Statistical Analysis Using Combined Data Sources

Ray Chambers
Centre for Statistical and Survey Methodology
University of Wollongong

JPSM Presentation, University of Maryland, April 7, 2011

What Are ‘The Data’?

The classical perspective - a (random) rectangular window on the population of interest…

A Messier Real World

Multiple sources of data with varying levels of aggregation

Typically a distorted window on the target population + windows on related populations

Example 1

- Sample values Y and X for two correlated variables at t_1
- Register values of X at time t_{k-1} available at time t_k
- Focus on estimation of population total of Y at t_1

Example 2

- Values of Y from register A
- Values of X from register B
- Sample of units from register A linked to register B
- Interest in modelling Y - X relationship at register level

Example 3

- Values of Y from register A
- Values of X from register B
- Registers probabilistically linked
- Interest in modelling Y - X relationship at register level

Example 4

- Values of Y and auxiliaries X and C from survey A
- Values of Y and auxiliaries Z and C from survey B
- Marginal estimates for Y based on combined sample required

Example 5

- Values of 'accurate' zero-one variable Y from a small survey A
- Values of 'rough' zero-one approximation X from a much larger survey B
- Small area estimates of Y required

Example 6

- Values of Y and Z from a large national survey
- Values of X and Z from another, distinct, large national survey
- Values of Y, X and Z from a small, non-representative, third survey
- National model relating Y, X and Z required

How do we tackle inference using these complex data sources?

General Approach 1: Calibrated Sample Weighting

Sample weighting is a standard method of survey estimation

Many survey estimation systems use **calibrated** sample weights

- i.e. they are capable of **exactly** reproducing known population quantities

- Typically, these are either totals or means of auxiliary variables

- **Natural way of integrating external information into survey weights**

Closest Calibrated Weighting

Model-Assisted Approach

Deville and Särndal (1992) - Use calibrated sample weights w_i that are **closest** to the expansion weights π_i^{-1} (i.e. the weights that define the Horvitz-Thompson estimator for the survey variables)

A metric for 'closeness' $\quad Q = (\mathbf{w}_s - \boldsymbol{\pi}_s^{-1})'\boldsymbol{\Omega}(\mathbf{w}_s - \boldsymbol{\pi}_s^{-1})$

Minimising Q subject to calibration leads to **GREG weights**

- $\mathbf{w}_s = \boldsymbol{\pi}_s^{-1} + \boldsymbol{\Omega}^{-1}\mathbf{Z}_s'\left(\mathbf{Z}_s'\boldsymbol{\Omega}^{-1}\mathbf{Z}_s\right)^{-1}(\mathbf{Z}_U'\mathbf{1}_N - \mathbf{Z}_s'\boldsymbol{\pi}_s^{-1})$
- $\mathbf{Z}_U$ = matrix of population values of auxiliary variables
- $\mathbf{Z}_s$ = corresponding matrix of sample values
- $\boldsymbol{\Omega}$= positive definite matrix (typically $diag\left(\pi_i v_i\right)$)

Model-Based Calibration

$$\mathbf{y}_U = \mathbf{Z}_U \beta + \mathbf{e}_U$$

Unbiased estimation under this model is a good thing...

$$E\left(\hat{t}_y - t_y \middle| \mathbf{Z}_U\right) = 0$$

Calibration with respect to the auxiliary variables in Z_U is equivalent to unbiasedness with respect to the linear model $E(\mathbf{y}_U | \mathbf{Z}_U) = \mathbf{Z}_U \beta$

$$E\left(\hat{t}_y - t_y \middle| \mathbf{Z}_U\right) = E\left(\mathbf{w}_s' \mathbf{y}_s - \mathbf{1}_N' \mathbf{y}_U \middle| \mathbf{Z}_U\right) = \left(\mathbf{w}_s' \mathbf{Z}_s - \mathbf{1}_N' \mathbf{Z}_U\right) \beta = 0$$

So calibration is a good thing – provided the linear model assumption is valid!

Efficient Model-Based Calibration

Royall (1976) - Best linear unbiased predictor (BLUP) of t_y is defined by weights of the form

$$\mathbf{w}_s^{opt} = \mathbf{1}_n + \mathbf{H}'\left(\mathbf{Z}_U'\mathbf{1}_N - \mathbf{Z}_s'\mathbf{1}_n\right) + \left(\mathbf{I}_n - \mathbf{H}'\mathbf{Z}_s'\right)\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\mathbf{1}_{N-n}$$

.. $\mathbf{I}_n$ = identity matrix of order n

.. $\mathbf{1}_k$ = k-vector of one's

.. $\mathbf{H} = \left(\mathbf{Z}_s'\mathbf{V}_{ss}^{-1}\mathbf{Z}_s\right)^{-1}\mathbf{Z}_s'\mathbf{V}_{ss}^{-1}$

.. $\mathbf{V}_U \propto Var\left(\mathbf{y}_U\right) = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}$

Model unbiasedness ⇔ these weights are calibrated on Z_U

$$\mathbf{Z}_s'\mathbf{w}_s^{opt} = \mathbf{Z}_U'\mathbf{1}_N$$

Example 1: Calibrating to Out of Date Constraints

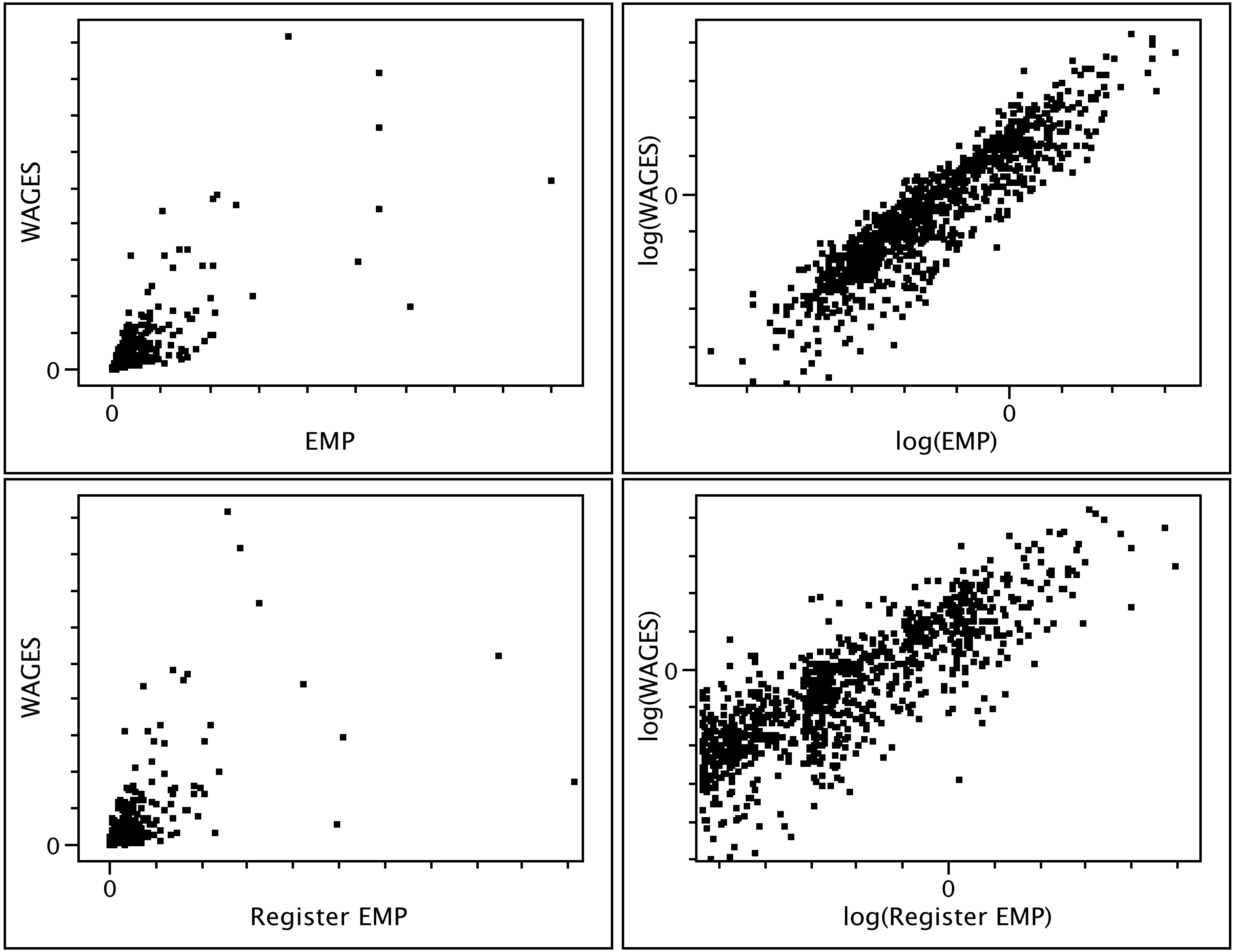
The population marginal information used to define a calibration constraint is often not 'exact'...

We want on calibrate on X, but do not know the population total of this variable. Instead, we know the population total of Z, a variable that is 'approximately' the same as X

- Should we try to 'predict' the population total of X or just calibrate to Z?

Y = WAGES (wages bill for a business)
X = EMP (# employees of the business)
Z = Register EMP (# employees of the business at last update of the business register)

Sector K1 (N = 1005)



Preferred Model (A)

$$E(Y|X) = \alpha_X + \beta_X X$$
$$Var(Y|X) = \sigma_X^2 X^2$$

- Calculation of Preferred BLUP requires value of t_X (unknown)
- Use Predicted value of t_X?
- Or Substitute t_Z for t_X?

Alternative Model (B)

$$E(Y|Z) = \alpha_Z + \beta_Z Z$$
$$Var(Y|Z) = \sigma_Z^2 Z^2$$

- Calculation of this Alternative BLUP requires value of t_Z (known)

- $\sigma_X^2 < \sigma_Z^2$, so BLUP under (A) expected to be more efficient than BLUP under (B)
- Values of Y, X and Z available on the sample

Design-Based Simulations

	%RelBias				%RelRMSE			
	Pref	Alt	Sub	Pred	Pref	Alt	Sub	Pred
STR1: stratified on Z / equal allocation / across stratum estimation								
GREG	0.44	0.24	-5.91	-0.43	11.32	13.29	12.19	11.93
BLUP	-5.69	-5.75	-11.50	-6.48	10.03	11.81	13.83	11.27
STR2: stratified on Z / equal allocation / within stratum estimation								
Both	1.07	-0.76	-7.21	-0.85	12.02	14.00	12.49	12.45
PPZ: no stratification / probability proportional to Z sampling								
GREG	0.16	0.05	-6.34	-1.73	9.98	11.32	11.40	11.44
BLUP	-3.86	-2.64	-9.68	-5.57	8.42	11.23	11.96	10.57

Reality?

Overall accuracy is all well and good, but a more practical requirement is that our estimate should deviate as little as possible from its preferred value...

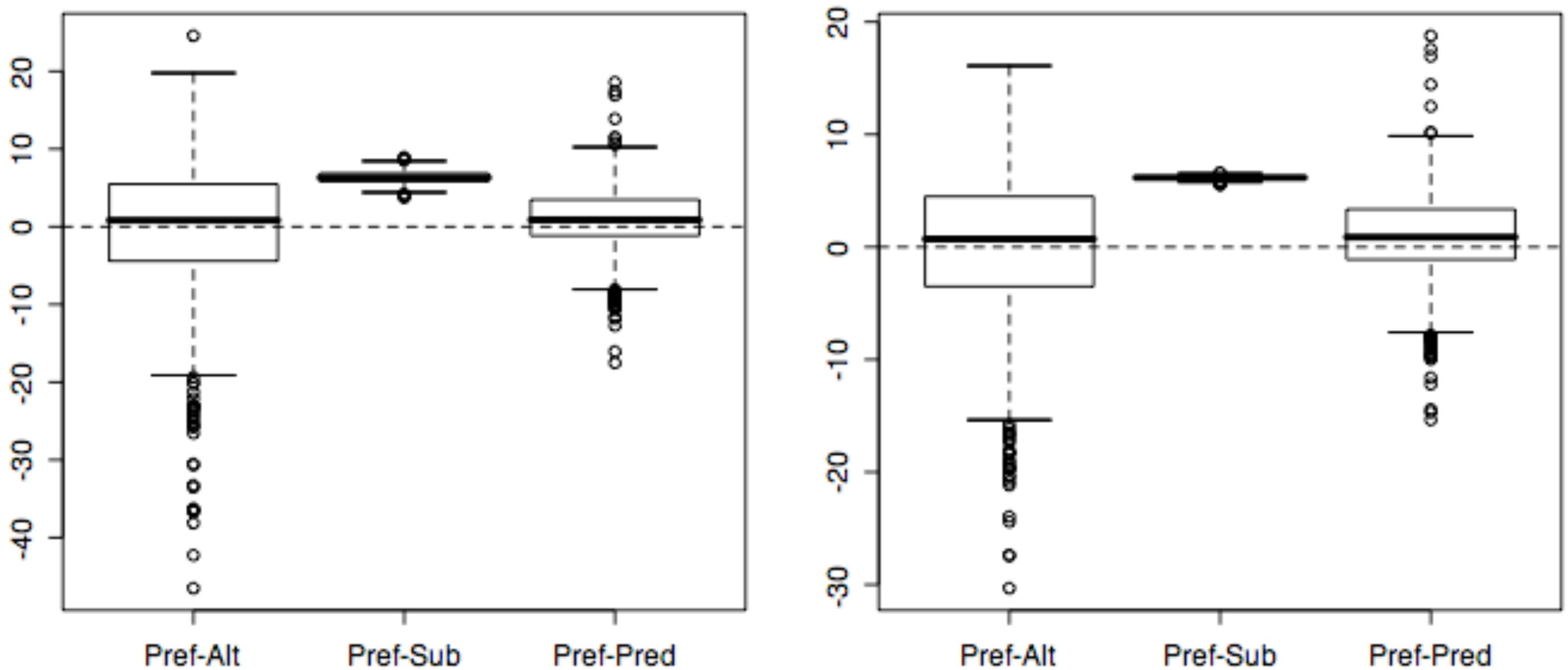
- preferred value is estimate defined by calibrating on X

- we want to **minimise revision** when t_X does become available...

Distributions of %RelDiff for STR1

GREG

BLUP



General Approach 2: Likelihood-Based Information Pooling

The data we have

A (confusing) mix of individual Y-values, values of other, related, variables, summary statistics, metadata (e.g. data definitions), paradata (e.g. information about how the data were obtained, sample weights, auxiliary data for the target population), related data from other surveys and other populations, etc, etc ....

The data we'd like to have (for likelihood inference)

A clean 'rectangular' database with unit record data for the target population and related populations

The Missing Information Principle

Likelihood-based inference using a 'messy' observed dataset $\mathbf{d}_s$ can be achieved by carrying out likelihood-based inference using a larger 'clean' dataset $\mathbf{d}_U$ but with the sufficient statistics defined by $\mathbf{d}_U$ replaced by their expected values given $\mathbf{d}_s$

Note

1. It doesn't matter what $\mathbf{d}_U$ is. The only requirement is that $\mathbf{d}_s$ (the data we have) is a subset of $\mathbf{d}_U$ (the data we would like to have)
2. First developed (**Orchard and Woodbury, 1972**) for inference with missing data, and forms basis for EM algorithm (**Dempster, Laird and Rubin, 1977**) used widely with missing data
3. Application to analysis of survey data by **Breckling *et al* (1994)**
4. Basis for data augmentation algorithms used to generate posterior distributions (**Tanner and Wong, 1987**)

MIP Identities

Provided the ideal data $\mathbf{d}_U$ **include** the available data $\mathbf{d}_s$, the available data score sc_s for the parameter Θ of the distribution of $\mathbf{d}_U$ is the conditional expectation, given these data, of the ideal data score sc_U for Θ, i.e.

$$sc_s = E\left\{\partial_\Theta \log f(\mathbf{d}_U) \middle| \mathbf{d}_s\right\} = E_s\left(sc_U\right)$$

Furthermore, the available data information $info_s$ for Θ is the conditional expectation, given these data, of the ideal data information $info_U$ for Θ **minus** the corresponding conditional variance of the ideal data score sc_U, i.e.

$$\begin{aligned} info_s &= E\left\{-\partial_{\Theta\Theta} \log f(\mathbf{d}_U) \middle| \mathbf{d}_s\right\} - Var\left\{\partial_\Theta \log f(\mathbf{d}_U) \middle| \mathbf{d}_s\right\} \\ &= E_s\left(info_U\right) - Var_s\left(sc_U\right) \end{aligned}$$

Example 2: Combining Survey Data and Marginal Population Information – Comparing the MIP With Calibrated Weighting

Motivating Scenario

Population U is such that values y_i and x_i of two scalar variables, Y and X are stored on separate registers, each of size N. A sample s of n units from one register is linked to the other via a unique common identifier, thus defining n matched (y_i, x_i) pairs

Aim To use these linked sample data to estimate the parameters α, β and σ^2 that characterise the population regression model

$$y_i = \alpha + \beta x_i + \sigma\varepsilon_i$$

where $\varepsilon_i \sim iid\ N(0,1)$

Extra Information

Register summary data are available. In particular, we know the **population means** $\bar{y}_U$ and $\bar{x}_U$ of Y and X...

- OLS estimates are **no longer** the 'full information' MLEs

- Can use the MIP to combine this population marginal information with the survey data to obtain full information MLEs

- Use E_s and Var_s to denote conditioning on sample values of Y and X + population means of these variables

MIP $\Rightarrow$ components of the **available data** score function are

$$sc_{1s} = \frac{1}{\sigma^2}\sum_U \left\{E_s(y_i) - \alpha - \beta x_i\right\}$$

$$sc_{2s} = \frac{1}{\sigma^2}\sum_U x_i \left\{E_s(y_i) - \alpha - \beta x_i\right\}$$

$$sc_{3s} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\left[\sum_U \left\{E_s(y_i) - \alpha - \beta x_i\right\}^2 + \sum_U Var_s(y_i)\right]$$

For non-sample *i*

$$y_i \,|\, \bar{x}_{U-s}, \bar{y}_{U-s} \sim N\left\{\bar{y}_{U-s} + \beta(x_i - \bar{x}_{U-s}), \sigma^2\left(1 - \frac{1}{N-n}\right)\right\}$$

Leads to an available data score with components

$$sc_{1s} = \frac{1}{\sigma^2}\left\{\sum_s (y_i - \alpha - \beta x_i) + (N-n)\left(\overline{y}_{U-s} - \alpha - \beta \overline{x}_{U-s}\right)\right\}$$

$$sc_{2s} = \frac{1}{\sigma^2}\left\{\sum_s x_i (y_i - \alpha - \beta x_i) + (N-n)\overline{x}_{U-s}\left(\overline{y}_{U-s} - \alpha - \beta \overline{x}_{U-s}\right)\right\}$$

$$sc_{3s} = -\frac{(n+1)}{2\sigma^2} + \frac{1}{2\sigma^4}\left\{\sum_s (y_i - \alpha - \beta x_i)^2 + (N-n)\left(\overline{y}_{U-s} - \alpha - \beta \overline{x}_{U-s}\right)^2\right\}$$

Full information MLEs defined by setting these score components to zero and solving for α, β and σ^2

Full Information MLEs

$$\hat{\beta}_{fimle} = \frac{\sum_s (x_i - \bar{x}_s)(y_i - \bar{y}_s) + n\bar{x}_s(\bar{y}_s - \bar{y}_U) + (N-n)\bar{x}_{U-s}(\bar{y}_{U-s} - \bar{y}_U)}{\sum_s (x_i - \bar{x}_s)^2 + n\bar{x}_s(\bar{x}_s - \bar{x}_U) + (N-n)\bar{x}_{U-s}(\bar{x}_{U-s} - \bar{x}_U)}$$

$$\hat{\alpha}_{fimle} = \bar{y}_U - \hat{\beta}_{fimle}\bar{x}_U$$

and

$$\hat{\sigma}^2_{fimle} = \frac{1}{n+1}\sum_s \left(y_i - \hat{\alpha}_{fimle} - \hat{\beta}_{fimle}x_i\right)^2 + (N-n)\left(\bar{y}_{U-s} - \hat{\alpha}_{fimle} - \hat{\beta}_{fimle}\bar{x}_{U-s}\right)^2$$

Pseudo-Likelihood Inference

Kish and Frankel (1974), **Binder (1983)**, **Godambe and Thompson (1986)**

$f(\mathbf{y}_U;\theta)$ = probability density of population Y values

- **If $\mathbf{y}_U$ were observed**, θ would be estimated by solution of $sc_U = \partial_\theta \log f(\mathbf{y}_U;\theta) = 0$
- For any specified value of θ, sc_U defines a **finite population parameter** ('census score'), which we can estimate from the sample data
  - sc_w = sample-weighted estimator of sc_U
  - **maximum pseudo-likelihood estimator** (MPLE) of θ is solution to $sc_w = 0$

Calibration + Pseudo-Likelihood

Assume SRSWOR. There are **three** calibration constraints

- the population size N
- population mean of X
- population mean of Y

$$\mathbf{w}^{cal} = \frac{N}{n}\mathbf{1}_n + N\left[\mathbf{1}_n\ \mathbf{y}_s\ \mathbf{x}_s\right]\begin{bmatrix} \mathbf{1}'_n\mathbf{1}_n & \mathbf{1}'_n\mathbf{y}_s & \mathbf{1}'_n\mathbf{x}_s \\ \mathbf{y}'_s\mathbf{1}_n & \mathbf{y}'_s\mathbf{y}_s & \mathbf{y}'_s\mathbf{x}_s \\ \mathbf{x}'_s\mathbf{1}_n & \mathbf{x}'_s\mathbf{y}_s & \mathbf{x}'_s\mathbf{x}_s \end{bmatrix}^{-1}\begin{pmatrix} 0 \\ \bar{y}_U - \bar{y}_s \\ \bar{x}_U - \bar{x}_s \end{pmatrix}$$

$$\hat{\beta}_{cal} = \left\{\sum\nolimits_s w_i^{cal} x_i (x_i - \bar{x}_{ws})\right\}^{-1} \sum\nolimits_s w_i^{cal} x_i (y_i - \bar{y}_{ws})$$

$$\hat{\alpha}_{cal} = \bar{y}_{ws} - \hat{\beta}_{cal}\bar{x}_{ws}$$

$$\hat{\sigma}^2_{cal} = N^{-1}\sum\nolimits_s w_i^{cal}(y_i - \hat{\alpha}_{cal} - \hat{\beta}_{cal}x_i)^2$$

Model-Based Simulations

% relative efficiencies with respect to 5% trimmed RMSE of MPLE (π^{-1}-weighted)

Parameter	Sampling Scheme	N = 500, n = 20		N = 1000, n = 50		N = 5000, n = 200	
		CAL	MIP	CAL	MIP	CAL	MIP
α	SRS	102.97	133.97	127.36	144.75	143.01	149.52
	PPX	74.51	477.16	70.40	502.14	76.36	622.86
	PPY	118.46	200.88	143.12	210.23	158.57	221.53
β	SRS	81.29	105.89	89.71	101.95	96.11	100.54
	PPX	27.18	198.40	26.07	224.58	27.96	270.35
	PPY	62.77	109.14	73.07	110.25	80.85	116.96
σ^2	SRS	84.00	102.34	93.54	100.12	99.43	100.08
	PPX	57.74	130.88	70.07	142.59	81.59	146.35
	PPY	61.58	98.88	71.09	100.96	87.09	102.44

Example 3: Modelling Probability-Linked Data

Fellegi and Sunter (1969) “Record Linkage is a solution to the problem of recognizing those records in two files which represent identical persons, objects, or events...”

- Y-register contains values of scalar random variable Y
- $\mathbf{X}$-register contains values of vector random variable $\mathbf{X}$
- Modelling of (Y, $\mathbf{X}$) relationship straightforward given a random sample of (Y, $\mathbf{X}$) values. But, we do not have such a sample – instead **probabilistic record linkage** used to link records from the two registers

Toy Assumptions

- Both registers contain N records, with no duplications
- **Complete linkage**: All records on both registers can be linked
- Categorical 'blocking' variable Z recorded on both registers
  - measured without error on both
  - takes Q distinct values $q = 1, 2, \ldots, Q$
  - M_q records in each register with $Z = q$ (so $N = \sum_q M_q$)
  - Linkage errors can only occur within 'blocks'
- We index the records on the linked data set in exactly the same way as we index the $\mathbf{X}$-register

A Model for Linkage Error

$$\mathbf{y}_q^* = \mathbf{A}_q \mathbf{y}_q$$

$\mathbf{A}_q$ is an unknown **random permutation matrix** of order M_q, i.e. entries of $\mathbf{A}_q$ are either zero or one, with a value of one occurring just once in each row and column

- $E_X(\mathbf{A}_q) = \mathbf{E}_q$ (assumed known for the moment)

Non-Informative Linkage $\mathbf{A}_q \perp \mathbf{y}_q \big| \mathbf{X}_q \Rightarrow E_X(\mathbf{y}_q^*) = \mathbf{E}_q E_X(\mathbf{y}_q)$

Exchangeable Linkage Errors

Since linkage process maximises the probability that a 'declared link' is a 'true link', **correct** linkages should be more likely than **incorrect** linkages...

- $\Pr(\text{correct linkage in block } q) = \lambda_q$
- $\Pr(\text{wrong linkage in block } q) = \gamma_q$

$$\mathbf{E}_q = E_X\left(\mathbf{A}_q\right) = \begin{bmatrix} \lambda_q & \gamma_q & \cdots & \gamma_q \\ \gamma_q & \lambda_q & \cdots & \gamma_q \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_q & \gamma_q & \cdots & \lambda_q \end{bmatrix}$$

Estimating Functions with Linked Data

Unbiased estimating function given **correctly-linked** data

$$\mathbf{H}(\theta) = \sum_{i=1}^{N} \mathbf{G}_i(\theta)\left\{y_i - f_i(\mathbf{x}_i;\theta)\right\} = \sum\nolimits_q \mathbf{G}_q(\theta)\left\{\mathbf{y}_q - \mathbf{f}_q(\mathbf{X}_q;\theta)\right\}$$

When used with **probability-linked** data, this becomes

$$\mathbf{H}^*(\theta) = \sum\nolimits_q \mathbf{G}_q(\theta)\left\{\mathbf{y}_q^* - \mathbf{f}_q(\mathbf{X}_q;\theta)\right\}$$

Estimating function is no longer unbiased...

$$E_X\left\{\mathbf{H}^*(\theta_0)\right\} = \sum_q \mathbf{G}_q(\theta_0)\left\{\left(\mathbf{E}_q - \mathbf{I}_q\right)\mathbf{f}_q(\mathbf{X}_q;\theta_0)\right\} \neq \mathbf{0}$$

A bias-corrected estimating function

$$\begin{aligned}\mathbf{H}_{adj}(\theta) &= \mathbf{H}^*(\theta) - \sum_q \mathbf{G}_q(\theta)\left\{\left(\mathbf{E}_q - \mathbf{I}_q\right)\mathbf{f}_q(\mathbf{X}_q;\theta)\right\} \\ &= \sum_q \mathbf{G}_q(\theta)\left\{\mathbf{y}_q^* - \mathbf{E}_q\mathbf{f}_q(\mathbf{X}_q;\theta)\right\}\end{aligned}$$

Variance estimation: Standard Taylor series approximation, leading to a plug-in **'sandwich'** estimator

Logistic Regression

$$\sum_q \mathbf{G}_q(\beta)\left\{\mathbf{y}_q^* - \mathbf{E}_q\mathbf{f}_q(\mathbf{X}_q;\beta)\right\} = \mathbf{0}$$

M (defines MLE when data are correctly linked): $\mathbf{G}_q(\beta) = \mathbf{X}_q'$

A (leads to unbiased estimator in linear model): $\mathbf{G}_q(\beta) = \mathbf{X}_q'\mathbf{E}_q'$

C (second-order optimal)

$$\mathbf{G}_q(\beta) = \partial_\beta\left\{E_X\left(\mathbf{y}_q^*\right)\right\}\left\{Var_X\left(\mathbf{y}_q^*\right)\right\}^{-1} = \mathbf{X}_q'\mathbf{D}_q(\beta)\mathbf{E}_q'\Sigma_q^{-1}(\beta)$$

where

$$\mathbf{D}_q(\beta) = diag\left[f_i(\beta)\{1 - f_i(\beta)\}; i \in q\right]$$

$$\Sigma_q(\beta) = Var\left(\mathbf{y}_q^* \middle| \mathbf{X}_q\right)$$

Some Simulation Results

- Three blocks, $M_1 = 1500$, $M_2 = 300$ and $M_3 = 200$, with independent exchangeable linkage errors in each block
- Two scenarios
  - λ_q correctly specified ($\lambda_1 = \mathbf{1.0}$, $\lambda_2 = \mathbf{0.95}$, $\lambda_3 = \mathbf{0.75}$)
  - λ_q estimated by $\hat{\lambda}_q = \min\left\{m_q^{-1}\left(m_q - 0.5\right), \max\left(M_q^{-1}, l_q\right)\right\}$, with l_q equal to the number of correct links in a random sample of $m_q = 20$ linked records in each of blocks 2 and 3
- $logit\left\{E\left(y_i \middle| x_i\right)\right\} = 1 - 5x_i$, with $x_i \sim Uniform[0,1]$

Focus on Estimation of Slope Parameter

Estimator	Relative Bias	Relative RMSE	Coverage
Scenario 1: Linkage Probabilities Correctly Specified			
Perfect Linkage	1.99	9.27	95.4
Naive	-8.91	11.70	73.2
M	2.68	12.34	96.1
A	2.68	12.26	96.3
C	2.44	11.20	95.4
Scenario 2: Linkage Probabilities Estimated			
Perfect Linkage	2.37	9.87	96.8
Naive	-8.53	11.83	76.0
M	5.74	20.10	97.6
A	3.31	15.77	96.8
C	2.57	12.53	95.6

Related Results

- MIP - **Chambers (2009)**, **Chambers *et al.* (2009)**
- Sample — Register linkage, with non-linkage + linkage errors
    - **Kim and Chambers (2009)**
- Allowing for linkage errors (Register — Register) in linear mixed modelling
    - **Samart and Chambers (2010)**
- Longitudinal linkage with errors (Sample — multiple Registers)
    - **Kim and Chambers (2009)**

The Other Examples

Example 4: Merkouris (*JASA*, 2004) "Combining independent regression estimators from multiple surveys"

- Suggests composite GREG estimator: $\phi\hat{t}_{Ay}^{GREG(X,C)} + (1-\phi)\hat{t}_{By}^{GREG(Z,C)}$
- Alternative BLUP based on

$$\begin{pmatrix} \mathbf{y}_{As} \\ \mathbf{y}_{Bs} \end{pmatrix} = \begin{bmatrix} \mathbf{X}_{As} & \mathbf{C}_{As} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{Bs} & \mathbf{Z}_{Bs} \end{bmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{e}_{As} \\ \mathbf{e}_{Bs} \end{pmatrix}?$$

- Or perhaps

$$\begin{pmatrix} \mathbf{y}_{As} \\ \mathbf{y}_{Bs} \end{pmatrix} = \begin{bmatrix} \mathbf{X}_{As} & \mathbf{C}_{As} & \hat{\mathbf{Z}}_{As} \\ \hat{\mathbf{X}}_{Bs} & \mathbf{C}_{Bs} & \mathbf{Z}_{Bs} \end{bmatrix} \boldsymbol{\beta} + \begin{pmatrix} \mathbf{e}_{As} \\ \mathbf{e}_{Bs} \end{pmatrix}?$$

Example 5: Elliot and Davis (*Applied Statistics*, 2005) "Obtaining cancer risk factor prevalence estimates in small areas: combining data from two surveys"

- Base SAE on survey B data, but use survey A data to modify survey B weights to get rid of the Y-X bias
- Implemented via **propensity weight adjustments** that ensure marginal probability that $Y = 1$ in area g is the same for both survey A and survey B
- Use of propensity-based bias correction increases variances relative to survey B estimators that do not use this correction. However, reduction in bias leads to smaller MSE

If it is possible to identify values of X for survey A sample, then this is item non-response for Y in survey B. MIP can be applied directly.

Example 6: Strauss, Carroll, Bortnick, Menkedick & Schultz (*Biometrics*, 2001): "Combining data sets to predict the effects of regulation of environmental lead exposure in housing stock"

Y = log(blood lead concentration) - NHANES, Rochester Study
H = log(environmental exposure) as measured by HUD
G = log(environmental exposure) as measured by Rochester Study
C = exposure variable common to HUD and Rochester Study

Target Model $Y = \alpha + \beta H + \gamma C + e$

Gaussian measurement error model + Rochester/HUD data used to get estimates of β and γ. Marginal data on Y from NHANES then used to estimate α

Complex, but MIP should be applicable ...

THANK YOU!

References

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review,* **51**, 279-292.

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M. and Welsh, A.H. (1994). Maximum likelihood inference from survey data. *International Statistical Review*, **62**, 349 - 363.

Chambers, R. (2009). Regression analysis of probability-linked data. *Official Statistics Research Series*, Vol **4**, No. 2, Statistics New Zealand, Wellington. (http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm)

Chambers, R., Chipperfield, J., Davis, W. and Kovacevic, M. (2009). Inference based on estimating equations and probability-linked data. *Working Paper* 18-09, Centre for Statistical and Survey Methodology.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**, 1-37.

Deville, J.C. and Särndal, C.E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, **87**, 376-382.

Felligi, I.P. and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64**, 1183-1210

Godambe, V.P. and Thompson, M.E. (1986). Parameters of super populations and survey population: their relationship and estimation. *International Statistical Review,* **54**, 37-59.
Kim, G. and Chambers, R. (2009). Regression analysis under incomplete linkage. *Working Paper* 17-09, Centre for Statistical and Survey Methodology.
Kim, G. and Chambers, R. (2010). Regression analysis for longitudinally linked data. *Working Paper* 22-10, Centre for Statistical and Survey Methodology.
Kish, L. and Frankel, M.R. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society, Series B,* **36**, 1-37.
Orchard, T. and Woodbury, M.A. (1972). A missing information principle: theory and application. *Proc. 6th Berkeley Symp. Math. Statist.*, **1**, 697-715.
Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*}, **71**, 657-664.
Samart, K. and Chambers, R. (2010). Fitting linear mixed models using linked data. *Working Paper* 18-10, Centre for Statistical and Survey Methodology.
Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, **82**, 528-550.