Robustness of population size estimates against violation of the independence assumption

Susanna C. Gerritse
University Utrecht

Peter G. M. van der Heijden
University Utrecht, University of Southampton

Bart F. M. Bakker
Statistics Netherlands, VU University Amsterdam

An important quality aspect of Censuses is the degree of coverage of the population. When administrative registers are used, undercoverage can be estimated by linking two or more registers and subsequently estimating the number of individuals that is missed by each of the registers. Adding this estimated number to the observed number of individuals yields a population size estimate. This is also known as the capture-recapture method. The standard approach uses loglinear models that most often rely on the assumption that being in the first register is statistically independent from being in the second register. If covariates are available, the independence assumption can be replaced with a less strict independence assumption conditional on covariates. However, both independence and independence conditional on covariates are rarely met. In this paper we will investigate the robustness of the population size estimate under dependence. The results show that violation of the independence assumption can lead to seriously biased estimates.

Key words: Capture-recapture method, population size estimate, sensitivity analysis, Census

1. Introduction

The preparations of the outcomes of the 2011-round of the Census are in progress. An increasing number of countries use administrative data to collect the necessary information. There are countries who are repeating this method such as Denmark, Finland and the Netherlands, and more than ten European countries that are using administrative data for the first time (Valente, 2010). Under Census regulation a quality report is obligatory, and one of the aspects that needs to be addressed is the undercoverage of the Census data. This asks for an estimate of the size of the population.

If one wants to estimate the size of a population, loglinear capture-recapture methods are commonly used (Fienberg, 1972). In countries with a Census based on administrative data, the approach mostly used is to find two registers and treating these as the captured and recaptured data. This method includes linking the individuals in the registers and subsequently estimating the number of individuals missed by both registers.

However, the outcome of the capture-recapture method depends on some assumptions underlying the data. In particular it is assumed that inclusion in the captured is independent of inclusion in the recaptured data (Chao, Tsay, Lin, Shau, & Chao, 2001; Van der Heijden, Whittaker, Cruyff, Bakker, & Van der Vliet, 2012). This independence assumption could easily be violated. Under dependence between registers the inclusion probability of one register is related to the inclusion probability of the other register. Under positive dependence individuals in the captured data have a higher probability of also being in the recaptured data, resulting in an underestimation of the population size estimate. Under negative dependence the opposite holds (Hook & Regal, 1995). These results are well known, but the size of these biases have not been studied before.

Independence is an unverifiable assumption, i.e., it cannot be verified from the data. The loglinear independence model for the linked captured and recaptured data has three parameters whereas there are only three counts. As such the independence model is a maximal or saturated model, because the observed counts are equal to the fitted counts. Thus we cannot assess dependence from the maximal model. The situation of a maximal model also holds when covariates of individuals are taken into account and we operate under the loglinear conditional independence model. Yet we are interested in the size of the impact of mild or severe violations of (conditional) independence on the population size estimate.

In this manuscript we propose a general approach to sensitivity analyses under the loglinear model framework. As far as we know, sensitivity analyses have not been carried out in the context of capture-recapture. Where in the maximal model specific interaction parameters are equal to zero, we impute fixed values departing from zero for these parameters and investigate the impact on the population size estimate.

We proceed as follows. In section two we will discuss the loglinear independence model with two registers and conduct a sensitivity analysis for violation of the independence assumption. In section three we will discuss a two register model including a covariate and conduct a sensitivity analysis. We use two data sources to illustrate the robustness of the capture-recapture method, that have been provided by Statistics Netherlands. The GBA (Gemeentelijke BasisAdministratie) is the official Dutch register containing information on people legally residing in the Netherlands and the HKS (HerkenningsDienst Systeem) is a police register of all suspects of criminal offenses. We refer the reader to Van der Heijden et al. (2012) for more details on the registers.

2. Two registers

The simplest population size estimation model makes use of two registers, 1 and 2. Let variables A and B respectively denote inclusion in registers 1 and 2. Let the levels of A be indexed by i ($i = 0, 1$) where $i = 0$ stands for "not included in register 1", and $i = 1$, stands for "included in register 1". Similarly, let the levels of B be indexed by j ($j = 0, 1$). Expected values are denoted by m_{ij}. Observed values are denoted by n_{ij} with $n_{00} = 0$, because there are no observations for the cases that belong in the population but were not present in either of the registers.

As an illustration we will use two registers of Statistics Netherlands, the GBA and the HKS, on people with an Afghan, Iranian or Iraqi nationality living in the

Netherlands in 2007 (shown in Table, 1a; Van der Heijden et al., 2012), and on people with a Polish nationality living in the Netherlands in 2009 (shown in Table (1b) Van der Heijden, Cruyff, & Van Gils, 2011).

Table 1:: The observed values for the two nationalities

(a) Middle Eastern people in 2007

	HKS	
GBA	1	0
1	1,085	26,254
0	255	-

(b) Polish people in 2009

	HKS	
GBA	1	0
1	374	39,488
0	1,373	-

Independence of A and B implies that the odds ratio $\theta = 1 = m_{00}m_{11}/m_{10}m_{01}$. Rewriting the odds ratio and replacing the expected values by observed values will give us maximum likelihood estimate (mle):

$$\hat{m}_{00} = \frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \frac{n_{10}n_{01}}{n_{11}}. \tag{1}$$

We now consider the case where two registers are dependent instead of independent. For two dependent registers we denote the loglinear model by $\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$, with identifying restrictions $\lambda_1^A = \lambda_1^B = \lambda_{11}^{AB} = \lambda_{10}^{AB} = \lambda_{01}^{AB} = 0$. Then the parameters λ, λ_0^A, λ_0^B and λ_{00}^{AB} are to be estimated. However there are only three observed counts so the fourth parameter λ_{00}^{AB} cannot be estimated. The approach we advocate is to impute fixed values for λ_{00}^{AB} into the model, and to estimate parameters for λ, λ_0^A, λ_0^B.

Let us denote the expected value for cell (0,0) when the odds ratio in the population is θ by $m_{00(\theta)}$. It follows that $\theta = m_{00(\theta)}m_{11}/m_{10}m_{01} = \exp(\lambda_{00}^{AB})$. Replacing expected by fitted and observed values we get:

$$\hat{m}_{00(\theta)} = \theta\frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \theta\frac{n_{10}n_{01}}{n_{11}} = \theta\hat{m}_{00}. \tag{2}$$

Equation (2) shows that $\hat{m}_{00(\theta)}$ can be found simply by multiplying the estimate under independence, $\hat{m}_{00}$, with θ. It is clear that we cannot use (2) to estimate $\hat{m}_{00}$ as θ is unknown. However (2) allows us to study the impact of a violation of the independence assumption as a function of θ. For this purpose, let n be the total of observed cases, $n = n_{01} + n_{10} + n_{11}$, let $\hat{N}$ be the population size estimated under $\theta = 1$, thus $\hat{N} = n + \hat{m}_{00}$, and define $\hat{N}_{(\theta)}$ as the estimated population size under an odds ratio of size θ, where $\hat{N}_{(\theta)} = n + \hat{m}_{00(\theta)} = n + \theta\hat{m}_{00}$. We report the relative bias as $\hat{N}/\hat{N}_{(\theta)}$ that shows the bias expressed as a proportion when the odds ratio in the population is θ.

Under independence between A and B for the Middle Eastern people $\hat{m}_{00} = 6,170.30$, and $\hat{N} = 27,594 + 6,170.30 = 33,764.30$. For the Polish people under independence between A and B $\hat{m}_{00} = 144,965.30$, and $\hat{N} = 41,235 + 144,965.30 = 186,200.30$

To investigate the robustness of $\hat{m}_{00}$ under dependence we vary θ from 0.5

to 2. Table 2 shows $\hat{m}_{00(\theta)}$, the population size estimate $\hat{N}_{(\theta)}$ and the estimated relative bias $\hat{N}/\hat{N}_{(\theta)}$ for both datasets. For the Middle Eastern people, under a dependence of $\theta = 0.5$, $\hat{m}_{00(\theta)}$ is $3,085.15$ cases lower than under independence, and for a dependence of $\theta = 2$, $\hat{m}_{00}$ is $6,170.30$ cases higher than under independence. So for an odds ratio ranging from 0.5 to 2, $\hat{N}$ deviates from $\hat{N}_{(\theta)}$ between 10 and -15 percent respectively. However for the Polish people, $\hat{N}_{(\theta)}$ deviates from $\hat{N}$ between 64 and -44 percent. This is a much larger difference than for the people with a Middle Eastern nationality.

Table 2:: Robustness analysis of the population size estimate for the people residing in the Netherlands in 2007 with a Middle Eastern nationality (upper part) and for people with a Polish nationality in 2009 (lower part).

	Odds Ratio	0.50	0.67	1.00	1.50	2.00
	$\hat{m}_{00(\theta)}$	3,085.15	4,113.53	6,170.30	9,255.45	12,340.60
Middle	$\hat{N}_{(\theta)}$	30,679.15	31,707.53	33,764.30	36,849.45	39,934.60
Eastern	$\hat{N}/\hat{N}_{(\theta)}$	1.10	1.06	1.00	0.92	0.85
	$\hat{m}_{00}$	72,482.65	96,643.53	144,965.30	217,447.90	289,930.60
Polish	$\hat{N}_{(\theta)}$	113,717.60	137,878.50	186,200.30	258,682.90	331,165.60
	$\hat{N}/\hat{N}_{(\theta)}$	1.64	1.35	1.00	0.72	0.56

We conclude that the population size estimate can be fairly accurate as with the Middle Eastern people, whereas for the Polish people the population size estimate is highly inaccurate.

3. Two registers with covariates

If covariates are available, the generally non-feasible independence assumption can be replaced with a less strict independence assumption conditional on covariates. This assumption is less stringent because it can take into account heterogeneous inclusion probabilities over the levels of the covariate. Using covariates has the advantage that we can investigate the size of the population over the levels of the covariate.

Assume we have observed covariate X_1, we then assume independence conditional on X_1. Let $\hat{m}_{ijx}$ denote the fitted values for A, B and X_1, where the levels of X_1 are indexed by x. The loglinear conditional independence model becomes log $m_{ijx} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{jx}^{BX} + \lambda_{ix}^{AX}$, and has the identifying restrictions that every parameter equals zero when i, j or $x = 1$.

When assuming dependence between A and B conditional on X, parameters λ_{ij}^{AB} and λ_{ijx}^{ABX} are free. Again, since we have no way of verifying conditional independence in the data we carry out a sensitivity analysis by imputing a fixed parameter λ_{ij}^{AB} or fixed parameters $(\lambda_{ij}^{AB} + \lambda_{ijx}^{ABX})$ into the model.

For the data in Table 1a covariate X_1 is gender, with $x = 1$ is males and $x = 2$ is females. Under independence conditional on X_1 there are two zero counts for cases not found in either register, both for males and females. Let $n_x = n_{10x} + n_{01x} + n_{11x}$ and $\hat{N}_x = n_x + \hat{m}_{00x}$, where $\hat{m}_{00x}$ is defined as $\hat{m}_{00x} = n_{10x} * n_{01x}/n_{11x}$. Then the

population size estimate when assuming independence between A and B conditional on X_1, for $x = 1, 2$, is $\hat{N} = \sum_{x=1}^{2} n_x + \sum_{x=1}^{2} \hat{m}_{00x}$.

Under dependence between A and B given X, the relation between the odds ratio for x, denoted by θ_x, and the loglinear parameters is:

$$\theta_x = \frac{m_{11x} m_{00x(\theta)}}{m_{10x} m_{01x}} = \exp(\lambda_{00}^{AB} + \lambda_{00x}^{ABX}). \tag{3}$$

When we assume that the dependence for $x = 1$ is identical to the dependence for $x = 2$, then:

$$\theta = \frac{m_{111} m_{001(\theta)}}{m_{101} m_{011}} = \frac{m_{112} m_{002(\theta)}}{m_{102} m_{012}} = \exp(\lambda_{00}^{AB}). \tag{4}$$

Under dependence of size θ, $\hat{m}_{00x(\theta)}$ becomes:

$$\hat{m}_{00x(\theta)} = \theta \frac{n_{10x} n_{01x}}{n_{11x}} = \theta \hat{m}_{00x}, \tag{5}$$

and thus the population estimate under dependence becomes

$$\hat{N} = \sum_{x=1}^{2} n_x + \theta * \sum_{x=1}^{2} \hat{m}_{00x}. \tag{6}$$

The sensitivity analysis for the two nationalities under an independence conditional on covariates is shown in Table 3. Under conditional independence the number of missed individuals is 5,696.14 for the Middle Eastern people and 101,00.00 for the Polish people. The population size estimations are 33,299.14 and 142,240 respectively. Note that, conditional independence does not imply a marginal independence, since $1,085 * 5,696.14/26,254 * 255 = 0.92$. The upper part of Table 3 shows the results of the robustness analysis for the Middle Eastern people, whereas the lower part of Table 3 shows the Polish data. The model for the Middle Eastern people is conditional on gender and the model for the Polish people is conditional on gender, age and residential area.

For the Middle Eastern people under a relatively large dependence, ranging from $\theta = 0.5$ tot $\theta = 2$, there will only be a 9 percent overestimation and a 15 percent underestimation respectively compared to $\hat{N}_{(\theta)}$ when $\theta = 1$. For the Polish people with an odds ratio ranging from 0.5 to 2, $\hat{N}$ deviates from $\hat{N}_{(\theta)}$ between an overestimation of 55 and an underestimation of 42 percent. Again, this is a much larger difference than for the people with a Middle Eastern nationality. Thus dependence can seriously bias the population size estimator, even though in some cases a high dependence can give a fairly accurate population size estimate.

4. Conclusion

We have shown for two different datasets that the population size estimate under dependence could be fairly robust as well as not robust at all. The difference in robustness for both datasets can be explained as follows. Under independence $\hat{m}_{00} = n_{01} n_{10} / n_{11}$. This shows that an important factor in the size of $\hat{m}_{00}$ is the size of n_{11}: the smaller the overlap n_{11} in the two registers, the larger the probability of being missed by both registers. Under dependence $\hat{m}_{00}$ is simply multiplied with θ,

Table 3:: Robustness analysis for the people with a Middle Eastern nationality residing in the Netherlands in 2007 (upper part), and the people with a Polish nationality residing in the Netherlands in 2009 (lower part).

	Odds Ratio	0.50	0.67	1.00	1.50	2.00
	$\hat{m}_{00}$	2,848.05	3,797.40	5,696.14	8,544.15	11,392.2
Middle	$\hat{N}_{(\theta)}$	30,450.05	31,400.40	33,299.14	36,147.15	38,995.20
Eastern	$\hat{N}/\hat{N}_{(\theta)}$	1.09	1.06	1.00	0.92	0.85
	$\hat{m}_{00}$	50,502.50	67,336.67	101,005.00	151,507.50	202,010.00
Polish	$\hat{N}_{(\theta)}$	91,737.50	108,571.70	142,240.00	192,742.50	243,245.00
	$\hat{N}/\hat{N}_{(\theta)}$	1.55	1.31	1.00	0.74	0.58

thus when the overlap n_{11} is smaller and hence the missed part of the population $\hat{m}_{00}$ is larger, the population size estimator will be less robust. To illustrate for the Middle Eastern people $1,085/33,764.3 = 0.032$, so 3.2 percent of the population is in the overlap and for the Polish people $374/186,200 = 0.002$, so only 0.2 percent of the population is in the overlap. This reflects that the Polish people are, much more so than for people from the Middle Eastern countries under study, in the position that they work on a temporary basis without living permanently in The Netherlands. By law, it is approved for people from European Union countries like Poland to work without a working and living permit. This is not the case with the Middle Eastern countries. Therefore, the frequency table differs between both nationalities which gives a relatively high estimation of the missed population of the Polish people compared to the Middle Eastern countries.

References

Chao, A., Tsay, P. K., Lin, S.-H., Shau, W.-Y., & Chao, D.-Y. (2001). Tutorial in biostatistics. the application of capture-recapture models of epidemiological data. *Statistics in Medicine*, *20*, 3123 - 3157.

Fienberg, S. E. (1972). The multiple recapture census for closed populations and incomplete 2k contingency tables. *Biometrika*, *59*, 409 - 439.

Hook, E. B., & Regal, R. R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews*, *17*, 243 - 264.

Valente, P. (2010). Main results of the unece / unsd survey on the 2010 / 2011 round of censuses in the unece region. *Eurostat*.

Van der Heijden, P. G. M., Cruyff, M. J. L. F., & Van Gils, G. (2011). Aantallen geregistreerde en niet-geregistreerde burgers uit MOE-landen die in Nederland verblijven. Rapportage schattingen 2008 en 2009. The number of registered and non-registered citizens from MOE-countries residing in the Netherlands. Reporting estimations 2008 and 2009. *The Hague, Ministry of Social Affairs and Employment*.

Van der Heijden, P. G. M., Whittaker, J., Cruyff, M. J. L. F., Bakker, B. F. M., & Van der Vliet, H. N. (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*, 831 852.