Intermediate group project report

Zhenzhen Ge and Linfeng He

November 29, 2016

Following our prior project proposal, we have processed to the stage of data gathering and analyzing. The sources of information included, but not limited to, various government agencies such as FBI, United Census Bureau, National Center for Education Statistics, etc. Our finding has not been fully finalized yet as we are still investigating. Our early results has suggested that the bachelor graduation rates have quite explanatory powers over the income per capital among states.

Data Description

As it has been iterated, our study would be centered on the socio-economical factors of each state. Thus, numbers like the population of every state, would play a significant role in our study. We have collected and selected eight variables which we believe which would be important to achieving our desirable outcomes:

1. Pop: The population of the state[1]
2. Percap.Consum: Personal Consumption Expenditures by state, 2015[2]
3. LifeExp: Average life expectancy from birth (all values expressed in years)[3]
4. HS.Rate: High school graduation rate, latest[4]
5. BA.Rate: Bachelor school graduation rate, latest[4]
6. Area: Land area of the state[5]
7. Violent.Crime: The violent crime figures include the offenses of murder, rape, robbery and aggravated assault[6]
8. Percap.Income: State Personal Income[7]

[1] Wikipedia: https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population

[2] Bureau of Economic Analysis: http://www.bea.gov/newsreleases/regional/pce/pce_newsrelease.htm

[3] Institute for Health Metrics and Evaluation: http://www.healthdata.org/us-health/data-download

[4] U.S. Department of Education: https://nces.ed.gov/programs/digest/d15/tables/dt15_104.85.asp?current=yes

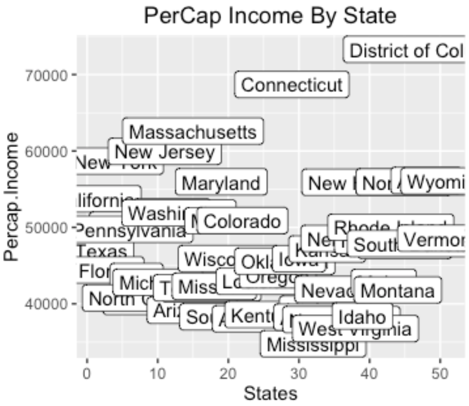
[5] Bureau of U.S. Census: https://www.census.gov/geo/reference/geoguide.html

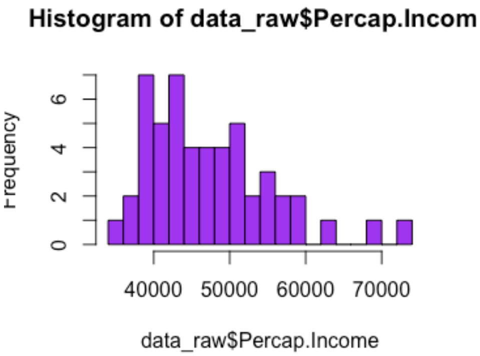
[6] FBI Uniform Crime Report: https://ucr.fbi.gov/crime-in-the-u.s/2015/crime-in-the-u.s.-2015/tables/table-4

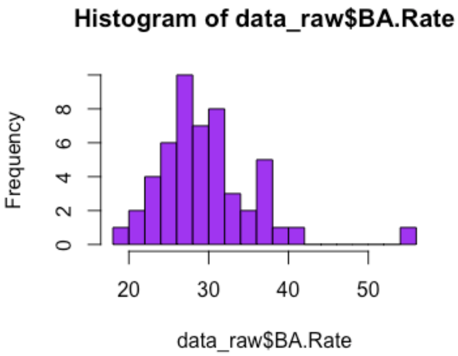
[7] Bureau of Economic Analysis: http://www.bea.gov/newsreleases/regional/spi/sqpi_newsrelease.htm

Preliminary Analysis

When we completed the data gathering, we first take a glance at the overall data, such as the distribution of personal income, bachelor graduation rate, and other variables by states, to get a general understanding of the data. As we have speculated that, there are inequalities in population (obviously), area, consumption per capital, bachelor degree rate, violent crime rate and income per capital among states. Whereas the life expectancy and high school graduation rate seem to be more constant (since they are quite substantial already).







Methodologies

1. Model fit and hypothesis test

We fit models with per capita income as the response variable, and all others as predictor variable. By evaluating p-values, we can see two variables(bachelor graduation rate and personal consumption) are significant at level 0.5.

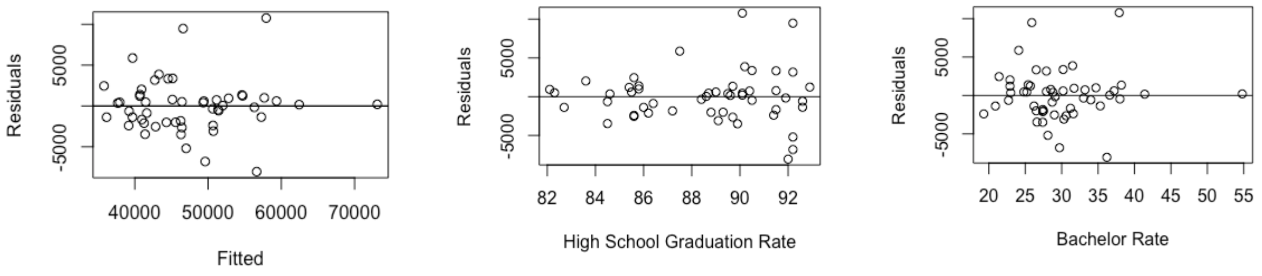
```
Call:
lm(formula = Percap.Income ~ Pop + Percap.Consum + LifeExp + HS.Rate + BA.Rate + Area + Violent.Crime, data_raw)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.496e+04  3.051e+04   1.146   0.2582
Pop            2.715e-04  3.712e-04   0.731   0.4685
Percap.Consum  1.092e+00  1.650e-01   6.620 4.57e-08 ***
LifeExp       -4.612e+02  4.889e+02  -0.943   0.3507
HS.Rate       -4.320e+01  2.877e+02  -0.150   0.8813
BA.Rate        3.138e+02  1.450e+02   2.164   0.0361 *
Area          -5.923e-04  2.184e-03  -0.271   0.7875
Violent.Crime -3.135e-02  8.929e-02  -0.351   0.7272
```

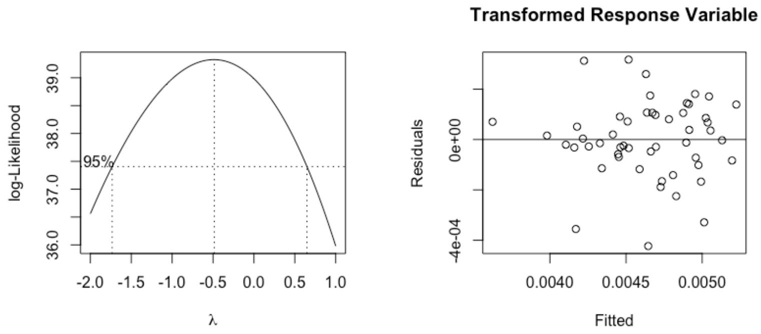
2. Diagnostics

- Constant Variance

We also want to check whether our assumptions of this linear model are correct. So we run series of tests to check whether constant variance holds under current model.

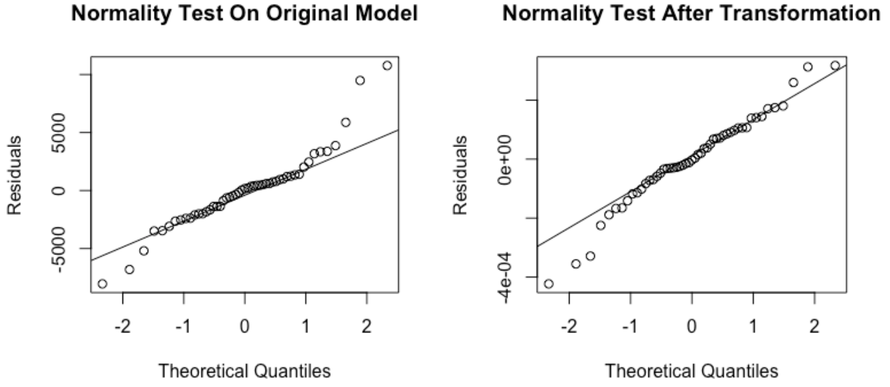


Then conduct Box-Cox method to find the best transformation for the response variable. After the transformation, we get improved plot which has more constant variance, however the improvement is minor.



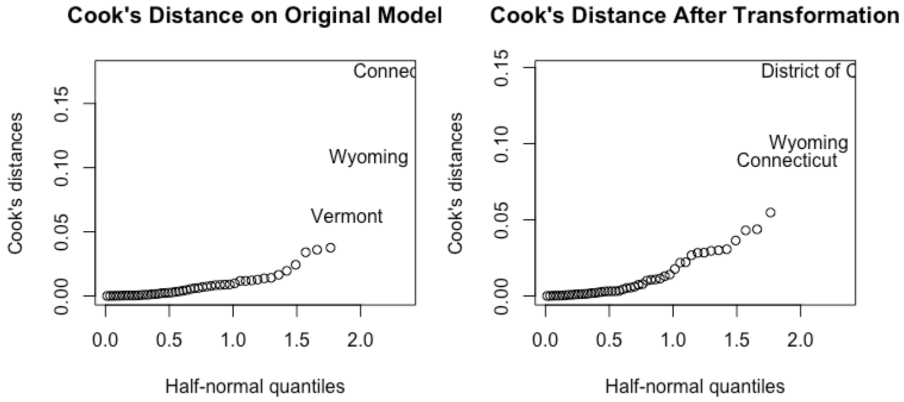
- Normality

We check both models whether normality holds. On both QQ plots we can see there might be heavy tail issues.



- Outliers

We look at possible outliers by plot their cook's distance. The values are all smaller than 1, therefore we may assume no outliers among the states.



3. Robust regression

Since we have heavy tail issues in our data, we will try to use robust method to estimate our model.

4. Model selection

Next we will implement model selection using Information Criterion method.

5. Inference from final model

We shall conclude our findings basing on final model. Furthermore, we also want to see whether different area fits quite different model. Despite of the limited information we have gathered thus far, we are inspired to bring the scale of studies to the regional level. Specifically, we may split all states into west part and east part and repeat some steps to investigate. Also, we will extend our response variable not just income per capital but to the rest of variables whichever we deem necessary, and test whether there are more stories to be told about the relationships among variables.