DE GRUYTER OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 507–513, http://dx.doi.org/10.1515/JOS-2015-0030

Discussion

Ray Chambers[1]

1. Introduction

I am very grateful for the opportunity to contribute to this special issue of the Journal of Official Statistics by commenting on the articles in it. In particular, I have chosen to focus my comments on the articles by Burger et al., Gerritse et al., Tuoto and Di Consiglio, and Zhang, because these authors, to a greater or lesser extent, tackle measurement-error issues that are important emerging features of official statistics methodology.

2. Comments on Burger et al. article

I start with the article by Burger et al. This addresses the important issue of industry misclassification when records from a survey and an administrative data source are combined. In particular, the article considers a business survey application where in fact a census is carried out, in the sense that there a 100% survey of large businesses is conducted, with data for the remaining medium and small businesses extracted from a tax register. To quote the authors, "Because no samples are drawn and missing data are imputed, no complicated design-based or model-based estimators are required to make inference about the target population." This of course ignores the whole minefield of imputation bias and variability, as well as the usual conceptual issues that arise when two variables ostensibly referring to the same thing are measured in two different ways. But once one pushes this (huge) elephant out of the living room, then the issue of errors in the industry classification of the units in the two sources can be considered. The article introduces a simple model for misclassification errors within a group of industries that is the same as the simple exchangeable model for linkage errors introduced by Neter et al. (1965), and used as the basis for bias correction in that context in a series of papers starting with Chambers (2009). However, the authors of this article are not interested in bias correction *per se*, focusing instead on bootstrap simulation of the extent of the bias and the increase in variability that arise under a multinomial version of this simple model. Here their results are sobering, indicating quite significant increases in both bias and variability even when the data meet the quality specifications of an internal Service Level Agreement (SLA) on classification accuracy. Interestingly, the results in the article show that because higher levels of accuracy in classifying small to medium businesses lead to reductions in bias relative to expected levels under the SLA, there is in fact a large bias-variance trade-off to be made in terms of allocating resources for carrying out the classification. No information is provided on how this trade-off can be (was?) eventually resolved, but, again quoting the authors, "the current paper provides insight into the sensitivity of mixed source statistics to a source-specific nonsampling error." Much more research needs to be done,

[1]National Institute for Applied Statistics Research, Wollongong NSW 2522, Australia. Email: ray@uow.edu.au

particularly in terms of developing robust misclassification bias corrections for the outputs from the application. This is particularly the case since these outputs appear to be an important component of the information used in determining gross domestic product. In this context, the work on bias correction for linkage errors may prove useful, see Kim and Chambers (2012).

3. Comments on Gerritse et al. article

The remaining three articles all focus on a different type of measurement error, introduced when two or more data sources, each with incomplete coverage of a population of interest, are linked in order to estimate the total population size. This of course is the classical census coverage problem, and the so-called dual-system estimation (DSE) methodology for dealing with it is now well established. The article by Gerritse et al. uses the DSE as a jumping-off point, providing a nice overview of the main issues that arise when using this approach, and particularly focusing on the problems that arise when the union of the two sources is a subset of the population of interest (so undercoverage is the focus) and the key assumption of independent coverage errors for the two data sources is in fact incorrect. In this context it helps to introduce some notation, so let $A = 1(0)$ denote the event that a population unit (of some agreed type) is included (not included) in the first data source, and let $B = 1(0)$ denote the same two events for the second data source. Put $N_A(N_B)$ equal to the known counts of population units with $A = 1$ $(B = 1)$ and put L equal to the set of linked units, with X_{11} equal to the linked count, that is the number of units with $A = 1$ and $B = 1$. The DSE for the unknown total population size N is then $(N_A \times N_B)/X_{11}$, and can be easily shown to be the method of moments estimator for N under a number of assumptions, a crucial one of which is independent 'capture' events for the same population unit relative to the two data sources.

There have been a variety of suggestions in the literature on reducing the bias that ensues when the two data sources are in fact not independent. However, as the authors emphasise, "Independence is an unverifiable assumption, that is, it cannot be verified from the data used for the estimation of the population size." Consequently, given the available data, all one can do is carry out numerical exercises based on the data at hand to demonstrate sensitivity to failure of this assumption, or carry out studies to investigate bias under simulated conditions. Following Brown et al. (1999, 2006) these authors take the first approach and investigate the sensitivity of the DSE estimates obtained by linking records on the Dutch Population Register with records on a police register. Like Brown et al. (2006), the approach is based on perturbing the odds ratio in a log-linear model for the complete cross classification of the target population, though the methodology presented in the article extends this model to one of Poisson counts and also considers the case where heterogeneous coverage probabilities arise because of covariate information from one or both of the data sources. As one would expect, the higher the achieved coverage, the less sensitive are the DSE-based methods to break down in the independence assumption. This is nicely illustrated in the application described in the article, where a realistic variation in the odds ratio leads to biases in the range -15% to $+9\%$ for the estimated counts of people of either gender and with an Afghan, Iraqi, and Iranian nationality two years previously, compared with biases in the range -42% to $+58\%$ for

the corresponding estimates of people with Polish nationality. As the authors point out, the main reason for this difference is the fact that Dutch and EU law ensure that the overall coverage of the first group by the two data sources is much higher than the corresponding coverage of the second group. However, the fact that such biases can occur is a salutary reminder that failure of model assumptions can have a much more dramatic impact when one is dealing with measurement error than, for example, when one is using regression models for prediction in a 'pure' sample-survey context.

4. Comments on Tuoto and Di Consiglio article

Turning now to the article by Tuoto and Di Consiglio, we see that these authors consider exactly the same situation as that considered by Gerritse et al. but in this case focus on a different measurement-error problem, that of linkage errors when the two data sources are integrated to obtain X_{11}. These authors also use a different nomenclature from that used in Gerritse et al., referring to the DSE estimator as the Petersen estimator, reflecting its origin in estimating the sizes of wild animal populations in the late nineteenth century. As in Gerritse et al., there is an (unspoken) assumption of multinomial sampling throughout, allowing the straightforward development of estimators from moments of unknown quantities. In addition to the definition of L and X_{11}, define $A - L$ as the set of X_{10} population units on A but not on B, that is, X_{10} is the number of records found to be only on list A. Similarly, define $B - L(X_{01})$ to be the set (number) of records found to be only on list B. Then $N_A = X_{11} + X_{10}$ and $N_B = X_{11} + X_{01}$. Under independence and perfect linkage,

$$E(X_{11}) = N \Pr(\text{record in } A) \Pr(\text{record in } B)$$

while

$$E(N_A) = N \Pr(\text{record in } A) = N\tau_1$$

$$E(N_B) = N \Pr(\text{record in } B) = N\tau_2$$

so, using a 'hat' to denote an estimate,

$$\widehat{\Pr}(\text{record in } A) = \hat{\tau}_1 = X_{11}/N_B$$

$$\widehat{\Pr}(\text{record in } B) = \hat{\tau}_2 = X_{11}/N_A$$

and therefore, setting $M = X_{11} + X_{10} + X_{01}$, we have $E(M) = N(\tau_1 + \tau_2 - \tau_1\tau_2)$. The Petersen estimator of N follows by replacing the unknown parameters in this expression by their moment estimates, leading to

$$\hat{N} = M/(\hat{\tau}_1 + \hat{\tau}_2 - \hat{\tau}_1\hat{\tau}_2)$$

It is straightforward to see that this estimator is identical to the DSE defined earlier.

However, the reality in most cases is that there are errors in linking, in the sense that records common to both lists are not matched, as well as matched records that are incorrectly matched. This problem is (partially) addressed by Ding and Fienberg (1994),

who assume incorrect matching is only from A to B. In this context, one can define

$$\alpha = \Pr(\text{correct match}) = \Pr(\text{match is a record from } L)$$

$$\beta = \Pr(\text{incorrect link}|\text{match}) = \Pr(A - L \text{ record matched to } B \text{ record})$$

It follows that

$$\Pr(L \text{ unit linked}) = \alpha \Pr(L \text{ unit}) = \alpha\tau_1\tau_2$$

$$\Pr(A - L \text{ unit linked}) = \beta \Pr(A - L \text{ unit}) = \beta\tau_1(1 - \tau_2)$$

$$\Pr(B - L \text{ unit linked}) = 0$$

and so

$$E(X_{11}) = N(\alpha\tau_1\tau_2 + \beta\tau_1(1 - \tau_2))$$

$$E(X_{10}) = E(N_A) - E(X_{11}) = N(\tau_1 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2))$$

$$E(X_{01}) = E(N_B) - E(X_{11}) = N(\tau_2 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2))$$

Since a population unit that is not on either data set cannot be matched to one that is, it follows that $M = X_{11} + X_{10} + X_{01}$ is the number of unique population units identified in the union of the two data sources, with

$$E(M) = N(\tau_1 + \tau_2 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2))$$

Assuming estimates of α and β are available from the linking process, the Ding and Fienberg estimator of N is the method of moments estimator derived from this identity, with τ_1 and τ_2 replaced by their moment-based estimates, which must then satisfy

$$\hat{\tau}_1 = N_A/\hat{N} = (N_A/M)(\hat{\tau}_1 + \hat{\tau}_2 - (\alpha - \beta)\hat{\tau}_1\hat{\tau}_2 - \beta\hat{\tau}_1)$$

$$\hat{\tau}_2 = N_B/\hat{N} = (N_B/M)(\hat{\tau}_1 + \hat{\tau}_2 - (\alpha - \beta)\hat{\tau}_1\hat{\tau}_2 - \beta\hat{\tau}_1)$$

Solving for $\hat{\tau}_1$ and $\hat{\tau}_2$ based on these identities, we obtain

$$\hat{\tau}_1 = (X_{11} - N_A\beta)/(N_B(\alpha - \beta))$$

and

$$\hat{\tau}_2 = (X_{11} - N_A\beta)/(N_A(\alpha - \beta))$$

It is straightforward to see that in the case of no linkage error, that is $\alpha = 1$ and $\beta = 0$, the Ding and Feinberg estimator defined by $E(M)$ above reduces to the Petersen estimator.

The article by Tuoto and Di Consiglio extends this idea to also allow linkage errors from B to A. In order to do this, these authors assume that the probability of this happening is the same as the probability of incorrect matching from A to B (i.e., β). Then, following the same approach as that underpinning the Ding and Fienberg estimator, it can be seen that

$$E(X_{11}) = N(\alpha\tau_1\tau_2 + \beta\tau_1(1 - \tau_2) + \beta\tau_2(1 - \tau_1))$$

$$E(X_{10}) = E(N_A) - E(X_{11}) = N(\tau_1 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2) - \beta\tau_2(1 - \tau_1))$$

$$E(X_{01}) = E(N_B) - E(X_{11}) = N(\tau_2 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2) - \beta\tau_2(1 - \tau_1)),$$

so collecting terms

$$E(M) = N(\tau_1 + \tau_2 - \alpha\tau_1\tau_2 - \beta\tau_1(1 - \tau_2) - \beta\tau_2(1 - \tau_1))$$

The same argument as used by Ding and Fienberg then leads to

$$\hat{\tau}_1 = N_A/\hat{N} = (\beta M + X_{11}(\beta - 1))/(N_A(2\beta - \alpha))$$

$$\hat{\tau}_2 = N_B/\hat{N} = (\beta M + X_{11}(\beta - 1))/(N_B(2\beta - \alpha)).$$

Substitution of these expressions for $\hat{\tau}_1$ and $\hat{\tau}_2$ into the method of moments estimator of N defined by the preceding expression for $E(M)$ leads to the adjusted estimator for N defined by Expression (13) in the article.

As noted by Tuoto and Di Consiglio, the main advantage of (13) over the standard Ding and Fienberg approach is bias reduction when β is non-negligible. However, this assumes symmetry of incorrect matching between A and B, which is debatable and should be possible to generalise. Also, the approach depends on having access to good estimates of linkage-error probabilities, which can require audit samples. In this context it is important to note that these values of α and β must be such that the estimate $\hat{N}$ of N defined by (13) in the article satisfies the consistency restrictions defined by the Fréchet inequalities,

$$\max(N_A, N_B) \leq \hat{N} \leq \min(N_A, N_B)(\alpha\hat{\tau}_1\hat{\tau}_2 + \beta\hat{\tau}_1(1 - \hat{\tau}_2) + \beta\hat{\tau}_2(1 - \hat{\tau}_1))^{-1}.$$

5. Comments on Zhang article

Finally, I turn to the article by Zhang. This considers another possible source of measurement error when a population size is estimated by linking two or more data sources. In this case the author tackles the situation where two population lists (or registers) are linked in order to estimate the size of a population that is partially captured by each list. The twist is that these lists also include units that are not from the population of interest. In other words, there is both undercoverage as well as overcoverage when the two lists are linked. We can characterise this situation using the schematic below. This shows a target population U of (unknown) size N, partially covered by two linked lists, denoted as usual by A and B. Without loss of generality we denote membership of $A(B)$ by

	$U = 1$		
	$B = 1$	$B = 0$	
$A = 1$	N_{11}	N_{10}	N_A
$A = 0$	N_{01}	N_{00}	$N - N_A$
	N_B	$N - N_B$	N
	$U = 0$		
	$B = 1$	$B = 0$	
$A = 1$	K_{11}	K_{10}	K_A
$A = 0$	K_{01}	0	$K - K_A$
	K_B	$K - K_B$	K

the binary event $A = 1$ $(B = 1)$. Similarly, membership of U is denoted by the binary event $U = 1$.

The author refers to the set of $N + K$ units covered by this schematic as the target-list universe and assumes an underlying multinomial distribution for the cell counts defining it. Note the structural zero for the (000) cell, since the target-list universe cannot contain such units. The author also assumes

- An independent coverage survey with only undercoverage error (all surveyed units are $U = 1$) but with unknown target population coverage. That is,

$$\pi = \Pr\left(\text{unit in } U \text{ included in sample}\right)$$

  is unknown. This will be the case if the framework used to select the sample for the coverage survey is a subset of U.

- Perfect linking of A and B as well as linking of coverage survey units to A and B. Consequently, $X_{11} = N_{11} + K_{11}$, $X_{01} = N_{01} + K_{01}$ and $X_{10} = N_{10} + K_{10}$ are known, as is the corresponding breakdown of the survey counts, which we denote n_{11}, n_{10}, n_{01} and n_{00}, with the usual interpretation.

Note that there is no assumption of independence between A and B. The aim is to use these data to estimate N.

Let τ_{jk} denote the conditional probability that a randomly sampled unit from the target-list universe has $A = j$ and $B = k$ given that it is a member of the target population, that is, has $U = 1$. Then, under the assumed multinomial model for the target-list universe, the linked list counts satisfy $E(N_{jk}) = X_{jk}\tau_{jk}$, and for the corresponding linked sample counts, $E(n_{jk}) = X_{jk}\tau_{jk}\pi$, with

$$E(n_{00}) = E(N)\pi - E(n_{11}) - E(n_{10}) - E(n_{01})$$

Unfortunately, without knowing the value of π, the equation for $E(n_{00})$ above shows that the available data are insufficient to identify N given the assumed multinomial model for the target-list universe. Another identifying assumption is needed. In the article, the author uses a log-linear model characterisation of the problem to investigate alternative approaches to resolving this identification problem, with the most promising of these based on a 'pseudo-independence' assumption for the list universe defined by the union of A and B. This is where the probability of a nontarget population unit in this universe being linked is the product of the corresponding probabilities of a nontarget population unit being on either list, see Equation (11) in the article. The author argues that this assumption is reasonable when the lists are of high quality, that is, there are few target population units missed by them, and derives the method of moments estimators of these probabilities, see Equation (13). The corresponding method of moments estimator of N then follows from standard arguments.

6. Some Concluding Observations

From the perspective of a commentator, all four articles reviewed above have a common focus. They all consider problems that arise when situations corresponding to nonstandard measurement error scenarios arise in official statistics. The way they tackle these problems

is different. The first two articles, by Burger et al. and Gerritse et al., use sensitivity analysis and simulation to illustrate the extent of the problem when standard statistical methods (which ignore the measurement error) are used. As we see, their findings are sobering. The glass is definitely half empty. The articles by Tuoto and Di Consiglio and by Zhang are more along the lines of the glass being half full. Both focus on remedial action, extending the models underpinning the standard methods to accommodate the measurement error. Their results are encouraging, in the sense that they show that these errors can be dealt with in a systematic way. However, they are far from being the final word on the matter. Both tackle the estimation problem, but leave the (hard!) inference problem for later. The reason for this is clear – unlike the well-known sample error structure that is implicit in conventional official statistics, modern official statistics is increasingly eschewing sampling or minimising the use of (expensive) samples, instead using a variety of linking and combining techniques to create what is hopefully something like a 'census' of the population of interest. As these authors clearly demonstrate, this can be a fool's paradise. The errors implicit in linking (or even more importantly, nonlinking), as well as misspecification errors in the implicit models underpinning the estimates derived from these data, can be considerable. The four articles in this issue that I have commented on here represent significant steps towards development of a methodological framework for inference in such situations. It is quite obvious that such a framework will depend on modelling assumptions, so the classical design-based inference paradigm that has for so long served so well in official statistics is irrelevant. What we see here is evidence that the model-based inference paradigm for official statistics that is taking its place needs to be applied with a strong dose of common sense, and a good knowledge of the frailties of the models used. The insurance provided by design-controlled randomisation is no longer available.

7. References

Brown, J.J., O. Abbott, and I.D. Diamond. 2006. "Dependence in the 2001 One-Number Census Project." *Journal of the Royal Statistical Society Series (Statistics in Society)* 169: 883–902. Doi: http://dx.doi.org/10.1111/j.1467-985X.2006.00431.x

Brown, J.J., I.D. Diamond, R.L. Chambers, L.J. Buckner, and A.D. Teague. 1999. "A Methodological Strategy for a One-Number Census in the UK." *Journal of the Royal Statistical Society Series A* 162: 247–267.

Chambers, R. 2009. "Regression Analysis of Probability-Linked Data." *Official Statistics Research Series*. Available at: http://www.statisphere.govt.nz/official-statistics-research/series/vol-4.htm.

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

Kim, G., and R. Chambers. 2012. "Regression Analysis Under Incomplete Linkage." *Computational Statistics and Data Analysis* 56: 2756–2770.

Neter, J., E.S. Maynes, and R. Ramanathan. 1965. "The Effect of Mismatching on the Measurement of Response Error." *Journal of the American Statistical Association* 60: 1005–1027.