Population size estimation based on erroneous capture-recapture data

Li-Chun Zhang
University of Southampton, Southampton, UK; Statistics Norway, Oslo, Norway
E-mail: L.Zhang@soton.ac.uk

Abstract

Log-linear models have long been used for population size estimation based on capture-recapture data subjected to under enumeration. We extend the scope to allow for additional erroneous enumeration, by introducing a concept of pseudo conditional independence and developing new classes of log-linear and -odds models accordingly. For selection among models that achieve the same fitted log-likelihood, which can occur for incomplete categorical data, we explore the use of latent discrepancy measures. Potential applications include replacing traditional costly Census with alternative register-based approaches, sizing of dynamic clandestine population using relevant but erroneous lists, evaluation of diagnostic efficiency of multiple tests, *etc.*

Key Words: Coverage of register, pseudo conditional independence, log-linear and -odds models, maximum likelihood estimation, latent log-likelihood discrepancy.

1 Introduction

Our aim is to estimate the unknown size N of a target population $U = \{1, 2, ..., N\}$ based on the following set-up of list-survey capture-recapture (CR) data.

1.1 Erroneous list and coverage of register

Suppose we have a *list* of units, denoted by L. Let x be the known number of units in L. For any unit $i \in L$, let $I_{iU} = 1$ if $i \in U$ and $I_{iU} = 0$ otherwise. We refer to the probability $\xi = P(I_{iU} = 1 | i \in L)$ as the *hit* rate, and to the probability $\theta = P(I_{iU} = 0 | i \in L) = 1 - \xi$ as the *error* rate. Let y be the number of *hits*, i.e. units in L that belong to U. Let r be the number of *errors*, i.e. units in L that do not belong to U. We have $E(y|x) = x\xi = x(1-\theta) = x - E(r|x)$. We refer to the probability that a unit in U belongs to L as the *catch* rate of the list, denoted by γ. If L derives from a register, we may refer to $N - y$ as its *under-coverage* and r its *over-coverage*. The three parameters (ξ, θ, γ) are structurally constrained with each other:

$$E(y|N) = N\gamma = x\xi = x(1-\theta) = E(y|x)$$

We stipulate an independent coverage survey, denoted by S. Let n be the number of population units that are captured in S. Due to non-response or under-enumeration, we have $n < N$ and $\gamma_s = E(n|N)/N$ as the catch rate of S. In principle, S needs not to be an enumeration survey that aims to catch every unit of the population, and sub-sampling of $S \subset U$ can potentially provide a viable approach, whereby we have $E(n|N) = \pi_s \gamma_s$ where π_s denotes the sampling probability. But for simplicity we shall assume $\pi_s = 1$ in this paper. Let m be the number of population units missed out by S, i.e. $N = n + m$. Let (n_L, m_L) be the number of units in L that are captured or missing by S. The list-survey CR data are related to each other as follows:

$$(N) = n + (m),\ x = (y) + (r),\ n_L + (m_L) = x - (r),\ n_{S(L)} = n - n_L \text{ and } (m_{S(L)}) = (m) - (m_L)$$

where a number in parenthesis is unobserved, and one without is assumed to be observed.

To complete the set-up of erroneous CR data, we assume multiple lists $L_1, L_2, ...L_K$ together with a single survey S. Each list may contain hits and errors as above. All importantly, we assume that it is possible to identity whether a given unit belongs to any of $L_1, L_2, ...L_K$ and S.

1.2 Potential applications

The set-up extends the traditional setting of Census (L) in combination with independent coverage survey (S), or the alternative scenario where the Census is replaced by a Central Population Register. The existing dual-frame estimation approach may no longer be acceptable provided non-negligible list errors. An obvious option in such cases is to apply follow-up (or dependent) sampling of L in addition to independent coverage survey - see Elkin *et al* (2012) for a review of related census practices. However, the multiple-list independent-survey set-up can be motivated in a situation where multiple registers may possibly be used in place of the Census, but list-dependent sampling is prohibited by regulations of individual data protection.

Applications to CR data outside of the Census context can be conceived. For instance, the target population may be both clandestine and dynamic, such as the active drug-users. A relevant list may be the records of previously treated drug-users at a clinic, which can be erroneous if the patient is no longer a drug-user. Similarly, enumeration errors can occur in other relevant lists.

Finally, useful applications may be possible with some adaption of the set-up. For instance, each list may consist of the units with a positive test result, and several tests can be conducted. Only the test-positive units are subjected to a comprehensive diagnosis, which reveals the errors and suggests potential misses of each test including the final diagnosis. A model for predicting the hit (or error) rate given the test results and the overall catch rate may then be of interest.

1.3 Multi-list CR data as incomplete contingency table

Let $U_+ = U \cup_{k=1}^K L_k$ be the list-survey universe. Let $R = U_+ \setminus U$. We illustrate with $K = 2$. Put

$$N_{jkl}^{12R} = \sum_{i \in U_+} (I_{i1} = j)(I_{i2} = k)(I_{iR} = l) \quad \text{for} \quad j = 0, 1 \quad k = 0, 1 \quad l = 0, 1$$

which is a 2-list incomplete contingency table with structural-zero cells $(N_{000}^{12R}, N_{001}^{12R})$ of units neither in L_1 nor L_2. Put $\rho_{jk} = \log(\theta_{jk}) - \log(1 - \theta_{jk}) = \log \mu_{ij1}^{12R} - \log \mu_{jk0}^{12R}$ for $\mu_{jkl}^{12R} = E(N_{jkl}^{12R})$. Table 1 provides an overview of the hierarchical models, with suitable parameter λ's and α's on the log-scale, where the structural incompleteness leads to $\lambda_1^R = 0$. The log-linear and -odds models on the same row are equivalent to each other given the binary indicator variables.

Table 1: Hierarchical log-linear and -odds models for 2-list incomplete contingency table

Log-linear Restriction	Log-odds model	Model assumption
-	$\rho_{jk} = j\alpha_1 + k\alpha_2 + jk\alpha_{12}$	Saturated model
$\lambda_{jkl}^{12R} = 0$	$\rho_{jk} = j\alpha_1 + k\alpha_2$	Null log-linear interaction; main-effects log odds
$\lambda_{kl}^{2R} = \lambda_{jkl}^{12R} = 0$	$\rho_{jk} = \alpha_j$	I_{i2} independent of I_{iR} given $I_{i1} = j$; $\theta_{jk} = \text{logit}^{-1}(\alpha_j)$
$\lambda_{jl}^{1R} = \lambda_{jkl}^{12R} = 0$	$\rho_{jk} = \alpha_k$	I_{i1} independent of I_{iR} given $I_{i2} = k$; $\theta_{jk} = \text{logit}^{-1}(\alpha_k)$
$\lambda_{jl}^{1R} = \lambda_{kl}^{2R} = \lambda_{jkl}^{12R} = 0$	$\rho_{jk} = \lambda_1^R = 0$	(I_{i1}, I_{i2}) independent of I_{iR}; $\theta_{jk} = 0.5$

It is clear that the model of independence between (I_{i1}, I_{i2}) and I_{iR} is hardly applicable. Next, $\theta_{jk} = \theta_j$ amounts to $\theta_{10} = \theta_{11} \neq \theta_{01}$, i.e. the error rate is the same for any unit in L_1 regardless whether it belongs to L_2 or not, which would only hold in special cases. Similarly for $\theta_{jk} = \theta_k$. Finally, the assumption of null 2nd-order interaction $\lambda_{jkl}^{12R} = 0$ implies $\theta_{11} > \theta_{10}\theta_{01}$. Now, provided a majority of the units are in both lists, it is conceivable that θ_{01} and θ_{10} can be much higher than the marginal error rate of either list, say, θ_1 and θ_2, while θ_{11} is much lower than either θ_1 or θ_2. Indeed, for large $(\theta_{10}, \theta_{01})$ but θ_{11} close to 0, the assumption $\theta_{11} = \theta_1\theta_2$ can fit the data much better that a model that entails $\theta_{11} > \theta_{10}\theta_{01}$. But $\theta_{11} = \theta_1\theta_2$ can not be expressed as a log-linear (or log-odds) model for 2-list CR data framed as incomplete contingency tables.

1.4 Pseudo conditional independence

The assumption $\theta_{12} = \theta_1\theta_2$ provides in fact a modeling alternative to the list-dependent sampling approach denied to us by stipulation (Section 1.2). To be explicit, this amounts to assume

$$P(I_{iU} = 0|I_{i1} = I_{i2} = 1) = P(I_{iU} = 0|I_{i1} = 1)P(I_{iU} = 0|I_{i2} = 1) \tag{1}$$

If we refer to $P(I_{iU} = 0|I_{i1} = I_{i2} = 1)$ as the *joint* conditional probability of $I_{iU} = 0$ because it is its probability conditional on the joint event $(I_{i1} = 1) \cap (I_{i2} = 1)$, and to $P(I_{iU} = 0|I_{i1} = 1)$ or $P(I_{iU} = 0|I_{i2} = 1)$ as the *marginal* conditional probability given the respective marginal event $I_{i1} = 1$ or $I_{i2} = 1$, then assumption (1) states that *the joint conditional probability is the product of the marginal conditional probabilities.* In contrast, conditional independence states that *the conditional joint probability is the product of the conditional marginal probabilities*, such as $P(A \cap B|C) = P(A|C)P(B|C)$. Hence, we shall refer to (1) as an assumption of *pseudo conditional independence (PCI)* for list errors between L_1 and L_2. Notice that, similarly but not equivalently, the PCI assumption for list hits between L_1 and L_2 can be given as

$$P(I_{iU} = 1|I_{i1} = I_{i2} = 1) = P(I_{iU} = 1|I_{i1} = 1)P(I_{iU} = 1|I_{i2} = 1) \quad \Leftrightarrow \quad \xi_{12} = \xi_1\xi_2$$

The sequels contain the following: in Section 2 log-linear and -odds models are developed on the concept of pseudo conditional independence, where estimation is only briefly explained due to limited space; in Section 3 model selection utilizing latent discrepancy measures is explored.

2 Model and estimation

2.1 Multi-list CR table

First we frame the data as *complete capture-recapture (CR) table.* Given K lists, denote by Ω_K the set of all *non-empty* subsets of $\{1, 2, ..., K\}$. For example, $\Omega_2 = \{\{1, 2\}, \{1\}, \{2\}\}$ and $\Omega_3 = \{\{1, 2, 3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1\}, \{2\}, \{3\}\}$. For $\omega \in \Omega_K$ and $\omega^c = \{1, ..., K\} \setminus \omega$, put

$$x_{\omega(\omega^c)} = \sum_{i \in U_L} \left(\prod_{k\in\omega} I_{ik}\right)\left(\prod_{k\in\omega^c}(1 - I_{ik})\right) \qquad \text{and} \qquad x_\omega = x_{\omega(\emptyset)}$$

where $U_L = \cup_{k=1}^K L_k$ is the *list universe.* The K-list CR table is given by $\mathbf{x} = (x_\omega)^T_{\omega\in\Omega_K}$. Whereas the index $\omega(\omega^c)$ corresponds directly to the contingency table set-up with the structural-zero cell $x_{(1,2,...,K)} \equiv 0$. Table 2 illustrates the notational correspondence for $K = 2$ and 3.

2.2 Log-linear and -odds models for multi-list CR table

Denote by θ_Ω the error rate among the x_ω units for $\omega \in \Omega_K$, and by ξ_ω the hit rate. We define the saturated log-linear model of the error rates of the K-list CR table as follows:

$$E(r_\omega|x_\omega) = x_\omega\theta_\omega \qquad \text{and} \qquad \log\theta_\omega = \eta_\omega = \sum_{a\in\Omega(\omega)} \alpha_a \tag{2}$$

Table 2: Index of 2-list and 3-list contingency and CR tables. Structural zero cell (*).

Contingency table		CR table	
(I_{i1}, I_{i2})	$\omega(\omega^c)$	(I_{i1}, I_{i2})	ω
(1, 1)	12	(1, 1)	12
(1, 0)	1(2)	(1, -)	1
(0, 1)	2(1)	(-, 1)	2
(0, 0)*	(12)*		

Contingency table		CR table	
(I_{i1}, I_{i2}, I_{i3})	$\omega(\omega^c)$	(I_{i1}, I_{i2}, I_{i3})	ω
(1, 1, 1)	123	(1, 1, 1)	123
(1, 1, 0)	12(3)	(1, 1, -)	12
(1, 0, 1)	13(2)	(1, -, 1)	13
(0, 1, 1)	23(1)	(-, 1, 1)	23
(1, 0, 0)	1(23)	(1, -, -)	1
(0, 1, 0)	2(13)	(-, 1, -)	2
(0, 0, 1)	3(12)	(-, -, 1)	3
(0, 0, 0)*	(123)*		

where $\Omega(\omega)$ is the *set* of non-empty subsets of ω. For the saturated log-linear model of the hit rates, we only need to replace r_ω with y_ω and θ_ω with ξ_ω in (2). Whereas, for the corresponding (but not equivalent) log-odds model we simply use the logit link function instead, i.e.

$$E(r_\omega|x_\omega) = x_\omega\theta_\omega \qquad \text{and} \qquad \text{logit } \theta_\omega = \eta_\omega = \sum_{a\in\Omega(\omega)} \alpha_a \tag{3}$$

For $K = 2$, the saturated log-linear model (2) amounts to $\log\theta_1 = \alpha_1$, $\log\theta_2 = \alpha_2$ and $\log\theta_{12} = \alpha_1 + \alpha_2 + \alpha_{12}$. There are 3 parameters for the 3 free observations (r_1, r_2, r_{12}) conditional on (x_1, x_2, x_{12}). A more parsimonious main-effects model is given by $\alpha_{12} = 0$, such that

$$\log\theta_1 = \alpha_1 \quad \text{and} \quad \log\theta_2 = \alpha_2, \quad \text{and} \quad \log\theta_{12} = \alpha_1 + \alpha_2 \quad \Leftrightarrow \quad \theta_{12} = \theta_1\theta_2$$

where θ_1 is the marginal error rate of L_1, and θ_2 that of L_2, i.e. the assumption of PCI (1).

Table 3: Generic log-linear models for error- or hit-rates of 3-list CR table

Model Restriction	Model Interpretation
-	Saturated model
$\alpha_{123} = 0$	Null 2nd-order PCI-interaction among (L_1, L_2, L_3)
$\alpha_{12} = 0$	PCI between L_1 and L_2
$\alpha_{123} = -\alpha_{12}$	Conditional PCI between L_1 and L_2 given L_3
$\alpha_{12} = \alpha_{123} = 0$	Both PCI and conditional PCI between L_1 and L_2
$\alpha_{12} = \alpha_{13} = \alpha_{123} = 0$	PCI between L_1 and (L_2, L_3)
$\alpha_{12} = \alpha_{13} = \alpha_{23} = \alpha_{123} = 0$	Mutual PCI between L_1, L_2 and L_3

Table 3 provides a summary of the hierarchical log-linear models for 3-list CR tables in analogy to the hierarchical log-linear models for 3-way contingency tables. For instance, the model of PCI between L_1 and L_2 is given by $\alpha_{12} = 0$ and $\theta_{12} = \theta_1\theta_2$ as before. Whereas we refer to $\alpha_{12} = -\alpha_{123}$ as the model of *conditional PCI* between L_1 and L_2 given L_3, since we can write

$$\begin{aligned}\eta_{123} &= \log\theta_3 + \log\theta_{12|3} = \log\theta_3 + \log\theta_{1|3} + \log\theta_{2|3} = \log\theta_3 + \log(\theta_{13}/\theta_3) + \log(\theta_{23}/\theta_3) \\ &= \eta_3 + (\eta_{13} - \eta_3) + (\eta_{23} - \eta_3) = \alpha_3 + (\alpha_1 + \alpha_{13}) + (\alpha_2 + \alpha_{23})\end{aligned}$$

The model of both PCI and conditional PCI between L_1 and L_2 is then given by $\alpha_{12} = \alpha_{123} = 0$. Notice that the quantities $\theta_{12|3}$, $\theta_{1|3}$ and $\theta_{2|3}$ are *not* conditional probabilities. We call them the *pseudo conditional probabilities* that are defined by the conditional probability calculus, i.e.

$$\theta_{\omega|a} \overset{\text{def}}{=} \theta_{\{\omega,a\}}/\theta_a \qquad \text{for} \quad \omega \neq a \in \Omega_k$$

This enables us to express the relationships between the proper conditional probabilities θ_ω's by means of conditional probability calculus. Finally, we notice that no PCI of any kind is induced by $\alpha_{123} = 0$. By analogy to the log-linear model for contingency tables, we shall refer to this as the assumption of null 2nd-order *PCI-interaction* among (L_1, L_2, L_3).

Obviously the log-odds models (3) do not have the PCI interpretations as the log-linear models. Instead, they can simply be described in terms of the *list interactions* on the log scale. For instance, $\alpha_{123} = 0$ amounts to assuming null 2nd-order list interaction among (L_1, L_2, L_3), whereas $\alpha_{12} = \alpha_{13} = \alpha_{23} = \alpha_{123} = 0$ yields the main-effects log-odds model. The addition of log-odds models increase the flexibility in applications. For instance, suppose the data deviate from the PCI assumption in the direction of $\theta_{12} > \theta_1\theta_2$, then the log-odds model of $\alpha_{12} = 0$ may achieve good fit instead, since $\theta_{12} > \theta_1\theta_2$ is an inherent property of the logit link in this case.

2.3 Maximum likelihood estimation (MLE)

The complete-data model for list-survey CR data consists of two independent parts: (I) the list errors (or hits) under a log-linear or -odds model described above, (II) the survey catches as a sequence of Bernoulli trails with catch rate γ_s. In particular, we adopt the negative binomial model $(n_{S(L)}, \gamma_s)$ for $m_{(SL)} = N - y_L - n_{S(L)}$ conditional on $(n_{S(L)}, y_L)$, where $m_{(SL)}$ is the number of units missed by all the $K+1$ data sets. Notice that N is then treated as a random variable instead of an unknown constant, and $n_{S(L)}$ as if it were fixed in advance. This entails the loss of one degree of freedom in the data, while equally reducing the number of parameters by one. Due to the limited space, details of the development and subsequent MLE are omitted here.

3 Model selection

For illustration consider the data in Table 4. The complete-data list-hits and errors are given in the right-side half of the table, where $(x_1, x_2, x_{12}) = (1050, 1200, 900)$. The hits and errors among the units in only of the lists are held fixed, but the joint list error r_{12} is allowed to vary and, in accordance, the empirical joint list error rate r_{12}/x_{12}, marginal error rate r_1/x_1 of L_1 and r_2/x_2 of L_2. The list-survey CR data are given to the left of these, where we vary the empirical list and survey catch rate $\gamma_L = y_L/N$ and $\gamma_s = n_S/N$, respectively. Data under the various settings of Table 5 are generated according to the scheme depicted in Table 4 as follows: $(x_{1(2)}, x_{2(1)}, x_{12}, r_{1(2)}, r_{2(1)}) \xrightarrow{r_{12}} (y_{12}, y_L) \xrightarrow{\gamma_L} N \xrightarrow{\gamma_s} (n_{1(2)}, n_{2(1)}, n_{12}, n_{(12)})$.

Table 4: Complete-data list-hit and -error and list-survey CR data. Data allowed to vary for illustration: joint list error (r_{12}), empirical survey catch rate (γ_s), empirical list catch rate (γ_L).

	List-survey CR data		Complete-data list CR data		
Index $\omega(\omega^c)$	Observed $n_{\omega(\omega^c)}$	Unobserved $m_{\omega(\omega^c)}$	Hit $y_{\omega(\omega^c)}$	Error $r_{\omega(\omega^c)}$	List $x_{\omega(\omega^c)}$
1(2)	$63\gamma_s$	$63(1-\gamma_s)$	63	87	150
2(1)	$78\gamma_s$	$78(1-\gamma_s)$	78	222	300
12	$(900 - r_{12})\gamma_s$	$(900 - r_{12})(1-\gamma_s)$	$900 - r_{12}$	r_{12}	900
(12)	$(y_L/\gamma_L - y_L)\gamma_s$	$(y_L/\gamma_L - y_L)(1-\gamma_s)$	$y_L \ (= y_{1(2)} + y_{2(1)} + y_{12})$		

Table 5: Estimates $(\hat{N}, \hat{r}_{12})$ for $\hat{r}_{12} = E(r_{12}|n_{12}; \hat{\boldsymbol{\psi}})$ assuming $\alpha_{12} = 0$. LLR-selection marked †

	(N, r_{12}, n)			
$(\gamma_L, \gamma_s) = (0.95, 0.95)$	$(1087, 8, 1032)$	$(1077, 18, 1023)$	$(1066, 28, 1012)$	$(1056, 38, 1003)$
Log-linear $(\hat{N}, \hat{r}_{12})$	$(1074, 18.2)^\dagger$	$(1077, 18.0)^\dagger$	$(1078, 17.7)$	$(1081, 17.5)$
Log-odds $(\hat{N}, \hat{r}_{12})$	$(1061, 29.4)$	$(1063, 28.9)$	$(1065, 28.3)^\dagger$	$(1068, 27.8)^\dagger$
	(N, r_{12}, n)			
$(\gamma_L, \gamma_s) = (0.80, 0.95)$	$(1291, 8, 1226)$	$(1279, 18, 1215)$	$(1266, 28, 1202)$	$(1254, 38, 1191)$
Log-linear $(\hat{N}, \hat{r}_{12})$	$(1276, 18.2)^\dagger$	$(1279, 18.0)^\dagger$	$(1281, 17.7)$	$(1283, 17.5)$
Log-odds $(\hat{N}, \hat{r}_{12})$	$(1260, 29.4)$	$(1263, 28.9)$	$(1265, 28.3)^\dagger$	$(1268, 27.8)^\dagger$
	(N, r_{12}, n)			
$(\gamma_L, \gamma_s) = (0.90, 0.98)$	$(1148, 8, 1125)$	$(1137, 18, 1114)$	$(1126, 28, 1104)$	$(1114, 38, 1092)$
Log-linear $(\hat{N}, \hat{r}_{12})$	$(1135, 18.2)^\dagger$	$(1137, 18.0)^\dagger$	$(1139, 17.7)$	$(1141, 17.5)$
Log-odds $(\hat{N}, \hat{r}_{12})$	$(1125, 26.0)^*$	$(1123, 28.9)$	$(1125, 28.4)^\dagger$	$(1127, 27.8)^\dagger$

Note: Fitted log-likelihood equal to maximum achievable log-likelihood except marked *

Both the log models (2) and (3) with $\alpha_{12} = 0$ are fitted to obtain $(\hat{N}, \hat{r}_{12})$, where $\hat{r}_{12} = E(r_{12}|n_{12}; \hat{\boldsymbol{\psi}})$ and $\hat{\boldsymbol{\psi}}$ denotes the MLE of the corresponding model parameters. We notice that the number of parameters equals to the degrees of freedom in data, and the fitted log-likelihood achieves the maximum attainable value everywhere in Table 5 except for the case marked by $*$. However, the resulting $\hat{N}$ is not the same. Indeed, had the complete list-data been observable, one would have been able to distinguish the two by the log-likelihood ratio (LR) test.

Now, the "true" data that would have exactly satisfied the model assumptions exist in the present setting, which is $r^0_{12} = 18$ for the log-linear model and $r^0_{12} = 28.4$ for the log-odds model. These can be found by simulation with hypothetical complete list-data. Generally, when the assumed model is not true, Vuong (1989) shows that the MLE based on IID observations converges to the so-called pseudo-true value, i.e. the parameter value of the assumed model that minimizes the Kullback-Leibler information criterion. For models of the exponential family of distributions with missing data, we show that $\hat{r}_{12}$ converges to the *pseudo-true data* r^*_{12}, i.e. the value of r_{12} that would have yielded the pseudo-true parameter value as the complete-data MLE.

The estimates in Table 5 are consistent with these theoretical result. We obtain $\hat{r}_{12} = 18$ under the log-linear model when the data are generated with $r_{12} = r^0_{12} = 18$, and $\hat{r}_{12} \doteq 28.4$ under the log-odds model when the data are generated close to r^0_{12} at $r_{12} = 28$. Otherwise, neither of the models is true, so that $\hat{r}_{12}$ is an estimate of the pseudo-true data rather than the "true" data, such as $\hat{r}^*_{12} = 18.2$ for the log-linear model when the "true" data is $r_{12} = 8$, *etc.*

The estimates marked by † in Table 5 are identified by a *latent LR* criterion based on these results, which prefers the log-linear model (A) to the log-odds model (B) if $\ell_A(\hat{\boldsymbol{\psi}}; r^0_{12}) - \ell_A(\hat{\boldsymbol{\psi}}; \hat{r}^*_{12}) < \ell_B(\hat{\boldsymbol{\psi}}; r^0_{12}) - \ell_B(\hat{\boldsymbol{\psi}}; \hat{r}^*_{12})$, where ℓ_A is the log-likelihood of the log-linear model based on the complete list-data and ℓ_B that of the log-odds model. The log-odds model is selected in the opposite case.

Finally, the survey catch n as a naive estimate of N marks the bottom-line of performance for alternative estimators. Provided the best acceptable model is selected, $\hat{N}$ is better than n.

References

[1] Elkin, M., Dent, P. and Rahman, N. (2012). *A Review of International Approaches to Estimating and Adjusting for Under- and Over-coverage*. ONS Internal Report.

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