

[1. TITLE] The Truth Behind Income: Analysis of Socioeconomic Factors in Populous U.S. Counties  
[AUTHOR] Anirban Chowdhury ([achowdh1@andrew.cmu.edu](mailto:achowdh1@andrew.cmu.edu))  
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh PA, 15213

## [ABSTRACT]

It is widely known that a person's income can depend on several factors of their environment; but the actual relationships between these socioeconomic variables and income are not widely known and discovering these trends is the goal of this research project. We use data collected by Kutner et al (2005) regarding county demographic information like percent unemployed, population in certain age bins, income, percent below poverty level, etc. from the 440 most populated counties in the US from 1990 to 1992. We examined this data through regression analysis; we fit a multiple linear regression model with interactions in order to estimate the relationships between income per capita and the other socioeconomic factors in our dataset. From this model, we were able to estimate coefficients with decent predictive accuracy and valid assumptions and we discovered that, in line with our intuition, higher percentage of bachelor's degrees corresponded with increased income per capita and higher percentage of unemployment corresponded with decreased income per capita (controlling for other variables). We paid particular attention to crime, and discovered that when looking at total crime and marginalizing over all other variables the relationship with income is significant (although later on we saw crime does not lead to any increase in predictive accuracy). In context, our findings mean that there are in fact important relationships between socioeconomic factors and income and even the geographic location. These findings can be used to proactively target counties at risk of having a low per capita income for policy-based relief efforts.

[KEY WORDS] Regression analysis, income, unemployment, education

## [INTRODUCTION]

Several factors define a person's income: education level, location, occupation, etc. However, there are also socioeconomic factors outside of an individual's control that also influence their income. The purpose of this research project is to explore these external variables and assess their impact on income per capita in the U.S. In particular, areas with high crime could be related to low income (as low income per capita could encourage crime). This relationship is worth investigating. Similarly, factors like poverty prevalence, age distribution, and average education can also have an impact on an area's per capita income. In this research, we examine which of these socioeconomic factors are relevant via multiple linear regression. We are interested in multiple research questions: what relationships exist between different socioeconomic variables, the specific relationship and relevance in using crime to model income, and how we can predict income using all the variables available to us.

## [DATA]

Kutner et al (2005) collected socioeconomic data from the 440 most populated counties in the US. Each row in the dataset corresponds to a county (there are some duplicate county names because certain county names collide across states, e.g. Jefferson county is a name in 7 states) and there are 14 total other variables. There are also only 48 states instead of 50. This is to be expected however, the data only contains the most populous counties, so states in the west that do not have many people in any county would not show up. So, even though there are only 373 counties, there are exactly 440 county-state pairings. Thus we are not concerned by this apparent inconsistency.

Another data issue comes from the sample bias. Since we only look at populous counties, it's entirely possible that the distribution of these socioeconomic factors and their relationship with per capita income differs in lower populated states or counties. A quick examination of the available data shows us

that Alaska and Wyoming, two very sparsely populated states have no representation in the dataset. Additionally, populated states like NY, PA, CA, and FL have much more representation in the dataset than others. So, we should take efforts to mitigate this bias in later work.

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: CDI Data from Kutner et al.

Table 1 presents a general description of all the variables present in the dataset we used. In particular, we have multiple socioeconomic statistics like % unemployed, % below poverty level, and geographic area. First, we can take a look at summary statistics for the quantitative variables.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
land.area	15.0	451.250	656.50	1.041411e+03	946.750	20062.0
pop	100043.0	139027.250	217280.50	3.930109e+05	436064.500	8863164.0
pop.18_34	16.4	26.200	28.10	2.856841e+01	30.025	49.7
pop.65_plus	3.0	9.875	11.75	1.216977e+01	13.625	33.8
doctors	39.0	182.750	401.00	9.879977e+02	1036.000	23677.0
hosp.beds	92.0	390.750	755.00	1.458627e+03	1575.750	27700.0
crimes	563.0	6219.500	11820.50	2.711162e+04	26279.500	688936.0
pct.hs.grad	46.6	73.875	77.70	7.756068e+01	82.400	92.9
pct.bach.deg	8.1	15.275	19.70	2.108114e+01	25.325	52.3
pct.below.pov	1.4	5.300	7.90	8.720682e+00	10.900	36.3
pct.unemp	2.2	5.100	6.20	6.596591e+00	7.500	21.3
per.cap.income	8899.0	16118.250	17759.00	1.856148e+04	20270.000	37541.0
tot.income	1141.0	2311.000	3857.00	7.869273e+03	8654.250	184230.0

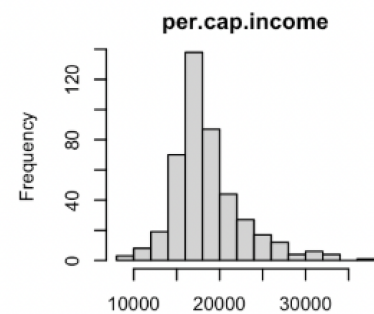
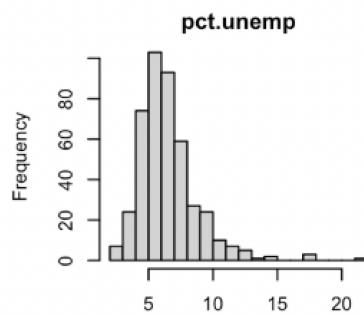
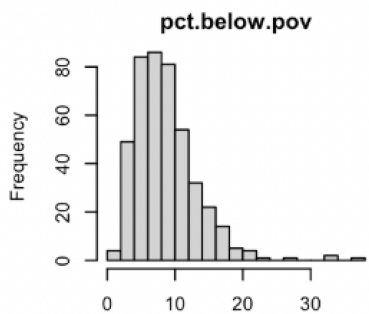
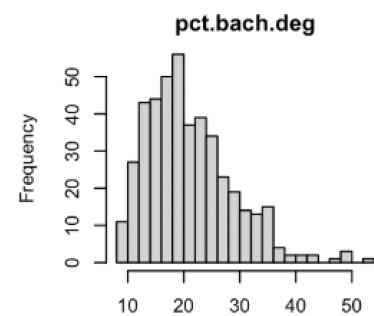
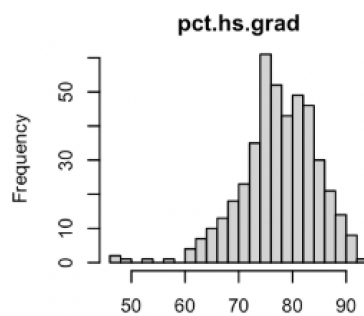
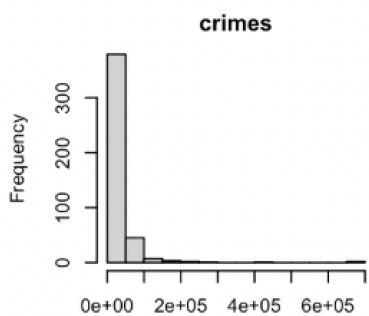
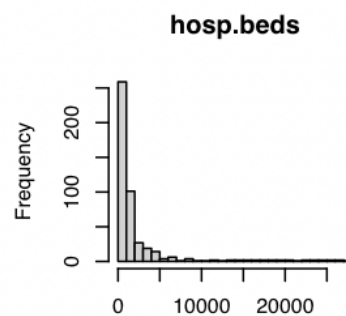
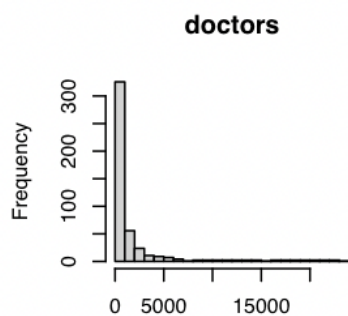
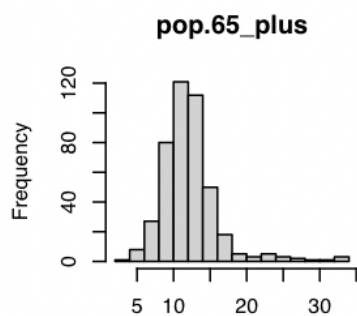
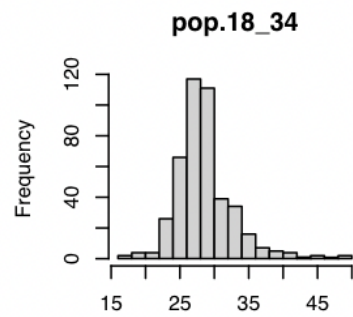
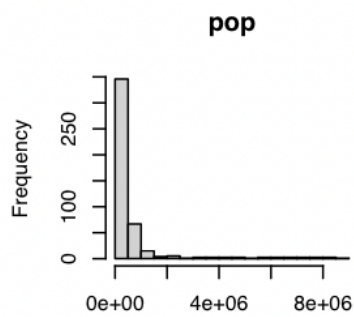
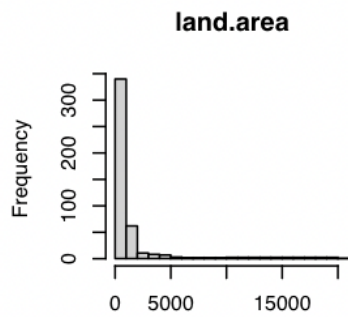
TABLE 2: Quantitative Variable Summaries

Var1	Freq
NC	0.2454545
NE	0.2340909
S	0.3454545
W	0.1750000

TABLE 3: Summary of Region

Table 2 shows a 5-number summary for every quantitative variable and Table 3 shows the frequencies for each Region in the dataset. See Appendix Page 1-3 for the table for State. The table for County is omitted because county is our row identifier.

Next, we explore what the general distributions for all these variables look like.



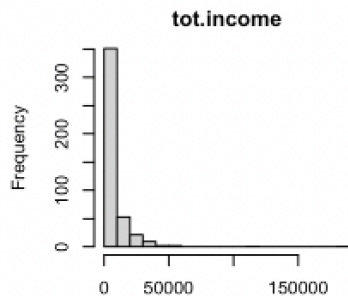


FIGURE 1: Histograms of Quantitative Variables

From the histograms, we can see several relevant variables are skewed. Since we are interested in examining crime rate, we look closely at the distributions of crime and population and find that they are both right skewed. Since these will likely be involved in our model (either crimes or crimes per capita) it is important to keep in mind that a transformation might be necessary. the same is true for total income, but not per capita income. A lot of the other aggregate statistics (total doctors, hospital beds) also look right skewed.

We can next examine the relationships between different variables to assess which ones exist and whether or not they align with intuition.

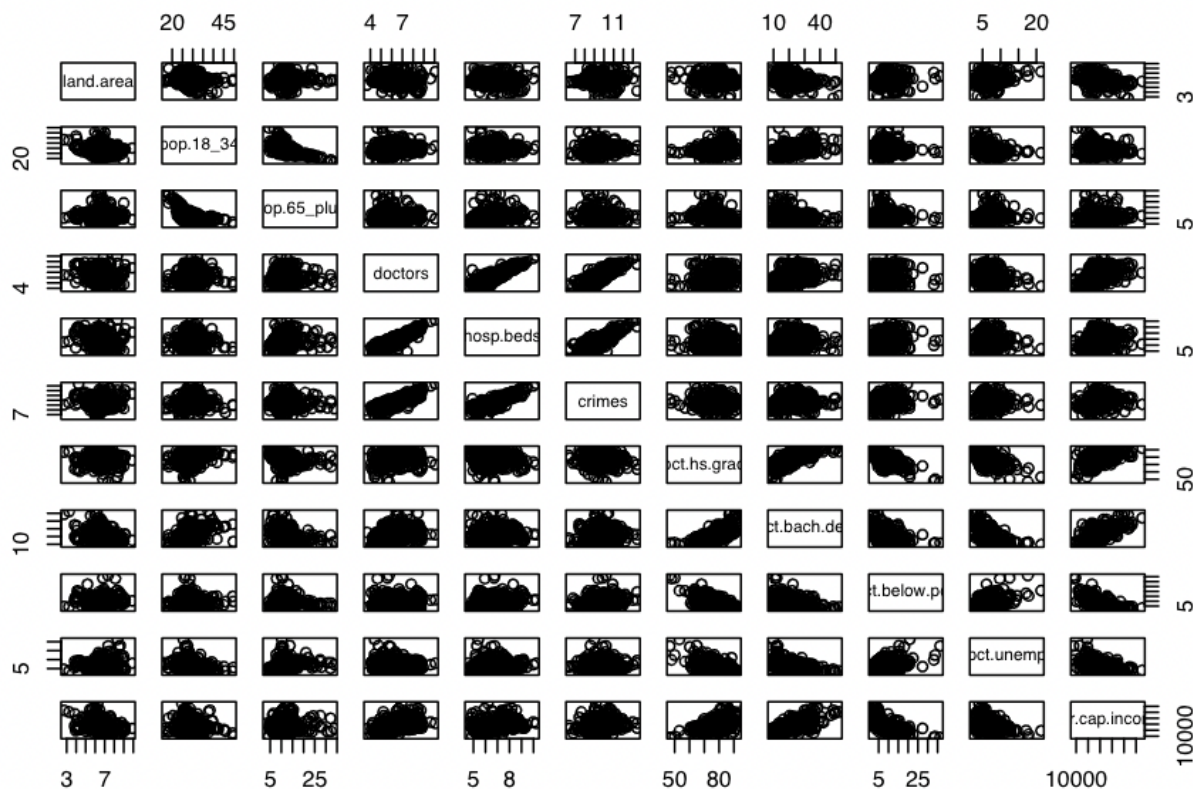


FIGURE 2: Pairs Plot



Figure 2 shows a pairs plot of all the quantitative variables. We can assess the relationships between certain variables in our data by examining the corresponding index in the scatterplot matrix. For example, we see a negative relationship between % unemployed and % bachelors, which makes sense as we would expect less people to be unemployed in areas where more people have higher education (better odds of getting a job). Similarly, we see a strong positive relationship between hospital beds and crimes, which could be because high-crime areas need to support more injuries and patients. Somewhat surprisingly, there does not seem to be any relationship between crime and unemployment or poverty, indicating that high-crime areas are not necessarily impoverished and crime is not necessarily always because of money. Next, we can examine how geographic regions impact certain variables.

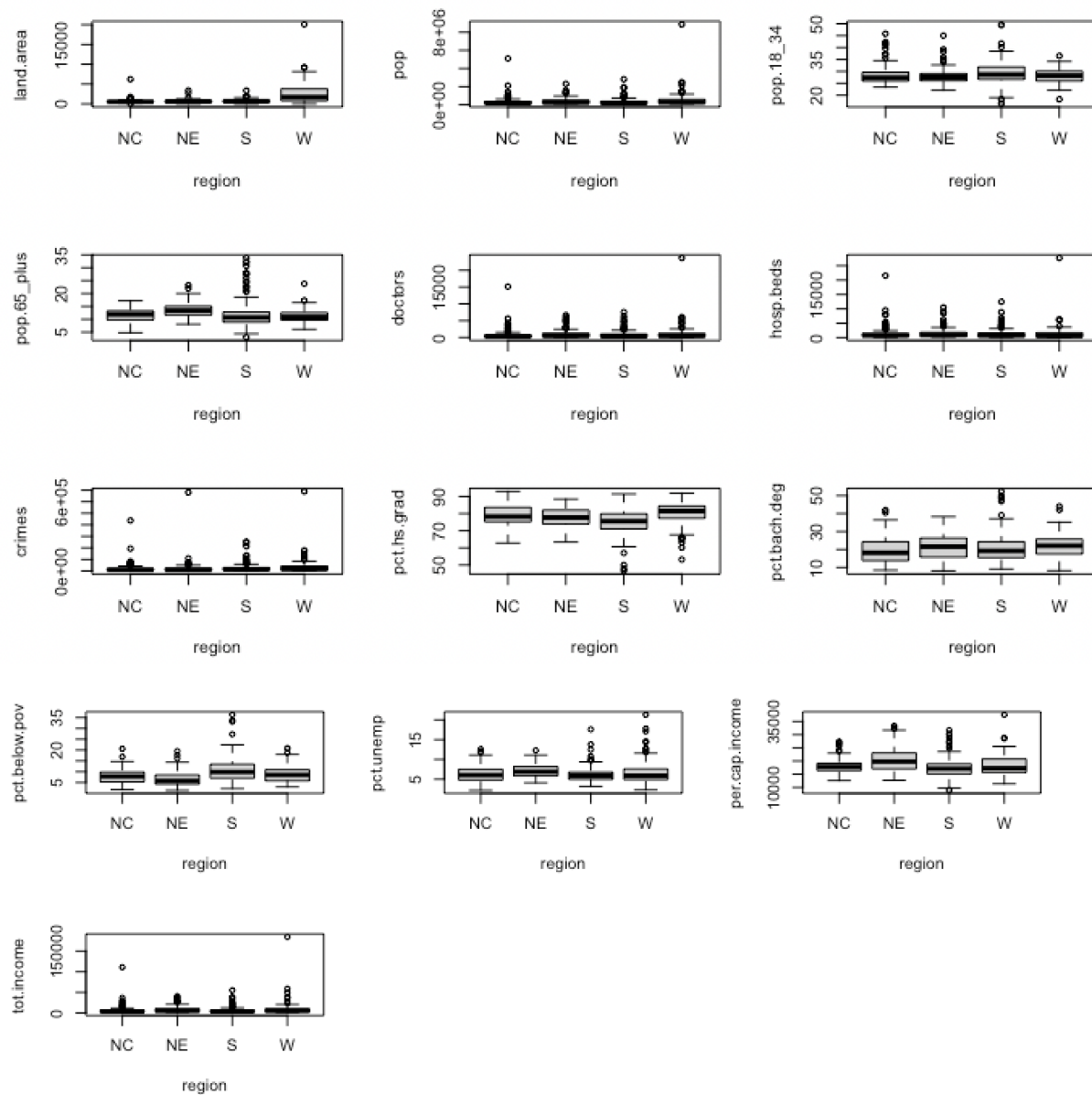
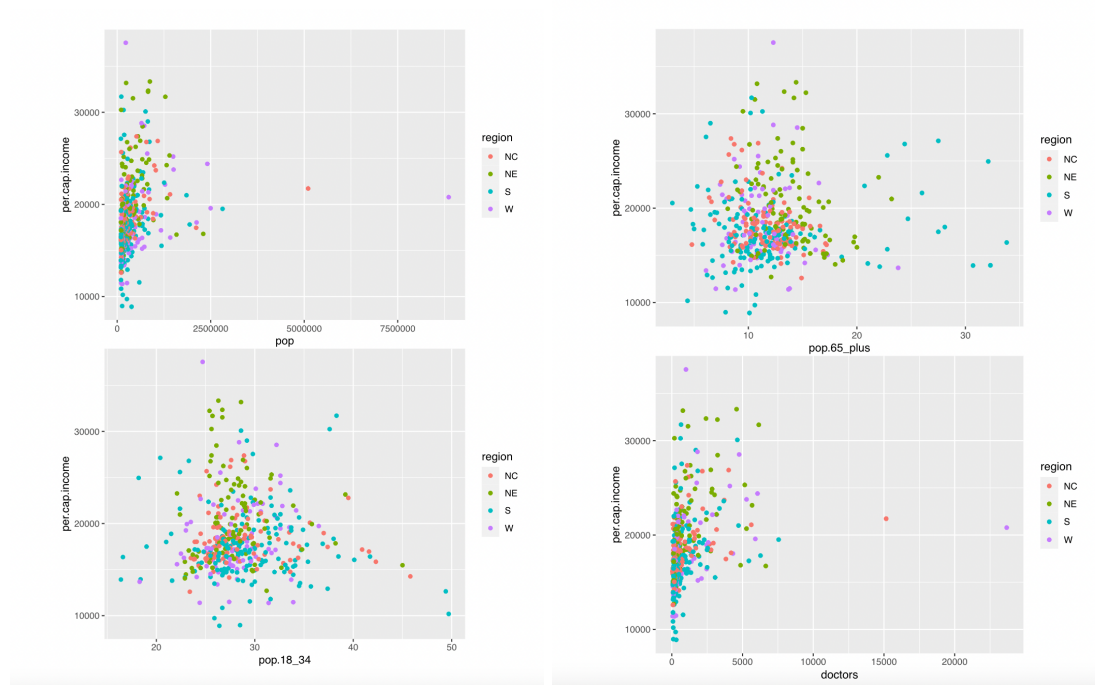


FIGURE 3: Boxplots of Region

From Figure 3, we can see that the median and quantiles of several variables change with region. It looks like the per capita income is higher in the NE and W regions, which makes sense as the tech hubs in San Francisco and NYC likely pull up these respective groups. Interestingly, there also seems to be a lower high school grad % in the South. Of course, we see a higher land area in the west, as those states are much larger.

Finally, we can examine the relationships between the categorical and quantitative variables. In particular, we are interested in predicting income so we can examine how the relationship between income and other variables changes due to region.





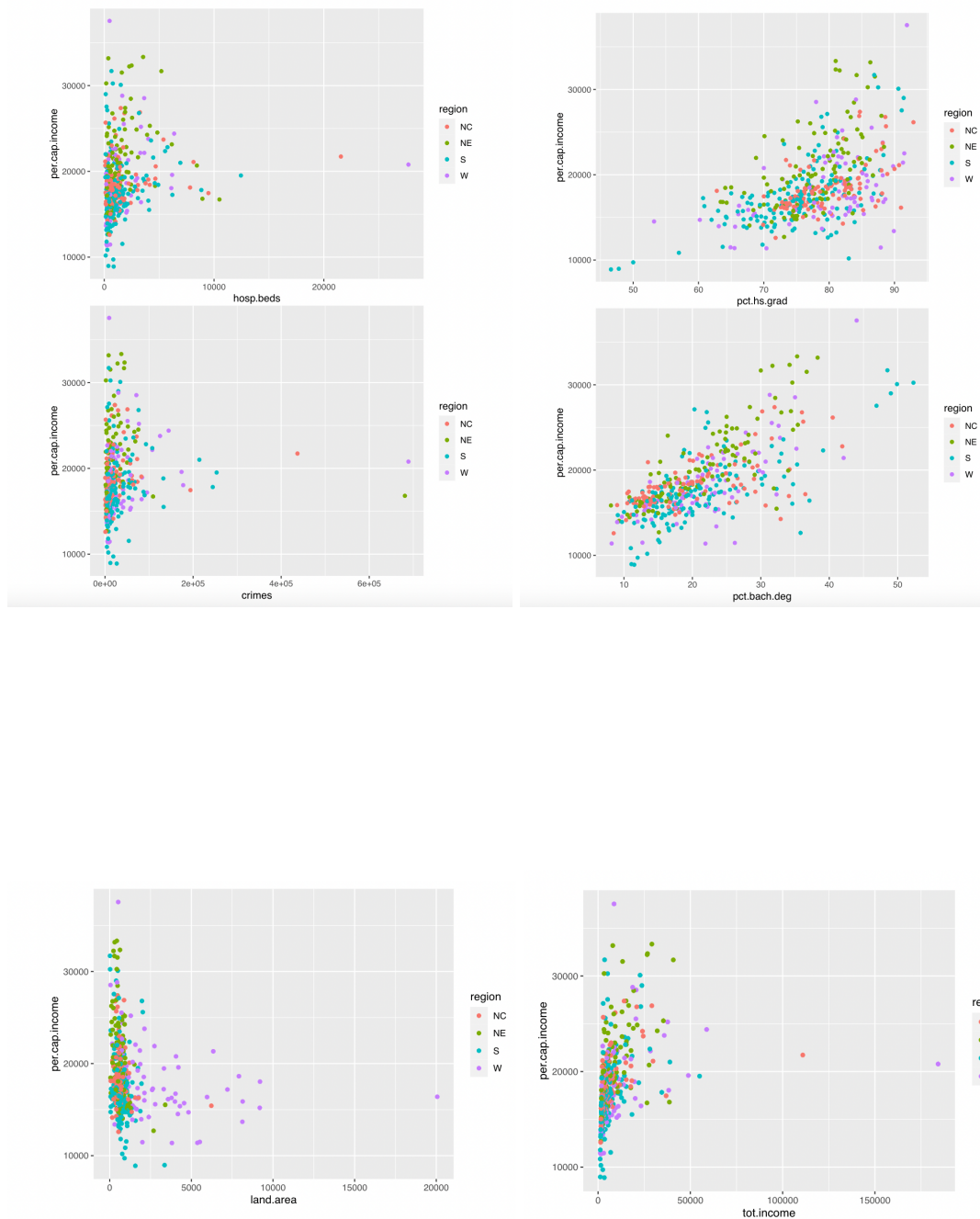


FIGURE 4: How Region affects Income vs Other Variables

Figure 4 presents a series of colored scatterplots that outline how the relationship between income and the other variables changes when region is conditioned on. In a lot of these plots, because the data is such high variance, it is difficult to specifically make out a meaningful difference in trend. When looking at crime specifically, since the relationship between crime and income is of particular interest to us, we do not see any apparent changes in slope across the four regions. However, it is important we still include interactions in our model to account for less apparent relationships that we cannot pick up from the scatterplots.

## [METHODS]

We conducted this analysis in 2 parts. In the first, we exclusively examine the effect of crime on income per capita without conditioning for any other variables and we assess whether or not interactions should be included and whether or not region influences this relationship.

The first step in the data analysis procedure for this work was the assessment of the relationship between crime and income, and how this relationship changes across regions. To this end, we first fit 2 models: a regression of income against crime and region, and a regression of income against crime, region, and the interaction between crime and region. We use ANCOVA and a partial F test to first assess whether or not the interactions are necessary, then we examine the significance of the crime and region coefficients. We repeat this analysis with a newly constructed crime per capita variable (crime / total population) in order to assess if the per capita crime value is more meaningful. We repeated the fitting and ANCOVA procedure above for this new variable as well.

Before conducting any modeling or analysis, we first refer back to Figure 1 to assess any necessary transformations of the data. Because doctors, land area, hospital beds, and crimes were all right skewed, we first mutated these variables with a log transformation. Additionally, we dropped state and county from the dataset (county is an identifier, and state was not helpful in any analysis).

For the modeling procedure, we first fit a multiple linear regression model of per capita income against all of the variables (not including crimes per capita, we just used crimes) and all interactions between these variables and region. Because many coefficients were insignificant and we had evidence of multicollinearity, we experimented with a few model selection approaches. Since stepwise regression approximates all-subsets, we proceeded with BIC criterion stepwise regression first. We then compared this model with the one chosen by all-subsets and found that they had fidelity. So, we proceed with the model agreed upon by stepwise and all subsets, then add region and interactions back in the model. Finally, we experiment with removing certain interactions and reassessing the explainability of the model with ANOVA tests. Because we are interested in interpretability and statistical claims about the estimated coefficients, we do not try penalized regression.

Next, we use marginal model plots, diagnostics, and multicollinearity assessments to validate model assumptions and come to meaningful conclusions about the relationships we are interested in.

## [RESULTS]

First, we explore the output of our analysis focusing on crime. When fitting a multiple linear regression model against total crimes and region, we get the output in figure 5.

```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9661.0 -2260.7  -618.3  1650.0 19492.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.811e+04  3.784e+02  47.846 < 2e-16 ***
## crimes       8.915e-03  3.188e-03   2.797  0.00539 **
## regionNE     2.286e+03  5.325e+02   4.293  2.17e-05 ***
## regionS     -8.606e+02  4.868e+02  -1.768  0.07782 .
## regionW     -1.428e+02  5.796e+02  -0.246  0.80548
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3866 on 435 degrees of freedom
## Multiple R-squared:  0.1011, Adjusted R-squared:  0.09288
## F-statistic: 12.24 on 4 and 435 DF,  p-value: 1.946e-09
```

FIGURE 5: Crimes and Income, controlling for Region

Note that we also tried fitting the above model with interaction, but a partial F test tells us that this model does not explain the data any better than the reduced one above (see appendix pages 13-15 for details). From this output, it appears that there is a positive relationship between crimes and income. Specifically, when controlling for region, we estimate an unit increase in crime corresponds with an average increase of  $8.9\text{e-}3$  units of per capita income. We also expect this relationship to be different based on region, as the NE term is significant. In context, we expect a unit increase in crime in a NE county to correspond to an increase of  $8.9\text{e-}3 + 2.3\text{e}+3 \approx 2.3\text{e}+3$  in units of per capita income. So, NE counties have a much steeper positive relationship on average between crime and income as opposed to NC (the base class). We can also try this analysis with crimes per capita instead of total crime (see appendix pages 13-15 for output). Upon doing so, we see no significant relationship between crime and income, and the interactions are still not meaningful. Because this provides us with no explicit benefit, we do not transform crimes in our model.

Our next step to answer the research questions is to build a model to predict income per capita. First, we remove population, total income, state, and county from the model. This is because population and total income are both directly related to the response and population is likely collinear with variables like `pop.18_34` (total population with age between 18 and 34), and state is not helpful in modeling (we tried some experiments including it but found it to not improve the fit). After these dataset mutations, we first fit a full model with all terms, including interactions. The model explained about 84% of the variation in the response, but had issues with multicollinearity and several insignificant terms. So, we proceed with variable selection. Since variable selection algorithms often do not handle categorical data and interaction

terms well, we will first select for the quantitative variables, then add region and interactions back in to improve the fit if necessary. We first tried stepwise regression with BIC, as it is a heuristic for all subsets.

```
##
## Call:
## lm(formula = per.cap.income ~ I(crimes/pop) + region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8634  -2300   -631   1710  19332
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18006.04     537.04  33.528 < 2e-16 ***
## I(crimes/pop)  5773.20    7520.41   0.768  0.4431
## regionNE      2354.70     541.97   4.345 1.74e-05 ***
## regionS       -927.45     512.31  -1.810  0.0709 .
## regionW        -34.92     586.03  -0.060  0.9525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3898 on 435 degrees of freedom
## Multiple R-squared:  0.08622,    Adjusted R-squared:  0.07782
## F-statistic: 10.26 on 4 and 435 DF,  p-value: 6.007e-08
```

FIGURE 6: BIC Stepwise Output

From the output above, we can see that the approximate best model in terms of BIC is a regression of income per capita against land area, population with ages from 18 to 34, doctors, % high school grads, %with bachelor's degrees, % below poverty, and % unemployed. When using all-subsets regression with a BIC search criterion, we arrive at an identical model. We could have also experimented with LASSO, but because penalized regression is less interpretable in context we chose to proceed with the models we found above.

Next, we add region and all interactions back into the model. Because several of the coefficients in this model are insignificant and it is difficult to assess whether the interactions are actually meaningful. The new fitted model is presented below. There are multiple interaction terms where every level is insignificant, so we have justification to try a new model with these terms taken out.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	26365.00	6037.12	4.367	1.59e-05	***
pop.18_34	-329.35	55.17	-5.970	5.13e-09	***
doctors	909.94	199.06	4.571	6.42e-06	***
pct.hs.grad	-118.62	70.93	-1.672	0.095222	.
pct.bach.deg	342.72	61.69	5.555	4.98e-08	***
pct.below.pov	-443.21	77.52	-5.718	2.08e-08	***
pct.unemp	327.90	103.56	3.166	0.001659	**
regionNE	8345.82	7647.63	1.091	0.275781	
regionS	-6483.53	6598.24	-0.983	0.326374	
regionW	27386.55	8918.87	3.071	0.002278	**
pop.18_34:regionNE	-126.80	78.30	-1.619	0.106111	
pop.18_34:regionS	86.68	64.82	1.337	0.181907	
pop.18_34:regionW	12.98	87.27	0.149	0.881820	
doctors:regionNE	-45.68	283.48	-0.161	0.872067	
doctors:regionS	-61.54	244.55	-0.252	0.801426	
doctors:regionW	-95.17	280.63	-0.339	0.734689	
pct.hs.grad:regionNE	-116.11	88.98	-1.305	0.192660	
pct.hs.grad:regionS	58.68	79.07	0.742	0.458459	
pct.hs.grad:regionW	-336.02	96.68	-3.476	0.000564	***
pct.bach.deg:regionNE	249.06	81.75	3.047	0.002464	**
pct.bach.deg:regionS	-16.38	67.49	-0.243	0.808307	
pct.bach.deg:regionW	179.27	75.49	2.375	0.018024	*
pct.below.pov:regionNE	22.11	108.66	0.203	0.838870	
pct.below.pov:regionS	126.06	86.95	1.450	0.147850	
pct.below.pov:regionW	-281.84	115.71	-2.436	0.015283	*
pct.unemp:regionNE	-156.06	157.18	-0.993	0.321346	
pct.unemp:regionS	-242.09	140.18	-1.727	0.084928	.
pct.unemp:regionW	-377.14	144.17	-2.616	0.009224	**

FIGURE 7: Selected Quantitative Variables with Region and All Interactions

In particular, the interaction between doctors and region and the interaction between population between 18 and 34 and region are insignificant at all levels, so we first try to remove these terms. We fit a new model with only the remaining interactions and region, and of course all of the quantitative variables.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	26275.113	5452.666	4.819	2.02e-06	***
pop.18_34	-311.555	23.585	-13.210	< 2e-16	***
doctors	875.820	87.533	10.006	< 2e-16	***
pct.hs.grad	-121.194	68.159	-1.778	0.076114	.
pct.bach.deg	341.063	45.709	7.462	5.00e-13	***
pct.below.pov	-444.613	72.731	-6.113	2.24e-09	***
pct.unemp	333.017	103.703	3.211	0.001424	**
regionNE	5156.517	6563.365	0.786	0.432517	
regionS	-6025.465	5981.880	-1.007	0.314380	
regionW	26275.154	7729.723	3.399	0.000741	***
pct.hs.grad:regionNE	-106.707	83.943	-1.271	0.204366	
pct.hs.grad:regionS	79.713	76.703	1.039	0.299297	
pct.hs.grad:regionW	-324.506	91.069	-3.563	0.000408	***
pct.bach.deg:regionNE	206.265	56.213	3.669	0.000275	***
pct.bach.deg:regionS	-9.271	50.357	-0.184	0.854025	
pct.bach.deg:regionW	173.775	56.612	3.070	0.002284	**
pct.below.pov:regionNE	-18.314	100.827	-0.182	0.855957	
pct.below.pov:regionS	159.038	79.973	1.989	0.047393	*
pct.below.pov:regionW	-273.588	112.101	-2.441	0.015079	*
pct.unemp:regionNE	-178.932	157.437	-1.137	0.256388	
pct.unemp:regionS	-305.351	138.413	-2.206	0.027921	*
pct.unemp:regionW	-373.807	143.196	-2.610	0.009367	**

FIGURE 8: Final Model

Figure 8 shows the coefficients for the new model, and we first notice that all of the interactions and the region variable are significant at one level at least, so we include them all. When examining the VIFs, we see that there are several variables with high VIFs, but this is often the case when using interaction terms and despite the multicollinearity issues we still have interpretable coefficients in expected directions. This model explains the data equivalently well when compared to the model with all interactions and better than the model without region and no interactions (see appendix pages 22-25). So, we proceed with this model and validate its assumptions. The added variable plots and model diagnostics show us that the assumptions are valid; we see no poor leverage points and the residual scatter is normal and centered around 0. There is some deviation from the normal line in the qq plot, but overall the assumptions seem well met. The transformations we applied also fit the data well as shown in the output from the marginal model plots.



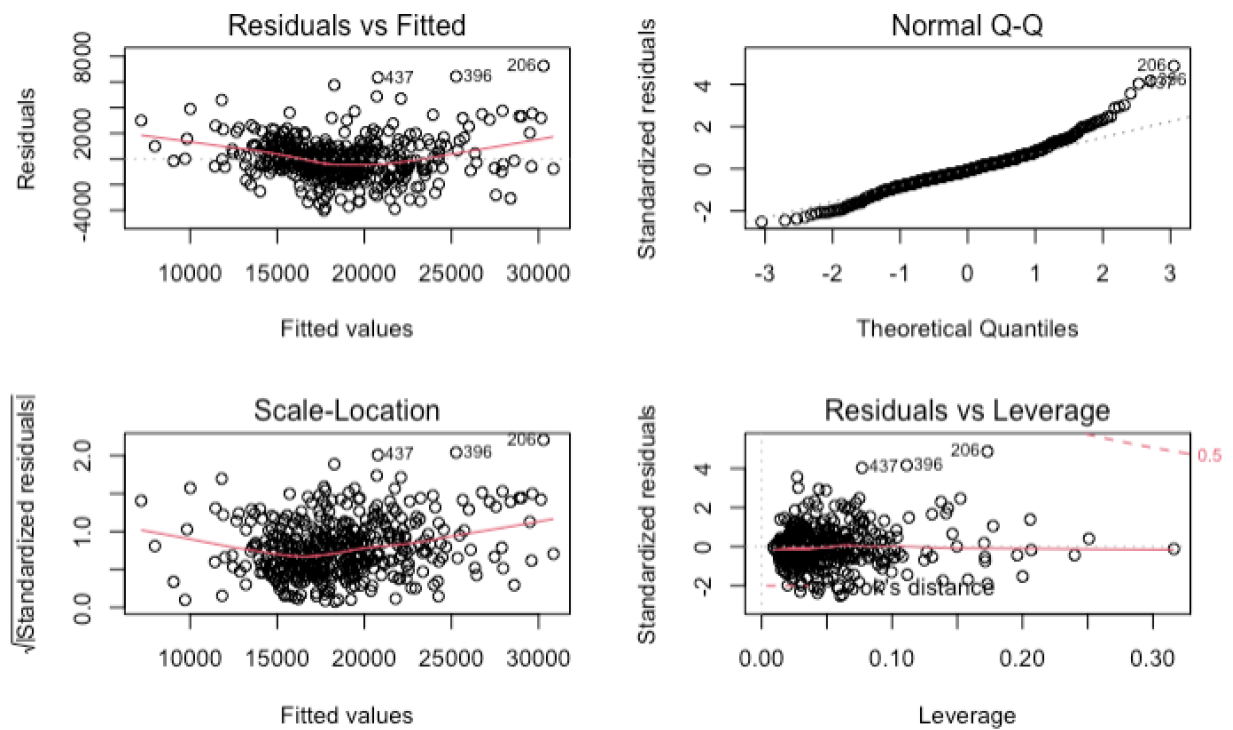


FIGURE 9: Diagnostics

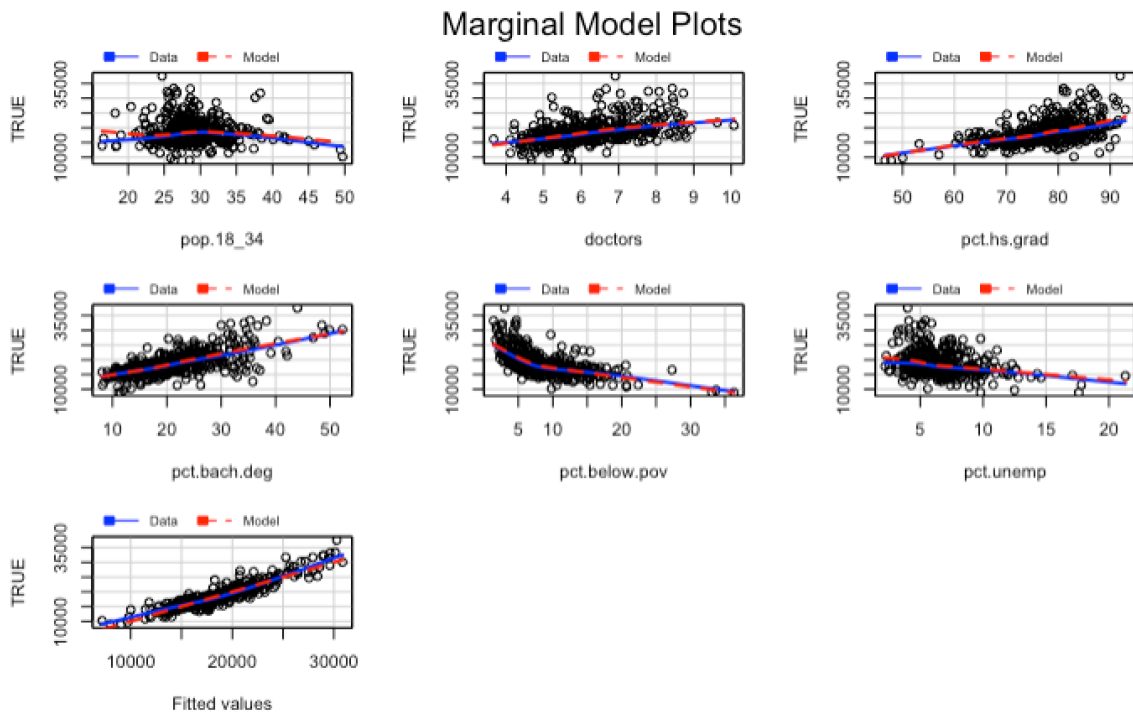


FIGURE 10: Marginal Model Plots

Finally, we interpret some coefficients in context.



- Controlling for the other variables in the model, we expect a 1% increase in percent below poverty to correspond with a decrease in per capita income of 444 units on average for an NC county.
- Controlling for the other variables in the model, we expect a 1% increase in bachelor's degree holders to correspond with a 341 unit increase in income per capita on average for an NC county.
- We expect the increase in income per capita that corresponds with a 1% increase in bachelor's degree holders to be 206 units larger for a county in NE as opposed to one in NC when controlling for the other predictors.

One peculiar thing with this model is that there is a positive relationship between unemployment and income, which could be due to multicollinearity or the fact that the relationship is different across different regions (as some of the interaction terms are negative and have high magnitude).

## [DISCUSSION]

We addressed 4 separate questions throughout this analysis. Through our EDA and initial experiments and plotting of the dataset, we were able to identify key insights about various relationships within the data. To summarize: we saw a negative relationship between % unemployed and % bachelors, which makes sense as higher education can increase one's market value or job prospects. Similarly, we saw a strong positive relationship between hospital beds and crimes, which could be because high-crime areas need to support more injuries and patients. However, there did not seem to be any relationship between crime and unemployment or poverty, indicating that high-crime areas are not necessarily impoverished and crime is not necessarily always because of money. Next, through ANCOVA, we analyzed the relationship between crime and income and how this relationship changes over region. We determined that region does not affect this relationship, and that there is a significant positive relationship between crime and income when marginalizing (i.e. not controlling for) all the other predictors. This relationship is not apparent when addressing crimes per capita, however, which could be important for a stakeholder to consider.

A large part of this work focused on the modeling problem, where we successfully developed a multiple linear regression model to determine what the relationships are between per capita income and various other socioeconomic factors. We found that, in line with our intuition, increased poverty in the area corresponds with decreased income per capita, and that an increase in the proportion of the population with higher education corresponds with an increase in per capita income. We also noticed that the relationship between bachelor degree and income is steeper in the NE region as opposed to NC, which could mean that a college education goes further in the northeast than it does elsewhere.

Although we were able to extract meaningful information from our approach, there were some important limitations. The interaction terms introduced multicollinearity that caused the estimate for unemployment to have a sign that did not agree with our intuition; a possible remediation would be a more extensive model selection procedure or consideration for dropping a few variables that might not impact performance. There was also some deviation from the normal line in the qq plot; we could have employed a transformation of the response (e.g. with Box Cox) in order to remediate this. For this analysis, response transformations were omitted to keep the coefficients as straightforward to interpret as possible.

Overall, this work gives important insights into the socioeconomic factors behind income in various regions. Using this model, stakeholders would be able to identify what could lead to low income in certain counties for targeted remediation through policy-based efforts. There is also potential for future

work. In particular, we could examine more variables and consider more transformations (e.g. per capita crimes) or try a more flexible model. We could also try more in-depth transformations to try to validate the assumptions better. Further research could include analysis of how these relationships have changed over time or assessing a causal relationship between socioeconomic status and per capita income.

Our final question was to address missing counties and states and determine whether or not that was an issue. This was discussed briefly in the DATA section and we summarize our findings here. We noticed that there were only 48/50 states present, but this is reasonable since the data only contained the 440 most populous counties. Next, we noticed that there were less counties than rows, but this is because some states have the same county names (so, we have 440 unique state-county name pairs). Thus, we are not concerned by these inconsistencies.

Something that is concerning is the inherent bias in the data we used to fit the model. Since we only looked at the most populous counties in the U.S. it could be the case that the model is not generalizable to low-population areas. One future extension of this work could be to analyze the socioeconomic differences between high population and low population counties and examine if the model will generalize well. For example, if the relationship between income and things like unemployment, crime, etc. is markedly different in less populated states then our model would not be able to explain this. When examining the states, we find that Alaska and Wyoming, 2 very low-density states are missing. Additionally, our model contains a population variable, so it is entirely possible that it cannot generalize with respect to this variable.

[BIBLIOGRAPHY]

1. Kutner, Michael H. 2005. *Applied Linear Statistical Models / Michael H. Kutner ... Et Al.*. Boston: McGraw-Hill Irwin.

# 36-617 Project 01 Technical Appendix

Anirban Chowdhury

10/17/2021

[Q1]

[1.A]

```
library(knitr)
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5      v purrr  0.3.4
## v tibble  3.1.3      v dplyr  1.0.7
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   2.0.1      v forcats 0.5.1
```

```
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
```

```
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
```

```
##      recode
```

```
## The following object is masked from 'package:purrr':
```

```
##
```

```
##      some
```

```
library(MASS)
```

```
##
```

```
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##      select
```

```
library(leaps)
cdi = read.table('cdi.dat')

(kable(sort(table(cdi$state)/nrow(cdi))))
```

Var1	Freq
DC	0.0022727
ID	0.0022727
MT	0.0022727
ND	0.0022727
SD	0.0022727
VT	0.0022727
WV	0.0022727
AR	0.0045455
DE	0.0045455
NM	0.0045455
NV	0.0045455
HI	0.0068182
KY	0.0068182
MS	0.0068182
NE	0.0068182
RI	0.0068182
KS	0.0090909
NH	0.0090909
OK	0.0090909
UT	0.0090909
AZ	0.0113636
ME	0.0113636
OR	0.0136364
AL	0.0159091
MN	0.0159091
CT	0.0181818
MO	0.0181818
TN	0.0181818
CO	0.0204545
GA	0.0204545
LA	0.0204545
VA	0.0204545
MD	0.0227273
WA	0.0227273
MA	0.0250000
SC	0.0250000
WI	0.0250000
IN	0.0318182
IL	0.0386364
MI	0.0409091
NC	0.0409091
NJ	0.0409091

Var1	Freq
NY	0.0500000
OH	0.0545455
TX	0.0636364
FL	0.0659091
PA	0.0659091
CA	0.0772727

```
kable(table(cdi$region)/nrow(cdi))
```

Var1	Freq
NC	0.2454545
NE	0.2340909
S	0.3454545
W	0.1750000

```
#kable(table(cdi$county)/nrow(cdi))
```

```
tab = summary(cdi$land.area)
name = c('land.area')
for (i in 5:16){
  name = c(name, names(cdi)[i])
  tab = rbind(tab, summary(cdi[,i]))
}
rownames(tab) = name
kable(tab)
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
land.area	15.0	451.250	656.50	1.041411e+03	946.750	20062.0
pop	100043.0	139027.250	217280.50	3.930109e+05	436064.500	8863164.0
pop.18_34	16.4	26.200	28.10	2.856841e+01	30.025	49.7
pop.65_plus	3.0	9.875	11.75	1.216977e+01	13.625	33.8
doctors	39.0	182.750	401.00	9.879977e+02	1036.000	23677.0
hosp.beds	92.0	390.750	755.00	1.458627e+03	1575.750	27700.0
crimes	563.0	6219.500	11820.50	2.711162e+04	26279.500	688936.0
pct.hs.grad	46.6	73.875	77.70	7.756068e+01	82.400	92.9
pct.bach.deg	8.1	15.275	19.70	2.108114e+01	25.325	52.3
pct.below.pov	1.4	5.300	7.90	8.720682e+00	10.900	36.3
pct.unemp	2.2	5.100	6.20	6.596591e+00	7.500	21.3
per.cap.income	8899.0	16118.250	17759.00	1.856148e+04	20270.000	37541.0
tot.income	1141.0	2311.000	3857.00	7.869273e+03	8654.250	184230.0

There are no NA values in this table. We present summary statistics for each variable. For categorical data, we present a table of the frequency of each class. For quantitative data, we present 5 number summaries. For county, most of the frequency counts are 1, so there is no reason to include this table. There are 373 unique values out of 440 rows.

```
sort(table(cdi$county), decreasing = T)[1:20]
```

```
##
##  Jefferson Montgomery Washington Cumberland      Jackson      Lake      Clark
##        7          6          5          4          4          4          3
##  Hamilton      Kent      Madison      Marion Middlesex      Monroe      Orange
##        3          3          3          3          3          3          3
##      Wayne      York      Allen      Bay      Butler      Calhoun
##        3          3          2          2          2          2
```

```
cdi[cdi$county == "Jefferson",1:3]
```

```
##      id      county state
## 66  66 Jefferson  KY
## 68  68 Jefferson  AL
## 107 107 Jefferson  LA
## 110 110 Jefferson  CO
## 202 202 Jefferson  TX
## 271 271 Jefferson  MO
## 399 399 Jefferson  NY
```

```
cdi[cdi$county == "Montgomery",1:3]
```

```
##      id      county state
## 48  48 Montgomery  MD
## 58  58 Montgomery  PA
## 80  80 Montgomery  OH
## 230 230 Montgomery  AL
## 254 254 Montgomery  TX
## 438 438 Montgomery  TN
```

```
cdi[cdi$county == "Washington",1:3]
```

```
##      id      county state
## 156 156 Washington  OR
## 234 234 Washington  PA
## 312 312 Washington  MN
## 389 389 Washington  AR
## 402 402 Washington  RI
```

So, some county names repeat because different states have the same names for certain counties.

```
length(unique(paste(cdi$state, cdi$county, sep = "-")))
```

```
## [1] 440
```

We can see that we have exactly 440 county-state pairings, justifying our argument that the counties that repeat are just collisions across different states. This does not impede our modeling process.

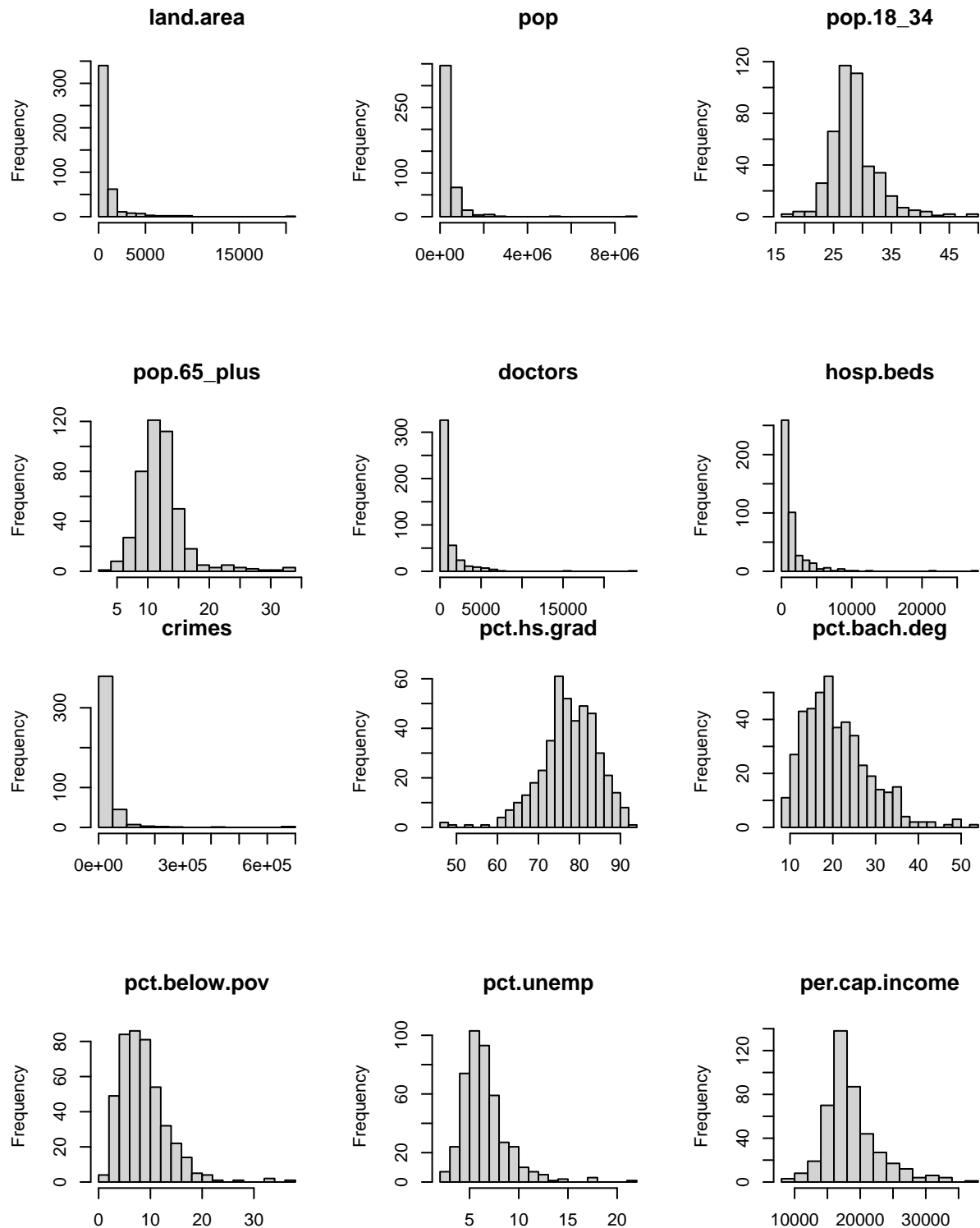


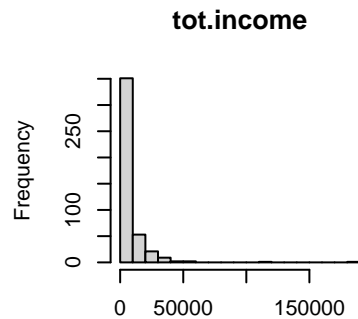
```

par(mfrow = c(2,3))
for (i in 4:16){

  hist(cdi[,i], main = names(cdi)[i], xlab = '', breaks = 20)
}

```





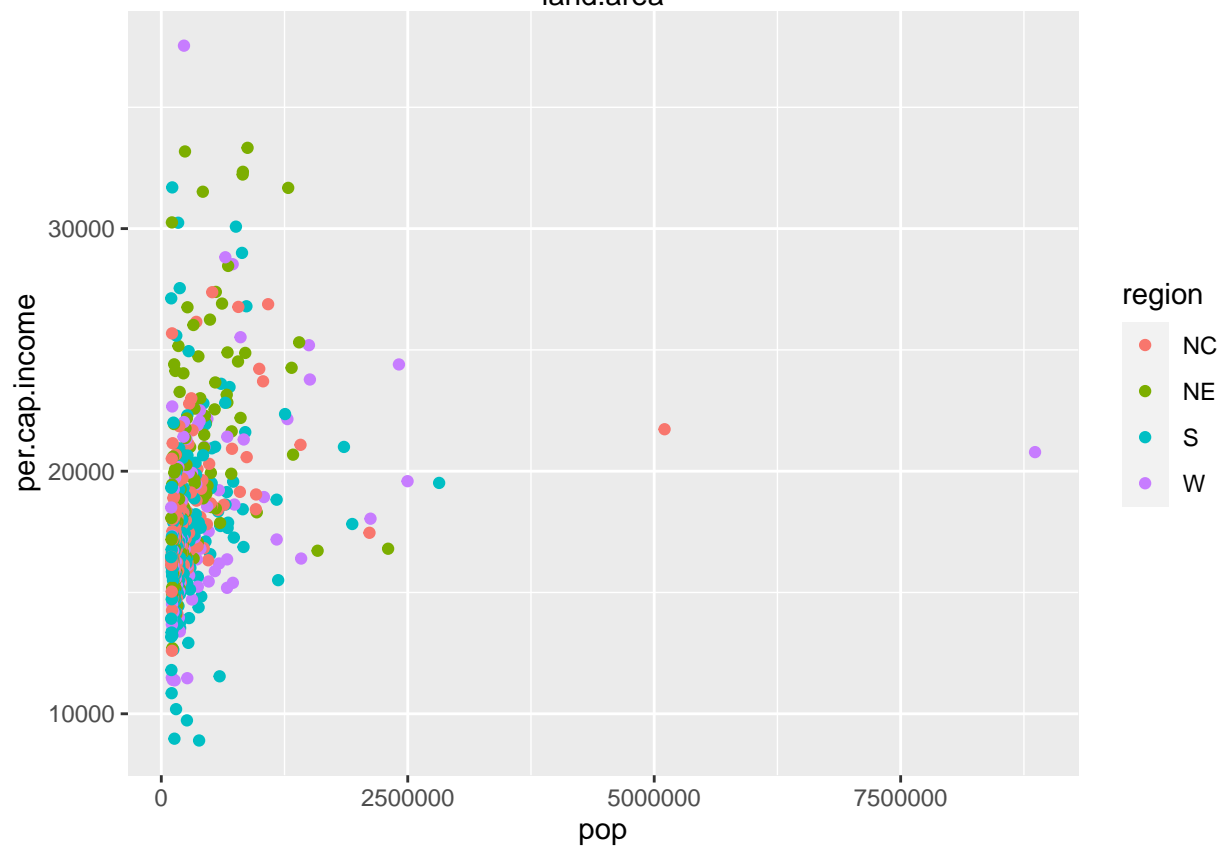
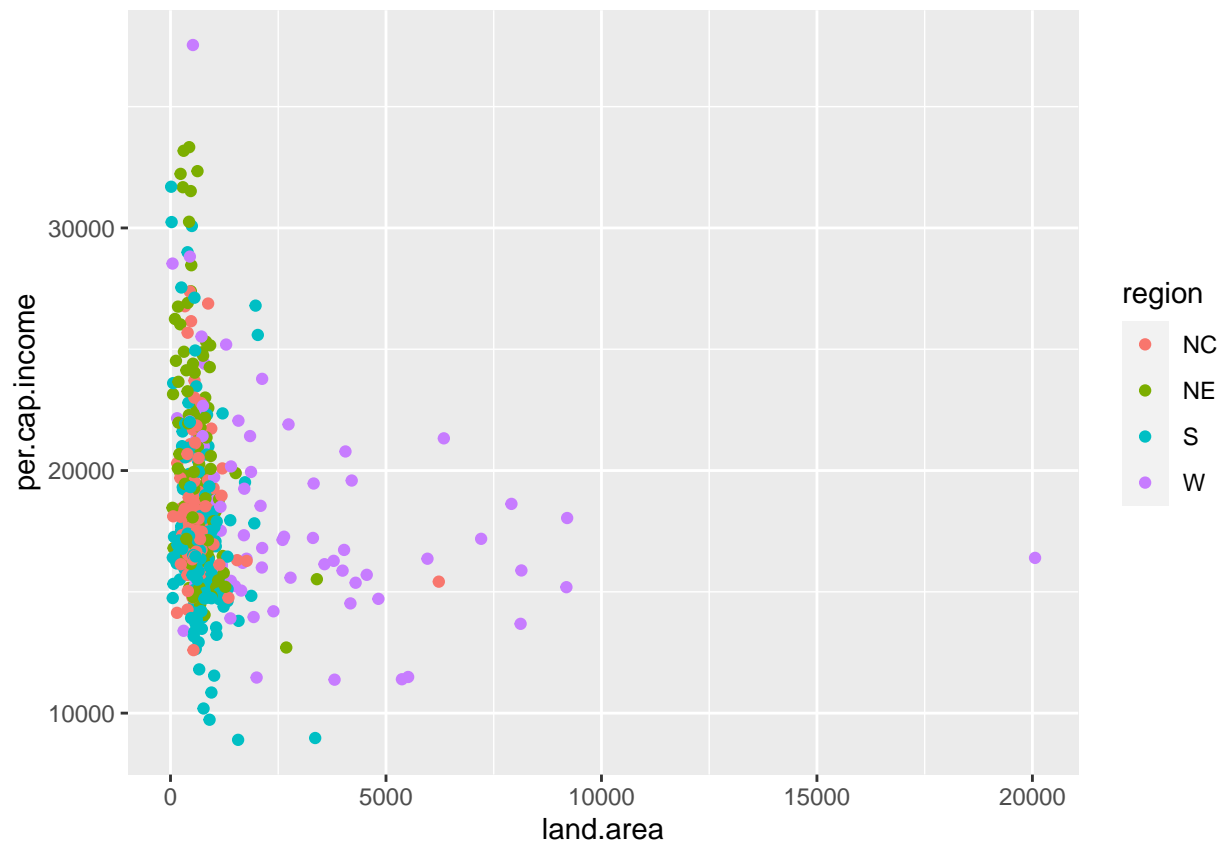
From the histograms, we can see several relevant variables are skewed. Since we are interested in examining crime rate, we look closely at the distributions of crime and population and find that they are both right skewed. Since these will likely be involved in our model (either crimes or crimes per capita) it is important to keep in mind that a transformation might be necessary. the same is true for total income, but not per capita income. A lot of the other aggregate statistics (total doctors, hospital beds) also look right skewed.

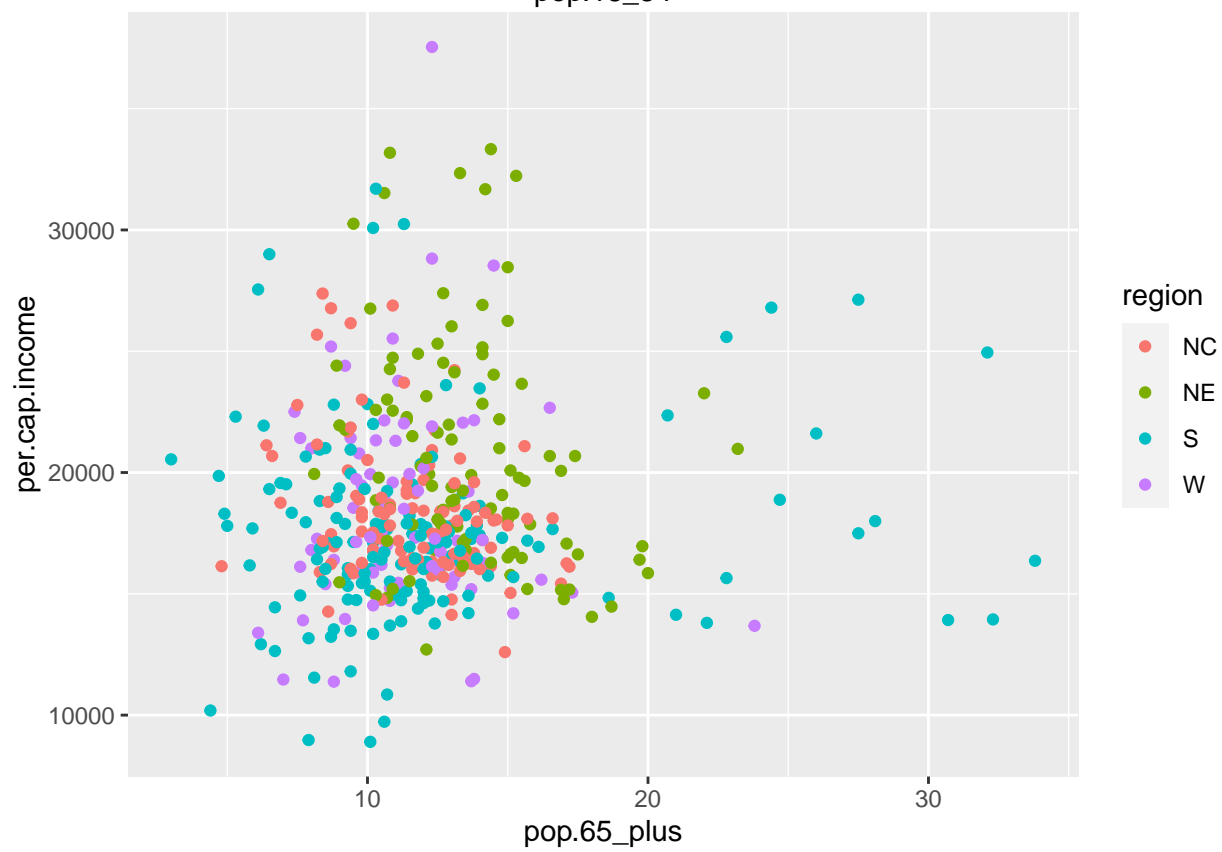
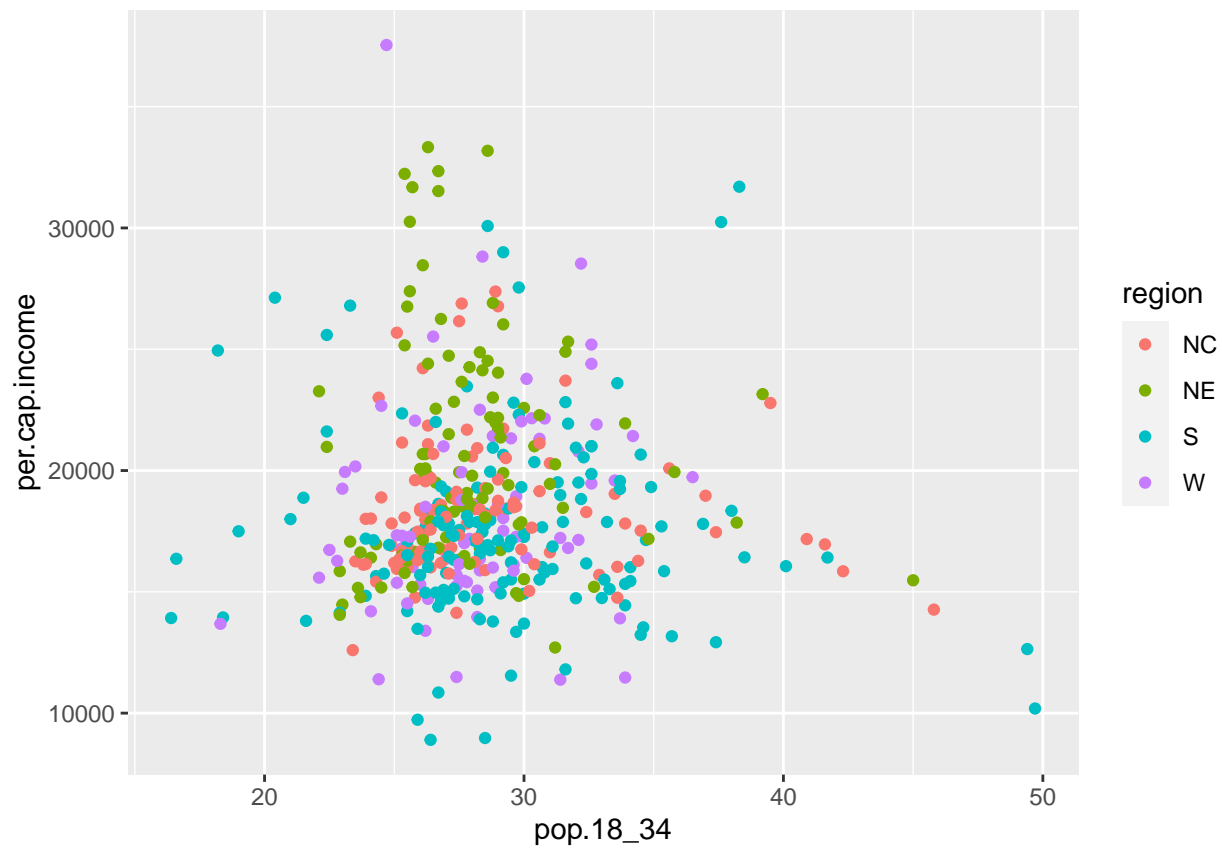
We see some important relationships among the quantitative variables here. We see negative trends with unemployment rate and poverty rate, and a positive trend with high school graduation rate. Interestingly, we do not see any relationship with crimes per capita (omitted for consiseness) or total crimes.

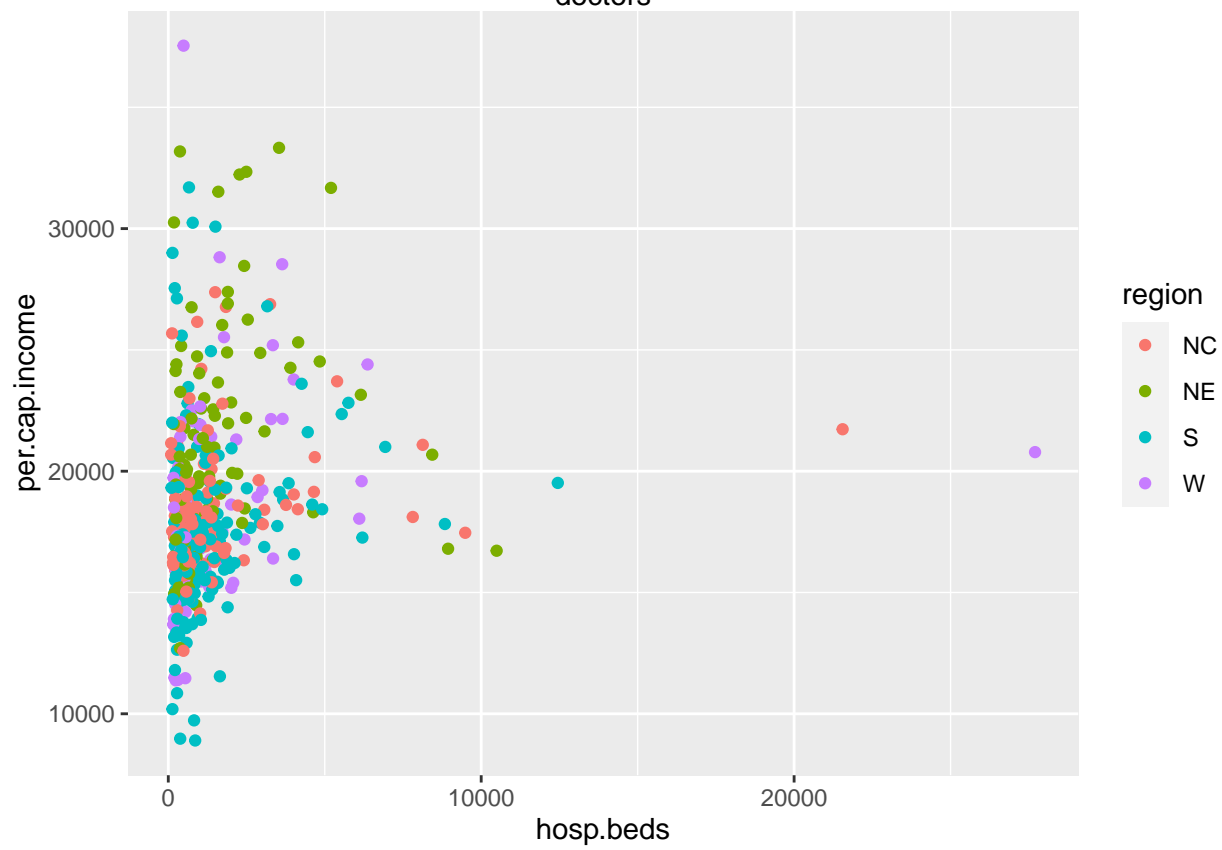
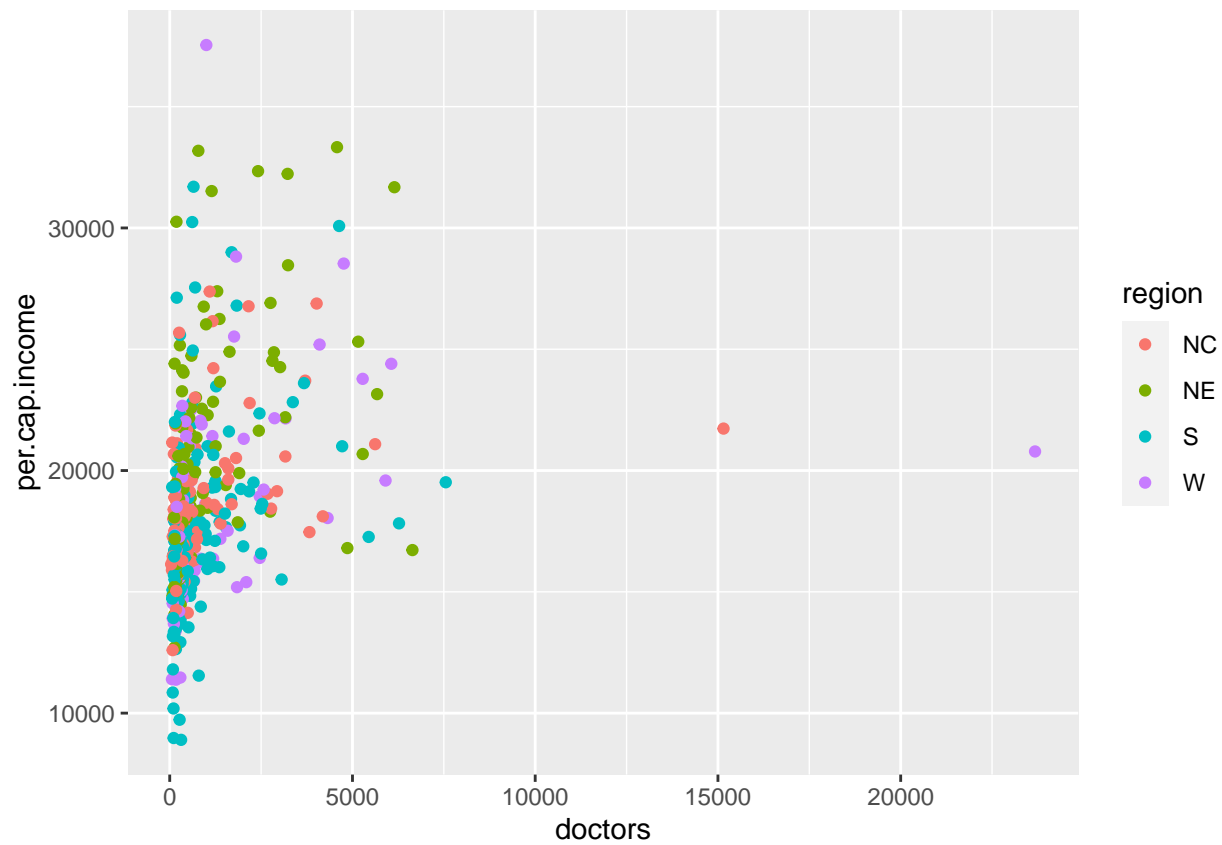
```
par(mfrow = c(3, 4))

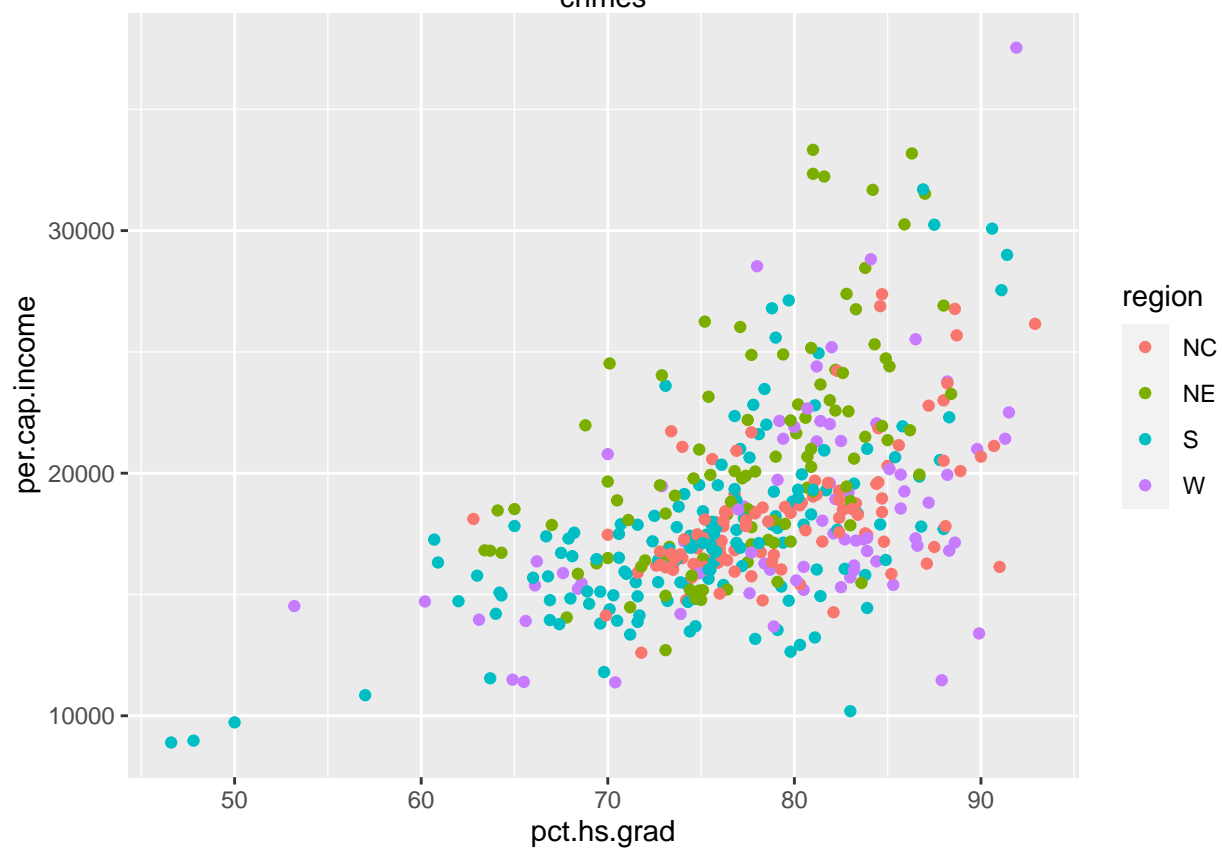
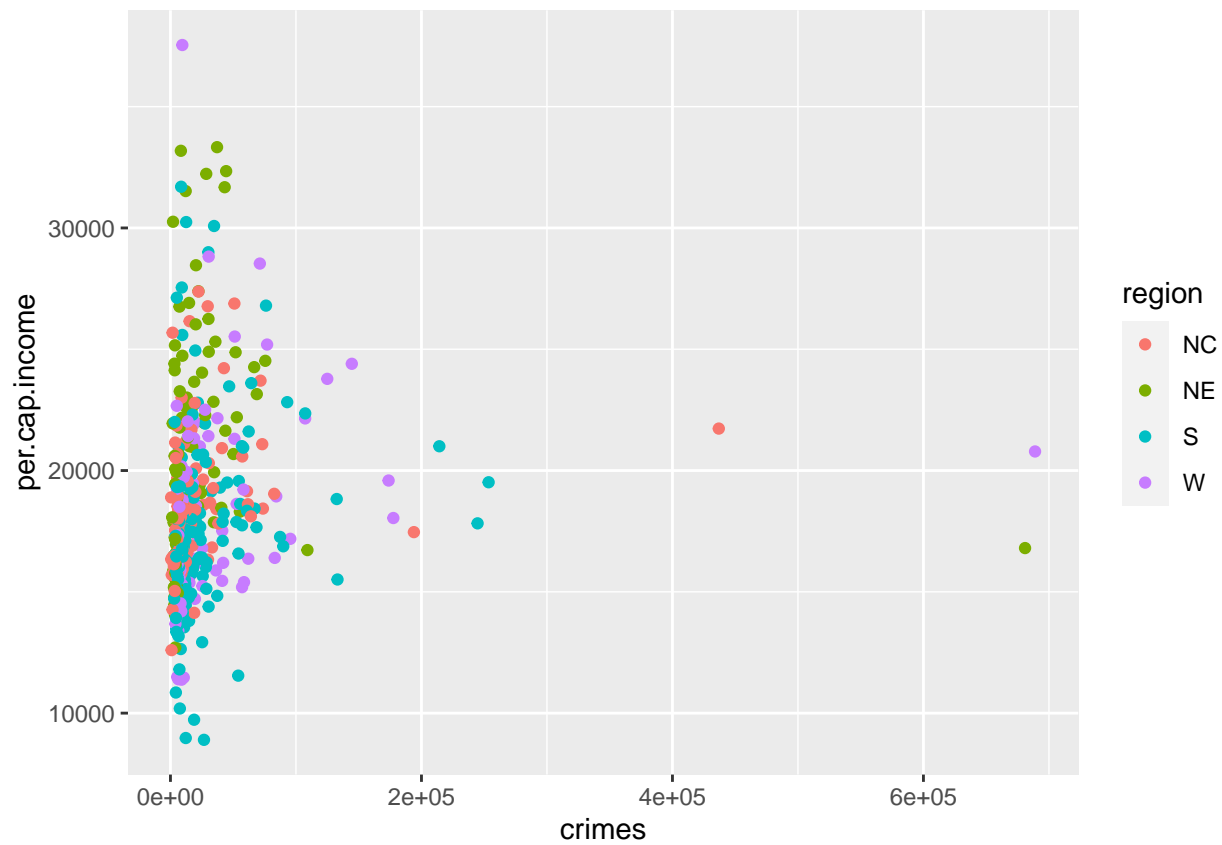
library(ggplot2)

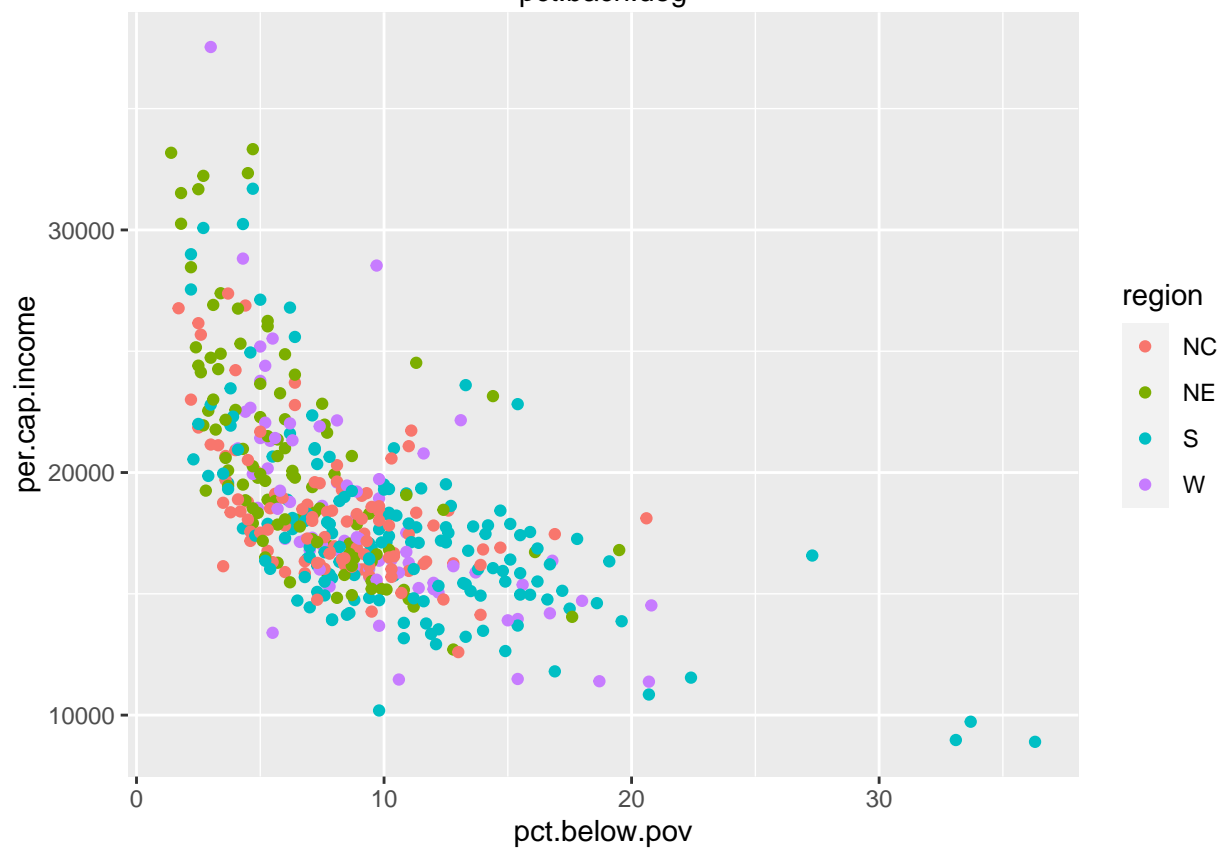
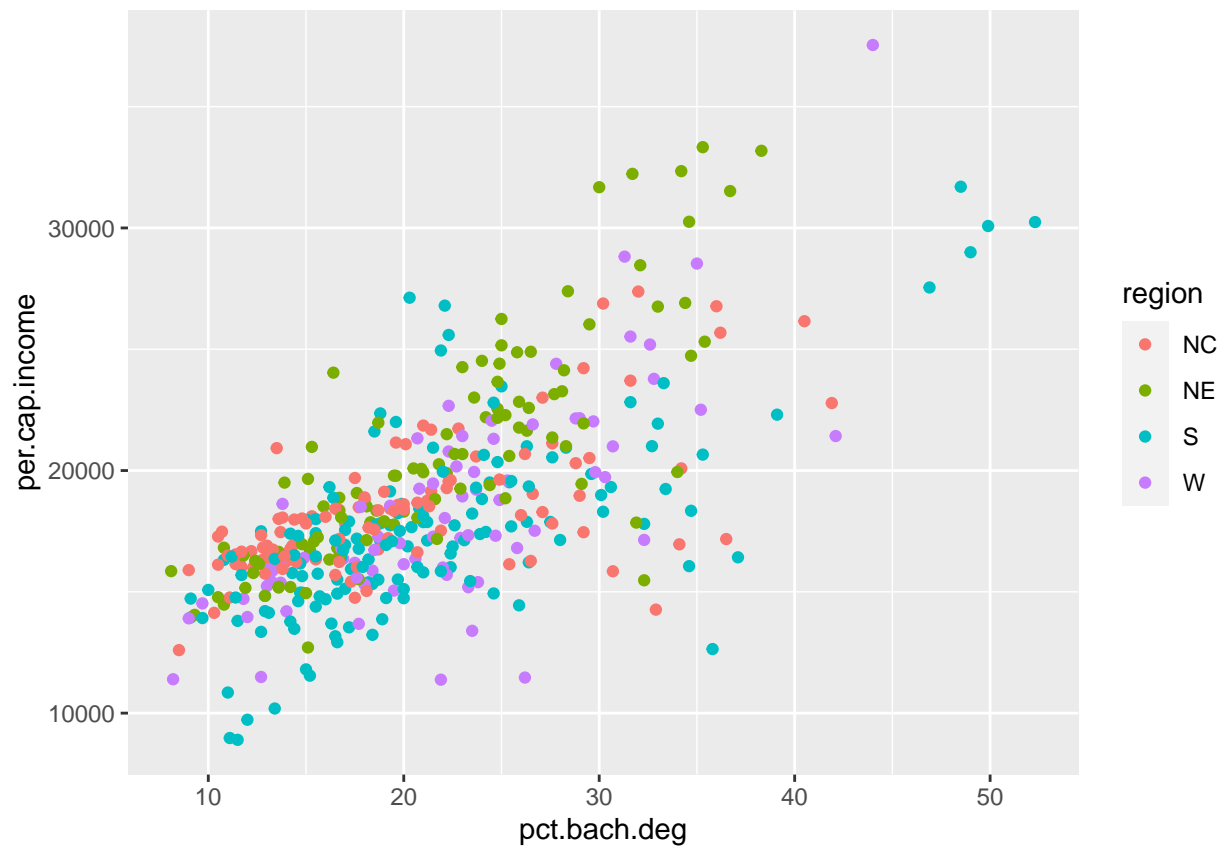
for (i in 4:16){
  if(i != 15){
    name = colnames(cdi)[i]
    print(ggplot(data = cdi) + geom_point(aes(x = cdi[,i],
                                              y = per.cap.income,
                                              color = region)) + xlab(name))
  }
}
```



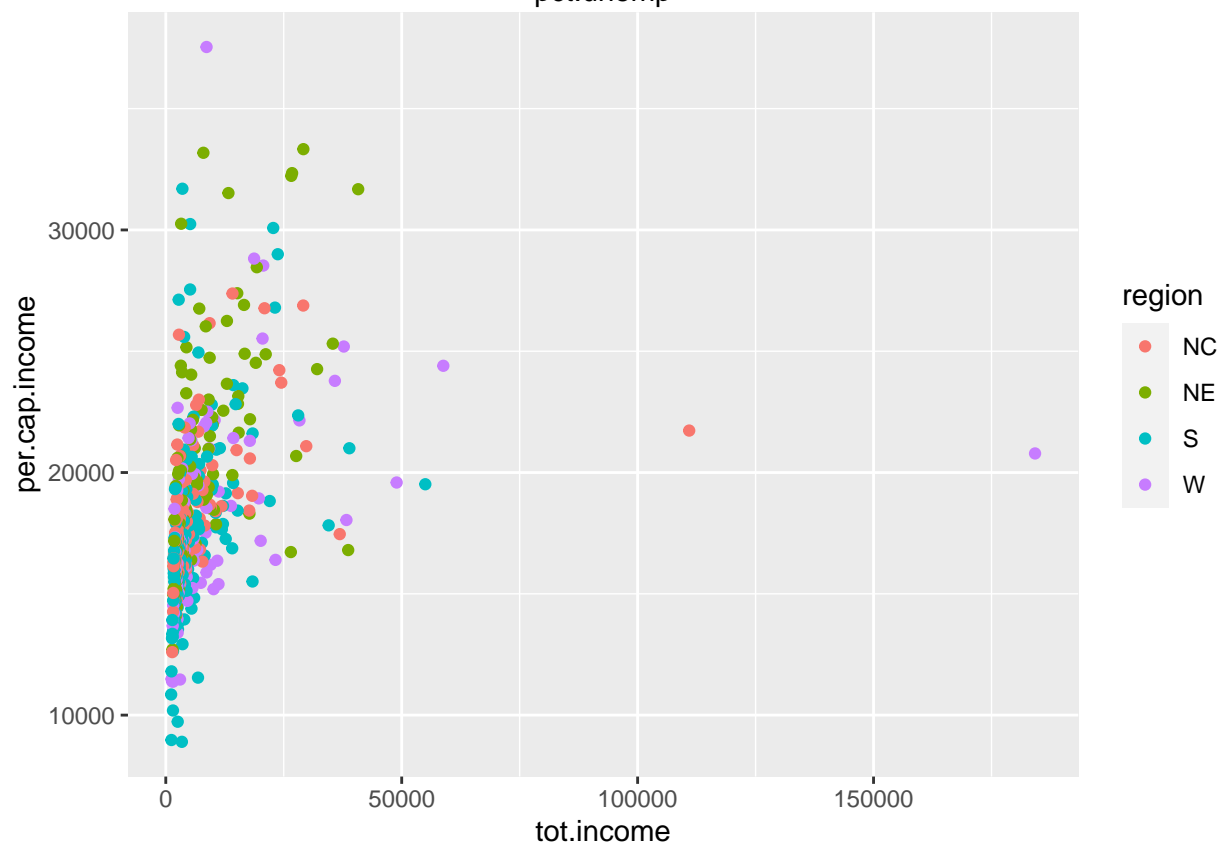
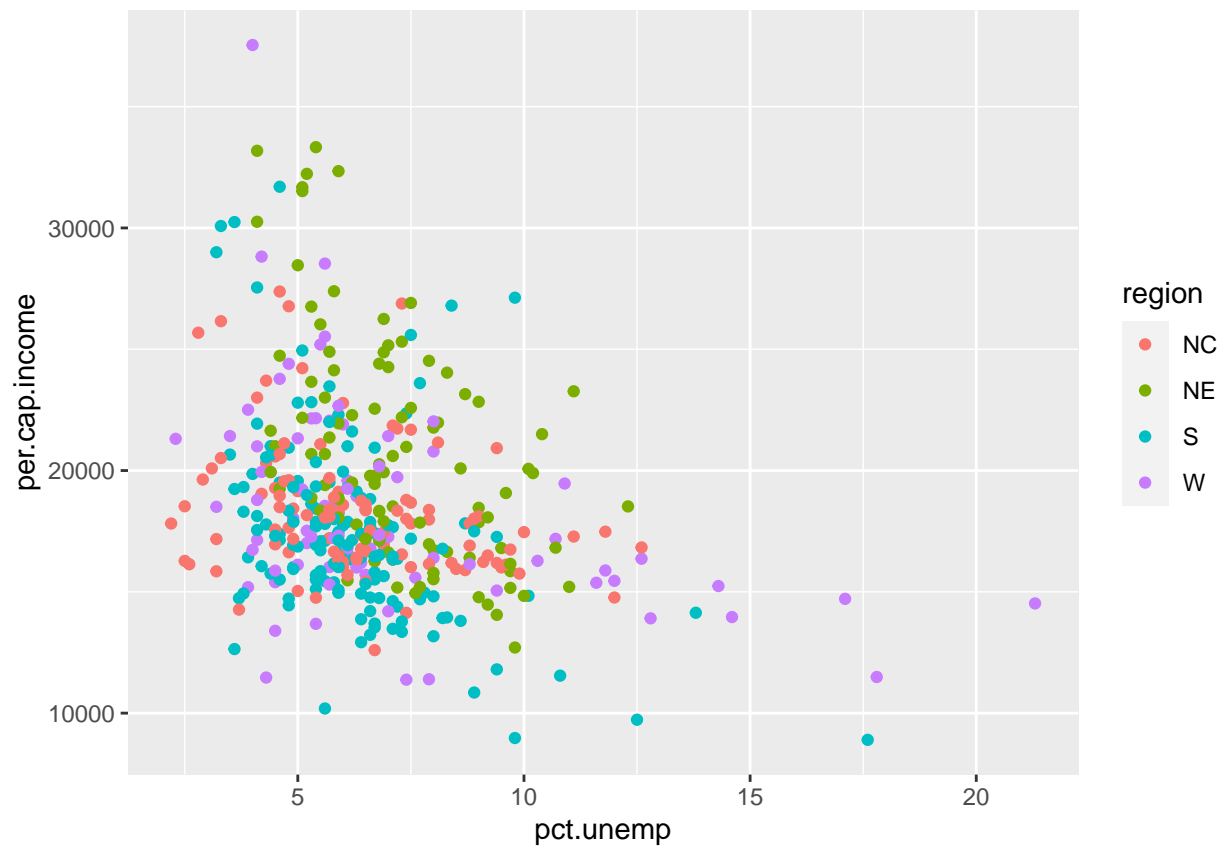




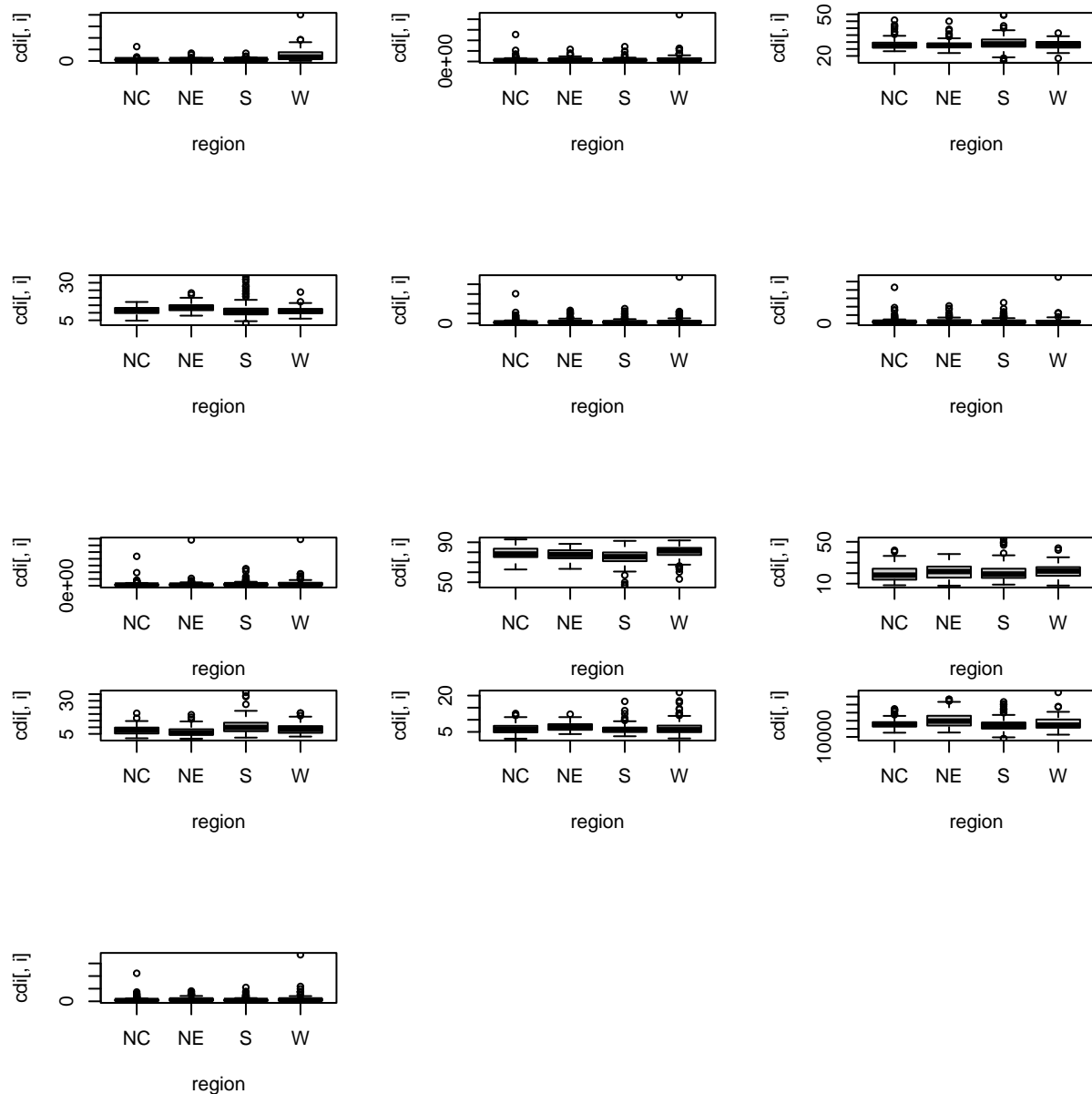








```
par(mfrow = c(3,3))
for (i in 4:(ncol(cdi) - 1)){
  boxplot(cdi[,i] ~ region, yllab = names(cdi)[i], data = cdi)}
```



We see a different distribution of income in different regions but the medians are close and quantiles overlap, so it's unlikely this variable contains relevant information to our modeling problem.

### [1.B]

Looking at the scatterplot above, we do not see evidence for interaction based on region in the relationship between income and crimes. We just see random scatter in all the groups. For this reason, we do not include an interaction term in the regression of income against crimes. (if we do try to include one, it is not significant and the ANOVA indicates that the model with interaction does not explain the data any better).

```
lm1_nointer = lm(per.cap.income ~ crimes + region , data = cdi)
lm1 = lm(per.cap.income ~ crimes + region + region : crimes, data = cdi)
anova(lm1, lm1_nointer)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ crimes + region + region:crimes
## Model 2: per.cap.income ~ crimes + region
##   Res.Df      RSS Df Sum of Sq    F Pr(>F)
## 1     432 6438799739
## 2     435 6501791845 -3 -62992106 1.4088 0.2396
```

```
summary(lm1_nointer)
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9661.0 -2260.7  -618.3  1650.0 19492.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.811e+04  3.784e+02  47.846  < 2e-16 ***
## crimes       8.915e-03  3.188e-03   2.797  0.00539 **
## regionNE     2.286e+03  5.325e+02   4.293  2.17e-05 ***
## regionS     -8.606e+02  4.868e+02  -1.768  0.07782 .
## regionW     -1.428e+02  5.796e+02  -0.246  0.80548
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3866 on 435 degrees of freedom
## Multiple R-squared:  0.1011, Adjusted R-squared:  0.09288
## F-statistic: 12.24 on 4 and 435 DF, p-value: 1.946e-09
```

Above we fit a linear regression of income against total crimes and region. Our analysis tells us that in fact there is surprisingly a significant positive relationship between crimes and income because the estimated coefficient is  $> 0$  and the p value is  $< 0.05$ , so for a unit increase in total crime we expect to see an increase in  $8.9e-03$  units of income per capita. However, since we're using income per capita, it could be better to use crime per capita instead of total crime to maintain consistency and keep the variables on a similar scale.

```
lm2_nointer = lm(per.cap.income ~ I(crimes/pop) + region , data = cdi)
lm2 = lm(per.cap.income ~ I(crimes/pop) + region
          + region : I(crimes/pop), data = cdi)
anova(lm2_nointer, lm2)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ I(crimes/pop) + region
## Model 2: per.cap.income ~ I(crimes/pop) + region + region:I(crimes/pop)
```

```
##      Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1      435 6609753963
## 2      432 6607856753   3   1897210 0.0413 0.9888
```

```
summary(lm2_nointer)
```

```
##
## Call:
## lm(formula = per.cap.income ~ I(crimes/pop) + region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8634  -2300   -631   1710  19332
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18006.04     537.04  33.528 < 2e-16 ***
## I(crimes/pop)  5773.20    7520.41   0.768  0.4431
## regionNE      2354.70     541.97   4.345 1.74e-05 ***
## regionS       -927.45     512.31  -1.810  0.0709 .
## regionW       -34.92     586.03  -0.060  0.9525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3898 on 435 degrees of freedom
## Multiple R-squared:  0.08622,    Adjusted R-squared:  0.07782
## F-statistic: 10.26 on 4 and 435 DF,  p-value: 6.007e-08
```

When computing against crime per capita instead, the analysis looks a little bit different. In this model, there is no evidence for a linear relationship between crime per capita and income because the p value is > 0.05. Again, we use the model without interaction because including these terms does not help us explain the data any better.

Because we get no explicit benefit from using crimes per capita and this could lead to complications in multicollinearity when including population or population by age as a dependent variable in the model later on, we proceed with just crimes as a predictor.

## [1.C]

Clearly, the models we fit above are inadequate. So, we must expand our search space and consider more predictors and transformations of these predictors as needed. Because crimes, hospital beds, doctors, land area, population, and total income are all skewed right, we can first take a log transformation to improve the distributions to better satisfy model assumptions.

When considering how region affects the relationship of the quantitative predictors and income, we can look at the above plots and find that the different regions seem to follow random scatter and do not separate into groups for any of these income/predictor pairings. However, we do include interactions to see if after relevant transformations and variable inclusions these relationships are relevant. We also omit state and county from this analysis for interpretability: considering factors with so many levels increases the difficulty of both the modeling problem and the interpretation problem, and we're already capturing geographic information in the region variable. We tried fitting a model with state, but saw no benefit and ended up with the same final model after variable selection.

First, since income per capital is directly related to population and total income, we first remove these from the data as to only include relevant predictors. We also remove county, state, and id for simplicity. Then, we should transform the variables that are clearly skewed right: doctors, hospital beds, land area, and crimes.

```

cdi$county = NULL
cdi$state = NULL
cdi$id = NULL
cdi$tot.income = NULL
cdi$pop = NULL
cdi$doctors = log(cdi$doctors )
cdi$hosp.beds = log(cdi$hosp.beds )
cdi$crimes = log(cdi$crimes )
cdi$land.area = log(cdi$land.area )

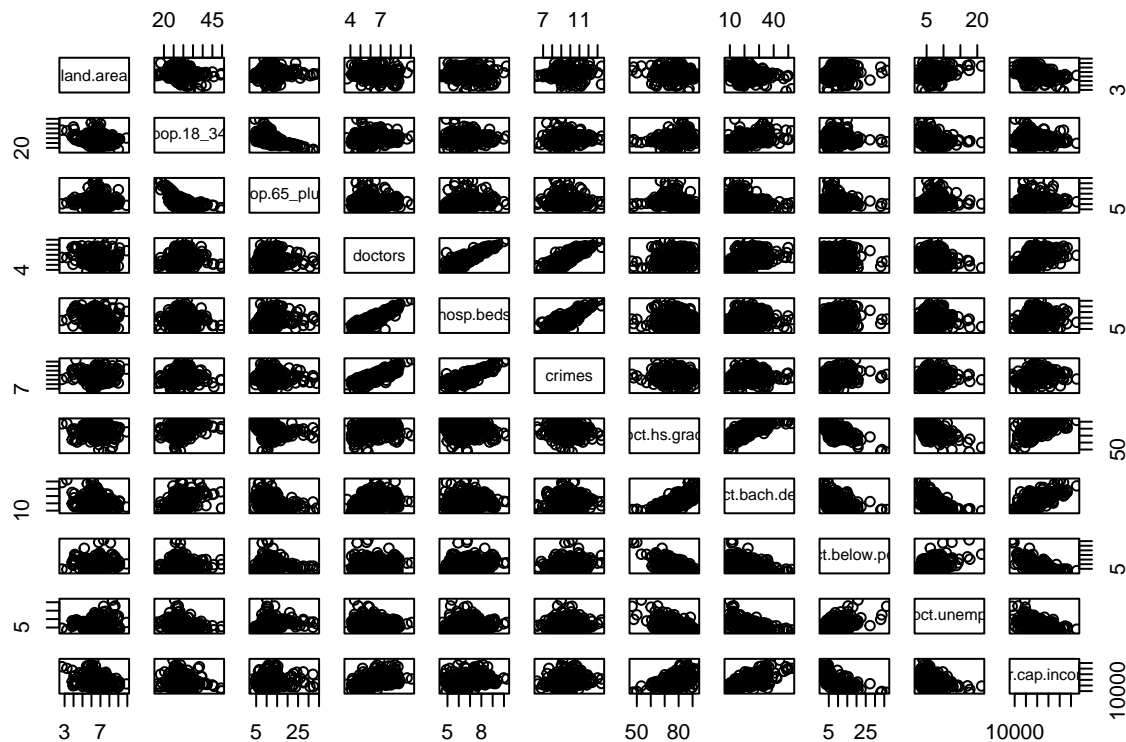
lm_full = lm(per.cap.income ~ . , data = cdi)
lm_full_inter = lm(per.cap.income ~ .*region , data = cdi)
summary(lm_full_inter)

##
## Call:
## lm(formula = per.cap.income ~ . * region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4181.8  -830.2   -81.0    692.4   6272.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    27527.1301   6384.3709   4.312 2.05e-05 ***
## land.area      -611.6857    327.1091  -1.870 0.062225 .
## pop.18_34      -316.3239     59.8129  -5.289 2.04e-07 ***
## pop.65_plus    -13.9216    106.8299  -0.130 0.896383
## doctors         925.6080    497.4879   1.861 0.063547 .
## hosp.beds     -151.8263    534.6754  -0.284 0.776590
## crimes         157.4794    307.8606   0.512 0.609266
## pct.hs.grad    -89.3150     72.3469  -1.235 0.217734
## pct.bach.deg    319.9115     64.6938   4.945 1.13e-06 ***
## pct.below.pov  -437.9730     81.6343  -5.365 1.38e-07 ***
## pct.unemp       332.6328    112.3044   2.962 0.003242 **
## regionNE       7646.2210   8456.2360   0.904 0.366433
## regionS      -3655.1215   7090.0721  -0.516 0.606473
## regionW      39542.1668  10137.8967   3.900 0.000113 ***
## land.area:regionNE    19.4306    431.3388   0.045 0.964092
## land.area:regionS   -165.4512    376.3480  -0.440 0.660450
## land.area:regionW    331.4142    391.5202   0.846 0.397796
## pop.18_34:regionNE  -195.9275     88.0449  -2.225 0.026623 *
## pop.18_34:regionS     58.5780     72.6074   0.807 0.420279
## pop.18_34:regionW    -0.4859     99.8203  -0.005 0.996119
## pop.65_plus:regionNE -148.1345    136.9211  -1.082 0.279957
## pop.65_plus:regionS    52.8801    112.7167   0.469 0.639226
## pop.65_plus:regionW   -93.9925    144.1454  -0.652 0.514736
## doctors:regionNE    -794.6074    806.4630  -0.985 0.325079
## doctors:regionS     174.8199    643.1920   0.272 0.785917

```

```
## doctors:regionW      2132.3982  1039.3420   2.052 0.040858 *
## hosp.beds:regionNE    679.9715   864.6849   0.786 0.432115
## hosp.beds:regionS    -360.4192   657.9997  -0.548 0.584172
## hosp.beds:regionW    -64.6518   835.6308  -0.077 0.938369
## crimes:regionNE      237.7094   492.7801   0.482 0.629799
## crimes:regionS       34.2973   449.5959   0.076 0.939231
## crimes:regionW     -2320.8276   764.3259  -3.036 0.002552 **
## pct.hs.grad:regionNE  -75.2925    92.5338  -0.814 0.416319
## pct.hs.grad:regionS   46.6164    79.7802   0.584 0.559345
## pct.hs.grad:regionW  -367.7820    96.6265  -3.806 0.000163 ***
## pct.bach.deg:regionNE 222.0039    89.2939   2.486 0.013322 *
## pct.bach.deg:regionS  -21.5795    70.7407  -0.305 0.760488
## pct.bach.deg:regionW  118.9315    81.6123   1.457 0.145833
## pct.below.pov:regionNE -8.6187   111.4515  -0.077 0.938399
## pct.below.pov:regionS 173.1631    91.6347   1.890 0.059527 .
## pct.below.pov:regionW -231.5163   119.0365  -1.945 0.052492 .
## pct.unemp:regionNE    -129.2935   160.3883  -0.806 0.420653
## pct.unemp:regionS     -293.2009   149.4599  -1.962 0.050493 .
## pct.unemp:regionW    -375.3741   148.6425  -2.525 0.011948 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1554 on 396 degrees of freedom
## Multiple R-squared:  0.8678, Adjusted R-squared:  0.8535
## F-statistic: 60.48 on 43 and 396 DF,  p-value: < 2.2e-16
```

```
pairs(cdi[, -ncol(cdi)])
```



Right away we see a much better fit. The R squared is much higher and many variables (crimes not included) are significant. Most of the interaction terms do not seem useful. From this set of variables, we can examine the VIFs to determine any multicollinearity.

```
library(rms)
```

```
## Loading required package: Hmisc

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##      src, summarize

## The following objects are masked from 'package:base':
##
##      format.pval, units

## Loading required package: SparseM

##
## Attaching package: 'SparseM'

## The following object is masked from 'package:base':
##
##      backsolve

##
## Attaching package: 'rms'

## The following objects are masked from 'package:car':
##
##      Predict, vif
```

```
rms::vif(lm_full_inter)
```

##	land.area	pop.18_34	pop.65_plus
##	14.78608	11.42796	33.08557
##	doctors	hosp.beds	crimes
##	58.90931	52.33403	20.18326
##	pct.hs.grad	pct.bach.deg	pct.below.pov
##	46.84266	44.59529	26.28063
##	pct.unemp	regionNE	regionS
##	12.53665	2336.85258	2071.79787
##	regionW	land.area:regionNE	land.area:regionS
##	2704.60022	243.42055	242.17880



```
##      land.area:regionW      pop.18_34:regionNE      pop.18_34:regionS
##      228.99105              201.97400              193.46201
##      pop.18_34:regionW      pop.65_plus:regionNE      pop.65_plus:regionS
##      211.66898              120.17780              99.87398
##      pop.65_plus:regionW      doctors:regionNE      doctors:regionS
##      74.68390              888.95914              659.82570
##      doctors:regionW      hosp.beds:regionNE      hosp.beds:regionS
##      1200.19163            1174.25715            825.77659
##      hosp.beds:regionW      crimes:regionNE      crimes:regionS
##      837.84799            690.72890            791.22928
##      crimes:regionW      pct.hs.grad:regionNE      pct.hs.grad:regionS
##      1520.77118            1695.92697            1506.10587
##      pct.hs.grad:regionW      pct.bach.deg:regionNE      pct.bach.deg:regionS
##      1578.47478            141.22802            112.38018
##      pct.bach.deg:regionW      pct.below.pov:regionNE      pct.below.pov:regionS
##      96.20519              23.25079              55.42150
##      pct.below.pov:regionW      pct.unemp:regionNE      pct.unemp:regionS
##      38.53981              46.53163              40.49845
##      pct.unemp:regionW
##      37.18172
```

There is indeed some multicollinearity issue within doctors and hospital beds, and crimes is also high. Most of the interaction terms have very high VIFs, as expected. Because of this, it is necessary to attempt some modeling selection. We try stepwise with a backwards direction using BIC as our criterion, as this is a heuristic for all subsets and should give us a reasonable model. When selecting variables, we choose to first remove region and interactions because selection algorithms are not robust to categorical data. We experiment with stepwise and all subsets search.

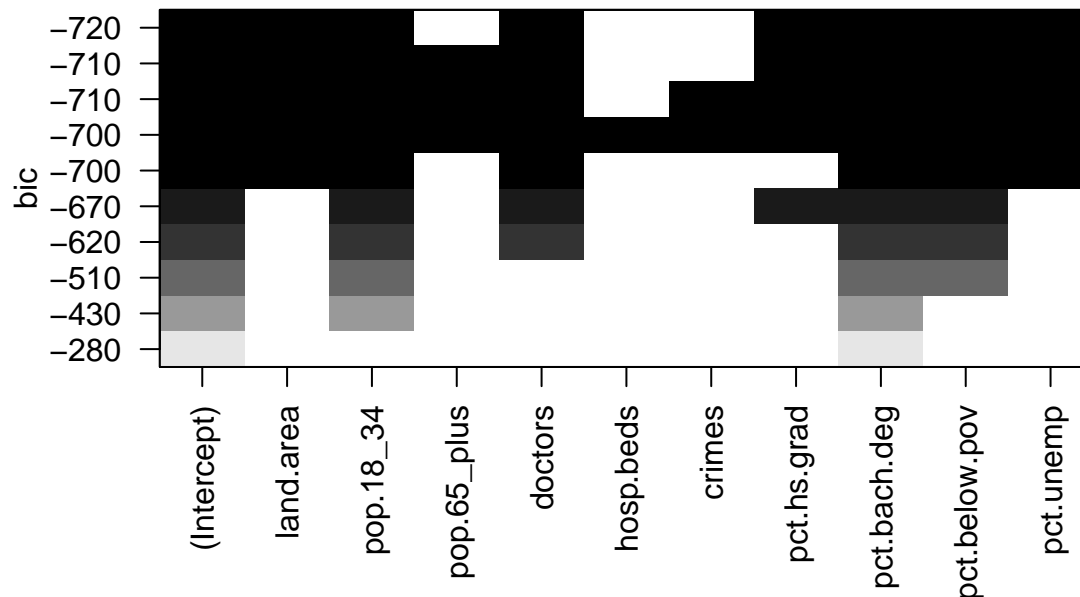
```
lm_full_noregion = lm(per.cap.income ~ . -region, data = cdi)

lm_step_noregion = stepAIC(lm_full_noregion, k = log(nrow(cdi)), trace = F)
summary(lm_step_noregion)
```

```
##
## Call:
## lm(formula = per.cap.income ~ land.area + pop.18_34 + doctors +
##      pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5688.4 -1015.1  -123.4   892.2  8260.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28748.60   1944.84  14.782 < 2e-16 ***
## land.area    -683.89     99.76   -6.855 2.47e-11 ***
## pop.18_34    -300.39     23.21  -12.942 < 2e-16 ***
## doctors      1000.90     83.92   11.926 < 2e-16 ***
## pct.hs.grad  -116.80     22.60   -5.168 3.63e-07 ***
## pct.bach.deg   371.01     19.31   19.214 < 2e-16 ***
## pct.below.pov -427.27     26.28  -16.258 < 2e-16 ***
## pct.unemp     251.44     45.47    5.530 5.56e-08 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1713 on 432 degrees of freedom
## Multiple R-squared:  0.8248, Adjusted R-squared:  0.822
## F-statistic: 290.6 on 7 and 432 DF,  p-value: < 2.2e-16
```

```
lm_sub_noregion = regsubsets(per.cap.income ~ .-region, data = cdi, nvmax = 15)
plot(lm_sub_noregion)
```



```
coef(lm_sub_noregion, 1:8)[[7]]
```

```
## (Intercept) land.area pop.18_34 doctors pct.hs.grad
## 28748.6035 -683.8873 -300.3892 1000.9013 -116.8039
## pct.bach.deg pct.below.pov pct.unemp
## 371.0053 -427.2673 251.4416
```

```
lm_sub_res = lm(per.cap.income ~ pop.18_34 + doctors + pct.hs.grad +
pct.bach.deg + pct.below.pov + pct.unemp, data = cdi)
summary(lm_sub_res)
```

```
##
## Call:
## lm(formula = per.cap.income ~ pop.18_34 + doctors + pct.hs.grad +
## pct.bach.deg + pct.below.pov + pct.unemp, data = cdi)
##
## Residuals:
## Min 1Q Median 3Q Max
## -4923.9 -1070.2 -131.4 944.1 8211.5
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26910.05 2025.98 13.282 < 2e-16 ***
```

```
## pop.18_34      -288.72      24.35 -11.859 < 2e-16 ***
## doctors        1002.53      88.27  11.358 < 2e-16 ***
## pct.hs.grad    -152.96      23.12  -6.617 1.09e-10 ***
## pct.bach.deg    392.00      20.05  19.549 < 2e-16 ***
## pct.below.pov  -459.48      27.20 -16.896 < 2e-16 ***
## pct.unemp       203.03      47.24   4.298 2.13e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1801 on 433 degrees of freedom
## Multiple R-squared:  0.8058, Adjusted R-squared:  0.8031
## F-statistic: 299.4 on 6 and 433 DF,  p-value: < 2.2e-16
```

We actually find that stepwise and all subsets arrive at the same model. So, we now take this subset of quantitative variables and add back region and interactions to see if there will be an improvement in the fit.

```
library(MASS)
library(leaps)

lm_sub_region_res = lm(per.cap.income ~ (pop.18_34 +
                                         doctors + pct.hs.grad +
                                         pct.bach.deg + pct.below.pov +
                                         pct.unemp) * region, data = cdi)

summary(lm_sub_region_res)
```

```
##
## Call:
## lm(formula = per.cap.income ~ (pop.18_34 + doctors + pct.hs.grad +
##   pct.bach.deg + pct.below.pov + pct.unemp) * region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4136.4  -915.1   -94.0    747.2   7161.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    26365.00    6037.12   4.367 1.59e-05 ***
## pop.18_34       -329.35     55.17  -5.970 5.13e-09 ***
## doctors          909.94    199.06   4.571 6.42e-06 ***
## pct.hs.grad     -118.62     70.93  -1.672 0.095222 .
## pct.bach.deg     342.72     61.69   5.555 4.98e-08 ***
## pct.below.pov   -443.21     77.52  -5.718 2.08e-08 ***
## pct.unemp        327.90    103.56   3.166 0.001659 **
## regionNE        8345.82   7647.63   1.091 0.275781
## regionS       -6483.53   6598.24  -0.983 0.326374
## regionW       27386.55   8918.87   3.071 0.002278 **
## pop.18_34:regionNE  -126.80     78.30  -1.619 0.106111
## pop.18_34:regionS    86.68     64.82   1.337 0.181907
## pop.18_34:regionW    12.98     87.27   0.149 0.881820
## doctors:regionNE    -45.68    283.48  -0.161 0.872067
## doctors:regionS    -61.54    244.55  -0.252 0.801426
```

```
## doctors:regionW      -95.17      280.63   -0.339  0.734689
## pct.hs.grad:regionNE -116.11      88.98   -1.305  0.192660
## pct.hs.grad:regionS    58.68      79.07    0.742  0.458459
## pct.hs.grad:regionW  -336.02      96.68   -3.476  0.000564 ***
## pct.bach.deg:regionNE  249.06      81.75    3.047  0.002464 **
## pct.bach.deg:regionS   -16.38      67.49   -0.243  0.808307
## pct.bach.deg:regionW   179.27      75.49    2.375  0.018024 *
## pct.below.pov:regionNE  22.11     108.66    0.203  0.838870
## pct.below.pov:regionS  126.06      86.95    1.450  0.147850
## pct.below.pov:regionW -281.84     115.71   -2.436  0.015283 *
## pct.unemp:regionNE    -156.06     157.18   -0.993  0.321346
## pct.unemp:regionS     -242.09     140.18   -1.727  0.084928 .
## pct.unemp:regionW     -377.14     144.17   -2.616  0.009224 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1626 on 412 degrees of freedom
## Multiple R-squared:  0.8494, Adjusted R-squared:  0.8396
## F-statistic: 86.08 on 27 and 412 DF,  p-value: < 2.2e-16
```

When we add back region and interactions, we get a couple terms where every level is insignificant. Using our best judgement we attempt to remove these variables and reassess the fit.

```
lm_sub_region_res_small = lm(per.cap.income ~ (pop.18_34 +
      doctors + pct.hs.grad +
      pct.bach.deg + pct.below.pov +
      pct.unemp) * region -
      doctors:region - pop.18_34:region, data = cdi)
summary(lm_sub_region_res_small)
```

```
##
## Call:
## lm(formula = per.cap.income ~ (pop.18_34 + doctors + pct.hs.grad +
##   pct.bach.deg + pct.below.pov + pct.unemp) * region - doctors:region -
##   pop.18_34:region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4004.1  -890.1  -124.3   754.6  7260.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    26275.113    5452.666   4.819 2.02e-06 ***
## pop.18_34      -311.555     23.585  -13.210 < 2e-16 ***
## doctors         875.820     87.533  10.006 < 2e-16 ***
## pct.hs.grad    -121.194     68.159   -1.778  0.076114 .
## pct.bach.deg     341.063     45.709    7.462 5.00e-13 ***
## pct.below.pov  -444.613     72.731   -6.113 2.24e-09 ***
## pct.unemp       333.017    103.703    3.211 0.001424 **
## regionNE       5156.517    6563.365    0.786 0.432517
## regionS       -6025.465    5981.880   -1.007 0.314380
## regionW       26275.154    7729.723    3.399 0.000741 ***
## pct.hs.grad:regionNE  -106.707     83.943   -1.271 0.204366
```

```
## pct.hs.grad:regionS      79.713      76.703      1.039 0.299297
## pct.hs.grad:regionW     -324.506      91.069     -3.563 0.000408 ***
## pct.bach.deg:regionNE    206.265      56.213      3.669 0.000275 ***
## pct.bach.deg:regionS      -9.271      50.357     -0.184 0.854025
## pct.bach.deg:regionW     173.775      56.612      3.070 0.002284 **
## pct.below.pov:regionNE   -18.314     100.827     -0.182 0.855957
## pct.below.pov:regionS    159.038      79.973      1.989 0.047393 *
## pct.below.pov:regionW   -273.588     112.101     -2.441 0.015079 *
## pct.unemp:regionNE      -178.932     157.437     -1.137 0.256388
## pct.unemp:regionS       -305.351     138.413     -2.206 0.027921 *
## pct.unemp:regionW       -373.807     143.196     -2.610 0.009367 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1638 on 418 degrees of freedom
## Multiple R-squared:  0.845, Adjusted R-squared:  0.8372
## F-statistic: 108.5 on 21 and 418 DF,  p-value: < 2.2e-16
```

Our reduced model has a significant coefficient for at least one level for all interactions and categorical variables. Next, we see if it lost any explainability through ANOVA.

```
anova(lm_sub_region_res, lm_sub_region_res_small)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ (pop.18_34 + doctors + pct.hs.grad + pct.bach.deg +
##   pct.below.pov + pct.unemp) * region
## Model 2: per.cap.income ~ (pop.18_34 + doctors + pct.hs.grad + pct.bach.deg +
##   pct.below.pov + pct.unemp) * region - doctors:region - pop.18_34:region
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1      412 1089211145
## 2      418 1121004010 -6 -31792865 2.0043 0.06398 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the full model does not necessarily explain any of the variation in the response better than the reduced model. So, we are able to remove these insignificant terms without losing modeling power.

A natural next question is to experiment with removing more interactions and reassessing the fit. We tried a few combinations of interactions to take out, but all of them resulted in a worse fit, as per the output below.

```
lm_sub_region_res_smaller = lm(per.cap.income ~ (pop.18_34 + doctors + pct.hs.grad + pct.bach.deg + pct
anova(lm_sub_region_res_smaller, lm_sub_region_res_small)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ (pop.18_34 + doctors + pct.hs.grad + pct.bach.deg +
##   pct.below.pov + pct.unemp) * region - doctors:region - pct.unemp:region -
##   pop.18_34:region
## Model 2: per.cap.income ~ (pop.18_34 + doctors + pct.hs.grad + pct.bach.deg +
##   pct.below.pov + pct.unemp) * region - doctors:region - pop.18_34:region
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1      421 1142203076
```

```
## 2    418 1121004010 3 21199066 2.6349 0.04943 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, the full model with more interaction terms explains the data better than the one without these terms. So, we do not have a justification for removing any more interaction terms.

As a final check, we can assess whether or not including region and the interactions actually improved the fit with another ANOVA.

```
anova(lm_sub_res, lm_sub_region_res_small)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ pop.18_34 + doctors + pct.hs.grad + pct.bach.deg +
##   pct.below.pov + pct.unemp
## Model 2: per.cap.income ~ (pop.18_34 + doctors + pct.hs.grad + pct.bach.deg +
##   pct.below.pov + pct.unemp) * region - doctors:region - pop.18_34:region
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      433 1.405e+09
## 2      418 1.121e+09 15  2.84e+08 7.0597 8.025e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output above shows a significant F score and thus including region and the interactions we chose did in fact help us with the fit.

```
summary(lm_sub_region_res_small)
```

```
##
## Call:
## lm(formula = per.cap.income ~ (pop.18_34 + doctors + pct.hs.grad +
##   pct.bach.deg + pct.below.pov + pct.unemp) * region - doctors:region -
##   pop.18_34:region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4004.1  -890.1  -124.3   754.6  7260.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    26275.113    5452.666   4.819 2.02e-06 ***
## pop.18_34      -311.555     23.585 -13.210 < 2e-16 ***
## doctors         875.820     87.533  10.006 < 2e-16 ***
## pct.hs.grad    -121.194     68.159  -1.778 0.076114 .
## pct.bach.deg     341.063     45.709   7.462 5.00e-13 ***
## pct.below.pov  -444.613     72.731  -6.113 2.24e-09 ***
## pct.unemp       333.017    103.703   3.211 0.001424 **
## regionNE       5156.517    6563.365   0.786 0.432517
## regionS      -6025.465    5981.880  -1.007 0.314380
## regionW       26275.154    7729.723   3.399 0.000741 ***
## pct.hs.grad:regionNE  -106.707     83.943  -1.271 0.204366
## pct.hs.grad:regionS    79.713     76.703   1.039 0.299297
```

```
## pct.hs.grad:regionW      -324.506      91.069   -3.563 0.000408 ***
## pct.bach.deg:regionNE    206.265      56.213    3.669 0.000275 ***
## pct.bach.deg:regionS     -9.271      50.357   -0.184 0.854025
## pct.bach.deg:regionW     173.775      56.612    3.070 0.002284 **
## pct.below.pov:regionNE  -18.314     100.827   -0.182 0.855957
## pct.below.pov:regionS    159.038      79.973    1.989 0.047393 *
## pct.below.pov:regionW   -273.588     112.101   -2.441 0.015079 *
## pct.unemp:regionNE      -178.932     157.437   -1.137 0.256388
## pct.unemp:regionS       -305.351     138.413   -2.206 0.027921 *
## pct.unemp:regionW       -373.807     143.196   -2.610 0.009367 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1638 on 418 degrees of freedom
## Multiple R-squared:  0.845, Adjusted R-squared:  0.8372
## F-statistic: 108.5 on 21 and 418 DF,  p-value: < 2.2e-16
```

Above is the summary for the final model we choose. Next, we validate its assumptions and examine multicollinearity.

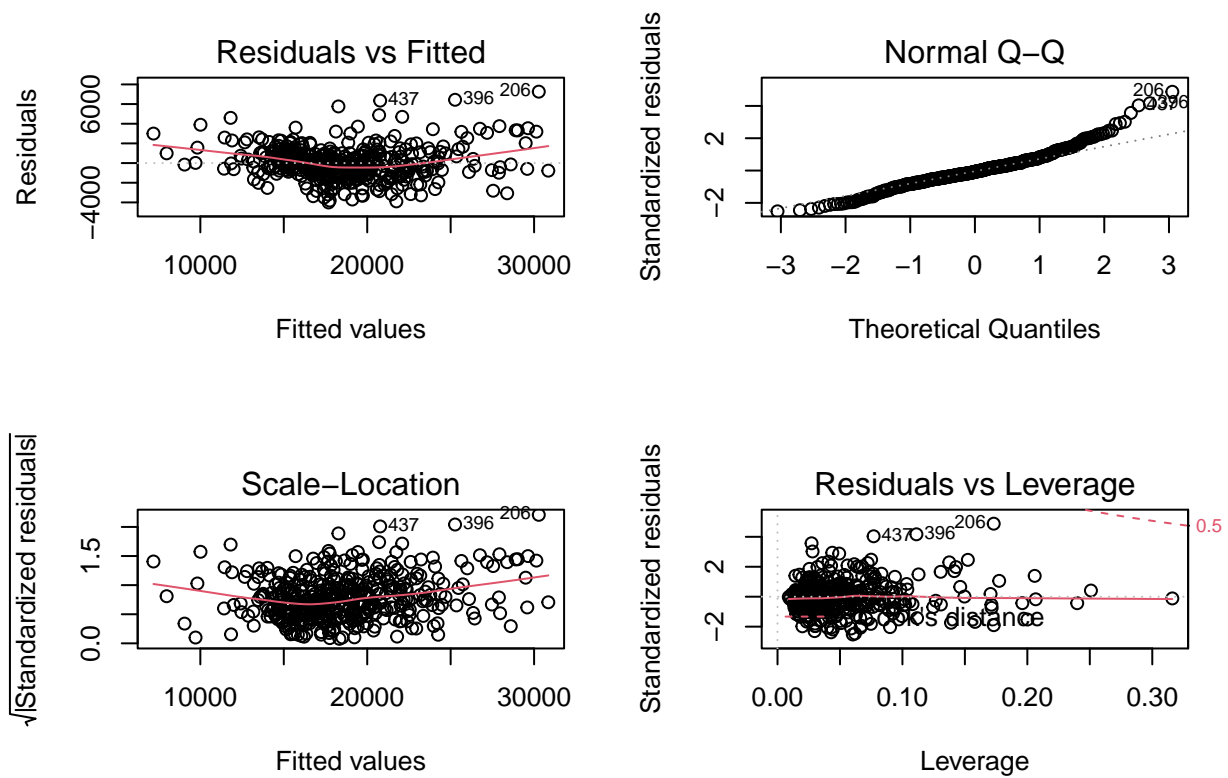
```
vif(lm_sub_region_res_small)
```

```
##          pop.18_34          doctors          pct.hs.grad
##          1.599381          1.641602          37.424800
##          pct.bach.deg          pct.below.pov          pct.unemp
##          20.038883          18.777267          9.622177
##          regionNE          regionS          regionW
##          1267.175237          1327.479835          1415.277701
##          pct.hs.grad:regionNE          pct.hs.grad:regionS          pct.hs.grad:regionW
##          1256.253171          1253.138967          1262.106587
##          pct.bach.deg:regionNE          pct.bach.deg:regionS          pct.bach.deg:regionW
##          50.380733          51.259727          41.669411
##          pct.below.pov:regionNE          pct.below.pov:regionS          pct.below.pov:regionW
##          17.128812          37.997555          30.766221
##          pct.unemp:regionNE          pct.unemp:regionS          pct.unemp:regionW
##          40.357617          31.264306          31.060862
```

As expected, we see some multicollinearity. However, because this is due in part to the inclusion of interactions, because BIC approximates an explanatory and parsimonious model, and because most of the coefficients are meaningful in sign, we choose to keep the variables we have.

We are focused on an interpretable model, so we do not inspect LASSO as a regularized model does not give us the necessary statistical information. As a result of these analysis, we consider the backward step/all subsets model plus region and a few interactions to be our best and move forward with validating its assumptions.

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



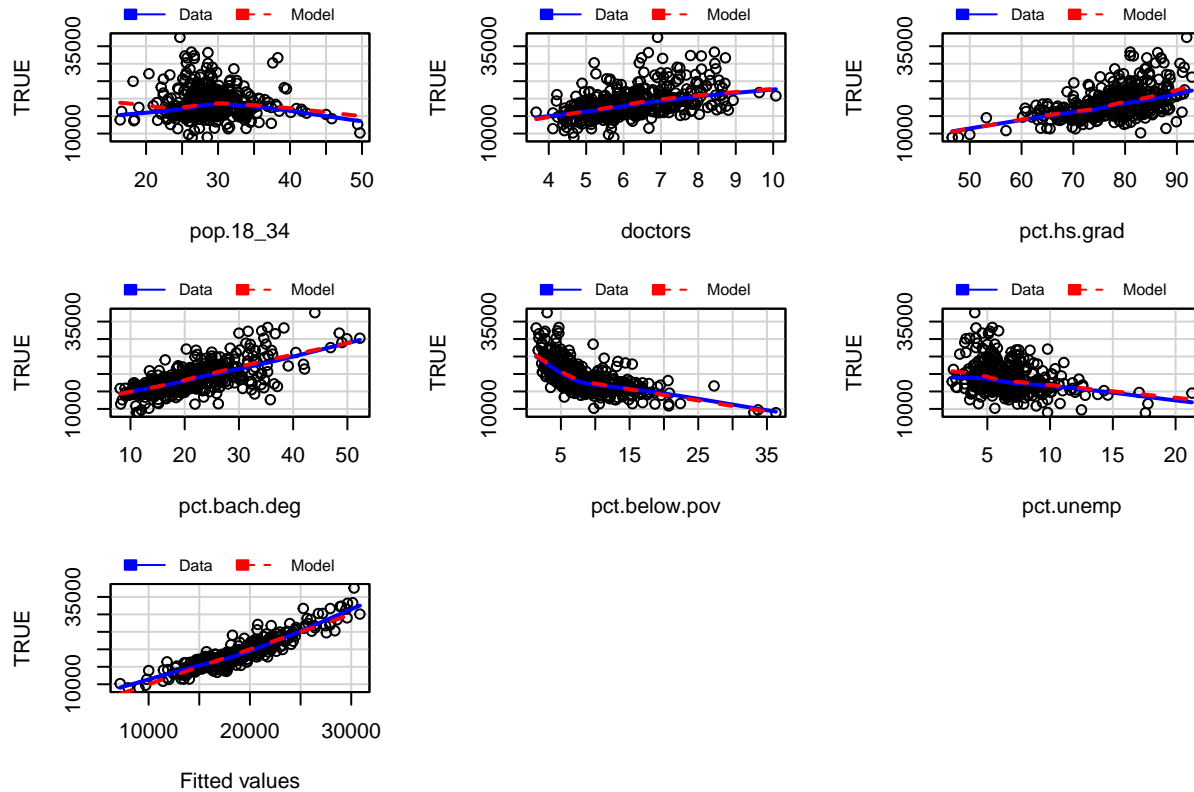
```
library(car)

mmps(lm_sub_region_res_small, terms = ~.-region)
```

```
## Warning in mmps(lm_sub_region_res_small, terms = ~. - region): Interactions and/
## or factors skipped
```



## Marginal Model Plots



From the diagnostics and marginal model plots above, we can see that the assumptions are mostly valid. While there is some deviation from the normal line in the qq plot in the top right, there is only random scatter in the standardized residuals, and most of them are close to the qq line besides the top right. There are only a few apparent outliers or poor leverage points. Furthermore, the marginal model plots show that the fits of all the quantitative predictors in the model are accurate, as the estimated curves are close to each other. So, the transformations we chose, while less interpretable, did in fact improve the model assumptions and validity. In this instance, we choose to sacrifice interpretability because a powerful model that satisfies all assumptions can still let us make concise and relevant arguments about the relationships between the predictors and the response. Additionally, the log function is monotonic so an increase in a log predictor coefficient, can be qualitatively interpreted in the same direction as the natural predictor coefficient.

Throughout our model selection process, we had to make several tradeoffs in terms of interpretability, modeling assumptions, and predictive power. We chose to find a model that satisfied at least a baseline of all three. We did not use any complex transformations and only implemented log transforms, which are simple to interpret in context. Our final model has a high R squared, but it is not the highest we saw; the full model had a higher adjusted R-squared but broke some assumptions due to the VIFs. We made the tradeoff of reliable estimates in exchange for more explainability of the response. We can still interpret these coefficients in context despite the VIFs. Additionally, we are more inclined to include all terms, even multicollinear ones, because our selection came from BIC so these variables are necessary in approximating the true model. Our model has decently valid assumptions and interpretable coefficients, so it meets all relevant criteria for this problem. Although not the “best” in any one category, its usefulness in interpretability and predictive power, along with its validity will make this model a valuable tool for any stakeholder.