Robust regression using probabilistically linked data

Suojin Wang, Department of Statistics, Texas A&M University

Based on joint work with Ray Chambers, Nicola Salvati, Enrico Fabrizi, M.G. Ranalli

Probabilistic record linkage

- 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 ...”;
- Variables present in both files (matching variables) are used to link records in order to maximize the probability that they refer to the same unit.
- Record linkage is now very widely used.
    - Medical and epidemiological applications predominate.
    - Trusted Third Party (TTP) data linkage by the Western Australia Data Linkage Unit led to 708 research outputs over 1995-2003 (Brook et al., 2008).

Measurement error in the American Community Survey

- Boudreaux et al. (2015) use linkage to examine measurement error in Medicaid coverage for the 2009 American Community Survey (ACS).
  - Sample size was over 4 million persons.
- ACS records linked to enrollment records from the Medicaid Statistical Information System (MSIS)
  - Only 78.4% of linked records that were coded as enrolled for Medicaid on MSIS were also recognized as Medicaid enrollments by the ACS.

Using linked individual patient data to identify COVID risk factors

- OpenSAFELY analytics platform provides access to linked patient data from all the hospital registers of the UK National Health Service.
  - This dataset is large – 20 billion rows of data for about 58 million patients.
  - Only aggregated results are viewable by researchers.
- This linked data resource was used to provide insights into the risk factors associated with Covid-related infection, hospitalisation and mortality during the early stages of the pandemic in the UK (Williamson et al., 2020; Mathur et al., 2021).

Record linkage is not perfect

- Linked data are obtained by integrating two or more distinct data sources.
- Measurement errors can arise because the data held on the contributing sources are not precisely the data that would be collected from a study carried out on a single target population.
- Not all records in the different sources can be linked.
- Not all matches identified by linkage processes are “perfect”.
- In these cases probabilistic linkage methods are typically used.
    - Correct linkage rates of 75%-95% have often been reported in past studies.

Bias due to linkage errors

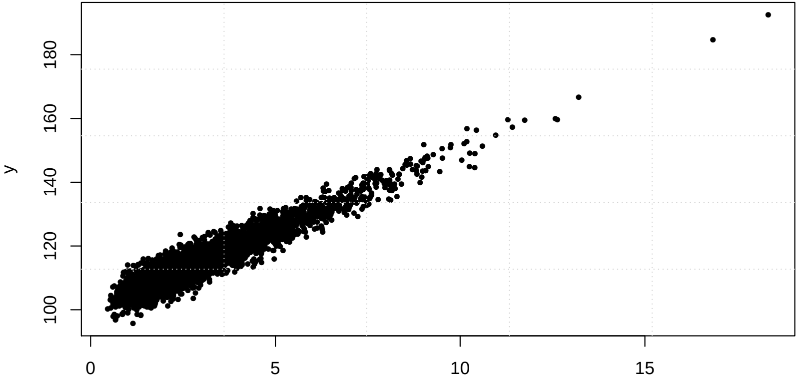
- Linkage errors lead to bias when the linked data are used to fit statistical models for the "correctly linked" data.
- Standard estimation methods (e.g., ordinary least squares) need to be modified to remove this bias.
  - Requires analyst to incorporate knowledge about the statistical characteristics of the linkage process into a model for the linked data.
  - The appropriate model for inference given linked data should combine a model for the linkage error with a model for the process underpinning the correctly linked data.

Linkage errors can be confounded with other model errors

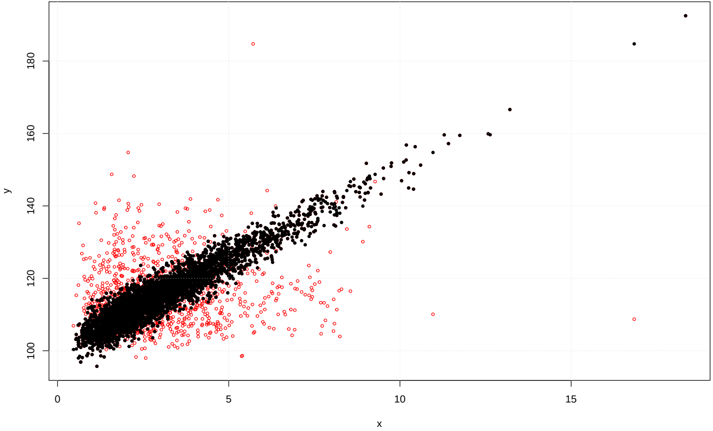
- Increased likelihood of model misspecification when inferential models are based on the linked data alone.

  $\Longrightarrow$ "linkage robust" statistical approach
- For example, linkage errors can lead to outliers in the linked data and thus in the sample.
  - Sample outliers caused by linkage errors are non-representative, i.e., they are not true values.
  - This can lead to biases even when modern outlier robust estimation methods are used.

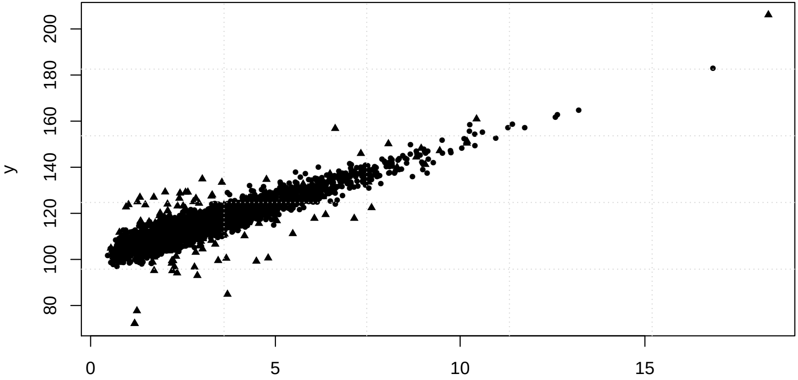
Linear regression illustration – no linkage errors



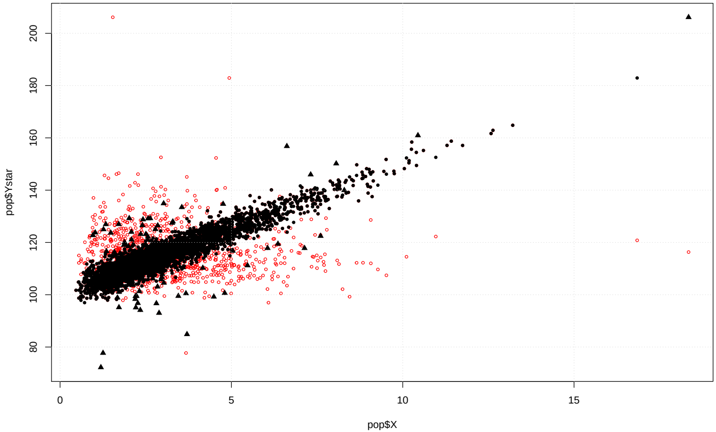
Linear regression illustration – only linkage errors



Linear regression illustration – only outliers



Linear regression illustration – linkage errors + outliers



Notation and assumptions - 1

- Initial focus on linear regression using linked data from two population registers:
  - $\mathcal{Y}$ register (target variable y).
  - $\mathcal{X}$ register (covariates X).
- Linked register is composed of records $(y^{\star}, \mathrm{X})$.
- Both registers are $1-1$ and complete.
- The $\mathcal{Y}$ and $\mathcal{X}$ registers have complete coverage of the same population U of size N, with no duplicates.
- $\mathcal{X}$ register includes a set of identifiers which can be used to partition the linked register into Q blocks, with each block containing the records for N_q individuals.
  - There is no between blocks linkage error.

Notation and assumptions - 2

- Individual linked sample $(y^\star, \mathrm{X})$ values in block q.
- Auxiliary information from linked register
  - Block q averages $\bar{\mathrm{x}}_q$ of covariates from the $\mathcal{X}$ register.
  - Block q average $\bar{y}_q^\star$ for $y^\star$. Since linkage is $1-1$ and complete, $\bar{y}_q^\star = \bar{y}_q$.
- Linkage paradata: Limited information about the accuracy of the linkage process (possibly derived from an audit sub-sample).

Modeling the linkage error - 1

- Under $1-1$ and complete linkage, $\mathrm{y}_q^\star = \mathrm{A}_q \mathrm{y}_q$, where $\mathrm{A}_q = [a_{jk}^q]$ is a latent random permutation matrix of order N_q.
- e.g.,

$$\mathrm{y}_q = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} \text{ and } \mathrm{A}_q = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \to \mathrm{y}_q^\star = \begin{pmatrix} y_3 \\ y_2 \\ y_1 \\ y_4 \\ y_5 \end{pmatrix}$$

- Partition $\mathrm{X}_q = \begin{bmatrix} \mathrm{X}_{sq} \\ \mathrm{X}_{rq} \end{bmatrix}$, $\mathrm{y}_q = \begin{bmatrix} \mathrm{y}_{sq} \\ \mathrm{y}_{rq} \end{bmatrix}$, $\mathrm{y}_q^\star = \begin{bmatrix} \mathrm{y}_{sq}^\star \\ \mathrm{y}_{rq}^\star \end{bmatrix}$ and $\mathrm{A}_q = \begin{bmatrix} \mathrm{A}_{sq} \\ \mathrm{A}_{rq} \end{bmatrix}$.
- $\mathrm{y}_{sq}^\star$ and X_q are known with $\mathrm{y}_{sq}^\star = \mathrm{A}_{sq}\mathrm{y}_q$; y_q is not observed.

Modeling the linkage error - 2

- A simple (unrealistic but pragmatic) linkage error model within block q for secondary analysis (Chambers, 2009) is the Exchangeable Linkage Errors (ELE) model: with $j, k = 1 \ldots, N_q$, $j \neq k$,

$$\begin{aligned} Pr(\text{correct linkage in block } q) &= Pr(a^q_{jj} = 1) = \lambda_q, \\ Pr(\text{incorrect linkage in block } q) &= Pr(a^q_{jk} = 1) = \gamma_q = \frac{1-\lambda_q}{N_q - 1}. \end{aligned}$$

Then

$$E_A(\mathrm{A}_{sq}) = \mathrm{T}_{sq} = \left[(\lambda_q - \gamma_q)\mathrm{I}_{n_q} \,\middle|\, 0_{rq} \right] + \gamma_q 1_{n_q} 1'_{N_q},$$

$$E_{A,M}(\mathrm{y}^\star_{sq}) = \mathrm{T}_{sq}\mathrm{X}_q\boldsymbol{\beta} = \mathrm{X}^\star_{sq}\boldsymbol{\beta}.$$

Least squares regression (OLS)

- We assume homoskedastic regression errors.
- The naive estimator of the regression coefficients is then

$$\hat{\boldsymbol{\beta}} = \left(\sum_q \mathsf{X}_{sq}^T \mathsf{X}_{sq} \right)^{-1} \sum_q \mathsf{X}_{sq}^T \mathsf{y}_{sq}^{\star}.$$

- This ignores the linked nature of the data, as well as any sample outliers.
- It is biased unless $\mathsf{A}_q = \mathsf{I}_{N_q}$ (no linkage error) since

$$E_{A,M}(\hat{\boldsymbol{\beta}}) = (\sum_q \mathsf{X}_{sq}^T \mathsf{X}_{sq})^{-1} \sum_q \mathsf{X}_{sq}^T \mathsf{X}_{sq}^{\star} \boldsymbol{\beta} \neq \boldsymbol{\beta}.$$

Outlier robust regression

- Linkage errors can lead to outliers in the sample data.
- Estimate $\boldsymbol{\beta}$ with outlier robust weighting:

  Solution of $\sum_q \mathrm{X}_{sq} \mathrm{W}^{\psi}_{sq} (\mathrm{y}^{\star}_{sq} - \mathrm{X}_{sq}\boldsymbol{\beta}) = 0$, where W^{ψ}_{sq} is a n-diagonal weight matrix

$$w_j = \frac{\psi(s^{-1}(y^{\star}_{jq} - \mathrm{x}^T_{jq}\boldsymbol{\beta}))}{s^{-1}(y^{\star}_{jq} - \mathrm{x}^T_{jq}\boldsymbol{\beta})}.$$

  Here s is a robust estimate of the scale of the residuals and ψ is a bounded influence function. Standard choices are the Huber (c=1.345) and Biweight (c=4.685) influence functions.

LE bias corrected regression

- Uses linked sample data $+ \bar{x}_q$ (Kim & Chambers, 2012).
- Unbiased estimating equation for $\boldsymbol{\beta}$:

$$\sum_q G_{sq}(y^\star_{sq} - E_{A,M}(y^\star_{sq})) = 0$$

  - G_{sq} = weighting matrix, $E_{A,M}(y^\star_{sq}) = X^\star_{sq}\boldsymbol{\beta}$, $X^\star_{sq} = \{(\lambda_q - \gamma_q)X_{sq} + \gamma_q N_q 1_{n_q}\bar{x}_q^T\}$.
- Three standard choices for the weighting matrix (Chambers, 2009):
  - Least squares weighting: $G_{sq} = X^T_{sq}$.
  - Lahiri and Larsen (2005) weighting: $G_{sq} = X^{\star T}_{sq}$.
  - Best linear weighting under ELE: $G_{sq} = X^{\star T}_{sq}(\sigma^2_e I_{n_q} + V_{sq})$, where $V_{sq} = E_A(A_{sq} f_{sq} f'_{sq} A'_{sq})$ can be approximated and $f_{sq} = X_{sq}\boldsymbol{\beta}$.
- σ^2_e can be estimated by the method of moments.

Outlier robust version of LE bias corrected regression

Solution to $\sum_q G_{sq} W_{sq}^{\psi\star}(y_{sq}^\star - X_{sq}^\star \boldsymbol{\beta}) = 0$, where

- $W_{sq}^{\psi\star}$ is a diagonal matrix of weights defined by component-wise division of the vector $\psi\left\{\Sigma_{sq}^{-1/2}(y_{sq}^\star - X_{sq}^\star \boldsymbol{\beta})\right\}$ by the vector $\Sigma_{sq}^{-1/2}(y_{sq}^\star - X_{sq}^\star \boldsymbol{\beta})$.
- Σ_{sq} is a robust estimate of $Var(y_{sq}^\star - X_{sq}^\star \boldsymbol{\beta})$.
- ψ is a bounded influence function (Huber or Biweight) with tuning parameter set as required.

Gaussian approximation to MLE under ELE

- Chambers & Diniz da Silva (2020): Data: $\tilde{\mathrm{y}}_q = (\mathrm{y}_{sq}^{\star T}, \bar{\mathrm{y}}_q)^T$ and X_q
  Gaussian copula approximation to the joint distribution of $\tilde{\mathrm{y}}_q$ + application of MIP leads to MLEs for $\boldsymbol{\beta}$ and σ_e^2 based on an augmented Gaussian model with

$$E(\tilde{\mathrm{y}}_q|\tilde{\mathrm{X}}_q) = \tilde{\mathrm{X}}_q\boldsymbol{\beta} \text{ and } V(\tilde{\mathrm{y}}_q|\tilde{\mathrm{X}}_q) = \sigma_e^2\Omega_q,$$

$$\tilde{\mathrm{X}}_q = \begin{pmatrix} \mathrm{X}_{sq}^\star \\ \bar{\mathrm{x}}_q^T \end{pmatrix}, \quad \Omega_q = \begin{bmatrix} (\mathrm{I}_{n_q} + \sigma_e^{-2}\mathrm{V}_{sq}) & \left\{N_q^{-1}(\lambda_q - \gamma_q) + \gamma_q\right\} 1_{n_q} \\ \left\{N_q^{-1}(\lambda_q - \gamma_q) + \gamma_q\right\} 1_{sq}^T & N_q^{-1} \end{bmatrix}.$$

- Bias corrected MLE for $\boldsymbol{\beta}$ under the augmented model is solution to

$$\sum_q \tilde{\mathrm{X}}_q^T \hat{\Omega}_q^{-1}(\tilde{\mathrm{y}}_q - \tilde{\mathrm{X}}_q\boldsymbol{\beta}) = 0.$$

- σ_e^2 can be estimated by the corresponding MLE.

Robustified Gaussian MLE

- Estimating equation for RMLE for $\boldsymbol{\beta}$ uses a robust version of $\hat{\Omega}_q$:

$$\sum_q \tilde{X}_q^T \hat{H}_{wq}^{a,b}(\tilde{y}_q - \tilde{X}_q\boldsymbol{\beta}) = 0,$$

$$\hat{H}_{wq}^{a,b} = (\hat{W}_q^{a,b})^{1/2}\hat{\Omega}_q^{-1}(\hat{W}_q^{a,b})^{1/2}$$

$$\hat{W}_q^{a,b} = \mathrm{Diag}\left[\frac{\psi\{(\hat{\sigma}_{sq}^{\star\psi})^{-1}(y_{sq}^{\star} - X_{sq}^{\star}\hat{\boldsymbol{\beta}}); k = a\}}{(\hat{\sigma}_{sq}^{\star\psi})^{-1}(y_{sq}^{\star} - X_{sq}^{\star}\hat{\boldsymbol{\beta}})}, \frac{\phi\{(\hat{\sigma}_q^{\phi})^{-1}N_q(\bar{y}_q - x_q\hat{\boldsymbol{\beta}}); k = b\}}{(\hat{\sigma}_q^{\phi})^{-1}N_q(\bar{y}_q - x_q\hat{\boldsymbol{\beta}})}\right],$$

  $\hat{\sigma}_{sq}^{\star\psi}$ and $\hat{\sigma}_q^{\phi}$ are robust estimators of the scale based on the sample and non-sample residuals, a and b are tuning constant values for the influence functions.

- Corresponding RMLE for σ_e^2 is also provided.

- Estimators of β

... Naive estimator - OLS

... Outlier robust M-estimator - ROB

... Best lnear weighting under ELE - BD

... Robustified best linear weighting under ELE - BE

... Gaussian MLE under ELE - MLE

... Robustified Gaussian MLE under ELE - RMLE

Alternative approaches

- Zhang and Tuoto (2021) propose a pseudo-OLS method for secondary linear regression analysis, where neither the matching variables nor the unlinked records are available to the analyst, and develop a diagnostic test for the assumption of non-informative linkage errors.
- Slawski and Ben-David (2019) assume the existence of mismatches for a proportion α of the observations without assuming further knowledge of the linkage process (including the value of α). They obtain an estimate of the regression coefficients by solving a penalized least squares optimization problem.

Linear regression simulations

- $\boldsymbol{\beta} = (1,3)^T$, $\sigma_e^2 = 64$
- Simulation set up
  - ... 30 blocks, $N_q = 50$
  - ... SRSWOR from linked register, $n_q = 5$
  - ... ELE-based linkage errors
  - ... $\lambda_1 = 1$ (B1-B20), $\lambda_2 = 0.9$ (B21-B26), $\lambda_3 = 0.7$ (B27-B30)
  - ... X: $N(10, 16)$ (B1-B20), $N(5, 16)$ (B21-B26), $N(2, 16)$ (B26-B30)
  - ... Outliers in the regression errors drawn from $N(50, 36)$
  - ... Scenarios:
    (a) 0% (no outliers); (b) 0% in B1-B20, 5% in B21-B30
- Tuning constants
  - ... Biweight RMLE: $a = 4.685$, $b = 7$
  - ... Huber RMLE: $a = 1.345$, $b = 3$

Intercept *X-Y* Simulations

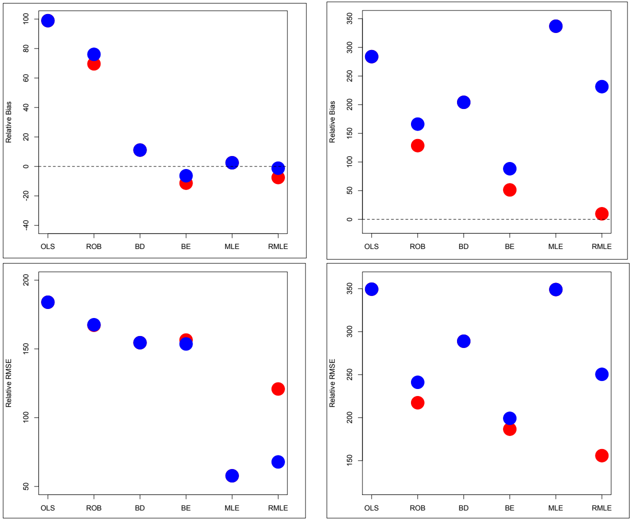
No outliers

5% outliers in B21-B30

Hu
Bi

RBias

RRMSE

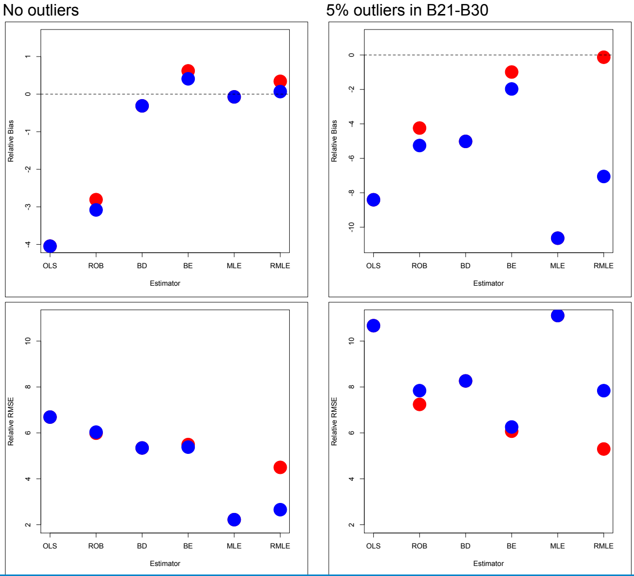


Slope *X-Y* Simulations

● Hu
● Bi

RBias

RRMSE



Remarks

- Linkage errors lead to outliers even if the population does not contain outliers.
- Traditional robust methods like ROB are good for dealing with population outliers but are not as effective when dealing with outliers generated by linkage errors.
- The linkage error bias correction approaches work very well when outliers are due to linkage errors, with MLE somewhat more efficient. However, they are not robust to real population outliers.
- Their robust versions, particularly with Biweight weighting, seem to be better able to deal with combined linkage errors and population outliers, with RMLE the superior performer.

Extensions to small area estimation (SAE)

- In SAE we are interested in estimating domain means

$$m_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$$

  or other domain descriptive quantities.
- Typically y_{ij} values are from a sample survey (with n_i possibly too small to estimate most of the m_i with adequate precision).
- The linkage error problem occurs also in SAE when linked data are used.
- We extended the LE bias corrected robust regression approaches to the SAE setting with a real data application.

Current and future research on linked data

- Relaxing the assumption of no linkage errors across blocks.
- More flexible linkage error models (allowing for incomplete linkage outcomes).
- LE-adjustment when no (or unreliable) information about the quality of linkage is released.