Analysis of integrated data for Official Statistics

Li-Chun Zhang[1,2,3]

[1] ***University of Southampton (L.Zhang@soton.ac.uk)***

[2] ***Statistisk sentralbyraa, Norway***

[3] ***Universitetet i Oslo***

June 2019, Firenze

Introduction

A long-standing topic:

Survey data for Official Statistics

- *complex* survey data
- *missing* survey data
- *latent* (survey) data
- ..., SAE, ...

Problem of selection: $S_r \overset{missing}{\subset} S \overset{complex}{\subset} U$

Problem of measurement: $\Pr\Big(y_{i,obs} \neq \overset{latent}{y_i}\Big) > 0$

Some other increasingly important topics:

$\underline{\quad ? \quad}$ *data for Official Statistics*

Some other increasingly important topics:

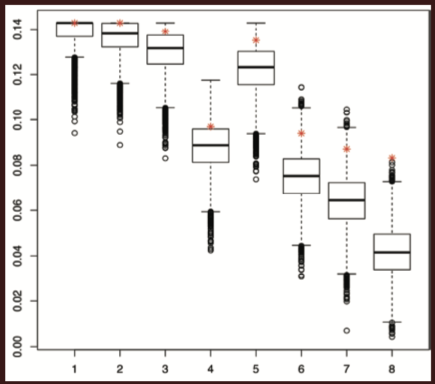
__?__ data for Official Statistics

- *Big*
- *Integrated*
- *Network*

Integration: "...data from multiple sources are combined to enable statistical inference, or to generate new statistical data for purposes that cannot be served by each source on its own..." — *Zhang & Chambers (2019, Preface)*

Chapman & Hall/CRC
Statistics in the Social and Behavioral Sciences Series

Analysis of Integrated Data



Edited by
Li-Chun Zhang
Raymond L. Chambers

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Analysis of Integrated Data
Zhang • Chambers

AID Contents

1. Introduction (Chambers)
2. On secondary analysis of datasets that cannot be linked without errors (Zhang)
3. Capture-recapture methods in the presence of linkage errors (Di Consiglio et al.)
4. An overview of uncertainty and estimation in statistical matching (Conti et al.)
5. Auxiliary variable selection in a statistical matching problem (D'Orazio et al.)
6. Minimal inference from incomplete 2 x 2-tables (Zhang & Chambers)
7. Dual and multiple system estimation with fully and partially observed covariates (Van der Heijden et al.)
8. Estimating population size in multiple record systems with uncertainty of state identification (Di Cecco)
9. Log-linear models of erroneous list data (Zhang)
10. Sampling design and analysis using geo-referenced data (Filipponi et al.)

The problem in "Analysis of Integrated Data"

A fundamental issue in Official Statistics:

Population – Entity

An immediate problem when combining multiple sources:

Entity ambiguity

- *not* possible to state with certainty that the integrated source corresponds to the target population of interest
- *lack* of an *identified* population set of target units or an *observed* subpopulation set of such units

The problem in "Analysis of Integrated Data"

Three generic settings of entity ambiguity:

- imperfect *record linkage* or *entity resolution*

  e.g. Fellegi and Sunter (1969), Christensen (2012)

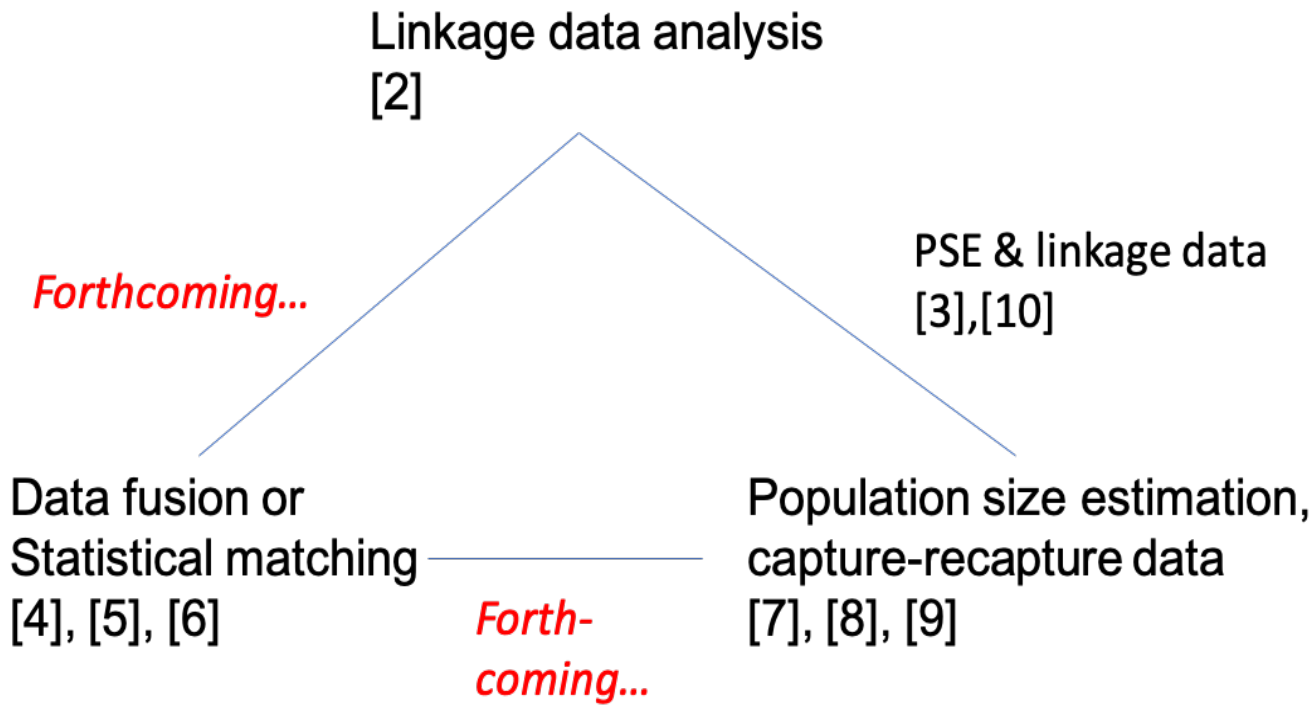
- *data fusion* or *statistical matching*: non-overlapping sources, joint information created from marginal info.

  e.g. D'Orazio et al. (2006), Wakefield (2004)

- *population size estimation* or *capture-recapture* data: population misclassified or erroneously covered in sources

  e.g. van der Heijden et al. (2012), Zhang (2015), Zhang (2011)

The problem in "Analysis of Integrated Data"



Census 2011 England and Patient Register (C-PR)

Type	Pass	# **Links**	**False Linkage Rate**
Deterministic	1	30780660	0.00011
	2	11733197	0.00389
	3	1513471	0.00561
	4	2444838	0.00375
	5	1346432	0.00748
	6	121483	0.00886
	7	1007293	0.00100
	8	825069	0.01485
	9	35432	0.00100
Probabilistic	1	511239	0.02948
	2	*298645*	*0.07165*
Total		*50617759*	

Entity ambiguity in linkage data

Datasets: $A = \{a\}$, $|A| = n_A$ and $B = \{b\}$, $|B| = n_B$

Ambiguity: the set of (unique) underlying entities

Entities: matched AB, unmatched A_u and B_u, where

$$\min(n_A, n_B) \leq |A_u| + |AB| + |B_u| \leq n_A + n_B$$

Record linkage $\mapsto$ linked set $\widetilde{AB}$ as estimated AB

Probabilistic: error in key-variables causing linkage error

Fellegi & Sunter (1969) based on the space of *pairs*:

$$A \times B = \{\underline{\text{M}}\text{atched pairs}\} \cup \{\underline{\text{U}}\text{nmatched pairs}\}$$

NB. subsets M and U not disjoint in terms of entities

FS-paradigm unsuitable for analysis of linkage data

A challenge: MLE by EM algorithm?

Modelling observed data X_A, Y_B, K_A, K_B given AB:

$$f(X_A, Y_B|AB; \psi, \eta, \theta) = \prod_{A_u} f(x_a; \psi) \cdot \prod_{B_u} f(y_b; \eta) \cdot \prod_{AB} f(x_a, y_b; \theta)$$

$$f(K_A, K_B|AB, X, Y) = \prod_{A_u} f(k_a|x_a) \cdot \prod_{B_u} f(k_b|y_b) \cdot \prod_{AB} f(k_a, k_b|x_a, y_b)$$

e.g. for key-variable error that is completely random:

$$f(K_A, K_B|AB, X, Y) = \prod_{A_u} f(k_a) \cdot \prod_{B_u} f(k_b) \cdot \prod_{AB} f(k_a, k_b)$$

DeGroot and Goel (1980): correlation ρ of bivariate normal?

- observe $x_{n\times1}$ and $y_{n\times1}$; true $y_M = \omega y$ via permutation matrix ω
- ω as unknown parameter: MLE of ρ on the boundary – bad
- ω as missing variables: integration $\mapsto$ likelihood, $n = 5$ – weird

A challenge: MLE by EM algorithm?

Q: Can use EM-algorithm in general? Seems not...

Assume complete match space $|AB| = |A| = |B|$ for simplicity:

$$y_i = x_i^\top \beta + \epsilon_i \qquad \text{for} \quad i \in AB$$

$$y_M = \omega y_B \qquad X_M = [X_A \; : \; \omega X_B]$$

$$\text{complete data } (\omega, z) \qquad \text{observed } z = (K_A, K_B, X_A, X_B, y_B)$$

$$\hat{\beta} = E(X_M^\top X_M|z)^{-1} E(X_M^\top y_M|z) \quad \text{known nontrivial } f(\omega|K_A, K_B)$$

$$= \begin{bmatrix} X_A^\top X_A^\top & X_A^\top E(\omega|z) X_B \\ X_B^\top E(\omega^\top|z) X_A & X_B^\top X_B \end{bmatrix}^{-1} \begin{bmatrix} X_A^\top E(\omega|z) \\ X_B^\top \end{bmatrix} y_B$$

$$E(\hat{\beta}|z_{(Y)}) = \begin{bmatrix} X_A^\top X_A^\top & X_A^\top E(\omega|z_{(Y)}) X_B \\ X_B^\top E(\omega^\top|z_{(Y)}) X_A & X_B^\top X_B \end{bmatrix}^{-1}$$

$$\begin{bmatrix} X_A^\top E(\omega|z_{(Y)}) E(\omega^\top|z_{(Y)}) X_A & X_A^\top E(\omega|z_{(Y)}) X_B \\ X_B^\top E(\omega^\top|z_{(Y)}) X_A & X_B^\top X_B \end{bmatrix} \beta$$

Entity ambiguity in data fusion

Datasets: y_A, $A \subset U$ and z_B, $B \subset U$; $A \cap B = \emptyset$

Unknown: the joint distribution of (y, z) *for* $i \in U$

Data fusion generates, say, a dataset for $A \cup B$:

$$\widetilde{[y\ z]} = \begin{bmatrix} y_A & z^*_A \\ y^*_B & z_B \\ (y_\emptyset) & (z_\emptyset) \end{bmatrix} \quad vs. \quad \begin{bmatrix} y_{A \setminus B} & z^*_{A \setminus B} \\ y^*_{B \setminus A} & z_{B \setminus A} \\ y_{A \cap B} & z_{A \cap B} \end{bmatrix}$$

Ambiguity: can $\widetilde{[y\ z]}$ *be a dataset from* $[y\ z]_U$ *at all?*

Is missing data otherwise beyond this ambiguity?

NB. add link. data ambiguity if $|A \cap B| > 0$ but uncertain $A \cap B$

Study of uncertainty space

Measure of *uncertainty space*: the set of joint $f_{Y,Z}(y, z)$ that is compatible with the marginal $f_Y(y)$ and $f_Z(z)$

e.g. in Statistical matching: Kadane (1978), Moriarity and Scheuren (2001), D'Orazio *et al.* (2006), Rässler and Kiesel (2009) and Conti *et al.* (2012, 2013)

Can be studied without sampling uncertainty, e.g.

$$L = \max\big(0,\ \Pr(Y = 1) + \Pr(Z = 2) - 1\big) \leq \Pr(Y = 1, Z = 2)$$

$$U = \min\big(\Pr(Y = 1),\ \Pr(Z = 2)\big) \geq \Pr(Y = 1, Z = 2)$$

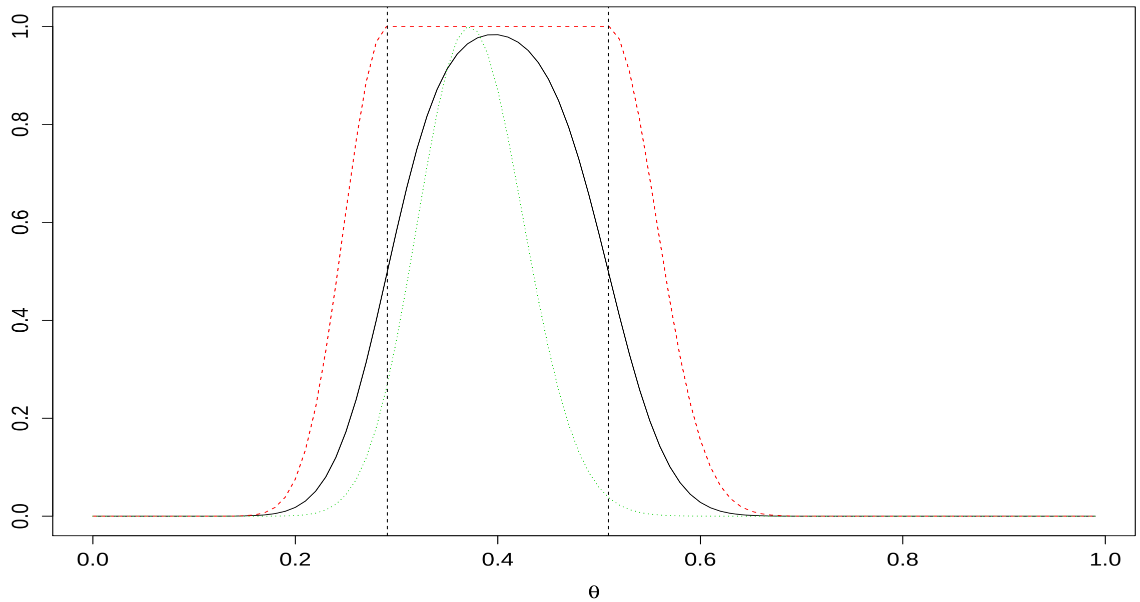
Can estimate the bound given finite sample: $(\hat{L}, \hat{U})$

Can we say something about θ^*, a given point in Θ?

Minimal inference from finite sample

Missing binary data: $(n_{11}, n_{01}, n_{+0}) = (32, 54, 24)$

Target	Observed $(R = 1)$	Missing $(R = 0)$	Total
$Y = 1$	n_{11}	–	–
$Y = 0$	n_{01}	–	–
Total	n_{+1}	n_{+0}	n



Dashed: profile likelihood; dotted: MCAR likelihood; <u>solid: observed corroboration</u>
vertical dotted lines (left, right): $(\widehat{L}, \widehat{U})$

Corroboration

We define the *corroboration function* of θ, for $\theta \in \Theta$, to be

$$c(\theta; \psi) = \Pr\big(\theta \in (\widehat{L}, \widehat{U}); \psi\big)$$

where the probability is evaluated with respect to $f(n_{11}, n_{01}, n_{+0}; \psi)$.

The *actual corroboration* at the true, identifiable ψ_0 is given by

$$c_0(\theta) = c(\theta; \psi_0)$$
$$c_0(\theta_0) = c(\theta_0; \psi_0) = \text{Confidence level of } (\widehat{L}, \widehat{U})$$

The *observed (ML) corroboration* is given (via MLE $\widehat{\psi}$) as

$$\widehat{c}(\theta) = c(\theta; \widehat{\psi}).$$

The higher the corroboration of θ, the harder it is to reject it.

Corroboration Test (CT)

Null hypothesis $H_A : \theta^* \in (L_0, U_0)$ against $H_B : \theta^* \notin (L_0, U_0)$

The Likelihood Ratio Test is inapplicable.

Test statistic: $T_n = 1$ if $\theta^* \in \text{Interior}(\widehat{\Theta}_n)$ and $T_n = 0$ if $\theta^* \notin \widehat{\Theta}_n$

CT: reject H_A if $T_n = 0$. With asymptotic power

$$\lim_n \beta_n(\theta^*) = 1 - \lim_n \Pr(T_n = 1; \psi_0) = 1 - \lim_n c_n(\theta^*; \psi_0)$$

Type-I error if H_A is true, but we reject H_A: $\Pr(\text{Type-I}) \to 0$

Type-II error if H_B is true, but we do not reject H_A: $\Pr(\text{Type-II}) \to 0$

Theo.: CT of obs. power $1 - \widehat{c}(\theta^*)$ *is strongly Chernoff-consistent.*

$\theta^* = \Pr(Y = 1)$	0.2	0.3	0.4	0.5	0.6
Observed corroboration $\widehat{c}(\theta^*)$	0.018	0.583	0.985	0.576	0.028
Profile LR$(\theta^*, 0.4)$	0.076	1	1	1	0.156

Entity ambiguity in capture-recapture data

Datasets: $A = \{a\}$, $|A| = n_A$ and $B = \{b\}$, $|B| = n_B$

Ambiguity: population U *of unknown size* N*, where*

$$\begin{cases} A \cup B \setminus U \neq \emptyset \\ U \setminus (A \cup B) \neq \emptyset \end{cases}$$

Erroneous enumeration in $A \cup B$, or population domain misclassification with $A = \cup_{d=1}^{D} A_d$ and $B = \cup_{d'=1}^{D} B_{d'}$

NB. existing log-linear models for U only (Fienberg, 1972)

NB. additional linkage data ambiguity if uncertain $A \cap B$, or fusion data ambiguity if $A \cap B = \emptyset$

Entity ambiguity in capture-recapture data

(GBA, WWB, LADIS)	Count	(GBA, WWB, LADIS)	Count
(1, 1, 1)	30	(0, 1, 1)	175
(1, 1, 0)	495	(0, 1, 0)	2792
(1, 0, 1)	24	(0, 0, 1)	654
(1, 0, 0)	999	(0, 0, 0)	$m_\emptyset$

- Persons extracted from the Dutch population register (GBA), who are registered at an institute which hosts homeless people
- Persons in a register of social benefit (WWB), who do not have a permanent place of residence
- Persons who are homeless according to the National Alcohol and Drugs Information System (LADIS).

Coumans et al. (2017): missing-by-all $\hat{m}_\emptyset = 12589$; using covariates gender, age, place, country of origin; standard log-linear modelling

Log-linear model of list-population universe

List-population universe: $A \cup U$, for $A = \cup_{k=1}^{K} A_k$

Log-linear model of erroneous enumeration in $A \cup U$:

$$\text{logit } \theta_\omega = \log \mu_{\omega 0} - \log \mu_{\omega 1} = \sum_{\nu \in \Omega(\omega)} \lambda_\nu$$

$$\omega \subseteq \{1, ..., K\} \qquad 0/1 = \text{out/in } U$$

$$\Omega(\omega) = \text{all non-empty subsets of } \omega$$

$$\theta_\omega = \Pr(i \notin U \mid i \in \bigcap_{k \in \omega} A_k, i \notin \bigcup_{k \notin \omega} A_k)$$

NB. ω for cross-classified list domain, e.g.

$$K = 4 : \ \omega = \{1, 4\} \ \Leftrightarrow \ \text{cross-classification} = (1001)$$

$K = 2,\ \omega = \{1, 2\},\ \lambda_{11} = 0$: for cross-classified domain

$$\text{logit } \theta_{(11)} = \text{logit } \theta_{(10)} + \text{logit } \theta_{(01)}$$

Log-linear models of pseudo conditional independence

Pseudo conditional independence (PCI) with $K = 2$:

$$\log\theta_{11} = \log\theta_{1+} + \log\theta_{+1} \quad \Leftrightarrow \quad \theta_{11} = \theta_{1+}\theta_{+1}$$

$$\Pr(i \notin U \mid i \in A_1, i \in A_2) = \Pr(i \notin U \mid i \in A_1)\Pr(i \notin U \mid i \in A_2)$$

In contrast, an example of conditional independence:

$$\Pr(i \in A_1, i \in A_2 \mid i \in U) = \Pr(i \in A_1 \mid i \in U)\Pr(i \in A_2 \mid i \in U)$$

Hierarchical log-linear PCI models, illustrated:

Model Restrictions	Model Interpretation
-	Saturated model
$\alpha_{123} = 0$	Null 2nd-order PCI-interaction
$\alpha_{12} = \alpha_{123} = 0$	PCI between A_1 and A_2 given A_3
$\alpha_{12} = \alpha_{13} = \alpha_{123} = 0$	PCI between A_1 and (A_2, A_3)
$\alpha_{12} = \alpha_{13} = \alpha_{23} = \alpha_{123} = 0$	Mutual PCI between A_1, A_2 and A_3

Capture-recapture data with over- and under-count

Bipartition of lists: with or without erroneous enum.

$$\begin{array}{ccc} & \{A_1, A_2, ..., A_K\} & \\ & \swarrow \qquad \searrow & \\ \{A_1, ..., A_J\} & & \{A_{J+1}, ..., A_K\} \\ \supset & & = \\ \{\mathcal{S}_1, ..., \mathcal{S}_J\} & & \{\mathcal{S}_{J+1}, ..., \mathcal{S}_K\} \quad \subset U \end{array}$$

Erroneous enum. in *marginally classified* list domain:

$$\log \theta_{\omega+} = \sum_{\nu \in \Omega(\omega)} \alpha_\nu \qquad \text{for } \omega \subseteq \{1, ..., J\}$$

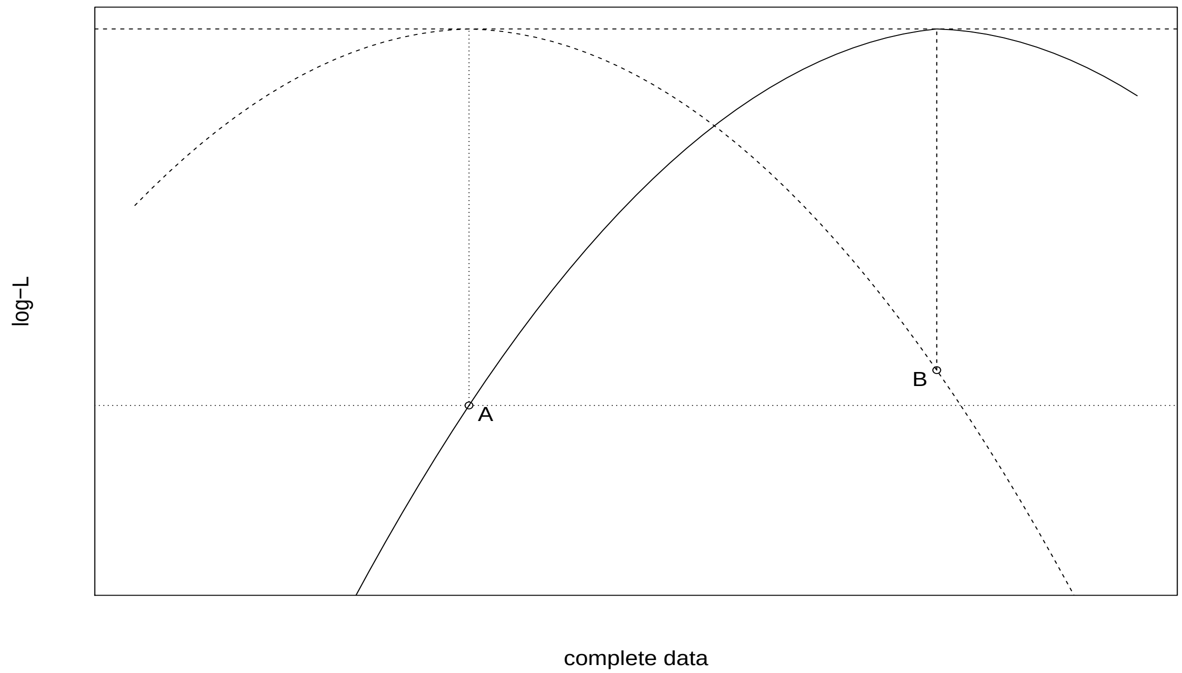
and log-linear model of $\{\mathcal{S}_1, ..., \mathcal{S}_K\} \subset U$ (Fienberg, 1972):

$$\log \mu_\omega = \sum_{\nu \in \Omega(\omega)} \lambda_\nu \qquad \text{for } \omega \subseteq \{1, ..., K\}$$

		$\mathcal{S}$ = GBA, A_1 = WWB, A_2 = LADIS					
Mod.†	Deviance	$\hat{\theta}_{1+}$	$\hat{\theta}_{2+}$	$\hat{\theta}_{12+}$	$\hat{\gamma}$	$\hat{m}_{\emptyset}$	$n_{\emptyset}$
log	0.03	0.007	0.589	0.004	0.151	5597	999
logit	0.01	0.030	0.602	0.045	0.155	5447	999
		$\mathcal{S}$ = WWB, A_1 = GBA, A_2 = LADIS					
Mod.	Deviance	$\hat{\theta}_{1+}$	$\hat{\theta}_{2+}$	$\hat{\theta}_{12+}$	$\hat{\gamma}$	$\hat{m}_{\emptyset}$	$n_{\emptyset}$
log	0.98	0.657	0.767	0.504	0.987	38	2792
logit	9.83	0.129	0.425	0.099	0.388	4409	2792
		$\mathcal{S}$ = LADIS, A_1 = GBA, A_2 = WWB					
Mod.*	Deviance	$\hat{\theta}_{1+}$	$\hat{\theta}_{2+}$	$\hat{\theta}_{12+}$	$\hat{\gamma}$	$\hat{m}_{\emptyset}$	$n_{\emptyset}$
log	0.03	0.399	0.005	0.002	0.059	*10434*	654
logit	0.03	0.401	0.009	0.006	0.059	*10383*	654

e.g. † & ∗ "equivalent" (Vuong, 1989) by LRT selection

Capture-recapture data with over- and under-count



Latent Likelihood Ratio Criterion (LLRC), illustrated: Models with saturated ℓ_C horizontally aligned (top dash); Maximum $\hat{\ell}_C$ of different models based on *pseudo-true* data under model A (curve dash) and B (curve solid); half deviances of A (vertical dot) and B (vertical dash).

Capture-recapture data with over- and under-count

LLRC selection marked †:

	Model pseudo-true	Model fitted	Latent deviance
(I)	(S = GBA, log)	(S = GBA, logit)	138.04
	(S = GBA, logit)	(S = GBA, log)†	137.59
(II)	(S = LADIS, log)	(S = LADIS, logit)	14.22
	(S = LADIS, logit)	(S = LADIS, log)†	12.73
(III)	(S = GBA, log)	(S = LADIS, log)†	10208.12
	(S = LADIS, log)	(S = GBA, log)	19247.94
(IV)	(S = GBA, logit)	(S = LADIS, log)†	10599.38
	(S = LADIS, log)	(S = GBA, logit)	19476.44
(V)	(S = GBA, log)	(S = LADIS, logit)†	10225.69
	(S = LADIS, logit)	(S = GBA, log)	19310.55
(VI)	(S = GBA, logit)	(S = LADIS, logit)†	10524.25
	(S = LADIS, logit)	(S = GBA, logit)	19500.54

[1] Christen, P. (2012). A survey of indexing techniques for scalable record linkage and deduplication. *ISEE Transactions on Knowledge and Data Engineering*, **24**.

[2] Conti, P.L., Marella, D. and Scanu, M. (2012). Uncertainty analysis in statistical matching. *Journal of Official Statistics*, vol. **28**, pp. 69 - 88.

[3] Conti, P.L., Marella, D. and Scanu, M. (2013). Uncertainty analysis for statistical matching of ordered categorical variables. *Computational Statistics & Data Analysis*, vol. **68**, pp. 311-325.

[4] Coumans, A.M., Cruyff, M., Van der Heijden, P. G. M., Wolf, J. and Schmeets, H. (2017). Estimating homelessness in the Netherlands using a capture-recapture approach. *Social Indicators Research* **130** pp. 189–212.

[5] DeGroot, M.H. and Goel, P.K. (1980). Estimation of the correlation coefficient from a broken random sample. *The Annals of Statistics*, **8**, 264–278.

[6] D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. Chichester: Wiley.

[7] Fellegi, I.P. and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, **64**, 1183-121 0.

[8] Fienberg, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika* **59** 409–439.

[9] Moriarity, C. and Scheuren, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, vol. **17**, pp. 407-422.

[10] Rässler, S. and Kiesl, H. (2009). How useful are uncertainty bounds? Some recent theory with an application to Rubin's causal model. In *Proceedings of the 57th Sessions of the International Statistical Institute*.

[11] Van der Heijden, P. G. M., Whittaker, J., Cruyff, M. J. L. F., Bakker, B., and van der Vliet, R. (2012). People born in the Middle East but residing in the Netherlands: Invariant population size estimates and the role of active and passive covariates. *Annals of Applied Statistics*, **6**, 831-852.

[12] Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57** 307–333.

[13] Wakefield, J. (2004). Ecological inference for 2×2 tables. (With discussions). *Journal of the Royal Statistical Society, Series A*, vol. **167**, pp. 385-445.

[14] Zhang, L.-C. (2015). On modelling register coverage errors. *Journal of Official Statistics*, **31**, 381-396.

[15] Zhang, L.-C. (2011). A unit-error theory for register-based household statistics. *Journal of Official Statistics*, **27**, 415-432.