DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 381–396, http://dx.doi.org/10.1515/JOS-2015-0023

On Modelling Register Coverage Errors

Li-Chun Zhang[1]

Register data that originate from administrative or other secondary sources are increasingly being used to generate statistical outputs directly. The coverage of the input datasets is an important issue in this respect. Traditionally capture-recapture models have been used to deal with multiple list enumerations subjected to undercoverage errors. The aim of this article is to scope possible approaches to modelling capture-recapture data with *additional* overcoverage error. Attention is primarily given to model interpretations and conditions under which a model may provide a plausible basis for estimation and uncertainty evaluation. The setting with two list enumerations is examined in depth as it is the most common in practice. Models that can be extended to include more than two lists are identified. An additional independent coverage survey with *only* undercoverage error is always needed for estimation. Potential application to census coverage-error adjustment is discussed.

Key words: List error and catch; log-linear model; pseudoconditional independence.

1. Introduction

More and more often, register data that originate from administrative or other secondary sources are being used to generate statistical outputs directly, instead of merely supplying auxiliary information for sample surveys and census. The recent round of census provides examples of this development in a number of European countries. The coverage of the input registers has a direct bearing on the population size statistics and, in the next instance, statistics about the various characteristics of interest (Zhang 2012).

A register has undercoverage of the target population if there exist population units that are not listed in the register; it has overcoverage if not all the units in the register belong to the target population. Capture-recapture (CR) models for population size estimation (e.g., Fienberg 1972; Cormack 1989; IWGDMF 1995a and 1995b) can be used to deal with the *undercoverage* errors that exist in multiple registers. A notable application is census underenumeration adjustment using an independent U-sample coverage survey to generate recapture data. See for example Wolter (1986), Hogan (1993), Brown et al. (2011), Renaud (2007), and Nirel and Glickman (2009). Note that the term *list* (e.g., Wolter 1986) is more natural than *register* in this context, as well as in a number of situations outside official statistics, such as sizing of wildlife, hard-to-reach or clandestine populations. The two terms list and register will be used interchangeably in this article.

[1]Department of Social Statistics and Demography/S3RI, University of Southampton, Highfield, Southampton SO17 1BJ, UK Email: L.Zhang@soton.ac.uk. and Statistics Norway, Akersveien 26, PB 8131 Dep, Oslo 2213, Norway Email: lcz@ssb.no.

Acknowledgment: I thank James Brown, Peter van der Heijden, two referees and the Associate Editors for stimulating discussion, healthy criticism as well as encouragement.

When it comes to overcoverage, the standard census adjustment approach is to deploy a separate *O-sample*, selected from the census reports, to directly estimate the overcoverage rate. No explicit statistical model is applied to the O-sample, in contrast to the U-sample. Moreover, fieldwork for the O-sample can be limited or totally absent – see for example Renaud (2007) for an account of the Swiss census. On the one hand, this helps to bring down the cost; on the other hand, spurious coverage errors such as duplicate reports and misreports of census residence area can to a large extent be assessed based on record matching and clerical checks without any fieldwork. However, the ability to detect *erroneous* enumeration, that is, reports of nonexistent or out-of-scope cases, may be reduced as a result.

A modelling approach to include both under- and overcoverage errors can thus have direct relevance to the census methodology. It may potentially provide a means to assess as well as to adjust for erroneous census enumerations, provided additional register enumerations from secondary sources. For example, the Office for National Statistics in the UK is currently investigating the use of administrative data for the future provision of population statistics (ONS 2013). The same goes for those countries where the traditional census enumeration has already been replaced by population registers (e.g., Israel, Switzerland), but the O-sample deploys only limited fieldwork or no fieldwork at all.

Moreover, applications to CR data in a range of situations can be conceived. For instance, the target population may be clandestine and dynamic, such as active drug users. Relevant lists may be available from the police, clinics, and various nongovernmental organisations. Erroneous enumeration can occur in all these lists. Or, consider multiple screening procedures, each generating a list of the units with a positive test result. Only the test-positive units are subjected to a comprehensive examination, which may reveal both erroneous enumerations and underenumerations in each list. A model for predicting the errors of each test as well as the combined test results may then be of interest.

In the sequels we investigate some possible approaches to modelling two-list CR data in the presence of both over- and undercoverage errors. Section 2 briefly sets out the CR model underlying the dual-system estimator (DSE) in use for census undercoverage, as expounded in Wolter (1986). The modelling approach is extended to include the overcoverage error in Section 3. All possible standard log-linear modelling alternatives for crossclassified counts are examined, as well as an approach based on the concept of pseudoconditional independence. The emphasis is on the modelling strategy, the interpretation and the conditions under which a model may provide a plausible basis for statistical estimation and uncertainty evaluation. Models that can readily be generalised to include more than two lists are identified. In Section 4 the different models are compared to each other, using artificial CR datasets that seem relevant for the setting of census population size estimation with additional administrative register data. Discussions will be given in Section 5 regarding the future work that is needed to establish a viable estimation methodology for the census or census-like population statistics.

2. Homogeneity Model for Dual-System Estimation

Wolter (1986) discussed several CR models for census undercoverage errors. The *homogeneity* model described below underpins the DSE currently in use in a number of

countries. References to the assumptions as stated by Wolter are cited and given in parentheses.

Let target population U be of unknown size N. Let A and B be two lists, both of which aim to enumerate U. Let the probability that a unit in U belongs to a particular *list domain* be given as below:

		List B: in	List B: out	
List A	in	p_{11}	p_{10}	p_{1+}
	out	p_{01}	p_{00}	p_{0+}
		p_{+1}	p_{+0}	1

Each unit is assumed to follow independently ("Autonomous Independence") the multinomial distribution ("Multinomial") with probability p_{ab} for being included in the list domain (a, b), for $a, b = 1, 0, +$. Note that U_{00} refers to the units that are neither enumerated in A nor B. Let the list-domain size N_{ab} be observed except for N_{00} and $N = N_{++}$, that is, the matching of list A and B is error free ("Matching"). All the units in list A and B can be identified ("Nonresponse"). Neither list A nor B contain overcoverage errors ("Spurious Events"). Finally, under the assumption that the event of being enumerated in list A is independent of that in B ("Causal Independence"), the probability p_{ab} is given by

$$p_{ab} = p_{a+}p_{+b} \tag{1}$$

For application to census undercoverage adjustment, let A be the census data and B the independent coverage-survey data. To avoid additional details, we assume that the coverage survey aims to enumerate the *whole* population at the sampled locations, such as census blocks or postcode areas, so that the missing survey enumerations are not due to sample selection, and the estimation below may be repeated for the target population at *each* sampled location. Because there is a time lag between the two list enumerations in practice, one needs to assume that the target population remains the same ("Closure").

A large-sample estimator of N and (p_{1+}, p_{+1}) in (1) is given by

$$(\hat{N},\, \hat{p}_{1+},\, \hat{p}_{+1}) = \left(\frac{N_{1+}N_{+1}}{N_{11}},\, \frac{N_{11}}{N_{+1}},\, \frac{N_{11}}{N_{1+}}\right)$$

(e.g., Wolter 1986). In particular, $\hat{N}$ is the so-called Dual-System Estimator (DSE). Among others this may be motivated as the method-of-moments estimator (MME) based on the set of moment equations:

$$\begin{cases} E(N_{11}) = Np_{1+}p_{+1} \\ E(N_{1+}) = Np_{1+} \\ E(N_{+1}) = Np_{+1} \\ E(N_{00}) = N - E(N_{1+}) - E(N_{+1}) + E(N_{11}) \end{cases} \tag{2}$$

Note that the last equation is merely a tautology since N_{00} is nonobservable, such that there are in effect only three equations.

3. Model with Additional Overcoverage Errors

3.1. Target-List Universe

Erroneous enumerations in census correspond to reports of nonexistent or out-of-scope cases, such as newborns after the census reference period that are mistakenly recorded in the census. Out-of-scope newborns can equally occur in lists originating from administrative sources, such as when the entry time point of a record is misreported. More often, though, erroneous register enumerations happen because an individual leaves the target population without deregistering. For instance, someone may have moved abroad without notifying their general practitioner and thus becomes an erroneous enumeration in the Patient Register for the census. Likewise, the same individual may fail to notify the election office, and become an erroneous enumeration in the Electoral Register, say, until the next time this person takes part in the general election from abroad.

Generally speaking, therefore, it is unlikely to be the case that overcoverage errors are independent across multiple registers. Moreover, erroneous enumerations may be more extensive in the administrative registers than in the census. For example, the Patient Register enumeration of the population of England and Wales is over four percent higher than the Census 2011 population estimate (ONS 2013). In other words, if unaccounted for, erroneous register enumeration is potentially a source of large bias.

The homogeneity model above is defined for the units in the target population alone. Erroneous list enumeration implies that there are units included in list A or B, or both, which are *not* in the target population U. One needs to extend the reference set to the *target-list universe*, denoted by $U^* = U \cup A \cup B$. Let the probability that a unit in U^* belongs to a particular *target-list domain* be given as below:

			List B: in	List B: out	
In U	List A	in	p_{111}	p_{110}	p_{11+}
		out	p_{101}	p_{100}	p_{10+}
			p_{1+1}	p_{1+0}	p_{1++}
			List B: in	List B: out	
Out of U	List A	in	p_{011}	p_{010}	p_{01+}
		out	p_{001}	—	p_{001}
			p_{0+1}	p_{010}	

Each unit in U^* is assumed to follow independently ("Autonomous Independence") the multinomial distribution ("Multinomial") with probability p_{uab}, for $u, a, b = 1, 0, +$, except for $(u, a, b) = (0, 0, 0)$ which is *not* part of the target-list universe. Let N_{uab} be the

size of the corresponding target-list domain, where $N_{000} \equiv 0$, that is a structural zero. The target population is given by $U = U^*_{1++}$ and its size by $N = N_{1++}$ in this notation. Let N_{uab} be observed for $(u, a, b) = (+, 1, 1)$, $(+, 1, 0)$ or $(+, 0, 1)$, that is the matching of list A and B is errorfree ("Matching"), and let all the list units be identified ("Nonresponse").

Thus, all the assumptions of the homogeneity model are retained, except for the three of "Spurious Events", "Closure" and "Causal Independence". This is of course not to say that the other assumptions are all beyond criticism. But they are not dealt with in this article. In particular, we modify the assumption of "Spurious Events" to exclude all other overoverage errors, such as duplicate reports, but allow for erroneous list enumeration. The "Closure" assumption is no longer necessary, because we now allow for erroneous list enumerations. What remains to be explored are the possibilities of replacing the assumption of "Causal Independence" (1).

3.2. Moment Equations Given Additional Survey Enumeration

The seven parameters of the multinomial distribution are not estimable given only three observed list-domain counts N_{+11}, N_{+10} and N_{+01}. Assume that there exists an additional coverage survey, denoted by S, which (I) has *only* undercoverage error so that all the units in S belong to U, and (II) can be matched to list A and B *without* errors.

The following additional notations seem convenient. Let n_{ab} be the observed number of units in S that belong to the list domain (a, b). Assume that the event of being enumerated in S is independent of the inclusion in the lists, such that

$$\pi_S = P\big(i \in S | i \in U^*_{1ab}\big) = P(i \in S) \tag{3}$$

It follows that $E(n_{ab}) = E(N_{1ab})\pi_S$. Consider two possible decompositions

$$E(N_{1ab}) = E(N)P\big(i \in U^*_{1ab} | i \in U\big) = E(N_{+ab})P\big(i \in U | i \in U^*_{+ab}\big) \tag{4}$$

for $(a, b) \neq (0, 0)$. The first conditional probability that unit $i \in U$ is in the list domain (a, b) will be referred to as the corresponding list *catch rate*, short handed as

$$\pi_{ab} = p_{1ab}/p_{1++}$$

for $a, b = 1, 0, +$. The second conditional probability is given by one minus the conditional probability that a unit in the list domain (a, b) is an erroneous enumeration, for $(a, b) \neq (0, 0)$, to be referred to as the corresponding list *error rate* and short handed as

$$\theta_{ab} = p_{0ab}/p_{+ab} = p_{0ab}/(p_{1ab} + p_{0ab})$$

Given that our interest is to see how the erroneous enumerations can be modelled, it will be useful to observe a set of moment equations, conditional on $\mathbf{x} = (x_{11}, x_{10}, x_{01})$ defined

by $x_{ab} = N_{+ab}$, given in terms of the list error rates:

$$\begin{cases} E(n_{11}|\mathbf{x}) = x_{11}(1-\theta_{11})\pi_S \\ E(n_{10}|\mathbf{x}) = x_{10}(1-\theta_{10})\pi_S \\ E(n_{01}|\mathbf{x}) = x_{01}(1-\theta_{01})\pi_S \\ E(n_{00}|\mathbf{x}) = \big(E(N|\mathrm{x}) - x_{11}(1-\theta_{11}) - x_{10}(1-\theta_{10}) - x_{01}(1-\theta_{01})\big)\pi_S \end{cases} \tag{5}$$

Notice that, since the unknown quantity $E(N|\mathbf{x})$ appears only in the last equation, this last equation can only be used to derive an estimate of $E(N|\mathbf{x})$ given the other parameter estimates. There are four parameters in the first three equations of (5). At least one additional assumption is needed from the different models, which can be compared to each other in terms of how they transform the first three equations. The strategy now is to examine systematically the possible log-linear models for, respectively, the target universe U, the target-list universe U^* and the *list universe*, denoted by $U_L = A \cup B$.

3.3. A Log-Linear Model of U

The list catch rates are defined for the units in U, conditional on which the N_{1ab}s form a two-way contingency table with fixed total N. The saturated log-linear model is

$$\log\, \pi_{ab} = \lambda + \lambda_a^A + \lambda_b^B + \lambda_{ab}^{AB}$$

(e.g., Agresti 2013). The largest nonsaturated model is given by

$$\lambda_{ab}^{AB} = 0 \Leftrightarrow \pi_{ab} = \pi_{a+}\pi_{+b} \Leftrightarrow \pi_{11}\pi_{00} = \pi_{10}\pi_{01} \tag{6}$$

that is the event of being enumerated in List A is independent of that in B. Given that $E(n_{ab}|N) = N\pi_{ab}\pi_S$, Model (6) implies

$$\begin{aligned} E(N_{111}|N) &= E(N_{11+}|N)E(N_{1+1}|N)/N \\ E(n_{11}|N)E(n_{00}|N) &= E(n_{10}|N)E(n_{01}|N) \end{aligned}$$

the latter of which can be checked given the n_{ab}s.

As discussed previously, one does not really expect (6) to hold for example between the census and the Patient Register, or between the Patient and the Electoral Registers, and so on. Still, to see the implications of (6) on the list error rates, let $\theta_{1+} = p_{01+}/p_{+1+}$ be the probability that a unit in list A is erroneous and $\theta_{+1} = p_{0+1}/p_{++1}$ that a unit is erroneous in list B. Combining (6) with decompositions like (4), we have

$$\frac{(1-\theta_{11})}{(1-\theta_{1+})(1-\theta_{+1})} = \frac{E(x_{1+})E(x_{+1})}{E(x_{11})E(N)} \tag{7}$$

On account of (7), we refer to (6) as an *incidental* model of the list error mechanism, in the sense that it imposes constraints between the list error rate and the target population size N. For instance, under (6), we have $N = E(N_{11+}|N)E(N_{1+1}|N)/E(N_{111}|N)$.

Since $N_{111} \leq N_{+11} = x_{11}$, and $N_{111} = N_{11+} - N_{110} \geq N_{11+} - N_{+10} = N_{11+} - x_{10}$, and $N_{111} = N_{1+1} - N_{101} \geq N_{1+1} - N_{+01} = N_{1+1} - x_{01}$, we must have

$$\frac{E(N_{11+}|N)E(N_{1+1}|N)}{E(x_{11}|N)} \leq N \leq \min\left(\frac{E(N_{11+}|N)E(N_{1+1}|N)}{E(N_{11+}|N) - E(x_{10}|N)}, \frac{E(N_{11+}|N)E(N_{1+1}|N)}{E(N_{1+1}|N) - E(x_{01}|N)}\right)$$

Now that each list error rate is a conditional probability *within* the list universe, such constraints on the target population size are unwarranted in general.

3.4. Log-Linear Models for Target-List Universe

The saturated log-linear model of p_{uab} of the target-list universe U^* is given by

$$\log p_{uab} = \lambda + \lambda_u^U + \lambda_a^A + \lambda_b^B + \lambda_{ua}^{UA} + \lambda_{ub}^{UB} + \lambda_{ab}^{AB} + \lambda_{uab}^{UAB}$$

Without losing generality, we shall set all the λs to zero except those with all their subscripts equal to one. The structural zero cell, that is, $p_{000} = 0$, can be accommodated by dropping the parameter λ, such that the seven parameters of the saturated model are $\left(\lambda_1^U, \lambda_1^A, \lambda_1^B, \lambda_{11}^{UA}, \lambda_{11}^{UB}, \lambda_{11}^{AB}, \lambda_{111}^{UAB}\right)$.

The largest nonsaturated hierarchical model is the one with $\lambda_{111}^{UAB} = 0$, denoted by $[UA][UB][AB]$, where

$$\begin{aligned}
p_{100} &= \exp\left(\lambda_1^U\right) \\
p_{010} &= \exp\left(\lambda_1^A\right) \\
p_{110} &= \exp\left(\lambda_1^U + \lambda_1^A + \lambda_{11}^{UA}\right) \\
p_{001} &= \exp\left(\lambda_1^B\right) \\
p_{101} &= \exp\left(\lambda_1^U + \lambda_1^B + \lambda_{11}^{UB}\right) \\
p_{011} &= \exp\left(\lambda_1^A + \lambda_1^B + \lambda_{11}^{AB}\right) \\
p_{111} &= \exp\left(\lambda_1^U + \lambda_1^A + \lambda_1^B + \lambda_{11}^{UA} + \lambda_{11}^{UB} + \lambda_{11}^{AB}\right)
\end{aligned}$$

It follows that

$$\log \frac{p_{011}}{p_{111}} = \log \frac{p_{010}}{p_{110}} + \log \frac{p_{001}}{p_{101}} + \log p_{100}$$

The three log ratios correspond to the log odds of list error in list domain $(1, 1)$, $(1, 0)$ and $(0, 1)$, respectively, denoted by $\operatorname{logit}\theta_{11}$, $\operatorname{logit}\theta_{10}$ and $\operatorname{logit}\theta_{01}$, whereas p_{100} is the proportion of target-population units outside of the list universe. In terms of the list error rates, then, the model amounts to the following assumption

$$\operatorname{logit}\theta_{11} = \operatorname{logit}\theta_{10} + \operatorname{logit}\theta_{01} + (\log E(N_{100}) - \log(N_{+++})) \tag{8}$$

which is an incidental model, just like (6). Since there are no compelling reasons why the conditional probabilities of erroneous enumeration within the list universe must depend on the number of target units *outside* of it, Model (8) cannot be of general use.

It is possible to further reduce the log-linear model. But this would only result in incidental models based on implausible assumptions. For instance, under model $[UA][AB]$

with $\lambda_{11}^{UB} = 0$ in addition, we would have

$$\frac{p_{001}}{p_{101}} = \frac{1}{p_{100}} \quad \text{and} \quad \frac{p_{010}}{p_{110}} = \frac{p_{011}}{p_{111}} = \frac{1}{p_{100}\exp\left(\lambda_{11}^{UA}\right)}$$

3.5. Log-Linear Models for List Universe

To separate p_{100} from the list error mechanism, consider now modelling the list universe $U_L = A \cup B$ with the conditional probabilities, for $(a, b) \neq (0, 0)$ and $u = 0, 1$,

$$q_{uab} = p_{uab}/(1 - p_{100})$$

The saturated log-linear model of q_{uab} is given by

$$\log q_{uab} = \lambda + \lambda_u^U + \lambda_a^A + \lambda_b^B + \lambda_{ua}^{UA} + \lambda_{ub}^{UB} + \lambda_{ab}^{AB} + \lambda_{uab}^{UAB}$$

Without losing generality, we shall set all the λs to zero except those with all their subscripts equal to one. There are two structural-zero cells in U_L, namely, $q_{000} = q_{100} = 0$, which can be accommodated by dropping the parameters λ and λ_1^U, such that the six parameters of the saturated model are $\left(\lambda_1^A, \lambda_1^B, \lambda_{11}^{UA}, \lambda_{11}^{UB}, \lambda_{11}^{AB}, \lambda_{111}^{UAB}\right)$.

The largest nonsaturated hierarchical model is the one with $\lambda_{uab}^{UAB} = 0$, where

$$\begin{aligned}
q_{010} &= \exp\left(\lambda_1^A\right) \\
q_{110} &= \exp\left(\lambda_1^A + \lambda_{11}^{UA}\right) \\
q_{001} &= \exp\left(\lambda_1^B\right) \\
q_{101} &= \exp\left(\lambda_1^B + \lambda_{11}^{UB}\right) \\
q_{011} &= \exp\left(\lambda_1^A + \lambda_1^B + \lambda_{11}^{AB}\right) \\
q_{111} &= \exp\left(\lambda_1^A + \lambda_1^B + \lambda_{11}^{UA} + \lambda_{11}^{UB} + \lambda_{11}^{AB}\right)
\end{aligned}$$

In terms of the log odds of erroneous enumeration, that is, logit θ_{11}, logit θ_{10} and logit θ_{01}, this amounts to the following assumption, for $(a, b) \neq (0, 0)$,

$$\text{logit } \theta_{ab} = a\gamma_A + b\gamma_B \Leftrightarrow \text{logit } \theta_{11} = \text{logit } \theta_{10} + \text{logit } \theta_{01} \tag{9}$$

This is a 'standard' null second-order interaction assumption, that is, $\lambda_{uab}^{UAB} = 0$, of the three-way classification of the list units. It is *not* an incidental model. Whether or not plausible for the particular data of concern, it is a model that can *not* be disregarded *a priori*, and it can readily be extended to situations involving more than two lists, where the log-linear model of the extended list universe can be put down similarly.

We note that further reduction of Model (9) would only result in less plausible assumptions. For instance, under model $[UA][AB]$ with $\lambda_{11}^{UB} = 0$ in addition, we have

$$\frac{q_{001}}{q_{101}} = 1 \text{ and } \frac{q_{010}}{q_{110}} = \frac{q_{011}}{q_{111}} = \exp\left(-\lambda_{11}^{UA}\right)$$

that is, the error rate is simply 0.5 for the units in B but not A, and it is the same for all the units in A whether they belong to list B or not, which seems unwarranted in general.

3.6. Two Alternative Log-Linear Models for List Universe

So far (9) is the only model of list erroneous enumeration that (i) does not involve incidental assumptions about the target population size, and (ii) can be extended to include more than two lists. When a list error rate is low, its logit does not differ much from its log. For instance, for a ten percent error rate, we have $\text{logit}\,0.1 = -2.2$ compared to $\log 0.1 = -2.3$. Replacing logit in (9) with log leads to the following log-linear model

$$\log \theta_{ab} = a\alpha_A + b\alpha_B \Leftrightarrow \log \theta_{11} = \log \theta_{10} + \log \theta_{01} \Leftrightarrow \theta_{11} = \theta_{10}\theta_{01} \qquad (10)$$

for $(a, b) = (1, 1), (1, 0), 0, 1)$, that is, the error rate of the units in both A and B is the product of the error rate of the units in only A (but not B) and that of the units in only B (but not A). That is, for $i \in U_L$,

$$P(i \notin U | i \in A \cap B) = P(i \notin U | i \in A \setminus B) P(i \notin U | i \in B \setminus A)$$

Clearly, every extension of (9) to the situation with more than two lists gives rise to a corresponding model (10), as the two differ only in the choice of the link function. Provided low error rates, the two are expected to yield nearly the same fit to the data. But the difference can become greater if some or all of the error rates are appreciable.

Now, consider the scenario where list A and B have high quality so that both have low erroneous enumerations, that is, both $\theta_{1+} = p_{01+}/p_{+1+}$ and $\theta_{+1} = p_{0+1}/p_{++1}$ are small, and both have high catch rates, so that the list domain $(1, 1)$ is much larger than domain $(1, 0)$ or $(0, 1)$. It then seems natural to expect the error rate to be even lower among the units in both A and B, that is, $\theta_{11} < \theta_{1+}$ and $\theta_{11} < \theta_{+1}$, while the error rates among the units that belong to only one list are comparatively high, that is, $\theta_{10} > \theta_{1+}$ and $\theta_{01} > \theta_{+1}$. It is thus worth considering $\theta_{11} = \theta_{1+}\theta_{+1}$ as an alternative to $\theta_{11} = \theta_{10}\theta_{01}$ above, that is,

$$\log \theta_{11} = \log \theta_{1+} + \log \theta_{+1} \Leftrightarrow \theta_{11} = \theta_{1+}\theta_{+1} \qquad (11)$$

The main difference is that θ_{11} can be much lower under (11) than under (10).

It should be noted that Model (11) does *not* belong to the standard log-linear models for cross classified counts based on the concept of conditional independence. The examination of the possible standard log-linear models above empirically verifies this for the two-list setting. Generically speaking, denote by X, Y and Z any three random events. A conditional independence assumption among them must be of the form

$$P(X \cap Y | Z) = P(X|Z)P(Y|Z)$$

that is, the *conditional joint* probability is the product of the *conditional marginal* probabilities. If we put X as erroneous enumeration for $i \in U_L$, and Y as its inclusion in list A and Z as its inclusion in B, then (11) has the form

$$P(X | Y \cap Z) = P(X|Y)P(X|Z)$$

that is, the *joint conditional* probability is the product of the *marginal conditional* probabilities. We refer to this as an assumption of *pseudoconditional independence* (PCI).

It is possible to develop classes of log-linear models that extend (11) to list CR data involving more than two lists. But we shall not go into the details here. Instead, let us look

at a heuristic example of why Model (11) may be more suitable than (10) *when the quality of the list enumerations is high*. Assume two lists that have *no* erroneous enumerations at all and $N_{+11} = N_{+1+} = N_{++1}$, in which case we have $\theta_{11} = \theta_{1+} = \theta_{+1} = 0$ while $(\theta_{10}, \theta_{01})$ do not exist. In other words, Model (11) holds but (10) is not applicable. Suppose now two units leave the population. First, in the ideal case, the two events are registered in both lists so that $(N_{+11}, N_{+1+}, N_{++1})$ are all reduced by two. Then, Model (11) still holds and (10) remains inapplicable. Next, suppose some lack of updating, such that the one event is registered in list A but not B, and the other is registered in B but not A. Then, we still have $\theta_{11} = 0$, but $\theta_{10} = \theta_{10} = 1$, and $\theta_{1+} = 1/(N_{+1+} - 1)$ and $\theta_{+1} = 1/(N_{++1} - 1)$. Model (10) errs much more than (11), because the difference between $\theta_{11} = 0$ and $\theta_{10}\theta_{01} = 1$ is much larger than the difference between $\theta_{11} = 0$ and $\theta_{1+}\theta_{+1} = 1/[(N_{+1+} - 1)(N_{++1} - 1)]$. One can go through the other possibilities of imperfect updating, and one will find that the Model (11) either holds or errs only little.

Both Model (10) and (11) can be fitted given survey data S. For the two-list setting, it is convenient to derive the MME from (5) directly (Appendix). We have

$$\hat{\theta}_{10} = \frac{x_{01}}{n_{01}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{10}}{x_{10}}\right) \quad \text{and} \quad \hat{\theta}_{01} = \frac{x_{10}}{n_{10}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{01}}{x_{01}}\right) \tag{12}$$

for Model (10), and

$$\hat{\theta}_{1+} = \frac{x_{+1}}{n_{+1}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{1+}}{x_{1+}}\right) \quad \text{and} \quad \hat{\theta}_{+1} = \frac{x_{1+}}{n_{1+}}\left(\frac{n_{11}}{x_{11}} - \frac{n_{+1}}{x_{+1}}\right) \tag{13}$$

for Model (11). Any estimated error rate that is negative will be replaced by 0.

4. Simulations

4.1. Range of Fitting

First we explore numerically the differences between the models outlined above, in order to better appreciate the conditions under which a good fit can be achieved for list CR data.

Consider the two-list CR data in Table 1. In Example (I), the number of units is 1,000 in list A and 1,200 in B and 900 in both A and B. The number of erroneous units is 50 in list A and 80 in B. The number of erroneous units among those in both A and B is left to vary, denoted by r_{11}. The number of erroneous units among those in A but not B is then $50 - r_{11}$,

Table 1. Two numerical examples of two-list CR data with under- and overcoverage

		A	B	A and B	A but not B	B but not A
(I)	List enumeration	1,000	1,200	900	100	300
	No. erroneous units	50	80	r_{11}	$50 - r_{11}$	$80 - r_{11}$
		A	B	A and B	A but not B	B but not A
(II)	List enumeration	1,200	1,350	900	300	450
	No. erroneous units	250	400	r_{11}	$250 - r_{11}$	$400 - r_{11}$

Table 2. Values of r_{11} at which models fit perfectly for data in Table 1

Model	Example (I)		Example (II)	
	r_{11}	$(\theta_{10},\ \theta_{01},\ \theta_{11})$	r_{11}	$(\theta_{10},\ \theta_{01},\ \theta_{11})$
(9)	33	(0.170, 0.157, 0.0367)	184	(0.220, 0.480, 0.207)
(10)	30	(0.200, 0.167, 0.0333)	155	(0.317, 0.544, 0.172)
(11)	3	(0.470, 0.257, 0.0033)	56	(0.208, 0.296, 0.062)

and it is $80 - r_{11}$ among those in B but not A. By varying r_{11}, the idea is to see when the Models (9), (10) and (11) appear most plausible. The case is similar for Example (II).

More specifically, for Example (I), Model (9) fits the CR data perfectly when, for some $1 \leq r_{11} \leq 49$, we have $\text{logit}(r_{11}/900) = \text{logit}((50 - r_{11})/100) + \text{logit}((80 - r_{11})/300)$, which occurs at $r_{11} = 33$. Model (10) fits perfectly at $r_{11} = 30$, where $\log(r_{11}/900) = \log((50 - r_{11})/100) + \log((80 - r_{11})/300)$, whereas Model (11) fits perfectly at $r_{11} = 3$, where $\log(r_{11}/900) = \log(50/900) + \log(80/1200)$. The corresponding errors rates are summarized in Table 2. Similarly for Example (II).

The situations that are favorable to Models (9) and (10) are seen to be fairly similar for relatively low error rates such as in Example (I). The one fits best at $r_{11} = 33$ and the other at 30. However, the difference between the two becomes larger as the error rates increase. In Example (II), the one fits best at $r_{11} = 184$ and the other at 155. Also the corresponding error rates are seen to differ more in this case.

Next, Model (11) is more suitable in situations where relatively more erroneous enumerations occur among the units that belong to only one list, while erroneous enumeration is much less probable for units in both lists. In Example (I), the PCI assumption (11) fits best when $r_{11} = 3$ and $\theta_{11} = 0.0033$, the latter of which is *much* lower than the marginal error rates $\theta_{1+} = 0.050$ and $\theta_{+1} = 0.067$. The contrast between θ_{11} on the one hand and $(\theta_{10}, \theta_{01})$ on the other is much larger than under model (9) or (10). The contrast is reduced as the error rates increase in Example (II). But the situation where Model (11) would be plausible is still quite different from those for the other two models.

In conclusion, both Models (10) and (11) are additions to the standard log-linear model (9) rooted in the concept of conditional independence. In particular, Model (11) provides an alternative in situations where there is a large contrast between the overcoverage error among the units in both lists and that among the units in only one list. The aim of the discussion above is to illustrate when the different models might be applicable and how they relate to each other.

4.2. Adjustment of Census Erroneous Enumeration

As mentioned earlier, adjustment of census erroneous enumeration traditionally requires a separate O-sample in addition to the independent U-sample for undercoverage adjustment. In theory, an O-sample selected from the list enumerations can be used to estimate the error rates $(\theta_{11}, \theta_{01}, \theta_{10})$. This requires making a strong assumption that fieldwork is able to identify *all* the erroneous list enumerations in the O-sample. It would also imply extra cost, although to some extent this can be controlled by the choice of the O-sample size. On both accounts, it seems of interest if the modelling

approach considered in this article can potentially provide useful adjustment of census erroneous enumeration *without* the need for conducting the fieldwork. The possibility is explored here.

Assume three datasets: census, denoted by A, register enumeration processed from administrative sources, denoted by B, and an independent undercoverage survey, denoted by S. Without losing generality, we shall suppose that the survey S attempts to enumerate everyone in the selected areas. This yields the two-list one-survey setting in each surveyed area. The following assumptions and observations are worth noting:

- The census erroneous enumeration rate is expected to be relatively low. We assume that the range of the marginal error rate θ_{1+} of the census (i.e., List A) is reasonably covered by the following set of values: $\theta_{1+} = 0.2\%, 0.5\%, 1\%$.
- The register enumeration can have a higher, even much higher, marginal error rate θ_{+1}. We shall explore the following set of values: $\theta_{+1} = 1\%, 5\%, 10\%, 20\%$.
- Provided independent survey (Equation 3), we have $E(n) = E(N)\pi_S = E(N_{1++})\pi_S$ where n is the total survey enumeration, and $E(n - n_{00}) = E(N_{1++} - N_{100})\pi_S$ where n_{00} is the number of individuals enumerated in S that do not belong to list A nor B. Thus, the *overall* list catch rate can be given by

$$\frac{E(N - N_{100})}{E(N)} = \frac{E(N_{1++} - N_{100})}{E(N_{1++})} = \frac{E(n - n_{00})}{E(n)}$$

  and estimated by $1 - n_{00}/n$, irrespective of the error rates. An important implication is that the *relative* bias induced by the misspecification of a *nonincidental* erroneous enumeration model is unrelated to the target population size N.

- Provided the theoretical value of θ_{11} in addition to θ_{1+} and θ_{+1}, a straightforward simulation approach to evaluate the potential bias of an error model is to repeatedly generate $\mathbf{n} = (n_{11}, n_{10}, n_{01}, n_{00})$ under some given value of π_S, conditional on the target-list universe, and calculate the average of $\hat{N}$ over all the repetitions. More convenient, however, is to fit the moment Equations (5) just *once* to the expected values of $\mathbf{n}$, denoted by $\dot{\mathbf{n}}$, and use the difference between the corresponding $\hat{N}(\dot{\mathbf{n}})$ and N as an approximation to the model bias. This has two advantages: firstly, it makes it clear that the result is invariant to the arbitrary choice of π_S, which cancels out on both sides of the equations in (5) at $\dot{\mathbf{n}} = E(\mathbf{n}|U^*)$; secondly, the result is not subjected to the Monte Carlo errors of the repeated sampling approach.

For comparison to the equally cost-efficient approach without extra fieldwork associated with the O-sample, we consider the DSE based on census A and undercoverage survey S, that is ignoring the potential erroneous census enumerations. Corresponding to the expected survey enumeration $\dot{n}$, this is given by

$$\dot{N}_{DSE} = \dot{n}x_{1+}/\dot{n}_{1+} \approx E(\hat{N}_{DSE}|U^*)$$

Clearly, the relative bias of this unadjusted DSE is simply θ_{1+}, because the hypothetical unbiased DSE is then given by $\dot{n}x_{1+}(1 - \theta_{1+})/\dot{n}_{1+}$.

Table 3. Range of relative bias under Model (10) and (11) for census enumeration error adjustment. Census enumeration = 1,000, register enumeration = 1,200, census-register enumeration = 900. Error rate of census errra (θ_{1+}), register enumeration (θ_{+1}), census-register enumeration (θ_{11}), where $0 < \theta_{11} < \theta_{1+}$. All numbers in %.

Model (10)	Register error rate			
Census error rate	1	5	10	20
0.2	(0.078, 0.078)	(−0.11, −0.11)	(−0.48, −0.48)	(−3.4, −3.4)
0.5	(−0.038, 0.43)	(−0.88, 0.32)	(−2.5, 0.095)	(−16, −1.6)
1	(−0.25, 1)	(−2.3, 1)	(−6.3, 1)	(−38, 1)
Model (11)	**Register error rate**			
Census error rate	1	5	10	20
0.2	(0.11, 0.11)	(0.11, 0.11)	(0.1, 0.1)	(0.089, 0.089)
0.5	(0.11, 0.45)	(0.091, 0.44)	(0.068, 0.44)	(0.014, 0.43)
1	(0.1, 1)	(0.065, 1)	(0.012, 1)	(−0.11, 1)

Table 3 gives the range of relative bias under the Model (10) and (11), respectively. For each combination of $(\theta_{1+}, \theta_{+1})$, the number of erroneous enumeration N_{011} among the units in both A and B (i.e., the census-register enumeration) is bounded upwards by $\min(N_{+1+}\theta_{1+}, N_{++1}\theta_{+1})$ for the given target-list universe. In the simulation setting here, this is always equal to the integer $N_{+1+}\theta_{1+} = x_{1+}\theta_{1+}$. Each possible N_{011} yields a different target population size $N = N_{1++}$, a corresponding 'joint' error rate $\theta_{11} = N_{011}/x_{11} = N_{011}/N_{+11}$, and a set of expected survey enumerations $\dot{\mathbf{n}}$. The relative bias of a model is given by $\hat{N}(\dot{\mathbf{n}})/N - 1$, where $\hat{N}$ is derived from (12) under Model (10) and (13) under Model (11). As explained above, this relative bias is invariant towards any arbitrary but admissible choice of the survey catch rate π_S and the overall list catch rate adopted in the simulation. The relative biases corresponding to $N_{011} = 1$ and $N_{011} = x_{1+}\theta_{1+} - 1$, respectively, yield the range of relative bias reported in Table 3.

Take first the results for Model (10) in the upper half of Table 3. At $\theta_{1+} = 0.2\%$ and with census enumeration being 1,000, there are only two erroneous census enumerations, and the DSE has a relative bias of 0.2%. Only $N_{011} = 1$ is in the range to be examined, so that the lower and upper ends of the relative bias range coincide in this case. As the register error rate θ_{+1} increases, the estimate of N_{011} increases under Model (10), to the extent that it is 31.6 when the register error rate is 20%, leading to a large negative bias −3.4% due to model misspecification. Next, at $\theta_{1+} = 0.5\%$, the two end points correspond to $N_{011} = 1$ and $N_{011} = 4$. Model (10) is most misleading at the lower end, as the exploration in Subsection 4.1 has indicated, where the estimate of N_{011} is 142.6, leading to a disastrous negative relative bias for N. The performance becomes even worse at $\theta_{1+} = 1\%$, where large negative bias already occurs somewhere between $\theta_{+1} =$ 1% and 5%. At the upper end, where $N_{011} = 9$, the MME (12) is initially negative and needs to be truncated to 0, that is, no census erroneous enumeration at all. The model estimate $\hat{N}$ then becomes the same as the DSE, and has the same relative bias which is equal to θ_{1+}.

In short, when misspecified, Model (10) can lead to grave negative bias in situations where both the census and the register have non-negligible error rates but the error rate is much lower among the census-register enumeration. For example, at $(\theta_{1+}, \theta_{+1}) = (1\%, 5\%)$, the negative bias of Model (10) would be larger in absolute value than the bias of the DSE for all $\theta_{11} < 0.4\%$.

Turning now to Model (11), we notice immediately that its bias is in *no* case larger than that of the DSE. At $\theta_{1+} = 0.2\%$ and $N_{011} = 1$, the estimate of N_{011} increases from 0.007 at $\theta_{+1} = 1\%$ to 0.2 at $\theta_{+1} = 20\%$. In absolute terms, however, such differences have essentially no bearing on the resulting bias, which is about half of that of the DSE across the range of θ_{+1}. Next, at $\theta_{1+} = 0.5\%$, the model predicted value of N_{011} would be somewhere between 0 and 1 for all the values of θ_{+1} here. As N_{011} increases from 1 and 4, the fitted N_{011} (and N_{01+}) decreases steadily towards 0, resulting in the bias to increase towards that of the DSE. The case is similar at $\theta_{1+} = 1\%$, where Model (11) removes almost all the bias of the DSE as $N_{011} \rightarrow 1$, while tending towards the DSE as $N_{011} \rightarrow 9$.

Thus, it looks like Model (11) is a more robust choice than (10) for potential adjustment of census erroneous enumeration using an additional list enumeration derived from administrative sources. Within the plausible range of marginal error rates of the census and register enumerations (e.g., in Table 3), the PCI assumption (11) removes essentially all the bias of the census-survey DSE as the number of erroneous enumerations among the units in both the census and the register (i.e., N_{011}) tends to zero. At the other other end, as the latter tends towards its upper bound, that is, $N_{011} \rightarrow \min(N_{01+}, N_{0+1})$, the bias of the model estimate increases towards that of the DSE.

5. Summary and Discussion

Above we have considered some approaches to modelling erroneous enumeration as a type of overcoverage error. Two types of nonincidental models of the list universe are identified. The first of these consists of standard log-linear models, such as (9), and the associated models using alternative link functions, such as (10). The second of these refers to a class of log-linear models that build on the concept of pseudoconditional independence. The two types of models are suitable for different error mechanisms of the data, and are therefore complementary to each other in practice.

One possible application is the adjustment of census erroneous enumeration based on an independent coverage survey and an additional register enumeration processed from administrative sources. Simulations under what seems to be the plausible range of the census and register error rates suggest that Model (11) is robust towards misspecification of the error rate among the ones enumerated in both the census and the register. The potential bias is bounded upwards by the bias of the DSE that ignores erroneous enumeration.

Of course, further investigation should also take into account the variance of the DSE compared to that of the adjusted model estimator. Simulation on the historic census and register data will be necessary. Moreover, it is important to consider the over and undercoverage adjustments hand in hand. Various authors have considered the so-called triple-system estimator (TSE) based on census, register and coverage survey for adjusting under-coverage. See Griffin (2014) for a recent update. A traditional motivation for the TSE is the possibility to relax the "Causal Independence" assumption (1).

An independent survey, however, is needed in the two-list setting that allows for overcoverage errors. There is simply not enough degree of freedom otherwise. The tension needs to be resolved.

An approach to census-like population statistics without the census is a more ambitious goal. To start with, the census may be replaced by an "improved administrative file" (i.e., register), as some countries have done already. A modelling approach can be used to assess and potentially adjust the erroneous register enumeration, provided very little or no fieldwork associated with the O-sample. It also opens up the possibility for using several input registers instead of one combined register.

Appendix

Method-of-Moment Estimator (MME)

Dividing the first equation in (5) by the second and third, respectively, we obtain

$$\begin{cases} n_{11}(x_1 - r_1) = n_1(x_{11} - r_{11}) = n_1 x_{11}(1 - (r_1 r_2)/(x_1 x_2)) \\ n_{11}(x_2 - r_2) = n_2(x_{11} - r_{11}) = n_2 x_{11}(1 - (r_1 r_2)/(x_1 x_2)) \end{cases}$$

where $(n_1, n_2) = (n_{10}, n_{01})$, $(x_1, x_2) = (x_{10}, x_{01})$ and $(r_1, r_2) = \left(x_{10}\hat{\theta}_{10}, x_{01}\hat{\theta}_{01}\right)$ under Model (10), and $(n_1, n_2) = (n_{1+}, n_{+1})$, $(x_1, x_2) = (x_{1+}, x_{+1})$ and $(r_1, r_2) = \left(x_{1+}\hat{\theta}_{1+}, x_{+1}\hat{\theta}_{+1}\right)$ under Model (11). Note the symmetry between r_1 and r_2. We have

$$ar_1^2 - br_1 + c = 0 \quad \text{where } (a, b, c) = \left(\frac{n_2}{n_1 x_1 x_2}, \frac{n_{11}}{x_{11} n_1} + \frac{n_2}{n_1 x_2} - \frac{1}{x_1}, \frac{n_{11} x_1}{x_{11} n_1} - 1\right)$$

After some algebra we obtain

$$\Delta = b^2 - 4ac = \left(-\frac{n_{11}}{x_{11} n_1} + \frac{n_2}{n_1 x_2} + \frac{1}{x_1}\right)^2 \quad \text{so that} \quad \frac{b + \sqrt{\Delta}}{2a} \equiv x_1$$

It follows that the admissible r_1 and, by symmetry, r_2 are given by

$$r_1 = \frac{x_2}{n_2}\left(\frac{n_{11}}{x_{11}} x_1 - n_1\right) \quad \text{and} \quad r_2 = \frac{x_1}{n_1}\left(\frac{n_{11}}{x_{11}} x_2 - n_2\right)$$

We obtain r_1/x_1 as $\hat{\theta}_{10}$ under (10) or $\hat{\theta}_{1+}$ under (11). The case is similar for r_2. We obtain $\hat{\theta}_{11}$ according to either Model (10) or (11). Next, we obtain $\hat{\pi}_S = (x_1 - r_1)/n_1 = (x_2 - r_2)/n_2$, and $\hat{N}$ on substituting these parameter estimates into the last equation of (5). Linear approximation yields the variance of the MME.

6. References

Agresti, A. 2013. *Categorical Data Analysis*, 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc.

Brown, J., O. Abbott, and Paul A. Smith. 2011. "Design of the 2001 and 2011 Census Coverage Surveys for England and Wales." *Journal of the Royal Statistical Society Series A (Statistics in Society)* 174: 881–906. Doi: http://dx.doi.org/10.1111/j.1467-985X.2011.00697.x.

Cormack, R.M. 1989. "Log-Linear Models for Capture-Recapture." *Biometrics* 45: 395–413. Doi: http://dx.doi.org/10.2307/2531485.

Fienberg, S.E. 1972. "The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables." *Biometrika* 59: 409–439. Doi: http://dx.doi.org/10.1093/biomet/59.3.591.

Griffin, R.A. 2014. "Potential Uses of Administrative Records for Triple System Modeling for Estimation of Census Coverage Error in 2020." *Journal of Official Statistics* 30: 177–189. Doi: http://dx.doi.org/10.2478/jos-2014-0012.

Hogan, H. 1993. "The Post-Enumeration Survey: Operations and Results." *Journal of the American Statistical Association* 88: 1047–1060. Doi: http://dx.doi.org/10.1080/01621459.1993.10476374.

IWGDMF – International Working Group for Disease Monitoring and Forecasting. 1995a. "Capture-recapture and Multiple-record Systems Estimation I: History and Theoretical Development." *American Journal of Epidemiology* 142: 1047–1058. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7485050 (accessed 15 July 2015).

IWGDMF – International Working Group for Disease Monitoring and Forecasting. 1995b. "Capture-recapture and Multiple-record Systems Estimation 2: Applications." *American Journal of Epidemiology* 142: 1059–1068. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7485051.

Nirel, R. and H. Glickman. 2009. "Sample Surveys and Censuses." In *Sample Surveys: Design, Methods and Applications*, Vol. 29A, edited by D. Pfeffermann and C.R. Rao. pp. 539–565.

ONS – Office for National Statistics. 2013. *Beyond 2011: Producing Population Estimates Using Administrative Data: In Practice*. ONS Internal Report, available at: http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/index.html (accessed 15 July 2015).

Renaud, A. 2007. "Estimation of the Coverage of the 2000 Census of Population in Switzerland: Methods and Results." *Survey Methodology* 33: 199–210.

Wolter, K. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81: 338–346. Doi: http://dx.doi.org/10.1080/01621459.1986.10478277.

Zhang, L.-C. 2012. "Topics of Statistical Theory for Register-based Statistics and Data Integration." *Statistica Neerlandica* 66: 41–63. Doi: http://dx.doi.org/10.1111/j.1467-9574.2011.00508.x.

Received January 2014
Revised November 2014
Accepted November 2014