Data Harmonization in Survey Data Integration

January 25, 2015

Second FCSM/WSS Workshop on Quality of Integrated Data

Don Jang

NORC at the UNIVERSITY of CHICAGO

Outline

- Data harmonization
  - What/Why/Challenges/Process
- Example: SESTAT
- Some thoughts

What is Data Harmonization?

- It is the process of mapping and synchronizing data derived from multiple sources into a coherent data file for analysis.

Why Harmonize Data?

- Enhance data utility and quality
- Increase micro and macro level estimation capacity
- Minimize response burden and costs of survey programs
- Create further opportunities for comparative analysis

Challenges for Data Harmonization

- **Disparate data**
  - Data linkage
  - Consistency/comparability
- **Variation of data items measured**
  - Survey questions
  - Administrative records
  - Transactional records
- **Cost benefit analysis: value and expense**
- **Representativeness**
  - Missing data
  - Bias
- **No universal data quality measures**
  - Fit for use

Overview of Scientists and Engineers Statistical Data System (SESTAT)

- U.S. National Science Foundation's System of Workforce Surveys
- Three surveys:
  - National Survey of Recent College Graduates (NSRCG)
  - Survey of Doctorate Recipients (SDR)
  - National Survey of College Graduates (NSCG)
- A comprehensive and integrated system of information on the employment, education, and demographic characteristics of scientists and engineers in the U.S.
- For survey details, https://www.nsf.gov/statistics/sestat/

SESTAT Component Surveys

- NSRCG: Individuals who received a bachelor's or master's degree in an S&E field during the previous two academic years from a U.S. institution
- SDR: Individuals who received a doctorate in an S&E field from a U.S. institution
- NSCG
  - All other S&E bachelor's or master's degrees
  - Individuals who received a doctorate in an S&E field from a foreign institution
  - Non-S&E degrees but working in S&E occupations as of the survey reference date
  - Covers about 90 percent of the SESTAT population

Benefits to Harmonizing Multiple Survey Data:

- Broadens cross-sectional coverage
  - Creates nationally representative sample for the entire S&E population
- Increases temporal coverage
  - Produces trend estimates over time
- Adds a longitudinal component
  - Allows career tracking of S&E individuals over time

Process for Data Harmonization

Coordination and Standardization

- Survey questions
- Variable naming conventions, for example:
  - Variable: working for pay as of the survey reference date
  - NSCG and SDR: A1
  - NSRCG B1
  - SESTAT variable name WRKG
- Variable formats
- Harmonize variables over time when choice of categories for a question changed over time
- Coordination of sample designs and data collection protocols

Process for Data Harmonization (Cont'd)

Guidelines

- Standardize **coding** schemes for education and employer institutions; degree level; field of degrees; occupations, etc.
- Apply consistent **editing** rules across three surveys
- Guidelines for **response rate** calculations

Process for Data Harmonization (Cont'd)

Coordinating Statistical Data Processing and Estimation Procedures

- Imputation guidelines for a set of auxiliary variables, imputation order, and imputation methods
- Weighting procedures on a set of auxiliary variables, response propensity models, and adjustments
- Variance estimation methods

Process for Data Harmonization (Cont'd)

- Data Products
  - Multiplicity adjustments
  - Weights for cross-sectional and longitudinal analyses
  - Variance estimation for cross-sectional, trend, and longitudinal analyses
- Documentation
  - Metadata
  - Methodology reports
  - Micro data

Discussion

- SESTAT example shows that integration of multiple survey data for a federal statistics program requires major resources in time, $$, and staff.
- Consider data harmonization as a sparate Program
  - If a federal statistics program involves the integration of multiple data, the entire data life cycle needs to be planned and established to include the necessary harmonization and integration activities.

Thank You!

NORC
at the UNIVERSITY of CHICAGO

insight for informed decisions™