Ask the Experts

Wider applications for dual and multiple system estimation

Peter G. M. van der Heijden[1] **and Maarten Cruyff**[2]

[1]Utrecht University, the Netherlands, and University of Southampton, UK,
p.g.m.vanderheijden@uu.nl
[2]Utrecht University, the Netherlands, m.cruyff@uu.nl

Abstract

Dual and multiple system estimation can be usefully applied in much more general settings than usually considered. We consider settings such as making use of covariates that are not available in all of the lists, and lists that cover different (but overlapping) periods in time, lists that cover different (but overlapping) regions, and lists that cover different (but overlapping) age ranges.

Keywords: Dual system estimation; multiple system estimation; covariates; capture-recapture

1 Introduction

We want to show here that dual and multiple system estimation can be applied in much more general settings than usually considered.

In dual system estimation the aim is to estimate a population size. See Table 1. Two lists of individuals, say lists A and B, are linked. Being in list A (B) is denoted by 1, and not being in list A (B) by 0. There are 259 individuals in A and in B, 539 individuals in A but not in B, and 91 individuals in B but not in A. The individuals not in A and in B are missed by both lists, and the aim is to estimate their number.

Important assumptions are perfect linkage of the lists, a homogeneous probability to be included in at least one of the lists (for the other list the inclusion probabilities may be heterogenous over individuals), and independence between the inclusion probabilities of A and B. Due to the assumption of independence, the estimated missing count for the cell (0,0) is 539 * 91 / 259. The independence model in dual system estimation can also be denoted as a loglinear independence model. This has the advantage that generalizations of dual system estimation to situations where background characteristics of individuals are taken into account, and more than two lists are used, are easily described. We will denote loglinear models by placing the variables that constitute the higher order margins between square brackets. So here the independence model is denoted by $[A][B]$.

Table 1: *Contingency table after linking list A and list B*

	$B=1$	$B=0$
$A=1$	259	539
$A=0$	91	0

Dual system estimation plays a role in the census, where the lists are the census survey (A) and the so-called census coverage survey (B). In a census context the overlap in cell (1,1) is relatively large in comparison to (1,0) and (0,1), and thus the estimate in cell (0,0) is relatively small. Therefore a violation of the independence assumption will have only a minor effect on the population size (compare Gerritsen et al., 2015a, where the resulting bias is quantified). But in other contexts this is not necessarily the case.

As said, dual system estimation assumes independence of the inclusion probability of list A and of list B. This can be easily violated. It may be, for example, that in both lists there is heterogeneity of inclusion probabilities in the sense that some individuals have lower probabilities to be on both lists and other individuals have higher probabilities. Think of young men, that are harder to include in both the census survey as well as in the census coverage survey. Two important ways to deal with such violations are (i) using covariates, and (ii) using more than one list.

For covariates one can think of variables such as age and gender. Then the assumption becomes that inclusion in list A is independent of inclusion in list B for each combination of gender (denoted by, for example, X_1) and age (X_2) separately. In terms of loglinear models this would mean that we would have to fit the loglinear model $[AX_1X_2][BX_1X_2]$. Using this loglinear model would solve the young men problem that we just discussed.

For a third list one may think of using, for example, a police register C with apprehensions. This would mean that we have a $2\times2\times2$ table with one missing cell. Thus there are seven counts, and a loglinear model with seven parameters may be fit: $[AB][AC][BC]$. In this model the independence assumption is replaced by the assumption that there is no three factor interaction. This would mean that the odds ratio between the census survey and the census coverage survey is identical for individuals included in the police register and individuals not included in the police register. This is a much less demanding assumption than independence.

These results are well know and described in detail in, for example, Bishop, Fienberg and Holland (1975), and the International Working Group on Disease Monitoring and Forecasting (1995). We now move to new grounds, namely to settings where the covariates are not available in each list, and to settings where the lists cover populations that only partly overlap.

2 Missing covariates

Consider dual system estimation. When two lists are linked, covariates that are present in only one of the lists will be missing for individuals that are not on that list. Consider Panel 1 in Table 2 (see Van der Heijden et al., 2012, 2018, for details). List A is part of the population register, with individuals born in Iran, Irak or Afghanistan, and B is the corresponding part of the police register. Marital status (denoted as X_1) may be a covariate of interest but it is only collected in the population register A. Hence it is missing for those individuals only in B. Police region where apprehended, covariate X_2, is only collected in the police register B and therefore missing in A. For the individuals both

Table 2: *Covariate X_1 (Marital status) is only observed in population register A and X_2 (Police region where apprehended) is only observed in police register B*

Panel 1: Observed counts of All individuals

		$B=1$		$B=0$
		$X_2=0$	$X_2=1$	X_2 missing
$A=1$	$X_1=0$	259	539	13,898
	$X_1=1$	110	177	12,356
$A=0$	X_1 missing	91	164	-

Panel 2: Fitted values under $[AX_2][X_1X_2][BX_1]$

		$B=1$		$B=0$	
		$X_2=0$	$X_2=1$	$X_2=0$	$X_2=1$
$A=1$	$X_1=0$	259.0	539.0	4,510.8	9,387.2
	$X_1=1$	110.0	177.0	4,735.8	7,620.3
$A=0$	$X_1=0$	63.9	123.5	1,112.4	2,150.2
	$X_1=1$	27.1	40.5	1,167.9	1,745.4

in A and B the relation between marital status and police region where apprehended is know, see the upper left four cells. We would like to estimate the missing values on marital status and police region where apprehended for those individuals that are in only one of the registers. Panel 2 provides these estimates. The estimates are produced using the EM algorithm assuming missing at random assumption. The model used is $[AX_2][X_1X_2][BX_1]$. The model has 8 parameters, and where 8 is also the number of counts in the table in Panel 1. Due to the term X_1X_2 the odds ratio between X_1 and X_2 in Panel 1 is projected to obtain estimates for the missing data. Notice that, for example, 4,510.8 + 9,387.2 = 13,898, so the 13,898 are spread out over the two cells.

A more complicated application concerns Polish individuals in the Dutch population register, see Table 3 taken from Gerritse et al. (2015b). We want to know the number of individuals born in Poland being in the Netherlands. For the census, it is important to split this number up in individuals having usual residence and those who have not. There are three lists, namely the population register, the employment register and the police register, hence we have a triple system estimation problem. For the first two lists the covariate usual residence can be derived, but for the crime suspects register this information is missing. Hence there are two missing cells, for which we know the sum: 1,043. This number is the number of individuals that are only in the crime suspects register. Using a loglinear model in the EM algorithm this number is spread out over the two missing cells, and subsequently for the two (no, no, no) cells an estimate can be found that completes the population size estimation problem.

These two examples illustrate that it is possible to use covariates that are not available in each of the registers. From a substantive point of view this is important as it provides population size estimates broken down over the levels of the covariate, and this provides further insight into the constitution of the population. From a statistical point of view van der Heijden et al. (2018) show that, when the missing at random assumption holds, making use of the covariates lead to estimates with good properties in terms of RMSE.

Table 3: *Polish individuals by the population register, the employment register and the crime suspects register, by usual residence. The counts for the two cells labeled "missing" add up to 1,043.*

			Crime suspects	
Usual residence	Population	Employment	Yes	No
No	Yes	Yes	32	3,523
		No	34	3,225
	No	Yes	149	60,190
		No	missing	0
Yes	Yes	Yes	183	21,309
		No	195	14,052
	No	Yes	81	20,216
		No	missing	0

3 Different but overlapping populations

Dual and multiple system estimation can also be used when lists refer to different but overlapping populations. The key is to consider it as a missing data problem.

As a first example, assume that the lists cover different time periods. In Zwane et al. (2004) we study the size of the population of babies having Spina Bifida. There are six lists and the lists cover different time periods. For the lists not covering the full time period, the missing entries in the contingency table are estimated with EM, assuming missing at random. This latter assumption means that relations between lists found in the full time periode are projected to the time period where some of the lists are missing, an assumption that is plausible.

One may also think of one list covering the north and the middle part of a country and another list the middle and south part. Using EM one can estimate the counts for the south part for the first list, and for the north part for the second list. The relation between both lists in the middle part is used for such projections.

As a last example, one can also think of lists covering different age ranges. For example, one could have a census list and a driving licence list. Here the young will not have a driving licence, but the (non-)overlap between the census and the driving license list can projected to this young age group.

4 Conclusion

We hope to have shown that dual and multiple system estimation can be applied in wider context than is usually considered. We would welcome to see such applications and are happy to advice!

References

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W (1975). *Discrete Multivariate Analysis, Theory and Practice*. McGraw-Hill, New York. doi: 10.1007/ 978-0-387-72806-3.

Gerritse, S. C., van der Heijden, P. G. M., and Bakker, B. F. M. (2015a). Sensitivity of population size estimation for violating parametric assumptions in loglinear models. *Journal of Official Statistics*, **31**(3), 357-379. doi=10.1515/jos-2015-0022.

Gerritse, S. C. , Bakker, B. F. M., and van der Heijden, P. G. M. (2015b). Different methods to complete datasets used for capturerecapure estimation: estimating the number of usual residents in the Netherlands. *Statistical Journal of IAOS*, **31**(4), 613-627. doi: 10.3233/SJI-150938.

International Working Group for Disease Monitoring and Forecasting (1995). Capturerecapture and multiple record systems estimation. Part i. History and theoretical development. *American Journal of Epidemiology*, **142**, 1059-1068. doi: 10.1093/oxfordjournals.aje.a117558.

Van der Heijden, P. G. M., Whittaker, J. Cruyff, M., Bakker, B. F. M., and Van der Vliet, R. (2012). People born in the middle east but residing in the netherlands: Invariant population size estimates and the role of active and passive covariates. *The Annals of Applied Statistics*, **6**(3), 831-852. doi: 10.1214/12-AOAS536.

Van der Heijden, P. G. M., Smith, P. A., Cruyff, M. and Bakker, B. F. M. (2018). An overview of population size estimation where linking registers results in incomplete covariates, with an application to mode of transport of serious road casualties. *Journal of Official Statistics*, **34**(1), 239-263.

Zwane, E., Van der Pal-de Bruin, K., and Van der Heijden, P. G. M. (2004). The multiple-record systems estimator when registrations refer to different but overlapping populations. *Statistics in Medicine*, **23**, 2267-2281. doi: 10.1002/sim.1818.