Checking Simulation Program for Theoretical MSE of Simple Mean, RP Model Predictor, and Expanded Model Predictor of a Realized Random Cluster Means

Ed Stanek

Introduction

We describe programs and a check of the expressions for the theoretical MSE of predictors in two stage cluster sampling. This check is motivated by an apparent mistake in the simulation. Using the simulation programs ced07p20.sas, and the simulations indicated by ced07p25.sas, we evaluate the predictor in many possible samples, and also the theoretical MSE of the predictor. If the average squared error (MSE) is calculated based on the predictor for the various samples, we should end up with an expression equal to the theoretical MSE.

Comparison of the MSE based on multiple sample Predictors with the Theoretical MSE

In the program ced07p25.sas, we compute a term that is supposed to correspond to a standardized difference between the simulated MSE and the theoretical MSE. In our check, we first simply compare the average squared difference between the predictor and the true value with the theoretical MSE. We do this in the program ced07p27.sas.

Table 2 below summarizes what should be the average squared error from the simulation, and the theoretical MSE from formula. Notice that for the sample mean, the results agree. For the others, the results do not agree. We investigate why there is this difference next.

Source: ced07p27.sas on 6/18/2007 by EJS
Table 2x. List of theoretical MSE for predictors

<table>
<thead>
<tr>
<th>Obs</th>
<th>trial_n</th>
<th>P1_TMSE</th>
<th>PSMSE1</th>
<th>P2_TMSE</th>
<th>PSMSE2</th>
<th>P3_TMSE</th>
<th>PSMSE3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25000</td>
<td>0.82010</td>
<td>0.53648</td>
<td>2.25</td>
<td>2.24991</td>
<td>1.78448</td>
<td>1.68948</td>
</tr>
</tbody>
</table>

Evaluation of the RP MSE

We first consider expressions for the theoretical RP MSE, and try to resolve why the simulated values differ. From the program ced07p27.sas, we compute the theoretical MSE for the RP model as

\[ \text{IF f_sd=0 THEN p3_tmse} = (1-fs)*\left(\frac{v_e}{(\&sn*sm)}\right) + ((\&sn-1)/\&sn)*(1-k_rp)*v_c; \]

This expression in given on page 13 of c07ed25.sas, and attributed to the manuscript by Stanek and Singer (2004), such that

\[ \text{MSE}(\hat{t}_b) = (1-f)\left[\frac{\sigma_e^2}{nm} + \left(\frac{n-1}{n}\right)(1-k)\sigma^2\right] \]
with \( k = \frac{m \sigma'^2}{m \sigma'^2 + \sigma^2} \) where \( \sigma'^2 = \sigma^2 - \frac{\sigma^2}{M} \), \( \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2 \), \( \bar{Y}_i = \frac{\sum_{j=1}^{m} Y_{ij}}{m} \), and \( \bar{Y} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij}}{nm} \).

We can check the results of the theoretical MSE directly. For the simulation, parameters correspond to those given in Table 1a. In addition, \( \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2 = 5 \), \( \sigma'^2 = \sigma^2 - \frac{\sigma^2}{M} = 5 - \frac{5}{20} = \frac{95}{20} = 4.75 \) so that

\[
k = \frac{m \sigma'^2}{m \sigma'^2 + \sigma^2} = \frac{2 \left( \frac{95}{20} \right)}{2 \left( \frac{95}{20} \right) + 5} = \frac{9.5}{14.5} = 0.6551724.
\]

Using these results,

\[
MSE(\hat{T}_B) = (1 - f) \left[ \frac{\sigma^2}{nm} + \left( \frac{n-1}{n} \right) (1-k) \sigma^2 \right]
= \left( 1 - \frac{2}{20} \right) \left[ \frac{5}{5(2)} + \left( \frac{5-1}{5} \right) (1- \frac{9.5}{14.5}) \right]
= 0.9 \left[ 0.5 + \left( \frac{4}{5} \right) \left( \frac{5}{14.5} \right) \right]
= 0.9 \left[ 0.5 + \frac{20}{14.5} \right]
= 0.9 \left[ \frac{27.25}{14.5} \right]
= 1.691379
\]

Notice that this expression matches the average MSE based on the simulated values of the RP predictor, but not the theoretical MSE from the simulation. This indicates that there is an error in the program that calculated the theoretical MSE.

Table 1a. Data that Define Cluster, Unit, and Sample Sizes for the Population (first 10 clusters)

<table>
<thead>
<tr>
<th># Clusters= 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>s</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>M(s)</td>
</tr>
<tr>
<td>m(s)</td>
</tr>
<tr>
<td>v(s)</td>
</tr>
</tbody>
</table>

Max M(s)= 20   Sample Clusters= 5

Ave f(s) 0.10  Std f(s) 0.00
Ave k(s) 0.91  Std k(s) 0.00

Source: ced07p27.sas on 6/18/2007 by EJS
Simulation Validation

We discovered in the process an error in the computer code such that a macro variable was named in one case ns, and in other case sn. This caused the mistakes. We corrected the error in the program ced07p28.sas. Below is the result of a simulation of 25000 trials that compares the theoretical MSE with the MSE based on the average squared errors of the predictors.

Table 2x. List of theoretical MSE for predictors

<table>
<thead>
<tr>
<th>Obs</th>
<th>trial_n</th>
<th>P1_TMSE</th>
<th>PSMSE1</th>
<th>P2_TMSE</th>
<th>PSMSE2</th>
<th>P3_TMSE</th>
<th>PSMSE3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25000</td>
<td>0.53412</td>
<td>0.53648</td>
<td>2.25</td>
<td>2.24991</td>
<td>1.69138</td>
<td>1.68948</td>
</tr>
</tbody>
</table>

Notice that the theoretical MSE and the empirical MSE are nearly the same for each setting.