Incorporating cyclical effects and time-varying covariates in models for single-source capture-recapture data

Thomas Husken[1], Maarten Cruyff[1], Peter van der Heijden[12]

[1] Utrecht University, the Netherlands
[2] University of Southampton, England

E-mail for correspondence: `t.f.husken@uu.nl`

Abstract: The objective of capture-recapture analysis is to estimate the size of an elusive population, for which the zero-truncated Poisson model is a basic model. We extend this model to the more general recurrent events model to include cyclical effects and time-varying covariates. An application to police data on victims of domestic violence provides strong evidence for the presence of weekly and seasonal cyclical effects on the rate of police reports.

Keywords: single-source capture-recapture; cyclical effects; time-varying covariates; recurrent events model; zero-truncation.

1 Introduction

The zero-truncated Poisson model is a well-established model for the analysis of single-source capture-recapture data. Such data typically arise when each observation of a member of an elusive population is recorded in a registration file. Counting the number of records for each individual population member yields a zero-truncated count distribution, because population members with a zero count are not in the register. Under the assumption that the counts follow a Poisson distribution, an estimate of the Poisson parameter can be obtained that in turn can be used to estimate the frequency of the zero count. Relevant covariates can be included to model individual differences in Poisson parameters, this leads to the zero-truncated Poisson regression model (TPR) (see Cruyff & van der Heijden, 2013; van der Heijden et al., 2003).

Like any type of events data, single-source capture-recapture data can exhibit seasonal or cyclical patterns. For example, a homeless person may be more likely

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

to stay in a homeless shelter during winter than in the summer and a problematic drug user may have a higher probability of being admitted to the hospital in the weekend than during the week. However, the TPR is unable to incorporate these types of cyclical effects.

In this paper we present a method that allows for the inclusion of cyclical effects in single-source capture-recapture data. In this method, the TPR is extended to the zero-truncated recurrent events model (TREM), which includes a time dimension. This dimension makes the model more general than the TPR, since it allows for the inclusion of time-varying covariates and cyclical effects. The resulting model can accommodate a wide variety of effects: time-invariant, cyclical, time-varying, and interactions thereof.

2 Method

The TREM is an extension of the TPR that allows for the modelling of time-varying covariates and cyclical effects in single-source capture-recapture data. The likelihood of the TREM is given by

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \prod_{j=1}^{y_i} \lambda_{ij} \left(\frac{e^{-\Lambda_i(\tau)}}{1 - e^{-\Lambda_i(\tau)}} \right) \right\}, \tag{1}$$

where $\Lambda_i = \sum_{t=1}^{\tau} \lambda_{it}$, and $\ln \lambda_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_p x_{itp}$ (Cook & Lawless, 2007, p. 273-278). Here, x_{itp} specifies the value covariate p takes at time point t for person i. Additionally, y_i is the total number of captures over the observation period for individual i. Note that this allows for time-varying covariates since x_{itp} may vary over time.

Cyclical effects are modelled by adding a cosine term to the linear predictor of the TREM:

$$\ln \lambda_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_p x_{itp} + \alpha \cos\left(\frac{2\pi}{k} t - \theta \right), \tag{2}$$

where α is the amplitude and θ the horizontal shift. The period k is a constant and determines how often the cyclical component peaks. The cosine term in Equation (2) is non-linear and therefore difficult to estimate, but can be rewritten to a linear function. The most common method is using a trigonometric identity to parametrise the cyclical effect as

$$\alpha \cos\left(\frac{2\pi}{k} t - \theta \right) = \beta_{\cos} \cos\left(\frac{2\pi}{k} t \right) + \beta_{\sin} \sin\left(\frac{2\pi}{k} t \right), \tag{3}$$

from Cryer & Chan (1994, Ch. 3, p. 34). The final expression can be easily included in the linear predictor of the recurrent events model, where $\cos(\frac{2\pi}{k}t)$ and $\sin(\frac{2\pi}{k}t)$ are entered as covariates. The interpretation of the two cyclical regression coefficients is not very intuitive, but they can be transformed back in terms of α and θ:

$$\begin{aligned} \alpha &= \sqrt{\beta_{\cos}^2 + \beta_{\sin}^2}, \\ \theta &= \operatorname{arctan2}(\beta_{\sin}, \beta_{\cos}), \end{aligned} \tag{4}$$

which is also known as a polar transformation. The linear parametrisation of Equation (3) has the advantage that interaction effects with time-invariant covariates can be included almost in the classical manner. This provides the option to include main effects of time-invariant covariates, cyclical effects, and interactions between these two simultaneously.

Parameter estimates of the TREM are obtained by optimizing the loglikelihood given by

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{y_i} \ln \lambda_{ij} - \sum_{i=1}^{n} \Lambda_i - \sum_{i=1}^{n} \ln[1 - e^{-\Lambda_i}]. \tag{5}$$

Analytical closed form expressions for the score function and Hessian of the TREM can be derived from the loglikelihood (see Hu & Lawless, 1996). These expressions are then used to set up a Newton-Raphson algorithm. Standard errors are obtained through the observed information matrix.

Given the parameter estimates $\hat{\boldsymbol{\beta}}$ of the TREM, the Horvitz-Thompson population size estimate is obtained as

$$\hat{N} = \sum_{i=1}^{n} \frac{I_i}{1 - e^{-\hat{\Lambda}_i}}, \tag{6}$$

where $I_i = 1$ if case i is observed in the sample and $I_i = 0$ otherwise. The denominator $1 - e^{-\hat{\Lambda}_i}$ is the probability that a population member is observed in the sample. The variance of $\hat{N}$ is calculated through the Delta method, as presented in van der Heijden et al. (2003).

3 Application to domestic violence data

The application is a data set of domestic violence victims from the Netherlands in the period 2004 - 2006. The response of interest is the number of times a police report was filed for domestic violence for a certain individual. Although information on perpetrators of domestic violence is also available, we do not focus on that group in this paper. Hence, our population of interest is defined as victims of domestic violence. There are a total of 56,575 observed victims of domestic violence in the period 2004 - 2006. These data are made available by the Dutch national police.

The variables gender and age are available as subject-specific covariates. Gender is included as a time-invariant covariate. Age is modelled with time-varying linear and quadratic contrasts, meaning that an individual can move from one age group to another during the observation period. The age categories are: 0-17, 18-29, 30-39, 40-49, 50+.

Cyclical effects with periods 366 and 7 are included, representing seasonal and weekly effects, respectively. Interaction effects between the cyclical week effect and the linear and quadratic age effects allow each age group to have a different cyclical week effect. Furthermore, a linear effect of time is added in the final model to allow for an increase or decrease of capture probabilities over the time period of three years. Finally, an interaction effect of the cyclical season effect and the linear time effect is included so that the cyclical seasonal is allowed to vary over time.

TABLE 1. Regression coefficients and point and interval estimate of the population size for the TREM model fit to the domestic violence data

Variable	Coding	$\hat{\boldsymbol{\beta}}$	**SE**
Intercept		-8.63	0.03 ***
Gender	(male = 0, female = 1)	0.63	0.03 ***
Age	Linear	0.26	0.03 ***
	Quadratic	-0.50	0.02 ***
$\cos_{366}$		-0.06	0.01 ***
$\sin_{366}$		-0.02	0.01 **
$\cos_7$		0.05	0.01 ***
$\sin_7$		-0.04	0.01 ***
Age (Linear)*$\cos_7$		-0.03	0.01 *
Age (Linear)*$\sin_7$		0.03	0.01
Age (Quadratic)*$\cos_7$		0.12	0.01 ***
Age (Quadratic)*$\sin_7$		0.01	0.01
Time		0.34	0.01 ***
Time*$\cos_{366}$		0.01	0.01
Time*$\sin_{366}$		0.16	0.01 ***
$\hat{N}$		211,155	
95%-CI		206,460 - 215,848	

Note: $*** = p < 0.001$, $** = p < 0.01$, $* = p < 0.05$.

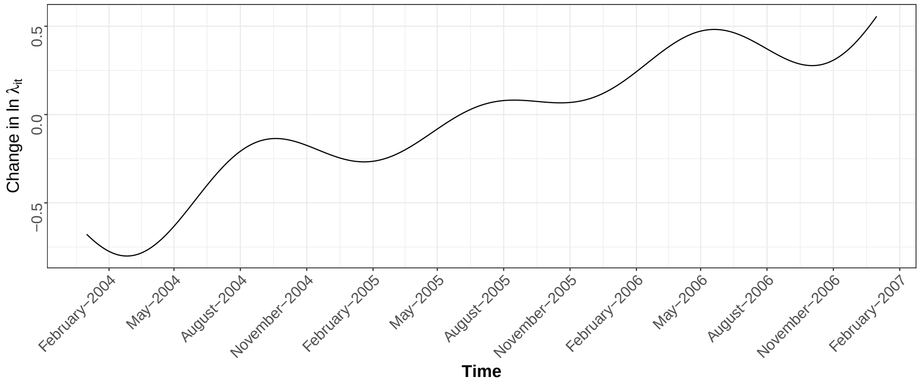

FIGURE 1. Fitted general cyclical trend for the domestic violence application

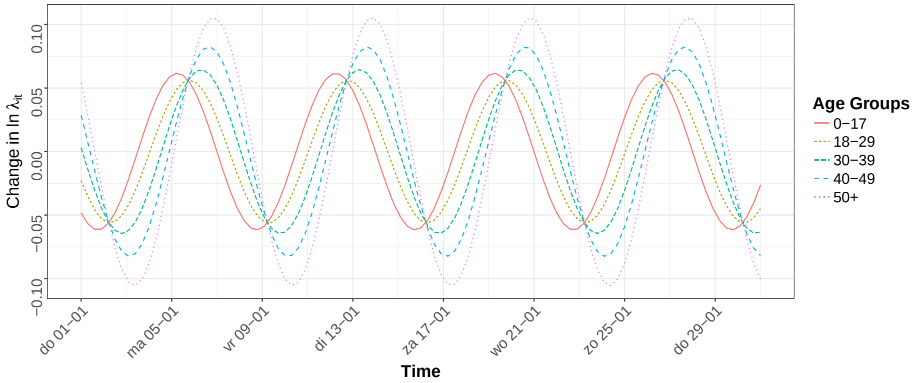

FIGURE 2. Fitted cyclical week effect and its interaction with age for the domestic violence application for the month January 2004

Table 1 shows the regression coefficients and point and interval estimate of the population size for the model fitted to the data. The effects of gender and age are significant. Women are more likely to be mentioned as a victim in a police report for domestic violence than men, and there is a quadratic effect of age.

Both the cyclical main effects are significant, indicating the presence of seasonal and weekly variation in capture probabilities. Additionally, the interaction effects of age (linear and quadratic) with the cosine terms of the cyclical week effect are significant, so that the cyclical week effect is different for each age group.

A positive effect of time is found, indicating that the capture probabilities increase over the course of the observation period. One of the two interaction components of time with the cyclical season effect is also significant, so that the cyclical season effect is different over the three years. The population size estimate of domestic violence victims in the time period 2004 - 2006 is 211,155 (95%-CI: 206,460 - 215,848).

The general cyclical trend of the fitted model (omitting cyclical week effects) is presented in Figure 1. This trend consists of three components: the cyclical season main effect, the linear effect of time, and the interaction between the two. In general, we can say that the cyclical season effect is stronger in 2004 and 2006 than in 2005, and that the capture probabilities increase over the course of the observation period. Additionally, the cyclical season effect in 2006 peaks in May, while in the 2004 and 2005 the effect peaks in September.

In Figure 2, the cyclical week effect and the interaction of this effect with age is presented. These effects are plotted for the month January in 2004, and repeat throughout the length of the observation window. The age group 30-39 is the reference group in this analysis, represented by the green curve. The strength of the cyclical week effect is lowest for this reference group. The cyclical week effect is strongest for the 50+ age group, while the strength of the cyclical effects of the other groups is somewhere in between. For all age groups, the cyclical week effect peaks after the weekend. The groups 40-49 and 50+ peak on Wednesday, and the other groups on Monday (18-29) and Tuesday (0-17 and 30-39).

References

Cook, R.J. & Lawless, J.F. (2007). *The statistical analysis of recurrent events.* Springer Science & Business Media.

Cruyff, M.J. & van der Heijden, P.G.M. (2013). *Sensitivity Analysis and Calibration of Population Size Estimates Obtained with the Zero-Truncated Poisson Regression Model.* Statistical Modelling, 14(5): 361-373

Cryer, J.D. & Chan, K. (2008). *Time series analysis.* Princeton university press.

Hu, X. J. & Lawless, J. F. (1996). *Estimation of rate and mean functions from truncated recurrent event data.* Journal of the American Statistical Association, 91(433):300-310.

van der Heijden, P.G.M., Bustami, R., Cruyff, M.J., Engbersen, G. & van Houwelingen, H.C. (2003). *Point and interval estimation of the population size using the truncated Poisson regression model.* Statistical Modelling, 3(4): 305-322.