Model-Assisted Domain Estimation

Dan Liao and Phillip Kott

RTI International

Federal Committee on Statistical Methodology (FCSM)
Research Conference
November 4-6, 2013
Washington Convention Center
Introduction

- Domain Estimation: estimation of population quantities (e.g. totals or means) for the desired population subgroups in a descriptive survey

- Context: Design-based estimation
  - the randomness is introduced by the sampling design
  - mainly used for domains whose sample size is reasonably large (for small domains, small area estimation is often used)

- Use of auxiliary information: model-assisted approach (Särndal et al., 1992)
Use of Auxiliary Data

- With high-quality auxiliary information, it is possible to obtain better accuracy for domain estimates.
  - accurate
  - moderately or highly correlated with the domain variables

- Different types of auxiliary data
  - *population-level* aggregates (e.g. from population census, other official statistics)
  - *unit-level* auxiliary data (e.g. from administrative records)
  - *domain-level* aggregates (e.g. from State registers)
  - *intermediate-level* aggregates (e.g. from first-phase sample surveys)
Notations

Let

- $U$ be the population (of $N$ elements)
- $S$ the sample (no nonresponse for simplicity)
- $y_k$ the survey value of interest for element $k$
- $x_k$ a vector of calibration variables
- $d_k$ a domain-membership indicator
- $w_k$ calibration weights for which $\sum_S w_k x_k = \sum_U x_k$
We are interested in estimating the population mean in the domain,

\[ U: M_d = \frac{\sum_U d_k y_k}{\sum_U d_k} \]

- with either the **calibration estimator**

\[ S: m_{d, ca} = \frac{\sum_S w_k d_k y_k}{\sum_S w_k d_k} \]

- or the **model-assisted estimator**

\[ S: y_k \sim x_k^T b \]

\[ U: m_{d, ma} = \frac{\sum_U d_k x_k^T b}{\sum_U d_k} = \frac{\sum_U d_k x_k^T [\sum_S (w_j x_j x_j^T)^{-1} \sum_S w_j x_j y_j] / \sum_U d_k}{\sum_U d_k} \]

(Design weights often replace calibration weights)
Application 1: Combing Information from Administrative Records with Sample Surveys

Sample Survey
- $X$
- $y$
- Design Weight $\times$ Adjustment

Calibration Estimator

$y = X^T b$

Administrative Records
- $X$

Model-Assisted Estimator

$X^T b = \hat{y}$
Bias Measure

- **Calibration estimator**, $m_{d, ca}$, is **design consistent** (if the sample size in the domain is large enough).

- **Model-assisted estimator**
  
  - When there is a $\lambda$ such that $\lambda^T x_k = d_k$ for all $k$ (e.g., when $d_k$ is a component of $x_k$),
    
    $$m_{d, ca} \approx m_{d, ma} \quad (\approx \text{means asymptotically equal})$$
  
  - Otherwise, **model-assisted estimator**, $m_{d, ma}$, is nearly unbiased (in some sense) when:
    
    $$E(y_k | x_k, d_k) = x_k^T \beta.$$
If the model is correct in the domain (H₀), the idealized test statistic:

\[
T^* = \frac{\sum_S w_k d_k (y_k - x_k^T \beta)}{\sum_S w_k d_k}
\]

has expectation (nearly) zero.

- Estimated test statistic:

\[
T = \frac{\sum_S w_k d_k (y_k - x_k^T b)}{\sum_S w_k d_k}
= \frac{\sum_S w_k d_k q_k}{\sum_S w_k d_k}
\]

This can be treated as a calibrated mean and the estimated variance be computed with WTADJUST in SUDAAN.
Variance Estimation

- **Calibration Estimator**

  \( m_{d, ca} \) is a calibrated mean within a domain, estimating its variance is straightforward with WTADJUST.

- **Model-Assisted Estimator**

  \[
  \text{var}(m_{d, ma}) = \left( \sum_s w_j d_j \right)^{-2} \text{var}(\sum_s w_k z_k),
  \]

  where \( z_k = \sum_U d_j x_j^T \sum_s (w_j x_j x_j^T)^{-1} x_k (y_k - x_k^T b). \)

  - \( \text{var}(\sum_s w_k z_k) \) can be estimated with WTADJUST
Example: 2010 Natality Data

- Data File: 2010 Natality Public Use File
  - Excluding foreign residents
  - Excluding records with missing values in the following variables:
    - DBWT: Birth Weight
    - UBFACIL: Facility Type
    - UPREVUS: Number of Prenatal Visits
    - COMBGEST: Gestational Age
    - MAGER: Mother’s Age
  - Select 1 out of 100 records (to reduce the data size)

- Population Size: N = 38,358
- Variable of Interest ($y_k$): Baby’s Birth Weight
Sample Selection

• 14 Strata:
  ➢ FACIL2 (2 facility types)
  ➢ GEST7G (7 gestational age groups)

  n=500 for each stratum in hospital; n=50 for each stratum in the other facility types

*FACIL2
1=Hospital; 2=Others (e.g. Freestanding Birthing Center or Clinic/Doctor’s Office, Residence)

*GEST7G
1=18-36 weeks, 2=37 weeks, 3=38 weeks, 4=39 weeks, 5=40 weeks, 6=41 weeks, 7=42+ weeks
Calibration Variable \( (x_k) \):

- Mother’s Race (four categories),
- Mother’s Age (continuous), and
- Infant Sex

Calibration Method: Generalized Raking

\[
w_k = w_k^{\text{original}} \exp(x_k^T b)
\]

(Other methods could have been used)
### Domain Estimates: Mother’s Race

- **Mother’s Race: Black**

*when domain variable is part of calibration variables*

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Mean</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calibration Estimator</td>
<td>3125.86</td>
<td>45.14</td>
</tr>
<tr>
<td>Variance estimation accounted for calibration (PROC WTADJUST)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance estimation NOT accounted for calibration</td>
<td>45.22</td>
<td></td>
</tr>
<tr>
<td>Model-Assisted Estimator</td>
<td>3079.16</td>
<td>44.60</td>
</tr>
<tr>
<td>Proper Variance Estimation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naïve Variance Estimation (treating ŷ as true value)</td>
<td>8.10</td>
<td></td>
</tr>
<tr>
<td>Bias Measure of the Model-Assisted Estimate</td>
<td>0</td>
<td>44.88</td>
</tr>
<tr>
<td>Variance estimation accounted for calibration (PROC WTADJUST)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

P-value of the bias measure: 1.000
### Domain Estimates: Gestational Age

#### Gestational Age

*(when domain variable is NOT part of the calibration variables)*

<table>
<thead>
<tr>
<th>Gestational Age</th>
<th>Calibration Estimator</th>
<th>Model-Assisted Estimator</th>
<th>Bias Measure*</th>
<th>P-Value of the Bias Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>Mean</td>
<td>SE</td>
</tr>
<tr>
<td>≤ 36 weeks</td>
<td>2573.59</td>
<td>60.26</td>
<td>2531.91</td>
<td>16.36</td>
</tr>
<tr>
<td>37-38 weeks</td>
<td>3205.85</td>
<td>28.91</td>
<td>3200.72</td>
<td>16.02</td>
</tr>
<tr>
<td>39 weeks</td>
<td>3437.19</td>
<td>33.98</td>
<td>3391.67</td>
<td>15.94</td>
</tr>
<tr>
<td>40 weeks</td>
<td>3418.95</td>
<td>34.42</td>
<td>3454.89</td>
<td>15.89</td>
</tr>
<tr>
<td>41 weeks</td>
<td>3507.68</td>
<td>32.98</td>
<td>3517.14</td>
<td>16.09</td>
</tr>
<tr>
<td>≥42 weeks</td>
<td>3490.92</td>
<td>42.46</td>
<td>3450.13</td>
<td>16.01</td>
</tr>
</tbody>
</table>

* Bias Measure of the Model-Assisted Estimate
Domain Estimates: Mother’s Age

Mother’s Age

*(when domain variable is correlated with the calibration variables)*

<table>
<thead>
<tr>
<th>Mother’s Age</th>
<th>Calibration Estimator</th>
<th>Model-Assisted Estimator</th>
<th>Bias Measure*</th>
<th>P-Value of the Bias Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SE</td>
<td>Mean</td>
<td>SE</td>
</tr>
<tr>
<td>≤ 19</td>
<td>3103.55</td>
<td>62.86</td>
<td>3139.02</td>
<td>34.91</td>
</tr>
<tr>
<td>20-24</td>
<td>3221.35</td>
<td>33.55</td>
<td>3216.27</td>
<td>21.47</td>
</tr>
<tr>
<td>25-29</td>
<td>3343.99</td>
<td>35.55</td>
<td>3298.23</td>
<td>16.37</td>
</tr>
<tr>
<td>30-34</td>
<td>3318.32</td>
<td>35.05</td>
<td>3313.47</td>
<td>20.03</td>
</tr>
<tr>
<td>≥35</td>
<td>3289.98</td>
<td>44.98</td>
<td>3279.74</td>
<td>31.90</td>
</tr>
</tbody>
</table>

* Bias Measure of the Model-Assisted Estimate
Conclusions

- **Design Consistency**
  - When computing a domain estimate, a calibration estimator is design-consistent.
  - A model-assisted estimator is asymptotically design-consistent, only when *domain variable is a component of the calibration variables*.

- **Bias Measure for Model-Assisted Estimator**
  - When *the domain variable is NOT a component of the calibration vector*, a *proper test* should be performed to assess the potential magnitude and significance of the bias of the model-assisted estimate.
Conclusions (continued)

- **Variance Estimation**
  - When the domain variable is a component of the calibration variables, the calibration estimator performs similarly to the model-assisted estimator (both the estimates and SE of estimates are similar; both methods are asymptotically unbiased).
  - When the domain variable is NOT a component of the calibration variables, if the model-assisted estimate is NOT biased, then the model-assisted estimate has smaller SEs (i.e. more efficient) than the calibration estimate. We can test for a potential bias.
Application 2: Two-Phase Sample Survey

2\textsuperscript{nd} Phase
- \textbf{X}
- \textbf{y}
- 2\textsuperscript{nd} phase weight
  × Adjustment

Calibration Estimator

1\textsuperscript{st} Phase
- \textbf{X}
- 1\textsuperscript{st} phase weight

\textbf{X}\textsuperscript{T}\beta = \hat{y}

Model-Assisted Estimator

\textbullet \ Note: the variance estimators for two-phase sample surveys are different from Application 1, because of uncertainties caused by 1\textsuperscript{st} phase survey.
Contact Information

- Dan Liao
dliao@rti.org

- Phillip Kott
pkott@rti.org