Mathematical Statistics
Stockholm University

Estimating transition intensities in pure birth processes sampled in time

Åke Svensson,
Stockholm University and The Swedish Institute for Infectious Disease Control

Research Report 2005:11

ISSN 1650-0377

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Research Report **2005:11**,
http://www.math.su.se/matstat

Estimating transition intensities in pure birth processes sampled in time

Åke Svensson,

Octobre 2005

Abstract

We consider a situation where the state of a large number of independent pure birth processes, with common birth intensities, are observed at two time points. Estimates are derived, both when no restriction is placed on the intensities, and when they are assumed to follow a certain parametric model which assumes that the birth intensities are increasing. This is an assumptions that is sometimes made in the study of dynamic networks, where the births are interpreted as the occurrence of new edges. In this paper the theoretical results are applied to data from a survey of how the number of sexual contacts develops in time. The situation is generalized so that the birth intensities also depend on individual random factors. Different goodness-of-fit test of the models are considered.

Keywords: Branching processes, sexual networks.

*This is a technical report that contains the theoretical statistical background for an analysis of data from Swedish and Norwegian surveys or sexual behavior. The final results will be presented in forthcoming reports together with Fredrik Liljeros, Stockholm university, and Birgitte Freiesleben de Blasio, university of Oslo.

Estimating transition intensities in pure birth processes sampled in time

Åke Svensson

The Swedish Institute for Infectious Disease Control and Stockholm University

November 1, 2005

Abstract

We consider a situation where the state of a large number of independent pure birth processes, with common birth intensities, are observed at two time points. Estimates are derived, both when no restriction is placed on the intensities, and when they are assumed to follow a certain parametric model which assumes that the birth intensities are increasing. This is an assumptions that is sometimes made in the study of dynamic networks, where the births are interpreted as the occurrence of new edges. In this paper the theoretical results are applied to data from a survey of how the number of sexual contacts develops in time. The situation is generalized so that the birth intensities also depend on individual random factors. Different goodness-of-fit test of the models are considered.

1. Introduction

The purpose of this paper is to discuss how to estimate jump intensities from observations on several independent counting processes, that are observed at (two) fixed time points. Each of n process, $N_1(\cdot), N_2(\cdot), \ldots, N_n(\cdot)$, counts the number events of some kind in time. For convenience we assume that the first observation is taken at time $t = 0$. Thus, $N_i(0)$ is the number of events in the i'th processes that has occurred before time $t = 0$. Each process is also observed a second (non–random) time $T_i > 0$, that may be different for the different processes. The analysis will thus be based on the observations $(N_i(0), N_i(T_i))$, $i = 1, \ldots, n$.

Observations of this kind may occur in many situations. In section 5 the results are applied to data from a survey on the number of sexual partners. A random sample of persons have answered a question concerning the number of new partners during the last year and the number of previous partners. The value of the individual processes at time $t = 0$ is thus the number of previous partners, and the values after one year the sum of previous and new partners. In this case all T_i equal one year.

We will regard the processes as pure birth processes, there the time till a new event depends on the number of previous events. The time a process stays in state j is assumed to be exponentially distributed and the times in the different stages are assumed to be independent.

The statistical analysis will be based on two models, one with fixed effects, and one which involves individual random effects. First we will assume that the intensity of a jump

from stage j to $j+1$, denoted by π_j, is common for all n processes. Estimates of of the jump intensities will be derived both in a non-parametric setting, i.e. no relation between the values of the intensities is assumed, and in a parametric model where $\pi_j = \pi_j(\theta)$, $k = 1, 2, \ldots$, are assumed to be given as values of a function of a low-dimensional parameter θ. In particular we will be interested in the case where $\pi_0 = \beta$ and $\pi_j = \gamma j^\delta$ when $j \geq 1$. In this case the parameter $\theta = (\beta, \gamma, \delta)$.

In the application studied in section no auxiliary variables, such as age or social-economic factors, are used in the analysis. One could expect that there is a considerable variation between individuals that makes the assumption that all person have the same jump intensities questionable. For this reason a second, generalized, model is considered. It is assumed that there is an individual (random) multiplicative factor κ_i, which makes the jump intensity vector for individual i equal to $\kappa_i \pi_j$, $j = 0, \ldots$. The multiplicative factors are assumed to be drawn independently from a gamma distribution.

In section 2 we will calculate the likelihood and derive estimates and their asymptotic properties under the assumptions of fixed jump intensities. In section 3 we will give corresponding results for the model with individual random factors. Different ways to measure goodness–of–fit are discussed in section 4. Finally the theoretical results are applied in section 5.

2. Model with fixed jump intensities

The statistical interest is focused on the vector of jump intensities is $\pi = (\pi_0, \pi_1, \ldots)$. The pure birth process model implies that the time a process stays in state j is exponentially distributed with mean $1/\pi_j$, and that the times in the different stages are independent (cf Feller (1968)). This property is used to derive the likelihood.

2.1 *Likelihoods*

Assume that we have a birth process with birth intensities, $\mu_0, \mu_1, \ldots$ starting in state $N(0) = 0$. Let τ_i be the time the process spends in state i, $i = 1, 2, \ldots$. Then

$$\Pr(N_i(T) = v) = \Pr(\tau_0 \leq T, \ldots, \sum_{j=0}^{v-1} \tau_j \leq T, \sum_{j=0}^{v} \tau_j > T). \tag{2.1}$$

Since τ_j, $j = 1, \ldots, v$ are independent and exponentially distributed random variables we find that this probability can be written as:

$$\Pr(N_i(T) = v) = \prod_{j=0}^{v-1} \mu_j \int\int_{\sum_{j=0}^{v} t_i = T} \exp(-\sum_{j=0}^{v} \mu_j t_j) dt_0 \ldots dt_v. \tag{2.2}$$

A more explicit expression for these probabilities can be obtained either by successive partial integrations or by applying Kolmogorovs forward equations (cf. Vadeby (2004)).

$$C_v(\mu_0, \ldots, \mu_v, T) = \int\int_{\sum_{j=0}^{v} t_j = T} \exp(-\sum_{j=0}^{v} \mu_j t_j) dt_0 \ldots dt_v = \sum_{k=0}^{v} \frac{\exp(-T\mu_k)}{\prod_{j \neq k} (\mu_j - \mu_k)}. \tag{2.3}$$

If some of the intensities are equal, i.e. $\mu_k = \mu_{k'}$ then the expression should be interpreted as a limit when $\mu_k \to \mu_{k'}$. Assume that $\mu_0, \ldots, \mu_v$ takes the m distinct values $\tilde{\mu}_1, \ldots, \tilde{\mu}_m$ with the frequencies $r_1, \ldots, r_m$, where $\sum r_i = v$, then

$$C_v(\mu_0, \ldots, \mu_v, T) = \prod_{i=1}^{m} \frac{T_i^{r_i-1}}{(r_i-1)!} \frac{\partial^{r_i-1}}{\partial \tilde{\mu}_i^{r_i-1}} \sum_{k=1}^{m} \frac{\exp(-T\tilde{\mu}_k)}{\prod_{j\neq k}(\tilde{\mu}_j - \tilde{\mu}_k)}. \tag{2.4}$$

It is well-known that if all intensities are equal, i.e., $\tilde{\mu} = \mu_0 = \mu_1, \ldots = \mu_v$ the probabilities are given by Poisson probabilities and $C_v(\tilde{\mu}, \ldots, \tilde{\mu}, T) = T^v \exp(-T\tilde{\mu})/v!$.

We can use the expressions (2.2) and (2.3 or 2.4) to calculate the probability that a birth processes with initial value $N_i(0) = s_i$ is in state s_i+v_i at time $t = T_i$. The probability is obtained by inserting $T = T_i$, $v = v_i$, and $\mu_j = \pi_{s_i+j}$, $j = 0, \ldots, v$ in (2.3 or 2.4).

After some simplifications, the following expression for the logarithm of the likelihood, for a set of observed independent processes, is obtained:

$$\sum_j a_j \ln(\pi_j) + \sum_i \ln(C_{v_i}(\pi_{s_i}, \ldots, \pi_{s_i+v_i}, T_i)), \tag{2.5}$$

where a_j is the number of observed jumps from stage j, $j = 0, 1, \ldots$.

The time R_{ij} that the i'th process spends in state j is important when deriving estimates. It turns out that this time is closely related to the functions C. In fact if $s_i \leq j \leq s_i + v_i$ then (cf. (2.2))

$$-\frac{\partial \ln(C_{v_i}(\pi_{s_i}, \ldots, \pi_{s_i+v_i}, T_i))}{\partial \pi_j} = \mathrm{E}_\pi(R_{ij} \mid N_i(0) = s_i, N_i(T_i) = s_i + v_i). \tag{2.6}$$

2.2 *Non-parametric estimation of intensities*

Without any parametric assumptions on the structure of the birth intensities a formal derivation of the ML-estimates yields the ML-equations:

$$\frac{a_j}{\pi_j} = \sum_{i=1}^{n} \mathrm{E}_\pi(R_{ij} \mid N_i(0), N_i(t_i)), \tag{2.7}$$

$j = 0, 1, \ldots$.

It should be noted that these equations do not always have a unique solution. From (2.3) it is clear that $\mathrm{E}_\pi(R_{ij})$ are symmetric in the parameters $\pi_{s_i}, \ldots, \pi_{s_i+v_i}$. If every process that spends time in state j also spends time in state $j+1$ the right-hand term of the equation (2.7) takes the same value if we switch π_j and π_{j+1}. In the following we will resolve this indeterminacy by (arbitrarily) assuming that all intensities in such states have the same value.

A convenient algorithm to solve (2.7) is suggested by the EM-algorithm. This algorithm calculates the expectations in the right-hand term of the equation for given values of the parameters π_j and updates these values iteratively by solving (2.7).

We can not immediately apply standard asymptotic theory for ML-estimates since we have a infinite number of unknown parameters. However, under suitable regularity assumptions it is possible to prove that the estimate of π_j is asymptotically normal distributed provided the corresponding a_j is large. We can not expect all a_j to be large. In fact there will be no jump from the the state $\max(s_i + N_i(T_i))$, even if some time is spent in that state.

2.3 *Estimates in the parametric model*

We will consider the parametric model in which

$$\pi_j(\beta, \gamma, \delta) = \begin{cases} \beta & \text{if } j = 0, \\ \gamma j^{\delta} & \text{if } j \geq 1. \end{cases} \tag{2.8}$$

Using (2.6) we see that the ML-estimates of the three parameters δ, β and γ solve the system of equations:

$$\begin{aligned} a_0 &= \beta \sum_j \frac{d\pi_j}{d\beta} \sum_i \mathrm{E}_{\pi}(R_{ij} \mid N_i(0), N_i(T_i)) \\ &= \beta \sum_i \mathrm{E}_{\pi(\beta,\gamma,\delta)}(R_{i0} \mid N_i(0), N_i(T_i)) \\ \sum_{j\geq 1} a_j &= \gamma \sum_j \frac{d\pi_j}{d\gamma} \sum_i \mathrm{E}_{\pi}(R_{ij} \mid N_i(0), N_i(T_i)) \\ &= \gamma \sum_{j\geq 1} \sum_i \mathrm{E}_{\pi(\beta,\gamma,\delta)}(R_{ij} \mid N_i(0), N_i(T_i)) j^{\delta} \\ \sum_{j\geq 1} a_j \ln(j) &= \sum_j \frac{d\pi_j}{d\delta} \sum_i \mathrm{E}_{\pi}(R_{ij} \mid N_i(0), N_i(t_i))) \\ &= \sum_{j\geq 1} \sum_i \mathrm{E}_{\pi(\beta,\gamma,\delta)}(R_{ij} \mid N_i(0), N_i(T_i)) j^{\delta} \ln(j). \end{aligned} \tag{2.9}$$

Also in this case a convenient algorithm is suggested by the EM-algorithm. The expected values for the times in the states are calculated for given values of (β, γ, δ), using equation (2.6). The parameters are updated by solving the equations (2.9).

Standard ML-theory implies that these estimates are asymptotically normal distributed with a variance-covariance matrix which is the inverse of the Fisher information matrix. Applying the expressions (2.3) and (2.5) some calculations yield that the elements in the Fisher information matrix for the parameters (β, γ, δ) can be derived as:

$$
\begin{aligned}
I(1,1) &= -\mathrm{E}(a_0)/\beta^2 + \sum_i \mathrm{Var}_\pi(R_{i0}),\\
I(1,2) = I(2,1) &= \sum_i \sum_{j\geq 1} j^\delta \mathrm{Cov}_\pi(R_{ij}, R_{i0}),\\
I(1,3) = I(3,1) &= \sum_i \sum_{j\geq 1} \gamma j^\delta \ln(j) \mathrm{Cov}_\pi(R_{ij}, R_{i0}),\\
I(2,2) &= -\sum_{j\geq 1} \mathrm{E}(a_j)/\gamma^2 + \sum_i \sum_{j\geq 1} \sum_{k\geq 1} j^\delta k^\delta \mathrm{Cov}_\pi(R_{ij}, R_{ik}),\\
I(2,3) = I(3,2) &= -\sum_i \sum_{j\geq 1} j^\delta \ln(j) \mathrm{E}_\pi(R_{ij}) + \gamma \sum_i \sum_{j\geq 1} \sum_{k\geq 1} j^\delta k^\delta \ln(k) \mathrm{Cov}_\pi(R_{ij}, R_{ik}),\\
I(3,3) &= -\gamma \sum_i \sum_{j\geq 1} j^\delta (\ln(j))^2 \mathrm{E}_\pi(R_{ij}) + \gamma^2 \sum_i \sum_{j\geq 1} \sum_{k\geq 1} j^\delta \ln(j) k^\delta \ln(k) \mathrm{Cov}_\pi(R_{ij}, R_{ik}).
\end{aligned}
\tag{2.10}
$$

The information matrix may be approximated by inserting the estimate of the parameters and of the vector of intensities. The asymptotic variance-covariance matrix of the parameter estimates are then derived as the inverse of the resulting approximate matrix.

2.4 *Numerical approximations of the expected times in states and their covariances*

The formula (2.6) gives an exact expression of the expected time in a state given the initail and final state. A straightforward use of the expression may cause serious numerical problems since the expression is evaluated as a sum of positive and negative numbers that may be quite large. An alternative method is to approximate the expected time using Monte Carlo simulations. This can be done choosing $v+1$-dimensional random vectors $(t_0, t_1, \ldots, t_v)$ uniformly distributed on the simplex $\sum_{i=1}^{v} t_i = T_i$, then

$$
\exp(-\sum_{i=0}^{v} \mu_i t_i) \tag{2.11}
$$

has the expectation $C_v(\mu_0, \ldots, \mu_v, T)$. The random variable

$$
t_j \exp(-\sum_{i=0}^{v} \mu_i t_i) \tag{2.12}
$$

has the expectation $-\partial C_v(\mu_0, \ldots, \mu_v, T)/\partial \mu_j$.

By repeating this procedure a large number of times and taking means of the resulting random vectors given by (2.11) and (2.12), we can approximate $C_{v_i}(\pi_{s_i}, \ldots, \pi_{s_i+v_i}, T_i)$ and $\partial C_{v_i}(\pi_{s_i}, \ldots, \pi_{s_i+v_i}, T_i)/\partial \pi_j$ with arbitrary precision. Taking the ratio of these means we get an approximation of $\mathrm{E}_\pi(R_{ij} \mid N_i(0) = s_i, N_i(T_i) = s_i + v_i)$.

A simple way to generate random vectors for the simulation is to choose $v+1$ independent random variables, that are exponentially distributed with mean 1 and let $t_i = T_i z_i / \sum_j z_j$.

This Monte Carlo method can also be used also to derive numerical approximations of the Fisher information, which involves second moments of the times R_{ij}.

2.5 *Confidence intervals of estimates in the parametric model*

There are several ways to study the precision of estimates of parameters. We will here suggest three possible ways to derive confidence intervals

- asymptotic theory for ML-estimates,
- partial likelihood for the parameter δ, and
- parametric bootstrap simulations.

Provided that the number of observed processes are large it should be possible to use asymptotic results for ML-estimates. This requires that certain regularity conditions have to be satisfied. We will not investigate this further in this paper. The asymptotic theory, implies that the variances of the parameter estimates can be approximated using the inverse of the Fisher information matrix (2.10).

The partial likelihood for the parameter δ is given by the maximum attainable likelihood, by varying the parameters β and γ, for different values of the parameter δ (cf. Barndorff-Nielsen and Cox (1989)). Confidence intervals for the parameter δ are derived from this partial likelihood as the set of δ-values that has a log likelihood that does not differ to much from the maximal log likelihood. Partial likelihoods for the other two parameters is obtained in a similar way.

The parametric bootstrap simulations are obtained by simulating birth and death processes with the given starting values $N_i(0)$ and the estimated parameter values (cf Efron and Tibshirani (1993)). For each simulation the parameters are re-estimated. The empirical distribution of thee bootstrapped estimates are used to evaluate the properties of the estimators.

3. Parametric model with random factors

In this model we assume that an individual has the jump intensities $\kappa\pi_i$, $i = 0, 1, \ldots$, where κ is a gamma distributed random variable with parameters (α, α). We will here only consider the parametric model where π_i are given by (2.8). The density function of κ is

$$g_\alpha(\kappa) = \frac{\alpha^\alpha}{\Gamma(\alpha)}\kappa^{\alpha-1}\exp(-\alpha\kappa), \tag{3.1}$$

$\mathrm{E}(\kappa) = 1$ and $\mathrm{Var}(\kappa) = 1/\alpha$.

The probability that a pure birth process, with jump intensities $\kappa\mu_0, \kappa\mu_1, \ldots$, that starts in $N_i(0) = 0$ end in state $N_i(T) = v$, is

$$\Pr(N_i(T) = v) = \prod_{j=0}^{v-1}\mu_j \int_0^\infty \left[\iint_{\sum_{j=0}^{v} t_i = T} \kappa^v \exp(-\kappa \sum_{j=0}^{v} \mu_j t_j) dt_0 \ldots dt_v \right] g_\alpha(\kappa) d\kappa. \tag{3.2}$$

Simple calculations, using (2.2) and (2.3) yields that

$$\Pr(N_i(T) = v) = \prod_{j=0}^{v-1} \mu_j \sum_{k=0}^{v} \frac{(1 + T\mu_k/\alpha)^{-\alpha}}{\prod_{j \neq k} (\mu_j - \mu_k)}. \tag{3.3}$$

In this model the functions

$$\begin{aligned} D_v(\mu_0, \ldots, \mu_v, T, \alpha) &= \int_0^\infty \left[\iint_{\sum_{j=0}^{v} t_j = T} \kappa^v \exp(-\kappa \sum_{j=0}^{v} \mu_j t_j) dt_0 \ldots dt_v \right] g_\alpha(\kappa) d\kappa \\ &= \sum_{k=0}^{v} \frac{(1 + T\mu_k/\alpha)^{-\alpha}}{\prod_{j \neq k} (\mu_j - \mu_k)} \end{aligned} \tag{3.4}$$

plays a role corresponding to the functions C_v in the model with fixed jump intensities. Here

$$-\frac{\partial \ln(D_{v_i}(\pi_{s_i}, \ldots, \pi_{s_i+v_i}, T_i, \alpha))}{\partial \pi_j} = \mathrm{E}_\pi(\kappa_i R_{ij} \mid N_i(0) = s_i, N_i(T_i) = s_i + v_i). \tag{3.5}$$

The logarithm of the likelihood for a set of observed independent processes equals

$$\sum_j a_j \ln(\pi_j) + \sum_i \ln(D_{v_i}(\pi_{s_i}, \ldots, \pi_{s_i+v_i}, T_i, \alpha)). \tag{3.6}$$

3.1 *Estimates in the parametric model with random effects*

Compared with the model with fixed effect the model with random gamma-distributed individual multiplicative effects has one further parameter, namely α. An ML-estimate of this parameter can be obtained by considering the profile-log-likelihood, i.e., the maximum of the log-likelihood (3.6), for given values of α. The profile likelihood is obtained by solving likelihood equations similar to the equations (2.9), where $\mathrm{E}_\pi(R_{ij} \mid N_i(0), N_i(T_i)))$ is replaced by the expectations $\mathrm{E}_{\pi,\alpha}(\kappa_i R_{ij} \mid N_i(0), N_i(T_i)))$.

In the same way the Fisher information matrix, for a fixed value of α, is obtained from a formula similar to (2.10) where the first and second moments are calculated for $\kappa_i R_{ij}$ instead of R_{ij}.

3.2 *Numerical approximations of expected times in state and their covariances*

The last expression in (3.4) can be used to simulate an estimate of the expected value of $\kappa_i R_{ij}$. Random vectors $(t_1, \ldots, t_{v+1})$ that are uniformly distributed on the simplex $\sum t_i = 1$ are generated in the way suggested in section (2.4).

The random variable

$$\frac{\Gamma(\alpha+v)\alpha^{\alpha}}{\Gamma(\alpha)(\alpha+\sum\mu_i t_i)^{\alpha+v}} \tag{3.7}$$

has the expectation $D_v(\mu_o,\ldots,\mu_v,T,\alpha)$ and

$$t_j\frac{\Gamma(\alpha+v+1)\alpha^{\alpha}}{\Gamma(\alpha)(\alpha+\sum\mu_i t_i)^{\alpha+v+1}} \tag{3.8}$$

has the expectation $-\partial D_v(\mu_o,\ldots,\mu_v,T,\alpha)$. Producing a large number of such estimates we can approximate $\mathrm{E}_{\mathrm{pi},\alpha}(\kappa_i R_{ij} \mid N_i(0)=s_i, N_i(T_i)=s_i+v_i)$ as the ratio between the means of the estimates obtained from (3.7) and (3.8). In a similar way approximations of the elements of the Fisher information matrix can be obtained.

4. Goodness-of-fit tests based on likelihood ratios

The suggested models are very simple and uses a low-dimensional parameter. It is thus necessary to investigate the fit of the model. In the following discussion of the goodness-of-fit tests we will assume that all values of T_i are the same, i.e. that all processes are observed after an identical time lap, T.

We will derive a measure of deviance between the parametric model and the observed data by comparing the parametric model with a crude model where we assume that the final state of the process only depends on the initial state, i.e.

$$\Pr(N_i(T)=f \mid N_i(0)=s)=p_{sf}.$$

Let n_s denote the number of processes that starts in state s, and n_{sf} the number that starts in state ends in state f. The log of the likelihood ratio for the parametric model versus this crude model is

$$-2\ln(LR)=\sum\ln(\Pr(N_i(T)=f_i \mid N_i(0)=s_i)-\sum n_{s_i f_i}\ln(n_{s_i f_i}/n_{s_i}). \tag{4.1}$$

Here the probabilities $\Pr(N_i(T)=f_i \mid N_i(0)=s_i)$ are calculated using (2.2) or (3.2), depending on if a model with fixed or random effects are considered, with the estimated parameter values. The deviance is often used to evaluate the fit of a simple model. In regular cases it is asymptotically χ^2-distributed if the null hypothesis holds . In this case it is not clear that this is a good approximation and it is not clear how to calculate the relevant degree of freedom. As a standard the degree of freedom used is the difference of the dimension of the statistics that are used to calculate the likelihoods in the two models. Here the crude model has, in principle, an infinite number of parameters.

An alternative way to find the distribution of the deviance, under the null hypothesis, is to apply a (bootstrap) simulation technique. A number of values of n-dimensional vectors $(N_1(T),\ldots,N_n(T))$ are simulated using the observed values of $N_i(0)$ and the estimated values of the probabilities in the model. For each such simulation a value of a deviance is calculated. This result in an empirical distribution of simulated deviances. The observed deviance is then compared with this distribution.

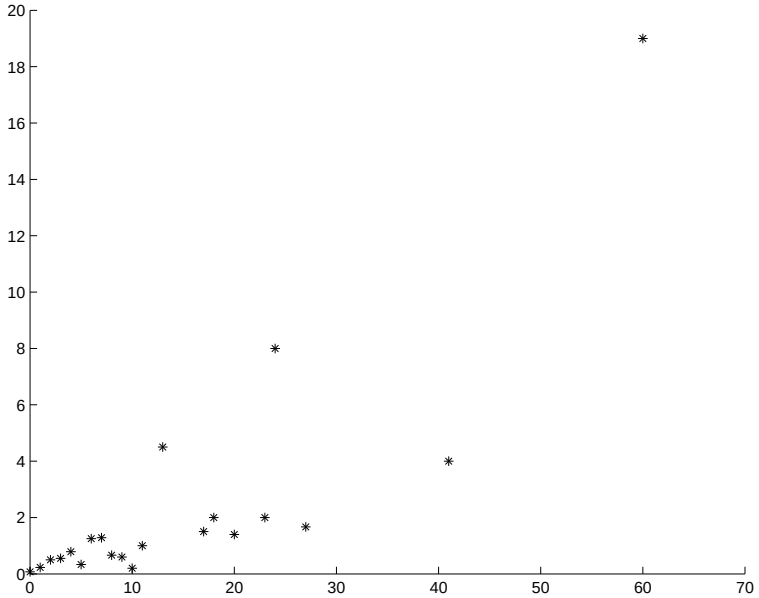

Figure 5.1. Relation between the mean number of new partners year 0-4, initial state (s_i), and the number of new partners year 5, ($f_i - s_i$)

5. An example analyzed

We will consider the following example. The observations come from a study of the sexual behavior. In this study 801 males, have been asked for their number of new sexual partners during the last 5 years. The starting values s_i are the number of new sexual partners year 0-4, and the final value f_i is the number of new partners year 0-5. The births corresponds to new partners and the model describes how the number of new sexual partners within one year relates to the history of partners during the preceding 4 years. In this situation all individuals are observed during one year, i.e, all values of T_i are equal, $T_i = 1$.

The data is illustrated in figure 5.1, which shows the relation between the the starting value and the number of new partners the last year by plotting the means of the number of new partners for all persons with the same initial number of partners.

This figure seems to indicate that $f_i - s_i = N_i(1) - N_i(0)$ grows with $s_i = N_i(0)$.

In the following analysis we have not included the extreme observations with $s_i = 60$ and $f_i = 80$. Thus 800 observations remain.

5.1 *Estimates in the non-parametric model and the parametric model with fixed effects*

The non-parametric estimates of the jump intensities are shown in figure 5.2. We only show the estimates for $j \leq 12$, since the non-parametric estimates behaves very unstable for larger values of j. The oscillating behavior of the non-parametric estimates is due to the fact that estimates of subsequent intensities are highly negatively correlated.

The estimates and the confidence intervals are summarized in table (5.1). A histogram of 1000 bootstrapped estimated of the parameter δ is given in figure 5.1.

The estimates and confidence intervals are summarized in the table 5.1.

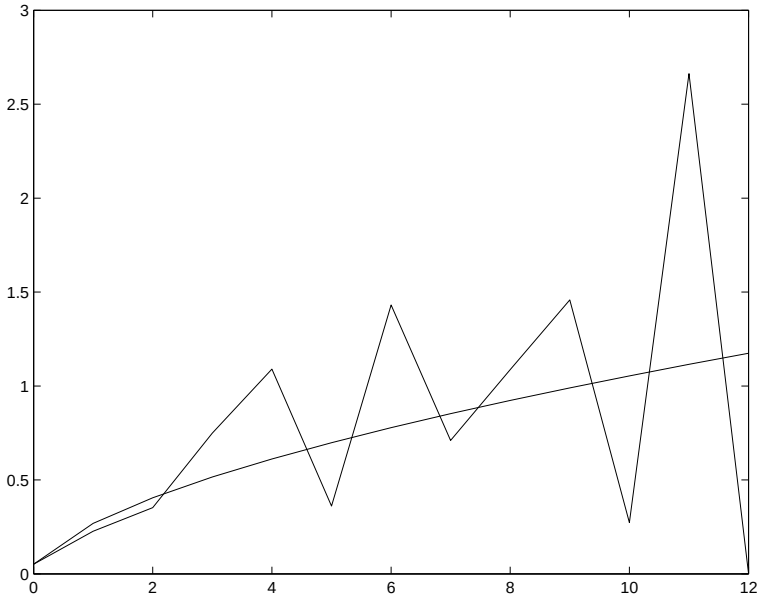

Figure 5.2. Estimates of birth intensities for the parametric model (with fixed effects) and the non-parametric model

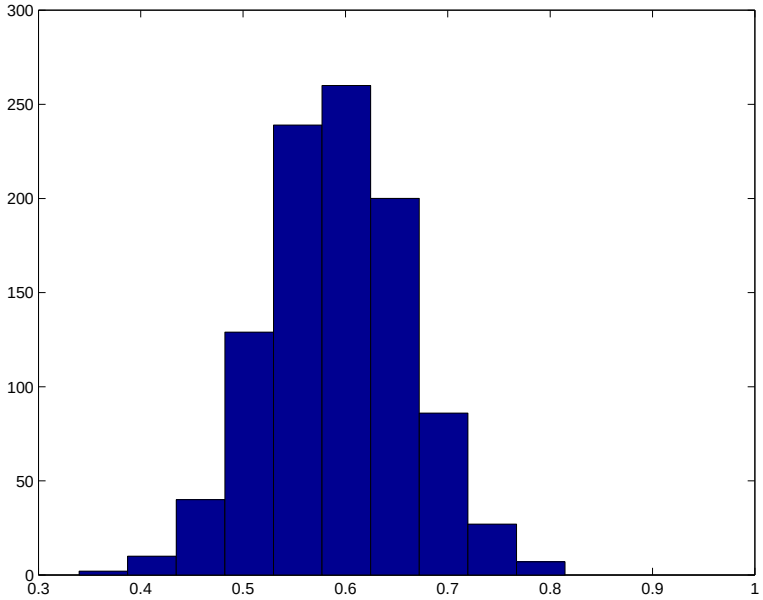

Figure 5.3. Histogram of 1000 bootstrapped estimates of δ

Table 5.1

Estimates for parametric model with fixed effects and 95 % confidence intervals, based on 800 observations.

parameter	estimate	normation	asymptotic 95 % ci	profile likelihood 95 % ci	Bootstrap 95 % ci
β	0.052	per year	(0.032,0.072)	(0.035, 0.073)	(0.034, 0.070)
γ	0.27	per year	(0.20,0.34)	(0.23,0.31)	(0.20,0.34)
δ	0.59	-	(0.46,0.73)	(0.46, 0.72)	(0.46, 0.73)

Table 5.2

Estimates for parametric model with random factors and 95 % confidence intervals.

parameter	estimate	normation	Bootstrap 95 % ci
β	0.053	per year	(0.034, 0.077)
γ	0.26	per year	(0.18,0.35)
δ	0.59		(0.39, 0.74)

5.2 *Estimates in the parametric model with random factors*

Compared to the model with fixed effects the model with gamma-distributed random factors involves one extra parameter. In figure the profile likelihood for this parameter is illustrated. The ML-estimate of $\alpha = 1.09$. The profile likelihood gives the 95 % confidence interval $[0.62, 2.01]$.

In table 5.2 the estimates of the other parameters are given with confidence intervals derived from parametric bootstrapping. As can be expected the confidence intervals are somewhat broader than the confidence interval derived from the fixed model. However the estimates does not differ much. This is illustrated by figure 5.2, which gives the estimated jumps intensities from different states for different values of κ. The dotted line illustrates the jump intensities estimated from the model with fixed effects. This is very close to the estimates with $\kappa = 1$.

In figure 5.2 the bootstrapped estimates of δ are illustrated.

5.3 *Deviances of the parametric models*

We will compare the two parametric models with a crude model using simulations as described above. The results are summarized in table 5.3.

From the table 5.3 it is clear that the observed deviance, when all observations are used, is considerable larger than can be expected for the model with fixed effects. It obviously indicates that the parameter estimates and their confidence intervals have to interpreted with great care. However, the model with random effects seems to have a reasonable fit as

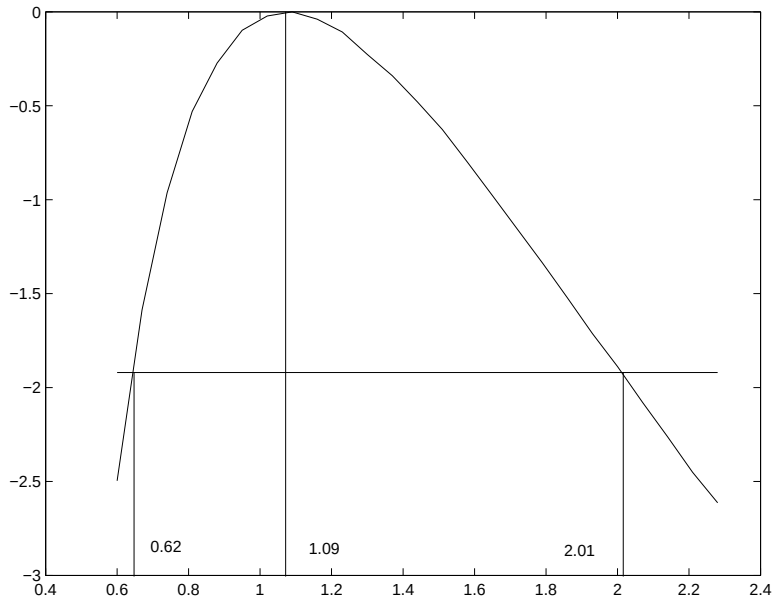

Figure 5.4. Profile likelihood for the parameter, α, in the gamma-distribution of the random factors

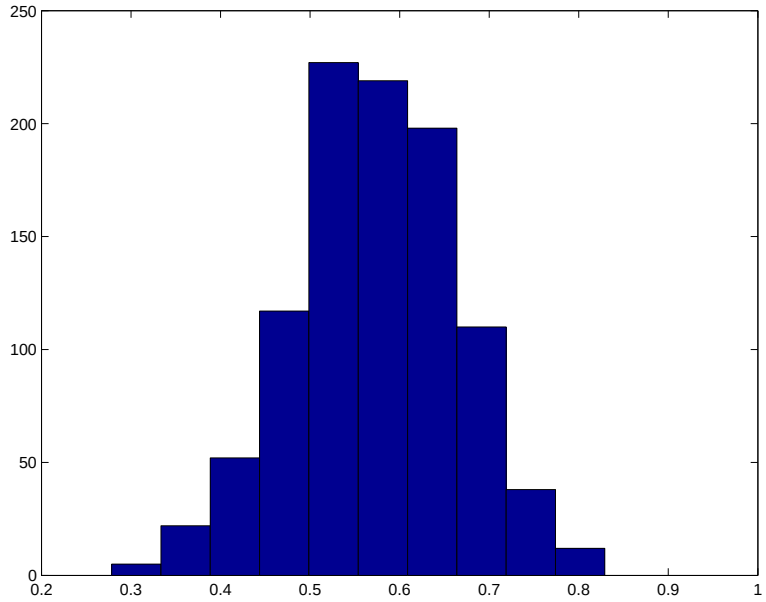

Figure 5.5. Histogram of 1000 bootstrapped estimates of δ

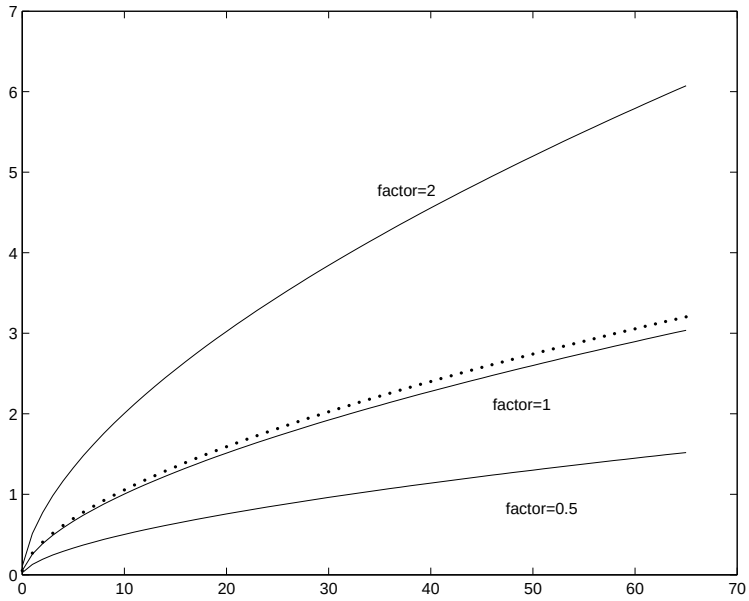

Figure 5.6. Estimated jump intensities for individuals with different factor (solid lines) and estimated jump intensities for the model with fixed effects (dotted line).

Table 5.3
Deviances for the two parametric models

observations	deviance (relative the basic model)	p-value estimated from simulations	mean of bootstrapped deviances	std of bootstrapped deviances
fixed effects	162.9	<0.01	114.5	10.7
random factor	128.7	≈ 0.15	114.5	13.1

measured by the deviance.

An analysis, based on residuals, which is not presented here indicates that the lack of fit of the fixed effect model is not caused by a single or a few outlying observations.

5.4 *Comparisons of statistics and their estimated expected values*

A alternative way to evaluate the fit of a model is to study differences between observed values of particularly interesting statistics and the value that could be expected if the model was true. We will here use estimates from the analysis with all observations, i.e no potential outliers are removed. Of course, the result depends on the test statistics used.

First we consider the number of jumps from different states. Table 5.4 gives both the observed value and the expected value that can be calculated from the parametric model (under the assumption that the estimated parameter values are the true values). It can be observed that the two models yields very similar expected counts and seem to be in very good agreement with the observed numbers.

A second statistic is the number of new partners (regardless of the number of previous partners). Table 5.5 gives the observed counts and their estimated expected values. There seems to be a rather large deviation between the observations and the estimates from the model with fixed effects. However, the estimated number of new partners derived from the random effect model seems to be closed to what is observed. The χ^2–value for the deviation is 1.82. This value correspond to the 0.40-percentile in a χ^2-distribution with 2 degrees of freedom.

The difference between the fit of the models to these two statistics can possibly be explained by the fact that the number of jumps from specific states depends on the mean intensities over different individuals, which is the same in the two models, whereas the number of jumps of different individuals also reflects the variation in the birth intensities between individuals.

Acknowledgements

This reports contains the theoretical statistical background for an analysis of data from Swedish and a Norwegian surveys of sexual behaviors. The complete analysis will be given in forthcoming reports together with Fredrik Liljeros, Stockholm university and Birgitte Freiesleben de Blasio, University of Oslo.

Table 5.4

Number of jumps from states 0 till 20

state	Observed No. of jumps from state	Expected No. fixed effects	Expected No. random effects
0	26	26.2	26.2
1	25	28.8	27.6
2	22	25.8	25.0
3	23	17.3	17.7
4	21	12.9	13.4
5	8	11.0	11.3
6	6	5.5	6.4
7	5	5.6	6.0
8	4	3.7	4.2
9	6	4.4	4.5
10	2	5.0	4.7
11	2	2.7	2.9
12	0	2.4	2.6
13	2	2.4	2.5
14	2	1.0	1.3
15	2	1.1	1.3
16	2	0.5	0.7
17	5	3.2	2.8
18	3	2.5	2.2
19	1	1.2	1.3
20	3	4.5	3.8

Table 5.5

Observed and expected number of new partners

No of new partners	Observed	estimated fixed effects	estimated random effects
0	683	664.7	684.0
1	80	95.4	75.1
2	18	27.5	23.2
3	9	8.4	9.1
4	4	2.6	4.1
>4	6	1.4	4.7
total	800	800	800

References

Barndorff-Nielsen, O. and Cox, D. (1989). *Asymptotic techniques: for use in statistics*. Chapman and Hall.

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

Feller, W. (1968). *An introduction to probability theory and its applications, Vol I*. Wiley, 3rd edition.

Vadeby, A. (2004). Modeling of relative collision safety including driver characteristics. *Accident Anal Prev* **36**, 909–917.