Introduction

Instructor: Yuta Toyama

Last updated: 2020-05-10

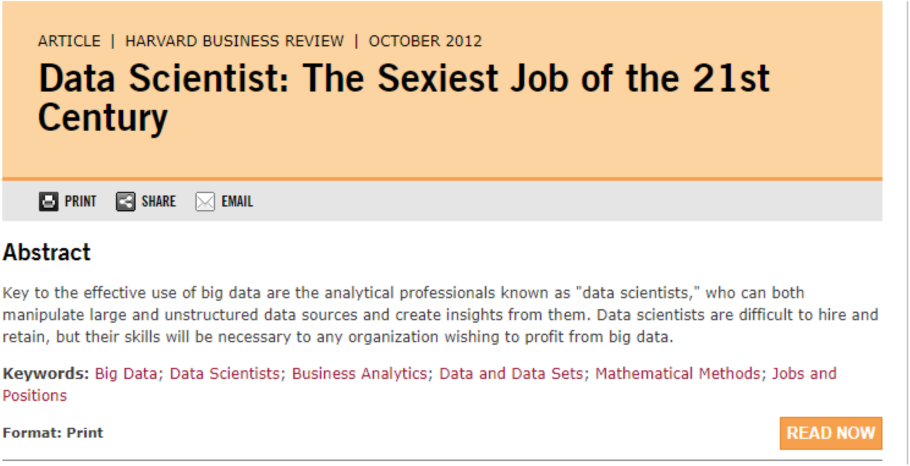
1. Introduction to the course

2. Introduction of `R` and `R studio`

Section 1

Introduction to the course

Data analysis is more important than ever



ARTICLE | HARVARD BUSINESS REVIEW | OCTOBER 2012

Data Scientist: The Sexiest Job of the 21st Century

Abstract

Key to the effective use of big data are the analytical professionals known as "data scientists," who can both manipulate large and unstructured data sources and create insights from them. Data scientists are difficult to hire and retain, but their skills will be necessary to any organization wishing to profit from big data.

Keywords: Big Data; Data Scientists; Business Analytics; Data and Data Sets; Mathematical Methods; Jobs and Positions

Format: Print

- Davenport, Thomas H., and D. J. Patil. “Data Scientist: The Sexiest Job of the 21st Century.” Harvard Business Review 90, no. 10 (October 2012): 70–76. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Goal of the Course

1. Learn **program evaluation and/or causal inference approach** in econometrics
2. Learn how to conduct empirical analysis using **R language**.

Causal Questions in Economics and Politics

- Causality: X causes/affects/impacts Y
- Many questions are causal!
  - How much does an additional year of schooling increase your wage ?
  - How does online advertisement affect sales of products
  - Do mergers between firms increase product prices
  - Does democracy cause economic growth
  - Does higher turnout benefit Democrats in presidential election?
- Use data to infer the causal effects of A on B

Does **correlation** imply **causality**?

- Suppose that X and Y are moving together (**correlated**).
- Examples:
  - 1: Cities with many police officers have more crimes (positive correlation).
  - 2: Those who went to college earn more money by 10%.
- Questions:
  - Does this mean "X causes Y"?
  - Is the magnitude of the correlation equal to that of causal effect?

Three cases

1. Causality
2. Reverse causality
3. Third factor (often called ‘spurious correlation’)

Case: Is Avigan effective treatment for COVID-19?

- Some Japanese celebrities who caught COVID-19 report their own experience that Anti-flu drug Avigan was effective for recovery.
  - An episode from Junichi Ishida: https://www.chunichi.co.jp/chuspo/article/entertainment/news/CK2020042302100136.html
- Does this mean Avigan is silver bullet for COVID-19?
- Why do we care?
  - Side effects of Avigan, resource allocation of public funds, etc.

Difficulty 1: Counterfactual

- **Counterfatual outcome** is never observed
- Even without Avigan, a patient might have gotten better.

Difficulty 2: Endogenous selection of treatment

- Suppose we have data on patients who got and did not get Avigan in their treatment.
- Does comparison of those people give you treatment effect of Avigan?
- Issue: Those who get Avigan and who do not get might be quite different in other aspects.

Solution: Clinical Trial

- Collect patients and **randomly** choose patients who are given the treatment.
  - Treatment group and control group.
- Compare the outcome of treatment and control group to see the effectiveness of treatment.
- Fujifilm is now conducting clinical trial on Avigan for COVID-19. https://asia.nikkei.com/Business/Pharmaceuticals/Fujifilm-starts-clinical-trial-on-Avigan-for-coronavirus

Why do we need to learn computation ?

1. Conduct statistical and empirical analysis using your own data set
   1.1 Construct the data set
   1.2 Describe the data
   1.3 Run regression or estimate an economic object
   1.4 Make tables and figures that show the results of your analysis.

2. Confirm implications from econometric theory through numerical simulations. - Ex. Asymptotic theory considers the case when the sample size is large enough (i.e., $N \to \infty$) - Law of large numbers, central limit theorem - How well is the asymptotic approximation? - So called **Monte Carlo simulations**

Why do we use R?

- Many alternatives: Stata, Matlab, Python, etc. . .

1. Free software!!
   - Stata is expensive.
   - Campus-wide licence for Matlab is available.
2. Good balance of flexibility and easy-to-use for econometrics
   - Stata is easy to use for econometrics, but hard to write your own program.
   - Matlab is the opposite.
   - You can do everything with R, including data construction, regression analysis, and complicated structural estimation.
3. Many users
   - Popular in data science.
   - Many packages being developed (especially machine learning methods)

Course Plan

1. Review of Statistics (1 week)
2. Linear Regression (3 weeks)
3. Instrumental Variable Regression (2 weeks)
4. Panel Data (2 weeks)
5. Program Evaluation and Causal Inference Methods (4 weeks)
   - 5.1 Randomized control trial
   - 5.2 Matching method
   - 5.3 Difference-in-differences
   - 5.4 Regression discontinuity design

Reference

- Lecture notes are based on
  1. Christoph Hanck, Martin Arnold, Alexander Gerber and Martin Schmelzer "Introduction to Econometrics with R" https://www.econometrics-with-r.org/
  2. Wooldridge "Introduction to Econometrics"
  3. Angrist and Pischke "Mastering Metrics"
- Other useful reference for Econometrics
  - Angrist and Pischke "Mostly Harmless Econometrics"
    - Japanese translation also available.
- Other useful reference for R programming
  - Wickham and Grolemund "R for Data Science" https://r4ds.had.co.nz/
    - Japanese translation also available.

Relation to other courses

- This course does **NOT** cover
  1. Discrete choice models
  2. Machine learning methods
  3. Structural estimation

- I recommend the following courses to learn these topics
  - Econometrics II or Applied Econometrics by Prof. Hoshino (for topic 1 and 2)
  - Economic Study (Microeconometrics) by me (for topic 1 and 3)
  - Advanced Econometrics by Prof. Ueda and Prod. Dendup (for topic 1)

Section 2

Introduction of R and R studio

Getting Started

- You can use R/R studio in the PC room.
- However, I strongly recommend you install R/Studio in your laptop and bring it to the class.
- Install in the following order
  1. R: https://www.r-project.org/
  2. Rstudio: https://www.rstudio.com/
- Now open Rstudio.

Helps

- The RStudio team has developed a number of “cheatsheets” for working with both R and RStudio.
- This particular cheatsheet for Base R will summarize many of the concepts in this document.

Quick tour of Rstudio

- There are four panels
  1. Source: Write your own code here.
  2. Console:
  3. Environment/History:
  4. Files/Plots/Packages/Help:
- In the Source panel,
  - Write your own code.
  - Save your code in `.R` file
  - Click `Run` command to run your entire code.
- In the concole panel,
  - After clicking `Run` in the source panel, your code is evaluated.
  - You can directly type your code here to implement.

Basic Calculations

To get started, we'll use R like a simple calculator.

Addition, Subtraction, Multiplication and Division

Math	R	Result
$3 + 2$	`3 + 2`	5
$3 - 2$	`3 - 2`	1
$3 \cdot 2$	`3 * 2`	6
$3/2$	`3 / 2`	1.5

```
1 + 3
```

```
## [1] 4
```

Exponents

Math	R	Result
3^2	`3 ^ 2`	9
$2^{(-3)}$	`2 ^ (-3)`	0.125
$100^{1/2}$	`100 ^ (1 / 2)`	10
$\sqrt{100}$	`sqrt(100)`	10

Mathematical Constants

Math	R	Result
π	`pi`	3.1415927
e	`exp(1)`	2.7182818

Logarithms

- Note that we will use ln and log interchangeably to mean the natural logarithm.
- There is no `ln()` in `R`, instead it uses `log()` to mean the natural logarithm.

Math	R	Result
$\log(e)$	`log(exp(1))`	1
$\log_{10}(1000)$	`log10(1000)`	3
$\log_2(8)$	`log2(8)`	3
$\log_4(16)$	`log(16, base = 4)`	2

Trigonometry

Math	R	Result
$\sin(\pi/2)$	`sin(pi / 2)`	1
$\cos(0)$	`cos(0)`	1

Getting Help

- In using R as a calculator, we have seen a number of functions: `sqrt()`, `exp()`, `log()` and `sin()`.
- To get documentation about a function in R, simply put a question mark in front of the function name and RStudio will display the documentation, for example:

```
?log
?sin
?paste
?lm
```

Installing Packages

- One of the main strengths of R as an open-source project is its package system.
- To install a package, use the `install.packages()` function.
  - Think of this as buying a recipe book from the store, bringing it home, and putting it on your shelf.

```
install.packages("ggplot2")
```

- Once a package is installed, it must be loaded into your current R session before being used.
  - Think of this as taking the book off of the shelf and opening it up to read.

```
library(ggplot2)
```

- Once you close R, all the packages are closed and put back on the imaginary shelf.
- The next time you open R, you do not have to install the package again, but you do have to load any packages you intend to use by invoking `library()`.