 \times \mathrm{regionW} * \mathrm{pct.hs.grad} \\ &- 0.02 \times \mathrm{regionW} * \mathrm{pct.below.pov} - 0.01 \times \mathrm{regionS} * \mathrm{pct.unemp} - 0.01 \times \mathrm{regionW} \\ &* \mathrm{pct.unemp} + 0.06 \times \mathrm{regionW} * \mathrm{log.per.cap.hosp.bed} + \varepsilon\end{aligned}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.587197285	0.1313299362	80.615263	1.129184e-261
log.land.area	-0.032852917	0.0050767575	-6.471240	2.652408e-10
pop.18_34	-0.016461929	0.0013619800	-12.086763	3.717908e-29
pop.65_plus	-0.003388813	0.0014657445	-2.312008	2.124889e-02
log.crimes	0.047582348	0.0041507151	11.463651	9.923162e-27
pct.hs.grad	-0.005385007	0.0011214092	-4.802000	2.172430e-06
pct.bach.deg	0.018467491	0.0008952562	20.628163	3.174743e-66
pct.below.pov	-0.028071786	0.0014295288	-19.637090	9.394425e-62
pct.unemp	0.012797690	0.0023074528	5.546241	5.106290e-08
log.per.cap.hosp.beds	0.055269285	0.0094950419	5.820857	1.146190e-08

Table 8: Estimated coefficients and standard errors for model (1).

```
Analysis of Variance Table

Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
    log.crimes + pct.hs.grad + pct.bach.deg + pct.below.pov +
    pct.unemp + log.per.cap.hosp.beds
Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
    log.crimes + pct.hs.grad + pct.bach.deg + pct.below.pov +
    pct.unemp + log.per.cap.hosp.beds + region + pct.hs.grad:region +
    pct.below.pov:region + pct.unemp:region + log.per.cap.hosp.beds:region
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    430 3.0686
2    415 2.5113 15   0.55731 6.1397 1.062e-11 ***
```

Table 9: Estimated coefficients and standard errors for model (2).

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	10.2968941	0.2584674	39.838	< 2e-16	***
log.land.area	-0.0361999	0.0055526	-6.519	2.05e-10	***
pop.18_34	-0.0159934	0.0013109	-12.201	< 2e-16	***
pop.65_plus	-0.0009692	0.0015013	-0.646	0.518910	
log.crimes	0.0485077	0.0041196	11.775	< 2e-16	***
pct.hs.grad	-0.0058902	0.0024860	-2.369	0.018278	*
pct.bach.deg	0.0180357	0.0009314	19.365	< 2e-16	***
pct.below.pov	-0.0230162	0.0039675	-5.801	1.31e-08	***
pct.unemp	0.0126859	0.0049170	2.580	0.010223	*
log.per.cap.hosp.beds	0.0058340	0.0170453	0.342	0.732327	
regionNE	0.0133618	0.3228946	0.041	0.967012	
regionS	0.0866043	0.2740041	0.316	0.752109	
regionW	1.7063132	0.3788133	4.504	8.67e-06	***
pct.hs.grad:regionNE	0.0035491	0.0029754	1.193	0.233634	
pct.hs.grad:regionS	0.0021277	0.0025963	0.820	0.412964	
pct.hs.grad:regionW	-0.0134941	0.0036328	-3.715	0.000231	***
pct.below.pov:regionNE	-0.0057364	0.0054144	-1.059	0.290002	
pct.below.pov:regionS	0.0026802	0.0044069	0.608	0.543390	
pct.below.pov:regionW	-0.0202029	0.0057138	-3.536	0.000452	***
pct.unemp:regionNE	-0.0051337	0.0074663	-0.688	0.492103	
pct.unemp:regionS	-0.0148402	0.0067267	-2.206	0.027920	*
pct.unemp:regionW	-0.0138082	0.0068760	-2.008	0.045274	*
log.per.cap.hosp.beds:regionNE	0.0359182	0.0293619	1.223	0.221912	
log.per.cap.hosp.beds:regionS	0.0407353	0.0209645	1.943	0.052685	.
log.per.cap.hosp.beds:regionW	0.0616266	0.0274782	2.243	0.025440	*

Table 10: Analysis of variance table 3.

Model (1) shows that we got final model from all subsets method with nine predictors. All the coefficients of predictors have significantly different from zero.

For Model (2), shows interaction with region, and some of coefficients of predictors have significantly different from zero.

However, most of the coefficients are small, and some seem to have the wrong sign (e.g. pct.hs.grad and pct.unemp, log.crimes).

Both of the ANOVA F test from Table 9 and AIC, we prefer Model (1). On the other hand, BIC prefers the simpler model. Besides, Model (2) had higher R^2_{adj} and better residual diagnostic plots (Appendix 3.1, p. 21-31).

Based above, all subsets method selects Model (2) as a better model.

- **Stepwise Regression**

Stepwise regression with without interaction with region choosed the model with nine predictors.

Model (3):

$$\begin{aligned}\mathrm{log.per.cap.income} &= 10.91 - 0.03 \times \mathrm{log.land.area} - 0.02 \times \mathrm{pop.18\_34} - 0.003 \times \mathrm{pop.65\_plus} \\ &+ 0.04 \times \mathrm{log.crimes} - 0.005 \times \mathrm{pct.hs.grad} + 0.02 \times \mathrm{pct.bath.deg} \\ &- 0.03 \times \mathrm{pct.below.pov} + 0.01 \times \mathrm{pct.uemp} + 0.08 \times \mathrm{log.per.cap.doctors} + \varepsilon\end{aligned}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.9050376	0.1426061	76.470	< 2e-16
log.land.area	-0.0322550	0.0048983	-6.585	1.33e-10
pop.18_34	-0.0162008	0.0013128	-12.340	< 2e-16
pop.65_plus	-0.0038544	0.0013884	-2.776	0.00574
log.crimes	0.0389684	0.0042899	9.084	< 2e-16
pct.hs.grad	-0.0050247	0.0010892	-4.613	5.24e-06
pct.bach.deg	0.0150274	0.0009569	15.704	< 2e-16
pct.below.pov	-0.0276564	0.0013299	-20.796	< 2e-16
pct.unemp	0.0130256	0.0022235	5.858	9.32e-09
log.per.cap.doctors	0.0822644	0.0105519	7.796	4.86e-14

Table 11: Estimated coefficients and standard errors for model (3).

Model (3) shows that all coefficients of the predictors have significantly different from zero. However, most of the coefficients are small, and some still seem to have the wrong sign (e.g. pct.hs.grad and pct.unemp, log.crimes). We can see there is one variable log.per.cap.doctors is different from Model (1) where it has log.per.cap.hosp.beds instead.

Based on output of VIF test, residual diagnostics plots and marginal model plots (Appendix 3.2, p. 31-39), we found very similar output to Model (1) since we only have one variable different from Model (1).

However, in this case, we decided not to add the interaction terms because after tried, both of BIC-based model and AIC-based model, there are some statistically significant interactions, but all of the interaction terms with coefficients are close to 0. Thus, it is not useful to consider interaction terms on per-capita income.

Based above, stepwise regression selects Model (3) as the best model.

- **LASSO**

LASSO output ended with two models (the second model with region included), shown here with estimated regression coefficients:

Model (5):

$$\begin{aligned}\mathrm{log.per.cap.income} \\ &= 10.36 - 0.04 \times \mathrm{log.land.area} - 0.01 \times \mathrm{pop.18\_34} + 0.04 \times \mathrm{log.crimes} \\ &+ 0.01 \times \mathrm{pct.bath.deg} - 0.02 \times \mathrm{pct.below.pov} + 0.01 \times \mathrm{pct.uemp} \\ &+ 0.06 \times \mathrm{log.per.cap.doctors} + \varepsilon\end{aligned}$$

Model (6):

$$\begin{aligned}\mathrm{log.per.cap.income} \\ &= 9.99 - 0.03 \times \mathrm{log.land.area} - 0.02 \times \mathrm{pop.18\_34} + 0.04 \times \mathrm{log.crimes} \\ &+ 0.01 \times \mathrm{pct.bath.deg} - 0.02 \times \mathrm{pct.below.pov} + 0.01 \times \mathrm{pct.uemp} \\ &+ 0.03 \times \mathrm{log.per.doctors} + 0.68 \times \mathrm{regionW} + 0.30 \times \mathrm{regionS} + 0.38 \times \mathrm{regionNE} \\ &- 0.01 \times \mathrm{regionNE} * \mathrm{pct.unemp} - 0.01 \times \mathrm{regionS} * \mathrm{pct.unemp} + 0.04 \times \mathrm{regionS} \\ &* \mathrm{log.per.cap.doctors} + 0.1 \times \mathrm{regionW} * \mathrm{log.per.cap.doctors} + \varepsilon\end{aligned}$$

However, most of the coefficients are small, and some still seem to have the wrong sign in Model (5) and Model (6) (e.g. log.crimes).

Both of the ANOVA F test from Table 13 and AIC prefer Model (6). On the other hand, BIC prefers the simpler model. Besides Model (6) had higher R^2_{adj} than Model (5) and better residual diagnostic plots (Appendix 3.3, p. 39-47).

Based above, LASSO selects Model (6) as the best model.

```
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        10.3595873  0.0937621 110.488  < 2e-16 ***
log.land.area      -0.0369618  0.0049080  -7.531 2.97e-13 ***
pop.18_34          -0.0145057  0.0011457 -12.661  < 2e-16 ***
log.crimes          0.0413618  0.0043814   9.440  < 2e-16 ***
pct.bach.deg        0.0134060  0.0008351  16.053  < 2e-16 ***
pct.below.pov      -0.0239156  0.0011335 -21.099  < 2e-16 ***
pct.unemp           0.0148977  0.0021934   6.792 3.67e-11 ***
log.per.cap.doctors 0.0731657  0.0099539   7.350 9.97e-13 ***
```

Table 12: Estimated coefficients and standard errors for model (5).

```
Analysis of Variance Table

Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.crimes + 
    pct.bach.deg + pct.below.pov + pct.unemp + log.per.cap.doctors
Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.crimes + 
    pct.bach.deg + pct.below.pov + pct.unemp + log.per.cap.doctors + 
    region + region:pct.unemp + region:log.per.cap.doctors
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)    
1    432 3.0785                                  
2    423 2.7345  9   0.34399 5.9123 8.56e-08 ***
```

Table 13: Analysis of variance table 4.

```
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  9.9901410  0.1156985  86.346  < 2e-16
log.land.area               -0.0296356  0.0055557  -5.334 1.57e-07
pop.18_34                   -0.0150500  0.0011322 -13.293  < 2e-16
log.crimes                   0.0442597  0.0043688  10.131  < 2e-16
pct.bach.deg                 0.0137848  0.0008232  16.746  < 2e-16
pct.below.pov               -0.0206012  0.0012481 -16.507  < 2e-16
pct.unemp                    0.0144703  0.0042471   3.407 0.000719
log.per.cap.doctors          0.0280807  0.0141504   1.984 0.047852
regionNE                     0.3772241  0.1318038   2.862 0.004419
regionS                      0.2985624  0.1067612   2.797 0.005401
regionW                      0.6787587  0.1610012   4.216 3.04e-05
pct.unemp:regionNE          -0.0139355  0.0062731  -2.221 0.026846
pct.unemp:regionS           -0.0118148  0.0053499  -2.208 0.027751
pct.unemp:regionW            0.0039229  0.0049671   0.790 0.430100
log.per.cap.doctors:regionNE  0.0408476  0.0222811   1.833 0.067463
log.per.cap.doctors:regionS   0.0401736  0.0175183   2.293 0.022325
log.per.cap.doctors:regionW   0.1159527  0.0266833   4.346 1.74e-05
```

Table 14: Estimated coefficients and standard errors for model (6).

Final Model

From Table 15, we get all the information about our models from three methods. Based on the information, both of the R^2_{adj}, AIC prefer Model (2). On the other hand, BIC prefers the simpler Model (3). There is not much information for residual diagnostic plots since as mentioned before, they are all good for all the models.

While in this case, we prefer with Model (2) based on most of the criteria.

Method	Model	R^2_{adj}	AIC	BIC
All Subsets	(2)	0.8584	-972.3557	-866.0996
Stepwise Regression	(3)	0.8422	-938.9712	-894.0166
LASSO	(6)	0.8487	-950.8978	-877.3358

Table 15: Summary table of all models.

4.4. Missing Values

In this case, we only have 440 observations from 48 states (including Washington DC) in the dataset, while there are approximately 3000 counties from 51 states (including Washington DC) in the United States[3]. In addition, there are many of states in this dataset only with few counties recorded, and the rich and poverty level are different with each county and state.

However, in this project we do not need to be worried about missing states or the missing counties. Because considering the above reasons, our analysis did not include state and county as a predictor for the regression analysis.

[3] Nelson, G. D., & Rae, A. (2016). An economic geography of the United States: From commutes to megaregions. *PLoS ONE*, *11*(11), 1–23. https://doi.org/10.1371/journal.pone.016608.

5. Discussion

5.1 Investigate the data one pair of variables at a time and find variables which seem to be related to other variables

We investigated two-variable relationships and concluded (Appendix 1, p. 9: Correlation Matrix):

- Total personal income and Total population are highly correlated.
- Total serious crimes, Number of hospital beds and Number of active physicians are also reasonably highly correlated with each other.
- Percent high school graduates and Percent bachelor's degree are both positively correlated with Per-capita income.
- Percent below poverty level, Percent unemployment are both negatively correlated with Per-capita income.
- Percent high school graduates and Percent bachelor's degree, Percent below poverty level, Percent unemployment are moderately highly correlated with each other.

5.2 Build a regression model that predicts per-capita income of 1990 CDI population from crime rate and region of the country

When considered to build a regression model that predicts per-capita income of 1990 CDI population (in dollars) from crime rate and region of the country.

For each version of the income and crime variables (number of crimes and per-capita crime), we considered interactions. We examined residual diagnostic plots, ANOVA F-tests and we used AIC, BIC and R^2_{adj} to get the best way to measure crime rate is total serious crimes, the best model shown here with estimated regression coefficients:

$$\mathrm{log.per.cap.income} = 9.19 + 0.07 \times \mathrm{log.crimes} + 0.10 \times \mathrm{regionNE} - 0.09 \times \mathrm{regionS} + \varepsilon$$

5.3 Develop the best multiple regression model to predict per-capita income of 1990 CDI population from other variables associated with the county's economic, health and social well-being

We selected the best regression model to predict per-capita income (in dollars) from other variables associated with the county's economic, health and social well-being by using all subsets method, stepwise regression as well as LASSO method.

Then, we compared model performance based on AIC, BIC and R^2_{adj}. We found that the model that predicts per-capita income best is model from all subsets with interaction added, shown here with estimated regression coefficients:

$$\begin{aligned}\mathrm{log.per.cap.income} &\\ &= 10.30 - 0.04 \times \mathrm{log.land.area} - 0.02 \times \mathrm{pop.18\_34} - 0.0009 \times \mathrm{pop.65\_plus} \\ &+ 0.05 \times \mathrm{log.crimes} - 0.005 \times \mathrm{pct.hs.grad} + 0.018 \times \mathrm{pct.bath.deg} \\ &- 0.02 \times \mathrm{pct.below.pov} + 0.01 \times \mathrm{pct.uemp} + 0.05 \times \mathrm{log.per.cap.hosp.beds} \\ &+ 0.71 \times \mathrm{regionW} + 0.09 \times \mathrm{regionS} - 0.01 \times \mathrm{regionW} * \mathrm{pct.hs.grad} \\ &- 0.02 \times \mathrm{regionW} * \mathrm{pct.below.pov} - 0.01 \times \mathrm{regionS} * \mathrm{pct.unemp} - 0.01 \times \mathrm{regionW} \\ &* \mathrm{pct.unemp} + 0.06 \times \mathrm{regionW} * \mathrm{log.per.cap.hosp.bed} + \varepsilon\end{aligned}$$

Here are somethings to interpret:

- For every 1% increase in land area, there is 0.07% decrease in expected per-capita income.
- For every 1 unit increase in percent of population aged 18-34, there is 0.02 decrease in expected per-capita income.
- For every 1 unit increase in percent high school graduates, there is 0.005 decrease in expected per-capita income.
- For every 1% increase in total serious crimes, there is 0.05% increase in expected per-capita income.
- For every 1 unit increase in percent bachelor's degrees, there is 0.018 increase in expected per-capita income.
- For every 1 unit increase in percent below poverty level, there is 0.02 decrease in expected per-capita income.
- For every 1 unit increase in percent unemployment, there is 0.01 increase in expected per-capita income.
- For every 1% increase in per-capita hospital beds, there is 0.05% decrease in expected per-capita income.
- The interaction between Western region and percent high school graduates is negatively correlated with per-capita income.
- The interaction between Western region and percent below poverty level is negatively correlated with per-capita income.
- The interaction between Southern/Western region and percent unemployment is negatively correlated with per-capita income.
- The interaction between Western region and per-capita hospital beds is positively correlated with per-capita income.
- In the main effects for region, the Western and Southern part of the U.S. is positively correlated with per-capita income.

5.4 Illustrate the reasons about if we should be worried about either the missing states or the missing counties

In this case, we only have limited 440 observations from 48 states (including Washington DC) in the dataset, which are just small part of the United States. In addition, there are many of states in this dataset only with few counties recorded, and the rich and poverty level are different with each county and state.

In this project we do not need to be worried about missing states or the missing counties. Because considering the above reasons, our analysis did not include state and county as a predictor for the regression analysis.

5.5 Limitations

A key limitation of this work is that the data is quite old, from 2005. This limits the generalizability of the results to the present time. Besides, as we have mentioned above, in the final model there still are obvious shortcomings such as most of the coefficients are small, and some still seem to have the wrong sign but we needed to add these predictors since they are some significant interactions. Also, the data only captured limited 440 observations from 48 states (including Washington DC) in the dataset, which are just small parts of the United States.

Hence, we need additional up to date dataset to draw a more precise conclusion. It might be also useful to have extra test set of data to test the final model, in this case we could better verify if the final model is a good fit with less negative effects from limitations.

References

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

Nelson, G. D., & Rae, A. (2016). An economic geography of the United States: From commutes to megaregions. *PLoS ONE*, *11*(11), 1–23. https://doi.org/10.1371/journal.pone.0166083

Technical Appendix

ZIYAN (OLIVIA) WANG

10/7/2021

Appendix 1. Initial Data Import & Exploration

```
library(tinytex)
library(tidyverse)
library(kableExtra)
library(GGally)
library(grid)
library(gridExtra)
library(ggplotify)
library(reshape2)
library(Hmisc)
cdi <- read.table("cdi.dat",header=T)
```

The following is a table with the usual one-dimensional summary statistics.

```
cdi_con <- cdi[,-c(1,2,3,17)]
summary(cdi_con)
```

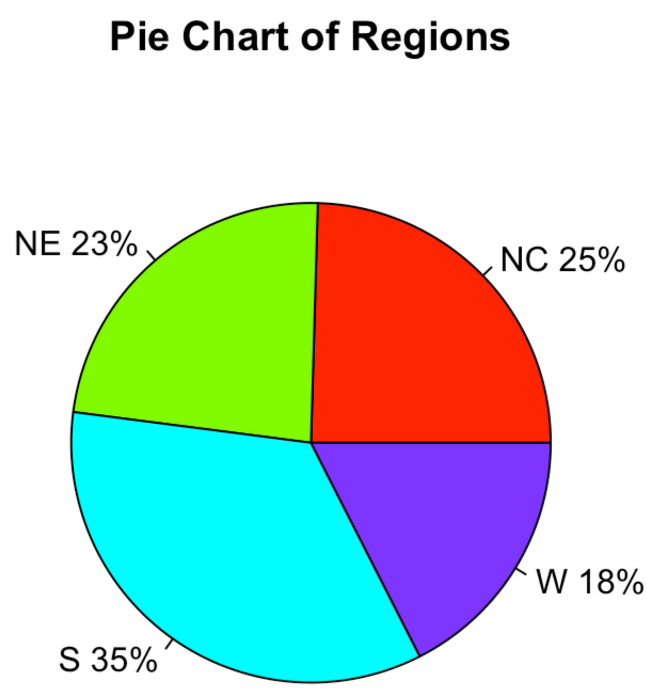
```
##    land.area           pop            pop.18_34      pop.65_plus
##  Min.   :   15.0   Min.   : 100043   Min.   :16.40   Min.   : 3.000
##  1st Qu.:  451.2   1st Qu.: 139027   1st Qu.:26.20   1st Qu.: 9.875
##  Median :  656.5   Median : 217280   Median :28.10   Median :11.750
##  Mean   : 1041.4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

This is a table for categorical variable region.

```
table(cdi$region)
```

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

```
# pie charts for some of categorical variables

# region
slices_region <- c(108, 103, 152, 77)
lbls <- c("NC", "NE", "S", "W")
pct <- round(slices_region/sum(slices_region)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices_region,labels = lbls, col=rainbow(length(lbls)),
   main="Pie Chart of Regions")
```




```
# Indicate where (in which variables) there is missing data
cdi[!complete.cases(cdi),]
```

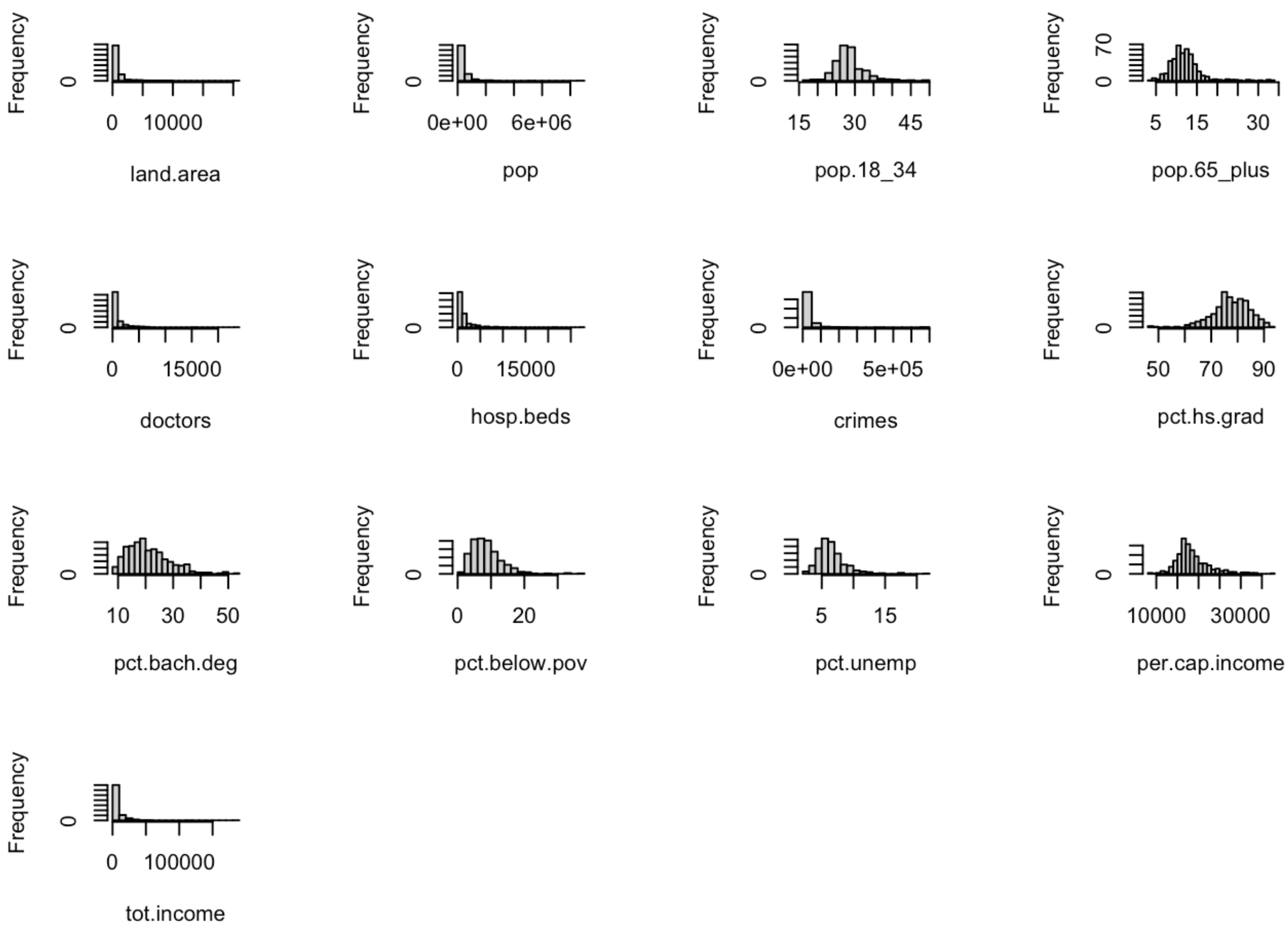
0 rows | 1-10 of 17 columns

```
# no missing data
```

There do not appear to be any missing values in the data!

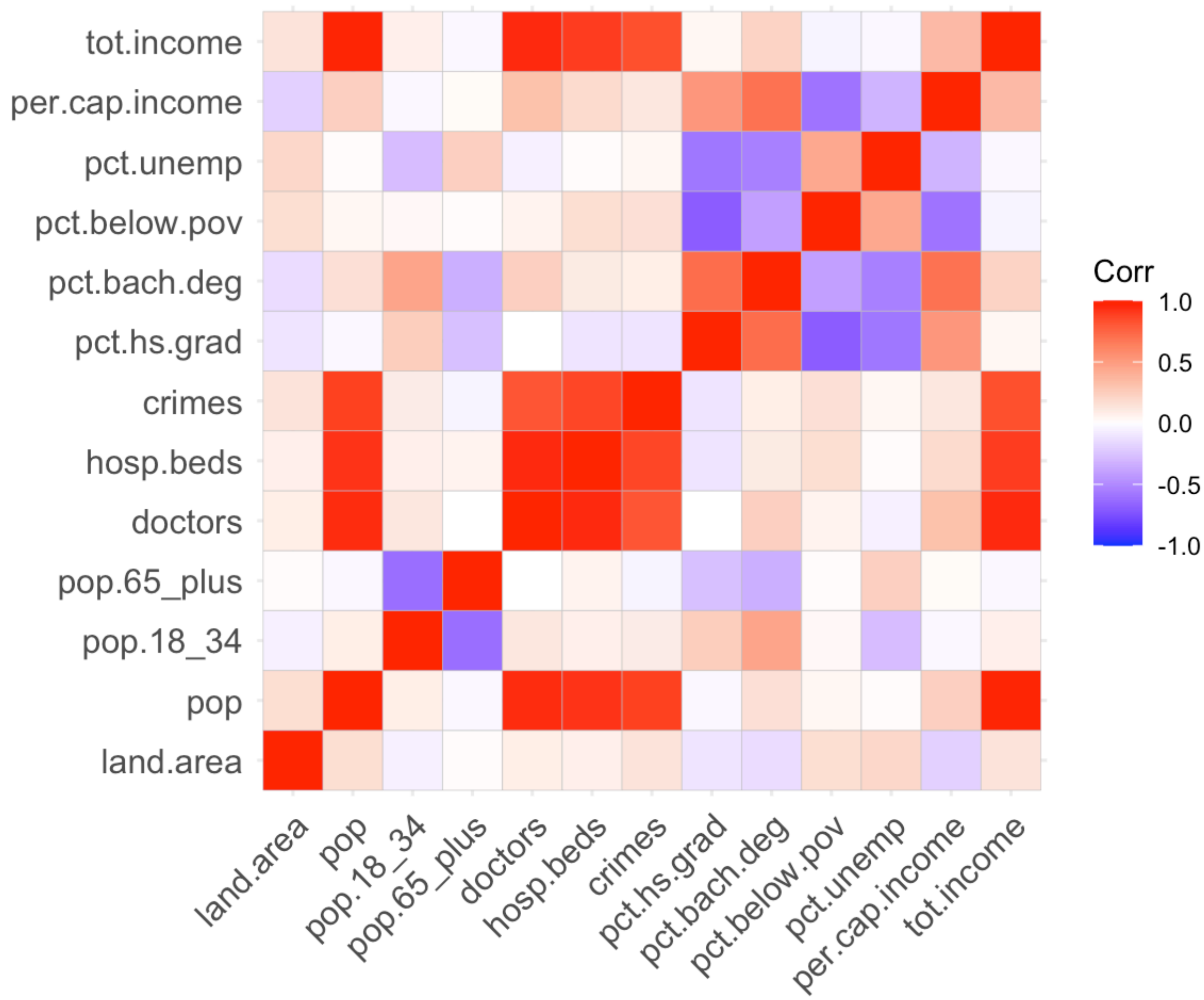
```
cdi_con <- data.frame(cdi_con)
```

```
par(mar = c(4, 4, 4, 4))
hist(cdi_con)
```

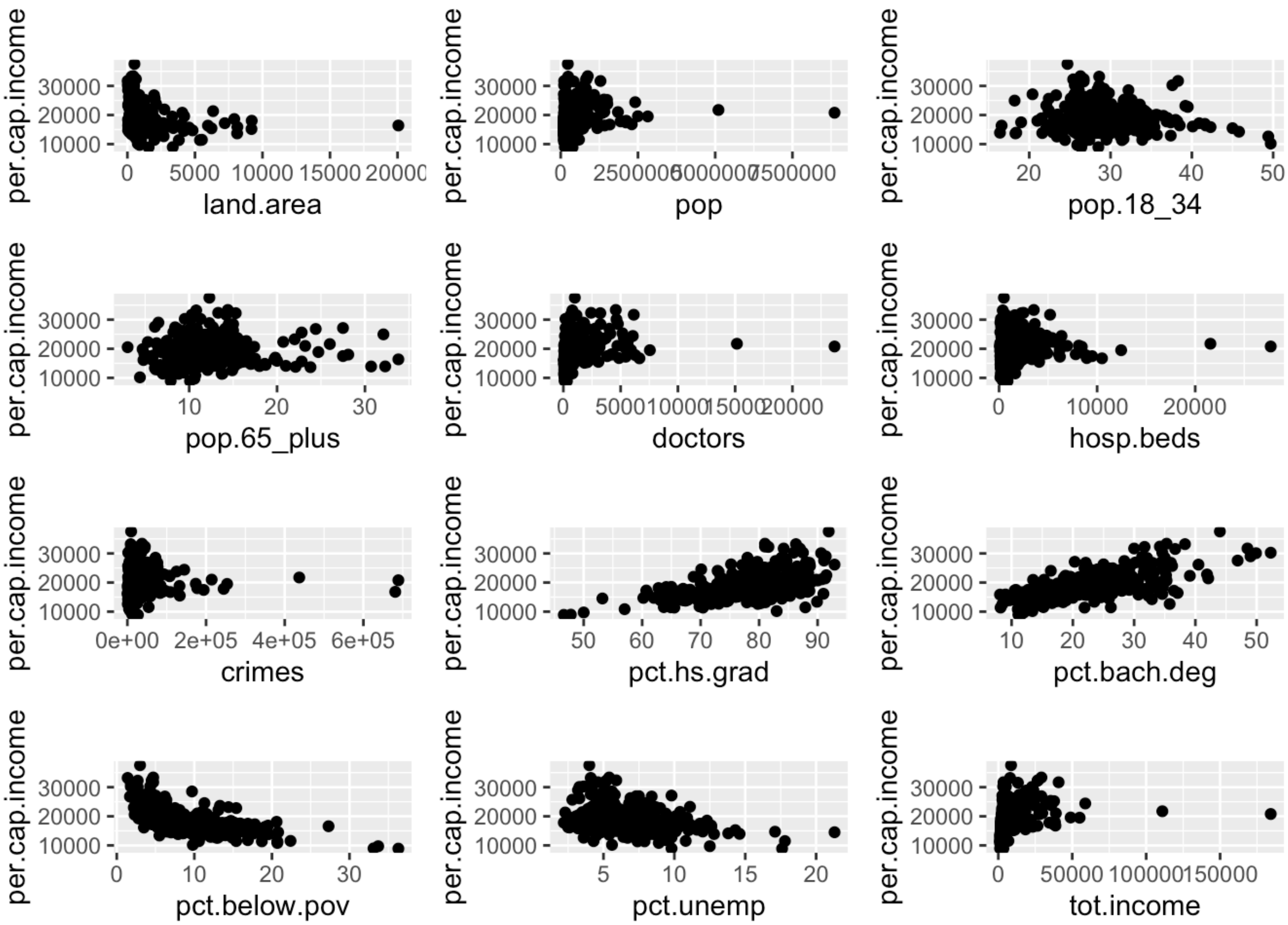


- In this situation, we did not consider id since is just the same as the row numbers of the data frame, it is not useful for data analysis. Besides, we put aside state and county first, since there are total of 48 unique values of state and 373 unique values of county, they are a lot.

```
library(ggcorrplot)
cor_matrix = cor(cdi_con)
ggcorrplot(cor_matrix)
```

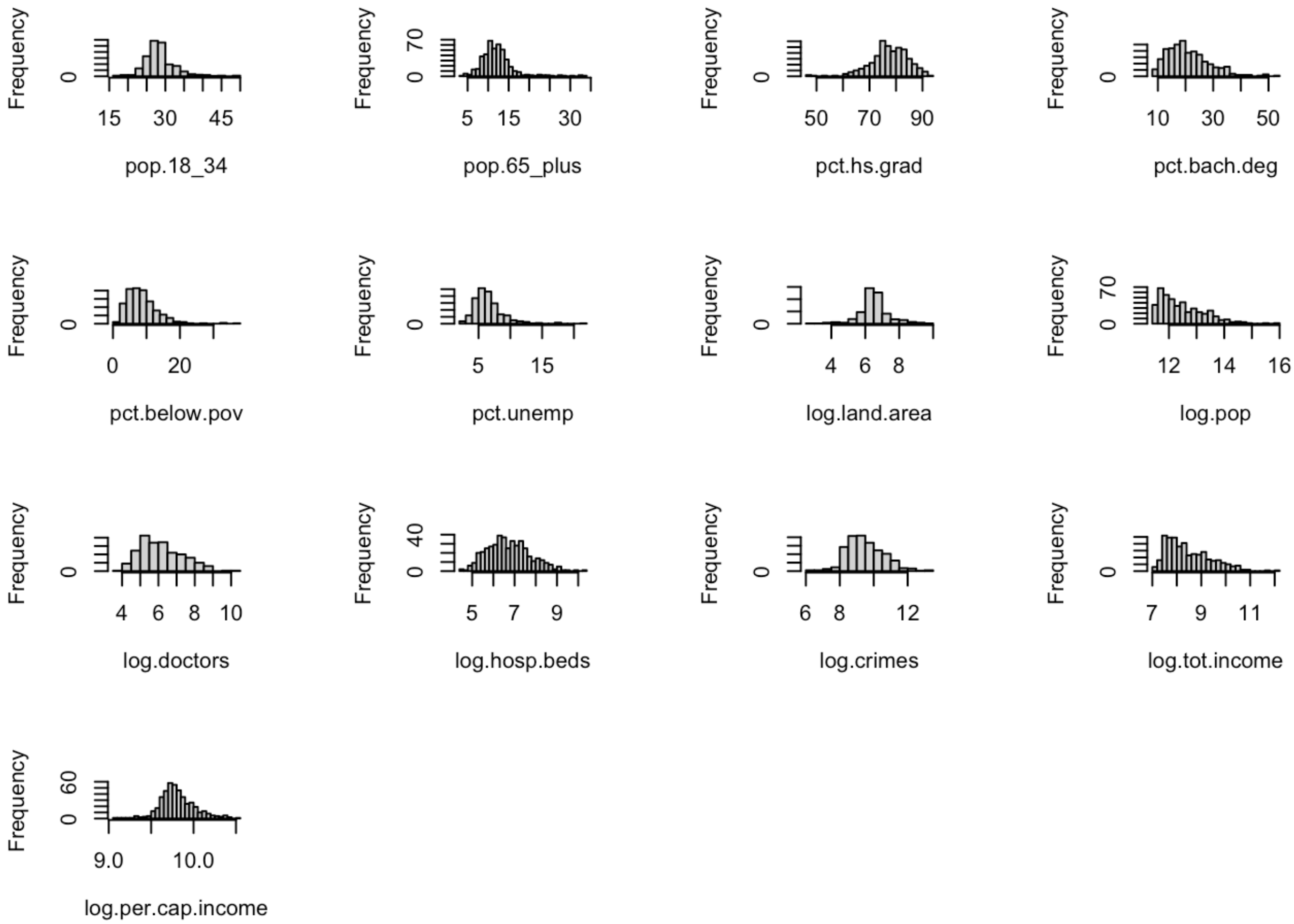



```
scatter_plot <- function(df,yvar="per.cap.income") {
    result <- NULL
    y.index <- grep(yvar,names(df))
    for (xvar in names(df)[-y.index]) {
      d <- data.frame(xx=df[,xvar],yy=df[,y.index])
      if(mode(df[,xvar])=="numeric") {
        p <- ggplot(d,aes(x=xx,y=yy)) + geom_point() +
  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

grid.arrange(grobs=scatter_plot(cdi_con))
```

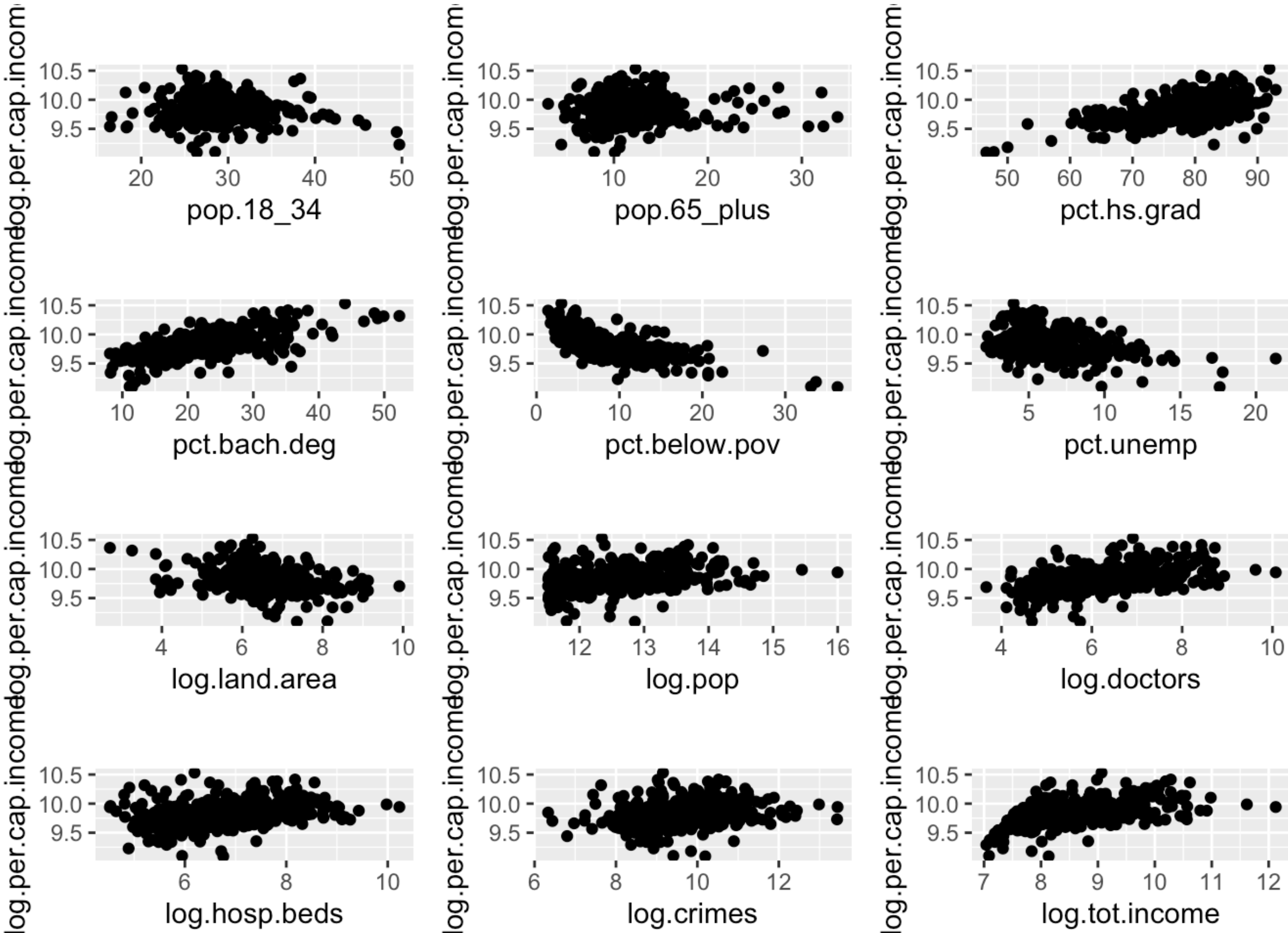



```
attach(cdi_con)
cdi2 = cdi_con

cdi2$log.land.area = log(land.area)
cdi2$log.pop = log(pop)
cdi2$log.doctors = log(doctors)
cdi2$log.hosp.beds = log(hosp.beds)
cdi2$log.crimes = log(crimes)
cdi2$log.tot.income = log(tot.income)
cdi2$log.per.cap.income = log(per.cap.income)

cdi2 = cdi2[,-c(1,2,5,6,7,12,13)]
detach()
```

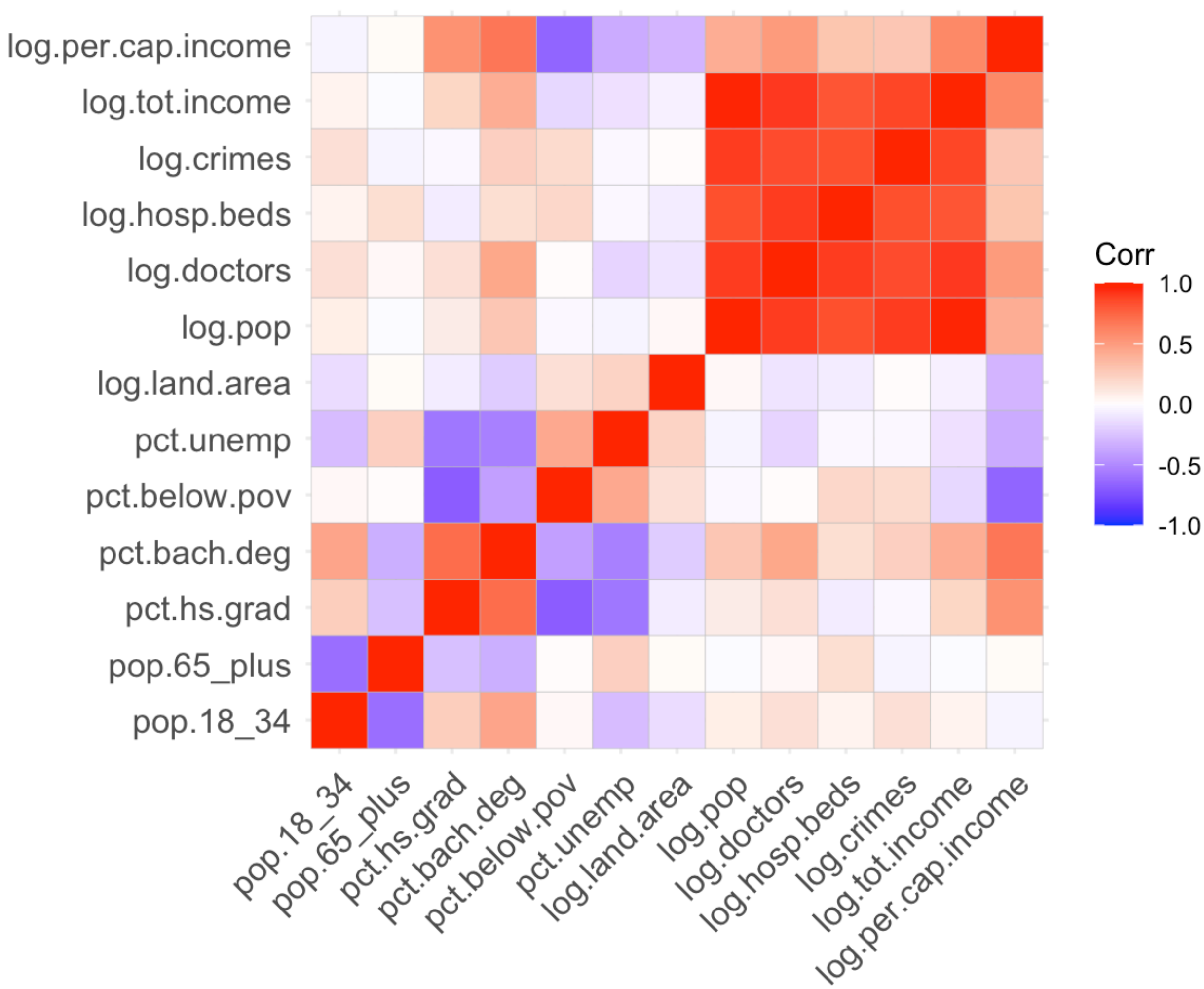
```
par(mar = c(4, 4, 4, 4))
hist(cdi2)
```




```
grid.arrange(grobs=scatter_plot(cdi2,"log.per.cap.income"))
```




```
cor_matrix = cor(cdi2)
ggcorrplot(cor_matrix)
```



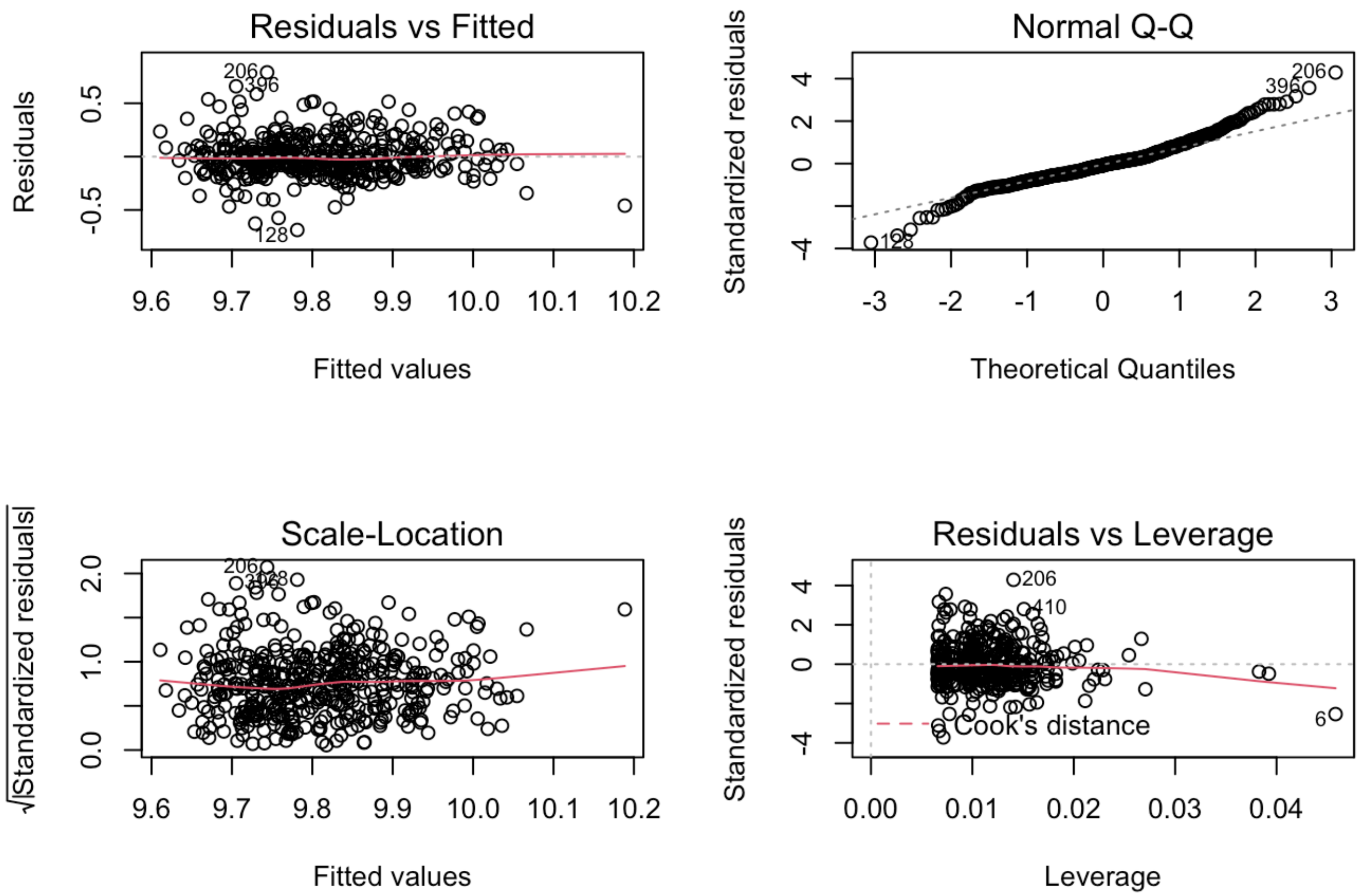
- The good news is that after we took logarithms of seven variables mentioned above, the skewing seems to be largely controlled, and the correlations are little stronger than before, more importantly, we can see from the scatter plots that linear relationships we analyzed above are also stronger than they used to be.

Appendix 2. Regression Analysis I

```
region = cdi$region
modelA <- lm(log.per.cap.income ~ log.crimes + region, data=cdi2)
summary(modelA)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.crimes + region, data = cdi2)
##
## 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
```

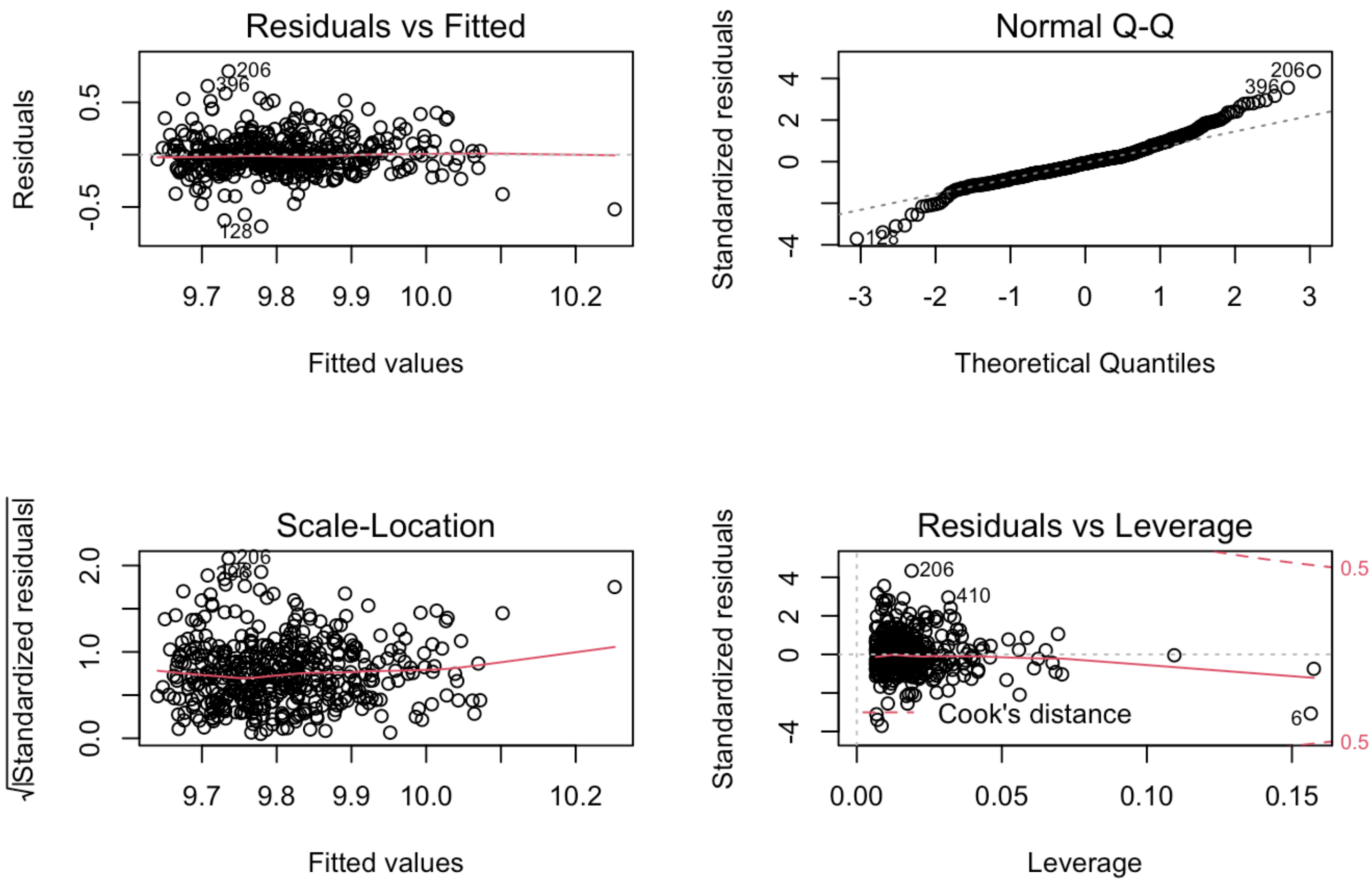
```
par(mfrow=c(2,2))
plot(modelA)
```




```
modelB <- lm(log.per.cap.income ~ log.crimes + region + log.crimes * region, data=cdi
2)
summary(modelB)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.crimes + region + log.crimes *
##     region, data = cdi2)
##
## 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
```

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




```
anova(modelA,modelB)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	435	14.94889	NA	NA	NA	NA
2	432	14.87212	3	0.07677825	0.7434092	0.5266378

2 rows

```
coef(summary(modelA))
```

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  9.18843110 0.079812437 115.125305 0.000000e+00
## log.crimes   0.06669491 0.008421114   7.919963 2.002771e-14
## regionNE     0.10445836 0.025531314   4.091382 5.110827e-05
## regionS     -0.08698350 0.023617956  -3.682939 2.595887e-04
## regionW     -0.05527965 0.028167096  -1.962561 5.033416e-02
```

We prefer Model A:

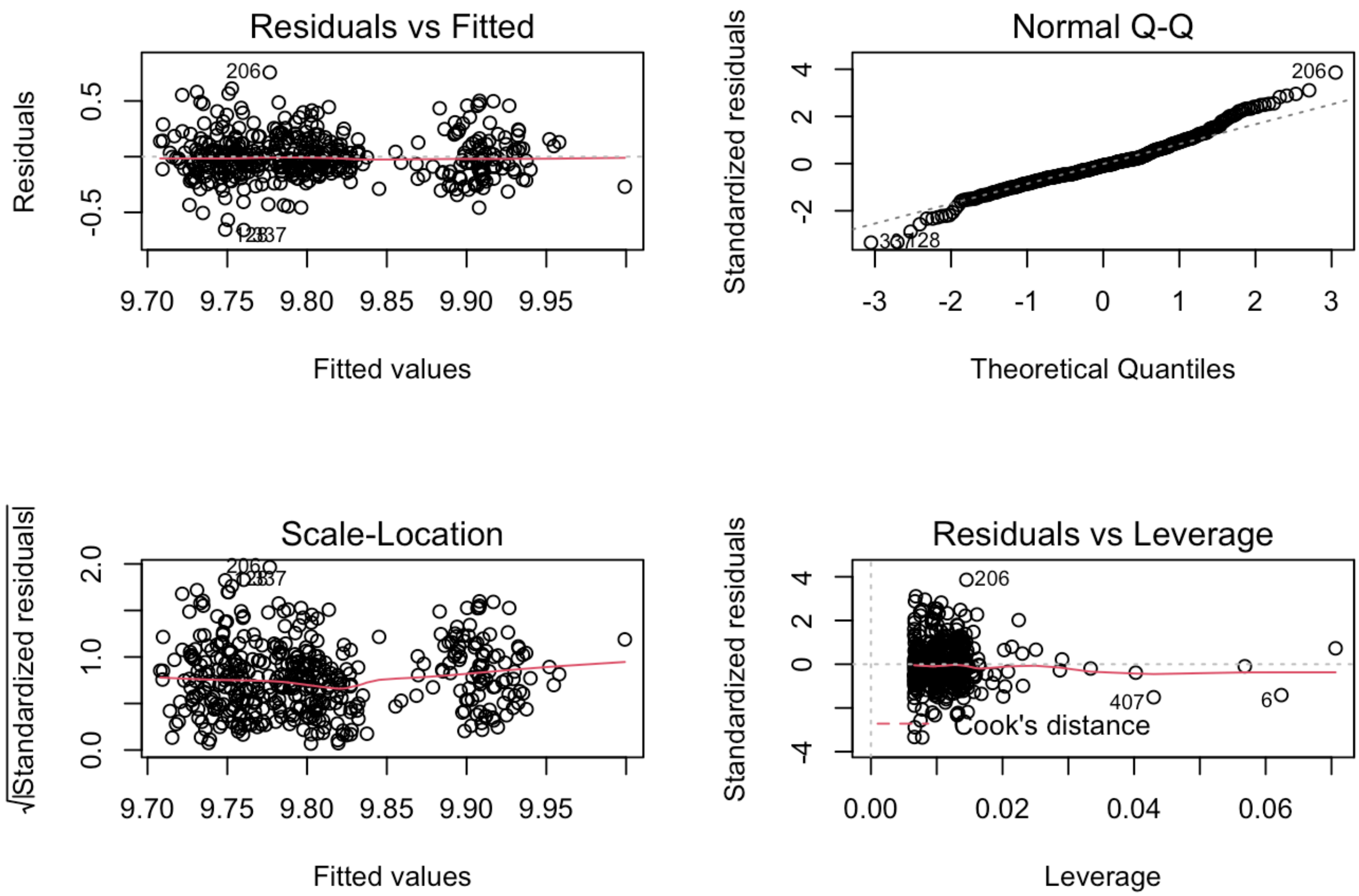
- From residual plots, none of them are awful, so we could trust F-test to compare these models:
- The P-value is 0.5266 which is larger than 0.05, there is no strong evidence against the reduced model in favour of the full model. Thus, we prefer the Model A.
- Model A had R^2_{adj}=0.1959 which is larger than Model B with R^2_{adj}=0.1945 and better residual diagnostic plots.

```
attach(cdi_con)
per.cap.crime <- crimes/pop
log.per.cap.crimes <-log(per.cap.crime)
detach()
```

```
modelC <- lm(log.per.cap.income ~ log.per.cap.crimes + region, data=cdi2)
summary(modelC)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.per.cap.crimes + region,
##     data = cdi2)
##
## 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.cap.crimes  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
```

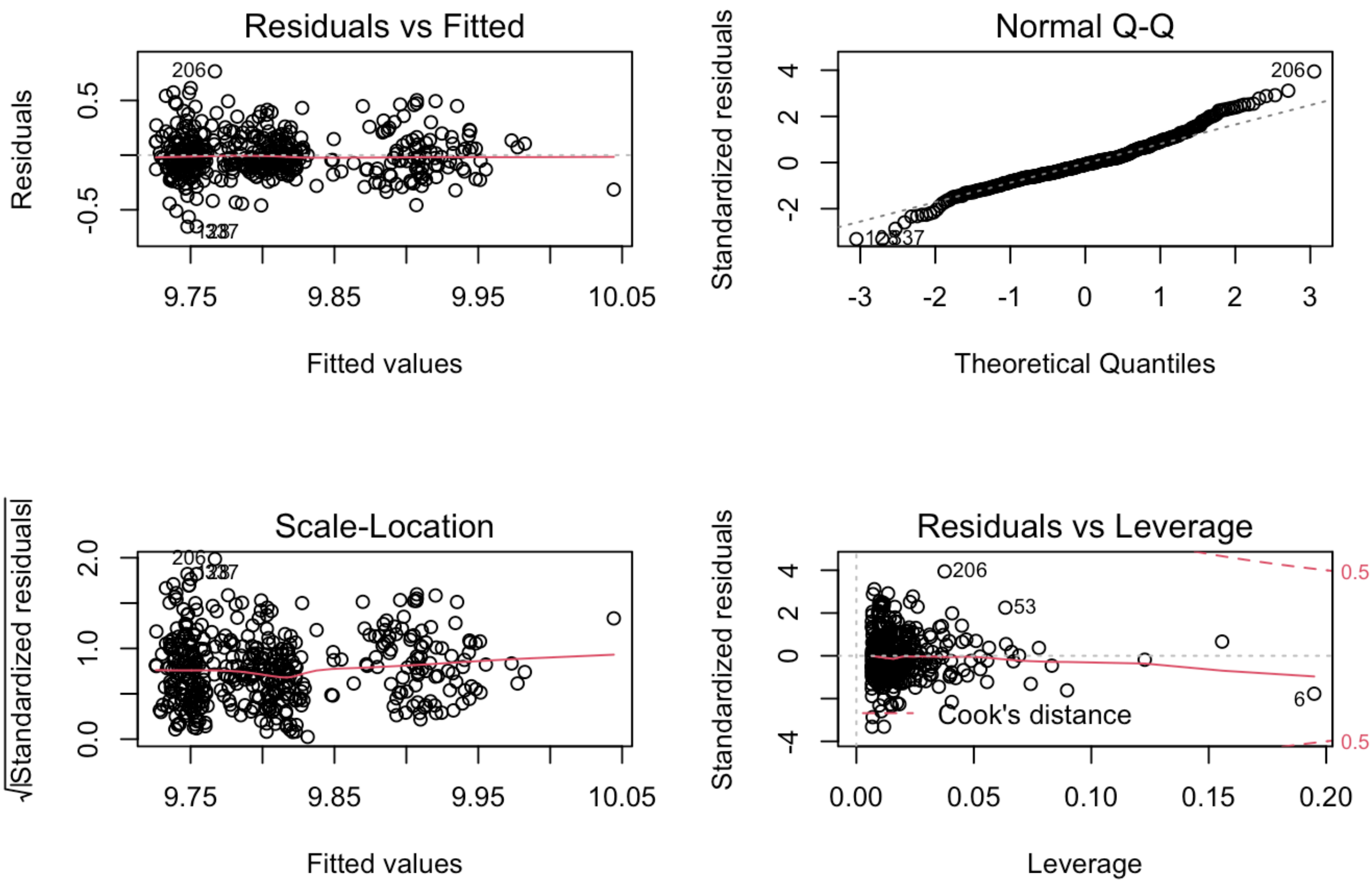
```
par(mfrow=c(2,2))
plot(modelC)
```




```
modelD <- lm(log.per.cap.income ~ log.per.cap.crimes + region + log.per.cap.crimes *
region, data=cdi2)
summary(modelD)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.per.cap.crimes + region +
##     log.per.cap.crimes * region, data = cdi2)
##
## 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.cap.crimes          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.cap.crimes:regionNE  0.02924    0.05232   0.559    0.577
## log.per.cap.crimes:regionS  -0.01104    0.05554  -0.199    0.843
## log.per.cap.crimes: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
```

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




```
anova(modelC,modelD)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	435	16.95241	NA	NA	NA	NA
2	432	16.92833	3	0.02408026	0.2048376	0.8930392

2 rows

```
coef(summary(modelC))
```

```
##                        Estimate Std. Error     t value     Pr(>|t|)
## (Intercept)         9.93628386 0.06933762 143.3029302 0.000000e+00
## log.per.cap.crimes  0.04242970 0.02147872   1.9754301 4.885103e-02
## regionNE            0.11456811 0.02760334   4.1505159 3.992234e-05
## regionS            -0.07455751 0.02624295  -2.8410493 4.707758e-03
## regionW            -0.02425540 0.03001832  -0.8080196 4.195209e-01
```

We prefer Model C:

- Again, residual plots look ok. So we could trust F-test to compare these models:
- The P-value is 0.893 which is larger than 0.05, there is no strong evidence against the reduced model in favour of the full model. Thus, we prefer the Model C.
- Besides, Model C had R^2_{adj}=0.08814 which is larger than Model D with R^2_{adj}=0.0831 and better residual diagnostic plots.

```
AIC(modelA,modelC)
```

	df <dbl>	AIC <dbl>
modelA	6	-227.4746
modelC	6	-172.1347

2 rows

```
BIC(modelA,modelC)
```

	df <dbl>	BIC <dbl>
modelA	6	-202.9539
modelC	6	-147.6140

2 rows

We compared these two winners (Model A vs. Model C), by using AIC or BIC and R_adj^2, and the output shows that Model A is the best model, by either AIC or BIC, notice that Model A (R^2_{adj}=0.1959) is also with higher than Model C (R^2_{adj}=0.08814). Thus, the best way to measure crime rate is total serious crimes.

Appendix 3. Regression Analysis II

```
attach(cdi2)
log.per.cap.doctors = log.doctors - log.pop
log.per.cap.hosp.beds = log.hosp.beds - log.pop
cdi2["log.per.cap.doctors"] = log.per.cap.doctors
cdi2["log.per.cap.hosp.beds"] = log.per.cap.hosp.beds
detach()
```

```
change <- c(grep("log.pop",names(cdi2)),grep("log.tot.income",names(cdi2)),
          grep("log.doctors",names(cdi2)), grep("log.hosp.beds",names(cdi2)))
cdi3 <- cdi2[,-change]
```

```
cdi3
```

	pop.18_34 <dbl>	pop.65_plus <dbl>	pct.hs.grad <dbl>	pct.bach.deg <dbl>	pct.below.pov <dbl>	pct.une... <dbl>	log.lan
1	32.1	9.7	70.0	22.3	11.6	8.0	8.3
2	29.2	12.4	73.4	22.8	11.1	7.2	6.8
3	31.3	7.1	74.9	25.4	12.5	5.7	7.4
4	33.5	10.9	81.9	25.3	8.1	6.1	8.3
5	32.6	9.2	81.2	27.8	5.2	4.8	6.6
6	28.3	12.4	63.7	16.6	19.5	9.5	4.2
7	29.2	12.5	81.5	22.1	8.8	4.9	9.1
8	27.4	12.5	70.0	13.7	16.9	10.0	6.4
9	27.1	13.9	65.0	18.8	14.2	8.7	7.5
10	32.6	8.2	77.1	26.3	10.4	6.1	6.7

1-10 of 440 rows | 1-8 of 12 columns Previous **1** 2 3 4 5 6 ... 44 Next

```
names(cdi3)
```

```
##  [1] "pop.18_34"             "pop.65_plus"           "pct.hs.grad"
##  [4] "pct.bach.deg"          "pct.below.pov"         "pct.unemp"
##  [7] "log.land.area"         "log.crimes"            "log.per.cap.income"
## [10] "log.per.cap.doctors"   "log.per.cap.hosp.beds"
```

Appendix 3.1. All subsets:

```
library(leaps)
library(car)
```

```
model1 <- regsubsets(log.per.cap.income ~ ., data=cdi3, nvmax=10)
model1.summary <- summary(model1)
print(best.model <- which.min(model1.summary$bic))
```

```
## [1] 9
```

Model1 shows that we got final model from all subsets method with nine predictors. All the predictors have significantly different from zero.

```
tmp <- cdi3[,model1.summary$which[best.model,][-1]]
model1 <- lm(log.per.cap.income ~ .,data=tmp)
summary(model1)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ ., data = tmp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34745 -0.04808 -0.00580  0.04953  0.24686
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           10.5871973  0.1313299  80.615  < 2e-16 ***
## pop.18_34             -0.0164619  0.0013620 -12.087  < 2e-16 ***
## pop.65_plus           -0.0033888  0.0014657  -2.312   0.0212 *
## pct.hs.grad           -0.0053850  0.0011214  -4.802 2.17e-06 ***
## pct.bach.deg           0.0184675  0.0008953  20.628  < 2e-16 ***
## pct.below.pov         -0.0280718  0.0014295 -19.637  < 2e-16 ***
## pct.unemp              0.0127977  0.0023075   5.546 5.11e-08 ***
## log.land.area         -0.0328529  0.0050768  -6.471 2.65e-10 ***
## log.crimes             0.0475823  0.0041507  11.464  < 2e-16 ***
## log.per.cap.hosp.beds  0.0552693  0.0094950   5.821 1.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08448 on 430 degrees of freedom
## Multiple R-squared:  0.8364, Adjusted R-squared:  0.833
## F-statistic: 244.3 on 9 and 430 DF,  p-value: < 2.2e-16
```

```
model1.summary$which[best.model,]
```

```
##           (Intercept)             pop.18_34           pop.65_plus
##                  TRUE                  TRUE                  TRUE
##           pct.hs.grad          pct.bach.deg         pct.below.pov
##                  TRUE                  TRUE                  TRUE
##             pct.unemp         log.land.area            log.crimes
##                  TRUE                  TRUE                  TRUE
##   log.per.cap.doctors log.per.cap.hosp.beds
##                  TRUE                 FALSE
```

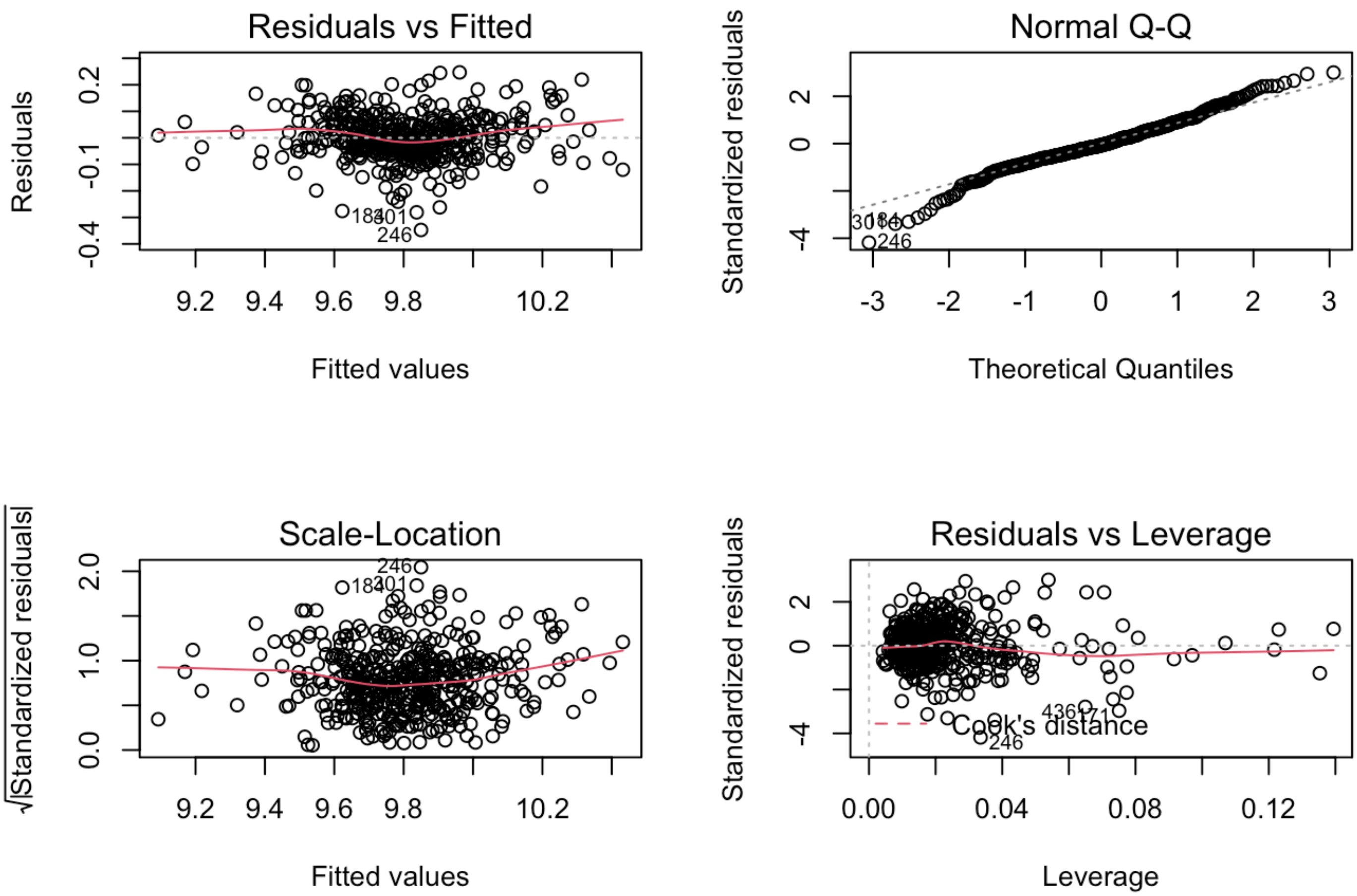
```
tmp <- cdi3[,model1.summary$which[best.model,][-1]]
model1 <- lm(log.per.cap.income ~ .,data=tmp)
summary(model1)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ ., data = tmp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34745 -0.04808 -0.00580  0.04953  0.24686
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           10.5871973  0.1313299  80.615  < 2e-16 ***
## pop.18_34             -0.0164619  0.0013620 -12.087  < 2e-16 ***
## pop.65_plus           -0.0033888  0.0014657  -2.312   0.0212 *
## pct.hs.grad           -0.0053850  0.0011214  -4.802 2.17e-06 ***
## pct.bach.deg           0.0184675  0.0008953  20.628  < 2e-16 ***
## pct.below.pov         -0.0280718  0.0014295 -19.637  < 2e-16 ***
## pct.unemp              0.0127977  0.0023075   5.546 5.11e-08 ***
## log.land.area         -0.0328529  0.0050768  -6.471 2.65e-10 ***
## log.crimes             0.0475823  0.0041507  11.464  < 2e-16 ***
## log.per.cap.hosp.beds  0.0552693  0.0094950   5.821 1.15e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08448 on 430 degrees of freedom
## Multiple R-squared:  0.8364, Adjusted R-squared:  0.833
## F-statistic: 244.3 on 9 and 430 DF,  p-value: < 2.2e-16
```

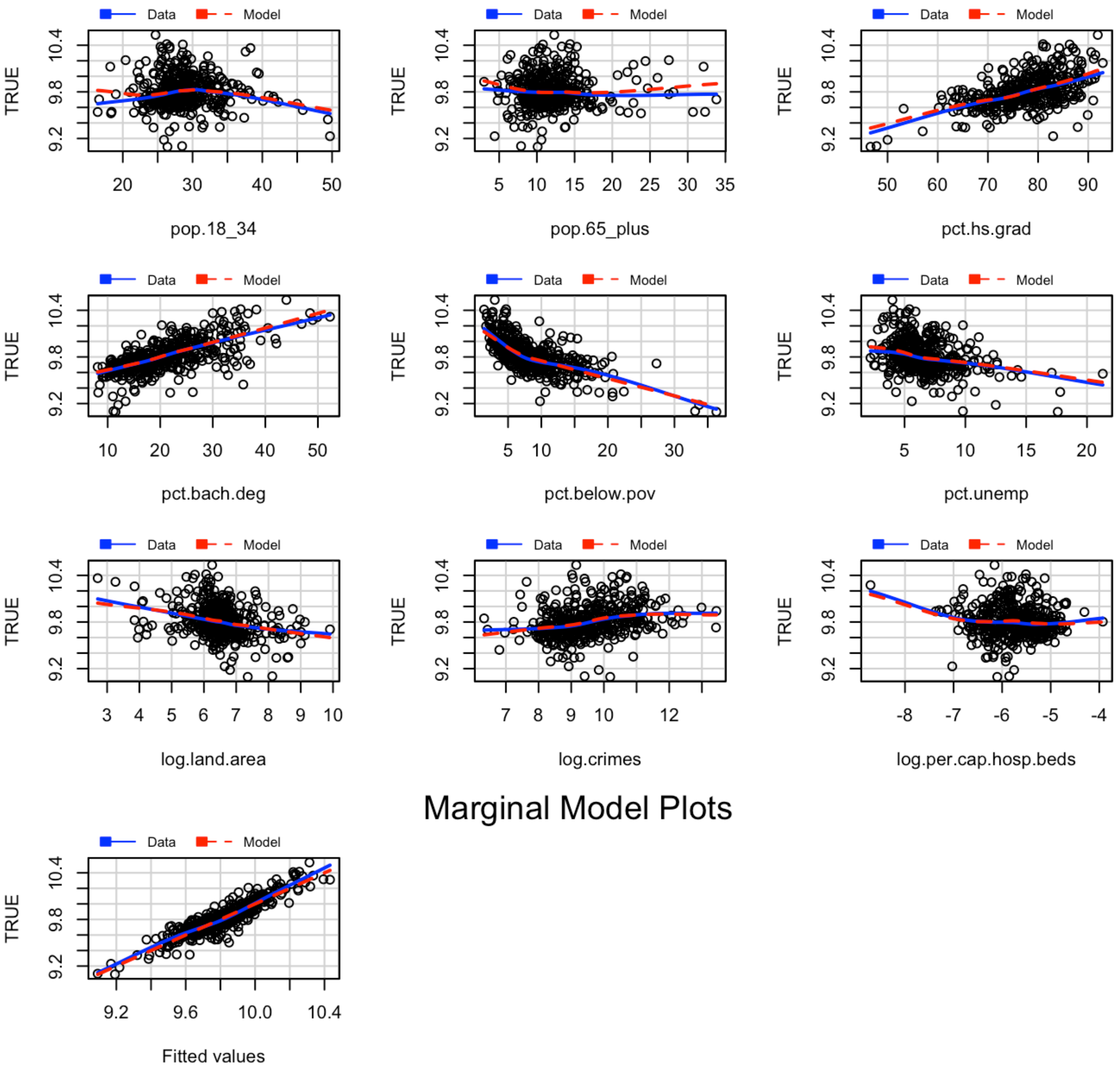
```
vif(model1)
```

```
##             pop.18_34           pop.65_plus           pct.hs.grad
##              2.004378              2.106824              3.807054
##          pct.bach.deg         pct.below.pov             pct.unemp
##              2.888796              2.726057              1.790244
##         log.land.area            log.crimes log.per.cap.hosp.beds
##              1.204756              1.241045              1.682366
```

```
par(mfrow=c(2,2))
plot(model1)
```


Residuals vs Fitted
Residuals
Fitted values
184
401
246
Normal Q-Q
Standardized residuals
Theoretical Quantiles
301
184
246
Scale-Location
√|Standardized residuals|
Fitted values
246
184
401
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
436
171
246

```
mmps(model1)
```


Data
Model
TRUE
pop.18_34
pop.65_plus
pct.hs.grad
pct.bach.deg
pct.below.pov
pct.unemp
log.land.area
log.crimes
log.per.cap.hosp.beds
Marginal Model Plots
Fitted values

Model1:

- Based on output of VIF test and residual diagnostics plots, none of the VIFs seem excessively large (>5).
- The diagnostic plots don't show much except that the QQ plot suggests left bottom tails are a bit longer than expected for the normal distribution.
- We can also check for marginal model plots：The nonparametric estimates of each pair-wise relationship are marked as solid curves, while the smooths of the fitted values are marked as dashed curves, we see that there is very good agreement between the two sets of fits in Figure above. Thus, at this stage we shall continue under the assumption that model1 is a valid model. Hence we are not missing any important transformations, interactions.

```
tmp <- cbind(tmp,region=cdi$region)
model2 <- lm(log.per.cap.income ~ .  *region,data=tmp)
summary(model2)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ . * region, data = tmp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23397 -0.04171 -0.00361  0.03726  0.33097
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  10.2429515  0.3786718  27.050  < 2e-16 ***
## pop.18_34                    -0.0168359  0.0028736  -5.859 9.75e-09 ***
## pop.65_plus                   0.0018837  0.0055522   0.339 0.734592
## pct.hs.grad                  -0.0054352  0.0034482  -1.576 0.115767
## pct.bach.deg                  0.0188517  0.0025344   7.438 6.27e-13 ***
## pct.below.pov                -0.0220087  0.0041772  -5.269 2.25e-07 ***
## pct.unemp                     0.0131414  0.0054321   2.419 0.015999 *
## log.land.area                -0.0316165  0.0158436  -1.996 0.046662 *
## log.crimes                    0.0416444  0.0080812   5.153 4.03e-07 ***
## log.per.cap.hosp.beds         0.0029689  0.0191249   0.155 0.876713
## regionNE                      0.4063638  0.5044450   0.806 0.420972
## regionS                       0.0106892  0.4186595   0.026 0.979643
```

```
## regionW                          1.7835135  0.5597231   3.186 0.001553 **
## pop.18_34:regionNE              -0.0039839  0.0042798  -0.931 0.352487
## pop.18_34:regionS               -0.0003271  0.0035084  -0.093 0.925755
## pop.18_34:regionW                0.0057118  0.0048329   1.182 0.237959
## pop.65_plus:regionNE            -0.0081761  0.0068786  -1.189 0.235287
## pop.65_plus:regionS             -0.0030926  0.0058488  -0.529 0.597264
## pop.65_plus:regionW             -0.0017433  0.0071757  -0.243 0.808175
## pct.hs.grad:regionNE            -0.0004295  0.0044627  -0.096 0.923383
## pct.hs.grad:regionS              0.0048379  0.0038268   1.264 0.206892
## pct.hs.grad:regionW             -0.0167689  0.0046661  -3.594 0.000367 ***
## pct.bach.deg:regionNE            0.0037752  0.0036167   1.044 0.297200
## pct.bach.deg:regionS            -0.0044228  0.0028387  -1.558 0.120024
## pct.bach.deg:regionW             0.0038663  0.0033317   1.160 0.246553
## pct.below.pov:regionNE          -0.0052012  0.0057023  -0.912 0.362255
## pct.below.pov:regionS            0.0039897  0.0047176   0.846 0.398233
## pct.below.pov:regionW           -0.0230691  0.0059122  -3.902 0.000112 ***
## pct.unemp:regionNE              -0.0012360  0.0079253  -0.156 0.876141
## pct.unemp:regionS               -0.0196719  0.0073481  -2.677 0.007731 **
## pct.unemp:regionW               -0.0141753  0.0072644  -1.951 0.051715 .
## log.land.area:regionNE          -0.0012015  0.0208317  -0.058 0.954034
## log.land.area:regionS           -0.0132422  0.0182744  -0.725 0.469103
## log.land.area:regionW            0.0138966  0.0190266   0.730 0.465586
## log.crimes:regionNE              0.0069092  0.0125217   0.552 0.581408
## log.crimes:regionS               0.0100411  0.0109961   0.913 0.361712
## log.crimes:regionW              -0.0072645  0.0129979  -0.559 0.576541
## log.per.cap.hosp.beds:regionNE  0.0424622  0.0327132   1.298 0.195030
## log.per.cap.hosp.beds:regionS   0.0380187  0.0237547   1.600 0.110284
## log.per.cap.hosp.beds:regionW   0.0699933  0.0312524   2.240 0.025665 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07659 on 400 degrees of freedom
## Multiple R-squared:  0.875,  Adjusted R-squared:  0.8628
## F-statistic: 71.76 on 39 and 400 DF,  p-value: < 2.2e-16
```

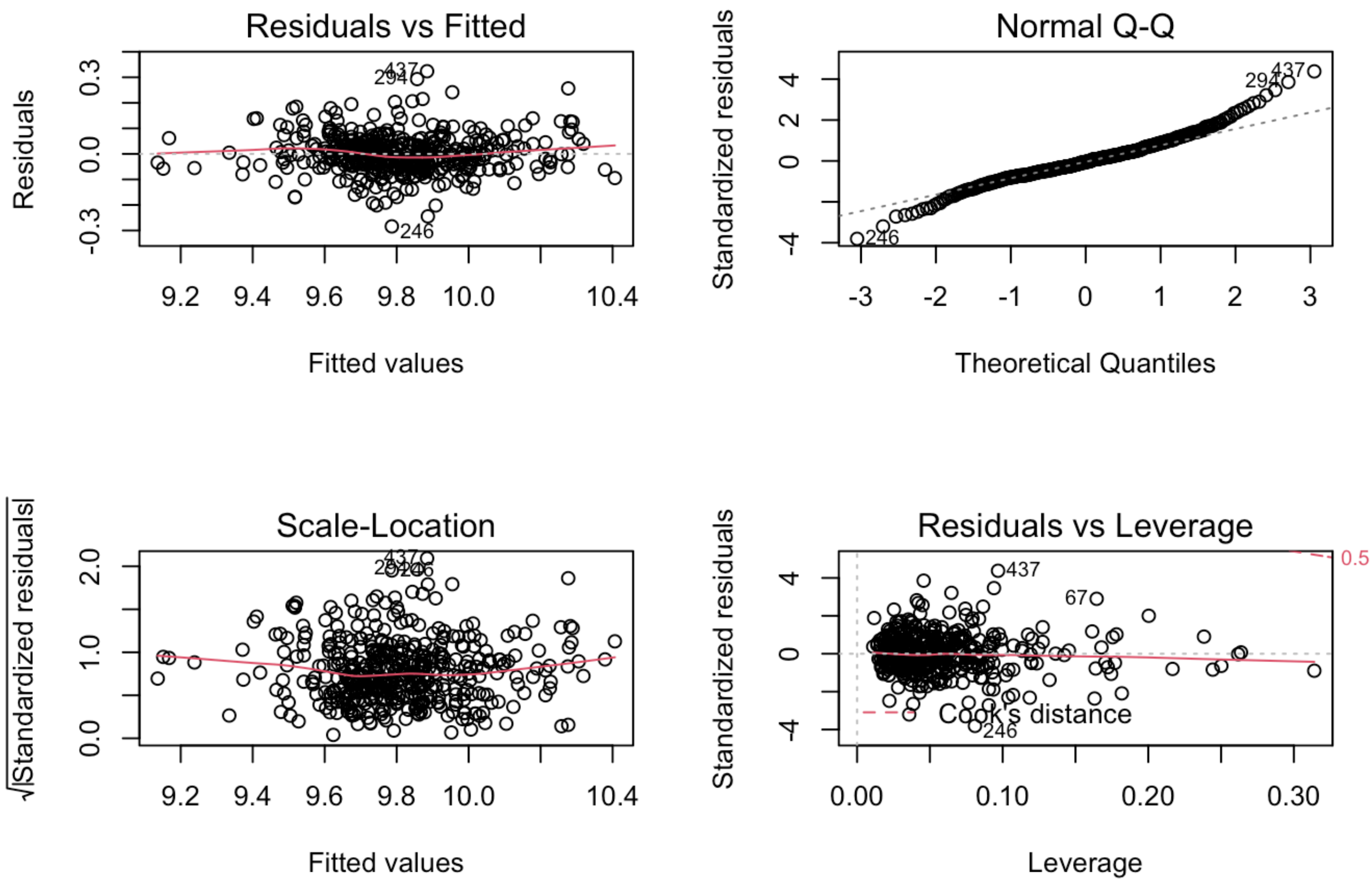
```
model2 <- update(model2,. ~ . - region:log.land.area -
                   region:pop.18_34 - region:pop.65_plus -
                   region:log.crimes - region:pct.bach.deg)
summary(model2)
```

```
##
## 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.crimes + log.per.cap.hosp.beds + region + pct.hs.grad:region +
##     pct.below.pov:region + pct.unemp:region + log.per.cap.hosp.beds:region,
##     data = tmp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28398 -0.04361 -0.00358  0.03845  0.32371
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    10.2968941  0.2584674  39.838  < 2e-16 ***
## pop.18_34                      -0.0159934  0.0013109 -12.201  < 2e-16 ***
## pop.65_plus                    -0.0009692  0.0015013  -0.646 0.518910
## pct.hs.grad                    -0.0058902  0.0024860  -2.369 0.018278 *
## pct.bach.deg                    0.0180357  0.0009314  19.365  < 2e-16 ***
## pct.below.pov                  -0.0230162  0.0039675  -5.801 1.31e-08 ***
## pct.unemp                       0.0126859  0.0049170   2.580 0.010223 *
## log.land.area                  -0.0361999  0.0055526  -6.519 2.05e-10 ***
## log.crimes                      0.0485077  0.0041196  11.775  < 2e-16 ***
## log.per.cap.hosp.beds           0.0058340  0.0170453   0.342 0.732327
## regionNE                        0.0133618  0.3228946   0.041 0.967012
## regionS                         0.0866043  0.2740041   0.316 0.752109
## regionW                         1.7063132  0.3788133   4.504 8.67e-06 ***
## pct.hs.grad:regionNE            0.0035491  0.0029754   1.193 0.233634
## pct.hs.grad:regionS             0.0021277  0.0025963   0.820 0.412964
## pct.hs.grad:regionW            -0.0134941  0.0036328  -3.715 0.000231 ***
## pct.below.pov:regionNE         -0.0057364  0.0054144  -1.059 0.290002
## pct.below.pov:regionS           0.0026802  0.0044069   0.608 0.543390
## pct.below.pov:regionW          -0.0202029  0.0057138  -3.536 0.000452 ***
## pct.unemp:regionNE             -0.0051337  0.0074663  -0.688 0.492103
## pct.unemp:regionS              -0.0148402  0.0067267  -2.206 0.027920 *
## pct.unemp:regionW              -0.0138082  0.0068760  -2.008 0.045274 *
## log.per.cap.hosp.beds:regionNE  0.0359182  0.0293619   1.223 0.221912
## log.per.cap.hosp.beds:regionS   0.0407353  0.0209645   1.943 0.052685 .
## log.per.cap.hosp.beds:regionW   0.0616266  0.0274782   2.243 0.025440 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07779 on 415 degrees of freedom
## Multiple R-squared:  0.8661, Adjusted R-squared:  0.8584
## F-statistic: 111.9 on 24 and 415 DF,  p-value: < 2.2e-16
```

```
vif(model2)
```

```
##                                     GVIF Df GVIF^(1/(2*Df))
## pop.18_34                   2.189669e+00  1        1.479753
## pop.65_plus                 2.606491e+00  1        1.614463
## pct.hs.grad                 2.206442e+01  1        4.697278
## pct.bach.deg                3.687123e+00  1        1.920188
## pct.below.pov               2.476362e+01  1        4.976306
## pct.unemp                   9.586823e+00  1        3.096260
## log.land.area               1.699594e+00  1        1.303685
## log.crimes                  1.441699e+00  1        1.200708
## log.per.cap.hosp.beds       6.393792e+00  1        2.528595
## region                      5.577172e+08  3       28.690339
## pct.hs.grad:region          8.683300e+07  3       21.043314
## pct.below.pov:region        1.054253e+04  3        4.682640
## pct.unemp:region            1.531130e+04  3        4.983128
## log.per.cap.hosp.beds:region 8.722892e+06  3       14.347517
```

```
par(mfrow=c(2,2))
plot(model2)
```



Model2:

- Based on output of VIF test and residual diagnostics plots, none of the VIFs of single variables seem excessively large (>5). But we have interaction terms with large VIFs.
- The diagnostic plots are much better especially that the QQ plot suggests left bottom tails are fixed for the normal distribution.

```
anova(model1,model2)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	430	3.068649	NA	NA	NA	NA
2	415	2.511340	15	0.5573092	6.139706	1.062341e-11

2 rows

```
AIC(model1,model2)
```

	df <dbl>	AIC <dbl>
model1	11	-914.1705
model2	26	-972.3557

2 rows

```
BIC(model1,model2)
```

	df <dbl>	BIC <dbl>
model1	11	-869.2160
model2	26	-866.0996

2 rows

All subsets prefer Model2:

Both of the ANOVA F test and AIC prefer Model2. On the other hand, BIC prefers the simpler model. BIC chooses a simpler model than other methods since of larger penalty, however BIC prefers simpler model rather than a highly predictive model. Besides Model2 had R^2_{adj}=0.8584 which is larger than Model1 with R^2_{adj}=0.833 and better residual diagnostic plots. Based above, all subsets method selects Model2 as a better model.

Appendix 3.2. Stepwise Regression:

```
library(MASS)
```

```
n <- 440
base <- lm(log.per.cap.income ~ .,data=cdi3)
summary(base)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ ., data = cdi3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33791 -0.04523 -0.00547  0.04854  0.24619
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           10.919833   0.142991  76.367  < 2e-16 ***
## pop.18_34             -0.016418   0.001323 -12.408  < 2e-16 ***
## pop.65_plus           -0.004318   0.001435  -3.008  0.00278 **
## pct.hs.grad           -0.005115   0.001091  -4.689 3.69e-06 ***
## pct.bach.deg           0.015539   0.001039  14.957  < 2e-16 ***
## pct.below.pov         -0.028166   0.001389 -20.278  < 2e-16 ***
## pct.unemp              0.013431   0.002245   5.982 4.64e-09 ***
## log.land.area         -0.031414   0.004940  -6.359 5.21e-10 ***
## log.crimes             0.039623   0.004318   9.176  < 2e-16 ***
## log.per.cap.doctors    0.071061   0.013789   5.153 3.91e-07 ***
## log.per.cap.hosp.beds  0.015209   0.012063   1.261  0.20807
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08207 on 429 degrees of freedom
## Multiple R-squared:  0.846,  Adjusted R-squared:  0.8424
## F-statistic: 235.6 on 10 and 429 DF,  p-value: < 2.2e-16
```

```
# backward elimination based on AIC
backAIC <- step(base,direction="backward", data=cdi3)
```

```
## Start:  AIC=-2189.26
## 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.crimes + log.per.cap.doctors + log.per.cap.hosp.beds
##
##                          Df Sum of Sq    RSS     AIC
## - log.per.cap.hosp.beds  1   0.01071 2.9005 -2189.6
## <none>                               2.8898 -2189.3
## - pop.65_plus            1   0.06097 2.9507 -2182.1
## - pct.hs.grad            1   0.14812 3.0379 -2169.3
## - log.per.cap.doctors    1   0.17889 3.0686 -2164.8
## - pct.unemp              1   0.24107 3.1308 -2156.0
## - log.land.area          1   0.27237 3.1621 -2151.6
## - log.crimes             1   0.56713 3.4569 -2112.4
## - pop.18_34              1   1.03700 3.9268 -2056.3
## - pct.bach.deg           1   1.50695 4.3967 -2006.6
## - pct.below.pov          1   2.76985 5.6596 -1895.5
##
## Step:  AIC=-2189.64
## 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.crimes + log.per.cap.doctors
##
##                        Df Sum of Sq    RSS     AIC
## <none>                             2.9005 -2189.6
## - pop.65_plus          1   0.05199 2.9525 -2183.8
## - pct.hs.grad          1   0.14356 3.0440 -2170.4
## - pct.unemp            1   0.23148 3.1320 -2157.8
## - log.land.area        1   0.29249 3.1930 -2149.4
## - log.per.cap.doctors  1   0.40998 3.3104 -2133.5
## - log.crimes           1   0.55659 3.4571 -2114.4
## - pop.18_34            1   1.02717 3.9276 -2058.2
## - pct.bach.deg         1   1.66340 4.5639 -1992.2
## - pct.below.pov        1   2.91729 5.8178 -1885.4
```

```
# backward elimination based on BIC
backBIC <- step(base,direction="backward", data=cdi3, k=log(n))
```

```
## Start:  AIC=-2144.31
## 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.crimes + log.per.cap.doctors + log.per.cap.hosp.beds
##
##                         Df Sum of Sq    RSS     AIC
## - log.per.cap.hosp.beds  1   0.01071 2.9005 -2148.8
## <none>                               2.8898 -2144.3
## - pop.65_plus            1   0.06097 2.9507 -2141.2
## - pct.hs.grad            1   0.14812 3.0379 -2128.4
## - log.per.cap.doctors    1   0.17889 3.0686 -2124.0
## - pct.unemp              1   0.24107 3.1308 -2115.1
## - log.land.area          1   0.27237 3.1621 -2110.8
## - log.crimes             1   0.56713 3.4569 -2071.6
## - pop.18_34              1   1.03700 3.9268 -2015.5
## - pct.bach.deg           1   1.50695 4.3967 -1965.7
## - pct.below.pov          1   2.76985 5.6596 -1854.6
##
## Step:  AIC=-2148.77
## 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.crimes + log.per.cap.doctors
##
##                       Df Sum of Sq    RSS     AIC
## <none>                             2.9005 -2148.8
## - pop.65_plus          1   0.05199 2.9525 -2147.0
## - pct.hs.grad          1   0.14356 3.0440 -2133.6
## - pct.unemp            1   0.23148 3.1320 -2121.1
## - log.land.area        1   0.29249 3.1930 -2112.6
## - log.per.cap.doctors  1   0.40998 3.3104 -2096.7
## - log.crimes           1   0.55659 3.4571 -2077.6
## - pop.18_34            1   1.02717 3.9276 -2021.5
## - pct.bach.deg         1   1.66340 4.5639 -1955.4
## - pct.below.pov        1   2.91729 5.8178 -1848.6
```

```
# forward selection based on AIC
mint <- base
forwardAIC <- step(mint,scope=list(lower=~1,
upper=~log.land.area + log.per.cap.doctors
          + log.per.cap.hosp.beds + pop.18_34 + pop.65_plus
          + pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
          + log.crimes),
direction="forward", data=cdi3)
```

```
## Start:  AIC=-2189.26
## 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.crimes + log.per.cap.doctors + log.per.cap.hosp.beds
```

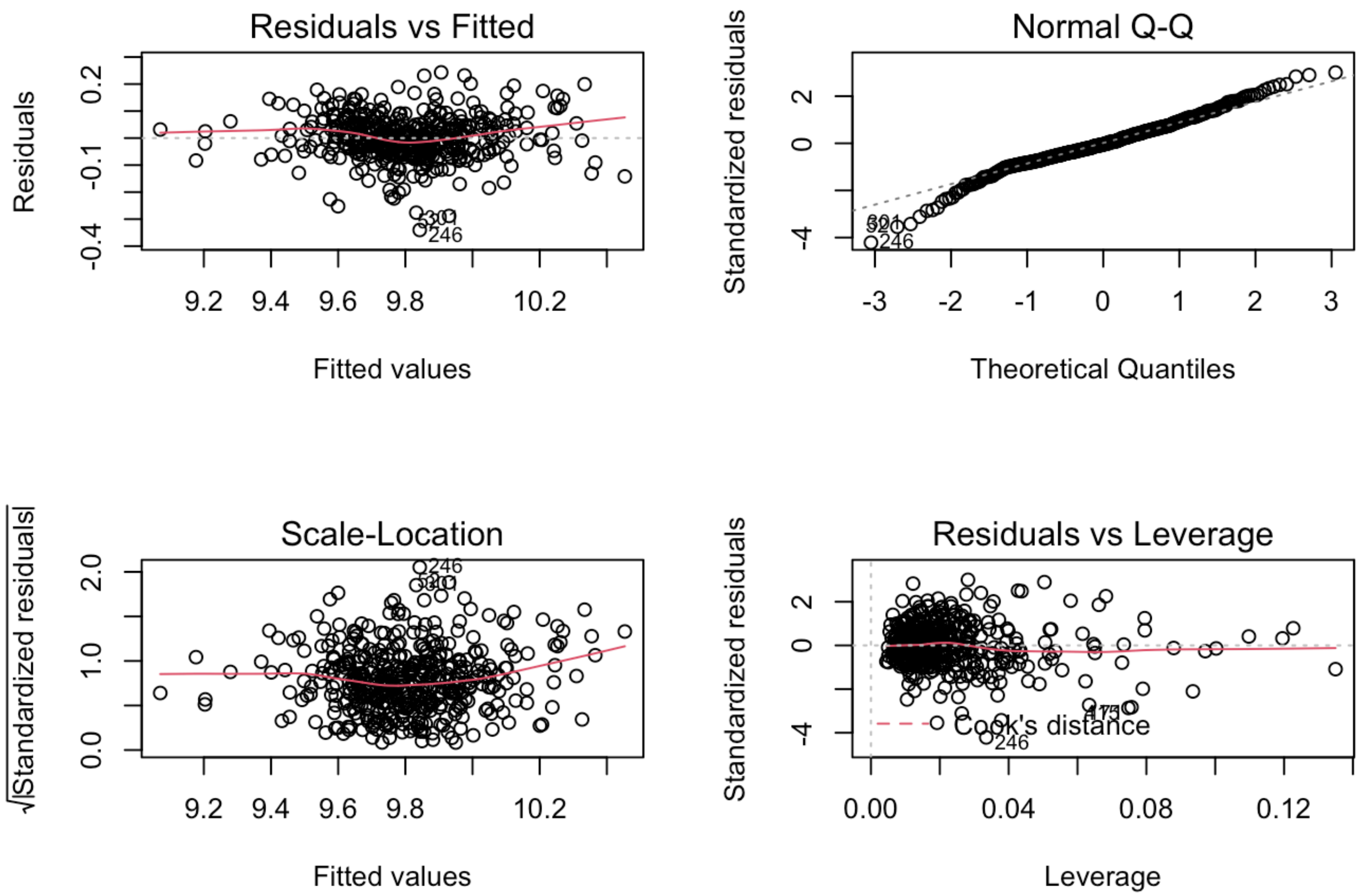
```
# forward selection based on BIC
forwardBIC <- step(mint,scope=list(lower=~1,
upper=~log.land.area + log.per.cap.doctors
          + log.per.cap.hosp.beds + pop.18_34 + pop.65_plus
          + pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
          + log.crimes),
direction="forward", data=cdi3,k=log(n))
```

```
## Start:  AIC=-2144.31
## 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.crimes + log.per.cap.doctors + log.per.cap.hosp.beds
```

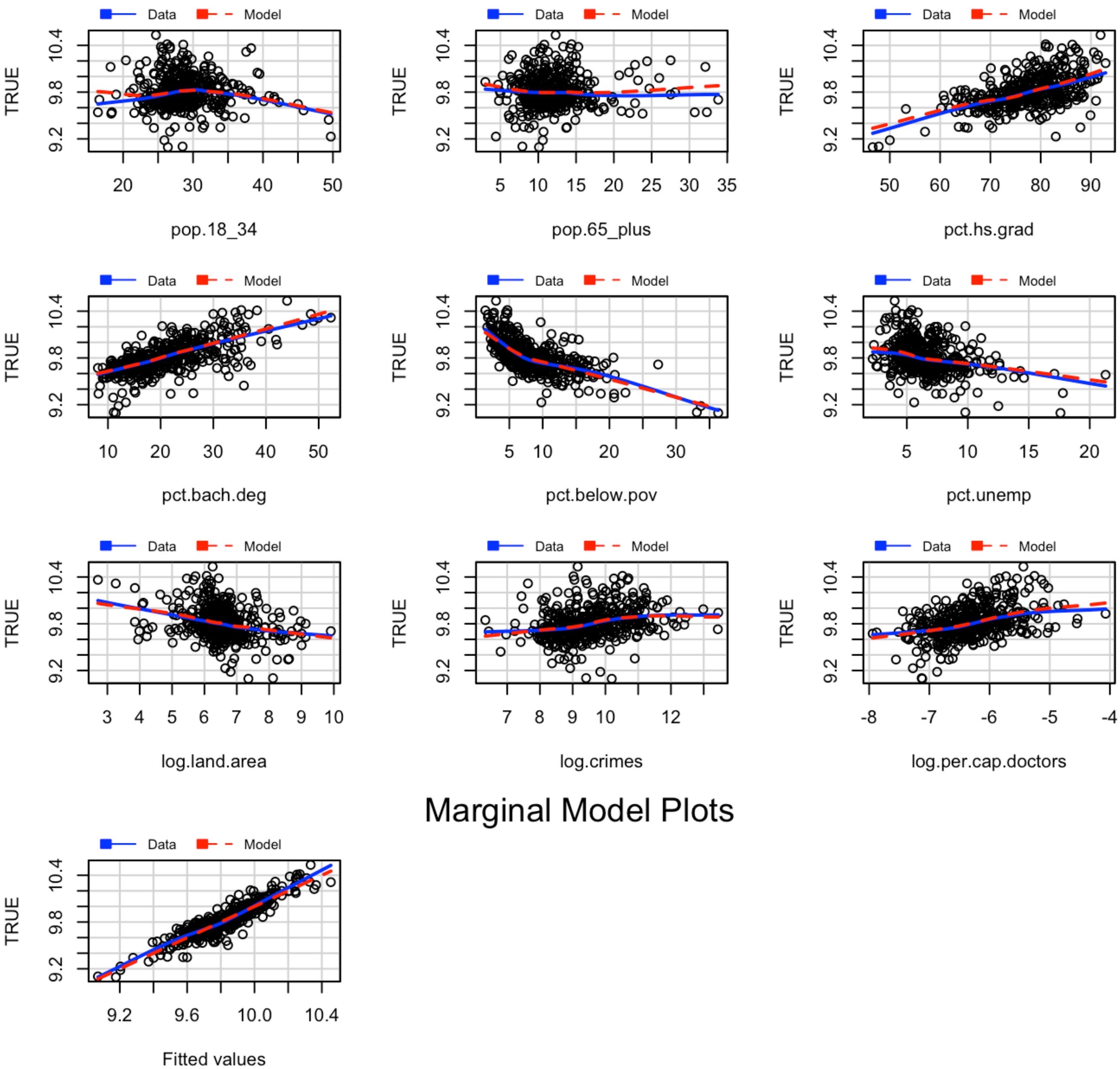
```
model3 = lm(log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
    pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
    log.crimes + log.per.cap.doctors, data=cdi3)
summary(model3)
```

```
##
## 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.crimes + log.per.cap.doctors, data = cdi3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34023 -0.04671 -0.00384  0.04929  0.24325
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         10.9050376  0.1426061  76.470  < 2e-16 ***
## pop.18_34           -0.0162008  0.0013128 -12.340  < 2e-16 ***
## pop.65_plus         -0.0038544  0.0013884  -2.776  0.00574 **
## pct.hs.grad         -0.0050247  0.0010892  -4.613 5.24e-06 ***
## pct.bach.deg         0.0150274  0.0009569  15.704  < 2e-16 ***
## pct.below.pov       -0.0276564  0.0013299 -20.796  < 2e-16 ***
## pct.unemp            0.0130256  0.0022235   5.858 9.32e-09 ***
## log.land.area       -0.0322550  0.0048983  -6.585 1.33e-10 ***
## log.crimes           0.0389684  0.0042899   9.084  < 2e-16 ***
## log.per.cap.doctors  0.0822644  0.0105519   7.796 4.86e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08213 on 430 degrees of freedom
## Multiple R-squared:  0.8454, Adjusted R-squared:  0.8422
## F-statistic: 261.3 on 9 and 430 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(model3)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
246

```
mmps(model3)
```


Marginal Model Plots
Data
Model
TRUE
pop.18_34
pop.65_plus
pct.hs.grad
pct.bach.deg
pct.below.pov
pct.unemp
log.land.area
log.crimes
log.per.cap.doctors
Fitted values

```
vif(model3)
```

```
##          pop.18_34         pop.65_plus         pct.hs.grad        pct.bach.deg 
##           1.970357            1.999920            3.799458            3.491976 
##       pct.below.pov          pct.unemp       log.land.area          log.crimes 
##           2.495965            1.758740            1.186578            1.402528 
## log.per.cap.doctors 
##           2.241848
```

- Because in this case, forward selection based on AIC and BIC choosed the same full model, so we don't consider any more.
- Both of backward elimination with AIC or BIC without interaction with region choosed the same model with nine predictors named Model3.

Model3:

- Based on output of VIF test and residual diagnostics plots, none of the VIFs seem excessively large (>5).
- The diagnostic plots don't show much except that the QQ plot suggests left bottom tails are a bit longer than expected for the normal distribution.
- We can also check for marginal model plots：The nonparametric estimates of each pair-wise relationship are marked as solid curves, while the smooths of the fitted values are marked as dashed curves, we see that there is very good agreement between the two sets of fits in Figure above. Thus, at this stage we shall continue under the assumption that model3 is a valid model. Hence we are not missing any important transformations, interactions.
- However, in this case, we decided not to add the interaction terms because after tried, both of BIC-based model and AIC-based model, there are some statistically significant interactions, but all of the interaction terms with coefficients are close to 0. Thus, it is not useful to consider interaction terms on per-capita income.

Stepwise Regression prefer Model3:

- Based above, stepwise regression selects Model3 as the best model.

Appendix 3.3. LASSO:

```r
library(glmnet)
library(arm)
```

```r
log.per.cap.income = cdi3$log.per.cap.income
predictors_1 = cdi3[,-which(colnames(cdi3)=="log.per.cap.income")]
result <- cv.glmnet(as.matrix(predictors_1), log.per.cap.income, alpha=1)
c(lambda.1se=result$lambda.1se,lambda.min=result$lambda.min)
```

```
##   lambda.1se   lambda.min
## 0.0044721426 0.0002278171
```

```r
cbind(coef(result,s=result$lambda.min),
             coef(result,s=result$lambda.1se)
            )
```

```
## 11 x 2 sparse Matrix of class "dgCMatrix"
##                                s1          s1
## (Intercept)           10.886855545 10.399926240
## pop.18_34             -0.016190349 -0.012919846
## pop.65_plus           -0.004005526  .
## pct.hs.grad           -0.004865861 -0.001107900
## pct.bach.deg           0.015354305  0.012879960
## pct.below.pov         -0.027841640 -0.023284784
## pct.unemp              0.013201304  0.009365068
## log.land.area         -0.031467555 -0.030501888
## log.crimes             0.039488208  0.037655518
## log.per.cap.doctors    0.071646247  0.067554011
## log.per.cap.hosp.beds  0.013251410  .
```

```r
model4 = lm(log.per.cap.income ~ ., data=cdi3)
summary(model4)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ ., data = cdi3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33791 -0.04523 -0.00547  0.04854  0.24619
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           10.919833   0.142991  76.367  < 2e-16 ***
## pop.18_34             -0.016418   0.001323 -12.408  < 2e-16 ***
## pop.65_plus           -0.004318   0.001435  -3.008  0.00278 **
## pct.hs.grad           -0.005115   0.001091  -4.689 3.69e-06 ***
## pct.bach.deg           0.015539   0.001039  14.957  < 2e-16 ***
## pct.below.pov         -0.028166   0.001389 -20.278  < 2e-16 ***
## pct.unemp              0.013431   0.002245   5.982 4.64e-09 ***
## log.land.area         -0.031414   0.004940  -6.359 5.21e-10 ***
## log.crimes             0.039623   0.004318   9.176  < 2e-16 ***
## log.per.cap.doctors    0.071061   0.013789   5.153 3.91e-07 ***
## log.per.cap.hosp.beds  0.015209   0.012063   1.261  0.20807
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08207 on 429 degrees of freedom
## Multiple R-squared:  0.846,  Adjusted R-squared:  0.8424
## F-statistic: 235.6 on 10 and 429 DF,  p-value: < 2.2e-16
```

```
model5 = lm(log.per.cap.income ~ log.land.area+pop.18_34+log.crimes+pct.bach.deg+pct.
below.pov+pct.unemp+log.per.cap.doctors, data=cdi3)
summary(model5)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     log.crimes + pct.bach.deg + pct.below.pov + pct.unemp + log.per.cap.doctors,
##     data = cdi3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36330 -0.04765 -0.00822  0.05292  0.23976
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         10.3595873  0.0937621 110.488  < 2e-16 ***
## log.land.area       -0.0369618  0.0049080  -7.531 2.97e-13 ***
## pop.18_34           -0.0145057  0.0011457 -12.661  < 2e-16 ***
## log.crimes           0.0413618  0.0043814   9.440  < 2e-16 ***
## pct.bach.deg         0.0134060  0.0008351  16.053  < 2e-16 ***
## pct.below.pov       -0.0239156  0.0011335 -21.099  < 2e-16 ***
## pct.unemp            0.0148977  0.0021934   6.792 3.67e-11 ***
## log.per.cap.doctors  0.0731657  0.0099539   7.350 9.97e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08442 on 432 degrees of freedom
## Multiple R-squared:  0.8359, Adjusted R-squared:  0.8333
## F-statistic: 314.4 on 7 and 432 DF,  p-value: < 2.2e-16
```

```
region = cdi$region
model7 = lm(log.per.cap.income ~ log.land.area+pop.18_34+log.crimes+pct.bach.deg+pct.
below.pov+pct.unemp+log.per.cap.doctors
            +region+region:log.land.area+region:pop.18_34+region:log.crimes+region:pc
t.bach.deg
            +region:pct.below.pov+region:pct.unemp+region:log.per.cap.doctors, data=c
di3)
summary(model7)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     log.crimes + pct.bach.deg + pct.below.pov + pct.unemp + log.per.cap.doctors +
##     region + region:log.land.area + region:pop.18_34 + region:log.crimes +
##     region:pct.bach.deg + region:pct.below.pov + region:pct.unemp +
##     region:log.per.cap.doctors, data = cdi3)
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max 
## -0.289245 -0.044942 -0.003703  0.043653  0.293967 
## 
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                 10.1261688  0.2063454  49.074  < 2e-16 ***
## log.land.area               -0.0374364  0.0152589  -2.453 0.014568 *  
## pop.18_34                   -0.0158199  0.0026016  -6.081 2.75e-09 ***
## log.crimes                   0.0403349  0.0083737   4.817 2.06e-06 ***
## pct.bach.deg                 0.0140435  0.0021832   6.433 3.52e-10 ***
## pct.below.pov               -0.0214105  0.0036763  -5.824 1.17e-08 ***
## pct.unemp                    0.0157620  0.0053751   2.932 0.003553 ** 
## log.per.cap.doctors          0.0336327  0.0185940   1.809 0.071219 .  
## regionNE                    -0.0931416  0.3266599  -0.285 0.775687    
## regionS                      0.2937961  0.2520485   1.166 0.244444    
## regionW                      0.5260589  0.3282842   1.602 0.109830    
## log.land.area:regionNE      -0.0048414  0.0198434  -0.244 0.807369    
## log.land.area:regionS        0.0007884  0.0180702   0.044 0.965222    
## log.land.area:regionW        0.0281667  0.0187678   1.501 0.134181    
## pop.18_34:regionNE          -0.0016785  0.0036840  -0.456 0.648911    
## pop.18_34:regionS           -0.0002452  0.0030833  -0.080 0.936654    
## pop.18_34:regionW            0.0080890  0.0042552   1.901 0.058009 .  
## log.crimes:regionNE          0.0146038  0.0128116   1.140 0.255000    
## log.crimes:regionS           0.0015680  0.0117394   0.134 0.893810    
## log.crimes:regionW          -0.0061137  0.0137404  -0.445 0.656594    
## pct.bach.deg:regionNE        0.0036638  0.0030119   1.216 0.224514    
## pct.bach.deg:regionS        -0.0016234  0.0024601  -0.660 0.509689    
## pct.bach.deg:regionW        -0.0021836  0.0030404  -0.718 0.473049    
## pct.below.pov:regionNE      -0.0039333  0.0052617  -0.748 0.455169    
## pct.below.pov:regionS        0.0038647  0.0040103   0.964 0.335763    
## pct.below.pov:regionW       -0.0067391  0.0050382  -1.338 0.181768    
## pct.unemp:regionNE          -0.0051113  0.0079228  -0.645 0.519201    
## pct.unemp:regionS           -0.0212288  0.0072308  -2.936 0.003514 ** 
## pct.unemp:regionW            0.0056345  0.0066002   0.854 0.393779    
## log.per.cap.doctors:regionNE -0.0058385  0.0313818  -0.186 0.852499    
## log.per.cap.doctors:regionS   0.0326286  0.0244448   1.335 0.182690    
## log.per.cap.doctors:regionW   0.1343280  0.0348759   3.852 0.000136 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.07857 on 408 degrees of freedom
## Multiple R-squared:  0.8658, Adjusted R-squared:  0.8556 
## F-statistic: 84.88 on 31 and 408 DF,  p-value: < 2.2e-16
```

```
model6 = lm(log.per.cap.income ~ log.land.area+pop.18_34+log.crimes+pct.bach.deg+pct.
below.pov+pct.unemp+log.per.cap.doctors
            +region+region:pct.unemp+region:log.per.cap.doctors, data=cdi3)
summary(model6)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     log.crimes + pct.bach.deg + pct.below.pov + pct.unemp + log.per.cap.doctors +
##     region + region:pct.unemp + region:log.per.cap.doctors, data = cdi3)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.308544 -0.048249 -0.003535  0.049019  0.260536
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  9.9901410  0.1156985  86.346  < 2e-16 ***
## log.land.area               -0.0296356  0.0055557  -5.334 1.57e-07 ***
## pop.18_34                   -0.0150500  0.0011322 -13.293  < 2e-16 ***
## log.crimes                   0.0442597  0.0043688  10.131  < 2e-16 ***
## pct.bach.deg                 0.0137848  0.0008232  16.746  < 2e-16 ***
## pct.below.pov               -0.0206012  0.0012481 -16.507  < 2e-16 ***
## pct.unemp                    0.0144703  0.0042471   3.407 0.000719 ***
## log.per.cap.doctors          0.0280807  0.0141504   1.984 0.047852 *
## regionNE                     0.3772241  0.1318038   2.862 0.004419 **
## regionS                      0.2985624  0.1067612   2.797 0.005401 **
## regionW                      0.6787587  0.1610012   4.216 3.04e-05 ***
## pct.unemp:regionNE          -0.0139355  0.0062731  -2.221 0.026846 *
## pct.unemp:regionS           -0.0118148  0.0053499  -2.208 0.027751 *
## pct.unemp:regionW            0.0039229  0.0049671   0.790 0.430100
## log.per.cap.doctors:regionNE  0.0408476  0.0222811   1.833 0.067463 .
## log.per.cap.doctors:regionS   0.0401736  0.0175183   2.293 0.022325 *
## log.per.cap.doctors:regionW   0.1159527  0.0266833   4.346 1.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0804 on 423 degrees of freedom
## Multiple R-squared:  0.8543, Adjusted R-squared:  0.8487
## F-statistic:   155 on 16 and 423 DF,  p-value: < 2.2e-16
```

```
anova(model5,model6)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)

	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	432	3.078484	NA	NA	NA	NA
2	423	2.734499	9	0.3439852	5.912347	8.55976e-08

2 rows

```
AIC(model5,model6)
```

	df	AIC
	<dbl>	<dbl>
model5	9	-916.7626
model6	18	-950.8978

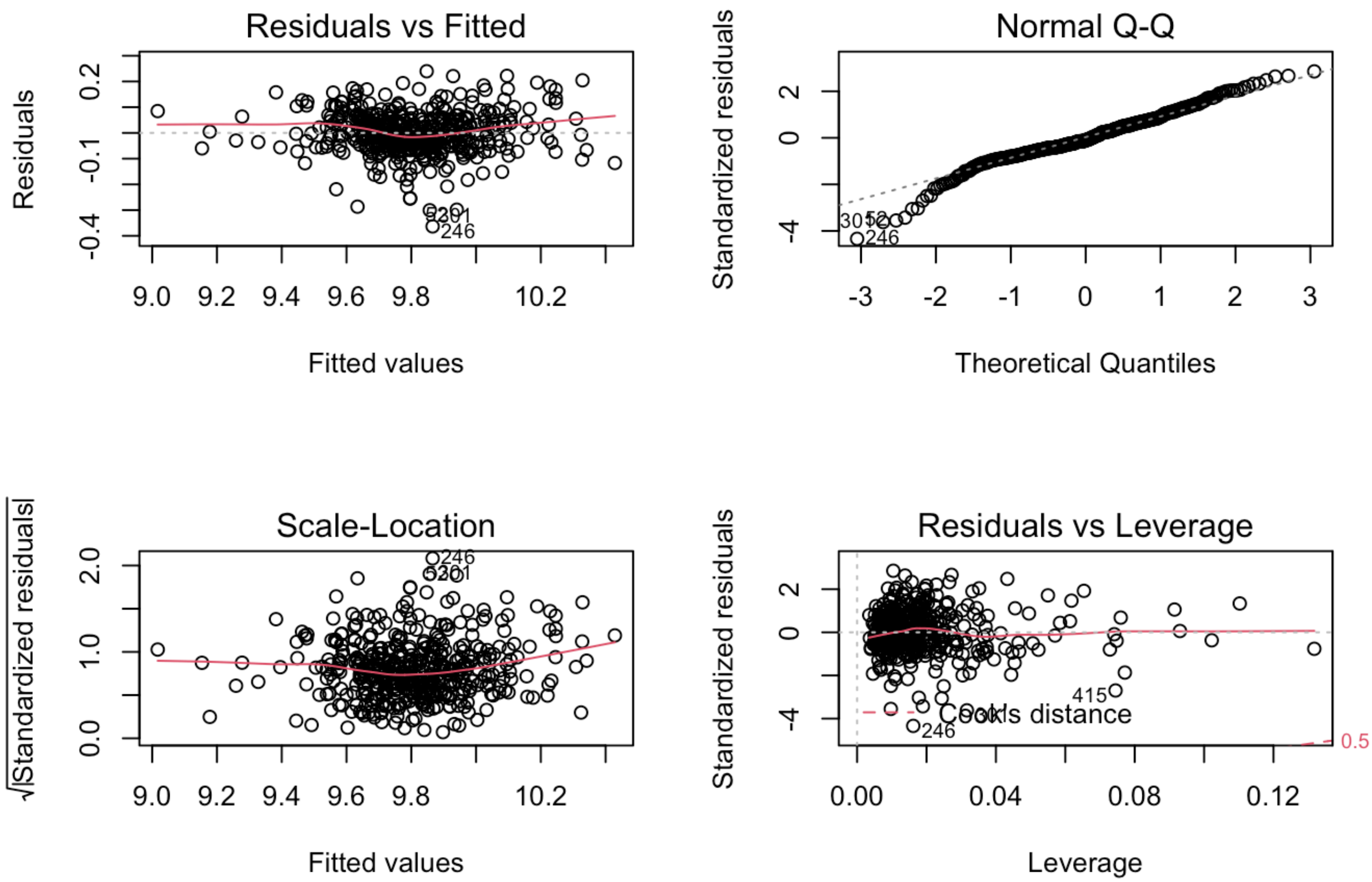
2 rows

```
BIC(model5,model6)
```

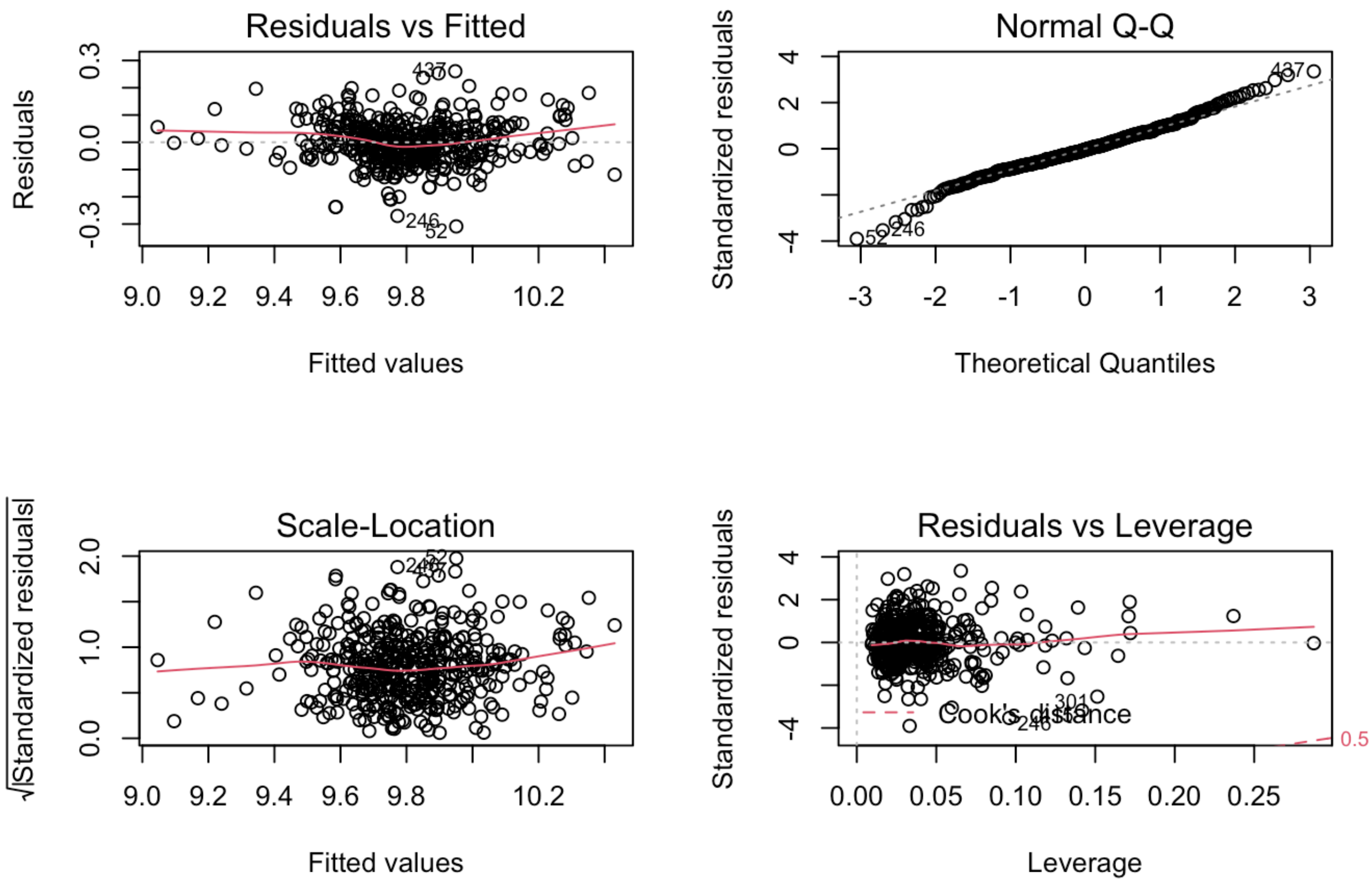
	df	BIC
	<dbl>	<dbl>
model5	9	-879.9816
model6	18	-877.3358

2 rows

```
par(mfrow=c(2,2))
plot(model5)
```




```
par(mfrow=c(2,2))
plot(model6)
```



Based on model summary for Model4 and Model5, we found there is one predictor log.per.cap.hosp.beds which is not statistically significant in Model4 which suggests that Model4 could be overfit.

- However, Model5 with all predictors are statistically significant. Thus, we prefer Model5.

After we added region interaction terms we got Model6:

- However, most of the coefficients are small, and some still seem to have the wrong sign in Model5 and Model6 (e.g. log.crimes).
- The diagnostic plots are much better especially that the QQ plot suggests left bottom tails are fixed in Model5 before for the normal distribution.

LASSO prefer Model6 as the best model:

- Both of the ANOVA F test and AIC prefer Model6. On the other hand, BIC prefers the simpler model. Besides Model6 had R^2_{adj}=0.8487 which is larger than Model5 with R^2_{adj}=0.8333 and better residual diagnostic plots, the QQ plot suggests left bottom tails has been fixed within Model6.