Statistical Modelling 2003; **3**: 305–322

Point and interval estimation of the population size using the truncated Poisson regression model

Peter GM van der Heijden[1], Rami Bustami[1], Maarten JLF Cruyff[1], Godfried Engbersen[2] and Hans C van Houwelingen[3]
[1]Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands
[2]Faculty of Social Sciences, Erasmus University Rotterdam, Rotterdam, The Netherlands
[3]Department of Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands

Abstract: A method is presented to derive point and interval estimates of the total number of individuals in a heterogenous Poisson population. The method is based on the Horvitz–Thompson approach. The zero-truncated Poisson regression model is fitted and results are used to obtain point and interval estimates for the total number of individuals in the population. The method is assessed by performing a simulation experiment computing coverage probabilities of Horvitz–Thompson confidence intervals for cases with different sample sizes and Poisson parameters. We illustrate our method using capture–recapture data from the police registration system providing information on illegal immigrants in four large cities in the Netherlands.

Key words: capture–recapture; Horvitz–Thompson confidence interval; parametric bootstrap; population size estimation; truncated Poisson regression model

Data and Programs available from: http://stat.uibk.ac.at/SMIJ
Received September 2001; **revised** January 2003, May 2003; **accepted** May 2003

1 Introduction

Registration files can be used to generate a list of individuals from some population. If each time that an observation of a population member occurs is registered but, for one reason or another, some population members are not observed at all, the list will be incomplete and will show only part of the population. In this paper a method is presented to estimate the size of a population and its characteristics in terms of a number of covariates, from incomplete registration lists.

As an example we discuss the estimation of the number of illegal immigrants in The Netherlands from police records. These records contain information on the number of times each illegal immigrant was apprehended by the police and they are incomplete because the illegal immigrants who were never apprehended do not appear in them. Two other examples from the field of criminology that our method was applied to are the estimation of the size of the population of drunken drivers and illegal gun owners in

Address for correspondence: PGM van der Heijden, Department of Methodology and Statistics, Utrecht University, PO Box 80.140, 3508 TC Utrecht, The Netherlands. Email: p.vanderHeijden@fss.uu.nl

10.1191/1471082X03st057oa

The Netherlands based on police records on the number of apprehended individuals (Van der Heijden *et al.*, 2003). One could also think of noncriminal applications, such as the estimation of the number of individuals with a certain illness from a registration of doctor visits, the number of researchers working in a particular research area from a registration of published papers, the number of research groups working in a particular area from a registration of patents, or the number of potential clients in a hotel chain from a registration of visits of clients. All these applications have in common that, due to the nature of the registration data, a zero count cannot be observed and the data are truncated.

For the estimation of the number of illegal immigrants in The Netherlands (Van der Heijden *et al.*, unpublished manuscript, 1997), police records are available for 1995, for four cities in The Netherlands: Amsterdam, Rotterdam, The Hague and Utrecht. The records are used to derive count data on how often each illegal immigrant is apprehended by the police.

These count data can be considered as a special form of capture–recapture data. In traditional methods it is assumed that the count of every individual is generated by the same Poisson distribution. Let y_i be the number of times individual $i(i = 1, \dots, N_{\text{obs}})$ is apprehended $(y_i = 0, 1, \dots)$. Then the number of times he/she is apprehended follows a Poisson distribution

$$P(y_i|\lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!} \tag{1.1}$$

which is determined by the Poisson parameter $\lambda(\lambda > 0)$. Since we are using registration data, we do not know the number of individuals who are apprehended zero times, f_0, but we can estimate their number from the observed frequencies $f_k(k > 0)$ by assuming that f_k is generated by a truncated Poisson distribution. Once an estimate $\hat{f}_0$ is obtained, we are able estimate the size of the population $\hat{N}$ by adding $\hat{f}_0$ to the number of apprehended illegal immigrants.

In this paper we adopt a more general approach. We use the truncated Poisson regression model (see, for example, Cameron and Trivedi, 1998; Gurmu, 1991; Long, 1997; Winkelmann, 1997), in which the logarithm of the Poisson parameter λ is a linear function of a number of characteristics (covariates) known for an observed individual. The regression model allows individuals with different characteristics to have different Poisson parameters. In the statistical literature, this is referred to as observed heterogeneity, where the term 'observed' indicates that the Poisson parameter of an individual is not influenced by unobserved variables. Here, for each observed individual i, $\hat{f}_{0i}$ is estimated from his estimated Poisson parameter $\hat{\lambda}_i$ and added up to obtain $\hat{f}_0 = \sum_i f_0$, $i = 1, \dots, N_{\text{obs}}$.

The following assumptions are involved in estimating the population size. The first assumption is that the count of an individual is generated by a Poisson distribution. This assumption can be better understood by studying the genesis of the Poisson distribution (see Johnson *et al.*, 1993, for an overview). We present here a few elements. (i) The Poisson distribution with parameter λ is generated as the limit of a binomial distribution with success probability p and N realizations, where N tends to infinity and

p tends to zero, while Np remains finite and equal to λ. However, even for small N and small p the approximation of the Poisson distribution to the binomial distribution is already quite good, for example the Poisson distribution with $\lambda = 0.1$ approximates the binomial distributions with the $N = 3$ and $p = 0.033$ or for $N = 10$ and $p = 0.01$ quite closely. (ii) Johnson *et al.* (1993, referring to Charlier, 1905) note that the probability of success does not have to be constant for the Poisson limit to hold. (iii) A related result is that if X_1 is a realization of a Poisson distribution with Poisson parameter λ_1, and X_2 is a realization of a Poisson distribution with Poisson parameter λ_2, then $X_1 + X_2$ is a realization of a Poisson distribution with Poisson parameter $\lambda_1 + \lambda_2$. These three elements show that a Poisson distribution for an individual can be generated under a rather broad set of circumstances. Elements (ii) and (iii) show us that the probability of apprehension does not have to be constant over time. However, it should be noted that if the probability of (new) apprehension increases or decreases due to an apprehension or a nonapprehension, the independence structure of the binomial distribution is violated, and consequently so is that of the Poisson distribution. So, for example, if an illegal immigrant changes his behaviour following an apprehension and if as a result of that his probability to be apprehended decreases, the assumption of the Poisson distribution is violated. An extreme instance of such a violation is that the illegal immigrant is apprehended and forced to leave the country.

The second assumption in our approach to estimate the size of the population is that the Poisson parameters are linearly related to a set of observed covariates. This presupposes that one is able to measure the relevant covariates, and this is not necessarily the case. If there are any relevant covariates that are unobserved, the assumption is violated. However, this kind of violation can become evident as a result of analysing the data: in the Poisson regression model, the conditional variance is equal to the conditional mean, but relevant unobserved covariates result in overdispersion, that is, the situation that the conditional variance is larger than the mean. We return to this later.

For the example of the illegal immigrants the first assumption is severely violated if, once apprehended, they are forced to leave the country. If they are expelled effectively, they have a low probability to return and be apprehended again. However, in The Netherlands, illegal immigrants who are apprehended by the police often cannot be effectively expelled because either they refuse to mention their nationality, or their home country does not cooperate in receiving them back. In these cases the police requests them to leave the country, but it is unlikely that they will abide by this request.

In the 1995 police records of the abovementioned four large cities it was registered whether the apprehended illegal immigrant was effectively expelled or not. In the data that was available to us an illegal immigrant was coded as ‘effectively expelled’ when he was transported back to his home country by airplane, by boat or by car. An illegal immigrant was coded as ‘ineffectively expelled’ if his file reported he was ‘sent away’, ‘set out in a southern direction’ (being left at the border of Belgium, where they can re-enter The Netherlands immediately), ‘left with destination unknown,’ ‘we cannot solve this problem,’ or ‘transported by car into the direction of the border.’ Lastly, an illegal immigrant was coded as ‘other’ when he was brought to a centre for asylum seekers or a psychiatric hospital. For more details, see van der Leun *et al.* (1998). In total 4392 illegal immigrants were filed, 1880 of whom could not be effectively expelled, 2036 were effectively expelled, and for 476 illegal immigrants the reason was 'other' or missing in his file. The apprehension data are given in Table 1. Note that, although 'effectively expelled' illegal immigrants have a much lower frequency of re-apprehension, re-apprehension is still possible when these illegal immigrants re-enter The Netherlands after having been transported back to their home country.

For our analysis, we will consider illegal immigrants who were not effectively expelled (further abbreviated as IINEE), as for those the two assumptions discussed above are not *a priori* unrealistic. The following covariates were available to us: age, gender, country and reason for being apprehended. To give some insight in the data, we present the apprehension frequencies for each of the levels of the covariates in Table 2.

The zero-truncated Poisson regression model provides an estimator for f_0, the number of IINEE that were not apprehended by the police, and, by adding the IINEE that were actually apprehended, their total number in the population. The relevance of these estimators increases if their confidence interval is known. For simple truncated Poisson regression models (with categorical covariates), such confidence intervals have already been derived for subpopulations obtained by subdividing the data according to all categorical covariate combinations (see Zelterman, 2002).

In this paper we extend this work in a number of ways: by (1) proposing overall confidence intervals for the population size, (2) estimating those intervals by fitting the truncated Poisson regression model with covariates that can be both categorical as well as continuous, (3) using more parsimonious models so that we are not forced to incorporate all categorical covariate combinations, but can restrict our models to include, for example, main effects only, (4) studying characteristics of the whole population as well as of subpopulations (e.g., the probability that members of subpopulations are apprehended), and (5) assessing model fit. Extensions (1) to (3) are not trivial problems since we not only have to take into account individual sample fluctuations, but also the probability of an individual to be observed or not. The method that we use to solve this problem is based on the Horvitz–Thompson estimator (Kendall and Stuart, 1991, p. 173).

In section 2 we review traditional capture–recapture methods employing the homogeneous Poisson model to estimate the number of unobserved individuals in the population. In section 3 the Horvitz–Thompson method is presented and applied to the homogeneous Poisson model. The zero-truncated Poisson regression model is reviewed in section 4 and the Horvitz–Thompson point and interval estimation method for this model is presented in section 5. Assessment and performance of the method is done using a simulation experiment as described in section 6. Application and data analysis are presented in section 7. Section 8 is devoted to a brief and general discussion.

Table 1 Illegal immigrants data: observed frequencies for the three groups

Group	f_1	f_2	f_3	f_4	f_5	f_6	Total
Not effectively expelled	1645	183	37	13	1	1	1880
Effectively expelled	1999	33	2	1	1		2036
Other missing	430	41	5				476
Total	4074	257	44	14	2	1	4392

Table 2 Illegal immigrants not effectively expelled: observed frequencies for the covariate categories

Covariate category	f_1	f_2	f_3	f_4	f_5	f_6	Total
>40 years	105	6					111
<40 years	1540	177	37	13	1	1	1769
Female	366	24	6	1	1		398
Male	1279	159	31	12		1	1482
Turkey	90	3					93
North Africa	838	146	28	9	1	1	1023
Rest of Africa	229	11	3				243
Surinam	63	1					64
Asia	272	9	1	2			284
America, Australia	153	13	5	2			173
Being illegal	224	29	5	1			259
Other reason	1421	154	32	12	1	1	1621

2 Traditional capture–recapture methods

The zero-truncated Poisson distribution is defined by a probability function conditional on $y > 0$, that is

$$P(y_i|y_i > 0, \lambda) = \frac{P(y_i|\lambda)}{P(y_i > 0|\lambda)} = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!(1 - \exp(-\lambda))}, \qquad y_i = 1, 2, \ldots \tag{2.1}$$

with $p(y_i > 0 \mid \lambda) = 1 - \exp(-\lambda)$, $i = 1 \ldots, N_{\mathrm{obs}}$. An estimate $\hat{\lambda}$ for λ can be obtained by fitting the truncated Poisson distribution using methods discussed for truncated Poisson regression in section 4 (the truncated Poisson regression models should then only have an intercept and no covariates). The estimate $\hat{\lambda}$ can be used to estimate the probability of an individual not to be observed, $\hat{p}_0 = \exp(-\hat{\lambda})$. The number of unobserved individuals (individuals who were not apprehended but had a positive probability to be apprehended), is denoted by $\hat{f}_0$ and can be calculated as

$$\hat{f}_0 = \frac{\hat{p}_0}{1 - \hat{p}_0} N_{\mathrm{obs}}$$

where N_{obs} is the number of observed individuals in the sample.

3 Horvitz–Thompson point and interval estimation of the total number of individuals: homogeneous Poisson case

Consider the zero-truncated homogeneous Poisson model defined by (2.1). A point estimate for the total number of individuals in the population may be defined as (Kendall and Stuart, 1991, p. 173)

$$\hat{N} = \sum_{i=1}^{N} \frac{I_i}{p(\lambda)} \tag{3.1}$$

Here $I_i = 1$ if individual i is present and $I_i = 0$ otherwise, and $p(\lambda) = 1 - \exp(\lambda)$ is the probability of an individual to be present in the sample. The estimate $\hat{\lambda}$ is obtained from fitting the zero-truncated homogeneous Poisson model (2.1). This estimator is (approximately) unbiased if, conditional on $I_1, \ldots, I_N$, $1/p(\hat{\lambda})$ is an (approximately) unbiased estimator of $1/p(\lambda)$. This holds if the number of observations N_{obs} is not too small.

The variance of $\hat{N}$ can be decomposed in two parts

$$\mathrm{var}(\hat{N}) = E[\mathrm{var}(\hat{N}|I_1, \ldots, I_N)] + \mathrm{var}(E[\hat{N}|I_1, \ldots, I_N]) \tag{3.2}$$

The first term in (3.2) reflects the sampling fluctuation in the truncated Poisson distribution conditional on $I_1, \ldots, I_N$ and is estimated by $\mathrm{var}(\hat{N}|I_1, \ldots, I_N)$. This term $\mathrm{var}(\hat{N}|I_1, \ldots, I_N)$ can be estimated using the δ-method. This leads to

$$\hat{\mathrm{var}}(\hat{N}|I_1, \ldots, I_N) = \left(\sum_{i=1}^{N} I_i \frac{\partial}{\partial\lambda}\frac{1}{p(\lambda)}\right)^T \sigma^2(\lambda)\left(\sum_{i=1}^{N} I_i \frac{\partial}{\partial\lambda}\frac{1}{p(\lambda)}\right)\Bigg|_{\lambda=\hat{\lambda}} \tag{3.3}$$

where $\sigma^2(\lambda) = \mathrm{var}_\lambda(\hat{\lambda}|I_1, \ldots, I_N)$. (This notation allows for high-dimensional λ.)

For our case, we have $\sum_{i=1}^{N} I_i = N_{\mathrm{obs}}$. So (3.3) can be re-written as

$$\hat{\mathrm{var}}(\hat{N}|I_1, \ldots, I_N) = \left(N_{\mathrm{obs}} - \frac{\exp(-\hat{\lambda})}{(1 - \exp(-\hat{\lambda}))^2}\right)^2 \sigma^2(\hat{\lambda}) \tag{3.4}$$

The second term in (3.2) reflects the variation in the obtained sample. If the number of observations N_{obs} is not too small, we can safely assume that

$$E[\hat{N}|I_1, \ldots, I_N] = \sum_{i=1}^{N} \frac{I_i}{p(\lambda)} \tag{3.5}$$

The variance of this expression is given by

$$\mathrm{var}\left(\sum_{i=1}^{N} \frac{I_i}{p(\lambda)}\right) = \sum_{i=1}^{N} \frac{p(\lambda)(1 - p(\lambda))}{p^2(\lambda)} = \sum_{i=1}^{N} \frac{1 - p(\lambda)}{p(\lambda)} \tag{3.6}$$

In this expression, N is unknown. An unbiased estimator of the expression above using only the available observations, is given by

$$\sum_{i=1}^{N} I_i \frac{1 - p(\lambda)}{p^2(\lambda)} = N_{\mathrm{obs}} \frac{1 - p(\lambda)}{p^2(\lambda)} \tag{3.7}$$

Finally, using $p(\lambda) = 1 - \exp(-\lambda)$ and plugging in the estimated $\hat{\lambda}$ we obtain

$$\hat{\text{var}}(E[\hat{N}|I_1, \ldots, I_N) = N_{\text{obs}} \frac{\exp(-\hat{\lambda})}{(1 - \exp(-\hat{\lambda}))^2} \tag{3.8}$$

The variance of λ in (3.4), $\sigma^2(\lambda)$, is estimated from the derivatives of the log-likelihood of the truncated Poisson distribution. Consider a random sample $Y_1, \ldots, Y_{N_{\text{obs}}}$ from the truncated Poisson distribution with parameter λ. Then the log-likelihood is defined by

$$\ell = \sum_{i=1}^{N_{\text{obs}}} y_i \log \lambda - N_{\text{obs}}\lambda - N_{\text{obs}} \log(1 - \exp(-\lambda)) - \log \prod_{i=1}^{N_{\text{obs}}} y_i! \tag{3.9}$$

The estimated variance of λ is

$$\hat{\sigma}^2(\lambda) = -\left(\frac{\partial^2 \ell}{\partial \lambda^2}\right)^{-1}$$

The first derivative of the log-likelihood (3.9) w.r.t. λ is

$$\frac{\partial \ell}{\partial \lambda} = \sum_{i=1}^{N_{\text{obs}}} y_i \lambda^{-1} - N_{\text{obs}} - \frac{N_{\text{obs}} \exp(-\lambda)}{1 - \exp(-\lambda)}$$

and the second derivative is (after simplification)

$$\frac{\partial^2 \ell}{\partial \lambda^2} = -\sum_{i=1}^{N_{\text{obs}}} y_i \lambda^{-2} + \frac{N_{\text{obs}} \exp(-\lambda)}{(1 - \exp(-\lambda))^2}$$

So the estimated variance of λ is

$$\hat{\sigma}^2(\lambda) = -\left(\frac{\partial^2 \ell}{\partial \lambda^2}\right)^{-1} = \left(\sum_{i=1}^{N_{\text{obs}}} y_i \lambda^{-2} - \frac{N_{\text{obs}} \exp(-\lambda)}{(1 - \exp(-\lambda))^2}\right)^{-1} \tag{3.10}$$

So the total variance in (3.2) is now obtained from (3.4) and (3.6), that is,

$$\begin{aligned} \text{var}(\hat{N}) = {} & \left(N_{\text{obs}} \frac{\exp(-\lambda)}{(1 - \exp(-\lambda))^2}\right)^2 \left(\sum_{i=1}^{N_{\text{obs}}} y_i \lambda^{-2} - \frac{N_{\text{obs}} \exp(-\lambda)}{(1 - \exp(-\lambda))^2}\right)^{-1} \\ & + N_{\text{obs}} \frac{\exp(-\lambda)}{(1 - \exp(-\lambda))^2} \end{aligned} \tag{3.11}$$

For large values of N_{obs}, the variance of the maximum likelihood (ML) estimator of λ is estimated by (see Johnson *et al.*, 1993)

$$\hat{\sigma}^2(\lambda) \approx \lambda(1 - \exp(-\lambda))^2(1 - \exp(-\lambda) - \lambda\exp(-\lambda))^{-1}N_{\text{obs}}^{-1} \tag{3.12}$$

Note that the variances in (3.10) and (3.12) coincide at the ML estimate of λ, that is, when

$$\frac{\partial\ell}{\partial\lambda} = 0$$

which implies that

$$\sum_{i=1}^{N_{\text{obs}}} y_i = \frac{N_{\text{obs}}\lambda}{1 - \exp(-\lambda)}$$

Thus

$$\begin{aligned}\hat{\sigma}^2(\lambda) &= \left(\frac{N_{\text{obs}}\lambda}{1 - \exp(-\lambda)}\lambda^{-2} - \frac{N_{\text{obs}}\exp(-\lambda)}{(1 - \exp(-\lambda))^2}\right)^{-1}\\ &= \left(\frac{N_{\text{obs}}[1 - \exp(-\lambda) - \lambda\exp(\lambda)]}{\lambda(1 - \exp(-\lambda))^2}\right)^{-1}\end{aligned}$$

which is equal to (3.12). Expressions (3.1) and (3.11) for $\hat{N}$ and var($\hat{N}$), respectively, are computed by replacing the parameter λ in these expressions by its estimate $\hat{\lambda}$ obtained from fitting the zero-truncated homogeneous Poisson distribution (2.1). The total variance in (3.11) can be used to compute a 95% confidence interval for N: $\hat{N} \pm 1.96\text{SD}(\hat{N})$, with $\text{SD}(\hat{N}) = \sqrt{\text{var}(\hat{N})}$. In section 6 a simulation experiment is reported that assesses the stability of these variance estimators.

4 The zero-truncated Poisson regression model

Let $Y_1, \ldots, Y_{N_{\text{obs}}}$ be a random sample from the zero-truncated Poisson distribution with parameter λ_i, $i = 1, \ldots, N_{\text{obs}}$. Consider the regression model (Cameron and Trivedi, 1998)

$$\log(\lambda_i) = \boldsymbol{\beta}^T\boldsymbol{x}_i \tag{4.1}$$

where $\boldsymbol{\beta} = (\alpha, \beta_1, \ldots, \beta_p)^T$, and $\boldsymbol{x}_i$ is a vector of covariate values for subject i, that is $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^T$. The log-likelihood is given by

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{N_{\text{obs}}} [y_i\log(\lambda_i) - \lambda_i - \log(1 - \exp(\lambda_i)) - \log(y_i!)]$$

Model (4.1) can be fitted by maximizing the likelihood directly, for example using a Newton–Raphson procedure. This method is incorporated in Poisson regression routines where it is possible to specify truncation, for example in the computer program LIMDEP (Greene, 1995). We will elaborate on this method below. An alternative method that employs the EM-algorithm is proposed by Dietz and Böhning (2000). They show how standard software for Poisson regression can be employed to fit the truncated Poisson regression model. Interestingly, they obtain an estimate of the population size as a by-product of their estimation procedure.

The score function for the Newton–Raphson procedure is

$$U(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

The current value of the parameter vector $\boldsymbol{\beta}^{(t)}$ is updated by

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \boldsymbol{W}(\boldsymbol{\beta}^{(t)})^{-1}\boldsymbol{U}(\boldsymbol{\beta}^{(t)})$$

with $\boldsymbol{W}$ the observed information matrix, that is,

$$\boldsymbol{W}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \tag{4.2}$$

Fitting model (4.1) provides an estimator for the unknown parameter λ_i for the sampled individuals and thus for the probability to be present, $p(\lambda_i)$, $i = 1, \ldots, N_{\text{obs}}$.

5 Horvitz–Thompson point and interval estimation of the total number of individuals: heterogeneous Poisson case

The fit of model (4.1) can be used to derive the Horvitz–Thompson estimator for the total number of individuals in a heterogeneous Poisson population which is then defined by

$$\hat{N} = \sum_{i=1}^{N} \frac{I_i}{p(\boldsymbol{x}_i, \boldsymbol{\beta})} \tag{5.1}$$

where $I_i = 1$ if present and 0, otherwise. As in the homogeneous case, the variance of $\hat{N}$ is given by

$$\text{var}(\hat{N}) = E[\text{var}(\hat{N}|I_i)] + \text{var}(E[\hat{N}|I_i]) \tag{5.2}$$

The first term in (5.2) reflects the sampling fluctuation in the truncated Poisson distribution conditional on $I_1, \ldots, I_N$, and is estimated by $\text{var}(\hat{N}|I_1, \ldots, I_N)$. The latter can be estimated using the δ-method. This leads to

$$\hat{\text{var}}(\hat{N}|I_1, \ldots, I_N) = \left(\sum_{i=1}^{N_{\text{obs}}} \frac{\partial}{\partial \hat{\boldsymbol{\beta}}} \frac{1}{p(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})}\right)^T (\boldsymbol{W}(\hat{\boldsymbol{\beta}}))^{-1} \left(\sum_{i=1}^{N_{\text{obs}}} \frac{\partial}{\partial \hat{\boldsymbol{\beta}}} \frac{1}{p(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})}\right) \tag{5.3}$$

with $\boldsymbol{W}(\hat{\boldsymbol{\beta}})$ the observed information matrix obtained in (4.2), $p(\boldsymbol{x}_i, \boldsymbol{\beta}) = 1 - \exp(-\lambda_i) = 1 - \exp(-\exp(\boldsymbol{\beta}^T \boldsymbol{x}_i))$, the probability of an individual i to be observed in the sample, and

$$\sum_{i=1}^{N_{\text{obs}}} \frac{\partial}{\partial \boldsymbol{\beta}} \frac{1}{p(\boldsymbol{x}_i, \boldsymbol{\beta})} = \sum_{i=1}^{N_{\text{obs}}} \frac{-\boldsymbol{x}_i \exp(\log(\lambda_i) - \lambda_i)}{(1 - \exp(-\lambda_i))^2}$$

The second term in (5.2) reflects the variation in the obtained sample. Using the same argument as in section 3, we obtain

$$\text{var}(E[\hat{N}|I_1, \ldots, I_N]) \approx \text{var}\left(\sum_{i=1}^{N} I_i \frac{1}{p(\boldsymbol{x}_i, \boldsymbol{\beta})}\right) \tag{5.4}$$

which is estimated from the observed data by

$$\hat{\text{var}}(E[\hat{N}|I_1, \ldots, I_N]) = \sum_{i=1}^{N} I_i \frac{1 - p(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})}{p^2(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})} \tag{5.5}$$

Expression (5.1) for $\hat{N}$ is computed by replacing the parameter vector $\boldsymbol{\beta}$ by its ML estimate $\hat{\boldsymbol{\beta}}$ obtained from fitting the zero-truncated Poisson regression model (4.1). The total variance in (5.2) is estimated by adding expressions (5.3) and (5.5), and can be used to compute a 95% confidence interval for N: $\hat{N} \pm 1.96\text{SD}(\hat{N})$.

We have written a GAUSS-386i (GAUSS, version 3.2.8) procedure that fits the truncated Poisson regression model and computes Horvitz–Thompson point and interval estimates for the total number of individuals in the population.

The Horvitz–Thompson point estimate has the interesting property that removing covariates from a model tends to decrease the population size estimate. We prove this phenomenon for the simple model with a dichotomous covariate. We will show that this model will yield a larger estimated probability of a zero count, and hence a larger estimated population size than the model without the dichotomous covariate.

Assume a population is built up from two subpopulations of equal size, one with Poisson parameter $X = x + \delta$ and one with Poisson parameter $X = x - \delta$, with $\delta \neq 0$. In this population the probability of a zero count equals $E(\exp(-X)) = \exp(-x)[\exp(\delta) + \exp(-\delta)]/2$. Assume now a misspecified model for this population, namely the model where the fact that this population consists of two subpopulations is ignored.

Under this misspecified model the probability of a zero count equals $\exp(E(-X)) = \exp(-x)$. The result now follows by the convexity of the exponential function since

$$E[\exp(-X)] = \frac{\exp\{-(x-\delta)\} + \exp\{-(x+\delta)\}}{2} > \exp\left\{\frac{-(x-\delta)-(x+\delta)}{2}\right\}$$
$$= \exp(-x) = \exp(E[-X])$$

a special case of Jensen's inequality. This is illustrated in Figure 1. The figure also illustrates that the proof generalizes in a straightforward way to the situation where the subpopulations are not of equal size (the straight line is above the curved line). Also, the proof can be easily generalized to covariates with more than two categories, and the situation of more than one covariate.

In the data sets that we have analysed thus far, the removal of covariates also decreases the variance of the population size estimate, but we have no proof that this property holds in general.

Note that the expectation of the model with the covariate approaches that of the model without the covariate as δ approaches zero. So the smaller the difference in λ between the subpopulations, the less the effect on the population size estimate will be when the covariate discriminating between the subpopulations is omitted from the model. This also implies that covariates that contribute significantly to the fit of the model (which can be assessed with a chi-square test or a criterion like the AIC)

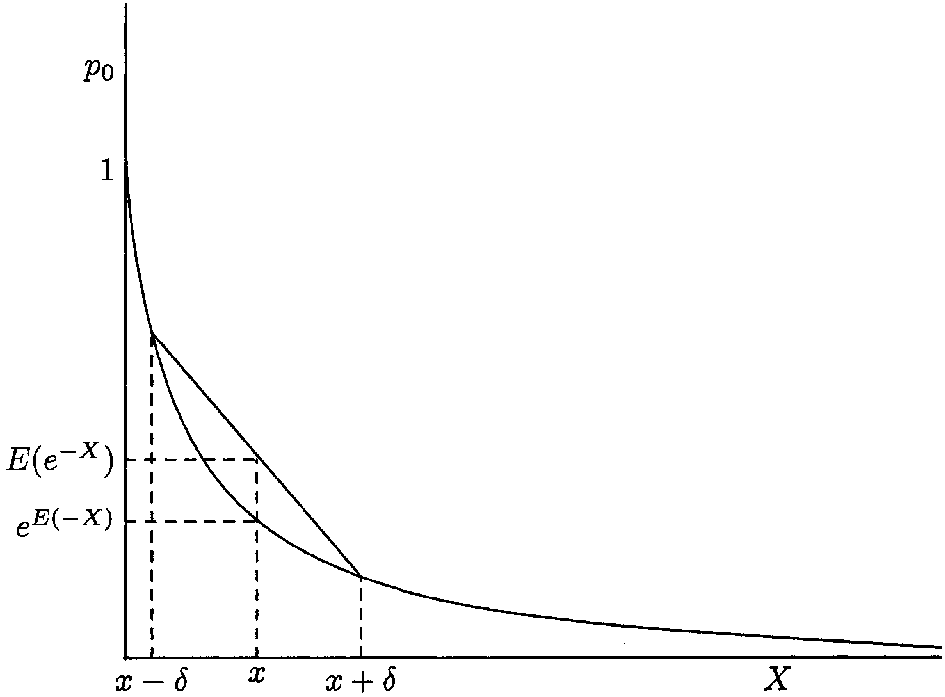

Figure 1 Expectation of the zero count for a model with and a model without a dichotomous covariate.

introduce observed heterogeneity in the model and therefore significantly reduce the amount of overdispersion. Nonsignificant covariates, however, do not significantly reduce the amount of overdispersion and may therefore be left out. For the process of model selection this means that the best model is the model with all significant covariates included and, for reasons of parsimony and smaller confidence intervals, all nonsignificant covariates left out.

Finally, from Jensen's inequality we may infer that, if the best model is still overdispersed, the population size estimate will be an underestimate of the true population size. To check if this is the case, the observed and the fitted marginal frequencies can be compared. For this purpose we use Pearson residuals, which are computed as (observed – fitted)/$\sqrt{}$(fitted). Also, a Lagrange multiplier test proposed by Gurmu (1991) can be used to test for overdispersion in the model as a result of unobserved heterogeneity. It compares the model fit of the Poisson model with alternative models with an extra dispersion parameter included, such as the negative binomial regression model. The test statistic is chi-square distributed with one degree of freedom.

6 A simulation experiment

To assess the performance of the Horvitz–Thompson method, an experiment was carried out to investigate the coverage probability of the Horvitz–Thompson confidence interval (Table 3). At the same time we evaluated the coverage probability of the confidence interval obtained by using parametric bootstrapping (see, for example Efron and Tibshirani, 1993). The experiment is performed using a homogenous Poisson model (with intercept only), and is carried out as follows:

1) A sample of size $N = 100, 250, 500, 1000$ is drawn from a nontruncated homogenous Poisson distribution with parameters $\lambda = 0.5, 1, 1.5, 2, 2.5$.
2) After omitting the zero count, for each of the above 20 observed samples of size N_{obs}, an EM-algorithm is applied to fit a truncated homogenous Poisson distribution to obtain an estimate $\hat{f}_0$ for f_0, the zero-count, as well as an estimate $\hat{\lambda}$ for the Poisson parameter λ. Thus, $\hat{N} = N_{\text{obs}} + \hat{f}_0$.

Table 3 Coverage probabilities of Horvitz–Thompson (HT) 95% confidence intervals and 95% confidence intervals generated from 500 parametric bootstrap samples (Boot)

(λ, N)	100	250	500	1000
0.5	*HT*: 0.93 *Boot*: 0.88	*HT*: 0.96 *Boot*: 0.89	*HT*: 0.95 *Boot*: 0.92	*HT*: 0.94 *Boot*: 0.95
1	*HT*: 0.96 *Boot*: 0.92	*HT*: 0.95 *Boot*: 0.95	*HT*: 0.94 *Boot*: 0.93	*HT*: 0.95 *Boot*: 0.93
1.5	*HT*: 0.95 *Boot*: 0.94	*HT*: 0.94 *Boot*: 0.95	*HT*: 0.96 *Boot*: 0.95	*HT*: 0.95 *Boot*: 0.96
2	*HT*: 0.96 *Boot*: 0.96	*HT*: 0.96 *Boot*: 0.95	*HT*: 0.95 *Boot*: 0.96	*HT*: 0.95 *Boot*: 0.95
2.5	*HT*: 0.97 *Boot*: 0.96	*HT*: 0.94 *Boot*: 0.94	*HT*: 0.96 *Boot*: 0.95	*HT*: 0.92 *Boot*: 0.93

3) Horvitz–Thompson 95% confidence intervals are computed.
4) For each of the above 20 observed samples, 500 bootstrap samples are drawn from a nontruncated homogenous Poisson distribution with $\hat{N}$ and $\hat{\lambda}$ obtained in 2). 95% bootstrap confidence intervals are obtained using the percentile method. Note that by drawing samples of size $\hat{N}$ from a nontruncated distribution instead of drawing samples of size N_{obs} from a truncated distribution, we take into account that there are two sources for the variance of $\hat{N}$ (see (3.2)).
5) Steps 1 to 4 are repeated 500 times.
6) Coverage probabilities were calculated as the proportion of confidence intervals containing the original sample size N. These probabilities were obtained for both the Horvitz–Thompson confidence interval and the bootstrap confidence interval.

The results summarized in Table 3 indicate that the Horvitz–Thompson confidence interval has a higher coverage probability than that of the bootstrap confidence interval when both λ and N are small ($\lambda = 0.5$ and $N = 100$, 250). For other values of λ and N, bootstrap confidence intervals and Horvitz–Thompson confidence intervals are comparable.

In general, the simulation results indicate that the Horvitz–Thompson confidence interval performs well for different values of N and λ in the homogeneous Poisson case. It is likely that the method also works well if there are (a limited number of) categorical covariates as considered in this paper. The total population size is the sum of the population sizes for each covariate pattern. In a saturated model, each pattern has its own Poisson parameter and the estimated population sizes per pattern are completely independent. The variance of the sum is equal to the sum of the variances. Since the simulation results show the validity of the variance estimate for each pattern, it is *a fortiori* valid for the total. Modelling the Poisson parameters via a regression model with categorical covariates can only help to improve the estimator and reduce the variance. It is like combining patterns with similar probabilities. Therefore, we are confident that the simulation results also carry over to this situation.

The situation gets more complicated if there are many possible covariate patterns. The Poisson parameter for rare patterns can become very hard to estimate. This can be detrimental if the Poisson parameter λ is small, indicating a small capture probability. How to handle this situation is the subject of further research.

Similar arguments hold for continuous covariates. If there are no outlying patterns, regression models could be thought of as partitioning the covariate space in homogeneous subgroups. For each subgroup the simulation results apply and hence we trust that our procedure is also valid for the total sample size. The situation is different if there are extreme patterns. The Poisson parameters for such patterns might be hard to estimate and sensitive to the regression model used. Problems may arise if such patterns have low capture probability. Our simulation study does not quite cover this situation. It is also the subject of further research.

7 Data analysis

Consider the IINEE data described in section 1. The response of interest is the number of times an individual is apprehended by the police. The following four variables were available as covariates in the truncated Poisson regression model: *nationality* (N), *gender* (G), *age* (A) and *reason* for being apprehended (R) (compare Table 2) (unfortunately, we do not have continuous covariates in this example, but the method proposed in this paper applies in the same way). We recoded the variable *nationality*, which had six categories, by creating five dummy variables considering America and Australia as the reference category.

Table 4 shows the results of the zero-truncated Poisson regression analysis on the full model, containing all four available covariates. It shows the maximum likelihood estimates of the regression parameters together with their corresponding standard errors and P-values. The variables *gender*, *age* and *nationality* (Turkey, Rest of Africa, Surinam or Asia) contribute significantly to the average number of times an individual is apprehended by the police. The results show that male individuals and individuals who are less than 40 years of age are, on average, more frequently apprehended by the police. Individuals from Turkey, rest of Africa, Surinam and Asia are less frequently apprehended than those from America and Australia. The variable *reason* for being apprehended appears to have no impact on the average number of times an individual is apprehended by the police.

For the purpose of model selection, we fitted several truncated Poisson regression models. We used expression (5.1) to obtain the point estimates of the total number in the population and the variance in (5.2) to compute the 95% confidence intervals. The results are shown in Table 5. The null model yields the lowest estimate of the total number of IINEE ($\hat{N} = 7080$). The corresponding 95% Horvitz–Thompson confidence interval is (6363, 7797). The largest estimate of N, $\hat{N} = 12\,691$, is obtained by fitting the full model of Table 5. These estimates illustrate the theoretical result of section 5 that, in a sequence of nested models, the more covariates that are added to the model, the higher the point estimate of N is expected to become.

In order to compare the various models we also computed AIC- values and performed likehood-ratio tests for the models in Table 5. Just as in other areas of statistical modelling, model selection is a difficult issue, and more research is needed to find out which approach is best in the area we are discussing here. However, in line with

Table 4 Truncated Poisson regression model fit to the IINEE data

Regression parameters	MLE	SE	P-value*
Intercept	−2.317	0.449	
Gender (male = 1, female = 0)	0.397	0.163	0.015
Age (<40 years = 1, >40 years = 0)	0.975	0.408	0.017
Nationality			
(Turkey)	−1.675	0.603	0.006
(North Africa)	0.190	0.194	0.328
(Rest of Africa)	−0.911	0.301	0.003
(Surinam)	−2.337	1.014	0.021
(Asia)	−1.092	0.302	<0.001
(American and Australia)	0.000		
Reason (being illegal = 1, other reason = 0)	0.011	0.162	0.946

Log-likelihood = −848.448.
*P-value for Wald test.

Table 5 Estimated $\hat{N}$ and *HT* 95% confidence intervals for *N* obtained from fitting different truncated Poisson regression models. Model comparisons using the likelihood-ratio test and AIC criterion are also given. $\chi^2_{(1)}$ is the Lagrange multiplier test testing for overdispersion

Model	AIC	G^2	df	P*	$\chi^2_{(1)}$	$\hat{N}$	CI
Null	1805.9				106.0	7080	6363–7797
G	1798.3	9.6	1	0.002	99.7	7319	6504–8134
G+A	1789.0	11.2	1	<0.001	93.7	7807	6637–8976
G+A+N	1712.9	86.1	5	<0.001	55.0	12690	7186–18194
G+A+N+R	1714.9	0.004	1	0.949	55.0	12691	7185–18198

**P*-value for likelihood-ratio test. HT = Horvitz–Thompson.

the theoretical result just illustrated, in order to obtain an accurate estimate for N it is crucial that covariates that take into account a significant amount of the heterogeneity are included in the final model. Therefore one could start with all available covariates (the full model $\mathrm{G}+\mathrm{A}+\mathrm{N}+\mathrm{R}$) and then drop covariates that turn out not to be significant (backward elimination). A model choice can then be based on likelihood-ratio tests (which can be used if the models are nested), or the AIC criterion (which can also be used when the models are non-nested).

The likelihood-ratio test in Table 5 shows that the variable *reason* for being caught can be dropped from the full model ($G^2 = 0.004$, $\mathrm{df} = 1$, $P = 0.949$). From the resulting model ($\mathrm{G}+\mathrm{A}+\mathrm{N}$) the variable *nationality* cannot be dropped ($G^2 = 86.1$, $\mathrm{df} = 5$, $P < 0.001$), nor can the variables *gender* and *age* (not shown here). Since the AIC criterion also favours this model and our choice of $\hat{N}$ should be based on the best-fitting model, our best estimate seems to be that of the model ($\mathrm{G}+\mathrm{A}+\mathrm{N}$), $\hat{N} = 12\,690$. When models are misspecified (e.g., the null model and the models in the second and the third row of Table 5) their results, including the value of $\hat{N}$, should not be interpreted.

A way of examining the goodness of fit of a model is to compare the observed and the estimated frequencies by looking at the Pearson residuals, as presented in Table 6. The residuals for $k = 2$, $k = 4$ and $k = 6$ seem rather large, indicating some lack of fit. The Lagrange multiplier test of Gurmu (1991) (see section 5) suggests that there still remains some unobserved heterogeneity that cannot be ignored ($\chi^2 = 55.0$, $\mathrm{df} = 1$). Therefore we must conclude that the population size estimate $\hat{N} = 12\,690$ should be interpreted as an underestimate of the true population size. It is an object of further

Table 6 Observed and estimated counts for illegal immigrants for model ($\mathrm{G}+\mathrm{A}+\mathrm{N}$)

k	Observed	Estimated	Residuals
0	0	10810.4	
1	1645	1612.6	0.81
2	183	233.7	−3.32
3	37	30.1	1.25
4	13	3.2	5.42
5	1	0.3	1.31
6	1	0.0	6.57

Table 7 Comparisons between observed and estimated *N* for subgroups based on model (G + A + N)

Subgroup	Observed	Estimated	Observed/Estimated
Males	1482	8880.10	0.167
Females	398	3811.40	0.104
Individuals with *age* <40 years	1769	10506.72	0.168
Individuals with *age* >40 years	111	2184.73	0.051
Individuals from Turkey	93	1740.03	0.053
Individuals from North Africa	1023	3055.23	0.335
Individuals from Rest of Africa	243	2058.00	0.118
Individuals from Surinam	64	2387.75	0.027
Individuals from Asia	284	2741.96	0.104
Individuals from America and Australia	173	708.47	0.244
Individuals caught for reason Being illegal	259	1631.68	0.159
Individuals caught for Other reason	1621	11509.77	0.147

study whether the sign and the size of specific residuals are indications of the size of the discrepancy between the population size estimate and the true population size.

It is also possible to make comparisons between observed and estimated number of individuals for subgroups in the data. Table 7 shows such comparisons based on the model fit of model (G + A + N). Note that for all subgroups the Horvitz–Thompson estimate of the number of individuals is much larger than the number of individuals observed in the data. This indicates that the probability that illegal individuals are not apprehended is high for all subgroups in the population. Moreover, it is clear that male individuals, individuals who are less than 40 years of age, individuals from North Africa have larger probability to be apprehended, a confirmation of what was observed in Table 4.

8 Discussion

The Horvitz–Thompson method was presented to estimate the total number of individuals in a heterogeneous Poisson population. The truncated Poisson regression model was utilized to estimate f_0, the number of individuals who were not apprehended by the police, but have a positive probability to be apprehended. The Horvitz–Thompson method was assessed using a simulation experiment evaluating a truncated homogeneous Poisson distribution and for this situation it was proved to be appropriate (it is likely that these results generalize to Poisson regression, but some specific questions still need to be investigated; see section 6). It was proven that in a series of nested models, models including more significant covariates yield a larger point estimate for the population size.

In this approach the following assumptions were made. First, it is assumed that each individual count is a realization of a Poisson distribution. In principle, this assumption holds under rather broad circumstances as explained in section 1. The main threat to this assumption is that, due to an apprehension or a nonapprehension, the probability of the next apprehension increases or decreases. For example, an illegal immigrant who is never apprehended might become careless, thus increasing his probability to be apprehended, or, once apprehended he might become more careful so that his probability to be apprehended decreases. It may also be that an apprehension influences the behaviour of police officers, in the sense that the probability of a next apprehension increases. Whether any of these examples is true or not is a topic that needs to be investigated separately.

The second assumption is that the logarithm of each individual Poisson parameter is a function of covariates, or in other words, that the covariates are sufficient to describe the individual differences in the Poisson parameters. However, it will often occur that not every relevant covariate is measured so that this assumption is violated. Whether this is the case can be checked from the analysis. First, a Lagrange multiplier test discussed in section 5 tests if there remains unobserved heterogeneity, that is, heterogeneity that is additional to that taken into account by the available covariates. Second, the observed counts may deviate from the fitted counts. If it turns out that the second assumption is violated, the point estimate of the population size is to be interpreted as a lower bound.

In the case where the second assumption is violated, the difference between the lower bound estimate and the true population size is not known. One way to investigate this is to use a model that accounts for unobserved heterogeneity (overdispersion) between individuals, such as the zero-truncated negative binomial regression model, not used in this work. Such models take into account other sources of heterogeneity between individuals that are not observed in the data in terms of covariates. The zero-truncated negative binomial model incorporates overdispersion (which is accounted for by including an additional parameter α in the model) in the sense that the truncated variance of the negative binomial exceeds the truncated variance of the Poisson. The Poisson is a limiting case of the negative binomial and is obtained as $\alpha \to 0$ (see Grogger and Carson, 1991; Greene, 1997, for more details). An implementation of the Horvitz–Thompson method to results from fitting such models will also be the subject of a future publication.

Acknowledgement

The authors wish to thank the Dutch Ministry of Justice for their financial support and for making the police registration data available for statistical research.

References

Cameron AC, Trivedi P (1998) *Regression analysis of count data.* Cambridge: Cambridge University Press.

Charlier CVL (1905) Die zweite Form des Fehlergesetzes. *Arkiv fur Matematik, Astronomi och Fysik*, **15**(2), 1–35.

Dietz E, Böhning D (2000) On estimation of the Poisson parameter in zero-modified Poisson models. *Computational Statistics and Data Analysis*, **34**, 441–59.

Efron B, Tibshirani R (1993) *An introduction to the bootstrap.* London: Chapman and Hall.

Greene WH (1995) *LIMDEP version 7.0, user's manual.* New York: Econometric Software, Inc.

Greene WH (1997) *Econometric analysis*, third edition. New Jersey: Prentice-Hall International, Inc.

Grogger JT, Carson RT (1991) Models for truncated counts. *Journal of Applied Econometrics*, **6**(2), 225–38.

Gurmu S (1991). Tests for detecting overdispersion in the positive Poisson regression model. *Journal of Business & Economic Statistics*, **9**(2), 215–22.

Johnson NL, Kotz S, Kemp AW (1993) *Univariate discrete distributions*, 2nd edition. New York: Wiley.

Kendall M, Stuart A (1991) *Advanced theory of statistics*, 2nd edition. London: Charles Griffen & Company Limited.

Long J (1997) *Regression models for categorical and limited dependent variables.* California, USA: Sage.

Van der Heijden PGM, Zelterman D, Engbersen G, van der Leun J (1997) Estimating the number of illegals in the Netherlands with the truncated Poisson regression model. Unpublished manuscript.

Van der Heijden PGM, Cruyff, MJLF, Van Houwelingen HC (2003) Estimating the size of a criminal population from police records using the truncated Poisson regression model. *Statistica Neerlandica*, **57**(3), 1–16.

van der Leun J, Engbersen en Peter van der Heijden G (1998) *Illegaliteit en criminaliteit: schattingen, aanhoudingen en uitzettingen.* Rotterdam: Erasmusuniversiteit, vakgroep Sociologie.

Winkelmann R (1997) *Econometric analysis of count data.* Berlin, Heidelberg: Springer-Verlag.

Zelterman D (2002) *Advanced log-linear models using SAS.* Cary, NC: SAS Institute.