DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 31, No. 3, 2015, pp. 357–379, http://dx.doi.org/10.1515/JOS-2015-0022

Sensitivity of Population Size Estimation for Violating Parametric Assumptions in Log-linear Models

Susanna C. Gerritse[1], *Peter G.M. van der Heijden*[2], *and Bart F.M. Bakker*[3]

An important quality aspect of censuses is the degree of coverage of the population. When administrative registers are available undercoverage can be estimated via capture-recapture methodology. The standard approach uses the log-linear model that relies on the assumption that being in the first register is independent of being in the second register. In models using covariates, this assumption of independence is relaxed into independence conditional on covariates. In this article we describe, in a general setting, how sensitivity analyses can be carried out to assess the robustness of the population size estimate. We make use of log-linear Poisson regression using an offset, to simulate departure from the model. This approach can be extended to the case where we have covariates observed in both registers, and to a model with covariates observed in only one register. The robustness of the population size estimate is a function of implied coverage: as implied coverage is low the robustness is low. We conclude that it is important for researchers to investigate and report the estimated robustness of their population size estimate for quality reasons. Extensions are made to log-linear modeling in case of more than two registers and the multiplier method.

Key words: Capture-Recapture methodology; dual-system estimation; sensitivity analysis; census; Poisson log-linear regression.

1. Introduction

For the Census of 2011, an increasing number of countries used administrative data to collect the necessary information. Under census regulations 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, capture-recapture methods, making use of log-linear models, are commonly used (Fienberg 1972; Bishop et al. 1975; Cormack 1989; International Working Group for Disease Monitoring and Forecasting 1995). These methods go by different names, such as mark-recapture methods, dual-system methods or dual-record system methods. In this

[1] Utrecht University, Methods and Statistics, Padualaan l4, Utrecht 3584 CH, The Netherlands and University of Southampton, UK. Email: sc.gerritse@gmail.com
[2] Utrecht University, Methods and Statistics, Padualaan l4, Utrecht 3584 CH, The Netherlands and University of Southampton, UK. Email: P.G.M.vanderheijden@uu.nl
[3] Statistics Netherlands, Methodology, P.O.Box 24500, 2490 HA, The Hague, The Netherlands and VU University, Netherlands. Email: bfm.bakker@cbs.nl
Acknowledgments: An earlier version of this article was presented at the 59th World Statistics Congress, 25 – 30 August 2013. Rik van der Vliet and Peter-Paul de Wolf for their valuable comments on earlier drafts of this article. Olav Laudy for his help in writing the SPSS routine and Raymond Chambers for pointing out the link to the work of James Brown. We would also like to express our gratitude for the valuable comments of our reviewers and editors.

article we use the label capture-recapture. In countries with a traditional census a postenumeration survey could be organised to collect recaptured data, as was the case for instance in the United Kingdom (Brown et al. 1999; ONS 2012), and in the U.S. (Wolter 1986; Bell 1993; Nirel and Glickman 2009). In this case, a survey with a relatively small sample size is linked to the census data. In countries with a census based on administrative data, the approach used most is to find two registers and treating these as the captured and recaptured data. The 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 heavily on some assumptions underlying the data. In particular, if two sources are used, it is usually assumed that inclusion in the captured data is independent of inclusion in the recaptured data. A second assumption deals with homogeneity versus heterogeneity of inclusion probabilities. If there is one source of heterogeneity it is assumed that at least for one of the two sources the inclusion probabilities are homogeneous (Chao et al. 2001; Zwane and Van der Heijden 2004). If there are two sources of heterogeneity (two covariates), the estimates are unbiased if the inclusion probabilities of the first source vary with one source of heterogeneity, and the inclusion probabilities of the second source vary with a second source of heterogeneity, but the two sources of heterogeneity are statistically independent (Seber 1982, 86). The remaining two assumptions are that the population is closed and that the registers are perfectly linked.

The assumption of independence between two registers is very strict and can easily be violated. Under dependence between registers, the inclusion probability of one register is related to the inclusion probability of the other register. Then, 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. Additionally, under negative dependence the opposite holds (Hook and Regal 1995).

Independence is an unverifiable assumption, that is, it cannot be verified from the data used for the estimation of the population size. The log-linear independence model for the linked captured and recaptured data has three parameters, whereas there are only three counts. Because the observed counts are equal to fitted counts, the independence model is the saturated model (compare van der Heijden et al. 2012). Thus we cannot assess dependence from the saturated model. One way of reducing the impact of the strict independence assumption is to replace it with the lesser strict assumption of independence conditional on covariates. Adding covariates enables us to reduce heterogeneity introduced to the model due to the specific covariate, adjusting the population size estimate for the better. The situation of a saturated model also holds when covariates of individuals are taken into account and we operate under the log-linear conditional independence model. However, we are interested in what the impact of mild or severe violations of (conditional) independence is on the population size estimate. It does not necessarily have to be the case that violation of the (conditional) independence assumption results in a substantive bias in the population size estimate. It is of important to also assess what happens when the other assumptions are violated. However, looking at all assumptions at once is very complex. In this article, we will thus focus on the violation of the independence assumption, assuming all other assumptions to be met.

We propose a general approach to sensitivity analyses under the log-linear model framework using a log-linear Poisson regression, a special case of the generalized linear

model. Where in the saturated model specific interaction parameters are equal to zero, we impute fixed values departing from zero for these parameters, thus simulating dependence, and investigate the impact on the population size estimate. As the log-linear interaction parameters are closely related to the (conditional) odds ratio, there is a clear interpretation for the values to which we fix the parameters.

Similar findings come from the research of Brown et al. (1999), where the census was linked to a Post Enumeration Survey to assess under- and overcoverage (cf. also Wolter 1986; Bell 1993). Brown et al. (1999) used a fixed odds ratio of 0.1 and 10 to investigate the impact of simulated dependence on the population size estimate. They showed that fixed dependence can seriously bias the population size estimate under the independence assumption. Results like these are valuable, since they give insight to the size of the impact of violated independence. However, research into the robustness of the population size estimator under violation of independence is non standard. As far as we know, other research on the impact of the violation of independence involves simulation studies, an already known population size estimate or uses multiple sources (Wolter 1986; Bell 1993; Cormack et al. 2000; Hook and Regal, 1992, 1997, 2000; Brown et al. 2006; Baffour et al. 2013).

We extend the results of Brown et al. (1999) by, instead of using the standard log-linear model, working under a log-linear Poisson regression where we simulate a fixed dependence using offsets. In simulating dependence by adding a fixed offset value to the log-linear model, we can compare the population size estimate under independence to the population size estimate under a 'true' dependence. Additionally we extend our two-register independence model to the case with covariates observed in both registers (fully observed covariates) and covariates observed in only one register (partially observed covariates).

Partially observed covariates are usually ignored because including them would lead to missing values in the other register. However, ignoring these covariates when they actually are related to the inclusion probability of the register results in a biased population size estimate (Zwane and van der Heijden 2007). In assuming missing at random (MAR) we can impute the missing values of the partially observed covariate in the other register and use this covariate to replace the strict independence assumption with independence conditional on covariates. For partially observed covariates the log-linear model is easily extendable, so that we can also conduct sensitivity analyses in this context.

We proceed as follows. In section 2 we will discuss the log-linear model for a capture-recapture model with two registers without covariates. In Section 3 we will discuss a two-register capture-recapture model and conduct a sensitivity analysis on two registers with a conditional independence. In Section 4 the independence assumption will be conditional on partially observed covariates, where a covariate has been observed in only one register. Here the sensitivity analysis is on the dependence of the partially observed covariate on the register, thus whether the covariate influences the inclusion probability of the register. Section 5 provides some extensions made to a specific model, namely for models for three registers, the multiplier method and confidence intervals.

We use two data sources to illustrate the robustness of capture-recapture methodology, which have been provided by Statistics Netherlands. We chose not to make a simulation study because researchers in the field of capture-recapture use real data and we wanted to make the impact of a possible dependence relevant to such researchers. The first data

source is the GBA (Gemeentelijke Basisadministratie) which is the official Dutch Population Register containing demographic information on the 'de jure' population. The 'de jure' population differs from the 'de facto' population, the latter also containing residents who have immigrated from other countries of the European Union and did not register as such, immigrants who (are planning to) stay less than four months and illegal immigrants. An important part of the difference between the 'de jure' and the 'de facto' population is the group of temporary workers from eastern Europe, in particular Poland. The second data source is the HKS (Herkenningsdienst systeem), which is a police register of all persons suspected of known offenses. We refer the reader to van der Heijden et al. (2012) for more details on the registers.

2. Two Registers Without Covariates

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 to the population but were not present in either of the registers.

Recall that one of the assumptions in population size estimation is that the probability of being in the first register is independent of the probability of being in the second register. Under independence, the log-linear model for the counts n_{01}, n_{10} and n_{11} is:

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B \tag{1}$$

where we used the identifying restrictions $\lambda_0^A = \lambda_0^B = 0$. There are two ways to derive the estimate of the missed part of the population. First, by $\hat{m}_{00} = \exp(\hat{\lambda})$, and second, by using the property that the odds ratio under independence is 1, that is, $m_{00}m_{11}/m_{10}m_{01} = 1$ so that:

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

For the first way of estimating the missed portion of the population we need an estimate of λ in (1). There are several ways to estimate the parameters in (1), and it suits our purposes later on to use the generalized linear model. We assume that n_{ij} follow a Poisson distribution; a log link connects the expected values m_{ij} to the linear predictor. In terms of matrices and vectors we get

$$\log \begin{pmatrix} m_{11} \\ m_{10} \\ m_{01} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda_1^A \\ \lambda_1^B \end{pmatrix} \tag{3}$$

where the right-hand side of (3) leads to a vector with elements $\left[\lambda + \lambda_1^A + \lambda_1^B, \lambda + \lambda_1^A, \lambda + \lambda_1^B\right]$. Thus the estimates of λ, λ_1^A and λ_1^B will get us estimates

$\hat{m}_{11}, \hat{m}_{10}$ and $\hat{m}_{01}$ of which also the missed portion of the population $\hat{m}_{00}$ is found by log $(\hat{m}_{00}) = \hat{\lambda}$, so that $\hat{m}_{00} = \exp(\hat{\lambda})$.

However, the problem with using the independence model is that independence is an unverifiable assumption, that is, we can not verify independence from the data. Thus the Poisson log-linear model for independence works under the assumption that the interaction parameter $\lambda_{ij}^{AB} = 0$. As noted before, this assumption could be violated and the population size estimate under independence may well be inaccurate. We are interested in what happens to the population size estimate when we assume independence when actually the inclusion probabilities of inclusion in registers 1 and 2 are dependent.

The approach we advocate is to include a fixed interaction parameter $\tilde{\lambda}_{ij}^{AB}$ in the model, where the tilde indicates that the interaction parameter is not estimated but fixed. By choosing interesting values for $\tilde{\lambda}_{ij}^{AB}$ we can conduct a sensitivity analysis on the population size estimate. The log-linear model then becomes:

$$\log m_{ij} = \lambda + \lambda_i^A + \lambda_j^B + \tilde{\lambda}_{ij}^{AB} \tag{4}$$

where we used the identifying restrictions $\tilde{\lambda}_{00}^{AB} = \tilde{\lambda}_{10}^{AB} = \tilde{\lambda}_{01}^{AB} = 0$. In matrix terms we get:

$$\log \begin{pmatrix} m_{11} \\ m_{10} \\ m_{01} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda_1^A \\ \lambda_1^B \\ \tilde{\lambda}_{11}^{AB} \end{pmatrix} \tag{5}$$

The log-linear model for independence is a special case of this saturated model when $\lambda_{ij}^{AB} = \tilde{\lambda}_{ij}^{AB} = 0$. Dependence can be introduced to log-linear models by fixing $\tilde{\lambda}_{ij}^{AB}$ to anything but 0. In software for Poisson regression, Model (4) and (5) can be fit by entering $\tilde{\lambda}_{ij}^{AB}$ as a so-called offset. When $\tilde{\lambda}_{ij}^{AB} \neq 0$, $\hat{\lambda}$ in (5) differs from $\hat{\lambda}$ in (3).

Note that interesting values for $\tilde{\lambda}_{ij}^{AB}$ can be chosen using a direct relationship between λ_{ij}^{AB} and the odds ratio θ, which is:

$$\theta = \frac{m_{11} m_{00}}{m_{10} m_{01}} = \exp \tilde{\lambda}_{11}^{AB}. \tag{6}$$

Using the Poisson log-linear model with an offset is a general approach for carrying out a sensitivity analysis. The approach is general in the sense that it can be applied in more complicated log-linear models, for example when it is desirable to investigate violations of more than one assumption simultaneously (cf. the models discussed in Subsection 4.2). For completeness we also discuss a second method that is simpler but less general.

The second way of estimating the missed portion of the population is by using odds ratios directly, as has been done in Brown et al. (1999). We show this second way to give a full overview of the method. This also provides for simpler notation, which we will use in the rest of the article. Under independence, the odds ratio $m_{11} m_{00} / m_{10} m_{01} = 1$, and by rewriting and replacing the expected values with observed values, we get maximum

likelihood estimate (2). We can impute dependence by making the odds ratio $\theta \neq 1$. Thus $\theta = m_{11}m_{00}/m_{10}m_{01}$, and

$$\hat{m}_{00(\theta)} = \theta\frac{\hat{m}_{10}\hat{m}_{01}}{\hat{m}_{11}} = \theta\frac{n_{10}n_{01}}{n_{11}} = \theta\hat{m}_{00}. \tag{7}$$

Note that $\hat{m}_{00(\theta)}$ can be found simply by multiplying the estimate under independence, $\hat{m}_{00}$, with θ. Both approaches, the log-linear Poisson regression with an offset and the odds ratio, yield the same $\hat{m}_{00}$. We will use the odds ratio to denote dependence as it provides a simpler notation than the interaction parameter $\tilde{\lambda}_{ij}^{AB}$.

The methods just described allow us to study the impact of a violation of the independence assumption as a function of θ. To get the population size estimate, 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 dependence of size θ, $\hat{N}_{(\theta)} = n + \hat{m}_{00(\theta)} = n + \theta\hat{m}_{00}$. It follows that under negative dependence (i.e., $\theta < 1$), $\hat{N}$ will be an overestimation compared to $\hat{N}_{(\theta)}$, and under a positive dependence (i.e., $\theta > 1$), $\hat{N}$ will be an underestimation compared to $\hat{N}_{(\theta)}$. The bias will be smaller the closer θ is to 1.

Assume that Register 1 has a better coverage of the population than Register 2. Then when $n_{11}/(n_{11} + n_{01})$ is high the observed coverage is high, and vice versa. Brown et al. (2006) showed that as the observed coverage increases, the number of individuals that are missed by Register 1 reduces and $n_{11}/n_{10}n_{01}$ increases so that $n_{10}n_{01}/n_{11} = \hat{m}_{00}$ decreases. Then, the implied coverage of Register 1 is high, so that $\hat{m}_{00}$ is reasonably robust to dependence. When the observed coverage decreases, the number of individuals missed by Register 1 increases and $n_{11}/n_{10}n_{01}$ decreases. Then the implied coverage of Register 1 will be low, so that $\hat{m}_{00}$ is less robust to dependence.

To illustrate, we use two registers of Statistics Netherlands, the GBA and the HKS, on people with Afghan, Iranian, or Iraqi (AII) nationality living in the Netherlands in 2007 (shown in Table 1; van der Heijden et al. 2012), and on people with a Polish nationality living in the Netherlands in 2009 (shown in Table 1; van der Heijden et al. 2011).

For the people with Afghan, Iraqi, and Iranian nationality $\hat{m}_{00} = 6,\ 170$ under independence between the registers GBA and HKS. The population size estimated under $\theta = 1$ becomes $\hat{N} = 27,594 + 6,170 = 33,764$. Then, under dependence between the registers GBA and HKS the estimated population size becomes $\hat{N}_{(\theta)} = 27,594 + (\theta^{*}6,170)$, see (7).

To investigate the robustness of the estimate under dependence we vary θ from 0.5 to 2. In the log-linear Poisson regression approach this corresponds to using offsets varying between log(0.5) and log(2). Table 2 shows $\hat{m}_{00(\theta)}$, the population size estimate $\hat{N}(\theta)$, the estimated

Table 1. The observed values for the two nationalities, with the Afghan, Iraqi, and Iranian people residing in the Netherlands in 2007 on the left, and the Polish people residing in the Netherlands in 2009 on the right.

AII	HKS		Polish	HKS	
GBA	1	0	GBA	1	0
1	1,085	26,254	1	374	39,488
0	255	-	0	1,445	-

Table 2. Sensitivity analysis of the population size estimate for the people residing in the Netherlands in 2007 with Afghan, Iraqi, and Iranian nationality (upper panel) and for people with Polish nationality in 2009 (lower panel).

		Odds ratio				
		0.50	0.67	1.00	1.50	2.00
AII	$\hat{m}_{00(\theta)}$	3,085	4,114	6,170	9,255	12,341
	$\hat{N}_{(\theta)}$	30,679	31,708	33,764	36,849	39,935
	$\hat{N}/\hat{N}_{(\theta)}$	1.10	1.06	1.00	0.92	0.85
	se	223	293	441	647	864
Polish	$\hat{m}_{00(\theta)}$	76,284	101,712	152,567	228,851	305,135
	$\hat{N}_{(\theta)}$	117,591	143,019	193,874	270,158	346,442
	$\hat{N}/\hat{N}_{(\theta)}$	1.65	1.36	1.00	0.72	0.56
	se	4,473	6,024	8,787	13,630	17,866

relative bias $\hat{N}/\hat{N}_{(\theta)}$ and the bootstrapped standard error (se) of the estimate for both nationalities (details about the parametric bootstrap are provided in Subsection 5.3). As can be seen from the upper panel of Table 2, for the people with Afghan, Iraqi, and Iranian nationality under a dependence of $\theta = 0.5$, the estimate $\hat{m}_{00(\theta)}$ is half the size of the population size estimate under independence, and for a dependence of $\theta = 2$ the estimate $\hat{m}_{00}$ is twice the size of the population size estimate under independence. If in the population the registers are dependent with a true size θ, the population size estimate under independence varies between a ten percent overestimation and a 15 percent underestimation. Thus when the true $\theta \neq 1$ our population size estimate under independence remains fairly accurate.

However, for the Polish people the population size estimate under dependence is not robust. As can be seen from the lower panel of Table 2, if in the population the registers are dependent with a true size θ, the population size estimate under independence deviates between a 65 percent overestimation and 44 percent underestimation. Thus when the true $\theta \neq 1$, the population size estimate under independence for the Polish people is not robust.

The most important reason why the population size estimate deviates this much is because the implied coverage of the people with Afghan, Iraqi, and Iranian nationality is smaller than for the individuals with a Polish nationality. For example, 1,085 is $1,085/(1,085 + 255) = 0.81$, thus 81 percent of implied coverage of the GBA measured by the HKS. By contrast, for the individuals with Polish nationality the implied coverage of the GBA is only 21 percent, confirming the research by Brown et al. (2006) that as the observed coverage increases, the implied coverage increases and thus the population size estimate is more robust against dependence.

The estimated standard error of $\hat{N}_{(\theta)}$ is mainly determined by the size of $\hat{m}_{00(\theta)}$, and this explains the sharp rise of the standard error from $\theta = .50$ to $\theta = 2.00$ and the difference in standard error between the individuals with Afghan, Iraqi, and Iranian nationality and the individuals with Polish nationality.

3. Two Registers With Fully Observed Covariates

Covariates were first introduced to capture-recapture by Alho (1990) to reduce the heterogeneity resulting from individual differences on that covariate. As such, if covariates

are available, the generally nonfeasible independence assumption can be replaced with a less strict conditional independence assumption, where independence is conditional on covariates (Bishop et al. 1975; van der Heijden et al. 2012). This assumption is less stringent because it can take into account inclusion probabilities that are heterogeneous over the levels of the included covariate. Another advantage of using covariates is that it allows us to investigate the characteristics of the missing portion of the population.

Suppose we have observed covariate X, where the levels of X are indexed by x, $(x = 0,\ 1)$. Under independence conditional on X, there are two zero counts for cases not found in either register, namely for $x = 0$ and for $x = 1$. Let m_{ijx} denote the expected values for A, B and X. The log-linear model for independence for two registers and covariate X is

$$\log m_{ijx} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{ix}^{AX} + \lambda_{jx}^{BX}, \tag{8}$$

with identifying restrictions that a parameter equals zero when i or j or $x = 0$. When assuming independence between A and B conditional on X, $\lambda_{ij}^{AB} = \lambda_{ijx}^{ABX} = 0$. We use the notation of Bishop et al. (1975) to denote hierarchical log-linear models, that is, we denote this model as $[AX][BX]$.

In Section 2 we discussed two ways to estimate population sizes in a sensitivity analysis, namely one using an offset in a Poisson log-linear model and another using odds ratios directly. Here we only discuss the first way as it is more general. We assume that n_{ijx} follow a Poisson distribution and a log link connects the expected value m_{ijx} to the linear predictor.

It is important to note that in this context, too, sensitivity analyses are useful for assessing the impact of assumptions that are not verifiable from the data under study. Here conditional independence is the unverifiable assumption, since model $[AX][BX]$ is the saturated model. By contrast, model violations for more restricted models are verifiable in the data, for example for a model such as $[A][BX]$. Hence, the impact of interaction between A and X does not have to be investigated via a sensitivity analysis. However, when there may be dependence between A and B, a sensitivity analysis is useful.

We model dependence in the data by adding fixed parameters $\tilde{\lambda}_{ij}^{AB} + \tilde{\lambda}_{ijx}^{ABX}$ to Model (8). We again work under the saturated model, as the number of parameters to be estimated is equal to the number of observed parameters:

$$\log m_{ijx} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_x^X + \lambda_{ix}^{AX} + \lambda_{jx}^{BX} + \tilde{\lambda}_{ij}^{AB} + \tilde{\lambda}_{ijx}^{ABX}, \tag{9}$$

with the additional restrictions that parameters $\tilde{\lambda}_{ij}^{AB}$ and $\tilde{\lambda}_{ijx}^{ABX}$ equal zero when i or j or $x = 0$.

Under dependence between A and B given X, the association between the odds ratio θ_x and the log-linear parameters is:

$$\theta_x = \frac{m_{11x}m_{00x}}{m_{10x}m_{01x}} = \exp\left(\tilde{\lambda}_{11}^{AB} + \tilde{\lambda}_{11x}^{ABX}\right). \tag{10}$$

When we assume that dependence for $x = 0$ is identical to dependence for $x = 1$, then:

$$\theta = \frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}} = \exp\left(\tilde{\lambda}_{11}^{AB}\right). \tag{11}$$

Table 3. The observed values for the Afghan, Iraqi, and Iranian people, males on the left panel and females on the right panel.

Males	HKS		Females	HKS	
GBA	1	0	GBA	1	0
1	972	14,883	1	113	11,371
0	234	-	0	21	-

We estimate (9) using log-linear Poisson regression with for cell (1,1,0) the offset $\tilde{\lambda}_{11}^{AB}$ and for cell (1,1,1) the offset $\tilde{\lambda}_{11}^{AB} + \tilde{\lambda}_{111}^{ABX}$. After estimating (9), estimates for the missed portions of the population are found by $\hat{m}_{000} = \exp(\hat{\lambda})$ and $\hat{m}_{001} = \exp\left(\hat{\lambda} + \hat{\lambda}_1^X\right)$.

Table 3 shows the data for the Afghan, Iraqi, and Iranian people distributed over males ($x = 0$) and females ($x = 1$). Under conditional independence, $\hat{m}_{000} = 3,583$ and $\hat{m}_{001} = 2,113$. Taken together, both registers missed 5,696 cases. Note that conditional independence does not imply marginal independence under model $[AX][BX]$, since the marginal odds ratio $1,085^*5,696/26,254^*255 = 0.92$, and hence shows dependence (under marginal independence it would be equal to 1).

We estimate the parameters in (9) with a Poisson regression with $\tilde{\lambda}_{ijx}^{ABX} = 0$, so that the odds ratio of the males equals the odds ratio of the females (cf. (11)). The upper panel of Table 5 shows the results of the sensitivity analysis for the people with Afghan, Iraqi, and Iranian nationality in 2007 and the covariate gender. If in the population the registers are dependent with a true size θ, the population size estimate under independence varies between a nine percent overestimation to a 15 percent underestimation. As $\hat{m}_{00(\theta)}$ is relatively small, the standard error is relatively small. Thus when the true $\theta = 0.5$ but we estimate under $\theta = 1$, the population size estimate under independence is fairly robust.

For the people with a Polish nationality residing in the Netherlands in 2009 the covariate gender is also used. Under conditional independence, the estimate $\hat{m}_{00x} = 144,548$. The lower panel of Table 5 shows the sensitivity analysis of the population size estimator under conditional independence. If in the population the registers are dependent with a true size θ, the population size estimate under independence ranged between a 58 percent overestimation and a 42 percent underestimation. Thus when the true $\theta \neq 1$, the population size estimate deviates greatly from the population size estimate under $\theta = 1$, indicating that for this dataset the population size estimate under independence is not robust.

We note that this example uses a covariate with only two levels. One can easily extend this to covariates with more levels. Assume covariate W has three levels, where the levels of W are indexed by w ($w = 0,\ 1,\ 2$). Then there are three zero counts, namely for $w = 0$, $w = 1$ and $w = 2$. One can estimate the zero counts using Equation (10), where estimates

Table 4. The observed values for the Polish people, males on the left panel and females on the right panel.

Males	HKS		Females	HKS	
GBA	1	0	GBA	1	0
1	313	19,152	1	61	20,336
0	1,349	-	0	96	-

Table 5. Sensitivity analysis for the people with Afghan, Iraqi, and Iranian (AII) nationality residing in the Netherlands in 2007 (upper panel), and the people with Polish nationality residing in the Netherlands in 2009 (lower panel), conditional on gender.

		Odds ratio				
		0.50	0.67	1.00	1.50	2.00
AII	$\hat{m}_{00}$	2,848	3,797	5,696	8,544	11,392
	$\hat{N}_{(\theta)}$	30,442	31,391	33,290	36,138	38,986
	$\hat{N}/\hat{N}_{(\theta)}$	1.09	1.06	1.00	0.92	0.85
	Se	292	390	576	863	1144
Polish	$\hat{m}_{00}$	57,274	76,365	114,548	171,821	229,095
	$\hat{N}_{(\theta)}$	98,581	117,672	155,855	213,128	270,402
	$\hat{N}/\hat{N}_{(\theta)}$	1.58	1.32	1.00	0.73	0.58
	Se	3,814	5,088	7450	11,465	15,135

for the missed portions of the population are found by $\hat{m}_{000} = \exp(\hat{\lambda})$ and $\hat{m}_{001} = \exp\left(\hat{\lambda} + \hat{\lambda}_1^W\right)$ and $\hat{m}_{002} = \exp\left(\hat{\lambda} + \hat{\lambda}_2^W\right)$.

4. Two Registers With Partially Observed Covariates

In Section 3 we used covariates that are present in both registers (fully observed covariates) to replace the strict independence assumption with an independence assumption conditional on covariates. However, a register usually also has a set of variables that are only measured in one register and not in the other register (partially observed covariates). Partially observed covariates in A are usually ignored because including them leads to missing data in B for those individuals that are not in A, and vice versa. When these covariates are related to the inclusion probability, ignoring the partially observed covariates can lead to a biased population size estimate (Zwane and van der Heijden 2007; van der Heijden et a1. 2012).

4.1. Partially Observed Covariates

Partially observed covariates can be approached as a missing data problem (Zwane and van der Heijden 2007). If we assume MAR mechanism for the data, then we can use the Expectation-Maximization (EM) algorithm to estimate the missing values of the partially observed covariate of register 1 (and 2) for the individuals not present in Register 1 (and 2). MAR assumes that the probability of missingness depends only on the observed variables in the capture-recapture model (Little and Rubin 1987). When the assumption of MAR has been satisfied, the EM algorithm will give unbiased estimates.

Suppose register 1 has the covariate X_1, indexed by $k(k = 0, 1)$, where the values for X_1 are missing for $A = 0$ because X_1 is not in register 2. Assume that register 2 has the covariate X_2, indexed by $l(l = 0, 1)$, where the values for X_2 are missing for $B = 0$ because X_2 is not in register 1. The log-linear conditional independence model for two registers, with two partially observed covariates X_1 and X_2, is denoted as

$$\log m_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2}, \quad (12)$$

Table 6. Expected values for two registers and two partially observed covariates.

		B = 1		B = 0	
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
A = 1	$X_1 = 1$	m_{1111}	m_{1110}	m_{1011}	m_{1010}
	$X_1 = 0$	m_{1101}	m_{1100}	m_{1001}	m_{1000}
A = 0	$X_1 = 1$	m_{0111}	m_{0110}	m_{0011}	m_{0010}
	$X_1 = 0$	m_{0101}	m_{0100}	m_{0001}	m_{0000}

with identifying restrictions $\lambda_{ij}^{AB} = \lambda_{ik}^{AX_1} = \lambda_{jl}^{BX_2} = \lambda_{ijk}^{ABX_1} = \lambda_{ijl}^{ABX_2} = \lambda_{ijkl}^{ABX_1X_2} = 0$. The conditional independence model is denoted by $[AX_2][BX_1][X_1X_2]$. Inclusion of the parameter $\lambda_{il}^{AX_2}$ instead of the parameter $\lambda_{ik}^{AX_1}$ may seem counterintuitive, but no interaction for A and X_1 can be identified as the levels of X_1 do not vary over individuals for which $A = 0$, and similarly for B and X_2 (Zwane and van der Heijden 2007).

Table 6 illustrates that two registers with two covariates lead to 16 cells. However, because our covariates are only partially observed, columns $X_2 = 1$ and $X_2 = 0$ for $B = 0$ are collapsed, just as rows $X_1 = 1$ and $X_1 = 0$ for $A = 0$ are collapsed. In other words, we do not observe counts for m_{0111} and m_{0101} but only one count for the sum $m_{0111} + m_{0101}$, and similarly for $m_{0110} + m_{0100}$, $m_{1011} + m_{1010}$ and $m_{1001} + m_{1000}$. Note that we have no observed values for m_{0011}, m_{0001}, m_{0010} and m_{0000}, as these refer to individuals who are in neither of the registers. Thus model $[AX_2][BX_1][X_1X_2]$ is saturated with eight observed values and eight parameters to be estimated.

Using the EM algorithm we first estimate the four missing cells, that is, the cells that are missing because the covariates are only partially observed. In the E-step we spread out the four sums $m_{0111} + m_{0101}$, $m_{0110} + m_{0100}$, $m_{1011} + m_{1010}$ and $m_{1001} + m_{1000}$ over the eight cells to get an expectation for the missing data. In the M-step we estimate log-linear model (12) to the completed table of twelve cells. For estimation, we assume that the twelve counts follow a Poisson distribution and a log link connects the expected counts to the linear predictor. The resulting estimates are then used for the E-step where in the M-step, following (12), we estimate the parameters again.

To illustrate we once more use the data on the people with Afghan, Iraqi, and Iranian nationality residing in the Netherlands in 2007 with two partially observed covariates (van der Heijden et al. 2012). The GBA has the partially observed covariate marital status (X_1), where $X_1 = 1$ denotes either being married or living together and $X_1 = 0$ denotes either unmarried, divorced or widowed. The HKS has the partially observed covariate police region (X_2), where $X_2 = 1$ denotes residing in one of the five biggest cities of the Netherlands (i.e., Amsterdam, Rotterdam, Utrecht, The Hague, and Eindhoven) and $X_2 = 0$ denotes residing in the rest of the country.

Due to the log-linear model used, the first four observed values remain unchanged for each iteration (for $GBA = 1$ and $HKS = 1$). The upper panel of Table 7 shows the observed counts and the lower panel of Table 7 shows the fitted counts after convergence of the EM algorithm. As an example, the observed value of 91 (for $X_2 = 1$, where X_1 values are missing under $GBA = 0$) is spread out into the values 64 for $X_1 = 1$ and 27 for $X_1 = 0$. After convergence, the unobserved part of the population is estimated. In total,

Table 7. Data for the Afghan, Iraqi, and Iranian people residing in the Netherlands in 2007, spread out over the partially observed covariates marital status X_1 and police region X_2

Panel 1: The observed counts

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

Panel 2: The fitted frequencies

		HKS = 1		HKS = 0	
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
GBA = 1	$X_1 = 1$	259	539	4,511	9,387
	$X_1 = 0$	110	177	4,736	7,620
GBA = 0	$X_1 = 1$	64	123	1,112	2,150
	$X_1 = 0$	27	41	1,168	1,745

we estimate that there were 33,770 individuals with Afghan, Iraqi, and Iranian nationality residing in the Netherlands in 2007.

4.2. Sensitivity Analyses

We again make use of a sensitivity analysis to investigate the unverifiable assumption of independence conditional on partially observed covariates. Model violations for more restricted models are verifiable in the data. For example, using a model such as $[AX_2][BX_1]$ allows us to investigate absence of interaction $\lambda_{kl}^{X_1X_2}$ in the data. Thus the impact of an interaction between X_1 and X_2 does not need to be investigated via a sensitivity analysis. However, in this context (12) is the saturated model and therefore model violations such as dependence between A and X_1, between B and X_2, and between A and B are unverifiable, rendering it useful to conduct a sensitivity analysis. Note that in the previous sections we used a sensitivity analysis to assess the interaction between the two registers. In this section we assess not only the interaction between A and B, but also the interaction between the register and its partially observed covariate. To exemplify, we introduce an interaction parameter that simulates dependence between the GBA and marital status. Such a dependence would imply that marital status influences the inclusion probability of being in the GBA.

The log-linear model for an interaction between A and B would be:

$$\log m_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2} + \tilde{\lambda}_{ij}^{AB}, \qquad (13)$$

with additional identifying restrictions that $\tilde{\lambda}_{ij}^{AB} = 0$ when i or j equals 0. Here $\exp\left(\tilde{\lambda}_{ij}^{AB}\right)$ is the conditional odds ratio for the interaction between A and B.

Assume the partially observed covariate marital status is related to the inclusion probability of the GBA, thus $\lambda_{ik}^{AX_1} \neq 0$. Because the interaction between A and X_1 is

unverifiable from the data, the fixed parameter $\tilde{\lambda}_{ik}^{AX_1}$ has been added to the log-linear model (12). We continue to work under the saturated model:

$$\log m_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2} + \tilde{\lambda}_{ik}^{AX_1}, \qquad (14)$$

with additional identifying restrictions that $\tilde{\lambda}_{ik}^{AX_1} = 0$ when i or k equals 0. The same can be done for the interaction between B and X_2. When the partially observed covariate X_2 is related to the inclusion probability of register B, $\lambda_{jl}^{BX_2} \neq 0$. We add fixed parameter $\tilde{\lambda}_{jl}^{BX_2}$ to the log-linear model. The log-linear model then becomes:

$$\log m_{ijkl} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^{X_1} + \lambda_l^{X_2} + \lambda_{il}^{AX_2} + \lambda_{jk}^{BX_1} + \lambda_{kl}^{X_1X_2} + \tilde{\lambda}_{jl}^{BX_2}, \qquad (15)$$

with additional identifying restrictions that $\tilde{\lambda}_{jl}^{BX_2} = 0$ when j or l equals 0. We can estimate (13), (14) and (15) via Poisson regressions with offsets. Note that in modeling these relationships we have to fix the offset variable on a log scale. Then we can estimate the portions of the population that both registers have missed by $\hat{m}_{0000} = \exp(\hat{\lambda})$, $\hat{m}_{0010} = \exp\left(\hat{\lambda} + \hat{\lambda}_1^{X_1}\right)$, $\hat{m}_{0001} = \exp\left(\hat{\lambda} + \hat{\lambda}_1^{X_2}\right)$ and $\hat{m}_{0011} = \exp\left(\hat{\lambda} + \hat{\lambda}_1^{X_1} + \hat{\lambda}_1^{X_2} + \hat{\lambda}_{11}^{X_1X_2}\right)$.

The upper panel of Table 8 shows the sensitivity analysis for the interaction between A and B, the middle panel shows the sensitivity analysis for the interaction between A and X_1 and the lower panel shows the sensitivity analysis for the interaction between B and X_2 for the Afghan, Iraqi, and Iranian people. As can be seen, for the interaction between A and B, the relative bias is similar to the bias found in Tables 2 and 5. If in the population the GBA and marital status are dependent with a true size θ, the estimation under independence deviates between a 2.22 percent overestimation to a 2.89 percent underestimation, and the estimation under independence between the HKS and police region deviates between a 0.23 percent underestimation and a 0.19 percent overestimation. Thus for the interactions AX_1 and BX_2, when the true $\theta \neq 1$, the population size estimate under independence remains fairly robust.

We have done the same for the people with Polish nationality residing in the Netherlands in 2009. The observed values are shown in the upper panel of Table 9 and the expected

Table 8. Sensitivity analysis of the population size estimate for the people residing in the Netherlands in 2007 with an Afghan, Iraqi, and Iranian nationality with the interaction A and X_1 (upper panel) and the interaction between B and X_2 (lower panel).

		Odds ratio				
		0.50	0.67	1.00	1.50	2.00
AB	$\hat{m}_{00(\theta)}$	3.088	4,117	6,176	9,264	12,352
	$\hat{N}_{(\theta)}$	30.682	31,711	33,770	36,858	39,946
	$\hat{N}/\hat{N}_{(\theta)}$	1.10	1.06	1.00	0.92	0.85
AX1	$\hat{m}_{00(\theta)}$	5,443	5,711	6,176	6,736	7,179
	$\hat{N}_{(\theta)}$	33,037	33,305	33,770	34,330	34,773
	$\hat{N}/\hat{N}_{(\theta)}$	1.0222	1.0140	1.00	0.9837	0.9711
BX2	$\hat{m}_{00(\theta)}$	6,253	6,220	6,176	6,136	6,112
	$\hat{N}_{(\theta)}$	33,847	33,814	33,770	33,730	33,706
	$\hat{N}/\hat{N}_{(\theta)}$	0.9977	0.9987	1.00	1.0012	1.0019

Table 9. The observed counts for the people with Polish nationality residing in the Netherlands in 2009 (upper panel) and the fitted frequencies spread out over the partially observed covariates (lower panel).

Panel 1: The observed counts

		HKS = 1		HKS = 0
		$X_2 = 1$	$X_2 = 0$	X_2 missing
GBA = 1	$X_1 = 1$	111	188	25,416
	$X_1 = 2$	32	43	14,072
GBA = 0	$X_1 = 1$	603	842	

Panel 2: The fitted frequencies

		HKS = 1		HKS = 0	
		$X_2 = 1$	$X_2 = 0$	$X_2 = 1$	$X_2 = 0$
GBA = 1	$X_1 = 1$	111	188	9,435	15,981
	$X_1 = 2$	32	43	6,004	8,068
GBA = 0	$X_1 = 1$	468	685	39,787	58,250
	$X_1 = 2$	135	157	25,318	29,408

frequencies are shown in the lower panel of Table 9. Again a sensitivity analysis has been conducted, which is shown in Table 10. Just as with the individuals with Afghan, Iraqi, and Iranian nationality, the estimates and thus the relative bias under dependence between A and B remains unchanged (cf. Tables 2 and 5). If in the population the GBA and marital status are dependent with a true size θ, the population size estimate under independence ranges from a seven percent overestimation to a nine percent underestimation (upper panel). The estimate under independence between the HKS and police region deviates from a two percent underestimation to a two percent overestimation (lower panel). Thus when the true $\theta \neq 1$, the population size estimate under independence remains fairly robust.

Table 10. Sensitivity analysis of the population size estimate for the the people residing in the Netherlands in 2009 with Polish nationality with the interaction between A and X_1 (upper panel) and the interaction between B and X_2 (lower panel).

		Odds ratio				
		0.50	0.67	1.00	1.50	2.00
AB	$\hat{m}_{00(\theta)}$	76,381	101,842	152,762	229,143	305,524
	$\hat{N}_{(\theta)}$	117,688	143,149	194,069	270,450	346,832
	$\hat{N}/\hat{N}_{(\theta)}$	1.65	1.36	1.00	0.71	0.56
AX1	$\hat{m}_{00(\theta)}$	139,494	144,238	152,762	163,584	172,582
	$\hat{N}_{(\theta)}$	180,801	185,545	194,069	204,891	213,889
	$\hat{N}/\hat{N}_{(\theta)}$	1.07	1.05	1.00	0.95	0.91
BX2	$\hat{m}_{00(\theta)}$	156,616	155,004	152,762	150,707	149,429
	$\hat{N}_{(\theta)}$	197,923	196,311	194,069	192,014	190,736
	$\hat{N}/\hat{N}_{(\theta)}$	0.98	0.99	1.00	1.01	1.02

Under the use of partially observed covariates it becomes clear why the log-linear Poisson regression provides a more general approach than using odds ratios to implement the sensitivity analyses. When using log-linear Poisson regression the process becomes vastly simpler, in that the offset can be set to any number per cell. When multiple different offsets are in use, the log-linear Poisson regression allows for this complexity, whereas implementing odds ratios may become gruesome.

5. Miscellany

5.1. Extension to Multiple Sources

One way to make the impact of possible violations of the independence assumption less severe is by conditioning on covariates, as we have seen in Section 3 and 4. Another way to make the impact of possible violations of the independence assumption less severe is by adding registers, when more registers are available (cf. Baffour et al. 2013). Assume we have three registers 1, 2 and 3, where the variables A, B and C respectively stand for inclusion in the registers. We denote the expected values m_{ijp} where $i, j, p = 1$ stand for the inclusion into Registers 1, 2 and 3 respectively and where $i, j, p = 0$ stands for the absence in registers 1, 2 and 3.

For three variables, the saturated log-linear model is denoted by

$$\log\ m_{ijp} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_p^C + \lambda_{ij}^{AB} + \lambda_{ip}^{AC} + \lambda_{jp}^{BC}, \tag{16}$$

with identifying restrictions that a parameter equals zero when i,j or $p = 0$. We assume that interaction parameter $\lambda_{ijp}^{ABC} = 0$. Model $[AB][BC][AC]$ is the saturated model, as the number of observed parameters equals the number of parameters to be estimated. With d registers, we assume that the d-factor interaction is absent.

For estimation, assume that n_{ijp} follow a Poisson distribution and a log link connects the expected value m_{ijp} to the linear predictor. We can estimate the parameters in (16) via a Poisson log-linear regression.

Model $[AB][BC][AC]$ assumes that odds conditional on a third variable are equal, for example for the odds ratio between A and B given C we find

$$\frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}}. \tag{17}$$

Model (16) assumes that for estimation with odds ratios under saturated model $[AB][BC][AC]$ we get:

$$\frac{\hat{m}_{010}\hat{m}_{001}\hat{m}_{100}\hat{m}_{111}}{\hat{m}_{011}\hat{m}_{110}\hat{m}_{101}} = \frac{n_{010}n_{001}n_{100}n_{111}}{n_{011}n_{110}n_{101}} = \hat{m}_{000}. \tag{18}$$

An estimate for $\hat{m}_{000}$ is easily derived from (17) as $[AB][AC][BC]$ is the saturated model in this context; absence of the three-factor interaction is an unverifiable assumption as it cannot be verified in the data. More restricted models such as $[AB][AC]$ are verifiable in the data. However, we can investigate the robustness of the population size estimate against violations of the assumption that the three-factor interaction is absent by fixing the

interaction parameter to anything but 0, that is, $\tilde{\lambda}^{ABC}_{ijp} \neq 0$. Thus the log-linear model becomes:

$$\log\ m_{ijp} = \lambda + \lambda^A_i + \lambda^B_j + \lambda^C_p + \lambda^{AB}_{ij} + \lambda^{AC}_{ip} + \lambda^{BC}_{jp} + \tilde{\lambda}^{ABC}_{ijp}, \tag{19}$$

with the additional identifying restriction where parameter $\tilde{\lambda}^{ABC}_{ijp}$ equals zero when i or j or $p = 0$. The population size estimate under (19) can be estimated using Poisson log-linear regression with parameter $\tilde{\lambda}^{ABC}_{ijp}$ as an offset.

Under dependence between A and B given C, the association between the odds ratio θ and the log-linear parameters is:

$$\theta^{(p=0)}_{AB} = \frac{m_{110}m_{000}}{m_{100}m_{010}} = \exp\ (\lambda^{AB}_{11}), \tag{20}$$

and:

$$\theta^{(p=1)}_{AB} = \frac{m_{111}m_{001}}{m_{101}m_{011}} = \exp\ (\lambda^{AB}_{11} + \lambda^{ABC}_{111}). \tag{21}$$

When we assume that the odds ratio between A and B is the same for $p = 0$ and $p = 1$, we get

$$\theta_{AB} = \frac{m_{110}m_{000}}{m_{100}m_{010}} = \frac{m_{111}m_{001}}{m_{101}m_{011}} = \exp\ (\lambda^{AB}_{11}). \tag{22}$$

When more registers are available we can use these extra registers to reduce the impact of violations of the independence assumption. As we have shown, the log-linear model is easily generalizable to multiple registers.

5.2. *Multiplier Method*

The multiplier method is an alternative method to estimate the size of a population and it is used, amongst others, in drug use research and HIV prevalence (European Monitoring Centre for Drugs and Drug Addiction (EMCDDA) 1997; Cruts and van Laar 2010; Temurhan et al. 2011). Multiplier methods are user-friendly for their mathematical simplicity, and absence of linkage, and are straightforward to use. At least two data sources are needed to use the multiplier method, usually a comprehensive register and a survey. For example, assume we wish to estimate the number of Polish people residing in the Netherlands in 2013. We assume that everyone has an equal chance of going to a hospital, thus we go to hospitals to assess how many Polish patients there are, and ask them whether they are in the GBA. Then assume the data we found is the data from Table 11. There are 200 Polish people, of which 150 are in the GBA. Thus $p(\text{GBA}\ |\text{Hospital}) = 0.75$. If a total of 40,000 Polish people are registered, in the GBA, this means our actual total should be $40,000/0.75 = 53,333$ and we missed $53,333 - 40,000 = 13,333$ people who are not registered in the GBA.

The multiplier method can also be explained from the perspective of capture-recapture methods. Using the counts provided above, we have n_{11}, n_{01} and n_{1+} so that $n_{1+} - n_{11} = n_{10}$ and Equation (2) gives $\left(39,850^{*}50\right)/150 = 13,283$. Then $\hat{N} = 150 + 50 + 39,850 + 13,283 = 53,333$, which is the exact same value as we got above. A sensitivity analysis could be conducted using Equation (7).

Table 11. Artificial observed data for the Polish people in the hospital

		Hospital		
		1	0	
GBA	1	150	39,850	40,000
	0	50	-	-
		200	-	-

The attractiveness of the multiplier method lies in the absence of the linkage of two sources. When estimating hidden or hard-to-reach populations, it is likely that it is difficult to obtain identifying variables to link the individuals in the samples. The absence of linkage is what makes the multiplier method different from capture-recapture. However, it has to be kept in mind that the multiplier method also relies on the underlying assumptions that being in the hospital is statistically independent from being in the GBA, and that it relies on individuals reporting their GBA status accurately when being admitted to the hospital.

5.3. *Confidence Intervals*

Apart from robustness, another aspect of the usefulness of a point estimate is its confidence interval. Parametric bootstrap confidence intervals can be used to find these confidence intervals in a simple way when dealing with incomplete contingency tables. In a parametric bootstrap sample, the estimate $\hat{m}_{00(\theta)}$ for cell (0, 0) is used in the multinomial probabilities. So for Table 1, the four probabilities are $n_{11}/\hat{N}_{(\theta)}$, $n_{10}/\hat{N}_{(\theta)}$, $n_{01}/\hat{N}_{(\theta)}$ and $\hat{m}_{00(\theta)}/\hat{N}_{(\theta)}$. A sample with size $\tilde{\lambda}_{ij}^{AB}$ is drawn with replacement. This yields four counts $n_{11}^{b=1}$, $n_{01}^{b=1}$, $n_{10}^{b=1}$ and $n_{00}^{b=1}$. The first bootstrap population size estimate $\hat{N}^{b=1}$ is found using only $n_{11}^{b=1}$, $n_{01}^{b=1}$, $n_{10}^{b=1}$, that is, ignoring $n_{00}^{b=1}$, and estimating $\hat{m}_{00(\theta)}^{b=1}$. This is repeated 10,000 times, yielding 10,000 bootstrap population size estimates. From these, 2.5 and 97.5 percentile scores are derived.

To exemplify we constructed a parametric bootstrapping confidence interval on the data presented in Section 2, which can be found in Table 12. The R code for the parametric bootstrap confidence interval can be found in Appendix A.3.

To compare, we also constructed the asymptotic confidence estimate $CI = \hat{m}_{00} + / - z_{(.975)}\left(\sqrt{\hat{\mathrm{Var}}(n)}\right)$, where $\hat{\mathrm{Var}}(n) = \left(n_{1+}n_{+1}n_{10}n_{01}\right)/\left((n_{11})^3\right)$ (Bishop et al. 1975). The estimated confidence interval for the Afghan, Iraqi, and Iranian people under independence is $32,905.44 - 34,623.16$, which is close to the bootstrapped confidence interval.

Table 12. Confidence intervals

Odds Ratio	AII	Polish
0.50	30,254 – 31,132	109,529 – 127,022
0.67	31,156 – 32,288	132,278 – 155,837
1.00	32,931 – 34,654	177,476 – 212,431
1.50	35,607 – 38,125	245,439 – 298,960
2.00	38,292 – 41,682	314,212 – 384,579

6. Discussion

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. Deviations from independence when implied coverage is low (and thus $\hat{m}_{00}$ is high) result in bigger deviations from the population size estimate under fixed dependence than when the implied coverage is higher. Thus the estimate becomes less robust and this makes the situation worse. For the Afghan, Iraqi, and Iranian people the population size estimate did not change much when dependence was introduced; it also remained fairly robust whether or not we assumed conditional independence on fully observed covariates. However, for the Polish people, the implied coverage is small, resulting in a higher $\hat{m}_{00}$ so that the deviation from independence will be large. The resulting lack of robustness makes it even worse. Not only did the population size estimate under independence change dramatically under fixed dependence, adding a covariate to replace the strict independence assumption with the less strict independence assumption conditional on covariates changed the population size estimate but did not improve the robustness.

This reflects the fact that Polish people, much more than people from Afghanistan, Iraq, and Iran, are in the position that they work on a temporary basis without living permanently in the Netherlands. By law, it is permitted for people from European Union countries like Poland to work in the Netherlands without a work and living permit. This is not the case for people from Afghanistan, Iraq, and Iran. Therefore, the coverage of the GBA differs between both nationalities, which gives a relatively high estimation of the missed population of the Polish people compared to the Afghan, Iraqi, and Iranian people. Additionally, because we multiply $\hat{m}_{00}$ with θ, it follows that a bigger $\hat{m}_{00}$ will result in a bigger $\hat{m}_{00\theta}$ than a smaller $\hat{m}_{00}$ would when multiplied with the same θ.

We also showed how to investigate robustness of the population size estimate in models with partially observed covariates. For the example we used, the population size estimate was relatively insensitive to violation of specific conditional independence assumptions. Since adding covariates reduces heterogeneity and gives the opportunity to assess how the population is divided over the levels of the covariate, it is useful to include a partially observed covariate.

In this article we assumed that the only assumption that was violated was the independence assumption. However, violation of other assumptions could also have a large impact on the population size estimate. In particular, research on violation of the assumptions that the registers are perfectly linked as well as that the population is closed during the observation period is needed to draw conclusions on the usefulness of the capture-recapture method for estimating the undercoverage of census data.

We have chosen a range of odds ratio from 0.5 to 2. To our knowledge, it is not possible to get an accurate estimation of what a realistic θ value would be, since it is impossible to ascertain θ from the data. One way of dealing with the strict independence assumption is by adding a third register, hence using another source to estimate θ, as has been done by Brown et al. (2006) who created an adjustment factor based on a third source for the census.

In conclusion, it is important to assess the size of the implied coverage of one of the registers. We have shown that lack of robustness under dependence is easily established when implied coverage is low. However, when implied coverage is high the population size estimate remains fairly robust. Thus, instead of accepting the population size estimate as it is, researchers should report on the robustness of their estimate.

7. Appendix

To estimate the population size under log-linear models, we have used Poisson regression with an offset in SPSS and R.

A.1. R Code

Below is given the R code to get estimates $\hat{m}_{00kl}$ in the EM algorithm, for the Polish data only.

```
##Give the data
data = c(111,188,32,43,12708,12708,7036,7036,301.5,421,301.5,421) ## Polish data
data = data*10000
freqitx = freqit1 = data

## Design matrix
A = c(1,1,1,1,1,1,1,1,0,0,0,0)
B = c(1,1,1,1,0,0,0,0,1,1,1,1)
X1 = c(1,1,0,0,1,1,0,0,1,1,0,0)
X2 = c(1,0,1,0,1,0,1,0,1,0,1,0)

## OR for independence
offst = c(0,0,0,0,0,0,0,0,0,0,0,0)
for (i in 1:50000){
glm = glm(freqitx ~ A*X2 + B*X1 + X1*X2, offset=offst, family=poisson)
freqdata = c(data[1:4])
freqfit = glm$fitted.values[5:12]
freqitx = c(freqdata,freqfit)
freqitx = round(freqitx)}

## Parameter estimates under independence
par = glm$coefficients
m0011 = as.numeric(exp(par[1]+par[3]+par[5]+par[8]))
m0010 = as.numeric(exp(par[1]+par[5]))
m0001 = as.numeric(exp(par[1]+par[3]))
m0000 = as.numeric(exp(par[1]))
matrix = matrix(c(glm$fitted.values[1],glm$fitted.values[2],
glm$fitted.values[5],glm$fitted.values[6],glm$fitted.values[3],glm$fitted.values[4],
glm$fitted.values[7], glm$fitted.values[8], glm$fitted.values[9], glm$fitted.values[10],
m0011,m0010,glm$fitted.values[11],glm$fitted.values[12],m0001,m0000),4,4,byrow
= TRUE)
N = sum(matrix)
```

```
## Define the offsets. Here we only give an example for the offsets of BX2 = 0.5
offst1 = c(-0.6931472,0,-0.6931472,0,0,0,0,0,-0.6931472,0,-0.6931472,0)
## Iterative GLM Loop for the EM algorithm
for (i in 1:50000){
glm = glm(freqitx ~ A*X2 + B*X1 + X1*X2, offset = offst1, family=poisson)
freqdata = c(data[1:4])
freqfit = glm$fitted.values[5:12]
freqitx = c(freqdata,freqfit)
freqitx = round(freqitx)}

## Calculation of estimated missed frequencies
par = glm$coefficients
m0011 = as.numeric(exp(par[1] + par[3] + par[5] + par[8]))
m0010 = as.numeric(exp(par[1] + par[5]))
m0001 = as.numeric(exp(par[1] + par[3]))
m0000 = as.numeric(exp(par[1]))

m00comp = m0011 + m0010 + m0001 + m0000
PSE = sum(data)+ m00comp
print(m00comp)
print(sum(data)+ m00comp)
print(N/PSE)
```

A.2. SPSS Syntax

```
compute freqitx = freqit1.
compute freqitx = rnd(freqitx).
execute.
DEFINE EM_PGLM()
!DO !l = 1 !TO 10000.
GENLIN freqitx BY A B X1 X2 (ORDER = ASCENDING)
/MODEL A B X1 X2 A*X2 B*X1 X1*X2 INTERCEPT = YES OFFSET = offst05
DISTRIBUTION = POISSON LINK = LOG
/SAVE MEANPRED (pred_val).
compute diff = ABS(freqit1-pred_val).
means diff.
compute freqitx = pred_val.

IF((A = 1)&(B = 1)&(X1 = 1)&(X2 = 1))freqitx = freqit1.
IF((A = 1)&(B = 1)&(X1 = 2)&(X2 = 1))freqitx = freqit1.
IF((A = 1)&(B = 1)&(X1 = 1)&(X2 = 2))freqitx = freqit1.
IF((A = 1)&(B = 1)&(X1 = 2)&(X2 = 2))freqitx = freqit1.

COMPUTE freqitx = rnd(freqitx).
execute.
delete variables pred_val.
```

```
!DOEND
!ENDDEFINE.
##run the macro
EM_PGLM.
```

A.3. R Code Parametic Bootstrap

The R code presented below represents the parametric bootstrap for the Polish data from Table 1

```
data = c(374, 39488, 1445) ## Polish data
theta = 2
m00 = (data[2]*data[3])/data[1]
m00theta = m00*theta
datacomp = sum(data,m00theta)
## The estimate of N, under an offset theta
n = sum(data)
N = n + m00theta
##The relative bias under an offset theta
(n + m00)/N
## Parametric bootstrap
NN = c(N)
p = matrix(c(data/datacomp, m00theta/datacomp),1)
set.seed(N)
library(combinat)
databoot = rmultinomial(rep(NN, 10000),p)
m00boot = theta* (databoot[,2]*databoot[,3])/databoot[,1]
nboot = databoot[,1:3]
Nboot = m00boot + nboot[,1] + nboot[,2] + nboot[,3]
quantile(Nboot, c(0.025, 0.5, 0.975), type = 1)
sd = function(x) sqrt(var(x))
sd(Nboot)
```

8. References

Alho, J.M. 1990. "Logistic Regression in Capture-recapture Models." *Biometrics* 46: 623–635. Doi: http://dx.doi.org/10.2307/2532083.

Baffour, B., J.J. Brown, and P.W.F. Smith. 2013. "An Investigation of Triple System Estimators in Censuses." *Statistical Journal of the International Association for Official Statistics* 29: 53–68. Doi: http://dx.doi.org/10.3233/SJI-130760.

Bell, W.R. 1993. "Using Information from Demographic Analysis in Post-enumeration Survey Estimation." *Journal of the American Statistical Association* 88: 1106–1118. Doi: http://dx.doi.org/10.1080/01621459.1993.10476381.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. *Discrete multivariate analysis*. Cambridge, MA: MIT Press.

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 A* 169: 883–902.

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.

Chao, A., P.K. Tsay, S.-H. Lin, W.-Y. Shau, and D.-Y. Chao. 2001. "Tutorial in Biostatistics. The Application of Capture-recapture Models of Epidemiological Data." *Statistics in Medicine* 20: 3123–3157. Doi: http://dx.doi.org/10.1002/sim.996.

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

Cormack, R.M., Y.-F. Chang, and G.S. Smith. 2000. "Estimating Deaths from Industrial Injury by Capture-recapture: A Cautionary Tale." *International Journal of Epidemiology* 29: 1053–1059. Doi: http://dx.doi.org/10.1093/ije/29.6.1053.

Cruts, A.A.N., and M.W. van Laar. 2010. *Aantal Problematische Harddrugsgebruikers in Nederland*. Utrecht: Trimbos Instituut.

European Monitoring Centre for Drugs and Drug Addiction (EMCDDA). 1997. *Methodological Pilot Study of Local Level Prevalence Estimates*. Lisbon: EMCDDA.

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.

Hook, E.B., and R.R. Regal. 1992. "The Value of Capture-recapture Methods Even for Apparent Exhaustive Surveys." *American Journal of Epidemiology* 135: 1060–1067.

Hook, E.B., and R.R. Regal. 1995. "Capture-recapture Methods in Epidemiology: Methods and Limitations." *Epidemiologic Reviews* 17: 243–264.

Hook, E.B., and R.R. Regal. 1997. "Validity of Methods for Model Selection. *Weighting for Model Uncertainty, and Small Sample Adjustment in Capture-recapture Estimation." American Journal of Epidemiology* 145: 1138–1144. Available at: http://aje.oxfordjournals.org/content/145/12/1138.full.pdf (accessed July 2015).

Hook, E.B., and R.R. Regal. 2000. "Accuracy of Alternative Approaches to Capture-recapture Estimates of Disease Frequency: Internal Validity Analysis of Data from Five Sources." *American Journal of Epidemiology* 152: 771–779. Doi: http://dx.doi.org/10.1093/aje/152.8.771.

International Working Group for Disease Monitoring and Forecasting. 1995. "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 July 2015).

Little, R.J., and D.B. Rubin. 1987. *Statistical Analysis with Missing Data*. Hoboken, NJ: John Wiley and Sons.

Nirel, R., and H. Glickman. 2009. "Sample Surveys and Censuses." *Sample surveys: Design Methods and Applications* 29A: 539–565.

Office of National Statistics (ONS). 2012. *The 2011 Census Coverage Assessment and Adjustment Process*. London: Office for National Statistics.

Seber, G.A.F. 1982. *The Estimation of Animal Abundance and Related Parameters*. London: Edward Arnold.

Temurhan, M., R. Meijer, S. Choenni, M. van Ooyen-Houben, G. Cruts, and M. van Laar. 2011. "Capture-recapture Method for Estimating the Number of Problem Drug Users: The Case of the Netherlands." *In Proceedings of the Intelligence and Security Informatics Conference (EISIC)*, 12–14 september, 2011, 46–51. Available at: http://www.computer.org/csdl/proceedings/eisic/2011/4406/00/4406a046-abs.html (accessed July 2015).

Van der Heijden, P.G.M., M.J.L.F. Cruy, and G. van Gils. 2011. Aantallen Geregistreerde en Nietgeregistreerde 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*. Available at: http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2013/01/14/aantallen-geregistreerde-en-niet-geregistreerde-burgers-uit-moe-landen-die-in-nederland-verblijven.html [in Dutch] (accessed July 2015).

Van der Heijden, P.G.M., J. Whittaker, M.J.L.F. Cruy, B.F.M. Bakker, and H.N. van der Vliet. 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. Doi: http://dx.doi.org/10.1214/12-AOAS536.

Wolter, K.M. 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.

Zwane, E.N., and P.G.M. van der Heijden. 2004. "Semiparametric Models for Capture-recapture Studies with Covariates." *Computational Statistics and Data Analysis* 47: 729–743. Doi: http://dx.doi.org/10.1016/j.csda.2003.11.010.

Zwane, E.N., and P.G.M. van der Heijden. 2007. "Analysing Capture-recapture Data When Some Variables of Heterogeneous Catchability Are Not Collected or Asked in All Registries." *Statistics in Medicine* 26: 1069–1089. Doi: http://dx.doi.org/10.1002/sim.2577.

Received December 2013
Revised October 2014
Accepted October 2014