Computer Applications for Small Area Estimation, Part 1

Jerzy Wieczorek
Small Area Estimation Research Group, CSRM
1/17/2013

Outline

- Basic approach for continuous area-level data
- The estimates we need and how to get them
- SAS example "by hand" with `PROC REG` and `DATA` steps
- SAS example "automated" with `PROC MIXED`
- Further resources
- Plan for next time

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 $\sigma_i^2 + \sigma_M^2$
  $M\hat{S}E \approx mean(\sigma_i^2 + \sigma_M^2) \approx mean(\sigma_i^2) + \sigma_M^2$
  $\sigma_M^2 \approx M\hat{S}E - mean(\sigma_i^2)$
- Adjust for estimation of β too

Estimating model variance

- Prasad-Rao estimator:
  $\sigma_M^2 \approx M\hat{S}E - mean(\sigma_i^2)$
  $\hat{\sigma}_M^2 = M\hat{S}E - \sum \sigma_i^2(1-h_i)/(m-p)$
- 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)

Put it all together "by hand"

- See example SAS code, "SAE_AreaLevel_ByHand.sas"
- Estimate σ_M^2 using OLS in `PROC REG`, run WLS in `PROC REG` for final β estimates, compute shrinkage weights, estimate $\hat{Y}_i$ and its standard error.
- Check regression residual plots; compare CVs of direct and SAE estimates.

"Automate it" with `PROC MIXED`

- See example SAS code, "SAE_AreaLevel_ProcMixed.sas"
- Put in an initial guess for σ^2_M, then let `PROC MIXED` estimate it for you using REML
- `PROC MIXED` also produces the shrinkage estimates and their standard errors for you
- Check plots of Marginal Studentized Residuals

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!)

Further resources

- SAS code and more examples: Mukhopadhyay & McDowell (2011), Sheu & Suzuki (2001)
- Area-level model: Fay & Herriot (1979)
- SAE books: Rao (2003), Longford (2005), Mukhopadhyay (1998)
- Small area group and visiting scholars in CSRM; SAMB and SAEB in SEHSD

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