The Use of Administrative Data in the Edit and Imputation Process

Natalie Shlomo
Israel Central Bureau of Statistics
natalies@cbs.gov.il

Two main uses of administrative data in the survey process:

1. A direct statistical source for building censuses and registers, and calculating auxiliary data for size variables in pps sampling and for calculating sample weights by calibrating to known totals.

2. A secondary statistical source for improving survey processes, understanding and modeling non-response, editing and imputation.

Administrative data as a direct statistical source

Editing administrative data:

- Specific problems when editing administrative data include record linkage errors, discrepancies in the definitions of the statistical units, smaller units report periodically (i.e., revenue data).
- Selective editing indicators calculated for flagging extreme values, coding errors, inconsistencies between monthly reports and between variables in the different data sources.

Imputing on administrative data:

1. Linked sources of data provide very good covariates for imputation models, especially for imputing revenue based on wages and employees.

2. Imputation strategy based on ratio estimates within homogeneous strata, or using previous data on the same entity.

Example shown on Israel CBS business register based on linking two administrative sources:

* VAT – revenue data on businesses

* Social Security – wages and employees of businesses

Administrative data as a secondary statistical source

Improving survey processes:

Based on a population register, census or other data source (i.e., income tax file), link data to respondents and non respondents in the survey.

- Better understand underlying mechanisms of non-response by gathering information on non-respondents for improving imputation and estimation procedures.
- Edits and logical checks on respondents to improve the quality and accuracy of the data.

- For non-respondents:
  - ➢ Analysis of characteristics of non-respondents by filling in demographic, geographic and other variables from administrative sources.
  - ➢ Compute response probabilities based on logistic regression on auxiliary data to predict response status.
  - ➢ Impute within homogeneous weighting or imputation classes based on response probabilities or target variables
  - ➢ Variances can be adjusted to take into account imputed values.

- Respondents:
  - ➢Build record combinations from different data sources, calculate total variable field scores based on the sum of individual field scores equal to the probability that the value of the variable is the correct one.
  - ➢Choose record with lowest number of edit failures and highest total variable field score.
  - ➢Impute first missing values and target variables, and correct inconsistent records that failed edits, using appropriate imputation method and taking into account auxiliary data.

Summary

1. The use of administrative data in survey processing can improve the quality of the statistical output, provided that the administrative data is itself of good quality and its limitations are known and can be dealt with.

2. By incorporating different sources of administrative data, edit failures can automatically be corrected, data can be obtained for non-respondents and target variables can be imputed based on more complete covariates and better imputation models.

Thank You