Impact of 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 region variables are fairly related to per-capita income, and that the best model involves non-collinear variables like number of doctors, percentage of bachelor degrees, percentage of unemployment as well as some added interaction terms with region being significant predictors. The best model could be improved further by exploring two-way interaction terms of quantitative variables, and obtaining additional data for better 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, which measures 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 in income across the different counties. A county's prosperity can be influenced by various economic, health as well as 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.

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.

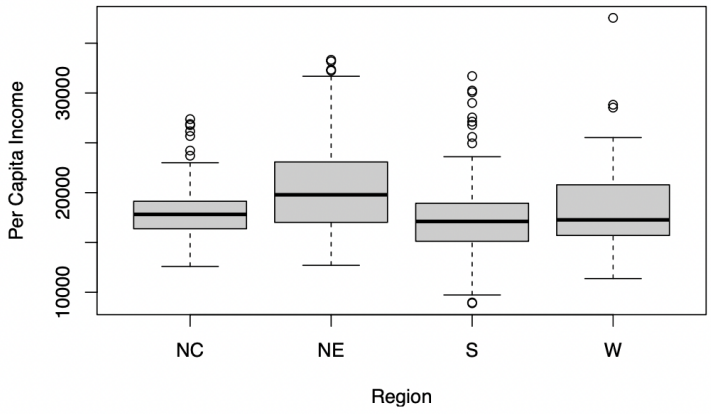


Figure 1: Per-capita income per region

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.

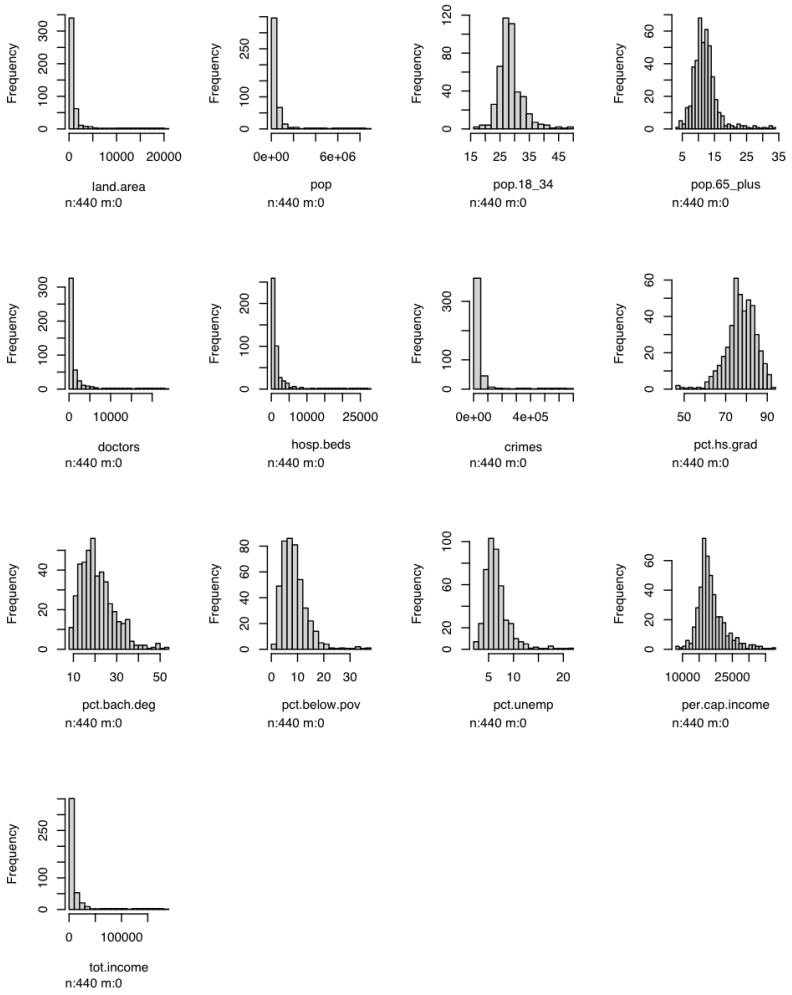


Figure 2: Histogram distributions of quantitative variables

Methods

We will address 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 should be related to crime rate, and that this relationship may be different in different regions of the country, we built regression models to predict *per.cap.income* from *crimes* and *region*. We considered regression models using logarithmic transformations of the response variable *per.cap.income* and the quantitative predictor *crimes*, and further considered models with and without the *region* additive term, as well as the interaction terms between *crimes* and *region*.

We evaluated the validity of the models through residual diagnostic plots, as well as 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 crime rate variable with per-capita crime rate given by *crimes/pop* to investigate 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 the response and predictors, were evaluated to assess whether they needed transformations or not. The variables that were highly skewed and needed logarithmic transformations were:

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

Only logarithmic transformations were used, not only because some of the variables had slight skewdness that were negligible, 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 predictor variables 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 also one that satisfied key modeling assumptions as well as one that was interpretable in the context of social science and economics.

Thereafter, the categorical variable *region* was added back to determine its significance in predicting per-capita income. Both additive terms and interactions terms were considered, and we observed the coefficient summary to check if any indicators for *region* was statistically significant. Here, if any indicator for a categorical variable or its interactions terms seemed important then we chose to keep the whole categorical variable. If none seemed important, then we dropped the whole variable.

Similar to the second research question, F tests (ANOVA) as well as the AIC and BIC values were used to compare the fits of models containing different subsets of variables. We evaluated the validity of the models through residual diagnostic plots, and assessed the significance of each coefficients through model summaries.

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 page 7 of the Technical Appendix suggests that:

- *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*, where *per.cap.income* = *tot.income*/*pop*.
- 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 we would need to address in our analysis.

We further looked at the boxplot of per-capita income per region in Figure 2, 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 to 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*, as well as the additive / interaction terms with the *region* categorical variable (details in page 8 and 9 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,

$$log.per.cap.income = 0.054 \cdot log.crimes + 9.29 \quad (1)$$

As seen in page 8 of the Technical Appendix, both coefficients were statistically significant with low p values, and the R squared value was 0.079, meaning approximately 7.9% of the total variability in the response variable was explained by the model. 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,

$$log.per.cap.income = 0.067 \cdot log.crimes + 0.1 \cdot regionNE - 0.09 \cdot regionS - 0.06 \cdot regionW + 9.19 \quad (2)$$

Page 8 and 9 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, meaning approximately 20.3% of the total variability in the response variable was explained by the model. 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 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 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 for all three models (page 10, 11 and 12 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 including normality of the residuals and a few existing 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. F test Analysis of Variance (ANOVA) was performed on the models as seen in Table 3, to really justify whether the interaction terms in the third model were effective or not. The second model with the additive *region* variable without interaction terms 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)

2.4 Per-capita crimes

We also substituted the *crimes* variable with per-capita crimes and fit the three identical regression models in 2.1, 2.2 and 2.3. Page 13 and 14 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 when compared to the models using the raw *crimes* variable. Further, the *log.per.capita.crimes* predictor variable was no longer as significant in the three models as *log.crimes* was in the previous three models.

The residual diagnostic plots in pages 16, 17 and 18 of the Technical Appendix showed little differences from the previous 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 (with additive *region* and no interactions) from the raw *log.crime* and *log.per.capita.crime* models. The results in page 18 of the Technical Appendix shows that the model with raw *crimes* had smaller AIC and BIC values (smaller the better).

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

Figure 3 shows the histogram plots for the quantitative variables after the logarithmic transformations were applied. It could be observed that a lot of the skewing has improved. Further, the correlation heatmap after logarithmic transformation shown in Page 21 of the Technical Appendix, suggested that the correlations between transformed variables remained relatively similar, but a bit stronger than that between un-transformed variables.

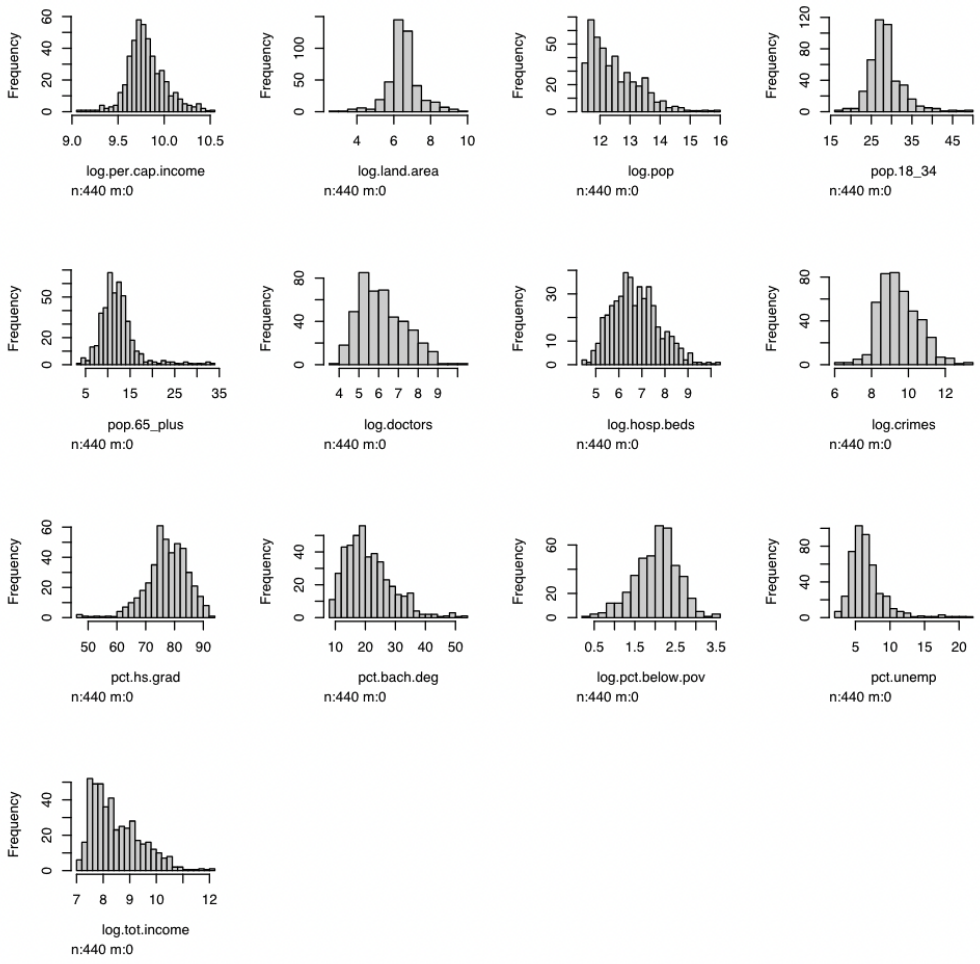


Figure 3: Histogram distributions of quantitative variables after transformation

To start off, page 22 of the Technical Appendix shows the coefficient summary of the full model including all the quantitative variables plus the region categorical variable (note that *id*, *county*, *state*, *pop* and *tot.pop* were excluded). The resulting R squared value was 0.8394, meaning that 83.94% of the total variability of the response variable could be explained by the model. It also seemed like 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 sign.

Table 4 shows the VIF of each predictor variables. The full model (with all variables) suffered from multicollinearity with some predictors having large VIFs that exceed 5. In particular, *log.doctors* and *log.hosp.beds* had VIFs of 15.3 and 12.1 respectively, and this makes sense because the more doctors you have in a county, the more hospital beds you would expect to see. *log.crimes*

is also a predictor with a high VIF of 6.24.

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 predictor variables

Variable selection - all-subsets

We now look at the results for the all-subsets variable selection method on the model.

In Figure 4, we can see a graphical summary of the variable subsets chosen by all-subsets method and the corresponding BIC values. The dark squares indicate which variables are included in the model that has the BIC value on the left. Eventually, the all-subsets method chose 6 variables that gave the lowest BIC value of -747.68 (page 24 of Technical Appendix).

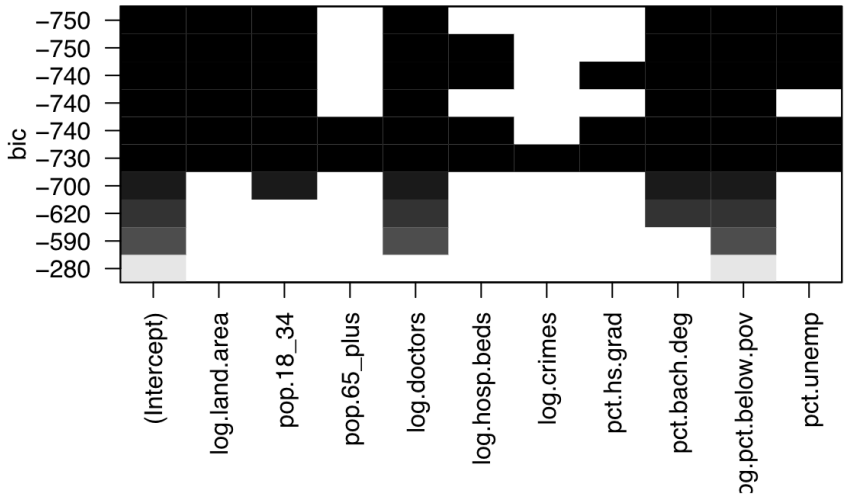


Figure 4: All-subsets graphical plot

In Table 5, we can see the coefficient summary of the model using the variables chosen by the all-subsets method. The R squared value turned out to be 0.834, meaning roughly 83.4% of the total variability of our response variable could be explained by the model. All coefficients were also statistically significant with low p values. But at the same time, the coefficient estimates seemed to be quite small.

	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 25 of the Technical Appendix, we could see that none of the chosen 6 predictor variables had an excessively large VIF, signaling that multicollinearity was no longer an issue.

The residual diagnostic plots for the all-subsets model in page 26 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 27 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 28 and 29 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.

Variable selection - stepwise BIC

We now look at the results for the stepwise variable selection method on the model. 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 32 and 33 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.

Variable selection - LASSO regression

We now look at the results for the LASSO regression method on the model. Figure 5 shows the plot of Mean-Squared Error (MSE) vs Log(λ), where the dotted lines show the optimal number of

variables with the lowest MSE.

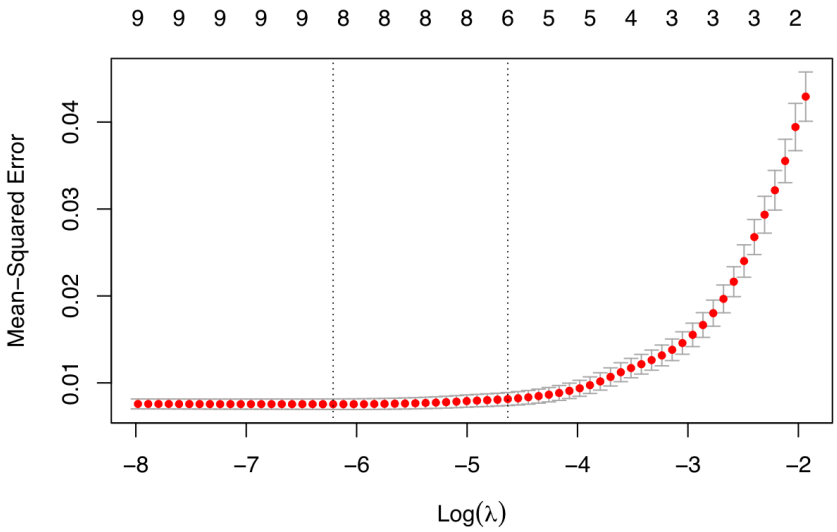


Figure 5: MSE vs Logλ

Table 7 below shows the values for *lambda.min*, the best λ value found by cross-validation, and *lambda.1se*, the value of λ that is 1 standard deviation larger than *lambda.min*. We chose to use the λ value of *lambda.1se* (0.0097), since it could protect against capitalization on chance.

lambda.1se	lambda.min
0.009740676	0.002003182

Table 7: lambda.min and lambda.1se

Table 8 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 38 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 8: lambda.min and lambda.1se

The only difference between the all-subsets method and LASSO regression was that all-subsets chose to include *pct.unemp* rather than *pop.65_plus* which LASSO regression did. Page 39 of the Technical Appendix shows the F test (ANOVA) result to determine which of the two variables is significant, given the other 5 common variables are fixed. We can see that the all-subsets model with *pct.unemp* turned out to be more significant with a low p value.

Adding 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 for an additive term as well as interaction terms.

Page 30 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 interaction terms and categoric variables if any of the indicators for the specific categorical variable was statistically significant. If none were significant, we decided to drop the whole group of interactions.

Table 9 below shows the resulting model with the added *region* dummy 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, and the coefficients still remained quite small in magnitude.

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

To further justify the use of the *region* variable as well as the chosen interaction terms, the F test (ANOVA) was performed on three models. Page 31 and 32 of the Technical Appendix suggested

that the model with some *region* interaction terms added is statistically significant with a very low p value, and was better than the base all-subsets model as well as the base model with only the additive *region* variable. Furthermore, the AIC for the new model with some interaction terms turned out to be lower than that of the base all-subsets model, while the BIC value was the opposite in result. This is interpretable since BIC tends to favor simpler models , while AIC favors more complex models in theory.

Interpreting the final model

The chosen final model was the base all-subsets model with an additive *region* and some interaction terms added as shown in Table 9. Although more complex than the base model, the categorical variable *region* was proven to be fairly important, as the median per-capita income was observed to be fairly different across the 4 regions (Figure 2). The interaction terms are 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, AIC and BIC values also suggested that the interaction terms were valuable when predicting per-capita income.

Along with the final model's coefficient summary in Table 9, we also looked at diagnostic plots produced in Figure 6 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.

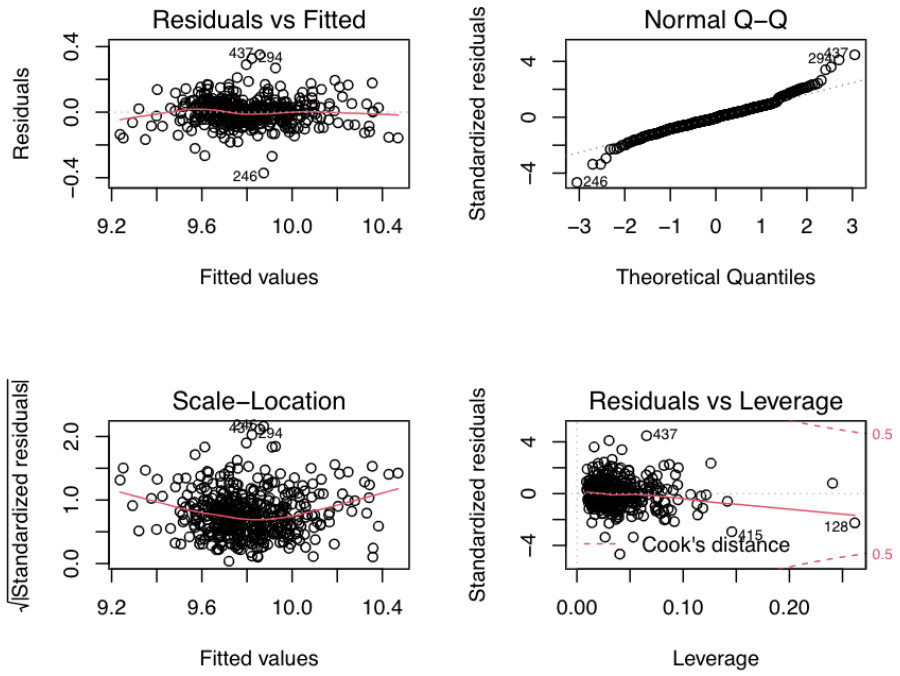


Figure 6: Residual diagnostic plots for final model

We also looked at the plot of Y (*log.per.cap.income*) vs the fitted values $\hat{Y}$ produced in Figure 7. 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.

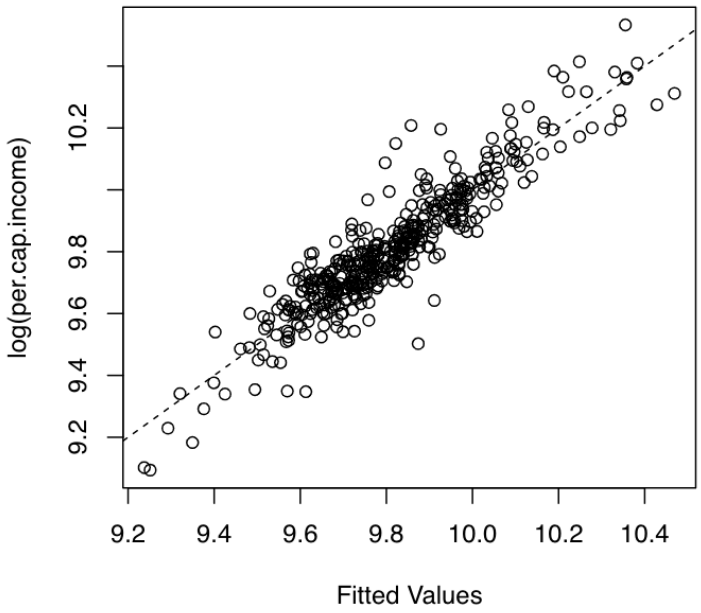


Figure 7: Plot of Y vs $\hat{Y}$

Finally, we attempted to interpret the resulting coefficients of the final model (Table 9):

- For every 1% increase in a county's land area, there is a 0.03% decrease in expected per-capita income. (We might conjecture that this could be due to an urban-rural contrast: rural counties tend to be biggerthan 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 expected2% 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).
- The percent of population that are high school graduates doesn't have much effect, except in the South, where a one percentage point increase in hs 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.

Discussion

1. Relationship between each individual pair of variables

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

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 be differ in different regions of the country, we looked at 3 different models including an additive region and interaction terms with region.

Looking at the summaries of the three models (page 8 and 9 of the Technical Appendix) we were able to recognize a positive correlation between per-capita income and total crimes percentage-wise. Overall, the log transformations on the 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 (base model plus additive region variable) turned out to be the most significant according to the F test.

We also attempted to see if changing the crimes variable to per-capita crimes helped in any way, because per-capita crimes would be in the same comparable scale as per-capita income, thereby potentially leading to better interpretability. However, changing the variable this way resulted in a significant decrease in R squared value, meaning that the model using per-capita crimes was not able to explain the variability of per-capita income as well as the model using the raw total crimes. The AIC and BIC values also seemed to increase, suggesting that the trade-off between interpretability and model fit was not equal in value when using per-capita crimes. As a result, we could see that sticking to the raw crimes variable was the better choice.

Due to the fact that all the interaction terms in the third model did not turn out to be significant, we could also conclude that the relationship between per-capita income and total crimes did not differ significantly in different regions of the country. However, we were able to figure out that the additive region variables themselves were significant enough to be valuable in the model when none of the interaction terms were involved.

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

In order to address and solve the problem of non-linearity and non-normality of the variables, we performed logarithmic transformations on certain quantitative variables. This in turn improved a lot of the problems of skewness, while keeping the correlation between variables roughly unchanged. At the same time, we 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 and knowledgeable about 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 regression, stepwise regression and LASSO regression - to counter this. All three methods produced similar optimal subsets of significant variables that gave a minimum value of BIC when fitted into a model. But through an F-test of overall significance, we were able to find out that the variable subset chosen by all-subset regression and stepwise regression produced the best fitting valid model. It was understandable that the variable *log.hosp.beds* was eliminated in all three methods due to its high collinearity with *log.doctors*.

All three methods also chose to exclude *pop.65_plus*, since it would probably have been highly correlated to its counterpart variable *pop.18_34*. 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, given all other quantitative variables were ignored. 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 that was previously hypothesized to be an important indicator of per-capita income. With the additive term and some 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 variable (in the interaction) 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 the expected sign.

Limitations and future work

One of the evident limitations 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 suggests 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 more complex models that could improve the validity of the model. The usefulness of some interactions terms including the region categorical variable is another evidence that there could be unidentified interaction terms that could additionally enhance the model.

Another limitation that needs to be addressed is that only 440 counties (including those that have duplicate county names in different states), were considered out of the total of approximately 3,000 counties in the US. Since we are only working with a certain sample of a population, there is always the possibility that the data is biased and is 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. The entire analysis in this report is written with the assumption that random sampling was performed on the data and that the presented results could be used to infer about the population. If given additional time, we could possibly investigate further on how 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, or per state, to see if there are any large deviations in the summaries.

Another area we could look into further would be the *state* categorical 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 are entirely nested

within regions (perfect collinearity), 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 data (more counties) to use as tests sets that could compare some of the models we found. We are using reasonable methods for variable selection, but since our entire data set is in fact our training sample, there is a big possibility 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, for example. If we were able to cross-validate on some new or hold-out data, we might be able to better distinguish the best models, at least 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.)

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)
```

Looking at the plot below showing how many unique values there are for each variables, *county* is a categorical variable that has 373 values (almost equal to number of rows). *state* is another categorical variable that has 48 values, which is a lot, so I would set this variable aside during my analysis. *id* of course is another categorical variable that is just the same as the number of rows, so I will exclude it from my analysis.

```
apply(dplyr::select(cdi, id,county,state,region),2,function(x) {length(unique(x))}) %>%
  kbl(booktabs=T,col.names="unique values",caption=" ") %>%
  kable_classic(full_width=F)
```

Below is the summary statistics for the categorical variable *region*, which I will be using for my analysis. It can be observed that most of the counties are in the South region, while the least are in the West region. The low number of counties in the West region could be indicative of under-sampling or that the land is larger so there are fewer counties to sample from. The high number of counties in the South region could be indicative of over-sampling or perhaps the South has a lot of counties covering smaller land areas.

```
# Summary statistics for categorical variables
tmp <- rbind(with(cdi,table(region)))
row.names(tmp) <- "Freq"
tmp %>% kbl(booktabs=T,caption=" ") %>% kable_classic(full_width=F)
```

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.

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 2:

	unique values
id	440
county	373
state	48
region	4

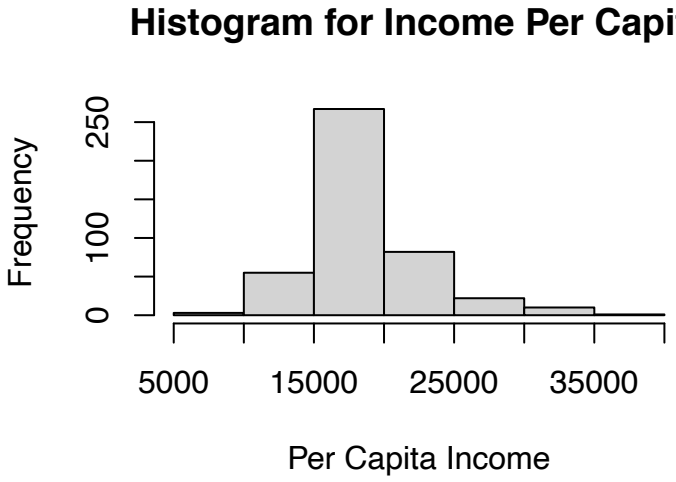
Table 3:

	NC	NE	S	W
Freq	108	103	152	77

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

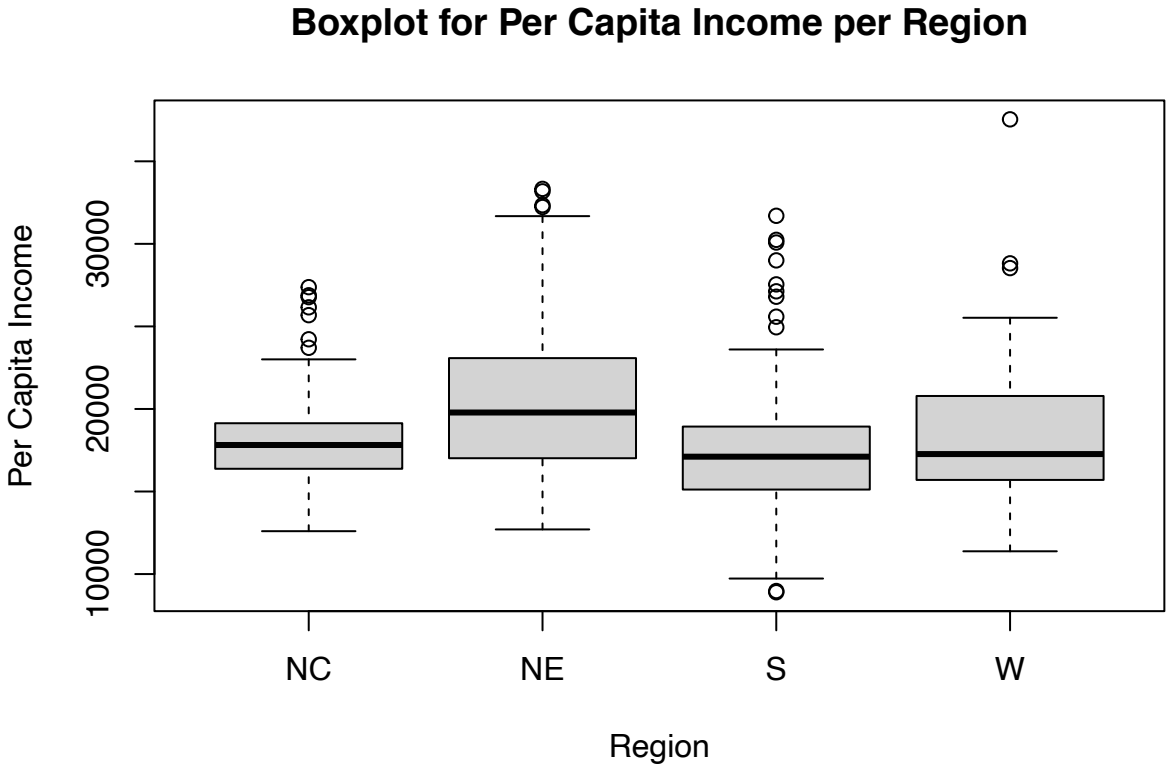
```
##             id         county          state      land.area            pop
##          FALSE          FALSE          FALSE          FALSE          FALSE
##      pop.18_34    pop.65_plus        doctors      hosp.beds         crimes
##          FALSE          FALSE          FALSE          FALSE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

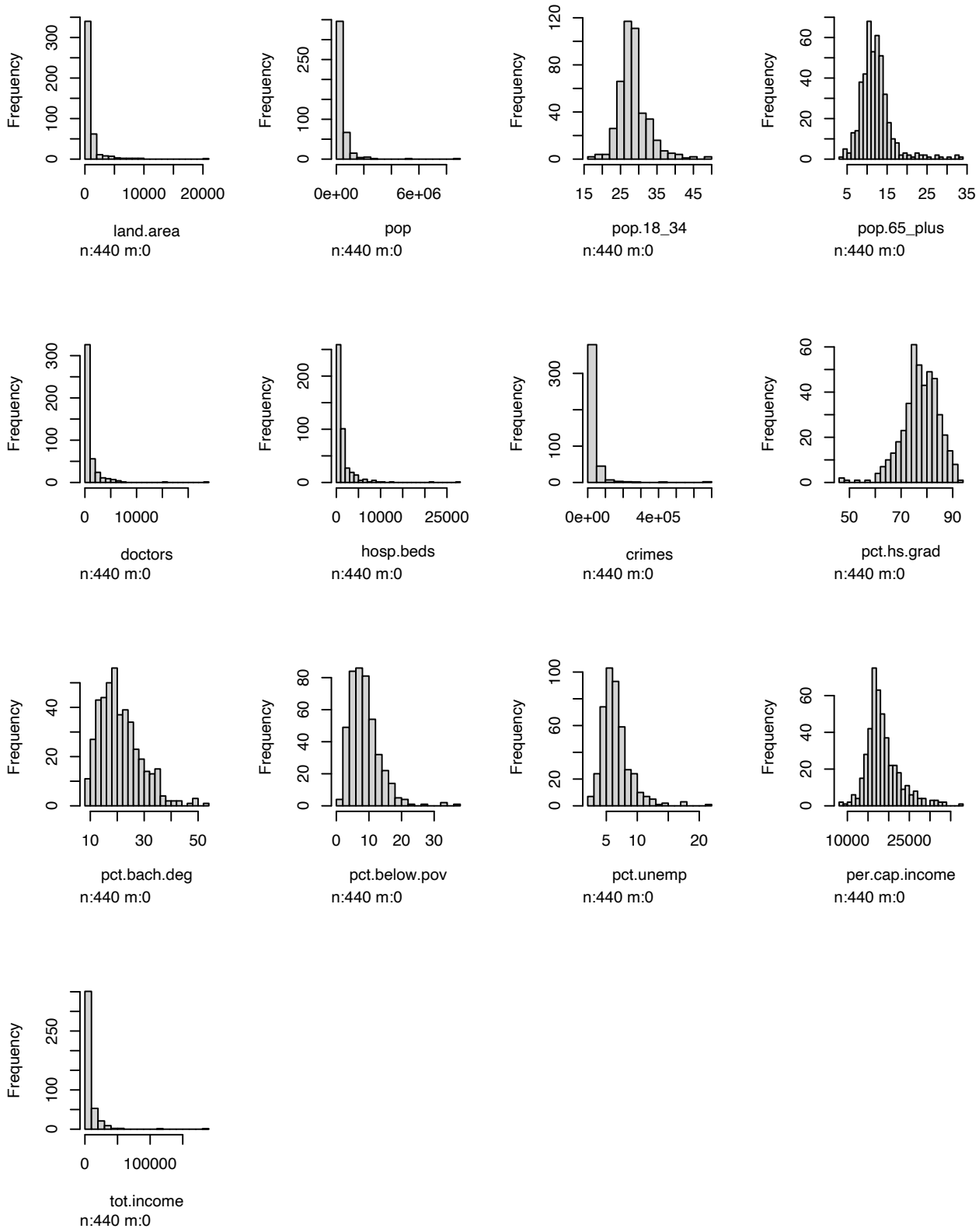


We can also explore if there are any major differences in per capita income in the 4 regions. Upon looking at the plot below, we can deduce that the median from the North East region is the highest, and also has the largest Interquartile Range. Overall, there seems to be some minor difference in median per capita income between all 4 regions.

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



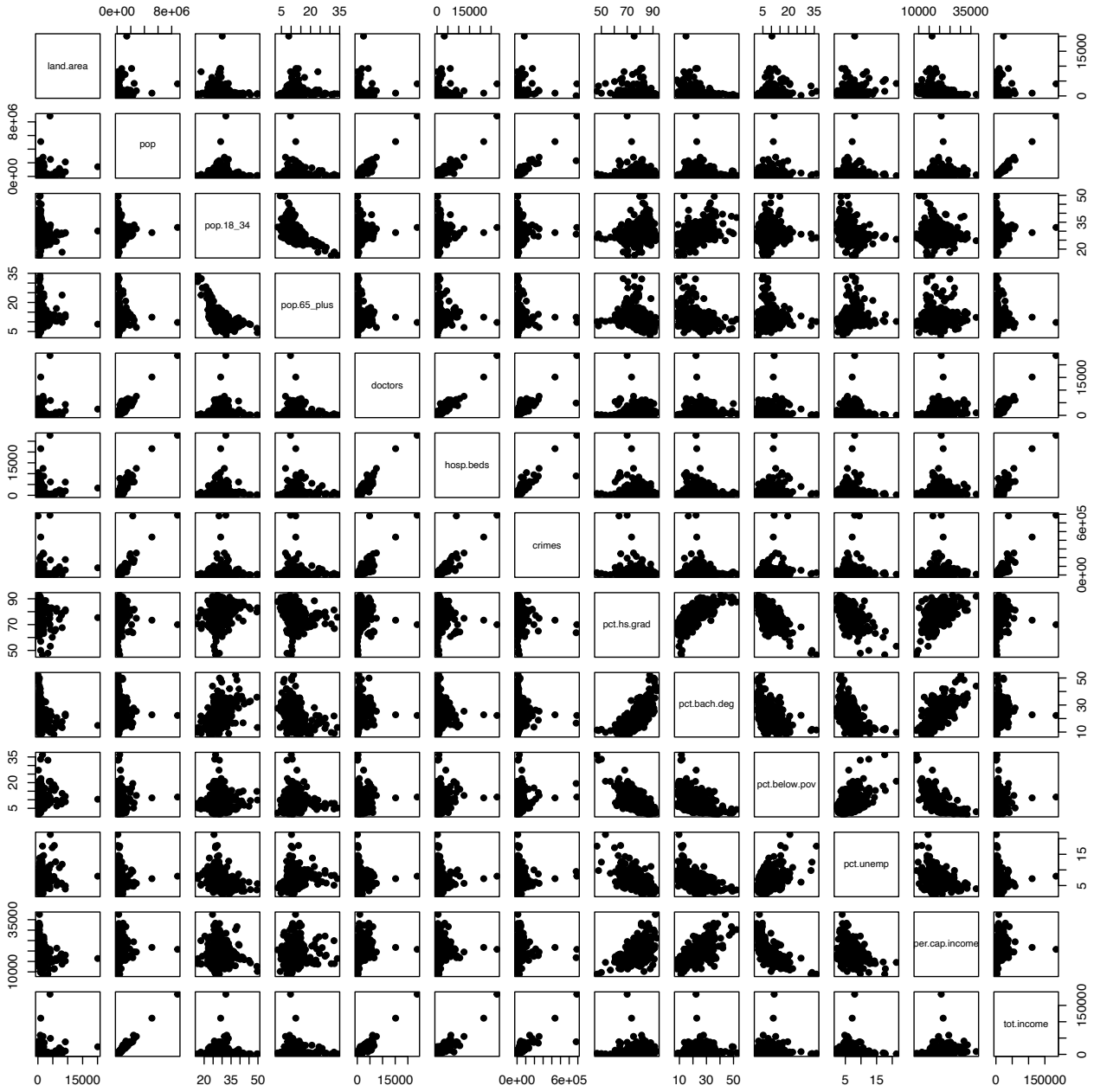
We can also look below at the histogram distribution of other predictor variables. We can observe that there are severely skewed variables like *land.area*, *pop*, *doctors*, *hosp.beds*, *crimes* and *tot.income*. The rest of the predictors are not perfectly normal, but rather seem to be slightly skewed either to the right or left.



/

Next, we can look at a scatter plot matrix to identify overall relationships between the variables. Our response variable *per.cap.income* seems to be relatively linearly related to some of the predictors including

pct.back.grad and *pct.hs.grad* while some predictors show a curved relationship like *pct.below.pov* and *pct.unemp*. The rest of the predictors show a skewed or random relationship.



We can also take a look at the correlation matrix heatmap to understand if there are any correlations between the predictors.
We can make the following conclusions from the correlation matrix:

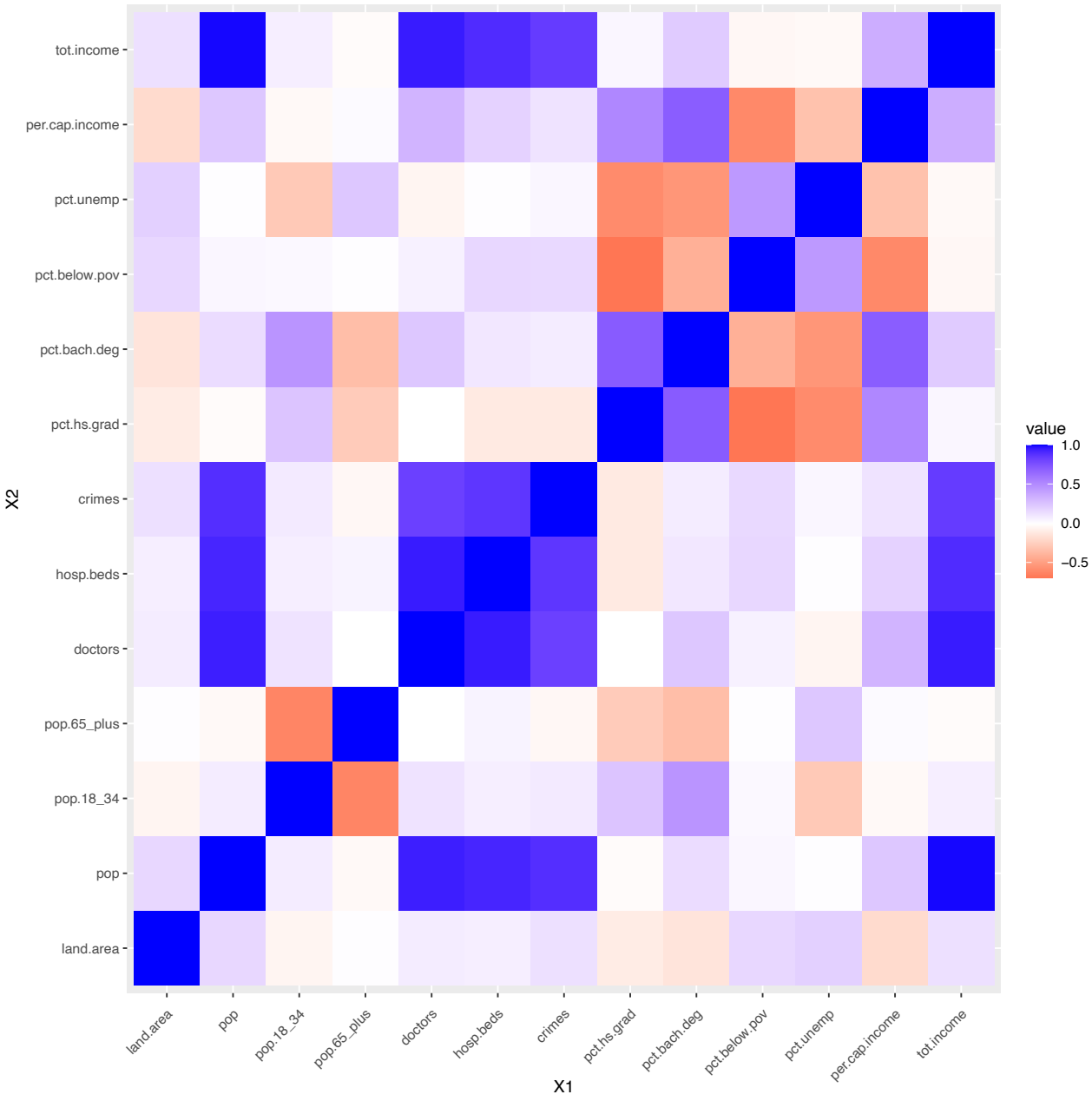
- *tot.income* and *pop* are highly correlated. This is expected because the response variable *per.cap.income* is a deterministic function of *pop* and *tot.income*, where *per.cap.income = tot.income / pop.*
- both are reasonably highly correlated with *crimes*, *hosp.beds* and *doctors*
- *pct.hs.grad* 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 as strong a correlation as others, *pct.hs.grad* and *pct.bach.deg* are negatively correlated to *pct.unemp*, and this 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.
- the three variables *crimes*, *hosp.beds* and *doctors* seem strongly correlated with one another. Although not obvious for *crime*, *doctors* and *host.beds* would be expected to be correlated because the more doctors/physicians you have in a county, the more hospital beds you would expect to see.

These 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 will build a regression model that predicts per-capita income from the crime rate and region of the country. I will be exploring models with and without the interaction term. Note that I will apply a **log transformation** to the *crimes* variable as it is severely right skewed. I will be applying a **log transformation** to both the response variable *per.cap.income* and our predictor *crimes* to make them more normally distributed.

```
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 )
```

Let us first look at the summaries of both 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
```

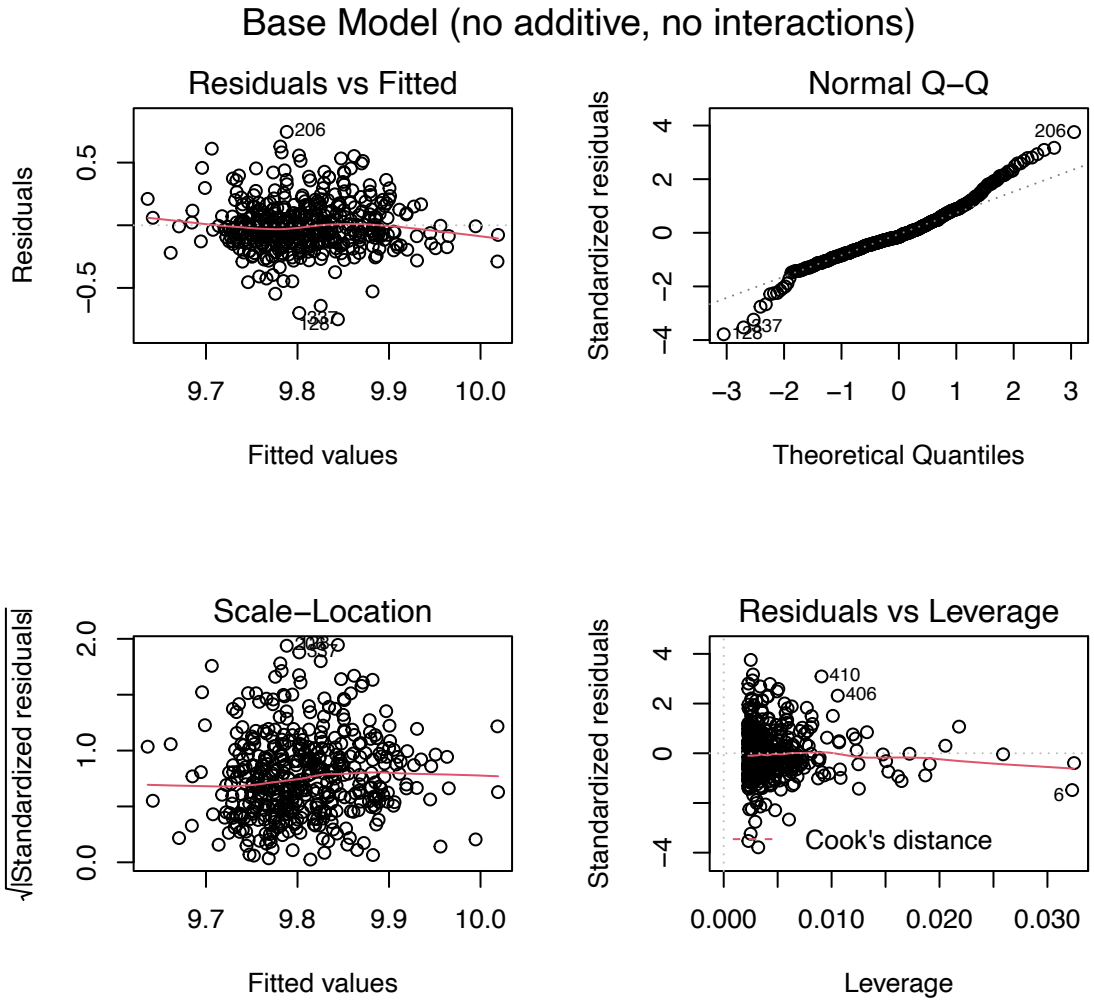
For model 1 (without region nor interaction terms), we can see that the R squared value is 0.079, meaning that 7.9% of the total variability in our response variable can be explained by the model. All coefficients are also statistically significant, with very low p values.

For model 2 (with region and without interaction terms), the R squared value increased to 0.2032. All coefficients are also statistically significant, with very low p values except for the coefficient for *regionW*, which has a p value of 0.05.

For model 3 (with interaction terms), the R squared value increased slightly to 0.2073. Only the coefficient for *log(crimes)* is statistically significant with a low p-value. The rest of the coefficients including all the interactions terms have a high p value.

There seems to be a fairly positive correlation between log(*per.cap.income*) and log(*crimes*). For a unit percentage increase in *crimes*, you would get approximately a 0.07% increase in the *per.cap.income*. This interpretation is possible since we decided to log transform both the response and predictor variable.

Let us now take a look at the residual diagnostic plots for the three models, to see if the models are valid.



Starting from the top-left plot of **Residuals vs Fitted**, we can observe that the residuals are not too far from 0. Also, the residuals do not seem to have a distinct shape along the fitted values, and is approximately equally and randomly spread out around the dashed line. This provides enough evidence that the residuals have a relatively constant variance.

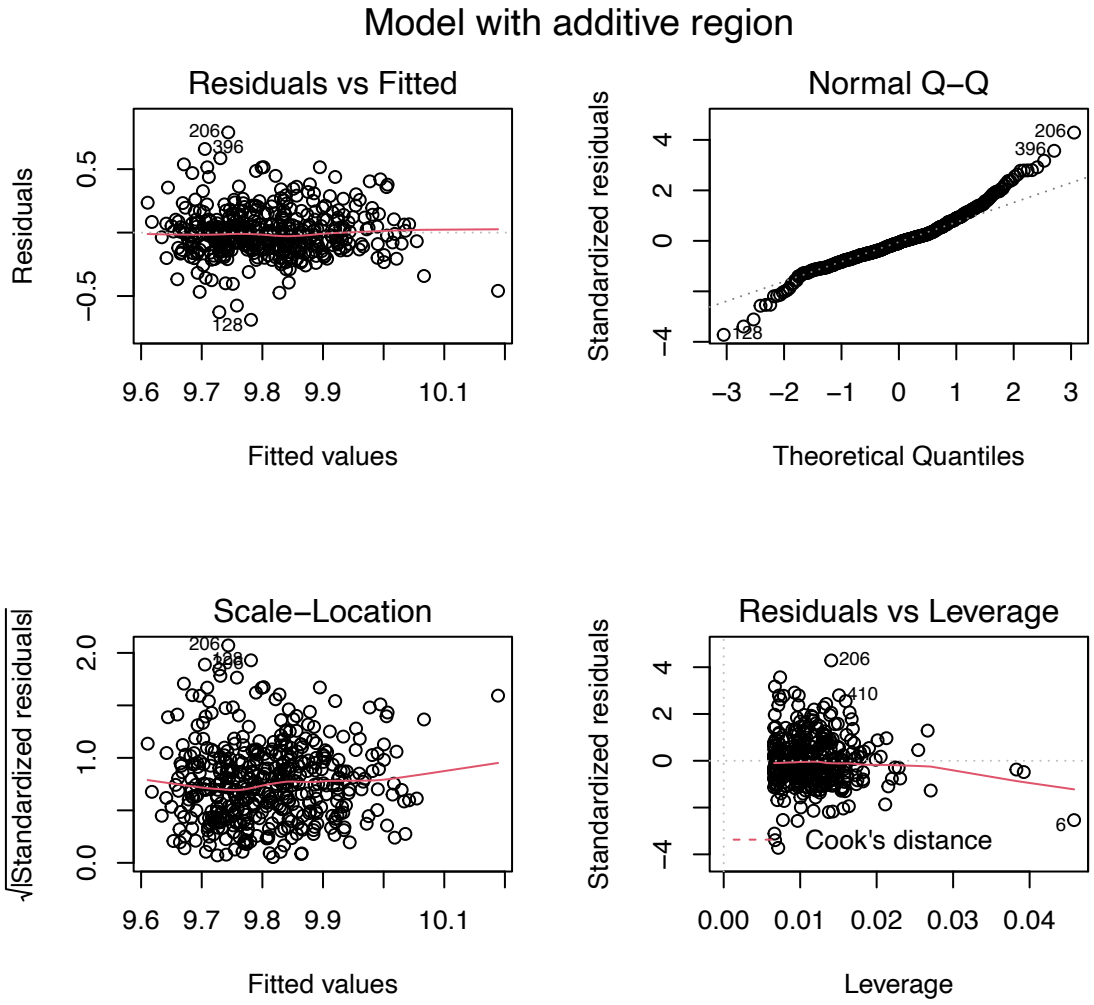
The top-right **Normal Q-Q** plot shows that the residuals seem to follow a straight line without much deviations, except that it has a heavy right tail and a heavy left tail. We can say that the residuals do not deviate heavily from a normal distribution.

The bottom-left plot of $\sqrt{|Standardized\ Residuals|}$ **vs Fitted**, shows that the residuals are spread approximately equally along the ranges of the predictor. The line is relatively horizontal and does not show any specific shape. The residuals are also fairly randomly spread out, suggesting constant error variance.

Finally, the bottom-right plot of **Residuals vs Leverage** shows that there are no points that lie outside the

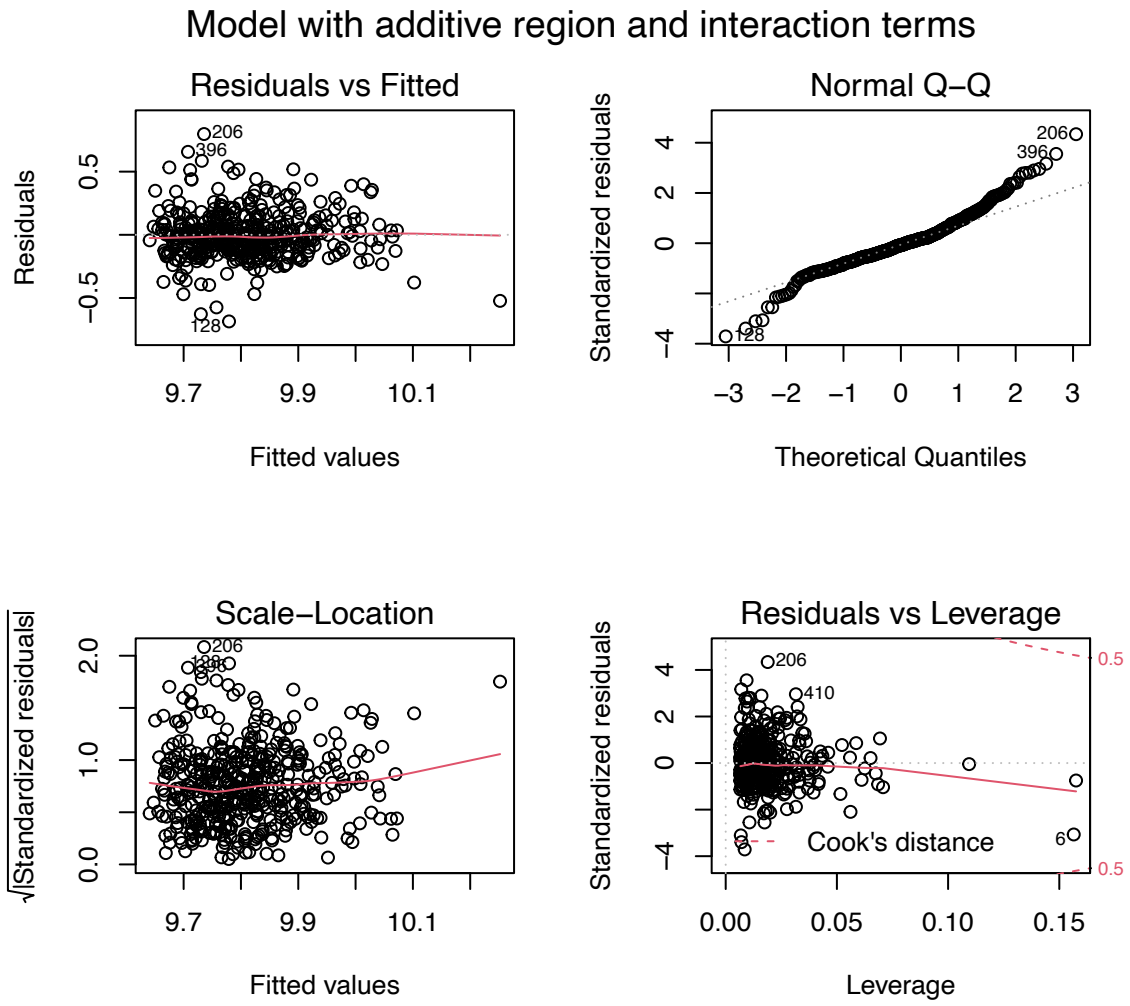
dashed Cook's distance line, suggesting that there are no highly influential points that strongly affect the model. But we do see a few points with standardized residuals absolute value of 2 or higher, suggesting that there are some outliers that we may need to look at closer. I would particularly take a closer look at data point 6, as it is a high leverage point and an outlier.

Below is the residual diagnostic plot for the model **with the interaction terms**.



The residual diagnostic plots do not show much of a difference compared to that of the model without interaction terms. Residuals seem to have relatively constant variance, QQ plot shows that the residuals are fairly normal around the center but is has quite a heavy right tail and a slightly heavy left tail. Finally, we see a few outliers where the standardized residuals have values larger than the absolute value of 2, but in general we do not see any extreme/highly influential points outside the Cook's distance.

Now lets look at the residual diagnostic plot for the model **with the 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.

Below is the result of the F-test. It looks like Model 2 ("additive" model with no interaction terms) is statistically significant with low p-value, and is doing the best. As a result, I will **not** be using the interaction terms in my model, but will still use *region* as an additive term.

```
## 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
```

Using model 2 as our optimal model, we can interpret the following: * All across the U.S, for every 1% increase in crimes, we expect an increase in 0.07% increase in per capita income. * Different regions of the country have different baselines for per capita income. NC region has $exp(9.19) = \$9798.65$, NE has $exp(9.19 + 0.010) = \$10829.18$, and so forth. S has \$8955.29, and W has \$9228.02. All of the region baselines are, according to the model, significantly different from the NC baseline.

Per capita crimes

I will now attempt to see whether my answer changes when I change the *crimes* variable to "per-capita crimes" (number of crimes / population). 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 )
```

Similar to before, let us start by looking at the summaries of both 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
```

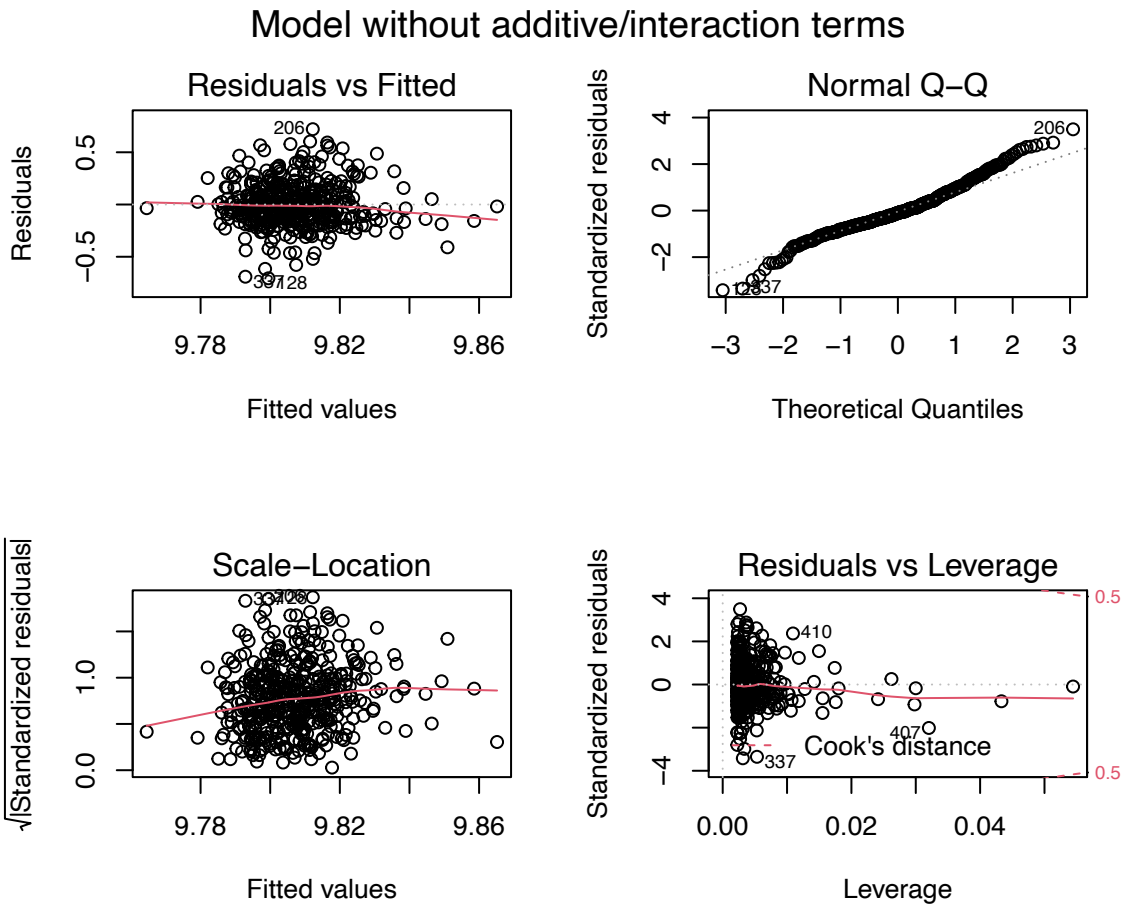
```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(per.cap.crimes) + region,
##     data = cdi)
##
```

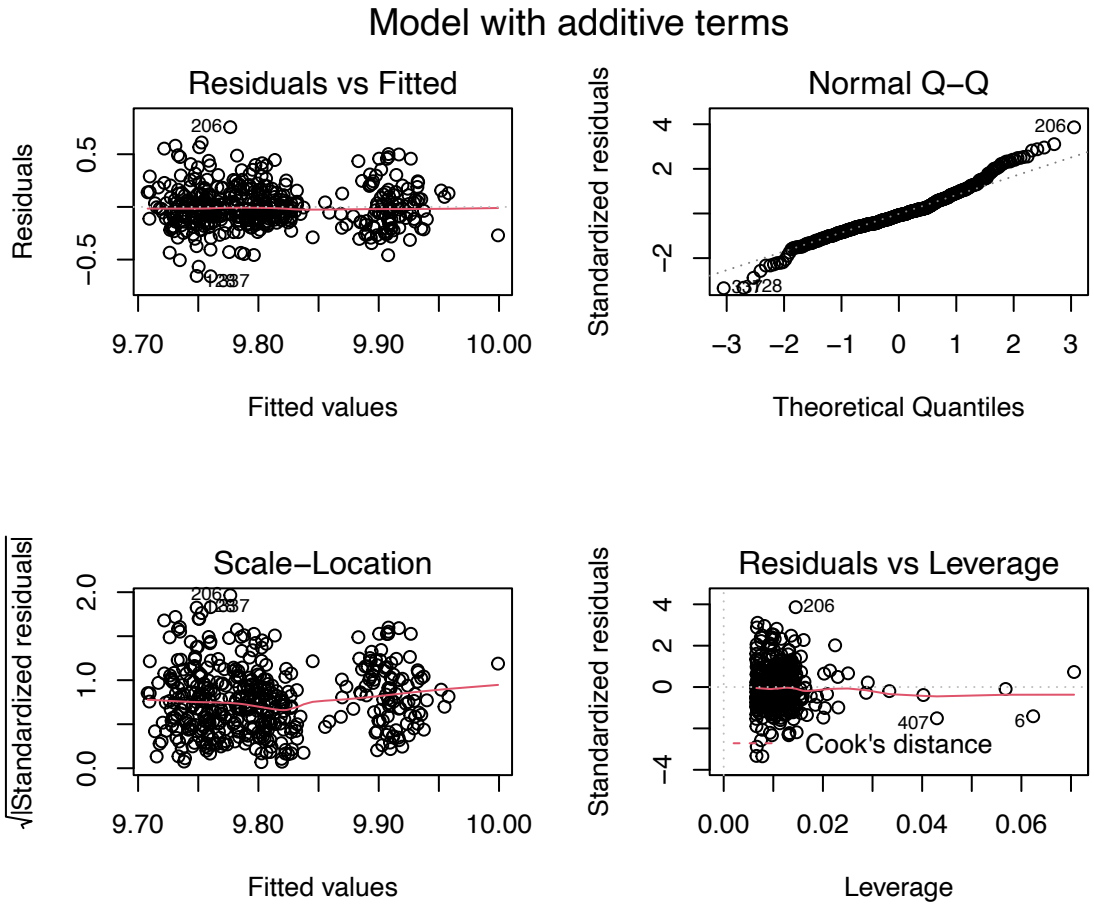
```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65832 -0.11431 -0.01548  0.10838  0.75657
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          9.93628    0.06934 143.303  < 2e-16 ***
## log(per.cap.crimes)  0.04243    0.02148   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
```

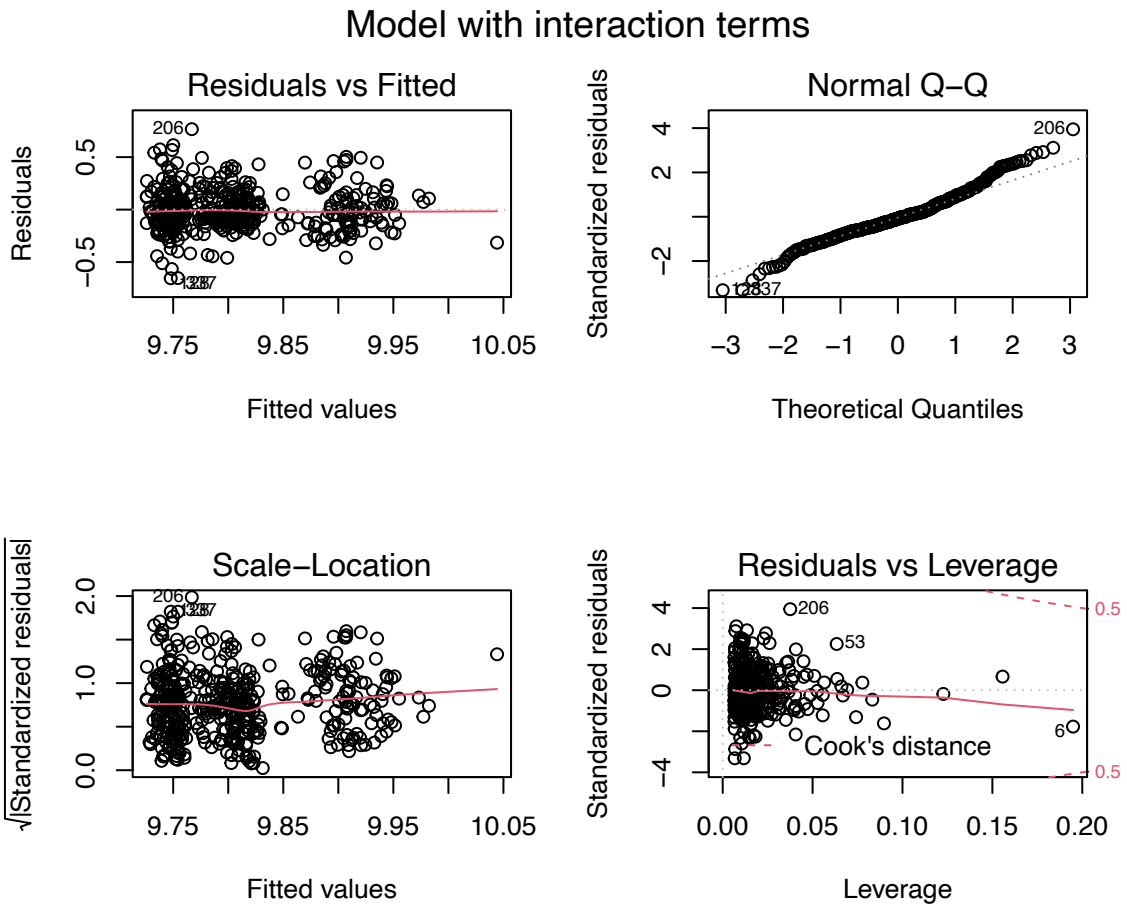
```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(per.cap.crimes) * region,
##     data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65410 -0.11829 -0.01708  0.10399  0.76628
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (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
```

Overall, the R squared values decreased significantly compared to the models using the raw *crimes* predictor variable. We also notice that for the second model (with additive), the coefficient for the *log(per.cap.crimes)* variable was very close to being not statistically significant. This suggests that the model became more reliant on the dummy variables rather than the crime rate variable. For the third model (with interaction terms), we see that none of the coefficients are statistically significant. The residual standard error for the models also show an overall increase when compared to the previous 3 models using the raw *crimes* variable.

Now, lets look at the residual diagnostic plots for the three models to assess their validity:


Model without additive/interaction terms
Residuals vs Fitted
Residuals
Fitted values
Normal Q–Q
Standardized residuals
Theoretical Quantiles
Scale–Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage


Model with additive terms
Residuals vs Fitted
Residuals
Fitted values
Normal Q–Q
Standardized residuals
Theoretical Quantiles
Scale–Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage



All three models have similar residual diagnostic plots. Starting from the top-left plot of **Residuals vs Fitted**, we can observe that the residuals are not too far from 0. Also, the residuals are approximately equally and randomly spread out around the dashed line. But we do see that the residuals are shaped into 3 distinct groups. This provides enough evidence to question the validity of the model although the residuals have constant variance.

The top-right **Normal Q-Q** plot shows that the residuals seem to follow a straight line except for the heavy right tail. This suggests that the residuals show no big departures from normality.

The bottom-left plot of $\sqrt{|Standardized\ Residuals|}$ **vs Fitted**, shows that the standardized residuals are spread approximately equally along the ranges of the predictor. But similar to before, we can observe a distinct shape of the standardized residual, seemingly grouped into 3 distinct categories (probably due to the *region* dummy variables being significant).

Finally, the bottom-right plot of **Residuals vs Leverage** shows that there are no points that lie outside the dashed Cook's distance line, suggesting that there are no highly influential points that strongly affect the model. But we do see several points with standardized residuals absolute value of 2 or higher, suggesting that there are some outliers that we may need to look at closer.

To conclude, the three new models do not show a significant change in terms of the residual diagnostic plots (except for the 3 groups shown in the residual plots), and could be considered fairly valid due to its constant error variance. The coefficient for *per.cap.crimes* seem to be no longer highly significant as before and the residual standard error has increased when compared to the model using the raw *crimes* variable. Changing

crimes to *per.cap.crimes* is not a good idea. But I would actually use the human intuition that using Per Capita Crimes could be better, since our response variable is Per Capita Income. It would make sense to be consistent with the response variable and use Per Capita Crimes, and this would especially be easier to interpret when explaining to clients or collaborators. Also, since we are only looking at two variables right now (*crimes* and *region*), I would be more willing to accept the tradeoff and follow what the statistical result is telling me.

Let us use AIC and BIC to compare between Model 2 and Model 5 to see if this is true.

```
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
```

It turns out that model 2 is still better with smaller AIC and BIC values.

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

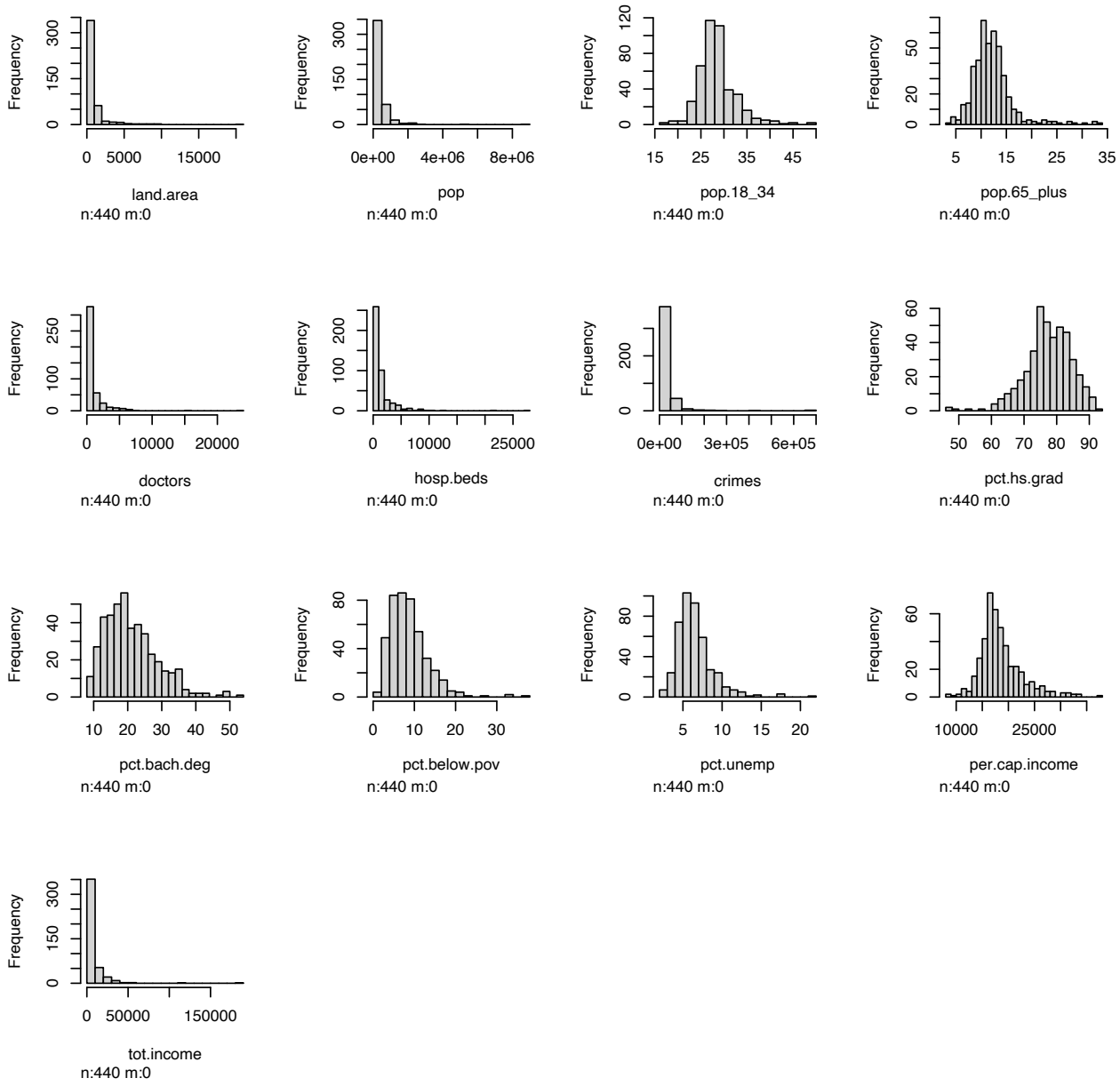
Research question 3

I will be dropping the *id*, *county* and *state* column and will be focusing on the rest of the 13 predictor variables. *id* column is just an incremented number from 1 to 440, *county* also has 373 unique values, which makes it not desirable to have when doing regression analysis. I also chose to exclude *state* for now as it has 48 unique values and thought it wouldn't add much to explaining the variability of our response variable as much as the other predictors could.

```
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

Let us revisit the histograms I generated for non-dummy variables in **Research question 1**



A lot of the variables seem to be skewed to the right, while some seem to be either slightly skewed or relatively normally distributed. I will be performing log transformations for those variables that are skewed, but will leave the others as is. It is easier to explain logarithmic transformations to a social scientist in terms of percentage-change concepts. The more untransformed the variables are, the easier it will be for the social scientist to think about the models I present. I will perform log transformations on the variables below:
* per.cap.income * land.area * pop * doctors * hosp.beds * crimes * pct.below.pov * tot.income

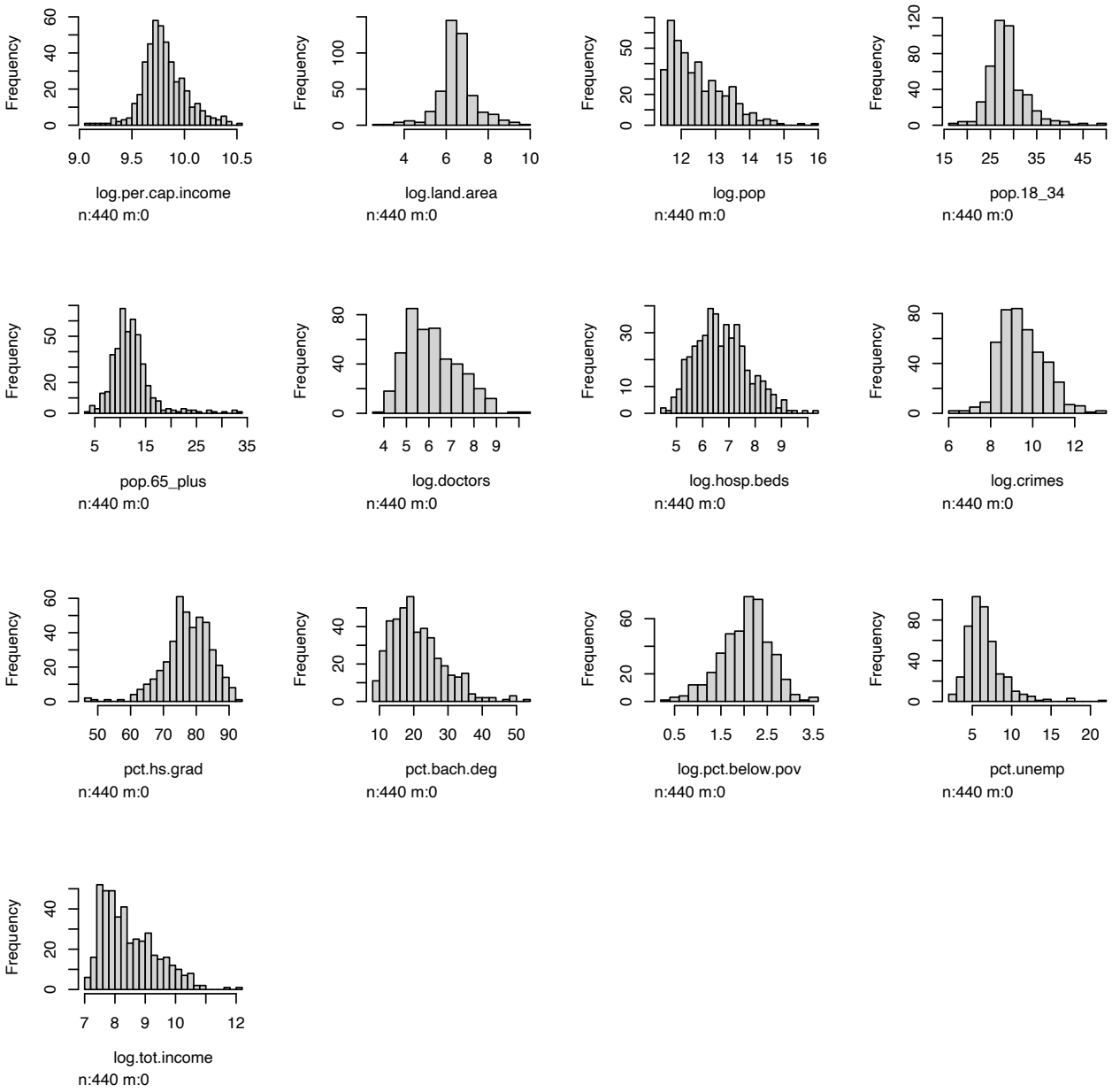
```
# Transform the variables
cdi_transformed <- cdi_new

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="")
}
```

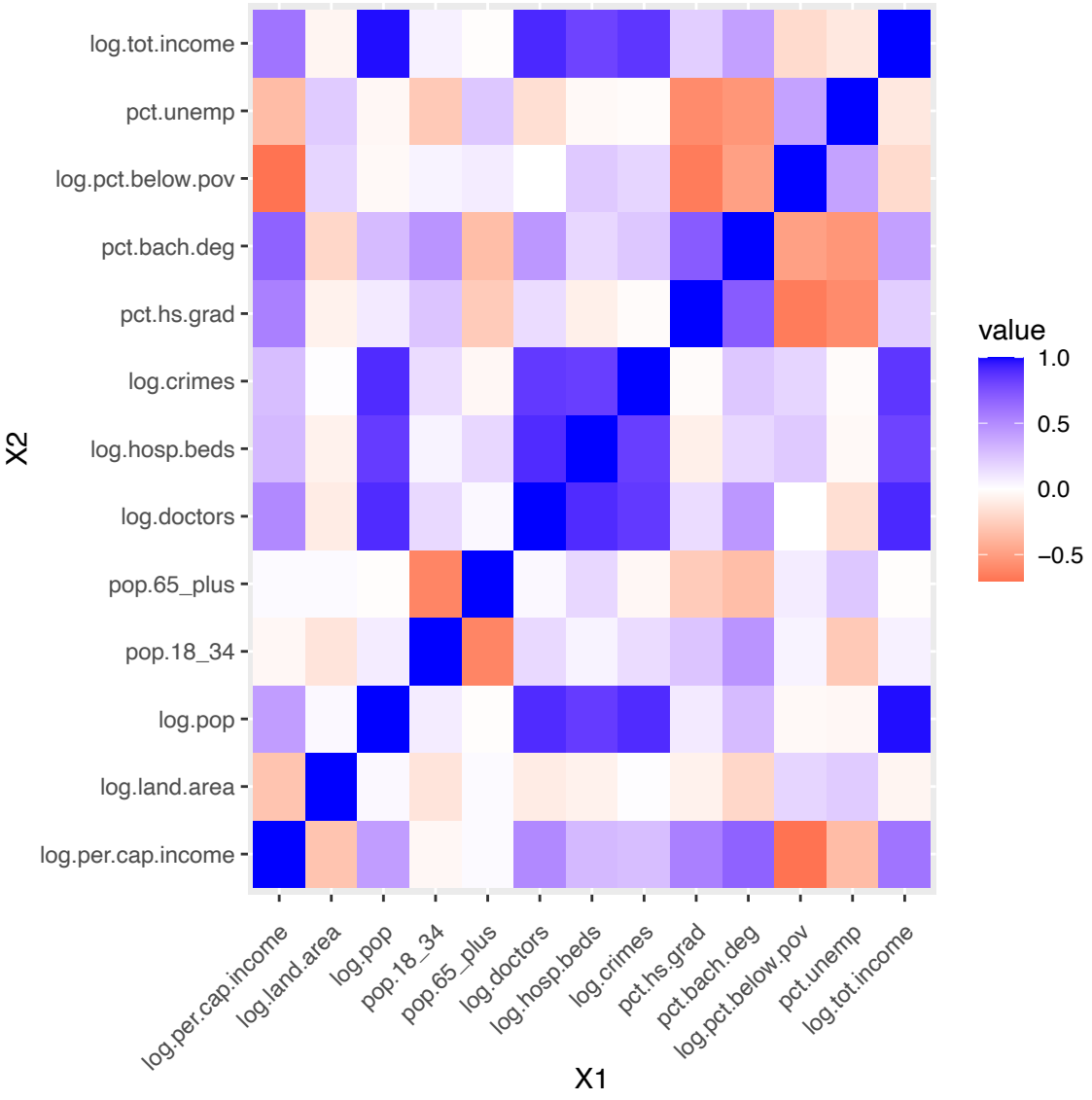
Let us take another look at the histogram plot for the variables below. It seems like a lot of the skewings have been improved except for *log.pop* and *log.tot.income*. Note that I will not be using *log.pop* nor *log.tot.income*, since our response variable *per.cap.income* is a deterministic function of *pop* and *tot.income*.



Let us quickly check the correlation matrix heatmap again to see if anything changed. The correlations seem to be similar, but a bit stronger than the correlations for untransformed 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

Let us now fit a full model including all the variables including the *region* categorical variable (changed to factor). Note that I have excluded the *log.pop* and *log.tot.income* variables since *log.per.cap.income* is a deterministic function of the two variables, and we won't be able to learn about anything associated with *log.per.cap.income* with these predictors included. The summary can be seen below. We can see that the R squared value is 0.8394, meaning that 83.94% of the total variability of the response variable can be explained by the model. It also seems like some predictors are statistically significant with low p values, while some are not significant. Predictors like *pct.hs.grad* and *pct.unemp* seem to have the wrong coefficients with opposite signs. These are signs that variable selection methodologies would help.

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

```r
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

My initial hypothesis is that there would be highly collinear predictor variables in our full model, and that some of them would have to be dropped. Let us first look at the Variance Inflation Factors (VIF) for each of the predictors below.

```
##     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
```

Looking at the VIF's, the model seems to suffer from **multicollinearity** with some predictors having large variance inflation factors. A number of these variance inflation factors exceed 5, the cut-off often used, and so the associated regression coefficients are poorly estimated due to multicollinearity.

The predictors *log.doctors* and *log.hosp.beds* also have fairly high VIF values of 15.3 and 12.1 respectively. This also makes sense because the more doctors you have in a county, the more hospital beds you would expect to see.

The model suffers from multicollinearity, but is not too bad as originally thought. It would however make sense to carry out variable selection methodologies to select the best variables for the model. Note that the categorical predictor *region* is not considered for now, because none of our variable selection methods are capable of dealing with the group of indicator variables associated with a categorical variable. I will come back to this variable later.

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*.

```
region_var <- cdi_transformed$region

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

names(cdi_transformed)
```

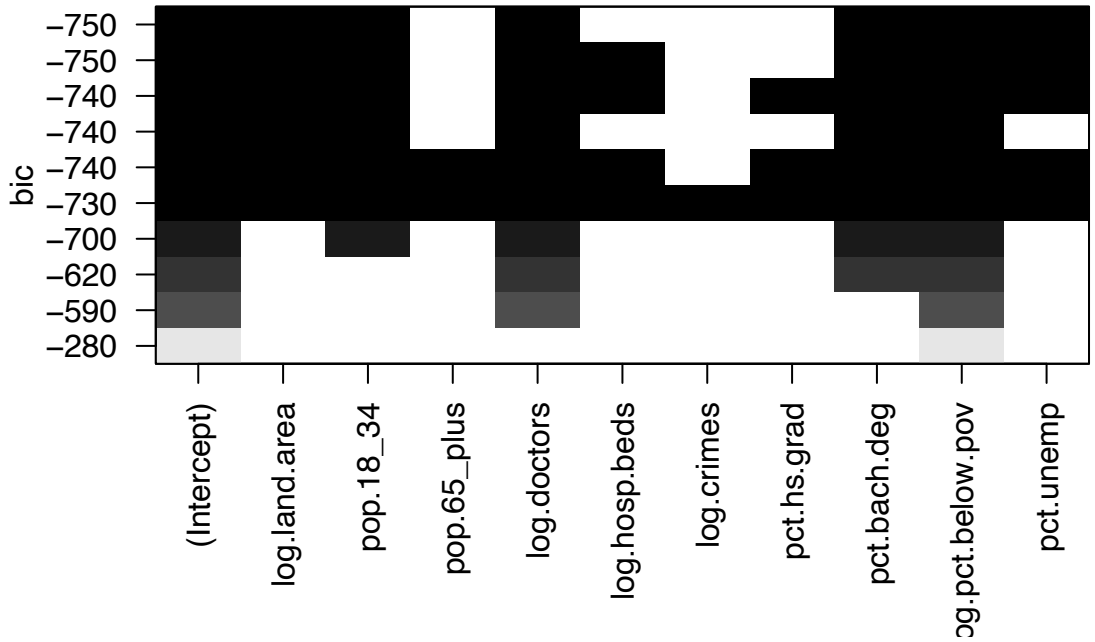
```
##  [1] "log.per.cap.income" "log.land.area"      "pop.18_34"
##  [4] "pop.65_plus"        "log.doctors"        "log.hosp.beds"
##  [7] "log.crimes"         "pct.hs.grad"        "pct.bach.deg"
## [10] "log.pct.below.pov"  "pct.unemp"
```

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

All subsets

I will first try out the all subsets method. 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)
```



Below we can see that the all subsets method chose 6 variables that gave the lowest BIC. We can see the coefficients for each of the variables chosen.

```
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]]
```

Let us explore the summary for these coefficients when fit to a new model below. All coefficients seem to be statistically significant with low p values. The coefficient estimates themselves are quite small, and *pct.unemp* still seems to have the wrong sign.

```
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
```

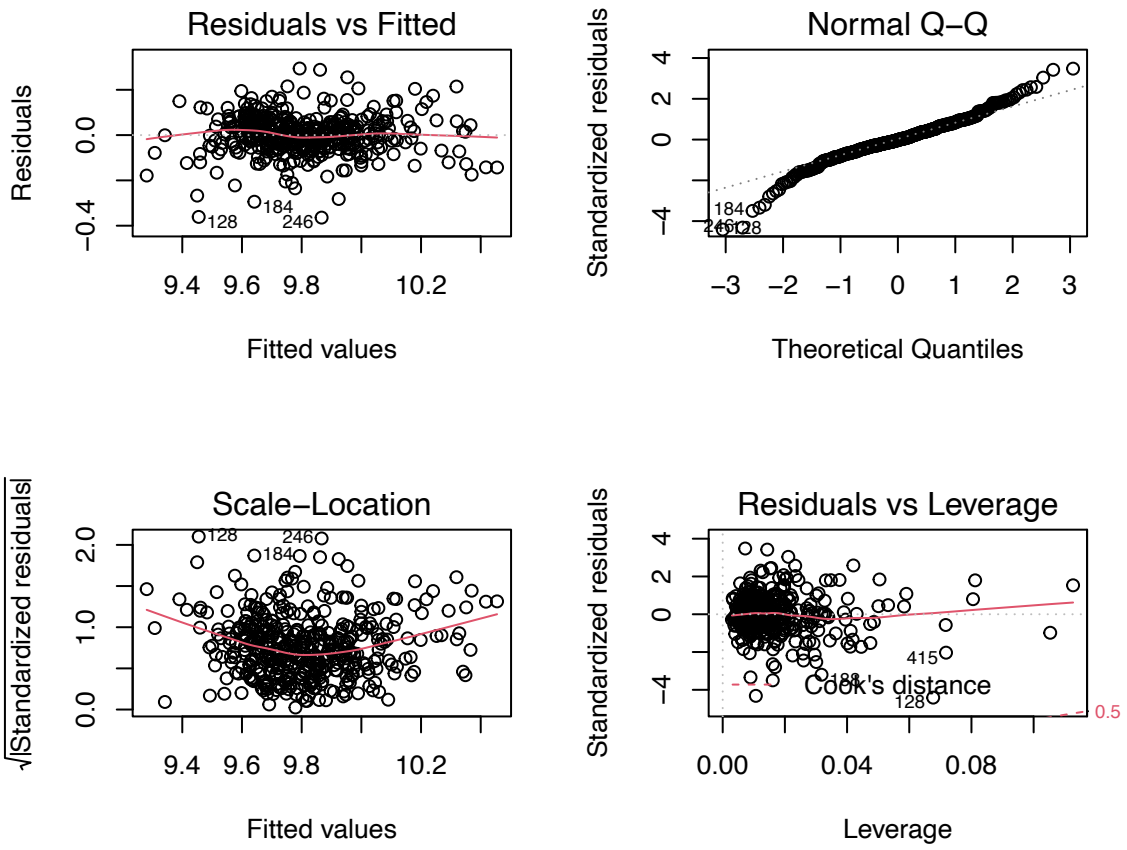
```
##                       Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)       10.095545110 0.0516863528 195.323225 0.000000e+00
## log.land.area     -0.036212594 0.0048198661  -7.513195 3.336311e-13
## pop.18_34         -0.012026824 0.0011883729 -10.120413 9.445833e-22
## log.doctors        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)
```

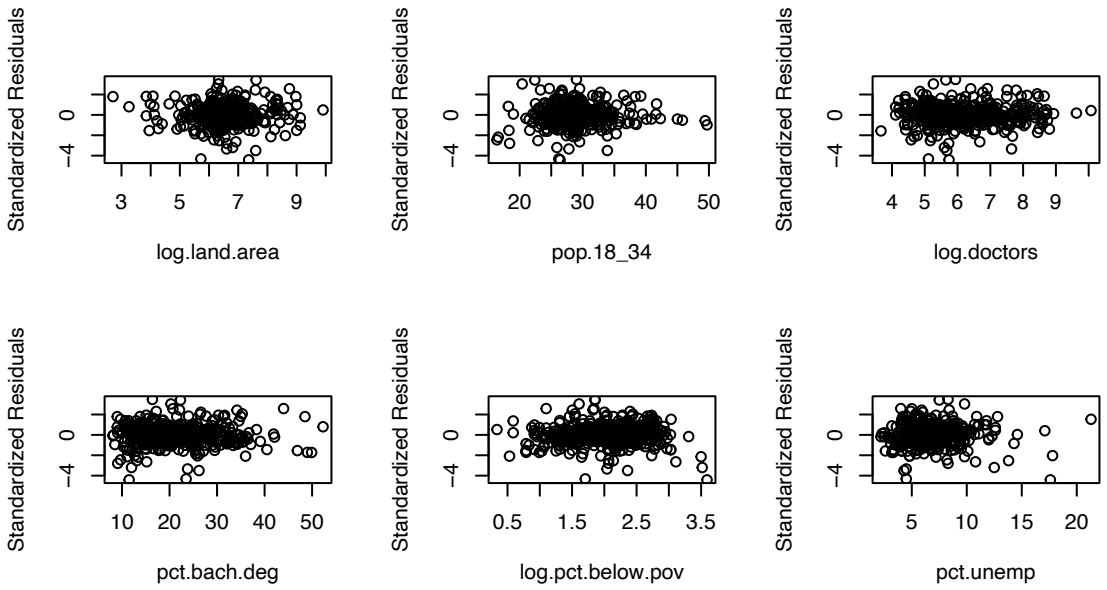
```
##     log.land.area         pop.18_34       log.doctors      pct.bach.deg
##            1.0778            1.5145            1.4052            2.8085
## log.pct.below.pov         pct.unemp
##            1.7505            1.4990
```

Now lets look at the residual diagnostic plots. Except for the fact that the normal Q-Q plot shows that the left and right tails are a little bit heavy, the rest of the plots seem to be ok.



Next, let us take a look at the standardized residual plots against each of the predictor variables below.We notice that the residuals for all the of the plots are relatively randomly scattered. This is good evidence that the model is valid.

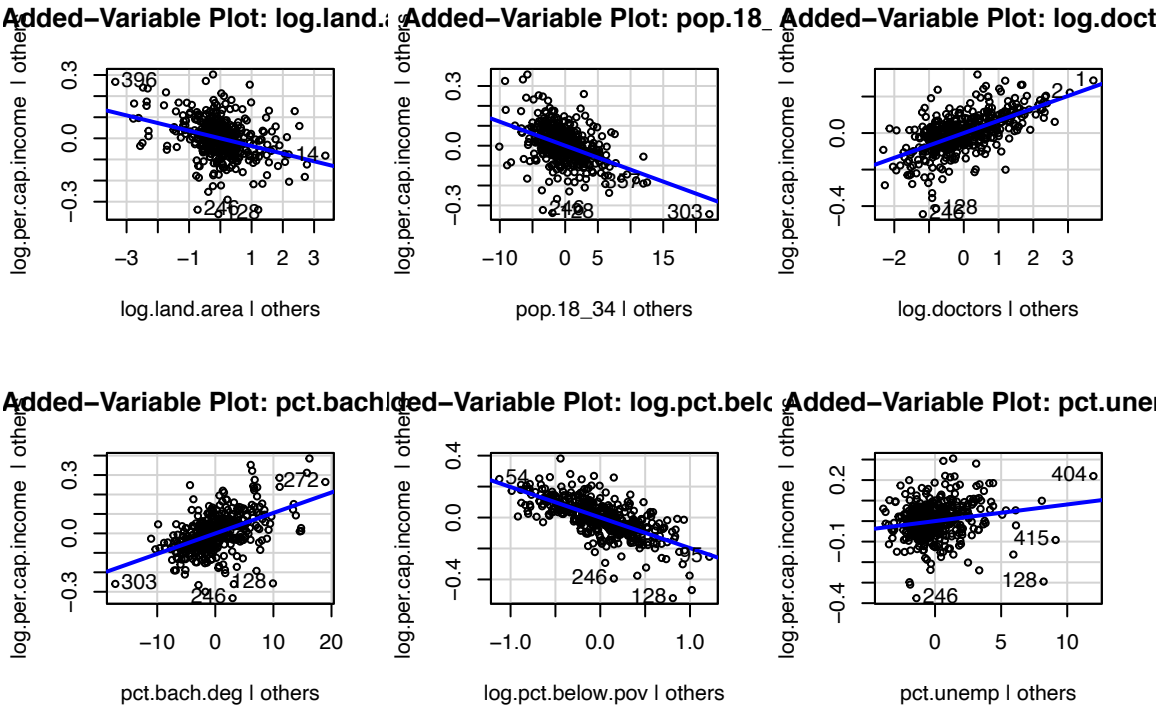
```
## The following objects are masked from cdi_transformed (pos = 3):
##
##     log.crimes, log.doctors, log.hosp.beds, log.land.area,
##     log.pct.below.pov, log.per.cap.income, pct.bach.deg, pct.hs.grad,
##     pct.unemp, pop.18_34, pop.65_plus
```



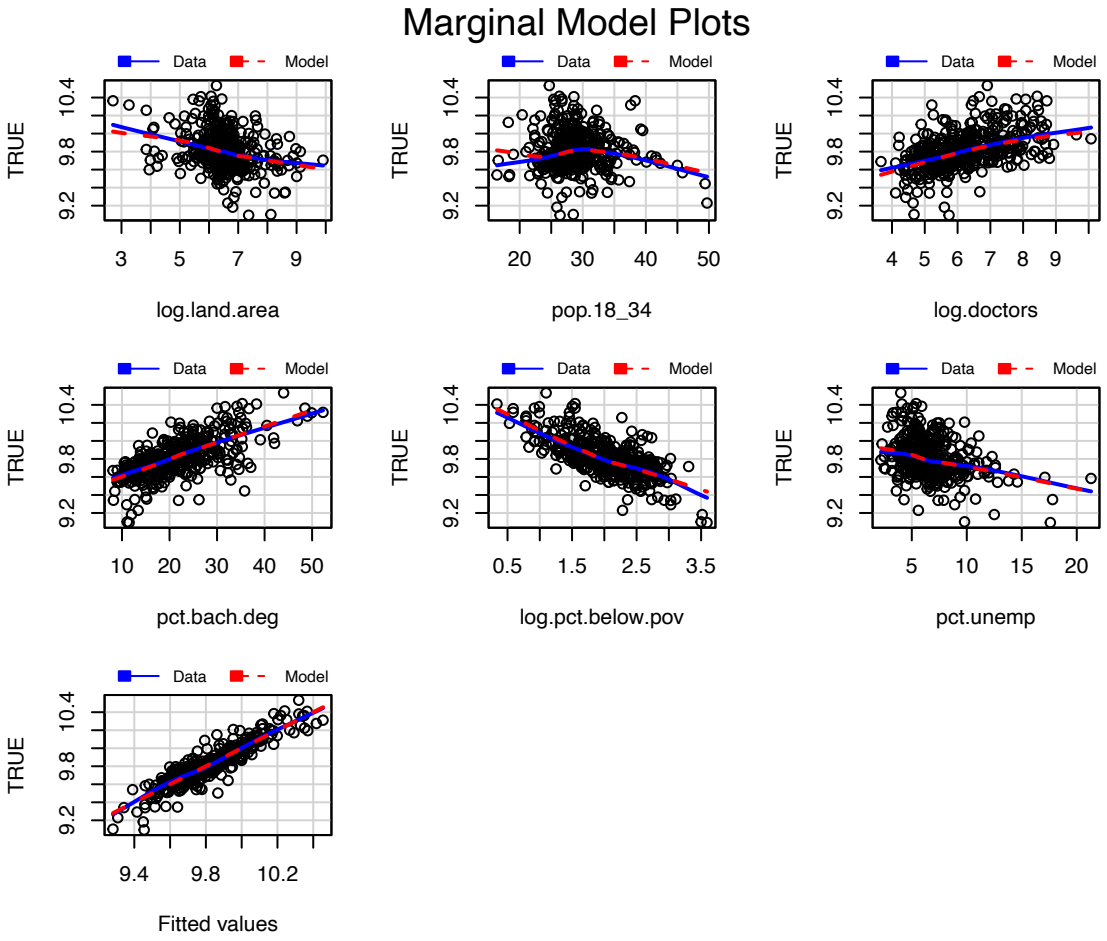
Next, let us look at the added variable plot below. The added variable plots indicate that linear models would be appropriate and all the predictors are important, when adjusted for the effects of the other predictors. This further indicates that the model is valid.

```
## 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
```



We can also take a look at the marginal plots. The plots all look good, as we can see the blue curved lines tend to line up well with the red dashed model-based curves.



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. If none is significant, I will be dropping the whole variable. As a result:
* 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 original model with 6 variables generated from all subsets, and our model with region interaction terms involved. We will use ANOVA F test, AIC and BIC values. The F test suggests that the model with some region interaction terms is statistically significant, and 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
```

We can see that ANOVA (F test) and AIC favor the model with the additive region and some interactions included. But BIC favors the first model without region, because BIC tends to favor simpler models (it goes for parsimonious explanatory model rather than predictive model). I would say that *all_subsets_model_with_some_region* is the optimal choice here.

Variable Selection - Stepwise Regression

Now, I will attempt to carry out **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. All predictors are statistcally significant with low p values.

```
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 ***
## 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, and compare the results with the all subsets and previous stepwise BIC we explored. It seems like the AIC and BIC for the stepwise BIC with interaction terms is the lowest.

```
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()
```

	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

Lets now look at the stepwise BIC model with two way interaction to see if it is actually worth to include two way interaction terms.

```
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 woul 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.

Since the stepwise BIC regression came up with the same predictors as all subsets did, we will get the same results when considering the *region* categorical variable. We would end up with the all subsets model with some significant *region* interaction terms involved (*all_subsets_model_with_some_region*).

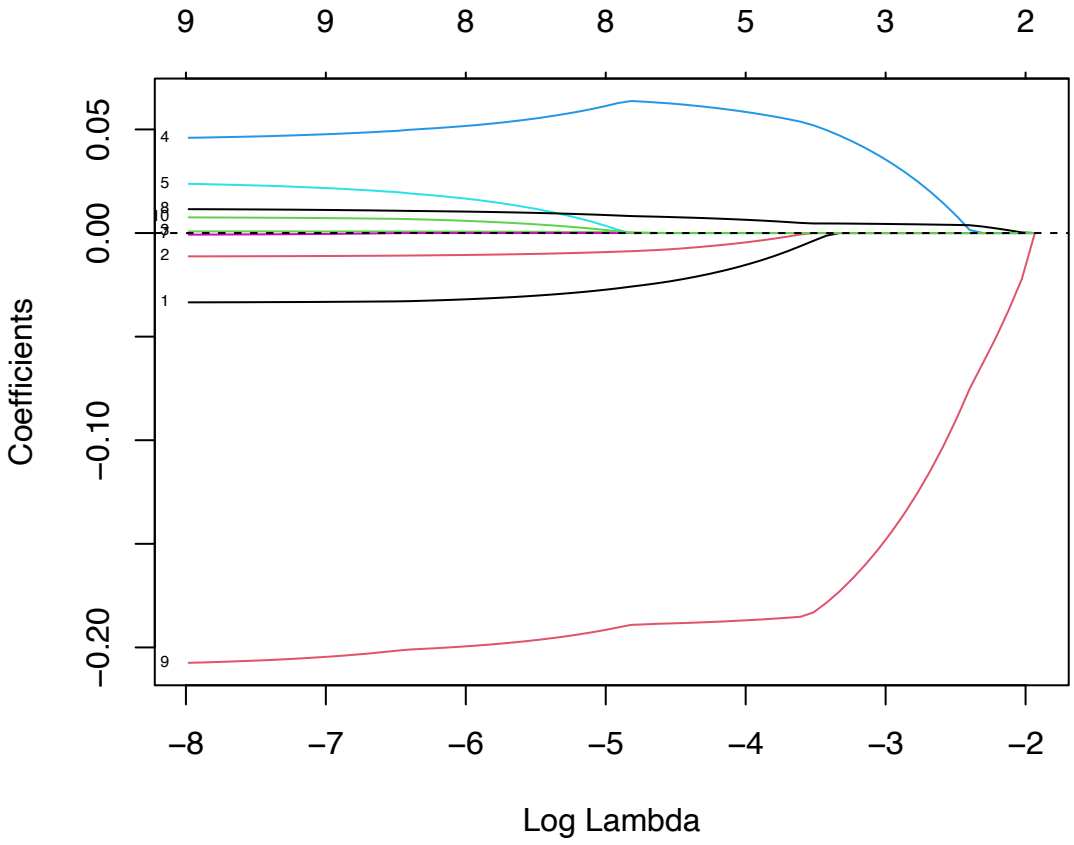
Variable Selection - LASSO regression

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)
```

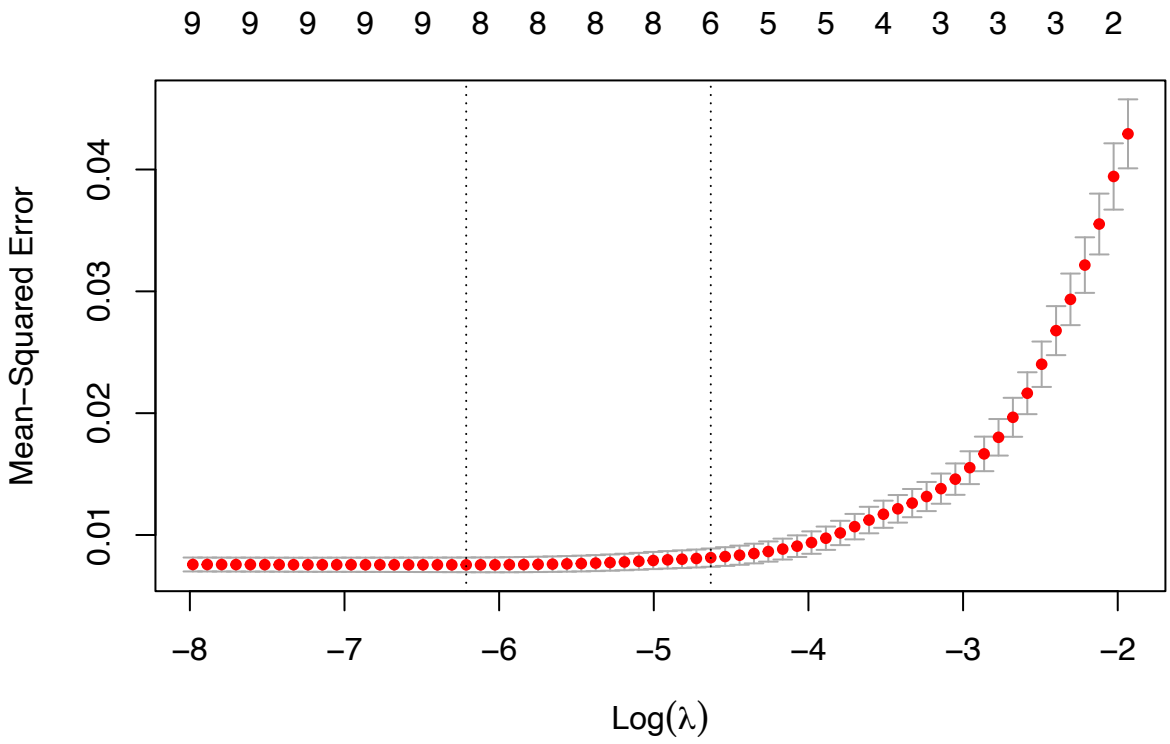
The plot shows how many non-zero variables are in the model at the top. So at a log Lambda of -4, the model has 5 variables.

```
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)
```



lambda.min, the best value found by cross-validation and **lambda.1se**, which is the value of lambda that is one SE larger than lambda.min, is seen below. I will be using **lambda.1se** of 0.0097 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. It looks like the R squared value is still high with a value of 0.833, and the residual standard error is relatively small. All predictor coefficients except for *pop.65_plus* is statistically significant with low p values.

```
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
```

We can immediately see that our all subsets model with *pct.unemp* turned out to be more significant. The predictor *pop.65_plus* does not add much to the model, so we will stick with our final all subsets model we have identified before, with the *regions* categorical variable plus some interaction terms added.

It makes sense that *hosp.beds* was removed, since it would be highly correlated to the *doctors* variable. Interestingly, the variable *crimes* seemed to be insignificant when predicting the Per Capita Income of a County.

Final model with region variable and assessing validity of model

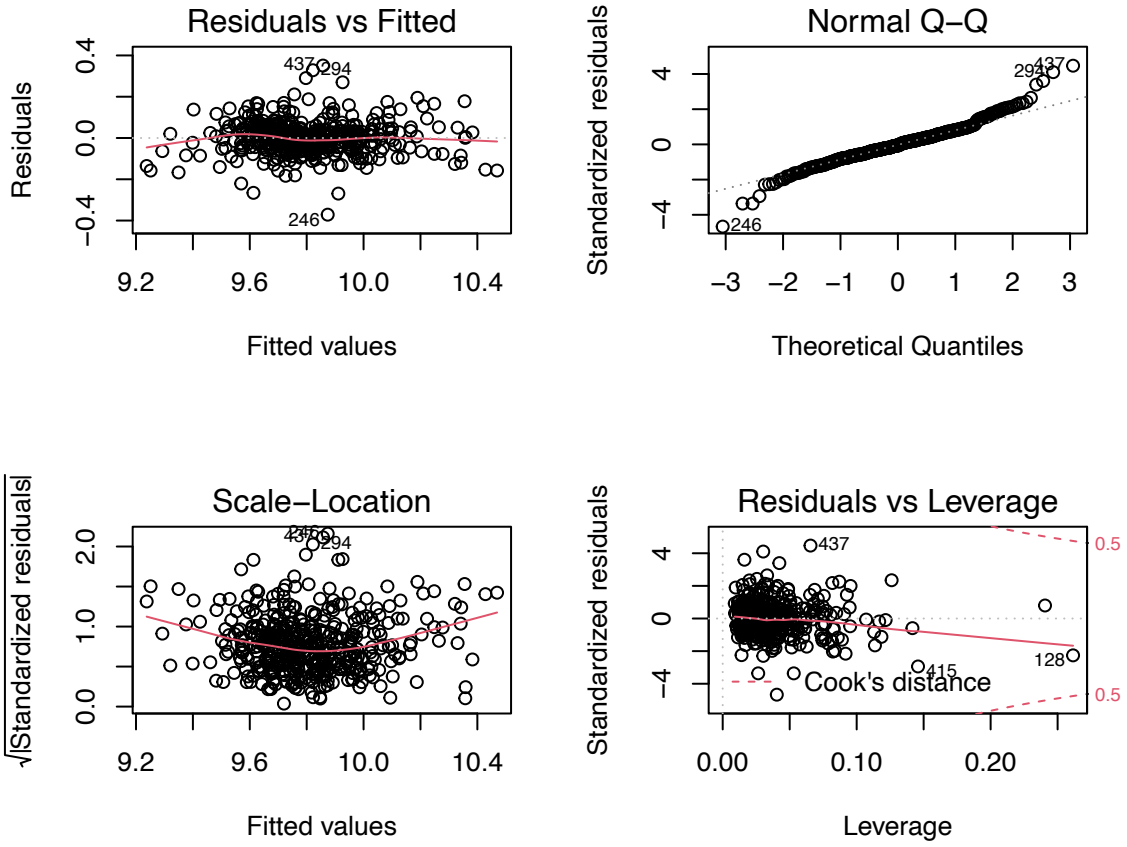
Looking at the boxplot of *Per Capita Income* and *Region* that I produced in **part (a)**, it can easily be seen that there is some variability and differences in per capita income between the difference regions. Due to this I thought that the *region* variable could add something meaningful to the model and not including it could make the model suffer from ommitted variable bias. Thus, I chose to include the *region* categorical variable in my final model.

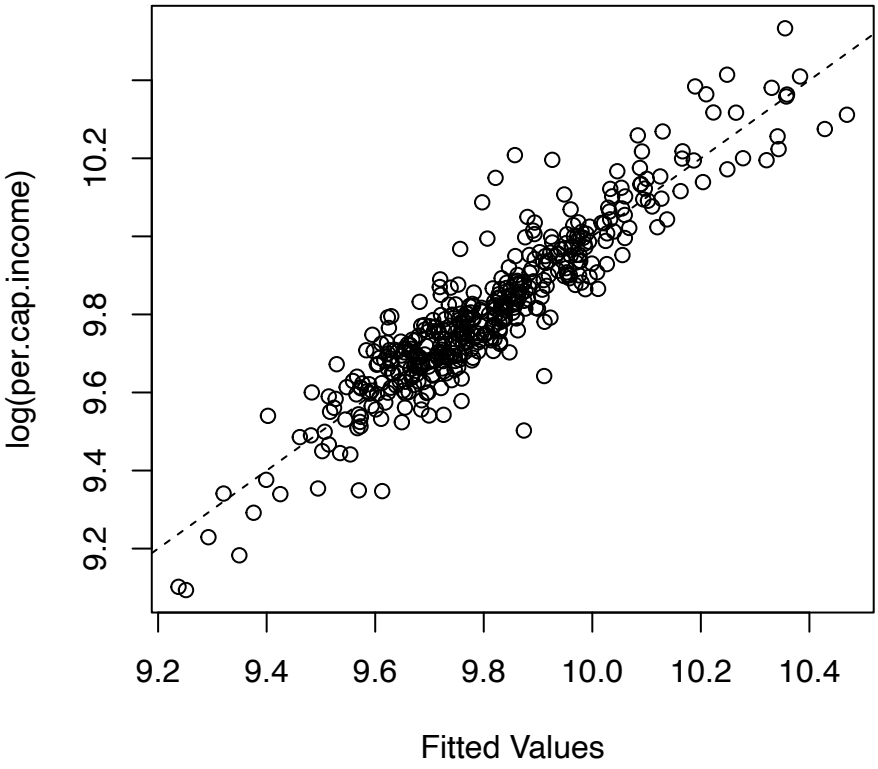
The summary of my final model can be seen below. It looks like the R squared value is around 0.85 and didn't change much. Interpretation on the coefficients will be done in the IMRAD report.

```
##                                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
```




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