Recommendations on SAE FH model

1 – Direct survey estimates and synthetic estimates

2 - Composite estimates and shrinkage weights

3 - Ingredients needed to apply F-H model

4 - Model checking

5 - References

Direct survey estimates

- $y_i \sim N(Y_i, \sigma_i^2)$ (y_i is weighted estimate for area i, not individual survey response; Y_i is true value)

- If direct estimates' sampling variances are unacceptably large, small area modeling may help.

- Census Bureau cutoff: want majority of the CVs of key estimates to be < 30%
$\hat{CV} = \sigma_i / y_i$
http://www.census.gov/quality/standards/standardf1.html

Synthetic estimates?

$$Y_i \sim N(X_i^T \beta, \sigma_M^2)$$

- Replacing y_i with $X^T \hat{\beta}$ as your estimate can improve noisiest estimates, but can also change your "good" estimates too much: largest areas should already have small $y_i - Y_i$ and may have comparatively larger $X_i^T \beta - Y_i$

Direct for largest, else synthetic?

$$\hat{Y}_i = w_i X_i^T \beta + (1 - w_i) y_i$$

- Each w_i is 0 or 1.

  $w_i = 0$: use direct estimate, ignore regression

  $w_i = 1$: use regression, ignore direct estimate

Composite / shrinkage estimates

$$\hat{Y}_i = w_i X_i^T \beta + (1 - w_i) y_i$$

- w_i ranges continuously between 0 and 1.
  w_i near 0: $\hat{Y}_i$ near y_i
  w_i near 1: $\hat{Y}_i$ near $X^T \hat{\beta}$

Shrinkage weights

$$\hat{Y}_i = w_i X_i^T \beta + (1 - w_i) y_i$$

- w_i ranges continuously between 0 and 1.
  w_i near 0: σ_i^2 low, trust y_i more
  w_i near 1: σ_i^2 high, trust $X^T \hat{\beta}$ more

- $y_i \sim N(Y_i, \sigma_i^2)$ and $Y_i \sim N(X_i^T \beta, \sigma_M^2)$

  so
  $$y_i \sim N(X_i^T \beta, \sigma_i^2 + \sigma_M^2)$$
  $$w_i = \sigma_i^2 / (\sigma_i^2 + \sigma_M^2)$$

Ingredients needed

$$\hat{Y}_i = \hat{w}_i X_i^T \hat{\beta} + (1 - \hat{w}_i)y_i$$

$$\hat{w}_i = \sigma_i^2 / (\sigma_i^2 + \hat{\sigma}_M^2)$$

- We have y_i and σ_i^2 from the survey (really an estimate, $\hat{\sigma}_i^2$, but we treat it as known); we have X from auxiliary / administrative data; we can estimate β using WLS regression of $y_i \sim N(X_i^T \beta, \sigma_i^2 + \sigma_M^2)$

- Just need a way to estimate σ_M^2

Estimating model variance

- Several good estimators; "REML" usually best

- "Prasad-Rao" simpler (for illustration only!)

- Under model $y_i \sim N(X_i^T \beta, \sigma_i^2 + \sigma_M^2)$
  regression MSE estimates average of $\sigma_i^2 + \sigma_M^2$
  $$\hat{MSE} \approx mean(\sigma_i^2 + \sigma_M^2) \approx mean(\sigma_i^2) + \sigma_M^2$$
  $$\sigma_M^2 \approx \hat{MSE} - mean(\sigma_i^2)$$

- Adjust for estimation of β too

Estimating model variance

- Prasad-Rao estimator:

$$\sigma_M^2 \approx M\hat{S}E - mean(\sigma_i^2)$$

$$\hat{\sigma}_M^2 = M\hat{S}E - \sum_i \sigma_i^2 (1 - h_i)/(m - p)$$

- h_i is i^{th} diagonal element
  of hat matrix, $X^T(X^TX)^{-1}X$;
  m is nr of areas; p is nr of parameters

- "Iterative": need OLS to get MSE to plug in
  here, then can WLS for actually estimating β

Standard errors of the estimates

- Standard errors of the new estimates should account for σ_i^2, σ_M^2, and estimation of β

$$g_1 = w_i \sigma_M^2 = \sigma_M^2 \sigma_i^2 / (\sigma_i^2 + \sigma_M^2)$$

$$g_2 = w_i^2 Var(X^T \hat{\beta})$$

$$\widehat{MSE}(\hat{Y}_i) = g_1 + g_2$$

- More advanced: can also add g_3, a term to account for estimation of σ_M^2; see Rao (2003)

Model checking

- Shrinkage weights: $\hat{w}_i$ are not all near 0 or 1?

- Model variance: if $\hat{\sigma}^2_M$ too close to 0, σ^2_i may be overestimates

- Raking factors: is sum of county-level SAE estimates close to state-level direct estimate?

- Compare to a "truth deck" (full census or admin data): check if point estimates and MSE are good, CI coverage is nominal, etc.

Complications

- Your data are not normal as given, but are approximately normal on a transformed scale?

$$\log(y_i) \sim N(X_i^T \beta, \sigma_i^2 + \sigma_M^2)$$

  Then need to correct for bias when transforming estimates back to original scale.

- Your data are not normal at all, but rather binomial, Poisson, etc.? Hierarchical Bayes modeling is more flexible (next time!)

Question from last time

- Is normality required for (non-Bayesian) Fay-Herriot model? **No**. This would be enough:

- $y_i = Y_i + e_i\,, Y_i = X_i^T \beta + u_i\,,$

- $E(e_i) = E(u_i) = 0\,, V(e_i) = \sigma_i^2\,, V(u_i) = \sigma_M^2$

- **Without** normality, we can estimate BLUP (best linear unbiased predictor) and get moment-based variance estimates. **With** normality, BLUP is also BP, and we can use MLE or REML.

Bibliography

- Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74, 269-277. http://www.jstor.org/stable/10.2307/2286322

- Longford, N.T. (2005). Missing Data And Small-area Estimation: Modern Analytical Equipment for the Survey Statistician. New York: Springer.

- Mukhopadhyay, P. (1998). Small Area Estimation in Survey Sampling. New Delhi: Narosa Pub House.

- Mukhopadhyay, P.K., and McDowell, A. (2011). Small Area Estimation for Survey Data Analysis Using SAS Software. Proceedings of the SAS Global Forum 2011 Conference. http://support.sas.com/resources/papers/proceedings11/336-2011.pdf

- Rao, J.N.K. (2003). Small Area Estimation. New York: Wiley.

- Sheu, C., and Suzuki, S. (2001). Meta-Analysis Using Linear Mixed Models. Behavior Research Methods, vol. 33, issue 2, 102-107. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.28.9389&rep=rep1&type=pdf

- Prasada N.G.D, Rao J.N.K (1990) The estimation of the mean squared error of the small area estimators, J. Am., Statistic. Ass. 85, 163-171.
- Pratesi M, Salvati N (2008) Small area estimation: the Eblup estimator based on spatially correlated random effects, Stats Methods Appl, 17, 114-131.
- Pratesi M., Salvati N. (2009), "Small area estimation in the presence of correlated random area effects", *Journal of Official Statistics*, 25, 37-53.
- Rao, J.N.K. and Molina, I. (2015) Small Area Estimation, 2nd edn. Wiley, Hoboken, New Jersey.
- Singh B.B., Shukla G.K, Kundu D. (2005) Spatio temporal models in small area estimation. Surv. Method. 31, 2, 183-195
- Maruhenda J, Morales D, Del Carmen Pardo M (XXXX) Information criteria for Fay and Herriot model selection, Comp. Stat. and Data An., XX, XX-XX
- Molina I, Salvati N, Pratesi M (2009) Boostrap for estimating the MSE of spatial Eblup, Computational Statistics, 24, 441-458.

- Giusti C, Marchetti S, Pratesi M, Salvati N (2012) Semiparametric Fay Herriot model using penalized spines, J. Ind. Soc. Agric. Stat. 66 (1), 1-14.
- Maruhenda J, Molina I, Morales D (2013) Small Area Estimation with spatio-temporal Fay and Herriot models, Comp. Stat. and Data An., 58, 308-325
- Benedetti R., Pratesi M, Salvati N (2013) Local stationarity in small area estimation models, Stats Methods Appl, XX, 22-81
- Pratesi M. (2015), "Spatial Disaggregation and Small Area Estimation Methods for Agricultural Surveys: Solutions and Perspectives", Technical Report in the Global Strategy Publications.
- Pratesi M., (2016) (ed) Analysis of Poverty Data by Small Area Estimation, New York, Wiley.
- Pratesi M., Salvati N. (2016), Introduction on measuring poverty at local level using SAE methods, in Pratesi M (ed) Analysis of Poverty Data by Small Area Estimation, New York, Wiley