What Determines Per-Capita Income?

Stefano Molina *

October 18, 2021

Abstract

We try to analyze the relation between demographic variables and per-capita income for a select sample of counties. The data is obtained from Kutner et al. 2005 and has demographic data for 440 counties across 48 states. We perform some linear model variable selection methods to define the set of variables that bests fits the data. The resulting method is a good fit for predicting per-capita income, but may be missing enough data for a robust prediction.

Keywords: linear regression, variable selection methods, LASSO

*gmolinam@andrew.cmu.edu

1 Introduction

Using data from Kutner et al. 2005, we want to analyze the influence that demographic variables have on the per-capita income of counties. We make some sense of the relations between all the variables, test a theory that crimes and region make a good fit for predicting per-capita income, and look for a model to predict per-capita income based on the rest of the demographic variables.

The questions we want to answer are:

- Which variables seem to be related?
- Can crime or crime rate and region be a good set of predictors for per-capita income?
- How does a good fitting model for per-capita income looks based on a combination of the variables from the data?
- Does having a small set of counties from the total number of counties in the US matter for the model?

2 Data

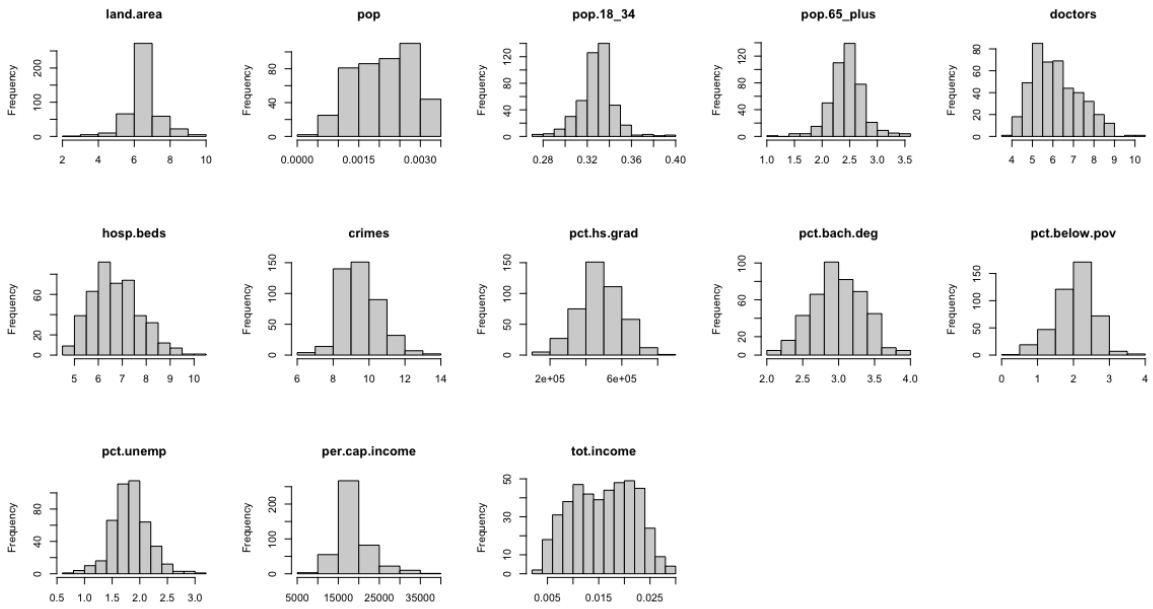
The data was obtained from Kutner et al. 2005, it contains county demographic information(CDI) for 440 counties accross the country for 1990-1992. The data includes geographic information as well as numerical variables related to the population's characteristics. Some histograms are shown in Figure 1 to make sense of the distributions of each numerical variable and determine for the further sections if transformations are needed for them.

The variables and their definitions, according to Kutner et al. 2005 are as follow:

- id - Identification number, ranging from 1 to 440
- county - County name
- state - State name
- land.area - Land area (square miles)
- pop.18.34 - Percent of CDI aged 18 to 34
- pop.65.plus - Percent of CDI aged 65 or older
- doctors - Number of professionally active nonfederal physicians during 1990
- hosp.beds - Total number of beds, cribs, and bassinets during 1990
- crimes - Total number of serious crimes in 1990, including murder, rape, rob- bery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
- pct.hs.grad - Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
- pct.bac.deg - Percent of adult population (persons 25 years old or older) with bachelor's degree
- pct.below.pov - Percent of 1990 CDI population with income below poverty level
- pct.unemp - Percent of 1990 CDI population that is unemployed
- per.cap.income - Per-capita income of 1990 CDI population (in dollars)
- tot.income - Total personal income of 1990 CDI population (in millions of dollars)

- region - Geographic region classification used by the US Bureau of the Cen- sus, 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)

Figure 1: Distributions of the numeric variables



Variable	Unique Values	NA Values
county	373	0.00
state	48	0.00
region	4	0.00
county/state	440	0.00

Table 1: Number of unique text variables and NA values

Variable	Min	Mean	Median	Max	NAs
crimes	563.00	27111.62	11820.50	688936.00	0
doctors	39.00	988.00	401.00	23677.00	0
hosp.beds	92.00	1458.63	755.00	27700.00	0
land.area	15.00	1041.41	656.50	20062.00	0
pct.bach.deg	8.10	21.08	19.70	52.30	0
pct.below.pov	1.40	8.72	7.90	36.30	0
pct.hs.grad	46.60	77.56	77.70	92.90	0
pct.unemp	2.20	6.60	6.20	21.30	0
per.cap.income	8899.00	18561.48	17759.00	37541.00	0
pop	100043.00	393010.92	217280.50	8863164.00	0
pop.18_34	16.40	28.57	28.10	49.70	0
pop.65_plus	3.00	12.17	11.75	33.80	0
tot.income	1141.00	7869.27	3857.00	184230.00	0

Table 2: Summary for the numeric variables

3 Methods

To answer the research questions, multiple statistical methods were used and tested.

The relation between the variables was analyzed with a correlation matrix between all of the numeric variables. As a 13×13 matrix may take too long to analyze and find relations, a correlation matrix was also used. This plot shows a divergent color scale for the values of the matrix, which makes it easy to find which variables are highly correlated and in which direction.

To test the question of whether crimes (or crime rate) and region can explain per-capita income, I used linear regression with and without interaction for both variables. The same process was done for crime rate and region and their results were compared.

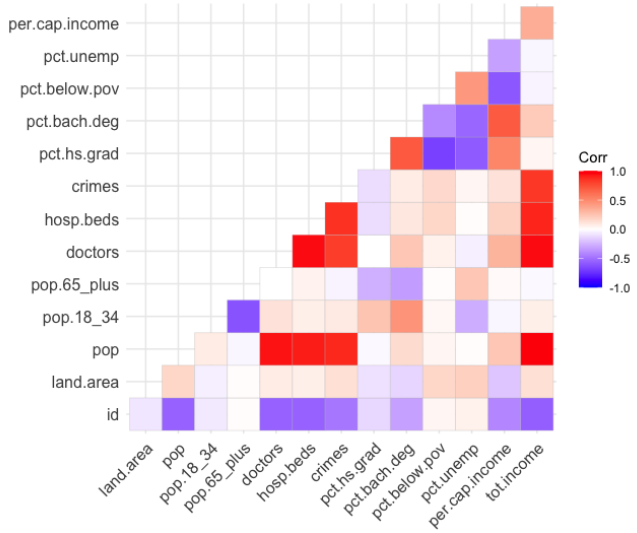
The selection of the model that best predicts per-capita income according to the needs of the study used multiple statistical analyses and model selection methods. The first thing was using the plots from Figure 1 as a reference to determine that some variables need a transformation to work correctly in the linear model. Also, taking into account the correlations plot, I did some variable selection based on the Variance Inflation Factor (VIF) to help the regression work more accurately. As the remaining variables were still not easy to interpret as a whole and the possibility that some of them were not relevant for the model, I used model selection procedures that help to choose the variables that make the best fit for the model. The procedures I used were: stepwise selection, testing all subsets, and LASSO regression. Each procedure was run independently and their results were compared to analyzed their similarities and then choose the model that made the most sense based on them.

4 Results

First, to see how the variables relate between each other, I used a correlation matrix to make a correlation plot. Figure 2 shows the relations between each combination of variables in a two-color scale in a way that the intensity and color of the cell will tell the sign and magnitude of the correlation.

We can see that some variables are highly correlated to population and total income.

Figure 2: Correlation plot of the variables



These variables are: crimes, hospital beds, and doctors. Other set of variables that have high correlations are the percentage of high school graduates, bachelors degree holders, percentage below poverty levels, and percentage of unemployment. These values could bring collinearity problems to the linear models that use these variables and should be analyzed to choose if any of them should be omitted.

To address the question whether crime(or crime rate) is related to per-capita income and the relation is different across regions, I used a linear model with crime with and without interaction with region. The model with interactions doesn't seem to add new information as their p-values are not statistically significant and the coefficients for the other variables doesn't change, as shown in Table 1. I would suggest to keep the model without interactions.

I also calculated the model with crime rate with and without interaction with region. First, it is important to notice that for both models, the crime rate variable is not statistically significant. Just as in the first model, adding interactions does not add relevant information to the model since all of the interactions are not statistically significant.

	Without interactions	With interactions
crimes	0.009**	0.014
	(0.003)	(0.008)
region: NC	18106.910***	18004.776***
	(378.438)	(409.242)
region: NE	20392.947***	20578.242***
	(387.980)	(401.869)
region: S	17246.353***	16948.446***
	(325.170)	(383.090)
region: W	17964.083***	17948.240***
	(458.849)	(488.476)
crimes × region: NE/NC		-0.013
		(0.010)
crimes × region: S/NC		0.006
		(0.011)
crimes × region: W/NC		-0.004
		(0.009)
R-squared	0.959	0.959
N	440	440

Significance: $***: p < 0.001$; $**: p < 0.01$;
$*: p < 0.05$

Table 3: Influence of crimes and regions in per-capita income

I checked if the numerical variables, except ID, needed transformations. I decided that

if a suggested power transformation was below 1/3, a log transformation would be used. Most of the variables needed some kind of transformation according to the tests performed, as shown in Appendix. Also, I omitted the variable *county* since it was of no use because the *id* variable was already a unique identifier for each observation.

After that, I calculated a linear regression including all the variables from the dataset and tried to use the Variance Inflation Factor to try to see if there was collinearity between some of the variables. I found that apparently, **state** and **region** were colinear and had to decide which to use so the VIF function would work. This made sense since the summary for the original regression omitted all the region categories. I tried omitting one variable at a time and found that omitting region has a higher R^2, but looking at the output of the regression, it would be hard to interpret all the coefficients for states since not all them were statistically significant. I decided to keep region.

Having addressed the first problem due to collinearity, the VIF function shows that population, crimes, and tot.income have a $GVIV^{\frac{1}{2df}}$ above 5, which means they should be removed. This should probably raise some flags for any social scientist: why would anyone omit the two variables that are directly related to the response variable? It is possible to argue that these two variables are already the ones that generate the response variable, and using only them could give an almost perfect fit. This is a valid point that should be addressed reminding that this would lead to overfitting and the model would not be useful for further analysis of prediction. I chose to omit these variables to avoid potential overfitting.

Finally, I ran the regressions for all possible models: all the variables, stepwise selection,

all subsets and LASSO. A comparison of the coefficients for the models is available in the Appendix. A first impression is that the three methods drop the hospital beds variable. The all subsets model also drops the population 65 or older and the LASSO model the percentage of high school graduates and region. Since the all subsets model is using one of the regions, I assume all of them should be considered. I would consider discarding LASSO for the sake of simplicity. It is now important to consider comparing the all subsets and stepwise selection models. Using analysis of variance to compare the models, the all subsets models is favored and ultimately selected. The summary statistics for this model are shown in Table 3 and its diagnostics plots in the Appendix.

One of our questions was whether having just one small sample of the 3000 counties of the US could be a problem for the model. Considering that according to the 1990 US Census (Bureau 2000), there were almost 250 million people living in the US at that time, which means roughly a 70% of the population is represented in the dataset. Now, looking at the dataset, the minimum value for the counties' population is 100,000. With this information, we can calculate the average population of the remaining counties, which will be close to 30,000 people. Considering that the average county for the data has a population of almost 400,000, the data may not give a good model for low populated counties and the predictions would not be expected to be accurate. On the other hand, the missing states are Iowa, Arkansas and Wyoming, which are low populated states.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0578687923	0.002	28.28	0.00
land.area	0.0004356824	0.000	6.09	0.000
pop.18_34	-0.0392049187	0.004	-9.46	0.000
doctors	-0.0006614169	0.000	-10.10	0.000
pct.hs.grad	0.0000000043	0.000	4.62	0.000
pct.bach.deg	-0.0037636758	0.000	-11.21	0.000
pct.below.pov	0.0029677158	0.000	18.19	0.000
pct.unemp	-0.0006668757	0.000	-3.02	0.002
crime_rate	-0.0041960885	0.001	-3.19	0.001
regionNE	0.0002694790	0.000	1.56	0.118
regionS	0.0007652065	0.000	4.83	0.000
regionW	0.0001450657	0.000	0.77	0.440

Table 4: Coefficient information for the selected model

5 Discussion

The data set contains two principal groups of variables: those related to population and those related to labor and education. Because of this evident relations, it made sense that there would exist some medium to high correlation between some of them. The first group makes sense because they usually grow as population increases either because more populations means more need for hospital beds and doctor or because higher concentrations of people increases the possibility of crimes. The second group makes sense too because of the implications of education of labor and education in the outcome of a population's income.

The high correlations for these variables did create some problems of multicollinearity for the regressions and must be taken into account for future works.

The relation between crimes and total income was evident from the correlations plot and it would be expected that a linear model between them and regions will have a good fit.

Looking at both models, it is possible to determine that the model including number of crimes is a more appropriate model since all its variables are statistically significant. This model shows that the number of crimes and the region are good predictors for per-capita income accounting for almost 96% of its variability. The difference between the two models is that for the second, crimes is divided by total population which may be having some influence on the dynamic between the variables as per-capita income is already divided by total population. The diagnostic plots for the model including crimes are shown in the Appendix. They may have some issues with some observations as the Q-Q plot has some skewed values in the right tail and also some high influence points that have high leverage and are outliers for the standardized residuals. These two points are the Los Angeles county in California and Kings county in New York, which are not surprising to have some extreme values for both crime rate and per-capita income.

The model for predicting per-capita income needed some use of transformations to work correctly and some variable selection based on variance inflation factors but the final results look to be fitting the data correctly. From the comparison between the models on the Appendix, it can be seen that the three models are almost the same with just one variable differing between them. Ultimately, the LASSO model was discarded for simplicity and because it was omitting region. The analysis of variance showed that the all subsets model

was a better fit than the stepwise selection, and from its summary it is possible to see that all of the selected variables are significantly different from 0 except two of the regions. From the diagnostic plots, there appear to be some observations that do not follow the normal distribution in the right side of the Q-Q plot, but besides that the model looks like a good fit: the residuals vs fitted plot doesn't have a distinguishable pattern, the scale-location plot looks like the variance is constant, and there are no high influence points on the data.

Considering the size of the sample of counties that are in the data, it could be worth noting that the model may not work for predicting counties with low density of population as there are no counties with these characteristics in the data. The missing states should not be a problem for the model because they are not part of it.

For future research, it would be useful to find a sample of small counties in order to make the model better and suitable for any kind of prediction.

References

Bureau, U.S. Census. 2000. *Population Change and Distribution 1990 to 2000.* `https://www.census.gov/prod/2001pubs/c2kbr01-2.pdf`.

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

A results

	Without interactions	With interactions
crime rate	5773.202	4379.070
	(7520.413)	(15893.507)
region: NC	18006.045***	18077.294***
	(537.039)	(895.208)
region: NE	20360.741***	20406.331***
	(493.620)	(641.617)
region: S	17078.598***	17066.941***
	(618.848)	(975.221)
region: W	17971.122***	17407.303***
	(637.921)	(1770.432)
crime$_rate$× region: NE/NC		288.387
		(20184.661)
crime$_rate$× region: S/NC		1558.919
		(20556.112)
crime$_rate$× region: W/NC		10655.542
		(32322.408)
R-squared	0.958	0.958
N	440	440

Significance: $***: p < 0.001$; $**: p < 0.01$;
$*: p < 0.05$

Table 5: Influence of crimes and regions in per-capita income

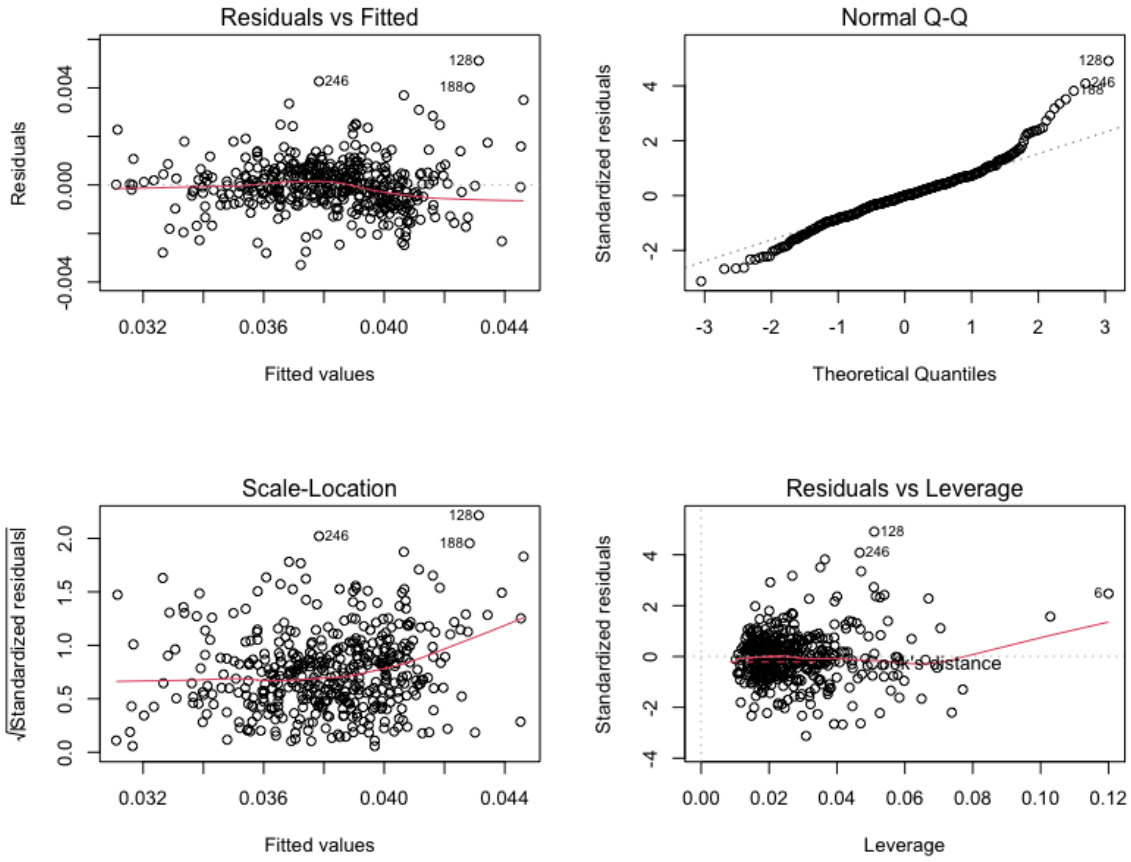
	Variable	powerTransform
1	land.area	0.00
2	pop	-0.58
3	pop.18_34	-0.39
4	pop.65_plus	-0.01
5	doctors	-0.22
6	hosp.beds	-0.15
7	crimes	-0.13
8	pct.hs.grad	3.07
9	pct.bach.deg	-0.03
10	pct.below.pov	0.18
11	pct.unemp	-0.11
12	per.cap.income	-0.37
13	tot.income	-0.44
14	crime_rate	0.38
15	per.cap.income3	1.11

Table 6: Suggested power transformations

	Variable	Regression	StepAIC	All.subsets	LASSO
1	(Intercept)	0.0566601482	0.0564247421	0.0576659074	0.0500578839
2	land.area	0.0004362029	0.0004258883	0.0004453434	0.0003436730
3	pop.18_34	-0.0300780792	-0.0301527936	-0.0390517356	-0.0188090642
4	pop.65_plus	-0.0006339199	-0.0006800881		-0.0006558449
5	doctors	-0.0004573955	-0.0006240007	-0.0006523899	-0.0007807830
6	hosp.beds	-0.0001972827			
7	pct.hs.grad	0.0000000040	0.0000000040	0.0000000038	
8	pct.bach.deg	-0.0038462755	-0.0037521099	-0.0036134548	-0.0021363354
9	pct.below.pov	0.0031025404	0.0030443762	0.0029215797	0.0025800139
10	pct.unemp	-0.0007144710	-0.0006974828	-0.0005820122	-0.0002134638
11	regionNE	0.0003415626	0.0003564918		
12	regionS	0.0006782244	0.0007183960	0.0006472652	
13	regionW	0.0000080228	0.0001036392		
14	crime_rate	-0.0040852207	-0.0042785823	-0.0045828114	-0.0001170040

Table 7: Comparison between variable selection models

Figure 3: Diagnostic plots for the all subsets model



B code

```
library(dplyr)
library(leaps)
library(car)
library(MASS)
library(glmnet)
library(purrr)
library(tidyverse)
library(ggcorrplot)

cdi <- read.table("/Users/Stefano_1/Documents/CMU/Applied_Linear_Models
/Projects/cdi.dat")

chars <- apply(cdi[,unlist(lapply(cdi, is.character))],
               2, function(x)length(unique(x)))
chars_na <- apply(cdi[,unlist(lapply(cdi, is.character))],
               2, function(x)sum(is.na(x)))

county.states <- data.frame(variable = "county/state",
                            unique.values = length(unique(paste0(cdi$county,
                            "-", cdi$state))),
                            na.values = 0)
```

```
chars_df <- data.frame(variable = names(chars),
                       unique.values = unname(chars),
                       na.values = unname(chars_na)) %>%
  bind_rows(county.states)


sum_nas <- function(x){sum(is.na(x))}

num_vars <- cdi %>%
  select_if(negate(is.character)) %>%
  pivot_longer(!id) %>%
  dplyr::select(-id) %>%
  group_by(name) %>%
  summarise_all(list(min = min,
                     mean = mean,
                     median = median,
                     max = max,
                     NAs = sum_nas))

par(mfrow = c(3,5))

for(i in names(cdi)[2:17]){
  if(is.character(cdi[[i]])){next}
  hist(cdi[[i]], main = i, xlab = "")
}
```

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

ggcorrplot(cor, type = "lower")

cdi$crime_rate <- cdi$crimes/cdi$pop

powerTransform(cdi$crimes)
powerTransform(cdi$per.cap.income)
powerTransform(cdi$crime_rate)

cdi$per.cap.income3 <- cdi$per.cap.income^(-1/3)

reg1 <- lm(per.cap.income ~ crimes + region - 1, data = cdi)

summary(reg1)

reg1.2 <- lm(per.cap.income ~ crimes + region + crimes:region - 1,
             data = cdi)

summary(reg1.2)

par(mfrow = c(2,2))
plot(reg1, which = 1)
plot(reg1, which = 2)
```

```
plot(reg1, which = 3)
plot(reg1, which = 5)

reg2 <- lm(per.cap.income ~ crime_rate + region - 1,
           data = cdi)

summary(reg2)

reg2.2 <- lm(per.cap.income ~ crime_rate + region + crime_rate:region - 1,
data = cdi)

summary(reg2.2)

ap_cdi <- apply(cdi[,-c(1,2,3,17)], 2, powerTransform)

ap_cdi <- lapply(ap_cdi, function(x){x$lambda}) %>% unlist()

names(ap_cdi) <- substr(names(ap_cdi), 1, (nchar(names(ap_cdi))-10))

ap_cdi <- data.frame(Variable = names(ap_cdi),
                     powerTransform = unname(ap_cdi))
ap_cdi <-ap_cdi[,-15]

cdi$land.area <- log(cdi$land.area)
cdi$pop       <- cdi$pop^(-1/2)
```

```
cdi$pop.18_34 <- cdi$pop.18_34^(-1/3)
cdi$pop.65_plus <- log(cdi$pop.65_plus)
cdi$doctors    <- log(cdi$doctors)
cdi$hosp.beds <- log(cdi$hosp.beds)
cdi$crimes     <- log(cdi$crimes)
cdi$pct.hs.grad <- cdi$pct.hs.grad^3
cdi$pct.bach.deg <- log(cdi$pct.bach.deg)
cdi$pct.below.pov <- log(cdi$pct.below.pov)
cdi$pct.unemp <- log(cdi$pct.unemp)
cdi$tot.income <- cdi$tot.income^(-1/2)
cdi$crime_rate <- cdi$crime_rate^(1/3)

cdi_final <- cdi[, -c(1,2,3,15)]

reg3 <- lm(per.cap.income3~., data = cdi_final)

alias(reg3)
vif(reg3) #looks like we should remove pop and total income

cdi_final <- cdi_final %>% dplyr::select(-pop, - tot.income, -crimes)
# cdi_final <- cdi_final %>% dplyr::select(-pop, -tot.income)

reg3 <- lm(per.cap.income3~., data = cdi_final)

reg3_stepaic <- stepAIC(reg3, trace = FALSE)
```

```
coef(reg3_stepaic)

reg3_subsets <- regsubsets(per.cap.income3~. ,
                  data =cdi_final,
                  really.big = T,
                  nvmax = 10)
reg3_subsetsum <- summary(reg3_subsets)
coef_all_subsets <- coef(reg3_subsets,
                  which.min(reg3_subsetsum$bic))

last <- ncol(cdi_final)
reg3_lasso <- cv.glmnet(data.matrix(cdi_final[,-last]),
                        cdi_final[,last],
                        alpha = 1)

lasso_coefs <- cbind(coef(reg3_lasso, s=reg3_lasso$lambda.min),
                     coef(reg3_lasso, s=reg3_lasso$lambda.1se))

coef_lasso <- as.matrix(coef(reg3_lasso, s=reg3_lasso$lambda.1se))
lasso_coef <- as.matrix(coef(reg3_lasso))[coef_lasso !=0]
coef_lasso <- rownames(coef_lasso)[coef_lasso !=0]
coef_lasso <- data.frame(Variable =coef_lasso,
                         LASSO = lasso_coef)

cdi_coefs <- data.frame(Variable = names(reg3$coefficients),
```

```
                                Regression = unname(coef(reg3))) %>%
  full_join(data.frame(Variable = names(coef(reg3_stepaic)),
                               StepAIC = unname(coef(reg3_stepaic)))) %>%
  full_join(data.frame(Variable = names(coef_all_subsets),
                               All.subsets = unname(coef_all_subsets))) %>%
  full_join(coef_lasso)

summary(reg3_stepaic)
aux <- names(coef_all_subsets)[-1]
aux <- aux[!startsWith(aux,"region")]
aux <- c(aux, "per.cap.income3", "region")
aux <- cdi_final[,aux]

reg3_all_subsets <- lm(per.cap.income3~., data=aux)
summary(reg3_all_subsets)

anova(reg3_stepaic, reg3_all_subsets)

par(mfrow = c(2,2))
plot(reg3_all_subsets, which = 1)
plot(reg3_all_subsets, which = 2)
plot(reg3_all_subsets, which = 3)
plot(reg3_all_subsets, which = 5)
```