Introduction to Data Science, Project 2 Report

Deng Kaisheng

Code running guidance

1. Put "P02_Corporate tax.xlsx" in the code file path.
2. Put all the ACS data files in a folder called "ACS_DP02 data", and put all the csv files in the folder. Put the folder in the code file path.
3. The relative paths are like this:

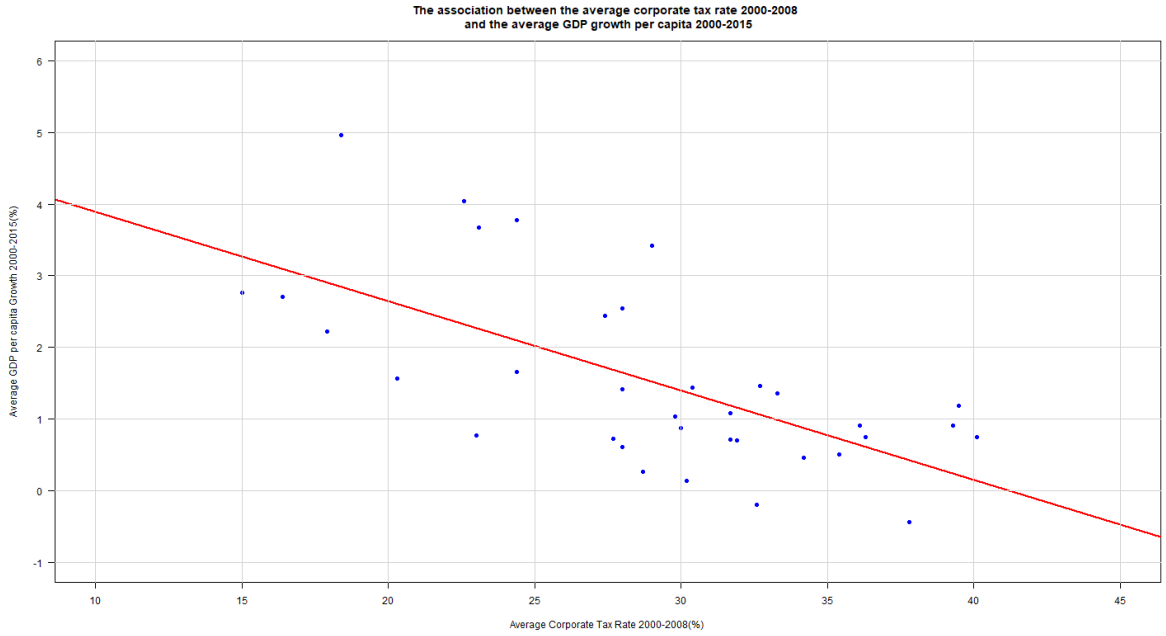
名称	修改日期	类型	大小
ACS_DP02 data	2019/7/27 14:12	文件夹	
P02_Corporate tax.xlsx	2019/7/25 11:15	Microsoft Excel 工...	19 KB
task_A.R	2019/7/27 13:52	R 文件	3 KB
task_B.R	2019/7/27 11:42	R 文件	6 KB

Then run the code, the output files will be in the same path of the code files.

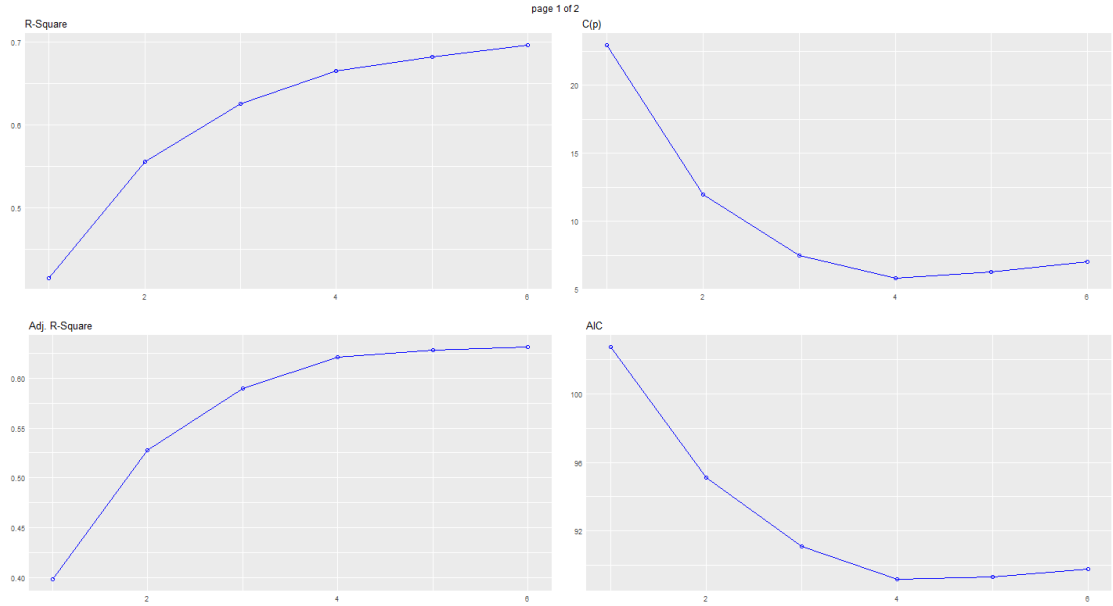
task A

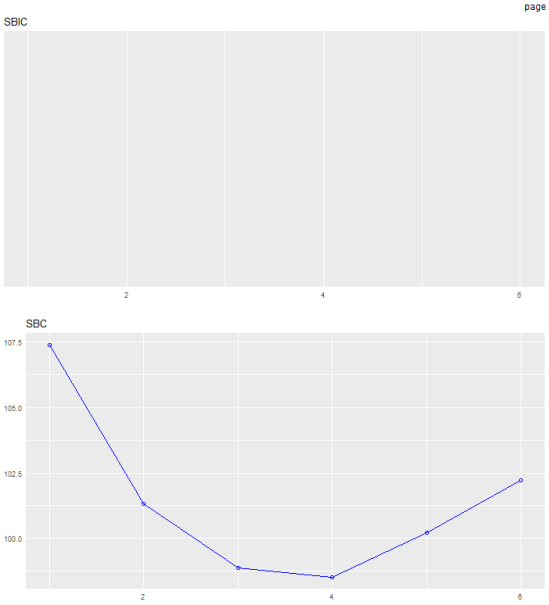
Full code see file "task_A.R"

- Using the variables given in the 3 equations in page 55, I trained 3 linear models, and their summaries **fit the conclusion in the paper perfectly**.
- Based on equation 3, if a country has corporate tax rate = 20%, GDP per capita in 2000 = $10,000, and debt to GDP ratio = 35%, **the hypothetical GDP per capita growth will be 3.24%.**
- I used R code to plot a figure almost the same as figure 4. **See file "Figure4_using R.png"**



- Think in the Next: Why do I use corporate tax rates averaged from 2000 to 2008 instead of from 2000 to 2015?
    - I think one reason is the complex interaction between the different factors.
    - **Another reason, I guess, is that there should be some time for the independent variables to take effect.**
- I used *olsrr* package to do model selection works. **The report of model selection is in file "Model_Selection_Report.csv**. In the column "predictors", the variables are the most capable to train the model are listed. We can reach the following conclusions:
    - When it comes to Adjusted-R Squared value(adjr), the most effective model uses 6 predictors: ctax, ypc2000, dty, trade, ihc, and y2000.
    - When it comes to AIC value(aic), the most "balanced"(complexity and accuracy) model uses 4 predictors: ctax, ypc2000, trade, and ihc.
    - I also visualize the comparations between these models. See file "Model_Selection_Analysis 1.png" and "Model_Selection_Analysis 2.png"





- And then **I compared the model with 4 predictors and that with 6 predictors, the results are like below**,(see file "Model_Summary_4_Predictors.png" and "Model_Summary_6_Predictors.png")

```
Call:
lm(formula = ypcg ~ ctax + ypc2000 + trade + ihc, data = corporateTax)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3348 -0.5379 -0.1456  0.5344  1.8154

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.499e+00  1.355e+00   1.106  0.27750
ctax        -8.525e-02  2.414e-02  -3.531  0.00136 **
ypc2000     -3.779e-05  8.326e-06  -4.539 8.54e-05 ***
trade        6.596e-03  3.499e-03   1.885  0.06912 .
ihc          1.038e+00  3.795e-01   2.735  0.01038 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7871 on 30 degrees of freedom
Multiple R-squared:  0.6657,    Adjusted R-squared:  0.6211
F-statistic: 14.93 on 4 and 30 DF,  p-value: 8.013e-07
```

```
Call:
lm(formula = ypcg ~ . - country, data = corporateTax)

Residuals:
     Min       1Q   Median       3Q      Max
-1.15884 -0.52871  0.02501  0.28981  1.66408

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.066e+00  1.387e+00   1.490 0.147501
ctax        -7.851e-02  2.934e-02  -2.675 0.012323 *
ypc2000     -3.734e-05  8.245e-06  -4.528 0.000101 ***
dty         -6.490e-03  5.431e-03  -1.195 0.242140
trade        6.510e-03  3.732e-03   1.744 0.092120 .
ihc          8.866e-01  3.852e-01   2.301 0.029024 *
y2000        1.025e-04  9.113e-05   1.124 0.270355
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7762 on 28 degrees of freedom
Multiple R-squared:  0.6965,    Adjusted R-squared:  0.6314
F-statistic: 10.71 on 6 and 28 DF,  p-value: 3.474e-06
```

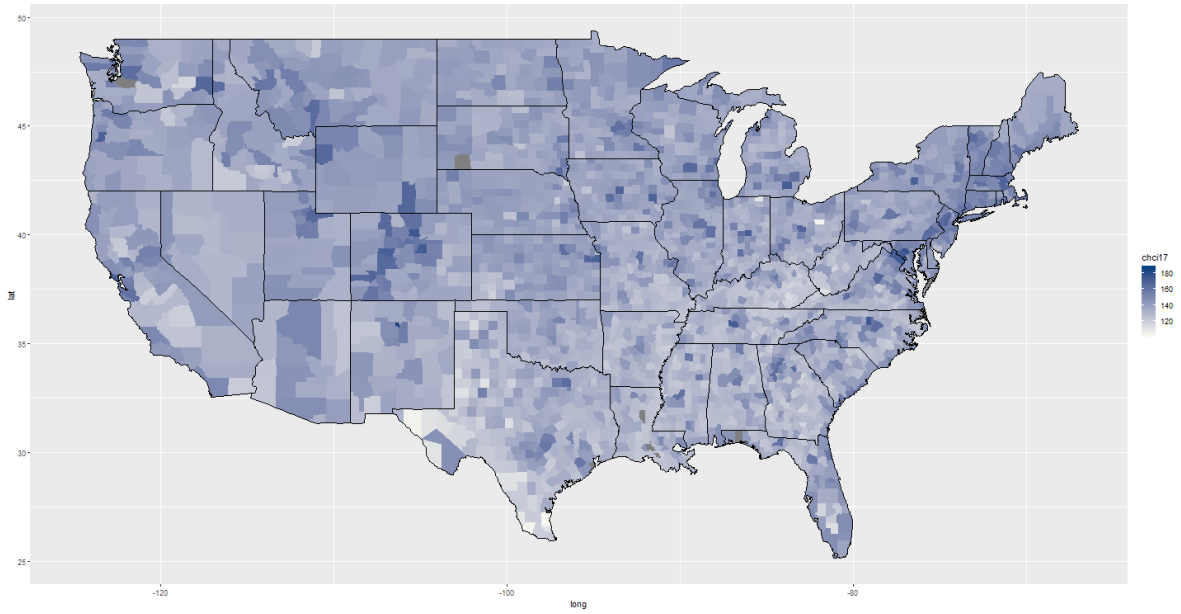
- It can be found that **in the 6-predictors' model, the p-value of Intercept, dty and y2000 are not within a acceptable range, which actually means they are not suitable in this model.**
- However, **in the 4-predictors' model, all the 4 predictors' p-value are within an acceptable range, which ensure the "mistaken probability" less than 5%**
- What's more, the increase of adjr of 6-predictors from 4-predictors are actually not very obvious. **And because of the lower AIC, I believe the 4-predictors' model will perform better than the 6-predictors' one in generalization.**
- According to the previous analysis, I personally prefer the model with 4-predictors, the equation is:

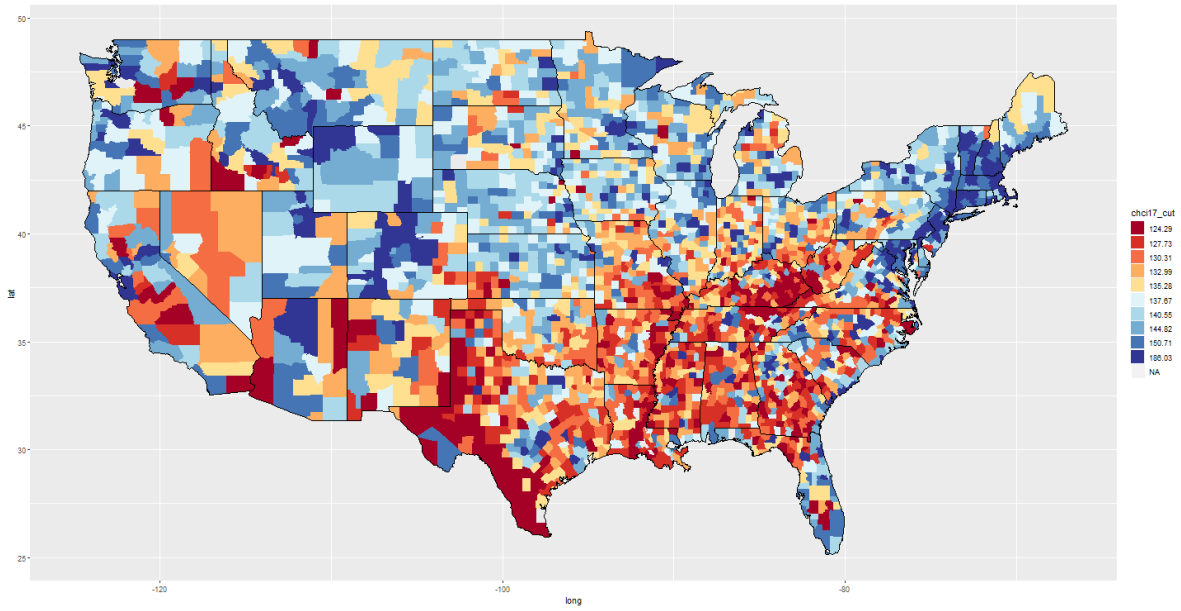
  ypcg = 1.499 - 0.085 * ctax - 0.0000378 * ypc2000 + 0.0066 * trade + 1.038 * ihc

task B

Full code see file "task_B.R"

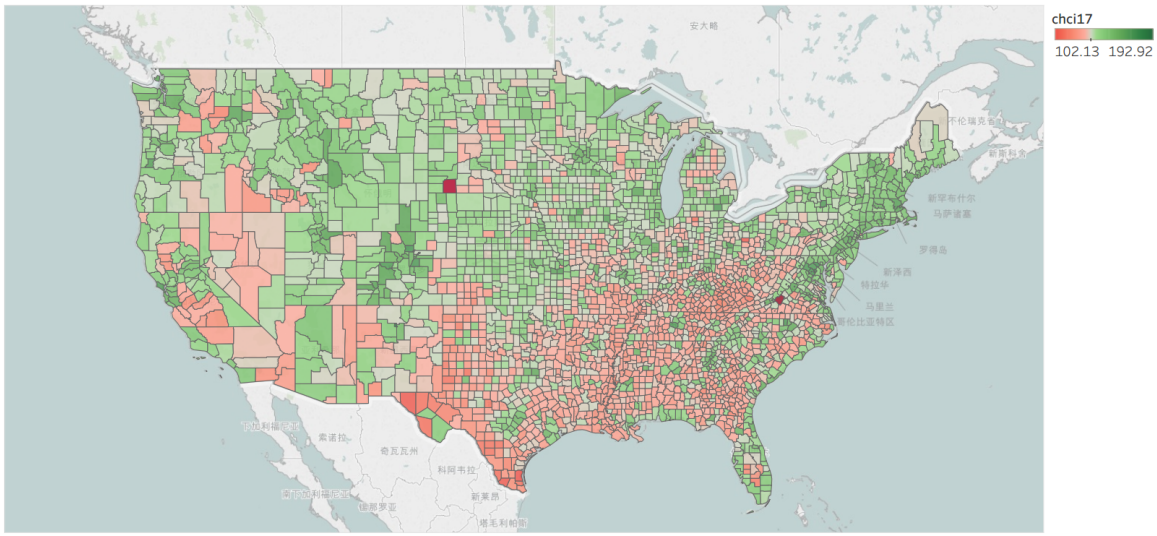
- The output data frame is written in file "CHCI_POP.csv"
- The counties with top 10 highest chcigr is listed in file "CHCIGR_top10.csv"
- The counties with top 10 highest popgr is listed in file "POPGR_top10.csv"
- Using *ggplot()* function I plotted two figures, one with gradient color and the other with divergent color. See file "CHCI_GradientColor.png" and file "CHCI_DivergentColor.png"



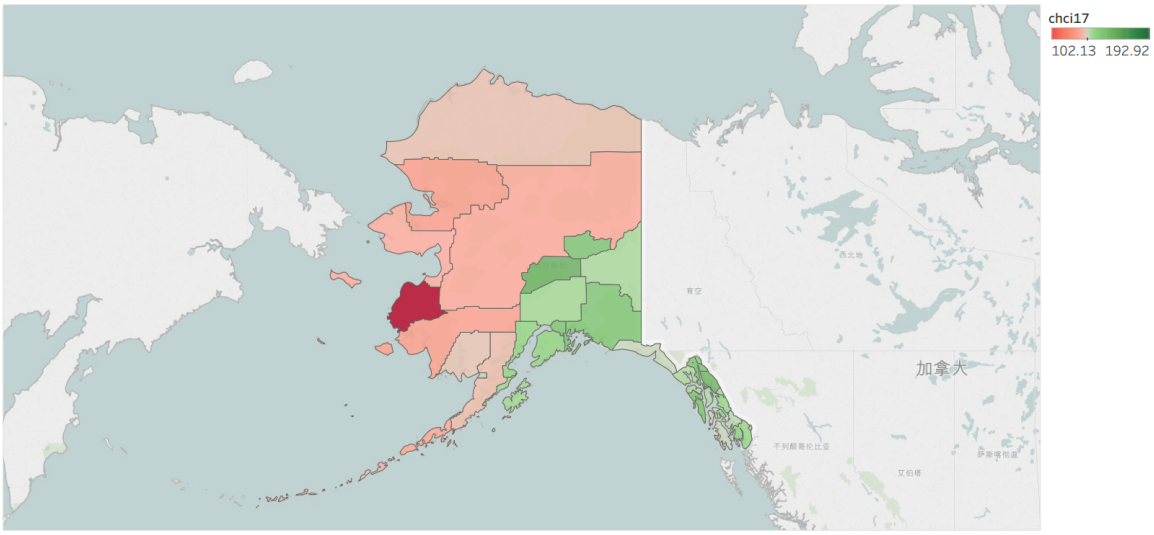


- **There are three figures plotted by Tableau, and I exported all of them to PDF files.** See file "CHCI2017 Of Mainland, United States.pdf", "CHCI2017 Of Alaska, United States.pdf" and "CHCI2017 Of Hawaii, United States.pdf", the screen cuts are like below:

CHCI2017 Of Mainland, United States

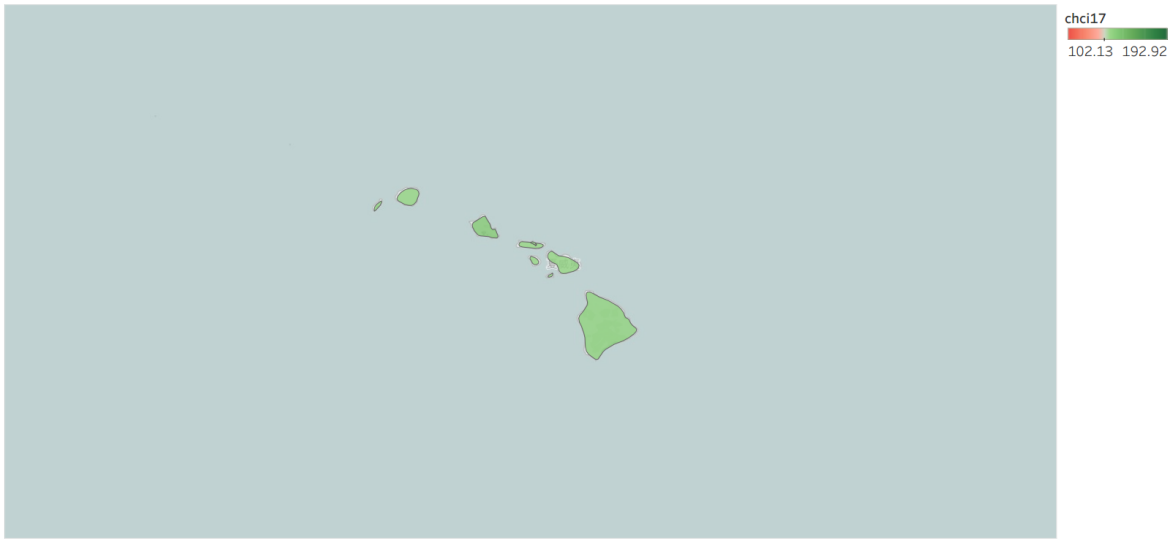


CHCI2017 Of Alaska, United States



Map based on Longitude (generated) and Latitude (generated). Color shows details about chci17. Details are shown for MyFips and county.

CHCI2017 Of Hawaii, United States



Map based on Longitude (generated) and Latitude (generated). Color shows details about chci17. Details are shown for MyFips and county.

References

[ColorBrewer: Color Advice for Maps]

[Tableau Help: Create a Simple Calculated Field]

[Maps in R: Plotting data points on a map]