Statistics Canada Statistique Canada

Canada

Statistics Canada
www.statcan.gc.ca

Analyzing Linked Data Sets: Understanding the Association between Linkage Errors and Analysis

ITSEW 2014, Washington

Karla Fox

- Record Linkage
  - Probabilistic Record Linkage
  - Deterministic Record Linkage
- Linkage Errors
  - Types of Errors
  - Source of Errors
  - Estimating Errors
- Analysis of a Linked File
  - Impact of Errors
  - Methods that Adjust for Errors

What is Record Linkage?

- Fellegi & Sunter (1969):

  "... [is] a solution to the problem of recognizing those records in two files which represent identical persons, objects, or events (said to be matched)."

- Exact Matching (Deterministic or Probabilistic)
  - To match the same individuals

- Statistical Matching
  - To create a joint distribution from several marginal distributions

Deterministic Linkage

- Unique Key
- String Edit
- Sound or Phonetic
- Distance Measures
- Bi-partite Graphs

- Supervised-Unsupervised Learning Methods

Probabilistic Record Linkage

- **Define**

$$A_1 = \{\textbf{true matches}\}$$
$$A_2 = \{\textbf{possible matches}\}$$
$$A_3 = \{\textbf{non-matches}\} \quad \gamma(a,b) = \textbf{comparison of record pair } (a,b)$$

- **Classification Problem**
  - **Each record pair (a,b) is to be classified as a match (A_1) or a possible-match (A_2) non-match (A_3) based on a likelihood ratio**

$$\frac{P(\gamma(a,b)|M)}{P(\gamma(a,b)|U)}$$

- **An Optimal Linkage Rule is defined as one that minimizes the A2 'possible' links for fixed error levels in A1 and A3**

Basic Approach

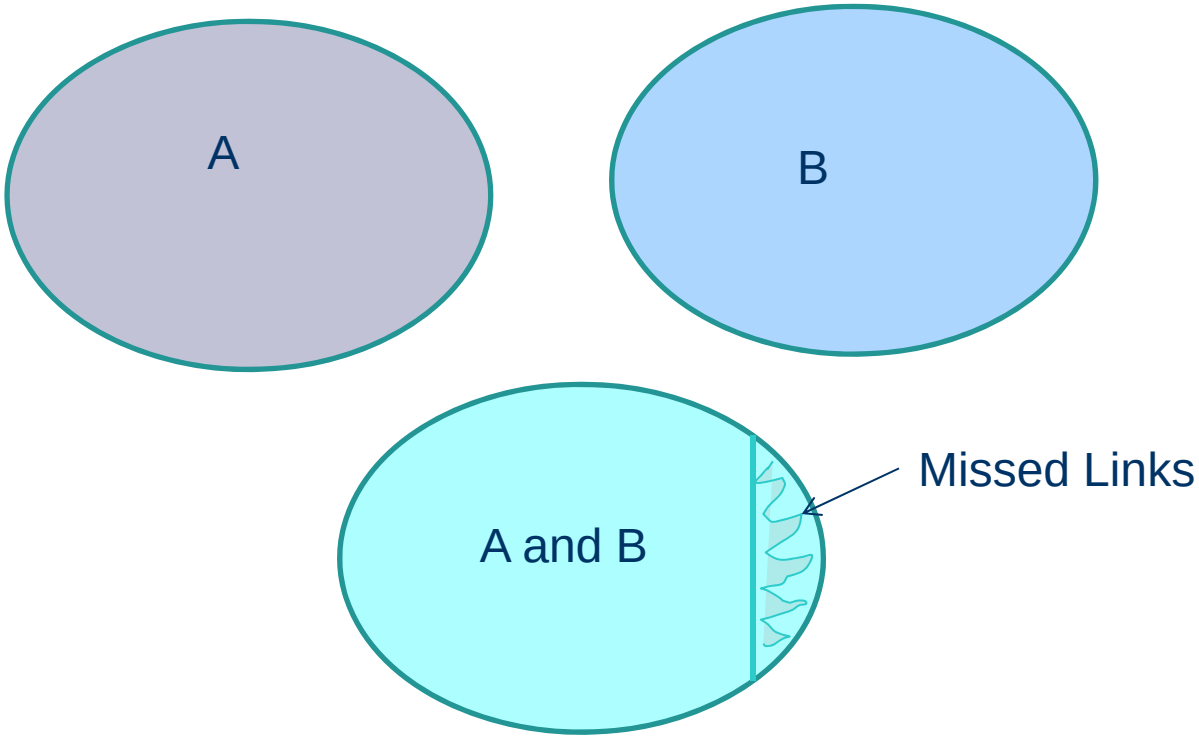
- **We will create a Cartesian product from the two files.**
  - **We do optimize things and only compare units that make sense to compare using Blocks.**
- **We will review each pair and using some form of *rule* decide if the pair is a true pair or not.**

Linkage Errors: Exact Matching

- **False Match: A record is incorrectly matched (false positive)**
  - **Impossible link: Records with no true match that are matched**
  - **Incorrect link: Record with a true match matched to an incorrect match (within the set or outside the set)**

- **Missed Match: A record is not matched when it should be (false negative)**

The Trivial Case



A and B are the same population

Missed Links

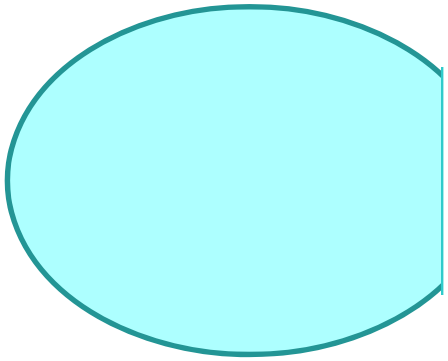
- **Links that were missed can be analyzed as a standard missing data problem**
  - **There will be a missing link mechanism that could be:**

    Missing Completely at Random

    Missing at Random

    Non-Ignorable

The Other Error we have in Linkage



John Smith, 43
John Smith, 43

J. Smith, 43
John Smith, 43

Incorrect Links are PERMUTATIONS, if we are lucky.
If we are not lucky they are links to data outside of our study population.

Linkage Errors

		Actual		
		Matched	Unmatched	
Inferred	Matched	T^+	F^+ *	PPV= $(T^+)/(T^+ + F^+)$
	Unmatched	F^- ***	T^-	NPV= $(T^-)/(T^- + F^-)$
		MMR $=(F^-)/(T^+ + F^-)$ (Type I Error) Sensitivity=1-MMR	FMR$=(F^-)/(T^+ + F^-)$ (Type II Error) Specificity=1-FMR	

Recapping Linkage Errors

- **Incorrect Links**
  - **These are a permutation of the correct set (when there is no coverage issues)**

    Or
  - **These are a link to a incorrect set (when we have coverage issues or we link when there should be no link)**
- **Missed Links**
  - **These are missing data problem**
  - **But without a known N as with survey non-response**

What Gives Rise to the Errors?

- **Heterogeneity/Homogeneity**
- **File Size**
- **Number of Variables for Matching**
- **Quality of the Variables for Matching**
- **Thresholds set in the Matching Methods**

Analysis with linked data

We analyze the linked file in order to understand or make inferences about X and Y as if they both observed from the same unit.

Analysis with Linked Data

As with the other errors we have talked about here RL errors affect:

- **Point Estimates**
- **Variance**
- **Hypothesis Testing**
  - **Type I**
  - **Type II**

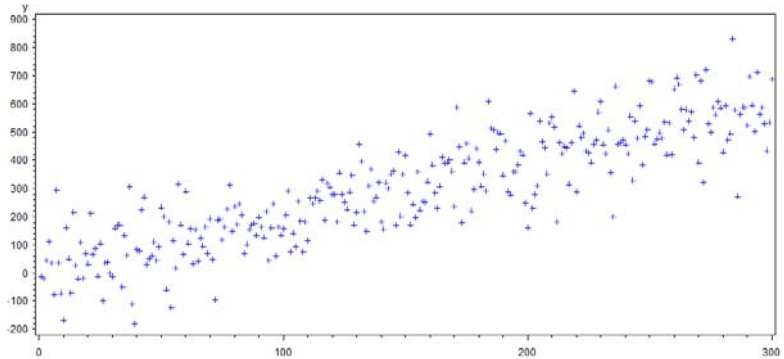
Rates:

If the events are measured in one data set and the population at risk comes from the linked file we have:

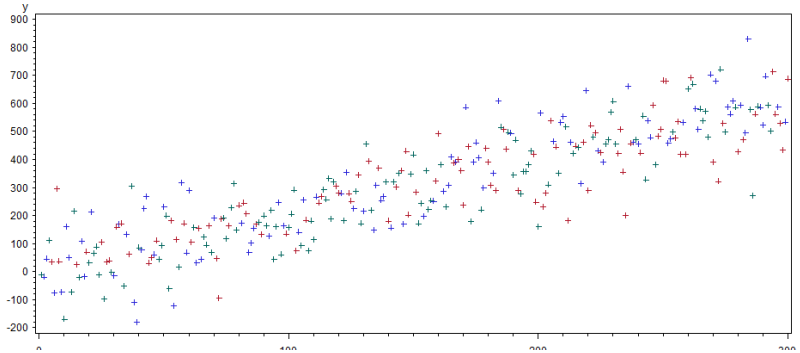
- **False Positive is when a unit of the data set is incorrectly labelled to have the event (incorrect link)**
- **False Negative is when a unit does in fact have the event but is considered to not have the event (missed link)**
- **False Positives will increase the ratio**
- **False Negatives will decrease the ratio**

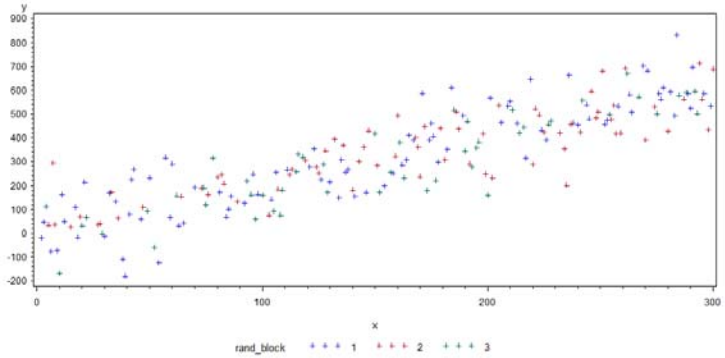
A simple regression illustration

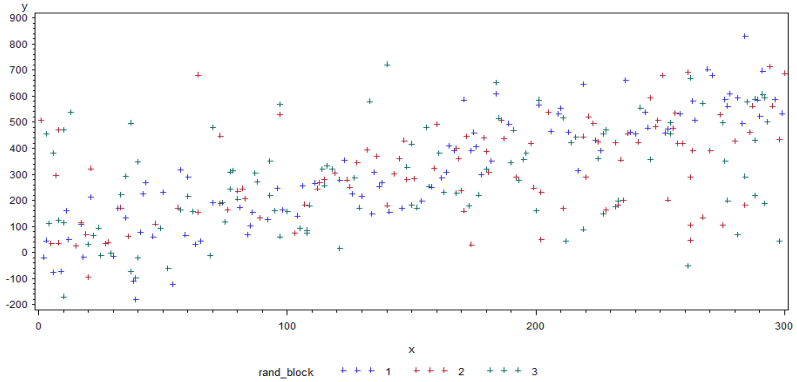
300 data points, simple ratio (y=2x+e)



Put the data into random blocks







Source: Modified version of M. Kovacevic's simulations

Permutations in one-to-one matching

- **For one-to-one matching, (assuming the linkage set has no over-coverage) the assignment matrix has the following properties:**
  - **Each row of the assignment matrix contains only one 1 with other values set to zero. The same holds for the columns.**
  - **When an *i-th diagonal element is equal to 1 it means that there was a correct record linkage of the i-th unit.***
  - **The assignment matrix for the entire file is a block diagonal matrix**

Analysis with Linked Data: Approaches

- **Y is observed in one data set and X is observed in another**
  - **Divide Y into blocks where Pr(correct linkage) = λ_q and Pr(incorrect linkage) = γ_q**
  - **Model the relationship between true y and matched y using an random permutation matrix of order M_q based on parameter estimates of a RL model**

$$y_q^* = A_q y_q \;\; ; \;\; E_X(A_q) = E_q \qquad \textbf{then}$$

$$E_q = (\lambda_q - \gamma_q) I_q + \gamma_q 1_q 1_q^T$$

Methods that Adjust for Record Linkage Errors in Analysis

- **Generalized Estimating Equation**

If we assume no linkage errors we have

$$H^*(\hat{\theta}^*) = \sum_q G_q(\hat{\theta}^*)\left\{y_q^* - f_q(\hat{\theta}^*)\right\} = 0$$

- **Consider only incorrect links (no missing links, no impossible links)**
- **Linkage is non-informative**

with linkage errors the biased corrected version

$$H^*_{adj}(\theta) = \sum_q G_q(\theta)\left\{y_q^* - E_q f_q(\theta)\right\} = 0$$

Methods that Adjust for Record Linkage Errors in Analysis

- **Lahari-Larsen**

We let $\theta \equiv \beta$ **and** $f_q(\beta) = X_q \beta$ **with** $G_q = X_q^T E_q^T$

- **BLU Estimator**

We let $\theta \equiv \beta$ **and** $f_q(\beta) = X_q \beta$ **with** $G_q = X_q^T E_q^T \Sigma_q^{-1}$

- **EBLUP (Chambers et al)**

Iterate between the BLUE β **and** $\hat{\sigma}^2$ **where**

$$\hat{\sigma}^2 = N^{-1}\left\{\sum_q (y_q^* - f_q)^T (y_q^* - f_q) - 2\sum_q f_q^T (I_q - E_q) f_q\right\}$$

Methods that Adjust for Record Linkage Errors in Analysis

- **Missing Links in Outcome Studies (Wang et al)**
  - **Only those with the outcome can be linked so the missing mechanism is non-ignorable , then assume missing linkage among those with the outcome is random and is due to erroneous or incomplete records**
  - **Estimate missing from the outcome dataset**

$$\hat{q} = \sum_{i=1}^{n} \frac{Y_i}{N}$$

  - **Where Y_i is 1 if there is a link and N is the total number of cases**
  - **MLE when the missing rate is known**
  - **GEE when the missing rate is estimated**

Open Research

Better Models for Linked Data:

- There needs to be work on models adjusted with both missed links and incorrect links.

Better Estimation of Linkage Errors:

- Estimation of linkage errors
  - Pair-wise models
  - Group-wise models
- Analysis under known linkage errors
- Joint linkage and analysis/estimation

Questions?