Estimating the scale of modern slavery in the United Kingdom in 2013

Thomas Uriot

June 6, 2021

Abstract

In this article, the population size of modern slaves in the UK, in 2013, is estimated. A capture-recapture approach is used and data is collected from 4 different sources. The overlapping observations are then used to estimate the unobserved part of the population. A log-linear model is fitted to the data using R and the expected frequencies are estimated. It is shown that the 95% confidence interval for the population size of modern slaves is [9450:12678]. This way of collecting data fails to give an estimate for more restricted geographical areas as well as particular types of modern slavery.

Keywords: modern slavery; United Kingdom; Epidemiology; log-linear; population size

1 Introduction

The traditional definition of a slave from the Oxford Dictionary is: a slave is a person who is the legal property of another and is forced to obey them. At the time of traditional slavery, it would have been simple to estimate the total number of slaves, as they were seen as commodities and thus recorded in official agreements between the seller and the buyer. Nowadays, traditional slavery -legally owning a person- is forbidden in all countries in the world. However, another form of slavery called 'modern slavery' still exists today. The illegal nature of modern slavery makes it very hard to know how many people are modern slaves, in other words, they form a hidden population. With 1,746 reported cases in 2013, it represents a 47% increases from 2012. Of course, the fact that there are more reported cases in 2013, does not mean that there are more people affected in 2013 than in 2012 but it remains a worrying figure[1]. According to the U.S Department of State, modern slavery includes a variety of activities: forced labor, sex trafficking, bonded labor, involuntary domestic servitude and child soldiers[2]. In order to estimate the number of modern slaves, one needs to collect information from different sources or lists and uses the overlapping observations within the sources. It is preferable to be restricted to a certain geographical area in order to collect data since it would be very complicated to find sources with any overlap if the region of study was too large.

Taking several sources and using the overlapping individuals in order to estimate the size of the population of interest is a common method in epidemiological studies. Jacqueline E.Tate and Michael G. Hudgens (2006) have used a two and three-stage sampling design to obtain size estimates of populations at risk of HIV. Another paper by Matthew Hickman et al. (2014) is closely related to this article. The authors collected data from three different sources and used log-linear models in order to estimate the total number of drug users in England. Note that the authors warn against the assumption of the

[1] https://modernslavery.co.uk/index.html

[2] http://www.state.gov/j/tip/what/

highest-order interaction being 0, which can lead to "seriously biased estimates".

This article is based on the paper by Bernard Silverman (2014) and the analysis builds on the 2013 National Crime Agency (NCA) Strategic Assessment, which collected data from 6 sources that identify potential victims of human trafficking (modern slavery). In the study, the authors combined 2 sources (police force and national crime agency) because they contained very similar information, leaving 5 different sources. Here, only 4 of those sources are used as a matter of simplification. The sources used are: local authority, non-government organisation, police force and government organisation. Before the study, the only known figure for the number of modern slaves in the UK was the number of potential victims. The aim of this study is to build a generalised linear model with Poisson distribution (log link function) and then use the bootstrap method to come up with a 95% confidence interval for the population size. Being able to provide a good estimate is the first step for the government to take appropriate measures to fight human trafficking and modern slavery.

2 Data structure and modeling

2.1 Data collection

As mentioned in the introduction, the data was collected from 4 different sources, in the United Kingdom, in 2013. The sources are

S_1 - Local authority
S_2 - Non-government organisation
S_3 - Police force
S_4 - Government organisation

and the numbers of potential victims of modern slavery are recorded in Table 1 below, according to each list, including the overlaps.

	S_4	1		0	
	S_3	1	0	1	0
$S_1 = 1$	$S_2 = 1$	1	1	1	15
	$S_2 = 0$	0	3	19	54
$S_1 = 0$	$S_2 = 1$	4	19	62	464
	$S_2 = 0$	76	703	1006	x

Table 1: Contingency table for the number of potential victims from all sources.

As a bit of notation, let n_{1111} denote the number of individuals seen in all sources and n_{0000} be the number of unseen individuals. Thus, looking at Table 1, $n_{1111} = 1$ and $n_{0000} = x$. The subscripts of n correspond to S_1, S_2, S_3 and S_4 respectively, such that $n_{1001} = 3$ is the number of individuals seen in both S_1 and S_4 but unseen in S_2 and S_3. Therefore the total number of individuals n is

$$n = n_{1000} + n_{0100} + n_{0010} + n_{0001} + n_{0000} = 54 + 464 + 1006 + 703 + x. \tag{1}$$

The aim is then to estimate x. There are two equivalent ways of doing this: maximizing the likelihood function and finding the maximum likelihood estimates or using a generalised linear model. In this article, the later is used.

2.2 Modeling

Let μ_{ijkl} denote the expected cell frequencies, for i, j, k, l either 1 or 0. For example,

$$E[n_{1000}] = \mu_{1000}.$$

For multinomial sampling, under independence, the μ_{ijkl} have the structure

$$\mu_{ijkl} = n\pi_{ijk+}\pi_{ij+l}\pi_{i+kl}\pi_{+jkl},$$

where π_{ijkl} is the probability of observing an individual and the + subscript denotes the sum over either all i, j, k or l. For instance,

$$\pi_{+jkl} = \pi_{0jkl} + \pi_{1jkl}.$$

Thus $log(\mu_{ijkl})$ has additive form. The independence assumption is equivalent to the highest order interaction term being 0. Furthermore, in this study, three-way interactions between sources are assumed to be non-significant and only two-way interactions will be kept into the model. Therefore, the log-linear model is

$$log(\mu_{ijkl}) = \alpha + \beta_i + \beta_j + \beta_k + \beta_l + \beta_{(ik)} + \beta_{(ij)} + \beta_{(il)} + \beta_{(jl)} + \beta_{(jk)} + \beta_{(kl)}, \qquad (2)$$

with the corner-point constraints that each parameter with a 0 subscript is equal to 0. The estimated expected frequencies are denoted by $\hat{\mu}_{ijkl}$ and then

$$log(\hat{\mu}_{ijkl}) = \hat{\alpha} + \hat{\beta}_i + \hat{\beta}_j + \hat{\beta}_k + \hat{\beta}_l + \hat{\beta}_{(ik)} + \hat{\beta}_{(ij)} + \hat{\beta}_{(il)} + \hat{\beta}_{(jl)} + \hat{\beta}_{(jk)} + \hat{\beta}_{(kl)}. \qquad (3)$$

Here, the aim is to estimate μ_{0000} where $\hat{\mu}_{0000} = \exp\hat{\alpha}$ because of the corner-point constraints. Then, it is possible to estimate the total population n by substituting $x = \exp(\hat{\alpha})$ in equation (1). In order to estimate the parameters in (3) and come up with a confidence interval, the software R was used and the code is provided in Appendix A. The model deemed to give the best fit to the data is the one with the lowest AIC (Aikaike's criterion). The formula to calculate AIC is

$$AIC = 2\ln(L) + 2p. \qquad (4)$$

where L is the likelihood function evaluated at the maximum likelihood estimators and the total number of parameters estimated in the model is denoted by p.
Note that even if the AIC of the submodel is equal or slightly higher, one would still keep the submodel since it would be more parsimonious. A parsimonious model is a model that offers a good fit to the data with the least number of parameters. This is because it is more convenient to use the model and it is more suitable to real life applications. For instance. if the model of choice includes many factors, one will need to find and individual for whom all these factors are known or measurable which can be difficult in some cases. The first block of code computes the parameters of the log-linear model using the $glm()$ function. The second block implements the bootstrap method to compute a confidence interval for the population size.

3 Analysis and Results

3.1 Model selection

Firstly, I started from the most complex model in (2) and then from the simplest additive model with no two-way interactions. Both methods led to the same result. In the first case, one has to fit the model with all the two-way interactions and then fit model (2) six times with a different two-way interaction removed each time. Once that the first two-way interaction is removed, one has to repeat the process and eliminated another interaction until no more improvement (either a smaller AIC or a more parsimonious model with a similar AIC) is possible. In Table 2 below, the AIC of the selected models at each step are shown. This is for the first method, working from the most complex model to the best fit. Let us label the following models

$$log(\mu_{ijkl}) = \alpha + \beta_i + \beta_j + \beta_k + \beta_l + \beta_{(ik)} + \beta_{(ij)} + \beta_{(il)} + \beta_{(jl)} + \beta_{(jk)}, \quad (5)$$

$$log(\mu_{ijkl}) = \alpha + \beta_i + \beta_j + \beta_k + \beta_l + \beta_{(ik)} + \beta_{(ij)} + \beta_{(jl)} + \beta_{(jk)}, \quad (6)$$

$$log(\mu_{ijkl}) = \alpha + \beta_i + \beta_j + \beta_k + \beta_l + \beta_{(ij)} + \beta_{(jl)} + \beta_{(jk)}. \quad (7)$$

Models	AIC
(2)	101.4
(5)	99.41
(6)	97.67
(7)	96.83

Table 2: AIC for the selected models after each elimination step.

I found that model (7) gave the best fit to the data with the lowest AIC and the lowest number of interactions. When one removes any of the interaction terms left in (7), the fit became significantly worse with a much larger AIC. Thus, I decided that the gain in parsimony from leaving another two-way interaction out was not justified.

3.2 Results

Using the output from R, the estimates of the parameters of the model (7) are given in Table 3 below.

	Parameters							
	α	$\beta_{i=1}$	$\beta_{j=1}$	$\beta_{k=1}$	$\beta_{l=1}$	$\beta_{(i=1,j=1)}$	$\beta_{(j=1,l=1)}$	$\beta_{(j=1,k=1)}$
Estimates	9.06	-5.06	-2.91	-2.14	-2.51	1.50	0.89	-0.56

Table 3: Estimates of the parameters for model (7).

Thus, the estimated number of unobserved (hidden) modern slaves is

$$\hat{\mu}_{0000} = \exp(9.06) \approx 8604.$$

Plugging this number back in (1), the estimated population size n is 10831. However, one is more interested in a confidence interval rather than a point estimate. Using the bootstrap code provided in Appendix A, I found that the 95% confidence interval for the unobserved individuals was [7223:10451]. Each time, the bootstrap method gives a different interval, but the variation is negligible due to the high number of draws. Finally, adding the observed individuals gives a 95% confidence interval for the total number of modern slaves of [9450:12678].

4 Discussion

The population size of modern slaves in the UK in 2013 is estimated to be between 9450 and 12678 people, at a 95% confidence level. Using the point estimate of 10831 that I found earlier, the ratio of modern slaves against the total population of the UK is approximately 1:6000, in 2013. However, there are several limits in practice with this method. It is not possible to know how many people are affected by which "type" of modern slavery that are mentioned in the introduction. Furthermore, all that the study provides is the total number of modern slaves across the UK and we cannot know which country, region or even city is more affected, within the UK. It is of interest to know where the activities are the most concentrated in order to fight them more efficiently. I suggest that it may be useful to focus on a particular type of modern slavery across the 4 sources of data and do the same analysis for each type. As far as the geographical location is concerned, one could use regional sources of data (Scotland, London or cities with more than a certain population size for instance) and compare it with the global figure I have found to know the proportion. Combining both the type of modern slavery and more restricted areas within the UK may not be very informative as the number of observations obtained from such sources would be too small to give a good estimate.

Appendix A

4.1 GLM model

```
# Use glm function to fit Poisson model to data

# Read in the data:

n <- c(1,1,1,15,0,3,19,54,4,19,62,464,76,703,1006)

# Read in the relationship between response variables and log-linear terms
# n = X theta (X is design matrix)
# Create an indicator variable which corresponds
# to the data points in the contingency table

X1 <- c(1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0)
X2 <- c(1,1,1,1,0,0,0,0,1,1,1,1,0,0,0,0)
X3 <- c(1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0)
X4 <- c(1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0)

#Model selection starting with all two-way interactions

mlepo <- glm(n~(X1+X2+X3+X4+X1:X2+X1:X3+X1:X4+X2:X3+X2:X4+X3:X4),
family=poisson(link="log"))
mlepo

mlepo <- glm(n~(X1+X2+X3+X4+X1:X2+X1:X3+X1:X4+X2:X3+X2:X4),
family=poisson(link="log"))
mlepo

mlepo <- glm(n~(X1+X2+X3+X4+X1:X2+X1:X3+X2:X3+X2:X4),
family=poisson(link="log"))
mlepo

mlepo <- glm(n~(X1+X2+X3+X4+X1:X2+X1:X3+X2:X4),
family=poisson(link="log"))
mlepo

#Model selection starting from the additive model

mlepo <- glm(n~(X1+X2+X3+X4+X1),
family=poisson(link="log"))
mlepo

mlepo <- glm(n~(X1+X2+X3+X4+X1:X2),
family=poisson(link="log"))
```

```
mlepo

mlepo <- glm(n~(X1+X2+X3+X4+X1:X2+X1:X3),
family=poisson(link="log"))
mlepo

mlepo <- glm(n~(X1+X2+X3+X4+X1:X2+X4:X2+X1:X3),
family=poisson(link="log"))
mlepo
```

4.2 Bootstrap

```
#Only change from the original code is the inclusion of the new log-linear model

# To calculate a bootstrap estimate of unobserved cell
#

bootpo <- function(bootsamples,n=data) {

  # Set up vector to store the bootstrapped MLE values:

  bootpo <- array(0,bootsamples+1)

  # Calculate empirical estimate of cell probabilities for observed individuals:

  p <- n/sum(n)

  # Simulate bootstrap replicates, calculate MLEs of theta
  # terms and simulate new unobserved cell

  for (i in 1:bootsamples) {
    nboot <- rmultinom(1,sum(n),p)
    temp <- glm(nboot~(X1+X2+X3+X4+X1:X2+X1:X3+X2:X4),
    family=poisson(link="log"))
    bootpo[i] <- exp(temp$coefficients[1])
  }

  # Add in the MLE for N from the observed data
  # (these are included in bootstrap replicates)

  bootpo[bootsamples+1] <- exp(mlepo$coefficients[1])

  # Output the 95% CI as the lower and upper 2.5% quantiles of
  # MLE estimates from bootstrapped sample

  quantile(bootpo,c(0.025,0.975))

}
```

References

Tate, E. and Michael, B. H. (2006) Estimating Population Size with Two and Three-Stage Sampling Designs. *American Journal of Epidemiology*.

Matthew, H. (1999) Estimating the prevalence of problem drug use in inner London: a discussion of three capture-recapture studies. Addiction. **94**, 1653-1662.

Colin Aitken. (2015) Categorical data lecture notes. Chapter 5, 1-5.