An analysis of factors affecting per capita income in the United States

Bhoomika Moorjani
bmoorjan@andrew.cmu.edu

18 October 2021

ABSTRACT

Per capita income is an important determinant of economic development in different regions of a country. This study aims to study the factors that affect per capita income based on county demographic information for 440 of the most populous counties in the United states for the years 1990 and 1992. We found that there is a weak positive relationship between per capita income and crime which doesn't change with region but the effect of change in crime on per capita income varies with region. Among other variables, percent of population aged between 18 and 34, land area, percentage of population below poverty line have an inverse relationship with per capita income while number of doctors, percent unemployed and percent with bachelor's degree are positively associated with per capita income. However, since the data only provides information about 440 of the most populous counties in the United States, we can't be certain about drawing these conclusions about all 3006 counties.

INTRODUCTION

Per capita income i.e., the ratio of total personal income to the total population is an important measure of the standard of living of a population. Since different regions of the United States aren't equally developed, per-capita income is an important tool that is used by economists to compare the relative performance of different regions. This study tries to determine the effect of 11 variables associated with the county's economic, health and social well-being on per capita income and how effective they are in changing the per capita income between counties. This paper will address the following questions:

Part One: How are the demographic variables in the dataset related to each other?
Part Two: How is per capita income of a county related to the number of crimes and crime rate ?
Part Three: How can we predict per capita income of a county from variables associated with its economic, health and social well-being?

DATA

The data taken from Kutner et al. (2005) provides county demographic information (CDI) for 440 of the most populous counties in the United States for the years 1990 and 1992. Each line of dataset has an identification number, county name, state abbreviation and provides information on 14 variables for each county. Counties with missing variables were deleted from the dataset. Data is available in the file cdi.dat which can be accessed from the Project 1 folder on Canvas.

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

Table 1

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 1990 CDI population aged 18-34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 or older
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)

METHODS

Our analysis, consisting of four parts, was carried out using the R language and environment for statistical computing.

Part One: We visually compared summary statistics (Table 2 and 3 in appendix 1), histograms (Figure 1 and 5 in appendix 1), correlation plots (Figure 2 and 6 in appendix 1) , scatter plot (Figure 3 and 7 in appendix 1) and box plots (Figure 4 in appendix 1). Then we leveraged log transformation to clean up skewness in the variables.

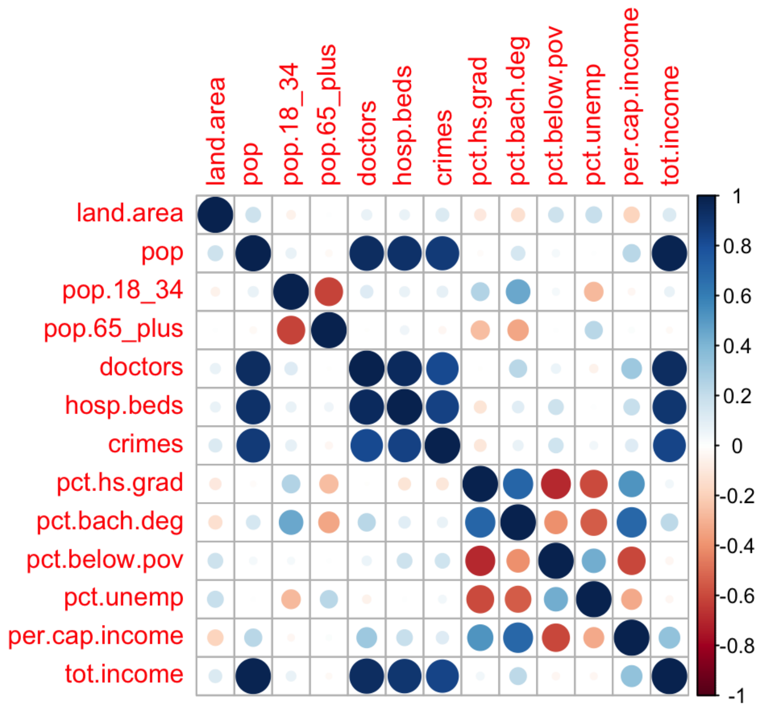
Part Two: To analyse the relationship between per capita income and crime, we tried fitting a linear regression model to the data with log(per capita income) as our response variable and log(crime) as our explanatory variable (Model 1.1 in appendix 2). Next, we included the region variable in our model to examine its effect on the relationship between per capita income and crime. Since region is a categorical variable, we considered two situations: one, where it produces an additive change in per capita income (Model 1.2 in appendix 2) and the second, where it changes the size of effect of number of crimes on per capita income (Model 2.3 in appendix 2).
Crime rate which is crime on a per capita basis is often used to make comparisons between regions as it adjusts for population size. We tried using log(crime rate) instead of log(crimes) in the models mentioned earlier to see if it is a better predictor of per capita income (Models 2.1, 2.2, 2.3 in appendix 2). We then used the model selection criteria including Analysis of Variance (ANOVA) test, Akaike's Information Criteria (AIC), Bayesian Information Criteria (BIC) and diagnostic plots to select the model that best fits the data.

Part Three: We used stepwise regression, a step-by-step iterative process of constructing a model that involves selection of potential explanatory variables to predict per capita income and testing for statistical significance after each iteration to arrive at the final model (Model 3.2 in appendix 3).
We then calculated the variance inflation factor (VIF) to measure the amount of multicollinearity in our set of explanatory variables. Additionally, we visually inspected the diagnostic plots, marginal model plot and added variable plot to examine how well our model fits the data. Similar to part two, the region variable was included in our model to examine its effect on the relationship between per capita income

and all other variables (Model 3.3 in appendix 3). We further improved this model by excluding the explanatory variables which didn't have statistically significant coefficients (Model 3.4 in appendix 3). We then compared Model 3.4 i.e, with region variable and Model 3.2 i.e, without region variable to see which one was a better fit.

RESULTS

Part One:



We can draw the following conclusions from the above correlation plot(figure 2 in appendix 1):

1. **Population and Total Income:** These two variables are highly correlated as expected. Counties with larger populations will be able to generate more income.
2. **Number of active physicians, number of hospital beds and number of serious crimes:** These three variables are positively correlated to each other and with population and total income as expected. Counties with larger populations will need healthcare infrastructure with greater capacity and the volume of crimes in these counties will be higher.
3. **Land Area**: One might think that the land mass of a county would enable the county to produce more total income and as a result more per capita income than the others. However, the findings of this paper conclude that land area is not related to total income and has a very weak negative correlation with per capita income.
4. **Percent of population aged 18-34 and 65 or older:** These two variables are negatively correlated as both are a subset of the total population. These are also related to the percentage of the population with a bachelor's degree, although the correlation isn't very strong. If the population is younger i.e., percentage of the population aged 18-34 is higher, more people will have a bachelor's degree.

5. **Percent of high school graduates and percent with bachelor's degrees:** These two variables are positively correlated to each other and per capita income. Also, they are negatively correlated to the percentage of population below the poverty line and percentage of population unemployed. Counties where more people live below the poverty line, fewer people graduate from high school and get bachelor's degrees. As a result, they won't be able to find jobs and hence more people will be unemployed. On the other hand, in counties with lower percentages of population living in poverty, more people will graduate from high school, get bachelor's degrees, find jobs and per capita income will be higher.

Part Two:

Model 1.2 in Appendix 2 : Log (Per Capita Income) = Baseline + 0.07 * Log (Crime)

According to the above model, we can expect a 0.07% increase in the baseline per capita income for every 1% increase in the number of crimes. The intercept in our model i.e., baseline per capita income varies for different regions and is as follows:
S Region = USD 8,955.293
W Region = USD 9,228.022
NC Region = USD 9,798.651
NE Region = USD 10,509.13
The coefficients for interaction between region and crime had very low p-values i.e, were statistically significant. Therefore, all of these region baselines are significantly different from the NC baseline. The model reports a R-squared value of 0.2032 which suggests that the association between per capita income and crime is not very strong.

Since crime rate is often used by Economists to make comparisons between regions as it adjusts for population size, we tried using log(crime rate) instead of log(crimes) in the above model.

Model 2.2 in Appendix 2 : Log (Per Capita Income) = Baseline + 0.04 * Log (Crime rate)

According to the second model, we can expect a 0.04% increase in the baseline per capita income for every 1% increase in the crime rate. The baseline per capita incomes by region are:
S Region = USD 19,341.34
W Region = USD 20,332.99
NC Region = USD 20,743.74
NE Region = USD 23,155.79
The second model reports a R-squared value of 0.09 which is lower than the previous one and the coefficient of crime rate is not statistically significant. Additionally, since the AIC and BIC values for the first model were lower, we chose the first model.

Part Three:

7 variables selected by stepwise regression to predict log(per capita income) were log(land.area), log(pop.18_34), log(pop.65_plus), log(doctors), log(pct.bach.deg), log(pct.below.pov), log(pct.unemp)

We introduced interactions with region variables and eliminated the terms that were not statistically significant including log(land.area) * region, log(doctors)* region and log(pct.bach.deg) * region.

We compared the two models - with and without interaction with region variable using ANOVA test, AIC and BIC. ANOVA test and AIC are in favour of the model with region terms but BIC seems to favour the smaller model without region variable.

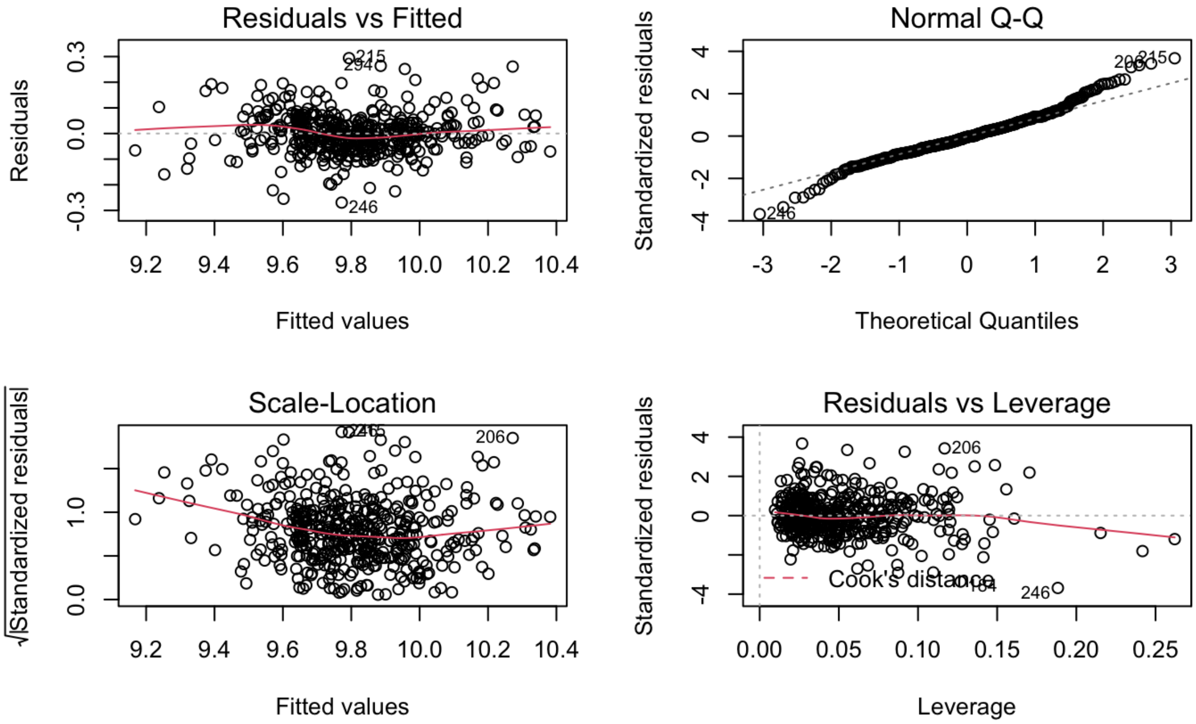
We chose the below model with region interaction terms:

Model 3.2 in Appendix 3 : log(per.cap.income) ~ log(land.area) + log(pop.18_34) * log(region) + log(pop.65_plus) * region + log(doctors) + log(pct.bach.deg) + log(pct.below.pov) * region + log(pct.unemp) * region + region

Below table summarises statistically significant coefficients from our final model:

1% ⬆ in below explanatory variable	**Effect on per capita income** *Baseline per capita income = USD 31,257.04 (NC, NE, S) and USD 6,634.244 (W)*
Land area	0.04% ⬇
Percent of population aged 18-34	0.38% ⬇ (NC, NE, S) 0.01% ⬆ (W)
Doctors	0.06% ⬆
Percent with bachelor’s degree	0.25% ⬆
Percent below poverty line	0.16% ⬇
Percent unemployed	0.11% ⬆ (NC, NE, W) 0.07% ⬇ (S)

The model reports a R-squared value of 0.8577. The below diagnostic plots suggest that modelling assumptions are satisfied and the model is a good fit for the data.



DISCUSSION

In the first part of this paper, we've summarized the relationships between 13 demographic variables for 440 most populous counties in the United States for the year 1990. From our model in part 2, we can conclude that there is a weak positive relationship between per capita income and crime which doesn't change with region but the effect of change in crime on per capita income varies with region. In the third part, we've created a model to predict the change in per capita income as a result of changes in variables associated with the economic, health and social well-being of the county.

While comparing this relationship for different regions, we notice that the West seems to have a significantly lower baseline per capita income compared to the other regions. From table 3 in appendix 1, we observe that West has the least number of counties sampled but they cumulatively cover the maximum land area. This could be because the counties in the west are larger in terms of land area covered. Based on this, we might conjecture that these counties are mostly rural areas where farming, animal husbandry and/or mining are the primary economic activities which typically generate lower per capita income. This would also explain why we can expect a decrease of 0.04% in the per capita income for an increase of 1% in land area.

An increase of 1% in the percentage of population aged 18-34 is associated with a 0.38% decrease in per capita income across the US except for in the West where the baseline per capita income is significantly lower to begin with. We might conjecture that this segment of the population is not at their peak earning

capacity or are mainly employed in blue-collar jobs and their lower incomes bring down the average income.
For every 1% increase in doctors, we can expect a 0.06% increase in per capita income. This could be because doctors earn high incomes which drives up the per capita income in the county. Alternatively, a county having more doctors could mean the population of the county is large which is why the total income and hence per capita income is higher. On the contrary, counties with a higher percentage of population living below the poverty line tend to have lower per capita income as people living below the poverty line don't make significant contributions to the total income of the county.
For every 1% increase in percentage of population with a bachelor's degree, we can expect a 0.25% increase in per capita income as this segment of the population will be able to find high-paying jobs.

Since the data provides information about only the 440 of the most populous counties, we can't be sure about making inferences for all the 3006 counties in the US based on our analysis of the sample dataset. We need to check if our sample dataset is representative of the population. We can do this by comparing the values of demographic indicators in this dataset to their national average.

Technical Appendix

APPENDIX 1

```
#Table 2 and 3
cdinumeric <- cdidata[,-c(1,2,3,17)]
apply(cdinumeric,2,function(x) c(summary(x),SD=sd(x))) %>% as.data.frame %>% t() %>%
  round(digits=2) %>% kbl(booktabs=T,caption="Table 2") %>% kable_classic()
```

```
eda_region <- cdidata %>% group_by(region) %>%
  summarise("Number of Counties" = length(county), "Number of States" = length(unique(state)), "Land Are
eda_region %>% kbl(booktabs=T,caption="Table 3") %>% kable_classic(full_width=F)
```

There are several variables in Table 2 with mean greater than median indicating possible right-skewing. For the region variable (Table 3), West has the least number of counties sampled but they cumulatively cover the maximum land area. This could be because the counties in the west are larger in terms of land area covered. Similarly, number of counties sampled is the highest in South most likely due to small size of counties in the region.

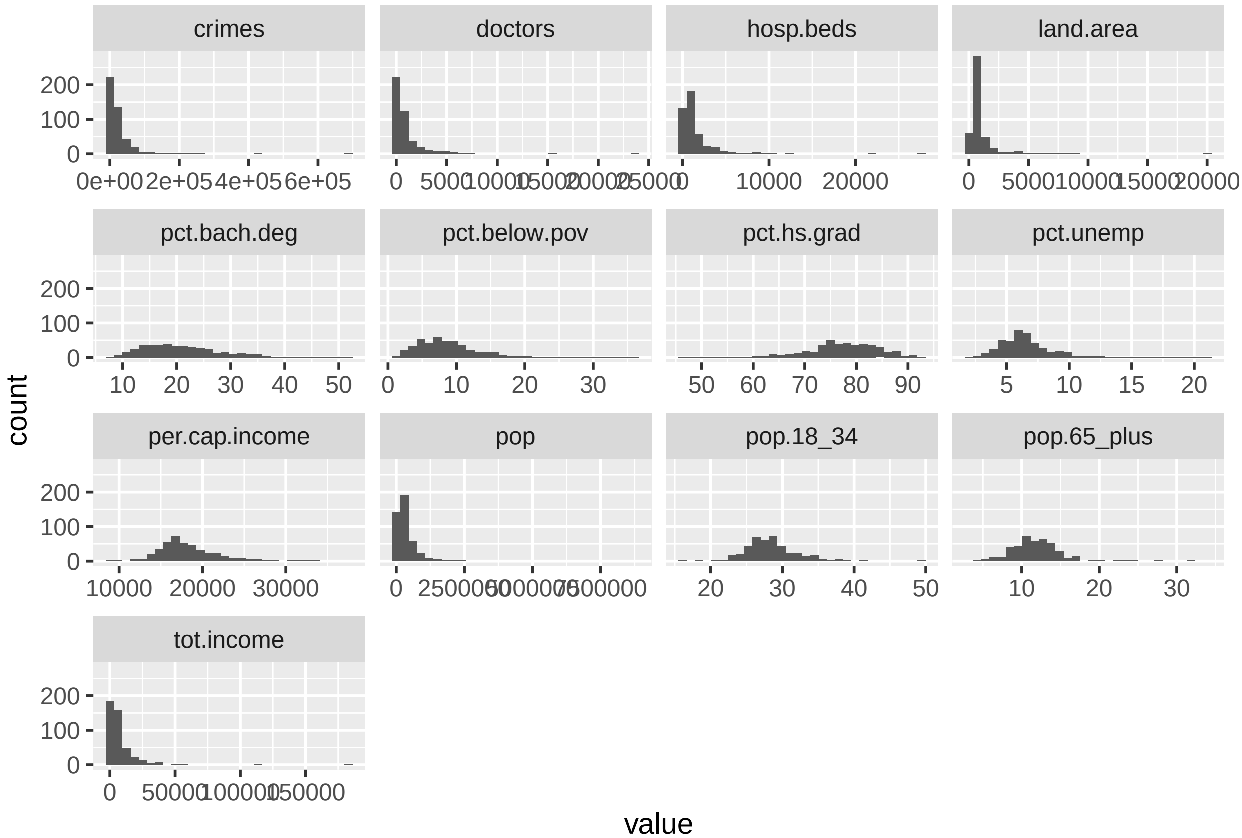
```
#Figure 1
cdigood <- data.frame(cdinumeric,region=cdidata$region)
ggplot(gather(cdinumeric), aes(value)) +
  geom_histogram(bins=30) +
  facet_wrap(~key, scales = 'free_x')
```

Table 1: Table 2

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
land.area	15.0	451.25	656.50	1041.41	946.75	20062.0	1549.92
pop	100043.0	139027.25	217280.50	393010.92	436064.50	8863164.0	601987.02
pop.18_34	16.4	26.20	28.10	28.57	30.02	49.7	4.19
pop.65_plus	3.0	9.88	11.75	12.17	13.62	33.8	3.99
doctors	39.0	182.75	401.00	988.00	1036.00	23677.0	1789.75
hosp.beds	92.0	390.75	755.00	1458.63	1575.75	27700.0	2289.13
crimes	563.0	6219.50	11820.50	27111.62	26279.50	688936.0	58237.51
pct.hs.grad	46.6	73.88	77.70	77.56	82.40	92.9	7.02
pct.bach.deg	8.1	15.28	19.70	21.08	25.33	52.3	7.65
pct.below.pov	1.4	5.30	7.90	8.72	10.90	36.3	4.66
pct.unemp	2.2	5.10	6.20	6.60	7.50	21.3	2.34
per.cap.income	8899.0	16118.25	17759.00	18561.48	20270.00	37541.0	4059.19
tot.income	1141.0	2311.00	3857.00	7869.27	8654.25	184230.0	12884.32

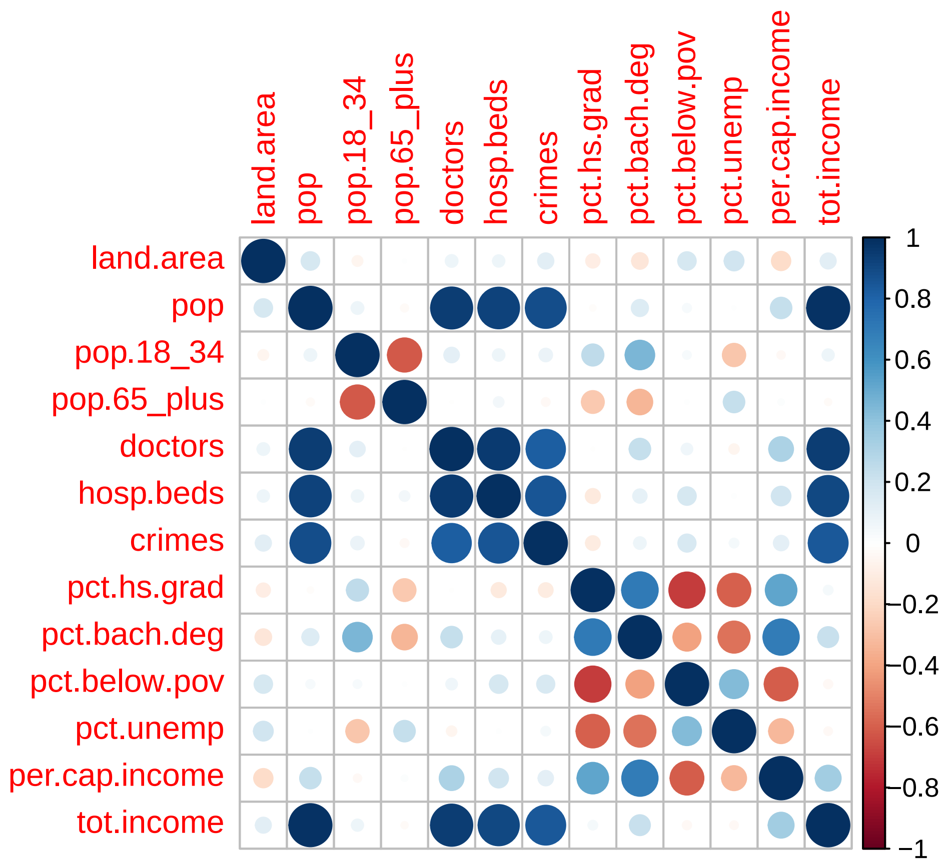
Table 2: Table 3

region	Number of Counties	Number of States	Land Area	Population
NC	108	11	68372	37386529
NE	103	10	67518	40770956
S	152	16	110446	50008592
W	77	11	211885	44758728



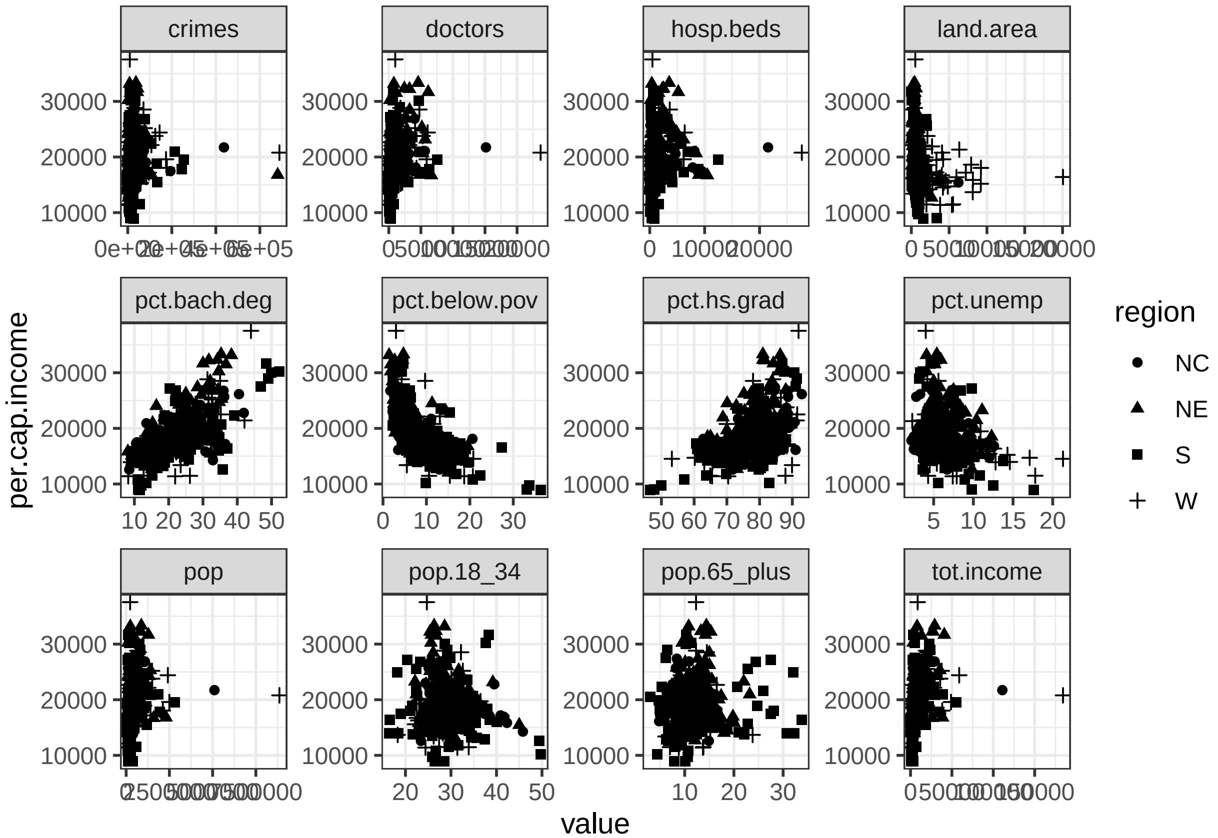
The histograms in figure 1 suggest some of the variables including crimes, doctors, hosp.beds, land.area, pop and tot.income are severely right skewed.

```
# Figure 2
corrplot(cor(cdinumeric))
```



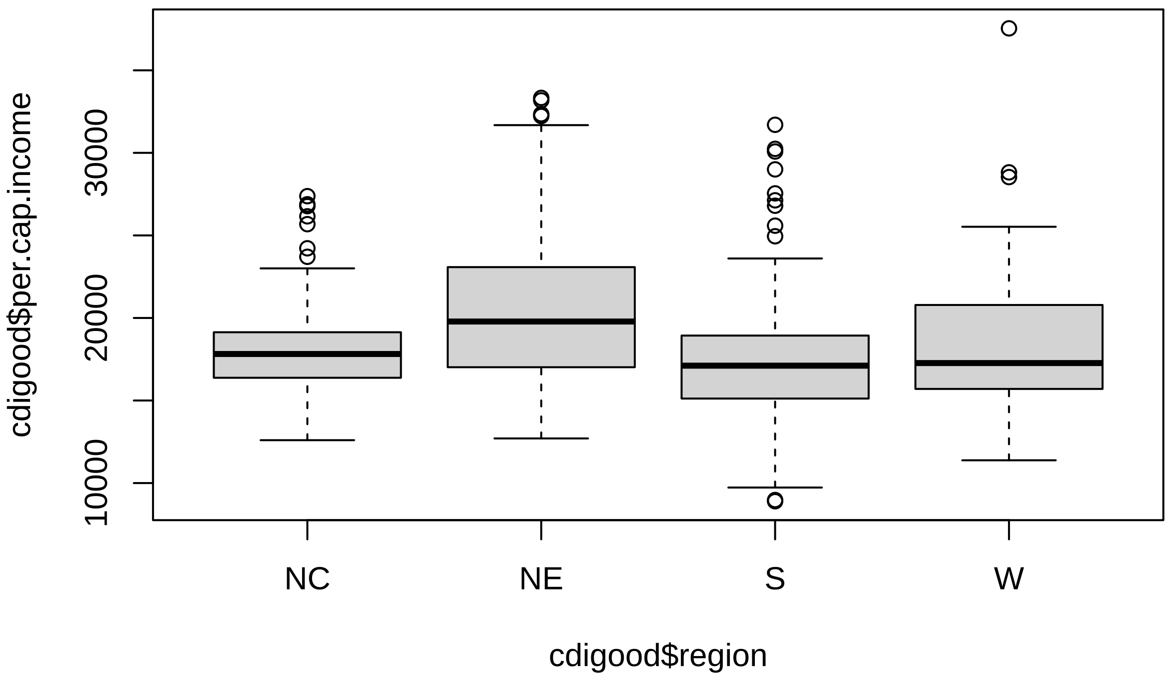
From the correlation plot in figure 2, we observe that : (i) tot.income and pop are highly correlated (ii) both are reasonably correlated with crimes, hosp.beds and doctors (iii) the three variables crimes, hosp.beds and doctors seem strongly correlated with one another (iv) pct.hs.grad and pct.bach.deg are moderately correlated with one another and positively correlated with per.cap.income (v) pct.below.pov and pct.unemp are moderately correlated with one another and negatively correlated with per.cap.income, pct.hs.grad and pct.bach.deg

```
#Figure 3
cdigood %>%
  gather(-per.cap.income,-region, key = "var", value = "value") %>%
  ggplot(aes(x=value, y = per.cap.income, shape=region))+
  geom_point() +
  facet_wrap(~var, scales = "free") +
  theme_bw()
```



The scatter plots in figure 3 suggest pct.hs.grad, pct.bach.deg, pct.below.pov and pct.unemp are going to be most effective in predicting per.cap.income which is in line with our conclusion from the correlation plot in figure 2.

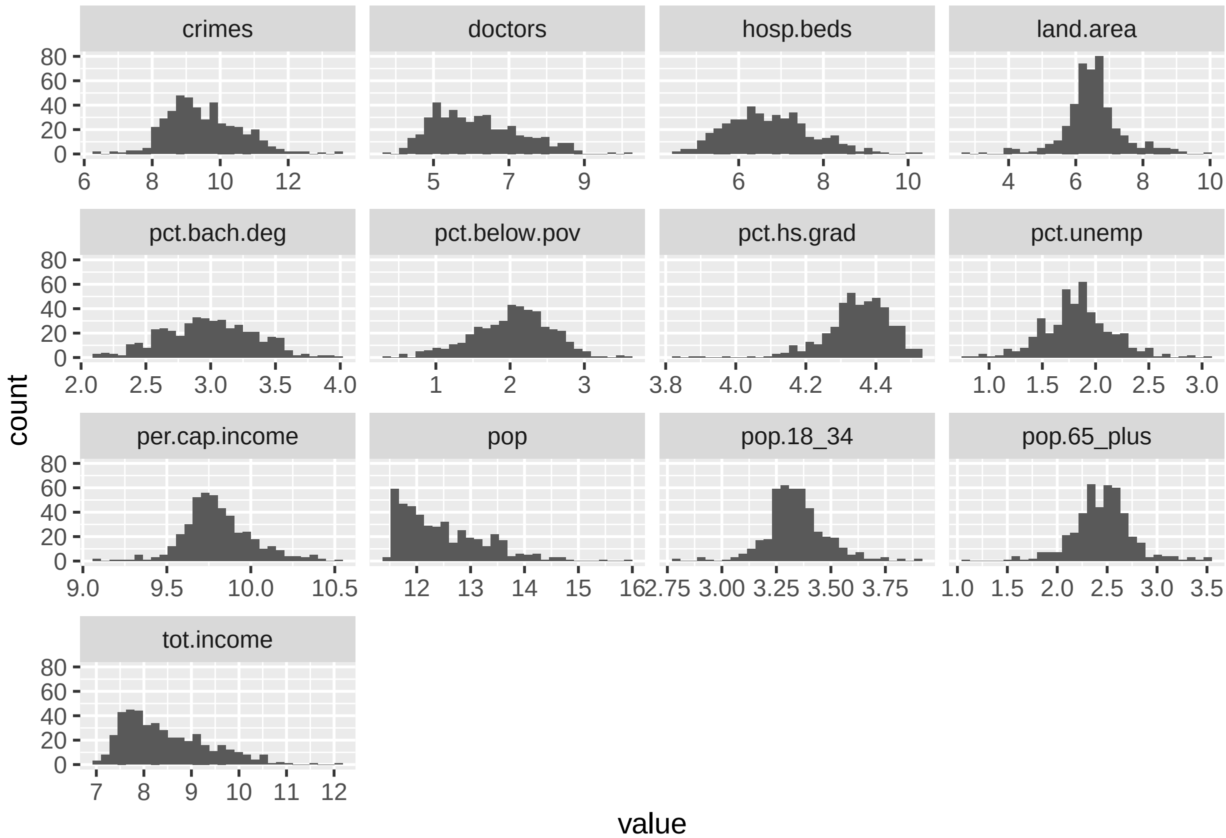
```
#Figure 4
boxplot(cdigood$per.cap.income~cdigood$region)
```



The box plot in figure 4 suggests there is greater variability in the per.cap.income in the North East and West regions. There are a few large outliers in the North Central and South regions which need to be investigated.

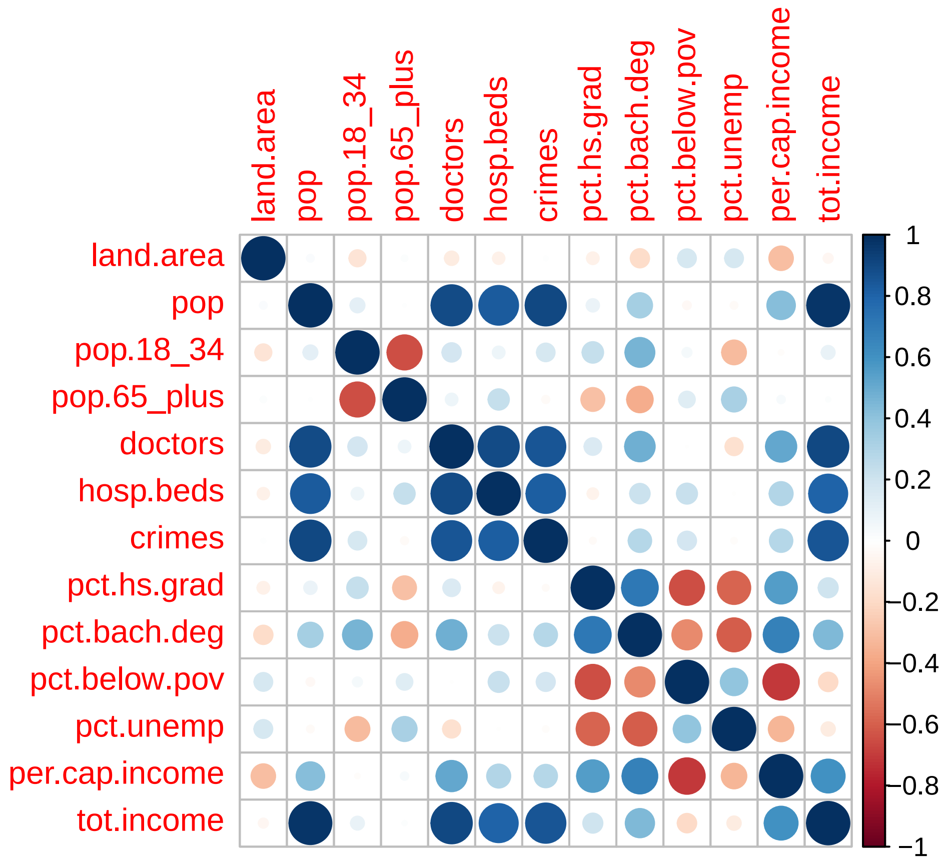
To address heavy skewing in some of the variables identified in figure 1, we have log transformed them. To ensure we can interpret the coefficients in the regression models consistently as percentage change in per.cap.income for a 1% change in the corresponding explanatory variable, we've log transformed the variables with minor skewing as well.

```
#Figure 5
cdilogs <- data.frame(log(cdinumeric))
ggplot(gather(cdilogs), aes(value)) +
  geom_histogram(bins=30) +
  facet_wrap(~key, scales = 'free_x')
```



The histograms in figure 5 suggest, log transformations have brought the skewing under control in all the variables except pop and tot.income. Since per.cap.income = tot.income / pop, we are going to be using only per.cap.income and don't need to work on reducing the skewness in tot.income and pop.

```
#Figure 6
corrplot(cor(cdilogs))
```




```
#Figure 7
cdilogsgood <- data.frame(cdilogs, region=cdidata$region)
cdilogsgood %>%
  gather(-per.cap.income,-region, key = "var", value = "value") %>%
  ggplot(aes(x=value, y = per.cap.income, shape=region))+
  geom_point() +
  facet_wrap(~var, scales = "free") +
  theme_bw()
```

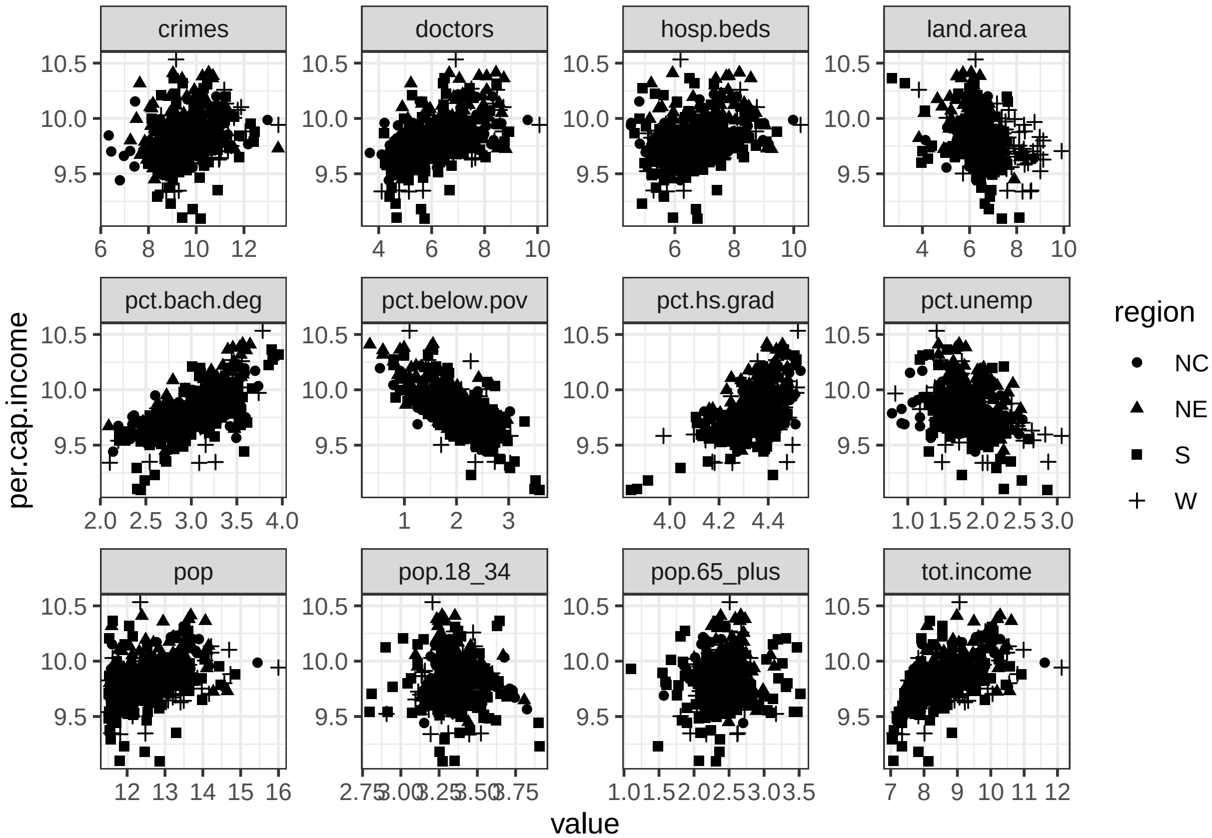


Figure 6 and 7 indicate stronger relationship between transformed variables compared to figure 2 and 3 for untransformed variables.

APPENDIX 2

```
#Crimes
model1.1 <- lm(cdilogsgood$per.cap.income ~ cdilogsgood$crimes)
model1.2 <- lm(cdilogsgood$per.cap.income ~ cdilogsgood$crimes + cdilogsgood$region)
model1.3 <- lm(cdilogsgood$per.cap.income ~ cdilogsgood$crimes * cdilogsgood$region)
anova(model1.1,model1.2,model1.3)
```

```
## Analysis of Variance Table
##
## Model 1: cdilogsgood$per.cap.income ~ cdilogsgood$crimes
## Model 2: cdilogsgood$per.cap.income ~ cdilogsgood$crimes + cdilogsgood$region
## Model 3: cdilogsgood$per.cap.income ~ cdilogsgood$crimes * cdilogsgood$region
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    438 17.271
## 2    435 14.949  3   2.32194 22.4823 1.523e-13 ***
## 3    432 14.872  3   0.07678  0.7434    0.5266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The additive model1.2 with no interactions is the best as it has the lowest p-value.

```
#Crime Rate
cdilogsgood$crime.rate <- cdilogsgood$crimes - cdilogsgood$pop
model2.1 <- lm(cdilogsgood$per.cap.income ~ cdilogsgood$crime.rate)
model2.2 <- lm(cdilogsgood$per.cap.income ~ cdilogsgood$crime.rate + cdilogsgood$region)
model2.3 <- lm(cdilogsgood$per.cap.income ~ cdilogsgood$crime.rate * cdilogsgood$region)
anova(model2.1,model2.2,model2.3)
```

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

The additive model2.2 with no interactions is the best as it has the lowest p-value.

```
#Comparing 1.2 and 2.2
AIC(model1.2, model2.2)
```

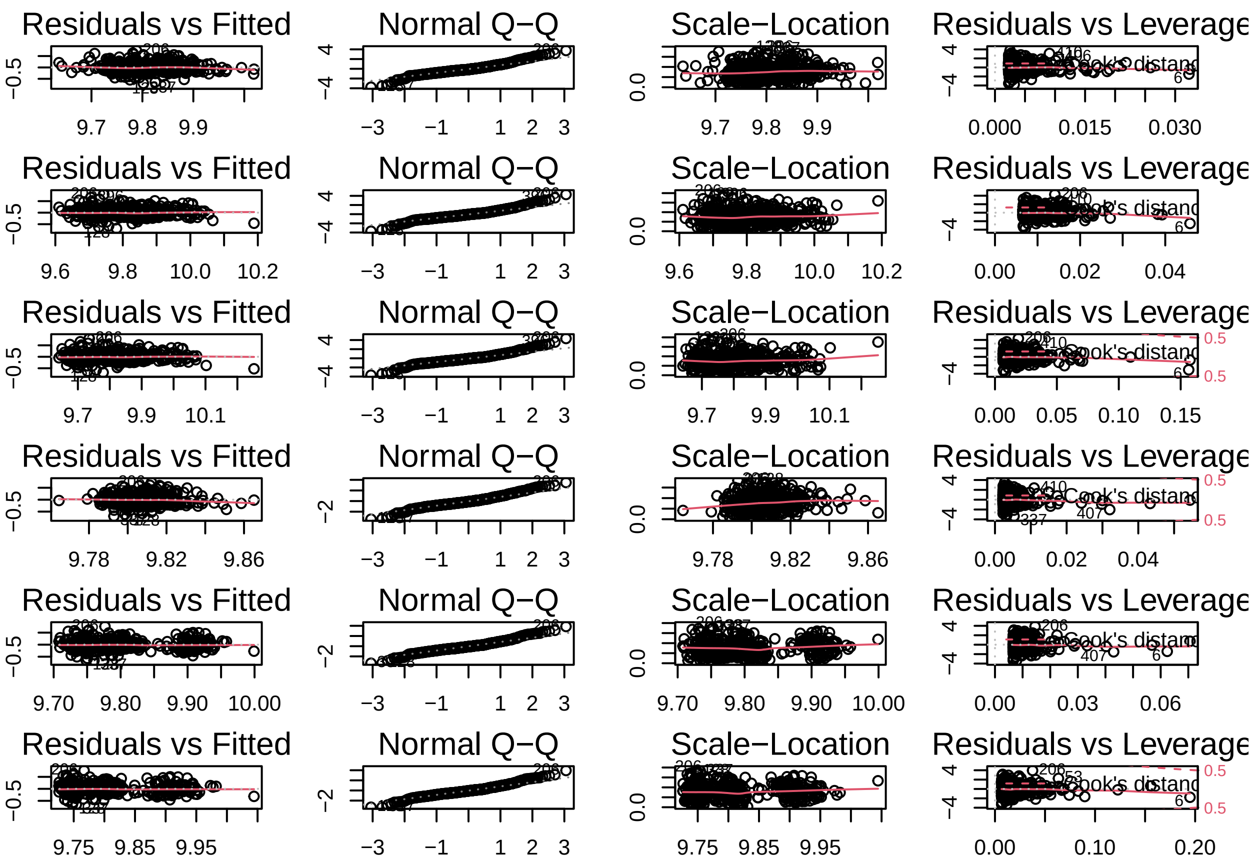
```
##          df       AIC
## model1.2  6 -227.4746
## model2.2  6 -172.1347
```

```
BIC(model1.2, model2.2)
```

```
##          df       BIC
## model1.2  6 -202.9539
## model2.2  6 -147.6140
```

Model 1.2 seems better as it has lower AIC and BIC compared to model 2.2

```
oldmar <- par()$mar
par(mfrow=c(6,4))
par(mar=c(2,2,2,2))
invisible(lapply(list(model1.1,model1.2,model1.3,model2.1,model2.2,model2.3), function(x) plot(x,cex.ma
```




```
round(coef(summary(model1.2)),2)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  9.19       0.08  115.13     0.00
## cdilogsgood$crimes           0.07       0.01    7.92     0.00
## cdilogsgood$regionNE         0.10       0.03    4.09     0.00
## cdilogsgood$regionS         -0.09       0.02   -3.68     0.00
## cdilogsgood$regionW         -0.06       0.03   -1.96     0.05
```

Across the US, for every 1% increase in crimes, we can expect a 0.07% increase in the per capita income. The intercept in our model i.e., baseline per capita income varies for different regions and is as follows: NC Region = exp(9.19) = USD 9,798.651 NE Region = exp(9.19 + 0.07) = USD 10,509.13 S Region = exp(9.19 - 0.09) = USD 8,955.293 W Region = exp(9.19 - 0.06) = USD 9,228.022 All of these region baselines are significantly different from the NC baseline. In conclusion, the relationship between per capita income and crime doesn't change with region but the effect of change in crime on per capita income varies with region.

APPENDIX 3

Since per capita income = total income / population, we're not going to consider total income and population as they would be perfectly collinear with per capita income which would result in our analysis not being able to pick up on any other variables which might be related to per capita income.

```
#Stepwise Variable Selection
cdilogsgood2 <- cdilogsgood[,c(-2,-13,-15)]
model3.1 <- lm(per.cap.income~.-region, data = cdilogsgood2)
step_result_aic <- stepAIC(model3.1,scope=list(lower = ~ 1, upper = ~ .), k=2,trace=F)
```

```r
step_result_bic <- stepAIC(model3.1,scope=list(lower = ~ 1, upper = ~ .), k=log(440),trace=F)
step_result_aic
```

```
##
## Call:
## lm(formula = per.cap.income ~ land.area + pop.18_34 + pop.65_plus +
##     doctors + pct.bach.deg + pct.below.pov + pct.unemp, data = cdilogsgood2)
##
## Coefficients:
##   (Intercept)      land.area      pop.18_34    pop.65_plus        doctors
##       9.96961       -0.03615       -0.26275        0.05126        0.06192
##  pct.bach.deg  pct.below.pov      pct.unemp
##       0.24047       -0.20534        0.07847
```

```r
step_result_bic
```

```
##
## Call:
## lm(formula = per.cap.income ~ land.area + pop.18_34 + pop.65_plus +
##     doctors + pct.bach.deg + pct.below.pov + pct.unemp, data = cdilogsgood2)
##
## Coefficients:
##   (Intercept)      land.area      pop.18_34    pop.65_plus        doctors
##       9.96961       -0.03615       -0.26275        0.05126        0.06192
##  pct.bach.deg  pct.below.pov      pct.unemp
##       0.24047       -0.20534        0.07847
```

```r
model3.2 <- lm(per.cap.income ~ land.area + pop.18_34 + pop.65_plus + doctors + pct.bach.deg + pct.belo
summary(model3.2)
```

```
##
## Call:
## lm(formula = per.cap.income ~ land.area + pop.18_34 + pop.65_plus +
##     doctors + pct.bach.deg + pct.below.pov + pct.unemp, data = cdilogsgood2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33852 -0.04799 -0.00399  0.04646  0.28265
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.969607   0.180765  55.152  < 2e-16 ***
## land.area     -0.036153   0.004890  -7.394 7.48e-13 ***
## pop.18_34     -0.262751   0.044501  -5.904 7.17e-09 ***
## pop.65_plus    0.051263   0.019327   2.652  0.00829 **
## doctors        0.061921   0.004597  13.469  < 2e-16 ***
## pct.bach.deg   0.240466   0.020847  11.535  < 2e-16 ***
## pct.below.pov -0.205343   0.010142 -20.247  < 2e-16 ***
## pct.unemp      0.078475   0.016267   4.824 1.95e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

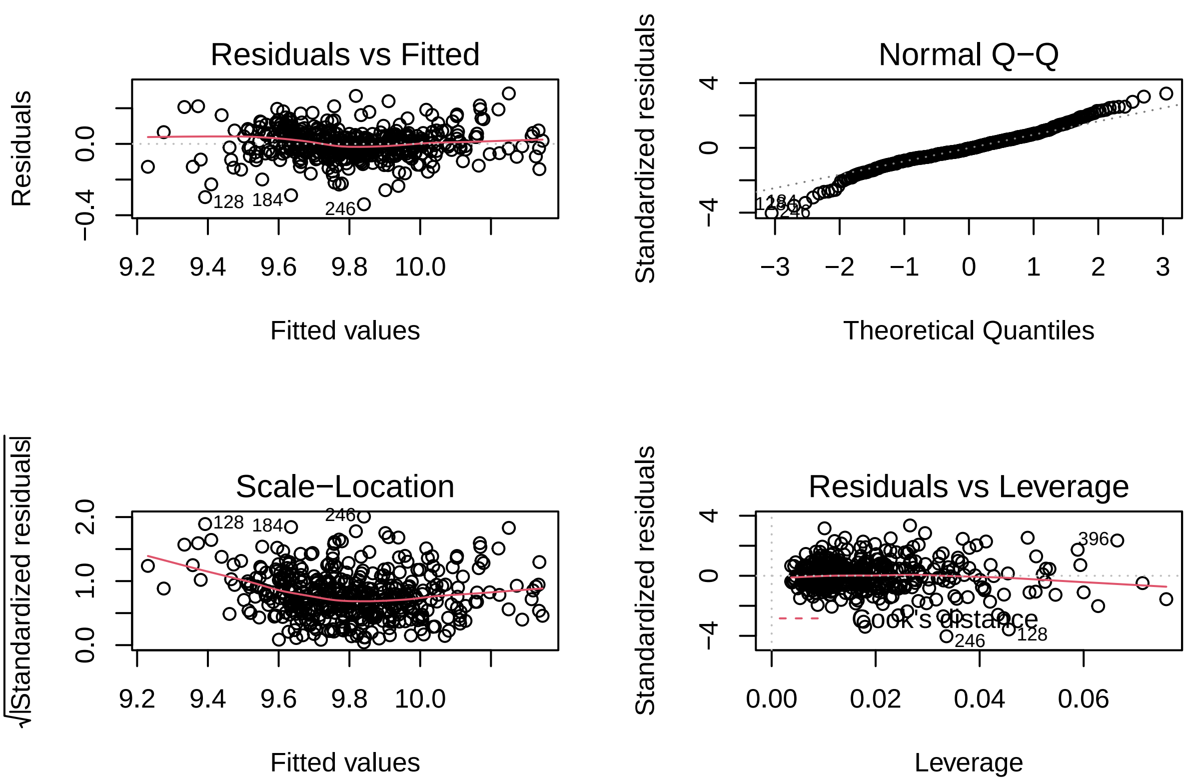
```
## Residual standard error: 0.0855 on 432 degrees of freedom
## Multiple R-squared:  0.8317, Adjusted R-squared:  0.829
## F-statistic: 304.9 on 7 and 432 DF,  p-value: < 2.2e-16
```

All the coeffecients in this model have small p-values and hence are statistically significant.

```
#Variable Inflation Factor (VIF)
vif(model3.2)
```

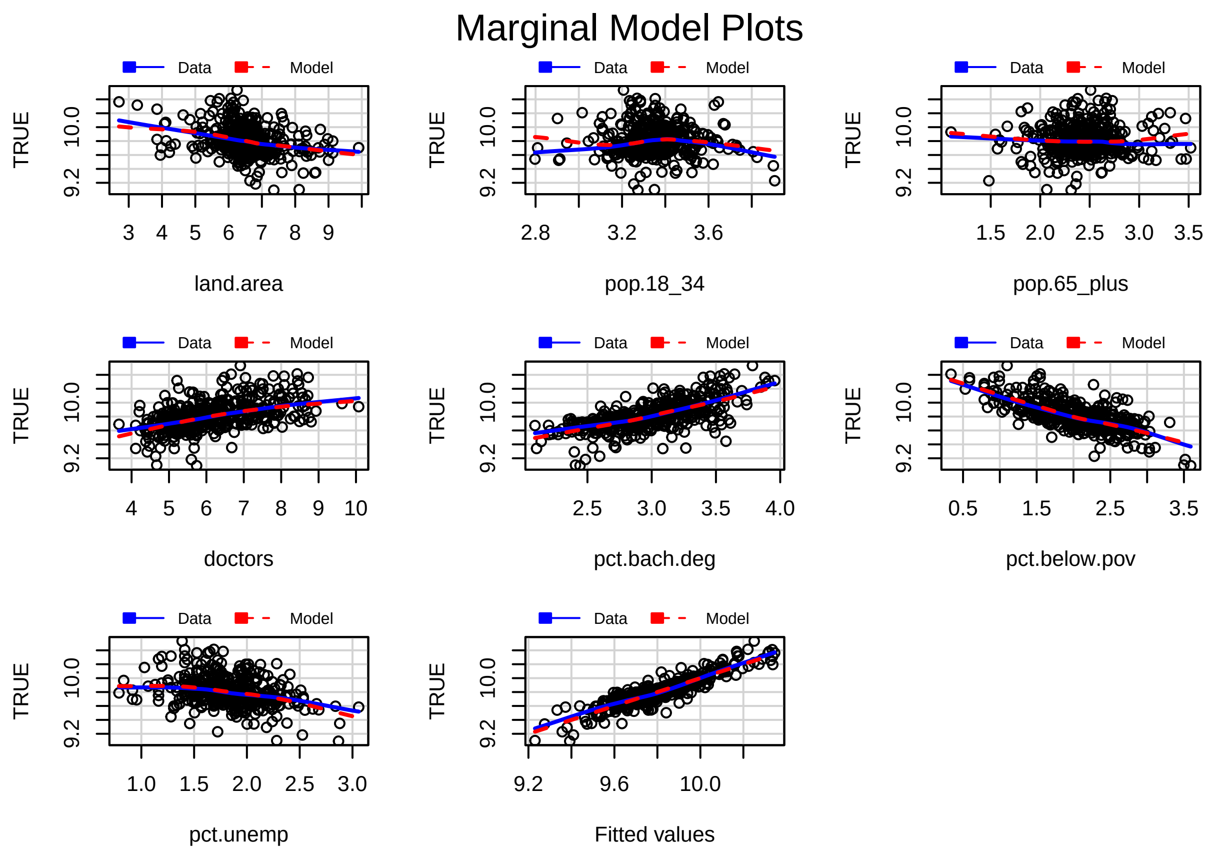
```
##      land.area      pop.18_34   pop.65_plus       doctors   pct.bach.deg
##       1.091025       2.361069      2.056001      1.661277       3.278912
## pct.below.pov     pct.unemp
##       1.738272       1.691526
```

```
#Diagnostic Plots
par(mfrow=c(2,2))
plot(model3.2)
```



None of the VIFs are large suggesting none of our explanatory variables are collinear. Diagnostic Plots suggest this model is a good fit except for the long tails in the Q-Q plot.

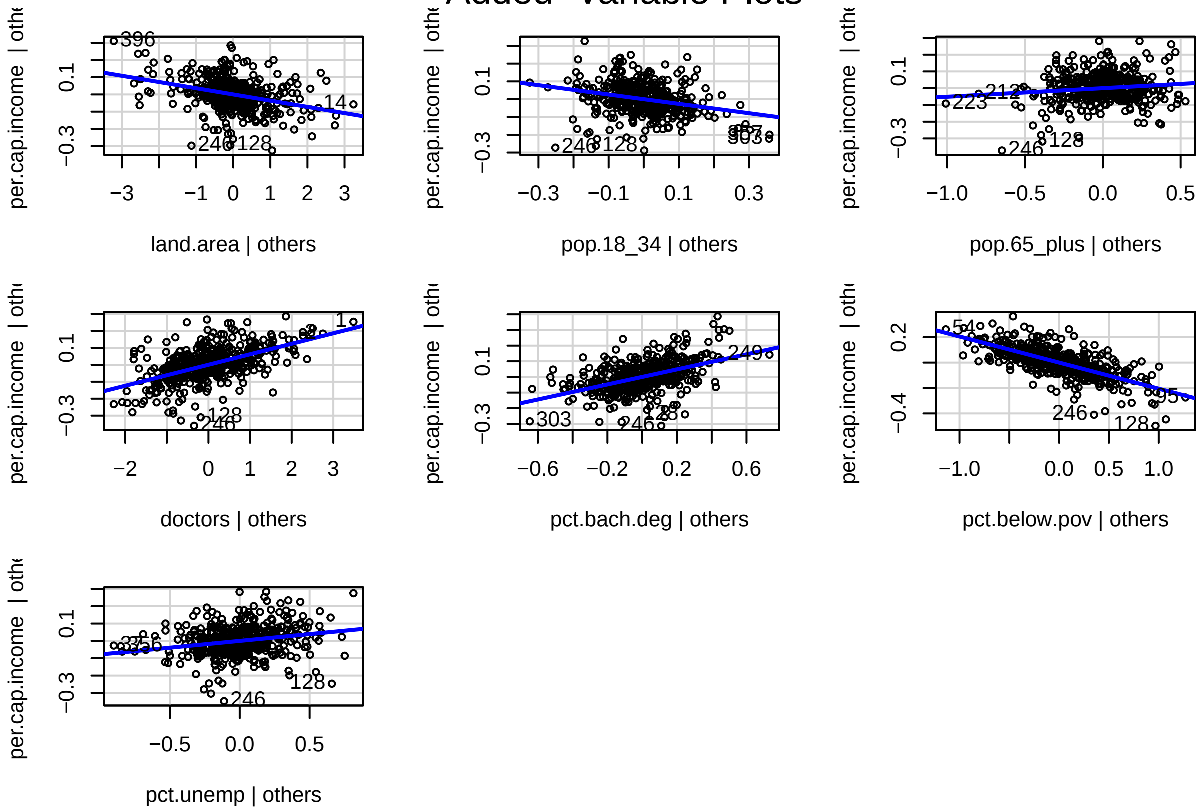
```
mmps(model3.2)
```



The predicted values of our model line up with the smooth fit function in the above marginal model plots verifying that our model is adequate.

```
avPlots(model3.2)
```

Added−Variable Plots




```r
# Interaction with region
model3.3 <- lm(per.cap.income ~ land.area*region + pop.18_34*region + pop.65_plus*region + doctors*regi
summary(model3.3)
```

```
##
## Call:
## lm(formula = per.cap.income ~ land.area * region + pop.18_34 *
##     region + pop.65_plus * region + doctors * region + pct.bach.deg *
##     region + pct.below.pov * region + pct.unemp * region, data = cdilogsgood2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.254691 -0.049853 -0.000922  0.044357  0.303808
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.283962   0.373483  27.535  < 2e-16 ***
## land.area              -0.023492   0.015512  -1.514 0.130696
## regionNE                0.748693   0.599447   1.249 0.212391
## regionS                -0.179624   0.486522  -0.369 0.712170
## regionW                -1.398741   0.570203  -2.453 0.014582 *
## pop.18_34              -0.360009   0.095952  -3.752 0.000201 ***
## pop.65_plus             0.004331   0.057370   0.075 0.939859
## doctors                 0.053251   0.009921   5.367 1.34e-07 ***
## pct.bach.deg            0.232317   0.052510   4.424 1.24e-05 ***
## pct.below.pov          -0.161190   0.027863  -5.785 1.44e-08 ***
```

```
## pct.unemp              0.103420   0.032969   3.137 0.001831 **
## land.area:regionNE     -0.020053   0.019879  -1.009 0.313694
## land.area:regionS      -0.015861   0.018085  -0.877 0.380984
## land.area:regionW      -0.010614   0.018787  -0.565 0.572417
## regionNE:pop.18_34     -0.165104   0.147606  -1.119 0.263991
## regionS:pop.18_34       0.065480   0.119085   0.550 0.582717
## regionW:pop.18_34       0.431197   0.152000   2.837 0.004784 **
## regionNE:pop.65_plus   -0.086165   0.085583  -1.007 0.314626
## regionS:pop.65_plus     0.080647   0.063586   1.268 0.205413
## regionW:pop.65_plus     0.173733   0.078432   2.215 0.027307 *
## regionNE:doctors        0.010825   0.014277   0.758 0.448760
## regionS:doctors        -0.003434   0.012762  -0.269 0.788017
## regionW:doctors         0.011984   0.013995   0.856 0.392325
## regionNE:pct.bach.deg   0.050602   0.071876   0.704 0.481819
## regionS:pct.bach.deg    0.055036   0.061820   0.890 0.373845
## regionW:pct.bach.deg   -0.076665   0.071624  -1.070 0.285080
## regionNE:pct.below.pov -0.015638   0.039120  -0.400 0.689544
## regionS:pct.below.pov  -0.013670   0.032229  -0.424 0.671686
## regionW:pct.below.pov  -0.141760   0.044487  -3.187 0.001550 **
## regionNE:pct.unemp     -0.023153   0.054711  -0.423 0.672376
## regionS:pct.unemp      -0.154959   0.046817  -3.310 0.001016 **
## regionW:pct.unemp       0.032366   0.046484   0.696 0.486649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0809 on 408 degrees of freedom
## Multiple R-squared:  0.8577, Adjusted R-squared:  0.8469
## F-statistic: 79.31 on 31 and 408 DF,  p-value: < 2.2e-16
```

We've chosen to keep the statistically significant interactions namely : region:pop.18_34, region:pop.65_plus, region:pct.below.pov , region:pct.unemp and drop the others.

```
model3.4 <- lm(per.cap.income ~ land.area + pop.18_34*region + pop.65_plus*region + doctors + pct.bach.d
summary(model3.4)
```

```
##
## Call:
## lm(formula = per.cap.income ~ land.area + pop.18_34 * region +
##     pop.65_plus * region + doctors + pct.bach.deg + pct.below.pov *
##     region + pct.unemp * region + region, data = cdilogsgood2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.269397 -0.046548 -0.003837  0.042689  0.293960
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.035e+01  3.680e-01  28.136  < 2e-16 ***
## land.area              -3.644e-02  5.567e-03  -6.547 1.73e-10 ***
## pop.18_34              -3.827e-01  8.625e-02  -4.438 1.17e-05 ***
## regionNE                5.515e-01  5.837e-01   0.945  0.34526
## regionS                -2.020e-01  4.608e-01  -0.438  0.66130
## regionW                -1.551e+00  5.370e-01  -2.889  0.00407 **
```

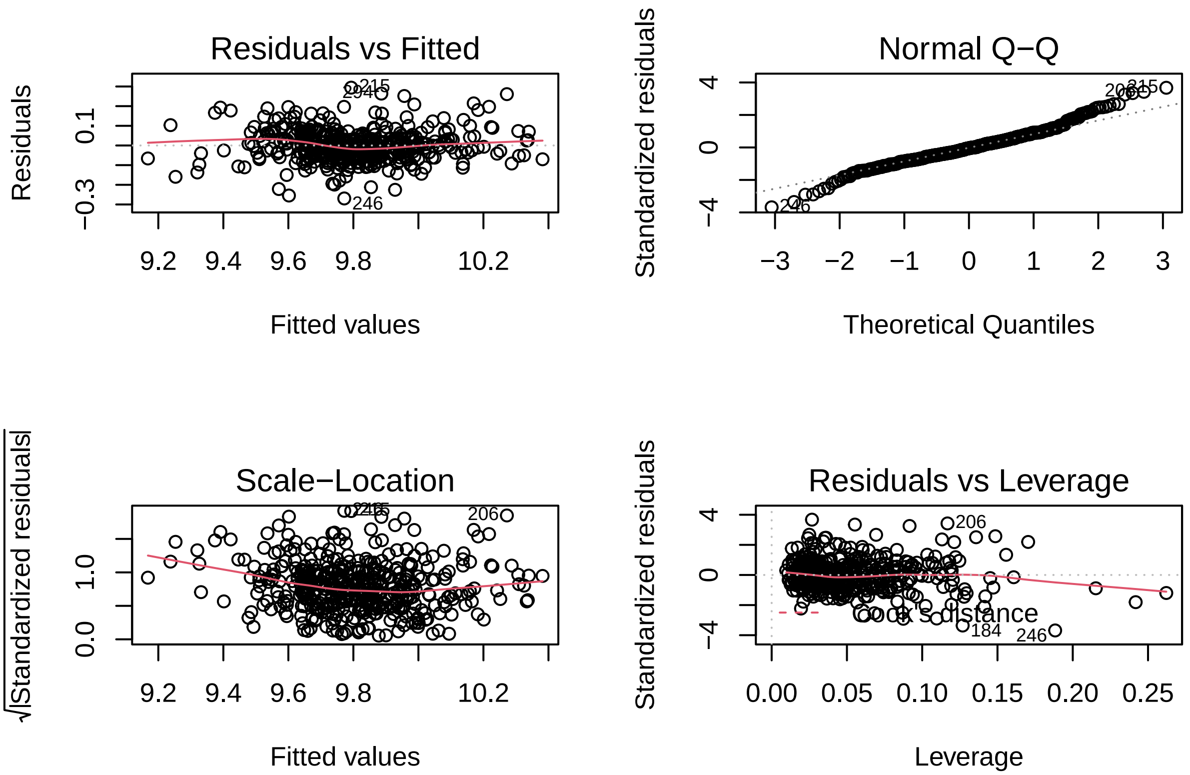
```
## pop.65_plus            1.928e-05  5.508e-02   0.000  0.99972
## doctors                5.856e-02  4.546e-03  12.881  < 2e-16 ***
## pct.bach.deg           2.485e-01  2.114e-02  11.758  < 2e-16 ***
## pct.below.pov         -1.611e-01  2.543e-02  -6.336 6.13e-10 ***
## pct.unemp              1.147e-01  2.906e-02   3.947 9.27e-05 ***
## pop.18_34:regionNE    -7.171e-02  1.308e-01  -0.548  0.58375
## pop.18_34:regionS      1.048e-01  1.070e-01   0.979  0.32809
## pop.18_34:regionW      3.886e-01  1.271e-01   3.058  0.00238 **
## regionNE:pop.65_plus  -6.190e-02  8.197e-02  -0.755  0.45060
## regionS:pop.65_plus    7.851e-02  6.049e-02   1.298  0.19506
## regionW:pop.65_plus    1.620e-01  7.422e-02   2.183  0.02960 *
## regionNE:pct.below.pov -2.797e-02  3.278e-02  -0.853  0.39397
## regionS:pct.below.pov  -2.283e-02  2.906e-02  -0.786  0.43245
## regionW:pct.below.pov  -1.101e-01  3.933e-02  -2.800  0.00535 **
## regionNE:pct.unemp     -5.637e-02  5.036e-02  -1.119  0.26359
## regionS:pct.unemp      -1.788e-01  4.169e-02  -4.288 2.24e-05 ***
## regionW:pct.unemp       4.810e-02  4.113e-02   1.169  0.24292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08113 on 417 degrees of freedom
## Multiple R-squared:  0.8537, Adjusted R-squared:  0.846
## F-statistic: 110.6 on 22 and 417 DF,  p-value: < 2.2e-16
```

```
vif(model3.4)
```

```
##                                GVIF Df GVIF^(1/(2*Df))
## land.area             1.570613e+00  1        1.253241
## pop.18_34             9.850708e+00  1        3.138584
## region                1.072087e+10  3       46.957508
## pop.65_plus           1.854436e+01  1        4.306317
## doctors               1.804582e+00  1        1.343347
## pct.bach.deg          3.743243e+00  1        1.934746
## pct.below.pov         1.213910e+01  1        3.484121
## pct.unemp             5.997470e+00  1        2.448973
## pop.18_34:region      2.331384e+09  3       36.414050
## region:pop.65_plus    1.184919e+07  3       15.098997
## region:pct.below.pov  5.395667e+04  3        6.147155
## region:pct.unemp      2.366034e+05  3        7.864483
```

Adding interactions has created collinearity but since they still have low p-values and are statistically significant we retain them.

```
par(mfrow=c(2,2))
plot(model3.4)
```



Disgnostic plots are very identical to the ones for the model without region variable.

```
#Comparing model with and without region interactions
anova(model3.2, model3.4)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ land.area + pop.18_34 + pop.65_plus + doctors +
##     pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: per.cap.income ~ land.area + pop.18_34 * region + pop.65_plus *
##     region + doctors + pct.bach.deg + pct.below.pov * region +
##     pct.unemp * region + region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    432 3.1580
## 2    417 2.7445 15    0.4135 4.1884 3.172e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(model3.2, model3.4)
```

```
##          df       AIC
## model3.2  9 -905.5407
## model3.4 24 -937.2897
```

```r
BIC(model3.2, model3.4)
```

```
##          df       BIC
## model3.2  9 -868.7597
## model3.4 24 -839.2072
```

ANOVA test and AIC are in favour of the model with region terms but BIC seems to favour the smaller model without region varible. We've chosen model 3.2 with the region terms.

```r
round(summary(model3.4)$coef,2)
```

```
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)               10.35       0.37   28.14     0.00
## land.area                 -0.04       0.01   -6.55     0.00
## pop.18_34                 -0.38       0.09   -4.44     0.00
## regionNE                   0.55       0.58    0.94     0.35
## regionS                   -0.20       0.46   -0.44     0.66
## regionW                   -1.55       0.54   -2.89     0.00
## pop.65_plus                0.00       0.06    0.00     1.00
## doctors                    0.06       0.00   12.88     0.00
## pct.bach.deg               0.25       0.02   11.76     0.00
## pct.below.pov             -0.16       0.03   -6.34     0.00
## pct.unemp                  0.11       0.03    3.95     0.00
## pop.18_34:regionNE        -0.07       0.13   -0.55     0.58
## pop.18_34:regionS          0.10       0.11    0.98     0.33
## pop.18_34:regionW          0.39       0.13    3.06     0.00
## regionNE:pop.65_plus      -0.06       0.08   -0.76     0.45
## regionS:pop.65_plus        0.08       0.06    1.30     0.20
## regionW:pop.65_plus        0.16       0.07    2.18     0.03
## regionNE:pct.below.pov    -0.03       0.03   -0.85     0.39
## regionS:pct.below.pov     -0.02       0.03   -0.79     0.43
## regionW:pct.below.pov     -0.11       0.04   -2.80     0.01
## regionNE:pct.unemp        -0.06       0.05   -1.12     0.26
## regionS:pct.unemp         -0.18       0.04   -4.29     0.00
## regionW:pct.unemp          0.05       0.04    1.17     0.24
```