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)
  - references of design-based estimation for domains: Yates (1953, 1960), Durbin (1958), Hartley (1959), Lehtonen and Veijanen (2009)
- 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
- $\mathbf{x}_k$ a vector of calibration variables
- d_k a domain-membership indicator
- w_k calibration weights for which $\sum_S w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$

RTI INTERNATIONAL

Two Domain Estimators

We are interested in estimating the population mean in the domain,

$$U: \; M_d = \sum_U d_k y_k \;/\; \sum_U d_k$$

- with either the **calibration estimator**

$$S: \; m_{d,\,ca} = \sum_S w_k d_k y_k \;/\; \sum_S w_k d_k$$

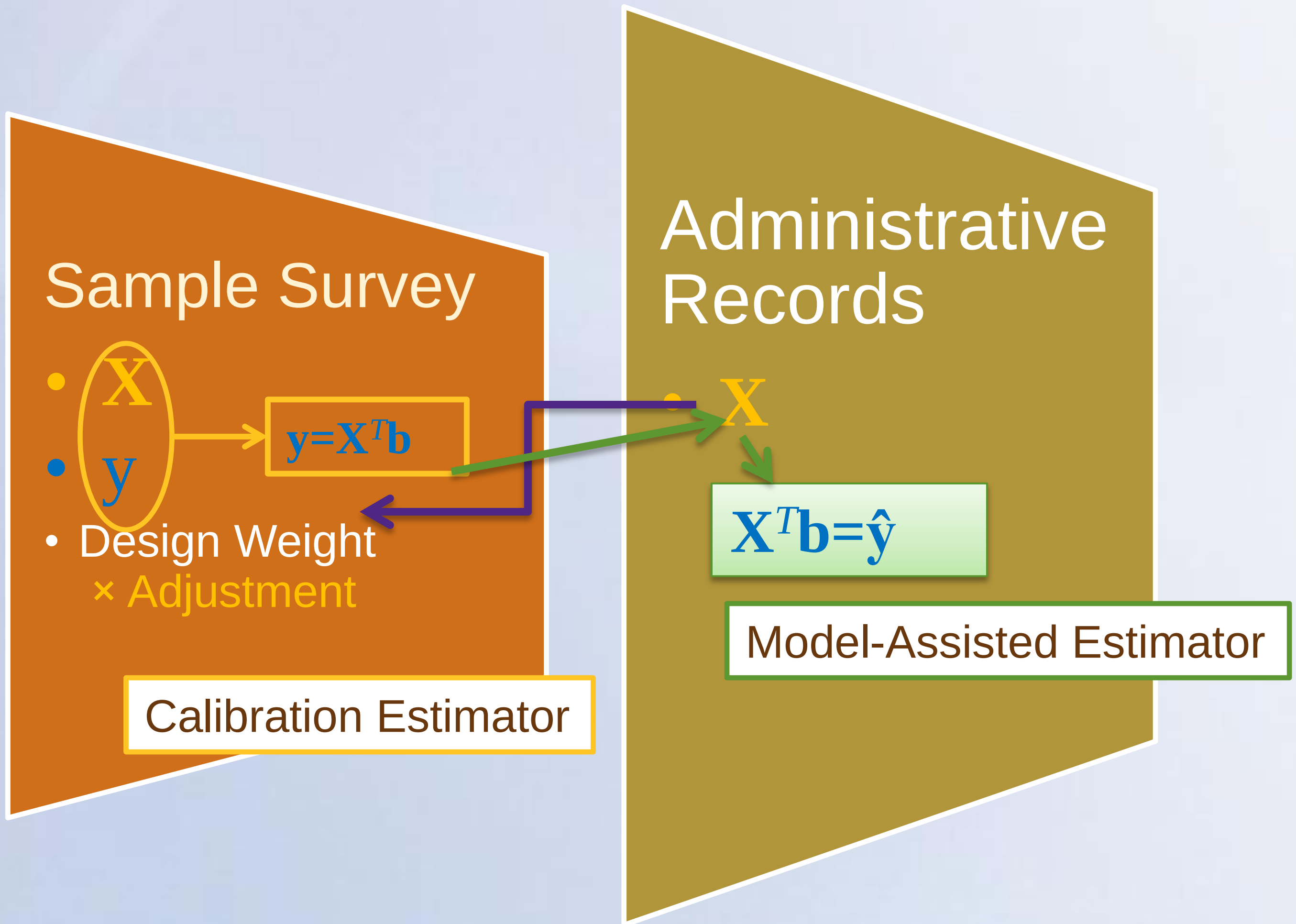
- or the **model-assisted estimator**

$$S: \; y_k \sim \mathbf{x}_k^T \mathbf{b} \qquad \hat{y}_k$$

$$U: \; m_{d,\,ma} = \sum_U d_k \boxed{\mathbf{x}_k^T \mathbf{b}} \;/\; \sum_U d_k = \sum_U d_k \mathbf{x}_k^T \left[\sum_S (w_j \mathbf{x}_j \mathbf{x}_j^T)^{-1} \sum_S w_j \mathbf{x}_j y_j\right] / \sum_U d_k$$

(design weights often replace calibration weights)

Application 1: Combing Information from Administrative Records with Sample Surveys



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 $\boldsymbol{\lambda}$ such that $\boldsymbol{\lambda}^T\mathbf{x}_k = d_k$ for all k (e.g., when d_k is a component of $\mathbf{x}_k$),

    $$m_{d,\,ca} \approx m_{d,\,ma}$$ (≈ means asymptotically equal).
  - Otherwise, model-assisted estimator, $m_{d,\,ma}$, is nearly unbiased (in some sense) when:

    $$\mathrm{E}(y_k \mid \mathbf{x}_k, d_k) = \mathbf{x}_k^T \boldsymbol{\beta}.$$

Bias Measure

If the model is correct in the domain (H_0), the idealized test statistic:

$$T^* = \sum_S w_k d_k (y_k - \mathbf{x}_k^T \boldsymbol{\beta}) / \sum_S w_k d_k$$

has expectation (nearly) zero.

- Estimated test statistic:

$$T = \sum_S w_k d_k (y_k - \mathbf{x}_k^T \mathbf{b}) / \sum_S w_k d_k$$

$$= \sum_S w_k d_k q_k / \sum_S w_k d_k$$

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

$$\mathrm{var}(m_{d,\,ma}) = \left(\sum_S w_j d_j\right)^{-2} \mathrm{var}\left(\sum_S w_k z_k\right),$$

where $z_k = \left[\sum_U d_j \mathbf{x}_j^T \sum_S (w_j \mathbf{x}_j \mathbf{x}_j^T)^{-1}\right] \mathbf{x}_k (y_k - \mathbf{x}_k^T \mathbf{b})$.

– $\mathrm{var}(\sum_S w_k z_k)$ 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

Calibration Variable ($\mathbf{x}_k$):

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

Calibration Method: Generalized Raking

$$w_k = w_k^{original} exp(\mathbf{x}_k{}^T \mathbf{b})$$

(Other methods could have been used)

Domain Estimates: Mother's Race

- Mother's Race: Black

(when domain variable is part of calibration variables)

Estimator		Mean	SE
Calibration Estimator	Variance estimation accounted for calibration (PROC WTADJUST)	3125.86	45.14
	Variance estimation NOT accounted for calibration		45.22
Model-Assisted Estimator	Proper Variance Estimation	3079.16	44.60
	Naïve Variance Estimation (treating $\hat{y}$ as true value)		8.10
Bias Measure of the Model-Assisted Estimate	Variance estimation accounted for calibration (PROC WTADJUST)	0	44.88

P-value of the bias measure: 1.000

Domain Estimates: Gestational Age

- Gestational Age

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

Gestational Age	Calibration Estimator		Model-Assisted Estimator		Bias Measure*	P-Value of the Bias Measure
	Mean	**SE**	**Mean**	**SE**		
≤ 36 weeks	2573.59	60.26	2531.91	16.36	-706.15	0.000
37-38 weeks	3205.85	28.91	3200.72	16.02	66.04	0.020
39 weeks	3437.19	33.98	3391.67	15.94	149.96	0.000
40 weeks	3418.95	34.42	3454.89	15.89	139.27	0.000
41 weeks	3507.68	32.98	3517.14	16.09	233.26	0.000
≥42 weeks	3490.92	42.46	3450.13	16.01	215.80	0.000

* Bias Measure of the Model-Assisted Estimate

Domain Estimates: Mother's Age

- Mother's Age

  (when domain variable is correlated with the calibration variables)

Mother's Age	Calibration Estimator		Model-Assisted Estimator		Bias Measure*	P-Value of the Bias Measure
	Mean	SE	Mean	SE		
≤ 19	3103.55	62.86	3139.02	34.91	-98.04	0.115
20-24	3221.35	33.55	3216.27	21.47	-12.26	0.709
25-29	3343.99	35.55	3298.23	16.37	65.75	0.062
30-34	3318.32	35.05	3313.47	20.03	5.05	0.885
≥35	3289.98	44.98	3279.74	31.90	-55.42	0.202

* 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

2nd Phase

- **X**
- **y**
- 2nd phase weight ×Adjustment

Calibration Estimator

1st Phase

- **X**
- **1st phase weight**

$\mathbf{X}^T\boldsymbol{\beta}=\hat{\mathbf{y}}$

Model-Assisted Estimator

- **Note: the variance estimators for two-phase sample surveys are different from Application 1, because of uncertainties caused by 1st phase survey.**

Contact Information

- Dan Liao
  dliao@rti.org

- Phillip Kott
  pkott@rti.org