Analysis on the Average Income per Person in US

Zhuoheng Han
Department of Statistics and Data Science, Carnegie Mellon University
zhouhenh@andrew.cmu.edu

Abstract

In this paper, we address that whether per-capital income is related to crimes and region and find the best model predicting per-capital income. We use the county demographic information (CDI) dataset to help us solve questions. We build regression models and compare the performance to decide the best model. Our final model predicting per.capita.income contains coefficients pop.18_34, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, region, land.area, and doctors. In order to improve the analysis, we need to research on other counties since the dataset only contains 1/9 total counties in US.

Introduction

Nowadays, social scientists are interested in determining per-capita income to evaluate the life quality of the population. In this paper, we are discussing how average income per person was related to other variables associated with the county's economic, health and social well-being from cdi.dat. We address four research questions:

- Which variables seem to be related to other variables in the dataset? Which are not? Are these relationships reasonable?
- Prove or Disprove a theory that per-capita income should be related to crime rate, and that relationship may be different in different regions of the country. Does it matter if you use number of crimes or (number of crimes)/(population) in your analysis?
- Find the best model predicting per-capita income.
- There are 51 states and around 3000 counties in US, but 48 states and 440 counties are represented in the dataset. Should we be worried about either the missing states or the missing counties? Why or why not?

Data

The cdi.dat is taken from Kutner et al. (2005). There are total 17 columns and 440 rows. It provides county demographic information (CDI) for 440 most populous counties in the United States. Each line of the dataset provides information for a single county. There are no missing values in this dataset. The definition of each variable is given below:

1. id: Identification number 1–440
2. county: County name
3. state: Two-letter state abbreviation
4. land.area: Land area (square miles)
5. pop: Estimated 1990 CDI total population
6. pop.18_34: Percent of 1990 CDI population aged 18–34

7. pop.65_plus: Percent of 1990 CDI population aged 65 or old

8. doctors: Number of professionally active non-federal doctors during 1990

9. hosp.beds: Total number of beds, cribs, and bassinets during 1990

10. 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. pct.hs.grad: Percent of adult population (persons 25 years old or older) who completed 12 or more years of school

12. pct.bach.deg: Percent of adult population (persons 25 years old or older) with bachelor's degree

13. pct.below.pov: Percent of 1990 CDI population with income below poverty level

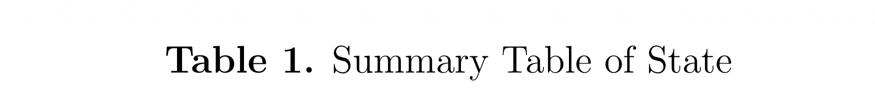
14. pct.unemp: Percent of 1990 CDI population that is unemployed

15. per.cap.income: Per-capita income (i.e. average income per person) of 1990 CDI population (in dollars)

16. tot.income: Total personal income of 1990 CDI population (in millions of dollars)

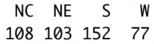
17. region: Geographic region classification used by the US Bureau of the Census, NE (northeast region of the US), NC (north-central region of the US), S (southern region of the US), and W (Western region of the US)

Variables id and county are unique for each row, which means we can ignore these two variables when doing data analysis. Below are the summary tables for two category variables state and region:



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

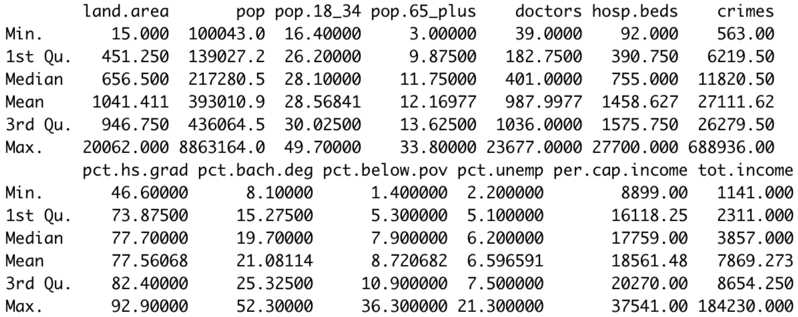
Table 1. Summary Table of State



```
 NC  NE   S   W 
108 103 152  77 
```

Table 2. Summary Table of Geographic Region

Then, the summary of numerical variables are shown below.



```
        land.area       pop pop.18_34 pop.65_plus    doctors hosp.beds    crimes
Min.       15.000  100043.0  16.40000     3.00000    39.0000    92.000    563.00
1st Qu.   451.250  139027.2  26.20000     9.87500   182.7500   390.750   6219.50
Median    656.500  217280.5  28.10000    11.75000   401.0000   755.000  11820.50
Mean     1041.411  393010.9  28.56841    12.16977   987.9977  1458.627  27111.62
3rd Qu.   946.750  436064.5  30.02500    13.62500  1036.0000  1575.750  26279.50
Max.    20062.000 8863164.0  49.70000    33.80000 23677.0000 27700.000 688936.00
        pct.hs.grad pct.bach.deg pct.below.pov pct.unemp per.cap.income tot.income
Min.       46.60000      8.10000      1.400000  2.200000        8899.00   1141.000
1st Qu.    73.87500     15.27500      5.300000  5.100000       16118.25   2311.000
Median     77.70000     19.70000      7.900000  6.200000       17759.00   3857.000
Mean       77.56068     21.08114      8.720682  6.596591       18561.48   7869.273
3rd Qu.    82.40000     25.32500     10.900000  7.500000       20270.00   8654.250
Max.       92.90000     52.30000     36.300000 21.300000       37541.00 184230.000
```

Table 3. Summary Table of Numerical Variables

Methods

In order to find the relationship between each variable, we plot the correlation matrix on the variables. We build regression models on number of crimes or (number of crimes)/(population) and use ANOVA and AIC to find the best model to inspect whether per-capita income is related to crime rate and region. We build all-subsets regression, stepwise AIC regression, and stepwise BIC regression to find the best model to find the best model predicting per-capita income. Finally, we apply EDA method on the region to compare whether it follows our understanding.

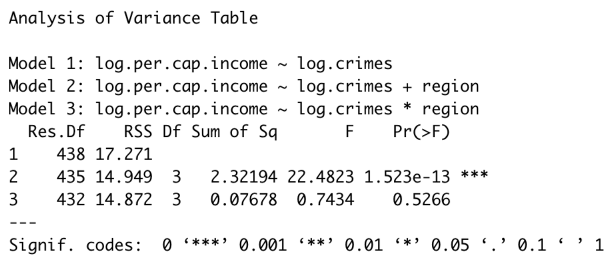
Results

Relationship Between Variables

From the correlation matrix plot (Appendix **Figure 1.**), we can find that pop is highly correlated with tot.income, doctors, hosp.beds, and crimes. That is no surprise since more population result in more total incomes; more population result in more people choosing to be doctors; more hospital beds are needed for more population; and more crimes might occur due to the more population. Also, three variables doctors, hosp.beds, and crimes are strongly correlated with one another, which is reasonable because more hospitals beds are needed if there exist more crimes and result in more doctors to take care. per.cap.income is kind of highly correlated with pct.hs.grad, pct.bach.deg (postively correlated) and pct.below.pov, pct.unemp (negatively correlated); all four of these variables are moderately highly correlated with one another. This is also reasonable since people with higher degree have more chance to be employed and always earn more.

Analysis on Income and Crime in Different Region

Before building models, we need to transform the skewed data first. Since logarithms clean up a lot of the skewing in the data, we use log-transform on land.area, pop, doctors, hosp.beds, crimes, per.cap.income, and tot.income variables. Then there are three models to think about.
lm(log.per.cap.income $\sim$ log.crimes),
lm(log.per.cap.income $\sim$ log.crimes + region), and
lm(log.per.cap.income $\sim$ log.crimes * region)

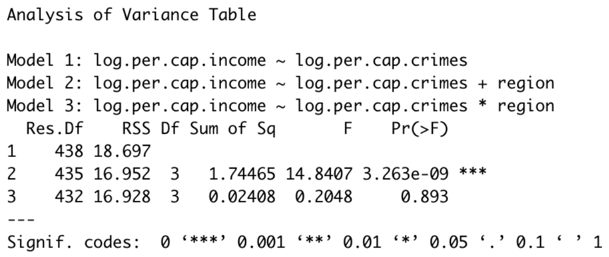


```
Analysis of Variance Table

Model 1: log.per.cap.income ~ log.crimes
Model 2: log.per.cap.income ~ log.crimes + region
Model 3: log.per.cap.income ~ log.crimes * 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
```

Table 4. ANOVA on three models

From the ANOVA result, we can find that model lm(log.per.cap.income $\sim$ log.crimes + region) is the best among those three models since its p-value $= 1.523e - 13 < 0.05$.
In order to compare this with a model involving per-capita crime, we construct a new variable log.per.cap.crimes, which is equal to log.crimes - log.pop. Once again, there are three models to think about.
lm(log.per.cap.income $\sim$ log.per.cap.crimes),
lm(log.per.cap.income $\sim$ log.per.cap.crimes + region), and
lm(log.per.cap.income $\sim$ log.per.cap.crimes * region).



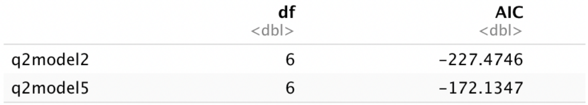
```
Analysis of Variance Table

Model 1: log.per.cap.income ~ log.per.cap.crimes
Model 2: log.per.cap.income ~ log.per.cap.crimes + region
Model 3: log.per.cap.income ~ log.per.cap.crimes * 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
```

Table 5. ANOVA on three models

From the ANOVA result, we can find that model lm(log.per.cap.income ∼ log.per.cap.crimes + region) is the best among those three models since its p-value $= 3.263e - 09 < 0.05$.

To compare two winners, we use AIC because the two winners are not nested models.



	df <dbl>	AIC <dbl>
q2model2	6	-227.4746
q2model5	6	-172.1347

Table 6. AIC between winner models

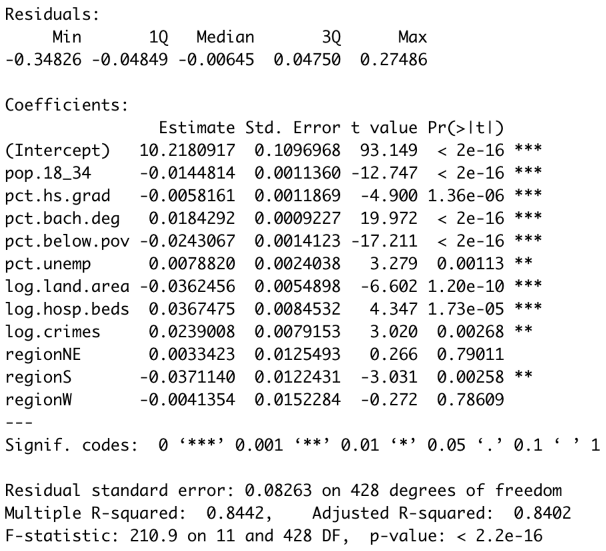
From the AIC result, it shows that lm(log.per.cap.income ∼ log.crimes + region) is the best model since AIC value of this model is smaller. The level of income varies with region in the US, but is not related to crime.

Best Model Predicting Income per Person

From the Data section, id and county variables are not useful so we decide to drop these two variables. Also, we take log.pop and log.tot.income out of consideration, since log.per.cap.income = log.tot.income - log.pop, which is a deterministic function of those two variables. Lastly, state and region are two category variables for the location, and region contains states geographically so we decide to drop off state to avoid duplicate information.

First, we start with all-subsets regression. Based on the all-subsets plot (Appendix **Figure 2.**), we can find that the best model is with coefficients pop.18_34, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, regionS, log.land.area, and log.doctors.

Based on the rule of thumb: if any indicator for a categorical variable seems important (e.g. a statistically significant coefficient), then keep the whole categorical variable, so we will keep region in the final model. Below are the summary statistics of the final model using all.subsets.



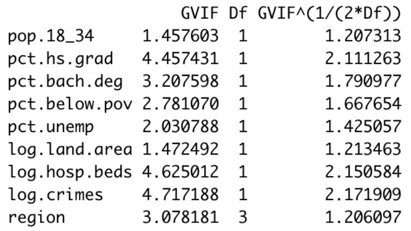
```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.34826 -0.04849 -0.00645  0.04750  0.27486 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   10.2180917  0.1096968  93.149  < 2e-16 ***
pop.18_34     -0.0144814  0.0011360 -12.747  < 2e-16 ***
pct.hs.grad   -0.0058161  0.0011869  -4.900 1.36e-06 ***
pct.bach.deg   0.0184292  0.0009227  19.972  < 2e-16 ***
pct.below.pov -0.0243067  0.0014123 -17.211  < 2e-16 ***
pct.unemp      0.0078820  0.0024038   3.279  0.00113 ** 
log.land.area -0.0362456  0.0054898  -6.602 1.20e-10 ***
log.hosp.beds  0.0367475  0.0084532   4.347 1.73e-05 ***
log.crimes     0.0239008  0.0079153   3.020  0.00268 ** 
regionNE       0.0033423  0.0125493   0.266  0.79011    
regionS       -0.0371140  0.0122431  -3.031  0.00258 ** 
regionW       -0.0041354  0.0152284  -0.272  0.78609    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08263 on 428 degrees of freedom
Multiple R-squared:  0.8442,	Adjusted R-squared:  0.8402 
F-statistic: 210.9 on 11 and 428 DF,  p-value: < 2.2e-16
```

Table 7. Summary statistics for all-subsets regression model

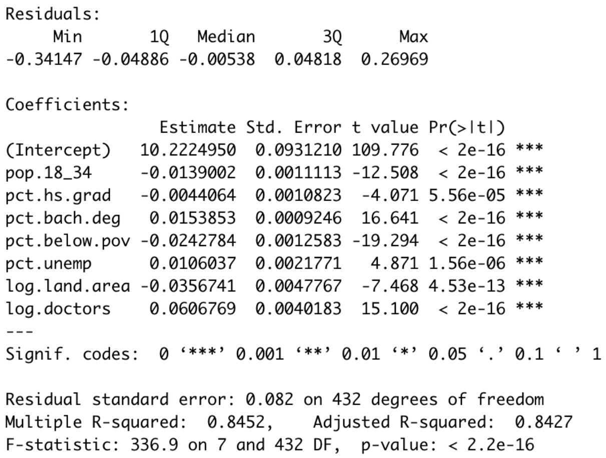
We can find variables are statistically significant. In order to further decide whether it is a good model, we apply VIF method to check the multicollinearity and plot the diagnostic plots (Appendix **Figure 3.**)



```
                  GVIF Df GVIF^(1/(2*Df))
pop.18_34     1.457603  1        1.207313
pct.hs.grad   4.457431  1        2.111263
pct.bach.deg  3.207598  1        1.790977
pct.below.pov 2.781070  1        1.667654
pct.unemp     2.030788  1        1.425057
log.land.area 1.472492  1        1.213463
log.hosp.beds 4.625012  1        2.150584
log.crimes    4.717188  1        2.171909
region        3.078181  3        1.206097
```

Table 8. VIF for the all-subsets regression model

None of the VIF values seem excessively large, i.e., there is no multicollinearity issue that need to be addressed.
From Residuals vs Fitted plot, residuals are randomly distributed around 0. From Q-Q plot, it suggests both the left and the right tails are a bit longer than expected for the normal distribution. From the Scale-Location plot, it has constant variance. From Residuals vs Leverage plot, there is no high influential points that might influence the model performance.
Then, we consider stepwise regression using AIC and BIC. We can find that the best model using BIC stepwise is with coefficients pop.18_34, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, log.land.area, and log.doctors. Below are the summary statistics of the final model using BIC stepwise.



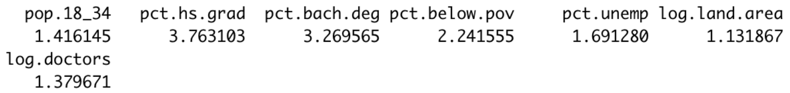
```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.34147 -0.04886 -0.00538  0.04818  0.26969 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)   10.2224950  0.0931210 109.776  < 2e-16 ***
pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
pct.hs.grad   -0.0044064  0.0010823  -4.071 5.56e-05 ***
pct.bach.deg   0.0153853  0.0009246  16.641  < 2e-16 ***
pct.below.pov -0.0242784  0.0012583 -19.294  < 2e-16 ***
pct.unemp      0.0106037  0.0021771   4.871 1.56e-06 ***
log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
log.doctors    0.0606769  0.0040183  15.100  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.082 on 432 degrees of freedom
Multiple R-squared:  0.8452,	Adjusted R-squared:  0.8427 
F-statistic: 336.9 on 7 and 432 DF,  p-value: < 2.2e-16
```

Table 9. Summary statistics for stepwise BIC model

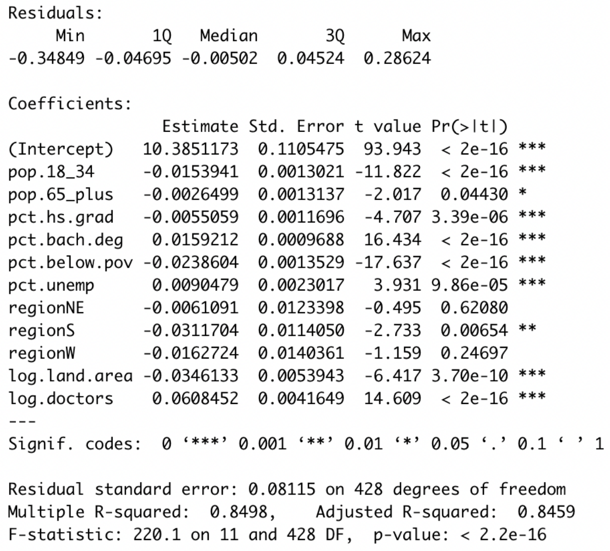
We can find variables are statistically significant. In order to further decide whether it is a good model, we apply VIF method to check the multicollinearity and plot the diagnostic plots (Appendix **Figure 4.**)



```
    pop.18_34   pct.hs.grad  pct.bach.deg pct.below.pov     pct.unemp log.land.area 
     1.416145      3.763103      3.269565      2.241555      1.691280      1.131867 
  log.doctors 
     1.379671 
```

Table 10. VIF for stepwise BIC regression model

None of the VIF values seem excessively large, i.e., there is no multicollinearity issue that need to be addressed.
From Residuals vs Fitted plot, residuals are randomly distributed around 0. From Q-Q plot, it suggests the left tails are a bit longer than expected for the normal distribution. From the Scale-Location plot, it has constant variance. From Residuals vs Leverage plot, there is no high influential points that might influence the model performance.
We can find that the best model using AIC stepwise is with coefficients pop.18_34, pop.65_plus, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, region, log.land.area, and log.doctors. Below are the summary statistics of the final model using AIC stepwise.



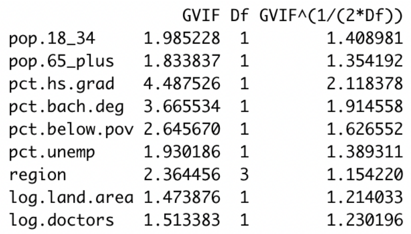
```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.34849 -0.04695 -0.00502  0.04524  0.28624 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   10.3851173  0.1105475  93.943  < 2e-16 ***
pop.18_34     -0.0153941  0.0013021 -11.822  < 2e-16 ***
pop.65_plus   -0.0026499  0.0013137  -2.017  0.04430 *  
pct.hs.grad   -0.0055059  0.0011696  -4.707 3.39e-06 ***
pct.bach.deg   0.0159212  0.0009688  16.434  < 2e-16 ***
pct.below.pov -0.0238604  0.0013529 -17.637  < 2e-16 ***
pct.unemp      0.0090479  0.0023017   3.931 9.86e-05 ***
regionNE      -0.0061091  0.0123398  -0.495  0.62080    
regionS       -0.0311704  0.0114050  -2.733  0.00654 ** 
regionW       -0.0162724  0.0140361  -1.159  0.24697    
log.land.area -0.0346133  0.0053943  -6.417 3.70e-10 ***
log.doctors    0.0608452  0.0041649  14.609  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08115 on 428 degrees of freedom
Multiple R-squared:  0.8498,	Adjusted R-squared:  0.8459 
F-statistic: 220.1 on 11 and 428 DF,  p-value: < 2.2e-16
```

Table 11. Summary statistics for stepwise AIC model

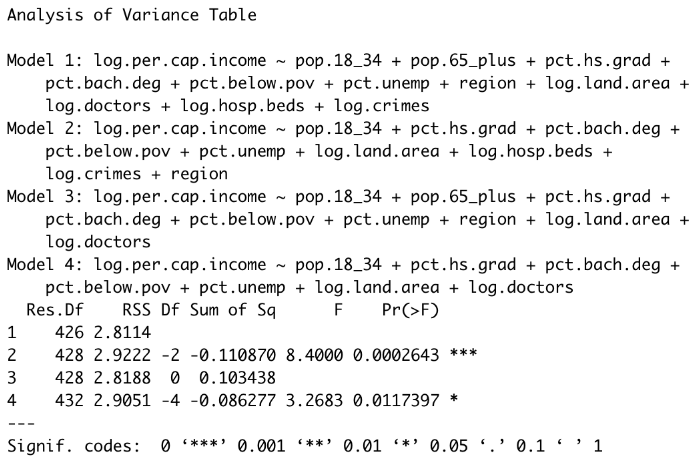
We can find variables are statistically significant. In order to further decide whether it is a good model, we apply VIF method to check the multicollinearity and plot the diagnostic plots (Appendix **Figure 5.**)



```
                  GVIF Df GVIF^(1/(2*Df))
pop.18_34     1.985228  1        1.408981
pop.65_plus   1.833837  1        1.354192
pct.hs.grad   4.487526  1        2.118378
pct.bach.deg  3.665534  1        1.914558
pct.below.pov 2.645670  1        1.626552
pct.unemp     1.930186  1        1.389311
region        2.364456  3        1.154220
log.land.area 1.473876  1        1.214033
log.doctors   1.513383  1        1.230196
```

Table 12. VIF for stepwise AIC regression model

None of the VIF values seem excessively large, i.e., there is no multicollinearity issue that need to be addressed.
From Residuals vs Fitted plot, residuals are randomly distributed around 0. From Q-Q plot, it suggests both the left tails and right tails are a bit longer than expected for the normal distribution. From the Scale-Location plot, it has constant variance. From Residuals vs Leverage plot, there is no high influential points that might influence the model performance.
In order to find the best model using these three methods, we use ANOVA with full model. We find that the best model is all-subsets regression model lm(log.per.cap.income ∼ pop.18_34 + pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp + region + log.land.area + log.doctors).



```
Analysis of Variance Table

Model 1: log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
    pct.bach.deg + pct.below.pov + pct.unemp + region + log.land.area +
    log.doctors + log.hosp.beds + log.crimes
Model 2: log.per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
    pct.below.pov + pct.unemp + log.land.area + log.hosp.beds +
    log.crimes + region
Model 3: log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
    pct.bach.deg + pct.below.pov + pct.unemp + region + log.land.area +
    log.doctors
Model 4: log.per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
    pct.below.pov + pct.unemp + log.land.area + log.doctors
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    426 2.8114
2    428 2.9222 -2 -0.110870 8.4000 0.0002643 ***
3    428 2.8188  0  0.103438
4    432 2.9051 -4 -0.086277 3.2683 0.0117397 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 13. ANOVA table

Analysis on Missing States and Counties

Based on these 440 counties, we would say if these counties can represent the around 3000 counties in United States, then we do not need to worry about the missing counties. However, since the counties are 440 of the most populous counties in the United States, it is not randomly sampled, which might cause the bias. In order to check the feasibility, we plot two boxplots (Appendix **Figure 6.**&**Figure 7.**) which are per capital income in different region and population in different regions. We can find that the median of per capital income in NE region is the highest, which is reasonable since northeastern US are economically developed area. the sum of population in NE region is higher than that in NC region, which conflicts the fact that NE region has the lowest region. Thus, we might worry about the missing counties.

Discussion

According to the results, per.capita.income is not highly correlated with other variables except income and pop. If we only keep per.capita.income, crimes, and region, we can find that there is no strong relationship between per.capita.income and crimes, but per.capita.income varies in different regions. Our final model predicting per.capita.income contains coefficients pop.18_34, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, region, log.land.area, and log.doctors. Even though the performance of model is well, we still need to worry about the missing counties.

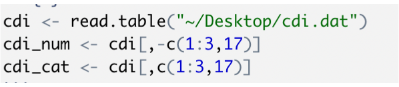
If we have no time limitation, we are going to use LASSO and Ridge regression with cross-validation when analyzing the dataset. In this way, we might be able to distinguish which model is the best model better, at least in terms of prediction error. Also, one more weakness is that there are only 440 counties in our dataset, which are 1/9 of all counties in US. It might be biased since we use this dataset to build a model predicting the per capital income in US. It would be an improvement if we spend more time researching on other counties. Taking other counties into consideration will help us address the issues much better, especially the last research question.

Reference

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

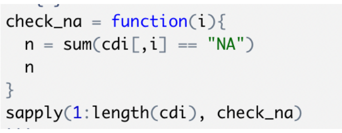
Appendix

Import the data and divide the dataset into category variables and numeric variables.



```
cdi <- read.table("~/Desktop/cdi.dat")
cdi_num <- cdi[,-c(1:3,17)]
cdi_cat <- cdi[,c(1:3,17)]
```

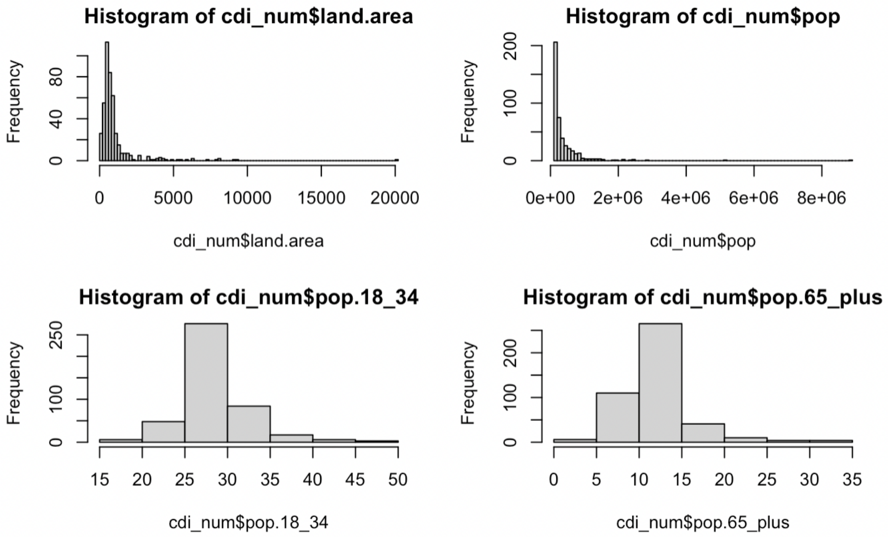
Check NA.



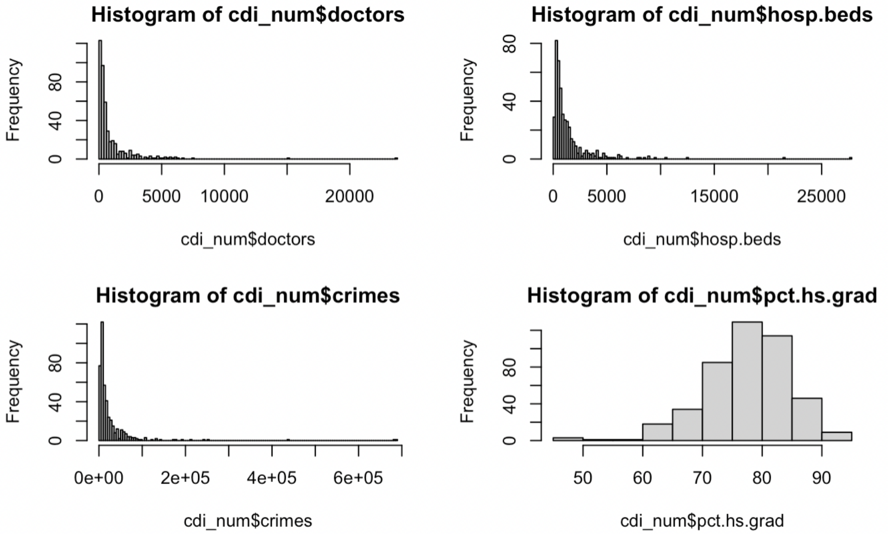
```
check_na = function(i){
  n = sum(cdi[,i] == "NA")
  n
}
sapply(1:length(cdi), check_na)
```

Plot histograms of each numeric variables.

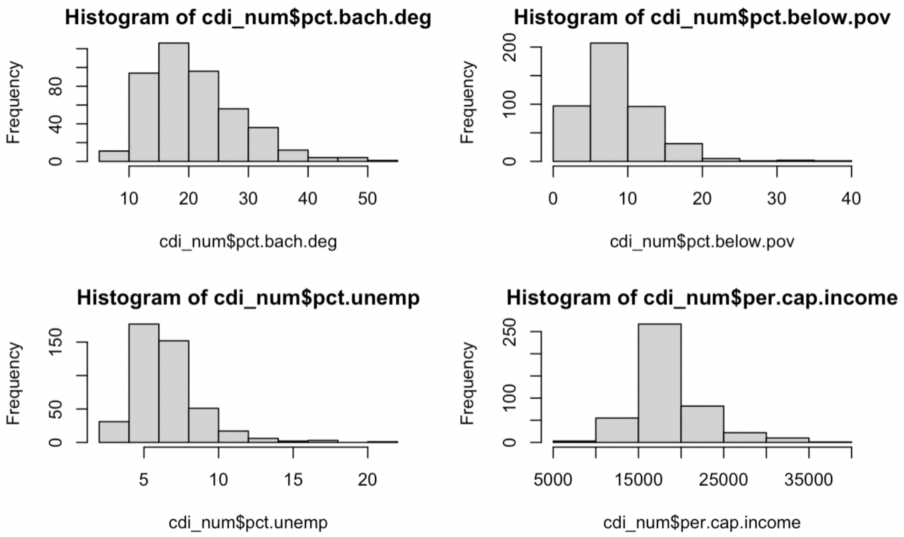
```r
par(mfrow = c(2,2))
hist(cdi_num$land.area, breaks = 100)
hist(cdi_num$pop, breaks = 100)
hist(cdi_num$pop.18_34)
hist(cdi_num$pop.65_plus)
hist(cdi_num$doctors, breaks = 100)
hist(cdi_num$hosp.beds, breaks = 100)
hist(cdi_num$crimes, breaks = 100)
hist(cdi_num$pct.hs.grad)
hist(cdi_num$pct.bach.deg)
hist(cdi_num$pct.below.pov)
hist(cdi_num$pct.unemp)
hist(cdi_num$per.cap.income)
hist(cdi_num$tot.income, breaks = 100)
```



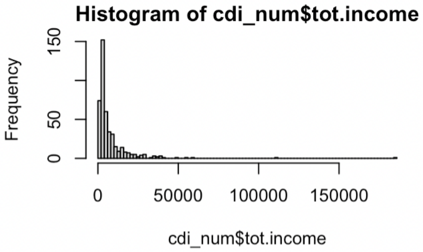
Hist 1.



Hist 2.

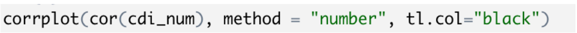


Hist 3.



Hist 4.

In order to find the relationships between each variable, we plot the correlation matrix.



```
corrplot(cor(cdi_num), method = "number", tl.col="black")
```

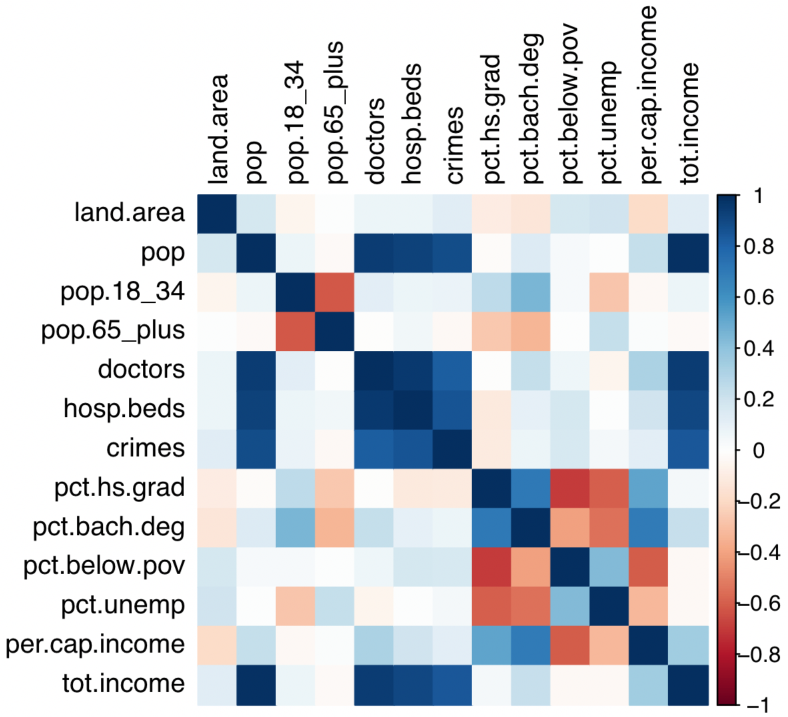
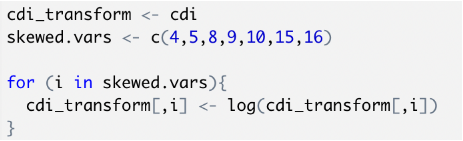


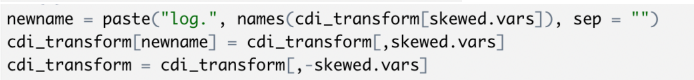
Figure 1. Correlation Matrix

Log transform on land.area, pop, doctors, hosp.beds, crimes, per.cap.income, and tot.income variables and rename the column name.



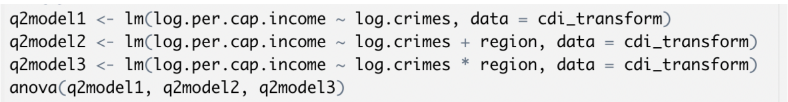
```
cdi_transform <- cdi
skewed.vars <- c(4,5,8,9,10,15,16)

for (i in skewed.vars){
  cdi_transform[,i] <- log(cdi_transform[,i])
}
```

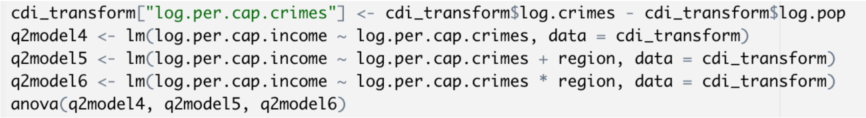


```
newname = paste("log.", names(cdi_transform[skewed.vars]), sep = "")
cdi_transform[newname] = cdi_transform[,skewed.vars]
cdi_transform = cdi_transform[,-skewed.vars]
```

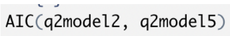
Then we build three models on crime number and region and three models on crime number/population and region.



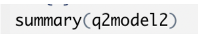
```
q2model1 <- lm(log.per.cap.income ~ log.crimes, data = cdi_transform)
q2model2 <- lm(log.per.cap.income ~ log.crimes + region, data = cdi_transform)
q2model3 <- lm(log.per.cap.income ~ log.crimes * region, data = cdi_transform)
anova(q2model1, q2model2, q2model3)
```



```
cdi_transform["log.per.cap.crimes"] <- cdi_transform$log.crimes - cdi_transform$log.pop
q2model4 <- lm(log.per.cap.income ~ log.per.cap.crimes, data = cdi_transform)
q2model5 <- lm(log.per.cap.income ~ log.per.cap.crimes + region, data = cdi_transform)
q2model6 <- lm(log.per.cap.income ~ log.per.cap.crimes * region, data = cdi_transform)
anova(q2model4, q2model5, q2model6)
```



```
AIC(q2model2, q2model5)
```



```
summary(q2model2)
```

In order to find the best model, we first drop useless variables id, county, log.tot.income, log.pop, and state.

```
cdi_final <- cdi_transform[, -which(names(cdi_transform) %in% c("id", "county", "state",
"log.pop", "log.tot.income", "log.per.cap.crimes"))]
```

Apply all-subsets regression

```
all.subsets <- regsubsets(log.per.cap.income ~ ., cdi_final, nvmax = 10)
plot(all.subsets)
```

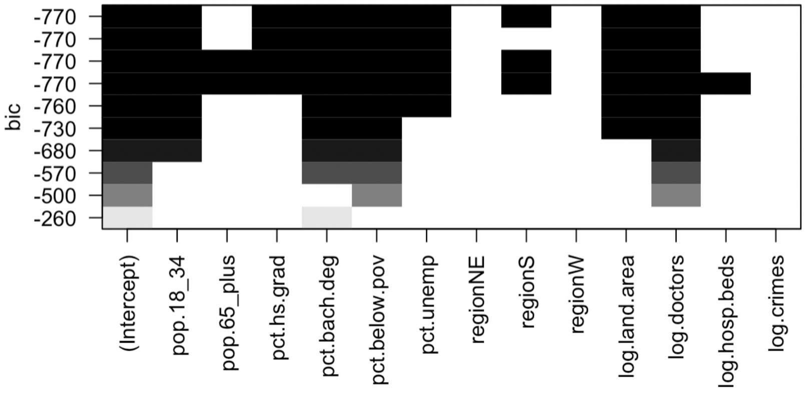


Figure 2. All-subsets Plot

```
tmp <- cdi_final[,all.subsets.summary$which[best.model,][-1]]
tmp["log.per.cap.income"] <- cdi_final["log.per.cap.income"]
tmp["region"] <- cdi_final["region"]
all.subsets.model <- lm(log.per.cap.income ~ .,data = tmp)
summary(all.subsets.model)
```

```
vif(all.subsets.model)
```

```
par(mfrow=c(2,2))
plot(all.subsets.model)
```

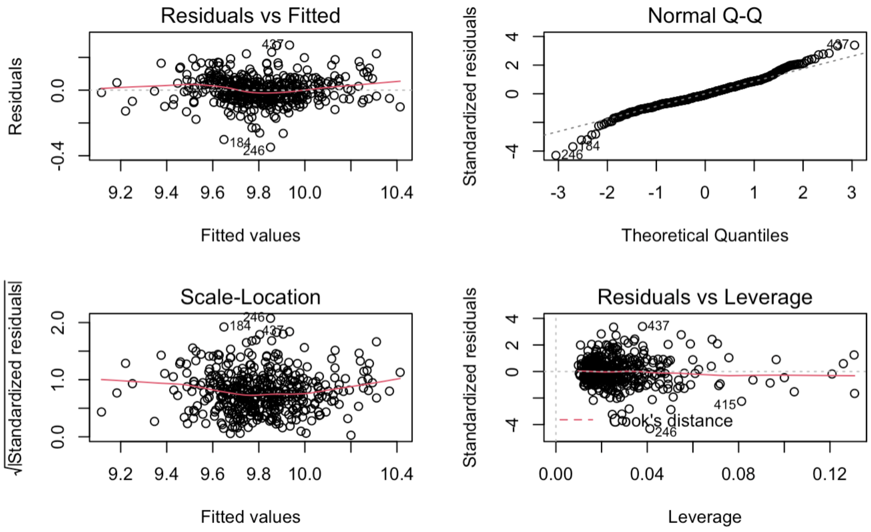


Figure 3. Diagnostic plots for All-subsets Regression Model

Apply stepwise BIC

```
model.BIC <- stepAIC(lm(log.per.cap.income ~ .,data = cdi_final), direction = "both", k =
log(dim(cdi_final)[1]))
summary(model.BIC)
```

```
vif(model.BIC)
```

```
par(mfrow=c(2,2))
plot(model.BIC)
```

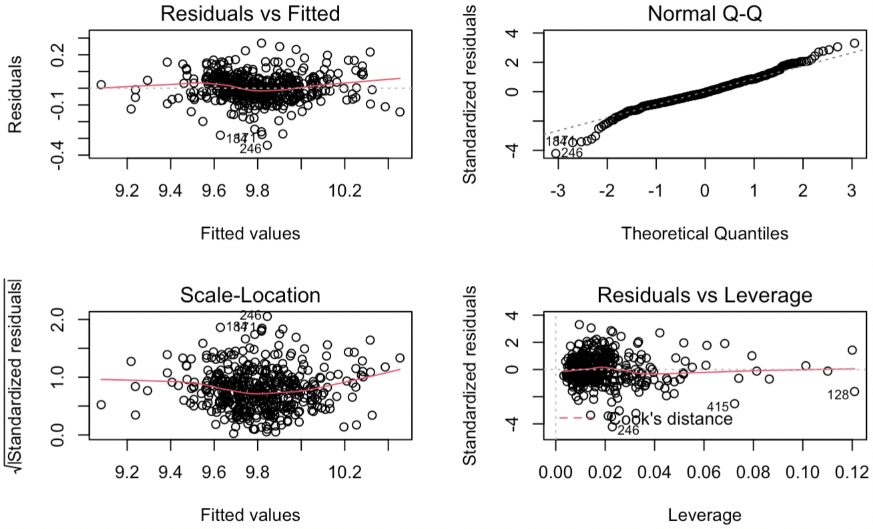


Figure 4. Diagnostic plots for Stepwise BIC Regression Model

Apply stepwise AIC

```
model.AIC <- stepAIC(lm(log.per.cap.income ~ .,data = cdi_final), direction = "both", k = 2)
summary(model.AIC)
```

```
vif(model.AIC)
```

```
par(mfrow=c(2,2))
plot(model.AIC)
```

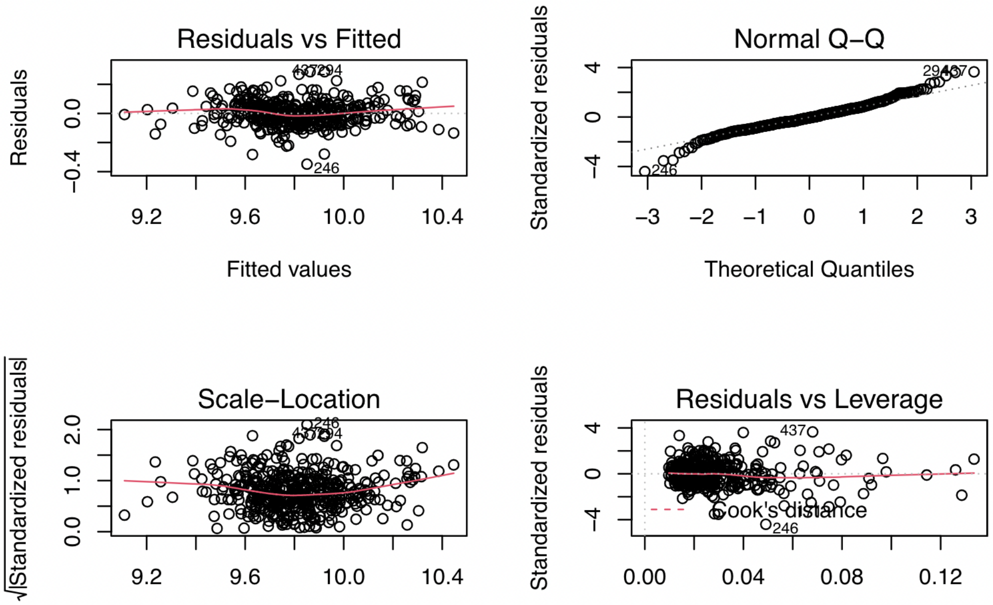
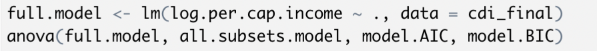
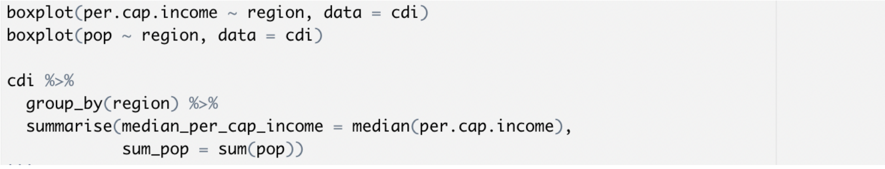


Figure 5. Diagnostic plots for Stepwise AIC Regression Model



```
full.model <- lm(log.per.cap.income ~ ., data = cdi_final)
anova(full.model, all.subsets.model, model.AIC, model.BIC)
```



```
boxplot(per.cap.income ~ region, data = cdi)
boxplot(pop ~ region, data = cdi)

cdi %>%
  group_by(region) %>%
  summarise(median_per_cap_income = median(per.cap.income),
            sum_pop = sum(pop))
```

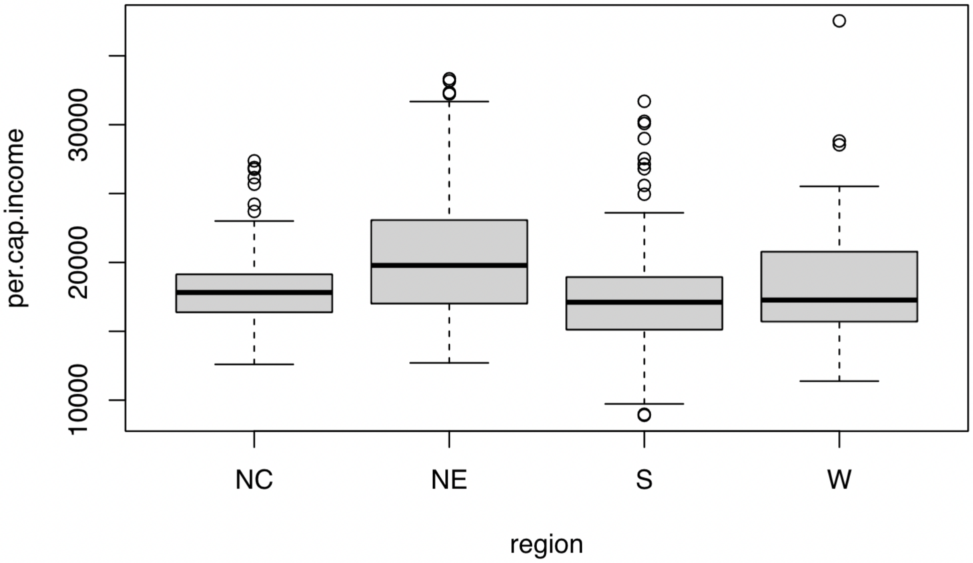


Figure 6. Boxplot of per Capital Income in Different region

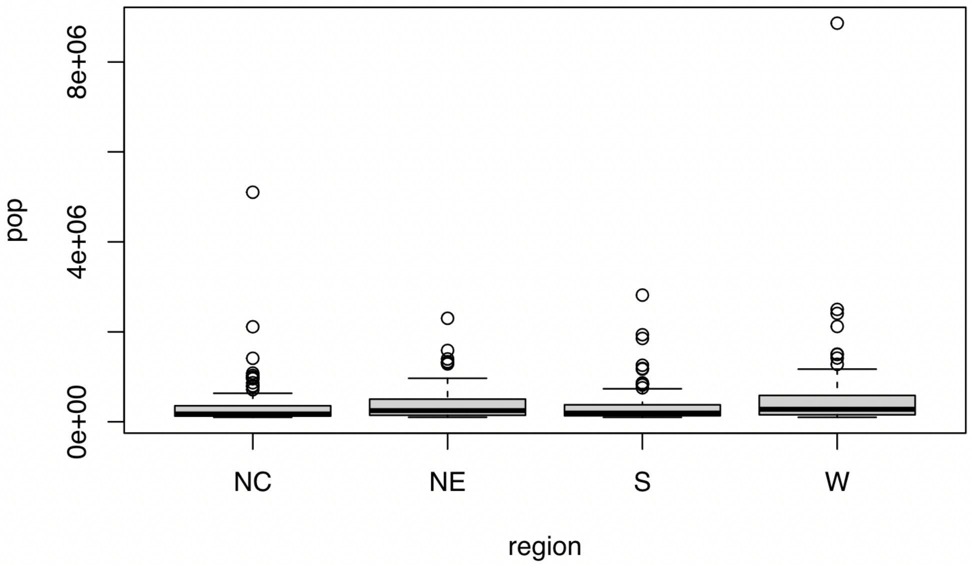


Figure 7. Boxplot of population in Different region

```
## # A tibble: 4 x 3
##   region median_per_cap_income  sum_pop
##   <chr>                  <dbl>    <int>
## 1 NC                     17817 37386529
## 2 NE                     19785 40770956
## 3 S                      17110 50008592
## 4 W                      17268 44758728
```

Table 1.