Impact of a County's Economic, Health and Social Well-being on Per Capita Income

Lee, Woo Chan
woochanl@andrew.cmu.edu

Department of Statistics and Data Science, Carnegie Mellon University

October 2021

Abstract

We address several questions related to the association between average income per person and a county's economic, health and social well being. We examine data on countys' demographic information (Kutner et al.), using exploratory data analyses and a variety of techniques in linear regression and optimal variable selection. We find that the total crime rate and the region variable are fairly related to per-capita income, and that the best model involves non-collinear significant variables such as the number of doctors, percentage of bachelor degrees, percentage of unemployment as well as some interaction terms with region. Some missing observations in counties and states could be a cause for concern that needs to be addressed further, and it would be worthwhile in the future to explore two-way interaction terms, as well as obtain additional data for cross-validation.

Introduction

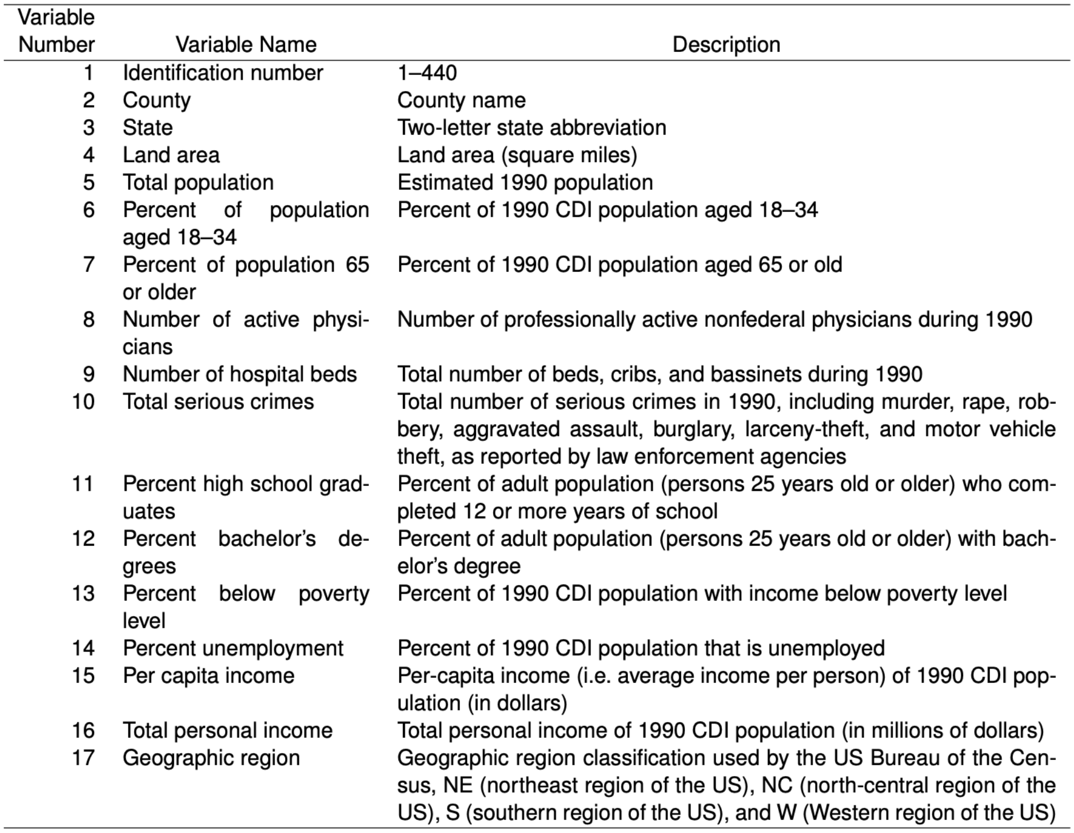
There are numerous indicators that social economists use to measure prosperity and wealth across the world, and one such widely used metric is the Per Capita Income, measuring the average income per person in a given state or region. Income inequality across US counties is a widely known issue (Sommeiller, et al. 2016), and it would be useful to understand the factors that might affect this disparity across the different counties. A county's prosperity can be influenced by various economic, health, and social factors. The goal of this paper is to investigate the relationship between average income per person and variables associated with a county's economic, health and social well-being, as well as find an optimal regression model that can explain the associations.

In particular, we will:

- Explore the relationship between each individual pair of variables
- Examine how crime rates and region affects per-capita income
- Find the best model to predict per-capita income from the full list of variables
- Examine whether the missing states and counties from the data is a cause for concern

Data

The data is taken from Kutner et al. (2005): 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 Table 1. The total number of observations is 440, and there are no observed "NA" values across the dataset.



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 from Kutner et al. (2005)

The summary statistics of the quantitative variables are given in Table 2.

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

Table 2: Summary statistics for quantitative variables

Figure 1 below shows a box plot of per-capita income across the different regions. The median per-capita income across the 4 regions show slight variation.

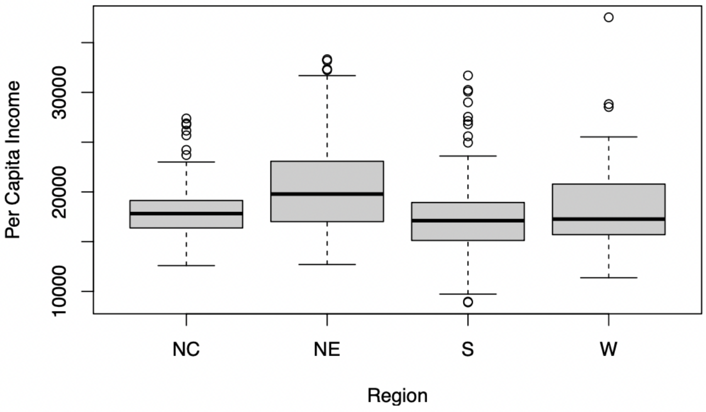


Figure 1: Per-capita income per region

The histogram distributions of each quantitative variable is shown in Figure 2. The figure shows that our response variable *per.cap.income* is slightly skewed to the right, and most of the other predictor variables are severely right-skewed as well.

Out of the 3 categorical variables *county*, *state*, and *region*, we only used the *region* variable. The reason for this was because the combination of *county* and *state* represented one observation of each unique county, adding up to 440, the total number of rows in the dataset. A large number of unique values would not be useful for data analysis, and it was a reasonable decision to leave out *county* and *state* from consideration.

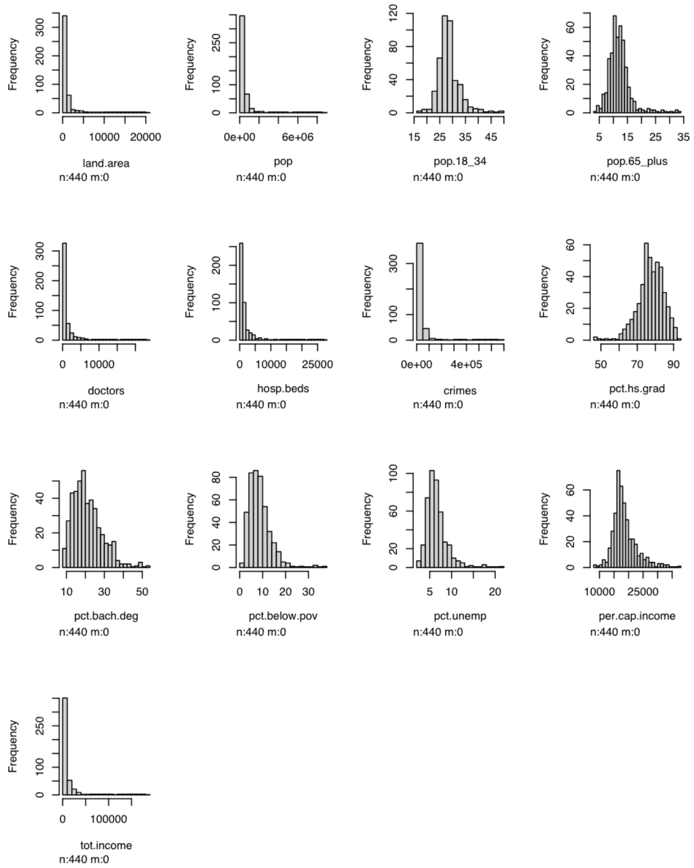


Figure 2: Histogram distributions of quantitative variables

Methods

Below we will outline the methods used for each research questions defined in the introduction section.

1. Relationship between each individual pair of variables

A correlation heatmap was used to explore the correlation between all quantitative variables, and deduce whether multicollinearity was an issue in the dataset. Box plots were also used to determine the relationship between categorical and quantitative variables.

2. Examine how crimes and region affects per-capita income

In order to evaluate the theory that per-capita income is related to crime rate, and that this relationship may vary in different regions of the country, we built regression models to predict *per.cap.income* from *crimes* and *region*. We used logarithmic transformations of the response variable *per.cap.income* and the quantitative predictor *crimes*, and further considered models with the additive *region* variable, as well as the interactions between *crimes* and *region*.

We evaluated the validity of the models through residual diagnostic plots, and assessed the significance of each coefficients in order to come up with an optimal combination of the *crimes* and *region* predictors. F tests (ANOVA) as well as the AIC and BIC values were used to compare the fits of different models.

In addition, we also attempted to replace the *crimes* variable with per-capita crime rate given by *crimes/pop* to observe if there was any change in model fit.

3. Finding the best model to predict per-capita income

The histogram plots for all 13 quantitative variables (including response variable and predictors), were evaluated to assess whether transformations were needed or not. The variables that were highly skewed and needed logarithmic transformations were:

- per.cap.income (Response variable)
- land.area
- pop
- doctors
- hosp.beds
- crimes
- pct.below.pov
- tot.income

Only the logarithmic transformations were used, not only because some of the variables had slight skewdness, but also because logarithmic transformations tend to be easier to interpret in terms of percentage-change concepts. Considering the audience of this analysis, the more untransformed the variables are, the easier it will be to comprehend about the models presented in this report.

Also note that the predictor variables *log.pop* and *log.tot.income* were dropped from the analysis, since our response variable *log.per.cap.income* is a deterministic function of both predictors. More specifically, *per.cap.income = tot.income/pop*.

We also looked at the Variance Inflation Factors (VIF) for each of the predictors to assess the severity of multicollinearity when all predictors were considered. Consequently, variable selection methods such as all-subsets, stepwise and LASSO regression were used to choose the optimal subset of quantitative variables that produced the best fitting model. The ”best” model was one that satisfied key modeling assumptions as well as being interpretable in the context of social science and economics.

The categorical variable *region* was later added back to determine its significance in predicting per-capita income. Both additive terms and interactions terms were considered. If any coefficient indicator for *region* or its interaction terms seemed important, then we chose to keep the whole set of interaction variables.

Finally, F tests (ANOVA) and AIC / BIC values were used to compare the fits of models with different variable subsets. We evaluated the validity of the models through residual diagnostic plots, and assessed the significance of each coefficients through model summaries.

4. Addressing the missing counties and states

We used simple exploratory data analysis on the county, state and region variables to find out whether the missing observations would cause any problems. Exploring the difference in composition of these variables as well as some intelligent conjecture would be able to help address this problem further.

Results

Below are the results for each of the research questions defined in the introduction section.

1. Relationship between each individual pair of variables

The correlation matrix heatmap on Figure 3 below suggests that:

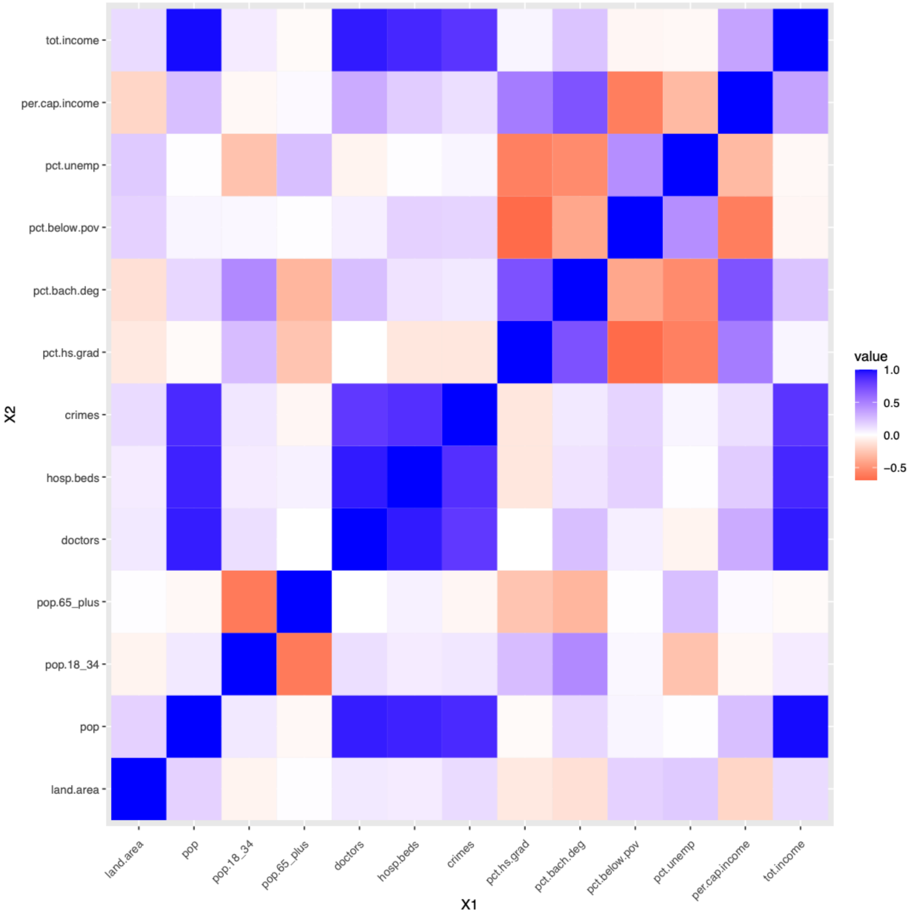


Figure 3: Correlation matrix heatmap

- *tot.income* and *pop* are highly correlated. This is expected because the response variable *per.cap.income* is a deterministic function of *pop* and *tot.income*.
- both *tot.income* and *pop* are also highly correlated with *crimes*, *hosp.beds* and *doctors*
- *pct.hs.grad* and *pct.bach.deg* have moderately high correlation, and this is expected because a person is more likely to hold a bachelor's degree if he/she also graduated from high school.
- Although not a very strong correlation, *pct.hs.grad* and *pct.bach.deg* are negatively correlated to *pct.unemp*, which makes sense because people who graduated from high school as well as those who hold a bachelor's degree are less likely to be unemployed.
- *hosp.beds* and *doctors* are strongly correlated with one another. This is expected because the more doctors / physicians you have in a county, the more hospital beds you would expect to see.

These observations indicate that multicollinarity might be a problem that we would need to address further.

In Figure 2, we looked at the boxplot of per-capita income across each region, and noticed that the median and inter-quartile range of per-capita income was fairly different across the 4 regions. This suggested that the categorical variable *region* could potentially be important in predicting per-capita income.

2. Examine how crimes and region affects per-capita income

We considered a total of 3 regression models using the log-transformed variables *log.per.cap.income* and *log.crimes*, and the interactions with the *region* variable (details in page 7 and 8 in Technical Appendix).

2.1 Base model with only crimes variable

The base regression model involving *log.per.cap.income* and *log.crimes* had the estimated regression coefficients,

$$log.per.cap.income = 0.054 \cdot log.crimes + 9.29 \tag{1}$$

As seen in page 8 of the Technical Appendix, the coefficient for *log.crimes* was statistically significant with a low p value, and the R squared value was 0.079. A unit percentage increase in total crimes led to roughly a 0.054% increase in per-capita income.

2.2 Model with additive region variable

The regression model involving *log.per.cap.income*, *log.crimes*, and the additive *region* variable had the estimated regression coefficients,

$$log.per.cap.income = 0.067 \cdot log.crimes + 0.1 \cdot regionNE - 0.09 \cdot regionS - 0.06 \cdot regionW + 9.19 \tag{2}$$

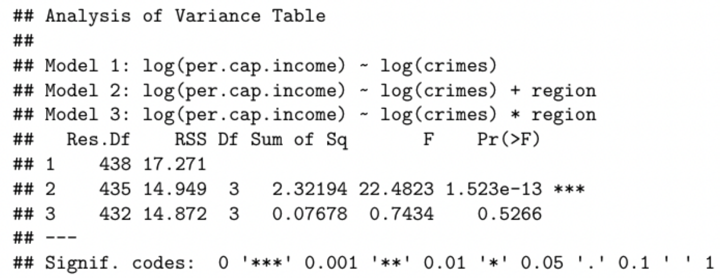
Page 8 of the Technical Appendix shows us that all of the coefficients were statistically significant with low enough p values, and the R squared value increased significantly to 0.2032. A unit percentage increase in total crimes led to roughly a 0.067% increase in per-capita income.

2.3 Model with additive region and interaction terms

Our third regression model added interaction terms between the variables *region* and *log.crimes*. Page 8 and 9 of the Technical Appendix shows that only the coefficient for *log.crimes* was statistically significant, while all other variables had high p values. The R squared value was roughly similar to our previous additive model with a value of 0.2073. A unit percentage increase in total crimes led to roughly a 0.051% increase in per-capita income.

The residual diagnostic plots in page 9 and 10 of the Technical Appendix suggested that all models were fairly valid, conforming to the key assumptions of linear regression. But they also had some minor limitations such as non-normality of the residuals and some influential points that needed further inspection.

Introducing the *region* variable in the second and third model significantly increased the R squared value, suggesting that *region* would be an important variable to keep. In Table 3, the F test was performed on the three models to justify whether the interaction terms were effective or not. The second model with additive *region*, but without interactions turned out to be doing the best with a very low 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
```

Table 3: F test (ANOVA)

Overall, the three models suggested that there is a positive correlation between per-capita income and crime rate.

2.4 Per-capita crimes

We also substituted the *log.crimes* variable with log(per-capita crimes) and fit the three identical regression models in 2.1, 2.2 and 2.3. Page 11 and 12 in the Technical Appendix shows that the R squared values for all three models decreased significantly and the residual standard errors showed an overall increase. Further, the *log.per.capita.crimes* predictor was no longer as significant in the new three models as *log.crimes* was in the original three models.

The residual diagnostic plots in page 13 and 14 of the Technical Appendix showed little differences compared to the original three models, suggesting that the new models with per-capita crimes were fairly valid as well.

We also used AIC and BIC values to compare between each of the second models (additive *region* without interactions) using raw *log.crime* and *log.per.capita.crime*. The results in page 15 of the Technical Appendix show that the model with raw *crimes* had smaller AIC and BIC values (smaller the better), therefore suggesting that transforming *crimes* into *per-capita crimes* was not a good idea.

3. Finding the best model to predict per-capita income

Figure 4 shows the histogram plots for quantitative variables after the logarithmic transformations were applied. It can be observed that a lot of the skewness have improved. Further, the correlation heatmap after logarithmic transformation shown in Page 17 of the Technical Appendix, suggested that the correlations between transformed variables remained relatively similar, but a bit stronger.

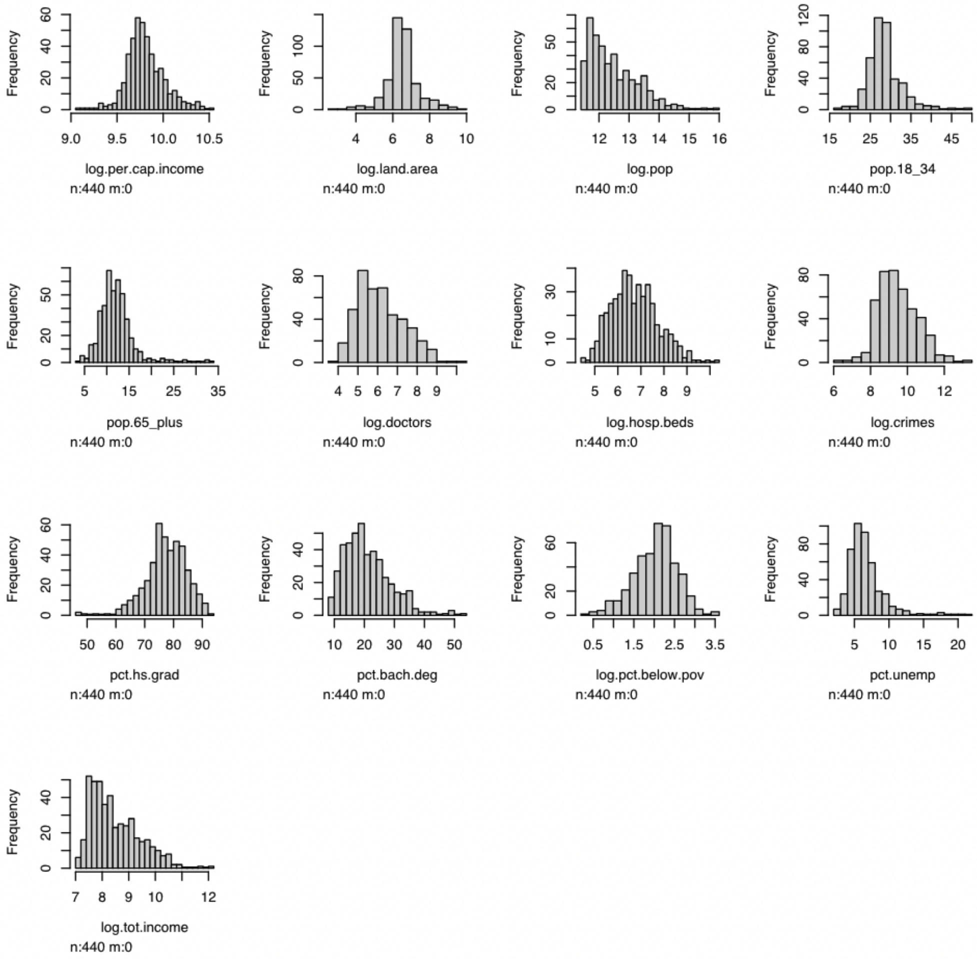


Figure 4: Histogram distributions of quantitative variables after transformation

To start off, we fitted a full model including all the quantitative variables plus the region categorical variable (note that *id*, *county*, *state*, *pop* and *tot.income* were excluded). Pages 17 and 18 of the Technical Appendix shows the coefficient summary of this full model. The resulting R squared value was 0.8394, meaning that 83.94% of the total variability of the response variable was explained by the model. Some predictors were statistically significant with low p values, while some were not. Further, predictors like *pct.hs.grad* and *pct.unemp* even seemed to have the wrong coefficients with opposite signs.

Table 4 shows the VIF for each predictor variables. The full model suffered from multicollinearity, where some predictors had large VIFs that exceeded 5. In particular, *log.doctors* and *log.hosp.beds* had VIFs of 15.3 and 12.1 respectively. *log.crimes* also turned out to have a high VIF of 6.24. The problem of multicollinearity suggested that optimal variable selection methods were needed.

```
    log.land.area         pop.18_34       pop.65_plus       log.doctors
           1.4826            2.2287            2.0155           15.3220
    log.hosp.beds        log.crimes       pct.hs.grad      pct.bach.deg
          12.0950            6.2380            3.9041            4.3764
log.pct.below.pov         pct.unemp          regionNE           regionS
           3.0729            2.0074            1.9315            2.2908
          regionW
           2.3844
```

Table 4: VIF for each predictor variables

3.1 Variable selection - All Subsets

We will now look at the results for the all-subsets variable selection method. The all-subsets method eventually chose 6 variables that gave the lowest BIC value of -747.68 (more details on page 18, 19, and 20 of the Technical Appendix).

In Table 5, we can see the coefficient summary of the model using the 6 variables chosen. The R squared value turned out to be 0.834, and all coefficients were statistically significant with low p values. But at the same time, the coefficient estimates seemed to be quite small in magnitude.

```
                      Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)       10.095545110 0.0516863528 195.323225 0.000000e+00
log.land.area     -0.036212594 0.0048198661  -7.513195 3.336311e-13
pop.18_34         -0.012026824 0.0011883729 -10.120413 9.445833e-22
log.doctors        0.067772351 0.0041934269  16.161567 2.586913e-46
pct.bach.deg       0.010523423 0.0008860605  11.876641 2.352204e-28
log.pct.below.pov -0.197474797 0.0100936858 -19.564191 1.560595e-61
pct.unemp          0.008109587 0.0021194105   3.826341 1.492036e-04
```

Table 5: Coefficient summary for all-subsets

In page 20 of the Technical Appendix, we can see that none of the chosen 6 predictors had an excessively large VIF, signaling that multicollinearity was no longer an issue.

The residual diagnostic plots for the all-subsets model in page 21 of the Technical Appendix suggested that the model was valid, except for a minor limitation that the left and right tails of the Normal Q-Q plot were a little bit heavy.

The standardized residual plots against each of the predictors in page 21 of the Technical Appendix showed that the residuals for all plots were relatively randomly scattered, further suggesting the validity of the model.

The added variable plots and marginal plots is shown in page 22 and 23 of the Technical Appendix respectively. And they both further add to the fact that the chosen predictors were appropriate and that the model was valid.

3.2 Variable selection - stepwise BIC

We now look at the results for the stepwise variable selection method. Note that BIC is the information criteria used.

The selection procedure as well as the BIC value at each step can be seen in page 24 and 25 of the Technical Appendix. Table 6 shows the coefficient summary of the variables chosen by the stepwise method.

```
(Intercept)        10.0955451  0.0516864 195.323  < 2e-16 ***
log.land.area      -0.0362126  0.0048199  -7.513 3.34e-13 ***
pop.18_34          -0.0120268  0.0011884 -10.120  < 2e-16 ***
log.doctors         0.0677724  0.0041934  16.162  < 2e-16 ***
pct.bach.deg        0.0105234  0.0008861  11.877  < 2e-16 ***
log.pct.below.pov -0.1974748  0.0100937 -19.564  < 2e-16 ***
pct.unemp           0.0081096  0.0021194   3.826 0.000149 ***
```

Table 6: Coefficient summary for stepwise

We can see that stepwise method chose the same subset of variables as the all-subsets method did. All predictors had coefficients that were statistically significant with low p values.

3.3 Variable selection - LASSO regression

We now look at the results for the LASSO regression (more details in page 27, 28 and 29 of the Technical Appendix). Note that we chose to use the λ value of *lambda.1se*, which is 1 standard deviation larger than the best λ value found through cross-validation, since it could protect against capitalization on chance.

Table 7 shows the variable subset chosen by LASSO regression and their coefficient summaries. All variables except *pop.65_plus* had statistically significant coefficients with low p values. The full summary table from Page 30 of the Technical Appendix shows that the R squared value was 0.829, which was not too different from the all-subsets and stepwise models.

```
(Intercept)        10.1116202  0.0628428 160.903  < 2e-16 ***
log.land.area      -0.0337191  0.0049006  -6.881 2.10e-11 ***
pop.18_34          -0.0116315  0.0014289  -8.140 4.24e-15 ***
pop.65_plus         0.0014938  0.0013571   1.101    0.272
log.doctors         0.0673303  0.0043416  15.508  < 2e-16 ***
pct.bach.deg        0.0096519  0.0008668  11.135  < 2e-16 ***
log.pct.below.pov -0.1911802  0.0100990 -18.931  < 2e-16 ***
```

Table 7: Coefficient summary for LASSO

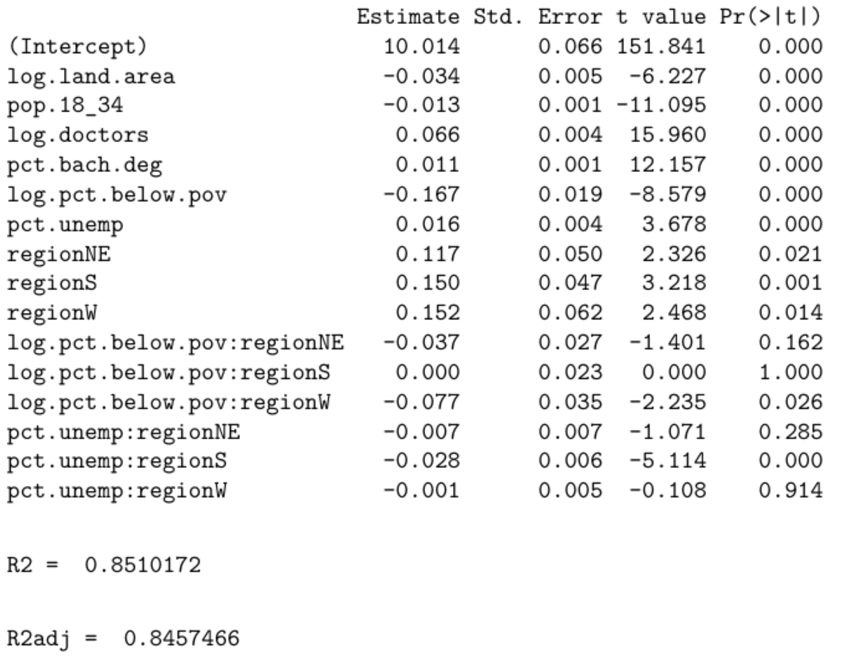
The only difference between the all-subsets method and LASSO regression was that all-subsets model chose to include *pct.unemp* rather than *pop.65_plus* which LASSO regression did. Page 31 of the Technical Appendix shows the F test (ANOVA) result suggesting that the all-subsets model with *pct.unemp* turned out to be a better choice.

3.4 Adding back the region variable

After figuring out the optimal subset of variables through the chosen variable selection methodologies, the categorical variable *region* was brought back to be considered as additive and interaction terms.

Page 31 and 32 of the Technical Appendix shows the coefficient summary of adding the additive *region* variable as well as all possible interaction terms with the existing quantitative variables. We decided to keep the entire group of interaction terms if any of the indicators for the specific categorical variable was statistically significant. If none were significant, we dropped the whole group of interactions.

Table 8 below shows the resulting model with the added *region* variables and selected interaction terms. All the main effects and interaction terms that involve *region* have at least one significant term. The R squared value slightly increased to 0.851.



```
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    10.014      0.066 151.841    0.000
log.land.area                  -0.034      0.005  -6.227    0.000
pop.18_34                      -0.013      0.001 -11.095    0.000
log.doctors                     0.066      0.004  15.960    0.000
pct.bach.deg                    0.011      0.001  12.157    0.000
log.pct.below.pov              -0.167      0.019  -8.579    0.000
pct.unemp                       0.016      0.004   3.678    0.000
regionNE                        0.117      0.050   2.326    0.021
regionS                         0.150      0.047   3.218    0.001
regionW                         0.152      0.062   2.468    0.014
log.pct.below.pov:regionNE     -0.037      0.027  -1.401    0.162
log.pct.below.pov:regionS       0.000      0.023   0.000    1.000
log.pct.below.pov:regionW      -0.077      0.035  -2.235    0.026
pct.unemp:regionNE             -0.007      0.007  -1.071    0.285
pct.unemp:regionS              -0.028      0.006  -5.114    0.000
pct.unemp:regionW              -0.001      0.005  -0.108    0.914

R2 =  0.8510172

R2adj =  0.8457466
```

Table 8: Additive region and some interaction terms added to all-subsets model

To further justify the use of the *region* variable and the chosen interaction terms, the F test (ANOVA) was performed on three models (details in page 33 and 34 of the Technical Appendix). The model in Table 8 was statistically significant with a very low p value, and was better than the base all-subsets model as well as the all-subsets model with only the additive *region* variable.

Furthermore, in page 34 of the Technical Appendix, the model with some chosen interaction terms turned out to have lower AIC values than the base all-subsets model, while the BIC value was opposite in result. This is interpretable since BIC tends to favor simpler models , while AIC favors more complex models in theory.

3.5 Interpreting the final model

The chosen final model was the all-subsets model with some chosen interaction terms as shown previously in Table 8. Although more complex than the base model, the categorical variable *region* was proven to be fairly important as we saw when interpreting Figure 2. The interaction terms were also not too difficult to interpret, given that they simply indicate the quantitative variables interacting with different parts of the region (NC, NE, S, W). The F test and AIC values also suggested that the interaction terms were valuable when predicting per-capita income.

Along with the final model's coefficient summary in Table 8, we also looked at diagnostic plots produced in Figure 5 below, and saw that the model was fairly valid since it conformed to the key assumption of constant error variance, but had heavy right and left tails similar to the base all-subsets model. There was also a specific high leverage outlier point that needed further inspection.

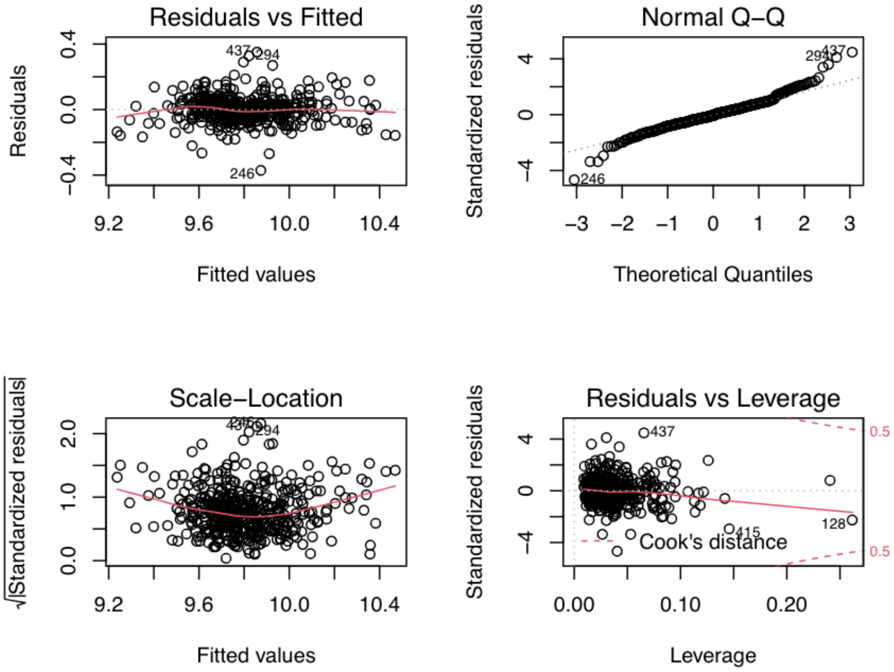


Figure 5: Residual diagnostic plots for final model

We also looked at the plot of Y (*log.per.cap.income*) vs the fitted values $\hat{Y}$ produced in Figure 6. We could see that the straight-line fit to this plot (displayed as a dashed line) provides a fairly good fit, although not perfect. This further suggests that the model is valid.

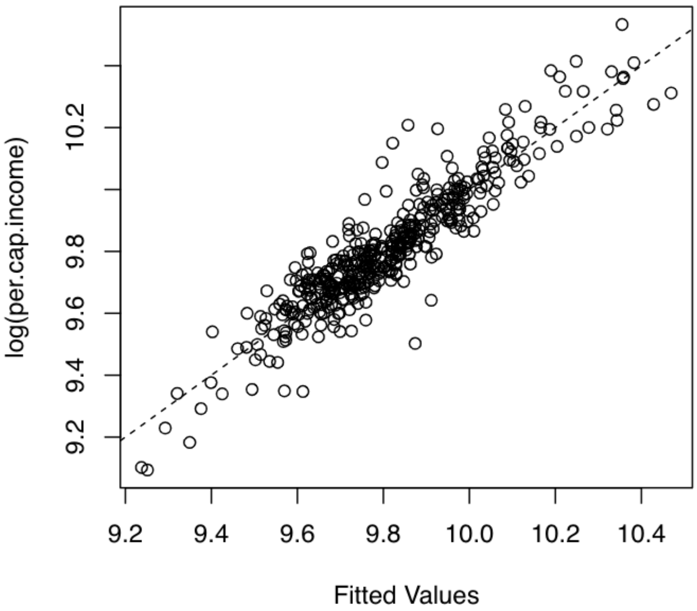


Figure 6: Plot of Y vs $\hat{Y}$

Finally, we attempted to interpret some of the resulting coefficients of the final model in Table 8:

- For every 1% increase in a county's land area, there is a 0.03% decrease in expected per-capita income. (This could be due to an urban-rural contrast: rural counties tend to be bigger than urban counties).
- 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. (This could be because 18–34 year olds are not at peak earning capacity yet and so their lower incomes drags down the per-capita income).
- For every 1% increase in the number of doctors in a county, the expected per-capita income increases by about 0.06% (This could be because doctors are well-paid and could be big contributors to the per-capita income).
- For every 1 percentage point increase in the percent of people with bachelors degree, there is an expected 1% increase in per-capita income (This could be because a bachelor's degree can make a person more employable, thereby increasing earning potential and contribution to per-capita income).
- In the main effect for region, and in several of the interactions for region, each of the 4 regions show slight variation, but does not show significant deviation from each other.

4. Addressing the missing counties and states

It can be argued that the missing states and counties can be a cause for concern as our data may not be fully representative of the whole population. In our data, only 440 counties (including those that have duplicate county names in different states) were considered out of approximately 3,000 counties in the US. Furthermore, the frequency table for the state variable in page 1 of the technical appendix suggests that 3 states (Alaska, Iowa and Wyoming) were excluded from the dataset.

Upon looking at some online resources on county population [4], we found that the top 100 most populous counties were included as a subset in our data. Even though there was no clear statement on how the dataset was sampled, we could hypothesize that Kutner simply chose to include the 440 most populous counties.

Page 1 of the technical appendix also shows the frequency table of the region variable. Most counties turned out to be in the South region, while the least were in the West. There could have been lack of sampling in the West and over-sampling in the South. Another reason for this could be that the land areas of counties are just larger in the West (fewer counties to sample from), while the land area of counties are smaller in the South (more counties that cover similar land areas). But overall, the imbalance in samples between the 4 region categories was not too large.

Given the fact that region, state and county are all categorical variables of similar nature but at different scale levels, it can be argued that the more aggregated region variable is already a fairly good predictor with relatively balanced sample sizes. Thus, including more granular variables like county and state would be unnecessary. Our main argument would be that the missing observations would not be a big cause for concern when trying to obtain a general big picture of predictors that affect per-capita income.

Discussion

Below are the recap of our findings for each of the research questions.

1. Relationship between each individual pair of variables

The correlation heatmap suggested that there were several variables that were correlated, hinting a potential problem of multicollinearity. We also found that per-capita income was quite varied across the 4 different regions defined by the *region* variable.

2. Examine how crimes and region affects per-capita income

In order to assess the theory that per-capita income is related to crime rate, and that this relationship may vary in different regions of the country, we looked at 3 different models including additive and interaction terms with the *region* variable.

The coefficient summaries of the three models indicated a positive correlation between per-capita income and total crimes. The log transformations on both variables allowed us to interpret that a unit percentage increase in total crime led to roughly a 0.06% increase in per-capita income. All three models were also fairly valid, but in the end, the second model (additive region variable without interactions) turned out to be the best model.

We also attempted to see if changing the crimes variable to per-capita crimes helped in any way. This would allow per-capita crimes to be in the same comparable scale as per-capita income, thereby leading to better interpretability. However, this resulted in a significant decrease in R squared value, as well as an increase in the AIC and BIC values, suggesting that the trade-off between interpretability and model fit was not equal enough to justify using *per-capita crimes* in place of the raw *crimes*.

Lastly, the interaction terms in the models did not turn out to be significant. This would mean that the relationship between per-capita income and crimes did not vary significantly in different regions of the country. However, the additive region variables were individually significant enough to be valuable in the model given none of the interaction terms were involved.

3. Finding the best model to predict per-capita income

Logarithmic transformations on certain variables improved a lot of the problems of skewness, while keeping correlations between variables roughly unchanged. We also attempted to maintain easy interpretability of the variables by only applying logarithmic transformations when absolutely needed, while keeping as many untransformed variables as possible. This facilitates explaining the models to anyone who is interested in the social science and economics field but less knowledgeable about technical matters.

However, the problem of multicollinearity remained, and we used three different variable selection methodologies - all subsets, stepwise, and LASSO regression - to counter this. All three methods produced similar optimal subsets of significant variables that generated a minimum value of BIC. But through the F-test of overall significance, we were able to find out that the variable subset chosen by all-subset and stepwise regression produced the best model.

All three methods chose to exclude *pop.65_plus*, since it would probably have been highly correlated to its counterpart variable *pop.18_34*. It was also understandable that *log.hosp.beds* was eliminated in all three methods due to its high collinearity with *log.doctors*. The one variable that was unexpectedly eliminated from all three methods was *log.crimes*, because in our second research question we saw that the variable was pretty significant in predicting per-capita income. This implies that when other variables are involved, *log.crimes* becomes relatively insignificant.

The final best model was chosen after adding back the region variable, as it was previously hypothesized to be an important indicator of per-capita income. With the additive term and some chosen interaction terms added in, the final model gave an R squared value of 0.851, meaning that roughly 85.1% of the total variance of per-capita income was explained by the model.

Noticing the improvement in model fit through not only the R squared value but also the AIC and BIC values, we were able to deduce that keeping some interaction terms were justifiable. We figured that the resulting trade-off of added complexity did not severely impact the interpretability of the model, since the interaction terms simply corresponded to the differing relationship between the quantitative variables and per-capita income in different regions.

The final model turned out to be moderately parsimonious, and most of the estimated coefficients, except for *pct.unemp* had expected signs (plus / minus).

4. Addressing the missing counties and states

In our exploratory data analysis, we found that 3 states (Alaska, Iowa and Wyoming) were excluded from the observations, and that only 440 counties out of approximately 3,000 total counties in the US were included. Through preliminary research, it was hypothesized that the 440 observations were the top 440 most populous counties in the US.

Because the region variable captured a more aggregated view of counties and states (of similar categorical nature), our analysis assumed that the missing values were not a big cause for concern. The key argument here was that the missing observations would be unnecessary when trying to obtain a big picture on variables that affect per-capita income.

Limitations and future work

An opposing argument that can be made to Research Question 4 is that the missing states and counties could in fact contain important information, thus making our entire analysis biased, and create an unrealistic picture of variables that affect per-capita income. Since we are only working with a certain sample of a population, there is always the possibility that the data is biased and not representative of the entire 3,000 counties of population.

However, this problem would be mitigated if the 440 samples in the dataset were selected randomly. If given additional time, we could possibly investigate further on how exactly the CDI data (Kutner et al. (2005)) was collected, focusing on the sampling methods for the selected counties. It would also be wise to compare the summary statistics of the given dataset with that of the overall population to see if there are any large deviations in results.

Another evident limitation in a lot of the models explored in this analysis was that the residual diagnostic plots were never perfect. The slight curves in the center of the residual plots as well as the heavy right and left tails of the Q-Q plot suggested that further improvements in the model can be made. In the future, we would look into the two-way interaction terms between quantitative variables and consider more complex models that can improve the validity of the model. The usefulness of interaction terms with the region variable is an evidence that there could be unidentified interaction terms with the potential to enhance the model.

If given additional time, we could also further explore the *state* variable, since 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. However, states have perfect collinearity with region, and if we were to use *state* as a categorical variable in our models, it would only make sense to exclude the *region* variable.

Finally, it would be useful to have additional observations to use as test sets to compare some of the models we found. We are using reasonable methods for variable selection, but since our entire dataset is in fact our training sample, there is a big possibility for overfitting noise in the data. If we were able to cross-validate on some new data, we might be able to better compare and determine the best models in terms of prediction error.

References

[1] Sommeiller, et al. (2016), ”Income inequality in the U.S. by state, metropolitan area, and county”, Economic Policy Institute, https://www.epi.org/publication/income-inequality-in-the-us/

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

[3] Sheather, S. J. (2009). A modern approach to regression with R. (Springer eBooks.)

[4] List of the most populous counties in the United States . (2021) . Wikipedia . https://en.wikipedia.org/wiki/List_of_the_most_populous_counties_in_the_United_States

Technical Appendix

Lee, Woo Chan

10/15/2021

Research question 1

Below are the summary statistics for all continuous variables in the dataset.

```
# Summary statistics of continuous variables
cdi_cat <- cdi %>%
  dplyr::select(state, region, county)


cdi_con <- cdi[,-c(1,2,3,17)]## get rid of id, county, state and (for now)
apply(cdi_con,2,function(x) c(summary(x),SD=sd(x))) %>%
  as.data.frame %>% t() %>%
  round(digits=2) %>%
  kbl(booktabs=T,caption=" ") %>%
  kable_classic()
```

```
#summary(cdi_con)
```

Below is the summary statistics for the categorical variable *region*.

```
table(cdi$region)
```

```
##
##  NC  NE   S   W
## 108 103 152  77
```

Below is the frequency table for the *state* variable.

```
table(cdi$state)
```

```
##
## AL AR AZ CA CO CT DC DE FL GA HI ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC
##  7  2  5 34  9  8  1  2 29  9  3  1 17 14  4  3  9 11 10  5 18  7  8  3  1 18
## ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV
##  1  3  4 18  2  2 22 24  4  6 29  3 11  1  8 28  4  9  1 10 11  1
```

Below is the table of region vs state.

Table 1:

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

```
table(cdi$state, cdi$region)
```

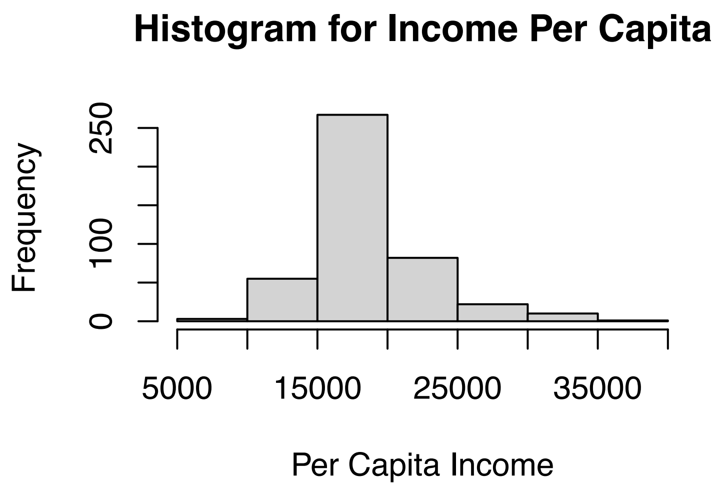
```
##
##      NC NE  S  W
##   AL  0  0  7  0
##   AR  0  0  2  0
##   AZ  0  0  0  5
##   CA  0  0  0 34
##   CO  0  0  0  9
##   CT  0  8  0  0
##   DC  0  0  1  0
##   DE  0  2  0  0
##   FL  0  0 29  0
##   GA  0  0  9  0
##   HI  0  0  0  3
##   ID  0  0  0  1
##   IL 17  0  0  0
##   IN 14  0  0  0
##   KS  4  0  0  0
##   KY  0  0  3  0
##   LA  0  0  9  0
##   MA  0 11  0  0
##   MD  0  0 10  0
##   ME  0  5  0  0
##   MI 18  0  0  0
##   MN  7  0  0  0
##   MO  8  0  0  0
##   MS  0  0  3  0
##   MT  0  0  0  1
##   NC  0  0 18  0
##   ND  1  0  0  0
##   NE  3  0  0  0
##   NH  0  4  0  0
##   NJ  0 18  0  0
##   NM  0  0  0  2
##   NV  0  0  0  2
##   NY  0 22  0  0
##   OH 24  0  0  0
##   OK  0  0  4  0
##   OR  0  0  0  6
##   PA  0 29  0  0
##   RI  0  3  0  0
##   SC  0  0 11  0
##   SD  1  0  0  0
##   TN  0  0  8  0
##   TX  0  0 28  0
##   UT  0  0  0  4
##   VA  0  0  9  0
##   VT  0  1  0  0
##   WA  0  0  0 10
##   WI 11  0  0  0
##   WV  0  0  1  0
```

The table below indicates that there are no observed "NA" values in any of the columns. This is because the data was cleaned beforehand by the instructor.

```
# Find NA values
contains_any_na <- sapply(cdi, function(x) any(is.na(x)))
print(contains_any_na)
```

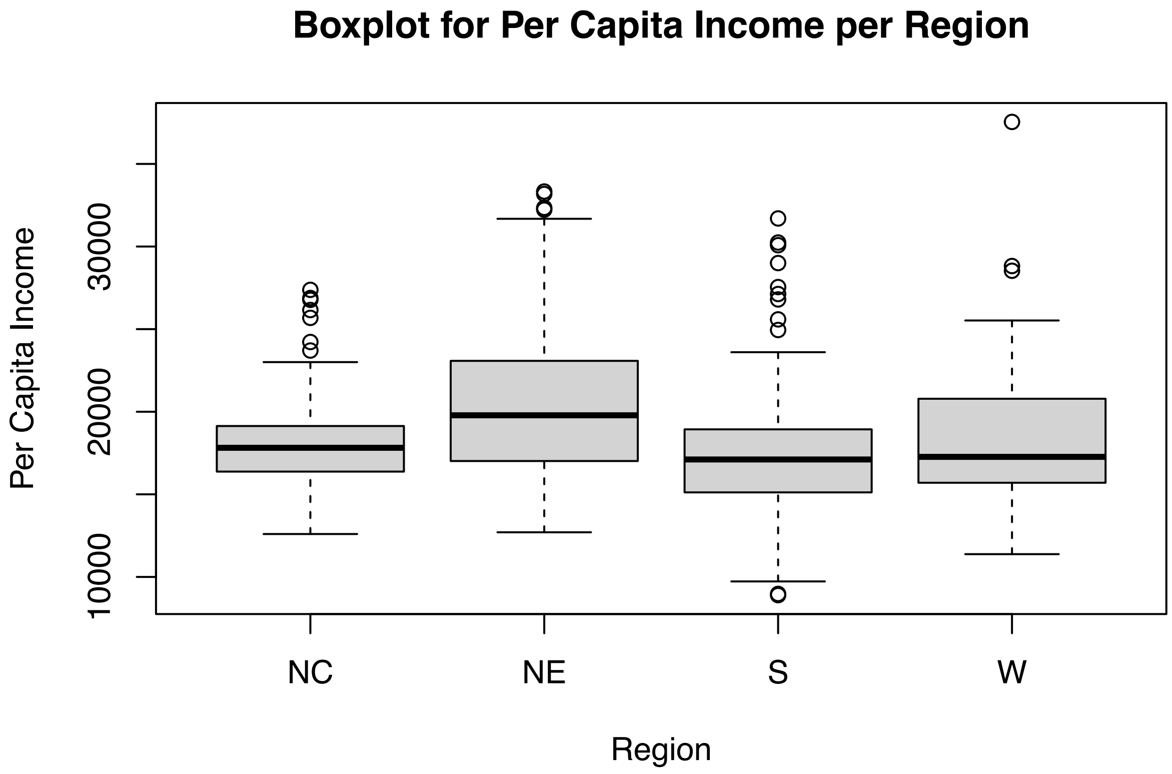
```
##             id         county          state      land.area            pop
##          FALSE          FALSE          FALSE          FALSE          FALSE
##      pop.18_34    pop.65_plus        doctors      hosp.beds         crimes
##          FALSE          FALSE          FALSE          FALSE          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
```

From the histogram below, we can see that our response variable *per.cap.income* is a little bit skewed to the right, but still relatively normally distributed.

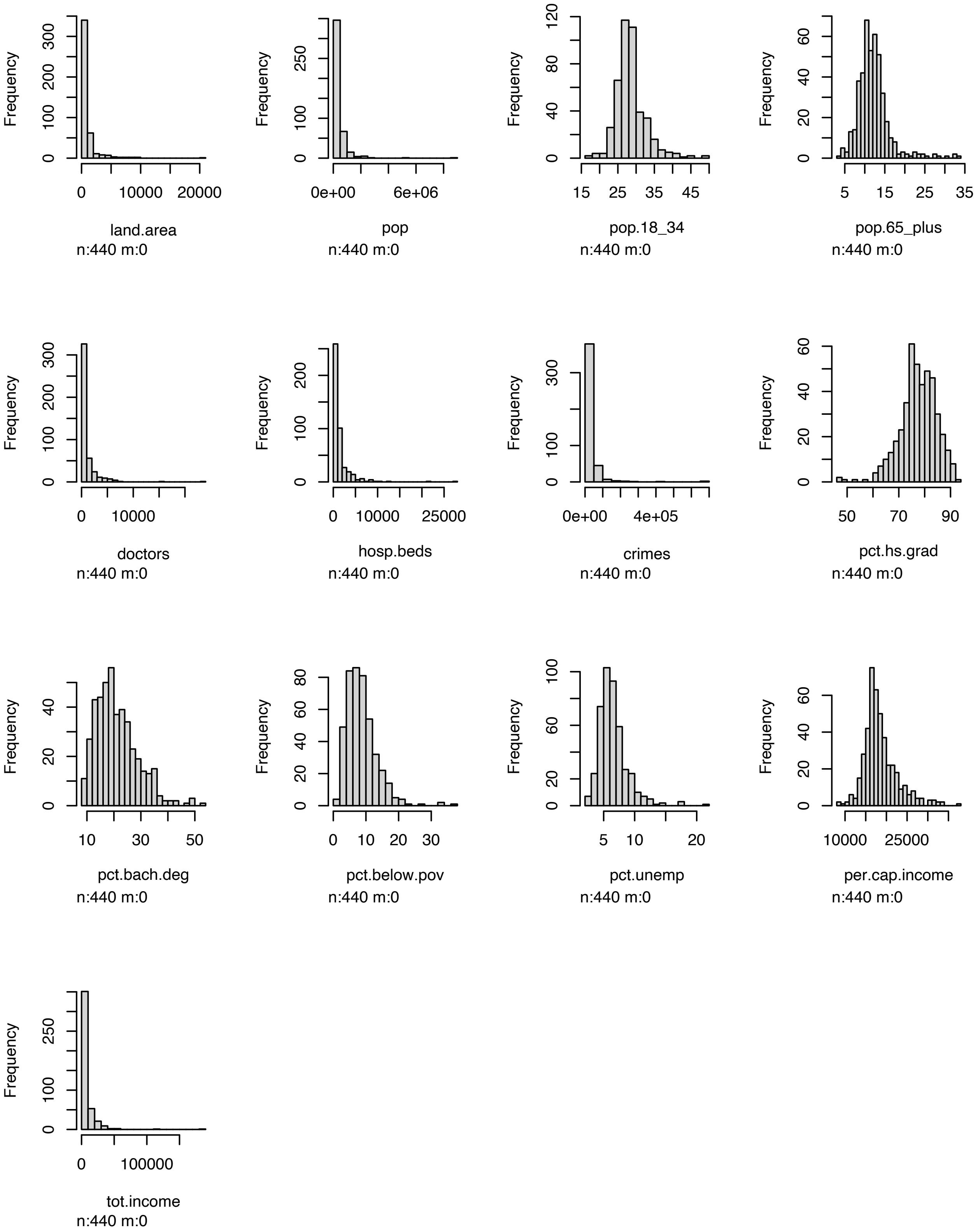


Below shows the box plot of per capita income in the 4 regions. The median from the North East region is the highest, and also has the largest Interquartile Range. Overall, there seems to be some difference in median per capita income between all 4 regions.

```
cdi$region <- factor(cdi$region)
boxplot(cdi$per.cap.income ~ cdi$region, ylab = "Per Capita Income", xlab="Region", main="Boxplot for Pe
```

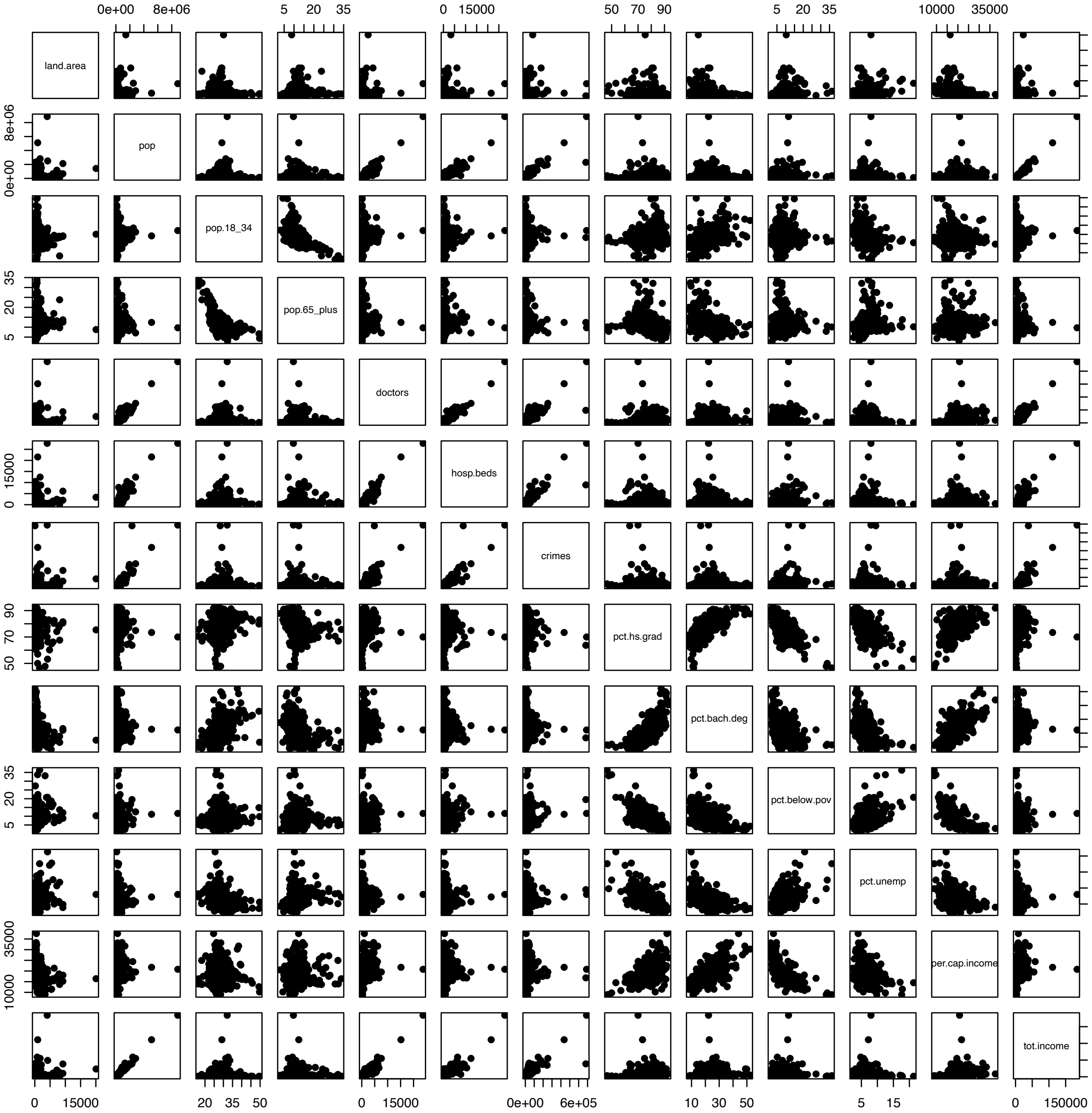


We can also look below at the histogram distribution of other predictor variables. We can observe that there are severely skewed variables like *land.area*, *pop*, *doctors*, *hosp.beds*, *crimes* and *tot.income*.



/

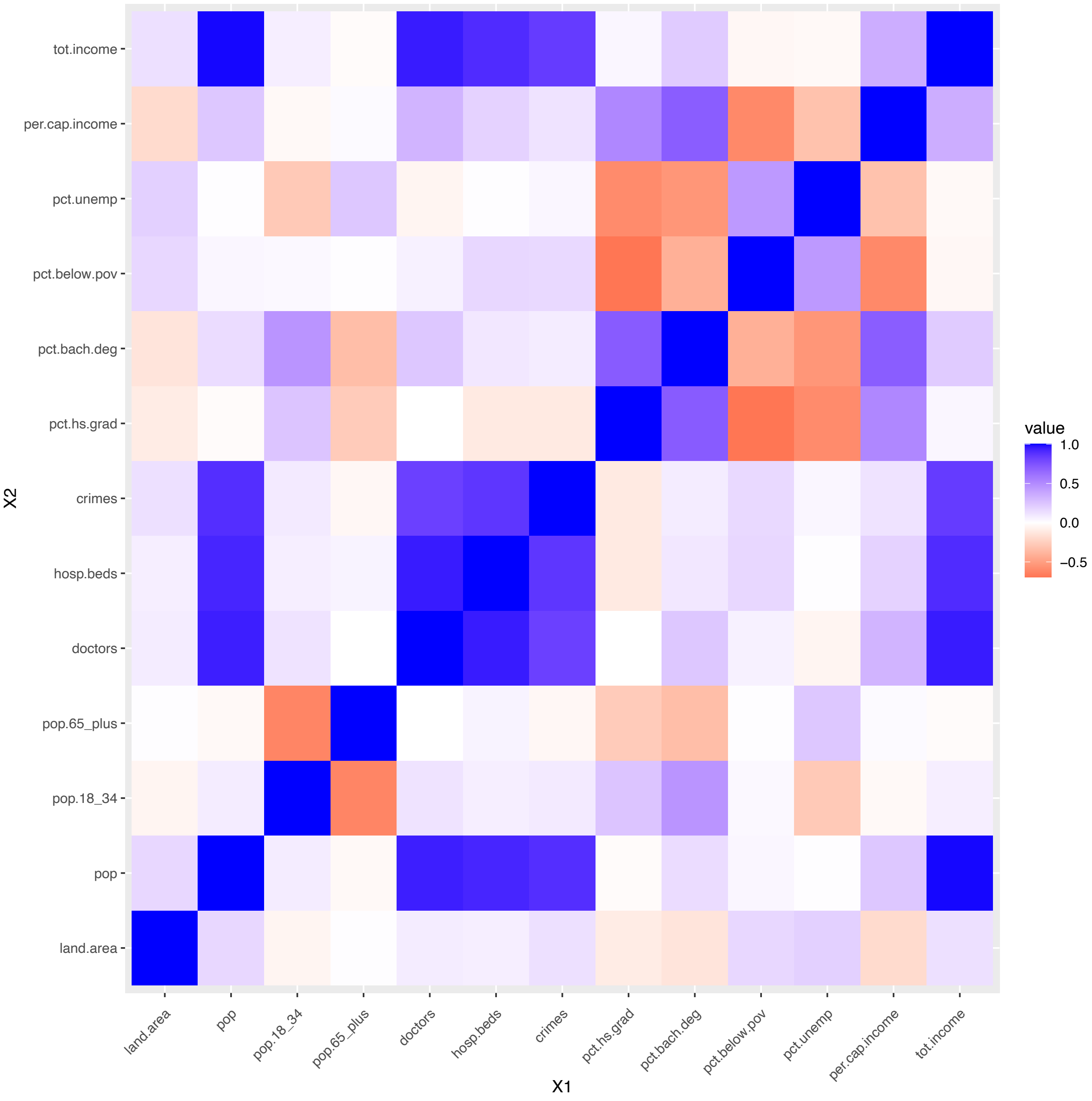
Below is a scatter plot matrix to identify overall relationships between the variables.



Below is the correlation matrix heatmap to understand if there are any correlations between the predictors.The observations suggest that we may run into multi-collinearity problems when we start fitting models.

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE

## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE
```



Research question 2

Interaction term vs additive term (region variable)

I fit below the ordinary model, the additive model and the interaction model with *crimes* and *region*.

```
income_fit1 <- lm(log(per.cap.income) ~ log(crimes), cdi)
income_fit2 <- lm(log(per.cap.income) ~ log(crimes) + region, cdi)
income_fit3 <- lm(log(per.cap.income) ~ log(crimes)*region, cdi )
```

Below is the summary of the 3 models.

```
summary(income_fit1)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(crimes), data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75042 -0.11569 -0.02976  0.09597  0.74498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.295146   0.083764  110.97  < 2e-16 ***
## log(crimes) 0.053858   0.008758    6.15 1.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1986 on 438 degrees of freedom
## Multiple R-squared:  0.07948,    Adjusted R-squared:  0.07738
## F-statistic: 37.82 on 1 and 438 DF,  p-value: 1.752e-09
```

```
summary(income_fit2)
```

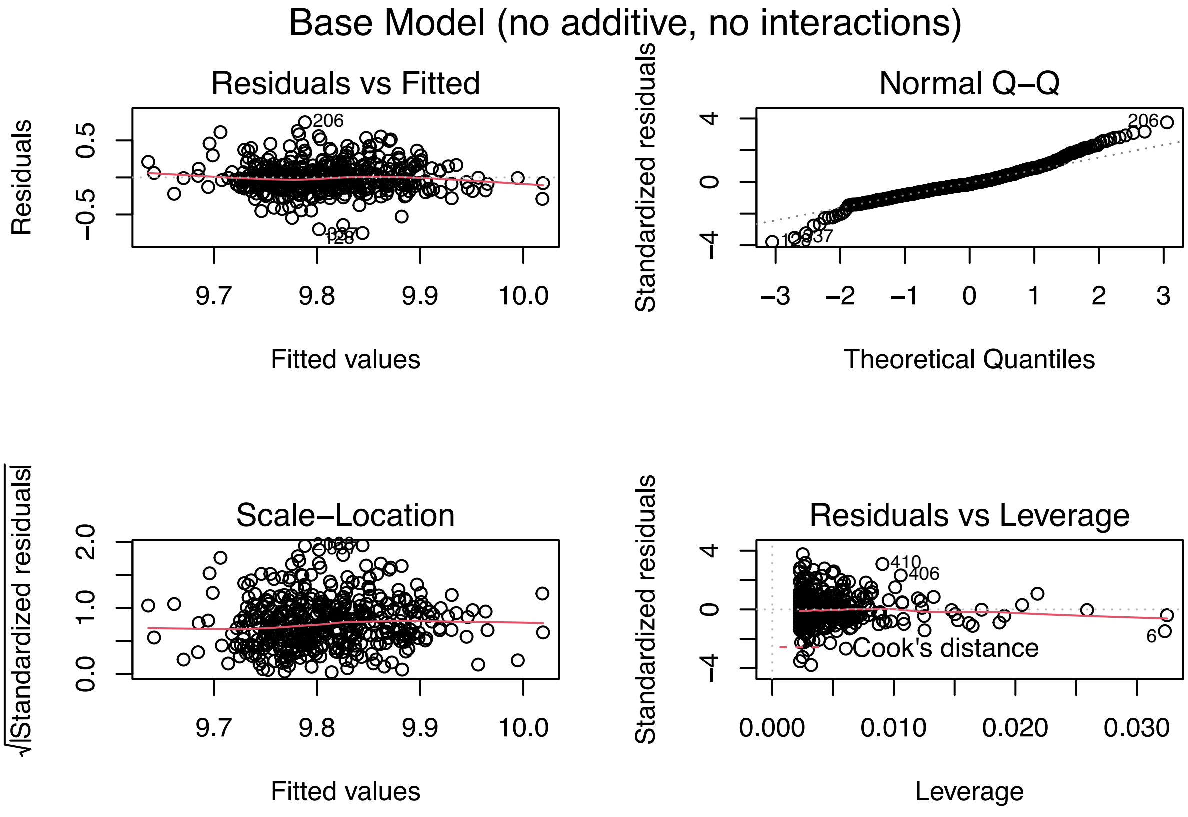
```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(crimes) + region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68757 -0.10557 -0.01422  0.08905  0.78946
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.188431   0.079812 115.125  < 2e-16 ***
## log(crimes)  0.066695   0.008421   7.920 2.00e-14 ***
## regionNE     0.104458   0.025531   4.091 5.11e-05 ***
## regionS     -0.086983   0.023618  -3.683  0.00026 ***
## regionW     -0.055280   0.028167  -1.963  0.05033 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1854 on 435 degrees of freedom
## Multiple R-squared:  0.2032, Adjusted R-squared:  0.1959
## F-statistic: 27.74 on 4 and 435 DF,  p-value: < 2.2e-16
```

```
summary(income_fit3)
```

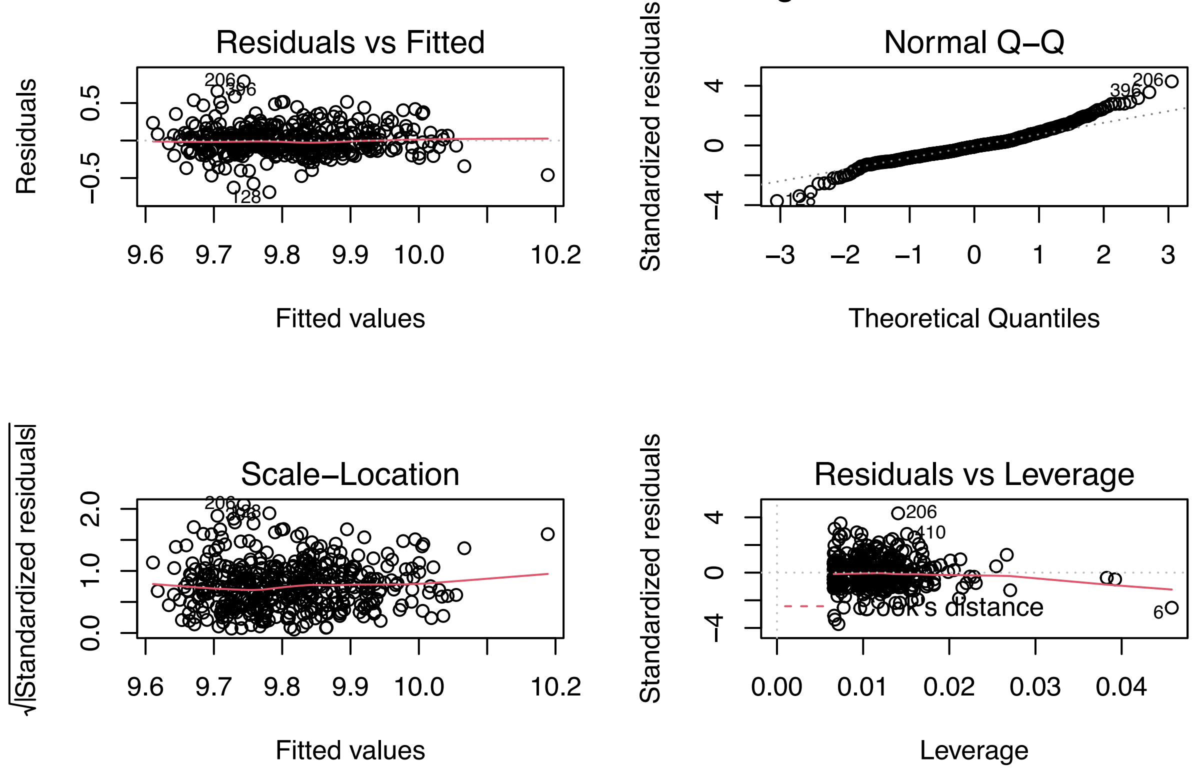
```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(crimes) * region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68552 -0.10418 -0.01444  0.08302  0.79755
```

```
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.33677    0.14579  64.044  < 2e-16 ***
## log(crimes)           0.05064    0.01566   3.233  0.00132 **
## regionNE             -0.18407    0.21515  -0.856  0.39272
## regionS              -0.19717    0.21211  -0.930  0.35312
## regionW              -0.31439    0.24465  -1.285  0.19947
## log(crimes):regionNE  0.03122    0.02311   1.351  0.17749
## log(crimes):regionS   0.01211    0.02228   0.544  0.58696
## log(crimes):regionW   0.02727    0.02523   1.081  0.28028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1855 on 432 degrees of freedom
## Multiple R-squared:  0.2073, Adjusted R-squared:  0.1945
## F-statistic: 16.14 on 7 and 432 DF,  p-value: < 2.2e-16
```

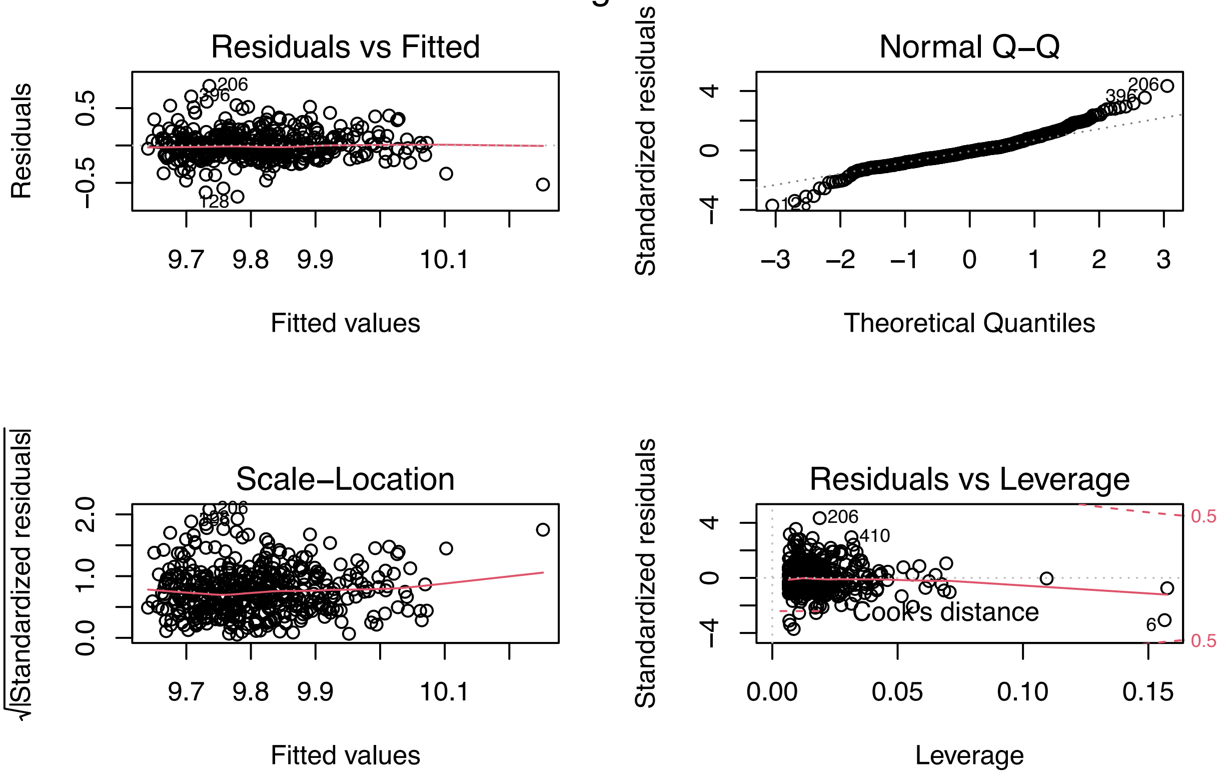
Below are the residual plots for the three models to assess validity.



Model with additive region



Model with additive region and interaction terms



Overall, all three models seem to be fairly valid especially since they conform to the key assumption for linear regression of constant error variance, but they do have some limitations as well.

To really justify the use of the additive and interaction terms, I will be taking a look at the F-tests to compare the models.

```
## 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
```

Per capita crimes

I will first make a new column describing "per-capita crimes".

```
cdi <- cdi %>%
  mutate(
    per.cap.crimes = crimes / pop
  )
```

Next, I will fit the 3 models (with additive, with interaction terms) again:

```
income_fit4 <- lm(log(per.cap.income) ~ log(per.cap.crimes), cdi)
income_fit5 <- lm(log(per.cap.income) ~ log(per.cap.crimes) + region, cdi)
income_fit6 <- lm(log(per.cap.income) ~ log(per.cap.crimes)*region, cdi )
```

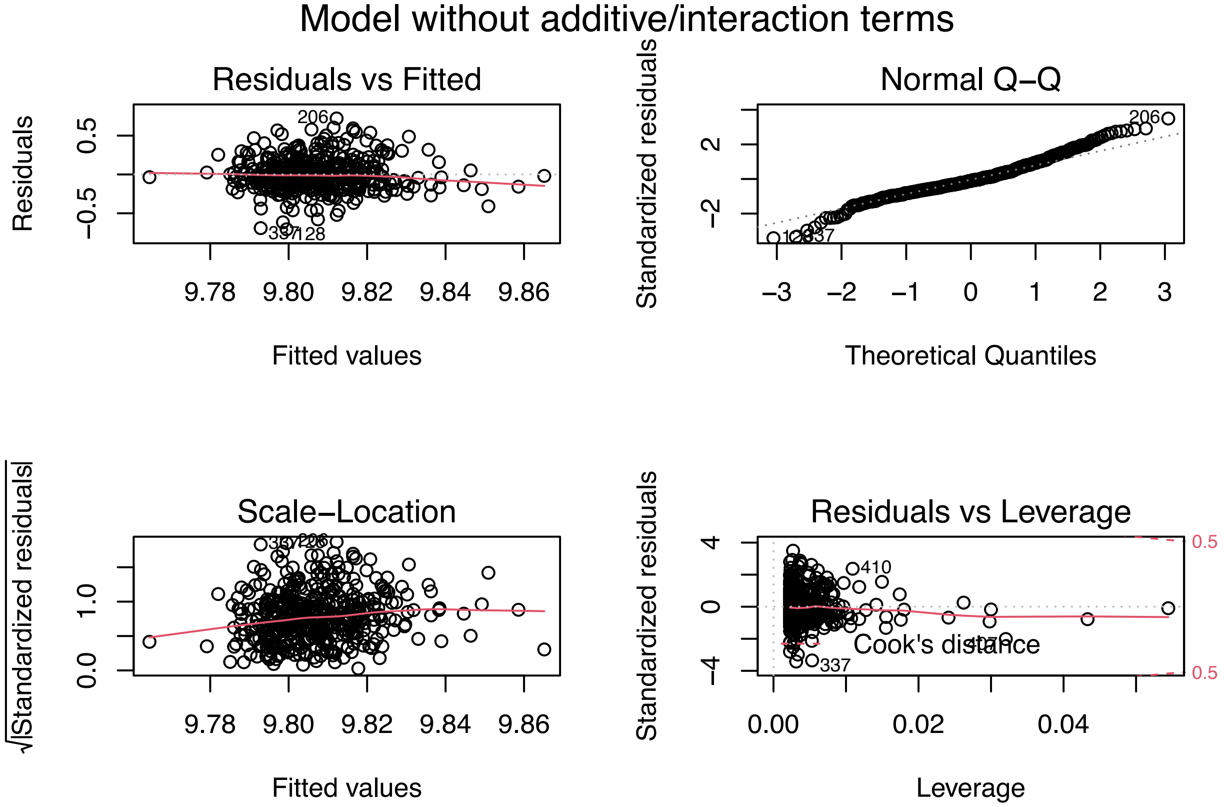
Below are the summaries of the three models

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(per.cap.crimes), data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7058 -0.1242 -0.0221  0.1066  0.7210
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.73510    0.05908 164.765   <2e-16 ***
## log(per.cap.crimes) -0.02417    0.01959  -1.233    0.218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2066 on 438 degrees of freedom
## Multiple R-squared:  0.003461,   Adjusted R-squared:  0.001186
## F-statistic: 1.521 on 1 and 438 DF,  p-value: 0.2181
```

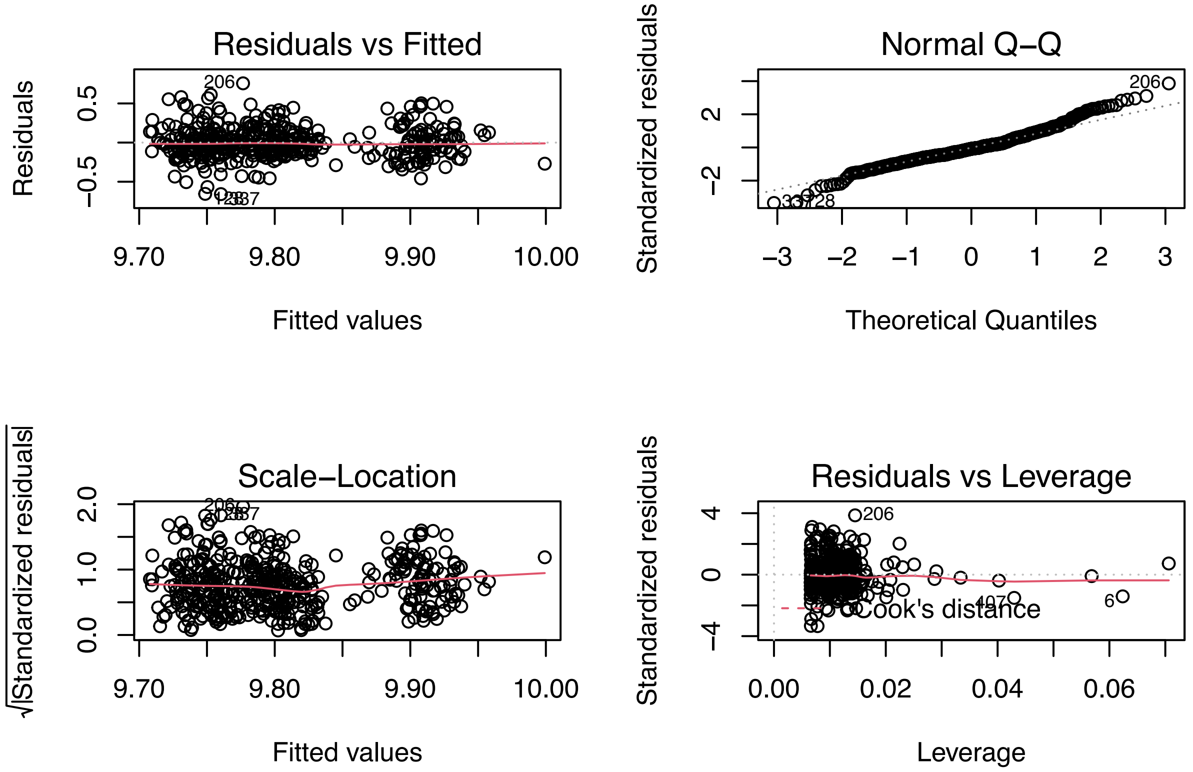
```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(per.cap.crimes) + region,
##     data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65832 -0.11431 -0.01548  0.10838  0.75657
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.93628    0.06934 143.303  < 2e-16 ***
## log(per.cap.crimes)  0.04243    0.02148   1.975  0.04885 *
## regionNE             0.11457    0.02760   4.151 3.99e-05 ***
## regionS             -0.07456    0.02624  -2.841  0.00471 **
## regionW             -0.02426    0.03002  -0.808  0.41952
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1974 on 435 degrees of freedom
## Multiple R-squared:  0.09645,    Adjusted R-squared:  0.08814
## F-statistic: 11.61 on 4 and 435 DF,  p-value: 5.776e-09
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(per.cap.crimes) * region,
##     data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65410 -0.11829 -0.01708  0.10399  0.76628
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   9.91177    0.10503  94.367   <2e-16 ***
## log(per.cap.crimes)           0.03454    0.03327   1.038    0.300
## regionNE                      0.21007    0.17165   1.224    0.222
## regionS                      -0.10137    0.16072  -0.631    0.529
## regionW                       0.07689    0.26753   0.287    0.774
## log(per.cap.crimes):regionNE  0.02924    0.05232   0.559    0.577
## log(per.cap.crimes):regionS  -0.01104    0.05554  -0.199    0.843
## log(per.cap.crimes):regionW   0.03495    0.09268   0.377    0.706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.198 on 432 degrees of freedom
## Multiple R-squared:  0.09773,    Adjusted R-squared:  0.08311
## F-statistic: 6.685 on 7 and 432 DF,  p-value: 1.575e-07
```

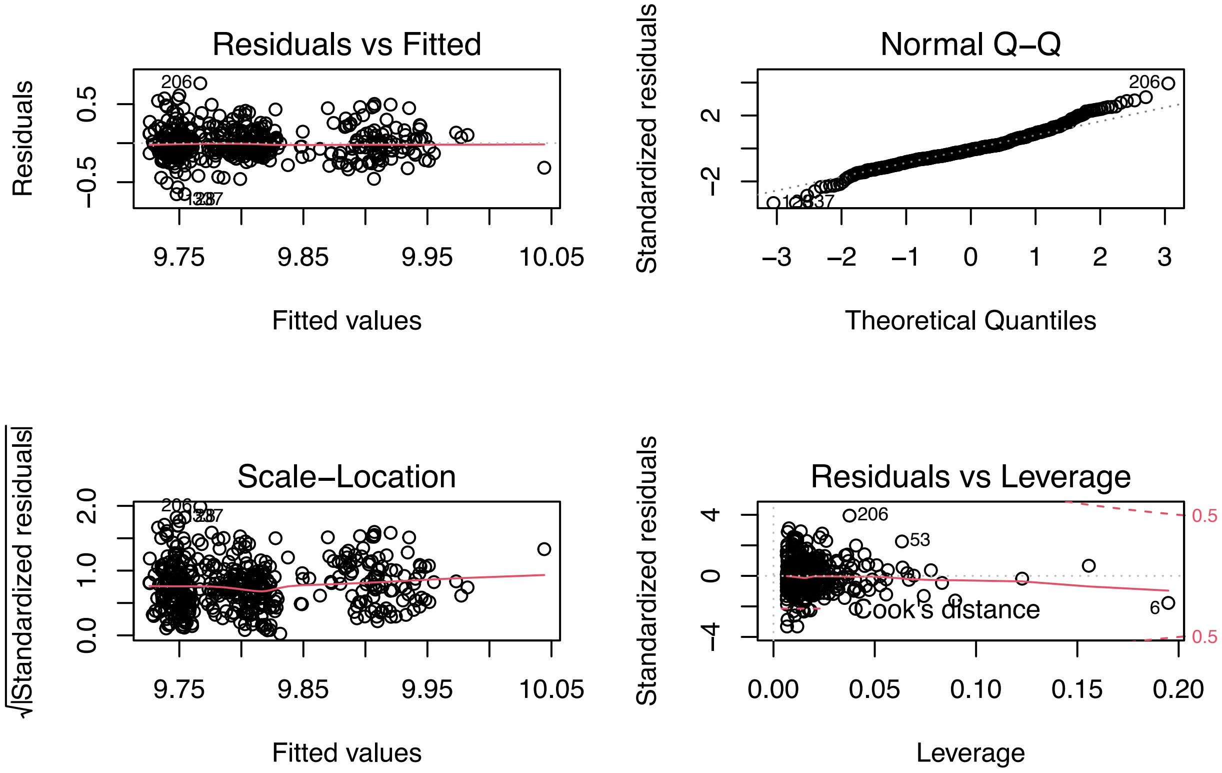
Below is the residual diagnostic plots for the three fitted models.



Model with additive terms



Model with interaction terms



Below we compare the AIC and BIC values between Model 2 and Model 5.

```
AIC(income_fit2, income_fit5)
```

```
##             df       AIC
## income_fit2  6 -227.4746
## income_fit5  6 -172.1347
```

```
BIC(income_fit2, income_fit5)
```

```
##             df       BIC
## income_fit2  6 -202.9539
## income_fit5  6 -147.6140
```

My final model would use *per.cap.crimes* and **not** include any interaction terms.

Research question 3

I have dropped the *id*, *county* and *state* column and will be focusing on the rest of the 13 predictor variables.

```
cdi_new <- read.table("/Users/lee14257/Desktop/CMU MSP/Applied Linear Models/HW/hw06/cdi.dat") %>%
  as_tibble() %>%
  dplyr::select(c(-id, -county, -state)) %>%
  dplyr::select(per.cap.income, everything())
```

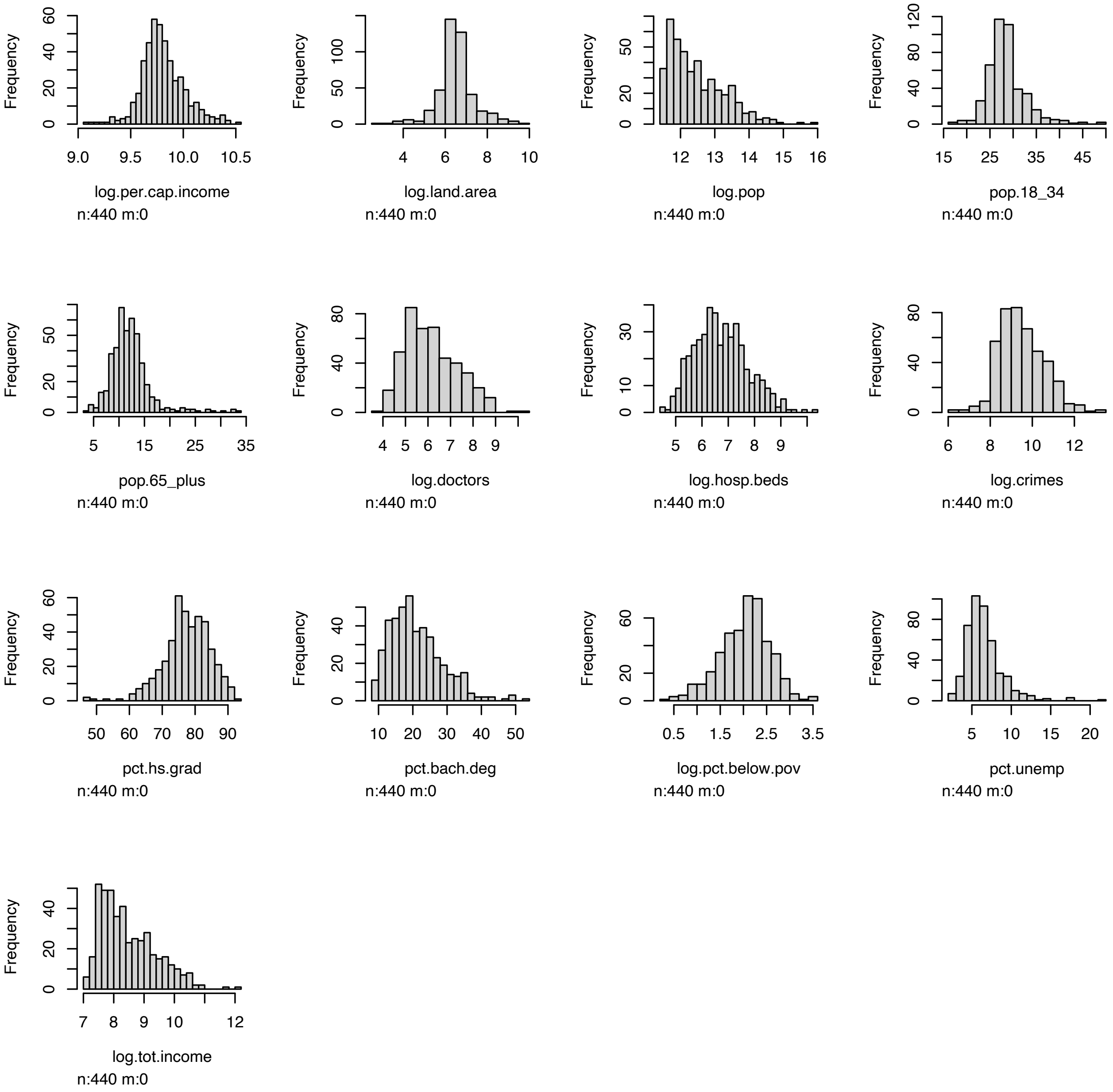
Transformations of variables

```
# Transform the variables
cdi_transformed <- cdi_new

skewed.vars <- c("per.cap.income","land.area", "pop", "doctors", "hosp.beds", "crimes", "tot.income","p

for (tmp in skewed.vars) {
  loc <- grep(paste("^",tmp,"$",sep=""),names(cdi_transformed))
  cdi_transformed[,loc] <- log(cdi_transformed[,loc])
  names(cdi_transformed)[loc] <- paste("log.",names(cdi_transformed)[loc],sep="")
}
```

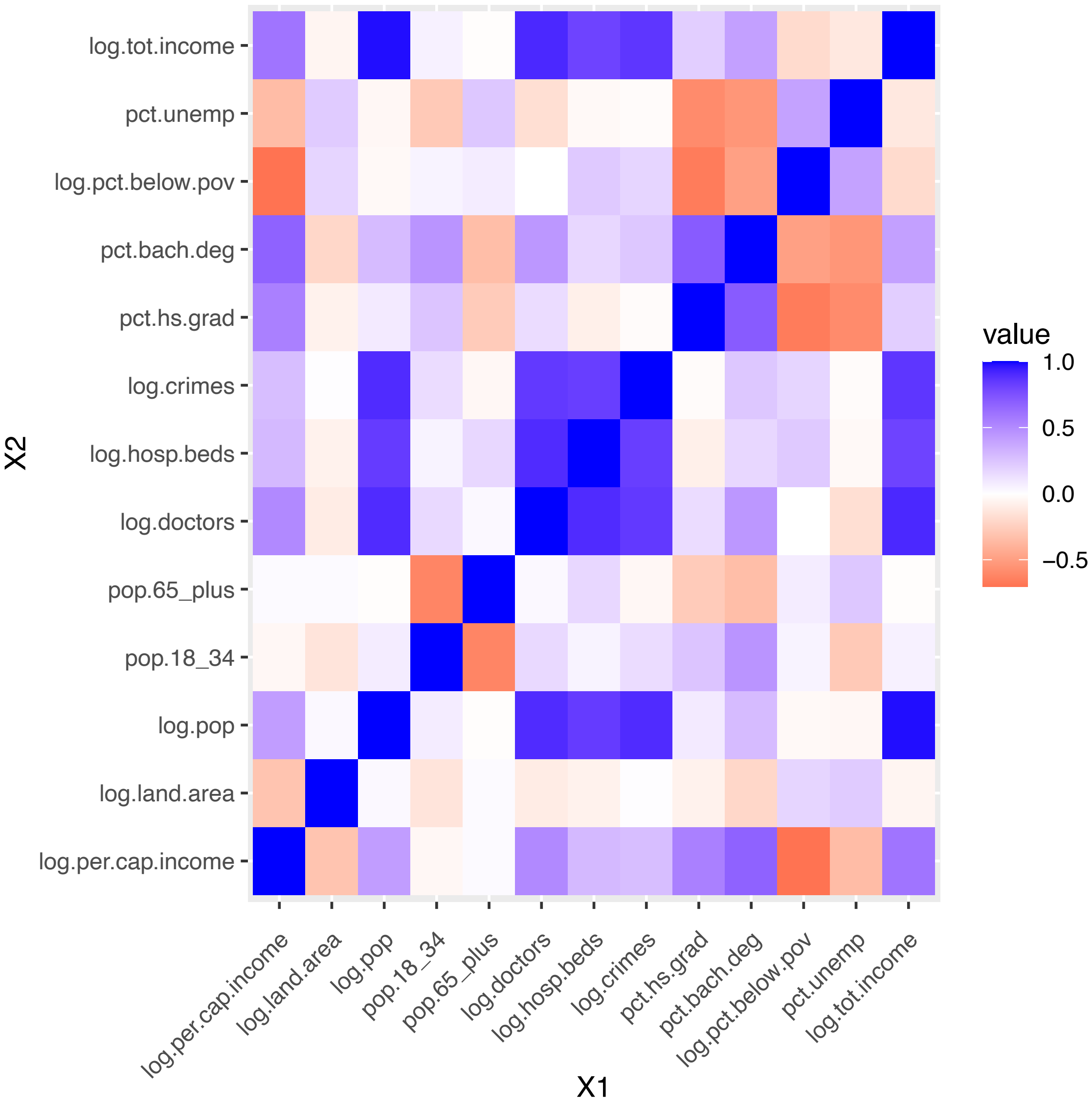
Below we see the histogram plot for the variables after undergoing transformations. A lot of the skewings have improved.



Below is the correlation matrix of the transformed variables.

```
## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE

## Warning in type.convert.default(X[[i]], ...): 'as.is' should be specified by the
## caller; using TRUE
```



Fitting best model

Below we fit a full model including all the variables including the *region* categorical variable (changed to factor).

```
cdi_transformed <- cdi_transformed %>%
  dplyr::select(-log.pop, -log.tot.income)
```

```
full_cdi_model1 <- lm(log.per.cap.income ~ ., data = cdi_transformed)
summary(full_cdi_model1)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ ., data = cdi_transformed)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.36144 -0.04299 -0.00126  0.04709  0.30283 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       10.190127   0.117672  86.598  < 2e-16 ***
## log.land.area     -0.036627   0.005606  -6.533 1.84e-10 ***
## pop.18_34         -0.011184   0.001430  -7.823 4.11e-14 ***
## pop.65_plus        0.001334   0.001427   0.934  0.35060    
## log.doctors        0.036404   0.013732   2.651  0.00833 ** 
## log.hosp.beds      0.024767   0.013912   1.780  0.07575 .  
## log.crimes         0.006864   0.009263   0.741  0.45913    
## pct.hs.grad       -0.002179   0.001130  -1.927  0.05462 .  
## pct.bach.deg       0.012531   0.001097  11.424  < 2e-16 ***
## log.pct.below.pov -0.206448   0.013262 -15.566  < 2e-16 ***
## pct.unemp          0.005502   0.002432   2.262  0.02418 *  
## regionNE          -0.002301   0.013158  -0.175  0.86128    
## regionS           -0.027867   0.012760  -2.184  0.02952 *  
## regionW            0.007510   0.016292   0.461  0.64507    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.08409 on 426 degrees of freedom
## Multiple R-squared:  0.8394, Adjusted R-squared:  0.8345 
## F-statistic: 171.3 on 13 and 426 DF,  p-value: < 2.2e-16
```

Multicollinearity and VIF

Below is the Variance Inflation Factors (VIF) for each of the predictors.

```
##     log.land.area         pop.18_34       pop.65_plus       log.doctors 
##            1.4826            2.2287            2.0155           15.3220 
##     log.hosp.beds        log.crimes       pct.hs.grad      pct.bach.deg 
##           12.0950            6.2380            3.9041            4.3764 
## log.pct.below.pov         pct.unemp          regionNE           regionS 
##            3.0729            2.0074            1.9315            2.2908 
##           regionW 
##            2.3844
```

Variable Selection - All Subsets

After removing *region* from the data, we end up with 10 total predictor variables along with the response variable *log.per.cap.income*. We fit the new model below without region.

```
region_var <- cdi_transformed$region

cdi_transformed <- cdi_transformed %>%
  dplyr::select(-region)
```

```
names(cdi_transformed)
```

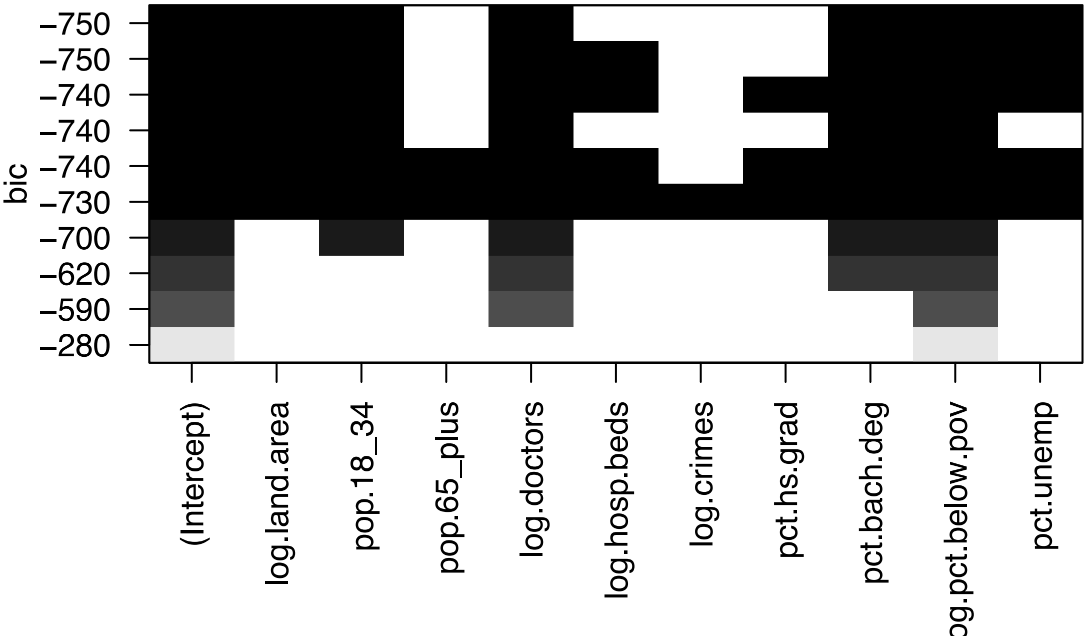
```
##  [1] "log.per.cap.income" "log.land.area"      "pop.18_34"
##  [4] "pop.65_plus"        "log.doctors"        "log.hosp.beds"
##  [7] "log.crimes"         "pct.hs.grad"        "pct.bach.deg"
## [10] "log.pct.below.pov"  "pct.unemp"
```

```
# Fit new model with the existing variables
full_cdi_model2 <- lm(log.per.cap.income ~ ., data=cdi_transformed)
```

All subsets

Below are the results for all subsets. In the plot, the dark squares indicate which variables are in the model that has the BIC values on the left. The darker the squares, the better the model.

```
all_subsets_1 <- regsubsets(log.per.cap.income ~ ., data=cdi_transformed,nvmax=10)
plot(all_subsets_1)
```



All subsets method chose 6 variables that gave the lowest BIC. Below are the coefficient values for the chosen variables.

```
all_subsets_1.summary <- summary(all_subsets_1)
all_subsets_1.summary$bic
```

```
##  [1] -284.6733 -593.3658 -624.5119 -697.5023 -739.1367 -747.6815 -746.1704
##  [8] -741.1124 -735.3797 -729.2930
```

```
tmp <- cdi_transformed %>% dplyr::select(-log.per.cap.income)
min(all_subsets_1.summary$bic)
```

```
## [1] -747.6815
```

```
print(best.model <- which.min(all_subsets_1.summary$bic))
```

```
## [1] 6
```

```
coef(all_subsets_1,best.model)
```

```
##       (Intercept)     log.land.area         pop.18_34       log.doctors 
##      10.095545110      -0.036212594      -0.012026824       0.067772351 
##      pct.bach.deg log.pct.below.pov         pct.unemp 
##       0.010523423      -0.197474797       0.008109587
```

```
cdi_transformed_allsubsets <- tmp[,all_subsets_1.summary$which[best.model,][-1]]
```

```
all_subsets_model <- lm(log.per.cap.income ~ log.land.area + pop.18_34 + 
                          log.doctors + pct.bach.deg + log.pct.below.pov + 
                          pct.unemp, data=cdi_transformed)

summary(all_subsets_model)$coefficients
```

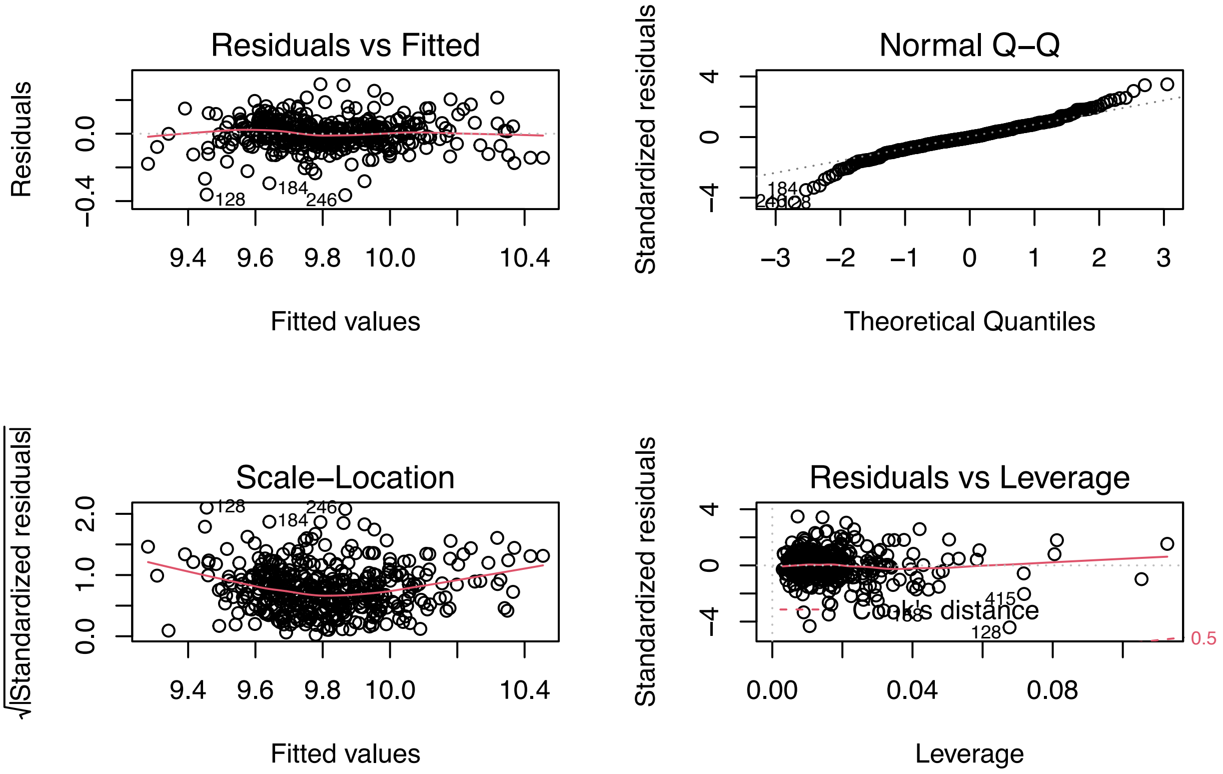
```
##                       Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)       10.095545110 0.0516863528 195.323225 0.000000e+00
## log.land.area     -0.036212594 0.0048198661  -7.513195 3.336311e-13
## pop.18_34         -0.012026824 0.0011883729 -10.120413 9.445833e-22
## log.doctors        0.067772351 0.0041934269  16.161567 2.586913e-46
## pct.bach.deg       0.010523423 0.0008860605  11.876641 2.352204e-28
## log.pct.below.pov -0.197474797 0.0100936858 -19.564191 1.560595e-61
## pct.unemp          0.008109587 0.0021194105   3.826341 1.492036e-04
```

Let us take another look at the VIF for each variables. None of the variables seem to have an excessively large value.

```
vif(all_subsets_model)
```

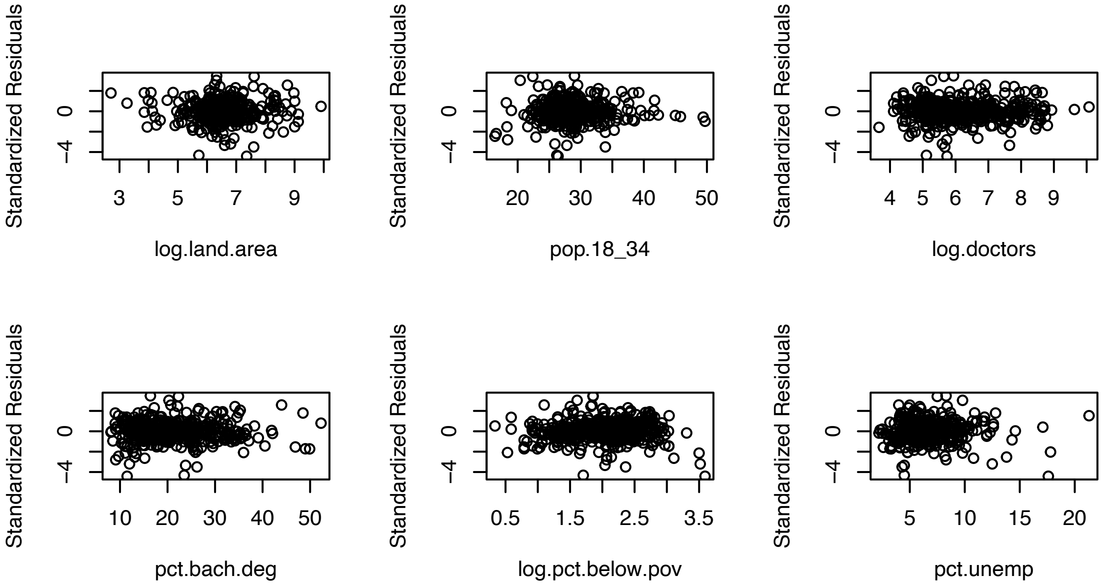
```
##     log.land.area         pop.18_34       log.doctors      pct.bach.deg 
##            1.0778            1.5145            1.4052            2.8085 
## log.pct.below.pov         pct.unemp 
##            1.7505            1.4990
```

Below are the residual diagnostic plots.



Below is the standardized residual plots against each of the predictor variables.

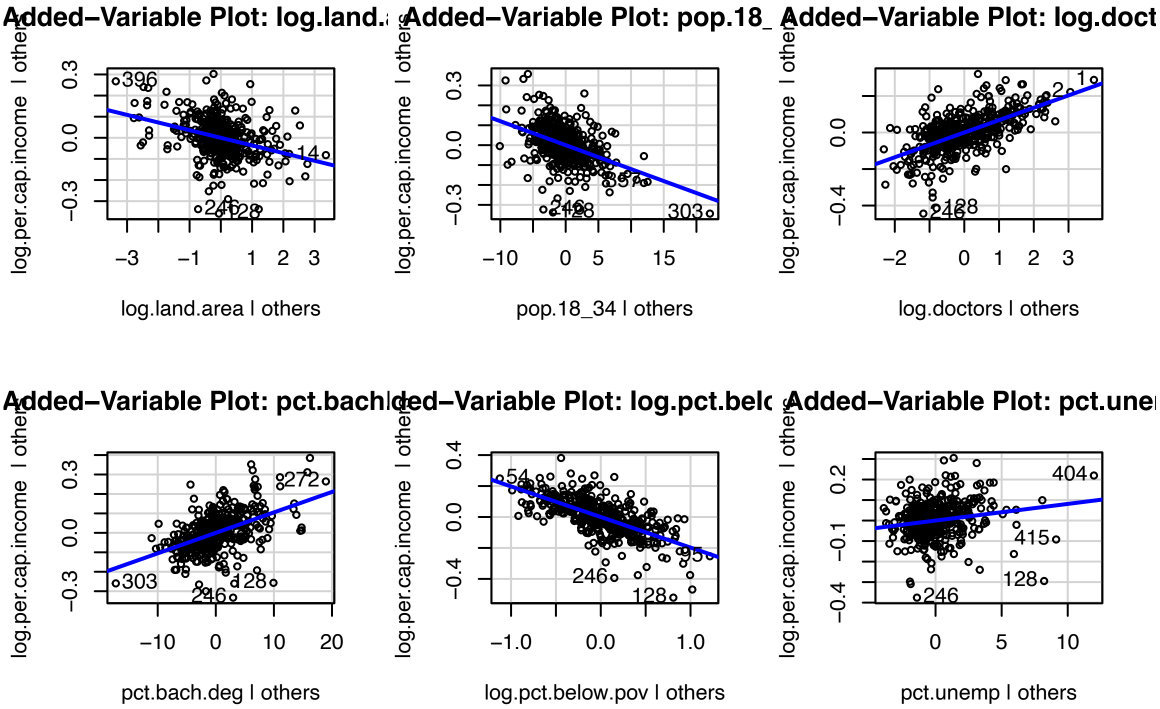
```
## The following objects are masked from cdi_transformed (pos = 3):
##
##     log.crimes, log.doctors, log.hosp.beds, log.land.area,
##     log.pct.below.pov, log.per.cap.income, pct.bach.deg, pct.hs.grad,
##     pct.unemp, pop.18_34, pop.65_plus
```



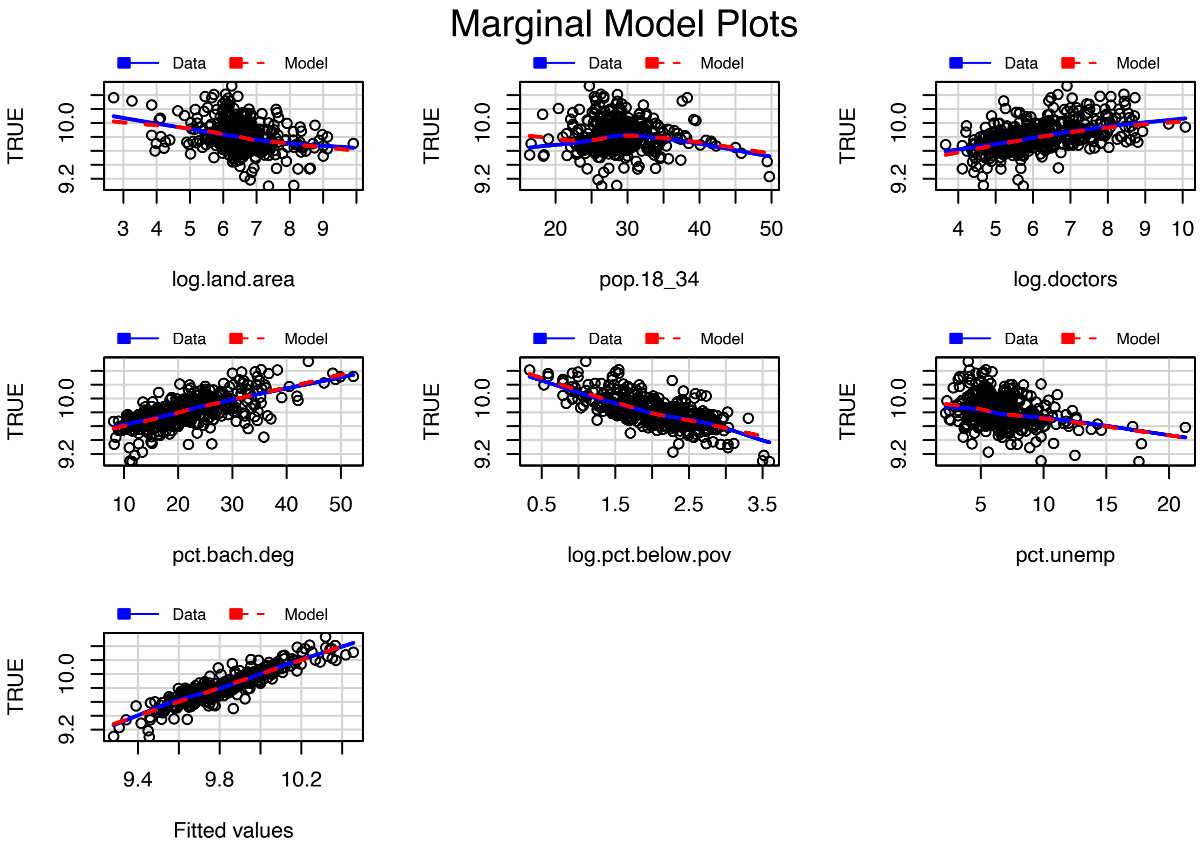
Below is the added variable plot.

```
## The following objects are masked from cdi_transformed (pos = 3):
##
##     log.crimes, log.doctors, log.hosp.beds, log.land.area,
##     log.pct.below.pov, log.per.cap.income, pct.bach.deg, pct.hs.grad,
##     pct.unemp, pop.18_34, pop.65_plus
```

```
## The following objects are masked from cdi_transformed (pos = 4):
##
##     log.crimes, log.doctors, log.hosp.beds, log.land.area,
##     log.pct.below.pov, log.per.cap.income, pct.bach.deg, pct.hs.grad,
##     pct.unemp, pop.18_34, pop.65_plus
```



Below is the marginal plot.



Variable Selection - Stepwise Regression

Below is the **Stepwise Regression** in both directions (backward elimination and forward selection) using BIC as the information criterion.

```
# Stepwise
n=dim(cdi)[1]
stepwise_BIC_cdi <- stepAIC(full_cdi_model2, direction = "both", k=log(n))
```

```
## Start:  AIC=-2117.47
## log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
##     log.doctors + log.hosp.beds + log.crimes + pct.hs.grad +
##     pct.bach.deg + log.pct.below.pov + pct.unemp
##
##                     Df Sum of Sq    RSS     AIC
## - log.crimes         1   0.00000 3.0715 -2123.6
## - pop.65_plus        1   0.00228 3.0738 -2123.2
## - pct.hs.grad        1   0.00693 3.0785 -2122.6
## - log.hosp.beds      1   0.02872 3.1003 -2119.5
## <none>                           3.0715 -2117.5
## - log.doctors        1   0.08411 3.1557 -2111.7
## - pct.unemp          1   0.08441 3.1560 -2111.6
## - log.land.area      1   0.31856 3.3901 -2080.1
## - pop.18_34          1   0.46483 3.5364 -2061.6
## - pct.bach.deg       1   0.88030 3.9518 -2012.7
## - log.pct.below.pov  1   2.23344 5.3050 -1883.1
##
## Step:  AIC=-2123.55
## log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
##     log.doctors + log.hosp.beds + pct.hs.grad + pct.bach.deg +
##     log.pct.below.pov + pct.unemp
##
##                     Df Sum of Sq    RSS     AIC
## - pop.65_plus        1   0.00247 3.0740 -2129.3
## - pct.hs.grad        1   0.00693 3.0785 -2128.7
## - log.hosp.beds      1   0.02908 3.1006 -2125.5
## <none>                           3.0715 -2123.6
## - pct.unemp          1   0.08492 3.1565 -2117.6
## + log.crimes         1   0.00000 3.0715 -2117.5
## - log.doctors        1   0.10550 3.1770 -2114.8
## - log.land.area      1   0.32228 3.3938 -2085.7
## - pop.18_34          1   0.46596 3.5375 -2067.5
## - pct.bach.deg       1   0.88809 3.9596 -2017.9
## - log.pct.below.pov  1   2.26551 5.3371 -1886.5
##
## Step:  AIC=-2129.29
## log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     log.hosp.beds + pct.hs.grad + pct.bach.deg + log.pct.below.pov +
##     pct.unemp
##
##                     Df Sum of Sq    RSS     AIC
## - pct.hs.grad        1   0.00720 3.0812 -2134.3
## - log.hosp.beds      1   0.03324 3.1073 -2130.6
## <none>                           3.0740 -2129.3
## + pop.65_plus        1   0.00247 3.0715 -2123.6
## + log.crimes         1   0.00020 3.0738 -2123.2
## - pct.unemp          1   0.08620 3.1602 -2123.2
## - log.doctors        1   0.10338 3.1774 -2120.8
## - log.land.area      1   0.32972 3.4037 -2090.5
## - pop.18_34          1   0.70581 3.7798 -2044.4
## - pct.bach.deg       1   0.88757 3.9616 -2023.8
## - log.pct.below.pov  1   2.26830 5.3423 -1892.2
##
## Step:  AIC=-2134.34
## log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     log.hosp.beds + pct.bach.deg + log.pct.below.pov + pct.unemp
##
##                     Df Sum of Sq    RSS     AIC
## - log.hosp.beds      1   0.03221 3.1134 -2135.9
## <none>                           3.0812 -2134.3
## + pct.hs.grad        1   0.00720 3.0740 -2129.3
## + pop.65_plus        1   0.00274 3.0785 -2128.7
## + log.crimes         1   0.00018 3.0810 -2128.3
## - log.doctors        1   0.10822 3.1894 -2125.2
## - pct.unemp          1   0.11531 3.1965 -2124.3
## - log.land.area      1   0.37266 3.4539 -2090.2
## - pop.18_34          1   0.71435 3.7956 -2048.7
## - pct.bach.deg       1   0.99368 4.0749 -2017.4
## - log.pct.below.pov  1   2.67750 5.7587 -1865.3
##
## Step:  AIC=-2135.86
## log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.bach.deg + log.pct.below.pov + pct.unemp
##
##                     Df Sum of Sq    RSS     AIC
## <none>                           3.1134 -2135.9
## + log.hosp.beds      1   0.03221 3.0812 -2134.3
## + pop.65_plus        1   0.00694 3.1065 -2130.8
## + pct.hs.grad        1   0.00617 3.1073 -2130.6
## + log.crimes         1   0.00000 3.1134 -2129.8
## - pct.unemp          1   0.10527 3.2187 -2127.3
## - log.land.area      1   0.40588 3.5193 -2088.0
## - pop.18_34          1   0.73646 3.8499 -2048.5
## - pct.bach.deg       1   1.01423 4.1277 -2017.9
## - log.doctors        1   1.87809 4.9915 -1934.3
## - log.pct.below.pov  1   2.75216 5.8656 -1863.3
```

```
anova(all_subsets_model, stepwise_BIC_cdi)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.bach.deg + log.pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.bach.deg + log.pct.below.pov + pct.unemp
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1    433 3.1134
## 2    433 3.1134  0         0
```

Below are the predictor variables that the **stepwise** procedure selected. We can see that stepwise regression using BIC chose the same variables as the all subsets method did.

```
summary(stepwise_BIC_cdi)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     log.doctors + pct.bach.deg + log.pct.below.pov + pct.unemp,
##     data = cdi_transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36433 -0.04268 -0.00228  0.04802  0.29399
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       10.0955451  0.0516864 195.323  < 2e-16 ***
## log.land.area     -0.0362126  0.0048199  -7.513 3.34e-13 ***
## pop.18_34         -0.0120268  0.0011884 -10.120  < 2e-16 ***
## log.doctors        0.0677724  0.0041934  16.162  < 2e-16 ***
## pct.bach.deg       0.0105234  0.0008861  11.877  < 2e-16 ***
```

	df	AIC	BIC
all_subsets_model	8	-913.7973	-881.1031
stepwise_BIC_cdi	8	-913.7973	-881.1031
stepwise_BIC_cdi_interactions	18	-1067.1890	-993.6271

```
## log.pct.below.pov -0.1974748  0.0100937 -19.564  < 2e-16 ***
## pct.unemp          0.0081096  0.0021194   3.826 0.000149 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0848 on 433 degrees of freedom
## Multiple R-squared:  0.8341, Adjusted R-squared:  0.8318
## F-statistic: 362.7 on 6 and 433 DF,  p-value: < 2.2e-16
```

```
cat("\nR2 = ",summary(stepwise_BIC_cdi)$r.squared)
```

```
##
## R2 =  0.8340571
```

```
cat("\nR2adj = ",summary(stepwise_BIC_cdi)$adj.r.squared)
```

```
##
## R2adj =  0.8317576
```

Now lets look at a model using stepwise BIC with two way interaction terms considered.

```
stepwise_BIC_cdi_interactions <- stepAIC(full_cdi_model2,scope=list(lower = ~ 1, upper = ~ .^2),
                      k=log(dim(cdi_transformed)[1]),                          ## BIC penalty.
                      trace=F)

comparison <- cbind(
AIC(all_subsets_model,stepwise_BIC_cdi,stepwise_BIC_cdi_interactions),
BIC(all_subsets_model,stepwise_BIC_cdi,stepwise_BIC_cdi_interactions))
comparison <- comparison[,-3]
names(comparison) <- c("df","AIC","BIC")
comparison %>% kbl(booktabs=T) %>% kable_classic()
```

```
round(summary(stepwise_BIC_cdi_interactions)$coef,2)
```

```
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          11.78       0.36   32.35     0.00
## log.land.area                         0.10       0.03    3.03     0.00
## pop.18_34                            -0.03       0.01   -5.24     0.00
## pop.65_plus                          -0.03       0.00  -10.01     0.00
## log.doctors                          -0.05       0.03   -1.68     0.09
## log.hosp.beds                         0.00       0.02    0.04     0.97
## pct.hs.grad                          -0.02       0.00   -8.38     0.00
## pct.bach.deg                          0.02       0.00    5.02     0.00
## log.pct.below.pov                  -0.65       0.12   -5.21     0.00
## pct.unemp                           0.01       0.00    5.12     0.00
## pop.65_plus:pct.bach.deg            0.00       0.00    9.71     0.00
## pct.hs.grad:log.pct.below.pov       0.01       0.00    8.02     0.00
## pct.bach.deg:log.pct.below.pov      0.00       0.00   -4.03     0.00
## log.land.area:log.pct.below.pov    -0.04       0.01   -3.86     0.00
## log.land.area:pct.bach.deg          0.00       0.00   -2.61     0.01
## pop.18_34:log.doctors               0.00       0.00    2.61     0.01
## log.hosp.beds:log.pct.below.pov     0.02       0.01    2.47     0.01
```

```
cat("\nR2 = ",summary(stepwise_BIC_cdi_interactions)$r.squared)
```

```
##
## R2 =  0.8881038
```

```
cat("\nR2adj = ",summary(stepwise_BIC_cdi_interactions)$adj.r.squared)
```

```
##
## R2adj =  0.8838713
```

Although there is a decrease in AIC and BIC as well as increase in R squared value, I would still be disinclined to include the interaction terms, since the improvement is pretty small compared to all the variables and interaction terms added to the model. It would be worth to discuss this with the social scientist, but would also be hard to explain the meaning behind these interaction terms.

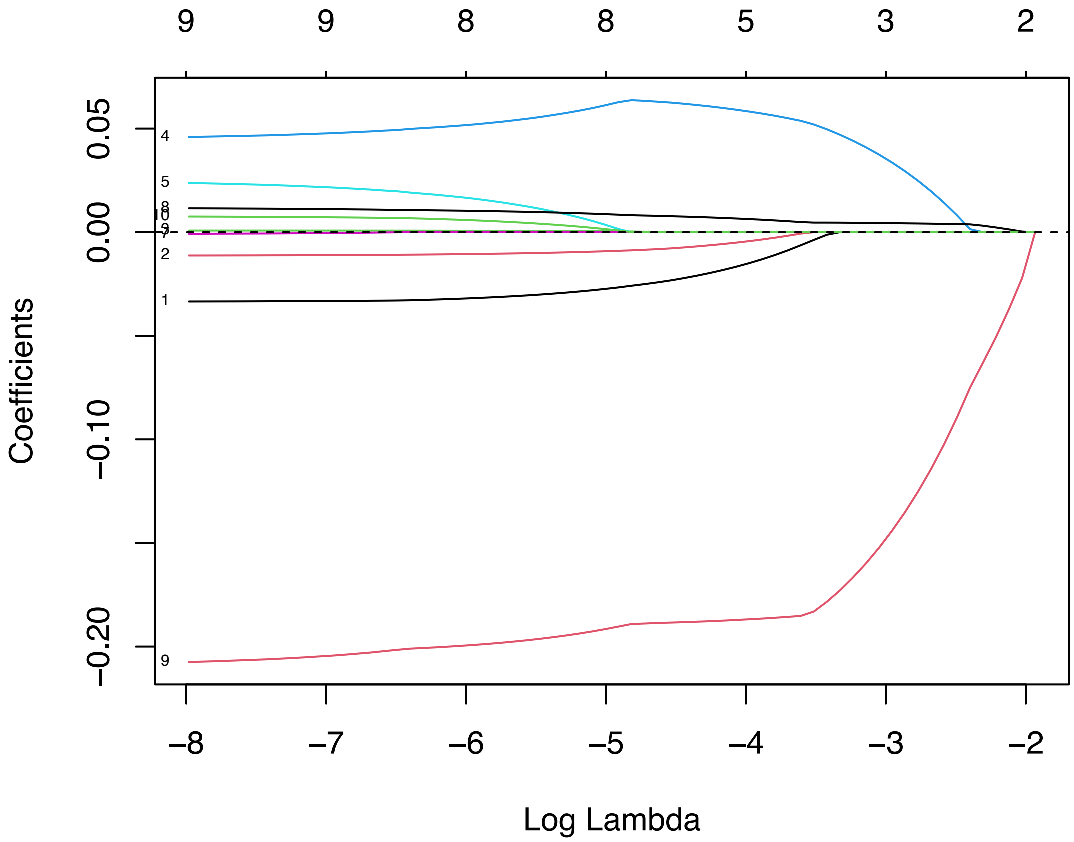
Variable Selection - LASSO regression

Let us try another variable selection method called LASSO regression. Note that the variable *region* was removed since LASSO cannot make use of categorical variables.

```
set.seed(1000)
#cdi_transformed_num <- cdi_transformed %>%
  #dplyr::select(-region)
  #mutate(region = as.numeric(region))
x.full_cdi <- as.matrix(cdi_transformed[,-1])
y.full_cdi <- as.matrix(cdi_transformed[,1])
fit.lasso_cdi <- glmnet(x.full_cdi, y.full_cdi)
```

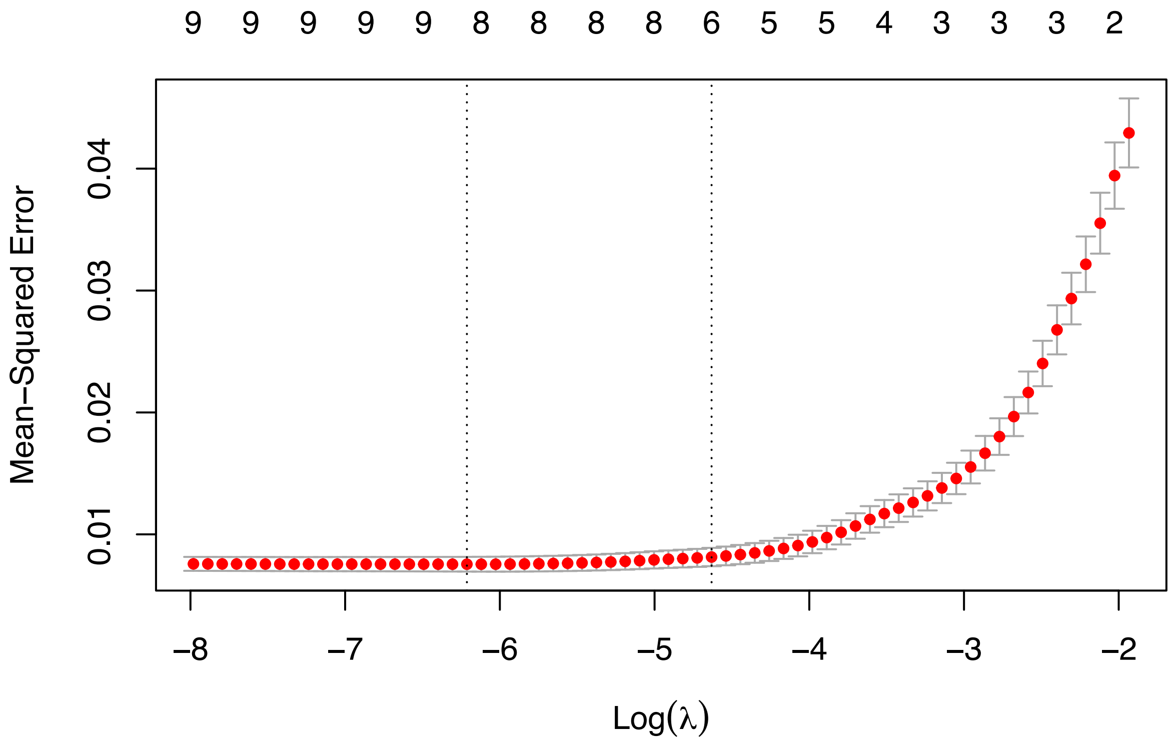
The plot shows how many non-zero variables are in the model at the top. So at a log Lambda of -4, the model has 5 variables.

```
plot(fit.lasso_cdi, xvar="lambda", label=T)
abline(h=0, lty=2)
```



Below is the plot of MSE vs Log Lambda.

```
result_cdi <- cv.glmnet(x.full_cdi, y.full_cdi)
plot(result_cdi)
```



I will be using **lambda.1se** of 0.0097, which is the value of lambda that is one SE larger than lambda.min, since it can protect against capitalization on chance.

```
c(lambda.1se = result_cdi$lambda.1se, lambda.min = result_cdi$lambda.min)
```

```
##  lambda.1se  lambda.min
## 0.009740676 0.002003182
```

Below we can see the variable selection results using LASSO and the lambda value I chose (*lambda.1se*) vs *lambda.min*.

```
tmp <- cbind(coef(result_cdi, s=result_cdi$lambda.min), coef(result_cdi, s=result_cdi$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)         10.0271525139      10.0272981586
## log.land.area       -0.0325044714      -0.0242271305
## pop.18_34           -0.0107392687      -0.0081926462
## pop.65_plus          0.0007238627       0.0001013881
## log.doctors          0.0506163765       0.0629713628
## log.hosp.beds        0.0180587422        .
## log.crimes           .                   .
## pct.hs.grad          .                   .
## pct.bach.deg         0.0105312491       0.0078837697
## log.pct.below.pov   -0.2003705751      -0.1885968546
## pct.unemp            0.0063668044        .
```

Below is the summary of the resulting model using variables selected from LASSO.

```
full_cdi_model_lasso <- lm(log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus + log.doctors +
                             pct.bach.deg + log.pct.below.pov, data = cdi_transformed)
summary(full_cdi_model_lasso)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     pop.65_plus + log.doctors + pct.bach.deg + log.pct.below.pov,
##     data = cdi_transformed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.36564 -0.04698 -0.00367  0.04932  0.30155
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       10.1116202  0.0628428 160.903  < 2e-16 ***
## log.land.area     -0.0337191  0.0049006  -6.881 2.10e-11 ***
## pop.18_34         -0.0116315  0.0014289  -8.140 4.24e-15 ***
## pop.65_plus        0.0014938  0.0013571   1.101    0.272
## log.doctors        0.0673303  0.0043416  15.508  < 2e-16 ***
## pct.bach.deg       0.0096519  0.0008668  11.135  < 2e-16 ***
## log.pct.below.pov -0.1911802  0.0100990 -18.931  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0861 on 433 degrees of freedom
## Multiple R-squared:  0.8289, Adjusted R-squared:  0.8266
## F-statistic: 349.7 on 6 and 433 DF,  p-value: < 2.2e-16
```

Let us now compare the predictors selected from stepwise method and LASSO.

```
## # A tibble: 10 x 3
##    Variables         stepwise_final LASSO
##    <chr>                      <int> <int>
##  1 log.land.area                  1     1
##  2 pop.18_34                      1     1
##  3 pop.65_plus                    0     1
##  4 log.doctors                    1     1
##  5 log.hosp.beds                  0     0
##  6 log.crimes                     0     0
##  7 pct.hs.grad                    0     0
##  8 pct.bach.deg                   1     1
##  9 log.pct.below.pov              1     1
## 10 pct.unemp                      1     0
```

```
allsubset_lasso_common_model<-lm(log.per.cap.income ~ log.land.area + pop.18_34 +
                                     log.doctors + pct.bach.deg +
                                   log.pct.below.pov, cdi_transformed)
```

We can notice that the all subsets and LASSO regression chose the same 5 variables (*log.land.area, pop.18_34, log.doctors, pct.bach.deg, log.pct.below.pov*), except for the fact that LASSO chose to additionally include the predictor *pop.65_plus*, while our final stepwise regression model chose *pct.unemp* instead. We can quickly perform an ANOVA F test on the models to see which one is the most significant.

```
anova(allsubset_lasso_common_model, all_subsets_model, full_cdi_model_lasso)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.bach.deg + log.pct.below.pov
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.bach.deg + log.pct.below.pov + pct.unemp
## Model 3: log.per.cap.income ~ log.land.area + pop.18_34 + pop.65_plus +
##     log.doctors + pct.bach.deg + log.pct.below.pov
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    434 3.2187
## 2    433 3.1134  1  0.105273 14.641 0.0001492 ***
## 3    433 3.2097  0 -0.096292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adding back region

Lastly, we can check if adding back the *region* variable helps in any way. We will be keeping the categorical variable if any indicators for the categorical variable is statistically significant.
* Keep: `region`, `region:pct.below.pov`, `region:pct.unemp`

- Drop: `region:log.land.area`, `region:pop.18_34`, `region:log.doctors`, `region:pct.bach.deg`

```
cdi_transformed_allsubsets <- cbind(cdi_transformed_allsubsets, log.per.cap.income = cdi_transformed$log
tmp <- cbind(cdi_transformed_allsubsets,region=cdi$region)
all_subsets_model_with_region <- lm(log.per.cap.income ~ .*region,data=tmp)
summary(all_subsets_model_with_region)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ . * region, data = tmp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33212 -0.04534 -0.00384  0.04414  0.34554
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 9.9858194  0.1251014  79.822  < 2e-16 ***
## log.land.area              -0.0230428  0.0153531  -1.501  0.13416
## pop.18_34                  -0.0127476  0.0028646  -4.450 1.11e-05 ***
## log.doctors                 0.0537441  0.0091796   5.855 9.77e-09 ***
## pct.bach.deg                0.0112314  0.0023924   4.695 3.64e-06 ***
## log.pct.below.pov          -0.1554738  0.0250420  -6.209 1.31e-09 ***
## pct.unemp                   0.0146486  0.0050548   2.898  0.00396 **
## regionNE                    0.1183333  0.1870451   0.633  0.52732
## regionS                     0.3339204  0.1555420   2.147  0.03239 *
## regionW                    -0.1049334  0.1831194  -0.573  0.56694
## log.land.area:regionNE     -0.0198535  0.0197240  -1.007  0.31474
## log.land.area:regionS      -0.0182742  0.0178437  -1.024  0.30638
## log.land.area:regionW      -0.0007866  0.0187013  -0.042  0.96647
## pop.18_34:regionNE         -0.0012844  0.0040299  -0.319  0.75010
## pop.18_34:regionS          -0.0025245  0.0033247  -0.759  0.44811
## pop.18_34:regionW           0.0044403  0.0044363   1.001  0.31746
## log.doctors:regionNE        0.0068329  0.0133119   0.513  0.60802
## log.doctors:regionS         0.0105406  0.0116884   0.902  0.36769
## log.doctors:regionW         0.0209585  0.0130712   1.603  0.10961
## pct.bach.deg:regionNE       0.0031476  0.0032855   0.958  0.33862
## pct.bach.deg:regionS       -0.0012692  0.0027056  -0.469  0.63923
## pct.bach.deg:regionW        0.0003701  0.0032104   0.115  0.90827
## log.pct.below.pov:regionNE -0.0211976  0.0366029  -0.579  0.56282
## log.pct.below.pov:regionS  -0.0067038  0.0297971  -0.225  0.82210
## log.pct.below.pov:regionW  -0.0914301  0.0412887  -2.214  0.02735 *
## pct.unemp:regionNE         -0.0036546  0.0077360  -0.472  0.63688
## pct.unemp:regionS          -0.0313720  0.0066655  -4.707 3.44e-06 ***
## pct.unemp:regionW           0.0018297  0.0062413   0.293  0.76954
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08054 on 412 degrees of freedom
## Multiple R-squared:  0.8576, Adjusted R-squared:  0.8482
## F-statistic: 91.87 on 27 and 412 DF,  p-value: < 2.2e-16
```

Thus we arrive at the following model. All the main effects and interaction terms that involve *region* have at least one significant term and the R squared (0.85) and residual standard error did not change too much.

```
all_subsets_model_with_some_region <- update(all_subsets_model_with_region,
                                                . ~ . - region:log.land.area -
                                                  region:pop.18_34 - region:log.doctors -
                                               region:pct.bach.deg)
summary(all_subsets_model_with_some_region)
```

```
## 
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 + 
##     log.doctors + pct.bach.deg + log.pct.below.pov + pct.unemp + 
##     region + log.pct.below.pov:region + pct.unemp:region, data = tmp)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.37137 -0.04631 -0.00436  0.04248  0.35086 
## 
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                1.001e+01  6.595e-02 151.841  < 2e-16 ***
## log.land.area             -3.423e-02  5.498e-03  -6.227 1.15e-09 ***
## pop.18_34                 -1.295e-02  1.167e-03 -11.095  < 2e-16 ***
## log.doctors                6.597e-02  4.133e-03  15.960  < 2e-16 ***
## pct.bach.deg               1.079e-02  8.874e-04  12.157  < 2e-16 ***
## log.pct.below.pov         -1.668e-01  1.944e-02  -8.579  < 2e-16 ***
## pct.unemp                  1.569e-02  4.266e-03   3.678 0.000265 ***
## regionNE                   1.172e-01  5.038e-02   2.326 0.020509 *  
## regionS                    1.503e-01  4.669e-02   3.218 0.001388 ** 
## regionW                    1.525e-01  6.177e-02   2.468 0.013972 *  
## log.pct.below.pov:regionNE -3.723e-02  2.658e-02  -1.401 0.162087    
## log.pct.below.pov:regionS  -1.069e-05  2.294e-02   0.000 0.999628    
## log.pct.below.pov:regionW  -7.733e-02  3.459e-02  -2.235 0.025919 *  
## pct.unemp:regionNE        -7.459e-03  6.964e-03  -1.071 0.284734    
## pct.unemp:regionS         -2.835e-02  5.543e-03  -5.114 4.78e-07 ***
## pct.unemp:regionW         -5.860e-04  5.418e-03  -0.108 0.913929    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.08119 on 424 degrees of freedom
## Multiple R-squared:  0.851,  Adjusted R-squared:  0.8457 
## F-statistic: 161.5 on 15 and 424 DF,  p-value: < 2.2e-16
```

Now lets compare the allsubsets model with 6 variables, and our model with region interaction terms added. Below are the results for F-test, AIC and BIC values. The results show that it is worth to add these terms rather than the base model.

```
# ANOVA
all_subsets_model_add_region <- lm(log.per.cap.income ~ log.land.area + pop.18_34 +
                          log.doctors + pct.bach.deg + log.pct.below.pov +
                          pct.unemp + region, data=tmp)
anova(all_subsets_model, all_subsets_model_add_region, all_subsets_model_with_some_region)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.bach.deg + log.pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.bach.deg + log.pct.below.pov + pct.unemp + region
## Model 3: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.bach.deg + log.pct.below.pov + pct.unemp + region + log.pct.below.pov:region +
##     pct.unemp:region
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    433 3.1134
## 2    430 3.0760  3   0.03739 1.8905  0.1305
## 3    424 2.7952  6   0.28082 7.0994 3.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# AIC comparison
AIC(all_subsets_model, all_subsets_model_add_region, all_subsets_model_with_some_region)
```

```
##                                    df       AIC
## all_subsets_model                   8 -913.7973
## all_subsets_model_add_region       11 -913.1134
## all_subsets_model_with_some_region 17 -943.2351
```

```
# BIC comparison
BIC(all_subsets_model, all_subsets_model_add_region, all_subsets_model_with_some_region)
```

```
##                                    df       BIC
## all_subsets_model                   8 -881.1031
## all_subsets_model_add_region       11 -868.1589
## all_subsets_model_with_some_region 17 -873.7599
```

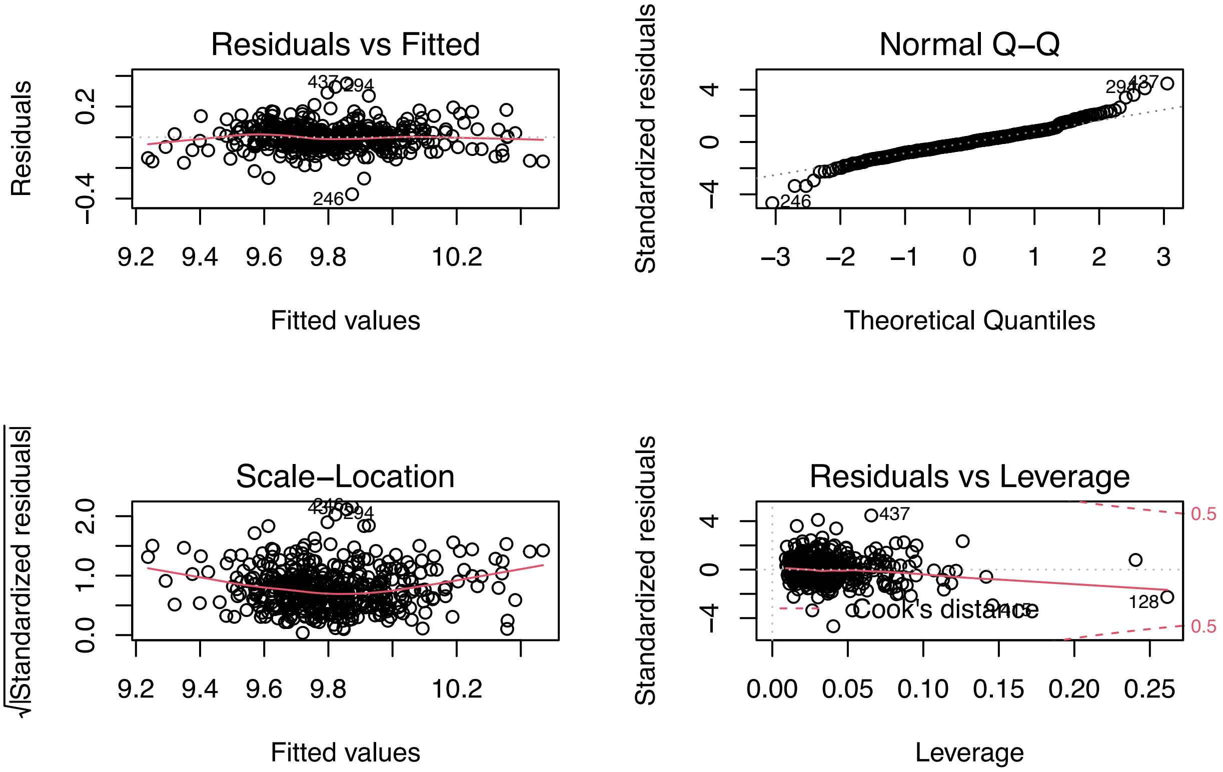
The *all_subsets_model_with_some_region* seems to be the optimal choice here.

Final model with region variable and assessing validity of model

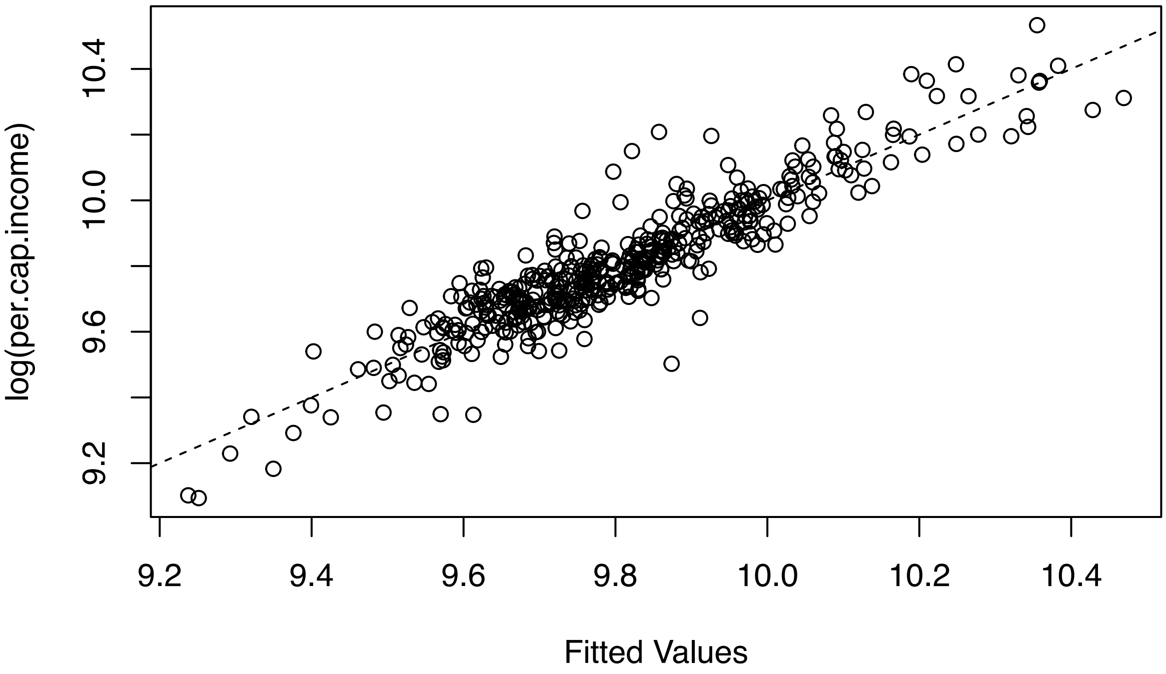
The summary of my final model (with *region*) can be seen below.

```
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   10.014      0.066 151.841    0.000
## log.land.area                 -0.034      0.005  -6.227    0.000
## pop.18_34                     -0.013      0.001 -11.095    0.000
## log.doctors                    0.066      0.004  15.960    0.000
## pct.bach.deg                   0.011      0.001  12.157    0.000
## log.pct.below.pov             -0.167      0.019  -8.579    0.000
## pct.unemp                      0.016      0.004   3.678    0.000
## regionNE                       0.117      0.050   2.326    0.021
## regionS                        0.150      0.047   3.218    0.001
## regionW                          0.152      0.062   2.468    0.014
## log.pct.below.pov:regionNE      -0.037      0.027  -1.401    0.162
## log.pct.below.pov:regionS        0.000      0.023   0.000    1.000
## log.pct.below.pov:regionW       -0.077      0.035  -2.235    0.026
## pct.unemp:regionNE              -0.007      0.007  -1.071    0.285
## pct.unemp:regionS               -0.028      0.006  -5.114    0.000
## pct.unemp:regionW               -0.001      0.005  -0.108    0.914
##
## R2 =  0.8510172
##
## R2adj =  0.8457466
```

Below are the residual diagnostic plots.



Below is the plot of Y_hat vs Y.


log(per.cap.income)
9.2
9.6
10.0
10.4
9.2
9.4
9.6
9.8
10.0
10.2
10.4
Fitted Values