Regression Analysis to Predict Income Using Socio-economic Factors at Regional Level

Ziyan Xia
Department of Statistics and Data Science, Carnegie Mellon University
zxia2@andrew.cmu.edu

Abstract:

It's always an interesting topic for social scientists to learn how average income was related to other variables associated with the county's economic, health and social well-being. To address this question, we used a county demographic information (CDI) for 440 of the most populous counties in the United State as our data. Methods including exploratory data analysis (EDA), All Subset and Stepwise Regression as well as ANOVA were applied to the CDI data to find the relationship between these socioeconomic variables and select the best model we need. Using these methods, we found that per capita income is best predicted from logged land area, Percent of population aged 18–34, logged number of active physicians, Percent high school graduates, Percent bachelor's degrees, Percent below poverty level, Percent unemployment, region and the interactions between region and Percent high school graduates, Percent bachelor's degrees, Percent below poverty level and Percent unemployment. Further researches about detecting high-order interactions and whether to include more states and county data should be done.

1. Introduction

Knowing what effects average income per person will be helpful for policy making and social science studies. We all know some variables such as education levels, unemployment rate are associated with average income per person of the population at an intuitive level. However, there are more variables we can use and themselves might be associates with each other. Therefore, it's always an interesting topic for social scientists to learn how average income was related to other variables associated with the county's economic, health and social well-beings. To answer the question, we break it to four small questions and they are as follows.

1. **Relationship between variables:**
   Is there any pairwise relationships between those economic, health and social well-beings associated variables that is quite surprising? If there is, can we explain in terms of their meanings?

2. **Crime rate and Region:**

Is it true that controlling all the other variables, average income per person should be related crime rate and this relationship vary by different regions like Northeast and South? Here should we use number of crimes or (number of crimes)/(population)?

3. **Predict Income:**
   Can we find the best model to predict average income per person using selected variables?

4. **Whether missingness matters:**
   Should we be worried about either the missing states or the missing counties? Why or why not?

2. 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.

We started by checking the definition of CDI data we will be using. Table 1 is the definition of variables. From Table 1, It is quite obvious that population and total person income are both directly related to average income per person and total population is likely to be related with variables like total population with age between 18 and 34.

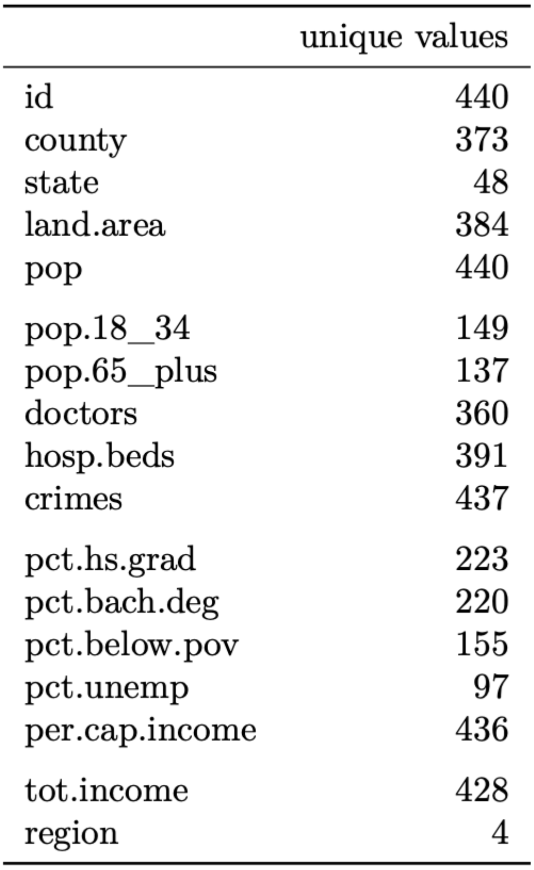
Table 1: Variable definitions for CDI data from Kutner et al. (2005). Original source: Geospatial and Statistical Data Center, University of Virginia.

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)

Then we would like to check the unique values of each variable. Table 2 is the unique values of each variable. From Table 2, It looks like ID is just the same as the row number for each row of the CDI data and therefore not useful for data analysis. Variable State has 48 values, which is also a lot and County is a categorical variable with nearly as many unique values (373) as rows in the CDI data frame (440). As mentioned before, each row represents single county, we will explore more on this afterwards. There is no missing data in CDI data. [See Technical Appendix, Page 4]

Table 2: Unique values for each variable

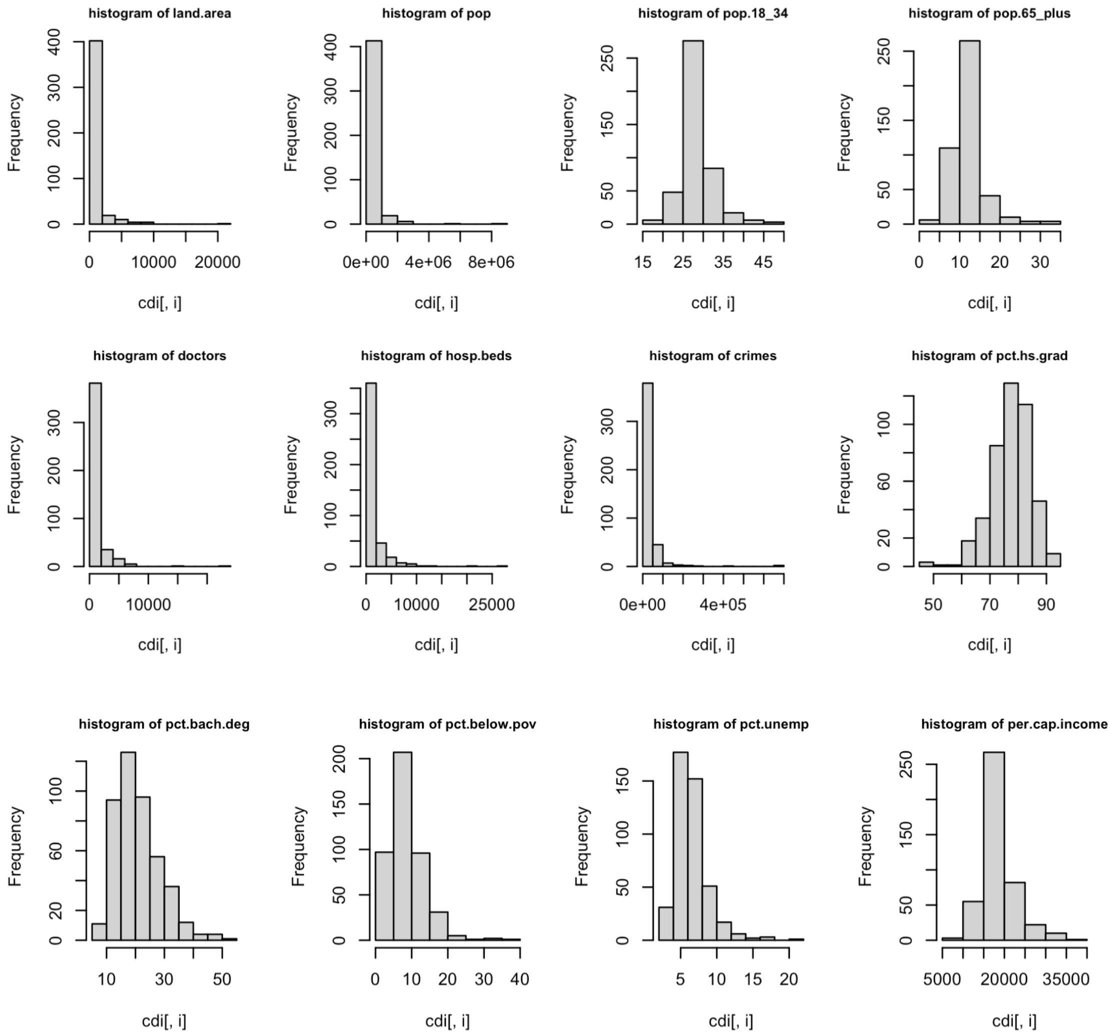


	unique values
id	440
county	373
state	48
land.area	384
pop	440
pop.18_34	149
pop.65_plus	137
doctors	360
hosp.beds	391
crimes	437
pct.hs.grad	223
pct.bach.deg	220
pct.below.pov	155
pct.unemp	97
per.cap.income	436
tot.income	428
region	4

To explore the distribution of numeric variables, we made a summary table for all numeric variables and histograms for all numeric variables. Table 3 is the summary statistics of all numeric variables. From Table 3, there are several variables with Mean substantially larger than Median, indicating possible right skewed. Figure 1 is the histograms for all numeric variables and it further confirmed that there are several right skewed variables, which are land area, population, number of active physicians, number of hospital beds, total serious crimes, and total personal income, and maybe average income per person. These variables may need some log transformation before modelling. After applying log transformation to some skewed variables, the distribution of these variables looks more normal [See Technical Appendix, Page 3-4].

Table 3: Summary Statistics for numeric variables

	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



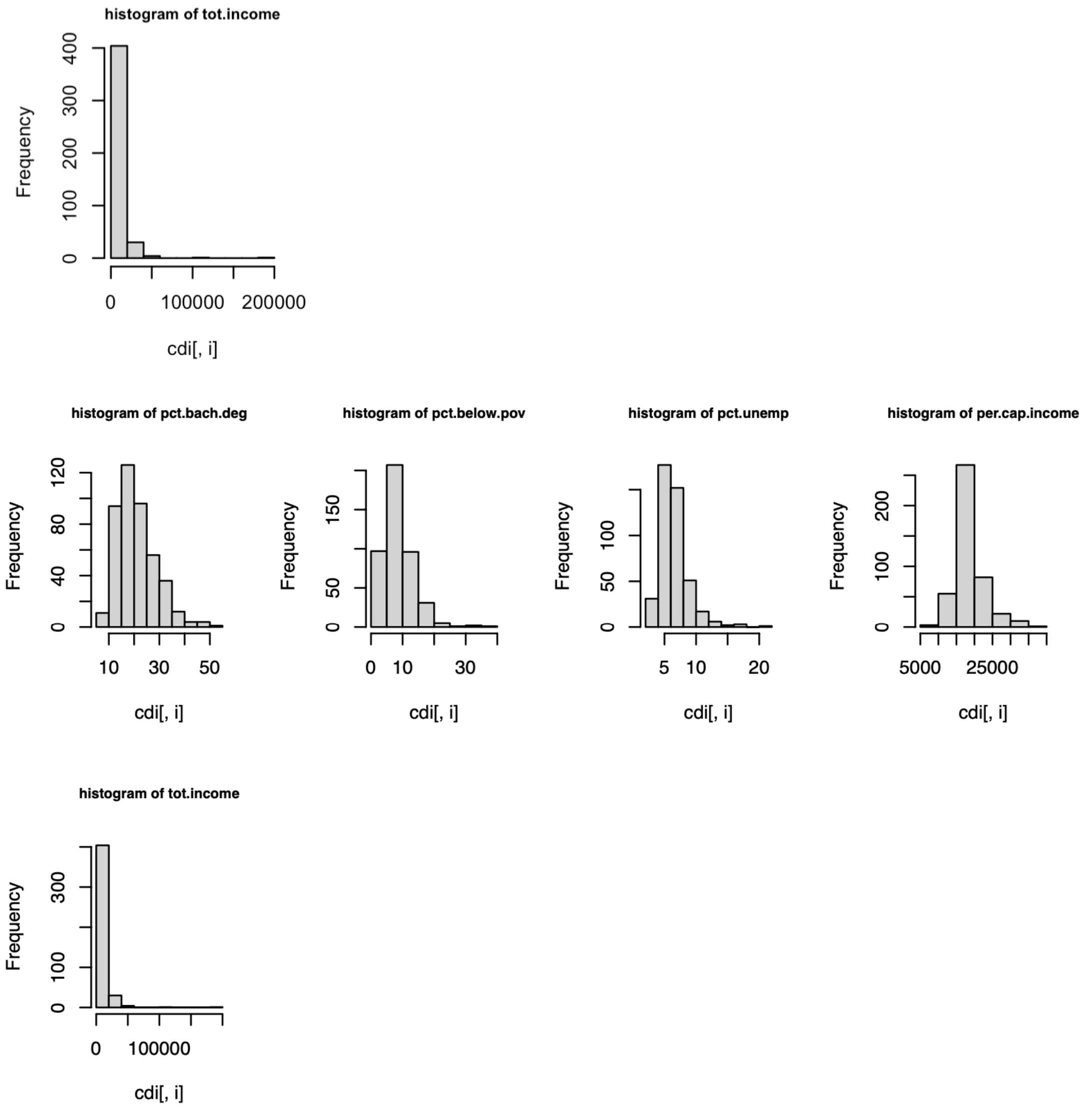


Figure 1: Histograms of all numeric variables

As Table 2 indicates that maybe multiple observations per county are really single observations from counties with the same name in different states, to further investigate this, we made a table that combines county with state. Table 4 is a table that combines county with state. Table 2 and Table 4 show that if we combine county with state, we get 440 unique values, which means some counties in different states have the same name. We only have one observation per unique county and there aren't multiple observations per county are really single observations from counties with the same name in different states. Therefore, county is not a useful 2 variable to include in models.

Table 4: Combine County with State

Counties 1-110	Counties 111-220	Counties 221-330	Counties 331-440
Ada ID	Ector TX	Lycoming PA	Rockingham NH
Adams CO	El_Dorado CA	Macomb MI	Rockland NY
Aiken SC	El_Paso CO	Macon IL	Rowan NC
Alachua FL	El_Paso TX	Madison AL	Rutherford TN
Alamance NC	Elkhart IN	Madison IL	Sacramento CA
Alameda CA	Erie NY	Madison IN	Saginaw MI
Albany NY	Erie PA	Mahoning OH	Salt_Lake UT
Alexandria_City VA	Escambia FL	Manatee FL	San_Bernardino CA
Allegheny PA	Essex MA	Marathon WI	San_Diego CA
Allen IN	Essex NJ	Maricopa AZ	San_Francisco CA
Allen OH	Fairfax_County VA	Marin CA	San_Joaquin CA
Anderson SC	Fairfield CT	Marion FL	San_Luis_Obispo CA
Androscoggin ME	Fairfield OH	Marion IN	San_Mateo CA
Anne_Arundel MD	Fayette KY	Marion OR	Sangamon IL
Arapahoe CO	Fayette PA	Martin FL	Santa_Barbara CA
Arlington_County VA	Florence SC	Maui HI	Santa_Clara CA
Atlantic NJ	Forsyth NC	McHenry IL	Santa_Cruz CA
Baltimore MD	Fort_Bend TX	McLean IL	Sarasota FL
Baltimore_City MD	Franklin OH	McLennan TX	Saratoga NY
Barnstable MA	Franklin PA	Mecklenburg NC	Sarpy NE
Bay FL	Frederick MD	Medina OH	Schenectady NY
Bay MI	Fresno CA	Merced CA	Schuylkill PA
Beaver PA	Fulton GA	Mercer NJ	Sedgwick KS
Bell TX	Galveston TX	Mercer PA	Seminole FL
Benton WA	Gaston NC	Merrimack NH	Shasta CA
Bergen NJ	Genesee MI	Middlesex CT	Shawnee KS
Berks PA	Gloucester NJ	Middlesex MA	Sheboygan WI
Berkshire MA	Greene MO	Middlesex NJ	Shelby TN
Bernalillo NM	Greene OH	Midland TX	Smith TX
Berrien MI	Greenville SC	Milwaukee WI	Snohomish WA

Table 5: Frequency table of region

	NC	NE	S	W
Freq	108	103	152	77

Besides from making a histogram of region, we also made a frequency table of region, which is table 5. We found that the majority of observations are in the Southern region, followed by the North-central region and northeast regions and finally the western region.

4. Methods

To learn how average income was related to other variables associated with the county's economic, health and social well-being, there are four questions to answer. Our methods to answer these questions are as follows:

1. **Relationship between variables:**
To answer this question, we draw a correlation plot with different color indicating whether is positive or negative correlation and with different shades of a specific color indicating different values of correlation. The categorical variables were excluded here because their categories are too many and therefore hard to interpret.

2. **Crime rate and Region:**
For this question, after applying log transformation to per capita income and total serious crimes, we first created a model which regresses response variable per capita income solely on total serious crimes as our baseline model. Then we redid the regression process adding categorical variable region and then the interaction term between total serious crimes. We used AVONA and partial F test to test the whether the region and the interaction between total serious crimes we added is needed in the model.
Then we replaced the logged total serious crimes with logged crime rate, which is calculated by applying log transformation to (total serious crime / total population). After that, we redid the previous process about testing whether to add region and the interaction between region and crimes rate in the model. After these two processes, we will have two models to compare and they are both selected by AVOVA and Partial F test. At this time, we selected the best model by comparing AIC and BIC of these two models because they are not nested models so ANOVA doesn't work for this situation. For the final step, we used diagnostics plots of this best model to decide whether this is a good fit.

3. **Predict Income:**
As skewness of some variables is discovered as a potential problem in the histograms we draw before, we first log transformed number of active physicians, land area, number of hospital beds, and total serious crimes because they were all right skewed. As we mentioned some collinearity between variables and some meaningless variables as well as some categorical variables with too many unique values, we dropped total population, ID, total personal income, state and county in our modelling here.
For modelling process, we first fitted a linear regression model that regressed per capita income on all the other variables left. Then we conducted Stepwise (using BIC), All Subsets Regression to select the best model without considering the interaction between numeric predictor variables and region. After selecting the best model without interactions, added-variables plots and marginal model plots were used to evaluate the performance of the model. Then we used ANOVA methods to decide which interaction terms between numeric variables to keep based on their significance. Finally, we evaluated the performance of the final best model by diagnostics plots.

4. **Whether missingness is important:**

Whether missingness of the states or counties matter will be analyzed by evaluating the missing states in the states that appear in the CDI data and by Table 4 which combine county with state. We would also consider the results of previous model fitting.

5. Results

1. Relationship between variables:

To explore the relationship between these numeric variables, we made a correlation plot of them. Figure 2 is the correlation plot of numeric variables in CDI data. From this correlation plot, there are many variables that are highly correlated with others. Table 6 is a table that selects some variables which have obvious correlation with others.

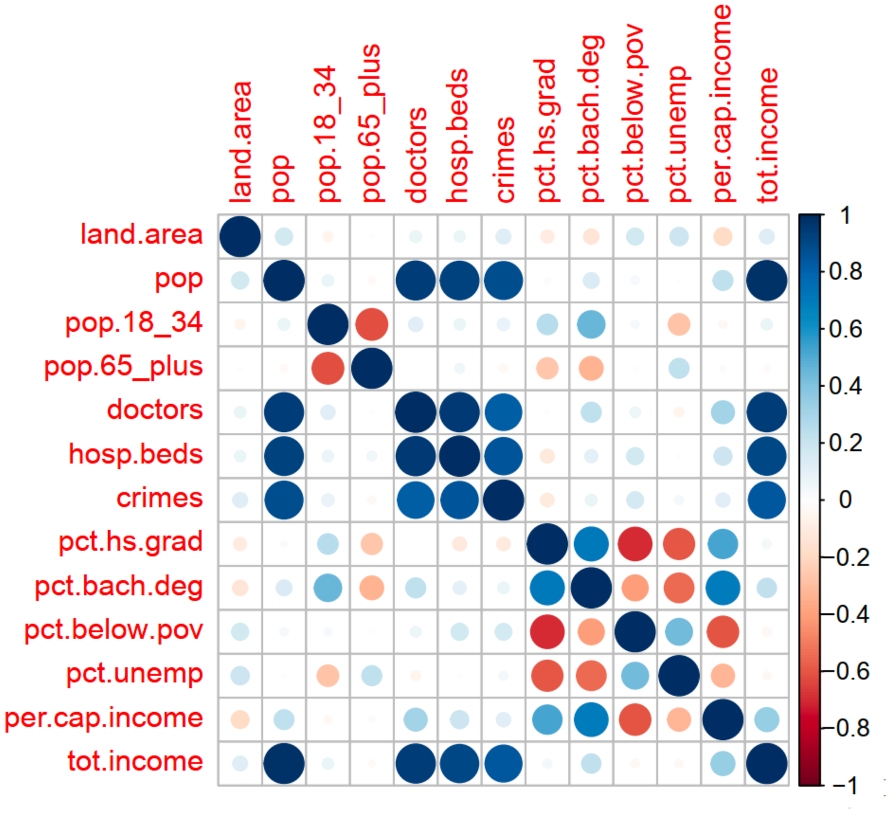


Figure 2: The Correlation Plot of all numeric variables

Table 6: Correlation of the variables

Variables	Obvious Correlation	Correlation
Total population	Number of active physicians Number of hospital beds Total serious crimes Total personal income	0.94 0.92 0.89 0.99
Percent of population aged 18–36	Percent of population 65 or older	-0.62

Number of hospital beds	Total population	0.92
	Number of hospital beds	0.95
	Total serious crimes	0.86
	Total personal income	0.90
Number of active physicians	Total population	0.94
	Number of hospital beds	0.95
	Total serious crimes	0.82
	Total personal income	0.95
Total serious crimes	Number of active physicians	0.82
	Number of hospital beds	0.86
	Total population	0.89
	Total personal income	0.84
Percent high school graduates	Percent bachelor’s degrees	0.70
	Percent below poverty level	-0.69
	Percent unemployment	-0.59
Percent below poverty level	Percent high school graduates	-0.69
	Per capita income	-0.60
Per Capita Income	Percent bachelor’s degrees	0.70
	Percent below poverty level	-0.60
	Percent high school graduates	0.52

From Table 6 and Figure 2, we found that:

- Total personal income and Total population are highly correlated, which is not surprising.
- Both of Total personal income and Total population are reasonably highly correlated with Total serious crimes, Number of hospital beds and Number of active physicians.
- The three variables Total serious crimes, Number of hospital beds and Number of active physicians seem strongly correlated with one another.
- Per Capita Income isn’t really highly correlated with anything, but the best possibilities seem to be Percent high school graduates, Percent bachelor’s degrees (positively correlated with Per Capita Income) and Percent below poverty level, Percent unemployment (negatively correlated with Per Capita Income); all four of these variables are moderately highly correlated with one another

2. **Crime rate and Region**:

To answer question 2, we created 6 models: regress per capita income on total serious crimes (Model 1, baseline model), regress per capita income on total crimes and region (Model 2), regress per capita income on total crimes and region with interaction between total serious crimes and region (Model 3), regress per capita income on crime rate (Model 4, baseline model), regress per capita income on crime rate and region (Model 5), regress per capita income on crime rate and region with interaction between crime rate and region (Model 6).

Table 8 shows the AIC and BIC for these 6 models and Table 7 shows the p values for added terms of multiple ANOVA tests. AIC and BIC are model comparison criterion and the less the AIC and BIC are, the better the model. For ANOVA test, we use p value of partial F test to decide whether to add terms in the model, if the p value is very small, that means the term is statistically significant and should be added in the model. Here, "*" means interaction between the first variable and the second variables and "~" here means regressing the left variable on the right variables. [See Technical Appendix, Page 8].

Table 7: F Statistics for ANOVA test

ANOVA baseline model	Added term	P Value
per capita income ~ crimes	region	1.523e-13
per capita income ~ crimes + region	region*crimes	0.5266
per capita income ~ crime rate	region	3.263e-09
per capita income ~ crime rate + region	crime rate*region	0.893

Table 7 indicates that either the interaction between total serious crimes and region or the interaction between crime rate and region are not necessary in the model as their coefficients are not statistically significant. However, region should be added in the model as a predictor variable.

Table 8 shows the AIC and BIC of these models, as our final model should be chosen from Model 2 and Model 5, by comparing the AIC and BIC of these two models, our final best model to predict per capita income from crimes is the Model 2 as both AIC and BIC of Model 2 are smaller.

Table 8: AIC and BIC for 6 models

Model	AIC	BIC
per capita income ~ crimes	-169.9	-157.7
per capita income ~ crimes + region	-227.5	-203.0
per capita income ~ crimes + region + crimes*region	-223.8	-187.0
per capita income ~ crime rate	-122.8	-122.8
per capita income ~ crime rate + region	-172.1	-147.6
per capita income ~ crime rate + region + crime rate*region	-166.8	-130.0

The interpretations of the final best model are, as follows: [See Technical Appendix, Page 9].

- All across the US, for every 1% increase in per-capita crime, there is an associated 0.04% increase in per-capita income
- The regional baseline salaries are: NC: $20,743.74, NE: $23,155.79, S: $19,341.34, and W: $20,332.99. All but the W region have baselines that are, according to the model, significantly different from the NC baseline.
- Again, the level of salary varies with region, but not the way it varies with crime, according to the model.

3. Predict Income:

From Table 1, as per capita income is actually just total personal income/population, so we cannot use total personal income in our regression analysis, also for this reason, we should exclude total population in the fitted model. Besides, the ID, Country, State will be no helpful for prediction but only added complexity to the model so it will be better to exclude them in the regression analysis.

For this question, we decided to first conduct variables selection without categorical variable. Testing which interaction terms to include in the model will be the final step. We first regressed logged per capita income on all the other variables except region (logged land area, logged number of physicians, logged number of hospital bed, logged total serious crimes, Percent of population aged 18–36, Percent of population 65 or older, Percent bachelor's degrees, Percent below poverty level, Percent unemployment, Percent high school graduates) as our baseline model. Then we used All Subsets method to compare the BIC of each subset and selected the model that has the smallest BIC as our best model [See Technical Appendix, Page 10].

To test whether the variable transformation and selection makes sense as well as whether model is a good fit, diagnostics plots, added-variable plots and marginal model plots were made to evaluate the model performance.

For the diagnostics plots of our final model, the Residuals vs Fitted plot shows that there is not nonlinear pattern in the model but there are some outliers detected; the normal Q-Q plot shows that the residuals are normally distributed and the assumption that errors are normally distributed holds, but there are some points that deviated from the diagonal line a lot; the scale-location plot shows that the residuals are spread equally along the ranges of predictors and the residuals have constant variance; the residuals vs leverage plot shows that there are no influential points in the model [See Technical Appendix, Page 12].

The Added-Variable plots show that the model won't need any transformation to the variables included in the model. The marginal model plots look very good – the blue data-based curves line up well with the red model-based curves. We don't seem to be missing any important transformations, interactions, etc [See Technical Appendix, Page 13-14].

Therefore, for this All Subsets method, the best one is the model that regresses logged per capita income on logged land area, logged number of physicians, Percent of population aged 18–36, Percent bachelor's degrees, Percent below poverty level, Percent unemployment, Percent high school graduates.

Except for All Subsets, we also used Stepwise Regression to do the variables regression and this time we also excluded categorical variable region. The best model selected by Stepwise Regression using BIC is the same as the best one selected by All Subsets. The best model selected by Stepwise Regression using AIC is slightly different. Compared to the previous best model, it has one more variable, which is Percent of population 65 or older. To decide whether to include this variable in the model, an ANOVA test was conducted to address this problem. [See Technical Appendix, Page 14]. The result rejects the null hypothesis that Percent of population 65 or older is not needed in the model. We also checked the summary of the model [See Technical Appendix, Page 15].

Although the coefficient on Percent of population 65 or older is significantly different from zero, its effect on expected per-capita income appears to be quite small. Considering that our model should best reflects the social science and the meaning of the variables, we decided to exclude it from the model.

The final step is to decide whether to include any interaction between the remaining numeric variables and the categorical variables region. We did same thing as we did for addressing question 2—we conducted multiple AVOVA tests to select interaction terms. Table 9 is the result of the ANOVA tests. The baseline model is the best model selected from All Subsets and Stepwise Regression. The column [Interaction Term] contains each interaction term we tested and its corresponding p value in column [P Value]. "*" here means interaction between the first variable and the second variables. The interactions here are all pairwise.

Table 9: Multiple ANOVA Test Result

Interaction Term	P Value
logged land area*region	0.09528
Percent of population aged 18–34*region	0.01744
logged number of active physicians*region	0.02433
Percent high school graduates *region	0.0002411
Percent bachelor's degrees *region	0.0001938
Percent below poverty level *region	0.006056
Percent unemployment *region	0.0002352

Based on the Table 9, 6 out of 7 interaction terms have their p values less than 0.05. If the p value is very small, here we mean less than 0.05, it means this interaction is significantly different from zero. However, adding so many significant interaction terms will make our model hard to interpret. Therefore, we changed the significance level from 0.05 to 0.01 to do the interaction terms selection, by this way we are able to reduce the significant interaction terms to just 3, which are interaction between region and Percent high school graduates, Percent bachelor's degrees, Percent below poverty level and Percent unemployment.

Now the final model is the model as follows:
Y-Response Variable:
logged per capita income
X-Predictor Variables:
logged land area
Percent of population aged 18–34
logged number of active physicians
Percent high school graduates
Percent bachelor's degrees
Percent below poverty level
Percent unemployment
region

Percent high school graduates*region (interaction)
Percent bachelor's degrees*region (interaction)
Percent below poverty level*region (interaction)
Percent unemployment*region (interaction)

The interpretation of the coefficients of each predictor variables are as follows [See Technical Appendix, Page 18-19]:

- For every 1% increase in a county's land area, there is a 0.03% decrease in expected per-capita income.
- For every 1% increase in the number of doctors in a county, the expected per-capita income increases by about 0.06%.
- For every 1% increase in the percent of the population aged 18–34, there is an expected 2% drop in per-capita income.
- Percent high school graduates doesn't have much effect, except in the South, where a one percentage point increase in has graduates induces an expected 2% decrease in per-capita income.
- 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.

We also checked whether this final model is a good fit using diagnostics plots [See Technical Appendix, Page 19]. For the diagnostics plot of our final model, the Residuals vs Fitted plot shows that there is not nonlinear pattern in the model but there are some outliers detected; the normal Q-Q plot shows that the residuals are normally distributed and the assumption that errors are normally distributed holds, but there are some points that deviated from the diagonal line a lot; the scale-location plot shows that the residuals are spread equally along the ranges of predictors and the residuals have constant variance; the residuals vs leverage plot shows that there are no influential points in the model. Hence, the final model is a good fit.

4. Whether missingness matters

After extracting the unique values of variable State, we found that state Alaska, Iowa, and Wyoming are missing states in our data. These states are perhaps less populous compared to other states. The modelling process before shows that there are a quite a lot variables associated with region and these interactions are useful when predict per capita income. It seems that the missing data matters because region plays a large part in our prediction and the model might not be applied to these states because of the missing data.

Both "Baltimore MD" and "Baltimore City MD" are listed in Table 4, which makes me wonder these two data points are really independent. Therefore, if adding more county data, they may be highly relevant just like these two [See Technical Appendix, Page 19-20].

6. Discussion

The analysis shows some reasonable relationships between variables. For example, total population is related with number of doctors, total serious crimes and also total incomes. The higher total population is, the higher number of doctors, total crimes and also total incomes. Also, it makes sense that the higher percentage of high school graduates, the lower the percentage of population below poverty level and the percentage of unemployment. However, it is quite surprising that Per Capita Income isn't really highly correlated with anything.

Our analysis found that there is no need to include interactions in the model to predict per capita income use crimes/crime rate and region but it's better to include the categorical variable region in our analysis. The results for question 2 answered the second question that if we ignore all other variables, per-capita income should be related to crime rate, and that this relationship may be different in different regions of the country (Northeast, Northcentral, South, and West).

We tackled the problem about skewness by using log transformation and successfully found a model that is overall a good fit to predict per capita income using the variables in CDI data. For this analysis, we conjecture that the positive significant coefficient of land area in the model could be due to an urban-rural contrast because rural counties tend to be bigger than urban ones. We also think that the surprising negative coefficient of Percent of population aged 18–36 in the model might be due to percent 18–34 years old are not at peak earning capacity yet and so perhaps their lower incomes drag down the per-capita average. The positive coefficient of number of active physicians in the model makes sense as doctors are well-paid and could be big contributors to the per-capita average income. It's quite interesting that after adding interactions, the previous significant variable Percent high school graduates don't have much effects. It might depend on whether college graduates are counted as a subset of has 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. Also, we only tested pairwise interactions but didn't test whether there should be higher order interactions. It seems hard to test higher order interactions just use regression analysis so for further research, more methods may be applied to this dataset to do this.

Throughout the model selection process, some tradeoffs were made when selecting interpretability. We didn't include all the terms that will be considered statistically significant and we only did log transformation to some extremely skewed variables but left the other quite skewed variables untransformed.

Missing states and counties seem to be a problem. Some variables may vary between states and counties, the distribution of these variables may be different in different states and counties. We also found that data points in our data might not be independent. There should be further research after collecting more missing data.

Finally, the CDI data was collected 30 years ago, which means the model we built might not be quite useful nowadays. This is also a limitation.

References

Kutner, M.H., Nachsheim, C.J., Neter, J. & Li, W. (2005) *Applied Linear Statistical Models, Fifth Edition*. NY: McGrawHill/Irwin

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

Technical Appendix

Ziyan Xia

10/9/2021

Part A: Exploratory Data Analysis

```r
cdi<-read.table("/Users/ceciliaxia/Desktop/cdi.dat")
```

Table 2: Unique values for each variable

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

It looks like ID is just the same as the row number for each row of the CDI data and therefore not useful for data analysis. Variable State has 48 values, which is also a lot and County is a categorical variable with nearly as many unique values (373) as rows in the CDI data frame (440). As mentioned before, each row represents single county, we will explore more on this afterwards.

Table 3: Summary Statistics for numeric variables

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

Table 3 is the summary statistics of all numeric variables. From Table 3, there are several variables with Mean substantially larger than Median, indicating possible right skewed.

Figure 1: Histograms of all numeric variables

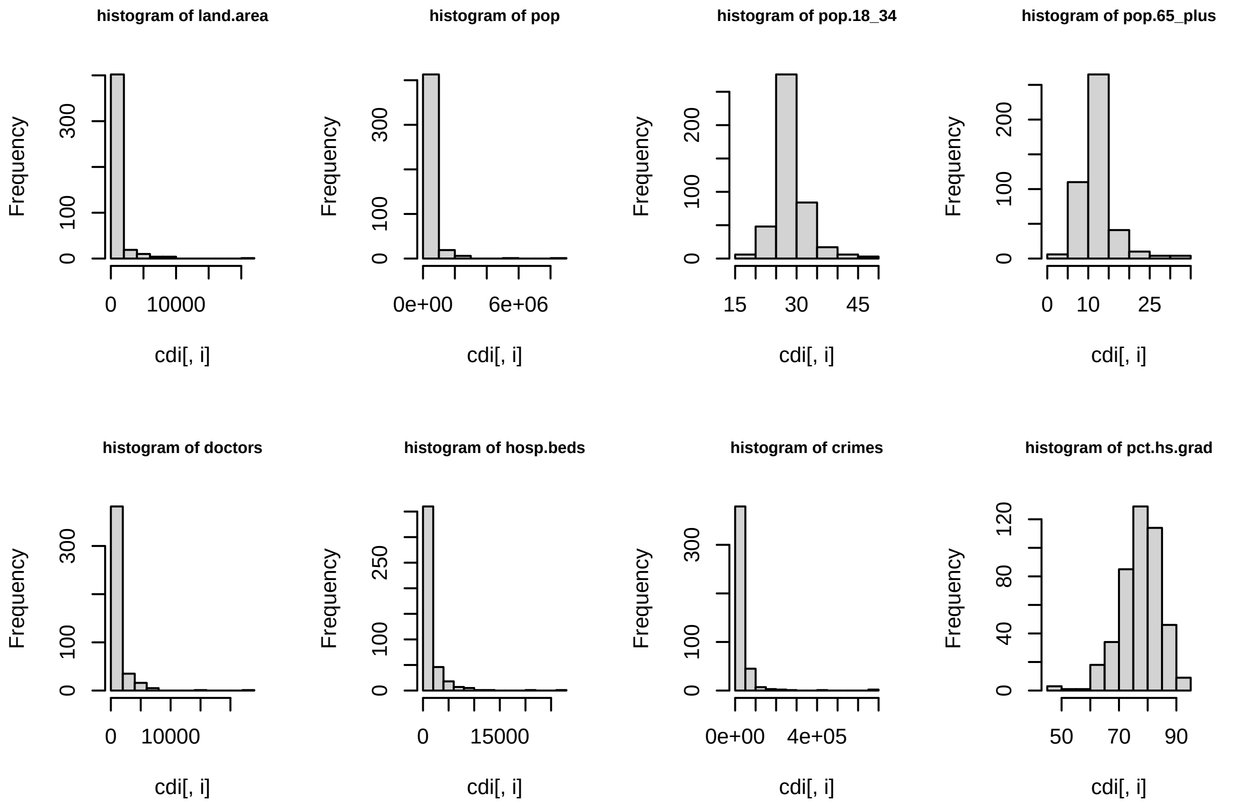
```r
par(mfrow=c(2,4))
par(mfrow=c(2,4))
for (i in c(4:16)){
  hist(cdi[,i],main=paste("histogram of",colnames(cdi)[i]),cex.main=0.8)
}
```

Table 1:

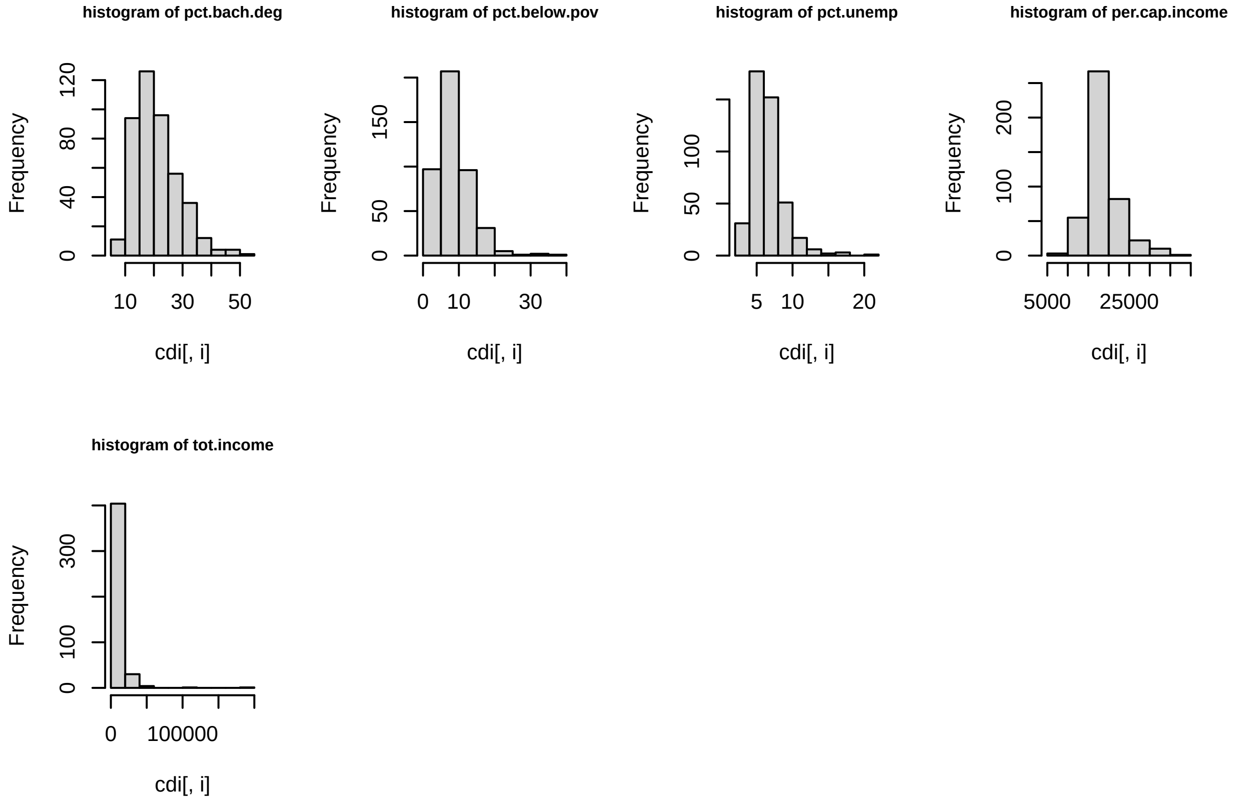
	unique values
id	440
county	373
state	48
land.area	384
pop	440
pop.18_34	149
pop.65_plus	137
doctors	360
hosp.beds	391
crimes	437
pct.hs.grad	223
pct.bach.deg	220
pct.below.pov	155
pct.unemp	97
per.cap.income	436
tot.income	428
region	4

Table 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



```r
par(mfrow=c(2,4))
```



```r
for (i in c(4,8,9,10,15)){
  hist(log(cdi[,i]),main=paste("histogram of logged",colnames(cdi)[i]),cex.main=0.8)
}
```

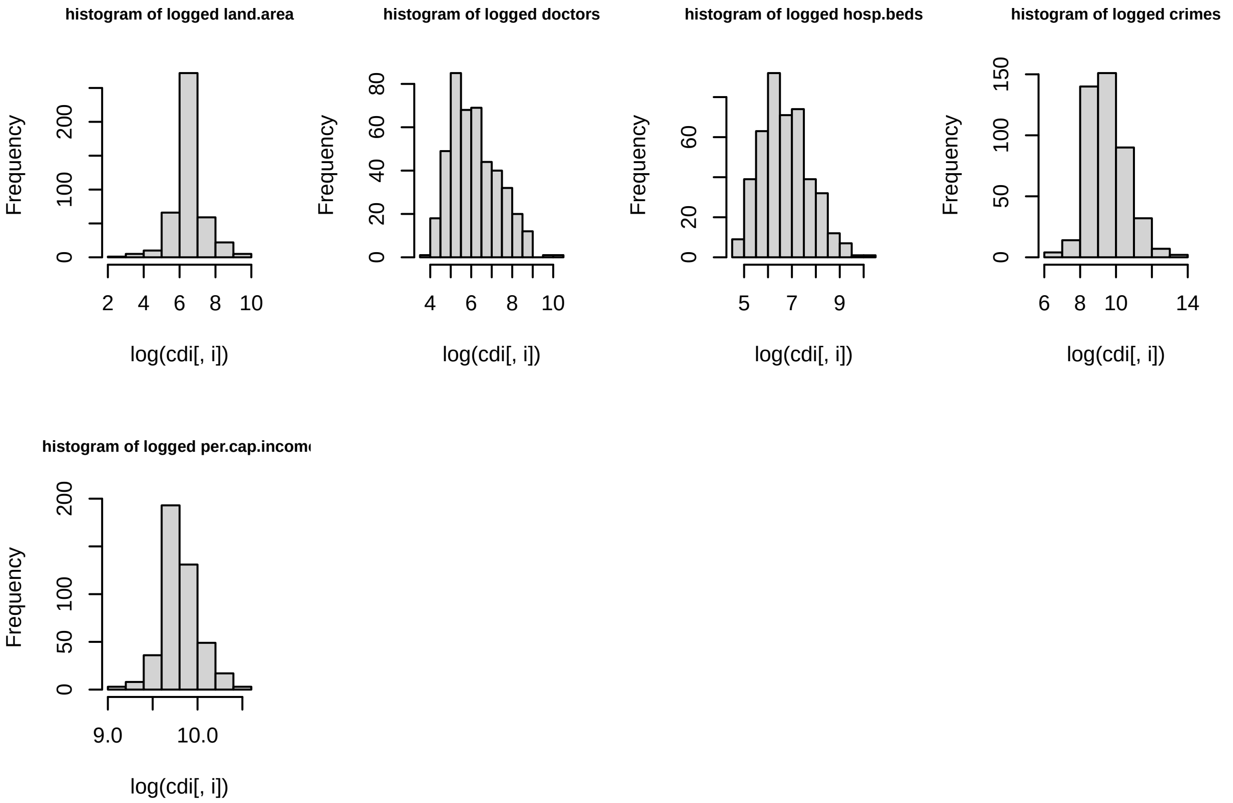


Figure 1 confirmed that there are several right skewed variables, which are land area, population, number of active physicians, number of hospital beds, total serious crimes, and total personal income, and maybe average income per person. These variables may need some log transformation before modeling. After applying log transformation to some skewed variabels, the skeweness was fixed.

Detect NAs

Indicate where (in which variables) there is missing data (NA's)

```r
apply(cdi,2,function(x) any(is.na(x)) )
```

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

There do not appear to be any missing values in the data!

Table 4: Combine County with State

```r
county.state <- with(cdi,paste(county,state))
tmp <- as.data.frame(matrix(sort(county.state),ncol=4))
names(tmp) <- paste("Counties",c("1-110","111-220","221-330","331-440"))
```

```
tmp[1:30,] %>% kbl(booktabs=T,longtable=T,caption=" ") %>% kable_classic(full_width=F)
```

Table 3:

Counties 1-110	Counties 111-220	Counties 221-330	Counties 331-440
Ada ID	Ector TX	Lycoming PA	Rockingham NH
Adams CO	El_Dorado CA	Macomb MI	Rockland NY
Aiken SC	El_Paso CO	Macon IL	Rowan NC
Alachua FL	El_Paso TX	Madison AL	Rutherford TN
Alamance NC	Elkhart IN	Madison IL	Sacramento CA
Alameda CA	Erie NY	Madison IN	Saginaw MI
Albany NY	Erie PA	Mahoning OH	Salt_Lake UT
Alexandria_City VA	Escambia FL	Manatee FL	San_Bernardino CA
Allegheny PA	Essex MA	Marathon WI	San_Diego CA
Allen IN	Essex NJ	Maricopa AZ	San_Francisco CA
Allen OH	Fairfax_County VA	Marin CA	San_Joaquin CA
Anderson SC	Fairfield CT	Marion FL	San_Luis_Obispo CA
Androscoggin ME	Fairfield OH	Marion IN	San_Mateo CA
Anne_Arundel MD	Fayette KY	Marion OR	Sangamon IL
Arapahoe CO	Fayette PA	Martin FL	Santa_Barbara CA
Arlington_County VA	Florence SC	Maui HI	Santa_Clara CA
Atlantic NJ	Forsyth NC	McHenry IL	Santa_Cruz CA
Baltimore MD	Fort_Bend TX	McLean IL	Sarasota FL
Baltimore_City MD	Franklin OH	McLennan TX	Saratoga NY
Barnstable MA	Franklin PA	Mecklenburg NC	Sarpy NE
Bay FL	Frederick MD	Medina OH	Schenectady NY
Bay MI	Fresno CA	Merced CA	Schuylkill PA
Beaver PA	Fulton GA	Mercer NJ	Sedgwick KS
Bell TX	Galveston TX	Mercer PA	Seminole FL
Benton WA	Gaston NC	Merrimack NH	Shasta CA
Bergen NJ	Genesee MI	Middlesex CT	Shawnee KS
Berks PA	Gloucester NJ	Middlesex MA	Sheboygan WI
Berkshire MA	Greene MO	Middlesex NJ	Shelby TN
Bernalillo NM	Greene OH	Midland TX	Smith TX
Berrien MI	Greenville SC	Milwaukee WI	Snohomish WA

Table 4 show that if we combine county with state, we get 440 unique values, which means some counties in different states have the same name. We only have one observation per unique county and there aren't multiple observations per county are really single observations from counties with the same name in different states. Therefore, county is not a useful 2 variable to include in models.

Table 5: Frequency table of region

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

Besides from making a histogram of region, we also made a frequency table of region, which is table 5. We found that the majority of observations are in the Southern region, followed by the North-central region and northeast regions and finally the western region.

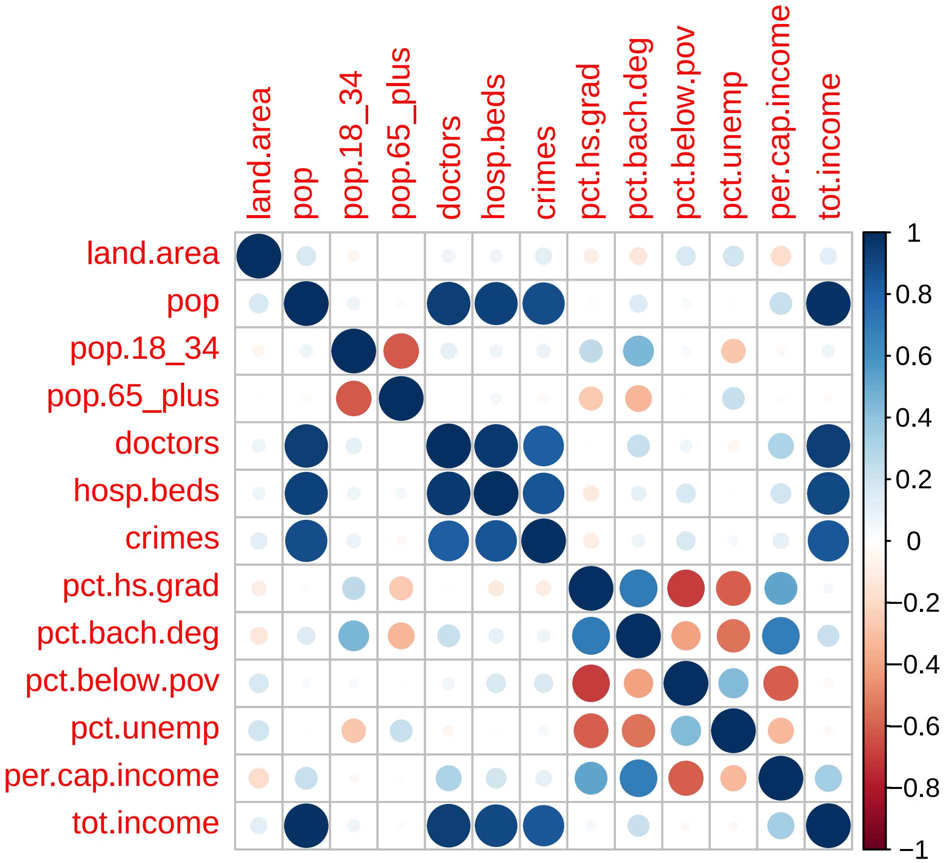
Part B:Relationship between variables

Table 4:

	NC	NE	S	W
Freq	108	103	152	77

Figure 2: The Correlation Plot of all numeric variables

```r
corrplot::corrplot(cor(cdi[,c(4:16)]))
```



To explore the relationship between these numeric variables, we did a correlation plot of them. Figure 2 is the correlation plot of numeric variables in CDI data. From this correlation plot, there are many variables that are highly correlated with others. For example, the correlation between pop and doctors is really high, which is not surprising. However, there is some surprising relationship in this plot and we will elaborate more on this for results part.

Correlation between variables

```r
cor(cdi[,c(4:16)])
```

```
##                   land.area          pop   pop.18_34  pop.65_plus      doctors
## land.area       1.000000000  0.173083353 -0.05487812  0.005770871  0.078074657
## pop             0.173083353  1.000000000  0.07837212 -0.029037393  0.940248591
## pop.18_34      -0.054878125  0.078372117  1.00000000 -0.616309639  0.119699240
## pop.65_plus     0.005770871 -0.029037393 -0.61630964  1.000000000 -0.003128630
## doctors         0.078074657  0.940248591  0.11969924 -0.003128630  1.000000000
## hosp.beds       0.073047270  0.923738360  0.07453191  0.053278417  0.950464395
## crimes          0.129475371  0.886331846  0.08994063 -0.035290324  0.820459477
## pct.hs.grad    -0.098598111 -0.017426900  0.25058429 -0.268251758 -0.004248085
## pct.bach.deg   -0.137237736  0.146813850  0.45609703 -0.339228765  0.236765466
## pct.below.pov   0.171343348  0.038019509  0.03397551  0.006578474  0.064136254
## pct.unemp       0.199209277  0.005351703 -0.27852706  0.236309411 -0.050516116
## per.cap.income -0.187715132  0.235610188 -0.03164843  0.018590706  0.316134625
## tot.income      0.127074261  0.986747626  0.07116151 -0.022733151  0.948110571
##                   hosp.beds      crimes  pct.hs.grad pct.bach.deg pct.below.pov
## land.area       0.073047270  0.12947537 -0.098598111  -0.13723774   0.171343348
## pop             0.923738360  0.88633185 -0.017426900   0.14681385   0.038019509
## pop.18_34       0.074531907  0.08994063  0.250584290   0.45609703   0.033975512
## pop.65_plus     0.053278417 -0.03529032 -0.268251758  -0.33922877   0.006578474
## doctors         0.950464395  0.82045948 -0.004248085   0.23676547   0.064136254
## hosp.beds       1.000000000  0.85684988 -0.111916382   0.10042653   0.172793840
## crimes          0.856849883  1.00000000 -0.106328401   0.07707652   0.164405659
## pct.hs.grad    -0.111916382 -0.10632840  1.000000000   0.70778672  -0.691750483
## pct.bach.deg    0.100426534  0.07707652  0.707786723   1.00000000  -0.408423848
## pct.below.pov   0.172793840  0.16440566 -0.691750483  -0.40842385   1.000000000
## pct.unemp       0.007523992  0.04355675 -0.593595788  -0.54090691   0.436947236
## per.cap.income  0.194808180  0.11753914  0.522996133   0.69536186  -0.601725039
## tot.income      0.902061545  0.84309805  0.043355729   0.22223013  -0.038739339
##                   pct.unemp per.cap.income  tot.income
## land.area       0.199209277    -0.18771513  0.12707426
## pop             0.005351703     0.23561019  0.98674763
## pop.18_34      -0.278527058    -0.03164843  0.07116151
## pop.65_plus     0.236309411     0.01859071 -0.02273315
## doctors        -0.050516116     0.31613462  0.94811057
## hosp.beds       0.007523992     0.19480818  0.90206155
## crimes          0.043556752     0.11753914  0.84309805
## pct.hs.grad    -0.593595788     0.52299613  0.04335573
## pct.bach.deg   -0.540906913     0.69536186  0.22223013
## pct.below.pov   0.436947236    -0.60172504 -0.03873934
## pct.unemp       1.000000000    -0.32214439 -0.03387633
## per.cap.income -0.322144395     1.00000000  0.34768161
## tot.income     -0.033876330     0.34768161  1.00000000
```

From the ouptut,

• Total personal income and Total population are highly correlated, which is not surprising. • Both of Total personal income and Total population are reasonably highly correlated with Total serious crimes, Number of hospital beds and Number of active physicians. • The three variables Total serious crimes, Number of hospital beds and Number of active physicians seem strongly correlated with one another. • Per Capita Income isn't really highly correlated with anything, but the best possibilities seem to be Percent high school graduates, Percent bachelor's degrees (positively correlated with Per Capita Income) and Percent below poverty level, Percent unemployment (negatively correlated with Per Capita Income); all four of these variables are moderately highly correlated with one another

Part C:Crime rate and Region

As skewness of some variables is discovered as a potential problem in the histograms we draw before, we first log transformed number of active physicians, land area, number of hospital beds, and total serious crimes because they were all right skewed.

```
cdigood <- data.frame(cdinumeric,region=cdi$region)
cdigood$log.land.area<-log(cdigood$land.area)
cdigood$log.doctors<-log(cdigood$doctors)
```

```
cdigood$log.hosp.beds<-log(cdigood$hosp.beds)
cdigood$log.crimes<-log(cdigood$crimes)
cdigood$log.per.cap.income<-log(cdigood$per.cap.income)
```

To answer question 2, we created 6 models: regress per capita income on total serious crimes (Model 1, baseline model), regress per capita income on total crimes and region (Model 2), regress per capita income on total crimes and region with interaction between total serious crimes and region (Model 3), regress per capita income on crime rate (Model 4, baseline model), regress per capita income on crime rate and region (Model 5), regress per capita income on crime rate and region with interaction between crime rate and region (Model 6).

```
ancova.01 <- lm(log.per.cap.income ~ log.crimes,data=cdigood)
ancova.02 <- lm(log.per.cap.income ~ log.crimes + region,data=cdigood)
ancova.03 <- lm(log.per.cap.income ~ log.crimes * region,data=cdigood)
anova(ancova.01,ancova.02, ancova.03)
```

```
## 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
```

```
attach(cdigood)
per.cap.crime <- crimes/pop
log.per.cap.crimes <-log(per.cap.crime)
detach()
ancova.04 <- lm(log.per.cap.income ~ log.per.cap.crimes,data=cdigood)
ancova.05 <- lm(log.per.cap.income ~ log.per.cap.crimes + region,data=cdigood)
ancova.06 <- lm(log.per.cap.income ~ log.per.cap.crimes * region,data=cdigood)
anova(ancova.04,ancova.05, ancova.06)
```

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

The ANOVA test result indicates that either the interaction between total serious crimes and region or the interaction between crime rate and region are not necessary in the model as their coefficients are not statistically significant. However, region should be added in the model as a predictor variable.

```
data.frame(AIC=AIC(ancova.01,ancova.02,ancova.03,ancova.04,ancova.05,ancova.06),
BIC=BIC(ancova.01,ancova.02,ancova.03,ancova.04,ancova.05,ancova.06))[,-3] %>%
kbl(booktabs=T,col.names=c("df","AIC","BIC")) %>% kable_classic(full_width=F)
```

	df	AIC	BIC
ancova.01	3	-169.9466	-157.6863
ancova.02	6	-227.4746	-202.9539
ancova.03	9	-223.7402	-186.9593
ancova.04	3	-135.0340	-122.7737
ancova.05	6	-172.1347	-147.6140
ancova.06	9	-166.7601	-129.9792

This table shows the AIC and BIC of these models, as our final model should be chosen from Model 2 and Model 5, by comparing the AIC and BIC of these two models, our final best model to predict per capita income from crimes is the Model 2 as both AIC and BIC of Model 2 are smaller.

The final best model is ancova.02

```
summary(ancova.02)
```

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

• All across the US, for every 1% increase in per-capita crime, there is an associated 0.04% increase in per-capita income (so, slightly smaller effect, but still statistically significant; should that matter?). • The regional baseline salaries are: NC: $20,743.74, NE: $23,155.79, S: is $19,341.34, and W: $20,332.99. All but the W region have baselines that are, according to the model, significantly different from the NC baseline. • Again, the level of salary varies with region, but not the way it varies with crime, according to the model.

Part D: Predict Income

As we mentioned some collinearity between variables and some meaningless variables as well as some categorical variables with too many unique values, we dropped total population, ID, total personal income, state and county in our modeling here.

The variables we used in the model

```
new<-cdigood[,c(3,4,8,9,10,11,14:19)]
names(new)
```

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

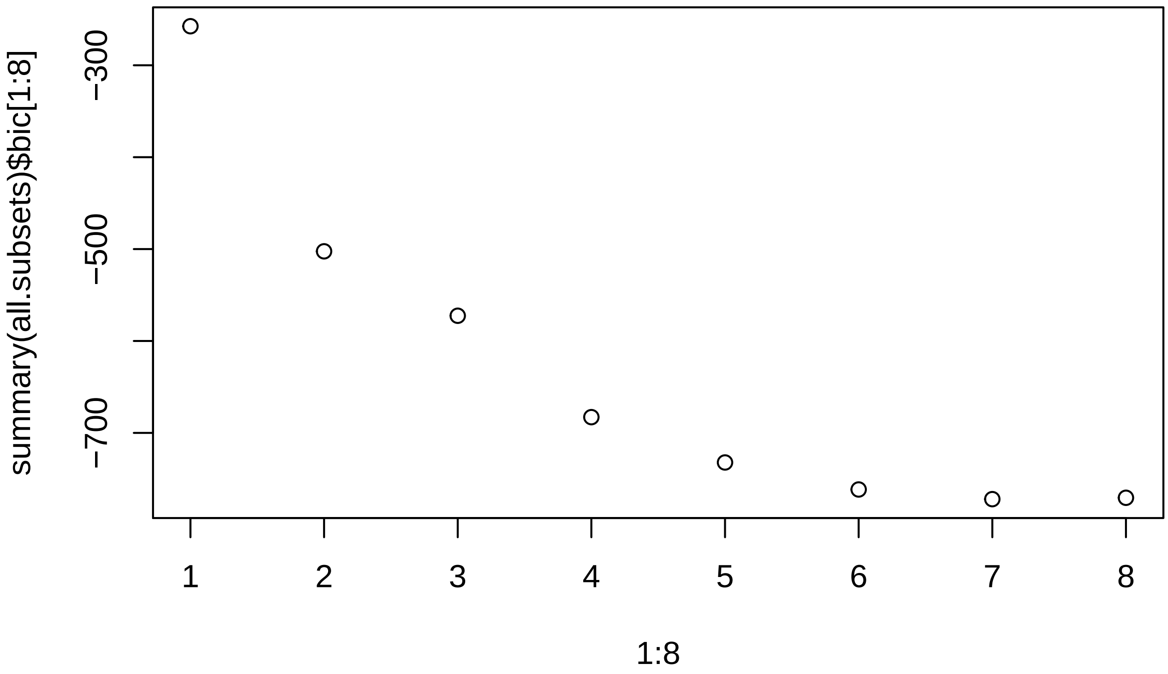
As per capita income is actually just total personal income/population, so we cannot use total personal income in our model, also for this reason, we should exclude total population in our model.

Also, the ID, Country, State will be no helpful for prediction but only added complexity to the model so I exclude them.

Conduct variable selection using BIC in All Subsets

Plot the BIC over each model and extract the coefficients of the one with smallest BIC

```
plot(1:8,summary(all.subsets)$bic[1:8])
```



```
summary(all.subsets)$bic[6:8]
```

```
## [1] -761.5908 -772.0715 -770.5990
```

```
coef(all.subsets,6:8)
```

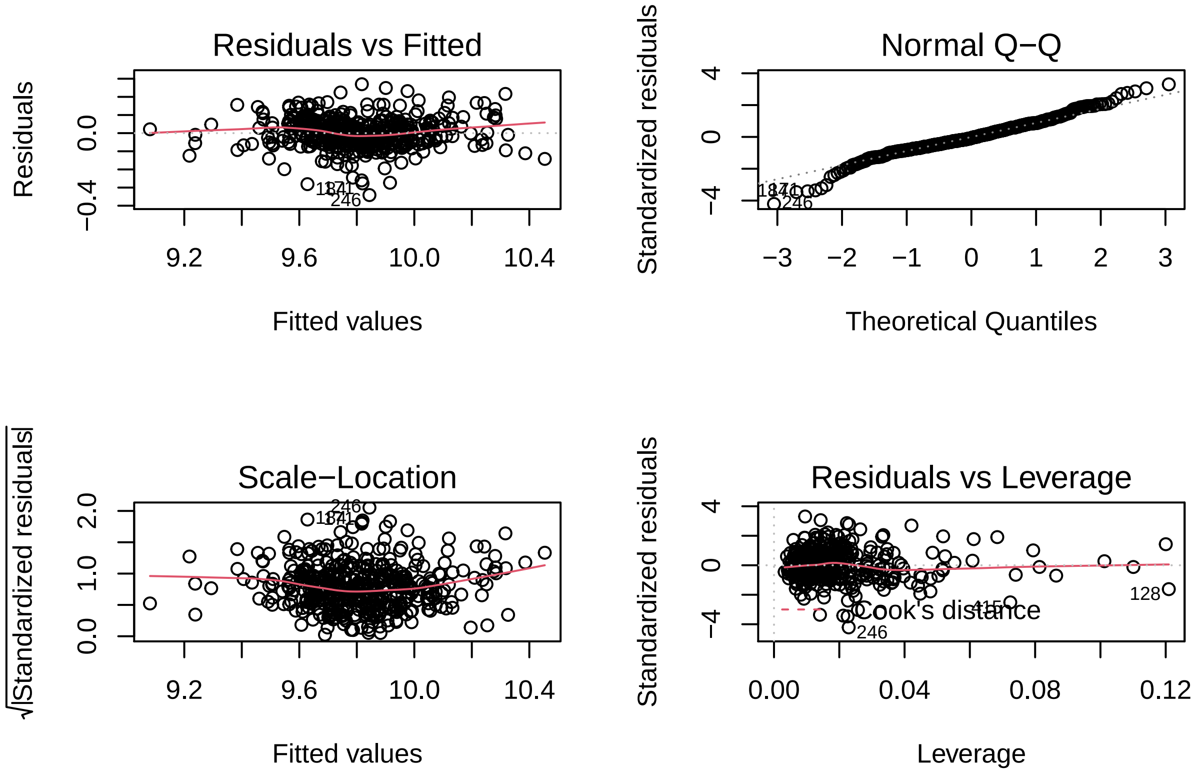
```
## [[1]]
##   (Intercept)     pop.18_34  pct.bach.deg pct.below.pov     pct.unemp
##    9.90343798   -0.01409166    0.01341559   -0.02138922    0.01290540
## log.land.area   log.doctors
##   -0.04021183    0.06286862
##
## [[2]]
##   (Intercept)     pop.18_34   pct.hs.grad  pct.bach.deg pct.below.pov
##  10.222495041  -0.013900201  -0.004406396   0.015385301  -0.024278371
##     pct.unemp log.land.area   log.doctors
##   0.010603691  -0.035674062   0.060676872
##
## [[3]]
##   (Intercept)     pop.18_34   pop.65_plus   pct.hs.grad  pct.bach.deg
##  10.315966592  -0.015348817  -0.002766377  -0.004657948   0.015214937
## pct.below.pov     pct.unemp log.land.area   log.doctors
##  -0.024614405   0.010768825  -0.036493494   0.062605267
```

Name the best model as best1 and evaluate its performance by diagnostics plots.

```
best1<-lm(log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors + pct.hs.grad + pct.bach.deg + p
summary(best1)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp, data = new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34147 -0.04886 -0.00538  0.04818  0.26969
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.2224950  0.0931210 109.776  < 2e-16 ***
## log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
## pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
## log.doctors    0.0606769  0.0040183  15.100  < 2e-16 ***
## pct.hs.grad   -0.0044064  0.0010823  -4.071 5.56e-05 ***
## pct.bach.deg   0.0153853  0.0009246  16.641  < 2e-16 ***
## pct.below.pov -0.0242784  0.0012583 -19.294  < 2e-16 ***
## pct.unemp      0.0106037  0.0021771   4.871 1.56e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.082 on 432 degrees of freedom
## Multiple R-squared:  0.8452, Adjusted R-squared:  0.8427
## F-statistic: 336.9 on 7 and 432 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(best1)
```

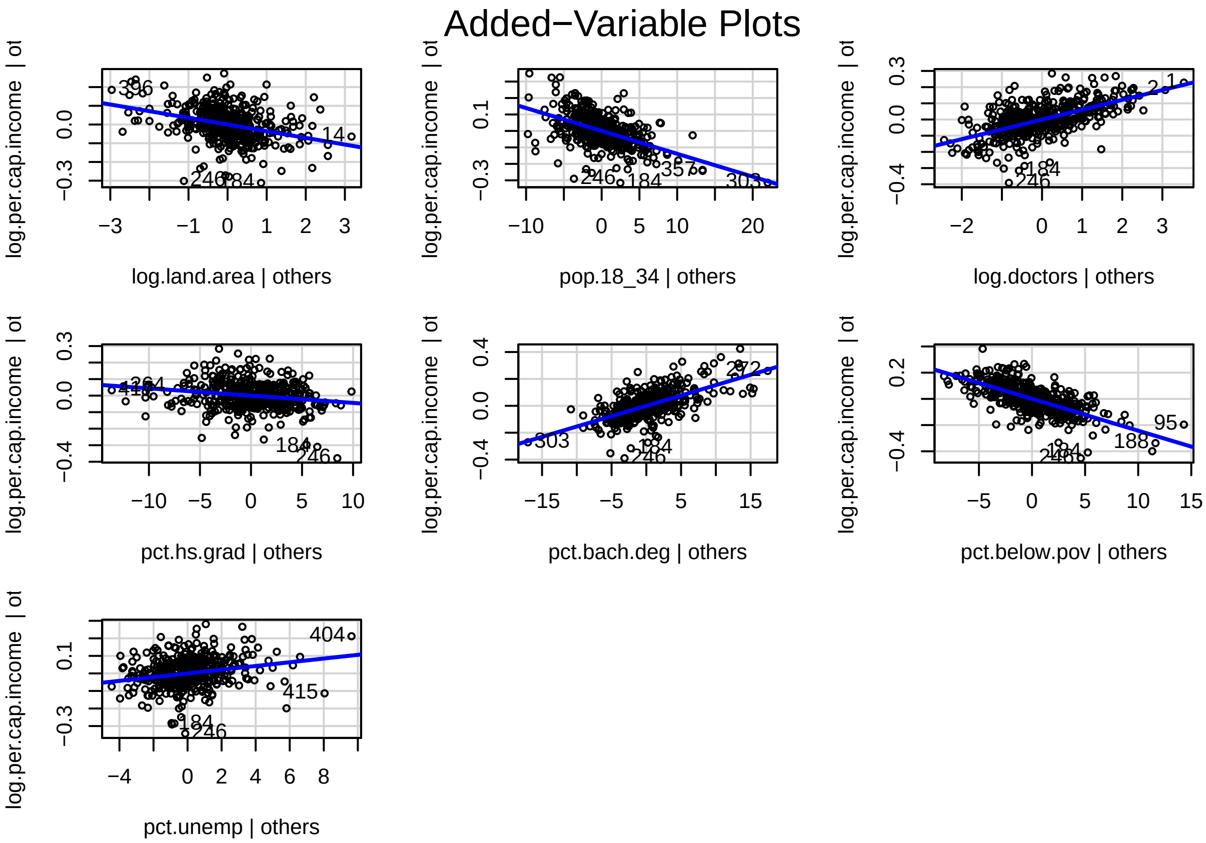


For the diagnostics plot of our final model, the Residuals vs Fitted plot shows that there is not nonlinear pattern in the model but there are some outliers detected; the normal Q-Q plot shows that the residuals are normally distributed and the assumption that errors are normally distributed holds, but there are some points that deviated from the diagonal line a lot; the scale-location plot shows that the residuals are spread equally along the ranges of predictors and the residuals have constant variance; the residuals vs leverage plot shows that there are no influential points in the model.

To test whether the variable transformation and selection makes sense as well as whether model is a good fit, diagnostics plots, added-variable plots and marginal model plots were made to evaluate the model performance.

evaluate its performance by marginal model plots

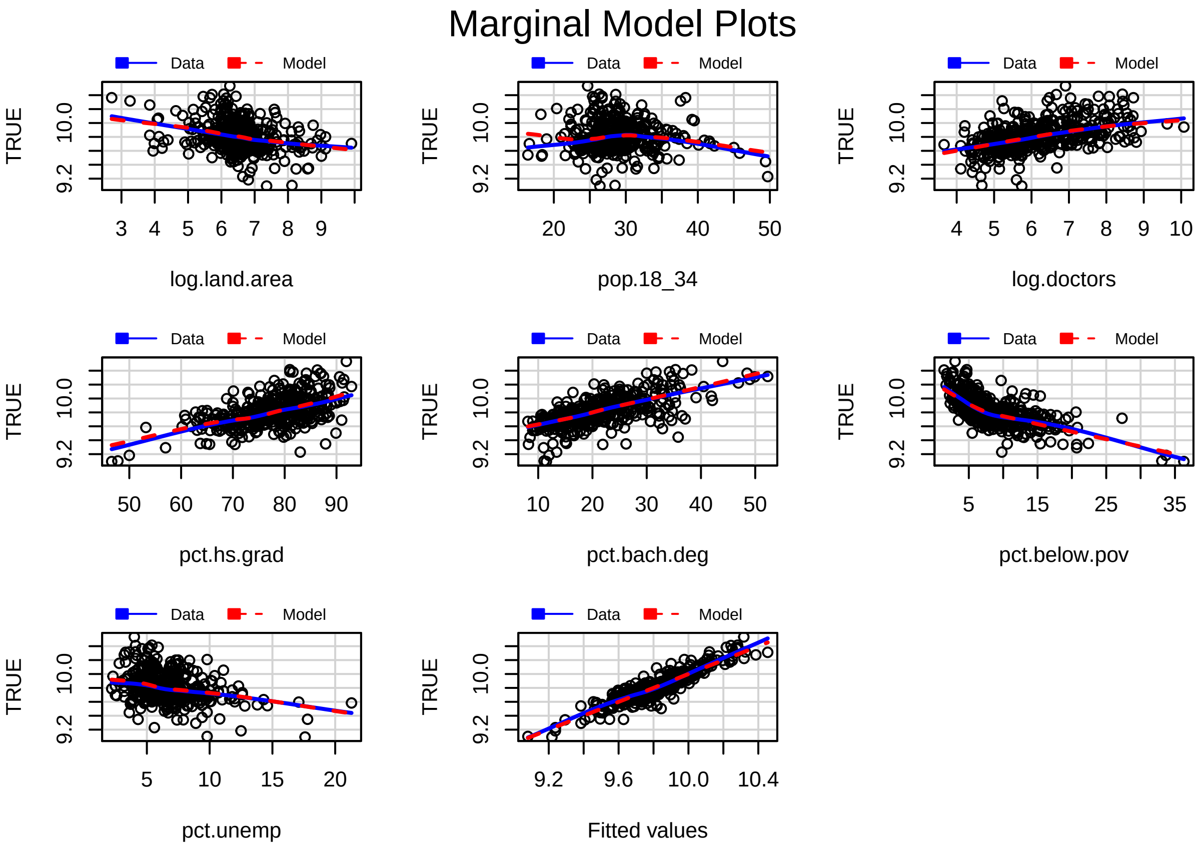
```
avPlots(best1)
```



The Added-Variable plots show that the model won't need any transformation to the variables included in the model.

evaluate its performance by marginal model plots.

```
mmps(best1)
```



The marginal model plots look very good – the blue data-based curves line up well with the red model-based curves. We don't seem to be missing any important transformations, interactions, etc.

Therefore, for this All Subsets method, the best one is to regress logged per capita income on logged land area, logged number of physicians, Percent of population aged 18–36, Percent bachelor's degrees, Percent below poverty level, Percent unemployment, Percent high school graduates.

Except for All Subsets, we also used Stepwise Regression to do the variables regression and this time also exclude categorical variable region.

Conduct variable selection using BIC in Stepwise Regression

```
stepwise.base <- lm(log.per.cap.income ~.-region,data=new)
step.result.01.bic <- stepAIC(stepwise.base,scope=list(lower = ~ 1, upper = ~ .),k=log(dim(new)[1]),tra
anova(best1,step.result.01.bic)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp + log.land.area + log.doctors
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1    432 2.9051
## 2    432 2.9051  0         0
```

The best model selected by Stepwise Regression using BIC is the same as the best one selected by All Subsets.

Conduct variable selection using AIC in Stepwise Regression

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

The best model selected by Stepwise Regression using AIC is slightly different.

Model Selection using ANOVA

```
anova(best1,step.result.01.aic)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##     log.doctors
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    432 2.9051
## 2    431 2.8748  1  0.030306 4.5437 0.03361 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compared to the previous best model, it has one more variable, which is Percent of population 65 or older. To decide whether to include this variable in the model, an ANOVA test was conducted to address this problem. The result rejects the null hypothesis that Percent of population 65 or older is not needed in the model. We also checked the summary of the model

```
summary(step.result.01.aic)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##     log.doctors, data = new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35756 -0.04551 -0.00543  0.04844  0.27399
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.3159666  0.1025858 100.559  < 2e-16 ***
## pop.18_34     -0.0153488  0.0012988 -11.818  < 2e-16 ***
## pop.65_plus   -0.0027664  0.0012978  -2.132   0.0336 *
## pct.hs.grad   -0.0046579  0.0010843  -4.296 2.15e-05 ***
## pct.bach.deg   0.0152149  0.0009242  16.462  < 2e-16 ***
## pct.below.pov -0.0246144  0.0012631 -19.488  < 2e-16 ***
## pct.unemp      0.0107688  0.0021696   4.963 9.99e-07 ***
## log.land.area -0.0364935  0.0047728  -7.646 1.36e-13 ***
## log.doctors    0.0626053  0.0041029  15.259  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08167 on 431 degrees of freedom
## Multiple R-squared:  0.8468, Adjusted R-squared:  0.8439
## F-statistic: 297.7 on 8 and 431 DF,  p-value: < 2.2e-16
```

Although the coefficient on Percent of population 65 or older is significantly different from zero, its effect on expected per-capita income appears to be quite small. Considering that our model should best reflects the social science and the meaning of the variables, we decided to exclude it from the model.

The final step is to decide whether to include any interaction between the remaining numeric variables and the categorical variables region. We did same thing as we did for addressing question 2 — we conducted multiple AVOVA test for selection of interaction terms. We tested the interaction between the categorical variable region and all the other continuous variables pair by pair and select every interaction whose coefficient is statistically significant (here I mean the one whose p vale is with more than two stars in the summary)

```
fit_final<-lm(log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors + pct.hs.grad + pct.bach.deg

m1<-lm(log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors + pct.hs.grad + pct.bach.deg + pct.

m2<-lm(log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors + pct.hs.grad + pct.bach.deg + pct.

m3<-lm(log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors + pct.hs.grad + pct.bach.deg + pct.

m4<-lm(log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors + pct.hs.grad + pct.bach.deg + pct.

m5<-lm(log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors + pct.hs.grad + pct.bach.deg + pct.

m6<-lm(log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors + pct.hs.grad + pct.bach.deg + pct.

m7<-lm(log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors + pct.hs.grad + pct.bach.deg + pct.

anova(fit_final,m1)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     log.land.area * region
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    432 2.9051
## 2    426 2.8328  6  0.072284 1.8117 0.09528 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_final,m2)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     pop.18_34 * region
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    432 2.9051
## 2    426 2.8025  6    0.1026 2.5994 0.01744 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_final,m3)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     log.doctors * region
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    432 2.9051
## 2    426 2.8082  6  0.096909 2.4502 0.02433 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_final,m4)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     +pct.hs.grad * region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    432 2.9051
## 2    426 2.7350  6   0.17004 4.4142 0.0002411 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_final,m5)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     pct.bach.deg * region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    432 2.9051
## 2    426 2.7318  6   0.17329 4.5039 0.0001938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_final,m6)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     pct.below.pov * region
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    432 2.9051
## 2    426 2.7850  6   0.12011 3.0621 0.006056 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(fit_final,m7)
```

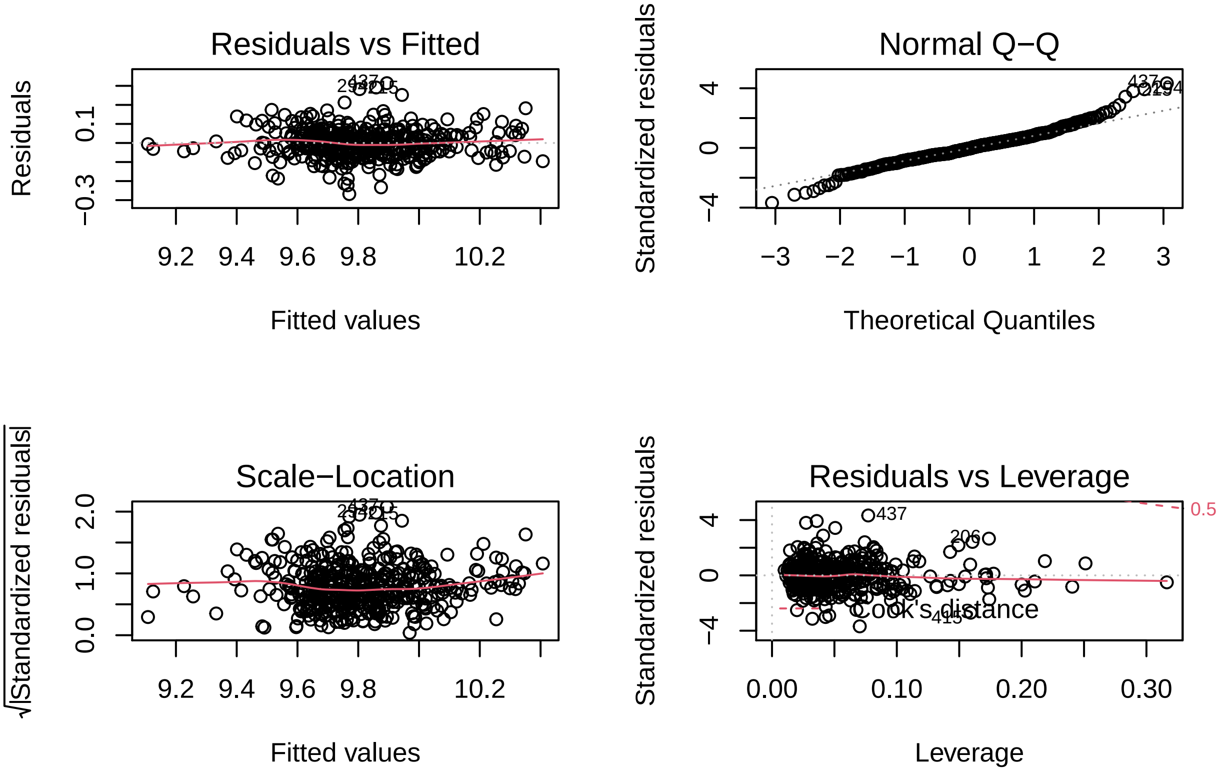
```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     pct.unemp * region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    432 2.9051
## 2    426 2.7347  6   0.17041 4.4243 0.0002352 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6 out of 7 interaction terms have their p values less than 0.05. If the p value is very small, here we mean less than 0.05, it means this interaction is significantly different from zero. However, adding so many significant interaction terms will make our model hard to interpret. Therefore, we changed the significance level from 0.05 to 0.01 to do the interaction terms selection, by this way we are able to reduce the significant interaction terms to just 3, which are interaction between region and Percent high school graduates, Percent bachelor's degrees, Percent below poverty level and Percent unemployment.

```
m_final<-lm(log.per.cap.income ~ log.land.area + pop.18_34 + log.doctors + pct.hs.grad + pct.bach.deg +
summary(m_final)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.land.area + pop.18_34 +
##     log.doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp + pct.hs.grad * region + pct.bach.deg * region +
##     pct.below.pov * region + pct.unemp * region, data = new)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.268015 -0.043459 -0.002511  0.039967  0.313939
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.125260   0.251582  40.246  < 2e-16 ***
## log.land.area          -0.034569   0.005376  -6.430 3.50e-10 ***
## pop.18_34              -0.015404   0.001087 -14.170  < 2e-16 ***
## log.doctors             0.055342   0.004034  13.720  < 2e-16 ***
## pct.hs.grad            -0.002503   0.003151  -0.794 0.427456
## pct.bach.deg            0.014208   0.002108   6.741 5.24e-11 ***
## pct.below.pov          -0.023634   0.003351  -7.054 7.30e-12 ***
## pct.unemp               0.017787   0.004783   3.719 0.000228 ***
## regionNE                0.219429   0.302526   0.725 0.468661
## regionS                -0.062648   0.276125  -0.227 0.820627
## regionW                 1.629351   0.357633   4.556 6.86e-06 ***
## pct.hs.grad:regionNE   -0.003640   0.003876  -0.939 0.348271
## pct.hs.grad:regionS     0.002014   0.003539   0.569 0.569690
## pct.hs.grad:regionW    -0.018916   0.004204  -4.499 8.85e-06 ***
## pct.bach.deg:regionNE   0.005905   0.002618   2.256 0.024611 *
## pct.bach.deg:regionS   -0.001298   0.002321  -0.559 0.576352
## pct.bach.deg:regionW    0.006326   0.002620   2.415 0.016183 *
## pct.below.pov:regionNE -0.002435   0.004647  -0.524 0.600488
## pct.below.pov:regionS   0.007137   0.003686   1.937 0.053482 .
## pct.below.pov:regionW  -0.015224   0.005169  -2.945 0.003407 **
## pct.unemp:regionNE     -0.007967   0.007255  -1.098 0.272761
## pct.unemp:regionS      -0.024668   0.006377  -3.868 0.000127 ***
## pct.unemp:regionW      -0.019757   0.006603  -2.992 0.002935 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07545 on 417 degrees of freedom
## Multiple R-squared:  0.8735, Adjusted R-squared:  0.8668
## F-statistic: 130.9 on 22 and 417 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(m_final)
```



For the diagnostics plot of our final model, the Residuals vs Fitted plot shows that there is not nonlinear pattern in the model but there are some outliers detected; the normal Q-Q plot shows that the residuals are normally distributed and the assumption that errors are normally distributed holds, but there are some points that deviated from the diagonal line a lot; the scale-location plot shows that the residuals are spread equally along the ranges of predictors and the residuals have constant variance; the residuals vs leverage plot shows that there are no influential points in the model. Hence, the final model is a good fit.

Part E: Whether missingness matters

Alaska, Iowa, and Wyoming are the states that are missing and these are states that are perhaps less populous.

```
unique(cdi$state)
```

```
##  [1] "CA" "IL" "TX" "NY" "AZ" "MI" "FL" "PA" "WA" "OH" "MA" "MN" "MO" "WI" "CT"
## [16] "HI" "TN" "NJ" "VA" "IN" "MD" "NV" "UT" "KY" "AL" "GA" "DC" "OK" "RI" "OR"
## [31] "NC" "LA" "NM" "CO" "DE" "NE" "KS" "AR" "NH" "SC" "MS" "ME" "WV" "ID" "VT"
## [46] "SD" "MT" "ND"
```

```
length(unique(cdi$state))
```

```
## [1] 48
```