Introduction to administrative register data

Christian Dudel

May 17, 2023

Materials

Materials are available from GitHub, also mirrored on OSF:

- https://github.com/christiandudel/ThinkData2023
- https://osf.io/h6knq/

What will be covered in this course?

1. Broad overview of benefits and challenges of ARD
2. Handling of ARD using R

What will NOT be covered in this course?

- Every possible pro and con of ARD
- Every possible computational challenge
- Stata, Python, SPSS, SAS, ...
- Analysis

Goals

At the end of this course...

- ...you have a basic idea when ARD is appropriate
- ...you have a basic idea how to handle ARD

Target audience: No experience with ARD

Prerequisites

- No experience of ARD required
- Basic statistical knowledge
- Some experience using statistical software (R or other software)

Course schedule

9:30-11:30 Introduction

13:00-14:30 Handling of big data, classic ARD

15:00-16:30 Complex ARD

Contact

- Email: dudel@demogr.mpg.de
- Office: 358 (3rd floor, east wing)
- Twitter: @c_dudel
- Mastodon @c_dudel@mstdn.social
- Website: http://www.christiandudel.com

My experience with ARD

- Birth register data from many countries
- Social security register data from Spain, Italy, US
- Health register data from Scotland
- Matched survey and social security register data from Germany
- (Combined data from several registers for Norway)

What is ARD? (1)

- No consistent definition in literature
- No consistent terminology: administrative register data, register data, administrative data, administrative register, administrative records, ...

What is ARD? (2)

- "Administrative" = Derived from administrative system
- "Register" = Run continuously, full target population
- "Data" = Quantitative, rectangular data

Examples of ARD

- Population registers
- Birth registers, death registers
- Migration registers
- Tax registers
- Social security registers
- Student registers
- Health registers
- Company/establishment registers
- Housing/building registers
- Vehicle licensing registers

Benefits of ARD

- Size: not just a sample, often large
- Participation: often compulsory (legally required), sometimes highly incentivized
- Nonresponse: Often no (or very few) missing values

ARD in demographic research

- ARD has a very long history in demography
- "Classic" demography often used ARD...
- ... but mostly restricted to vital registration data

ARD in social science research (1)

- ARD is a (mostly) recent development in social science research
- According to Google Scholar: 1,200 publications in 1990-94 with 'register data'...
- ... while in 2015-2019 there were 16,900 publications

ARD in social science research (2)

Why did ARD only become popular recently?

- Availability (digitization)
- Supply and demand
- Computational power

What makes ARD special?

- Found data: Not collected for research purposes
- Found data: Often messy, fragmented, semi-systematic
- Big data: Often large and complex

Challenges of ARD

- Ethical
- Legal
- Technical
- Practical
- Quality

Challenges: Ethical

- Informed consent
- Misuse of registers

Challenges: Legal

- Data protection laws
- Limited access and control

Challenges: Technical

- Size: Requires a lot of computing power
- Complexity: Handling difficult because of fragmentation

Challenges: Practical

- Documentation: Completeness
- Language: Data, documentation, experts

Challenges: Quality

- Total survey error framework: difference between true value of statistic and value derived from survey is due to two main sources of errors, measurement and representation
- Sources of errors can be further decomposed into error components
- Can also be applied to ARD

Total survey error framework: Components

- Representation: Coverage error, sampling error, nonresponse error (unit/item), adjustment error
- Measurement: Validity, measurement error, processing error
- To what extent do these apply to ARD?

Data handling process

- Discovery: Learn about the data
- Structuring: Bring it in a format ready for analysis
- Cleaning: Edit variables, create new variables, etc.
- Enriching: Combine with other data sources
- Validating: Did the previous steps work as planned?
- Analysis: Run your analysis

Readings: General

- https://doi.org/10.1016/j.ssresearch.2016.04.015
- https://doi.org/10.7758/rsf.2019.5.2.01
- https://doi.org/10.1111/j.1467-9574.2011.00508.x
- https://doi.org/10.3917/popu.1302.0215

Readings: Data quality (examples)

- https://doi.org/10.1111/aogs.14445
- https://doi.org/10.1007/s00181-008-0238-6
- https://www150.statcan.gc.ca/n1/pub/12-001-x/2015001/article/14151-eng.pdf