

# The Analysis of Panel Data with a Flexible Frailty

by

Parvin Dehghani, M.Sc

A thesis submitted to the Faculty of Graduate and Postdoctoral  
Affairs in partial fulfillment of the requirements for the degree of

Master of Science in Statistics

In

Ottawa Institute of Mathematics and Statistics  
School of Mathematics and Statistics

Carleton University  
Ottawa, Ontario  
January, 2018

©Copyright  
2018, Parvin Dehghani

## **Abstract**

Analysis of panel data under the non-homogeneous Poisson mixed process is investigated. The ordinary Poisson process model is not suitable for the analysis of count data in the presence of overdispersion. For handling the overdispersion problem, a multiplicative finite mixture of gamma distributed random effects is assigned to each individual in the model which gives a non-homogeneous Poisson mixed model. The parameter of the mixed Poisson process (baseline intensity function) is modeled by using psplines and the parameters of the model are estimated by means of the iterative Expectation-Maximization (EM) algorithm. The statistical properties of the proposed count model are investigated through simulation study and the model is used to analyze the Cherry Bark Tortrix Moth data.

## **Acknowledgements**

I would like to express my sincere gratitude to my graduate supervisor Dr Jason Nielsen for his continuous support of my master study. He showed me new ways to approach and solve research problems in data science. His patient guidance, advice, feedback and ideas have been invaluable.

I am thankful to the school of Mathematics and Statistics for providing all the required facilities and computer resources. I would like to send my special regards to Nicki Gaertner and Kevin Crosby for all the assistance they have provided me during my graduate study.

I must also express my very profound gratitude to my family members for providing me with continuous support of my study throughout my life.

# Table of Contents

<b>Abstract.....</b>	<b>ii</b>
<b>Acknowledgements .....</b>	<b>iii</b>
<b>Table of Contents</b>	<b>iv</b>
<b>1 Chapter: Introduction .....</b>	<b>1</b>
<b>2 Chapter: Background.....</b>	<b>3</b>
2.1 Splines	3
2.1.1 B-Splines	3
2.1.2 Smoothing Splines (Penalized Splines)	7
2.1.3 P-Splines	9
2.1.4 Selection of Smoothing Parameter	11
2.2 Finite Mixture Models	12
2.3 Recurrent Event Data	13
2.3.1 Heterogeneity	14
2.4 Expectation-Maximization Algorithm	15
<b>3 Chapter: Methodology</b>	<b>22</b>
3.1 Poisson Process	22
3.1.1 Non-homogeneous Poisson Process	24
3.1.2 Proportional Intensity Assumption	26
3.1.3 Models for Heterogeneity	27
3.2 Semiparametric Analysis for Recurrent Events	29
3.2.1 Semiparametric Nonhomogeneous Poisson Process Model	29

<b>4 Chapter: Analysis</b>	<b>36</b>
4.1 Simulation	36
4.2 Data Analysis	43
<b>5 Chapter: Conclusion</b>	<b>48</b>
<b>References</b>	<b>50</b>

## List of Figures

1	A display of b-splines up to degree three	6
2	A plot of the frailty's density function against the estimated function in a simulation study	38
3	A plot of the baseline intensity function versus the estimated baseline intensity function in a simulation study	39
4	The bias of the baseline intensity estimator in a simulation study	40
5	Bias for the proposed probability model's parameters	42
6	The plot of the frailty's density function modeled by cherry bark moth data	44
7	The plot of the baseline intensity function modeled by cherry bark moth data	47

# Chapter 1

## Introduction

A clinical trial is a set of experiments that are designed to evaluate the effects of biomedical or behavioral interventions such as drugs, treatments or devices on individual subjects in clinical research studies. In biomedical studies, an event is defined as a recurrent event if it can occur more than once for a participant. Two types of recurrent event data are event time data and panel data for which all individuals are followed over time. Panel data is a more popular kind of data where the total number of occurred events in a period of time is registered but for event time data, the exact time of occurrence should be recorded which is expensive and not always possible. Panel data used extensively in econometric, health and social sciences, is extracted from a surveyed sample of individuals throughout the time, and so represents multiple observations per individual in a given sample. In biostatistics context, panel data usually shows the number of events in a time interval or the number of transitions between a few states over discrete time points. In this thesis, analysis and modeling of panel data is explored by using Poisson process defined as a special case of a renewal process in probability theory. In a renewal process, a set of independent and identical exponentially distributed inter-arrival times lead to a Poisson count model. Over-dispersion is an important problem of modeling count data by a Poisson process in which the mean should be equal to the variance.

In order to take over-dispersion into account in the Poisson process, a specific random effect for each individual is added to the Poisson model. Therefore the data are modeled by a mixed Poisson model where a probability density function of subject's random effect

is added to the Poisson model. The final model is constructed based on density function of a random effect and the baseline intensity function (parameter) of the non-homogeneous Poisson process that is estimated by non-parametric models such as b-splines. All parameters of the model are estimated by using the iterative Expectation-Maximization algorithm in which each iteration contains two processes: Expectation step and Maximization step. Following estimation of parameters, a simulation study is done for evaluation of mixed Poisson model based on varying number of time panels and individuals.

In the chapter 2 (background), we will have a review of some basic concepts used in this research thesis for statistical modeling and analysis. The background chapter will cover some explanations about splines, finite mixture models, recurrent event data and the Expectation-Maximization algorithm. In chapter 3 (Methodology), the structure of a non-homogeneous Poisson mixed effect model will be explained in detail. The methodology will explain how a finite mixture of Negative binomial is constructed by using the concepts of non-homogeneous Poisson process, proportional intensity function, heterogeneity, finite mixture models and semiparametric analysis for recurrent event data. Chapter 4 (analysis) is applying the probability model obtained in chapter 3 for the simulation study to evaluate the statistical properties of the proposed method. In the analysis section, the results are represented in the form of some charts meanwhile we investigate the bias of all parameters based on different number of time panels and individuals. After interpreting the goodness of the probability model, the Cherry Bark Moth data is analyzed.

## **Chapter 2**

### **Background**

#### **2.1 Splines**

Classical modeling tools such as linear regression (Multiple regression, General Linear model) and nonlinear regression (Generalized Linear Models including: Log-linear models, models with linear and non-linear parameters) can be insufficient when trying to interpret nonlinear/complex relationship between variables. Smoothing splines a modern nonparametric analysis method has a high degree of flexibility and make no assumption about the relationship between variables. Instead, the relationship is constructed by a set of coefficients and basis functions that span a linear function space. This property of splines is especially useful for high dimensional input problems where we may need more complicated models. Some primary definitions and application of splines in data analysis and modeling will be presented in this chapter.

##### **2.1.1 B-splines**

A spline function  $f(x)$  is a piecewise polynomial function which is derived by partitioning the domain into some disjoint intervals and representing the function by a separate polynomial in each interval. The polynomials are constrained to join smoothly at places known as knots. The major properties of spline functions are the continuity of the polynomials and existence of some continuous derivatives at the knots. After indicating the location of the knots and polynomials in each interval, the spline function  $f(x)$  is:

$f(x) = \sum_{i=1}^m \beta_i B_i(x)$  where  $B_i(x)$  is the  $i$ th basis function and  $\beta_i$  is the  $i$ th corresponding

coefficient. As can be seen the spline function  $f(x)$  is a linear combinations of basis functions and coefficients. The order of the spline, the number of knots and their placement are key factors of a spline function. In general, a b-spline (a short form of basis spline) of order- $M$  is a piecewise polynomial of order  $M$  with continuous derivatives up to order  $M-2$  for a given set of knots. The b-spline basis functions can be obtained by the de Boor (1978) recursion formula which shows high efficiency in numerical computation even with a large number of knots.

According to de Boor formula, if  $S_n(x) = \sum_i \beta_i B_{i,n}(x)$  is a b-spline of order- $n$  with  $(a,b)$  defined as the domain and the sequence of  $k$ -inner knots  $t_1, t_2, \dots, t_k$ , then the  $i$ th b-spline basis function of order- $m$  is:

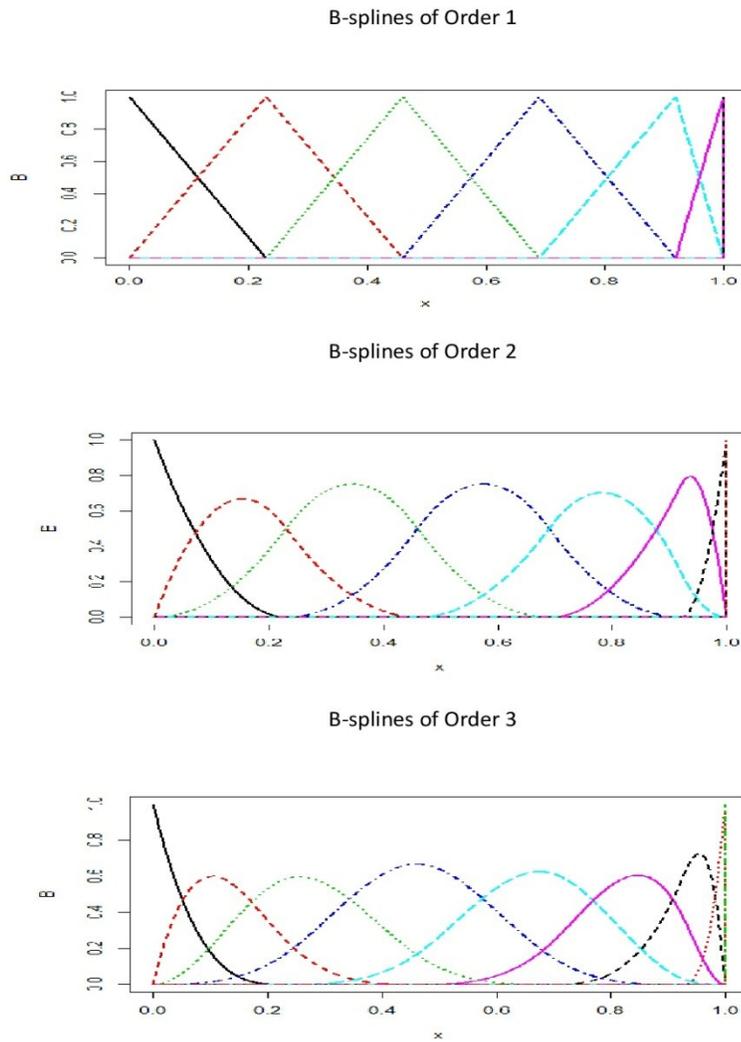
$$B_{i,m}(x) = \frac{x - t_i}{t_{i+m-1} - t_i} B_{i,m-1}(x) + \frac{t_{i+m} - x}{t_{i+m} - t_{i+1}} B_{i+1,m-1}(x) \text{ for } i=1, \dots, k+2n-m$$

Each basis function of order  $m$  ( $m = 1, 2, 3, \dots$ ) is recursively obtained based on basis function of order  $m-1$  which starts the recursion. The basis function of order-1 is  $B_{i,1}(x)$  for  $t_i < x < t_{i+1}$  where  $i=1, \dots, 2k-n$ , otherwise it is equal to zero.

The sequence of basis functions spans the space of splines for the given knots. For example, a cubic spline where  $n=4$ ,  $B_{i,4}$ ,  $i=1, \dots, k+4$  creates  $k+4$  cubic spline basis functions for  $k$  knots ( $t_1, t_2, \dots, t_k$ ). The plots of b-spline basis functions are useful for interpretation. Figure 2.1 represents the b-spline of order one to three. It is visible that other than degree one, all polynomial pieces are continuous at inner knots. We also need to mention as the degree of splines is increasing, each polynomial overlaps with more

basis functions in its neighborhood in the domain of the b-spline, but close to boundary of splines less overlap is observed. For example, a first-order b-spline basis function overlaps with two neighbors, while a b-spline basis function of order-2 overlaps with four neighbors. Regarding the aforementioned expressions about b-spline, some general characteristics of a b-spline of order- $m$  are as follows:

- 1) It includes  $m+1$  polynomial pieces where each of them is of degree  $m$ .
- 2) All polynomials are connected at  $m$  inner knots.
- 3) It has continuous derivative up to order  $m-1$  at the joining points.
- 4) Any basis function on the domain spanned by  $m+2$  knots, is positive, otherwise the basis function is zero.
- 5) The basis function overlaps with  $2m$  closest neighbors (polynomials) except at the boundary knots.
- 6) The  $m+1$  b-spline basis functions are non-zero for any given  $x$  on the domain.



**Figure 2.1:** A display of B-spline plots up to degree three on the domain  $(0, 1)$  with equidistant inner knots. The plots clearly express how a b-spline overlaps with polynomial pieces located at its neighborhood for given knots.

### 2.1.2 Smoothing Splines (Penalized Splines)

Smoothing splines a non-parametric method of analysis play an important role in modern research in the area of statistics and data science (Eilers and Marx;1996). They are extensively used for finding flexible methods of estimation of regression models and therefore are applicable to high dimensional data modeling. The other reason for their popularity is that smoothing splines can detect properties in the data which may not be clearly shown by parametric models. Regarding the smoothness of splines, selecting an optimal number of knots with their positions has always been the main issue of many research works, since a lot of knots create over-fitting and small number of knots restricts flexibility and leads to under-fitting. There are several approaches for optimizing the number of knots. One of the methods discussed by Friedman and Silverman gives an automatic scheme for the number of knots and their position but leads to a complicated numerical problem (Friedman and Silverman, 1989). Another idea is to use a great number of knots and a penalty on the second derivative in order to control over-fitting (O'Sullivan, 1986). Before we start with smoothing regression splines, we briefly mention some points about the linear model, least square estimation and relevant definitions.

**Definition:** A linear regression model is an approach for showing the relationship between a dependent variable  $Y$  and explanatory variables  $X$  (which can be more than one). A linear model in matrix form is:  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , where  $\mathbf{Y}$  is the vector of observed values,  $\mathbf{X}$  is the fixed and known values of explanatory variables,  $\boldsymbol{\beta}$  defined as the vector of unknown parameters and  $\boldsymbol{\varepsilon}$  the vector of unobserved errors for  $\mathbf{Y} \in \mathbb{R}^n$ ,  $\mathbf{X} \in \mathbb{R}^{n \times p}$ ,  $\boldsymbol{\beta} \in \mathbb{R}^p$ ,  $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ .

The goal is to find the best estimated value of the vector of parameters  $\beta$  by minimizing the sum of the squared distances between the observed vector  $Y$  and  $X\beta$ . This method is known as Least Square (LS) in which the vector of parameter is estimated by minimizing:

$$Q(\beta) = (Y - X\beta)^T (Y - X\beta) = \sum_{i=1}^n (y_i - x_i^T \beta)^2 \quad (2.1)$$

for given data  $y_i, x_{i1}, x_{i2}, \dots, x_{ip}, i=1, \dots, n$ .

The solution of 2.1 is obtained by setting the derivatives of 2.1 with respect to  $\beta$  equal to

$$\text{zero: } \frac{\partial Q(\beta)}{\partial(\beta)} = \frac{\partial[(Y - X\beta)^T (Y - X\beta)]}{\partial\beta} = -2(Y^T X)^T + 2X^T X\beta = 0 \text{ giving}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \text{ the least square solution.}$$

Now the minimization procedure for estimation is discussed for b-spline by smoothing and regularization. If the function  $f(x)$  is estimated by b-spline function, then the best estimator of  $f(x)$  is the one which minimizes the residual sum of squares where the penalized version of least squares (2.1) for linear models is introduced as follows:

$$\begin{aligned} RSS(\beta, \lambda) &= \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \int [f''(t)]^2 dt \\ &= \sum_{i=1}^n (y_i - \sum_{j=1}^N B_j \beta_j(x_i))^2 + \lambda \int [\sum_{j=1}^N B_j \beta_j''(t)]^2 dt \end{aligned} \quad (2.2)$$

In equation 2.2, the first term captures the sum of the squared distance between observed and fitted values, the second term penalizes curvature and the smoothing parameter ( $\lambda > 0$ ) controls the tradeoff between the two terms. There are two special points regarding the values of  $\lambda$ :

i)  $\lambda=0$  imposes no restriction,  $f(x)$  will interpolate the data if  $N=n$ .

ii)  $\lambda \rightarrow \infty$  results in ordinary linear regression if  $f(x)$  is a cubic spline, as the penalty term results in the solution to the differential equation  $f''(x) = 0$ .

If  $\mathbf{B}$  is the design matrix of basis spline functions and  $\boldsymbol{\beta}$  defined as the vector of parameters with dimension  $N$ , the equation (2.2) in form of matrix will be:

$$RSS(\boldsymbol{\beta}, \lambda) = (Y - B\boldsymbol{\beta})^T (Y - B\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\psi} \boldsymbol{\beta} \quad (2.3)$$

where  $B_{ij} = B_j(x_i)$  and  $\boldsymbol{\psi}_{jr} = \int B_j''(t) B_r''(t) dt$ . The estimator of parameter vector  $\boldsymbol{\beta}$  is the solution to:

$$\frac{\partial RSS(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}} = -2(Y^T B)^T + 2[B^T B + \lambda \boldsymbol{\psi}] \boldsymbol{\beta} = 0 \quad (2.4)$$

$$\text{which is given by: } \hat{\boldsymbol{\beta}} = (B^T B + \lambda \boldsymbol{\psi})^{-1} B^T Y \quad (2.5)$$

Finally the estimated smoothing spline is:

$$\hat{f}(x) = \sum_{j=1}^N \hat{\boldsymbol{\beta}}_j B_j(x) = B \hat{\boldsymbol{\beta}} = B(B^T B + \lambda \boldsymbol{\psi})^{-1} B^T Y = S_\lambda Y \quad (2.6)$$

It is clear that the estimated function is a linear function of observations  $\mathbf{Y}$ . In this relationship  $\mathbf{S}_\lambda$  is the smoothing matrix and the sum of diagonal elements of  $\mathbf{S}_\lambda$  gives the smoothing spline's effective degree of freedom,  $\text{Trace}(\mathbf{S}_\lambda) = df_\lambda$ .

### 2.1.3 P-Splines

In this part a numerical approximation for the integral of the squared higher derivatives of the fitted spline function is discussed. In the method studied by Eilers and Marx (1996), a difference penalty on the coefficients of adjacent b-splines with equally spaced knots is used instead of integration over squared derivatives. This method is known as P-spline.

P-splines don't suffer from boundary effects, keeping the moments of data (mean /

variance) unchanged and have easier numerical computation procedures. For smoothing splines the penalty term requires the derivatives of b-spline function and thus we will have to review the formulas for derivatives of b-splines presented by Eilers and Marx (1996).

For a defined spline in form of  $f(x) = \sum_{i=1}^m \beta_i B_i(x)$ , the first divided difference is

expressed as:

$$h \sum_j \beta_j B'_j(x, m) = \sum_j \beta_j B_j(x, m-1) - \sum_j \beta_{j+1} B_{j+1}(x, m-1) = -\sum_j \Delta \beta_{j+1} B_j(x, m-1) \quad (2.7)$$

In the formula (2.7), h is the distance between knots and  $\Delta \beta_j = \beta_j - \beta_{j-1}$ .

The next relationship for second derivative can be approximated by finite differences:

$$h^2 \sum_j \beta_j B''_j(x; q) = \Delta^2 \sum_j \beta_j B_j(x; q-2) \quad (2.8)$$

where  $\Delta^2 \beta_j = \Delta \Delta \beta_j = \beta_j - 2\beta_{j-1} + \beta_{j-2}$ .

Assume a sequence of data points  $(x_i, y_i)$ ,  $i=1, \dots, N$  and a set of B-splines  $B_j(\cdot)$ ,  $j=1, \dots, n$ .

Now we are aiming to minimize the least square function:

$$Q = \sum_{i=1}^N [y_i - \sum_{j=1}^n \beta_j B_j(x_i)]^2 \quad (2.9)$$

Based on O'Sullivan's approach (1986), a large number of knots and a penalty on the second derivative of fitted curve are considered for (2.9). So Q is represented in the new form of:

$$Q = \sum_{i=1}^N [y_i - \sum_{j=1}^n \beta_j B_j(x_i)]^2 + \lambda \int_{x_{\min}}^{x_{\max}} [\sum_{j=1}^n \beta_j B''_j(x)]^2 dx \quad (2.10)$$

We know the integration is defined on squared second derivative, however it needs to be mentioned that lower or higher orders of derivative are possible.

By applying formula (2.8) and higher order finite differences of the coefficients of adjacent b-splines into (2.10), we obtain the equivalent form of Q as follows:

$$Q = \sum_{i=1}^N [y_i - \sum_{j=1}^n \beta_j B_j(x_i)]^2 + \lambda \sum_{j=r+1}^n (\Delta^2 \beta_j)^2 \quad (2.11)$$

where the first term is the sum of squared residuals (deviance between observed values and the fitted b-splines). The second term called penalty term includes the smoothing parameter  $\lambda$  and sum of squared finite differences of order-2 on the coefficients of b-splines.

#### 2.1.4 Selection of Smoothing Parameter

So far we have interpreted the spline functions and least square equations for an assumed amount of the smoothing parameter  $\lambda$ . Now we are going to discuss the methods of selecting  $\lambda$  which controls the smoothness of the fitted curve.

For smoothing splines, the smoothing parameter is in close relationship with degree of freedom and the number of knots. The relationship between smoother matrix, smoothing parameter and degree of freedom is in linear form of  $\text{Trace}(S_\lambda) = df_\lambda$ . The smoother matrix  $S_\lambda$  is a symmetric and positive semi-definite matrix, so it can be rewritten in an Eigen decomposition form which is a function of  $\lambda$ . For an automatic method of selection, a few values are selected for degree of freedom. For each degree of freedom, a corresponding  $\lambda$  is derived through the aforementioned relationship  $\text{Trace}(S_\lambda) = c$ . Then the optimal amount of  $\lambda$  is obtained based on some criteria like F-test, residual plots, cross validation and so on. One of the criteria used for obtaining the optimal value of  $\lambda$  is AIC for which we need the log-likelihood function (which can be replaced by deviance) and the dimension of vector of estimated parameters.  $AIC = dev(v; \beta, \lambda) + 2 \dim(\beta, \lambda)$

where the dimension of the vector of parameters is the trace of smoother matrix (Hastie and Tibshirani;1990). Another way for getting optimal  $\lambda$  is cross validation (CV),

$$CV(\lambda) = \sum_{i=1}^N \left( \frac{y_i - \hat{y}_i}{1 - s_{ii}} \right)^2 \quad (2.12)$$

or generalized cross validation (GCV) (Wahba;1990),

$$GCV(\lambda) = \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{(N - \sum_{i=1}^N s_{ii})^2} \quad (2.13)$$

where  $s_{ii}$  is the  $i$ th diagonal element of the smoothing matrix. The optimal value of  $\lambda$  is the one which minimizes  $CV(\lambda)$  or  $GCV(\lambda)$ .

## 2.2 Finite Mixture Models

In this thesis a finite mixture of continuous random variables will be utilized. A finite mixture model is a convex combination of individual distributions (components) and nonnegative weights that sum to one. So it is a powerful and flexible method for approximating more complex distributions. Finite mixture models have many applications in areas of statistical modeling, analysis, machine learning, survival analysis, clustering and classification.

**Definition:** Let vector  $\mathbf{X} = (X_1, \dots, X_j, \dots, X_L)$  be a sequence of random variables with observed values  $\mathbf{x} = (x_1, \dots, x_j, \dots, x_L)$ . The probability density function of a mixture model which is the convex combination of  $I$  probability density functions is given as:

$$p(x|\Theta) = \sum_{i=1}^I a_i p_i(x|\theta_i) \quad (2.14)$$

where  $p_i(x | \theta_i)$  is the probability density function of the  $i$ th component, corresponding weight  $a_i$  (the probability of being in the  $i$ th component) with constraint  $\sum_{i=1}^I a_i = 1, a_i \geq 0$  and  $\Theta = (a_1, \dots, a_I, \theta_1, \dots, \theta_I)$  is the vector of parameters.

A simple interpretation of finite mixture model is to say that the random variable  $X$  is generated from  $I$  distinct random processes where each of the processes has a probability density function  $p_k$  and proportion of observations  $a_k$  from that associated process.

### 2.3 Recurrent Event Data

Recurrent event data is a popular type of longitudinal data for which subjects/individuals are followed over time and the number of observed events per subject is recorded. There has been enormous growth in application and analysis of recurrent event data by researchers in many fields, particularly medicine and clinical studies (Dean and Nielsen, 2007). For recurrent event data studies, the overall period of observation for  $i$ th subject will be denoted by  $(T_{i0}, T_{ie_i})$  and  $T_{i0} < T_{i1} < \dots < T_{ie_i}$  where  $T_{ij} \in (T_{i0}, T_{ie_i}]$ ,  $j=1, 2, \dots, e_i$  show the sequence of follow-up times at which the occurrence of an event is investigated. In this thesis  $M$  will denote the number of objects in the study. Recurrent event data generally come in two forms:

1) **Event time data:** This kind of data represents the exact occurrence times of the event under investigation for all  $M$  subjects in the study. The set of recorded event times for  $i$ th individual is  $\{w_{ijl}\}$  where  $i=1, 2, \dots, M$ ,  $j=1, 2, \dots, e_i$ ,  $l=1, 2, \dots, N_{ij}$ , and  $\{w_{ijl}\}$  represents the time of the  $l$ th event in the interval  $(T_{i(j-1)}, T_{ij}]$  for the  $i$ th individual/object.

2) **Panel data:** Panel data are the number of events  $N_{ij}$  in a time interval  $(T_{i(j-1)}, T_{ij}]$ ,  $j=1,2,\dots,e_i$  for each individual  $i$ ,  $i=1,\dots, M$ .

Note:  $N_{ij}$  is the number of occurrences for the  $i$ th individual in the  $j$ th time interval and

the total number of events per individual is  $N_i = \sum_{j=1}^{e_i} N_{ij}$ .

The start time of the study and observation is denoted by  $T_{i0}$  which is not necessarily the same for all subjects since the entry time into the study may change based on different factors. In this research work all start times are assumed to be zero ( $T_{i0}=0$ ) entry into the study, but the followup times  $T_{ij}$  can be different for individuals. Recording the time of event for each individual is not always possible, as following all objects continuously over a specific range of time may be expensive and/or intrusive. For this reason panel data studies are a more popular design used by many researchers and organizations (Dean and Nielsen, 2007).

### 2.3.1 Heterogeneity

Heterogeneity is a concept often applied in statistical analysis where the statistical properties of a dataset or several datasets are described. Heterogeneity is usually the source of overdispersion in data analysis of counts and it is clear after the analysis of the results. Overdispersion results when the variance of the data is bigger than the expected value and is a common occurrence when modeling count data (McCullagh and Nelder, 1989). In statistical sciences, the task is to find a suitable parametric model for fitting the observed data. In the selection procedure for a model, the underdispersion or overdispersion of the data should be taken into account since a number of models can not accommodate the high rate of variability of data.

In econometric subjects one of the familiar models for count data is the Poisson model which has the restriction that the mean should be equal to the variance, but in many studies the variance of the observed data exceeds the mean (overdispersion). The phenomenon of overdispersion is a common issue in count data analysis (Dean, 1998) and it can lead to unreliable inference or underestimation of the standard error. Suppose a sequence of random variables  $N_i, i=1,2,\dots,M$  are independent and identically Poisson distributed with parameter  $\mu_i$ :

$$N_i \stackrel{iid}{\sim} \text{Poisson}(\mu_i) \text{ where } \mu_i = \exp(X_i^T \beta).$$

As we mentioned for overdispersed data in a Poisson process, variance is greater than mean  $V(N_i) > E(N_i)$ . To take overdispersion into account, a specific random effect for each individual is considered in the Poisson model:

$$\forall i: N_i | v_i \stackrel{iid}{\sim} \text{Poisson}(v_i \mu_i) \text{ where } v_i \text{ is a set of specific random effects for all}$$

individuals following the same distribution,  $v_i \stackrel{iid}{\sim} f_v, i = 1,2,\dots,M$ .

For example if specific random effects are following gamma distribution, then the unconditional model  $N_i$  will be a mixed effect model called the Negative binomial.

More details about mixed effect models will be presented in chapter 3.

## 2.4 Expectation-Maximization Algorithm

In the statistical sciences, the Expectation-Maximization algorithm is an efficient iterative procedure for finding maximum likelihood estimation of parameters in the presence of missing data. In the EM algorithm, each iteration includes two steps: the (Expectation) E-step and the (Maximization) M-step. The E-step constructs conditional expectation of log

likelihood for estimating and updating the missing data given the observed data and the current estimates of the model parameters. In the M-step, parameters are computed by maximizing the complete log likelihood given the current prediction of the missing data. The EM algorithm is a very useful method for estimation of parameters in mixture of models and compound distributions (McLachlan and Krishnan, 2007).

Let  $\mathbf{Y}$  denote the vector of observed values,  $\boldsymbol{\theta}$  is a set of parameters and  $\mathbf{Z}$ , the vector of missing data. Now the general approach of EM algorithm is briefly explained in four steps:

1) The conditional probability distribution  $p(w|y, \theta^m)$  for the complete data set

$\mathbf{W}=(\mathbf{Y},\mathbf{Z})$  is derived based on observed values ( $\mathbf{Y}$ ) and current value of parameters  $\boldsymbol{\theta}^m$ .

2) By using the conditional distribution function found in step 1, the conditional expected log-likelihood of the complete data is obtained as follows:

$$Q(\theta| \theta^m) = \int_{\chi(y)} \log p(w| \theta) f(w| y, \theta^m) dy = E_{w|y, \theta^m} [\log p(W| \theta)]$$

(2.15)

where the integral is calculated over the set of observed values  $\chi(y)$  and the function Q depends only on the current estimate of  $\boldsymbol{\theta}$  ( $\boldsymbol{\theta}^m$ ) and observed values ( $\mathbf{y}$ ).

3) In this step we compute the new value of  $\boldsymbol{\theta}$  which maximizes the Q function derived in step 2. The new estimation is called  $\boldsymbol{\theta}^{m+1}$ .

4) A standard stopping criteria to indicate convergence to the mle is given by

$$\|\boldsymbol{\theta}^{m+1} - \boldsymbol{\theta}^m\| < \delta \text{ for some } \delta > 0. \text{ Since the parameter values are maximizing the log-}$$

likelihood function  $l(\theta) = \log p(y| \theta)$ , another stopping rule that can be used is:

$$|l(\boldsymbol{\theta}^{m+1}) - l(\boldsymbol{\theta}^m)| < \delta \text{ for some } \delta > 0. \text{ If the criteria doesn't hold, } \boldsymbol{\theta}^m \text{ is replaced by } \boldsymbol{\theta}^{m+1}$$

$(\boldsymbol{\theta}^m = \boldsymbol{\theta}^{m+1})$  and the algorithm goes back to step 1. The iterations will continue until the standard criteria holds and we get the maximum likelihood estimates of the parameters. The EM algorithm is presented in four steps where a combination of steps 1 and 2 form the E-step and step 3 gives M-step. We are going to clarify this algorithm by using an example in which both E-step and M-steps can be described in closed form, however in many cases it may not be possible to find closed form for the steps and we need to follow other procedures such as nested EM algorithm.

**Example:** Suppose a set of independent and identically distributed random variables of size  $n$ ,  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$  can be modeled as a mixture of two normal distributions,

$Y_i \stackrel{iid}{\sim} F_Y; i = 1, \dots, n$  where,

$$f_Y(y | \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p) = p \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(y - \mu_1)^2}{2\sigma_1^2}\right\} + (1 - p) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(y - \mu_2)^2}{2\sigma_2^2}\right\} \quad (2.16)$$

We consider a random variable  $Z_i, i=1, \dots, n$  defined as an indicator function for clustering the observed data into two groups,  $Z_i=1$  if the  $i$ th observation belongs to the first cluster, otherwise it will be zero and the data belongs to the second cluster ( $Z_i=0$ ). So it is assumed that  $Z_i; i=1, \dots, n$  is a sequence of independent and Bernoulli distributed random variables of size  $n$  with probability of success  $p$  in each trial,

$$Z_i \stackrel{iid}{\sim} f_{Z_i}(z_i, p) = p^{z_i} (1 - p)^{1-z_i}; i = 1, \dots, n$$

The indicator variable is grouping the observations into two clusters where the data are following normal distribution in the first cluster with parameters  $\mu_1$  and  $\sigma_1^2$ , in the 2nd cluster with parameters  $\mu_2$  and  $\sigma_2^2$ .

If  $Z_i=1$  then  $f_{Y_i|Z_i=z_i}(y_i) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(y_i - \mu_1)^2}{2\sigma_1^2}\right\}$  and

If  $Z_i=0$  then  $f_{Y_i|Z_i=z_i}(y_i) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(y_i - \mu_2)^2}{2\sigma_2^2}\right\}$

Our aim is to derive the maximum likelihood estimation of the parameters

$\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p)^T$ . At first the complete log-likelihood function is found through

the likelihood function given clustered data. The complete likelihood function is given

by:

$$L_{comp}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \prod_{i=1}^n [f_{Y_i|Z_i=z_i}(y_i)]^{z_i} [f_{Y_i|Z_i=1-z_i}(y_i)]^{1-z_i} \quad (2.17)$$

and the complete log-likelihood function by

$$\begin{aligned} l_{comp}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) &= \ln(L_{comp}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)) = \sum_{i=1}^n [z_i \ln[f_{Y_i|Z_i=z_i}(y_i)] + (1-z_i) \ln[f_{Y_i|Z_i=1-z_i}(y_i)]] \\ &= \sum_{i=1}^n z_i \left[ -\ln(\sqrt{2\pi}) - \frac{1}{2} \ln(\sigma_1^2) - \frac{(y_i - \mu_1)^2}{2\sigma_1^2} \right] + (1-z_i) \left[ -\ln(\sqrt{2\pi}) - \frac{1}{2} \ln(\sigma_2^2) - \frac{(y_i - \mu_2)^2}{2\sigma_2^2} \right] \end{aligned} \quad (2.18)$$

The mle of the parameters  $\mu_1, \mu_2, \sigma_1^2$  and  $\sigma_2^2$  is computed via partial derivatives of the

log-likelihood with respect to each element of the vector  $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ . So ML

estimators are derived by setting each of the following equations to zero:

$$\frac{\partial}{\partial \mu_1} l_{comp}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = 0 \Rightarrow \hat{\mu}_1 = \frac{\sum_{i=1}^n z_i y_i}{\sum_{i=1}^n z_i} \quad (2.19)$$

$$\frac{\partial}{\partial \mu_2} l_{comp}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = 0 \Rightarrow \hat{\mu}_2 = \frac{\sum_{i=1}^n (1-z_i) y_i}{\sum_{i=1}^n (1-z_i)} \quad (2.20)$$

$$\frac{\partial}{\partial \sigma_1^2} l_{comp}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = 0 \Rightarrow \hat{\sigma}_1^2 = \frac{\sum_{i=1}^n z_i (y_i - \mu_1)^2}{\sum_{i=1}^n z_i} \quad (2.21)$$

$$\frac{\partial}{\partial \sigma_2^2} l_{comp}(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = 0 \Rightarrow \hat{\sigma}_2^2 = \frac{\sum_{i=1}^n (1 - z_i) (y_i - \mu_2)^2}{\sum_{i=1}^n (1 - z_i)} \quad (2.22)$$

The maximum likelihood estimation of the parameter  $p$  is calculated based on the predicted values of the random variable  $Z_i$ ;  $i=1, \dots, n$ .

$$Z_i \stackrel{iid}{\sim} f_{Z_i}(z_i, p) = p^{z_i} (1 - p)^{1 - z_i}; i = 1, \dots, n$$

$$\text{The likelihood function is: } L(p, z) = \prod_{i=1}^n p^{z_i} (1 - p)^{1 - z_i}$$

$$\text{The log-likelihood function is: } l(p, z) = \ln L(p, z) = \sum_{i=1}^n z_i \ln(p) + (1 - z_i) \ln(1 - p)$$

If  $Z_i$ 's are known, then  $p$  is estimated as follows:

$$\frac{\partial}{\partial p} l_z = 0 \Rightarrow \hat{p} = \frac{\sum_{i=1}^n z_i}{n} \quad (2.23)$$

However if the set of indicator random variables  $Z_i$ 's are unobserved data points and we don't know the membership of observed data ( $y_i$ ), then the missing data are estimated by using Bayes theorem where the formulas form the E-step of the EM algorithm:

$E[i \text{ in } 1^{\text{st}} \text{ cluster} \mid Y] = 1 \cdot P(Z_i=1 \mid Y) + 0 \cdot P(Z_i=0 \mid Y) = (1)P(Z_i=1 \mid Y)$

$$= \frac{f_{y_i|z_i=1}(y_i) \cdot p(z_i=1)}{f_Y(y_i)} = \frac{pf_{y_i|z_i=1}(y_i)}{pf_{y_i|z_i=1}(y_i) + (1-p)f_{y_i|z_i=0}(y_i)}$$

$$= \frac{p \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(y_i - \mu_1)^2}{2\sigma_1^2}\right\}}{p \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(y_i - \mu_1)^2}{2\sigma_1^2}\right\} + (1-p) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(y_i - \mu_2)^2}{2\sigma_2^2}\right\}}$$

So, the predicted  $Z_i$  in the E-step of algorithm is:

$$z_i = \frac{p \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(y_i - \mu_1)^2}{2\sigma_1^2}\right\}}{p \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(y_i - \mu_1)^2}{2\sigma_1^2}\right\} + (1-p) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(y_i - \mu_2)^2}{2\sigma_2^2}\right\}}$$

Finally the EM algorithm is ready to be performed based on the derived formulas as follows:

1) The ordered observed data are clustered to two groups where the size of each group depends on initial value of  $p$  in the algorithm. Based on initial value of  $p=.5$  in this research work, the ordered observed data  $y_{(1)}, y_{(2)}, \dots, y_{(n)}$  are clustered to two groups of equal sizes. Consider  $z_i=1$  for first half of ordered data  $y_1, \dots, y_{\lfloor \frac{n}{2} \rfloor}$  and  $z_i=0$  for second half of ordered data  $y_{\lfloor \frac{n}{2} \rfloor + 1}, \dots, y_n$ . If  $p=.25$ , then 25 percent of data in the first group and 75 percent in the second group.

2) M-step: The equations 2.19, 2.20, 2.21, 2.22 and 2.23 are used to compute maximum likelihood estimations of parameters  $(\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{p})$ .

3) E-step:  $z_i$ 's are updated as follows.

$$\hat{z}_i = \frac{\hat{p} \frac{1}{\sqrt{2\pi\hat{\sigma}_1}} \exp\left\{-\frac{(y_i - \hat{\mu}_1)^2}{2\hat{\sigma}_1^2}\right\}}{\hat{p} \frac{1}{\sqrt{2\pi\hat{\sigma}_1}} \exp\left\{-\frac{(y_i - \hat{\mu}_1)^2}{2\hat{\sigma}_1^2}\right\} + (1 - \hat{p}) \frac{1}{\sqrt{2\pi\hat{\sigma}_2}} \exp\left\{-\frac{(y_i - \hat{\mu}_2)^2}{2\hat{\sigma}_2^2}\right\}}$$

4) The algorithm is repeated until:

$$\|\theta^{j+1} - \theta^j\| < \delta \quad \text{or} \quad |l(\theta^{j+1}) - l(\theta^j)|$$

## Chapter 3

### Methodology

#### 3.1 Poisson Process

Suppose a set of nonnegative random variables  $\{X_1, X_2, \dots\}$  are independently and identically distributed, then the counting process  $\{N(t); t \geq 0\}$  which can be defined as  $N(t) = \max \{n : S_n \leq t\}$  with  $\sum_{j=1}^n X_j = S_n$  is known as a renewal process. Renewal processes are extensively used for modeling occurrences of events happening randomly over time where the waiting times between events are assumed to be independent and identically distributed (Winkelmann, 1995). These temporal events are called renewals or arrivals. Such an assumption makes the renewal process a flexible and powerful tool. Its most important role is in the structure of complex stochastic models where more than one renewal process is used. For example the arrivals can be customers arriving at an insurance company, natural events like tornados or earthquakes, emission of particles from radioactive sources and so on.

A special case of the renewal process arises if the non-negative inter-arrival (waiting) times between consequent events are exponentially distributed and is called a Poisson process. In this case the number of events (renewals) in a specified range of time is Poisson distributed. Working with the exponential distribution is relatively easier than other distribution families because of its simpler form of probability density function (single parameter distribution), moment generating function, mean and variance. Due to this simple form of exponential density function, estimation of parameter and convolution of independent exponential random variable is more straightforward than the other

distributions (Ross, 2010). The most important property of exponential density function is that it is memoryless which is defined as follows:

$$P(X > s+t | X > t) = \frac{P(X > s+t, X > t)}{P(X > t)} = P(X > s)$$

or,

$$P(X > s+t) = P(X > s)P(X > t)$$

If X is the lifetime of an object, then the memoryless property represents the probability of the object surviving for at least t+s years on condition that it has lived t years is equal to the probability that it survives for s years (Ross,2010).

**Definition:** The stochastic process  $\{N(t) | t \geq 0\}$  is referred to as a Poisson process with rate  $\lambda$  when the sequence of waiting times  $\{T_1, T_2, \dots\}$  are independently identically distributed with exponential distribution with the parameter  $\lambda$  ( $E(T_i) = \frac{1}{\lambda}$ ).

Note that  $T_k$  represents the time between the (k-1)th and kth events. The probability distribution function of  $N(t)$  can be derived in theory as follows:

$$N(T) < n \Leftrightarrow S_n \geq T \quad (3.1)$$

then,

$$P(N(T) < n) = P(S_n \geq T) = 1 - F_n(T) \quad (3.2)$$

where  $F_n(T)$  is the cumulative distribution function of the arrival time of the nth event

that is computed by  $S_n = \sum_{j=1}^n T_j$  in renewal theory. The following equation shows a very

important relationship between count model and inter-arrival time distribution:

$$P(N_T = n) = P(N_T < n+1) - P(N_T < n) = P(S_{n+1} \geq T) - P(S_n \geq T) = F_n(T) - F_{n+1}(T) \quad (3.3)$$

Equation 3.3 illustrates that the probability density function of the number of events in the time interval  $(0, T)$  can be derived based on the cumulative distribution function of  $S_n$  shown in formulas 3.1 and 3.2.

### 3.1.1 Non-homogeneous Poisson Process

In the previous section we briefly discussed the concept of the Poisson process which is characterized by a constant rate of renewals  $\lambda$ . Some physical processes fail to be modeled by a Poisson process because the assumption of stationary doesn't hold. For example, if the purpose is modeling the arrival times of cars at a specific street, the mean arrival rate of cars may not be the same during different hours of a day which implies an ordinary Poisson process cannot be used. It is of great importance to consider a more general form of the process with non-constant arrival rate. A non-homogeneous Poisson process has the properties of an ordinary Poisson process, except for the fact that the average arrival rate is a function of time,  $\lambda(t)$ . As usual, let  $N(t, t+h)$  denote the number of random points (events) in the time interval  $(t, t+h]$  for  $t \geq 0$  and  $h \geq 0$ .

**Definition and basic properties:** Assume  $N(t) = N(0, t)$ , then the counting process  $\{N(t), t \geq 0\}$  is known as a non-homogeneous Poisson process with intensity rate  $\lambda(t) \geq 0$  if it satisfies in following conditions:

- 1)  $N(0) = 0$
- 2)  $\{N(t), t \geq 0\}$  has independent increments
- 3)  $P(N(t, t+h) = 0) = 1 - \lambda(t)h + o(h)$
- 4)  $P(N(t, t+h) = 1) = \lambda(t)h + o(h)$
- 5)  $P(N(t, t+h) \geq 2) = o(h)$

where  $o(h)$  represents any function of  $h$ ,  $g(h)$  which is going toward zero faster than  $h$ ,

$\lim_{h \rightarrow 0} \frac{g(h)}{h} = 0$ . Conditions (3) to (5) highlight the memoryless property of the process

which says that the occurrence of an event in interval  $(t, t+h]$  doesn't impact the occurrence of other events in another disjoint time interval (prior to/after time  $t$ ).

As we mentioned beforehand, for a non-homogeneous Poisson process with non-negative rate function  $\lambda(t)$ , the number of events in a time interval follows a Poisson distribution.

The parameter of the process which depends on the time interval is written as follows:

$$\Lambda(t) = \int_0^t \lambda(u) du \quad (3.4)$$

where  $\Lambda(t)$  is a continuous function of  $t$  and is called the cumulative intensity function.

By using conditions 1-5, it can be shown (Ross, 1983) that the process  $N(t, t+h)$  has a Poisson distribution with probability mass function defined as:

$$P\{N(t, t+h) = n\} = \frac{\left( \int_t^{t+h} \lambda(u) du \right)^n \exp \left\{ - \int_t^{t+h} \lambda(u) du \right\}}{n!}, \quad n=0, 1, 2, \dots \quad (3.5)$$

So the random variable  $N(t)$  follows Poisson distribution with the parameter  $\int_t^{t+h} \lambda(u) du$ .

The likelihood function based on  $M$  observable event times is built under the assumptions that waiting times between events is a sequence of independent and identically distributed random variables and by starting the process at time  $t_i$ , the next event will be occurring in  $(t_{i+1}, t_{i+1}+\Delta t]$  with the probability of:

$$\begin{aligned} P\{N(t_i, t_{i+1}) = 0 \text{ and } N(t_{i+1}, t_{i+1} + \Delta t) = 1\} &= P\{N(t_i, t_{i+1}) = 0\} P\{N(t_{i+1}, t_{i+1} + \Delta t) = 1\} \\ &= \exp \left\{ - \int_{t_i}^{t_{i+1}} \lambda(u) du \right\} [\lambda(t_{i+1}) \Delta t + o(\Delta t)] = \lambda(t_{i+1}) \exp \left\{ - \int_{t_i}^{t_{i+1}} \lambda(u) du \right\} \Delta t + o(\Delta t) \end{aligned} \quad (3.6)$$

Now by consideration of an observation period over fixed time interval  $(0, T_e]$  and follow up times  $0 = t_0 \leq t_1 < t_2 < \dots < t_M \leq T_e$  for  $M$  events, the likelihood function is given by :

$$L = \left( \prod_{i=1}^N \lambda(t_i) \exp \left\{ - \int_{t_{i-1}}^{t_i} \lambda(u) du \right\} \right) \exp \left\{ - \int_{T_N}^{T_e} \lambda(u) du \right\} = \left\{ \prod_{i=1}^N \lambda(t_i) \right\} \exp \{ - \Lambda(T_e) \} \quad (3.7)$$

The joint conditional probability density function of event times  $T_1, T_2, \dots, T_n$  given the number of randomly occurring events in interval  $(0, T_e]$  ( $N(T_e) = n$ ) is given by:

$$f_{T|N}(t_1, t_2, \dots, t_n | N(T_e) = n) = n! \prod_{i=1}^n \frac{\lambda(t_i)}{\Lambda(T_e)} \quad (3.8)$$

The equation 3.8 can be interpreted as the distribution of a sequence of order statistics of size  $n$  with density function  $\frac{\lambda(t)}{\Lambda(T_e)} I(0 < t < T_e)$  where  $I(B)$  is defined as indicator

function of the event  $B$ .

### 3.1.2 Proportional Intensity Assumption

This chapter focuses on proportional intensity models similar to proportional hazard models discussed as a class of survival models in statistics (Lawless, 1982). A proportional hazard model consists of two parts: the baseline hazard function ( $\lambda_0(t)$ ), describing the risk of an individual at baseline level of covariates ( $x=0$ ) and the effect parameters  $g(x)$ , showing a proportional increase or decrease in risk in response to  $x$ .

Now for our purpose in this research work, the hazard function is replaced by intensity function of a count process to deal with recurring event data. The proportional intensity models is in fact an extension of the Cox model into a stochastic process in which each individual is assigned a different intensity that is proportional to other individual's intensities. In other words, the ratio of intensities for two individuals with corresponding

covariate vectors  $X_1$  and  $X_2$  is  $\lambda(t | x_1) / \lambda(t | x_2)$  which remains constant in time. In a general form, the intensity function is given by:

$$\lambda(t | x) = \lambda_0(t)g(x) \quad (3.9)$$

where  $\lambda_0(t)$  is the baseline intensity function and  $g(x)$  can be any 1-1 function of the covariates, however the most common form of  $g(x)$  in survival analysis, analysis of recurrent event data and Poisson process is  $g(x) = e^{x'\beta}$  (Lawless,1982).

### 3.1.3 Models for Heterogeneity

The concept of overdispersion was briefly explained in chapter 2. One of the popular models for count data is Poisson model in which the restriction of the equality between the mean and the variance doesn't allow us to use it for modeling overdispersed data. In order to take overdispersion into account in the model, a specific random effect for each object is considered in the model where the discussion of mixed effect model appears in statistical analysis. Regarding proportional intensity models, one suggested model for accommodating the excess variance is to consider a unique intensity for each individual,  $\lambda_i(t) = \nu_i \lambda(t)$ . As can be seen, each individual has a specific random effect  $\nu_i$  with probability density function  $f(\nu)$  depending on parameter  $\tau$ .

It will be assumed that the expectation of individual random effect is  $E(\nu_i) = 1$  since either the  $x_i$  or  $\lambda_0(t)$  includes an intercept term. Then under multiplicative proportional intensity assumption,  $\{N_i(t) | \nu_i, t \geq 0\}$  is a count process following the Poisson distribution with intensity function  $\lambda_i(t | \nu_i, x_i) = \nu_i \lambda_0(t) e^{x_i'\beta}$ . The likelihood function for each subject is built by application of 3.7, so the likelihood of the  $i$ th subject whose  $N_i$  observations

happen at event times  $t_{ij}$  where  $0 \leq t_{i1} < t_{i2} < \dots < t_{iN_i} \leq T_{ie_i}$  and  $L_i = \int_0^\infty L_{\{N_i(t) | v_i\}} f(v_i) dv_i$

where  $L_{\{N_i(t) | v_i\}}$  is the likelihood of the Poisson process with intensity

$\lambda_i(t | v_i, x_i) = v_i \lambda_0(t) e^{x_i \beta}$  for given  $v_i$ . Finally, the likelihood function based on event times is rewritten as:

$$\begin{aligned} L_i &= \int_0^\infty \left[ \prod_{j=1}^{N_i} \lambda(t_{ij} | v_i, x_i) \right] \exp \{-\Lambda(T_{ie_i} | v_i, x_i)\} f(v_i) dv_i \\ &= \int_0^\infty \left[ \prod_{j=1}^{N_i} v_i \lambda_0(t_{ij}) e^{x_i \beta} \right] \exp \{-v_i \Lambda_0(T_{ie_i}) e^{x_i \beta}\} f(v_i) dv_i \\ &= \prod_{j=1}^{N_i} \left( \frac{\lambda_0(t_{ij})}{\Lambda_0(T_{ie_i})} \right) \int_0^\infty \left[ v_i \Lambda_0(T_{ie_i}) e^{x_i \beta} \right]^{N_i} \exp \{-v_i \Lambda_0(T_{ie_i}) e^{x_i \beta}\} f(v_i) dv_i \end{aligned} \quad (3.10)$$

Equation 3.10 clearly represents that  $N_i | v_i$  has a Poisson distribution with mean  $v_i \mu_i$

where  $\mu_i = \Lambda_0(T_{ie_i}) e^{x_i \beta}$ , so the marginal distribution of  $N_i$  is determined by the

distribution of  $v_i$ . It is noticeable that in the new conditional probability model each

individual has the same random effect which induces a correlation between the individual

panel counts. One of the common options for distribution of  $v_i$  is gamma with

probability density function,

$$p(v_i) = \frac{1}{\Gamma\left(\frac{1}{\tau}\right) \tau^{\frac{1}{\tau}}} v_i^{\frac{1}{\tau}-1} \exp\left\{-\frac{v_i}{\tau}\right\} I(v_i > 0, \tau > 0) \quad (3.11)$$

It can be shown that  $N_i$  is negative binomially distributed with mean  $\mu_i$  and variance

$\mu_i(1 + \tau \mu_i)$  (Lawless, 1987a). In other words, negative binomial is an extended version of

Poisson distribution (with the overdispersion parameter  $\tau$ ) which handles the assumption

of overdispersion of the data.

### 3.2 Semiparametric Analysis for Recurrent Events

In this section we aim to interpret the semiparametric analysis of recurrent events. The semiparametric model and its application for modeling a nonhomogeneous Poisson process with a multiplicative proportional intensity assumption including covariate affects are described. B-splines will be used to model the baseline intensity function  $\lambda_0(t)$  plus individual specific random effects  $\nu_i$  will be used to handle heterogeneity. In the following part, the estimation procedure of model parameters and the structure of the full likelihood function will be discussed.

#### 3.2.1 Semiparametric Nonhomogeneous Poisson Process Model

Consider a group of  $M$  independent random processes  $N_1(t), N_2(t), \dots, N_M(t)$  where  $N_i(t)$  denotes the number of randomly occurred events for the  $i$ th subject in time interval  $(0, t]$ . Suppose for each subject, the intensity function  $\lambda_i(t)$  has multiplicative and proportional intensity assumption of the form  $\lambda_i(t) = \nu_i \lambda_0(t) e^{\mathbf{x}_i \beta}$  where  $\lambda_0(t)$  is the baseline intensity,  $\beta$  is a vector of regression parameters,  $\mathbf{x}_i$  is a vector of covariates and  $\nu_i$  a specific random effect for the  $i$ th subject respectively. A probability density function  $f(\nu_i; \tau)$  is considered for the specific random effects  $\nu_i > 0$ . Now we continue the discussion under the assumptions that  $E(\nu_i) = 1$  and variance  $V(\nu_i) = \tau$  to handle heterogeneity. Before constructing the likelihood function for all  $M$  individuals some notations should be recalled. Subject  $i$  is followed during time period  $(0, T_{ie_i}]$  with followup times  $0 = T_{i0} < T_{i1} < \dots < T_{ie_i}$ ,  $N_i$  is the number of events experienced by  $i$ th individual at times

$\{w_{ijl}\}$ ,  $j=1,2,\dots,e_i$ ,  $l=1, 2, \dots, N_{ij}$  where  $N_i = \sum_{j=1}^{e_i} N_{ij}$  and  $N_{ij}$  denotes the number of events

in the  $j$ th followup period  $[T_{i(j-1)}, T_{ij}]$ . Then the likelihood function is:

$$\begin{aligned} L_e &= \prod_{i=1}^M \int_0^\infty L_{\{N_i(t)|v_i\}} f(v_i; \tau) dv_i \\ &= \prod_{i=1}^M \left( \prod_{j=1}^{e_i} \prod_{l=1}^{N_{ij}} \frac{\lambda_0(w_{ijl})}{\Lambda_0(T_{ie_i})} \right) \int_0^\infty [v_i \mu_i]^{N_i} \exp \{-v_i \mu_i\} f(v_i; \tau) dv_i \end{aligned} \quad (3.12)$$

where  $\mu_i = E(N_i) = \Lambda_0(T_{ie_i}) e^{x_i^T \beta}$ .

The likelihood for panel data  $\{N_{ij} = N_i(T_{i(j-1)}, T_{ij})\}$ ,  $j=1,\dots,e_i$  is constructed with some changes in 3.12 by calculating the probability of  $N_i$  events for which  $N_{i1}$  events happening in first followup period  $(0, T_{i1}]$ ,  $N_{i2}$  happening in second period  $(T_{i1}, T_{i2}]$  and so on. Then the likelihood for recurrent event panel data is:

$$L_p = \prod_{i=1}^M \left\{ \binom{N_i}{N_{i1}, \dots, N_{ie_i}} \prod_{j=1}^{e_i} \left( \frac{\mu_{ij}}{\mu_i} \right)^{N_{ij}} \right\} \int_0^\infty (v_i \mu_i)^{N_i} \exp \{-v_i \mu_i\} f(v_i; \tau) dv_i \quad (3.13)$$

where  $\mu_{ij} = E(N_{ij}) = [\Lambda_0(T_{ij}) - \Lambda_0(T_{i(j-1)})] e^{x_i^T \beta}$  and  $\mu_i = \sum_{j=1}^{e_i} \mu_{ij} = \Lambda_0(T_{ie_i}) e^{x_i^T \beta}$ .

As can be seen, the likelihood functions based on event times 3.12 and panel counts 3.13 can be classified into two components (Lawless; 1978a). The likelihoods in a more clear

form are written as  $L = L_\psi L_{MP}$ .  $L_\psi$  for event times data is:

$$L_\psi = \prod_{i=1}^M \prod_{j=1}^{e_i} \prod_{l=1}^{N_{ij}} \frac{\lambda_0(w_{ijl})}{\Lambda_0(T_{ie_i})} \quad (3.14)$$

and for panel data,

$$L_\psi = \prod_{i=1}^M \left\{ \binom{N_i}{N_{i1}, \dots, N_{ie_i}} \prod_{j=1}^{e_i} \left( \frac{\Lambda_0(T_{ij}) - \Lambda_0(T_{i(j-1)})}{\Lambda_0(T_{ie_i})} \right)^{N_{ij}} \right\} \quad (3.15)$$

We can define  $L_{\psi}$  as the probability kernel for conditional probability density of the event times  $\{w_{ijl}\}$  or panel data  $\{N_{ij}\}$  on condition that  $N_i$  events have occurred.

For both event times and panel counts,  $L_{MP}$  is given by:

$$\begin{aligned} L_{MP} &= \prod_{i=1}^M \int_0^{\infty} [v_i \mu_i]^{N_i} \exp\{-v_i \mu_i\} f(v_i; \tau) dv_i \\ &= \prod_{i=1}^M \int_0^{\infty} [v_i \Lambda_0(T_{ie_i}) e^{x_i^T \beta}]^{N_i} \exp\{-v_i \Lambda_0(T_{ie_i}) e^{x_i^T \beta}\} f(v_i; \tau) dv_i \end{aligned} \quad (3.16)$$

Also  $L_{MP}$  is described as a mixed Poisson kernel for the probability of  $N_i$  occurred events in followup interval  $(0, T_{ie_i}]$ ,  $i=1, \dots, M$ .

As we said, the likelihood function depends on the probability distribution of individual specific random effect  $v_i$ . Several options can be selected for this distribution, for example Inverse-Gaussian (Dean, Lawless and Willmot, 1989), gamma (Lawless, 1987a) and log-Normal. Gamma is a popular distribution since gamma distribution for random effect in Poisson model gives Negative binomial and is conjugate. So  $N_i$  is unconditionally follows negative binomial with mean  $\mu_i$  and variance  $\mu_i(1 + \tau\mu_i)$  when the random effect  $v_i$  is gamma distributed with mean 1 and variance  $\tau$ . In other words, the resulting mixed Poisson distribution is negative binomial with overdispersion magnitude of  $\tau\mu_i^2$  under the gamma assumption which has the probability density function,

$$L_{MP} = L_{NB} = \prod_{i=1}^M \frac{\Gamma(N_i + \tau^{-1})}{(N_i!) \Gamma(\tau^{-1})} \left( \frac{\tau\mu_i}{1 + \tau\mu_i} \right)^{N_i} \left( \frac{1}{1 + \tau\mu_i} \right)^{\tau^{-1}} \quad (3.17)$$

Finally, the likelihood of the model is defined as  $L = L_{\psi} L_{MP}$  where  $L_{\psi}$  and  $L_{NB}$  were defined in 3.14, 3.15 and 3.17.

It was mentioned that there are several choices for specific random effect distribution including Inverse-Gaussian, log-Normal and gamma distributions. Gamma distribution for random effect in the Poisson model gives Negative binomial which relaxes overdispersion assumption of the data in the model. If a gamma distribution  $(\frac{1}{\tau}, \tau)$  assigns to random effect distribution, the random effect belongs to one parameter distribution families which may not handle some shapes of data rightly. Instead, a mixture of gamma distributions is discussed for specific random effect. A mixture of gamma distributions with k, k=2, 3, 4, ... components for random effect makes the model more complicated. For simplicity, a mixture of two gamma distributions (a three parameter family) is used for specific random effect in this research thesis which can handle different shapes of data including right-tailed, left-tailed and bimodal data sets. A mixture of gamma distributions is a conjugate prior to mixture of Negative binomial family which can fit a number of other families like gamma and Inverse-Gaussian. Suppose the specific random effect is represented as a mixture of two gamma

distributions with parameters  $(\frac{1}{\tau_1}, \tau_1)$  and  $(\frac{1}{\tau_2}, \tau_2)$  :

$$p(v_i) = pf(v_i; \tau_1) + (1 - p)f(v_i; \tau_2) \quad (3.18)$$

where p is the probability parameter of the Bernoulli random variable  $Z_i$  showing the membership of subject specific random effects. The random effect follows gamma

distribution  $(\frac{1}{\tau_1}, \tau_1)$  with probability p, otherwise it has gamma distribution  $(\frac{1}{\tau_2}, \tau_2)$  with corresponding probability (1-p). The probability density functions stated in 3.18 are defined as,

$$f(v_i; \tau_1) = \frac{1}{\Gamma\left(\frac{1}{\tau_1}\right) \tau_1^{\frac{1}{\tau_1}}} v_i^{\frac{1}{\tau_1}-1} \exp\left(-\frac{v_i}{\tau_1}\right); v_i > 0, \tau_1 > 0$$

and,

$$f(v_i; \tau_2) = \frac{1}{\Gamma\left(\frac{1}{\tau_2}\right) \tau_2^{\frac{1}{\tau_2}}} v_i^{\frac{1}{\tau_2}-1} \exp\left(-\frac{v_i}{\tau_2}\right); v_i > 0, \tau_2 > 0$$

By using equation 3.18 where the specific random effect is modeled as a mixture of two Gamma distributions, the  $L_{MP}$  function in equation 3.16 will be a mixture of two negative binomial density functions based on parameters  $\tau_1$  and  $\tau_2$  which is describing as follows:

$$\begin{aligned} L_{MP} &= \prod_{i=1}^M \int_0^{\infty} [v_i \mu_i]^{N_i} \exp\{-v_i \mu_i\} f(v_i; \tau) dv_i \\ &= \prod_{i=1}^M \int_0^{\infty} [v_i \mu_i]^{N_i} \exp\{-v_i \mu_i\} [pf(v_i; \tau_1) + (1-p)f(v_i; \tau_2)] dv_i \\ &= \prod_{i=1}^M p \int_0^{\infty} [v_i \mu_i]^{N_i} \exp\{-v_i \mu_i\} f(v_i; \tau_1) dv_i + (1-p) \int_0^{\infty} [v_i \mu_i]^{N_i} \exp\{-v_i \mu_i\} f(v_i; \tau_2) dv_i \\ &= \prod_{i=1}^M \left[ p \frac{\Gamma(N_i + \tau_1^{-1})}{(N_i!) \Gamma(\tau_1^{-1})} \left(\frac{\tau_1 \mu_i}{1 + \tau_1 \mu_i}\right)^{N_i} \left(\frac{1}{1 + \tau_1 \mu_i}\right)^{\tau_1^{-1}} + (1-p) \frac{\Gamma(N_i + \tau_2^{-1})}{(N_i!) \Gamma(\tau_2^{-1})} \left(\frac{\tau_2 \mu_i}{1 + \tau_2 \mu_i}\right)^{N_i} \left(\frac{1}{1 + \tau_2 \mu_i}\right)^{\tau_2^{-1}} \right] \end{aligned}$$

Regarding the Expectation-Maximization algorithm explained in chapter 2, a Bernoulli distributed indicator function  $Z_i$  can be defined for clustering the individuals into two groups. An individual belongs to first group ( $Z_i=1$ ) with probability  $p$  if it follows Negative binomial distribution with mean  $\mu_i$  and variance  $\mu_i(1 + \tau_1 \mu_i)$ , otherwise it will be in the second group following Negative binomial with mean  $\mu_i$  and variance

$\mu_i(1 + \tau_2\mu_i)$ . Now if the  $Z_i$ 's are known and we are aware of the grouping of all objects,

then  $L_{MP}$  can be rewritten as follows:

$$\begin{aligned}
L_{MP} &= \prod_{i=1}^M \int_0^{\infty} [v_i \mu_i]^{N_i} \exp\{-v_i \mu_i\} f(v_i; \tau) dv_i \\
&= \prod_{i=1}^M \left[ p \frac{\Gamma(N_i + \tau_1^{-1})}{(N_i!) \Gamma(\tau_1^{-1})} \left( \frac{\tau_1 \mu_i}{1 + \tau_1 \mu_i} \right)^{N_i} \left( \frac{1}{1 + \tau_1 \mu_i} \right)^{\tau_1^{-1}} + (1-p) \frac{\Gamma(N_i + \tau_2^{-1})}{(N_i!) \Gamma(\tau_2^{-1})} \left( \frac{\tau_2 \mu_i}{1 + \tau_2 \mu_i} \right)^{N_i} \left( \frac{1}{1 + \tau_2 \mu_i} \right)^{\tau_2^{-1}} \right]
\end{aligned} \tag{3.19}$$

The total likelihood function for panel data based on a mixture of gamma distributed random effects is then:

$$\begin{aligned}
L(\beta, \tau_1, \tau_2) &= \prod_{i=1}^M \left\{ \binom{N_i}{N_{i1}, \dots, N_{ie_i}} \prod_{j=1}^{e_i} \left( \frac{\Lambda_0(T_{ij}) - \Lambda_0(T_{i(j-1)})}{\Lambda_0(T_{ie_i})} \right)^{N_{ij}} \right\} \\
&= \left[ p \frac{\Gamma(N_i + \tau_1^{-1})}{(N_i!) \Gamma(\tau_1^{-1})} \left( \frac{\tau_1 \mu_i}{1 + \tau_1 \mu_i} \right)^{N_i} \left( \frac{1}{1 + \tau_1 \mu_i} \right)^{\tau_1^{-1}} + (1-p) \frac{\Gamma(N_i + \tau_2^{-1})}{(N_i!) \Gamma(\tau_2^{-1})} \left( \frac{\tau_2 \mu_i}{1 + \tau_2 \mu_i} \right)^{N_i} \left( \frac{1}{1 + \tau_2 \mu_i} \right)^{\tau_2^{-1}} \right]
\end{aligned} \tag{3.20}$$

If  $Z_i$ 's are known, the complete likelihood function for M-step of EM algorithm is:

$$\begin{aligned}
L_{comp}(\beta, \tau_1, \tau_2) &= \prod_{i=1}^M \left\{ \binom{N_i}{N_{i1}, \dots, N_{ie_i}} \prod_{j=1}^{e_i} \left( \frac{\Lambda_0(T_{ij}) - \Lambda_0(T_{i(j-1)})}{\Lambda_0(T_{ie_i})} \right)^{N_{ij}} \right\} \\
&= \left[ \frac{\Gamma(N_i + \tau_1^{-1})}{(N_i!) \Gamma(\tau_1^{-1})} \left( \frac{\tau_1 \mu_i}{1 + \tau_1 \mu_i} \right)^{N_i} \left( \frac{1}{1 + \tau_1 \mu_i} \right)^{\tau_1^{-1}} \right]^{Z_i} \left[ \frac{\Gamma(N_i + \tau_2^{-1})}{(N_i!) \Gamma(\tau_2^{-1})} \left( \frac{\tau_2 \mu_i}{1 + \tau_2 \mu_i} \right)^{N_i} \left( \frac{1}{1 + \tau_2 \mu_i} \right)^{\tau_2^{-1}} \right]^{1-Z_i}
\end{aligned} \tag{3.21}$$

The baseline cumulative intensity function  $\Lambda_0(t)$  is a function of time derived based on baseline intensity function  $\lambda_0(t)$  shown in 3.4. The baseline intensity function is modeled as a continuous function of t:

$\lambda_0(t) = \exp(X'\beta)$  where  $\mathbf{X}$  is the design vector and  $\boldsymbol{\beta}$  is the vector of parameters. The Exponential function always keeps the parameter of Poisson process positive since positive mean is one of the restrictions of Poisson models. In this thesis for representing the baseline intensity function as a continuous and smooth function of time,  $X'\beta$  is calculated via p-spline interpreted in chapter two. So the cumulative intensity function is computed as follows:

$$\Lambda_0(t+h) - \Lambda_0(t) = \int_t^{t+h} \exp(\lambda_0(u))du = \int_t^{t+h} \exp(B^T(u)\psi)du$$

where  $\mathbf{B}(u)$  is the vector of basis spline functions and  $\boldsymbol{\psi}$  is the vector of parameters which will be estimated by EM algorithm.

## Chapter 4

### Analysis

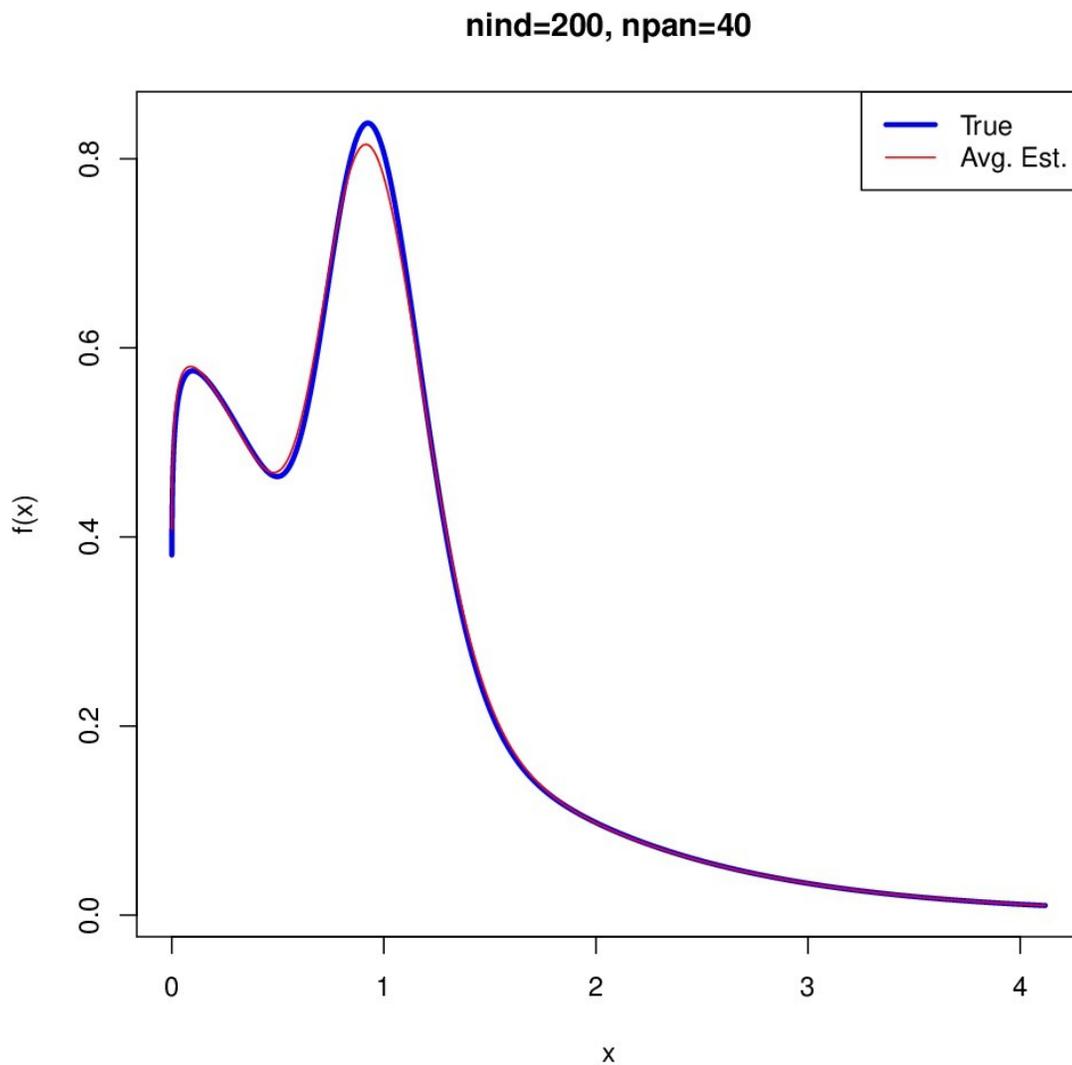
#### 4.1 Simulation

In the statistical inference area, there are some closed form formulas for estimation of parameters under consideration in probability models, and these estimators are helpful for some purposes such as construction of confidence interval, standard error and son on. However estimation of parameters is not always simple. One of the standard tools for estimating the statistical properties of estimators is statistical simulation which is employed in several approaches like bootstrapping (estimation and measures of accuracy are calculated by random sampling), Monte Carlo method (data are repeatedly generated from a specified model until the true estimation appears) and generation of estimated posterior distribution (the posterior distribution of parameter is derived by knowledge of prior distribution based on current data). The results of simulation are used to evaluate the statistical properties of estimators in probability models. In this research thesis 5000 datasets are generated from a non-homogeneous mixed Poisson process in which the parameter of the Poisson process is a function of time and the frailty's distribution is a mixture of two Gamma density functions. The parameters of the probability model are estimated by applying the generated data and initial values of model's parameters in the Expectation-Maximization algorithm. The results of simulations are presented in graphical forms for easier understanding.

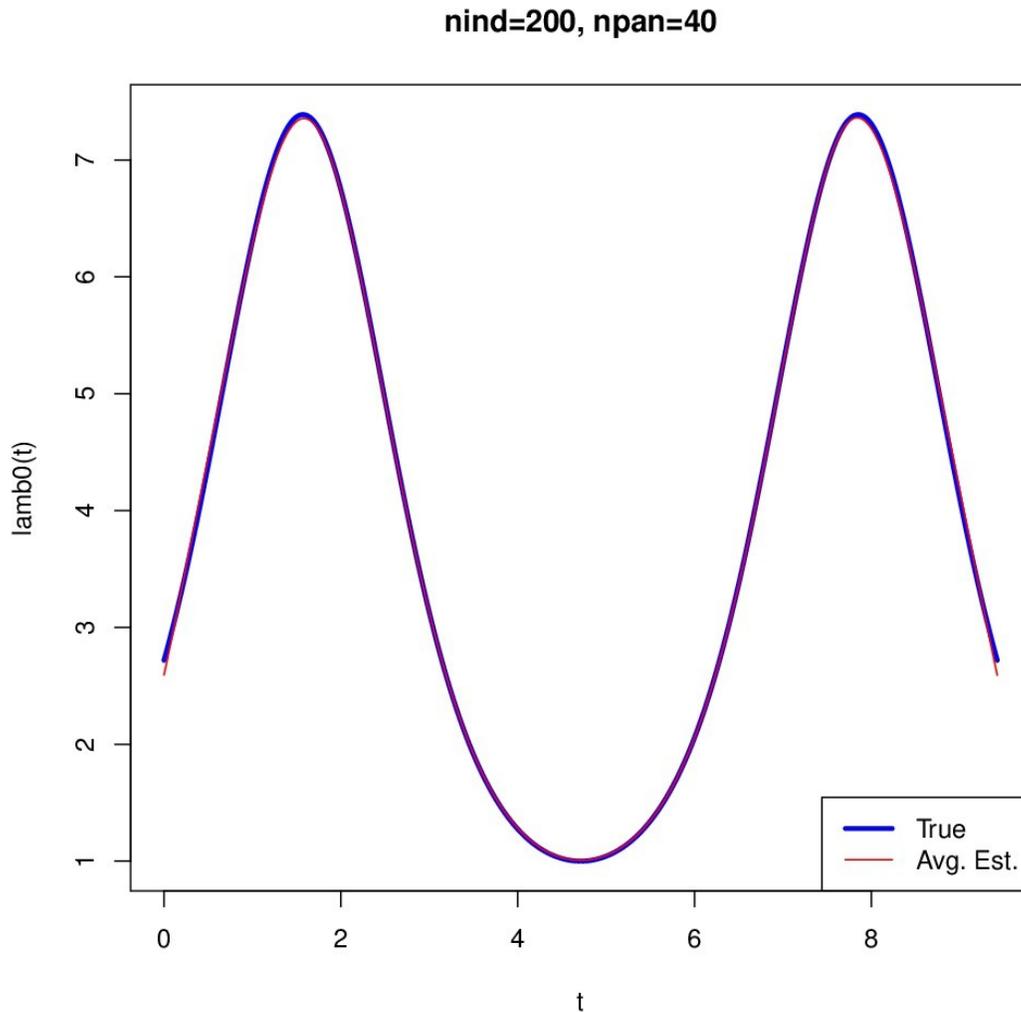
Charts 4.1 and 4.2 are the results of plotting the true frailty's density function  $f(v)$  and the true baseline intensity function  $\lambda_0(t)$  in blue curves against the average of 5000 estimated functions in red curves which is derived from simulation by using the EM

algorithm. The results of simulation obtained over 40 time panels and 200 individuals. In each plot,  $nind$  and  $npan$  is representing the number of individuals and the number of panels (time intervals) respectively. As can be seen in the chart 4.1, in spite of deviation from some points of the blue line, the estimated (red) line is alongside the true (blue) line implying that the probability model fits the data well. The same result is observed in the plot 4.2 where the estimated baseline intensity function is almost along the true one highlighting the reliability of the baseline intensity function.

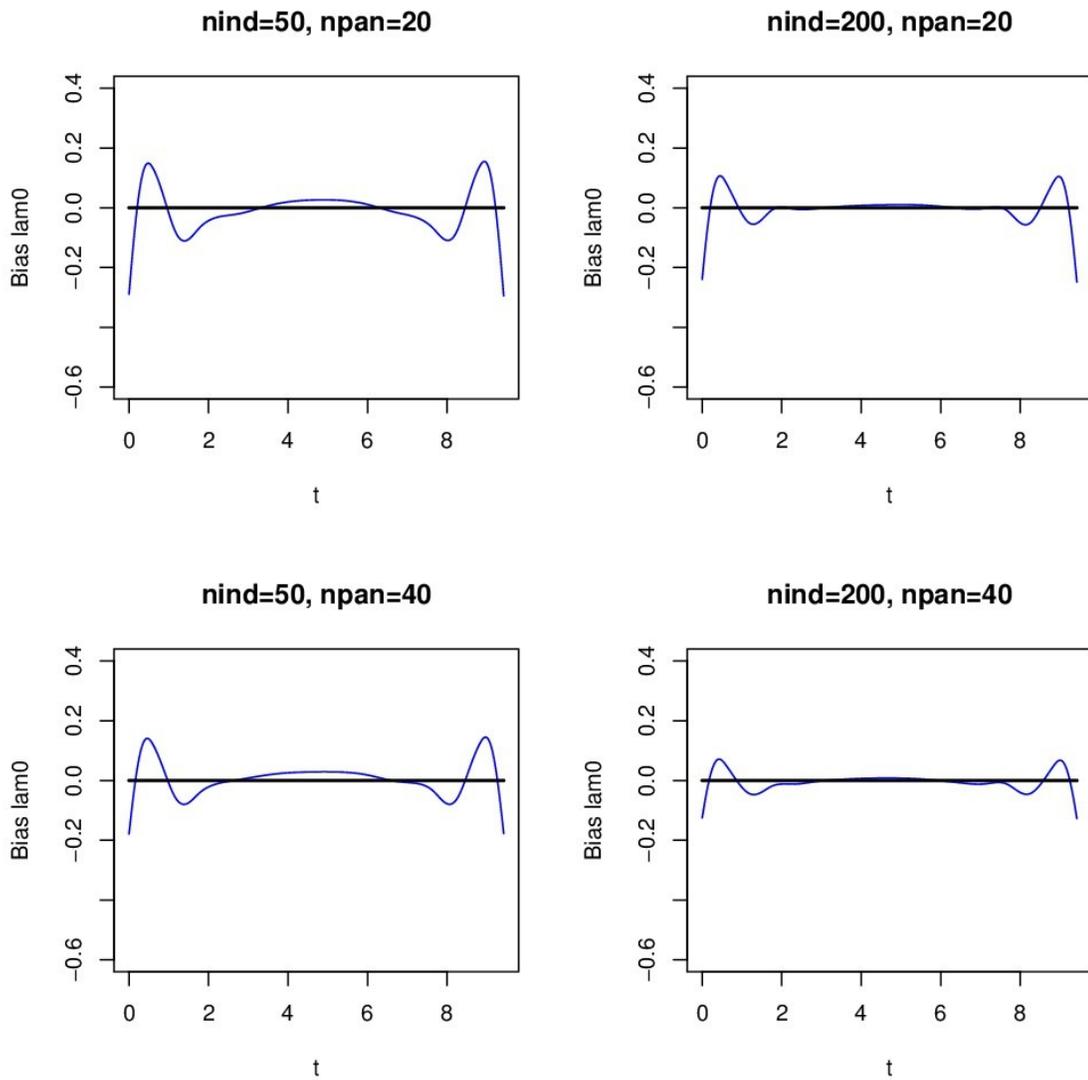
Chart 4.3 is demonstrating the resulting bias of the baseline intensity estimator in which each point is the difference between the true baseline intensity and the expectation of estimated values against time based on varying number of time panels and individuals. A comparison among the top two panels shows that the distance between baseline intensity function and the expectation of estimated values is getting very close to zero when the number of individuals is increasing from 50 to 200 for the same number of panels ( $npan=20$ ). It can be observed that the bias of the baseline intensity estimator is decreasing while the number of time panels is increasing from 20 in top left panel to 40 in the lower left panel. The same result is seen in the top right and lower right panels where an increase in the number of panels from 20 to 40 gives less bias for the estimator based on the same number of individuals. Finally it is noticeable that among all panels in the chart 4.3, the lower right panel gives the best result for the bias of the baseline intensity estimator representing that the estimator is getting more efficient as the number of data and time panels are getting larger.



**Figure 4.1:** A plot of the true frailty's density function in blue curve and the average of estimated frailty's density functions in red curve obtained from the simulation study with 200 individuals and 40 time panels. The plot explains the estimated curve fits the data well.

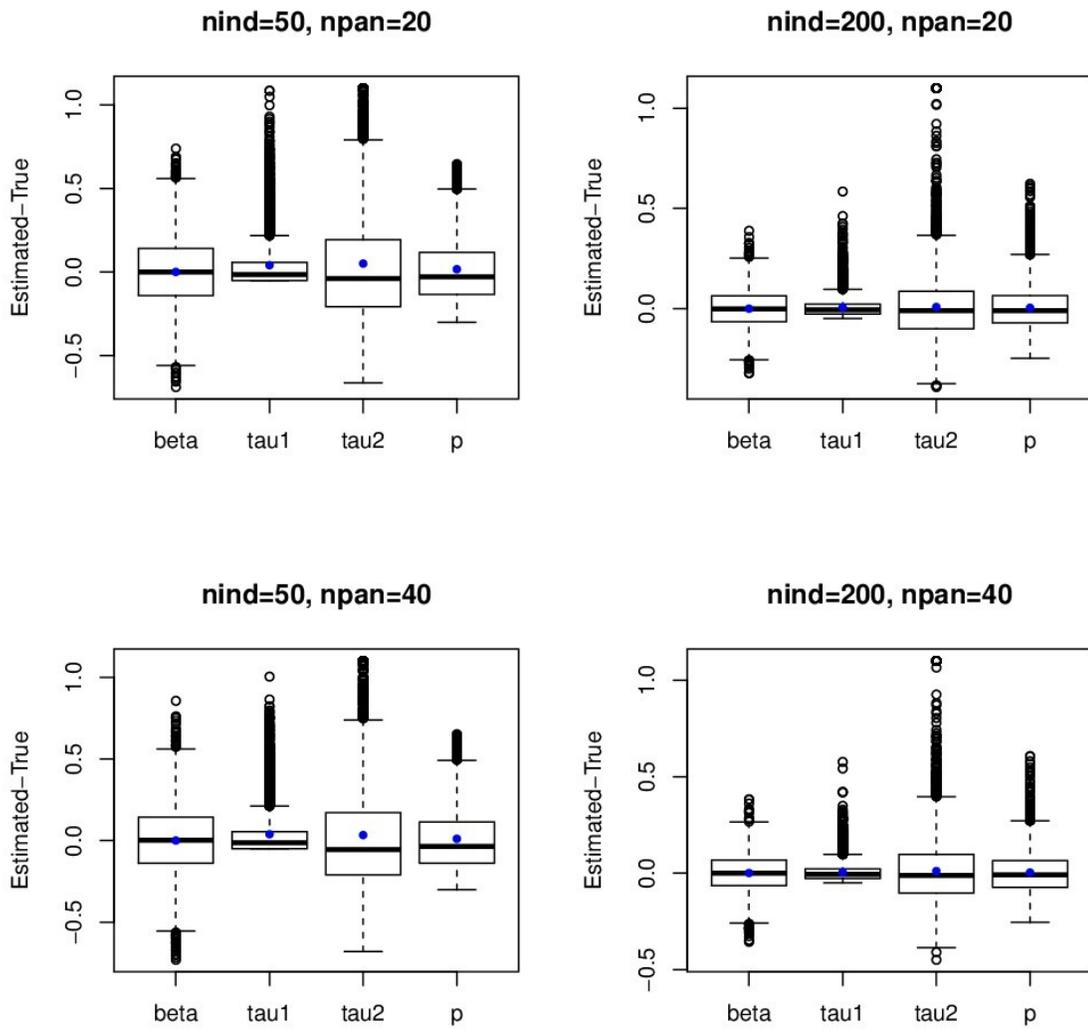


**Figure 4.2:** A plot of the true baseline intensity function in blue curve and the average of estimated intensity functions in red curve over time derived from simulation study with 200 individuals and 40 time panels. This plot highlights the reliability of the intensity function since the estimated function is plotted almost alongside the true function through time.



**Figure 4.3:** A matrix of plots showing the bias of the baseline intensity estimator over time in the simulation study with different number of individuals and time panels. A comparison among all plots represents that the bias of the intensity estimator is getting smaller while the number of individuals and time panels are getting larger.

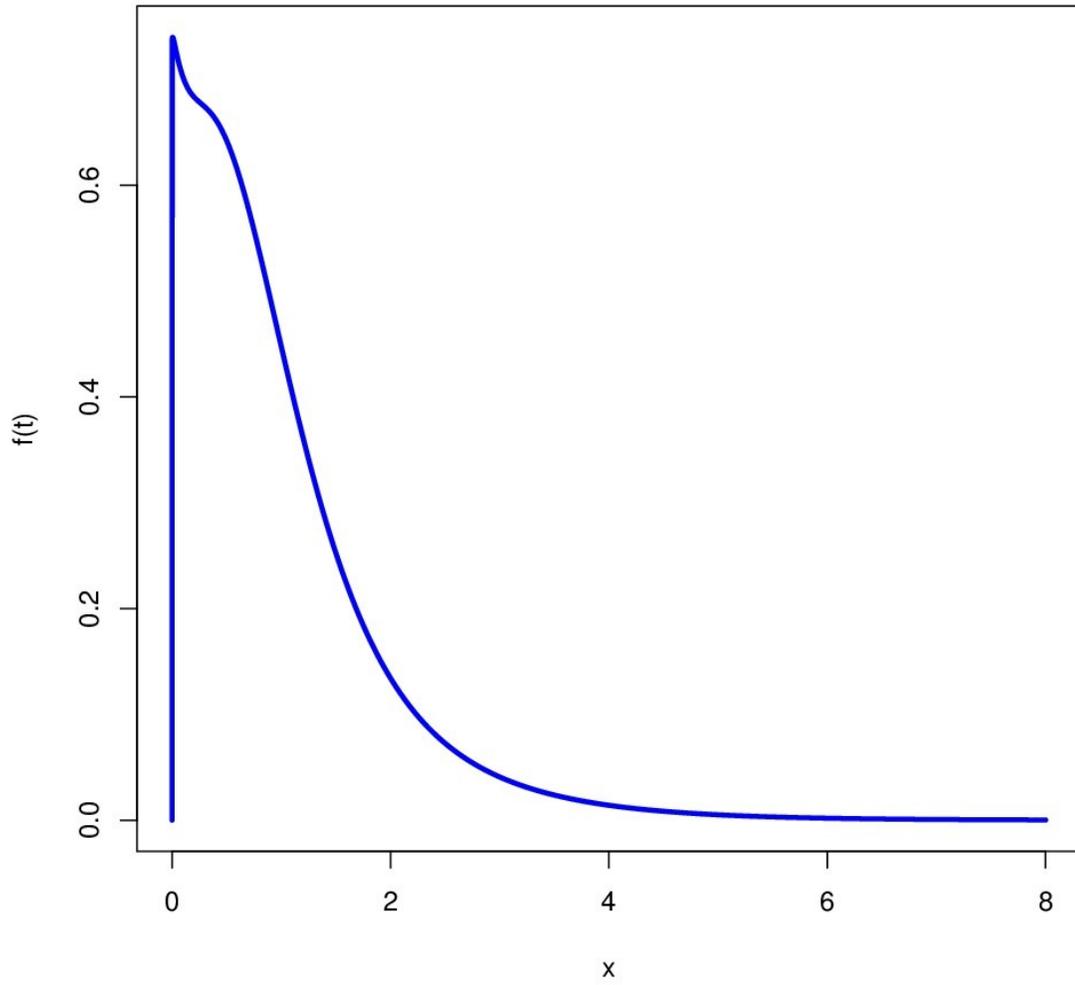
Chart 4.4 is a collection of classified boxplots representing the distribution of the bias of four parameters  $\beta, \tau_1, \tau_2$  and  $p$  based on the number of individuals and time panels in the simulation study. Every parameter's boxplot includes a set of points where each of them is the difference between the true and the estimated parameter values. Also the blue point and the black line in the box display the mean and the median of the (other points) bias of that parameter respectively where the equality of the mean and the median shows a symmetry in the distribution of points. A comparison between the top two plots explains that the bias of all estimators is more symmetric in upper right plot with 200 individuals than the bias in the upper left plot with 50 individuals at the same level of time panels ( $n_{pan}=20$ ). In the upper left plot the parameters  $\tau_1$  and  $\tau_2$  are overestimated since the mean exceeds the median. It is clear that there are not considerable changes in the distribution of bias when the number of time panels is increasing from 20 in the upper left plot to 40 in lower left plot with the same number of individuals ( $n_{ind}=50$ ). The similar results are observed for the two right plots with 200 individuals where an increase in the number of time panels from 20 in the upper right plot to 40 in the lower right plot doesn't affect the shape of distribution considerably. A total illustration of the chart implies that an increase in the number of individuals gives better estimation of parameters than the increase in the number of time panels.



**Figure 4.4:** A matrix of boxplots displaying the bias of the estimators for the parameters  $\beta, \tau_1, \tau_2$  and  $p$  in the simulation study. The charts are classified based on the number of individuals and time panels.

## 4.2 Data Analysis

So far we have evaluated the goodness of the Poisson mixed model by comparison of the true and the estimated values of parameters, baseline intensity function and frailty's model. It was found that in the simulation study, the proposed model looks well. Now we are going to fit the frailty's probability model to the Cherry Bark Tortrix Moth study data set. The aim of this study was to investigate the effect of pheromones on the mating patterns of the Cherry Bark Moth (*Enarmonia fomesana*). The pheromone in the study was demonstrated to be competitive with caged virgin females in seducing males into traps. The hypothesis under the study is that the mate seeking males are confused with location of females in the trees upon the release of pheromones. In order to examine the hypothesis, 20 cherry trees were considered in the experiment where all trees are attached with pheromone-baited trap in similar locations. All trees were provided with scent dispensers, 10 trees were chosen randomly as treatment group and their dispensers filled with female pheromones, the rest of 10 trees were considered as control group. The observations had shown that males were attracted to all of the traps filled with female baits but disappointed when they were trying to find the traps in treatment trees. The traps were cleaned once per week and the number of trapped moths was counted. We have a panel structure of 19 follow up intervals producing weekly counts. In this study, the baits were refreshed at 3 week intervals. Following the evaluation of the proposed count model and the baseline intensity function by using the simulation study, the Cherry Bark Moth data are going to be analyzed by the defined models. In chart 4.5 the random effect's probability model plotted against values of  $x$  is showing an increasing trend when  $x$  is around zero and then converges towards zero as  $x$  is getting larger.



**Figure 4.5:** A display of the frailty's density function formed by the Cherry Bark Moth data

As we explained in chapters 2 and 3, the Cherry Bark Moth data is modeled by the proposed models (3.21) in the chapter 3 and all parameters are estimated by the iterative Expectation- Maximization algorithm in R. The maximum likelihood estimators of the parameters derived based on the Cherry bark data are  $\hat{\beta} = 3.355$ ,  $\hat{\tau}_1 = .314$ ,  $\hat{\tau}_2 = .995$  and  $\hat{p} = .242$  with corresponding standard errors  $SE(\hat{\beta}) = .378$ ,  $SE(\hat{\tau}_1) = .380$ ,  $SE(\hat{\tau}_2) = .378$  and  $SE(\hat{p}) = .203$ . In this research, we are interested for estimating the 95 percent confidence interval of all parameters in the form of (estimate-1.96SE, estimate+1.96SE). A 95 percent confidence interval for each of the parameters by using the maximum likelihood estimator and standard error is:

$$p(2.614 < \beta < 4.096) = .95$$

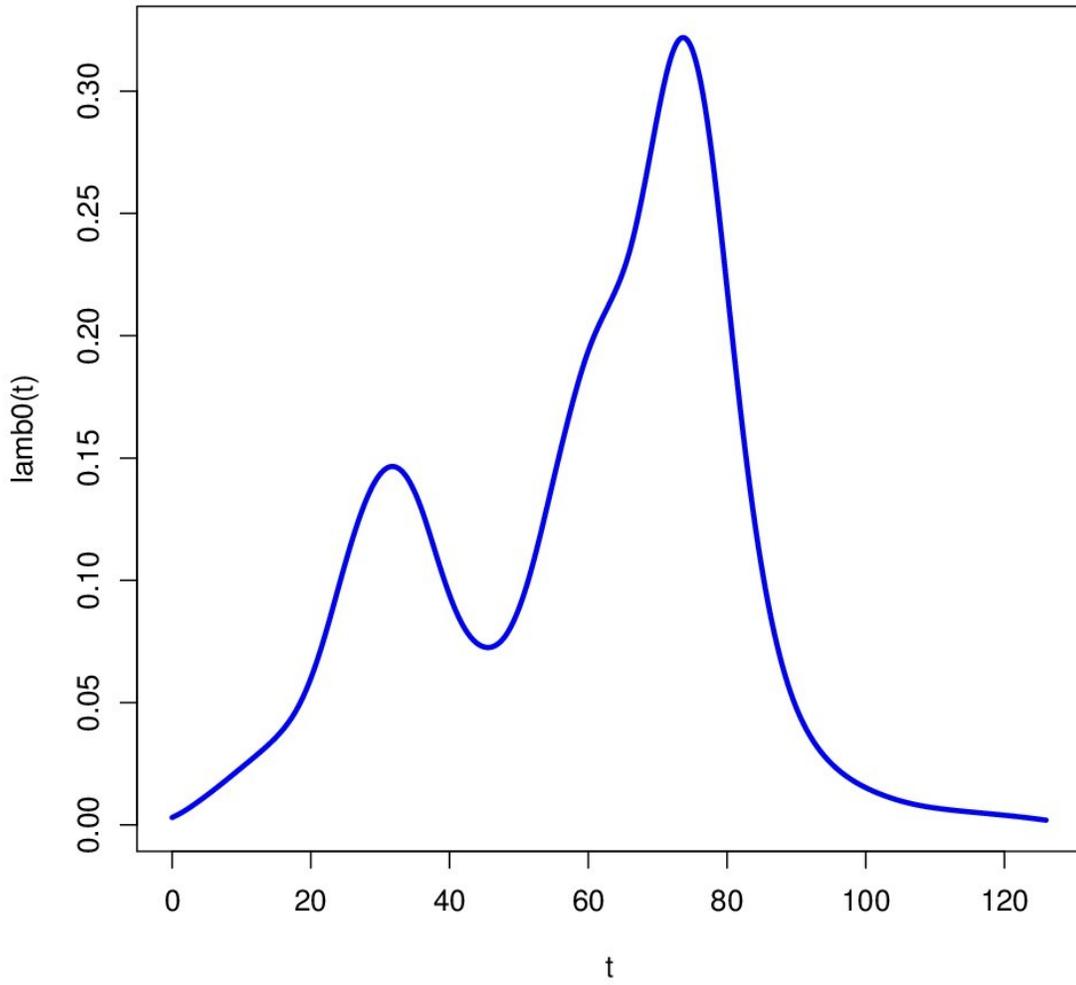
$$p(-.431 < \tau_1 < 1.059) = .95$$

$$p(.245 < \tau_2 < 1.736) = .95$$

$$p(-.156 < p < .640) = .95$$

The standard error of the parameters  $\beta, \tau_1, \tau_2$  are around .38 and the probability parameter has the lowest amount of dispersion with the estimate of  $SE(p)=.203$ . The 95 percent confidence intervals of the parameters  $\tau_1$  and  $p$  includes zero which determines that the values of the parameters are likely to be zero or near zero. As we know, there are two groups of control and treatment in the study and the amount of the parameter  $\beta$  is used for interpretation of the rate of the trapped moths. The 95 percent confidence interval of  $\beta$  doesn't include zero showing that it is very unlikely the amount of parameter goes toward zero and so the pheromone treatment in this study is significant at the 5 percent level of significance. The exponential function in the baseline intensity function explains how the rate of catching moths is increasing when the amount of  $\beta$  is big.

The chart 4.6 is displaying the baseline intensity function which is modeled based on p-spline function and formed by The Cherry Bark Moth data. As we mentioned beforehand, the baseline intensity function is expressing the rate of trapped moths by time  $t$ . The plot is showing some fluctuations against time, but finally converges towards zero for larger values of  $t$ . As can be seen, there are two peaks in the chart which interpret the changes in the trend of the rate of trapped moths during different time intervals. At first the rate of trapped moths is increasing from zero to around 15 percent which means more moths are looking for mating and consequently the breeding is increasing. The second peak is the highest point of the chart where the rate of trapped moths reaches above 30 percent implying the highest rate of breeding. At the end of the followup time intervals the baseline function is going toward zero which means the breeding is over and no more male moth is looking for mating. The p-spline function consists of a matrix of basis functions  $\mathbf{B}$  and the vector of parameters  $\mathbf{\Psi}$  estimated by using The EM algorithm.



**Figure 4.6:** A plot of the Baseline intensity function formed by the Cherry Bark Moth data

## Chapter5

### Conclusion

Panel data can be found in many fields of science including biostatistics, social sciences, economics, medical sciences and so on. The analysis of panel data is of great importance since the results will affect any future decisions or policy. In this research thesis count data referring to the number of events occurring in a time interval which is a popular type of data and is recorded over the time intervals. A Poisson model is usually chosen for analysis of counts which requires the mean to be equal to the variance of the data.

However, we often find data which doesn't satisfy this condition and the variance exceeds the mean of the data known as overdispersion. The data analysis by a Poisson model in the presence of overdispersion leads to under estimation of the standard error. An alternative approach is to assign a multiplicative specific random effect to each individual in a Poisson model where all individuals are sharing the same distribution. Among some possible choices, a gamma distributed random effect can be considered which gives a Negative binomial probability model. To create a more flexible model, a finite mixture of gamma distributions was considered for the random effect which gives a finite mixture of Negative binomial distribution with the same mean. The parameters of the final model are estimated using the iterative Expectation-Maximization algorithm. Then the statistical properties of the final model were evaluated in a simulation study.

In the simulation study, we obtained good statistical results for the proposed probability model explained in the chapter 3. It was shown that the estimation of parameters is affected by the total number of individuals but the total number of panels affected the baseline intensity function. We will be interested in the future to consider more

components (a mixture of three or four probability functions) and more complex structure for the individual random effect's distribution which can make the model more flexible.

## References

- [1] De Boor, C. A Practical Guide to Splines, Springer New York, vol.27.1978
- [2] Eilers, P.H.C., Marx, B.D. Flexible Smoothing with B-splines and Penalties. Statistical Science, vol.11, No.2.1996. pp.89-121
- [3] Friedman, J.H., Silverman, B.W. Flexible Parsimonious Smoothing and Additive Modeling. Journal of the American Statistical Association, vol.31, No.1.1989. pp.3-21
- [4] O'Sullivan, F., A Statistical Perspective on ||| – Posed Inverse Problems. Statistical Science, vol.1, No.4.1986, pp.502-518
- [5] Hastie, T., Tibshirani, R., Generalized Additive Models, First edition, Chapman & Hall, London (1990)
- [6] Dean, C.B. and Nielsen, J.D., Generalized linear Mixed models (2007). Journal of the Lifetime Data Analysis, vol.13, pp.497-512
- [7] Nelder, J.A., McCullagh, P., Generalized Linear Models. Second edition, Springer Science + Business Media, New York, NY.1989
- [8] Dean, C.B. (1998), Overdispersion, Encyclopedia of Biostatistics, Wiley, P. Armitage and T. Colton (eds.), pp.467-472
- [9] McLachlan, G.J., Krishnan, T., The Em Algorithm and Extensions. Second Edition, John Wiley & Sons, New York, 2007,
- [10] Winkelmann, R., Duration dependence and dispersion in count data models. Journal of Business and Economic Statistics, Vol.13, No.4.1995, pp.467-474
- [11] Ross, S.M., Introduction to probability models, 10<sup>th</sup> edition, Elsevier, California (2010)

[12] Lawless, J.F. (1982), *Statistical Models and Methods for Lifetime Data*. John Wiley and Sons, New York, 580 pp.

[13] Lawless, J.F., Negative Binomial and Mixed Poisson Regression. *The Canadian Journal of Statistics*, vol.15, 1987, pp.209-225

[14] Lawless, J.F., Ridge and Related Regression Estimators: Theory and Practice. *Journal of Communications in Statistics*, vol.7, No.2.1978, pp.139-164

[15] Dean, C.B., Lawless, J.F., Willmot, G.E., A mixed-Poisson inverse Gaussian Regression model, *Canadian Journal of Statistics*, vol.17, No.2.1989, pp.171-181