Introduction to administrative register data

Christian Dudel

May 17, 2023

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)

Materials

Materials are available from GitHub, also mirrored on OSF:

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

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

Course schedule

9:30-11:30 Introduction

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

15:00-16:30 Complex ARD

What is ARD?

- "Administrative" = Derived from administrative system
- "Register" = Covers 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 register
- Housing/building registers
- Vehicle licensing

Benefits of ARD

- Often large, not just a sample
- Good coverage: Participation is compulsory/legally required and/or highly incentivized
- Often high data quality with 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
- https://doi.org/10.3917/popu.1302.0215

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: May be large and complex

Challenges of ARD

- Ethical: Informed consent
- Legal: Difficult access
- Technical: Data size/complexity
- Practical: Missing documentation
- Analytic: Limited variables/measurement
- Quality: Data quality