Project-1

Yuqing Xu

10/17/2021

Social science and economic factors that can affect per-capita income in each US county

Abstract

The project focuses on the relationship between income per capita and other variables associated with the economic, health, and social well-being values in each county in the US. The file cdi.dat is taken from Kutneret al. (2005) and provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. The project uses stepwise variable selection in both direction, and all subsets regression to build two potentially best fitting models; diagnostic plots, AIC and BIC values, and some interpretation under social and economic context are used to select the best one as the final fitting model. $per.cap.income\ stateCA + stateNJ + stateUT + log(land.area) + log(doctors) + log(pct.bach.deg) + log(pct.below.pov) + pop.18_34$ is the final model selected, which has smaller AIC and BIC values, and can be reasonably interpreted for people in social science backgrounds, and also has the numeric variables with the signs of correlations matching the expectatation. For furthre exploration on this topic, the effect of region/state variable with other variables as interaction terms should be addressed, and some missing data should be collected and added on if possible to eliminate bias.

Introduction

Income is a factor that can reflect people's living quality, and thus the average income can be a factor that shows how well people live in a certain area, which is important in many social problems. The project mainly focuses on how income per capita was related to other variables associated with the county's economic, health, and social well-being like population, crimes, and education. And this research problem may give some perspectives on how to improve in any aspect to improve the income overall. The questions will be addressed related to the topic in this project are: 1. Is there any relation between variables in datasets? 2. How per-capita income was related to crime number, and does different regions of the country matter for this relationship? Is it better or more reasonable to use (number of crimes)/(population)? 3. Find the best model predicting per-capita income from the other variables, which best reflects the social science and the meaning of the variables. 4. Should we be worried about either the missing states or the missing counties?

Data

The file cdi.dat is taken from Kutneret al. (2005). The data 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 variables we are going to use to build models are: id: Identification

number from 1 to 440
county: County name
state: Two-letter state abbreviation
land.area: Land area (square miles)
pop: Estimated 1990 total population
pop.18_34: Percent of 1990 CDI population aged 18–34
pop.65_plus: Percent of 1990 CDI population aged 65 or old
doctors: Number of professionally active nonfederal physicians during 1990 hosp.beds: Total number of hospital beds, cribs, and bassinets during 1990
crimes: Total number of serious crimes in 1990, including murder, rape, rob-bery, aggravated assault, bur-glary, larceny-theft, and motor vehicletheft, as reported by law enforcement agencies
pct.hs.grad: Percent of adult population (persons 25 years old or older) who com-pleted 12 or more years of school pct.bach.deg: Percent of adult population (persons 25 years old or older) with bach-elor's degree
pct.below.pov: Percent of 1990 CDI population with income below poverty level
pct.unemp: Percent of 1990 CDI population that is unemployed Y variable=per.cap.income: Per-capita income (i.e. average income per person) of 1990 CDI pop-ulation (in dollars) tot.income: Total personal income of 1990 CDI population (in millions of dollars)
region: Geographic region classification used by the US Bureau of the Cen-sus, NE (northeast region of the US), NC (north-central region of theUS), S (southern region of the US), and W (Western region of the US)

We would not consider id and county as variables, because there is no duplicated value in these two variables and they are more like a identifier for each row of dataset. For each numerical variable, we can see the summary in the table below:

plots will be added after revision

For each categorical variable (state and region), we can see the summary in the tables below below:

plots will be added after revision

Then we can make histogram of each numeric variables to see the distribution:

plots will be added after revision

We can see that land.area, pop, doctors, hosp.beds, crimes, pct.bach.deg, pct.below.pov, pct.unemp, and tot.income are skewed to right, and pct.hs.grad is skewed to left. Then, we can do a correlation plot to check the collinearity.

plots will be added after revision

From the plot we can see that tot.income and pop are highly correlated, and they both are reasonably highly correlated with crimes, hosp.beds and doctors, and these three are also strongly correlated with each other. per.cap.income isn't really highly correlated with anything, but it has some positive correlation with pct.hs.grad, pct.bach.deg and some negative correlation with pct.below.pov, pct.unemp. And pct.hs.grad, pct.bach.deg, pct.below.pov, and pct.unemp. are moderately correlated with each other.

These correlations are expected. total income = per-capita income * population, so it is reasonable that they are correlated, and huge population in some way means that more people are criming, and also more people being doctors, and this county will need more beds in the hospital for this large amount of people. At the same time, more crimes means that more people are hurt and thus more doctors and hospital beds needed. Also, per-capita income has positive correlation with pct.hs.grad, pct.bach.deg because people with high school or college education are more likely to get high income than those who do not complete at least 12 years of school. And per-capita income has negative correlation with pct.below.pov, and pct.unemp because the increase of people with income below poverty level and people unemployed means that they are getting really low income.

Method

Fristly, we want to address the question that if we ignore all other variables, whether per-capita income should be related to crime number, and also that this relationship may be different in different regions of the country, which statistically means that if we will need to add interaction between region and crime number. At the beginning, we can apply log transformation on the heavily right-skewed variables: land.area, pop, doctors, hosp.beds, crimes, pct.bach.deg, pct.below.pov, pct.unemp, and tot.income; and we apply power transformation on the left skewed variable pct.hs.grad. Tranformation is applied to put the tails in and make the distribution of these variables more normal. Then, we will fit the first model: linear model $log(per - capitaincome)\ log(crimenumber)$ to see if crime number is a significant variable for per-capita income. For the second model, region variable is added: $log(per-capitaincome)\ log(crimenumber)+region$. For the third model, an internaction term is added: $log(per - capitaincome)\ log(crimenumber) + region + log(crimenumber) * region$. Then we can apply an anova test to see whch one fits the data best to decide if we need the region term or the interaction term. After noticing that per-capita income is total income/total population, we want to know if per-capita crime works better than total crime number in terms of using per-capita variable to predict per-capita variable. But before we compare crimme number with per-capita crime, we will need to figure out if region is still important with it is with per-capita crime. Thus, the model $log(per - capitaincome)\ log(per - capitacrime)$ is fitted to see the significance of per-capita crime. Then we seperately add region and interaction term to this model as above. Then anova test is applied to these three models to choose among $log(per - capitaincome)\ log(per - capitacrime)$, $log(per - capitaincome)\ log(per - capitacrime)+region$, and $log(per-capitaincome)\ log(per-capitacrime)+region+region*log(per-capita)$. The two best models selected from two sets of models are compared by diagnostic plots, AIC, BIC, and summary to choose the final one, which will be used in the later process.

For the further selection of variables and models, total income and population are removed because per-capita income = total income/total population. Then, we fit $per - capitaincome\ state + log(land.area) + log(doctors) + log(hosp.beds) + log(crimes) + log(pct.bach.deg) + log(pct.below.pov) + log(pct.unemp) + (pct.hs.grad)^2 + pop.18_34 + pop.65_plus + region$, which are the other 10 variables left. For marginal model plots for the variables in this model, we can see if the expected value from model for one specific variable can match the expected value from nonparametric regression procedure, which shows if there is unusual pattern on data that higher degree terms or more interaction terms are needed. If the data and model fit well on the plot, then we can do variable selection on the variables we already have. For variable selection, all subsets method, which maximizes r squared value and minimizes cp and BIC valeus, and stepwise selection are used. Then the two models are compared by diagnostic plots, AIC, BIC, and how variables can be interpreted in the aspect of social science and economy.

Results

For the model comparison to check the significance of crime number on per-capita income, and the exploration of how region can affect the relationship between per-capita income and crime number in part 2) of Code Appendix, by the anova test on three models $log(per - capitaincome)\ log(crimenumber)$, the model with summative region variable: $log(per - capitaincome)\ log(crimenumber) + region.$ and the model with the interaction term: $log(per - capitaincome)\ log(crimenumber) + region + log(crimenumber) * region$, we find that the model that per-capita income with only the crime number is not enough when comparing to model that per-capita income with both crime number and region. Also, as the p-value for model $log(per-capitaincome)\ log(crimenumber)+region+log(crimenumber)*region$ is about 0.39, which means that we can accept the hypothesis that the model with region but without interaction term is enough for per-capita income. Thus, $log(per-capitaincome)\ log(crimenumber)+region$ is selected, which means that dummy variable region produces only additive changes in log(Y). Then, when per-capita crime is considered in the model in part 3) of the Code Appendix, we get the same result that $log(per-capitaincome)\ log(per-capitacrime) + region$ is enough and interaction between per-capita crime and region is also unnecessary. Then we come to the choice between total crime number and per-capita crime in the part 4) of Code Appendix. To decide which one is better between fit 1: $log(per-capitaincome)\ log(crimenumber)+region$ and fit 3: $log(per - capitaincome)\ log(per - capitacrime) + region$, based on the summary, we can see

that the r squared value is larger for model 1 than model 3, and also the log(per.capita.crime) is not a significant variable in model 3. For the diagnostic plots, residuals vs.fitted plot for model 1 looks good as there is no pattern and the mean is at about 0; there are some points off the line at the two sides; there is a slightly upward pattern in the scale-location plot; there is no point with both high residuals and high leverage. For model 3, the plots are quite similar with those of model 1, except that points on residuals and scale-location plots are roughly clustered into three groups, so that maybe regions affect the model more for model 3. Also, when comparing their AIC values and BIC values, we can see that fit 1 has both lower AIC value and BIC value, which makes it better than fit 3. Overall, I will choose model 1 $log(per-capitaincome)$ $log(crimenumber)+region$ for its lower AIC and BIC values and higher r squared value. Thus, I think use total crime number is better than per-capita crime to predict per-capita income.

plots will be added after revision

Then, we are going to do variable selection, which corresponds to part 5) of Code Appendix, from variables in model log(per-capita income)~state+log(land.area)+log(doctors)+log(hosp.beds)+log(crimes)+log(pct.bach.deg)+log(pct.be + pop.18_34 + pop.65_plus + region. By marginal model plots in part 5) (a) of code, we can see that the expected value from model for one specific variable can match the expected value from nonparametric regression procedure, which means that there is no unusual pattern on data that higher degree terms or more interaction terms are needed. Thus, we will continue using the variables in this model and do variable selection on these variables we already have.

mmps plots will be added after revision

Stepwise selection in both directions is used in part 5) (b) of code, which iteratively adding and removing predictors in the predictive model in order to find the subset of variables in the data set resulting in the model that has lowest prediction error. And at this step we get the model step_model: $per.cap.income\ state + log(land.area) + log(doctors) + log(pct.bach.deg) + log(pct.below.pov) + pct.hs.grad + pop.18_34 + pop.65_plus$. All subsets method is used in part 5) (c), which tests all possible subsets of the set of potential independent variables. For the model that maximizes squared value and minimizes cp and BIC values, we get model with 8 variables for all three requirements, fit_8: $per.cap.income\ stateCA + stateNJ + stateUT + log(land.area) + log(doctors) + log(pct.bach.deg) + log(pct.below.pov) + pop.18_34$.

plots will be added after revision

In part 5) (d) of the code, the two models selected are compared in several ways.

diagnostic plots will be added after revision

From the two sets of diagnostic plots, we can find it is hard to choose based on them. Both resduals vs.fitted plots and scale-location plots has a slight upward concavity patterns shown. And from both normal q-q plots, we can see that they all have some points off the line at the very right side, which also corresponds to the residuals vs.leverage plots that they both have points with either high residuals or high leverage, but no point with both high residuals and high leverage. They have quite similar r squared values and it is also hard to choose based on this single value. However, fit_8 has smaller AIC and BIC values.

aic bic value plot will be added after revision

mmps for fit_8 plots will be added after revision

Also, I will choose model fit_8 as a better one as it has far less variables (there are many variables "hidden" under the state in step_model), and each state can result its own additive change on Y, which is somehow overfitting from my perspective, and also may cover some impact from region, which is a more generalized area factor. The marginal model plots are also shown above, which gives that no more higher degree

terms and interaction terms are needed for this set of variables. In this model, apart from those numeric variables that may contribute to the prediction of per.capita.income, three states are also extracted as special representatives, which may have especially high or low per capita income in this dataset (which is not big thus may be biased in some way) that can affect the model. Not only how many infrastructures/features this state has can be related to its per capita income, sometimes the state itself also means something, and I think this really related to social science and economics of a state other than pure statistical inference. For the numeric variables, the positive corrlation between y variable and doctors, pct.bach.deg, and the negative correlation between y variable and pct.below.pov can both fit the correlation plot and our expectation at the very beginning, which makes this model reasonable in both statiscal and social aspects.

Discussion

For the correlation between the variables, from the correlation plots we can see that tot.income and pop are highly correlated, and they both are reasonably highly correlated with crimes, hosp.beds and doctors, and these three are also strongly correlated with each other.

per.cap.income isn't really highly correlated with anything, but it has some positive correlation with pct.hs.grad, pct.bach.deg and some negative correlation with pct.below.pov, pct.unemp. And pct.hs.grad, pct.bach.deg, pct.below.pov, and pct.unemp. are moderately correlated with each other. These correlations are expected. total income = per-capita income * population, so it is reasonable that they are correlated, and huge population in some way means that more people are criming, and also more people being doctors, and this county will need more beds in the hospital for this large amount of people. At the same time, more crimes means that more people are hurt and thus more doctors and hospital beds needed. Also, per-capita income has positive correlation with pct.hs.grad, pct.bach.deg because people with high school or college education are more likely to get high income than those who do not complete at least 12 years of school. And per-capita income has negative correlation with pct.below.pov, and pct.unemp because the increase of people with income below poverty level and people unemployed means that they are getting really low income.

From the two anova test, we can know that total crime number and region are significant variables for per-capita income, which means that per-capita income has relationship with both of them. Also, as the interaction term is not significant, each region results in different additive change on per-capita income. And I think using total crime numbers instead of per-capita crime is better for predicting the per-capita income after comparing the diagnostic plots, summaries, AIC, and BIC values of two models. And the final model I select from all subsets and stepwise model selection methods is $per.cap.income\ stateCA + stateNJ + stateUT + log(land.area) + log(doctors) + log(pct.bach.deg) + log(pct.below.pov) + pop.18_34$, which has smaller AIC and BIC values, and also includes states that can be considered as representations somehow in the social science perspective, and also has the numeric variables with either positive or negative correlations that can match the expectated correlations at the very beginning.

Strengths: Estimated coefficients have the expected sign. The model is confirmed by stepwise and All subsets procedures. Variables are either in their original scale, or are transformed by logarithm or power, and final model is concise, which makes explaining the models to people good at social science & economics but not statistics easier. Weakness: The residual diagnostic plots are just OK. Do not have time on exploring how the interaction term between region or state, and other variables can result in different slopes for variables predicting the per-capita income. (might be addressed in the final draft)

There is also an important problem is that there are missing states and missing counties in the datasets (48/51 states and 373/3000 counties appear in the dataset). We should be care about it because there are too many missing state such that there might be many other variables that are significant in these states and might be considered as not valued in this project. Also, the missing data can result in some missing information in the state and region variable, which might be biased under what we have now.

Thus, in the future research on this same topic, how region/state variable can interactively change per-capita income with other variables should be concerned, and some missing data should be collected and added on if possible. ## some highlights on the key points will be added

Reference

Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. (2005) Applied Linear Statistical Models. 5th Edition, McGraw-Hill, Irwin, New York.

Code Appendix

1) Data summary and EDA

```
cdi <- read.table("cdi.dat")
#View(cdi)
```

```
#colnames(cdi)
```

```
attach(cdi)
table <- matrix(c(summary(land.area),
summary(pop),
summary(pop.18_34),
summary(pop.65_plus),
summary(doctors),
summary(hosp.beds),
summary(crimes),
summary(pct.hs.grad),
summary(pct.bach.deg),
summary(pct.below.pov),
summary(pct.unemp),
summary(per.cap.income),
summary(tot.income)), ncol = 6, byrow = TRUE)
detach(cdi)
```

```
rownames(table)<-c("land.area", "pop", "pop.18_34", "pop.65_plus", "doctors", "hosp.beds", "crimes", "p

colnames(table) <- c("Min.", "1st Qu.",  "Median","Mean", "3rd Qu.", "Max.")
table
```

```
##                    Min.    1st Qu.    Median         Mean    3rd Qu.      Max.
## land.area          15.0    451.250    656.50 1.041411e+03    946.750   20062.0
## pop            100043.0 139027.250 217280.50 3.930109e+05 436064.500 8863164.0
## pop.18_34          16.4     26.200     28.10 2.856841e+01     30.025      49.7
## pop.65_plus         3.0      9.875     11.75 1.216977e+01     13.625      33.8
## doctors            39.0    182.750    401.00 9.879977e+02   1036.000   23677.0
## hosp.beds          92.0    390.750    755.00 1.458627e+03   1575.750   27700.0
## crimes            563.0   6219.500  11820.50 2.711162e+04  26279.500  688936.0
## pct.hs.grad        46.6     73.875     77.70 7.756068e+01     82.400      92.9
## pct.bach.deg        8.1     15.275     19.70 2.108114e+01     25.325      52.3
## pct.below.pov       1.4      5.300      7.90 8.720682e+00     10.900      36.3
## pct.unemp           2.2      5.100      6.20 6.596591e+00      7.500      21.3
## per.cap.income   8899.0  16118.250  17759.00 1.856148e+04  20270.000   37541.0
## tot.income       1141.0   2311.000   3857.00 7.869273e+03   8654.250  184230.0
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
table_1 <- data.frame(summary(as.factor(cdi$state)))
table_1 <- table_1 %>% rename(frequency=summary.as.factor.cdi.state..)
table_1
```

```
##    frequency
## AL         7
## AR         2
## AZ         5
## CA        34
## CO         9
## CT         8
## DC         1
## DE         2
## FL        29
## GA         9
## HI         3
## ID         1
## IL        17
## IN        14
## KS         4
## KY         3
## LA         9
## MA        11
## MD        10
## ME         5
## MI        18
## MN         7
## MO         8
## MS         3
## MT         1
## NC        18
## ND         1
## NE         3
## NH         4
## NJ        18
## NM         2
## NV         2
## NY        22
## OH        24
```

```
## OK          4
## OR          6
## PA         29
## RI          3
## SC         11
## SD          1
## TN          8
## TX         28
## UT          4
## VA          9
## VT          1
## WA         10
## WI         11
## WV          1
```

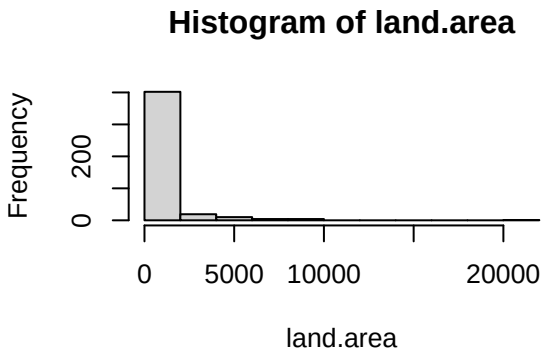
```
table_2 <- data.frame(summary(as.factor(cdi$region)))
table_2 <- table_2 %>% rename(frequency=summary.as.factor.cdi.region..)
table_2
```

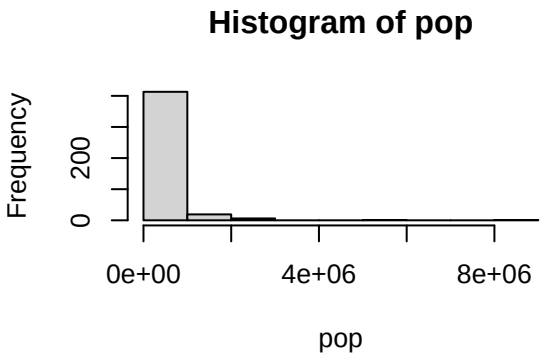
```
##    frequency
## NC       108
## NE       103
## S        152
## W         77
```

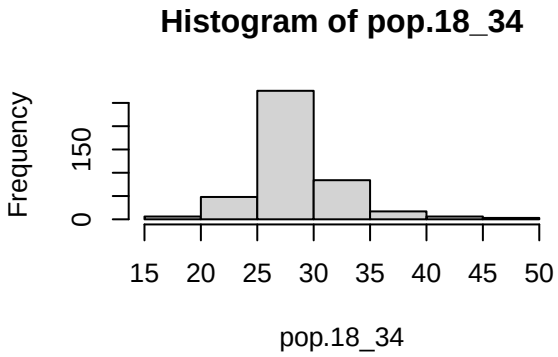
```
#is.na(cdi)
```

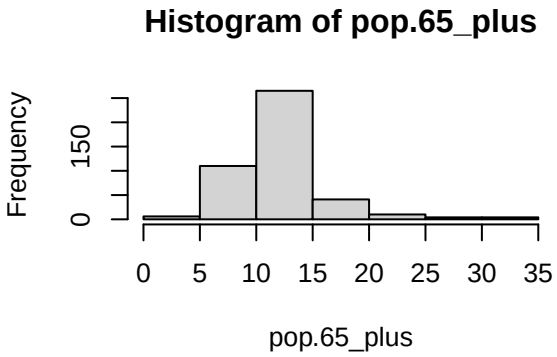
There is no missing data (NA's).

```
attach(cdi)
par(mfrow=c(2,2))
hist(land.area)
hist(pop)
hist(pop.18_34)
hist(pop.65_plus)
```

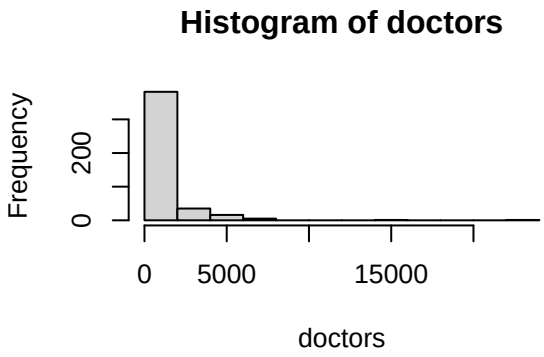


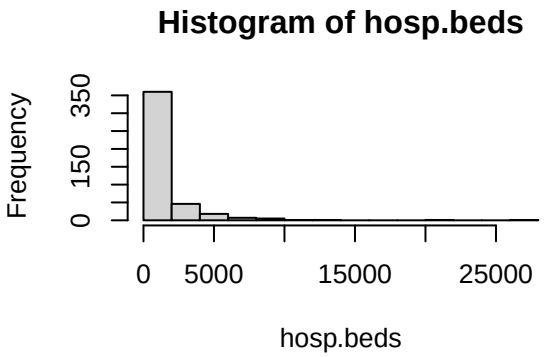


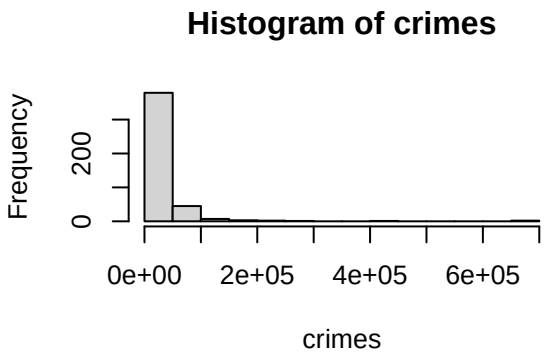


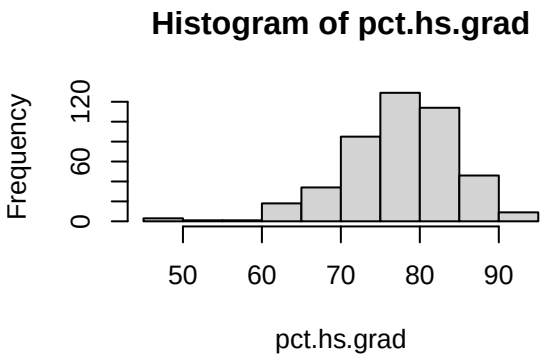



```
hist(doctors)
hist(hosp.beds)
hist(crimes)
hist(pct.hs.grad)
```

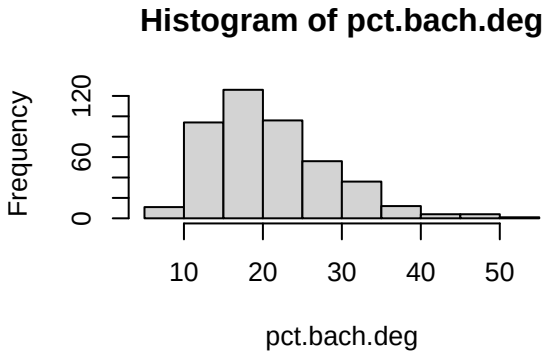


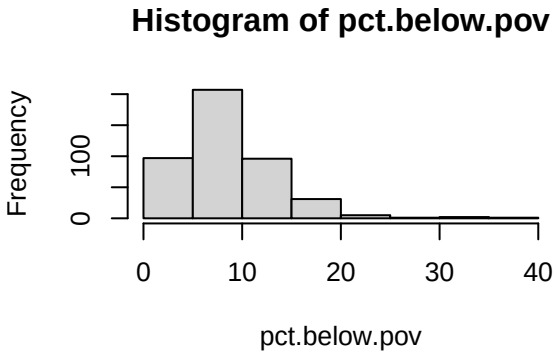


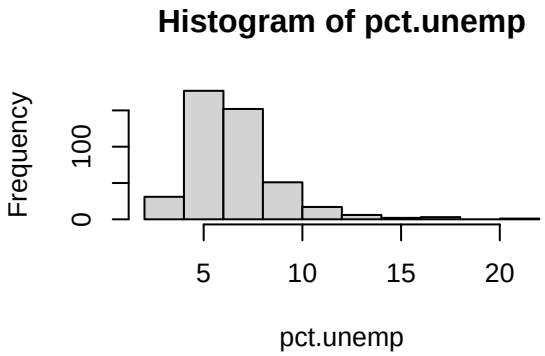


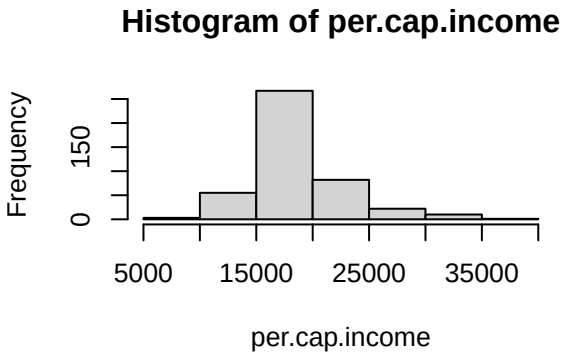



```
hist(pct.bach.deg)
hist(pct.below.pov)
hist(pct.unemp)
hist(per.cap.income)
```

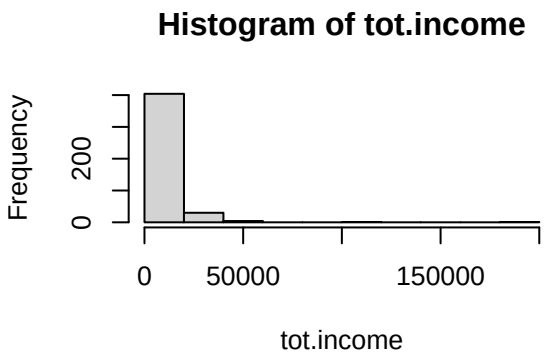









```
hist(tot.income)
detach(cdi)
```

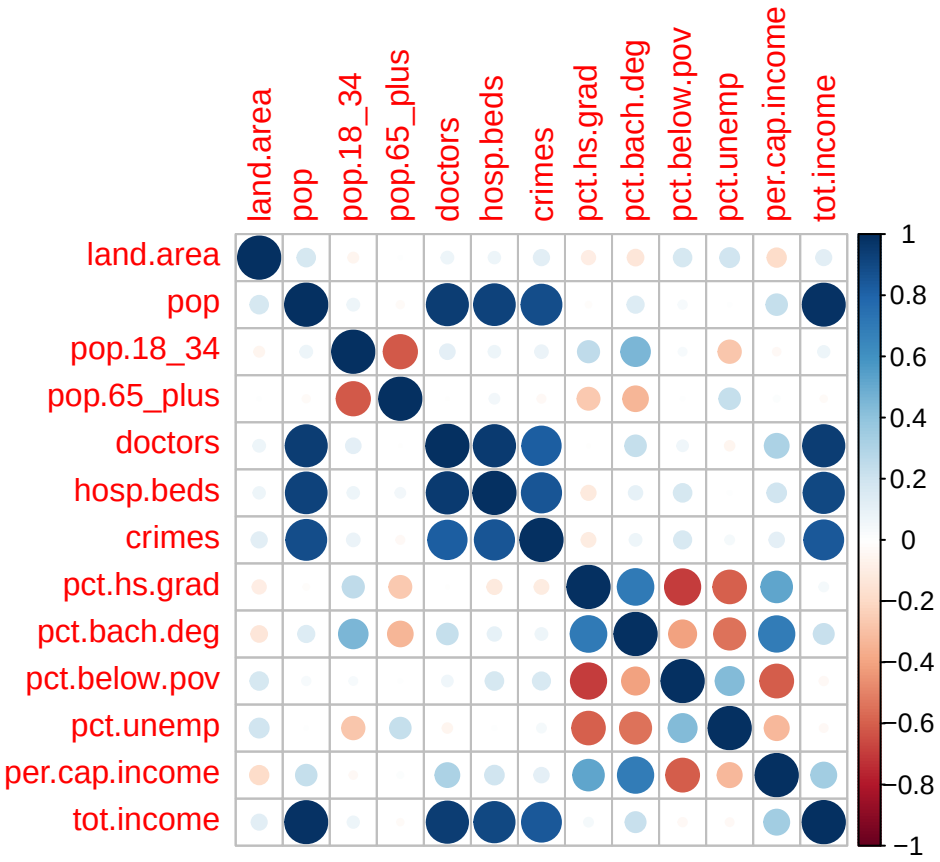


We can see that land.area, pop, doctors, hosp.beds, crimes, pct.bach.deg, pct.below.pov, pct.unemp, and tot.income are highly skewed to right, thus we may need to apply log transformation to them to put the tails in. Also, pct.hs.grad is somehow skewed to the left, so we apply a power transformation of degree 2 to it.

```
library("corrplot")
```

```
## corrplot 0.90 loaded
```

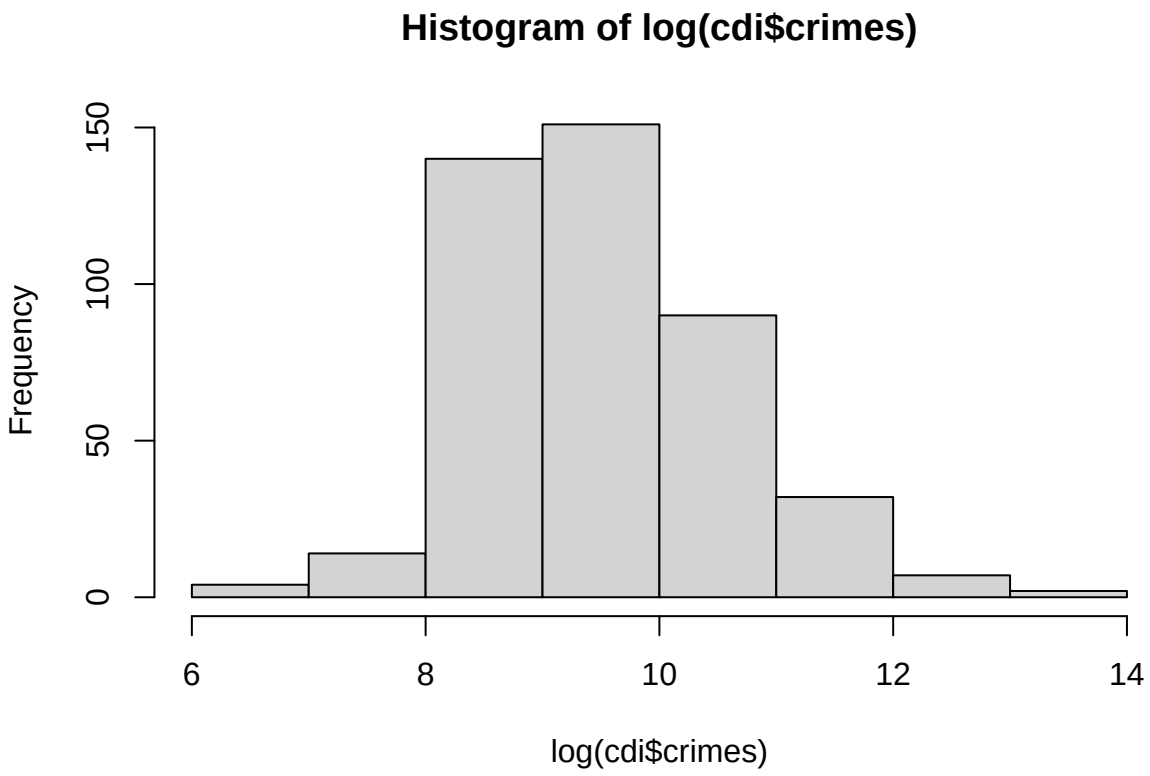
```
corr <- cor(cdi[4:16])
corrplot(corr, method = "circle")
```



From the correlation plot above, we can see that the dots are dark means that the corresponding two variables are highly correlated, so when the highly correlated variables occur in the same model, we will need to consider if there is any confounding variables or missing variables that may mislead the model.

2) How crimes and region are related to per-capita crime

```
hist(log(cdi$crimes))
```



Log transformation is applied to crimes.

```
fit <- lm(per.cap.income~log(crimes), data = cdi)
summary(fit)
```

```
##
## Call:
## lm(formula = per.cap.income ~ log(crimes), data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10358.7  -2292.5   -867.7   1489.4  19330.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8972.8     1651.1   5.435 9.14e-08 ***
## log(crimes)   1009.0      172.6   5.845 9.90e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3914 on 438 degrees of freedom
## Multiple R-squared:  0.07236,    Adjusted R-squared:  0.07024
## F-statistic: 34.16 on 1 and 438 DF,  p-value: 9.901e-09
```

```
fit <- lm(per.cap.income~log(crimes), data = cdi)
fit1 <- lm(per.cap.income~log(crimes)+region, data = cdi)
```

```
summary(fit)
```

```
##
## Call:
## lm(formula = per.cap.income ~ log(crimes), data = cdi)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10358.7  -2292.5   -867.7   1489.4  19330.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8972.8     1651.1   5.435 9.14e-08 ***
## log(crimes)   1009.0      172.6   5.845 9.90e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3914 on 438 degrees of freedom
## Multiple R-squared:  0.07236,    Adjusted R-squared:  0.07024
## F-statistic: 34.16 on 1 and 438 DF,  p-value: 9.901e-09
```

```
summary(fit1)
```

```
##
## Call:
## lm(formula = per.cap.income ~ log(crimes) + region, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9229.2 -2183.6  -502.4  1339.3 20110.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6870.8     1582.5   4.342 1.76e-05 ***
## log(crimes)   1237.3      167.0   7.411 6.61e-13 ***
## regionNE      2284.9      506.2   4.514 8.21e-06 ***
## regionS      -1354.4      468.3  -2.892  0.00402 **
## regionW       -768.2      558.5  -1.376  0.16968
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3676 on 435 degrees of freedom
## Multiple R-squared:  0.1875, Adjusted R-squared:  0.1801
## F-statistic:  25.1 on 4 and 435 DF,  p-value: < 2.2e-16
```

For this original model, we cansee that log(crimes) is a significant variable for per.capita.income, and each percent increase of crime leads to about 12 increase of per.capita.income. Then we check if the model becomes better with interaction term in the model.

```
fit2<-lm(per.cap.income~log(crimes)+region+(log(crimes):region), data = cdi)
summary(fit2)
```

```
##
## Call:
## lm(formula = per.cap.income ~ log(crimes) + region + (log(crimes):region),
##     data = cdi)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10810  -2127   -533   1187  20202
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9634.8     2888.0   3.336 0.000923 ***
## log(crimes)              938.1      310.3   3.024 0.002648 **
## regionNE               -4544.8     4262.1  -1.066 0.286870
## regionS                -2595.4     4201.9  -0.618 0.537117
## regionW                -4784.6     4846.6  -0.987 0.324093
## log(crimes):regionNE     738.8      457.8   1.614 0.107313
## log(crimes):regionS      141.8      441.4   0.321 0.748223
## log(crimes):regionW      426.0      499.8   0.852 0.394467
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3676 on 432 degrees of freedom
## Multiple R-squared:  0.1931, Adjusted R-squared:  0.1801
## F-statistic: 14.77 on 7 and 432 DF,  p-value: < 2.2e-16
```

It seems that r squared value does not change a lot with interaction term. Then we use anova test to check H0: fit1(no interaction term) is enough or Ha: reject H0 so that we will need an interaction term.

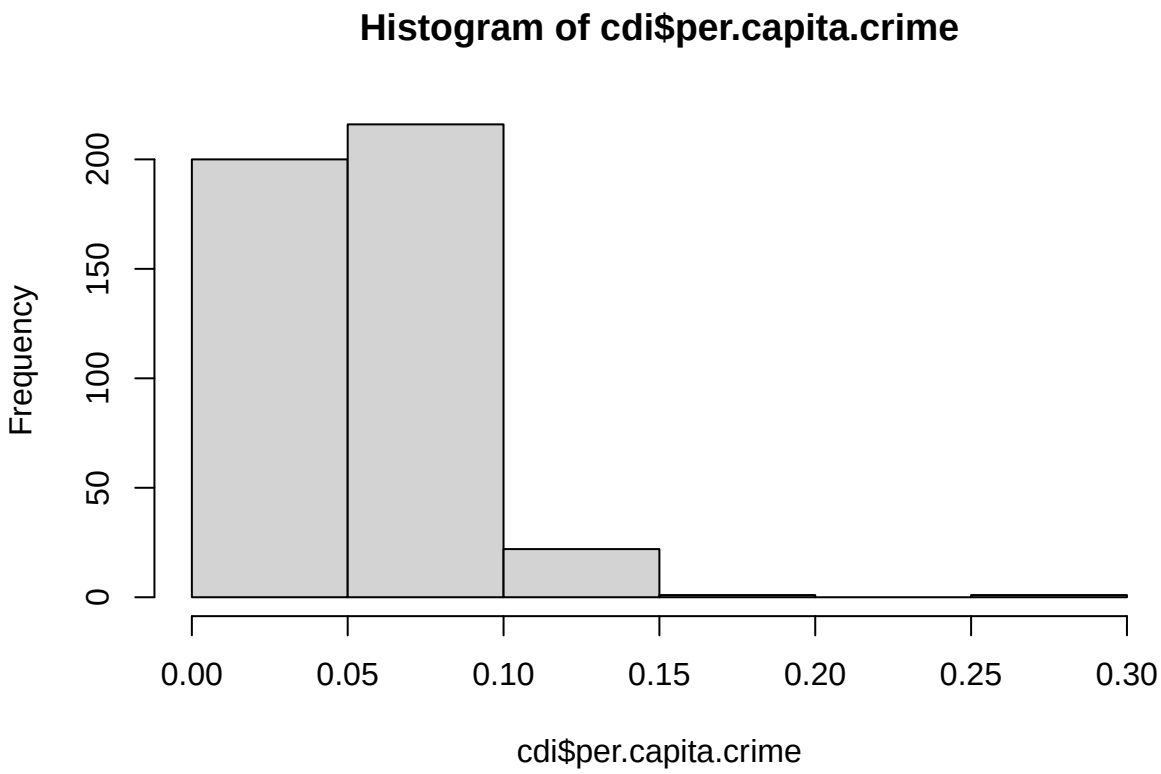
```
anova(fit, fit1, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ log(crimes)
## Model 2: per.cap.income ~ log(crimes) + region
## Model 3: per.cap.income ~ log(crimes) + region + (log(crimes):region)
##   Res.Df        RSS Df Sum of Sq       F    Pr(>F)
## 1    438 6710024435
## 2    435 5876801559  3 833222876 20.5579 1.807e-12 ***
## 3    432 5836388967  3  40412592  0.9971     0.394
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3) How per-capita crime and region are related to per-capita crime

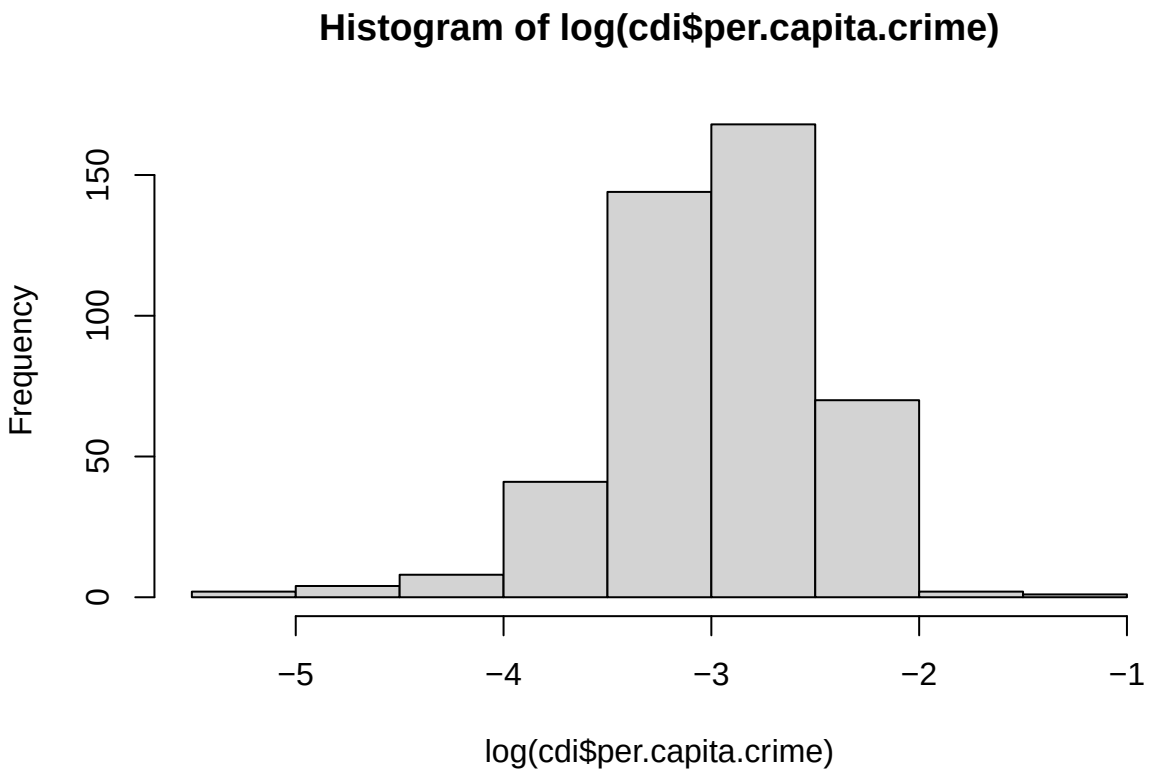
```
cdi["per.capita.crime"] = cdi$crimes/cdi$pop
```

```
hist(cdi$per.capita.crime)
```



We can see that the per.capita.crime is highle right skewed, so we may apply a log transformation on it, and it looks better now.

```
hist(log(cdi$per.capita.crime))
```




```
fit <- lm(per.cap.income~log(per.capita.crime), data = cdi)
fit3 <- lm(per.cap.income~log(per.capita.crime)+region, data = cdi)
summary(fit3)
```

```
##
## Call:
## lm(formula = per.cap.income ~ log(per.capita.crime) + region,
##     data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8725.5 -2270.1  -639.8  1768.3 19455.1
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            20349.7     1366.2  14.895  < 2e-16 ***
## log(per.capita.crime)    659.9      423.2   1.559   0.1197
## regionNE                2444.2      543.9   4.494 8.98e-06 ***
## regionS                -1073.8      517.1  -2.077   0.0384 *
## regionW                 -158.0      591.5  -0.267   0.7895
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3890 on 435 degrees of freedom
## Multiple R-squared:  0.09007,    Adjusted R-squared:  0.0817
## F-statistic: 10.76 on 4 and 435 DF,  p-value: 2.501e-08
```

```
fit4 <- lm(per.cap.income~log(per.capita.crime)+region+(log(per.capita.crime):region), data = cdi)
summary(fit4)
```

```
##
## Call:
## lm(formula = per.cap.income ~ log(per.capita.crime) + region +
##     (log(per.capita.crime):region), data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8600.4 -2312.3  -653.3  1735.2 19486.5
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     19913.6     2069.5   9.622   <2e-16 ***
## log(per.capita.crime)             519.4      655.5   0.792    0.429
## regionNE                         4585.8     3382.0   1.356    0.176
## regionS                         -1705.7     3166.7  -0.539    0.590
## regionW                           525.7     5271.3   0.100    0.921
## log(per.capita.crime):regionNE    653.1     1030.9   0.634    0.527
## log(per.capita.crime):regionS    -253.5     1094.4  -0.232    0.817
## log(per.capita.crime):regionW     227.9     1826.0   0.125    0.901
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3900 on 432 degrees of freedom
## Multiple R-squared:  0.09147,    Adjusted R-squared:  0.07675
## F-statistic: 6.213 on 7 and 432 DF,  p-value: 6.001e-07
```

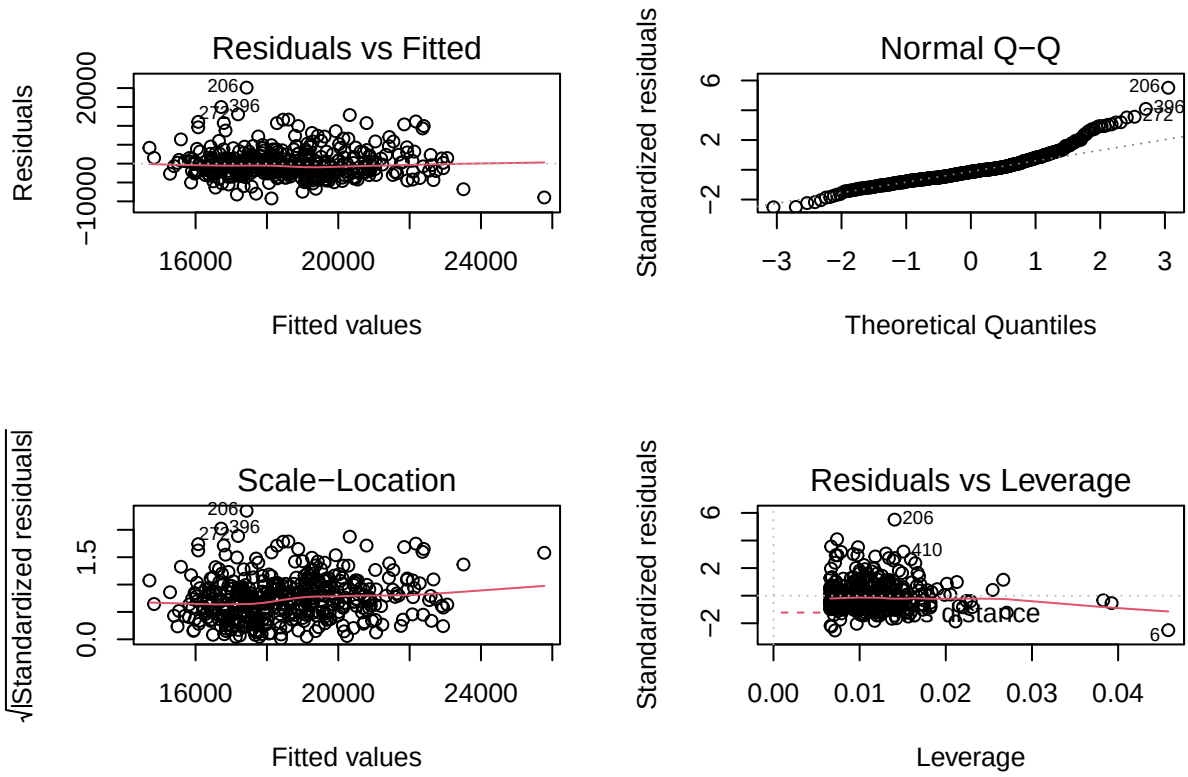
```
anova(fit, fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ log(per.capita.crime)
## Model 2: per.cap.income ~ log(per.capita.crime) + region
## Model 3: per.cap.income ~ log(per.capita.crime) + region + (log(per.capita.crime):region)
##   Res.Df        RSS Df Sum of Sq       F    Pr(>F)
## 1    438 7200895643
## 2    435 6581927659  3 618967984 13.5627 1.797e-08 ***
## 3    432 6571800580  3  10127079  0.2219    0.8812
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

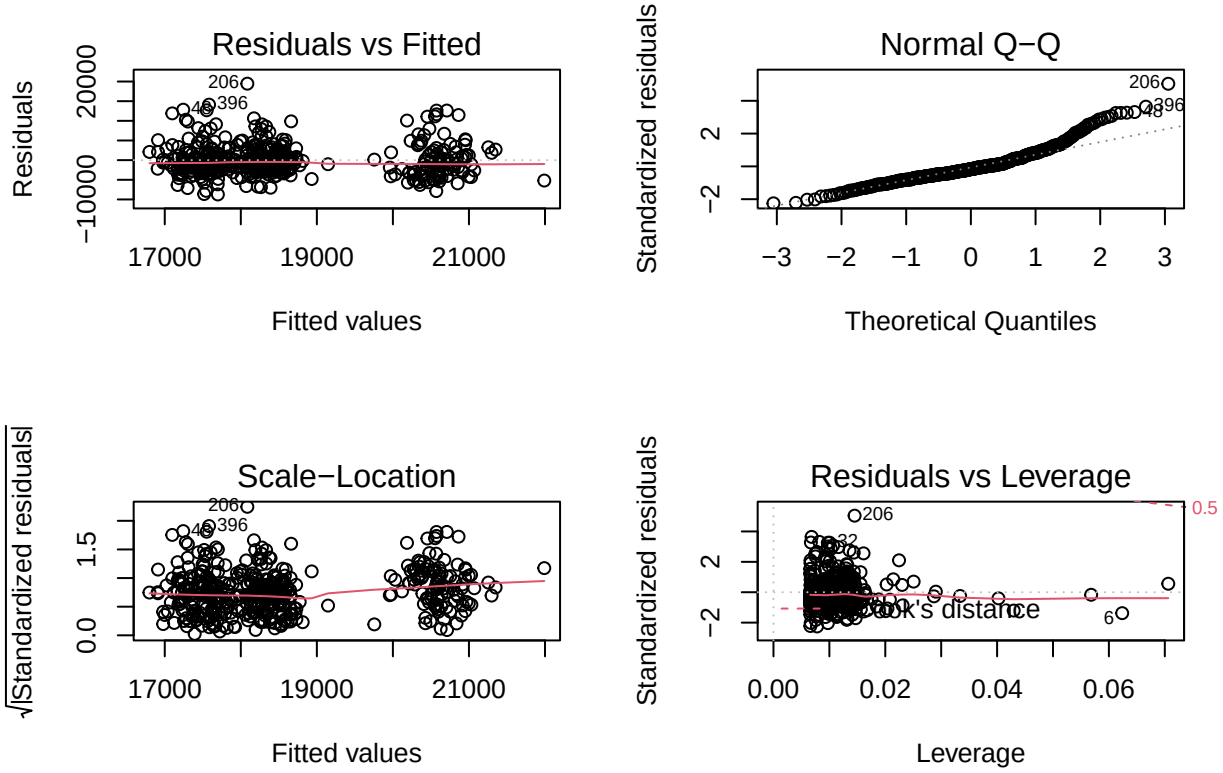
We can see that for model per.cap.income ~ log(per.capita.crime) + region, we also do not need to add the interaction term as the p-value is $0.8812 > 0.05$.

4) Choose between total crime number and per-capita crime

```
par(mfrow = c(2,2))
plot(fit1)
```




```
par(mfrow = c(2,2))
plot(fit3)
```




```
AIC(fit1, fit3)
```

```
##      df      AIC
## fit1  6 8479.968
## fit3  6 8529.826
```

```
BIC(fit1, fit3)
```

```
##      df      BIC
## fit1  6 8504.488
## fit3  6 8554.347
```

To decide which one of fit 1 and fit 3, apart from the summary of two models, we will also need to check the diagnostic plots for them. Based on the summary, we can see that the r squared value is larger for model 1, per.cap.income ~ log(crimes) + region, than model 3, per.cap.income ~ log(per.capita.crime) + region, and also the log(per.capita.crime) is not a significant variable in model 3. For the diagnostic plots, residuals vs.fitted plot for model 1 looks good as there is no pattern and the mean is at about 0; there are some points off the line at the sides; there is a slightly upward pattern in the scale-location plot; there is no points with both high residuals and high leverage. For model 3, the plots are quite similar with those of model 1, except that points on residuals and scale-location plots are roughly clustered into three groups, so that maybe regions affect the model more for this one. Also, when comparing their AIC values and BIC values, we can see that fit 1 has both lower AIC value and BIC value, which makes it better than fit 3. Overall, I will choose model 1 for its lower AIC and BIC values and higher r squared value.

5) Variable selection process

(a) Check if terms other than linear terms are needed

We can see that land.area, pop, doctors, hosp.beds, crimes, pct.bach.deg, pct.below.pov, pct.unemp, and tot.income are highly skewed to right, thus we may need to apply log transformation to them to put the tails in. Also, pct.hs.grad is somehow skewed to the left, so we apply a power transformation of degree 2 to it. Also, as per capita income = total income/population, and we also can see the high correlation among them in the correlation plot. Thus, I decide to remove these two features.

```
fit_all <- lm(per.cap.income ~ state+log(land.area)+log(doctors)+log(hosp.beds)+log(crimes)+log(pct.bach
```

```
summary(fit_all)
```

```
##
## Call:
## lm(formula = per.cap.income ~ state + log(land.area) + log(doctors) +
##     log(hosp.beds) + log(crimes) + log(pct.bach.deg) + log(pct.below.pov) +
##     log(pct.unemp) + (pct.hs.grad)^2 + pop.18_34 + pop.65_plus +
##     region, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3908.7  -844.0   -27.7   623.6  8490.5
##
## Coefficients: (3 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)        16369.8798  2998.2197   5.460 8.60e-08 ***
## stateAR             -816.2239  1274.6542  -0.640 0.522329
## stateAZ             -415.5665  1000.2078  -0.415 0.678024
## stateCA             1840.1749   720.6152   2.554 0.011049 *
## stateCO              336.2731   839.0895   0.401 0.688821
## stateCT             1331.7400   899.9490   1.480 0.139752
## stateDC             1212.0393  1754.8243   0.691 0.490180
## stateDE              297.5976  1296.0024   0.230 0.818504
## stateFL             -724.4511   740.5308  -0.978 0.328552
## stateGA              509.2918   824.3549   0.618 0.537071
## stateHI              818.1313  1154.1649   0.709 0.478849
## stateID             -426.7534  1713.9600  -0.249 0.803505
## stateIL              891.6149   743.8954   1.199 0.231436
## stateIN             -174.0558   763.1091  -0.228 0.819700
## stateKS              100.5636  1024.8091   0.098 0.921881
## stateKY             -438.2043  1111.6388  -0.394 0.693656
## stateLA              -93.6529   812.7084  -0.115 0.908319
## stateMA              269.8435   890.5957   0.303 0.762061
## stateMD              170.6599   852.7861   0.200 0.841493
## stateME             -114.7022   961.3243  -0.119 0.905087
## stateMI             1254.5347   767.5692   1.634 0.102993
## stateMN             -403.6193   881.8037  -0.458 0.647414
## stateMO              -56.6052   845.7622  -0.067 0.946674
## stateMS             -926.6000  1106.3883  -0.837 0.402835
## stateMT              154.8184  1719.4422   0.090 0.928303
## stateNC              165.4056   735.6449   0.225 0.822221
## stateND             -442.6620  1742.5573  -0.254 0.799609
```

```
## stateNE            -1148.9577  1161.5976  -0.989 0.323231
## stateNH             -322.4703  1058.1262  -0.305 0.760718
## stateNJ             2081.4210   774.0865   2.689 0.007483 **
## stateNM            -1433.0215  1300.1348  -1.102 0.271064
## stateNV             4043.1655  1326.2465   3.049 0.002459 **
## stateNY              337.5600   724.7554   0.466 0.641655
## stateOH              218.5772   717.3272   0.305 0.760753
## stateOK             -874.0729  1013.1787  -0.863 0.388842
## stateOR            -1221.2424   931.5207  -1.311 0.190638
## statePA             -528.4056   724.8216  -0.729 0.466439
## stateRI            -2259.2581  1164.8775  -1.939 0.053179 .
## stateSC             -126.2580   779.0764  -0.162 0.871343
## stateSD                0.7241  1742.5440   0.000 0.999669
## stateTN             -311.4113   827.2506  -0.376 0.706798
## stateTX              109.1065   681.5759   0.160 0.872903
## stateUT            -4147.9832  1038.2338  -3.995 7.76e-05 ***
## stateVA              697.0655   900.3459   0.774 0.439280
## stateVT            -1075.1663  1733.0427  -0.620 0.535370
## stateWA             -346.8057   837.6247  -0.414 0.679081
## stateWI               14.2667   808.3708   0.018 0.985928
## stateWV             -564.0978  1703.0126  -0.331 0.740648
## log(land.area)      -738.8776   129.6795  -5.698 2.43e-08 ***
## log(doctors)        1051.3357   291.8377   3.602 0.000357 ***
## log(hosp.beds)        -4.4032   290.0307  -0.015 0.987895
## log(crimes)         -140.6072   197.3067  -0.713 0.476508
## log(pct.bach.deg)   5876.5738   565.9031  10.384  < 2e-16 ***
## log(pct.below.pov) -3457.7067   307.9783 -11.227  < 2e-16 ***
## log(pct.unemp)       661.9531   515.9128   1.283 0.200245
## pct.hs.grad          -51.2852    25.7906  -1.989 0.047467 *
## pop.18_34           -239.7679    30.4659  -7.870 3.69e-14 ***
## pop.65_plus           64.2804    34.4946   1.863 0.063160 .
## regionNE                  NA         NA      NA       NA
## regionS                   NA         NA      NA       NA
## regionW                   NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1584 on 382 degrees of freedom
## Multiple R-squared:  0.8675, Adjusted R-squared:  0.8478
## F-statistic:  43.9 on 57 and 382 DF,  p-value: < 2.2e-16
```

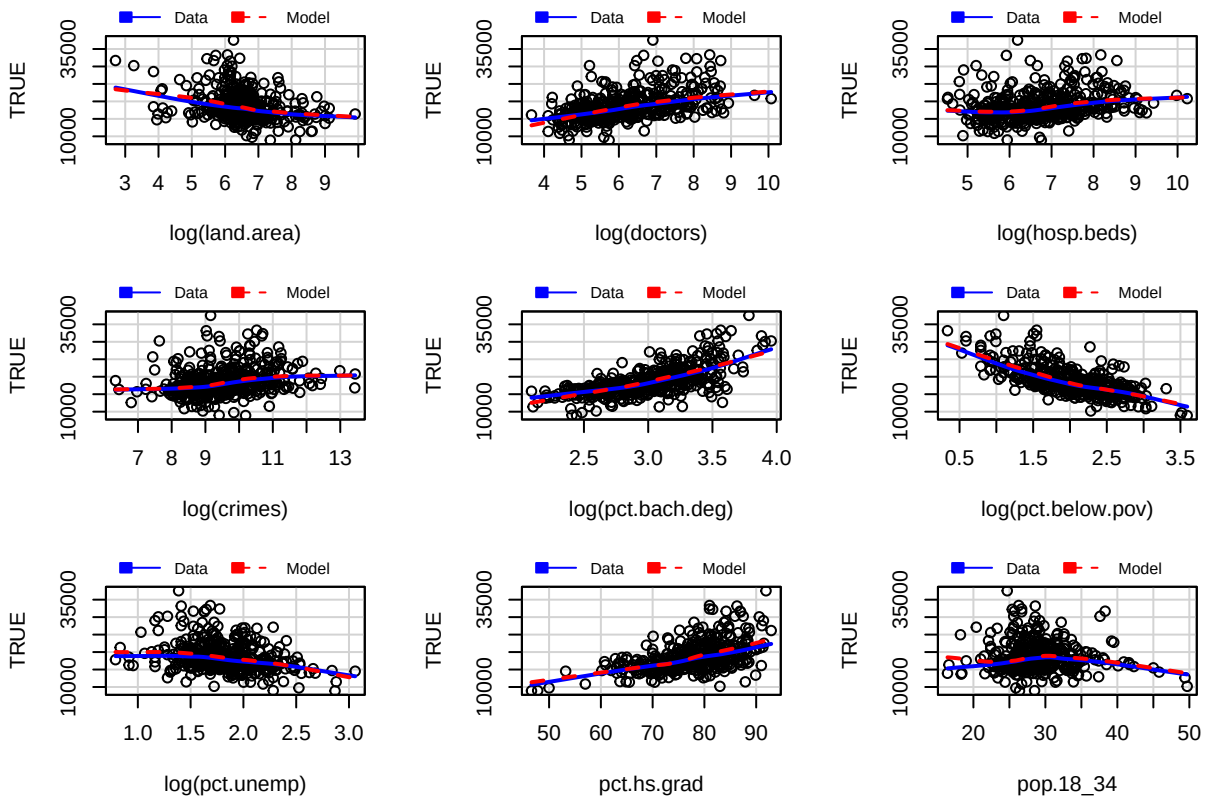
Here I want to explore if there is any interaction needed for this model instead of this simple multiple regression by using marginal model plots to see if the expected value from model for one specific variable can match the expected value from nonparametric regression procedure.

```
library(car)
```

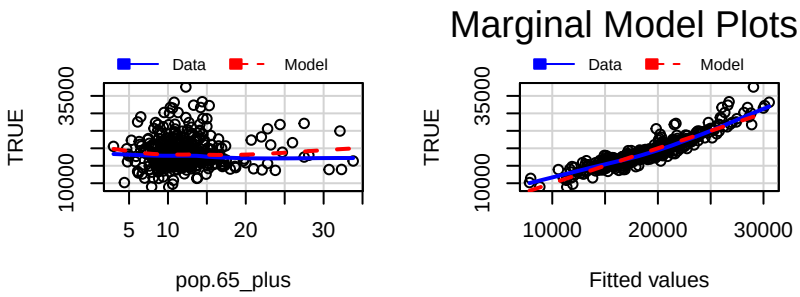
```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
```

```
##
##     recode
```

```
mmps(lm(per.cap.income ~ as.factor(state)+ log(land.area) + log(doctors) + log(hosp.beds) + log(crimes)
```




```
## Warning in mmps(lm(per.cap.income ~ as.factor(state) + log(land.area) + :
## Interactions and/or factors skipped
```



From the plots we can see they might not need to add any interaction term as each of them fits really well. Here we start doing a variable selection by stepwise selection and all subsets selection and make a comparison.

(b) Stepwise selection

```
library(leaps)
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
step_model <- stepAIC(fit_all, direction = "both", trace = FALSE)
summary(step_model)
```

```
##
## Call:
## lm(formula = per.cap.income ~ state + log(land.area) + log(doctors) +
##     log(pct.bach.deg) + log(pct.below.pov) + pct.hs.grad + pop.18_34 +
##     pop.65_plus, data = cdi)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3854.7  -872.9   -20.3   636.1  8458.1
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        17734.85    2230.26   7.952 2.06e-14 ***
## stateAR             -928.79    1269.82  -0.731  0.46496
## stateAZ             -454.08     975.99  -0.465  0.64201
## stateCA             2073.30     674.07   3.076  0.00225 **
## stateCO              292.48     818.69   0.357  0.72110
## stateCT             1683.78     855.29   1.969  0.04971 *
## stateDC             1556.73    1729.83   0.900  0.36872
## stateDE              451.17    1281.52   0.352  0.72499
## stateFL             -641.59     717.36  -0.894  0.37168
## stateGA              381.73     816.18   0.468  0.64026
## stateHI              429.48    1101.81   0.390  0.69690
## stateID             -498.60    1708.29  -0.292  0.77054
## stateIL             1050.29     731.97   1.435  0.15213
## stateIN             -151.20     758.25  -0.199  0.84206
## stateKS               73.58    1018.06   0.072  0.94242
## stateKY             -410.22    1106.85  -0.371  0.71113
## stateLA              -83.33     810.51  -0.103  0.91817
## stateMA              804.89     791.77   1.017  0.30999
## stateMD              304.04     824.65   0.369  0.71256
## stateME              133.28     942.24   0.141  0.88759
## stateMI             1504.34     735.15   2.046  0.04141 *
## stateMN             -335.43     875.20  -0.383  0.70174
## stateMO               18.58     841.62   0.022  0.98240
## stateMS             -847.79    1100.37  -0.770  0.44150
## stateMT              243.36    1712.19   0.142  0.88705
## stateNC               92.85     725.45   0.128  0.89822
## stateND             -675.52    1711.48  -0.395  0.69328
## stateNE            -1531.12    1123.76  -1.362  0.17384
## stateNH               52.72    1019.66   0.052  0.95880
## stateNJ             2300.45     754.41   3.049  0.00245 **
## stateNM            -1410.09    1280.86  -1.101  0.27163
## stateNV             3979.03    1320.67   3.013  0.00276 **
## stateNY              531.70     707.75   0.751  0.45296
## stateOH              314.76     709.77   0.443  0.65768
## stateOK             -835.11    1006.92  -0.829  0.40741
## stateOR            -1192.83     906.66  -1.316  0.18908
## statePA             -333.85     706.30  -0.473  0.63672
## stateRI            -1912.20    1127.58  -1.696  0.09072 .
## stateSC             -232.99     768.61  -0.303  0.76196
## stateSD             -337.53    1706.45  -0.198  0.84331
## stateTN             -343.22     824.47  -0.416  0.67742
## stateTX              100.41     671.53   0.150  0.88121
## stateUT            -4185.85    1031.51  -4.058 6.00e-05 ***
## stateVA              817.21     871.90   0.937  0.34920
## stateVT             -801.42    1707.87  -0.469  0.63916
## stateWA             -249.74     811.64  -0.308  0.75848
## stateWI               37.45     804.03   0.047  0.96287
## stateWV             -394.97    1695.45  -0.233  0.81592
```

```
## log(land.area)      -728.96     129.22  -5.641 3.27e-08 ***
## log(doctors)         910.07      91.18   9.981  < 2e-16 ***
## log(pct.bach.deg)   5760.40     514.29  11.201  < 2e-16 ***
## log(pct.below.pov) -3375.45     289.69 -11.652  < 2e-16 ***
## pct.hs.grad          -57.33      25.19  -2.276  0.02337 *
## pop.18_34           -245.78      30.04  -8.182 4.10e-15 ***
## pop.65_plus           64.71      32.63   1.983  0.04809 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1581 on 385 degrees of freedom
## Multiple R-squared:  0.8669, Adjusted R-squared:  0.8482
## F-statistic: 46.43 on 54 and 385 DF,  p-value: < 2.2e-16
```

(c) All subsets selection

```
all_subsets <- regsubsets(per.cap.income ~ state + log(land.area) + log(doctors) +
    log(pct.bach.deg) + log(pct.below.pov) + pct.hs.grad + pop.18_34 +
    pop.65_plus, data = cdi, really.big = TRUE)
```

```
cdi_sum <- summary(all_subsets)
data.frame(
  adj_r2 = which.max(cdi_sum$adjr2),
  cp = which.min(cdi_sum$cp),
  bic = which.min(cdi_sum$bic)
)
```

```
##   adj_r2 cp bic
## 1      8  8   8
```

```
coef(all_subsets, 1:8)
```

```
## [[1]]
##        (Intercept) log(pct.below.pov)
##          29237.812          -5254.348
##
## [[2]]
##        (Intercept)       log(doctors) log(pct.below.pov)
##          18350.851           1768.104          -5249.416
##
## [[3]]
##        (Intercept)       log(doctors)  log(pct.bach.deg) log(pct.below.pov)
##          12369.806           1434.751           2211.344          -4545.989
##
## [[4]]
##        (Intercept)       log(doctors)  log(pct.bach.deg) log(pct.below.pov)
##         12359.3280          1266.7217          4217.8824         -3814.2941
##          pop.18_34
##           -225.1977
##
## [[5]]
##        (Intercept)     log(land.area)       log(doctors)  log(pct.bach.deg)
```

```
##         17441.6279          -718.6741          1222.1396          4228.8712
## log(pct.below.pov)          pop.18_34
##         -3591.8006          -246.5155
##
## [[6]]
##        (Intercept)            stateCA     log(land.area)       log(doctors)
##         19161.8368          2023.8250          -926.3830          1136.5993
##  log(pct.bach.deg) log(pct.below.pov)          pop.18_34
##          4289.0459         -3564.8647          -254.6027
##
## [[7]]
##        (Intercept)            stateCA            stateUT     log(land.area)
##         18885.5656          1971.8515         -4945.9804          -918.7941
##       log(doctors)  log(pct.bach.deg) log(pct.below.pov)          pop.18_34
##          1113.5936          4443.6745         -3512.9550          -259.8475
##
## [[8]]
##        (Intercept)            stateCA            stateNJ            stateUT
##         17714.7383          1986.3657          2317.9648         -4871.2723
##     log(land.area)       log(doctors)  log(pct.bach.deg) log(pct.below.pov)
##          -832.0761          1073.6250          4520.9466         -3328.7360
##          pop.18_34
##          -254.6016
```

```
cdi["stateCA"] <- ifelse(cdi$state == "CA", 1, 0)
cdi["stateNJ"] <- ifelse(cdi$state == "NJ", 1, 0)
cdi["stateUT"] <- ifelse(cdi$state == "UT", 1, 0)
fit_8 <- lm(per.cap.income ~ stateCA + stateNJ + stateUT + log(land.area) + log(doctors) + log(pct.bach
summary(fit_8)
```

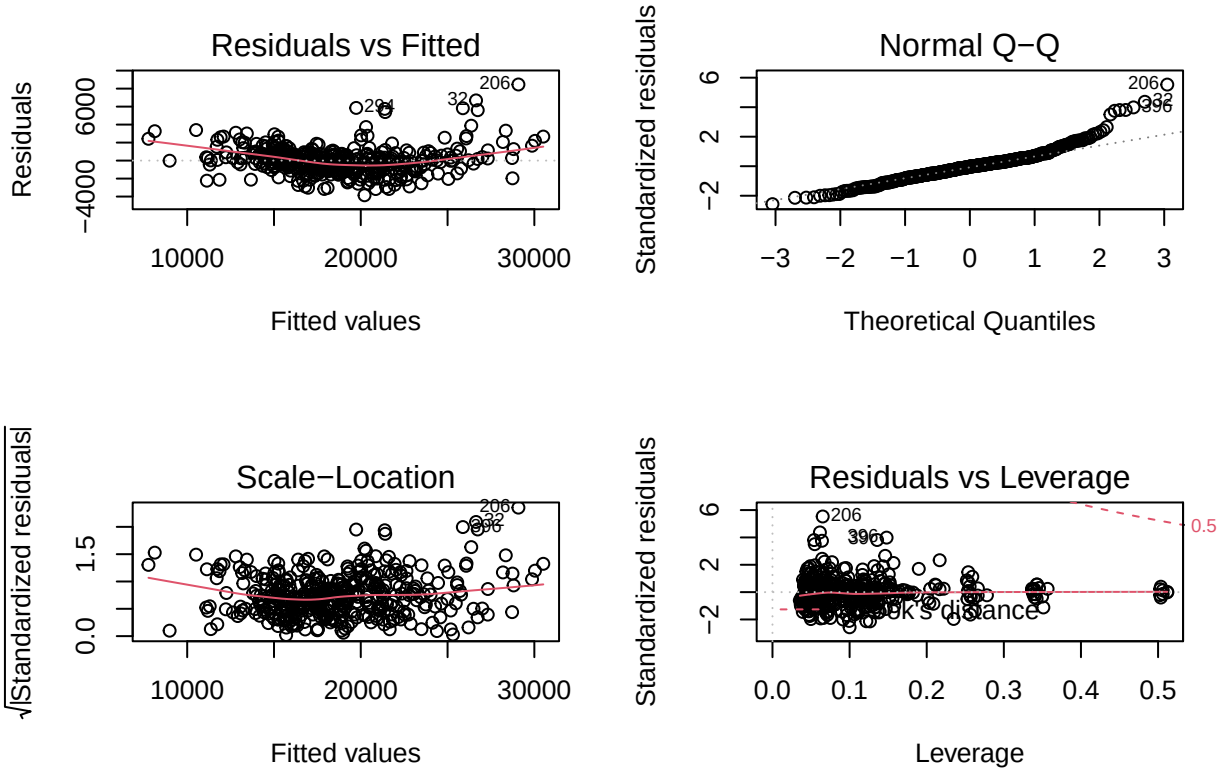
```
##
## Call:
## lm(formula = per.cap.income ~ stateCA + stateNJ + stateUT + log(land.area) +
##     log(doctors) + log(pct.bach.deg) + log(pct.below.pov) + pop.18_34,
##     data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3661.4 -1003.1  -160.0   777.7  8463.7
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        17714.74    1222.13  14.495  < 2e-16 ***
## stateCA             1986.37     314.54   6.315 6.72e-10 ***
## stateNJ             2317.96     413.48   5.606 3.70e-08 ***
## stateUT            -4871.27     828.75  -5.878 8.33e-09 ***
## log(land.area)      -832.08      99.60  -8.355 9.09e-16 ***
## log(doctors)        1073.63      85.77  12.518  < 2e-16 ***
## log(pct.bach.deg)   4520.95     366.41  12.338  < 2e-16 ***
## log(pct.below.pov) -3328.74     196.76 -16.918  < 2e-16 ***
## pop.18_34           -254.60      23.18 -10.984  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1645 on 431 degrees of freedom
## Multiple R-squared:  0.8388, Adjusted R-squared:  0.8358
## F-statistic: 280.4 on 8 and 431 DF,  p-value: < 2.2e-16
```

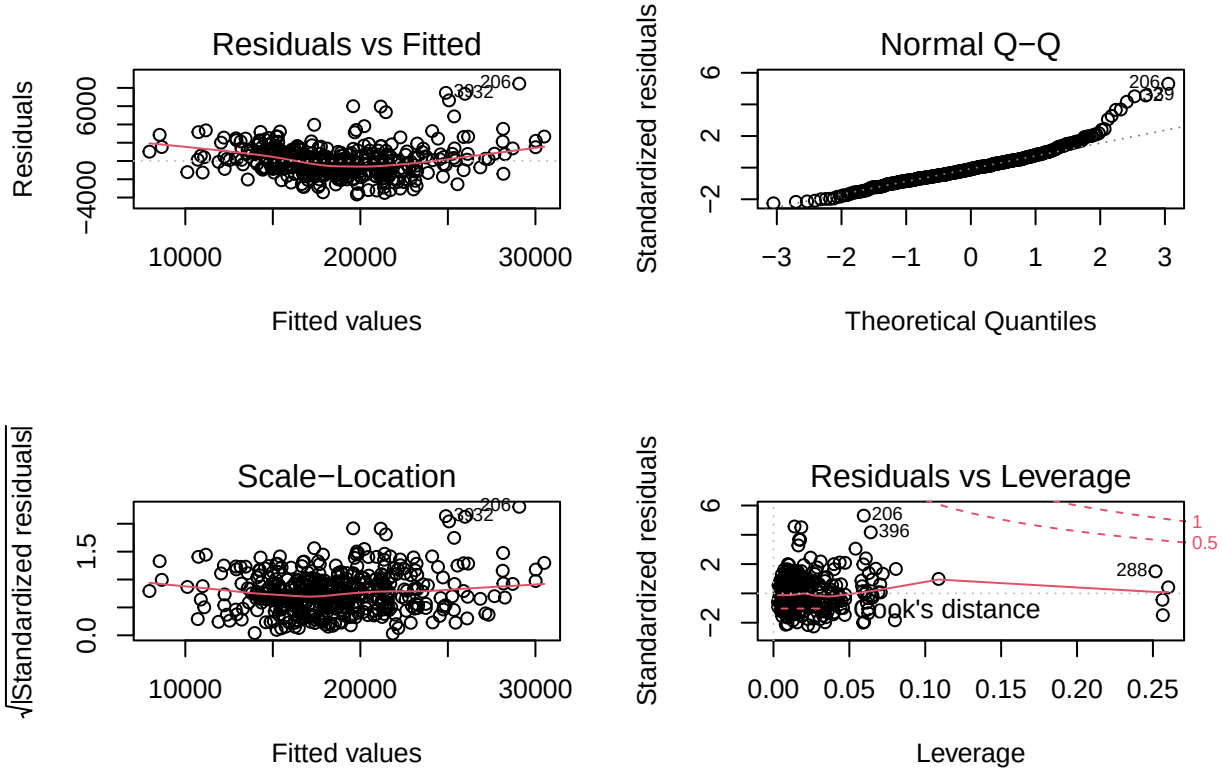
(d) Model comparison

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

```
## Warning: not plotting observations with leverage one:
##   73, 232, 233, 339, 356, 388, 429
```




```
par(mfrow=c(2,2))
plot(fit_8)
```



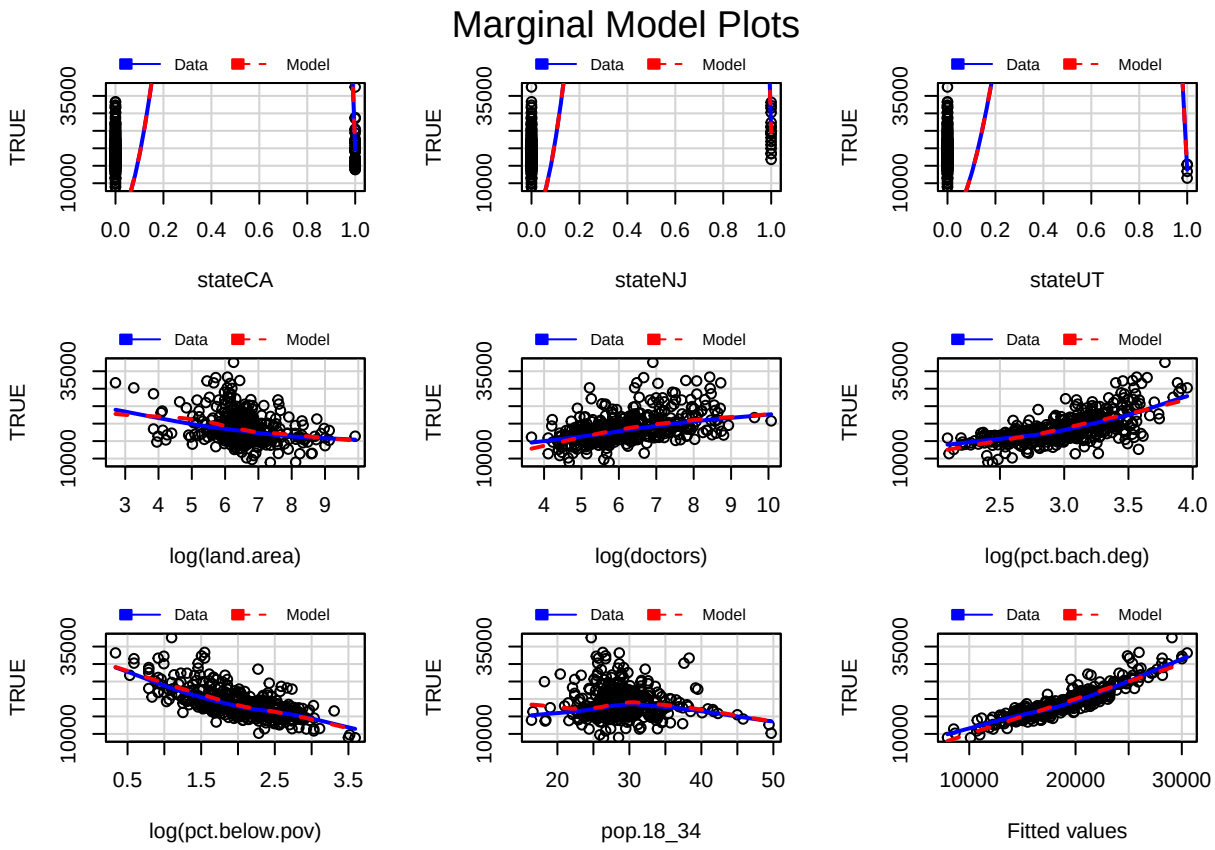

```
AIC(step_model, fit_8)
```

```
##            df      AIC
## step_model 56 7784.069
## fit_8      10 7776.268
```

```
BIC(step_model, fit_8)
```

```
##            df      BIC
## step_model 56 8012.928
## fit_8      10 7817.135
```

```
mmps(fit_8)
```




```
coefficients(fit_8)
```

```
##        (Intercept)             stateCA             stateNJ             stateUT
##         17714.7382          1986.3657          2317.9648         -4871.2723
##     log(land.area)       log(doctors)  log(pct.bach.deg) log(pct.below.pov)
##          -832.0761          1073.6250          4520.9466         -3328.7360
##          pop.18_34
##          -254.6016
```

step_model : per.cap.income ~ state + log(land.area) + log(doctors) + log(pct.bach.deg) + log(pct.below.pov) + pct.hs.grad + pop.18_34 + pop.65_plus, has r squared value of 0.8482 fit_8: per.cap.income ~ stateCA + stateNJ + stateUT + log(land.area) + log(doctors) + log(pct.bach.deg) + log(pct.below.pov) + pop.18_34, has r squared value of 0.8358 From the two sets of diagnostic plots, we can find it is hard to choose based on those. Both resduals vs.fitted plots and scale-location plots has a slight upward concavity patterns shown. And from both normal q-q plots, we can see that they all have some points off the line at the very right side, which also corresponds to the residuals vs.leverage plots that they both have points with either high residuals or high leverage, but no point with both high residuals and high leverage. They have quite similar r squared values and it is also hard to choose based on this single value. However, fit_8 has smaller AIC and BIC values. Also, I will choose model fit_8 as a better one as it has far less variables (there are many variables "hidden" under the state in step_model). In this model, apart from those numeric variables that may contribute to the prediction of per.capita.income, three states are also extracted as special representatives, which may have especially high or low per capita income in this dataset (which is not big thus may be biased in some way) that can affect the model. Not only how many infrastructures/features this state has can be related to its per capita income, sometimes the state itself also means something, and I think this really relaed to social science and economics of a state other than pure statistical inference.