Exploring Factors of Economic Health and Social Well-Being Among Populous US Counties

Alana Willis
Department of Statistics and Data Science
Carnegie Mellon University
alanaw@andrew.cmu.edu

October 18, 2021

Abstract

With over 3,000 counties spread across 51 states, including the District of Columbia, the United States has a very diverse population with a combination of social and economic factors that affect it. We are interested in finding the best model to predict average income per capita and the relationships between these factors based on 14 different characteristics. The data for this study has been collected for 440 of the most populous counties in the United States from the years 1990 and 1992 and is provided by the Geospatial and Statistical Data Center located at the University of Virginia. We performed general EDA and variable transformations in order to use regression methods such as Partial F-Tests, AIC/BIC, residual plots, VIF's, all subsets regression, stepwise regression, and lasso regression with cross validation. We discovered that many of the variables share associations and there are a few collinearity issues. Across the United States for every 1% increase in per-capita crime, there is a 0.04% increase in per-capita income and baseline per-capita income differs significantly by region, except for the in the West. We were able to fit a model predicting per capita income with high predictive power while accounting for interpretability using both main effects and interactions. Overall, the findings are descriptive in describing how a county's economic health and social well-being can affect average income per capita.

1 Introduction

A county is described as a governmental unit in the United States that is bigger than a city but smaller than a state. With over 3,000 counties spread across 51 states, including the District of Columbia, the United States has a very diverse population with a combination of social and economics factors that affect it. Social scientists are interested in looking at historical data to learn how average income per capita is associated with a county's economic health and social well-being based on factors such as population, land area, crime, education, and many

other determinants. In addition to answering the main question posed above, we hope to also address the following concerns:

1. Which variables seem to be related to which other variables in the data? Are these relationships expected and can these findings be explained in terms of the meanings of the variables?
2. If we ignore all other variables, what is the relationship between per-capita income and crime rate and how is it affected by different regions of the country?
3. Should we be worried about either the missing states or the missing counties? Why or why not?

2 Data

The data for this study has been collected for 440 of the most populous counties in the United States from the years 1990 and 1992. It is provided by the Geospatial and Statistical Data Center located at the University of Virginia. Each line of the data set has an identification number attached to a county and state abbreviation followed by information on 14 different characteristics, all described in table 1. Counties with missing data were deleted from the data set prior to this study.

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

Table 1: Variable definitions for CDI data from Kutner et al. (2005).

To create an ideal dataset for future modeling a few variables were immediately removed. ID was removed due to its replication of the row numbers. State and county were removed because they both contained too many unique values to be useful categorically, leaving region as the sole categorical variable (Appendix 1). Majority of the counties are located in the Southern region with 152 and the least in the Western region with 77 counties. The full distribution of region can be seen in table 2.

In addition to the variables in table 1, we created a new variable called per-capita crime from total crime and total population. Both per-capita income and per-capita crime are measured in thousands. Per-capita income has a mean of approximately $18,560 and a standard deviation of $4,060. Per-capita crime has a mean of approximately 57,290 and a standard deviation of 27,330. A table of summary statistics for all continuous variables can be found under Appendix 1.

Figure 1 displays distributions for all continuous variables in the dataset. A substantial amount of the variables have a prominent right skew which will be discussed later in the study. Figure 2 displays the associations between all of the continuous variables and the primary response variable, per-capita income. Due to the skewed distributions of some of the variables there are plots showing non linear associations which will also be assessed later in the study.

region	Freq.
NC	108
NE	103
S	152
W	77

Table 2: Frequency distribution of Region

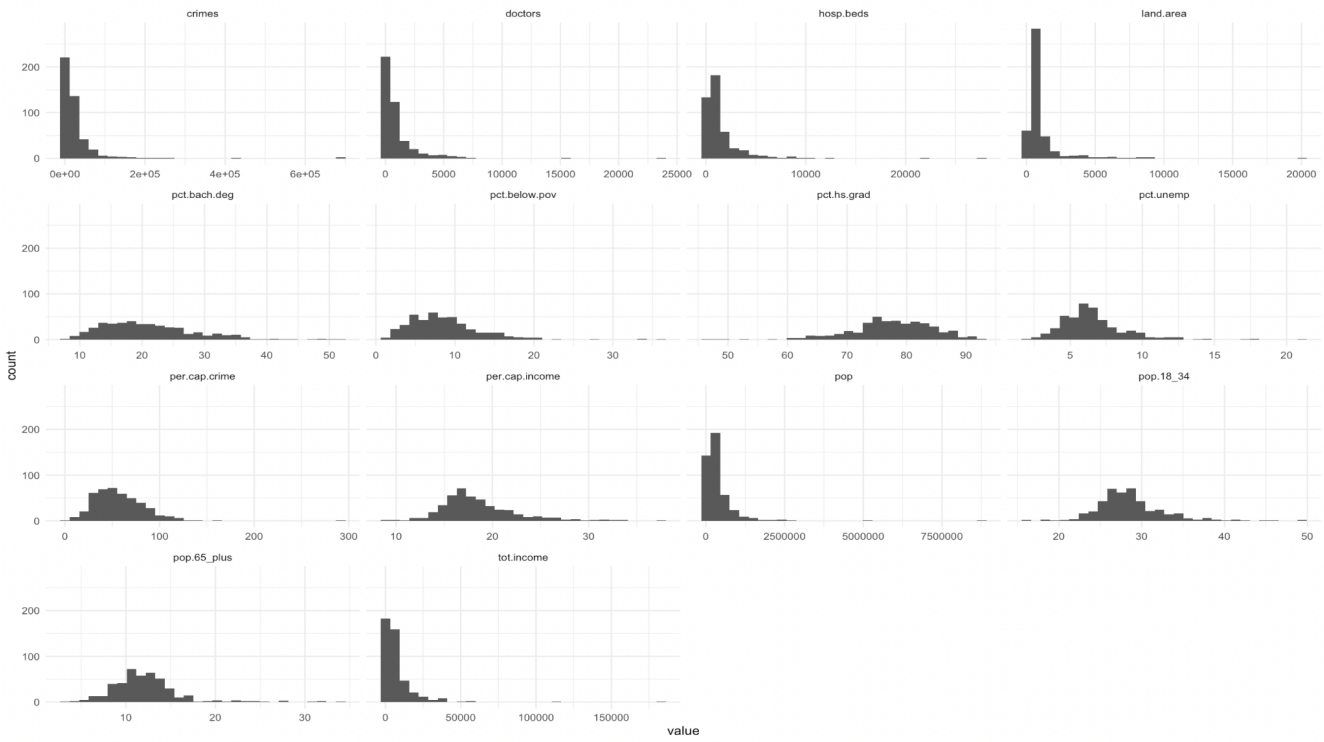


Figure 1: Distributions of continuous variables

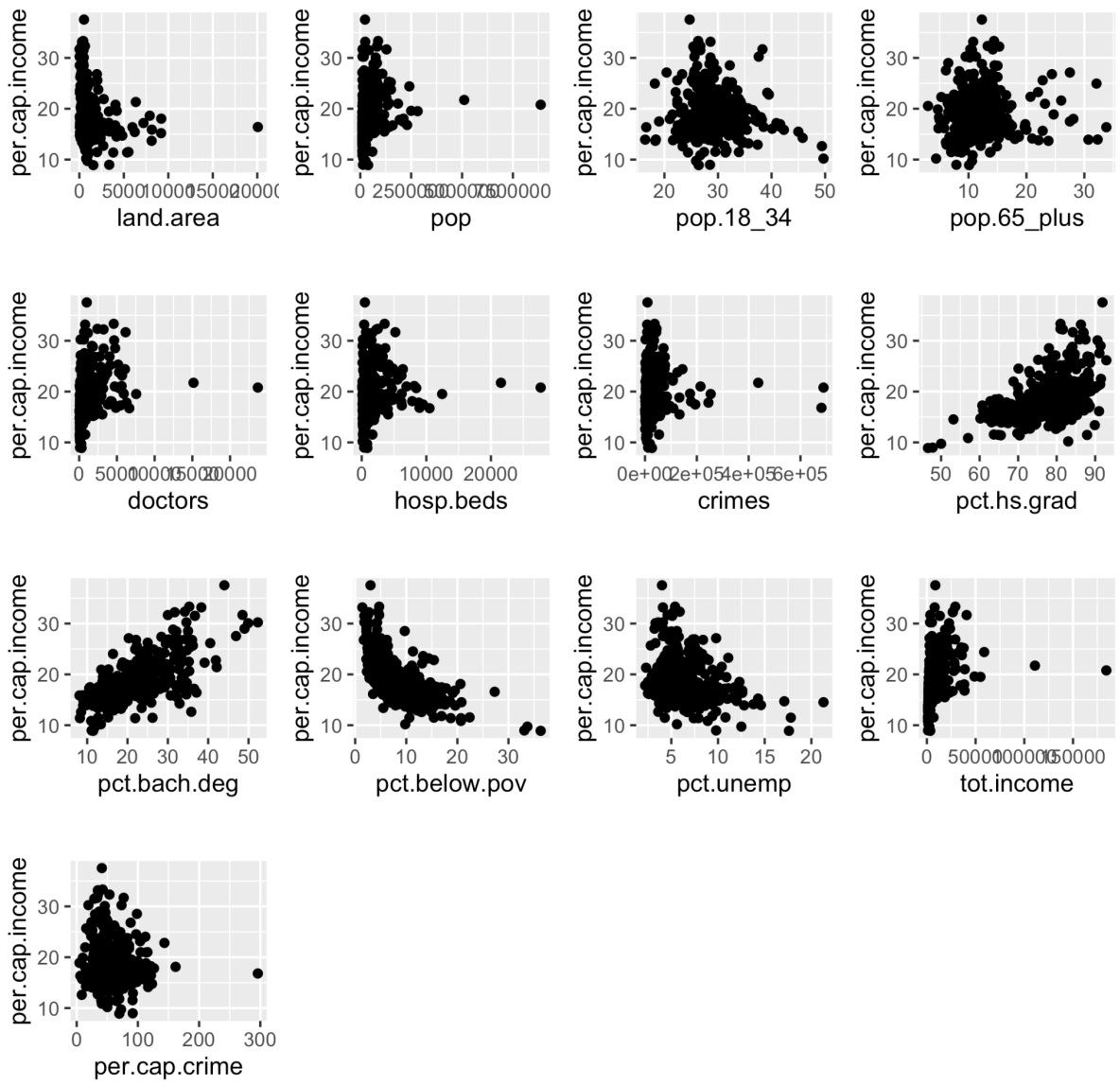


Figure 2: Relationship between per.cap.income and continuous variables

3 Methods

All statistical modeling and visualizations were made using the R language and environment for statistical computing (R Core Team, 2017).

Before beginning to answer any of the concerns of the social scientists, we first decided to transform variables that were skewed, keeping interpretability in mind, in hopes of creating better distributions and providing accurate model results. After transformations, we wanted to characterize the associations between variables. To do this we created a heat map correlation matrix to check for high correlations among the predictors and raise awareness to any problems of multicollinearity.

Next we considered two regression models. For the first model, predicting per-capita income based on crime rate and region, we assessed for a model that was best in terms of interpretability and predictive power. We looked at differences between additive models and interaction models using Partial F-Tests, AIC/BIC, and checking model assumptions with residual plots. Details on the model selection can be found under Appendix 3. For the second model, we wanted to find the best combination of variables to predict per-capita income. To do this we used a collection of variable selection techniques such as checking VIF's for collinearity problems, all subsets regression, stepwise regression, and lasso regression with cross validation. Details on the model selection can be found under Appendix 3.

To complete the analysis we considered the possible implication of the missing states and counties in the dataset.

4 Results

Based on the histograms (Figure 1) and scatter plots (Figure 2) we decided to transform the variables crimes, tot.income, land.area, doctors, hosp.beds, per.cap.crime, and per.cap.income with log transformations. Other variables like pct.unemp, pce.below.pov, and pop could have also used transformations, however to adhere to the integrity of interpretability we only considered those that improved with a log transformation. The distribution of the newly transformed variables can be seen in figure 3 and the scatter plots in figure 4. The predictors now show clear linear associations and can safely be used in modeling with the main response variable, log.per.cap.income. The log transformed variables will be used throughout the duration of the modeling.

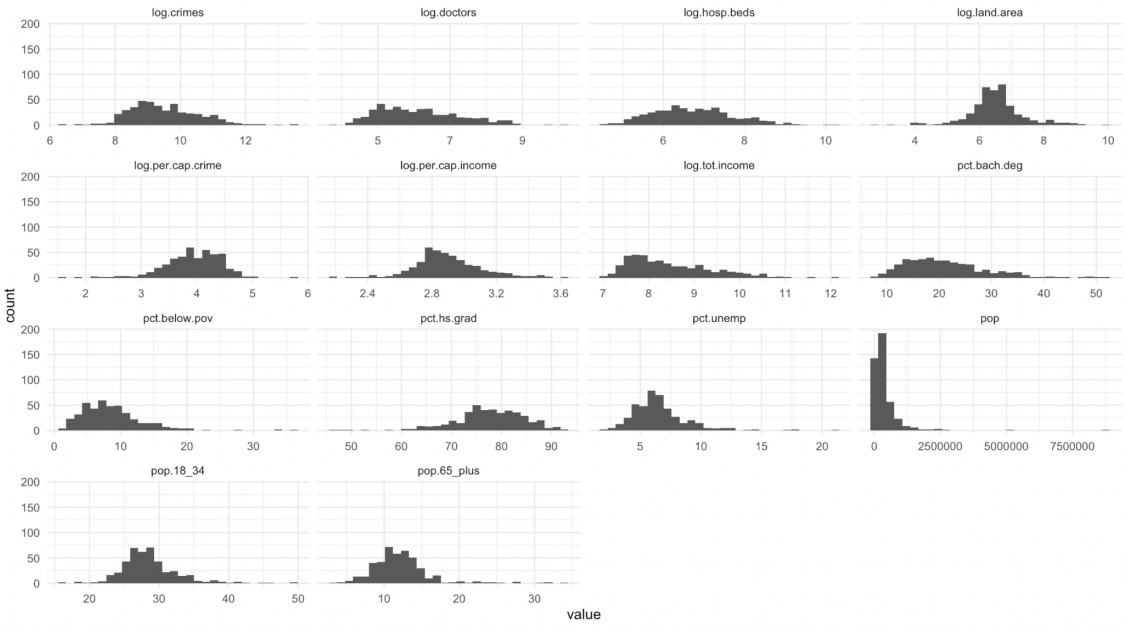


Figure 3: Distributions of transformed continuous variables

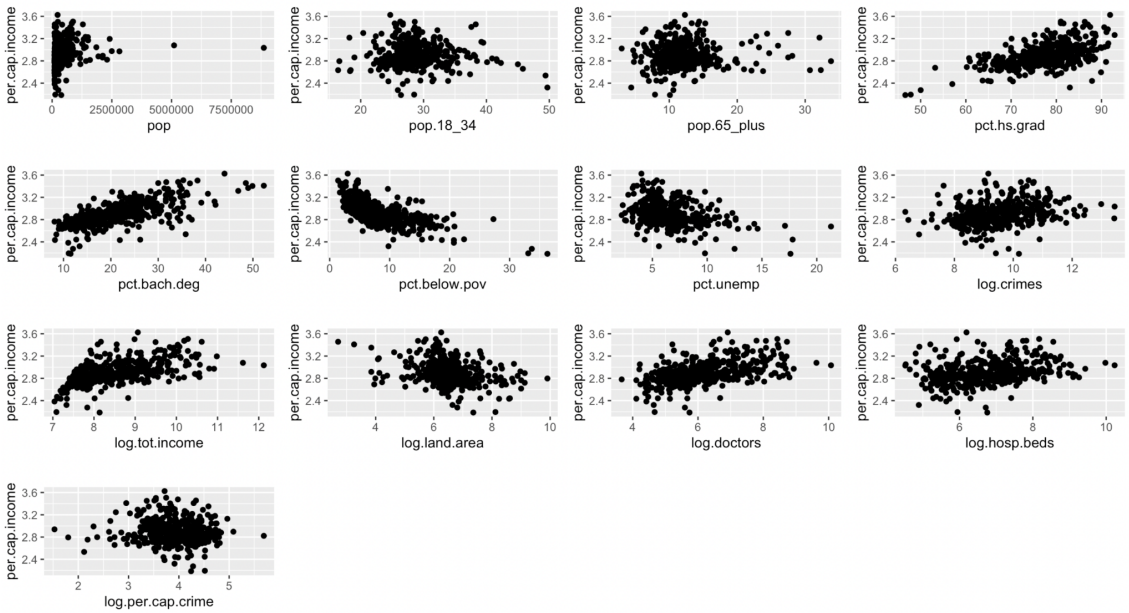


Figure 4: Relationship between log.per.cap.income and transformed continuous variables

The heat map correlation matrix, seen in figure 5, shows some interesting associations between variables. Many of the associations are to be expected and can be explained in terms of the variable descriptions. The log.doctors variable has some strong positive associations with log.crime, log.tot.income, and log.hosp.beds. Doctors probably have a high income thus influencing the total income, the amount of hospital beds is related to the number of doctors there are to attend to them, and doctors are treating criminals and their victims when needed. The log.hosp.beds variable shares the same strong positive associations with log.crimes, log.tot.income, and log.doctors. The variable log.tot.income has other strong positive correlations with pop and log.crimes.

In terms of collinearity issues, there is strong association between log.crimes and log.per.cap.crime. This is expected due to the fact that crimes was used to calculate per.cap.crime. There is also a strong positive correlation between pct.bach.deg and pct.hs.grad since those who have graduated highschool are included in those who have bachelor degrees. We assume these collinearities will be addressed in the variable selection methods. The main response variable log.per.cap.income has some moderate associations which means there is hope that good models can be made with it. The strongest of these being pct.hs.grad, pct.bach.deg, and pct.below.pov.

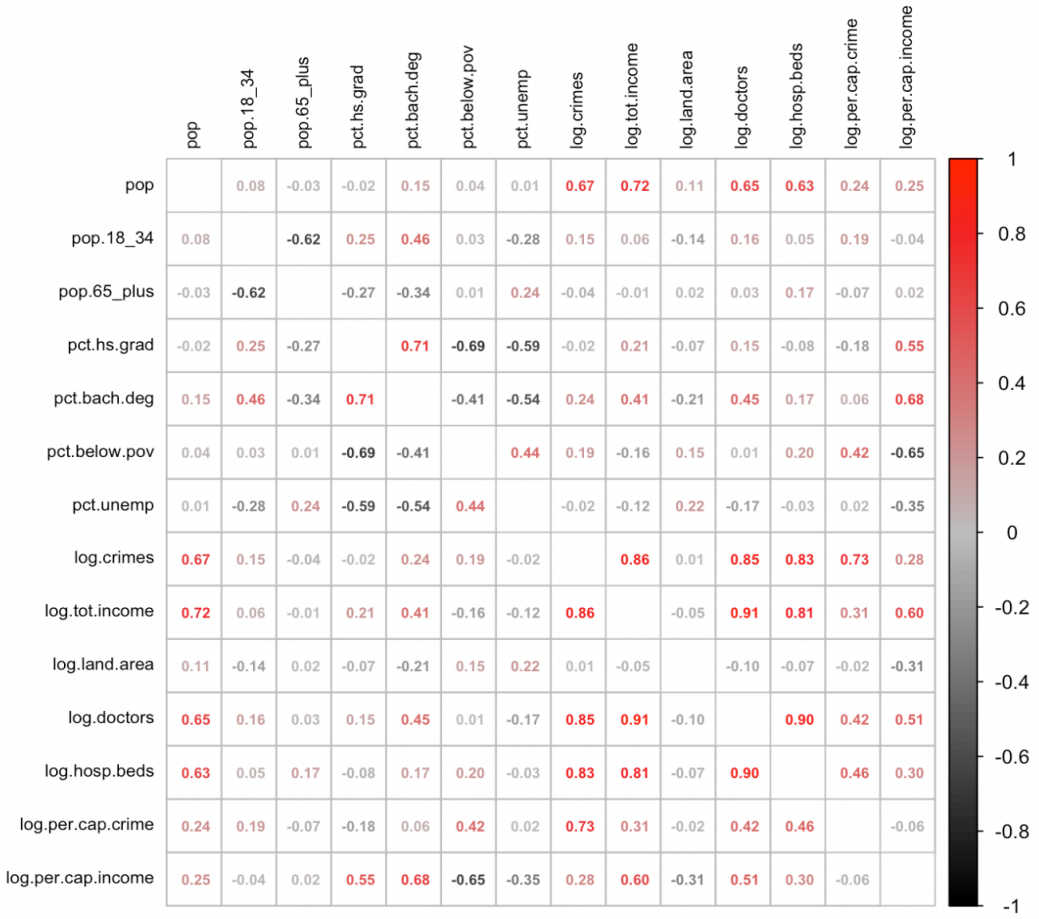


Figure 5: Heat map correlation matrix of variables.

There are two main factors we wanted to consider when trying to find a model to predict per.cap.income based on crime and region. One, is the additive or the interaction model preferred and two, is per.cap.crime a better predictor than crimes. As a reminder, log transformations were done previously to crimes, per.cap.crime, and per.cap.income and will be used in the models. We fit a total of 6 linear models seen in table 3; 3 with log.crimes and 3 with log.per.cap.crime including each null model. The residual plots for all models are acceptable meaning the models meet assumption requirements and we can continue on to using partial f tests to compare. All regression outputs and residual plots can be found under Appendix 3.

Name	**Model**
null	log.per.cap.income ~ log.crimes
null.1	log.per.cap.income ~ log.per.cap.crime
lm.1	log.per.cap.income ~ log.crimes + region

lm.2	log.per.cap.income ~ log.per.cap.crime + region
lm.3	log.per.cap.income ~ log.crimes*region
lm.4	log.per.cap.income ~ log.per.cap.crime*region

Table 3: Models predicting log.per.cap.income.

We performed two f tests; one for the models with log.crimes and one for the models with log.per.cap.crime. Both tests preferred the additive model over the interaction model. These outputs can be found under Appendix 3. To compare the two additive models we used AIC and BIC estimations. Both AIC and BIC estimations prefer the model using log.crimes as shown in table 4. Since interpretability is one of the most important things to keep in mind, we decided to pick the model using log.per.cap.crime over the model using log.crimes despite the results of the AIC and BIC estimates. This way both the response and the predictor are in the same units and have the same transformation.

	df	AIC	BIC
lm.1	6	-227.4746	-202.9539
lm.2	6	-172.1347	-147.6140

Table 4: AIC and BIC estimations for additive models

As stated before the residual plots satisfy modeling assumptions coupled with the enhanced interpretability we are confident on choosing lm.2 (table 3) as the final and best model. The regression output is shown in table 5. Across the United States for every 1% increase in per-capita crime, there is an 0.04% increase in per-capita income. Baseline per-capita incomes differ by region with North Central (NC) being $15,490, North East (NE) being $17,290, South (S) being $14,440, and West (W) being $15,180. All regions, except for the W, have baseline incomes that are significantly different from the NC baseline. According to the model, income level varies by region but not in the way it varies with crime.

Variable	**Coefficient Estimate**	**P-Value**
Intercept	2.74	< 2e-16 ***
log.per.cap.crime	0.04	0.05 *
regionNE	0.11	3.99e-5 ***

regionS	-0.07	0.005 **
regionW	-0.02	0.42

Table 4: Regression output for model lm.2 (final model).

To begin finding the best model to predict log.per.cap.income based on all predictors, we first decided to remove pop, log.tot.income, and log.crimes from the dataset. Pop and log.tot.income were used in the calculation of log.per.cap.income thus removing them was necessary. We used the VIF calculations (Appendix 3) to check our logic and see if any other variables should be removed due to collinearity. Both log.per.cap.crime and log.crimes had high VIF's, as log.crimes was used in the calculation of log.per.cap.crime, but log.per.cap.crime was significantly lower and it is in the same units as the response, thus we kept it over log.crimes. None of the variable selection methods are particularly good at working with categorical variables, so region was left out of the original variable selection but will be added back later.

We chose all subsets regression as the primary variable selection method because it considers all possible combinations of models. Using BIC generated by the all subsets regression, we determined the best model was model 7 with the lowest BIC of -772.0715. Table 5 shows the variables in model 7. The complete all subsets output can be found in the technical appendix under Appendix 3. To check the validity of the model we refit it in a standard linear regression to show the residual plots. The residual plots show that the model is valid as it meets modeling assumptions. The regression output and residual plots can be found under Appendix 3. Lastly, to confirm this base model we looked at the marginal model plots to check for any missing interactions and transformations. The plots, shown in the technical appendix (Appendix 3), also upholds the validity of our model.

Model 7
pop.18_34
pct.hs.grad
pct.bach.deg
pct.below.pov
pct.unemp
log.land.area
log.doctors

Table 5: Variables in the best model selected by all subsets regression.

Since we have a good starting model, we wanted to introduce region. To do so we fit a model with region interacting with all of the variables in model 7. The regression output and residual plots can be found under Appendix 3. From the output we dropped log.land.area:region, pop.18_34:region, and log.doctors:region because none of the interactions were significant. The AIC and BIC estimates (Appendix 3) that were calculated for each model showed that both prefer the model with interactions, thus that is the final model.

Variable	**Coefficient Estimate**	**P-Value**
Intercept	3.22	< 2e-16 ***
pop.18_34	-0.02	< 2e-16 ***
pct.hs.grad	-0.003	0.43
pct.bach.deg	0.01	5.24e-11 ***
pct.below.pov	-0.02	7.30e-12 ***
pct.unemp	0.02	0.0002 ***
log.land.area	-0.03	3.50e-10 ***
log.doctors	0.06	< 2e-16 ***
regionNE	0.22	0.47
regionS	-0.06	0.82
regionW	1.63	6.86e-06 ***
pct.hs.grad:regionNE	-0.004	0.35
pct.hs.grad:regionS	0.002	0.57
pct.hs.grad:regionW	-0.02	8.85e-06 ***
pct.bach.deg:regionNE	0.006	0.02 *
pct.bach.deg:regionS	-0.001	0.58
pct.bach.deg:regionW	0.006	0.02 *
pct.below.pov:regionNE	-0.002	0.60
pct.below.pov:regionS	0.007	0.05 .
pct.below.pov:regionW	-0.015	0.003 **

pct.unemp:regionNE	-0.008	0.27
pct.unemp:regionS	-0.02	0.0001 ***
pct.unemp:regionW	-0.02	0.003 **

Table 6: Regression output for final model from all subsets with interaction with region.

Next we repeated the same process with stepwise regression, starting without region and adding it in later. The BIC model is the exact same as the all subsets model which helps us confirm that the all subsets model is valid. The AIC model added pct.65_plus however, we know when we add the region interaction the number of terms increases thus BIC may be a better indicator since it is primarily used for larger models. The stepwise BIC model with the region interaction produces the exact same model as well, again helping us confirm our final model. As a last check, we performed lasso regression with cross validation and the 1se lambda. This variable selection technique also found the same base model as the all subsets and would likely find the same interaction model with region. The full outputs for both the stepwise regression and the lasso regression can be seen in the technical appendix under Appendix 3.

Now that we have confirmed a final model we can begin to interpret the outputs in table 6. For the sake of focusing on the most important factors, we will only interpret the significant variables. For every 1 percentage point increase in the percent of the population between 18 and 34 , per capita income decreases by 1.98%. This is most likely due to the fact that younger people make less money. For every 1 percentage point increase in the percent of the population that has a bachelor's degree, per capita income increases by 1%. This can most likely be attributed to the fact that more education can lead to higher income. For every 1 percentage point increase in the percent of the population that is below the poverty rate, per capita income decreases by 1.98%. For every 1 percentage point increase in the percent of the population that is unemployed, per capita income increases by 2.02%. This is an interesting finding as we may have expected the opposite. A possible explanation is that government assistance for people in poverty is resulting in the increase. For every 1% increase in land area, there is a 0.03% decrease in per-capita income. For every 1% increase in the amount of doctors, there is a 0.06% increase in per-capita income. This is attributed to the high income of doctors compared to the general population. The baseline per capita income for the West is $5,100. The West is also the only region with a significant difference from the North Central region in each interaction. This could be due to the low sample size for the Western region, thus each observation is weighted more heavily. The North East region has a significant difference in income from the North Central region in terms of the percentage of those with bachelor's degrees. The Southern region has a significant difference in income from the North Central region in terms of the percentage of people that are unemployed. Using a larger confidence interval the same applies to those below the poverty line. There are probably a lot of lurking variables that are influencing these results thus we can not directly characterize these differences.

A general rule of thumb when assessing whether a sample is large enough to be generalized to the population is if the sample size is at least 10% of the population. In our case, there are 373 counties present in the dataset out of a total of 3,000 counties in the US. Following the general rule, we can say that our findings are generalizable and we should not worry greatly about the missing states and counties. The only concern is that we do not know if the sample was randomly selected or if all 3,000 counties were sampled but these 373 were the only ones that remained. Also there are 48 out of 51 (including Washington D.C. as a state) represented which is way more than 10%. Lastly, in all of our modeling we used region as a location classifier instead of county or state. All the regions are represented in the dataset, thus missing data is not a big concern.

5 Discussion

In this study, we were examining four main concerns related to average income per capita and its association with a county's economic health and social well-being. First, we wanted to characterize the relationships of all the variables in the dataset independently and explain them in terms of the variable descriptions. We determined that many of the associations are to be expected and can be explained in terms of the variable descriptions. Due to the similarity of the variables and possible collinearities doctors and hospital beds share some of the same strong correlations. Total income and per-capita crime both had high correlations to some variables due to them being used in the calculations of other variables. The education variables, pct.bach.deg and pct.hs.grad, had high correlations since those who have graduated highschool are included in those who have bachelor degrees. Per capita income made a great response variable because of its moderate associations with the majority of the variables.

Second, we wanted to see the relationship between per-capita income and crime rate and how it is affected by different regions of the country. Instead of using the raw crime rate variable we created the per capita crime variable so that both the response and the predictor were measured in the same units. We found that across the United States for every 1% increase in per-capita crime, there is a 0.04% increase in per-capita income and baseline per-capita income differs significantly by region, except for the West. Overall, income level varies by region but not in the way it varies with crime.

Third, we decided to not worry too much about the missingness of counties and states in the dataset. More than 10% of states and counties are represented meaning the sample can be generalizable to the population. We also used region as our main variable to describe location and all regions were represented in the dataset.

Lastly, the most important concern we addressed was to find the best combination of variables to predict per-capita income. Using all subsets as our primary variable selection method and checking with other selection methods, we were able to fit a model with high predictive power while accounting for interpretability. The final model determined that possibly due to the younger population making less money, for every 1 percentage point increase in the percent of the population between 18 and 34, per capita income decreases by 1.98%. It also

found that for every 1 percentage point increase in the percent of the population that has a bachelor's degree, per capita income increases by 1%. This can most likely be attributed to the fact that more education usually leads to higher income. A 1 unit percent increase of the population that is below the poverty rate decreases per capita income by 1.98%. For every 1 percentage point increase in the percent of the population that is unemployed, per capita income increases by 2.02%. This is an interesting finding as we may have expected the opposite. A possible explanation is that government assistance for people in poverty is actually resulting in an increase in average income. For every 1% increase in land area, there is a 0.03% decrease in per-capita income. For every 1% increase in the amount of doctors, there is a 0.06% increase in per-capita income. This is most likely attributed to the high income of doctors compared to the general population. Possibly due to the low sample size, the West is the only region with a significant difference from the North Central region in each interaction. The North East region has a significant difference in income from the North Central region in terms of the percentage of those with bachelor's degrees. The Southern region has a significant difference in income from the North Central region in terms of the percentage of people that are unemployed. Generally, there are probably lurking variables that are influencing these results contributing to these differences.

We wanted to be more accurate in our interpretations of the region interactions however, we would like the social scientists to be able to understand the findings. We left them at broad generalizations that are still useful and hope in the future we can explore them more thoroughly. Had time permitted we would have also liked to explore interactions between the continuous variables. This may have led to an even more accurate model with higher predictive power. We also would have considered a way to work with the state variable as our main categorical location variable instead of region. Using state would provide even more distinct results on how per capita income is affected given the different factors measured in the data.

Overall, all findings are descriptive and fairly accurate in describing how a county’s economic health and social well-being can affect average income per capita.

References

Ford. C., (2017), *"Interpreting Log Transformations in a Linear Model"*, University of Virginia. URL https://data.library.virginia.edu/interpreting-log-transformations-in-a-linear-model/.

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

R Core Team (2017), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RStudio Team (2020), *R Studio: 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.

Appendices: CDI Analyses

Alana Willis

10/18/2021

Contents

Appendix 1. Initial Data/Library Imports & Exploration

Loaded packages needed for the analyses along with the initial data set.

```
library(gtsummary)
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(leaps)
library(MASS)
```

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

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

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

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

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

## The following object is masked from 'package:purrr':
##
##     some
```

```
library(glmnet)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-2
```

```
library(kableExtra)
```

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

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

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(grid)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

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

```
library(ggplotify)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(ggplot2)
library(arm)
```

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

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

```
## Working directory is /Users/alanawillis/Desktop/Fall_21/Applied Linear Models
```

```
##
## Attaching package: 'arm'
```

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

```
library(corrplot)
```

```
## corrplot 0.90 loaded
```

```
##
## Attaching package: 'corrplot'
```

Table 1: Summary of Continuous Variables

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
land.area	15.0	451.25	656.50	1041.41	946.75	20062.00	1549.92
pop	100043.0	139027.25	217280.50	393010.92	436064.50	8863164.00	601987.02
pop.18_34	16.4	26.20	28.10	28.57	30.02	49.70	4.19
pop.65_plus	3.0	9.88	11.75	12.17	13.62	33.80	3.99
doctors	39.0	182.75	401.00	988.00	1036.00	23677.00	1789.75
hosp.beds	92.0	390.75	755.00	1458.63	1575.75	27700.00	2289.13
crimes	563.0	6219.50	11820.50	27111.62	26279.50	688936.00	58237.51
pct.hs.grad	46.6	73.88	77.70	77.56	82.40	92.90	7.02
pct.bach.deg	8.1	15.28	19.70	21.08	25.33	52.30	7.65
pct.below.pov	1.4	5.30	7.90	8.72	10.90	36.30	4.66
pct.unemp	2.2	5.10	6.20	6.60	7.50	21.30	2.34
per.cap.income	8.9	16.12	17.76	18.56	20.27	37.54	4.06
tot.income	1141.0	2311.00	3857.00	7869.27	8654.25	184230.00	12884.32
per.cap.crime	4.6	38.10	52.43	57.29	72.60	295.99	27.33

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

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

Started by creating a data set with only the continuous variables to be used to created the table of summary statistics.

Created table of unique values of each variable. This is mostly to see which categorical variable will be best to include in the modeling later in the analyses.

Created table of summary statistics for continuous variables and a frequency table for the best categorical variable, region.

```
cdinumeric <- cdi[,-c(1,2,3,17)] %>%
  mutate(per.cap.crime=(cdi$crimes/cdi$pop)*1000, per.cap.income=per.cap.income/1000)



apply(cdinumeric,2,function(x) c(summary(x),SD=sd(x))) %>%
  as.data.frame %>%
  t() %>%
  round(digits=2) %>%
  kbl(booktabs=T,caption="Summary of Continuous Variables") %>%
  kable_minimal()
```

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

Table 2: Unique Values

	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

Table 3: Frequency of Region

region	Freq.
NC	108
NE	103
S	152
W	77

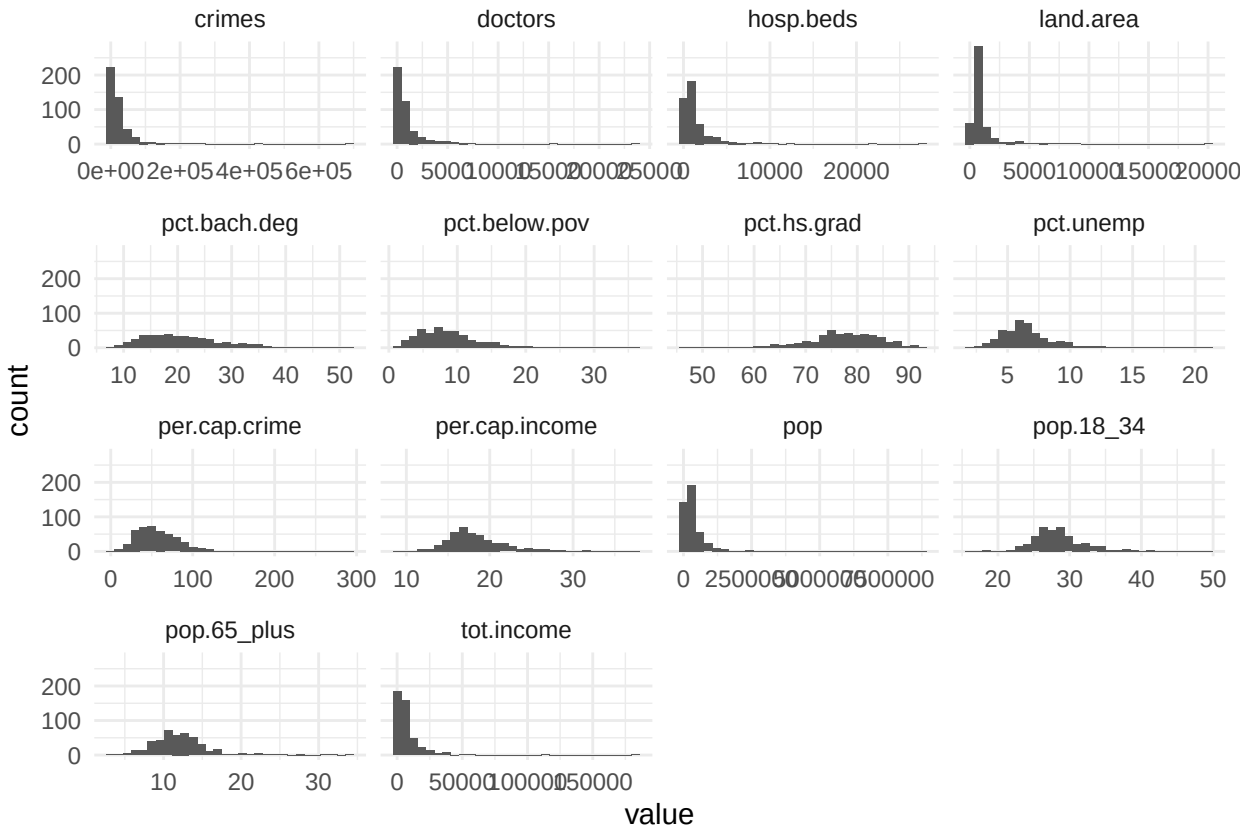
```
cdi_region <- cdi %>%
  dplyr::select(region) %>%
  group_by(region) %>%
  summarise(Freq.=n())

cdi_region %>%
  kbl(booktabs=T,caption="Frequency of Region", align = "ll") %>%
  kable_minimal()
```

Created per.cap.crime variable and mutated per.cap.income so that they are measured in the same units.

Visualized variables to see if any transformations were needed using histograms and scatter plots.

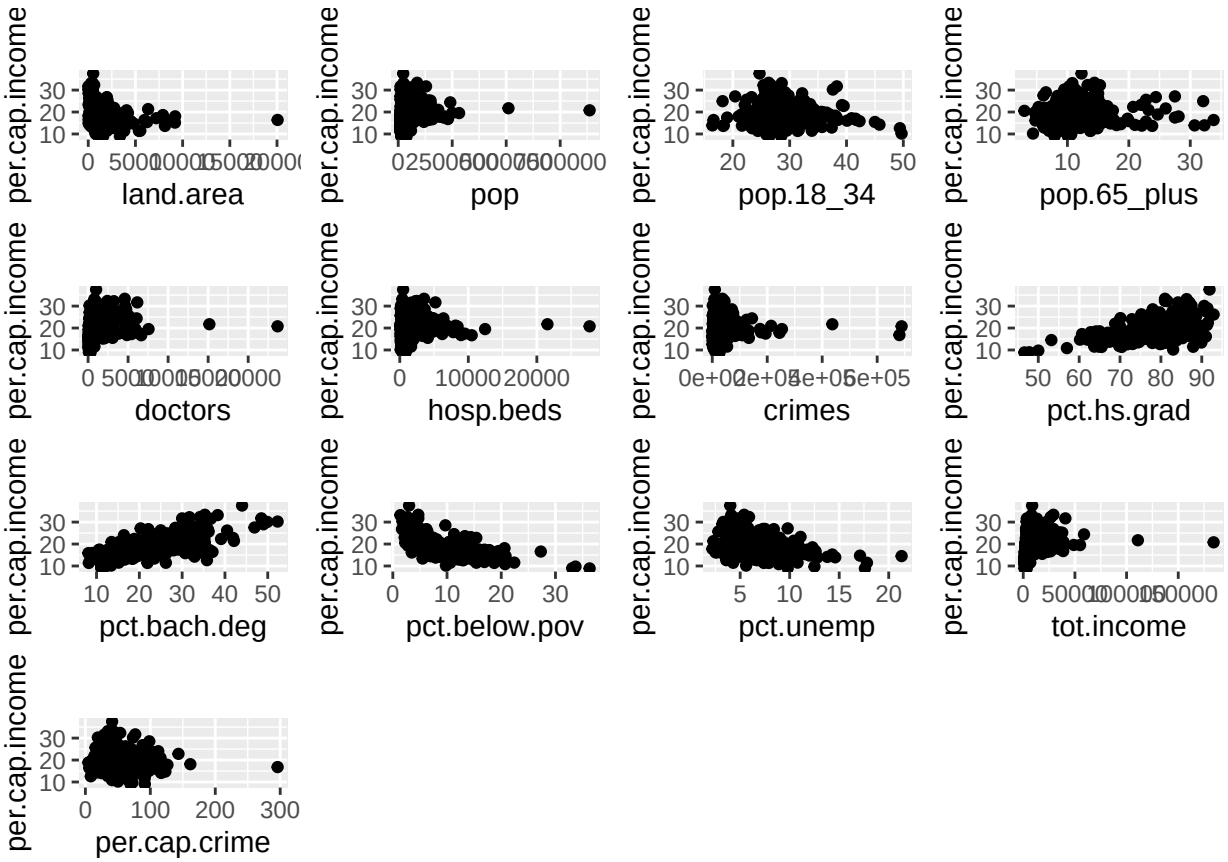
```
cdinumeric <- cdi[,-c(1,2,3,17)] %>%
  mutate(per.cap.crime=(cdi$crimes/cdi$pop)*1000, per.cap.income=per.cap.income/1000)



ggplot(gather(cdinumeric), aes(value))+
  geom_histogram(bins=30) +
```

```
  facet_wrap(~key, scales ='free_x')+
  theme_minimal()
```




```
cdinumeric.reg <-data.frame(cdinumeric,region=cdi$region)

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)
```

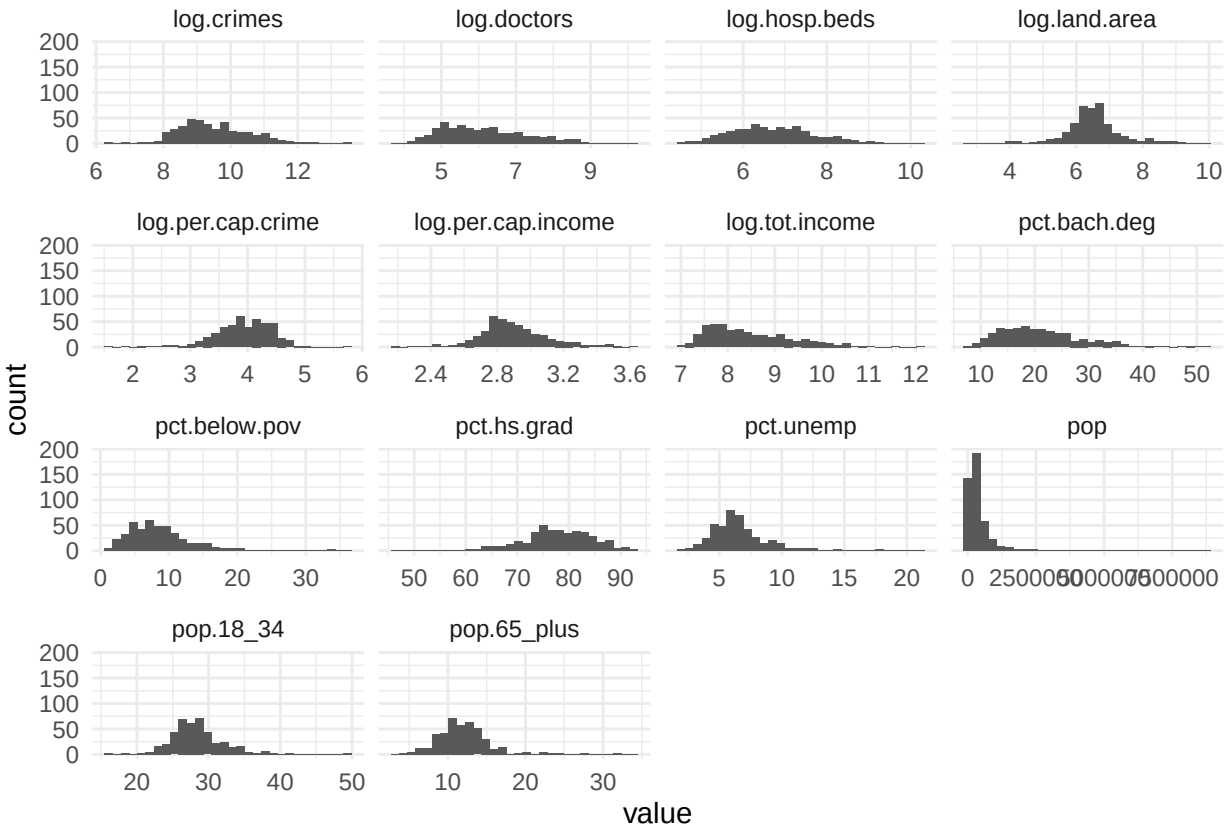
```
  }
save <- grid.arrange(grobs=scatter.builder(cdinumeric))
```



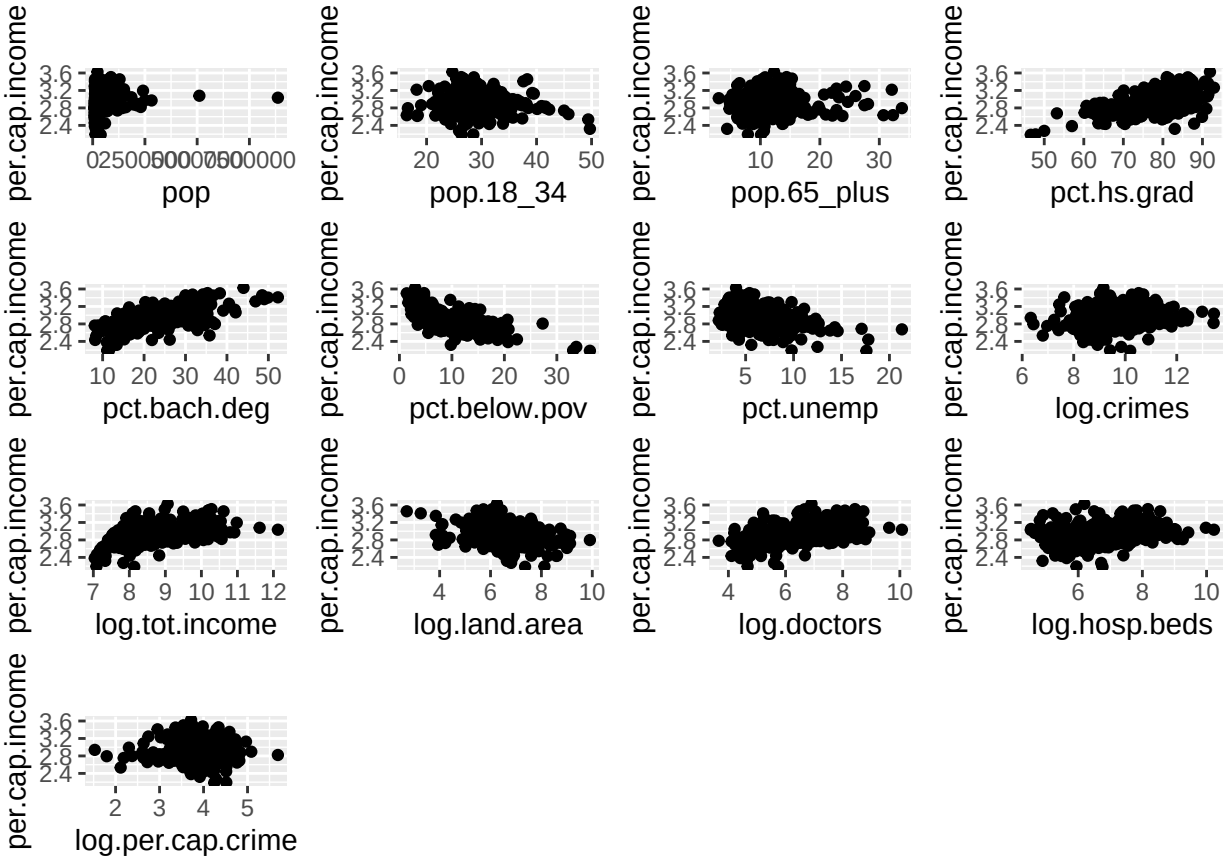
Appendix 2. Transformations and Correlations

Created a new data set with transformed variables and displayed the new histograms.

```
cdilogs <- cdinumeric %>%
  mutate(log.crimes=log(crimes),
         log.tot.income=log(tot.income),
         log.land.area=log(land.area),
         log.doctors=log(doctors),
         log.hosp.beds=log(hosp.beds),
         log.per.cap.crime=log(per.cap.crime),
         log.per.cap.income=log(per.cap.income)) %>%
  dplyr::select(-c(crimes, tot.income, land.area, doctors, hosp.beds,per.cap.crime, per.cap.income))

cdilogs.reg <- data.frame(cdilogs,region=cdi$region)

ggplot(gather(cdilogs), aes(value))+
  geom_histogram(bins=30) +
  facet_wrap(~key, scales ='free_x')+
  theme_minimal()
```

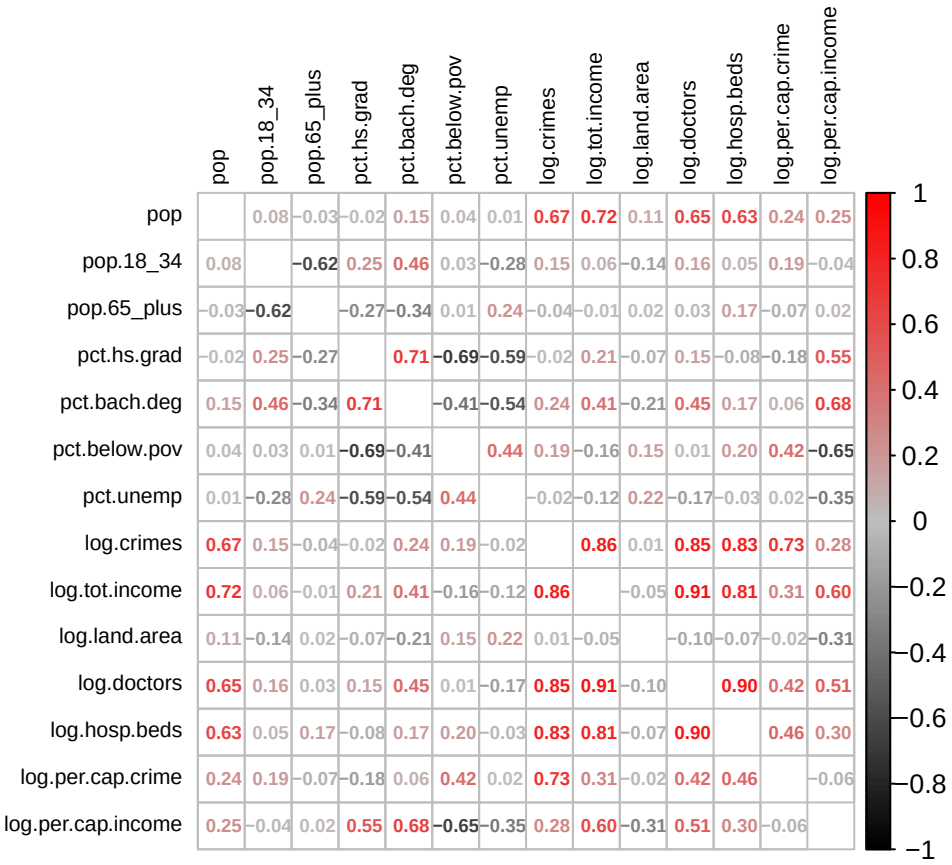



```
grid.arrange(grobs=scatter.builder(cdilogs))
```



Created a heat map correlation matrix to check for high correlations among the predictors and raise awareness to any problems of multicollinearity.

```
m <- cor(cdilogs)
msave <- corrplot(m,
                  method=c("number"),
                  diag=FALSE,
                  number.cex=0.6,
                  tl.cex = .7,
                  tl.col = "black",
                  col = colorRampPalette(c("black","gray","red"))(100))
```



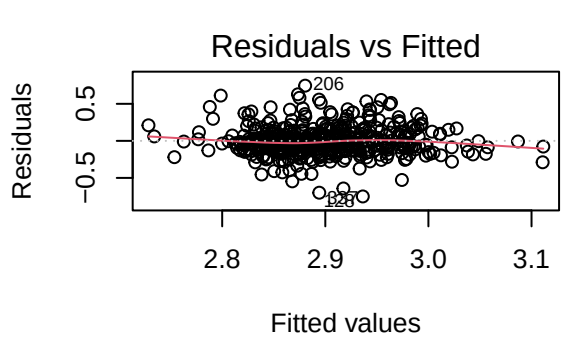
Appendix 3. Variable Selection and Regression Analysis

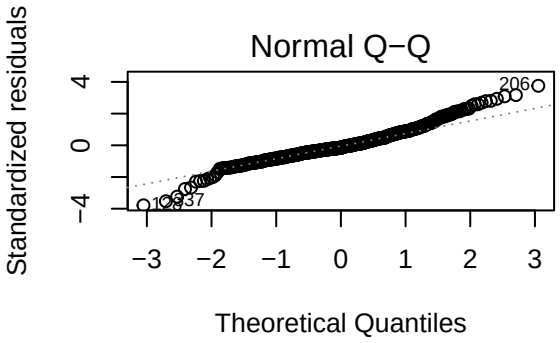
Per Capita Income ~ Per Capita Crime and Region

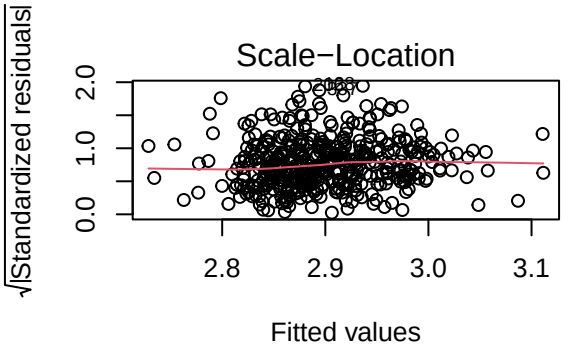
The first modeling was done to predict per-capita income based on crime rate and region.

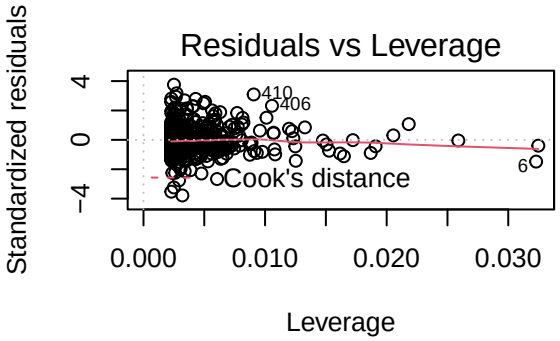
We fit a total of 6 linear models seen in table 3; 3 with log.crimes and 3 with log.per.cap.crime including each null model. The best model was assesed using Partial F-Tests, AIC/BIC, and checking model assumptions with residual plots.

```
null <- lm(log.per.cap.income ~ log.crimes, data=cdilogs.reg)
par(mfrow=c(2,2))
plot(null)
```

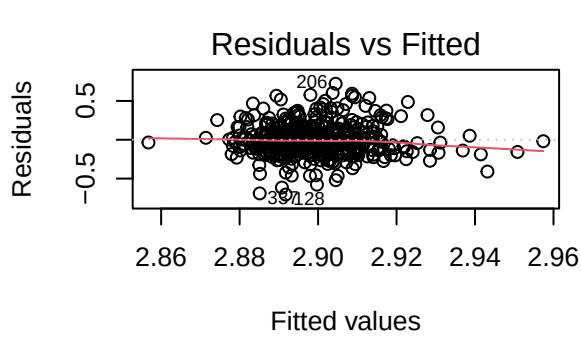


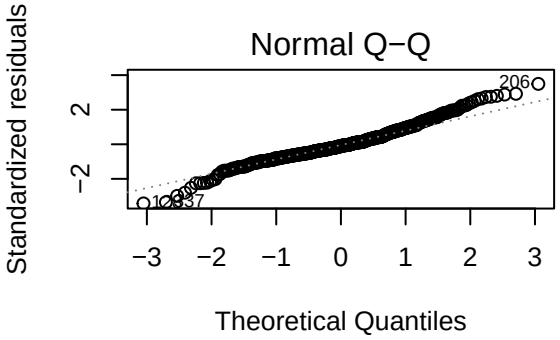


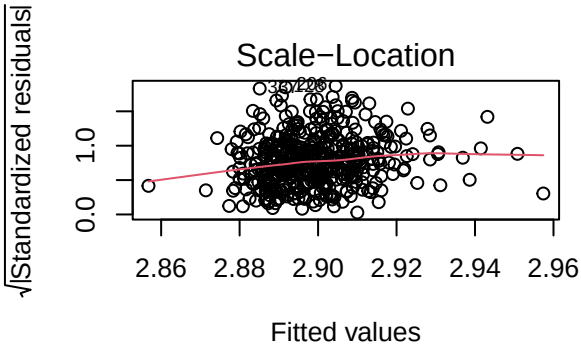


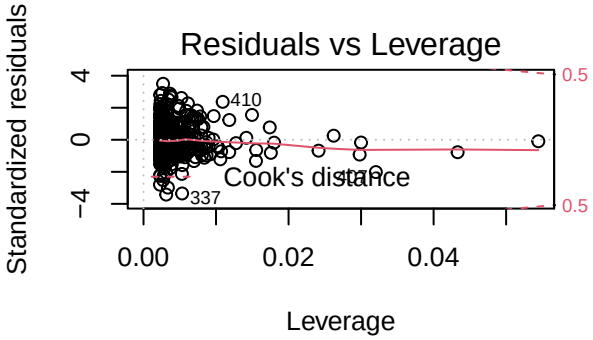



```
null.2 <- lm(log.per.cap.income ~ log.per.cap.crime, data=cdilogs.reg)
par(mfrow=c(2,2))
plot(null.2)
```





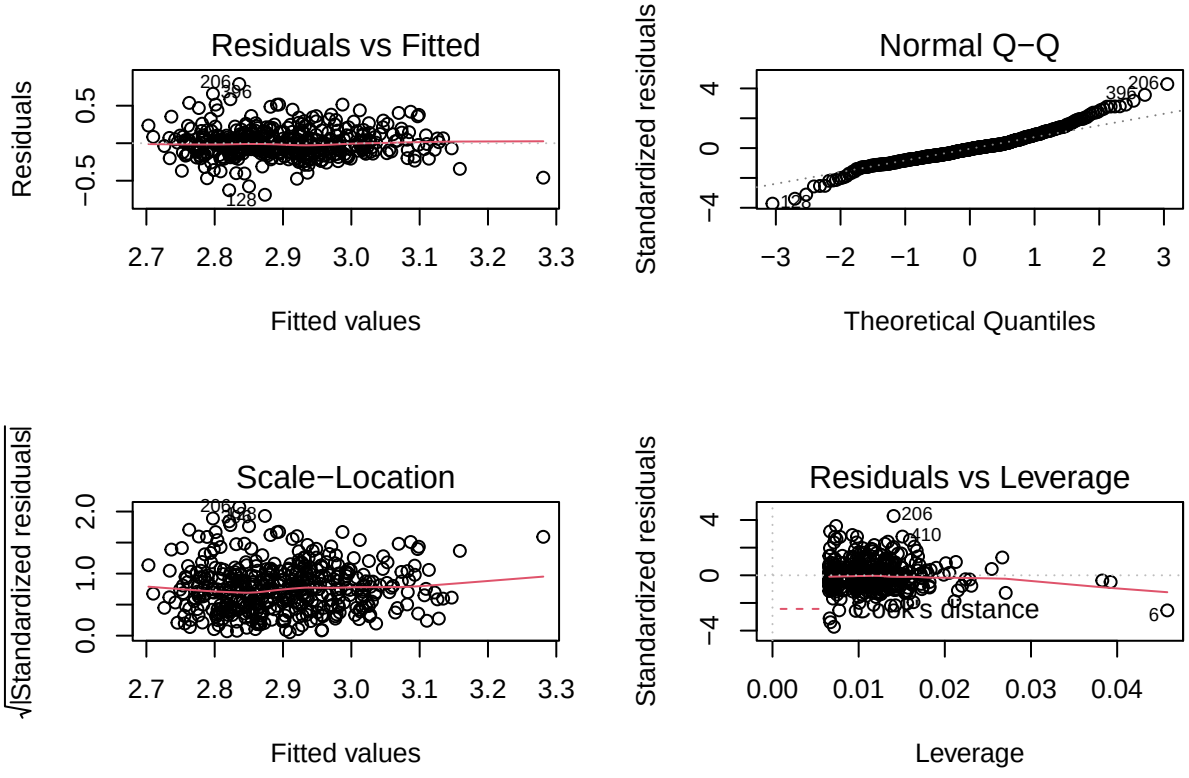




```
lm.1 <- lm(log.per.cap.income ~ log.crimes + region, data=cdilogs.reg)
summary(lm.1)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.crimes + region, data = cdilogs.reg)
##
## 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)  2.280676   0.079812  28.575  < 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
```

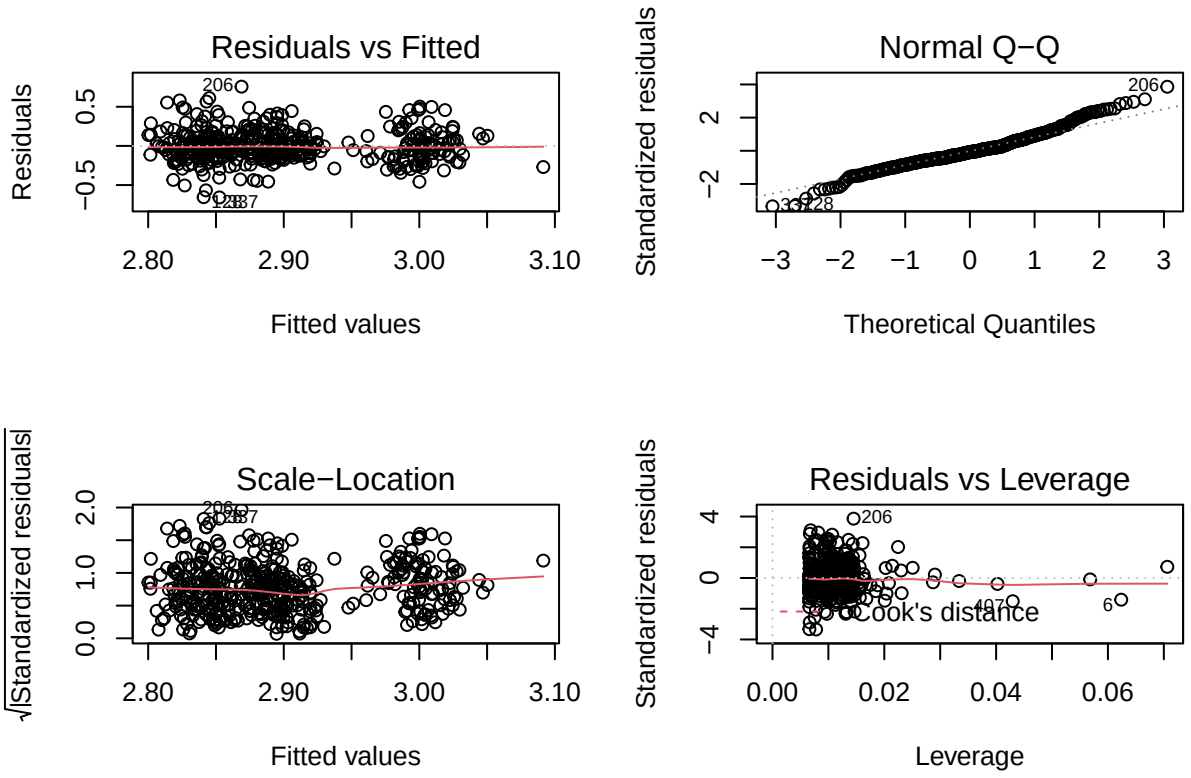
```
par(mfrow=c(2,2))
plot(lm.1)
```



```
lm.2 <- lm(log.per.cap.income ~ log.per.cap.crime + region, data=cdilogs.reg)
summary(lm.2)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.per.cap.crime + region,
##     data = cdilogs.reg)
##
## 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)        2.73543    0.08386  32.617  < 2e-16 ***
## log.per.cap.crime  0.04243    0.02148   1.975  0.04885 *
## regionNE           0.11457    0.02760   4.151 3.99e-05 ***
## regionS           -0.07456    0.02624  -2.841  0.00471 **
## regionW           -0.02426    0.03002  -0.808  0.41952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1974 on 435 degrees of freedom
## Multiple R-squared:  0.09645,    Adjusted R-squared:  0.08814
## F-statistic: 11.61 on 4 and 435 DF,  p-value: 5.776e-09
```

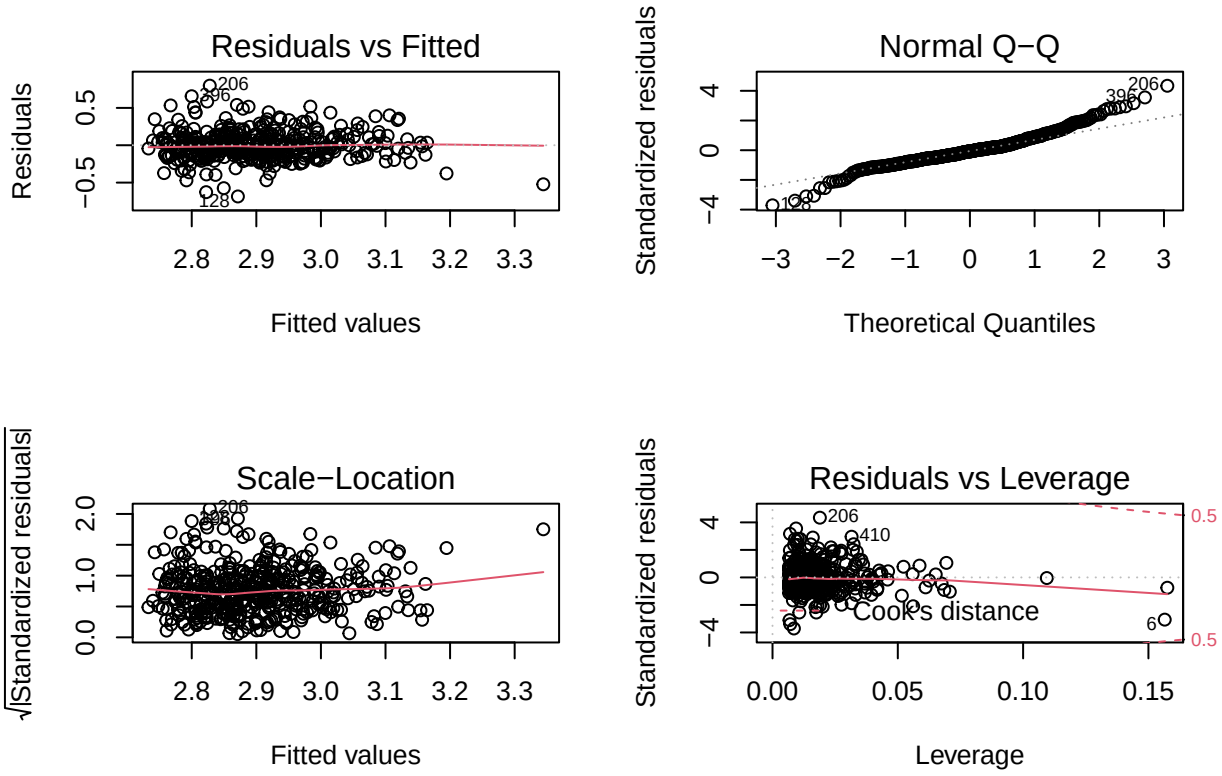
```
par(mfrow=c(2,2))
plot(lm.2)
```



```r
lm.3 <- lm(log.per.cap.income ~ log.crimes*region, data=cdilogs.reg)
summary(lm.3)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.crimes * region, data = cdilogs.reg)
##
## 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)          2.42902    0.14579  16.661  < 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
```

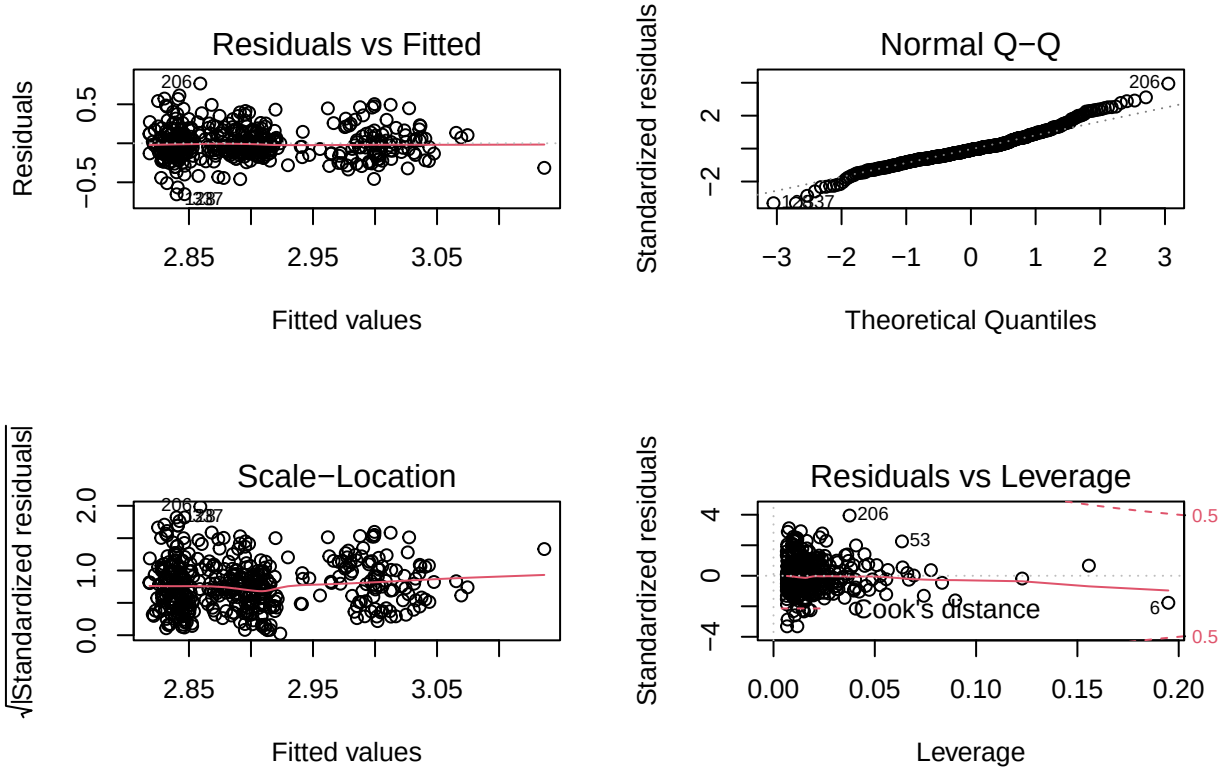
```r
par(mfrow=c(2,2))
plot(lm.3)
```




```
lm.4 <- lm(log.per.cap.income ~ log.per.cap.crime*region, data=cdilogs.reg)
summary(lm.4)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.per.cap.crime * region,
##     data = cdilogs.reg)
##
## 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)                  2.765458   0.127953  21.613   <2e-16 ***
## log.per.cap.crime            0.034535   0.033270   1.038    0.300
## regionNE                     0.008122   0.194080   0.042    0.967
## regionS                     -0.025139   0.226872  -0.111    0.912
## regionW                     -0.164523   0.375669  -0.438    0.662
## log.per.cap.crime:regionNE   0.029236   0.052324   0.559    0.577
## log.per.cap.crime:regionS   -0.011035   0.055544  -0.199    0.843
## log.per.cap.crime:regionW    0.034948   0.092675   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(lm.4)
```




```
anova(null.2, lm.2, lm.4)
```

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

```
anova(null, lm.1, lm.3)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.crimes
## Model 2: log.per.cap.income ~ log.crimes + region
## Model 3: log.per.cap.income ~ log.crimes * region
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 17.271
## 2    435 14.949  3   2.32194 22.4823 1.523e-13 ***
```

	df	AIC	BIC
lm.1	6	-227.4746	-202.9539
lm.2	6	-172.1347	-147.6140

```
## 3    432 14.872  3   0.07678  0.7434    0.5266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
data.frame(AIC=AIC(lm.1,lm.2),BIC=BIC(lm.1,lm.2))[,-3] %>%
  kbl(booktabs=T,col.names=c("df","AIC","BIC")) %>%
  kable_minimal(full_width=F)
```

Per Capita Income ~ All Predictors

Created new data sets, one without region and one with region to be used in the different varaible selection techniques.

We also removed pop, log.tot.income, and log.crimes from the dataset. Pop and log.tot.income were used in the calculation of log.per.cap.income thus removing them was necessary. We used the VIF calculation to check our logic and see if any other variables should be removed due to collinearity.

```
cdilogs.mod <- cdilogs %>%
  dplyr::select(-c(log.crimes, log.tot.income, pop))

cdilogs.mod.reg <-data.frame(cdilogs.mod,region=cdi$region)
```

```
vif <- as.data.frame(vif(lm(log.per.cap.income ~ ., data=cdilogs)))
vif
```

```
##                   vif(lm(log.per.cap.income ~ ., data = cdilogs))
## pop                                                      2.506851
## pop.18_34                                                2.745454
## pop.65_plus                                              2.187081
## pct.hs.grad                                              4.050146
## pct.bach.deg                                             6.295664
## pct.below.pov                                            5.442001
## pct.unemp                                                1.967483
## log.crimes                                             187.000319
## log.tot.income                                         125.673001
## log.land.area                                            1.372039
## log.doctors                                             17.324503
## log.hosp.beds                                            9.716302
## log.per.cap.crime                                       41.543524
```

All Subsets

Performed an all subsets regression on all the predictors (not including region). Then used BIC to pick the best model.

Once the best model was chosen we refit it in a standard linear regression to show the residual plots to check the validity of the model.

As a last check we looked at the marginal model plots to catch any possible missing interactions or transformations.

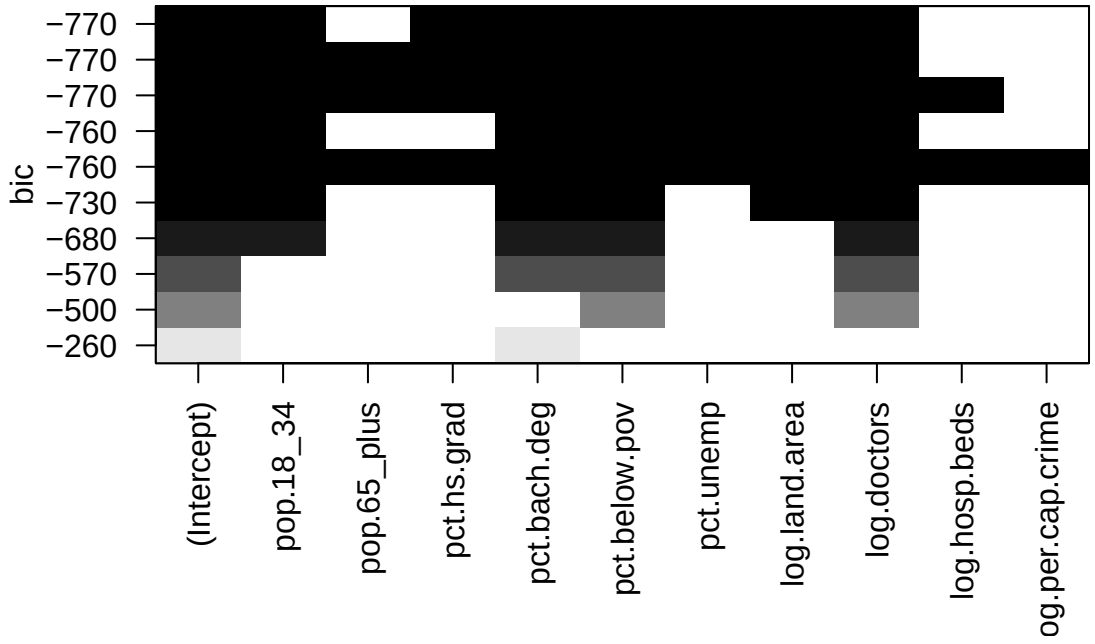
```
all.subs <- regsubsets(log.per.cap.income ~ ., data=cdilogs.mod, nvmax = 10)
outputs <- summary(all.subs)
outputs$adjr2
```

```
##  [1] 0.4570147 0.6923589 0.7406912 0.8005025 0.8237080 0.8369929 0.8426532
##  [8] 0.8439334 0.8441776 0.8441968
```

```
outputs$bic
```

```
##  [1] -257.5260 -502.4302 -572.5538 -682.8532 -732.1894 -761.5908 -772.0715
##  [8] -770.5990 -766.2235 -761.2151
```

```
plot(all.subs)
```




```
coef(all.subs, 7)
```

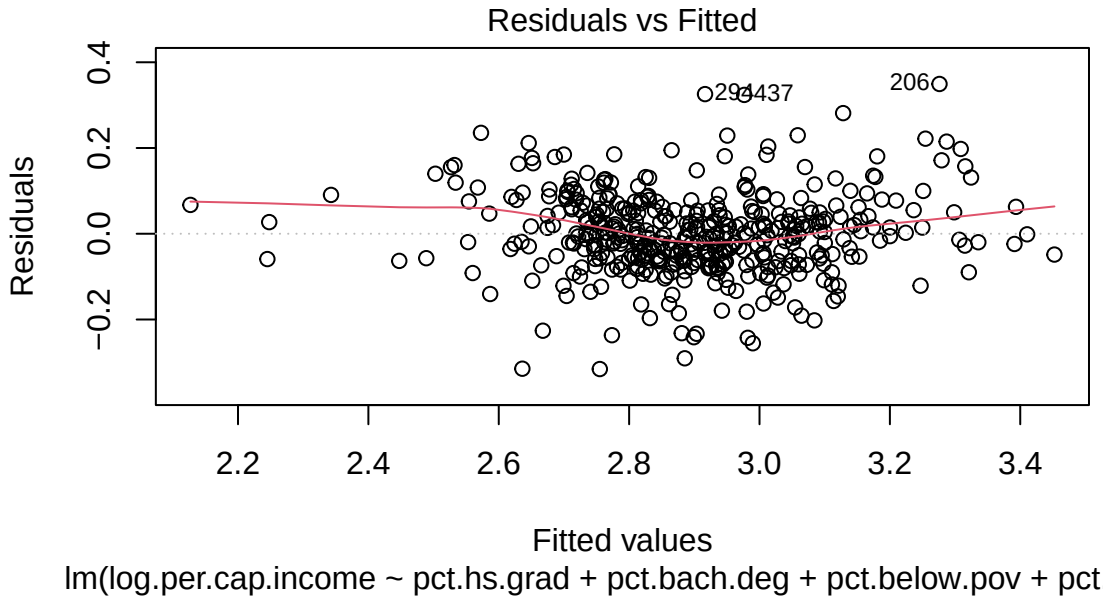
```
##   (Intercept)     pop.18_34   pct.hs.grad  pct.bach.deg pct.below.pov
##   3.314739762  -0.013900201  -0.004406396   0.015385301  -0.024278371
##     pct.unemp log.land.area   log.doctors
##   0.010603691  -0.035674062   0.060676872
```

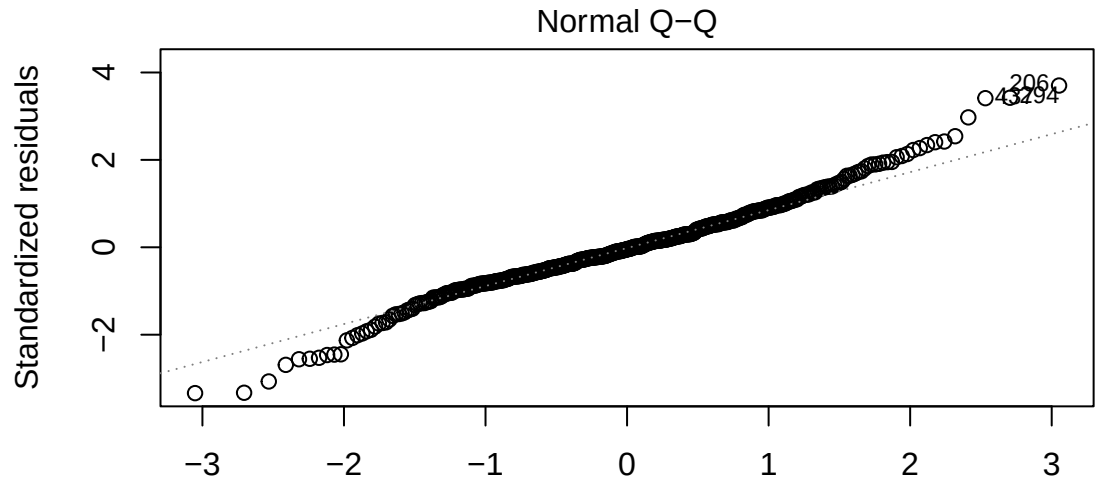
```
best.allsubs <- lm(log.per.cap.income ~ pct.hs.grad + pct.bach.deg+ pct.below.pov+ pct.unemp+ log.land.a
summary(best.allsubs)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pct.hs.grad + pct.bach.deg +
```

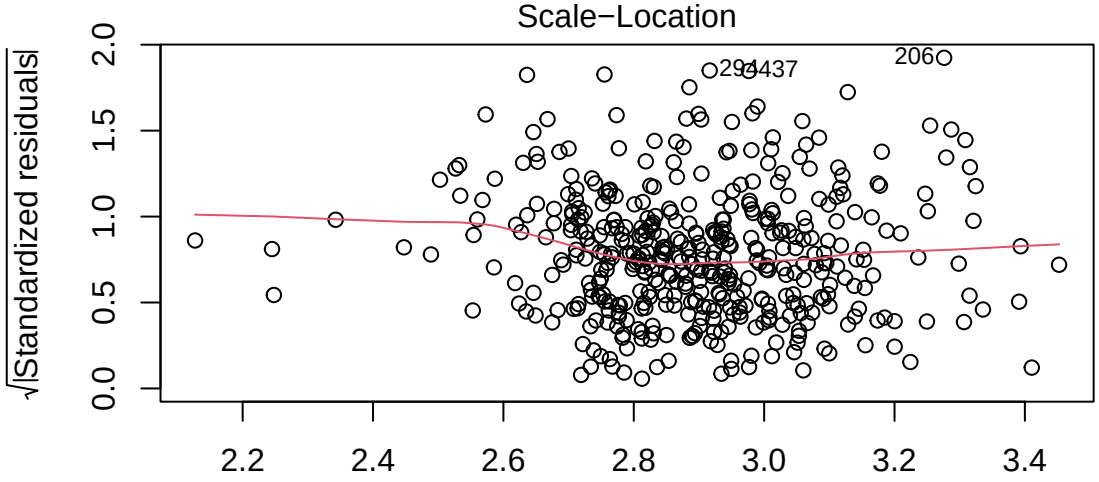
```
##     pct.below.pov + pct.unemp + log.land.area + log.doctors,
##     data = cdilogs.mod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31559 -0.05702 -0.00446  0.05381  0.34933
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.002583   0.104586  28.709  < 2e-16 ***
## pct.hs.grad   -0.004979   0.001261  -3.950 9.12e-05 ***
## pct.bach.deg   0.011336   0.001010  11.228  < 2e-16 ***
## pct.below.pov -0.028776   0.001406 -20.470  < 2e-16 ***
## pct.unemp      0.013421   0.002524   5.317 1.70e-07 ***
## log.land.area -0.031293   0.005554  -5.635 3.16e-08 ***
## log.doctors    0.066681   0.004651  14.337  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0956 on 433 degrees of freedom
## Multiple R-squared:  0.7891, Adjusted R-squared:  0.7862
## F-statistic:   270 on 6 and 433 DF,  p-value: < 2.2e-16
```

```
plot(best.allsubs)
```

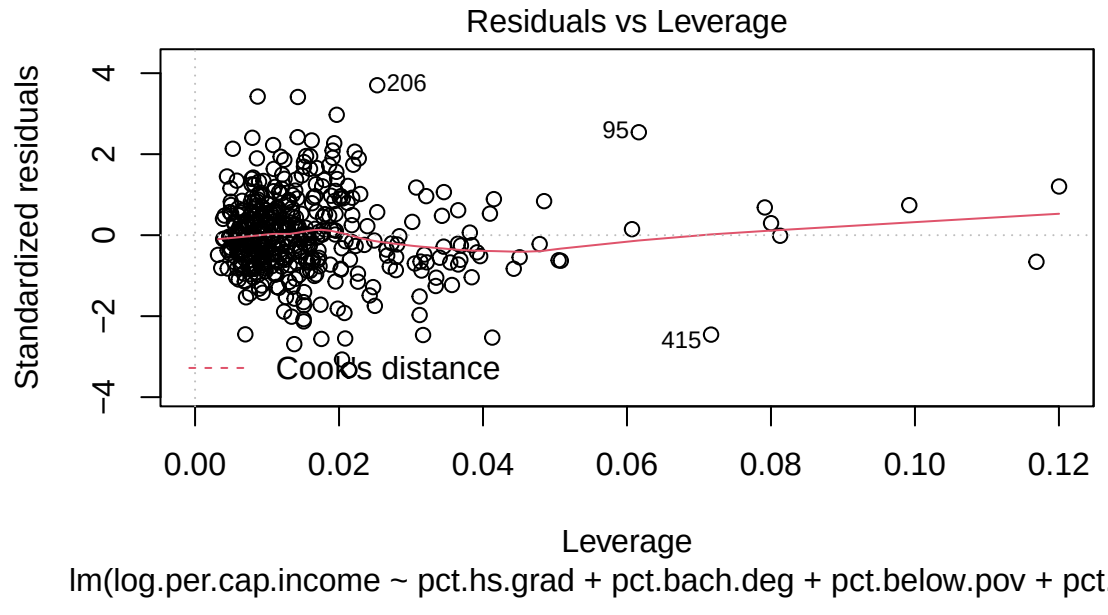




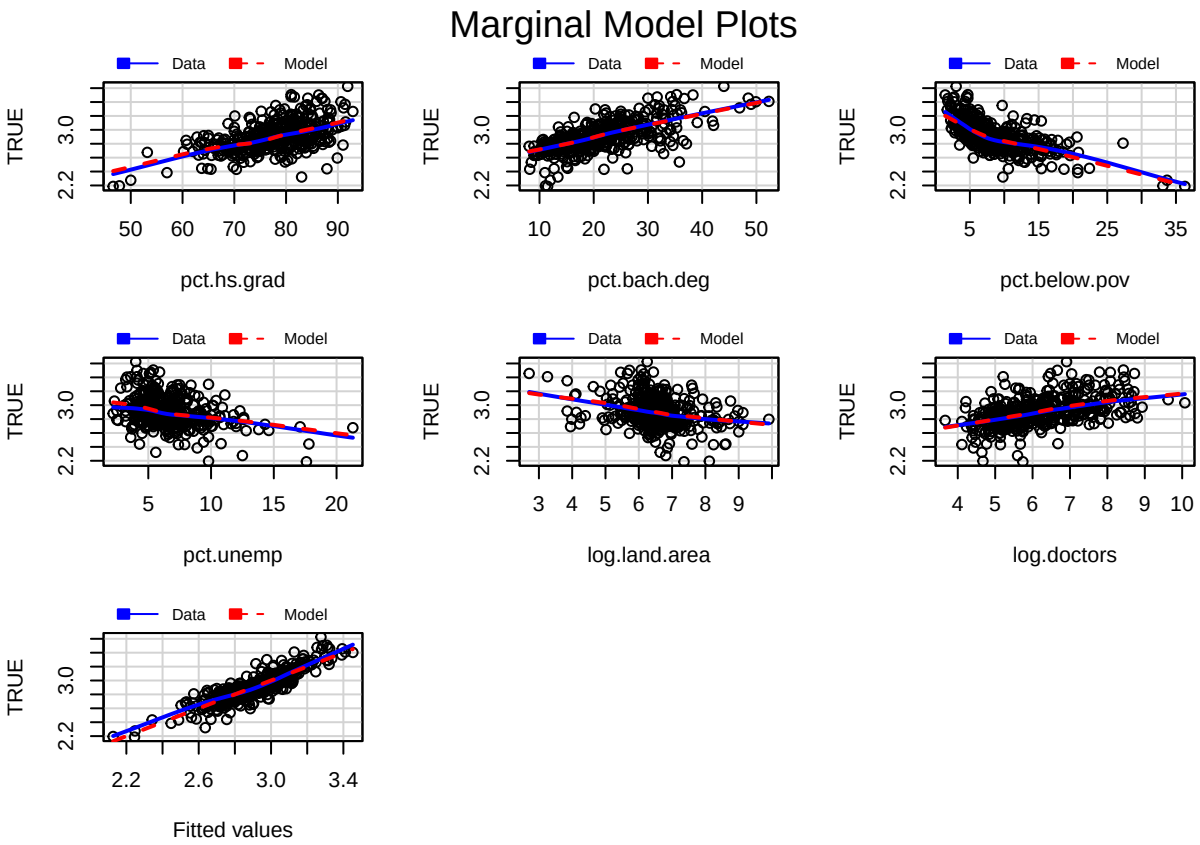
Theoretical Quantiles
lm(log.per.cap.income ~ pct.hs.grad + pct.bach.deg + pct.below.pov + pct.un ...



Fitted values
lm(log.per.cap.income ~ pct.hs.grad + pct.bach.deg + pct.below.pov + pct.un ...




```
mmps(best.allsubs)
```



Here we performed a linear regression using the best model found previously and added in region as an interaction term.

From the output we dropped log.land.area:region, pop.18_34:region, and log.doctors:region because none of the interactions were significant.

Lastly we calculated the AIC and BIC for the base model and the model with the interactions.

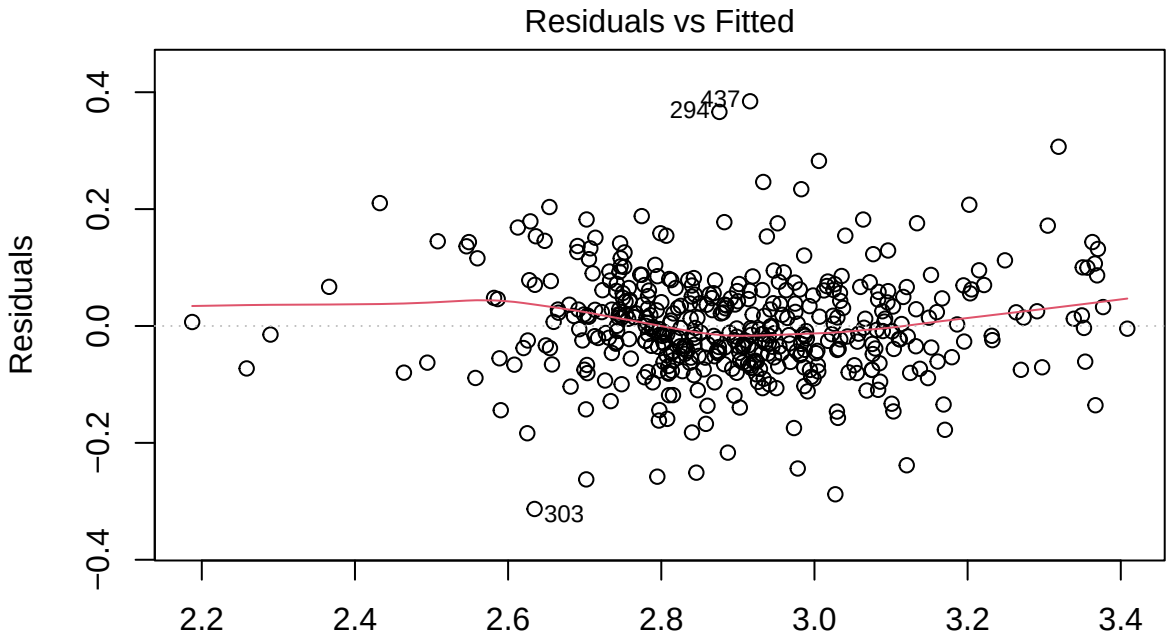
```
allsubscdi <- cdilogs.mod.reg %>%
  dplyr::select(c(pct.hs.grad, pct.bach.deg, pct.below.pov,  pct.unemp, log.land.area, log.doctors, reg


all.subs.reg <- lm(log.per.cap.income ~ .*region, data=allsubscdi)
summary(all.subs.reg)
```

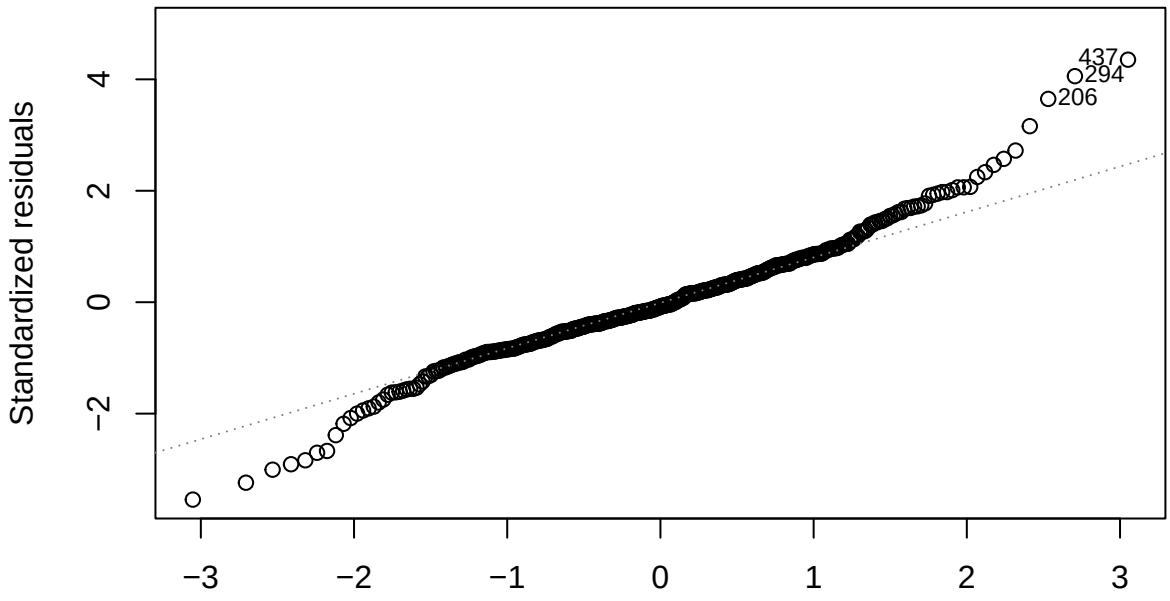
```
##
## Call:
## lm(formula = log.per.cap.income ~ . * region, data = allsubscdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30870 -0.05222 -0.00779  0.04812  0.38824
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.660788   0.318735   8.348 1.06e-15 ***
## pct.hs.grad               0.001377   0.004019   0.343 0.732091
## pct.bach.deg              0.004575   0.002884   1.586 0.113481
## pct.below.pov            -0.030093   0.004141  -7.267 1.86e-12 ***
## pct.unemp                 0.021790   0.005827   3.739 0.000211 ***
## log.land.area            -0.049229   0.017992  -2.736 0.006484 **
## log.doctors               0.075264   0.010301   7.306 1.44e-12 ***
## regionNE                  0.321322   0.400946   0.801 0.423357
## regionS                   0.278106   0.356333   0.780 0.435564
## regionW                   1.718705   0.484734   3.546 0.000437 ***
## pct.hs.grad:regionNE     -0.004202   0.005208  -0.807 0.420172
## pct.hs.grad:regionS      -0.006188   0.004442  -1.393 0.164385
## pct.hs.grad:regionW      -0.021368   0.005438  -3.929 9.99e-05 ***
## pct.bach.deg:regionNE     0.007756   0.004116   1.884 0.060201 .
## pct.bach.deg:regionS      0.006104   0.003271   1.866 0.062689 .
## pct.bach.deg:regionW      0.013758   0.003797   3.624 0.000327 ***
## pct.below.pov:regionNE   -0.001940   0.005834  -0.332 0.739694
## pct.below.pov:regionS     0.005392   0.004594   1.174 0.241126
## pct.below.pov:regionW    -0.011175   0.006370  -1.754 0.080144 .
## pct.unemp:regionNE       -0.012784   0.008830  -1.448 0.148437
## pct.unemp:regionS        -0.016194   0.007723  -2.097 0.036618 *
## pct.unemp:regionW        -0.023618   0.008101  -2.916 0.003745 **
## log.land.area:regionNE   -0.000923   0.023994  -0.038 0.969333
## log.land.area:regionS     0.024831   0.020654   1.202 0.229973
## log.land.area:regionW     0.027277   0.021694   1.257 0.209329
## log.doctors:regionNE     -0.007492   0.014961  -0.501 0.616822
## log.doctors:regionS      -0.006312   0.012964  -0.487 0.626599
## log.doctors:regionW      -0.036409   0.014798  -2.460 0.014289 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09124 on 412 degrees of freedom
## Multiple R-squared:  0.8172, Adjusted R-squared:  0.8052
## F-statistic: 68.22 on 27 and 412 DF,  p-value: < 2.2e-16
```

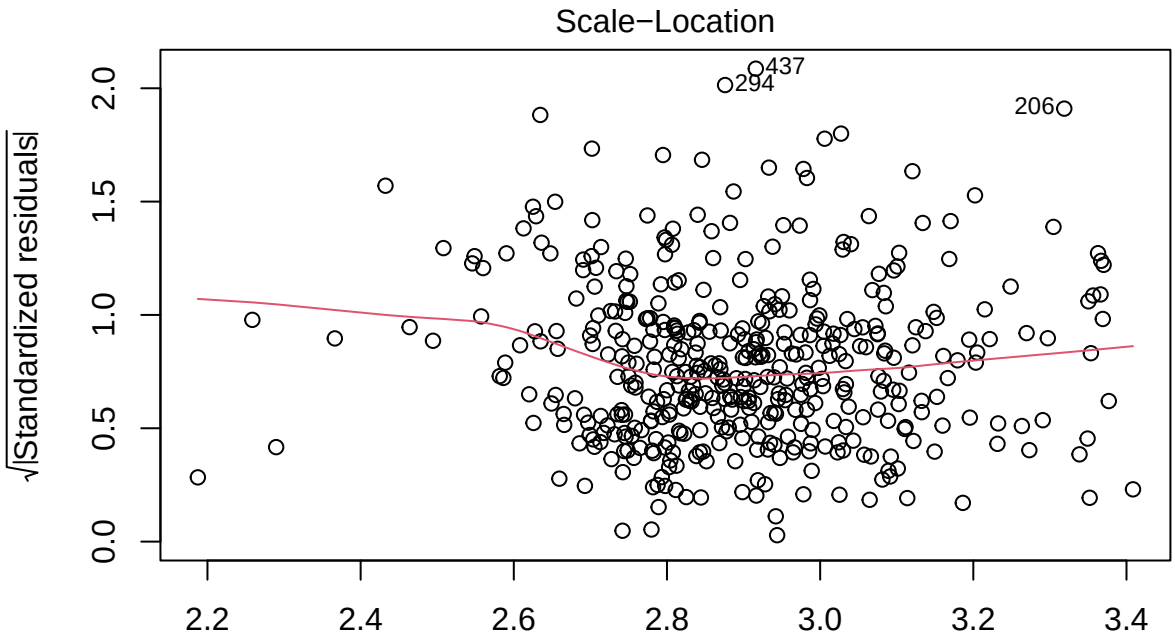
```
bestreg.allsubs <- update(all.subs.reg,. ~ . - region:log.land.area -region:pop.18_34 - region:log.docto
summary(bestreg.allsubs)
```

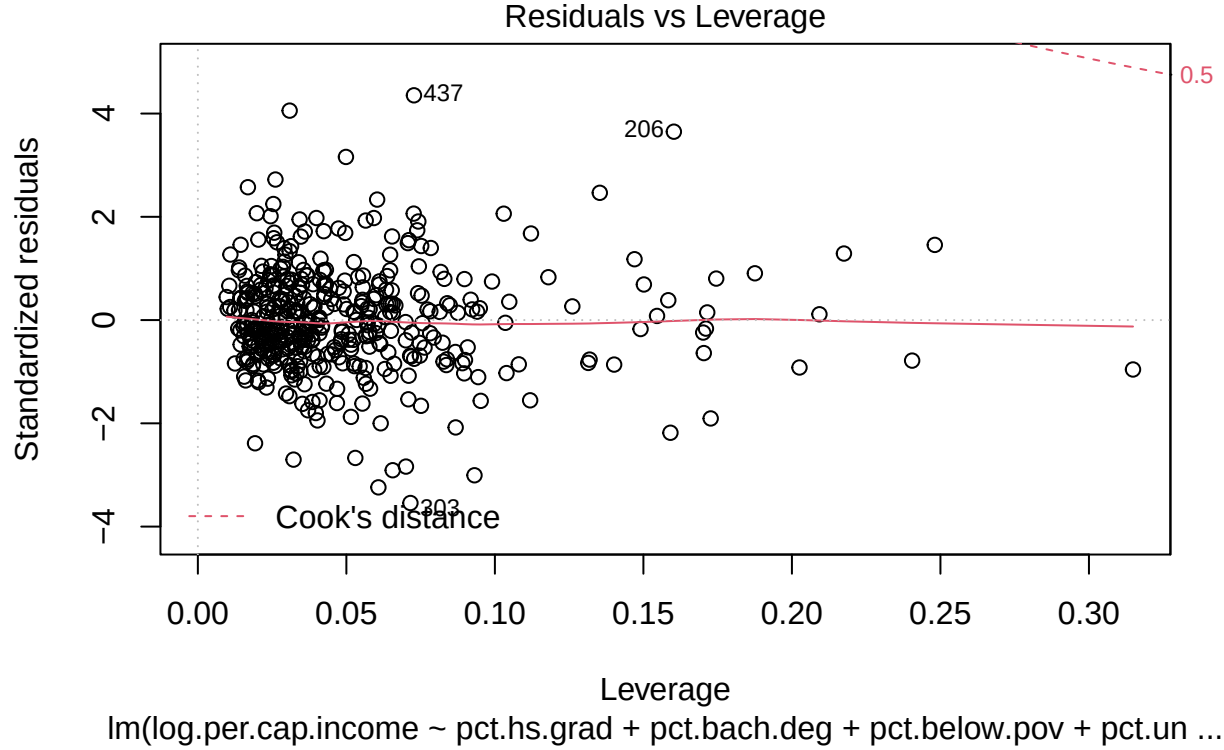
```
##
## Call:
## lm(formula = log.per.cap.income ~ pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log.land.area + log.doctors +
##     region + pct.hs.grad:region + pct.bach.deg:region + pct.below.pov:region +
##     pct.unemp:region, data = allsubscdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31312 -0.05090 -0.00701  0.04855  0.38446
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.7322450  0.3030020   9.017  < 2e-16 ***
## pct.hs.grad             -0.0006368  0.0038273  -0.166 0.867935
## pct.bach.deg             0.0066250  0.0024782   2.673 0.007805 **
## pct.below.pov           -0.0289166  0.0040482  -7.143 4.08e-12 ***
## pct.unemp                0.0212730  0.0058071   3.663 0.000281 ***
## log.land.area           -0.0321779  0.0065327  -4.926 1.21e-06 ***
## log.doctors              0.0642046  0.0048446  13.253  < 2e-16 ***
## regionNE                 0.2726560  0.3677538   0.741 0.458862
## regionS                  0.2696895  0.3344733   0.806 0.420522
## regionW                  1.3193081  0.4339614   3.040 0.002513 **
## pct.hs.grad:regionNE    -0.0042558  0.0047117  -0.903 0.366927
## pct.hs.grad:regionS     -0.0040689  0.0042706  -0.953 0.341256
## pct.hs.grad:regionW     -0.0161196  0.0051053  -3.157 0.001707 **
## pct.bach.deg:regionNE    0.0078019  0.0031785   2.455 0.014512 *
## pct.bach.deg:regionS     0.0041325  0.0027834   1.485 0.138386
## pct.bach.deg:regionW     0.0086859  0.0031788   2.732 0.006554 **
## pct.below.pov:regionNE -0.0023178  0.0056493  -0.410 0.681805
## pct.below.pov:regionS    0.0045764  0.0044753   1.023 0.307098
## pct.below.pov:regionW   -0.0091066  0.0062621  -1.454 0.146632
## pct.unemp:regionNE      -0.0127592  0.0088101  -1.448 0.148299
## pct.unemp:regionS       -0.0153619  0.0077114  -1.992 0.047010 *
## pct.unemp:regionW       -0.0199812  0.0080274  -2.489 0.013194 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09172 on 418 degrees of freedom
## Multiple R-squared:  0.8126, Adjusted R-squared:  0.8032
## F-statistic:  86.3 on 21 and 418 DF,  p-value: < 2.2e-16
```

```
plot(bestreg.allsubs)
```


Residuals vs Fitted
Residuals
Fitted values
lm(log.per.cap.income ~ pct.hs.grad + pct.bach.deg + pct.below.pov + pct.un ...
437
294
303
0.4
0.2
0.0
−0.2
−0.4
2.2
2.4
2.6
2.8
3.0
3.2
3.4


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log.per.cap.income ~ pct.hs.grad + pct.bach.deg + pct.below.pov + pct.un ...
437
294
206
4
2
0
−2
−3
−2
−1
0
1
2
3






```
data.frame(AIC=AIC(best.allsubs,bestreg.allsubs),BIC=BIC(best.allsubs,bestreg.allsubs))[,-3] %>%
  kbl(booktabs=T,col.names=c("df","AIC","BIC")) %>%
  kable_minimal(full_width=F)
```

	df	AIC	BIC
best.allsubs	8	-808.2839	-775.5897
bestreg.allsubs	23	-830.2400	-736.2441

Stepwise

Next we repeated the same process with stepwise regression, starting without region and adding it in later. We did both an AIC and BIC model to compare the results to the all subsets model.

```
stepwise_AIC <- stepAIC(lm(log.per.cap.income ~ .,data=cdilogs.mod),direction="both",k=2, trace = FALSE)
summary(stepwise_AIC)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##     log.doctors, data = cdilogs.mod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35756 -0.04551 -0.00543  0.04844  0.27399
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.4082113  0.1025858  33.223  < 2e-16 ***
## pop.18_34     -0.0153488  0.0012988 -11.818  < 2e-16 ***
## pop.65_plus   -0.0027664  0.0012978  -2.132   0.0336 *
## pct.hs.grad   -0.0046579  0.0010843  -4.296 2.15e-05 ***
## pct.bach.deg   0.0152149  0.0009242  16.462  < 2e-16 ***
## pct.below.pov -0.0246144  0.0012631 -19.488  < 2e-16 ***
## pct.unemp      0.0107688  0.0021696   4.963 9.99e-07 ***
## log.land.area -0.0364935  0.0047728  -7.646 1.36e-13 ***
## log.doctors    0.0626053  0.0041029  15.259  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08167 on 431 degrees of freedom
## Multiple R-squared:  0.8468, Adjusted R-squared:  0.8439
## F-statistic: 297.7 on 8 and 431 DF,  p-value: < 2.2e-16
```

```
stepwise_BIC <- stepAIC(lm(log.per.cap.income ~ .,data=cdilogs.mod),direction="both",k=log(nrow(cdilogs
summary(stepwise_BIC)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log.land.area + log.doctors,
##     data = cdilogs.mod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.34147 -0.04886 -0.00538  0.04818  0.26969
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.3147398  0.0931210  35.596  < 2e-16 ***
## pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
## pct.hs.grad   -0.0044064  0.0010823  -4.071 5.56e-05 ***
## pct.bach.deg   0.0153853  0.0009246  16.641  < 2e-16 ***
## pct.below.pov -0.0242784  0.0012583 -19.294  < 2e-16 ***
## pct.unemp      0.0106037  0.0021771   4.871 1.56e-06 ***
## log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
## log.doctors    0.0606769  0.0040183  15.100  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.082 on 432 degrees of freedom
## Multiple R-squared:  0.8452, Adjusted R-squared:  0.8427
## F-statistic: 336.9 on 7 and 432 DF,  p-value: < 2.2e-16
```

```
stepwisereg_AIC <- stepAIC(lm(log.per.cap.income ~ .*region,data=allsubscdi),direction="both",k=2, trac
summary(stepwisereg_AIC)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log.land.area + log.doctors +
##     region + pct.hs.grad:region + pct.bach.deg:region + pct.below.pov:region +
##     pct.unemp:region + log.doctors:region, data = allsubscdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30870 -0.05143 -0.00880  0.04925  0.38721
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.6169464  0.3162672   8.274 1.78e-15 ***
## pct.hs.grad              0.0005046  0.0039373   0.128 0.898084
## pct.bach.deg             0.0048970  0.0028700   1.706 0.088704 .
## pct.below.pov           -0.0300625  0.0041432  -7.256 1.98e-12 ***
## pct.unemp                0.0209967  0.0057832   3.631 0.000318 ***
## log.land.area           -0.0312781  0.0065154  -4.801 2.21e-06 ***
## log.doctors              0.0750128  0.0103045   7.280 1.69e-12 ***
## regionNE                 0.3403227  0.3982715   0.854 0.393321
## regionS                  0.3613728  0.3496537   1.034 0.301964
## regionW                  1.8283250  0.4763204   3.838 0.000143 ***
## pct.hs.grad:regionNE    -0.0050473  0.0049370  -1.022 0.307215
## pct.hs.grad:regionS     -0.0052400  0.0043633  -1.201 0.230467
## pct.hs.grad:regionW     -0.0203938  0.0053744  -3.795 0.000170 ***
## pct.bach.deg:regionNE    0.0089349  0.0039266   2.275 0.023387 *
## pct.bach.deg:regionS     0.0055619  0.0032415   1.716 0.086937 .
## pct.bach.deg:regionW     0.0127607  0.0036991   3.450 0.000619 ***
## pct.below.pov:regionNE -0.0015959  0.0058296  -0.274 0.784405
## pct.below.pov:regionS    0.0054667  0.0045948   1.190 0.234820
## pct.below.pov:regionW   -0.0107409  0.0063541  -1.690 0.091705 .
```

```
## pct.unemp:regionNE     -0.0121377  0.0088015  -1.379 0.168625
## pct.unemp:regionS      -0.0151578  0.0076783  -1.974 0.049033 *
## pct.unemp:regionW      -0.0228089  0.0080709  -2.826 0.004940 **
## log.doctors:regionNE   -0.0063761  0.0149549  -0.426 0.670071
## log.doctors:regionS    -0.0059443  0.0129673  -0.458 0.646900
## log.doctors:regionW    -0.0352067  0.0147637  -2.385 0.017543 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09129 on 415 degrees of freedom
## Multiple R-squared:  0.8157, Adjusted R-squared:  0.805
## F-statistic: 76.52 on 24 and 415 DF,  p-value: < 2.2e-16
```

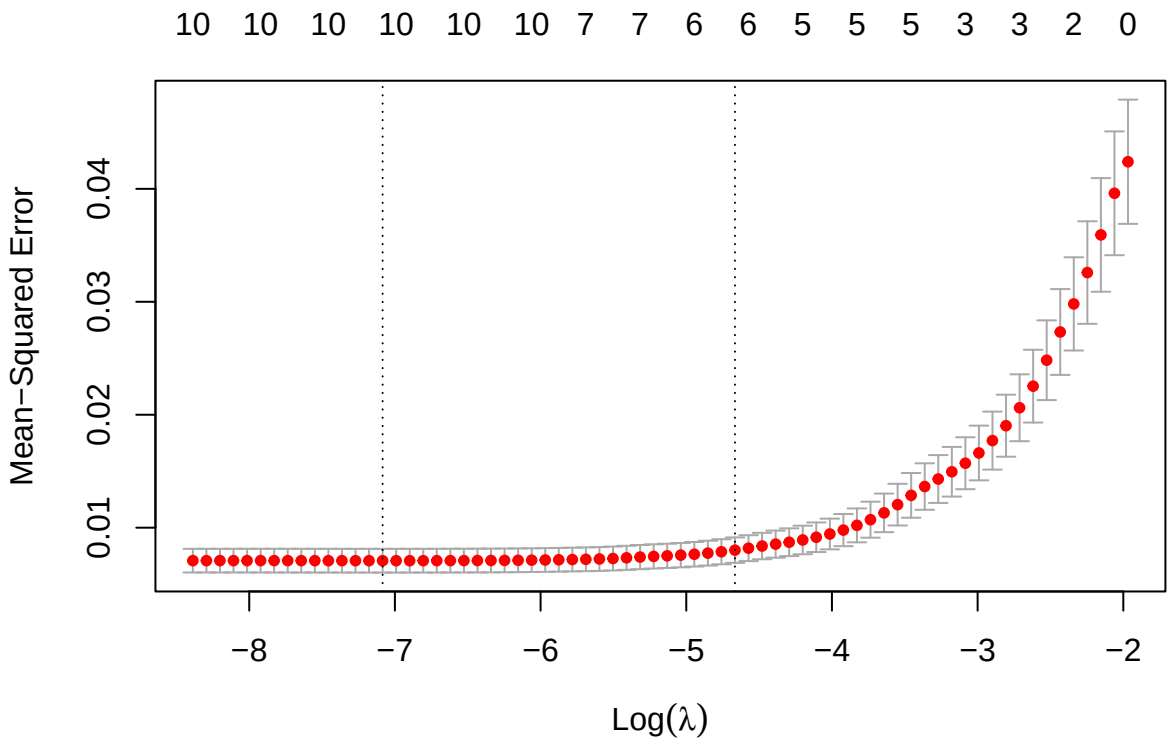
```
stepwisereg_BIC <- stepAIC(lm(log.per.cap.income ~ .*region,data=allsubscdi),direction="both",k=log(nrow
summary(stepwisereg_BIC)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log.land.area + log.doctors +
##     region + pct.bach.deg:region, data = allsubscdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33309 -0.05200 -0.00590  0.04898  0.35443
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.170626   0.119109  26.619  < 2e-16 ***
## pct.hs.grad            -0.006084   0.001345  -4.525 7.84e-06 ***
## pct.bach.deg            0.008093   0.001465   5.525 5.74e-08 ***
## pct.below.pov          -0.027787   0.001496 -18.575  < 2e-16 ***
## pct.unemp               0.010820   0.002706   3.999 7.50e-05 ***
## log.land.area          -0.028486   0.006395  -4.454 1.08e-05 ***
## log.doctors             0.064023   0.004634  13.815  < 2e-16 ***
## regionNE               -0.180452   0.039225  -4.600 5.57e-06 ***
## regionS                -0.099337   0.034709  -2.862  0.00442 **
## regionW                -0.119614   0.047127  -2.538  0.01150 *
## pct.bach.deg:regionNE   0.008450   0.001778   4.752 2.76e-06 ***
## pct.bach.deg:regionS    0.003516   0.001557   2.258  0.02442 *
## pct.bach.deg:regionW    0.005557   0.001987   2.797  0.00539 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09308 on 427 degrees of freedom
## Multiple R-squared:  0.8028, Adjusted R-squared:  0.7973
## F-statistic: 144.9 on 12 and 427 DF,  p-value: < 2.2e-16
```

Lasso

As a last check, we performed lasso regression with cross validation and the 1se lambda.

```
y <- cdilogs.mod$log.per.cap.income
x <- data.matrix(cdilogs.mod[,-11])

lasso.cdi <- cv.glmnet(x,y, alpha = 1)
plot(lasso.cdi)
```




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

```
##   lambda.1se   lambda.min
## 0.0094134316 0.0008379979
```

```
cbind(coef(lasso.cdi),coef(lasso.cdi,s=lasso.cdi$lambda.1se),coef(lasso.cdi,s=lasso.cdi$lambda.min))
```

```
## 11 x 3 sparse Matrix of class "dgCMatrix"
##                              s1           s1           s1
## (Intercept)        2.959121800  2.959121800  3.299856539
## pop.18_34         -0.010777113 -0.010777113 -0.014863142
## pop.65_plus        .            .           -0.002225481
## pct.hs.grad        .            .           -0.003950077
## pct.bach.deg       0.010843351  0.010843351  0.015103087
## pct.below.pov     -0.019273623 -0.019273623 -0.024464019
## pct.unemp          0.002733261  0.002733261  0.010499721
## log.land.area     -0.028388957 -0.028388957 -0.035461355
## log.doctors        0.057616865  0.057616865  0.052026942
## log.hosp.beds      .            .           0.010027064
## log.per.cap.crime  .            .           0.006802218
```

```
bestmodel.lasso <- glmnet(x,y, alpha = 1, lambda = lasso.cdi$lambda.1se)
coef(bestmodel.lasso)
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## (Intercept)        2.958937225
## pop.18_34         -0.010766676
## pop.65_plus        .
## pct.hs.grad        .
## pct.bach.deg       0.010835039
## pct.below.pov     -0.019278288
## pct.unemp          0.002730723
## log.land.area     -0.028387892
## log.doctors        0.057635089
## log.hosp.beds      .
## log.per.cap.crime  .
```