Class 2. Data sources

Nic Rivers

Advanced quantitative research methods, API6319
Fall 2019

Data basics

Level of aggregation:

Aggregate data are data for groups of individuals. For example, an aggregate data set might indicate that the total sales of gasoline in Canada in 2017 was 43.6 million litres.

Micro data are data for individuals. For example, a micro data set might indicate that I spent $946 on gasoline in 2018.

Types of data sets

1. Cross-sectional data (multiple units are compared at a single unit in time)
2. Time-series data (a single unit is compared over time)
3. Longitudinal/panel data (multiple units are compared over time)

In each data set, it is important to keep track of:

- The unit of observation (normally stored in rows of the data set)
- The variables (normally a variable occupies a column of the data set)

Notation

- It is conventional to use specialized notation when referring to our data
- We write the variable followed by an subscript, where the subscript refers to the unit of observation
- For example, if the variable is income, and the unit of observation is the country, we would write:

$$\text{Income}_{\text{country}}$$

to refer to show that the variable income is measured for each country.

- More commonly, we would use an abbreviation:

$$Y_i$$

where Y is shorthand for income, and where i is the set of countries in the data.

Cross-sectional data

Country	Crime rate[1]	Per capita income[2]
Canada	1.6	40,541
United States	4.8	48,387
Finland	2.1	36,236
South Africa	32	10,973
Honduras	78	4,345
...	...	...

Unit of observation country

Variables crime rate; per capita income

Notation $\text{crime}_{\text{Canada}} = 1.6$

[1] Intentional homicide rate per 100,000 population, 2010. Source: UNODC Homicide Statistics

[2] At purchasing power parity, 2011. Source: International Monetary Fund.

Time-series data

Year	Smoking percent[3]
1965	42.5
1970	39.5
1975	36.7
1980	33.6
1985	30.4
1990	28.2
1995	25.0
2000	23.1
2005	17.3
2010	16.2

Unit of observation year

Variables smoking rate

Notation $\text{smoke}_{2000} = 23.1$

[3] Percent of Canadian adults 15 years and older that are daily smokers, OECD Health Statistics

Longitudinal/panel data

Country	Year	Per capita CO_2 emissions[4]
Canada	1990	16.2
Canada	2000	17.5
Canada	2008	16.4
US	1990	19.1
US	2000	20.0
US	2008	17.5
China	1990	2.2
China	2000	2.7
China	2008	5.3
...	...	...

Unit of observation country-year

Variables per capita CO_2 emissions

Notation $\text{carbon}_{\text{China},2008} = 5.3$

[4] Per capita CO_2 emissions in metric tons of CO_2 per capita, Source: CDIAC

Aggregate data sources: Canada

- Statistics Canada collects aggregate data on all aspects of the Canadian population and economy: https://www150.statcan.gc.ca/n1/en/type/data
- For example:
  - Youth commencing correctional services, https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3510000401
  - Average expenditures on innovation activities, https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3310018501
  - Arthritis, https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310009606
  - Retail trade sales of motor vehicle and parts dealers, https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=2010000803
  - ...
- You can download the entire table as a csv file and use this in R.
- (There's an R package for this! $\rightarrow$ `CANSIM2R`)

Aggregate data sources: International

- World Bank https://data.worldbank.org/
- World Health Organization http://apps.who.int/gho/data/node.home
- UN http://data.un.org/Default.aspx
- ...

Micro data sources: Canada

- Statistics Canada conducts a national Census every five years, and also conducts about 350 surveys that cover virtually all aspects of Canadian life. (see a list here: https://www.statcan.gc.ca/eng/survey/list)
- As a university-affiliated researcher, you have access to most of the data that these surveys collect through the Public Use Microdata Files (PUMF).
- You can find these PUMFs on the library website, through ODESI (https://search1.odesi.ca/#/)

ODESI data sources: Examples

Canadian Community Health Survey The CCHS is a cross-sectional survey that collects information related to health status, health care utilization and health determinants for the Canadian population.

General Social Survey Canada's General Social Survey (GSS) program was designed as a series of independent, annual, cross-sectional surveys, each covering one topic in-depth. The overall objectives of the program were, and continue to be, to gather data on social trends in order to monitor changes in the living conditions and well being of Canadians, and to provide information on specific social policy issues.

Labour Force Survey The Labour Force Survey provides estimates of employment and unemployment which are among the timeliest and important measures of performance of the Canadian economy.

Census The long-form census is gathered from 1/5 of Canadians every 5 years, and provides information about housing, work, income, family, and demographics.

Survey of Household Spending Detailed estimates of how Canadians spend their money.

...

Confidential files vs. PUMFs

- PUMFs are micro-data files that have been anonymized–any potential identifying information is removed.
- Statistics Canada is very conservative in this respect–removes detailed geographical information, aggregates continuous variables (income, age, etc.) into much coarser groups.
- Most of what you'll likely require to conduct your study on micro-data is likely included in PUMFs.
- BUT, the confidential data files contain some information that is not available in the PUMFs. They are available in the Research Data Centre in the library. https://crdcn.org/carleton-ottawa-outaouais-rdc-cool-rdc. You need to apply for permission to use the files.

International micro-data files

- World Bank has a micro-data repository https://microdata.worldbank.org/index.php/home. For example, Bangladesh Smallholder Survey 2016 https://microdata.worldbank.org/index.php/catalog/2839/get-microdata
- Demographic and Health Survey series: over 400 comparable surveys administered across 90 countries. https://www.dhsprogram.com/
- Inter-university Consortium for Political and Social Research (ICPSR) https://www.icpsr.umich.edu/icpsrweb/ICPSR/
- ...

Other places

- City open data sets: e.g., http://data.ottawa.ca/
- NGOs: e.g., https://www.fraserinstitute.org/school-performance
- Administrative data: e.g., https://www.ices.on.ca/Data-and-Privacy/ICES-data
- Monitoring data: e.g., https://globalfishingwatch.org/data-blog/our-data-in-bigquery/
- ...

Getting data into R

- Once you find the data, save it as a .csv file into an appropriate directory (the /data directory of your project or assignment folder).
- Load the tidyverse package: `library(tidyverse)`
- Bring into R using: `my_data <- read_csv("file name and directory here.csv")`
- Alternatively, click on file–import dataset–from text(readr) and click the file