Survey calibration and two-phase designs

Thomas Lumley
(and James Dai)

UW Biostatistics

WNAR 2007

Outline

- What are survey calibration estimators
- Relationship to RRZ efficient estimators
- Calibration and ‘estimated weights’ paradox
- Calibration for straightforward AIPW estimators
- Semiparametric models with small codimension

Underlying question: **where does the information come from in semiparametric models with high codimension**

Estimating a total

Population size N, sampling size n, sampling probabilities π_i, sampling indicators R_i.

Goal: estimate

$$\sum_{i=1}^{N} y_i$$

Horvitz–Thompson estimator:

$$\hat{T} = \sum_{R_i=1} \frac{1}{\pi_i} y_i$$

To estimate parameters θ replace y_i with loglikelihood $\ell_i(\theta)$ or estimating functions $U_i(\theta)$.

Auxiliary information

HT estimator is inefficient when some population data are available.

Suppose x_i known for all i.

Fit $y_i \sim x_i\beta$ by (probability weighted) least squares to get $\widehat{\beta}$. Let r^2 be the proportion of variation explained.

$$\hat{T}_{reg} = \sum_{R_i=1} \frac{1}{\pi_i}(y_i - x_i\widehat{\beta}) + \sum_{i=1}^{N} x_i\widehat{\beta}$$

ie, HT estimator for sum of residuals plus population sum of fitted values.

Auxiliary information

For large n, N and with conditions on moments and sampling

$$\text{var}\left[\hat{T}_{reg}\right] = r^2\text{var}\left[\hat{T}\right] + O(N/\sqrt{n}) = \left(r^2 + O(n^{-1/2})\right)\text{var}\left[\hat{T}\right]$$

and the relative bias of $\hat{T}_{reg}$ is $O(1/n)$

The lack of bias does not require any assumptions about the distribution of $Y|X$

$\widehat{\beta}$ is consistent for the population least squares β, for which the mean residual is zero by construction.

Auxiliary information

Since $\hat{\beta}$ is linear in y_i we can write $\hat{T}_{reg}$ as

$$\hat{T}_{reg} = \sum_{R_i=1} w_i y_i = \sum_{R_i=1} \frac{g_i}{\pi_i} y_i$$

for some (ugly) g_i.

$\hat{T}_{reg}$ is an IPW estimator using ‘calibrated’ or ‘tuned’ (French: *calage*) weights.

General calibration problem: for a suitable metric $d(\,,\,)$ find g_i minimizing $\sum d(g_i, 1)$ and satisfying the calibration equations

$$\sum_{R_i=1} \frac{g_i}{\pi_i} x_i = \sum_{i=1}^{N} x_i$$

[Deville et al, JASA 1993; JNK Rao et al, Sankhya 2002]

Two-phase studies

Sample N individuals and measure some variables, then subsample n individuals with probabilities π_i and measure more individuals

Calibration problem: find g_i minimizing $\sum d(g_i, 1)$ and satisfying the calibration equations

$$\sum_{R_i=1} \frac{g_i}{\pi_i} x_i = \sum_{i=1}^{N} x_i$$

as in single-phase sample.

To estimate parameters θ, replace y_i with $U_i(\theta)$ and solve

$$\sum_{R_i=1} \frac{g_i}{\pi_i} U_i(\theta) = 0$$

RRZ estimators

Class of calibration estimators same as augmented IPW estimators of Robins, Rotnitzky, & Zhao.

Class contains the efficient estimator in the non-parametric phase-1 model (efficient design-based estimator) — the most efficient estimator that is consistent for the same limit as if we had complete data.

Typically not fully efficient if model assumptions are imposed at phase 1 — eg, Cox model assumes infinitely many constraints at phase 1, and efficient two-phase estimator is known (Mark & Katki, 2006; Nan 2004) and is more efficient than calibration estimator.

Estimated weights

The fact that

$$\sum_{R_i=1} \frac{1}{\pi_i} U_i(\theta) = 0$$

is strictly less efficient than an estimator based on estimated sampling probabilities $\widehat{\pi}_i$

$$\sum_{R_i=1} \frac{1}{\pi_i} U_i(\theta) = 0$$

is widely regarded as a paradox.

But as calibration estimators the gain in efficiency is intuitively clear: it is the same as the gain in efficiency from adjusting for pre-randomization variables in a randomized trial

Estimated weights

Issue is not ‘estimated weights’ vs ‘known weights’ but ‘known totals’ vs ‘estimated totals’ for the auxiliary variables by imposing the calibration equations.

Information is gained to the extent that $U_i(\theta)$ is predictable from auxiliary variables, and this gain can be large.

Choice of auxiliaries

The other heuristic gain from the calibration viewpoint is in choosing predictors for estimating π.

The regression formulation shows that the predictors should have strong linear relationships with $U_i(\theta)$.

If $U_i(\theta)$ is of a form such as

$$z_i w_i (y_i - \mu_i(\theta))$$

then z_i is approximately uncorrelated with U_i

So, don't use a variable correlated with a phase-2 predictor as a calibration variable, use a variable correlated with the phase-2 score function.

Survival example

Kulich & Lin (JASA, 2004) gave an example of a (pretend) two-phase study of Wilm's Tumor

- Phase 1: age at onset, tumor diameter, stage, histology (from local lab), time to relapse

- Phase 2: histology from gold-standard central lab.

Phase 2 is a stratified subsample base on relapse and on local-lab histology report.

[In fact, gold-standard histology available for everyone, so we can compare efficiency to full-cohort analysis] Kulich & Lin (following

Survival example

RRZ) used local-lab histology and other phase-1 variables in logistic regression to estimate $\widehat{\pi}$ (in fact, separately for each time point).

Effectively the same as calibration using phase-1 variables as auxiliary variables

Alternative:

- estimate misclassification in local-lab histology using phase two.
- Fit a measurement-error model to phase one using Prentice (1982) first-order approximation.
- Use the individual estimating functions U_i^* as calibration variables.

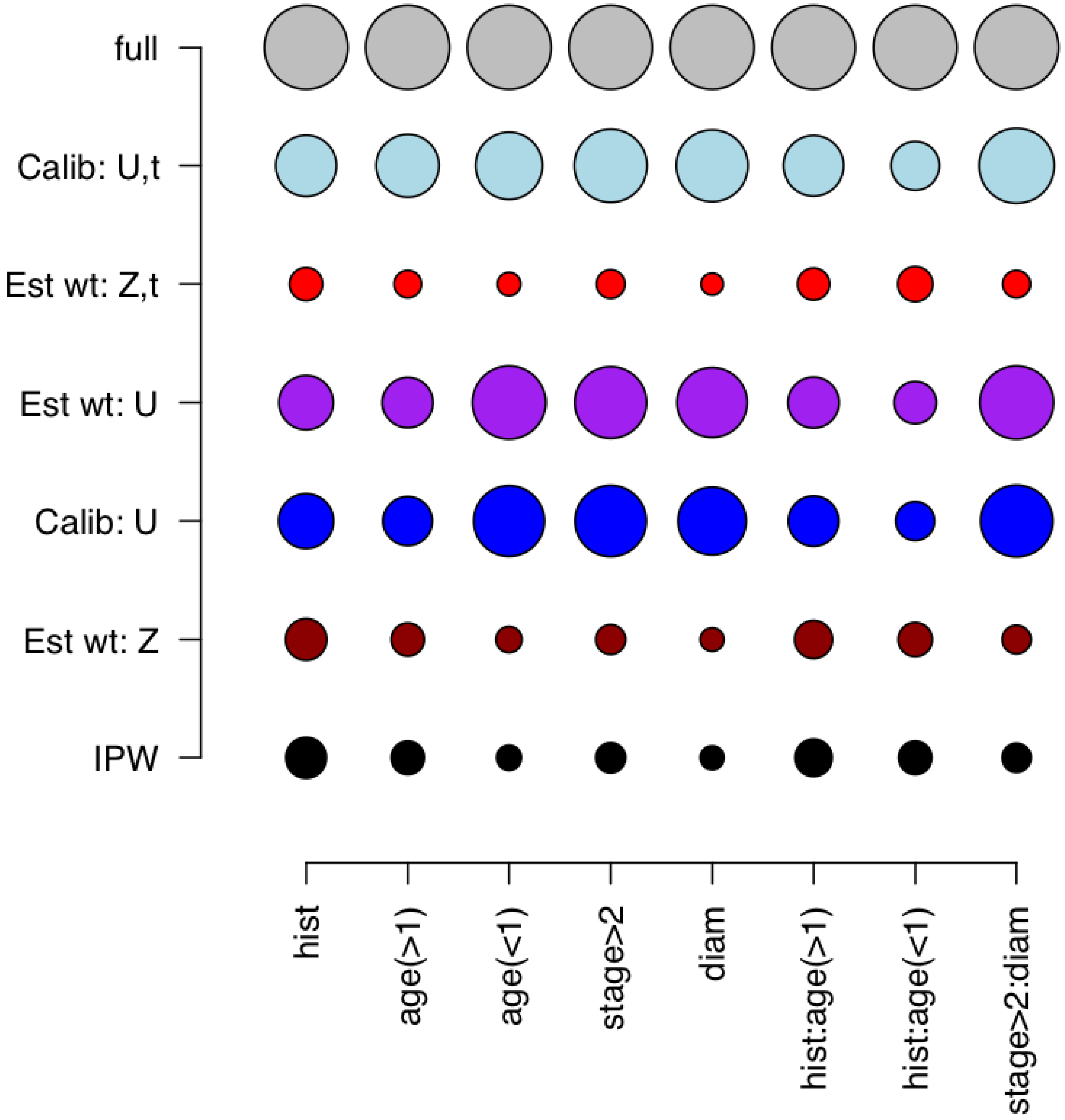
U_i^* and U_i should be more strongly correlated than local-lab histology and U_i.

Results: key

- U vs Z: auxiliary variable is estimating function (U) or raw variable (Z)
- Calib vs Est wt: calibration or weights estimated from probabilities from fitted logistic regression.
- t: indicates time-dependent weights (3 time periods)
- IPW: plain Horvitz-Thompson estimator
- full: Full cohort

Area of circle indicated estimated relative efficiency. Rows are estimators, columns are parameters.

Results


full
Calib: U,t
Est wt: Z,t
Est wt: U
Calib: U
Est wt: Z
IPW
hist
age(>1)
age(<1)
stage>2
diam
hist:age(>1)
hist:age(<1)
stage>2:diam

Gene-environment independence

Calibration can also be used to fit semiparametric models of finite codimension.

An example is a case–control study of drug–gene interaction

$$\text{logit } P[Y=1] = \alpha + \beta_g I(gene) + \beta_d I(drug) + \gamma I(drug \cap gene)$$

Often plausible that drug and genetic variant are independent in the population.

Gene-environment independence

For $2 \times 2 \times 2$ table and rare events, case-only estimation exploits the independence [Piegorsch et al, 1994]

		G	
	E	0	1
Case	0	a	b
	1	c	d
Control	0	e	f
	1	g	h

case-control estimator:

$$\frac{ad/bc}{eh/gf}$$

case-only estimator:

$$\frac{ad}{bc}$$

Calibration to independence

Find g_i so that

$$\sum_{R_i=1,\, y=0} \frac{g_i}{\pi_i}(gene-\mu_g)(drug-\mu_d) = E\left[(gene-\mu_g)(drug-\mu_d)\right] = 0$$

This reproduces the case-only estimator exactly. Same approach can be used with more complex models, by calibrating interaction terms to zero.

For discrete data there are only finitely many terms, for continuous data use basis functions such as polynomials or splines.

For rare events calibration to independence is almost fully efficient, implying that no information comes from the structure of the logistic model.

Calibration to independence

For non-rare events calibration is slightly more complicated, and is not fully efficient. It captures about half the efficiency gain possible from the NPMLE. (Dai 2007+)

Likelihood-based approaches to this problem exist (?abound): Chatterjee, Carrol, McNeney, Graham.

Calibration is interesting because it is easy and because it tells us about the sources of information.

Conclusions

Survey calibration is a useful way to think about RRZ-style estimators: intuition is easier

Calibration is a useful way to construct estimators in models of low co-dimension

Calibration estimators are typically usefully more efficient than the HT estimator (though not always)

Calibration estimators are often less efficient than semiparametric likelihood estimators, but the extra information really does come from the phase-one model assumptions.