Statistica Neerlandica (2003) Vol. 57, nr. 3, pp. 289–304

# Estimating the Size of a Criminal Population from Police Records Using the Truncated Poisson Regression Model

Peter G.M. van der Heijden*

*Faculty of Social Sciences, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands*

Maarten Cruyff

*Faculty of Social Sciences, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands*

Hans C. van Houwelingen

*Department of Medical Statistics, Leiden University Medical Center, P.O. Box 9600, 2300 RC Leiden, The Netherlands*

The truncated Poisson regression model is used to arrive at point and interval estimates of the size of two offender populations, i.e. drunk drivers and persons who illegally possess firearms. The dependent capture–recapture variables are constructed from Dutch police records and are counts of individual arrests for both violations. The population size estimates are derived assuming that each count is a realization of a Poisson distribution, and that the Poisson parameters are related to covariates through the truncated Poisson regression model. These assumptions are discussed in detail, and the tenability of the second assumption is assessed by evaluating the marginal residuals and performing tests on overdispersion. For the firearms example, the second assumption seems to hold well, but for the drunk drivers example there is some overdispersion. It is concluded that the method is useful, provided it is used with care.

*Key Words and Phrases:* capture–recapture, truncated Poisson regression, overdispersion, population size estimation.

## 1 Introduction

For many policy reasons, it is important to know the size of specific delinquent populations. One reason is that it provides insight into the threat these populations may pose on society. Another reason is that it gives an estimate of the workload of the police.

*P.vanderHeijden@fss.uu.nl.

However, estimating the size of a delinquent population may be problematic for various reasons. Counting the number of crimes from police records may lead to a dark number problem. It may be that the crime is registered but the offender is not known or, as is often the case with victimless crimes, the crime is not registered at all. In victim surveys people report the number of times they have been the victim of a particular crime, such as robbery or burglary. Based on that information an estimate can be obtained of the total number of these crimes. However, victim surveys do not provide an estimate of the number of offenders, since they usually are unknown to the victim, nor do they provide insight into victimless offences. Self-report studies can potentially estimate the size of a delinquent population since people are simply asked whether they are a member of this type of population. Problems related to self-report studies are: the difficulty of obtaining a representative sample, the risk of socially desirable answers and the need for large samples if offences are infrequent. For a more elaborate comparison and an overview of the literature on police registrations, victim surveys and self-report studies, we refer the reader to WITTEBROOD and JUNGER (2002).

In this paper we discuss a way of estimating the number of offenders from police data. The data we use are from the Dutch police registration system HKS (Herkennings Dienst Systeem). Offences committed by a known offender are registered in this system. Each report of an offence has an offender identification, so it is possible to construct an offender-based data set. Let's say we want to estimate the number of specific type of offenders, such as people who illegally possess guns. We can construct an offender-based data set that will have the number of offenders apprehended once in possession of a gun, the number apprehended twice, three times, and so on. Note that offenders who were never apprehended are not part of this offender-based data set. Yet, if we could estimate their number we would have an estimate of the total number of illegal gun owners.

The aim is to estimate the number of offenders never apprehended, using the data about offenders apprehended at least once. We derive these estimates under two assumptions. First, the number of apprehensions is a realization of a Poisson distribution. Second, the logarithm of the Poisson parameter for an offender is a linear function of covariates. We discuss these assumptions in greater detail at the end of the introduction.

At this point we want to indicate how we can estimate the size of the population never apprehended using these assumptions. Say we take an offender whose Poisson parameter specifies his probability of being apprehended at least once at 0.25, nonetheless he is apprehended. This implies that, for this one offender, we estimate that there are three other offenders who have not been apprehended. By performing this trick for every offender who is apprehended, and adding up all the individual estimates, we obtain an estimate of the total number of offenders who are not apprehended, and this solves our problem.

The methods employed in this paper originate from the field of biology, where they are used to estimate animal abundance. The data we use are a specific form of capture–recapture data. In capture–recapture data captures and recaptures are made

at specific time points, and for each animal seen at least once there is a capture history. For example if there are five capture times, a history could be 01101 if the animal is seen at captures 2, 3 and 5 and not seen at captures 1 and 4. Here, we use only the total number of times someone is captured since we are collecting data in continuous time. Typically, in the biological application area, covariate information is not available or not used, leading to a basic model in which the Poisson parameters are assumed to be homogenous over the animals. For an overview of this area, we refer the reader to SEBER (1982, chapter 4; 1986), CHAO (1988) and ZELTERMAN (1988, 2001, chapter 7). In the statistical literature, this problem is also known as the estimation of the size of a truncated sample (SANATHANAN, 1977), the estimation of the number of (unseen) species (EFRON and THISTED, 1976; BUNGE and FITZPATRICK, 1993), or the estimation of the size of a population using samples of size one (WILSON and COLLINS, 1992).

In criminology there are some early studies by GREENE and STOLLMACK (1981), who use arrest data to estimate the number of adults committing felonies and misdemeanors in Washington D.C. in 1974/5, ROSSMO and ROUTLEDGE (1990) who estimate migrating (or fleeing) fugitives in 1984, and prostitutes in 1986/7, both in Vancouver, and COLLINS and WILSON (1990) who use arrest data to estimate the number of adult and juvenile car thieves in the Australian capital territory in 1987. None of these early studies devote systematic attention to covariate information on the apprehended individuals or to confidence intervals for the point estimates of the number of individuals who are not apprehended. Our early statistical work in criminology is also in this vein (VAN DER HEIJDEN *et al.*, 1993; SMIT *et al.*, 1993).

In more recent work, we have incorporated covariate information by using the truncated Poisson regression model, which is well known in econometrics (GREENE, 1997, chapter 19; CAMERON and TRIVEDI, 1998, chapter 4; LONG, 1997, chapter 8). We have developed a method of estimating a frequency for the zero-count and a confidence interval for this point estimate (VAN DER HEIJDEN *et al.*, in press). Thus far we have experience with this method in estimating the number of illegal immigrants in the four largest Dutch cities in 1995 (see VAN DER LEUN *et al.*, 1998; see HOOGTEIJLING, 2002, for a critical evaluation of this method) and in the Netherlands in general from 1997 to 2000 (see ENGBERSEN *et al.*, 2002), and with the number of opiate users in Rotterdam in 1994 (SMIT, TOET and VAN DER HEIJDEN, 1996). In this paper we present two further examples, i.e. on the number of person who illegally own firearms and the number of car drivers under the influence of alcohol.

### 1.1 *Assumptions*

As the methodology originates from the field of biology and we are using it in the field of criminology, we discuss the assumptions of the methodology in greater detail. Obviously, assumptions that are realistic for animals may not also be realistic for human offenders.

The first assumption is that the number of times an individual is apprehended is a realization of a Poisson distribution. JOHNSON, KOTZ and KEMP (1993) discuss the

genesis of the Poisson distribution and state that it was originally derived by Poisson 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 $\lambda$. It turns out that even for small $N$, the Poisson distribution approximates the binomial distribution reasonably well if $p$ is sufficiently small. For example, for $N = 3$ and $p = 0.033$, and for $N = 10$ and $p = 0.01$, the probabilities of counts 0, 1, 2 and 3 are already very close to those of a Poisson distribution with $\lambda = 0.1$, and similarly for $N = 3$ and $p = 0.0033$, and for $N = 10$ and $p = 0.001$ with $\lambda = 0.01$.

Referring to CHARLIER (1905), JOHNSON *et al.* (1993) note that the probability of success $p$ does not have to be constant for the Poisson limit to hold. So, generally speaking, it follows for the type of applications we are discussing that individuals do not need to have a constant probability to be apprehended, but it suffices if they could be apprehended a number of times. A property of the Poisson distribution related to the result of CHARLIER (1905) is that if $X_1$ is a realization of a Poisson distribution with Poisson parameter $\lambda_1$, and $X_2$ is a realization of a Poisson distribution with Poisson parameter $\lambda_2$, then $X_1 + X_2$ is a realization of a Poisson distribution with Poisson parameter $\lambda_1 + \lambda_2$. So again we see that the probability of being apprehended need not be constant: if we split up the full period of data collection into a larger number of sub-periods and in each of these sub-periods the count is generated by some Poisson distribution, then the sum of these counts will also be generated by a Poisson distribution. For drunk driving this means an individual does not always have to be drunk when he is driving, but it suffices that this happens at least three times in the period of data collection. Similarly, someone who illegally owns a gun does not need to have it with him all the time, a small number of times suffices to consider his count to be generated by a Poisson distribution.

We note that the Poisson assumption is valid only if a change in the individual Poisson parameter is unrelated to any prior apprehensions or non-apprehensions. This follows from the independence of subsequent trials in a binomial distribution. For example, if someone who illegally owns an gun is apprehended and subsequently buys a new one but carries it around less often or does not buy a new one at all, the resulting change in capture probability is a violation of the Poisson assumption. Similarly, if someone is apprehended with a gun and the police then keep him under close surveillance, the increase in capture probability is a result of the apprehension and the Poisson assumption is violated. Or if someone who is driving under the influence of alcohol is not caught and starts doing so more frequently, his probability increases as a function of not being apprehended and such an increase violates the Poisson assumption. In the biostatistical literature, this is known as positive contagion (if the probability increases) or negative contagion (if the probability decreases).

Closely related to the contagion issue there is the problem of an open or closed population. A population is closed if the number of offenders is constant over the period of data collection, and is open if offenders may enter or leave the population

during this period. Given what has been noted above, it is clear that the population may be open as long as entering or leaving it is not related to apprehension or non-apprehension. For example, detention following an apprehension removes the person from the population and excludes the possibility of any subsequent apprehensions, and can therefore be seen as an extreme case of negative contagion.

So far we have only discussed the Poisson assumption pertaining to an individual count. The second assumption follows from using a regression model, in which the logarithm of the Poisson parameters is a linear function of covariates. In the regression model, the Poisson parameters are still assumed to be homogeneous for individuals with identical values on the covariates, but they are allowed to be heterogeneous for individuals with different values. Since the differences in Poisson parameters are assumed to be completely determined by the observed covariates, this is referred to as observed heterogeneity. So even if the count of every individual is Poisson distributed, the assumption of Poisson regression is violated if, in addition to observed heterogeneity, there are differences in the Poisson parameters that cannot be explained by the observed covariates. This additional heterogeneity is called unobserved heterogeneity.

In the Poisson regression model that only has observed heterogeneity, the conditional mean (i.e. the mean conditional on the covariates) is equal to the conditional variance, whereas if there is unobserved heterogeneity, the conditional variance is larger than the conditional mean. This is referred to as overdispersion.

In conclusion, the most important violations of the Poisson assumptions in criminological applications are contagion and overdispersion.

The contagion problem will probably be larger for some offences than for others, and an indication of its importance can be obtained by studying the behavior of offenders as well as police officers. If no additional information is available on their behavior, it seems best to interpret the results with caution. Overdispersion can be assessed in the data as a result of the analysis, and we return to this and its interpretation in section 2.2.

We start with an introduction to the data we use for the analysis, i.e. police records, and then we discuss zero-truncated Poisson regression. This is followed by the two examples and we end with a critical evaluation of our method.

## 2 Method

### 2.1 *Police records*

Since the early 1980s the Dutch police have registered all violations against more than 70 different criminal laws, provided the offender is known. From these records, files over the 1996 to 2001 period have been made available to us for violations against the laws regarding drunk driving and illegal ownership of firearms. These files present two problems. First, the quality of administration varies considerably over the 25 Dutch police regions. This is why we selected for the analyses the five

police regions reported to maintain the highest administrative quality. These regions are Rotterdam Rijnmond, Mid-Gelderland, Mid-Holland, South Holland South and Mid-West Brabant. It is unclear whether these five police regions are representative of the 25 Dutch police regions. Secondly, due to the police registration behavior using the registration system to help them find suspects (and not to conduct statistical analyses), the registration files contain double entries for the same apprehension. Since double entries result in distorted frequency distributions, population estimates based on these files would be biased. Fortunately, at the time of our study the police had just completed a version of the registration files in which double entries were eliminated as far as possible.

The registration files contain all the violations in the 1996 to 2001 period for drunk driving or illegal possession of firearms. We constructed dependent capture–recapture variables by counting the number of times each person was apprehended for drunk driving and illegal ownership of firearms, respectively. The dependent variables are computed over fixed periods of time ranging from one to five years. For illegal ownership of firearms we present an analysis of the data for the 1998 to 1999 period, and for drunk driving for the year 2000.

We used the background variables age, gender, age of first offence and police region as covariates. From the number of apprehensions for the remaining criminal offences, we constructed six covariates to measure each person's criminal history. These covariates have been constructed according to a standard categorization employed by Statistics Netherlands (see HULS *et al.*, 2000), distinguishing between violations related to violence, hard drugs, property, vandalism, traffic and violations of special "economic" laws (e.g. laws regulating working and environmental conditions). We computed the covariates by adding the number of apprehensions for the remaining offences to the corresponding covariate. The covariates are measured over the five-year period before the last measurement year of the dependent variable. The five-year period was taken to ensure comparability of criminal histories, since minor offences are deleted from the files after five years without any further registration.

Since the distributions of the six covariates measuring criminal history were very skewed, we transformed them before entering them into the regression model. We used the transformation $\log(1+x)$, since preliminary analysis showed that this logarithmic transformation led to better predictions of the dependent variables.

Only the "age of first offence" variable had missing values, about one half percent of the total number of observations. These values were imputed by the age of the subject minus the mean difference between age and age of first offence in the sample.

### 2.2 *Zero-truncated Poisson regression*

We start by introducing the zero-truncated Poisson distribution, and then work out zero-truncated Poisson regression.

Consider the data in Table 1. We denote the number of individuals apprehended $k$ times by $f_k$ ($k = 1,\ldots,K$). Let $y_i$ be the number of times individual $i$ ($i = 1,\ldots,N_{obs}$) is

Table 1. Observed and estimated counts for illegal possession of firearms (left panel) and for drunk driving (right panel).

| $k$ | Observed | Estimated | Residuals | Observed | Estimated | Residuals |
|---|---|---|---|---|---|---|
| 0 | 0 | 60,084.0 | – | 0 | 104,352.0 | – |
| 1 | 2,561 | 2,558.9 | 0.04 | 8,877 | 8,847.2 | 0.32 |
| 2 | 72 | 76.4 | −0.50 | 481 | 534.4 | −2.31 |
| 3 | 5 | 2.6 | 1.48 | 52 | 34.0 | 3.08 |
| 4 | – | – | – | 8 | 2.9 | 2.98 |
| 5 | – | – | – | 1 | 0.4 | 1.06 |

apprehended ($y_i = 0,1,\ldots$). Under the Poisson assumptions the probability that an individual is apprehended a specific number of times

$$P(y_i|\lambda) = \frac{\exp(-\lambda)\lambda^{y_i}}{y_i!}, \tag{1}$$

is determined by the Poisson parameter $\lambda(\lambda > 0)$. Note the lack of subscript for $\lambda$. At this stage the individual Poisson parameters are still assumed to be homogeneous in the population.

Since we are using registration data, we do not have an observed frequency of the individuals who are not apprehended, $f_0$, and this frequency needs to be estimated. For this purpose we assume that the observed frequencies $f_k$ $(k = 1,\ldots,K)$ are generated by a Poisson distribution truncated at zero.

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

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

with $p(y_i > 0|\lambda) = 1 - \exp(-\lambda)$, $i = 1,\ldots,N$. Assume we have an estimate $\hat{\lambda}$ for $\lambda$. This estimate can be used to find the probability of an individual not being observed, $\hat{p}_0 = \exp(-\hat{\lambda})$. The number of unobserved individuals (those individuals who were not apprehended but had a positive probability of being apprehended) is denoted by $\hat{f}_0$ and can be calculated as

$$\hat{f}_0 = \frac{\hat{p}_0}{1 - \hat{p}_0} N_{obs}, \tag{3}$$

where $N_{obs}$ is the number of observed individuals in the sample. An estimate of the population size, $\hat{N}$, is then obtained by

$$\hat{N} = \hat{f}_0 + N_{obs}. \tag{4}$$

VAN DER HEIJDEN *et al.* (in press) work out a Horvitz–Thompson point and interval estimate for $\hat{N}$. The point estimate for $N$ is

$$\hat{N} = \sum_{i=1}^{N} \frac{I_i}{p(\lambda)}, \tag{5}$$

where $I_i = 1$ if individual $i$ is in the sample and $I_i = 0$, otherwise, and $p(\lambda) = 1 - \exp(\lambda)$ is the probability of an individual being present in the sample. This probability can be estimated by replacing the parameter $\lambda$ with its estimated value $\hat{\lambda}$ obtained from fitting the zero-truncated homogeneous Poisson model (2).

The variance of $\hat{N}$ is given by

$$\text{var}(\hat{N}) = E[\text{var}(\hat{N}\,|\,I_i)] + \text{var}(E[\hat{N}\,|\,I_i]). \tag{6}$$

The first term in (6) reflects the uncertainty in the estimate $\lambda$, given the observed individuals. An estimate is obtained by the well-known delta method. The second term reflects the effect of the variability in the number of observed individuals. It is a well-known term in survey sampling. It depends on the $p(\lambda)$ values of all individuals and can be estimated from the observed individuals by a Horvitz–Thompson type estimator. For details on how the variance in (6) is estimated, we refer to VAN DER HEIJDEN *et al.* (in press). The first term in (6) will dominate the variance $\text{var}(\hat{N})$ if $\lambda$ is large, because the probability of being observed is large, and the second term in (6) will dominate the variance $\text{var}(\hat{N})$ if $\lambda$ is small, because then there is a large probability of not being observed.

We now introduce zero-truncated Poisson regression, where the Poisson parameter is a function of one or more covariates. Let $Y_1,\ldots,Y_N$ be a random sample from a zero-truncated Poisson distribution with parameter $\lambda_i$, $i = 1,\ldots,N$. Let $y_1,\ldots,y_N$ be the realizations of $Y_1,\ldots,Y_N$. In our applications $y_i$ is the observed number of times individual $i$ is found in the police registration system for different violations of a specific law, e.g. drunk driving. In the Poisson regression model (see for example CAMERON and TRIVEDI, 1998; GREENE, 1997; LONG, 1997), the Poisson parameter $\lambda_i$ of individual $i$ is a function of a covariate vector $\mathbf{x}_i = (1,x_{i1},\ldots,x_{ip})^{\mathrm{T}}$ as

$$\log(\lambda_i) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i, \tag{7}$$

where $\boldsymbol{\beta} = (\alpha,\beta_1,\ldots,\beta_p)^{\mathrm{T}}$. The Poisson parameter is now given with subscript $i$, since the covariates in the regression model introduce observed heterogeneity in the Poisson parameters. This heterogeneity is called observed since it is determined completely by the covariates, and individuals with identical covariate values will also have an identical Poisson parameter.

This model (7) provides an estimator for the unknown parameter $\lambda_i$ for the sampled individuals and thus for the probability of being present, $p(\lambda_i)$, $i = 1,\ldots,N_{obs}$. VAN DER HEIJDEN *et al.* (in press) derive the Horvitz–Thompson estimator for the total number of individuals in a heterogeneous Poisson population which is defined by

$$\hat{N} = \sum_{i=1}^{N} \frac{I_i}{p(\mathbf{x}_i, \boldsymbol{\beta}^*)}, \tag{8}$$

where $I_i = 1$ if the individual is present and 0 otherwise. The variance of $\hat{N}$ can be split into two parts as in (6) estimated in much the same way as the variance of (7). For details, we refer to VAN DER HEIJDEN *et al.*, (in press).

For every fitted model the deviance, also known as the likelihood ratio, is calculated as −2 times log (likelihood of the current model/likelihood of the saturated model). In a contingency table context, it is possible to evaluate the fit of a model if the cells in the table are reasonably filled. However, if there are continuous covariates, the deviance cannot be used to assess the absolute fit of a model. Then, if we compare two nested models, the difference between the deviances can be used to assess the relative fit of the two models: this difference is chi-square distributed and allows us to assess the significance of the explanatory variables that are in the more complex model but not in the less complex model.

The Poisson distribution is characterized by equality of mean and variance. An important reason for overdispersion (the variance exceeds the mean) is unobserved heterogeneity. This type of heterogeneity is called unobserved since the differences in the individual Poisson parameters cannot be explained by measured covariates. For the truncated Poisson regression model, a Lagrange multiplier test on overdispersion was developed by GURMU (1991). It compares the model fit of the Poisson model against alternative models with an extra dispersion parameter included, such as the negative binomial model. The test statistic is chi-square distributed with one degree of freedom. We have used this test to assess and compare the degree of overdispersion in different models.

Another way to assess model fit is to compare the distribution of the count variable observed in the data with the distribution estimated by the model. For this purpose we use Pearson residuals, i.e. residuals computed as (observed–fitted)/sqrt (fitted).

We now discuss an interesting property of the truncated Poisson regression model when used for the estimation of population sizes, i.e. a model with fewer covariates has a lower estimate of the population size $N$. This can be explained as follows. Assume that a truncated regression model with one dichotomous covariate is true, and that the model without this covariate is not true. This latter model is consequently misspecified. We now show that this misspecified model will have a lower estimate of the population size than the true model. As can be deduced from equation (5), the estimate of the population size is directly related to the estimate of the probability of the zero count $p_0$, which is $p_0 = \exp(-\lambda)$. Let the two Poisson parameters in the true model be $(x+\delta)$ and $(x-\delta)$, and let the Poisson parameter in the misspecified model be $x$. Then Figure 1 illustrates that $\exp(E(-\lambda))$ is smaller than $E(\exp(-\lambda))$, which is also known as the Jensen inequality. Generalization to a situation with more covariates is straightforward. In practice, this means significant covariates should always be included in the model. If not, the model is misspecified and the estimated population size will be too low. Also assume all the observed covariates are included in the final model, but there is still overdispersion, i.e. the observed heterogeneity is taken into account, but there is remaining unobserved heterogeneity. This means the final model is misspecified, and the final estimate of the population size can only be considered as a lower bound.

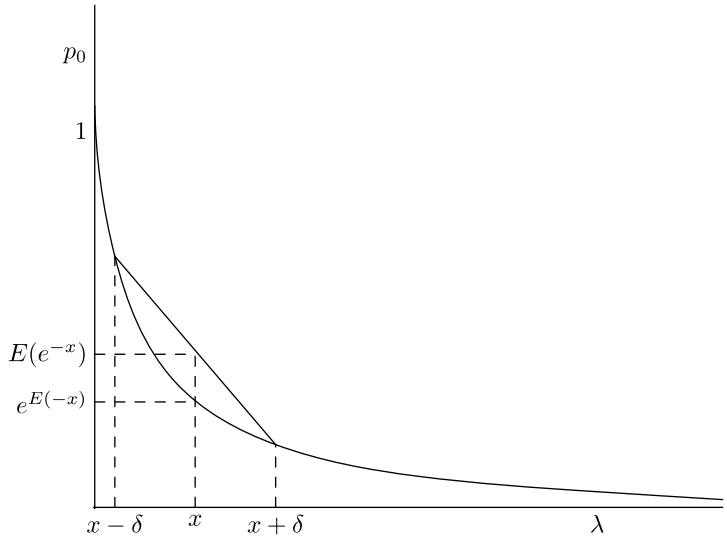


Fig. 1. Illustration showing that a model with a dichotomous covariate has higher estimates of $\hat{N}$ than a model without a covariate.

## 3 Examples

We present two examples. The first example pertains to the illegal possession of firearms, and the second example to drunk driving.

In the introduction we saw that there are two main threats to the validity of the assumptions we are making, i.e. contagion and overdispersion. Overdispersion can be assessed from the data, and this will be discussed below.

This leaves us with the problem of contagion, i.e. the probability of an apprehension or non-apprehension may increase as the result of an apprehension or non-apprehension of the same offender. This probability can also increase as the result of a behavior change on the part of the offender (e.g. if the offender finds out the punishment following an arrest is relatively minor, or might decrease if the punishment is evaluated as unpleasant), or as a result of a change in the behavior on the part of the police. For both the offences studied in this section, we have investigated the change in police behavior by studying arrest reports and interviewing police officers about a sample of offenders apprehended at least twice. In this study it became clear that contagion resulting from a change in police behavior, if it is present at all, cannot be large. For drunk driving, this is because most apprehensions take place at random, since the police apprehend most offenders by stopping and checking all passing drivers. Many offenders are apprehended as the result of their own behavior (such as displaying a gun in a public place), guns are found when they commit other offences so here the apprehensions do not take place due to any random initiative taken by police officers.

Table 1 shows the observed recapture frequencies for illegal ownership of firearms, as $f_1 = 2{,}561$, $f_2 = 72$ and $f_3 = 5$. Note that the sample variance in the dependent variable is small, so the information is limited. In Table 2 we report a model search procedure. The null model has a deviance of 738.1. We then add the gender variable to the model but the deviance does not become significantly better ($\Delta$ deviance $= 0.1$, for 1 d.f.). If we add the two age variables, i.e. current age and age at first offence, the fit improves significantly ($\Delta$ deviance is 10.2, for 2 d.f.). Adding six count variables for criminal history (violence, drugs, economic offences, traffic violations, property crimes and vandalism) also improves the fit ($\Delta$ deviance is 29.3, for 6 d.f.), but adding four dummy variables for the five police regions does not improve the fit further, so the distribution of the counts of the regions do not seem to differ if we control for sex, age variables and criminal histories. An indication that the model fits well is given by the match between observed and expected frequencies: the Pearson residuals are relatively small.

Another indication of fit is given in Table 2 by the test for overdispersion, i.e. a violation of the Poisson assumption (column headed by with $\chi^2_{(1)}$, and $p$-value in column 5). This test indicates whether there is evidence for unobserved heterogeneity, given the inclusion of covariates that take observed heterogeneity into account. This test shows that once we include the six criminal history covariates, there is no evidence in the data for further unobserved heterogeneity, so the Poisson assumption is not violated.

For the most comprehensive model $\hat{f}_0 = 60{,}084.0$, so for this model the total population size is estimated as $\hat{N} = 62{,}722$ (C.I. is 43,973–81,471). The population size estimates for the other models are shown in Table 2. These estimates illustrate the typical result, proven at the end of section 2.2, that the more covariates (i.e. the more observed heterogeneity) there are in the model, the larger the estimate for $\hat{N}$.

Table 2. Model comparisons Illegal possession of firearms in top panel, drunk driving in bottom panel.

| Model | Deviance | $\Delta$ dev. | d.f. | $\chi^2_{(1)}$ | $p$ | $\hat{N}$ | C.I. |
|---|---|---|---|---|---|---|---|
| Null model | 738.1 | – | – | 6.4 | 0.01 | 44,201 | (34,828–53,574) |
| + Gender | 738.1 | 0.1 | 1 | 6.4 | 0.01 | 44,239 | (34,843–53,634) |
| + Age var. | 727.9 | 10.2 | 2 | 5.4 | 0.02 | 48,244 | (36,863–59,625) |
| + Violations | 698.6 | 29.3 | 6 | 2.0 | 0.16 | 60,782 | (43,321–78,244) |
| + Regions | 696.3 | 2.3 | 4 | 1.8 | 0.18 | 62,722 | (43,973–81,471) |
| Null model | 4,652.4 | – | – | 115.4 | 0.00 | 78,710 | (72,738–84,682) |
| + Gender | 4,634.8 | 17.5 | 1 | 110.3 | 0.00 | 82,319 | (74,828–89,809) |
| + Age var. | 4,515.2 | 19.7 | 2 | 73.1 | 0.00 | 99,124 | (87,609–110,639) |
| + Violations | 4,389.3 | 125.8 | 6 | 30.7 | 0.00 | 108,097 | (95,581–120,612) |
| + Regions | 4,358.4 | 30.9 | 4 | 24.6 | 0.00 | 113,771 | (99,857–127,685) |

Deviance is the likelihood ratio. $\Delta$ dev. follows a chi-square distribution with degrees of freedom equal to the number of added covariates and can be used to assess the relative fit of two nested models. $\chi^2_{(1)}$ with its $p$ value refers to the Gurmu test for overdispersion, the test is chi-square distributed with one degree of freedom. $\hat{N}$ gives the estimated population size of the model, with C.I. the 95 percent confidence interval.

Table 3. Raw regression coefficients, standard errors, $t$ tests and exponents of regression coefficients for illegal possession of firearms (left panel) and for drunk driving (right panel).

| | $b$ | s.e. | $t$ | Exp($b$) | $b$ | s.e. | $t$ | Exp($b$) |
|---|---|---|---|---|---|---|---|---|
| Intercept | −3.82 | 0.40 | −9.67 | 0.02 | −1.67 | 0.15 | −11.20 | 0.19 |
| Male | 0.00 | – | – | 1.00 | 0.00 | – | – | 1.00 |
| Female | 0.35 | 0.46 | 0.75 | 1.42 | −.46 | 0.23 | −2.01 | 0.63 |
| Age first offence | −.02 | 0.02 | −1.31 | 0.98 | −.03 | 0.01 | −5.58 | 0.97 |
| Age | 0.03 | 0.02 | 2.01 | 1.03 | 0.01 | 0.01 | 1.55 | 1.01 |
| Violence | 0.34 | 0.18 | 1.86 | 1.41 | 0.10 | 0.11 | 0.93 | 1.10 |
| Drugs | 0.21 | 0.24 | 0.90 | 1.24 | −.13 | 0.23 | −.54 | 0.88 |
| Economic | −.68 | 0.84 | −.80 | 0.51 | −.31 | 0.57 | −.54 | 0.74 |
| Traffic | 0.25 | 0.24 | 1.04 | 1.29 | 0.66 | 0.07 | 9.63 | 1.94 |
| Property | 0.53 | 0.14 | 3.81 | 1.69 | 0.20 | 0.08 | 2.50 | 1.22 |
| Vandalism | 0.02 | 0.23 | 0.10 | 1.02 | 0.24 | 0.12 | 2.07 | 1.27 |
| Rotterdam Rijnmond | 0.00 | – | – | 1.00 | 0.00 | – | – | 1.00 |
| Gelderland Center | −.16 | 0.32 | −.48 | 0.85 | −.61 | 0.16 | −3.77 | 0.54 |
| Holland Center | −.44 | 0.45 | −1.00 | 0.64 | −.56 | 0.15 | −3.87 | 0.57 |
| South Holland South | 0.06 | 0.34 | 0.17 | 1.06 | −.35 | 0.13 | −2.61 | 0.71 |
| Center and West Brabant | −.34 | 0.31 | −1.10 | 0.71 | −.13 | 0.10 | −1.25 | 0.88 |

We also see that if covariates are added that are not significant, $\hat{N}$ does not become substantially larger.

Table 3 reports the raw regression coefficients (b) for the most comprehensive model. We also included the exponents of the raw regression coefficients since these are more easily interpretable. These indicate the (multiplicative) factor the expected count in the dependent variable changes by as a result of a unit change in the covariate, given that all the other covariates are held constant. For a general discussion of the interpretation of these coefficients, we refer to LONG (1997, p. 223–226, 241). Long distinguishes an interpretation in terms of expected counts and an interpretation in terms of the distribution of counts. We first give an interpretation in terms of expected counts, and then give an interpretation in terms of the truncated distribution (discussion of Table 4).

Given the model selection procedure, it is not surprising that the sex variable is not significant. The sign of the regression coefficient for age of first offence is in the

Table 4. Observeda $N$, estimated $N$ and estimated probability of being observed for some subgroups: for illegal possession of firearms (left panel) and for drunk driving (right panel).

| | Obs. $N$ | Est. $N$ | Prob. | Obs. $N$ | Est. $N$ | Prob. |
|---|---|---|---|---|---|---|
| Male | 2,496 | 60,030.3 | 0.042 | 8,738 | 99,070.3 | 0.088 |
| Female | 142 | 2,691.9 | 0.053 | 681 | 14,700.9 | 0.046 |
| Rotterdam Rijnmond | 1,172 | 25,401.7 | 0.046 | 3,960 | 37,481.2 | 0.106 |
| Gelderland Center | 423 | 10,120.7 | 0.042 | 962 | 15,699.9 | 0.061 |
| Holland Center | 237 | 7,093.5 | 0.033 | 1,260 | 21,151.4 | 0.060 |
| South Holland South | 262 | 4,669.1 | 0.056 | 1,278 | 18,081.9 | 0.071 |
| Center and West Brabant | 544 | 15,437.3 | 0.035 | 1,959 | 21,356.8 | 0.092 |

expected direction (the lower the age at the first offence, the larger the count) but not significant. Age is significant (for each year, the expected count increases by a factor of 1.03). The log transformations of the counts for the violation histories provide some evidence of a positive relation between violence and property crimes and the expected number of times someone is caught for illegal possession of firearms.

For the purposes of illustration we also show the estimated probabilities of being apprehended at least once in the left panel of Table 4. We only do so for the categorical covariates. All the estimates are around 0.042, the overall probability of being apprehended at least once. Estimates for the different levels do not differ much, and this is not surprising given that Table 2 shows that neither sex nor region contribute significantly to the model.

The second example pertains to drunk driving. Here the variance in the counts is larger than for the firearms example (see Table 1). Also the sample size is much larger. The observed and estimated frequencies are not as close as in the previous example, as is also indicated by the higher values of the residuals.

The model search reported in the bottom panel of Table 2 shows that now all the sets of variables increase the fit of the model, in particular the criminal histories. Also, note that the estimated population sizes increase substantially with each model. What makes this example different from the firearms example is that, in the final model, the $\chi^2_{(1)}$-test shows that there is still evidence of overdispersion (unobserved heterogeneity). This result is in line with the larger residuals in Table 1. Additional covariates are needed to take the unobserved heterogeneity into account, but unfortunately they are not available. However, since every additional covariate leads to an increase in $\hat{N}$, this means we should interpret the estimate of 113,771 of the final model as a lower bound estimate of the true population size, so the true population size is estimated to be larger than 113,771. See the discussion at the end of section 2.2.

We now turn to the interpretation of the regression coefficients. For drunk driving the sex variable is significant (for a female the expected count is 0.68 times that of a man), as is the age at the first offence and the three criminal history covariates 'other traffic violations', 'property crimes' and 'vandalism'. So the more often someone is caught for other traffic violations, the higher the expected count for drunk driving. The police regions also differ significantly. The number of estimated apprehensions in Rotterdam Rijnmond is about 1.85 times higher than in Gelderland Center (1.85 is the inverse of the exponent of the regression coefficient of Gelderland Center).

Table 4 shows interesting differences between the estimated probabilities of being apprehended at least once: males seem to be caught almost twice as often as females. Also, in Rotterdam Rijnmond the police seems to be more effective than in Gelderland Center and Holland Center (the probabilities of being apprehended at least once being 11 percent as compared with 6 and 6 percent, respectively).

## 4 Discussion

We have shown how police records can be used to estimate the size of criminal populations. These estimates can be used to evaluate the effectiveness of the police forces, and grant insight into differential arrest rates (COLLINS and WILSON, 1990) for different groups.

Even though the definition of the data is straightforward, the data are contaminated with errors. The reason is that the police do not collect these data for the purpose of conducting statistical analyses; they do so facilitate the process of apprehending individuals. Therefore the registration is not always as careful as it should be. For example, extensive data cleaning was required to minimize the likelihood of incorrect double counts (i.e. the same apprehension appearing twice in the system). Clearly, an incorrect double count decreases $f_k$ by 1 but increases $f_{k+1}$ by 1, so that there appears to be more recaptures than there actually are. The result is that the estimated zero count, $\hat{f}_0$, is too low. Although careful attention has been paid to eliminating incorrect double counts, it is possible that there are still a few in the data.

As regards the meaning of our estimate $\hat{N}$, one might wonder what it stand for? The firearms example may make clear why its meaning is not as straightforward as one may think. Imagine someone with a gun safely buried somewhere in his home. This individual has a zero probability of being apprehended, and we cannot generalize from the individuals who actually are apprehended to this type of individual. Basically, we can only generalize from the apprehended individuals to similar individuals who are not apprehended *but who are in principle apprehensible*. Thus the estimate $\hat{N}$ does not stand for the total number of illegal gun owners, it only stands for the *apprehensible* ones. On the other hand, do non-apprehensible individuals really exist and, if so, do these individuals belong to the population of interest? To have a zero capture probability, an individual has to hide his gun that well that it no longer poses any threat to society. We stress that the population estimate is still useful, since it is represents the number of individuals who pose a threat to society and is thus a good indication of the police workload (we may not reasonably expect the police to apprehend non-apprehensible individuals).

We have extensively discussed the assumptions of the model in section 1.1. The main threat to the validity of the model outcomes are contagion and overdispersion. Contagion refers to the capture probabilities changing as the result of apprehensions or non-apprehensions. We have given a few examples of why this might occur in the context of police registration data. In section 3 we discuss the possible behavior change of police officers, and conclude that our additional research (not reported here in detail) leads to the conclusion that it is not likely to be a serious problem. However, we do not have any insight into the contagion resulting from a possible change in the behavior of offenders. In general, positive contagion makes the observed counts too large, which leads to an underestimation of the population size. Negative contagion will make the observed counts too small, which leads to an

overestimation of the population size. In section 3 we discuss the openness of the population in the context of negative contagion, and conclude that leaving the population as a result of an apprehension is a violation of the Poisson assumption.

Overdispersion may result from unobserved heterogeneity. Unobserved heterogeneity becomes evident as a result of the analysis, and there is evidence of overdispersion in the drunk driving example, but not in the firearms example. The result is that the estimated population size can only be interpreted as an estimate of the lower bound for the true population size.

One last area where violation of the model assumptions may occur is in the dependence of the observations (COLLINS and WILSON, 1990), which would occur if certain apprehensions involve more than one individual at the same time. This type of violation can, in principle, be checked in the police registration system.

Concluding, we presented a method that provides an estimate of the population size, but the assumptions underlying the method need careful consideration as violations may seriously distort the estimate. We think it is advisable to compare the estimates found with estimates from other sources, if they are available.

**Acknowledgements**

This research has been partly funded by the Ministry of Justice, The Netherlands. We gratefully acknowledge the comments of guest editor Catrien Bijleveld and an anonymous referee, and the help of Leen Prins of the National Police Force Service (KLPD) in preparing the data files.

**References**

BUNGE, J. and M. FITZPATRICK (1993), Estimating the number of species: a review, *Journal of the American Statistical Association* **88**, 364–373.

CAMERON, A.C., and P.K. TRIVEDI (1998), *Regression analysis of count data, Econometric Society Monographs No. 30*, Cambridge, Cambridge University Press.

CHAO, A. (1988), Estimating animal abundance with capture frequency data, *Journal of Wildlife Management* **52**, 295–300.

CHARLIER, C.V.L. (1905), Die zweite Form des Fehlergesetzes, *Arkiv fur Matematik, Astronomi och Fysik* **2**, 1–35.

COLLINS, M.F., and R.M. WILSON (1990), Automobile theft: estimating the size of the criminal population, *Journal of Quantitative Criminology* **6**, 395–409.

EFRON, B. and R. THISTED (1976), Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika* **63**, 435–447.

ENGBERSEN, G., R. STARING, J. VAN DER LEUN, J. DE BOOM, P. VAN DER HEIJDEN and M. CRUYFF (2002), *Illegale vreemdelingen in Nederland. Omvang, overkomst, verblijf en uitzetting*, RISBO, Erasmus University, Rotterdam, The Netherlands.

GREENE, W.H. (1997), *Econometric analysis (3rd ed.)*, Prentice Hall, New Jersey.

GREENE, M.A., and S. STOLLMACK (1981), Estimating the number of criminals, in: J.A. FOX (ed.), *Models in quantitative criminology*, Academic Press, New York, 1–24.

GURMU, S. (1991), Test for detecting overdispersion in the positive Poisson regression model, *Journal of Business and Economic Statistics* **9** 215–222.

HOOGTEIJLING, E. (2002), *Raming van het aantal niet in de GBA geregistreerden*., Voorburg, Statistics Netherlands.

HULS, F.W.M., M.M. SCHREUDERS, M.H. TER HORST-VAN BREUKELEN and F.P. VAN TULDER (2000), *Criminaliteit en Rechtshandhaving 2000* **189**, 409–410.

JOHNSON, N.L., S. KOTZ and A.W. KEMP (1993), *Univariate discrete distributions*, (2nd ed.), Wiley, New York.

LONG, J.S. (1997), *Regression models for categorical and limited dependent variables*, CA Sage, Thousand Oaks.

ROSSMO, D.K. and R. ROUTLEDGE (1990), Estimating the size of criminal populations, *Journal of quantitative criminology* **6**, 293–314.

SANATHANAN, L. (1977), Estimating the size of a truncated sample, *Journal of the American Statistical Association* **72**, 669–672.

SEBER, G.A.F. (1982), *The estimation of animal abundance*, 2nd ed., Griffin and Co., London.

SEBER, G.A.F. (1986), A review of estimating animal abundance, *Biometrics* **42**, 267–292.

SMIT, F., P.G.M. van DER HEIJDEN and G. VAN GILS (1993), Enkele weinig gebruikte methoden om de omvang van criminaliteit te schatten, *Tijdschrift voor Criminologie* **36**, 97–119.

SMIT, F., J. TOET and P.G.M. VAN DER HEIJDEN (1996), City Report Rotterdam: Estimating the number of opiate users in Rotterdam using statistical models for incomplete count data, in: *European Monitoring Centre for Drugs and Drug Addiction (EMCDDA), Methodological pilot study of local level prevalence estimates*, EMCDDA, Lisbon, Portugal, pp. 49–69.

VAN DER HEIJDEN, P.G.M., F. SMIT and G. VAN GILS (1993), *Schattingen van het aantal slachtofferloze delicten*., Politia Nova **3**, Ministry of Internal Affairs, The Netherlands.

VAN DER HEIJDEN, P.G.M., R. BUSTAMI, M.J.L.F. CRUYFF, G. ENGBERSEN and H. VAN HOUWELINGEN (in press). Point and interval estimation of the population size using the truncated Poisson regression model, *Statistical Modelling*.

VAN DER LEUN, J., G. ENGBERSEN and P.G.M. VAN DER HEIJDEN (1998), *Illegaliteit en criminaliteit: schattingen, aanhoudingen en uitzettingen*, Erasmus University, Department of Sociology, Rotterdam.

WILSON, R.M. and M.F. COLLINS (1992), Capture–recapture estimation with samples of size one using frequency data, *Biometrika* **79**, 543–553.

WITTEBROOD K. and M. JUNGER (2002), Trends in violent crime: a comparison between police statistics and victimization surveys, *Social Indicators Research* **59**, 153–173.

ZELTERMAN, D. (1988), Robust estimation in truncated discrete distributions with application to capture-recapture experiments, *Journal of Statistical Planning and Inference* **18**, 225–237.

ZELTERMAN, D. (2001), *Selected applications for categorical data. Advanced log-linear models using GENMOD, in Series: SAS books by users*.

Received: August 2002. Revised: December 2002.