Small Area Estimation under the Growth Curve model

Innocent Ngaruye,

Linköping University, Sweden

Outline

- Introduction
- The model formulation
- Estimation of model parameters
- Prediction of random effects
- Simulation study example
- Further research
- Some references

Introduction

- The term *Growth Curve Modeling* has been used in different contexts to refer to a wide array of statistical models for repeated measures data.

- It has long played a significant role in empirical research within the developmental sciences, particulary in studying between-individual differences and within-individual patterns of change over time.

Introduction (cont'd)

- We propose to apply this model 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, $t_1, ..., t_p$ from the finite population U of size N partitioned into m disjoint subpopulations or domains $U_1, ..., U_m$ called *small areas* of sizes $N_i, i = 1, ..., m$ 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)

- The model at Small Area level is given by

$$\begin{aligned} \mathbf{Y}_i =& \mathbf{A}\mathbf{B}_i\mathbf{C}_i + \mathbf{1}\gamma'\mathbf{X}_i + \mathbf{1}\mathbf{u}_i' + \mathbf{E}_i, \\ \mathbf{u}_i &\sim \mathcal{N}_{N_i}(\mathbf{0}, \boldsymbol{\sigma}_u^2\mathbf{I}), \\ \mathbf{E}_i &\sim \mathcal{N}_{p,N_i}(\mathbf{0}, \boldsymbol{\sigma}_e^2\mathbf{I}, \mathbf{I}_{N_i}), \end{aligned} \tag{1}$$

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

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

The model formulation (cont'd)

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

$$\underbrace{\mathbf{Y}}_{p\times N} = \underbrace{\mathbf{A}}_{p\times q}\underbrace{\mathbf{B}}_{q\times mk}\underbrace{\mathbf{C}}_{mk\times N} + \underbrace{\mathbf{1}\gamma'[\mathbf{I}_r : \mathbf{I}_r : \cdots : \mathbf{I}_r]}_{p\times mr}\underbrace{\mathbf{X}}_{mr\times N} + \underbrace{\mathbf{1}}_{p\times 1}\underbrace{\mathbf{u}'}_{1\times N} + \underbrace{\mathbf{E}}_{p\times N}$$

or

$$\mathbf{Y} = \mathbf{ABC} + \mathbf{1}\gamma'\mathbf{DX} + \mathbf{1u}' + \mathbf{E}, \qquad (2)$$

$$\text{for } \mathbf{D} = [\mathbf{I}_r : \mathbf{I}_r : \cdots : \mathbf{I}_r]$$

Estimation of model parameters

- In order to transform (2) to a model which is easier to estimate, we transform the design matrix $\mathbf{A}$ into a new matrix $\mathbf{A}_1$ with two parts $\mathbf{A}_1 = [\mathbf{1} : \mathbf{H}]$ and the parameter matrix into a new matrix $\mathbf{\Xi} = [\xi_1 : \mathbf{\Xi}_2]$ comformably such that

$$\mathcal{C}(\mathbf{A}) = \mathcal{C}(\mathbf{1}) \oplus \mathcal{C}(\mathbf{H}) \text{ with } \mathcal{C}(\mathbf{H}) = \mathcal{C}(\mathbf{1})^{\perp} \cap \mathcal{C}(\mathbf{A})$$

- One way of this transformation is given below

$$\mathbf{A} = \begin{pmatrix} 1 & t_1 & \cdots & t_1^{q-1} \\ 1 & t_2 & \cdots & t_2^{q-1} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & t_p & \cdots & t_p^{q-1} \end{pmatrix} \longrightarrow \mathbf{A}_1 = \begin{pmatrix} 1 & t_1 - \overline{t} & \cdots & t_1^{q-1} - \overline{t^{q-1}} \\ 1 & t_2 - \overline{t} & \cdots & t_2^{q-1} - \overline{t^{q-1}} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & t_p - \overline{t} & \cdots & t_p^{q-1} - \overline{t^{q-1}} \end{pmatrix}$$

Estimation of model parameters (cont'd)

- We come up with the model

$$\mathbf{Y} = \mathbf{1}\xi_1'\mathbf{C} + \mathbf{H}\boldsymbol{\Xi}_2\mathbf{C} + \mathbf{1}\gamma'\mathbf{D}\mathbf{X} + \mathbf{1}\mathbf{u}' + \mathbf{E}$$

and make a one-to-one transformation

$$\begin{pmatrix} \mathbf{1}'\mathbf{Y} \\ \mathbf{H}'\mathbf{Y} \\ \mathbf{A}^{o\prime}\mathbf{Y} \end{pmatrix} = \begin{pmatrix} p\xi_1'\mathbf{C} + p\gamma'\mathbf{D}\mathbf{X} + p\mathbf{u}' + \mathbf{1}'\mathbf{E} \\ \mathbf{H}'\mathbf{H}\boldsymbol{\Xi}_2\mathbf{C} + \mathbf{H}'\mathbf{E} \\ \mathbf{A}^{o\prime}\mathbf{E} \end{pmatrix},$$

where $\mathbf{A}^o$ for a matrix $\mathbf{A}$ is such that $\mathbf{A}^{o\prime}\mathbf{A} = \mathbf{0}$ and $\mathcal{C}(\mathbf{A}^o) = \mathcal{C}(\mathbf{A})^\perp$.

Estimation of model parameters (cont'd)

After calculation, the maximum likelihood estimators are given by

$$\widehat{\mathbf{\Xi}}_2 = \Big(\mathbf{H}'\mathbf{H}\Big)^{-}\mathbf{H}'\mathbf{Y}\mathbf{C}'\Big(\mathbf{C}\mathbf{C}'\Big)^{-} + \Big(\mathbf{H}'\mathbf{H}\Big)^{o}\mathbf{T}_1 + \mathbf{H}'\mathbf{H}\mathbf{T}_2\Big(\mathbf{C}\mathbf{C}'\Big)^{o'}$$

$$\widehat{\gamma}' = \frac{1}{p}\Big[\mathbf{1}'\mathbf{Y}\mathbf{X}'\mathbf{D}' - \mathbf{1}'\mathbf{Y}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-}\mathbf{C}\mathbf{X}'\mathbf{D}' - p\mathbf{T}_3\Big(\mathbf{C}\mathbf{C}'\Big)^{o}\mathbf{C}\mathbf{X}'\mathbf{D}'\Big]$$

$$\times\Big[\mathbf{D}\mathbf{X}\mathbf{X}'\mathbf{D}' - \mathbf{D}\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-}\mathbf{C}\Big]^{-}$$

$$\widehat{\xi}_1' = \Big(\frac{1}{p}\mathbf{1}'\mathbf{Y} - \widehat{\gamma}'\mathbf{D}\mathbf{X}\Big)\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-} + \mathbf{T}\Big(\mathbf{C}\mathbf{C}'\Big)^{o}$$

for some matrices $\mathbf{T}, \mathbf{T}_1, \mathbf{T}_2$ and $\mathbf{T}_3$ of proper sizes.

Estimation of model parameters (cont'd)

- Once $\widehat{\xi_1'}$ and $\widehat{\boldsymbol{\Xi}_2}$ are obtained, we can then find the parameter matrix $\mathbf{B}$ by solving the linear system

$$\mathbf{1}\widehat{\xi_1'}\mathbf{C} + \mathbf{H}\widehat{\boldsymbol{\Xi}_2}\mathbf{C} = \mathbf{A}\widehat{B}\mathbf{C}.$$

Since, the matrices $\mathbf{A}$ and $\mathbf{C}$ are of full rank, then

$$\widehat{\mathbf{B}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\left(\mathbf{1}\widehat{\xi_1'}\mathbf{C} + \mathbf{H}\widehat{\boldsymbol{\Xi}_2}\mathbf{C}\right)\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}.$$

Estimation of model parameters (cont'd)

- Given the covariance structure of $\mathbf{Y}$

$$\boldsymbol{\Sigma} = \mathbf{1}\boldsymbol{\Sigma}_u\mathbf{1}' + \boldsymbol{\Sigma}_e = m\sigma_u^2\mathbf{1}\mathbf{1}' + \sigma_e^2\mathbf{I}_p,$$

and its inverse

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_e^2}\left(\mathbf{I}_p - \frac{m\sigma_u^2}{mp\sigma_u^2 + \sigma_e^2}\mathbf{1}\mathbf{1}'\right).$$

- We find the maximum likelihood estimator of the variance component axpressed by

$$\widehat{\sigma}_u^2 = \frac{\mathrm{tr}\left\{\mathbf{1}\mathbf{1}'\mathbf{W}\right\} - Np\sigma_e^2}{Nmp^2},$$

where

$$\mathbf{W} = (\mathbf{Y} - \mathbf{ABC} - \mathbf{1}\gamma'\mathbf{DX})(\mathbf{Y} - \mathbf{ABC} - \mathbf{1}\gamma'\mathbf{DX})'.$$

Prediction of random effects

- Under the theory of linear model and normal distribution, the best linear predictor of u that minimizes the mean square error is the conditional mean $E[\mathbf{u}|\mathbf{Y}]$ given by

$$E[\mathbf{u}|\mathbf{Y}] = E[\mathbf{u}] + Cov(\mathbf{u}', \mathbf{Y})Cov^{-1}(\mathbf{Y})(\mathbf{Y} - E[\mathbf{Y}]).$$

- Thus,

$$\begin{aligned}\widehat{\mathbf{u}} =& \widehat{\sigma}_u^2 \mathbf{1}'\widehat{\boldsymbol{\Sigma}}^{-1}(\mathbf{Y} - \mathbf{A}\widehat{\mathbf{B}}\mathbf{C} - \mathbf{1}\widehat{\gamma}'\mathbf{D}'\mathbf{X}) \\ =& \frac{\widehat{\sigma}_u^2}{mp\widehat{\sigma}_u^2 + \sigma_e^2}\mathbf{1}'(\mathbf{Y} - \mathbf{A}\widehat{\mathbf{B}}\mathbf{C} - \mathbf{1}\widehat{\gamma}'\mathbf{D}'\mathbf{X})\end{aligned}$$

Simulation study Example

We consider 6 small areas and draw a sample with the following sample sizes.

Table : Sample sizes

Area	Group 1	Group 2	Total
1	n_{11}=52	n_{12}=48	n_1=100
2	n_{21}=60	n_{22}=60	n_2=120
3	n_{31}=30	n_{32}=40	n_3=70
4	n_{41}=46	n_{42}=22	n_4=68
5	n_{51}=65	n_{52}=65	n_5=130
6	n_{61}=50	n_{62}=62	n_6=112
m=6	g_1=303	g_2=297	n=600

We assume $p = 4$ and $r = 3$.

Simulation study Example (cont'd)

The design matrices are

$$\mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} \longrightarrow \mathbf{H} = \begin{pmatrix} -1.5 \\ -0.5 \\ 0.5 \\ 1.5 \end{pmatrix},$$

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & & \mathbf{0} \\ & \cdots & \\ \mathbf{0} & & \mathbf{C}_6 \end{pmatrix} \quad \text{for} \quad \mathbf{C}_i = \left(\mathbf{1}'_{n_{i1}} \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} : \mathbf{1}'_{n_{i2}} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right),$$

$$i = 1, \cdots 6;$$

Simulation study Example (cont'd)

The parameter matrices are

$$\xi_1' = \begin{pmatrix} 20 & 21 & 22 & 23 & 24 & 25 & 26 & 27 & 28 & 29 & 30 & 31 \end{pmatrix},$$
$$\boldsymbol{\Xi}_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \end{pmatrix},$$

$$\begin{aligned}\mathbf{B} =& \mathbf{A}^{-}\left(\mathbf{1}\xi_1'\mathbf{C} + \mathbf{H}\boldsymbol{\Xi}_2\mathbf{C}\right)\mathbf{C}^{-} \\ =& \begin{pmatrix} 17.5 & 16 & 14.5 & 13 & 11.5 & 10 & 8.5 & 7 & 5.5 & 4 & 2.5 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \end{pmatrix},\end{aligned}$$

and

$$\gamma = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \sigma_u^2 = 5, \quad \sigma_e^2 = 6.$$

Simulation study Example (cont'd)

Then, the data are generated from

$$\mathbf{Y} \sim N_{p,n}(\mathbf{ABC} + \mathbf{1}\gamma'\mathbf{DX}, \mathbf{\Sigma}, \mathbf{I}_n),$$

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

$$\widehat{\xi_1}' = (\ 20.2534 \quad 21.6548 \quad 22.5961 \quad 23.6486 \quad 24.4233 \quad 25.0374 \quad 25.998$$
$$28.5361 \quad 29.9077 \quad 30.3292 \quad 31.1121 \ )$$
$$\widehat{\mathbf{\Xi}}_2 = (\ 1.1151 \quad 2.0824 \quad 3.0320 \quad 3.6376 \quad 4.6384 \quad 5.7882 \quad 7.0238 \quad 7.87$$
$$9.0386 \quad 10.1256 \quad 10.8561 \quad 11.9422 \ )$$

Simulation study Example (cont'd)

$$\widehat{\mathbf{B}} = \begin{pmatrix} 17.4657 & 17.3902 & 14.0748 & 14.5546 & 12.8274 & 10.5669 & 8.4390 \\ 1.1151 & 2.0824 & 3.0320 & 3.6376 & 4.6384 & 5.7882 & 7.0238 \end{pmatrix.$$

$$\begin{matrix} 5.9397 & 4.5936 & 3.1890 & 1.2566 \\ 9.0386 & 10.1256 & 10.8561 & 11.9422 \end{matrix} \Bigg)$$

$$\widehat{\sigma_u^2} = 5.0061, \quad \widehat{\gamma} = \begin{pmatrix} 1.0093 \\ 1.9501 \\ 3.0469 \end{pmatrix}, \quad \mathbf{ABC} = \begin{pmatrix} 18.5 & \cdots & 18 & \cdots & 13 \\ 19.5 & \cdots & 20 & \cdots & 25 \\ 20.5 & \cdots & 22 & \cdots & 37 \\ 21.5 & \cdots & 24 & \cdots & 49 \end{pmatrix}$$

$$\widehat{\mathbf{ABC}} = \begin{pmatrix} 18.5808 & \cdots & 18.4726 & \cdots & 13.1988 \\ 19.6959 & \cdots & 20.5550 & \cdots & 25.1410 \\ 20.8110 & \cdots & 22.6373 & \cdots & 37.0833 \\ 21.9261 & \cdots & 24.7197 & \cdots & 49.0255 \end{pmatrix}$$

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

Note: This work is being jointly conducted in the framework of my PhD research together with my supervisors:

Dietrich von Rosen and Martin Singull to whom I owe my aknowledgements.

THANKS !!!!!!!!