Using American Community Survey Data to Improve Estimates from Smaller U.S. Surveys through Bivariate Small Area Estimation Models

William R. Bell, U.S. Census Bureau
Carolina Franco, NORC at the University of Chicago

November 3, 2021

Introduction

- We investigate the potential for borrowing strength from ACS estimates via bivariate small area estimation models to reduce variances of estimates from smaller U.S. surveys.

Introduction

- We investigate the potential for borrowing strength from ACS estimates via bivariate small area estimation models to reduce variances of estimates from smaller U.S. surveys.
- The population quantities measured by ACS and the smaller survey must be related, but **are not assumed to be the same.**

Introduction

- We investigate the potential for borrowing strength from ACS estimates via bivariate small area estimation models to reduce variances of estimates from smaller U.S. surveys.
- The population quantities measured by ACS and the smaller survey must be related, but **are not assumed to be the same.**
- The bivariate modeling is very simple!

Introduction

- We investigate the potential for borrowing strength from ACS estimates via bivariate small area estimation models to reduce variances of estimates from smaller U.S. surveys.
- The population quantities measured by ACS and the smaller survey must be related, but **are not assumed to be the same.**
- The bivariate modeling is very simple!
- No covariates from auxiliary information are used!

Introduction

- We investigate the potential for borrowing strength from ACS estimates via bivariate small area estimation models to reduce variances of estimates from smaller U.S. surveys.
- The population quantities measured by ACS and the smaller survey must be related, but **are not assumed to be the same.**
- The bivariate modeling is very simple!
- No covariates from auxiliary information are used!
- We illustrate huge reductions in variances for our examples!

Introduction

- We investigate the potential for borrowing strength from ACS estimates via bivariate small area estimation models to reduce variances of estimates from smaller U.S. surveys.
- The population quantities measured by ACS and the smaller survey must be related, but **are not assumed to be the same.**
- The bivariate modeling is very simple!
- No covariates from auxiliary information are used!
- We illustrate huge reductions in variances for our examples!
- Note: Examples use data prior to the COVID-19 pandemic.

Three U.S. surveys

- **American Community Survey (ACS)**
  - Samples approximately 3.5 million addresses each year.
  - Many topics: demographics, income, health insurance, housing, disabilities, occupations, employment, education, etc
  - Produces 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**
  - Approximately 37,000 households and 70,000 persons in 2008 panel.

Three applications

1. **NHIS estimates of U.S. state uninsured rates**.
   ACS variable: Previous 1-year estimates of U.S. state uninsured rates (timing, questions asked, and the mode of survey delivery and design differ from NHIS).
   - In a CNSTAT workshop report, Kenney and Lynch (2010) noted a consensus that "the NHIS produces the most valid [insurance] coverage estimates," but also that "the [NHIS] sample size is too small to produce precise annual state and substate estimates for most states"
2. **SIPP estimates of U.S. state disability rates**.
   ACS variable: 1-year estimate of state disability rates
   - 2008 SIPP panel asked more detailed questions about disability than did ACS.
3. **ACS 1-year county estimates of poverty rates of school-aged children**
   - 2nd variable: Previous ACS 5-year estimates – these have a larger sample size, but are less current.

Bivariate Gaussian model

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

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

$$\begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix} \overset{ind}{\sim} N(0, \mathbf{V}_i), \quad \mathbf{V}_i = \begin{bmatrix} v_{i,11} & 0 \\ 0 & v_{i,22} \end{bmatrix}$$

- y_{1i} is the direct estimate of the population characteristic from the smaller survey, and y_{2i} is the related ACS direct estimate.
- Naive approach: include y_{2i} as a covariate in the model for Y_{1i}, but this ignores its sampling error! See Bell, Chung, Datta, Franco (2019)
- Parameterize model via $\mu_1, \mu_2, \sigma_{11}, \sigma_{22}$ and $\rho = \text{corr}(u_{1i}, u_{2i})$

Prediction when model parameters are known

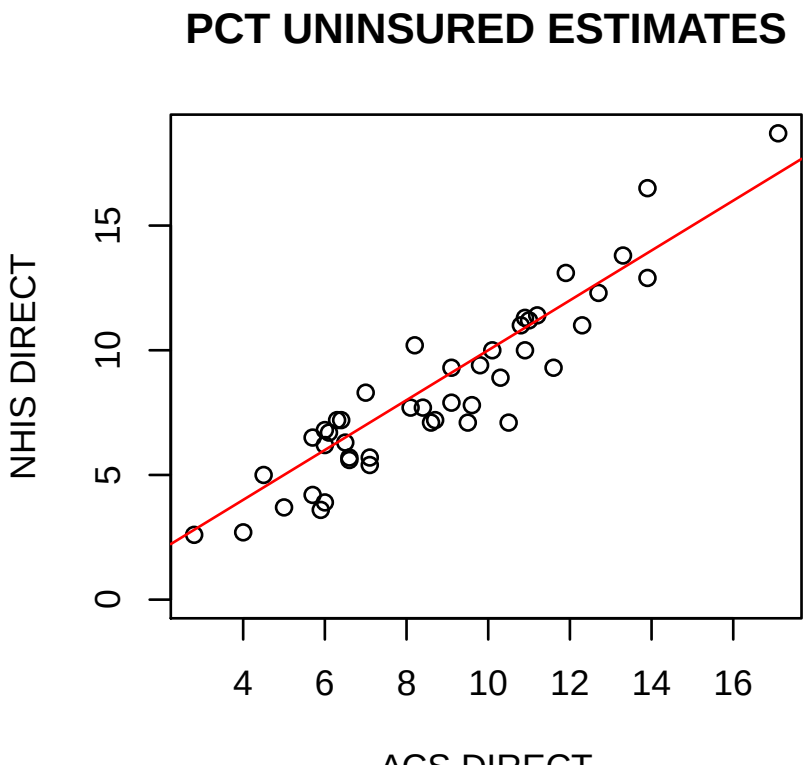
In matrix notation $\mathbf{y}_i = (\mathbf{Y}_i) + \mathbf{e}_i = (\boldsymbol{\mu} + \mathbf{u}_i) + \mathbf{e}_i$

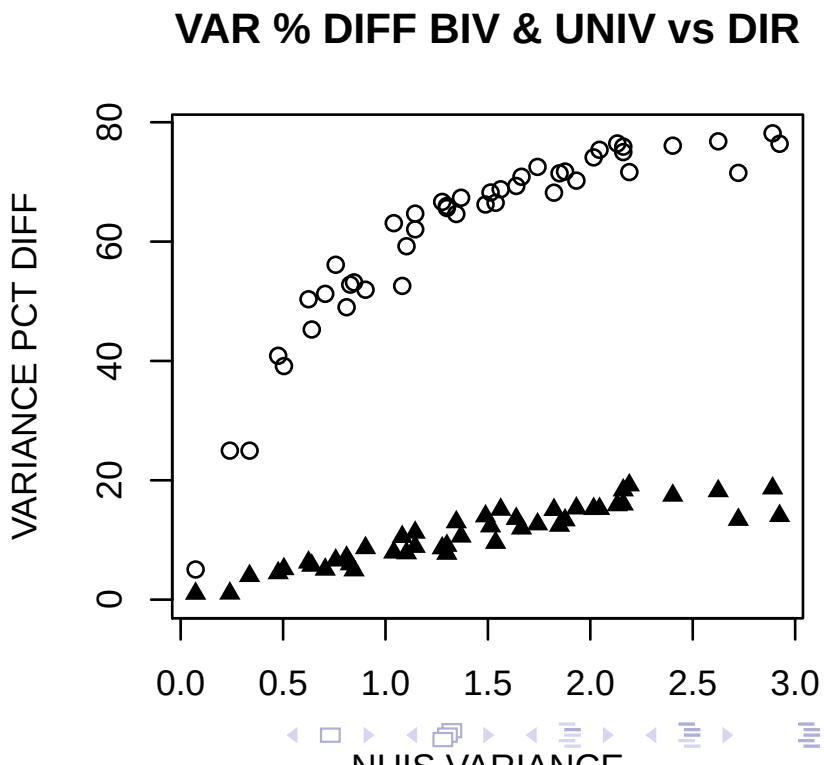
- $\hat{\mathbf{Y}}_i^{BP} = E(\mathbf{Y}_i|\mathbf{y}_i) = \boldsymbol{\mu} + \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \mathbf{V}_i)^{-1}(\mathbf{y}_i - \boldsymbol{\mu})$
- $MSE(\hat{\mathbf{Y}}_i^{BP}) = Var(\mathbf{Y}_i|\mathbf{y}_i) = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \mathbf{V}_i)^{-1}\boldsymbol{\Sigma}$
- 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 on the parameters.

Application 1: NHIS estimates of U.S. state uninsured rates

- y_{1i} = 2016 NHIS estimate, y_{2i} = 2015 ACS estimate
- Smoothing of NHIS direct sampling variances is applied
- Only 45 direct NHIS state estimates were published due to "considerations of sample size and precision."





Application 1: NHIS variance decreases from modeling

$\hat{\rho} = E(\rho | \{y_{1i}\}, \{y_{2i}\}) = .97$

model	percentage variance reductions				
	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

The results may suggest that it might be possible to publish estimates for more states using a bivariate model, due to the substantially lowered variances.

Application 2: 2010 SIPP state total disability estimates

- y_{1i} = SIPP estimate, y_{2i} = ACS estimate
- Smoothing of SIPP direct variances is applied
- $\hat{\rho} = .96$

model	percentage variance reductions				
	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.

Application 3: ACS 1-year estimates of county poverty rates for school-age children

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

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

Alternative bivariate models

- Because our applications involve proportions, we also fit univariate and bivariate versions of two other models
  - Binomial logit normal model (Franco and Bell 2015):
    - Binomial assumption for sampling model using effective sample size and effective number of successes
    - logit transformation of true proportions for linking model
  - Unmatched sampling and linking model (You and Rao 2002):
    - Gaussian assumption for sampling model
    - logit tansformation of true proportions for linking model
- These alternatives show large % reductions of variances similar, overall, to the Gaussian model.
- Point predictions (posterior means) are similar accross models, but posterior variances differ.
- Franco and Bell (2021) propose a model selection criterion, but the results are not definitive. More research is needed.

Concluding remarks

- Large variance decreases are possible by using bivariate models to borrow strength from ACS estimates to improve estimates from smaller U.S. surveys, provided ρ is high!
  - Given the wide range of population characteristics estimated by ACS, many smaller U.S. surveys estimate some closely related (or ostensibly the same) characteristics for which this borrowing of information could achieve substantial variance reductions.
- Generally little or no benefits can be realized by trying to borrow information from a smaller survey to improve the estimates from a much larger survey.
- Bivariate model is simple, easy to apply
- Further research is needed on model comparison statistics

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 authors and not of the U.S. Census Bureau or the National Opinion Research Center.

References

- Bell, William R., Chung, Hee-Cheol, Datta, Gauri S., and Franco, Carolina (2019), "Measurement Error in Small Area Estimation: Functional Versus Structural Versus Naive Models," Survey Methodology, 45, 61–80
- Franco, Carolina and Bell, William R. (2015), "Borrowing Information Over Time in Binomial/Logit Normal Models for Small Area Estimation," Statistics in Transition (new series) and Survey Methodology, joint issue on Small Area Estimation, 16, 563–584, available at http://stat.gov.pl/en/sit-en/issues-and-articles-sit/previous-issues/volume-16-number-4-december-2015/
- Franco, Carolina and Bell, William R. (2021). "Using American Community Survey Data to Improve Estimates from Smaller U. S. Surveys through Bivariate Small Area Estimation Models" Journal of Survey Statistics and Methodology, to appear.

References

- Kenney, Genevieve and Lynch, Victoria (2010). Monitoring Children's Health Insurance Coverage Under CHIPRA Using Federal Surveys, chapter 8 in Databases for Estimating Health Insurance Coverage for Children: A Workshop Summary. Thomas J. Plewes, rapporteur. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
- You, Yong and Rao, J. N. K. (2002), "Small Area Estimation Using Unmatched Sampling and Linking Models," *The Canadian Journal of Statistics*, **30**, 3–15.