Small Area Estimation for Multivariate repeated measures data

LinStat 20014, Linköping University

Innocent Ngaruye
Dietrich von Rosen, Martin Singull

Linköping University,Sweden

Outline

1. Introduction
2. The model formulation
3. Estimation of model parameters
4. Prediction of random effects
5. Simulation study example
6. Further research
7. Some references

Introduction

- Small Area Estimation (SAE) theory is concerned with solving the following problems

  1. How to produce reliable estimates of characteristics of interest, (total, means, quantiles, etc...) for small areas or domains, based on small samples or even no samples taken from these areas.
  2. How to assess the estimation or prediction error

- The only possible solution to the estimation problem to improve direct estimates is to "*borrow strength*" from other related data sets, either from similar areas, or relevant "*auxiliary information*" obtained from a recent census or some other administrative records.

Introduction (cont'd)

- We propose to apply the multivariate linear regression model for repeated measurements in SAE settings to get a model which borrows strength across both small areas and over time by incorporating simultaneously the effects of areas and time interaction.

- This model accounts for repeated surveys, group individuals and random effects variation. The estimation is discussed with a likelihood based approach and a simulation study is conducted.

The model formulation (cont'd)

- We consider repeated measurements on variable of interest y for p time points, $t_1, \ldots, t_p$ from the finite population U of size N partitioned into m disjoint subpopulations or domains $U_1, \ldots, U_m$ called *small areas* of sizes N_i, $i = 1, \ldots, m$ such that $\sum_{i=1}^{m} N_i = N$.

- We also assume that in every area, there are k different groups of units of size N_{ig} for goup g such that $\sum_{i}^{m} \sum_{g=1}^{k} N_{ig} = N$.

- We draw a sample of size n in all small areas such that the sample of size n_i is observed in area i and $\sum_{i}^{m} \sum_{g=1}^{k} n_{ig} = n$ and we suppose that we have auxiliary data $\mathbf{x}_{ij}$ of r variables (covariates) available for each population unit j in all m small areas.

The model formulation (cont'd)

- Consider the simple linear regression model for unit j in the ith area at a given time t

$$y_{ijt} = \beta_0 + \beta_1 t + \cdots + \beta_q t^{q-1} + \boldsymbol{\gamma}'\mathbf{x}_{ij} + u_{it} + e_{ijt}. \qquad (1)$$
$$j = 1, \cdots, N_i; i = 1, \cdots, m; t = 1, \cdots, p$$

where $e_{ijt} \sim \mathcal{N}(0, \sigma_e^2)$ and is independent of u_{it}.

- Varying t, the model can be written as

$$\mathbf{y}_{ij} = \mathbf{A}\boldsymbol{\beta} + \mathbf{1}_p \boldsymbol{\gamma}'\mathbf{x}_{ij} + \mathbf{u}_i + \mathbf{e}_{ij} \qquad (2)$$

where $\mathbf{u}_i \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma}_u)$ and $\mathbf{e}_{ij} \sim \mathcal{N}_p(\mathbf{0}, \sigma_e^2\mathbf{I})$

The model formulation (cont'd)

- Collecting the vectors $\mathbf{y}_{ij}$ for all units in small area i for different groups, the model for each Small Area is given by

$$\begin{aligned}\mathbf{Y}_i =& \mathbf{ABC}_i + \mathbf{1}_p\gamma'\mathbf{X}_i + \mathbf{u}_i\mathbf{z}_i + \mathbf{E}_i,\\ &\mathbf{u}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_u),\\ &\mathbf{E}_i \sim \mathcal{N}_{p,N_i}(\mathbf{0}, \sigma_e^2\mathbf{I}, \mathbf{I}_{N_i}),\end{aligned} \tag{3}$$

where $\mathbf{A}$ and $\mathbf{C}_i$ are resectively *within-individual* and *between-individual design matrices for fixed effects* given by

$$\mathbf{A} = \begin{pmatrix} 1 & t_1 & \cdots & t_1^{q-1} \\ 1 & t_2 & \cdots & t_2^{q-1} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & t_p & \cdots & t_p^{q-1} \end{pmatrix}, \mathbf{C}_i = \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & \ldots & \vdots & \vdots & \ldots & \vdots & \vdots & \ldots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}$$

The model formulation (cont'd)

- The corresponding model for all small areas can be expressed as

$$
\begin{aligned}
\mathbf{Y} =& \mathbf{ABHC} + \mathbf{1}_p \gamma' \mathbf{X} + \mathbf{UZ} + \mathbf{E} \\
& \mathbf{E} \sim \mathcal{N}_{p,N}(\mathbf{0}, \boldsymbol{\Sigma}_e, \mathbf{I}_N), \quad \mathbf{U} \sim \mathcal{N}_{p,m}(\mathbf{0}, \boldsymbol{\Sigma}_u, \mathbf{I}_m) \\
& vec(\mathbf{Y}) \sim \mathcal{N}_{pN}\Big(vec(\mathbf{ABHC} + \mathbf{1}_p \gamma' \mathbf{X}), \boldsymbol{\Sigma} \Big)
\end{aligned}
\tag{4}
$$

where

$$
\mathbf{H} = [\mathbf{I}_r : \mathbf{I}_r : \cdots : \mathbf{I}_r], \quad \boldsymbol{\Sigma} = \mathbf{Z}'\mathbf{Z} \otimes \boldsymbol{\Sigma}_u + \mathbf{I}_N \otimes \boldsymbol{\Sigma}_e, \quad \boldsymbol{\Sigma}_e = \sigma_e^2 \mathbf{I}
$$

σ_e^2 is assumed to be known.

Estimation of model parameters

- The model (4) is not a matrix normal distribution and then we transform it into two indepent components $\mathbf{Y}[\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1} : \mathbf{C}'^{o}] = [\mathbf{V} : \mathbf{W}]$ with

$$\mathbf{V} = \mathbf{ABH} + \mathbf{1}_p\gamma'\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1} + (\mathbf{UZ} + \mathbf{E})\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}$$
$$\mathbf{W} = \mathbf{1}_p\gamma'\mathbf{X}\mathbf{C}'^{o} + \mathbf{E}\mathbf{C}'^{o}$$

so that

$$Vec(\mathbf{V}) \sim \mathcal{N}_{pmk}\Big(Vec(\mathbf{ABH} + \mathbf{1}_p\gamma'\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}, \mathbf{\Psi}\Big)$$
$$\mathbf{W} \sim \mathcal{N}_{p,N-mk}\Big(\mathbf{1}_p\gamma'\mathbf{X}\mathbf{C}'^{o}, \mathbf{\Sigma}_e, \mathbf{I}_{N-mk}\Big)$$

where

$$\mathbf{\Psi} = (\mathbf{C}\mathbf{C}')^{-1}\mathbf{C}\mathbf{Z}'\mathbf{Z}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1} \otimes \mathbf{\Sigma}_u + (\mathbf{C}\mathbf{C}')^{-1} \otimes \mathbf{\Sigma}_e$$

$\mathbf{A}^{o}$ denotes any matrix of full rank spanning $\mathcal{C}(\mathbf{A})^{\perp}$

Estimation of model parameters (cont'd)

- By simultaneous decomposition of

$$(\mathbf{CC}')^{-1}\mathbf{CZ}'\mathbf{ZC}'(\mathbf{CC}')^{-1} = \mathbf{RDR}' \quad \text{and} \quad (\mathbf{CC}')^{-1} = \mathbf{RR}'$$

a further transformation gives $\mathbf{VR}^{-1'} = [\mathbf{V}^1 : \mathbf{V}^2]$ so that we have

$$\mathbf{V}^1 \sim \mathcal{N}_{p,m}\Big(\mathbf{ABHR}_1, \mathbf{\Sigma}_u, \mathbf{I}_m\Big),$$
$$\mathbf{V}^2 \sim \mathcal{N}_{p,mk-m}\Big(\mathbf{ABHR}_2, \mathbf{\Sigma}_e, \mathbf{I}_{mk-m}\Big)$$

where $\mathbf{R} = [\mathbf{R}_1 : \mathbf{R}_2]$ comformably to the structure of $\mathbf{D}$, that is $\mathbf{D} = \begin{pmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$ and $\mathbf{R}_1$ and $\mathbf{R}_2$ are such that $[\mathbf{R}_1 : \mathbf{R}_2]^{-1'} = [\mathbf{R}^1 : \mathbf{R}^2]$

Estimation of model parameters (cont'd)

Then, the maximum likelihood estimators (MLE) are given by

$$
\begin{aligned}
\widehat{\gamma}_{\text{MLE}} &= \frac{1}{p}\left(\mathbf{X}\mathbf{C}'^{o}(\mathbf{C}'^{o})'\mathbf{X}'\right)^{-1}\mathbf{X}\mathbf{C}'^{o}(\mathbf{C}'^{o})'\mathbf{Y}'\mathbf{1}_p \\
\widehat{\mathbf{B}}_{\text{MLE}} &= (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{V}^2\mathbf{R}^{2'}\mathbf{H}'(\mathbf{H}\mathbf{R}^2\mathbf{R}^{2'}\mathbf{H}')^{-} + \widehat{\mathbf{T}_1}(\mathbf{H}\mathbf{R}^2\mathbf{R}^{2'}\mathbf{H}')^{o'} \\
\widehat{\mathbf{\Sigma}}_{u\text{MLE}} &= \frac{1}{m}(\mathbf{V}^1 - \mathbf{A}\widehat{\mathbf{T}_1}\mathbf{C}_s)(\mathbf{V}^1 - \mathbf{A}\widehat{\mathbf{T}_1}\mathbf{C}_s)' - \mathbf{\Sigma}_e
\end{aligned}
$$

where

$$
\begin{aligned}
\widehat{\mathbf{T}_1} &= (\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-}\mathbf{A}'\mathbf{S}^{-1}\mathbf{V}^1\mathbf{C}_s'(\mathbf{C}_s\mathbf{C}_s')^{-} \\
\mathbf{S} &= \mathbf{V}^1(\mathbf{I} - \mathbf{C}_s'(\mathbf{C}_s\mathbf{C}_s')^{-}\mathbf{C}_s)\mathbf{V}^{1'} \\
\mathbf{C}_s &= (\mathbf{H}\mathbf{R}^2\mathbf{R}^{2'}\mathbf{H}')^{o'}\mathbf{H}\mathbf{R}_1
\end{aligned}
$$

Prediction of random effects

- Given the model

$$\mathbf{Y} = \mathbf{ABHC} + \mathbf{1}_p\gamma'\mathbf{X} + \mathbf{UZ} + \mathbf{E}$$

- By maximizing the joint density $f(\mathbf{Y}, \mathbf{U})$ with respect to $\mathbf{U}$ assuming the covariance matrices $\mathbf{\Sigma}_u$ and $\mathbf{\Sigma}_e$ to be known, we find the prediction of $\mathbf{U}$

$$\widehat{\mathbf{U}} = \left(\mathbf{\Sigma}_e\mathbf{\Sigma}_u^{-1} + \mathbf{I}_p\right)^{-1}(\mathbf{Y} - \mathbf{A}\widehat{\mathbf{B}}_{\mathrm{MLE}}\mathbf{HC} - \mathbf{1}_p\widehat{\gamma}'_{\mathrm{MLE}}\mathbf{X})\mathbf{Z}'$$

which leads to

$$\widehat{\mathbf{U}} = \left(\mathbf{\Sigma}_e\widehat{\mathbf{\Sigma}}_{u\mathrm{MLE}}^{-1} + \mathbf{I}_p\right)^{-1}(\mathbf{Y} - \mathbf{A}\widehat{\mathbf{B}}_{\mathrm{MLE}}\mathbf{HC} - \mathbf{1}_p\widehat{\gamma}'_{\mathrm{MLE}}\mathbf{X})\mathbf{Z}'$$

for $\mathbf{\Sigma}_u$ replaced by its estimator $\widehat{\mathbf{\Sigma}}_{u\mathrm{MLE}}$.

Simulation study Example

We consider 8 small areas and draw a sample, we assume $p = 4$ and $r = 3$

Table : Sample sizes

Area	Group 1	Group 2	Group 3	Total
1	n_{11}=12	n_{12}=18	n_{13}=16	n_1=46
2	n_{21}=21	n_{22}=23	n_{23}=12	n_2=56
3	n_{31}=10	n_{32}=20	n_{33}=15	n_3=45
4	n_{41}=16	n_{42}=24	n_{43}=17	n_4=57
5	n_{51}=24	n_{52}=26	n_{53}=21	n_5=71
6	n_{61}=20	n_{62}=12	n_{63}=28	n_6=60
7	n_{71}=27	n_{72}=13	n_{73}=14	n_7=54
8	n_{81}=20	n_{82}=14	n_{83}=27	n_8=61
m=8	g_1=150	g_2=150	g_3=150	n=450

Simulation study Example (cont'd)

The design matrices are

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}, \qquad \mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathbf{C}_8 \end{pmatrix}$$

$$\text{for} \quad \mathbf{C}_i = \left(\mathbf{1}'_{n_{i1}} \otimes \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} : \mathbf{1}'_{n_{i2}} \otimes \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} : \mathbf{1}'_{n_{i3}} \otimes \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right),$$

$i = 1, \cdots 8;$

Simulation study Example (cont'd)

The parameter matrices are

$$\mathbf{B} = \begin{pmatrix} 8 & 9 & 10 \\ 11 & 12 & 13 \end{pmatrix}, \quad \gamma = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

The sampling variance is assumed to be $\sigma_e^2 = 0.4$ and the covariance for random effects to be

$$\boldsymbol{\Sigma}_u = \begin{pmatrix} 1.6 & 1.7 & 1.8 & 2.6 \\ 1.7 & 2.8 & 1.3 & 3.8 \\ 1.8 & 1.3 & 3.9 & 3.1 \\ 2.6 & 3.8 & 3.1 & 9.2 \end{pmatrix}$$

Simulation study Example (cont'd)

Then, we generate the data from

$$Vec(\mathbf{Y}) \sim \mathcal{N}_{pn}(Vec(\mathbf{ABHC} + \mathbf{1}_p\gamma'\mathbf{X}), \mathbf{\Sigma}, \mathbf{I}_n)$$

where the matrix of covariates $\mathbf{X}$ is generated with random elements. The following MLEs are obtained:

$$\widehat{\mathbf{B}} = \begin{pmatrix} 8.0578 & 9.0730 & 10.0350 \\ 11.0388 & 12.0140 & 13.0388 \end{pmatrix}, \quad \widehat{\gamma} = \begin{pmatrix} 1.0166 \\ 1.9542 \\ 2.9752 \end{pmatrix}$$

Further research

- After obtaining all unkown parameters, then we can find directly the target small area characteristics of interest such as the small area totals and samall area means
- In further research, we want to test the efficiency, the distribution and all properties of the estimators
- We wish also to study the possible time correlation

Some references

- Bai Peng, *Exact distribution of MLE of covariance matrix in GMANOVA-MANOVA model*, J. Science in China, 2005
- Battese, G.E, R.M. and W.A. Fuller, *An error-components model for prediction of county crop areas using survey and satellite data*, American Statistical Association, 1988.
- Danny Pfeffermann *Small Area Estimation-New Develooments and Directions*. J. International Statistical Review, 2002.
- G. Datta, P. Lahiri, T. Maiti, K. Lu, *Hierarchical Bayes estimation of unemployment rates for the states of the US*, Journal of the American Statistical Association 94 (1999)
- G.K. Robinson, *That BLUP Is a Good Thing: The estimation of Random Effects*, Statistical Science, Vol. 6, 1991
- J.N.K. Rao, *Small Area Estimation*. Willey, 2003.

Some references

- T. Kollo and D. von Rosen, *Advanced Multivariate Statistics with matrices*. Springer, 2005.
- Kari Nissinen, *Small Area Estimation with Linear Mixed Models from Unit-level panel and Rotating panel data*. PhD Thesis, Jyväskylä University, 2009..
- M. Ghosh and J.N.K. Rao, *Small Area Estimation: An Appraisal*, J. Statistical Science, 1994.
- R. Chambers and R. G. Clark, *An introduction to Model-Based Survey Sampling with Applications*. Oxford, 2012.
- Robb J. Muirhead, *Aspects of Multivariate Statistical theory* , Wiley 2005.
- Tatsuya Kubokawa and Muni S. Srivastava, *Prediction in Multivariate Mixed Linear Models*, J. Japan Statist. Soc, 2003

THANKS !!!!!!!!