“Discussion” of talk by Trivellore Raghunathan

William R. Bell

U.S. Census Bureau
Research and Methodology Directorate

Third FCSM Workshop on Quality of Blended Data
February 26, 2018

Small Area Estimation: Fay-Herriot Model (1979)

$$\begin{aligned} y_i &= Y_i + e_i \\ &= (\alpha + \mathbf{x}_i'\beta + u_i) + e_i \end{aligned}$$

- Y_i = population target for area i
- y_i = direct survey estimate of Y_i
- e_i = sampling errors $\sim$ ind. $N(0, v_i)$ with v_i estimated
- $\mathbf{x}_i$ = vector of regression variables for area i
- β = vector of regression parameters (α = intercept)
- u_i = area i random effect (model error) $\sim$ $i.i.d.$ $N(0, \sigma_u^2)$, and independent of e_i.

Model Fitting and Prediction

Model fitting by REML or via Bayesian treatment

- calibrates covariates to predict Y_i via $\hat{\alpha} + \mathbf{x}_i'\hat{\beta}$

Best Linear Unbiased Prediction (BLUP)

- Given values for σ_u^2 and the v_i:

$$\hat{Y}_i = h_i y_i + (1 - h_i)(\hat{\alpha} + \mathbf{x}_i'\hat{\beta})$$

$$\text{where} \qquad h_i = \frac{\sigma_u^2}{\sigma_u^2 + v_i} \propto \frac{1}{v_i} = \frac{1}{\mathrm{var}(Y_i - y_i)}$$

$$1 - h_i = \frac{v_i}{\sigma_u^2 + v_i} \propto \frac{1}{\sigma_u^2} = \frac{1}{\mathrm{var}(Y_i - \mathbf{x}_i'\beta)}$$

United States Census Bureau

What is needed to make this work?

- define the pop characteristic of interest, Y_i – the "target"
- unbiased survey estimate y_i of Y_i
  - often, the target is defined as what y_i is estimating ($Y_i \equiv E(y_i)$)
  - also need decent estimates of sampling error variances, $v_i = \text{var}(e_i)$
- covariate(s) x_i with a consistent (linear) relation to Y_i

$$Y_i = \alpha + \beta x_i + u_i$$

  - note that x_i need not actually estimate Y_i

What is needed to make this work (continued)?

- If another survey estimate y_{2i} is a candidate as a covariate, use the bivariate FH model instead

$$y_{1i} = Y_{1i} + e_{1i} = (\mathbf{x}'_{1i}\beta_1 + u_{1i}) + e_{1i}$$

$$y_{2i} = Y_{2i} + e_{2i} = (\mathbf{x}'_{2i}\beta_2 + u_{2i}) + e_{2i}$$

where $\mathrm{Var}(u_{1i}) = \sigma_1^2$, $\mathrm{Var}(u_{2i}) = \sigma_2^2$, $\mathrm{Var}(e_{1i}) = v_{1i}$, and $\mathrm{Var}(e_{2i}) = v_{2i}$, and we make the same sort of assumptions as before. To this we add

$$\mathrm{Cov}(e_{1i}, e_{2i}) = v_{12,i} \text{ (or 0)} \qquad \mathrm{Cov}(u_{1i}, u_{2i}) = \sigma_{12}.$$

Note that $Y_{1i} \neq Y_{2i}$.

United States Census Bureau

What does this have in common with what Raghu talked about?

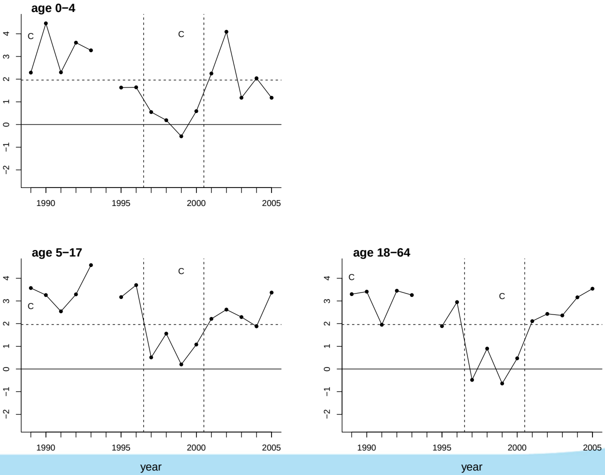
1. Need a data source that defines the estimation target Y_i.
   - Role of direct survey estimate y_i in SAE.
   - Appears to be NHANES data, or NHANES + MCBS, for Raghu's project.
2. Need covariates with a consistent relation to Y_i.
   - SAIPE uses covariates related to poverty obtained from tabulations of tax data, SNAP participants data.
   - Raghu has MCBS data and various other sources.
3. Model fitting and prediction calibrates the covariates to predict Y_i.
   - For Raghu's project, prediction involves multiple imputation.

The general elements should be common to other efforts to combine data sources to produce estimates.

What problems can arise for this general approach?

- No data source is suitable for defining the population target of interest. *(All data sources are substantially biased with respect to the desired target.)*
  - SAE reduces variances of survey estimates; won't address bias problems in y_i.
- Different relations between target Y_i and covariates x_i across observations i.
  - x_i may not be consistently defined or measured across observations i.
  - x_i may be unavailable for some observations i.
  - SAIPE example: free and reduced price lunch data
  - SAIPE example: effect of welfare reform on SNAP data
- Poor estimates of sampling variances v_i of y_i

SAIPE state poverty rate models (CPS data) t-stats for the SNAP participation rate coefficient



Some thoughts on transparency when combining data sources

Ideal (I suppose)

1. Release all data sources used
2. Release software used
3. Thoroughly document estimation and prediction methods

Obstacles to achieving ideal transparency

Data obstacles

- Confidential data sources
  - Does it help to release the non-confidential sources?
  - Are PUMS helpful when full data set cannot be released?
- Direct survey estimates that do not meet publication standards (samples too small, std errors too high, etc.). What are the options?
  - Release those direct estimates that meet publication standards; suppress those that don't.
  - Waive/relax standards and release all direct estimates, noting they have high std errors?
    - Note that small samples yield imprecise survey estimates (high true std errors) *and* imprecise estimates of std errors (some will be significantly too low)

United States Census Bureau

Obstacles to achieving ideal transparency

Software obstacles?

1. Could confidentiality of software ever be a problem?
2. In some cases substantial parts of software could be devoted to installation specific I/O which would be irrelevant for outsiders.

Some thoughts on transparency when combining data sources

Documentation of methodology

1. Who is the audience – statisticians or data users?
2. Two-tiered approach
   - general documentation for data users
   - links to more detailed technical documentation
3. Documentation typically also has an internal audience.

United States™ Census Bureau

Some thoughts on transparency when combining data sources

Time and effort required to achieve transparency

1. Why write detailed technical documentation for data users if none will read it?
2. Perhaps provide some documentation on request.

United States Census Bureau

Disclaimer

Any views expressed here are those of the author and not necessarily those of the U.S. Census Bureau.