

PREDICTING REALIZED RANDOM EFFECTS IN CLUSTER SAMPLES OF FINITE POPULATIONS

Edward J. Stanek III

Department of Biostatistics and Epidemiology, UMASS, Amherst, MA

Julio D. Singer

Department of Statistics, University of São Paulo, Brazil

Example 1:

In a 6-week period, what is the average daily Saturated Fat intake for a subject?

Data: 3 24-hr diet recalls on study subjects

Example 2:

In a week, how much Leisure time Physical Activity does a subject have?

In a month??

Data: 3 24-hr Physical Activity recalls per subject

Problem Summary

Interest: Parameter for a cluster (an average).

$$\mu_s = \frac{1}{M} \sum_{t=1}^M y_{st}$$

Data: Observe “part” of the parameter on some clusters.
(via 2-stage cluster sampling)

Questions:

How do we ‘estimate’ (or predict) the parameter for a ‘realized’ cluster?

Does the ‘time period’ that defines the parameter matter?

What do people usually do?

1. Use a Fixed Effect Model and the Sample Mean:

For subject “ s ” in the sample:
$$\bar{Y}_s = \frac{\sum_{j=1}^m Y_{sj}}{m}$$

2. Use Mixed Models and a BLUP:

For the i^{th} selected subject:
$$\hat{p}_i = \hat{\mu} + k_i (\bar{Y}_i - \hat{\mu})$$

3. Other???

Table 1. Sample Data for ID=677

<u>Sample Date</u>	<u>Sat. Fat Intake gm/d</u>	<u>Leisure METs hrs/d</u>
11 SEP 95	89.33	0.0
16 SEP 95	34.10	5.5
18 SEP 95	<u>46.74</u>	<u>3.5</u>
Average	56.7 (28.9)	3.0 (2.8)
BLUP	45.6 (6.19)	2.4 (1.4)

To Construct the **BLUP** (n=567, m=3):

Saturated Fat (gm/d):

Overall Mean: 25.2 g

Shrinkage Constant (K): 0.65

$$\text{BLUP} = 25.2 + 0.65 (56.7 - 25.2) = 45.6$$

Leisure Activity (METs/d):

Overall Mean: 1.95 g

Shrinkage Constant (K): 0.44

$$\text{BLUP} = 1.95 + 0.44 (3.0 - 1.95) = 2.41$$

To get the shrinkage constant:

	<u>Subject Variance</u>	<u>Day to day Variance</u>
Sat Fat:	108.2	175.3
Leisure Time:	3.2	12.3

$$k = \frac{\sigma^2}{\sigma^2 + \frac{\sigma_e^2}{m}}$$

Do the estimators (predictors) change with different targets? No!

How might we account for “Finiteness”?

Start with the parameter for Subject s :

$$\begin{aligned}\mu_s &= \frac{1}{M} \sum_{t=1}^M y_{st} \\ &= \frac{1}{M} \left(\sum_{t \in \text{sample}} y_{st} + \sum_{t \notin \text{sample}} y_{st} \right).\end{aligned}$$

Predict values “not observed” for a realized subject.

Use a Mixed Model: $Y_{ij} = (\mu + B_i) + E_{ij}$

$T_i = \mu + B_i =$ Mean for i^{th} selected subject

Strategy

Assumptions: 2-stage cluster sampling
without replacement, equal size clusters

Random Permutation model

i.e. indicator RV's from sampling:

$$Y_{ij} = \sum_{s=1}^N \sum_{t=1}^M U_{is} U_{jt}^{(s)} y_{st}$$

Target RV:
$$T_i = \sum_{s=1}^N U_{is} \mu_s$$

Summarize RVs in the Population:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{Z}\mathbf{B} + \mathbf{E}$$

Use Prediction Approach (Royall, 1976)

1. Divide \mathbf{Y} into Sample, and Remainder
2. Require Predictor to be:
 - i. Linear function of sample
 - ii. Unbiased predictor of T_i
3. Find coefficients that minimize EMSE

Results:

$$RP - Model : \hat{T} = \frac{m}{M} \bar{Y}_i + \left(\frac{M - m}{M} \right) \left[\bar{Y} + k (\bar{Y}_i - \bar{Y}) \right]$$
$$BLUP : \hat{p}_i = \bar{Y} + k_i (\bar{Y}_i - \bar{Y})$$

Example:		Time Period		
		7 Days	30 Days	1 Year
Sat. Fat (g/d):	Mixed Model:	45.6	45.6	45.6
	RP-Model:	49.3	46.4	45.7
Leisure:	Mixed Model:	2.41	2.41	2.41
	RP-Model:	2.56	2.44	2.41

Interpretation: BLUPs predict 'unobserved' values!

What about when there is response error?

For the Random Permutation Model:

$$\hat{T} = \left(\frac{m}{M}\right) \left[\bar{Y} + k_r^* (\bar{Y}_i - \bar{Y}) \right] + \left(\frac{M-m}{M}\right) \left[\bar{Y} + k^* (\bar{Y}_i - \bar{Y}) \right]$$
$$\hat{T} = \left(\frac{m}{M}\right) \bar{Y}_i + \left(\frac{M-m}{M}\right) \left[\bar{Y} + k (\bar{Y}_i - \bar{Y}) \right]$$

where

k_r^* = Shrinkage constant depending on Resp. error

k^* = Shrinkage constant depending on SSU and Resp.
Error

Why is this Important?

Real Problems are defined for Real Populations.
(often finite)

Results are intuitive .

Minimal Assumptions are required-
(2-stage sampling)

Results enable Mixed models to be used in Finite
Population Settings

Extensions have recently been developed
Unbalanced cluster sizes
SRS-simple addition of covariate.

Outstanding Questions:

Extensions: Covariates in 2-stage sampling
Other sampling designs
Missing data, etc

What is meant by a 'Realized Random Effect'?

Is there a difference between:

the mean for a selected cluster (i^{th} selection)

the mean for cluster s

Is a richer random variable framework needed?