Bootcamp on Administrative Data Analysis (BADA), Summer 2024
Bootcamp en analyse de données administratives (BADA), été 2024

IN-CLASS PORTION / PORTION EN CLASSE (AM)

An initiative of the *Quebec Interuniversity Centre on Social Statistics* (QICSS) and the *Social Statistics Study Group (SSSG)* at INRS

Course Outline
Version: May 7, 2024

Course development and training: Xavier St-Denis, PhD (INRS)
Contact: xavier.st-denis@inrs.ca

Location: QICSS Université de Montréal (Métro Côte-des-Neiges)
3535 chemin Queen-Mary, room 420
Montréal (Québec)
H3V 1H8
Dates: June 10-14, 2024.

Languages: English + French (documentation and presentations will be mostly in English)

Description: The BADA is a summer school intended for graduate students and university or non-university researchers wishing to carry out analyzes with administrative or linked data from Statistics Canada. The summer school aims to allow participants to develop a familiarity with the nature and content of the main administrative databases available in Statistics Canada's Research Data Centres (RDCs). The summer school will focus on three types of data: personal tax data; linkages between survey data and administrative data; and program data (e.g., education and immigration). This in-class portion will focus on the content and limitations of key variables; strategies for record linkages and analytical sample construction; and data innovations such as deriving indirect measures of various dimensions (fertility, family structure, employment, health, etc.) from various administrative datasets. Each day will focus on a specific dataset or category of data.

DAY 0: ADVANCED STATA PROGRAMMING FOR LONGITUDINAL DATA ANALYSIS APPLIED TO CANADIAN ADMINISTRATIVE DATA

Summary

This full day of technical training is designed for individuals with a working knowledge of Stata who wish to develop advanced skills necessary for the analysis of large and complex longitudinal datasets. Specifically, this Stata programming training will benefit researchers who wish to analyse Canadian administrative datasets and related record linkages. The workshop will provide detailed examples of Stata codes as well as exercises that will facilitate data preparation, data management, and the completion of data validation and description steps necessary before performing regression analysis.

Overview

1. The structure of large and complex longitudinal datasets
    a. Basics of longitudinal data analysis
    b. Identifiers and the relational structure of some longitudinal datasets
    c. Commands and features: merge, append, loops and macros
2. Describing life courses using longitudinal datasets
    a. Hierarchical properties of longitudinal data (time, individuals and families)
    b. Preparing longitudinal data for analysis
    c. Deriving key time-variant and time-invariant variables
    d. Commands: bysort, egen, _n, _N, reshape
3. Brief overview of stored/returned values, matrices, and programs
4. Practical workshop: change in a life course outcome, a comparison between two groups
    a. Case no1: cohort data
    b. Case no2: balanced data

Prerequisite

Prior knowledge of Stata is necessary. This activity is not suitable for Stata beginners and will not include an introduction to Stata.

Participants should bring their own laptop with Stata on it if they have one. Those who do not have access to a laptop and a Stata licence can contact Luc St-Pierre (acces@ciqss.ca) to arrange for a laptop to be provided to them on the day of the training.

DAY 1: INTRODUCTION TO ADMINISTRATIVE DATA ANALYSIS WITH A FOCUS ON INCOME TAX DATA

Summary

In this first session, we will introduce key concepts and basic features of administrative data, with a focus on Canadian tax data such as the Longitudinal Administrative Databank (LAD) and other datasets derived from the T1 Family Files (T1FF). Emphasis will be put on the longitudinal and relational nature of the data. The session will conclude with a detailed description of variables on income, program participation, and wealth in tax data.

Overview

1. Introduction

    a. Administrative data vs survey data vs « big data »
    b. What are the different types of administrative data and the process behind administrative data generation?
    c. Basic concepts in administrative data, with a focus on tax data: Family structure, income, tax information, geography

2. Administrative data as longitudinal data: a refresher

    a. Tax data: the example of the Longitudinal Administrative Databank (LAD)
    b. The structure of other tax and administrative datasets

3. Income and wealth in tax data

    a. Definitions and types of income
        i. Source of income
        ii. Pre- vs post-taxes and transfers
        iii. Individual vs family income
    b. Examples:
        i. Poverty and low-income
        ii. Income inequality and top earners in Canada
    c. Wealth in tax data

4. Guest lecture by Antoine Genest-Grégoire (Université de Sherbrooke) on using tax data from a fiscal research perspective

Main datasets discussed

- Datasets derived from the T1FF
- Longitudinal Administrative Database (LAD)
- T4 and other tax slips with information related to income and wealth

DAY 2: STUDYING FAMILIES USING TAX DATA

Summary

Canadian tax data includes detailed information on family structure and family relationships. In this session based on recent research in demography, sociology and economics, we will discuss what types of family dynamics can be observed cross-sectionally and longitudinally in Canadian administrative data, with a focus on data derived from personal tax records. Topics include couple dissolution and reconstituted families, same-sex couples, measures of fertility, and extended kinship networks.

Overview

1. Introduction to T1 Family Files (T1FF) data

    a. Family concept and identifier
    b. Family structure and family relationship variables in tax data
    c. Potential, challenges and issues when using T1FF data
    d. Family concepts in other administrative datasets (IMDB, etc.)

2. Couples in tax data

    a. Marital status
    b. Same-sex couples
    c. Couple formation and dissolution
    d. "First events", life course analysis, and cohort data

3. Direct and indirect observation of children and fertility

4. Kinship beyond the nuclear family in tax data

    a. Observing parents (and grandparents): the example of research on intergenerational mobility and multigenerational relationships
    b. Adult siblings, cousins, uncles, and aunts

5. Guest lecture by Winnie Yang (McGill), on family structures and same-sex couples in tax data

Main datasets discussed

- Datasets derived from the T1FF
- Longitudinal Administrative Database (LAD)

DAY 3: RECORD LINKAGES (LISA, IMDB AND OTHER CASES)

Summary

This session will allow participants to develop a familiarity with record linkage approaches. The main characteristics, properties, and limitations of record linkages will be presented, with a focus on linkages between survey and tax data, federal immigration data (IMDB) and tax data, and external program data integration. The session will also include practical examples guiding participants who wish to perform and validate the quality of record linkages.

Overview

1. Introduction to record linkages

    a. Different linkage structures
    b. Properties of record linkages
    c. How to perform and validate a record linkage

2. Three approaches to performing record linkages:

    a. Linkages between survey data and T1FF data: the Longitudinal and International Study of Adults (LISA)
    b. Linkage between the T1FF and other administrative datasets: the longitudinal Immigration Database (IMDB)
    c. Linkage between external program data and administrative datasets from Statistics Canada

3. Guest speaker: Marcus Fraga (Université Laval) on the longitudinal Immigration Database (IMDB)

Main datasets discussed:

- Longitudinal Immigration Database (IMDB)
- Longitudinal and International Study of Adults (LISA) linked with T1FF and other administrative datasets
- General Social Survey (GSS) linked with the T1FF
- CCHS linked with the T1FF

DAY 4: INNOVATIONS WITH DATA INTEGRATION AND INDIRECT MEASUREMENT

Summary

This session aims to present two types of innovations and challenges. First, it will introduce various linkage platforms that integrate a large number of administrative, survey, and Census datasets, including in some cases provincial data. The potential and limitations of these platforms will be discussed, with a focus on the Education and Labour Market Longitudinal Platform (ELMLP) as a key example. Second, it will discuss strategies used in published research to fill the gaps when it comes to measuring factors and dimensions that do not appear even in the most complex linkage platforms. This includes information on geographic mobility, employment characteristics, and emigration outside of Canada.

Overview

1. Innovations in data integration: linkage platforms and linkage with provincial administrative data

    a. Research on student pathways and outcomes with the Education and Labour Market Longitudinal Platform (ELMLP)
    b. Creating counterfactual groups for immigration research with the Longitudinal Immigration Statistical Environment (LISE)
    c. Quick overview of other platforms (CEEDD/BEAM, CanCHEC, crime and justice data)

2. Challenges with indirect measurement and measurement error in tax data

    a. Place of residence and geographic mobility
    b. Employment and labour supply
    c. Emigration
    d. Other topics to be confirmed (health, participation to postsecondary education, Indigenous identity)

3. Guest speaker: Laëtitia Renée (Université de Montréal) on the Business Employee Analytical Microdata (BEAM)

Main datasets discussed

- Education and Labour Market Longitudinal Platform (ELMLP) including the Postsecondary Education Information System (PSIS), the Registered Apprenticeships Information System (RAIS), and provincial K-12 data (British Columbia, New Brunswick, and other)
- Longitudinal Immigration Statistical Environment (LISE) including the IMDB, LAD and DAD
- Datasets derived from the T1FF