Investigating Associations between Average Income Per Person and Other Variables Related to County's Economy, Health and Social Well-being

Yueni Wang | Carnegie Mellon University | yueniw@andrew.cmu.edu

Abstract:

The motivation of this study is to discover how average income per person is related to other variables associated with county's economic, health and regions with a finalized statistical model. County demographic information (CDI) data for 440 of the most populous counties in the United States was obtained to conduct the research. Using multi linear regression, stepwise selections, LASSO and ANOVA tests, I compared and justified the models I obtained to get my final effective model. Based on my results, there happen to be clear linear relationships between average income per person and unemployment rates, degrees earned and other factors. These findings might need further adjustments because of possible associations, collinearity between predictors and lacking data from other states also affects the performance of final model.

Introduction:

How much money can people living in the U.S. expect to earn across different counties and states? There's is no simple answer to this question because there are many possible factors that potentially affect per capita income. To study for the possible factors that might affect per capita income, statistical models and methods are used to analyze data from various counties in the U.S. In addition, the study addresses the following problems:

1. Are there any relationships between each pair of variables? What are the possible reasons that they are related?
2. What does the crime and region from the data say about a theory that, if we ignore all other variables, per-capita income should be related to crime rate, and that this relationship may be different in different regions of the country (Northeast, North-central, South, and West)? How about crime rate per capita?
3. What is the best model predicting per-capita income from the other variables? Among all models, which one is most clearly indicated by the data and most of interest in terms of social sciences?
4. What is the impact of missing counties and states in the data when evaluating the model's performance and the precision of our conclusion?

With the utilization of statistical methods and CDI dataset, the following sections would address these problems in detail.

Data:

The CDI data we used in the study is taken from Kutneret al. (2005)1: It provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state

abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The definitions of the variables are given in the following table:

Variable Number	Variable Name	Description
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)

Table 1: Variable definitions for CDI data

Based on the output above, we generate a table of continuous variables, also as a description of important statistical values. Their minimum values, maximum values, median, mean, sd(standard deviation), 1st and 3rd quantile are displayed in the summary table.

	Min	Max	Median	1st Quantile	3rd Quantile	Mean	Sd
id	1.0	440.0	220.5	110.8	330.2	220.5	127.1613
land.area	15.0	20062.0	656.5	451.2	946.8	1041.4	1549.922
pop	100043	8863164	217280	139027	436064	393011	601987
pop.18_34	16.40	49.70	28.10	26.20	30.02	28.57	4.191083
pop.65_plus	3.000	33.800	11.750	9.875	13.625	12.170	3.992666
doctors	39.0	23677.0	401.0	182.8	1036.0	988.0	1789.75
hosp.beds	92.0	27700.0	755.0	390.8	1575.8	1458.6	2289.134
crimes	563	688936	11820	6220	26280	27112	58237.51
pct.hs.grad	46.60	92.90	77.70	73.88	82.40	77.56	7.015159
pct.bach.deg	8.10	52.30	19.70	15.28	25.32	21.08	7.654524
pct.below.pov	1.400	36.300	7.900	5.300	10.900	8.721	4.656737
pct.unemp	2.200	21.300	6.200	5.100	7.500	6.597	2.337924
per.cap.income	8899	37541	17759	16118	20270	18561	4059.192
tot.income	1141	184230	3857	2311	8654	7869	12884.32

Table2: Description of important statistical values for continuous variables

	NC	NE	S	W
Freq	108	103	152	77

Table 3: Descriptive statistics for categorial variable region

There are several variables with Mean substantially larger than Median (land.area, pop, doctors, hosp.beds, crimes, per.cap.income, and total.income), indicating possible right-skewing in their distribution. There are no variables with Mean substantially smaller than Median. In the table specially developed for the frequency of region variable, we also see some pattern. For the region variable (see technical appendix page2) it might be of some interest that the most counties are in the South (region 'S') and the least are in the west (region 'W'). The low number of counties in the West could be indicative of a lack of sampling in the West, or it could be that counties are just larger (in land area) in the West, so there are fewer counties to sample from. Similarly, the high number of counties in the South could be indicative of over-sampling, or perhap the South simply has a lot of counties that cover only small land areas.

From the frequency tables, we have further insight of the three categorial variables. It is apparent that there is not many duplicates or identical values in county names and the frequency of each name has a common range of 1 to 3, which stands for large variation regarding county locations. The frequency table of states has higher frequencies for each state's names, and even higher frequency for the four regions because there are only four regions in the dataset.

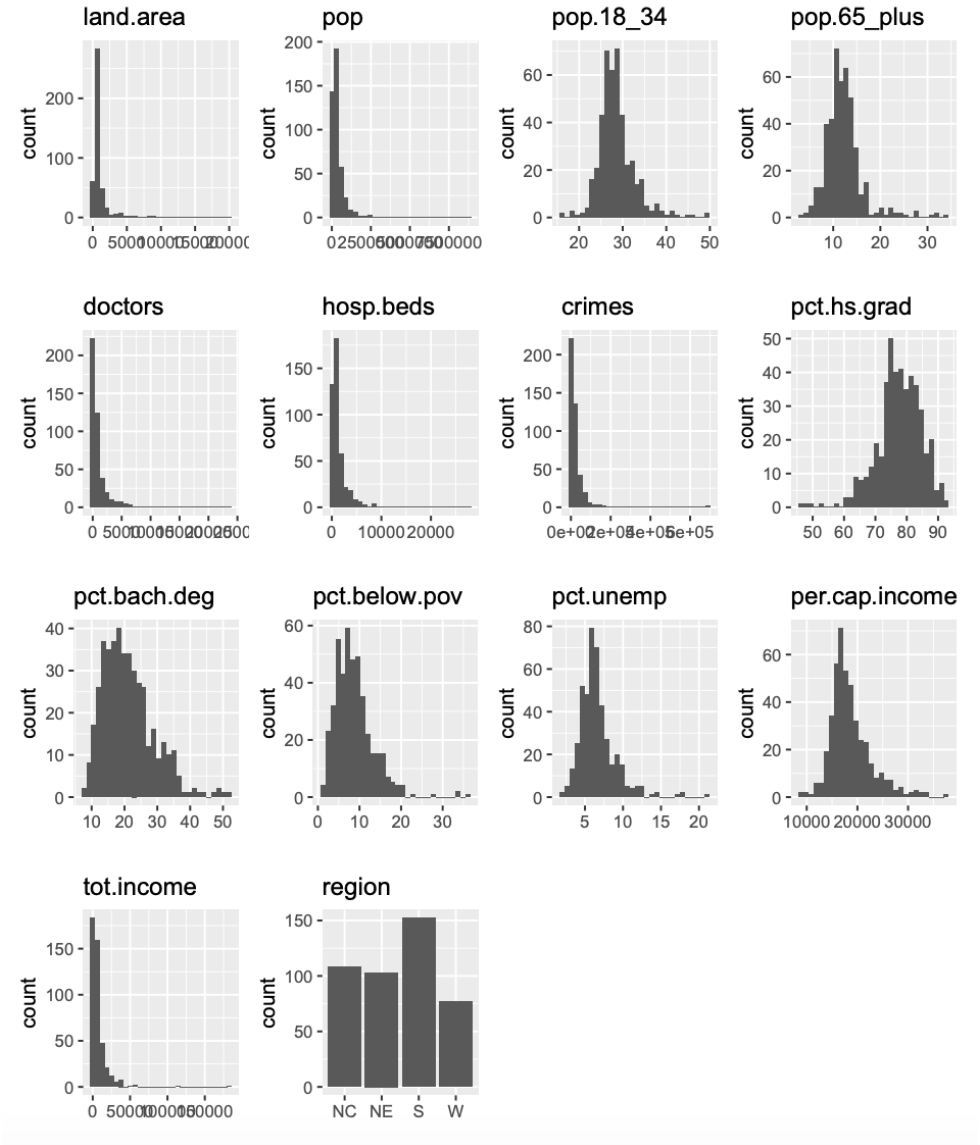
In addition to the numerical values, there are categorial variables included in this dataset.

	unique values
id	440
county	373
state	48
land.area	384
pop	440
pop.18_34	149
pop.65_plus	137
doctors	360
hosp.beds	391
crimes	437
pct.hs.grad	223
pct.bach.deg	220
pct.below.pov	155
pct.unemp	97
per.cap.income	436
tot.income	428
region	4

Table4: Unique values of each variable

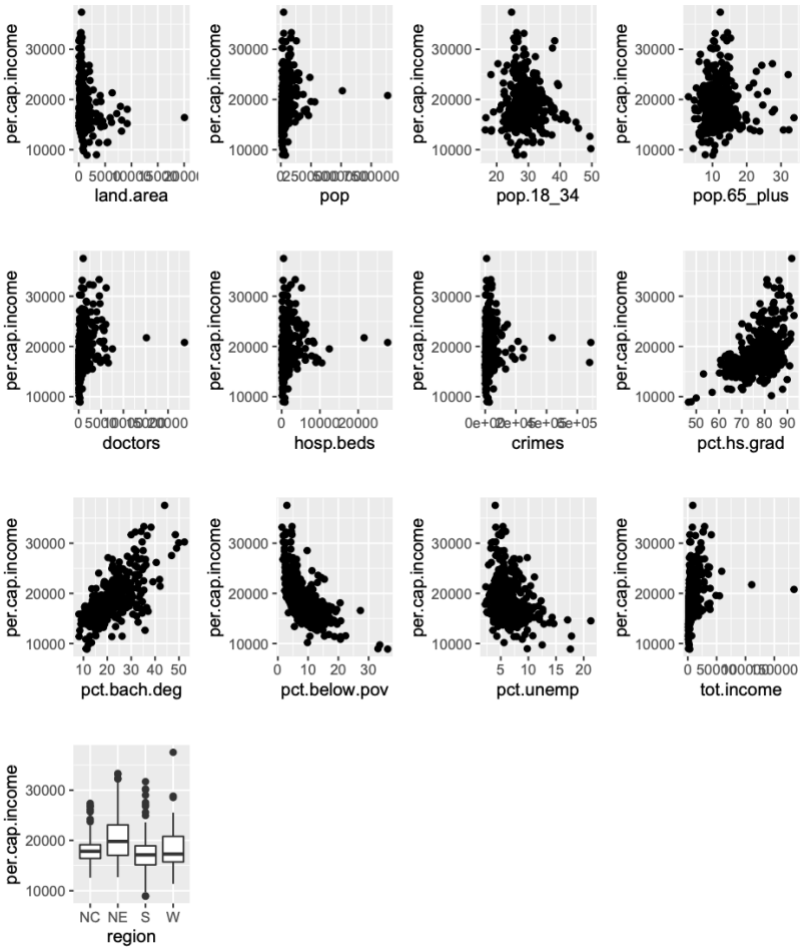
State has 48 values which is normally plenty enough to affect the model we plan to develop in this study, but it is hard to investigate on the features of each state. Therefore, it might not be included in the models. County is a categorical variable with nearly as many unique values (373) as rows in the cdidata data frame (440). A little more exploration shows that if I combine county with state, I get 440 unique values: some counties in different states have the same name.

I want to do some further data analysis by capturing univariate distributions using histograms as follow:



Plot 1: Histograms on each variable

Based on the univariate table which is also known as histogram, It looks like the variables that will really need attention (because they are severely right-skewed) are land.area, pop, doctors, hosp.beds, crimes, tot.income, and maybe per.cap.income.



Plot 2: Scatterplot of each variable

The scatterplot also provides us with valuable information of this set of data. The best possibilities for predicting per.cap.income are the same variables we identified from the correlation matrix: pct.hs.grad, pct.bach.deg,pct.below.pov, and pct.unemp. The last plot shows how per.cap.income varies across the four regions of the country. There is a lot of overlap in the boxplots, but the Northeast and the West seem to be doing a little better than the North Central and South regions.

After these initial investigations on the dataset, we are about to perform statistical methods to find our desired model and answer the rest of our research questions.

Methods:

To examine the first research question, I chose to do a visualization of correlation plot between each pair of variables from the dataset. The aim of developing a correlation plot is to check the potential correlation and collinearity that might appear between the predictors.

To address the second research question regarding the relationship between per capita income and crimes along with regions, the first method I used was log-transforming the two

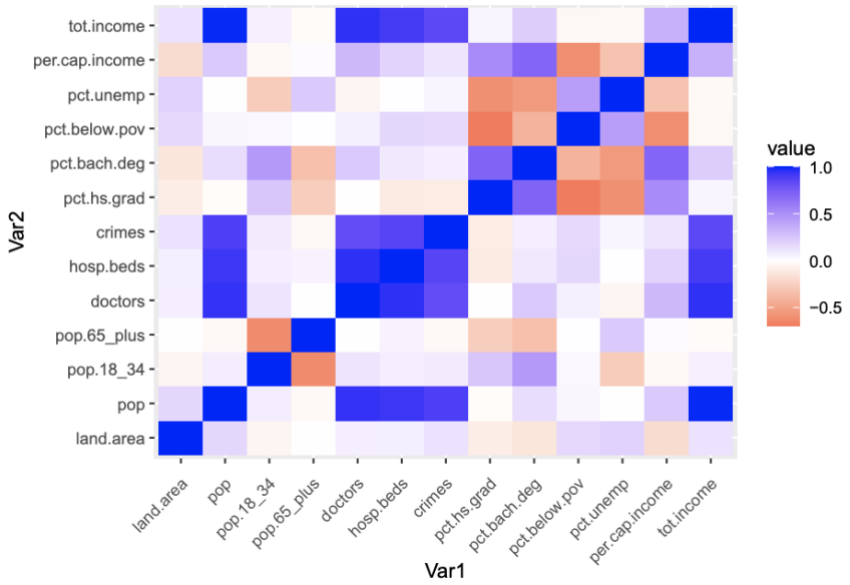
numerical variables that do not look normal in the above histograms to get a better fit. After this, I fitted three linear regression models with interaction terms being in one of them.

Anova test was used to figure out whether the model with interaction terms would have a better performance over other models. The AIC value was also compared in each two of the models.

To answer the third research question, log-transformation was again used on the variables that needed to be transformed. I performed stepwise backward selection comparing the AIC and BIC values of all subsets. LASSO method was also used to further justify our final chosen model. Looking at the diagnostic plots and the marginal variable plots generated by above methods, the results are listed as follow.

Results:

The results generated by the above methods are listed in this section.



Plot 3: Correlation plot between variables

We can make the following conclusions from the correlation matrix mentioned in the first method:

• tot.income and pop are highly correlated (no surprise there)

• both are reasonably highly correlated with crimes, hosp.beds and doctors

• the three variables crimes, hosp.beds and doctors seem strongly correlated with one another

• per.cap.income isn't really highly correlated with anything, but the best possibilities seem to be pct.hs.grad, pct.bach.deg (postively correlated with per.cap.income) and pct.below.pov, pct.unemp (negatively correlated with per.cap.income); all four of these variables are moderately highly correlated with one another

After performing log transformations, we have more normalized data which would help us in fitting the data.

Fitting the six models (three with the original crimes value, three with the per capita crime ratio), we find that both the second model with log crimes/per capita crimes value and region has the best performance. (see appendix page 15-18) The second model has better performance than model 1 with a significant p-value.

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.crimes
## Model 2: log.per.cap.income ~ log.crimes + region
## Model 3: log.per.cap.income ~ log.crimes * region
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 17.271
## 2    435 14.949  3   2.32194 22.4823 1.523e-13 ***
## 3    432 14.872  3   0.07678  0.7434    0.5266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table5: ANOVA result for percapita income w. crime+region models

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.per.cap.crimes
## Model 2: log.per.cap.income ~ log.per.cap.crimes + region
## Model 3: log.per.cap.income ~ log.per.cap.crimes * region
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 18.697
## 2    435 16.952  3   1.74465 14.8407 3.263e-09 ***
## 3    432 16.928  3   0.02408  0.2048     0.893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table6: ANOVA result for percapita income w. per capita crime+region models

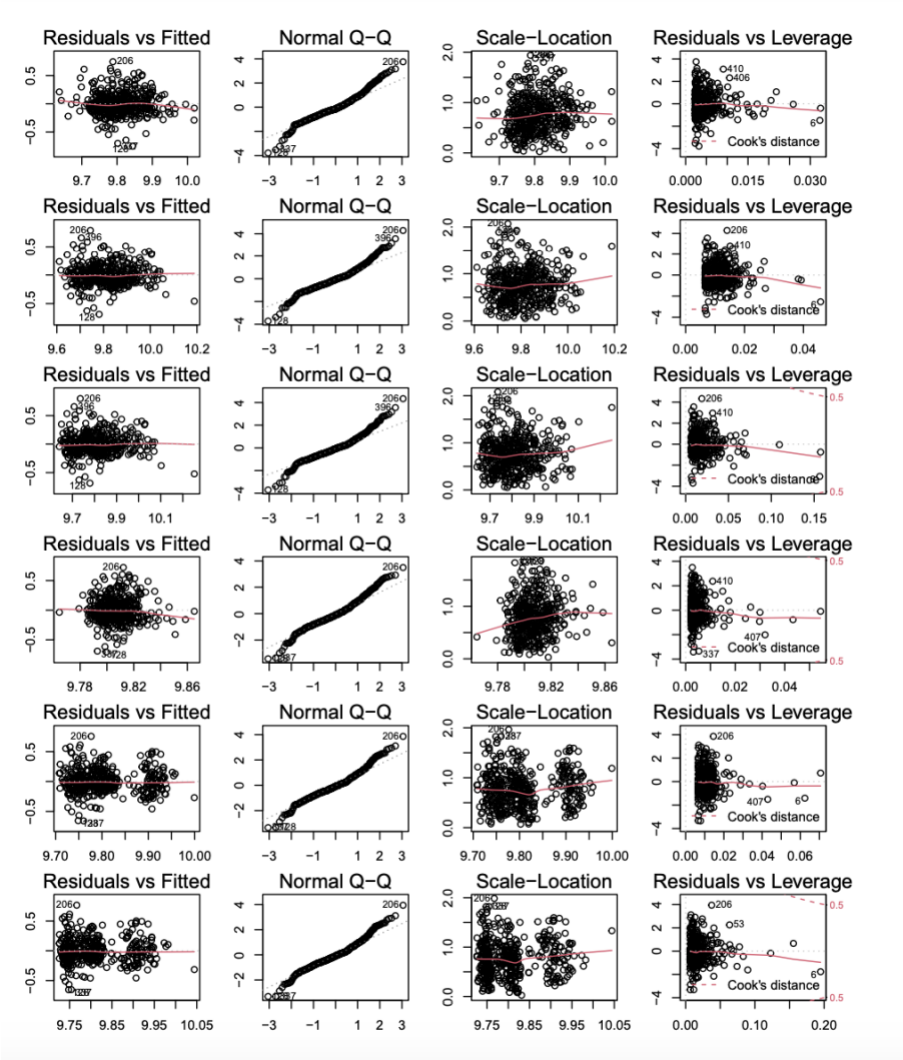
On the other hand, to further compare model2 and same model with per capita crimes, based on the AIC and BIC value, the second model with original crimes data as predictor has better performance with both smaller AIC and BIC values (see technical appendix page 18).

We can interpret the model as follows:

• All across the US, for every 1% increase in crimes, we expect a 0.07% increase in per-capita income, om average (this increase is statistically significant, but is it practically significant?).

• Different regions of the country have different baseline per-capita incomes however: In the NC region, the baseline salary is exp(9.19) = $9,798.65. In the NE it is exp(9.19+0.010) = $10,829.18, and so forth, so in the S it is $8,955.29, and in the W it is $9,228.02. All these region baselines are, according to the model, significantly different from the NC baseline.

• according to the model, the level of salary varies with region in the US, but the way it is related to crime does not.



Plot 4: Model diagnostic plots

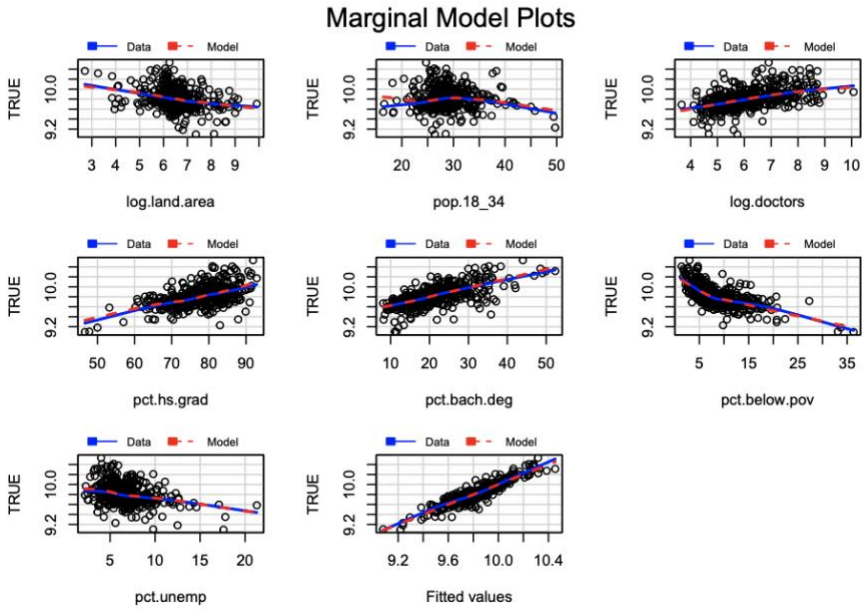
The diagnostic plots (see technical appendix page 19) also indicate that the second model has the best performance compared with other models.

To examine the third research question, after performing necessary log transformations, we fit the model first using all subsets and look at the complete model's BIC values while selecting the predictors with lowest BIC values to compose a best model (see technical appendix page 20).

```
##                  Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)   10.222495041 0.0931210074 109.776465 1.127483e-317
## log.land.area -0.035674062 0.0047767371  -7.468291  4.533156e-13
## pop.18_34     -0.013900201 0.0011113007 -12.508046  7.514862e-31
## log.doctors    0.060676872 0.0040183327  15.100012  1.133432e-41
## pct.hs.grad   -0.004406396 0.0010822796  -4.071403  5.558448e-05
## pct.bach.deg   0.015385301 0.0009245509  16.640838  2.100590e-48
## pct.below.pov -0.024278371 0.0012583372 -19.294011  2.812246e-60
## pct.unemp      0.010603691 0.0021771148   4.870525  1.564524e-06
```

Table7: Coefficients of our final model

According to the model summary, all p values are significant. After this, we also look at diagnostic plots and marginal model plots to justify our choice (see tech appendix page 21).



Plot5: Marginal model plots of final model

The marginal plots and diagnostic plots look very good. The VIF table also proved the model is a good fit. To investigate the importance of region, we take region as a factor in our model and perform the above selection (see tech appendix page 20-22).

Here are our findings on the final model:

• For every 1% increase in a county's land area, there is a 0.03% decrease in expected per-capita income. 27 (We might conjecture that this could be due to an urban-rural contrast: rural counties tend to be bigger than urban ones).

• For every 1% increase in the number of doctors in a county, the expected per-capita income increases by about 0.06%. That makes sense; doctors are well-paid and could be big contributors to the per-capita average income.

• For every 1 percentage point increase in the percent of the population aged 18–34, there is an expected 2% drop in per-capita income. (We might conjecture that this is because 18–34-year-olds are not at peak earning capacity yet and so perhaps their lower incomes drags down the per-capita average).

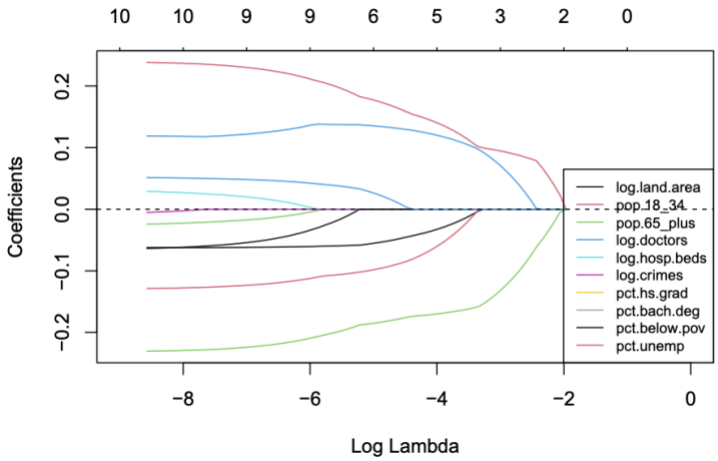
• percent of the population that are high school graduates doesn't have much effect, except in the South, where a one percentage point increase in graduates induces an expected 2% decrease in per-capita income. It might depend on whether college graduates are counted as a subset of hs graduates rather than counting them separately, or it might have something to do with some unique feature of economics in the southern region of the US.

• In the main effect for region, and in several of the interactions for region, the West shows up as deviating significantly from the North Central part of the US.

In BIC-based model, there are two interactions that appear to be statistically significant, but their practical effect is almost zero on per-capita income. The story is similar in the AIC model, which has many more interaction terms, but only two with an effect as large as 0.01. Although both interactions models produced big jumps in AIC and BIC (much bigger than 10!), the improvement in R2 and R2 adj is small, for all the terms that have been added to the models. For these reasons I might be willing to discuss these interactions with the social scientist, but I am disinclined to include them in a final model. If I stick with the model found by all-subsets and stepAIC with a BIC penalty, then my conclusions about adding interactions with region will also be the same, and I will once again be led to all.subsets.01.final.with.some.region, which has some interesting and mostly-interpretable structure.

Stepwise selection and LASSO are much the same as above method: we fit a complete model first and then use stepAIC or BIC selection to take out unnecessary variables (see technical appendix 22-24) while making ANOVA comparisons.

The final model returned was basically the same as first selection. Depending on which lambda value we use, the LASSO model would have different variables (see technical appendix 24-26).



Plot6: LASSO visualization

Discussion:

The first research question targets the possible correlation in CDI data, and we see that there is solid reason to concern about collinearity. Observations suggest that we may run into multi-collinearity problems when we start fitting models, but we could still make acceptable models after performing some transformations and adjustments to predictors.

To answer the second question: are there potential relationships between per capita income and crime rates/regions, our final model would be the modelb1, with response variable with crime and region. First, we look at the model summaries, modelb1 has an adjusted r-squared value of 0.09288. The second modelb2 which takes interaction term into consideration does not return any significant p-values in those interactions, which implies non-significance of these interaction terms. The adjusted r-squared value is 0.09543 which is a little bit higher than modelb1.If we look at the diagnostic plots for the two models, we see that the residual plots have a random pattern. The normal QQ plot looked fine with most points lying on the normal

regression line. There is slight downward pattern in the scale-location plot which indicates violation in constant variance. Three points stand out in leverage points which might be outliers. When we transform the crimes per capita as predictor, there is not so much difference than model b1 and b2 because the adjusted r-squared value did not improve. This might be caused by the reason that crime does not do much contribution to response variable. If we look at the diagnostic plots for the two models, we see that the residual plots have a random pattern. The normal QQ plot looked fine with most points lying on the normal regression line. There are slight downward pattern in the scale-location plot which indicates violation in constant variance. Three points stand out in leverage points which might be outliers.

We construct ANOVA table to compare model b1 and b2, model b3 and b4. Both test generate insignificant p-values which failed to reject the null hypothesis that the compared models has no difference in performance. There is no preference towards model containing interaction term over model without interaction. Therefore, we tend to choose model that is not overfitting, and the adjusted r-squared value is higher in model1 compared to model3.

As a conclusion, we chose modelb1 as our final model. The final model indicates that crime_per_capita has no difference than crimes itself, and neither crimes nor crime_per_capita does not contribute to the per capita income too much. Per one unit of crime per capita would result in 5773.2 increase in per capita income but the predictor is still not significant.In modelb1, the interpretation for coefficient is: With every unit of increase in regionNE, there is expected to be 2286 increase in per capita income. With every unit of increase in regionS, there is expected to be -860.6 increase in per capita income. With every unit of increase in regionW, there is expected to be -142.8 increase in per capita income. With every unit of increase in crimes, there is expected to be 8915 increases in per capita income. If all predictors happen to be zero, per capita income will be 1811.

To answer the third question of choosing an appropriate final model with per capita income as the response variable, we look at the above results generated by stepwise selection. Based on AIC selection, our final model contains seven predictors regarding the prediction of log per capita income. The added variable plots also indicate the importance of adding new variable in terms of choosing these predictors to compose the final model.

Based on the interaction terms check, we found out that there are three interaction terms that matter. Therefore, in the final model, we decided to take two of these interaction variables into consideration.

Here are a few of the model's pluses and minuses, and some tradeoffs:

• Pluses:

– The model is parsimonious, and most of the estimated coefficients have the expected sign.

– The model is confirmed by stepwise and lasso procedures.

– Those procedures also found more complex models with somewhat better fit, but improvements in fit seemed small compared to the added complexity of the model.

– All of the variables are either in their original scale, or the have been replaced with their logarithm. This facilitates explaining the models to anyone who is interested in and knowledgeable about the social science & economics but less knowledgeable about technical matters.

• Minuses:

– The coefficient on pct.unemp seems to go the wrong way, and the coefficient on 'pct.hs.grad' is quite small, statistically and practically (it remains in the model because there is a noticable interaction that it participates in).

– The residual diagnostic plots are just OK. The fact that stepwise regression found some well-fitting models with interactions between continuous variables suggests exploring those more complex models in the future.

– I did not explore the state variable. Some of the relationship between these demographic variables and per capita income might be explainable in terms of varying economic policy from one state to the next. (If one includes state in the model, one could take out region because the two are perfectly collinear (states are entirely nested within regions).

Finally, it would be very useful to have additional data to compare some of the models we found. We are using reasonable methods for variable selection, but since it is all within-sample (our entire data set is our training sample), there is ample room for overfitting noise in the data. Some of our inferences about which variables to leave in or take out may be based on overly optimistic standard error estimates.

In our chosen final model (as an answer to question 3), we have apparent tradeoff that, after adding the interaction variables, not all the variables in the model appeared to be important (having a p-value larger than 0.05). However, based on the statistical meaning of the model and social meaning of model, I still decided to include these two interaction variables because the interaction terms made each single predictor more significant. This could be proven based on the summary statistics, with most significant terms having p-value approaching zero and a high adjusted r-squared value of 0.8415 which is 84.15% variation explained. Also, our final model has lowest AIC value. If we look at the diagnostic plots for the two models, we see that the residual plots have a random pattern which is good. The normal QQ plot looked fine with most points lying on the normal regression line. There are no downward or upward pattern in the scale-location plot which indicates no violation in constant variance. There are also no obvious outliers in our final model.

In social terms of meaning, our final model indicates that there is an apparent positive linear relationship between per capita income and per capita unemployment, one unit of unemployment could cause exp (0.106094) unit of change in per capita income. Also, other estimation coefficient would mean the same thing when talking about positive relationship (per capita bachelor's degree, doctors, population 65_plus, and the interaction term of per capita unemployment with region S). The negative linear relationship between per capita income and per capita below poverty, population aged between 18 and 34 and land area could be identified as follow: one unit increase of per capita below poverty level could cause exp (0.1971) increase in per capita income.

The study still needs adjustments because there are still some apparent flaws that revealed by the current methods. First, we only performed log transformation which might not be enough for the study to be universal. Box-cox transformations and ridge models could be fitted to this study and see the difference. In addition, we intentionally excluded the states and county variables when fitting the model. There could be some interactions between them and other predictors that potentially affect the result. Finally, the limited data size prevents us from exploring further possible associations and relationships between our response variable and the predictor.

As to address the final problem in this study, we know that there are only 48 states presented in the dataset and only 373 out of 3000 counties represented in the dataset. Based on

what we found in above methods, we could tell that the missing interactions and correlations could be a huge problem in investigating per capita income. Based on the published article of Association of Household Income with Life Expectancy and Cause-Specific Mortality in Norway (2005-2015), we see that there are also other important factors missing along with regional data. Therefore, it is reasonable to worry about the missing data and the impact brought to our final model decision. A possible way of solving this problem is to expand our dataset while gathering more data from reliable sources to perform above statistical analysis.

References:

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

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

Technical Appendix

Yueni Wang

10/29/2021

Data:

```
#Summary table:
summary(cdi)
```

```
##        id            county             state             land.area
##  Min.   :  1.0   Length:440         Length:440         Min.   :   15.0
##  1st Qu.:110.8   Class :character   Class :character   1st Qu.:  451.2
##  Median :220.5   Mode  :character   Mode  :character   Median :  656.5
##  Mean   :220.5                                         Mean   : 1041.4
##  3rd Qu.:330.2                                         3rd Qu.:  946.8
##  Max.   :440.0                                         Max.   :20062.0
##       pop             pop.18_34      pop.65_plus        doctors
##  Min.   : 100043   Min.   :16.40   Min.   : 3.000   Min.   :   39.0
##  1st Qu.: 139027   1st Qu.:26.20   1st Qu.: 9.875   1st Qu.:  182.8
##  Median : 217280   Median :28.10   Median :11.750   Median :  401.0
##  Mean   : 393011   Mean   :28.57   Mean   :12.170   Mean   :  988.0
##  3rd Qu.: 436064   3rd Qu.:30.02   3rd Qu.:13.625   3rd Qu.: 1036.0
##  Max.   :8863164   Max.   :49.70   Max.   :33.800   Max.   :23677.0
##    hosp.beds           crimes         pct.hs.grad     pct.bach.deg
##  Min.   :   92.0   Min.   :   563   Min.   :46.60   Min.   : 8.10
##  1st Qu.:  390.8   1st Qu.:  6220   1st Qu.:73.88   1st Qu.:15.28
##  Median :  755.0   Median : 11820   Median :77.70   Median :19.70
##  Mean   : 1458.6   Mean   : 27112   Mean   :77.56   Mean   :21.08
##  3rd Qu.: 1575.8   3rd Qu.: 26280   3rd Qu.:82.40   3rd Qu.:25.32
##  Max.   :27700.0   Max.   :688936   Max.   :92.90   Max.   :52.30
##  pct.below.pov      pct.unemp      per.cap.income    tot.income
##  Min.   : 1.400   Min.   : 2.200   Min.   : 8899   Min.   :  1141
##  1st Qu.: 5.300   1st Qu.: 5.100   1st Qu.:16118   1st Qu.:  2311
##  Median : 7.900   Median : 6.200   Median :17759   Median :  3857
##  Mean   : 8.721   Mean   : 6.597   Mean   :18561   Mean   :  7869
##  3rd Qu.:10.900   3rd Qu.: 7.500   3rd Qu.:20270   3rd Qu.:  8654
##  Max.   :36.300   Max.   :21.300   Max.   :37541   Max.   :184230
##     region
##  Length:440
##  Class :character
##  Mode  :character
##
##
##
```

```
#Calculating the sd value of each predictor:
sd(cdi$id)
```

Table 1:

id	county	state	land.area	pop	pop.18_34	pop.65_plus	doctors	hosp.beds	crimes
1	Los_Angeles	CA	4060	8863164	32.1	9.7	23677	27700	688936
2	Cook	IL	946	5105067	29.2	12.4	15153	21550	436936
3	Harris	TX	1729	2818199	31.3	7.1	7553	12449	253526
4	San_Diego	CA	4205	2498016	33.5	10.9	5905	6179	173821
5	Orange	CA	790	2410556	32.6	9.2	6062	6369	144524
6	Kings	NY	71	2300664	28.3	12.4	4861	8942	680966

Table 2:

id	pct.hs.grad	pct.bach.deg	pct.below.pov	pct.unemp	per.cap.income	tot.income	region
1	70.0	22.3	11.6	8.0	20786	184230	W
2	73.4	22.8	11.1	7.2	21729	110928	NC
3	74.9	25.4	12.5	5.7	19517	55003	S
4	81.9	25.3	8.1	6.1	19588	48931	W
5	81.2	27.8	5.2	4.8	24400	58818	W
6	63.7	16.6	19.5	9.5	16803	38658	NE

```
sd(cdi$land.area)
sd(cdi$pop)
sd(cdi$pop.18_34)
sd(cdi$pop.65_plus)
sd(cdi$doctors)
sd(cdi$hosp.beds)
sd(cdi$crimes)
sd(cdi$pct.hs.grad)
sd(cdi$pct.bach.deg)
sd(cdi$pct.below.pov)
sd(cdi$pct.unemp)
sd(cdi$per.cap.income)
sd(cdi$tot.income)
```

```
#Further data summary:
head(cdi[,1:10]) %>% kbl(booktabs=T,caption=" ") %>% kable_classic()
```

```
head(cdi[,c(1,11:17)]) %>% kbl(booktabs=T,caption=" ") %>% kable_classic()
```

```
#Check unique values:
apply(cdi,2,function(x) {length(unique(x))}) %>%
kbl(booktabs=T,col.names="unique values",caption=" ") %>%
kable_classic(full_width=F)
```

```
#Table for categorial variable region:
tmp <- rbind(with(cdi,table(region)))
row.names(tmp) <- "Freq"
tmp %>% kbl(booktabs=T,caption=" ") %>% kable_classic(full_width=F)
```

```
#Getting the frequency table:
count(cdi,'county')
```

Table 3:

	unique values
id	440
county	373
state	48
land.area	384
pop	440
pop.18_34	149
pop.65_plus	137
doctors	360
hosp.beds	391
crimes	437
pct.hs.grad	223
pct.bach.deg	220
pct.below.pov	155
pct.unemp	97
per.cap.income	436
tot.income	428
region	4

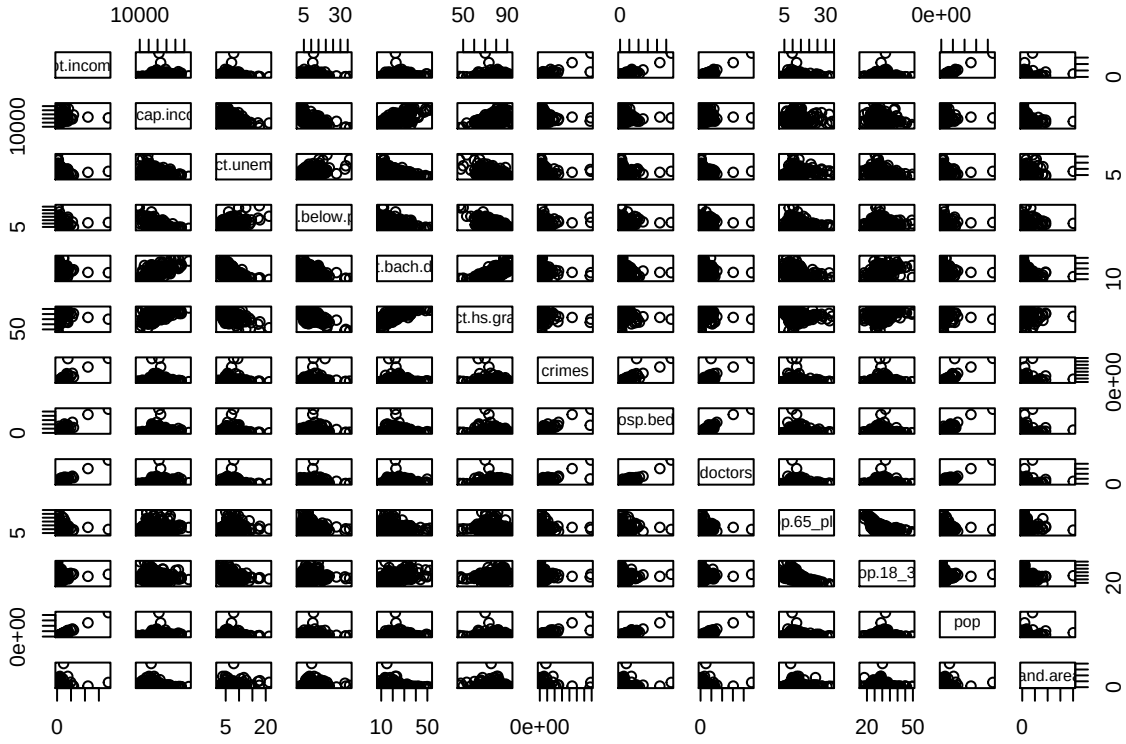
Table 4:

	NC	NE	S	W
Freq	108	103	152	77

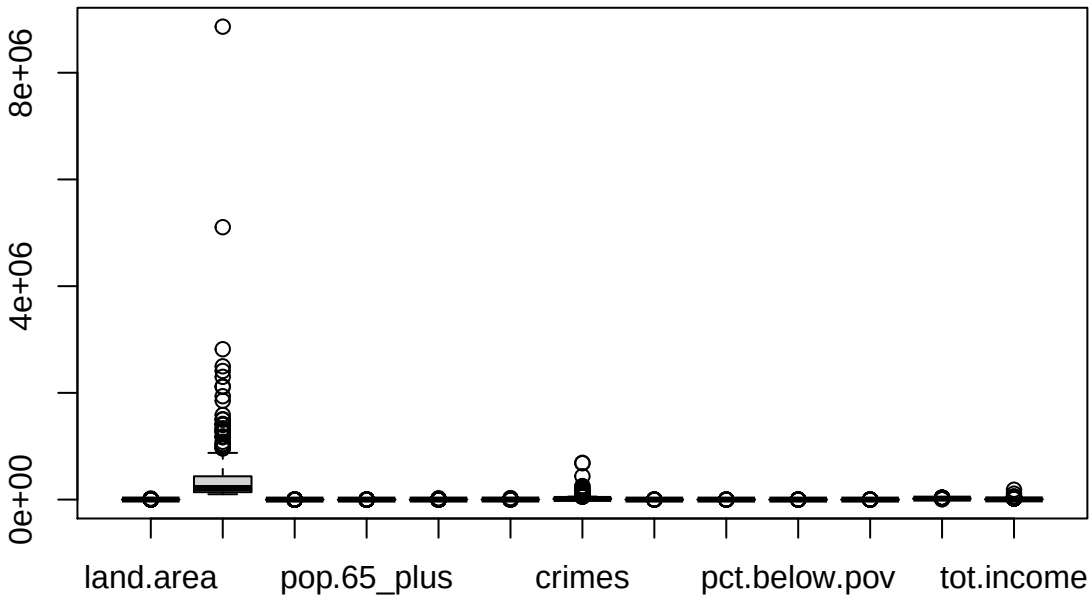
```
#Check for NA values(There's none according to the outcome):
apply(cdi,2,function(x) any(is.na(x)) )
```

```
##             id         county          state      land.area            pop
##          FALSE          FALSE          FALSE          FALSE          FALSE
##      pop.18_34    pop.65_plus        doctors      hosp.beds         crimes
##          FALSE          FALSE          FALSE          FALSE          FALSE
##    pct.hs.grad   pct.bach.deg  pct.below.pov      pct.unemp per.cap.income
##          FALSE          FALSE          FALSE          FALSE          FALSE
##     tot.income         region
##          FALSE          FALSE
```

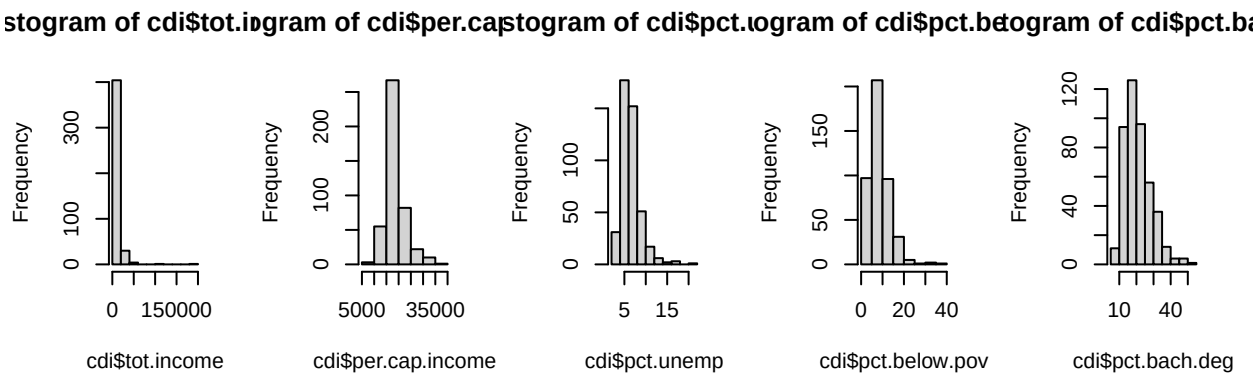
```
#Plots for the data:
attach(cdi)
pairs(tot.income ~per.cap.income+pct.unemp+pct.below.pov+pct.bach.deg+pct.hs.grad+crimes+hosp.beds+docto
```

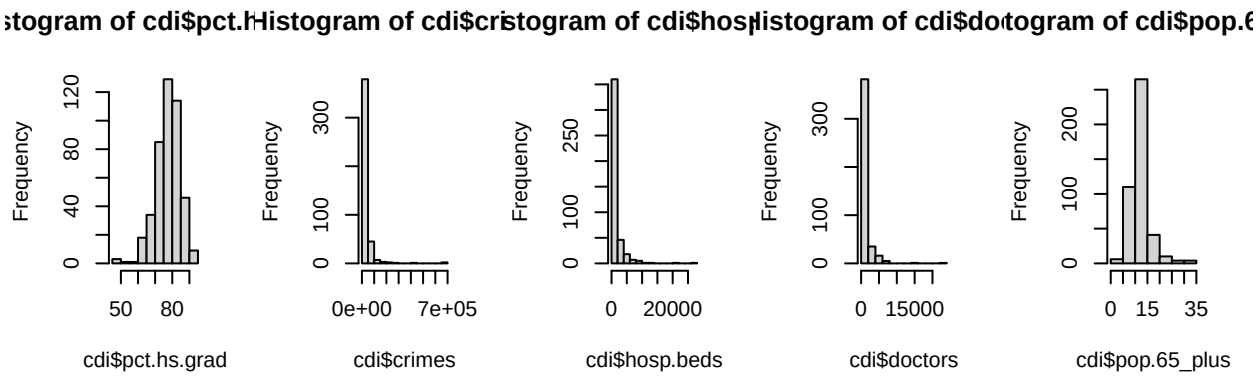



```
boxplot(cdi[,4:16])
```

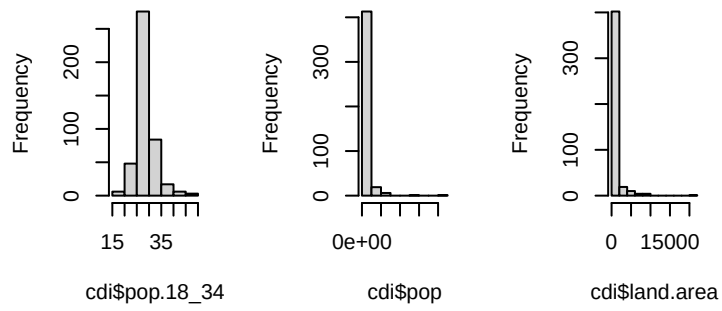



```
#Check the normalities:
par(mfrow = c(2,5))
hist(cdi$tot.income)
hist(cdi$per.cap.income)
hist(cdi$pct.unemp)
hist(cdi$pct.below.pov)
hist(cdi$pct.bach.deg)
hist(cdi$pct.hs.grad)
hist(cdi$crimes)
hist(cdi$hosp.beds)
hist(cdi$doctors)
hist(cdi$pop.65_plus)
```



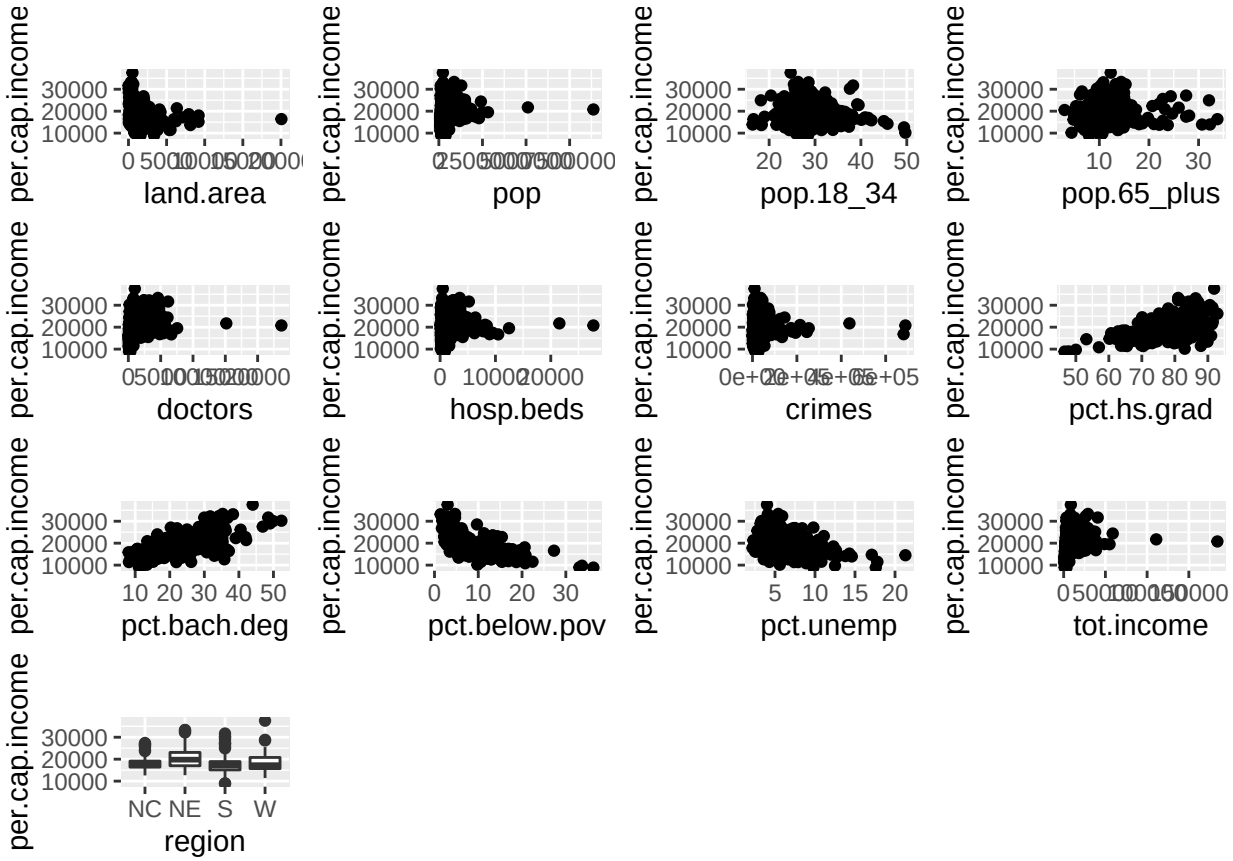



```
hist(cdi$pop.18_34)
hist(cdi$pop)
hist(cdi$land.area)
```



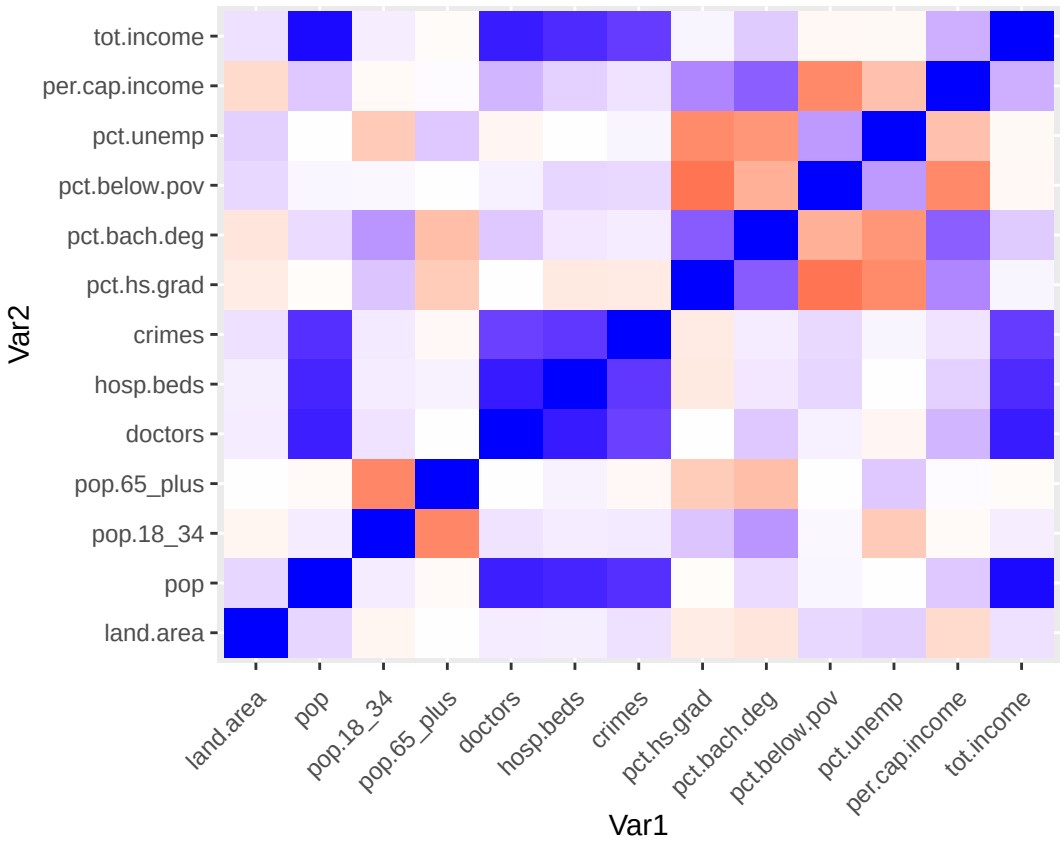

```
cdinumeric <- cdi[,-c(1,2,3,17)]
cdigood <- data.frame(cdinumeric,region=cdi$region)
scatter.builder <- function(df,yvar="per.cap.income") {
  result <- NULL
  y.index <- grep(yvar,names(df))
  for (xvar in names(df)[-y.index]) {
    d <- data.frame(xx=df[,xvar],yy=df[,y.index])
    if(mode(df[,xvar])=="numeric") {
      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(cdigood))
```



Methods and Results:

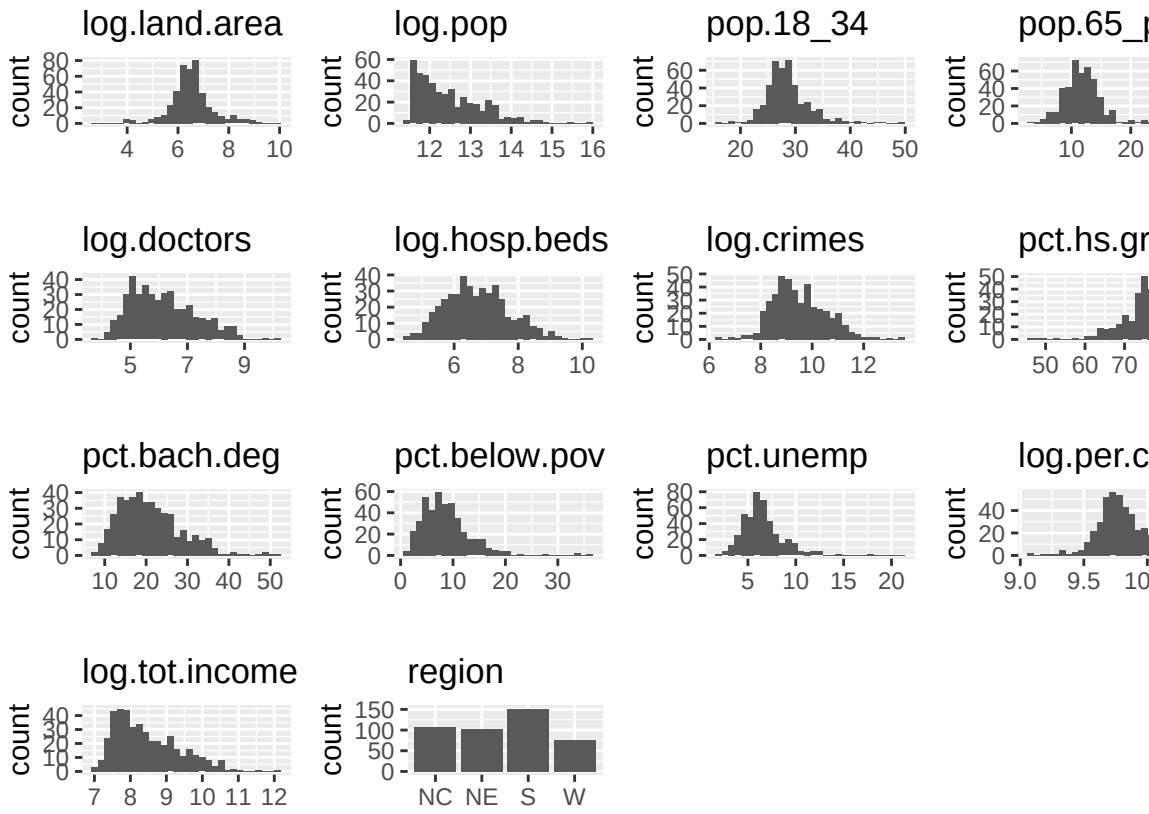
```
#Correlation plot:
cdinumeric <- cdi[,-c(1,2,3,17)]
corgraph <- function(df) {
cormat <- cor(df)
melted_cormat <- melt(cormat) ## need library(reshape2) for this...
ggplot(data = melted_cormat, aes(x=Var1, y=Var2, fill=value)) +
geom_tile() +
theme(axis.text.x = element_text(angle = 45,vjust=0.9,hjust=1)) +
scale_fill_gradient2(low="red",mid="white",high="blue")
}
corgraph(cdinumeric)
```



Research Question 1:

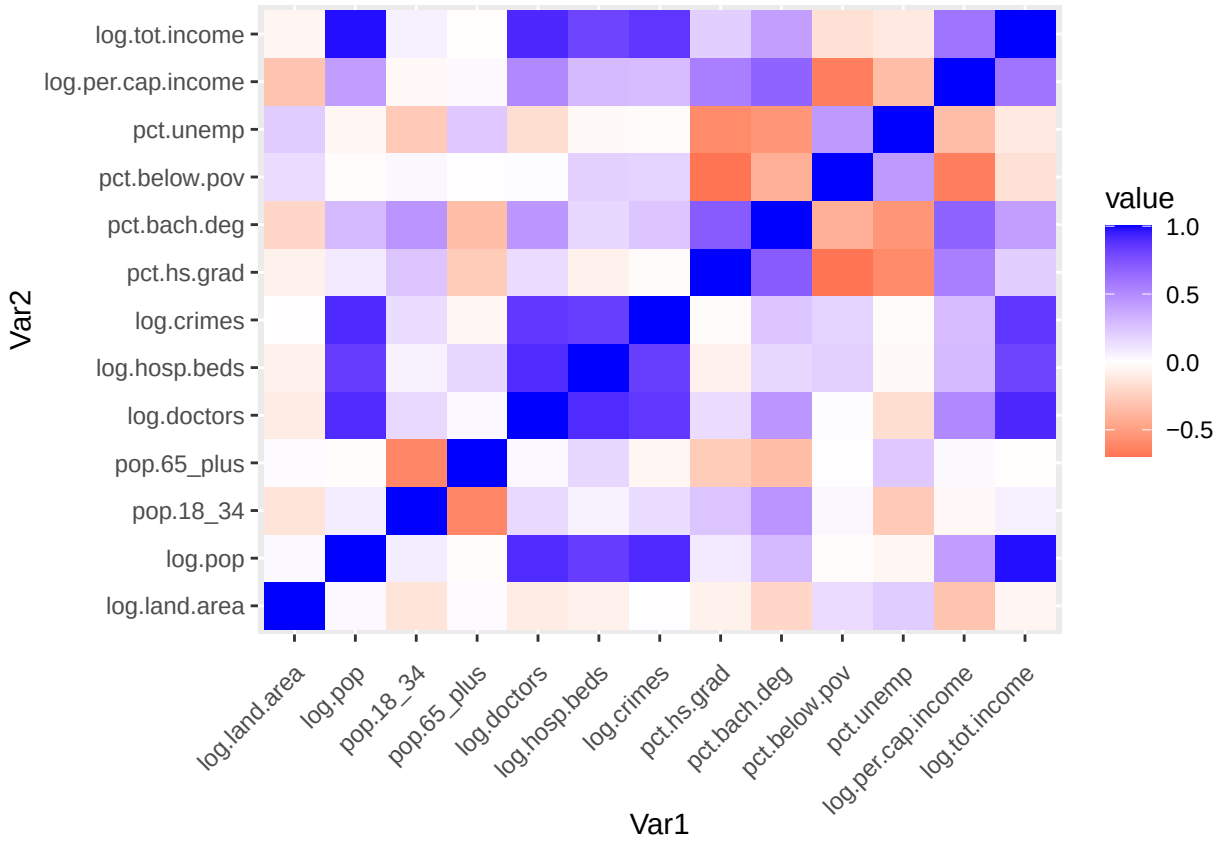
```
#Trasform the data to make it look normal:
cdilogs <- cdigood
skewed.vars <- c("land.area", "pop", "doctors", "hosp.beds", "crimes", "tot.income","per.cap.income")
for (tmp in skewed.vars) {
  loc <- grep(paste("^",tmp,"$",sep=""),names(cdilogs))
  cdilogs[,loc] <- log(cdilogs[,loc])
  names(cdilogs)[loc] <- paste("log.",names(cdilogs)[loc],sep="")
}

hist.builder <- function(df) {
  result <- NULL
  for (var in names(df)) {
    d <- data.frame(dd=df[,var])
    if(mode(df[,var])=="numeric") {
      p <- ggplot(d,aes(x=dd)) + geom_histogram() +
        ggtitle(var) + xlab("")
} else {
  p <- ggplot(d,aes(x=dd)) + geom_bar() +
    ggtitle(var) + xlab("")
  }
    result <- c(result,list(p))
}
  return(result)
}
grid.arrange(grobs=hist.builder(cdilogs))
```



Research Question 2:

```
corgraph(cdilogs[,-grep("region",names(cdilogs))])
```



```r
#Three models that are compared using ANOVA
modelb0 <- lm(log.per.cap.income ~ log.crimes,data=cdilogs)
modelb1 <- lm(log.per.cap.income ~ log.crimes + region,data=cdilogs)
modelb2 <- lm(log.per.cap.income ~ log.crimes * region,data=cdilogs)
anova(modelb0,modelb1,modelb2)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.crimes
## Model 2: log.per.cap.income ~ log.crimes + region
## Model 3: log.per.cap.income ~ log.crimes * region
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 17.271
## 2    435 14.949  3   2.32194 22.4823 1.523e-13 ***
## 3    432 14.872  3   0.07678  0.7434    0.5266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
modelb00 <- lm(log.per.cap.income ~ log.per.cap.crimes,data=cdilogs)
modelb3 <- lm(log.per.cap.income ~ log.per.cap.crimes + region,data=cdilogs)
modelb4 <- lm(log.per.cap.income ~ log.per.cap.crimes * region,data=cdilogs)
anova(modelb00,modelb3,modelb4)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.per.cap.crimes
## Model 2: log.per.cap.income ~ log.per.cap.crimes + region
## Model 3: log.per.cap.income ~ log.per.cap.crimes * region
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 18.697
## 2    435 16.952  3   1.74465 14.8407 3.263e-09 ***
## 3    432 16.928  3   0.02408  0.2048     0.893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Choose best model
AIC(modelb1,modelb3)
```

```
##         df       AIC
## modelb1  6 -227.4746
## modelb3  6 -172.1347
```

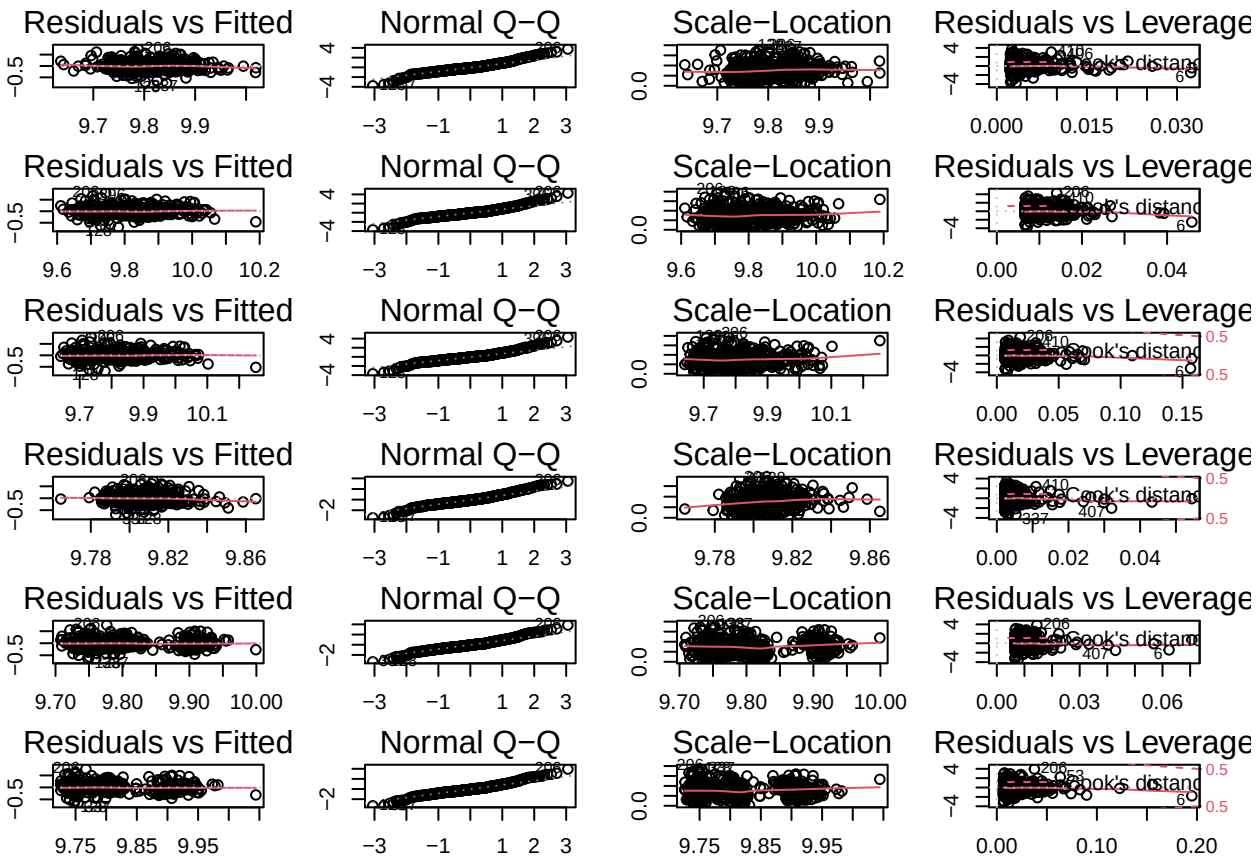
```r
BIC(modelb1,modelb3)
```

```
##         df       BIC
## modelb1  6 -202.9539
## modelb3  6 -147.6140
```

```r
round(coef(summary(modelb1)),2)
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9.19       0.08  115.13     0.00
## log.crimes      0.07       0.01    7.92     0.00
## regionNE        0.10       0.03    4.09     0.00
## regionS        -0.09       0.02   -3.68     0.00
## regionW        -0.06       0.03   -1.96     0.05
```

```
#Diagnostic plots:
oldmar <- par()$mar
par(mfrow=c(6,4))
par(mar=c(2,2,2,2))
invisible(lapply(list(modelb0,modelb1,modelb2,modelb00,modelb3,modelb4),
function(x) plot(x,cex.main=0.5)))
```




```
omit <- c(grep("log.pop",names(cdilogs)),grep("log.tot.income",names(cdilogs)))
cdilogred <- cdilogs[,-omit]
cdilogred.cont <- cdilogred[,-grep("region",names(cdilogred))]
names(cdilogred.cont)
```

Research Question 3:

```
##  [1] "log.land.area"      "pop.18_34"          "pop.65_plus"
##  [4] "log.doctors"        "log.hosp.beds"      "log.crimes"
##  [7] "pct.hs.grad"        "pct.bach.deg"       "pct.below.pov"
## [10] "pct.unemp"          "log.per.cap.income"
```

```
all.subsets.01 <- regsubsets(per.cap.income ~.,data=cdilogred.cont,nvmax=10)
all.subsets.01.summary <- summary(all.subsets.01)
print(best.model <- which.min(all.subsets.01.summary$bic))
```

```
## [1] 7
```

```
coef(all.subsets.01,best.model)
```

```
##        (Intercept)           pop.18_34      log.hosp.beds        pct.hs.grad
##      -170945.50317           -28.67957          -217.47742           -29.89116
##       pct.bach.deg      pct.below.pov          pct.unemp log.per.cap.income
##           61.99245           53.58504           40.46374        19584.61600
```

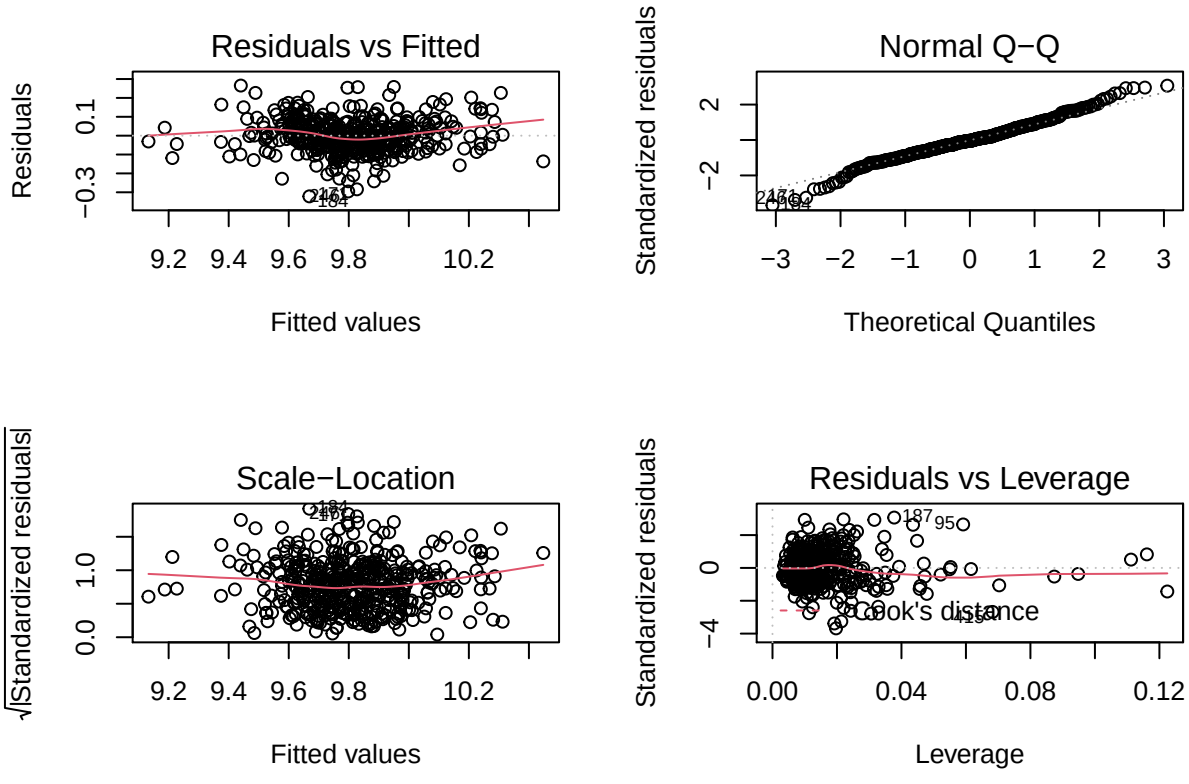
```
tmp <- cdilogred.cont[,all.subsets.01.summary$which[best.model,][-1]]
all.subsets.01.final.model <- lm(log.per.cap.income ~ .,data=tmp)
summary(all.subsets.01.final.model)$coef
```

```
##                   Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)   10.048956045 0.1015621774  98.943881 2.087901e-299
## pop.18_34     -0.013464995 0.0011914046 -11.301781  3.942244e-26
## log.hosp.beds  0.061713255 0.0045254164  13.637033  1.760463e-35
## pct.hs.grad   -0.006519522 0.0011306497  -5.766173  1.543887e-08
## pct.bach.deg   0.019347874 0.0009060899  21.353150  1.242875e-69
## pct.below.pov -0.026977105 0.0013448835 -20.059065  9.001190e-64
## pct.unemp      0.009230576 0.0023171894   3.983522  7.964658e-05
```

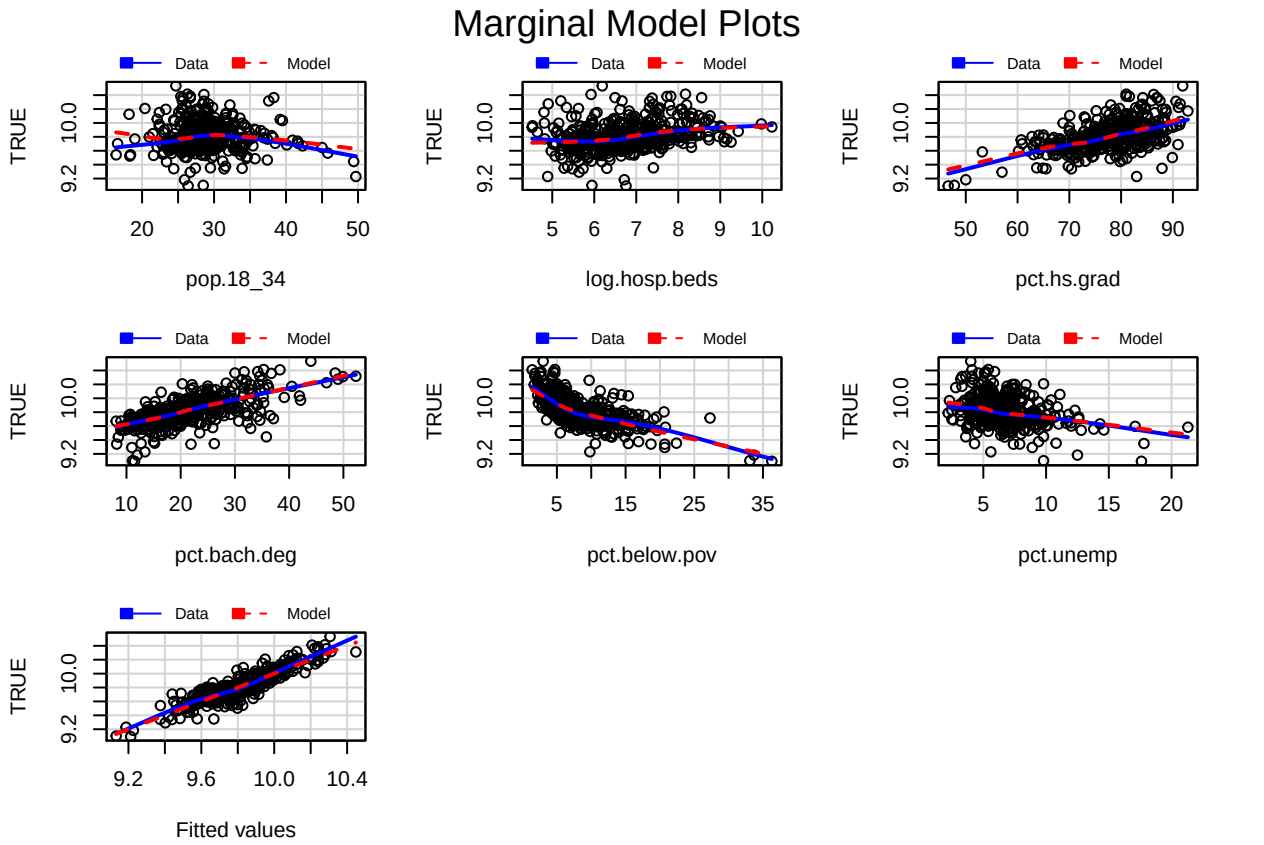
```
vif(all.subsets.01.final.model)
```

```
##     pop.18_34 log.hosp.beds   pct.hs.grad  pct.bach.deg pct.below.pov
##      1.406245      1.162743      3.548307      2.713120      2.212191
##     pct.unemp
##      1.655289
```

```
par(mfrow=c(2,2))
plot(all.subsets.01.final.model)
```



```
mmps(all.subsets.01.final.model)
```




```
tmp <- cbind(tmp,region=cdilogred$region)
all.subsets.01.final.with.region <- lm(log.per.cap.income ~ .*region,data=tmp)
all.subsets.01.final.with.some.region <-update(all.subsets.01.final.with.region,
. ~ . - region:log.land.area -region:pop.18_34 - region:log.doctors)
summary(all.subsets.01.final.with.some.region)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + log.hosp.beds +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     region + log.hosp.beds:region + pct.hs.grad:region + pct.bach.deg:region +
##     pct.below.pov:region + pct.unemp:region, data = tmp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22996 -0.04697 -0.00359  0.04498  0.32685
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.120027   0.285330  35.468  < 2e-16 ***
## pop.18_34              -0.015138   0.001198 -12.635  < 2e-16 ***
## log.hosp.beds           0.050030   0.010067   4.970 9.81e-07 ***
## pct.hs.grad            -0.006270   0.003446  -1.819 0.069565 .
## pct.bach.deg            0.018593   0.002388   7.786 5.58e-14 ***
## pct.below.pov          -0.024982   0.003874  -6.448 3.16e-10 ***
```

```
## pct.unemp               0.017158   0.005156   3.328 0.000953 ***
## regionNE                0.133176   0.358821   0.371 0.710716
## regionS                -0.332535   0.314030  -1.059 0.290249
## regionW                 1.482059   0.422614   3.507 0.000503 ***
## log.hosp.beds:regionNE  0.007687   0.014259   0.539 0.590094
## log.hosp.beds:regionS   0.005009   0.012307   0.407 0.684217
## log.hosp.beds:regionW  -0.002427   0.013981  -0.174 0.862266
## pct.hs.grad:regionNE   -0.003748   0.004330  -0.865 0.387313
## pct.hs.grad:regionS     0.005332   0.003866   1.379 0.168608
## pct.hs.grad:regionW    -0.016794   0.004711  -3.565 0.000406 ***
## pct.bach.deg:regionNE   0.007382   0.003117   2.369 0.018308 *
## pct.bach.deg:regionS   -0.002528   0.002677  -0.944 0.345531
## pct.bach.deg:regionW    0.006398   0.003035   2.108 0.035625 *
## pct.below.pov:regionNE -0.001302   0.005338  -0.244 0.807446
## pct.below.pov:regionS   0.006831   0.004242   1.610 0.108120
## pct.below.pov:regionW  -0.018081   0.005713  -3.165 0.001665 **
## pct.unemp:regionNE     -0.007742   0.007853  -0.986 0.324727
## pct.unemp:regionS      -0.024036   0.006884  -3.491 0.000532 ***
## pct.unemp:regionW      -0.019897   0.007191  -2.767 0.005911 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08141 on 415 degrees of freedom
## Multiple R-squared:  0.8534, Adjusted R-squared:  0.8449
## F-statistic: 100.6 on 24 and 415 DF,  p-value: < 2.2e-16
```

```
#Stepwise regression:
stepwise.base <- lm(log.per.cap.income ~ .,data=cdilogred.cont)
## try to duplicate all-subsets with BIC
step.result.01.bic <- stepAIC(stepwise.base,
scope=list(lower = ~ 1, upper = ~ .),
k=log(dim(cdilogred.cont)[1]),
trace=F)
anova(all.subsets.01.final.model,step.result.01.bic)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ pop.18_34 + log.hosp.beds + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    433 3.3703
## 2    432 2.9051  1   0.46518 69.176 1.187e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
step.result.01.aic <- stepAIC(stepwise.base,
scope=list(lower = ~ 1, upper = ~ .),
k=2,
trace=F)
anova(all.subsets.01.final.model,step.result.01.aic)
```

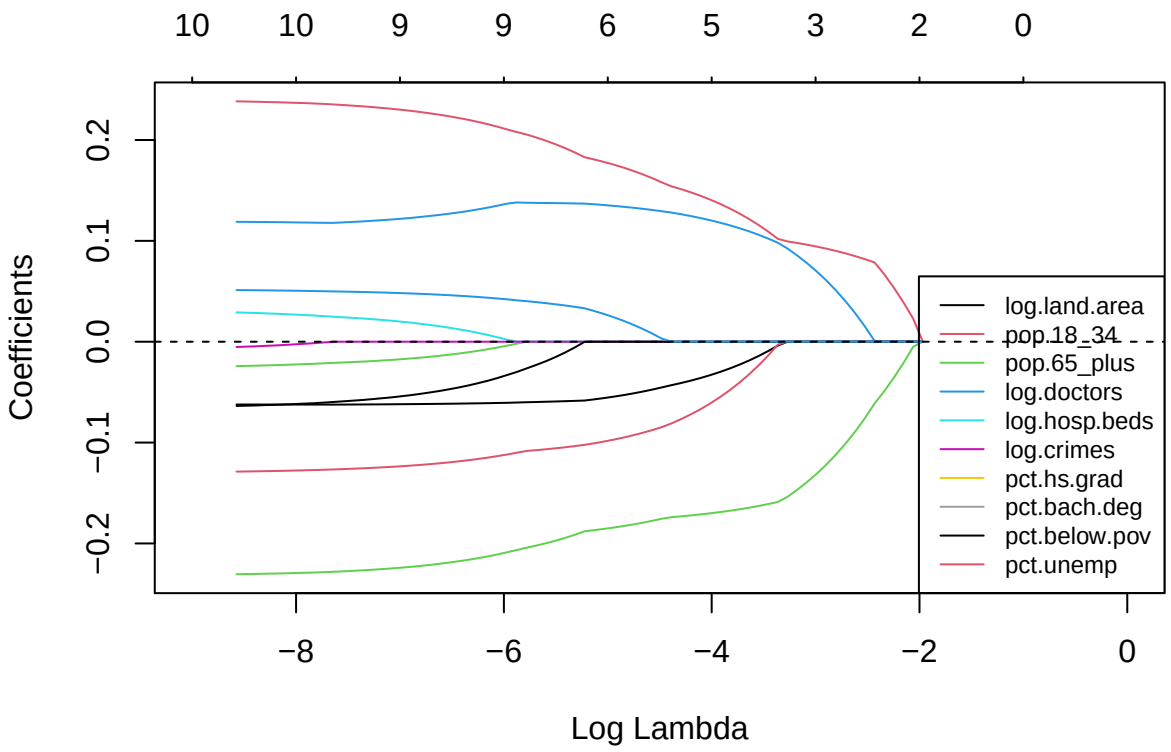
```
## Analysis of Variance Table
##
```

```
## Model 1: log.per.cap.income ~ pop.18_34 + log.hosp.beds + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    433 3.3703
## 2    431 2.8748  2   0.49549 37.143 1.31e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

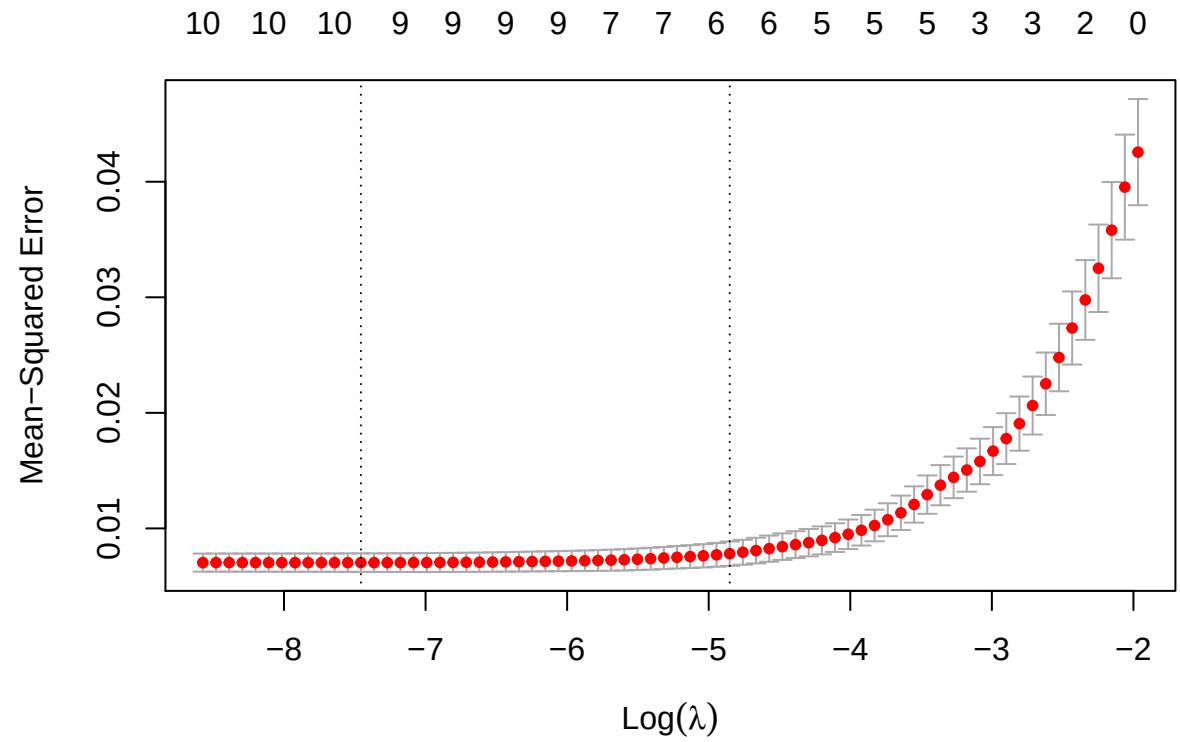
```
#Check if interaction term is important
step.result.02.bic <- stepAIC(stepwise.base,scope=list(lower = ~ 1, upper = ~ .^2),k=log(dim(cdilogred.
step.result.02.aic <- stepAIC(stepwise.base,scope=list(lower = ~ 1, upper = ~ .^2),k=2, trace=F)
comparison <- cbind(
AIC(all.subsets.01.final.model,step.result.01.aic,step.result.01.bic,
step.result.02.aic,step.result.02.bic),
BIC(all.subsets.01.final.model,step.result.01.aic,step.result.01.bic,
step.result.02.aic,step.result.02.bic))
comparison <- comparison[,-3]
names(comparison) <- c("df","AIC","BIC")
comparison %>% kbl(booktabs=T) %>% kable_classic()
```

	df	AIC	BIC
all.subsets.01.final.model	8	-878.9204	-846.2262
step.result.01.aic	10	-944.8883	-904.0206
step.result.01.bic	9	-942.2740	-905.4931
step.result.02.aic	27	-1064.7253	-954.3824
step.result.02.bic	12	-1020.6026	-971.5613

```
#Lasso:
loc <- grep("log.per.cap.income",names(cdilogred.cont))
y <- cdilogred.cont[,loc]
X <- apply(as.matrix(cdilogred.cont[,-loc]),2,function(x) rescale(x,"full"))
Xnames <- dimnames(X)[[2]]
lasso.result <- glmnet(X,y)
plot(lasso.result,xvar="lambda",xlim=c(-9,0))
abline(h=0,lty=2)
legend('bottomright',lty=1,col=1:length(Xnames),legend=Xnames,cex=0.75)
```




```
cv.lasso.result <- cv.glmnet(X,y)
plot(cv.lasso.result)
```




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

```
##   lambda.1se   lambda.min
## 0.0078151963 0.0005775994
```

```r
tmp <- cbind(coef(cv.lasso.result,s=cv.lasso.result$lambda.min),
coef(cv.lasso.result,s=cv.lasso.result$lambda.1se)
)
dimnames(tmp)[[2]] <- c("lambda(minMSE)","lambda(minMSE+1se)")
tmp
```

```
## 11 x 2 sparse Matrix of class "dgCMatrix"
##               lambda(minMSE) lambda(minMSE+1se)
## (Intercept)       9.80695459         9.80695459
## log.land.area    -0.06225867        -0.05299404
## pop.18_34        -0.12592460        -0.09508877
## pop.65_plus      -0.02018055         .
## log.doctors       0.11878438         0.13382474
## log.hosp.beds     0.02367462         .
## log.crimes        .                  .
## pct.hs.grad      -0.05845813         .
## pct.bach.deg      0.23405039         0.17274100
## pct.below.pov    -0.22720022        -0.18282023
## pct.unemp         0.04952110         0.02085727
```

```r
#Our final model could be with region interaction factors:
summary(all.subsets.01.final.with.some.region)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + log.hosp.beds +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     region + log.hosp.beds:region + pct.hs.grad:region + pct.bach.deg:region +
##     pct.below.pov:region + pct.unemp:region, data = tmp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22996 -0.04697 -0.00359  0.04498  0.32685
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             10.120027   0.285330  35.468  < 2e-16 ***
## pop.18_34               -0.015138   0.001198 -12.635  < 2e-16 ***
## log.hosp.beds            0.050030   0.010067   4.970 9.81e-07 ***
## pct.hs.grad             -0.006270   0.003446  -1.819 0.069565 .
## pct.bach.deg             0.018593   0.002388   7.786 5.58e-14 ***
## pct.below.pov           -0.024982   0.003874  -6.448 3.16e-10 ***
## pct.unemp                0.017158   0.005156   3.328 0.000953 ***
## regionNE                 0.133176   0.358821   0.371 0.710716
## regionS                 -0.332535   0.314030  -1.059 0.290249
## regionW                  1.482059   0.422614   3.507 0.000503 ***
## log.hosp.beds:regionNE   0.007687   0.014259   0.539 0.590094
## log.hosp.beds:regionS    0.005009   0.012307   0.407 0.684217
## log.hosp.beds:regionW   -0.002427   0.013981  -0.174 0.862266
## pct.hs.grad:regionNE    -0.003748   0.004330  -0.865 0.387313
## pct.hs.grad:regionS      0.005332   0.003866   1.379 0.168608
## pct.hs.grad:regionW     -0.016794   0.004711  -3.565 0.000406 ***
## pct.bach.deg:regionNE    0.007382   0.003117   2.369 0.018308 *
## pct.bach.deg:regionS    -0.002528   0.002677  -0.944 0.345531
```

```
## pct.bach.deg:regionW    0.006398   0.003035   2.108 0.035625 *
## pct.below.pov:regionNE -0.001302   0.005338  -0.244 0.807446
## pct.below.pov:regionS   0.006831   0.004242   1.610 0.108120
## pct.below.pov:regionW  -0.018081   0.005713  -3.165 0.001665 **
## pct.unemp:regionNE     -0.007742   0.007853  -0.986 0.324727
## pct.unemp:regionS      -0.024036   0.006884  -3.491 0.000532 ***
## pct.unemp:regionW      -0.019897   0.007191  -2.767 0.005911 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08141 on 415 degrees of freedom
## Multiple R-squared:  0.8534, Adjusted R-squared:  0.8449
## F-statistic: 100.6 on 24 and 415 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(all.subsets.01.final.with.some.region)
```