How demographic factors affect per capita income in the most populous counties in the United States from 1990 and 1992

***Author**: Hongsheng Xie*
Department of Statistics and Data Science, Carnegie Mellon University
hongshex@andrew.cmu.edu

Abstract

In this paper, we will use a dataset containing demographic factors to explore the correlation between each pair of variables, factors that affect income per person significantly, and the model that can predict per capita income appropriately. County Demographic Information (CDI) data collected by Kutner et al. (2005) is used and it contains demographic features across the most populous counties in United States. By doing exploratory data analysis, multiple linear regression, diagnosis visualization, and stepwise variable selection, several factors that affect income significantly are summed up. We find that crime is one surprising factor which does not have a direct correlation with income in the correlation graph, and we find the set of predictor variables which performs the best with variable selection methods. is studied independently, and the reason why crime works is explained. The results for all four questions are stated in the conclusion part, and the limitations of the models and data, for example no data for less populated counties, are discussed as well.

Introduction

For most people around the world, income and salary is one of the most important issue because money can improve the quality of people's life. Some social scientists are interested in looking at the historical data, to learn how average income per person was related to other variables associated with the county's economic, health and social well-being. With the dataset containing variables including distribution of age, education background, crime situation, etc., a statistical analysis can be developed to generate a professional conclusion that what factors play a great role for per capital income. Using data containing data of age, education background, crime situation, etc. from Kutner et al. (2005), this paper attempts to provide a multiple linear regression analysis to study the relationship between per capita income and other variables and build a variable set which could develop the most appropriate prediction model.

We will answer these following questions:

1. Relationships between the variables: Which variables seem to be related to which other variables in the data and which are not? Are there any surprising correlations?
2. Crime and crime rate: What is the relationship between crime and per capita income? Does this relationship depend on the region of the country (Northeast, Northcentral, South, and West)? Which form of crime data perform better? number of crimes or (number of crimes)/(population)?
3. What is the best model predicting per-capita income from the other variables?
    - Best reflects the social science and the meaning of the variables
    - Best satisfies modeling assumptions
    - Is most clearly indicated by the data
    - Can be explained to someone who is more interested in social, economic and health factors than in mathematics and statistics.

4. The data set used for this paper only contains data for the most popular counties. Should we be worried about either the missing states or the missing counties?

Data

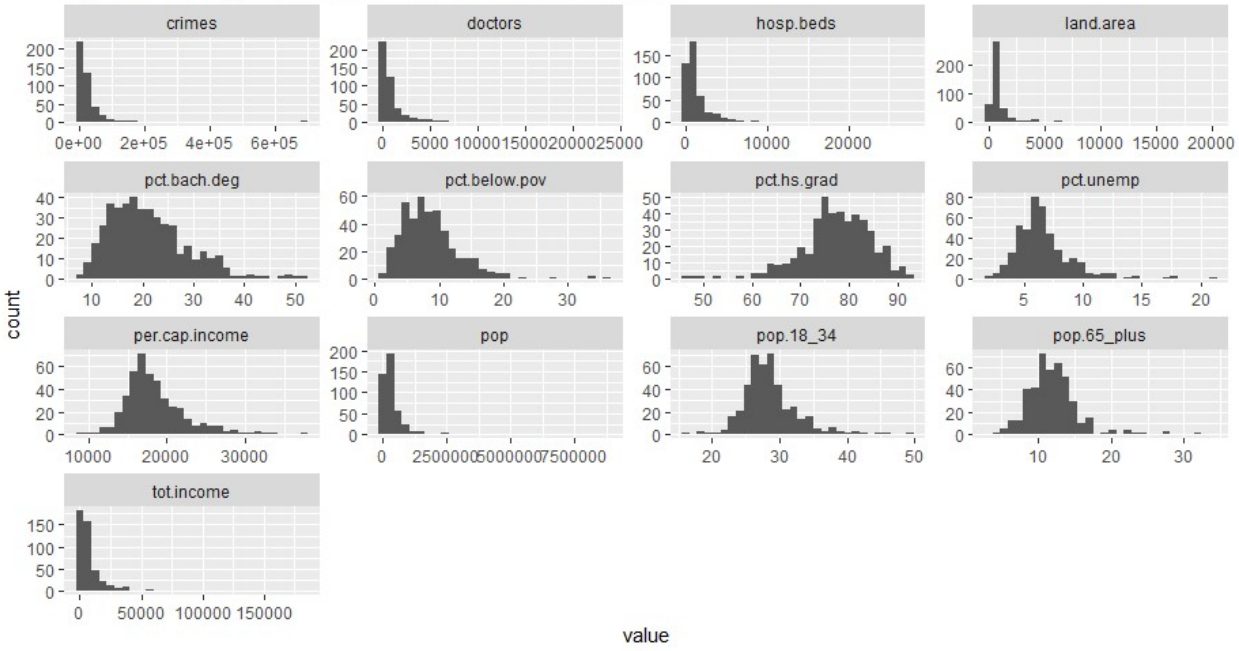
The data used to study factors affecting average income per person is taken from Kutneret al. (2005)1: It provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The definitions and sample size of the variables are given in Table 1 on this document.

Variable definitions for CDI data from Kutner et al. (2005)		
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	Percentage of population aged 18-34	Percent of 1990 CDI population aged 18–34
7	Percentage of population 65 or older	Percent of 1990 CDI population aged 65 or old
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI population that is unemployed
15	Per capita income	Per-capita income (i.e. average income per person) of 1990 CDI population (in dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification used by the US Bureau of the Census, NE (northeast region of the US), NC (north-central region of the US), S (southern region of the US), and W (Western region of the US)
Original source: Geospatial and Statistical Data Center, University of Virginia.		

Table 1

We examine the numeric variables and develop a description table for min, max, mean, etc. of the numeric variables. After doing exploratory data analysis, the distribution bar plots provide information on the distributions for the quantitative variables. We find out that several variables are apparently right skewed: Total Serious Crimes, Number of Active Physicians, Number of Hospital Beds, Land Area, Total Population, and Total personal income. And there is no missing value in the data set.

Figure 1: Histograms of continuous variables for CDI data



Except variable region, other categorical variables have a lot of unique values, so we only consider region in the categorical variables. Table 2 show number of samples for each of the four regions.

NC	NE	S	W
108	103	152	77

Table 2

Methods

For question one, we firstly do logarithm transformation on the apparently right skewed variables (Total Serious Crimes, Number of Active Physicians, Number of Hospital Beds, Land Area, Total Population, Per capita income, and Total personal income). Logarithm transformation is an appropriate way to make the distribution of variables normal because the transformed variables can be easily interpreted in real world. Then correlation graph is used to show the correlations between each pair of variables. Pairs with correlated coefficients are studied, and surprised correlations are explained with analysis.

For question two, we study whether crimes or per.cap.crime works better in the regression model with per.cap.income. We firstly fit a linear regression model for per capita income by logarithm transformed crime. We then include dummy variables for region in the model and test the linear models with and without interaction. After we chose the best model using crimes and that using per.cap.crime, we compare the two best models with AIC and BIC tests to check which model works better. Then we use

diagnosis plots to check if the regression model has some leverage points and unusual residual patterns or not.

For question three, because county, state and id are useless categorical variables, these variables are not considered. Besides, because per.cap.income is a deterministic function of population and total income, using these two variables will disturb other variables seriously. So, we will not consider these two variables and temporarily exclude Region. Then a multivariable regression model is fitted with logarithm transformation on the variables which skew extremely. By checking BIC, the best model with lowest BIC is selected. Then we use VIFs, residual diagnostics, and marginal model plots to check whether there is some obvious deficiency. Then we analyze if we should include variable Region and the interactions or not. We define a rule that if any interaction between a value of categorical variable and a numeric variable is significant, we keep the whole categorical variable; otherwise, we drop the interaction with that numeric variable. Then we use the rule to decide which interactions should be included. After we choose the final model, we use Anova, AIC, and BIC to check our selection. After we use all-subset to decide our variable subset for the final model, we use stepwise multivariable selection as another way. We both consider the model considering only the variables and the model considering interactions.

For question four, we firstly look at the number of unique states in our data set to make sure the data contains all 48 states in the continental U.S. Then we compare the average population of counties in our data set and the average population of counties in U.S. to consider the meaning of missing counties.

Results

Question 1. Relationships between the variables

From the bar plots for transformed variable (Figure 2), we find that the distribution for the continuous variables is now all relatively normal, so the models and plots analyses generated by these variables will be more convincing. From the correlation table (Figure 3), correlation for each pair of variables is shown. Log.Tot.Income has a strong positive correlation with log.pop, log.doctor, log.hosp.beds, and log.crimes, and log.per.capital.income. A reasonable person would expect a strong correlation between population and total income, number of active physicians(doctor), hospital beds because more people always mean larger Gross Domestic Product, more physicians, more hospitals and hospital beds and then larger total income. However, he/she may not expect a correlation between total income and crime. Considering number of crimes is calculated by population multiplying criminal rate, crime may have a strong positive correlation with total income because the positive correlation between crime and population. Pct.unemp has a strong negative correlation with pct.hs.grad and pct.bach.deg. It can be explained that high school and bachelor graduated students have sufficient knowledge to help them find a job. Seeing most of the correlations for a pair of variables are reasonable, we focus on the surprising ones. There are strong positive correlations between log.crimes and log.doctors and between log.crimes and log.hosp.beds. They are surprising because physician is a career with very low criminal rate in common sense. These surprising results appear because both of crimes and doctor have a strong correlation with population. High crimes always mean high population, and high population will then cause large number of active physicians. This logic also works for the correlation between log.hosp.beds and log.crimes. From the pair scatter plot in appendix Q1, it also shows apparent patterns between the pairs of variables which are strong correlated in the correlation graph.

Figure 2: Histograms of transformed continuous variables for CDI data

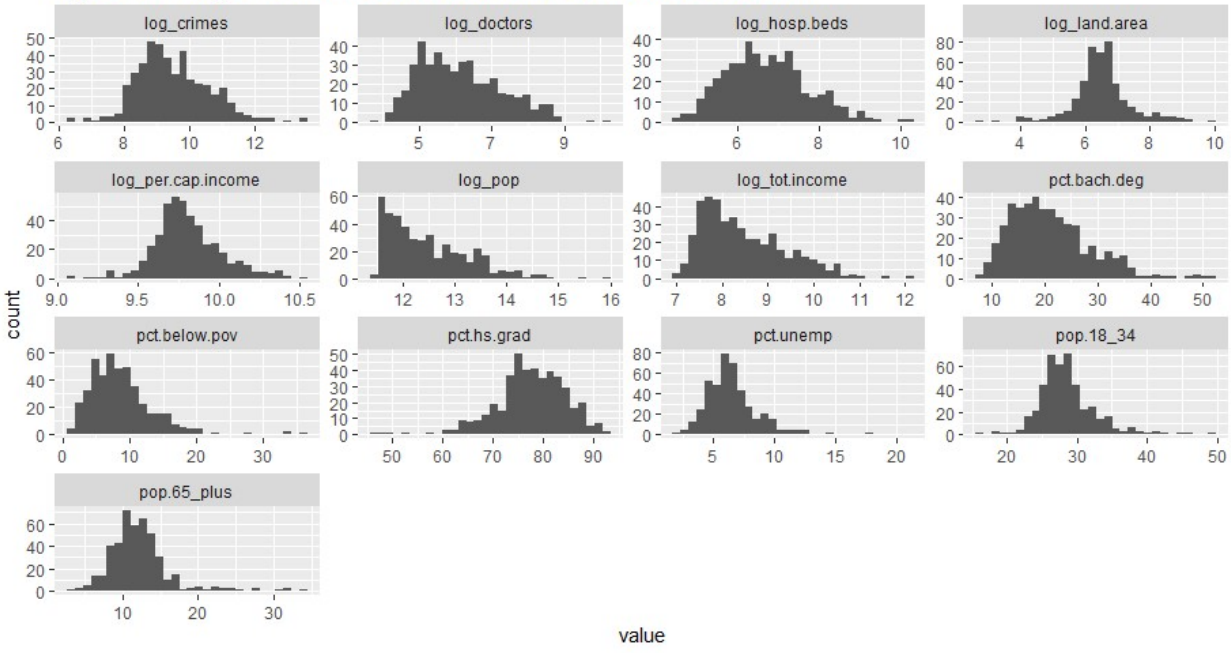
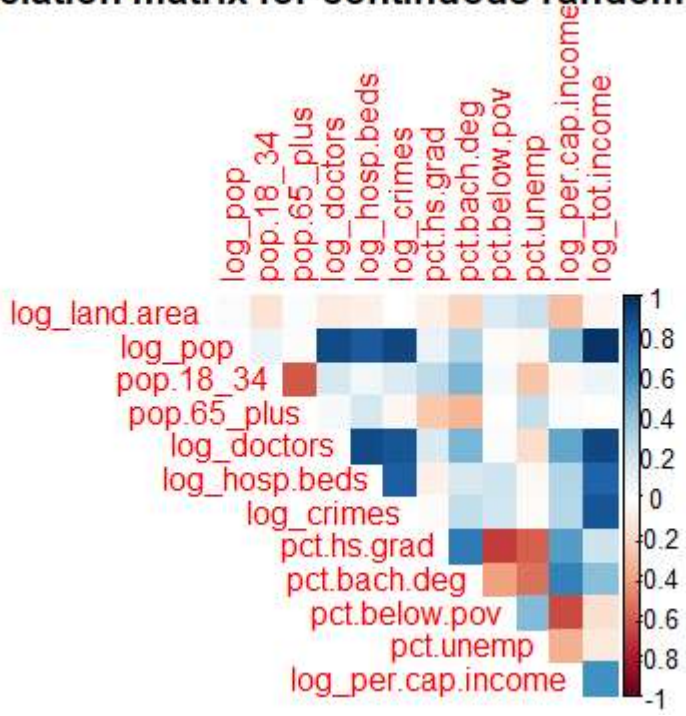


Figure 3: Correlation matrix for continuous random variables



After classifying all data by four regions, the summary tables show that the region with largest crimes median is W, the median for S is a bit less, and the medians for NC and NE are around 70% of median of W. However, means of crimes are quite different. Means of crimes in S, NC, and NE are similar, but the mean of crimes for W region is much larger. This means that the distribution graph for W, NC, NE region skews to the right and a small part of quite high values increase mean value significantly. Compared to

other three regions, S region has a more normal distribution. The difference between mean and median is not quite big.

Question 2. Crime and crime rate

Three models are fitted to investigate how crime relates to per capita income. The first model regresses logarithm of per capital income on logarithm of crimes. The second model regresses logarithm of per capital income on logarithm of crimes and region. The third model regresses logarithm of per capital income on logarithm of crimes, region, and the interaction between logarithm of crimes and region. It shows that a regression model with logarithm of crimes and region has a lowest P value and does the best. Three similar regression models using logarithm of per capital crime, region, and their interaction are also tested, and it shows that a model with both logarithm of per-capita crime and regions, without interaction works the best. Comparing the two best models and using AIC and BIC test models, we find that the model using total crime numbers performs better in both tests. For every 1% increase in crimes, we expect a 0.07% increase in per capital income, on average. With the coefficient estimates, the baseline per capital income in each region could be estimated. In the NC region, the baseline salary is $9,798.65. In the NE it is $10,829.18. In the S it is $8,955.29, and in the W, it is $9,228.02.

Then we use diagnosis plots to explore the characteristics of the two models with the best performance. In the model regresses on logarithm of crimes, there is not an apparent pattern in the residual vs fitted plot. The values of residual are centralized on 0. Also, the regression line is quite horizontal. There are significant left and right tails in the Q-Q plot, which means the residuals are not perfectly normal distributed. For the residual vs leverage plot, there are either some high leverage points or high residual points. Unlike the diagnosis plots for logarithm of crimes, the residuals vs fitted plot in the logarithm of crime rate diagnosis plots shows that the residuals are in two apparent clusters, which is a significant pattern that is needed to study. For the Q-Q plot and leverage vs residuals plot, this model has the similar attributes with those in the logarithm of crime model. Because the model analyses suggest that the crime model performs slightly better than the crime rate mode and has slightly better AIC and BIC, we prefer to use the model regresses on logarithm of crimes and region.

Question 3. Best model for predicting per capita income

We first use all-subset method to find out which subset of variable performs the best. Without the category variables (including region at the beginning) which can hardly show the correlation with logarithm of per capital income and the variables which will disturb the performance of other variables (population and total income), the multiple linear regression model shows that only logarithm of hosp beds and logarithm of crimes are not significant. By checking the lowest BIC with any numbers of variables. We find that the subset including log of land area, pop.18_34, log of doctors, pct high school grad, pct bachelor deg, pct below poverty, and pct unemployment has the lowest BIC among all the possible subsets. In VIFs test, there is not very big VIF value. In residual diagnosis, only the Q-Q plot has some negative pattern that there are both large right and left tail. In marginal model plots, the model also performs well: data-based curves line up well with model-based curves.

After select the best subset model, we need to consider whether region is an important variable in the variable selecting process. Under the role we defined, we keep region, interactions region:pct.hs.grad,

region:pct.below.pov, region:pct.unemp, and region:pct.bach.deg, and drop region:log.land.area, region:pop.18_34, and region:log.doctors. Using Anova, AIC, and BIC test, we find that except BIC test which always prefer easier models, other two models both prefer the model containing region. So, it is better to include the categorical variable region in the final regression model.

In stepwise multivariable selection, it shows that model with log.land.area, pop.65_plus, pop.18_34, log.doctor, pct.hs.grad, pct.bach.deg, pct.below.pov, and pct.unemp has the lowest AIC and is the best model with AIC backward stepwise selection, which adds pop.65_plus to the final model we get in the all-subset method. The model generated by BIC backward stepwise selection is the same as the final model in all-subset method. Because the estimate coefficient for pop.65_plus is very small, it will not make a great impact on the prediction of log of per.cap.income. And we do not need to worry about the difference. After we use stepwise selection method considering only the variables, we look for the best model considering interactions. Then we compare the performances of models with and without interactions (Table 3). It shows that both AIC and BIC like models with some interactions.

	df	AIC	BIC
backAIC	10	-944.89	-904.02
backBIC	9	-942.27	-905.49
backAIC2	27	-1064.73	-954.38
backBIC2	12	-1020.6	-971.56

Table 3

Question 4. Impact of missing states and missing counties

The data set is stratified across the 48 states in the continental U.S., so we can be sure that we are not excluding the states in the continental U.S. as we analyze the subset of the population. However, the data set does not include samples from Alaska, Hawaii, and Washington D.C. We can assume that the national information in Washington D.C. is similar as the states nearby, but Alaska and Hawaii are two individual states far away from the continental U.S. We consider whether it will be better to consider the information from these two states, but we think that these two states do not have a strong similarity with other states, or even significantly different from other states in another word. So, we think the lack of states will increase the effectiveness of our model and conclusion instead.

The mean population of counties in the data set is 393010.9 and there are 373 counties in total. From pewtrusts, it is said that there are 3,142 counties in the U.S. in 1990s. Considering the population in 1990 is 249.6 million, the average population of all counties in 1990 is 79439.8 which is around one fifth of the average population of counties in our data set. Considering counties with more population always have higher development degree, it would be safer to assume that the characteristics of counties in our data set are not similar with those of counties not in the data since their geography likely requires that their local economies and communities are driven by different aspects than the rest of the country. So, we conclude that it is more likely that the analyses and models in this research only explore the features of cities. If we want to explore the features of counties in America in general, more information for the small counties is necessary.

Discussion

According to the result, we find that per capital income is deterministic by population and the total income of that region. After excluding these two factors, per capital income will be influenced by population, number of physicians, number of hospital bed, number of crimes, and percentage below poverty. If we ignored other variables and only consider the number of crime or crime rate and region, we will find that per capital income will not be influenced significantly by crime compared to the combination of model regresses crime and region together. It shows that region is an important variable, and interactions between each numerical variable and different regions also perform significantly different. When we study the impact of crime and crime rate with region on per capital income, all the test results show that total number of crimes performs better. For the model selection part, by using all-subset and stepwise selection method, we conclude that per capital income will be significantly influenced by land area, population from 18 to 34 + number of physicians + percentage of high school grad, percentage of bachelor's degree, percentage below poverty, and percentage of unemployment. Population over 65 is a factor that can truly influence per capital income, but because of the small estimate coefficient, the impact will be small.

For the lack of some states and counties, we think that the missing states (Alaska, Hawaii, and Washington D.C.) will not disturb our conclusion about how demographic factors affect per capita income. Washington D.C has the similar characteristics with states surround, and Alaska and Hawaii have so special geographical positions that considering these two states will disturb the result instead. However, the missing counties can be regarded as a weakness of our research. Studying only the most populous counties will probably conclude the features for only urban regions. We need to consider the less populous parts in U.S. to find out the domestic characteristics.

To induct an unbiases prediction, eliminating the shortages of the data set is critical. Future work that could be done to handle these limitations include updating county demographic information and expanding the dataset to include more counties in the country. This might avoid potentially biased conclusion by not collecting data in the less populated counties.

References

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

Sheather, S.J. (2009), A Modern Approach to Regression with R. New York: Springer Science + Business Media LLC.

Appendix

Hongsheng Xie

10/29/2021

Question 1

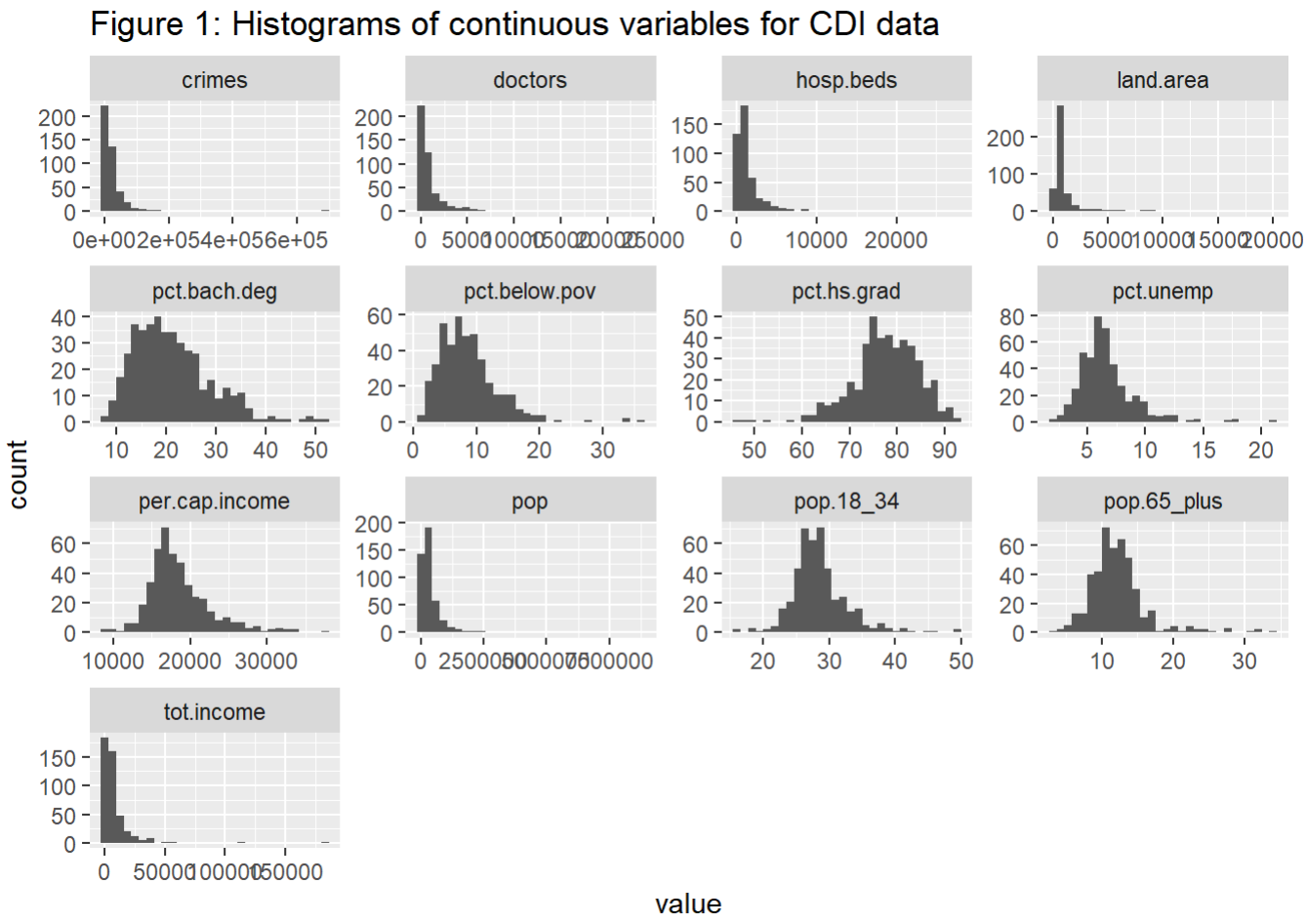
```
continue_series <- c(4,5,6,7,8,9,10,11,12,13,14,15,16)
category_series <- c(2, 3, 17)
df <- data.frame("Name" = colnames(data1),
                 "Min" = rep(0,17),
                 "1st Qu" = rep(0,17),
                 "Medidan" = rep(0,17),
                 "Mean" = rep(0,17),
                 "3rd Qu" = rep(0,17),
                 "Max" = rep(0,17))
for (i in continue_series) {
  s <- summary(data1[,i])
  df[i,2] <- s[1]
  df[i,3] <- s[2]
  df[i,4] <- s[3]
  df[i,5] <- s[4]
  df[i,6] <- s[5]
  df[i,7] <- s[6]
}
for (i in category_series) {
  df[i,c(2,7)] <- c(NA,NA,NA,NA,NA,NA)
}

knitr::kable(df[continue_series,], caption = "Table 2")
```

Table 2

	Name	Min	X1st.Qu	Medidan	Mean	X3rd.Qu	Max
4	land.area	15.0	451.250	656.50	1.041411e+03	946.750	20062.0
5	pop	100043.0	139027.250	217280.50	3.930109e+05	436064.500	8863164.0
6	pop.18_34	16.4	26.200	28.10	2.856841e+01	30.025	49.7
7	pop.65_plus	3.0	9.875	11.75	1.216977e+01	13.625	33.8
8	doctors	39.0	182.750	401.00	9.879977e+02	1036.000	23677.0
9	hosp.beds	92.0	390.750	755.00	1.458627e+03	1575.750	27700.0
10	crimes	563.0	6219.500	11820.50	2.711162e+04	26279.500	688936.0
11	pct.hs.grad	46.6	73.875	77.70	7.756068e+01	82.400	92.9
12	pct.bach.deg	8.1	15.275	19.70	2.108114e+01	25.325	52.3
13	pct.below.pov	1.4	5.300	7.90	8.720682e+00	10.900	36.3
14	pct.unemp	2.2	5.100	6.20	6.596591e+00	7.500	21.3
15	per.cap.income	8899.0	16118.250	17759.00	1.856148e+04	20270.000	37541.0
16	tot.income	1141.0	2311.000	3857.00	7.869273e+03	8654.250	184230.0

```
ggplot(gather(data1[,continue_series]), aes(x = value)) +
  geom_histogram() +
  facet_wrap(~key, scales = 'free') +
  labs(title = "Figure 1: Histograms of continuous variables for CDI data")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
table(data1[,17])
```

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

```
length(unique(data1[,2]))
```

```
## [1] 373
```

```
length(unique(data1[,3]))
```

```
## [1] 48
```

```
length(unique(data1[,17]))
```

```
## [1] 4
```

```
NC <- data1[data1$region == "NC",]
NE <- data1[data1$region == "NE",]
S <- data1[data1$region == "S",]
W <- data1[data1$region == "W",]

length(NC$id)
```

```
## [1] 108
```

```
length(NE$id)
```

```
## [1] 103
```

```
length(S$id)
```

```
## [1] 152
```

```
length(W$id)
```

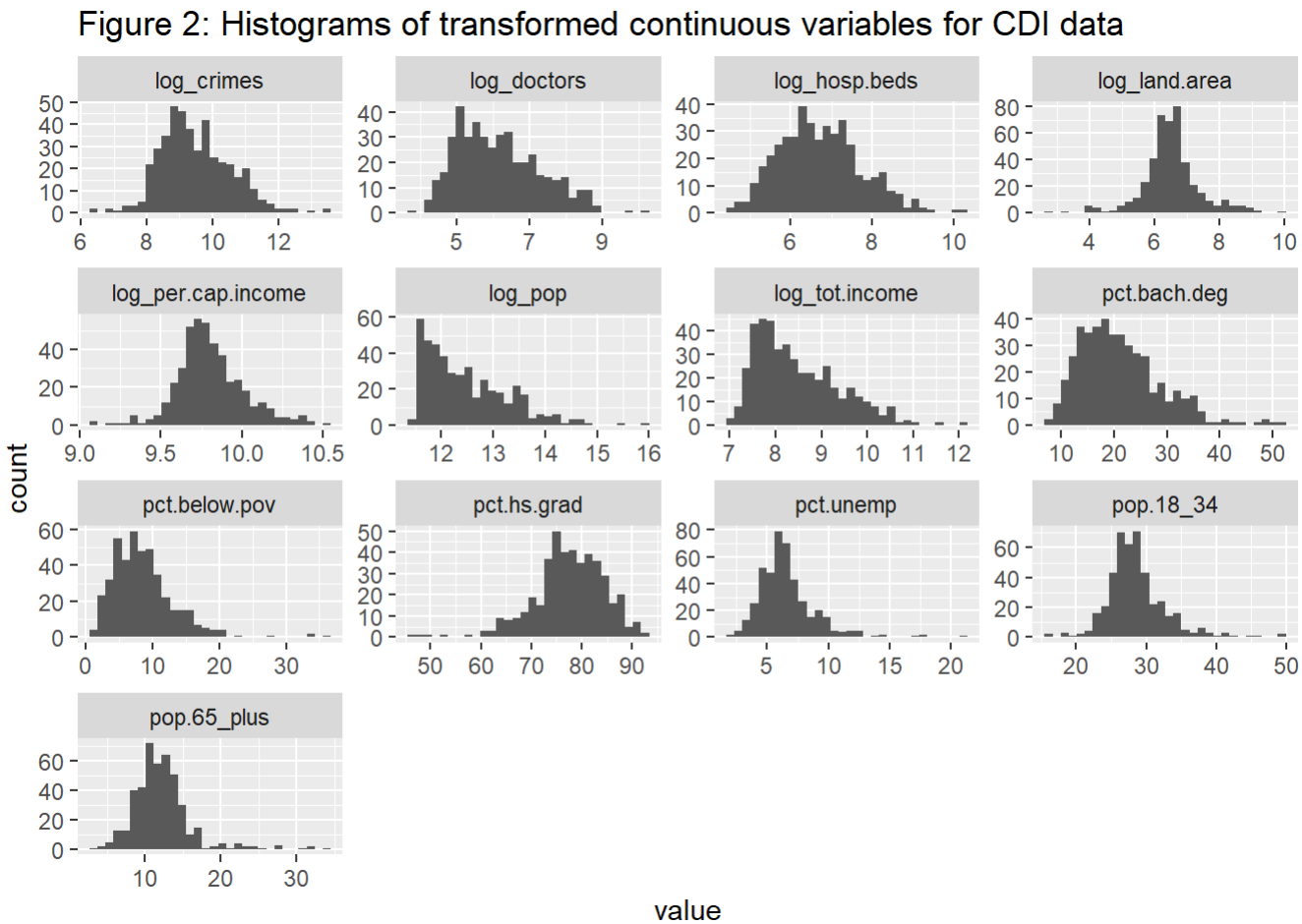
```
## [1] 77
```

```
df_na <- data.frame("IS.NA" = rep(TRUE,17),
                    "NA amount" = rep(0,17))

for (i in c(1:17)) {
  l <- data1[,i]
  s <- sum(is.na(l))
  if (s > 0) {
    df_na[i,1] = TRUE
    df_na[i,2] = s
  } else {
    df_na[i,1] = FALSE
    df_na[i,2] = 0
  }
}

# There is no missing value for any column.
```

```
data1$crimes <- log(data1$crimes)
data1$doctors <- log(data1$doctors)
data1$hosp.beds <- log(data1$hosp.beds)
data1$land.area <- log(data1$land.area)
data1$pop <- log(data1$pop)
data1$tot.income <- log(data1$tot.income)
data1$per.cap.income <- log(data1$per.cap.income)

data1 <- rename(data1, "log_crimes" = "crimes")
data1 <- rename(data1, "log_doctors" = "doctors")
data1 <- rename(data1, "log_hosp.beds" = "hosp.beds")
data1 <- rename(data1, "log_land.area" = "land.area")
data1 <- rename(data1, "log_pop" = "pop")
data1 <- rename(data1, "log_tot.income" = "tot.income")
data1 <- rename(data1, "log_per.cap.income" = "per.cap.income")
```

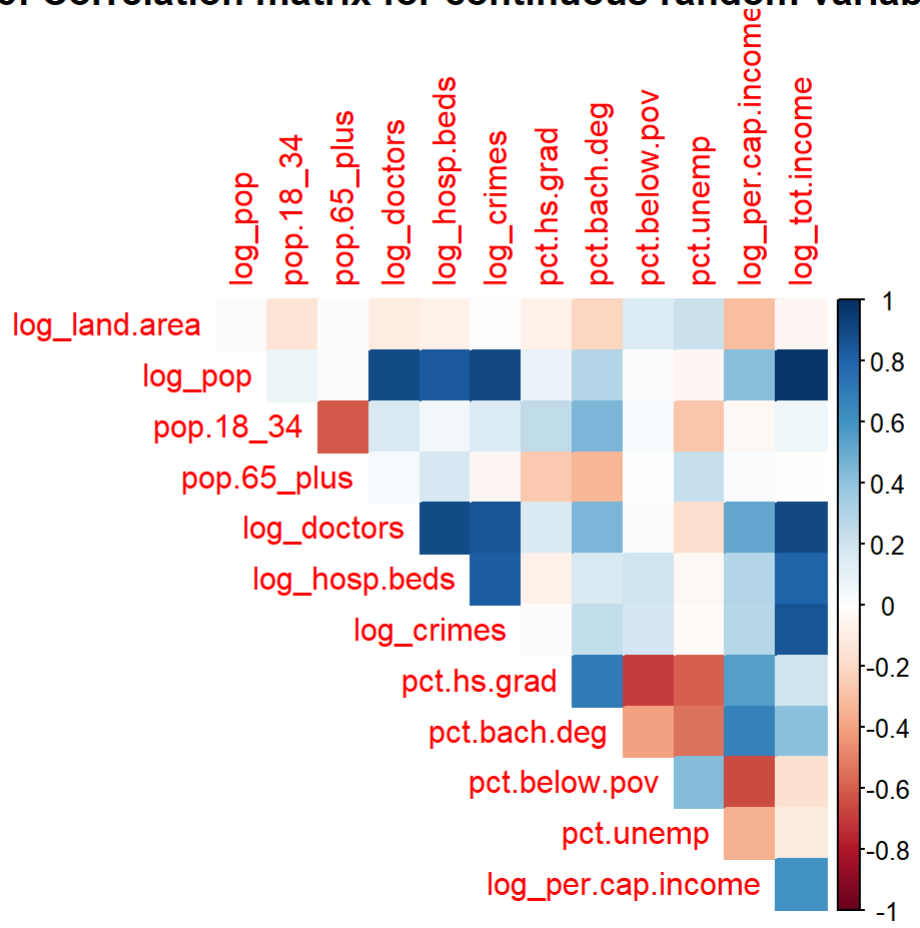
```
ggplot(gather(data1[,continue_series]), aes(x = value)) +
  geom_histogram() +
  facet_wrap(~key, scales = 'free') +
  labs(title = "Figure 2: Histograms of transformed continuous variables for CDI data")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

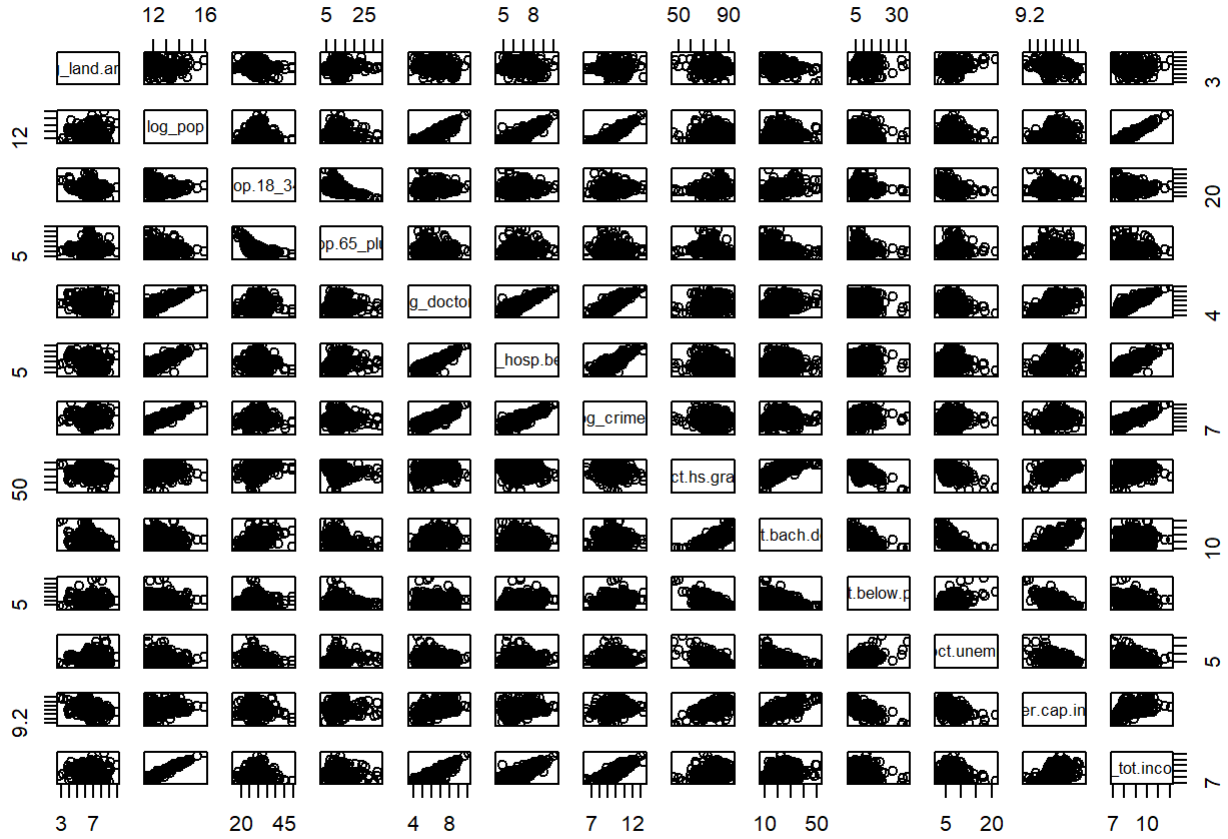


```
corrplot(cor(data1[,continue_series]), color = T, type = "upper",
title = "Figure 3: Correlation matrix for continuous random variables",
mar=c(0,0,1,0), diag = F, method = "color")
```

Figure 3: Correlation matrix for continuous random variables




```
pairs(~log_land.area+log_pop+pop.18_34+pop.65_plus+log_doctors+log_hosp.beds+log_crimes+pct.hs.g
rad+pct.bach.deg+pct.below.pov+pct.unemp+log_per.cap.income+log_tot.income, data = data1)
```




```
# The scatter matrix shows that the variables about amount like land area, population, amount of hospital beds are closely related because generally, more land area means more population and mo re crimes, hospitals and facilities. The percentage variables mostly skew to the right, and this phenomenon needs further study.
```

Question 2

```
per.cap.crimes <- exp(data1$log_crimes - data1$log_pop)
log_per.cap.crimes = data1$log_crimes - data1$log_pop
# ratio between two values equals the exponent of difference between logarithm of these two valu es.
l1 <- lm(log_per.cap.income ~ log_crimes, data = data1)
l1_r <- lm(log_per.cap.income ~ log_crimes + region, data = data1)
l1_inter <- lm(log_per.cap.income ~ region*log_crimes, data = data1)

anova(l1,l1_r,l1_inter)
```

```
## 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 ~ region * log_crimes
##   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
```

```
# It looks like ANCOVA model with no interactions and contain region as dummy value is doing the best.

l2<- lm(data1$log_per.cap.income ~ log_per.cap.crimes)
l2_CR <- lm(data1$log_per.cap.income ~ log_per.cap.crimes + data1$region)
l2_CR_inter <- lm(data1$log_per.cap.income ~ log_per.cap.crimes*region, data = data1)

anova(l2,l2_CR,l2_CR_inter)
```

```
## Analysis of Variance Table
##
## Model 1: data1$log_per.cap.income ~ log_per.cap.crimes
## Model 2: data1$log_per.cap.income ~ log_per.cap.crimes + data1$region
## Model 3: data1$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
```

```
# It looks like ANCOVA model with no interactions and contain region as dummy value is doing the best.


# If we want to compare these two winners, we need to use AIC or BIC, because the two winners are not nested models (you can’t get one from the other by imposing one or more linear constraints).

AIC(l1_r,l2_CR)
```

```
##       df       AIC
## l1_r   6 -227.4746
## l2_CR  6 -172.1347
```

```
BIC(l1_r,l2_CR)
```

```
##       df       BIC
## l1_r   6 -202.9539
## l2_CR  6 -147.6140
```

```
# It is apparent that model using log_crimes has both smaller AIC and BIC.

coef(summary(l1_r))
```

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  9.18843110 0.079812437 115.125305 0.000000e+00
## log_crimes   0.06669491 0.008421114   7.919963 2.002771e-14
## regionNE     0.10445836 0.025531314   4.091382 5.110827e-05
## regionS     -0.08698350 0.023617956  -3.682939 2.595887e-04
## regionW     -0.05527965 0.028167096  -1.962561 5.033416e-02
```

```
# So, according to the model, the level of salary varies with region in the US, but the way it i
s related to crime does not.

par(mfrow=c(2,2))
plot(l1_r)
```

```
formula(l1_r)
```

```
## log_per.cap.income ~ log_crimes + region
```

```
# There is not an apparent pattern in the residual vs fitted plot. The values of residual are ce
ntralized on 0. Also the regression line is quite horizontal.There are significant left and righ
t tails in the Q-Q plot, which means the residuals are not perfectly normal distributed. For the
residual vs leverage plot, there are either some high leverage points and high residual points.
plot(l2_CR)
```

```
formula(l2_CR)
```

```
## data1$log_per.cap.income ~ log_per.cap.crimes + data1$region
```

```
# Unlike the diagnosis plots for log_crimes, the residuals vs fitted plot in the log_crime_rate
 diagnosis plots shows that the residuals are in two apparent clusters, which is a significant p
attern that is needed to study. For the Q-Q plot and leverage vs residuals plot, this model has
 the similar attributes with those in the log_crime model.
```

Question 3 #ALL-SUBSET

```
l <- lm(log_per.cap.income~log_land.area+pop.18_34+pop.65_plus+log_doctors+log_hosp.beds+log_cri
mes+pct.hs.grad+pct.bach.deg+pct.below.pov+pct.unemp, data = data1)

summary(l)
```

```
##
## Call:
## lm(formula = log_per.cap.income ~ log_land.area + pop.18_34 +
##     pop.65_plus + log_doctors + log_hosp.beds + log_crimes +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp, data = data1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35561 -0.04712 -0.00846  0.04522  0.27681
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   10.313312   0.108644  94.927  < 2e-16 ***
## log_land.area -0.035641   0.004825  -7.386 7.95e-13 ***
## pop.18_34     -0.015542   0.001306 -11.897  < 2e-16 ***
## pop.65_plus   -0.003309   0.001371  -2.413   0.0162 *
## log_doctors    0.052055   0.012502   4.164 3.79e-05 ***
## log_hosp.beds  0.016215   0.012008   1.350   0.1776
## log_crimes    -0.004066   0.007831  -0.519   0.6039
## pct.hs.grad   -0.004738   0.001086  -4.363 1.61e-05 ***
## pct.bach.deg   0.015712   0.001027  15.305  < 2e-16 ***
## pct.below.pov -0.024945   0.001303 -19.138  < 2e-16 ***
## pct.unemp      0.011130   0.002186   5.091 5.34e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08168 on 429 degrees of freedom
## Multiple R-squared:  0.8475, Adjusted R-squared:  0.8439
## F-statistic: 238.4 on 10 and 429 DF,  p-value: < 2.2e-16
```

```
# Only logarithm of hosp_bed and logarithm crimes are not significant.
```

```
library(leaps)
library(car)
all.subsets1 <- regsubsets(log_per.cap.income ~ log_land.area+pop.18_34+pop.65_plus+log_doctors+
log_hosp.beds+log_crimes+pct.hs.grad+pct.bach.deg+pct.below.pov+pct.unemp, data = data1)

summary(all.subsets1)
```

```
## Subset selection object
## Call: regsubsets.formula(log_per.cap.income ~ log_land.area + pop.18_34 +
##     pop.65_plus + log_doctors + log_hosp.beds + log_crimes +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp, data = data1)
## 10 Variables  (and intercept)
##               Forced in Forced out
## log_land.area     FALSE      FALSE
## pop.18_34         FALSE      FALSE
## pop.65_plus       FALSE      FALSE
## log_doctors       FALSE      FALSE
## log_hosp.beds     FALSE      FALSE
## log_crimes        FALSE      FALSE
## pct.hs.grad       FALSE      FALSE
## pct.bach.deg      FALSE      FALSE
## pct.below.pov     FALSE      FALSE
## pct.unemp         FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          log_land.area pop.18_34 pop.65_plus log_doctors log_hosp.beds
## 1  ( 1 ) " "           " "       " "         " "          " "
## 2  ( 1 ) " "           " "       " "         "*"          " "
## 3  ( 1 ) " "           " "       " "         "*"          " "
## 4  ( 1 ) " "           "*"       " "         "*"          " "
## 5  ( 1 ) "*"           "*"       " "         "*"          " "
## 6  ( 1 ) "*"           "*"       " "         "*"          " "
## 7  ( 1 ) "*"           "*"       " "         "*"          " "
## 8  ( 1 ) "*"           "*"       "*"         "*"          " "
##          log_crimes pct.hs.grad pct.bach.deg pct.below.pov pct.unemp
## 1  ( 1 ) " "        " "         "*"          " "           " "
## 2  ( 1 ) " "        " "         " "          "*"           " "
## 3  ( 1 ) " "        " "         "*"          "*"           " "
## 4  ( 1 ) " "        " "         "*"          "*"           " "
## 5  ( 1 ) " "        " "         "*"          "*"           " "
## 6  ( 1 ) " "        " "         "*"          "*"           "*"
## 7  ( 1 ) " "        "*"         "*"          "*"           "*"
## 8  ( 1 ) " "        "*"         "*"          "*"           "*"
```

```
best.model <- which.min(summary(all.subsets1)$bic)

coef(all.subsets1,best.model)
```

```
##   (Intercept) log_land.area     pop.18_34   log_doctors   pct.hs.grad
##  10.222495041  -0.035674062  -0.013900201   0.060676872  -0.004406396
##  pct.bach.deg pct.below.pov     pct.unemp
##   0.015385301  -0.024278371   0.010603691
```

```
continue_series_use = c(4,6,7,8,9,10,11,12,13,14,15)
data_use <- data1[,continue_series_use]
tmp <- data_use[,summary(all.subsets1)$which[best.model,][-1]]
final_model <- lm(log_per.cap.income ~ .,data=tmp)
summary(final_model)$coef
```

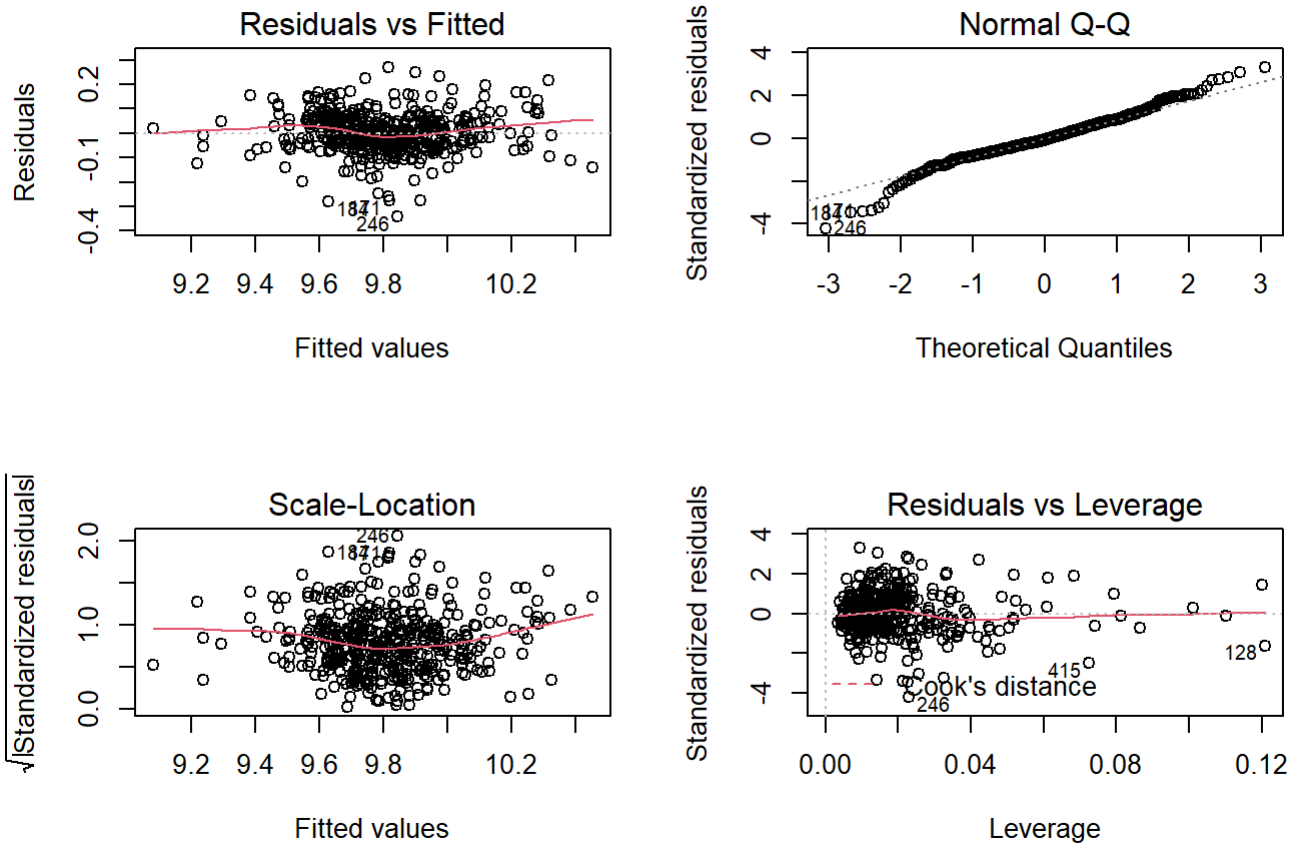
```
##                   Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)   10.222495041 0.0931210074 109.776465 1.127483e-317
## log_land.area -0.035674062 0.0047767371  -7.468291  4.533156e-13
## pop.18_34     -0.013900201 0.0011113007 -12.508046  7.514862e-31
## log_doctors    0.060676872 0.0040183327  15.100012  1.133432e-41
## pct.hs.grad   -0.004406396 0.0010822796  -4.071403  5.558448e-05
## pct.bach.deg   0.015385301 0.0009245509  16.640838  2.100590e-48
## pct.below.pov -0.024278371 0.0012583372 -19.294011  2.812246e-60
## pct.unemp      0.010603691 0.0021771148   4.870525  1.564524e-06
```

VIFs, diagnosis and mmp check

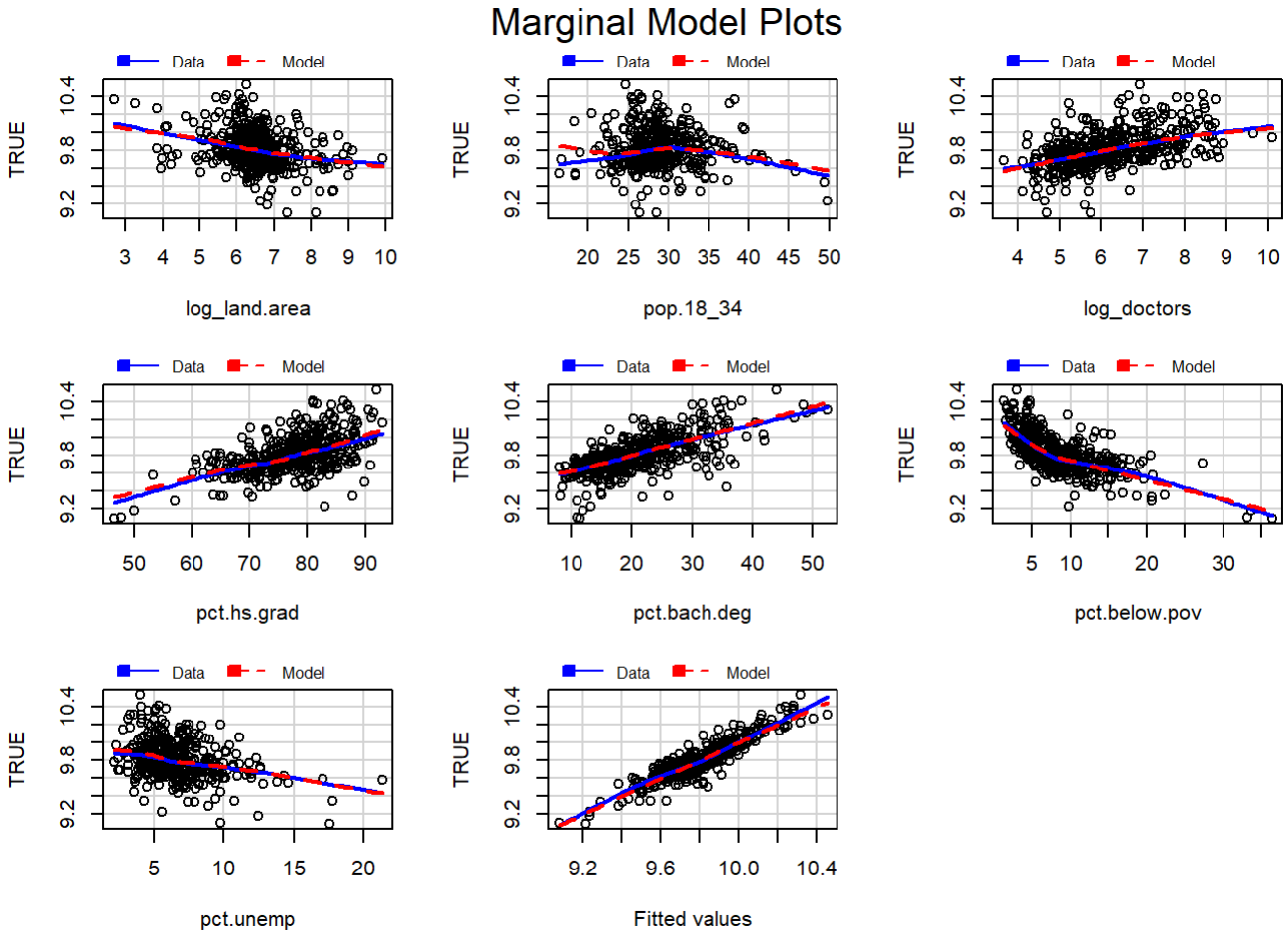
```
vif(final_model)
```

```
## log_land.area     pop.18_34   log_doctors   pct.hs.grad  pct.bach.deg
##      1.131867      1.416145      1.379671      3.763103      3.269565
## pct.below.pov     pct.unemp
##      2.241555      1.691280
```

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




```
mmps(final_model)
```



Check Region

```
tmp <- cbind(tmp,region=data1$region)
model_region <- lm(log_per.cap.income ~ .*region,data=tmp)
summary(model_region)
```

```
##
## Call:
## lm(formula = log_per.cap.income ~ . * region, data = tmp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.250782 -0.042332 -0.002298  0.040559  0.313570
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              10.1244260  0.2826240  35.823  < 2e-16 ***
## log_land.area            -0.0364187  0.0151355  -2.406 0.016564 *
## pop.18_34                -0.0147940  0.0026043  -5.681 2.55e-08 ***
## log_doctors               0.0544169  0.0093221   5.837 1.08e-08 ***
## pct.hs.grad              -0.0024773  0.0034110  -0.726 0.468088
## pct.bach.deg              0.0140833  0.0029254   4.814 2.09e-06 ***
## pct.below.pov            -0.0237085  0.0036234  -6.543 1.81e-10 ***
## pct.unemp                 0.0180393  0.0048923   3.687 0.000257 ***
## regionNE                  0.3243992  0.3577081   0.907 0.365004
## regionS                  -0.0345856  0.3131668  -0.110 0.912116
## regionW                   1.5043946  0.4226868   3.559 0.000416 ***
## log_land.area:regionNE   -0.0037179  0.0201435  -0.185 0.853656
## log_land.area:regionS    -0.0047582  0.0174155  -0.273 0.784825
## log_land.area:regionW     0.0151234  0.0181871   0.832 0.406154
## pop.18_34:regionNE       -0.0024780  0.0036873  -0.672 0.501939
## pop.18_34:regionS        -0.0008777  0.0030680  -0.286 0.774970
## pop.18_34:regionW         0.0014122  0.0040925   0.345 0.730220
## log_doctors:regionNE     -0.0046251  0.0132571  -0.349 0.727359
## log_doctors:regionS       0.0043337  0.0114401   0.379 0.705019
## log_doctors:regionW      -0.0034863  0.0131576  -0.265 0.791173
## pct.hs.grad:regionNE     -0.0037529  0.0044150  -0.850 0.395813
## pct.hs.grad:regionS       0.0021198  0.0037853   0.560 0.575790
## pct.hs.grad:regionW      -0.0190188  0.0045881  -4.145 4.13e-05 ***
## pct.bach.deg:regionNE     0.0069429  0.0040312   1.722 0.085776 .
## pct.bach.deg:regionS     -0.0015774  0.0032000  -0.493 0.622328
## pct.bach.deg:regionW      0.0071026  0.0036374   1.953 0.051541 .
## pct.below.pov:regionNE   -0.0014134  0.0050896  -0.278 0.781381
## pct.below.pov:regionS     0.0072764  0.0040739   1.786 0.074827 .
## pct.below.pov:regionW    -0.0161639  0.0054271  -2.978 0.003071 **
## pct.unemp:regionNE       -0.0083596  0.0073758  -1.133 0.257720
## pct.unemp:regionS        -0.0249396  0.0065867  -3.786 0.000176 ***
## pct.unemp:regionW        -0.0201466  0.0067713  -2.975 0.003101 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0759 on 408 degrees of freedom
## Multiple R-squared:  0.8747, Adjusted R-squared:  0.8652
## F-statistic: 91.91 on 31 and 408 DF,  p-value: < 2.2e-16
```

```
# If any interaction between a value of categorical variable and a numeric variable is significa
nt, we keep the whole categorical variable. Otherwise we drop the interaction with that numeric
 variable.

final_model_region <- update(model_region, . ~ . - region:log_land.area -
region:pop.18_34 - region:log_doctors)
summary(final_model_region)
```

```
##
## Call:
## lm(formula = log_per.cap.income ~ log_land.area + pop.18_34 +
##     log_doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp + region + pct.hs.grad:region + pct.bach.deg:region +
##     pct.below.pov:region + pct.unemp:region, data = tmp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.268015 -0.043459 -0.002511  0.039967  0.313939
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.125260   0.251582  40.246  < 2e-16 ***
## log_land.area          -0.034569   0.005376  -6.430 3.50e-10 ***
## pop.18_34              -0.015404   0.001087 -14.170  < 2e-16 ***
## log_doctors             0.055342   0.004034  13.720  < 2e-16 ***
## pct.hs.grad            -0.002503   0.003151  -0.794 0.427456
## pct.bach.deg            0.014208   0.002108   6.741 5.24e-11 ***
## pct.below.pov          -0.023634   0.003351  -7.054 7.30e-12 ***
## pct.unemp               0.017787   0.004783   3.719 0.000228 ***
## regionNE                0.219429   0.302526   0.725 0.468661
## regionS                -0.062648   0.276125  -0.227 0.820627
## regionW                 1.629351   0.357633   4.556 6.86e-06 ***
## pct.hs.grad:regionNE   -0.003640   0.003876  -0.939 0.348271
## pct.hs.grad:regionS     0.002014   0.003539   0.569 0.569690
## pct.hs.grad:regionW    -0.018916   0.004204  -4.499 8.85e-06 ***
## pct.bach.deg:regionNE   0.005905   0.002618   2.256 0.024611 *
## pct.bach.deg:regionS   -0.001298   0.002321  -0.559 0.576352
## pct.bach.deg:regionW    0.006326   0.002620   2.415 0.016183 *
## pct.below.pov:regionNE -0.002435   0.004647  -0.524 0.600488
## pct.below.pov:regionS   0.007137   0.003686   1.937 0.053482 .
## pct.below.pov:regionW  -0.015224   0.005169  -2.945 0.003407 **
## pct.unemp:regionNE     -0.007967   0.007255  -1.098 0.272761
## pct.unemp:regionS      -0.024668   0.006377  -3.868 0.000127 ***
## pct.unemp:regionW      -0.019757   0.006603  -2.992 0.002935 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07545 on 417 degrees of freedom
## Multiple R-squared:  0.8735, Adjusted R-squared:  0.8668
## F-statistic: 130.9 on 22 and 417 DF,  p-value: < 2.2e-16
```

```
# Now every numeric variable has at least one significant interaction with region.
```

Check models with and without region

```
anova(final_model,final_model_region)
```

```
## Analysis of Variance Table
##
## Model 1: log_per.cap.income ~ log_land.area + pop.18_34 + log_doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log_per.cap.income ~ log_land.area + pop.18_34 + log_doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##     region + pct.hs.grad:region + pct.bach.deg:region + pct.below.pov:region +
##     pct.unemp:region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    432 2.9051
## 2    417 2.3736 15   0.53148 6.2249 6.673e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
AIC(final_model, final_model_region)
```

```
##                    df       AIC
## final_model         9  -942.274
## final_model_region 24 -1001.178
```

```
BIC(final_model, final_model_region)
```

```
##                    df       BIC
## final_model         9 -905.4931
## final_model_region 24 -903.0957
```

```
# Anova and AIC tests perfer the model with region and the BIC tests prefer the model without re
gion.
```

Stepwise Variables Selection

```
stepwise_base <- lm(log_per.cap.income ~ ., data = data_use)

backAIC <- step(stepwise_base,direction="backward", data = data_use, k=2)
```

```
## Start:  AIC=-2193.54
## log_per.cap.income ~ log_land.area + pop.18_34 + pop.65_plus +
##     log_doctors + log_hosp.beds + log_crimes + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## - log_crimes     1   0.00180 2.8636 -2195.3
## - log_hosp.beds  1   0.01216 2.8740 -2193.7
## <none>                       2.8618 -2193.5
## - pop.65_plus    1   0.03884 2.9006 -2189.6
## - log_doctors    1   0.11565 2.9775 -2178.1
## - pct.hs.grad    1   0.12699 2.9888 -2176.4
## - pct.unemp      1   0.17289 3.0347 -2169.7
## - log_land.area  1   0.36392 3.2257 -2142.9
## - pop.18_34      1   0.94423 3.8060 -2070.1
## - pct.bach.deg   1   1.56251 4.4243 -2003.8
## - pct.below.pov  1   2.44318 5.3050 -1924.0
##
## Step:  AIC=-2195.27
## log_per.cap.income ~ log_land.area + pop.18_34 + pop.65_plus +
##     log_doctors + log_hosp.beds + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## - log_hosp.beds  1   0.01116 2.8748 -2195.6
## <none>                       2.8636 -2195.3
## - pop.65_plus    1   0.03709 2.9007 -2191.6
## - pct.hs.grad    1   0.12662 2.9902 -2178.2
## - log_doctors    1   0.12889 2.9925 -2177.9
## - pct.unemp      1   0.17123 3.0348 -2171.7
## - log_land.area  1   0.37492 3.2385 -2143.1
## - pop.18_34      1   0.94270 3.8063 -2072.1
## - pct.bach.deg   1   1.59514 4.4587 -2002.4
## - pct.below.pov  1   2.47345 5.3371 -1923.3
##
## Step:  AIC=-2195.55
## log_per.cap.income ~ log_land.area + pop.18_34 + pop.65_plus +
##     log_doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       2.8748 -2195.6
## - pop.65_plus    1   0.03031 2.9051 -2192.9
## - pct.hs.grad    1   0.12309 2.9978 -2179.1
## - pct.unemp      1   0.16432 3.0391 -2173.1
## - log_land.area  1   0.38995 3.2647 -2141.6
## - pop.18_34      1   0.93157 3.8063 -2074.1
## - log_doctors    1   1.55295 4.4277 -2007.5
## - pct.bach.deg   1   1.80755 4.6823 -1982.9
## - pct.below.pov  1   2.53302 5.4078 -1919.5
```

```
anova(final_model, backAIC)
```

```
## Analysis of Variance Table
##
## Model 1: log_per.cap.income ~ log_land.area + pop.18_34 + log_doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log_per.cap.income ~ log_land.area + pop.18_34 + pop.65_plus +
##     log_doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    432 2.9051
## 2    431 2.8748  1  0.030306 4.5437 0.03361 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# backward AIC add pop.65_plus compare to the all-subset method final model

backBIC <- step(stepwise_base,direction="backward", data = data_use, k=log(dim(data_use)[1]))
```

```
## Start:  AIC=-2148.59
## log_per.cap.income ~ log_land.area + pop.18_34 + pop.65_plus +
##     log_doctors + log_hosp.beds + log_crimes + pct.hs.grad +
##     pct.bach.deg + pct.below.pov + pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## - log_crimes     1   0.00180 2.8636 -2154.4
## - log_hosp.beds  1   0.01216 2.8740 -2152.8
## - pop.65_plus    1   0.03884 2.9006 -2148.7
## <none>                       2.8618 -2148.6
## - log_doctors    1   0.11565 2.9775 -2137.2
## - pct.hs.grad    1   0.12699 2.9888 -2135.6
## - pct.unemp      1   0.17289 3.0347 -2128.9
## - log_land.area  1   0.36392 3.2257 -2102.0
## - pop.18_34      1   0.94423 3.8060 -2029.2
## - pct.bach.deg   1   1.56251 4.4243 -1963.0
## - pct.below.pov  1   2.44318 5.3050 -1883.1
##
## Step:  AIC=-2154.4
## log_per.cap.income ~ log_land.area + pop.18_34 + pop.65_plus +
##     log_doctors + log_hosp.beds + pct.hs.grad + pct.bach.deg +
##     pct.below.pov + pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## - log_hosp.beds  1   0.01116 2.8748 -2158.8
## - pop.65_plus    1   0.03709 2.9007 -2154.8
## <none>                       2.8636 -2154.4
## - pct.hs.grad    1   0.12662 2.9902 -2141.4
## - log_doctors    1   0.12889 2.9925 -2141.1
## - pct.unemp      1   0.17123 3.0348 -2134.9
## - log_land.area  1   0.37492 3.2385 -2106.3
## - pop.18_34      1   0.94270 3.8063 -2035.3
## - pct.bach.deg   1   1.59514 4.4587 -1965.7
## - pct.below.pov  1   2.47345 5.3371 -1886.5
##
## Step:  AIC=-2158.77
## log_per.cap.income ~ log_land.area + pop.18_34 + pop.65_plus +
##     log_doctors + pct.hs.grad + pct.bach.deg + pct.below.pov +
##     pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## - pop.65_plus    1   0.03031 2.9051 -2160.2
## <none>                       2.8748 -2158.8
## - pct.hs.grad    1   0.12309 2.9978 -2146.4
## - pct.unemp      1   0.16432 3.0391 -2140.4
## - log_land.area  1   0.38995 3.2647 -2108.9
## - pop.18_34      1   0.93157 3.8063 -2041.3
## - log_doctors    1   1.55295 4.4277 -1974.8
## - pct.bach.deg   1   1.80755 4.6823 -1950.2
## - pct.below.pov  1   2.53302 5.4078 -1886.8
##
## Step:  AIC=-2160.25
## log_per.cap.income ~ log_land.area + pop.18_34 + log_doctors +
```

```
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       2.9051 -2160.2
## - pct.hs.grad    1   0.11147 3.0165 -2149.8
## - pct.unemp      1   0.15952 3.0646 -2142.8
## - log_land.area  1   0.37507 3.2801 -2112.9
## - pop.18_34      1   1.05209 3.9572 -2030.3
## - log_doctors    1   1.53330 4.4384 -1979.8
## - pct.bach.deg   1   1.86219 4.7673 -1948.4
## - pct.below.pov  1   2.50333 5.4084 -1892.9
```

```
#log_land.area + pop.18_34 + log_doctors + pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unem
p
anova(final_model, backBIC)
```

```
## Analysis of Variance Table
##
## Model 1: log_per.cap.income ~ log_land.area + pop.18_34 + log_doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
## Model 2: log_per.cap.income ~ log_land.area + pop.18_34 + log_doctors +
##     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1    432 2.9051
## 2    432 2.9051  0         0
```

```
summary(backAIC)$coef
```

```
##                   Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)   10.315966592 0.1025857893 100.559411 2.359405e-301
## log_land.area -0.036493494 0.0047727720  -7.646184  1.360706e-13
## pop.18_34     -0.015348817 0.0012987646 -11.818014  4.136902e-28
## pop.65_plus   -0.002766377 0.0012977992  -2.131591  3.360555e-02
## log_doctors    0.062605267 0.0041029328  15.258663  2.438771e-42
## pct.hs.grad   -0.004657948 0.0010843088  -4.295776  2.153275e-05
## pct.bach.deg   0.015214937 0.0009242442  16.462032  1.361311e-47
## pct.below.pov -0.024614405 0.0012630840 -19.487544  4.083797e-61
## pct.unemp      0.010768825 0.0021696234   4.963454  9.990989e-07
```

```
# The estimate coefficient for pop.65_plus is extremely small, so the impact on the final predic
tion of per.cap.income is very small.b
```

Stepwise with interaction

```
backAIC2 <- stepAIC(stepwise_base,scope=list(lower = ~ 1, upper = ~ .^2),
k=2, trace=F)
backBIC2 <- stepAIC(stepwise_base,scope=list(lower = ~ 1, upper = ~ .^2),
k=log(dim(data_use)[1]), trace=F)

comparison <- cbind(AIC(backAIC,backBIC,backAIC2,backBIC2),
BIC(backAIC,backBIC,backAIC2,backBIC2))

comparison <- comparison[,-3]

names(comparison) <- c("df","AIC","BIC")
comparison %>% kbl(booktabs=T) %>% kable_classic()
```

	df	AIC	BIC
backAIC	10	-944.8883	-904.0206
backBIC	9	-942.2740	-905.4931
backAIC2	27	-1064.7253	-954.3824
backBIC2	12	-1020.6026	-971.5613

Question 4

```
length(unique(data1$state))
```

```
## [1] 48
```

```
length(unique(data1$county)) # number of counties in the data set
```

```
## [1] 373
```

```
mean(exp(data1$log_pop)) # Average population in counties in the data set
```

```
## [1] 393010.9
```