UCLA Extension Data Science Intensive

Instructor: William Yu

Project 2

Submit your results in one R script for Part A and B with explanations and any outputs you got (in an excel file) through Canvas.

A. Calculate the City Human Capital Index (CHCI) from the American Community Survey (ACS) Data

- Browse the following webpage about CHCI developed by me:
- http://www.anderson.ucla.edu/centers/ucla-anderson-forecast/projects-and-partnerships/city-human-capital-index
- We want to calculate the CHCI by a weighting average of educational attainment of local adult residents. The data is from the 5-year American Community Survey summary file (DP02) from 2009 to 2017.
- Download all the ACS data into your computer. The data is shown by all the counties in the U.S. _metadata is the documentation for variables. _with_ann is the data file.
- Hint: To import an ideal data format from the original data, you could use "skip" setting in read.csv function, e.g. `acsdp2.5y17 = read.csv("ACS_17_5YR_DP02_with_ann.csv", skip=1)`
- Note: You can import the data by an easy and straightforward way, repeat the above code for each year. Or you can try an advanced way by using a loop, which could be difficult.
- You need to use the following variables: #2: Id2, #3: Geography, #232: Adult population, #238: % in less than 9^{th} grade, #242: % in 9^{th} to 12^{th} grade, #246, #250, #254, #258, and #262.
- Hint: the formula for calculating CHCI is: 50 * % in less than 9^{th} grade + 100 * % in 9^{th} to 12^{th} grade + … (see the table below).
- Add two variables into the data frame: CHCI growth from 2009 to 2017, and CHCI growth rate from 2009 to 2017.
- Your expected output should look like chci.xlsx attached in the folder.
- Reorder the data based on the following variables (e.g. you can use "arrange" function under library dplyr.
- In the class (in D02e_map), we plot the U.S. county map with color for CHCI in 2016 (period: '12-'16; midyear: 2014). Now apply it to plot the similar chart for CHCI in 2017 in R and save it into a separate tab in the excel output file.
- Plot a similar colorful county map in Tableau and save it into your excel output file as well. Note: Tableau can automatically recognize "county" variable as a graphic variable. But it might not help you to get the map we want to see. Instead, you could use "id" or "fips5" variable (5-digit number). And then click "geographic role" and choose "county" to force Tableau to recognize it as a geographic variable. But for 5 or 6 states, the id begins with 0 digit and it will be presented as only 4-digit number. As a result, Tableau cannot recognize it. Solution: in the csv file, you could save the county fips5 as a text

variable in B2 by typing " =text(A2, "00000")" and save the file as a xlsx file. (As an example, see Column B in chci.xlsx.)

- Tell a simple story from the chart you plotted.

The CHCI is calculated based on the following education attainment data (**multiplied by 10**) for adult residents in each area:

Education Attainment	Assigned Schooling Year	Calculation weight
Less than 9th grade	5	**50**
9th to 12th grade, no diploma	10	**100**
High school graduate (includes equivalency)	12	**120**
Some college, no degree	13	**130**
Associate's degree	14	**140**
Bachelor's degree	19	**190**
Graduate or professional degree	23	**230**

B. Regression: An Example of Turning A Small Dataset Into Knowledge

- Read my article: "Will A Lower Corporate Tax Rate Boost Economic Growth?"
- In the Course Website, download p02_Corporate tax.csv and save it into your computer.
- Write a R code to answer the following questions:
- On Page 55, there are three equations. Replicate these regression results by using R. Note: Beta 2 should be -0.00002.
- Based on Equation 3, use its coefficients (alpha and betas) to predict a hypothetical GDP per capita growth rate when a country has a corporate tax rate = 20%, GDP per capita in 2000 = $10,000, and debt to GDP ratio = 35%.
- Plot a chart similar to Figure 4 by using R. (The red line is a regression, fitted line). Note: the R output chart you got might be a bit different from that (a bit distorted) in the book. No worries.
- *Think in the Next:* Why do I use corporate tax rates averaged **from 2000 to 2008** instead of **from 2000 to 2015**?
- The dataset provides more variables (description as follows). Play around by adding these variables. And present the best model (using adj. R2) and briefly explain the result.

	OECD 35 Countries
Variable	Description
ypcg	GDP per capita growth rate, average from 2000 to 2015
ctax	Corporate tax rate (%), average from 2000 to 2008
ypc2000	GDP per capita (US$) in 2000
dty	Debt to GDP ratio (%), average from 2000 to 2008
trade	Trade (imports and exports) as percentage of GDP (%)
ihc	Index of country's human capital
y2000	GDP in 2000 (economy size, US billion $)