One for the Money: Understanding How Demographics Impacts Personal Income in American Counties

Caleb Pena, cpea@andrew.cmu.edu

October 30, 2021

Abstract

This paper analyzes the effect of demographic factors such as age, education, and crime rates on personal income. We use data from a 1990 report from the University of Virginia's Geospatial and Statistical Data Center. We fit several multiple regression models using advanced variable selection techniques and compared the results. Lasso regression turned out to be a particularly useful tool. We found that crime rates are not a significant predictor of income after controlling for population. In addition, our models provide evidence of the importance of the land area in predicting a county's income level. However, our analysis is limited by the fact the data only comes from relatively urban counties. Further study is needed to understand how these relationships play out in rural settings.

Introduction

Identifying and interpreting key indicators of economic health is an important task for policy makers. Personal income is an especially important metric since it serves as a useful proxy for standard of living. Understanding what demographic factors influence per capita income gives economists and other analysts a better idea of what kind of policy changes can help improve quality of life.

In this paper we will analyze historical data from 1990 to better understand how variables such as poverty, unemployment, or education impact income level. For this exercise we have limited our toolset to methods discussed in Sheather (2009). In particular, we have been asked to focus our research by answering the following four questions:

- Which pairs of variables are clearly related and what is the nature of those relationships?
- Is a county's crime rate associated with its per capita income? Does this association vary regionally?
- What is the best model for predicting per capita income?
- Has missing data skewed the dataset and our subsequent analysis?

Data

We study these questions using data collected by the Geospatial and Statistical Data Center of the University of Virginia as cited in Kutner et. al. (2013). The dataset includes 3 categorical and 13 numeric variables measuring demographic features of 440 of the most populous counties in the US. The data only tracks counties with at least 100,000 residents; the rest are omitted. Because of this requirement, three states are unrepresented. The two states with the lowest population density (US Census Bureau, 2021), Alaska and Wyoming, do not make the cut. Suprisingly, Iowa is also missing. Although it is near the middle of state population rankings, it spreads its population fairly evenly across 99 counties. It also has no major cities so none of its counties meet the strict requirement of our dataset.

Table 1 provides a brief summary of what kind of information was tracked.

Table 1: Description of Data

Variable Name	Description
county	Name of county
state	Name of state
land_area	Area in square miles
pop	Estimated 1990 population
pop_18_34	Percent of population aged 18-34
pop_65_plus	Percent of population aged 65+
doctors	Number of active physicians in 1990
hosp_beds	Number of hospital beds
crimes	Number of serious crimes
pct_hs_grad	Percent of adult population who graduated high school
pct_bach_deg	Percent of adult population with bachelor's degrees
pct_below_pov	Percent of population living below the poverty line
pct_unemp	Percent of population that is unemployed
per_cap_income	Per capita income in dollars
tot_income	Total personal income in dollars
region	Geographic region (S, W, NE, NC)

These variables exist on widely different scales. Some, like the percentage variables, are bounded between 0 and 100. Others, like the population variable and its close correlates, have ranges in the tens of thousands. Additional summary information can be found in Table 2 below. Of particular note is that the means of the variables `crimes`, `doctors`, `hosp_beds`, `land_area`, `pop`, and `tot_income` are all substantially higher than their respective medians. This is indicative of right-skewing that could potentially affect our analysis. We address this concern more in the next section. A deeper breakdown of the data including scatterplots and histograms can be found in part A of the technical appendix (pp. 2-5).

Table 2: Numerical Summaries

Variable	Mean	Median	SD	Minimum	Maximum
crimes	27111.62	11820.50	58237.51	563.0	688936.0
doctors	988.00	401.00	1789.75	39.0	23677.0
hosp_beds	1458.63	755.00	2289.13	92.0	27700.0
land_area	1041.41	656.50	1549.92	15.0	20062.0
pct_bach_deg	21.08	19.70	7.65	8.1	52.3
pct_below_pov	8.72	7.90	4.66	1.4	36.3
pct_hs_grad	77.56	77.70	7.02	46.6	92.9
pct_unemp	6.60	6.20	2.34	2.2	21.3
per_cap_income	18561.48	17759.00	4059.19	8899.0	37541.0
pop	393010.92	217280.50	601987.02	100043.0	8863164.0
pop_18_34	28.57	28.10	4.19	16.4	49.7
pop_65_plus	12.17	11.75	3.99	3.0	33.8
tot_income	7869.27	3857.00	12884.32	1141.0	184230.0

Methods

We broke our analysis into four subsections related to each of the four questions we were asked to address. The first was to explore and describe key pairwise relationships amongst the variables. To do this we constructed a heatmap of the Pearson correlation coefficients. The stronger pairs were then investigated visually using scatterplots with smoothing curves.

Next we were asked to identify whether or not crime rates have a direct impact on per capita income in different regions of the country. To answer this we built a multiple linear model with and without interaction terms between the categorical region variable and crime rate. This was attempted using both the total number of crimes and the per capita crime rate. We also visually inspected these regression

lines using the R package ggplot2.

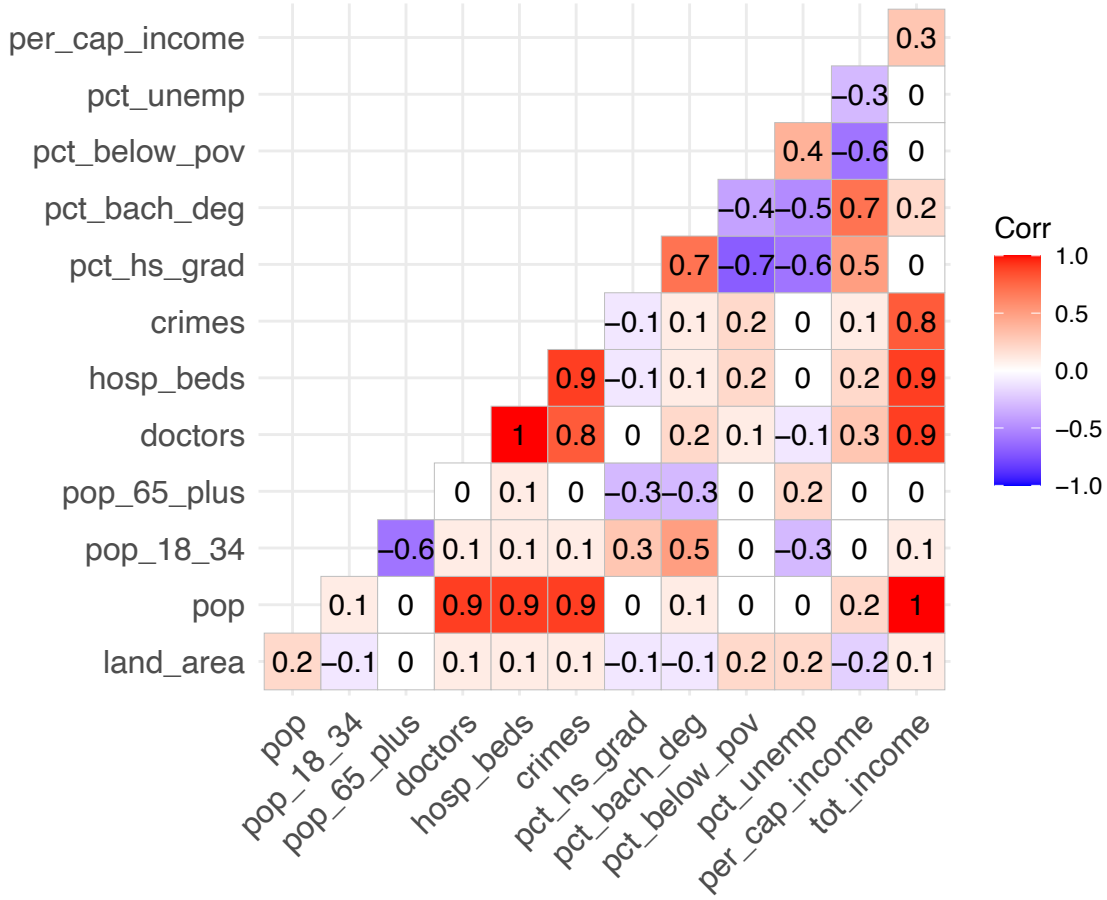
Third, we incorporated the remaining variables into our analysis and compared multiple models to find the one that best fit the shape of the data. We used stepwise regression and lasso regularization to select the most predictive variables. Then we compared these models using a combination of diagnostic plots, measures of fit, and tools to identify multi-colinearity. The code for this section may be found in Part C of the technical appendix (pp 10-13).

Finally, to determine how representative the data is we considered the mechanism that left certain counties out of the dataset. We used the MCAR, MAR, and MNAR paradigm discussed in van Buuren (2018).

Results

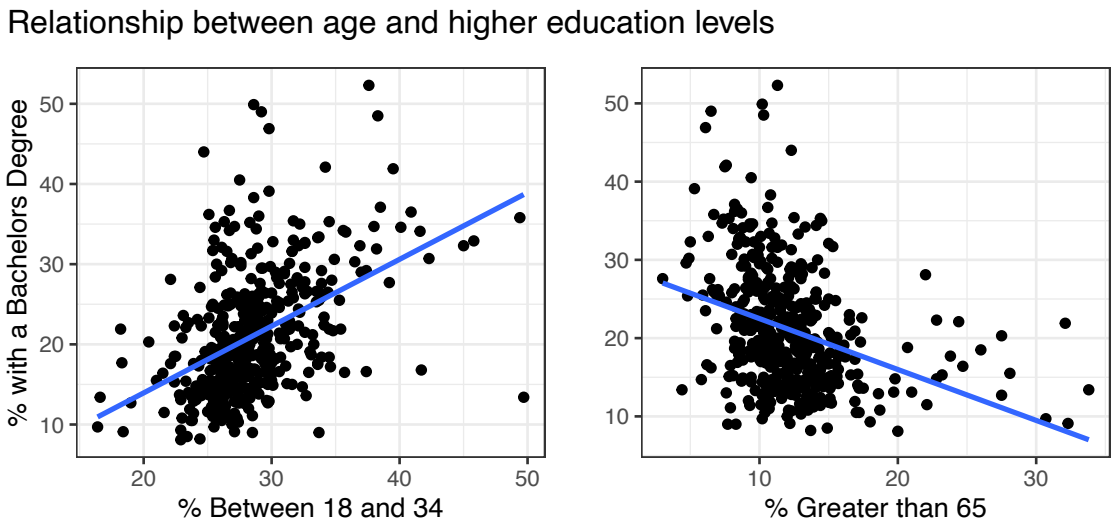
Pairwise relationships

Figure 1 shows the pairwise Pearson correlations between the variables. Most of these relationships are unsurprising. Total income has a positive relationship with total crimes, total hospital beds, and total doctors since these measures are primarily determined by total population. In addition, counties that are more educated score higher in both percentage of high school graduates and percent with a bachelor's degree. And higher per capita income is associated with a lower unemployment and poverty rate.



(Figure 1)

A less obvious result is the relationship between age and education. There is a strong linear relationship between the percentage of the population that is between 18 and 34 and the percent with a bachelor's degree. The opposite relationship holds for the percent of the population greater than 65. These relationships are plotted in Figure 2. Interestingly, despite the strong relationship between education and economic health, neither of the age variables show are closely linearly related to per capita income or poverty.



(Figure 2)

Modeling using just crime rate

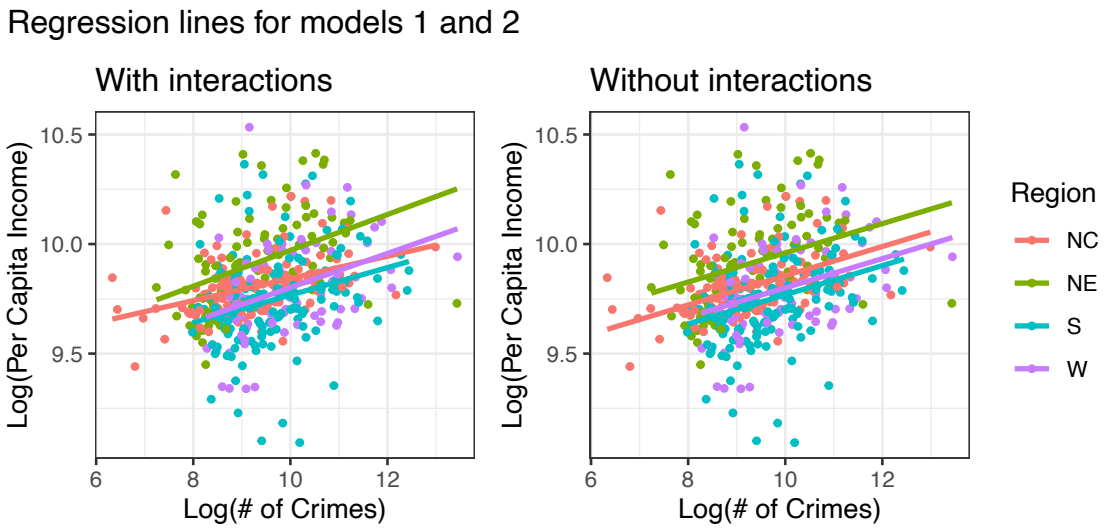
As is often the case with variables closely related to population, both the total crimes and per capita income variables are heavily right-skewed. To address this we took their logarithms and fit the following model:

$$log(PerCapitaIncome) = log(Crimes) + Region + error \quad (1)$$

Model 1 assumes that a change in the crime rate has the same impact regardless of region. We relaxed this assumption for model 2 by adding interaction terms:

$$log(PerCapitaIncome) = log(Crimes) * Region + error \quad (2)$$

The resulting regression lines can be seen in Figure 3. Modeling regional variation does appear to be important but the slopes are all very similar. The model with the interaction terms does not appear to add any meaninful information to the model. This was confirmed when we conducted a series of t-tests on the coefficients of the interaction terms. None of them had values that were significantly different from zero. Thus, we maintain that model 1 better captures the relationship between crime rate and per capita income.



(Figure 3)

Table 3: Summary of Model 1

Term	Estimate	Std Error	Statistic	P Value
(Intercept)	9.1884311	0.0798124	115.125305	0.0000000
log(crimes)	0.0666949	0.0084211	7.919963	0.0000000
regionNE	0.1044584	0.0255313	4.091382	0.0000511
regionS	-0.0869835	0.0236180	-3.682939	0.0002596
regionW	-0.0552796	0.0281671	-1.962561	0.0503342

The estimated slope for log(crimes) in model 1 is 0.067 as shown in Table 3. Since this coefficient is small, Sheather tells us we can interpret it using percentages (Sheather 2009). That is, a 1% change in crimes correponds with a roughly 0.067% change in per capita income. Even though this impact is statistically significant, it is still small in an absolute sense. More importantly, the direction of the association goes against our intution as well as established research on crime rates and income levels. This is likely because model 1 does not take into account population. We will see below that once we control for this omitted variable crime rates no longer predict income well.

Modeling with all variables

Next we tried to find the best model including a wider subset of the variables. We immediately excluded the `county` and `state` variables since each has so many factor levels that including them would certainly cause the model to overfit. On the other hand, understanding regional variation was important to our collaborators. For both of the methods below we coerced the algorithms to include `region` variable.

The numerical variables were almost all right-skewed. The amount of skewing varied so using Box-Cox transformations would likely results in a more formally valid model but for the sake of interpretability we chose to apply simple log transformations to each. The high school graduation variable shows left-skewing so we applied a power 2 transformation.

Two variable selection methods were used. Lasso regression and bidirectional stepwise regression beginning with the saturated model. The results from these methods are displayed in Table 4 below.

Table 4: Coefficient Estimates

Term	Stepwise	Lasso
log_crimes	0.0165337	-
log_doctors	0.0680864	0.0533612
log_hosp_beds	-	-
log_land_area	-0.0353510	-0.0345217
log_pop	-0.0432642	-
log_pct_bach	0.3833255	0.2957180
log_pct_ue	0.0950896	0.0509220
pct_below_pov	-0.0256074	-0.0197658
sq_pct_hs	-0.0000489	-0.0000124
pop_18_34	-0.0159102	-0.0123949
pop_65_plus	-0.0035269	-
regionW	0.0004903	-0.0193872
regionS	-0.0253447	-0.0344077
regionNC	0.0195700	-0.0059318
regionNE	-	-

The lasso model is much more parsimonious, dropping the `crimes`, `population`, and `65_plus` variables. The stepwise model keeps these variables though a quick look at the p-values shows that the coefficient for `crimes` is not significantly different from zero ($p = 0.146$). Neither model found the number of hospital beds in a county to be predictive of income.

The estimates for some of these coefficients are surprisingly similar between both methods. Take for example the `log_land_area` variable. Both models suggest a coefficent of about -0.035. That is, a 100% increase in the square footage of a given county (i.e. if it doubles) corresponds with a 3.5% decrease in per capita income. Similarly, a county located in the south (the only significant region variable) is expected to

have a per capita income level about 2.5-3.4% lower than a similar county in the north east, our base-line region.

Although both models are useful, we prefer the lasso regression for two reasons. First, the smaller model is preferred for the sake of parsimony. Secondly, because stepwise does not do an exhaustive model search, we somewhat arbitrarily chose to begin our search with the saturated model. But a different starting point could have led to very different results. Lasso does not suffer from this drawback.

Evaluating Missing Data

To answer how problematic the missing data is to our analysis, we have to consider what kind of mechanism leads to the data being absent. Van Buuren (2018) describes all missing data as falling into one of three categories. It can be missing completely at random (MCAR). This means their level of missingness does not depend on any of the observed or unobserved variables of interest. In this case, indepdence will ensure the missing data will not add any kind of bias to our estimates (though the smaller sample size may lead to a higher variance). Second, the data may be considered missing at random (MAR). This occurs when the amount of missingness depends on some variable but we track and control for that variable in our model. Within these recognized categories, the data is MCAR.

Unfortunately, neither of these cases apply to our research. Instead, our data is missing not at random (MNAR). That is, it's missingness is determined by a relevant variable (total population) and we cannot control for this pattern because those counties with fewer than 100,000 residents are completely omitted not partially missing. Thus, bias is definitely a risk we need to account for.

One way to address this shortcoming would be through some imputation method but since the data is missing implicitly this is not an option. Our only option is limit the scope of our conclusions to regions with medium to high population. So long as we contextualize our analysis properly our conclusions can still be valid.

Discussion

Now that we have discussed the results of our analysis, we are prepared to answer the questions set out at the beginning. First, we learned through studying correlation that most economic indicators, including poverty level and unemployment, relate closely to personal income. We also identified a surprising inverse relationship between age and education. It's possible this may be the result of changing economic circumstances. Commentators and pundits commonly talk about how jobs that previous generations could hold down with just a high school diploma now require a college degree. Our data suggests this trend may have impacted how the younger generation has valued pursuing higher education.

The second question asked whether the crime rate is related to per capita income. Contrary to common perceptions, our analysis showed that crime rates have a very weak (though statistically significant) association with income level. But even this signal disappears once other variables are controlled for. Poorer communities have been known to struggle to attract investment because potential investors worry about crime impacting the business opportunity. Helping investors realize this association is stereotypical but not always grounded in reality can help alleviate the economic stress these communities face.

Third, we set out to find the best model that predicts per capita income without artificially limiting the predictor space. We preferred the lasso regression model that included the variables `log(doctors)`, `log(land_area)`, `log(pct_bach)`, `log(pct_ue)`, `log(pct_below_poverty)`, `(pct_hs)^2`, `pop_between_18and34`, and `region`. Perhaps our most important takeaway is how impactful the land area of a county is. Larger counties are expected to have significantly lower income even after population has been controlled for (indirectly through the `doctors` proxy).

Finally, our analysis of the missing data showed our collaborators in the social sciences are rate to suspect the missing data might bias our results. However, by properly contextualizing our results we can make sure we don't draw conclusions too broadly.

Before generalizing our results the reader should be aware of two possible limitations. First, the data our study relies on is nearly 30 years old. Even in 1990 we identified a changing environment with respect to the essential education variables. This effect might be more significant today and it would be helpful to see more recent data. Secondly, urbanization is not well accounted for our in research. Although it is trivial to add a population density to our model, this would have limited value until data is collected on the nearly 2600 rural counties not included in the dataset.

References

Buuren, S. van. (2018). *Flexible imputation of missing data*. CRC Press, Taylor & Francis Group.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2013). *Applied Linear Statistical Models*. McGraw-Hill Education (India) Private Limited.

Sheather, S. J. (2009). *A modern approach to regression with R. Springer*.

US Census Bureau, (2021, October 8). Historical population density data (1910-2020). Census.gov. Retrieved October 30, 2021, from https://www.census.gov/data/tables/time-series/dec/density-data-text.html.

Technical Appendix

Part A: Data description.

The dataset contains 440 row and 16 columns

```
dim(cdi)
```

```
## [1] 440  16
```

The code below shows that the dataset has the same number of rows before and after removing rows with missing data. From this we can infer there are no NAs in the dataset.

```
nrow(cdi) == nrow(na.omit(cdi))
```

```
## [1] TRUE
```

The county column can be omitted since it is (when combined with state) a key column with a unique value for every row in the data. The other two categorical variables (state and region) have their frequencies displayed in tables 1 and 2. We also include a series of numerical summaries of the data in table 3.

```
count(cdi, state) %>% arrange(desc(n)) %>% head(10) %>%
  knitr::kable(caption = "Number of Counties in the Top 10 States")
```

Table 1: Number of Counties in the Top 10 States

state	n
CA	34
FL	29
PA	29
TX	28
OH	24
NY	22
MI	18
NC	18
NJ	18
IL	17

```
count(cdi, region) %>% arrange(desc(n)) %>%
  knitr::kable(caption = "Number of Counties in Each Region")
```

Table 2: Number of Counties in Each Region

region	n
S	152
NC	108
NE	103
W	77

```
cdi %>%
  select(where(is.numeric)) %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value") %>%
  group_by(variable) %>%
  summarise(mean = mean(value, na.rm = T),
            median = median(value, na.rm = T),
```

```
            sd = sd(value, na.rm = T),
            min = min(value, na.rm = T),
            max = max(value, na.rm = T)) %>%
  mutate(across(-variable, function(x) round(x, 2))) %>%
  knitr::kable(caption = "Numerical Summaries")
```

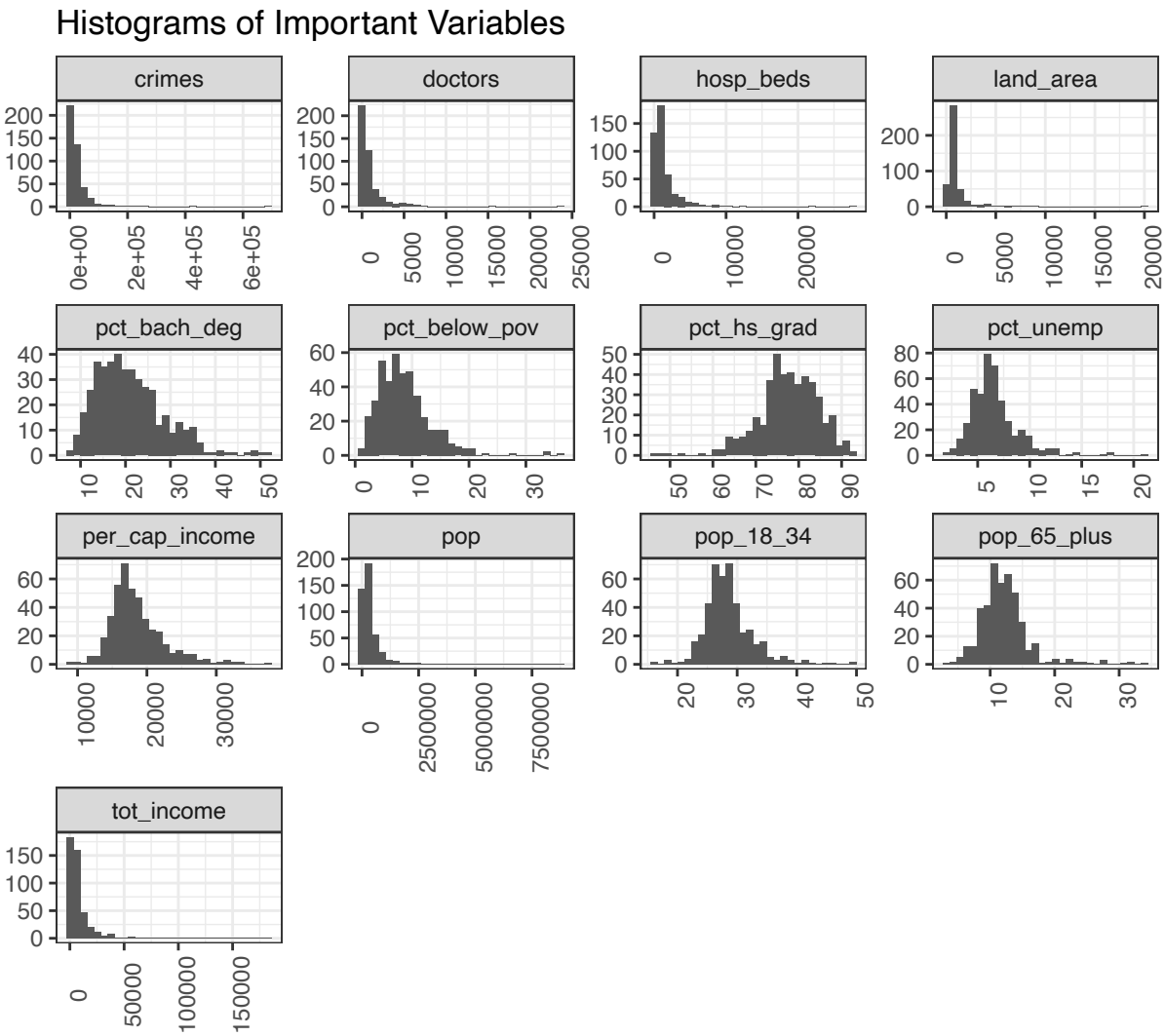
Table 3: Numerical Summaries

variable	mean	median	sd	min	max
crimes	27111.62	11820.50	58237.51	563.0	688936.0
doctors	988.00	401.00	1789.75	39.0	23677.0
hosp_beds	1458.63	755.00	2289.13	92.0	27700.0
land_area	1041.41	656.50	1549.92	15.0	20062.0
pct_bach_deg	21.08	19.70	7.65	8.1	52.3
pct_below_pov	8.72	7.90	4.66	1.4	36.3
pct_hs_grad	77.56	77.70	7.02	46.6	92.9
pct_unemp	6.60	6.20	2.34	2.2	21.3
per_cap_income	18561.48	17759.00	4059.19	8899.0	37541.0
pop	393010.92	217280.50	601987.02	100043.0	8863164.0
pop_18_34	28.57	28.10	4.19	16.4	49.7
pop_65_plus	12.17	11.75	3.99	3.0	33.8
tot_income	7869.27	3857.00	12884.32	1141.0	184230.0

EDA

We begin by examining the marginal distributions of the numeric variables. Figure 1 shows significant right skewing on most of the variables. This is unsurprising since we anticipate many of the "count" variables like the numeber of crimes or doctors will be closely correlated the right-skewed population variable. The percentage of high school graduates is the only column that appears left skewed.

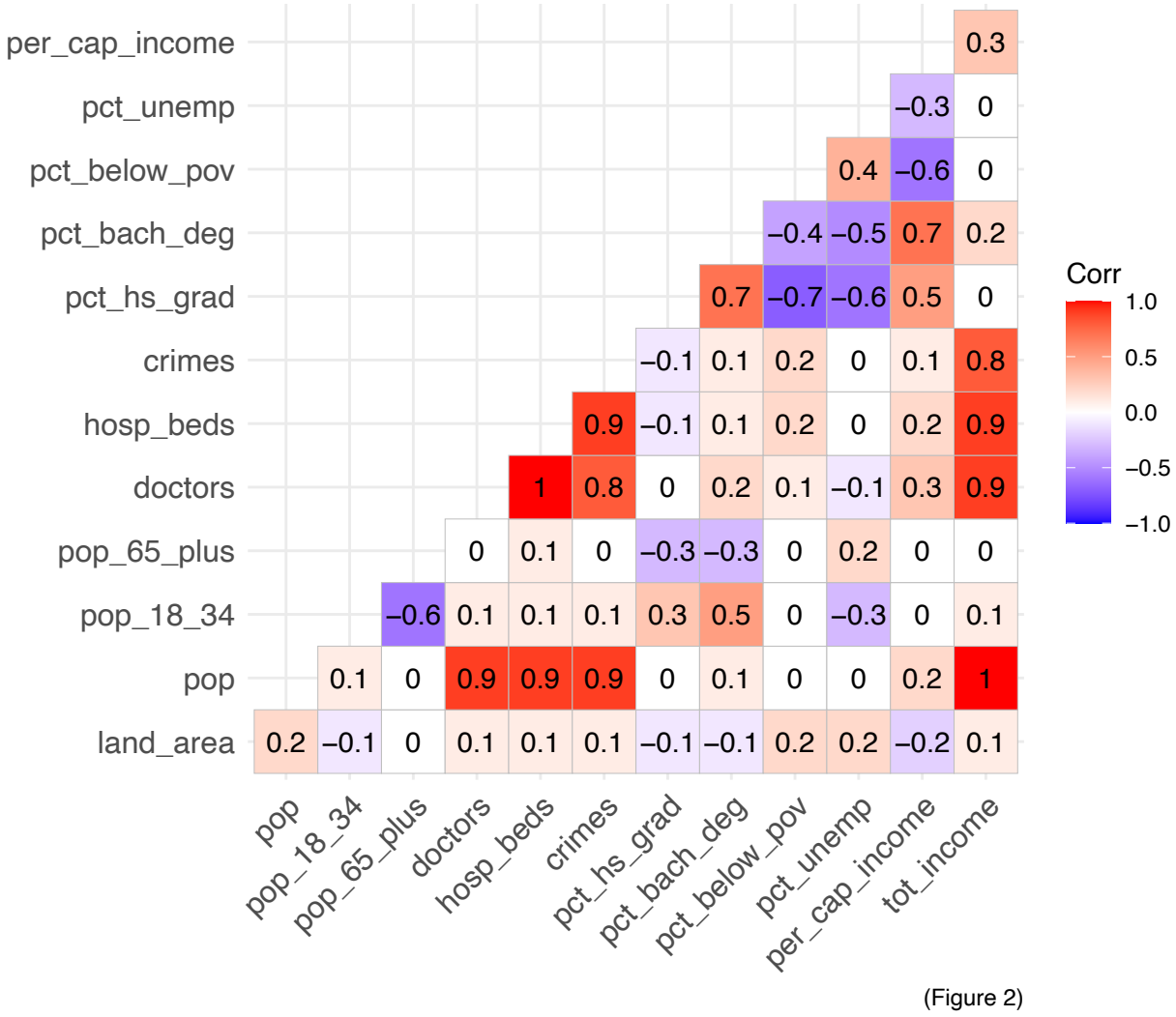
```
cdi %>%
  select(where(is.numeric)) %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value") %>%
  ggplot() +
  geom_histogram(aes(value)) +
  labs(x = "", y = "", title = "Histograms of Important Variables", caption = "(Figure 1)") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 90)) +
  facet_wrap(vars(variable), scales = "free")
```



(Figure 1)

Figure 2 confirms our suspicions that population, total income, number of crimes, doctors, and hospital beds are all closely positively correlated. We also learn that the education variables tend to be negatively correlated with the economic health variables. That is, a county with better educated residents tends to perform better economically.

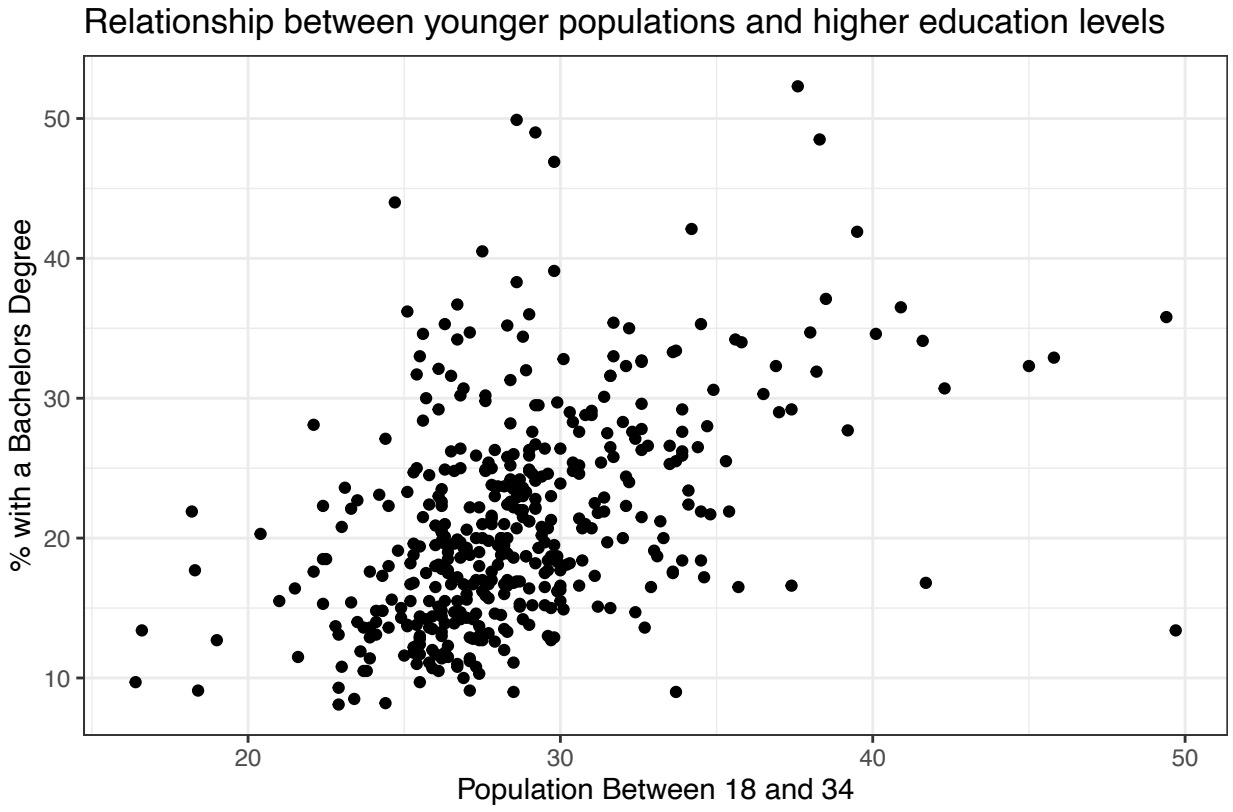
```
cdi %>%
  select(where(is.numeric)) %>%
  cor() %>%
  ggcorrplot::ggcorrplot(type = "lower", lab = T, digits = 1) +
  labs(caption = "(Figure 2)")
```



(Figure 2)

Most of these relationships are unsurprising. What is less intuitive is the relationship between counties with high population between 18 and 34 and the percentage of residents with bachelors degrees. These variables have a pearson correlation of 0.46. One possible takeaway is that younger generations are expected to have a higher education level than their predcessors. The plot in Figure 3 displays this relationship graphically.

```
cdi %>%
  ggplot() +
  geom_point(aes(pop_18_34, pct_bach_deg)) +
  labs(x = "Population Between 18 and 34", y = "% with a Bachelors Degree",
       title = "Relationship between younger populations and higher education levels",
       caption = "(Figure 3)") +
  theme_bw()
```



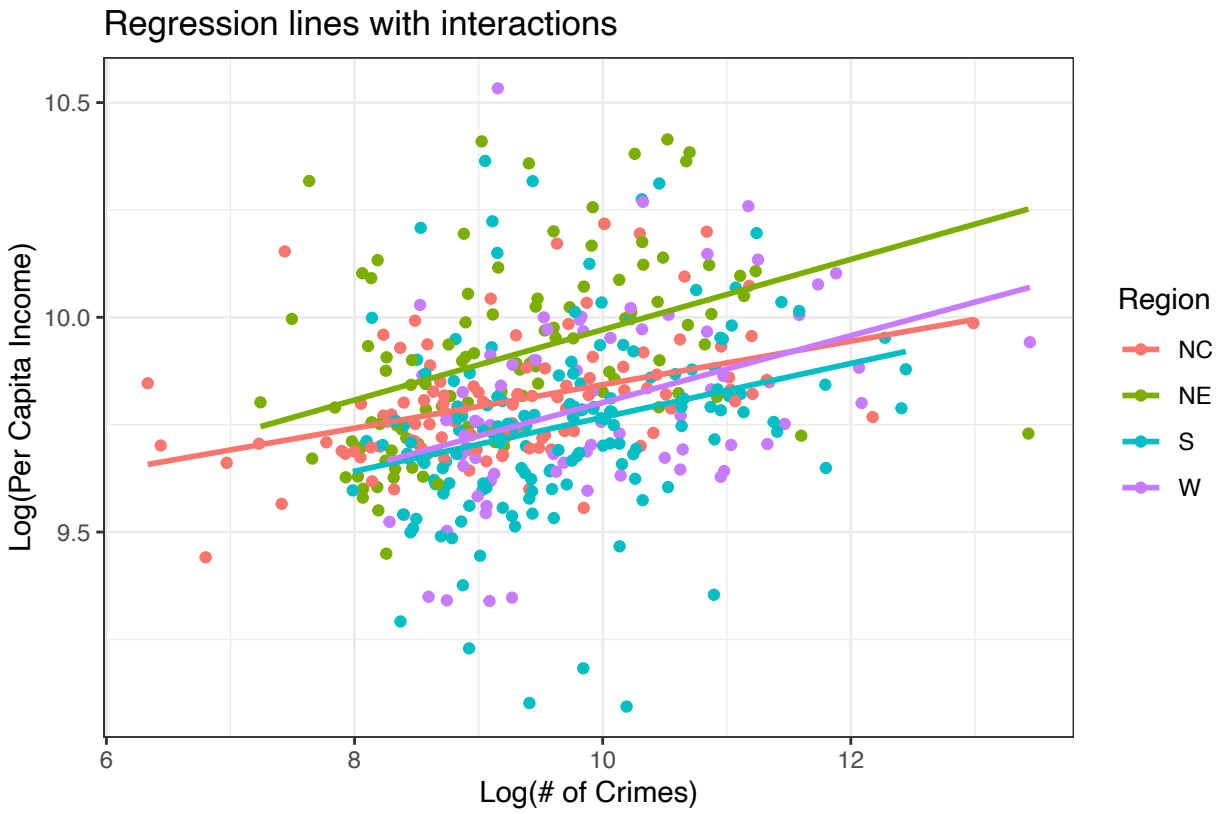
(Figure 3)

Part B: Modeling with just crime

As we saw in part A, `crimes` and `per_cap_income` are right-skewed so we applied a log transform to both and fit a linear model with region as a potential confounder. Figures 4 and 5 show what these regressions would look like with and without interaction effects. In the first figure, we can see the lines are mostly parallel. The NC region has the most distinct slope but this is possibly caused by the influential points off to the left.

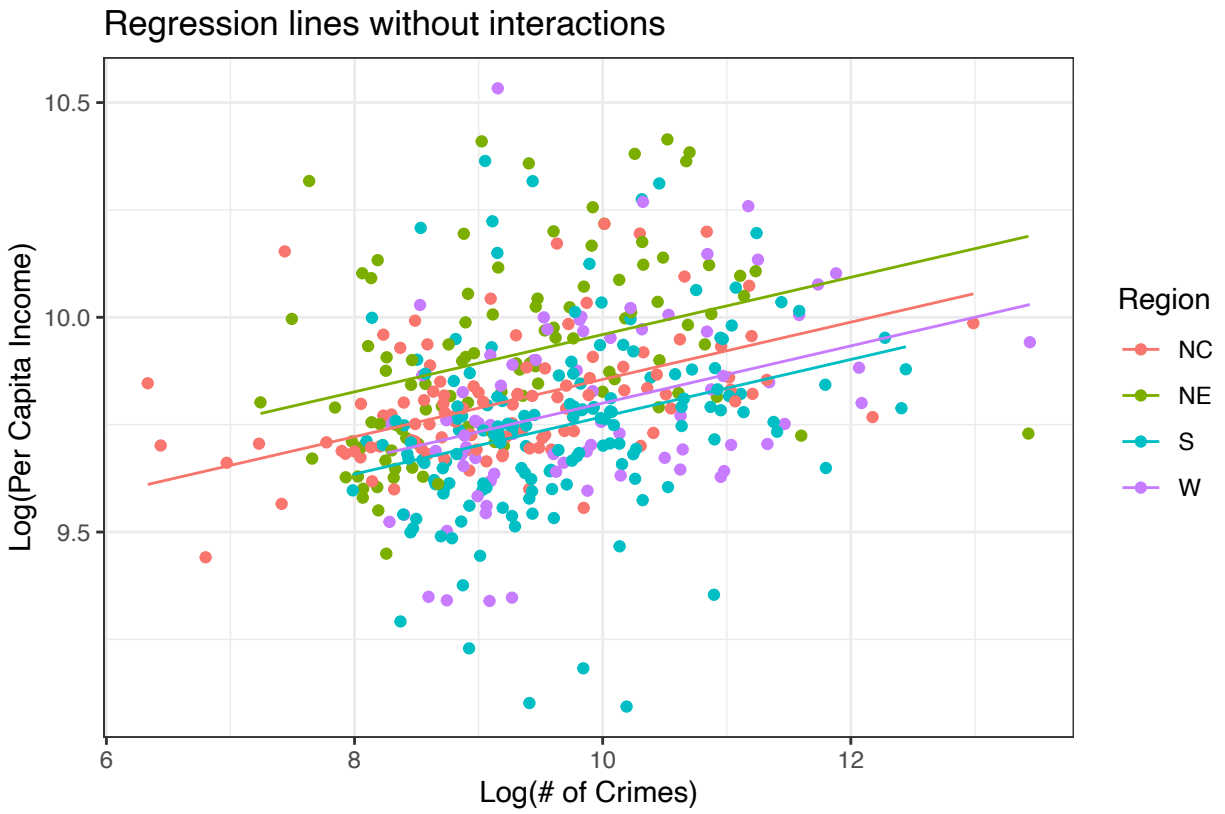
```
mod_wo_interactions <- lm(log(per_cap_income) ~ log(crimes) + region, data = cdi)
mod_w_interactions <- lm(log(per_cap_income) ~ log(crimes)*region, data = cdi)

cdi %>%
  ggplot(aes(log(crimes), log(per_cap_income), color = region)) +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  labs(x = "Log(# of Crimes)",y = "Log(Per Capita Income)",
       color = "Region", caption = "(Figure 4)",
       title = "Regression lines with interactions") +
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



(Figure 4)

```
broom::augment(mod_wo_interactions, cdi) %>%
  ggplot(aes(x = log(crimes), color = region)) +
  geom_point(aes(y=log(per_cap_income))) +
  geom_line(aes(y=.fitted)) +
  labs(x = "Log(# of Crimes)",y = "Log(Per Capita Income)",
       color = "Region", caption = "(Figure 5)",
       title = "Regression lines without interactions") +
  theme_bw()
```



(Figure 5)

The summary output below confirms that the difference in slopes is negligible. None of the interaction terms are significant and the p-value from the partial F-test is 0.5266 much higher than $\alpha = 0.05$. Thus, we are unable to detect a significant improvement to the model by including interaction terms.

```
summary(mod_w_interactions)
```

```
##
## 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
```

```
summary(mod_wo_interactions)
```

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

```
anova(mod_w_interactions, mod_wo_interactions)
```

```
## Analysis of Variance Table
##
## Model 1: log(per_cap_income) ~ log(crimes) * region
## Model 2: log(per_cap_income) ~ log(crimes) + region
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    432 14.872
## 2    435 14.949 -3 -0.076778 0.7434 0.5266
```

We also learn from the above output that, counter intuitively, per capita income is weakly positively related to the crime rate. Specifically, a 1% increase in the crime rate is associated with a 0.067% increase in per capita income.

Just to be thorough, we also attempted substituting per capita crime rate for total crimes and received very similar results - β_1 declined slightly from 0.067 to 0.042. Unfortunately, the explanatory power of the model was almost halved in the process. The per capita rate is likely better for our model since it allows us to adjust for the potential confounder population we identified in part A.

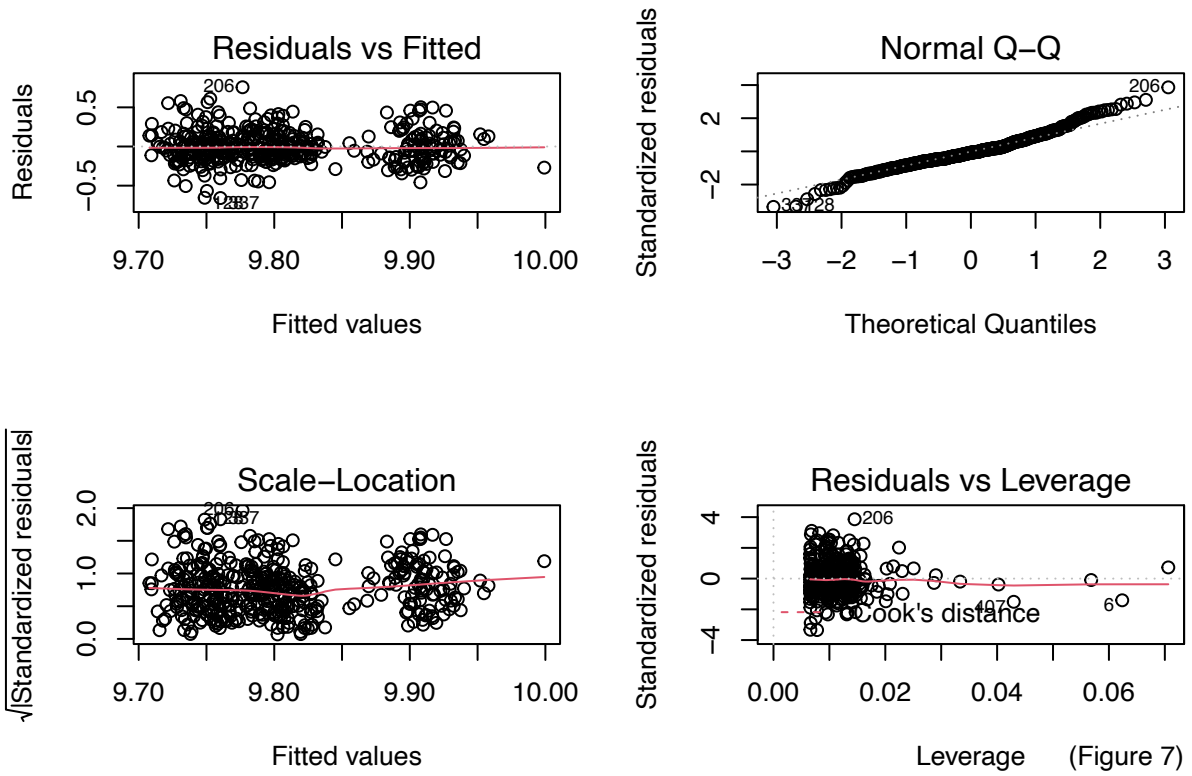
```
model_3 <- lm(log(per_cap_income) ~ log(per_capita_crimes) + region, data = mutate(cdi, per_capita_cr
summary(model_3)
```

```
##
## Call:
## lm(formula = log(per_cap_income) ~ log(per_capita_crimes) + region,
##     data = mutate(cdi, per_capita_crimes = as.numeric(crimes)/as.numeric(pop)))
##
```

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

Figure 7 shows us more work will need to be done. The residuals vs fit plot shows linearity but it also hints that there might be omitted variable bias since the residuals form two distinct clusters. The variance appears constant by the residuals deviate significantly from the normal distribution.

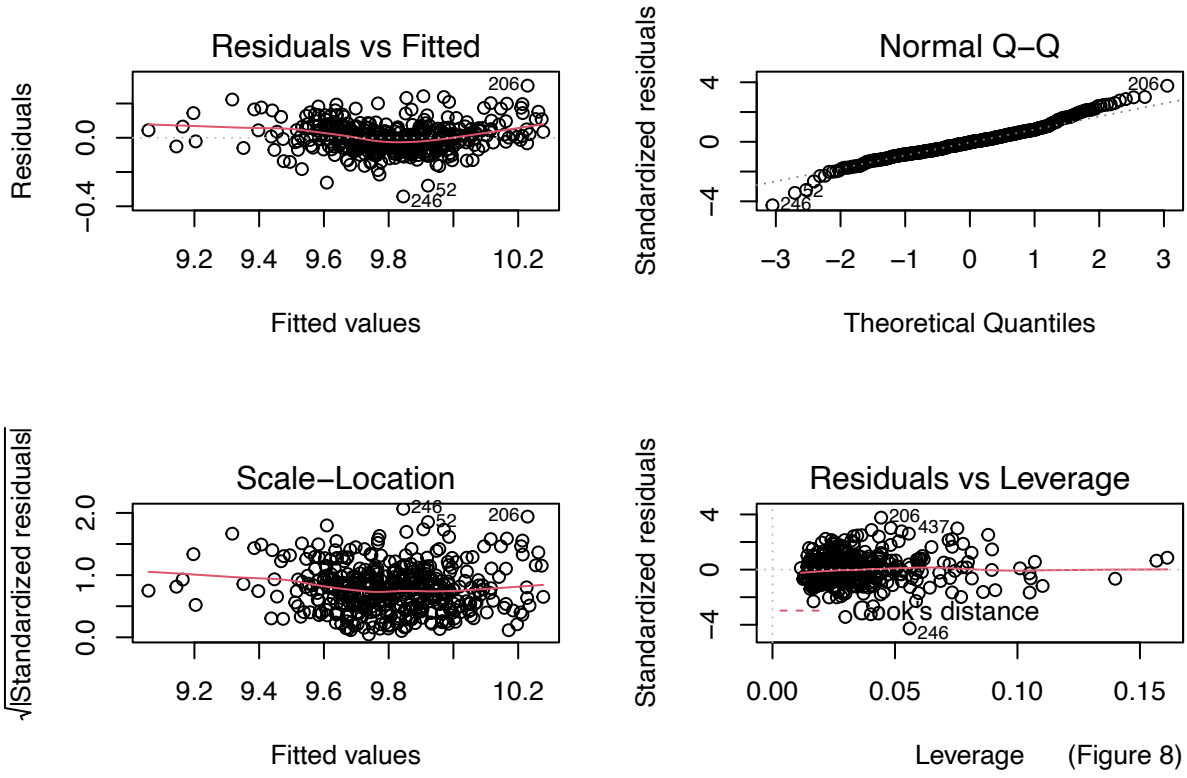
```
par(mfrow = c(2,2))
plot(model_3)
title(sub="(Figure 7)", adj=1, line=3, font=2)
```



Part C: Modelling with all the variables

In order to avoid making transformations that will be difficult to explain to a non-specialist, we will limit ourselves to just log and power 2 transforms. We square the `pct_hs_grad` column to address the left-skewing and take the log of all the others except `pct_below_pov`, `pop_18_34`, and `pop_65_plus`. Income is dropped since it is a simply function of population and per capita income. The resulting saturated model has the residual patterns shown in Figure 8. Some non-linear relationship has not been captured by our model as shown in the residual vs fit plot. The residuals are also not perfectly normally distributed - there is some evidence of overdispersion.

```
cdi_model_data <- cdi %>%
  mutate(regionW = as.numeric(region == "W"),
       regionS = as.numeric(region == "S"),
       regionNC = as.numeric(region == "NC"),
       regionNE = as.numeric(region == "NE")) %>%
  transmute(log_crimes = log(crimes),
            log_doctors = log(doctors),
            log_hosp_beds = log(hosp_beds),
            log_land_area = log(land_area),
            log_pop = log(pop),
            log_pcap_income = log(per_cap_income),
            log_pct_bach = log(pct_bach_deg),
            log_pct_ue = log(pct_unemp),
            pct_below_pov,
            sq_pct_hs = pct_hs_grad^2,
            pop_18_34,
            pop_65_plus,
            regionW, regionS, regionNC, regionNE)

full_model_cdi <- lm(log_pcap_income ~ ., data = cdi_model_data)

par(mfrow = c(2,2))
plot(full_model_cdi)
title(sub="(Figure 8)", adj=1, line=3, font=2)
```



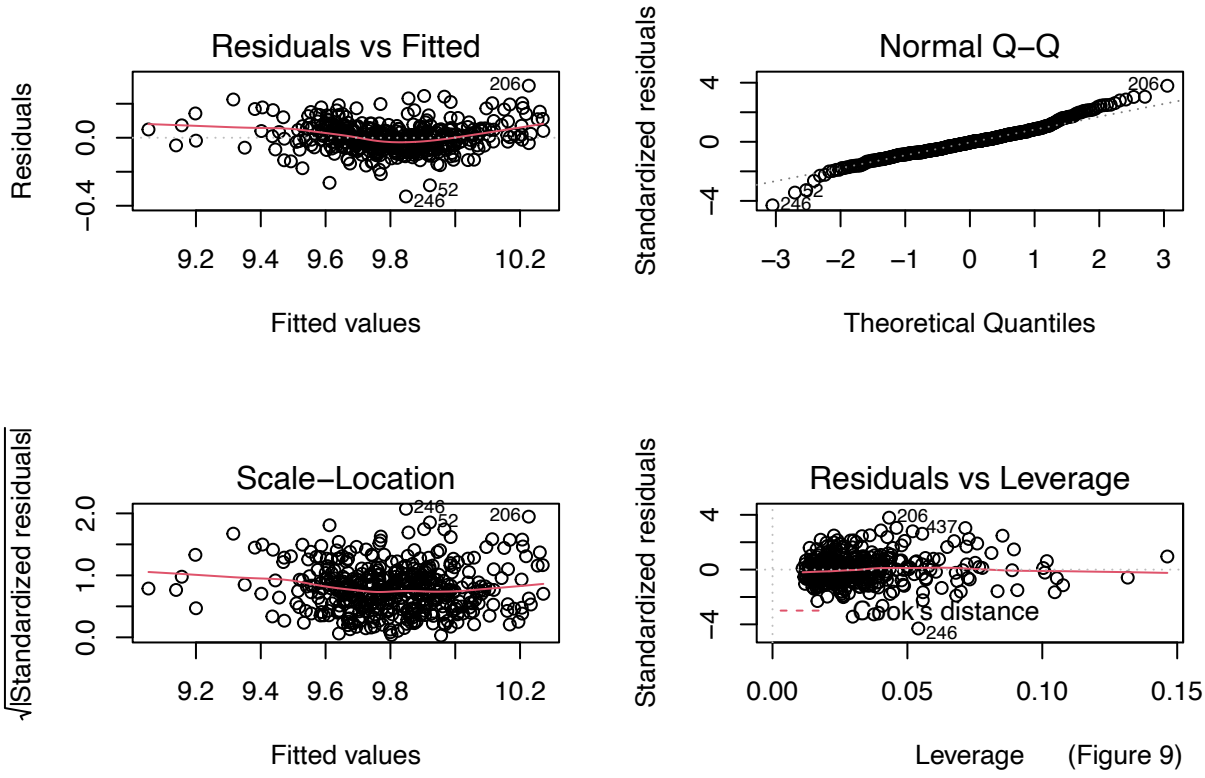
(Figure 8)

To address some of these shortcomings, and to deal with the multi-collinearity we are confident is present, we use stepwise regression to winnow down the predictor space. This eliminates the `log_hosp_beds` and `log_crimes` variables. None of the remaining variables have VIFs greater than 10 so we can be fairly confident multicollinearity is no longer a pressing concern.

```
best_cdi_model <- step(full_model_cdi, direction = "both",
                       scope = list(
                         lower = log_pcap_income ~ regionW + regionS + regionNC + regionNE,
                         upper = log_pcap_income ~ .))
```

Unfortunately, the same concerns are present in the new diagnostic plots below. This suggests we will need to consider other approaches, perhaps lasso or another regularization technique.

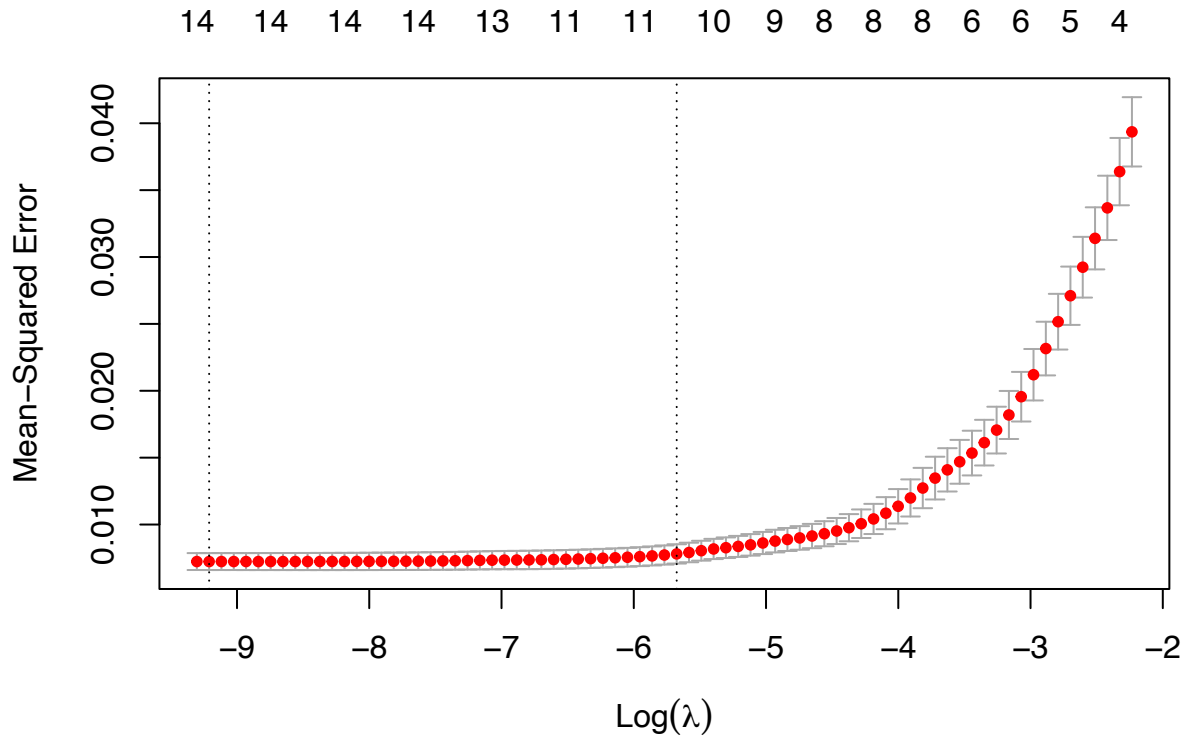
```
par(mfrow = c(2,2))
plot(best_cdi_model)
title(sub="(Figure 9)", adj=1, line=3, font=2)
```



(Figure 9)

LASSO regression

```
lasso_cv <- cv.glmnet(x = as.matrix(cdi_model_data[,-c(6,16)]), y = cdi_model_data[,6],
                      penalty.factor = c(rep(1,11), rep(0,3)))
plot(lasso_cv)
```



```
lasso_mod <- glmnet(x = as.matrix(cdi_model_data[,-c(6,16)]), y = cdi_model_data[,6],
                    lambda = lasso_cv$lambda.1se,
                    penalty.factor = c(rep(1,11), rep(0,3)))

original_vars <- tibble(term = names(cdi_model_data))
coefs_step <- broom::tidy(best_cdi_model)[,1:2]
coefs_table <- original_vars %>%
  left_join(coefs_step) %>%
  left_join(broom::tidy(lasso_mod)[,c(1,3)], by = "term")
```

```
## Joining, by = "term"
```

```
names(coefs_table) <- c("Term", "Stepwise", "Lasso")
knitr::kable(coefs_table)
```

Term	Stepwise	Lasso
log_crimes	0.0165337	NA
log_doctors	0.0680864	0.0533612
log_hosp_beds	NA	NA
log_land_area	-0.0353510	-0.0345217
log_pop	-0.0432642	NA
log_pcap_income	NA	NA
log_pct_bach	0.3833255	0.2957180
log_pct_ue	0.0950896	0.0509220
pct_below_pov	-0.0256074	-0.0197658
sq_pct_hs	-0.0000489	-0.0000124
pop_18_34	-0.0159102	-0.0123949
pop_65_plus	-0.0035269	NA
regionW	0.0004903	-0.0193872
regionS	-0.0253447	-0.0344077
regionNC	0.0195700	-0.0059318
regionNE	NA	NA