Small Area Estimation via Heteroscedastic Nested-Error Regression

Jiming Jiang & Thuan Nguyen

University of California,Davis, USA
and Oregon Health & Science University, Portland, USA

Presenter: Thuan Nguyen

09/02/2013

Introduction

- Small area estimation explores the idea of "borrowing strength" via statistical modeling.
- One important class of these models are the nested-error regression (NER) model.
- Battese *et al.* (1988) discussed data from 12 Iowa counties obtained from the 1978 June Enumerative Survey of the U.S. Department of Agriculture as well as data obtained from land observatory satellites on crop areas.
- The objective was to predict mean hectares of crops per segment for the 12 counties using the satellite information.

Nested-Error Regression (NER)

The NER model may be described as follows:
Consider sampling from finite subpopulations
$P_i = \{Y_{ik}, k = 1, \ldots, N_i\}, i = 1, \ldots, m.$
Suppose that auxiliary data $X_{ikl}, k = 1, \ldots, N_i, l = 1, \ldots, p$ are available for each P_i.
We assume that the following super-population NER model (Battese *et al.* 1988):
$Y_{ik} = \mathbf{X}'_{ik}\beta + v_i + e_{ik},\ i = 1, \ldots, m,\ k = 1, \ldots, N_i,$ where
$\mathbf{X}_{ik} = (X_{ikl})_{1 \le l \le p}$,
v_i's are domain-specific random effects, and e_{ik}'s are additional errors, such that the random effects and errors are independent with $v_i \sim N(0, \sigma_v^2)$ and $e_{ik} \sim N(0, \sigma_e^2)$.
We are interested in estimating the finite population mean of P_i, $\mu_i = N_i^{-1} \sum_{k=1}^{N_i} Y_{ik}$.

Nested-Error Regression (NER), cont.

Under the NER model, the BP of μ_i is

$$\mathrm{E}_{\mathrm{M},\psi}(\mu_i|y) = N_i^{-1}\{\sum_{j=1}^{n_i} y_{ij} + \sum_{k\notin I_i} \mathrm{E}_{\mathrm{M},\psi}(Y_{ik}|y_i)\},$$

which can be expressed as

$$\tilde{\mu}_i(\psi) = \bar{\mathbf{X}}_i'\beta + \left\{\frac{n_i}{N_i} + \left(1-\frac{n_i}{N_i}\right)\frac{n_i\sigma_v^2}{\sigma_e^2 + n_i\sigma_v^2}\right\}(\bar{y}_{i\cdot} - \bar{\mathbf{x}}_{i\cdot}'\beta),$$

where $\mathrm{E}_{\mathrm{M},\psi}$ denotes the model-based conditional expectation.

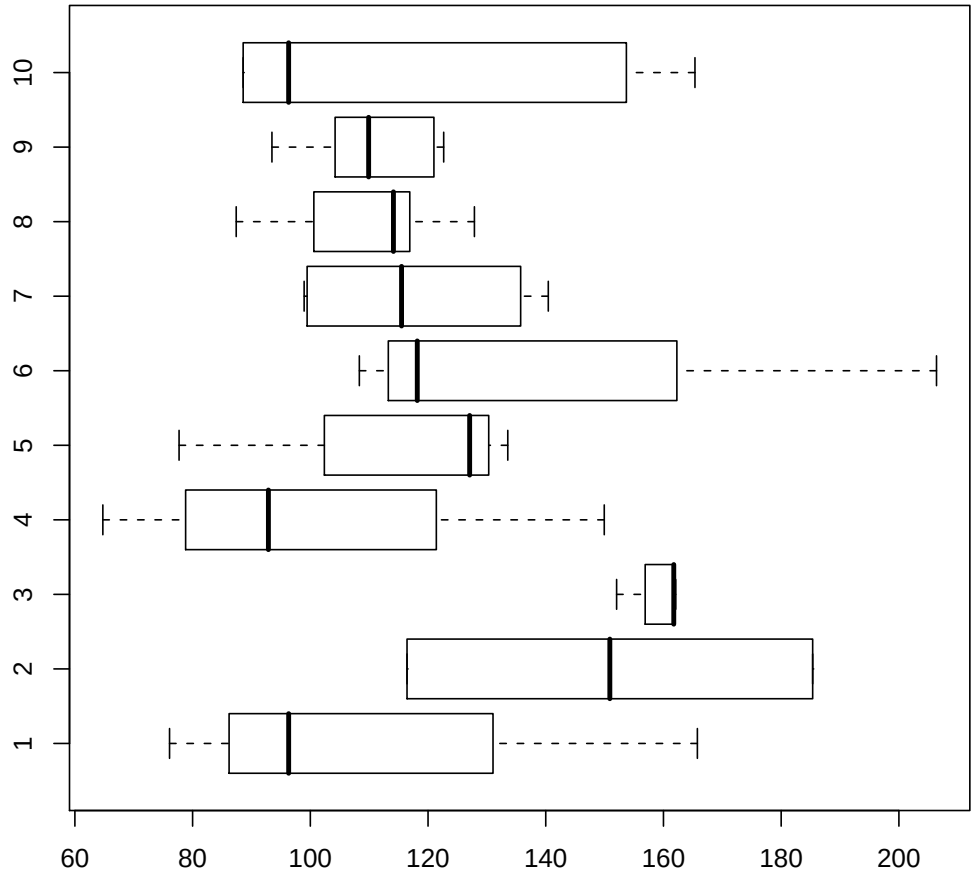
Nested-Error Regression (NER), cont.

- Under the NER model, the variance of Y_{ik} is a constant, $\sigma^2 = \sigma_v^2 + \sigma_e^2$. In practice, this assumption may not be valid.

- *Example:* Consider the corn data of Battese *et al.* (1988) mentioned above. To illustrate the within-area variation, we combine the first three counties (which have a single obs. within each county) to form the first subpopulation. The rest of the subpopulations consist of counties 4–12.

- Consider $y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + v_i + e_{ij}$, $i = 1, \ldots, 10, j = 1, \ldots, n_i$, where y_{ij} is the jth sampled hectare in area i; x_{ij1} and x_{ij2} are the corresponding numbers of pixels classified by the satellite as corn and soybeans, respectively.

Figure 1: Boxplots of the Iowa Crops Data



Heteroscedastic Nested-Error Regression (HNER)

- On the other hand, the expression of the BP depends only on the ratio of the variances, $\gamma = \sigma_v^2/\sigma_e^2$, rather than the variances themselves.

- In other words, the BP is unchanged even if σ_v^2, σ_e^2 depend on i, the index of the subpopulation, provided that $\gamma = \sigma_{v,i}^2/\sigma_{e,i}^2$ is a constant. This offers some potential flexibility in modeling the variance. The latter is called a heteroscedastic NER (HNER) model.

- More specifically, the following questions are of interest:

  (1) Under the HNER model, does the NER MLE of γ remain consistent? Note that γ is all we need in computing the BP.
  (2) The same question regarding the HNER MLE.

Heteroscedastic Nested-Error Regression (HNER), cont.

- Ignoring the heteroscedasticity can lead to inconsistent estimation of the within-cluster correlation, or equivalently, the variance ratio γ.
- The maximum likelihood estimators (MLEs) of the fixed effects and within-cluster correlation are consistent in a heteroscedastic nested-error regression (HNER) model with completely unknown within-cluster variances under mild conditions.
- See Jiang, J. and Nguyen, T. (2012), Small area estimation via heteroscedastic nested-error regression, *The Canad. J. Statist.* 40, 588-603.

Simulation Study

- Our theoretical study shows that the HNER MLE is consistent, while the NER MLE of γ may be inconsistent in a HNER situation.

- However, consistency is an estimation property. How much is the difference in the consistency property translated into that in terms of the predictive performance? We set up a simulation study to investigate.

- Consider the following simple model:
  $y_{ij} = \beta_1 + v_i + e_{ij}, i = 1, \ldots, m_1, j = 1, 2, 3$ and
  $y_{ij} = \beta_2 + v_i + e_{ij}, i = m_1 + 1, \ldots, m, j = 1, \ldots, 8$, where
  $m = 2m_1$.

- The true values of β_1, β_2 are 1 and -1, respectively.

Simulation Study, cont.

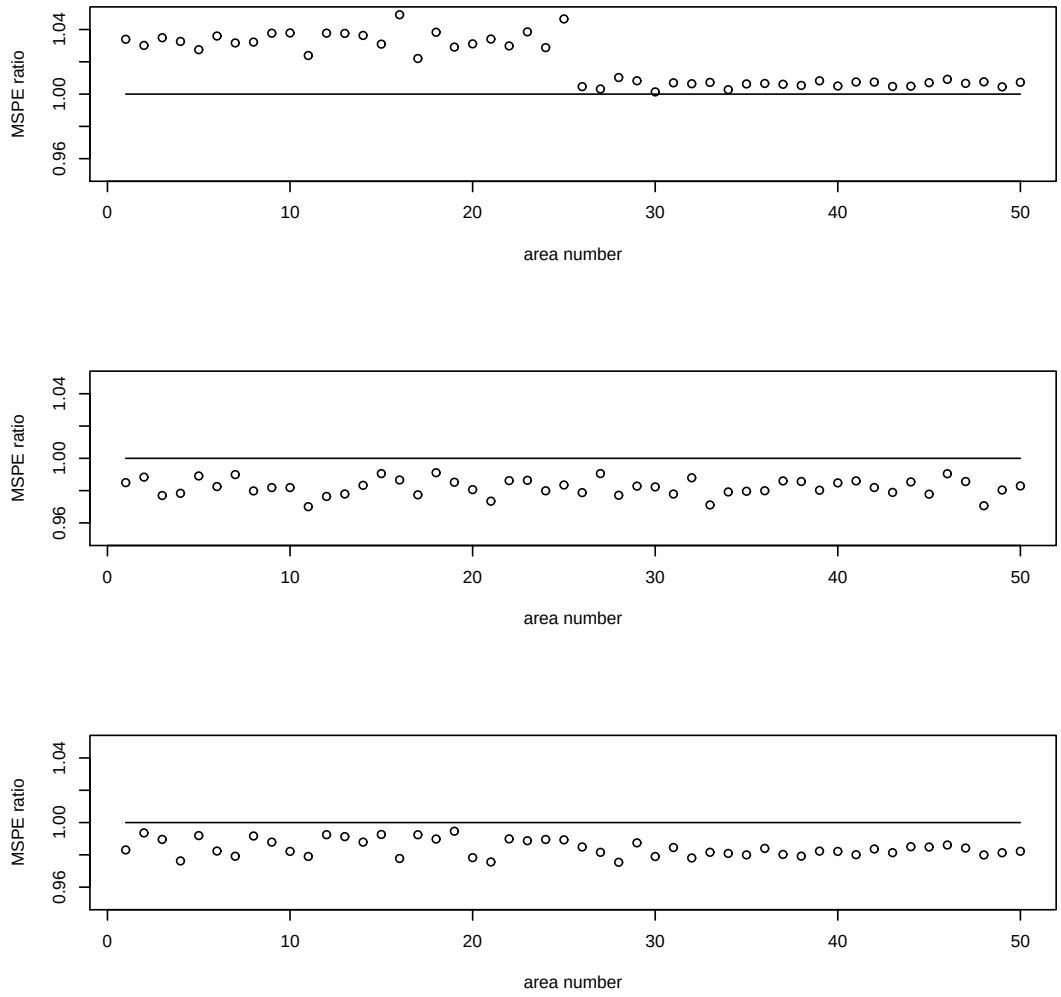
- The v_i's and e_{ij}'s satisfiy the assumption of the HNER model with the true value of γ equal to 1.

- Three scenarios of σ_i's are considered:
  (I) $\sigma_i = 0.2, 1 \leq i \leq m$;
  (II) $\sigma_i = 0.2, 1 \leq i \leq m_1$, and $\sigma_i = 0.8, m_1 + 1 \leq i \leq m$; and
  (III) $\sigma_i, 1 \leq i \leq m_1$ are generated from the Uniform$[0.2, 0.3]$ distribution, while $\sigma_i, m_1 + 1 \leq i \leq m$ are generated from the Uniform$[0.8, 0.9]$ distribution, in each simulation run.

- We consider $m = 50$ in this case. Due to the relatively large number of small areas, we present the results by plots.

- The MSPEs are evaluated over $K = 5000$ simulation runs.

Figure 2


MSPE ratio
0.96
1.00
1.04
0
10
20
30
40
50
area number
MSPE ratio
0.96
1.00
1.04
0
10
20
30
40
50
area number
MSPE ratio
0.96
1.00
1.04
0
10
20
30
40
50
area number

Measure of Uncertainty–Area Specific MSPE

- Although consistent estimators of $\sigma_i^2, 1 \leq i \leq m$ are not needed for (2) as a point predictor, it is a different story when it comes to measure of uncertainty.
- This is because the area-specific MSPE depends on not just β and γ (or ρ), but also on σ_i^2.
- Furthermore, when $\sigma_i^2, 1 \leq i \leq m$ are completely unknown, it is impossible to estimate them consistently no matter what method is used (this is because the effective sample size for estimating σ_i^2 is n_i, which is supposed to be bounded in SAE).

Measure of Uncertainty–Area Specific MSPE

- Therefore, we make an additional assumption that the σ_i^2's can be treated as random variables. More specifically, we assume the following:

- *A1.* $\sigma_i^2, 1 \leq i \leq m$ are random variables so that there is a known division, $\{1, \ldots, m\} = S_1 \cup \cdots \cup S_q$, such that $\mathrm{E}(\sigma_i^2) = \phi_t, i \in S_t, 1 \leq t \leq q$, where $\phi_1, \ldots, \phi_q$ are unknown.

- *A2.* Conditional on $\sigma_i^2, 1 \leq i \leq m$, we have the HNER.

- *A3.* $y_i, i = 1, \ldots, m$ are marginally independent.

- Under assumptions *A1*—*A3*, a second-order unbiased area-specific MSPE can be obtained by using the jackknife method of Jiang, Lihiri & Wan (2002).

Partial Results of MSPE Estimation

$m = 20$				$m = 50$			
Area	MSPE	$\widehat{\text{MSPE}}$	%RB	Area	MSPE	$\widehat{\text{MSPE}}$	%RB
1	.0179	.0244	36.3	1	.0174	.0180	3.4
2	.0194	.0242	25.0	2	.0170	.0179	5.3
3	.0196	.0242	23.8	3	.0167	.0180	7.8
4	.0186	.0246	32.4	4	.0161	.0179	11.5
5	.0192	.0240	25.0	5	.0182	.0183	0.2
11	.0861	.0963	11.8	26	.0818	.0837	2.2
12	.0838	.0967	15.4	27	.0792	.0837	5.8
13	.0902	.0989	9.6	28	.0807	.0835	3.6
14	.0810	.0944	16.6	29	.0823	.0838	1.8
15	.0799	.0973	21.7	30	.0766	.0838	9.4

Iowa crops data (revisited)

- Recall that, for the Iowa crops data, we combine the first three counties, which have a single observation for each county, to form the first small area.
- One reason for doing so is to make sure that the conditions for our theorems [omitted; see Jiang and Nguyen (2012)] are satisfied.
- The HNER MLEs for $\beta_k, k = 0, 1, 2$ and γ are found to be 67.78, 0.24, -0.14, and 0.79, respectively.

  As a comparison, the corresponding NER MLEs are 19.72, 0.36, -0.03, and 0.12, respectively.

Notes

- An inspection of the sample variances suggests two groups: those above 1000 and those below, that is, $S_1 = \{1, 2, 4, 6, 10\}$ and $S_2 = \{3, 5, 7, 8, 9\}$.
- This is also supported by the boxplots (Fig. 1).
- Thus, $q = 2$ in this case. The jackknife MSPE estimates are obtained, and the square roots of the MSPE estimates are reported as measures of uncertainty.
- As comparisons, the EBLUPs based on the NER MLEs and the square roots of their jackknife MSPE estimates (Jiang *et al.* 2002) are also reported.

Iowa crops data revisited

EBLUPs and measures of uncertainty (areas 1–5):

Area	1	2	3	4	5
EBLUP	113	111	141	107	110
$\sqrt{\widehat{\mathrm{MSPE}}}$	15.1	15.0	12.6	14.0	13.1
EBLUP_1	120	116	134	107	117
$\sqrt{\widehat{\mathrm{MSPE}}_1}$	8.9	11.4	15.1	10.0	9.4

Iowa crops data revisited, cont.

EBLUPs and measures of uncertainty (areas 6–10):

Area	6	7	8	9	10
EBLUP	122	113	120	106	128
$\sqrt{\widehat{\mathrm{MSPE}}}$	11.9	9.7	12.4	10.5	13.1
EBLUP_1	122	110	125	115	131
$\sqrt{\widehat{\mathrm{MSPE}}_1}$	8.9	10.8	9.0	14.0	8.5

Notes

- It is seen that, while the values of the two EBLUPs are fairly close, the HNER-based MSPE estimates are larger than the NER-based MSPE estimates for most of the small areas.
- This seems to make sense, as the NER based estimation has ignored the heteroscedasticity altogether.
- For example, the HNER-based MSPE estimates are larger than the NER-based MSPE estimates for all of the small areas in group S_1 (those whose sample variances are above 1000);
- The HNER-based MSPE estimates are smaller than the NER-based ones for 3 out 5 small areas in group S_2 (those whose sample variances are below 1000).
- In fact, the two largest NER-based MSPE estimates (areas 3 and 9) both occur in S_2, which seems a bit counterintuitive, especially in view of Fig. 1.