Prediction of recurrent events

by

Marc Fredette

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2004

Author's declaration for electronic submission of a thesis

I hereby declare that I am the sole author of this thesis. This is a true copy of my thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In this thesis, we will study issues related to prediction problems and put an emphasis on those arising when recurrent events are involved. First we define the basic concepts of frequentist and Bayesian statistical prediction in the first chapter. In the second chapter, we study frequentist prediction intervals and their associated predictive distributions. We will then present an approach based on asymptotically uniform pivotals that is shown to dominate the plug-in approach under certain conditions.

The following three chapters consider the prediction of recurrent events. The third chapter presents different prediction models when these events can be modeled using homogeneous Poisson processes. Amongst these models, those using random effects are shown to possess interesting features. In the fourth chapter, the time homogeneity assumption is relaxed and we present prediction models for non-homogeneous Poisson processes. The behavior of these models is then studied for prediction problems with a finite horizon. In the fifth chapter, we apply the concepts discussed previously to a warranty dataset coming from the automobile industry. The number of processes in this dataset being very large, we focus on methods providing computationally rapid prediction intervals. Finally, we discuss the possibilities of future research in the last chapter.

Acknowledgements

First, I would like to thank my supervisor Jerry Lawless for his guidance, insights, and support. You have no idea how much I learned from all our chats and I consider myself lucky that I had a supervisor like you. Thank you for helping me to fulfill my dream!

I wish to thank my thesis committee: Dr. Gord Willmot, Dr. Mary Thompson, Dr. Mahesh Pandey, and Dr. Bill Meeker (Iowa State University) for their dedication in reviewing my thesis and for their helpful comments and suggestions.

This research was supported by scholarships from the *Natural Sciences and Engineering Research Concil of Canada* (NSERC) and by *Le Fonds pour la Formation de Chercheurs et l'Aide à la Recherche du Québec* (FCAR). Both scholarships allowed me to focus my energies on completing this research.

Un merci bien spécial à mes amis avec qui j'ai partagé cette intéressante aventure que fut ce doctorat: Aurélie, merci pour les beaux souvenirs. Merci aussi à Chantal, Christian, Cody, Jonathan, Kristina, Louise et Philippe....WE DID IT!!!

Finalement, je veux remercier ma famille, on sait bien que ca n'a pas toujours été facile pour nous tous au cours des dernières années. Cependant, non seulement vous n'avez jamais cessé de m'appuyer et de m'encourager, mais de plus vous ne m'avez jamais fait sentir coupable d'être loin à des moments où j'aurai dû (voulu) être près...je vous aime.

Contents

1 Statistical Prediction **1**

1.1 Terminology 2

1.1.1 Point predictor and prediction interval 2

1.1.2 Tolerance interval 4

1.2 Prediction Intervals Under a Frequentist Framework 6

1.2.1 Pivotal Method 7

1.2.2 Test-Inversion Method 9

1.2.3 Plug-in Method 10

1.2.4 Predictive density and likelihood method 11

1.3 Bayesian Framework 12

1.3.1 Prior and Posterior Distributions 13

1.3.2 Bayesian Predictive Density 14

1.3.3 Point and Set Predictions 15

1.3.4 Non-informative prior 17

1.3.5 Empirical Bayes Methods 18

1.4 Thesis Outline 19

2 Predictive Distributions and Calibration **21**

2.1 Predictive Distributions 22
2.2 Calibration 24
2.3 A Proposed Approach 26
2.4 Illustrations 32
2.4.1 Exponential distribution 32
2.4.2 Log-normal distribution 36

3 Prediction Models for Homogeneous Poisson Processes 39
3.1 Prediction of Recurrent Events 40
3.2 Prediction of Homogeneous Poisson Processes 41
3.2.1 Point Prediction 43
3.2.2 Prediction intervals 45
3.3 Prediction Models using Random Effects 49
3.3.1 Random Effects Model 49
3.3.2 Complete Specification of the Random Effects Model 55
3.4 Simulations 58
3.4.1 Point Prediction 59
3.4.2 Set Prediction 64

4 Prediction Models for Nonhomogeneous Poisson Processes 75
4.1 Point Predictors and Prediction Intervals 77
4.2 Random Effects Model 79
4.3 Empirical study 86

5 Prediction of Warranty Claims 103
5.1 Motivating Dataset 103
5.2 Prediction Model Proposed 106

5.3 Model Fitting 113
5.3.1 Starting Values 113
5.3.2 Numerical Results 117
5.3.3 Prediction Intervals 125
5.4 Calibration 131
5.5 Calibration using asymptotic normality 137

6 Future Research and Other Topics **141**
6.1 Robustness and sensitivity 141
6.2 Model extensions 143
6.3 Other topics 144

List of Tables

3.1 Sets of fixed rates used in the simulations. 60
3.2 Predictor for each method. 61
3.3 Comparison of the discrepancies of point predictors. 62
3.4 Comparison of the discrepancies using different distributions for the rates. 65
3.5 Comparison of the average KL distance for different methods. 66
3.6 Comparison of the average KL when the rates are random. 67
3.7 Coverage proportions (and average lengths) of 90% prediction intervals. . . 69
3.8 Coverage proportions of 90% prediction intervals (random rates). 73

4.1 Score equations for m.l.e.'s. 78
4.2 A list of NHPP models. 87
4.3 Maximum likelihood estimates for the LOG dataset. 93
4.4 Absolute error of point predictors (LOG dataset). 94

5.1 Number of cars with the same number of claims. 105
5.2 Number of claims observed at every given time. 118
5.3 Estimates of a, b, and c. 119
5.4 Estimates of β when $q = 4$. 120
5.5 Correlation matrix of $\hat{\beta}$ (t=571). 121

5.6 Distribution of the total claims amongst all the cars. 122
5.7 Approximated coverage probability of 95% plug-in intervals. 135

List of Figures

2.1 Functions $\tilde{G}(u)$ and $\tilde{F}_p(y|x)$, exponential distribution. 34
2.2 The loss in coverage probability induced by using the plug-in approach. . . 35
2.3 Functions $\tilde{G}(u)$ and $\tilde{F}_p(y|x)$, log-normal distribution. 38

3.1 Empirical coverage probabilities of 90% prediction intervals. 71

4.1 Recurrence data plots. 89
4.2 Real and plug-in 90% prediction intervals for the simulated LOG dataset. . 91
4.3 Plug-in 90% prediction intervals for the simulated LOG dataset. 92
4.4 Real and plug-in 90% prediction intervals for the simulated EXP dataset. . 96
4.5 Real and plug-in 90% prediction intervals for the simulated POW dataset. 97
4.6 Plug-in 90% prediction intervals for the TUMOR dataset. 98
4.7 Calibrated 90% prediction intervals for the simulated EXP dataset. 99
4.8 Calibration curves for the EXP dataset 100
4.9 Calibrated 90% prediction intervals for the TUMOR dataset. 101

5.1 Histogram of the occurrence times. 105
5.2 Warranty claims occurrences (time of sale is the origin). 107
5.3 Quantile-quantile plot of the $\hat{u}_{ij}$'s (t=571). 123
5.4 Quantile-quantile plots of the $\hat{u}_{ij}$'s. 124

5.5 Histogram of the occurrence times (t=571). 126
5.6 Histograms of the occurrence times. 127
5.7 Non-calibrated 95% prediction intervals 128
5.8 Non-calibrated 95% prediction intervals using $q = 1, \ldots, 4$ 130
5.9 Non-calibrated 95% prediction intervals with different $f(t; \beta)$ 132
5.10 Calibrated 95% prediction intervals. 134
5.11 Calibration curves. 136
5.12 Quantile-quantile plots between $\widehat{\mathbf{U}}$ and $\mathbf{U}$. 138
5.13 Comparison of calibration methods. 140

Chapter 1

Statistical Prediction

In traditional statistical analysis, we use the information contained in a sample to make inferences about the population where this sample was taken from. Usually, these inferences are based on estimates, confidence intervals or hypothesis tests for parameters of a specified model. If this model describes adequately the population, the analysis containing these inferences are appropriate for most scientific problems. On the other hand, we also encounter problems where the understanding of the population's behavior is not of interest by itself; it is a means of foretelling future events. We call such problems prediction problems. Since prediction problems are rarely a case of logical deduction, the use of probabilistic and statistical tools are inevitable in any scientific approach used to solve them. Many early papers deal with prediction; for example Pearson (1920), Baker (1935), de Finetti (1937), or Wilks (1942). However, even if prediction problems are often encountered, statisticians are now devoting most of their attention to inferential problems. An interesting discussion about the neglect of prediction analysis can be found in the preface of Aitchison & Dunsmore's (1975) book and this issue is also discussed in Geisser (1993).

Statistical prediction can be applied in many domains such as engineering, industry,

business, and medicine. In each of these domains, it can be used for planning purposes (predict the total medical cost of a population, predict a future number of insurance claims), for process monitoring (predict the number of nuclear scrams in a power plant), or for decision making (software debugging, determination of a maintenance policy). Since the basic prediction concepts used in these applications are rarely mentioned in introductory statistical books, this chapter will be devoted to defining these concepts. The definitions in this chapter were derived from Aitchison & Dunsmore (1975), Nelson (1982), and Meeker & Escobar (1998, Chapter 12).

1.1 Terminology

Let us assume that a set of random variables $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ was drawn and is used to predict the future random variable $\mathbf{Y}$. We will usually assume that $\mathbf{Y}$ is univariate, though it may be a function of a multivariate outcome (eg. a sum, a mean, an order statistic). The density functions of $\mathbf{X}$ and $\mathbf{Y}$ are specified up to a common vector parameter θ and will be denoted by $f_{\mathbf{X}}(x;\theta)$ and $f_{\mathbf{Y}}(y;\theta)$. Note that when no confusion can arise, these densities will be denoted by $f(x;\theta)$ and $f(y;\theta)$ instead.

1.1.1 Point predictor and prediction interval

When we want to infer about an unknown parameter, two popular approaches are to find a point estimate or an interval of possible values for this parameter. In a prediction problem, we proceed in a similar way to predict the future random value $\mathbf{Y}$: we can either try to find a point predictor or a prediction interval.

Definition 1.1. *A **predictor** $\widehat{\mathbf{Y}}(x)$ is a function of the already observed x that is used to predict the realization of the random variable $\mathbf{Y}$.*

Both $\widehat{\mathbf{Y}}(\mathbf{X})$ and $\mathbf{Y}$ are random variables and their difference, is called the prediction error. When this error has an expectation of 0 we say that $\widehat{\mathbf{Y}}(\mathbf{X})$ is an unbiased predictor for $\mathbf{Y}$. However, when we are looking for a predictor of a future random value the bias is not the only concern; we also want the variance of the prediction error to be as small as possible.

Definition 1.2. *An unbiased predictor $\widehat{\mathbf{Y}}(\mathbf{X})$ is called the* ***the best unbiased predictor*** *for $\mathbf{Y}$ if its prediction error, $\widehat{\mathbf{Y}}(\mathbf{X}) - \mathbf{Y}$, has a smaller variance than any other unbiased predictors of $\mathbf{Y}$.*

Among all the concepts introduced in this section, the prediction interval is the one we will use the most often throughout this thesis.

Definition 1.3. *An interval with lower and upper endpoints $L(x)$ and $U(x)$ is called a $(1-\alpha)$* ***prediction interval*** *for $\mathbf{Y}$ if*

$$P[L(\mathbf{X}) \leq \mathbf{Y} \leq U(\mathbf{X})] \;=\; 1-\alpha. \tag{1.1}$$

When $\dim(\mathbf{Y}) > 1$, a region $R(x)$ such that

$$P[\mathbf{Y} \in R(\mathbf{X})] \;=\; 1-\alpha,$$

is called a $(1-\alpha)$ ***prediction region*** *for the vector $\mathbf{Y}$. The quantity $1-\alpha$ is called the* ***coverage probability*** *of the prediction interval or prediction region.*

This definition means that under repeated sampling of both $\mathbf{X}$ and $\mathbf{Y}$, y will be included in $[L(x), U(x)]$, or $R(x)$, in $(1-\alpha)100\%$ of the samples. If Definition 1.2 suggests an appropriate way to compare unbiased predictors, it is more difficult to compare prediction regions having the same coverage probability. Nevertheless, a possible way to quantify their efficiency is to compare their expected volume. For example, when many $(1-\alpha)$

prediction intervals are available, it seems appropriate to use the prediction interval having the smallest value for $\mathbb{E}[U(\mathbf{X}) - L(\mathbf{X})]$.

In general, the distribution of the three random variables $L(\mathbf{X})$, $\mathbf{Y}$, and $U(\mathbf{X})$ will depend on an unknown parameter θ. Therefore, it may be not possible to make (1.1) independent of θ. In this case, we will call $[L(x), U(x)]$ a $(1-\alpha)$ prediction interval for $\mathbf{Y}$ if

$$P[L(\mathbf{X}) \leq \mathbf{Y} \leq U(\mathbf{X}); \theta] \quad \geq \quad 1 - \alpha,$$

for all θ.

Because the preceding concepts are similar to those used to estimate a characteristic (*i.e.* parameter) of a distribution, many prediction problems go unrecognized and are incorrectly treated as estimation problems. This can lead to important errors. For example, suppose that we wish to predict a new observation coming from the linear regression model $\mathbf{Y} = Z\beta + \varepsilon$. If this prediction problem is incorrectly treated as an estimation problem, one could use the previously observed sample to obtain a confidence interval for $\mathbb{E}[\mathbf{Y}] = Z\beta$. Because it does not take into account the variability of ε, this procedure will likely give intervals that are too narrow and thus has a coverage probability below the desired level.

1.1.2 Tolerance interval

Another type of interval also used for prediction purposes is the tolerance interval. Many early papers deal with these intervals, for example Wilks (1941), Wald (1942), and Scheffe & Tukey (1945). We will present here two types of tolerance intervals for the random variable $\mathbf{Y}$; both are determined from the previously observed x.

Definition 1.4. *Let $C_{L,U}(x)$ be the coverage probability of the interval $[L(x), U(x)]$ given*

that $\mathbf{X} = x$*, i.e.*

$$C_{L,U}(x) = P[L(\mathbf{X}) \leq \mathbf{Y} \leq U(\mathbf{X}) \mid \mathbf{X} = x].$$

Therefore, $C_{L,U}(\mathbf{X})$ *is a random variable taking values between 0 and 1. In general, it also depends on* θ *and we may write* $C_{L,U}(x;\theta)$ *to remind ourselves of this.*

1. *If* $P[C_{L,U}(\mathbf{X}) \geq \beta] = \gamma$*, then* $[L(x), U(x)]$ *is called a* β***-content tolerance interval with confidence*** γ *for* $\mathbf{Y}$*.*

2. *If* $\mathbb{E}[C_{L,U}(\mathbf{X})] = \beta$*, then* $[L(x), U(x)]$ *is called a* β***-expectation tolerance interval*** *for* $\mathbf{Y}$*.*

The following proposition shows the equivalence between a $(1-\alpha)$-expectation tolerance interval and a $(1-\alpha)100\%$ prediction interval.

Proposition 1.1. *The interval* $[L(x), U(x)]$ *is a* $(1-\alpha)$ *prediction interval for* $\mathbf{Y}$ *if and only if it is also a* $(1-\alpha)$*-expectation tolerance interval for* $\mathbf{Y}$*.*

Proof. First we have,

$$\begin{aligned} \mathbb{E}[C_{L,U}(\mathbf{X})] &= \int_X C_{L,U}(x) f(x;\theta) dx \\ &= \int_X P[L(\mathbf{X}) \leq \mathbf{Y} \leq U(\mathbf{X}) \mid \mathbf{X} = x] f(x;\theta) dx \\ &= P[L(\mathbf{X}) \leq \mathbf{Y} \leq U(\mathbf{X})], \end{aligned}$$

where $\int_X$ means that we integrate over the sample space of $\mathbf{X}$. The equivalence between these two types of intervals is then clear from Definition 1.3 and 1.4. □

The main difference between the two types of tolerance intervals is that the β-content tolerance interval with confidence γ uses two parameters to describe the level of uncertainty: γ quantifies the uncertainty due to the sampling of $\mathbf{X}$, while β is due to the randomness of $\mathbf{Y}$. This is not the case for the β-expectation tolerance interval; it uses only one parameter to incorporate both the sampling variability of $\mathbf{X}$ and $\mathbf{Y}$. The β-expectation tolerance intervals are used when we want to predict a single realization of $\mathbf{Y}$. On the other hand, if we wish to obtain an interval where at least a given proportion of units will lie within this interval with a fixed confidence, a β-content tolerance interval with confidence γ is more appropriate.

1.2 Prediction Intervals Under a Frequentist Framework

As we mentioned previously, sometimes it is not possible to find a $(1-\alpha)$ prediction interval with a coverage probability independent of the unknown parameter θ. We then have to find a procedure such that the coverage probability is never less than $1-\alpha$ for all possible values of θ. Unfortunately, this approach often provides trivial prediction intervals (*i.e.* intervals including all the possible values of $\mathbf{Y}$).

Example 1.1. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be exponentially distributed with a common* θ*. Suppose that we want to find a positive constant* c *such that* $x \pm c$ *is a 90% prediction interval for* $\mathbf{Y}$*. The coverage probability of this interval will be a function of the unknown* θ*:*

$$\begin{aligned} P[\mathbf{X}-c \leq \mathbf{Y} \leq \mathbf{X}+c] &= \int_0^\infty \left(\int_{x-c}^{x+c} f(y;\theta)dy \right) f(x;\theta)dx \\ &= \int_0^\infty \left(F_{\mathbf{Y}}(x+c;\theta) - F_{\mathbf{Y}}(x-c;\theta) \right) f(x;\theta)dx \end{aligned}$$

$$\begin{aligned} &= \int_0^\infty \left[e^{-\theta(x-c)} - e^{-\theta(x+c)}\right] \theta e^{-\theta x} dx \\ &= (e^{\theta c} - e^{-\theta c}) \int_0^\infty \theta e^{-2\theta x} dx \\ &= (e^{\theta c} - e^{-\theta c})/2 \\ &= \sinh(\theta c). \end{aligned}$$

Since $\lim_{\theta \to 0} \sinh(\theta c) = 0$ *when* c *is fixed, the only way we can have a coverage probability never smaller than 0.9 for all* θ *is to set* $c = \infty$*, which is a trivial prediction interval for* $\mathbf{Y}$.

In this section, we will present methods to find non-trivial prediction intervals under a frequentist framework (*i.e.* when θ is unknown, but assumed constant). This exhaustive list of methods can be divided in four categories: the pivotal method, the test-inversion method, the plug-in method, and methods using approximate predictive densities. Only the first two methods allow us to find exact prediction intervals, whereas the others are used to find approximate prediction intervals. Methods using a Bayesian approach are also available and will be presented in the next section.

1.2.1 Pivotal Method

An easy way to obtain exact prediction intervals is available when there exists a random variable $Z(\mathbf{X}, \mathbf{Y})$ whose distribution does not depend on θ. Suppose that this variable, called a pivotal random variable, is such that we can find some a and b where

$$P[a \leq Z(\mathbf{X}, \mathbf{Y}) \leq b] = 1 - \alpha.$$

Then, we can find an exact prediction interval for $\mathbf{Y}$ in terms of $\mathbf{X}$ if this equation can be inverted into the form (1.1).

Pivotals are often used when $\mathbf{Y}$ has a location-scale distribution. We recall that a random variable has such a distribution, with location parameter μ and scale parameter σ, when its density function $f(t;\, \mu,\, \sigma)$ can be written as

$$f(t;\, \mu,\, \sigma) \;\; = \;\; \frac{1}{\sigma} g\left(\frac{t-\mu}{\sigma}\right).$$

When $\mathbf{Y}$ and $\mathbf{X}$ have such a distribution, we can show that the random variable $Z(\mathbf{X}, \mathbf{Y}) = [\mathbf{Y} - \hat{\mu}(\mathbf{X})]/\hat{\sigma}(\mathbf{X})$ is a pivotal random variable when $(\hat{\mu}(x), \hat{\sigma}(x))$ is the maximum likelihood estimate of (μ, σ). In fact, this is true for all estimates of (μ, σ) having certain invariance properties. Note that the maximum likelihood estimates are no longer satisfying these properties under certain censoring schemes. A proof of this can be found in Lawless (2003, Appendix E).

Now let $q_{_Z}(\alpha)$ be the parameter-free α quantile of $Z(\mathbf{X}, \mathbf{Y})$. Then,

$$P[q_{_Z}(\alpha_1) \leq \frac{\mathbf{Y} - \hat{\mu}(\mathbf{X})}{\hat{\sigma}(\mathbf{X})} \leq q_{_Z}(1-\alpha_2)] \;\; = \;\; 1 - \alpha,$$

when $\alpha_1 + \alpha_2 = \alpha$. By rearranging these terms we have,

$$P[\hat{\mu}(\mathbf{X}) + \hat{\sigma}(\mathbf{X}) q_{_Z}(\alpha_1) \leq \mathbf{Y} \leq \hat{\mu}(\mathbf{X}) + \hat{\sigma}(\mathbf{X}) q_{_Z}(1-\alpha_2)] \;\; = \;\; 1 - \alpha,$$

which gives us a $(1-\alpha)$ prediction interval for $\mathbf{Y}$. When the quantiles of $Z(\mathbf{X}, \mathbf{Y})$ cannot be obtained analytically, we can approximate them using a Monte Carlo method; a procedure is suggested in Meeker & Escobar (1998, Section 12.4). The prediction interval is then approximate but only because of Monte Carlo error, which can be made arbitrarily small.

This method allows us to find a non-trivial prediction interval for the problem presented in Example 1.1.

Example 1.2. *(Example 1.1 revisited) Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be exponentially distributed with a common rate* θ*. Now suppose that we want a 90% prediction interval for* $\mathbf{Y}$*. Since* $\mathbf{Y}$ *has*

a scale distribution with $\sigma = \theta^{-1}$ *and* x *is the maximum likelihood estimate of* σ*, we know that* $\mathbf{Z} = \mathbf{Y}/\mathbf{X}$ *is parameter-free. In fact, we can easily show that* $f(z) = (1+z)^{-2}$ *and* $q_Z(\alpha) = (1-\alpha)^{-1} - 1 = (q_Z(1-\alpha))^{-1}$*. Thus,*

$$P[\mathbf{X}q_Z(0.05) \leq \mathbf{Y} \leq \mathbf{X}(q_Z(0.05))^{-1}] \;\; = \;\; 0.90$$

and $[x/19, 19x]$ *is then a 90% prediction interval for* $\mathbf{Y}$*. In Example 1.1 we were not able to find a non-trivial prediction interval of the form* $x \pm c$*, but the pivotal method provided us one of the form* $[c^{-1}x, cx]$*.*

1.2.2 Test-Inversion Method

Another method uses hypothesis tests' critical regions to obtain exact prediction intervals. Unfortunately, this approach is usually only applicable to a limited number of problems. First, we suppose that the distributions of $\mathbf{X}$ and $\mathbf{Y}$ depend only on a common parameter θ, but that the values $\theta = \theta_1$ for $\mathbf{X}$ and $\theta = \theta_2$ for $\mathbf{Y}$ may be different. Then, we consider the null hypothesis $H_0 : \theta_1 = \theta_2$. If we can find a critical region $Q_{1-\alpha}$ of level $1-\alpha$, we have

$$P[(\mathbf{X}, \mathbf{Y}) \in Q_{1-\alpha} | H_0] = 1 - \alpha,$$

for all $\theta\,(= \theta_1 = \theta_2)$. Then if the region $Q_{1-\alpha}$ can be projected onto the subspace $\mathbf{X} = x$, this new region $R_{1-\alpha}(x)$ will be an exact $(1-\alpha)$ prediction region since

$$P[\mathbf{Y} \in R_{1-\alpha}(\mathbf{X}); \theta] = 1 - \alpha$$

for all θ, where

$$R_{1-\alpha}(x) = \{y; (x, y) \in Q_{1-\alpha}\}.$$

According to the critical region chosen, this method can give many different prediction regions. For a specified alternative hypothesis H_A, it is somewhat natural to use a critical region that is uniformly or locally most powerful. However, it is not usually obvious to determine which alternative hypothesis should be used to obtain an efficient prediction region. Again, among many candidates, the one having the smallest expected volume could be selected.

A method often providing similar results was suggested in Faulkenberry (1973). This method requires the existence of a sufficient statistic for θ coming from the random vector $(\mathbf{X}, \mathbf{Y})$. If such a statistic $S(\mathbf{X}, \mathbf{Y})$ is available, we try to find an interval $[L(x), U(x)]$ such that

$$P[L(\mathbf{X}) \leq \mathbf{Y} \leq U(\mathbf{X}) | S(\mathbf{X}, \mathbf{Y}) = s] = 1 - \alpha$$

for all s. Because of the sufficiency of $S(\mathbf{X}, \mathbf{Y})$, this probability statement is independent of θ. The prediction intervals are then equivalent under the conditional and unconditional distributions:

$$\begin{aligned} P[L(\mathbf{X}) \leq \mathbf{Y} \leq U(\mathbf{X})] &= \mathbb{E}\left[P[L(\mathbf{X}) \leq \mathbf{Y} \leq U(\mathbf{X}) | S(\mathbf{X}, \mathbf{Y}) = s]\right] \\ &= \mathbb{E}[1 - \alpha] \\ &= 1 - \alpha. \end{aligned}$$

1.2.3 Plug-in Method

Now let $q_{Y|x}(\alpha; \theta)$ be the α quantile of $\mathbf{Y} | \mathbf{X} = x$. Then, any intervals $[q_{Y|x}(\alpha_1; \theta), q_{Y|x}(1 - \alpha_2; \theta)]$, where $\alpha_1 + \alpha_2 = \alpha$, have a probability $1 - \alpha$ of containing $\mathbf{Y}$. To obtain a prediction interval for $\mathbf{Y}$, a naive method frequently applied uses the values drawn from $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$ to get an estimate of θ and substitutes it in $q_{Y|x}(\alpha; \theta)$. This interval

$[q_{Y|x}(\alpha_1;\hat{\theta}(x)), q_{Y|x}(1-\alpha_2;\hat{\theta}(x))]$ is often called a plug-in prediction interval. This interval is not an exact $(1-\alpha)$ prediction interval, but provided that $\hat{\theta}(\mathbf{X})$ has good asymptotic properties, it will be an approximate $(1-\alpha)$ prediction interval when n is sufficiently large. For example, when $\hat{\theta}(x)$ is the maximum likelihood estimate (m.l.e.), we know that $[q_{Y|x}(\alpha_1;\hat{\theta}(\mathbf{X})), q_{Y|x}(1-\alpha_2;\hat{\theta}(\mathbf{X}))]$ will tend towards $[q_{Y|x}(\alpha_1;\theta), q_{Y|x}(1-\alpha_2;\theta)]$ as n increases because the m.l.e. converges towards θ almost surely under some regularity conditions. Note that most of the time, m.l.e.'s are the estimates used to obtain plug-in prediction intervals.

Plug-in prediction intervals are too narrow for most problems because they ignore the uncertainty in $\hat{\theta}(\mathbf{X})$ relative to θ. Therefore, these intervals usually have a coverage probability smaller than $1-\alpha$. This problem is more important for small or moderate values of n, since as n increases, this problem will eventually disappear. This deficiency of the plug-in intervals can be partially solved by calibrating the intervals, a technique we will discuss in the next chapter.

1.2.4 Predictive density and likelihood method

When the density function of $\mathbf{Y}|x$ is fully known (*i.e.* parameter free), it is straightforward to find exact prediction intervals for $\mathbf{Y}$. Thus, a possible approach to find approximate prediction intervals is to estimate this density. Such a function, denoted by $\tilde{f}_p(y|x)$, uses the already observed x in order to be as close as possible to $f(y|x;\theta)$. Approximate prediction intervals are then obtained by finding an interval $[a,b]$ such that

$$\int_a^b \tilde{f}_p(y|x)dy \;=\; 1-\alpha.$$

The plug-in method can also be seen as a method using an estimated predictive density where $\tilde{f}_p(y|x) = f(y|x;\hat{\theta}(x))$.

A possible approximate predictive density, presented in Lejeune & Faulkenberry (1982), is called the maximum likelihood predictive density (or MLPD). It is similar to the one inspired by the plug-in method but uses $\hat{\theta}(x, y)$ instead of $\hat{\theta}(x)$ to estimate θ. This density is given by

$$\tilde{f}_p(y|x) = k(x) \sup_{\theta} f(x, y; \theta),$$

where $k(x)$ is a normalizing constant. This approach is related to the argument followed in Kalbfleisch & Sprott (1970) for the marginalization problem in multi-parameter likelihood. The nuisance parameter here is θ, and the "parameter" of interest is now the random variable $\mathbf{Y}$. Predictive densities will be studied in greater detail in the next chapter.

Finally, there is a method which provides a likelihood function for $\mathbf{Y}$ based on the observed value of x. A predictive density can then be derived from this predictive likelihood to find approximate prediction intervals. This method was developed independently by both Lauritzen and Hinkley (*cf.* Lauritzen (1974) and Hinkley (1979)). However, this method requires the existence of a sufficient statistics for $(\mathbf{X}, \mathbf{Y})$ and gives trivial results unless this statistic provides a genuine reduction of $(\mathbf{X}, \mathbf{Y})$.

Most of these methods are not easy to handle, and they all have theoretical shortcomings in some settings. The Bayesian approach below also provides predictive densities which converge to $f(y|x; \theta)$ as the number of $\mathbf{X}_i$'s increases.

1.3 Bayesian Framework

While a frequentist framework specifies a model for the observable data given an unknown but fixed parameter θ, a Bayesian framework treats this θ as a random quantity. We will see in this section how prediction problems are relatively straightforward under this framework. We will first define important prediction concepts inherent to this framework

like the prior, posterior, and predictive distributions. Then, we will see how to find point and set prediction from these distributions. Finally, we will present the non-informative and empirical Bayes approaches.

1.3.1 Prior and Posterior Distributions

To explicitly incorporate the uncertainty about θ, we assign plausibilities on the various possible values of θ through a density function denoted by $\pi(\theta)$. This density is often called the prior density function of θ. The introduction of a prior density on θ suggests that this parameter is a random variable. However, we rarely assume that θ has such a random interpretation. We usually assume that θ is an unknown constant associated with a prior density through subjective probability. The combination of this prior with $f(x \mid \theta)$ provides us a density for $\mathbf{X}$ that is independent of θ and usually completely specified. This marginal density function of $\mathbf{X}$ is denoted by

$$m(x) \quad = \quad \int_{\theta} f(x \mid \theta)\pi(\theta)d\theta.$$

We will see how a Bayesian approach usually simplifies the prediction problems. However, a difficulty introduced with this approach is that we now have to select a prior density function. This choice must be a nice judgment between mathematical tractability and an appropriate plausibility assessment for θ.

Although the incorporation of a prior distribution for θ gives us a marginal distribution on $\mathbf{X}$, we would like to point out that such an approach is different than modeling $\mathbf{X}$ directly through the density function $m(x)$, since the latter does not allow us to update the plausibility of a certain θ once we observe x. This updating is done using Bayes' theorem:

$$\pi(\theta \mid x) \quad = \quad \frac{f(x \mid \theta)\pi(\theta)}{m(x)}.$$

This density is called the posterior density function of θ given x. It means that our prior belief about the plausibility of a value of θ is re-evaluated once we observe that $\mathbf{X} = x$.

1.3.2 Bayesian Predictive Density

We can now make plausibility assessments about the unknown θ once we observe x, but our goal is to make the same kind of assessments about the future value of $\mathbf{Y}$. The essence of the Bayesian approach to the prediction problem is to determine a density function for y given the outcome x. Note that such a function is often used in credibility theory (Herzog 1999). It can be seen as the average density of $\mathbf{Y}$ weighted by the updated plausibilities of the possible values of θ given x. We then use this predictive density to obtain predictors and prediction intervals.

Definition 1.5. *Let* $\mathbf{Y}$ *be a random variable with density function* $f(y \mid \theta)$. *Given a density* $f(x \mid \theta)$ *for* $\mathbf{X}$ *and a prior density* $\pi(\theta)$ *for* θ, *the* ***Bayesian predictive density function*** *for* $\mathbf{Y}$ *given that* $\mathbf{X} = x$ *is defined by*

$$\begin{aligned} \tilde{f}_p(y|x) &= \int_\theta f(y \mid x, \theta)\pi(\theta \mid x)d\theta \\ &= \frac{\int_\theta f(y \mid x, \theta)f(x \mid \theta)\pi(\theta)d\theta}{\int_\theta f(x \mid \theta)\pi(\theta)d\theta}. \end{aligned}$$

This is a density for $\mathbf{Y}$ given x but not the value of θ. This is one of the main reasons why a Bayesian approach is often used in prediction: by integrating over θ, we eliminate a parameter that is in fact a nuisance parameter. There are obvious similarities between frequentist and Bayesian predictive density functions. For example, it is shown in Lejeune & Faulkenberry (1982) that the MLPD can be identical to a predictive density using an appropriate prior.

1.3.3 Point and Set Predictions

In a Bayesian setting, the way we will choose a point predictor for the outcome of $\mathbf{Y}$ will depend on how we view the consequences of being wrong. This is assessed by specifying a loss function, $L[\widehat{\mathbf{Y}}(x), y]$, which assigns a non-negative loss of predicting y with $\widehat{\mathbf{Y}}(x)$. Given that $\mathbf{X} = x$, the goal will be to select the point predictor minimizing the Bayes expected loss

$$\mathbb{L}[\widehat{\mathbf{Y}}(x)] \;\; = \mathbb{E}_{\mathbf{Y}}[L[\widehat{\mathbf{Y}}(x), \mathbf{Y}]] \;\; = \int_Y L(\widehat{\mathbf{Y}}(x), y)\tilde{f}_p(y|x)dy.$$

A loss function frequently used is the quadratic loss $L[\widehat{\mathbf{Y}}(x), y] = \Big(\widehat{\mathbf{Y}}(x) - y\Big)^2$. The Bayes expected loss is then

$$\begin{aligned}
\mathbb{L}[\widehat{\mathbf{Y}}(x)] &= \int_Y (\widehat{\mathbf{Y}}(x) - y)^2 \tilde{f}_p(y|x)dy \\
&= \int_Y (\widehat{\mathbf{Y}}(x) - \mathbb{E}[\mathbf{Y} \mid x] + \mathbb{E}[\mathbf{Y} \mid x] - y)^2 \tilde{f}_p(y|x)dy \\
&= \int_Y (\widehat{\mathbf{Y}}(x) - \mathbb{E}[\mathbf{Y} \mid x])^2 \tilde{f}_p(y|x)dy + \int_Y (y - \mathbb{E}[\mathbf{Y} \mid x])^2 \tilde{f}_p(y|x)dy \\
&= \Big(\widehat{\mathbf{Y}}(x) - \mathbb{E}[\mathbf{Y} \mid x]\Big)^2 + \mathbb{V}ar[\mathbf{Y} \mid x].
\end{aligned}$$

The Bayes expected loss is then minimized when $\widehat{\mathbf{Y}}(x) = \mathbb{E}[\mathbf{Y} \mid x]$. This predictor is also unbiased under this Bayesian setting:

$$\begin{aligned}
\mathbb{E}[\widehat{\mathbf{Y}}(\mathbf{X}) - \mathbf{Y}] &= \mathbb{E}\left[\mathbb{E}[\mathbf{Y} \mid \mathbf{X}]\right] - \mathbb{E}[\mathbf{Y}] \\
&= \mathbb{E}[\mathbf{Y}] - \mathbb{E}[\mathbf{Y}] \\
&= 0.
\end{aligned}$$

Two other loss functions are also common: the all-or-nothing loss and the linear loss. The following propositions give us the optimal predictor for both losses; the proofs are presented in Aitchison & Dunsmore (1975, Chapter 3.1).

Proposition 1.2. *Let the all-or-nothing loss function be the limiting case of*

$$L[\widehat{\mathbf{Y}}(x), y] = \begin{cases} 0 & \text{if } |\widehat{\mathbf{Y}}(x) - y| < \epsilon \\ 1 & \text{otherwise,} \end{cases}$$

as $\epsilon \to 0$. *The predictor minimizing the Bayes expected loss is the mode of the predictive density function* $\tilde{f}_p(y|x)$.

Proposition 1.3. *Let the linear loss function be given by*

$$L[\widehat{\mathbf{Y}}(x), y] = \begin{cases} \alpha(y - \widehat{\mathbf{Y}}(x)) & \text{if } \widehat{\mathbf{Y}}(x) < y, \\ \beta(\widehat{\mathbf{Y}}(x) - y) & \text{otherwise.} \end{cases}$$

The predictor minimizing the Bayes expected loss is the $\alpha/(\beta+\alpha)$ *quantile of the predictive density function* $\tilde{f}_p(y|x)$.

When a prediction interval is sought, it is easier to find one once we incorporate a prior distribution on θ, the main difficulty under a frequentist approach being that the coverage probability usually depends on θ. Let $q_{Y|x}(\alpha)$ be the α quantile of the predictive density function $\tilde{f}_p(y|x)$. If $\alpha_1+\alpha_2 = \alpha$, we call $[q_{Y|x}(\alpha_1), q_{Y|x}(1-\alpha_2)]$ a Bayesian $(1-\alpha)$ prediction interval for $\mathbf{Y}$ since

$$\begin{aligned} P[q_{Y|\mathbf{X}}(\alpha_1) \le \mathbf{Y} \le q_{Y|\mathbf{X}}(1-\alpha_2)] &= \int_{\mathbf{X}} \left(\int_{q_{Y|x}(\alpha_1)}^{q_{Y|x}(1-\alpha_2)} \tilde{f}_p(y|x) dy \right) m(x) dx \\ &= \int_{\mathbf{X}} [1-\alpha] m(x) dx \\ &= 1 - \alpha. \end{aligned}$$

When θ is believed to be a random variable, a Bayesian $(1-\alpha)$ prediction interval have a clear interpretation: under repeated sampling of θ, $\mathbf{X}$, and $\mathbf{Y}$, this interval will contain y in $(1-\alpha)$ of the samples. The intervals also have a conditional coverage probability of

$1-\alpha$ for any observed x. However, these prediction intervals do not have a frequentist interpretation when a prior distribution is used to represent the state of knowledge of a fixed θ. Then, a Bayesian prediction interval is not an exact prediction interval in the sense of Definition 1.3.

Clearly, $(1-\alpha)$ prediction intervals for a given predictive density function are not unique. It is then of interest to find the shortest $(1-\alpha)$ prediction interval. This interval is given by the set $\mathcal{S}_k$ of all y such that $\tilde{f}_p(y|x) > k$ where k has to satisfy

$$\int_{y \in \mathcal{S}_k} \tilde{f}_p(y|x)dy \;\; = \;\; 1-\alpha.$$

1.3.4 Non-informative prior

We already mentioned that a prior density is chosen either for its mathematical convenience or its ability to represent the plausibilities of θ. A prior motivated by a realistic argument is often constructed from expert's knowledge (Campodónico & Singpurwalla 1995, for example), but if that type of information is not available, a prior is usually selected because of its mathematical convenience. In that case, we may try to find a prior which contains little or no information about θ. Such a prior aims not to favor a value of θ over another. These priors are usually called non-informative priors in the literature but terms like vague or diffuse priors could be more appropriate. This approach being relatively objective, it may be more attractive than the usual subjective Bayesian approach.

Since we do not want to favor a value of θ over another, a logical prior could be the flat prior $\pi(\theta) = c$ for all θ. Unfortunately, this distribution is often improper, in that $\int_\theta \pi(\theta)d\theta = \infty$, and then does not seem adequate to be used as a prior density. However, if this prior leads to a proper posterior for $\theta \mid x$, we can still use it to find a predictive density function. Another problem with this type of non-informative prior is its variability under reparameterization. Usually, a flat prior on θ does not lead to a flat prior on a function of

θ, which does not seem to be in accordance with a non-informativeness principle.

Another type of non-informative prior is called Jeffreys' prior (Jeffreys 1961, p. 181). This prior is usually easy to compute and is invariant under transformation. Jeffreys' prior is given by

$$\pi(\theta) \;=\; \sqrt{\det \mathcal{I}(\theta)}, \tag{1.2}$$

where $\mathcal{I}(\theta)$ is the expected Fisher information matrix with components,

$$\mathcal{I}_{ij}(\theta) \;=\; -\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\mathbf{X} \mid \theta)\right]$$

for $i, j = 1, \ldots, \dim(\theta)$. This type of prior has an interesting frequentist property: for prediction problems where $\mathbf{X}$ is independent of $\mathbf{Y}$, Datta, Mukerjee, Ghosh & Sweeting (2000) showed that if there is a one-sided Bayesian $(1-\alpha)$ prediction interval having a frequentist coverage probability of $1 - \alpha + o(n^{-1})$, then the prior used was necessarily Jeffreys' prior.

1.3.5 Empirical Bayes Methods

Another type of prior distribution can also provide posteriors where the information is coming essentially from the data. They are the priors chosen using an empirical Bayes approach. This approach was first introduced in Robbins (1955). We will briefly present here the parametric approach. A complete description of non-parametric empirical Bayes methods can be found in Carlin & Louis (1996, Section 3.2).

A parametric empirical Bayes approach uses a family of prior distributions that is indexed by a vector parameter η. The marginal density function of $\mathbf{X}$, $m(x \mid \eta)$, is then used to find a point estimate $\hat{\eta}(x)$. Then, we treat $\pi(\theta \mid \hat{\eta}(x))$ as a completely specified prior distribution for θ. Due to the form of the marginal likelihood for parametric empirical

Bayes models, the EM algorithm (Dempster, Laird & Rubin 1977) can be effective to find these maximum likelihood estimates. In some other cases (Gaver & O'Muircheartaigh 1987, for example), estimates using a moment matching method are used.

Under this approach, a prediction interval for $\mathbf{Y}$ will be determined from the quantiles of the predictive density function

$$\tilde{f}_p(y \mid x; \hat{\eta}) \quad = \quad \int_\theta f(y \mid x, \theta)\pi(\theta \mid x; \hat{\eta})d\theta.$$

Obviously, the problems we already encountered with the plug-in prediction intervals will reappear here, since these new intervals also ignore the uncertainty about η. Carlin & Louis (1996, Section 3.5) describe some approaches to correct this. In most of the cases, we can easily associate them, from a mathematical point of view, with frequentist methods discussed previously to deal with the dependency of the coverage probability on θ. Nevertheless, it is expected that the ignorance of η will not affect the coverage probability to the same extent as the ignorance of θ does.

1.4 Thesis Outline

In this thesis, we will study issues related to prediction problems and put an emphasis on those arising when recurrent events are involved. We will mostly consider problems where interval predictions are sought but we will also address some issues regarding point prediction. Amongst others, here are two motivational examples:

- In a carcinogenicity experiment, a group of individuals are observed during a certain amount of time and the number of tumors developed by each individual is recorded. During such a study, is of interest for the research team to predict the remaining number of tumors for one or some of the individuals in the study group. Such an example will be considered in Chapter 4.

- Information about a fleet of automobiles under warranty is collected over the years. For planning purposes, it is of interest for the company to predict, at any given time, the remaining number of warranty claims until all the warranties expire. Such a dataset will be studied in Chapter 5.

In the second chapter, we will study frequentist prediction intervals and associated predictive distributions. Pivotal methods for obtaining intervals and predictive distributions are discussed and shown to possess advantages that include a type of optimality. The following three chapters will consider the prediction of recurrent events. In the third chapter, we will present different prediction models when these events can be modeled using homogeneous Poisson processes. Amongst these models, we will see that those using random effects are robust to different types of model misspecifications. In the fourth chapter, the time homogeneity assumption is relaxed and we will present prediction models for non-homogeneous Poisson processes. The behavior of these models will then be studied for prediction problems with a finite horizon. In the fifth chapter, we will apply the concepts discussed in the previous chapters to a warranty dataset coming from the automobile industry. The number of processes in this dataset being very large, we will propose some methods providing computationally rapid prediction intervals. Finally, we will discuss our future research in the last chapter.

Chapter 2

Predictive Distributions and Calibration

The purpose of this chapter is to discuss the usefulness of predictive distributions and to assess the adequacy of their associated prediction intervals. In particular, we will present a predictive distribution, derived through a process called calibration, that can be obtained in many settings and has several nice properties.

In the first section, we will discuss how different prediction approaches can be derived from predictive distributions and propose a criterion to compare them. In the ensuing section, we will present calibration approaches, some obtained with asymptotic expansions and others with bootstrap techniques. In the third section, we will describe an approach which, in terms of average Kullback-Leibler distance, dominates the plug-in approach under certain conditions. We will illustrate the methodology in the last section.

2.1 Predictive Distributions

Bayesian predictive distributions were presented in Section 1.3.2 and we briefly mentioned some frequentist ones in Section 1.2.4. In order to unify both paradigms, we will now loosely define a predictive distribution as any distribution independent of θ constructed from the observed data x in order to make predictive statements about the random variable $\mathbf{Y}$. Amongst others, such a definition was used in Harris (1989) and Barndorff-Nielsen & Cox (1996). Its cumulative distribution function (c.d.f.) will be denoted by $\tilde{F}_p(y|x)$ and its density by $\tilde{f}_p(y|x)$.

Without loss of generality, we will now consider one-sided prediction intervals of the form $(-\infty, L(\mathbf{X})]$ and simply denote by $L_\alpha(\mathbf{X})$ an exact (or approximate) α prediction interval. We can use a predictive density to obtain such an interval by finding the α quantile of its distribution, *i.e.*

$$\tilde{F}_p(L_\alpha(x)|x) \quad = \quad \alpha.$$

Therefore, plotting $\tilde{F}_p(y|x)$ along the range of $\mathbf{Y}|x$ will provide the realized (exact or approximate) α prediction intervals for all α $(0 < \alpha < 1)$.

For example, the predictive density

$$\tilde{f}_p(y|x) \quad = \quad f(y|x; \hat{\theta}(x)), \tag{2.1}$$

can be used to find the plug-in prediction intervals presented in Section 1.2.3, the density

$$\tilde{f}_p(y|x) \quad = \quad \int_\theta f(y|x,\theta)\pi(\theta|x)d\theta, \tag{2.2}$$

can be used to find the Bayesian intervals mentioned in Section 1.3.3, and when $Z(\mathbf{X}, \mathbf{Y})$ has a distribution that is independent of θ, the pivotal method presented in Section 1.2.1 has the predictive c.d.f.

$$\tilde{F}_p(y|x) \quad = \quad P[Z(\mathbf{X}, \mathbf{Y}) \leq Z(x, y)].$$

When a prediction interval is sought, it is not easy to compare the adequacy of different predictive methods. The only suggestion made so far was to select, among a class of methods providing exactly the same coverage probability, the method giving the shortest intervals (see Section 1.1.1). However, even when a such an idea can be applied, it can give different results for any θ and any α. Nevertheless, since the vast majority of predictive methods can be expressed through a predictive distribution, we can compare them using these distributions. This comparison can be done by evaluating how close each density is from the target density $f(y|x;\theta)$. A very common approach (e.g. Aitchison (1975), Murray (1977), Ng (1980), Harris (1989), Vidoni (1995), and Komaki (1996)) has been to assess this relative closeness with the average Kullback-Leibler (KL) distance (Kullback & Leibler 1951)):

$$D[\tilde{f}_p(y|x), f(y|x;\theta)] \;=\; \mathbb{E}\left[\log\left(\frac{f(\mathbf{Y}|\mathbf{X};\theta)}{\tilde{f}_p(\mathbf{Y}|\mathbf{X})}\right)\right]. \tag{2.3}$$

Provided that we can obtain (or approximate) this distance, we can now compare different methods and select the best one. This method can give different result for different values of θ but no longer depends on the coverage probability α chosen. Furthermore, we will see in Section 2.3 that we can find a method that is optimal, according to this criterion, amongst a large class of predictive densities. Note that other criteria can be used, for example Smith (1999) considered the mean squared error $\mathbb{E}[(\tilde{f}_p(\mathbf{Y}|\mathbf{X}) - f(\mathbf{Y}|\mathbf{X};\theta))^2]$.

Even if (2.3) is a function of the unknown θ, we can sometimes show that a method dominates another for all possible values of θ. Examples of this were presented in Aitchison (1975): assuming that $\mathbf{X}$ and $\mathbf{Y}$ are both gamma random variables with a common known shape parameter, or both multi-normal with the same mean and variance, we can always find a prior distribution such that the Bayesian predictive density (2.2) will always dominate the plug-in density (2.1) according to (2.3). Since the plug-in method ignores

the uncertainty in $\hat{\theta}(\mathbf{X})$ relative to θ, the inadequacy of this method has to be expected for some settings. However, we will present in the next section a procedure which usually helps to "move" a predictive density closer to the target distribution. Such a procedure, usually called calibration, is very often used in conjunction with the plug-in method.

2.2 Calibration

A family of prediction intervals $L_\alpha(\mathbf{X})$ indexed by α is obtained from a predictive c.d.f. such that $\tilde{F}_p(L_\alpha(x)|x) = \alpha$ for any α. Therefore, α does not necessarily represent the coverage probability of the prediction interval: we would like to obtain intervals with a frequentist coverage probability of α, or a Bayesian posterior probability of α, but a lot of predictive distributions do not satisfy this. Let $H(\alpha;\theta)$ be the actual coverage probability of $L_\alpha(x)$, *i.e.*

$$P[\mathbf{Y} \leq L_\alpha(\mathbf{X});\theta] \;=\; H(\alpha;\theta).$$

When the prediction intervals are associated with an exact (frequentist) prediction method, $H(\alpha;\theta) = \alpha$ for any α and θ, but usually $H(\alpha;\theta)$ converges to α as the information about θ contained in $\mathbf{X}$ increases. In practice, since θ is unknown, $H(\alpha;\theta)$ is usually approximated by a function $\tilde{H}(\alpha)$ of α only. This procedure is often called calibration and provides an approximate coverage probability of $\tilde{H}(\alpha)$ for the prediction interval $(-\infty, L_\alpha(x)]$.

Calibration can be done using asymptotic expansions or simulations. It is often used in conjunction with the plug-in method. For example, Cox (1975) and Barndorff-Nielsen & Cox (1996) considered the plug-in approach of Section 1.2.3 and showed that under some regularity conditions the coverage probability can be written as

$$H(\alpha;\theta) \;=\; \alpha + \frac{d(\theta)}{n} + O(n^{-3/2}), \tag{2.4}$$

where $n = \dim(\mathbf{X})$. The idea is then to find a new interval, often in the same family of prediction intervals (say $L_{\alpha'}(x)$), that would absorb (or reduce) the $d(\theta)/n$ term in (2.4). Vidoni (1995) and Komaki (1996) also used asymptotic expansions to improve the plug-in method.

The idea of approximating $H(\alpha;\theta)$ by simulation is usually more tractable (e.g. Harris (1989), Beran (1990), and Meeker & Escobar (1998, Section 12.6)). For illustrative purposes, we present here an algorithm based on parametric bootstrapping presented in Meeker & Escobar (1998, Section 12.6). The idea is to obtain a value α' such that the approximated coverage probability $\tilde{H}(\alpha')$ equals α. Note that we modified the original algorithm to also consider the case where $\mathbf{X}$ and $\mathbf{Y}$ are not independent:

1. Choose a value of α, say α_0 (preferably greater than α).
2. Simulate $x_1^*, \ldots, x_B^*$ from $f(x^*;\hat{\theta}(x))$ where $\hat{\theta}(x)$ is the m.l.e. based on the observed sample x.
3. Simulate $y_1^*, \ldots, y_B^*$ from the conditional densities $f(y|x_i^*;\hat{\theta}(x))$.
4. Obtain the m.l.e.'s $\hat{\theta}(x_i^*)$ where $i = 1, \ldots, B$.
5. Compute the B plug-in α_0 prediction intervals based on $(x_i^*, \hat{\theta}(x_i^*))$.
6. Approximate the coverage probability function at α_0 by:

$$\tilde{H}(\alpha_0) \;=\; \frac{1}{B}\sum_{i=1}^{B} \mathbb{I}(y_i^* \leq L_{\alpha_0}(x_i^*)).$$

7. Repeat steps 2-6 for different values of α_0 until an α' such that $\tilde{H}(\alpha') = \alpha$ is found.

Note that when the conditional c.d.f. of $\mathbf{Y}$ is easily obtained, we can eliminate the Monte Carlo error due to the sampling of the y^*'s. We forego the third step in this algorithm and

approximate the coverage probability function by:

$$\tilde{H}(\alpha_0) \;=\; \frac{1}{B}\sum_{i=1}^{B} F_{Y|x_i^*}(L_{\alpha_0}(x_i^*);\hat{\theta}(x_i^*)).$$

Performing such an algorithm can require a very substantial amount of computational time, especially when the simulated datasets are large or if a numerical optimization procedure is required to obtain the m.l.e.'s. However, we can significantly reduce the computational time by modifying the 7th step: we can keep the x_i^*'s, $\hat{\theta}(x_i^*)$'s, and y_i^*'s obtained the first time steps 2-4 were performed and repeat only steps 5 and 6 for different values of α_0. This modification can increase the Monte Carlo error but will not introduce any bias.

This algorithm gives us a single prediction interval with an approximate coverage probability of α. In the next section, we will propose a prediction method which also uses a calibration procedure but now gives prediction intervals for all α simultaneously.

2.3 A Proposed Approach

We already discussed that prediction based on pivotal quantities leads to exact prediction intervals. Let us assume that the random variable $\mathbf{Y}|x$ is continuous for all x. It is well known that $F(\mathbf{Y}|\mathbf{X};\theta)$ is uniformly distributed on $[0,1]$ for any θ. Therefore, we are now proposing to base our prediction of $\mathbf{Y}$ on the quantity

$$\mathbf{U} \;=\; F(\mathbf{Y}|\mathbf{X};\hat{\theta}(\mathbf{X})). \tag{2.5}$$

This quantity can be exactly pivotal but is at least asymptotically pivotal in most settings; under some regularity conditions $\hat{\theta}(\mathbf{X})$ is consistent for θ and thus U is asymptotically uniform on $[0,1]$. In this section, we will discuss several nice properties of predictions based on (2.5).

Let $G(u;\theta)$ be the c.d.f. of $\mathbf{U}$ and $\tilde{G}(u)$ an approximation of $G(u;\theta)$ independent of θ such that

$$\tilde{G}(u) \;=\; \begin{cases} G(u;\theta) \text{ if } \mathbf{U} \text{ is a pivotal,} \\ G(u;\hat{\theta}(x)) \text{ otherwise.} \end{cases}$$

Now let u_α be the α quantile based on $\tilde{G}$ and $q_{\mathbf{Y}|x}(p;\theta)$ the p quantile of $\mathbf{Y}|x$. We have

$$\begin{aligned} \tilde{G}(u_\alpha) &= P[\mathbf{U} \le u_\alpha; \hat{\theta}(x)] \\ &= \int_{X^*} P[F(\mathbf{Y}|x^*;\hat{\theta}(x^*)) \le u_\alpha; \hat{\theta}(x)] f(x^*;\hat{\theta}(x)) dx^* \\ &= \int_{X^*} P[F(\mathbf{Y}|x^*;\hat{\theta}(x^*)) \le F_{\mathbf{Y}|x^*}(q_{\mathbf{Y}|x^*}(u_\alpha;\hat{\theta}(x^*));\hat{\theta}(x^*));\hat{\theta}(x)] f(x^*;\hat{\theta}(x)) dx^* \\ &= \int_{X^*} P[\mathbf{Y} \le q_{\mathbf{Y}|x^*}(u_\alpha;\hat{\theta}(x^*))|x^*;\hat{\theta}(x)] f(x^*;\hat{\theta}(x)) dx^* \\ &= P[\mathbf{Y} \le q_{\mathbf{Y}|\mathbf{X}}(u_\alpha;\hat{\theta}(\mathbf{X}));\hat{\theta}(x)]. \end{aligned} \tag{2.6}$$

Therefore, an α prediction interval is easily obtained and has the plug-in form

$$L_\alpha(x) \;=\; q_{\mathbf{Y}|x}(u_\alpha;\hat{\theta}(x)).$$

The associated predictive c.d.f. is given by

$$\tilde{F}_p(y|x) \;=\; \tilde{G}(F(y|x;\hat{\theta}(x)))$$

and for any α

$$\begin{aligned} \tilde{F}_p(L_\alpha(x)|x) &= \tilde{G}(F_{\mathbf{Y}|x}(q_{\mathbf{Y}|x}(u_\alpha;\hat{\theta}(x));\hat{\theta}(x))) \\ &= \alpha. \end{aligned}$$

We can also see from (2.6) that this approach yields exact prediction intervals for all α only when $\tilde{G}(u) = G(u;\theta)$ for all θ, *i.e.* when U is exactly pivotal.

Another nice feature of predictions based on (2.5) is that the coverage probability associated with $L_\alpha(x)$ is simply the c.d.f. of U:

$$\begin{aligned} H(\alpha;\theta) &= P[\mathbf{Y} \leq q_{\mathbf{Y}|\mathbf{X}}(u_\alpha;\hat{\theta}(\mathbf{X}));\theta] \\ &= P[\mathbf{U} \leq u_\alpha;\theta] \\ &= G(u_\alpha;\theta). \end{aligned}$$

Therefore the calibration procedure is done by approximating the c.d.f. of $\mathbf{U}$ by $\tilde{G}$.

We now present an algorithm to perform such a calibration. We note that unlike the algorithm presented in the previous section, this one provides calibrated prediction intervals simultaneously for all α.

1. Simulate $x_1^*, \ldots, x_B^*$ from $f(x^*;\hat{\theta}(x))$ where $\hat{\theta}(x)$ is the m.l.e. based on the observed sample x.

2. Simulate $y_1^*, \ldots, y_B^*$ from the conditional densities $f(y|x_i^*;\hat{\theta}(x))$.

3. Obtain the m.l.e.'s $\hat{\theta}(x_i^*)$ where $i = 1, \ldots, B$.

4. Obtain $u_i^* = F(y_i^*|x_i^*, \hat{\theta}(x_i^*))$, so that the vector $(u_1^*, \ldots, u_B^*)$ is then a random sample of $\mathbf{U}$ based on $\theta = \hat{\theta}(x)$.

5. Approximate $\tilde{G}$ with the empirical c.d.f.

$$\tilde{G}(u) \simeq \frac{1}{B}\sum_{i=1}^{B} \mathbb{I}(u_i^* \leq u). \tag{2.7}$$

6. For any α, $u_{[B\alpha]} \simeq u_\alpha$, and an approximate α prediction interval for $\mathbf{Y}$ is given by $q_{\mathbf{Y}|x}(u_{[B\alpha]};\hat{\theta}(x))$.

The Monte Carlo error due to the sampling variability of the y^*'s can be eliminated if we can replace (2.7) with

$$\tilde{G}(u) \simeq \frac{1}{B}\sum_{i=1}^{B} P[F(Y|x_i^*;\hat{\theta}(x_i^*)) \leq u;\hat{\theta}(x)].$$

If one would choose to approximate the distribution of $\mathbf{U}$ with its $U(0,1)$ asymptotic distribution, the associated predictive c.d.f. would then be

$$\begin{aligned}\tilde{F}_p(y|x) &= \tilde{G}(F(y|x;\hat{\theta}(x))) \\ &= F(y|x;\hat{\theta}(x)),\end{aligned}$$

the predictive c.d.f. of the plug-in approach. Therefore, interesting features are revealed when we plot $\tilde{G}(u)$ for all $u \in [0,1]$ versus the c.d.f. of a $U(0,1)$: for any α, the difference $\alpha - \tilde{G}(\alpha)$ indicates the approximated loss in terms of coverage probability when the plug-in method is used. Also, such a plot can serve as a diagnostic tool to indicate whether the calibration was necessary. This can be especially useful when predictions are made at different times during a longitudinal study. As soon as we observe no significant differences between $\tilde{G}(u)$ and u, it appears reasonable to use the simple plug-in method for all subsequent predictions.

Another nice property is that when $F(\mathbf{Y}|\mathbf{X};\hat{\theta}(\mathbf{X}))$ is exactly pivotal, we can show that its associated predictive density

$$\tilde{f}_p(y|x) = \tilde{g}(F(y|x;\hat{\theta}(x)))f(y|x;\hat{\theta}(x)), \tag{2.8}$$

where $\tilde{g}(u) = d\tilde{G}(u)/du$, dominates the plug-in predictive density with respect to (2.3).

Theorem 2.1. *(Lawless & Fredette 2004) When* $\mathbf{U} = F(\mathbf{Y}|\mathbf{X};\hat{\theta}(\mathbf{X}))$ *is pivotal, the predictive density (2.8) has an average Kullback-Leibler distance (2.3) at least as small as* $\hat{f}_p(y|x) = f(y|x,\hat{\theta}(x))$*, the predictive density of the plug-in approach.*

Proof. Let Δ be the difference between the two distances, and we have,

$$\begin{aligned}
\Delta &= D[\hat{f}_p(y|x), f(y|x;\theta)] - D[\tilde{f}_p(y|x), f(y|x;\theta)] \\
&= \mathbb{E}\left[\log\left(\frac{\tilde{f}_p(\mathbf{Y}|\mathbf{X})}{\hat{f}_p(\mathbf{Y}|\mathbf{X})}\right)\right] \\
&= \mathbb{E}\left[\log(\tilde{g}(F(\mathbf{Y}|\mathbf{X};\hat{\theta}(\mathbf{X}))))\right] \\
&= \mathbb{E}\left[\log(\tilde{g}(\mathbf{U}))\right] \\
&= \int_0^1 g(u;\theta)\log(g(u;\hat{\theta}(x)))du \\
&= \int_0^1 g(u;\theta)\log(g(u;\theta))du,
\end{aligned}$$

since $\mathbf{U}$ is a pivotal. But for any densities $v(u;\theta)$ on $[0,1]$,

$$\int_0^1 g(u;\theta)\log\left(\frac{g(u;\theta)}{v(u;\theta)}\right)du \;\geq\; 0,$$

for any θ (Kullback & Leibler 1951). Therefore, by taking $v(u;\theta) = 1$, we have $\Delta \geq 0$. □

When $\mathbf{U}$ is only asymptotically pivotal, some insight into the comparison of $\hat{f}_p$ and $\tilde{f}_p$ can be gained by writing

$$\begin{aligned}
\Delta(\theta) &= \int_0^1 g(u;\theta)\log(\tilde{g}(u))du \\
&= \int_0^1 g(u;\theta)\log\left(\frac{g(u;\theta)}{1}\right)du - \int_0^1 g(u;\theta)\log\left(\frac{g(u;\theta)}{\tilde{g}(u)}\right)du,
\end{aligned}$$

where both integrals are non-negative. The first integral is small when the density $g(u;\theta)$ is close to the uniform density on $[0,1]$ and the second is small when $\tilde{g}(u)$ closely approximates $g(u;\theta)$. The value of $\Delta(\theta)$ can be either positive or negative when $\mathbf{U}$ is not pivotal, but it

will be positive when $\tilde{g}(u)$ is closer to $g(u;\theta)$ than is the uniform density. We conjecture that this should be true in most settings.

In addition to the fact that it dominates the plug-in approach, it can be shown that the predictive density derived from a pivotal quantity can be optimal amongst a large class of predictive densities.

Theorem 2.2. *(Lawless & Fredette 2004) Let* $\mathbf{W} = q(\mathbf{X}, \mathbf{Y})$ *be a pivotal quantity with c.d.f.* $G(w)$*. Amongst all the predictive distributions that are functions of this pivotal, the predictive c.d.f.*

$$\tilde{F}_p(y|x) \;\; = \;\; G(q(x,y))$$

has the density giving the smallest average Kullback-Leibler distance (2.3).

Proof. First we note that the minimization of (2.3) is equivalent to the maximization of

$$J(\tilde{f}_p) = \mathbb{E}[\log(\tilde{f}_p(\mathbf{Y}|\mathbf{X}))].$$

Now let $\tilde{F}_p(y|x) = R(w)$ for any c.d.f. $R(w)$, thus $\tilde{f}_p(y|x) = r(w)|\partial w/\partial y|$ and

$$J(\tilde{f}_p) = \mathbb{E}[\log(r(\mathbf{W}))] + K(\theta),$$

where $K(\theta)$ is the same for any densities $r(w)$. Now since

$$\mathbb{E}[\log(r(\mathbf{W}))] = \int_W \log(r(w))g(w)dw,$$

we can see that $J(\tilde{f}_p)$ is maximized when $r(w) = g(w)$. □

By taking $r(w) = 1$ over $0 \leq w \leq 1$, we can see that the truth of Theorem 2.2 implies the truth of Theorem 2.1.

2.4 Illustrations

We will now illustrate the methodology with two simple examples. In the first one, (2.5) is exactly pivotal and $\tilde{G}(u)$ can be obtained analytically. In the second example, (2.5) is asymptotically pivotal and its distribution must be approximated. In both cases, $\mathbf{X}$ and $\mathbf{Y}$ are independent but we will consider settings involving dependence in Chapter 4 and 5.

2.4.1 Exponential distribution

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from the exponential distribution with c.d.f. $F(x;\theta) = 1-\exp\{-x/\theta\}$ and let $\mathbf{Y}$ be an independent future observation with the same distribution. The m.l.e. of θ based on $\mathbf{X}$ is $\hat{\theta}(x) = \overline{x}$ and (2.5) is

$$\mathbf{U} \;=\; 1 - \exp\{\mathbf{Y}/\hat{\theta}(\mathbf{X})\}$$

We can show that $\mathbf{Y}/\hat{\theta}(\mathbf{X})$ has a F distribution with $(2, 2n)$ degrees of freedom (Lawless 2003, Section 4.6) and so $\mathbf{U}$ is exactly pivotal for all n. The density function of $W = \mathbf{Y}/\hat{\theta}(X)$ being $(1 + w/n)^{-(n+1)}$, an easy calculation shows that $\mathbf{U}$ has c.d.f.

$$\tilde{G}(u) \;=\; 1 - (1 - \frac{1}{n}\log(1-u))^{-n},$$

which converges towards $G(u) = u$ as n goes to ∞. The α quantile of this distribution is given by

$$u_\alpha \;=\; 1 - \exp\{n[1 - (1-\alpha)^{\frac{-1}{n}}]\}$$

and a one-sided u_α plug-in prediction interval will have an exact coverage probability of α.

The associated predictive c.d.f. is given by

$$\tilde{F}_p(y|x) \;=\; \tilde{G}(F(y|x;\hat{\theta}(x)))$$

$$= \quad 1 - \left(1 + \frac{y}{n\hat{\theta}(x)}\right)^{-n}. \tag{2.9}$$

Any exact prediction interval can be obtained directly from this c.d.f. For example, a 95% equal-tailed prediction interval is given by the .025 and the .975 quantiles of (2.9). Approximate plug-in prediction intervals are obtained the same way by using the quantiles of the predictive c.d.f.

$$\hat{F}_p(y|x) \quad = \quad 1 - \exp\left\{\frac{-y}{\hat{\theta}(x)}\right\}. \tag{2.10}$$

We can see that (2.9) converges towards (2.10) as n goes to ∞. This means that when n is sufficiently large, the prediction intervals provided by the plug-in method will be similar to the exact ones.

The left panel of Figure 2.1 shows $\tilde{G}(u)$ for $n = 10$ and 30 and the right panel shows the associated predictive c.d.f. (2.9) along with the plug-in predictive c.d.f. (2.10), all with $\hat{\theta}(x)$ taken equal to 1. There is a small difference between the exact method and the plug-in method when $n = 10$ and they appear similar when $n = 30$. However, we must keep in mind that prediction intervals are typically associated with upper and lower tails of the distribution. Therefore, prediction intervals can differ substantially even if the c.d.f.'s appear to be similar. For example, we can see from the right panel of Figure 2.1 that the exact upper bound of a 99% one-sided prediction interval is 5.85 when $n = 10$ and 4.98 when $n = 30$ but the plug-in approach gives 4.61 in both cases.

Figure 2.2 shows $\alpha - \tilde{G}(\alpha)$ for $n = 10$ and $n = 30$. It represents the loss in coverage probability induced by using a one-sided α plug-in prediction interval instead of the exact one. Obviously, the loss is smaller with $n = 30$ but it is interesting to note that even if the plug-in prediction bound can differ substantially from the exact one when $n = 10$, the loss in coverage probability is never greater than 3%.

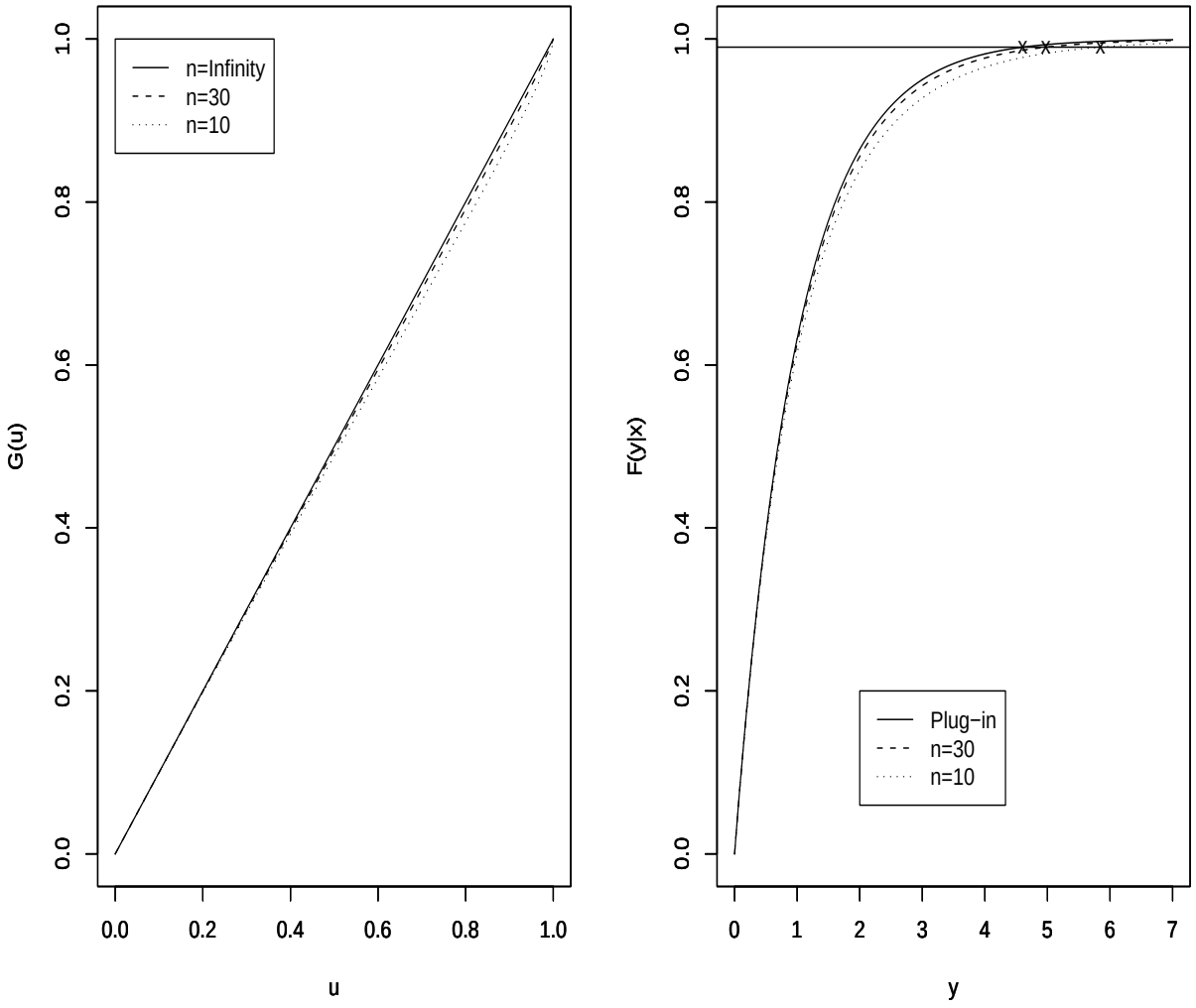

Figure 2.1: Functions $\tilde{G}(u)$ and $\tilde{F}_p(y|x)$, exponential distribution.

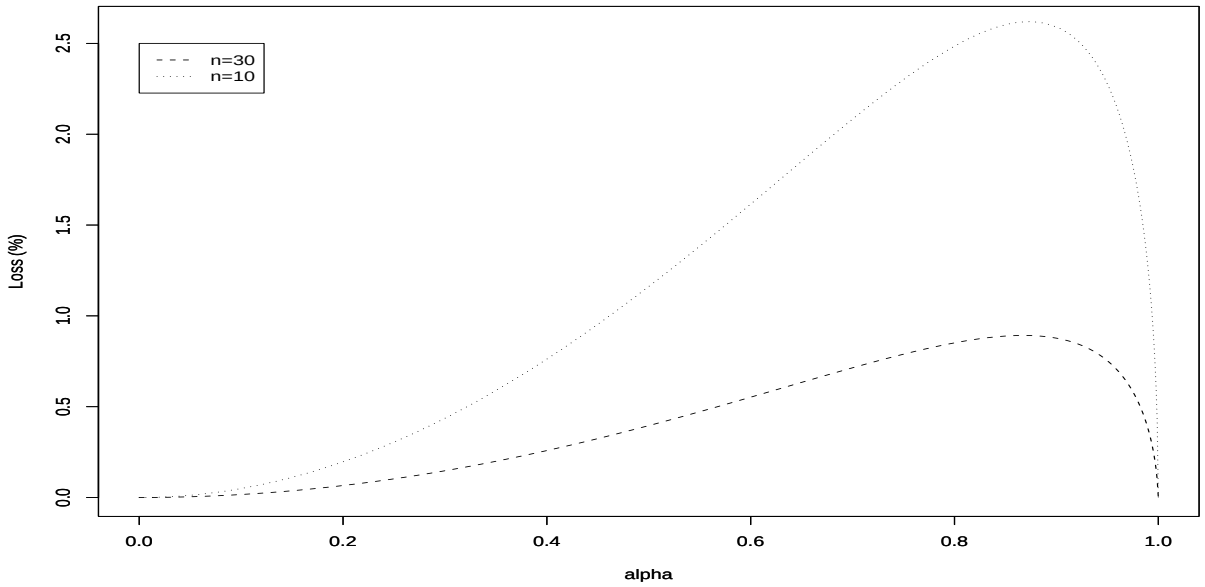

Figure 2.2: The loss in coverage probability induced by using the plug-in approach.

Under this setting, the average KL distance can be obtained analytically for both $\tilde{f}_p(y|x)$ and $\hat{f}_p(y|x)$. Using the fact that the expectation of a Fisher$(2, 2n)$ distribution is $n/(n-1)$ and that the expectation of the logarithm of a sum of n exponential with rate θ is $\psi(n) + \log(\theta)$, where $\psi(n) = \Gamma'(n)/\Gamma(n)$, we can show that the distance between $\hat{f}_p(y|x)$ and $f(y;\theta)$ is

$$\begin{aligned} D(\hat{f}_p(y|x), f(y|x)) &= \mathbb{E}\left[\log\left(\frac{e^{\frac{-\mathbf{Y}}{\theta}}/\theta}{e^{\frac{-\mathbf{Y}}{\hat{\theta}(\mathbf{X})}}/\hat{\theta}(\mathbf{X})}\right)\right] \\ &= \psi(n) + \frac{n}{n-1} - \log(n) - 1, \end{aligned}$$

and

$$\begin{aligned} D(\tilde{f}_p(y|x), f(y|x)) &= \mathbb{E}\left[\log\left(\frac{e^{\frac{-\mathbf{Y}}{\theta}}/\theta}{\frac{\partial}{\partial y}\tilde{F}_p(y|x)}\right)\right] \\ &= (n+1)\psi(n+1) - n\psi(n) - \log(n) - 1. \end{aligned}$$

Now since $\psi(n+1) = \psi(n) + 1/n$, we have

$$D(\hat{f}_p(y|x), f(y|x)) - D(\tilde{f}_p(y|x), f(y|x)) \;=\; \frac{1}{n(n-1)},$$

which converges towards 0 but, as stated by Theorem 2.1, this difference is always positive.

Finally, using the fact that $\psi(n+1) = log(n) + \mathcal{O}(n^{-1})$, we can show that the asymptotic behavior of the relative loss, with respect to the average KL distance, of using $\hat{f}_p$ instead of $\tilde{f}_p$ is

$$\begin{aligned}
\frac{D(\hat{f}_p(y|x), f(y|x)) - D(\tilde{f}_p(y|x), f(y|x))}{D(\tilde{f}_p(y|x), f(y|x))} &= \frac{1/[n(n-1)]}{\psi(n) + (n+1)/n - \log(n) - 1} \\
&= \frac{\mathcal{O}(n^{-2})}{\mathcal{O}(n^{-1})} \\
&= \mathcal{O}(n^{-1}).
\end{aligned}$$

2.4.2 Log-normal distribution

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample of log-normal random variables with parameters μ and σ and let $\mathbf{Y}$ be an independent future observation with the same distribution. We will obtain prediction intervals for $\mathbf{Y}$ based on a dataset presented in Lawless (1982, page 228). For our example, we will use the failure times of 15 bearings plus the censoring times for the 8 bearings who had not failed after 80 million cycles. Even though the log-normal distribution has a location-scale form for $\mathbf{Y}' = \log(\mathbf{Y})$, exact prediction intervals cannot be obtained here because of the time censoring. Therefore, we will predict $\mathbf{Y}$ based on the approximate pivotal

$$\begin{aligned}
U &= F(\mathbf{Y}; \hat{\mu}(\mathbf{X}), \hat{\sigma}(\mathbf{X})) \\
&= \Phi\left(\frac{\log(\mathbf{Y}) - \hat{\mu}(\mathbf{X})}{\hat{\sigma}(\mathbf{X})}\right),
\end{aligned}$$

where Φ is the c.d.f. of a standard normal distribution.

Based on the x observed, we approximated the c.d.f. of U with $\tilde{G}(u)$ using the algorithm presented in Section 2.3. The simulated x^*'s were also censored at 80 million cycles. The associated predictive c.d.f. is then given by

$$\tilde{F}_p(y|x) \;=\; \tilde{G}\left(\Phi\left(\frac{\log(y)-\hat{\mu}(x)}{\hat{\sigma}(x)}\right)\right).$$

The left panel of Figure 2.3 shows $\tilde{G}(u)$ and the right panel shows $\tilde{F}_p(y|x)$. We also added the functions $G(u) = u$ and $\hat{F}_p(y|x) = \Phi(\frac{\log(y)-\hat{\mu}(x)}{\hat{\sigma}(x)})$ on these panels. Again, we can see that $\tilde{G}(u)$ is very close to the c.d.f. of a $U(0,1)$ and the predictive c.d.f. is quite close to the plug-in c.d.f. Nevertheless, these two c.d.f's can provide quantiles that are substantially different. For example, a 95% prediction bound is 157.12 using the plug-in approach but 174.02 using an approximate pivotal. The latter quantile corresponds to the 96.7% prediction bound of the plug-in approach. Meeker & Escobar (1998, Section 12.6) obtained the same calibrated 95% prediction bound using the algorithm presented in Section 2.2. However, our proposed algorithm has the advantage of providing α prediction bounds for any $0 \leq \alpha \leq 1$ simultaneously with a single simulation.

In these two examples, the exact or approximated loss in coverage probability of using a plug-in approach is never greater than 2.5%. However, we will consider problems in Chapter 4 and 5 where the loss will be much greater. This is probably due to the fact that the number of unknown parameters will then be greater than one or two.

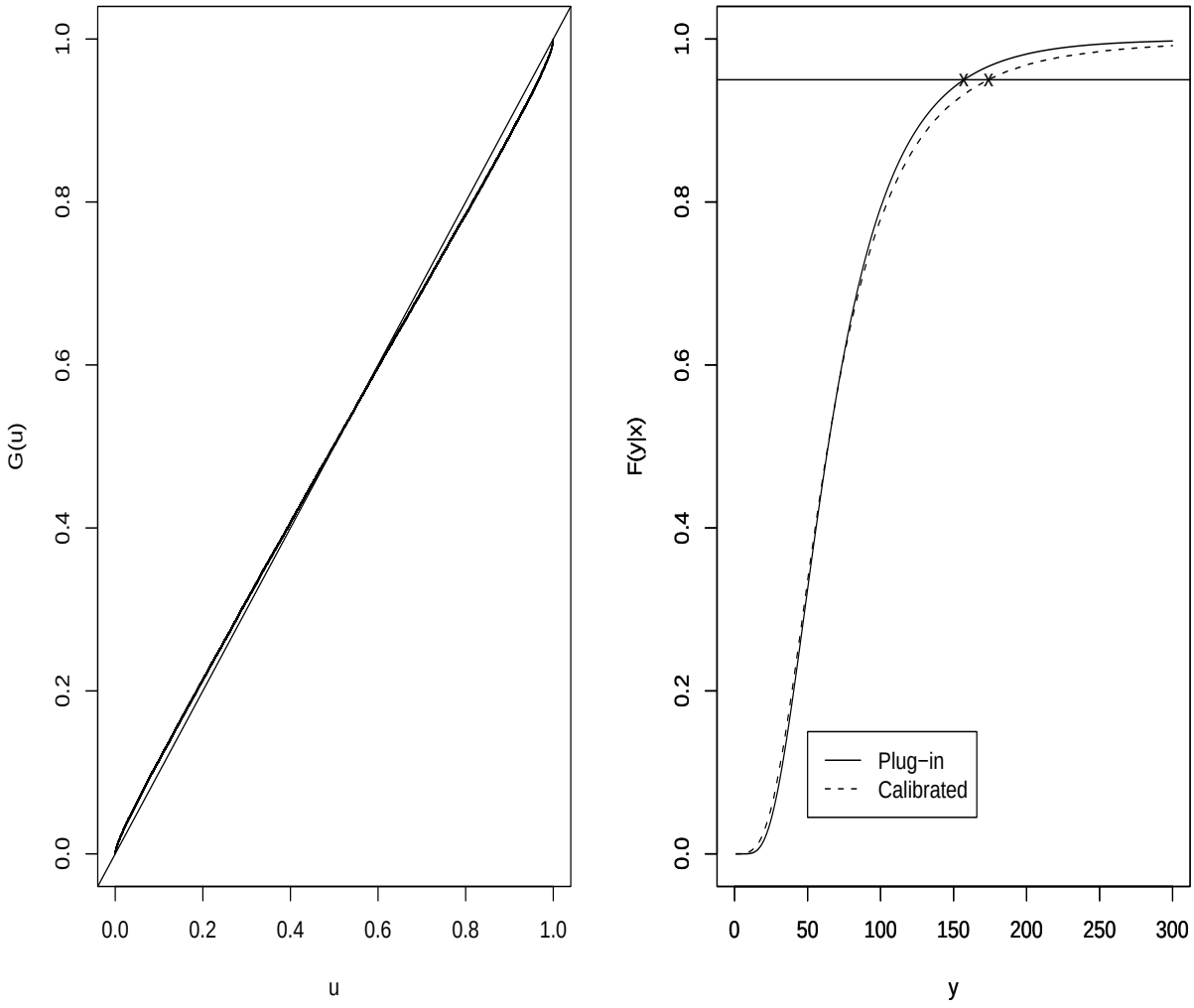

Figure 2.3: Functions $\tilde{G}(u)$ and $\tilde{F}_p(y|x)$, log-normal distribution.

Chapter 3

Prediction Models for Homogeneous Poisson Processes

We will now consider the prediction of recurrent events, events which occur repeatedly over time. Examples of recurrent events include successive tumors in cancer studies (Gail, Santner & Brown 1980), automobile warranty claims (Lawless & Nadeau 1995), failures of software (Raftery 1987), or scrams in a nuclear power plant (Martz, Parker & Rasmuron 1999).

We will now propose models to predict the number of future events for subjects already under observation. In this chapter, the intensity function that we will use to model these occurrences will be the one corresponding to a homogeneous Poisson process. First, we will present different ways used to predict the future number of events using such a process. Then, we will propose an alternative approach which uses a random effects model. Finally, we will perform various simulations to compare these different approaches.

3.1 Prediction of Recurrent Events

Let $\mathbf{N}(s,t)$ be the random variable representing the number of events occurring for a subject in the time interval $[s,t]$; we write $\mathbf{N}(t)$ for $\mathbf{N}(0,t)$. We will now consider continuous time processes where two events cannot occur simultaneously. Many different types of such processes are discussed in the literature (see Snyder & Miller (1991) or Grandell (1997)), but Poisson processes and renewal processes are the two most popular types used by statisticians to model such recurrent events. We can distinguish them through the intensity function

$$\lambda(t|H(t)) \;=\; \lim_{\Delta_t \to 0} \frac{P[\mathbf{N}(t,t+\Delta_t)=1|H(t)]}{\Delta_t}, \tag{3.1}$$

where $H(t)$ denotes the history of the process up to time t. Note that conditional on $H(0)$, (3.1) fully specifies the process $\{N(t), t>0\}$. Renewal processes make the assumption that (3.1) depends only on the time elapsed since the last event. These processes are semi-Markovian. On the other hand, the Poisson processes are Markovian because (3.1) depends only on t. The intensity, or rate, function is then simply denoted by $\lambda(t)$, and

$$\mathbf{N}(t) \sim \mathcal{PP}(\lambda(t))$$

means that $\mathbf{N}(t)$ is a Poisson process with rate function $\lambda(t)$. Note that when $\lambda(t)$ does not depend on t, $\mathbf{N}(t)$ is called a homogeneous Poisson process (HPP).

This Markovian assumption leads to the following well known properties of Poisson processes.

Proposition 3.1. *If* $\mathbf{N}(t) \sim \mathcal{PP}(\lambda(t))$*, then for all non-negative integers* n*,*

$$P[\mathbf{N}(s,r)=n] \;=\; \frac{e^{-\Lambda(s,r)}[\Lambda(s,r)]^n}{n!},$$

where $s<r$ *and* $\Lambda(s,r)=\int_s^r \lambda(u)du$*. Therefore,* $\mathbf{N}(s,r)$ *is a Poisson random variable with rate* $\Lambda(s,r)$*.*

Proposition 3.2. *If* $\mathbf{N}(t) \sim \mathcal{PP}(\lambda(t))$*, then* $\mathbf{N}(s_1, r_1)$ *and* $\mathbf{N}(s_2, r_2)$ *are independent Poisson random variables for any non-overlapping intervals* (s_1, r_1) *and* (s_2, r_2)*.*

For prediction problems involving recurrent events, these two properties combined with the fact that recurrent events data are often interval-censored make Poisson processes relatively easy to use. On the other hand, the distribution of the future number of events can be difficult to obtain when renewal processes are used. Since Poisson processes are much easier to use than renewal processes, we thus recommend the use of Poisson processes to predict recurrent events unless the Markovian assumption is clearly unreasonable.

Using the concepts presented in the previous chapters, we can now find point predictors and prediction intervals for occurrences arising from Poisson processes. We will consider homogeneous Poisson processes in this chapter and nonhomogeneous Poisson processes in Chapter 4.

3.2 Prediction of Homogeneous Poisson Processes

Suppose that we have k individuals and $\mathbf{N}_i(s, t)$ denotes the number of events occurring for the individual i in the time interval $[s, t]$. When these processes are time-homogeneous we can write

$$\mathbf{N}_i(t) \sim \mathcal{PP}(\lambda_i), \tag{3.2}$$

where the processes are independent and $i = 1, \ldots, k$. The rates can be either identical or different (*i.e.* $\lambda_i \neq \lambda_j$ for at least one i and one j). We will also suppose here that each process is observed up to a fixed time t_{1i} and that the interest is to predict the sum of the $\mathbf{N}_i(t_{1i}, t_{2i})$'s over a subset $\mathcal{S}$ of the k processes. It is clear from Proposition 3.1 that $\sum_{i \in \mathcal{S}} \mathbf{N}_i(t_{1i}, t_{2i})$ has a Poisson distribution with rate $\sum_{i \in \mathcal{S}} \lambda_i(t_{2i} - t_{1i})$.

According to the notation defined in Chapter 1, this is a prediction problem with $\mathbf{X} = (\mathbf{N}_1(t_{11}), \ldots, \mathbf{N}_k(t_{1k}))$ and $\mathbf{Y} = \sum_{i \in \mathcal{S}} \mathbf{N}_i(t_{1i}, t_{2i})$. Note that $(\mathbf{N}_1(t_{11}), \ldots, \mathbf{N}_k(t_{1k}))$ and $(\mathbf{N}_1(t_{11}, t_{21}), \ldots, \mathbf{N}_k(t_{1k}, t_{2k}))$ will be denoted by $\mathbf{N}(t_1)$ and $\mathbf{N}(t_1, t_2)$ respectively, whereas $\sum_{i \in \mathcal{S}} \mathbf{N}_i(t_{1i}, t_{2i})$ will be denoted by $N_{\mathcal{S}}(t_1, t_2)$.

Let τ_{ij} be the time of the jth occurrence coming from the ith process. Once we observe $\mathbf{N}(t_1)$ and the set of occurrence times τ, it is relatively straightforward to derive the likelihood function for the λ_i's from model (3.2) (Cox & Lewis 1966, Section 3.3):

$$\begin{aligned} L(\lambda_1, \ldots, \lambda_k | (N(t_1), \tau)) &= \prod_{i=1}^{k} \left(\prod_{j=1}^{N_i(t_{1i})} \lambda_i(\tau_{ij}) \right) \exp\{-\int_0^{t_{1i}} \lambda_i(u) du\} \qquad (3.3) \\ &= \prod_{i=1}^{k} \left(\prod_{j=1}^{N_i(t_{1i})} \lambda_i \right) \exp\{-\lambda_i t_{1i}\} \\ &= \left(\prod_{i=1}^{k} \lambda_i^{N_i(t_{1i})} \right) \exp\left\{ \sum_{i=1}^{k} -\lambda_i t_{1i} \right\}. \end{aligned}$$

We can see that

$$L(\lambda_1, \ldots, \lambda_k | (N(t_1), \tau)) = L(\lambda_1, \ldots, \lambda_k | N(t_1)),$$

which means that $N(t_1)$ is sufficient for $\lambda_1, \ldots, \lambda_k$ when the Poisson processes are time homogeneous.

The m.l.e. of λ_i is given by

$$\hat{\lambda}_i = \begin{cases} \dfrac{N_i(t_{1i})}{t_{1i}} & \text{if } \lambda_i \neq \lambda_j \ \forall j \neq i, \\ \dfrac{\sum_{j:\lambda_j = \lambda_i} N_j(t_{1j})}{\sum_{j:\lambda_j = \lambda_i} t_{1j}} & \text{otherwise.} \end{cases}$$

We will see that this estimate of λ_i is often used to obtain point and set predictions for $N_{\mathcal{S}}(t_1, t_2)$.

3.2.1 Point Prediction

The following proposition will show that an optimal point predictor for $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ is available under model (3.2). This proposition is valid only when all the λ_i's are different, but can be easily modified if some or all the λ_i's are known to be identical.

Proposition 3.3. *Let $\mathbf{N}_i(t) \sim \mathcal{PP}(\lambda_i)$ and $\mathcal{S}$ be a subset of $(1, \ldots, k)$. If $\lambda_i \neq \lambda_j \; \forall\, i, j = 1, \ldots, k$, an unbiased point predictor for $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ is given by $\sum_{i \in \mathcal{S}} (t_{2i} - t_{1i})\hat{\lambda}_i$, where $\hat{\lambda}_i = N_i(t_{1i})/t_{1i}$ is the m.l.e. of λ_i. Furthermore, among all the unbiased predictors obtained using the already observed sample $N(t_1)$, none of them has a prediction error with a smaller variance.*

Proof. First let us show that $\sum(t_{2i} - t_{1i})\hat{\lambda}_i$ is an unbiased predictor for $\sum \mathbf{N}_i(t_{1i}, t_{2i})$. Note that throughout this proof, the sums are always taken over all the observations in the subset $\mathcal{S}$.

$$\begin{aligned} \mathbb{E}\left[\sum (t_{2i} - t_{1i})\hat{\lambda}_i\right] &= \sum (t_{2i} - t_{1i}) \mathbb{E}\left[\frac{\mathbf{N}_i(t_{1i})}{t_{1i}}\right] \\ &= \sum (t_{2i} - t_{1i})\lambda_i \\ &= \mathbb{E}\left[\mathbf{N}_{\mathcal{S}}(t_1, t_2)\right]. \end{aligned}$$

We know from Proposition 3.2 that the random vectors $\mathbf{N}(t_1)$ and $\mathbf{N}(t_1, t_2)$ are independent. The variance of the prediction error is then given by

$$\begin{aligned} \mathbb{V}ar\left[\sum [(t_{2i} - t_{1i})\hat{\lambda}_i - \mathbf{N}_i(t_{1i}, t_{2i})]\right] &= \mathbb{V}ar\left[\sum (t_{2i} - t_{1i})\frac{\mathbf{N}_i(t_{1i})}{t_{1i}}\right] + \mathbb{V}ar\left[\sum \mathbf{N}_i(t_{1i}, t_{2i})\right] \\ &= \sum (t_{2i} - t_{1i})^2 \frac{\lambda_i}{t_{1i}} + \mathbb{V}ar\left[\sum \mathbf{N}_i(t_{1i}, t_{2i})\right]. \end{aligned} \tag{3.4}$$

We will now show that $\sum (t_{2i} - t_{1i})^2 \frac{\lambda_i}{t_{1i}}$ corresponds to the Cramer-Rao lower bound for the variance of an unbiased estimator of $\mathbb{E}[\mathbf{N}_{\mathcal{S}}(t_1, t_2)]$. Since the second term in the right hand

side of (3.4) is common for all unbiased predictors, this will be sufficient to prove that no unbiased predictor of $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ based on $N(t_1)$ can have a prediction error with a smaller variance. The Cramer-Rao bound for the multivariate parameter case is given in Kendall & Allen (1977, page 16):

$$\sum_i \sum_j \frac{\partial g(\lambda)}{\partial \lambda_i} \frac{\partial g(\lambda)}{\partial \lambda_j} \mathcal{I}_{ij}^{-1},$$

where $g(\lambda)$ is the expectation of $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ and the $\mathcal{I}_{ij}^{-1}$'s are the components of the inverse of Fisher's information matrix for $\{\lambda_i : i \in \mathcal{S}\}$. We can see that

$$\begin{aligned} \frac{\partial g(\lambda)}{\partial \lambda_i} &= \frac{\partial}{\partial \lambda_i} \sum (t_{2i} - t_{1i}) \lambda_i \\ &= t_{2i} - t_{1i} \end{aligned}$$

and $\mathcal{I}$ is a diagonal matrix with components

$$\begin{aligned} \mathcal{I}_{ij} &= \mathbb{E}\left[\left(\frac{\partial}{\partial \lambda_i} \log P[\mathbf{N}(t_1) = N(t_1); \lambda] \right) \left(\frac{\partial}{\partial \lambda_j} \log P[\mathbf{N}(t_1) = N(t_1); \lambda] \right) \right] \\ &= \mathbb{E}\left[\left(-t_{1i} + \frac{\mathbf{N}_i(t_{1i})}{\lambda_i} \right) \left(-t_{1j} + \frac{\mathbf{N}_j(t_{1j})}{\lambda_j} \right) \right] \text{ (from Proposition 3.1)} \\ &= -t_{1i} t_{1j} + \mathbb{E}\left[\frac{\mathbf{N}_i(t_{1i}) \mathbf{N}_j(t_{1j})}{\lambda_i \lambda_j} \right] \\ &= \begin{cases} \dfrac{t_{1i}}{\lambda_i} & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \end{aligned}$$

Therefore, the Cramer-Rao lower bound can be written as

$$\begin{aligned} \sum_i \sum_j \left(\frac{\partial g(\lambda)}{\partial \lambda_i} \right) \left(\frac{\partial g(\lambda)}{\partial \lambda_j} \right) \mathcal{I}_{ij}^{-1} &= \sum (t_{2i} - t_{1i})^2 \frac{\lambda_i}{t_{1i}} \\ &= \mathbb{V}ar\left[\sum (t_{2i} - t_{1i}) \hat{\lambda}_i \right]. \end{aligned}$$

Note that the usual regularity conditions are satisfied here. □

We can see that the predictor $\sum_{i \in \mathcal{S}}(t_{2i} - t_{1i})N_i(t_{1i})/t_{i1}$ is no longer optimal when at least two λ_i's are identical. This points out the importance of correctly assessing the homogeneity/heterogeneity of the rates. The proof of the preceding proposition also suggests a general method to find optimal predictors when the two samples of interest are independent: a predictor will be the best unbiased predictor if it is an efficient estimator of the expectation of the quantity to predict.

3.2.2 Prediction intervals

We saw in Section 1.2 that the literature provides some methods to find exact prediction intervals. However, little work has been done on finding prediction intervals for discrete random variables (Patel & Samaranayake 1991). One of the main obstacles is that pivotal quantities are rarely available for these random variables. We will now present some ways to find prediction intervals for HPP's. However, some restrictions have to be made on the rates or on the observation times in order to find exact prediction intervals.

The quantity $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ taking only integer values, an exact $1 - \alpha$ prediction interval for a discrete random variable is defined as an interval $[L(N(t_1)), U(N(t_1))]$ with integer endpoints such that

$$P[L(\mathbf{N}(t_1)) \leq \mathbf{N}_{\mathcal{S}}(t_1, t_2) \leq U(\mathbf{N}(t_1)); \lambda] \;\; \geq \;\; 1 - \alpha$$

for any vector $\lambda = \{\lambda_i : i = 1, \ldots, k\}$, but the coverage probability is smaller than $1 - \alpha$ if we increase $L(\mathbf{N}(t_1))$ or decrease $U(\mathbf{N}(t_1))$. Note that the endpoints can be randomized to obtain an interval with a coverage probability of exactly $1 - \alpha$.

Exact $1 - \alpha$ prediction intervals for $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ are available if either the rates or the observation times are identical for all the processes in $\mathcal{S}$. These two conditions are defined as followed,

- **Condition 1:** $\lambda_i = \lambda^* \; \forall i \in \mathcal{S}$.
- **Condition 2:** $(t_{1i}, t_{2i}) = (t_1^*, t_2^*) \; \forall i \in \mathcal{S}$.

The random variable $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ will have a Poisson distribution with rate $\lambda^* \sum_{i \in \mathcal{S}} (t_{2i} - t_{1i})$ or $(t_2^* - t_1^*) \sum_{i \in \mathcal{S}} \lambda_i$ under Condition 1 or 2 respectively. Note that both conditions are trivially satisfied when we want to predict the number of occurrences of a single observation.

One method to obtain an exact prediction interval is given in Faulkenberry (1973). It uses the sufficient statistic method presented at the end of Section 1.2.2. First, we see that once we observe $(N_i(t_{1i}), N_i(t_{1i}, t_{2i}))$ for all the observations in $\mathcal{S}$, $N_{\mathcal{S}}(t_2)$ is a sufficient statistic for λ^* when Condition 1 is true or for $\sum_{i \in \mathcal{S}} \lambda_i$ when Condition 2 is true. The conditional distribution of $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ given $N_{\mathcal{S}}(t_2)$ is binomial with density

$$f(N_{\mathcal{S}}(t_1, t_2) | N_{\mathcal{S}}(t_2)) \;=\; \binom{N_{\mathcal{S}}(t_2)}{N_{\mathcal{S}}(t_1, t_2)} \left[\frac{\sum_{i \in \mathcal{S}} (t_{2i} - t_{1i})}{\sum_{i \in \mathcal{S}} t_{2i}} \right]^{N_{\mathcal{S}}(t_1, t_2)} \left(\frac{\sum_{i \in \mathcal{S}} t_{1i}}{\sum_{i \in \mathcal{S}} t_{2i}} \right)^{N_{\mathcal{S}}(t_1)}.$$

Using this density, it is shown in Faulkenberry (1973) that an exact $1-\alpha$ prediction interval procedure is given by the biggest integer a and the smallest integer b such that

$$\sum_{i=0}^{b} f(i | N_{\mathcal{S}}(t_1) + b) \;\geq\; 1 - \alpha_1$$

and

$$\sum_{i=0}^{a} f(i | N_{\mathcal{S}}(t_1) + a) \;<\; \alpha_2,$$

where $\alpha_1 + \alpha_2 = \alpha$. Using similar approaches, Vit (1973) and Nelson (1969) found the same prediction interval.

Approximate prediction intervals are also available by using the plug-in method presented in Section 1.2.3. These intervals will be close to exact prediction intervals when $N_{\mathcal{S}}(t_1)$ is large enough and can be calibrated using one of the algorithms presented in

Chapter 2. Under Condition 1, an interval is obtained by estimating λ^* with $\widehat{\lambda^*} = N_{\mathcal{S}}(t_1)/\sum_{i\in\mathcal{S}} t_{1i}$ and finding the appropriate quantiles of a Poisson distribution with rate $\widehat{\lambda^*}\sum_{i\in\mathcal{S}}(t_{2i}-t_{1i})$. When Condition 2 is true, we estimate $\sum_{i\in\mathcal{S}}\lambda_i$ with $N_{\mathcal{S}}(t_1)/t_1^*$ and find the quantiles of a Poisson distribution with rate $(t_2^*-t_1^*)N_{\mathcal{S}}(t_1)/t_1^*$. Note that additional conditions for the processes in $\mathcal{S}^c$ can provide better prediction intervals. For example, if all the rates in $\{1,\ldots,k\} = \mathcal{S}\cup\mathcal{S}^c$ are assumed to be identical, an approximate prediction interval for $\mathbf{N}_{\mathcal{S}}(t_1,t_2)$ is obtained by replacing $\hat{\lambda} = \sum_{i\in\mathcal{S}} N_i(t_{1i})/\sum_{i\in\mathcal{S}} t_{1i}$ with $\hat{\lambda} = \sum_{i=1}^k N_i(t_{1i})/\sum_{i=1}^k t_{1i}$.

The maximum likelihood predictive density (MLPD) presented in Section 1.2.4 can also be used to obtain an approximate prediction interval for $\mathbf{N}_{\mathcal{S}}(t_1,t_2)$ under Condition 1 or 2. However, this method usually provides intervals that are not easy to compute. We can show that the MLPD for $\mathbf{N}_{\mathcal{S}}(t_1,t_2)$ given that $N_{\mathcal{S}}(t_1)=x$ can be written as

$$\tilde{f}_p(y|x) = k(x)\frac{\exp(-y)}{y!}(x+y)^{(x+y)}\left[\frac{\sum_{i\in\mathcal{S}}(t_{2i}-t_{1i})}{\sum_{i\in\mathcal{S}} t_{2i}}\right]^y,$$

where $k(x)$ is a normalizing constant. This constant can be approximated by using the fact that

$$\log(k(x)^{-1}) = \left[e^{-1}\frac{\sum_{i\in\mathcal{S}}(t_{2i}-t_{1i})}{\sum_{i\in\mathcal{S}} t_{2i}}\right]\log\left(\mathbb{E}[(x+Z)^{(x+Z)}]\right),$$

where Z has a Poisson distribution with rate $e^{-1}\sum_{i\in\mathcal{S}}(t_{2i}-t_{1i})/\sum_{i\in\mathcal{S}} t_{2i}$. When $N_{\mathcal{S}}(t_2)$ is large, Lejeune (1975) showed that this MLPD can be approximated by a negative binomial with parameters $x+1/2$ and $\sum_{i\in\mathcal{S}} t_{1i}/\sum_{i\in\mathcal{S}} t_{2i}$, *i.e.*

$$\tilde{f}_p(y|x) \simeq \frac{\Gamma(y+x+\frac{1}{2})}{\Gamma(x+\frac{1}{2})y!}\left(\frac{\sum_{i\in\mathcal{S}} t_{1i}}{\sum_{i\in\mathcal{S}} t_{2i}}\right)^{x+\frac{1}{2}}\left[\frac{\sum_{i\in\mathcal{S}}(t_{2i}-t_{1i})}{\sum_{i\in\mathcal{S}} t_{2i}}\right]^y.$$

We will often mention the negative binomial distribution in this thesis. Note that we are always referring to the distribution where the first parameter is not necessarily an integer.

It is often of interest to predict $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ when the rates and the observation times are different (*i.e.* when Conditions 1 and 2 are not true). However, it does not seem possible to use the sufficient statistic method to find an exact prediction interval. Nevertheless, when all the $N_i(t_{1i})$'s in the subset $\mathcal{S}$ are sufficiently large, the plug-in approach can provide an approximate prediction interval with a coverage probability close to $1-\alpha$. Since $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ has a Poisson distribution with rate $\sum_{i \in \mathcal{S}}(t_{2i}-t_{1i})\lambda_i$, we can estimate this rate by replacing the λ_i's with their m.l.e.'s $\hat{\lambda}_i = N_i(t_{1i})/t_{1i}$. An approximate $1-\alpha$ prediction interval is then obtained by finding the quantiles of this distribution. Unfortunately, the actual coverage probability may be far from $1-\alpha$ when some of the $\hat{\lambda}_i$'s are highly variable.

It is also possible to use the MLPD method when Conditions 1 and 2 are not true, but the maximization of the joint density of $(\mathbf{N}(t_1), \mathbf{N}_{\mathcal{S}}(t_1, t_2))$ with respect to $\{\lambda_i : i \in \mathcal{S}\}$ is often difficult to obtain. Note also that the predictive likelihood approach, presented in Section 1.2.4, is not of interest here since the sufficiency principle does not provide a genuine reduction of the data.

It can also be useful to use a Bayesian approach to obtain approximate prediction intervals. A natural prior for the λ_i's is the conjugate gamma distribution. However, it is also possible to use Jeffreys' non-informative prior (1.2). This prior can be written as

$$\pi(\lambda_i) \;=\; \frac{1}{\sqrt{\lambda_i}}.$$

This density is improper (*i.e.* $\int_{\lambda_i} \pi(\lambda_i) d\lambda_i = \infty$), but we can show that the posterior distribution is always a (proper) gamma. The Bayesian approach being similar, from a mathematical point of view, to the random effects model proposed in the next section, we will forego here the presentation of the corresponding posterior and predictive densities.

3.3 Prediction Models using Random Effects

We saw in the previous section that prediction problems involving Poisson processes are not handled easily via known frequentist methods. Exact prediction intervals are only available when the rates or the observation times are identical, a strict assumption in many practical situations. Furthermore, when we relax these assumptions, the approximate prediction intervals are only adequate when all the $N_i(t_{1i})$'s are large. In order to find better approximate prediction intervals, we will propose in this section an approach using random effects to model the different rates.

3.3.1 Random Effects Model

Again, let $\mathbf{N}_i(t)$ be a HPP with an unknown rate λ_i. Now, we will assume that the λ_i's are unobservable i.i.d. random effects. This may seem like a stringent assumption but unobservable rates are often believed to be random in some sense. For example, the rates may be affected by various unobserved covariates, random events, or shocks. If we let the λ_i's be gamma random variables, we have the following random effects model:

$$\begin{aligned} \mathbf{N}_i(t)|\lambda_i &\sim \mathcal{PP}(\lambda_i), \\ \lambda_i &\sim \text{Gamma}(a, b) \end{aligned} \tag{3.5}$$

where $i = 1, \ldots, k$. The density of a particular λ_i is then given by

$$\frac{b^a e^{-\lambda_i b} \lambda_i^{a-1}}{\Gamma(a)},$$

with $\mathbb{E}[\lambda_i] = a/b$ and $\mathbb{V}ar[\lambda_i] = a/b^2$. By assuming such a distribution on the unobservable λ_i's, we now have a model with only 2 unknown parameters, a and b, instead of k unknown parameters in model (3.2).

Although they look similar from a mathematical point of view, random effects models are different from Bayesian models. Random effects models assume an actual physical model on unobservable quantities, while Bayesian models incorporate distributions on these quantities based on subjective probabilities. Nevertheless, both approaches assume a random distribution on unobservable quantities and many Bayesian concepts presented in Section 1.3 can be used with random effects models.

Let $\pi(\lambda; a, b)$ be the density function of the random effects. Once we observe $N(t_1)$, the conditional density of the random effects is

$$\begin{aligned}
\pi(\lambda|N(t_1); a, b) &= \frac{P[\mathbf{N}(t_1) = N(t_1)|\lambda]\pi(\lambda; a, b)}{\int_\lambda P[\mathbf{N}(t_1) = N(t_1)|\lambda]\pi(\lambda; a, b)d\lambda} \\
&= \prod_{i=1}^{k} \frac{e^{-\lambda_i(b+t_{1i})}\lambda_i^{a+N_i(t_{1i})-1}}{\int_{\lambda_i} e^{-\lambda_i(b+t_{1i})}\lambda_i^{a+N_i(t_{1i})-1}d\lambda_i} \\
&= \prod_{i=1}^{k} \frac{(b+t_{1i})^{a+N_i(t_{1i})}e^{-\lambda_i(b+t_{1i})}\lambda_i^{a+N_i(t_{1i})-1}}{\Gamma(a+N_i(t_{1i}))}.
\end{aligned}$$

The $\lambda_i|N(t_1)$'s are then independent Gamma$(a + N_i(t_{1i}), b + t_{1i})$ random variables. Since only $N_i(t_{1i})$ and t_{1i} affect the conditional distribution of λ_i, we will use $\lambda_i|N_i(t_{1i})$ instead of $\lambda_i|N(t_1)$ to represent this random variable. Clearly, the choice of a gamma distribution in model (3.5) was motivated by its nice mathematical properties when used with Poisson processes. Nevertheless, we will see in Section 3.4 that predictors and prediction intervals obtained using this model seem appropriate even when the real random effects are not gamma.

Using this conditional density, the density function of $\mathbf{N}_s(t_1, t_2)$ given $N(t_1)$ can be found:

Proposition 3.4. *Under the model given in (3.5), the density function of* $\mathbf{N}_s(t_1, t_2)$ *given*

$N(t_1)$ is a convolution of $|\mathcal{S}|$ negative binomials with parameters $a+N_i(t_{1i})$ and $(b+t_{1i})/(b+t_{2i})$.

Proof. First, let us find the density of $\mathbf{N}_i(t_{1i}, t_{2i})$ given $N(t_1)$:

$$\begin{aligned} P[\mathbf{N}_i(t_{1i}, t_{2i}) = n | N(t_1); a, b] &= \int_{\lambda_i} P[\mathbf{N}_i(t_{1i}, t_{2i}) = n|\lambda_i]\pi(\lambda_i|N_i(t_{1i}); a, b)d\lambda_i \\ &= \int_{\lambda_i} \left(\frac{e^{-(t_{2i}-t_{1i})\lambda_i}[(t_{2i}-t_{1i})\lambda_i]^n}{n!} \right) \times \\ &\quad \left(\frac{(b+t_{1i})^{a+N_i(t_{1i})} e^{-\lambda_i(b+t_{1i})} \lambda_i^{a+N_i(t_{1i})-1}}{\Gamma(a+N_i(t_{1i}))} \right) d\lambda_i \\ &= \frac{\Gamma(a+N_i(t_{1i})+n)}{\Gamma(a+N_i(t_{1i}))n!} \left(\frac{t_{2i}-t_{1i}}{b+t_{2i}} \right)^n \left(\frac{b+t_{1i}}{b+t_{2i}} \right)^{a+N_i(t_{1i})}. \end{aligned}$$

Which is the density of negative binomial with parameters $a+N_i(t_{1i})$ and $(b+t_{1i})/(b+t_{2i})$.

Therefore, the density function of $\sum_{i \in \mathcal{S}} \mathbf{N}_i(t_{1i}, t_{2i})$ given $N(t_1)$ is a convolution of $|\mathcal{S}|$ negative binomials and its density, denoted by $p(n|N(t_1))$, is

$$\begin{aligned} p(n|N(t_1)) &= P[\mathbf{N}_{\mathcal{S}}(t_1, t_2) = n | N(t_1); a, b] \\ &= \sum_{\{z_i : \sum_{\mathcal{S}} z_i = n\}} \prod_{i \in \mathcal{S}} \frac{\Gamma(a+N_i(t_{1i})+z_i)}{\Gamma(a+N_i(t_{1i}))z_i!} \left(\frac{t_{2i}-t_{1i}}{b+t_{2i}} \right)^{z_i} \times \\ &\quad \left(\frac{b+t_{1i}}{b+t_{2i}} \right)^{a+N_i(t_{1i})}. \end{aligned} \tag{3.6}$$

□

The density (3.6) is obtained by summing over $\binom{n+|\mathcal{S}|-1}{n}$ terms, which may not be feasible when $|\mathcal{S}|$ or n is large. However, this problem in principle can be solved by using a recursive formula to find $p(n|N(t_1))$.

Proposition 3.5. *Under the model given in (3.5), the density function of* $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ *given* $N(t_1)$ *can be written as*

$$\begin{aligned} p(n|N(t_1); a, b) &= P[\mathbf{N}_{\mathcal{S}}(t_1, t_2) = n|N(t_1); a, b] \\ &= \begin{cases} \displaystyle\prod_{i \in \mathcal{S}} \left(\frac{b + t_{1i}}{b + t_{2i}}\right)^{(a+N_i(t_{1i}))} & \text{if } n = 0, \\ \displaystyle\sum_{j=0}^{n-1} p(j|N(t_1); a, b) \left(\frac{H^{(n-1-j)}(0)}{(n-1-j)!}\right) & \text{if } n \geq 1, \end{cases} \end{aligned}$$

where

$$H^{(j)}(0) = \sum_{i \in \mathcal{S}} (a + N_i(t_{1i})) \left(\frac{t_{2i} - t_{i1}}{b + t_{2i}}\right)^{j+1}.$$

A proof of this, which uses the probability generating function of $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ given $N(t_1)$, can be easily derived from Klugman, Panger & Wilmot (2004, Example 4.60).

When n is very large, it can be more convenient to approximate $p(n|N(t_1); a, b)$ instead of using a recursive formula. Such an approximation can be done by generating convolutions of gamma random variables. This is due to the fact that the predictive density function can be written as

$$\begin{aligned} p(n|N(t_1); a, b) &= \int_{\lambda} P[\mathbf{N}_{\mathcal{S}}(t_1, t_2) = n|\lambda] \pi(\lambda|N(t_1); a, b) d\lambda \\ &= \int_{\lambda} P\left[\text{Poisson}\left(\sum_{i \in \mathcal{S}} (t_{2i} - t_{1i})\lambda_i\right) = n\right] \pi(\lambda|N(t_1); a, b) d\lambda \\ &= \int_0^{\infty} P\left[\text{Poisson}(u) = n\right] \psi(u; a, b) du \\ &= \int_0^{\infty} \frac{e^{-u} u^n}{n!} \psi(u; a, b) du, \end{aligned}$$

where $\psi(u; a, b)$ is the density of the convolution of $|\mathcal{S}|$ Gamma$(a + N_i(t_{1i}), (b + t_{1i})/(t_{2i} - t_{1i}))$ random variables. Once we generate B convolutions of gammas $u_1^*, \ldots, u_B^*$, we can

approximate the predictive density function with

$$p(n|N(t_1);a,b) \;\simeq\; \sum_{i=1}^{B} \frac{e^{-u_i^*}(u_i^*)^n}{n!}.$$

An asymptotic expansion for $p(n|N(t_1);a,b)$ can also be derived using a result given in Bender (1974) about asymptotic expansion for convolutions of random variables. For any $i \in \mathcal{S}$ we can show that:

$$p(n|N(t_1);a,b) \;\simeq\; \frac{\Gamma(a+N_i(t_{1i})+n)}{\Gamma(a+N_i(t_{1i}))n!}\left(\frac{t_{2i}-t_{1i}}{b+t_{2i}}\right)^n\left(\frac{b+t_{1i}}{b+t_{2i}}\right)^{a+N_i(t_{1i})} \times C\left(\frac{b+t_{2i}}{t_{2i}-t_{1i}}\right), \tag{3.7}$$

where

$$C(s) \;=\; \prod_{j\in\mathcal{S}\setminus\{i\}}\left[1+(1-s)\frac{t_{2j}-t_{1j}}{b+t_{1j}}\right]^{-(a+N_j(t_{1j})}.$$

When a and b are known (the case where a and b are unknown will be considered in the next subsection), methods presented in Section 1.3.3 can be used to find unbiased predictors and exact prediction intervals. For example, we can see that

$$\begin{aligned}
\mathbb{E}[\mathbf{N}_{\mathcal{S}}(t_1,t_2)|N(t_1)] &= \mathbb{E}\left[\mathbb{E}[\mathbf{N}_{\mathcal{S}}(t_1,t_2)|\lambda]|N(t_1)\right] \\
&= \mathbb{E}\left[\sum_{i\in\mathcal{S}}\lambda_i(t_{2i}-t_{1i})|N(t_1)\right] \\
&= \sum_{i\in\mathcal{S}}\mathbb{E}[\lambda_i|N_i(t_{1i})](t_{2i}-t_{1i}) \\
&= \sum_{i\in\mathcal{S}}\left(\frac{a+N_i(t_{1i})}{b+t_{1i}}\right)(t_{2i}-t_{1i})
\end{aligned} \tag{3.8}$$

and

$$\mathbb{E}\left[\mathbb{E}[\mathbf{N}_{\mathcal{S}}(t_1,t_2)|\mathbf{N}(t_1)]-\mathbf{N}_{\mathcal{S}}(t_1,t_2)\right] \;=\; \mathbb{E}[\mathbf{N}_{\mathcal{S}}(t_1,t_2)]-\mathbb{E}[\mathbf{N}_{\mathcal{S}}(t_1,t_2)]$$

$$= \quad 0.$$

Thus, provided with the full knowledge of a and b, (3.8) is an unbiased point predictor for $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$. It can also be derived from Section 1.3.3 that this predictor minimizes

$$\mathbb{E}[(\mathbf{N}_{\mathcal{S}}(t_1, t_2) - \widehat{\mathbf{N}}_{\mathcal{S}}(t_1, t_2))^2 | \mathbf{N}(t_1)]$$

among all predictors $\widehat{\mathbf{N}}_{\mathcal{S}}(t_1, t_2)$.

When a and b are known, exact prediction intervals for $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ are available by using quantiles of the predictive density $\tilde{f}_p(n|N(t_1)) = p(n|N(t_1); a, b)$. Let $q_{\mathbf{N}_{\mathcal{S}}|N(t_1)}(\alpha; a, b)$ be the α quantile of this predictive distribution. Then, if $\alpha_1 + \alpha_2 = \alpha$,

$$\begin{aligned} P[q_{\mathbf{N}_{\mathcal{S}}|\mathbf{N}(t_1)}(\alpha_1; a, b) \leq \mathbf{N}_{\mathcal{S}}(t_1, t_2) \quad &\leq \\ q_{\mathbf{N}_{\mathcal{S}}|\mathbf{N}(t_1)}(1 - \alpha_2; a, b)] \quad &= \quad \sum_{N(t_1)} P[q_{\mathbf{N}_{\mathcal{S}}|N(t_1)}(\alpha_1; a, b) \leq \mathbf{N}_{\mathcal{S}}(t_1, t_2) \leq \\ &\qquad q_{\mathbf{N}_{\mathcal{S}}|N(t_1)}(1 - \alpha_2; a, b); a, b] P[\mathbf{N}(t_1) = N(t_1); a, b] \\ &= \quad \sum_{N(t_1)} (1 - \alpha) P[\mathbf{N}(t_1) = N(t_1)] \\ &= \quad 1 - \alpha. \end{aligned}$$

So, the interval $[q_{\mathbf{N}_{\mathcal{S}}|N(t_1)}(\alpha_1; a, b), q_{\mathbf{N}_{\mathcal{S}}|N(t_1)}(1-\alpha_2; a, b)]$ is an exact $1-\alpha$ prediction interval with respect to Definition 1.3; under repeated sampling of λ, $\mathbf{N}(t_1)$, and $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$, the interval obtained will contain $n(t_1, t_2)$ for $(1 - \alpha)100\%$ of theses samples.

To conclude this section, we would like to mention that this random effects model can be adapted if it is assumed that all the rates in $\mathcal{S}$ are identical. This model can written as

$$\begin{aligned} \mathbf{N}_i(t)|\lambda &\sim \mathcal{PP}(\lambda), \\ \lambda &\sim \text{Gamma}(a, b), \end{aligned}$$

for all i in $\mathcal{S}$. The conditional distribution of λ is then Gamma$(a + n(t_1), b + \sum_{i \in \mathcal{S}} t_{1i})$,

and the distribution of $\mathbf{N}_{\mathcal{S}}(t_1, t_2)|N(t_1)$ is negative binomial with density

$$p(n|N(t_1); a, b) = \frac{\Gamma(a + N(t_1) + n)}{\Gamma(a + N(t_1))\, n!} \left(\frac{\sum_{i\in\mathcal{S}} t_{2i} - \sum_{i\in\mathcal{S}} t_{1i}}{b + \sum_{i\in\mathcal{S}} t_{2i}} \right)^n \left(\frac{b + \sum_{i\in\mathcal{S}} t_{1i}}{b + \sum_{i\in\mathcal{S}} t_{2i}} \right)^{a+N(t_1)}.$$

This expression is also the Bayesian predictive distribution for a setting in which the common rate λ is given a Gamma(a, b) prior distribution; see the end of Section 3.2.2. Also, if it is possible to classify the observations in different groups, we can adapt this model to assign a common rate for all the processes within a group.

3.3.2 Complete Specification of the Random Effects Model

In the previous subsection, we used the fact that a model with random effects, like a Bayesian model, allows us to use posterior expectations and predictive densities to find unbiased predictors and exact prediction intervals. However, if it is common to model a state of knowledge through a fully specified prior distribution, it is a stringent assumption to state that the distribution of the random effects is fully known. Therefore, when we use a Gamma(a, b) distribution for the random effects λ, we should assume that none of these parameters are known. In this subsection, we will discuss ways to estimate the parameters a and b. These estimates will then substitute for the real parameters in the predictors and prediction intervals previously mentioned. If these estimates are consistent, we will then have approximately unbiased predictors and approximately exact prediction intervals. Note that it is also possible to calibrate these plug-in prediction intervals.

First, obvious candidates for $\hat{a}$ and $\hat{b}$ are the values maximizing the marginal likelihood of $N(t_1)$. This likelihood is given by

$$L(a, b|N(t_1)) = \int_\lambda P[\mathbf{N}(t_1) = N(t_1)|\lambda]\pi(\lambda|a, b)d\lambda$$

$$
\begin{aligned}
&= \prod_{i=1}^{k} \frac{t_{1i}^{N_i(t_{1i})}\, b^a}{N_i(t_{1i})!\Gamma(a)} \int_{\lambda_i} e^{-\lambda_i(b+t_{1i})} \lambda_i^{a+N_i(t_{1i})-1} d\lambda_i \\
&= \prod_{i=1}^{k} \frac{\Gamma(a+N_i(t_{1i}))}{\Gamma(a)N_i(t_{1i})!} \left(\frac{t_{1i}}{b+t_{1i}}\right)^{N_i(t_{1i})} \left(\frac{b}{b+t_{1i}}\right)^{a}.
\end{aligned}
$$

The $\mathbf{N}_i(t_{1i})$'s are then independent random variables, having a negative binomial distribution with parameters a and $b/(b+t_{1i})$. The marginal m.l.e.'s $\hat{a}_{mle}$ and $\hat{b}_{mle}$ can be found using usual maximization techniques. Also, some routines specifically written for the negative binomial distribution are available in standard statistical software.

Another way to obtain estimates of a and b is to match the first two unconditional centered moments with their respective empirical moments. Letting $\mathbf{R}_i$ be $\mathbf{N}_i(t_{1i})/t_{1i}$, we have

$$
\begin{aligned}
\mathbb{E}[\mathbf{R}_i] &= \mathbb{E}[\mathbb{E}[\mathbf{R}_i|\lambda_i]] \\
&= \frac{a}{b}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{V}ar[\mathbf{R}_i] &= \mathbb{E}[\mathbb{V}ar[\mathbf{R}_i|\lambda_i]] + \mathbb{V}ar[\mathbb{E}[\mathbf{R}_i|\lambda_i]] \\
&= \frac{a}{b}(t_{1i})^{-1} + \frac{a}{b^2}.
\end{aligned}
$$

The variance of $\mathbf{R}_i$ being different for each observation, we can replace $(t_{1i})^{-1}$ with its average value. We can now use the moment matching method to obtain the estimates

$$
\hat{a}_{mm} = \frac{(\overline{R})^2}{S_R^2 - \overline{R}\left(\overline{t_1^{-1}}\right)},
$$

$$\hat{b}_{mm} = \frac{\overline{R}}{S_R^2 - \overline{R}\left(\overline{t_1^{-1}}\right)},$$

where $\overline{t_1^{-1}}$ is the sample mean of $(t_{1i})^{-1}$, while $\overline{R}$ and S_R^2 are respectively the sample mean and variance of the R_i's. Note that when all the t_{1i}'s are identical, it is possible to show that the moment matching estimates and the m.l.e.'s partially agree in the sense that

$$\frac{\hat{a}_{mm}}{\hat{b}_{mm}} = \frac{\hat{a}_{mle}}{\hat{b}_{mle}}.$$

When plug-in methods are discussed in the literature, the estimates used always focus on how well they fit the data. This is done mostly through the maximization of a likelihood function or the minimization of a specified discrepancy. However, it can be argued that the ultimate goal is not to find a model that provides a good fit to the data but a model able to predict adequately upcoming observations. Clearly, the goodness of a fit and its ability to predict are usually not two competing goals but it may be of interest to find estimates focusing on the latter. For example, if we especially want to find point predictors for $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$, one can select parameters that will make the point prediction of the (known) $\mathbf{N}_{\mathcal{S}}(t^*, t_1)$ given $N(t^*)$ as precise as possible. These estimates will differ according to the value of t^* and the discrepancy chosen, but a possible choice is

$$(\hat{a}_{dis}, \hat{b}_{dis}) = \arg_{(a,b)}[minD(a,b)],$$

where

$$\begin{aligned} D(a,b) &= \sum_{i=1}^{k} \left(N_i(t_z^*, t_{1i}) - \widehat{N}_i(t_z^*, t_{1i})\right)^2 \\ &= \sum_{i=1}^{k} \left(N_i(t_z^*, t_{1i}) - \mathbb{E}[\mathbf{N}_i(t_z^*, t_{1i}) | N_i(t_z^*)]\right)^2 \\ &= \sum_{i=1}^{k} \left[N_i(t_z^*, t_{1i}) - \left(\frac{a + N_i(t_z^*)}{b + t_z^*}\right)(t_{1i} - t_z^*)\right]^2 \end{aligned} \tag{3.9}$$

and

$$t_z^* = \min\{t : \sum_{i=1}^{k} N_i(t) \geq z \sum_{i=1}^{k} N_i(t_{1i}) \text{ and } t \leq \min(t_{1i})\}.$$

If the choice of the preceding quadratic discrepancy appears reasonable for point prediction problems, the optimal choice of t_z^*, with $0 < z < 1$, requires more investigation.

It is expected that $\hat{a}_{dis}$ and $\hat{b}_{dis}$ should be close to $\hat{a}_{mle}$ and $\hat{b}_{mle}$ when the $N_i(t_{1i})$'s are large. Furthermore, we may encounter situations where these values will make the prediction model more robust to some types of misspecifications. For example, if the processes are not time homogeneous and have rates that are decreasing over time, it is expected that $\hat{a}_{dis}/\hat{b}_{dis}$ would be smaller than $\hat{a}_{mle}/\hat{b}_{mle}$, which should reduce the potential over-prediction.

We now have presented 3 methods to find estimates that will completely specify our random effects model: m.l.e.'s, moment matching estimates, and estimates minimizing a given discrepancy. However, if one does not wish to use plug-in methods, the only alternative seems to be to use a Bayesian model with prior distributions on a and b. It is highly likely that any choice of prior will require numerical integration to find the predictive density function (3.6); Ngai & Stroud (1994) proposed distributions that are computationally convenient for a similar model. Note also that priors on a and b will often be non-informative given the rare availability of a state of knowledge for parameters of an unobservable random quantity.

3.4 Simulations

In this section, we will study and compare, through extensive simulations, all the prediction methods presented so far. One of the main objectives will be to evaluate the robustness

of these methods with respect to different types of misspecifications. All the processes simulated will be HPP's but a lot of different types of rates will be used. They will be either small or large, identical or different, fixed or random, and when they are random they will be either gamma or non-gamma.

3.4.1 Point Prediction

For different sets of fixed $\lambda = \{\lambda_i : i = 1, \ldots, k\}$, we generated $B = 2,000$ samples from $k = 20$ Poisson processes. We assumed that these processes were observed up to the times $t_1 = \{5, 5.5, \ldots, 9.5, 10, \ldots, 14.0, 14.5\}$ and that we want to predict the number of occurrences for each process up to the times $t_2 = \{12.5, 12.5, \ldots, 12.5, 17.5, \ldots, 17.5, 17.5\}$. These numbers were chosen in order to represent different values of $t_{2i} - t_{1i}$. Because the observation times can always be rescaled into different units of time, their magnitudes are of no particular interest.

In order to compare the ability of each point prediction method for $\mathbf{N}_i(t_{1i}, t_{2i})$, we analyzed the following discrepancy

$$D = \sqrt{\frac{\sum_{i=1}^{k} \left(N_i(t_{1i}, t_{2i}) - \widehat{N}_i(t_{1i}, t_{2i}) \right)^2}{k}}, \tag{3.10}$$

where $\widehat{N}_i(t_{1i}, t_{2i})$ is the point predictor provided by the method chosen. The value of D represents, for a given sample of k processes, the root mean square error between the real value of $N_i(t_{1i}, t_{2i})$ and its predictor.

First, simulations were produced for 8 sets of fixed λ, they were arbitrarily chosen to reflect different orders of heterogeneity and magnitude (*cf.* Table 3.1). For each of these sets, 6 methods were compared: a plug-in method assuming Poisson processes with identical rates, another plug-in method assuming Poisson processes with different rates, a

λ	$\overline{\lambda}$	S^2_λ
$\{\lambda_i = 1\ \forall i\}$	1	0
$\{0.5, 0.55, \ldots, 1.5\}\backslash\{1\}$	1	0.1
$\{\lambda_i = \frac{2}{21} i\}$	1	0.3
$\{\lambda_i = \frac{2}{287} i^2\}$	1	0.8
$\{\lambda_i = 10\ \forall i\}$	10	0
$\{9.5, 9.55, \ldots, 10.5\}\backslash\{10\}$	10	0.1
$\{\lambda_i = 9 + \frac{2}{21} i\}$	10	0.3
$\{\lambda_i = 9 + \frac{2}{287} i^2\}$	10	0.8

Table 3.1: Sets of fixed rates used in the simulations.

Bayesian method based on Jeffreys' non-informative prior assuming that the λ_i's are i.i.d. with density

$$\pi(\lambda_i) \quad = \quad \frac{1}{\sqrt{\lambda_i}},$$

and the three methods discussed in Section 3.3.2. The formula used to obtain $\widehat{N}_i(t_{1i}, t_{2i})$ for each of these methods is given in Table 3.2. The first two methods use their respective best unbiased predictors (*cf.* Proposition 3.3). The Bayesian method uses the posterior expectation of $\mathbf{N}_i(t_{1i}, t_{2i})$ given $N_i(t_{1i})$. Finally, the predictors for the 3 methods using random effects are the plug-in predictors derived from (3.8). Table 3.2 also gives the symbols used throughout this section to designate each of the methods discussed.

The results of some of these simulations are presented in Table 3.3. For a given set λ and a given method, each cell contains the average value of (3.10) over the 2,000 samples. The last column contains results assuming the full knowledge of λ (*i.e.* $\widehat{N}_i(t_{1i}, t_{2i}) = (t_{2i} - t_{1i})\lambda_i$). Note that the discrepancies for the $\mathcal{G}(\frac{1}{2}, 0)$ method are not presented; they were always

Method	Notation	$\widehat{N}_i(t_{1i}, t_{2i})$
$\mathbf{N}_i(t) \sim \mathcal{PP}(\widehat{\lambda})$	$\mathcal{P}(\widehat{\lambda})$	$(t_{2i} - t_{1i}) \left(\frac{\sum_{i=1}^{k} N_i(t_{1i})}{\sum_{i=1}^{k} t_{1i}} \right)$
$\mathbf{N}_i(t) \sim \mathcal{PP}(\widehat{\lambda}_i)$	$\mathcal{P}(\widehat{\lambda}_i)$	$(t_{2i} - t_{1i}) \left(\frac{N_i(t_{1i})}{t_{1i}} \right)$
$\mathbf{N}_i(t)\|\lambda_i \sim \mathcal{PP}(\lambda_i)$ $\pi(\lambda_i) = 1/\sqrt{\lambda_i}$	$\mathcal{G}(\frac{1}{2}, 0)$	$(t_{2i} - t_{1i}) \left(\frac{0.5 + N_i(t_{1i})}{0 + t_{1i}} \right)$
$\mathbf{N}_i(t)\|\lambda_i \sim \mathcal{PP}(\lambda_i)$ $\lambda_i \sim$Gamma$(\hat{a}_{mle}, \hat{b}_{mle})$	$\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$	$(t_{2i} - t_{1i}) \left(\frac{\hat{a}_{mle} + N_i(t_{1i})}{\hat{b}_{mle} + t_{1i}} \right)$
$\mathbf{N}_i(t)\|\lambda_i \sim \mathcal{PP}(\lambda_i)$ $\lambda_i \sim$Gamma$(\hat{a}_{mm}, \hat{b}_{mm})$	$\mathcal{G}(\hat{a}_{mm}, \hat{b}_{mm})$	$(t_{2i} - t_{1i}) \left(\frac{\hat{a}_{mm} + N_i(t_{1i})}{\hat{b}_{mm} + t_{1i}} \right)$
$\mathbf{N}_i(t)\|\lambda_i \sim \mathcal{PP}(\lambda_i)$ $\lambda_i \sim$Gamma$(\hat{a}_{dis}, \hat{b}_{dis})$	$\mathcal{G}(\hat{a}_{dis}, \hat{b}_{dis})$	$(t_{2i} - t_{1i}) \left(\frac{\hat{a}_{dis} + N_i(t_{1i})}{\hat{b}_{dis} + t_{1i}} \right)$

Table 3.2: Predictor for each method.

$Poisson(\lambda_i)$	$\mathcal{P}(\hat{\lambda})$	$\mathcal{P}(\hat{\lambda}_i)$	$\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$	$\mathcal{G}(\hat{a}_{mm}, \hat{b}_{mm})$	$\mathcal{G}(\hat{a}_{dis}, \hat{b}_{dis})$	True
$\overline{\lambda} = 1,\ S_\lambda^2 = 0$	**2.284**	2.919	2.297	2.603	2.341	2.248
$\overline{\lambda} = 1,\ S_\lambda^2 = 0.1$	2.918	2.807	2.670	**2.605**	2.929	2.213
$\overline{\lambda} = 1,\ S_\lambda^2 = 0.3$	3.992	**2.651**	2.681	2.664	3.038	2.163
$\overline{\lambda} = 1,\ S_\lambda^2 = 0.8$	5.239	**2.554**	**2.557**	2.571	2.562	2.120
$\overline{\lambda} = 10,\ S_\lambda^2 = 0$	**7.190**	9.221	7.225	8.145	7.359	7.075
$\overline{\lambda} = 10,\ S_\lambda^2 = 0.1$	**7.498**	9.294	7.534	8.152	7.747	7.171
$\overline{\lambda} = 10,\ S_\lambda^2 = 0.3$	7.971	9.285	**7.911**	8.087	8.195	7.206
$\overline{\lambda} = 10,\ S_\lambda^2 = 0.8$	8.723	9.282	8.279	**8.197**	8.783	7.183

Table 3.3: Comparison of the discrepancies of point predictors.

similar to those obtained with the $\mathcal{P}(\hat{\lambda}_i)$ method. In this table, the smallest average discrepancy for a given set λ is written in bold font. For the cases where another method provides an expected discrepancy that is not significantly bigger, at the 1% significance level, this discrepancy is also written in bold font. These tests use the asymptotic multivariate normal distribution of each row of average discrepancies. Even if these tests do not take into account that we are actually doing multiple comparisons, it does not greatly affect the conclusions since a difference as small as 0.01 between two discrepancies is usually (statistically) significant. This is due to the fact that two discrepancies usually have a strong positive correlation.

The first thing we notice in Table 3.3 is the good performance of the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method: even when this method does not provide the smallest average discrepancy, its value is always close to the smallest one. This method seems to predict well for any values of $\overline{\lambda}$ and S_λ^2. Note that since the λ_i's are fixed, the notations $\overline{\lambda}$ and S_λ^2 are used to represent

the average and the heterogeneity of the λ_i's. We can also see in this table that the $\mathcal{P}(\hat{\lambda})$ and $\mathcal{P}(\hat{\lambda}_i)$ methods are not robust to the rate homogeneity/heterogeneity assumption. The $\mathcal{P}(\hat{\lambda})$ method, which assumes that all the λ_i's are identical, gives very high discrepancies when the rates are highly heterogeneous ($\overline{\lambda} = 1$ with S^2_λ =0.3 or 0.8) and the $\mathcal{P}(\hat{\lambda}_i)$ method, which assumes that all the λ_i's are different, fails to outperform the $\mathcal{P}(\hat{\lambda})$ method when the rates are moderately different ($\overline{\lambda} = 10$ with S^2_λ =0.1, 0.3 or 0.8). A good performance of the $\mathcal{P}(\hat{\lambda})$ and $\mathcal{P}(\hat{\lambda}_i)$ methods when the homogeneity/heterogeneity of the rates was incorrectly assumed was not expected, but the behavior of the $\mathcal{P}(\hat{\lambda}_i)$ method when $\overline{\lambda} = 10$ is still surprising. With a large expected number of observed events per process ranging from 45.0 ($\min t_{1i}\lambda_i = 5 \times 9.007$) to 170.9 ($\max t_{1i}\lambda_i = 15 \times 11.787$), it was expected that this method could provide $\hat{\lambda}_i$'s that are close to the true λ_i's.

The robustness of the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method to the rate homogeneity/heterogeneity assumption is easier to explain when we rewrite its predictor the following way:

$$\begin{aligned}
\widehat{N}_i(t_{1i}, t_{2i}) &= (t_{2i} - t_{1i}) \left(\frac{\hat{a}_{mle} + N_i(t_{1i}, t_{2i})}{\hat{b}_{mle} + t_{1j}} \right) \\
&= (t_{2i} - t_{1i}) \left[\frac{\hat{b}_{mle}}{\hat{b}_{mle} + t_{1i}} \left(\frac{\hat{a}_{mle}}{\hat{b}_{mle}} \right) + \frac{t_{1i}}{\hat{b}_{mle} + t_{1i}} \left(\frac{N_i(t_{1i})}{t_{1i}} \right) \right] \\
&= (t_{2i} - t_{1i}) \left[w_i \left(\frac{\hat{a}_{mle}}{\hat{b}_{mle}} \right) + (1 - w_i)\hat{\lambda}_i \right], \qquad (3.11)
\end{aligned}$$

where $w_i = \hat{b}_{mle}/(\hat{b}_{mle} + t_{1i})$. When the processes are $\mathcal{PP}(\lambda_i)$, we know that $\hat{\lambda}_i \xrightarrow{a.s.} \lambda_i$ as $N_i(t_{1i})$ goes to infinity, and when the processes are $\mathcal{PP}(\lambda)$, $\hat{a}_{mle}/\hat{b}_{mle} \xrightarrow{a.s.} \lambda$ as all the $N_i(t_{1i})$'s go to infinity. The predictor is then a mixture between estimates of the expectation of $\mathbf{N}_i(t_{1i}, t_{2i})$ under the homogeneous and the heterogeneous assumption. Furthermore, the weight w_i is close to one when $\hat{b}_{mle}$ is large, which is usually the case when the sample suggests that the rates are homogeneous. This feature is very similar to the shrinkage property of empirical Bayes point estimators Carlin & Louis (1996, Section 3.3).

Table 3.3 also compares the three methods using random effects. If the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method seems to be the most suitable one, the $\mathcal{G}(\hat{a}_{mm}, \hat{b}_{mm})$ method also performs very well when the rates are reasonably heterogeneous. The third method uses the estimates $\hat{a}_{dis}$ and $\hat{b}_{dis}$, which minimize (3.9) with z arbitrarily chosen to be 0.25. This method seems to predict adequately but not as well as the other two methods.

We also compared the average discrepancies of the $\mathcal{P}(\hat{\lambda})$, $\mathcal{P}(\hat{\lambda}_i)$, and $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ methods when the λ_i's are actual random variables. The results are presented in Table 3.4. The first thing we notice is that the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method is the one having the best performance to predict the $\mathbf{N}_i(t_{1i}, t_{2i})$'s but this was expected since this method actually treats λ as a random vector. However, a very interesting feature of the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method is that it seems robust to the real distribution of the random effects. In our simulations, we can actually show that there are no (statistically) significant differences, at the 1% level, between the discrepancies found with the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method when the rates are gamma, log-normal, or Weibull. We used these two other distributions because they are often used to model non-negative random variables.

In conclusion, it appears from these simulations that the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method is effective to find point predictors of the number of occurrences coming from HPP's. It seems very robust to the rate homogeneity/heterogeneity assumption and it seems also robust to the type of distribution of the rates when they are random variables.

3.4.2 Set Prediction

All the methods studied in this section can also be expressed via a predictive distribution for $\mathbf{N}_i(t_{1i}, t_{2i})$: a Poisson distribution with rate $(t_{2i} - t_{1i})\hat{\lambda}$ and $(t_{2i} - t_{1i})\hat{\lambda}_i$ respectively for the $\mathcal{P}(\hat{\lambda})$ and $\mathcal{P}(\hat{\lambda}_i)$ methods and a negative binomial distribution with parameters $(N_i(t_{1i}) + 0.5, t_{1i}/t_{2i})$ and $(N_i(t_{1i}) + \hat{a}_{mle}, (\hat{b}_{mle} + t_{1i})/(\hat{b}_{mle} + t_{2i}))$ respectively for the $\mathcal{G}(\frac{1}{2}, 0)$

Moments for λ_i	Distribution for λ_i	$\mathcal{P}(\hat{\lambda})$	$\mathcal{P}(\hat{\lambda}_i)$	$\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$	True
$\mathbb{E}[\lambda_i] = 1\,,\ \mathbb{V}ar[\lambda_i] = 0.1$	Gamma	2.830	2.961	**2.637**	2.257
	Log-normal	2.830	2.914	**2.608**	2.253
	Weibull	2.840	2.925	**2.635**	2.262
$\mathbb{E}[\lambda_i] = 1\,, \mathbb{V}ar[\lambda_i] = 0.3$	Gamma	3.756	2.956	**2.799**	2.286
	Log-normal	3.638	2.948	**2.775**	2.236
	Weibull	3.708	2.879	**2.731**	2.236
$\mathbb{E}[\lambda_i] = 1\,, \mathbb{V}ar[\lambda_i] = 0.8$	Gamma	5.127	2.907	**2.837**	2.238
	Log-normal	4.934	2.904	**2.820**	2.230
	Weibull	5.179	2.882	**2.820**	2.233
$\mathbb{E}[\lambda_i] = 10\,, \mathbb{V}ar[\lambda_i] = 0.1$	Gamma	**7.466**	9.329	7.492	7.179
	Log-normal	**7.405**	9.318	7.446	7.129
	Weibull	**7.561**	9.411	7.600	7.268
$\mathbb{E}[\lambda_i] = 10\,, \mathbb{V}ar[\lambda_i] = 0.3$	Gamma	7.816	9.212	**7.744**	7.119
	Log-normal	7.842	9.352	**7.821**	7.153
	Weibull	7.804	9.258	**7.765**	7.125
$\mathbb{E}[\lambda_i] = 10\,, \mathbb{V}ar[\lambda_i] = 0.8$	Gamma	8.652	9.408	**8.261**	7.188
	Log-normal	8.619	9.309	**8.182**	7.163
	Weibull	8.588	9.174	**8.100**	7.121

Table 3.4: Comparison of the discrepancies using different distributions for the rates.

Poisson(λ_i)	$\mathcal{P}(\hat{\lambda})$	$\mathcal{P}(\hat{\lambda}_i)$	$\mathcal{G}(\frac{1}{2}, 0)$	$\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$
$\overline{\lambda} = 1,\ S^2_\lambda = 0$	**0.30**	7.22	4.75	0.36
$\overline{\lambda} = 1,\ S^2_\lambda = 0.1$	6.24	6.95	4.63	**3.45**
$\overline{\lambda} = 1,\ S^2_\lambda = 0.3$	21.83	5.65	**4.45**	4.90
$\overline{\lambda} = 1,\ S^2_\lambda = 0.8$	47.93	5.11	**4.06**	4.46
$\overline{\lambda} = 10,\ S^2_\lambda = 0$	**0.26**	6.17	4.61	0.38
$\overline{\lambda} = 10,\ S^2_\lambda = 0.1$	**0.84**	6.28	4.61	0.93
$\overline{\lambda} = 10,\ S^2_\lambda = 0.3$	2.07	6.31	4.60	**1.79**
$\overline{\lambda} = 10,\ S^2_\lambda = 0.8$	4.33	6.26	4.69	**2.65**

Table 3.5: Comparison of the average KL distance for different methods.

and $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ methods. Therefore, when we know the real distribution of $\mathbf{N}_i(t_{1i}, t_{2i})$, we can compare the distance between these predictive densities and the real density of $\mathbf{N}_i(t_{1i}, t_{2i})$. This should give us an indication of the ability of each method to provide adequate prediction intervals. The distance used will be the average Kullback-Liebler distance given in (2.3) and it will be estimated by simulating $B = 2,000$ samples of $k = 20$ Poisson processes. Let $N^*_j(t_1)$ and $N^*_j(t_1, t_2)$ be the counts generated for the jth sample, $j = 1, \ldots, B$. The average KL distance is then estimated by

$$\widehat{\mathcal{D}}\left(\tilde{f}_p(N(t_1,t_2)|N(t_1)), f(N(t_1,t_2);\lambda)\right) \;=\; \sum_{j=1}^{B} \log\left[\frac{f(N^*_j(t_1,t_2);\lambda)}{\tilde{f}_p(N^*_j(t_1,t_2)|N^*_j(t_1))}\right],$$

where $f(N^*_j(t_1, t_2); \lambda)$ is the joint density of k Poisson variables and $\tilde{f}_p(N^*_j(t_1, t_2)|N^*_j(t_1))$ is the predictive density obtained from each of the 4 methods mentioned above.

The results of these simulations are given in Table 3.5. We can see that the two methods adding extra variability on the fixed λ are usually performing better than the other two

Moments for λ_i	$\mathcal{P}(\hat{\lambda})$	$\mathcal{P}(\hat{\lambda}_i)$	$\mathcal{G}(\frac{1}{2},0)$	$\mathcal{G}(\hat{a}_{mle},\hat{b}_{mle})$
$\mathbb{E}[\lambda_i]=1\,,\mathbb{V}ar[\lambda_i]=0.1$	2.844	4.814	2.304	**0.435**
$\mathbb{E}[\lambda_i]=1\,,\mathbb{V}ar[\lambda_i]=0.3$	11.872	3.430	1.168	**0.206**
$\mathbb{E}[\lambda_i]=1\,,\mathbb{V}ar[\lambda_i]=0.8$	30.916	2.340	0.491	**0.136**
$\mathbb{E}[\lambda_i]=10\,,\mathbb{V}ar[\lambda_i]=0.1$	**0.313**	6.012	4.267	0.406
$\mathbb{E}[\lambda_i]=10\,,\mathbb{V}ar[\lambda_i]=0.3$	2.672	5.105	3.481	**0.489**
$\mathbb{E}[\lambda_i]=10\,,\mathbb{V}ar[\lambda_i]=0.8$	2.142	4.100	2.458	**0.424**

Table 3.6: Comparison of the average KL when the rates are random.

methods. We also see that when the $\mathcal{G}(\hat{a}_{mle},\hat{b}_{mle})$ method does not provide the closest predictive density, it is never far out, which is not the case for the predictive density of the $\mathcal{G}(\frac{1}{2},0)$ method. Because they ignore the uncertainty about λ, the $\mathcal{P}(\hat{\lambda})$ and $\mathcal{P}(\hat{\lambda}_i)$ methods are not expected to have predictive densities close to the true one when the number of occurrences is small. However, the poor performance of the $\mathcal{P}(\hat{\lambda}_i)$ when $\overline{\lambda}=10$ indicates that even when the expected number of events per observations is between 45.0 and 170.9, the uncertainty about λ cannot be neglected. Another problem with the $\mathcal{P}(\hat{\lambda}_i)$ method is the unavailability of a non-degenerate predictive density for $N_i(t_{1i},t_{2i})$ when $N_i(t_{1i})=0$ (*i.e.* $\hat{\lambda}_i=0/t_{1i}$). In our simulations the problem was avoided by using $0.5/t_{1i}$ instead of 0.

When the rates are actual random variables, we can see from Table 3.6 that the $\mathcal{G}(\hat{a}_{mle},\hat{b}_{mle})$ method usually outperforms the other 3 methods. Note that since the rates are Gamma(a,b) the function $f(N(t_1,t_2);\lambda)$ must be replaced by $f(N(t_1,t_2)|N(t_1);a,b)$, the joint density of k negative binomials.

If predictive densities are of interest, the effectiveness of a method to find adequate

prediction intervals should also be assessed by checking its ability to have the desired coverage probability. When the rates are fixed, it is expected that the $\mathcal{P}(\hat{\lambda})$ and the $\mathcal{P}(\hat{\lambda}_i)$ methods will have coverage probabilities below the desired level, especially when the number of occurrences is insufficient to neglect the uncertainty about λ. As for the $\mathcal{G}(\frac{1}{2}, 0)$ and the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ methods, it is not clear what effect the introduction of a random distribution on a fixed λ will have on their coverage probabilities. Note that since it is usually computationally intense to calibrate a single dataset, calibration was not performed on the thousands of samples generated for our simulations. Nevertheless, we will then find the methods which are the most likely to have coverage proportions close to the desired levels. Therefore, these methods will have to be calibrated less often and their prediction intervals will then be obtained more rapidly.

To estimate the actual coverage probability of the prediction intervals, we simulated $B = 2,000$ samples of $k = 20$ Poisson processes for different sets of fixed λ. For each of the processes simulated, we calculated a one-sided 90% prediction interval of the form $[0, L(N(t_1))]$ for each method already investigated and the one suggested by Faulkenberry (1973) and presented in Section 3.2.2. The quantity to predict being discrete, each interval was randomized. The results are contained in Table 3.7. In each cell, we have the proportion of the $2,000 \times 20 = 40,000$ counts that where included in the corresponding 90% prediction interval and the number in parentheses corresponds to the average length of these intervals. The first thing we notice is how close to 0.9 the coverage proportions are when the $\mathcal{G}(\frac{1}{2}, 0)$ and $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ methods are used. In both cases, it looks like the extra-variability added was relatively appropriate to calibrate these intervals. When $\overline{\lambda} = 10$, we can also see that all the methods except $\mathcal{P}(\hat{\lambda}_i)$ have a coverage proportion very close to the desired level. However, the $\mathcal{P}(\hat{\lambda})$ and the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ methods seem to use shorter intervals to reach the desired proportion. The results are different when $\overline{\lambda} = 1$. Except

Coverage Probability	$\mathcal{P}(\hat{\lambda})$	$\mathcal{P}(\hat{\lambda}_i)$	Faulk.	$\mathcal{G}(\frac{1}{2}, 0)$	$\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$
$\overline{\lambda} = 1,\ S_\lambda^2 = 0$	0.897 (7.774)	0.825 (7.738)	0.873 (8.605)	0.897 (9.035)	0.899 (7.836)
$\overline{\lambda} = 1,\ S_\lambda^2 = 0.1$	0.894 (8.397)	0.823 (7.472)	0.865 (8.231)	0.894 (8.689)	0.906 (8.385)
$\overline{\lambda} = 1,\ S_\lambda^2 = 0.3$	0.869 (8.878)	0.844 (7.225)	0.854 (7.790)	0.905 (8.296)	0.911 (8.288)
$\overline{\lambda} = 1,\ S_\lambda^2 = 0.8$	0.825 (9.431)	0.864 (6.889)	0.858 (7.194)	0.912 (7.765)	0.909 (7.639)
$\overline{\lambda} = 10,\ S_\lambda^2 = 0$	0.896 (61.285)	0.837 (61.283)	0.893 (63.986)	0.899 (64.330)	0.897 (61.467)
$\overline{\lambda} = 10,\ S_\lambda^2 = 0.1$	0.905 (61.824)	0.840 (61.085)	0.893 (63.766)	0.899 (64.112)	0.907 (61.993)
$\overline{\lambda} = 10,\ S_\lambda^2 = 0.3$	0.905 (62.238)	0.843 (60.884)	0.896 (63.550)	0.902 (63.896)	0.910 (62.421)
$\overline{\lambda} = 10,\ S_\lambda^2 = 0.8$	0.902 (62.770)	0.844 (60.718)	0.897 (63.369)	0.903 (63.717)	0.914 (62.936)

Table 3.7: Coverage proportions (and average lengths) of 90% prediction intervals.

for the $\mathcal{P}(\hat{\lambda})$ method when S_λ^2 is small, both the methods that do not use random effects have a coverage proportion under 90%. The poor performance of the method suggested by Faulkenberry (1973) is probably due to the fact that this method cannot be randomized properly.

It should be pointed out here that when the λ_i's are identical (*i.e.* $S_\lambda^2 = 0$), the m.l.e.'s $\hat{a}_{mle}$ and $\hat{b}_{mle}$ obtained are often very large ($> 10^6$). Because $\lambda_i = \lambda \ \forall i$, it is not surprising to get estimates such that $\hat{\mu} = \hat{a}_{mle}/\hat{b}_{mle}$ is close to λ and $\hat{\sigma}^2 = \hat{a}_{mle}/\hat{b}_{mle}^2$ is close to 0. However, when $\hat{a}_{mle}$ and $\hat{b}_{mle}$ are very large, it may not be possible to get a precise value of the predictive density function of $\mathbf{N}_i(t_{1i}, t_{2i})$ via any statistical software. Nevertheless, we can still obtain appropriate prediction intervals by finding quantiles of a Poisson distribution that is equivalent to the negative binomial when $\hat{a}_{mle}$ and $\hat{b}_{mle}$ are large:

$$
\begin{aligned}
P[\mathbf{N}_i(t_{1i}, t_{2i}) = n | N_i(t_{1i}); \hat{a}, \hat{b}] &= \frac{\Gamma(\hat{a} + N_i(t_{1i}) + n)}{\Gamma(\hat{a} + N_i(t_{1i}))n!} \left(\frac{t_{2i} - t_{1i}}{\hat{b} + t_{2i}}\right)^n \left(\frac{\hat{b} + t_{1i}}{\hat{b} + t_{2i}}\right)^{\hat{a}+N_i(t_{1i})} \\
&= \frac{(t_{2i} - t_{1i})^n}{n!} \left(\frac{\prod_{l=0}^{n-1}(\hat{a} + N_i(t_{1i}) + l)}{(\hat{b} + t_{2i})^n} \right) \times \\
&\quad \left(\frac{\hat{b} + t_{2i}}{\hat{b} + t_{1i}}\right)^{-(\hat{a}+N_i(t_{1i}))} \\
&\simeq \frac{(t_{2i} - t_{1i})^n}{n!} \left(\frac{\hat{a}}{\hat{b}}\right)^n \left(1 + \frac{t_{2i} - t_{1i}}{\hat{b}}\right)^{-\hat{a}} \\
&= \frac{(t_{2i} - t_{1i})^n}{n!} \left(\frac{\hat{a}}{\hat{b}}\right)^n \left(1 + \frac{\frac{\hat{a}}{\hat{b}}(t_{2i} - t_{1i})}{\hat{a}} \right)^{-\hat{a}} \\
&\simeq P\left[\text{Poisson}\left(\frac{\hat{a}}{\hat{b}}(t_{2i} - t_{1i})\right) = n\right].
\end{aligned}
$$

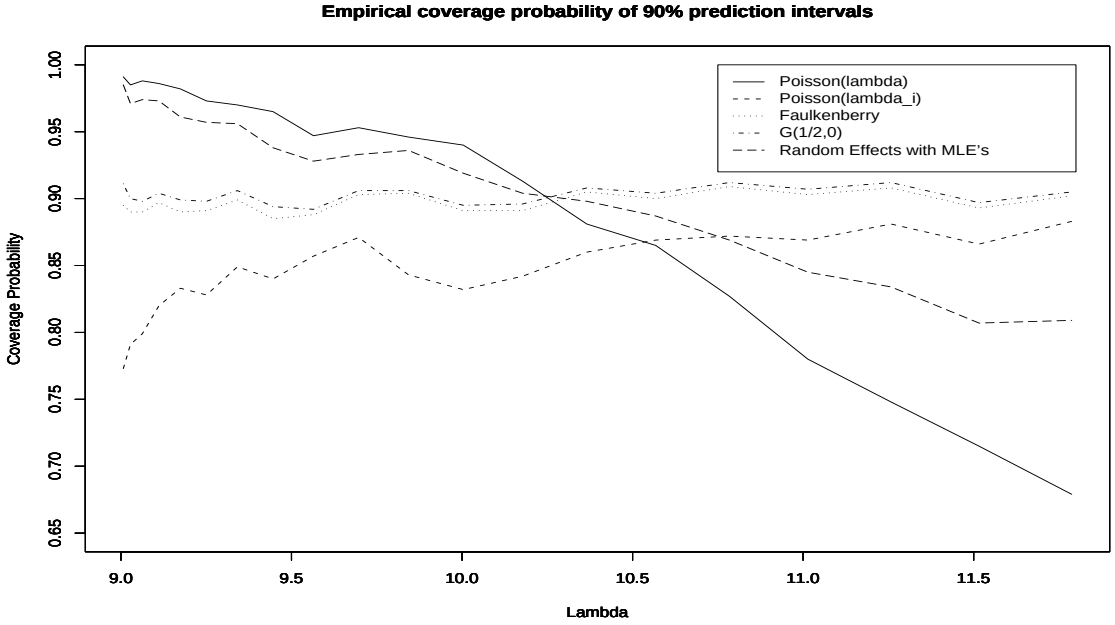

Figure 3.1: Empirical coverage probabilities of 90% prediction intervals.

A deeper study of these simulations also reveals a very interesting feature of the $\mathcal{G}(\frac{1}{2}, 0)$ method: when the rates are fixed but different, a study of the coverage proportion with respect to the real value of the rates indicates that while the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method has a coverage proportion that decreases when the value of λ_i increases, the $\mathcal{G}(\frac{1}{2}, 0)$ method seems robust to the value of λ_i. Figure 3.1 presents this result for the case where $\overline{\lambda} = 10$ and $S^2_\lambda = 0.8$. It seems to indicate an undesirable behavior for the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method: it provides an adequate prediction for an individual count when population average criteria are considered (KL distance and coverage proportions) but does not necessarily predict well for each individual unit. This feature can also probably explain why the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method usually gives shorter intervals than the $\mathcal{G}(\frac{1}{2}, 0)$ method even if they usually have similar coverage proportions: the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method is covering more counts when the

λ_i's are small, but it is covering fewer counts when longer prediction intervals are needed.

When the rates are unobservable random variables, the 5 methods studied in Table 3.7 should have coverage probabilities below the desired level. Nevertheless, the $\mathcal{G}(\frac{1}{2}, 0)$ and $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ methods should give better results since they treat the rates as random. This is what we see in the simulations presented in Table 3.8. When the rates are random, the 5 methods give coverage proportions below 90%. However, it is interesting to see that there is a small loss of precision when we neglect the uncertainty about the random effects distribution: the coverage proportions given by the $\mathcal{G}(\frac{1}{2}, 0)$ and $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ methods are never smaller than 89%. Also, it is interesting to note that these 2 methods give similar results whether the rates are gamma, log-normal, or Weibull. The average lengths of the prediction intervals were calculated but are not included in this table, the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method always gave intervals having a smaller average length than the $\mathcal{G}(\frac{1}{2}, 0)$ method.

In conclusion, the preceding simulations revealed very interesting properties of the $\mathcal{G}(\hat{a}_{mle}, \hat{b}_{mle})$ method. When the rates are fixed, this method gives relatively precise predictors whether the rates are identical or different, and the variability added by this method seems to give a reasonable calibration. When the rates are random, the real distribution of the rates seems to have a small influence on this method. Finally, the effect of replacing the unknown parameters by m.l.e.'s appears to be negligible.

Moments for λ_i	Distribution for λ_i	$\mathcal{P}(\hat{\lambda})$	$\mathcal{P}(\hat{\lambda}_i)$	Faulk.	$\mathcal{G}(\frac{1}{2},0)$	$\mathcal{G}(\hat{a}_{mle},\hat{b}_{mle})$
$\mathbb{E}[\lambda_i]=1\,,\mathbb{V}ar[\lambda_i]=0.1$	Gamma	0.855	0.826	0.872	0.897	0.889
	Log-normal	0.859	0.826	0.873	0.898	0.891
	Weibull	0.854	0.825	0.870	0.897	0.892
$\mathbb{E}[\lambda_i]=1\,,\mathbb{V}ar[\lambda_i]=0.3$	Gamma	0.819	0.827	0.869	0.898	0.896
	Log-normal	0.829	0.824	0.868	0.897	0.894
	Weibull	0.808	0.828	0.865	0.897	0.893
$\mathbb{E}[\lambda_i]=1\,,\mathbb{V}ar[\lambda_i]=0.8$	Gamma	0.788	0.835	0.861	0.901	0.898
	Log-normal	0.811	0.827	0.862	0.896	0.895
	Weibull	0.785	0.837	0.864	0.904	0.897
$\mathbb{E}[\lambda_i]=10\,,\mathbb{V}ar[\lambda_i]=0.1$	Gamma	0.893	0.840	0.895	0.901	0.897
	Log-normal	0.891	0.839	0.892	0.898	0.895
	Weibull	0.890	0.841	0.894	0.900	0.896
$\mathbb{E}[\lambda_i]=10\,,\mathbb{V}ar[\lambda_i]=0.3$	Gamma	0.883	0.841	0.893	0.899	0.894
	Log-normal	0.881	0.842	0.894	0.900	0.893
	Weibull	0.881	0.839	0.893	0.898	0.893
$\mathbb{E}[\lambda_i]=10\,,\mathbb{V}ar[\lambda_i]=0.8$	Gamma	0.827	0.781	0.898	0.908	0.869
	Log-normal	0.858	0.836	0.891	0.898	0.888
	Weibull	0.862	0.841	0.893	0.899	0.891

Table 3.8: Coverage proportions of 90% prediction intervals (random rates).

Chapter 4

Prediction Models for Nonhomogeneous Poisson Processes

In this chapter, we will study prediction models for Poisson processes where the time homogeneity assumption is relaxed. After defining such nonhomogeneous Poisson processes (NHPP's), we will propose some point predictors and prediction intervals for this type of process. Finally, we will see that using a random effects model can improve the precision of the predictions, when there is a sufficient degree of heterogeneity across the processes.

Throughout this chapter, the rate function defined in (3.1) will always be written under the parametric form

$$\lambda(t;\alpha,\beta) = \alpha f(t;\beta),$$

where α is a scalar and β is a vector of low dimension. Using this parameterization, the number of events in the time interval $[t_1, t_2]$, still denoted by $N(t_1, t_2)$ or $N(t_2)$ when $t_1 = 0$, has a Poisson distribution with rate $\alpha(F(t_2;\beta) - F(t_1;\beta))$, where $F(t;\beta) = \int_0^t f(u;\beta)du$.

This parameterization is convenient because $f(t;\beta)$ and α both measure different as-

pects of a NHPP. The function $f(t;\beta)$ describes the behavior of the rate function with respect to time. When $f(t;\beta)$ is an increasing (decreasing) function with respect to t events are likely to occur more (less) often as time goes by and when $f(t;\beta)$ is constant the process is a homogeneous Poisson process. Possible choices of functions f will be discussed in Section 4.3. On the other hand, α represents the expected number of events up to a certain horizon time T when $F(T;\beta) = 1$. For example, if $f(t;\beta)$ is a density function on $[0,\infty)$, α corresponds to $\mathbb{E}[\mathbf{N}(\infty)]$.

Like in the previous chapter, more than one Poisson process will be considered here. Rate heterogeneity between two (time) homogeneous Poisson processes can only be addressed by using two different scalars for their rates, but the heterogeneity between two NHPP's can be modeled in different ways. First, it may be reasonable to assume that different processes have the same behavior over time but not the same expected number of events. Such a scenario can be modeled using the rate function

$$\lambda_i(t;\alpha_i,\beta) = \alpha_i f(t;\beta), \tag{4.1}$$

where the subscript i denotes the ith process amongst k. Another way to model the heterogeneity is to consider models with rate function

$$\lambda_i(t;\alpha,\beta_i) = \alpha f(t;\beta_i). \tag{4.2}$$

Then, if $F(\infty;\beta) = 1$ for all β, the expected number of events is the same for all processes but their behaviors over time are different. Finally, we may want to use both types of heterogeneity jointly by using the parametric form

$$\lambda_i(t;\alpha_i,\beta_i) = \alpha_i f(t;\beta_i).$$

We believe that the first type of heterogeneity mentioned here can be applied to many practical situations. This model will therefore be considered often in this chapter.

4.1 Point Predictors and Prediction Intervals

Now we suppose that k processes with rate $\lambda_i(t;\alpha_i,\beta_i)$ are observed up to time t_{1i} and we wish to predict $\mathbf{N}_{\mathcal{S}}(t_1,t_2)$, the sum of all the $\mathbf{N}_i(t_{1i},t_{2i})$'s in $\mathcal{S}\subset\{1,\ldots,k\}$. Except for a few settings, it does not seem possible to find unbiased predictors for these processes. Nevertheless, since $\mathbf{N}_{\mathcal{S}}(t_1,t_2)$ has a Poisson distribution with rate $\sum_{i\in\mathcal{S}}\alpha_i(F(t_{2i};\beta_i)-F(t_{1i};\beta_i))$, a predictor can be obtained by estimating α and β. Although biased, the plug-in predictor

$$\widehat{\mathbf{N}}_{\mathcal{S}}(t_1,t_2)=\sum_{i\in\mathcal{S}}\hat{\alpha}_i(F(t_{2i};\hat{\beta}_i)-F(t_{1i};\hat{\beta}_i))$$

is adequate if the parameters are accurately estimated. In order to do so, we usually choose the parameters maximizing the likelihood

$$L(\alpha,\beta|(N(t_1),\tau))=\prod_{i=1}^{k}\left(\prod_{j=1}^{N_i(t_{1i})}\alpha_i f(\tau_{ij};\beta_i)\right)\exp\{-\alpha_i F(t_{1i};\beta_i)\},\tag{4.3}$$

where τ_{ij} is the time of the jth occurrence coming from the ith process. This likelihood function was easily derived from (3.3). Obviously, the m.l.e.'s will have different forms for each heterogeneity assumption. Table 4.1 lists the equations these estimates must satisfy for each form of heterogeneity. Note that the numbers of equations vary from $1+\dim(\beta)$ to $k(1+\dim(\beta))$ according to the form chosen. Clearly, the preciseness of the estimates will depend on the model selected. For example, model (4.1) should allow us to obtain an estimate for β that is more precise than those obtained for the α_i's, the reason being that all k processes are directly used to estimate β.

Similar problems are encountered when an exact prediction interval for $\mathbf{N}_{\mathcal{S}}(t_1,t_2)$ is sought. Besides some trivial cases, it does not seem possible to find an exact prediction interval using the methods presented in Section 1.2. An alternative is to construct an

Parameters	Score equations
(α, β)	$\hat{\alpha} = \dfrac{\sum_{i=1}^{k} N_i(t_{1i})}{\sum_{i=1}^{k} F(t_{1i};\hat{\beta})}$ $\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{N_i(t_{1i})} \dfrac{\frac{\partial}{\partial\beta} f(\tau_{ij};\hat{\beta})}{f(\tau_{ij};\hat{\beta})} = \sum_{i=1}^{k} N_i(t_{1i}) \left(\dfrac{\sum_{i=1}^{k} \frac{\partial}{\partial\beta} F(t_{1i};\hat{\beta})}{\sum_{i=1}^{k} F(t_{1i};\hat{\beta})} \right)$
$(\alpha_1, \ldots, \alpha_k, \beta)$	$\hat{\alpha}_i = \dfrac{N_i(t_{1i})}{F(t_{1i};\hat{\beta})}$ $\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{N_i(t_{1i})} \dfrac{\frac{\partial}{\partial\beta} f(\tau_{ij};\hat{\beta})}{f(\tau_{ij};\hat{\beta})} = \sum_{i=1}^{k} N_i(t_{1i}) \left(\dfrac{\frac{\partial}{\partial\beta} F(t_{1i};\hat{\beta})}{F(t_{1i};\hat{\beta})} \right)$
$(\alpha, \beta_1, \ldots, \beta_k)$	$\hat{\alpha} = \dfrac{\sum_{i=1}^{k} N_i(t_{1i})}{\sum_{i=1}^{k} F(t_{1i};\hat{\beta}_i)}$ $\displaystyle\sum_{j=1}^{N_i(t_{1i})} \dfrac{\frac{\partial}{\partial\beta_i} f(\tau_{ij};\hat{\beta}_i)}{f(\tau_{ij};\hat{\beta}_i)} = \sum_{l=1}^{k} N_l(t_{1l}) \left(\dfrac{\frac{\partial}{\partial\beta_i} F(t_{1i};\hat{\beta}_i)}{\sum_{l=1}^{k} F(t_{1l};\hat{\beta}_l)} \right)$
$(\alpha_1, \ldots, \alpha_k, \beta_1, \ldots, \beta_k)$	$\hat{\alpha}_i = \dfrac{N_i(t_{1i})}{F(t_{1i};\hat{\beta}_i)}$ $\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{N_i(t_{1i})} \dfrac{\frac{\partial}{\partial\beta_i} f(\tau_{ij};\hat{\beta}_i)}{f(\tau_{ij};\hat{\beta}_i)} = N_i(t_{1i}) \left(\dfrac{\frac{\partial}{\partial\beta_i} F(t_{1i};\hat{\beta}_i)}{F(t_{1i};\hat{\beta}_i)} \right)$

Table 4.1: Score equations for m.l.e.'s.

approximate prediction interval by using the fact that $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ has a Poisson distribution. For example, we can substitute the unknown parameters with their m.l.e.'s to obtain a plug-in prediction interval. Such an interval is expected to have a coverage probability below the desired one but it can be calibrated by simulating NHPP's with rates $\hat{\alpha}_i f(t; \hat{\beta}_i)$. Another type of approximate prediction interval for $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ is also available with a model like (4.2). We can modify Faulkenberry's (1973) approach (*cf.* Section 1.2.2) to find an exact prediction interval for a random variable having a Poisson distribution with rate $\sum_{i \in \mathcal{S}} \alpha(F(t_{2i}; \hat{\beta}_i) - F(t_{1i}; \hat{\beta}_i))$, a sufficient statistic for α now being available.

If Bayesian prediction intervals for NHPP's are sought, many different types of prior distributions are available in the literature (*c.f.* Kuo & Yang (1996) and Grandell (1997) for exhaustive lists). If the choice of a prior will clearly depend on the form of $f(t; \beta)$, some general points can be noted about Bayesian prediction approaches for NHPP's. First, it is probably a restrictive assumption, but since α and β usually measure different aspects of a Poisson process, they are usually modeled through independent priors. Also, due to the form of the likelihood in (4.3), the prior distribution on α (or on the α_i's according to the assumptions made) is usually gamma. Finally, we should point out that the introduction of a prior on β will often lead to the use of a numerical integration technique to obtain the predictive density.

4.2 Random Effects Model

When it is not reasonable to assume that the rates between the processes are identical, it is clear from Table 4.1 that estimates of α_i and β_i could be imprecise for some of the processes where only a little amount of information is available. Consequently, predictors and prediction intervals obtained using these estimates could be imprecise as well. There-

fore, we will present here a useful prediction model for NHPP's with different rates. This (rate) heterogeneity will be modeled using (4.1), where processes are affected by time in a similar manner but do not have the same expected number of occurrences over the same time interval.

As in Chapter 3, this model will use gamma random effects to incorporate heterogeneity between processes:

$$\begin{aligned} \mathbf{N_i}(t)|\alpha_i &\sim \mathcal{PP}(\alpha_i f(t;\beta)), \\ \alpha_i &\sim \text{Gamma}(a,b), \end{aligned} \tag{4.4}$$

where $i = 1,\ldots,k$. Predictions will be made using predictive densities derived from the density of $\mathbf{N}_{\mathcal{S}}(t_1,t_2)$ given $N(t_1)$ and the set of occurrence times $\tau = \{\tau_{ij} : i = 1,\ldots,k \text{ and } j = 1,\ldots,N_i(t_{1i})\}$. To find this density, we first need the conditional distribution of the random effects:

$$\begin{aligned} \pi(\alpha|(N(t_1),\tau);a,b,\beta) &= \frac{L(\alpha,\beta|(N(t_1),\tau))\pi(\alpha;a,b)}{\int_\alpha L(\alpha,\beta|(N(t_1),\tau))\pi(\alpha;a,b)d\alpha} \\ &= \prod_{i=1}^{k} \frac{\left[\left(\prod_{j=1}^{N_i(t_{1i})} \alpha_i f(\tau_{ij};\beta)\right) e^{-\alpha_i F(t_{1i};\beta)}\right] b^a e^{-\alpha_i b}\alpha_i^{a-1}\Gamma(a)^{-1}}{\int_{\alpha_i}\left[\left(\prod_{j=1}^{N_i(t_{1i})} \alpha_i f(\tau_{ij};\beta)\right) e^{-\alpha_i F(t_{1i};\beta)}\right] b^a e^{-\alpha_i b}\alpha_i^{a-1}\Gamma(a)^{-1}d\alpha_i} \\ &= \prod_{i=1}^{k} \frac{e^{-\alpha_i(b+F(t_{1i};\beta))}\alpha_i^{N_i(t_{1i})+a-1}}{\int_{\alpha_i} e^{-\alpha_i(b+F(t_{1i};\beta))}\alpha_i^{N_i(t_{1i})+a-1}d\alpha_i} \\ &= \prod_{i=1}^{k} \frac{(b+F(t_{1i};\beta))^{a+N_i(t_{1i})}e^{\alpha_i(b+F(t_{1i};\beta))}\alpha_i^{a+N_i(t_{1i})-1}}{\Gamma(a+N_i(t_{1i}))}. \end{aligned}$$

Then, the $\alpha_i|(N(t_1),\tau)$'s are independent Gamma$(a+N_i(t_{1i}), b+F(t_{1i};\beta))$ random variables. Note that the times of the occurrences do not affect the conditional distribution

of α_i; only the knowledge of $N_i(t_{1i})$ is required to determine its conditional distribution. Therefore, this random variable will be denoted by $\alpha_i|N_i(t_{1i})$.

Using this conditional density, we can find the density function for $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ given $N(t_1)$.

Corollary 4.1. *(of Proposition 3.4) Under the model given in (4.4), the predictive density function of* $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ *given* $N(t_1)$ *is a convolution of* $|\mathcal{S}|$ *negative binomials with parameters* $a + N_i(t_{1i})$ *and* $(b + F(t_{1i}; \beta))/(b + F(t_{2i}; \beta))$. *Such density can be written as*

$$\begin{aligned} p[n|N(t_1); a, b, \beta] &= P[\mathbf{N}_{\mathcal{S}}(t_1, t_2) = n|N(t_1); a, b, \beta] \\ &= \sum_{\{z_i: \sum_{i\in\mathcal{S}} z_i = n\}} \prod_{i\in\mathcal{S}} \frac{\Gamma(a + N_i(t_{1i}) + z_i)}{\Gamma(a + N_i(t_{1i}))z_i!} \left(\frac{F(t_{2i}; \beta) - F(t_{1i}; \beta)}{b + F(t_{2i}; \beta)}\right)^{z_i} \times \\ &\quad \left(\frac{b + F(t_{1i}; \beta)}{b + F(t_{2i}; \beta)}\right)^{a+N_i(t_{1i})}. \end{aligned} \tag{4.5}$$

We can see that the times of occurrences do not appear in (4.5). This means that even if the number of occurrences for each process is interval-censored on $[0, t_{1i}]$, it will not change the predictive density. Nevertheless, it may affect the preciseness of the estimate of the unknown β. We also note that when $f(t; \beta) = 1$, the model given in (4.4) is identical to model (3.5) for homogeneous Poisson processes. Then, the preceding density function is identical to the one given in (3.6) since $F(t; \beta) = t$.

The density (4.5) is obtained by summing over $\binom{n+|\mathcal{S}|+1}{n}$ terms, which may not be possible for some software when $|\mathcal{S}|$ or n are large. However, this problem can be solved by using a recursive formula to find $p(n|N(t_1))$.

Corollary 4.2. *(of Proposition 3.5) Under the model given in (4.4), the density function of* $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$ *given* $N(t_1)$ *can be written as*

$$p(n|N(t_1); a, b) = P[\mathbf{N}_{\mathcal{S}}(t_1, t_2) = n|N(t_1); a, b, \beta]$$

$$= \begin{cases} \displaystyle\prod_{i\in\mathcal{S}} \left(\frac{b+F(t_{1i};\beta)}{b+F(t_{2i};\beta)}\right)^{(a+N_i(t_{1i}))} & \text{if } n=0, \\ \displaystyle\sum_{j=0}^{n-1} p(j|N(t_1);a,b)\left(\frac{H^{(n-1-j)}(0)}{(n-1-j)!}\right) & \text{if } n\geq 1, \end{cases}$$

where

$$H^{(j)}(0) = \sum_{i\in\mathcal{S}}(a+N_i(t_{1i}))\left(\frac{F(t_{2i};\beta)-F(t_{i1};\beta)}{b+F(t_{2i};\beta)}\right)^{j+1}.$$

When n is very large, it can be more convenient to approximate $p(n|N(t_1);a,b)$ instead of using a recursive formula. Such an approximation can be done by generating convolutions of gamma random variables. Like in Chapter 3, we can show that

$$p(n|\tau;a,b,\beta) \simeq \sum_{i=1}^{B}\frac{e^{-u_i^*}(u_i^*)^n}{n!}. \tag{4.6}$$

when B is large. Here, u_i^* is a convolution of $|\mathcal{S}|$ Gamma$(a+N_i(t_{1i}),(b+F(t_{1i};\beta))/(F(t_{2i};\beta)-F(t_{1i};\beta)))$.

An asymptotic expansion for $p(n|N(t_1);a,b)$ similar to (3.7) can also be derived here. For any $i\in\mathcal{S}$ we can show that:

$$p(n|N(t_1);a,b) \simeq \frac{\Gamma(a+N_i(t_{1i})+n)}{\Gamma(a+N_i(t_{1i}))n!}\left(\frac{F(t_{2i},\beta)-F(t_{1i},\beta)}{b+F(t_{2i},\beta)}\right)^n\left(\frac{b+F(t_{1i};\beta)}{b+F(t_{2i};\beta)}\right)^{a+N_i(t_{1i})}\times C\left(\frac{b+F(t_{2i};\beta)}{F(t_{2i};\beta)-F(t_{1i};\beta)}\right),$$

where

$$C(s) = \prod_{j\in\mathcal{S}\setminus\{i\}}\left[1+(1-s)\frac{F(t_{2j};\beta)-F(t_{1j};\beta)}{b+F(t_{1j};\beta)}\right]^{-(a+N_j(t_{1j})}.$$

We will now present different ways to estimate the unknown parameters a, b, and β. These estimates will substitute for the real parameters in the plug-in predictive density

$\tilde{f}_p(n|N(t_1)) = p(n|N(t_1); \hat{a}, \hat{b}, \hat{\beta})$. If these estimates are consistent, we will then have approximately unbiased predictors and approximately exact prediction intervals. Note that it will then still be possible to calibrate these plug-in prediction intervals.

Obvious candidates for $\hat{a}$, $\hat{b}$, and $\hat{\beta}$ are those maximizing the marginal likelihood

$$\begin{aligned}
L(a,b,\beta|(N(t_1),\tau)) &= \int_\alpha L(\alpha,\beta|(N(t_1),\tau))\pi(\alpha;a,b)d\alpha \\
&= \int_\alpha \left[\prod_{i=1}^{k}\left(\prod_{j=1}^{N_i(t_{1i})}\alpha_i f(\tau_{ij};\beta)\right)e^{-\alpha_i F(t_{1i};\beta)}\right]\left(\prod_{i=1}^{k}\frac{b^a e^{-\alpha_i b}\alpha_i^{a-1}}{\Gamma(a)}\right)d\alpha \\
&= \prod_{i=1}^{k}\left[\frac{(\prod_{j=1}^{N_i(t_{1i})}f(\tau_{ij};\beta))b^a}{\Gamma(a)}\int_{\alpha_i} e^{-\alpha_i(b+F(t_{1i};\beta))}\alpha_i^{a+N_i(t_{1i})-1}d\alpha_i\right] \\
&= \prod_{i=1}^{k}\left[\left(\prod_{j=1}^{N_i(t_{1i})}f(\tau_{ij};\beta)\right)\left(\frac{b^a}{(b+F(t_{1i};\beta))^{a+N_i(t_{1i})}}\right)\times\right. \\
&\qquad \left.\left(\frac{\Gamma(a+N_i(t_{1i}))}{\Gamma(a)}\right)\right]. \qquad (4.7)
\end{aligned}$$

We can find them by solving the score equations,

$$\begin{aligned}
\sum_{i=1}^{k}\left[log\left(\frac{b}{b+F(t_{1i};\beta)}\right)+\mathbb{I}_{N_i(t_{1i})\neq 0}\sum_{l=0}^{N_i(t_{1i})-1}\frac{1}{a+l}\right] &= 0, \\
\sum_{i=1}^{k}\left[\frac{a}{b}-\frac{a+N_i(t_{1i})}{b+F(t_{1i};\beta)}\right] &= 0,
\end{aligned}$$

and

$$\sum_{i=1}^{k}\left[\sum_{j=1}^{N_i(t_{1i})}\frac{\frac{\partial}{\partial\beta}f(\tau_{ij};\beta)}{f(\tau_{ij};\beta)}-\frac{(a+N_i(t_{1i}))\frac{\partial}{\partial\beta}F(t_{1i};\beta)}{F(t_{1i};\beta)+b}\right] = 0.$$

These estimates will be denoted $\hat{a}_{mle}$, $\hat{b}_{mle}$, and $\hat{\beta}_{mle}$. Note that we have to use numerical methods for most forms of $f(t;\beta)$.

We already discussed in Chapter 3 that the ultimate goal in prediction problems is not just to find a model that provides a good fit to the data but that is able to predict adequately future observations. Clearly, this can also be done with NHPP's. For example, if we especially want to find point predictors for $\mathbf{N}_{\mathcal{S}}(t_1, t_2)$, we can select parameters that will make the point prediction of the (known) $\mathbf{N}_{\mathcal{S}}(t^*, t_1)$ given $N(t^*)$ as precise as possible. These estimates will differ according to the value of t^* and the discrepancy chosen, but a possible choice is

$$(\hat{a}_{dis}, \hat{b}_{dis}, \hat{\beta}_{dis}) \;=\; \arg_{(a,b,\beta)}[minD(a,b,\beta)],$$

where

$$\begin{aligned} D(a,b,\beta) &= \sum_{i=1}^{k} \left(N_i(t_z^*, t_{1i}) - \mathbb{E}[\mathbf{N}_i(t_z^*, t_{1i})|N_i(t_z^*)]\right)^2 \\ &= \sum_{i=1}^{k} \left[N_i(t_z^*, t_{1i}) - \left(\frac{a + N_i(t_z^*)}{b + F(t_z^*;\beta)}\right)(F(t_{1i};\beta) - F(t_z^*;\beta))\right]^2 \end{aligned}$$

and

$$t_z^* = \min\{t : \sum_{i=1}^{k} N_i(t) \geq z \sum_{i=1}^{k} N_i(t_{1i}) \text{ and } t \leq \min(t_{1i})\}.$$

With NHPP's, the estimation of unknown parameters based on their ability to predict can also be done in a different way. Instead of minimizing a certain discrepancy, we can select parameters maximizing the predictive density function given a portion of the data already observed. At time t_z^*, a fraction z of the $\sum_{i=1}^{k} N_i(t_{1i})$ events had been observed, if we let

$$M \;=\; \{(i,j) : \tau_{ij} > t_z^*\},$$

the density function for the remaining observations is then:

$$p(\tau_{ij} : (i,j) \in M|N(t_z^*); a, b, \beta) \;=\; \int_{\alpha} p(\tau_{ij} : (i,j) \in M|\alpha; a, b, \beta)\pi(\alpha|N(t_z^*); a, b, \beta)d\alpha$$

$$
\begin{aligned}
&= \prod_{i=1}^{k} \int_0^\infty \left(\prod_{\{\tau_{ij} > t_z^*\}} \alpha_i f(\tau_{ij};\beta) \right) e^{-\alpha_i (F(t_{1i};\beta) - F(t_z^*;\beta))} \times \\
&\qquad \left[\frac{(b + F(t_z^*;\beta))^{a+N_i(t_z^*)} e^{-\alpha_i (b+F(t_z^*;\beta))} \alpha_i^{a+N_i(t_z^*)-1}}{\Gamma(a + N_i(t_z^*))} \right] d\alpha_i \\
&= \prod_{i=1}^{k} \frac{\left(\prod_{\{\tau_{ij} > t_z^*\}} f(\tau_{ij};\beta) \right) (b + F(t_z^*;\beta))^{a+N_i(t_z^*)}}{\Gamma(a + N_i(t_z^*))} \times \\
&\qquad \int_0^\infty e^{-\alpha_i (b+F(t_{1i};\beta))} \alpha_i^{a+N_i(t_{1i})-1} d\alpha_i \\
&= \prod_{i=1}^{k} \left(\prod_{\{\tau_{ij} > t_z^*\}} f(\tau_{ij};\beta) \right) \left[\frac{(b + F(t_z^*;\beta))^{a+N_i(t_z^*)}}{(b + F(t_{1i};\beta))^{a+N_i(t_{i1})}} \right] \times \\
&\qquad \left(\frac{\Gamma(a + N_i(t_{i1}))}{\Gamma(a + N_i(t_z^*))} \right). \qquad (4.8)
\end{aligned}
$$

The estimates for a, b, and β are then

$$
(\hat{a}_{mpe}, \hat{b}_{mpe}, \hat{\beta}_{mpe}) = \arg_{a,b,\beta}(\max p(\{\tau_{ij} > t_z^*\} | N(t_z^*); a, b, \beta)),
$$

where *mpe* stands for maximum predictive estimate. Note that when z is close to zero, the estimates obtained are identical to $(\hat{a}_{mle}, \hat{b}_{mle}, \hat{\beta}_{mle})$.

We now have presented 3 methods to find estimates that will completely specify our random effects model: maximum likelihood estimation, estimation by minimizing a given discrepancy, and maximum predictive estimation. However, if one does not wish to use plug-in methods, the only alternative seems to be to use a Bayesian model with prior distributions on a, b, and β. It is highly likely that any choice of prior will require numerical integration to find the predictive density function. Note also that priors on a and b will often be non-informative given the rare availability of a state of knowledge for parameters of an unobservable random quantity.

The last thing we have to specify in the random effects model (4.4) is the function $f(t;\beta)$. If the parameters a, b, and β are chosen among any positive real numbers, the function $f(t;\beta)$ is usually chosen from a small number of functions. We suggest two ways to select such a function among a finite number of candidates. First, we can use the likelihood given in (4.7) and select the function providing the highest value of $L(\hat{a}_{mle}, \hat{b}_{mle}, \hat{\beta}_{mle}|(N(t_1), \tau))$. Also, for a given t_z^*, we can use the predictive density (4.8) and select the function maximizing $p(\{\tau_{ij} > t_z^*\}|N(t_z^*); \hat{a}_{mpe}, \hat{b}_{mpe}, \hat{\beta}_{mpe})$. Since $\dim(\beta)$ is not necessarily the same for all the functions considered, well known criteria like the AIC or the BIC can be used for both approaches.

4.3 Empirical study

In this section, we will study the methods presented so far. While this was done through extensive simulations in the previous chapter, different approaches for NHPP's will be studied by analyzing some simulated and real datasets. Instead of focusing on the robustness of random effects models with respect to various forms of misspecifications, we will study here the effect of data accumulation on prediction. We conjecture that most of the interesting features observed previously with processes that were time homogeneous should still be true here. We recall that the features observed were mainly the robustness of the methods using random effects to the rate homogeneity/heterogeneity assumption, and their robustness to the true distribution of the random effects.

All the functions $f(t;\beta)$ used in this section can be found in Table 4.2. These functions were chosen mostly because of their popularity in the literature and their ability to characterize different behaviors of NHPP's. Note that for all functions we have $\alpha, \beta > 0$. The first model, called the EXP model in this section, is often referred to as the Goel-Okumoto

MODEL	$\alpha f(t;\beta)$	$\alpha F(t;\beta)$
EXP	$\alpha\beta e^{-\beta t}$	$\alpha(1-e^{-\beta t})$
GAM	$\alpha\beta^2 t e^{-\beta t}$	$\alpha[1-(1+\beta t)e^{-\beta t}]$
LOG	$\alpha\beta/(\beta t+1)$	$\alpha\log(1+\beta t)$
POW	$\alpha\beta t^{\beta-1}$	αt^{β}

Table 4.2: A list of NHPP models.

model (Goel & Okumoto 1979) in the literature. For any value of β, the rate function is decreasing with respect to t and since $F(\infty;\beta)=1$, the expected number of occurrences in $[0,\infty)$ is finite and corresponds to α. The function associated with the second model also has a finite expectation on $[0,\infty)$. However, its rate function is increasing on $[0,1/\beta)$ and decreasing on $(1/\beta,\infty)$. This model will be referred to as the GAM model. The third model always has a decreasing rate but the expected number of events on $[0,\infty)$ is not finite, it will be referred to as the LOG model. The last model is the well known power law model and will be called the POW model. It is the most flexible model studied here. Like the LOG model its expectation is infinite, but its rate function can be either decreasing or increasing depending on whether β is smaller or greater than 1. By letting $\beta=1$, this model can also be used to model homogeneous Poisson processes.

We will now use real and simulated datasets to study the effect of data accumulation on point and set predictions. In order to do so, we will find predictors and prediction intervals for $\sum_{i=1}^{k}\mathbf{N}_i(t_{1i},t_{2i})$ with different values of t_{1i} converging towards t_{2i}. This situation, called finite horizon prediction, is commonly encountered since processes are often monitored throughout a long period of time. The parameters will be chosen to reflect different levels of heterogeneity and different magnitudes for the total number of events observed. These processes will be simulated using the following proposition. A proof of this proposition can

be found in Snyder & Miller (1991, Section 2.3).

Proposition 4.1. *Let n be the realization of a Poisson random variable with rate $\alpha F(T;\beta)$, where $F(t;\beta)$ is a positive-valued function that is invertible and increasing with respect to t. Now, let $u_1,\ldots,u_n$ be n independent realizations of an uniform distribution on $[0,1]$. Then, the n values t_i such that*

$$\frac{F(t_i;\beta)}{F(T;\beta)} \quad = \quad u_i$$

are the realizations, over the interval $[0,T]$, of a NHPP with rate $\alpha f(t;\beta)$ where $f(t;\beta) = \frac{\partial}{\partial t}F(t;\beta)$.

The first dataset is a simulation of $k = 25$ processes over the time interval $[0,100]$. We used the random effects model (4.4), with $f(t;\beta)$ corresponding to the LOG function. The parameters used were $a = 10$, $b = 2$, and $\beta = 0.05$. We chose these parameters to obtain approximately 10 events per subject and a 5-fold variation between the $N_i(100)$'s; a likely scenario in practical problems. We obtained $\sum_{i=1}^{25} N_i(100) = 250$ with the $N_i(100)$'s ranging from 3 to 17. This dataset and the following ones are presented in Figure 4.1. Note that the processes are sorted according to the number of occurrences, which explains the greater number of events in the upper part of the plots. The second dataset was simulated to obtain a similar variability but with a smaller number of events. The parameters used were $a = 10$, $b = 2$, $\beta = 0.02$, and we used the EXP model from Table 4.2. Like for the first dataset, we simulated $k = 25$ processes observed over the time interval $[0,100]$. We also decided to simulate NHPP's from model (4.1) where the α_i's are fixed effects. This was done with the third dataset of $k = 25$ processes over the time interval $[0,100]$. The parameters were $(\alpha_1,\alpha_2,\ldots,\alpha_{25}) = (0.4, 0.45,\ldots,1.6)$, $\beta = 0.5$, and the POW model was used. With these parameters, this dataset is expected to have a variability amongst the $N_i(100)$'s and a total number of events that are similar to those of the first dataset.

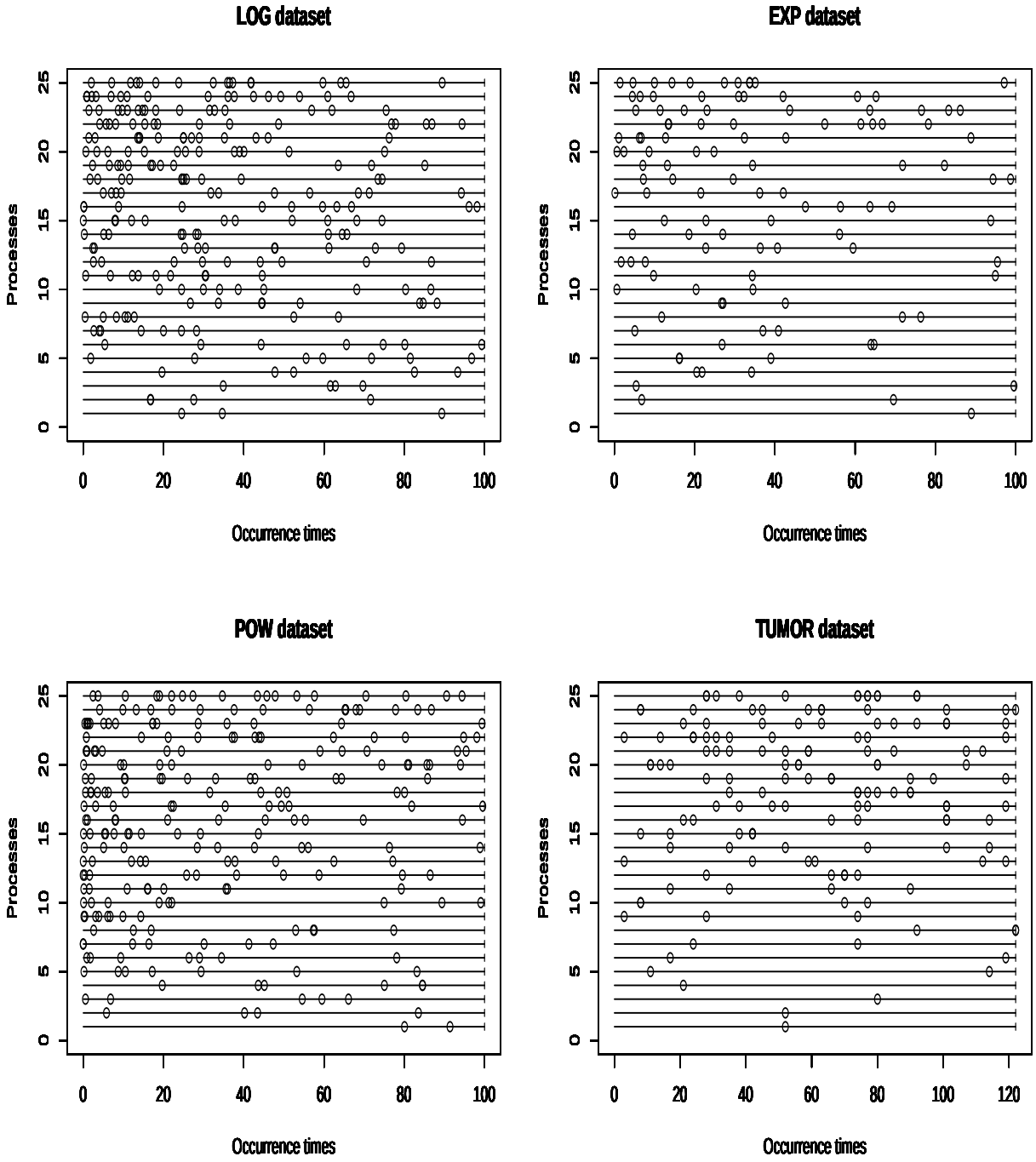

Figure 4.1: Recurrence data plots.

Finally, the fourth dataset corresponds to times of development of mammary tumors from a (control) group of 25 female rats observed over a period of 122 days (Gail et al. 1980).

To study the effect of data accumulation on the first dataset, Figure 4.2 and 4.3 contain 90% prediction intervals for $\sum_{i=1}^{25} \mathbf{N}_i(t_{2i})$, where $t_{2i} = 100$, using the different values $t_{1i} = (15, 20, \ldots, 95)$ for each process. Such predictions are useful in areas such as insurance or warranty coverage, where it may be desired to predict, say, annual total claims based on what we observed at different times of the year. The first part of Figure 4.2 presents plug-in prediction intervals using the m.l.e.'s $(\hat{\alpha}_1, \ldots, \hat{\alpha}_{25}, \hat{\beta})$ of the model without random effects (4.1), these estimates were obtained using the equations in the second row of Table 4.1. This figure also includes plug-in prediction intervals using $(\hat{a}_{mle}, \hat{b}_{mle}, \hat{\beta}_{mle})$ to specify the distribution of the random effects. Finally, the last part of Figure 4.2 shows exact prediction intervals using the real values of a, b, and β.

The first thing we notice in this figure is the similarity between the two types of plug-in prediction intervals: they vary similarly over time and always agree on whether or not they include the actual value of $\sum_{i=1}^{25} \mathbf{N}_i(100)$. However, the method using the random effects usually gives, as expected, wider prediction intervals. We can also see that the plug-in intervals from the random effects model are also wider than the prediction intervals using the real parameters: they contain approximately 10 more integers for the first intervals and only a few towards the end.

Figure 4.3 presents plug-in prediction intervals using the estimates $(\hat{a}_{dis}, \hat{b}_{dis}, \hat{\beta}_{dis})$ and $(\hat{a}_{mpe}, \hat{b}_{mpe}, \hat{\beta}_{mpe})$. In both cases, we used $z = 0.25$ to obtain t_z^*. We also tried $z = 0.5$ and $z = 0.75$ but the results obtained were less accurate, especially with $z = 0.75$. We can see that intervals using $(\hat{a}_{mpe}, \hat{b}_{mpe}, \hat{\beta}_{mpe})$ do not perform well. However, it is not the case for those using $(\hat{a}_{dis}, \hat{b}_{dis}, \hat{\beta}_{dis})$: they include the real value as often as the first two plug-in methods and their lengths are similar to the lengths of the real prediction intervals.

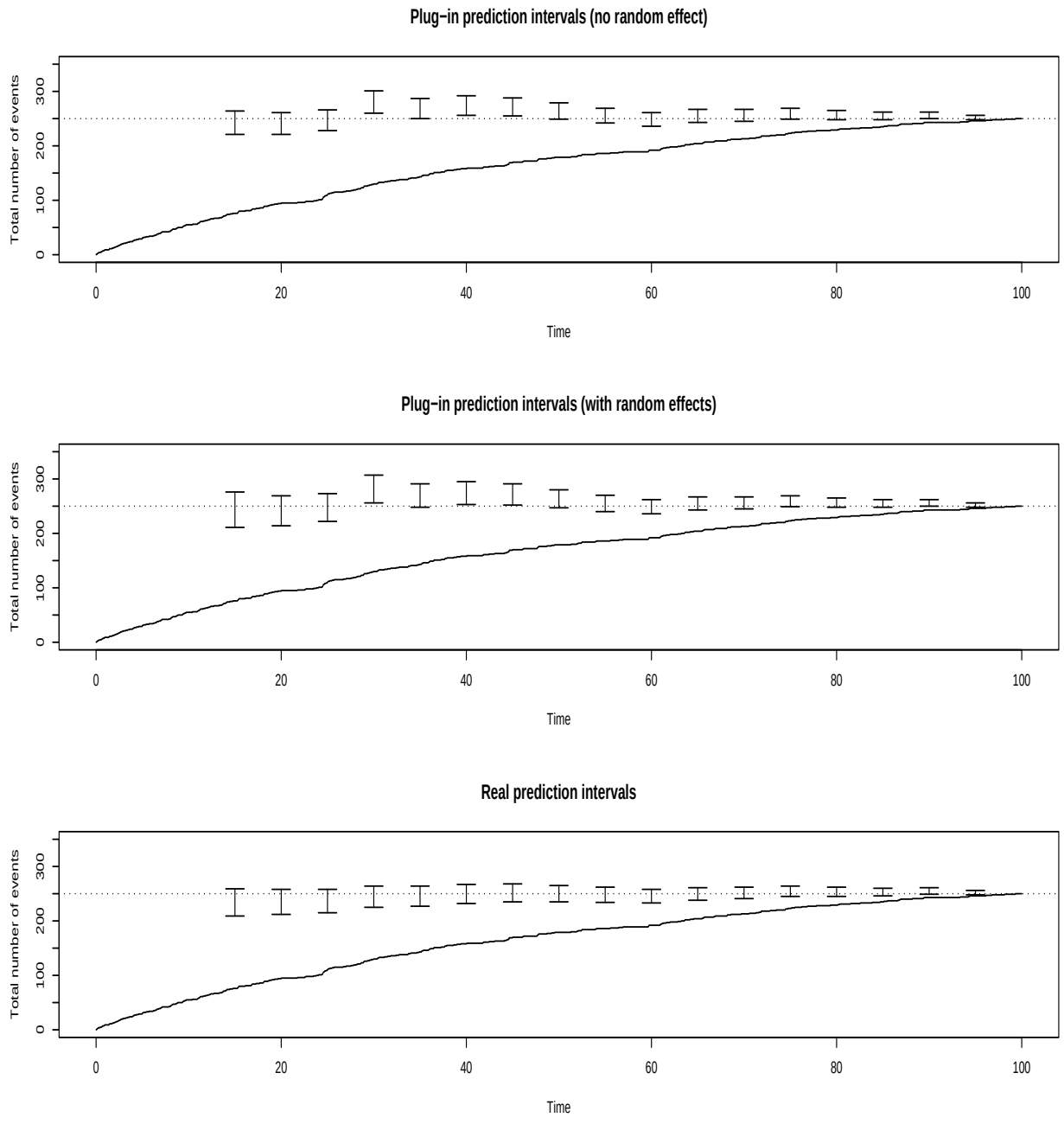

Figure 4.2: Real and plug-in 90% prediction intervals for the simulated LOG dataset.

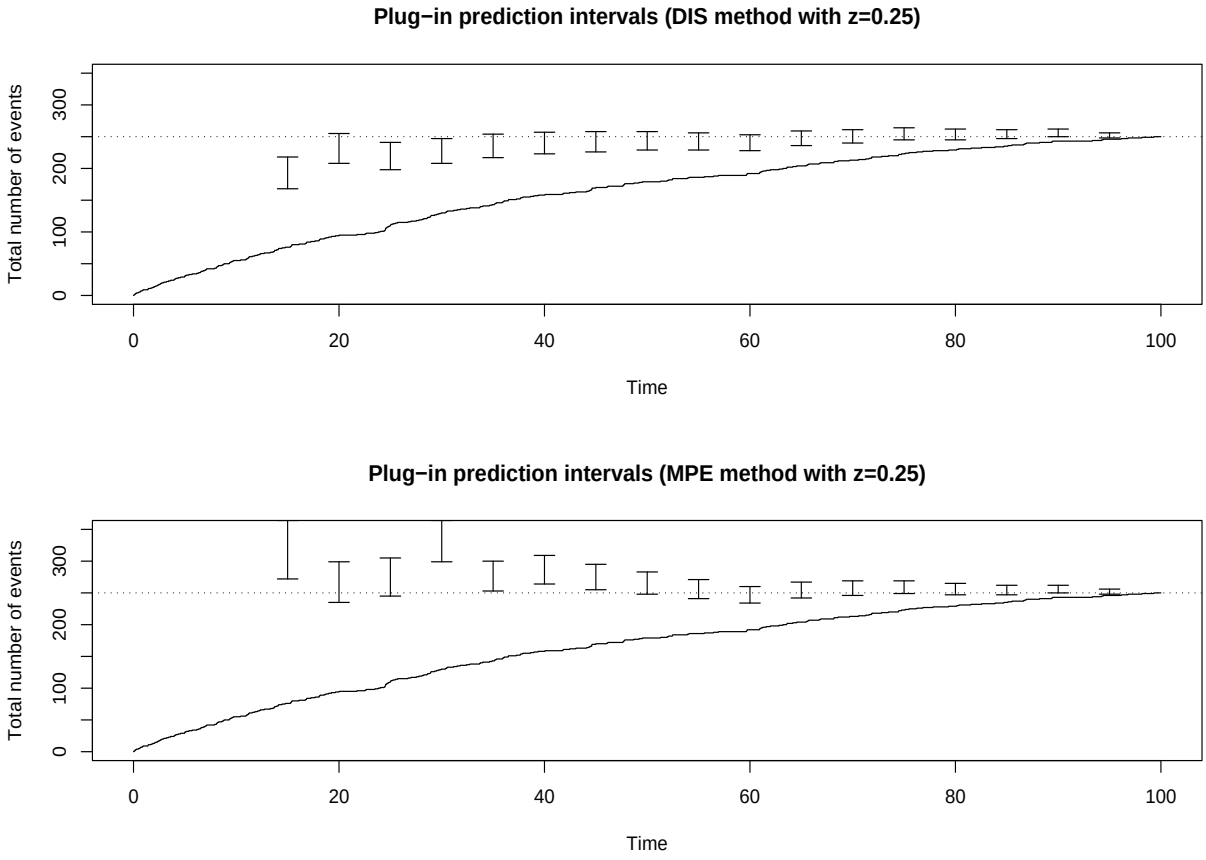

Figure 4.3: Plug-in 90% prediction intervals for the simulated LOG dataset.

t_{1i}	$\hat{u}$ $(u = 1.609)$	$\hat{v}$ $(v = 0.916)$	$\hat{\beta}$ $(\beta = 0.050)$
20	1.667	2.304	0.052
30	2.086	2.601	0.030
40	2.015	2.418	0.033
50	1.897	2.042	0.039
60	1.721	1.474	0.049
70	1.810	1.450	0.043
80	1.816	1.421	0.043
90	1.813	1.007	0.043
100	1.729	0.665	0.049

Table 4.3: Maximum likelihood estimates for the LOG dataset.

The study of this dataset also revealed interesting features about the m.l.e.'s $(\hat{a}_{mle}, \hat{b}_{mle}, \hat{\beta}_{mle})$. First, if we use the whole dataset to find 95% confidence intervals for $a = 10$ and $b = 2$, we can see that both intervals contain the real value but are relatively large, [8.49, 24.18] for a and [0.75, 5.05] for b. However, using the reparameterization $u = log(a/b)$ and $v = log(a/b^2)$, we obtain two 95% confidence intervals having different characteristics: a short confidence interval, [1.56, 1.90], for $u = 1.61$ and a larger one, [-0.33, 1.66], for $v = 0.92$. The parameters u and v being the logarithms of the mean and variance of the random effects, it seems that this dataset of 25 processes allows us to obtain a relatively accurate estimate for the mean of the (unobservable) random effects, but not for its variance. A similar feature is also shown in Table 4.3 which contains the estimates $(\hat{u}, \hat{v}, \hat{\beta})$ obtained using the data up to different values of t_{1i}. We can see that the mean is accurately estimated, even for small values of t_{1i}. However, it is not the case for

t_{1i}	$(\hat{\alpha}_1, \ldots, \hat{\alpha}_{25}, \hat{\beta})$	$(\hat{a}_{mle}, \hat{b}_{mle}, \hat{\beta}_{mle})$	$(\hat{a}_{dis}, \hat{b}_{dis}, \hat{\beta}_{dis})$	$(\hat{a}_{mpe}, \hat{b}_{mpe}, \hat{\beta}_{mpe})$
20	4.37	2.98	2.53	3.64
30	4.30	3.14	2.86	4.91
40	3.37	2.55	2.25	2.99
50	2.51	1.95	1.88	2.05
60	1.74	1.40	1.48	1.42
70	1.59	1.36	1.37	1.37
80	1.29	1.14	1.13	1.13
90	0.65	0.60	0.60	0.60

Table 4.4: Absolute error of point predictors (LOG dataset).

the estimates of the variance since they vary substantially as t_{1i} increases. This table also shows that β is accurately estimated for moderate values of t_{1i}. It is interesting to note that these values of $\hat{\beta}$ also correspond to the estimates of β obtained from the model without random effects. In this dataset, the difference between these two estimates is unobservable up to the third decimal.

Using this dataset, it appears that whether we are assuming that the α_i's are random or not, the approaches using m.l.e.'s are equivalent. However, it does not seem to be the case when we wish to predict only for a few processes. Table 4.4 presents the average distance between the actual value of $\mathbf{N}_i(t_{1i}, t_{2i})$ and its prediction. We can then see the advantage of using random effects models; the three methods using them have a better performance. The one using $(\hat{a}_{dis}, \hat{b}_{dis}, \hat{\beta}_{dis})$ usually gives the best prediction, followed closely by the one using $(\hat{a}_{mle}, \hat{b}_{mle}, \hat{\beta}_{mle})$.

Many of the features mentioned above regarding the first dataset are also true for the

second one. However, since the expected number of events per process is approximatively 50% smaller, it seems that it has affected the performance of the plug-in prediction intervals. We can see in Figure 4.4 that the 90% plug-in prediction intervals using the estimates $(\hat{a}_{mle}, \hat{b}_{mle}, \hat{\beta}_{mle})$ or $(\hat{a}_{dis}, \hat{b}_{dis}, \hat{\beta}_{dis})$ have coverage proportions that are far from the desired one. We must keep in mind that successive prediction intervals are not independent but we think that these results still suggests that these prediction intervals should be calibrated for this dataset. This will be done later in this section.

As we mentioned earlier, the third dataset used fixed values for $(\alpha_1, \ldots, \alpha_{25})$. It is interesting to note that all the features mentioned in the analysis of the first dataset are still valid for this dataset, where the total expected number of events was approximatively the same. It seems that whether or not the α_i's are random has little impact on the analysis. First, we can see from Figure 4.5 how well 90% plug-in prediction intervals perform with respect to the real prediction intervals. The plug-in intervals presented in this figure were obtained using the estimates $(\hat{a}_{mle}, \hat{b}_{mle}, \hat{\beta}_{mle})$, but the other three types of estimations gave very similar results. Like for the first dataset, the model using random effects gives m.l.e.'s of $u = log(a/b)$ that are relatively stable as t_{i1} increases. This is not the case for the estimates of $v = log(a/b^2)$, as they vary substantially over time. Finally, point predictors using random effects again have smaller average distances from the actual value of $\mathbf{N}_i(t_{1i}, t_{2i})$.

The fourth dataset is the only one that was not simulated. The POW model was used since it was usually the one giving the highest likelihood. The total number of tumors recorded was 149 and, like for the second dataset, it does not seem enough to be able to obtain adequate plug-in prediction intervals. Figure 4.6 shows plug-in prediction intervals using $(\hat{a}_{mle}, \hat{b}_{mle}, \hat{\beta}_{mle})$. Clearly, the coverage proportion is below the desired level.

By looking at the second and fourth datasets, it seems that when the total num-

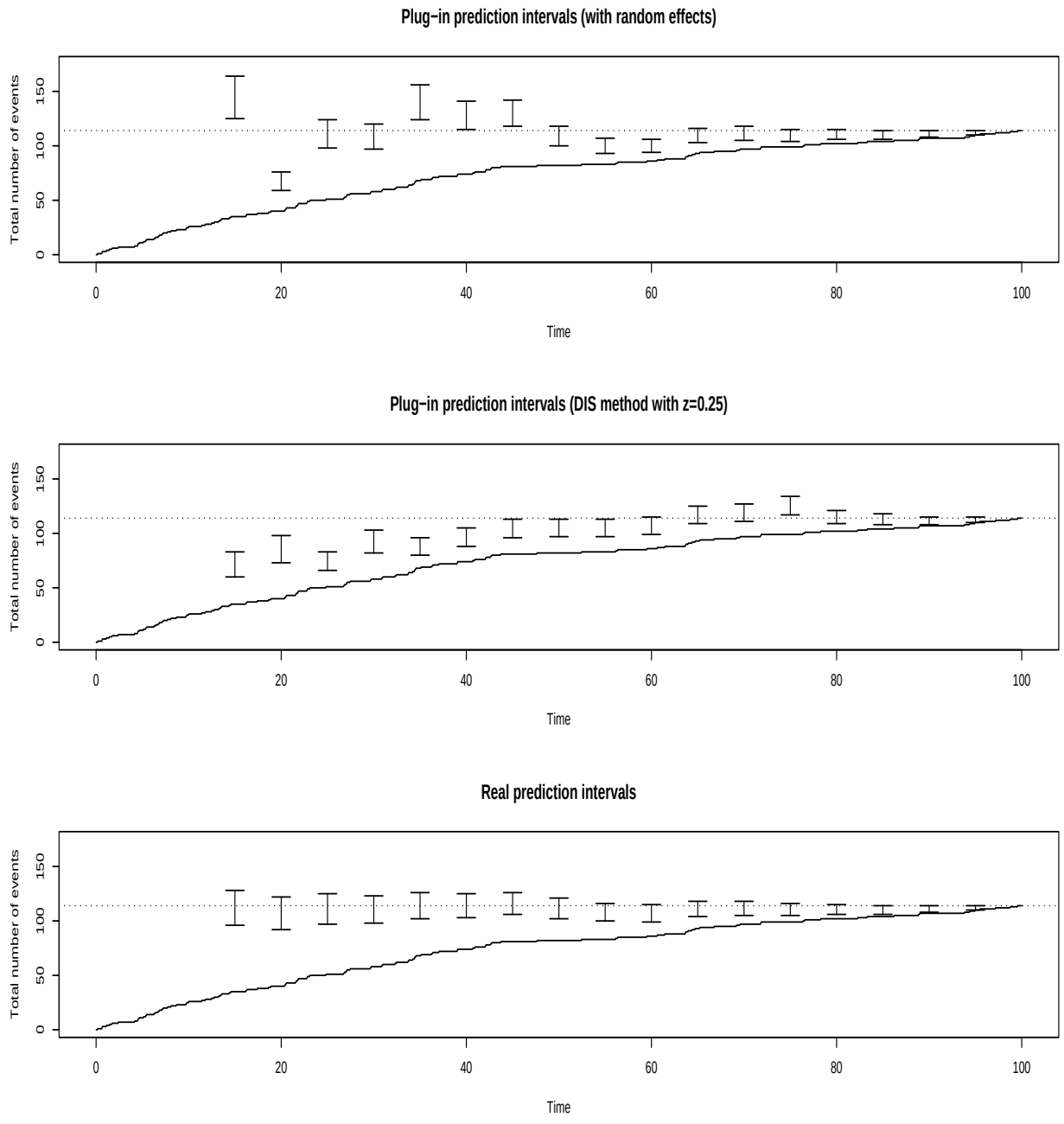

Figure 4.4: Real and plug-in 90% prediction intervals for the simulated EXP dataset.

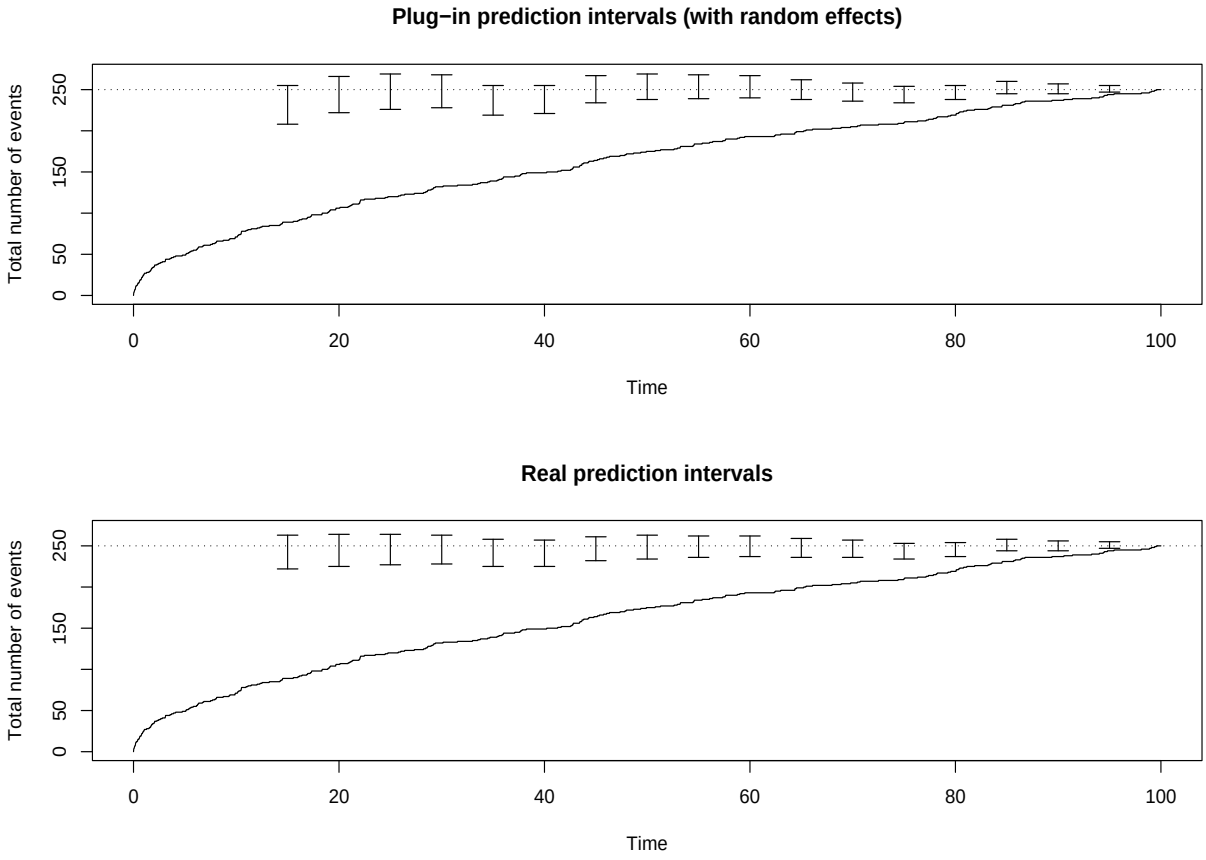

Figure 4.5: Real and plug-in 90% prediction intervals for the simulated POW dataset.

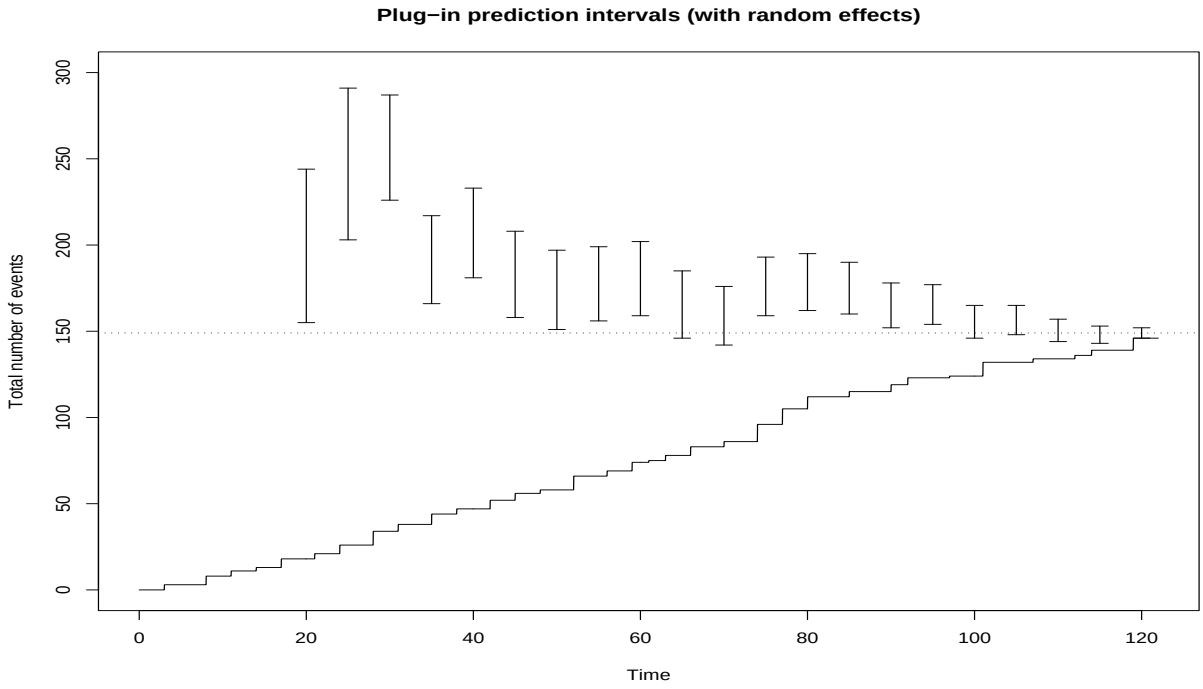

Figure 4.6: Plug-in 90% prediction intervals for the TUMOR dataset.

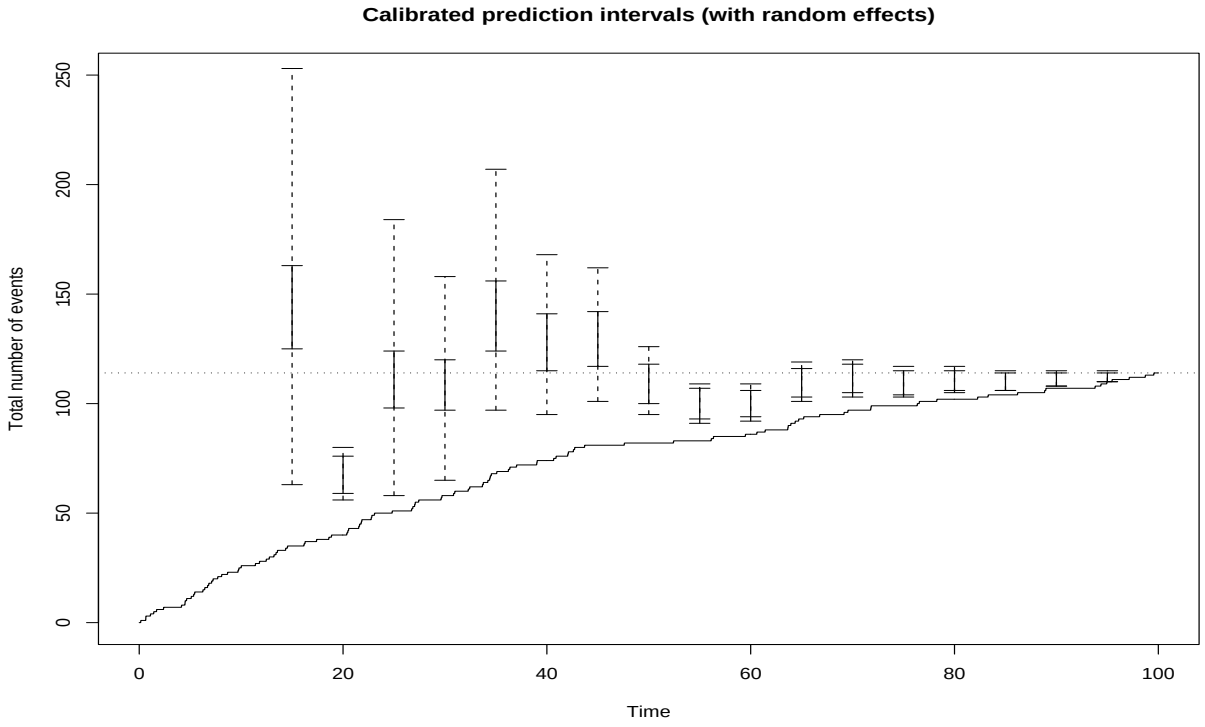

Figure 4.7: Calibrated 90% prediction intervals for the simulated EXP dataset.

ber of events is small, it is important to calibrate the plug-in prediction intervals. Figure 4.7 shows calibrated 90% prediction intervals for the second dataset when the estimates $(\hat{a}_{mle}, \hat{b}_{mle}, \hat{\beta}_{mle})$ are used. The values added by the calibration procedure are indicated by a dotted line. The early intervals are usually very large but these new intervals are clearly more appropriate. These calibrations were done by approximating the c.d.f. of $\mathbf{U} = F(\sum_{i=1}^{25} \mathbf{N}_i(t_1, 100) | \mathbf{N}(t_1); \hat{a}(\mathbf{N}(t_1)), \hat{b}(\mathbf{N}(t_1)), \hat{\beta}(\mathbf{N}(t_1)))$ using the algorithm described in Section 2.3 with $B = 1000$. Figure 4.8 shows the calibration curves for different values of t_1. These curves approach the c.d.f. of an uniform distribution as t_1 increases. However, this is no longer the case when t_1 is too close to 100. In such a case, the discrete random variable $\sum_{i=1}^{25} \mathbf{N}_i(t_1, 100)$ only has a few plausible values and the calibration approach is

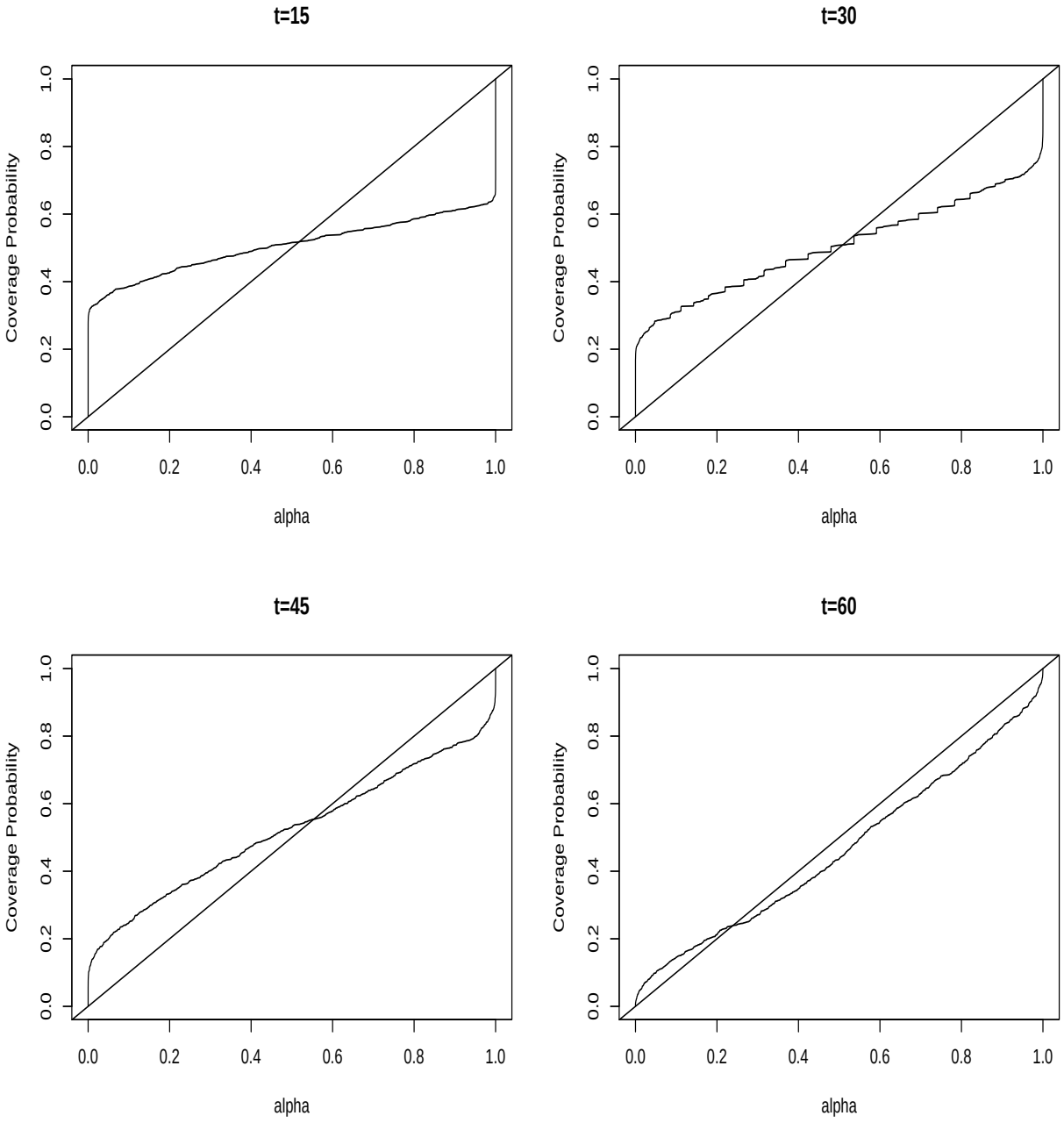

Figure 4.8: Calibration curves for the EXP dataset

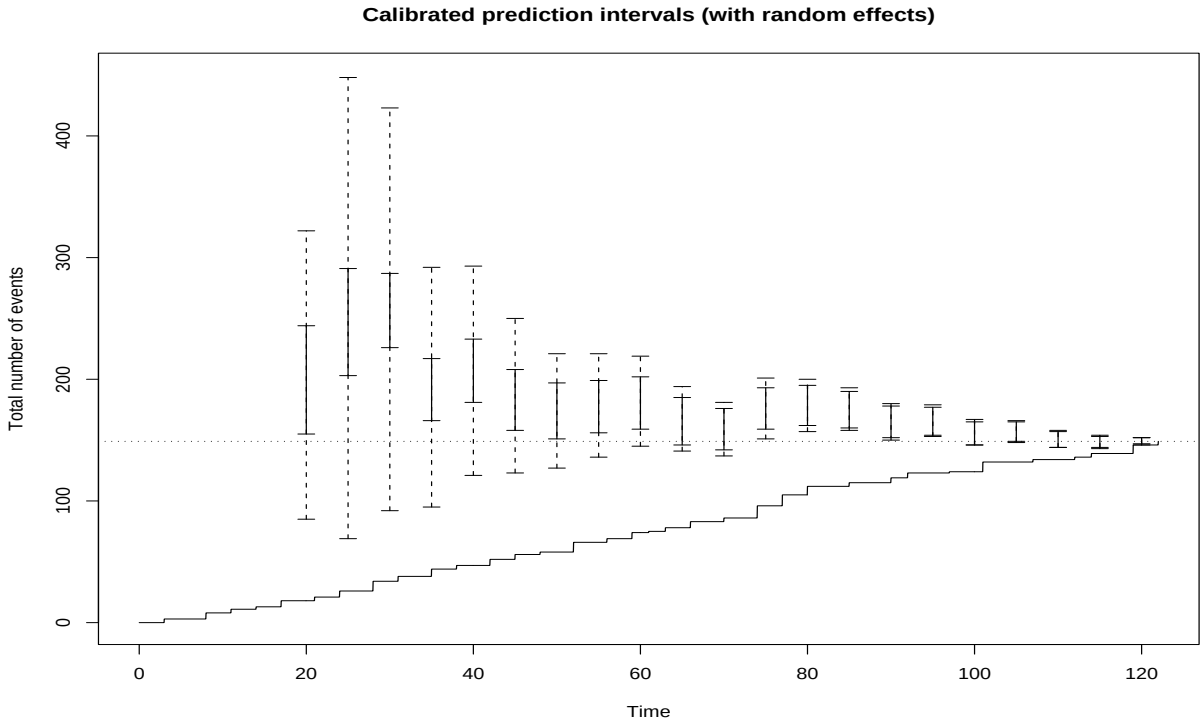

Figure 4.9: Calibrated 90% prediction intervals for the TUMOR dataset.

influenced by this fact.

Figure 4.9 shows the calibrated prediction interval for the fourth dataset. These new intervals are clearly more appropriate than those obtained with the plug-in approach but they still show some deficiencies: after 70 days, there is a sudden increase in the number of tumors and we see that it leads to an over-prediction of the final value during a certain amount of time. However, we don't know the real distribution of the processes in this dataset and if a calibration procedure can somewhat correct the uncertainty about the unknown parameters, it does not correct model imperfection. We will present in the next chapter a new function $f(t;\beta)$ where $\dim(\beta)$ is determined by the dataset itself. We re-calibrated this dataset using this function with $\dim(\beta) = 2$ and the prediction intervals

were then adequate.

Chapter 5

Prediction of Warranty Claims

With products under warranty, manufacturers often collect detailed claims data. When this dataset is maintained properly, it is of interest to predict the eventual total number of warranty claims using the data already observed. In this chapter, we will apply to a warranty dataset the methods discussed previously for finite horizon prediction problems. The number of processes in this dataset being very large, we will focus on methods providing computationally rapid prediction intervals.

First, we will introduce the dataset of interest. Then, we will propose a prediction model and discuss the assumptions made. Once the numerical results are presented, we will discuss possible calibration approaches that are taking into account the significant size of this dataset.

5.1 Motivating Dataset

First presented in Kalbfleisch, Lawless & Robinson (1991), this dataset contains warranty information on one subsystem for 36,683 cars of one model type. Over a span of 571 days,

the following times were recorded for each car: production time, sale time, claim time(s), and the times at which these claims were reported. Each car had a one year or 12,000 mile warranty, whichever came first.

We let $N_i(t)$ be the total number of warranty claims for the ith car t days after it was sold. Note that $N_i(0)$ is not necessarily equal to 0 as some claims could occur between the production day and the day of sale. Throughout this chapter, the quantity to predict will be $\sum_i \mathbf{N}_i(365)$, the total number of warranty claims for this fleet of cars. To be able to assess the validity of our predictions, we will only consider cars for which $\mathbf{N}_i(365)$ is known, *i.e.* cars sold at least 365 days before the end of the data collection. The resulting dataset contains 15,775 cars. Each of them had between 0 and 10 claims for a total of 2,620 claims. Only 44 of these claims occurred before the day of sale. Table 5.1 shows the distribution of total claims amongst all the cars. Unlike the datasets studied so far, we can see most of the cars (89%) never had a warranty claim, and only a few cars (1%) had more than 2 claims before the end of the warranty. Figure 5.1 is a histogram of the claim occurrence times during the year where each car is potentially under warranty. It appears that the rate of occurrence of claims increases over the first 100 days after the sale and decreases after that point. We can see a relatively abrupt change after approximately 250 days where the frequency decreases more rapidly; a possible explanation for this may be that a significant number of cars are then no longer under warranty because of the mileage drop-out. We can also notice a smaller frequency of claims between 40 and 80 days after the sale.

When analyzing such massive datasets, it is often interesting to study a figure like Figure 5.2. Instead of being grouped, each occurrence time is now plotted. The claims for each car are represented along an invisible line. The first car produced appears at the bottom while the last one, produced 207 days later, appears at the top. Although it is

Number of claims	Number of cars
0	13,987
1	1,243
2	379
3	103
4	34
5+	29
2,620	15,775

Table 5.1: Number of cars with the same number of claims.

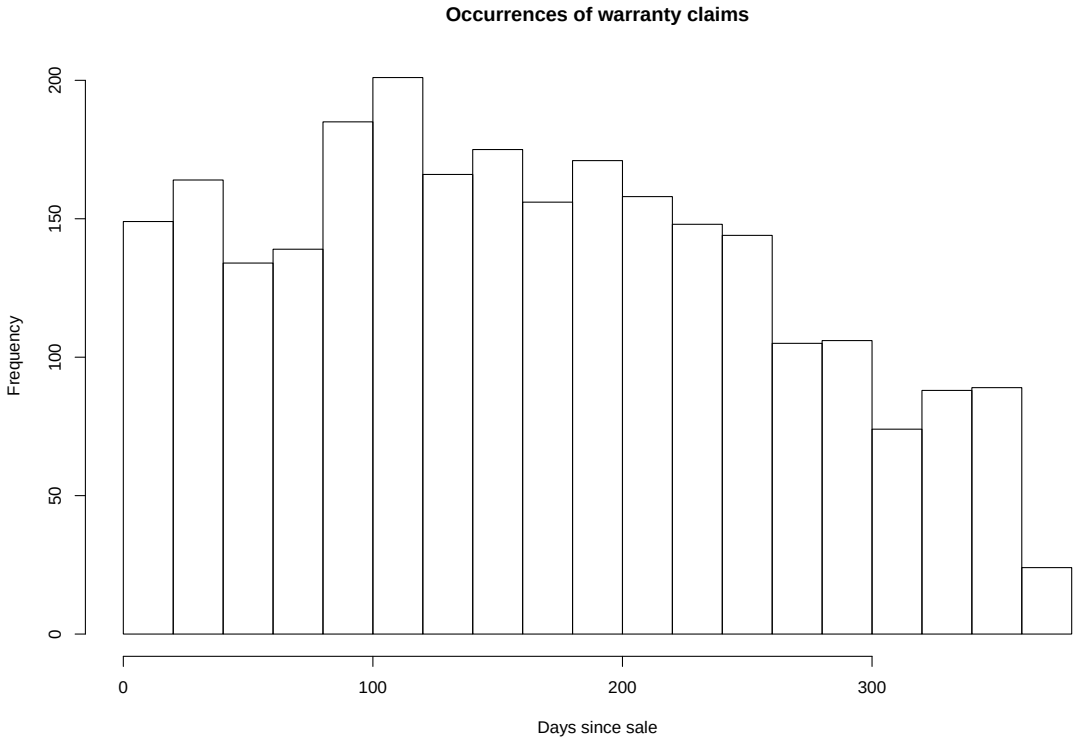

Figure 5.1: Histogram of the occurrence times.

less apparent, this figure indicates the same features as Figure 5.1. In addition, we can now see that the cars produced towards the end appear to have more claims than the first ones and that some cars manufactured during a certain early period had fewer claims than the others. We must keep in mind that most of the characteristics mentioned above are not known to the statistician facing this prediction problem. When all the information about the dataset is available, the quantity to be predicted is then completely determined. Therefore, the features mentioned will not be taken into account in our prediction model.

5.2 Prediction Model Proposed

We will now propose a model to predict the total number of warranty claims or, equivalently, the average number per vehicle. This model is a simple extension of model (4.4) which now takes into account that $\mathbf{N}_i(0)$, the number of claims before the ith car was sold, is not necessarily 0. First, we need to introduce the following notations:

$$\begin{aligned}
k &= \text{The total number of cars (15,775).} \\
\tau_{i,j} &= \text{The age of the } i\text{th car when the } j\text{th claim occurs.} \\
t_i^s &= \text{Sale time of the } i\text{th car.} \\
S_t &= \{i : t_i^s \leq t\} \\
&= \text{Set of cars already sold at time } t. \\
H_i(t) &= \text{Information available for car } i \text{ at time } t. \\
&= \begin{cases} \{t_i^s, N_i(0), N_i(0, t - t_i^s), \{\tau_{i,j} : j = 1, \ldots, N_i(0, t - t_i^s)\}\} & \text{if } i \in S_t, \\ \emptyset \quad \text{if } i \in S_t^c. \end{cases}
\end{aligned}$$

The prediction model proposed is then,

$$\mathbf{N}_i(0)|\alpha_i \sim \text{Poisson}(c\alpha_i),$$

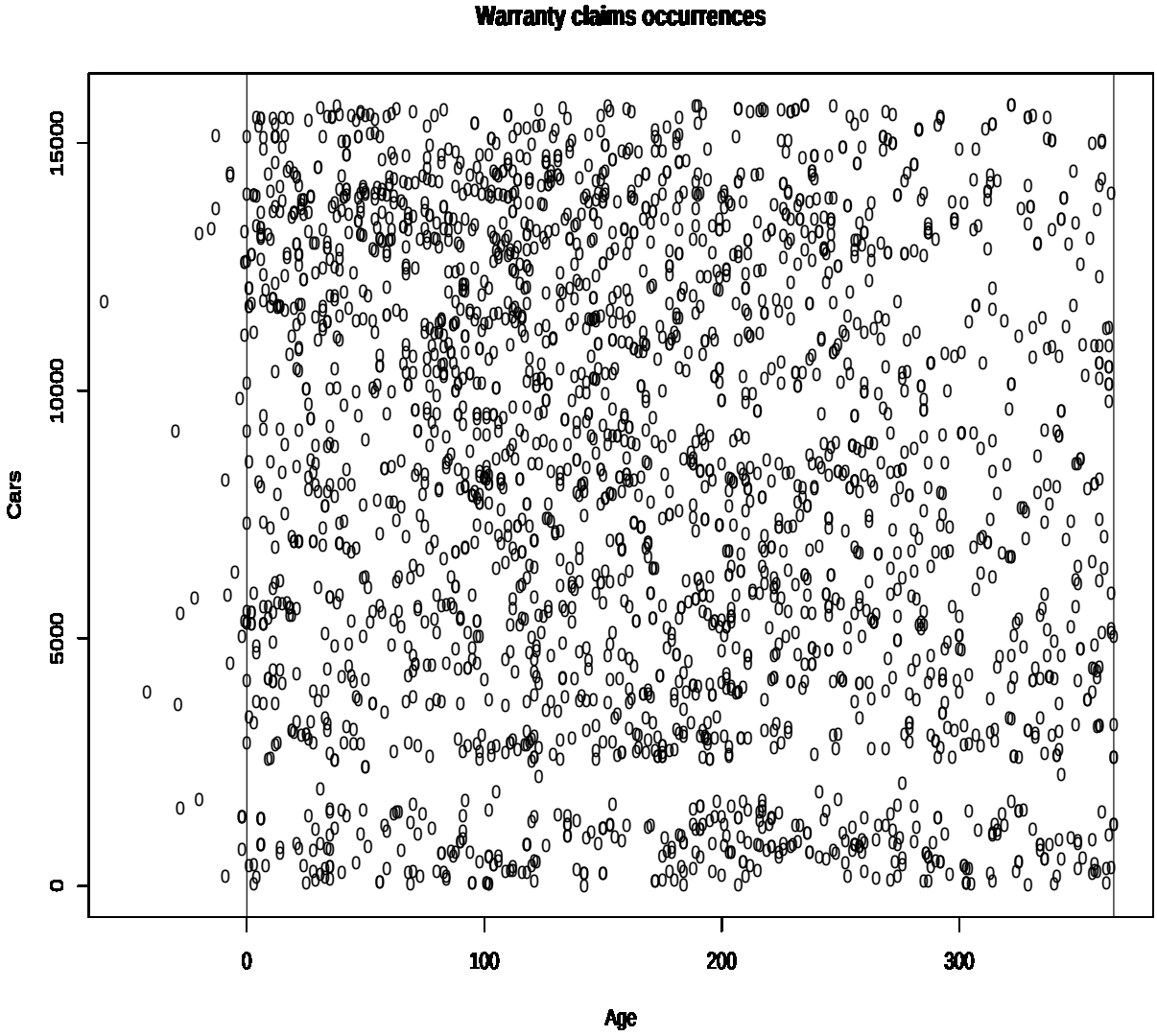

Figure 5.2: Warranty claims occurrences (time of sale is the origin).

$$\begin{aligned} \mathbf{N}_i(0,t)|\alpha_i &\sim \mathcal{PP}(\alpha_i f(t;\beta)), \\ \alpha_i &\sim \text{Gamma}(a,b), \end{aligned} \tag{5.1}$$

where $i = 1, \ldots, k$. In this new model, the unknown parameters are a, b, c, and β. All the assumptions made by using such a model, and suitable choices for $f(t;\beta)$, will be discussed later in this section.

Even with this extension, the predictive distribution will still be a convolution of negative binomials:

Proposition 5.1. *Using the model given in (5.1), the density function for the number of future warranty claims, given the information available at time t, is a convolution of $k - |S_{t-365}|$ negative binomials. $k - |S_t|$ of these negative binomials have parameters a and $b/(b+c+F(365;\beta))$, while the remaining $|S_t| - |S_{t-365}|$ ones have parameters $a + N_i(t - t_i^s)$ and $(b + c + F(t - t_i^s;\beta))/(b + c + F(365;\beta))$ where $i \in S_t \backslash S_{t-365}$.*

Proof. At time t, only the cars that have not been sold yet and those sold over the last 365 days are at risk of having any additional warranty claims. Therefore, the variable of interest is a sum of $k - |S_{t-365}|$ random variables:

$$\sum_{i \in S_t^c} \mathbf{N}_i(365) + \sum_{i \in S_t \backslash S_{t-365}} \mathbf{N}_i(t - t_i^s, 365), \tag{5.2}$$

where S_t^c is the set of cars still unsold at time t and $S_t \backslash S_{t-365}$ is the set of all cars sold the year before time t. When we condition on $\bigcup_{i=1}^k H_i(t)$, the total information available at time t, each random variable in the first summation of (5.2) has the density

$$\begin{aligned} p(n|H_i(t);a,b,c,\beta) &= p(n|\emptyset;a,b,c,\beta) \\ &= \int_{\alpha_i} P[\mathbf{N}_i(365) = n|\alpha_i;c,\beta]\pi(\alpha_i;a,b)d\alpha_i \end{aligned}$$

$$
\begin{aligned}
&= \int_{\alpha_i} \frac{\exp\{-\alpha_i(c+F(365;\beta))\}[\alpha_i(c+F(365;\beta))]^n}{n!} \times \\
&\qquad \frac{b^a \exp\{-\alpha_i b\}\alpha_i^{a-1}}{\Gamma(a)} d\alpha_i \\
&= \frac{\Gamma(a+n)}{n!\Gamma(a)} \left(\frac{b}{b+c+F(365;\beta)}\right)^a \left(\frac{c+F(365;\beta)}{b+c+F(365;\beta)}\right)^n,
\end{aligned}
$$

which is the density function of a negative binomial with parameters a and $b/(b+c+F(365;\beta))$.

As for the ith random variable in the second summation of (5.2), its density is

$$
\begin{aligned}
p(n|H_i(t);a,b,c,\beta) &= p(n|N_i(t-t_i^s);a,b,c,\beta) \\
&= \int_{\alpha_i} P[\mathbf{N}_i(t-t_i^s,365)=n|\alpha_i;c,\beta]\pi(\alpha_i|N_i(t-t_i^s);a,b)d\alpha_i \\
&= \int_{\alpha_i} \frac{\exp\{-\alpha_i(F(365;\beta)-F(t-t_i^s;\beta))\}[\alpha_i(F(365;\beta)-F(t-t_i^s;\beta))]^n}{n!} \times \\
&\qquad \frac{(b+c+F(t-t_i^s;\beta))^{a+N_i(t-t_i^s)}\exp\{-\alpha_i(b+c+F(t-t_i^s;\beta))\}}{\Gamma(a+N_i(t-t_i^s))} \times \\
&\qquad \alpha_i^{a+N_i(t-t_i^s)-1} d\alpha_i \\
&= \frac{\Gamma(a+N_i(t-t_i^s)+n)}{n!\Gamma(a+N_i(t-t_i^s))} \left(\frac{b+c+F(t-t_i^s;\beta)}{b+c+F(365;\beta)}\right)^{(a+N_i(t-t_i^s))} \times \\
&\qquad \left(\frac{F(365;\beta)-F(t-t_i^s;\beta)}{b+c+F(365;\beta)}\right)^n,
\end{aligned}
$$

which is the density function of a negative binomial with parameters $a+N_i(t-t_i^s)$ and $(b+c+F(t-t_i^s;\beta))/(b+c+F(365;\beta))$. □

Like the density given in Proposition 4.1, we can evaluate this density by using recursive formulae similar to the one given in Corollary 4.2 or by an approximation similar to (4.6). The number of claims to predict being quite large in this dataset, the approximate density will be used instead of the recursive formulae.

A model like (5.1) provides a relatively simple predictive distribution for the total number of claims. Such simplicity is especially important when the number of processes considered in the convolution is large. However, some assumptions have to be made in order to use this model. First, we assume that the total number of cars to be sold is known. It does not appear to be a very restrictive assumption. Based on previous years and early sales, the total number of cars to be produced is likely to be known quite rapidly and the number of unsold cars is probably negligible as the company will try to liquidate all cars unsold after a certain time. We are also assuming the absence of delay in the report of a claim. It is possible to extend our model to take this delay into account but the predictive density would then be more complex. In addition, this assumption should not be too restrictive if the reporting delays are reasonably small.

The two assumptions mentioned above will not affect the adequacy of the predictions made; they only affect the applicability of this model to other datasets. However, model (5.1) makes other assumptions that could affect the quality of the predictions. For example, it is expected, and verified as the data are collected, that the length of time between the production and the sale of a car has an impact on the eventual number of warranty claims. This is not directly accounted for in our prediction model but even a simple model extension to correct this would greatly complicate the predictive distribution. For example, let us consider the model:

$$\begin{aligned}
\mathbf{N}_i(0)|\alpha_i &\sim \text{Poisson}((t_i^s - t_i^p)c\alpha_i), \\
\mathbf{N}_i(0,t)|\alpha_i &\sim \mathcal{PP}((t_i^s - t_i^p)\alpha_i f(t;\beta)), \\
\alpha_i &\sim \text{Gamma}(a,b),
\end{aligned}$$

where t_i^p is the (known) production time of the ith car. Even if we estimate the distribution of $\mathbf{T}_i^s - t_i^p$ using simple Kaplan-Meier estimates, we can show that the predictive distribution

would then be a convolution of mixtures of many different negative binomials. Note also that since we are using the same random effect with $\mathbf{N}_i(0)$ and $\mathbf{N}_i(0,t)$, we are neglecting the driver effect. It would not greatly complicate the predictive distribution to unlink these effects but the results presented in this chapter are very similar under both models.

Another assumption that may affect the adequacy of our predictions is that the mileage drop-out is not accounted for. Our model assumes that each car is at risk during one complete year but it is not always the case; cars are no longer under warranty once they were used for more than 12,000 miles. We believe that this problem can be partially solved by using a flexible function $f(t;\beta)$ that would be decreasing quickly as we approach the end of the year.

We will now present different ways to estimate the unknown parameters a, b, c, and β. These estimates will substitute for the real parameters in the predictors and prediction intervals obtained from the distribution given by Proposition 5.1. First, we have to specify completely the form of the non-negative function $f(t;\beta)$. Since it is important to use a function that is very flexible, instead of using a function proposed in the literature where $\dim(\beta)$ is usually one or two, we are proposing the function

$$f(t;\beta) \;=\; \exp\{\beta_1 t + \beta_2 t^2 + \ldots + \beta_q t^q\}. \tag{5.3}$$

Even for moderate values of q, this function can have different changes in its (time) behavior with up to $q-1$ critical points. Note that in the cases where t and q are large, it is recommended to use $f(\log(1+t);\beta)$ instead of $f(t;\beta)$. A difficulty introduced by using this function is that $F(t;\beta) = \int_0^t f(u;\beta)du$ now has to be evaluated numerically when $q > 1$. However, simple numerical techniques like the trapezoidal method will give very adequate approximations of this function.

The estimates considered here will be the maximum likelihood estimates. Given the

information available at time t, the likelihood function is given by

$$\begin{aligned}
L(a,b,c,\beta|\bigcup_{i=1}^{k} H_i(t)) &= \int_{\underset{\sim}{\alpha}} L(\underset{\sim}{\alpha},c,\beta|\bigcup_{i=1}^{k} H_i(t))\pi(\underset{\sim}{\alpha};a,b)d\underset{\sim}{\alpha} \\
&= \prod_{i\in S_t}\int_{\alpha_i}\left[\frac{e^{-\alpha_i c}(\alpha_i c)^{N_i(0)}}{N_i(0)!}\right]\left(e^{-\alpha_i F(t-t_i^s;\beta)}\prod_{j=1}^{N_i(0,t-t_i^s)}\alpha_i f(\tau_{ij};\beta)\right)\times \\
&\quad \left[\frac{b^a e^{-\alpha_i b}\alpha_i^{a-1}}{\Gamma(a)}\right]d\alpha_i \\
&= \prod_{i\in S_t}\frac{c^{N_i(0)}b^a(\prod_{j=1}^{N_i(0,t-t_i^s)}f(\tau_{ij};\beta))}{N_i(0)!\Gamma(a)}\int_{\alpha_i} e^{-\alpha_i(c+F(t-t_i^s;\beta)+b)}\alpha_i^{N_i(0)+a-1}d\alpha_i \\
&= \prod_{i\in S_t} c^{N_i(0)}\left(\prod_{j=1}^{N_i(0,t-t_i^s)}f(\tau_{ij};\beta)\right)\left(\frac{\Gamma(a+N_i(t-t_i^s))}{N_i(0)!\Gamma(a)}\right)\times \\
&\quad \left[\frac{b^a}{(b+c+F(t-t_i^s;\beta))^{a+N_i(t-t_i^s)}}\right]. \qquad (5.4)
\end{aligned}$$

This likelihood function will also be used to find a suitable value for q, the dimension of the vector of unknown parameters β. We will initially fit model (5.1) with $q = 1$ and iterate as long as there is a substantial improvement in the likelihood function (5.4). Using the well-known fact that

$$-2\log\left(L(\hat{a},\hat{b},\hat{c},(\hat{\beta}_1,\ldots,\hat{\beta}_{q+1})|\bigcup_{i=1}^{k}H_i(t)) - L(\hat{a},\hat{b},\hat{c},(\hat{\beta}_1,\ldots,\hat{\beta}_q)|\bigcup_{i=1}^{k}H_i(t))\right) \qquad (5.5)$$

converges toward a $\mathcal{X}_1^2$ distribution as t goes to infinity, we will iterate q as long as (5.5) is greater than 3.84, the 95% quantile of this distribution. Our prediction model then has the great advantage that the number of parameters used in the function $f(t;\beta)$ is determined by the dataset itself.

5.3 Model Fitting

In this section, we will discuss issues regarding the fitting of our proposed prediction model. We will first propose ways to reduce the computational time required to obtain maximum likelihood estimates by using a conditional likelihood function and adequate starting values. Then, we will assess the adequacy of the fitted model and present the prediction intervals obtained.

5.3.1 Starting Values

With $q+3$ parameters and up to 15,775 processes, the time required by non-linear maximization routines to obtain the maximum likelihood estimates is non-negligible. However, we can reduce this time substantially by using a conditional likelihood function for the vector β. Let $g(\cdot)$ be the joint density function of the ages at which the ith car has been repaired over the interval $(0, t-t_i^s]$, the likelihood function can then be rewritten as

$$\begin{aligned} L\left(a,b,c,\beta|\bigcup_{i=1}^{k} H_i(t)\right) &= \prod_{i\in S_t} P[\mathbf{N}_i(0)=N_i(0);a,b,c]\times \\ &\quad g(\tau_{i,1},\ldots,\tau_{i,\mathbf{N}_i(0,t-t_i^s)};a,b,c,\beta) \\ &= \prod_{i\in S_t} P[\mathbf{N}_i(0)=N_i(0),\mathbf{N}_i(0,t-t_i^s)=N_i(0,t-t_i^s);a,b,c,\beta]\times \\ &\quad g(\tau_{i,1},\ldots,\tau_{i,N_i(0,t-t_i^s)}|N_i(0,t-t_i^s);a,b,c,\beta) \\ &= \left(\prod_{i\in S_t} P[\mathbf{N}_i(0)=N_i(0),\mathbf{N}_i(0,t-t_i^s)=N_i(0,t-t_i^s);a,b,c,\beta]\right)\times \\ &\quad \left(\prod_{i\in S_t}\prod_{j=1}^{N_i(0,t-t_i^s)} \frac{f(\tau_{ij};\beta)}{F(t-t_i^s;\beta)}\right), \end{aligned} \tag{5.6}$$

where $P[\mathbf{N}_i(0)=N_i(0),\mathbf{N}_i(0,t-t_i^s)=N_i(0,t-t_i^s);a,b,c,\beta]$ is the product of the probability function of 2 negative binomials with parameters $(a,\frac{b}{b+c})$ and $(a+N_i(0),\frac{b+c}{b+c+F(t-t_i^s;\beta)})$

respectively. The likelihood function is then the product of

$$L(a,b,c,\beta|N(0),N(0,t^s)) \;=\; P[\mathbf{N}(0),\mathbf{N}(0,t-t^s);a,b,c,\beta] \tag{5.7}$$

and

$$L_c(\beta|\tau(t)) \;=\; \prod_{i\in S_t}\prod_{j=1}^{N_i(0,t-t_i^s)}\frac{f(\tau_{ij};\beta)}{F(t-t_i^s;\beta)}, \tag{5.8}$$

where

$$\tau(t) \;=\; \{\tau_{ij} : i\in S_t, j=1,\ldots,N_i(0,t-t_i^s)\}.$$

One can see that $L_c(\beta|\tau(t))$ is the likelihood function for β when we condition on the total number of claims per car. Empirical studies suggest that (5.7) depends little on β. This means that the parameter β maximizing (5.8) will be close to the $\hat{\beta}$ obtained by maximizing (5.6). Thus, we are recommending to first find the one maximizing $L_c(\beta|\tau(t))$ and use this value (say $\hat{\beta}_c$) as a starting value to maximize the original likelihood. These two estimates being similar, we are essentially maximizing q parameters and then 3 instead of $q+3$ parameters all at once, a strategy that often leads to a substantial decrease in of computational time. Note that this conditional likelihood function can also be used instead of the original one to find an optimal value for q.

Once $\hat{\beta}_c$ is found, we can also obtain starting values for a, b, and c in the original likelihood. Let $\mathbf{R}_i$ be $\mathbf{N}_i(0,t-t_i^s)/F(t-t_i^s;\hat{\beta}_c)$, we then have

$$\begin{aligned}
\mathbb{E}[\mathbf{N}_i(0)] &= \mathbb{E}[\mathbb{E}[\mathbf{N}_i(0)|\alpha_i]] \\
&= c\,\frac{a}{b}, \\
\mathbb{E}[\mathbf{R}_i|N_i(0)] &= \mathbb{E}[\mathbb{E}[\mathbf{R}_i|\alpha_i]|N_i(0)] \\
&= \frac{a+N_i(0)}{b+c},
\end{aligned}$$

and

$$\begin{aligned}\mathbb{V}ar[\mathbf{R}_i|N_i(0)] &= \mathbb{E}[\mathbb{V}ar[\mathbf{R}_i|\alpha_i]|N_i(0)] + \mathbb{V}ar[\mathbb{E}[\mathbf{R}_i|\alpha_i]|N_i(0)] \\ &= \left(\frac{a+N_i(0)}{b+c}\right)\frac{1}{F(t-t_i^s;\hat{\beta}_c)} + \frac{a+N_i(0)}{(b+c)^2}.\end{aligned}$$

Using a technique similar to the moment matching approach presented in Section 3.3.2 we obtain the starting values:

$$\begin{aligned}\hat{a}_0 &= \frac{\overline{F}(t-t^s;\hat{\beta}_c)\overline{R}^2}{\overline{F}(t-t^s;\hat{\beta}_c)S_R^2 - \overline{R}} - \overline{N}(0), \\ \hat{b}_0 &= \frac{\hat{a}_0}{\overline{R}}, \\ \hat{c}_0 &= \frac{\overline{N}(0)}{\overline{R}}.\end{aligned} \tag{5.9}$$

Where

$$\begin{aligned}\overline{N}(0) &= \frac{\sum_{i\in S_t} N_i(0)}{|S_t|}, \\ \overline{F}(t-t^s;\hat{\beta}_c) &= \frac{\sum_{i\in S_t} F(t-t_i^s;\hat{\beta}_c)}{|S_t|},\end{aligned}$$

and $\overline{R}$ and S_R^2 are respectively the sample mean and variance of the R_i's.

To shorten the time required to obtain the maximum likelihood estimates, we can also reparameterize $f(t;\beta)$. The way this function is defined in (5.3) leads to high correlations between the components of $\hat{\beta}$ (or $\hat{\beta}_c$). Furthermore, a certain $\hat{\beta}_i$ obtained will differ substantially from the $\hat{\beta}_i$ obtained when we will iterate the value of q. This means that when we will maximize $L_c(\beta|\tau(t))$ with $\dim(\beta) = q+1$, we will not be able to use the vector $(\hat{\beta}_1,\ldots,\hat{\beta}_q)$ as a starting value for the first q components of the new vector $(\beta_1,\ldots,\beta_{q+1})$. To correct these problems, we are suggesting the following reparameterization:

$$f(t;\beta) = \exp\{\beta_1 L_1(t) + \ldots + \beta_q L_q(t)\}, \tag{5.10}$$

where

$$L_n(t) \;=\; e^t \frac{\partial^n}{\partial t^n}(t^n e^{-t}).$$

These polynomials, called Laguerre polynomials, are of interest here because they are orthogonal with respect to the inner product $< L_n, L_m >= \int_0^\infty e^{-t} L_n(t) L_m(t) dt$. In addition, they are easily obtained via the recursive formula

$$L_{n+2}(t) \;=\; [2(n+1) - t + 1] L_{n+1}(t) - (n+1)^2 L_n(t).$$

Using this reparameterization, the correlation between the components of $\hat{\beta}$ are reduced and the vector $((\hat{\beta}_1, \ldots, \hat{\beta}_q), 0)$ now provides better starting values to find $(\hat{\beta}_1, \ldots, \hat{\beta}_{q+1})$.

The following algorithm summarizes how to obtain maximum likelihood estimates when we incorporate the features mentioned in this subsection:

Step $q = 1$

- Find $\hat{\beta}_c$ using 0 as a starting value.
- Using this $\hat{\beta}_c$, find $(\hat{a}_0, \hat{b}_0, \hat{c}_0)$ given by (5.9).
- Find $(\hat{a}, \hat{b}, \hat{c}, \hat{\beta})$ using $(\hat{a}_0, \hat{b}_0, \hat{c}_0, \hat{\beta}_c)$ as starting values.

Step $q = i$ $(i \geq 2)$

- Let $\hat{\beta}_c^{i-1}$ be the $\hat{\beta}_c$ obtained in the previous step. We now obtain $\hat{\beta}_c$ using $(\hat{\beta}_c^{i-1}, 0)$ as starting values.
- If $-2 \log[L_c(\hat{\beta}_c) - L_c((\hat{\beta}_c^{i-1}, 0))] < 3.84$, we stop the algorithm and use the $(\hat{a}, \hat{b}, \hat{c}, \hat{\beta})$ found in the previous step. Otherwise, we use this new $\hat{\beta}_c$ to find $(\hat{a}_0, \hat{b}_0, \hat{c}_0)$.
- Find $(\hat{a}, \hat{b}, \hat{c}, \hat{\beta})$ using $(\hat{a}_0, \hat{b}_0, \hat{c}_0, \hat{\beta}_c)$ as starting values.

5.3.2 Numerical Results

In order to correct issues like range restrictions or inappropriate scalings, model (5.1) was reparameterized as

$$\begin{aligned} \mathbf{N}_i(0)|\alpha_i^{'} &\sim \text{Poisson}(\exp\{c^{'}\}\alpha_i^{'}), \\ \mathbf{N}_i(0,t)|\alpha_i^{'} &\sim \mathcal{PP}(\exp\{a^{'}\}\alpha_i^{'} f(t;\beta^{'})), \\ \alpha_i^{'} &\sim \text{Gamma}(\exp\{-b^{'}\}, \exp\{-b^{'}\}), \end{aligned} \tag{5.11}$$

where

$$f(t;\beta^{'}) = \exp\{\beta_1^{'} L_1(\log(1+t)) + \ldots + \beta_q^{'} L_q(\log(1+t))\}.$$

Under this reparameterization, $a^{'}$, $b^{'}$, and $c^{'}$ now represent the logarithms of $\mathbb{E}[\mathbf{N}_i(0,t)/F(t;\beta^{'})]$, $\mathbb{V}ar[\alpha_i^{'}]$, and $\mathbb{E}[\mathbf{N}_i(0)]$ respectively. Note that we now have, without loss of generality, $\mathbb{E}[\alpha_i^{'}] = 1$. Until the end of this chapter, we will always use these parameters instead of the previous ones. Thus, we will omit the $'$ superscript from now on to simplify the notation.

We develop predictions of the total number of warranty claims after 100 days and every subsequent 50 days. Table 5.2 shows the number of warranty claims observed at times $t = 100, 150, \ldots, 550$ of this 571 day long longitudinal process. Based on the information available at each given time, the optimal value for $q = \dim(\hat{\beta})$ was 1 after 100 and 150 days, 2 after 200 days, and 4 thereafter.

Table 5.3 shows the estimates obtained for a, b, and c throughout time. Although there is still a certain amount of uncertainty about $\hat{a}$ at the end of the study, the approximate 95% confidence interval being (-9.78,-8.24), its point estimation does not vary much during the study; with only 33% of the claims observed at $t = 250$, $\hat{a} = -8.90$ is very close to $\hat{a} = -9.01$ obtained at the end. As opposed to $\hat{a}$, there is only a small amount of

Time	Number of claims	%	$q = \dim(\hat{\beta})$
100	49	1.9%	1
150	184	7.0%	1
200	457	17.4%	2
250	874	33.4%	4
300	1,392	53.1%	4
350	1,791	68.4%	4
400	2,160	82.4%	4
450	2,426	92.6%	4
500	2,555	97.5%	4
550	2,615	99.8%	4
571	2,620	100.0%	4

Table 5.2: Number of claims observed at every given time.

Time	$\hat{a}$	$\hat{b}$	$\hat{c}$
100	-8.30	3.13	-6.16
150	-7.91	3.06	-6.10
200	-8.04	2.60	-5.89
250	-8.90	2.36	-5.88
300	-8.86	2.00	-5.88
350	-9.02	1.84	-5.88
400	-8.84	1.71	-5.88
450	-8.82	1.68	-5.88
500	-8.98	1.64	-5.88
550	-9.01	1.63	-5.88
571	-9.01	1.63	-5.88

Table 5.3: Estimates of a, b, and c.

Time	$\hat{\beta}$	$\hat{\beta}_c$
250	(-0.31, -0.53, 0.12, -0.04)	(-0.44, -0.40, 0.07, -0.03)
300	(-0.40, -0.47, 0.10, -0.04)	(-0.48, -0.22, 0.00, -0.02)
350	(-0.40, -0.57, 0.13, -0.04)	(-0.40, -0.36, 0.06, -0.03)
400	(-0.38, -0.48 , 0.10, -0.04)	(-0.34, -0.32, 0.06, -0.02)
450	(-0.36, -0.47, 0.10, -0.03)	(-0.33, -0.38, 0.08, -0.03)
500	(-0.40, -0.55, 0.12, -0.04)	(-0.40, -0.53, 0.11, -0.04)
550	(-0.40, -0.56, 0.13, -0.04)	(-0.39, -0.56, 0.12, -0.04)
571	(-0.40, -0.57, 0.13, -0.04)	(-0.40, -0.57, 0.13, -0.04)

Table 5.4: Estimates of β when $q = 4$.

uncertainty about $\hat{b}$ at the end of the study, its approximate 95% confidence interval being (1.53,1.73), and at least two thirds of the claims had to be observed before we could obtain a point estimate similar to the final one. Finally, since all the $N_i(0)$'s were observed in the first 250 days, we obtained stable and relatively precise estimates for $c = \log \mathbb{E}[\mathbf{N}_i(0)]$ quite rapidly.

For every time where $q = 4$ was chosen, Table 5.4 gives the estimate of β obtained and its initial value $\hat{\beta}_c$. As t increases, the difference between these estimates appears to be negligible. Such result suggests that little would be lost by estimating β only through its conditional likelihood and then to estimate a, b, and c via the profile likelihood $L_p(a, b, c|\hat{\beta}_c, N(0), N(0, t^s)) = L(a, b, c, \hat{\beta}_c|N(0), N(0, t^s))$ that can be derived from (5.7). This procedure should reduce the computational time unless it is desired to estimate the covariance between $(\hat{a}, \hat{b}, \hat{c})$ and $\hat{\beta}$.

Table 5.5 presents some results on the appropriateness of using a reparameterization

With reparameterization				Without reparameterization			
1.00	0.83	-0.01	0.65	1.00	-0.98	0.95	-0.91
	1.00	-0.53	0.96		1.00	-0.99	0.97
		1.00	-0.73			1.00	-0.99
			1.00				1.00

Table 5.5: Correlation matrix of $\hat{\beta}$ (t=571).

of $f(t;\beta)$ with Laguerre polynomials (5.10) instead of (5.3). One of the reasons why we suggested such parameterization was to reduce the correlations between the $\hat{\beta}_i$'s. This table shows that although some correlations are still large with the new parameterization, there is a clear amelioration from the original model.

To conclude this subsection, we will present some methods to asses how well model (5.11) fits our dataset. First we compared the distribution of the total claims amongst all the cars and their corresponding fitted values. Table 5.6 presents these results for some values of t. The only systematic departure is that our model always overestimates the number of cars without any claims. However, since the number of parameters estimated is greater than the possible modalities for the $N_i(t-t_i^s)$'s, we cannot use statistics like Pearson's goodness-of-fit statistic to test if such departure is significant. The only exception is at the end of the study since all the $\mathbf{N}_i(t-t_i^s)$'s are then identically distributed as negative binomials with parameters $\exp\{b\}$ and $\exp\{b\}/(\exp\{b\}+\exp\{c\}+\exp\{a\}F(365;\beta))$. We can then use Pearson's statistic and treat $\exp\{b\}/(\exp\{b\}+\exp\{c\}+\exp\{a\}F(365;\beta))$ as a single unknown parameter. The p-value of this test is 15.0% and thus our model seems adequate to model the $\mathbf{N}_i(365)$'s.

In addition to the total number of claims for each car, we are also interested to as-

$t = 200$	0	1	2	3	4+			
Obs.	14,767	321	45	10	3			
Fitted	14,798	285	48	11	4			
$t = 400$	0	1	2	3	4	5	6+	
Obs.	14,229	1,120	311	72	28	8	7	
Fitted	14,255	1,102	280	88	31	11	7	
$t = 571$	0	1	2	3	4	5	6	7+
Obs.	13,987	1,243	379	103	34	15	8	6
Fitted	14,008	1,249	339	113	41	16	6	5

Table 5.6: Distribution of the total claims amongst all the cars.

sess the validity of our function $f(t;\beta)$ to model the occurrence times. It is clear from Proposition 4.1, that at a given time t the set of all

$$u_{ij} = \frac{F(\tau_{ij};\beta)}{F(t - t_i^s;\beta)},$$

where $i = 1, \ldots, k$ and $j = 1, \ldots, N_i(0, t - t_i^s)$, would form a sample of independent observations uniformly distributed on $[0, 1]$. Therefore, if $\hat{\beta}$ is a precise estimate of β, the sample of all

$$\hat{u}_{ij} = \frac{F(\tau_{ij};\hat{\beta})}{F(t - t_i^s;\hat{\beta})},$$

should behave like a sample of independent uniforms when the model is right. Figure 5.3 shows the empirical quantiles of the $\hat{u}_{ij}$'s versus the theoretical quantiles of an $U(0, 1)$ using the complete dataset. This figure clearly suggests that $f(t;\beta)$ models adequately the τ_{ij}'s. Figure 5.4 shows this quantile-quantile plot when the model is fitted at times t=150, 250, 350, and 450. We can see the $\hat{u}_{ij}$'s also seem to be close to $U(0, 1)$ at some early points in

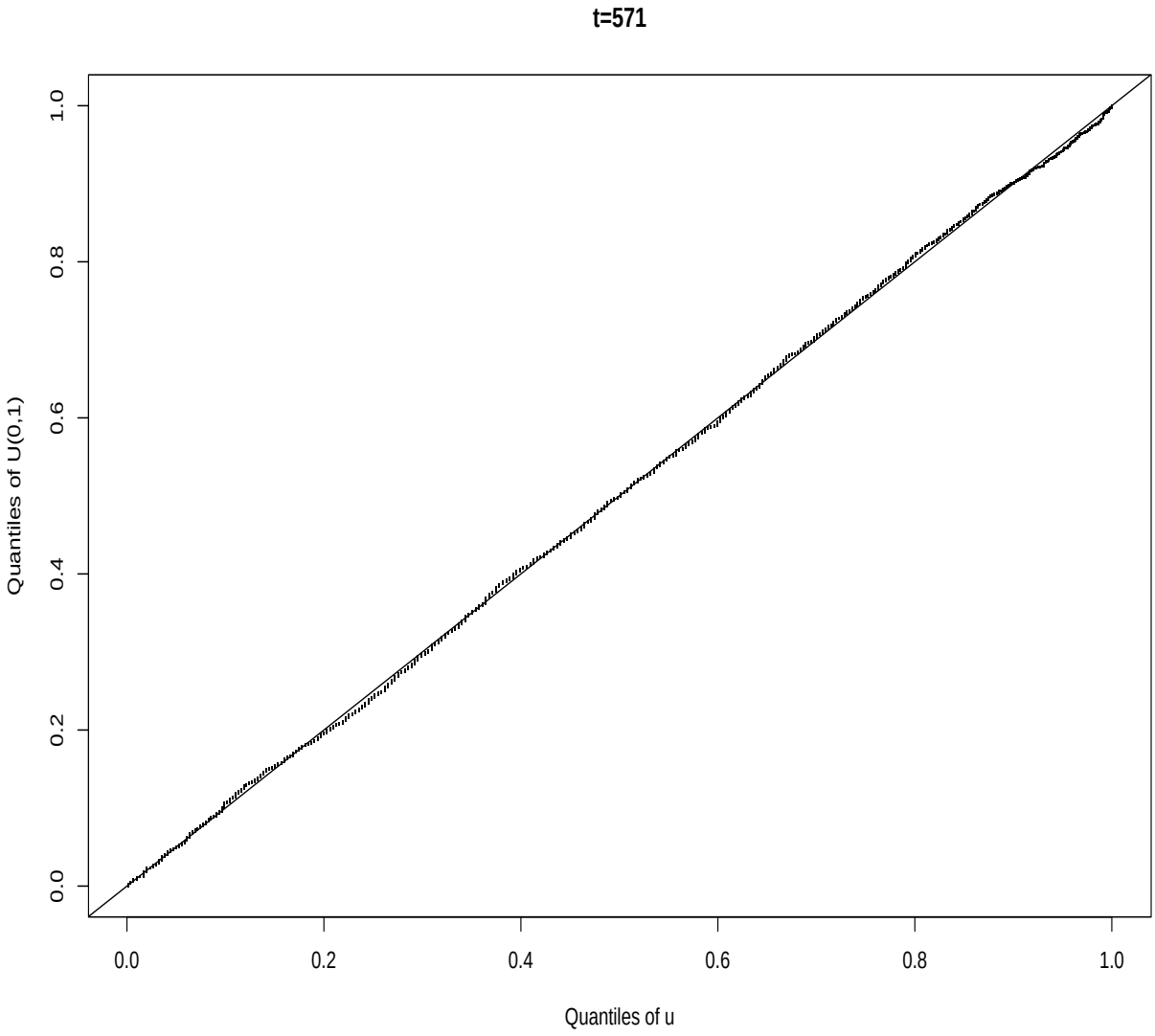

Figure 5.3: Quantile-quantile plot of the $\hat{u}_{ij}$'s (t=571).

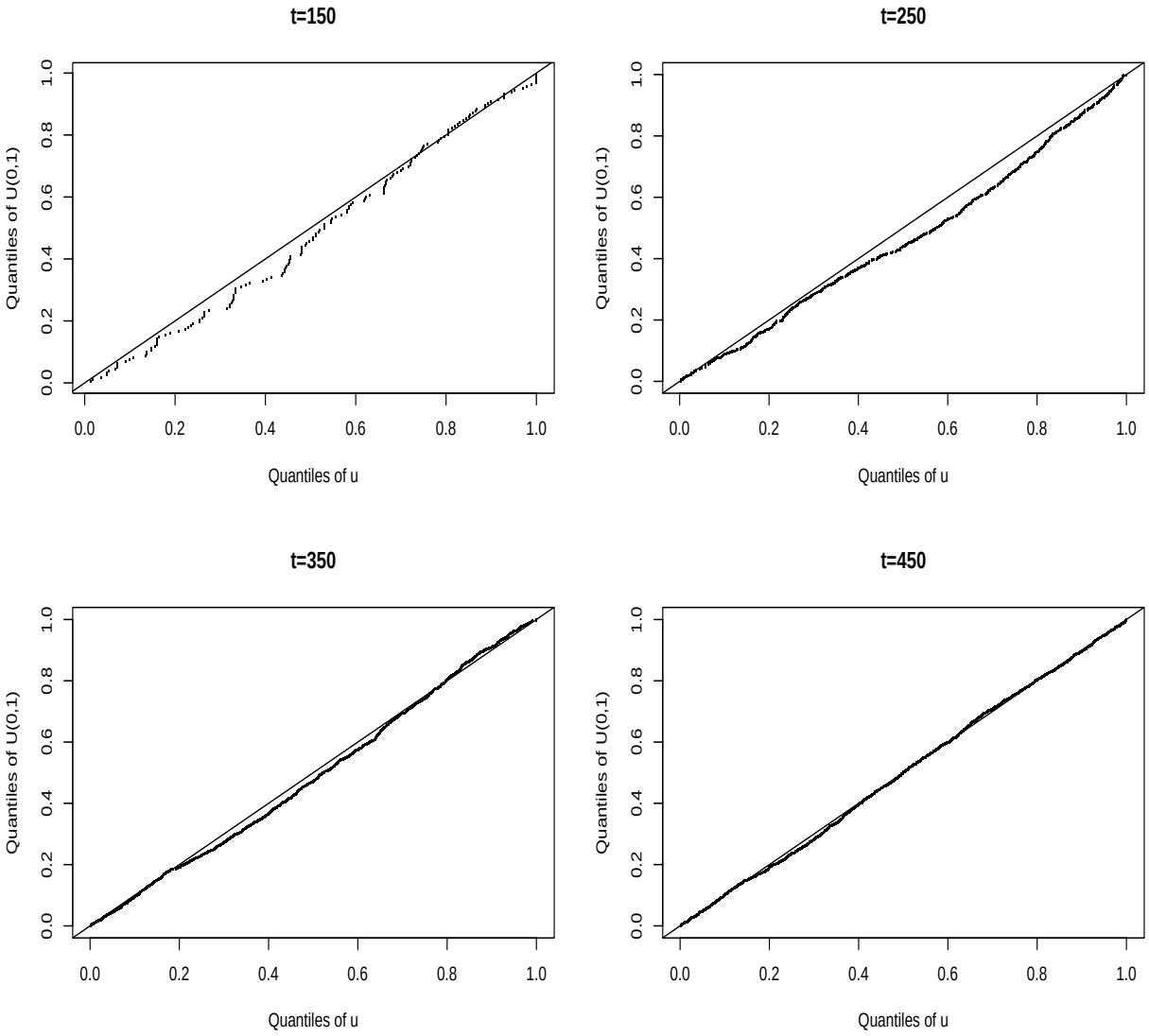

Figure 5.4: Quantile-quantile plots of the $\hat{u}_{ij}$'s.

the study.

At the end of the study, all the τ_{ij}'s are identically distributed with density function $f(t;\beta)/F(365;\beta)$. Thus, we also assessed the adequacy of $f(t;\beta)$ by comparing the estimated density with the histogram of the occurrence times. This is done in Figure 5.5 where the estimated density is plotted with q=1, 2, 3, and 4. It is clear in this figure that the estimated density with $q = 4$ is more adequate than the densities with smaller q. It is interesting to study the behavior of this function with 4 critical points: it suggests that buyers are using their warranty privileges very early after they bought their vehicles but less in the next couple of weeks. After that, our model suggests an increase in the number of claims for approximately 4-5 months and a decrease thereafter, probably because of the mileage drop-out.

Figure 5.6 shows that the adequacy of $f(t;\beta)$ when the model is fitted at earlier times also appears to be good. We should note that at these earlier times, the cars were not observed for the same amount of time and so late warranty claims are under-represented in these histograms. However this was corrected by comparing them with the function

$$f^*(u;\beta) \;=\; \frac{\sum_{i\in S_{t-u}} f(u;\beta)}{\sum_{i=1}^{k} F(t-t_i^s;\beta)}$$

instead of $f(t;\beta)$.

5.3.3 Prediction Intervals

Since our model seems adequate, we can now look at its ability to predict $\sum_{i=1}^{k} \mathbf{N}_i(365)$. The upper panel of figure 5.7 shows 95% non-calibrated plug-in prediction intervals using the information available after $100, 150, \ldots, 550$ days. The digit next to each interval is the optimal value of q at that point. Early prediction intervals are clearly unsatisfactory but the model starts to give better intervals halfway through the longitudinal study, when at

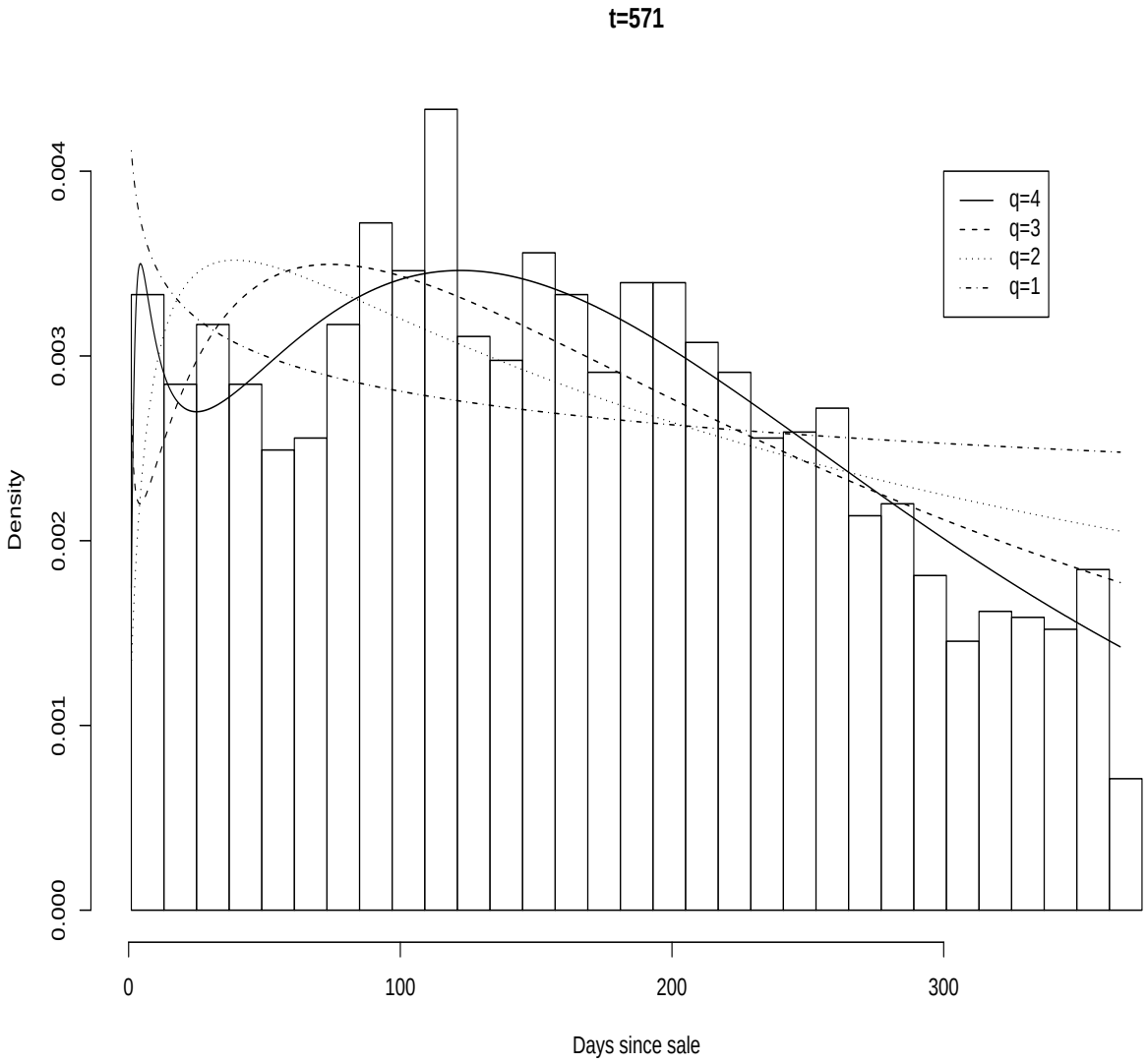

Figure 5.5: Histogram of the occurrence times (t=571).

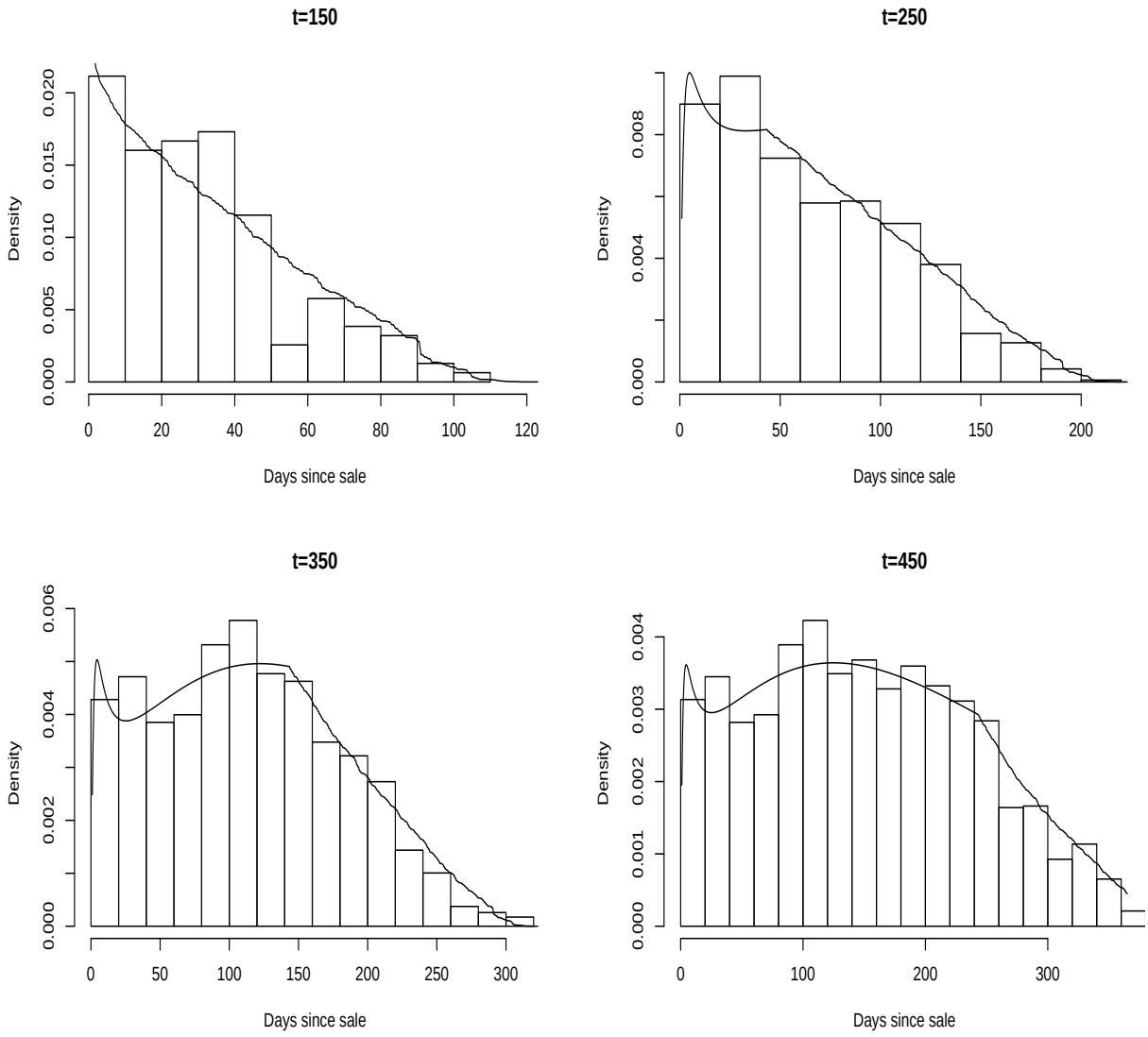

Figure 5.6: Histograms of the occurrence times.

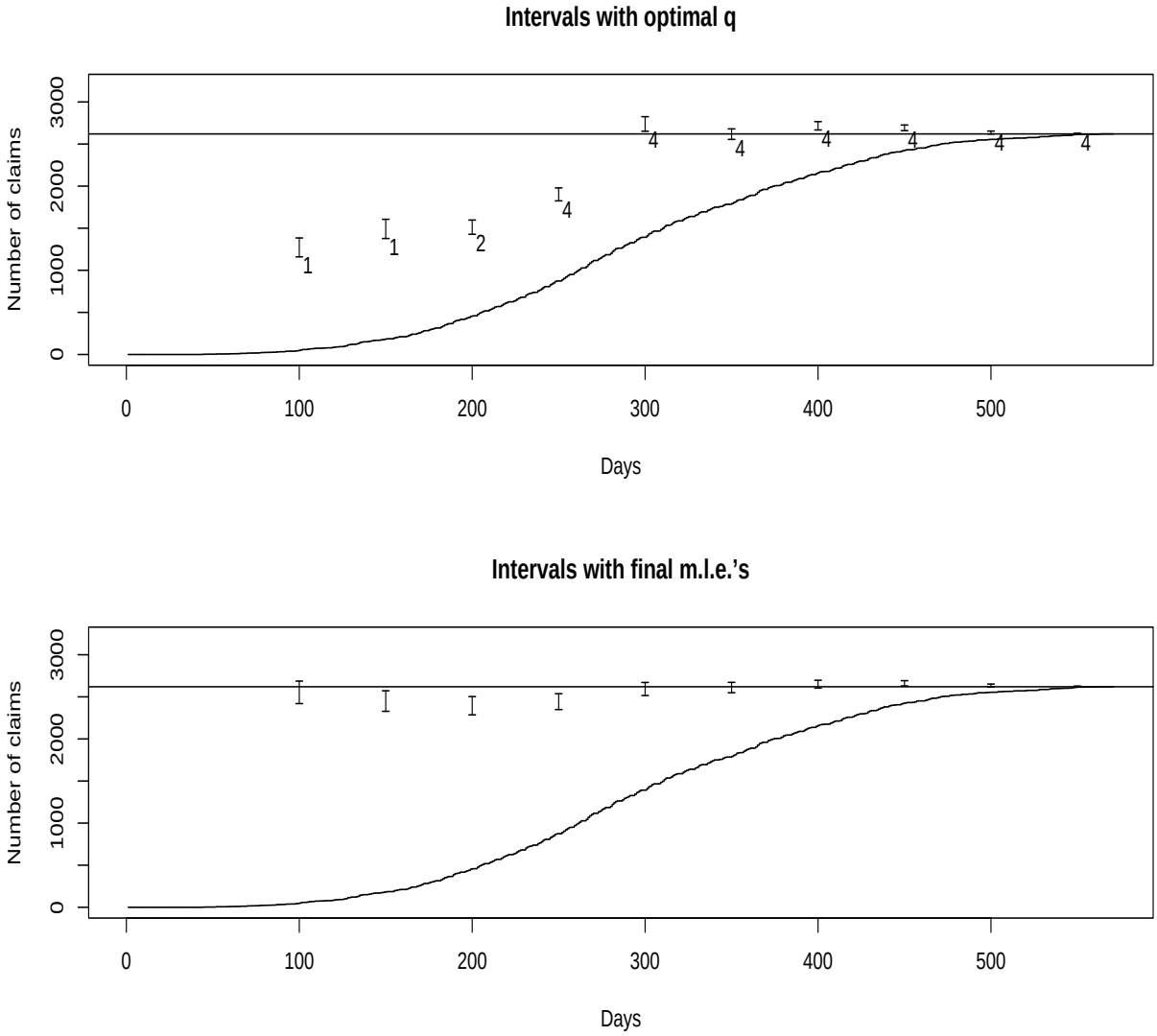

Figure 5.7: Non-calibrated 95% prediction intervals

least 50% of the claims are observed. Factors other than model imperfection can explain the poor performance of the early prediction intervals. For example, some of these intervals were obtained when few claims were observed and therefore provide inadequate m.l.e.'s to use with a plug-in method. In addition, we saw in Figure 5.2 that a group of cars did not behave like other cars in the sense that they had very few claims. Since these cars were produced at an early period, this could partially explain the under-prediction early in the study. The lower panel of Figure 5.7 also seems to indicate that the early under-prediction is more due to poor estimation of the unknown parameters than model imperfection. This panel shows 95% non-calibrated plug-in prediction intervals using the m.l.e.'s obtained with the complete dataset and we can see that early prediction would be adequate if good estimates could be used.

Even if they usually do not include the real value, we believe that the prediction intervals obtained after 300 days are predicting well. These non-calibrated intervals are quite close to the true value considering that we are ignoring the sampling variability in the estimation of 7 unknown parameters. We will calibrate these intervals in the next section and see that they will then contain the real value.

Figure 5.8 shows 95% non-calibrated plug-in prediction intervals for $q = 1, \ldots, 4$. It is interesting to notice that except for 2 cases ($t = 200$ and $t = 250$) the optimal value of q is giving here the best prediction intervals (*i.e.* closer to the true value). Also, when t is moderate or large the values of the likelihood functions based on models with $q = 4$ were much greater than models with $q < 4$ and we can see in Figure 5.8 that the models with $q = 4$ are providing much better prediction intervals than models with $q < 4$.

We think that the main advantage of using polynomials to model $\log f(t; \beta)$ is that the dimension of β can then be easily determined by the dataset itself. When comparing different types of function where β has the same dimension, we have no reason to believe

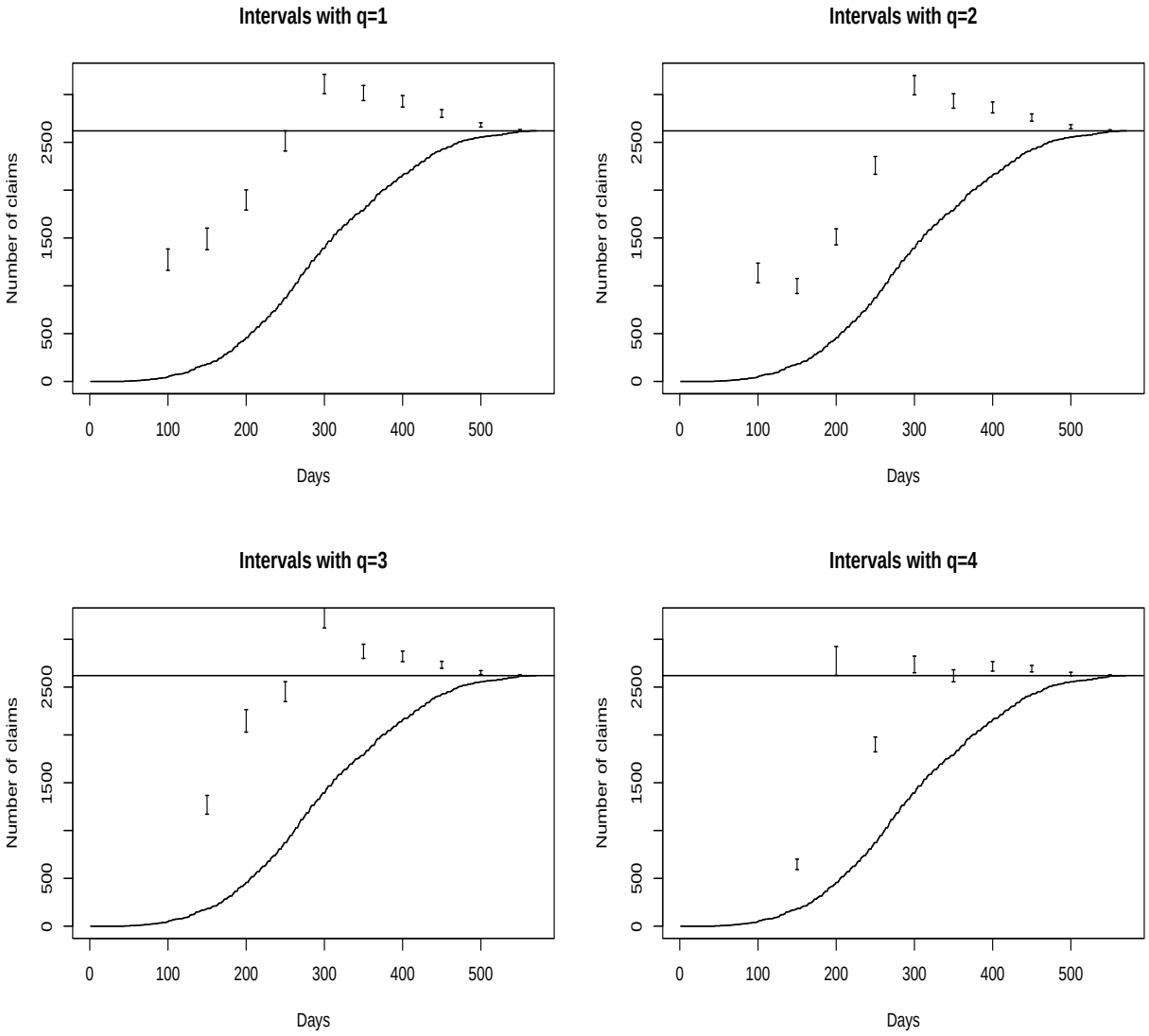

Figure 5.8: Non-calibrated 95% prediction intervals using $q = 1, \ldots, 4$

a priori that our log-polynomial model would be the most adequate. For example, we can see on Figure 5.9 95% non-calibrated plug-in prediction intervals using the EXP, GAM, LOG and POW models presented in Table 4.2. For all these models $\dim(\beta) = 1$ and we can see that some of these models are predicting as well as our model with $q = 1$. In practice, one should explore various parametric families in order to find one that describes the problem at hand well.

5.4 Calibration

From a theoretical point of view, it is clear that calibrated plug-in prediction intervals are more adequate than simple plug-in prediction intervals (see Theorem 2.1 for an example). However, practical prediction problems are mostly handled using non-calibrated intervals. Foregoing the fact that some scientists could be unaware of this procedure, the main reason why intervals are not calibrated is because this procedure can be relatively time consuming. This is especially true in problems like ours here, where the datasets are very large.

We will calibrate our intervals using the algorithm described in Section 2.3. The time required to perform this algorithm can be divided into three categories: the time required to simulate B datasets, the one required to obtain B sets of m.l.e.'s, and the time needed to approximate the predictive density by simulating convolutions of gammas. Amongst these three categories, the last one is of lesser importance and can be improved upon, when t is large enough, by using a recursive formula to calculate the predictive density. As for the first two categories, we will propose in this section some ways to reduce the time needed for each of them.

When we are simulating datasets, the time required to simulate the (gamma) random effects α_i and the (Poisson) counts $N_i(t - t_i^s)$ is negligible. However, the simulation of the

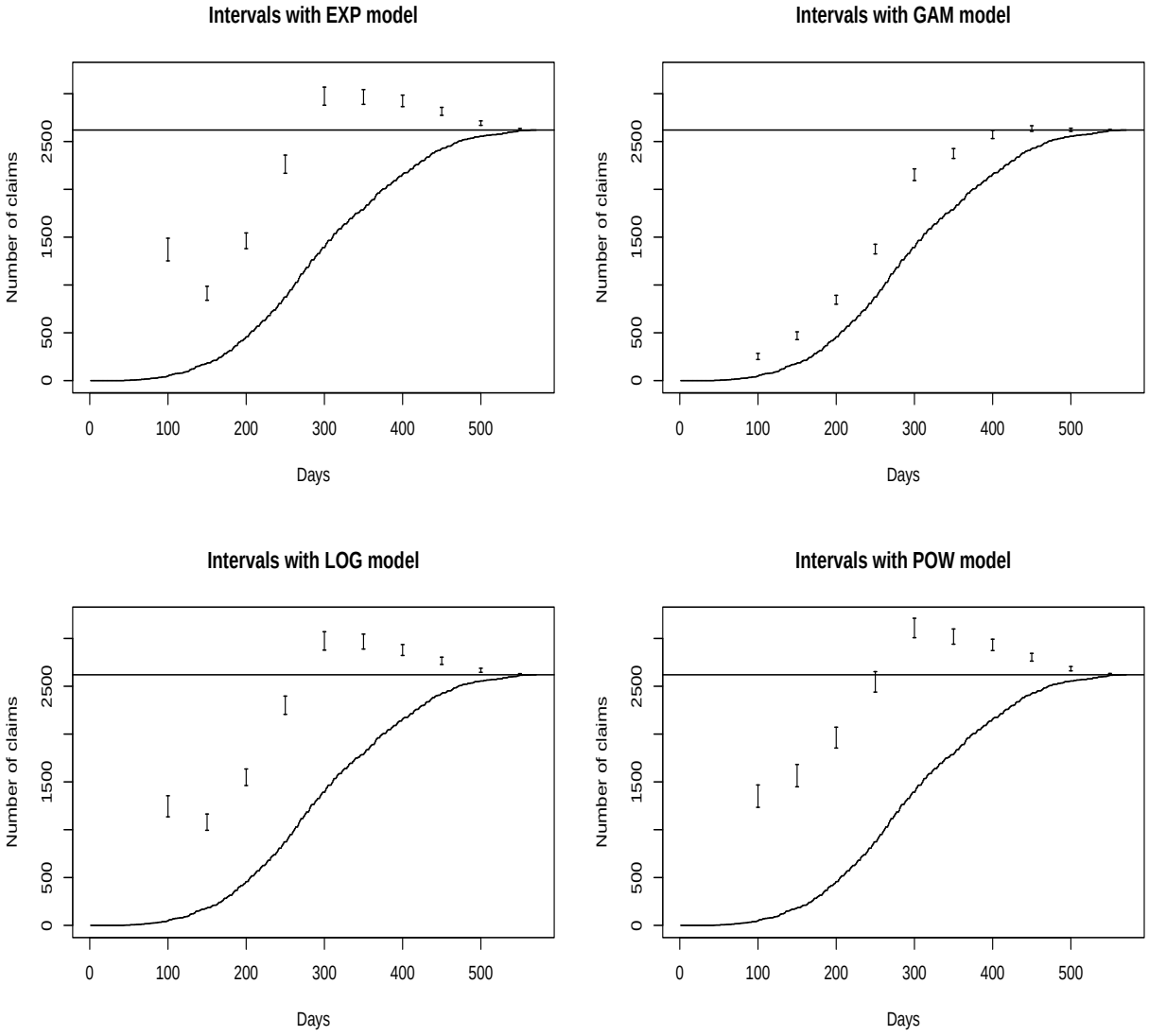

Figure 5.9: Non-calibrated 95% prediction intervals with different $f(t;\beta)$

claim times τ_{ij} can take a considerable amount of time. Each claim time is simulated by generating an uniform number on $[0, 1]$ and then finding the τ such that

$$u = \frac{F(\tau; \beta)}{F(t - t_i^s; \beta)}.$$

In our case, the function $F(\tau; \beta)$ cannot be inverted analytically and so all the τ_{ij}'s have to be found numerically.

We will now show how these times can be simulated more rapidly if we simulate discrete times instead of continuous ones. First, let us split the interval $[0, t-t_i^s]$ in M short intervals $[t_{l-1}, t_l]$ where $t_0 = 0$ and $t_M = t - t_i^s$. Given that a (simulated) claim occurred on $[0, t-t_i^s]$, we can show that

$$P[t_{l-1} \leq \tau_{ij} \leq t_l] = \frac{F(t_l; \hat{\beta}) - F(t_{l-1}; \hat{\beta})}{F(t - t_i^s; \hat{\beta})}. \tag{5.12}$$

Therefore, by simulating a discrete integer between 1 and M with probabilities given by (5.12), the associated claim time would be simulated from a distribution close to the real one when M is large enough. This procedure is significantly faster than the usual one since no numerical methods have to be used. In addition, the simulated claims will then look more like the observed ones since these claims were reported in days.

Using this method, we simulated $B = 1000$ datasets at times $t = 100, 150, \ldots, 550$ to calibrate the 95% plug-in prediction intervals first presented in the upper panel of Figure 5.7. These new intervals are presented in Figure 5.10. We can see that these intervals are now predicting well halfway through the process but they do not predict well early in the process. Considering that our model seems adequate and that we saw in the lower panel of Figure 5.7 that early prediction intervals using the final m.l.e.'s were good, we believe that early prediction intervals are inadequate because the calibration curve $G(u; a, b, c, \beta)$ (see Section 2.3) cannot be approximated closely by $\tilde{G}(u) = G(u; \hat{a}, \hat{b}, \hat{c}, \hat{\beta})$

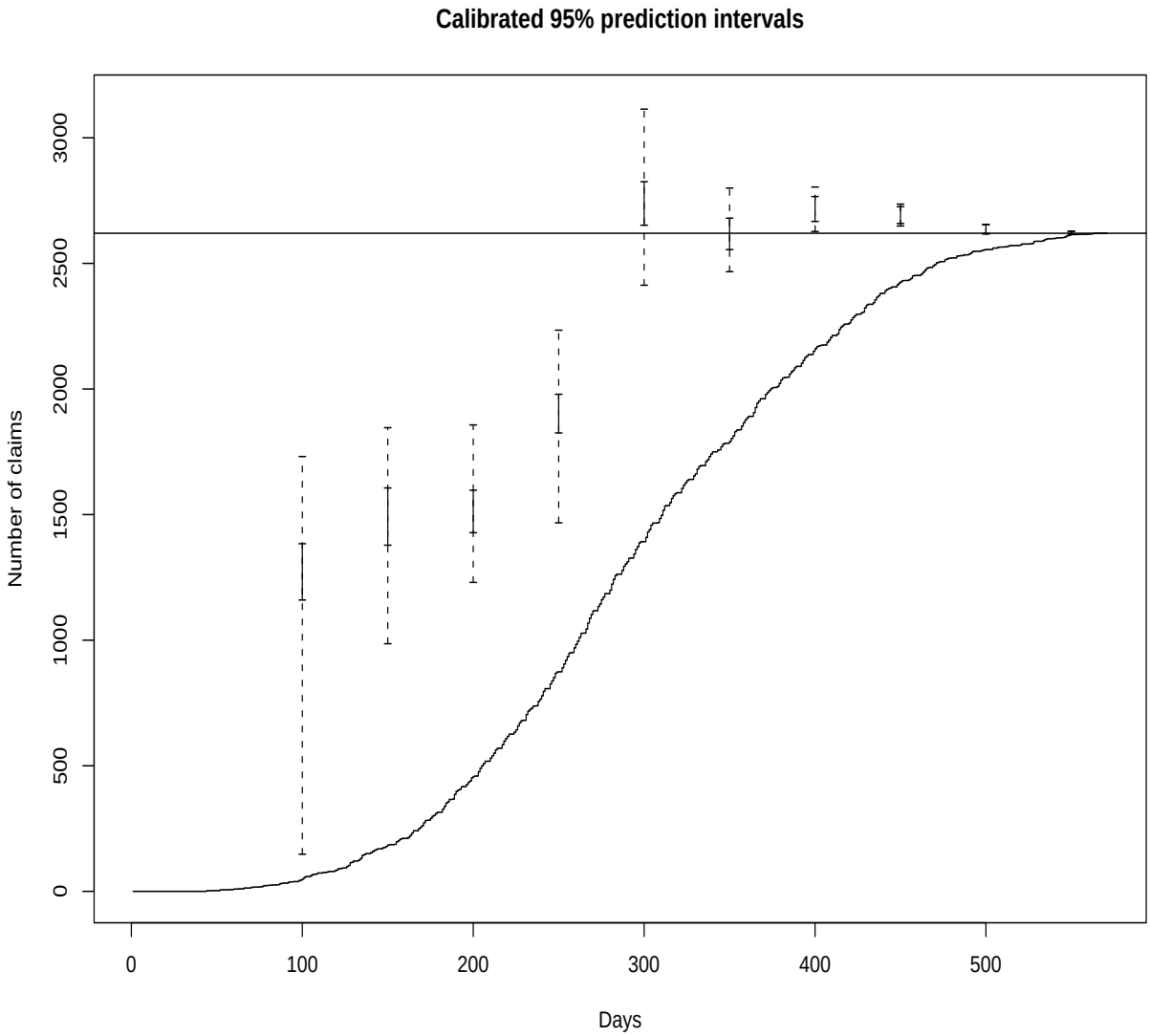

Figure 5.10: Calibrated 95% prediction intervals.

t	Coverage prob.
100	15.3%
150	34.6%
200	32.6%
250	25.6%
300	38.3%
350	53.6%
400	64.2%
450	75.8%
500	86.2%
550	96.0%

Table 5.7: Approximated coverage probability of 95% plug-in intervals.

early on or because $F(t;\beta)$ cannot extrapolate adequately. It is likely that adequate early predictions can only be achieved when some prior information is available.

The calibration curves calculated at different points in the process are presented in Figure 5.11. We can see that these curves are converging towards the c.d.f. of an uniform distribution but we can also see that unless we are very close to the end of the process, it is important to calibrate our intervals, especially for the lower quantiles. The importance of calibrating the intervals is also shown in Table 5.7 which shows $\tilde{G}(.975) - \tilde{G}(.025)$, the approximated coverage probability of an equal-tailed 95% plug-in prediction intervals. We can see that unless t is very close to 571, the non-calibrated prediction intervals do not appear to have an acceptable coverage probability.

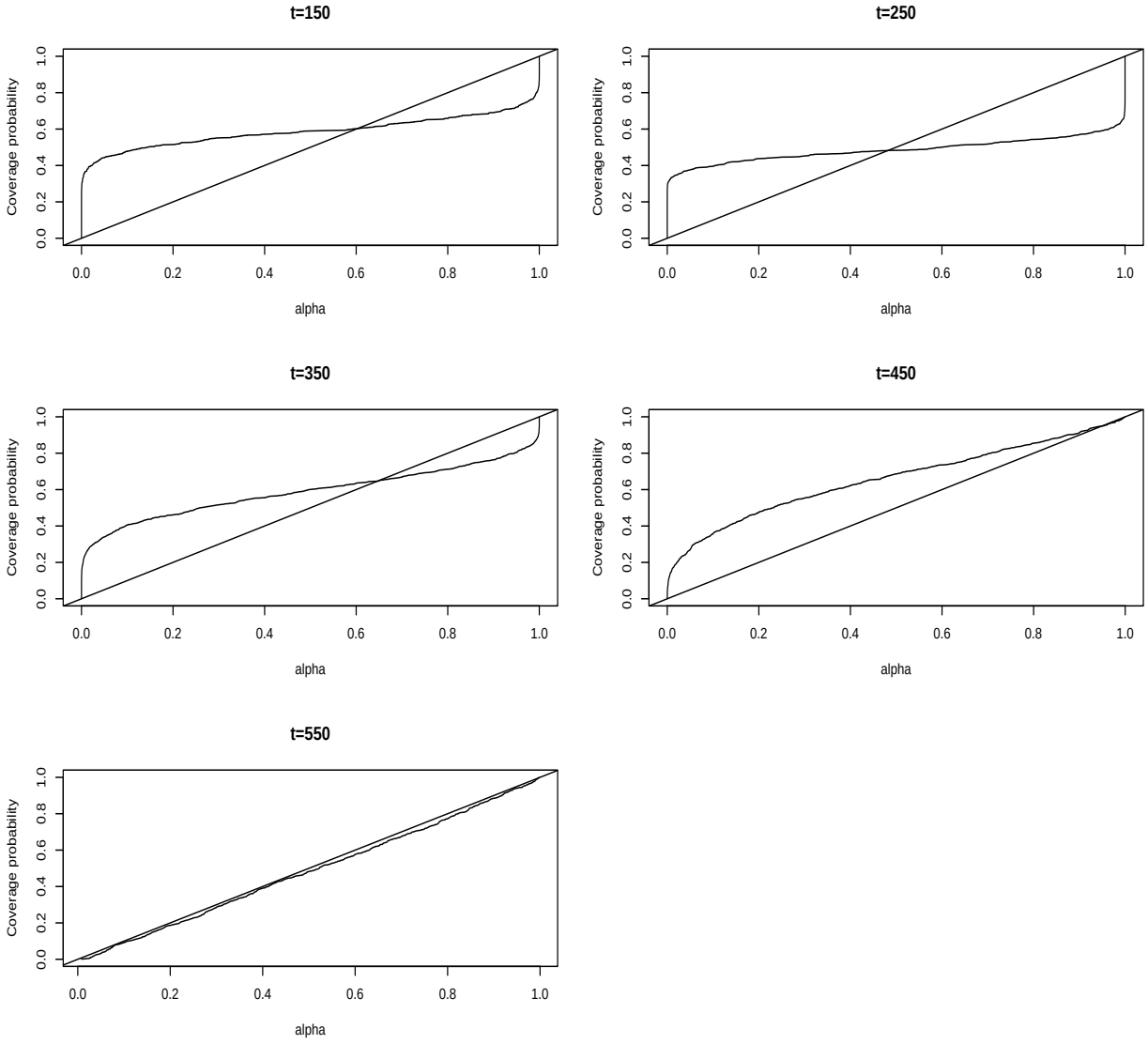

Figure 5.11: Calibration curves.

5.5 Calibration using asymptotic normality

We believe that it is clear from Table 5.7 and Figure 5.10 that plug-in prediction intervals should be calibrated. However, the amount of computational time required to calibrate these intervals is non negligible: using a 1395 Mhz processor, it took approximately 10 hours to calibrate the early intervals and approximately 25 hours for the following ones. The computational time required increases with t because more processes are incorporated in the likelihood function as time goes by, and the maximization of the B likelihood functions was by far what took most of the time to perform our algorithms. We will now propose a method that will provide us prediction intervals which are similar to the calibrated intervals, when t is large enough, but do not require the maximization of likelihood functions.

Let $\theta = (a, b, c, \beta)$ so that a calibrated prediction interval is obtained when we approximate the distribution of

$$\mathbf{U} \;=\; F(\sum_{i=1}^{k} \mathbf{N}_i(365)|(\mathbf{N}(0), \mathbf{N}(t^s)); \hat{\theta}(\mathbf{N}(0), \mathbf{N}(t^s), \tau(t))),$$

assuming that $\theta = \hat{\theta}(N(0), N(t^s), \tau(t))$. Since m.l.e.'s have an asymptotic normal distribution, the distribution of $\mathbf{U}$ is asymptotically equivalent to the distribution of

$$\widehat{\mathbf{U}} \;=\; F(\sum_{i=1}^{k} \mathbf{N}_i(365)|(\mathbf{N}(0), \mathbf{N}(t^s)); \hat{\theta}_N),$$

where $\hat{\theta}_N \sim \mathcal{N}(\theta, \mathcal{I}^{-1}(\theta))$ and $\mathcal{I}(\theta)$ is Fisher's information matrix. Therefore, calibrating the intervals by approximating the distribution of $\widehat{\mathbf{U}}$ instead of $\mathbf{U}$ should give us similar intervals when t is large enough. However, these new intervals will be obtained quite rapidly since we will not have to simulate the τ_{ij}'s and no likelihood functions have to be maximized.

At times $t = 100, 150, \ldots, 550$ we generated samples of B values from $\widehat{\mathbf{U}}$ and compared them with samples of $\mathbf{U}$ in Figure 5.12. We also obtained these new calibrated intervals

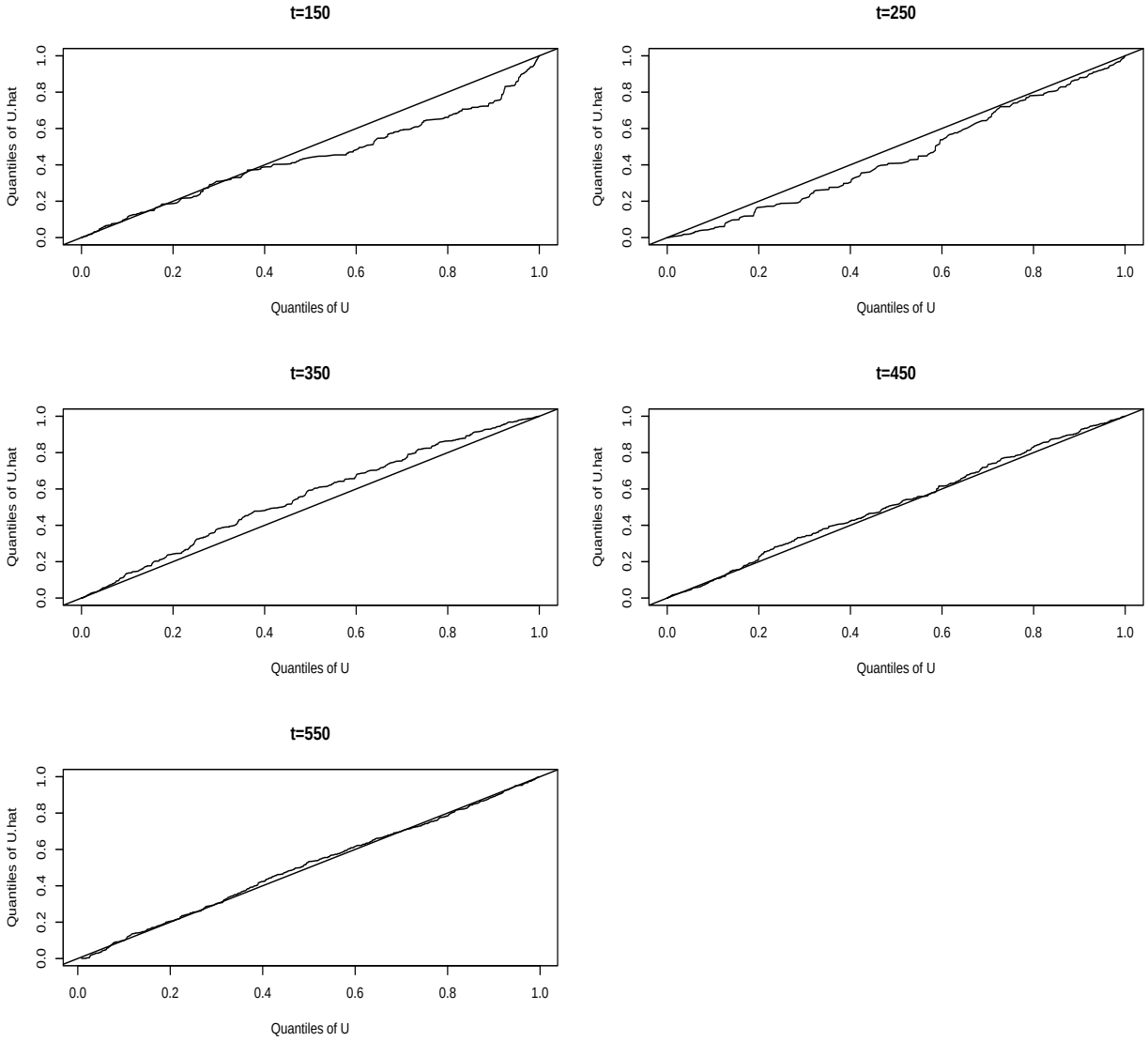

Figure 5.12: Quantile-quantile plots between $\widehat{\mathbf{U}}$ and $\mathbf{U}$.

and compared them with those based on **U** in Figure 5.13. As expected, both distributions are similar when t is large enough and so are their prediction intervals, especially the upper ends. It is interesting to point out that this approach starts to give similar intervals at the moment where the usual calibration approach starts to require a considerable amount of computational time.

For finite horizon problems like this one, the use of this new calibration approach allows us to recommend the following strategy to calculate prediction intervals: at first we obtain the three types of intervals (simple plug-in, calibrated plug-in, and calibrated plug-in using normal m.l.e.'s). When the two calibration approaches start to provide similar intervals, we can cease the calculation of the longest approach. Finally, when the simple plug-in intervals start to look like the approximated calibrated intervals, we can cease to calibrate them.

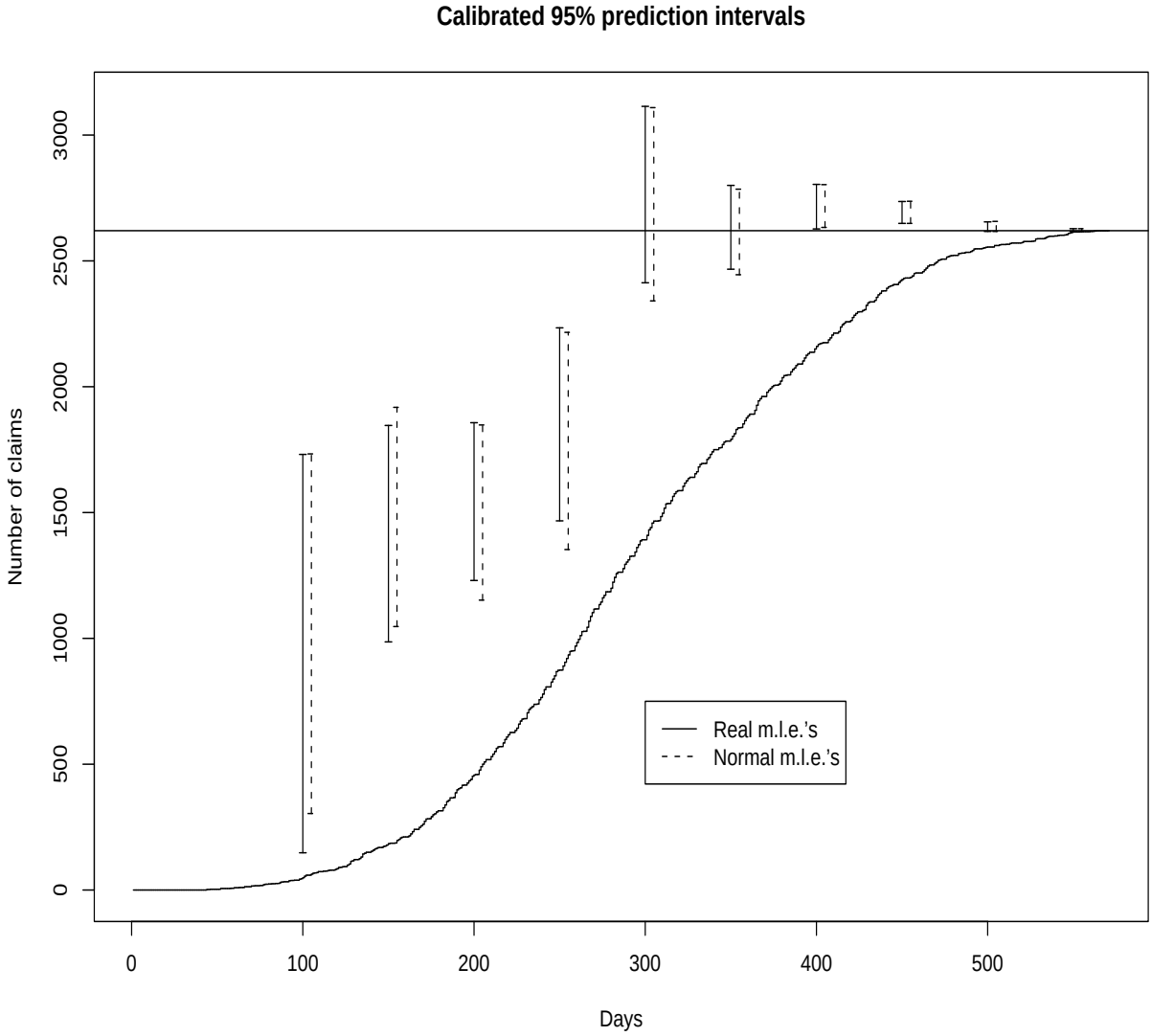

Figure 5.13: Comparison of calibration methods.

Chapter 6

Future Research and Other Topics

Interesting features about the prediction of recurrent events were revealed in Chapter 3, 4, and 5. Some of these features are related to specific issues arising from the use of plug-in methods with random effects models, such as their robustness, their sensitivity, and the determination of appropriate estimation procedures. In this chapter, we will explain how these features could be investigated further and discuss possible model extensions for NHPP's. Other related prediction topics that could be interesting to study will also be presented.

6.1 Robustness and sensitivity

We already discussed that it is relatively straightforward to find predictors and prediction intervals when models with random effects are used. These models are also very flexible: we can do predictions for unobserved processes by letting $t_{1i} = 0$, a predictive distribution is available even when $N_i(t_{1i}) = 0$, we can model different groups of processes with different parameters, and we can easily add covariates into the models when some are available.

However, one can argue that by adding a distribution on unobservable rates, models with random effects are highly parametric. It is then important to study the robustness of these models to various types of misspecifications. For example, this was done in (3.11) where we explained why predictors obtained from models with random effects are robust to the rate homogeneity/heterogeneity assumption. In addition, we saw that approaches using random effects are robust to the real distribution of these effects. Therefore, when the real random effects are not gamma, it would also be interesting to show if a plug-in prediction interval assuming gamma random effects has, when the total number of events observed is large, an approximative coverage probability of $1 - \alpha$.

In Section 3.4 and 4.3, we also saw that approaches assuming random effects can perform well when the real rates are fixed. Especially, we saw in Table 3.7 that this extra variability added can be just appropriate to have coverage proportions close to the desired level. Therefore, we would like to investigate when random effects approaches are appropriate to compensate for the uncertainty about the fixed rates. The work done in Datta et al. (2000) could be useful here. This paper deals with frequentist validity of Bayesian prediction methods and when the rates are fixed a plug-in random effects approach is actually an empirical Bayes method.

As for the calibration of prediction intervals, we saw that even if it usually provides better intervals by taking into account the uncertainty about the unknown parameters, it is still using estimated parameters to simulate the processes. Therefore, it would be interesting to study how sensitive this approach is.

6.2 Model extensions

Experts' knowledge or any other types of prior information were never considered in our prediction models. However, we can think of many scenarios where such information could be available. For example, warranty datasets collected in the previous years or experts' knowledge about the behavior of $f(t;\beta)$ can be available when we wish to predict automobile warranty claims. Therefore, we would like to develop some Bayesian prediction models for this type of problem. When the prior information is adequate, a Bayesian approach would be especially useful to improve our poor early predictions.

Experts' knowledge regarding the mileage drop-out could also be useful. For example, it could allow us to model this more explicitly by using

$$f(t;\beta) = f^*(t;\beta)p(t),$$

where $p(t)$ is a non-increasing function such that $p(0) = 1$ and $p(365) = 0$ which indicates the approximated proportion of vehicles believed to be under warranty after t days.

In addition to the Bayesian extensions, we would also like to extend our prediction models to take some other features into account. For example, models with covariates or models where the β's are also treated as random effects. Models using splines for $f(t;\beta)$ will also be considered.

Finally, little research has been done on goodness-of-fit tests when more than one NHPP are considered. When we assessed the adequacy of our model in Section 5.3.2, we did it by assessing the fit of the total counts and the occurrence times separately. We would like to find some goodness-of-fit tests able to asses the adequacy of both features simultaneously.

6.3 Other topics

Many other types of data are considered in prediction problems and we would like to study some of them. First, the work we did so far can easily be modified to deal with some cost prediction problems. Sometimes, costs arise over time from some random events and it is of interest to predict the total cost over a certain time interval. The prediction of the total cost of making repairs for cars under warranty is a good example. We can try to predict such a total by modeling the events with Poisson processes and the cost per event with a certain distribution, say $G(c)$. It should then be relatively straightforward to derive predictors and prediction intervals for the total cost.

We would also like to study prediction problems for multi-state models. Suppose, for example, that we observe a fixed number of individuals over time as they are moving from a state to another. It may be of interest to determine prediction regions for the number of individuals in each state in a future time. Such problems can become very challenging because of certain characteristics inherent to the datasets. Amongst many others, we can encounter problems where the processes have certain forms of spatial dependencies or where the data are aggregated (Kalbfleisch, Lawless & Vollmer 1983).

Bibliography

Aitchison, J. (1975), 'Goodness of prediction fit', *Biometrika* **62**, 547–554.

Aitchison, J. & Dunsmore, I. (1975), *Statistical Prediction Analysis*, Cambridge University Press.

Baker, G. (1935), 'The probability that the mean of a second sample will differ from the mean of a first sample by less than a certain multiple of the standard deviation of the first sample', *Annals of Mathematical Statistics* **6**, 197–201.

Barndorff-Nielsen, O. & Cox, D. (1996), 'Prediction and asymptotics', *Bernoulli* **2**, 319–340.

Bender, E. (1974), 'Asymptotic methods of enumaration', *SIAM Review* **16**, 485–515.

Beran, R. (1990), 'Calibrating prediction regions', *Journal of the American Statistical Association* **85**, 715–723.

Campodónico, S. & Singpurwalla, N. D. (1995), 'Inference and predictions from Poisson point processes incorporating expert knowledge', *Journal of the American Statistical Association* **90**, 220–226.

Carlin, B. P. & Louis, T. A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall.

Cox, D. (1975), Prediction intervals and empirical bayes confidence intervals, *in* J. Gani, ed., 'Perspectives in Probability and Statistics', Applied Probability Trust, pp. 47–56.

Cox, D. & Lewis, P. (1966), *The Statistical Analysis of Series of Events*, London: Methuen.

Datta, S. G., Mukerjee, R., Ghosh, M. & Sweeting, T. (2000), Bayesian prediction with approximate frequentist validity, Technical report, University of Surrey.

de Finetti, B. (1937), 'La prévision: ses lois logiques, ses sources subjectives', *Annales de l'Institut Henri Poincaré* **7**, 1–68.

Dempster, A., Laird, N. & Rubin, D. (1977), 'Maximum likelihood estimation from incomplete data via the em algorithm', *Journal of the Royal Statistical Association, Series B* **39**, 1–38.

Faulkenberry, D. G. (1973), 'A method of obtaining prediction intervals', *Journal of the American Statistical Association* **68**(342), 433–435.

Gail, M., Santner, T. & Brown, C. (1980), 'An analysis of comparative carcinogenesis experiments based on multiple times to tumor', *Biometrics* **36**, 255–266.

Gaver, D. P. & O'Muircheartaigh, I. G. (1987), 'Robust empirical bayes analyses of event rates', *Technometrics* **29**, 1–15.

Geisser, S. (1993), *Predictive Inference: An Introduction*, Chapman & Hall.

Goel, A. & Okumoto, K. (1979), 'Time-dependent error-detection rate model for software reliability and other performance measures', *IEEE Transactions on Reliability* **28**, 206–211.

Grandell, J. (1997), *Mixed Poisson Processes*, Chapman & Hall.

Harris, I. (1989), 'Predictive fit for natural exponential families', *Biometrika* **76**, 675–684.

Herzog, T. (1999), *Introduction to Credibility Theory*, 3rd edn, ACTEX Publications.

Hinkley, D. (1979), 'Predictive likelihood', *The Annals of Statistics* **7**, 718–728.

Jeffreys, H. (1961), *Theory of Probability*, 3rd edn, Oxford: University Press.

Kalbfleisch, J. D., Lawless, J. F. & Robinson, J. (1991), 'Methods for the analysis and prediction of warranty claims', *Technometrics* **33**, 273–285.

Kalbfleisch, J. D., Lawless, J. F. & Vollmer, W. (1983), 'Estimation in Markov models from aggregate data', *Biometrics* **39**, 907–918.

Kalbfleisch, J. D. & Sprott, D. (1970), 'Application of likelihood methods to models involving large numbers of parameters', *Journal of the Royal Statistical Association, Series B* **32**, 175–194.

Kendall, M. & Allen, S. (1977), *The Advanced Theory of Statistics, Inference and Relationship*, Vol. 2, 4th edn, Charles Griffin and company.

Klugman, S., Panger, H. & Wilmot, G. (2004), *Loss Models: From Data to Decisions*, 2nd edn, Wiley.

Komaki, F. (1996), 'On asymptotic properties of predictive distributions', *Biometrika* **83**, 299–313.

Kullback, S. & Leibler, R. (1951), 'On information and sufficiency', *Annals of Mathematical Statistics* **22**, 79–86.

Kuo, L. & Yang, T. Y. (1996), 'Bayesian computation for nonhomogeneous Poisson processes in software reliability', *Journal of the American Statistical Association* **91**(434), 763–773.

Lauritzen, S. (1974), 'Sufficiency, prediction and extreme models', *Scandinavian Journal of Statistics* **2**, 128–134.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, Wiley.

Lawless, J. F. (2003), *Statistical Models and Methods for Lifetime Data*, 2nd edn, Wiley.

Lawless, J. & Fredette, M. (2004), Frequentist prediction intervals and predictive distributions, submitted to *Biometrika*.

Lawless, J. & Nadeau, J. (1995), 'Some simple robust methods for the analysis of recurrent events', *Technometrics* **37**, 158–168.

Lejeune, M. (1975), A maximum likelihood approach to prediction with applications to binomial and Poisson populations, unpublished PhD dissertation, Department of Statistics, Oregon State University.

Lejeune, M. & Faulkenberry, D. G. (1982), 'A simple predictive density function', *Journal of the American Statistical Association* **77**, 654–657.

Martz, H. F., Parker, R. L. & Rasmuron, D. M. (1999), 'Estimation of trends in the scram rate at nuclear power plants', *Technometrics* **41**, 352–364.

Meeker, W. Q. & Escobar, L. A. (1998), *Statistical Methods for Reliability Data*, Wiley.

Murray, G. (1977), 'A note on the estimation of probability densitiy functions', *Biometrika* **64**, 150–152.

Nelson, W. (1969), Confidence intervals for the ratio of two Poisson means and predictor intervals for a Poisson random variable, Technical Report 69-C-118, General Electric Research and Development Center.

Nelson, W. (1982), *Applied Life Data Analysis*, Wiley.

Ng, V. (1980), 'On the estimation of parametric density functions', *Biometrika* **67**, 505–506.

Ngai, H. M. & Stroud, T. (1994), 'Hierarchical bayes simultaneous estimation of Poisson means', *Communications in Statistics -Theory and Methods* **23**(10), 2965–2991.

Patel, J. & Samaranayake, V. (1991), 'Prediction intervals for som discrete distributions', *Journal of Quality Technology* **23**, 270–278.

Pearson, K. (1920), 'The fundamental problem of practical statistics', *Biometrika* **13**, 1–16.

Raftery, A. (1987), 'Inference and prediction for a general order statistic model with unknown population size', *Journal of the American Statistical Association* **82**(400), 1163–1168.

Robbins, H. (1955), An empirical bayes approach to statistics, *in* U. of California Press, ed., '3rd Berkeley Symposium on Mathematical Statistics and Probability', pp. 157–164.

Scheffe, H. & Tukey, J. (1945), 'Non-parametric estimation. I. validation of order statistics', *Annals of Mathematical Statistics* **16**, 187–192.

Smith, R. (1999), Bayesian and frequentist approaches to parametric predictive inference, *in* A. D. J.M. Bernardo, J.O. Berger & A. S. (Eds.), eds, 'Bayesian Statistics 6', Oxford University Press, pp. 589–612.

Snyder, D. & Miller, M. (1991), *Random Point Processes in Time and Space*, Springer-Verlag.

Vidoni, P. (1995), 'A note on modified estimative prediction limits and distributions', *Biometrika* **85**, 949–953.

Vit, P. (1973), 'Interval prediction for a Poisson process', *Biometrika* **60**, 667–668.

Wald, A. (1942), 'Setting of tolerance limits when the sample is large', *Annals of Mathematical Statistics* **13**, 389–399.

Wilks, S. (1941), 'Determination of sample sizes for setting tolerance limits', *Annals of Mathematical Statistics* **12**, 91–96.

Wilks, S. (1942), 'Statistical prediction with special reference to the problem of tolerance limits', *Annals of Mathematical Statistics* **13**, 400–409.