Exploring the Relation between Per Capita Income and Various County Demographic Information

Megan Christy
Department of Statistics and Data Science, Carnegie Mellon University
mechrist@andrew.cmu.edu
29 October 2021

Abstract

We address the question of how to best predict per capita income from various county demographic information. We examine data on selected county demographic information (CDI) for 440 of the most populous counties in the United States between 1990 and 1992 from Kutner et al. (2005). We use visualizations, linear regression models, and various variable selection processes to answer the questions set out in the paper. We find that per capita income is best predicted from the percentage of the population aged 18-34, the percentage of high school graduates, the percentage of bachelor degrees, percentage below poverty rate, percentage unemployment, land area, the number of active physicians, the interaction between percentage bachelor degrees and region, the interaction between percentage below poverty level and region, and the interaction between percentage unemployment and region. This indicates that per capita income is impacted by a variety of geographic, health, social, and educational county variables.

1 Introduction

As income inequality continues to be a problem in the United States, social scientists are interested in determining what factors are associated with income. This is important because in order to address the effects of income inequality, we need to know which aspects of a particular county actually impact income. The main question we are aiming to answer is: what predicts per capita income? More specifically, we aim to create a model that predicts per capita income from various variables related to a county's geographic, health, social, and educational information. In addition to the main research question, we will address the following questions:

- Which variables are related to each other and which are not? Are these relationships what a reasonable person would expect?
- Ignoring all other variables, is per capita income related to crime rate, and is this relationship different in different regions? Does it matter if we use crimes or crimes per capita (crimes/population)?
- Should we be worried about missing states or missing counties?

2 Data

The data includes information on selected county demographic information (CDI) for 440 of the most populous counties in the United States between 1990 and 1992. It originally came from the Geospatial and Statistical Data Center at the University of Virginia, and we obtained it from Kutner et al. (2005) by downloading the cdi.dat file. The following variables were measured for each of the counties in the dataset in Table 1 below. The variable name is followed in parentheses by what the variable is called in the dataset file.

Variable Name	Variable Description
Identification number (id)	A unique identifier for each observation, 1-440
County (county)	Name of the county
State (state)	Two letter abbreviation for state
Land area (land.area)	Land area (in square miles)
Total population (pop)	Estimated 1990 county population
Percent of population aged 18-34 (pop.18_34)	Percent of 1990 county population aged 18-34
Percent of population 65 or older (pop.65_plus)	Percent of 1990 county population 65 or older
Number of active physicians (doctors)	Number of professionally active nonfederal physicians in 1990
Number of hospital beds (hosp.beds)	Total number of hospital beds, cribs, and bassinets during 1990
Total serious crimes (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
Percent high school graduates (pct.hs.grad)	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
Percent bachelor's degrees (pct.bach.deg)	Percent of adult population (persons 25 years old or older) with bachelor's degree

Percent below poverty level (pct.below.pov)	Percent of 1990 county population with income below poverty level
Percent unemployment (pct.unemp)	Percent of 1990 CDI population that is unemployed
Per capita income (per.cap.income)	Per-capita income (i.e. average income per person) of 1990 county population (in dollars)
Total personal income (tot.income)	Total personal income of 1990 county population (in millions of dollars)
Geographic region (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: Table of variable names and descriptions

To get an exploratory look at the data, we examine some summary statistics of each of the variables. We begin by calculating the number of unique values for the id, county, and state variables (see Technical Appendix, page 2). These variables have a lot of unique values (440, 373, and 48, respectively) and will not be very informative when we begin building models. We also calculate the proportions of observations belonging to each region (see Technical Appendix, page 2) and find that the majority of observations are in the Southern region (around 35%), followed by the North-central region and northeast regions (around 25% and 23%) and finally the western region (around 18%). Table 2 below includes summary statistics of the remaining variables:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
land.area	15.00	451.25	656.50	1041.41	946.75	20062.00
pop	100043	139027	217280	393011	436064	8863164
pop.18_34	16.4000	26.2000	28.1000	28.5684	30.0250	49.7000
pop.65_plus	3.0000	9.8750	11.7500	12.1698	13.6250	33.8000
doctors	39.000	182.750	401.000	987.998	1036.000	23677.000
hosp.beds	92.00	390.75	755.00	1458.63	1575.75	27700.00
crimes	563.0	6219.5	11820.5	27111.6	26279.5	688936.0
pct.hs.grad	46.6000	73.8750	77.7000	77.5607	82.4000	92.9000
pct.bach.deg	8.1000	15.2750	19.7000	21.0811	25.3250	52.3000
pct.below.pov	1.40000	5.30000	7.90000	8.72068	10.90000	36.30000
pct.unemp	2.20000	5.10000	6.20000	6.59659	7.50000	21.30000
per.cap.income	8899.0	16118.2	17759.0	18561.5	20270.0	37541.0
tot.income	1141.00	2311.00	3857.00	7869.27	8654.25	184230.00

Table 2: Table of summary statistics of quantitative variables

We see that the mean is larger than the median for several variables, indicating that there could be some skew in the distributions of the variables. This can also be seen in the histograms of these variables (see Technical Appendix, pages 5-7). In particular, it appears that land area, total population, number of active physicians, number of hospital beds, total serious crimes, and total personal income are skewed right.

3 Methods

3.1 Visualizing Which Variables are Related to Each Other

We began our analysis by looking at histograms of the quantitative variables, and we considered some transformations based on these visualizations. To address the first bullet-pointed question mentioned in the Introduction, we constructed scatterplots of each predictor variable versus per capita income and we examined the correlations between each of the predictors with themselves. We used side-by-side boxplots to examine the relationship between region and per capita income. Details of these analyses can be found in pages 4-14 of the Technical Appendix.

3.2 Modelling the Relationship between Per Capita Income and Crime Rate

To address the second bullet-pointed question, we consider three linear regression models predicting per capita income from crimes and region, and three linear regression models predicting per capita income from per capita crime and region. We used F-tests to determine the best model out of each set of three, then used AIC and BIC to compare the two best models according to the F-tests. Details of these analyses can be found in pages 15-24 of the Technical Appendix.

3.3 Building a model to Predict Per Capita Income

To address the main research question of building a model to predict per capita income, before using any formal variable selection methods, we began by fitting a linear regression model with all predictors except total population, total income, region, id, state, and county. We did not consider total population and total income because per capita income is a function of them. We did not consider id, state, and county because, as previously stated, there are too many unique values, so we determined that they would not be very useful in analysis. We began our analysis without region so we could focus on the quantitative variables since some variable selection processes struggle with categorical variables. We planned to reconsider region later in our analysis. After fitting the full model, we used VIF values and added variable plots to decide on which variables to remove from the model. After deciding on a model, we added region and all interactions with region, then dropped interactions that were not significant. We compared the model with and the model without interactions using an F-test. Details of these analyses can be found in pages 24-42 of the Technical Appendix.

Next, we considered models using all subsets regression, stepwise regression (once with an AIC penalty and once with a BIC penalty), and lasso (using cross validation to select lambda). We compared these models with the ones we made previously using F-tests, AIC, and BIC. We assessed model validity of all models using residual diagnostic plots and marginal model plots. Details of these analyses can be found in pages 42-57 of the Technical Appendix.

3.4 Addressing Missing States and Counties

Finally, to address the question of whether we should be concerned about missing states or missing counties, we determined which states are represented in the dataset and highlighted the three that are missing. Details of this can be found on page 57 of the Technical Appendix.

4 Results

4.1 Visualizing Which Variables are Related to Each Other

Based on the histograms of the quantitative variables, we decide to perform a log transformation on land area, total population, number of active physicians, number of hospital beds, total serious crimes, and total personal income to address extreme right skew. The log transformation allows for an intuitive percent change interpretation of the coefficients. For the most part, the log transformations appear to address the right skew and make the distributions closer to univariate normal (see histograms in Technical Appendix, page 12), except for population and total income. This is not too worrisome because per capita income is a function of these variables, so we did not include them in any models.

We use a correlation heat map (see Figure 1 below) to get a sense for which variables (after performing the transformations) are related to each other.

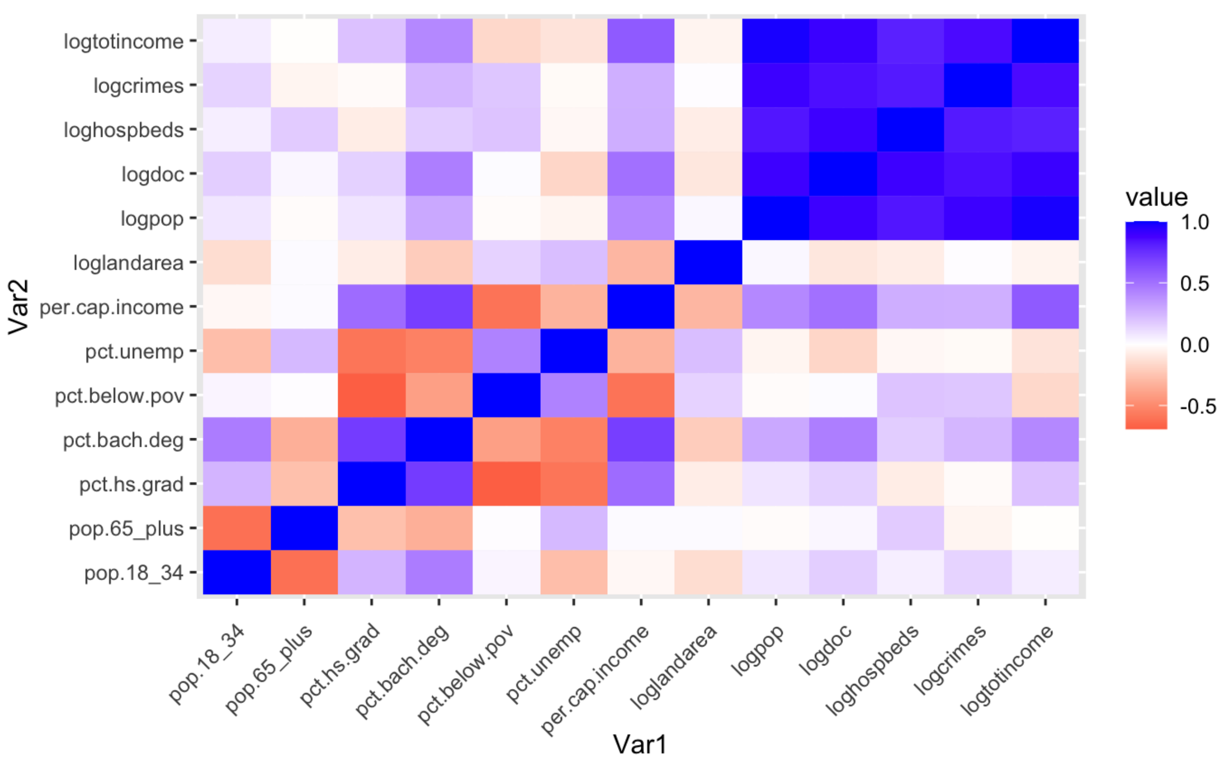


Figure 1: Correlation heat map of variables being considered for modelling

There appears to be a high correlation between the log(total income), log(crimes), log(hospital beds), log(doctors), and log(population). These relationships are generally what a reasonable person would expect, except maybe log(crimes) being related to the health variables. It appears that percent population 65 plus is not related to log(population), log(land area), and per capita income. We find it surprising that the percentage of the population 65 plus is not related to per capita income since income for the retired population is likely quite different from the income of other portions of the population.

We use scatter plots (see Technical Appendix, pages 8-10, 14) to evaluate the relationship between per capita income and the other variables. There appears to be a positive linear relationship between per capita income and percent of high school graduates and percent bachelor's degree, and a negative linear relationship between per capita income and % below poverty level. These relationships with per capita income are what we would expect. There also appears to be a slight positive linear trend between per capita income and log(doctors) that we did not see before. It also appears that log(total income) and per capita income have a positive linear relationship, which makes sense. We use side-by-side boxplots to look at the relationship between region and per capita income, and they show that the regions have similar median per capita incomes, but all suffer from skew and the spread of the distributions varies (see Technical Appendix, page 9).

4.2 Modelling the Relationship between Per Capita Income and Crime Rate

The second bullet-pointed research question asks us to look at the relationship between per capita income and crime (we will use log(crime)), and whether or not it varies by region. We found that the best model includes log(crimes) and region as additive terms according to the partial F-tests comparing the three potential models (see Technical Appendix, pages 14-19). This suggests that per-capita income is related to log(crime rate) and region, but that the relationship between per-capita income and log(crime rate) is not different in different regions.

We also looked at the relationship between per capita income and log(per capita crime), and whether or not it varies by region. We again found that the best model includes (log per capita crime) and region as additive terms according to the partial F-tests comparing the three potential models (see Technical Appendix, pages 19-23). This suggests that per-capita income is related to log(crime per capita) and region, but that the relationship between per-capita income and log(crime per capita) is not different in different regions.

We compare the model using log(crimes) and the model using log(crime per capita) and find that AIC and BIC prefer the model with log(crimes) (see Technical Appendix, page 23-24). However, using log(per capita crime) best answers the question because it makes the most sense for the variables to be on the same scale. Thus, we will use log(per capita crime) for the remainder of the analysis.

According to the model, we see that a 1% increase in log(per capita crime) is associated with a $6.60 dollar increase in per capita income on average. In the North-central region, the baseline per capita income is $20,349.70. In the Western region, the baseline per capita income is $20,191.70, and this is not significantly different from the north-central region baseline. The North Eastern and Southern baselines do differ significantly from the North-central baseline, with values of $22,793.90 and $19,275.90 respectively. See Technical Appendix, pages 20-21 for the model output.

4.3 Building a model to Predict Per Capita Income

To find the best model to predict per capita income, we fit a total of eleven models (see Technical Appendix, pages 24-57). We narrowed it down to two models to decide between:

- A model predicting per capita income from percent population 18-34, percentage high school graduates, percentage bachelor degrees, percentage below poverty level, percentage unemployment, land area, doctors, and region.

- The same model, but without the percentage of high school graduates.

We ultimately decided on the first listed model over the second because AIC and BIC preferred the first model (see Technical Appendix, page 52), and the first model was confirmed by the stepwise regression and lasso variable selection techniques (see Technical Appendix, pages 52-54). Table 3 provides the coefficient estimates and standard errors of the model:

Coefficient	Estimate	Standard Error
Intercept	30022.004	2049.377
Population 18-34	-289.194	22.690
Percentage high school graduates	-131.152	23.685
Percentage bachelor degrees	343.777	23.403
Percentage below poverty level	-489.843	55.665
Percentage unemployment	400.897	76.394
log(Land area)	-684.294	113.148
log(Doctors)	972.718	83.396
Percentage bachelor degrees:regionNE	93.222	20.719
Percentage bachelor degrees:regionS	8.234	17.029
Percentage bachelor degrees:regionW	33.849	19.161

Percentage below poverty level:regionNE	-8.576	78.924
Percentage below poverty level:regionS	144.065	59.163
Percentage below poverty level:regionW	29.330	80.877
Percentage unemployment:regionNE	-301.085	106.401
Percentage unemployment:regionS	-326.500	90.140
Percentage unemployment:regionW	-157.249	102.606

Table 3: Coefficient estimates and standard errors of final model

The relationship between the predictors and per capita income is significant. For a 1% increase in population aged 18 to 34, we would expect a $289.20 decrease in per capita income on average. For a 1% increase in the percentage of high school graduates, we would expect a $131.15 decrease in per capita income on average. For a 1% increase in land area, we would expect a $6.84 decrease in per capita income on average. For a 1% increase in doctors, we would expect an $9.72 increase in per capita income on average.

The interpretation of the predictors included in interaction terms is slightly different. For a 1% increase in percentage of bachelor degrees, we would expect $343.78 increase in per capita income on average for the North-central region. For that same 1% increase in percentage of bachelor degrees, we would expect a $437, $352.01, and $377.63 increase in per capita income on average for the Northeastern, Southern, and Western regions, respectively. For a 1% increase in percent below poverty level, we would expect a $489.84 decrease in per capita income on average for the North-central region. For that same 1% increase in percent below poverty level, we would expect a $498.42, $345.78, and $460.51 decrease in per capita income on average for the Northeastern, Southern, and Western regions, respectively. For a 1% increase in percent unemployment, we would expect an $400.90 increase in per capita income on average for the North-central region. For that same 1% increase in percent unemployment, we would expect a $99.81, $74.40, and $243.65 increase in per capita income on average for the Northeastern, Southern, and Western regions, respectively.

The residual diagnostic plot and marginal model plots generally support that the model is valid (see Technical Appendix, pages 56-7). The residuals appear to be randomly scattered with constant variance and mean zero and there do not appear to be any influential observations in the

Residuals vs Leverage plot. There are some deviations from the normal Q-Q plot. In the marginal model plots, the fitted loess curves tend to follow the fitted model curves well, which provides evidence that the model is valid. The R-squared of the model is 0.8445, indicating that the model accounts for 84.45% of the variation in per capita income. None of the VIF values are extremely high to suggest issues with multicollinearity (see Technical Appendix, page 55).

4.4 Addressing Missing States and Counties

To assess whether or not we should be worried about missing counties or states, we look at the unique values of state in the dataset and determine that Alaska, Iowa, and Wyoming are the states that are missing (see Technical Appendix, page 57). These are states that are perhaps less populous (or at least likely contain counties that are not very populous), so we should be worried that the states are missing because the results of our analysis may not generalize to these states and their counties.

5 Discussion

In this paper, we look at the relationship between a variety of predictors and per capita income. We find that there seems to be a high correlation between the following variables: log(total income), log(crimes), log(hospital beds), log(doctors), and log(population). We also find that the percentage of high school graduates, percentage of bachelor's degrees, log(doctors), and log(total income) seem to be positively linearly related to per capita income, and percentage below poverty level seems to be negatively linearly related to per capita income. These all seem to be relationships that we would reasonably expect. It looks like the percentage of population 65 plus does not seem to be related to log(population), log(land area), and per capita income. We think it is somewhat surprising that the percentage of population 65 plus is not related to per capita income since income for the retired population is likely quite different from the income of other portions of the population. A possible reason that we did not find a relationship may be that there is not much variance in the percentage of population 65 plus across the majority of the observations in the data, so it could be difficult to detect significant differences in per capita income.

We determine that per capita income is related to per capita crime rate and region, but the relationship between per capita income and per capita crime rate is not different for different regions.

We find that the "best" model to predict per capita income includes: population aged 18-34, percentage high school graduates, percentage bachelor degrees, percentage below poverty rate, percentage unemployment, log(land area), log(doctors), the interaction between percentage bachelor degrees and region, the interaction between percentage below poverty level and region, and the interaction between percentage unemployment and region. The model uses transformations and interaction terms that are not overly difficult to interpret, generally satisfies model assumptions, and seems to be a good fit for the data.

The model predicts an increasing relationship between per capita income and doctors and a decreasing relationship between per capita income and population 18-34, percentage of high school graduates, and land area. For a 1% decrease in population 18-34, we would expect a $289.20 decrease in per capita income on average, and the other coefficients of these variables are interpreted similarly (see Results). The predictors included in interaction terms are interpreted differently, for example for a 1% increase in percentage of bachelor degrees, we would expect $343.78 increase in per capita income on average for the North-central region. For that same 1% increase in percentage of bachelor degrees, we would expect a $437, $352.01, and $377.63 increase in per capita income on average for the Northeastern, Southern, and Western regions, respectively. The other coefficients of these variables are interpreted similarly (see Results).

We believe that we should be worried about missing states or counties in the dataset because the states that are missing are ones that likely have counties that are not extremely populated, and we do not know if the results would be the same for counties that are not as populated. This suggests that a weakness of these analyses is that the data only include the most populous counties in the United States. We have no way to know if our results would be the same for less populous counties. A future study could include counties of all sizes to get a better understanding of how to best predict per capita income for all counties in the United States.

It should also be noted that the data are around 30 years old at the writing of this report. Before using the results of this report to make present-day decisions, a study with more recent data should be performed.

In summary, despite these limitations, there is evidence that we can predict per capita income using various county geographic, health, social, and educational information. Having a better understanding of what impacts per capita income most could help policymakers generate new ideas about how to address income inequality in the United States.

6 References

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

Sheather, S.J. (2009), *A Modern Approach to Regression with R*. New York: Springer Science +

Business Media LLC.

36-617 Project 1 Technical Appendix

Megan Christy

10/29/2021

Technical Appendix

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

```
## Loading required package: carData
```

```
library(MASS)
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-2
```

```
library(ggplot2)
library(knitr)
library(vtable)
```

```
## Loading required package: kableExtra
```

```
library(reshape2)
library(kableExtra)
```

Initial Data Import and Exploration

```
# Reading in the data
cdi.dat = read.table("cdi.dat")
kable(head(cdi.dat[, c(1:8)]))
```

id	county	state	land.area	pop	pop.18_34	pop.65_plus	doctors
1	Los_Angeles	CA	4060	8863164	32.1	9.7	23677
2	Cook	IL	946	5105067	29.2	12.4	15153
3	Harris	TX	1729	2818199	31.3	7.1	7553
4	San_Diego	CA	4205	2498016	33.5	10.9	5905
5	Orange	CA	790	2410556	32.6	9.2	6062
6	Kings	NY	71	2300664	28.3	12.4	4861

```
kable(head(cdi.dat[, c(9:17)]))
```

hosp.beds	crimes	pct.hs.grad	pct.bach.deg	pct.below.pov	pct.unemp	per.cap.income	tot.income	region
27700	688936	70.0	22.3	11.6	8.0	20786	184230	W
21550	436936	73.4	22.8	11.1	7.2	21729	110928	NC
12449	253526	74.9	25.4	12.5	5.7	19517	55003	S
6179	173821	81.9	25.3	8.1	6.1	19588	48931	W
6369	144524	81.2	27.8	5.2	4.8	24400	58818	W
8942	680966	63.7	16.6	19.5	9.5	16803	38658	NE

I used the head function to get a first look at the variables in the dataset. Next, we will try to summarize the variables in the data set.

```r
# making summary tables for the variables
kbl(apply(cdi.dat[, c(1:3)],2,function(x) {length(unique(x))}), col.names = "Unique Values")
```

	Unique Values
id	440
county	373
state	48

```r
kbl(prop.table(sort(table(cdi.dat$region), decreasing = TRUE)), col.names = c("Region", "Proportion"))
```

Region	Proportion
S	0.3454545
NC	0.2454545
NE	0.2340909
W	0.1750000

```r
quant.vars = rbind(format(summary(cdi.dat$land.area), digits = 6),
                   format(summary(cdi.dat$pop), digits = 6),
                   format(summary(cdi.dat$pop.18_34), digits = 6),
                   format(summary(cdi.dat$pop.65_plus), digits = 6),
                   format(summary(cdi.dat$doctors), digits = 6),
                   format(summary(cdi.dat$hosp.beds), digits = 6),
                   format(summary(cdi.dat$crimes), digits = 6),
                   format(summary(cdi.dat$pct.hs.grad), digits = 6),
                   format(summary(cdi.dat$pct.bach.deg), digits = 6),
                   format(summary(cdi.dat$pct.below.pov), digits = 6),
                   format(summary(cdi.dat$pct.unemp), digits = 6),
                   format(summary(cdi.dat$per.cap.income), digits = 6),
                   format(summary(cdi.dat$tot.income), digits = 6))
rownames(quant.vars) = c("land.area", "pop", "pop.18_34", "pop.65_plus",
                         "doctors", "hosp.beds", "crimes", "pct.hs.grad",
                         "pct.bach.deg", "pct.below.pov", "pct.unemp",
                         "per.cap.income", "tot.income")
kable(quant.vars)
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
land.area	15.00	451.25	656.50	1041.41	946.75	20062.00
pop	100043	139027	217280	393011	436064	8863164
pop.18_34	16.4000	26.2000	28.1000	28.5684	30.0250	49.7000
pop.65_plus	3.0000	9.8750	11.7500	12.1698	13.6250	33.8000
doctors	39.000	182.750	401.000	987.998	1036.000	23677.000
hosp.beds	92.00	390.75	755.00	1458.63	1575.75	27700.00
crimes	563.0	6219.5	11820.5	27111.6	26279.5	688936.0
pct.hs.grad	46.6000	73.8750	77.7000	77.5607	82.4000	92.9000
pct.bach.deg	8.1000	15.2750	19.7000	21.0811	25.3250	52.3000
pct.below.pov	1.40000	5.30000	7.90000	8.72068	10.90000	36.30000
pct.unemp	2.20000	5.10000	6.20000	6.59659	7.50000	21.30000
per.cap.income	8899.0	16118.2	17759.0	18561.5	20270.0	37541.0
tot.income	1141.00	2311.00	3857.00	7869.27	8654.25	184230.00

I created tables to summarize the variables in the dataset. id, county, and state have a large number of unique values, so there isn't a ton we can do to summarize these variables. Looking at the region variable, we see that the largest proportion of observations are from the Southern region. We summarized the quantitative variables using basic summary statistics. We next will check if there is any missing data.

```
# checking for NAs
length(which(is.na(cdi.dat$id) == TRUE))
```

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

```
length(which(is.na(cdi.dat$county) == TRUE))
```

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

```
length(which(is.na(cdi.dat$state) == TRUE))
```

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

```
length(which(is.na(cdi.dat$land.area) == TRUE))
```

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

```
length(which(is.na(cdi.dat$pop) == TRUE))
```

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

```
length(which(is.na(cdi.dat$pop.18_34) == TRUE))
```

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

```
length(which(is.na(cdi.dat$pop.65_plus) == TRUE))
```

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

```
length(which(is.na(cdi.dat$doctors) == TRUE))
```

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

```
length(which(is.na(cdi.dat$hosp.beds) == TRUE))
```

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

```
length(which(is.na(cdi.dat$crimes) == TRUE))
```

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

```
length(which(is.na(cdi.dat$pct.hs.grad) == TRUE))
```

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

```
length(which(is.na(cdi.dat$pct.bach.deg) == TRUE))
```

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

```
length(which(is.na(cdi.dat$pct.below.pov) == TRUE))
```

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

```
length(which(is.na(cdi.dat$pct.unemp) == TRUE))
```

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

```
length(which(is.na(cdi.dat$per.cap.income) == TRUE))
```

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

```
length(which(is.na(cdi.dat$tot.income) == TRUE))
```

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

```
length(which(is.na(cdi.dat$region) == TRUE))
```

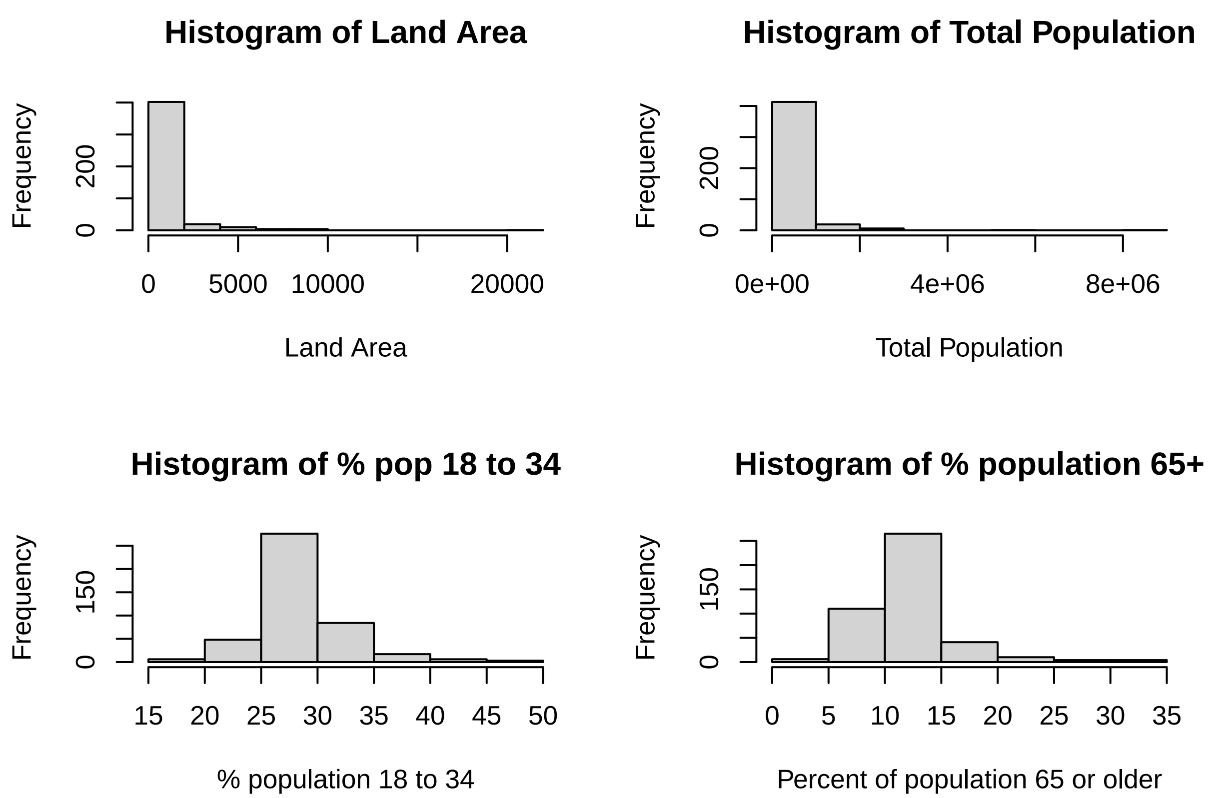
```
## [1] 0
```

There is no missing data. This makes sense since the description of the dataset notes that counties with missing data were deleted.

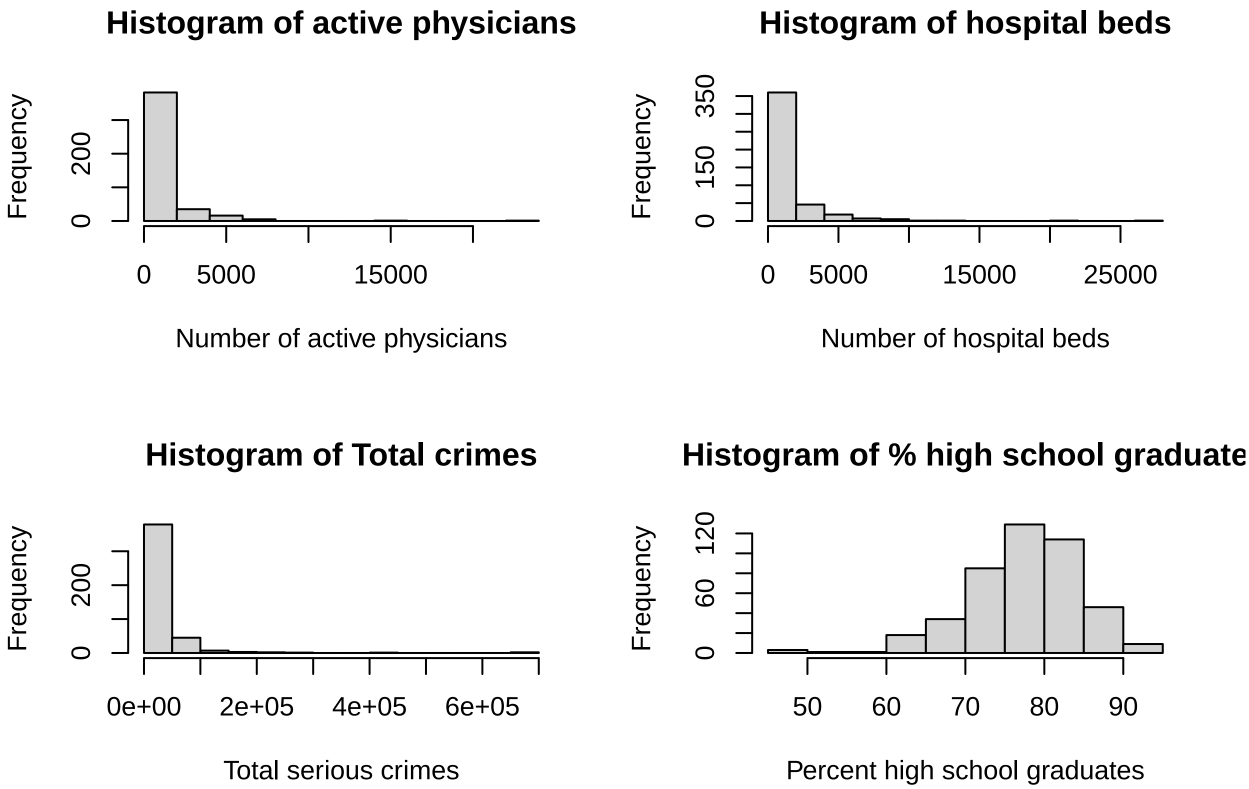
Getting a sense for which variables are and are not related

Next, we will perform EDA to look at the distributions of the variables. We construct histograms to look at the quantitative variables, and a bar plot to look at region. We did not look at id, state, or county since there are so many unique values of these variables.

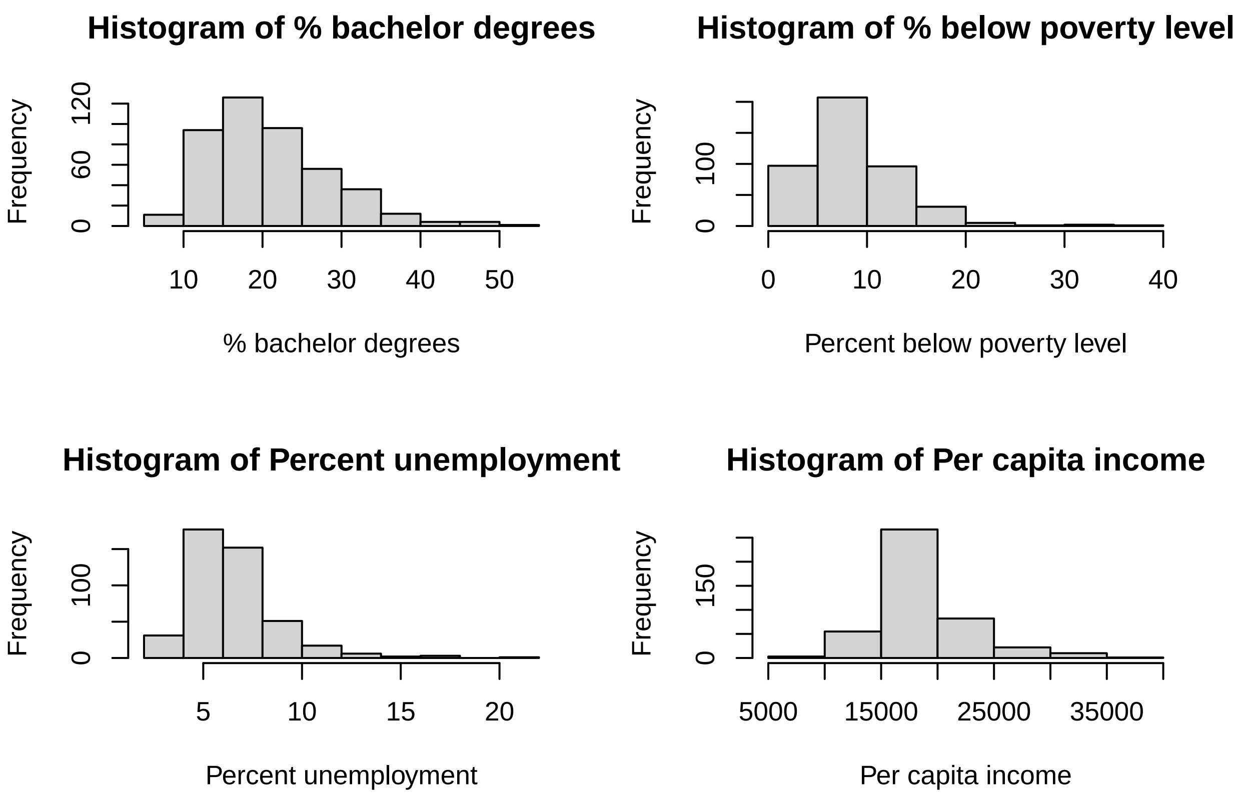
```
# EDA: Histograms of quantitative variables, bar plot of region
par(mfrow = c(2,2))
hist(cdi.dat$land.area, xlab = "Land Area", main = "Histogram of Land Area")
hist(cdi.dat$pop, xlab = "Total Population", main = "Histogram of Total Population")
hist(cdi.dat$pop.18_34, xlab = "% population 18 to 34",
     main = "Histogram of % pop 18 to 34")
hist(cdi.dat$pop.65_plus, xlab = "Percent of population 65 or older",
     main = "Histogram of % population 65+")
```



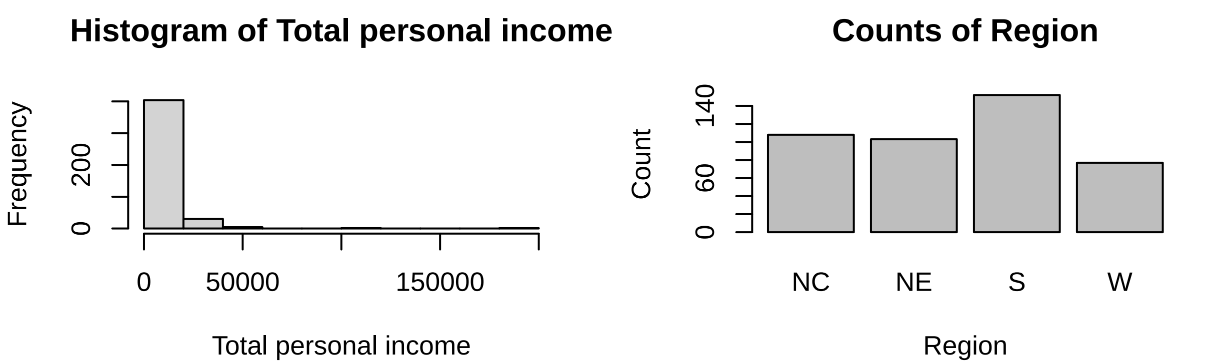

```
hist(cdi.dat$doctors, xlab = "Number of active physicians",
     main = "Histogram of active physicians")
hist(cdi.dat$hosp.beds, xlab = "Number of hospital beds",
     main = "Histogram of hospital beds")
hist(cdi.dat$crimes, xlab = "Total serious crimes",
     main = "Histogram of Total crimes")
hist(cdi.dat$pct.hs.grad, xlab = "Percent high school graduates",
     main = "Histogram of % high school graduate")
```




```
hist(cdi.dat$pct.bach.deg, xlab = "% bachelor degrees",
     main = "Histogram of % bachelor degrees")
hist(cdi.dat$pct.below.pov, xlab = "Percent below poverty level",
     main = "Histogram of % below poverty level")
hist(cdi.dat$pct.unemp, xlab = "Percent unemployment",
     main = "Histogram of Percent unemployment")
hist(cdi.dat$per.cap.income, xlab = "Per capita income",
     main = "Histogram of Per capita income")
```



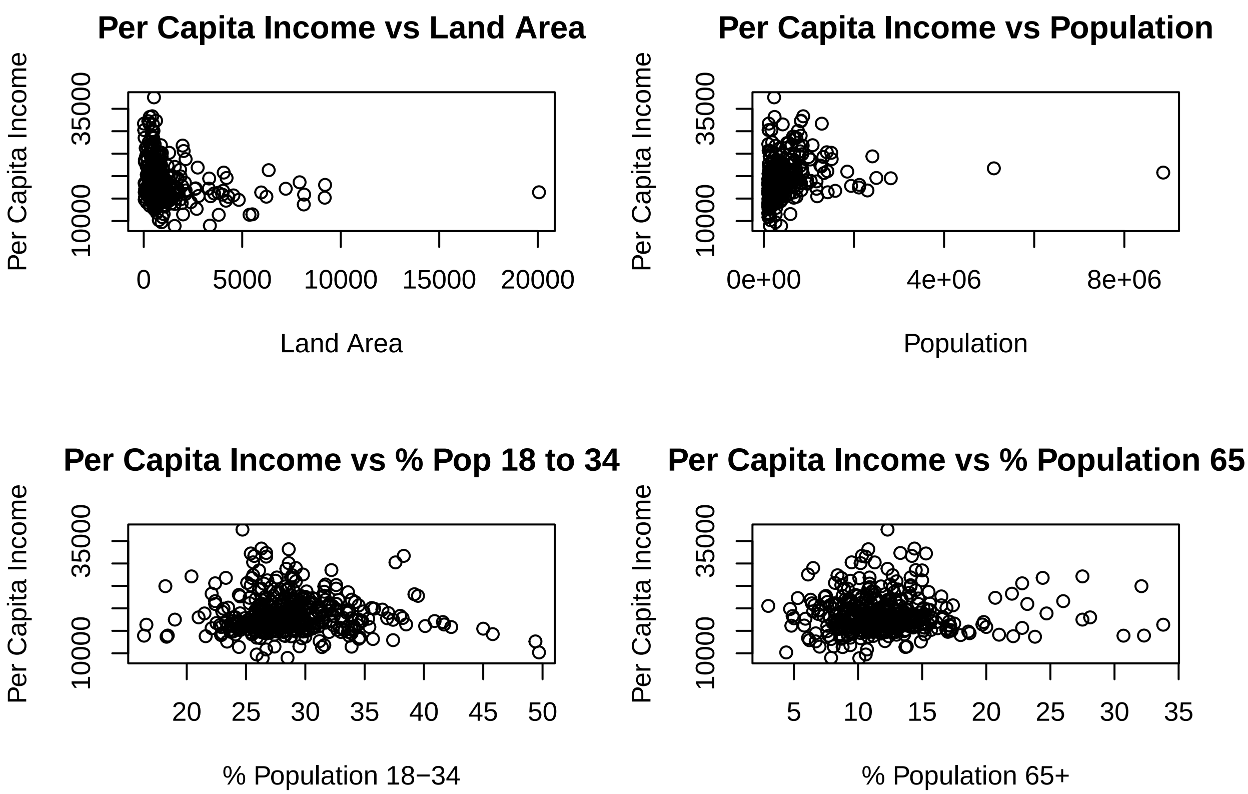

```
hist(cdi.dat$tot.income, xlab = "Total personal income",
     main = "Histogram of Total personal income")
barplot(table(cdi.dat$region), xlab = "Region", ylab = "Count",
        main = "Counts of Region")
```



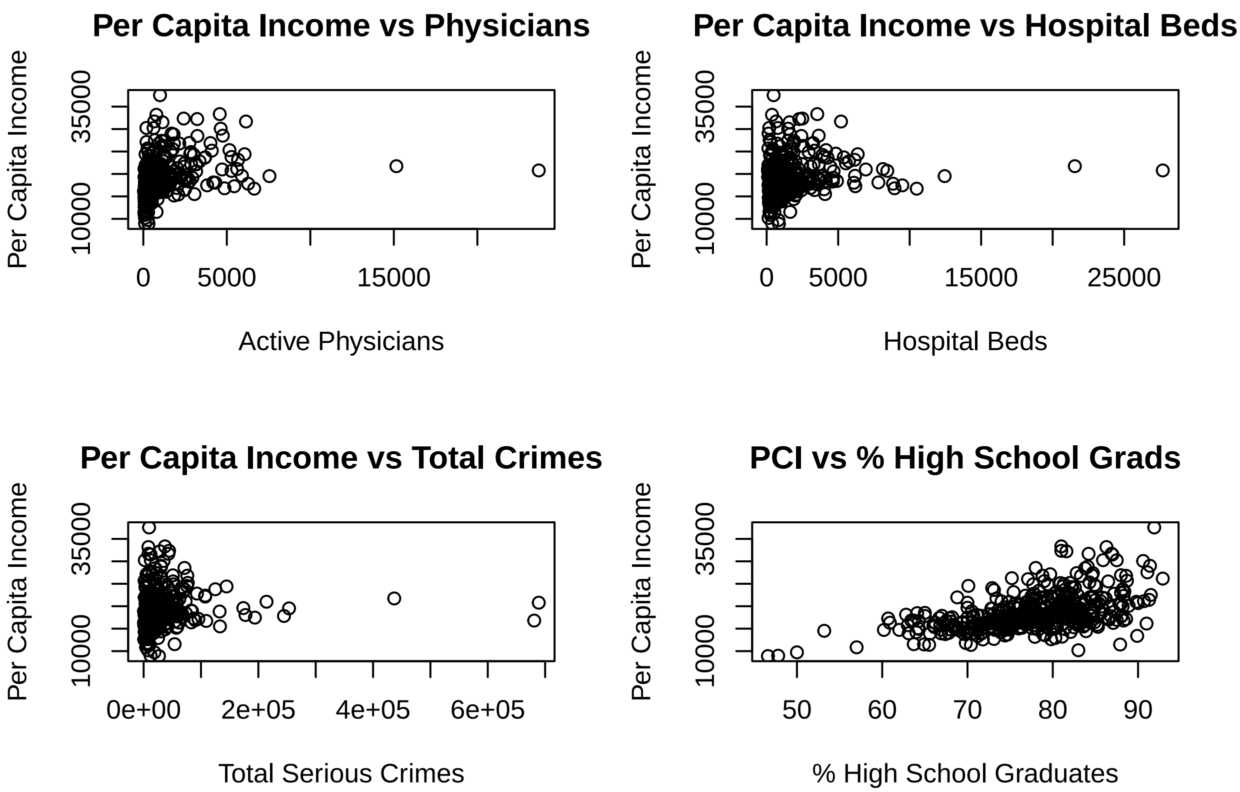
There is clear right skew in many of the histograms (like land area, total population, number of active physicians, number of hospital beds, total serious crimes, and total personal income), and we will consider transformations of these variables to address issues with skew. However, first we look at the relationship between our variable of interest (per.cap.income) and all the other untransformed variables.

```
# EDA: bivariate plots with per.cap.income
par(mfrow = c(2,2))
plot(cdi.dat$land.area, cdi.dat$per.cap.income, xlab = "Land Area", ylab = "Per Capita Income",
     main = "Per Capita Income vs Land Area")
plot(cdi.dat$pop, cdi.dat$per.cap.income, xlab = "Population", ylab = "Per Capita Income",
     main = "Per Capita Income vs Population")
plot(cdi.dat$pop.18_34, cdi.dat$per.cap.income, xlab = "% Population 18-34",
     ylab = "Per Capita Income",
```

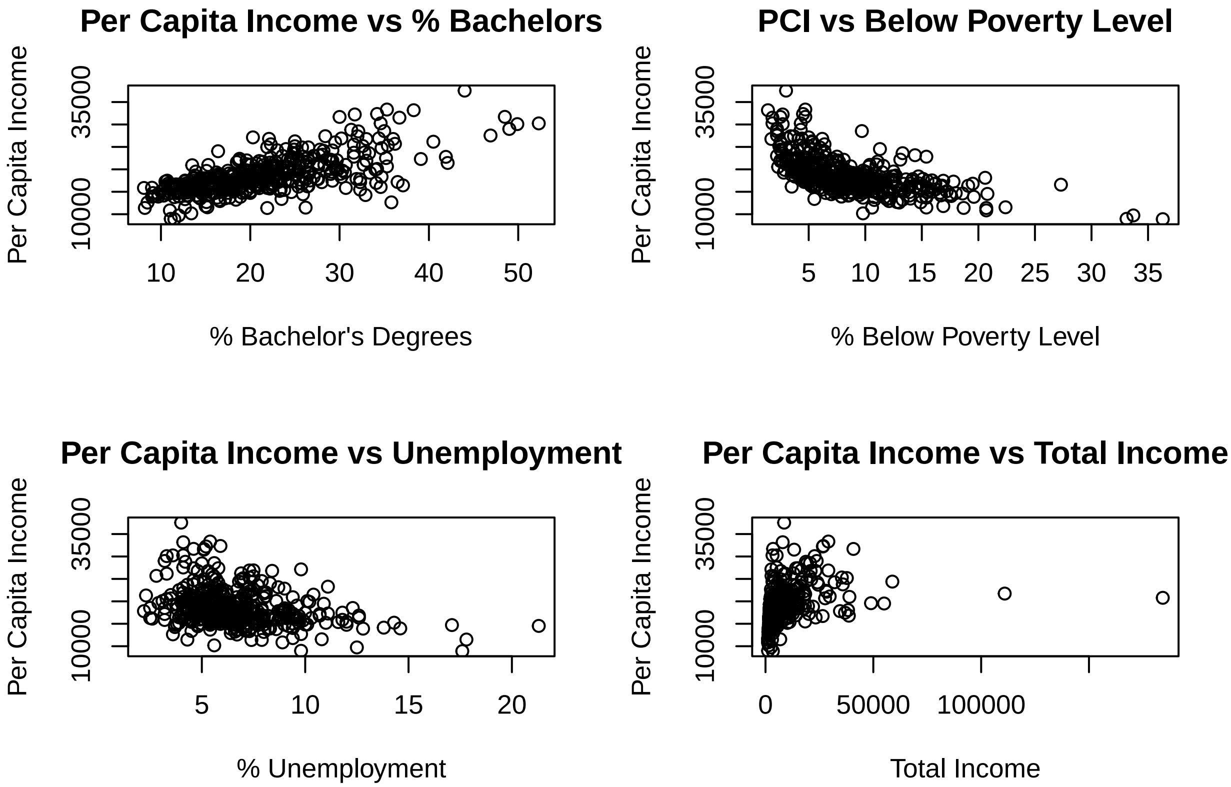
```
     main = "Per Capita Income vs % Pop 18 to 34")
plot(cdi.dat$pop.65_plus, cdi.dat$per.cap.income, xlab = "% Population 65+",
     ylab = "Per Capita Income",
     main = "Per Capita Income vs % Population 65+")
```



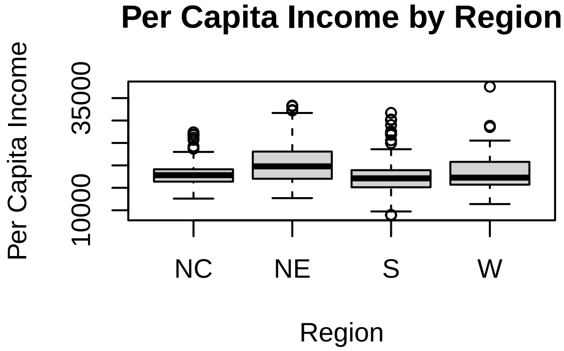

```
plot(cdi.dat$doctors, cdi.dat$per.cap.income, xlab = "Active Physicians",
     ylab = "Per Capita Income",
     main = "Per Capita Income vs Physicians")
plot(cdi.dat$hosp.beds, cdi.dat$per.cap.income, xlab = "Hospital Beds",
     ylab = "Per Capita Income",
     main = "Per Capita Income vs Hospital Beds")
plot(cdi.dat$crimes, cdi.dat$per.cap.income, xlab = "Total Serious Crimes",
     ylab = "Per Capita Income",
     main = "Per Capita Income vs Total Crimes")
plot(cdi.dat$pct.hs.grad, cdi.dat$per.cap.income, xlab = "% High School Graduates",
     ylab = "Per Capita Income",
     main = "PCI vs % High School Grads")
```




```
plot(cdi.dat$pct.bach.deg, cdi.dat$per.cap.income, xlab = "% Bachelor's Degrees",
     ylab = "Per Capita Income",
     main = "Per Capita Income vs % Bachelors")
plot(cdi.dat$pct.below.pov, cdi.dat$per.cap.income, xlab = "% Below Poverty Level",
     ylab = "Per Capita Income",
     main = "PCI vs Below Poverty Level")
plot(cdi.dat$pct.unemp, cdi.dat$per.cap.income, xlab = "% Unemployment",
     ylab = "Per Capita Income",
     main = "Per Capita Income vs Unemployment")
plot(cdi.dat$tot.income, cdi.dat$per.cap.income, xlab = "Total Income",
     ylab = "Per Capita Income",
     main = "Per Capita Income vs Total Income")
```



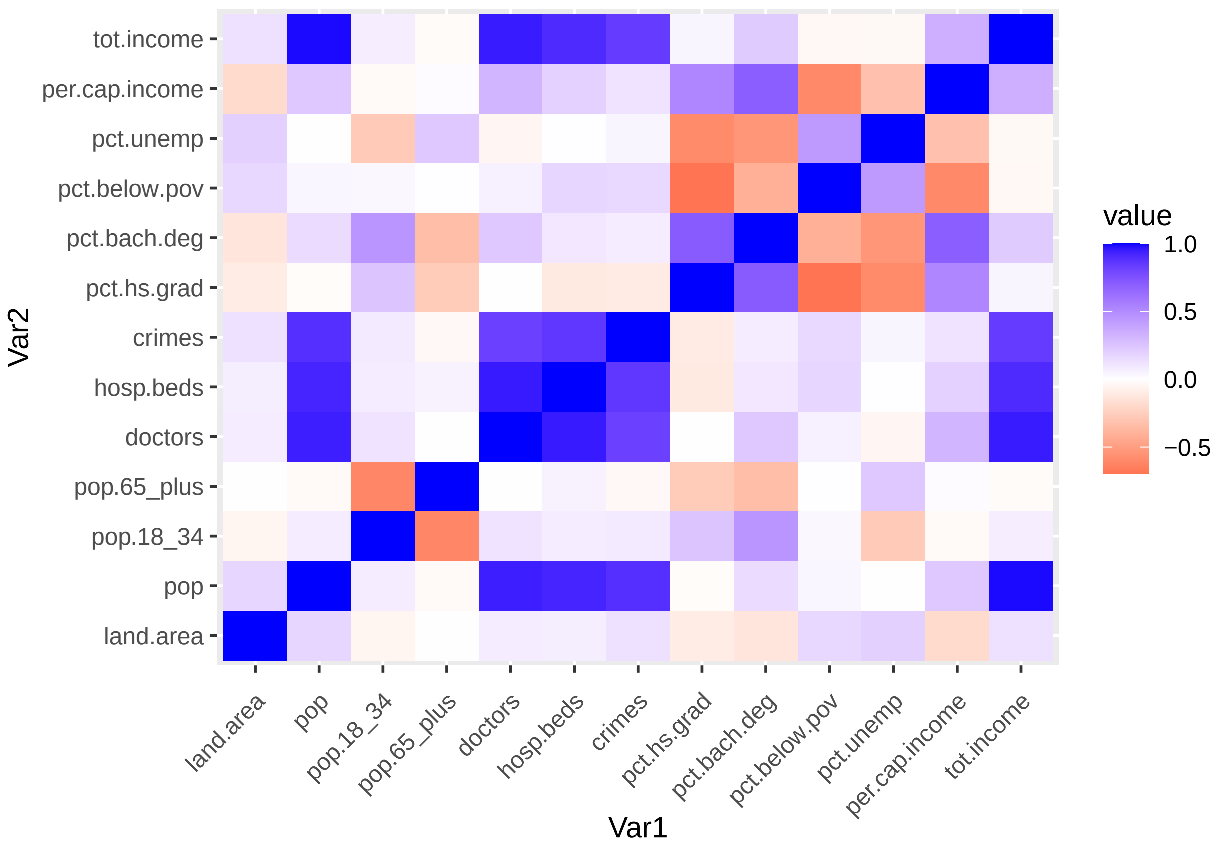

```
boxplot(per.cap.income ~ region, data = cdi.dat, xlab = "Region", ylab = "Per Capita Income",
        main = "Per Capita Income by Region")
```



Due to the skewed nature in many of the variables that we noticed previously, it is hard to get an idea of some relationships with per capita income. However, there appears to be a positive linear relationship between per capita income and % high school graduates and % bachelor's degree, and a negative linear relationship between per capita income and % below poverty level. The side by side boxplots show that the regions have similar median per capita incomes, but all suffer from skew and the spread of the distributions varies. It is also important to look at the relationships between the predictors with each other. We will construct a heatmap of the correlation to examine this since there are so many variables.

```
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="w
}
```

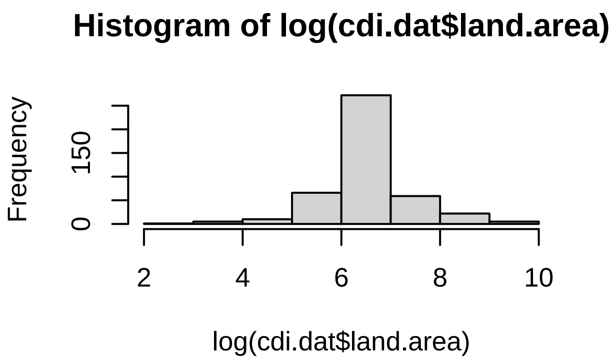
```r
corgraph(cdi.dat[,c(4:16)])
```

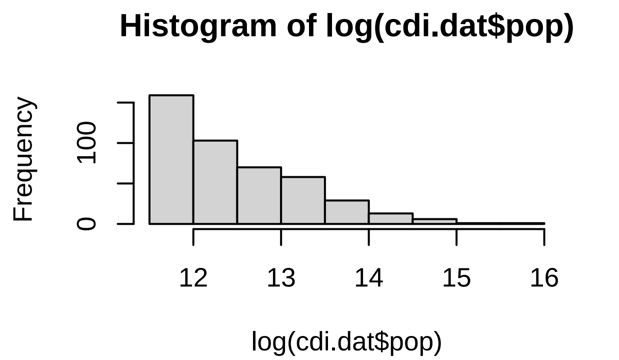


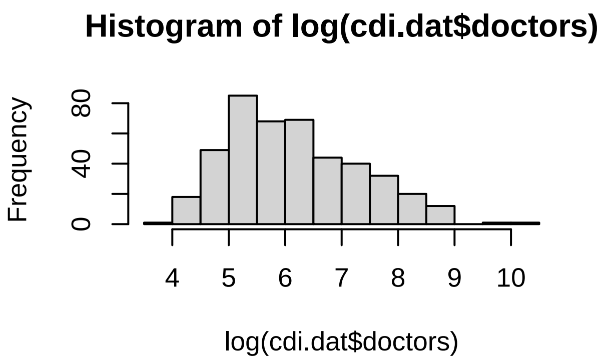
It appears that total income is highly correlated with population. It also appears that doctors, hospital beds, and crimes are highly correlated with each other, as well as with total income and population. All of these are what we would expect, except maybe crimes being related to the health variables. The fact that there is high correlation between predictors suggests that we will have issues with multicollinearity when we begin fitting models. Some examples of variables that are not related to each other include percent population 65 plus and total income, percent below poverty and total income, and percent unemployment and total income. These seem to be somewhat surprising. However, there does appear to be moderate correlation between these variables and per capita income.

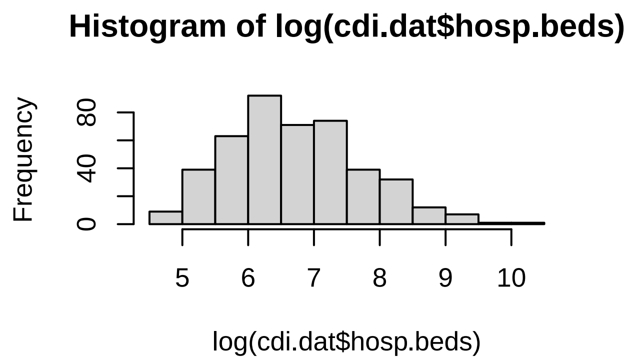
Now, we will consider transformations of the variables we identified previously. Because the variables suffer from right skew, we chose log transformations. One benefit of using a log transformation is that is had an intuitive percent change interpretability.

```r
# log transformations to address extreme right skew
par(mfrow = c(2,2))
hist(log(cdi.dat$land.area))
hist(log(cdi.dat$pop))
hist(log(cdi.dat$doctors))
hist(log(cdi.dat$hosp.beds))
```

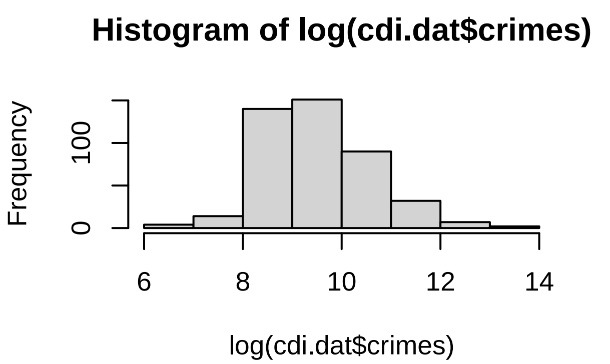


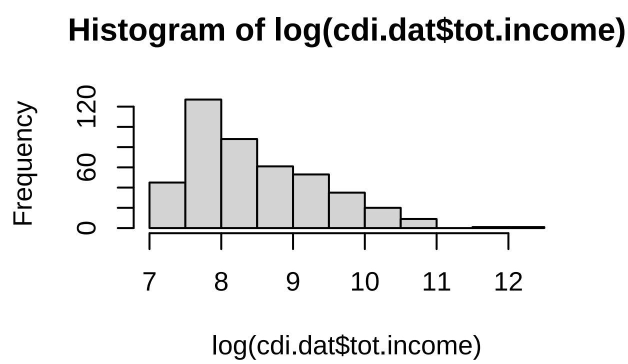






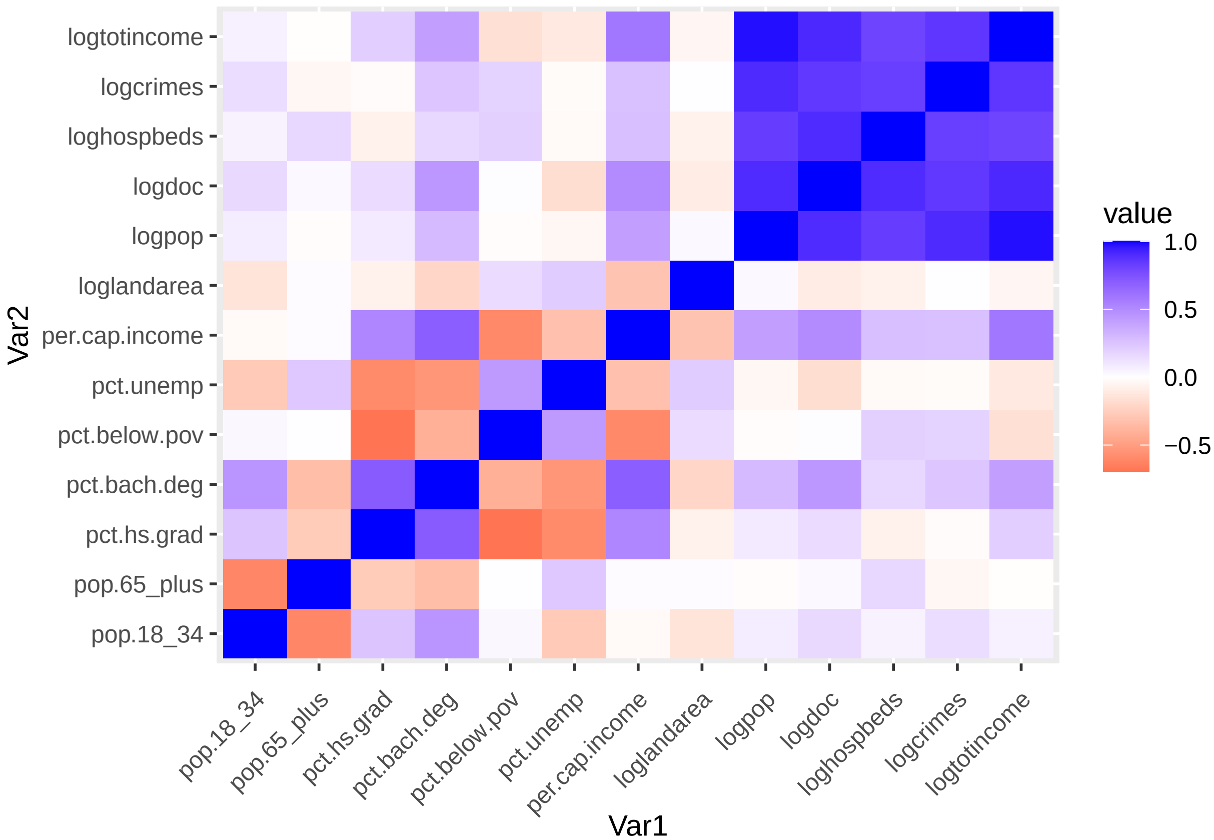

```
hist(log(cdi.dat$crimes))
hist(log(cdi.dat$tot.income))

# created new transformed variables
cdi.dat$loglandarea = log(cdi.dat$land.area)
cdi.dat$logpop = log(cdi.dat$pop)
cdi.dat$logdoc = log(cdi.dat$doctors)
cdi.dat$loghospbeds = log(cdi.dat$hosp.beds)
cdi.dat$logcrimes = log(cdi.dat$crimes)
cdi.dat$logtotincome = log(cdi.dat$tot.income)
```





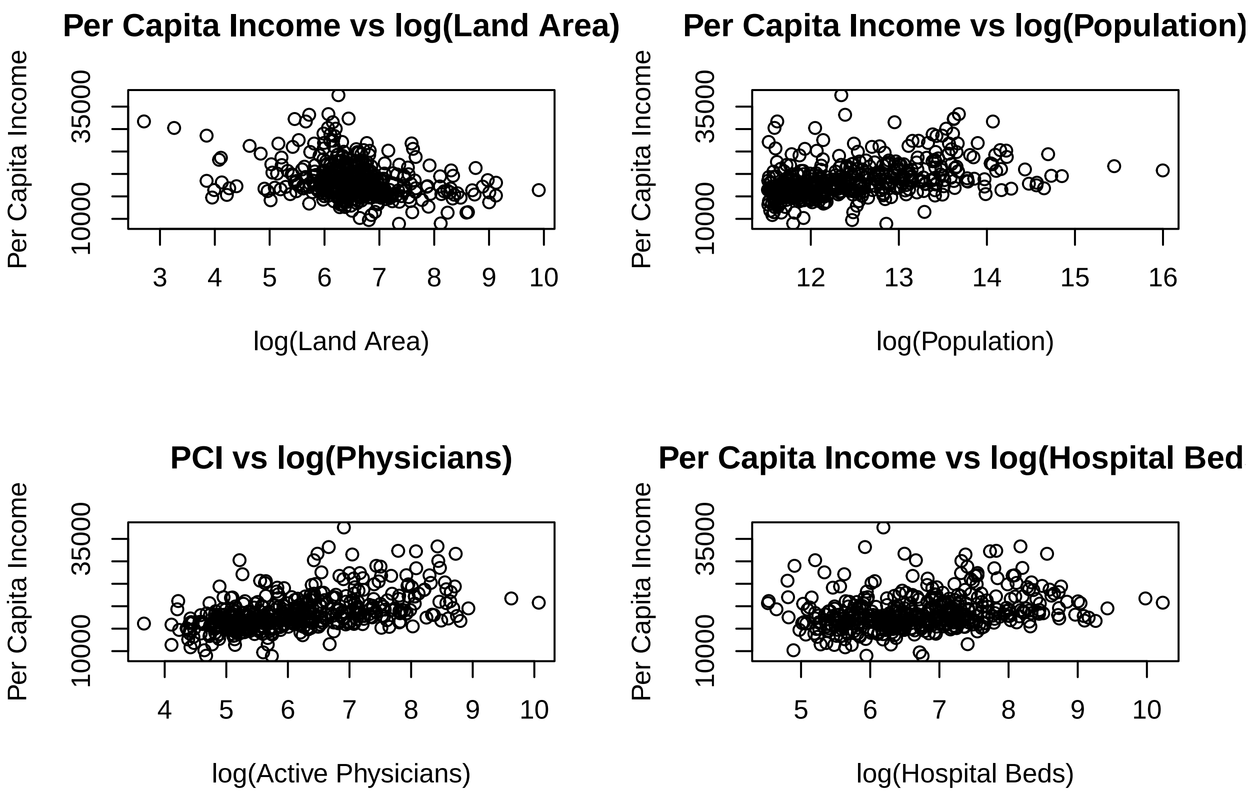
For the most part, the log transformations appear to address the right skew and make the distributions closer to univariate normal, except for population and total income. This is not too worrisome because per capita income is a function of these variables, so they likely will not be included in modelling. Next we reconstruct the heat map to examine correlations in the transformed data.

```
# heat map of correlation matrix (transformed)
corgraph(cdi.dat[,c(6:7, 11:15, 18:23)])
```

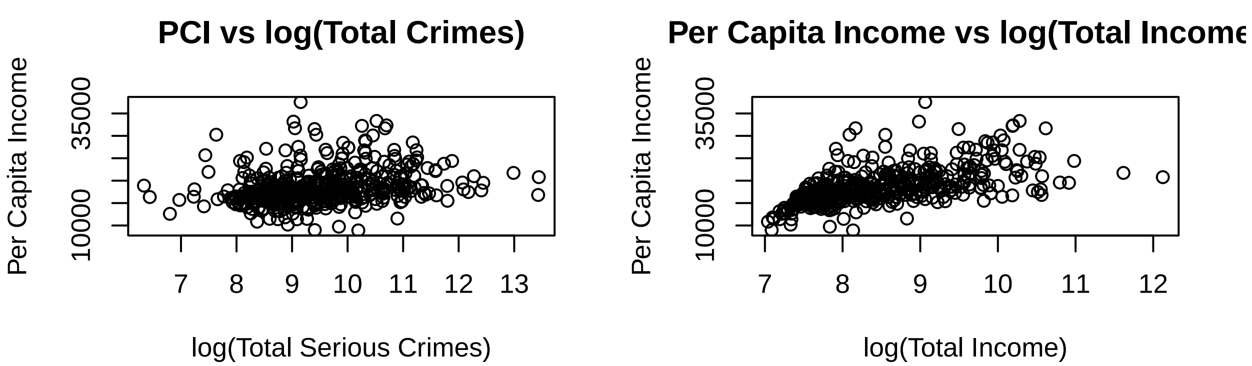


The heat map reveals similar to insights to what we saw with the untransformed data. There appears to be high correlation between the log(total income), log(crimes), log(hospital beds), log(doctors), and log(population). We also reconstruct the scatterplots between the transformed predictors and per capita income.

```
# EDA: bivariate plots with per.cap.income (transformed)
par(mfrow = c(2,2))
plot(cdi.dat$loglandarea, cdi.dat$per.cap.income, xlab = "log(Land Area)",
     ylab = "Per Capita Income",
     main = "Per Capita Income vs log(Land Area)")
plot(cdi.dat$logpop, cdi.dat$per.cap.income, xlab = "log(Population)",
     ylab = "Per Capita Income",
     main = "Per Capita Income vs log(Population)")
plot(cdi.dat$logdoc, cdi.dat$per.cap.income, xlab = "log(Active Physicians)",
     ylab = "Per Capita Income",
     main = "PCI vs log(Physicians)")
plot(cdi.dat$loghospbeds, cdi.dat$per.cap.income, xlab = "log(Hospital Beds)",
     ylab = "Per Capita Income",
     main = "Per Capita Income vs log(Hospital Beds)")
```




```r
plot(cdi.dat$logcrimes, cdi.dat$per.cap.income, xlab = "log(Total Serious Crimes)",
     ylab = "Per Capita Income",
     main = "PCI vs log(Total Crimes)")
plot(cdi.dat$logtotincome, cdi.dat$per.cap.income, xlab = "log(Total Income)",
     ylab = "Per Capita Income",
     main = "Per Capita Income vs log(Total Income)")
```



There appears to be a slight positive linear trend between per capita income and log(doctors) that we did not see before. It also appears that log(total income) and per capita income have a positive linear relationship, which makes sense.

For the rest of the analysis, we will use the dataset without id, county, and state (since we determined they aren't very useful since there are so many unique values) and with the transformations we performed.

```r
# constructing the final dataset
cdi.dat.final = cdi.dat[ ,c(6:7, 11:15, 17, 18:23)]
```

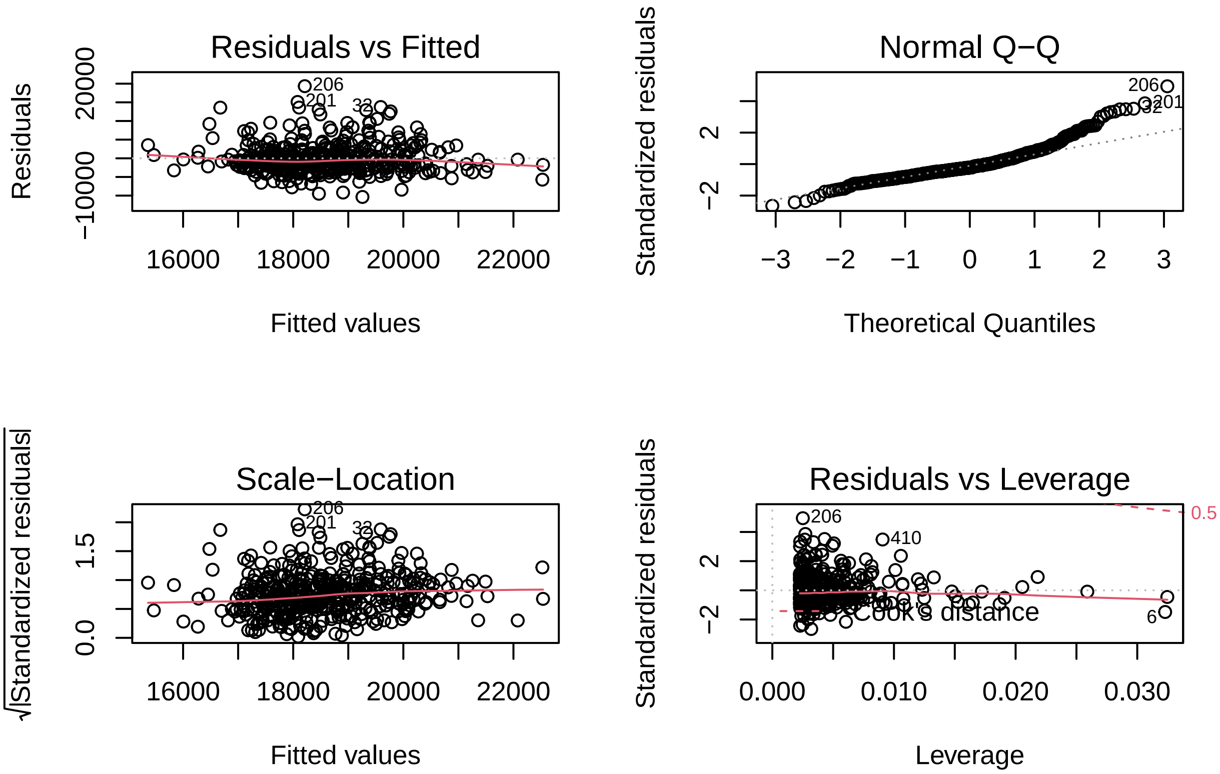
Looking at the relationship between per capita income and crime and region

The second research question asks us to look at the relationship between per capita income and crime (we will use log(crime)), and whether or not it varies by region. We begin by considering three models:

```
par(mfrow = c(2,2))
cdi.mod1 = lm(per.cap.income ~ logcrimes, data = cdi.dat.final)
summary(cdi.mod1)
```

```
##
## Call:
## lm(formula = per.cap.income ~ logcrimes, data = cdi.dat.final)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10358.7  -2292.5   -867.7   1489.4  19330.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8972.8     1651.1   5.435 9.14e-08 ***
## logcrimes     1009.0      172.6   5.845 9.90e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3914 on 438 degrees of freedom
## Multiple R-squared:  0.07236,    Adjusted R-squared:  0.07024
## F-statistic: 34.16 on 1 and 438 DF,  p-value: 9.901e-09
```

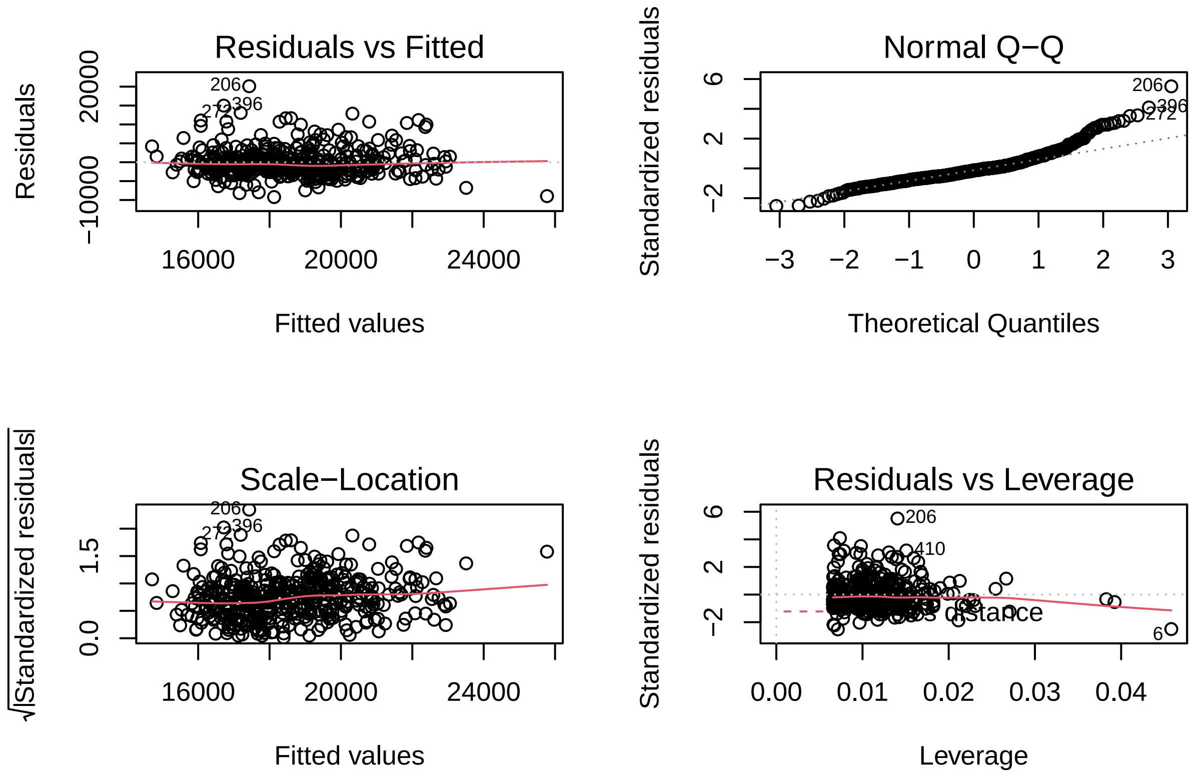
```
plot(cdi.mod1)
```




```
cdi.mod2 = lm(per.cap.income ~ logcrimes + factor(region), data = cdi.dat.final)
summary(cdi.mod2)
```

```
##
## Call:
## lm(formula = per.cap.income ~ logcrimes + factor(region), data = cdi.dat.final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9229.2 -2183.6  -502.4  1339.3 20110.9
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        6870.8     1582.5   4.342 1.76e-05 ***
## logcrimes          1237.3      167.0   7.411 6.61e-13 ***
## factor(region)NE   2284.9      506.2   4.514 8.21e-06 ***
## factor(region)S   -1354.4      468.3  -2.892  0.00402 **
## factor(region)W    -768.2      558.5  -1.376  0.16968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3676 on 435 degrees of freedom
## Multiple R-squared:  0.1875, Adjusted R-squared:  0.1801
## F-statistic:  25.1 on 4 and 435 DF,  p-value: < 2.2e-16
```

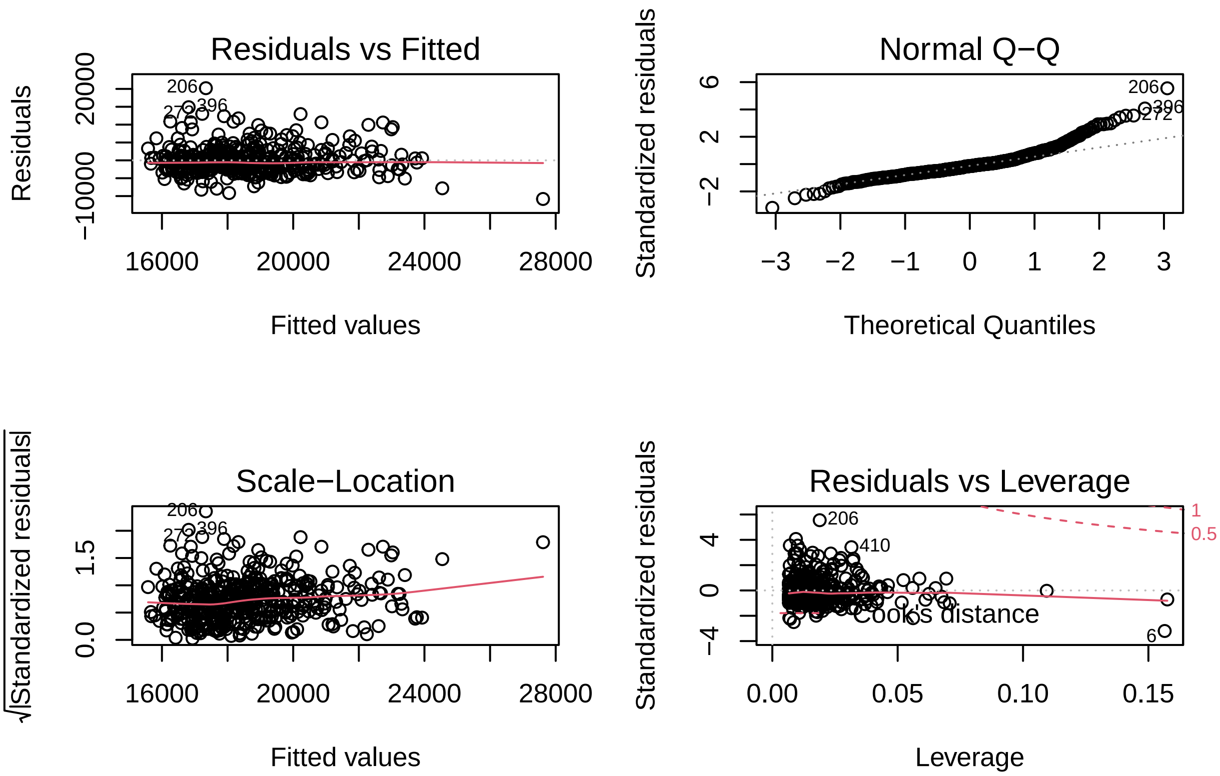
```
plot(cdi.mod2)
```




```
cdi.mod3= lm(per.cap.income ~ logcrimes * factor(region), data = cdi.dat.final)
summary(cdi.mod3)
```

```
##
## Call:
## lm(formula = per.cap.income ~ logcrimes * factor(region), data = cdi.dat.final)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10810  -2127   -533   1187  20202
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    9634.8     2888.0   3.336 0.000923 ***
## logcrimes                       938.1      310.3   3.024 0.002648 **
## factor(region)NE              -4544.8     4262.1  -1.066 0.286870
## factor(region)S               -2595.4     4201.9  -0.618 0.537117
## factor(region)W               -4784.6     4846.6  -0.987 0.324093
## logcrimes:factor(region)NE      738.8      457.8   1.614 0.107313
## logcrimes:factor(region)S       141.8      441.4   0.321 0.748223
## logcrimes:factor(region)W       426.0      499.8   0.852 0.394467
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3676 on 432 degrees of freedom
## Multiple R-squared:  0.1931, Adjusted R-squared:  0.1801
## F-statistic: 14.77 on 7 and 432 DF,  p-value: < 2.2e-16
```

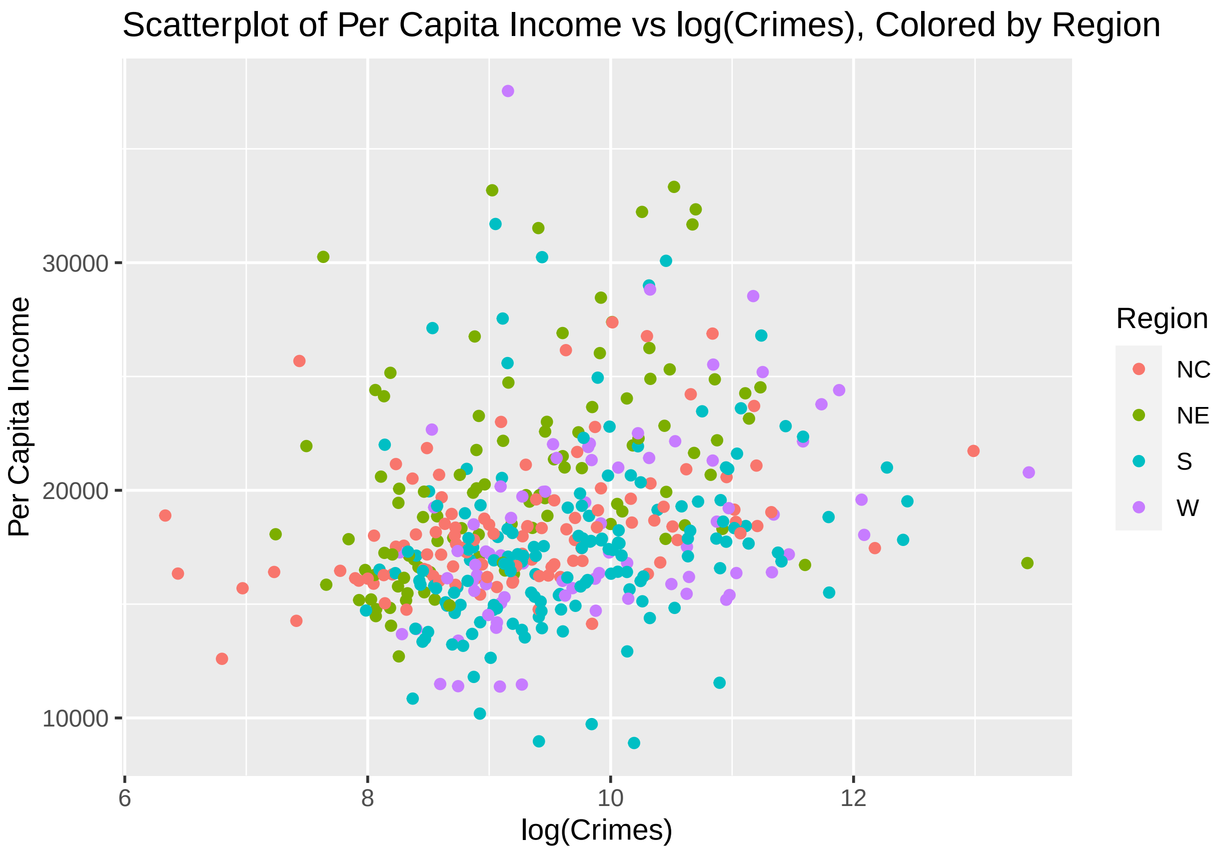
```
plot(cdi.mod3)
```




```
anova(cdi.mod1, cdi.mod2, cdi.mod3)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ logcrimes
## Model 2: per.cap.income ~ logcrimes + factor(region)
## Model 3: per.cap.income ~ logcrimes * factor(region)
##   Res.Df        RSS Df Sum of Sq       F    Pr(>F)
## 1    438 6710024435
## 2    435 5876801559  3 833222876 20.5579 1.807e-12 ***
## 3    432 5836388967  3  40412592  0.9971     0.394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(data = cdi.dat.final, aes(x = logcrimes, y = per.cap.income, color = factor(region))) +
  geom_point() +
  labs(x = "log(Crimes)", y = "Per Capita Income",
       title = "Scatterplot of Per Capita Income vs log(Crimes), Colored by Region",
       color = "Region")
```



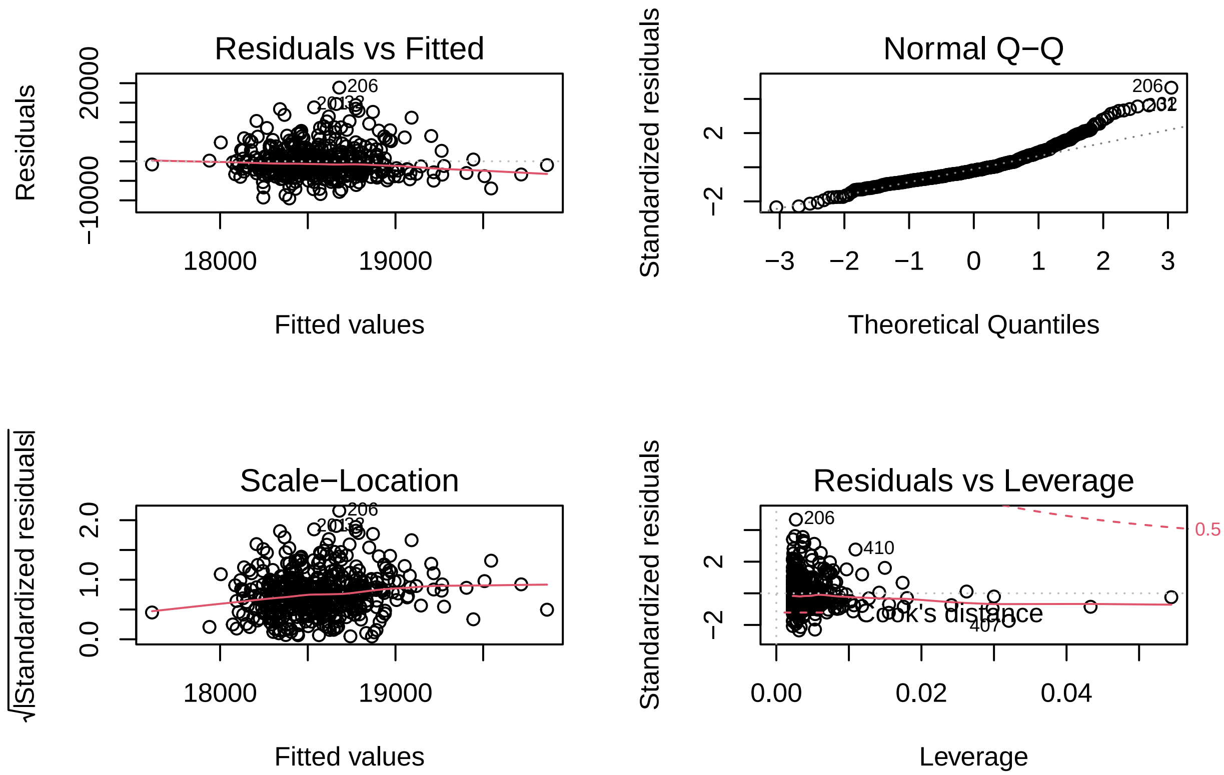
The diagnostic plots of these three models generally look acceptable (despite some large deviations in the Q-Q plot), so we can compare the them using F-tests. According to the tests, the second model (with log(crimes) and region as additive terms) is the best model. This suggests that per-capita income is related to log(crime rate) and region, but that the relationship between per-capita income and log(crime rate) is not different in different regions. This is supported visually by the scatterplot of per capita income vs log(crime rate) colored by region since it does not appear that the relationship changes based on region. According to this model, we see that a 1% increase in log(crimes) is associated with a \$12.37 increase in per capita income on average. In the North-central region, the baseline per capita income is 6,870.8 dollars. In the Western region, the baseline per capita income is 6,102.6 dollars, and this is not significantly different from the north-central region baseline. The North Eastern and Southern baselines do differ significantly from the North-central baseline, with values of 9155.7 dollars and 5516.4 dollars respectively. Now we will see if these answers change when considering log(per capita crime) instead of log(crimes).

```
cdi.dat.final$log.per.cap.crime = log(cdi.dat$crimes/cdi.dat$pop)

par(mfrow = c(2,2))
cdi.mod4 = lm(per.cap.income ~ log.per.cap.crime, data = cdi.dat.final)
summary(cdi.mod4)
```

```
##
## Call:
## lm(formula = per.cap.income ~ log.per.cap.crime, data = cdi.dat.final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9495.5 -2539.5  -782.9  1634.7 18861.4
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        16953.4     1159.5  14.621   <2e-16 ***
## log.per.cap.crime   -540.9      384.5  -1.407     0.16
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4055 on 438 degrees of freedom
## Multiple R-squared:  0.004496,   Adjusted R-squared:  0.002224
## F-statistic: 1.978 on 1 and 438 DF,  p-value: 0.1603
```

```
plot(cdi.mod4)
```

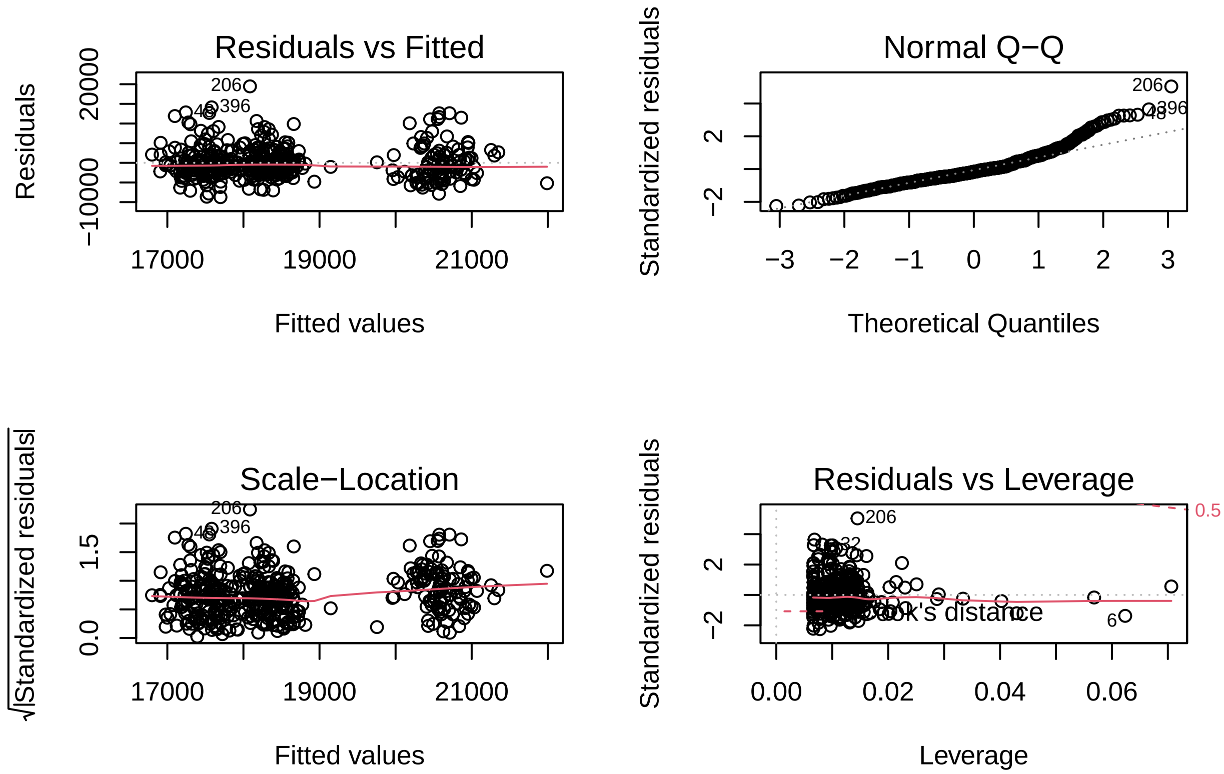



```
cdi.mod5 = lm(per.cap.income ~ log.per.cap.crime + factor(region), data = cdi.dat.final)
summary(cdi.mod5)
```

```
##
## Call:
## lm(formula = per.cap.income ~ log.per.cap.crime + factor(region),
##     data = cdi.dat.final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8725.5 -2270.1  -639.8  1768.3 19455.1
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        20349.7     1366.2  14.895  < 2e-16 ***
## log.per.cap.crime    659.9      423.2   1.559   0.1197
## factor(region)NE    2444.2      543.9   4.494 8.98e-06 ***
## factor(region)S    -1073.8      517.1  -2.077   0.0384 *
```

```
## factor(region)W     -158.0      591.5  -0.267   0.7895
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3890 on 435 degrees of freedom
## Multiple R-squared:  0.09007,    Adjusted R-squared:  0.0817
## F-statistic: 10.76 on 4 and 435 DF,  p-value: 2.501e-08
```

```
plot(cdi.mod5)
```

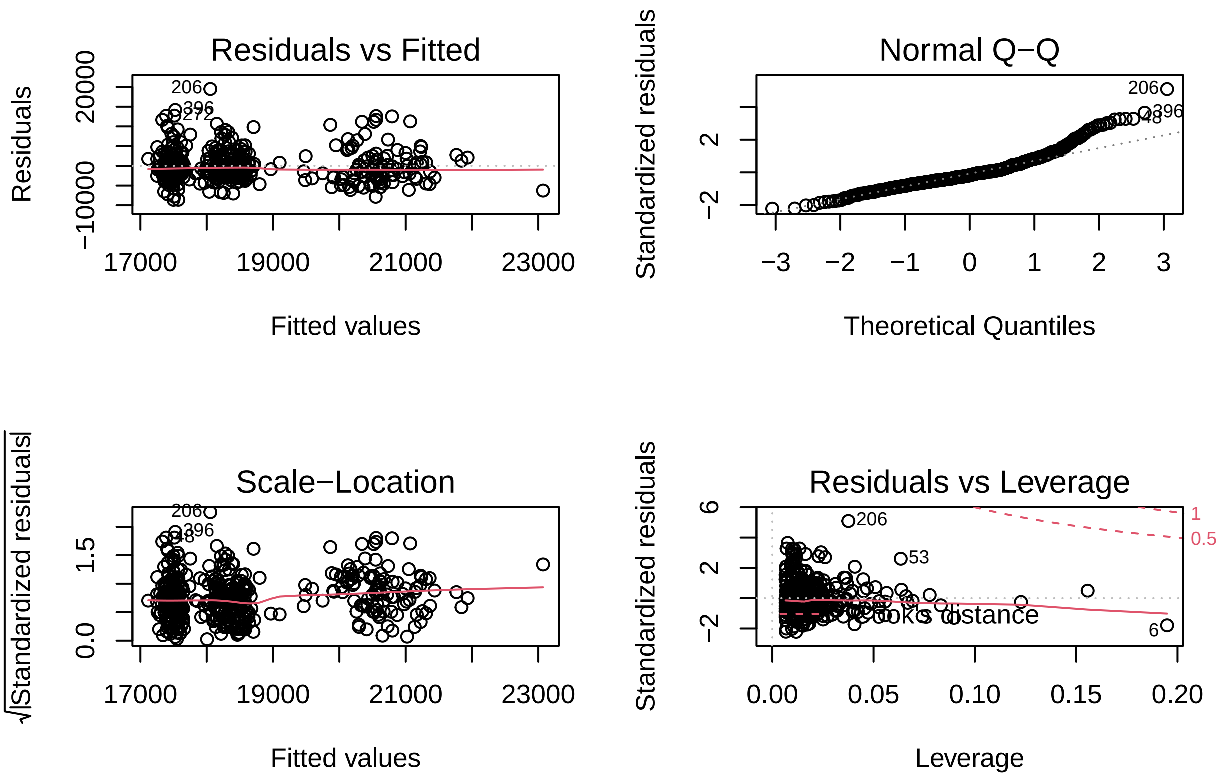



```
cdi.mod6 = lm(per.cap.income ~ log.per.cap.crime * factor(region), data = cdi.dat.final)
summary(cdi.mod6)
```

```
##
## Call:
## lm(formula = per.cap.income ~ log.per.cap.crime * factor(region),
##     data = cdi.dat.final)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8600.4 -2312.3  -653.3  1735.2 19486.5
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          19913.6     2069.5   9.622   <2e-16 ***
## log.per.cap.crime                      519.4      655.5   0.792    0.429
## factor(region)NE                      4585.8     3382.0   1.356    0.176
## factor(region)S                      -1705.7     3166.7  -0.539    0.590
## factor(region)W                        525.7     5271.3   0.100    0.921
## log.per.cap.crime:factor(region)NE     653.1     1030.9   0.634    0.527
```

```
## log.per.cap.crime:factor(region)S    -253.5     1094.4  -0.232    0.817
## log.per.cap.crime:factor(region)W     227.9     1826.0   0.125    0.901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3900 on 432 degrees of freedom
## Multiple R-squared:  0.09147,    Adjusted R-squared:  0.07675
## F-statistic: 6.213 on 7 and 432 DF,  p-value: 6.001e-07
```

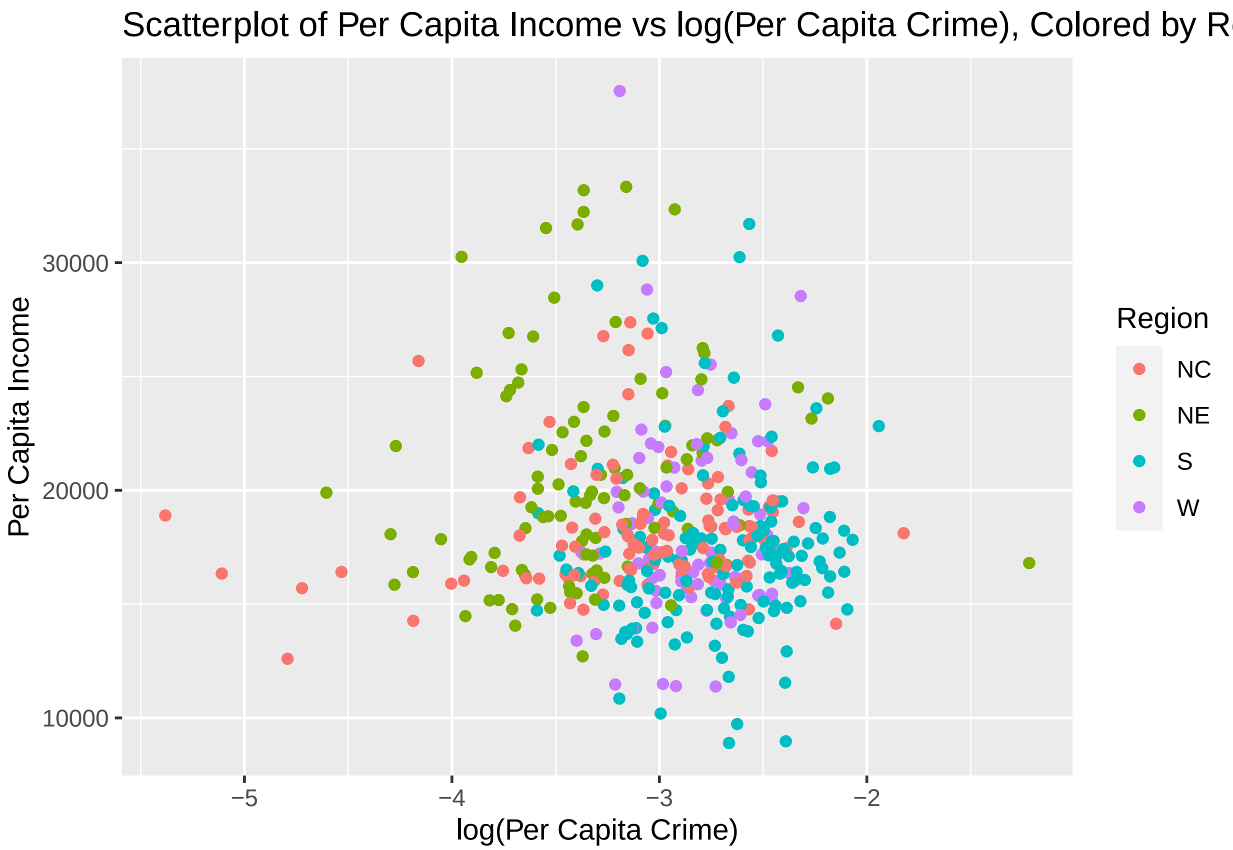
```
plot(cdi.mod6)
```




```
anova(cdi.mod4, cdi.mod5, cdi.mod6)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ log.per.cap.crime
## Model 2: per.cap.income ~ log.per.cap.crime + factor(region)
## Model 3: per.cap.income ~ log.per.cap.crime * factor(region)
##   Res.Df        RSS Df Sum of Sq       F    Pr(>F)
## 1    438 7200895643
## 2    435 6581927659  3 618967984 13.5627 1.797e-08 ***
## 3    432 6571800580  3  10127079  0.2219    0.8812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ggplot(data = cdi.dat.final, aes(x = log.per.cap.crime, y = per.cap.income,
                                 color = factor(region))) +
  geom_point() +
  labs(x = "log(Per Capita Crime)", y = "Per Capita Income",
       title = "Scatterplot of Per Capita Income vs log(Per Capita Crime), Colored by Region", color =
```



Despite some clustering in the residual plots and deviations in the QQ-plots, the diagnostic plots are not horrible and thus we can use F-tests to compare these models. The tests reveal that the model with additive terms of log(per capita crime) and region is the best model. This suggests that per-capita income is related to log(per capita crime) and region, but that the relationship between per capita income and log(per capita crime) is not different in different regions. This is supported visually by the scatterplot of per capita income vs log(per capita crime rate) colored by region since it does not appear that the relationship changes based on region. According to this model, we see that a 1% increase in log(per capita crime) is associated with a \$6.60 increase in per capita income on average. In the North-central region, the baseline per capita income is 20,349.7 dollars. In the Western region, the baseline per capita income is 20,191.7 dollars, and this is not significantly different from the north-central region baseline. The North Eastern and Southern baselines do differ significantly from the North-central baseline, with values of 22793.90 dollars and 19275.90 dollars respectively.

Now let's compare the two models (cdi.mod2 and cdi.mod5) we have interpreted:

```
AIC(cdi.mod2, cdi.mod5)
```

```
##          df      AIC
## cdi.mod2  6 8479.968
## cdi.mod5  6 8529.826
```

```
BIC(cdi.mod2, cdi.mod5)
```

```
##          df      BIC
## cdi.mod2  6 8504.488
## cdi.mod5  6 8554.347
```

Using quantitative metrics (AIC and BIC), it appears that model 2 (with just regular log(crimes)) is the better model. However, using log(per capita crime) best answers the question because it makes the most sense for the variables to be on the same scale. Thus, we will use log(per capita crime) for the remainder of the analysis.

```
cdi.dat.final = cdi.dat.final[, -13]
```

Building the "best" model to predict per capita income

Now we will try to find the best model predicting per capita income from the other variables. Since per capita income is a function of population and total income, we begin by removing log(population) and log(total income) from being considered in the model.

```
cdi.dat.modelling = cdi.dat.final[, -c(10,13)]
```

We will start by considering all variables except region, then add it later to see if it improves the model. Before using any formal variable selection methods, we tried building a model with all variables, then dropping variables based on VIFs and added variable plots.

```
# first try: all but region
mod1 = lm(per.cap.income ~. - region, data = cdi.dat.modelling)
summary(mod1)
```

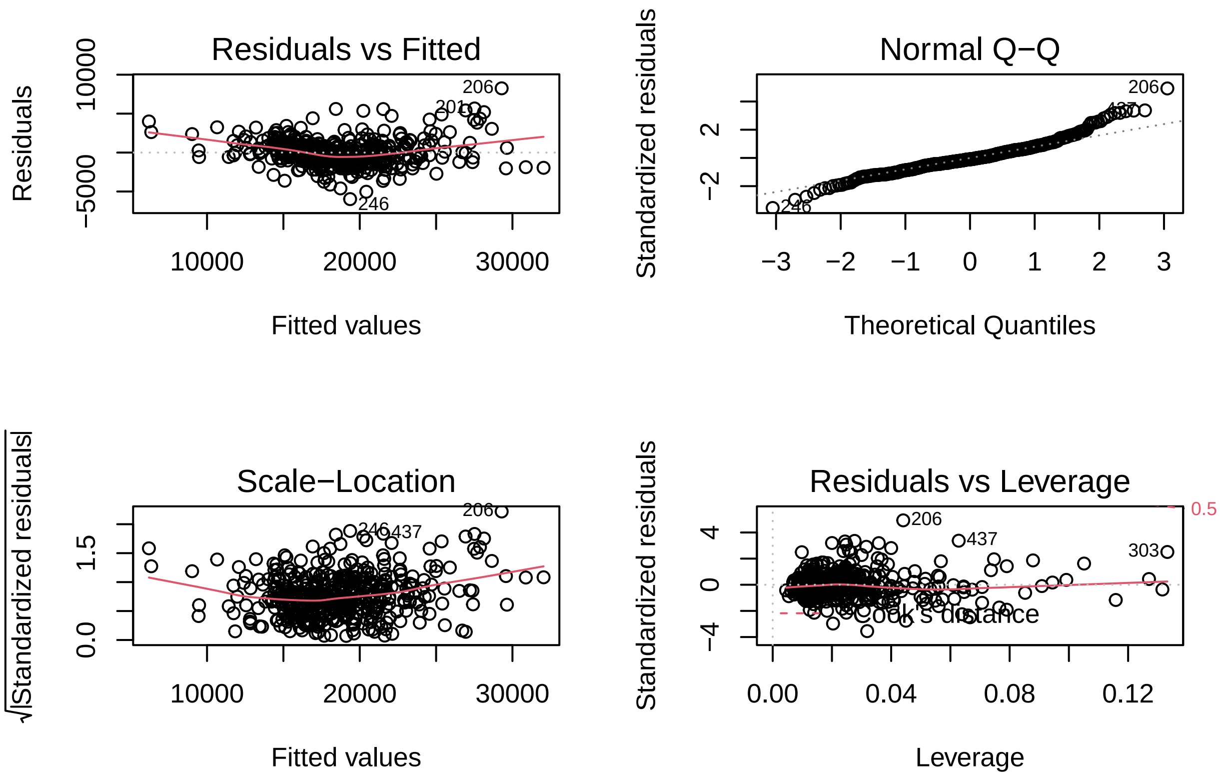
```
##
## Call:
## lm(formula = per.cap.income ~ . - region, data = cdi.dat.modelling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5976.3  -924.5  -157.8   911.5  8247.8
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       30019.06    2378.32  12.622  < 2e-16 ***
## pop.18_34          -325.01      27.44 -11.846  < 2e-16 ***
## pop.65_plus         -48.59      27.87  -1.743    0.082 .
## pct.hs.grad        -121.06      22.77  -5.317 1.70e-07 ***
## pct.bach.deg        368.42      21.39  17.226  < 2e-16 ***
## pct.below.pov      -430.06      28.96 -14.851  < 2e-16 ***
## pct.unemp           253.28      45.86   5.523 5.80e-08 ***
## loglandarea        -698.60     100.52  -6.950 1.37e-11 ***
## logdoc             1031.14     235.90   4.371 1.55e-05 ***
## loghospbeds          18.12     249.87   0.073    0.942
## log.per.cap.crime   -70.28     205.46  -0.342    0.732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1712 on 429 degrees of freedom
## Multiple R-squared:  0.8261, Adjusted R-squared:  0.8221
## F-statistic: 203.8 on 10 and 429 DF,  p-value: < 2.2e-16
```

```
vif(mod1)
```

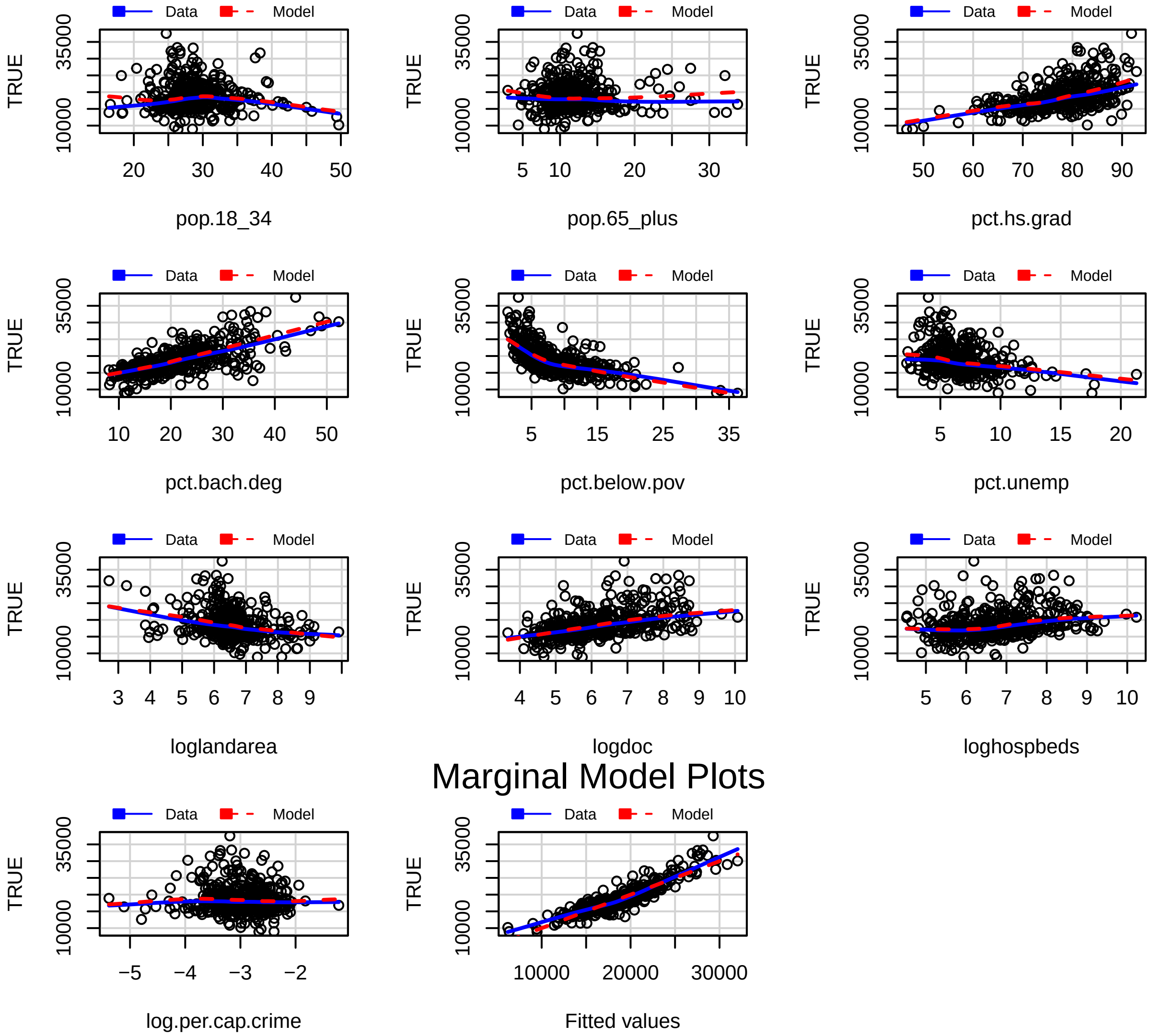
```
##        pop.18_34      pop.65_plus      pct.hs.grad     pct.bach.deg
##         1.979952         1.853750         3.820223         4.013215
##    pct.below.pov        pct.unemp      loglandarea           logdoc
##         2.723069         1.721429         1.149748        10.906872
```

```
##        loghospbeds log.per.cap.crime
##           9.410985          1.600834
```

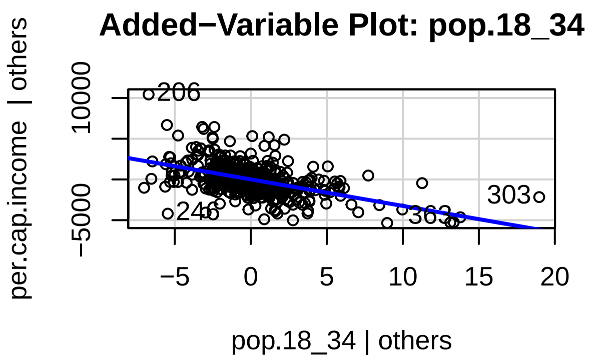
```
par(mfrow = c(2,2))
plot(mod1)
```

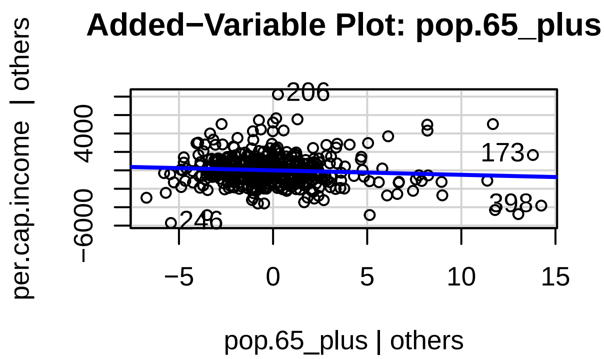


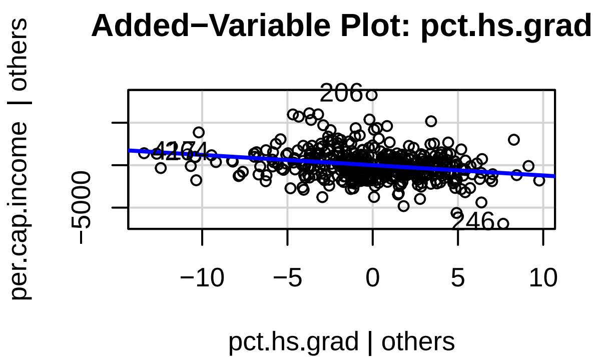

```
mmps(mod1)
```

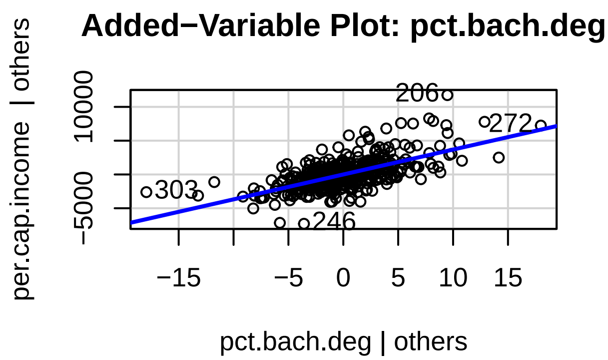



```
par(mfrow = c(2,2))
avPlot(mod1, "pop.18_34")
avPlot(mod1, "pop.65_plus")
avPlot(mod1, "pct.hs.grad")
avPlot(mod1, "pct.bach.deg")
```

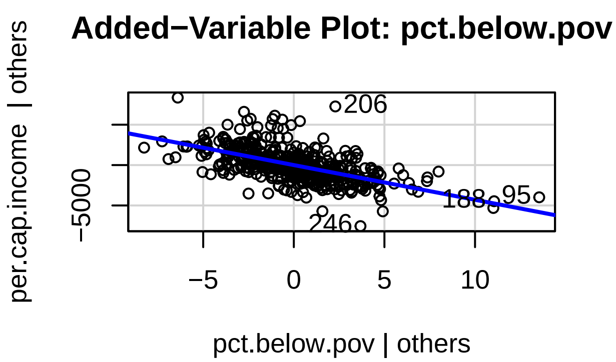


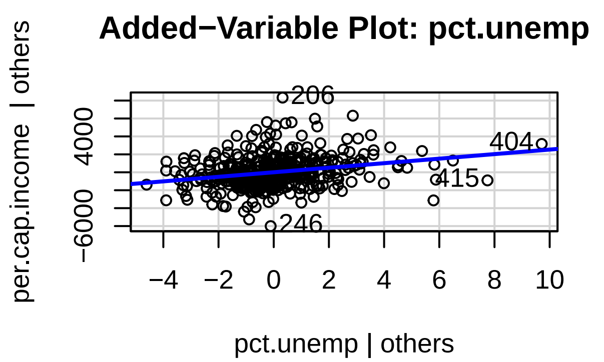


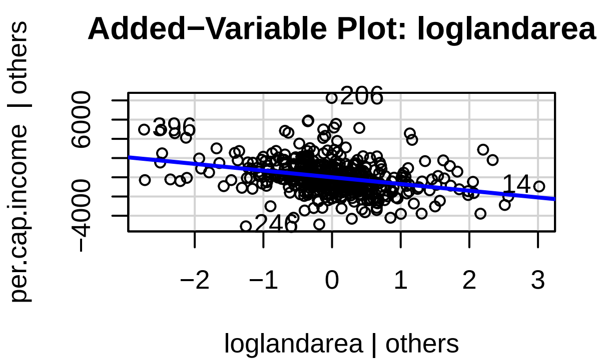


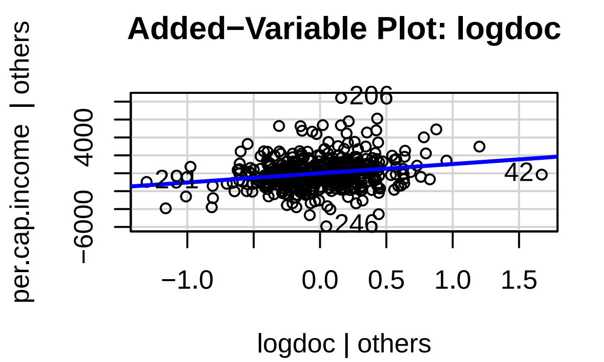



```r
avPlot(mod1, "pct.below.pov")
avPlot(mod1, "pct.unemp")
avPlot(mod1, "loglandarea")
avPlot(mod1, "logdoc")
```

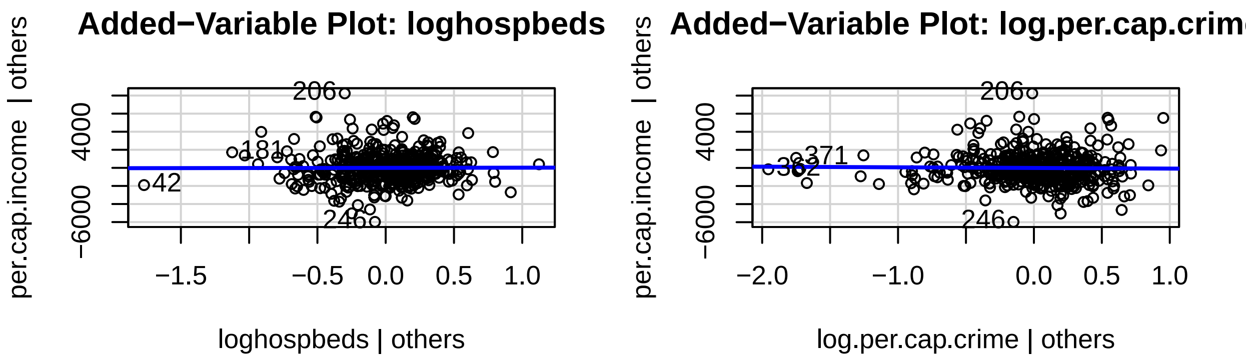








```r
avPlot(mod1, "loghospbeds")
avPlot(mod1, "log.per.cap.crime")
```



log(doctors) and log(hospbeds) have high values of VIF. We will try dropping log(hospbeds), since it does not appear to have a large effect on predicting per capita income according to the added variable plot. We will also try dropping pop.65_plus and log(per capita crime) because they do not appear to have a large effect on predicting per capta income according to their added variable plots.

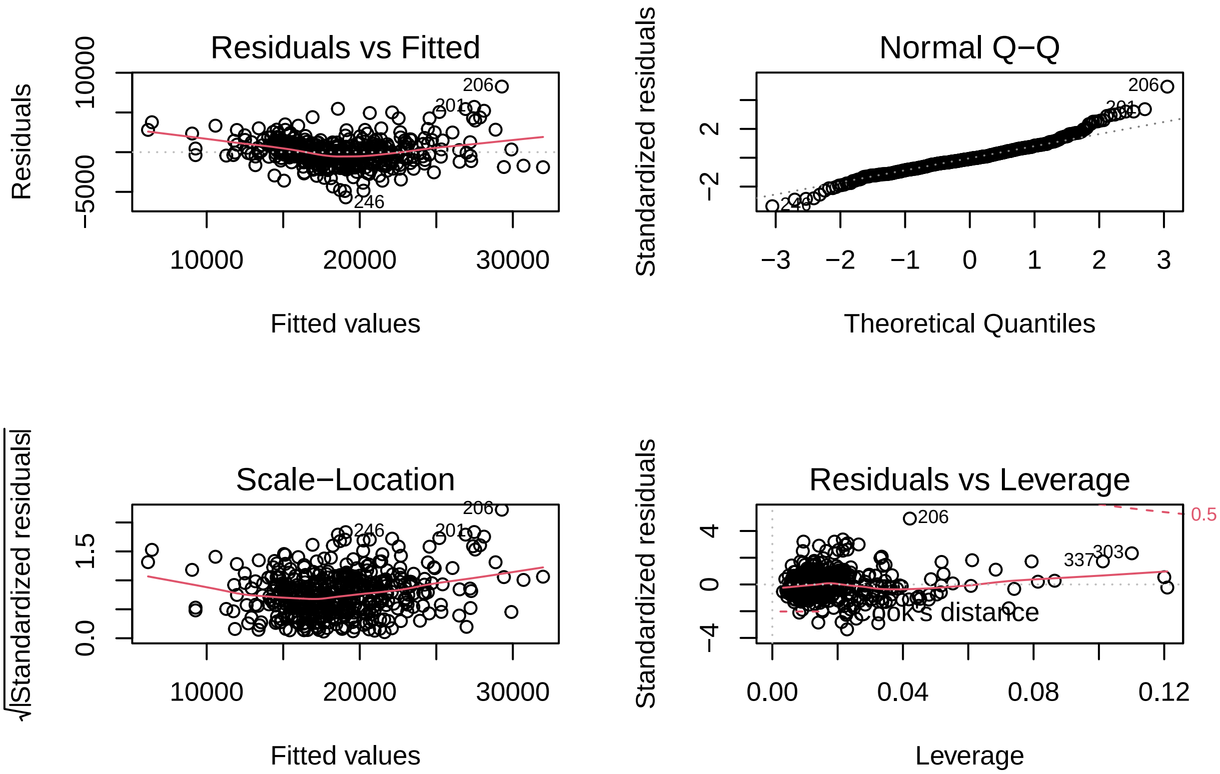
```r
mod2= lm(per.cap.income ~. - region -loghospbeds -pop.65_plus
         -log.per.cap.crime, data = cdi.dat.modelling)
summary(mod2)
```

```
##
## Call:
## lm(formula = per.cap.income ~ . - region - loghospbeds - pop.65_plus -
##     log.per.cap.crime, data = cdi.dat.modelling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5688.4 -1015.1  -123.4   892.2  8260.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28748.60    1944.84  14.782  < 2e-16 ***
## pop.18_34     -300.39      23.21 -12.942  < 2e-16 ***
## pct.hs.grad   -116.80      22.60  -5.168 3.63e-07 ***
## pct.bach.deg   371.01      19.31  19.214  < 2e-16 ***
## pct.below.pov -427.27      26.28 -16.258  < 2e-16 ***
## pct.unemp      251.44      45.47   5.530 5.56e-08 ***
## loglandarea   -683.89      99.76  -6.855 2.47e-11 ***
## logdoc        1000.90      83.92  11.926  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1713 on 432 degrees of freedom
## Multiple R-squared:  0.8248, Adjusted R-squared:  0.822
## F-statistic: 290.6 on 7 and 432 DF,  p-value: < 2.2e-16
```

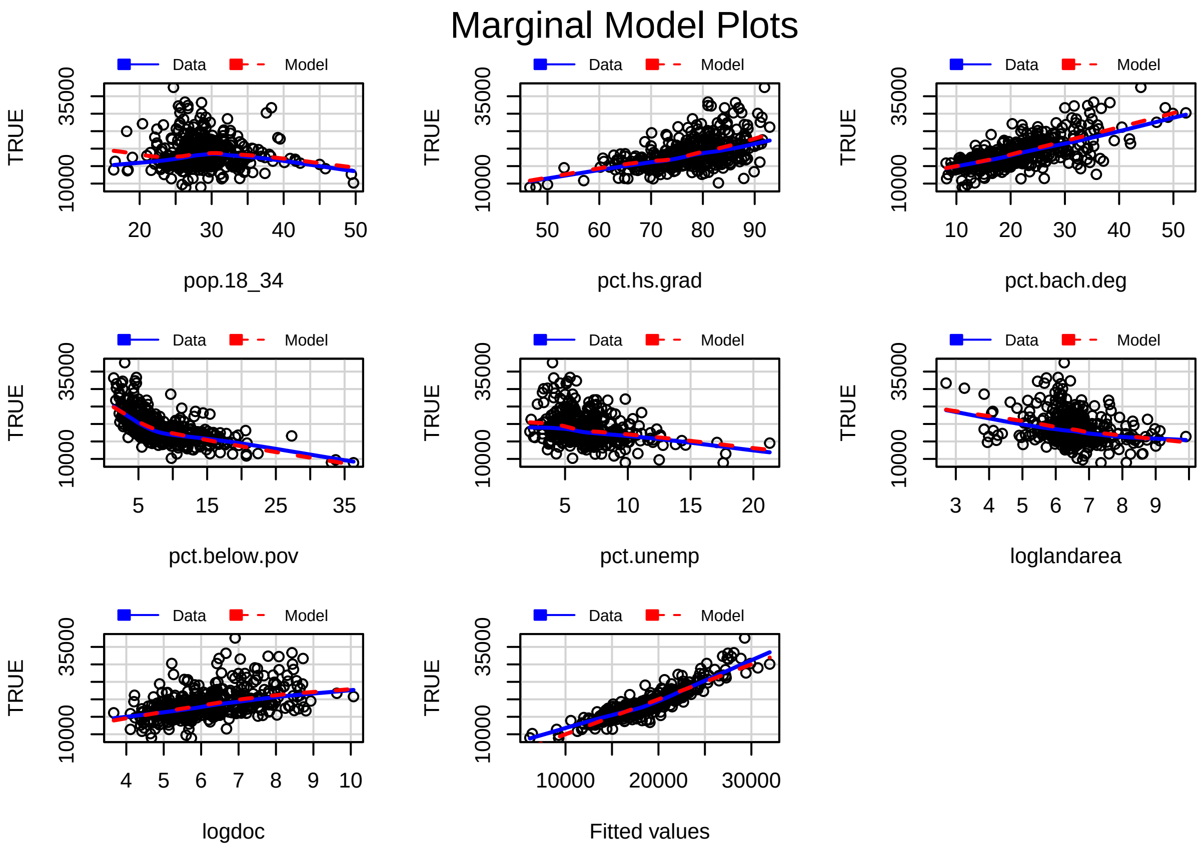
```r
vif(mod2)
```

```
##     pop.18_34   pct.hs.grad  pct.bach.deg pct.below.pov     pct.unemp
##      1.416145      3.763103      3.269565      2.241555      1.691280
##   loglandarea        logdoc
##      1.131867      1.379671
```

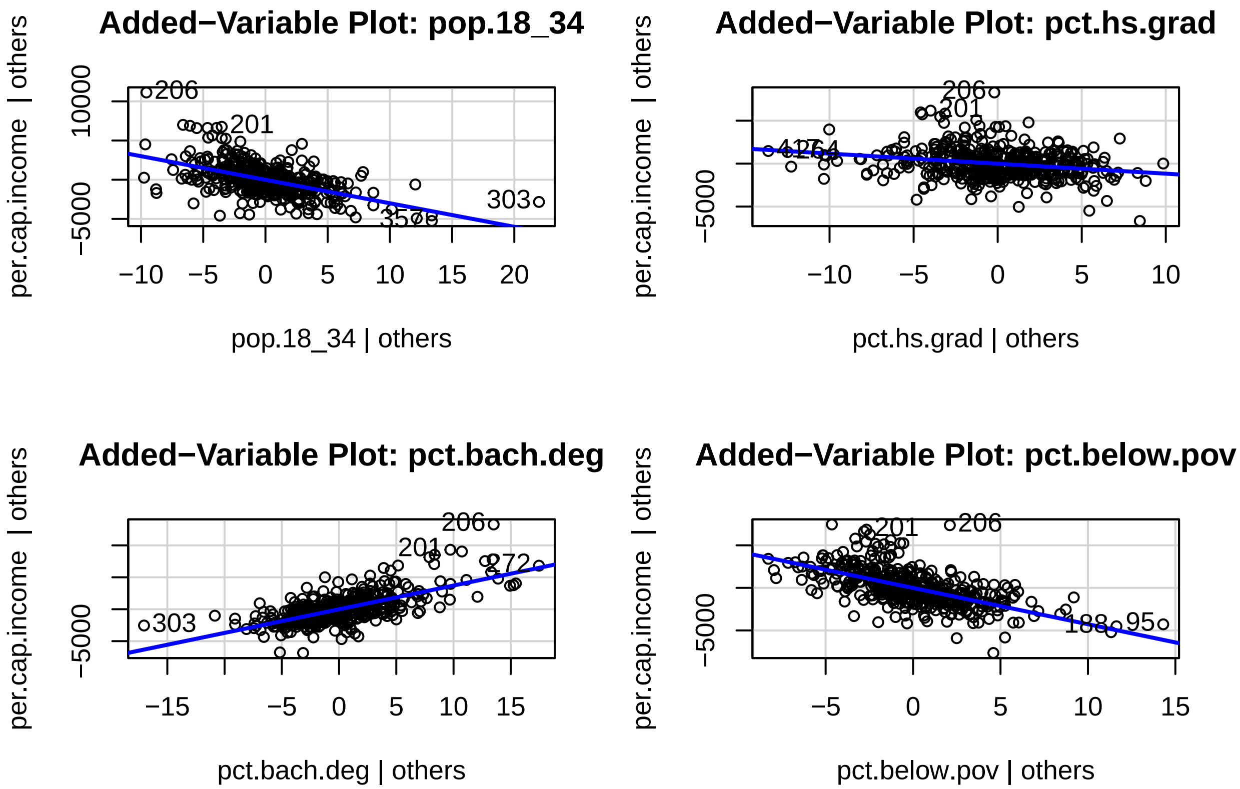
```r
par(mfrow = c(2,2))
plot(mod2)
```



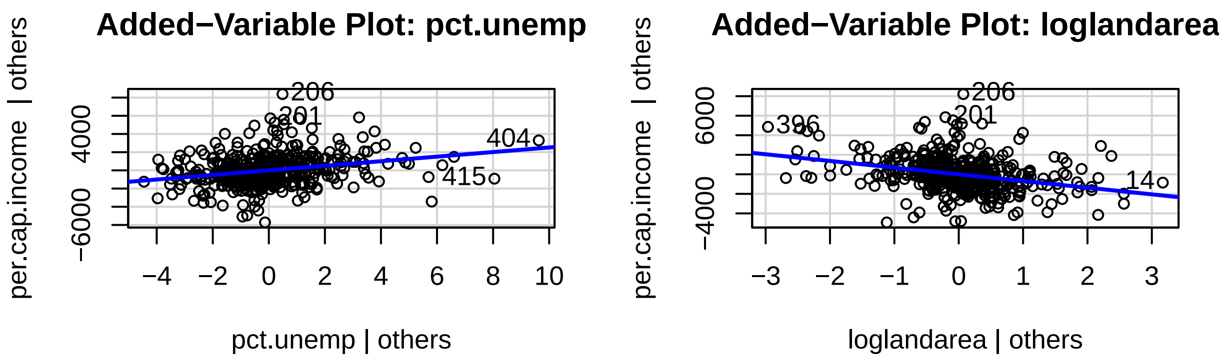

```r
mmps(mod2)
```

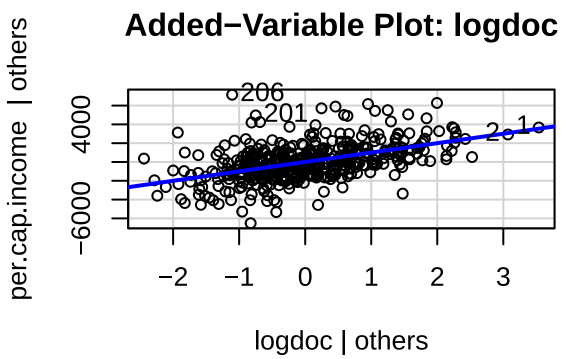



```
par(mfrow = c(2,2))
avPlot(mod2, "pop.18_34")
avPlot(mod2, "pct.hs.grad")
avPlot(mod2, "pct.bach.deg")
avPlot(mod2, "pct.below.pov")
```




```r
avPlot(mod2, "pct.unemp")
avPlot(mod2, "loglandarea")
avPlot(mod2, "logdoc")
```





The diagnostic plots of this model look decent- residuals appear to be randomly scattered with constant variance and mean zero and there do not appear to be any influential observations in the Residuals vs Leverage plot. There are some deviations from the normal Q-Q plot. In the marginal model plots, the fitted loess curves tend to follow the fitted model curves well, which provides evidence that the model is valid. The R-squared of this model is 0.8248, suggesting that the model predicts almost 83% of the variation in per capita income. Now let's try adding region back in. We will start by interacting region with every predictor in model 2.

```
cdi.dat.modelling$region = as.factor(cdi.dat.modelling$region)
mod3= lm(per.cap.income ~.*region -loghospbeds -pop.65_plus
         -log.per.cap.crime -loghospbeds:region
         - pop.65_plus:region -log.per.cap.crime:region,
         data = cdi.dat.modelling)
summary(mod3)
```

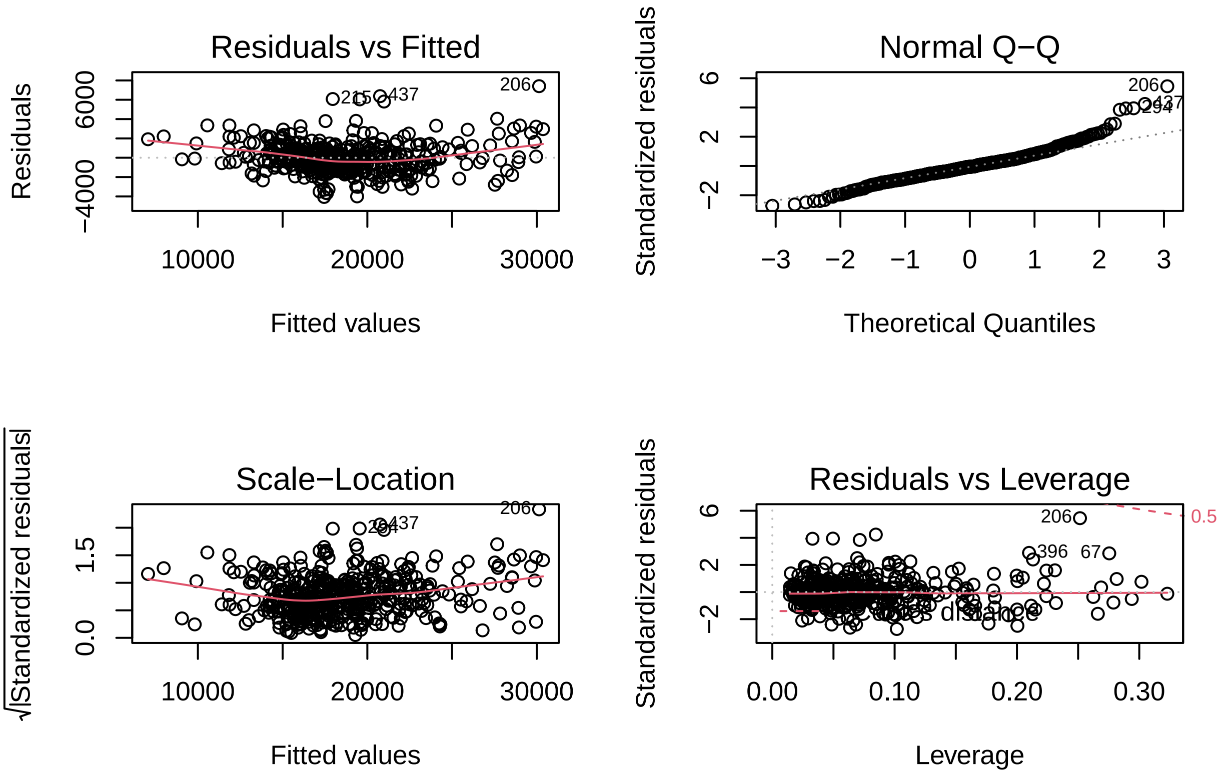
```
##
## Call:
## lm(formula = per.cap.income ~ . * region - loghospbeds - pop.65_plus -
##     log.per.cap.crime - loghospbeds:region - pop.65_plus:region -
##     log.per.cap.crime:region, data = cdi.dat.modelling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4056.1  -897.0   -84.8   700.4  7405.2
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            27308.461   5849.144   4.669 4.12e-06 ***
## pop.18_34               -312.381     53.897  -5.796 1.36e-08 ***
## pct.hs.grad              -82.747     70.594  -1.172 0.241822
## pct.bach.deg             320.180     60.544   5.288 2.02e-07 ***
## pct.below.pov           -451.618     74.989  -6.022 3.84e-09 ***
## pct.unemp                360.802    101.249   3.563 0.000409 ***
## regionNE                7763.075   7403.075   1.049 0.294970
## regionS                -2669.870   6481.255  -0.412 0.680602
## regionW                29543.350   8747.864   3.377 0.000803 ***
## loglandarea             -662.131    313.243  -2.114 0.035139 *
## logdoc                   942.933    192.930   4.887 1.47e-06 ***
## pop.18_34:regionNE      -129.753     76.312  -1.700 0.089836 .
## pop.18_34:regionS         40.044     63.495   0.631 0.528618
## pop.18_34:regionW         -1.841     84.697  -0.022 0.982673
## pct.hs.grad:regionNE     -92.924     91.373  -1.017 0.309766
## pct.hs.grad:regionS       39.501     78.341   0.504 0.614378
## pct.hs.grad:regionW     -366.740     94.955  -3.862 0.000131 ***
## pct.bach.deg:regionNE    215.314     83.429   2.581 0.010206 *
## pct.bach.deg:regionS     -14.386     66.226  -0.217 0.828142
## pct.bach.deg:regionW     168.985     75.279   2.245 0.025317 *
## pct.below.pov:regionNE    12.652    105.333   0.120 0.904452
## pct.below.pov:regionS    161.490     84.313   1.915 0.056145 .
## pct.below.pov:regionW   -251.406    112.319  -2.238 0.025739 *
## pct.unemp:regionNE      -184.710    152.648  -1.210 0.226965
## pct.unemp:regionS       -272.108    136.317  -1.996 0.046584 *
## pct.unemp:regionW       -409.218    140.138  -2.920 0.003693 **
## regionNE:loglandarea      35.169    416.887   0.084 0.932810
## regionS:loglandarea     -119.445    360.428  -0.331 0.740513
```

```
## regionW:loglandarea      214.668    376.398   0.570 0.568773
## regionNE:logdoc          -92.406    274.366  -0.337 0.736443
## regionS:logdoc          -101.199    236.763  -0.427 0.669293
## regionW:logdoc           -84.476    272.308  -0.310 0.756551
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1571 on 408 degrees of freedom
## Multiple R-squared:  0.8608, Adjusted R-squared:  0.8503
## F-statistic: 81.41 on 31 and 408 DF,  p-value: < 2.2e-16
```

```
vif(mod3)
```

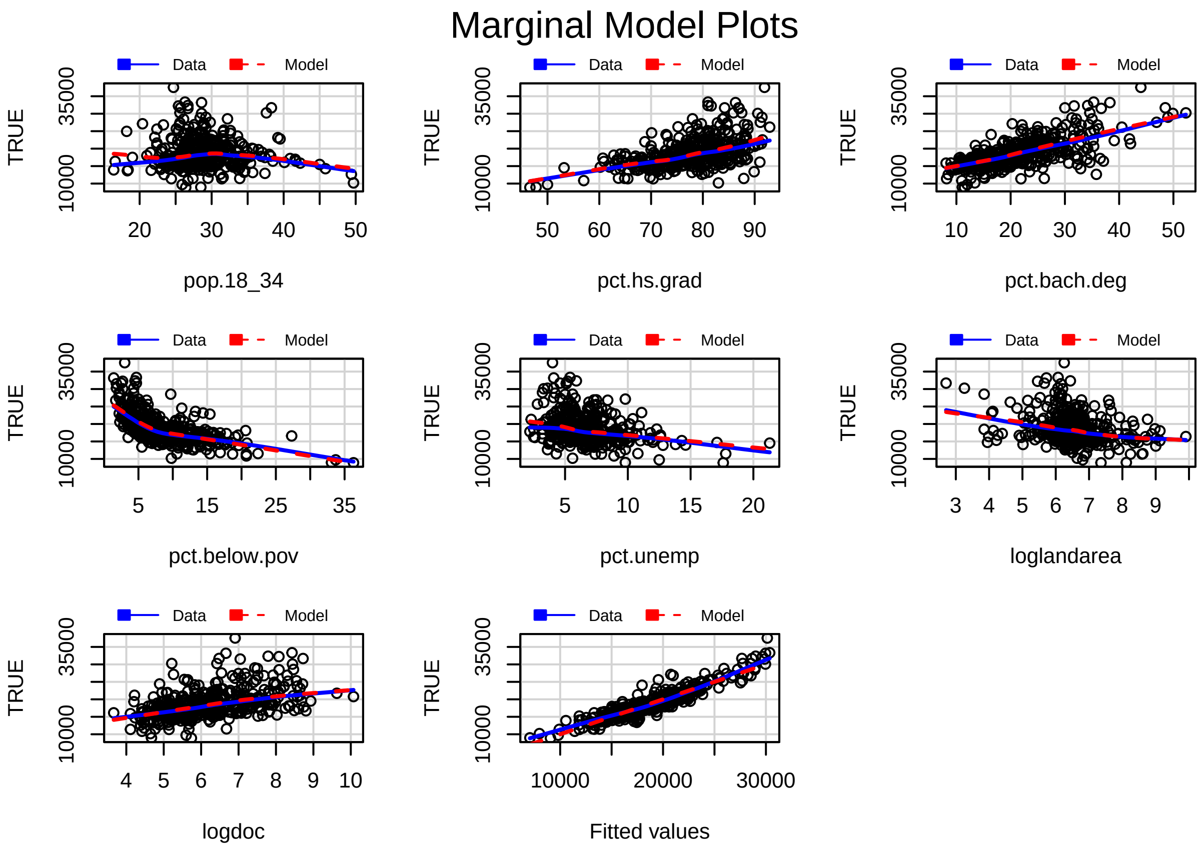
```
##                              GVIF Df GVIF^(1/(2*Df))
## pop.18_34            9.079164e+00  1        3.013165
## pct.hs.grad          4.363895e+01  1        6.605978
## pct.bach.deg         3.821558e+01  1        6.181876
## pct.below.pov        2.169820e+01  1        4.658133
## pct.unemp            9.970229e+00  1        3.157567
## region               1.051674e+09  3       31.889435
## loglandarea          1.326672e+01  1        3.642350
## logdoc               8.668627e+00  1        2.944253
## pop.18_34:region     8.656877e+05  3        9.762481
## pct.hs.grad:region   5.636531e+08  3       28.741007
## pct.bach.deg:region  1.642299e+05  3        7.400176
## pct.below.pov:region 9.294366e+03  3        4.585323
## pct.unemp:region     1.530451e+04  3        4.982759
## region:loglandarea   1.530409e+06  3       10.734980
## region:logdoc        1.875708e+05  3        7.565905
```

```
par(mfrow = c(2,2))
plot(mod3)
```




```
mmps(mod3)
```

```
## Warning in mmps(mod3): Interactions and/or factors skipped
```



Now we will drop the interactions in which none of the region factors significantly interact (at 0.05 level) with the other predictors. Thus, we will drop pop.18_34:region, log(landarea):region, and log(doctors):region.

```
mod4= lm(per.cap.income ~.*region -loghospbeds
         -pop.65_plus -log.per.cap.crime
         -loghospbeds*region - pop.65_plus*region
         -log.per.cap.crime*region -pop.18_34:region
         -region:loglandarea - region:logdoc,
         data = cdi.dat.modelling)
summary(mod4)
```

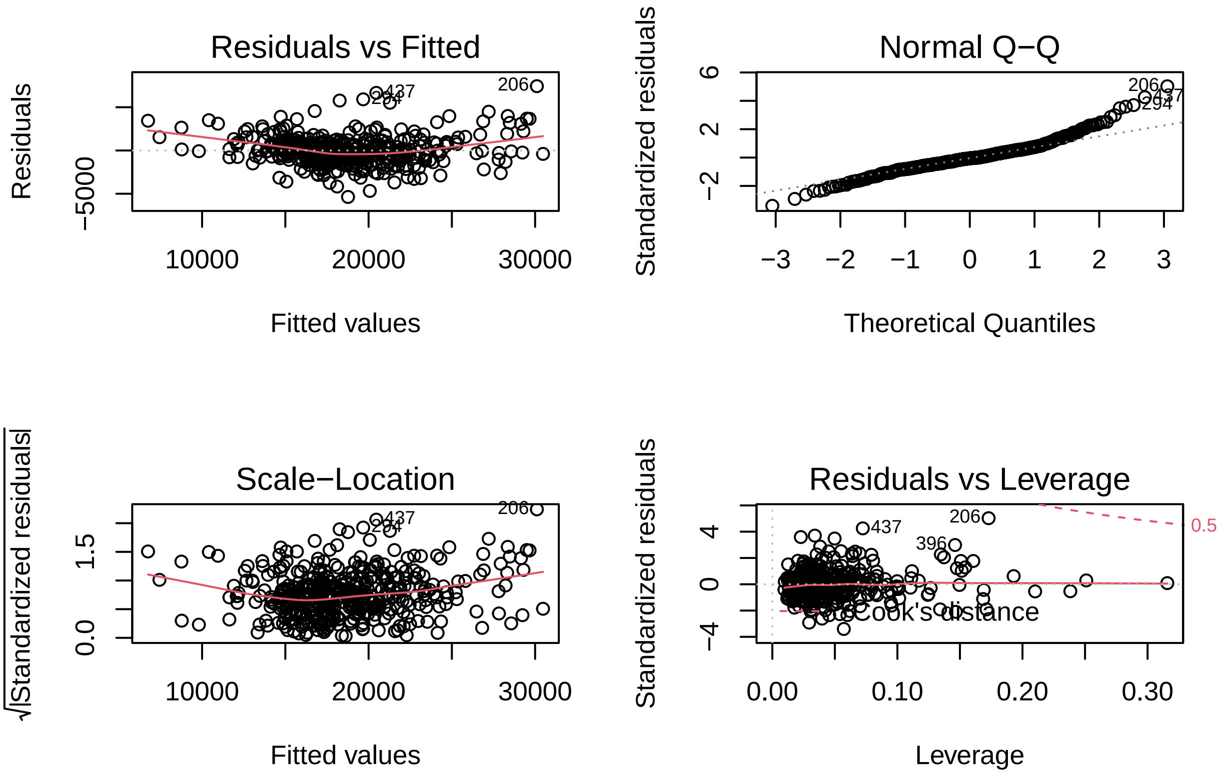
```
##
## Call:
## lm(formula = per.cap.income ~ . * region - loghospbeds - pop.65_plus -
##     log.per.cap.crime - loghospbeds * region - pop.65_plus *
##     region - log.per.cap.crime * region - pop.18_34:region -
##     region:loglandarea - region:logdoc, data = cdi.dat.modelling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5371.1  -889.0   -78.7   768.9  7433.5
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              29259.036   2096.017  13.959  < 2e-16 ***
## pop.18_34                 -296.182     23.091 -12.827  < 2e-16 ***
## pct.hs.grad               -112.909     28.275  -3.993 7.69e-05 ***
## pct.bach.deg               320.208     39.042   8.202 2.92e-15 ***
```

```
## pct.below.pov          -486.109     56.145  -8.658  < 2e-16 ***
## pct.unemp               349.084     99.927   3.493 0.000528 ***
## loglandarea            -644.272    114.426  -5.630 3.29e-08 ***
## logdoc                  979.643     84.786  11.554  < 2e-16 ***
## pct.hs.grad:regionNE    -27.045     22.744  -1.189 0.235063
## pct.hs.grad:regionS       2.662     18.588   0.143 0.886211
## pct.hs.grad:regionW     -32.867     20.790  -1.581 0.114649
## pct.bach.deg:regionNE   146.986     49.929   2.944 0.003421 **
## pct.bach.deg:regionS      6.671     42.395   0.157 0.875036
## pct.bach.deg:regionW    111.533     51.663   2.159 0.031425 *
## pct.below.pov:regionNE    4.728     79.452   0.060 0.952573
## pct.below.pov:regionS   151.665     59.276   2.559 0.010858 *
## pct.below.pov:regionW    62.241     83.449   0.746 0.456169
## pct.unemp:regionNE     -171.281    151.252  -1.132 0.258103
## pct.unemp:regionS      -350.505    132.797  -2.639 0.008614 **
## pct.unemp:regionW       -74.832    124.510  -0.601 0.548155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1626 on 420 degrees of freedom
## Multiple R-squared:  0.8465, Adjusted R-squared:  0.8395
## F-statistic: 121.9 on 19 and 420 DF,  p-value: < 2.2e-16
```

```
vif(mod4)
```

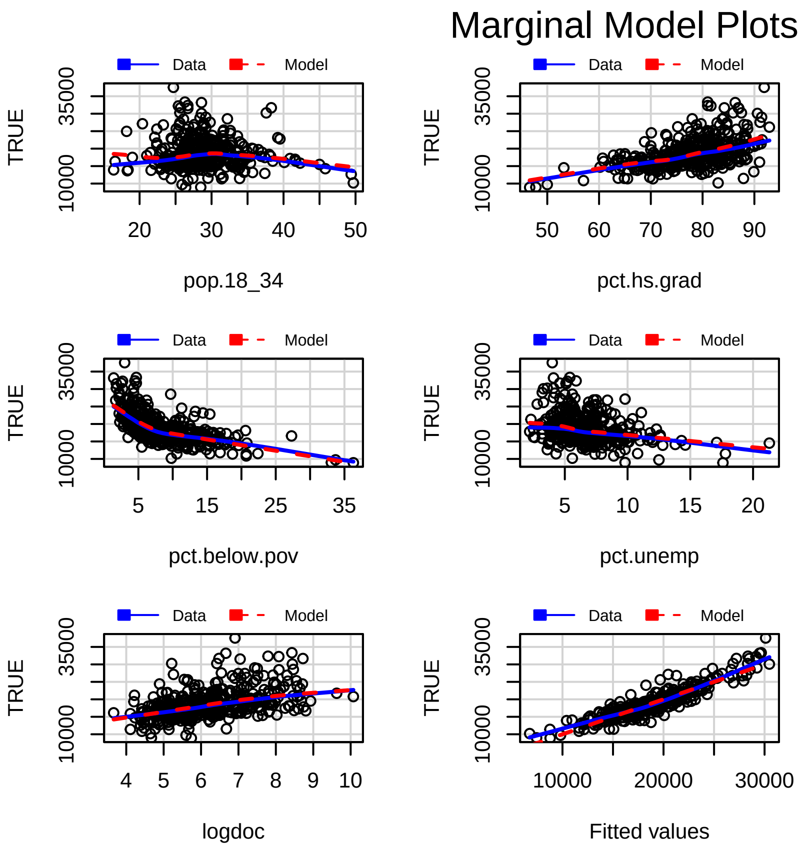
```
##                               GVIF Df GVIF^(1/(2*Df))
## pop.18_34             1.554818e+00  1        1.246923
## pct.hs.grad           6.531744e+00  1        2.555728
## pct.bach.deg          1.482682e+01  1        3.850561
## pct.below.pov         1.134809e+01  1        3.368692
## pct.unemp             9.060668e+00  1        3.010094
## loglandarea           1.651679e+00  1        1.285177
## logdoc                1.562000e+00  1        1.249800
## pct.hs.grad:region    1.223398e+05  3        7.045762
## pct.bach.deg:region   1.526584e+04  3        4.980659
## pct.below.pov:region  1.437926e+03  3        3.359614
## pct.unemp:region      8.001394e+03  3        4.472266
```

```
par(mfrow = c(2,2))
plot(mod4)
```




```
mmps(mod4)
```

```
## Warning in mmps(mod4): Interactions and/or factors skipped
```



Now the interaction with the percent high school graduates is not signficant, so we will drop this interactions too.

```
mod5= lm(per.cap.income ~.*region -loghospbeds
         -pop.65_plus -log.per.cap.crime
         -loghospbeds*region - pop.65_plus*region
         -log.per.cap.crime*region -pop.18_34:region
         -region:loglandarea - region:logdoc
         -pct.hs.grad:region, data = cdi.dat.modelling)
summary(mod5)
```

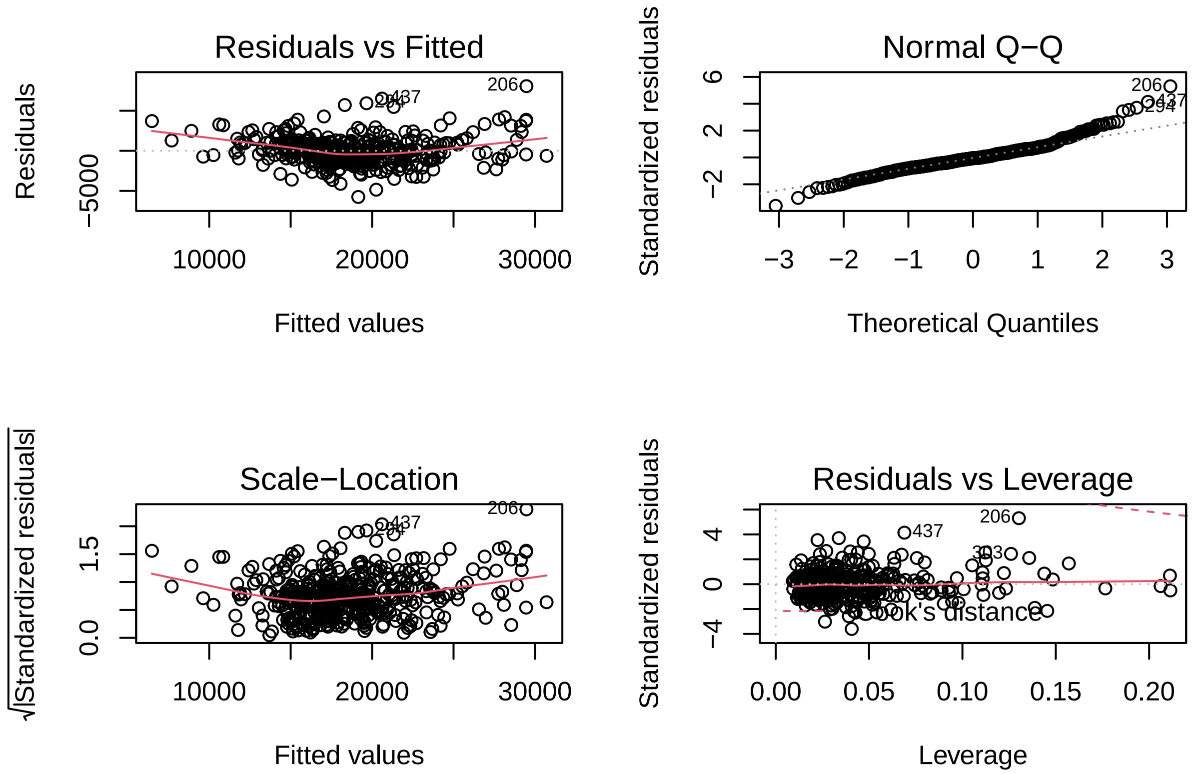
```
##
## Call:
## lm(formula = per.cap.income ~ . * region - loghospbeds - pop.65_plus -
##     log.per.cap.crime - loghospbeds * region - pop.65_plus *
##     region - log.per.cap.crime * region - pop.18_34:region -
##     region:loglandarea - region:logdoc - pct.hs.grad:region,
##     data = cdi.dat.modelling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5756.5  -936.6   -85.1   808.0  8065.5
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             30022.004   2049.377  14.649  < 2e-16 ***
## pop.18_34                -289.194     22.690 -12.745  < 2e-16 ***
## pct.hs.grad              -131.152     23.685  -5.537 5.40e-08 ***
```

```
## pct.bach.deg              343.777     23.403  14.689  < 2e-16 ***
## pct.below.pov            -489.843     55.665  -8.800  < 2e-16 ***
## pct.unemp                 400.897     76.394   5.248 2.44e-07 ***
## loglandarea              -684.294    113.148  -6.048 3.23e-09 ***
## logdoc                    972.718     83.396  11.664  < 2e-16 ***
## pct.bach.deg:regionNE      93.222     20.719   4.499 8.82e-06 ***
## pct.bach.deg:regionS        8.234     17.029   0.484 0.628988
## pct.bach.deg:regionW       33.849     19.161   1.767 0.078030 .
## pct.below.pov:regionNE     -8.576     78.924  -0.109 0.913519
## pct.below.pov:regionS     144.065     59.163   2.435 0.015301 *
## pct.below.pov:regionW      29.330     80.877   0.363 0.717046
## pct.unemp:regionNE       -301.085    106.401  -2.830 0.004881 **
## pct.unemp:regionS        -326.500     90.140  -3.622 0.000328 ***
## pct.unemp:regionW        -157.249    102.606  -1.533 0.126132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1631 on 423 degrees of freedom
## Multiple R-squared:  0.8445, Adjusted R-squared:  0.8386
## F-statistic: 143.6 on 16 and 423 DF,  p-value: < 2.2e-16
```

```
vif(mod5)
```

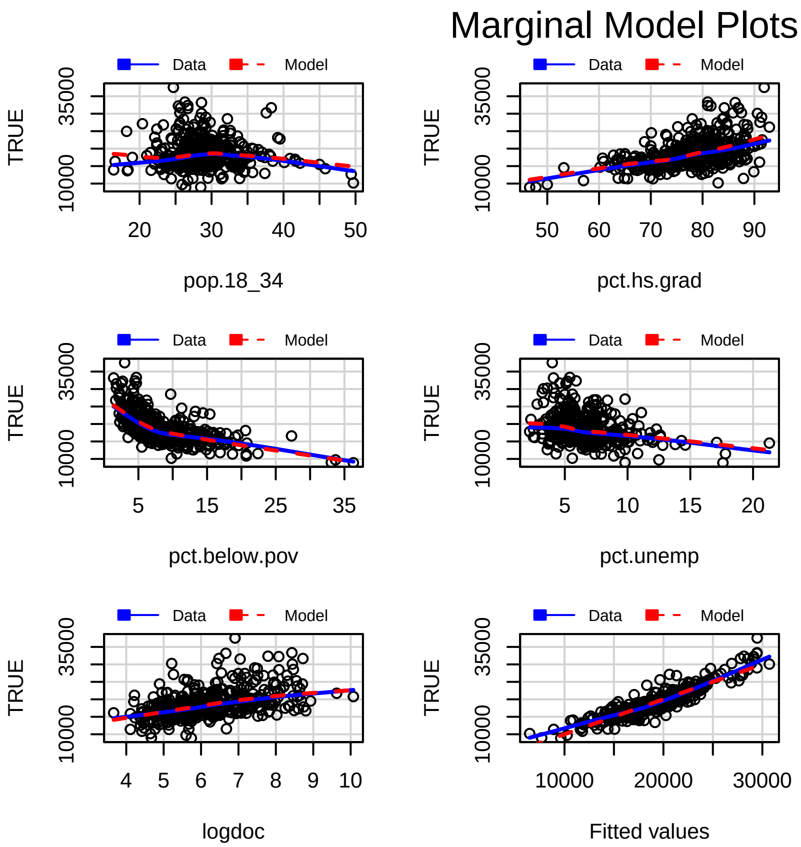
```
##                              GVIF Df GVIF^(1/(2*Df))
## pop.18_34                1.493142  1        1.221942
## pct.hs.grad              4.557977  1        2.134942
## pct.bach.deg             5.298435  1        2.301833
## pct.below.pov           11.093929  1        3.330755
## pct.unemp                5.266715  1        2.294932
## loglandarea              1.606195  1        1.267358
## logdoc                   1.502968  1        1.225956
## pct.bach.deg:region     71.343085  3        2.036535
## pct.below.pov:region  1297.624841  3        3.302617
## pct.unemp:region      1610.072857  3        3.423531
```

```
par(mfrow = c(2,2))
plot(mod5)
```




```
mmps(mod5)
```

```
## Warning in mmps(mod5): Interactions and/or factors skipped
```



The diagnostic plots of Model 5 look adequate, and none of the VIF values are concerning. Now let's compare Model 5 with Model 2 (the best model we found before considering interactions):

```
anova(mod2,mod5)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ (pop.18_34 + pop.65_plus + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + region + loglandarea + logdoc +
##     loghospbeds + log.per.cap.crime) - region - loghospbeds -
##     pop.65_plus - log.per.cap.crime
## Model 2: per.cap.income ~ (pop.18_34 + pop.65_plus + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + region + loglandarea + logdoc +
##     loghospbeds + log.per.cap.crime) * region - loghospbeds -
##     pop.65_plus - log.per.cap.crime - loghospbeds * region -
##     pop.65_plus * region - log.per.cap.crime * region - pop.18_34:region -
##     region:loglandarea - region:logdoc - pct.hs.grad:region
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    432 1267158222
## 2    423 1124708880  9 142449342 5.9528 7.432e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(mod2, mod5)
```

```
##      df      AIC
## mod2  9 7810.904
## mod5 18 7776.433
```

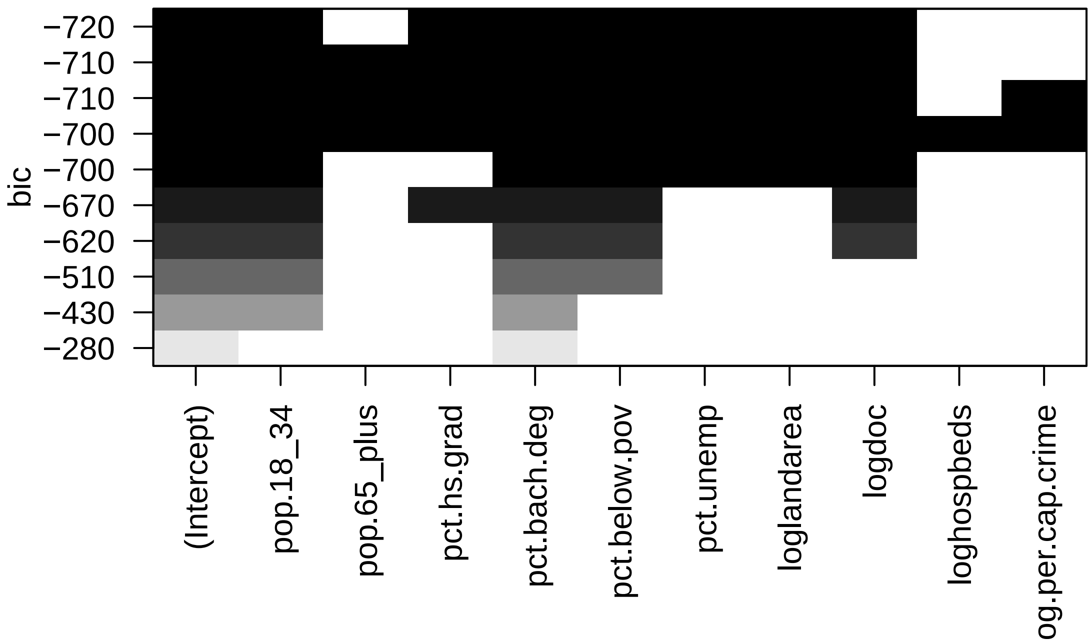
```
BIC(mod2, mod5)
```

```
##      df      BIC
## mod2  9 7847.685
## mod5 18 7849.995
```

According to the F-test and AIC, the model that includes interactions with region (Model 5) is the better model. Now, we will try some formal variable selection methods and compare them to Model 5.

We will start with all subsets regression and consider all variables except region.

```
all.subset = regsubsets(per.cap.income ~.-region, data = cdi.dat.modelling, nvmax = 10)
plot(all.subset)
```




```
best.subset = which.min(summary(all.subset)$bic)
coef(all.subset, best.subset)
```

```
##   (Intercept)     pop.18_34   pct.hs.grad  pct.bach.deg pct.below.pov
##    28748.6035     -300.3892     -116.8039      371.0053     -427.2673
##     pct.unemp   loglandarea        logdoc
##      251.4416     -683.8873     1000.9013
```

```
mod6 = lm(per.cap.income ~ pop.18_34 + pct.bach.deg +
            pct.below.pov + pct.unemp + loglandarea +
            logdoc, data = cdi.dat.modelling)
summary(mod6)
```

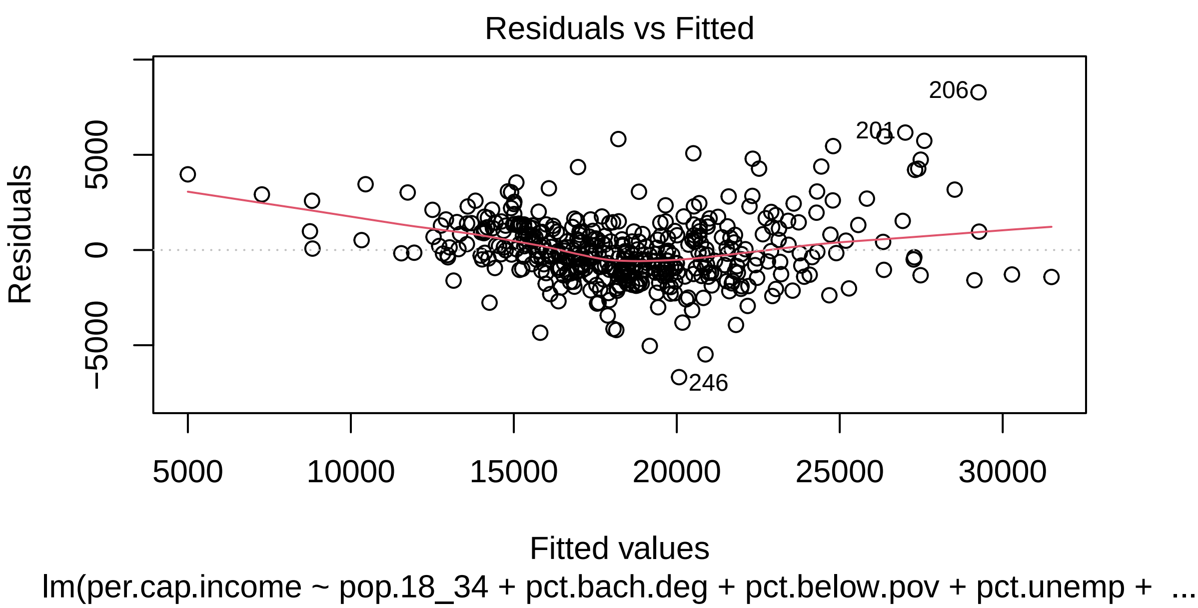
```
##
## Call:
## lm(formula = per.cap.income ~ pop.18_34 + pct.bach.deg + pct.below.pov +
##     pct.unemp + loglandarea + logdoc, data = cdi.dat.modelling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6676.3 -1054.7  -163.4   951.7  8283.0
##
```

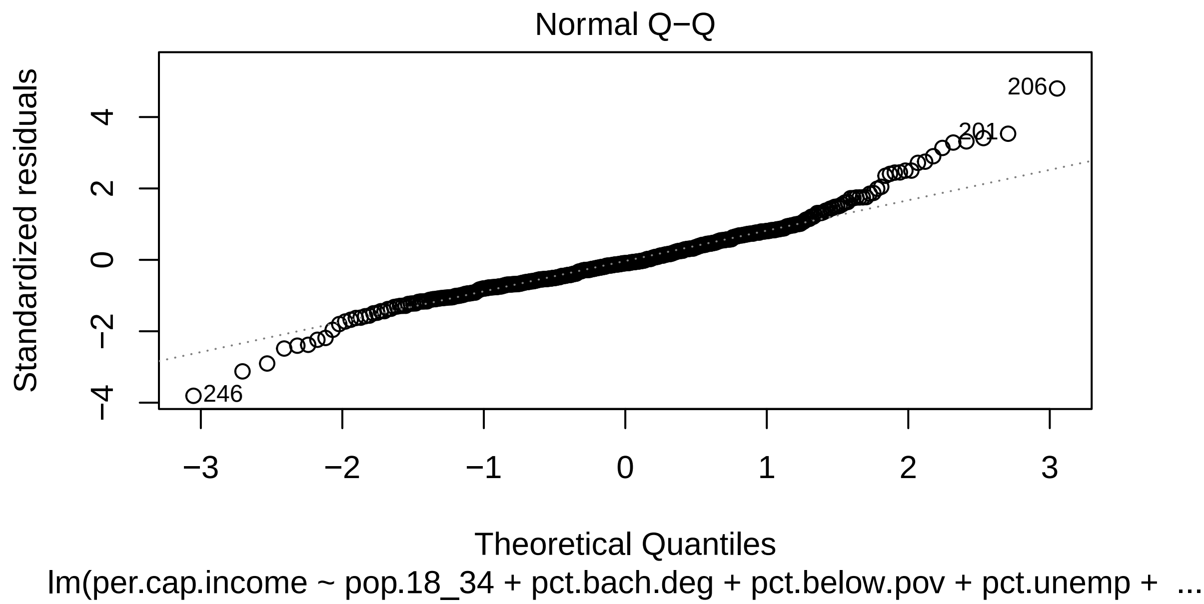
```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    20291.10    1081.32  18.765  < 2e-16 ***
## pop.18_34       -305.46      23.87 -12.798  < 2e-16 ***
## pct.bach.deg     318.79      16.94  18.823  < 2e-16 ***
## pct.below.pov   -350.68      22.34 -15.699  < 2e-16 ***
## pct.unemp        312.45      45.19   6.914 1.70e-11 ***
## loglandarea     -804.17      99.85  -8.054 7.83e-15 ***
## logdoc          1059.00      85.60  12.372  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1763 on 433 degrees of freedom
## Multiple R-squared:  0.814,  Adjusted R-squared:  0.8114
## F-statistic: 315.8 on 6 and 433 DF,  p-value: < 2.2e-16
```

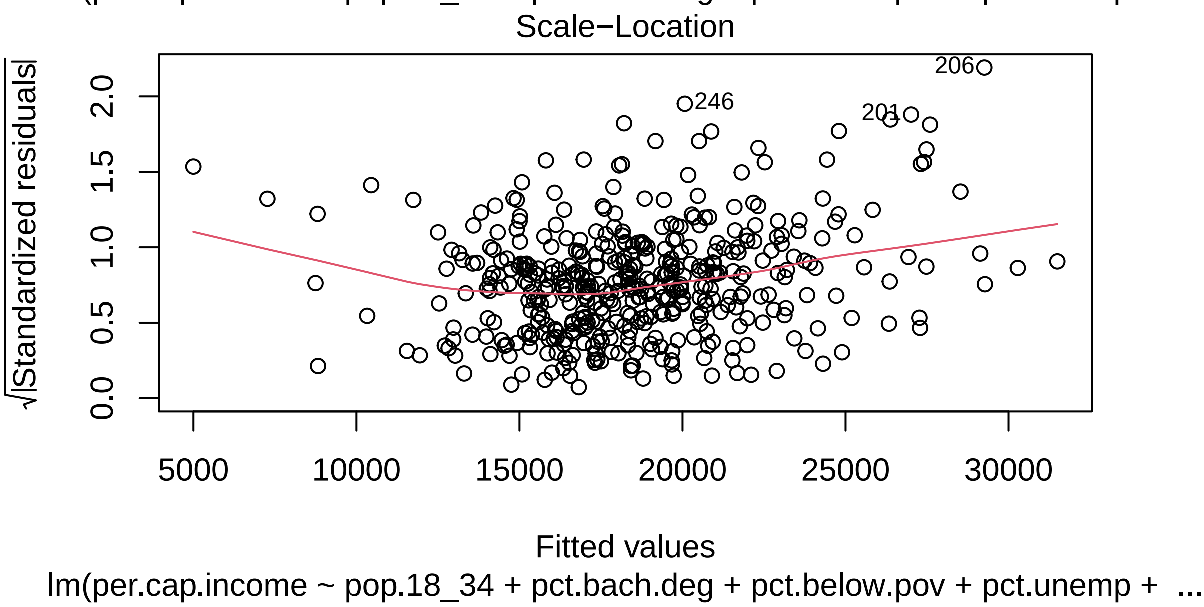
```
vif(mod6)
```

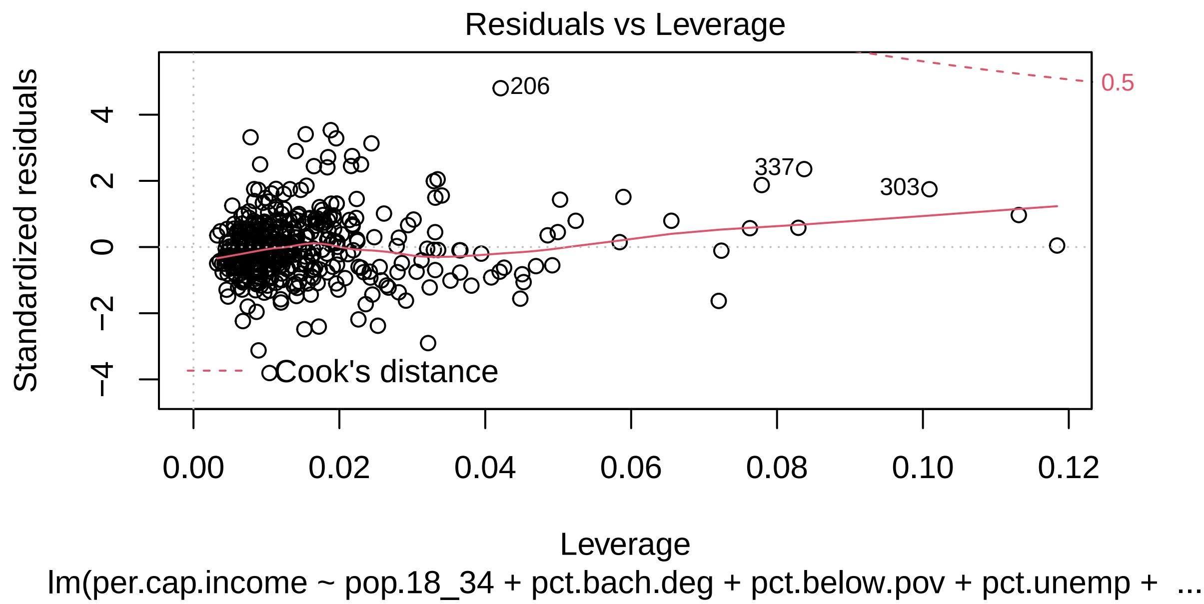
```
##     pop.18_34  pct.bach.deg pct.below.pov     pct.unemp   loglandarea
##      1.413609      2.374313      1.528688      1.577237      1.070246
##        logdoc
##      1.354909
```

```
#par(mfrow = c(2,2))
plot(mod6)
```

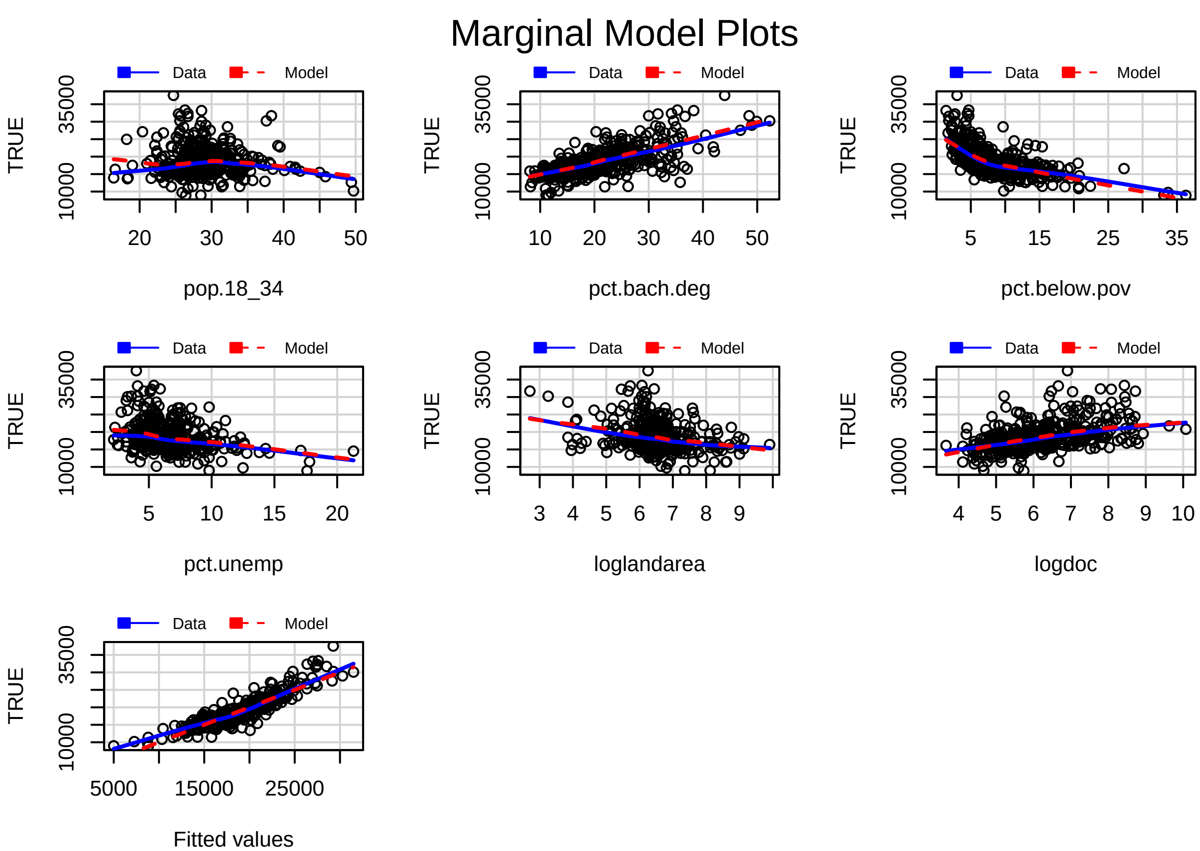



Normal Q−Q
Standardized residuals
Theoretical Quantiles
−3 −2 −1 0 1 2 3
−4 −2 0 2 4
206
201
246
lm(per.cap.income ~ pop.18_34 + pct.bach.deg + pct.below.pov + pct.unemp + ...


Scale−Location
√|Standardized residuals|
Fitted values
5000 10000 15000 20000 25000 30000
0.0 0.5 1.0 1.5 2.0
206
246
201
lm(per.cap.income ~ pop.18_34 + pct.bach.deg + pct.below.pov + pct.unemp + ...




```
mmps(mod6)
```



The model obtained from all subsets regression (Model 6) is nearly the same as our Model 2 (except model 2 contains percent high school graduates). This model looks good: the diagnostic plots look similar to what we have been getting, and there are no exceptionally high VIF values. Now let's try adding region back into this model:

```r
mod7 = lm(per.cap.income ~ pop.18_34 + pct.bach.deg +
            pct.below.pov + pct.unemp + loglandarea +
            logdoc + region + pop.18_34*region +
            pct.bach.deg*region + pct.below.pov:region +
            pct.unemp:region + loglandarea:region +
            logdoc:region , data = cdi.dat.modelling)
summary(mod7)
```

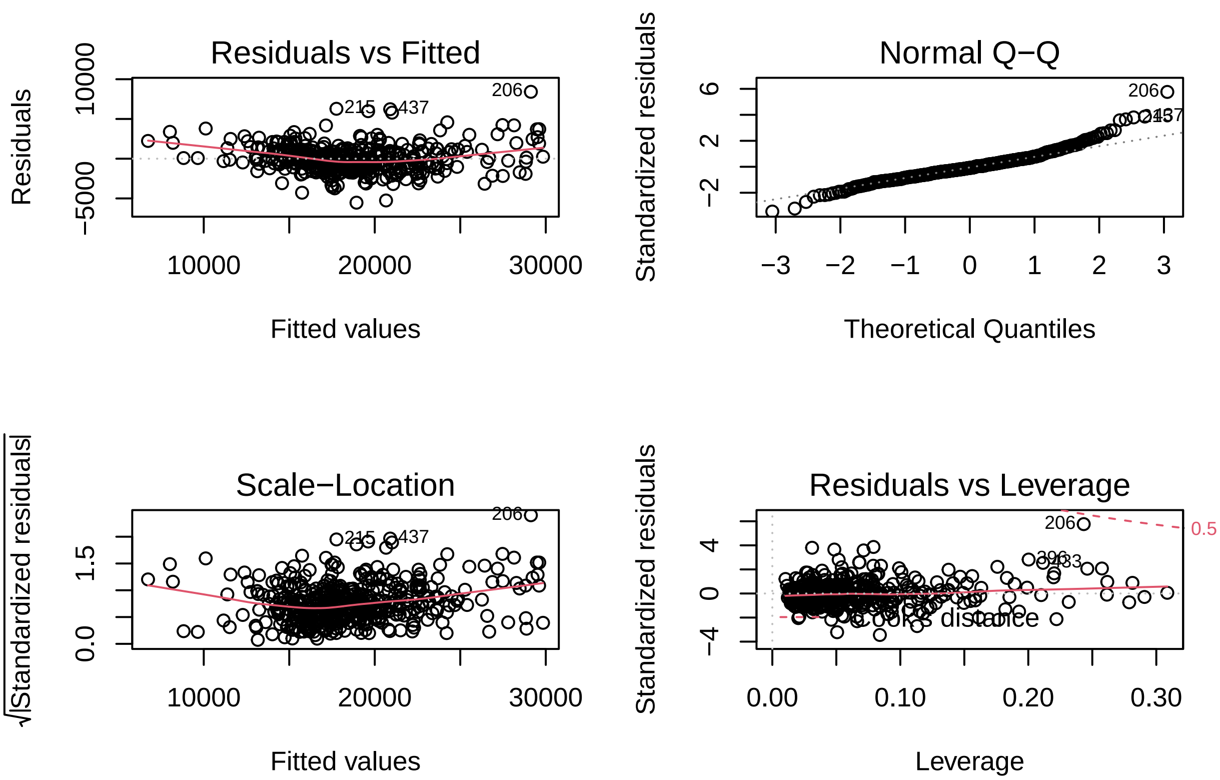
```
##
## Call:
## lm(formula = per.cap.income ~ pop.18_34 + pct.bach.deg + pct.below.pov +
##     pct.unemp + loglandarea + logdoc + region + pop.18_34 * region +
##     pct.bach.deg * region + pct.below.pov:region + pct.unemp:region +
##     loglandarea:region + logdoc:region, data = cdi.dat.modelling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5536.9  -984.4  -136.1   816.5  8412.8
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            21077.261   2606.471   8.087 6.92e-15 ***
## pop.18_34               -299.815     56.432  -5.313 1.77e-07 ***
## pct.bach.deg             270.751     46.413   5.834 1.10e-08 ***
## pct.below.pov           -414.758     72.732  -5.703 2.25e-08 ***
## pct.unemp                376.234    107.253   3.508 0.000501 ***
## loglandarea             -750.403    324.844  -2.310 0.021379 *
## logdoc                  1018.888    194.146   5.248 2.47e-07 ***
## regionNE                2955.442   3889.815   0.760 0.447815
## regionS                  931.588   3289.479   0.283 0.777164
## regionW                -6894.275   3688.967  -1.869 0.062347 .
## pop.18_34:regionNE      -112.289     80.022  -1.403 0.161304
## pop.18_34:regionS         15.476     66.100   0.234 0.815003
## pop.18_34:regionW         39.083     89.397   0.437 0.662204
## pct.bach.deg:regionNE    140.583     63.214   2.224 0.026696 *
## pct.bach.deg:regionS      16.946     52.400   0.323 0.746562
## pct.bach.deg:regionW      73.957     62.952   1.175 0.240751
## pct.below.pov:regionNE    13.961    106.555   0.131 0.895826
## pct.below.pov:regionS    156.104     79.296   1.969 0.049665 *
## pct.below.pov:regionW     60.810    102.443   0.594 0.553107
## pct.unemp:regionNE      -180.892    162.335  -1.114 0.265795
## pct.unemp:regionS       -287.268    144.956  -1.982 0.048170 *
## pct.unemp:regionW          5.909    134.143   0.044 0.964888
## loglandarea:regionNE    -248.676    417.925  -0.595 0.552154
## loglandarea:regionS      -56.632    375.965  -0.151 0.880341
## loglandarea:regionW      246.998    393.911   0.627 0.530979
## logdoc:regionNE           24.916    276.549   0.090 0.928255
## logdoc:regionS          -187.545    243.136  -0.771 0.440936
## logdoc:regionW           269.287    275.023   0.979 0.328082
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1678 on 412 degrees of freedom
## Multiple R-squared:  0.8396, Adjusted R-squared:  0.8291
```

```
## F-statistic: 79.87 on 27 and 412 DF,  p-value: < 2.2e-16
```

```r
vif(mod7)
```

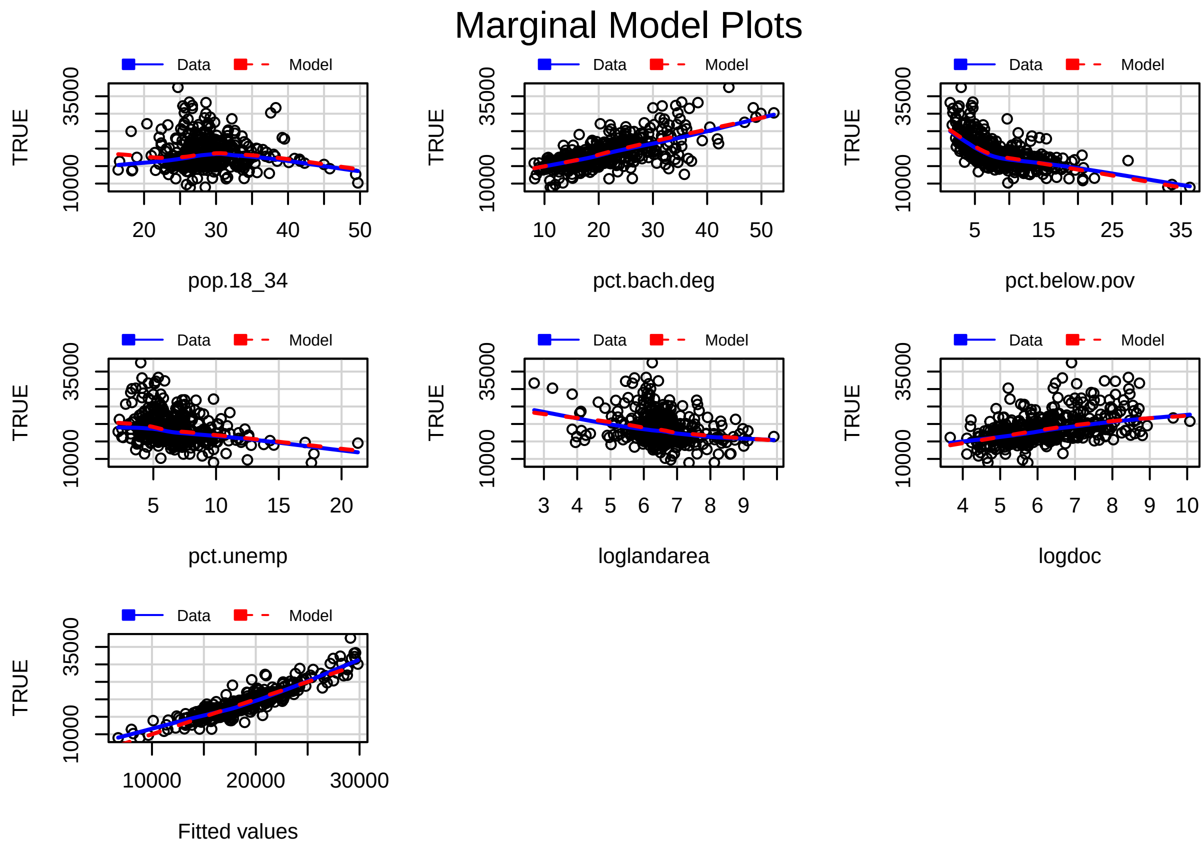
```
##                              GVIF Df GVIF^(1/(2*Df))
## pop.18_34            8.720000e+00  1        2.952965
## pct.bach.deg         1.967584e+01  1        4.435745
## pct.below.pov        1.788271e+01  1        4.228795
## pct.unemp            9.801640e+00  1        3.130757
## loglandarea          1.249992e+01  1        3.535523
## logdoc               7.690721e+00  1        2.773215
## region               1.363777e+07  3       15.456953
## pop.18_34:region     7.726646e+05  3        9.579257
## pct.bach.deg:region  3.552275e+04  3        5.733473
## pct.below.pov:region 4.286776e+03  3        4.030465
## pct.unemp:region     1.070165e+04  3        4.694346
## loglandarea:region   1.201715e+06  3       10.310987
## logdoc:region        1.430474e+05  3        7.231804
```

```r
par(mfrow = c(2,2))
plot(mod7)
```




```r
mmps(mod7)
```

```
## Warning in mmps(mod7): Interactions and/or factors skipped
```



Again, we will drop any interactions that are not significant.

```
mod8 = lm(per.cap.income ~ pop.18_34 + pct.bach.deg +
            pct.below.pov + pct.unemp +
            loglandarea + logdoc +
            pct.bach.deg:region + pct.below.pov:region +
            pct.unemp:region, data = cdi.dat.modelling)
summary(mod8)
```

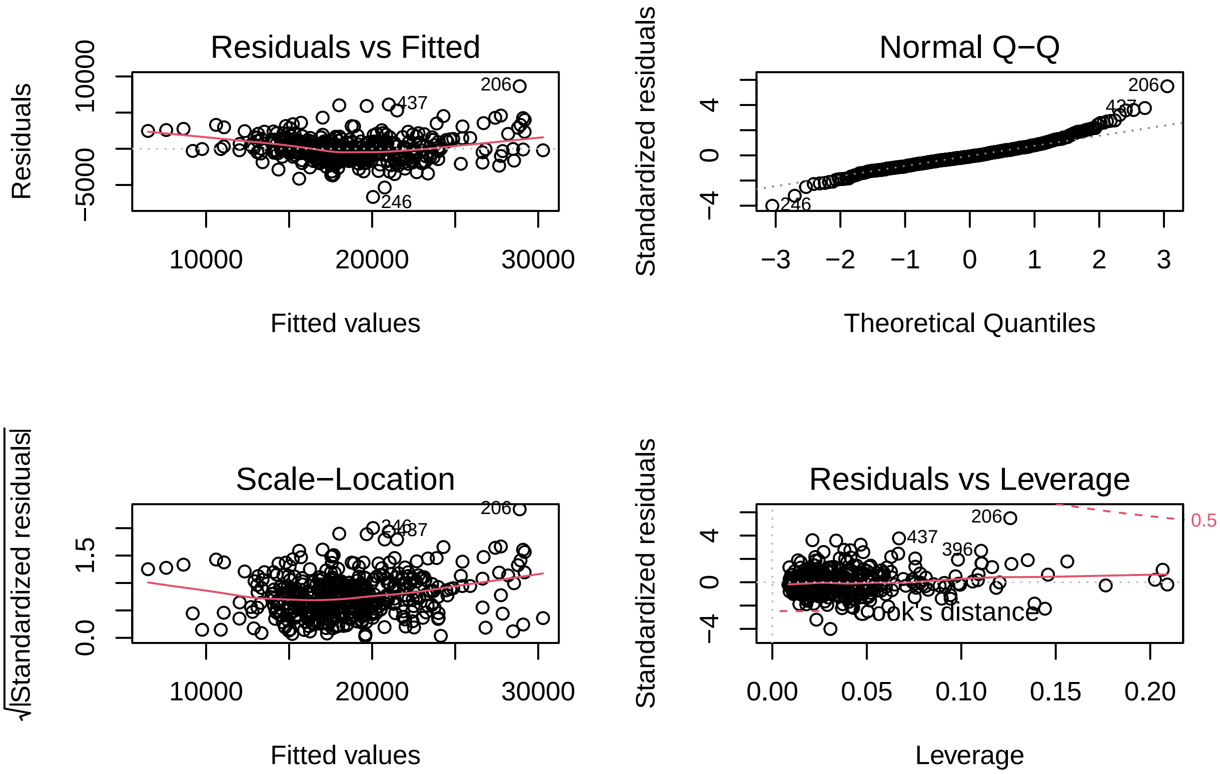
```
##
## Call:
## lm(formula = per.cap.income ~ pop.18_34 + pct.bach.deg + pct.below.pov +
##     pct.unemp + loglandarea + logdoc + pct.bach.deg:region +
##     pct.below.pov:region + pct.unemp:region, data = cdi.dat.modelling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6658.9  -978.8  -144.8   819.2  8657.8
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               20658.15    1197.58  17.250  < 2e-16 ***
## pop.18_34                  -295.55      23.44 -12.608  < 2e-16 ***
## pct.bach.deg                279.37      21.01  13.299  < 2e-16 ***
## pct.below.pov              -414.11      55.81  -7.419 6.48e-13 ***
## pct.unemp                   416.93      78.96   5.280 2.07e-07 ***
## loglandarea                -799.32     115.05  -6.948 1.41e-11 ***
## logdoc                     1046.16      85.17  12.284  < 2e-16 ***
```

```
## pct.bach.deg:regionNE    100.52      21.39   4.700 3.53e-06 ***
## pct.bach.deg:regionS      10.56      17.61   0.600  0.54914
## pct.bach.deg:regionW      19.22      19.63   0.979  0.32819
## pct.below.pov:regionNE   -11.80      81.64  -0.144  0.88518
## pct.below.pov:regionS    151.40      61.18   2.475  0.01373 *
## pct.below.pov:regionW     22.35      83.65   0.267  0.78943
## pct.unemp:regionNE      -263.36     109.83  -2.398  0.01693 *
## pct.unemp:regionS       -279.45      92.82  -3.010  0.00276 **
## pct.unemp:regionW        -83.51     105.24  -0.794  0.42791
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1687 on 424 degrees of freedom
## Multiple R-squared:  0.8332, Adjusted R-squared:  0.8273
## F-statistic: 141.2 on 15 and 424 DF,  p-value: < 2.2e-16
```

```
vif(mod8)
```

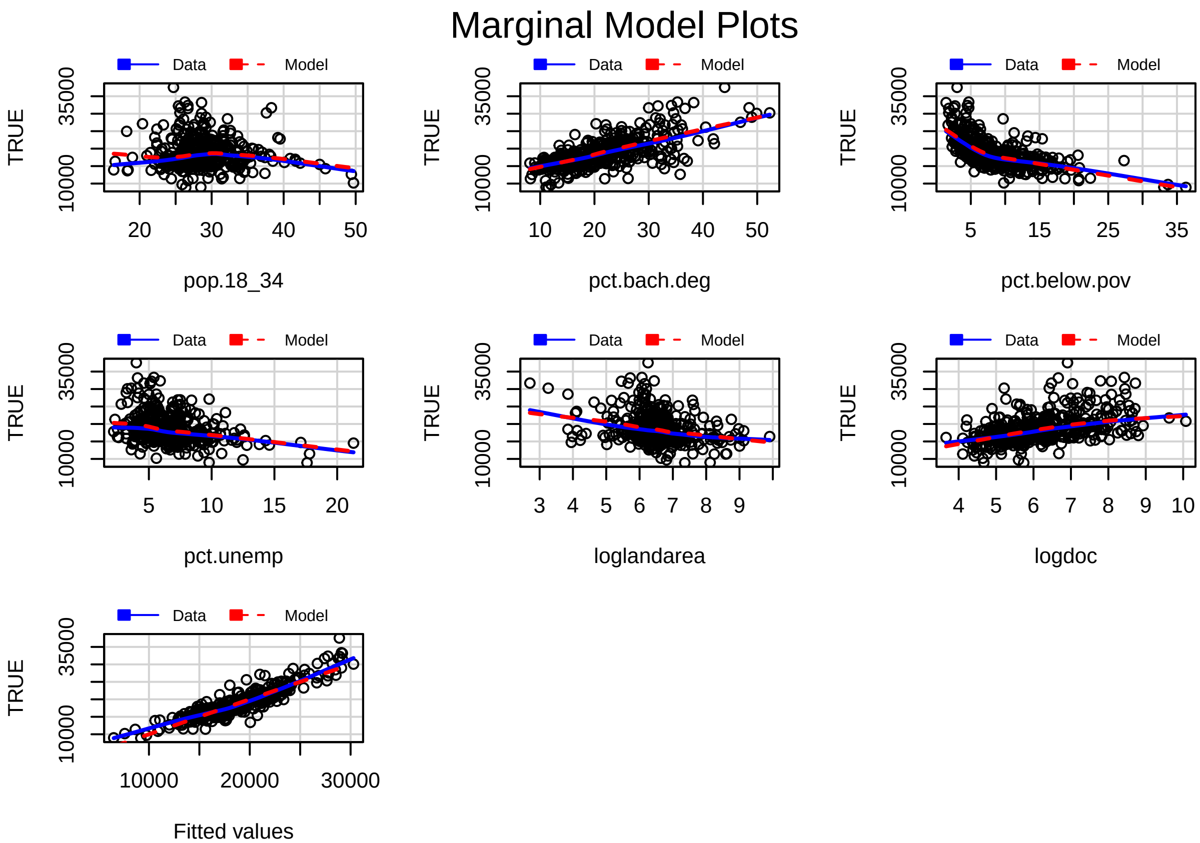
```
##                             GVIF Df GVIF^(1/(2*Df))
## pop.18_34               1.489321  1        1.220378
## pct.bach.deg            3.989720  1        1.997428
## pct.below.pov          10.424175  1        3.228649
## pct.unemp               5.259152  1        2.293284
## loglandarea             1.552061  1        1.245817
## logdoc                  1.464955  1        1.210353
## pct.bach.deg:region    68.266374  3        2.021628
## pct.below.pov:region 1294.927304  3        3.301472
## pct.unemp:region     1580.669692  3        3.413031
```

```
par(mfrow = c(2,2))
plot(mod8)
```




```r
mmps(mod8)
```

```
## Warning in mmps(mod8): Interactions and/or factors skipped
```



The diagnostic plots look good, and there are no dangerously high VIF values. Now let's compare model 6 (no interactions) to model 8.

```
anova(mod6,mod8)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ pop.18_34 + pct.bach.deg + pct.below.pov + pct.unemp +
##     loglandarea + logdoc
## Model 2: per.cap.income ~ pop.18_34 + pct.bach.deg + pct.below.pov + pct.unemp +
##     loglandarea + logdoc + pct.bach.deg:region + pct.below.pov:region +
##     pct.unemp:region
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    433 1345484610
## 2    424 1206238232  9 139246377 5.4384 4.465e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(mod6,mod8)
```

```
##      df      AIC
## mod6  8 7835.294
## mod8 17 7805.225
```

```
BIC(mod6,mod8)
```

```
##      df      BIC
## mod6  8 7867.988
## mod8 17 7874.700
```

According to the F-test and AIC, Model 8 is the better model. Now let's compare it to Model 5, the final model we landed on before performing formal variable selection. Note that the only difference between these models is that Model 5 includes percent high school graduates and Model 8 does not.

```
anova(mod5, mod8)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ (pop.18_34 + pop.65_plus + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + region + loglandarea + logdoc +
##     loghospbeds + log.per.cap.crime) * region - loghospbeds -
##     pop.65_plus - log.per.cap.crime - loghospbeds * region -
##     pop.65_plus * region - log.per.cap.crime * region - pop.18_34:region -
##     region:loglandarea - region:logdoc - pct.hs.grad:region
## Model 2: per.cap.income ~ pop.18_34 + pct.bach.deg + pct.below.pov + pct.unemp +
##     loglandarea + logdoc + pct.bach.deg:region + pct.below.pov:region +
##     pct.unemp:region
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    423 1124708880
## 2    424 1206238232 -1 -81529352 30.663 5.399e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(mod5, mod8)
```

```
##      df      AIC
## mod5 18 7776.433
## mod8 17 7805.225
```

```
BIC(mod5, mod8)
```

```
##      df      BIC
## mod5 18 7849.995
## mod8 17 7874.700
```

According to the F-test, Model 8 is an improvement over Model 5. However, AIC and BIC prefer Model 5.

Now we will try stepwise regression, and compare the results to Models 5 and 8:

```
# stepwise AIC and BIC
mod9 <- stepAIC(lm(per.cap.income ~ . -region,
                   data=cdi.dat.modelling),
                scope=list(lower = ~ 1, upper = ~ .),
k=2, trace=F)
mod9
```

```
##
## Call:
## lm(formula = per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + loglandarea +
##     logdoc, data = cdi.dat.modelling)
##
## Coefficients:
##   (Intercept)      pop.18_34    pop.65_plus    pct.hs.grad   pct.bach.deg
##      30366.74        -325.47         -47.89        -121.16         368.06
## pct.below.pov      pct.unemp    loglandarea         logdoc
##       -433.08         254.30        -698.07        1034.28
```

```
mod10 <- stepAIC(lm(per.cap.income ~ .-region,
                    data=cdi.dat.modelling),
                 scope=list(lower = ~ 1, upper = ~ .),
k=log(dim(cdi.dat.modelling)[1]), trace=F)
mod10
```

```
##
## Call:
## lm(formula = per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + loglandarea + logdoc, data = cdi.dat.modelling)
##
## Coefficients:
##   (Intercept)      pop.18_34    pct.hs.grad   pct.bach.deg  pct.below.pov
##       28748.6         -300.4         -116.8          371.0         -427.3
##     pct.unemp    loglandarea         logdoc
##         251.4         -683.9         1000.9
```

```
anova(mod9,mod10)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + loglandarea + logdoc
## Model 2: per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp + loglandarea + logdoc
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1    431 1258075791
## 2    432 1267158222 -1  -9082431 3.1115 0.07845 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

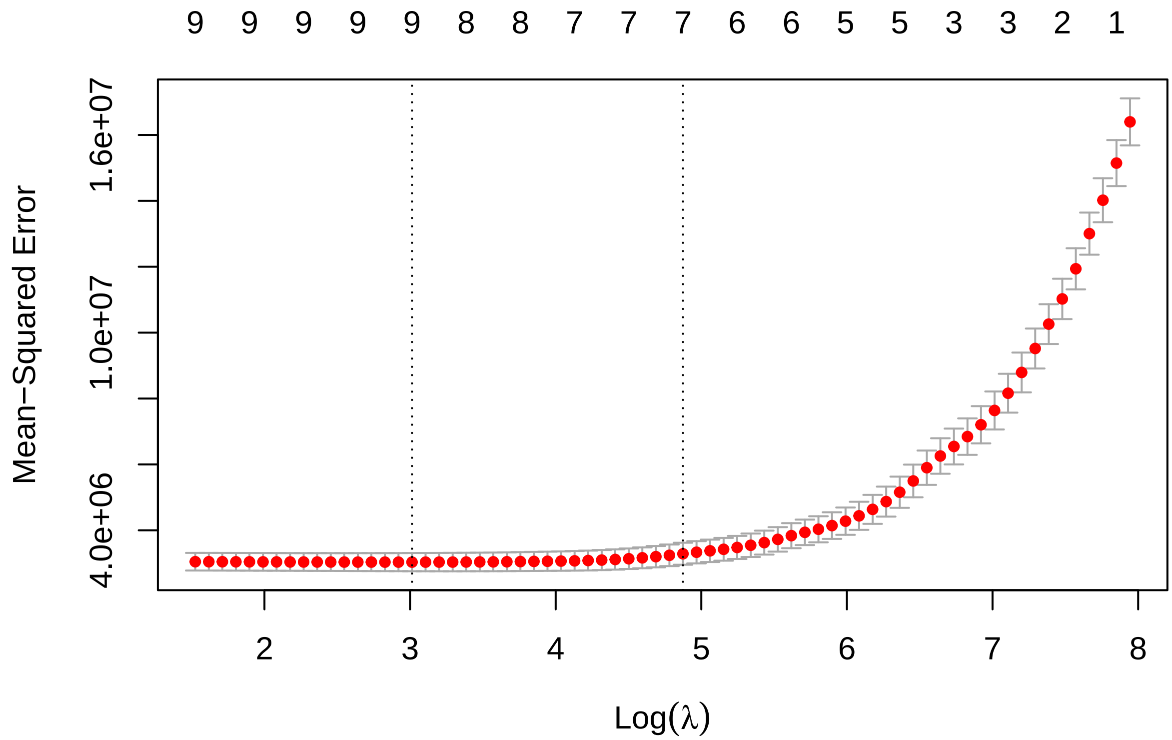
When using BIC, stepwise regression gives us the same model as model 2. Thus, if we were to add region and consider interactions, we would land on Model 5 again. When using AIC, the model is nearly the same but includes population 65 plus as a predictor. Since the F-test yields a p-value of 0.08 (so is significant at the 0.1 level) and we saw previously that population 65 plus does not add much to the prediction of per capita income through the added variable plots, we will say we prefer Model 10 to Model 9. Thus, once we consider region and interactions, we land on Model 5 again.

Now let's try Lasso:

```
library(arm)
```

```
## Loading required package: lme4

##
## arm (Version 1.12-2, built: 2021-10-15)

## Working directory is /Users/meganchristy/Downloads

##
## Attaching package: 'arm'

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

```
x = apply(as.matrix(cdi.dat.modelling[,c(-7,-8)]),2, function(x) rescale(x,"full"))
cv.glm.lasso = cv.glmnet(x, cdi.dat.modelling[,7])
plot(cv.glm.lasso)
```




```r
cbind(coef(cv.glm.lasso,s=cv.glm.lasso$lambda.min), coef(cv.glm.lasso,s=cv.glm.lasso$lambda.1se))
```

```
## 11 x 2 sparse Matrix of class "dgCMatrix"
##                             s1          s1
## (Intercept)        18561.48182 18561.4818
## pop.18_34          -2592.00502 -2170.1580
## pop.65_plus         -236.68817      .
## pct.hs.grad        -1439.15694  -143.4467
## pct.bach.deg        5450.83360  4403.3465
## pct.below.pov      -3856.84954 -3054.7903
## pct.unemp           1120.77664   775.3103
## loglandarea        -1193.84996 -1097.0419
## logdoc              2339.43544  2244.2729
## loghospbeds             .           .
## log.per.cap.crime    -21.98671      .
```

The model obtained by minimizing cross-validation error includes almost all the predictors, and the model that is 1 SE above that is the same as our Model 2. Thus, if we were to add region and consider interactions, we would land on Model 5 again.

Thus, we are left to compare Model 5 and Model 8. Since stepwise and lasso essentially found the same result to model 5, and AIC and BIC prefer Model 5 over Model 8, we will select Model 5. Our final model is:

```r
modfinal= lm(per.cap.income ~ . * region - loghospbeds - pop.65_plus -
    log.per.cap.crime - loghospbeds * region - pop.65_plus *
    region - log.per.cap.crime * region - pop.18_34:region -
    region:loglandarea - region:logdoc - pct.hs.grad:region,
    data = cdi.dat.modelling)
summary(modfinal)
```

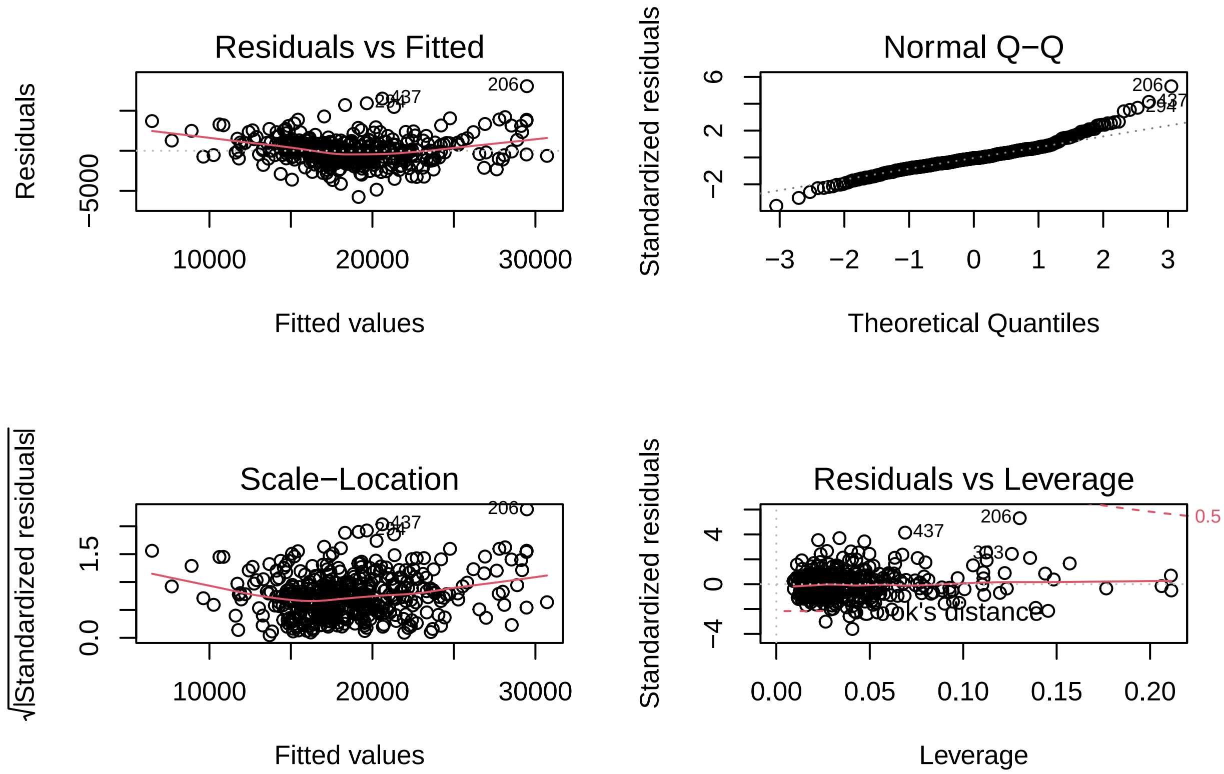
```
##
## Call:
```

```
## lm(formula = per.cap.income ~ . * region - loghospbeds - pop.65_plus -
##     log.per.cap.crime - loghospbeds * region - pop.65_plus *
##     region - log.per.cap.crime * region - pop.18_34:region -
##     region:loglandarea - region:logdoc - pct.hs.grad:region,
##     data = cdi.dat.modelling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5756.5  -936.6   -85.1   808.0  8065.5
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            30022.004   2049.377  14.649  < 2e-16 ***
## pop.18_34               -289.194     22.690 -12.745  < 2e-16 ***
## pct.hs.grad             -131.152     23.685  -5.537 5.40e-08 ***
## pct.bach.deg             343.777     23.403  14.689  < 2e-16 ***
## pct.below.pov           -489.843     55.665  -8.800  < 2e-16 ***
## pct.unemp                400.897     76.394   5.248 2.44e-07 ***
## loglandarea             -684.294    113.148  -6.048 3.23e-09 ***
## logdoc                   972.718     83.396  11.664  < 2e-16 ***
## pct.bach.deg:regionNE     93.222     20.719   4.499 8.82e-06 ***
## pct.bach.deg:regionS       8.234     17.029   0.484 0.628988
## pct.bach.deg:regionW      33.849     19.161   1.767 0.078030 .
## pct.below.pov:regionNE    -8.576     78.924  -0.109 0.913519
## pct.below.pov:regionS    144.065     59.163   2.435 0.015301 *
## pct.below.pov:regionW     29.330     80.877   0.363 0.717046
## pct.unemp:regionNE      -301.085    106.401  -2.830 0.004881 **
## pct.unemp:regionS       -326.500     90.140  -3.622 0.000328 ***
## pct.unemp:regionW       -157.249    102.606  -1.533 0.126132
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1631 on 423 degrees of freedom
## Multiple R-squared:  0.8445, Adjusted R-squared:  0.8386
## F-statistic: 143.6 on 16 and 423 DF,  p-value: < 2.2e-16
```

```r
vif(modfinal)
```

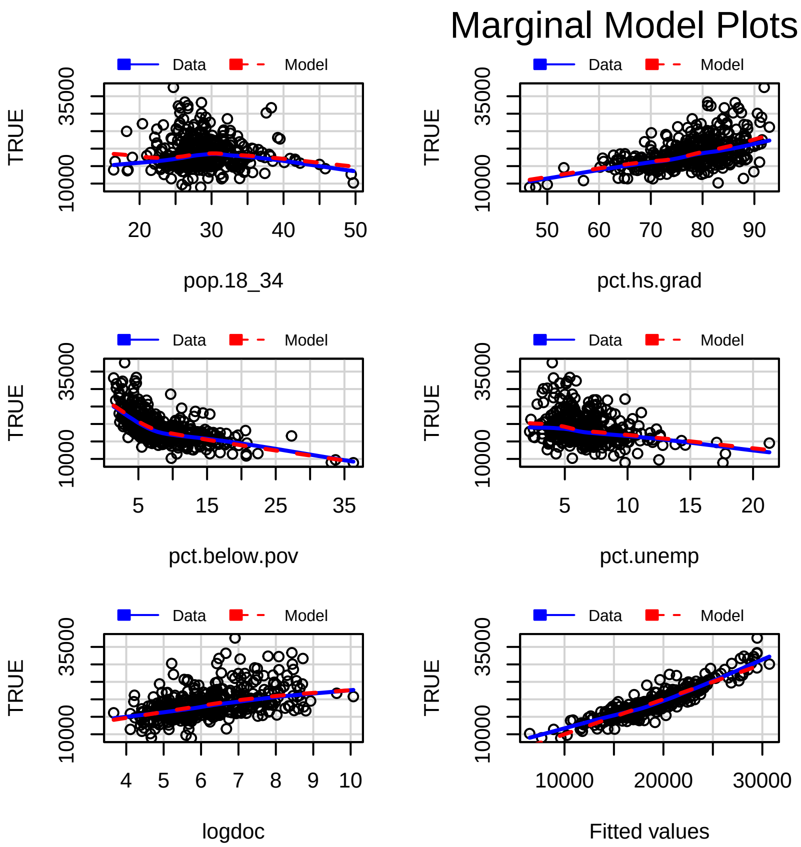
```
##                             GVIF Df GVIF^(1/(2*Df))
## pop.18_34               1.493142  1        1.221942
## pct.hs.grad             4.557977  1        2.134942
## pct.bach.deg            5.298435  1        2.301833
## pct.below.pov          11.093929  1        3.330755
## pct.unemp               5.266715  1        2.294932
## loglandarea             1.606195  1        1.267358
## logdoc                  1.502968  1        1.225956
## pct.bach.deg:region    71.343085  3        2.036535
## pct.below.pov:region 1297.624841  3        3.302617
## pct.unemp:region     1610.072857  3        3.423531
```

```r
par(mfrow = c(2,2))
plot(modfinal)
```




```
mmps(modfinal)
```

```
## Warning in mmps(modfinal): Interactions and/or factors skipped
```



This is the model we obtained by fitting the model with all predictors except region, dropping variables with multicollinearity issues or inadequate added variable plots, adding region and interactions in, then dropping insignificant interactions. It was confirmed through the all subsets, stepwise regression, and lasso procedures. The R-squared of the model is 0.8445, indicating that the model explains 84.45% of variability in per capita income. None of the VIF values are dangerously large, and the residual diagnostic plots are good enough.

Should we be worried about missing states or counties?

To address the fourth research question (should we be worried about missing states or missing counties), we determined the 48 unique values of state in the data in order to figure out which states are missing.

```
sort(unique(cdi.dat$state))
```

```
##  [1] "AL" "AR" "AZ" "CA" "CO" "CT" "DC" "DE" "FL" "GA" "HI" "ID" "IL" "IN" "KS"
## [16] "KY" "LA" "MA" "MD" "ME" "MI" "MN" "MO" "MS" "MT" "NC" "ND" "NE" "NH" "NJ"
## [31] "NM" "NV" "NY" "OH" "OK" "OR" "PA" "RI" "SC" "SD" "TN" "TX" "UT" "VA" "VT"
## [46] "WA" "WI" "WV"
```

The states that are missing are Alaska, Iowa, and Wyoming. These are states that are perhaps less populous (or at least contain counties that are not very populous), so we should be worried that the states are missing because the results of our analysis may not generalize to these states and their counties.