Discussion on "Using Models to Estimate Hog Production"

Gauri Sankar Datta

Mathematical Statistician, U.S. Census Bureau
and University of Georgia

In Consultation with Dr. Eric Slud
US Census Bureau
A CNSTAT Workshop
May 15, 2019
Washington, DC

Fay-Herriot Model: A Popular Model for Area-Level Data

1. Goal is to estimate θ_i, hog production for state i, $i = 1, \cdots, m$
2. State summary Y_i based on state sample *directly* estimates θ_i
3. Direct estimates are often not reliable
4. To develop reliable SAE, Fay and Herriot (1979) proposed a model for to "borrow strength" from other data source
5. **Sampling model:** $Y_i = \theta_i + e_i, \; e_i \overset{ind}{\sim} N(0, D_i), i = 1, \cdots, m$
6. Known sampling variances D_i: $D = \text{Diag}(D_1, \cdots, D_m)$
7. Fay-Herriot model connects θ_i to covariate x_i by a
8. **Linking model** : $\theta_i = x_i^T \beta + v_i, \; v_i \overset{ind}{\sim} N(0, \sigma_v^2), i = 1, \cdots, m$
9. The predictors x_i do not fully explain θ_i by linear regression
10. Random term v_i is the model error

Benchmarking of SAE Predictions

- SAE usually considers explicit use of models.
- These model-based estimates can differ widely from the direct estimates, especially for areas with very low sample sizes.
- One potential drawback of the model-based estimates is that when aggregated, the overall estimate for a larger geographical area may be quite different from the corresponding direct estimate, the latter being usually believed to be quite reliable.
- The problem can be more severe in the event of model failure.
- An overall agreement with the direct estimates at an aggregate level: often a political necessity to convince the legislators of the utility of small area estimates.
- A Bayesian benchmarking solution in Datta et al. (2011)

A Bayesian Benchmarking Solution

- $Y_1, \cdots, Y_m$: the direct estimators of the m small area means $\theta_1, \ldots, \theta_m$. Let $\mathbf{Y} = (Y_1, \cdots, Y_m)^T$, $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)^T$.
- Require estimators $\hat{\boldsymbol{\theta}}^{BM1} = (\hat{\theta}_1^{BM1}, \cdots, \hat{\theta}_m^{BM1})^T$ of $\boldsymbol{\theta}$ such that $\sum_{i=1}^m w_i \hat{\theta}_i^{BM1} = t$.
- t: either externally given or equals $\sum_{i=1}^m w_i \hat{\theta}_i$,where w_i are known weights.
- Bayesian approach: Minimize $\sum_{i=1}^m E[(\theta_i - e_i)^2 | \mathbf{y}]$ with respect to e_i's satisfying $\bar{e}_w = \sum_{i=1}^m w_i e_i = t$.
- $\hat{\theta}_i^B$ is the posterior mean of θ_i, $i = 1, \cdots, m$
- $\bar{\hat{\theta}}_w^B = \sum_{i=1}^m w_i \hat{\theta}_i^B$, $L = \sum_{i=1}^m w_i^2$.
- Benchmarked estimator: $\hat{\theta}_i^{BM1} = \hat{\theta}_i^B + L^{-1}(t - \bar{\hat{\theta}}_w^B) w_i$.

(MV) FH Measurement Error Model

$$\mathbf{Y}_i = \boldsymbol{\theta}_i + \mathbf{e}_i, \quad \boldsymbol{\theta}_i = \boldsymbol{\beta}\mathbf{x}_i + \boldsymbol{\delta}\mathbf{z}_i + \mathbf{v}_i,$$
$$\mathbf{X}_i = \mathbf{x}_i + \boldsymbol{\eta}_i, \qquad\qquad i = 1, \ldots, m$$

- $\mathbf{Y}_i$, $\boldsymbol{\theta}_i$ $s-$dimensional. $s = 1$ is univariate
- $s = 4$ to estimate statewide hog production $\boldsymbol{\theta}_i$ for the four weight groups
- **Covariates $\mathbf{x}_i$ are not observed, instead $\mathbf{X}_i$ are observed**. These covariates may be from a survey
- Covariates $\mathbf{z}_i$ are observed with no measurement error.
- $\boldsymbol{\beta}(s \times p)$, $\boldsymbol{\delta}(s \times q)$, and $\Sigma_v(s \times s)$ p.d.
- $\mathbf{e}_i \overset{ind}{\sim} N_s(0, \mathbf{D}_i) \perp \mathbf{v}_i \overset{iid}{\sim} N_s(0, \Sigma_v) \perp \boldsymbol{\eta}_i \overset{ind}{\sim} N_p(0, \mathbf{C}_i)$.
- Variance matrices $\mathbf{D}_i, \mathbf{C}_i$ are known and p.d.

Multivariate Cross-Sectional Time Series SAE Model*

- At time $t(= 1, \cdots, T)$, $\mathbf{Y}_{it}$ is an s-dimensional vector of direct estimators of some characteristics $\boldsymbol{\theta}_{it}$
- The problem is to estimate some function of the $\boldsymbol{\theta}_{it}$'s
- One possible MV C-S T-S extension of FH model given by Ghosh et al. (1996)
- $\mathbf{Y}_{it}|\boldsymbol{\theta}_{it} \stackrel{ind}{\sim} N_s(\boldsymbol{\theta}_{it}, \mathbf{D}_{it}), t = 1, \cdots, T, i = 1, \cdots, m$
- $\boldsymbol{\theta}_{it} = \mathbf{X}_{it}\boldsymbol{\alpha} + \mathbf{Z}_{it}\mathbf{b}_t + \mathbf{v}_{it}$, where $\mathbf{v}_{it} \stackrel{ind}{\sim} N_s(0, \boldsymbol{\Sigma}_v)$
- A Random Walk model for $\mathbf{b}_t$: $\mathbf{b}_t|\mathbf{b}_{t-1} \sim N(\mathbf{b}_{t-1}, \boldsymbol{\Sigma}_b)$
- An HB model may be based on diffuse priors for $\boldsymbol{\alpha}, \boldsymbol{\Sigma}_v$ and $\boldsymbol{\Sigma}_b$