A Different Paradigm Shift: Combining Administrative Data and Survey Samples for the Intelligent User

Phillip Kott (with Dan Liao)
RTI International

Washington Statistical Society Conference on Administrative Records for Best Possible Estimates
September 18, 2014

Introduction

- Polemics later.
  Our focus will mostly be on statistics.
- We propose using "model-assisted" estimates for domains when domain-specific survey data are sparse but useful auxiliary administrative data exist and when the domain estimates are not deemed biased.
- Calibration estimates are not useful in this context, while estimates that trade off bias and variance are overkill.
- Linearization is possible, but the jackknife is easier.
- If needed we can add errors to our predicted values (e.g., for estimating proportions and percentiles).

Notation

Let

- U be the population (of N elements)
- S the sample
- y_k the value of interest for survey element k,
- $\mathbf{x}_k$ a vector of administrative calibration variables
- δ_k a domain-membership indicator
- d_k design weight (after adjusting for selection biases)
- $w_k \approx d_k$ calibration weight for which $\sum_S w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$

Two Domain Estimators

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

$$Y_\delta = \sum_U \delta_k y_k.$$

- We could use a **calibration estimator**

$$\hat{Y}_{\delta,ca} = \sum_S w_k \delta_k y_k .$$

- Or this **model-assisted (or synthetic) estimator**

The model: $E(y_k) = \mathbf{x}_k^T \boldsymbol{\beta}$

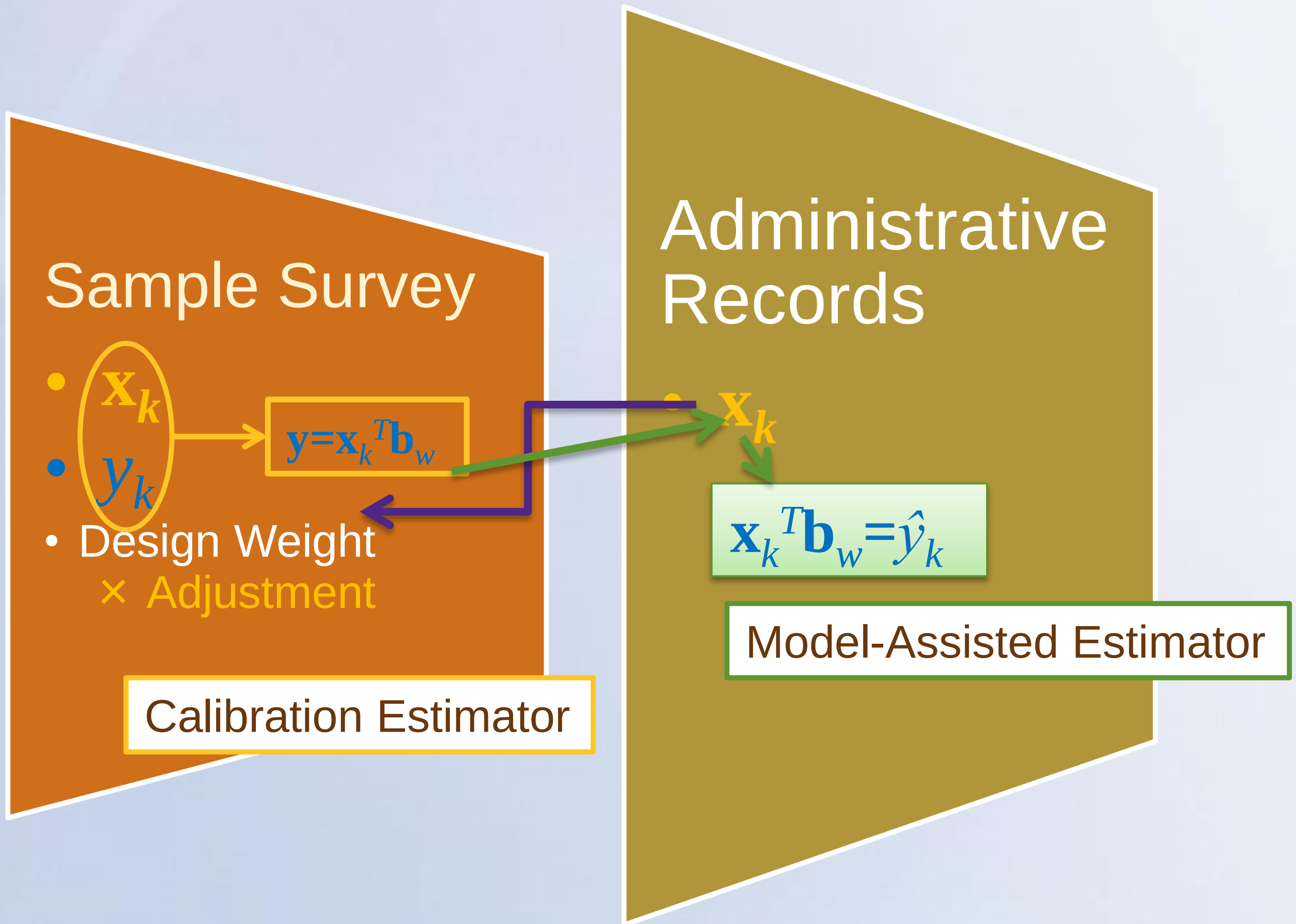
$\hat{y}_k$

$$\hat{Y}_{\delta,ma} = \sum_U \delta_k \mathbf{x}_k^T \mathbf{b}_w = \sum_U \delta_k \mathbf{x}_k^T \left[\sum_S (w_j \mathbf{x}_j \mathbf{x}_j^T)^{-1} \sum_S w_j \mathbf{x}_j y_j\right]$$

↑ ↑

(design weights can replace calibration weights)

Combining Information from Administrative Records with Sample Surveys



Bias Measure

- Calibration estimator, $\hat{Y}_{\delta,ca}$, is *design consistent* (when the sample size in the domain is large enough).
- Model-assisted estimator: $\hat{Y}_{\delta,ma} = \sum_U \delta_k \mathbf{x}_k{}^T \mathbf{b}_w$

  When there is a λ such that for all k $\lambda^T \mathbf{x}_k = \delta$,

$$\hat{Y}_{\delta,ma} = \sum_U \delta_k \mathbf{x}_k{}^T \mathbf{b}_w \approx \sum_S w_k \delta_k \mathbf{x}_k \mathbf{b}_w = \hat{Y}_{\delta,ca},$$

and the model-assisted estimator is nearly unbiased.

Otherwise, it is nearly unbiased (in some sense) only when $\mathrm{E}(y_k \mid \mathbf{x}_k, \delta_k) = \mathbf{x}_k{}^T \boldsymbol{\beta}$.

Bias Measure

More on the Magic Formula

When $\lambda^T \mathbf{x}_k = \delta_k$ for all k (e.g., when δ_k is a component of $\mathbf{x}_k$ and the corresponding component of λ is 1 while the others are all 0):

$$\sum_S w_k \delta_k \hat{y}_k = \sum_S w_k \delta_k \mathbf{x}_k{}^T \mathbf{b}_w$$

$$= \sum_S w_k \delta_k \mathbf{x}_k{}^T \left(\sum_S w_j \mathbf{x}_j \mathbf{x}_j{}^T\right)^{-1} \sum_S w_j \mathbf{x}_j y_j$$

$$= \sum_S w_k \lambda^T \mathbf{x}_k \mathbf{x}_k{}^T \left(\sum_S w_j \mathbf{x}_j \mathbf{x}_j{}^T\right)^{-1} \sum_S w_j \mathbf{x}_j y_j$$

$$= \sum_S w_k \lambda^T \mathbf{x}_k \mathbf{x}_k{}^T \left(\sum_S w_j \mathbf{x}_j \mathbf{x}_j{}^T\right)^{-1} \sum_S w_j \mathbf{x}_j y_j$$

$$= \lambda^T \sum_S w_j \mathbf{x}_j y_j$$

$$= \sum_S w_j \delta_j y_j = \hat{Y}_{\delta,ca}$$

Bias Measure

Otherwise, iff the model is correct in the domain (H_0), the idealized test statistic: $T^* = \sum_S w_k \delta_k (y_k - \mathbf{x}_k^T \boldsymbol{\beta})$ has expectation (nearly) zero.

- Estimated test statistic, the bias measure:

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

$$= \sum_S w_k \delta_k q_k$$

This can be treated as a calibrated mean and the estimated variance can be computed with WTADJUST in SUDAAN *but a jackknife would be better* (because $\mathbf{b}_w$ is random and finite-population correction is a nonissue).

Variance Estimation

- ## Calibration Estimator

  Estimating the combined variance of $\hat{Y}_{\delta,ca}$ (model and probability-sampling) is straightforward with WTADJUST if, say, $w_k = d_k exp(\mathbf{x}_k^T \mathbf{g})$.

- ## Model-Assisted Estimator

  $$\text{var}(\hat{Y}_{\delta,ma}) = \text{var}(\sum_U \delta_j \mathbf{x}_j^T \mathbf{b}_w) = \text{var}(\sum_S w_k z_k),$$

  where $z_k = [\sum_U \delta_j \mathbf{x}_j^T \sum_S (w_j \mathbf{x}_j \mathbf{x}_j^T)^{-1}] \mathbf{x}_k (y_k - \mathbf{x}_k^T \mathbf{b}_w),$

  and $\text{var}(\sum_S w_k z_k)$ can be estimated with WTADJUST, but …

Variance Estimation

Jackknifing is easier

(*if* finite-population correction can be ignored).

Effectively, it is the $\mathbf{b}_w$ that are computed, first with the original calibration weights, then with the replicate calibration weights.

Operationally, it is as if each of the $\hat{y}_k = \mathbf{x}_k \mathbf{b}_w$ in U are computed, first with the original calibration weights, then with the replicate calibration weights.

Example: Drug-Related ED Visits

A mostly-imaginary frame U of N = 6300 hospital emergency departments (EDs).

Each hospital has a previous annual number of ED visits, and is either *urban* or *non-urban*, *public* or *private.*

We have a stratified (16 strata) simple random sample of n = 346 EDs.

Stratification by region, urban/nonurban, and partially by public/private and size.

Stratum sample sizes range from 5 to 65.

Calibration Weighting

Initial Calibration Variables ($\mathbf{x}_k$):

- Regions (four categories),
- Frame visits (continuous), and
- Public/Private
- Urban/Nonurban

Calibration Weighting Method: Unconstrained Generalized Raking:

$$w_k = d_k exp(\mathbf{x}_k^T \mathbf{g})$$

Weights must be positive, unlike with linear calibration.

The Extended Delete a Group Jackknife

- List by the sample by stratum, then systematically assign each sampled unit to one of G = 30 groups.

- Initially set $d_k(r) = 0$ if $k \in$ Group r,

  $d_k(r) = N_h/n_{hr}$ if $k \notin$ Group r and $k \in$ Stratum h

  $d_k(r) = w_k$ otherwise.

- If stratum containing k has $n_h < 30$,

  replace 0 with $d_k[1 - (n_h - 1)Z_h]$ and

  replace N_h/n_{hr} with $d_k(1 + Z_h)$, where $Z_h^2 = 30/[29n_h(n_h - 1)]$.

The Extended Delete a Group Jackknife

The DAG Jackknife Variance Estimator for a estimator t is

$$v_{DAG} = \frac{29}{30}\sum_{r=1}^{30}(t_{(r)} - t)^2,$$

where $t_{(r)}$ is computed with the r'th set of weights which may themselves be calibrated – in our case to the same targets as the original sample.

There is no harm replacing t with the average of the $t_{(r)}$.

It's relative standard error is at most $\sqrt{(2/29)} \approx .26$

The Domains

Region (1, 2, 3, 4) × Public (1) or not (0)

Domain	Sample Size	Bias Measure	Standard Error	t value (Bias/SE)
All	346	-0.00000	0.00000	-0.11939
10	62	0.40960	0.52798	0.77579
11	97	-0.75017	0.97290	-0.77107
20	18	-0.74959	1.38844	-0.53988
21	36	0.27749	0.51398	0.53988
30	73	0.13164	0.04390	2.99848
31	5	-3.30938	1.10369	-2.99848
40	42	-0.21434	0.45655	-0.46949
41	13	0.33511	0.71378	0.46949

Standard errors were estimated with an extended dag jackknife.

Only Cell 31 had a bad *t* value with a linearized test.

The Estimates

Domain	Direct Estimate	Direct SE	Calibrated Estimate	Calibrated SE	Model-Assisted Estimate	Model-Assisted SE
All	55228	3951	52346	1325	52346	1325
10	11905	808	11436	774	11667	398
11	6149	575	5773	506	6475	321
20	1340	466	1212	369	644	276
21	16164	2677	15004	1669	15058	661
30	4336	229	4268	227	3987	202
31	96	32	102	35	207	36
40	8370	1145	7999	1010	8170	711
41	6868	1972	6551	1767	6137	320

All standard errors were estimated with an extended dag jackknife (with no finite-population correction).

The Estimates Redux

After adding a dummy calibration variable for Cell 30

	Direct		Calibrated		Model-Assisted	
Domain	Estimate	SE	Estimate	SE	Estimate	SE
All	55228	3951	52354	1328	52354	1328
10	11905	808	11426	778	11646	397
11	6149	575	5781	503	6497	325
20	1340	466	1211	369	617	280
21	16164	2677	15017	1677	15092	662
30	4336	229	4278	227	4112	205
31	96	32	96	32	90	29
40	8370	1145	7975	1007	8095	724
41	6868	1972	6571	1777	6206	322

The Estimates with All Cells in the Model

Domain	Our Model-Assisted Estimate	Our Model-Assisted SE	All Cells Model-Assisted Estimate	All Cells Model-Assisted SE
All	52354	1328	52345	1321
10	11646	397	11871	483
11	6497	325	6271	343
20	617	280	513	500
21	15092	662	15208	496
30	4112	205	4111	205
31	90	29	90	29
40	8095	724	7978	746
41	6206	322	6302	445

The *All Cells Model-Assisted Estimate* includes frame visits, an urban indicator, and eight cell indicators in the model.

Interpreting the Results

Calibration weighting greatly decreased the standard error of the estimate for all drug-related hospital visits, but only marginally within individual domains (cells).

What we have called a “model-assisted” estimator worked much better.

Estimates were biased in two cells, a bias that was removed by adding a cell identifier.

Adding all the cell identifiers tended to increase domain standard errors.

Discussion Points

- Isn’t what you proposed really just a synthetic estimator?
- Yes.
- Why use weights when estimating β?
- Because the sampling design may not be ignorable.
- It also makes the numbers add up across domains.
- Aren’t those test of bias weak?
- Yes. And absence of evidence is not evidence of absence.
- More testing is advisable.
- Empirical Bayes/Empirical BLUP/Hierarchical Bayes effectively model the bias when it cannot be assumed to be zero.

Discussion Points

- Why didn't calibration weighting work better?
- For a domain, one is effectively modeling $\delta_k y_k$ (or worse, $\delta(y_k - \bar{y}_\delta)$, when estimating means) as a function of the calibration variables.
- For calibration weighting to work well, one would need domain-specific calibration variables.
- Nearly pseudo-optimal calibration weighting would have worked a *little* better.
- What about estimating means?
- An intercept needs to be in the model, then the extension is trivial.

Discussion Points

- How do we estimate proportions and percentiles?
- We could replace the linear model with a logistic.
- Better would be to sort the weighted sample y_k by their $\mathbf{x}_k^T\mathbf{b}_w$ values and the frame $\hat{y}_k$ conformally.
  Then assign errors to the frame values from the sample values systematically.

- What if finite-population correction mattered (as it should have here)?
- We could have only predicted values for U–S using $\mathbf{b}_{w\text{-}1}$.
  Proper variance estimation is less clear.

Concluding Remarks

- We need to walk humbly with our data.
- Our estimates do no come from on high.
  They are fraught with potential errors,
  which we should make as clear to users as possible.
- We should redirect our estimation program to serve primarily intelligent users, rather than treating our target audience like they are dumber than dirt.
- As always, more research is needed (on variance estimation).

Some References

Kim, J.K. and Rao, J.N.K. (2012). Combining data from two independent surveys: a model-assisted approach. *Biometrika* 99(1), 85-100.

Kott, P.S. (2011). A nearly pseudo-optimal method for keeping calibration weights from falling below unity in the absence of nonresponse or frame errors. *Pakistan Journal of Statistics*, 27(4), 391–396.

Kott, P.S. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 17(4), 521–526.

Contact Information

- Phillip Kott
  pkott@rti.org

- Dan Liao
  dliao@rti.org