Analysis of county demographic information: predicting average income per person

Julia Keating
jhkeatin@andrew.cmu.edu

1 Abstract

In this paper we will analyze how average income per person is related to other variables associated with a county's economic, health and social well-being. Our data consists of selected county demographic information for 440 of the most populous counties in the United States. We will develop a model to predict per capita income from total number of crimes, as well as the best model to predict per capita income from all variables in the dataset using a variety of methods such as the all subsets method. Models will be compared using partial F tests and assessed for validity and goodness of fit through diagnostic plots, R^2 values, VIF's and BIC values. Our final models suggest that the relationship between per capita income and total number of crimes is different in different regions of the country. Additionally, we found that per capita income of a county is best predicted by the land area of the county, the percent of the county aged 18-34, the number of active physicians in the county, the percent of the adult population of the county who completed 12 or more years of school, the percent of the adult population of the county with bachelor's degrees, the percent of the population of the county with income below the poverty level, the percent of the population of the county that is unemployed, and the region of the United States the county is in.

2 Introduction

Social scientists are interested in looking at historical county demographic information to learn how average income per person was related to other variables associated with the county's economic, health and social well-being. To provide insight about these relationships, we will answer the following research questions:

1. Which variables are related within the data? Are all of the relationships what a reasonable person would expect, or are there some surprises?
2. Is per-capita income related to crime rate, and is this relationship different in different regions of the country? Does it matter if we use number of crimes or per capita crimes in our analysis?
3. What is the best model predicting per-capita income from the other variables?
4. Should we be worried about either the missing states or the missing counties in our dataset?

3 Data

Our data is taken from Kutneret al. (2005), and consists of selected county demographic information for 440 of the most populous counties in the United States. There are no counties with missing data in our dataset, which was verified by checking for NA's in the dataset. Our dataset contains 17 variables which are defined in Table 1. Our response variable for this analysis is per capita income. Table 2 contains a summary of the quantitative variables except for identification number which is not useful in our analysis. Tables 3, 4 and 5 contain summaries of the 3 categorical variables. We note that the county and state variables contain a large number of unique values (373 and 48, respectively). Thus, these variables are not very useful and we will exclude them in this analysis.

Figure 1 contains histograms for the quantitative variables excluding identification number. Note that the data for land.area, pop, doctors, hosp.beds, crimes, tot.income and per.cap.income are strongly right-skewed, which suggests we may want to apply transformations these variables later in our analysis.

Figure 2 displays the correlation matrix of the quantitative variables, excluding identification number, as a heatmap. We note that total personal income and total population are highly correlated, which is not surprising because we would generally expect a larger population to have a larger total personal income. We also note that total personal income, total population, active physicians, number of hospital beds and total serious crimes all appear to be fairly highly correlated with eachother. Finally, note that per capita income, our response variable, is not very highly correlated with any variables in the plot. Per capita income is most strongly correlated with percent bachelor's degrees, percent high school graduates, and percent below poverty level.

Variable number	Variable name	Definition
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 older
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, aggrevated 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 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)

Variable Name	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Land area	15.0	451.2	656.5	1041.4	946.8	20062.0
Total population	100043	139027	217280	393011	436064	8863164
Percent of population aged 18-34	16.40	26.20	28.10	28.57	30.02	49.70
Percent of population aged 65 or older	3.000	9.875	11.750	12.170	13.625	33.800
Number of active physicians	39.0	182.8	401.0	988.0	1036.0	23677.0
Number of hospital beds	92.0	390.8	755.0	1458.6	1575.8	27700.0
Total serious crimes	563	6220	11820	27112	26280	688936
Percent high school graduates	46.60	73.88	77.70	77.56	82.40	92.90
Percent bachelor's degrees	8.10	15.28	19.70	21.08	25.32	52.30
Percent below poverty level	1.400	5.300	7.900	8.721	10.900	36.300
Percent unemployment	2.200	5.100	6.200	6.597	7.500	21.300
Per capita income	8899	16118	17759	18561	20270	37541
Total personal income	1141	2311	3857	7869	8654	184230

Table 2: Summary of quantitative variables in dataset, excluding id

4 Methods

In order to determine which variables are related within the data, we began with simple exploratory data analysis. As a result of this analysis, we also decided to apply log transformations to the variables land area, total population, number of active physicians, number of hospital beds, crimes, total personal income and per capita income to address the extreme right skew in the distribution of these variables. See pages 6-8 in the Technical Appendix for more detail about the effects of this transformation on the distributions of the variables.

Frequency	Number of unique counties
1	334
2	23
3	10
4	3
5	1
6	1
7	1

Table 3: Summary of counties in the dataset

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

Table 4: Summary of states in the dataset

Region	Count
NC	108
NE	103
S	152
W	77

Table 5: Summary of regions in the dataset

To determine if per capita income should be related to crime rate, we first fit linear models which predict log(per capita income) from log(total serious crimes) as well as from log(total serious crimes) and geographic region, with and without an interaction term between log(total serious crimes) and geographic region. Next, to determine whether using crime rate, where crime rate equals total serious crimes divided by total population, instead of total serious crimes made a difference, we fit linear models which predict log(per capita income) from log(crime rate) in addition to log(crime rate) and geographic region, with and without an interaction term between log(crime rate) and geographic region. We assessed each of the six models using residual plots and compared models using partial F tests in addition to AIC and BIC values to decide which model was the best model.

We then shifted our focus in order to find the best model predicting log(per capita income) using all predictors. We excluded the variables identification number, county and state as they are not useful in our analysis. We also excluded the variables total population and total personal income which are directly related to our response variable, per capita income. We started by using the all subsets method on all remaining predictors, excluding region temporarily, with transformations as described previously. We fit a new model containing region, the predictors in the model chosen by the all subsets method, and interaction terms between all of these predictors and region. We removed interaction terms from the model which did not have statistically significant coefficients for any factor. We compared the resulting model to the model chosen by the all subsets method using a partial F test. We repeated this same process with stepwise regression using the AIC criterion and using the BIC criterion. The final three models were assessed using VIF values and residual plots. Model characteristics were taken into account in addition to AIC and BIC values to choose our final model.

Lastly, to decide whether the missing states and counties in our dataset are problematic, we compared the number of counties in each state to the number of counties in our dataset to see generally how representative our dataset is of all counties in the United States.

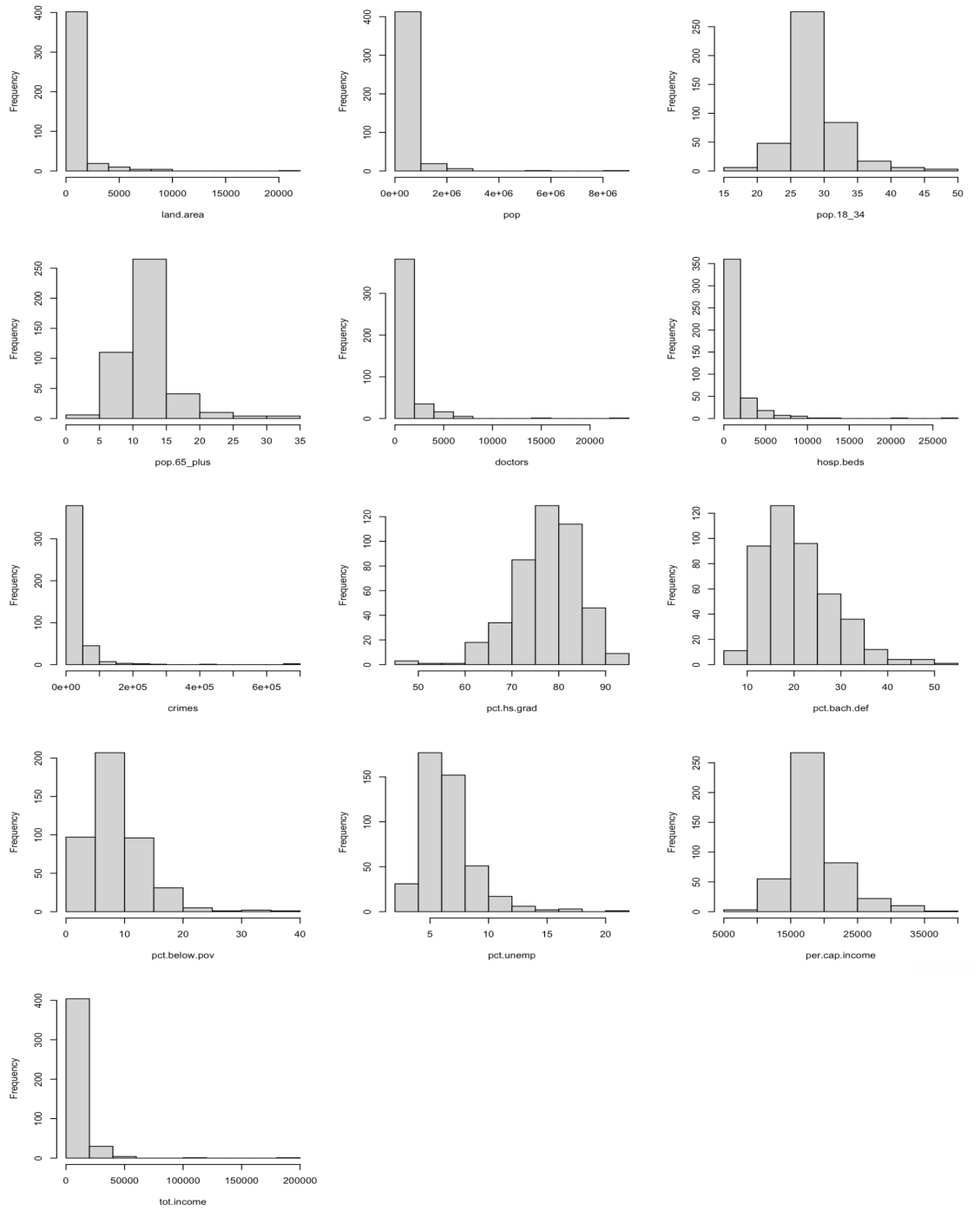


Figure 1: Histograms of quantitative variables in dataset, excluding id

5 Results

First, our exploratory data analysis revealed that total personal income and total population are highly correlated, which is not surprising because we would generally expect a larger population to have a larger total personal income. Additionally, we saw that total personal income, total population, active physicians, number of hospital beds and total serious crimes all appear to be fairly highly correlated with eachother. This is also not entirely surprising because we would expect larger, more populous counties to have larger hospitals and thus a greater number of physicians and hospital beds, as well as a larger number of total crimes if we assume that crime rate is relatively constant. Surprisingly, we saw that per capita income, our response variable, was not very highly correlated with any variables in the dataset. Per capita income was most strongly correlated with percent bachelor's degrees, percent high school graduates, and percent below poverty level. It makes sense that per capita income is fairly strongly correlated with the education variables, because we would expect more educated counties to have a higher income. It also is not surprising that per capita income is

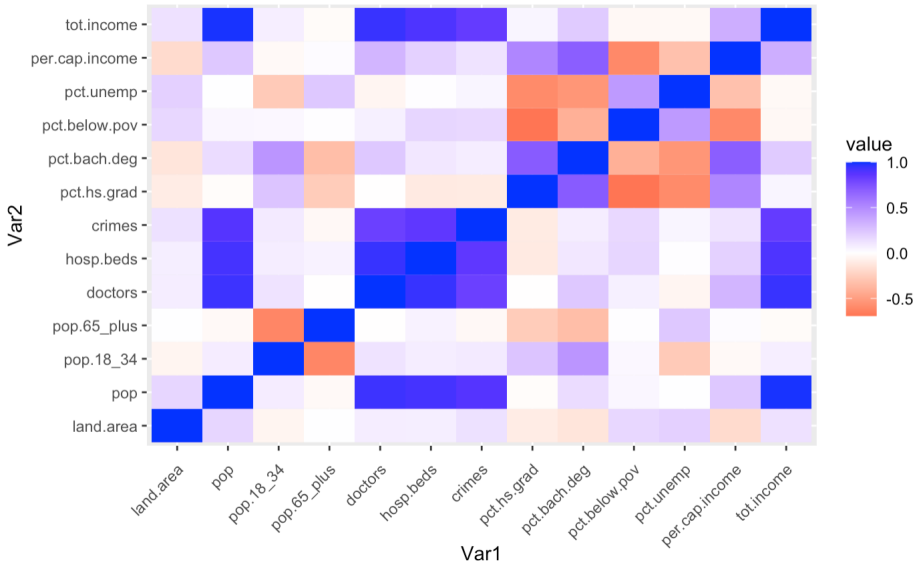


Figure 2: Heatmap of correlation matrix of quantitative variables, excluding id

related to percent below poverty level since we would expect a county with a large percent of the population with income below the poverty level to have a smaller per capita income, which is evidenced by the negative correlation between these variables.

We then fit a linear model which predicts log(per capita income) from log(crimes), a linear model which predicts log(per capita income) from log(crimes) and region, and a linear model which predicts log(per capita income) from log(crimes), region and the interaction between log(crimes) and region. We then fit the same three models but using log(crime rate) instead of log(crimes), where crime rate is crimes/total.pop. These models are shown below. Essentially, model (1) predicts that per capita income is the same for a set number of crimes regardless of region and for a set change in number of crimes, the change in per capita income is the same regardless of region. Model (2) predicts that per capita income is the same for a set number of crimes regardless of region but for a set change in number of crimes, the change in per capita income differs by region. Model (3) predicts that per capita income differs for a set number of crimes depending on region and for a set change in number of crimes, the change in per capita income differs by region. Models (4), (5) and (6) can be interpreted the same way but using crime rate instead of total number of crimes.

$$\log(\text{per.cap.income}) \sim \log(\text{crimes}) \tag{1}$$
$$\log(\text{per.cap.income}) \sim \log(\text{crimes}) + \text{region} \tag{2}$$
$$\log(\text{per.cap.income}) \sim \log(\text{crimes}) + \text{region} + \log(\text{crimes}) * \text{region} \tag{3}$$
$$\log(\text{per.cap.income}) \sim \log(\text{crime.rate}) \tag{4}$$
$$\log(\text{per.cap.income}) \sim \log(\text{crime.rate}) + \text{region} \tag{5}$$
$$\log(\text{per.cap.income}) \sim \log(\text{crime.rate}) + \text{region} + \log(\text{crime.rate}) * \text{region} \tag{6}$$

We will first consider models (1), (2) and (3). All three models appear to be equally valid models according to their residual plots. All three models are statistically significant, but model (3) contains several coefficients which are not statistically significant, unlike models (1) and (2). We also notice that model (1) has a much smaller R^2 value than models (2) and (3), which have very similar R^2 values. A partial F test between models (1) and (2) indicates that we have significant evidence that model (2) is a better fit for the data than model (1), i.e. the region term significantly improves the model. A partial F test between models (2) and (3) indicates that we do not have significant evidence that model (3) is a better fit for the data than model (2). Thus, we choose model (2) as the best model among models (1), (2) and (3). See pages 10-14 of the Technical Appendix for more detail about these models.

We will now consider models (4), (5) and (6). Residual plots suggest that model (4) is a valid model, however we see a somewhat concerning clustering pattern between the residuals and fitted values for models (5) and (6). We note that models (5) and (6) are statistically significant while model (4) is not. Also, we note that all coefficients except the intercept in models (4) and (6) are not statistically significant, while all coefficients in model (5) are significant except for one. Model (4) has a very small R^2 value compared to models (5) and

(6) which have very similar R^2 values. A partial F test between models (4) and (5) indicates that we have significant evidence that model (5) is a better fit for the data than model (4), i.e. the region term significantly improves the model. A partial F test between models (5) and (6) indicates that we do not have significant evidence that model (6) is a better fit for the data than model (5). Thus, we choose model (5) as the best model among models (4), (5) and (6). See pages 14-18 of the Technical Appendix for more detail about these models.

Lastly, we compare models (2) and (5). Model (2) is a statistically significant model and all coefficients in the model are significant as well. Model (2) is a valid model with an R^2 value of 0.1959. Model (5) is a statistically significant model and all coefficients in the model are significant except for one. Model (5) does not appear to be a completely valid model due to the clustering trend between residuals and fitted values, and the value of R^2 for this model is only 0.08814. For these reasons. We choose model (2) as the best model of the six models fit.

Next in our analysis, we applied the all subsets method to all variables except for id, county, state, log.pop and log.tot.income, including region. The model with the lowest BIC value produced by this method is as follows:

$$\begin{aligned}\log(\text{per.cap.income}) \sim \log(\text{land.area}) + \text{pop.18\_34} + \log(\text{doctors}) + \text{pct.hs.grad} + \text{pct.bach.deg} \\ + \text{pct.below.pov} + \text{pct.unemp} \end{aligned} \quad (7)$$

We then fit a model including all predictors in model (7), region, and the interactions between region and all predictors in model (7). We removed all interaction terms which did not contain a significant coefficient for at least one factor of region. The resulting model is as follows:

$$\begin{aligned}\log(\text{per.cap.income}) \sim \log(\text{land.area}) + \text{pop.18\_34} + \log(\text{doctors}) + \text{pct.hs.grad} + \text{pct.bach.deg} \\ + \text{pct.below.pov} + \text{pct.unemp} + \text{region} + \text{pct.hs.grad} * \text{region} \\ + \text{pct.below.pov} * \text{region} + \text{pct.unemp} * \text{region} \end{aligned} \quad (8)$$

A partial F test between models (7) and (8) indicates that we have significant evidence that model (8) is a better fit for the data than model (7), i.e. the region term and the included interactions with region significantly improve the model. See pages 26-29 of the Technical Appendix for more detail about this test. The estimated coefficients and their associated standard errors are provided in Table 6

Coefficient	Estimate	Standard Error
Intercept	10.242123935	0.2176557
log(land.area)	-0.038173762	0.0053996
pop.18_34	-0.014934657	0.0010897
log(doctors)	0.057228443	0.0040082
pct.hs.grad	-0.004353194	0.0024515
pct.bach.deg	0.015630966	0.0009715
pct.below.pov	-0.025202882	0.0032612
pct.unemp	0.019739969	0.0046254
regionNE	-0.052006957	0.2707173
regionS	-0.038971766	0.2383516
regionW	1.391048448	0.3408962
pct.hs.grad*regionNE	0.001768418	0.0029293
pct.hs.grad*regionS	0.001152511	0.0025618
pct.below.pov*regionNE	-0.014147323	0.0035826
pct.hs.grad*regionW	-0.001517033	0.0046143
pct.below.pov*regionS	0.007018461	0.0035199
pct.below.pov*regionW	-0.013791967	0.0051811
pct.unemp*regionNE	-0.012984072	0.0070423
pct.unemp*regionS	-0.023113781	0.0061365
pct.unemp*regionW	-0.021735737	0.0065225

Table 6: Estimated coefficients for model (8)

Model (8) appears to be a valid model. See pages 26-27 of the Technical Appendix for more detail about the validity of the model. Model (8) is also a good fit for the data since its R^2 value is equal to 0.8615, which is very close to 1. Additional models were fit through other variable selection methods, but we do not present

them here because model (8) was deemed to be a valid and good enough model which was much simpler than the other models fit. See pages 20-26 and 29-31 of the Technical Appendix for more detail about the other models considered in our analysis.

Lastly, we wanted to determine if we should be concerned about missing states or counties in the dataset. We do not believe that our dataset is representative of all counties in the US because the number of counties in our dataset for each state is not proportional to the actual number of counties in that state. For example, note that the state Texas contains 254 counties while the state Pennsylvania contains only 67 counties, but our dataset contains almost the same number of counties for Texas and Pennsylvania (28 and 29 respectively) (The Fact File, 2021). Also, note that the states California and Montana contain a similar number of counties (58 and 56 respectively) but there are 34 counties in California in our dataset and only 8 counties in Montana in our dataset (The Fact File, 2021). Thus, if we would like use our model to predict per capita income of any county in the United States, we could be concerned about the missing counties and states in our dataset.

6 Discussion

The first goal of our analysis was to determine what relationships existed between the variables in our dataset. We saw that total personal income and total population are highly correlated, which is not surprising because we would generally expect a larger population to have a larger total personal income. We also found that total personal income, total population, active physicians, number of hospital beds and total serious crimes all appear to be fairly highly correlated with eachother. This is also not entirely surprising because we would generally expect more populous counties to have more doctors, larger hospitals and a greater number of crimes. Our response variable, per capita income, did not appear to be very highly correlated with any variables, but was most strongly correlated with percent bachelor's degrees, percent high school graduates, and percent below poverty level. This is not surprising because we would generally expect more educated counties to have higher per capita incomes and counties with a high percent of the population with income below the poverty level to have a lower per capita income.

Next, we found that the best model predicting per capita income from total number of crimes included region but not the interaction between total number of crimes and region. Using crime rate instead of total number of crimes in the model did not make a significant difference. Our model suggests that the relationship between total number of crimes and per capita income is different in different regions of the country, with per capita income being smallest in the southern region and largest in the north-eastern region for a set total number of crimes. However, for a set change in the total number of crimes, our model estimates that the change in per capita income is the same regardless of region.

Then, we found that the best model predicting per capita income from all variables in the dataset was model (8). We chose this model as our best model because it was a valid model which was a good fit to the data and was not as complex as models produced by other methods. Model (8) implies that per capita income of a county is best predicted by the land area of the county, the percent of the county aged 18-34, the number of active physicians in the county, the percent of the adult population of the county who completed 12 or more years of school, the percent of the adult population of the county with bachelor's degrees, the percent of the population of the county with income below the poverty level, the percent of the population of the county that is unemployed, and the region of the United States the county is in. We did not consider interactions between quantitative variables in fitting this model because we did not want to produce an extremely complicated model, but some of these interactions could improve the model and should be considered in future research.

Lastly, we considered whether we should be concerned about missing counties and states in the dataset. We found that our dataset is not very representative of all counties in the United States. Thus, our analysis pertains only to the most populous counties in the United States, not all counties in the United States. Our model would likely be improved by collecting more data. Collecting data on more counties would help ensure that our dataset is representative of all counties in the US and that our results are applicable to all counties in the United States. In addition, a larger dataset might allow us to include the variable state in our model which might improve the fit.

7 References

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

The Fact File. 2021. List Of U.S. States And Number of Counties In Each - The Fact File. [online] Available at: ¡https://thefactfile.org/us-states-counties/¿ [Accessed 30 October 2021].

Technical Appendix

Read in the data and summarize quantitative variables

```
cdi <- read.delim("cdi.dat", header = TRUE, sep=" ")
head(cdi)
```

```
##   id      county state land.area     pop pop.18_34 pop.65_plus doctors
## 1  1 Los_Angeles    CA      4060 8863164      32.1         9.7   23677
## 2  2        Cook    IL       946 5105067      29.2        12.4   15153
## 3  3      Harris    TX      1729 2818199      31.3         7.1    7553
## 4  4   San_Diego    CA      4205 2498016      33.5        10.9    5905
## 5  5      Orange    CA       790 2410556      32.6         9.2    6062
## 6  6       Kings    NY        71 2300664      28.3        12.4    4861
##   hosp.beds crimes pct.hs.grad pct.bach.deg pct.below.pov pct.unemp
## 1     27700 688936        70.0         22.3          11.6       8.0
## 2     21550 436936        73.4         22.8          11.1       7.2
## 3     12449 253526        74.9         25.4          12.5       5.7
## 4      6179 173821        81.9         25.3           8.1       6.1
## 5      6369 144524        81.2         27.8           5.2       4.8
## 6      8942 680966        63.7         16.6          19.5       9.5
##   per.cap.income tot.income region
## 1          20786     184230      W
## 2          21729     110928     NC
## 3          19517      55003      S
## 4          19588      48931      W
## 5          24400      58818      W
## 6          16803      38658     NE
```

```
summary(cdi)
```

```
##        id           county             state             land.area
##  Min.   :  1.0   Length:440         Length:440         Min.   :   15.0
##  1st Qu.:110.8   Class :character   Class :character   1st Qu.:  451.2
##  Median :220.5   Mode  :character   Mode  :character   Median :  656.5
##  Mean   :220.5                                         Mean   : 1041.4
##  3rd Qu.:330.2                                         3rd Qu.:  946.8
##  Max.   :440.0                                         Max.   :20062.0
##       pop            pop.18_34      pop.65_plus        doctors
##  Min.   : 100043   Min.   :16.40   Min.   : 3.000   Min.   :   39.0
##  1st Qu.: 139027   1st Qu.:26.20   1st Qu.: 9.875   1st Qu.:  182.8
##  Median : 217280   Median :28.10   Median :11.750   Median :  401.0
##  Mean   : 393011   Mean   :28.57   Mean   :12.170   Mean   :  988.0
##  3rd Qu.: 436064   3rd Qu.:30.02   3rd Qu.:13.625   3rd Qu.: 1036.0
##  Max.   :8863164   Max.   :49.70   Max.   :33.800   Max.   :23677.0
##    hosp.beds           crimes        pct.hs.grad     pct.bach.deg
##  Min.   :   92.0   Min.   :   563   Min.   :46.60   Min.   : 8.10
##  1st Qu.:  390.8   1st Qu.:  6220   1st Qu.:73.88   1st Qu.:15.28
##  Median :  755.0   Median : 11820   Median :77.70   Median :19.70
```

```
##  Mean   : 1458.6   Mean   : 27112   Mean   :77.56   Mean   :21.08
##  3rd Qu.: 1575.8   3rd Qu.: 26280   3rd Qu.:82.40   3rd Qu.:25.32
##  Max.   :27700.0   Max.   :688936   Max.   :92.90   Max.   :52.30
##  pct.below.pov      pct.unemp      per.cap.income    tot.income
##  Min.   : 1.400   Min.   : 2.200   Min.   : 8899   Min.   :  1141
##  1st Qu.: 5.300   1st Qu.: 5.100   1st Qu.:16118   1st Qu.:  2311
##  Median : 7.900   Median : 6.200   Median :17759   Median :  3857
##  Mean   : 8.721   Mean   : 6.597   Mean   :18561   Mean   :  7869
##  3rd Qu.:10.900   3rd Qu.: 7.500   3rd Qu.:20270   3rd Qu.:  8654
##  Max.   :36.300   Max.   :21.300   Max.   :37541   Max.   :184230
##     region
##  Length:440
##  Class :character
##  Mode  :character
##
##
##
```

Summary of county

```
c <- data.frame(table(cdi$county))
unique(c$Freq)
```

```
## [1] 1 2 3 4 7 6 5
```

```
length(which(c$Freq==1))
```

```
## [1] 334
```

```
length(which(c$Freq==2))
```

```
## [1] 23
```

```
length(which(c$Freq==3))
```

```
## [1] 10
```

```
length(which(c$Freq==4))
```

```
## [1] 3
```

```
length(which(c$Freq==5))
```

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

```
length(which(c$Freq==6))
```

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

```
length(which(c$Freq==7))
```

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

We see that there are 334 out of 373 unique counties which only appear once in the dataset, so this variable is not very useful.

Summary of state

```
table(cdi$state)
```

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

Similarly, we see that there are several states which only appear once in the dataset, such as DC and ID, so this variable is not very useful

Summary of region

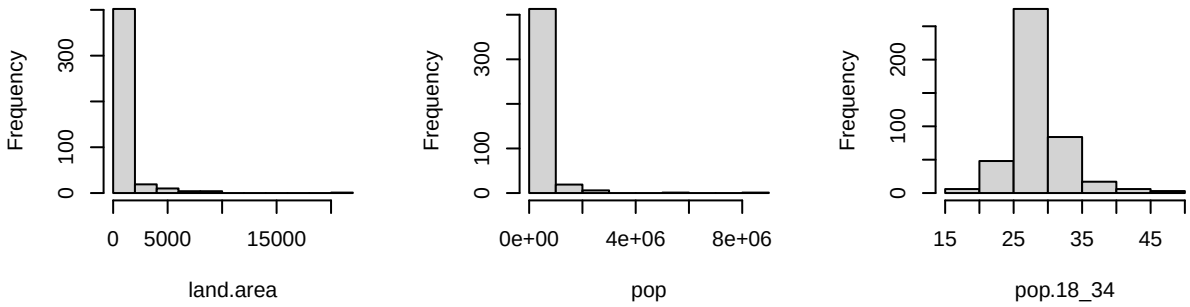
```
table(cdi$region)
```

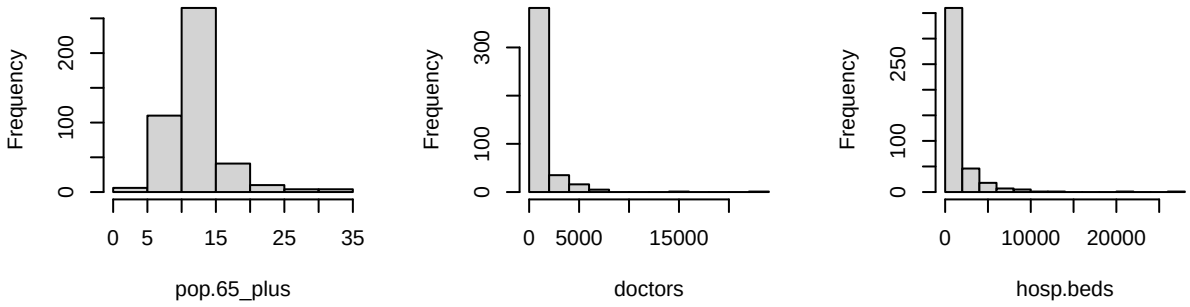
```
##
##  NC  NE   S   W
## 108 103 152  77
```

We have a fairly large number of observations for each region, so this variable may be useful.

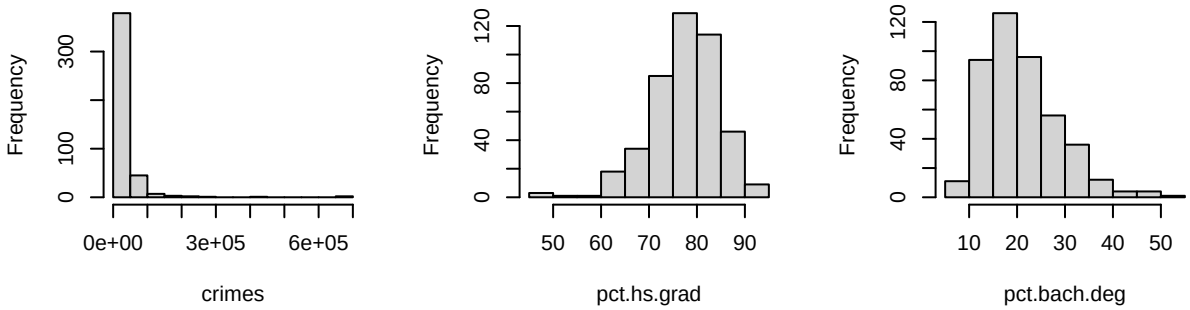
Histograms of untransformed quantitative variables

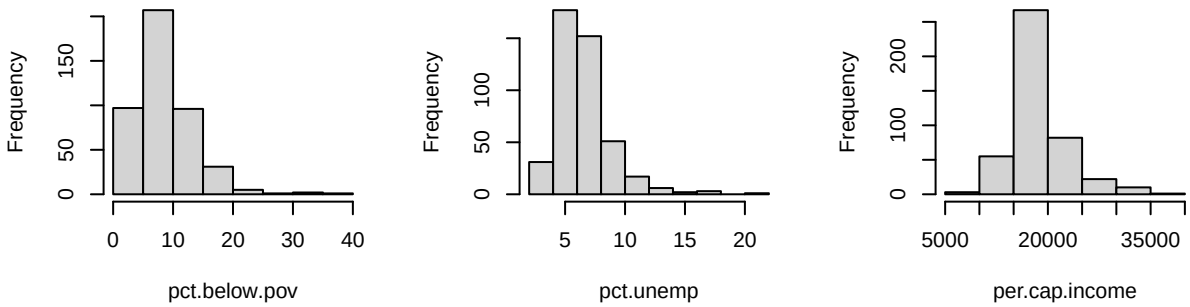
```
par(mfrow=c(2,3))
hist(cdi$land.area,main=NULL,xlab="land.area")
hist(cdi$pop,main=NULL,xlab="pop")
hist(cdi$pop.18_34,main=NULL,xlab="pop.18_34")
hist(cdi$pop.65_plus,main=NULL,xlab="pop.65_plus")
hist(cdi$doctors,main=NULL,xlab="doctors")
hist(cdi$hosp.beds,main=NULL,xlab="hosp.beds")
```



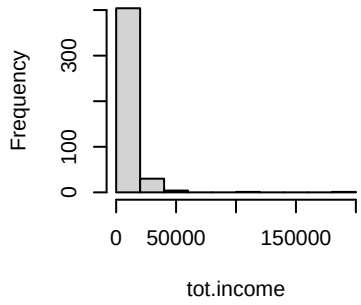



```
hist(cdi$crimes,main=NULL,xlab="crimes")
hist(cdi$pct.hs.grad,main=NULL,xlab="pct.hs.grad")
hist(cdi$pct.bach.deg,main=NULL,xlab="pct.bach.deg")
hist(cdi$pct.below.pov,main=NULL,xlab="pct.below.pov")
hist(cdi$pct.unemp,main=NULL,xlab="pct.unemp")
hist(cdi$per.cap.income,main=NULL,xlab="per.cap.income")
```





```
hist(cdi$tot.income,main=NULL,xlab="tot.income")
```



We see that several variables have extremely right-skewed distributions, so we may need to consider transformations later on.

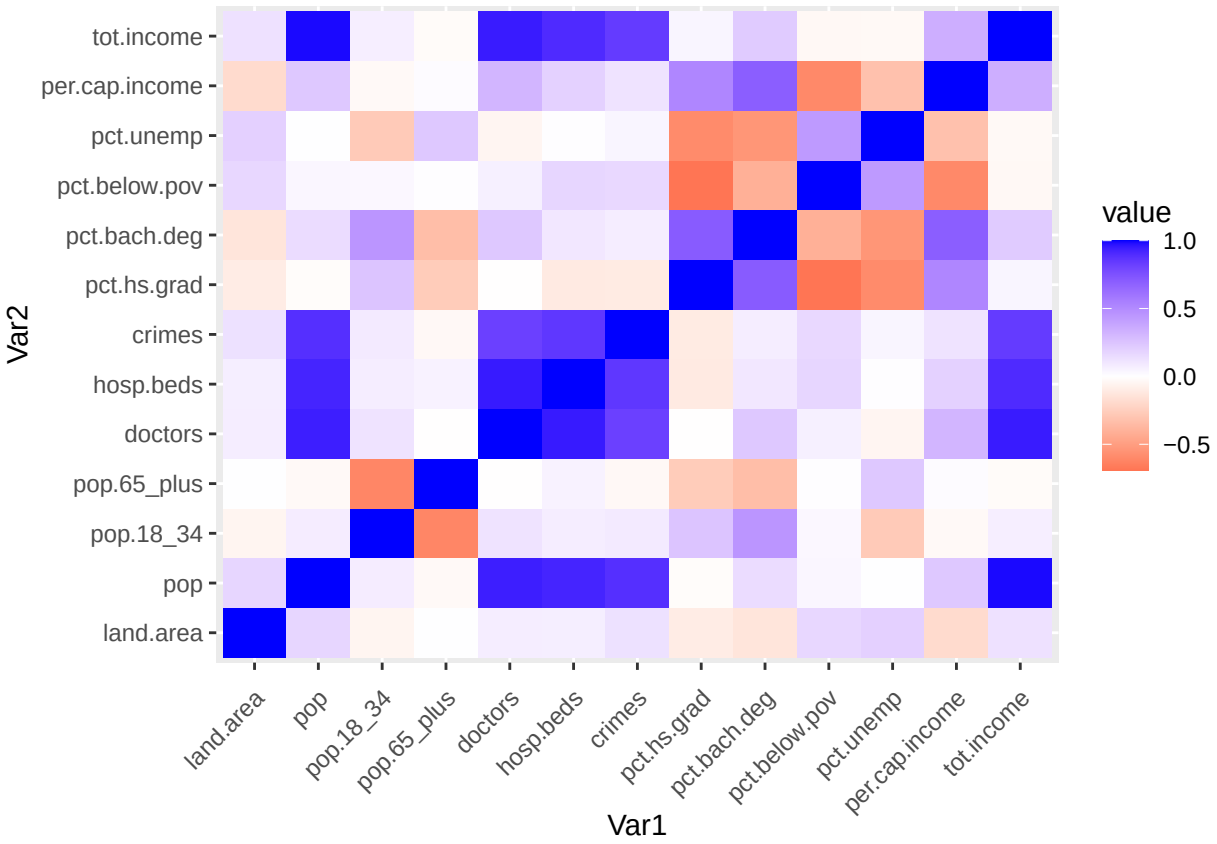
Heatmap of correlation matrix

```
library(reshape2)
```

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

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

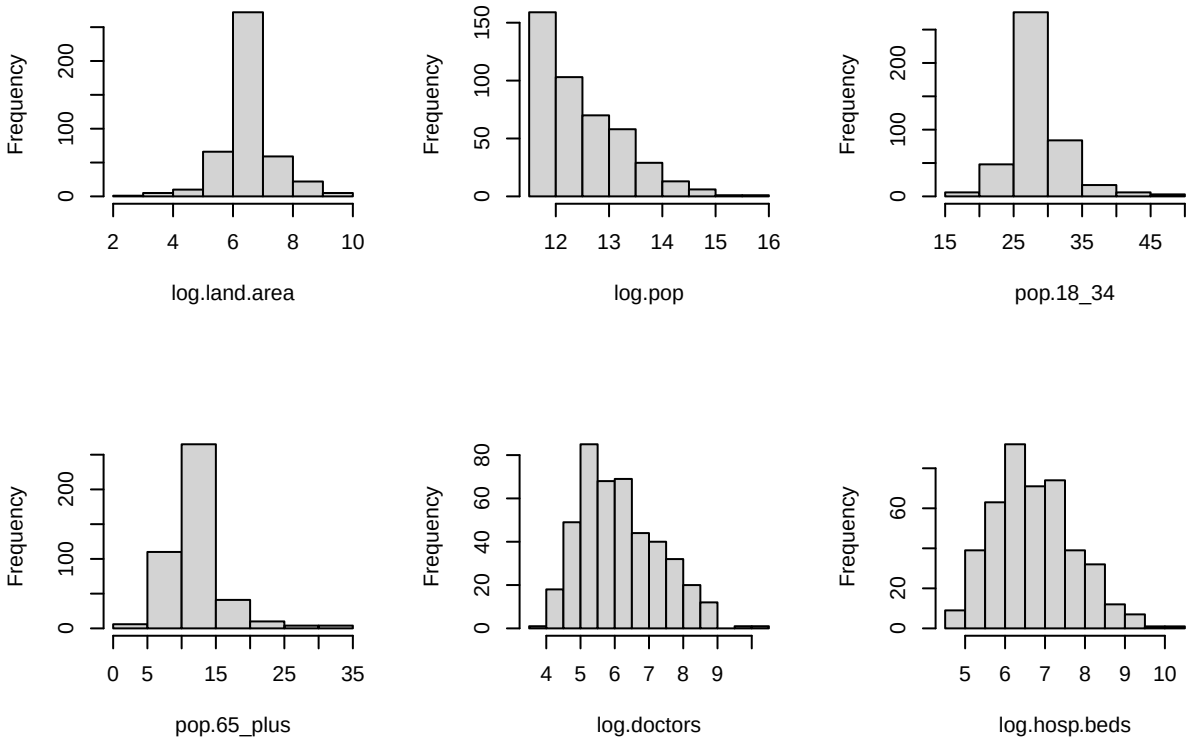
```
cdinumeric <- cdi[,-c(1,2,3,17)]
corgraph <-function(df) {
  cormat <- cor(df)
  melted_cormat <- melt(cormat)
  ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
    geom_tile() +
    theme(axis.text.x = element_text(angle = 45,vjust=0.9,hjust=1)) +
    scale_fill_gradient2(low="red",mid="white",high="blue")
  }

corgraph(cdinumeric)
```



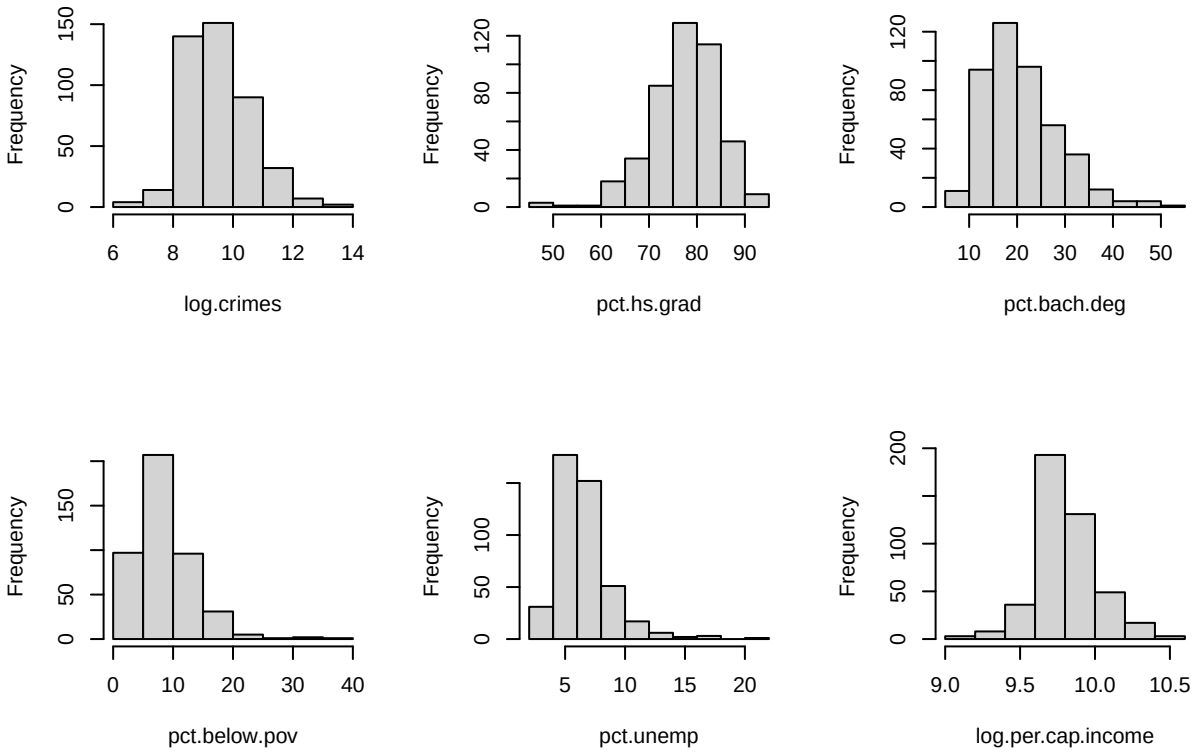
The variables tot.income, pop, doctors, hosp.beds and crimes are strongly correlated with eachother. The response variable, per.cap.income, is not strongly correlated with any of the variables.

Histograms of transformed quantitative variables

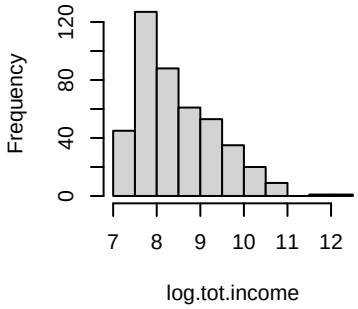
```
cdi.updated <- cdi
par(mfrow=c(2,3))
hist(log(cdi$land.area),main=NULL,xlab="log.land.area")
cdi.updated$land.area <- log(cdi$land.area)

hist(log(cdi$pop),main=NULL,xlab="log.pop")
cdi.updated$pop <- log(cdi$pop)

hist(cdi$pop.18_34,main=NULL,xlab="pop.18_34")

hist(cdi$pop.65_plus,main=NULL,xlab="pop.65_plus")

hist(log(cdi$doctors),main=NULL,xlab="log.doctors")
cdi.updated$doctors <- log(cdi$doctors)

hist(log(cdi$hosp.beds),main=NULL,xlab="log.hosp.beds")
```




```
cdi.updated$hosp.beds <- log(cdi$hosp.beds)

hist(log(cdi$crimes),main=NULL,xlab="log.crimes")
cdi.updated$crimes <- log(cdi$crimes)

hist(cdi$pct.hs.grad,main=NULL,xlab="pct.hs.grad")

hist(cdi$pct.bach.deg,main=NULL,xlab="pct.bach.deg")

hist(cdi$pct.below.pov,main=NULL,xlab="pct.below.pov")

hist(cdi$pct.unemp,main=NULL,xlab="pct.unemp")

hist(log(cdi$per.cap.income),main=NULL,xlab="log.per.cap.income")
```



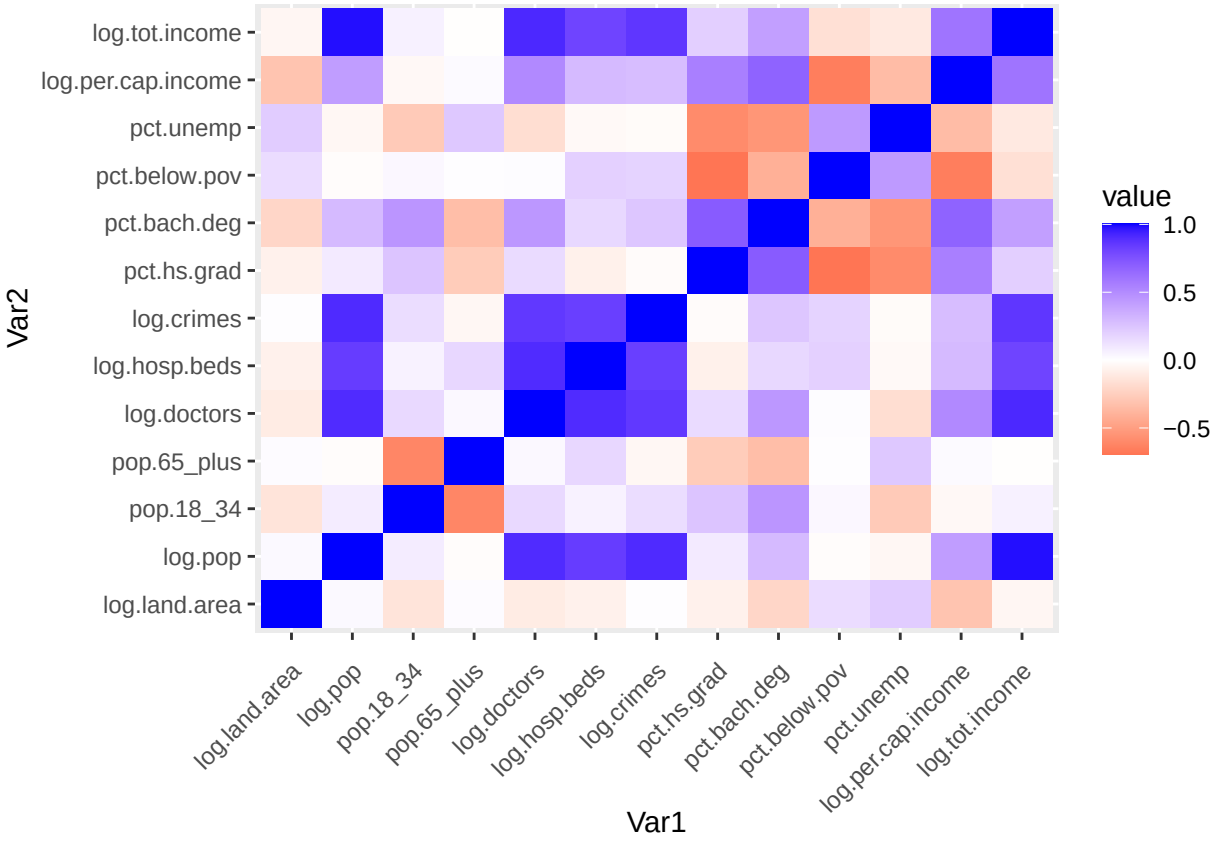

```
cdi.updated$per.cap.income <- log(cdi.updated$per.cap.income)

hist(log(cdi$tot.income),main=NULL,xlab="log.tot.income")
cdi.updated$tot.income <- log(cdi$tot.income)

cdi.updated <- cdi.updated[,-c(18,19,20,21)] %>%
  rename(log.pop = pop,
         log.land.area = land.area,
         log.doctors = doctors,
         log.hosp.beds = hosp.beds,
         log.crimes = crimes,
         log.tot.income = tot.income,
         log.per.cap.income = per.cap.income
         )
```



Applying log-transformations to the variables land.area, pop, doctors, hosp.beds, crimes, per cap.income and tot.income cause all variables to be somewhat normally distributed, but not perfectly. However, applying many different kinds of transformations to many variables will result in a model which is very difficult to understand, so we will leave it at that.

Heatmap of correlation matrix after transformations applied

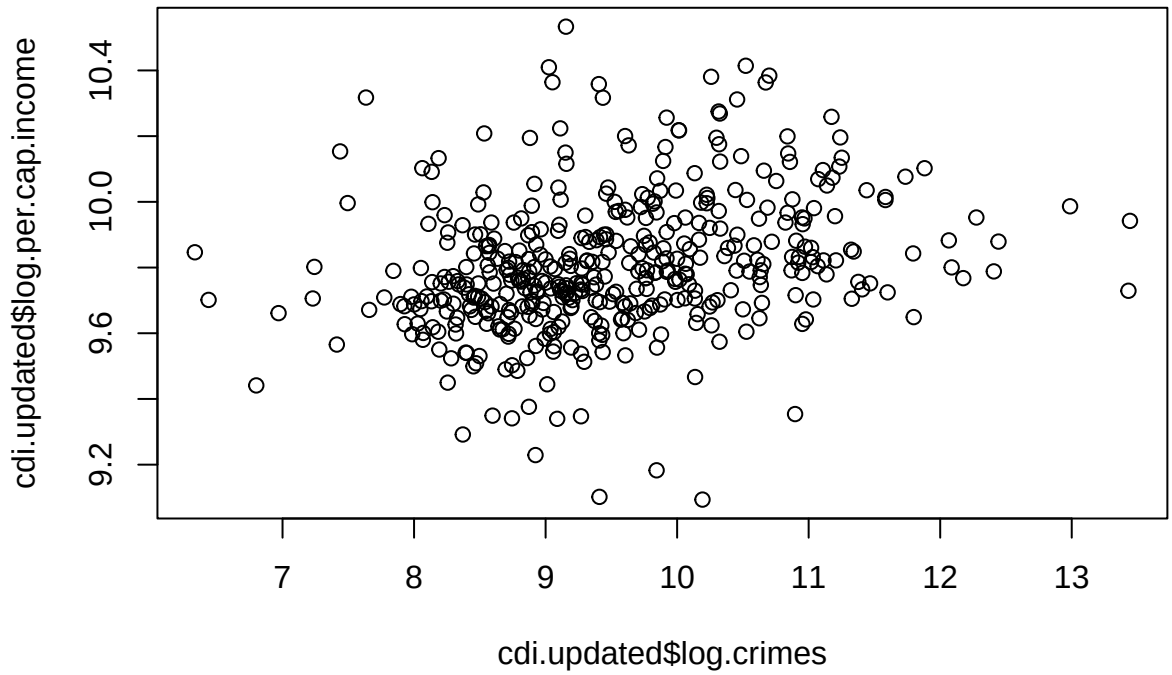
```
cdi.updatednumeric <- cdi.updated[,-c(1,2,3,17)]
corgraph(cdi.updatednumeric)
```



After applying the log-transformations, the relationships between the variables appear to be the same as before.

Relationship between log(crimes) and log(per.cap.income)

```
plot(cdi.updated$log.crimes,cdi.updated$log.per.cap.income)
```



We see a slightly positive linear relationship between log(per.cap.income) and log(crimes).

Find model to predict log(per.cap.income) from log(crimes)

```
linmod0 <- lm(log.per.cap.income~log.crimes,data=cdi.updated)
linmod1 <- lm(log.per.cap.income~log.crimes+region,data=cdi.updated)
linmod2 <- lm(log.per.cap.income~log.crimes*region,data=cdi.updated)
summary(linmod0)
```

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

```
summary(linmod1)
```

```
##
## Call:
```

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

```
summary(linmod2)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.crimes * region, data = cdi.updated)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68552 -0.10418 -0.01444  0.08302  0.79755
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.33677    0.14579  64.044  < 2e-16 ***
## log.crimes            0.05064    0.01566   3.233  0.00132 **
## regionNE             -0.18407    0.21515  -0.856  0.39272
## regionS              -0.19717    0.21211  -0.930  0.35312
## regionW              -0.31439    0.24465  -1.285  0.19947
## log.crimes:regionNE   0.03122    0.02311   1.351  0.17749
## log.crimes:regionS    0.01211    0.02228   0.544  0.58696
## log.crimes:regionW    0.02727    0.02523   1.081  0.28028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1855 on 432 degrees of freedom
## Multiple R-squared:  0.2073, Adjusted R-squared:  0.1945
## F-statistic: 16.14 on 7 and 432 DF,  p-value: < 2.2e-16
```

```
anova(linmod0,linmod1)
```

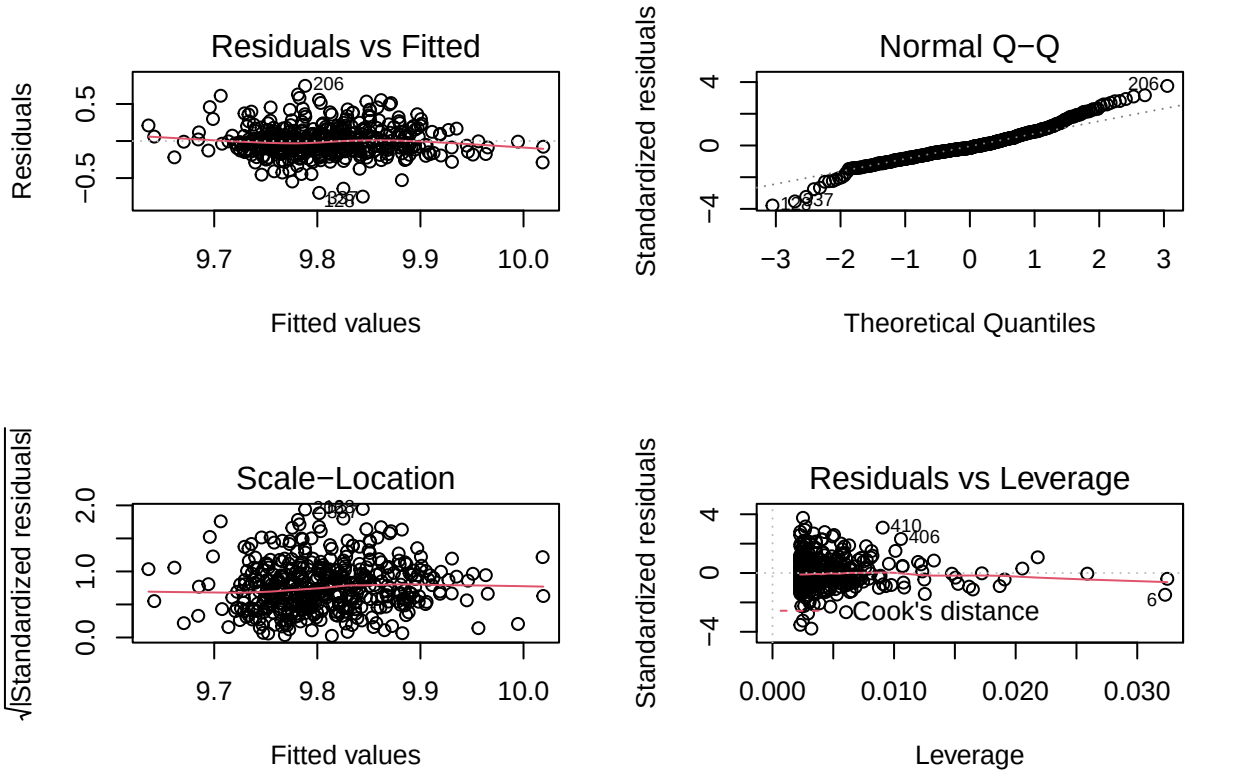
```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.crimes
## Model 2: log.per.cap.income ~ log.crimes + region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    438 17.271
```

```
## 2    435 14.949  3    2.3219 22.522 1.427e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

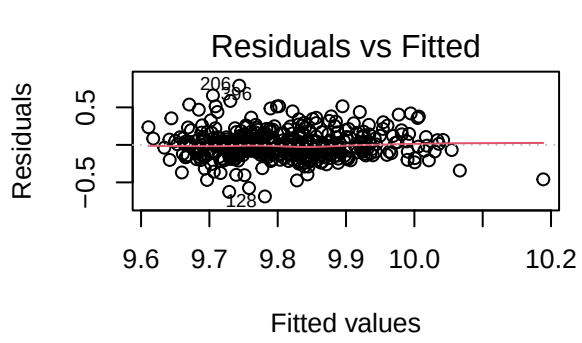
```
anova(linmod1,linmod2)
```

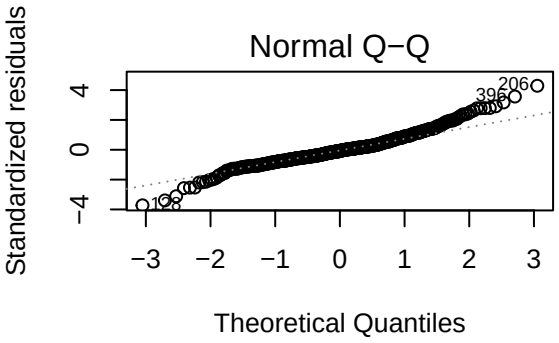
```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.crimes + region
## Model 2: log.per.cap.income ~ log.crimes * region
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    435 14.949
## 2    432 14.872  3  0.076778 0.7434 0.5266
```

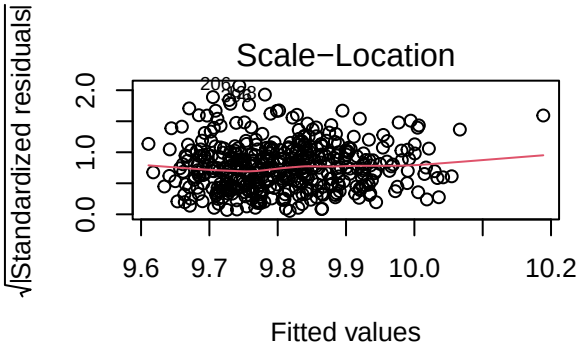
```
par(mfrow=c(2,2))
plot(linmod0)
```

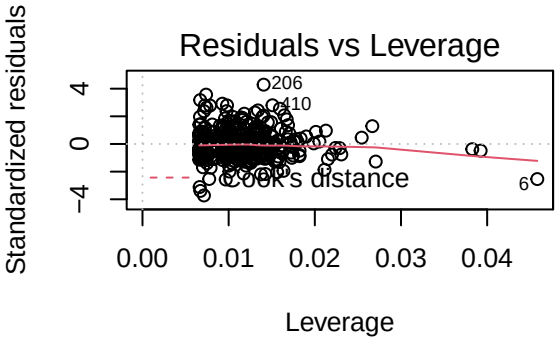



```
plot(linmod1)
```

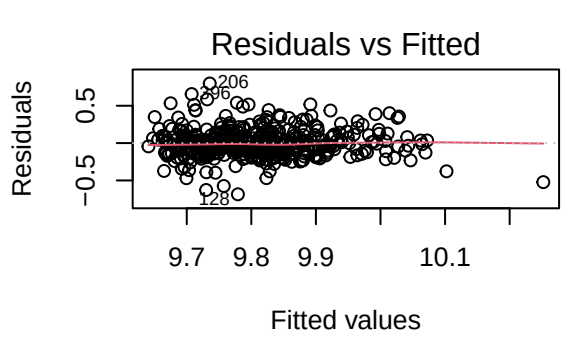


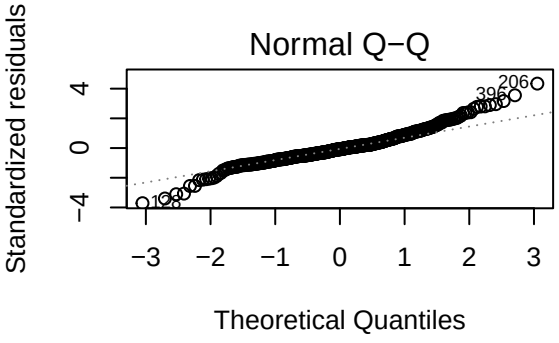


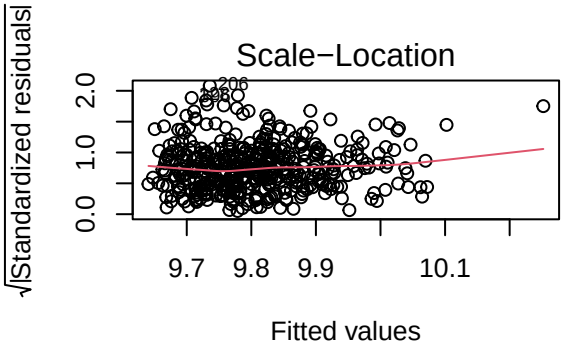


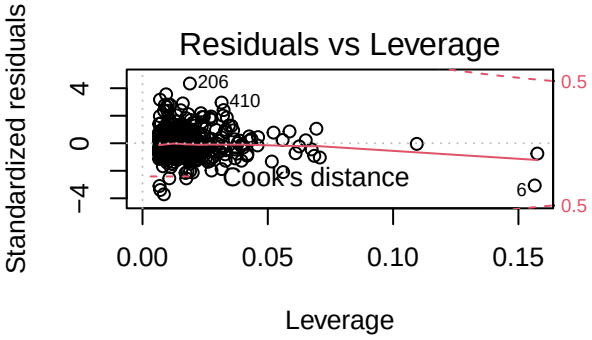



```
plot(linmod2)
```









Model (1): Model is statistically significant. All coefficients are statistically significant. The value of R-squared is small (0.07738). Model appears to be a fairly valid model. Residuals are fairly normally distributed with mean zero and constant variance and we do not see any concerning outliers.

Model (2): Model is statistically significant. All coefficients are statistically significant (except many regionW). The value of R-squared is somewhat small (0.1959). Model appears to be a fairly valid model. Residuals are fairly normally distributed with mean zero and constant variance and we do not see any concerning outliers.

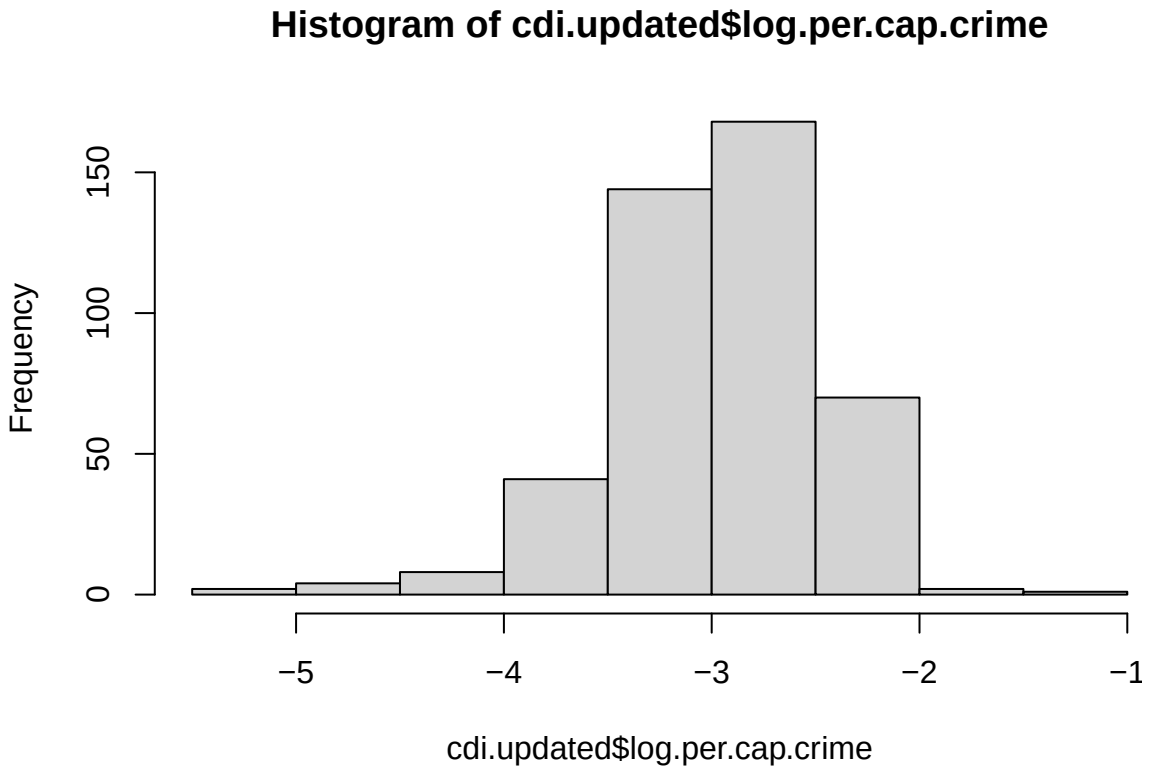
Model (3): Model is statistically significant. Some coefficients are statistically significant, but most are not. The value of R-squared is somewhat small (0.1945). Model appears to be a fairly valid model. Residuals are fairly normally distributed with mean zero and constant variance and we do not see any concerning outliers.

ANOVA test between models (1) and (2): We have significant evidence in favor of model (2)

ANOVA test between models (2) and (3): We do not have significant evidence in favor of model (3)

Find model to predict log(per.cap.income) from log(crime rate)

```
cdi.updated$log.per.cap.crime <- log((cdi$crimes)/(cdi$pop))
hist(cdi.updated$log.per.cap.crime)
```




```
linmod3 <- lm(log.per.cap.income~log.per.cap.crime,data=cdi.updated)
linmod4 <- lm(log.per.cap.income~log.per.cap.crime+region,data=cdi.updated)
linmod5 <- lm(log.per.cap.income~log.per.cap.crime*region,data=cdi.updated)
summary(linmod3)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.per.cap.crime, data = cdi.updated)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7058 -0.1242 -0.0221  0.1066  0.7210
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.73510    0.05908 164.765   <2e-16 ***
## log.per.cap.crime -0.02417    0.01959  -1.233    0.218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2066 on 438 degrees of freedom
## Multiple R-squared:  0.003461,   Adjusted R-squared:  0.001186
## F-statistic: 1.521 on 1 and 438 DF,  p-value: 0.2181
```

```
summary(linmod4)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.per.cap.crime + region,
##     data = cdi.updated)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65832 -0.11431 -0.01548  0.10838  0.75657
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        9.93628    0.06934 143.303  < 2e-16 ***
## log.per.cap.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
```

```
summary(linmod5)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.per.cap.crime * region,
##     data = cdi.updated)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65410 -0.11829 -0.01708  0.10399  0.76628
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  9.91177    0.10503  94.367   <2e-16 ***
## log.per.cap.crime            0.03454    0.03327   1.038    0.300
## regionNE                     0.21007    0.17165   1.224    0.222
## regionS                     -0.10137    0.16072  -0.631    0.529
## regionW                      0.07689    0.26753   0.287    0.774
## log.per.cap.crime:regionNE   0.02924    0.05232   0.559    0.577
```

```
## log.per.cap.crime:regionS  -0.01104    0.05554  -0.199    0.843
## log.per.cap.crime:regionW   0.03495    0.09268   0.377    0.706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.198 on 432 degrees of freedom
## Multiple R-squared:  0.09773,    Adjusted R-squared:  0.08311
## F-statistic: 6.685 on 7 and 432 DF,  p-value: 1.575e-07
```

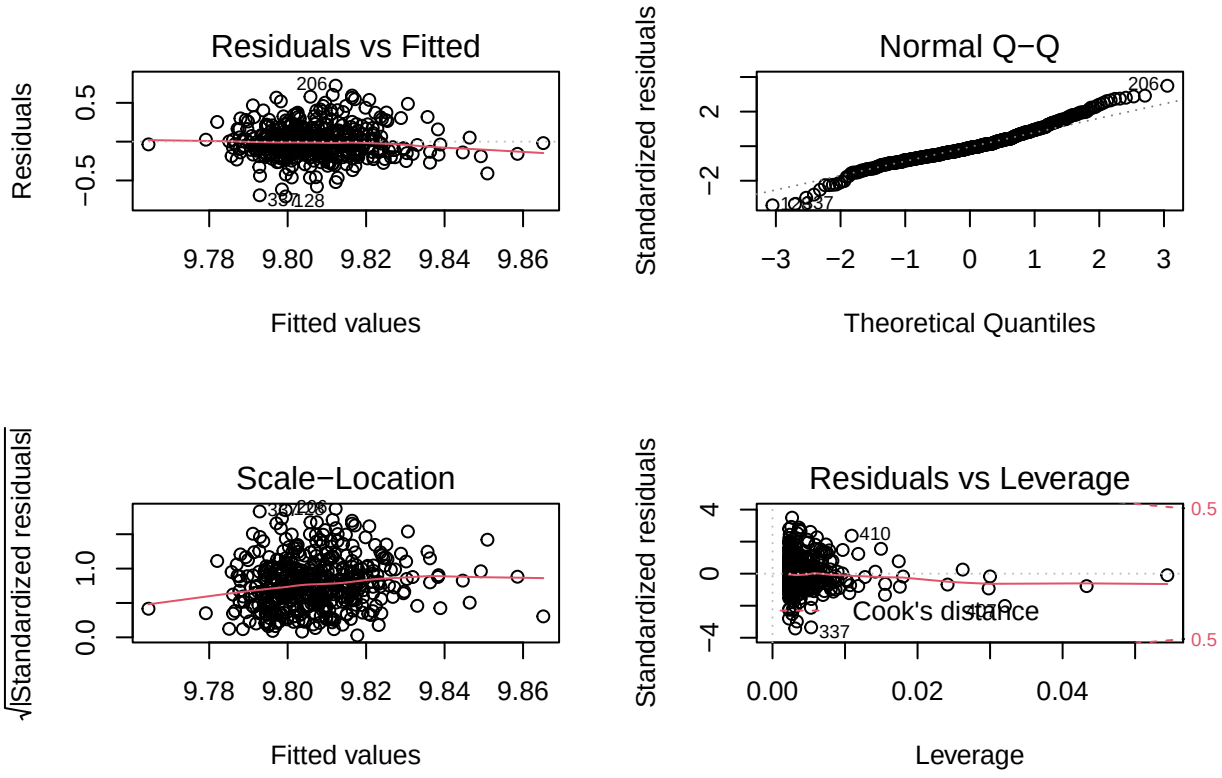
```
anova(linmod3,linmod4)
```

```
## 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
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    438 18.697
## 2    435 16.952  3    1.7447 14.923 2.907e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

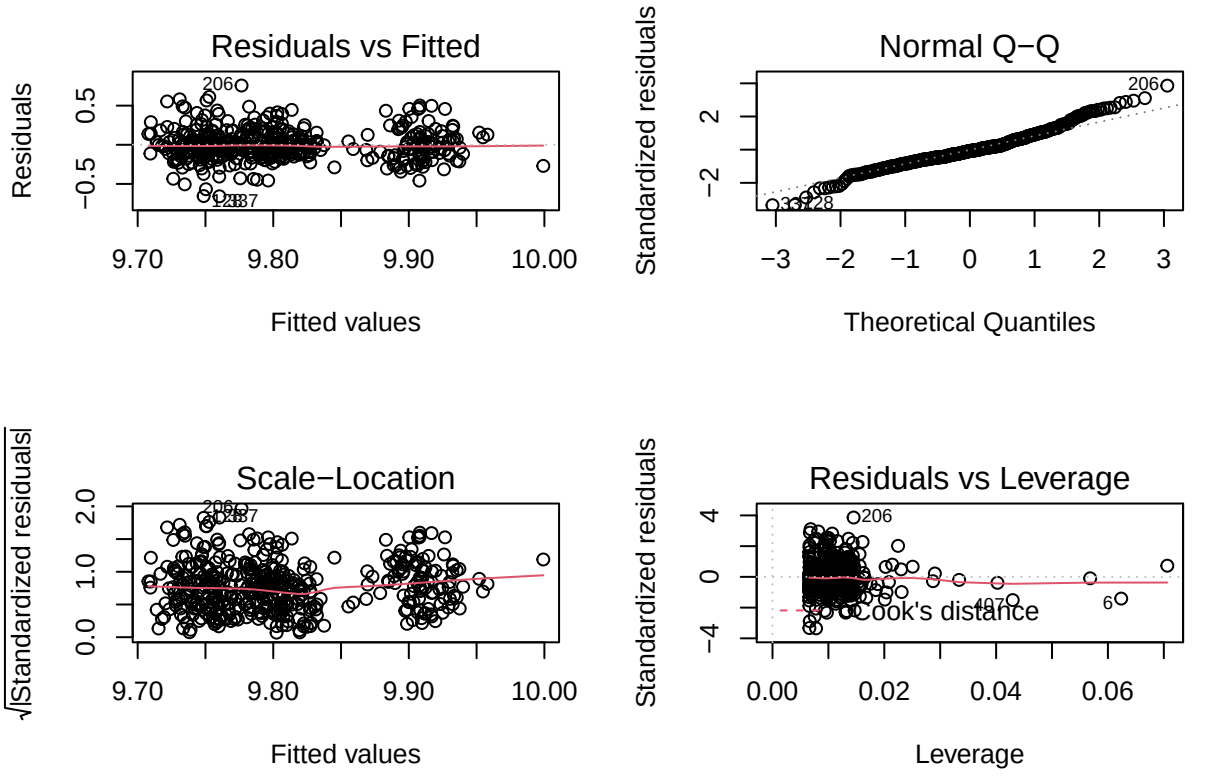
```
anova(linmod4,linmod5)
```

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

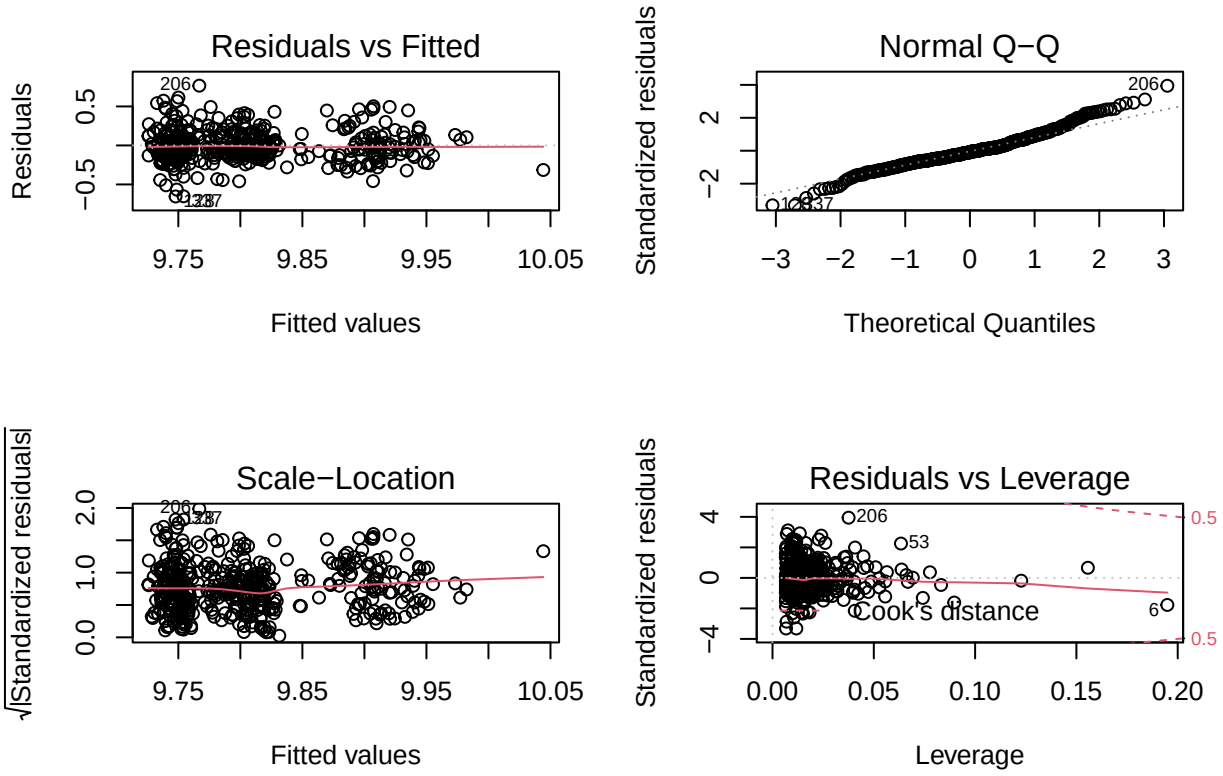
```
par(mfrow=c(2,2))
plot(linmod3)
```




```
plot(linmod4)
```




```
plot(linmod5)
```




```
cdi.updated$log.per.cap.crime <- NULL
```

Model (4): Model is not statistically significant. Coefficient of log(per.cap.crime) is not statistically significant. The value of R-squared is very small (0.001186). Model appears to be a fairly valid model. Residuals are fairly normally distributed with mean zero and constant variance and we do not see any concerning outliers.

Model (5): Model is statistically significant. All coefficients except for regionW are statistically significant. The value of R-squared is small (0.08814). Model appears to be a fairly valid model. Residuals are fairly normally distributed with mean zero and constant variance and we do not see any concerning outliers. However, there is a concerning clustering trend in the residuals.

Model (6): Model is statistically significant. Some coefficients are statistically significant, but most are not. The value of R-squared is small (0.08311). Model appears to be a fairly valid model. Residuals are fairly normally distributed with mean zero and constant variance and we do not see any concerning outliers. However, there is a concerning clustering trend in the residuals.

ANOVA test between models (4) and (5): We have significant evidence in favor of model (5)

ANOVA test between models (5) and (6): We do not have significant evidence in favor of model (6)

All subsets method on all variables except id, county, state, log.pop, log.tot.income, and region

```
library(leaps)
library(car)
cdi.good <- cdi.updated[,-c(1,2,3,5,16,17)]
all.subsets <- regsubsets(log.per.cap.income~.,cdi.good,nvmax=14)
s <- summary(all.subsets)
d <- data.frame(s$rss,s$bic)
d
```

```
##        s.rss     s.bic
## 1  10.164280 -257.5260
## 2   5.745665 -502.4302
## 3   4.831903 -572.5538
## 4   3.708866 -682.8532
## 5   3.269918 -732.1894
## 6   3.016538 -761.5908
## 7   2.905068 -772.0715
## 8   2.874761 -770.5990
## 9   2.863602 -766.2235
## 10  2.861804 -760.4131
```

The minimum BIC occurs when subset size equals 7.

Coefficients when subset size is 7

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

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

The variables in the model chosen by the all subsets method are land.area, pop.18_34, doctors, pct.hs.grad, pct.bach.deg, pct.below.pov, and pct.unemp.

Summary, VIFs and residual plots of model chosen by all subsets method

```
all.subsets.mod <- lm(log.per.cap.income~log.land.area+pop.18_34+log.doctors
                      +pct.hs.grad+pct.bach.deg+pct.below.pov+pct.unemp,data=cdi.good)
summary(all.subsets.mod)
```

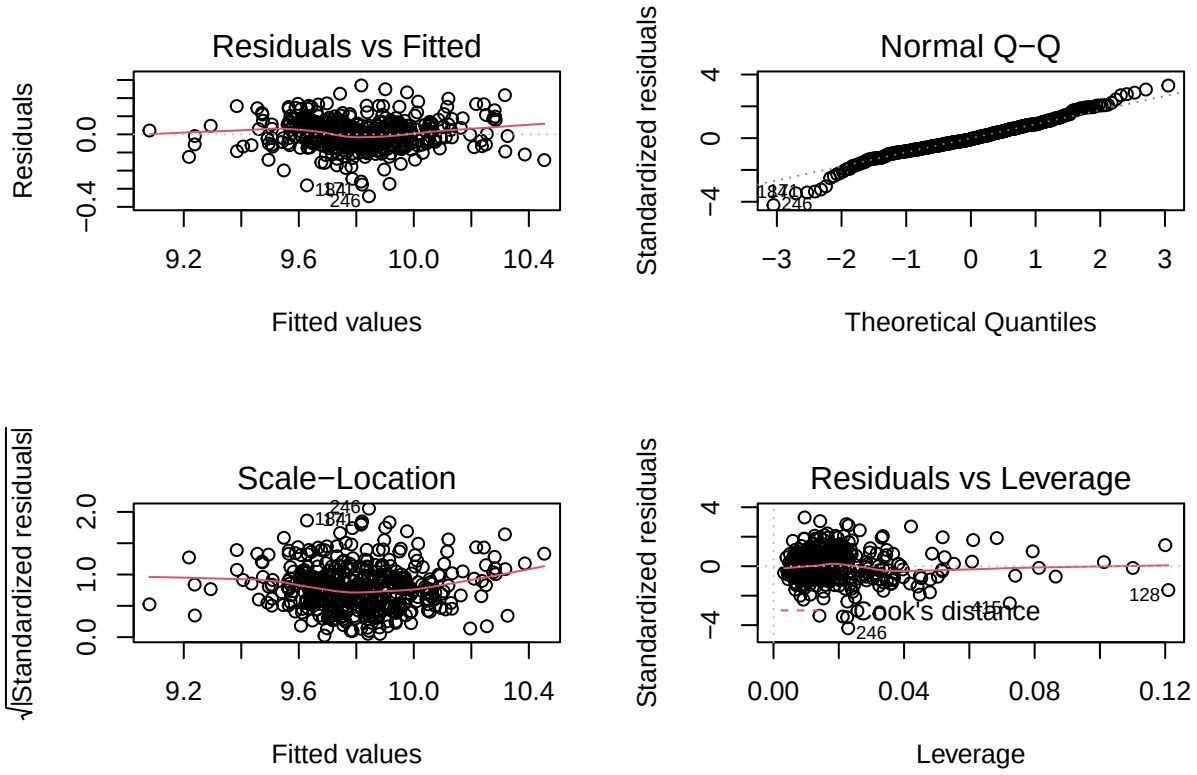
```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp, data = cdi.good)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34147 -0.04886 -0.00538  0.04818  0.26969
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.2224950  0.0931210 109.776  < 2e-16 ***
## log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
## pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
## log.doctors    0.0606769  0.0040183  15.100  < 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 ***
## ---
```

```
## 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
```

```
vif(all.subsets.mod)
```

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

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



Model is statistically significant. All coefficients are statistically significant. The value of R-squared is close to 1 (0.8427). Model appears to be a valid model. Residuals are fairly normally distributed with mean zero and constant variance and we do not see any concerning outliers.

Stepwise AIC and BIC method on all variables except id, county, state, log.pop, log.tot.income, and region

```
stepAIC(lm(log.per.cap.income~.,data=cdi.good),direction="both",k=2)
```

```
## Start:  AIC=-2193.54
## log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
##     log.doctors + log.hosp.beds + log.crimes + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
##
```

```
##                 Df Sum of Sq    RSS     AIC
## - log.crimes     1   0.00180 2.8636 -2195.3
## - log.hosp.beds  1   0.01216 2.8740 -2193.7
## <none>                       2.8618 -2193.5
## - pop.65_plus    1   0.03884 2.9006 -2189.6
## - log.doctors    1   0.11565 2.9775 -2178.1
## - pct.hs.grad    1   0.12699 2.9888 -2176.4
## - pct.unemp      1   0.17289 3.0347 -2169.7
## - log.land.area  1   0.36392 3.2257 -2142.9
## - pop.18_34      1   0.94423 3.8060 -2070.1
## - pct.bach.deg   1   1.56251 4.4243 -2003.8
## - pct.below.pov  1   2.44318 5.3050 -1924.0
##
## Step:  AIC=-2195.27
## log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
##     log.doctors + log.hosp.beds + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## - log.hosp.beds  1   0.01116 2.8748 -2195.6
## <none>                       2.8636 -2195.3
## + log.crimes     1   0.00180 2.8618 -2193.5
## - pop.65_plus    1   0.03709 2.9007 -2191.6
## - pct.hs.grad    1   0.12662 2.9902 -2178.2
## - log.doctors    1   0.12889 2.9925 -2177.9
## - pct.unemp      1   0.17123 3.0348 -2171.7
## - log.land.area  1   0.37492 3.2385 -2143.1
## - pop.18_34      1   0.94270 3.8063 -2072.1
## - pct.bach.deg   1   1.59514 4.4587 -2002.4
## - pct.below.pov  1   2.47345 5.3371 -1923.3
##
## Step:  AIC=-2195.55
## log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       2.8748 -2195.6
## + log.hosp.beds  1   0.01116 2.8636 -2195.3
## + log.crimes     1   0.00079 2.8740 -2193.7
## - pop.65_plus    1   0.03031 2.9051 -2192.9
## - pct.hs.grad    1   0.12309 2.9978 -2179.1
## - pct.unemp      1   0.16432 3.0391 -2173.1
## - log.land.area  1   0.38995 3.2647 -2141.6
## - pop.18_34      1   0.93157 3.8063 -2074.1
## - log.doctors    1   1.55295 4.4277 -2007.5
## - pct.bach.deg   1   1.80755 4.6823 -1982.9
## - pct.below.pov  1   2.53302 5.4078 -1919.5
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     pop.65_plus + log.doctors + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp, data = cdi.good)
```

```
##
## Coefficients:
##   (Intercept)  log.land.area      pop.18_34    pop.65_plus    log.doctors
##     10.315967      -0.036493      -0.015349      -0.002766       0.062605
##   pct.hs.grad   pct.bach.deg  pct.below.pov      pct.unemp
##     -0.004658       0.015215      -0.024614       0.010769
```

```
n <- dim(cdi.updated)[1]
stepAIC(lm(log.per.cap.income~.,data=cdi.good),direction="both",k=log(n))
```

```
## Start:  AIC=-2148.59
## log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
##     log.doctors + log.hosp.beds + log.crimes + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## - log.crimes     1   0.00180 2.8636 -2154.4
## - log.hosp.beds  1   0.01216 2.8740 -2152.8
## - pop.65_plus    1   0.03884 2.9006 -2148.7
## <none>                       2.8618 -2148.6
## - log.doctors    1   0.11565 2.9775 -2137.2
## - pct.hs.grad    1   0.12699 2.9888 -2135.6
## - pct.unemp      1   0.17289 3.0347 -2128.9
## - log.land.area  1   0.36392 3.2257 -2102.0
## - pop.18_34      1   0.94423 3.8060 -2029.2
## - pct.bach.deg   1   1.56251 4.4243 -1963.0
## - pct.below.pov  1   2.44318 5.3050 -1883.1
##
## Step:  AIC=-2154.4
## log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
##     log.doctors + log.hosp.beds + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## - log.hosp.beds  1   0.01116 2.8748 -2158.8
## - pop.65_plus    1   0.03709 2.9007 -2154.8
## <none>                       2.8636 -2154.4
## + log.crimes     1   0.00180 2.8618 -2148.6
## - pct.hs.grad    1   0.12662 2.9902 -2141.4
## - log.doctors    1   0.12889 2.9925 -2141.1
## - pct.unemp      1   0.17123 3.0348 -2134.9
## - log.land.area  1   0.37492 3.2385 -2106.3
## - pop.18_34      1   0.94270 3.8063 -2035.3
## - pct.bach.deg   1   1.59514 4.4587 -1965.7
## - pct.below.pov  1   2.47345 5.3371 -1886.5
##
## Step:  AIC=-2158.77
## log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## - pop.65_plus    1   0.03031 2.9051 -2160.2
## <none>                       2.8748 -2158.8
## + log.hosp.beds  1   0.01116 2.8636 -2154.4
```

```
## + log.crimes     1   0.00079 2.8740 -2152.8
## - pct.hs.grad    1   0.12309 2.9978 -2146.4
## - pct.unemp      1   0.16432 3.0391 -2140.4
## - log.land.area  1   0.38995 3.2647 -2108.9
## - pop.18_34      1   0.93157 3.8063 -2041.3
## - log.doctors    1   1.55295 4.4277 -1974.8
## - pct.bach.deg   1   1.80755 4.6823 -1950.2
## - pct.below.pov  1   2.53302 5.4078 -1886.8
##
## Step:  AIC=-2160.25
## log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       2.9051 -2160.2
## + pop.65_plus    1   0.03031 2.8748 -2158.8
## + log.hosp.beds  1   0.00438 2.9007 -2154.8
## + log.crimes     1   0.00014 2.9049 -2154.2
## - pct.hs.grad    1   0.11147 3.0165 -2149.8
## - pct.unemp      1   0.15952 3.0646 -2142.8
## - log.land.area  1   0.37507 3.2801 -2112.9
## - pop.18_34      1   1.05209 3.9572 -2030.3
## - log.doctors    1   1.53330 4.4384 -1979.8
## - pct.bach.deg   1   1.86219 4.7673 -1948.4
## - pct.below.pov  1   2.50333 5.4084 -1892.9
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp, data = cdi.good)
##
## Coefficients:
##   (Intercept)  log.land.area      pop.18_34    log.doctors    pct.hs.grad
##     10.222495      -0.035674      -0.013900       0.060677      -0.004406
##  pct.bach.deg  pct.below.pov      pct.unemp
##      0.015385      -0.024278       0.010604
```

Note that the model produced by stepwise regression with AIC is the same as the one produced by the all subsets method but with pct.65_plus added. The model produced by stepwise regression with BIC is identical to the model produced by the all subsets method.

Summary, VIF's and residual plots for model chosen by stepwise AIC method

```
stepAIC.mod <- lm(log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus
                  + log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov
                  + pct.unemp, data = cdi.good)
summary(stepAIC.mod)
```

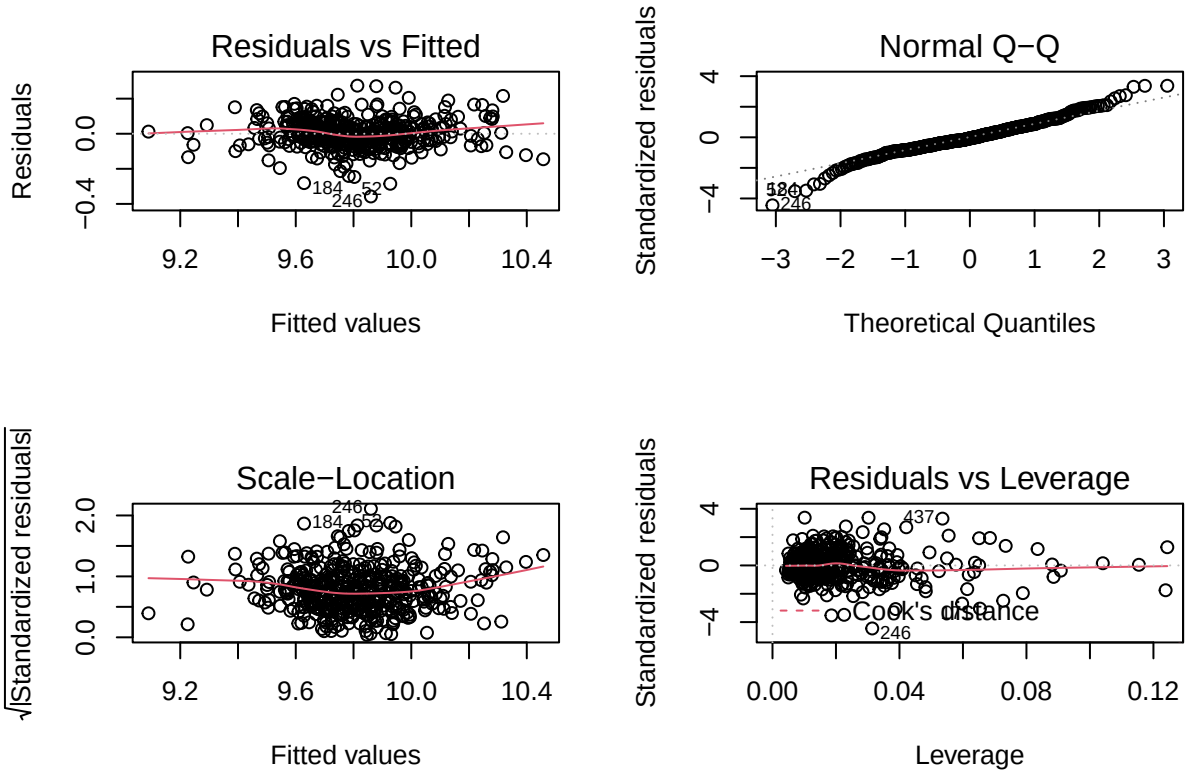
```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     pop.65_plus + log.doctors + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp, data = cdi.good)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35756 -0.04551 -0.00543  0.04844  0.27399
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.3159666  0.1025858 100.559  < 2e-16 ***
## log.land.area -0.0364935  0.0047728  -7.646 1.36e-13 ***
## pop.18_34     -0.0153488  0.0012988 -11.818  < 2e-16 ***
## pop.65_plus   -0.0027664  0.0012978  -2.132   0.0336 *
## log.doctors    0.0626053  0.0041029  15.259  < 2e-16 ***
## 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 ***
## ---
## 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
```

```
vif(stepAIC.mod)
```

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

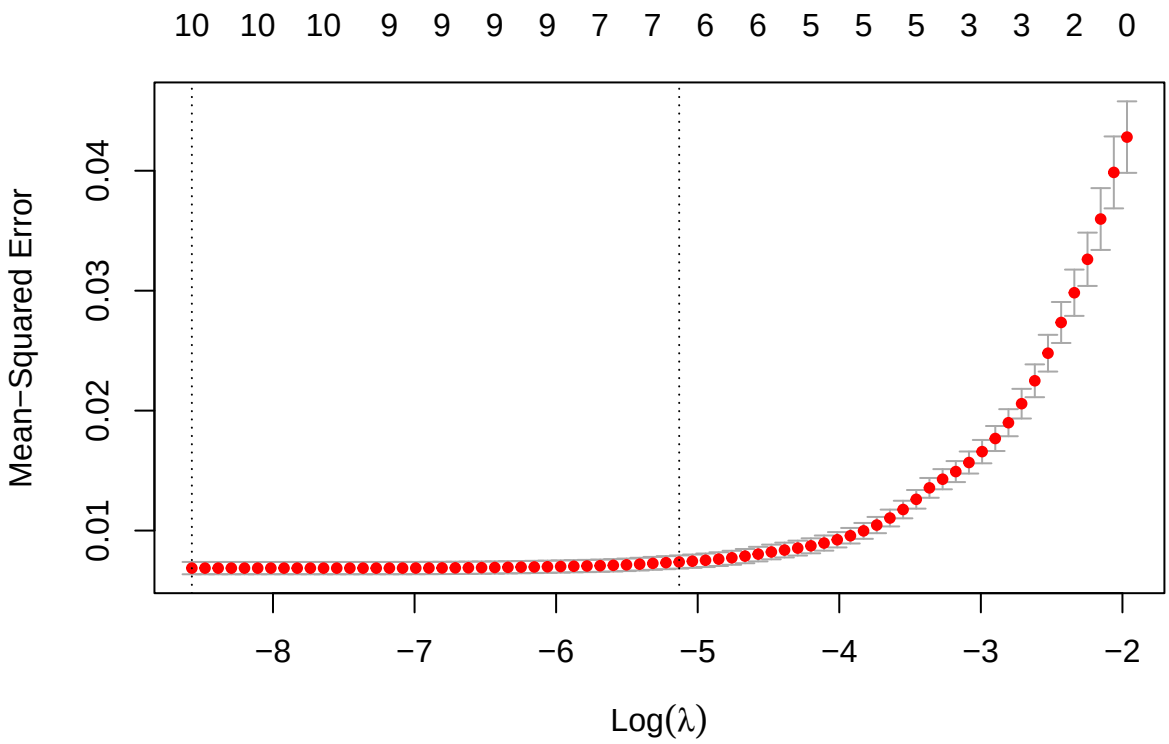
```
par(mfrow=c(2,2))
plot(stepAIC.mod)
```



Model is statistically significant. All coefficients are statistically significant. The value of R-squared is close to 1 (0.8439). Model appears to be a valid model. Residuals are fairly normally distributed with mean zero and constant variance and we do not see any concerning outliers. However, this model appears to be a fairly equally good fit as the model chosen by all subsets, so we will choose the simplest model. In this case, that is the model chosen by all subsets.

Lasso method on all variables except id, county, state, log.pop, log.tot.income, and region

```
result <- cv.glmnet(as.matrix(cdi.good[,-c(11)]),cdi.good[,11])
plot(result)
```




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

```
##   lambda.1se   lambda.min
## 0.0059119094 0.0001891378
```

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

```
## 11 x 3 sparse Matrix of class "dgCMatrix"
##                          s1           s1           s1
## (Intercept)    9.880596947  9.880596947 10.288836619
## log.land.area -0.032786921 -0.032786921 -0.035715268
## pop.18_34     -0.012013486 -0.012013486 -0.015364277
## pop.65_plus    .            .           -0.003027088
## log.doctors    0.059553442  0.059553442  0.051964486
## log.hosp.beds  .            .           0.014479819
## log.crimes     .            .          -0.002419004
## pct.hs.grad    .            .          -0.004537077
## pct.bach.deg   0.011803004  0.011803004  0.015567203
## pct.below.pov -0.020058121 -0.020058121 -0.024753098
## pct.unemp      0.006517379  0.006517379  0.010950352
```

Seems to be the same as the model chosen by all subsets again.

Now we look at all models above but including interactions with region

First, all subsets

```
cdi.good$region <- cdi.updated$region
all.subsets.mod.tmp <- lm(log.per.cap.income~(log.land.area
                                +pop.18_34+log.doctors+pct.hs.grad+pct.bach.deg
```

```
                                  +pct.below.pov+pct.unemp)*region,data=cdi.good)
summary(all.subsets.mod.tmp)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ (log.land.area + pop.18_34 +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp) * region, data = cdi.good)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.250782 -0.042332 -0.002298  0.040559  0.313570
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              10.1244260  0.2826240  35.823  < 2e-16 ***
## log.land.area            -0.0364187  0.0151355  -2.406 0.016564 *
## pop.18_34                -0.0147940  0.0026043  -5.681 2.55e-08 ***
## log.doctors               0.0544169  0.0093221   5.837 1.08e-08 ***
## pct.hs.grad              -0.0024773  0.0034110  -0.726 0.468088
## pct.bach.deg              0.0140833  0.0029254   4.814 2.09e-06 ***
## pct.below.pov            -0.0237085  0.0036234  -6.543 1.81e-10 ***
## pct.unemp                 0.0180393  0.0048923   3.687 0.000257 ***
## regionNE                  0.3243992  0.3577081   0.907 0.365004
## regionS                  -0.0345856  0.3131668  -0.110 0.912116
## regionW                   1.5043946  0.4226868   3.559 0.000416 ***
## log.land.area:regionNE   -0.0037179  0.0201435  -0.185 0.853656
## log.land.area:regionS    -0.0047582  0.0174155  -0.273 0.784825
## log.land.area:regionW     0.0151234  0.0181871   0.832 0.406154
## pop.18_34:regionNE       -0.0024780  0.0036873  -0.672 0.501939
## pop.18_34:regionS        -0.0008777  0.0030680  -0.286 0.774970
## pop.18_34:regionW         0.0014122  0.0040925   0.345 0.730220
## log.doctors:regionNE     -0.0046251  0.0132571  -0.349 0.727359
## log.doctors:regionS       0.0043337  0.0114401   0.379 0.705019
## log.doctors:regionW      -0.0034863  0.0131576  -0.265 0.791173
## pct.hs.grad:regionNE     -0.0037529  0.0044150  -0.850 0.395813
## pct.hs.grad:regionS       0.0021198  0.0037853   0.560 0.575790
## pct.hs.grad:regionW      -0.0190188  0.0045881  -4.145 4.13e-05 ***
## pct.bach.deg:regionNE     0.0069429  0.0040312   1.722 0.085776 .
## pct.bach.deg:regionS     -0.0015774  0.0032000  -0.493 0.622328
## pct.bach.deg:regionW      0.0071026  0.0036374   1.953 0.051541 .
## pct.below.pov:regionNE   -0.0014134  0.0050896  -0.278 0.781381
## pct.below.pov:regionS     0.0072764  0.0040739   1.786 0.074827 .
## pct.below.pov:regionW    -0.0161639  0.0054271  -2.978 0.003071 **
## pct.unemp:regionNE       -0.0083596  0.0073758  -1.133 0.257720
## pct.unemp:regionS        -0.0249396  0.0065867  -3.786 0.000176 ***
## pct.unemp:regionW        -0.0201466  0.0067713  -2.975 0.003101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0759 on 408 degrees of freedom
## Multiple R-squared:  0.8747, Adjusted R-squared:  0.8652
## F-statistic: 91.91 on 31 and 408 DF,  p-value: < 2.2e-16
```

```
all.subsets.mod.r <- lm(log.per.cap.income~log.land.area+pop.18_34+log.doctors
                      +pct.hs.grad+pct.bach.deg+pct.below.pov+pct.unemp+region
                      +pct.hs.grad*region+pct.below.pov*region+pct.unemp*region,
                      data=cdi.good)
summary(all.subsets.mod.r)
```

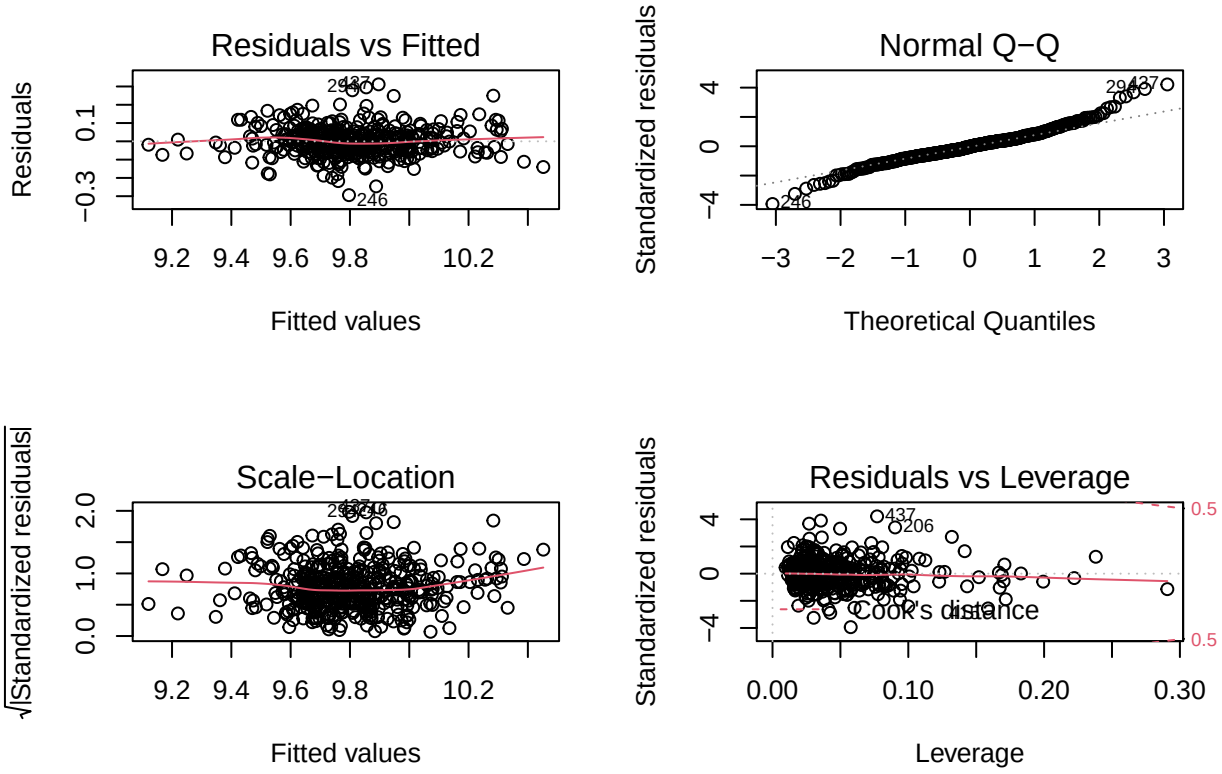
```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp + region + pct.hs.grad * region + pct.below.pov *
##     region + pct.unemp * region, data = cdi.good)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.294186 -0.043597 -0.001583  0.037667  0.311609
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             10.2421239  0.2176557  47.057  < 2e-16 ***
## log.land.area           -0.0381738  0.0053996  -7.070 6.51e-12 ***
## pop.18_34               -0.0149347  0.0010897 -13.705  < 2e-16 ***
## log.doctors              0.0572284  0.0040082  14.278  < 2e-16 ***
## pct.hs.grad             -0.0043532  0.0024515  -1.776 0.076501 .
## pct.bach.deg             0.0156310  0.0009715  16.090  < 2e-16 ***
## pct.below.pov           -0.0252029  0.0032612  -7.728 8.12e-14 ***
## pct.unemp                0.0197400  0.0046254   4.268 2.44e-05 ***
## regionNE                -0.0520070  0.2707173  -0.192 0.847750
## regionS                 -0.0389718  0.2383516  -0.164 0.870199
## regionW                  1.3910484  0.3408962   4.081 5.38e-05 ***
## pct.hs.grad:regionNE     0.0017684  0.0029293   0.604 0.546374
## pct.hs.grad:regionS      0.0011525  0.0025618   0.450 0.653024
## pct.hs.grad:regionW     -0.0141473  0.0035826  -3.949 9.20e-05 ***
## pct.below.pov:regionNE -0.0015170  0.0046143  -0.329 0.742493
## pct.below.pov:regionS    0.0070185  0.0035199   1.994 0.046808 *
## pct.below.pov:regionW   -0.0137920  0.0051811  -2.662 0.008066 **
## pct.unemp:regionNE      -0.0129841  0.0070423  -1.844 0.065929 .
## pct.unemp:regionS       -0.0231138  0.0061365  -3.767 0.000189 ***
## pct.unemp:regionW       -0.0217357  0.0065225  -3.332 0.000937 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07692 on 420 degrees of freedom
## Multiple R-squared:  0.8675, Adjusted R-squared:  0.8615
## F-statistic: 144.8 on 19 and 420 DF,  p-value: < 2.2e-16
```

```
vif(all.subsets.mod.r)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## log.land.area        1.643605e+00  1        1.282032
## pop.18_34            1.547481e+00  1        1.243978
## log.doctors          1.559981e+00  1        1.248992
## pct.hs.grad          2.194177e+01  1        4.684205
## pct.bach.deg         4.102307e+00  1        2.025415
```

```
## pct.below.pov         1.710982e+01  1        4.136402
## pct.unemp             8.675528e+00  1        2.945425
## region                2.454546e+08  3       25.022374
## pct.hs.grad:region    8.506975e+07  3       20.971486
## pct.below.pov:region  5.278685e+03  3        4.172736
## pct.unemp:region      1.108865e+04  3        4.722222
```

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




```
anova(all.subsets.mod,all.subsets.mod.r)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     region + pct.hs.grad * region + pct.below.pov * region +
##     pct.unemp * region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    432 2.9051
## 2    420 2.4853 12   0.41978 5.9117 1.555e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We add region and interactions between region and all predictors in the model. Then, we remove any interactions which are not significant. The resulting model includes region and the interactions between region and pct.hs.grad, pct.below.pov and pct.unemp. This model is statistically significant and most, but not all, coefficients are statistically significant. The value of R-squared is close to 1 (0.8615). Model appears to be a valid model. Residuals are fairly normally distributed with mean zero and constant variance and we

do not see any concerning outliers. ANOVA test indicates significant evidence in favor of the model with region and specified interaction terms included.

Stepwise AIC

```
stepAIC.mod.tmp <- lm(log.per.cap.income ~ (log.land.area + pop.18_34
                                            + pop.65_plus + log.doctors + pct.hs.grad
                                            + pct.bach.deg + pct.below.pov
                                            + pct.unemp)*region, data = cdi.good)
# summary(stepAIC.mod.tmp)
stepAIC.mod.r <- lm(log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus
                    + log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov
                    + pct.unemp + region + pct.hs.grad*region + pct.below.pov*region
                    + pct.unemp*region, data = cdi.good)
summary(stepAIC.mod.r)
```

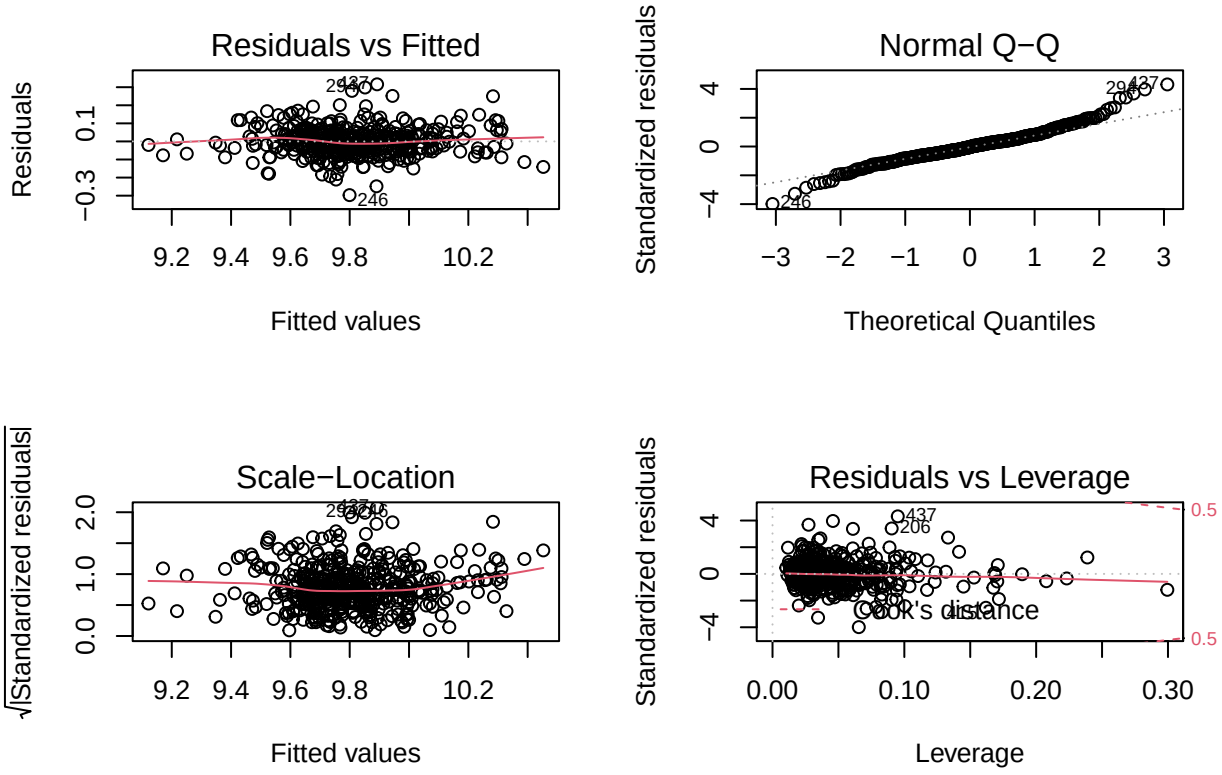
```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     pop.65_plus + log.doctors + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + region + pct.hs.grad * region +
##     pct.below.pov * region + pct.unemp * region, data = cdi.good)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.296995 -0.044232 -0.001996  0.037815  0.315861
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             10.2566413  0.2207090  46.471  < 2e-16 ***
## log.land.area           -0.0381759  0.0054049  -7.063 6.82e-12 ***
## pop.18_34               -0.0152143  0.0012851 -11.839  < 2e-16 ***
## pop.65_plus             -0.0005474  0.0013302  -0.412 0.680911
## log.doctors              0.0575509  0.0040879  14.078  < 2e-16 ***
## pct.hs.grad             -0.0043725  0.0024544  -1.782 0.075554 .
## pct.bach.deg             0.0156026  0.0009749  16.005  < 2e-16 ***
## pct.below.pov           -0.0250040  0.0033000  -7.577 2.29e-13 ***
## pct.unemp                0.0194964  0.0046677   4.177 3.60e-05 ***
## regionNE                -0.0535613  0.2710118  -0.198 0.843427
## regionS                 -0.0377252  0.2386069  -0.158 0.874449
## regionW                  1.3902000  0.3412401   4.074 5.53e-05 ***
## pct.hs.grad:regionNE     0.0017833  0.0029325   0.608 0.543437
## pct.hs.grad:regionS      0.0011333  0.0025647   0.442 0.658796
## pct.hs.grad:regionW     -0.0141407  0.0035861  -3.943 9.42e-05 ***
## pct.below.pov:regionNE -0.0016569  0.0046313  -0.358 0.720702
## pct.below.pov:regionS    0.0066566  0.0036315   1.833 0.067506 .
## pct.below.pov:regionW   -0.0139186  0.0051953  -2.679 0.007674 **
## pct.unemp:regionNE      -0.0126195  0.0071048  -1.776 0.076425 .
## pct.unemp:regionS       -0.0225118  0.0063144  -3.565 0.000406 ***
## pct.unemp:regionW       -0.0215869  0.0065390  -3.301 0.001045 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.077 on 419 degrees of freedom
## Multiple R-squared:  0.8676, Adjusted R-squared:  0.8613
## F-statistic: 137.3 on 20 and 419 DF,  p-value: < 2.2e-16
```

```
vif(stepAIC.mod.r)
```

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

```
par(mfrow=c(2,2))
plot(stepAIC.mod.r)
```




```
anova(stepAIC.mod,stepAIC.mod.r)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
```

```
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp + region + pct.hs.grad * region + pct.below.pov *
##     region + pct.unemp * region
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    431 2.8748
## 2    419 2.4843 12   0.39048 5.4881 1.02e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We add region and interactions between region and all predictors in the model. Then, we remove any interactions which are not significant. The resulting model includes region and the interactions between region and pct.hs.grad, pct.below.pov and pct.unemp. This model is statistically significant and most, but not all, coefficients are statistically significant. The value of R-squared is close to 1 (0.8613). Model appears to be a valid model. Residuals are fairly normally distributed with mean zero and constant variance and we do not see any concerning outliers. ANOVA test indicates significant evidence in favor of the model with region and specified interaction terms included.