Modeling per-capita income in U.S. counties

Zach Ohl – zohl@andrew.cmu.edu

Abstract

The goal of this study is to examine how the average income per person in U.S. counties is related to other geographic, economic, and social variables. The data consist of observations for the 440 most populous counties in the country and 14 variables for each county describing information from the years 1990 and 1992. We approached this topic by searching for a linear regression model that predicts average income based on some combination of the 14 variables. The model that we found included expected predictors such as bachelor's degree percentage, as well as some less expected terms, including interactions. The final model reveals some useful predictors for a county's average income per person, but caution must be taken before using this model to make predictions about the approximately 2600 other U.S. counties.

Introduction

Our goal is to discover how certain features of a county's economic health, physical healthcare and social well-being are related to the county's per capita income. The variables used in the study contain information on the county's geography, demographics, metrics of physical health, crime, education, and economic information. Using these variables, we have been given the following tasks:

1. List and describe any apparent relationships between variables. Explain the relationships in terms of the meanings of the variables, if possible.
2. Describe the predictive ability of a county's crime numbers and region of the country on its per-capita income.
3. Model per-capita income using a combination of the other variables. Choose a statistically valid model that reflects the meaning of the variables and can be interpreted by the social scientists who requested the study.
4. Describe the consequences of the missing states and counties from the data set.

Each section of the paper after Data is broken down according to these four questions.

Data

The data for this study comes from the textbook *Applied Linear Statistical Models*, by Kutner and others. their original source was the Geospatial and Statistical Data Center at the University of Virginia. The 1990-1992 dataset contains 440 observations, each representing a unique U.S. county. An initial look at the counties might suggest that county is a categorical variable, since multiple observations have the same value of county. But these actually represent duplicates of the same county name in different states, so all 440 county observations are unique (See Technical Appendix, page 1). The values of 17 variables are included in the dataset but we will use 14 of them—13 quantitative and 1 categorical variable. Variable definitions are listed in the table on the next page.

Variable definitions

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

Variable summaries are shown below.

Summary tables of the numeric variables

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

Summary of regions

	NC	NE	S	W
Freq	108	103	152	77

Methods

1.

We will look at summaries of each variable, check for missing values, and examine the distributions of each variable alone and against other variables to look for patterns. The shapes of single-variable distributions will be used to suggest transformations of variables. Correlations plots of paired variables will be used to look for high correlation between predictors and linear relationships among variables, especially between predictors and the response, per-capita income.

After identifying the relationships, we will examine the questions about whether the relationships are expected and can be explained using the meanings of the variables. The shapes of the distributions will also be used to suggest transformations for the models in research questions 2 and 3.

2.

To address this question, we will first find 3 models that predict per-capita income using region and crime (the raw crimes variable). One model will use just crime, one will use crime and region, and one will use those two predictors plus their interaction. The crime variable will be log-transformed to deal with its right skew.

We'll then repeat fitting the three models above, but with log(crimes) replaced by log(crimes/population). We will compare the 6 total models using ANOVA tests and information criteria, while making sure modeling assumptions are met according to the residual diagnostics.

3.

Before fitting models, we will transform most, but not all, of the predictors using log, because of the skews shown in their distributions. We may consider power transformations, especially on the one left-skewed variable, later in selection process once the variables have been narrowed down.

We will remove the variables for total income and population. They can be used to calculate to the response variable, per-capita income (total income/population), so there is no need to assess their relationship with the response variable. We'll also add a new variable, population density (population/land area). Based on prior knowledge of incomes in U.S. cities, we could imagine this new variable being significant. The population density variable will also be log-transformed due to right skew.

We'll then look for a model using all remaining variables except for state. We may look at state as a predictor later, but for now, a categorical variable that has 48 categories and is unlikely to be significant would only make the model selection messier. At first, we'll only use variables on their own with no interactions. If any of the region indicator variables are selected, we'll follow the convention of keeping all four region indicators.

Then we will proceed to variable selection using the 'all subsets' method, starting with the 12 numerical variables and one categorical (region) variable. We will check the diagnostics (variance inflation factors, four residual diagnostic plots, and marginal model plots) and compare the subsets of variables suggested when using Akaike information criterion (AIC) and when using Bayesian information criterion (BIC). Analysis of variance (ANOVA) and covariance (ANVOCA) may be used to compare sets and subsets of

predictors, when applicable. When selecting variables, we will keep in mind that the purpose of our model is understanding the predictors, in addition to making predictions.

We'll also look for a model with interactions included. A stepwise method may be used in place of all subsets. Again, we'll use information criteria including AIC and BIC, as well as ANOVA/ANCOVA tests to compare models. Similarly to before, if the interaction of a continuous variable and a region indicator variable is selected, we'll keep that continuous variable's interaction with all four region indicators. Knowing that interactions can make the model messy and difficult to interpret, we'll consider ways to reduce the model if the variable selection methods results in a complicated model.

If the winning models from both the interaction and non-interaction processes both meet modelling assumptions equally well, we'll choose between them by comparing adjusted r^2, AIC, and BIC.

Eventually we'll try a penalized regression method, LASSO regression–with interactions and without–and compare the results with the results from above.

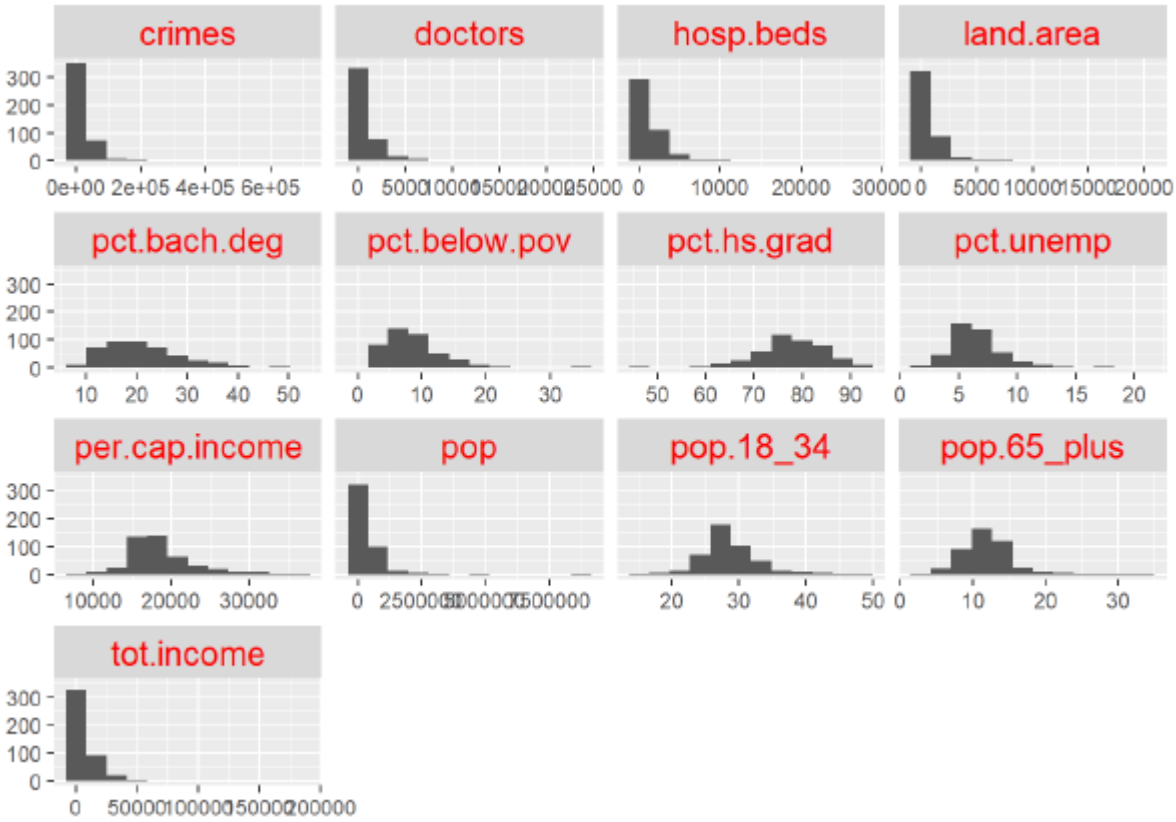
4.

To answer this question, we'll consider how the 440 counties in the dataset ended up there, what they have in common, and what the missing counties have in common. We'll use the limited information given, and if we have time, possibly look up more information on the included counties to check for any patterns.

Results

1.

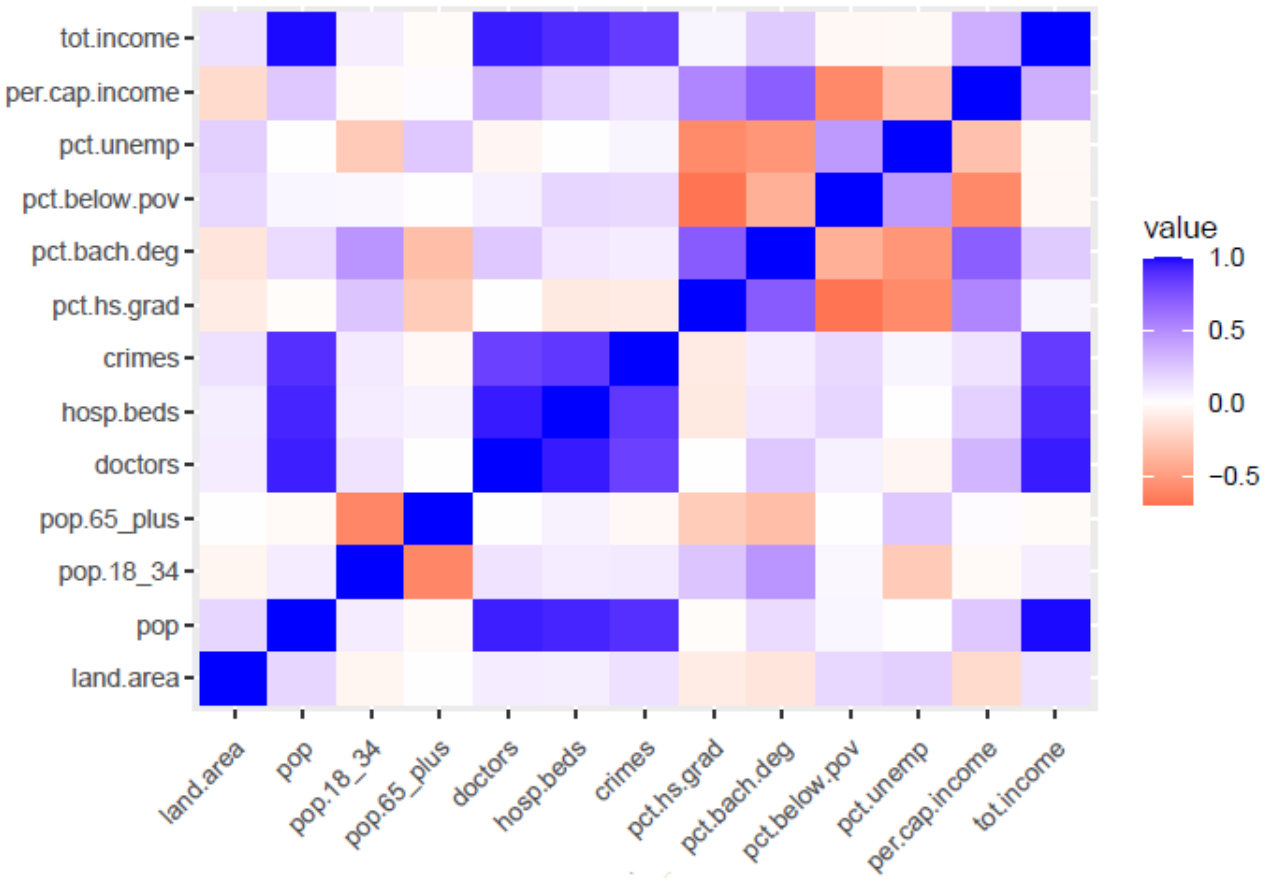
Distributions of the numeric variables are shown below.



The variables *crimes, doctors, hosp.beds, land.area, pop,* and *total.inc* are all heavily right skewed. Predictors *pct.bach.deg, pct.below.pov, pct.unemp, pop.65plus*, and the response, *per.cap.income* all appear only slightly right-skewed, while pct.hs.grad has a slight left skew. The last numeric predictor, *pop.18-34*, appears relatively symmetric.

The heatmap below visualizes the correlations between the variables:

Heatmap of variable correlations



Based on the correlation heat map above, there are apparent relations for pairs of variables you would expect, for instance, doctors vs hospital beds and population vs total income. Some interesting variables that have a relationship with *per.cap.income* include *pop*, *doctors, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp*, and *tot.income*. The relationship *per.cap.income* vs *pc.bach.deg* relationship looks very linear, while the pattern appears more curved for the plots of *per.cap.income* vs *pct.hs.grad, pct.below.pov*, and *pct.unemp*. The relationships between the response and the *pop, doctors*, and to*t.income* predictors just look vaguely positive—no real pattern is obvious.

A set of scatter plots (Tech. appendix, page 3) also illustrates some of these relationships.

2. All six models were similar with respect to residual diagnostics (see Tech. Appx., pages 5 and 8), so we used analysis of covariance to choose the best model out of each set of three. In each set of three models, the first only used some measure of crime, the second used crime and region, and the third used those two predictors plus their interaction.

For each set of three models, whether total crimes or crimes per capita was used as the crime measure, the best model was the second model (with crime, region, but interaction). Both of these models had coefficients with the same signs, so in that sense, it doesn't matter whether total crimes or crimes per

capita is used in the model. But we are still interested in which predictor results in a better model. To compare these two models and determine whether crimes or crimes per capita makes a better predictor, we looked at the adjusted r^2, AIC, and BIC for each model. All three measures pointed to the model using total crimes, not per-capita crimes, as shown in the Tech. Appx, page 9).

So, we chose the following model to answer the question:

log(per.cap.income) = 9.188 + 0.0667*log(crimes) + 0.104*regionNE – 0.0870*regionS – 0.0553*regionW

The coefficients can be interpreted as follows.

Crime: For every 1% increase in the number of crimes in a U.S. county, we expect about a 0.07% increase in per-capita income.

Region: The baseline incomes per capita by region are shown below.

North-central region: $9779.07
Northeast region: $10,850.86
Southern region: $8964.25
Western region: $9252.97

The model suggests that per-capita income is indeed related to crime, and that it is also related to region, but that the relationship between per-capita income and crime is not dependent on region.

3.

The variables that were log-transformed due to skewness were: the response variable, *per.cap.income*, and the predictors *land.area, pop.65_plus, doctors, hosp.beds, crimes, pct.bach.deg, pct.unemp*, and pop.dens. No power transformations were used, although we may still consider one for the left-skewed variable, *pct.hs.grad*, depending on the look of the marginal model plots or other diagnostics later on.

With these variables transformed, and the variables *pop*, *tot.income*, and *state* removed, we began variable selection using all subsets with no interactions included. We looked at the models that minimized both BIC and AIC (see Tech. Appx, pages 11 and 12). The only difference was an additional region indicator variable being selected by AIC, but we would include all region indicators if the variable region was to be included anyway, so there was no practical difference in the two models.

The variables included in the model are: *land.area*, *pop.18_34*, *doctors*, *pct.hs.grad* , *pct.bach.deg*, *pct.below.pov*, *pct.unemp*, and *region*. We examined variance inflation factors (VIFs), residual diagnostic plots, and marginal model plots, and all evidence showed a valid model (see Tech. Appx, page 13).

Then we looked for a model with variable interaction terms included. The all subsets method was not practical due to time and computing limits, so we used stepwise selection in both directions to choose variables. We initially tried selection algorithms using both AIC and BIC, but even using BIC, the resulting model was pretty complex, with lots of difficult interaction terms, and of course the AIC-selected model was even more complex. So decided to work with the BIC model and only check the AIC one if necessary.

In general, when choosing between models suggested by AIC and BIC measures, we'll lean toward the BIC-suggested model because our goal is understanding the relationship between the variables. However, we

also want the model that is most clearly indicated by the data, so if strange variable combinations significantly improve the model, we will attempt to include them and interpret them.

The stepwise-BIC model had 40 predictor terms, including interactions and all the different indicator variables for different regions. We wanted to eliminate some terms to make it more interpretable. Before that, we checked to make sure including interactions was necessary at all using an F-test. The test stat was very significant (see Tech. Appx, page 20), so we proceeded with trying to simplify the interaction model.

We started by removing interactions that were insignificant judging by their p-values or had coefficients less than 0.01. Four interactions were removed by these criteria—none of them seemed especially meaningful based on the definitions of the variables. We continued in this manner with decreasingly strict criteria (see Tech. Appx, pages 21-27). The BIC got a little worse from the first four term removals, but got better from certain other removals. Eventually we arrived at a model with 22 terms (called model_bic_1 in the Tech. Appx), including 9 continuous variable interactions, as well as the interaction between region and pct.bach.deg.

The diagnostics for this model (including residual plots and marginal model plots) all looked good for this model (see Tech. Appx, page XX). They only problem with it was high VIFs, which is to be expected from a model with many interaction terms.

Next, we used LASSO regression to find a model. Since we've already concluded at this point that interaction terms improve the model, we won't discuss LASSO regression with no interactions. For the regression with interactions, our strategy was to include the interactions from the final model found earlier, *model_bic_1*, and see if LASSO kept the terms in. The LASSO selection process kept almost all of the predictors we used before, but there was one interaction, *hosp.beds:pop.dens*, that it was not as insistent on keeping (see Tech. Appx, page 34).

In our previous final model, the coefficient for this predictor was less than 0.01 and had a p-value greater than 0.05. So, based on that information and on the LASSO results, we tried making this one final term removal from the previous model. The removal had no notable effect on BIC (see Tech. Appx, page 36), so we decided to use this model with 21 terms as our final model for per-capita income.

The best model predicting per-capita income from the other variables is described on the next page (we don't present it as a formula due to its length). Note that these predictors sum up to predict the log of per-capita income.

Model coefficients

Coefficient	Estimate
(Intercept)	14.225
pop.18_34	–0.0376
log(pop.65_plus)	–1.292
log(doctors)	–0.128
log(hosp.beds)	–0.1208
log(crimes)	–0.0778
pct.hs.grad	–0.0350
log(pct.bach.deg)	0.453
pct.below.pov	0.0416
log(pct.unemp)	0.0818
log(pop.dens)	0.0507
log(crimes)*log(land.area)	0.0175
log(pct.bach.deg)*log(land.area)	–0.0515
pct.below.pov*log(land.area)	–0.00609
pop.18_34*log(pop.dens)	0.00218
log(pop.65_plus)*log(hosp.beds)	0.0531
log(pop.65_plus)*pct.hs.grad	0.0118
log(doctors)*log(pct.bach.deg)	0.0579
log(pct.bach.deg)*pct.below.pov	–0.00954
log(pct.bach.deg)*regionNE	–0.00291
log(pct.bach.deg)*regionS	–0.0150
log(pct.bach.deg)*regionW	–0.00840

The coefficients can be interpreted as follows.

Percent of population aged 18-34: For every 1 unit increase in the of 18-34 population percent in a U.S. county, we expect about a 0.04% decrease in per-capita income.

Percent of population 65 or older: For every 1% increase in the 65+ population percent in a U.S. county, we expect about a 1.3% decrease in per-capita income.

Number of active physicians: For every 1% increase in the number of doctors in a U.S. county, we expect about a 0.13% decrease in per-capita income.

Number of hospital beds: For every 1% increase in the number of hospital beds in a U.S. county, we expect about a 0.12% decrease in per-capita income.

Total serious crimes: For every 1% increase in the number of crimes in a U.S. county, we expect about a 0.08% decrease in per-capita income.

Percent high school graduates: For every 1 unit increase in the of percent of a U.S. county that graduated high school, we expect about a 0.035% decrease in per-capita income.

Percent bachelor's degrees: For every 1% increase in the of percent of a U.S. county with a bachelor's degree, we expect about a 0.45% increase in per-capita income.

Percent below poverty: For every 1 unit increase in the of percent in a U.S. county, we expect about a 0.04% decrease in per-capita income.

Percent unemployment: For every 1% increase in the of percent of a U.S. county that is unemployed, we expect about a 0.08% increase in per-capita income.

Population density: For every 1% increase in the of population density of a U.S. county, we expect about a 0.05% increase in per-capita income.

The interaction terms are trickier to interpret, and we think it's simpler to interpret the signs of the coefficients rather than discuss specific unit or percent changes.

Interactions with positive coefficients: *crimes*land.area, pop.18_34:pop.dens, pop.65_plus*hosp.beds, pop.65_plus*pct.hs.grad, doctors*pct.bach.deg*

The positive coefficients of these continuous interaction terms indicate that as one variable increases, the slope of the other variable increases.

Interactions with negative coefficients: *pct.bach.deg*land.area, pct.below.pov*land.area, pct.bach.deg*pct.below.pov, pct.bach.deg*regionNE, pct.bach.deg*regionS, pct.bach.deg*regionW*

The negative coefficients of these continuous interaction terms indicate that as one variable increases, the slope of the other variable decreases. The negative coefficients for the continuous-categorical interactions are a little easier to explain. From the positive coefficient on the lone *pct.bach.deg* term, we know that that per-capita income increases as the percent of county residents with a bachelor's degree increases. The negative coefficients on the interactions of pct.bach.deg and *regionNE, regionS*, and *regionW* imply that the positive effect lessens in those three regions, compared to in *regionNC*.

4.

Based on the description of the dataset, we determined that the sample of counties in the dataset were not randomly selected from the population of U.S. counties. We have two important pieces of information about how the 440 counties in the dataset were chosen:

-They are the 440 most populous counties in the U.S. (with exceptions-see below).

-Any observations with missing values were deleted from the set.

Each of these facts are reasons that the sample of counties in the study is not random. There is plenty of reason to doubt that the most populous counties in the U.S. are representative of the rest of the approximately 2600 counties. The minimum county population in the dataset was over 10,000, but there are plenty of counties with only a few thousand people and even a few hundred people. One county in Hawaii has 86 people!

Deleting any observations with missing values of a variable is another non-random method of choosing counties. It is possible that the observations with missing data tend to have something in common and that the data is missing for a reason. This might overlap with the last problem since smaller counties are less likely to keep complete records.

Discussion

1.

Most of the relationships we noticed are expected. For instance, variables that represent population totals (total income, crimes, population, crimes, hospital beds, and doctors) and not percentages or per-capita measures are all correlated. None of the percentage or per-capita variables have super high correlations. One variable pair with a slight correlation is percentage of bachelor's degrees and per-capita income. We would expect this relationship to show up in the model. Both education measures, percent bachelor's degrees and percent high school graduates, are also inversely correlated with percent below poverty and percent unemployed. This should be expected since education is supposed to prevent both of those things. The inverse correlation between the young and old age groups is also expected. We can also see that per-capita income is inversely correlated with percent below poverty and percent unemployed. These relationships also make sense and we would expect to see them in the models.

2.

According to the model, per-capital income and crime have a positive relationship. Since we know that the people earning the higher incomes tend to commit less crimes, the relationship is most likely due to other variables, such as population-density. Bigger cities are likely to have more crimes as well as higher average incomes.

The previous two answers do not change when per-capita crime used instead of total crimes. The coefficient on crime is still positive (about 0.459) and analysis of covariance still suggests that the

interaction terms are unneeded ($F \approx 0.120$). However, the per-capita crime coefficient is less significant in that model and the crimes coefficient is significant in the model I chose, which the main reason I chose it.

Because higher per-capita income is not actually caused by crime, we are only interested in which variable predicts that income better: either total crimes or per-capita crimes. Because of lurking variables or other reasons, total crimes predicts per-capita income better than per-capita crimes. If we were using a predictor that was more likely to have a direct affect on income, like college education, it might make more sense to use the 'per-capita' version of both predictor and response for the sake of interpretability and consistency. But to answer the question asked in this part of the study, we chose the model with total crimes because of its superior predictive ability.

3.

The meanings of the predictor's coefficients were discussed above. Now we turn the discussion to whether or not these meanings make any sense. The answer is mixed. Some of the coefficients make intuitive sense, some can be justified with a little speculation, and some don't make much sense.

For example, percent bachelor's degrees had the largest positive coefficient. It was positive, which makes sense, since you would expect higher rates of college education to coincide with higher incomes. Similarly, percent of population 65 or older had the lowest negative coefficient. This makes sense since retired county residents who earn little to no income would be expected to decrease the per-capita income of a county.

Some coefficients that make less obvious sense but can be justified are numbers of active physicians and hospital beds. Both coefficients are negative. Our first thought is that these would be associated with wealth, but they could also be associated with higher percentages of elderly residents, and as we already mentioned, older residents lead to lower average incomes.

Some coefficients that are hard to make sense of are percent below poverty, Percent unemployment, and population density. The first two would be expected to have a negative effect on per-capita income, but their coefficients are positive. The population density variable was added specifically because we expected it to correlate to higher incomes, but its coefficient ended up negative.

The interaction terms are even more difficult to put in context. But some can be justified if you think about them. For example, knowing that percent bachelor's degrees has a positive association with income, you could see why increasing land area might reduce that slope—many of the higher paying positions in rural areas like independent contractor, small business owner, or even farmer, don't require a degree, while higher paying jobs in cities usually do. However, many of the other interactions are even more of a stretch.

This is probably the biggest weakness of the model: it is not parsimonious. Some of the predictors are confusing to interpret or even predict the opposite of what you would expect. Another weakness is the terms removed after stepwise selection. You could argue that if we're going to include predictors that are hard to interpret, we might as well just include all of the 40-50 predictors and make the model as accurate a predictor as possible. Another weakness is that we left out the state variable. The final model was

briefly checked with state included, and a few state indicator variables were significant (see Tech. Appx. page 37), but we didn't think it was worth putting every state in the model.

There are other weaknesses of the model that have more to do with the data the study was based on than the model itself. These weaknesses are discussed in the next section.

Some strengths of the model are its high adjusted r^2 and low BIC, and the fact that we were able to get the number of predictors down to 21 instead of 40-50. The model has good residual diagnostic plots, and marginal model plots that suggest the variable were modeled correctly. The model should also have high predictive ability.

4.

The counties in the dataset were not randomly selected from the population of U.S. counties. Furthermore, they represent a very specific subset of all U.S. counties—the most populous ones. We cannot assume that the models based on the largest counties would generalize to small or medium sized counties, or any actual random sample with a wide range of county populations. The deleted observations add to the non-random nature of the subset of counties.

Because of these reasons—yes, we should be worried about the missing counties. We suggest that any future studies on this topic use a randomly selected sample of U.S. counties instead of the most populous ones, or just use every U.S. county, since there are only about 3000 of them. We also recommend that counties with missing data are kept in the dataset so that whatever information they do contain can be used to create a more representative model.

References

Kutner, M.H., Nachsheim, C.J., Neter, J. & Li, W. (2005) Applied Linear Statistical Models, Fifth Edition. NY: McGraw-Hill/Irwin. Original source: Geospatial and Statistical Data Center, University of Virginia.

Modeling per-capita income in U.S. counties - Technical Appendix

Zach Ohl

Load packages:

```
library(tidyverse)  #
library(car)        #
library(dplyr)
library(ggplot2)    #
library(knitr)      #
library(kableExtra) #
library(reshape2)   #
library(reshape)    #
library(grid)       #
library(gridExtra)  # for grid.arrange
library(leaps)      # for all subsets
library(glmnet)     # for LASSO
library(arm)        #
```

Question 1

Read in data

```
income <- read.csv("C:/Users/Zachary Ohl/Desktop/CMU courses/Applied Linear Models/project 1/cd
i.dat",
    sep = "")
detach()
attach(income)
income_numeric <- dplyr::select(income, -c(1:3, 17))
```

Make sure no missing data. Visual inspections suggests no other types of NAs.

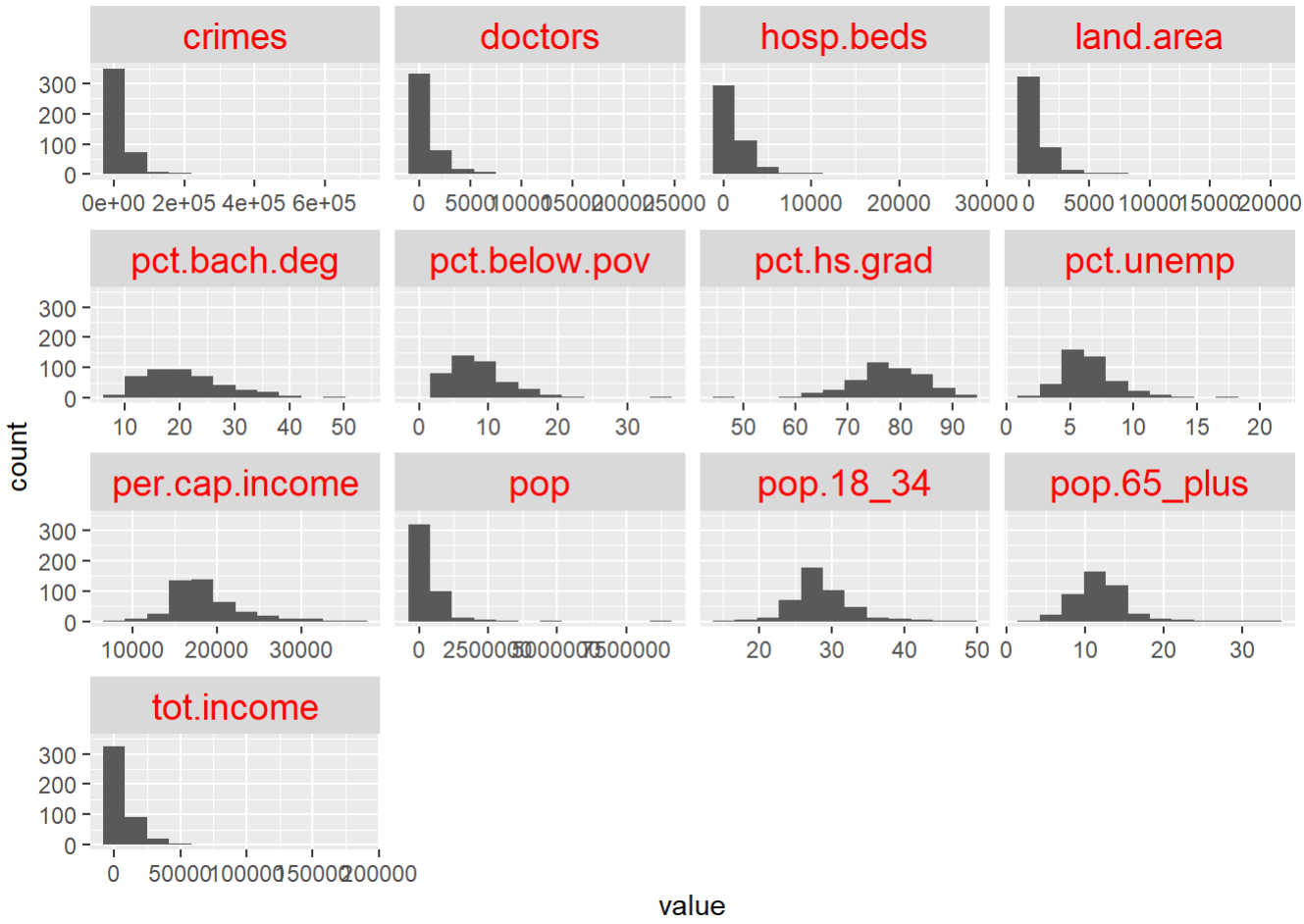
```
all(!is.na(income))
[1] TRUE
```

Check for unique values of county to make sure it's not a categorical variable.

```
length(unique(income$county))  #373 'unique' counties
[1] 373
length(unique(paste(income$county, income$state)))  #actually 440 unique countys
[1] 440
```

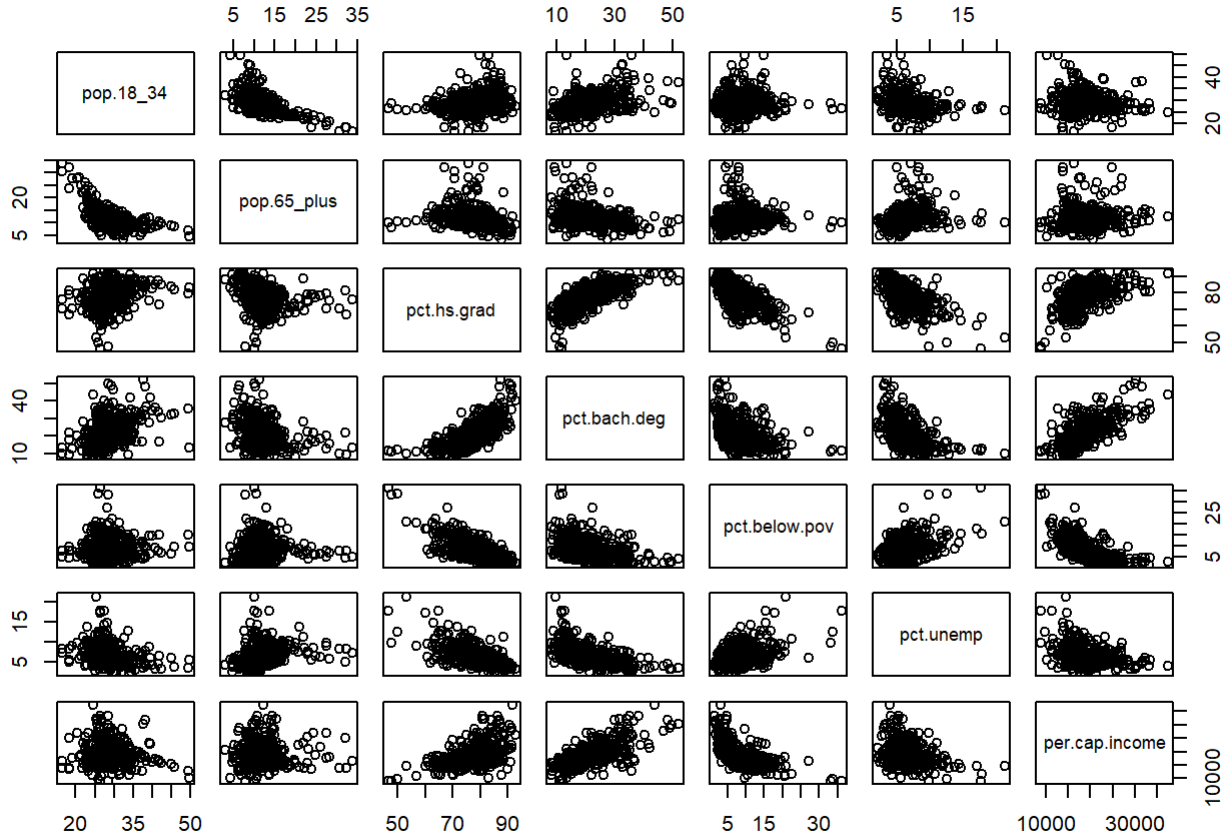
Numeric variable distributions:

```
ggplot(gather(income_numeric), aes(value)) + geom_histogram(bins = 12) +
    facet_wrap(~key, scales = "free_x") + theme(strip.text = element_text(size = 14,
    color = "red"))
```



Histograms of the numeric variables

```
pairs(income_numeric[, c(3, 4, 8:12)])  #
```



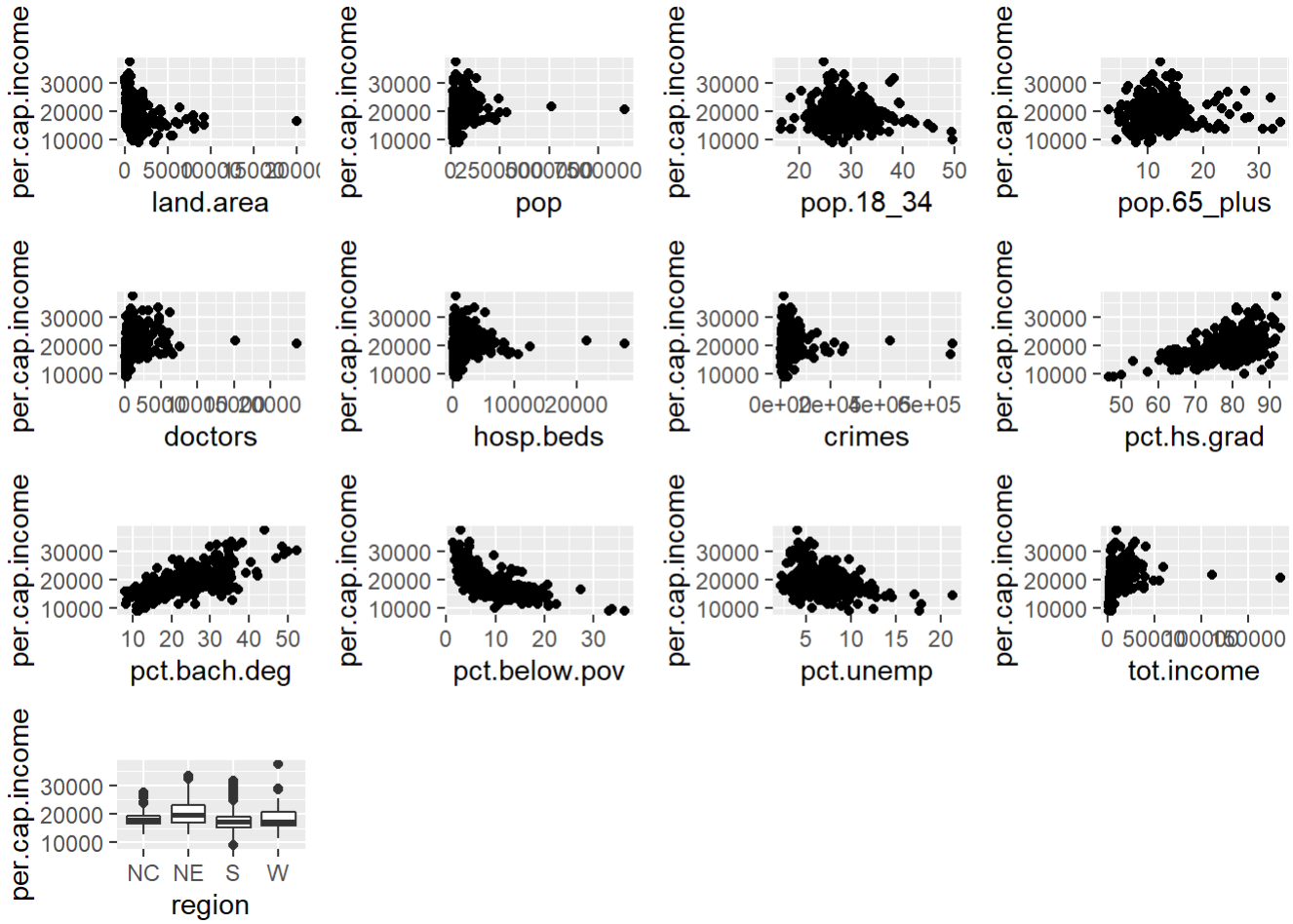
Pairs plots for some numeric variables

From looking at a matrix of every possible pairs, most of the pairs didn't show any significant pattern or correlation, as suggested by the heat plot above. Of the correlations that were apparent, most appeared linear, which could signal good potential for the model, as well as collinearity.

```
income_num_region <- data.frame(income_numeric, region = income$region)

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

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



Scatter plots of per-capita income vs. predictors

The scatter plots confirm the results the correlations heat plot. The region box plots show overlapping IQRs of each region, but there are two regions with noticeably higher per-capita incomes than the other two.

Preliminary variable transformations and addition/deletion

Based on the distributions of variables, make the following log-transformations of the response variable and many of the predictors before proceeding with the research questions.
Here we also remove variables we discussed removing (total income and population), the columns that aren't usable variables because they're unique (id, county), and the state variable for now, since we don't want 50 indicator variables messing up the model at this point. We also add the new predictor for population density, although it won't be used until question 3.

```
income_trans <- mutate(income, land.area = log(land.area), pop.65_plus = log(pop.65_plus),
    doctors = log(doctors), hosp.beds = log(hosp.beds), crimes = log(crimes),
    per.cap.income = log(per.cap.income), pct.bach.deg = log(pct.bach.deg),
    pct.unemp = log(pct.unemp), pop.dens <- log(pop/land.area)) %>%
    dplyr::select(-c(1:3, 5, 16))
```

Question 2

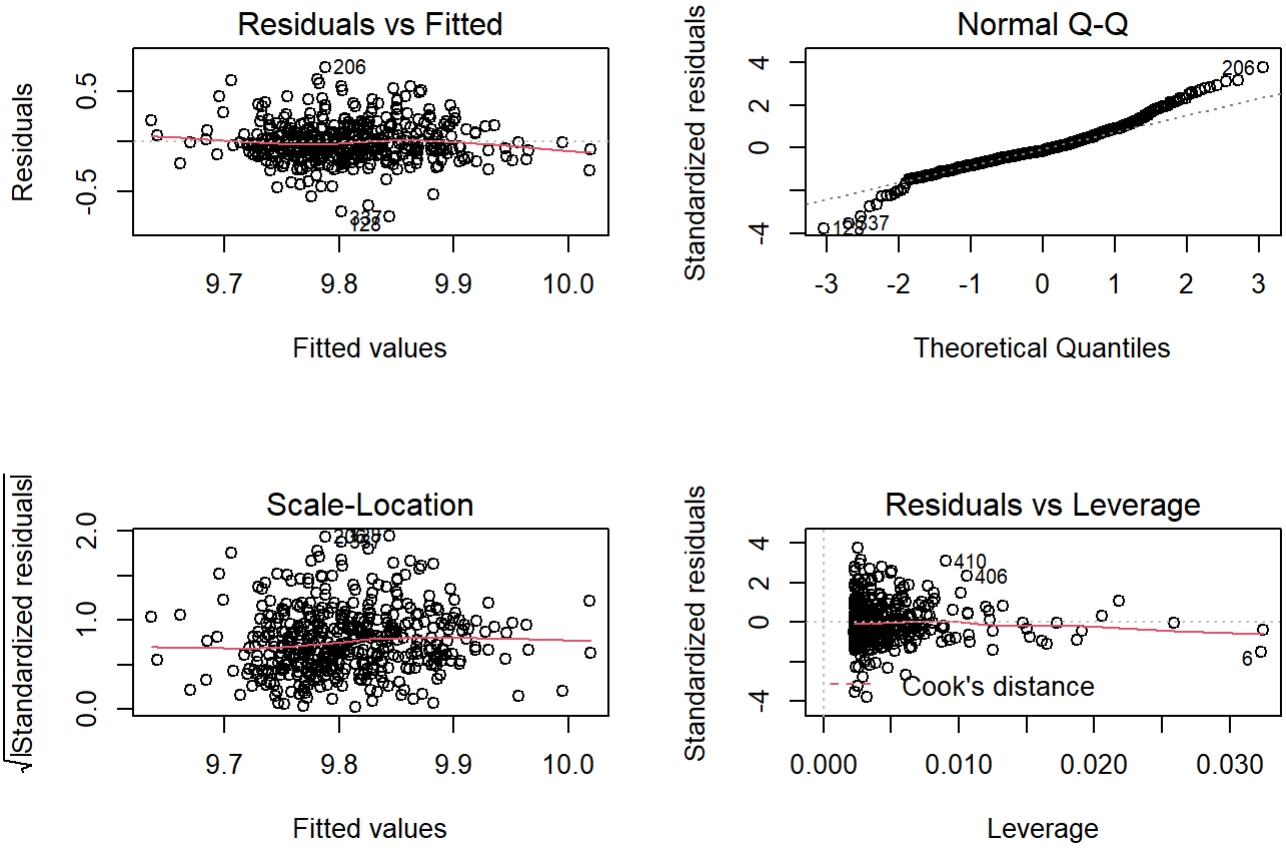
The three models using raw crime numbers are:

```
mod_crime1 <- lm(per.cap.income ~ crimes, data = income_trans)
mod_crime2 <- lm(per.cap.income ~ crimes + region, data = income_trans)
mod_crime3 <- lm(per.cap.income ~ crimes + region, data = income_trans)
```

Note that both numeric variables have been log-transformed.

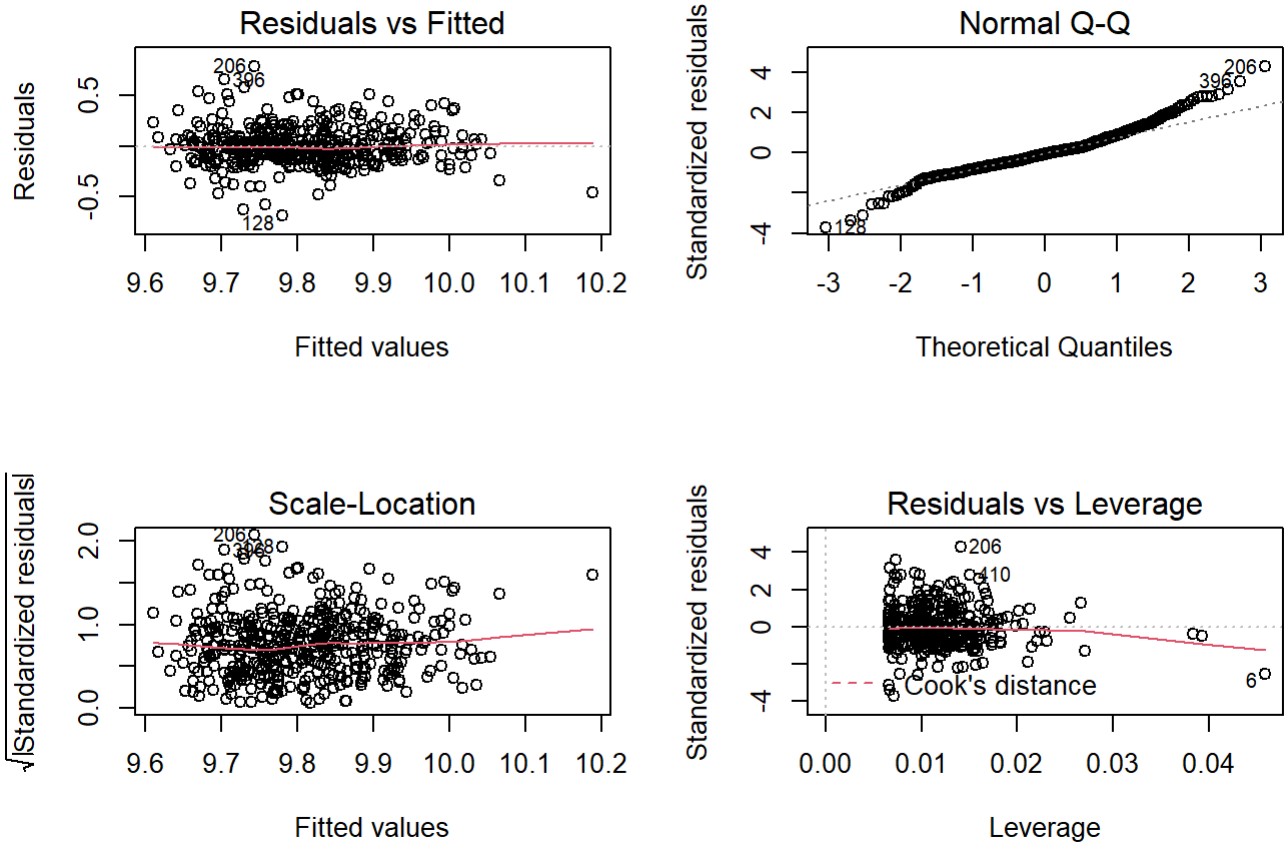
Based on the residual plots, nothing sticks out to eliminate either of the 3 models.
Residual diagnostics for mod_crime1 (raw crimes = only predictor) are below. The main issue is the residuals are both right and left skewed according to the QQ plot:

```
par(mfrow = c(2, 2))
plot(mod_crime1)
```



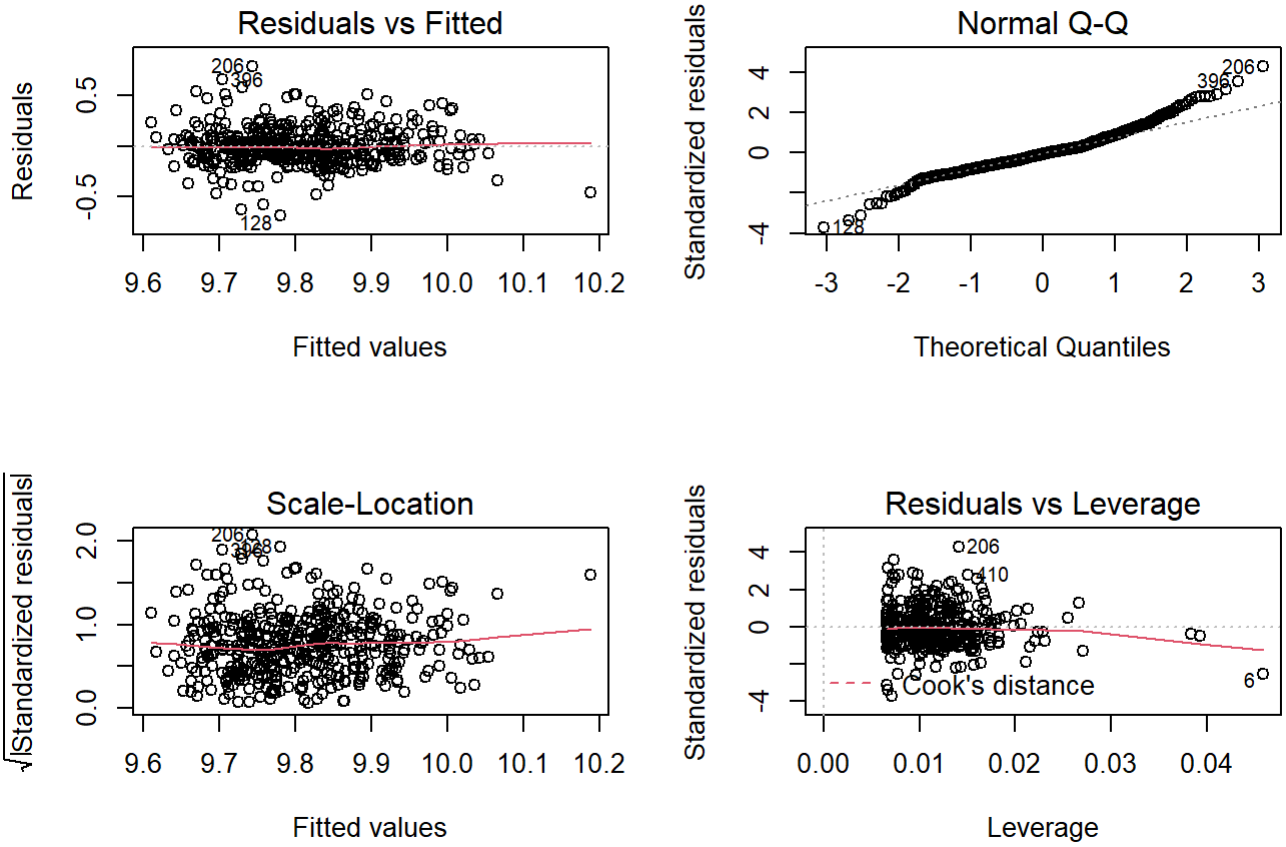
Residual diagnostics for mod_crime2 (predictors = raw crimes and region) are below. Again, the residuals are both right and left skewed according to the QQ plot:

```
par(mfrow = c(2, 2))
plot(mod_crime2)
```



Residual diagnostics for mod_crime3 (predictors = raw crimes, region, and their interactions) are below. The residuals are still skewed on both ends.

```
par(mfrow = c(2, 2))
plot(mod_crime3)
```



Because all 3 models have decent diagnostics with the same minor issue, use analysis of covariance to compare the models.

F-test for the 3 models:

```
anova(mod_crime1, mod_crime2, mod_crime3)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	438	17.27083	*NA*	*NA*	*NA*	*NA*
2	435	14.94889	3	2.321936	22.52212	1.426626e-13
3	435	14.94889	0	0.000000	*NA*	*NA*

3 rows

The F-stat for the 2nd model (region included but no interactions) is significant, while the F-stat for the 3rd model (with interactions) is not. So, based on ANCOVA, the 2nd model is preferred.

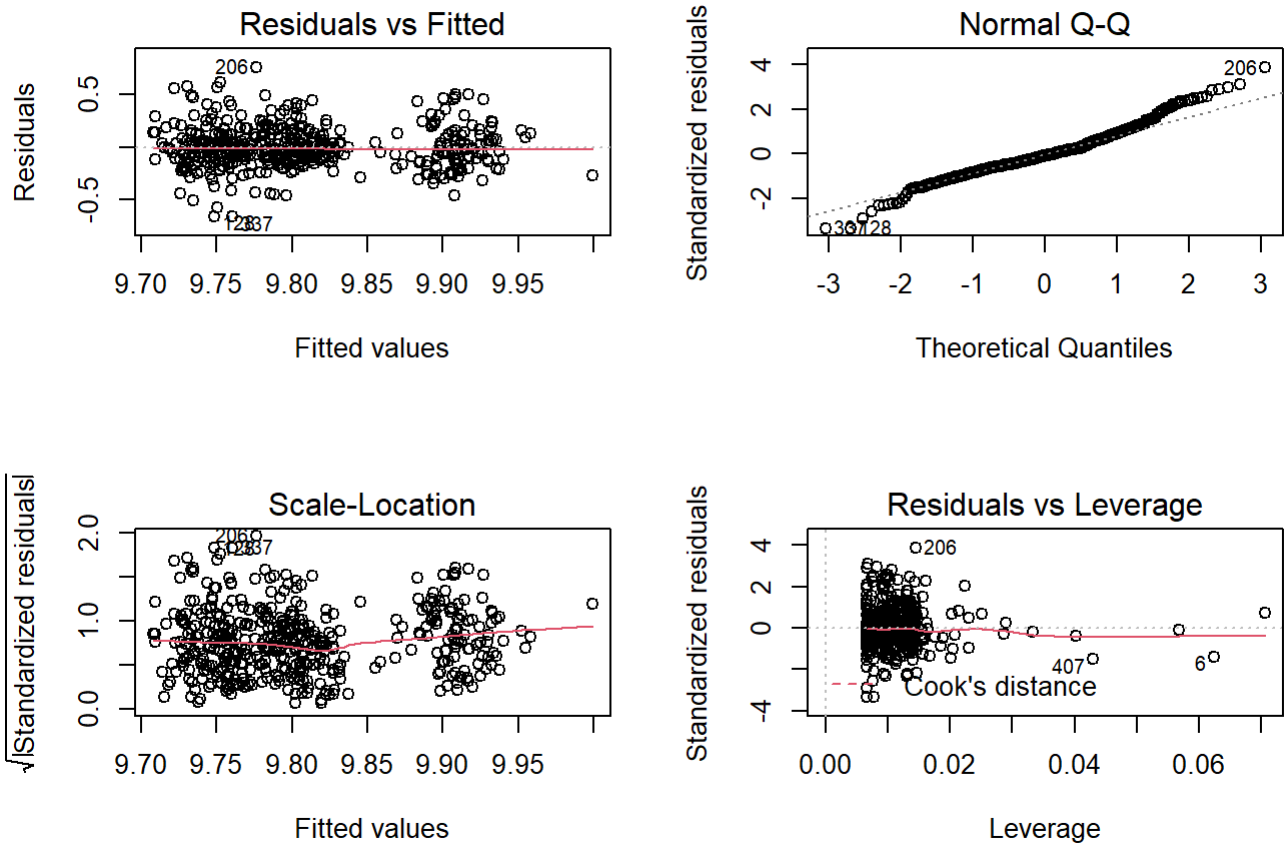
Now we repeat the process with the raw crime numbers replaced with crimes per capita.
The three models using crimes per capita are:

```
per.cap.crime <- income_trans$crimes - log(income$pop)
mod_crime_PC1 <- lm(per.cap.income ~ per.cap.crime, data = income_trans)
mod_crime_PC2 <- lm(per.cap.income ~ per.cap.crime + region, data = income_trans)
mod_crime_PC3 <- lm(per.cap.income ~ per.cap.crime + region, data = income_trans)
```

Note that both numeric variables have been log-transformed again. The new variable was created as shown because crimes were already logged and population was not.

The residual plots look virtually identical to the 3 sets of plots using raw crimes. For example, the diagnostics for mod_crime_PC2 (predictors = raw crimes and region) are below.

```
par(mfrow = c(2, 2))
plot(mod_crime_PC2)
```



Again, use analysis of covariance to compare the models.

F-test for the 3 models:

```
anova(mod_crime_PC1, mod_crime_PC2, mod_crime_PC3)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	438	18.69705	*NA*	*NA*	*NA*	*NA*

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
2	435	16.95241	3	1.744645	14.92258	2.906812e-09
3	435	16.95241	0	0.000000	*NA*	*NA*

3 rows

The F-test prefers the 2nd model (region included but no interactions) again.

Now we'll compare the 2 best models using adjusted r^2, AIC, and BIC.

```
comp2 <- cbind(c(summary(mod_crime2)$adj.r.squared, summary(mod_crime_PC2)$adj.r.squared),
    AIC(mod_crime2, mod_crime_PC2), BIC(mod_crime2, mod_crime_PC2))
comp2 <- comp2[, -c(2, 4, 6)]
names(comp2) <- c("Adjusted r^2", "AIC", "BIC")
comp2 %>%
    kbl(booktabs = T) %>%
    kable_classic()
```

	Adjusted r^2	AIC	BIC
mod_crime2	0.1959087	-227.4746	-202.9539
mod_crime_PC2	0.0881411	-172.1347	-147.6140

Adjusted r^2, AIC and BIC suggest the model with just total crime numbers rather than crimes per capita.

Chosen model:

```
summary(mod_crime2)

Call:
lm(formula = per.cap.income ~ crimes + region, data = income_trans)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.68757 -0.10557 -0.01422  0.08905  0.78946 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  9.188431   0.079812 115.125  < 2e-16 ***
crimes       0.066695   0.008421   7.920 2.00e-14 ***
regionNE     0.104458   0.025531   4.091 5.11e-05 ***
regionS     -0.086983   0.023618  -3.683  0.00026 ***
regionW     -0.055280   0.028167  -1.963  0.05033 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1854 on 435 degrees of freedom
Multiple R-squared:  0.2032,    Adjusted R-squared:  0.1959 
F-statistic: 27.74 on 4 and 435 DF,  p-value: < 2.2e-16
summary(mod_crime_PC2)

Call:
lm(formula = per.cap.income ~ per.cap.crime + region, data = income_trans)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.65832 -0.11431 -0.01548  0.10838  0.75657 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    9.93628    0.06934 143.303  < 2e-16 ***
per.cap.crime  0.04243    0.02148   1.975  0.04885 *  
regionNE       0.11457    0.02760   4.151 3.99e-05 ***
regionS       -0.07456    0.02624  -2.841  0.00471 ** 
regionW       -0.02426    0.03002  -0.808  0.41952    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1974 on 435 degrees of freedom
Multiple R-squared:  0.09645,   Adjusted R-squared:  0.08814 
F-statistic: 11.61 on 4 and 435 DF,  p-value: 5.776e-09
```

Question 3

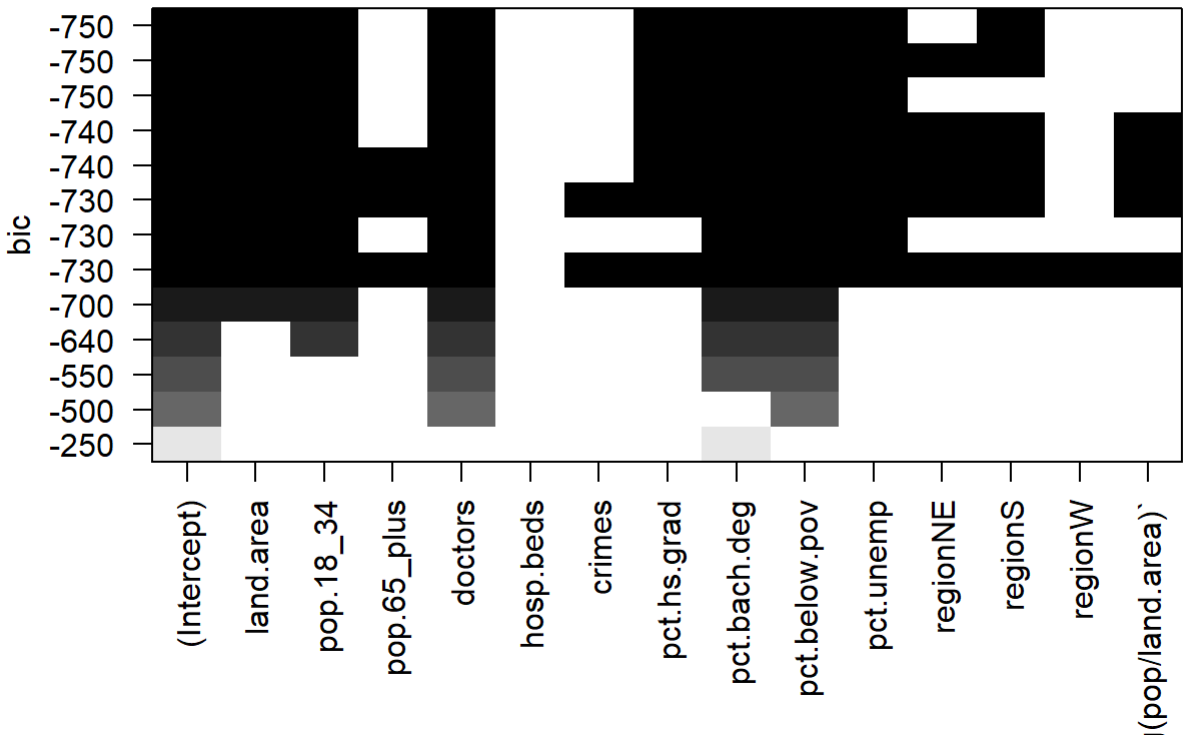
All subsets regression with no interactions

Find all possible subsets.

```
library(leaps)  #regsubsets() function doesn't seem to work unless i load this library right bef
ore using it
allsubs_mod1 <- regsubsets(per.cap.income ~ ., data = income_trans, nvmax = 13)
```

View graphical summary of all subsets, with results sorted by BIC.

```
plot(allsubs_mod1)
```

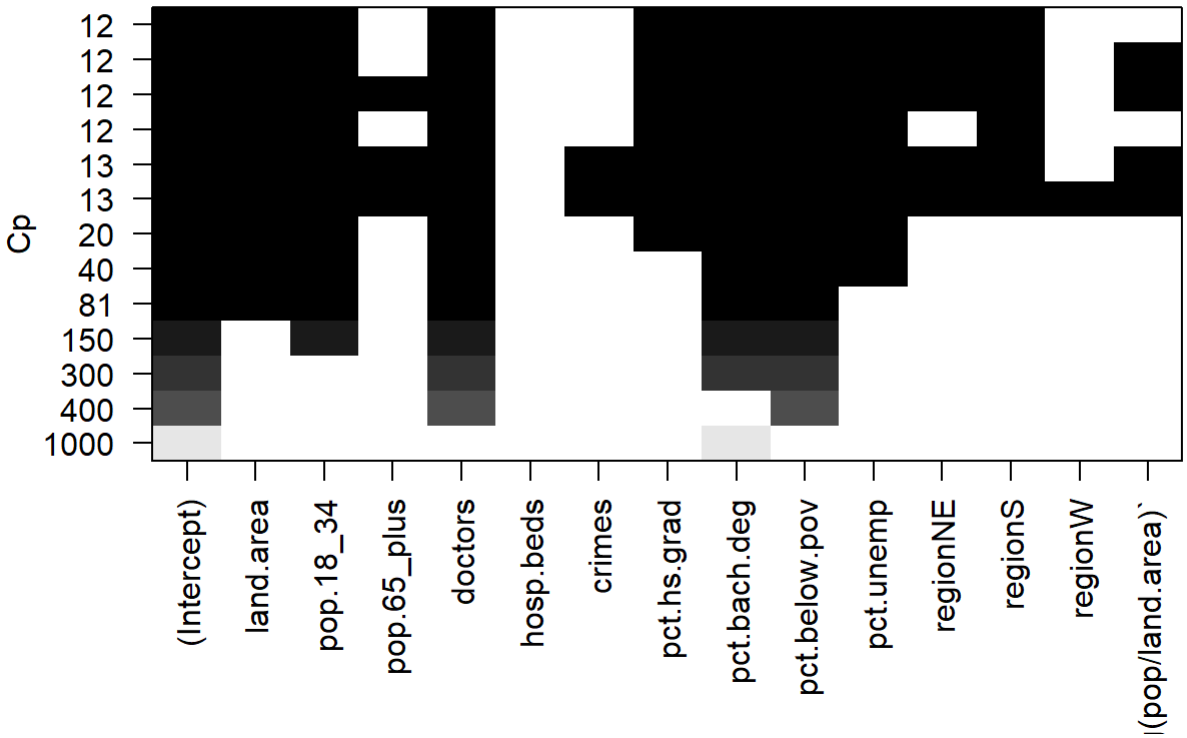


BIC for models containging different subsets of predictors

In the plot, the dark squares indicate which variables are in the model that has the BIC values on the left. So, the set of predictors resulting in the best (lowest) BIC is land.area, pop.18_34, doctors, pct.hs.grad, pct.bach.deg, pct.bel.pov, pct.unemp, and region.

View graphical summary of all subsets, with results sorted by AIC (Mallow's Cp is equivalent to AIC).

```
plot(allsubs_mod1, scale = "Cp")
```



BIC for models containging different subsets of predictors

The only extra variable selected when using AIC is an additional region indicator variable. But we would already be including all region indicator variables in the model anyway.

Fit model with predictors listed above (according to best BIC).

```
allsubs1_summary <- summary(allsubs_mod1)
best.allsubs <- which.min(allsubs1_summary$bic)
vars1 <- income_trans[, allsubs1_summary$which[best.allsubs, ][-1]]
allsubs_mod1_final <- lm(per.cap.income ~ ., data = vars1)
summary(allsubs_mod1_final)

Call:
lm(formula = per.cap.income ~ ., data = vars1)

Residuals:
    Min      1Q  Median      3Q     Max 
-6172.3 -1137.5  -180.7   811.8 10542.7 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   13825.57    2303.91   6.001 4.18e-09 ***
land.area      -785.89     121.51  -6.468 2.72e-10 ***
pop.18_34      -291.02      25.28 -11.512  < 2e-16 ***
doctors         841.27      97.40   8.637  < 2e-16 ***
pct.hs.grad    -153.57      27.54  -5.577 4.34e-08 ***
pct.bach.deg   8564.22     541.75  15.808  < 2e-16 ***
pct.below.pov  -421.72      30.72 -13.727  < 2e-16 ***
pct.unemp      1828.61     383.57   4.767 2.56e-06 ***
regionNE       -266.25     286.62  -0.929  0.35345    
regionS        -739.02     259.30  -2.850  0.00458 ** 
regionW        -102.04     319.23  -0.320  0.74940    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1846 on 429 degrees of freedom
Multiple R-squared:  0.7979,    Adjusted R-squared:  0.7932 
F-statistic: 169.4 on 10 and 429 DF,  p-value: < 2.2e-16
```

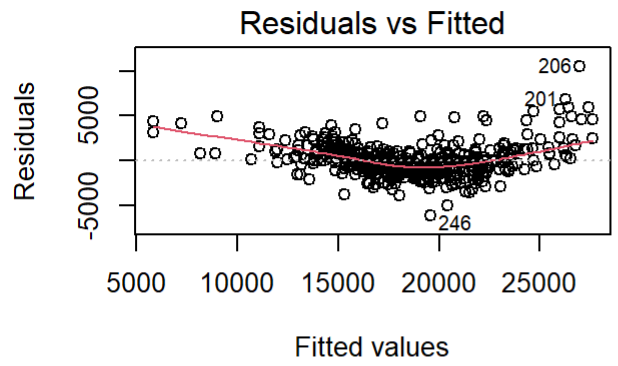
Check VIFs for collinearity.

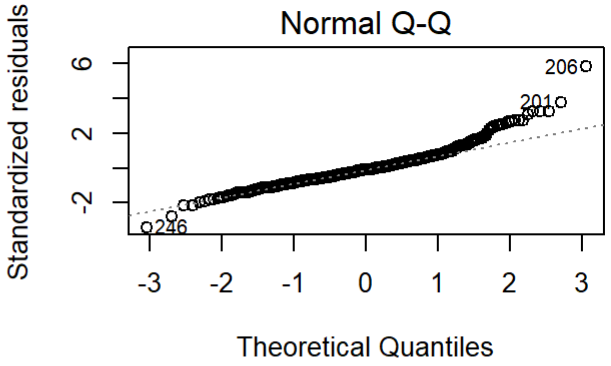
```
vif(allsubs_mod1_final)
                 GVIF Df GVIF^(1/(2*Df))
land.area     1.445259  1        1.202189
pop.18_34     1.446066  1        1.202525
doctors       1.599534  1        1.264727
pct.hs.grad   4.806990  1        2.192485
pct.bach.deg  4.749754  1        2.179393
pct.below.pov 2.636463  1        1.623719
pct.unemp     2.017310  1        1.420320
region        2.360480  3        1.153896
```

They all look fine. The biggest are pct.hs.grad and pct.bach.deg.

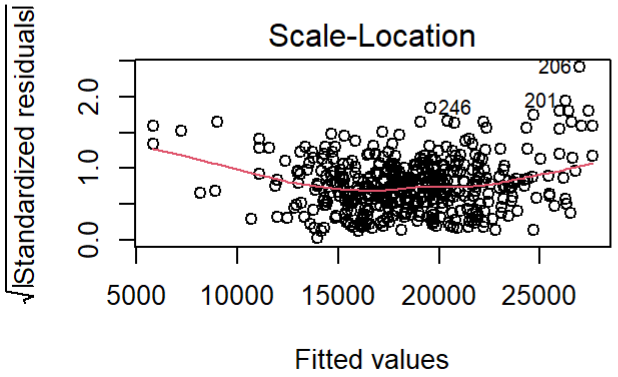
Check residual diagnostics.

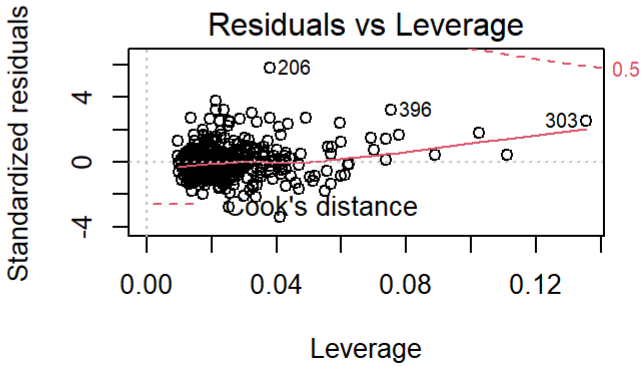
```
par(mfrow = c(2, 2))
plot(allsubs_mod1_final)
```





Not

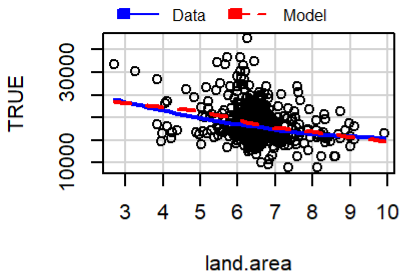


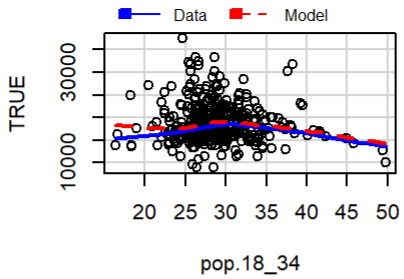


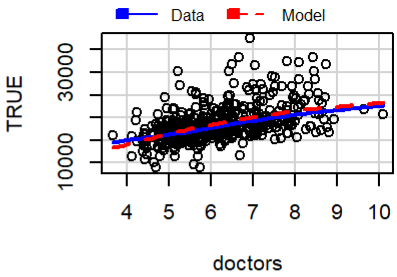
bad. QQ plot suggests residuals skew to the right a bit. Also, both the left plots show a slight dipping curve pattern, but it could be due to edge effects.

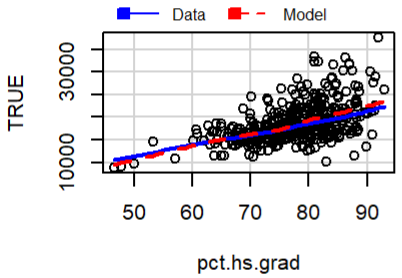
Check the marginal model plots.

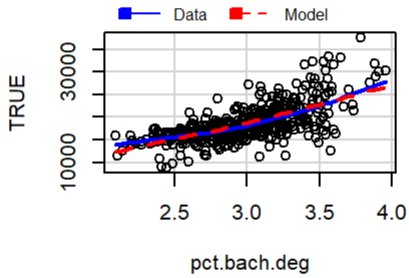
```
mmps(allsubs_mod1_final)
Warning in xy.coords(x, y, xlabel, ylabel, log): NAs introduced by coercion
Warning in min(x): no non-missing arguments to min; returning Inf
Warning in max(x): no non-missing arguments to max; returning -Inf
Error in plot.window(...): need finite 'xlim' values
```

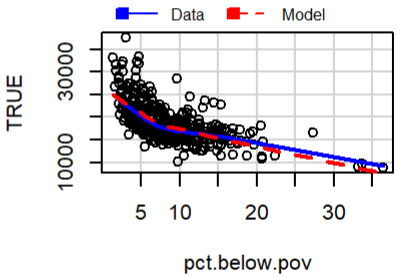




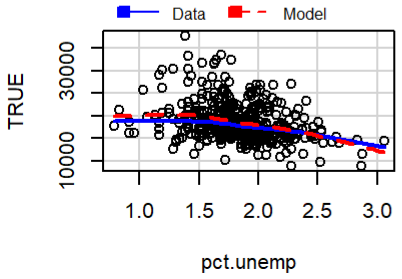








They



all look good.

All subsets regression WITH interactions

Find all possible subsets, if possible.

```
library(leaps)  #regsubsets() function doesn't seem to work unless i load this library right bef
ore using it
max_num <- 13
allsubs_mod_int_1 <- regsubsets(per.cap.income ~ .^2, data = income_trans,
    nvmax = max_num)
```

Will take too long. Try stepwise.

Stepwise selection in both directions:

Check results of AIC model:

```
summary(step_mod1_int_aic)

Call:
lm(formula = per.cap.income ~ land.area + pop.18_34 + pop.65_plus +
    doctors + hosp.beds + crimes + pct.hs.grad + pct.bach.deg +
    pct.below.pov + pct.unemp + region + `pop.dens <- log(pop/land.area)` +
    land.area:doctors + land.area:crimes + land.area:pct.bach.deg +
    land.area:pct.below.pov + pop.18_34:pct.hs.grad + pop.18_34:pct.bach.deg +
    pop.18_34:pct.below.pov + pop.18_34:pct.unemp + pop.65_plus:hosp.beds +
    pop.65_plus:crimes + pop.65_plus:pct.hs.grad + pop.65_plus:pct.below.pov +
    pop.65_plus:region + doctors:hosp.beds + doctors:crimes +
    doctors:pct.hs.grad + doctors:pct.bach.deg + doctors:pct.unemp +
    doctors:region + doctors:`pop.dens <- log(pop/land.area)` +
    hosp.beds:crimes + hosp.beds:pct.hs.grad + hosp.beds:pct.below.pov +
    hosp.beds:region + pct.hs.grad:pct.unemp + pct.hs.grad:region +
    pct.bach.deg:pct.below.pov + pct.bach.deg:region + pct.bach.deg:`pop.dens <- log(pop/land.ar
ea)` +
    pct.below.pov:region + pct.below.pov:`pop.dens <- log(pop/land.area)` +
    pct.unemp:`pop.dens <- log(pop/land.area)` + region:`pop.dens <- log(pop/land.area)`,
    data = income_trans)

Residuals:
      Min        1Q    Median        3Q       Max
-0.184402 -0.039750 -0.002335  0.039973  0.239112

Coefficients:
                                            Estimate Std. Error t value
(Intercept)                                13.8360306  2.3299836   5.938
land.area                                  -0.0090563  0.0753069  -0.120
pop.18_34                                  -0.0285447  0.0327407  -0.872
pop.65_plus                                -1.8141622  0.3957811  -4.584
doctors                                     0.1667722  0.2038879   0.818
hosp.beds                                  -0.9307014  0.2299586  -4.047
crimes                                     -0.1590750  0.0933184  -1.705
pct.hs.grad                                -0.0761900  0.0180989  -4.210
pct.bach.deg                                1.2199547  0.4375110   2.788
pct.below.pov                               0.0314657  0.0405657   0.776
pct.unemp                                   0.4372734  0.3314293   1.319
regionNE                                    0.6582747  0.4981974   1.321
regionS                                    -0.3522215  0.4065677  -0.866
regionW                                     2.0584403  0.5512922   3.734
`pop.dens <- log(pop/land.area)`            0.4676337  0.1850410   2.527
land.area:doctors                          -0.0224080  0.0094061  -2.382
land.area:crimes                            0.0313218  0.0095422   3.282
land.area:pct.bach.deg                     -0.0459515  0.0165293  -2.780
land.area:pct.below.pov                    -0.0048022  0.0014047  -3.419
pop.18_34:pct.hs.grad                       0.0005981  0.0003629   1.648
pop.18_34:pct.bach.deg                     -0.0084317  0.0034989  -2.410
pop.18_34:pct.below.pov                     0.0012205  0.0004472   2.729
pop.18_34:pct.unemp                        -0.0091137  0.0037235  -2.448
pop.65_plus:hosp.beds                       0.0560861  0.0265317   2.114
pop.65_plus:crimes                         -0.0400255  0.0257837  -1.552
pop.65_plus:pct.hs.grad                     0.0208523  0.0042911   4.859
```

```
pop.65_plus:pct.below.pov                          0.0205159  0.0075045   2.734
pop.65_plus:regionNE                              -0.0472866  0.0652523  -0.725
pop.65_plus:regionS                                0.0775740  0.0545136   1.423
pop.65_plus:regionW                               -0.1242079  0.0687318  -1.807
doctors:hosp.beds                                 -0.0315112  0.0106732  -2.952
doctors:crimes                                    -0.0448184  0.0127992  -3.502
doctors:pct.hs.grad                               -0.0051732  0.0019724  -2.623
doctors:pct.bach.deg                               0.1996827  0.0331832   6.018
doctors:pct.unemp                                  0.0820985  0.0292385   2.808
doctors:regionNE                                  -0.0774214  0.0395962  -1.955
doctors:regionS                                   -0.0466280  0.0323115  -1.443
doctors:regionW                                    0.1143320  0.0468215   2.442
doctors:`pop.dens <- log(pop/land.area)`           0.0299136  0.0143684   2.082
hosp.beds:crimes                                   0.0547335  0.0128403   4.263
hosp.beds:pct.hs.grad                              0.0048087  0.0021747   2.211
hosp.beds:pct.below.pov                            0.0121460  0.0024358   4.986
hosp.beds:regionNE                                 0.0853173  0.0351758   2.425
hosp.beds:regionS                                  0.0233441  0.0285970   0.816
hosp.beds:regionW                                 -0.0273571  0.0342049  -0.800
pct.hs.grad:pct.unemp                              0.0054139  0.0019162   2.825
pct.hs.grad:regionNE                              -0.0139852  0.0039587  -3.533
pct.hs.grad:regionS                               -0.0065048  0.0034542  -1.883
pct.hs.grad:regionW                               -0.0184918  0.0039458  -4.686
pct.bach.deg:pct.below.pov                        -0.0232609  0.0038571  -6.031
pct.bach.deg:regionNE                              0.1516838  0.0730496   2.076
pct.bach.deg:regionS                               0.2033754  0.0618616   3.288
pct.bach.deg:regionW                               0.2418542  0.0712114   3.396
pct.bach.deg:`pop.dens <- log(pop/land.area)`     -0.1417433  0.0486059  -2.916
pct.below.pov:regionNE                            -0.0108678  0.0048321  -2.249
pct.below.pov:regionS                              0.0004294  0.0038568   0.111
pct.below.pov:regionW                             -0.0110240  0.0053855  -2.047
pct.below.pov:`pop.dens <- log(pop/land.area)`    -0.0116059  0.0028391  -4.088
pct.unemp:`pop.dens <- log(pop/land.area)`        -0.0994389  0.0405416  -2.453
regionNE:`pop.dens <- log(pop/land.area)`          0.0088610  0.0373881   0.237
regionS:`pop.dens <- log(pop/land.area)`           0.0151116  0.0295414   0.512
regionW:`pop.dens <- log(pop/land.area)`          -0.1377896  0.0421575  -3.268
                                                  Pr(>|t|)
(Intercept)                                       6.53e-09 ***
land.area                                         0.904342
pop.18_34                                         0.383849
pop.65_plus                                       6.22e-06 ***
doctors                                           0.413895
hosp.beds                                         6.29e-05 ***
crimes                                            0.089082 .
pct.hs.grad                                       3.20e-05 ***
pct.bach.deg                                      0.005565 **
pct.below.pov                                     0.438427
pct.unemp                                         0.187849
regionNE                                          0.187196
regionS                                           0.386859
regionW                                           0.000218 ***
`pop.dens <- log(pop/land.area)`                  0.011905 *
land.area:doctors                                 0.017700 *
land.area:crimes                                  0.001125 **
```

```
land.area:pct.bach.deg                          0.005708 **
land.area:pct.below.pov                         0.000698 ***
pop.18_34:pct.hs.grad                           0.100149
pop.18_34:pct.bach.deg                          0.016438 *
pop.18_34:pct.below.pov                         0.006644 **
pop.18_34:pct.unemp                             0.014835 *
pop.65_plus:hosp.beds                           0.035176 *
pop.65_plus:crimes                              0.121414
pop.65_plus:pct.hs.grad                         1.73e-06 ***
pop.65_plus:pct.below.pov                       0.006555 **
pop.65_plus:regionNE                            0.469100
pop.65_plus:regionS                             0.155555
pop.65_plus:regionW                             0.071536 .
doctors:hosp.beds                               0.003350 **
doctors:crimes                                  0.000518 ***
doctors:pct.hs.grad                             0.009073 **
doctors:pct.bach.deg                            4.19e-09 ***
doctors:pct.unemp                               0.005245 **
doctors:regionNE                                0.051287 .
doctors:regionS                                 0.149826
doctors:regionW                                 0.015069 *
doctors:`pop.dens <- log(pop/land.area)`        0.038023 *
hosp.beds:crimes                                2.55e-05 ***
hosp.beds:pct.hs.grad                           0.027616 *
hosp.beds:pct.below.pov                         9.39e-07 ***
hosp.beds:regionNE                              0.015757 *
hosp.beds:regionS                               0.414836
hosp.beds:regionW                               0.424330
pct.hs.grad:pct.unemp                           0.004975 **
pct.hs.grad:regionNE                            0.000462 ***
pct.hs.grad:regionS                             0.060450 .
pct.hs.grad:regionW                             3.89e-06 ***
pct.bach.deg:pct.below.pov                      3.89e-09 ***
pct.bach.deg:regionNE                           0.038528 *
pct.bach.deg:regionS                            0.001105 **
pct.bach.deg:regionW                            0.000756 ***
pct.bach.deg:`pop.dens <- log(pop/land.area)`   0.003755 **
pct.below.pov:regionNE                          0.025082 *
pct.below.pov:regionS                           0.911418
pct.below.pov:regionW                           0.041350 *
pct.below.pov:`pop.dens <- log(pop/land.area)`  5.32e-05 ***
pct.unemp:`pop.dens <- log(pop/land.area)`      0.014628 *
regionNE:`pop.dens <- log(pop/land.area)`       0.812786
regionS:`pop.dens <- log(pop/land.area)`        0.609272
regionW:`pop.dens <- log(pop/land.area)`        0.001180 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06304 on 378 degrees of freedom
Multiple R-squared:  0.9199,    Adjusted R-squared:  0.907
F-statistic: 71.21 on 61 and 378 DF,  p-value: < 2.2e-16
```

Check results of BIC model:

```
summary(step_mod1_int_bic)

Call:
lm(formula = per.cap.income ~ land.area + pop.18_34 + pop.65_plus +
    doctors + hosp.beds + crimes + pct.hs.grad + pct.bach.deg +
    pct.below.pov + pct.unemp + region + `pop.dens <- log(pop/land.area)` +
    land.area:crimes + land.area:pct.bach.deg + land.area:pct.below.pov +
    pop.18_34:pct.below.pov + pop.65_plus:hosp.beds + pop.65_plus:pct.hs.grad +
    doctors:hosp.beds + hosp.beds:pct.bach.deg + hosp.beds:pct.below.pov +
    hosp.beds:`pop.dens <- log(pop/land.area)` + pct.hs.grad:pct.below.pov +
    pct.hs.grad:region + pct.bach.deg:pct.below.pov + pct.bach.deg:region +
    pct.bach.deg:`pop.dens <- log(pop/land.area)` + pct.below.pov:`pop.dens <- log(pop/land.are
a)`,
    data = income_trans)

Residuals:
      Min        1Q    Median        3Q       Max
-0.197676 -0.043717 -0.003027  0.039622  0.257992

Coefficients:
                                              Estimate Std. Error t value
(Intercept)                                 13.4030138  1.2776260  10.491
land.area                                    0.1626012  0.0555324   2.928
pop.18_34                                   -0.0211499  0.0025609  -8.259
pop.65_plus                                 -1.2380950  0.1970007  -6.285
doctors                                      0.2854667  0.0734082   3.889
hosp.beds                                   -0.9646417  0.1595885  -6.045
crimes                                      -0.0306957  0.0252537  -1.215
pct.hs.grad                                 -0.0328966  0.0055871  -5.888
pct.bach.deg                                 0.7501171  0.2548714   2.943
pct.below.pov                                0.1043814  0.0220345   4.737
pct.unemp                                    0.0562607  0.0156260   3.600
regionNE                                     0.1406865  0.1473416   0.955
regionS                                     -0.0961725  0.1347016  -0.714
regionW                                      0.0579870  0.1550816   0.374
`pop.dens <- log(pop/land.area)`            -0.0308830  0.1229578  -0.251
land.area:crimes                             0.0107801  0.0037889   2.845
land.area:pct.bach.deg                      -0.0785971  0.0136450  -5.760
land.area:pct.below.pov                     -0.0063417  0.0011753  -5.396
pop.18_34:pct.below.pov                      0.0009146  0.0002539   3.602
pop.65_plus:hosp.beds                        0.0494506  0.0148558   3.329
pop.65_plus:pct.hs.grad                      0.0117123  0.0020703   5.657
doctors:hosp.beds                           -0.0377165  0.0106797  -3.532
hosp.beds:pct.bach.deg                       0.1509009  0.0263037   5.737
hosp.beds:pct.below.pov                      0.0173220  0.0020418   8.484
hosp.beds:`pop.dens <- log(pop/land.area)`   0.0468497  0.0131358   3.567
pct.hs.grad:pct.below.pov                    0.0003469  0.0001234   2.811
pct.hs.grad:regionNE                        -0.0058292  0.0028574  -2.040
pct.hs.grad:regionS                         -0.0047366  0.0027096  -1.748
pct.hs.grad:regionW                         -0.0127232  0.0027838  -4.570
pct.bach.deg:pct.below.pov                  -0.0309729  0.0044555  -6.952
pct.bach.deg:regionNE                        0.1048833  0.0449020   2.336
pct.bach.deg:regionS                         0.1420931  0.0438547   3.240
```

```
pct.bach.deg:regionW                              0.3077393  0.0514393   5.983
pct.bach.deg:`pop.dens <- log(pop/land.area)`    -0.0699188  0.0293127  -2.385
pct.below.pov:`pop.dens <- log(pop/land.area)` -0.0158865  0.0023523  -6.754
                                                 Pr(>|t|)
(Intercept)                                       < 2e-16 ***
land.area                                        0.003604 **
pop.18_34                                        2.11e-15 ***
pop.65_plus                                      8.50e-10 ***
doctors                                          0.000118 ***
hosp.beds                                        3.40e-09 ***
crimes                                           0.224886
pct.hs.grad                                      8.22e-09 ***
pct.bach.deg                                     0.003436 **
pct.below.pov                                    3.00e-06 ***
pct.unemp                                        0.000357 ***
regionNE                                         0.340232
regionS                                          0.475659
regionW                                          0.708665
`pop.dens <- log(pop/land.area)`                 0.801812
land.area:crimes                                 0.004664 **
land.area:pct.bach.deg                           1.66e-08 ***
land.area:pct.below.pov                          1.16e-07 ***
pop.18_34:pct.below.pov                          0.000354 ***
pop.65_plus:hosp.beds                            0.000952 ***
pop.65_plus:pct.hs.grad                          2.91e-08 ***
doctors:hosp.beds                                0.000461 ***
hosp.beds:pct.bach.deg                           1.89e-08 ***
hosp.beds:pct.below.pov                          4.14e-16 ***
hosp.beds:`pop.dens <- log(pop/land.area)`       0.000405 ***
pct.hs.grad:pct.below.pov                        0.005179 **
pct.hs.grad:regionNE                             0.041996 *
pct.hs.grad:regionS                              0.081204 .
pct.hs.grad:regionW                              6.47e-06 ***
pct.bach.deg:pct.below.pov                       1.45e-11 ***
pct.bach.deg:regionNE                            0.019988 *
pct.bach.deg:regionS                             0.001294 **
pct.bach.deg:regionW                             4.83e-09 ***
pct.bach.deg:`pop.dens <- log(pop/land.area)`    0.017526 *
pct.below.pov:`pop.dens <- log(pop/land.area)` 5.02e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06597 on 405 degrees of freedom
Multiple R-squared:  0.9061,    Adjusted R-squared:  0.8982
F-statistic: 114.9 on 34 and 405 DF,  p-value: < 2.2e-16
```

First examine the smaller BIC model. If all good, we won't bother with the AIC one. Add back in the missing indicator variables for any region interaction if necessary. (not necessary).

The adjusted r^2 of this model with interactions is way better than the model without (R2 = 0.901 vs R2 = 0.793), but let's check if the additional terms are actually necessary using ANOVA.

```
anova(allsubs_mod1_final, step_mod1_int_bic)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	429	1.462014e+09	*NA*	*NA*	*NA*	*NA*
2	405	1.762498e+00	24	1462014016	13998019741	0

2 rows

The F-test statistic is extremely significant, so it seems that the interaction terms are actually needed.
However, there are 40 coefficients in this model, many of them strange interactions. The model is too complex to explain, so i'll look for reasons to remove terms, especially interactions.

I'll start removing any interactions with p > 0.05 or |coefficient| < 0.01.
Interactions that fit this criteria are:
pop.18_34:pct.below.pov
doctors:pct.below.pov
hosp.beds:pct.below.pov
pct.hs.grad:pct.unemp

Update model with these 4 removed.

```
step_mod1_int_bic_reduced <- update(step_mod1_int_bic, . ~ . - pop.18_34:pct.below.pov -
    doctors:pct.below.pov - hosp.beds:pct.below.pov - pct.hs.grad:pct.unemp)
summary(step_mod1_int_bic_reduced)

Call:
lm(formula = per.cap.income ~ land.area + pop.18_34 + pop.65_plus +
    doctors + hosp.beds + crimes + pct.hs.grad + pct.bach.deg +
    pct.below.pov + pct.unemp + region + `pop.dens <- log(pop/land.area)` +
    land.area:crimes + land.area:pct.bach.deg + land.area:pct.below.pov +
    pop.65_plus:hosp.beds + pop.65_plus:pct.hs.grad + doctors:hosp.beds +
    hosp.beds:pct.bach.deg + hosp.beds:`pop.dens <- log(pop/land.area)` +
    pct.hs.grad:pct.below.pov + pct.hs.grad:region + pct.bach.deg:pct.below.pov +
    pct.bach.deg:region + pct.bach.deg:`pop.dens <- log(pop/land.area)` +
    pct.below.pov:`pop.dens <- log(pop/land.area)`, data = income_trans)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21972 -0.04464 -0.00341  0.04344  0.25483

Coefficients:
                                              Estimate Std. Error t value
(Intercept)                                 12.7548183  1.3830270   9.222
land.area                                    0.1779038  0.0603481   2.948
pop.18_34                                   -0.0157454  0.0013671 -11.518
pop.65_plus                                 -1.1396001  0.2093886  -5.443
doctors                                      0.0827703  0.0753783   1.098
hosp.beds                                   -0.2595784  0.1481522  -1.752
crimes                                      -0.0417487  0.0274713  -1.520
pct.hs.grad                                 -0.0313975  0.0060244  -5.212
pct.bach.deg                                 0.5266091  0.2754300   1.912
pct.below.pov                                0.0677246  0.0220132   3.077
pct.unemp                                    0.0545445  0.0170071   3.207
regionNE                                     0.0126738  0.1589319   0.080
regionS                                     -0.1556942  0.1459824  -1.067
regionW                                      0.0331977  0.1678535   0.198
`pop.dens <- log(pop/land.area)`            -0.1679745  0.1328051  -1.265
land.area:crimes                             0.0116340  0.0041253   2.820
land.area:pct.bach.deg                      -0.0834208  0.0147960  -5.638
land.area:pct.below.pov                     -0.0079315  0.0012647  -6.271
pop.65_plus:hosp.beds                        0.0363390  0.0160391   2.266
pop.65_plus:pct.hs.grad                      0.0111950  0.0021999   5.089
doctors:hosp.beds                           -0.0034149  0.0107606  -0.317
hosp.beds:pct.bach.deg                       0.0333176  0.0246554   1.351
hosp.beds:`pop.dens <- log(pop/land.area)`   0.0080454  0.0133447   0.603
pct.hs.grad:pct.below.pov                    0.0002122  0.0001326   1.600
pct.hs.grad:regionNE                        -0.0043477  0.0030923  -1.406
pct.hs.grad:regionS                         -0.0035809  0.0029442  -1.216
pct.hs.grad:regionW                         -0.0108161  0.0030166  -3.586
pct.bach.deg:pct.below.pov                  -0.0184415  0.0044069  -4.185
pct.bach.deg:regionNE                        0.1087114  0.0488273   2.226
pct.bach.deg:regionS                         0.1331377  0.0477538   2.788
pct.bach.deg:regionW                         0.2618614  0.0557376   4.698
pct.bach.deg:`pop.dens <- log(pop/land.area)`   0.0194101  0.0299474   0.648
```

```
pct.below.pov:`pop.dens <- log(pop/land.area)` -0.0002698  0.0015068  -0.179
                                               Pr(>|t|)
(Intercept)                                     < 2e-16 ***
land.area                                      0.003383 **
pop.18_34                                       < 2e-16 ***
pop.65_plus                                    9.09e-08 ***
doctors                                        0.272825
hosp.beds                                      0.080509 .
crimes                                         0.129358
pct.hs.grad                                    2.98e-07 ***
pct.bach.deg                                   0.056584 .
pct.below.pov                                  0.002236 **
pct.unemp                                      0.001446 **
regionNE                                       0.936481
regionS                                        0.286817
regionW                                        0.843318
`pop.dens <- log(pop/land.area)`               0.206660
land.area:crimes                               0.005034 **
land.area:pct.bach.deg                         3.22e-08 ***
land.area:pct.below.pov                        9.15e-10 ***
pop.65_plus:hosp.beds                          0.023998 *
pop.65_plus:pct.hs.grad                        5.51e-07 ***
doctors:hosp.beds                              0.751140
hosp.beds:pct.bach.deg                         0.177340
hosp.beds:`pop.dens <- log(pop/land.area)`     0.546918
pct.hs.grad:pct.below.pov                      0.110361
pct.hs.grad:regionNE                           0.160504
pct.hs.grad:regionS                            0.224601
pct.hs.grad:regionW                            0.000377 ***
pct.bach.deg:pct.below.pov                     3.50e-05 ***
pct.bach.deg:regionNE                          0.026531 *
pct.bach.deg:regionS                           0.005552 **
pct.bach.deg:regionW                           3.60e-06 ***
pct.bach.deg:`pop.dens <- log(pop/land.area)`  0.517259
pct.below.pov:`pop.dens <- log(pop/land.area)` 0.857986
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07185 on 407 degrees of freedom
Multiple R-squared:  0.888, Adjusted R-squared:  0.8792
F-statistic: 100.8 on 32 and 407 DF,  p-value: < 2.2e-16
```

The model is still too complex to explain, so i'll try to remove more terms.
The region categorical variable is not significant on its own. It seems that its effect is captured in interactions with the two continuous variables, pct.hs.grad and pct.bach.deg. So try removing region and the 3 interactions with p > 0.05:
pop.18_34:doctors
pct.bach.deg:pop.dens
pct.below.pov:pop.dens

Update model with these removed.

```
step_mod2_int_bic_reduced <- update(step_mod1_int_bic_reduced, . ~ . -
    region - pop.18_34:doctors - pct.bach.deg:`pop.dens <- log(pop/land.area)` -
    pct.below.pov:`pop.dens <- log(pop/land.area)`)
summary(step_mod2_int_bic_reduced)

Call:
lm(formula = per.cap.income ~ land.area + pop.18_34 + pop.65_plus +
    doctors + hosp.beds + crimes + pct.hs.grad + pct.bach.deg +
    pct.below.pov + pct.unemp + `pop.dens <- log(pop/land.area)` +
    land.area:crimes + land.area:pct.bach.deg + land.area:pct.below.pov +
    pop.65_plus:hosp.beds + pop.65_plus:pct.hs.grad + doctors:hosp.beds +
    hosp.beds:pct.bach.deg + hosp.beds:`pop.dens <- log(pop/land.area)` +
    pct.hs.grad:pct.below.pov + pct.hs.grad:region + pct.bach.deg:pct.below.pov +
    pct.bach.deg:region, data = income_trans)

Residuals:
      Min        1Q    Median        3Q       Max
-0.226069 -0.045116 -0.001997  0.043606  0.259464

Coefficients:
                                          Estimate Std. Error t value
(Intercept)                             12.2370598  1.0323889  11.853
land.area                                0.2025881  0.0558824   3.625
pop.18_34                               -0.0157331  0.0013249 -11.875
pop.65_plus                             -1.1192164  0.1920231  -5.829
doctors                                  0.1006792  0.0732282   1.375
hosp.beds                               -0.3283370  0.1331207  -2.466
crimes                                  -0.0332015  0.0268382  -1.237
pct.hs.grad                             -0.0296441  0.0055288  -5.362
pct.bach.deg                             0.6817137  0.1413404   4.823
pct.below.pov                            0.0621402  0.0120118   5.173
pct.unemp                                0.0606900  0.0161983   3.747
`pop.dens <- log(pop/land.area)`        -0.1268483  0.0925689  -1.370
land.area:crimes                         0.0103228  0.0040330   2.560
land.area:pct.bach.deg                  -0.0879934  0.0135145  -6.511
land.area:pct.below.pov                 -0.0076810  0.0011777  -6.522
pop.65_plus:hosp.beds                    0.0406030  0.0151324   2.683
pop.65_plus:pct.hs.grad                  0.0105397  0.0020131   5.236
doctors:hosp.beds                       -0.0057875  0.0103983  -0.557
hosp.beds:pct.bach.deg                   0.0467878  0.0159549   2.932
hosp.beds:`pop.dens <- log(pop/land.area)`  0.0107765  0.0131257   0.821
pct.hs.grad:pct.below.pov                0.0003290  0.0001118   2.943
pct.hs.grad:regionNE                    -0.0041515  0.0017693  -2.346
pct.hs.grad:regionS                     -0.0063016  0.0016702  -3.773
pct.hs.grad:regionW                     -0.0113228  0.0019757  -5.731
pct.bach.deg:pct.below.pov              -0.0210221  0.0037390  -5.622
pct.bach.deg:regionNE                    0.1081013  0.0465941   2.320
pct.bach.deg:regionS                     0.1517551  0.0437551   3.468
pct.bach.deg:regionW                     0.2853825  0.0519334   5.495
                                        Pr(>|t|)
(Intercept)                              < 2e-16 ***
land.area                               0.000325 ***
pop.18_34                                < 2e-16 ***
```

```
pop.65_plus                                 1.13e-08 ***
doctors                                     0.169919
hosp.beds                                   0.014052 *
crimes                                      0.216754
pct.hs.grad                                 1.38e-07 ***
pct.bach.deg                                1.99e-06 ***
pct.below.pov                               3.60e-07 ***
pct.unemp                                   0.000205 ***
`pop.dens <- log(pop/land.area)`            0.171336
land.area:crimes                            0.010837 *
land.area:pct.bach.deg                      2.17e-10 ***
land.area:pct.below.pov                     2.04e-10 ***
pop.65_plus:hosp.beds                       0.007586 **
pop.65_plus:pct.hs.grad                     2.63e-07 ***
doctors:hosp.beds                           0.578118
hosp.beds:pct.bach.deg                      0.003550 **
hosp.beds:`pop.dens <- log(pop/land.area)` 0.412108
pct.hs.grad:pct.below.pov                   0.003437 **
pct.hs.grad:regionNE                        0.019428 *
pct.hs.grad:regionS                         0.000185 ***
pct.hs.grad:regionW                         1.93e-08 ***
pct.bach.deg:pct.below.pov                  3.48e-08 ***
pct.bach.deg:regionNE                       0.020825 *
pct.bach.deg:regionS                        0.000579 ***
pct.bach.deg:regionW                        6.85e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07172 on 412 degrees of freedom
Multiple R-squared:  0.887, Adjusted R-squared:  0.8796
F-statistic: 119.8 on 27 and 412 DF,  p-value: < 2.2e-16
```

After each variable removal, the adjusted r^2 slightly decreases, but the BIC doesn't always get worse.

```
cat("BIC of stepwise model: ", BIC(step_mod1_int_bic), "\n")
BIC of stepwise model:  -961.0289
cat("BIC after first variable group removal: ", BIC(step_mod1_int_bic_reduced),
    "\n")
BIC after first variable group removal:  -895.8444
cat("BIC after second variable group removal: ", BIC(step_mod2_int_bic_reduced))
BIC after second variable group removal:  -922.5173
```

Let's try removing a few more. We'll remove variables for the reasons above as well as variables that are borderline cases for the reasons above and additionally don't make any sense from an interpretation standpoint:
land.area (p >> 0.05)
pct.hs.grad:region (|each of the 3 coefficients| < 0.01)
land.area:pop.18_34 (|coefficient| < 0.01)
doctors:crimes (coefficient is pretty small and interaction doesn't make sense)
hosp.beds:crimes (coefficient is pretty small and interaction doesn't make sense)
doctors:hosp.beds (coefficient is pretty small and interaction doesn't make sense with negative coefficient)

```
step_mod3_int_bic_reduced <- update(step_mod2_int_bic_reduced, . ~ . -
    land.area - pct.hs.grad:region - land.area:pop.18_34 - doctors:crimes -
    hosp.beds:crimes - doctors:hosp.beds)
summary(step_mod3_int_bic_reduced)

Call:
lm(formula = per.cap.income ~ pop.18_34 + pop.65_plus + doctors +
    hosp.beds + crimes + pct.hs.grad + pct.bach.deg + pct.below.pov +
    pct.unemp + `pop.dens <- log(pop/land.area)` + land.area:crimes +
    land.area:pct.bach.deg + land.area:pct.below.pov + pop.65_plus:hosp.beds +
    pop.65_plus:pct.hs.grad + hosp.beds:pct.bach.deg + hosp.beds:`pop.dens <- log(pop/land.area)
` +
    pct.hs.grad:pct.below.pov + pct.bach.deg:pct.below.pov +
    pct.bach.deg:region, data = income_trans)

Residuals:
      Min        1Q    Median        3Q       Max
-0.288975 -0.047049 -0.001421  0.046313  0.255296

Coefficients:
                                          Estimate Std. Error t value
(Intercept)                             13.2852770  0.5970178  22.253
pop.18_34                               -0.0158794  0.0013448 -11.808
pop.65_plus                             -1.1274435  0.1960884  -5.750
doctors                                  0.0547358  0.0135027   4.054
hosp.beds                               -0.2777313  0.0651500  -4.263
crimes                                  -0.0802114  0.0232761  -3.446
pct.hs.grad                             -0.0363573  0.0053121  -6.844
pct.bach.deg                             0.5939322  0.1043820   5.690
pct.below.pov                            0.0486098  0.0122648   3.963
pct.unemp                                0.0804189  0.0164932   4.876
`pop.dens <- log(pop/land.area)`        -0.0747512  0.0381528  -1.959
crimes:land.area                         0.0169840  0.0034401   4.937
pct.bach.deg:land.area                  -0.0514006  0.0092544  -5.554
pct.below.pov:land.area                 -0.0057988  0.0011375  -5.098
pop.65_plus:hosp.beds                    0.0387445  0.0136085   2.847
pop.65_plus:pct.hs.grad                  0.0108588  0.0020218   5.371
hosp.beds:pct.bach.deg                   0.0479036  0.0132899   3.604
hosp.beds:`pop.dens <- log(pop/land.area)`  0.0033941  0.0041519   0.817
pct.hs.grad:pct.below.pov                0.0003511  0.0001124   3.123
pct.bach.deg:pct.below.pov              -0.0210726  0.0037352  -5.642
pct.bach.deg:regionNE                   -0.0055464  0.0041500  -1.336
pct.bach.deg:regionS                    -0.0152902  0.0039608  -3.860
pct.bach.deg:regionW                    -0.0095358  0.0047467  -2.009
                                        Pr(>|t|)
(Intercept)                              < 2e-16 ***
pop.18_34                                < 2e-16 ***
pop.65_plus                             1.73e-08 ***
doctors                                 6.02e-05 ***
hosp.beds                               2.50e-05 ***
crimes                                  0.000627 ***
pct.hs.grad                             2.76e-11 ***
pct.bach.deg                            2.40e-08 ***
```

```
pct.below.pov                              8.69e-05 ***
pct.unemp                                  1.54e-06 ***
`pop.dens <- log(pop/land.area)`           0.050748 .
crimes:land.area                           1.15e-06 ***
pct.bach.deg:land.area                     4.98e-08 ***
pct.below.pov:land.area                    5.22e-07 ***
pop.65_plus:hosp.beds                      0.004630 **
pop.65_plus:pct.hs.grad                    1.30e-07 ***
hosp.beds:pct.bach.deg                     0.000351 ***
hosp.beds:`pop.dens <- log(pop/land.area)` 0.414131
pct.hs.grad:pct.below.pov                  0.001914 **
pct.bach.deg:pct.below.pov                 3.11e-08 ***
pct.bach.deg:regionNE                      0.182120
pct.bach.deg:regionS                       0.000131 ***
pct.bach.deg:regionW                       0.045189 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07499 on 417 degrees of freedom
Multiple R-squared:  0.875, Adjusted R-squared:  0.8684
F-statistic: 132.7 on 22 and 417 DF,  p-value: < 2.2e-16
BIC(step_mod3_int_bic_reduced)
[1] -908.462
```

This model still has good BIC and the number of terms is becoming more manageable. Some will still be hard to explain, but we'll accept that consequence of having a good model that represents the data.

Final model based on stepwise BIC selection and manually removing variables:

```
model_bic_1 <- step_mod3_int_bic_reduced
```

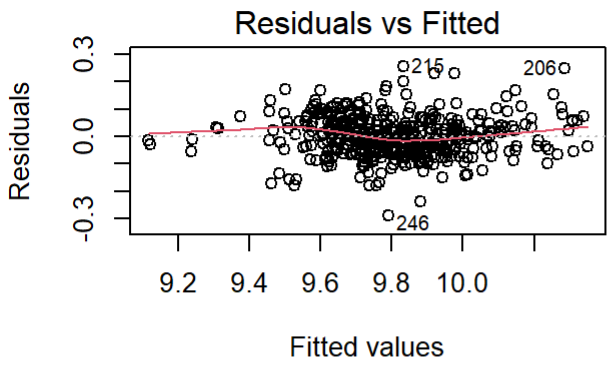
Check diagnostics for BIC reduced model. Check VIFs for collinearity.

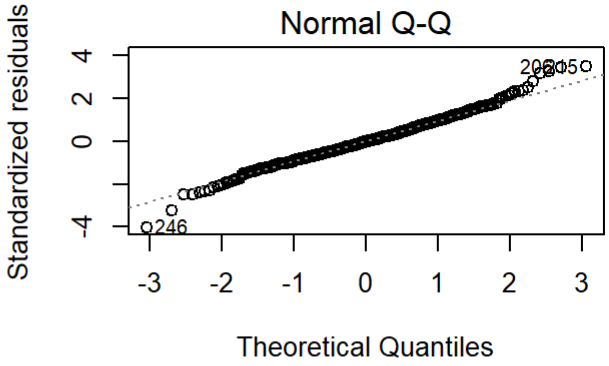
```
vif(model_bic_1)
                                            GVIF Df GVIF^(1/(2*Df))
pop.18_34                                2.479880  1        1.574763
pop.65_plus                            275.130515  1       16.587059
doctors                                 18.630616  1        4.316320
hosp.beds                              333.578814  1       18.264140
crimes                                  49.530158  1        7.037767
pct.hs.grad                            108.418607  1       10.412426
pct.bach.deg                           106.866761  1       10.337638
pct.below.pov                          254.667941  1       15.958319
pct.unemp                                2.260547  1        1.503511
`pop.dens <- log(pop/land.area)`        73.046796  1        8.546742
crimes:land.area                       118.592366  1       10.890012
pct.bach.deg:land.area                  66.243001  1        8.138980
pct.below.pov:land.area                113.732756  1       10.664556
pop.65_plus:hosp.beds                  173.142620  1       13.158367
pop.65_plus:pct.hs.grad                189.515818  1       13.766474
hosp.beds:pct.bach.deg                 243.984934  1       15.620017
hosp.beds:`pop.dens <- log(pop/land.area)` 347.120849  1       18.631179
pct.hs.grad:pct.below.pov               83.126242  1        9.117359
pct.bach.deg:pct.below.pov             169.578750  1       13.022241
pct.bach.deg:region                      4.844006  3        1.300771
```

There is plenty of collinearity, but that is expected with the interaction terms included.

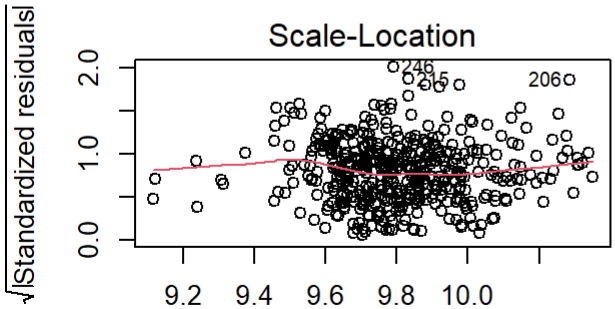
Check residual diagnostics.

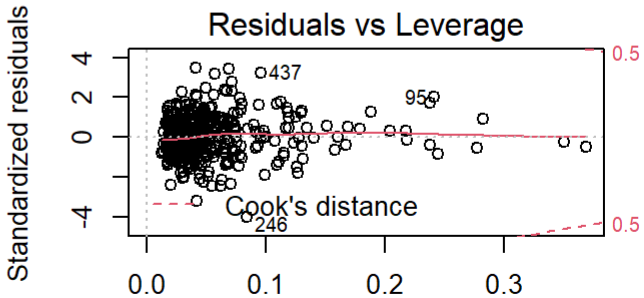
```
par(mfrow = c(2, 2))
plot(model_bic_1)
```





The

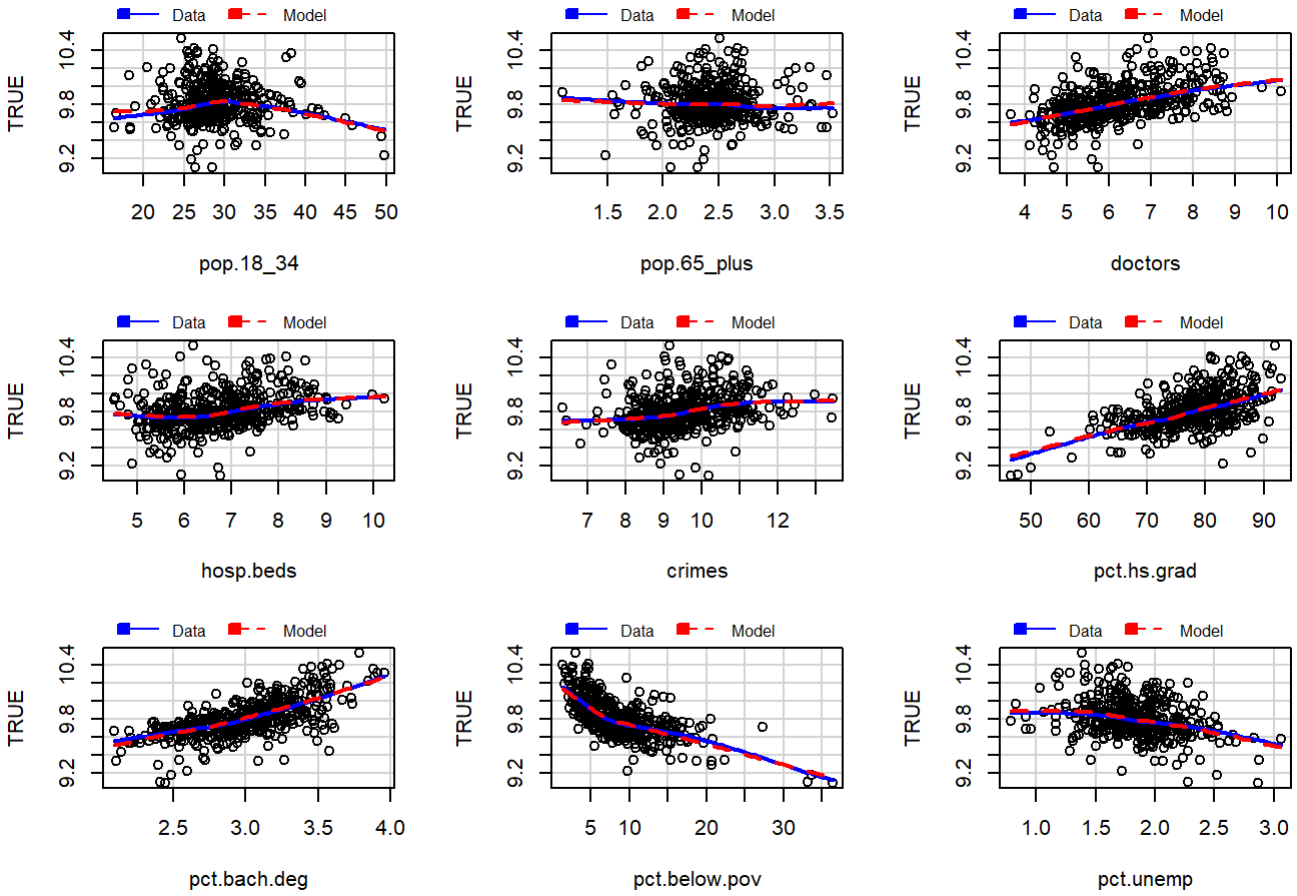




residual diagnostic plots look good for the most part. The QQ plot suggests residuals skew to the right a bit. Also, the scale-location plot shows a slight dipping curve pattern, but it could be due to edge effects.

Check the marginal model plots.

```
mmps(model_bic_1)
Error in `[.data.frame`(mf$mf.vars, , labels2[j]): undefined columns selected
```



The marginal model plots for the individual continuous predictors still look good.

#LASSO

Next, try LASSO regression to estimate coefficients and select variables simultaneously.

LASSO with no interactions:

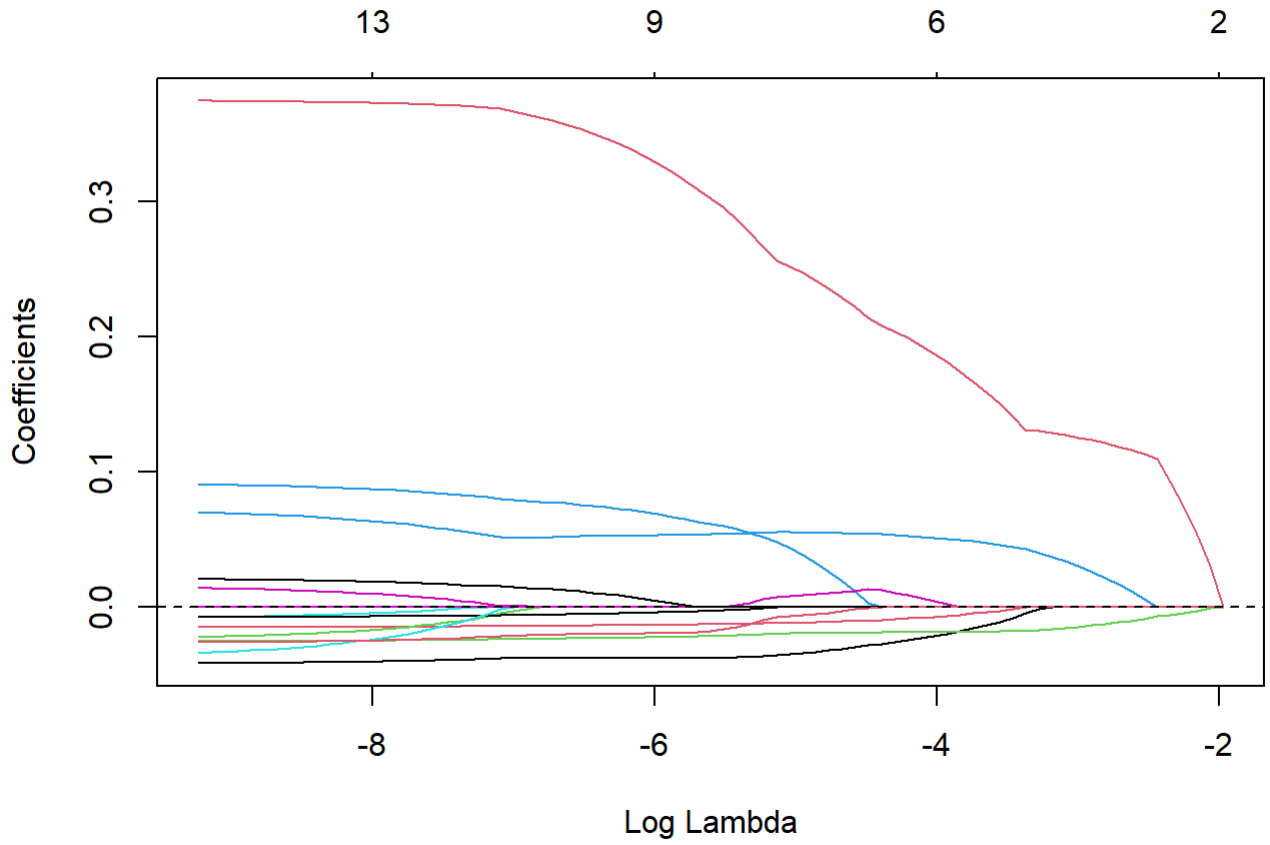
Make new data frame of predictors for LASSO, including region dummy variables.

```
lasso_predictors <- income_trans[, -c(11:12)]
lasso_predictors$region_NE <- ifelse(region == "NE", yes = 1, no = 0)
lasso_predictors$region_NC <- ifelse(region == "NC", yes = 1, no = 0)
lasso_predictors$region_S <- ifelse(region == "S", yes = 1, no = 0)
lasso_predictors$region_W <- ifelse(region == "W", yes = 1, no = 0)
sum(lasso_predictors[, 12:15])  #check that dummy values add up to 440
[1] 440
```

Find model:

```
lasso_mod1 <- glmnet(as.matrix(lasso_predictors), income_trans[, 11], alpha = 1)
```

```
plot(lasso_mod1, xvar = "lambda")
abline(h = 0, lty = 2)
```



Shrinkage plot for number of predictors at values of log(<U+03BB>)

It's not obvious where to cut the plot to decide on a number of predictors.

I'll try the minimum λ found by cross-validation and the λ value one standard error greater than that. Compare the selected coefficients using these two cutoffs:

```
cv_lasso_mod1 <- cv.glmnet(as.matrix(lasso_predictors), income_trans[,
    11], alpha = 1)

c(lambda.1se = cv_lasso_mod1$lambda.1se, lambda.min = cv_lasso_mod1$lambda.min)
  lambda.1se   lambda.min
4.878723e-03 9.802492e-05
coef_table <- cbind(coef(cv_lasso_mod1, s = cv_lasso_mod1$lambda.1se),
    coef(cv_lasso_mod1, s = cv_lasso_mod1$lambda.min))
dimnames(coef_table)[[2]] <- c("lambda(minMSE+1se)", "lambda(minMSE)")
coef_table
16 x 2 sparse Matrix of class "dgCMatrix"
                              lambda(minMSE+1se) lambda(minMSE)
(Intercept)                          9.386421156    9.876704535
land.area                           -0.036470934   -0.041247401
pop.18_34                           -0.012072867   -0.014786534
pop.65_plus                          .             -0.021829328
doctors                              0.054929516    0.069962187
hosp.beds                            .             -0.007256323
crimes                               .              0.014269024
pct.hs.grad                         -0.001085158   -0.007048349
pct.bach.deg                         0.277071025    0.374148236
pct.below.pov                       -0.020158096   -0.025074999
pct.unemp                            0.055240432    0.090650159
pop.dens <- log(pop/land.area)       .             -0.033565248
region_NE                            0.003580189    .
region_NC                            .              0.020832585
region_S                            -0.011933083   -0.025939862
region_W                             .              .
```

The list of coefficients in the model with the best λ found by cross-validation is similar to those found using all subsets with no interactions. The only additional coefficient kept here is the population density predictor. It is interesting that the coefficients are all way smaller in magnitude, but at least they have the same signs.

Next we'll try LASSO regression with interactions included. It would not be time-efficient to check every possible interaction, so we'll try including the interactions in the final model found earlier, model_bic_1, and see if LASSO keeps those same terms.

Add interaction columns to the predictor data frame:

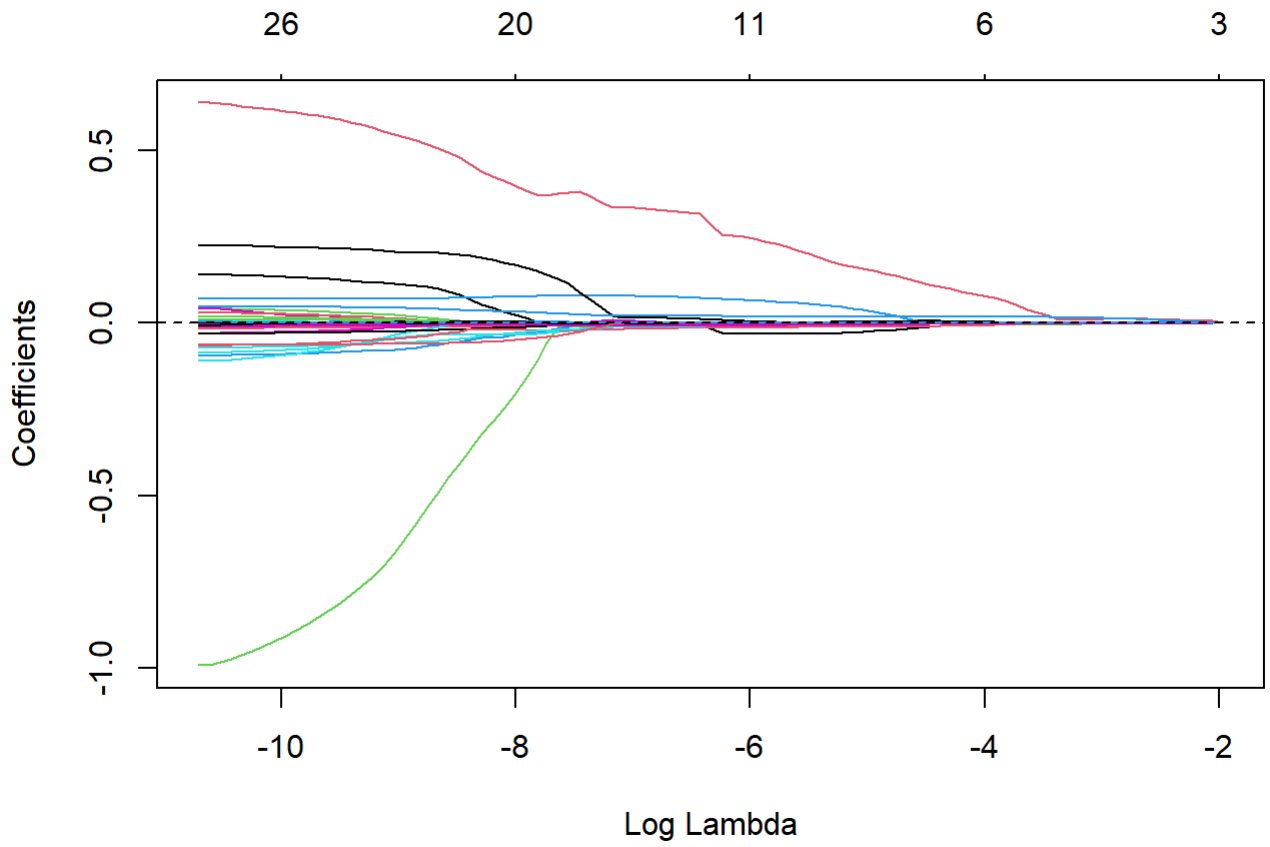
```
lasso_predictors_int <- lasso_predictors
lasso_predictors_int$"crimes:land.area" <- with(lasso_predictors, crimes *
    land.area)
lasso_predictors_int$"pct.bach.deg:land.area" <- with(lasso_predictors,
    pct.bach.deg * land.area)
lasso_predictors_int$"pct.below.pov:land.area" <- with(lasso_predictors,
    pct.below.pov * land.area)
lasso_predictors_int$"pop.18_34:pop.dens" <- with(lasso_predictors, pop.18_34 *
    `pop.dens <- log(pop/land.area)`)
lasso_predictors_int$"pop.65_plus:hosp.beds" <- with(lasso_predictors,
    pop.65_plus * hosp.beds)
lasso_predictors_int$"pop.65_plus:pct.hs.grad" <- with(lasso_predictors,
    pop.65_plus * pct.hs.grad)
lasso_predictors_int$"doctors:pct.bach.deg" <- with(lasso_predictors, doctors *
    pct.bach.deg)
lasso_predictors_int$"hosp.beds:pop.dens" <- with(lasso_predictors, hosp.beds *
    `pop.dens <- log(pop/land.area)`)
lasso_predictors_int$"pct.bach.deg:pct.below.pov" <- with(lasso_predictors,
    pct.bach.deg * pct.below.pov)
lasso_predictors_int$"pct.bach.deg:region_NE" <- with(lasso_predictors,
    pct.bach.deg * region_NE)
lasso_predictors_int$"pct.bach.deg:region_NC" <- with(lasso_predictors,
    pct.bach.deg * region_NC)
lasso_predictors_int$"pct.bach.deg:region_S" <- with(lasso_predictors,
    pct.bach.deg * region_S)
lasso_predictors_int$"pct.bach.deg:region_W" <- with(lasso_predictors,
    pct.bach.deg * region_W)
```

Find interaction model:

```
lasso_mod_int1 <- glmnet(as.matrix(lasso_predictors_int), income_trans[,
    11], alpha = 1)
```

```


plot(lasso_mod_int1, xvar = "lambda")
abline(h = 0, lty = 2)
```



Shrinkage plot for number of predictors at values of log(<U+03BB>)

Again the plot is a little too crowded to be helpful.

I'll try the minimum λ found by cross-validation and the λ value one standard error greater than that. Compare the selected coefficients using these two cutoffs:

```
cv_lasso_mod_int1 <- cv.glmnet(as.matrix(lasso_predictors_int), income_trans[,
    11], alpha = 1)

c(lambda.1se = cv_lasso_mod_int1$lambda.1se, lambda.min = cv_lasso_mod_int1$lambda.min)
  lambda.1se   lambda.min
3.312605e-04 2.230758e-05
coef_table_int <- cbind(coef(cv_lasso_mod_int1, s = cv_lasso_mod_int1$lambda.1se),
    coef(cv_lasso_mod_int1, s = cv_lasso_mod_int1$lambda.min))
dimnames(coef_table_int)[[2]] <- c("lambda(minMSE+1se)", "lambda(minMSE)")
coef_table_int
29 x 2 sparse Matrix of class "dgCMatrix"
                             lambda(minMSE+1se) lambda(minMSE)
(Intercept)                        10.2256508668   12.069203933
land.area                           0.0240853080    0.141107819
pop.18_34                          -0.0090573522   -0.014952362
pop.65_plus                        -0.2076720195   -0.990949637
doctors                            -0.0352001183   -0.093672335
hosp.beds                          -0.0008131937   -0.109713388
crimes                              .              -0.029753800
pct.hs.grad                        -0.0121580805   -0.030671987
pct.bach.deg                        0.3975955419    0.637599831
pct.below.pov                       .               0.041248148
pct.unemp                           0.0765623594    0.069280483
pop.dens <- log(pop/land.area)     -0.0334044168   -0.086285891
region_NE                           .               0.043009811
region_NC                           0.1685002847    0.225635891
region_S                           -0.0197462932   -0.064809343
region_W                            .              -0.009128751
crimes:land.area                    0.0037367947    0.009972066
pct.bach.deg:land.area             -0.0289429457   -0.071819773
pct.below.pov:land.area            -0.0014594885   -0.006100826
pop.18_34:pop.dens                 -0.0004773092    .
pop.65_plus:hosp.beds               .               0.030426003
pop.65_plus:pct.hs.grad             0.0024329002    0.009795877
doctors:pct.bach.deg                0.0333767190    0.048623709
hosp.beds:pop.dens                  .               0.003762037
pct.bach.deg:pct.below.pov         -0.0054921687   -0.009585310
pct.bach.deg:region_NE              0.0032086120   -0.006379927
pct.bach.deg:region_NC             -0.0489889933   -0.064037601
pct.bach.deg:region_S               .               0.017802476
pct.bach.deg:region_W               .               0.003728503
```

The model with the best λ found by cross-validation keeps every predictor available. The model with λ = 1 standard error above that keeps all except the hosp.beds:pop.dens interaction. Looking back at our original model found by stepwise selection and other criteria, this term probably would have been one of the next to be removed if we kept eliminating variables, since its coefficient is small and its interpretation is not obvious. We will try removing it from that model and see if it makes it worse.

```
model_bic_1_reduced <- update(model_bic_1, . ~ . - hosp.beds:`pop.dens <- log(pop/land.area)`)
summary(model_bic_1_reduced)

Call:
lm(formula = per.cap.income ~ pop.18_34 + pop.65_plus + doctors +
    hosp.beds + crimes + pct.hs.grad + pct.bach.deg + pct.below.pov +
    pct.unemp + `pop.dens <- log(pop/land.area)` + crimes:land.area +
    pct.bach.deg:land.area + pct.below.pov:land.area + pop.65_plus:hosp.beds +
    pop.65_plus:pct.hs.grad + hosp.beds:pct.bach.deg + pct.hs.grad:pct.below.pov +
    pct.bach.deg:pct.below.pov + pct.bach.deg:region, data = income_trans)

Residuals:
      Min        1Q    Median        3Q       Max
-0.288516 -0.046980 -0.001433  0.045396  0.254416

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     13.0958870  0.5500130  23.810  < 2e-16 ***
pop.18_34                       -0.0158753  0.0013442 -11.810  < 2e-16 ***
pop.65_plus                     -1.1360081  0.1957306  -5.804 1.28e-08 ***
doctors                          0.0543058  0.0134871   4.026 6.72e-05 ***
hosp.beds                       -0.2533694  0.0579113  -4.375 1.53e-05 ***
crimes                          -0.0824613  0.0231037  -3.569 0.000400 ***
pct.hs.grad                     -0.0366247  0.0052999  -6.910 1.81e-11 ***
pct.bach.deg                     0.5788608  0.1027001   5.636 3.20e-08 ***
pct.below.pov                    0.0487467  0.0122588   3.976 8.24e-05 ***
pct.unemp                        0.0801552  0.0164835   4.863 1.64e-06 ***
`pop.dens <- log(pop/land.area)` -0.0470517  0.0175277  -2.684 0.007555 **
crimes:land.area                 0.0172159  0.0034270   5.024 7.53e-07 ***
pct.bach.deg:land.area          -0.0515506  0.0092490  -5.574 4.48e-08 ***
pct.below.pov:land.area         -0.0059087  0.0011291  -5.233 2.64e-07 ***
pop.65_plus:hosp.beds            0.0398205  0.0135393   2.941 0.003452 **
pop.65_plus:pct.hs.grad          0.0108799  0.0020208   5.384 1.22e-07 ***
hosp.beds:pct.bach.deg           0.0505555  0.0128828   3.924 0.000102 ***
pct.hs.grad:pct.below.pov        0.0003558  0.0001122   3.170 0.001638 **
pct.bach.deg:pct.below.pov      -0.0209381  0.0037301  -5.613 3.62e-08 ***
pct.bach.deg:regionNE           -0.0057943  0.0041372  -1.401 0.162100
pct.bach.deg:regionS            -0.0157734  0.0039149  -4.029 6.65e-05 ***
pct.bach.deg:regionW            -0.0097614  0.0047368  -2.061 0.039943 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07496 on 418 degrees of freedom
Multiple R-squared:  0.8748,    Adjusted R-squared:  0.8685
F-statistic: 139.1 on 21 and 418 DF,  p-value: < 2.2e-16
BIC(model_bic_1)
[1] -908.462
BIC(model_bic_1_reduced)
[1] -913.8443
```

Judging by BIC, removing the interaction doesn't make the model any worse, so we will use model_bic_1_reduced as our final model.

```
formula(model_bic_1_reduced)
per.cap.income ~ pop.18_34 + pop.65_plus + doctors + hosp.beds +
    crimes + pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
    `pop.dens <- log(pop/land.area)` + crimes:land.area + pct.bach.deg:land.area +
    pct.below.pov:land.area + pop.65_plus:hosp.beds + pop.65_plus:pct.hs.grad +
    hosp.beds:pct.bach.deg + pct.hs.grad:pct.below.pov + pct.bach.deg:pct.below.pov +
    pct.bach.deg:region
```

Since this model was extensively modified after the first time we did an F-test against the no-interaction model, we will repeat that test to be sure additional terms are still helpful

```
anova(allsubs_mod1_final, model_bic_1_reduced)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	429	1.462014e+09	*NA*	*NA*	*NA*	*NA*
2	418	2.348570e+00	11	1462014016	23655474197	0

2 rows

Yes, interactions definitely still improve model.

Check state variable.

```
model_bic_1_reduced_state <- update(model_bic_1_reduced, . ~ . + state)
summary(model_bic_1_reduced_state)

Call:
lm(formula = per.cap.income ~ pop.18_34 + pop.65_plus + doctors +
    hosp.beds + crimes + pct.hs.grad + pct.bach.deg + pct.below.pov +
    pct.unemp + `pop.dens <- log(pop/land.area)` + state + crimes:land.area +
    pct.bach.deg:land.area + pct.below.pov:land.area + pop.65_plus:hosp.beds +
    pop.65_plus:pct.hs.grad + hosp.beds:pct.bach.deg + pct.hs.grad:pct.below.pov +
    pct.bach.deg:pct.below.pov + pct.bach.deg:region, data = income_trans)

Residuals:
      Min        1Q    Median        3Q       Max
-0.159784 -0.036323 -0.001063  0.033585  0.267836

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     12.5585775  0.5319937  23.607  < 2e-16 ***
pop.18_34                       -0.0149610  0.0013225 -11.313  < 2e-16 ***
pop.65_plus                     -0.8204872  0.1890940  -4.339 1.85e-05 ***
doctors                          0.0576344  0.0130762   4.408 1.37e-05 ***
hosp.beds                       -0.3198945  0.0536998  -5.957 5.97e-09 ***
crimes                          -0.0582705  0.0220786  -2.639 0.008660 **
pct.hs.grad                     -0.0249232  0.0053846  -4.629 5.10e-06 ***
pct.bach.deg                     0.2958131  0.1008434   2.933 0.003561 **
pct.below.pov                    0.0333586  0.0117308   2.844 0.004707 **
pct.unemp                        0.0302700  0.0220352   1.374 0.170360
`pop.dens <- log(pop/land.area)` -0.0429763  0.0172400  -2.493 0.013108 *
stateAR                         -0.0717269  0.0522307  -1.373 0.170497
stateAZ                         -0.0701355  0.1019876  -0.688 0.492080
stateCA                          0.0846229  0.1029001   0.822 0.411389
stateCO                          0.0469433  0.1063137   0.442 0.659069
stateCT                          0.2238444  0.0967387   2.314 0.021219 *
stateDC                         -0.0012413  0.0737181  -0.017 0.986575
stateDE                          0.1278245  0.1005480   1.271 0.204426
stateFL                         -0.0324425  0.0300542  -1.079 0.281081
stateGA                          0.0448196  0.0340597   1.316 0.189016
stateHI                          0.0571224  0.1047290   0.545 0.585785
stateID                          0.0271177  0.1207031   0.225 0.822364
stateIL                          0.3370519  0.0844020   3.993 7.85e-05 ***
stateIN                          0.2828936  0.0832485   3.398 0.000752 ***
stateKS                          0.2700808  0.0896963   3.011 0.002782 **
stateKY                         -0.0125570  0.0454497  -0.276 0.782484
stateLA                          0.0405662  0.0340453   1.192 0.234205
stateMA                          0.1786706  0.0974171   1.834 0.067443 .
stateMD                          0.0344394  0.0352973   0.976 0.329852
stateME                          0.1502111  0.0938217   1.601 0.110222
stateMI                          0.3349506  0.0830085   4.035 6.63e-05 ***
stateMN                          0.2819956  0.0890010   3.168 0.001660 **
stateMO                          0.3004945  0.0857598   3.504 0.000514 ***
stateMS                         -0.0601019  0.0454034  -1.324 0.186407
stateMT                          0.0671976  0.1195326   0.562 0.574340
stateNC                         -0.0289918  0.0302162  -0.959 0.337942
```

```
stateND                          0.3057926  0.1075048   2.844 0.004695 **
stateNE                          0.2644052  0.0949532   2.785 0.005634 **
stateNH                          0.1836131  0.0999980   1.836 0.067133 .
stateNJ                          0.2583325  0.0941256   2.745 0.006354 **
stateNM                         -0.0200531  0.1171666  -0.171 0.864199
stateNV                          0.2005017  0.1080014   1.856 0.064179 .
stateNY                          0.1606461  0.0917473   1.751 0.080779 .
stateOH                          0.2852448  0.0809395   3.524 0.000478 ***
stateOK                         -0.0643476  0.0413550  -1.556 0.120565
stateOR                         -0.0399292  0.1042857  -0.383 0.702026
statePA                          0.1580455  0.0875513   1.805 0.071858 .
stateRI                          0.0855589  0.1006041   0.850 0.395622
stateSC                         -0.0451459  0.0319177  -1.414 0.158070
stateSD                          0.3020037  0.1055114   2.862 0.004445 **
stateTN                         -0.0364300  0.0337671  -1.079 0.281350
stateTX                          0.0034236  0.0281756   0.122 0.903354
stateUT                         -0.1921065  0.1071642  -1.793 0.073846 .
stateVA                          0.0061035  0.0383594   0.159 0.873667
stateVT                          0.1371945  0.1201580   1.142 0.254280
stateWA                          0.0044716  0.1023333   0.044 0.965170
stateWI                          0.2985246  0.0828838   3.602 0.000359 ***
stateWV                         -0.0036079  0.0694324  -0.052 0.958586
crimes:land.area                 0.0135733  0.0032414   4.187 3.53e-05 ***
pct.bach.deg:land.area          -0.0432371  0.0088536  -4.884 1.55e-06 ***
pct.below.pov:land.area         -0.0039448  0.0010743  -3.672 0.000276 ***
pop.65_plus:hosp.beds            0.0440164  0.0121657   3.618 0.000338 ***
pop.65_plus:pct.hs.grad          0.0070743  0.0019805   3.572 0.000401 ***
hosp.beds:pct.bach.deg           0.0656261  0.0119543   5.490 7.48e-08 ***
pct.hs.grad:pct.below.pov        0.0002897  0.0001092   2.653 0.008317 **
pct.bach.deg:pct.below.pov      -0.0173658  0.0035371  -4.910 1.37e-06 ***
pct.bach.deg:regionNE            0.0401138  0.0303002   1.324 0.186358
pct.bach.deg:regionS             0.0892400  0.0263060   3.392 0.000768 ***
pct.bach.deg:regionW             0.0782516  0.0329593   2.374 0.018096 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06445 on 371 degrees of freedom
Multiple R-squared:  0.9179,    Adjusted R-squared:  0.9028
F-statistic: 60.98 on 68 and 371 DF,  p-value: < 2.2e-16
anova(model_bic_1_reduced, model_bic_1_reduced_state)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	418	2.348570	*NA*	*NA*	*NA*	*NA*
2	371	1.540833	47	0.8077365	4.137996	4.267607e-15

2 rows

F = 3.8993. P-value is significant, but this many indicators is not practical.