When are BLUPs Bad

Ed Stanek  UMass- Amherst
Julio Singer  USP- Brazil
George Reed  UMass- Worc.
Wenjun Li  UMass- Worc.

Outline

• Description of the Problem
  – Examples
• Basic Models & Ideas
• Approaches
• Results
• Conclusions/Further Research
Problem: Estimate/Predict Cluster Latent Value

Example: 11 Middle Schools
8-16 Classrooms/School
Question: How much bullying occurs at school?

Example: 166 School Districts
6-18 Schools/District
Question: Does the District have a sun-protection policy for students?

Example: 25,000 Members in an HMO
30 Days/Member before Cholesterol measure
Question: What is a Patient’s Ave Sat fat intake (g)?

Description of the Problem

Clustered Finite Population (cluster=group of subjects)

Data is available only on some clusters (schools)
Observational studies, multi-stage samples, group-randomized trials

For selected clusters, response is observed on a subset of subjects (students) (possibly with response error).

How to Best Estimate/Predict the Cluster’s Latent Value?
Basic Models & Ideas

Simple Response Error Model for a Subject:

School: \( s = 1, \ldots, N \)
Student: \( t = 1, \ldots, M \)

\[
Y_{stk} = Y_{st} + W_{stk}
\]

where

\[
E(W_{stk}) = 0
\]

Combining terms, the Response Error Model:

\[
Y_{stk} = \mu + \beta_s + \epsilon_{st} + (W_{stk})
\]

becomes

\[
Y_{ijk} = \mu + \left( B_i + E_{ij} + W_{ijk}^* \right)
\]

where

ith school effect \( B_i = \sum_{s=1}^{N} U_{is} \beta_s \)

jth student effect \( E_{ij} = \sum_{s=1}^{N} \sum_{t=1}^{M} U_{is} U_{jt} \epsilon_{st} \)

kth response error \( W_{ijk}^* = \sum_{x=1}^{N} \sum_{t=1}^{M} U_{is} U_{jt} W_{stk} \)
Model Properties

\[ Y_{ijk} = \mu + B_i + \frac{E_y}{\sigma_y} + W_{ijk} \]

Over Schools
\[ E \left( B_i \right) = 0 \quad \text{var} \left( B_i \right) = \sigma^2 \]

Over Students
\[ E \left( E_y \right) = 0 \quad \text{var} \left( E_y \right) = \sigma^2_e \]

The Latent Value of the School in the ith position:

\[ \mu + B_i = \sum_{s=1}^{N} U_{is} \mu_s \]

Approaches to Estimate/Predict Latent Value

1. Fixed Models
2. Mixed Models
3. Super-population Models
4. Random Permutation Models
1. Fixed Effect Models:

$$Y_{ij} = \mu + \beta_j + E_{ij}$$

(Random Sample of Students)

Latent Cluster Mean:

$$\mu + \beta_j$$

(i.e. School "s")

OLS Estimate:

$$\bar{Y}_s = \frac{1}{m} \sum_{j=1}^{m} Y_{sjk}$$

2. Mixed Models

$$Y_{ij} = \mu + B_i + E_{ij}$$

$$Y = X\mu + (ZB + E)$$

$$\text{var}(B) = G$$

$$\text{var}(E) = R$$

$$\Sigma = \text{var}(ZB + E) = ZGZ' + R$$

To predict $$\mu + B_i$$

Solution:

$$\hat{\mu} = \left(X\Sigma^{-1}X'\right)^{-1} X\Sigma^{-1}Y$$

$$\hat{B} = GZ'\Sigma^{-1}(Y - X\hat{\mu})$$

BLUP = $$\hat{\mu} + \hat{B}_i = \bar{Y} + k\left(\bar{Y}_i - \bar{Y}\right)$$

where

$$k = \frac{\sigma^2}{\sigma^2 + \sigma_e^2 / m}$$
3. Super-population Models

Super-population | Finite Population (Realization)
---|---
\[ Y_{ij} = \mu + B_i + E_{ij} \] | \[ y_{st} = \mu + \beta_s + \epsilon_{st} \]

Latent Value | Latent Value
---|---
\[ \mu + B_i = \frac{1}{M} \left( \sum_{j=1}^{m} Y_{ij} + \sum_{j=m+1}^{M} Y_{ij} \right) \] | \[ \mu_s = \frac{1}{M} \sum_{i=1}^{M} y_{st} \]

Breaking Up the Latent Value Parameter

\[ \mu + B_i = \frac{1}{M} \left( \sum_{j=1}^{m} Y_{ij} + \sum_{j=m+1}^{M} Y_{ij} \right) \]

\[ \mu + B_i = \frac{1}{M} \left[ \frac{\sum_{j=1}^{m} Y_{ij}}{m} + \left( M - m \right) \frac{\sum_{j=m+1}^{M} Y_{ij}}{M - m} \right] \]

\[ \mu + B_i = f \overline{Y}_{it} + (1-f) \overline{Y}_{it} \]

where \[ f = \frac{m}{M} \]
Superpopulation Model Prediction Approach

(Scott and Smith, 1969)

Sample: \[ \bar{Y}_{i,t} = \frac{1}{n} \sum_{j=1}^{n} Y_{ij} \]

Remainder: \[ \bar{Y}_{i,II} = \frac{1}{N-n} \sum_{j=n+1}^{N} Y_{ij} \]

\[ \mu + B_i = f\bar{Y}_{i,t} + (1-f)\bar{Y}_{i,II} \]

BLUP: \[ f \bar{Y}_i + (1-f) \hat{\bar{Y}}_{i,II} \]
\[ \hat{\bar{Y}}_{i,II} = \bar{y} + k (\bar{y}_i - \bar{y}) \]
\[ k = \frac{\sigma^2}{\sigma^2 + \sigma^2_e / m} \]

4. Random Permutation Model Predictors

Finite Population: \[ y_{st} = \mu + \beta_s + \epsilon_{st} \]

2-Stage Permutation: \[ Y_{ij} = \mu + B_i + E_{ij} \]

To Predict: \[ \mu + B_i = f\bar{Y}_{i,t} + (1-f)\bar{Y}_{i,II} \]

BLUP \[ f \bar{Y}_i + (1-f) \hat{\bar{Y}}_{i,II}^* \]
\[ \hat{\bar{Y}}_{i,II}^* = \bar{y} + k^* (\bar{y}_i - \bar{y}) \]

where \[ k^* = \frac{\sigma^2}{\sigma^2 + \sigma^2_e / m} \]
\[ \sigma^2 = \sigma^2 - \frac{\sigma^2_e}{M} \]
Approaches

1. Fixed: \( \bar{y}_i \)

2. Mixed: \( [\bar{y} + k(\bar{y}_i - \bar{y})] \)

3. Superpop. \( f \bar{y}_i + (1 - f) [\bar{y} + k(\bar{y}_i - \bar{y})] \)

4. RP Model \( f \bar{y}_i + (1 - f) [\bar{y} + k^* (\bar{y}_i - \bar{y})] \)

\[
k^* = \frac{\sigma^2}{\sigma^2 + \sigma_e^2 / m}
\]

\[
k^* = \frac{\sigma^2}{\sigma^2 + \sigma_e^2} / m
\]

\[
\sigma^2 = \sigma^2 - \frac{\sigma_e^2}{M}
\]

If variance components are known, then the RP Model Predictor is “best”-
(i.e. has small expected MSE)

How much better depends on:

Sampling fraction: \( f = \frac{m}{M} \)

Cluster-intra-class correlation: \( \rho(s) = \frac{\sigma^2}{\sigma^2 + \sigma_e^2} \)

Unit- Reliability: \( \rho(t) = \frac{\sigma_e^2}{\sigma_e^2 + \sigma_R^2} \)
Figure 1a. Percent Increase in Expected MSE for Mixed Model (—) and Scott and Smith Model (---) Predictors Relative to the Random Permutation Model Predictors of the Latent Value of a Realized Sample PSU

f=0.1  \(\rho(s)=0.1\)

Percent Increase in EMSE

Unit Intra-Class Correlation (\(\rho(t)\))

Figure 1b. Percent Increase in Expected MSE for Mixed Model (—) and Scott and Smith Model (---) Predictors Relative to the Random Permutation Model Predictors of the Latent Value of a Realized Sample PSU

f=0.5  \(\rho(s)=0.1\)

Percent Increase in EMSE

Unit Intra-Class Correlation (\(\rho(t)\))
What happens when variances must be estimated?

Are BLUPs still best?

Not always.
Simulation Study:
Two-stage finite population sampling

Settings:  N, M, n, m

Variance Components:  $\rho(s), \rho(t), \sigma^2, \sigma^2_e, \sigma^2_k$

Distribution of Clusters:  normal, uniform, beta, gamma

Unit Variance Prop Cluster Mean: yes/no

Simulation Results (No response Error)

N=3
N=5
N=10
N=25
N=50
N=100
N=500
N=1000
Simulation Results (with Response Error)
N=100, n=10, M=25, m=10

Clusters Units
Normal Normal
Gamma(2) Normal

Summary and Conclusions

eBLUPs can have higher MSE than the Cluster Ave
Small #s of sample clusters (n<5)

eRP predictors (while theoretically best), may have higher MSE than MM empirical predictors, or Super-population predictors.

Properties don’t seem to depend on distributions.

Simulations are limited- fall short of providing guidelines
Thanks