Socioeconomic Factors that Affect Per Capita Income

Mitchell Pudil[*]
mpudil@andrew.cmu.edu

10 October 2019

Abstract

We examine economic data on 440 of the most populous counties in the United States using the dataset provided by Kutner et al. (2005). We first look at correlations across several economic variables, including unemployment, per capita income, region, education, and so forth. We determine if the relationship between crime rate and per capita income differs by region of the United States. We attempt to find the best explainable model for per capita income given Kutner's dataset. We answer these questions through use of correlation matrices, VIF's, partial F-tests, and linear regression. We find evidence to support the claim that the relationship between crime rates and per capita income varies by region. We also find evidence suggesting that population density, education, unemployment, region, and several other demographic variables are related to per capita income. We end with a discussion regarding the drawbacks of the data and discuss measures, such as an increase in the number of states and counties represented, that should be taken in future research to improve the internal and external validity of the research presented.

1 Introduction

per capita income is an important subject matter for social scientists. It determines the way people live, and what opportunities are available to them. However, there is also much confusion about the relationships between per capita income and other socioeconomic factors such as crime, region, and education. For example, many argue that the relationship between crime and per capita income is dependent on the region of the United States one lives in. For all of these reasons, we make an attempt at answering the following questions

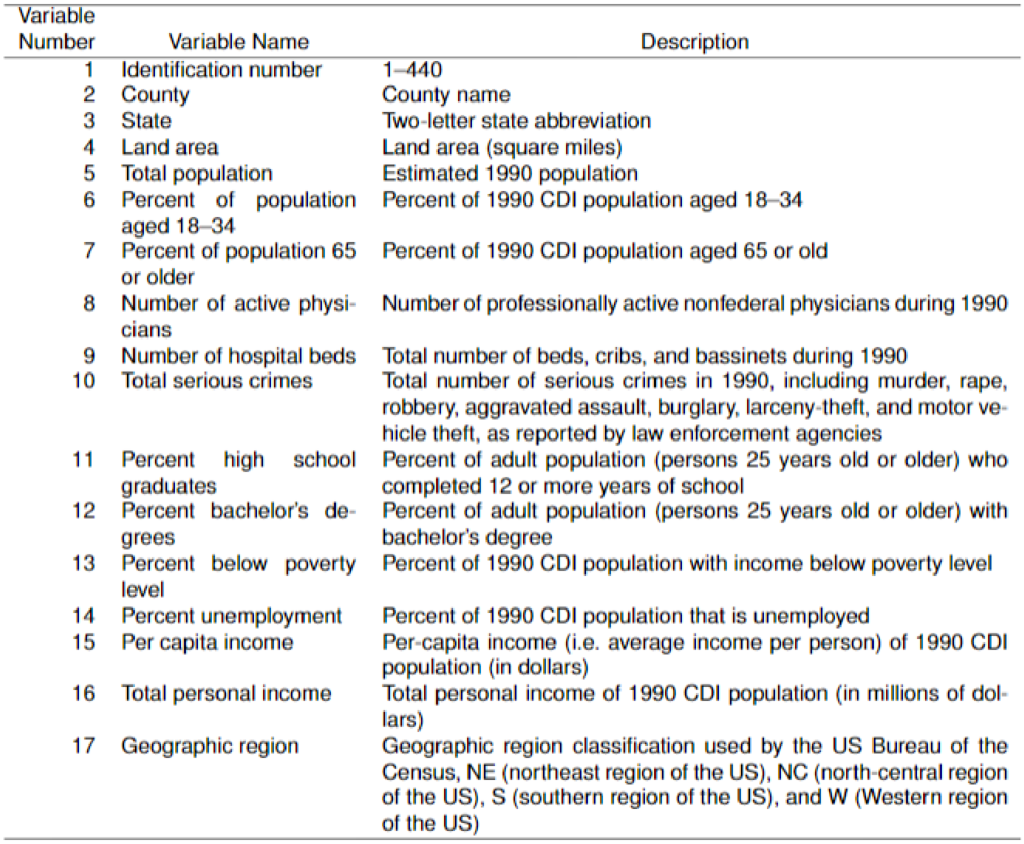
- What socioeconomic factors are related to each other, and are these relationships surprising or explainable?
- Does the relationship between crime and per capita income vary by region?
- Which explainable model predicts per capita income best?
- Is it worrisome that only 373 counties and 48 states are represented in the dataset?

[*] Master's Student of Statistical Practice at Carnegie Mellon University

This paper is structured as follows. Section 2 briefly describes the data that was used for this analysis along with the methodology that allows us to answer each question. Section 3 provides a more thorough explanation of how the methods were used to answer the first three questions listed above along with the results that were found. Section 4 answers the question regarding the sampling of the data and discusses the extent to which the data and the model are accurate. It also provides suggestions for future work. The paper ends with a code appendix, which contains more details regarding the process by which the questions were answered along with additional tables, figures, and summaries.

2 Methods

The data for this study come from Kutner et al. (2005) and contains 17 variables related to the 440 of the most populous county along with related economic, health, and social well-being statistics. The full list of initial variables is shown above as Figure 1 below. Recall, however, that some variables have been transformed to produce other variables and that a couple of the variables are not used at all in the final model.



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)

Figure 1: Description of Variables. Data collected from Kutner et al. (2005).

The first question regards the relationship between the economic and health-related factors available in the dataset. We used a correlation matrix, boxplots, and VIFs to show and quantify the relationships. For reasons later expressed in the discussion section, however, these relationships cannot be proven here to be causal. Furthermore, because of the small sample size and the fact that not all states were included in the model, no plots or correlations were analyzed between state and economic factors. The results section explains the many relationships between the economic variables.

Secondly, we are interested in whether or not the relationship between crime and per capita income vary by region. Briefly, to determine whether there was an effect, we regressed per capita income on crime, region, and an interaction between crime and region. This model was then compared to one without the interaction term via a partial F-test. The analysis is described more fully in the results section, and the specific regression and its results are shown on page 6 of the Appendix.

The next question regards finding the best model. In order to answer the question, we used linear regression, correlation matrices, and diagnostic plots to determine the optimal explainable model to predict per capita income. We used the R language for statistical computing (R Core Team, 2017). Along with the results section, the full process by which the best model was chosen is shown in detail on pages 7-22 of the Appendix.

Finally, we are interested in understanding the consequences of the sample of data we use, especially with regards to the subsets of counties used. In this case, most of the arguments for the issues are based on reasoning alone rather than figures or other analyses. Since this is the case, this question will be answered in the discussion section.

Throughout the analysis, various variable transformations and interactions were implemented. Briefly, we started by dividing variables such as doctors, crimes, and hospital beds by population to match other variables that were already in per capita terms. In the final model, population was divided by land area to have a population density variable, and then logged as it led to a better-predicted model and was still interpretable (see pages 1 and 9 of the Appendix). The population variable itself was taken out, along with land area and total income to avoid perfect multicollinearity with the other predictors or an R squared of 1. Interactions between region and unemployment rate and between region and the percentage of high school grads were implemented and stayed in the final model because of interpretability, statistical significance, and the fact that the results are interesting (see pages 14-15 of the Appendix). Variables already in terms of percentages or otherwise transformed to be in terms of percentages were not further transformed to maintain interpretability.

3 Results

We would now like to answer the first three questions asked in the introduction. For each question, we explain the statistical analysis and the results in the order parallel to the introduction and methods sections.

3.1 Relationships Among Economic Factors

The first question we are interested in is: which of our factors are related to which other factors? Moreover, if there are relationships between variables, do they make intuitive sense? To answer this question, we look to Figures 2 and 3. We notice that there are several relationships, both positive and negative, among the variables.

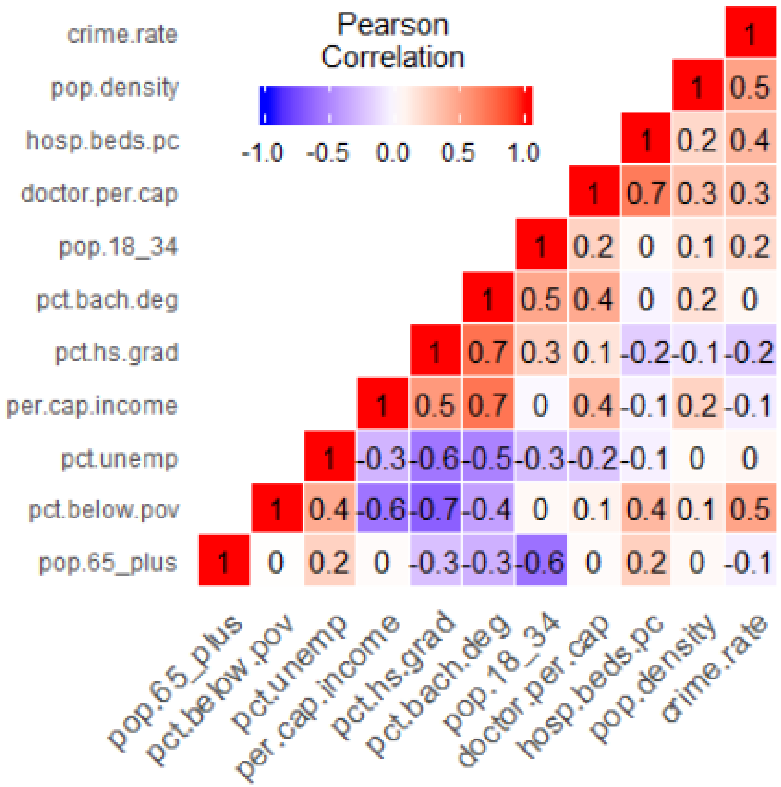


Figure 2: Correlation Matrix of Continuous Socioeconomic Variables

First, the number of hospital beds per capita, the number of doctors per capita, the population density, and the crime rate are all positively correlated with each other. The relationship crime rate and population density is a bit difficult to explain, since it could be argued that having more people around would lower the chances of getting shot since there would be more witnesses; however, there are so many types of crimes that can be committed that could theoretically be done regardless of the population density. While a formal analysis of this is outside the scope of this paper, other research articles such as Ladd (1992) have looked extensively at this issue.

We also find high levels of positive correlation among the percentage of high school grads, the percentage of bachelors degrees, and per capita income. This makes sense for a couple of reasons. First, in more well-off areas where people have more money, they can use the money to gain an education. On the other hand, those who graduate from high school or college tend to make more money, so their income increases. Muller (2002) showed that the lack of high school education accounts for the income inequality effect.

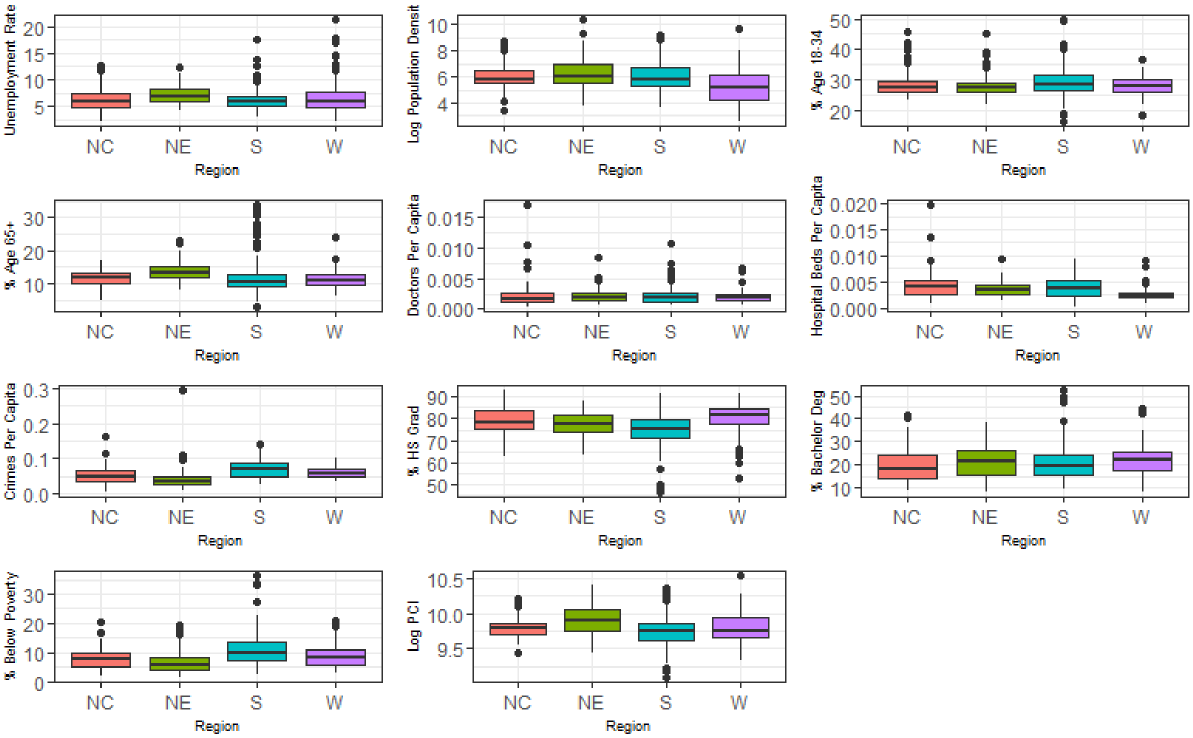


Figure 3: Boxplots of Continuous Variables by Region

Additionally, the unemployment rate and the education rates[1] are negatively related. This makes sense because one of the primary functions of education is to earn a job, so when a large portion of the population has an education, it is likely that the majority are employed as well. Mincer (1991) expounded on this idea and showed that the reduction of the incidence of unemployment is found to be far more important than the reduced duration of unemployment in creating the educational differentials in unemployment rates.

Unemployment rate is also positively related to the percentage below poverty level. This is intuitive as those without a job would usually be below poverty level. On the other hand, if poverty level is low, then the economy is doing well, and so unemployment would probably decrease. We also find that the percent of population above age 65 is negatively correlated to the population between age 18-34. This is trivial since if the population has a smaller percentage of younger people, it must have a larger proportion of middle-aged and older people.

Per capita income is also negatively related to the percentage below poverty and unemployment rate. It makes sense that per capita income is negatively related to poverty and unemployment rates because if the average person has a greater amount of money, then it is likely that most people have jobs and are above the poverty line.

In Figure 3, we notice that land area varies quite a bit in the Western United States. This makes sense since there are many large counties in the West, such as San Bernandino, LA, and Humboldt county. The South tends to have a large ranges for age, education, and percentage below poverty. Interestingly, there does not appear to be high differences in crime rates by region, although there are counties in each region with high crime rates.

[1] both high school grads and bachelors degrees

We also looked at VIFs[2]. None of the variables contained in the correlation matrix had a VIF over 4. However, when looking at the non-transformed variables, there are some variables such as the number of doctors and the number of hospital beds that are highly collinear. Further information regarding the VIFs of the models, see pages 1, 2 and 11 of the Appendix.

Overall, there appears to be many relationships between economic factors. However, these correlations alone cannot be interpreted as causal effects. Also, the correlations are constrained to the 1990-1992 years for more highly-populated areas.

3.2 Effect of Crime on Per Capita Income by Region

The next question we were interested in is if the relationship between crime and per capita income changes by region, disregarding all other variables. To answer this question, we first looked at a graph of the relationship between crimes and per capita income for each region (see Figure 4 below).

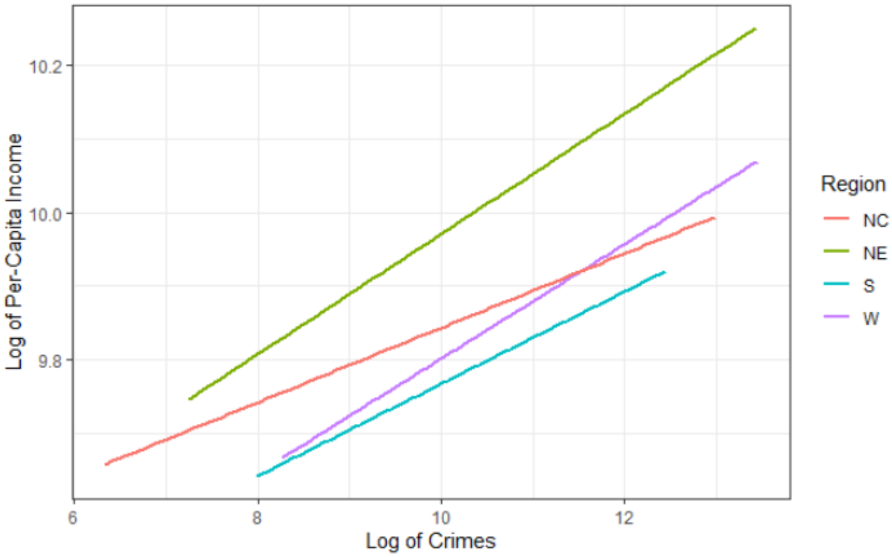


Figure 4: Relationship Between Crime and Per Capita Income, by Region

We notice that the slopes and intercepts for each line differ. The change in the intercept alone suggests that there is likely an effect of region on per capita income. A significant change in the slope, however, might suggest that the relationship between crime and per capita income varies by region. In Figure 4, it appears that there is a change in both the intercept and slope for each region. However, keeping in mind the small sample size of the data, it becomes less clear whether or not those changes are statistically significant.

Because of this uncertainty, we use transformation and interaction techniques to find the best interpretable model that would answer such question. Specifically, both the per capita income and crimes were logged as suggested by doing power transformations through the Box Cox method, and were rounded for interpretability. The interaction term then allows us to estimate the relationship between crime and per capita income for each region.

[2] Variable Inflation Factors, used to measure collinearity and defined as $\frac{1}{1-R_i^2}$

The model used to answer this question was:

$$\log(\mathrm{PCI}) = \beta_0 + \beta_1 \log(\mathrm{crimes}) + \alpha\,\mathrm{R} + \gamma\,\mathrm{CR} + \epsilon \tag{1}$$

In the regression above R represents the Northeast, South, and West regions of the United States, each of which are multiplied by some coefficient. CR represents the interaction between the log of crime and each of the regions. Page 6 in the Appendix shows that the relationship between crime and per capita income is significant as well as the relationship between region and per capita income.

Although the individual interaction terms are not statistically significant, it is necessary to conduct a partial F-test[3] against a model without the interaction terms to determine whether or not region (in general) affects the relationship between crime and per capita income. In other words, we need to find the probability that all of the estimates on the interaction terms are equal to 0. This partial F-test, shown in the Appendix on page 6, results in a p-value less than 0.001, suggesting that the relationship between crime and per capita income does vary by region. Unfortunately, because each of the interaction terms individually are not statistically significant, it is not possible to determine which region(s) affect(s) this relationship. However, we can conclude that there is some sort of effect of region on the relationship between crime and per capita income.

3.3 Best Explainable Model

Next, we would like to determine what the "best" model is to predict per capita income. In this case, we define "best" as one that best reflects the social science and the meaning of the variables. It also best satisfies the modelling assumptions. The "best" model should also be the one most indicated by the data, but that can also be explained to someone who is more interested in social, economic and health factors than in mathematics and statistics.

The process of determining such a model involved many steps, many of which were recursive. It involved determining the correct transformations and interactions with the best combination of interpretability and predictive power. It also involved looking at diagnostic plots, summary statistics, leverage points and outliers, and predictive summaries such as AIC, BIC, and adjusted R squared.

The final model used was:

$$\begin{aligned}\log(\mathrm{PCI}) = {} & \beta_0 + \beta_1 \log(\mathrm{Pop.Density}) + \beta_2\,\mathrm{Pop.18.to.34} + \beta_3\,\mathrm{Hosp.Beds.Per.Capita} + \\ & \beta_4\,\mathrm{Crime.Rate} + \beta_5\,\mathrm{Percent.HS.Grads} + \beta_6\,\mathrm{Percent.Bach.Degree} + \beta_7\,\mathrm{Percent.Below.Poverty} + \\ & \beta_8\,\mathrm{Percent.Unemployed} + \alpha\,\mathrm{R} + \gamma\,\mathrm{RU} + \delta\,\mathrm{RH} + \epsilon\end{aligned} \tag{2}$$

In the regression above, RU represents the interaction terms between region and unemployment. RH represents the interaction between the region and the percentage of high school graduates in the area. We began the process of finding this model by first dividing population by other variables[4] so that all predictors were consistent but also made logical sense. For example, the number of crimes was converted into crime rate and the number of hospital beds was converted into the number of hospital beds per capita. This model was compared to one where the only transformations performed

[3] A nested model F tests in which the reduced model is something other than the constant-only model

[4] In the case of population density, population was divided by land area

were ones suggested by the Box-Cox method that would also be easily interpretable (see page 8 of the Appendix). It was found that dividing population by other variables provided a model with lower AIC and BIC, and less collinearity (see page 14 of the Appendix).

Next, interaction terms between region and unemployment and between region and the percentage of high school graduates were included. Both interactions had at least one term that was statistically significant (see page 15 of the Appendix). This allowed us to reason that both the relationship between unemployment and per capita income as well as the relationship between education and per capita vary by region. Next, the doctors per capita variable and percentage of population over 65 were dropped because they were not statistically significant and there were other variables in the model that were similar to those variables (see pages 15-16 of the Appendix). This again decreased AIC and BIC, and did not affect adjusted R squared hardly at all. Stepwise regression implemented also showed that those variables should be dropped (see page 17 of the Appendix).

Finally, “bad”[5] leverage[6] points were examined. Only one was found, but after taking the observation out, prediction performance increased noticeably (see page 19 of the Appendix). This final model was also shown to have constant variance of error terms (see page 20 of the Appendix). The normality of the error terms was not perfectly normal, but likely due to the categorical variables included in the model. K-fold cross validation (where k=5) showed that the average error was 6.1%, which, while imperfect, is an acceptable estimation of per capita income (see pages 21-22 of the Appendix). The adjusted R squared for this model is 0.85, which helps to suggest excellent predictive power. We find statistically significant effects of the following variables: population density, percent of population between 18 and 34, hospital beds per capita, crime rate, percentage of population with a bachelor’s degree, Western United States[7], and some terms in the either of the interactions.

As implied earlier, the reason this model was chosen over several others was because it was the model with the best predictive power that was also easily explainable to social scientists. Of course, not all possible models were tested; doing so would be computationally difficulty. Now, this model did not have the lowest AIC or the highest R-squared. In fact, one potential model of simply regression per capita income on the ratio of total income and population would theoretically receive a perfect R-squared score. However, doing so would provide very little information regarding the effect of many other health or economic factors.

In conclusion, there appears to be several strong predictors of per capita income, both from health and economic statistics. The final model does a decent job at predicting per capita income using a variety of economic variables while providing interesting connections between the variables and per capita income.

4 Discussion

The model determined in the results section has many interesting results. Unfortunately, there are some problems and cautions that come as a result from using this particular data and the model itself. In order to acknowledge these issues, we must understand how the fact that our data represents only 48 states and 373 of the approximate 3000 counties in the United States affects our

[5] Leverage points that are outliers

[6] The cutoff for leverage that was used is $\frac{2(p+1)}{n}$ where p is the number of explanatory variables and n is the number of observations

[7] The NC region was used as a baseline

interpretation of the model.

Generally speaking, the number of states and counties appears to be a good representative sample. That is, most of the states are represented, and a smaller but still presumably useful number of counties are representative. However, with our particular questions, these data have major drawbacks which affect the internal and external validity of the experiment. First, the counties in the dataset were not randomly chosen; they were the most populous counties. Therefore, we cannot extrapolate the model to less populated counties, which unfortunately leaves out the majority of the United States. Along these same lines, because all of the data were collected from 1990 to 1992, much of the results are irrelevant for the modern day. For example, it is possible that the relationship between unemployment and per capita income, or the relationship between education and per capita income has changed over the nearly two decades. This could be because of changes in education, changes in the economy, or the differences in population.

Further, it is difficult to feel comfortable including state fixed effects in the model as doing so would take up many degrees of freedom. Additionally, many states such as DC, MT, ND, SD, VT, and WV are all represented by just one county, and other states are not represented at all (see page 22 of the Appendix). By not including state in the regression, we cannot accurately factor out the unobserved differences within each state, which weakens the internal validity of models (1) and (2).

We also find that the number of counties represented for each state is highly unbalanced, as they are correlated with population. With more counties, and with the rest of the states, we could have a better representative sample, and potentially be able to analyze the relationships between each state and the per capita income, and perhaps how the relationships between variables such as crime and per capita income varies by state.

The model itself is also a bit problematic, especially if causation is implied. While the data does produce an excellent adjusted R squared of approximately 0.86, we do have endogeneity problems. For example, the model is likely subject to reverse causality. It is difficult to know whether many of the predictors, such as unemployment or education, affects per capita income, or if it is the per capita income that affects education and unemployment. Likely, both relationships are true for the reasons previously mentioned. There are likely many more variables that we did not account for, such as state[8], year, and other socioeconomic, health, or even technological advances.

Disregarding the issues in the data, we were able to answer some of the questions we had posed. First, we were interested in understanding the relationships between the variables used in the dataset. We found that there is high positive correlation between the number of hospital beds per capita, the number of doctors per capita, total income population, and crime rate. We found that education and per capita income are highly correlated. We found that the unemployment rate and education rates are negatively correlated, and that unemployment rate is positively related to the percentage below the poverty level. We determined that the percentage of the population above 65 is negatively correlated to the population between 18-34. We also found that per capita income is negatively related to land area, percentage below poverty, and unemployment rate. Future research on this should either use more current data or data from the past few decades to truly understand the relationship between these variables as it is possible that the correlation between some of these variables change over time.

The second question we were interested in answering was whether the relationship between crime and per capita income depended on the region of the United States. While model (1) was a bit simplistic and did not control for many variables, the data suggested that region did play a role

[8] Which again could theoretically have been used if we had more observations

in the relationship between crime and per capita income. Further research here could be focused on determining how strong this relationship is currently, or even how the impact of region on the relationship between crime and income has changed across the years.

Finally, we were able to determine the ”best” model given our dataset. While it would have been computationally difficult to truly calculate all possible models, we did find a model that was both interpretable and that had high statistical prediction power but that best satisfied modelling assumptions. Future research regarding this question should include a much larger sample from all states and many more (if not all) counties from across many years. Again, if causation is to be inferred, then more sophisticated models such as instrumental variable (IV) regression should be used to counteract the endogeneity problems.

Overall, while there are issues with having this subset of counties and states represented, there are still some interesting connections we can make between many of the variables in the data. However, this model is likely better off to be used for prediction and correlation alone[9] than causation and inference.

References

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

Ladd, Helen F. 1992. “Population Growth, Density and the Costs of Providing Public Services.” Urban Studies 29(2):273–95.

Mincer, Jacob 1991. ”Education and Unemploymen.” NBER Working Paper No. w3838. Available at SSRN: https://ssrn.com/abstract=226736.

Muller, A. 2002. “Education, Income Inequality, and Mortality: a Multiple Regression Analysis.” Bmj 324(7328):23–23.

R Core Team (2017), *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

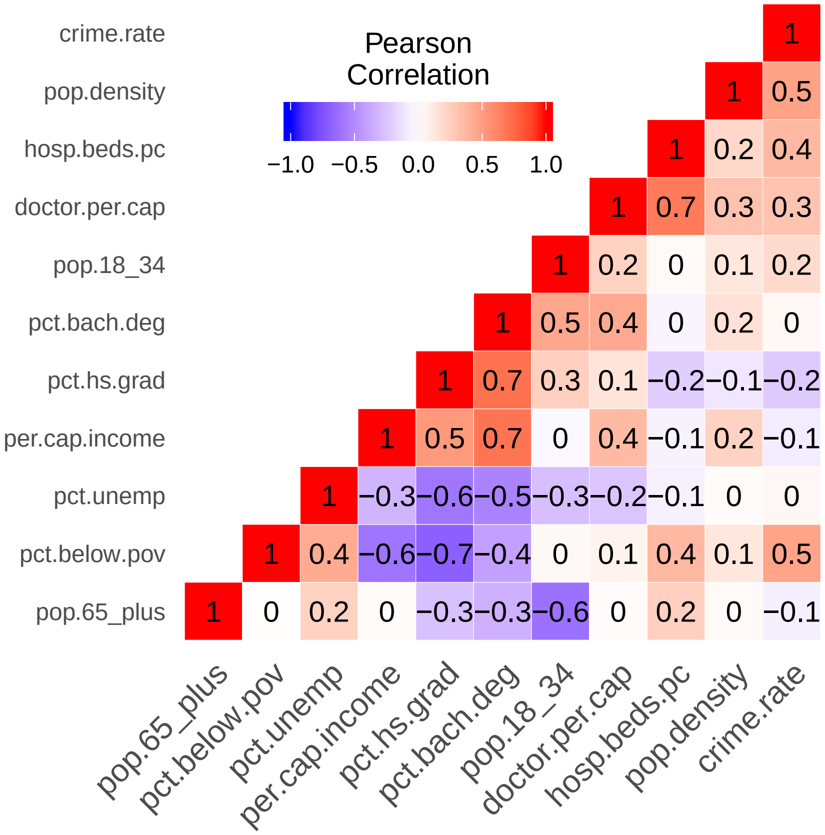
[9] And even then, with caution

Appendix

To start off, let's consider some variables that could be meaningful to include in the model from a social science perspective.

- Population density = land area / population
- Crime rate = crimes / population
- Percent doctors: Doctors / population
- Percent hospital beds: hospital beds / population

```
cdi <- read.table("C:/Users/Mitchell Pudil/Downloads/cdi.dat")[,-1] %>%
  mutate(pop.density = pop/land.area, crime.rate = crimes/pop,
         doctor.per.cap = doctors/pop,
         hosp.beds.pc = hosp.beds/pop)


heatmapCreator(cdi[,-c(1:4,7:9,15:16)])
```



```
vif(lm(per.cap.income~., data=cdi[,-c(1:4,7:9,15:16)]))
```

```
##      pop.18_34     pop.65_plus     pct.hs.grad    pct.bach.deg  pct.below.pov
##       1.977524        1.971229        3.767916        3.550070       3.008835
##      pct.unemp    pop.density     crime.rate doctor.per.cap    hosp.beds.pc
##       1.825396       1.470406       1.826209       3.164631       3.040984
```

We notice from the above correlation heatmap that: • The number of hospital beds, the number of doctors, the total income, population, and crimes are all positively correlated (see upper right)

- The percentage of high school grads, percentage of bachelors degrees and per capita income are highly related to each other

- The unemployment rate and the education rates (hs grads and bachelors degrees) are negatively related and positively related to the percentage below poverty level.

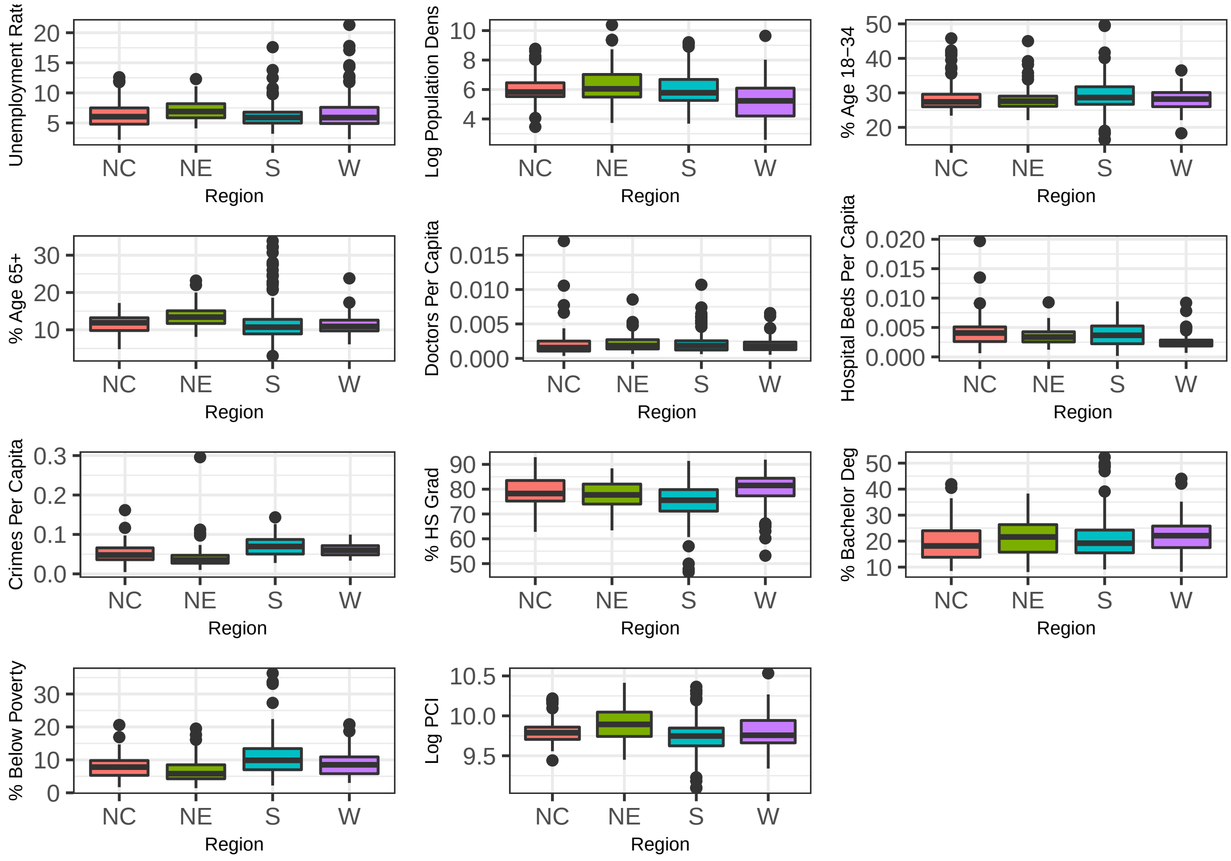
- The percent of population above 65 is negatively correlated to the population between 18-34.

- Per capita income is positively related to total income, but perhaps not as highly as expected. It is also negatively related to land area, percentage below poverty, and unemployment rate.

Let's look at the relationships between region and the quantitative variables.

```r
unemployplot <- ggplot(cdi, aes(x=region, y=pct.unemp, fill=region)) +
  geom_boxplot() + theme_bw() + labs(x="Region", y="Unemployment Rate") +
  theme(legend.position = "none") + theme(axis.title=element_text(size=7))


landplot <- ggplot(cdi, aes(x=region, y=log(pop/land.area), fill=region)) +
  geom_boxplot() + theme_bw() + labs(x="Region", y="Log Population Density") +
  theme(legend.position = "none") + theme(axis.title=element_text(size=7))

youngplot <- ggplot(cdi, aes(x=region, y=pop.18_34, fill=region)) +
  geom_boxplot() + theme_bw() + labs(x="Region", y="% Age 18-34") +
  theme(legend.position = "none") + theme(axis.title=element_text(size=7))


oldplot <- ggplot(cdi, aes(x=region, y=pop.65_plus, fill=region)) +
  geom_boxplot() + theme_bw() + labs(x="Region", y="% Age 65+") +
  theme(legend.position = "none") + theme(axis.title=element_text(size=7))


docplot <- ggplot(cdi, aes(x=region, y=doctors/pop, fill=region)) +
  geom_boxplot() + theme_bw() + labs(x="Region", y="Doctors Per Capita") +
  theme(legend.position = "none") + theme(axis.title=element_text(size=7))


hosbedsplot <- ggplot(cdi, aes(x=region, y=hosp.beds/pop, fill=region)) +
  geom_boxplot() + theme_bw() + labs(x="Region", y="Hospital Beds Per Capita") +
  theme(legend.position = "none") + theme(axis.title=element_text(size=7))


crimeplot <- ggplot(cdi, aes(x=region, y=crimes/pop, fill=region)) +
  geom_boxplot() + theme_bw() + labs(x="Region", y="Crimes Per Capita") +
  theme(legend.position = "none") + theme(axis.title=element_text(size=7))


hsplot <- ggplot(cdi, aes(x=region, y=pct.hs.grad, fill=region)) +
  geom_boxplot() + theme_bw() + labs(x="Region", y="% HS Grad") +
  theme(legend.position = "none") + theme(axis.title=element_text(size=7))
```

```
bachplot <- ggplot(cdi, aes(x=region, y=pct.bach.deg, fill=region)) +
  geom_boxplot() + theme_bw() + labs(x="Region", y="% Bachelor Deg") +
  theme(legend.position = "none") + theme(axis.title=element_text(size=7))


povertyplot <- ggplot(cdi, aes(x=region, y=pct.below.pov, fill=region)) +
  geom_boxplot() + theme_bw() + labs(x="Region", y="% Below Poverty") +
  theme(legend.position = "none") + theme(axis.title=element_text(size=7))


percapplot <- ggplot(cdi, aes(x=region, y=log(per.cap.income), fill=region)) +
  geom_boxplot() + theme_bw() + labs(x="Region", y="Log PCI") +
  theme(legend.position = "none") + theme(axis.title=element_text(size=7))



ggarrange(unemployplot, landplot, youngplot, oldplot, docplot, hosbedsplot,
          crimeplot, hsplot, bachplot, povertyplot, percapplot, ncol=3, nrow=4)
```



The above plot shows us that:

- Unemployment rate and land area varies most in the West

- There are a handful of outlier counties in the South where the percentage of high school graduates are less than 60%, but there are also several counties in the South where the percent of bachelor degrees in the population is above 35%.

- The median percentage of people by county that live below the poverty rate is highest in the South.
- There does not appear to be huge overall differences in crime rates by region, although there are counties in each region where the crime is high.

We can also look at VIFs which come from a linear regression of per capita income on all explanatory variables (other than county and state).

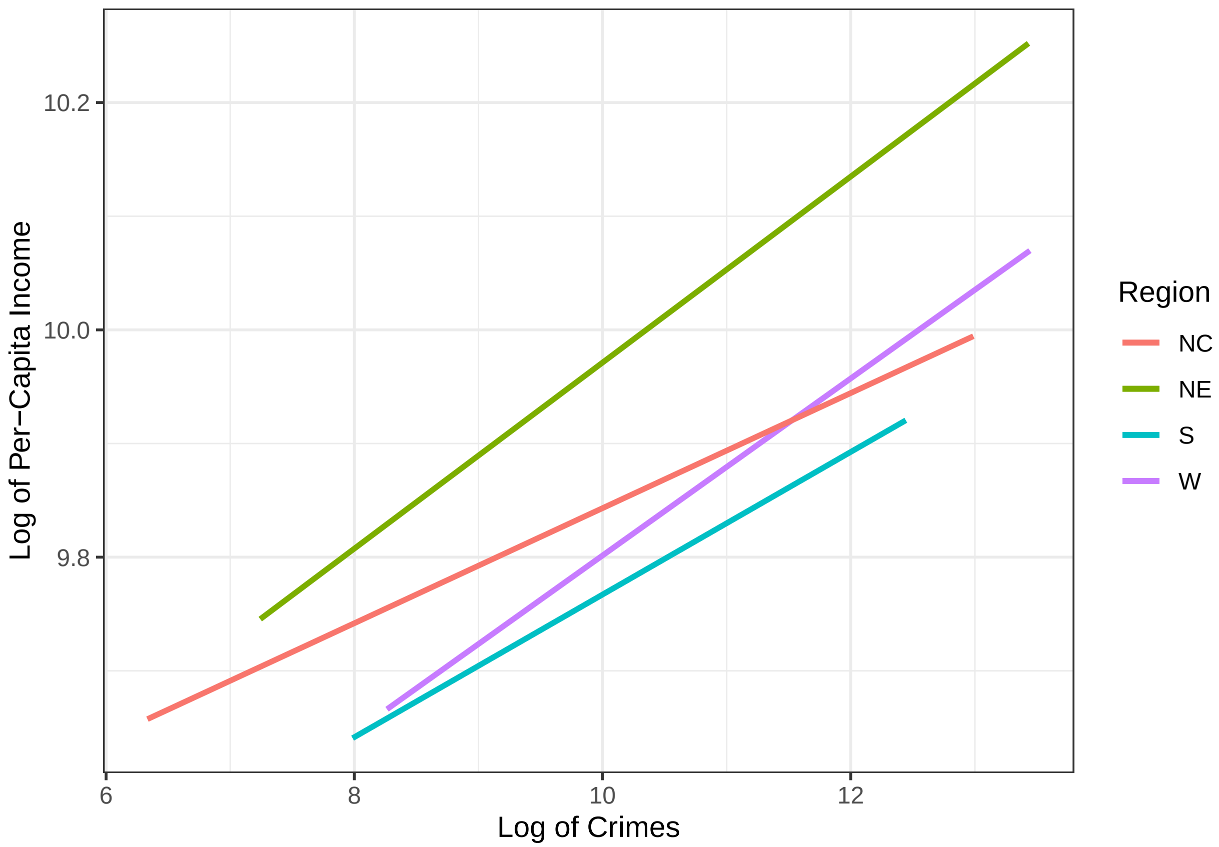
```
lm(per.cap.income ~., data=cdi[,-c(1:4,7:9,15:16)]) %>% vif %>% round(2)
```

```
##      pop.18_34    pop.65_plus    pct.hs.grad   pct.bach.deg  pct.below.pov
##           1.98           1.97           3.77           3.55           3.01
##      pct.unemp    pop.density     crime.rate doctor.per.cap   hosp.beds.pc
##           1.83           1.47           1.83           3.16           3.04
```

2. There is a theory that, if we ignore all other variables, per-capita income should be related to crime rate, and that this relationship may be different in different regions of the country (Northeast, Northcentral, South, and West). What do the data say?

Plot crime vs. per-capita income for each region

```
ggplot() +
  geom_smooth(data=subset(cdi, cdi$region=="W"),
              mapping=aes(x = log(crimes), y=log(per.cap.income), color="W"),
              method="lm", se = FALSE) +
  geom_smooth(data=subset(cdi, cdi$region=="NC"),
              mapping=aes(x=log(crimes), y=log(per.cap.income), color="NC"),
              se=FALSE, method="lm") +
  geom_smooth(data=subset(cdi, cdi$region=="S"),
              aes(x=log(crimes), y=log(per.cap.income), color="S"),
              se=FALSE, method="lm") +
  geom_smooth(data=subset(cdi, cdi$region=="NE"),
              aes(x=log(crimes), y=log(per.cap.income), color="NE"),
              se=FALSE, method="lm") +
  theme_bw() +
  labs(x="Log of Crimes", y="Log of Per-Capita Income") +
  labs(color="Region")
```



Obviously, this isn't sufficient to prove anything, but it shows there is likely a relationship between crime and per-capita income at the very least.

Check to see if any variables need to be transformed.

```
powerTransform(lm(per.cap.income~1, data=cdi))
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts
## argument ignored

## Estimated transformation parameter
##         Y1
## -0.3683365
```

```
powerTransform(lm(crimes~1, data=cdi))
```

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts
## argument ignored

## Estimated transformation parameter
##         Y1
## -0.1307109
```

It appears that logging both per capita income and crimes would be helpful here

```
crime_region_lm <- lm(log(per.cap.income) ~ log(crimes)*region, data=cdi)
summary(crime_region_lm)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(crimes) * region, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68552 -0.10418 -0.01444  0.08302  0.79755
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.33677    0.14579  64.044  < 2e-16 ***
## log(crimes)            0.05064    0.01566   3.233  0.00132 **
## regionNE              -0.18407    0.21515  -0.856  0.39272
## regionS               -0.19717    0.21211  -0.930  0.35312
## regionW               -0.31439    0.24465  -1.285  0.19947
## log(crimes):regionNE   0.03122    0.02311   1.351  0.17749
## log(crimes):regionS    0.01211    0.02228   0.544  0.58696
## log(crimes):regionW    0.02727    0.02523   1.081  0.28028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1855 on 432 degrees of freedom
## Multiple R-squared:  0.2073, Adjusted R-squared:  0.1945
## F-statistic: 16.14 on 7 and 432 DF,  p-value: < 2.2e-16
```

We find that there is a statistically significant effect of crime rate on per capita income. However, the relationship between crime rate and per capita income does not appear to vary by any region specifically. However, even this is not enough since we are really looking at all of the interactions at once. Thus we will need an F test to determine if the relationship between crime and per-capita income varies by region.

```
crime_lm <- lm(log(per.cap.income) ~ log(crimes), data=cdi)
anova(crime_region_lm, crime_lm)
```

```
## Analysis of Variance Table
##
## Model 1: log(per.cap.income) ~ log(crimes) * region
## Model 2: log(per.cap.income) ~ log(crimes)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    432 14.872
## 2    438 17.271 -6   -2.3987 11.613 4.538e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So even though individually, none of the interactions appeared statistically significant, there is very strong evidence suggesting that region does affect the relationship between crime and per-capita income.

3. Find the best model predicting per-capita income from the other variables (including possible transformations, interactions, etc.). Here "best" means a good compromise between

- Best reflects the social science and the meaning of the variables
- Best satisfies modeling assumptions
- Is most clearly indicated by the data
- Can be explained to someone who is more interested in social, economic and health factors than in mathematics and statistics.

Let's now determine what other transformations could be good by power transforming the other variables. We will do this for all continuous variables we have created up until now.

```
cdi.tf <- cdi[,-c(1,2,16)] # Variables dropped: county, state, pop,

pt <- data.frame(matrix(nrow=17, ncol=3, data=NA))
colnames(pt) <- c("Variable", "Suggested_Transformation", "Rounded_Transformation")
pt$Variable <- colnames(cdi.tf)[1:17]

for(i in 1:17){
  suppressWarnings(pt$Suggested_Transformation[i] <- round((powerTransform(lm(cdi.tf[,i]~1))$lambda),2)
}

pt$Rounded_Transformation <- c("log", -0.5, -0.5, rep("log", 4), 3,
                               rep("log", 3), -0.5, -0.5, "log", -0.5, rep("log", 2))
pt
```

```
##           Variable Suggested_Transformation Rounded_Transformation
## 1        land.area                     0.00                    log
## 2              pop                    -0.58                   -0.5
## 3        pop.18_34                    -0.39                   -0.5
## 4      pop.65_plus                    -0.01                    log
## 5          doctors                    -0.22                    log
## 6        hosp.beds                    -0.15                    log
## 7           crimes                    -0.13                    log
## 8      pct.hs.grad                     3.07                      3
## 9     pct.bach.deg                    -0.03                    log
## 10   pct.below.pov                     0.18                    log
## 11       pct.unemp                    -0.11                    log
## 12 per.cap.income                    -0.37                   -0.5
## 13      tot.income                    -0.44                   -0.5
## 14     pop.density                    -0.10                    log
## 15      crime.rate                     0.38                   -0.5
## 16 doctor.per.cap                    -0.23                    log
## 17    hosp.beds.pc                     0.23                    log
```

Now, there would be some problems with interpretations even if we just used the rounded transformations. For example, it is difficult to interpret coefficient on the log of the percentage of population with a bachelors degree or high school degree. It is also difficult to explain a cubic relationship. Further, we have variables that are aliased that we will want to remove before starting a regression. This includes two versions of the same variable as well as variables that are equal to a proportion of other variables (e.g. per-capita income = total income / population). Since there are a couple choices for which variable to remove, we will create two models: one where we don't divide the newly-created variables by population (model 1) and one where we do (model 2). It also not very interesting to use one measure of income (total income) to predict another measure of income (per capita income), so we will remove total income in both cases.

Our first two models then becomes:

```
lm1 <- lm(log(per.cap.income) ~ log(land.area) + pop.18_34 + pop.65_plus +
            log(doctors) + log(hosp.beds) + log(crimes) + pct.hs.grad +
            pct.bach.deg + pct.below.pov + pct.unemp + region, data=cdi)

lm2 <- lm(log(per.cap.income) ~ log(pop.density) + pop.18_34 + pop.65_plus +
            doctor.per.cap + hosp.beds.pc + crime.rate + pct.hs.grad + pct.bach.deg +
            pct.below.pov + pct.unemp + region, data=cdi)
```

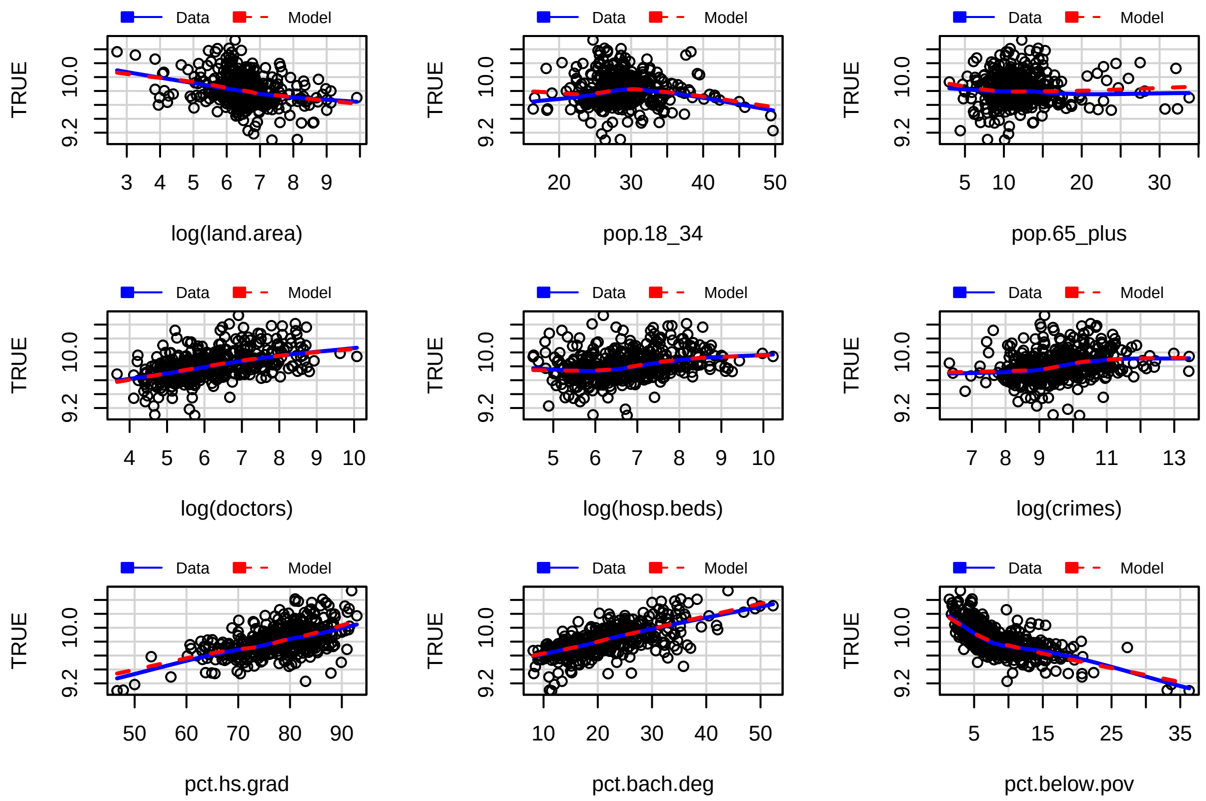
Let's compare the two models, starting with the first.

```
vif(lm1)
```

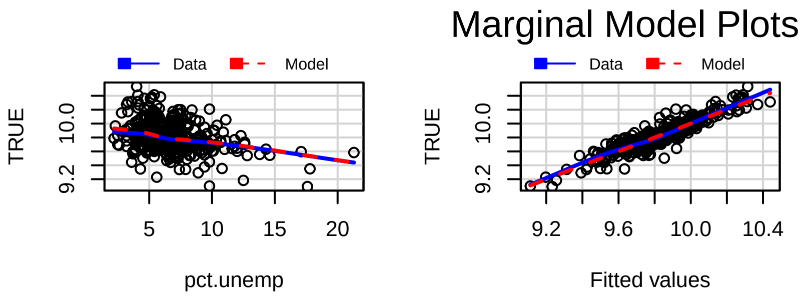
```
##                       GVIF Df GVIF^(1/(2*Df))
## log(land.area)  1.484666  1        1.218469
## pop.18_34       1.997041  1        1.413167
## pop.65_plus     2.053608  1        1.433041
## log(doctors)   15.046783  1        3.879018
## log(hosp.beds) 11.403602  1        3.376922
## log(crimes)     6.238869  1        2.497773
## pct.hs.grad     4.495620  1        2.120288
## pct.bach.deg    4.353907  1        2.086602
## pct.below.pov   2.878962  1        1.696750
## pct.unemp       2.059659  1        1.435151
## region          3.580885  3        1.236892
```

The vifs for lm1 are high for doctors, hospital beds, and crimes because of how closely related they all are.

```
mmps(lm1)
```

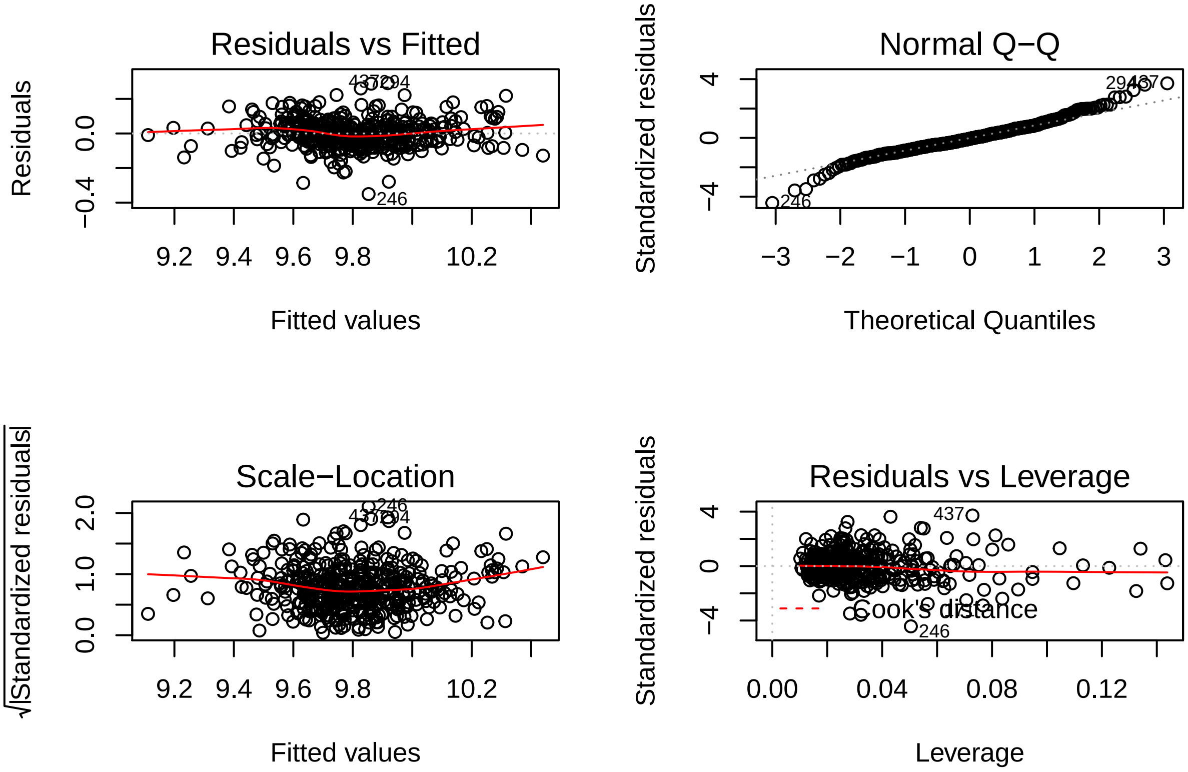


```
## Warning in mmps(lm1): Interactions and/or factors skipped
```



The marginal model plots for lm1 are fine. Let's look at the residual diagnostic plots

```r
par(mfrow = c(2,2))
plot(lm1)
```



Observations 294 and 437 look a bit off, but other than that, it seems fine. We'll return to determining whether or not we should remove any of the observations after we determine which model is better.

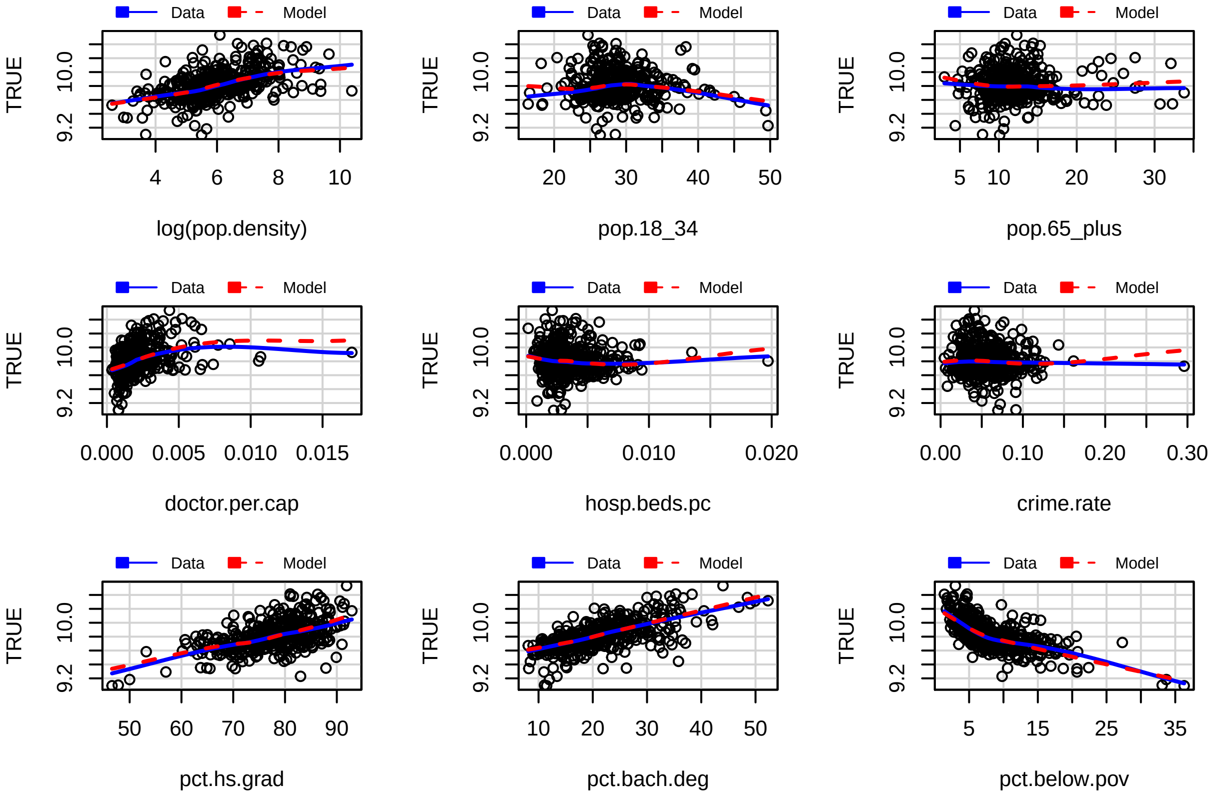
Now let's look at lm2.

```
vif(lm2)
```

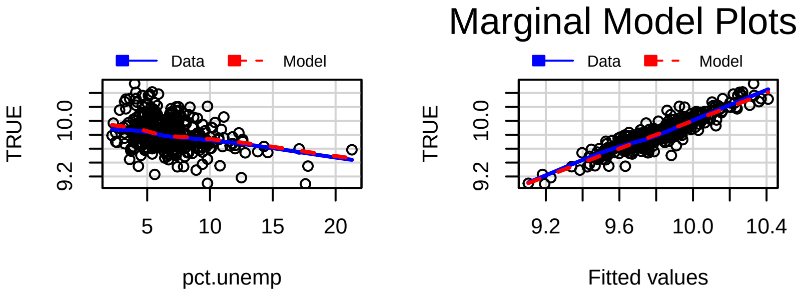
```
##                      GVIF Df GVIF^(1/(2*Df))
## log(pop.density) 1.816898  1        1.347924
## pop.18_34        2.027666  1        1.423961
## pop.65_plus      2.042058  1        1.429006
## doctor.per.cap   3.347513  1        1.829621
## hosp.beds.pc     3.546933  1        1.883330
## crime.rate       2.102241  1        1.449911
## pct.hs.grad      4.561990  1        2.135882
## pct.bach.deg     4.009568  1        2.002391
## pct.below.pov    3.474906  1        1.864110
## pct.unemp        2.077610  1        1.441392
## region           2.785499  3        1.186180
```

The VIFs don't look nearly as bad as in model 1 now that we have normalized the variables.

```
mmps(lm2)
```

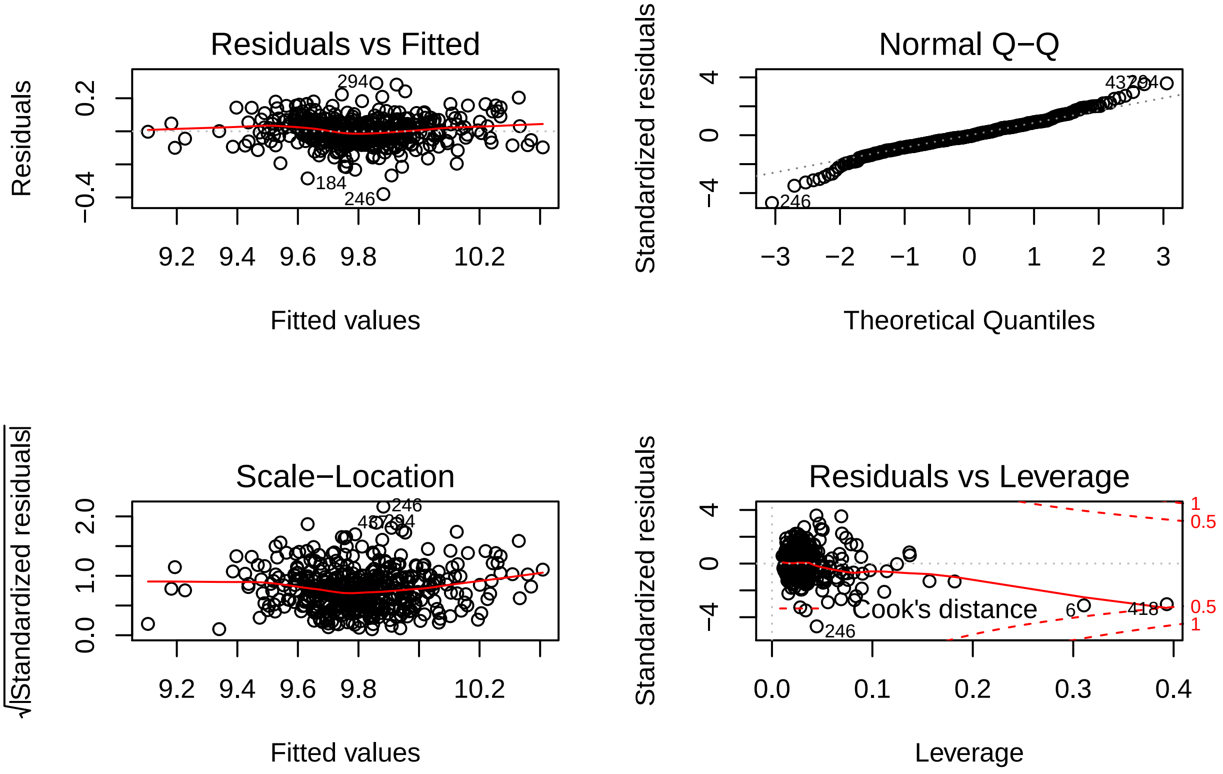


```
## Warning in mmps(lm2): Interactions and/or factors skipped
```



The marginal model plots still look fine.

```r
par(mfrow = c(2,2))
plot(lm2)
```



The residuals look about as good as in lm1. Note that the distribution of the error terms is not normal, but that is likely due to the categorical variables we have in the model. Let's also compare the xIC and adj R squared.

```
AICs = c(AIC(lm1), AIC(lm2))
BICs = c(BIC(lm1), BIC(lm2))
adjrsqs = c(summary(lm1)$adj.r.squared, summary(lm2)$adj.r.squared)
data.frame(model = 1:2, AIC=AICs, BIC=BICs, adj.r.sq = adjrsqs)
```

```
##   model       AIC       BIC  adj.r.sq
## 1     1 -944.7011 -883.3995 0.8455840
## 2     2 -926.5291 -865.2274 0.8390731
```

It appears that lm1 has high VIFs but lower xICs and adjusted r squared than model 2. Since the coefficients of lm1 are biased becuase of the high levels of collinearity, and the xIC and adjusted R squared of model 2 is still good, let's continue with model 2 for now.

Let's see if we can add some meaningful interactions that would help increased adjusted r squared, but also make sense to social scientists. A few that may be interesting are:

- Region * unemployment: does the relationship between unemployment and per-capita income vary by region?
- Region * hs.grad: does the relationship between schooling and per-capita income vary by region?

Let's test some of these out with model 3.

```
lm3 <- lm(log(per.cap.income) ~ log(pop.density) + pop.18_34 + pop.65_plus +
            doctor.per.cap + hosp.beds.pc + crime.rate + pct.bach.deg +
            pct.below.pov + region*pct.unemp + region*pct.hs.grad, data=cdi)

summary(lm3)
```

```
## 
## Call:
## lm(formula = log(per.cap.income) ~ log(pop.density) + pop.18_34 + 
##     pop.65_plus + doctor.per.cap + hosp.beds.pc + crime.rate + 
##     pct.bach.deg + pct.below.pov + region * pct.unemp + region * 
##     pct.hs.grad, data = cdi)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.33655 -0.04428 -0.00222  0.04389  0.30266 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           9.947906   0.207493  47.943  < 2e-16 ***
## log(pop.density)      0.045727   0.004682   9.766  < 2e-16 ***
## pop.18_34            -0.015730   0.001338 -11.754  < 2e-16 ***
## pop.65_plus          -0.001599   0.001428  -1.120 0.263421    
## doctor.per.cap        5.314435   4.661966   1.140 0.254954    
## hosp.beds.pc          9.807381   3.740140   2.622 0.009054 ** 
## crime.rate            0.698621   0.207195   3.372 0.000816 ***
## pct.bach.deg          0.017396   0.001041  16.715  < 2e-16 ***
## pct.below.pov        -0.024905   0.001631 -15.267  < 2e-16 ***
## regionNE             -0.127846   0.228166  -0.560 0.575561    
## regionS               0.222759   0.202721   1.099 0.272464    
## regionW               0.874110   0.271044   3.225 0.001358 ** 
## pct.unemp             0.018666   0.004963   3.761 0.000193 ***
## pct.hs.grad          -0.003850   0.002216  -1.737 0.083042 .  
## regionNE:pct.unemp   -0.013628   0.007284  -1.871 0.062059 .  
## regionS:pct.unemp    -0.015738   0.006587  -2.389 0.017316 *  
## regionW:pct.unemp    -0.021384   0.006893  -3.102 0.002051 ** 
## regionNE:pct.hs.grad  0.003026   0.002488   1.216 0.224535    
## regionS:pct.hs.grad  -0.002017   0.002226  -0.906 0.365295    
## regionW:pct.hs.grad  -0.009001   0.002937  -3.064 0.002322 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.08045 on 420 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8486 
## F-statistic: 130.5 on 19 and 420 DF,  p-value: < 2.2e-16
```

We find that there are significant interaction effect(s) between the proposed interactions, though not necessarily for all factors (regions). Since they are significant for at least one interaction, we will keep all the interactions in.

Personally, I think it's a bit confusing to have both the doctors variable and the number of hospital beds in the county. Plus, once accounting for the number of hospital beds, the doctors variable isn't statistically significant anyway. Also, the population age 65+ isn't statistically significant. Let's see what happens when we throw the doctors variable and the age 65+ variables away.

```
lm4 <- lm(log(per.cap.income) ~ log(pop.density) + pop.18_34 +
            hosp.beds.pc + crime.rate + pct.bach.deg + pct.below.pov +
            region*pct.unemp + region*pct.hs.grad, data=cdi)

summary(lm4)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(pop.density) + pop.18_34 +
##     hosp.beds.pc + crime.rate + pct.bach.deg + pct.below.pov +
##     region * pct.unemp + region * pct.hs.grad, data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33065 -0.04736 -0.00193  0.04359  0.29609
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            9.8539581  0.1971265  49.988  < 2e-16 ***
## log(pop.density)       0.0469743  0.0045918  10.230  < 2e-16 ***
## pop.18_34             -0.0148611  0.0011420 -13.013  < 2e-16 ***
## hosp.beds.pc          11.6169163  2.4531312   4.736 2.99e-06 ***
## crime.rate             0.6958819  0.2073343   3.356 0.000861 ***
## pct.bach.deg           0.0178729  0.0009613  18.592  < 2e-16 ***
## pct.below.pov         -0.0246904  0.0015744 -15.683  < 2e-16 ***
## regionNE              -0.1012159  0.2276768  -0.445 0.656866
## regionS                0.2780265  0.1997340   1.392 0.164660
## regionW                0.9157506  0.2690727   3.403 0.000729 ***
## pct.unemp              0.0191977  0.0049547   3.875 0.000124 ***
## pct.hs.grad           -0.0034443  0.0021969  -1.568 0.117670
## regionNE:pct.unemp    -0.0147966  0.0072358  -2.045 0.041484 *
## regionS:pct.unemp     -0.0173322  0.0064093  -2.704 0.007123 **
## regionW:pct.unemp     -0.0218397  0.0068908  -3.169 0.001639 **
## regionNE:pct.hs.grad   0.0027706  0.0024838   1.115 0.265292
## regionS:pct.hs.grad   -0.0026153  0.0021956  -1.191 0.234245
## regionW:pct.hs.grad   -0.0094566  0.0029173  -3.242 0.001283 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08051 on 422 degrees of freedom
## Multiple R-squared:  0.8542, Adjusted R-squared:  0.8483
## F-statistic: 145.5 on 17 and 422 DF,  p-value: < 2.2e-16
```

Let's now compare models 3 and 4.

```
AICs = c(AIC(lm3), AIC(lm4))
BICs = c(BIC(lm3), BIC(lm4))
adjrsqs = c(summary(lm3)$adj.r.squared, summary(lm4)$adj.r.squared)
data.frame(model = 3:4, AIC=AICs, BIC=BICs, adj.r.sq = adjrsqs)
```

```
##   model       AIC       BIC  adj.r.sq
## 1     3 -947.5293 -861.7071 0.8485684
## 2     4 -948.8011 -871.1524 0.8483487
```

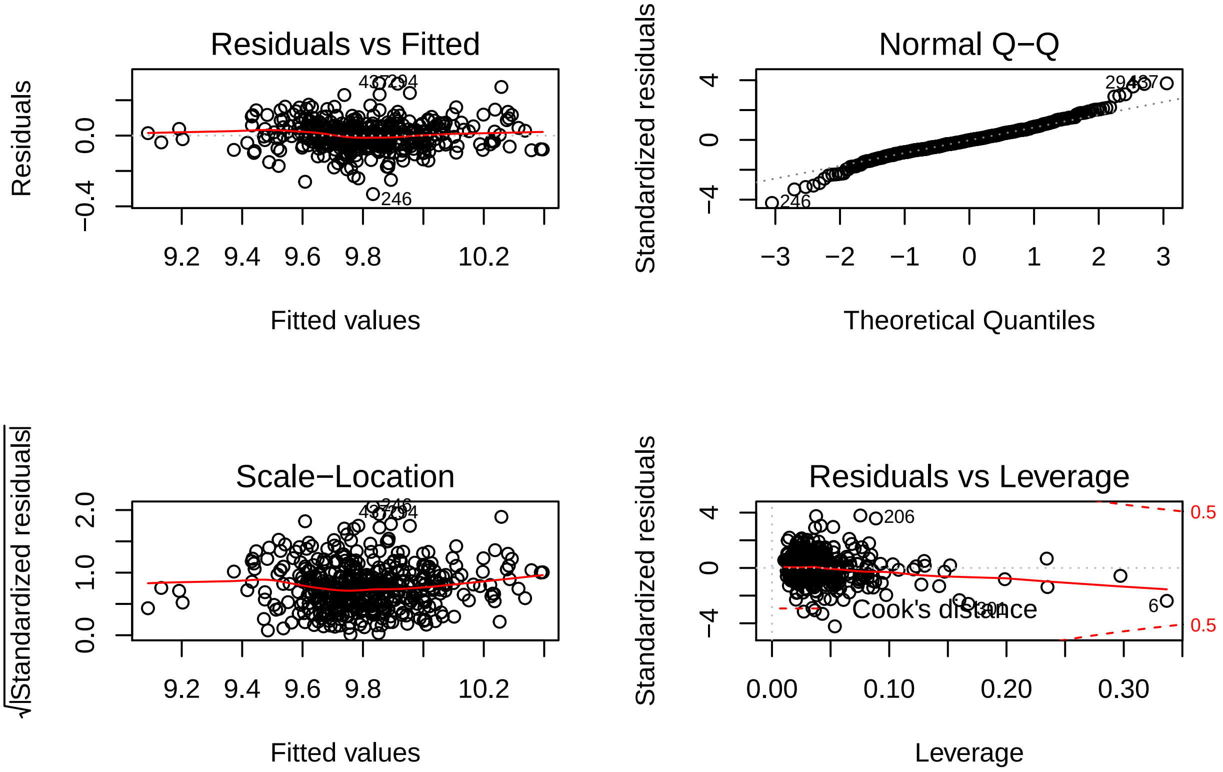
Model 4 appears to outperform model 3 (or at least be approximately equal to it in the case of adjusted r squared). It would also be interesting to see how stepwise regression simplifies the model.

```
step.model <- stepAIC(lm3, direction = "both",
                      trace = FALSE)

summary(step.model)
```

```
##
## Call:
## lm(formula = log(per.cap.income) ~ log(pop.density) + pop.18_34 +
##     hosp.beds.pc + crime.rate + pct.bach.deg + pct.below.pov +
##     region + pct.unemp + pct.hs.grad + region:pct.unemp + region:pct.hs.grad,
##     data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33065 -0.04736 -0.00193  0.04359  0.29609
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.8539581  0.1971265  49.988  < 2e-16 ***
## log(pop.density)      0.0469743  0.0045918  10.230  < 2e-16 ***
## pop.18_34            -0.0148611  0.0011420 -13.013  < 2e-16 ***
## hosp.beds.pc         11.6169163  2.4531312   4.736 2.99e-06 ***
## crime.rate            0.6958819  0.2073343   3.356 0.000861 ***
## pct.bach.deg          0.0178729  0.0009613  18.592  < 2e-16 ***
## pct.below.pov        -0.0246904  0.0015744 -15.683  < 2e-16 ***
## regionNE             -0.1012159  0.2276768  -0.445 0.656866
## regionS               0.2780265  0.1997340   1.392 0.164660
## regionW               0.9157506  0.2690727   3.403 0.000729 ***
## pct.unemp             0.0191977  0.0049547   3.875 0.000124 ***
## pct.hs.grad          -0.0034443  0.0021969  -1.568 0.117670
## regionNE:pct.unemp   -0.0147966  0.0072358  -2.045 0.041484 *
## regionS:pct.unemp    -0.0173322  0.0064093  -2.704 0.007123 **
## regionW:pct.unemp    -0.0218397  0.0068908  -3.169 0.001639 **
## regionNE:pct.hs.grad  0.0027706  0.0024838   1.115 0.265292
## regionS:pct.hs.grad  -0.0026153  0.0021956  -1.191 0.234245
## regionW:pct.hs.grad  -0.0094566  0.0029173  -3.242 0.001283 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08051 on 422 degrees of freedom
## Multiple R-squared:  0.8542, Adjusted R-squared:  0.8483
## F-statistic: 145.5 on 17 and 422 DF,  p-value: < 2.2e-16
```

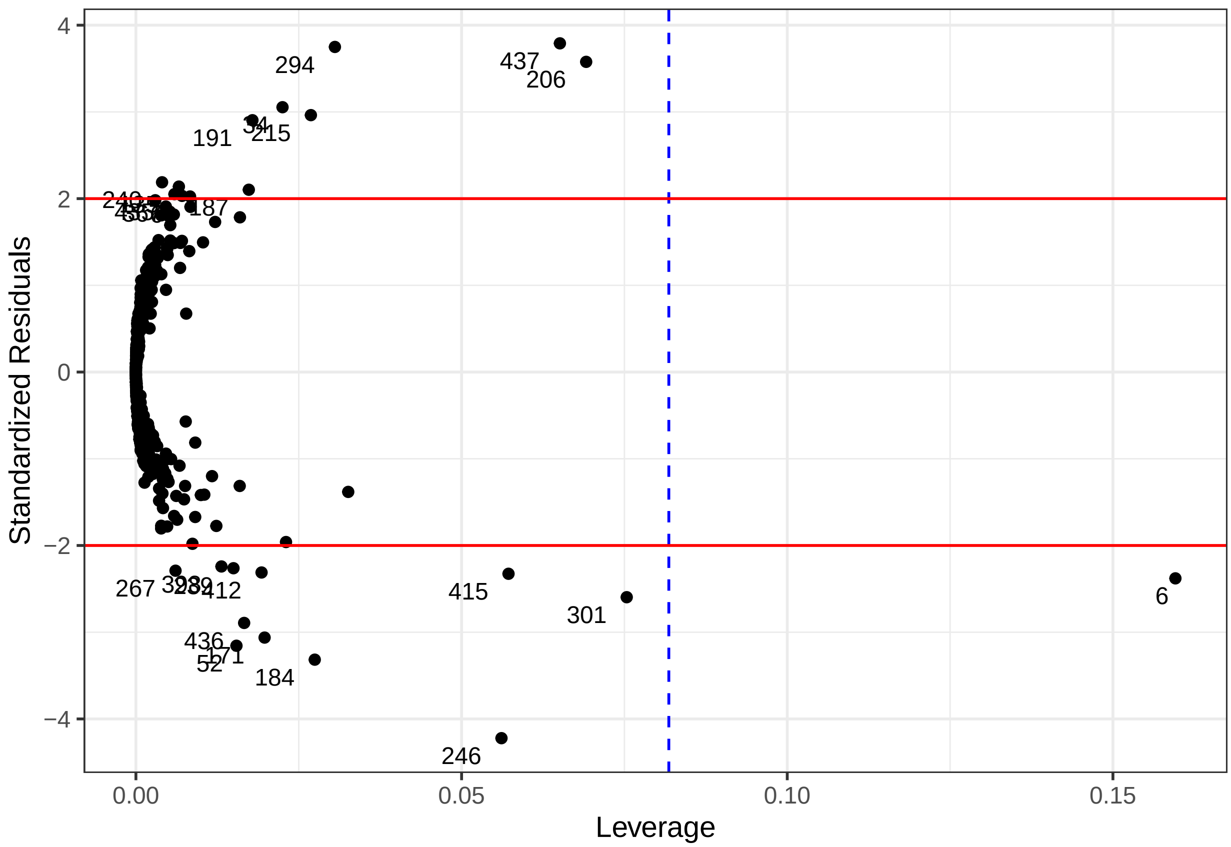
Interestingly, it is the same as lm4! So far, lm4 seems to be a great model. Let's make sure it satisfies our assumptions.

```
par(mfrow=c(2,2))
plot(lm4)
```



The residuals plots look good, but it's hard to tell if there are any bad leverage points. Let's draw our own graph based off of Sheather's suggestions for leverage cutoffs.

```
lev4 <- cooks.distance(lm4)
lev.cutoff <- 2*(17+1)/nrow(cdi) # Sheather suggests the cutoff for the leverage points to be 2*(p+1)/n

res4 <- stdres(lm4)
reslev4 <- data.frame(lev=lev4, res=res4, row=rownames(cdi))


ggplot(reslev4, aes(lev, res)) + geom_point() +
  geom_text(data=subset(reslev4, reslev4$lev > lev.cutoff | abs(reslev4$res) > 2) , aes(x=lev, y=res, la
  geom_vline(xintercept = lev.cutoff, linetype="dashed", col="blue") +
  geom_hline(yintercept = 2, col="red") +
  geom_hline(yintercept = -2, col="red") +
  labs(x="Leverage", y="Standardized Residuals")+
  theme_bw()
```



It appears that we have one bad leverage point: row 6, or Kings, NY. This is an outlier since the land area is only 71 square miles but it has a huge population of over 32,000. Let's create a model that does not have that observation and compare the two models.

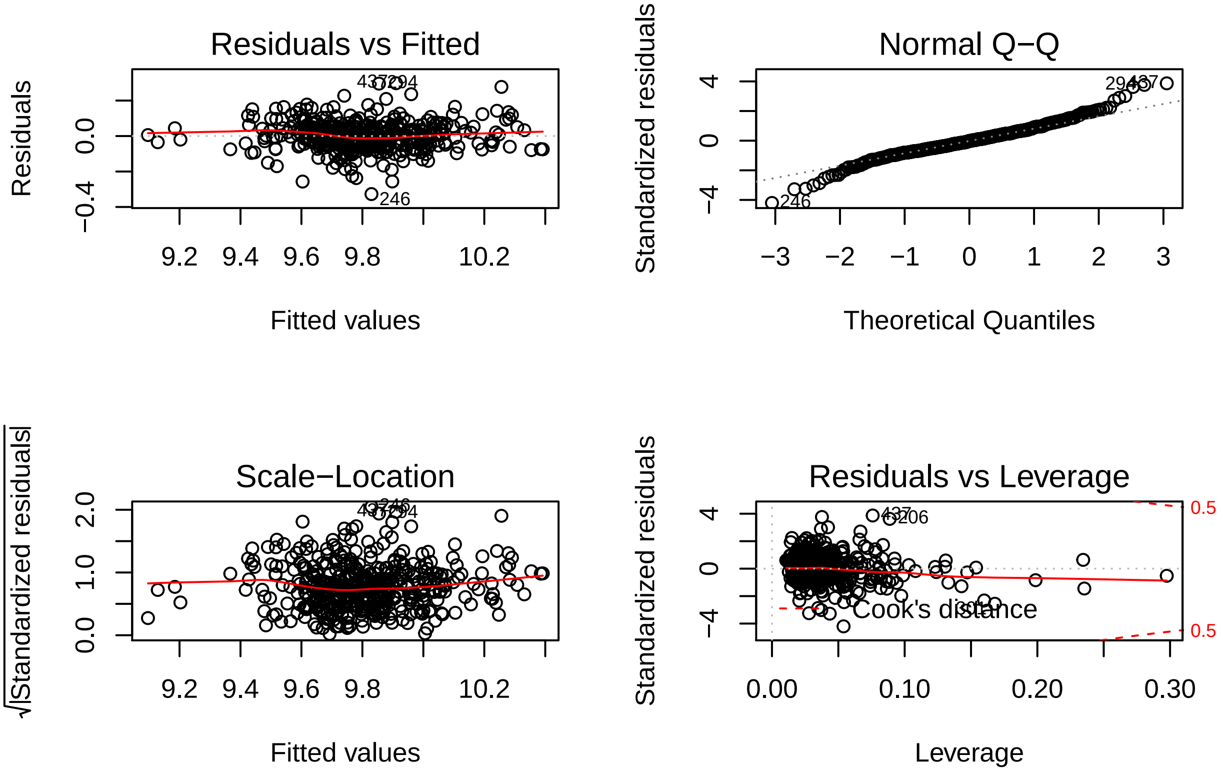
```
lm5 <- lm(log(per.cap.income) ~ log(pop.density) + pop.18_34 + hosp.beds.pc + crime.rate + pct.bach.deg

AICs = c(AIC(lm4), AIC(lm5))
BICs = c(BIC(lm4), BIC(lm5))
adjrsqs = c(summary(lm4)$adj.r.squared, summary(lm5)$adj.r.squared)
data.frame(model = 4:5, AIC=AICs, BIC=BICs, adj.r.sq = adjrsqs)
```

```
##   model       AIC       BIC  adj.r.sq
## 1     4 -948.8011 -871.1524 0.8483487
## 2     5 -951.4868 -873.8814 0.8503205
```

The model has now improved in xIC and r squared, and quite significantly too for just dropping one bad observation.

Let's test this model to make sure it passes all the tests.

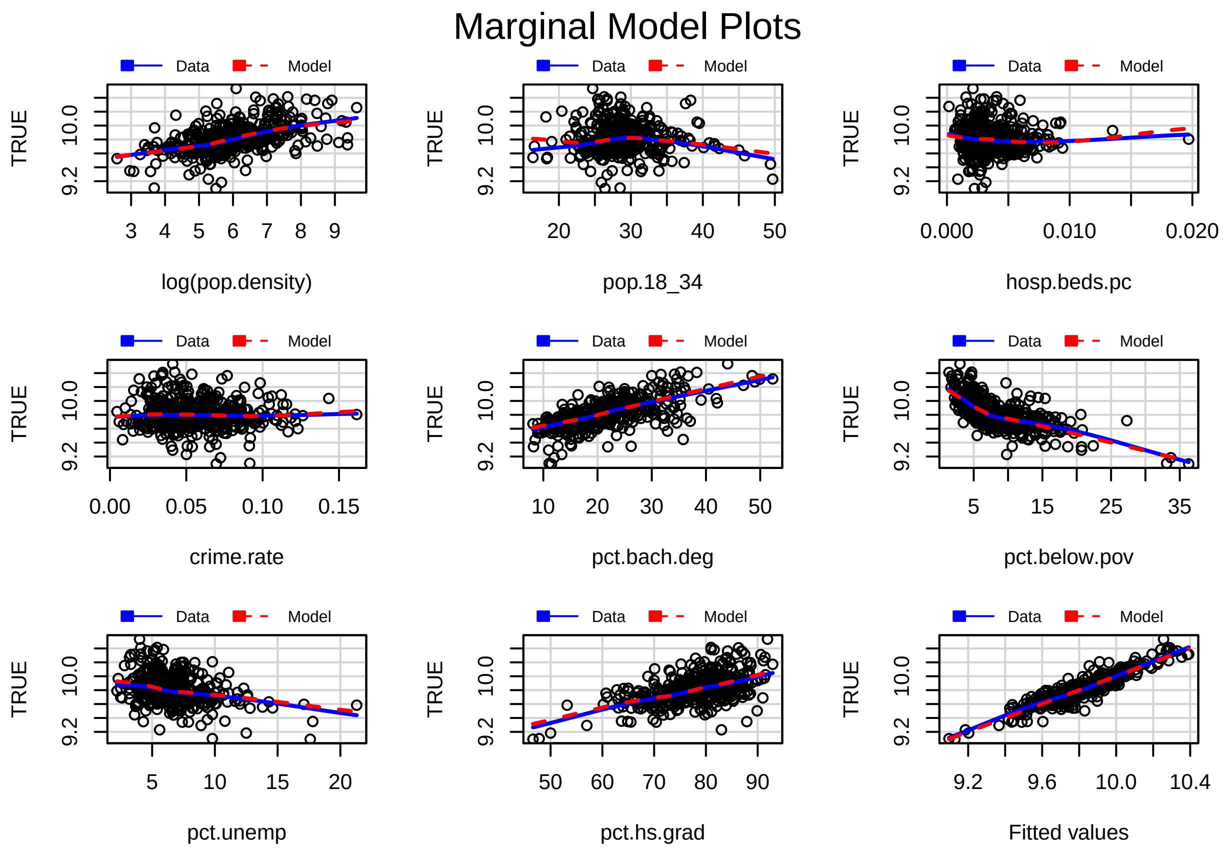
```
par(mfrow=c(2,2))
plot(lm5)
```



Again, there's not too much we can do about the normality because of the categorical variables, but everythiing else looks fine.

```
mmps(lm5)
```

```
## Warning in mmps(lm5): Interactions and/or factors skipped
```



The marginal model plots fit.

Let's now perform k-fold cross validation to test the predictive power of lm5.

```
set.seed(12)
dataset <- cdi[-6, ] #Create data frame


#install.packages("cvTools")
library(cvTools) #run the above line if you don't have this library
```

```
## Loading required package: lattice

## Loading required package: robustbase
```

```
k <- 5 #the number of folds

folds <- cvFolds(NROW(dataset), K=k)
mse <- rep(NA,5)
percent.off <- rep(NA,5)
for(i in 1:k){
  train <- dataset[folds$subsets[folds$which != i], ] #Set the training set
  validation <- dataset[folds$subsets[folds$which == i], ] #Set the validation set

  newlm <- lm(log(per.cap.income) ~ log(pop.density) + pop.18_34 +
                hosp.beds.pc + crime.rate + pct.bach.deg + pct.below.pov +
                region*pct.unemp + region*pct.hs.grad, data=train)
 #Get your new linear model (just fit on the train data)
 preds <- exp(predict(newlm,newdata=validation)) #Get the predicitons for the validation set (from the
 actual <- validation$per.cap.income
 mse[i] <- (actual-preds)^2 %>% sum %>% mean

 percent.off[i] <- mean(abs(preds-actual)/actual)*100

}


mean(mse)
```

```
## [1] 231943809
```

```
mean(percent.off)
```

```
## [1] 6.114681
```

This model (lm9) does a good job at prediction. It is off by about 6.1% on average.

4. A county is a governmental unit in the United States that is bigger than a city but smaller than a state. There are 50 states in the US, plus the District of Columbia, which is usually coded as a 51st state in data like this. There are 48 states represented in the data. There are approximately 3000 counties in the US, and 373 represented in the data set. Should we be worried about either the missing states or the missing counties? Why or why not?

See discussion section.

The following table looks at how many counties were represented by state. Note that not all states are included.

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

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