Westat

Improving Lives Through Research®

Discussion of "Leveraging External Data Sources to Improve Federal Government Surveys"

Jean Opsomer

Westat

August 6, 2024

Presentations

- **Stephanie Zimmer**: Age-Eligibility Oversampling to Reduce Screening Costs in a Multimode Survey (NSFG)
- **Jay Clark**: Enhancing Weighting in NHANES with External Data
- **Wan-Ying Chang**: Improving Survey Efficiency with Linked Data, the SDR Story

General Comments

- Decreasing response rates and rising costs are existential concerns for surveys as a primary data collection approach
- Survey organizations try to address this *directly* by:
  - more sophisticated methods to find and contact sampled units, including multi-mode and multi-frame methods
  - improve questionnaire design for access, legibility, etc
  - use of incentives (monetary or other)
  - reducing survey burden, by limiting questions to essential ones or splitting questionnaire
  - responsive/adaptive design in data collection
- Alternative option: sample from "preselected" respondents (e.g. representative panels)
  - greatly simplifies recruitment for a given survey, but doesn't solve the fundamental problem of nonresponse
  - introduces new issues of panel conditioning, straightlining, etc
- In general, it is critical to continue researching survey methods, survey data collection and survey sampling methods to maximize "yield" of surveys and control nonresponse bias

General Comments (2)

- Survey organizations can try to address these concerns *indirectly* by:
  - more efficiently using and/or creating frame information to target sample
  - use auxiliary variables to account for nonresponse in estimation
  - use auxiliary variables to improve precision and representativeness (calibration)
  - more sophisticated estimation approaches (e.g. double-robust methods)
  - append data from complementary non-survey sources
- The three presentations are excellent examples of some of these direct and indirect approaches in important U.S. government surveys

1. Zimmer

- Investigation of usefulness of imperfect frame information to improve efficiency of data collection
  - here: "vendor variables" in ABS frames
  - evaluated on "gold standard" external survey
  - impressive results for predicting eligibility
- What about the use of the vendor variables directly, esp. when you don't have a validation sample?
- What about other eligibility criteria besides age?
- Wishlist: work with vendors to assess quality and coverage of their appended variables, so that they can be improved over time?

2. Clark

- Example of how to deal with nonresponse after sample selection and data collection are complete
- In general population surveys, limited information available on the frame for use in modeling response mechanism and calibration
  - frame = list of postal delivery addresses (plus vendor variables)
  - controls = ACS/census info on households and individuals
- Geography and demographics are insufficient nonresponse adjustments, so important to look for additional variables. Here:
  - visual observation of DU (but need to be careful)
  - new modeled SDOH variables to spatially link to the DU
  - use of multi-phase collected variables to adjust "downstream" phases
- Income calibration is clearly also a nonresponse adjustment, even though it is used at the calibration step. Was that a conscious choice and if so, why?
- How do you "calibrate" qualitative information like visual observation?

3. Chang

- Example of getting data from other sources to supplement survey data, including
  - evaluation of measurement/recall errors of survey
  - evaluation of impact of linkage errors
  - potential for reducing respondent burden
- Major advantage: ability to obtain data that cannot be obtained through a survey (e.g. bibliometrics data, income/tax data, shopping record, medical records, etc)
- Challenges:
  - need to develop linkage methodology and measures of linkage error
  - assessment of data quality, since these data are often "found"
- Results show a high level of informative non-linking, so is it possible to create representative estimates?
- Can the propensity-to-link be modeled based on survey variables and/or external sources?

Final thoughts

- Datasets with weights as a primary way to disseminate datasets for descriptive and analytic inference remain important:
  - ensures consistency and representativeness
  - design-based framework still valid even if the "design" is only partly known
    - interpretation: conditioning on the observed data, or "modeling how you obtained the data, not the data themselves"
    - allows to separate the work of analyzing data into two component, each performed by subject-matter experts (SME): (1) model the inclusion in the dataset and (2) perform (weighted) statistical analysis to understand underlying population
- Continued challenges that need to be addressed
  - clearly communicating limitations and correct interpretation of data
  - transparency of what was adjusted for and what wasn't
  - NRBAs that include external validation whenever possible
  - inference: attempt to account for all major sources of error, including design, nonresponse, noncoverage, nonlinking, measurement error, etc

Final thoughts (2)

Contact: JeanOpsomer@westat.com