Regression Analysis using Probability Linked Data

Issues and Opportunities

Ray Chambers

Frontiers in Social Statistics Methodology Workshop
8 February 2017

NIASRA
NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

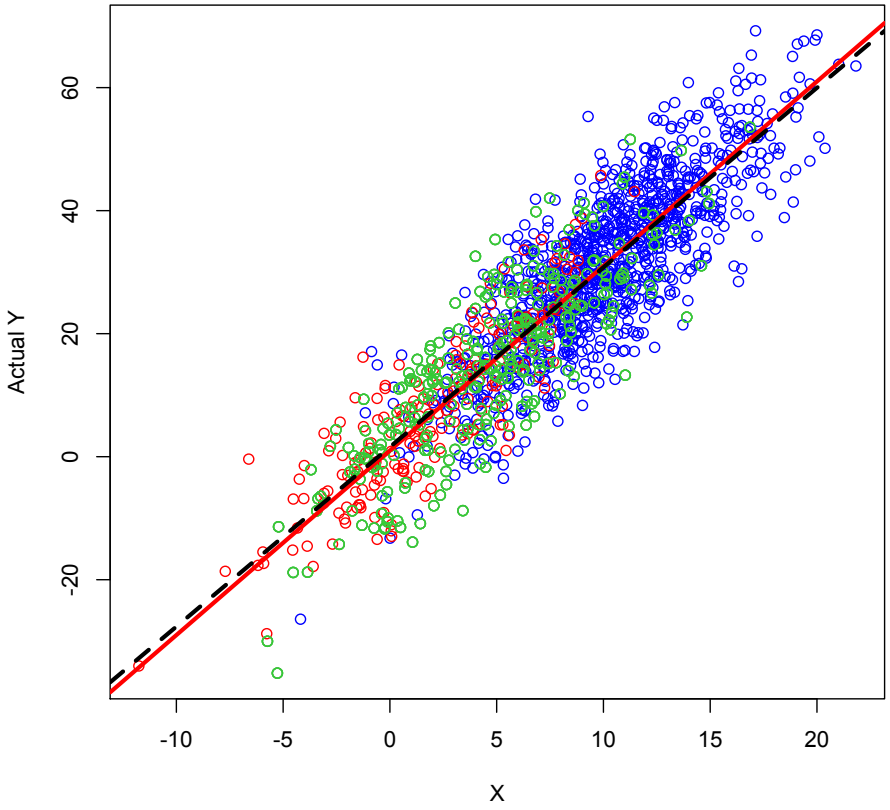
UNIVERSITY OF WOLLONGONG

Focus: Probabilistic Record Linkage

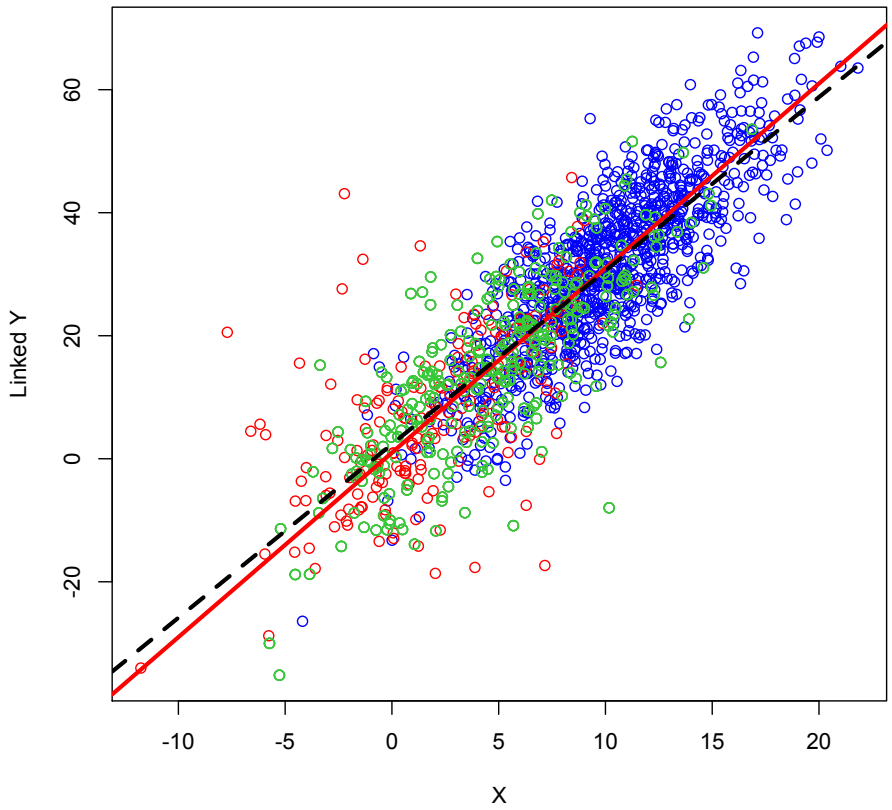
- Technique for linking records in two or more files covering the same population by maximising the probability that the data values in the linked record are from the same population unit
- Applications:
  - merging of large databases
  - elimination of duplicates in registers
  - generating longitudinal records from cross-sectional data
  - combining data from surveys & registers
- Problem: No common unique identifier
- Consequence: Possible linkage errors (linked record made up of data values from different population units)
  - attenuation of statistical relationships in the actual (correctly linked) population

X-Y Illustration - 1

Perfectly linked register data for 30 blocks, each of size 50

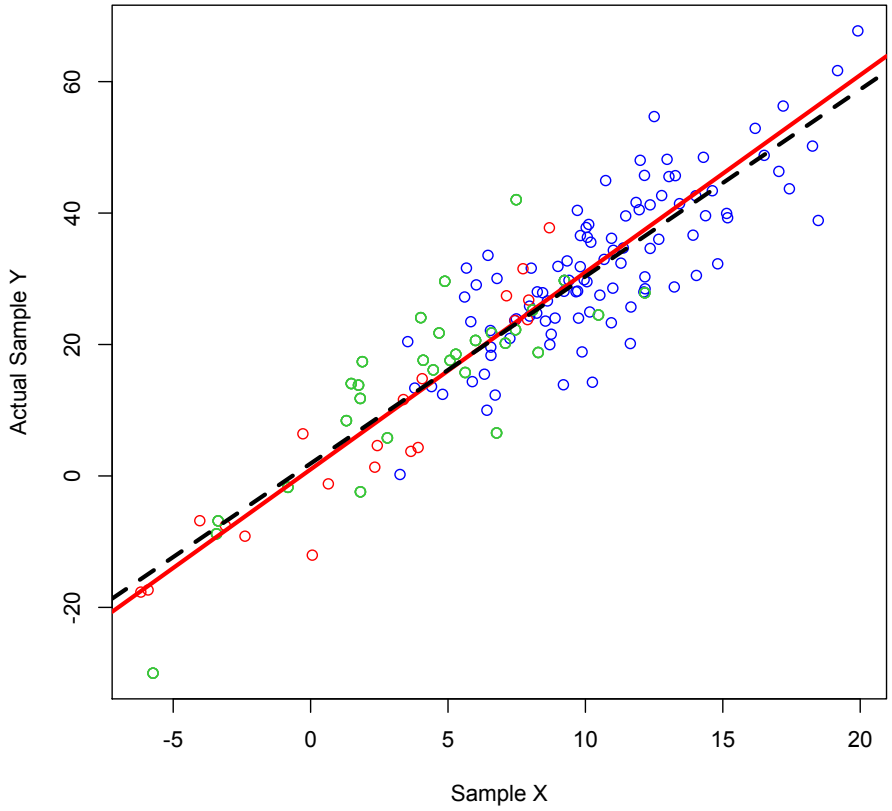


Probability-linked register data with correct linkage probs of 1.0 (1-20), 0.9 (21-26), 0.7 (27-30)

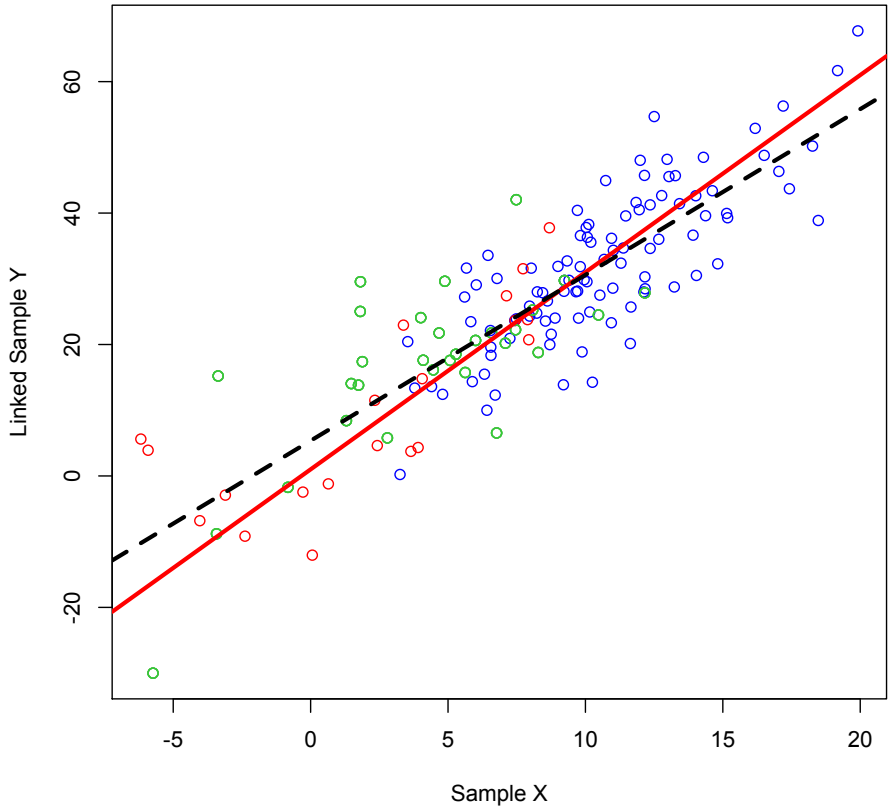


X-Y Illustration - 2

Perfectly linked sample data from the 30 blocks ($m_q = 5$)



Probability-linked sample data - correct linkage probabilities are 1.0 (1-20), 0.9 (21-26), 0.7 (27-30)

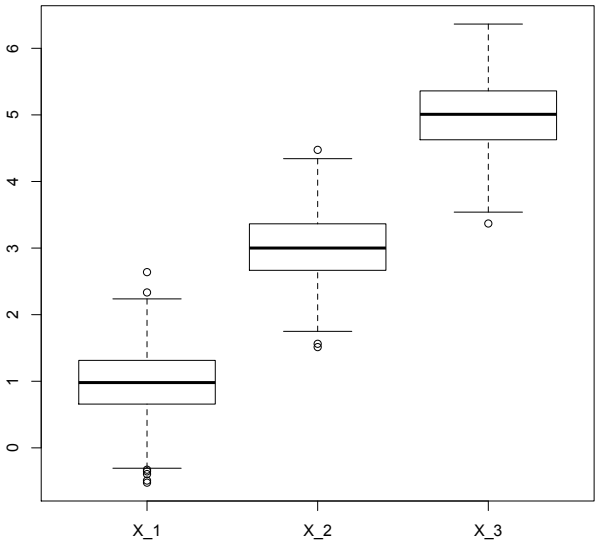


X-Domains Illustration - 1

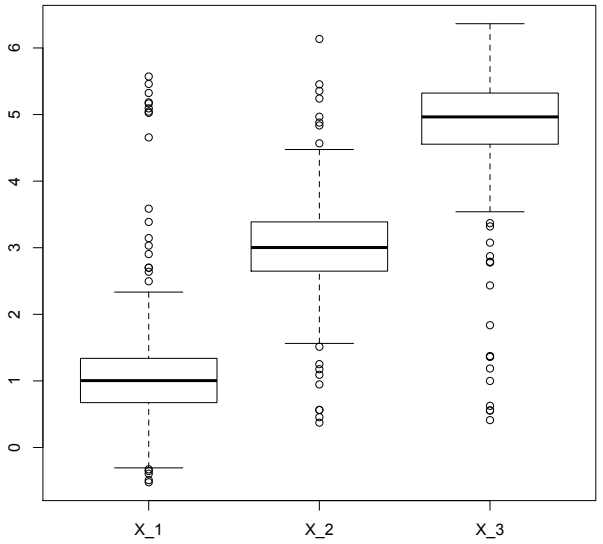
3 X-domains distributed unevenly across 3 blocks

Block	Pr(Correct Linkage)	X-Domain Proportions			M_q
		X=1	X=2	X=3	
1	1.0	0.5	0.3	0.2	1000
2	0.9	0.3	0.5	0.2	300
3	0.7	0.2	0.3	0.5	200
Domain Size (Average Y-value)		630 (1)	510 (3)	360 (5)	1500 (2.64)

Original register data

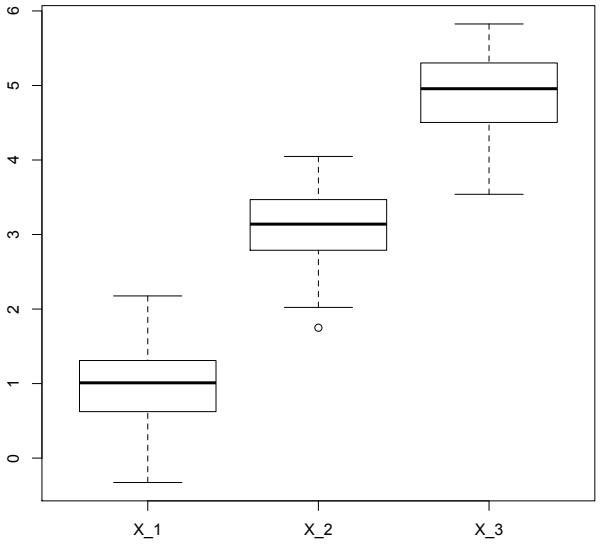


Probability-linked register data

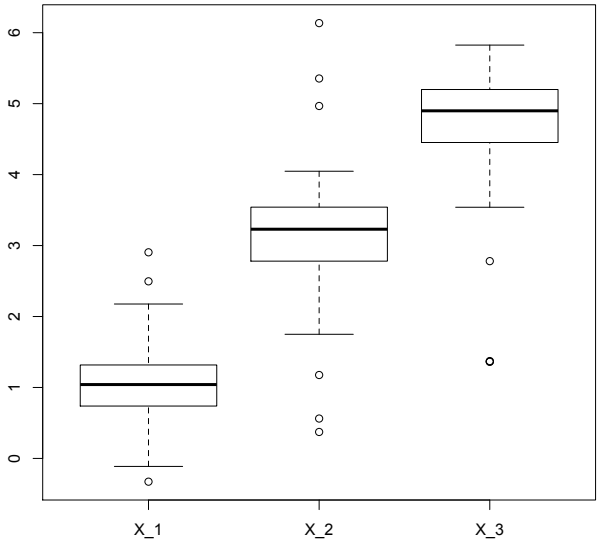


X-Domains Illustration - 2

Perfectly linked sample data: Independent sampling within 3 blocks ($m_q = 50$)



Probability-linked sample data: Block-specific correct linkage probabilities of 1.0, 0.9 and 0.7



Framework for Problem

- Focus on regression modelling using linked data from two population registers
  - … $\mathbf{X}$-register (covariates)
  - … $\mathbf{y}$-register (response)

- Linkage carried out probabilistically (Felligi-Sunter)
  - … linked data set is then pair $(\mathbf{y}^*, \mathbf{X})$

- Too expensive to completely link both registers, so a non-informative (given $\mathbf{X}$) method is used to select a sample of records from $\mathbf{X}$, which are then linked to records in $\mathbf{y}$
  - … identifiers used in this linking process have no errors

More precisely ...

1. Both registers are 1-1 and complete (i.e. they cover the same population)

2. Both registers can be separately partitioned into Q 'match blocks' or just 'blocks'
   - … no errors in block identifiers on each register
   - … records in different blocks can never be linked
   - … subscript of q denotes quantities specific to the q^{th} block

3. Linkage errors within a block are independent of regression errors for that block (non-informative linkage)

4. Sampling from $\mathbf{X}$ ($\mathbf{X}_{sq}$) then linking to $\mathbf{y}$ ($\mathbf{y}^{*}_{sq}$) is stochastically equivalent to directly sampling from $\left(\mathbf{y}^{*}_{q}, \mathbf{X}_{q}\right)$
   - … i.e. sample *then* link is stochastically equivalent to link *then* sample

5. 1-1 linkage, so all records are linkable (in theory)
   - … in practise there will be non-linkable records
   - … 'non-linkage' is at random given $\mathbf{X}_{q}$, so missing links are ignorable
   - … for linking samples to registers, missing links are equivalent to non-sampled records that could potentially be linked

6. Auxiliary information: Correct block averages $\bar{y}_{q}$ and $\bar{\mathbf{x}}_{q}$ are available for each block

Secondary Analyst Perspective

Available data are linked sample data $\left(\mathbf{y}_{sq}^{*}, \mathbf{X}_{sq}\right)$ + block average values $\bar{y}_q$ and $\bar{\mathbf{x}}_q$

Linkage Error Model (LEM)

- Under 1-1 and complete linkage, $\mathbf{y}_q^{*} = \mathbf{A}_q \mathbf{y}_q$, where $\mathbf{A}_q$ is a latent random permutation matrix, e.g.

$$\mathbf{y}_q = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} \text{ and } \mathbf{A}_q = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow \mathbf{y}_q^{*} = \begin{pmatrix} y_3 \\ y_2 \\ y_1 \\ y_4 \\ y_5 \end{pmatrix}$$

- Assumption that blocking is block specific allows us to drop q subscript
  - … all results from now on should be aggregated over blocks

- A simple (unrealistic but pragmatic) LEM for secondary analysis (Chambers, 2009) is the Exchangeable Linkage Errors (ELE) model:

$$\Pr\left\{y_i^* = y_j \middle| \mathbf{X}\right\} = \lambda_{ij} = \begin{cases} \lambda, & i = j, \\ \eta, & i \neq j \end{cases} \qquad i, j = 1, \ldots, M$$

- Within a block,
  - … correct linkages are more likely than incorrect ones
  - … all correct (incorrect) linkages are equally likely

- Consequences

... $E_{|X}(\mathbf{A}) = \mathbf{T} = \begin{bmatrix} \lambda & \eta & \eta & \eta & \eta \\ \eta & \lambda & \eta & \eta & \eta \\ \eta & \eta & \lambda & \eta & \eta \\ \eta & \eta & \eta & \lambda & \eta \\ \eta & \eta & \eta & \eta & \lambda \end{bmatrix}$

... since linkage is 1-1 and complete, $\eta = \dfrac{1-\lambda}{M-1}$

... put $\delta = \lambda - \eta$, then convenient to write $\mathbf{T} = \delta\mathbf{I} + \eta\mathbf{1}\mathbf{1}^T$

Aside: Missing Information Principle

- Correctly linked data are 'missing', and what we observe are the probabilistically linked data + auxiliary information (block averages)

  - … Missing Information Principle (MIP): MLE estimating equations (score functions) defined by replacing all sufficient statistics in the 'complete data' MLE estimating equations by their conditional expectations given the observed data (which includes the probabilistically linked data)

  - … Basis for well-known EM algorithm

  - … Chambers, Steel, Wang and Welsh (2012)

Applying the MIP to the Linkage Error Problem

- Partition linked population vector $\mathbf{y}^*$ into m sampled linked values $\mathbf{y}_s^*$ and $M-m$ non-sampled linked values, $\mathbf{y}_r^*$

$$\ldots \quad \mathbf{A} = \begin{pmatrix} \mathbf{A}_s \\ \mathbf{A}_r \end{pmatrix} \;\&\; \mathbf{T} = \begin{pmatrix} \mathbf{T}_s \\ \mathbf{T}_r \end{pmatrix}$$

$$\ldots \quad \mathbf{f} = E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\beta \Rightarrow \mathbf{f}_s = \mathbf{X}_s\beta$$

$$\ldots \quad \text{ELE} \Rightarrow \mathbf{T}_s = E(\mathbf{A}_s|\mathbf{X}) = cbind\left\{\delta\mathbf{I}_s + \eta\mathbf{1}_s\mathbf{1}_s^T,\ \eta\mathbf{1}_s\mathbf{1}_r^T\right\}$$

- Put $\tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y}_s^* \\ \bar{y} \end{pmatrix}$ and $\tilde{\mathbf{X}} = \begin{bmatrix} \delta \mathbf{X}_s + \eta M \mathbf{1}_s \bar{\mathbf{x}}^T \\ \bar{\mathbf{x}}^T \end{bmatrix}$

... FIMLE for β under normal errors and ELE is then

$$\hat{\beta} = \left(\tilde{\mathbf{X}}^T \mathbf{W}^{-1} \tilde{\mathbf{X}} \right)^{-1} \tilde{\mathbf{X}}^T \mathbf{W}^{-1} \tilde{\mathbf{y}}$$

where

$$\mathbf{W} = \begin{bmatrix} \mathbf{I}_s + \sigma^{-2} Var(\mathbf{A}_s \mathbf{f}) & M^{-1} \mathbf{T}_s \mathbf{1} \\ M^{-1} \mathbf{1}^T \mathbf{T}_s^T & M^{-1} \end{bmatrix}$$

which depends on σ^2 and $Var(\mathbf{A}_s \mathbf{f})$

- Put $\mathbf{c}^T = \begin{pmatrix} \mathbf{T}_s^T & M^{-1}\mathbf{1} \end{pmatrix}$. FIMLE of σ^2 is

$$\hat{\sigma}^2 = \left(trace\left(\mathbf{c}^T \mathbf{W}^{-1} \mathbf{c}\right)\right)^{-1} \left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\beta}\right)^T \mathbf{W}^{-1} \mathbf{c}\mathbf{c}^T \mathbf{W}^{-1} \left(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\beta}\right)$$

- $\mathbf{W}$ depends on σ^2 and $Var(\mathbf{A}_s\mathbf{f})$. In contrast

$$\mathbf{c}\mathbf{c}^T = \begin{bmatrix} \delta^2 \mathbf{I}_s + \eta\{2\delta + \eta M\}\mathbf{1}_s\mathbf{1}_s^T & M^{-1}(\delta + M\eta)\mathbf{1}_s \\ M^{-1}(\delta + M\eta)\mathbf{1}_s^T & M^{-1} \end{bmatrix}$$

depends only on λ

- Can solve for β and σ^2 iteratively, after substituting for λ
  - … requires expression for $Var(\mathbf{A}_s\mathbf{f})$

- Given linked registers, Chambers (2009) shows that under the ELE

$$Var(\mathbf{Af}) \approx diag\left((1-\lambda)\left\{\lambda\left(f_i-\overline{f}\right)^2+\overline{f}^{(2)}-\left(\overline{f}\right)^2\right\}\right)$$

  - … under non-informative sampling within blocks we replace all population averages by sample averages to obtain the corresponding sample approximation

$$Var(\mathbf{A}_s\mathbf{f}) \approx diag\left((1-\lambda)\left\{\lambda\left(f_i-\overline{f}_s\right)^2+\overline{f}_s^{(2)}-\left(\overline{f}_s\right)^2\right\};i\in s\right)$$

Alternative 1: Sample-Based Estimating Equation

- Based on linked sample data only (Kim & Chambers, 2012)
  - … $\mathbf{G}_s$ = user-specified GEE weighting matrix
  - … $\mathbf{w}_s$ = vector of sample weights for the linked sample units
  - … $\tilde{\mathbf{T}}_s = \delta \mathbf{I}_s + \eta \mathbf{1}_s \mathbf{w}_s^T$

- Sample-based estimating equation for β

$$\tilde{\mathbf{H}}_s = \mathbf{G}_s \left(\mathbf{y}_s^* - \tilde{\mathbf{T}}_s \mathbf{X}_s \beta \right) = \mathbf{0}$$

  - … $\hat{\beta} = \left(\mathbf{G}_s \tilde{\mathbf{T}}_s \mathbf{X}_s \right)^{-1} \left(\mathbf{G}_s \mathbf{y}_s^* \right)$
  - … sandwich approximation to variance
  - … moment estimation for σ^2

$$\hat{\sigma}^2 = m^{-1} \left\{ \left(\mathbf{y}_s^* - \mathbf{f}_s \right)^T \left(\mathbf{y}_s^* - \mathbf{f}_s \right) - 2 \mathbf{f}_s^T \left(\mathbf{I}_s - \mathbf{T}_{ss} \right) \mathbf{f}_s \right\}$$

where $\mathbf{T}_{ss} = \delta \mathbf{I}_s + \eta \mathbf{1}_s \mathbf{1}_s^T$

- Three standard choices for the estimation weighting matrix

  ... Least squares (LS) weighting $\mathbf{G}_s = \mathbf{X}_s^T$

  ... Lahiri-Larsen (LL) weighting $\mathbf{G}_s = \mathbf{X}_s^T \tilde{\mathbf{T}}_s^T$

  ... EBLUE (BL) weighting $\mathbf{G}_s = \mathbf{X}_s^T \tilde{\mathbf{T}}_s^T \hat{\Sigma}_s^{-1}$

Alternative 2: Calibrated Sample-Based Estimating Equation

- Uses linked sample data plus register information

... $\widehat{\mathbf{T}}_s = \left[\begin{array}{cc} \delta \mathbf{I}_s & \eta M \mathbf{1}_s \end{array} \right]_{m \times (m+1)}$

... $\widehat{\mathbf{X}}_s = \left[\begin{array}{c} \mathbf{X}_s \\ \overline{\mathbf{x}}^T \end{array} \right]_{(m+1) \times p}$

- Calibrated estimating equation

$$\widehat{\mathbf{H}}_s = \mathbf{G}_s \left(\mathbf{y}_s^* - \widehat{\mathbf{T}}_s \widehat{\mathbf{X}}_s \beta \right) + (M-m) \mathbf{g}_r \left(\overline{y}_r^* - \overline{\mathbf{x}}_r^T \beta \right) = \mathbf{0}$$

... $\hat{\beta} = \left(\mathbf{G}_s \widehat{\mathbf{T}}_s \widehat{\mathbf{X}}_s + (M-m) \mathbf{g}_r \overline{\mathbf{x}}_r^T \right)^{-1} \left(\mathbf{G}_s \mathbf{y}_s^* + (M-m) \mathbf{g}_r \overline{y}_r^* \right)$

... $\mathbf{g}_r = m \hat{\sigma}^2 \left\{ trace(\hat{\Sigma}_s) \right\}^{-1} \overline{\mathbf{x}}_r$

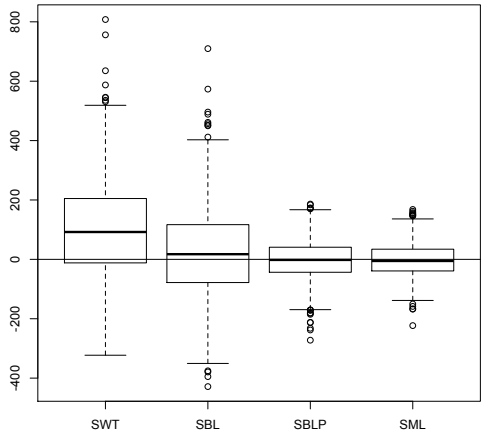
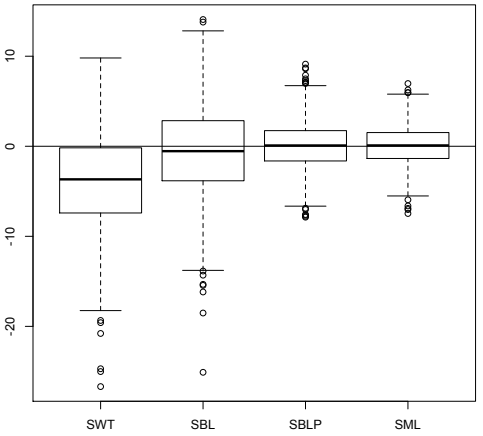
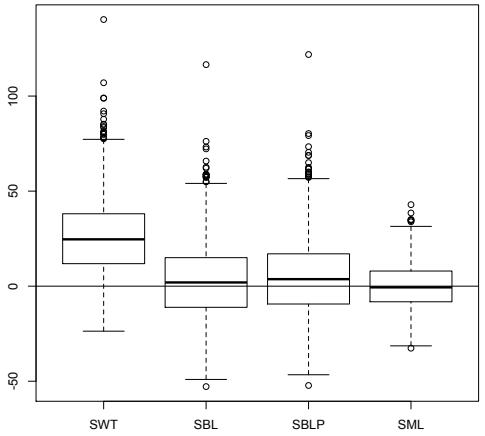
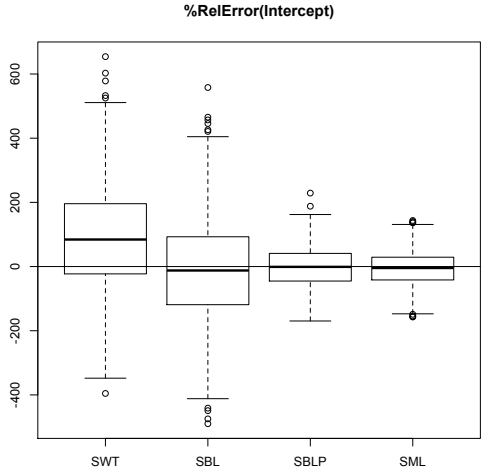
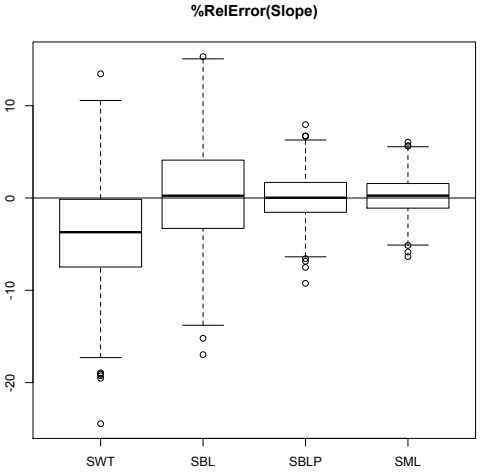
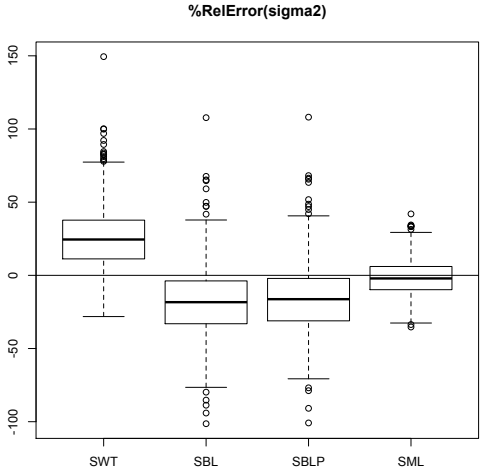
X-Y Simulations

- β = (1,3) and σ^2 = 64 - simulated data shown at start

- Simulation Scenario
  - … 30 blocks, $M_q = 50$
  - … SRSWOR samples, $m_q = 5$
  - … ELE-based linkage errors with λ = 1.0 (B1-B20), 0.9 (B21-B26), 0.7 (B27-B30)
  - … audit sample sizes: 2 each from B21-B30 (total 20)
  - … X: B1-B20: N(**10**,16), B21-B26: N(**5**,16), B27-B30: N(**2**,16)

- SWT Sample weighted estimator that ignores linkage errors
- SBL Solution to sample-based GEE, BL weighting
- SBLP Solution to calibrated GEE, BL weighting
- SML MLE

X-Y Simulations

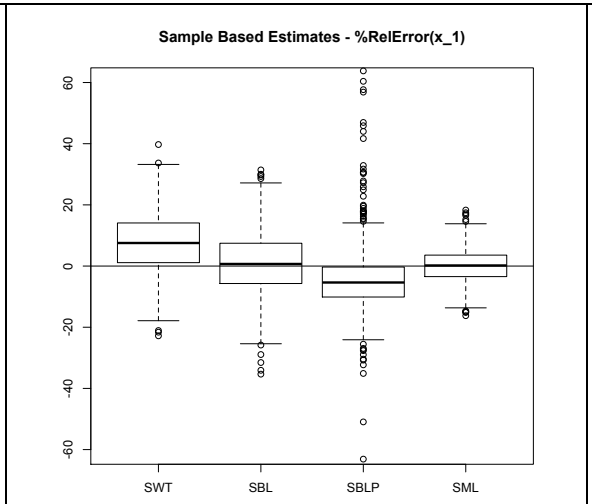
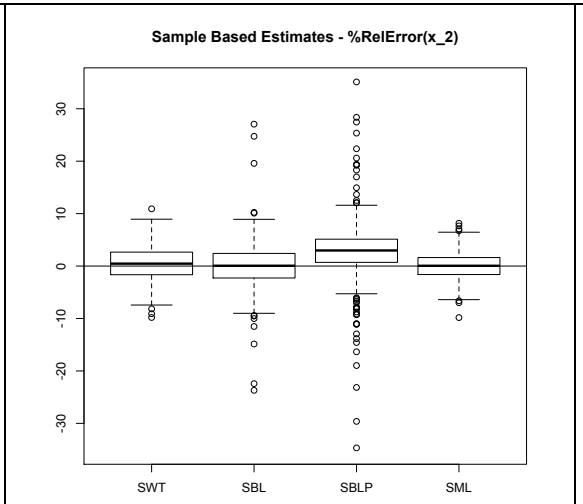
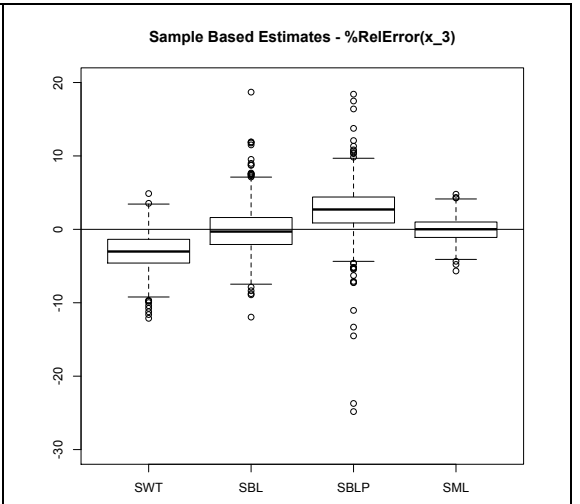
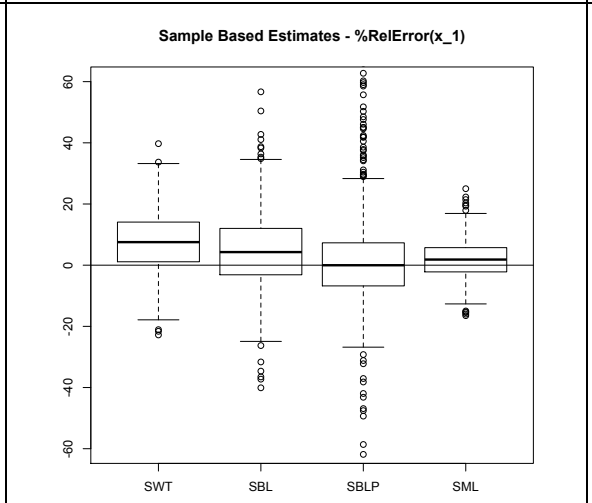
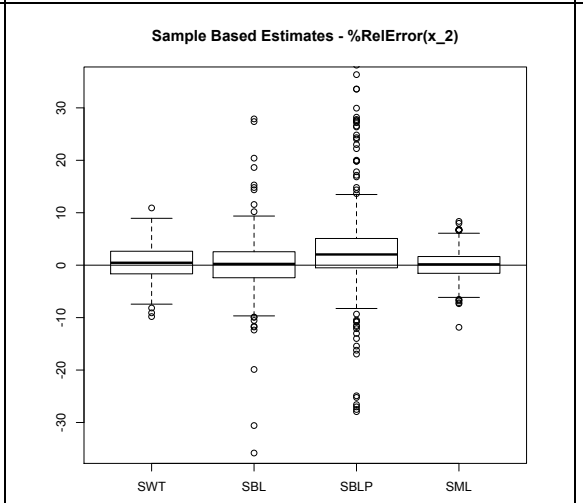
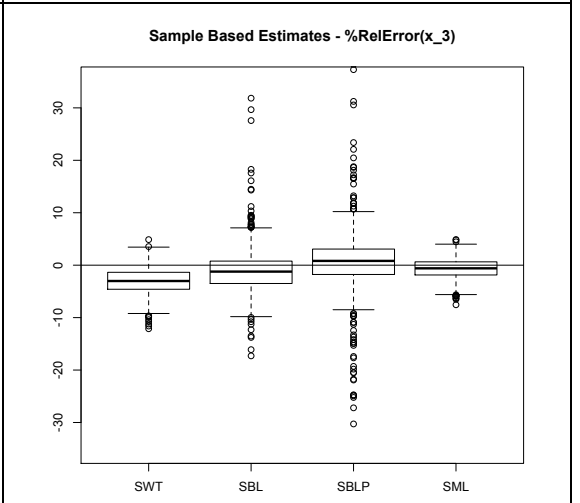
actual value of lambda used			
lambda estimated using audit sample			

X-Domain Simulations

- Scenario
  - 3 blocks, M = 1000, 300, 200
  - SRSWOR samples, m = 50, 50, 50
  - ELL-based linkage errors
  - audit sample sizes: 0, 8, 8 (total 16)
  - X: 3 categories distributed unevenly across blocks

Block	Pr(Correct Linkage)	X-Domain Proportions			M_q
		X=1	X=2	X=3	
1	1.0	0.5	0.3	0.2	1000
2	0.9	0.3	0.5	0.2	300
3	0.7	0.2	0.3	0.5	200
Domain Size (Average Y-value)		630 (1)	510 (3)	360 (5)	1500 (2.64)

X-Domain Simulations

actual value of lambda used			
lambda estimated using audit sample			

An Imputation/Weighting Solution?

James Chipperfield (ABS)

- Choose weighting matrix $\mathbf{G}_s = \mathbf{X}_s^T \tilde{\mathbf{T}}_s^{-1}$. Then

$$\hat{\beta} = \left(\mathbf{G}_s \tilde{\mathbf{T}}_s \mathbf{X}_s\right)^{-1} \left(\mathbf{G}_s \mathbf{y}_s^*\right) = \left(\mathbf{X}_s^T \mathbf{X}_s\right)^{-1} \left(\mathbf{X}_s^T \tilde{\mathbf{T}}_s^{-1} \mathbf{y}_s^*\right) = \left(\mathbf{X}_s^T \mathbf{X}_s\right)^{-1} \left(\mathbf{X}_s^T \hat{\mathbf{y}}_s^*\right)$$

  where $\hat{\mathbf{y}}_s^* = \left(\hat{y}_i^*\right) = \tilde{\mathbf{T}}_s^{-1} \mathbf{y}_s^*$ and $\hat{y}_i^* = \sum_{j \in s} u_{ij} y_j^*$ where $\tilde{\mathbf{T}}_s^{-1} = [u_{ij}]$

- When $\mathbf{X}_s$ corresponds to a set of domain inclusion indicators

$$\mathbf{X}_s^T \hat{\mathbf{y}}_s^* = \left(\sum_{i \in s} x_{ki} \hat{y}_i^*\right) = \left(\sum_{i \in s} x_{ki} \sum_{j \in s} u_{ij} y_j^*\right) = \left(\sum_{j \in s} \left(\sum_{i \in s} u_{ij} x_{ki}\right) y_j^*\right) = \left(\sum_{j \in s} \left(\underbrace{\sum_{i \in s_k} u_{ij}}_{w_{kj}}\right) y_j^*\right)$$

  ... can use ELE bootstrap to estimate $\tilde{\mathbf{T}}_s^{-1}$ if primary data linker, then release domain-specific weights w_{kj} and/or 'adjusted' $\hat{y}_i^*$

Next steps?

- Other useful models for data linkage errors?
- Extension to multi-linked registers
  - … Kim and Chambers (2012b, 2015)
- Privacy issues with LEMs?
- More complex (and more interesting) models?
  - … Samart and Chambers (2014) consider an extension of 2-register ELE to 'REML-type' fitting of a two-level model
- Real data applications?
  - … ABS has applied ideas to their linked data, but nothing else so far
- Software, software, software ....

REFERENCES

Chambers, R. (2009). Regression analysis of probability-linked data. *Statisphere*, Official Statistics Research Series, Volume 4.

Chambers, R.L. and Kim, G. (2015). Secondary analysis of linked data. Chapter 5 (pp. 83-108) in *Methodological Developments in Data Linkage .* (Editors K. Harron, H. Goldstein and C. Dibben). Wiley: Chichester.

Chambers, R.L., Steel, D.G., Wang, S. and Welsh, A.H. (2012). Maximum Likelihood Estimation for Sample Surveys. CRC Press: Boca Raton, Florida.

Chipperfield, J.O. and Chambers, R. (2015). Using the bootstrap to analyse binary data obtained via probabilistic linkage. *Journal of Official Statistics*, **31**, 397-414

Chipperfield, J. (2016). A weighting approach to making inference with probabilistically-linked data. Research Report, Australian Bureau of Statistics.

Kim, G. and Chambers, R. (2012a). Regression analysis under incomplete linkage. *Computational Statistics and Data Analysis*, **56**, 2756-2770.

Kim, G. and Chambers, R. (2012b). Regression analysis under probabilistic multi-linkage. *Statistica Neerlandica*, **66**, 64-79.

Kim, G. and Chambers, R. (2015). Unbiased estimation in the presence of correlated linkage error. *STAT*, **4**, 32–45.

Samart, K. and Chambers, R. (2014). Linear regression with nested errors using probability-linked data. *Australian and New Zealand Journal of Statistics*, **56,** 27-46.