

Relationships Between Average Income Per Person and County's Economic, Health and Social Well-being

Naijia Liu
naijial@andrew.cmu.edu

October 17, 2021

Abstract

We address the question of how average income per person was related to other variables associated with the county's economic, health and social well-being. We examine data on the county demographic information (CDI) for 440 of the most populous counties in the United States selected by Kutner et al. (2005). From exploratory data analysis, it appears that the average income per person is highly influenced by percent total personal income, total population, land area and region; and somewhat less by percent of adult population with bachelor's degree, Percent of adult population who completed 12 or more years of school, percent of population with income below poverty level and percent of population aged 18–34. Income per person is highly effected by county's population composition, economic and social-welling, especially education. Also there does existed a relationship between income per person and the geographic features of an county, including region and land area. Since the data and analysis only cover 373 counties, the study in this paper was also limited by the size of the data set.

1 Introduction

Per ca pita income has always been a topic of high concern among economists, statisticians and ordinary people, which may be influenced many factors. How will average income per person be related with economic, health and social well-being?

In this paper, we have been asked to explore the relationships between average per ca-pita income and county's economic, health, social Well-being, and predict it from all the influence factors. Relationships may vary depending on the geographical scope of the study, and the emphasis may also vary. Particularly, we are focusing on the counties.

In addition to answering the main question posed above, we will address the following questions:

- Is there any relationships between all the influence factors? If so, how they effect each other?
- Is there a relationship between per ca-pita income and crime rate? And how does this relationship differentiate in different regions of the country?
- Since we only cover 373 of approximately 3000 counties in the US, will the accuracy of the study affected by the missing counties? Why or why not?

2 Data

The data for this study is taken from Kutner et al. (2005). It provides selected county demographic information (CDI) for 440 of the most populous counties in the United States, and the information generally pertains to the years 1990 and 1992. The reader should refer to Kutner et al. (2005) for definitions, eligibility, inclusion/exclusion criteria, and so forth. See the relationships between all variables in Figure 1.

In all, 440 of the most populous counties' CDI in the United States are represented in the data available to us, and the variables were defined in the following Table 1:

In Figure 1 we show the relationships between all of the quantitative variables in the data. Some of them appear to be linearly correlated with the response variable per capital income, and the distribution

Variable Number	Variable Name	Description
1	Identification number	1–440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 or old
8	Number of active physicians	Number of professionally active non federal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor’s degrees	Percent of adult population (persons 25 years old or older) with bachelor’s degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI population that is unemployed
15	Per capita income	Per-capita income (i.e. average income per person) of 1990 CDI population (in dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification used by the US Bureau of the Census, NE (northeast region of the US), NC (north-central region of the US), S (southern region of the US), and W (Western region of the US)

Table 1: Definitions of All variables.

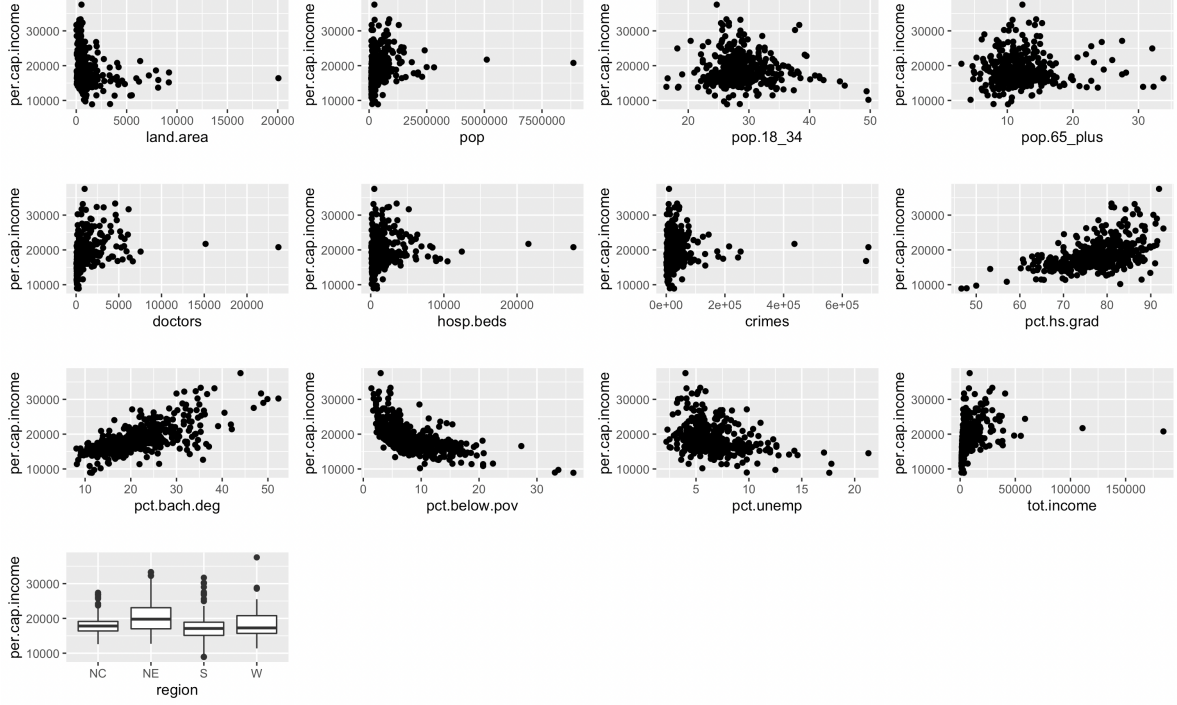


Figure 1: Scatterplot matrix of CDI(county demographic informations).

of some quantitative variables is highly skewed, indicating transformation. Table 2 and Table 3 provide us with a quick summaries of the numeric variables and character variables we will be using in this paper.

3 Methods

First, we did the visual comparison of exploratory scatter plots and histograms to check univariable distributions, using the R language and environment for statistical computing (R Core Team, 2017). And in order to identify any relationships may exist between all the influence factors, we have also examined the raw data in `cdi.dat` using exploratory scatter plots (see Figure 1) and a correlation plot. We used logarithm transformation on some variables to overcome non-linearity and skewness of variable distribution. Detailed R analyses can be found in Appendix 1.

Then, to address the question about whether per ca-pita income is related to crime rate, and how this relationship may be different in different regions of the country, we used three multiple regression models. This analysis can tell us about the existence of the relationship and the effect of region, after controlling for all the other predictor variables. After that, by replacing the number of crimes with per ca-pita crimes, we examined whether different scales of a variable affect its influence on predictor variable, and we compared two best models selected using AIC and BIC. Detailed R analyses can be found in Appendix 2.

Next, we considered multiple regression models, also in R, to fit the best model predicting per capita income from all the other variables except population and total income, which are two deterministic factors for per ca-pita income. And we converted land area, doctors, hospital beds, crimes and total income into per ca-pita scales for better interpretation. For further model selection, we first chose a best model with all possible subsets method by comparing AIC, AICc and BIC; then, another best model was chosen from step-wise regression; lastly, based on LASSO regression and cross-validation using mean-squared prediction error, we had another best model.

We examined case-wise residual plots, marginal model plots, partial-regression and a likelihood ratio test to select the best model. That model is used to interpret the effects of the predictor variables and to estimate per ca-pita income with county's economic, health, social Well-being. This final model

Numeric Variables				
Variables	Min.	Median	Mean	Max.
id	1.0	220.5	220.5	440.0
land.area	15.0	656.5	1041.4	20062.0
pop	100043	217280	393011	8863164
pop.18.34	16.40	28.10	28.57	49.70
pop.65.plus	3.000	11.750	12.170	33.800
doctors	39.0	401.0	988.0	23677.0
hosp.beds	92.0	755.0	1458.6	27700.0
pct.hs.grad	46.60	77.70	77.56	92.90
pct.bach.deg	8.10	19.70	21.08	52.30
pct.below.pov	1.400	7.900	8.721	36.300
pct.unemp	2.200	6.200	6.597	21.300
per.cap.income	8899	17759	18561	37541
tot.income	1141	3857	7869	184230

Table 2: Summary for Numeric Variables.

Character variables				
	NC	NE	S	W
Freq	108	103	152	77

Table 3: Summary for Character Variables.

also tells us about the effect of each individual predictor variable, after controlling for all the other predictor variables. Details of these analyses in R can be found in Appendices 3, 4 and 5.

4 Results

4.1 Exploratory Plots and Transformation

First, from the histograms of the distribution of raw data and Table 2, we found that land area, population, doctors, hospital beds, crimes ,total income need transformation, since they are severely right-skewed. And in the correlation matrix (Appendix 1), it suggest that we may run into multi-collinearity problems when we start fitting models, as three variables, including crimes, hospital beds and doctors, seem strongly correlated with one another. To address heavy skewing and potential leverage and influence issues, we only took the logarithms of the variables we identified earlier.

We also consider logarithms transformation for per ca-pita income. There are two reasons for this: 1) we can explain the logarithms in terms of percent-change concepts; 2) and the more untransformed variables there are, the easier it will be to understand. As we can see from Figure 2, the skewing seems to have largely been brought under control after transformation.

4.2 Relationship between Per Ca-pita Income and Crime Rate

Since logarithms cleaned up a lot of the skewing in the data, we will only use log-transformed variables from now on for regression. For the income and each version of crime variables, there are essentially three models to think about. First, for log.per.cap.income and log.crimes, the three models to consider are:

$$\log(per.cap.income) = \beta_0 + \beta_1 * \log(crimes) \quad (1)$$

$$\log(per.cap.income) = \beta_0 + \beta_1 * \log(crimes) + \beta_2 * region \quad (2)$$

$$\log(per.cap.income) = \beta_0 + \beta_1 * \log(crimes) * region \quad (3)$$

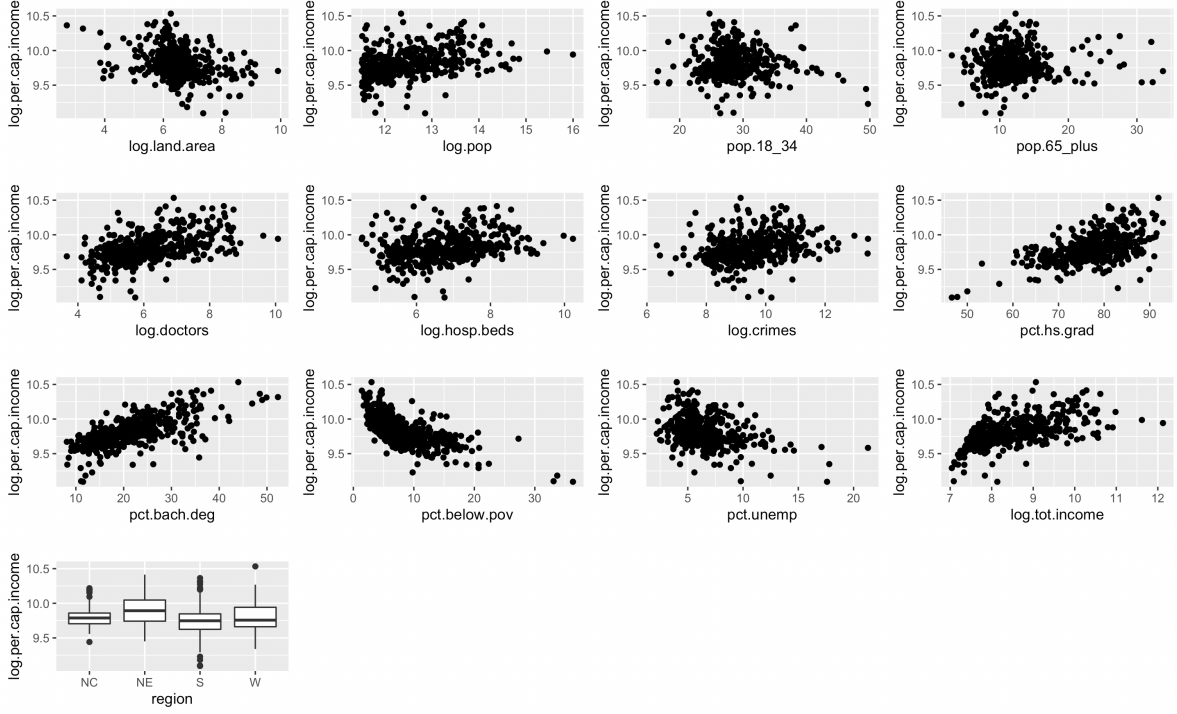


Figure 2: Scatterplot matrix of Transformed CDI(county demographic informations).

In order to compare this with a model involving “per-capita crime”, we have to construct a new variable, and we will look at the same three models with this new variable.

$$\log(\text{per.cap.income}) = \beta_0 + \beta_1 * \log(\text{per.cap.crimes}) \quad (4)$$

$$\log(\text{per.cap.income}) = \beta_0 + \beta_1 * \log(\text{per.cap.crimes}) + \beta_2 * \text{region} \quad (5)$$

$$\log(\text{per.cap.income}) = \beta_0 + \beta_1 * \log(\text{per.cap.crimes}) * \text{region} \quad (6)$$

We compared all 6 models with AIC or BIC, and from that, we get the result: the second model(named as ”ancova.02” in Appendix 2) has both the lowest AIC and the lowest BIC values (Appendix 2). However, there is an argument that per ca-pita crime is more comparable to, or at least on the same scale as, per ca-pita income. Next, in examining the residual plots for two models more closely (Appendix 2), both the diagnostic plots don’t show much except that the QQ plot suggests both the left and the right tails are a bit longer than expected for the normal distribution. Thus, for better interpretation, we finally decide on the model (5).

Table 4 gives the estimated coefficients for model (5) along with standard errors and the usual t-test for testing whether each coefficient is significantly different from zero. All across the US, for every 1% increase in per ca-pita crime, there is an associated 0.04% increase in per ca-pita income. And the regional baseline salaries are: NC: \$20,743.74, NE: \$23,155.79, S: is \$19,341.34, and W: \$20,332.99. All but the W region have baselines that are, according to the model, significantly different from the NC baseline. So, according to the model (5), the level of salary varies with region in the US, but the way it is related to crime does not.

4.3 Regression Analysis

To make the analysis simpler, we will not consider the specific geographic identifiers like: ”county” and ”state”, instead we will only use 4 different regions in ”Geographic region” as 4 dummy variables for identifying. Per ca-pita variables are more comparable to, or at least on the same scale as, per ca-pita income, so we will convert land area, doctors, hospital beds, crimes and total income into per

	Estimate	Std. Error	<i>t</i> value	$Pr(> t)$
β_0 : (Intercept)	9.94	0.07	143.30	0.00
β_1 : $\log(\text{per.cap.crimes})$	0.04	0.02	1.98	0.05
β_2 : regionNE	0.11	0.03	4.15	0.00
β_3 : regionS	-0.07	0.03	-2.84	0.00
β_3 : regionW	-0.02	0.03	-0.81	0.42

Table 4: Estimated coefficients for model(1).

ca-pita scales. And we took population and total income out of consideration for regression and since per ca-pita income is a deterministic function of them. And there are three methods we will be using for model selection, including all subsets, step-wise regression and LASSO.

4.3.1 All Subsets Regression

By locating the model with the lowest BIC, we found the best model 7 in all subsets regression.

$$\begin{aligned}
\log(\text{per.cap.incom}) = & \beta_0 + \beta_1 * \log(\text{per.cap.land.area}) + \beta_2 * \text{pop.1834} \\
& + \beta_3 * \text{pop.65plus} + \beta_4 * \log(\text{per.cap.doctors}) \\
& + \beta_5 * \text{pct.hs.grad} + \beta_6 * \text{pct.bach.deg} \\
& + \beta_7 * \text{pct.below.pov} + \beta_8 * \text{pct.unemp}
\end{aligned} \tag{7}$$

All the predictors have coefficients 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’). Thus, as mentioned before, we need to check the problem of multi-linearity before further investigation. None of the VIFs seem excessively large, and the diagnostic plots don’t show much except that the QQ plot suggests both the left and the right tails are a bit longer than expected for the normal distribution. The marginal model plots look very good – the blue data-based curves line up well with the red model-based curves. (Appendix 3)

Next, we added the interaction with region. Following the rule: if any indicator for a categorical variable seems important, then keep the whole categorical variable; and if none of them seem important, then drop the variable, we found another best model 8. After checking VIFs and diagnostics, and comparing to the best BIC model 7 that we obtained from subsets without the ‘region’ term (Appendix 4), we arrive at the final model 8 with interaction of region.

$$\begin{aligned}
\log(\text{per.cap.income}) = & \beta_0 + \beta_1 * \log(\text{per.cap.land.area}) + \beta_2 * \text{pop.1834} \\
& + \beta_3 * \text{pop.65plus} + \beta_4 * \log(\text{per.cap.doctors}) \\
& + \beta_5 * \text{pct.hs.grad} + \beta_6 * \text{pct.bach.deg} \\
& + \beta_7 * \text{pct.below.pov} + \beta_8 * \text{pct.unemp} \\
& + \beta_9 * \text{region} + \beta_{10} * \log(\text{per.cap.land.area} : \text{region}) \\
& + \beta_{11} * \log(\text{per.cap.doctors} : \text{region}) + \beta_{12} * (\text{pct.hs.grad} : \text{region}) \\
& + \beta_{13} * (\text{pct.below.pov} : \text{region}) + \beta_{15} * (\text{pct.unemp} : \text{region})
\end{aligned} \tag{8}$$

Table 5 gives the estimated coefficients for model (8) along with standard errors and the usual t-test for testing whether each coefficient is significantly different from zero.

As seen above, for every 1% increase in a county’s per ca-pita land area, there is a 0.05% decrease in expected per ca-pita income. This could be due to an urban-rural contrast: rural counties tend to be bigger than urban ones. However, in South and West, a one percentage point increase in county’s per ca-pita land area induces an expected 0.01% and 0.03% increase in per ca-pita income.

For every 1% increase in per ca-pita doctors in a county, the expected per ca-pita income increases by about 0.05%. For all regions, increase in the number of per ca-pita doctors induces increase in per ca-pita average income, especially in West, where a one percentage point increase in per ca-pita doctors induces an expected 0.12% increase in per ca-pita income. That makes sense; doctors are well-paid and could be big contributors to the per ca-pita average income.

	Estimate	Std. Error	<i>t</i> value	<i>Pr</i> ($> t $)
β_0 : (Intercept)	10.27	0.28	36.06	0.00
β_1 :log.per.cap.land.area	-0.05	0.01	-5.86	0.00
β_2 :pop.18 ₃₄	-0.02	0.00	-11.93	0.00
β_3 :pop.65 _{plus}	0.00	0.00	-1.33	0.19
β_4 :log.per.cap.doctors	0.05	0.02	3.10	0.00
β_5 :pct.hs.grad	0.00	0.00	-1.14	0.25
β_6 :pct.bach.deg	0.01	0.00	15.14	0.00
β_7 :pct.below.pov	-0.02	0.00	-6.03	0.00
β_8 :pct.unemp	0.02	0.00	3.72	0.00
β_9 :regionNE	-0.03	0.40	-0.08	0.93
β_{10} :regionS	0.32	0.32	1.00	0.32
β_{11} :regionW	2.49	0.42	5.91	0.00
β_{12} :log.per.cap.land.area:regionNE	-0.01	0.01	-0.88	0.38
β_{13} :log.per.cap.land.area:regionS	0.01	0.01	1.19	0.24
β_{14} :log.per.cap.land.area:regionW	0.03	0.01	2.23	0.03
β_{15} :log.per.cap.doctors:regionNE	0.01	0.03	0.23	0.82
β_{16} :log.per.cap.doctors:regionS	0.03	0.02	1.46	0.15
β_{17} :log.per.cap.doctors:regionW	0.12	0.03	3.98	0.00
β_{18} :pct.hs.grad:regionNE	0.00	0.00	0.43	0.67
β_{19} :pct.hs.grad:regionS	0.00	0.00	0.04	0.96
β_{20} :pct.hs.grad:regionW	-0.02	0.00	-4.59	0.00
β_{21} :pct.below.pov:regionNE	0.00	0.01	-0.89	0.38
β_{22} :pct.below.pov:regionS	0.00	0.00	0.78	0.44
β_{23} :pct.below.pov:regionW	-0.02	0.01	-3.54	0.00
β_{24} :pct.unemp:regionNE	-0.01	0.01	-1.30	0.19
β_{25} :pct.unemp:regionS	-0.02	0.01	-2.64	0.01
β_{26} :pct.unemp:regionW	-0.02	0.01	-2.51	0.01

Table 5: Estimated coefficients for model(8).

	df	AIC	BIC
all.subsets.final.model	10	-940.5743	-899.7065
stepwise.bic	10	-940.5743	-899.7065
stepwise.aic	11	-942.0976	-897.1431
stepwise.bic.2-way-inter	16	-1047.7658	-982.3774
stepwise.aic.2-way-inter	23	-1082.1848	-988.1890

Table 6: Models Comparison with AIC and BIC.

For every 1 percentage point increase in the percent of the population aged 18-34, there is an expected 2% drop in per ca-pita income. We might conjecture that this is because 18-34 year old are not at peak earning capacity yet and so perhaps their lower incomes drags down the per ca-pita average. However, percent of the population that are 65 years old and over doesn't have much effect on per ca-pita income.

Percent of the population that are high school graduates doesn't have much effect, except in the West, where a one percentage point increase in high school graduates induces an expected 2% decrease in per ca-pita income. We don't have a good explanation for this: It might depend on whether college graduates are counted as a subset of high school graduates rather than counting them separately, or it might have something to do with some unique feature of economics in the southern region of the US. And for every 1% increase in percent of the population that have a bachelor's degree, there is a 1% increase in expected per ca-pita income. This makes perfect sense, since higher education usually means higher income.

For every 1% increase in percent of percent below poverty level, there is a 2% decrease in expected per ca-pita income. This makes perfect sense, since percent of percent below poverty level must drag down the per ca-pita average.

And in the main effect for region, and in several of the interactions for region, the West shows up as deviating significantly from the North Central part of the US.

4.3.2 Stepwise Regression

First, we used two stepwise regressions, including AIC and BIC. And in the stepwise regression using the BIC criterion, we actually found the model that is exactly the same as the all-subsets model 8.

Stepwise regression using the AIC criterion added one more predictor, namely $\log(\text{per.cap.crimes})$. Although, $\log(\text{per.cap.crimes})$'s effect on expected per ca-pita income appears to be larger than zero too, it is only significantly different from zero under the significance level 0.1 (Appendix 4). So, we will not be adding $\log(\text{per.cap.crimes})$ into the model.

Next, we explored 2-way interactions briefly, and both AIC and BIC like models with some interactions, as we can see in Table 6. However, the model is just getting too complicated to explain to someone.

Naturally, we shall consider adding interactions into two models, and there are 2 interactions in BIC-based model and 7 interactions in AIC-based model, that appear to be statistically significant, but their practical effect is almost zero on per ca-pita income. Although both interactions models produced big jumps in AIC and BIC, the improvement in R^2 and R^2_{adj} is pretty small, for all the terms that have been added to the models. For these reasons we are willing to discuss these interactions with the social scientist, instead of including them in a final model.

In conclusion, we shall stick with the model found by all-subsets and stepAIC with a BIC penalty, and we will once again be led to model 8 we found based on the all-subsets, which has some interesting and mostly-interpretable structure.

4.3.3 LASSO Regression

The last variable selection method we'll consider is the LASSO. As we can see in the Figure 3, there is not such an obvious place to cut the shrinkage plot, to help determine what variables should be kept in the model.

Using cross-valuation, we got the model which minimizes 10-fold cross-validation error contains all 10 predictors (or sometimes 9— $\log.\text{crimes}$ can get left out—depending on what random folds we

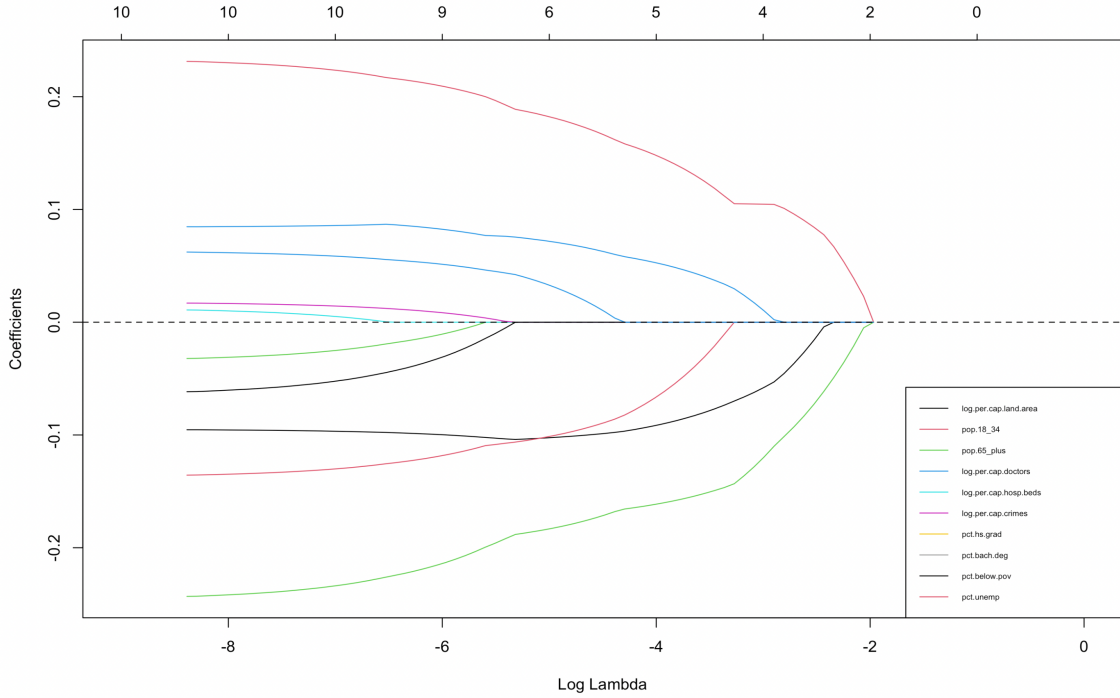


Figure 3: LASSO Variable Trace Plot.

get). The model 9 that is 1 SE has 2 less variable than the model that we first saw with all-subsets regression.

$$\begin{aligned}
 \log(\text{per.cap.incom}) = & \beta_0 + \beta_1 * \log(\text{per.cap.land.area}) + \beta_2 * \text{pop.18}_{34} \\
 & + \beta_4 * \log(\text{per.cap.doctors}) + \beta_5 * \text{pct.hs.grad} \\
 & + \beta_6 * \text{pct.bach.deg} + \beta_7 * \text{pct.below.pov} \\
 & + \beta_8 * \text{pct.unemp}
 \end{aligned} \tag{9}$$

Though F-test prefers the model that is 1 SE based on LASSO, both AIC and BIC prefers the original model based on the all-subsets regression (Appendix 5). So for better interpretation and the integrality of the model, we will settle on the very first model we got based on the all-subsets regression.

4.4 Discussion

Average per ca-pita income is positively correlated with per ca-pita doctors in a county and the percent of the population that have a bachelor's degree. We also found that per ca-pita doctors had the effect of about 0.05% per unit increase; respectively, every 1% increase in the percent of the population that have a bachelor's degree has the effect of about 1% increase in expected per ca-pita income. All make perfect sense, since doctors are well-paid jobs, and higher education usually means higher income.

Average per ca-pita income is negatively correlated with county's per ca-pita land area, the percent of the population aged 18-34 and the percent of percent below poverty level. For every 1% increase in a county's per ca-pita land area, there is a 0.05% decrease in expected per ca-pita income. And every 1 percentage point increase in the percent of the population aged 18-34 and the percent of percent below poverty level, there is respectively expected 2% and 2% and drop in per ca-pita income.

Overall, the percent of the population that are high school graduates doesn't have much effect, except in the West, where a one percentage point increase in high school graduates induces an expected 2% decrease in per ca-pita income. Though, we did not have a very good explanation for this. The same story is for the percent of the population that are 65 years old and over.

As discussed above, our results are basically consistent with common sense. More job opportunities, more well-paid jobs, more job opportunities, higher education and better social well-being are essential for increasing average per ca-pita income in a county. The composition of population has some effects on average per ca-pita income in a county, but this mostly is due to the income gap between various populations. Since we also took regions into consideration, it shows in our analysis that in several of the interactions for region, the West shows up as deviating significantly from the North Central part of the US.

Our study is limited since the data only covers 373 counties, and it would be very useful to have additional data to compare some of the models we found. We are using reasonable methods for variable selection, but since it is all within-sample (our entire data set is our training sample), there is ample room for over fitting noise in the data. Also just to simplify our analysis, we did not explore the state variable at all. Some of the relationship between these demographic variables and per ca-pita income might be explainable in terms of varying economic policy from one state to the next.

Some of our inferences about which variables to leave in or take out may be based on overly optimistic standard error estimates. For example, the coefficient on percent unemployment seems to go the wrong way, and the coefficient on percent high school graduates is quite small, statistically and practically (it remains in the model because there is a noticeable interaction that it participates in). If we were able to cross-validate on some new or hold-out data, we might be able to better distinguish what is the best model, at least in terms of prediction error. Moreover, the fact that stepwise regression found some well-fitting models with interactions between continuous variables suggests exploring those more complex models in the future.

References

- R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>.
- RStudio Team (2021). *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA URL <http://www.rstudio.com/>.
- Sheather, S.J. (2009), *A Modern Approach to Regression with R*. New York: Springer Science + Business Media LLC.
- Kutner, M.H., Nachsheim, C.J., Neter, J. Li, W. (2005), *Applied Linear Statistical Models, Fifth Edition*. NY: McGraw- Hill/Irwin.

Appendix

Naijia Liu

10/16/2021

Contents

Appendix 1. Initial Data Exploration & Transformation	1
Appendix 2. Analysis of Relationship between Per ca-pita income and Crime Rate	13
Appendix 3. Regression Analysis – All Subsets(without Regions)	15
Appendix 3. Regression Analysis – All Subsets(with Regions)	18
Appendix 4. Regression Analysis – Stepwise Regression	23
Appendix 5. Regression Analysis – LASSO	24

Appendix 1. Initial Data Exploration & Transformation

Read the data in, get a general sense of the variables, and make a “pairs” plot (scatterplot matrix) of the numerical variables. Note that “Per ca-pita income” is the response variable.

```
# loading the needed packages
```

```
library(Hmisc)
library(arm)
library(corrplot)
library(dplyr)
library(car)
library(alr3)
library(MASS)
library(leaps)
library(glmnet)
library(tidyverse)
library(kableExtra)
library(GGally)
library(grid)
library(gridExtra)
library(ggplotify)
library(reshape2)
```

```
cdi <- read.table("cdi.dat") %>% data.frame()
```

```
# Check to see how many unique values each variable has
```

```
apply(cdi,2,function(x) {length(unique(x))}) %>%
  kbl(booktabs=T,col.names="unique values",caption=" ") %>%
  kable_classic(full_width=F)
```

Table 1:

	unique values
id	440
county	373
state	48
land.area	384
pop	440
pop.18_34	149
pop.65_plus	137
doctors	360
hosp.beds	391
crimes	437
pct.hs.grad	223
pct.bach.deg	220
pct.below.pov	155
pct.unemp	97
per.cap.income	436
tot.income	428
region	4

There are two category of variables in data, we divided them into numeric and categorical variables for a more refined look.

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

# Summary Statistics for Numeric Variables
apply(cdinumeric, 2, function(x){c(summary(x), SD=sd(x))}) %>% as.data.frame %>% t() %>%
  round(digits=2) %>% kbl(booktabs=T, caption=" ") %>% kable_classic()
```

There are several variables with Mean substantially larger than Median (`land.area`, `pop`, `doctors`, `hosp.beds`, `crimes`, `per.cap.income`, and `total.income`), indicating possible right-skewing.

```
tmp <- rbind(with(cdi, table(region)))
row.names(tmp) <- "Freq"

# Summary Statistics for Character Variables (Region) which we will be using
tmp %>% kbl(booktabs=T, caption=" ") %>% kable_classic(full_width=F)
```

Then, we will look at the variables with a scatterplot matrix and histograms for each variable. If the histograms revealed especially long tails or wierd outliers, we might want to transform the data, recode or delete outliers, etc.

Univariable distributions:

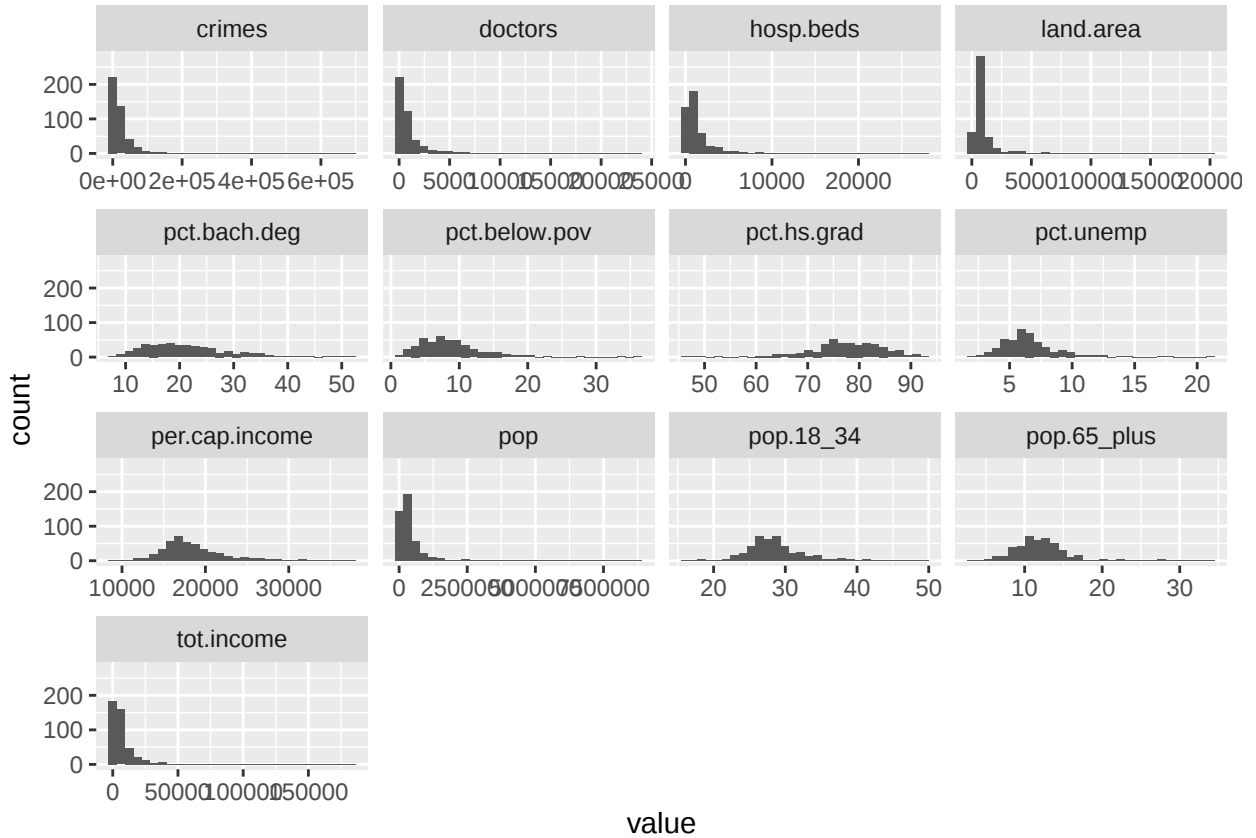
```
# histograms for all numeric predictor variables
ggplot(gather(cdinumeric), aes(value)) +
  geom_histogram(bins = 30) +
  facet_wrap(~key, scales = 'free_x')
```

Table 2:

	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 3:

	NC	NE	S	W
Freq	108	103	152	77



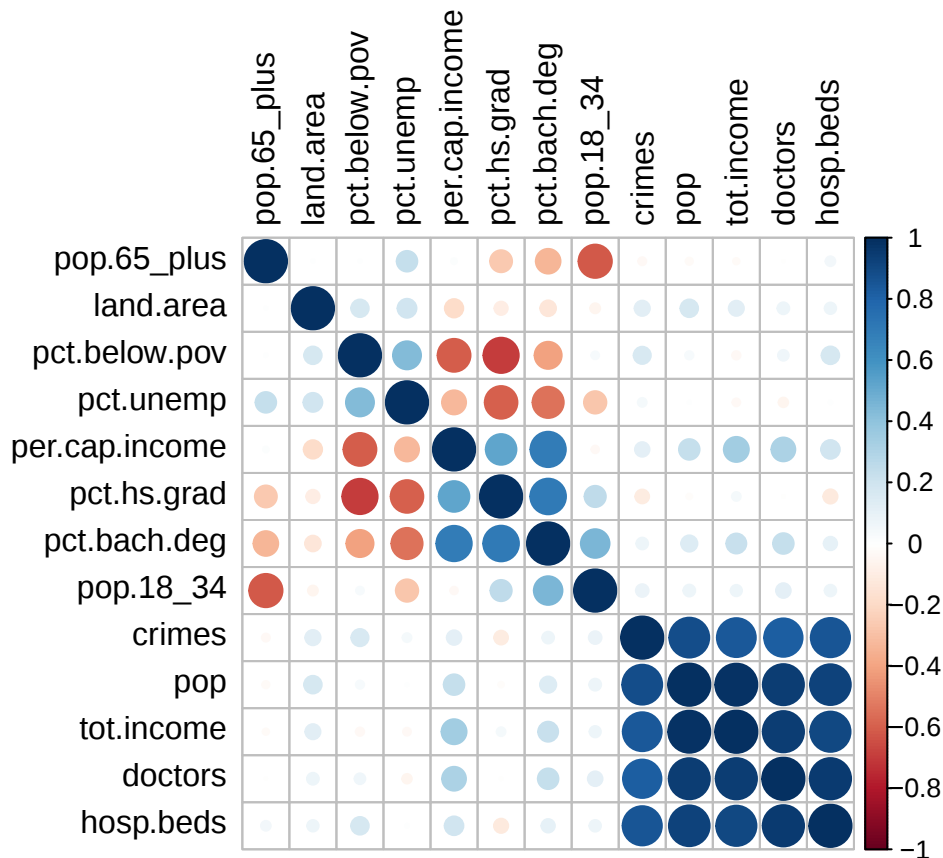
As we can see, `land.area`, `pop`, `tot.income`, `doctors`, `hosp.beds`, `crimes` and maybe `per.cap.income`

are severely skewed, which are the same variables that we identified from above. Thus, we need to consider transformations of these 7 predictor variables.

Bivariate relationships:

Heatmap of the correlation matrix with a large number of variables

```
co <- cor(cdinumeric)
corrplot::corrplot(co, order = "hclust", tl.col="black")
```



The correlation plot suggests that we may run into multi-collinearity problems. Now, we will look at the scatterplot matrix of the data.

```
# make a plot with all predictor variables
cdi2 <- data.frame(cdinumeric, region=cdi$region)

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

```

        ggtitle("") + xlab(xvar) + ylab(yvar)
      }
      result <- c(result,list(p))
    }
    return(result)
  }

grid.arrange(grobs=scatter.builder(cdi2))

```

We are going to take the logarithms of the variables that we identified earlier, to address heavy skewing and potential leverage and influence issues. The reason for using logarithms is that the transformed variables can be explained to the social scientist in terms of percent-change concepts.

```

cdilogs <- cdi2

skewed.vars <- c("land.area", "pop", "doctors", "hosp.beds", "crimes", "tot.income")

for (tmp in skewed.vars) {
  loc <- grep(paste("^",tmp,"$",sep=""),names(cdilogs))
  cdilogs[,loc] <- log(cdilogs[,loc])
  names(cdilogs)[loc] <- paste("log.",names(cdilogs)[loc],sep="")
}

```

Then, we start fitting the first model.

```

# regression
fit <- lm(per.cap.income ~ ., data = cdi2)
summary(fit)

##
## Call:
## lm(formula = per.cap.income ~ ., data = cdi2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7592.2  -879.1   -67.3    763.8   7714.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.887e+04  2.107e+03  13.703  < 2e-16 ***
## land.area    7.795e-02  6.267e-02   1.244  0.214279
## pop         -1.689e-02  1.364e-03 -12.377  < 2e-16 ***
## pop.18_34   -2.525e+02  2.640e+01  -9.566  < 2e-16 ***
## pop.65_plus -3.430e+01  2.635e+01  -1.302  0.193646
## doctors     -5.458e-01  2.175e-01  -2.509  0.012464 *
## hosp.beds    9.254e-01  1.522e-01   6.081  2.67e-09 ***
## crimes       1.082e-02  3.273e-03   3.307  0.001025 **
## pct.hs.grad -1.107e+02  2.352e+01  -4.708  3.40e-06 ***
## pct.bach.deg  3.624e+02  1.975e+01  18.347  < 2e-16 ***
## pct.below.pov -3.380e+02  2.945e+01 -11.479  < 2e-16 ***
## pct.unemp     1.631e+02  4.606e+01   3.542  0.000441 ***
## tot.income    7.251e-01  5.921e-02  12.246  < 2e-16 ***

```

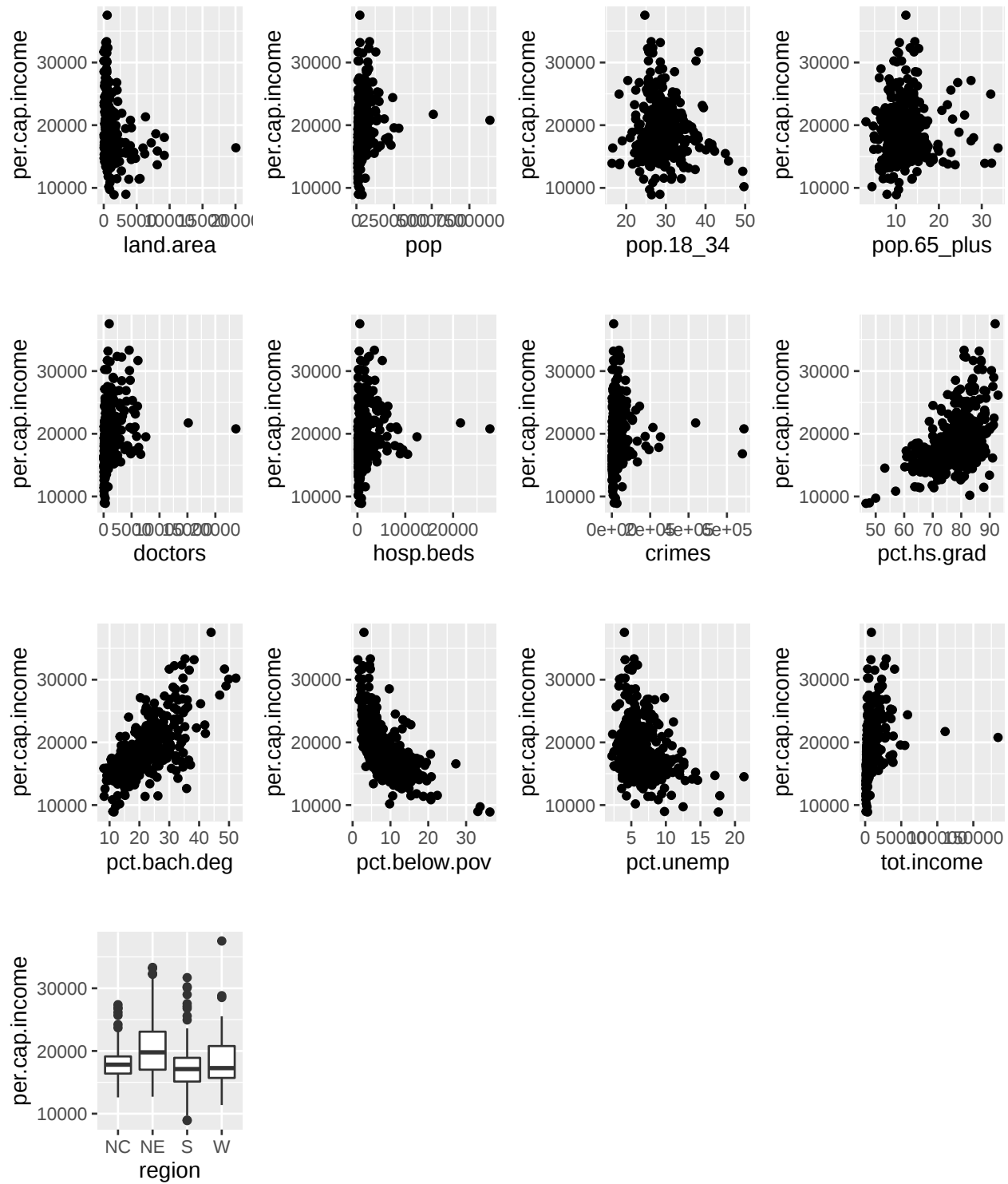
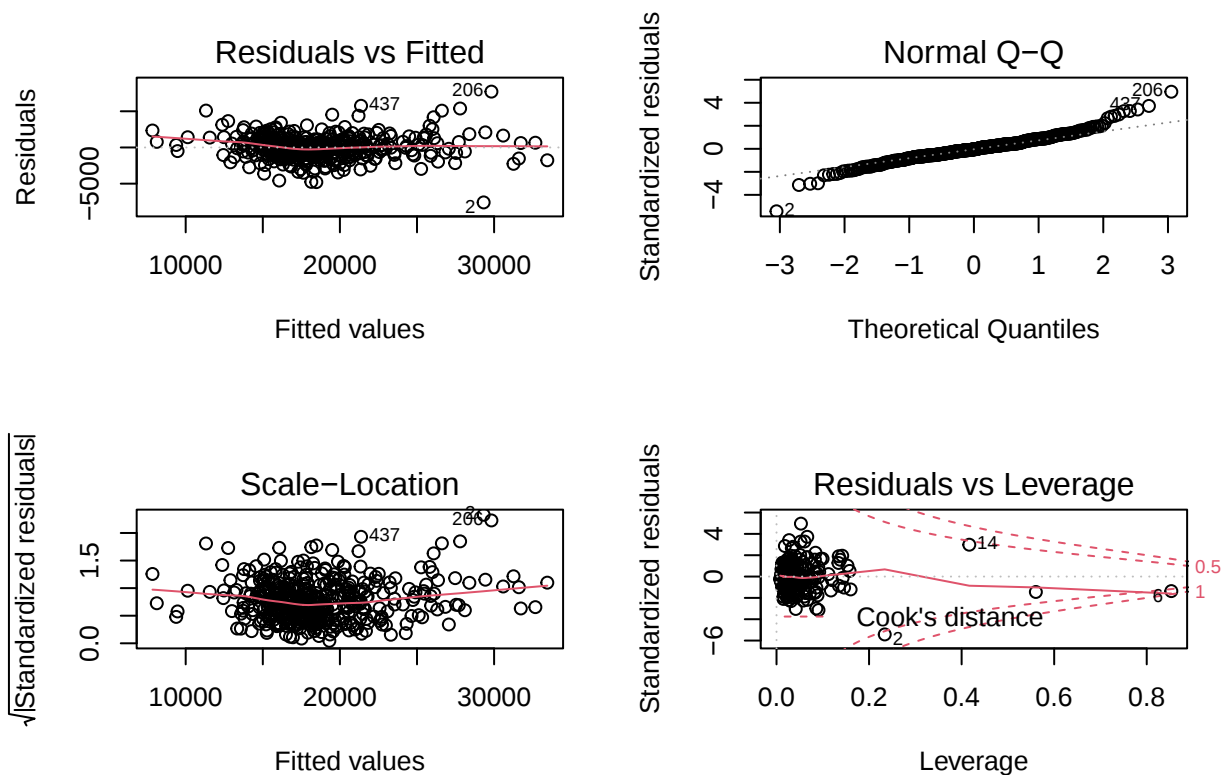


Figure 1: Scatter Plots with $y = \text{per.cap.income}$

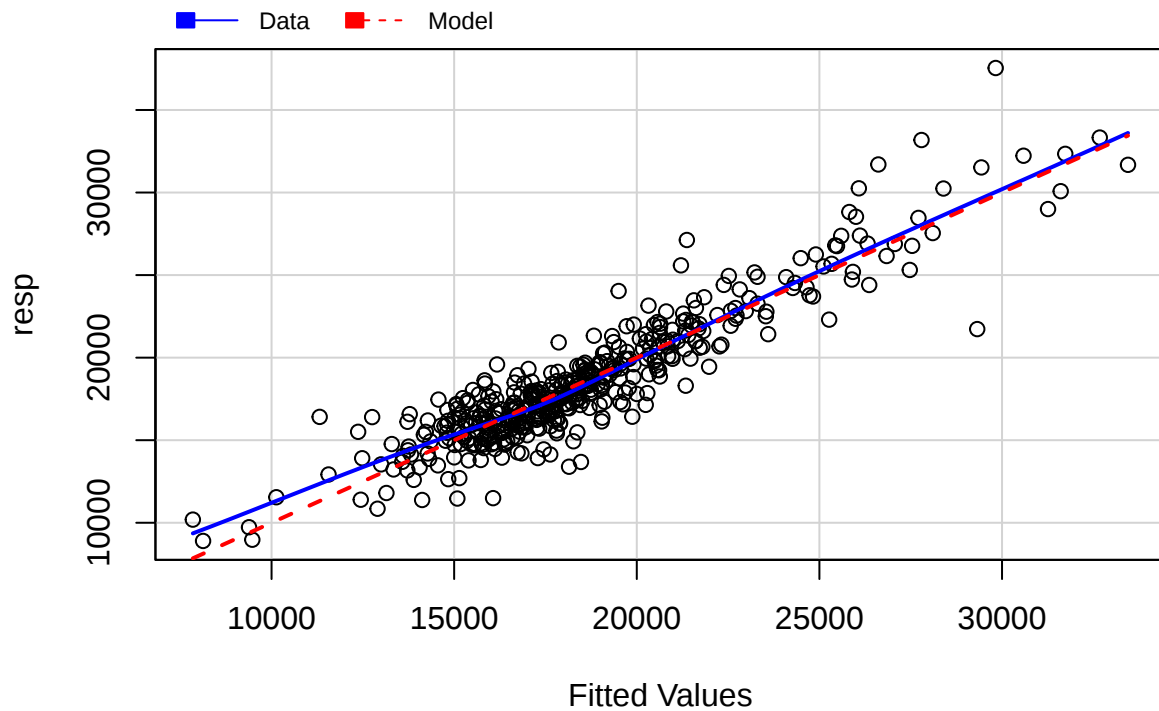
```
## regionNE      2.461e+02  2.432e+02   1.012 0.312099
## regionS       -3.727e+02  2.292e+02  -1.626 0.104739
## regionW       -2.157e+02  2.875e+02  -0.750 0.453528
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1598 on 424 degrees of freedom
## Multiple R-squared:  0.8502, Adjusted R-squared:  0.8449
## F-statistic: 160.5 on 15 and 424 DF,  p-value: < 2.2e-16
```

```
# the, we check the diagnostic plots
par(mfrow=c(2,2))
plot(fit)
```



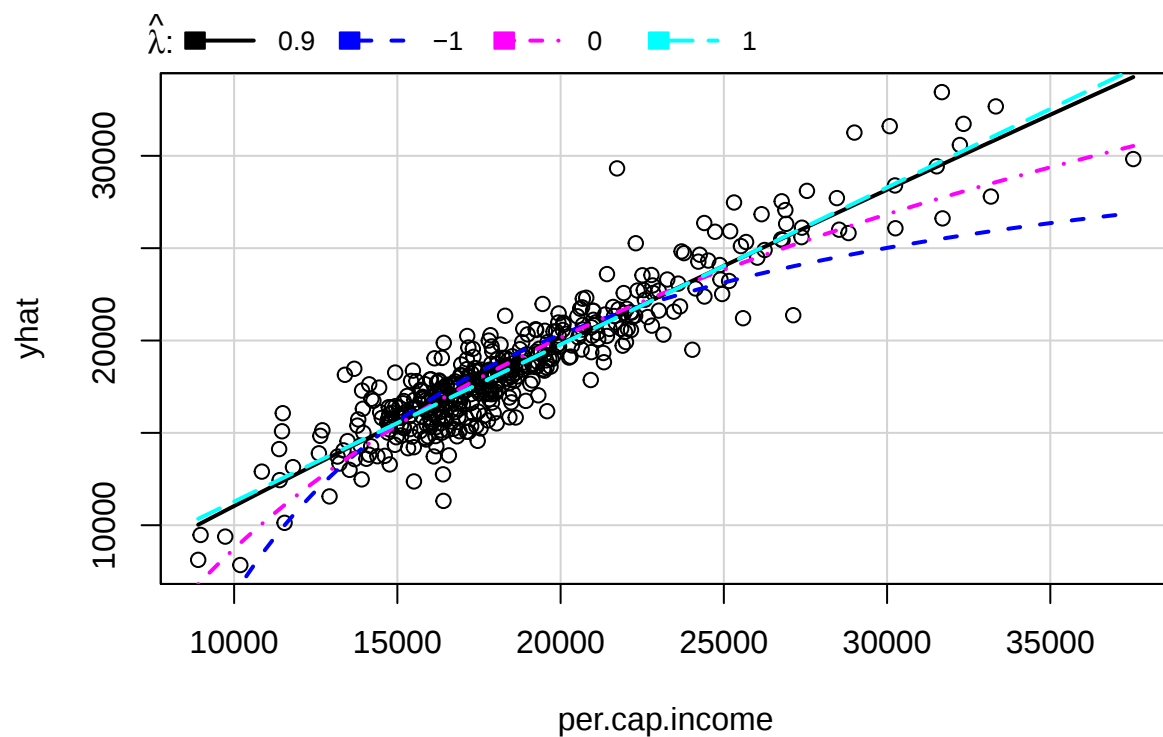
Both summary and the diagnostic plots look pretty great, except a little concave in standardized residuals against fitted value.

```
# marginal model plot for y and fitted values
par(mfrow=c(1,1))
m1 <- lm(cdi2$per.cap.income~fit$fitted.values)
mmp(m1,fit$fitted.values,xlab="Fitted Values",key=NULL)
```



Most part of marginal model plot for y and fitted values looks good, though there is departure in the left tail. Thus, we also consider transformation for per.cap.income.

```
# inverse response plot for the data
inverseResponsePlot(fit, key=TRUE)
```



```
##      lambda      RSS
```

```
## 1  0.8997968  919453119
## 2 -1.0000000 1511669678
## 3  0.0000000 1051021057
## 4  1.0000000  921021431
```

As we can see in the plot, $\lambda = 0$, which corresponds to natural logarithms provides a very good fit, as its fitted line is very close the optimal $\lambda = -0.11$'s. Thus, for better interpretation of the model, we choose natural logarithm transformation for y.

```
cdilogs <- cdi2

skewed.vars <- c("land.area", "pop", "doctors", "hosp.beds", "crimes", "tot.income", "per.cap.income")

for (tmp in skewed.vars) {
  loc <- grep(paste("^",tmp,"$",sep=""),names(cdilogs))
  cdilogs[,loc] <- log(cdilogs[,loc])
  names(cdilogs)[loc] <- paste("log.",names(cdilogs)[loc],sep="")
}
```

Let's check the transformed data.

```
# histogram of transformed data
hist.builder <- function(df) { ## creates a list of graphs
  result <- NULL
  for (var in names(df)) {
    d <- data.frame(dd=df[,var])
    if(mode(df[,var])=="numeric") {
      p <- ggplot(d,aes(x=dd)) + geom_histogram(bins = 30) +
        ggtitle(var) + xlab("")
    } else {
      p <- ggplot(d,aes(x=dd)) + geom_bar() +
        ggtitle(var) + xlab("")
    }
    result <- c(result,list(p))
  }
  return(result)
}

grid.arrange(grobs=hist.builder(cdilogs))
```

```
# Correlations after log transformations
co <- cor(cdilogs[,-grep("region",names(cdilogs))])
corrplot::corrplot(co, order = "hclust", tl.col="black")
```

```
# Scatterplots after log transformations
grid.arrange(grobs=scatter.builder(cdilogs,"log.per.cap.income"))
```

Except for log.pop, the skewing seems to have largely been brought under control.

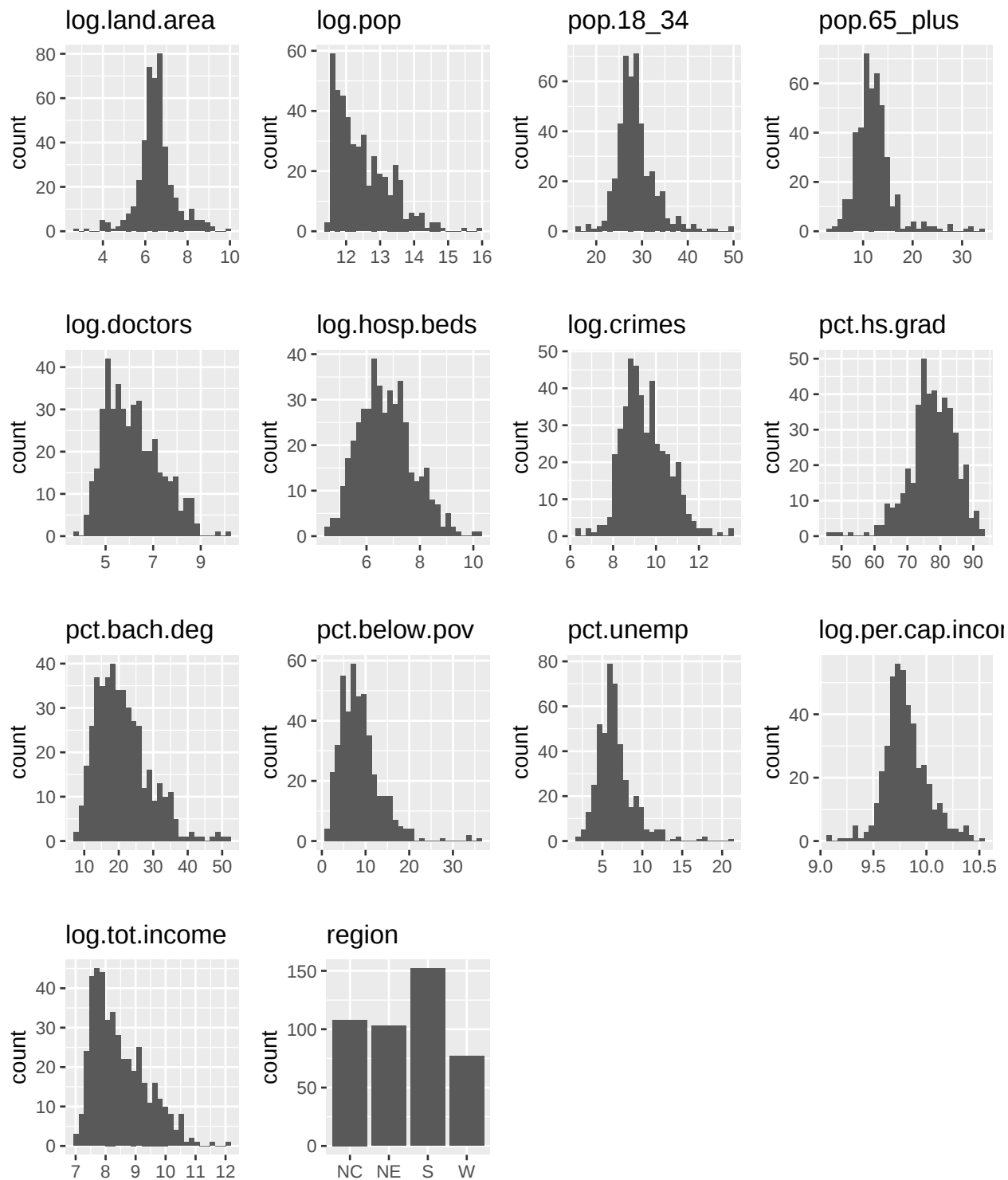


Figure 2: Histograms after log transformations

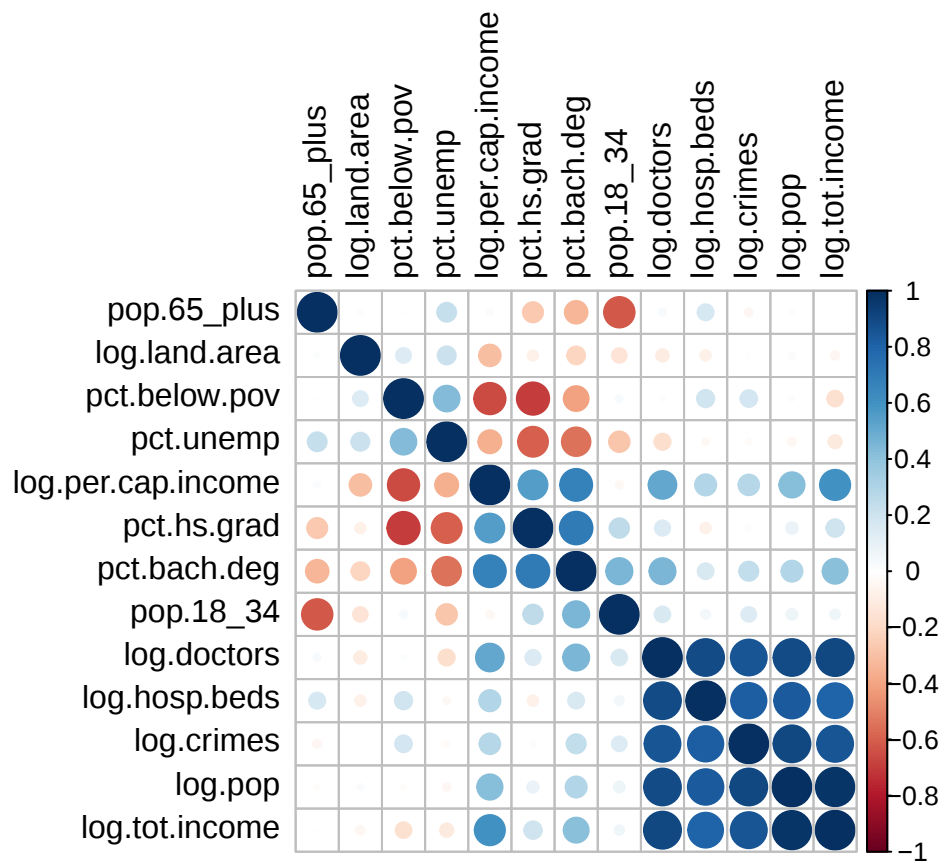


Figure 3: Correlations after log transformations

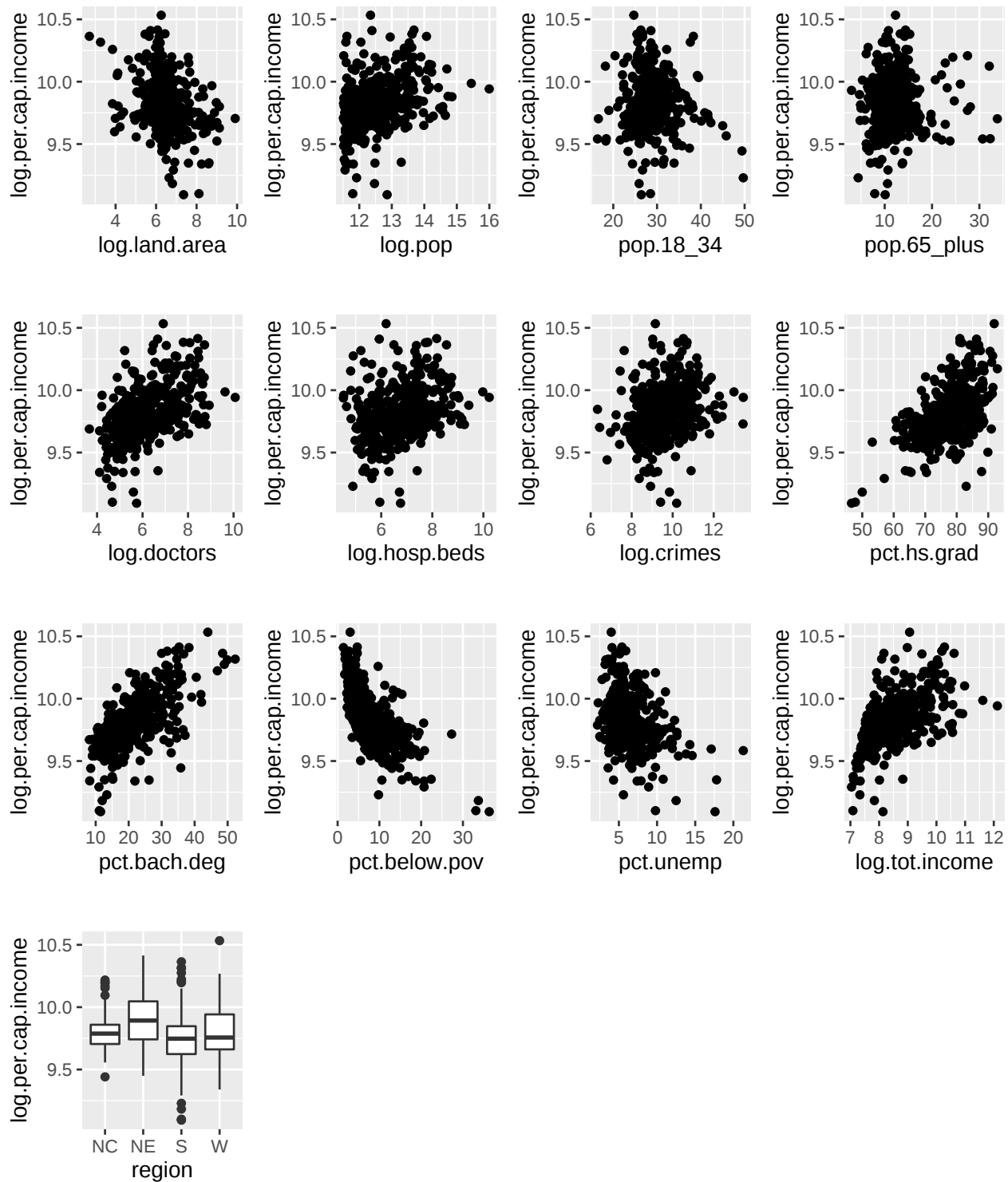


Figure 4: Scatterplots after log transformations

	df	AIC	BIC
ancova.01	3	-169.9466	-157.6863
ancova.02	6	-227.4746	-202.9539
ancova.03	9	-223.7402	-186.9593
ancova.04	3	-135.0340	-122.7737
ancova.05	6	-172.1347	-147.6140
ancova.06	9	-166.7601	-129.9792

Appendix 2. Analysis of Relationship between Per ca-pita income and Crime Rate

For log.per.cap.income and log.crimes, the six models to consider are:

```
ancova.01 <- lm(log.per.cap.income ~ log.crimes,data=cdilogs)

ancova.02 <- lm(log.per.cap.income ~ log.crimes + region,data=cdilogs)

ancova.03 <- lm(log.per.cap.income ~ log.crimes * region,data=cdilogs)

# regression model with per-capita crime
attach(cdi2)
log.per.cap.crimes <-log(crimes) - log(pop)
detach()

ancova.04 <- lm(log.per.cap.income ~ log.per.cap.crimes,data=cdilogs)

ancova.05 <- lm(log.per.cap.income ~ log.per.cap.crimes + region,data=cdilogs)

ancova.06 <- lm(log.per.cap.income ~ log.per.cap.crimes * region,data=cdilogs)
```

We could compare all 6 models with AIC or BIC. In this case, we get the same result: the additive model ancova.02 has both the lowest AIC and the lowest BIC values.

```
# table comparing all 6 models with AIC or BIC
data.frame(AIC=AIC(ancova.01,ancova.02,ancova.03,ancova.04,ancova.05,ancova.06),
BIC=BIC(ancova.01,ancova.02,ancova.03,ancova.04,ancova.05,ancova.06))[, -3] %>%
  kbl(booktabs=T,col.names=c("df","AIC","BIC")) %>% kable_classic(full_width=F)
```

However, there is an argument that per-capita crime is more comparable to, or at least on the same scale as, per-capita income, so we will briefly look at the second-best model, model 5 (anova.05), to see how it compares to model 2 (ancova.02):

```
oldmar <- par()$mar

par(mfrow=c(2,4))

par(mar=c(2,2,2,2))

invisible(lapply(list(ancova.02,ancova.05),
  function(x) plot(x,cex.main=0.5)))
```

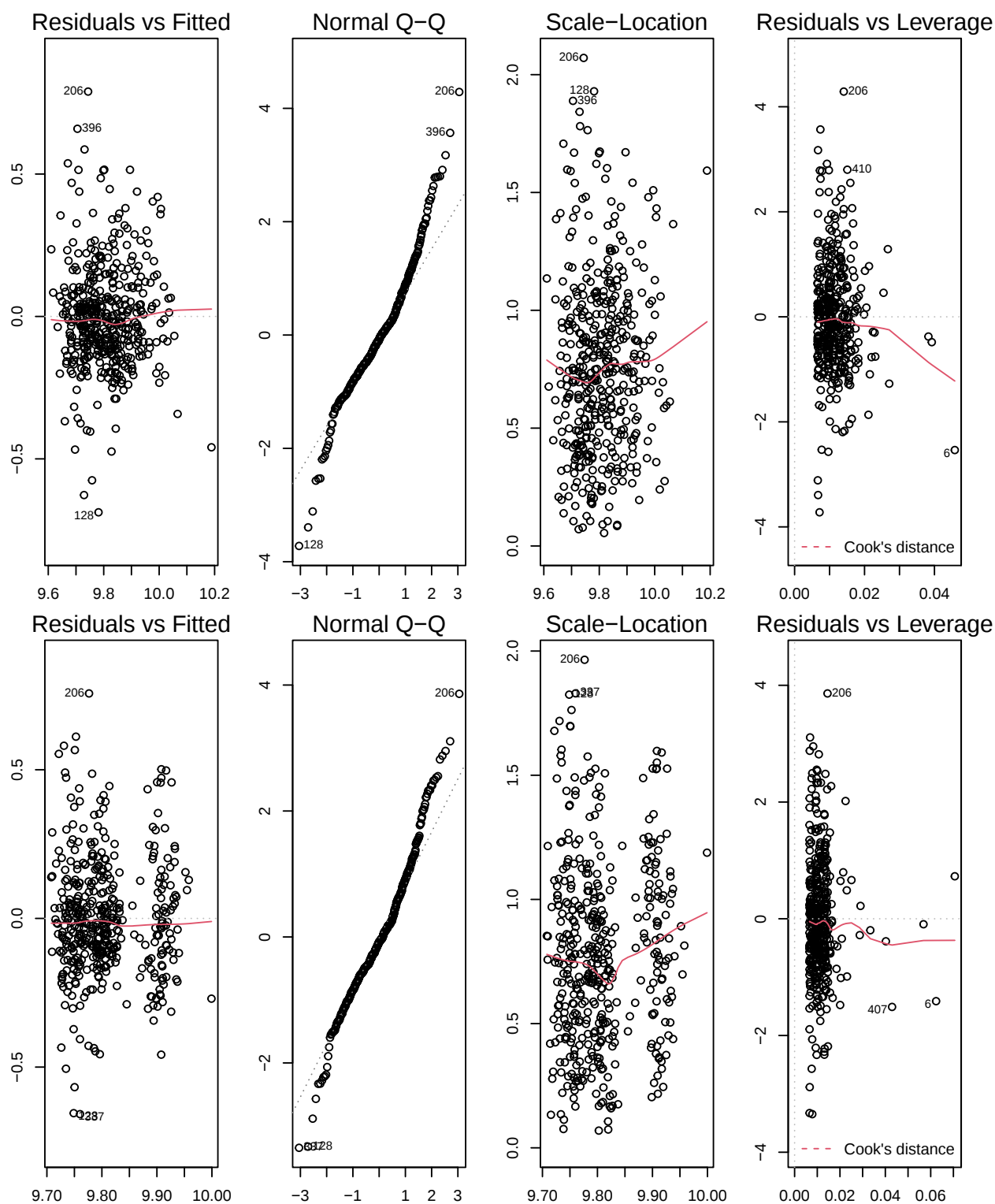


Figure 5: Residual diagnostics for all 2 ANCOVA models, in order: ancova.02, ancova.05

```
par(mar=oldmar)
```

Both the diagnostic plots don't show much except that the QQ plot suggests both the left and the right tails are a bit longer than expected for the normal distribution. Thus, for better interpretation, we finally decide on the model 5.

```
formula(ancova.05)
```

```
## log.per.cap.income ~ log.per.cap.crimes + region
```

```
round(coef(summary(ancova.05)),2)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.94        0.07  143.30    0.00
## log.per.cap.crimes  0.04        0.02   1.98    0.05
## regionNE          0.11        0.03   4.15    0.00
## regionS          -0.07        0.03  -2.84    0.00
## regionW          -0.02        0.03  -0.81    0.42
```

Appendix 3. Regression Analysis – All Subsets(without Regions)

Per-capita variables are more comparable to, or at least on the same scale as, per-capita income. So, we will convert land area, doctors, hospital beds, crimes and total income into per ca-pita scales.

```
# conversion for the data
cdiperlogs <- cdi2

loc <- grep(paste("^","pop","$",sep=""), names(cdiperlogs))
cdiperlogs[,loc] <- log(cdiperlogs[,loc])
names(cdiperlogs)[loc] <- paste("log.",names(cdiperlogs)[loc],sep="")

loc <- grep(paste("^","per.cap.income","$",sep=""), names(cdiperlogs))
cdiperlogs[,loc] <- log(cdiperlogs[,loc])
names(cdiperlogs)[loc] <- paste("log.",names(cdiperlogs)[loc],sep="")

vars <- c("land.area", "doctors", "hosp.beds", "crimes", "tot.income")

for (tmp in skewed.vars) {
  loc <- grep(paste("^",tmp,"$",sep=""),names(cdiperlogs))
  cdiperlogs[,loc] <- log(cdiperlogs[,loc]) - cdilogs$log.pop
  names(cdiperlogs)[loc] <- paste("log.per.cap.",names(cdiperlogs)[loc],sep="")
}

# take log.pop and log.tot.income out of consideration
omit <- c(grep("log.pop",names(cdilogs)),grep("log.tot.income",names(cdilogs)))
cdilogred <- cdiperlogs[,-omit]

# First, work *without* the `region` variable
cdilogred.cont <- cdilogred[,-grep("region",names(cdilogred))]
names(cdilogred.cont)
```

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

Locate the best model with the lowest BIC...

```
all.subsets.01 <- regsubsets(log.per.cap.income ~ ., data=cdilogred.cont,nvmax=10)

# coefficients and standard errors for the best model
all.subsets.01.summary <- summary(all.subsets.01)
best.model <- which.min(all.subsets.01.summary$bic)

tmp <- cdilogred.cont[,all.subsets.01.summary$which[best.model,][~1]]
all.subsets.01.final.model <- lm(log.per.cap.income ~ .,data=tmp)
summary(all.subsets.01.final.model)$coef
```

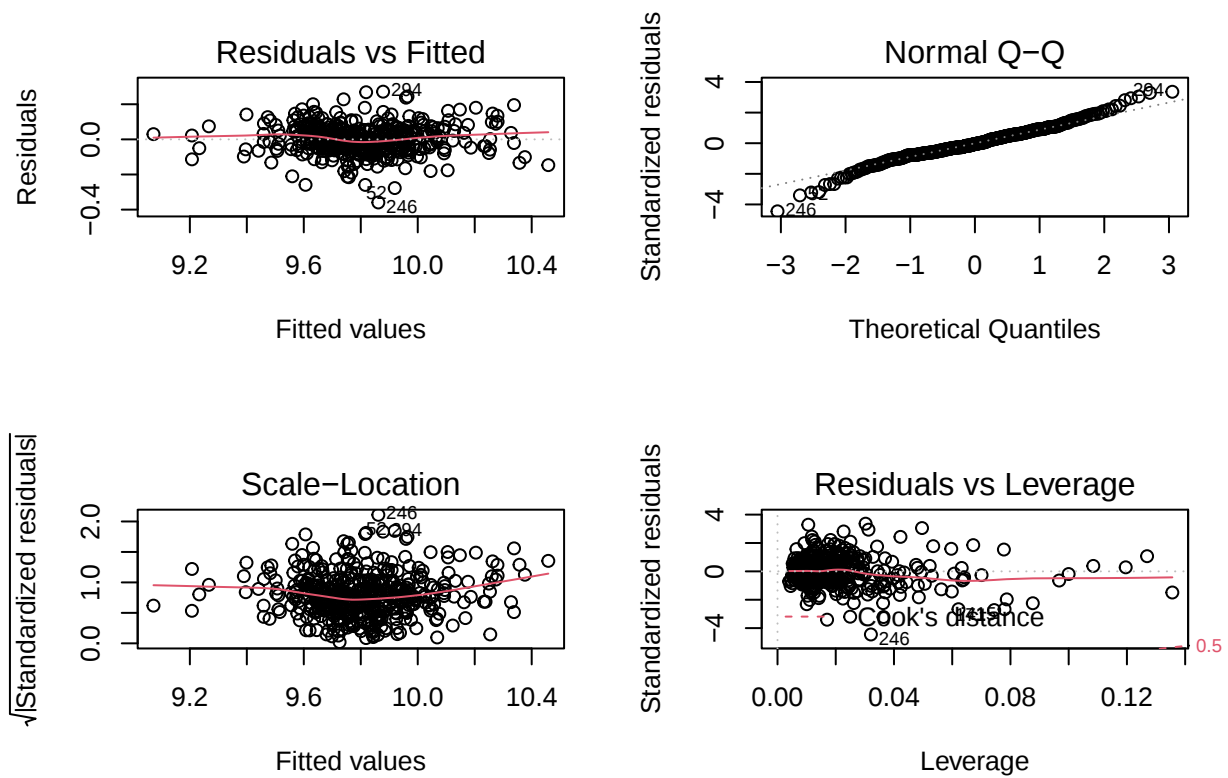
	Estimate	Std. Error	t value	Pr(> t)
## (Intercept)	10.785892327	0.142120643	75.892510	1.623250e-251
## log.per.cap.land.area	-0.042512925	0.004017583	-10.581717	2.035599e-23
## pop.18_34	-0.016109168	0.001306346	-12.331471	3.902227e-30
## pop.65_plus	-0.004040904	0.001384429	-2.918824	3.697586e-03
## log.per.cap.doctors	0.087212905	0.010349063	8.427131	5.350058e-16
## pct.hs.grad	-0.004522402	0.001087721	-4.157687	3.880188e-05
## pct.bach.deg	0.014868243	0.000958610	15.510211	1.984929e-43
## pct.below.pov	-0.025183140	0.001329506	-18.941728	1.177048e-58
## pct.unemp	0.012980498	0.002197574	5.906739	7.087581e-09

All the predictors have coefficients 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`). Before diving into what might be going on, let's at least check VIFs and residual diagnostics...

```
vif(all.subsets.01.final.model)
```

	pop.18_34	pop.65_plus
## log.per.cap.land.area	1.953668	1.991357
## log.per.cap.doctors	3.794826	3.509152
## pct.below.pov	pct.unemp	
##	1.720401	

```
par(mfrow=c(2,2))
# diagnostics for the best all subsets model
plot(all.subsets.01.final.model)
```

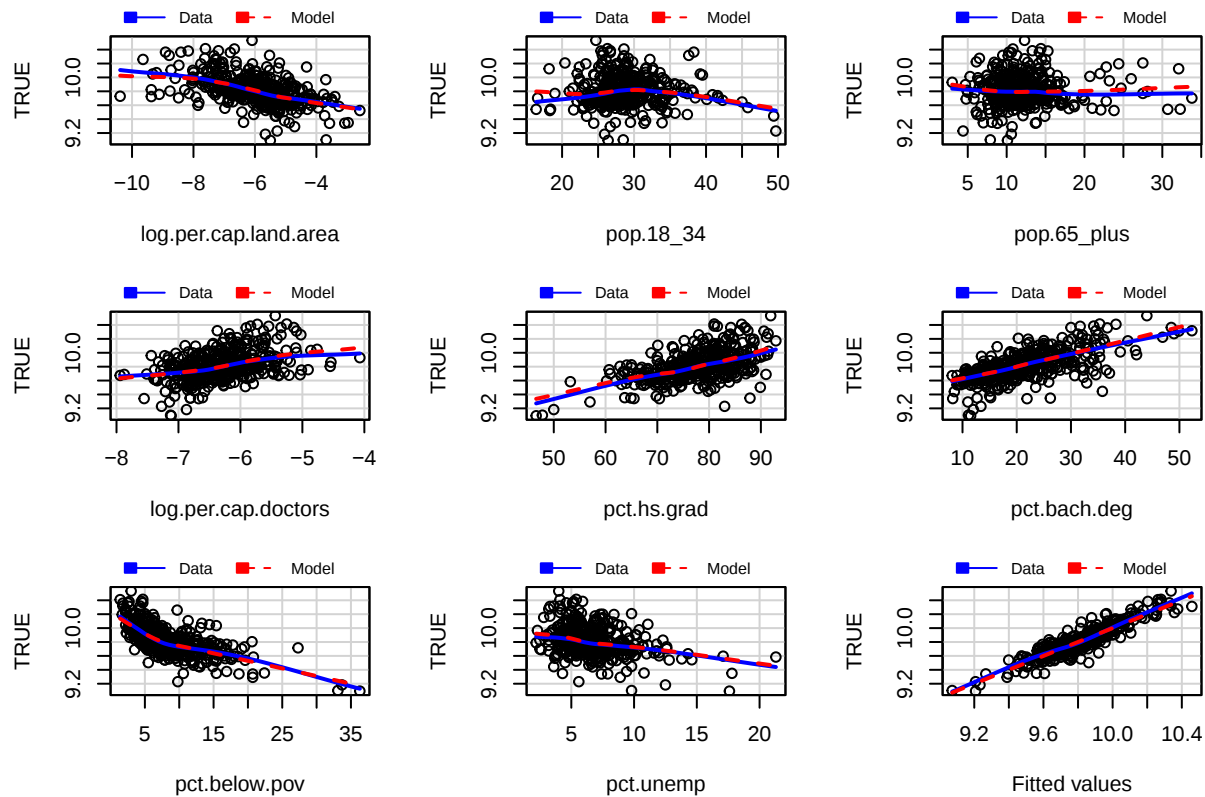


None of the VIFs seem excessively large, and the diagnostic plots don't show much except that the QQ plot suggests both the left and the right tails are a bit longer than expected for the normal distribution.

We can also look at marginal model plots, to see if we are missing a transformation, interaction, etc.

```
mmpr(all.subsets.01.final.model)
```

Marginal Model Plots



The marginal model plots look very good – the blue data-based curves line up well with the red model-based curves. We don't seem to be missing any important transformations, interactions, etc.

Appendix 3. Regression Analysis – All Subsets(with Regions)

```
# add regions into regression
tmp <- cbind(tmp,region=cdilogred$region)
all.subsets.01.final.with.region <- lm(log.per.cap.income ~ .*region,data=tmp)
summary(all.subsets.01.final.with.region)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ . * region, data = tmp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.243581 -0.043603 -0.000872  0.040701  0.295427
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.135880   0.375175  27.016 < 2e-16 ***
## log.per.cap.land.area -0.051212   0.009442  -5.424 1.01e-07 ***
## pop.18_34      -0.014841   0.002811  -5.279 2.13e-07 ***
## pop.65_plus      0.001018   0.005243   0.194 0.846093
## log.per.cap.doctors  0.044700   0.018788   2.379 0.017815 *
```

```

## pct.hs.grad          -0.002202    0.003450   -0.638  0.523774
## pct.bach.deg          0.014946    0.002924    5.111  4.95e-07 ***
## pct.below.pov        -0.022998    0.003912   -5.879  8.67e-09 ***
## pct.unemp             0.018630    0.005028    3.705  0.000241 ***
## regionNE              0.598161    0.511699    1.169  0.243105
## regionS               0.317762    0.423040    0.751  0.453006
## regionW               2.749818    0.561496    4.897  1.41e-06 ***
## log.per.cap.land.area:regionNE 0.002252    0.013452    0.167  0.867157
## log.per.cap.land.area:regionS 0.009275    0.012258    0.757  0.449697
## log.per.cap.land.area:regionW 0.039681    0.013379    2.966  0.003197 **
## pop.18_34:regionNE   -0.006445    0.004235   -1.522  0.128830
## pop.18_34:regionS    -0.001493    0.003409   -0.438  0.661732
## pop.18_34:regionW     0.002727    0.004663    0.585  0.558987
## pop.65_plus:regionNE -0.009749    0.006618   -1.473  0.141503
## pop.65_plus:regionS   -0.002213    0.005518   -0.401  0.688669
## pop.65_plus:regionW   -0.006832    0.006995   -0.977  0.329256
## log.per.cap.doctors:regionNE 0.009289    0.031949    0.291  0.771392
## log.per.cap.doctors:regionS 0.034939    0.024605    1.420  0.156377
## log.per.cap.doctors:regionW 0.119386    0.037063    3.221  0.001380 **
## pct.hs.grad:regionNE -0.003461    0.004415   -0.784  0.433613
## pct.hs.grad:regionS    0.001959    0.003807    0.514  0.607216
## pct.hs.grad:regionW   -0.019334    0.004587   -4.215  3.09e-05 ***
## pct.bach.deg:regionNE  0.005030    0.004053    1.241  0.215308
## pct.bach.deg:regionS   -0.002642    0.003201   -0.825  0.409608
## pct.bach.deg:regionW    0.003287    0.003677    0.894  0.371858
## pct.below.pov:regionNE -0.001206    0.005302   -0.228  0.820138
## pct.below.pov:regionS  0.005633    0.004404    1.279  0.201588
## pct.below.pov:regionW  -0.020050    0.005598   -3.582  0.000383 ***
## pct.unemp:regionNE     -0.007415    0.007575   -0.979  0.328233
## pct.unemp:regionS      -0.022264    0.006929   -3.213  0.001417 **
## pct.unemp:regionW      -0.018451    0.006889   -2.678  0.007699 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07476 on 404 degrees of freedom
## Multiple R-squared:  0.8797, Adjusted R-squared:  0.8692
## F-statistic: 84.37 on 35 and 404 DF,  p-value: < 2.2e-16

```

Follow a rule of thumb is: if any indicator for a categorical variable seems important (e.g. a statistically significant coefficient), then keep the whole categorical variable. If none of them seem important, then drop the variable. The same thing works for interactions with categorical variables. Thus, we arrive at

```

all.subsets.01.final.with.some.region <- update(all.subsets.01.final.with.region,
  . ~ . - region:pop.65_plus - region:pop.18_34 - region:pct.bach.deg)
summary(all.subsets.01.final.with.some.region)

```

```

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

```

```
##
## Residuals:
##      Min        1Q      Median        3Q      Max
## -0.258132 -0.045038 -0.001107  0.040222  0.295175
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.2679794  0.2847552  36.059 < 2e-16 ***
## log.per.cap.land.area -0.0499873  0.0085299  -5.860 9.46e-09 ***
## pop.18_34      -0.0154108  0.0012915 -11.933 < 2e-16 ***
## pop.65_plus    -0.0018873  0.0014218  -1.327 0.185117
## log.per.cap.doctors  0.0482811  0.0155591   3.103 0.002047 **
## pct.hs.grad    -0.0028499  0.0024993  -1.140 0.254832
## pct.bach.deg    0.0149133  0.0009848  15.143 < 2e-16 ***
## pct.below.pov  -0.0225437  0.0037358  -6.035 3.55e-09 ***
## pct.unemp       0.0181418  0.0048832   3.715 0.000231 ***
## regionNE       -0.0328403  0.3970457  -0.083 0.934121
## regionS         0.3236732  0.3242043   0.998 0.318689
## regionW         2.4940100  0.4221170   5.908 7.23e-09 ***
## log.per.cap.land.area:regionNE -0.0104496  0.0118883  -0.879 0.379924
## log.per.cap.land.area:regionS  0.0135607  0.0114089   1.189 0.235280
## log.per.cap.land.area:regionW  0.0268791  0.0120422   2.232 0.026145 *
## log.per.cap.doctors:regionNE  0.0062279  0.0272072   0.229 0.819057
## log.per.cap.doctors:regionS  0.0307224  0.0210523   1.459 0.145233
## log.per.cap.doctors:regionW  0.1193144  0.0299993   3.977 8.23e-05 ***
## pct.hs.grad:regionNE  0.0013270  0.0030688   0.432 0.665654
## pct.hs.grad:regionS  0.0001180  0.0026485   0.045 0.964471
## pct.hs.grad:regionW -0.0162969  0.0035498  -4.591 5.86e-06 ***
## pct.below.pov:regionNE -0.0044605  0.0050367  -0.886 0.376340
## pct.below.pov:regionS  0.0032457  0.0041796   0.777 0.437865
## pct.below.pov:regionW -0.0194333  0.0054923  -3.538 0.000448 ***
## pct.unemp:regionNE -0.0097674  0.0074850  -1.305 0.192644
## pct.unemp:regionS -0.0177148  0.0067103  -2.640 0.008606 **
## pct.unemp:regionW -0.0170169  0.0067854  -2.508 0.012529 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07595 on 413 degrees of freedom
## Multiple R-squared:  0.873, Adjusted R-squared:  0.865
## F-statistic: 109.2 on 26 and 413 DF, p-value: < 2.2e-16
```

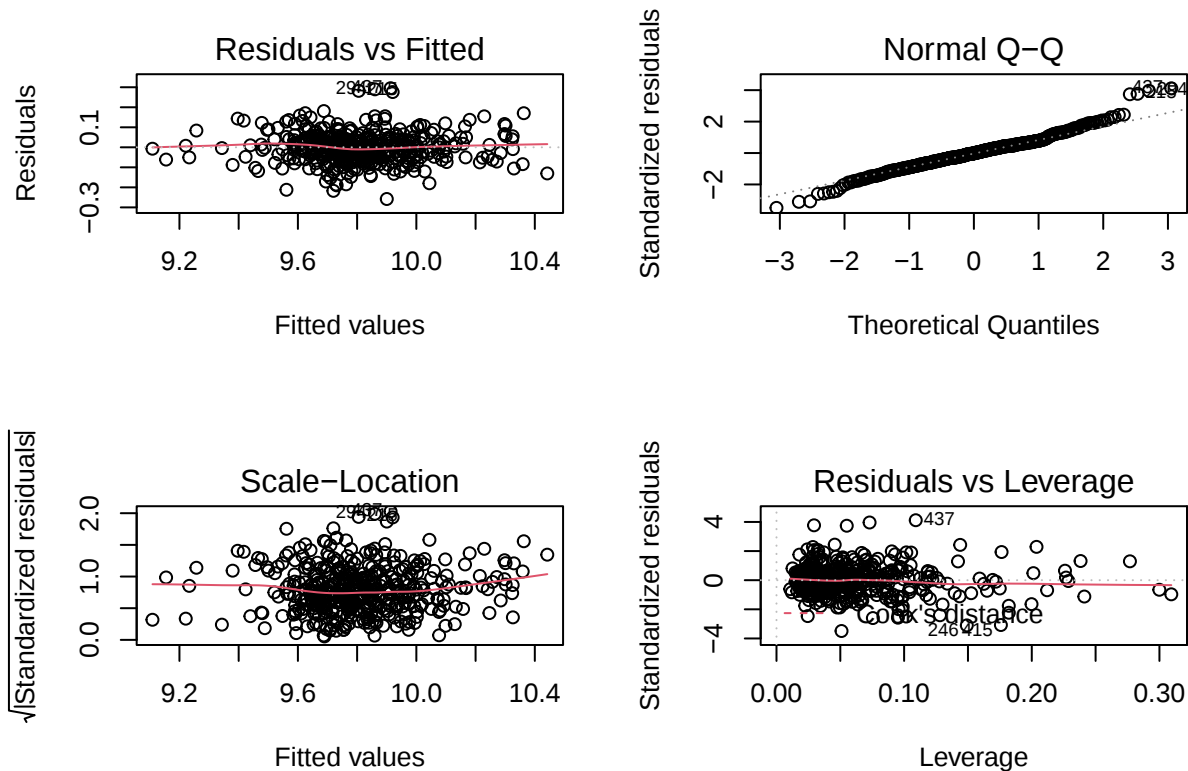
Let's quickly check VIFs and diagnostics.

```
vif(all.subsets.01.final.with.some.region)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## log.per.cap.land.area    7.460742e+00  1    2.731436
## pop.18_34                2.229411e+00  1    1.493121
## pop.65_plus              2.452456e+00  1    1.566032
## log.per.cap.doctors      5.699395e+00  1    2.387341
## pct.hs.grad              2.339401e+01  1    4.836735
## pct.bach.deg              4.324473e+00  1    2.079537
## pct.below.pov            2.303039e+01  1    4.798999
```

```
## pct.unemp                9.918770e+00  1      3.149408
## region                   1.985483e+09  3      35.452294
## log.per.cap.land.area:region 9.879657e+04  3      6.799187
## log.per.cap.doctors:region  2.069430e+07  3      16.569464
## pct.hs.grad:region         1.037548e+08  3      21.677109
## pct.below.pov:region       9.128360e+03  3      4.571571
## pct.unemp:region           1.689729e+04  3      5.065661
```

```
par(mfrow=c(2,2))
plot(all.subsets.01.final.with.some.region)
```



Then compare to the best BIC model that we obtained from stepwise without the `region` term.

```
# comparison
anova(all.subsets.01.final.model,all.subsets.01.final.with.some.region)

## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.per.cap.land.area + pop.18_34 + pop.65_plus +
##   log.per.cap.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##   pct.unemp
## Model 2: log.per.cap.income ~ log.per.cap.land.area + pop.18_34 + pop.65_plus +
##   log.per.cap.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##   pct.unemp + region + log.per.cap.land.area:region + log.per.cap.doctors:region +
##   pct.hs.grad:region + pct.below.pov:region + pct.unemp:region
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      431 2.9031
## 2      413 2.3825 18   0.52057 5.0133 2.38e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(all.subsets.01.final.model,all.subsets.01.final.with.some.region)
```

```
##                df      AIC
## all.subsets.01.final.model      10 -940.5743
## all.subsets.01.final.with.some.region 28 -991.5268
```

```
BIC(all.subsets.01.final.model,all.subsets.01.final.with.some.region)
```

```
##                df      BIC
## all.subsets.01.final.model      10 -899.7065
## all.subsets.01.final.with.some.region 28 -877.0971
```

The anova (F test) and AIC really like the model with the region terms in it. On the other hand, BIC prefers the simpler model. Finally, we will examine the table of estimated coefficients again for the model with some region terms in it.

```
formula(all.subsets.01.final.with.some.region)
```

```
## log.per.cap.income ~ log.per.cap.land.area + pop.18_34 + pop.65_plus +
##   log.per.cap.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##   pct.unemp + region + log.per.cap.land.area:region + log.per.cap.doctors:region +
##   pct.hs.grad:region + pct.below.pov:region + pct.unemp:region
```

```
round(summary(all.subsets.01.final.with.some.region)$coef,2)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)          10.27      0.28   36.06   0.00
## log.per.cap.land.area    -0.05      0.01   -5.86   0.00
## pop.18_34             -0.02      0.00  -11.93   0.00
## pop.65_plus            0.00      0.00   -1.33   0.19
## log.per.cap.doctors      0.05      0.02    3.10   0.00
## pct.hs.grad            0.00      0.00   -1.14   0.25
## pct.bach.deg            0.01      0.00   15.14   0.00
## pct.below.pov          -0.02      0.00   -6.03   0.00
## pct.unemp              0.02      0.00    3.72   0.00
## regionNE              -0.03      0.40   -0.08   0.93
## regionS               0.32      0.32    1.00   0.32
## regionW               2.49      0.42    5.91   0.00
## log.per.cap.land.area:regionNE -0.01      0.01   -0.88   0.38
## log.per.cap.land.area:regionS  0.01      0.01    1.19   0.24
## log.per.cap.land.area:regionW  0.03      0.01    2.23   0.03
## log.per.cap.doctors:regionNE  0.01      0.03    0.23   0.82
## log.per.cap.doctors:regionS  0.03      0.02    1.46   0.15
## log.per.cap.doctors:regionW  0.12      0.03    3.98   0.00
## pct.hs.grad:regionNE    0.00      0.00    0.43   0.67
## pct.hs.grad:regionS     0.00      0.00    0.04   0.96
## pct.hs.grad:regionW    -0.02      0.00   -4.59   0.00
## pct.below.pov:regionNE  0.00      0.01   -0.89   0.38
```

	df	AIC	BIC
all.subsets.01.final.model	10	-940.5743	-899.7065
step.result.01.aic	11	-942.0976	-897.1431
step.result.01.bic	10	-940.5743	-899.7065
step.result.02.aic	23	-1082.1848	-988.1890
step.result.02.bic	16	-1047.7658	-982.3774

```
## pct.below.pov:regionS      0.00      0.00      0.78      0.44
## pct.below.pov:regionW     -0.02      0.01     -3.54      0.00
## pct.unemp:regionNE        -0.01      0.01     -1.30      0.19
## pct.unemp:regionS         -0.02      0.01     -2.64      0.01
## pct.unemp:regionW         -0.02      0.01     -2.51      0.01
```

Appendix 4. Regression Analysis – Stepwise Regression

Two models in stepwise regression respectively using AIC and BIC.

```
stepwise.base <- lm(log.per.cap.income ~ ., data=cdilogred.cont)

## try to duplicate all-subsets with BIC
step.result.01.bic <- stepAIC(stepwise.base,
                             scope=list(lower = ~ 1, upper = ~ .),
                             k=log(dim(cdilogred.cont)[1]),
                             trace=F)

## now try with AIC
step.result.01.aic <- stepAIC(stepwise.base,
                             scope=list(lower = ~ 1, upper = ~ .),
                             k=2,
                             trace=F)
```

Well, now let's go for broke and see what happens if we just explore 2-way interactions briefly.

```
step.result.02.bic <- stepAIC(stepwise.base, scope=list(lower = ~ 1, upper = ~ .^2),
                             k=log(dim(cdilogred.cont)[1]),
                             trace=F) ## BIC penalty.

step.result.02.aic <- stepAIC(stepwise.base, scope=list(lower = ~ 1, upper = ~ .^2),
                             k=2,
                             trace=F) ## AIC penalty.

comparison <- cbind(
  AIC(all.subsets.01.final.model, step.result.01.aic, step.result.01.bic,
      step.result.02.aic, step.result.02.bic),
  BIC(all.subsets.01.final.model, step.result.01.aic, step.result.01.bic,
      step.result.02.aic, step.result.02.bic))
comparison <- comparison[, -3]
names(comparison) <- c("df", "AIC", "BIC")
comparison %>% kbl(booktabs=T) %>% kable_classic()
```

```

# compare the R2 of two models with interactions

# step.result.02.bic
cat("\nR2 = ",summary(step.result.02.bic)$r.squared)

##
## R2 = 0.8819856

cat("\nR2adj = ",summary(step.result.02.bic)$adj.r.squared)

##
## R2adj = 0.8780981

# step.result.02.aic
cat("\nR2 = ",summary(step.result.02.aic)$r.squared)

##
## R2 = 0.8942832

cat("\nR2adj = ",summary(step.result.02.aic)$adj.r.squared)

##
## R2adj = 0.8889721

```

Although both interactions models produced big jumps in AIC and BIC (much bigger than 10), the improvement in R^2 and R^2_{adj} is pretty small, for all the terms that have been added to the models. For these reasons we are willing to discuss these interactions with the social scientist, instead of including them in a final model.

Thus, we will stick with the best model found by all-subsets and stepAIC with a BIC penalty, then my conclusions about adding interactions with `region` will also be the same.

Appendix 5. Regression Analysis – LASSO

```

# shrinkage plot for LASSO
loc <- grep("log.per.cap.income",names(cdilogred.cont))

y <- cdilogred.cont[,loc]

X <- apply(as.matrix(cdilogred.cont[,-loc]),2,function(x) rescale(x,"full"))

Xnames <- dimnames(X)[[2]]

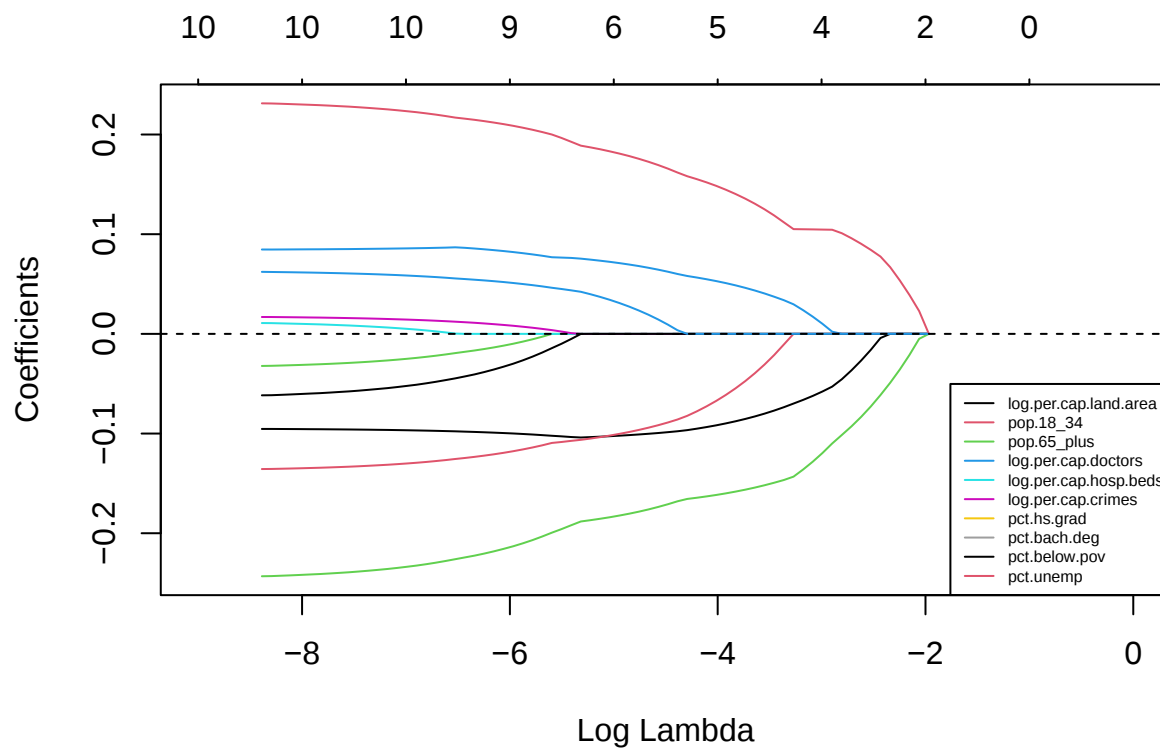
lasso.result <- glmnet(X,y)

plot(lasso.result,xvar="lambda",xlim=c(-9,0))

abline(h=0,lty=2)

legend('bottomright',lty=1,col=1:length(Xnames),legend=Xnames,cex=0.5)

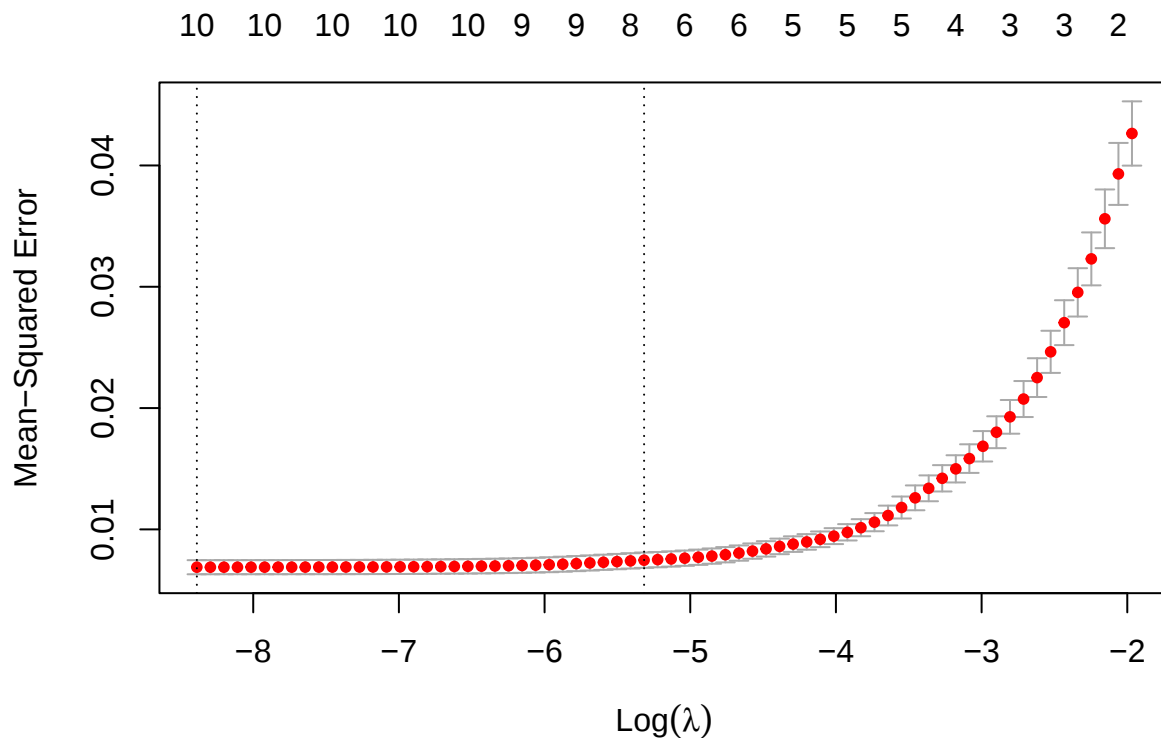
```



There is not such an obvious place to cut the shrinkage plot, to help determine what variables should be kept in the model.

```
# cross-validation
cv.lasso.result <- cv.glmnet(X,y)

plot(cv.lasso.result)
```



```
c(lambda.1se=cv.lasso.result$lambda.1se,lambda.min=cv.lasso.result$lambda.min)
```

```
##   lambda.1se   lambda.min
## 0.0049081710 0.0002278171
```

```
tmp <- cbind(coef(cv.lasso.result,s=cv.lasso.result$lambda.min),
             coef(cv.lasso.result,s=cv.lasso.result$lambda.1se)
            )
dimnames(tmp)[[2]] <- c("lambda(minMSE)", "lambda(minMSE+1se)")

tmp
```

```
## 11 x 2 sparse Matrix of class "dgCMatrix"
##               lambda(minMSE) lambda(minMSE+1se)
## (Intercept)      9.80695459      9.806955e+00
## log.per.cap.land.area -0.09536780 -1.039554e-01
## pop.18_34         -0.13561396 -1.062918e-01
## pop.65_plus       -0.03218231      .
## log.per.cap.doctors  0.08463289      7.547924e-02
## log.per.cap.hosp.beds  0.01086566      .
## log.per.cap.crimes   0.01693974      .
## pct.hs.grad       -0.06164630 -4.805525e-05
## pct.bach.deg        0.23126382      1.888050e-01
## pct.below.pov      -0.24313646 -1.881889e-01
## pct.unemp          0.06221462      4.221509e-02
```

It's interesting to note that the model which minimizes 10-fold cross-validation error contains all 10 predictors (or sometimes 9—log.crimes can get left out—depending on what random folds we get). The model that

is 1 SE has 2 less variable, which is pop.65_plus and pct.hs.grad than the model that we first saw with all-subsets regression.

Now let's compare these two models:

```
lasso.1se.model <- lm(log.per.cap.income ~ log.per.cap.land.area + pop.18_34
                      + log.per.cap.doctors + pct.bach.deg + pct.below.pov + pct.unemp,
                      data = cdilogred.cont)

anova(all.subsets.01.final.model,lasso.1se.model)

## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.per.cap.land.area + pop.18_34 + pop.65_plus +
##   log.per.cap.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##   pct.unemp
## Model 2: log.per.cap.income ~ log.per.cap.land.area + pop.18_34 + log.per.cap.doctors +
##   pct.bach.deg + pct.below.pov + pct.unemp
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      431 2.9031
## 2      433 3.0592 -2    -0.1561 11.588 1.254e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC(all.subsets.01.final.model,lasso.1se.model)

##                df      AIC
## all.subsets.01.final.model 10 -940.5743
## lasso.1se.model             8 -921.5292

BIC(all.subsets.01.final.model,lasso.1se.model)

##                df      BIC
## all.subsets.01.final.model 10 -899.7065
## lasso.1se.model             8 -888.8350

summary(lasso.1se.model)

##
## Call:
## lm(formula = log.per.cap.income ~ log.per.cap.land.area + pop.18_34 +
##   log.per.cap.doctors + pct.bach.deg + pct.below.pov + pct.unemp,
##   data = cdilogred.cont)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37735 -0.04922 -0.00842  0.05191  0.27781
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.2333158   0.0850955 120.257 < 2e-16 ***
## log.per.cap.land.area -0.0465305   0.0040070 -11.612 < 2e-16 ***
```

```
## pop.18_34          -0.0142544  0.0011372 -12.535 < 2e-16 ***
## log.per.cap.doctors  0.0769568  0.0097474   7.895 2.41e-14 ***
## pct.bach.deg        0.0134659  0.0008268  16.288 < 2e-16 ***
## pct.below.pov      -0.0216248  0.0011146 -19.402 < 2e-16 ***
## pct.unemp           0.0144028  0.0021723   6.630 1.00e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08405 on 433 degrees of freedom
## Multiple R-squared:  0.8369, Adjusted R-squared:  0.8347
## F-statistic: 370.4 on 6 and 433 DF,  p-value: < 2.2e-16
```

```
summary(all.subsets.01.final.model)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ ., data = tmp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35918 -0.04920 -0.00401  0.04835  0.27154
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.7858923   0.1421206   75.893 < 2e-16 ***
## log.per.cap.land.area -0.0425129   0.0040176  -10.582 < 2e-16 ***
## pop.18_34      -0.0161092   0.0013063  -12.331 < 2e-16 ***
## pop.65_plus    -0.0040409   0.0013844   -2.919  0.0037 **
## log.per.cap.doctors  0.0872129   0.0103491    8.427 5.35e-16 ***
## pct.hs.grad    -0.0045224   0.0010877   -4.158 3.88e-05 ***
## pct.bach.deg     0.0148682   0.0009586   15.510 < 2e-16 ***
## pct.below.pov   -0.0251831   0.0013295  -18.942 < 2e-16 ***
## pct.unemp       0.0129805   0.0021976    5.907 7.09e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08207 on 431 degrees of freedom
## Multiple R-squared:  0.8453, Adjusted R-squared:  0.8424
## F-statistic: 294.3 on 8 and 431 DF,  p-value: < 2.2e-16
```

Though F-test prefers the model that is 1 SE based on LASSO, both AIC and BIC prefers the original model based on the all-subsets regression. So for better interpretation and the integrality of the model, we will settle on the very first model we got based on the all-subsets regression.