DE GRUYTER OPEN

Journal of Official Statistics, Vol. 35, No. 1, 2019, pp. 279–283, http://dx.doi.org/10.2478/JOS-2019-0012

A Note on Dual System Population Size Estimator

Li-Chun Zhang[1]

Several countries are currently investigating the possibility of replacing the costly population census with a Population Data set derived from administrative sources, in combination with a purposely designed Population Coverage Survey. We formulate the assumptions of the dual system estimator in this context, and contrast them to the situation involving a census and a Census Coverage Survey.

Key words: Coverage error; undercount; capture-recapture; administrative data.

1. Introduction

The dual system estimator (DSE) has been used for adjusting population census undercoverage error with the help of a Census Coverage Survey (CCS). See Nirel and Glickman (2009) for a review. Wolter (1986) lists eight assumptions, that is, Assumption 1-6, 8, and 11, for the DSE in the most basic setting. Several countries are investigating the possibility of replacing the costly population census with a Population Data set (PD) derived from administrative sources, in combination with a purposely designed Population Coverage Survey (PCS). There is a need to pin down the assumptions of the DSE based on the PD and PCS, because the data generation process of a PD can be quite different and more difficult to model than that of a census. We propose to treat the PD as fixed and to consider the PCS as the only random source. This allows one to circumvent the problem of modelling the PD enumeration, where the mechanisms at the sources of the data may lie beyond the control of the statistician, and to focus on the design and implementation of the PCS, which is under the direct control of the statistician. In Section 2, we formulate the assumptions in the basic setting that is comparable to that of Wolter (1986). The advantages of the proposed conditional approach to the DSE, given the PD, will be explained in comparison to the traditional approach, where both the census and CCS are considered to be random. Some additional remarks on departures from the basic setting are given in Section 3, as well as some related ongoing developments.

2. The Four Assumptions for Consistency

Denote by $U = \{1, 2, \ldots, N\}$ the target population, which is of the unknown size N. Let A contain x enumeration records from the PD. Let S contain n records from the PCS. For each $i \in U$, let $\pi_i = E(\delta_i|A)$, where $\delta_i = 1$ if $i \in S$ and 0 otherwise. The notation $E(\cdot|A)$

[1] University of Southampton, Social Statistics and Demography, Highfield, Southampton, SO17 1BJ, United Kingdom. Email: L.Zhang@soton.ac.uk

emphasises that the enumeration in A or not is treated as fixed for all $i \in U$. In the most basic setting that is comparable to the DSE based on census and CCS, the following four assumptions are needed for the DSE based on PD and PCS:

(i) There are no duplicated records in A or S, and $A \cup S \subseteq U$.
(ii) The matched records between A and S can be identified without errors.
(iii) The capture probability in S is constant, that is, $\pi_i = \pi$ and $0 < \pi < 1$, for $i \in U$.
(iv) The captures in S are uncorrelated, that is, $Cov(\delta_i, \delta_j|A) = 0$ for $i \neq j \in U$.

According to (i), there are no duplicated or erroneous records in A or S, where a record i is erroneous if $i \notin U$. This is the same as Assumption 5 "Spurious Events" in Wolter (1986). The assumption (ii) combines Assumption 4 "Matching" and 7 "Nonresponse" in Wolter (1986). According to (iii), every population element has the same positive inclusion probability in the PCS. This is the same as Assumption 11 in Wolter (1986), except that it refers only to the capture probabilities in the PCS that are designed by the statistician, not the inclusion in the PD. Finally, the assumption (iv) is the same as Assumption 3 "Autonomous Independence" in Wolter (1986), except that it only pertains to the PCS enumeration, not the PD.

Let m be the number of matched records between A and S. Provided the assumptions (i) – (iii), we have $\mu_m = E(m|A) = \sum_{i \in A} \pi = x\pi$ and $\mu_n = E(n|A) = \sum_{i \in U} \pi = N\pi$, such that $x\mu_n/\mu_m = N$. Replacing μ_n and μ_m by n and m, respectively, we obtain the DSE

$$\hat{N} = xn/m. \tag{1}$$

It is important to notice that the only random source is the PCS that generates S, whereas we treat A (and x) as fixed, regardless of how complicated the data generation process may be that leads to the creation of A. In particular, the PD is allowed to have systematic undercoverage of the population, which is often the case with administrative registers. For instance, data set A may contain everyone that pays tax, but none who does not. The DSE (1) can still be motivated, because the estimated capture probability of the PCS among the tax payers, that is, m/x, can be extrapolated to the others, as long as the PCS satisfies assumption (iii). This provides additional flexibility, which is not permitted under the traditional approach to census and CCS. Moreover, treating A as fixed removes two of the three remaining assumptions of Wolter (1986). Assumption 2 "Multinomial" distribution of $(\delta_{i,census}, \delta_{i,CCS})$ is unnecessary, where $\delta_{i,census}$ and $\delta_{i,CCS}$ are the enumeration indicators of the census and CCS, now that inclusion or not in A is treated as fixed. Likewise, Assumption 8 "Causal Independence" between $\delta_{i,census}$ and $\delta_{i,CCS}$ is unnecessary, since the random variable δ_i cannot be correlated with a constant, that is, $i \in A$ or not.

Finally, Assumption 1 "Closure" of the population was used to ensure that the census and CCS aim at the same target population. In practice, it creates some tension to Assumption 8 "Causal Independence": to accommodate "Closure" one would conduct the CCS as close as possible to the census, yet doing so can potentially jeopardise "Causal Independence". The "Closure" assumption is no longer necessary, provided assumptions (i) and (iii) are satisfied. It is possible to implement any census population definition in the PCS, provided one can extract the data set A from the PD, which satisfies the assumption (i). For example, suppose the reference date is 11 November 2017 for a census night

population definition. The PCS may be deployed on the same day or immediately afterwards. Any member of the population has a chance to be enumerated in *S,* provided (iii), and no overcounting occurs, provided (i). Meanwhile, the processing of list *A* from the PD can take place both before and after 11 November 2017, aimed to satisfy assumption (i) and avoid erroneous enumeration. There is no need to assume that the target population itself is closed for a prolonged period after 11 November 2017.

Expanding $\hat{N}$ with respect to (n,m) around (μ_n,μ_m) yields

$$\hat{N}=N+\frac{x}{\mu_m}(n-\mu_n)-\frac{N}{\mu_m}(m-\mu_m)-\frac{x}{\mu_m^2}(n-\mu_n)(m-\mu_m)+\frac{N}{\mu_m^2}(m-\mu_m)^2+R_3, \quad (2)$$

where R_3 is the remainder. We have

$$\begin{aligned} E\left(\frac{\hat{N}}{N}|A\right)-1&=\left(1-\frac{x}{N}\right)\mu_m^{-2}V(m|A)+\frac{1}{N}E(R_3)\\ V(\hat{N})&\approx\frac{(N-x)^2}{\mu_m^2}V(m|A)+\frac{x^2}{\mu_m^2}V(n-m|A), \end{aligned} \quad (3)$$

where $V(m|A)=x\pi(1-\pi)$ and $V(n-m|A)=(N-x)\pi(1-\pi)$. Notice that we have used $Cov(n-m,m|A)=0$ and $Cov(n,m|A)=V(m|A)$, due to the assumption (iv). Provided $x/N=O(1)$ asymptotically, as $N\to\infty$, and $E(R_3)/N$ is of a lower order than the first term on the right-hand side of (3), we have $E\left(\frac{\hat{N}}{N}|A\right)\to 1$, because $V(m|A)/x=O(1)$ and $x/\mu_m=O(1)$. Now that $V\left(\frac{\hat{N}}{N}|A\right)\to 0$ in addition, the DSE (1) is such that $N/\hat{N}\to P_1$ asymptotically, as $N\to\infty$. The consistency of the DSE based on PD and PCS can thus be established under the assumptions (i) – (iv).

3. Additional Remarks

Below we consider potential departures from the four basic assumptions, taking them one by one in the reverse order.

Correlated PCS captures The assumption (iv) can be relaxed to allow correlated captures, such as intra-cluster correlated enumeration within the same household or building. Let the population U be partitioned into K clusters, denoted by $U=\cup_{k=1}^{K}U_k$.

(iv.c) $Cov(\delta_i,\delta_j)=0$ for $i\in U_k$ and $j\in U_l$, for $1\le k\ne l\le K$.

Provided (iv.c) instead of (iv), we have

$$E\left(\frac{\hat{N}}{N}|A\right)-1\approx\left(1-\frac{x}{N}\right)\mu_m^{-2}V(m|A)-\frac{x}{N}\mu_m^{-2}Cov(n-m,m|A),$$

where $V(m|A)=\sum_{i\in A}\pi_i(1-\pi_i)+\sum_{k=1}^{K}\sum_{i\ne j\in A_k}Cov(\delta_i,\delta_j)$, for $A_k=A\cap U_k$, and $Cov(n-m,m|A)=\sum_{k=1}^{K}\sum_{i\in A_k}\sum_{j\in A_k^c}Cov(\delta_i,\delta_j)$, for $A_k^c=U_k\backslash A_k$. Asymptotically, as $N\to\infty$, provided $x/N=O(1)$ as before, and $K/N=O(1)$ and $N_k=O(1)$, where N_k is the size of U_k which remains bounded asymptotically, the consistency of the DSE (1) is

retained. Moreover, the variance of $\hat{N}$ is now approximately given by

$$V(\hat{N}) = \frac{(N-x)^2}{\mu_m^2} V(m|A) + \frac{x^2}{\mu_m^2} V(n-m|A) - 2\frac{x(N-x)}{\mu_m^2} Cov(n-m, m|A),$$

where $V(n-m|A) = \sum_{i \in U \setminus A} \pi_i(1-\pi_i) + \sum_{k=1}^{K} \sum_{i \neq j \in A_k^c} Cov(\delta_i, \delta_j)$.

Heterogeneous PCS capture The assumption (iii) can be relaxed.

(iii.h) $\pi_i = \pi_h$ and $0 < \pi_h < 1$, for $i \in U_h$, where $U_1, \ldots, U_H$ form a post-stratification of the target population U.

Post-stratification is common in the practice of census-CCS DSE. Wolter (1986) introduces Assumption 7 "Post-stratification" to ensure that any variable used for the post-stratification is error-free. Provided this and the assumption (iii.h) instead of (iii), one may employ a post-stratified DSE based on the PD and PCS, which is given by

$$\hat{N}_p = \sum_{h=1}^{H} x_h n_h / m_h, \tag{4}$$

where x_h is the size of $A \cap U_h$, and n_h that of $S \cap U_h$, and m_h that of $A \cap S \cap U_h$. Asymptotically, as $N_h \to \infty$ for all $h = 1, \ldots, H$, we have $\hat{N}_p / N \xrightarrow{P} 1$.

(iii.a) $\bar{\pi}_A = \bar{\pi}_A^c$, where $\bar{\pi}_A = \sum_{i \in A} \pi_i / x$ and $\bar{\pi}_A^c = \sum_{i \in U \setminus A} \pi_i / (N-x)$ are the *average* capture probabilities among the population elements in and out of A, respectively.

According to (iii.a), the PCS does not have to achieve a constant capture probability across the population, which is less stringent than the assumption (iii). We have

$$x\frac{\mu_n}{\mu_m} = N\left(\frac{x}{N} + \left(1 - \frac{x}{N}\right)\frac{\bar{\pi}_A^c}{\bar{\pi}_A}\right) = N,$$

where $\mu_m = E(m|A) = x\bar{\pi}_A$, and $\mu_n = E(n|A) = N\left[(x/N)\bar{\pi}_A + (1-x/N)\bar{\pi}_A^c\right]$. The relative bias of the DSE is still given by (3), except that we now have $V(m|A) = \sum_{i \in A} \pi_i(1-\pi_i)$ and $V(n-m|A) = \sum_{i \in U \setminus A} \pi_i(1-\pi_i)$. Asymptotically, as $N \to \infty$, it converges to zero as before, so that the consistency property of the DSE (1) is retained.

(iii.ha) $\bar{\pi}_{A_h} = \bar{\pi}_{A_h}^c$, where $\bar{\pi}_{A_h} = \sum_{i \in A \cap U_h} \pi_i / x_h$ and $\bar{\pi}_{A_h}^c = \sum_{i \in U_h \setminus A} \pi_i / (N_h - x_h)$.

The assumption (iii.ha) combines (iii.h) and (iii.a), provided which the post-stratified DSE (4) retains its consistency property, as $N_h \to \infty$ for all $h = 1, \ldots, H$.

Linkage error The Matching assumption (ii) may be violated unless a unique identifier is available in both A and S, which can be used to link the records directly. See Ding and Fienberg (1994), Di Consiglio and Tuoto (2015) for a discussion in the presence of linkage errors. To adjust the DSE, one needs to obtain estimates of the relevant linkage error probabilities, which is not an easy task in practice. Moreover, heterogeneous linkage error probabilities may further complicate the treatment of heterogeneous catch probabilities (in S). ONS-M8 (2013) outlines a potential alternative approach, which is to match A and S at

a cluster level (such as address or dwelling) that is not affected by linkage errors. However, the approach requires an additional assumption that the PCS fully enumerates everyone in the captured clusters, which may be difficult to satisfy in practice.

Erroneous enumeration The assumption (i) is violated for A if it contains erroneous records. The traditional approach is to include an additional survey (sampled from A) to estimate the over-coverage rate (e.g., Nirel and Glickman 2009). In some recent works, models and methods are developed to accommodate erroneous records directly. Zhang (2015) considers log-linear models of two PDs, subjected to both erroneous and missing records, together with the PCS. Zhang and Dunne (2017) apply the trimmed DSE to Irish data to explore the potential over-coverage error of the PD. In situations where the PD is compiled from multiple administrative registers, it is possible to trim one or more source registers directly. Di Cecco et al. (2018) develop latent class models based on four or more enumeration lists, all of which may be subjected to erroneous enumeration.

In particular, the treatment of linkage error and erroneous enumeration are important research topics for the census transformation programmes in the coming years.

4. References

Di Cecco, D., M. Di Zio, D. Filipponi, and I. Rocchetti. 2018. "Population Size Estimation Using Multiple Incomplete Lists with Overcoverage." *Journal of Official Statistics* 34: 557–572. Doi: http://dx.doi.org/10.2478/JOS-2018-0026.

Ding, Y. and S.E. Fienberg. 1994. "Dual System Estimation of Census Undercount in the Presence of Matching Error." *Survey Methodology* 20: 149–158.

Di Consiglio, L. and T. Tuoto. 2015. "Coverage Evaluation on Probabilistically Linked Data." *Journal of Official Statistics* 31: 415–429. Doi: http://dx.doi.org/10.1515/JOS-2015-0025.

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: 539–565.

ONS-M8. 2013. *Beyond 2011: Producing Population Estimates Using Administrative Data: In Theory*. Avaiable at: https://www.ons.gov.uk/census/censustransformationprogramme/beyond2011censustransformationprogramme/reportsandpublications.

Wolter, K. 1986. "Some Coverage Error Models for Census Data." *Journal of the American Statistical Association* 81: 338–346. Doi: http://www.jstor.org/stable/2289222.

Zhang, L.-C. 2015. "On Modelling Register Coverage Errors." *Journal of Official Statistics* 31: 381–396. Doi: http://dx.doi.org/10.1515/JOS-2015-0023.

Zhang, L.-C. and J. Dunne. 2017. "Trimmed Dual System Estimation." In *Capture Recapture Methods for the Social and Medical Sciences*, edited by D. Böhning, J. Bunge, and P.v.d. Heijden: 239–259. Chapman and Hall/CRC.

Received November 2017
Revised April 2018
Accepted July 2018