Regression Analysis using Probability Linked Data
Issues and Opportunities

Ray Chambers

Frontiers in Social Statistics Methodology Workshop
8 February 2017
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
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)
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

<table>
<thead>
<tr>
<th>Block</th>
<th>Pr(Correct Linkage)</th>
<th>X-Domain Proportions</th>
<th>$M_q$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>0.5 0.3 0.2</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>0.9</td>
<td>0.3 0.5 0.2</td>
<td>300</td>
</tr>
<tr>
<td>3</td>
<td>0.7</td>
<td>0.2 0.3 0.5</td>
<td>200</td>
</tr>
</tbody>
</table>

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
  - X-register (covariates)
  - y-register (response)

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

- Too expensive to completely link both registers, so a **non-informative** (given \(X\)) method is used to select a **sample** of records from \(X\), which are then linked to records in \(y\)
  - 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 $X_X$ ($X_{sq}$) then linking to $y_y$ ($y^*_y$) is stochastically equivalent to directly sampling from $(y^*_y, X_y)$
   ... 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 $X_q$, 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 $\bar{x}_q$ are available for each block
Secondary Analyst Perspective

Available data are linked sample data \((y^*_s, X_{sq})\) + block average values \(\bar{y}_q\) and \(\bar{x}_q\)

Linkage Error Model (LEM)

- Under 1-1 and complete linkage, \(y^*_q = A_q y_q\), where \(A_q\) is a latent random permutation matrix, e.g.

\[
\begin{align*}
\mathbf{y}_q &= \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} \quad \text{and} \quad A_q = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \\
\Rightarrow y^*_q &= \begin{pmatrix} y_3 \\ y_2 \\ y_1 \\ y_4 \\ y_5 \end{pmatrix}
\end{align*}
\]
• 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:

  \[
  \Pr\{y_i^* = y_j | \mathbf{X}\} = \lambda_{ij} = \begin{cases} 
  \lambda, & i = j, \\
  \eta, & i \neq j 
  \end{cases} \quad i, j = 1, \ldots, M
  \]

• Within a block,
  ... correct linkages are more likely than incorrect ones
  ... all correct (incorrect) linkages are equally likely
Consequences

\[ E_x(A) = T = \begin{bmatrix} \lambda & \eta & \eta & \eta & \eta \\ \eta & \lambda & \eta & \eta & \eta \\ \eta & \eta & \lambda & \eta & \eta \\ \eta & \eta & \eta & \lambda & \eta \\ \eta & \eta & \eta & \eta & \lambda \end{bmatrix} \]

... since linkage is 1-1 and complete, \( \eta = \frac{1-\lambda}{M-1} \)

... put \( \delta = \lambda - \eta \), then convenient to write \( T = \delta I + \eta 11^T \)
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 $y^*$ into $m$ sampled linked values $y_s^*$ and $M - m$ non-sampled linked values, $y_r^*$

\[
... \quad A = \begin{pmatrix} A_s \\ A_r \end{pmatrix} \quad \text{&} \quad T = \begin{pmatrix} T_s \\ T_r \end{pmatrix}
\]

\[
... \quad f = E(y|X) = X\beta \Rightarrow f_s = X_s\beta
\]

\[
... \quad \text{ELE} \Rightarrow T_s = E(A_s|X) = cbind\left\{ \delta_{1s} + \eta_{1s}1_s^T, \eta_{1s}1_r^T \right\}
\]
• Put \( \tilde{y} = \begin{pmatrix} y_s^* \\ \bar{y} \end{pmatrix} \) and \( \tilde{X} = \begin{bmatrix} \delta X_s + \eta M 1_s \bar{x}^T \\ \bar{x}^T \end{bmatrix} \)

... FIMLE for \( \beta \) under normal errors and ELE is then

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

where

\[
W = \begin{bmatrix}
I_s + \sigma^{-2} Var(A_s f) & M^{-1} T_s 1 \\
M^{-1} T_s^T & M^{-1}
\end{bmatrix}
\]

which depends on \( \sigma^2 \) and \( Var(A_s f) \)
• Put $c^T = \begin{pmatrix} T_s^T & M^{-1}1 \end{pmatrix}$. FIMLE of $\sigma^2$ is

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

• $W$ depends on $\sigma^2$ and $\text{Var}(A_s f)$. In contrast

$$cc^T = \begin{bmatrix}
\delta^2I_s + \eta\{2\delta + \eta M\}1_s1_s^T & M^{-1}(\delta + M\eta)1_s \\
M^{-1}(\delta + M\eta)1_s^T & M^{-1}
\end{bmatrix}$$

depends only on $\lambda$
• Can solve for $\beta$ and $\sigma^2$ iteratively, after substituting for $\lambda$
  ... requires expression for $\text{Var}(\mathbf{A}_s \mathbf{f})$

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

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

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

$$\text{Var}(\mathbf{A}_s \mathbf{f}) \approx \text{diag}\left( (1 - \lambda) \left\{ \lambda \left( f_i - \bar{f}_s \right)^2 + \bar{f}^{(2)}_s - \left( \bar{f}_s \right)^2 \right\}; i \in s \right)$$
**Alternative 1: Sample-Based Estimating Equation**

- Based on **linked sample data** only (Kim & Chambers, 2012)
  - \( G_s \) = user-specified GEE weighting matrix
  - \( w_s \) = vector of sample weights for the linked sample units
  - \( \tilde{T}_s = \delta I_s + \eta 1_s w_s^T \)

- Sample-based estimating equation for \( \beta \)
  \[
  \tilde{H}_s = G_s \left( y^*_s - \tilde{T}_s X_s \beta \right) = 0
  \]
  - \( \hat{\beta} = \left( G_s \tilde{T}_s X_s \right)^{-1} \left( G_s y^*_s \right) \)
  - sandwich approximation to variance
  - moment estimation for \( \sigma^2 \)
  \[
  \hat{\sigma}^2 = m^{-1} \left\{ \left( y^*_s - f_s \right)^T \left( y^*_s - f_s \right) - 2 f_s^T \left( I_s - T_{ss} \right) f_s \right\}
  \]
  where \( T_{ss} = \delta I_s + \eta 1_s 1_s^T \)
• Three standard choices for the estimation weighting matrix

... Least squares (LS) weighting \[ G_s = X_s^T \]

... Lahiri-Larsen (LL) weighting \[ G_s = X_s^T \tilde{T}_s^T \]

... EBLUE (BL) weighting \[ G_s = X_s^T \tilde{T}_s^T \hat{\Sigma}_s^{-1} \]
Alternative 2: Calibrated Sample-Based Estimating Equation

- Uses linked sample data plus register information
  \[ \hat{T}_s = \begin{bmatrix} \delta I_s & \eta M 1_s \end{bmatrix}_{m \times (m+1)} \]
  \[ \hat{X}_s = \begin{bmatrix} X_s \\ \bar{x}^T \end{bmatrix}_{(m+1) \times p} \]

- Calibrated estimating equation
  \[ \hat{H}_s = G_s \left( y_s^* - \hat{T}_s \hat{X}_s \beta \right) + (M - m) g_r \left( \bar{y}_r^* - \bar{x}_r^T \beta \right) = 0 \]
  \[ \hat{\beta} = \left( G_s \hat{T}_s \hat{X}_s + (M - m) g_r \bar{x}_r^T \right)^{-1} \left( G_s y_s^* + (M - m) g_r \bar{y}_r^* \right) \]
  \[ g_r = m \hat{\sigma}^2 \left\{ \text{trace}(\hat{\Sigma}_s) \right\}^{-1} \bar{x}_r \]
X-Y Simulations

- $\beta = (1,3)$ and $\sigma^2 = 64$ - simulated data shown at start

- **Simulation Scenario**
  - ... 30 blocks, $M_q = 50$
  - ... SRSWOR samples, $m_q = 5$
  - ... ELE-based linkage errors with $\lambda = 1.0$ (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

<table>
<thead>
<tr>
<th>Block</th>
<th>Pr(Correct Linkage)</th>
<th>X-Domain Proportions</th>
<th>( M_q )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.0</td>
<td>X=1: 0.5, X=2: 0.3, X=3: 0.2</td>
<td>1000</td>
</tr>
<tr>
<td>2</td>
<td>0.9</td>
<td>X=1: 0.3, X=2: 0.5, X=3: 0.2</td>
<td>300</td>
</tr>
<tr>
<td>3</td>
<td>0.7</td>
<td>X=1: 0.2, X=2: 0.3, X=3: 0.5</td>
<td>200</td>
</tr>
</tbody>
</table>

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 $G_s = X_s^T \tilde{T}_s^{-1}$. Then
  $$\hat{\beta} = (G_s \tilde{T}_s X_s)^{-1} (G_s y_s^*) = (X_s^T X_s)^{-1} (X_s^T \tilde{T}_s^{-1} y_s) = (X_s^T X_s)^{-1} (X_s^T \hat{y}_s^*)$$
  where $\hat{y}_s^* = (\hat{y}_i^*) = \tilde{T}_s^{-1} y_s$ and $\hat{y}_i^* = \sum_{j \in s} u_{ij} y_j^*$ where $\tilde{T}_s^{-1} = [u_{ij}]$

- When $X_s$ corresponds to a set of domain inclusion indicators
  $$X_s^T \hat{y}_s^* = \left( \sum_{i \in s} x_{ki} \hat{y}_i^* \right) = \left( \sum_{i \in s} x_{ki} \sum_{j \in s} u_{ij} y_j^* \right) = \left( \sum_{j \in s} \left( \sum_{i \in s} u_{ij} x_{ki} \right) y_j^* \right) = \left( \sum_{j \in s} \left( \sum_{i \in s_{\mathcal{E}}} u_{ij} \right) y_j^* \right)$$

  ... can use ELE bootstrap to estimate $\tilde{T}_s^{-1}$ 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


