An approach to estimation and prediction with measurement error under simple random sampling

Ruitao Zhang, Ed stanek

Abstract

Measurement error in survey sampling refers to the error in survey responses arising from the method of data collection. How to predict the expected value of a sample unit (PSU) with measurement error is a basic but important problem that has not been addressed clearly before. The main challenge in this problem is that the notation used in previous work does not clearly distinguish whether the measurement error is associated with labeled subjects or sample indices. In this paper, we will explore this problem under simple random sampling with measurement error in a finite population. We first use sampling indicator random variables to define an expanded random permutation model. Then, we separate the permuted data into a sample and a remainder. Finally, we define the best linear unbiased predictor (BLUP) of a first-stage sample unit (PSU) by using the criteria that the predictor (Royall, 1976): 1. is linear in the sample. 2. is unbiased. 3. has minimum mean square error. One advantage of this method is that it gives smaller mean square error than other predictors proposed in the literature, including random permutation model. Another advantage is that it can clearly distinguish whether the measurement error is associated with labeled subjects or sample indices. We will explore how the estimator performs in practice in future work.

Introduction

- Motivation: Predicting the latent value of a sample subject when there is measurement error.
- Example: Predicting the latent value of a sample from a population with three subjects.
- Two errors exist: 1. from subjects. 2. from interviewers

Table 1. Pulse Rates for a Population of Three Subjects

<table>
<thead>
<tr>
<th>Sample Position (i)</th>
<th>Subject Label (s)</th>
<th>Pulse rates Per minute</th>
<th>Response Variance $\sigma_{se}^2$</th>
<th>Interviewer Variance $\sigma_{ie}^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>72</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>75</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>80</td>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

Methods

- Measurement error model for a population with $N$ subjects:
  \[ Y_{sk} = y_s + W_{sk}, s = 1, \ldots, N. \]
• Sampling indicator random variables: \( U_{is} \), which equals 1 when unit \( s \) is in position \( i \). Define \( Y_i = \sum_{s=1}^{N} U_{is} Y_s \).

• Random permutation model: \( Y_{ik}^* = \sum_{s=1}^{N} U_{is} Y_{sk} = Y_i + W_{ik}^* \), where \( W_{ik}^* = \sum_{s=1}^{N} U_{is} W_{sk} \).

• Expanded random permutation model: \( \bar{Y}^* = vec \begin{pmatrix} U_{11}Y_{1k} & U_{21}Y_{1k} & \cdots & U_{N1}Y_{1k} \\ U_{12}Y_{2k} & U_{22}Y_{2k} & \cdots & U_{N2}Y_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ U_{1N}Y_{Nk} & U_{2N}Y_{Nk} & \cdots & U_{NN}Y_{Nk} \end{pmatrix} \).
\[
\tilde{Y}^* = \text{vec} \left( \begin{array}{cccc}
U_{11}Y_{1k} & U_{21}Y_{1k} & \cdots & U_{N1}Y_{1k} \\
U_{12}Y_{2k} & U_{22}Y_{2k} & \cdots & U_{N2}Y_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
U_{1N}Y_{Nk} & U_{2N}Y_{Nk} & \cdots & U_{NN}Y_{Nk}
\end{array} \right) = \begin{pmatrix}
U_{11}Y_{1k} \\
\vdots \\
U_{1N}Y_{Nk} \\
U_{N1}Y_{1k} \\
\vdots \\
U_{NN}Y_{Nk}
\end{pmatrix}
\]

- Sample and the remainder: 
  \[
  \begin{pmatrix}
  \tilde{Y}_{II}^* \\
  \tilde{Y}_{Il}^*
  \end{pmatrix} = \begin{pmatrix}
  \left( I_n \otimes I_N \right) & \left( 0_{n(N-n)} \otimes I_N \right) \\
  \left( 0_{(N-n)n} \otimes I_{N-n} \right) & \left( I_{N-n} \otimes I_N \right)
  \end{pmatrix} \begin{pmatrix}
  \tilde{Y}^* \\
  \tilde{Y}^*
  \end{pmatrix}.
  \]

- Target: 
  \[P_i = g^t(\tilde{Y}_{II})\text{, where } g^t = (g^t_I \quad g^t_{II}) = (e^t_{il} \otimes I_N^t \quad e^t_{II} \otimes I_N^t)\text{.}\]

- Criteria for BLUP: 1. Linear in the sample. 2. Unbiased. 3. Minimum MSE.

Results

\[\hat{P}_i = Y_{ik}^* - (N-1) \left( k_i^* Y_{ik}^* + \frac{1}{n} \sum_{i=1}^{n} k_i^* Y_{ik}^* \right) \text{ for } i \leq n \text{ and } \hat{P}_i = \frac{1}{n} \sum_{i=1}^{n} Y_{ik}^* \text{ for } i > n\text{, where } Y_{ik}^* = \sum_{s=1}^{N} U_{is}Y_{sk}.\]
Conclusion

Advantage

- Smaller mean square error

\[
\text{var}(\hat{P}_i - P_i) = \frac{(N-1)^2 (n-1)}{nn^2 (N-1)} \left[ \sum_{x=1}^{N} \left( \sigma_{se}^2 k_x, y_x + \frac{\sum_{x=1}^{N} \sigma_{se}^2 k_x}{N(1-k)} k_x, y_x \right) \right]^2 - \frac{1}{N} \left[ \sum_{x=1}^{N} \left( \sigma_{se}^2 k_x, y_x + \frac{\sum_{x=1}^{N} \sigma_{se}^2 k_x}{N(1-k)} k_x, y_x \right) \right]^2
\]

\[- \frac{(N-1)}{(n-1)N^2} \left[ \sum_{x=1}^{N} \sigma_{se}^2 k_x + \frac{\left( \sum_{x=1}^{N} \sigma_{se}^2 k_x \right)^2}{N(1-k)} \right] + \frac{1}{N} \sum_{x=1}^{N} \sigma_{se}^2 \text{ for } i < n.\]

\[
\text{var}(\hat{P}_i - P_i) = \frac{N + Nn - 4n}{N(N-1)n} \left[ \sum_{x=1}^{N} y_x^2 - \frac{1}{N} \left( \sum_{x=1}^{N} y_x \right)^2 \right] \text{ for } i > n.
\]

- Clearly distinguish the measurement error associated with labeled subjects and sample indices.

Limitation

- Don’t know how the estimator performs in practice

Acknowledgement