Examining Sociological Effects on Per-Capita Income

Egan McClave, emcclave@andrew.cmu.edu

October 19, 2018

Contents

List of Figures **1**

List of Tables **1**

Abstract **2**

1 Introduction **2**

2 Methods **3**
2.1 Data Introduction . . . 3
2.2 Exploratory Data Analysis . . . 4
2.3 Visually Exploring the Effect of Crime . . . 7

3 Results **8**
3.1 Examining Relationships Between Variables . . . 8
3.2 Examining Effect of Crime on Per-Capita Income . . . 8
3.3 Examining Variables for Selection . . . 9
3.4 Examining Final Model . . . 9

4 Discussion **11**

5 References **12**

6 Code Appendix **13**
6.1 Exploratory Data Analysis . . . 13
6.2 Visually Exploring the Effect of Crime . . . 15
6.3 Examining Effect of Crime on Per-Capita Income . . . 17
6.4 Examining Variables for Selection . . . 17
6.5 Examining Final Model . . . 18

List of Figures

1 Distributions of Variables in `cdi.dat` . . . 4
2 Bivariate Analysis of Variables in `cdi.dat` . . . 5
3 Log Per Capita Income vs Every Numeric Variable (with `lm` model) . . . 6
4 Per-Capita Income vs Crime Rate by Region . . . 7
5 Log Per-Capita Income vs Log Crime Rate by Region . . . 7
6 Diagnostic Plots for Final Model . . . 10

List of Tables

1 Summary Statistics for Numeric Variables . . . 4
2 Initial and Final Transformations of Numeric Predictor Variables . . . 6
3 Summary Coefficients for Log Per-Capita Income on Crime Rate . . . 8
4 Summary Coefficients for Log Per-Capita Income on Log Crime Rate . . . 8
5 Variance Influcence Factors for Variables . . . 9
6 Summary Coefficients for Final Model . . . 11

Abstract

The purpose of this study was to examine which socioeconomic variables might be associated with the average income per person. We analyzed 440 instances of counties with features that describe their economic, health and social status over the course of a two year time period. Utilizing several methods we determined the optimal model that predicts Per-Capita Income contains variables on the county's land area, the percentage of residents between 18 and 34, the number of hospital beds, the total number of crimes in 1990, the percentage of adults who have graduated high school at a minimum, the number of residents who have earned a bachelor's degree, the percentage of residents with an income below the poverty line, and the percentage of residents who are unemployed. We also examined the impact the location of a county has on the relationship between Per-Capita Income and the county crime rate.

1 Introduction

Finding a job can be very difficult. The job market is a complex environment where opportunities are constantly changing. Jobs for an individual in one location are entriely different than jobs offered for said individual in another location - this means that the pay can be widly diverse as well. But how much of this difference in pay be attributed to simply a change in location? What if it is affected by other variables? This begs the question: how might the average income per person be related to other variables associated with their county's economic, health and social well-being?

This question is intimately dicussed between Social scientists. Social science is a large field of study and has many sub-branches specializing in a variety of topics. Sociology is one of these topics and focuses on the scientific study of society. Social scientists are researchers who use several methods to analyze anything from the relationships individuals have with one another to society as a whole. Thus, it is of interest to these researchers to determine this complex relationship. In addition to exploring the research goal defined above, we will address the following questions:

- Among the variety of variables, which pairs have a clear relationshp with one another? Are there any surprising relationships defined in the data?
- Is there any validation to the idea that per-capita income is dependent to crime rate? Is their relationship any different depending on the region of the country (Northeast, Northcentral, South, and West)?
- What might be the best clear and representative model for predicting per-capita income?
- Does having an unrepresentative dataset of the whole country invalidate any of the claims we make?

2 Methods

2.1 Data Introduction

This data is available in the `cdi.dat.txt` file from [Kutner, 2005]. It contains county demographic information (CDI) for 440 of the most populous counties in the United States between the years 1990 and 1992 [BEA, 1998]. Each record has 17 variables associated with it and they are discussed in more detail below. Several of the counties had missing values and thus were dropped from the datset for ease of analysis. The variable of interest in this study is `per.cap.income`. All analysis are done with the R language and the environment [R, 2017].

`id`	=	Identification number (1-44)
`county`	=	County name
`state`	=	Two-letter state abbreviation
`land.area`	=	Land area (square miles)
`pop`	=	Estimated 1990 population
`pop.18_34`	=	Percent of 1990 CDI population aged 18-34
`pop.65_plus`	=	Percent of 1990 CDI population aged 65 or older
`doctors`	=	Number of professionally active nonfederal physicians during 1990
`hosp.beds`	=	Total number of beds, cribs, and bassinets during 1990
`crimes`	=	Total number of serious crimes in 1990
`pct.hs.grad`	=	Percent of adult population (aged 25 or older) who completed 12 or more years of school
`pct.bach.deg`	=	Percent of adult population (aged 25 or older) with bachelor's degree
`pct.below.pov`	=	Percent of 1990 CDI population with income below poverty line
`pct.unemp`	=	Percent of 1990 CDI population that is unemployed
`per.cap.income`	=	Per-capita income of 1990 CDI population (in dollars)
`tot.income`	=	Total personal income of 1990 CDI population (in millions of dollars)
`region`	=	Geographic region of United States

2.2 Exploratory Data Analysis

For our analysis, we explored the relationship between variables by utilizing exploratory data analysis plots such as histograms and scatterplots accompanied with quantitative summary statistics. It is important to describe the distributions visually as it is useful in identifying the variables that might potentially require transformations. The plots in Figure 1 are the univariate histograms of all the variables. It appears that many of the variables are right skewed, few are left skewed and even fewer variables are appear normally distributed.

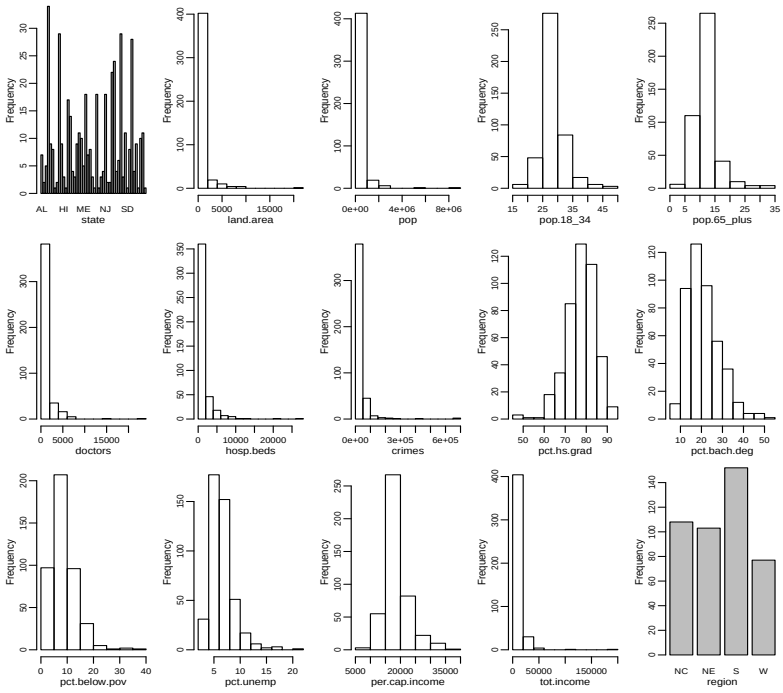


Figure 1: Distributions of Variables in `cdi.dat`

Table 1 displays the summary statistics (Min, 1st Q, Median, Mean, 3rd Q, Max, Std. Dev) for each of the 13 numeric variables. Similarily, the results from this table reconfirm the claims about Figure 1.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std. Dev
land.area	15.00	451.25	656.50	1.04E+03	946.75	20062.00	1.55E+03
pop	100043.00	139027.25	217280.50	3.93E+05	436064.50	8863164.00	6.02E+05
pop.18_34	16.40	26.20	28.10	2.86E+01	30.02	49.70	4.19E+00
pop.65_plus	3.00	9.88	11.75	1.22E+01	13.62	33.80	3.99E+00
doctors	39.00	182.75	401.00	9.88E+02	1036.00	23677.00	1.79E+03
hosp.beds	92.00	390.75	755.00	1.46E+03	1575.75	27700.00	2.29E+03
crimes	563.00	6219.50	11820.50	2.71E+04	26279.50	688936.00	5.82E+04
pct.hs.grad	46.60	73.88	77.70	7.76E+01	82.40	92.90	7.02E+00
pct.bach.deg	8.10	15.28	19.70	2.11E+01	25.33	52.30	7.65E+00
pct.below.pov	1.40	5.30	7.90	8.72E+00	10.90	36.30	4.66E+00
pct.unemp	2.20	5.10	6.20	6.60E+00	7.50	21.30	2.34E+00
per.cap.income	8899.00	16118.25	17759.00	1.86E+04	20270.00	37541.00	4.06E+03
tot.income	1141.00	2311.00	3857.00	7.87E+03	8654.25	184230.00	1.29E+04

Table 1: Summary Statistics for Numeric Variables

It is also equally as important to discern the relationships between pairs of variables in contrast to looking at univariate results. Figure 2 has plots for understanding these relationships between all the numeric variables. The lower diagonal contains scatterplots for pairs of variables along with a loess estimate curve in red and the upper diagonal contains the correlation for pairs of variables.

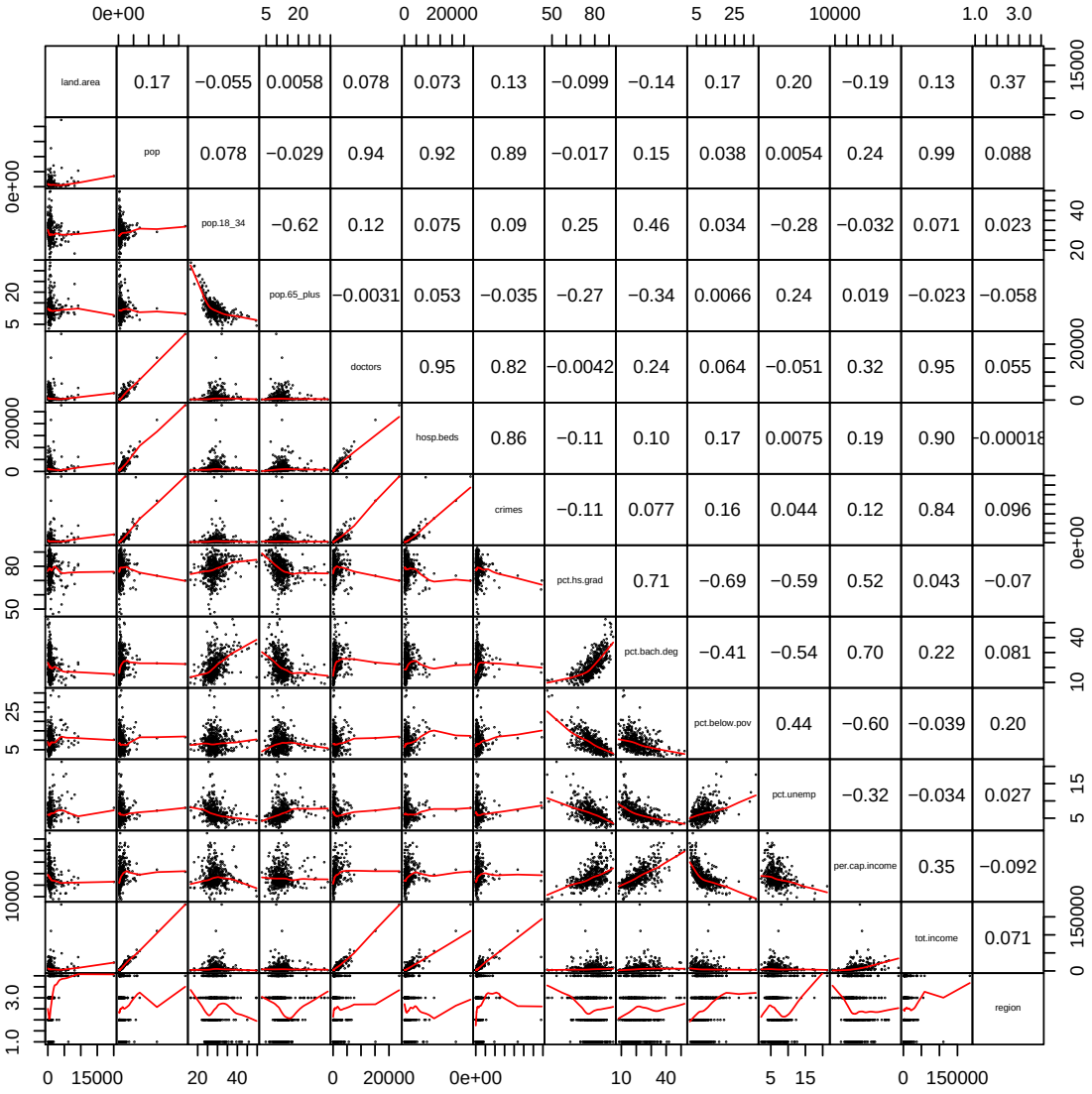
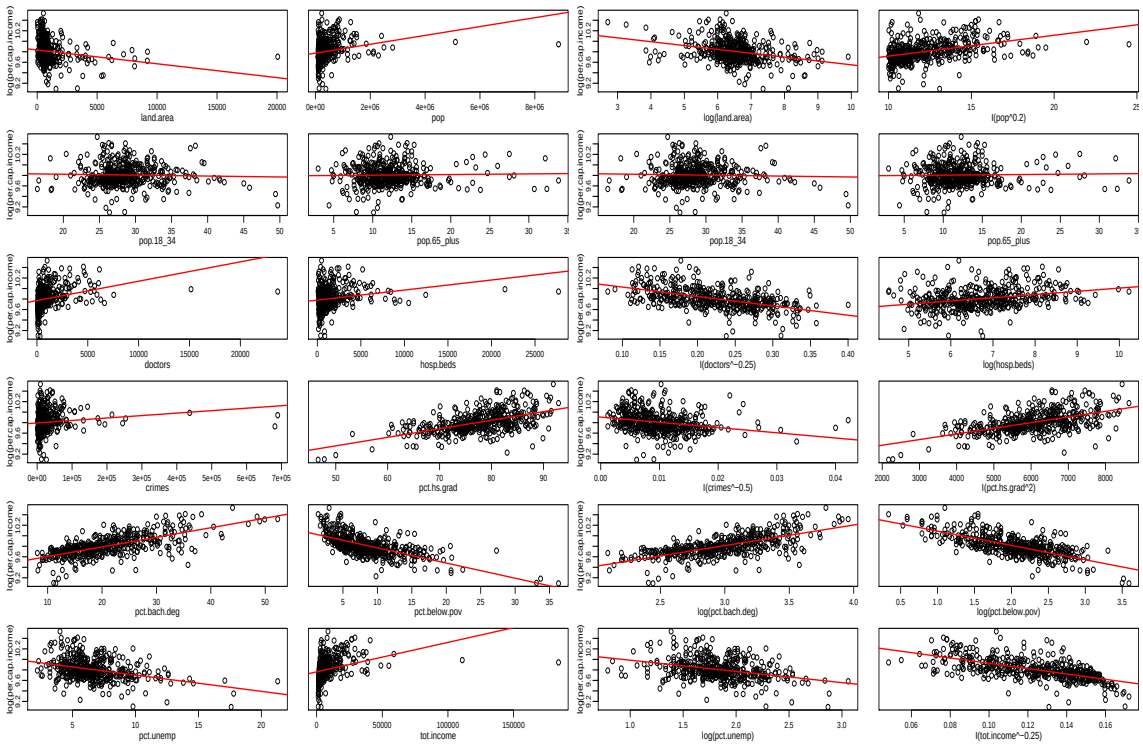


Figure 2: Bivariate Analysis of Variables in `cdi.dat`

Figure 3 display the relationship between log Per Capita Income and the other number variables. Subfigure (a) are scatterplots between `log(per.cap.income)` and the normal predictor variable. Subfigure (b) are the scatterplots between `log(per.cap.income)` and transformed predictor variables.



(a) Normal Predictors (b) Transformed Predictors

Figure 3: Log Per Capita Income vs Every Numeric Variable (with `lm` model)

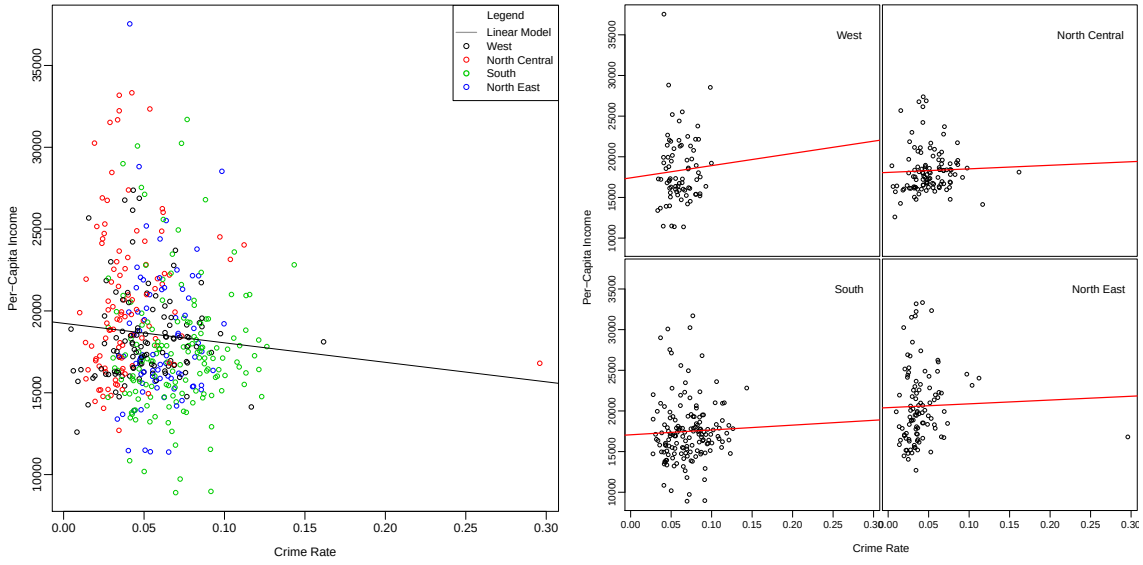
Initial	Final
`land.area`	`log(land.area)`
`pop`	$\texttt{pop}^{0.2}$
`pop.18_34`	`pop.18_34`
`pop.65_plus`	`pop.65_plus`
`doctors`	$\texttt{doctors}^{-0.25}$
`hosp.beds`	`log(hosp.beds)`
`crimes`	$\texttt{crimes}^{-0.5}$
`pct.hs.grad`	$\texttt{pct.hs.grad}^{2}$
`pct.bach.deg`	`log(pct.bach.deg)`
`pct.below.pov`	`log(pct.below.pov)`
`pct.unemp`	`log(pct.unemp)`
`tot.income`	$\texttt{tot.income}^{-0.25}$

Table 2: Initial and Final Transformations of Numeric Predictor Variables

It appears that a large percentage of the simple linear regression models benefit greatly from the transformations indicated in Table 2, the remaining do not appear to have been changed significantly.

2.3 Visually Exploring the Effect of Crime

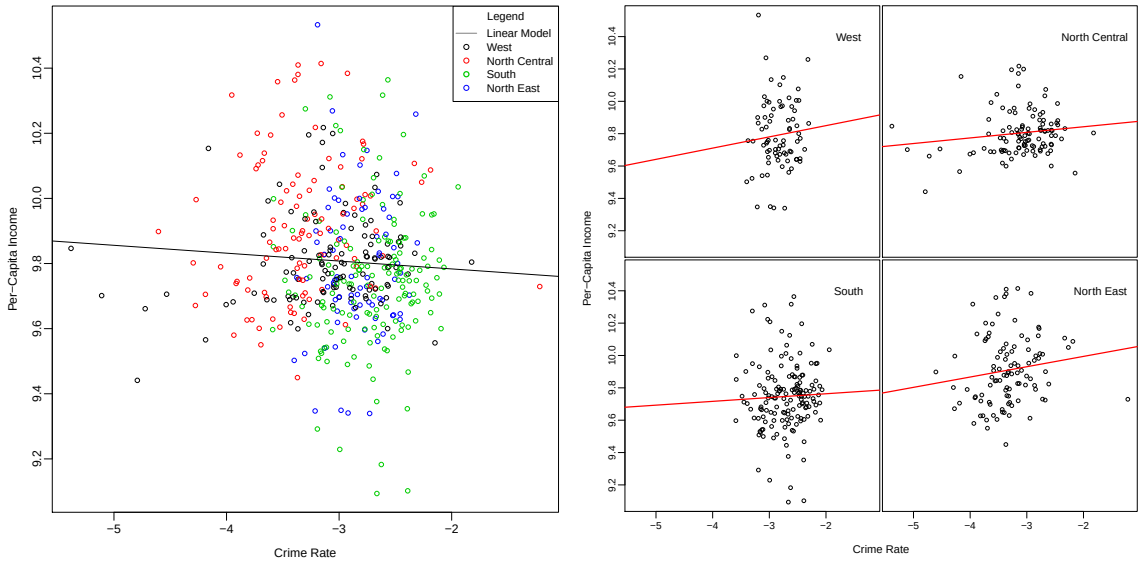
Since it is one of the motivating questions for this study we must examine the relationship between the crime rate ($\frac{\texttt{crime}}{\texttt{pop}}$) and per-capita income and the potential effect knowing the region of the country might have on the relationship. This is gone into detail in Figures 4 and 5 which look at this relationship on the normal and log-scale.



(a) Colored by Region

(b) Split by Region

Figure 4: Per-Capita Income vs Crime Rate by Region



(a) Colored by Region

(b) Split by Region

Figure 5: Log Per-Capita Income vs Log Crime Rate by Region

3 Results

3.1 Examining Relationships Between Variables

Figure 3 details the joint scatterplot of all pairs of variables as well as the correlation between the two varaibles. Very few pairs of variables have a high absolute measure of correlation. Other pairs have unexpected positive linear relationships such as `pop` with `doctors`, `hosp.beds`, `crimes` or `tot.income`. Other variables with similar relationships were `doctors` with `hosp.beds`, `crimes`, or `tot.income`, and `hosp.beds` with `crimes` or `tot.income` and lastly `crime` with `tot.income`.

There are also quite a few pairs that have an inversely proportional non-linear relationships, these are usually between percentages of the same variable so as one increases the other decreases (i.e. `pop.18_34` and `pop.65_plus`); others have a directly proportional relationship (i.e. `pct.hs.grad` `pct.bach.deg`). The unexpected non-linear relationships are between the population percentage variables and `per.unemp` variable.

Most of these relationships imply there is a multi-colinearity issue with the variables in the dataset and should be evaluated further in the model. This will be dicussed in Section 3.3.

3.2 Examining Effect of Crime on Per-Capita Income

	All	W	NC	S	NE
(Intercept)	19244.29	17407.30	18077.29	17066.94	20406.33
Crime Rate	-11919.27	15034.61	4379.07	5937.99	4667.46
Intercept Std Err	448.93	1945.61	630.92	960.98	763.78
Crime Rate Std Err	7074.51	30929.76	11201.35	12846.06	14811.63
Intercept P-value	0.00	0.00	0.00	0.00	0.00
Crime Rate P-value	0.09	0.63	0.70	0.64	0.75

Table 3: Summary Coefficients for Log Per-Capita Income on Crime Rate

	All	W	NC	S	NE
(Intercept)	9.74	9.99	9.91	9.81	10.12
Log Crime Rate	-0.02	0.07	0.03	0.02	0.06
Intercept Std Err	0.06	0.27	0.07	0.13	0.15
Log Crime Rate Std Err	0.02	0.09	0.02	0.05	0.04
Intercept P-value	0.00	0.00	0.00	0.00	0.00
Log Crime Rate P-value	0.22	0.47	0.14	0.62	0.14

Table 4: Summary Coefficients for Log Per-Capita Income on Log Crime Rate

Tables 3 and 4 represent the summary coefficents of the similar linear regression models plotted in Figures 4 and 5. Combining these results in the tables along with the figures they reference we can see that there is a slight negative linear relationship between `per.cap.income` and `crime rate`. However, once the data is split by `region`, the slopes of the models change signs. Figure 4 (b) gives a visual indication that the West has a slope that is significantly different from the other slopes, while Figure 5 (b) gives a visual indication that the slopes between the West and the North East are similar and the slopes between the North Central and the South are similar. We can confirm this by looking at the second row of the summary tables. There might be a relationship between `per.cap.income` and `crime rate` but it is most certainly affected by the influence of what region of the country the county is located in. It is also interesting to note the extreme outliers in some of the regions of the data that might have potentially skewed the results of the analysis. For example, there is a point in the North East that has a large crime rate value and a low Per-Capita Income which throws off the linear model for that plot.

3.3 Examining Variables for Selection

The model for initial consideration based on the transformations that were decided on in Section 2.2 is written in Equation 1. This model was then passed through the backwards step algorithm to determine if any variables should be removed based on BIC values. This filtered model is written in Equation 2. The variance-inflation factors (VIF) for this filtered model are displayed in Table 5.

$$\begin{aligned} \texttt{log(per.cap.income)} \sim\ & \texttt{log(land.area)} + \texttt{pop}^{0.2} + \texttt{pop.18\_34} + \texttt{pop.65\_plus} + \\ & \texttt{doctors}^{-0.25} + \texttt{log(hosp.beds)} + \texttt{crimes}^{-0.5} + \texttt{pct.hs.grad}^2 + \\ & \texttt{log(pct.bach.deg)} + \texttt{log(pct.below.pov)} + \texttt{log(pct.unemp)} + \\ & \texttt{tot.income}^{-0.25} + \texttt{region} \end{aligned} \tag{1}$$

$$\begin{aligned} \texttt{log(per.cap.income)} \sim\ & \texttt{log(land.area)} + \texttt{pop}^{0.2} + \texttt{pop.18\_34} + \texttt{log(hosp.beds)} + \\ & \texttt{crimes}^{-0.5} + \texttt{pct.hs.grad}^2 + \texttt{log(pct.bach.deg)} + \\ & \texttt{log(pct.below.pov)} + \texttt{log(pct.unemp)} + \texttt{tot.income}^{-0.25} \end{aligned} \tag{2}$$

	Values
log(land.area)	1.30
I(popˆ0.2)	16.45
pop.18_34	1.72
log(hosp.beds)	5.48
I(crimesˆ(-0.5))	3.05
I(pct.hs.gradˆ2)	3.57
log(pct.bach.deg)	4.39
log(pct.below.pov)	3.73
log(pct.unemp)	1.91
I(tot.incomeˆ-0.25)	26.12

Table 5: Variance Influcence Factors for Variables

Based on Table 5, there are several additional variables that should be removed due to a high VIF score. These variables are $\texttt{tot.income}^{-0.25}$ and $\texttt{pop}^{0.2}$. After removing these variables the final model that best clear and representative model for predicting Per Capita Income is written in Equation 3.

3.4 Examining Final Model

$$\begin{aligned} \texttt{log(per.cap.income)} \sim\ & \texttt{log(land.area)} + \texttt{pop.18\_34} + \texttt{log(hosp.beds)} + \\ & \texttt{crimes}^{-0.5} + \texttt{pct.hs.grad}^2 + \texttt{log(pct.bach.deg)} + \\ & \texttt{log(pct.below.pov)} + \texttt{log(pct.unemp)} \end{aligned} \tag{3}$$

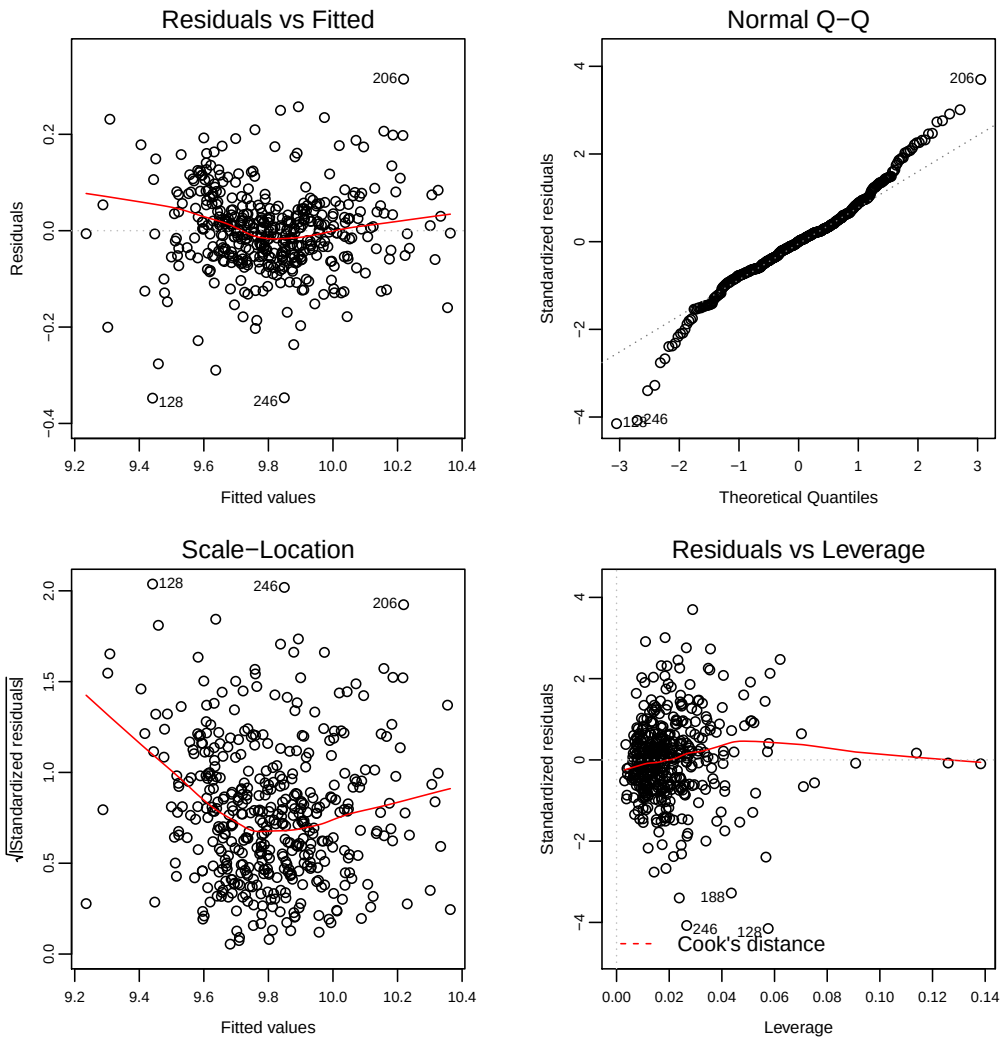


Figure 6: Diagnostic Plots for Final Model

Figure 6 depicts the diagnostic plots for Equation 3. The upper-left residual plot mostly has a constant amplitude and is centered around 0 which suggests the assumptions for this plot hold. The upper-right plot has a good chunk of the data following a normal distribution but the data does reveal skewness on both ends of the distribution which is cause for concern. The bottom-left plot does not reveal any pattern throughout the range of fitted values and the amplitude of the standardized residuals do not change which suggests the assumptions for this plot hold. The bottom-right plot indicates that none of the points are considered as outliers in our model. Overall, the resulting model appears to be statistically sound.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.43E+00	0.10	93.27	6.98E-288
log(land.area)	-3.35E-02	0.01	-6.66	8.53E-11
pop.18_34	-1.16E-02	0.00	-9.54	1.07E-19
log(hosp.beds)	5.87E-02	0.01	9.57	8.44E-20
I(crimes^-0.5)	-8.44E-01	1.16	-0.73	4.66E-01
I(pct.hs.grad^2)	-2.19E-05	0.00	-2.98	3.06E-03
log(pct.bach.deg)	3.31E-01	0.02	14.53	3.23E-39
log(pct.below.pov)	-2.27E-01	0.01	-19.29	3.34E-60
log(pct.unemp)	7.78E-02	0.02	4.65	4.52E-06

Table 6: Summary Coefficients for Final Model

The coefficients in Table 6 are interpretted as follows. While holding all other variables at a value of 0, the log transformed Per Capita income of a county has a value of 9.4330985. While holding all other variables constant, for a one percent increase in total land area of a county there is an expected -0.0334836 percent change in the Per Capita income for said county. While holding all other variables constant, for a one unit increase in the percentage of residents in a county between the ages of 18 and 34 there is an expected -0.0116304 change in the log transformed Per Capita income for said county. While holding all other variables constant, for a one unit increase in the log number of hospital beds in a county there is an expected 0.0586935 change in the log transformed Per Capita income for said county. While holding all other variables constant, for a one unit increase in $\frac{1}{\text{crime}^{0.5}}$ there is an expected -0.8436087 change in the log transformed Per Capita income for said county. While holding all other variables constant, for a one unit increase in squared percentage of adult residents who have completed at least 12 years of school there is an expected $-2.1877587 \times 10^{-5}$ change in the log transformed Per Capita income for said county. While holding all other variables constant, for a one percent increase in the percentage of adult residents who have earned a bachelor's degree in a county there is an expected 0.3313206 percent change in the transformed Per Capita income for said county. the log transformed Per Capita income for said county. While holding all other variables constant, for a one percent increase in the percentage of the population with income below the poverty line there is an expected -0.226856 percent change in the Per Capita income for said county. While holding all other variables constant, for a one percentage increase in the percentage of the population that is unemployed there is an expected 0.0778169 percent change in the Per Capita income for said county. For this model, the MSE and RMSE values are 71.65 and 8.46 respectively.

4 Discussion

Among the 13 predictor variables there were many that had some relationship to the main variable of interest, Per-Capita Income. Between the 13 variables there were a significant number of pairs that had some sort of relationship (linear/nonlinear) with one another. Section 3.1 explores this in-depth and discusses several relationships that potentially affect the model due to multi-colinearity issues and Section 3.3 attempts to construct a final model that to circumvent these colinearity issues.

Based on the results from Sections 2.3 and 3.2 there appears to be an intricate relationship between Per-Capita Income and Crime Rate. The full data shows there to be a negatively correlated relationship between the two variables but the individual scatterplots of these variables for each of the different regions of the country reveal otherwise. In Figures 4 (b) and 5 (b) there does exist some implicit positive relation between these variables contingent on the location of the county. These relationships are represented in Tables 3 and 4.

Section 3 discusses all methodology of choosing the final model which includes the circumvention of colinearity discussed previously. The final model for consideration is Equation 3. The transformations for the data appear just based on the scatterplots in Figure 3. The joint distribution plots look more normally distributed along the x-axis. The interpretability for each of these transformations is not too extreme either,

as discussed in Section 3.4. The assumptions for this model were mostly sound except for the normality assumption in the Q-Q plot in Figure 6 which was loosely held.

Since only 48 states were included in our dataset the results of our analysis should be taken with a grain of salt. Because we do not have access to the full geographic area of the United States the representative power of our model is not as useful. The missing states from the data are Alaska, Iowa and Wyoming so in these parts of the country where we do not have information the generalization ability of our model is very low. Furthermore, we know that the region of the country does have an impact on the Per-Capita Income for counties due to the discussion in Section 3.2, Thus, it would be in our interest to try and collect information on these states. If we consider that we used the most populous counties in the whole country then it should be alright to not have collected the 3300+ described in the Junker handout.

The analysis done in this report was conducted on only 440 observations during the course of 2 years. The results of this analysis would benefit greatly if the sample size were to grow not only in terms of increasing the time period the data was collected but also attempt to incldue the counties for the states that have missing data mentioned above. It would also be interesting to explore the relationship on region of the country more because the standard errors for the linear regression models for the seperate regions in Tables 3 and 4 are larger than the standard errors for the simple linear regression on the full dataset. These seems to suggest there is not enough observations for each of the individual regions to suggest a full relationship.

5 References

[BEA, 1998] BEA, B. o. E. A. (1998). *Regional Economic Analysis*. Geospatial & Statistical Data Center, University of Virginia, Charlottesvill, VA.

[Kutner, 2005] Kutner, M.H., N. (2005). *Applied Linear Statistical Models*. McGraw-Hill/Irwin.

[R, 2017] R, C. T. (2017). *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

6 Code Appendix

```
# Load packages
pkgs <- c('xtable', 'car')
lapply(pkgs, library, character.only=T)

# Load data
cdi <- read.table('./cdi.dat.txt')
i_vars <- c('land.area', 'pop', 'pop.18_34', 'pop.65_plus', 'doctors',
            'hosp.beds', 'crimes', 'pct.hs.grad', 'pct.bach.deg',
            'pct.below.pov', 'pct.unemp', 'per.cap.income',
            'tot.income', 'region')

cols <- c("#999999", "#E69F00", "#56B4E9", "#009E73",
          "#F0E442", "#0072B2", "#D55E00", "#CC79A7")

# Cache chunk options
opts_chunk$set(cache=T, autodep=T, cache.comments=F)
```

6.1 Exploratory Data Analysis

```
# Plot a 3x5 grid of plots of variables
par(mfrow=c(3,5), mar=c(2.5, 2.25, 0.5, 0.5) + 0.1, mgp=c(1.35, 0.5, 0),
    cex.axis=0.75, cex.lab=0.85, cex.main=0.75)
plot(cdi[,'state'], xlab='state', ylab='Frequency')
for (i in 4:16) {
  hist(cdi[,i], xlab=names(cdi)[i], main='')
}
plot(cdi[,'region'], xlab='region', ylab='Frequency')
```

```
# Displaying the summary statistics for the different numeric variables
var_summary <- t(apply(cdi[,4:16], 2, function(var) { c(summary(var), sd(var)) }))
colnames(var_summary)[7] <- 'Std. Dev'
print(xtable(var_summary, digits=c(0, 2, 2, 2, -2, 2, 2, -2), label='tab:var_summary',
             caption='Summary Statistics for Numeric Variables'),
      table.placement='H')
```

```
# Pairs correlation variable edited from pairs() documentation page
panel.cor <- function(x, y, cex.cor) {
  usr = par("usr")
  on.exit(par(usr))
  par(usr=c(0, 1, 0, 1))
  r = cor(x, y)
  txt = format(c(r, 0.123456789), digits=2)[1]
  txt = paste0('', txt)
  text(0.5, 0.5, txt)
}
```

```r
# Pairs plot of all numeric variables
pairs(cdi[,i_vars], lower.panel=panel.smooth, upper.panel=panel.cor,
      cex=0.1, cex.labels=0.52, gap=0)
```

```r
numeric_vars <- i_vars[which(i_vars != 'per.cap.income' & i_vars != 'region')]
# Plot a 6x2 grid of every variable vs per.cap.income
par(mfrow=c(6,2), mar=c(2.5, 2, 0.5, 0), mgp=c(1.25, 0.5, 0),
    cex.axis=0.75, cex.lab=0.90, cex.main=0.75)
for (i in seq_along(numeric_vars)) {
  if (! i %% 2) {
    with(cdi, plot(as.formula(paste('log(per.cap.income) ~', numeric_vars[i])),
                   yaxt='n', ylab=''))
  } else {
    with(cdi, plot(as.formula(paste('log(per.cap.income) ~', numeric_vars[i]))))
  }
  with(cdi, abline(lm(as.formula(paste('log(per.cap.income) ~', numeric_vars[i]))),
                   lwd=1.25, col='red'))
}

# Plot a 6x2 grid of every log transformed variable vs per.cap.income
with(cdi, plot(log(per.cap.income) ~ log(land.area)))
with(cdi, abline(lm(log(per.cap.income) ~ log(land.area)), lwd=1.25, col='red'))

with(cdi, plot(log(per.cap.income) ~ I(pop^0.2), yaxt='n', ylab=''))
with(cdi, abline(lm(log(per.cap.income) ~ I(pop^0.2)), lwd=1.25, col='red'))

with(cdi, plot(log(per.cap.income) ~ pop.18_34))
with(cdi, abline(lm(log(per.cap.income) ~ pop.18_34), lwd=1.25, col='red'))

with(cdi, plot(log(per.cap.income) ~ pop.65_plus, yaxt='n', ylab=''))
with(cdi, abline(lm(log(per.cap.income) ~ pop.65_plus), lwd=1.25, col='red'))

with(cdi, plot(log(per.cap.income) ~ I(doctors^-0.25)))
with(cdi, abline(lm(log(per.cap.income) ~ I(doctors^-0.25)), lwd=1.25, col='red'))

with(cdi, plot(log(per.cap.income) ~ log(hosp.beds), yaxt='n', ylab=''))
with(cdi, abline(lm(log(per.cap.income) ~ log(hosp.beds)), lwd=1.25, col='red'))

with(cdi, plot(log(per.cap.income) ~ I(crimes^-0.5)))
with(cdi, abline(lm(log(per.cap.income) ~ I(crimes^-0.5)), lwd=1.25, col='red'))

with(cdi, plot(log(per.cap.income) ~ I(pct.hs.grad^2), yaxt='n', ylab=''))
with(cdi, abline(lm(log(per.cap.income) ~ I(pct.hs.grad^2)), lwd=1.25, col='red'))

with(cdi, plot(log(per.cap.income) ~ log(pct.bach.deg)))
with(cdi, abline(lm(log(per.cap.income) ~ log(pct.bach.deg)), lwd=1.25, col='red'))

with(cdi, plot(log(per.cap.income) ~ log(pct.below.pov), yaxt='n', ylab=''))
with(cdi, abline(lm(log(per.cap.income) ~ log(pct.below.pov)), lwd=1.25, col='red'))
```

```r
with(cdi, plot(log(per.cap.income) ~ log(pct.unemp)))
with(cdi, abline(lm(log(per.cap.income) ~ log(pct.unemp)), lwd=1.25, col='red'))

with(cdi, plot(log(per.cap.income) ~ I(tot.income^-0.25), yaxt='n', ylab=''))
with(cdi, abline(lm(log(per.cap.income) ~ I(tot.income^-0.25)), lwd=1.25, col='red'))
```

6.2 Visually Exploring the Effect of Crime

```r
# Create crime rate varaible
cdi$crime.rate = with(cdi, crimes / pop)

# Plot per-capita income vs crime rate
par(mar=c(3.25, 2.75, 0, 0.5) + 0.1, mgp=c(2, 0.5, 0),
    cex.axis=0.95, cex.lab=0.95, cex.main=0.75)
with(cdi, plot(per.cap.income ~ crime.rate, cex=0.75, col=region,
               xlab='Crime Rate', ylab='Per-Capita Income'))
abline(lm_crime <- with(cdi, lm(per.cap.income ~ crime.rate), col=cols[1], lwd=1.5))

# Add legend for plot clarity
uni_regions <- unique(cdi$region)
legend('topright', c('Linear Model', 'West', 'North Central', 'South', 'North East'),
       lty=c(1, rep(0, 4)), pch=c(26, rep(1, 4)), lwd=c(1.5, 0, 0),
       col=c(cols[1], 1:4), cex=0.85, title='Legend')

# Plot grid of crime vs per-capita income for each region
## Plotting the West region values
par(mfrow=c(2,2), oma=c(4, 3.5, 0, 0), mar=c(0, 0, 0, 0) + 0.1, mgp=c(2, 0.5, 0),
    cex.axis=0.95, cex.lab=1.05, cex.main=0.75)
i_w <- which(cdi$region == 'W')
with(cdi[i_w,], plot(per.cap.income ~ crime.rate, cex=0.75, xaxt='n',
       xlim=range(cdi$crime.rate), ylim=range(cdi$per.cap.income)))
abline(lm_w <- with(cdi[i_w,], lm(per.cap.income ~ crime.rate)), lwd=1.5, col='red')
text(0.27, 35000, 'West')

## Plotting the North Carolina region values
i_nc <- which(cdi$region == 'NC')
with(cdi[i_nc,], plot(per.cap.income ~ crime.rate, cex=0.75, xaxt='n', yaxt='n',
       xlim=range(cdi$crime.rate), ylim=range(cdi$per.cap.income)))
abline(lm_nc <- with(cdi[i_nc,], lm(per.cap.income ~ crime.rate)), lwd=1.5, col='red')
text(0.25, 35000, 'North Central')

## Plotting the South region values
i_s <- which(cdi$region == 'S')
with(cdi[i_s,], plot(per.cap.income ~ crime.rate, cex=0.75,
       xlim=range(cdi$crime.rate), ylim=range(cdi$per.cap.income)))
abline(lm_s <- with(cdi[i_s,], lm(per.cap.income ~ crime.rate)), lwd=1.5, col='red')
text(0.27, 35000, 'South')

## Plotting the North East region values
```

```r
i_ne <- which(cdi$region == 'NE')
with(cdi[i_ne,], plot(per.cap.income ~ crime.rate, cex=0.75, yaxt='n',
       xlim=range(cdi$crime.rate), ylim=range(cdi$per.cap.income)))
abline(lm_ne <- with(cdi[i_ne,], lm(per.cap.income ~ crime.rate)), lwd=1.5, col='red')
text(0.26, 35000, 'North East')

# Add global axis to plot
title(xlab='Crime Rate', outer=T, line=2.25)
title(ylab='Per-Capita Income', outer=T, line=2.25)
```

```r
# Create crime rate varaible
cdi$crime.rate = with(cdi, crimes / pop)

# Plot log per-capita income vs log crime rate
par(mar=c(3.25, 2.75, 0, 0.5) + 0.1, mgp=c(2, 0.5, 0),
    cex.axis=0.95, cex.lab=0.95, cex.main=0.75)
with(cdi, plot(log(per.cap.income) ~ log(crime.rate), cex=0.75, col=region,
               xlab='Crime Rate', ylab='Per-Capita Income'))
abline(lm_crime_log <- with(cdi, lm(log(per.cap.income) ~ log(crime.rate)),
                            col=cols[1], lwd=1.5))

# Add legend for plot clarity
uni_regions <- unique(cdi$region)
legend('topright', c('Linear Model', 'West', 'North Central', 'South', 'North East'),
       lty=c(1, rep(0, 4)), pch=c(26, rep(1, 4)), lwd=c(1.5, 0, 0),
       col=c(cols[1], 1:4), cex=0.85, title='Legend')

# Plot grid of crime vs per-capita income for each region
## Plotting the West region values
par(mfrow=c(2,2), oma=c(4, 3.5, 0, 0), mar=c(0, 0, 0, 0) + 0.1, mgp=c(2, 0.5, 0),
    cex.axis=0.95, cex.lab=1.05, cex.main=0.75)
with(cdi[i_w,], plot(log(per.cap.income) ~ log(crime.rate), cex=0.75, xaxt='n',
       xlim=range(log(cdi$crime.rate)), ylim=range(log(cdi$per.cap.income))))
abline(lm_w_log <- with(cdi[i_w,], lm(log(per.cap.income) ~ log(crime.rate))),
       lwd=1.5, col='red')
text(-1.6, 10.4, 'West')

## Plotting the North Carolina region values
with(cdi[i_nc,], plot(log(per.cap.income) ~ log(crime.rate), cex=0.75, xaxt='n', yaxt='n',
       xlim=range(log(cdi$crime.rate)), ylim=range(log(cdi$per.cap.income))))
abline(lm_nc_log <- with(cdi[i_nc,], lm(log(per.cap.income) ~ log(crime.rate))),
       lwd=1.5, col='red')
text(-1.8, 10.4, 'North Central')

## Plotting the South region values
with(cdi[i_s,], plot(log(per.cap.income) ~ log(crime.rate), cex=0.75,
       xlim=range(log(cdi$crime.rate)), ylim=range(log(cdi$per.cap.income))))
abline(lm_s_log <- with(cdi[i_s,], lm(log(per.cap.income) ~ log(crime.rate))),
       lwd=1.5, col='red')
text(-1.6, 10.4, 'South')
```

```
## Plotting the North East region values
with(cdi[i_ne,], plot(log(per.cap.income) ~ log(crime.rate), cex=0.75, yaxt='n',
       xlim=range(log(cdi$crime.rate)), ylim=range(log(cdi$per.cap.income))))
abline(lm_ne_log <- with(cdi[i_ne,], lm(log(per.cap.income) ~ log(crime.rate))),
       lwd=1.5, col='red')
text(-1.7, 10.4, 'North East')

# Add global axis to plot
title(xlab='Crime Rate', outer=T, line=2.25)
title(ylab='Per-Capita Income', outer=T, line=2.25)
```

6.3 Examining Effect of Crime on Per-Capita Income

```
# Generating coefficients for all linear models
lms <- list(lm_crime, lm_w, lm_nc, lm_s, lm_ne)
log_lms <- list(lm_crime_log, lm_w_log, lm_nc_log, lm_s_log, lm_ne_log)

get_coefs <- function(acc, elem) {
  acc <- c(acc, coef(summary(elem))[,c(1,2,4)])
}

lms_coefs <- matrix(Reduce(get_coefs, lms, c()), nrow=length(lms), byrow=T)
dimnames(lms_coefs) <-
  list(c('All', 'W', 'NC', 'S', 'NE'), c('(Intercept)', 'Crime Rate',
       'Intercept Std Err', 'Crime Rate Std Err', 'Intercept P-value',
       'Crime Rate P-value'))
print(xtable(t(lms_coefs), label='tab:lms_coefs',
             caption='Summary Coefficients for Log Per-Capita Income on Crime Rate'),
      table.placement='H')

log_coefs <- matrix(Reduce(get_coefs, log_lms, c()), nrow=length(lms), byrow=T)
dimnames(log_coefs) <-
  list(c('All', 'W', 'NC', 'S', 'NE'), c('(Intercept)', 'Log Crime Rate',
       'Intercept Std Err', 'Log Crime Rate Std Err', 'Intercept P-value',
       'Log Crime Rate P-value'))
print(xtable(t(log_coefs), label='tab:log_coefs',
             caption='Summary Coefficients for Log Per-Capita Income on Log Crime Rate'),
      table.placement='H')
```

6.4 Examining Variables for Selection

```
# Build initial model
lm_init <- with(cdi, lm(log(per.cap.income) ~ log(land.area) + I(pop^0.2) + pop.18_34 +
  pop.65_plus + I(doctors^-0.25) + log(hosp.beds) + I(crimes^(-0.5)) + I(pct.hs.grad^2) +
  log(pct.bach.deg) + log(pct.below.pov) + log(pct.unemp) + I(tot.income^-0.25) + region))

# Filter variables using stepwise algorithm
```

```
lm_step <- step(lm_init, direction='both', k=log(nrow(cdi)), trace=F)

# Calculate vif scores and display values
vif_tab <- as.matrix(vif(lm_step))
colnames(vif_tab) <- c('Values')
print(xtable(vif_tab, label='tab:vif',
             caption='Variance Influcence Factors for Variables'),
      table.placement='H')
```

6.5 Examining Final Model

```
# Build final model
lm_final <- with(cdi, lm(log(per.cap.income) ~ log(land.area) + pop.18_34 +
  log(hosp.beds) + I(crimes^-0.5) + I(pct.hs.grad^2) + log(pct.bach.deg) +
  log(pct.below.pov) + log(pct.unemp)))

# Plot diagnostic graphs
par(mfrow=c(2,2), mar=c(3, 3, 2, 2.25) + 0.1, mgp=c(1.75, 0.5, 0),
    cex.axis=0.75, cex.lab=0.85, cex.main=0.75)
plot(lm_final)
```

```
# Printing coefficient estimates
final_coefs <- coef(summary(lm_final))
print(xtable(final_coefs, label='tab:final_coefs', digits=c(0, -2, 2, 2, -2),
             caption='Summary Coefficients for Final Model'),
      table.placement='H')

# Calculate MSE and RMSE for model
mse <- with(cdi, mean(per.cap.income - exp(lm_final$fitted.values)))
rmse <- sqrt(mse)
```