Harvard Extension Data Science

Dynamic Modeling and Forecasting in Big Data

Instructor: William Yu

Assignment 1

Submit your assignment in R Script/Python notebook or Jupyter notebook via Canvas before due date. You can add explanations in R script /Python notebook directly.

Part A: Linear Regression: An Example of Turning A Small Dataset Into Knowledge

- Read my article: "Will A Lower Corporate Tax Rate Boost Economic Growth?"
- In the assignment folder, download P02_Corporate tax.xlsx and save it into your computer.
- Write a code to answer the following questions:
- On Page 55, there are three equations. Replicate these regression results. 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. (The red line is a regression, fitted line). Note: the output chart you got might be a bit different from that (a bit distorted) in the book. No worries.
- Briefly explain why I use GDP per capita in 2000, not 2015? 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 $)

- Note: I used the difference of corporate tax rates across country (cross sectional) to imply whether a country's (i.e. U.S.) corporate tax rate changes (time series) will predict a change of economic growth rate. It is not a perfect dataset to use due to the lack of U.S. time series (constant corporate tax rate in the recent decades).

Part B: A Small Research Project: Learn How to Pull All Datasets Together

This question is optional for students pursuing undergraduate credit.

- Go to the folder of "Health in America Covid19 Variation" under Data and Script 2 in Week 2 and read my report, "Health in America: What explains the variation in COVID-19 mortality rate across the U.S.?" I used a simple linear regression / cross-sectional model to find the predictors for the COVID-19 accumulated case and death rates up to January 23, 2021.
- Open and run H02b_crossSection.R, in which I used it to run the regressions with the related datasets. To get the data, you need to have already activated your own API key in Line 30-32 and change the working directory. Follow my guidance in the script. Through the script, you can understand how a research is done from the beginning to the end. And you will also find that American Community Survey (ACS) is a great source to get background social, economic, and demographic data. In the future, you might find it useful.
- Go to USA Facts (https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/) to download the latest COVID-19 death data (under DOWNLOAD DATA). Use the latest accumulative deaths rate (**December 31, 2022**) by county as the new dependent variable and redo the Models 1 (Page 94) and 5 (Page 96). To simplify the assignment, keep all the explanatory variables the same and you don't need to update their values. The only data you need to update is the dependent variable. And run the regression models again.
  Note: This practice aims to provide insights into how a model's conclusions may or may not change over time with the introduction of new data.
- To simply your R code, you can source H02b_crossSection.R with the following:
  ```
  source("H02b_crossSection.R", echo = TRUE)
  ```

- Assignment submission:
  (1) One-page executive summary for your findings in a word or PDF file. In particular, take a look at what predictors changed their statistical significance and coefficient magnitude (Either single or double spaced, 11 or 12 font size is fine).
  (2) You can put all the visualization charts you want to add after the executive summary as an appendix.