Applying Bivariate Binomial-Logit Normal Models to Small Area Estimation

Carolina Franco and William R. Bell

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

SAE 2013, Bangkok, Thailand
September 2, 2013

United States Census Bureau

Disclaimer

This presentation and the paper are released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Introduction

- The U.S. Census Bureau's SAIPE (Small Area Income and Poverty Estimates) program estimates poverty for various age groups for states, counties, and school districts of the U.S.
- Our focus: poverty estimates of school-aged (5-17) children for counties.
- Inference is currently based on 1-year data from the American Community Survey (ACS), covariates from administrative records and a Census long form 2000 estimate.

The American Community Survey (ACS)

- Approximately 3 million addresses per year since 2005.
- Questions: demographic, income, disabilities, health insurance, etc.
- Sampling design: stratification, systematic sampling, clustering of persons, etc.
- Estimation procedure: basic weights undergo several adjustments to adjust for nonresponse, to calibrate to population controls, etc.
- Supplanted the census "long form", which sampled about 1/6 of the population every 10 years (last long-form in 2000)
- Publishes 1-year, 3-year, and 5-year estimates for billions of estimands each year.

United States Census Bureau

Why Consider a Bivariate Model in this Problem?

- County SAIPE model has traditionally used a previou Census long form estimate as an important regression variable.
- Census 2000 long form data increasingly out of date.
- The ACS 5-year estimate from the years prior to the production year may be a good alternative (Huang and Bell, 2012).
- Sampling error in census county estimates currently ignored in the modelling
- Due to the smaller sample size this is less acceptable with the ACS 5-year data.
- Bivariate model can allow for both sampling errors.

United States Census Bureau

The Fay-Herriot Model (1979)

- The model for m small areas:

$$y_i = Y_i + e_i \qquad i = 1, \ldots, m \tag{1}$$
$$Y_i = \mathbf{x}_i'\beta + u_i \tag{2}$$

- 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.
- $\mathbf{x}_i$ and β are the regression variables and coefficients.

United States Census Bureau

The SAIPE 5-17 County Production Poverty Model

- The model is of the form of (1) and (2) with a logarithmic transformation.
- y_i = log of the ACS estimate of the number of persons age 5-17 in poverty for county i.
- Y_i=log of the true number of persons age 5-17 in poverty in the county.
- β and σ_u^2 are estimated by ML.
- Prediction results are translated back from the log scale using properties of the lognormal distribution.

United States Census Bureau

The SAIPE 5-17 County Production Poverty Model–Regression Variables

- log of the number of "poor child exemptions" for the county, i.e., child exemptions claimed on tax returns whose adjusted gross income falls below the official poverty threshold;
- log of the number of county SNAP benefits recipients in July of the previous year;
- log of the estimated county population age 0-17 as of July 1;
- log of the total number of child exemptions in the county claimed on tax returns; and
- log of the Census 2000 county estimate of the number of children in poverty ages 5 to 17.

United States Census Bureau

Some Issues with the Current Production Model

- For some counties with small samples, the direct ACS estimate of the number of 5-17 year-olds in poverty is zero.
- Since logs cannot be taken of these zero estimates, such counties are dropped from the model fitting.
- Using the production model, one can still produce estimates for all counties.
- Our bivariate GLMM approach, which uses a generalized variance function (GVF) to estimate the sampling variances, does not require dropping any counties from the fitting.

United States Census Bureau

A Univariate Binomial/Logit Normal Model

- Let y_i be the sampled count, n_i the sample size, and p_i be the true proportion for county i
- Univariate Binomial/Logit Normal Model:

$$y_i|p_i, n_i \sim \text{Bin}(n_i, p_i) \qquad i = 1, \ldots, m \tag{3}$$

$$\text{logit}(p_i) = \mathbf{x}_i'\beta + u_i \tag{4}$$

- $\text{logit}(p_i) = \log[p_i/(1-p_i)]$, $u_i \sim N(0, \sigma_u^2)$.
- This model does not incorporate the complex sampling features of the data!

Use of Effective Sample Sizes

- Due to the complex sampling design, we use "effective" sample sizes $\tilde{n}_i$ and sample counts $\tilde{y}_i$ based on the design effect:

$$\tilde{n}_i = \tilde{p}_i(1 - \tilde{p}_i) \Big/ \widehat{\text{Var}}(\hat{p}_i)$$

$$\tilde{y}_i = \tilde{n}_i \times \hat{p}_i$$

- $\hat{p}_i$ are the direct ACS estimates; $\tilde{p}_i$ are preliminary estimates of p_i based on $\hat{p}_i$ defined such that they cannot be zero.
- We then substitute $(\tilde{n}_i, \tilde{y}_i)$ for (n_i, y_i) in the Binomial/Logit Normal Model, rounding to the nearest integer.

United States Census Bureau

The Bivariate Binomial/Logit Normal Model

$$\tilde{y}_{1i}|p_{1i}, \tilde{n}_{1i} \sim \text{Bin}(\tilde{n}_{1i}, p_{1i}) \qquad \tilde{y}_{2i}|p_{2i}, \tilde{n}_{2i} \sim \text{Bin}(\tilde{n}_{2i}, p_{2i}) \quad (5)$$

$$\text{logit}(p_{1i}) = \mathbf{x}'_{1i}\beta_1 + u_{1i} \qquad \text{logit}(p_{2i}) = \mathbf{x}'_{2i}\beta_2 + u_{2i} \quad (6)$$

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

for $i = 1, \ldots, m$.

- Application: $(\tilde{y}_{1i}, \tilde{n}_{1i})$, $(\tilde{y}_{2i}, \tilde{n}_{2i})$ are the effective sample counts of children aged 5-17 in poverty and effective sample sizes based on the 2011 ACS 1-year and the 2006-2010 ACS 5-year estimates.

United States Census Bureau

Comments

- $\tilde{y}_{1i}$ and $\tilde{y}_{2i}$ are assumed conditionally independent (given $p_{1i}, \tilde{n}_{1i}$ and $p_{2i}, \tilde{n}_{2i}$) since the ACS samples are drawn approximately independently each year.
- Unconditionally, these are dependent due to the correlation of the random effects u_{1i} and u_{2i}.
- To avoid excluding observations from the fitting, we use a Generalized Variance Function (GVF) to generate estimates of the sampling variance even for counties that have an observed count of zero.
- We use SAS's NLMIXED for fitting the model.

United States Census Bureau

The GVF–Introduction

- Using ACS direct sampling variances $\hat{S}_i^2$ for each survey, our GVF model is:

$$E(S_i^2) = \texttt{GVF}_\texttt{i} = \gamma_0(\texttt{p}_\texttt{i}(1-\texttt{p}_\texttt{i}))^{\gamma_1}(\texttt{Rw}_\texttt{i})^{\gamma_2}. \qquad (7)$$

- $Rw_i := \frac{\sum_{j=1}^{n_i} w_{ij}^2}{(\sum_{j=1}^{n_i} w_{ij})^2}$, where w_{ij} is the weight of household j in county i, and n_i is the sample size of county i.
- Rw_i is an estimate of the inverse of the effective sample size when there is no clustering (Kish, 1987).
- Only counties with $S_i^2 \neq 0$ that meet a minimum sample size threshold are used in the fitting.
- The log of equation (7) can be fitted as a linear model.

Intial Values for the GVF and Iterative Approach

- $\tilde{p}_i = logit^{-1}(x_i\hat{\eta})$ where $\hat{\eta}$ solves the optimization problem

$$min \sum_{i=1}^{m} (\hat{p}_i - logit^{-1}(x_i\eta))^2 \quad (8)$$

- $\hat{p}_i$ are the direct ACS estimates. Note $\tilde{p}_i$ cannot be zero.
- These $\tilde{p}_i$ are used in the GVF model to estimate $\gamma_0, \gamma_1, \gamma_2$
- We then use the fitted GVF model (7) to estimate $\texttt{GVF}_\texttt{i}$ for all counties.
- We fit the bivariate binomial/logit Normal model using these $\texttt{GVF}_\texttt{i}$ for the sampling variances $\widehat{Var}(\hat{p}_i)$.
- Iterative Approach: the $\tilde{p}_i$ are updated, repeat.

United States Census Bureau

Covariates Used in Bivariate Binomial/Logit Normal Model

- logit of the proportion of child exemptions "in poverty" for the county, i.e., the number of child exemptions claimed on tax returns whose adjusted gross income falls below the poverty threshold divided by the total number of child exemptions for the county;
- logit of an adjusted version of the county "tax child filer rate," which is defined as the number of child exemptions in the county claimed on tax returns divided by the county population age 0-17.
- logit of the ratio of county SNAP benefits recipients in July of the previous year to the county population of the previous year.

United States Census Bureau

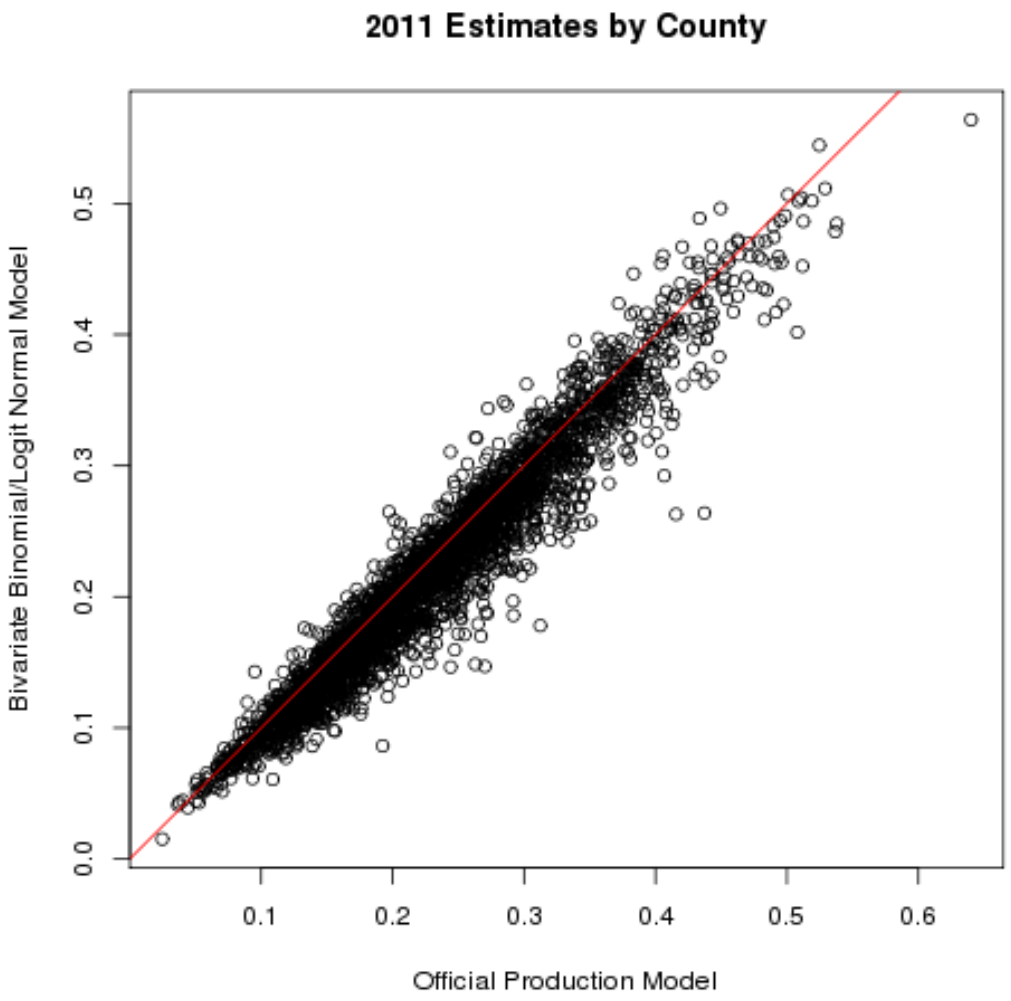
Regression Coefficients and Correlation Coefficient of Bivariate Binomial/Logit Normal Model Applied to the Year 2011

Par	Est	SE	Regression Variables (in logit scale)
β_{11}	0.73	0.03	Poverty Prop. of child exemptions 2011
β_{12}	-0.09^*	0.07	(Adjusted) county tax child filer rate 2011
β_{13}	0.30	0.02	Ratio of county SNAP recipients 2011
β_{21}	0.79	0.02	Poverty Prop. of child exemptions 2008
β_{22}	-0.20	0.02	(Adjusted) county tax child filer rate 2008
β_{23}	0.22	0.01	Ratio of county SNAP recipients 2008

Par	Est	SE	Description
ρ	0.3360	0.05	Correlation Coefficient of u_{1i} and u_{2i}

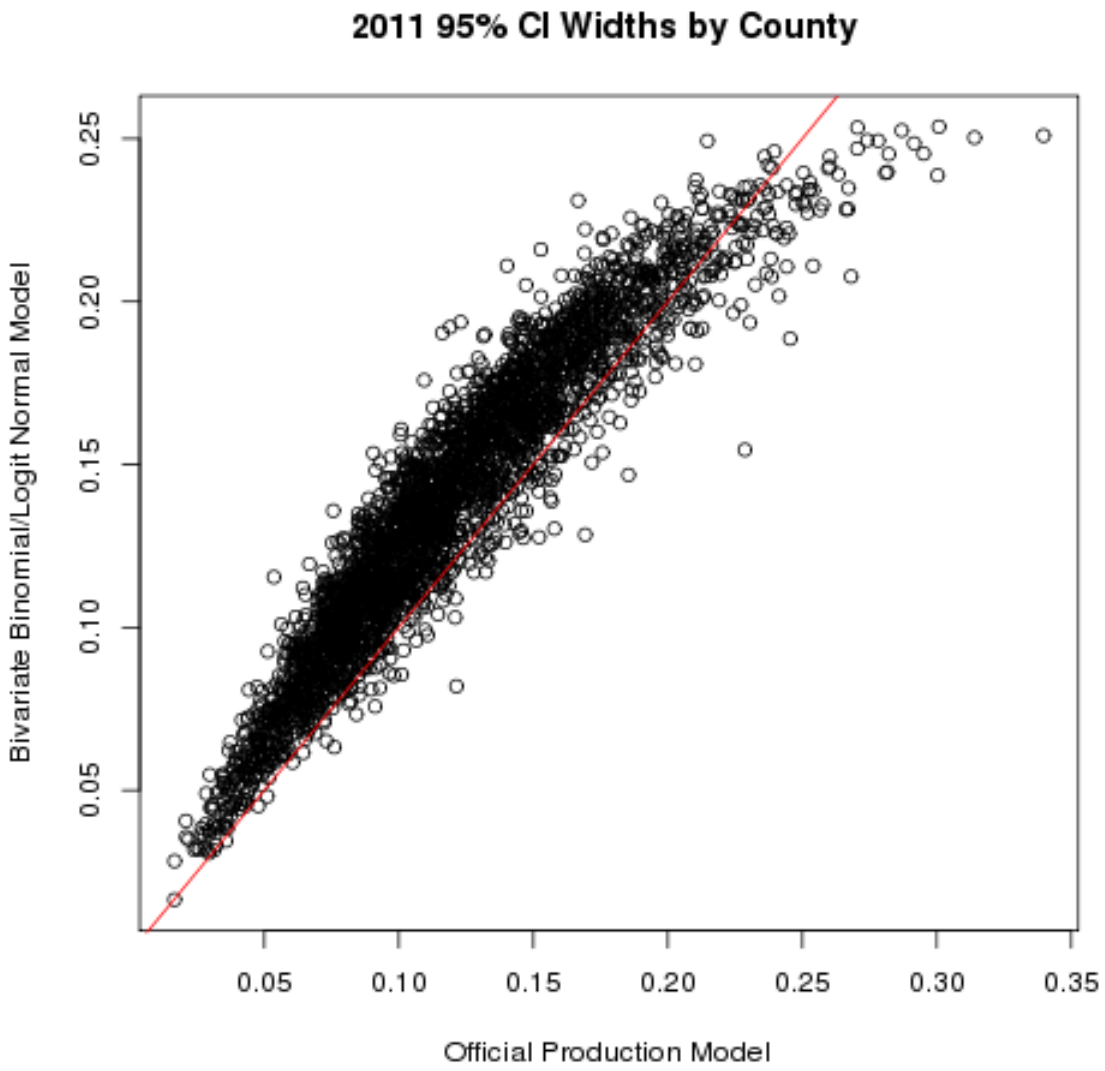
United States Census Bureau

Comparison of Bivariate Estimates with Estimates from Production Model



United States Census Bureau

Comparison of Prediction Interval Widths



Discussion

- The estimates of the Bivariate Binomial/Logit Normal Model are broadly similar to those of the current production model.
- The corresponding confidence intervals tend to be a little wider.
- Further investigation and comparisons to other alternative models are needed.

Future Research

Alternative Models:

- *Bivariate Log rate model*: Use a bivariate version of the linear Fay-Herriot model where y_{1i} is the log of the ACS estimated 5-17 poverty rate for county i, and y_{2i} is the log of the prior ACS 5-year estimate of the 5-17 poverty rate for county i.
- *Alternative link functions in the Bivariate GLMM model*: Substitute a different link function for the logit. Common alternatives include the probit and the log-log (Agresti 1990).
- *Unmatched sampling and linking models (You and Rao 2002)*: Replace the Binomial assumption with an assumption of normality.

Future Research

- *Nonlinear regression in the Fay-Herriot model*:Add the random effect directly to the model for the true proportions:

$$p_{1i} = \frac{\exp(\mathbf{x}'_{1i}\beta_1)}{1+\exp(\mathbf{x}'_{1i}\beta_1)} + u_{1i} \qquad p_{2i} = \frac{\exp(\mathbf{x}'_{1i}\beta_1)}{1+\exp(\mathbf{x}'_{1i}\beta_1)} + u_{2i}$$

- *Autoregressive Models*: Extend the Binomial-Logit Normal Model to model 1-year estimates for multiple years using a first-order autoregressive structure (AR(1))

United States Census Bureau

Thank you for your attention!

- Carolina.Franco@census.gov
- William.R.Bell@census.gov

United States Census Bureau

Selected Bibliography

- Ghosh, Malay; Natarajan, Kannan; Stroud, T. W. F.; and Carlin, Bradley P. (1998), "Generalized linear models for small-area estimation", *Journal of the American Statistical Association*, **93** , 273-282.
- Kish, L. (1987). Weighting in Deft^2. *The Survey Statistician*. June, 1987.
- Maples, J., (2012) "An Examination of the Relative Variance of Replicate Weight Variance Estimators for Ratios Through First-Order Expansions", 2012 Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods.

United States Census Bureau

Selected Bibliography

- Maples, Jerry J. and Bell, William R. (2007), "Small Area Estimation of School District Child Population and Poverty: Studying Use of IRS Income Tax Data," Research Report Number RRS2007-11, Statistical Research Division, U.S. Census Bureau, available at http://www.census.gov/srd/papers/pdf/rrs2007-11.pdf.
- Rao, J.N.K. (2003), *Small Area Estimation*, Hoboken, New Jersey: John Wiley
- 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.

United States Census Bureau