What Determines Per-Capita Income? A Study on Demographic and Greographic Variable Selection

Stefano Molina
gmolinam@andrew.cmu.edu

October 30, 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 linear model variable selection methods to define the set of variables that bests fits the data. The best model was selected using the all subsets method for variable selection, and it results in 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

1 Introduction

Income is a great matter of interest for social sciences. It helps understand the quality of life for people, but can also help understand the implications for the area they live in. Estimating income based on demographic characteristics can help understand the context that drives its variations. We are interested in analyzing how these social and geographic characteristics have influence in per-capita income at the county level. Using per-capita income helps reduce variation as it represents the income of the average citizen for the county.

Using data from Kutner et al. (2005), we want to analyze the influence that demographic variables have on the per-capita income of counties. Also, we are interested in whether the region where a county is located has some influence in this outcome. 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), which contains county demographic information(CDI) for 440 counties across the country from 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. Also, Tables 1 and 2 show some characteristics of the string and numeric variables, respectively.

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

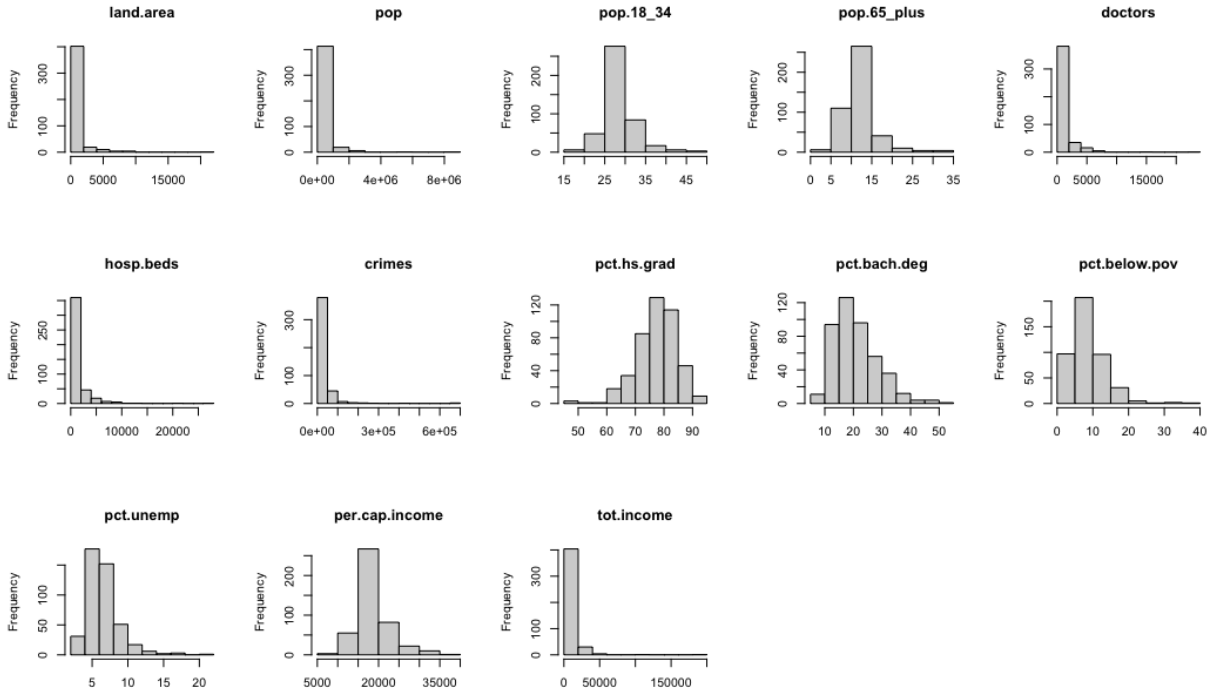
- id - Identification number, ranging from 1 to 440
- county - County name
- state - State name
- land.area - Land area (square miles)
- pop.18.34 - Percent of CDI aged 18 to 34
- pop.65.plus - Percent of CDI aged 65 or older
- doctors - Number of professionally active nonfederal physicians during 1990
- hosp.beds - Total number of beds, cribs, and bassinets during 1990
- crimes - Total number of serious crimes in 1990, including murder, rape, robbery, 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)

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

Figure 1: Distributions of the numeric variables



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

3.1 Relations between variables

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 heatmap 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.

3.2 The influence of crime on predicting per-capita income

To test the question of whether crimes (or crime rate, which is the number of crimes divided by the total population for each county) and region can explain per-capita income, a linear regression with and without interaction for both variables was used. The same process was done for crime rate and region and their results were compared.

3.3 Finding the best model

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 step 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, some variable selection was done based on the Variance Inflation Factor (VIF) to help the regression meet the no collinearity assumption needed for a regression to

work correctly. 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, some model selection procedures were used. These methods help to choose the variables that make the best fit for the model. The procedures that were used are: stepwise selection, testing all subsets, and LASSO regression. Stepwise selection consists on starting with the full (or empty) model and then removing (or adding) one variable at a time testing by a preselected criteria, usually the Akaike Selection Criterion (Sheather (2009)); testing all subsets, as the name suggests, finds the best model testing the regression results for all possible combinations of variables; and LASSO performs the regression penalizing for the size of the coefficients. Each procedure was run independently and their results were compared to analyze their similarities and then choose the model that made the most sense based on them.

3.4 Validity of the sample

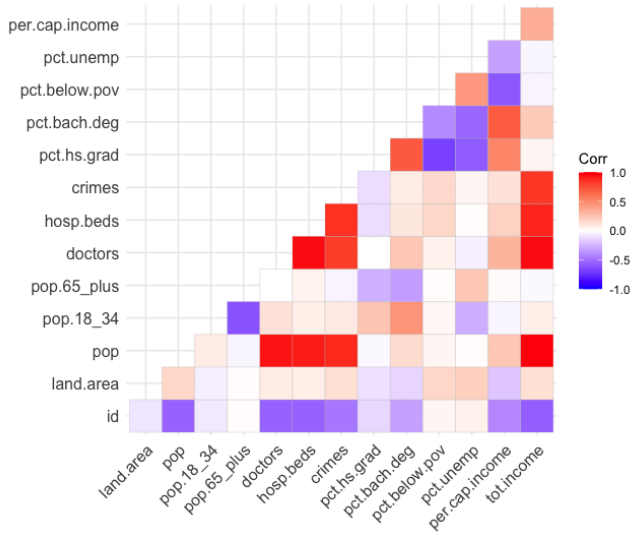
The last question is answered using mostly exploratory data analysis over the dataset and some lookup at the results of the 1990 US Census. It is developed in the results and addressed in the discussion.

4 Results

4.1 Relations between variables

First, to see how the variables relate to each other, a correlation matrix is used 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.

Figure 2: Correlation plot of the variables



We can see that some variables are highly correlated to population and total income. These variables are crimes, hospital beds, and doctors. Other sets of variables that have high correlations are the percentage of high school graduates, bachelor's 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.

4.2 The influence of crime on predicting per-capita income

To address the question of whether crime(or crime rate) is related to per-capita income and the relation is different across regions, linear models using crime with and without

interaction with region are used. 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 don't change, as shown in Table 1. It is suggested to keep the model without interactions. The selected model is:

$$per.capita.income \sim \beta_0 + \beta_1 crimes + \beta_2 region + \varepsilon$$

Also, the model with crime rate with and without interaction with region was calculated. 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. This model was the following:

$$per.capita.income \sim \beta_0 + \beta_1 crime\_rate + \beta_2 region + \varepsilon$$

The difference between the two models is that for the second, crimes are divided by the total population which may be having some influence on the dynamic between the variables as per-capita income is already divided by the total population. The diagnostic plots for the model including crimes are shown in the Appendix (page 28). 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.

	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

4.3 Finding the best model

To help the model perform adequately, all numeric variables were tested to check if they needed transformations. A transformation is usually needed if a variable is skewed and not meeting the normality assumption. If the suggested power transformation was lower than 1/3, a log transformation was used. Most of the variables needed some kind of transformation according to the tests, as shown in Table 6 (Appendix, page 20). Also, the *county* variable was omitted since its purpose was to identify each observation, just as the ID variable already did.

As there are 48 unique states in the sample, each had on average around 10 or 11 observations, which could be not enough to make good inferences and fitting. The state variables were omitted and the region variable was kept since it was already part of the question for the Section 4.2 models and the 110 average observations for each category should be enough for the model to make a good fit.

Other variables that showed collinearity problems were *population*, *crimes*, and *tot.income*. 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. The variables were omitted to avoid potential overfitting.

Finally, the regressions for all possible models were tried: all the variables, stepwise selection, all subsets, and LASSO (Sheather (2009)). A comparison of the coefficients for the

models is available in Table 7(Appendix, page 22). 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, it is assumed that all of them should be considered. As region is considered an important variable for this model and the LASSO model is discarding it, this variable selection is not considered. It is important to notice in Table 7 that using the variables selected by LASSO and adding the region variable would be a similar outcome to the stepwise selection model since the coefficient for *pct.hs.grad* is small. The next step is to compare the all subsets and stepwise selection models. Using analysis of variance to compare the models, the all subsets models is favored and selected. Considering some potential improvement, the region variable was interacted with every other numeric variable to test if any interaction produced coefficients that could still help the model. Three variables had significant interaction terms with region: pct.hs.grad^3, log(*pct.below.pov*), and log(*pct.unemp*). The interaction between region and pct.hs.grad^3 was finally discarded as the size of its coefficients were small compared to all the other coefficients. The new model was compared to the no-interactions model through an analysis of variance and it showed to have some improvement, as in Appendix (pages 35 and 36). The selected model is the following.

$$\begin{aligned}\log(\text{per.capita.income}) \sim \beta_0 + \beta_1 \log(\text{land.area}) + \beta_2 \text{pop.18\_34}^{-1/3} + \beta_3 \log(\text{doctors}) \\ + \beta_4 \text{pct.hs.grad}^3 + \beta_5 \log(\text{pct.bach.deg}) + \beta_6 \log(\text{pct.below.pov}) \\ + \beta_7 \log(\text{pct.unemp}) + \beta_8 \text{crime}\cdot\text{rate}^{1/3} + \beta_9 \text{region} \\ + \beta_{10} \log(\text{pct.below.pov}) * \text{region} + \beta_{11} \log(\text{pct.unemp}) * \text{region} + \varepsilon\end{aligned}$$

The summary statistics for this model are shown in Table 4 and its diagnostics plots in the Appendix(page 36).

To understand Table 4, it is worth remembering that the model tries to predict per-capita income, which as shown in Appendix (page 27) was transformed with the logarithm operation. From the results shown in Table 4, most results are easy to interpret, as the variables have log-transformations too. This is the case for *land.area, doctors, pct.bach.deg, pct.below.pov*, and *pct.unemp*, where a 1% increase for each variable, keeping everything else constant, will be translated to a 1% increase on per-capita income. For the rest of the variables, for which a power transformation was done, the interpretation is more complex. Given that per-capita income was transformed with the logarithm operation, one unit change of either $pop.18\_34^{-1/3}$, $pct.hs.grad^3$, or $crime\_rate^{1/3}$, will translate to a $\beta_i \times 100\%$ change in per-capita income. For the interaction terms, if the coefficient is significant, then the interaction coefficient is added to the corresponding variable coefficient according to the region. In the case of the region variable, the selection of any of the regions with significant coefficients will be added to the Intercept.

From the diagnostic plots for this model, shown in the Appendix(page 36), 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 don'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.

4.4 Validity of the sample

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 (U.S. 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)	7.94500	0.16034	49.55237	0.00000
log(land.area)	-0.03088	0.00555	-5.56059	0.00000
$pop.18\_34^{-1/3}$	3.38190	0.31819	10.62869	0.00000
log(doctors)	0.05060	0.00498	10.16779	0.00000
$\text{pct.hs.grad}^{3\dagger}$	-0.00000	0.00000	-5.15240	0.00000
log(pct.bach.deg)	0.32187	0.02634	12.21817	0.00000
log(pct.below.pov)	-0.20354	0.02004	-10.15691	0.00000
log(pct.unemp)	0.10491	0.02588	4.05436	0.00006
$\text{crime\_rate}^{1/3}$	0.29428	0.09907	2.97048	0.00314
regionNE	0.10468	0.07984	1.31118	0.19051
regionS	0.26921	0.06263	4.29879	0.00002
regionW	0.16637	0.06808	2.44371	0.01495
log(pct.below.pov):regionNE	-0.03068	0.02653	-1.15653	0.24812
log(pct.below.pov):regionS	-0.02285	0.02287	-0.99906	0.31834
log(pct.below.pov):regionW	-0.09979	0.03474	-2.87199	0.00428
log(pct.unemp):regionNE	-0.03745	0.04765	-0.78587	0.43239
log(pct.unemp):regionS	-0.15896	0.03572	-4.44970	0.00001
log(pct.unemp):regionW	0.01373	0.03859	0.35589	0.72210

† This variable had a coefficient sized 10^{-7}

Table 4: Coefficient information for the selected model

5 Discussion

The goal of this study was to analyze the influence of demographic and geographic variables in the per-capita income of a selected number of counties from the U.S. To do this, a variable selection regression model was used, which chose the following model:

$$\begin{aligned}\log(\text{per.capita.income}) \sim \beta_0 + \beta_1 \log(\text{land.area}) + \beta_2 \text{pop.18\_34}^{-1/3} + \beta_3 \log(\text{doctors}) \\ + \beta_4 \text{pct.hs.grad}^3 + \beta_5 \log(\text{pct.bach.deg}) + \beta_6 \log(\text{pct.below.pov}) \\ + \beta_7 \log(\text{pct.unemp}) + \beta_8 \text{crime.rate}^{1/3} + \beta_9 \text{region} \\ + \beta_{10} \log(\text{pct.below.pov}) * \text{region} + \beta_{11} \log(\text{pct.unemp}) * \text{region} + \varepsilon\end{aligned}$$

5.1 Relations between variables

The data set contains two principal groups of variables: those related to population (pop.18.34, pop.65.plus, doctors, hosp.beds, and crimes) and those related to labor and education (pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, and per.capita.income). Because of these relations, it made sense that there would exist some medium to high correlation between some of them. The first group, for which doctors; hospital beds; and crimes have high correlations, makes sense because they usually grow as the population increases either because more populations mean more need for hospital beds and doctors or because higher concentrations of people increase the possibility of crimes. The second group, for which percentage below poverty levels; percentage of bachelors degree holders; percentage of high school graduates; and percentage of unemployment have high correlations, 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.

5.2 The influence of crime on predicting per-capita income

The relation between crimes and per-capita 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. The selected model, which is

$$\text{per.capita.income} \sim \beta_0 + \beta_1 \text{crimes} + \beta_2 \text{region} + \varepsilon$$

that predicts per capita income based on crime rate and region with no interactions, shows that the number of crimes and the region are good predictors for per-capita income accounting for almost 96% of its variability. While it may seem like a good fit and no other variables may be needed to predict per-capita income, it would be interesting to first analyze the possibility that variation in crimes is not the cause of variation on per-capita income, but the opposite. If per-capita income is the cause for an increase in crimes, this model would become invalid, so it is worth analyzing if this is the correct approach for it. Considering additional variables, just as it is done in Sections 4.3 and 5.3 can help understand the influence that additional variables have when predicting per-capita income besides the number of crimes.

5.3 Finding the best model

While the final model was selected through a series of steps that helped to find the most adequate combination of explanatory variables, it is important to remind that the other models were almost as good and had similar coefficients for most of the shared variables, as seen on Table 7.

The model has a lot of potential for improvement, especially towards an easier to interpret result. It may be the case that power transformations help the model become a good fit but at the cost of not being able to easily explain the relations between the variables. For future works, it would be worth taking into account the interpretability needs before applying the suggested Box-Cox transformations suggested by the software. Also, it would be convenient to analyze further the differences between each models' selections and the reason for the variables being included or omitted since they may be important to the people that are interested in the model.

Finally, as suggested in Section 5.4, this model could become a better prediction for per-capita income if the data included more counties, especially those with low population density, to understand the true relation between the variables.

5.4 Validity of the sample

Considering the size of the sample of counties that are in the data (440 counties), 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 at the moment because they do not have highly populated cities, which could be the reason for them not being in the sample.

For future research, it would be useful to find a sample of small counties in order to make the model better and suitable for a more robust prediction, following the guidelines for county selection, it would be worth trying to replicate the small county sample given the original requirements.

References

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

Sheather, Simon. 2009. *A modern approach to regression with R.* Springer Science & Business Media.

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

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

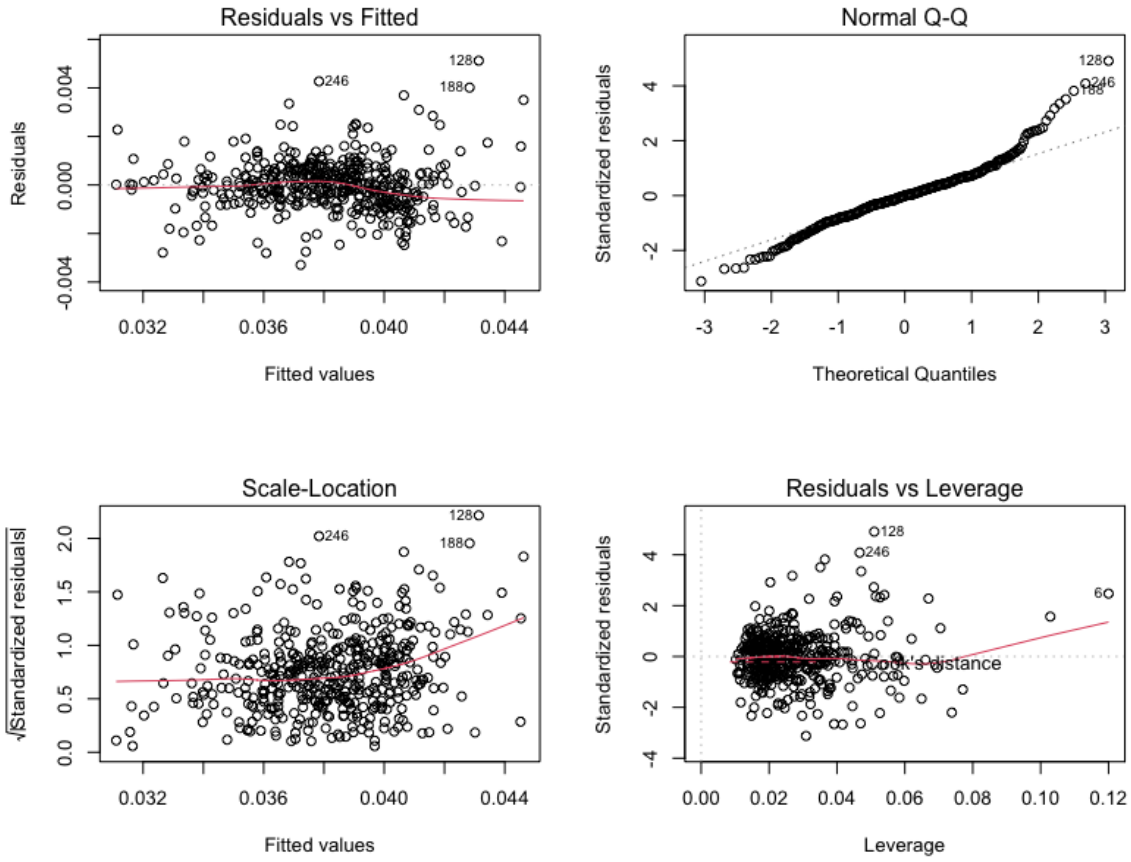
	Variable	powerTransform
1	land.area	0.00
2	pop	-0.58
3	pop.18_34	-0.39
4	pop.65_plus	-0.01
5	doctors	-0.22
6	hosp.beds	-0.15
7	crimes	-0.13
8	pct.hs.grad	3.07
9	pct.bach.deg	-0.03
10	pct.below.pov	0.18
11	pct.unemp	-0.11
12	per.cap.income	-0.37
13	tot.income	-0.44
14	crime_rate	0.38
15	per.cap.income3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

Table 6: Suggested power transformations

	Variable	Regression	StepAIC	All.subsets	LASSO
1	(Intercept)	8.2796428896	8.2920091903	8.1939523851	8.6391085734
2	land.area	-0.0350811881	-0.0345393452	-0.0351289003	-0.0299777874
3	pop.18_34	2.5296098723	2.5335347541	3.1569371942	1.8170547929
4	pop.65_plus	0.0444645424	0.0468898414		0.0503655355
5	doctors	0.0403910313	0.0491431016	0.0508781315	0.0604100936
6	hosp.beds	0.0103636165			
7	pct.hs.grad	-0.0000003493	-0.0000003476	-0.0000003308	
8	pct.bach.deg	0.3205310002	0.3155843119	0.3053851622	0.1983124214
9	pct.below.pov	-0.2434105672	-0.2403550928	-0.2309455770	-0.2071189339
10	pct.unemp	0.0638065575	0.0629141347	0.0541136732	0.0428964430
11	regionNE	-0.0262517681	-0.0270360269		
12	regionS	-0.0511571376	-0.0532674234	-0.0486149747	
13	regionW	0.0008576372	-0.0041652627		
14	crime_rate	0.2892592844	0.2994169182	0.3317425153	0.0563003953

Table 7: Comparison between variable selection models

Figure 3: Diagnostic plots for the all subsets model



Appendix B

```
cdi <- read.table("/Users/Stefano_1/Documents/CMU/Applied Linear Models/Projects/cdi.dat")
```

For the character variables, we will only display the number of distinct values.

```
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)
```

Table 1: Categorical variables and their total unique values

variable	unique.values	na.values
county	373	0
state	48	0
region	4	0
county/state	440	0

For the rest of the variables, we will show the minimum, mean, median and maximum of each column, and also the count of missing values (NAs). We will leave the id column out of the summary as it has 440 distinct values and its summary won't be useful.

```
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))
```

Table 2: Summary of numerical varaibles

name	min	mean	median	max	NAs
crimes	563.0	27111.62	11820.50	688936.0	0
doctors	39.0	988.00	401.00	23677.0	0
hosp.beds	92.0	1458.63	755.00	27700.0	0

name	min	mean	median	max	NAs
land.area	15.0	1041.41	656.50	20062.0	0
pct.bach.deg	8.1	21.08	19.70	52.3	0
pct.below.pov	1.4	8.72	7.90	36.3	0
pct.hs.grad	46.6	77.56	77.70	92.9	0
pct.unemp	2.2	6.60	6.20	21.3	0
per.cap.income	8899.0	18561.48	17759.00	37541.0	0
pop	100043.0	393010.92	217280.50	8863164.0	0
pop.18_34	16.4	28.57	28.10	49.7	0
pop.65_plus	3.0	12.17	11.75	33.8	0
tot.income	1141.0	7869.27	3857.00	184230.0	0

From the histograms, we can see that some of the variables need some kind of transformation in order to work correctly when a regression is performed with them on it. It is worth noting that the variables that are not percentages or divided by total population tend to be less skewed right than the variables that are just a single measurement. Also, since some of these variables are calculated dividing by population, it is expected that some of them will be correlated to population.

```
par(mfrow = c(3,5))

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

#savePlot("/Users/Stefano_1/Documents/CMU/Applied Linear Models/Projects/vars_hists",
#          type = "png")
```

From the correlations plot, we can see that indeed, some variables are highly correlated to population, but they are not the ones that we initially expected. These variables are: crimes, hospital beds, doctors, and total income, which make 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.

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

ggcorrplot(cor, type = "lower")
```

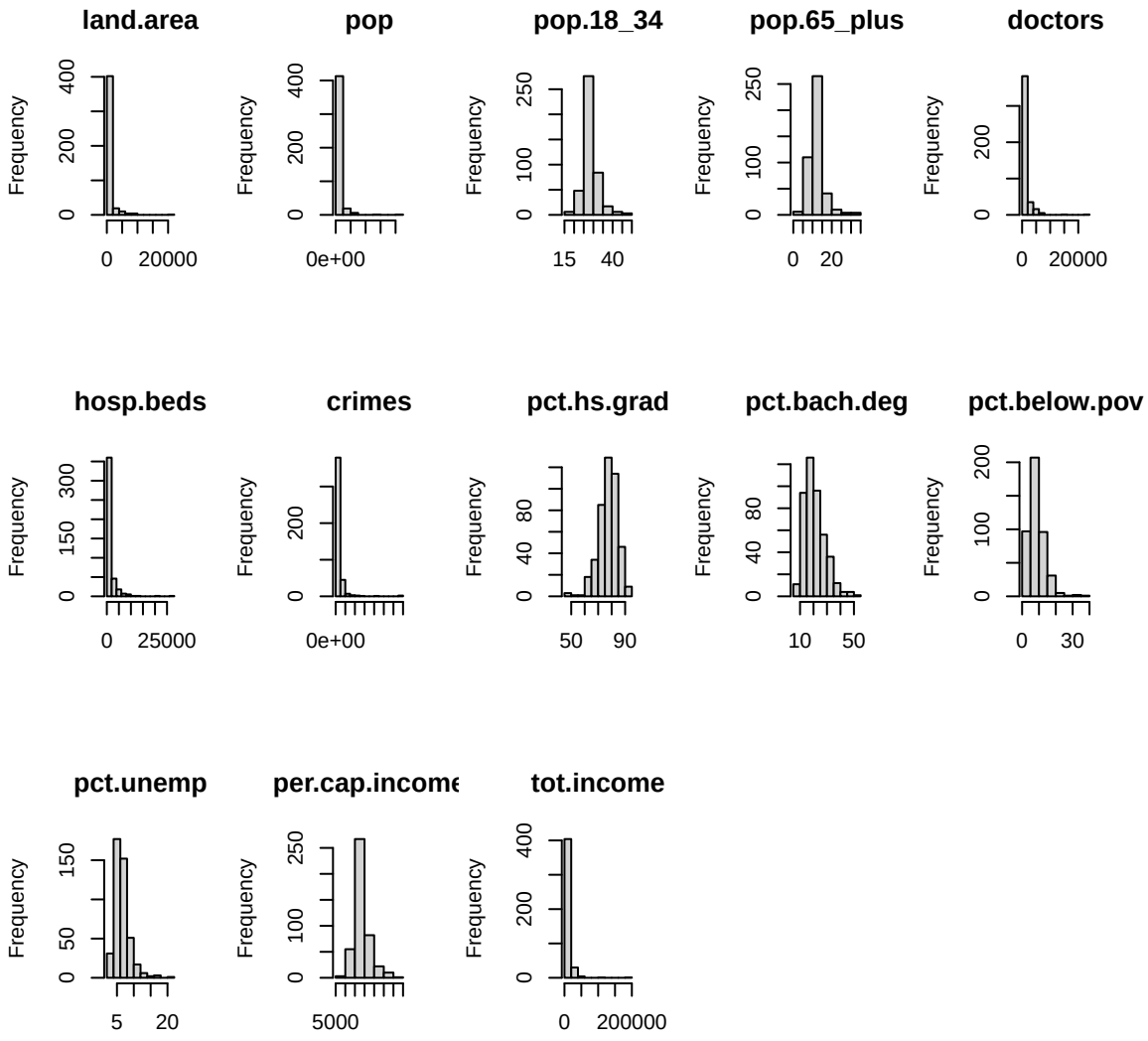
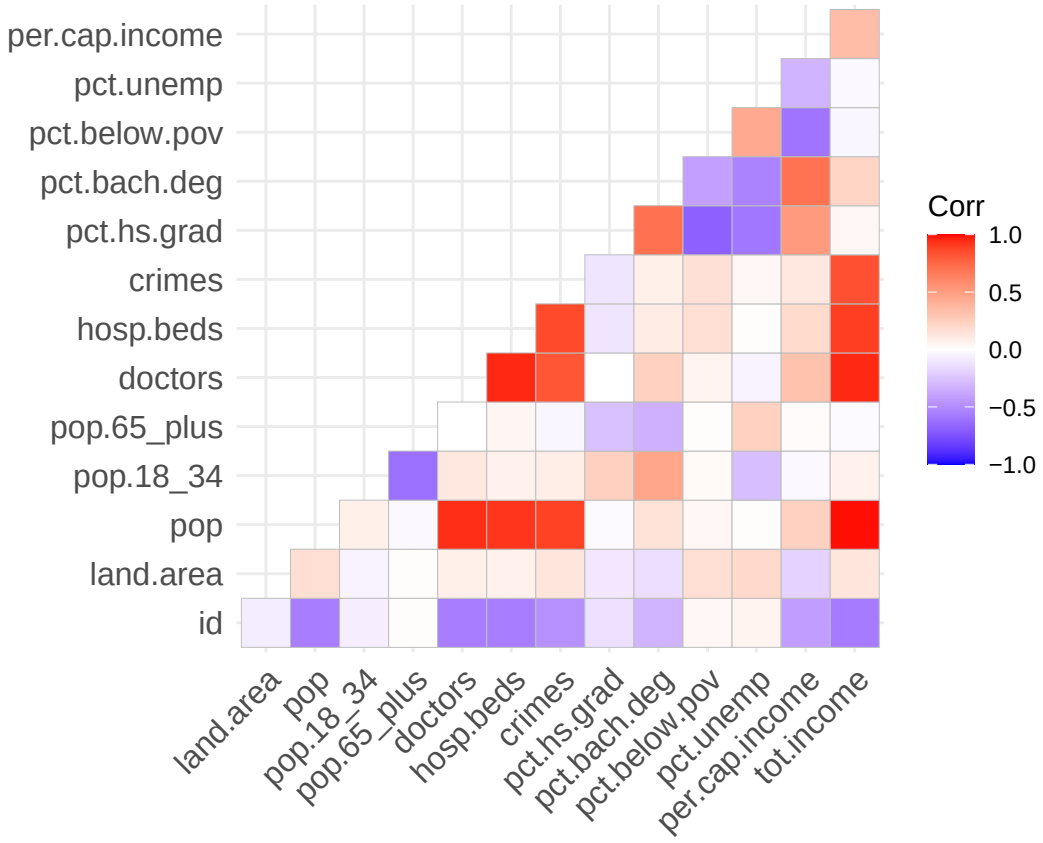


Figure 1: Scatter plots of response variable and each predictor.




```
# ggsave(ggcorrplot(cor, type = "lower"), "/Users/Stefano_1/Documents/CMU/Applied Linear Models/Project
```

This models were fitted for Sections 4.2 and 5.2

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

powerTransform(cdi$crimes)
```

```
## Estimated transformation parameter
## cdi$crimes
## -0.1307109
```

```
powerTransform(cdi$per.cap.income)
```

```
## Estimated transformation parameter
## cdi$per.cap.income
##         -0.3683365
```

```
powerTransform(cdi$crime_rate)
```

```
## Estimated transformation parameter
## cdi$crime_rate
##      0.3776893
```

```r
#cdi$per.cap.income3 <- cdi$per.cap.income^(-1/3)
cdi$per.cap.income3 <- log(cdi$per.cap.income)

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

summary(reg1)
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region - 1, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9661.0 -2260.7  -618.3  1650.0 19492.6
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## crimes   8.915e-03  3.188e-03   2.797  0.00539 **
## regionNC 1.811e+04  3.784e+02  47.846  < 2e-16 ***
## regionNE 2.039e+04  3.880e+02  52.562  < 2e-16 ***
## regionS  1.725e+04  3.252e+02  53.038  < 2e-16 ***
## regionW  1.796e+04  4.588e+02  39.150  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3866 on 435 degrees of freedom
## Multiple R-squared:  0.9591, Adjusted R-squared:  0.9586
## F-statistic:  2038 on 5 and 435 DF,  p-value: < 2.2e-16
```

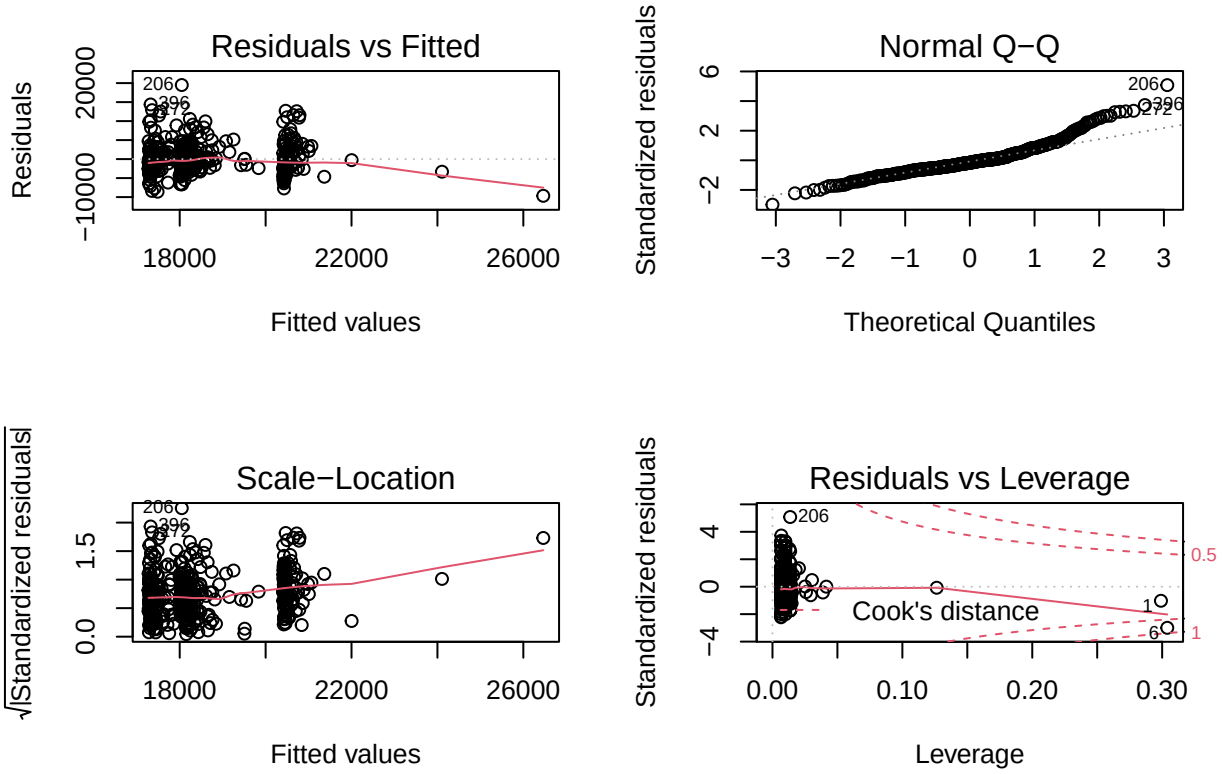
```r
reg1.2 <- lm(per.cap.income ~ crimes + region + crimes:region - 1, data = cdi)

summary(reg1.2)
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region + crimes:region -
##     1, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8582.4 -2225.2  -676.2  1563.4 19504.7
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## crimes            1.361e-02  7.882e-03   1.726   0.0851 .
## regionNC          1.800e+04  4.092e+02  43.995   <2e-16 ***
## regionNE          2.058e+04  4.019e+02  51.206   <2e-16 ***
## regionS           1.695e+04  3.831e+02  44.241   <2e-16 ***
## regionW           1.795e+04  4.885e+02  36.743   <2e-16 ***
## crimes:regionNE -1.272e-02  9.677e-03  -1.314   0.1895
## crimes:regionS   6.348e-03  1.136e-02   0.559   0.5765
## crimes:regionW  -4.295e-03  9.486e-03  -0.453   0.6509
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3861 on 432 degrees of freedom
## Multiple R-squared:  0.9595, Adjusted R-squared:  0.9587
## F-statistic:  1278 on 8 and 432 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(reg1, which = 1)
plot(reg1, which = 2)
plot(reg1, which = 3)
plot(reg1, which = 5)
```



The code for the second set of models is included but not run since it was found to not be useful.

```
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)
```

The following section of code corresponds to Sections 4.3 and 5.3. It performs power transformations on the variables, and then runs the variable selection models on the transformed data. The models are compared and their coefficients displayed.

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

```
## Warning in estimateTransform.default(X, Y, weights, family, ...): Convergence
## failure: return code = 52
```

```
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)
```

```
## Model :
## per.cap.income3 ~ land.area + pop + pop.18_34 + pop.65_plus +
##     doctors + hosp.beds + crimes + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + tot.income + region + crime_rate
```

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

```
##                    GVIF Df GVIF^(1/(2*Df))
## land.area      1.673307  1        1.293564
## pop           70.141295  1        8.375040
## pop.18_34      2.905496  1        1.704551
## pop.65_plus    2.977294  1        1.725484
## doctors       17.954510  1        4.237276
## hosp.beds     12.160895  1        3.487247
## crimes        51.549139  1        7.179773
## pct.hs.grad    4.992429  1        2.234374
## pct.bach.deg   6.870890  1        2.621238
## pct.below.pov  5.856157  1        2.419950
```

```
## pct.unemp      2.200697  1        1.483475
## tot.income    55.888481  1        7.475860
## region         4.594342  3        1.289349
## crime_rate    13.606346  1        3.688678
```

```
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)
```

```
##   (Intercept)     land.area     pop.18_34   pop.65_plus       doctors
##  8.292009e+00 -3.453935e-02  2.533535e+00  4.688984e-02  4.914310e-02
##   pct.hs.grad  pct.bach.deg pct.below.pov     pct.unemp      regionNE
## -3.475534e-07  3.155843e-01 -2.403551e-01  6.291413e-02 -2.703603e-02
##       regionS       regionW    crime_rate
## -5.326742e-02 -4.165263e-03  2.994169e-01
```

```
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$lambda.1se))
lasso_coef <- as.matrix(coef(reg3_lasso))[coef_lasso !=0]
coef_lasso <- rownames(coef_lasso)[coef_lasso !=0]
coef_lasso <- data.frame(Variable =coef_lasso,
                         LASSO = lasso_coef)
coefs_lasso <- c("per.cap.income3", coef_lasso$Variable[-1], "region")
cdi_lasoo <- cdi[,coefs_lasso]
reg3_lasso_coef <- lm(per.cap.income3~., data = cdi_lasoo)


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)
```

```
## Joining, by = "Variable"
## Joining, by = "Variable"
## Joining, by = "Variable"
```

Table 3: Coefficients for all the regression models

Variable	Regression	StepAIC	All.subsets	LASSO
(Intercept)	8.2796428896	8.2920091903	8.1939523851	8.82406577
land.area	-0.0350811881	-0.0345393452	-0.0351289003	-0.02702814
pop.18_34	2.5296098723	2.5335347541	3.1569371942	1.56335564
pop.65_plus	0.0444645424	0.0468898414	NA	0.04573100
doctors	0.0403910313	0.0491431016	0.0508781315	0.06118158
hosp.beds	0.0103636165	NA	NA	NA
pct.hs.grad	-0.0000003493	-0.0000003476	-0.0000003308	NA
pct.bach.deg	0.3205310002	0.3155843119	0.3053851622	0.17673355
pct.below.pov	-0.2434105672	-0.2403550928	-0.2309455770	-0.20128458
pct.unemp	0.0638065575	0.0629141347	0.0541136732	0.02082006
regionNE	-0.0262517681	-0.0270360269	NA	NA
regionS	-0.0511571376	-0.0532674234	-0.0486149747	NA
regionW	0.0008576372	-0.0041652627	NA	NA
crime_rate	0.2892592844	0.2994169182	0.3317425153	NA

The models for the stepwise and all subsets methods are compared in the following chunk of code.

```
#comparing the subsetaic and allsubsets

summary(reg3_stepaic)
```

```
##
## Call:
## lm(formula = per.cap.income3 ~ land.area + pop.18_34 + pop.65_plus +
##     doctors + pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     region + crime_rate, data = cdi_final)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.308172 -0.045235  0.002948  0.045195  0.283228
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.292e+00  1.639e-01  50.602  < 2e-16 ***
## land.area     -3.454e-02  5.530e-03  -6.246 1.02e-09 ***
## pop.18_34      2.534e+00  4.195e-01   6.039 3.37e-09 ***
## pop.65_plus    4.689e-02  2.037e-02   2.301 0.021852 *
## doctors        4.914e-02  5.179e-03   9.488  < 2e-16 ***
## pct.hs.grad   -3.476e-07  7.145e-08  -4.864 1.62e-06 ***
## pct.bach.deg   3.156e-01  2.593e-02  12.171  < 2e-16 ***
## pct.below.pov -2.404e-01  1.281e-02 -18.760  < 2e-16 ***
## pct.unemp      6.291e-02  1.708e-02   3.684 0.000259 ***
## regionNE      -2.704e-02  1.355e-02  -1.995 0.046640 *
## regionS       -5.327e-02  1.231e-02  -4.326 1.89e-05 ***
## regionW       -4.165e-03  1.458e-02  -0.286 0.775299
## crime_rate     2.994e-01  1.014e-01   2.953 0.003321 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.08273 on 427 degrees of freedom
## Multiple R-squared:  0.8442, Adjusted R-squared:  0.8399
## F-statistic: 192.9 on 12 and 427 DF,  p-value: < 2.2e-16
```

```
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)
```

```
##
## Call:
## lm(formula = per.cap.income3 ~ ., data = aux)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32708 -0.04415  0.00008  0.04476  0.28592
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.192e+00  1.588e-01  51.576  < 2e-16 ***
## land.area     -3.521e-02  5.549e-03  -6.346 5.65e-10 ***
## pop.18_34      3.158e+00  3.217e-01   9.817  < 2e-16 ***
## doctors        5.172e-02  5.082e-03  10.178  < 2e-16 ***
## pct.hs.grad   -3.662e-07  7.135e-08  -5.132 4.35e-07 ***
## pct.bach.deg   3.164e-01  2.606e-02  12.142  < 2e-16 ***
## pct.below.pov -2.351e-01  1.267e-02 -18.556  < 2e-16 ***
## pct.unemp      6.080e-02  1.714e-02   3.547 0.000432 ***
## crime_rate     2.937e-01  1.019e-01   2.883 0.004134 **
## regionNE      -2.104e-02  1.336e-02  -1.574 0.116167
## regionS       -5.649e-02  1.229e-02  -4.595 5.70e-06 ***
## regionW       -7.021e-03  1.460e-02  -0.481 0.630875
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08314 on 428 degrees of freedom
## Multiple R-squared:  0.8423, Adjusted R-squared:  0.8383
## F-statistic: 207.8 on 11 and 428 DF,  p-value: < 2.2e-16
```

```
anova(reg3_stepaic, reg3_all_subsets)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income3 ~ land.area + pop.18_34 + pop.65_plus + doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     region + crime_rate
## Model 2: per.cap.income3 ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + crime_rate + region
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    427 2.9224
## 2    428 2.9587 -1 -0.036248 5.2962 0.02185 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After choosing the best model with variable selection, it was worth trying to see if interacting the selected variables with region could potentially benefit the model. The mmplots show that some variables have room for improvement in terms of their fitting, so a model with interactions will be tried to be fitted.

```
#mmplot(reg3_all_subsets, inc.legend = FALSE)
```

An all-interaction model proved to be a better fit than the model with no interactions, so in order to develop a more understandable model, some interactions may be discarded to still get a good fit. The summary suggests that the variables that have significant coefficient in their interactions with region are: **pct.hs.grad**, **pct.below.pov**, and **pct.unemp**, so a model considering some, or all, of those could be considered.

```
reg3_all_subsets_int <- lm(per.cap.income3~.*region, data=aux)

summary(reg3_all_subsets_int)
```

```
##
## Call:
## lm(formula = per.cap.income3 ~ . * region, data = aux)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.258821 -0.044358 -0.005134  0.042640  0.279721
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.975e+00  3.921e-01  20.341  < 2e-16 ***
## land.area               -2.214e-02  1.554e-02  -1.425 0.154863
## pop.18_34                3.410e+00  8.025e-01   4.249 2.66e-05 ***
## doctors                  5.124e-02  1.107e-02   4.627 4.99e-06 ***
## pct.hs.grad             -3.698e-08  1.898e-07  -0.195 0.845601
## pct.bach.deg             2.350e-01  6.476e-02   3.629 0.000321 ***
## pct.below.pov           -1.675e-01  2.750e-02  -6.089 2.65e-09 ***
## pct.unemp                9.795e-02  3.275e-02   2.991 0.002951 **
## crime_rate               8.842e-02  1.689e-01   0.524 0.600868
## regionNE                -7.089e-01  5.876e-01  -1.206 0.228362
## regionS                 -1.870e-02  4.491e-01  -0.042 0.966808
## regionW                  9.417e-01  5.723e-01   1.646 0.100644
## land.area:regionNE       6.010e-04  2.106e-02   0.029 0.977250
## land.area:regionS       -2.055e-02  1.785e-02  -1.151 0.250289
## land.area:regionW       -5.620e-03  1.871e-02  -0.300 0.764087
## pop.18_34:regionNE       1.336e+00  1.162e+00   1.150 0.250838
## pop.18_34:regionS        5.421e-01  9.270e-01   0.585 0.558979
## pop.18_34:regionW       -1.320e+00  1.137e+00  -1.161 0.246338
## doctors:regionNE        -4.180e-03  1.543e-02  -0.271 0.786600
## doctors:regionS         -7.987e-03  1.408e-02  -0.567 0.570857
## doctors:regionW         -7.561e-03  1.628e-02  -0.464 0.642648
## pct.hs.grad:regionNE    -2.880e-07  2.591e-07  -1.111 0.267026
## pct.hs.grad:regionS     -2.328e-07  2.223e-07  -1.047 0.295695
## pct.hs.grad:regionW     -8.713e-07  2.447e-07  -3.561 0.000413 ***
```

```
## pct.bach.deg:regionNE   1.469e-01  9.520e-02   1.543 0.123568
## pct.bach.deg:regionS    9.146e-02  7.652e-02   1.195 0.232702
## pct.bach.deg:regionW    1.253e-01  8.416e-02   1.489 0.137283
## pct.below.pov:regionNE -4.042e-02  3.862e-02  -1.047 0.295944
## pct.below.pov:regionS  -4.532e-02  3.435e-02  -1.319 0.187832
## pct.below.pov:regionW  -2.281e-01  4.799e-02  -4.754 2.78e-06 ***
## pct.unemp:regionNE     -1.746e-02  5.342e-02  -0.327 0.743936
## pct.unemp:regionS      -1.488e-01  4.623e-02  -3.219 0.001391 **
## pct.unemp:regionW      -3.874e-02  4.835e-02  -0.801 0.423451
## crime_rate:regionNE     2.664e-01  2.546e-01   1.046 0.296005
## crime_rate:regionS      4.407e-01  2.517e-01   1.751 0.080709 .
## crime_rate:regionW      4.822e-01  4.164e-01   1.158 0.247538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07832 on 404 degrees of freedom
## Multiple R-squared:  0.8679, Adjusted R-squared:  0.8565
## F-statistic: 75.85 on 35 and 404 DF,  p-value: < 2.2e-16
```

```
anova(reg3_all_subsets, reg3_all_subsets_int)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income3 ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + crime_rate + region
## Model 2: per.cap.income3 ~ (land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + crime_rate + region) *
##     region
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    428 2.9587
## 2    404 2.4780 24   0.48068 3.2653 6.29e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Analyzing the model with interactions for the three variables and also the models discarding one of them, it looks like the best model, compared to the base model with no interaction, includes interaction with all three variables.

```
reg3_all_subsets_int <- lm(per.cap.income3~.+ pct.below.pov*region+pct.unemp*region, data=aux)

summary(reg3_all_subsets_int)
```

```
##
## Call:
## lm(formula = per.cap.income3 ~ . + pct.below.pov * region + pct.unemp *
##     region, data = aux)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33018 -0.04524 -0.00243  0.04518  0.28445
##
## Coefficients:
```

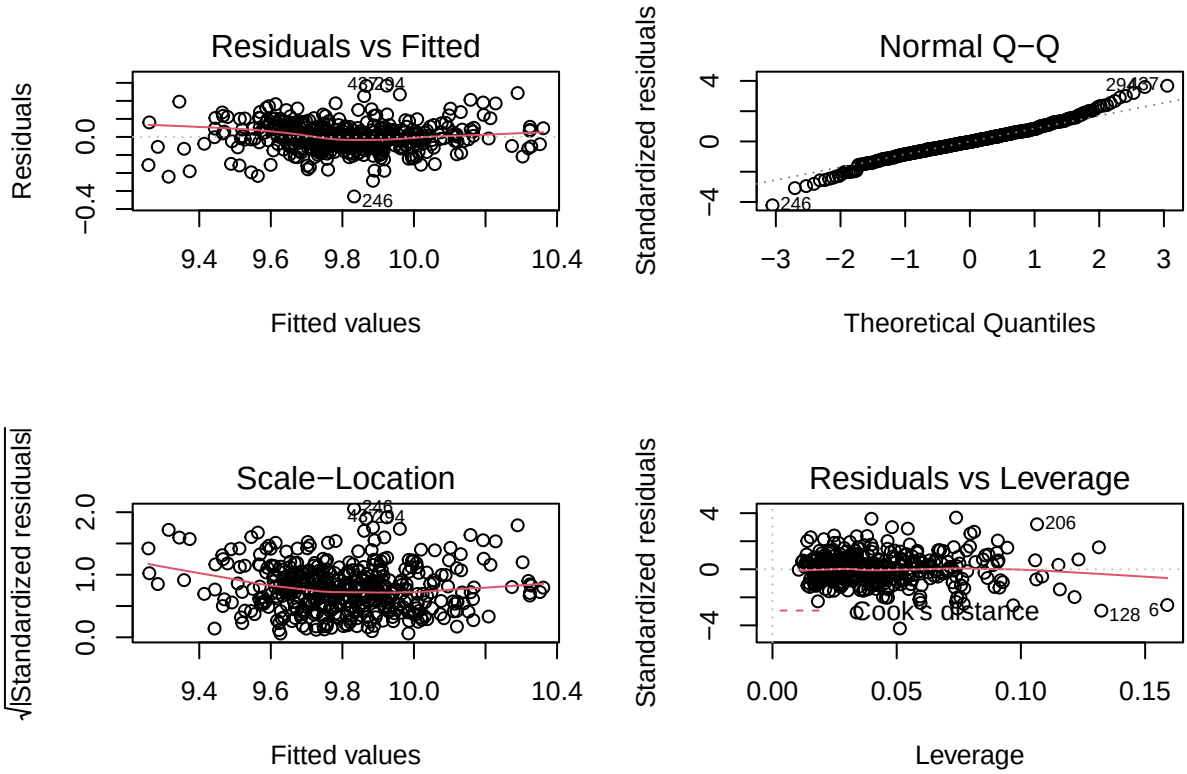
```
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              7.945e+00  1.603e-01  49.552  < 2e-16 ***
## land.area               -3.088e-02  5.554e-03  -5.561 4.78e-08 ***
## pop.18_34                3.382e+00  3.182e-01  10.629  < 2e-16 ***
## doctors                  5.060e-02  4.977e-03  10.168  < 2e-16 ***
## pct.hs.grad             -3.738e-07  7.255e-08  -5.152 3.96e-07 ***
## pct.bach.deg             3.219e-01  2.634e-02  12.218  < 2e-16 ***
## pct.below.pov           -2.035e-01  2.004e-02 -10.157  < 2e-16 ***
## pct.unemp                1.049e-01  2.588e-02   4.054 5.99e-05 ***
## crime_rate               2.943e-01  9.907e-02   2.970  0.00314 **
## regionNE                 1.047e-01  7.984e-02   1.311  0.19051
## regionS                  2.692e-01  6.263e-02   4.299 2.13e-05 ***
## regionW                  1.664e-01  6.808e-02   2.444  0.01495 *
## pct.below.pov:regionNE -3.068e-02  2.653e-02  -1.157  0.24812
## pct.below.pov:regionS  -2.285e-02  2.287e-02  -0.999  0.31834
## pct.below.pov:regionW  -9.979e-02  3.474e-02  -2.872  0.00428 **
## pct.unemp:regionNE     -3.745e-02  4.765e-02  -0.786  0.43239
## pct.unemp:regionS      -1.590e-01  3.572e-02  -4.450 1.10e-05 ***
## pct.unemp:regionW       1.373e-02  3.859e-02   0.356  0.72210
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08035 on 422 degrees of freedom
## Multiple R-squared:  0.8548, Adjusted R-squared:  0.8489
## F-statistic: 146.1 on 17 and 422 DF,  p-value: < 2.2e-16
```

```
anova(reg3_all_subsets, reg3_all_subsets_int)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income3 ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + crime_rate + region
## Model 2: per.cap.income3 ~ land.area + pop.18_34 + doctors + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + crime_rate + region +
##     pct.below.pov * region + pct.unemp * region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    428 2.9587
## 2    422 2.7243  6   0.23439 6.0513 4.322e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#mmplot(reg3_all_subsets_int, inc.legend = FALSE)
```

```
par(mfrow = c(2,2))
plot(reg3_all_subsets_int, which = 1)
plot(reg3_all_subsets_int, which = 2)
plot(reg3_all_subsets_int, which = 3)
plot(reg3_all_subsets_int, which = 5)
```


Residuals vs Fitted
Residuals
Fitted values
246
Normal Q−Q
Standardized residuals
Theoretical Quantiles
246
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
206