Borrowing strength from larger surveys to improve related estimates from smaller surveys using bivariate small area estimation models

Carolina Franco and William R. Bell

U.S. Census Bureau
Center for Statistical Research and Methodology

June 5, 2019

Introduction

- Investigate the potential of borrowing strength from larger surveys via bivariate small area estimation models through three illustrative applications.
- The quantities measured by the two surveys must be related, but **not necessarily the same**
- Ripe for implementation for U.S. applications using estimates from the American Community Survey (ACS), the largest US household survey, to improve other survey estimates
- Very simple!
- No covariates from auxiliary information needed!
- Huge reductions in variances!

Three US surveys

- **American Community Survey**
  - Samples approx. 3.5 million addresses each year.
  - Many topics: demographic, income, health insurance, housing, disabilities, occupations, employment, education, etc
  - Produces annual estimates based on 1 or 5 years of data.
- **National Health Interview Survey (NHIS)**
  - About 97,000 persons in sample for 2016 Early Release (ER) estimates.
  - Questions about a broad range of health topics through personal household interviews.
- **Survey of Income and Program Participation (SIPP) Disability Module**
  - Approx. 37,000 households and 70,000 persons in 2008 panel.
  - Detailed questions about disability.

Three applications

1. **NHIS estimates of US state uninsured rates**.
   ACS variable: Previous year's estimate of US state uninsured rates (timing, questions asked and the mode of survey delivery and design also differ).
2. **SIPP estimates of US state disability rates**.
   ACS variable: Estimate of state disability rates (types of disabilities and the time frames differ).
3. **ACS 1-yrcounty estimates (of anything! Take county rates of children in poverty to illustrate)**
   2nd variable: Previous ACS 5-yr estimates (larger sample size, but less current).

United States Census Bureau

Univariate Gaussian model

- For m small areas:

$$\begin{aligned} y_i &= Y_i + e_i \qquad i = 1, \ldots, m \\ Y_i &= \mu + u_i \end{aligned}$$

- Y_i is the population characteristic of interest for area i.
- y_i is the direct survey estimate of Y_i.
- e_i is the sampling error in y_i, generally assumed to be $N(0, v_i)$, independent with v_i known.
- u_i is the area i random effect, usually assumed to be *i.i.d.* $N(0, \sigma_u^2)$ and independent of the e_i
- Precedes Fay-Herriot: Stein (1956), Carter and Rolph (1974)

United States Census Bureau

Prediction in univariate Gaussian model

- Best predictor of Y_i (μ and σ_u^2 known):

$$\hat{Y}_i = (1 - \gamma_i) y_i + \gamma_i \mu$$

where

$$\gamma_i = \frac{v_i}{v_i + \sigma_u^2}$$

- Shrinkage to μ
- Smaller sampling variances imply more weight is placed on y_i.
- Parameters are not known: hierarchical Bayes or empirical Bayes approach.

United States Census Bureau

Bivariate Gaussian model

$$y_{1i} = Y_{1i} + e_{1i} = (\mu_1 + u_{1i}) + e_{1i}, \quad i = 1, \ldots, m.$$
$$y_{2i} = Y_{2i} + e_{2i} = (\mu_2 + u_{2i}) + e_{2i}$$

$$\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix} \overset{i.i.d}{\sim} N(0, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

$$\begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix} \overset{i.i.d}{\sim} N(0, \mathbf{V}_i), \quad \mathbf{V}_i = \begin{bmatrix} v_{i11} & 0 \\ 0 & v_{i22} \end{bmatrix}$$

- y_{1i} is direct estimate of characteristic of interest, y_{2i} is direct estimate from another survey of related characteristic
- We could have instead included y_{2i} as a covariate, but this would ignore sampling error! (see Bell, Chung, Datta, Franco, 2019)

Prediction when model parameters are known

In matrix notation $\mathbf{y}_i = (\mathbf{Y}_i) + \mathbf{e}_i = (\boldsymbol{\mu} + \mathbf{u}_i) + \mathbf{e}_i$

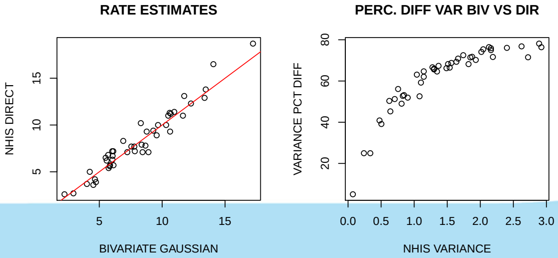
- $\hat{\mathbf{Y}}_i^{BP} = E(\mathbf{Y}_i|\mathbf{y}_i) = \boldsymbol{\mu} + \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \mathbf{V}_i)^{-1}(\mathbf{y}_i - \boldsymbol{\mu})$
- $MSE(\hat{\mathbf{Y}}_i^{BP}) = Var(\mathbf{Y}_i|\mathbf{y}_i) = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \mathbf{V}_i)^{-1}\boldsymbol{\Sigma}$
- We are interested in predicting Y_{1i} only, not Y_{2i}

In what follows, all models are given a hierarchical Bayes treatment (using JAGS) with diffuse priors

Application I: 2013 Uninsured rates for US States from NHIS borrowing from ACS

y_{1i}= NHIS estimate, 2016, y_{2i}= ACS estimate, 2015
Smoothing of NHIS direct sampling variances is applied.
Only 43 direct estimates published due to accuracy concerns.

Decrease in variance from the direct estimate of up to 78%, with a median decrease of 66%!!



MSE Decomposition when parameters are known

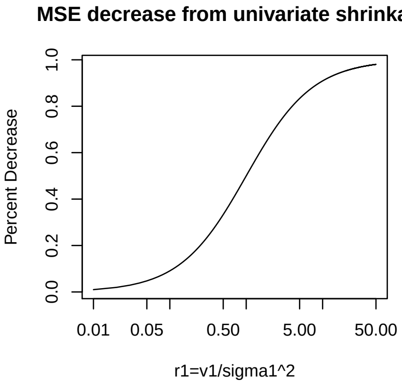
- Let $r_{1i} = \frac{v_{i1}}{\sigma_1^2}$, $r_{2i} = \frac{v_{i2}}{\sigma_2^2}$ and $\rho = corr(u_{1i}, u_{2i}) = \sigma_{12}/\sigma_1\sigma_2$

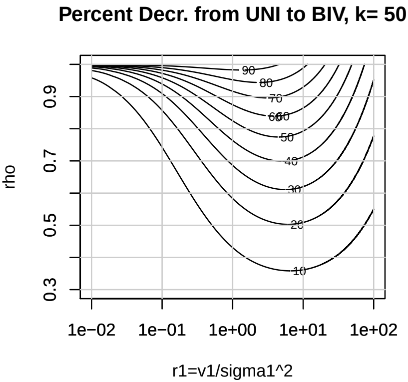
  % MSE/var Decrease Bivariate vs. Direct:

$$\underbrace{\left[\frac{r_{1i}}{1+r_{1i}}\right]}_{\text{\% Decr. UNI vs. DIR}} \times \left[1+\frac{1}{r_{1i}}\underbrace{\left(\frac{r_{1i}\rho^2}{(1+r_{1i})(1+r_{2i})-\rho^2}\right)}_{\text{\% Decr. BIV vs. UNI}}\right]$$

- Define $k_i = r_{1i}/r_{2i}$
- Note that when $\sigma_1^2 = \sigma_2^2$, $k_i = v_{1i}/v_{2i}$, so k_i can be thought of as a measure of relative accuracy or relative size of the surveys.

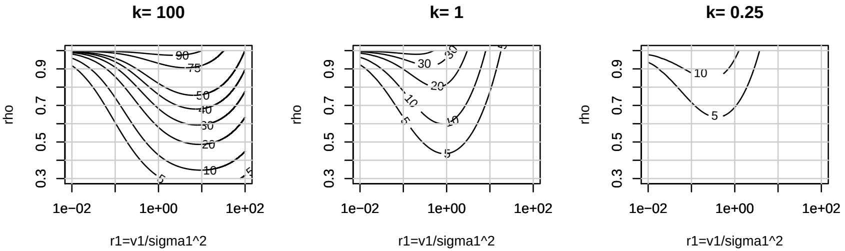
Plots of components of MSE decreases





Effect of changes in k_i on % Decrease BIV vs. UNI

- As k decreases, all else fixed, MSE reduction decreases.
- Suggests limited benefits from borrowing strength from *smaller* surveys.



Application I variance decrease decomposed

- $\hat{\rho}$= **.97**,
- r_{1i} max 0.25, k_i from 5 to 352, median 38

	percentage variance reductions				
model	**mean**	**1st q.**	**median**	**3rd q.**	**max**
univariate Gaussian	11	7	11	15	19
bivariate Gaussian	62	53	66	72	78

Table: Percent variance reductions from direct estimates for the univariate and bivariate models

May be able to publish more estimates using bivariate model, due to lowered variance

Application II: 2010 SIPP total disability

- y_{1i}= SIPP estimate y_{2i}= ACS estimate
- Smoothing of SIPP Direct Variances is Applied
- $\hat{\rho}$= **.96**
- r_{1i} max 3.75, third quartille 0.5; k_i median 32, max 180.

	percentage variance reductions				
model	**mean**	**1st q.**	**median**	**3rd q.**	**max**
univariate Gaussian	22	8	20	32	66
bivariate Gaussian	41	21	39	57	85

Table: Percent variance reductions from direct estimates for the univariate and bivariate models.

United States Census Bureau

Application III: ACS 1-yr estimates borrow from previous ACS 5-yr estimates

- 2012 county rates of children in poverty used as illustration (good regressors are available, but excluded here).
- y_{1i}= 2012 ACS 1yr est., y_{2i}= 2007-2011 ACS 5yr est.
- $\hat{\rho} = \mathbf{0.94}$, r_{1i} median 0.5, k_i median 4.

model	percentage variance reductions				
	mean	**1st q.**	**median**	**3rd q.**	**95 p.**
univariate Gaussian	33	17	32	47	65
bivariate Gaussian	62	54	67	74	81

Table: Percent variance reductions from direct estimates for the univariate and bivariate models

Other bivariate models

- Because applications are proportions, also fit univariate and bivariate versions of two other models
  - Binomial Logit Normal Model: Binomial assumpton for sampling model; logit transformation for linking model. Modification for design effect (Franco and Belll, 2013,2015)
  - Unmatched Sampling and Linking Model (Yu and Rao 2012): Gaussian Assumption for sampling model; logit tansformation for linking model
- Results on % differences were similar to that of the Gaussian models
- Predictions are similar accross models, but prediction standard errors differ
- Began working on model comparison, but difficult question

Concluding remarks

- Great variance decreases from borrowing strength from ACS to improve estimates from smaller surveys, provided ρ is high!
- Presumably not so great decreases when a larger survey borrows strength from a smaller one.
- Extremely simple method, easy to apply
- Future research: model comparison

Disclaimers

All U.S. Census Bureau disclosure avoidance guidelines have been followed and estimates have been approved for release by the U.S. Census Bureau Disclosure Review Board. DRB approval number: CBDRB-FY19-357.

The views expressed in this presentation are those of the author(s) and not the U.S. Census Bureau