The Area-Level Model: Some Recollections and Reflections

Robert E. Fay
Westat Inc. and JPSM/MPSM
March 15, 2024

Overview

- Fay and Herriot (1979)
- With Mamadou Diallo:
  - Rao and Yu (1994)
  - Application to National Crime Victimization Survey

And briefly

- sae2
- Australian Bureau of Statistics (2023)

Fay and Herriot (1979)

"Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data"

- Context
- The application
- The model
- Remarks

Context

- Stein (1955), James and Stein (1961) – inadmissibility of mean of multivariate normal for $m \geq 3$. “Stein’s paradox”
- Efron and Morris (1971, 1972, 1973a, 1973b, 1975) groundwork for practical applications
- Carter and Rolph (1974) “Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities”
- Ericksen (1973, 1974) – regression applied to survey data
- Madow, see Madow and Hansen (1975)

Revenue Sharing

- Congressional directive for Census Bureau to produce estimates of population and per capita income (PCI) for states and units of general-purpose local government
- PCI from 1970 Census 20% sample, updated using data from the Internal Revenue Service (IRS) and the Bureau of Economic Analysis (BEA) (Herriot 1977)
- Units with pop < 500 – used 1970 county PCI as base instead of direct estimates

Available data for small places

- For these small places or minor civil divisions (MCDs) most had
    - 20% sample PCI;
    - *Usually* value of housing on 100% basis, subject to quality checks
    - *Usually* IRS adjusted income/exemption, subject to quality checks

  as well as 20% sample PCI for the county

- Variance generalization for PCI estimates as a function of PCI estimate and estimated population size, based on eight states

The model

Population model

$$\theta_i = \mathbf{z}_i^T \beta + v_i$$

Sampling model

$$\hat{\theta}_i = \mathbf{z}_i^T \beta + v_i + e_i$$

where $E_m(v_i) = 0,\ V_m(v_i) = \sigma_v^2$ and

$$E_p(e_i|\theta_i) = 0,\ \ V_p(e_i|\theta_i) = \psi_i$$

The estimator for known $\boldsymbol{\sigma_v^2}$ *and* $\boldsymbol{\psi_i}$

Estimated regression parameters

$$\hat{\beta} = (Z^T V^{-1} Z)^{-1} Z^T V^{-1} \hat{\theta}$$

where V is a diagonal matrix with diagonal $\sigma_v^2 + \psi_i$

Best linear unbiased estimator (BLUP):

$$\theta_i^* = \mathbf{z}_i^T \hat{\beta} + \frac{\sigma_v^2}{\sigma_v^2 + \psi_i} (\hat{\theta}_i - \mathbf{z}_i^T \hat{\beta})$$

The EBLUP for known ψ_i

Alternatives to estimate σ_v^2:

- MLE (and now REML)
- Fitting of constants
- Method of moments

Empirical Best Linear Unbiased Estimator (EBLUP):

Plug in estimated σ_v^2 into BLUP

Decision to implement

- Calculated at state level and reviewed
- Decision to use as base for Revenue Sharing cleared at Division Chief level
- Limited special census data supported decision
- Revenue Sharing estimates already in production for these small places and MCDs.
  - Choose between revise base vs. no change
  - Different from choosing to publish estimates previously suppressed by reliability standards
- But for places 500-999, replaced direct estimates with modeled

Remarks

- Initially a “one-off”
  - Herriot successfully pushed for the 1980 Census to oversample population in small places and MCDs
- Highly constrained by computer resources:
  - Initial testing on about five (?) states but budget only allowed one full production run for the country
- FH (1979) reflected on possible variants after seeing results, esp., alternative functional relationship between $\widehat{N}_i$ and σ_v^2

... (fast forward)

See Rao (2003), Rao (2015), and other reviews for the growth in theory and applications of the area-level (and unit-level) SAE models

I will describe my recent involvement with one area-level application

National Crime Victimization Survey (NCVS)

National survey to measure criminal victimization *annually*

- Violent crime
  - Rape and sexual assault
  - Aggravated assault
  - Simple assault
  - Robbery
- Property crime
  - Burglary
  - Motor vehicle theft
  - (Other) Theft

Initially, collaboration with Mamadou Diallo

- Goal: estimates for states, (large) counties, and (large) metro areas
- Area-level model, estimated through EBLUP, using time series aspect of NCVS data, including multivariate nesting of types of crime
- Take advantage of long time series, e.g., 1999-2013, 2007-2018
- Account for sampling covariances over time
- Use FBI statistics, the Uniform Crime Reports (UCR), as auxiliary information
    - Census of reported crime but with missing data
    - Excludes crimes unreported to police

Cutting to the chase

- For 1999-2013 (states) 1998-2012 (large counties and large metro areas), published in 2015
- https://bjs.ojp.gov/library/publications/developmental-estimates-subnational-crime-rates-based-national-crime-0
- For 2007-2018 (states only), published in 2021
- https://bjs.ojp.gov/library/publications/constructing-and-disseminating-small-area-estimates-national-crime
- Published in the form of 3-year averages, although modeled at annual level

The FH model

Population model

$$\theta_i = \mathbf{z}_i^T \beta + \upsilon_i$$

Sampling model

$$\hat{\theta}_i = \mathbf{z}_i^T \beta + \upsilon_i + e_i$$

where $E_m(\upsilon_i) = 0,\ V_m(\upsilon_i) = \sigma_\upsilon^2$ and

$$E_p(e_i|\theta_i) = 0,\ \ V_p(e_i|\theta_i) = \psi_i$$

The Rao-Yu (1994) model

Population model

$$\theta_i = \mathbf{z}_{it}^T \beta + v_i + u_{it}$$

Sampling model

$$\hat{\theta}_i = \mathbf{z}_{it}^T \beta + v_i + u_{it} + e_{it}$$

where $E_m(v_i) = 0,\ V_m(v_i) = \sigma_v^2, E_p(e_{it}|\theta_{it}) = 0,\ \ V_p(\boldsymbol{e}_i|\boldsymbol{\theta}_i) = \boldsymbol{\psi}_i$ covariance matrix for each i, full matrix $\boldsymbol{\psi}$ block diagonal

u_{it} first-order auto-regressive series

$u_{it} = \rho u_{i,t-1} + \varepsilon_{it,} \quad |\rho| < 1,\ \varepsilon_{it} \sim^{ind} N(0, \sigma^2),$

The dynamic model

Population model

$$\theta_i = \mathbf{z}_{it}^T\beta + \rho^{t-1}v_i + u_{it}$$

Sampling model

$$\hat{\theta}_i = \mathbf{z}_{it}^T\beta + \rho^{t-1}v_i + u_{it} + e_{it}$$

where $E_m(v_i) = 0,\ V_m(v_i) = \sigma_v^2, E_p(e_{it}|\theta_{it}) = 0,\ \ V_p(\boldsymbol{e}_i|\boldsymbol{\theta}_i) = \boldsymbol{\psi}_i$ covariance matrix for each i, full matrix $\boldsymbol{\psi}$ block diagonal

u_{it} first-order auto-regressive series

$u_{it} = \rho u_{i,t-1} + \varepsilon_{it,} \quad |\rho| < 1,\ \varepsilon_{it} \sim^{ind} N(0, \sigma^2),$

Rao-Yu model, remarks

- Suited to short time series, e.g., sequence of years
- There are time-series alternatives for SAE: see, e.g., Rao and Molina (2015).
- Benefits from high $\rho < 1$, e.g., $\rho \approx 0.9$.
- Use of multiple years can increase the precision of the estimate of β, but changing regression relationships can be an issue.
- With Emily Berg and Elizabeth Petraglia, now investigating impact of changes in FBI crime statistics

Role of Time series in SAE

- From policy perspective, timeliness is very important for national statistics, *but*
- Time-series SAE may often give more weight to area effects (e.g., υ_i) than cross-sectional models

The sae2 *package*

- Meant as a footnote to the **sae** package accompanying Rao and Molina (2015)
- Initial contributions by Mamadou Diallo
- Implements the Rao-Yu and dynamic models using MLE and REML
- Allows for multivariate modeling (practically limited to 3 dimensions) and covariances of $\hat{\theta}_{it}$ over t. Rotating panel surveys, for example, require attention to covariances.
- Australian Bureau of Statistics (2023) is considering the package to model labor force characteristics.

Contact information

- bobfay@westat.com

References

- Australian Bureau of Statistics (2023, Oct. 30). A Rao-Yu model for small area estimation of labour force statistics. ABS https://www.abs.gov.au/statistics/research/rao-yu-model-small-area-estimation-labour-force-statistics. Accessed 13 Mar 2024.
- Carter, G.M. and Rolph, J.E. (1974). Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities. *Journal of the American Statistical Association*, 69, 880-885.
- Efron, B. and Morris, C. (1971). Limiting the Risk of Bayes and Empirical Bayes Estimators—Part I: The Bayes Case. *Journal of the American Statistical Association*, 66, 807-815.
- ______ (1972). Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case. *Journal of the American Statistical Association*, 67, 130-139.
- ______(1973a). Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach. *Journal of the American Statistical Association*, 68, 117-130.
- ______ (1973b). Combining Possibly Related Estimation Problems. *Journal of the Royal Statistical Society*, Ser.B, 35, 379-421.

References

- ______(1975). Data Analysis Using Stein's Estimator and Its Generalizations. *Journal of the American Statistical Association*, 70, 311-319.
- Ericksen, E. (1973). A Method of Combining Sample Survey Data and Symptomatic Indicators to Obtain Population Estimates for Local Areas. *Demography*, 10, 137-160.
- ______ (1974). A Regression Method for Estimating Population Changes for Local Areas. *Journal of the American Statistical Association*, 69, 867-875.
- Fay III, R.E. and Herriot, R.A. (1979). Estimation of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 74, 269-277.
- Herriot, R.A. (1977). Updating Per Capita Income for General Revenue Sharing. In *Small Area Statistics Papers, Series GE-41, No. 4*. U.S. Bureau of the Census.
- James, W. and Stein, C. (1961). Estimation with Quadratic Loss. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, 361-379.
- Madow and Hansen (1975). On Statistical Models and Estimation in Sample Surveys. In *Contributed Papers, 40th Session of the International Statistical Institute*, Warsaw, Poland, pp. 554-557.

References

- Rao, J.N.K. (2003). *Small Area Estimation*. John Wiley & Sons Inc. Hoboken, N.J.
- Rao. J.N.K and Molina, I (2015). *Small Area Estimation*. 2nd Ed. John Wiley & Sons Inc. Hoboken, N.J.
- Rao, J.N.K. and Yu, M. (1994). Small Area Estimation by Combining Time Series and Cross-Sectional Data. *Canadian Journal of Statistics*, 22, 511-526.
- Stein, C. (1955). Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*, Berkeley; University of California Press, 197-202.