**UNIVERSIDADE FEDERAL DE MINAS GERAIS**

**Instituto de Ciências Exatas**

**Programa de Pós-Graduação em Estatística**

**Frederico Machado Almeida**

# Solutions to the Monotone Likelihood in the Standard Mixture Cure Fraction Model

Belo Horizonte, MG - Brasil

2021

**Frederico Machado Almeida**

# Solutions to the Monotone Likelihood in the Standard Mixture Cure Fraction Model

Thesis submitted to the Departamento de Estatística of the Universidade Federal de Minas Gerais as a partial fulfillment of the requirements for the degree of Doctor in Statistics.

**Advisor:** Prof. Dr. Enrico Antônio Colosimo
**Co-Advisor:** Prof. Dr. Vinícius Diniz Mayrink

Belo Horizonte, MG - Brasil
2021

.

Almeida, Frederico Machado.

A447s Solutions to the monotone likelihood in the standard mixture cure fraction model [manuscrito] / Frederico Machado Almeida. – 2021.
xiii, 160 f. il.

Orientador: Enrico Antônio Colosimo.
Coorientador: Vinícius Diniz Mayrink.
Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f. 148–160.

1. Estatística – Teses. 2. Algoritmo EM – Teses. 3. Estatística matemática – Teses. 4. Inferência (Estatística) – Teses.I. Colosimo, Enrico Antônio. II. Mayrink, Vinícius Diniz. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. IV.Título.

CDU 519.2 (043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6ª Região nº 1510

**UNIVERSIDADE FEDERAL DE MINAS GERAIS**

**PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA**

ATA DA DEFESA DE TESE DE DOUTORADO DO ALUNO **FREDERICO MACHADO ALMEIDA**, MATRICULADO, SOB O Nº 2017667140, NO PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA, DO INSTITUTO DE CIÊNCIAS EXATAS, DA UNIVERSIDADE FEDERAL DE MINAS GERAIS, REALIZADA NO DIA 01 DE JULHO DE 2021

Aos 01 dia do mês de Julho de 2021, às 14h00, em reunião pública virtual 68 (conforme orientações para a atividade de defesa de tese durante a vigência da Portaria PRPG nº 1819) na sala https://us02web.zoom.us/j/89465316725?pwd=emJiOU5mbTRUb3IwbVY1WmR5SEd-Ndz09 do Instituto de Ciências Exatas da UFMG, reuniram-se os professores abaixo relacionados, formando a Comissão Examinadora homologada pelo Colegiado do Programa de Pós-Graduação em Estatística, para julgar a defesa de tese do aluno FREDERICO MACHADO ALMEIDA, nº matrícula 2017671140, intitulada: *"Solutions to the Monotone Likelihood in the Standard Mixture Cure Fraction Model"*, requisito final para obtenção do Grau de doutor em Estatística. Abrindo a sessão, o Senhor Presidente da Comissão, Prof. Enrico Antônio Colosimo, passou a palavra ao aluno para apresentação de seu trabalho. Seguiu-se a arguição pelos examinadores com a respectiva defesa do aluno. Após a defesa, os membros da banca examinadora reuniram-se reservadamente sem a presença do aluno e do público, para julgamento e expedição do resultado final. Foi atribuída a seguinte indicação:

(X) Aprovada.
( ) Reprovada com resubmissão do texto em ____ dias.
( ) Reprovada com resubmissão do texto e nova defesa em ____ dias.
( ) Reprovada.

| | |
|---|---|
| Prof. Enrico Antonio Colosimo-Orientador (EST/UFMG) | Prof. Vinícius Diniz Mayrink - Co-orientador EST/UFMG |
| Profa. Vera Lúcia Damasceno Tomazella (EST/USFCar)) | Prof. Mário de Castro Andrade Filho (ICMC/USP) |
| Prof. Fábio Nogueira Demarqui (EST/UFMG) | Prof. Wagner Barreto de Souza Pesquisador da KAUST-Arábia Saudita |

O resultado final foi comunicado publicamente ao aluno pelo Senhor Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente Ata, que será assinada por todos os membros participantes da banca examinadora. Belo Horizonte, 01 de julho de 2021.

Observações:

1. No caso de aprovação da tese, a banca pode solicitar modificações a serem feitas na versão final do texto. Neste caso, o texto final deve ser aprovado pelo orientador da tese. O pedido de expedição do diploma do candidato fica condicionado à submissão e aprovação, pelo orientador, da versão final do texto.
2. No caso de reprovação da tese com resubmissão do texto, o candidato deve submeter o novo texto dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca que então decidirão pela aprovação ou reprovação da tese.
3. No caso de reprovação da tese com resubmissão do texto e nova defesa, o candidato deve submeter o novo texto com a antecedência à nova defesa que o orientador julgar adequada. A nova defesa, mediante todos os membros da banca, deve ser realizada dentro do prazo estipulado pela banca, que deve ser de no máximo 6 (seis) meses. O novo texto deve ser avaliado por todos os membros da banca. Baseada no novo texto e na nova defesa, a banca decidirá pela aprovação ou reprovação da tese.

To my father Machado Almeida
(in memorial)

# Acknowledgements

Em primeiro lugar agradeço a Deus pelo dom de vida e, por me fortalecer dia após dia, durante toda a caminhada.

À minha família, em especial, aos meus pais, Machado Almeida (em memoria) e Matilde Marcelino pelos ensinamentos e valores morais que sempre implantaram em mim.

Aos meus orientadores, prof. Enrico A. Colosimo e prof. Vinícius D. Mayrink pela paciência e ensinamentos que de forma incansável vêm transmitindo durante todo esse tempo.

Agradeço a todos os professores do Departamento de Estatística do ICEx/UFMG pelos conhecimentos transmitidos durante os seis anos que estive vinculado ao programa. Ao professor Cristiano de Carvalho Santos, vai um especial agradecimento pelo excelente curso de Métodos Computacionais que foi fundamental no desenvolvimento desta tese.

A todos os professores do Departamento de Matemática e Informática da Universidade Eduardo Mondlane/Moçambique, em especial ao prof. Osvaldo André Loquiha, pelos ensinamentos, suporte e encorajamento.

Agradeço a todos os meus colegas e amigos, em especial ao Arthur, Guilherme(s), Thiago, Edson Raso, Erick, Rumenick e Jussiane, pelos momentos de convívio e aprendizado.

À Rogéria, secretaria do programa de pós-graduação em Estatística, vai um especial agradecimento pela ajuda incondicional que me proporcionou.

Por fim, agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro, sem o qual eu não teria condições para cursar o doutorado.

# Resumo

Modelos de sobrevivência para dados com fração de curados, são frequentes em pesquisas biomédicas. Em situações envolvendo eventos raros, onde é comum obter amostras pequenas com muitos tempos de censura, o processo de estimação dos coeficientes de regressão pode ser problemático, uma vez que algumas estimativas podem não assumir valores finitos. Este fenômeno é conhecido na literatura como o problema da Verosimilhança Monótona (VM), ocorrendo na presença de covariáveis categóricas fortemente desbalanceadas. A solução mais conhecida, é uma adaptação do método de Firth originalmente proposto para reduzir o viés dos estimadores de máxima verossimilhança. O método garante a obtenção de estimativas finitas a partir da penalização da função de verossimilhança, na qual o termo de penalidade pode ser interpretado como sendo a distribuição a priori invariante de Jeffreys, frequentemente usada em inferência Bayesiana. Estudos investigando a VM nos modelos de sobrevivência com fração de curados são escassos. Para solucionar o problema, nossa primeira proposta consiste em derivar a função escore modificada baseando-se no método de Firth. Nossa segunda contribuição consiste em investigar outras funções de penalidade (ou distribuições a priori) baseadas no enfoque Bayesiano. Um estudo de simulação Monte Carlo foi conduzido e indicou um bom desempenho em termos de inferência, especialmente para o caso Bayesiano. Uma análise foi conduzida para um conjunto de dados reais envolvendo pacientes com melanoma, atendidos no Hospital das Clínicas/UFMG. Esse conjunto de dados é relativamente novo e apresenta simultaneamente o problema da VM e fração de indivíduos curados.

Palavras-chaves: Modelos de fração de cura, Algoritmo EM, método de Firth, Inferência Bayesiana, Ligação logística, Melanoma.

# Abstract

Survival models for situations where some individuals are long-term survivors, immune or non-susceptible to the event of interest are extensively studied in biomedical research. Fitting a regression can be problematic in situations involving small sample sizes with many censored times, since the maximum likelihood estimates of some coefficients may be infinity. This phenomenon is commonly known as Monotone Likelihood (ML), occurring in the presence of many categorical and unbalanced covariates. A well-known solution is an adaptation of the Firth's method, originally created to reduce the maximum likelihood estimation bias. The method ensures finite estimates by penalizing the likelihood function, where the penalty term might be interpreted as the Jeffreys invariant prior, largely used in the Bayesian framework. The ML issue in the context involving mixture cure models is a topic rarely discussed in the literature, and it configures a central contribution of this work. In order to handle this point in such context, we propose to derive the adjusted score function based on the Firth method. The second major contribution is to investigate other flexible penalty functions (prior distributions), in which all inference procedures will be based on the posterior samples. An extensive Monte Carlo simulation study indicates good inference performance for the penalized estimates, especially in the Bayesian framework. The analysis is illustrated through a real application involving patients with melanoma assisted at the Hospital das Clínicas/UFMG. This is a relatively novel data set affected by the monotone likelihood issue and containing cured individuals.

Keywords: Cure rate models, EM algorithm, Firth method, Bayesian inference, Logistic link function, Melanoma.

# List of Abbreviations

| | |
|---|---|
| ACF | Autocorrelation Functions |
| AFT | Accelerated Failure Time |
| AFTMC | Accelerated Failure Time Mixture Cure |
| CAPES | Aperfeiçoamento de Pessoal de Nível Superior |
| CI | Confidence Interval |
| CP | Coverage Probability |
| CPU | Central Processing Unit |
| EM | Expectation and Maximization |
| $EM_{cs}(\%)$ | EM convergence samples proportion |
| FC | Firth Correction |
| HPD | Highest Posterior Density |
| ICEx | Instituto de Ciências Exatas |
| KM | Kaplan-Meier |
| MC | Monte Carlo |
| MCMC | Markov Chain Monte Carlo |
| MCSE | Monte Carlo Standard Error |
| M-H | Metropolis-Hastings |
| ML | Monotone Likelihood |
| MLE | Maximum Likelihood Estimate |

| | |
|---|---|
| NR | Newton-Raphson |
| PHM | Proportional Hazard Model |
| PTM | Promotion Times Model |
| Q-Q | Quantile-Quantile |
| RMSE | Mean Square Error |
| RB | Relative Bias |
| SE | Standard Error |
| SP | Separation Problem |
| SMCM | Standard Mixture Cure Model |
| SPMC | Semiparametric Mixture Cure |
| UFMG | Universidade Federal de Minas Gerais |
| VM | Verossimilhança Monótona |

Lifetime Data Analysis
https://doi.org/10.1007/s10985-020-09510-4

# Firth adjusted score function for monotone likelihood in the mixture cure fraction model

**Frederico Machado Almeida**[1] · **Enrico Antônio Colosimo**[1] · **Vinícius Diniz Mayrink**[1]

Received: 28 January 2020 / Accepted: 30 October 2020


## Abstract

Models for situations where some individuals are long-term survivors, immune or non-susceptible to the event of interest, are extensively studied in biomedical research. Fitting a regression can be problematic in situations involving small sample sizes with high censoring rate, since the maximum likelihood estimates of some coefficients may be infinity. This phenomenon is called monotone likelihood, and it occurs in the presence of many categorical covariates, especially when one covariate level is not associated with any failure (in survival analysis) or when a categorical covariate perfectly predicts a binary response (in the logistic regression). A well known solution is an adaptation of the Firth method, originally created to reduce the estimation bias. The method provides a finite estimate by penalizing the likelihood function. Bias correction in the mixture cure model is a topic rarely discussed in the literature and it configures a central contribution of this work. In order to handle this point in such context, we propose to derive the adjusted score function based on the Firth method. An extensive Monte Carlo simulation study indicates good inference performance for the penalized maximum likelihood estimates. The analysis is illustrated through a real application involving patients with melanoma assisted at the Hospital das Clínicas/UFMG in Brazil. This is a relatively novel data set affected by the monotone likelihood issue and containing cured individuals.

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s10985-020-09510-4) contains supplementary material, which is available to authorized users.

✉ Vinícius Diniz Mayrink
vdm@est.ufmg.br

Frederico Machado Almeida
falmeida856@gmail.com

Enrico Antônio Colosimo
enricoc@est.ufmg.br

[1] Departamento de Estatística, ICEx, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Belo Horizonte, MG 31270-901, Brazil

Published online: 13 November 2020

Biometrical Journal **00** (2020) 00, zzz–zzz / DOI: 10.1002/bimj.200100000

# Modified score function for monotone likelihood in the semiparametric mixture cure model

**Frederico M. Almeida** [1], **Enrico A. Colosimo** [1], and **Vinícius D. Mayrink** [*,1]

[1] Departamento de Estatística, ICEx, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, 31270-901, Av. Antônio Carlos, 6627, Brazil.

Received 00.00, revised 00.00, accepted 00.00

The cure fraction models are intended to analyze lifetime data from populations where some individuals are immune to the event under study, and allow a joint estimation of the distribution related to the cured and susceptible subjects, as opposed to the usual approach ignoring the cure rate. In situations involving small sample sizes with many censored times, the detection of non-finite coefficients may arise via maximum likelihood. This phenomenon is commonly known as monotone likelihood (ML), occurring in the Cox and logistic regression models when many categorical and unbalanced covariates are present. An existing solution to prevent the issue is based on the Firth correction, originally developed to reduce the estimation bias. The method ensures finite estimates by penalizing the likelihood function. In the context of mixture cure models, the ML issue is rarely discussed in the literature, therefore, this topic can be seen as the first contribution of our paper. The second major contribution, not well addressed elsewhere, is the study of the ML issue in cure mixture modeling under the flexibility of a semiparametric framework to handle the baseline hazard. We derive the modified score function based on the Firth approach and explore finite sample size properties of the estimators via a Monte Carlo scheme. The simulation results indicate that, the performance of coefficients related to the binary covariates are strongly affected to the imbalance degree. A real illustration, in the melanoma dataset, is discussed using a relatively novel data set collected in a Brazilian university hospital.

*Key words:* Cox regression; Cure rate; EM algorithm; Firth method; Melanoma.

## 1 Introduction

The cure rate models are a particular class of survival models developed for situations where the target population is divided in two groups: susceptible and non-susceptible subjects (also called "cured" or "long-term survivors"). In short, susceptible individuals will eventually experience the main event at some time point and the non-susceptible ones will never indicate the event, even after a long follow-up period. Consequently, the survival time of a non-susceptible case is considered to be infinity. For such type of data, standard lifetime models are not appropriate, because they do not account for the presence of the cure rate. In clinical studies, for example, the cured subjects are the patients with a favorable response to a particular treatment.

Cure models were designed to handle simultaneously the estimation related to the distribution of the susceptible cases (latency part) and the one explaining the proportion of cured subjects (incidence part). In the literature of cure rate models, two frameworks are extensively studied: $(i)$ the standard mixture cure model (SMCM) introduced by Boag (1949) and later extended by Berkson and Gage (1952) and $(ii)$ the promotion times model (PTM) proposed by Yakovlev and Tsodikov (1996). The present paper is focused on the SMCM, which is perhaps the most popular approach to jointly deal with the cured proportion and the probability of being uncured, especially in biomedical studies; see for instance Wang *et al.* (2018), Naseri *et al.* (2018) and Zaeran *et al.* (2019). This model is based on the assumption that the population

*Corresponding author: e-mail: vdm@est.ufmg.br, Phone:0000-0000 , Fax:0000-0000

 

# Contents

# Chapter 1

# INTRODUCTION

The problem of analyzing survival data arises in many fields of knowledge, sometimes with different names, but it does not imply on any real difference in the techniques. For example, in medical research (*survival analysis:* the event can be the death, hart attack, recurrence of a disease, among others); in economics (*duration analysis:* we may be interested to study the time until an unemployed person finds a new job); in demography (time to divorce); in finance (bank goes to bankrupt); in sociology (to study the time until a former prisoner is rearrested); insurance (warranty claim); in criminology (time from the acquittal of an individual until he commits a crime again); insurance (warranty claim); in engineering (*reliability analysis:* the time until a machine breaks down). Many other examples ca be indicated.

The response of interest is the time until the occurrence of some event, also known as survival time, which specifies the length of the follow-up time. By survival time, we mean for instance: years, months, weeks or days from the beginning of the follow-up of an individual until the event under study. Likewise, by event under study, or simply failure, we mean: the death, disease incidence, relapse from remission, recovery or any designed event of interest that may happen to an individual; see (Klein and Moeschberger, 2006; Colosimo and Giolo, 2006; Allison, 2010) for more details.

Survival analysis has two particular characteristics that differentiate it from other

statistical techniques. First, this area deals with time-to-event data, i.e., the response of interest, is an non-negative and continuous random variable, having an asymmetric distribution. Second, the survival data usually contain censored observations (also known as partially or incomplete data). Thus, an special treatment is required to model this type of data.

Censoring comes in many forms and occurs for many different reasons. The censoring mechanism may be classified into three types, *right censoring:* occurs when the final endpoint is only known to exceed a particular value of the survival time. This censoring mechanism, can be classified into three categories: *type I*, occurs when the censoring times are pre-specified, and they may vary from individual to individual. For situation involving the *type II censoring*, the study continues until the failure of a pre-established number of subjects. This type of censoring mechanism is rare in biomedical studies, but may be used in industrial settings, where time to failure of a device is of primary interest.

The *random censoring*, which is the most common scheme, especially in biomedical studies, occurs when a person withdraws from the study. Other censoring types are, the *left censoring*, occurring when the event time is smaller than the observed time, i.e., if we know that the event occurs at a time before a left bound, but we don't know when, and *interval censoring*, in which the failure time is only known to have occurred within a specified interval of time. For more details we suggested Colosimo and Giolo (2006) and Collett (2015).

A common assumption in usual survival analysis techniques is that every individual in the study will eventually experience the event of interest, if they are followed for a long time. However, this assumption is not observed in many real situations, such as in biomedical researches involving different types of cancer, such as leukemia, melanoma, prostate and breast cancer. That is, due to the great improvements observed in treatments, it is common to occur that, the end of the follow-up period, some individuals who responded in favor of the treatment. These subjects are commonly referred as cured, immune or long-term survivors, and their survival times are considered to be infinite.

The remaining individuals are called susceptible. The presence of cured subjects can be informally detected by a simple visual inspection of the Kaplan-Meier (KM) estimator (Kaplan and Meier, 1958). In other words, if cured individuals are present, then a long plateau containing numerous data points will be observed. In this case, we can be confident that (almost) all observations in the plateau correspond to cured individuals. Additionally, a formal test of the existence of the cure fraction was presented in Maller and Zhou (1994).

Models that incorporate the proportion of cured individuals are important in the literature, and provides a simultaneous estimation for the distribution related to the cured elements (incidence part) and the susceptible subjects (latency part). In the literature, two major approaches to constructing such models have been proposed. The first one, and widely used, is the standard mixture model (SMCM) proposed by Boag (1949) and later developed by Berkson and Gage (1952). An alternative to the SMCM is the *promotion times model* (PTM) proposed by Yakovlev and Tsodikov (1996), also known as the proportional hazards cure rate model. Some advantages and disadvantages inherent of these models are listed in Chen et al. (1999).

The SMCM, which will be the focus of this thesis, is formulated in terms of a mixture of components, one representing the proportion of cured subjects, and other related to the susceptible individuals. The popularity of this model is notable due to the high number of works in the literature using this model, such as, Sy and Taylor (2000), Peng and Dear (2000), Masud et al. (2018), Naseri et al. (2018), Wang et al. (2018), Zaeran et al. (2019) only to mention some papers. Additionally, the cure rate model proposed by Yakovlev and Tsodikov (1996), involves a structure of competitive risks, and were also considered in many works, such as Tsodikov et al. (2003), Zeng et al. (2006), Ma and Yin (2008), among others.

Different specifications for the latency and incidence parts can be considered. The parametric modeling for the latency part and the logistic regression into the incidence part were proposed by several authors, including Farewell (1982), Yamaguchi (1992), Peng et al. (1998) and Achcar et al. (2012). Therefore, great attention has been devoted to the

semiparametric approach for the latency distribution. Obviously, as the name suggests, the semiparametric mixture cure (SPMCM) employs the Cox and logistic regressions to deal with the distributions related to the susceptible and cured subjects. For example Taylor (1995), Sy and Taylor (2000), Peng and Dear (2000) and Cai et al. (2012) proposed a semiparametric approach to model the latency distribution based on the expectation-maximization (EM) algorithm. The authors in Kuk and Chen (1992) considered the semiparametric marginal likelihood method to estimate the unknown parameters based on a Monte Carlo (MC) approximation algorithm. The useful alternative of the SPMCM model is the accelerated failure time mixture cure (AFTMC) model, which employs the AFTM as the latency component instead of the SPMCM model (Li and Taylor, 2002; Corbière and Joly, 2007).

The model parameters in the SMCM are estimated via the maximum likelihood method based on fully specified parametric or semiparametric structure. According to Sy and Taylor (2000), infinite estimates for the regression coefficients may occur in the estimation procedure. This phenomenon is commonly referred as ML, occurring in survival regression models, and separation problem (SP), occurring in the logistic regression model. The ML (or SP) issue appears due to specific conditions in a data set. In this case, the likelihood function converges to a finite value while at least one of the parameter estimate diverges to $\pm$ infinity (Bryson and Johnson, 1981; Albert and Anderson, 1984; Heinze and Schemper, 2001). Particularly, the ML (or SP) issue arises in situations involving many unbalanced categorical covariates; that is, when all uncensored cases are associated with one level of a categorical covariate (in survival analysis) or when an explanatory variable perfectly predicts binary events or failures (in logistic regression). In situations involving only continuous covariates, the ML issue rarely occurs. However, Klein and Moeschberger (2006) state that this phenomenon may be observed specifically when a continuous covariate is ordered, having a high correlation with the event times.

In the literature, some important references dealing with the infinite estimation problem in situations involving a non-mixture model are Andersen (1991), Heinze and Schemper (2001), Fijorek and Sokołowski (2012), Lin et al. (2013), Almeida et al. (2018)

and Wu et al. (2018) only to mention some works. In the context of logistic regression, some few references are Silvapulle (1981), Heinze and Schemper (2002), Heinze and Ploner (2003) and Zorn (2005). Some inefficient solutions to circumvent divergent estimates are commented in Heinze and Schemper (2001) and Rainey (2016), which include the following options: $(i)$ omit the risk factor directly associated to the issue; $(ii)$ fit the standard approach by ignoring the ML influence; $(iii)$ use an *ad hoc* fitting approach, where artificial data is included through a multiple imputation method; $(iv)$ test other distributions for both parts of the model (latency and incidence) and, finally, $(v)$ apply an stratified analysis on the risk factors related to the divergent coefficients.

Removing the risk factor related to the ML issue (option 1) is probably the first choice in the moment that the researcher is faced with this problem. However, this approach is unsuitable, since the removal of any important covariate represents loss of information and, furthermore, it does not allow to adjusting the effects of the remaining factors in the regression structure. The second strategy is clearly not the best way, since it leads to an overestimation of the coefficient effect and a high uncertainty for the estimates. Additionally, the *ad hoc* adjustments (also known as data manipulation) may produce finite estimates (option 3). A simple adjustment of cell frequencies can have undesirable properties, see Agresti and Yang (1987) and Clogg et al. (1991) for more detailed explanation. Models whose parameters have different interpretations that are not risk-related (option 4) may be less appealing. In addition, there is no guarantees that divergent estimates will not occur in the chosen model. Finally, like other mentioned options, the stratified analysis (option 5), see Bryson and Johnson (1981), is not a good strategy because it does not allow us to assess the real effect of the risk factor on the model.

The most popular and efficient approach to deal with this issue was suggested in Heinze and Schemper (2001), which is an adaptation of the method originally created by Firth (1992) and Firth (1993), to reduce the bias of the maximum likelihood estimates. The method removes the first-order term of the asymptotic bias of the maximum likelihood estimate (MLE) by a suitable modification of the score function. The procedure

(also known as "Firth method/correction") leads to finite estimates by penalizing the likelihood function. In this case, the penalty term can be derived from a Jeffreys prior (Jeffreys, 1946) often used in the context of Bayesian inference. Some authors state that this approach may suffer with biased estimators and high standard errors, especially in situations involving the ML/SP problem; see also Pianto and Cribari-Neto (2011), Lin et al. (2013), Greenland and Mansournia (2015) and Kenne Pagui and Colosimo (2020) for more details.

Penalization is a very general method for stabilizing estimates, which has both, frequentist and Bayesian rationale. The Firth method is a well-known example of penalization, where the penalty term can be derived from a Jeffreys prior (Jeffreys, 1946). However, this approach is not perfect and may suffer with biased estimators and high standard errors, see also Pianto and Cribari-Neto (2011), Lin et al. (2013) and Greenland and Mansournia (2015) for additional details. Under the Bayesian inference, the authors Lin et al. (2013), Greenland and Mansournia (2015) and Almeida et al. (2018) studied the ML issue by using different penalty functions (prior distributions) in the context involving non-mixture models.

## 1.1 Motivation

The study developed in the present thesis was motivated by a real data set in the context of melanoma cancer. This is a relatively new data set previously investigated in few works such as Cherobin et al. (2018) and Almeida et al. (2018). In brief, melanoma is a severe skin cancer appearing when the melanocytes (pigment producers) start to grow out of control. This disease is predominant in white adults and it is a major cause of premature deaths, especially when diagnosed in advanced stages. As a result, the premature diagnosis and the identification of patients in the high-risk group to develop metastasis are fundamental strategies to reduce the mortality rate. The data were collected between 1995-2012 at the *Hospital das Clínicas* administered by the Universidade Federal de Minas Gerais in Brazil. Besides showing a real illustration, the main aim of this application is to assess the influence of five risk factors, concerning epidemiological

and histopathologic features, on the development of metastasis in patients diagnosed with invasive primary cutaneous melanoma. The data set consists of 221 patients with 33 of them developing metastasis, which implies in $\approx 85\%$ of censored subjects.

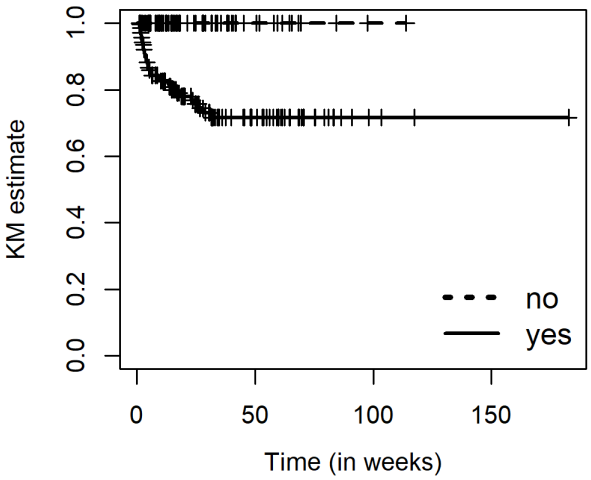


Figure 1.1: Kaplan-Meier estimate for the mitosis risk factor.

The survival curve in Figure 1.1 related to a very important covariate clearly shows a plateau indicating a cure proportion of approximately 70%. This configuration suggests the existence of long-term survivors in the melanoma data. On the other hand, the monotone likelihood issue might be seen in Figure 1.1, since the Kaplan-Meier curve for the level without mitosis is stable at 1. The problem emerges here due to the fact that the main event metastasis does not occur for the level "without mitosis". Further details about this data set will be given in Chapter 7.

## 1.2 Thesis Contributions

Solutions to divergent estimates in the context of cure fraction models have received little attention in the literature, and probably the works Almeida et al. (2021a) and Almeida et al. (2021b) are the first ones addressing this issue in the mentioned context. The reference Sy and Taylor (2000) emphasizes the occurrence of the ML problem in cure rate models, but no solution is discussed.

The main contributions of this thesis are the following: $(i)$ Handle the ML/SP by

modifying the score function based on the Firth correction (Firth, 1993); (*ii*) Investigate other flexible penalty functions (or prior distributions) to deal with this issue under the Bayesian framework; (*iii*) Conduct a sensitivity analysis, exploring different levels of the prior information; (*iv*) Develop some MC simulation studies considering different specifications for the latency distribution (parametric and semiparametric) to evaluate the finite sample proprieties of the penalized MLEs under the Bayesian and frequentist approaches. Finally, (*v*) Analyze the melanoma skin data set, which shows the presence of ML and long-term survivors.

## 1.3 Outline of the Thesis

The thesis is organized as follows: Chapter 2 describes the basic concepts of survival analysis, namely: the main censoring mechanism, likelihood function construction in the non-mixture approach, and other general elements related to the main survival models. Chapter 3 is devoted to present the cure fraction models, and the estimation procedures. Chapter 4 derives the modified score function based on the Firth approach and propose the standard errors estimation approach. Chapter 5 presents a basic concept of the Bayesian framework. The finite sample properties are investigated in Chapter 6. Chapter 7 illustrates the results related to the melanoma data analysis for both usual and penalized approaches. Finally, Chapter 8 presents the main conclusions and final remarks of this thesis.

# Chapter 2

# SURVIVAL ANALYSIS

## 2.1 Basic Concepts

The survival analysis encompasses a set of tools that aims to describe or analyze the time-to-event data. In this text, we will use the term *survival time* to refer to the response of interest, but it can also be referred to as the *failure time* or the *time-to-event*. In order to describe the non-mixture survival models, we begin denoting $T$ as a non-negative and absolutely continuous random variable, representing the survival time for some particular event from a homogeneous population.

Unlike to the usual statistical techniques, where the main functions of interest are the density function $f(t)$ and the cumulative distribution function $F(t) = P(T \leq t)$, the particularly useful function in survival analysis is named as *survival function*, which is the probability of an individual surviving until time $t$, given by

$$S(t) = P(T > t) = 1 - F(t), \quad t \geq 0. \tag{2.1}$$

Here, $t$ is the observed survival time, and $S(t) \in [0, 1]$ is a proper survival function, having the following proprieties: $(i)$ $S(t)$ is a monotone non-increasing and continuous function in $t$; $(ii)$ $S(t) = 1$, if $t = 0$, and $(iii)$ $S(t) \to 0$, when $t \to \infty$.

Another useful function in survival analysis, is the *hazard function* given by the following expression:

$$h(t) = \lim_{\Delta t \longrightarrow 0} \frac{P\left(t < T \leq t + \Delta t | T > t\right)}{\Delta t} = \frac{-\partial}{\partial t} \log S(t) \geq 0. \tag{2.2}$$

Note that, $h(t)\Delta t$ is approximately the instantaneous risk of an individual experiencing the event in $[t, t + \Delta t]$, given that he/she has survived up to $t$. Under the assumption that $T$ is absolutely continuous, the cumulative hazard is

$$H(t) = \int_0^t h(\zeta) d\zeta = -\log S(t), \quad t \geq 0. \tag{2.3}$$

For more detailed approach we recommend the following references Lawless (2003) and Klein and Moeschberger (2006). From Equation (2.3), the survival function in (2.1) can be expressed as $S(t) = \exp\{-H(t)\}$. Additionally, the probability density function has the form

$$f(t) = \frac{-\partial}{\partial t} \log S(t) = h(t) \times S(t), \tag{2.4}$$

where $f(t)$ is a proper density function with support $t \geq 0$. One of the main aims in survival analysis is to evaluate the effect of a set of risk factors on the response of interest. In brief, there are numerous types of regression models used to deal with survival data, and they can be classified according to their specification, such as "parametric models" (which can encompass for example, the Exponential, Weibull, Gamma, Log-normal, AFT models), "non-parametric" and "semiparametric modeling". In this case, the effect of potential risk factors into the survival times can be evaluated by incorporating a simple regression structure into the basic quantities described in Section 2.1. The next three sections are devoted to describe some widely used survival regression models.

## 2.2 Semiparametric Models

The Cox regression model (Cox, 1972), also known as the semiparametric model or proportional hazard model (PHM), is probably one of the most important statistical

models to analyze lifetime data. The first reason behind this popularity is the fact that the analyst does not need to know in advance the nature of the survival function. In addition, the analyst does not have to check if the assumptions, related to a given parametric distribution, are violated. The Cox regression model does not require a choice of some particular probability distribution to represent the survival time (it is a distribution-free model). The second reason is that the main estimation method associated with the PHM is a partial likelihood approach, i.e., the estimates depend only on the ranks of the event times, not their numerical values. This is an innovation in several ways. Third, even though the baseline hazard function is not specified, reasonably good estimates for the regression coefficients and adjusted survival curves can be obtained for a wide variety of data situations. Another way of saying this is that the Cox regression is a robust model, so that the obtained results will closely approximate these from the most appropriate parametric models. As additional attractive features of the PHM, one can mention that it is relatively easy to incorporate time-dependent variables, i.e., covariates that may change the value over the course of the observation period. The Cox regression model allows a kind of stratified analysis, which is very effective to control the nuisance covariates and can readily accommodate both discrete and continuous measurements of failure times, see Allison (2010), Guo (2010) and Kleinbaum and Klein (2012) for more general discussions. The conditional hazard function for the Cox regression model can be expressed as

$$h(t|\boldsymbol{\beta}, \mathbf{x}) = h_0(t) \exp\left(\mathbf{x}^\top \boldsymbol{\beta}\right), \tag{2.5}$$

where $h_0(t)$ is an arbitrary and unspecified baseline hazard function for continuous random variable $T$, which can be parametrically or non-parametrically specified. This quantity can be understood as the hazard function when all covariates are scaled to 0. The quantity $\mathbf{x} = (x_1, x_2, \cdots, x_p)^\top$ denotes the observed covariates vector of the $n \times p$ design matrix $\mathbf{X}$, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^\top$ is an unknown vector of regression coefficients. As previously described, the hazard rate in Equation (2.5) indicates potential failure rather than a survival probability.

The term *semiparametric* refers to the fact that the hazard function in (2.5) is a combination of the parametric component (i.e, the covariate effect part) and the unknown non-negative function $h_0(t)$. Additionally, the name *proportional hazard models* comes from the fact that the ratio of the hazards for two individuals $i$ and $j$ is constant over time provided that the covariates do not change over time. That is, the ratio depends only on the difference between their linear predictors at any time. Thus, it is conventional to assume that the effect on the covariates is multiplicative, leading to the hazard function. Denote by $\mathbf{x}_i$ and $\mathbf{x}_j$ the covariates vectors for subjects $i$ and $j$, respectively. From Equation (2.5), the proportional hazards assumption means that,

$$\frac{h_i(t|\boldsymbol{\beta}, \mathbf{x}_i)}{h_j(t|\boldsymbol{\beta}, \mathbf{x}_j)} = \frac{h_0(t)\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)}{h_0(t)\exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)} = \exp\left[(\mathbf{x}_i^\top - \mathbf{x}_j^\top)\boldsymbol{\beta}\right]. \tag{2.6}$$

Note that the term $\phi(\boldsymbol{\beta}) = \exp\left[(\mathbf{x}_i^\top - \mathbf{x}_j^\top)\boldsymbol{\beta}\right]$ in (2.6) does not depend on the survival time $t$, and because the covariates are not time dependent, the quantity $h_0(t)$ cancels out from this ratio, so that the hazards ratio for the two experimental subjects $i$ and $j$ is constant over the time. The main implication of this assumption is that the corresponding true survival functions, for all individuals, do not cross. If $\phi(\boldsymbol{\beta}) < 1$, the risk of failure is greater in individual $j$ than in $i$; similarly the risk of failure is greater for individual $i$ than in $j$, if $\phi(\boldsymbol{\beta}) > 1$. In order to incorporate a regression structure in the basic quantities described in Section 2.1, the conditional density and survival functions are given by

$$S(t|\boldsymbol{\beta}, \mathbf{x}) = S_0(t)^{\exp\left(\mathbf{x}^\top \boldsymbol{\beta}\right)}, \tag{2.7}$$

$$f(t|\boldsymbol{\beta}, \mathbf{x}) = h_0(t)\exp\left(\mathbf{x}^\top \boldsymbol{\beta}\right) \times S_0(t)^{\exp\left(\mathbf{x}^\top \boldsymbol{\beta}\right)}, \tag{2.8}$$

where $S_0(t) = \exp\left(-H_0(t)\right)$ is the baseline survival function. Note that, the quantity $H_0(t)$ denotes the cumulative baseline hazard function given in (3.18). The next two sections are devoted to present the estimation methods by taking into account the baseline hazard function and the partial likelihood function.

### 2.2.1 Likelihood Function

Assume that we have $n$ independent individuals, such that, the data have the form $\mathcal{D}_n = \{(t_i, \delta_i, \mathbf{x}_i), i = 1, \cdots, n\}$. For the $i$-th subject, the data consist in $t_i = \min\{T_i, C_i\}$ the observed failure time, where $T_i$ and $C_i$ denote the failure and censoring times, respectively. The observed survival times are such that $t_i$ is equal to $T_i$, if the lifetime is observed and $C_i$, if it is right-censored. The censoring indicator is: $\delta_i = 1_{\{T_i \leq C_i\}}$, taking the values $\delta_1 = 1$, if the failure occurs before the censoring time ($t_i = T_i$) or $\delta_1 = 0$, otherwise ($t_i = C_i$). Similarly, $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ip})^\top$ is the covariates vector for the $i$-th subject. A crucial assumption to construct the likelihood function in survival analysis is to assume that the survival and censoring times are independent. More specifically, denote by $\boldsymbol{\psi} = (\psi_1, \cdots, \psi_d)^\top$ the $d$-dimensional parameter space. Under the non-informative censoring mechanism, the likelihood function is

$$L(\boldsymbol{\psi}|\mathcal{D}_n) = \prod_{i=1}^{n} f(t_i|\mathbf{x}_i, \boldsymbol{\psi})^{\delta_i} S(t_i|\mathbf{x}_i, \boldsymbol{\psi})^{1-\delta_i} = \prod_{i=1}^{n} h(t_i|\boldsymbol{\psi}, \mathbf{x}_i)^{\delta_i} S(t_i|\boldsymbol{\psi}, \mathbf{x}_i). \quad (2.9)$$

In situations where the survival and censoring times are not independent, some specialized techniques must be invoked, see Kalbfleisch and Prentice (2002). From expression (2.9), its clear that the contribution of the $i$-th subject in the likelihood function is $f(t_i|\mathbf{x}_i, \boldsymbol{\psi})$ for the uncensored cases, and $S(t_i|\mathbf{x}_i, \boldsymbol{\psi})$ otherwise. From equations (2.7), (2.8) and (2.9), the log-likelihood function $\ell(\boldsymbol{\psi}|\mathcal{D}_n) = \log L(\boldsymbol{\psi}|\mathcal{D}_n)$ is

$$\begin{aligned} \ell(\boldsymbol{\psi}|\mathcal{D}_n) &= \sum_{i=1}^{n} \{\delta_i \log h(t_i|\boldsymbol{\psi}, \mathbf{x}_i) + \log S(t_i|\boldsymbol{\psi}, \mathbf{x}_i)\} \\ &= \sum_{i=1}^{n} \{\delta_i \left[\log h_0(t_i) + \mathbf{x}_i^\top \boldsymbol{\beta}\right] + \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) \log S_0(t_i)\}. \end{aligned} \quad (2.10)$$

Due to the dependence of the likelihood function in the quantity $h_0(t_i)$, a critical step to maximize (2.10) is to specify some parametric distribution for $h_0(t_i)$ or use some estimation approach that allows us to eliminate the nuisance function from the likelihood. In this work, two approaches will be described to deal with the baseline hazard function, namely: $(i)$ the parametric specification under the Weibull regression model and $(ii)$ the

semiparametric model based on the partial likelihood method. These two choices are perhaps the most applied regression models in survival analysis, due to their simplicity, flexibility, and the fact that both can be formulated in terms of proportional hazards assumption, see Cox (1972), Klein and Moeschberger (2006), Guo (2010) and Liu (2012) for more details.

The method which simplifies the parameter estimation by eliminating the unknown function $h_0(t_i)$ in the likelihood presented in (2.10) is one of the most commonly used approach. In short, the estimates of the unknown vector of regression coefficients $\boldsymbol{\beta}$ are obtained without having to specify any distribution for the baseline hazard function, meaning that, the time-dependent quantity in (2.10) is discarded, and the maximization is carried out on the remaining factor, called as *partial likelihood function*.

### 2.2.2 Partial Likelihood Function

The term partial likelihood is used because the likelihood formula considers only the probabilities for those subjects that fails, and it does not explicitly consider probabilities for the censored subjects. Thus, the likelihood for the Cox model does not consider probabilities for all subjects, and so it is called as *partial likelihood*. Based on the observed data set $\mathcal{D}_n = \{(t_i, \delta_i, \mathbf{x}_i), i = 1, \cdots, n\}$, assume that there are no ties between the event times, and that $t_{(1)} < t_{(2)} < \cdots < t_{(d^*)}$ denotes the ordered event times. Here, $\mathbf{x}_{(i)}$ is the covariate vector associated with the individual whose failure time is $t_{(i)}$. Under the non-informative censoring mechanism, Klein and Moeschberger (2006) present two general approaches to construct the partial likelihood function. The first one, by multiplying the conditional probabilities that an individual with covariates vector $\mathbf{x}_{(i)}$ fails at time $t_i$. In the second approach, the partial likelihood function can be derived by using the profile likelihood method based on the Equation (2.10), which results in the following expression

$$L(\boldsymbol{\beta}|\mathcal{D}_n) = \prod_{i=1}^{n} \left( \frac{\exp\left(\mathbf{x}_{(i)}^\top \boldsymbol{\beta}\right)}{\sum_{j \in R_{(t_i)}} \exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)} \right)^{\delta_i}, \tag{2.11}$$

where $R_{(t_{(i)})} = \{d^* : t_{d^*} \geq t_{(i)}\}$ denotes the risk set at time $t_{(i)}^-$, i.e., the set of all subjects

that are still under study at a time just prior to $t_i$, and $d^*$ denotes the number of distinct event times. The term in the denominator of Equation (2.11) is the sum of the values of $\exp\left(\mathbf{x}^\top \boldsymbol{\beta}\right)$, over all individuals that are at risk in the time point $t_{(i)}$. Note that the product in the partial likelihood function is taken over the individuals for whom the failure times have been recorded. However, censored subjects have contribution one in the likelihood. The corresponding partial log-likelihood function $\ell(\boldsymbol{\beta}|\mathcal{D}_n) = \log L(\boldsymbol{\beta}|\mathcal{D}_n)$ can be expressed as

$$\ell(\boldsymbol{\beta}|\mathcal{D}_n) = \sum_{i=1}^{n} \delta_i \left[ \mathbf{x}_{(i)}^\top \boldsymbol{\beta} - \log \sum_{j \in R_{(t_i)}} \exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right) \right]. \tag{2.12}$$

The maximization of the Equation (2.10) is carried out by using numerical methods, such as the Newton-Raphson (NR) procedure, largely used in the literature, see for example Rizzo (2007), Nash (2014). It is important to empathize that most of the main statistical softwares have convenient features, which enable the PHM to be fitted. They also provide the standard errors, hazard rate of the parameter estimates, among other quantities. The alternative partial likelihood function that allows us to accommodate tied observations or time-dependent covariates can be found for example in Klein and Moeschberger (2006), Colosimo and Giolo (2006) and Collett (2015).

## 2.3 Parametric Models

As discussed in Section 2.2, in the semiparametric approach the model parameters can be estimated without making any assumption about the distribution of the failure times. However, assumptions about how the covariates affect the survival experience are needed. As a result, the hazard function is not restricted to a specific functional form, and its shape is essentially determined by the actual data. The model has flexibility and widespread applicability (Collett, 2015).

Assumptions about how the covariates affect the survival experience are also needed here. Consequently, the functional form of the baseline hazard/survival function are completely specified according to some known probability distribution, except

for the values of the unknown parameters. These regression models are also useful to predict the survival, rather than to identify the covariates that influence the failure time. In some of them, such as Exponential and Weibull, the proportional hazards assumption is valid, meaning that a one unit change in an explanatory covariate causes a proportional change in the hazard. Similarly, in the AFT form, other special case of the model expressed in (2.5), a one unit change in an explanatory variable causes a proportional change in the survival time. An additional advantage of the AFT approach is that, the effect of the covariates on the survival can be described in absolute terms (e.g. numbers of years) rather than relative terms (a hazards ratio).

### 2.3.1 Weibull Regression Model

The Weibull regression model is one of the most used parametric models, and it can be seen as a special case of the Cox proportional hazards model, in which a functional form of $h_0(t)$ is completely specified. Before formulating the Weibull regression model, assume that the failure time $T$ has a two-parameters Weibull distribution, i.e., $T \sim \mathcal{W}(\alpha, \lambda^*)$, where $\lambda^* = \lambda \exp\left(\mathbf{x}^\top \boldsymbol{\beta}\right)$, being $\alpha$ and $\lambda = \exp(\beta_0)$, the shape and scale, respectively. Under this specification, the conditional probability density follows the particular parameterization presented in Ibrahim et al. (2001) and Collett (2015), which has the form:

$$f(t|\boldsymbol{\theta}, \mathbf{x}) = \alpha\lambda t^{\alpha-1} \exp\left(\mathbf{x}^\top \boldsymbol{\beta} - \lambda \exp\left(\mathbf{x}^\top \boldsymbol{\beta}\right) t^\alpha\right), \quad \alpha > 0\ ,\ \lambda > 0, \tag{2.13}$$

where $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \lambda, \alpha\right)^\top$ is a $(p+2)$ - dimensional parameters set for the latency part. The Exponential regression model is a special case when $\alpha = 1$ (which implies in constant hazards rates). From Equations (2.2) and (2.4), the conditional hazard and survival function are: $h(t|\boldsymbol{\theta}, \mathbf{x}) = \alpha\lambda t^{\alpha-1} \exp\left(\mathbf{x}^\top \boldsymbol{\beta}\right)$ and $S(t|\boldsymbol{\theta}, \mathbf{x}) = \exp\left(-\lambda \exp\left(\mathbf{x}^\top \boldsymbol{\beta}\right) t^\alpha\right)$, respectively. In this case, the baseline hazard function has the form $h_0(t) = \alpha\lambda t^{\alpha-1}$, which is monotonically increasing when $\alpha > 1$, decreasing when $0 < \alpha < 1$, and constant when $\alpha = 1$. Similarly, the baseline survival function has the form $S_0(t) = \exp(-\lambda t^\alpha)$. The parameters model might be obtained using the likelihood function presented in

Equation (2.9).

## 2.4 Brief Summary of the Chapter

In this chapter, we present some basics concepts about survival analysis, including: the censoring mechanism and some basic quantities. The regression models, described so far, assume that every individual in the population under study is susceptible to the event of interest and, therefore, the failure will occur, if the subject is followed long enough in terms of time. Because this assumption is not always valid in practice, the next chapter is devoted to introduce another class of regression models that are able to accommodate the limitations encountered in the usual survival techniques.

# Chapter 3

# CURE FRACTION MODELS

Survival models incorporating a cure fraction, commonly refereed to as cure rate models, become increasingly popular for analyzing data from cancer clinical trials for example. These models were developed to analyze the failure time data having a cured proportion. For such data, the conventional survival models are usually not appropriate, because they do not account for the possibility of cure.

The name *cure fraction model* is derived from its wide application in medical researches where, after a sufficiently long follow-up time, the failure may not occur for some individuals (Naseri et al., 2018; Wang et al., 2018; Zaeran et al., 2019). Some examples include studies involving patients with some breast cancer, non-Hodgkin's lymphoma, leukemia, prostate cancer, melanoma or head and neck cancer. In these cases, due to the high improvement in medical treatments, it is common to observe at the end of the study some individuals that responded favorably to the applied treatment. These subjects are known as cured, immune or long-term survivors, for which the event under study does not occur, even after a very long period of follow-up. Their survival times are considered to be infinity. The remaining individuals are said to be susceptible to the event under study, and their survival times are finite. In short, the susceptible individuals are those who will eventually experience the main event.

The cure fraction models may also be applied in different areas of knowledge.

For example, Yamaguchi (1992) illustrated several situations where this class of survival models may be appropriate, for example: some released prisoners will never be re-arrested (in sociology); some people never marry (in demography); some electronic component will never fail (in reliability, where the cure rate models are commonly refered as *limited-failure population life models*, see Meeker (1987)); in finance (some banks will never go to bankrupt); insurance (some warranty will never be claimed); in economics (some unemployed people will never find a new job). In this last example, the cure models are known as *split population models*), see Schmidt and Witte (1989).

## 3.1 Model Formulation

As mentioned in Chapter 1, there exists two major classes of cure fraction models. Before introducing these models, denote by $T^* < \infty$ the survival times for an individual in the susceptible group, and $Y$ is a partially observed random variable indicating that the subject will eventually ($Y = 1$) or never ($Y = 0$) experience the event of interest. Under the mixture population assumption, the survival time follow the decomposition (Liu et al., 2006, 2017):

$$T = YT^* + (1 - Y)\,\infty. \tag{3.1}$$

From expression (3.1), the survival times for the overall population are infinity for cured subjects ($Y = 0$). In this case, $T$ is an unobserved random variable, due to the existence of right censoring mechanism, and take the value $T^*$ for susceptible cases ($Y = 1$). Additionally, denote by $\mathbf{Z}$ another $n \times (q+1)$ covariates matrix for the incidence part, on which together with $\mathbf{X}$ described in Section 2.2, the distribution of the survival times defined in (3.1) may depend. In addition, denote by $\mathbf{z}$ the observed column vector representing a row of $\mathbf{Z}$, for which the first column contains 1's to accommodate an intercept. The covariates vector $\mathbf{z}$ might be identical, partially or completely different of $\mathbf{x}$.

Under the mixture modeling approach, the survival function for overall population $S_{pop}(t|\mathbf{x}, \mathbf{z}) = P(T > t|\mathbf{x}, \mathbf{z})$ has the form:

$$S_{pop}(t|\mathbf{x},\mathbf{z}) \;=\; 1 \;-\; \pi(\xi) \;+\; \pi(\xi)\; S(t|Y=1,\mathbf{x}), \tag{3.2}$$

where $S(t|Y=1,\mathbf{x})$ is a proper survival function for the susceptible group, similar to that given in Section 2.1, and sharing the same properties. In addition, $S_{pop}(t|\mathbf{x},\mathbf{z})$ is an improper survival function, having the following properties: $(i)$ $S_{pop}(t|\mathbf{x},\mathbf{z}) = 1$ when $t = 0$; $(ii)$ is a monotone non-increasing and continuous function in $t$, and $(iii)$ $\lim_{t\to\infty} S_{pop}(t|\mathbf{x},\mathbf{z}) = 1 - \pi(\xi) > 0$. In this case, $1-\pi(\xi)$ is the proportion of cured individuals. The structure in (3.2) can be classified as parametric, non-parametric and semiparametric mixture cure rate modeling according to the distribution specified to handle the susceptible group.

The basis of the mixture approach consists in considering that, the population under study is actually a mixture between susceptible and non-susceptible subjects. According to expression (2.2), the probability density function for overall population is given by $f_{pop}(t|\mathbf{x},\mathbf{z}) = \pi(\xi)\; f(t|Y=1,\mathbf{x})$, which is an improper density function, since it does not integrate 1. The hazard function is

$$h_{pop}(t|\mathbf{x},\mathbf{z}) = \frac{\pi(\xi)\; f(t|Y=1,\mathbf{x})}{1 \;-\; \pi(\xi) \;+\; \pi(\xi)\; S(t|Y=1,\mathbf{x})}. \tag{3.3}$$

The quantity $\pi(\xi)$ is the proportion of susceptible individuals. Under the standard mixture model, the logistic link function will be considered do evaluate $\pi(\xi)$. This decision is justified by the popularity of this link in clinical studies, and the simplicity to interpret regression coefficients affecting the log-odds of being cured or susceptible. Consequently, the proportion of susceptible subjects is

$$\pi(\xi) = P(T<\infty|\xi) = 1/\left[1+\exp(-\xi)\right],$$

where $\xi = \mathbf{z}^\top\boldsymbol{\gamma}$ is the linear predictor into the incidence distribution, and $\boldsymbol{\gamma} = (\gamma_0,\cdots,\gamma_q)^\top$ is an unknown vector of regression coefficients (Peng and Dear, 2000; Sy and Taylor, 2000; Cai et al., 2012; Masud et al., 2018). Obviously, other link functions can be considered, such as, the complementary log-log, $\log\left(-\log\left(1-\pi(\xi)\right)\right) = \mathbf{z}^\top\boldsymbol{\gamma}$, and the probit,

$\Phi^{-1}(\pi(\xi)) = \mathbf{z}^\top\boldsymbol{\gamma}$. The quantity $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. One of the main advantages of the SMCM is their simplicity in incorporating predictors in the regression structure through a link function for $\pi(\xi)$.

The promotion time model, developed by Yakovlev and Tsodikov (1996), is an alternative version of the SMCM, which is also largely discussed in the literature by several authors (Ibrahim and Laud, 1991; Chen et al., 1999; Ibrahim et al., 2001). In contrast with the SMCM, the PTM does not incorporate explicitly the partition of the population into the susceptible and the non-susceptible groups. The existence of the cured individuals is taken into account by assuming an improper survival function for the whole population. The PTM has an advantage over the previously presented approach, since it keeps the proportional hazards structure for the whole population, even in the presence of the covariates. This is one of the most desirable features, because it allows a straightforward interpretation of the covariates effects on the probability of cure, which does not occur in the SMCM.

The PTM were introduced due to its advantage in the biological interpretation of tumor cell growth. In order to explain the model, let $M$ be a random variable denoting the number of competing causes related to the occurrence of an event of interest, such that, $M$ follows a Poisson distribution with mean $\eta(\mathbf{z})$. Denote also by $W_l$, $l = 1, \cdots, M$ the promotion times for the $l$-th metastasis cells to produce detectable metastatic disease. Given $M = m$, the random variables $W_l$ are assumed to be independent and identically distributed with common distribution function $F(t) = 1 - S(t)$, which does not depend on $M$. Note that, $M$ and $W_l$ are two unobserved random quantities (latent variables). The observed time to cancer relapse is defined as $T = \min\{W_l, 0 \leq l \leq M\}$. When $M = 0$, we say that the particular subject is not susceptible to the event under study, i.e., there is no competitive causes in the occurrence of the event under study, and thus $P(W_0 = \infty) = 1$. This assumption allows us to assume that the survival times for such subjects are infinite. Following Ibrahim et al. (2001) and Yin and Ibrahim (2005), the survival function for the overall population is defined as

$$
\begin{aligned}
S_{pop}(t|\mathbf{z}) &= P(\text{no competing causes to the relapse at time t}) \\
&= P(M=0) + P(W_1 > t, \cdots, W_M > t, M \geq 1) \\
&= \exp\left(-\eta(\mathbf{z})\right) + \sum_{m=1}^{\infty} S(t)^m \frac{\eta(\mathbf{z})^m}{m!} \exp\left(-\eta(\mathbf{z})\right) \\
&= \exp\left[-\eta(\mathbf{z})(1-S(t))\right] \\
&= \exp\left[-\eta(\mathbf{z})F(t)\right]. \qquad (3.4)
\end{aligned}
$$

The quantity $\eta(\mathbf{z})$ is the link function with an intercept, and one often chooses $\eta(\mathbf{z}) = \exp\left(\mathbf{z}^\top \boldsymbol{\gamma}\right)$. Because $F(t)$ is a proper cumulative distribution function, $S_{pop}(t|\mathbf{z})$ follows the same properties of the survival function presented in (3.2). However, the proportion of cured subjects is given by $\lim_{t\to\infty} S_{pop}(t|\mathbf{z}) \equiv P(M=0) = \exp\left(-\eta(\mathbf{z})\right) > 0$. In this case, the cured fraction tends to 0, when $\eta(\mathbf{z}) \to \infty$, and tends to 1, when $\eta(\mathbf{z}) \to 0$. Note that the covariate effect may also be considered in $F(\cdot)$.

The probability density and hazard functions are $f_{pop}(t|\mathbf{z}) = \eta(\mathbf{z})f(t)\exp\left(-\eta(\mathbf{z})F(t)\right)$, with $f(t) = \partial F(t)/\partial t$, and $h_{pop}(t|\mathbf{z}) = \eta(\mathbf{z})f(t)$, respectively. Similar to the SMCM, different specification can be considered to model the cumulative distribution function $F(t)$: the piecewise exponential (Ibrahim et al., 2001; Castro et al., 2009; Demarqui et al., 2014) and the semiparametric or non-parametric specifications (Portier et al., 2017; Chen and Du, 2018; Lambert and Bremhorst, 2019). Under the proportional hazards assumption, denote by $\mathbf{z}_i$ and $\mathbf{z}_j$ the covariate vectors for the $i$-th and $j$-th subjects, respectively. Then, the quantity

$$
\frac{h_{pop}(t|\mathbf{z}_i)}{h_{pop}(t|\mathbf{z}_j)} = \frac{\eta(\mathbf{z}_i)f(t)}{\eta(\mathbf{z}_j)f(t)} = \frac{\eta(\mathbf{z}_i)}{\eta(\mathbf{z}_j)} = \exp\left((\mathbf{z}_i - \mathbf{z}_j)^\top \boldsymbol{\gamma}\right), \qquad (3.5)
$$

does not depend on $t$, which is equivalent to the proportional hazards assumption illustrated in (2.6). The biological interpretation of the promotion time cure model is presented in Chen et al. (1999). A general class of cure models were developed by Yin and Ibrahim (2005) through a Box-Cox transformation rewriting the population survival

function in the following manner

$$\frac{S_{pop}(t|\mathbf{z})^a - 1}{a} = -\eta(a, \mathbf{z})F(t), \quad 0 \leq a \leq 1. \tag{3.6}$$

Under a general class of cure rate models, the expression (3.6) lies in $\log S_{pop}(t|\mathbf{z}) = -\eta(0, \mathbf{z})F(t)$, which reduces to the (3.4), when $a \to 0$, and becomes to the expression (3.2), when $a = 1$. Additionally, the constant $a$ can mathematically assume any value on the real line; see Yin and Ibrahim (2005) for further details.

The frequentist estimates of the parameter set $\boldsymbol{\psi}$ can be obtained through routines to maximize the observed likelihood function. The MLEs, and the corresponding asymptotic properties are described in the next Sections.

## 3.2 Estimation Procedures

Denote by $\mathcal{D}_{obs} = \{(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i), i = 1, \cdots, n\}$ the observed data, which is similar to the data set defined in Section 2.2.1. Now, however, we include the $i$-th line of the covariates vector for the incidence part $\mathbf{z}_i$. Additionally, $\mathbf{x}_i$ is the covariates vector for the $i$-th subject affecting the latency distribution, and $t_i$ the observed survival time in the mixture population, given in expression (3.1). Under the standard mixture cure rate model, denote by $\omega_i^* = 1_{\{T_i^* \leq C_i\}}$ the failure indicator for the susceptible group. The censoring indicator for overall population is now defined as $\delta_i = 0$ for cured subjects ($Y_i = 0$) and $\delta_i = \omega_i^*$, otherwise. Here, $Y_i$ denotes the $i$-th row of the partially observed random vector $\mathbf{Y}$. The quantities $T_i^*$ and $C_i$ are respectively, the survival and censoring times for the $i$-th subject in the susceptible group.

Denote the full parameters set as $\boldsymbol{\psi} = \left(\boldsymbol{\theta}^\top, \boldsymbol{\gamma}^\top\right)^\top$, where $\boldsymbol{\theta}$ is the parameter subset for the latency part and $\boldsymbol{\gamma}$ is the subset for the incidence distribution. Note that the dimension of $\boldsymbol{\psi}$ varies according to the distribution specified to model the latency part, i.e., $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \lambda, \alpha\right)^\top$. We have $d = p + q + 3$, in the parametric mixture cure fraction model, and $\boldsymbol{\theta} \equiv \boldsymbol{\beta}$, with $d = p + q + 1$, for the semiparametric mixture cure fraction model. The intercept $\gamma_0$, included in $\boldsymbol{\gamma}$, reflects the overall incidence of $Y_i = 1$ for every

$i$, while the remaining terms in $\boldsymbol{\gamma}$ allows the covariates to influence the probability of an individual being susceptible. In addition, $\boldsymbol{\beta}$ reflects the effect of the covariate vectors on the time to event distribution (Farewell, 1986). Under the non-informative censoring mechanism, the observed likelihood function is

$$L(\boldsymbol{\psi}|\mathcal{D}_{obs}) = \prod_{i=1}^{n} \left[\pi(\xi_i) f(t_i|\boldsymbol{\theta}, Y_i = 1, \mathbf{x}_i)\right]^{\delta_i} \left[1 - \pi(\xi_i) + \pi(\xi_i) S(t_i|\boldsymbol{\theta}, Y_i = 1, \mathbf{x}_i)\right]^{1-\delta_i}. \quad (3.7)$$

Note that the contribution of the $i$-th subject in the likelihood is $\pi(\xi_i)\ f(t_i|\boldsymbol{\theta}, Y_i = 1, \mathbf{x}_i)$ for uncensored cases, and $1 - \pi(\xi_i) + \pi(\xi_i) S(t|\boldsymbol{\theta}, Y_i = 1, \mathbf{x}_i)$ otherwise. If we assume that all subjects are susceptible to the event under study, i.e., $\pi(\xi_i) = 1$, for every $i$, the likelihood function in (3.7) reduces to the likelihood of the standard approach given in Equation (2.9).

The MLE $\hat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ can be found by numerical maximization of the log-likelihood function $\ell(\boldsymbol{\psi}|\mathcal{D}_{obs}) = \log L(\boldsymbol{\psi}|\mathcal{D}_{obs})$. In many situations involving high dimensional parameter sets, the maximization of $\ell(\boldsymbol{\psi}|\mathcal{D}_{obs})$ under the frequentist approach may present numerical problems. A solution to circumvent this issue is to find the MLE using a method based on augmented data, i.e., a latent variable might be included in the likelihood construction (Sy and Taylor, 2000; Cho et al., 2001). In this strategy, the complete log-likelihood function can be factorized in two parts related to the latency and incidence structure of the model, which facilitates not only separated maximizations, but also the separated penalization of the likelihood functions as a solution to handle the infinite estimates problem. Due to the numerical problems reported in the literature, the maximization of the observed log-likelihood, and other advantages in using the complete data likelihood, the next section presents the estimation procedure based on the EM algorithm proposed by Dempster et al. (1977). The EM algorithm is also explored in the literature by Peng and Dear (2000), Fang et al. (2005), Cai et al. (2012), Masud et al. (2018) and He and Emura (2019), only to mention a few works.

## 3.3 The EM Algorithm

The EM algorithm is an iterative method widely used to find the MLEs in situations involving the presence of a partially observed data or missing information, also known as a latent variable. In order to use the EM algorithm, denote by $\mathcal{D}_c = \{\mathcal{D}_{obs}, \mathbf{Y}\}$, the augmented data, which contain the observed data $\mathcal{D}_{obs}$ and the vector of partially observed random variables $\mathbf{Y} = (Y_1, \cdots, Y_n)^\top$. The vector $\mathbf{Y}$ is said to be "partially observed" due to the following aspects: $(i)$ $\delta_i = 1 \Rightarrow Y_i = 1$ and $(ii)$ if $\delta_i = 0$, then $Y_i$ is an unobserved binary random variable (which can be either 0 or 1). Under the complete data approach, we can rewrite the density and survival functions for the $i$-th subject in the mixture population as: $[\pi(\xi_i)f(t_i|\boldsymbol{\theta}, \mathbf{x}_i)]^{Y_i}$ and $(1 - \pi(\xi_i))^{1-Y_i}[\pi(\xi_i)S(t_i|\boldsymbol{\theta}, \mathbf{x}_i)]^{Y_i}$, which are similar to the quantities presented in Section 3.1. In this case, under the non-informative censoring mechanism, the likelihood function based on the complete data is given by

$$L_c(\boldsymbol{\psi}|\mathcal{D}_c) = \prod_{i=1}^{n} [\pi(\xi_i)f(t_i|\boldsymbol{\theta}, \mathbf{x}_i)]^{Y_i\delta_i} \times ([1 - \pi(\xi_i)]^{1-Y_i} [\pi(\xi_i)S(t_i|\boldsymbol{\theta}, \mathbf{x}_i)]^{Y_i})^{1-\delta_i}.$$

After some algebras, and using the results in (2.5) and (2.7), the complete data likelihood function has the form

$$\begin{aligned} L_c(\boldsymbol{\psi}|\mathcal{D}_c) &= \prod_{i=1}^{n} \left[(1 - \pi(\xi_i))^{Y_i-1} h(t_i|\boldsymbol{\theta}, \mathbf{x}_i)^{Y_i}\right]^{\delta_i} \times ([1 - \pi(\xi_i)]^{1-Y_i} [\pi(\xi_i)S(t_i|\boldsymbol{\theta}, \mathbf{x}_i)]^{Y_i}) \\ &= \prod_{i=1}^{n} \pi(\xi_i)^{Y_i}(1 - \pi(\xi_i))^{1-Y_i} \times \prod_{i=1}^{n} \left[h_0(t_i) \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right]^{\delta_i} S_0(t_i)^{Y_i \exp\left(\mathbf{x}_i^\top \beta\right)} \\ &= L_1(\boldsymbol{\gamma}|\mathcal{D}_c)\ L_2(\boldsymbol{\theta}|\mathcal{D}_c). \end{aligned} \tag{3.8}$$

This is factored into two components: $L_1(\boldsymbol{\gamma}|\mathcal{D}_c)$ for the incidence, and $L_2(\boldsymbol{\theta}|\mathcal{D}_c)$ related to the latency distribution, respectively. The last expression were obtained by assuming that $\delta_i Y_i = \delta_i$. Additionally, the complete data log-likelihood function is given by $\ell_c(\boldsymbol{\psi}|\mathcal{D}_c) = \ell_1(\boldsymbol{\gamma}|\mathcal{D}_c) + \ell_2(\boldsymbol{\theta}|\mathcal{D}_c)$, where

$$\ell_1(\boldsymbol{\gamma}|\mathcal{D}_c) = \log L_1(\boldsymbol{\gamma}|\mathcal{D}_c) = \sum_{i=1}^{n} \left[Y_i \log \pi(\xi_i) + (1 - Y_i) \log(1 - \pi(\xi_i))\right] \text{ and}$$

$$\ell_2(\boldsymbol{\theta}|\mathcal{D}_c) = \log L_2(\boldsymbol{\theta}|\mathcal{D}_c) = \sum_{i=1}^{n} \left[\delta_i \left(\log h_0(t_i) + \mathbf{x}_i^\top \boldsymbol{\beta}\right) + Y_i \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) \log S_0(t_i)\right].$$

In contrast with the observed likelihood in Equation (3.7), the advantage of using the EM algorithm is that, the MLEs of $\boldsymbol{\theta}^\top$ and $\boldsymbol{\gamma}^\top$ will be obtained separately, since $\ell_1(\boldsymbol{\gamma}|\mathcal{D}_c)$ depends only on $\boldsymbol{\gamma}$, and $\ell_2(\boldsymbol{\theta}|\mathcal{D}_c)$ depends only on $\boldsymbol{\theta}$. Note that, $\ell_1(\boldsymbol{\gamma}|\mathcal{D}_c)$ is equivalent to the log-likelihood function of the logistic regression model. Similarly, apart from the latent variable $Y_i$, the log-likelihood function $\ell_2(\boldsymbol{\theta}|\mathcal{D}_c)$ is similar to that presented in Equation (2.10) based on the standard survival technique.

The EM algorithm consists of two steps: the expectation step (E-step) and the maximization step (M-step). In the E-step, we take the expectation of $\ell_c(\boldsymbol{\psi}|\mathcal{D}_c)$ with respect to the distribution of the latent variable $Y_i$, given the current value of $\boldsymbol{\psi}$ at the $m$-th iteration and the observed data $\mathcal{D}_{obs}$. Note that, when applying the logarithm in the complete likelihood function in Equation (3.8), the term $Y_i$ will appear within a sum involving a linear structure in $\ell_c(\boldsymbol{\psi}|\mathcal{D}_c)$. This linear structure is an important feature to compute the expected value of the complete data log-likelihood. That is, for censored cases, $\tilde{\ell}_c(\boldsymbol{\psi}|\mathcal{D}_c) = \mathbb{E}[\ell_c(\boldsymbol{\psi}|\mathcal{D}_c)|\boldsymbol{\psi}^{(m)}, \mathcal{D}_{obs}]$, where $\boldsymbol{\psi}^{(m)}$ denotes the current value of $\boldsymbol{\psi}$ in the $m$-th iteration. This expectation can be determined by the following sum $\tilde{\ell}_c(\boldsymbol{\psi}|\mathcal{D}_c) = \tilde{\ell}_1(\boldsymbol{\gamma}|\mathcal{D}_c) + \tilde{\ell}_2(\boldsymbol{\theta}|\mathcal{D}_c)$. The expected complete data log-likelihood for the incidence and latency part are, respectively

$$\tilde{\ell}_1\left(\boldsymbol{\gamma}|\mathcal{D}_{obs}, \mathbf{g}^{(m)}\right) = \sum_{i=1}^{n} \left[g_i^{(m)} \log \pi(\xi_i) + (1 - g_i^{(m)}) \log(1 - \pi(\xi_i))\right], \tag{3.9}$$

$$\tilde{\ell}_2\left(\boldsymbol{\theta}|\mathcal{D}_{obs}, \mathbf{g}^{(m)}\right) = \sum_{i=1}^{n} \left[\delta_i \left(\log h_0(t_i) + \mathbf{x}_i^\top \boldsymbol{\beta}\right) + g_i^{(m)} \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) \log S_0(t_i)\right], \tag{3.10}$$

where $\mathbf{g}^{(m)} = (g_1^{(m)}, \cdots, g_n^{(m)})$. For the $i$-th subject, the term $g_i^{(m)}$ denotes the expected value of the partially missing variable $Y_i$, given the observed data and the current value of the parameter set $\boldsymbol{\psi}$. For censored cases, we have the following result:

$$
\begin{aligned}
g_i^{(m)} &= \mathbb{E}\left(Y_i|\boldsymbol{\psi}^{(m)}, \mathcal{D}_{obs}\right) \\
&= \delta_i + (1-\delta_i)\, P\left(Y_i = 1|\boldsymbol{\psi}^{(m)}, T_i > t_i, \mathcal{D}_{obs}\right) \\
&= \delta_i + (1-\delta_i)\, P\left(Y_i = 1|\boldsymbol{\psi}^{(m)}, T_i > t_i, \delta_i = 0, \mathbf{x}_i, \mathbf{z}_i\right) \\
&= \delta_i + (1-\delta_i)\, \frac{P\left(\{Y_i = 1\} \cap (\boldsymbol{\psi}^{(m)}, T_i > t_i, \delta_i = 0, \mathbf{x}_i, \mathbf{z}_i)\right)}{P\left(\boldsymbol{\psi}^{(m)}, T_i > t_i, \delta_i = 0, \mathbf{x}_i, \mathbf{z}_i\right)} \\
&= \delta_i + (1-\delta_i)\, \frac{\pi(\xi_i) S_0(t_i)^{\exp\left(\mathbf{x}_i^\top \beta\right)}}{1 - \pi(\xi_i) + \pi(\xi_i) S_0(t_i)^{\exp\left(\mathbf{x}_i^\top \beta\right)}}\Bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}^{(m)}}. \qquad (3.11)
\end{aligned}
$$

The EM algorithm replaces the partially observed factor $Y_i$, in $\ell_1(\boldsymbol{\gamma}|\cdot)$ and $\ell_2(\boldsymbol{\theta}|\cdot)$, by its expected value $g_i^{(m)}$. Such as illustrated in expression (3.11), the contribution of each individual in the $\tilde{\ell}_c(\boldsymbol{\psi}|\mathcal{D}_c)$ varies between the censored and uncensored subjects. More specifically, the contribution is 1, if the $i$-th individual is uncensored, and $\pi(\xi_i^{(m)}) S_0^{(m)}(t_i)^{\exp\left(\mathbf{x}_i^\top \beta^{(m)}\right)} / \left[1 - \pi(\xi_i^{(m)}) + \pi(\xi_i^{(m)}) S_0^{(m)}(t_i)^{\exp\left(\mathbf{x}_i^\top \beta^{(m)}\right)}\right]$, otherwise. Note that $\xi_i^{(m)}$ is the value of the linear predictor $\xi_i = \mathbf{z}_i^\top \boldsymbol{\gamma}$ evaluated in the $m$-th iteration.

The M-step involves the separated maximization of $\tilde{\ell}_1(\boldsymbol{\gamma}|\mathcal{D}_{obs}, \mathbf{g}^{(m)})$ and $\tilde{\ell}_2(\boldsymbol{\theta}|\mathcal{D}_{obs}, \mathbf{g}^{(m)})$ with respect to the unknown parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, respectively. These two steps (E and M) iterate until a convergence is achieved. After the EM algorithm has converged, the quantity $\hat{\boldsymbol{\psi}}^{(m+1)}$ might be seen as the MLE of the parameters set $\boldsymbol{\psi}$, and the weights $g_i^{(m+1)}$ might be interpreted as the conditional probability that the $i$-th subject is susceptible to the main event, given the observed data. Under the standard approach, Equation (3.9) can be easily maximized by using a `glm` package available in `R` (R Core Team, 2021).

The solution to the score equation $\mathbf{U}(\boldsymbol{\psi}) = \mathbf{0}$ provides the MLE for the parameters set $\boldsymbol{\psi}$. In most cases, the resulting formula to obtained from the score equation cannot be expressed in a closed form, thus numerical methods such as the NR procedure are required to maximize simultaneously the (3.9) and (3.10). According to this approach, the MLE for the parameters set $\boldsymbol{\psi}$, at the $m$-th iteration, is given by,

$$
\hat{\boldsymbol{\psi}}^{(m+1)} = \hat{\boldsymbol{\psi}}^{(m)} + \mathbf{I}^{-1}(\hat{\boldsymbol{\psi}}^{(m)}) \mathbf{U}(\hat{\boldsymbol{\psi}}^{(m)}), \qquad (3.12)
$$

for $m = 0, 1, 2, \cdots$. The quantities $\mathbf{U}(\hat{\boldsymbol{\psi}}^{(m)})$ and $\mathbf{I}^{-1}(\hat{\boldsymbol{\psi}}^{(m)})$ are the scores vector, and observed information matrix, both evaluated at the point $\hat{\boldsymbol{\psi}}^{(m)}$. The iterative step in (3.12) is frequently started assuming $\hat{\boldsymbol{\psi}}^{(0)} = 0$ (for the regression coefficients), and 1 for the shape and scale parameters. For simplicity, we will rewrite the expected complete log-likelihood functions for the latency and incidence distributions as $\tilde{\ell}_1\left(\boldsymbol{\gamma}, \mathbf{g}^{(m)}\right)$ and $\tilde{\ell}_2\left(\boldsymbol{\theta}, \mathbf{g}^{(m)}\right)$, respectively. The $s$-th component of the usual score function for the incidence part $\mathbf{U}_1(\boldsymbol{\gamma}) = \partial\tilde{\ell}_1\left(\boldsymbol{\gamma}, \mathbf{g}^{(m)}\right)/\partial\boldsymbol{\gamma}$, has the following form

$$\mathbf{U}_1(\gamma_s) = \sum_{i=1}^{n} z_{si}\left[g_i^{(m)} - \pi(\xi_i)\right], \tag{3.13}$$

where $s = 0, 1, \cdots, q$. Likewise, denote by $\partial\pi(\xi_i)/\partial\gamma_b = z_{bi}\pi(\xi_i)\left[1 - \pi(\xi_i)\right]$ the derivative of $\pi(\xi_i)$ with respect to the parameter $\gamma_b$, for $b = 0, 1, \cdots, q$. The entry $(s, b)$ of the Fisher information matrix $\mathbf{I}_1(\boldsymbol{\gamma}) = \mathbb{E}\left[-\partial^2\tilde{\ell}_1\left(\boldsymbol{\gamma}, \mathbf{g}^{(m)}\right)/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top\right]$ is given by

$$\mathbf{I}_1(\gamma_{sb}) = \sum_{i=1}^{n} \pi(\xi_i)\left[1 - \pi(\xi_i)\right] z_{si} z_{bi}. \tag{3.14}$$

The Fisher information matrix, can also be rewritten as $\mathbf{I}_1(\boldsymbol{\gamma}) = \mathbf{z}^\top\mathbf{w}(\boldsymbol{\gamma})\mathbf{z}$, where $\mathbf{w}(\boldsymbol{\gamma}) = \text{diag}\left(\{\kappa_{2i}\}_{i=1,\cdots,n}\right)$ is a diagonal matrix. In addition, $\kappa_{2i} = \pi(\xi_i)\left[1 - \pi(\xi_i)\right]$ denotes the second-order cumulant, or simply the variance of the partially observed random variable. The difficult to maximize (3.10) depends on how the conditional baseline hazard function is specified. As previously stated, we propose to model the latency part using the parametric and semiparametric approaches. In this context, we present in the next two sections a general procedure to obtain the MLEs for both parametric and semiparametric mixture cure fraction models.

### 3.3.1 Parametric Mixture Cure Rate Model

The parametric cure rate model is obtained when the lifetime data for the susceptible group are modeled parametrically. As mentioned earlier, under the Weibull regression model, the conditional hazard and survival functions for the susceptible group are, respectively, $h_i(t|\cdot) = \alpha\lambda t_i^{\alpha-1}\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)$ and $S_i(t|\cdot) = \exp\left(-\lambda t_i^\alpha\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right)$. Consequently,

the expected log-likelihood function for the latency part is

$$\tilde{\ell}_2\left(\boldsymbol{\theta}|\mathcal{D}_{obs},\mathbf{g}^{(m)}\right)=\sum_{i=1}^{n}\left\{\delta_i\left[\mathbf{x}_i^\top\boldsymbol{\beta}+(\alpha-1)\log(t_i)+\log(\alpha\lambda)\right]-g_i^{(m)}\lambda t_i^\alpha\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right\}. \tag{3.15}$$

Following a similar idea of the incidence part, $r\in\{1,2,\cdots,p\}$, the score vector in the latency part $\mathbf{U}_2(\boldsymbol{\theta})=\left(\partial\tilde{\ell}_2\left(\boldsymbol{\theta},\mathbf{g}^{(m)}\right)/\partial\beta_r,\tilde{\ell}_2\left(\boldsymbol{\theta},\mathbf{g}^{(m)}\right)/\partial\lambda,\partial\tilde{\ell}_2\left(\boldsymbol{\theta},\mathbf{g}^{(m)}\right)/\partial\alpha\right)$ has the following components

$$\begin{aligned}
U_2(\beta_r) &= \sum_{i=1}^{n}x_{ri}g_i^{(m)}[\delta_i-\lambda t_i^\alpha\exp(\mathbf{x}_i^\top\boldsymbol{\beta})]\\
U_2(\lambda) &= \sum_{i=1}^{n}g_i^{(m)}\left[\frac{\delta_i}{\lambda}-t_i^\alpha\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right]\\
U_2(\alpha) &= \sum_{i=1}^{n}g_i^{(m)}\left[\delta_i\left(\frac{1}{\alpha}+\log(t_i)\right)-\lambda\exp(\mathbf{x}_i^\top\boldsymbol{\beta})t_i^\alpha\log(t_i)\right].
\end{aligned}$$

For convenience, we denote by $\mathbf{J}(\boldsymbol{\theta})$ the observed information matrix corresponding to the latency part, where $\mathbf{J}(\boldsymbol{\theta})=-\partial^2\tilde{\ell}_2\tilde{\ell}_2\left(\boldsymbol{\theta},\mathbf{g}^{(m)}\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ and $\mathbf{I}_2(\boldsymbol{\theta})=\mathbb{E}\left[\mathbf{J}(\boldsymbol{\theta})\right]$ is the expected information matrix. For $u\in\{1,2,\cdots,p\}$, the observed matrix $\mathbf{J}(\boldsymbol{\theta})$ has the following components

$$\begin{aligned}
\mathrm{J}_{\beta_r\beta_u} &= \sum_{i=1}^{n}g_i^{(m)}x_{ri}x_{ui}\lambda t_i^\alpha\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\\
\mathrm{J}_{\lambda\lambda} &= \sum_{i=1}^{n}g_i^{(m)}\left(\delta_i/\lambda^2\right)\\
\mathrm{J}_{\alpha\alpha} &= \sum_{i=1}^{n}g_i^{(m)}\left[\frac{\delta_i}{\alpha^2}+\lambda t_i^\alpha\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\log^2(t_i)\right]\\
\mathrm{J}_{\alpha\beta_r} &= \sum_{i=1}^{n}x_{ri}g_i^{(m)}t_i^\alpha\lambda\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\log(t_i)\\
\mathrm{J}_{\lambda\beta_r} &= \sum_{i=1}^{n}g_i^{(m)}x_{ri}t_i^\alpha\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\\
\mathrm{J}_{\lambda\alpha} &= \sum_{i=1}^{n}g_i^{(m)}t_i^\alpha\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\log(t_i).
\end{aligned}$$

From Equation (3.12), the iterative procedure for both (incidence and latency part) are expressed by $\boldsymbol{\gamma}^{(m+1)}=\boldsymbol{\gamma}^{(m)}+\mathbf{I}_1^{-1}(\boldsymbol{\gamma}^{(m)})\mathbf{U}_1(\boldsymbol{\gamma}^{(m)})$ and $\boldsymbol{\theta}^{(m+1)}=\boldsymbol{\theta}^{(m)}+\mathbf{I}_2^{-1}(\boldsymbol{\theta}^{(m)})\mathbf{U}_2(\boldsymbol{\theta}^{(m)})$,

respectively.

### 3.3.2 Semiparametric Mixture Cure Rate Model

The profile likelihood approach is the main procedure to construct the expected complete likelihood function in the cure rate models under the semiparametric specification for the latency part. Thus, assume that there is no ties, and the baseline hazard function is discrete with $h_{0i} = h_0(t_i)$, for $i \in \{1, 2, \cdots, n\}$, so that $H_0(t) = \sum_{i:t_{(i)} \leq t} h_{0i}$. Without loss of generality, suppose that the survival times are ordered $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$. Thus, the Equation (3.10) can be expressed in the following way

$$\begin{aligned} \tilde{\ell}_2\left(\boldsymbol{\beta}, \mathbf{h_0} | \mathcal{D}_{obs}, \mathbf{g}^{(m)}\right) &= \sum_{i=1}^{n}\left[\delta_i\left(\log h_{0i} + \mathbf{x}_{(i)}^{\top}\boldsymbol{\beta}\right) - \sum_{i=1}^{j} h_{0i} g_i^{(m)} \exp\left(\mathbf{x}_i^{\top}\boldsymbol{\beta}\right)\right] \\ &= \sum_{i=1}^{n}\left[\delta_i\left(\log h_{0i} + \mathbf{x}_{(i)}^{\top}\boldsymbol{\beta}\right) - h_{0i}\sum_{j=i}^{n} g_j^{(m)} \exp\left(\mathbf{x}_j^{\top}\boldsymbol{\beta}\right)\right]. \quad (3.16) \end{aligned}$$

Note that our first interest is to estimate $\boldsymbol{\beta}$. To this end, a commonly employed trick is to eliminate the nuisance parameter $\mathbf{h_0} = (h_{01}, h_{02}, \cdots, h_{0n})^{\top}$ by using the profile likelihood approach obtained from the full censored-data likelihood (Klein and Moeschberger, 2006). More specifically, we will first maximize $\tilde{\ell}_2\left(\boldsymbol{\beta}, \mathbf{h_0} | \mathcal{D}_{obs}, \mathbf{g}^{(m)}\right)$ with respect to the vector $\mathbf{h_0}$. The last expression is maximized when $h_{0i} = 0$, for $i \in \{1, \cdots, n\}$, except for time points at which the events under study occurs. For a fixed $\boldsymbol{\beta}$, the profile MLE of the baseline hazard function can be obtained by deriving (3.16) with respect to $h_{0i}$. That is

$$\frac{\partial \tilde{\ell}_2\left(\boldsymbol{\beta}, h_{01}, \cdots, h_{0n} | \mathcal{D}_{obs}, \mathbf{g}^{(m)}\right)}{\partial h_{0i}} = \frac{\delta_i}{h_{0i}} - \sum_{j=i}^{n} g_j^{(m)} \exp\left(\mathbf{x}_j^{\top}\boldsymbol{\beta}\right).$$

Finally, the profile MLE for $h_{0i}$ has the form

$$\hat{h}_{0i}\left(\boldsymbol{\beta}, \mathbf{g}^{(m)}\right) = \frac{\delta_i}{\sum\limits_{j \in R_{(t_i)}}^{n} g_j^{(m)} \exp\left(\mathbf{x}_j^{\top}\boldsymbol{\beta}\right)}. \quad (3.17)$$

Assuming that $T$ is absolutely continuous random variable and denoting by $R_{(t_i)} = \{d^* : t_{d^*} \geq t_{(i)}\}$ the risk set at time $t_{(i)}^-$ (also defined in Section 2.2.2), the Breslow-type estimator of the cumulative baseline hazard function, at the $m$-th iteration of the NR procedure, has the form:

$$\hat{H}_0^{(m+1)}\left(t|\boldsymbol{\beta}, \mathbf{g}^{(m)}\right) = \sum_{i:t_{(i)}\leq t}\left(\frac{\delta_i}{\sum\limits_{j\in R_{(t_i)}} g_j^{(m)} \exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}^{(m+1)}\right)}\right). \tag{3.18}$$

Note that, the expression (3.18) can be seen as a modification of the Nelson-Aalen (Nelson, 1969; Aalen, 1978) estimator. The term $\boldsymbol{\beta}^{(m+1)}$ is the estimate in $(m+1)$-th iteration. Based on (3.18), the conditional baseline survival function is given by

$$\hat{S}_0^{(m+1)}\left(t|\boldsymbol{\beta}, \mathbf{g}^{(m)}\right) = \exp\left[-\sum_{i:t_{(i)}\leq t}\left(\frac{\delta_i}{\sum\limits_{j\in R_{(t_i)}} g_j^{(m)} \exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}^{(m+1)}\right)}\right)\right]. \tag{3.19}$$

In cure fraction modeling, it is common to observe survival times greater than the largest uncensored time point $t_{(d^*)}$. In this case, the estimated baseline survival function in (3.19) does not tend to zero, when $t \to \infty$. In other words, $\hat{S}^{(m+1)}(t|y=1,\mathbf{x}) = \hat{S}_0^{(m+1)}(t_{(d^*)}+0|y=1)^{\exp\{\mathbf{x}^\top\boldsymbol{\beta}^{(m+1)}\}} > 0$, which determines an in improper distribution for the susceptible subjects, meaning that these individuals will have a nonzero probability to be free of failure. As a consequence, identifiability issues may occur in the semiparametric cure model, and the model is overparameterized. That is, the intercept term in the logistic predictor and the improper distribution of the susceptible individuals cannot be estimated. Even though the use of covariates in this mixture model solves some identification problems, the baseline group is still overparameterized. In order to deal with the identifiability issue, and consequently obtain a proper distribution for the susceptible subjects, the tail of the estimated survival function for such individuals must be fixed or completed (Peng, 2003).

In order to handle the upper tail of the estimated baseline survival function, and then avoid numerical instability, Taylor (1995) suggested to set $\hat{S}_0^{(m+1)}(t_{(d^*)}|y=1)$ to zero

for all $t > t_{(d^*)}$, meaning that, a zero-tail constraint is imposed on the estimated baseline survival function if there are censored survival times greater than the last uncensored time. However, this method, referred to as the Taylor tail completion (Taylor, 1995), implies that subjects with survival times greater than $t_{(d^*)}$ are always considered as cured. The method may overcompensate the bias from the identifiability issue, and its performance in cure models needs to be carefully examined.

Motivated by the situations previously described, we focused on the method proposed by Peng (2003) to estimate a proper baseline survival function $\hat{S}_0^{(m+1)}(t_{(d^*)}|\cdot)$ by setting $\hat{S}_0^{(m+1)}(t_{(d^*)} + 0|\cdot)$ as zero, smoothly. Thus, replacing (3.17) in (3.16) the profile likelihood function for $\boldsymbol{\beta}$ became proportional to

$$\tilde{\ell}_2\left(\boldsymbol{\beta}, S_0|\mathcal{D}_{obs}, \mathbf{g}^{(m)}\right) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_{(i)}^\top \boldsymbol{\beta} - \log\left(\sum_{j \in R_{(t_i)}} g_j^{(m)} \exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)\right)\right]. \tag{3.20}$$

Therefore, apart from the weights $g_j^{(m)}$, the profile likelihood function in (3.20) resembles the usual partial log-likelihood function in the Cox regression model presented in (2.11). In this case, the `coxph` function of the survival package (available in `R`) might be used to maximize $\tilde{\ell}_2(\boldsymbol{\beta}, S_0|\mathcal{D}_{obs}, \mathbf{g}^{(m)})$, with an additional offset variable $g_j^{(m)}$. A detailed discussion can be found in Sy and Taylor (2000) and Peng (2003). Another method to estimate the conditional baseline survival function is based on *the product-limit estimator*, largely discussed in Sy and Taylor (2000) and Peng and Dear (2000). This approach is based on a discrete distribution for the baseline survival function, and follows a similar argument presented in Kalbfleisch and Prentice (2002).

From the profile expected log-likelihood function given in (3.20), the score function $\mathrm{U}_2(\beta_r) = \partial\tilde{\ell}_2\left(\boldsymbol{\beta}, S_0|\mathcal{D}_{obs}, \mathbf{g}^{(m)}\right)/\partial\beta_r$ and the entry $(r, u)$ of the information matrix $\mathbf{I}_2(\boldsymbol{\beta}) = \mathbb{E}\left[-\partial^2\tilde{\ell}_2\left(\boldsymbol{\beta}, S_0|\mathcal{D}_{obs}, \mathbf{g}^{(m)}\right)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top\right]$ are written as

$$\mathrm{U}_2(\beta_r) = \sum_{i=1}^{n} \delta_i \left[ x_{(i)r} - \sum_{j \in R_{(t_i)}} x_{jr} w_j \right] \tag{3.21}$$

$$\mathrm{I}_2(\beta_{ru}) = \sum_{i=1}^{n} \delta_i \left[ \sum_{j \in R_{(t_i)}} x_{jr} x_{ju} w_j - \left( \sum_{j \in R_{(t_i)}} x_{jr} w_j \right) \left( \sum_{j \in R_{(t_i)}} x_{ju} w_j \right) \right], \tag{3.22}$$

where $w_i = g_i^{(m)} \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right) / \sum_{j \in R_{(t_i)}} g_j^{(m)} \exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)$, and $r, u \in \{1, 2, \cdots, p\}$ . The EM algorithm starts with initial values for $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\gamma}^{(0)}$ and $S_0^{(0)}$, and then iterates between the E and M steps until a convergence is achieved. The iterative procedure, when using the semiparametric structure the latency part, is summarized in Algorithm 1.

**Algorithm 1** EM algorithm for semiparametric cure rate model

**Input:** $\left(\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)}, S_0^{(0)}\right)^\top$, $\mathcal{D}_c = \{(t_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i, Y_i), i = 1, 2, \cdots, n\}$.

**Output:** The MLEs $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}, \hat{S}_0\right)^\top$.

1. Initialize the parameters $\left(\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)}, S_0^{(0)}\right)^\top$ and set $m = 1$.
2. **E-step:** Compute the weights $g_i^{(m)}$ using Equation ( 3.11).
   Compute $\tilde{\ell}_1\left(\boldsymbol{\gamma} | \mathcal{D}_{obs}, g_i^{(m)}\right)$ and $\tilde{\ell}_2\left(\boldsymbol{\beta}, S_0 | \mathcal{D}_{obs}, g_i^{(m)}\right)$ using (3.9) and (3.10), respectively.
   Compute the conditional baseline survival function $\hat{S}_0^{(m)}$ from expression (3.19).
3. **M-step:** Compute $\boldsymbol{\beta}^{(m+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\ \tilde{\ell}_2\left(\boldsymbol{\beta}, S_0 | \mathcal{D}_{obs}, g_i^{(m)}\right)$ and $\boldsymbol{\gamma}^{(m+1)} = \underset{\boldsymbol{\gamma}}{\operatorname{argmax}}\ \tilde{\ell}_1\left(\boldsymbol{\gamma} | \mathcal{D}_{obs}, g_i^{(m)}\right)$. Set $\boldsymbol{\psi}^{(m+1)} = \left(\boldsymbol{\beta}^{(m+1)}, \boldsymbol{\gamma}^{(m+1)}\right)^\top$.
4. **Convergence check:** IF $\left\|\boldsymbol{\psi}^{(m+1)} - \boldsymbol{\psi}^{(m)}\right\|^2 + \left\|S_0^{(m+1)} - S_0^{(m)}\right\|^2 < \zeta$,

   BREAK,

   ELSE,

   **Set** $m = m + 1$, and go to step (2).
5. **Return** $\boldsymbol{\psi}^{(m+1)}$.

The notation $\|\cdot\|$ denotes the Euclidean norm. In brief, the MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are

obtained in each iteration and then inserted in (3.11) and (3.19) to update these quantities. However, such as discussed in Chapter 1, due to the presence of the ML issue in the data set, one may observe divergent estimates for some regression coefficients directly related to the highly unbalanced dichotomous covariate ($\beta_1$ and $\gamma_1$). Consequently, the NR algorithm does not converge. In other words, the following functions $\tilde{\ell}_1\left(\boldsymbol{\gamma}|\mathcal{D}_{obs}, \mathbf{g}^{(m)}\right)$ and $\tilde{\ell}_2\left(\boldsymbol{\theta}, S_0|\mathcal{D}_{obs}, \mathbf{g}^{(m)}\right)$ may not have a maximum for some regression coefficients, meaning that, they are strictly monotonic towards $\pm\infty$, thus the score equations in (3.13) and (3.21) do not have a finite solution, i.e., $\mathbf{U}(\psi_k) \neq 0$ for some $k$ in $\{1, 2, \cdots, d\}$.

Figure 3.1 presents the behavior of the score functions for the regression coefficients related to the mitosis factor, which is promoting the ML issue in the melanoma data set mentioned in Chapter 1, for both latency and incidence part. Here, we are illustrating the profile score function for the parametric mixture cure rate model. The figure clearly shows that, the MLEs for the coefficients $\beta_1$ and $\gamma_1$, based on the usual score function, do not assume finite values, since the profiled score function is always above 0. A similar graph for the semiparametric mixture cure model can be found in Figure 4.2 of the Appendix B.2. A more detailed analysis of the real data application is presented in Chapter 7.

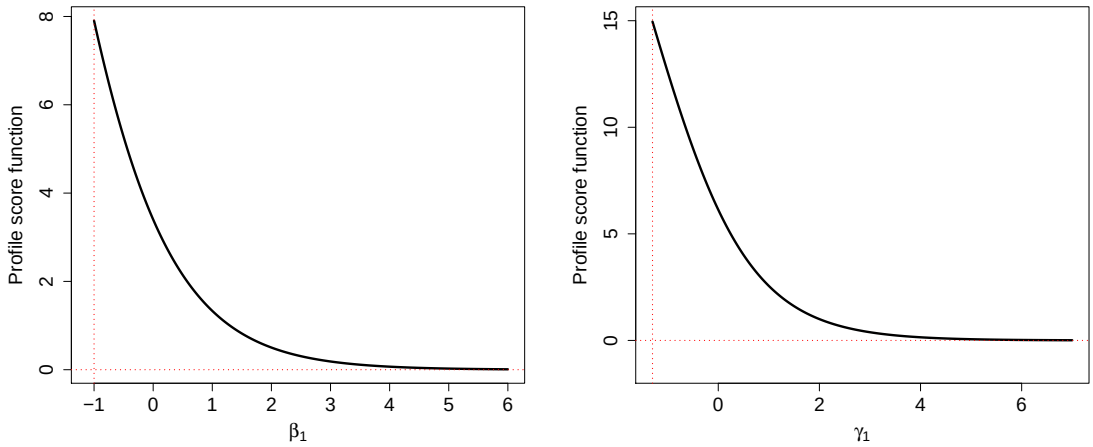


Figure 3.1: Behavior of the usual score functions for the coefficient related to the binary covariate mitosis in the latency part (left panel) and incidence part (right panel). The parametric mixture cure model is considered here.

## 3.4 Brief Summary of the Chapter

This chapter introduced the standard mixture cure fraction model, one of the most used regression models for situations involving mixture population. The fundamentals related to parameter estimation methods were also presented. The EM algorithm described in this chapter was fundamental in the sense that it allowed us to: (1) carry out a separate maximization of the two parts of the model and (2) derive an equivalent version of the partial likelihood function in the semiparametric mixture fraction cure model. Obviously, the result in (2) would not be obtained if we considered the observed likelihood in (3.7). In addition, the likelihood function based on the augmented data will allow us to penalize separately the latency and incidence distributions. This point will be discussed in the next chapter, where the modified version for the score functions will be derived as an strategy to avoid divergent estimates found when using the standard score functions in a data set affected by the ML issue.

# Chapter 4

# MODIFIED SCORE FUNCTIONS AND STANDARD ERROR ESTIMATION

The value $\hat{\psi}_k$ maximizing the expected complete log-likelihood functions $\tilde{\ell}_1\left(\boldsymbol{\gamma}, \mathbf{g}^{(m)}\right)$ or $\tilde{\ell}_2\left(\boldsymbol{\theta}, \mathbf{g}^{(m)}\right)$ has some important properties, if the usual regularity conditions are satisfied (Kosmidis, 2014). The regularity conditions are: $(i)$ the model must be identifiable; $(ii)$ the parameter set $\boldsymbol{\psi}$ must have a finite dimension $d$; $(iii)$ the parameter space cannot depend on the support of the distribution, i.e., $d$ cannot depend on the sample size $n$, and $(iv)$ there exists a sufficient number of expected log-likelihood derivatives under $\boldsymbol{\psi}$.

The identifiability problem in the cure rate models was discussed in Li et al. (2001) and Hanin and Huang (2014). In Chapter 1, we described some inefficient methods to deal with the ML/SP issue and other efficient methods based on the bias reduction approach. The bias is a fundamental property of an estimator, and can be defined as the difference between the estimated and the corresponding true parameter value. Additionally, small bias is a desirable result for a good estimator, i.e., if an experiment is repeated indefinitely, then the average of all estimates will be close to the true value to be estimated. Thus, bias of an estimator $\hat{\boldsymbol{\psi}}$ is defined as

$$b(\boldsymbol{\psi}) = \mathbb{E}_{\boldsymbol{\psi}}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) \simeq \frac{b_1(\boldsymbol{\psi})}{n} + \mathcal{O}(n^{-2}). \tag{4.1}$$

In this case, $n$ is the sample size and $b_1(\boldsymbol{\psi})$ is the first-order term in the asymptotic expansion of the MLE bias. The element $b_f(\boldsymbol{\psi})$, for $f = 1, 2, 3, \cdots$, denotes the $f$-th term in the bias expansion of $\hat{\boldsymbol{\psi}}$, which is $\mathcal{O}(1)$ as $n \to \infty$. The general approach for the bias expansion, presented in Equation (4.1), can be found in Bartlett (1953), Cox and Snell (1968) and Cordeiro and McCullagh (1991).

The literature devoted to this issue distinguishes the bias treatment into two methods: *explicit* and *implicit*. The explicit methods (or bias-correction methods) are such that the corrected MLE is obtained by subtracting the quantity in (4.1) from the $\hat{\boldsymbol{\psi}}$. In short, the bias-correction methods are based on two main steps: $(i)$ calculate the asymptotic bias in expression (4.1) and $(ii)$ subtract the asymptotic bias from the MLE. In this case, we have

$$\hat{\boldsymbol{\psi}}_{cor} = \hat{\boldsymbol{\psi}} - b(\hat{\boldsymbol{\psi}}). \tag{4.2}$$

The most popular bias-correction methods in the literature are the jackknife, bootstrap, and the methods which are based on the approximations of the bias function through asymptotic expansions, such as in (4.1). Two advantages of the explicit methods are: its simplicity for application, and the fact that the correction involves a two-steps procedure, as described above (Kosmidis, 2014). Likewise, some disadvantages of explicit methods are: due to its dependence on $\hat{\boldsymbol{\psi}}$, the corrected bias estimator, directly inherit the instabilities of the original MLE. In addition, these methods depend upon the finiteness of the MLE, i.e., the bias-correction methods are undefined when $\hat{\psi}_k$ is infinite, and also when the term $b(\boldsymbol{\psi})$ cannot be obtained in closed-form. In this case, the bias corrected estimate in (4.2) is undefined, due to the non-existence of the MLE. Furthermore, the explicit methods are inappropriate in situations involving sparse or unbalanced data, for which there is a non-zero probability that the MLE is infinite. This is the case for many categorical-response models (Silvapulle, 1981; Albert and Anderson, 1984; Bull et al., 2002; Heinze and Schemper, 2002), and also for survival regression models (Bryson and

Johnson, 1981; Heinze and Schemper, 2001) when the categorical response or the event under study are not associated with any level of the dichotomous covariates.

A solution to circumvent the mentioned problem is to use the implicit methods since the bias-reduction estimators are obtained here by solving the modified score equation. Loosely speaking, the implicit methods do not directly depend on the original MLEs, and thus, they are effectively well defined even when the MLEs does not exist. The Firth correction (Firth, 1993) which will be discussed in the next section, is one of the most widely used implicit methods.

## 4.1 Modified Score Functions

The proposed approach to deal with the ML issue relies on the modified score function based on the Firth correction (Firth, 1992, 1993). The Firth method is a preventive approach where the order term $\mathcal{O}(n^{-1})$ of the asymptotic bias can be removed from the original estimate $\hat{\psi}_k$ by a suitable modification of the score functions. The resulting modified estimators are first-order unbiased. The method was initially proposed to reduce the bias of MLEs in the generalized linear models, and became popular due to the works of Heinze and Schemper (2001) and Heinze and Schemper (2002), which present adaptations to handle the ML issue in both Cox and logistic regression models, respectively. Before describing the mentioned correction, consider the joint vector of the parameter set $\boldsymbol{\psi} = (\psi_1, \cdots, \psi_d)^\top$. The modified score function under the preventive approach has the form

$$\mathrm{U}^*(\psi_k) = \mathrm{U}(\psi_k) + \mathrm{A}(\psi_k), \quad \text{for} \quad k = 1, \cdots, d. \tag{4.3}$$

In this case, consider that $\mathrm{U}(\psi_k)$ is the standard score function and $\mathrm{A}(\psi_k)$ the penalty term such that, $\mathrm{A}(\psi_k)$ is $\mathcal{O}(1)$ as $n \to \infty$. This term represents the $k$-th component of the vector of penalties $\mathrm{A}(\boldsymbol{\psi}) = -\mathbf{I}(\boldsymbol{\psi})b(\boldsymbol{\psi})$, where $b(\boldsymbol{\psi})$ is the asymptotic bias expansion presented in (4.1). The general expression for the penalty term is given by:

$$\mathrm{A}(\psi_k) \;=\; \frac{1}{2}\sum_{u=1}^{d}\sum_{v=1}^{d} \mathrm{I}^{uv}\nu_{u,v,k} \;=\; \frac{1}{2}\mathrm{tr}\left[\mathbf{I}^{-1}(\boldsymbol{\psi})\left(\frac{\partial \mathbf{I}(\boldsymbol{\psi})}{\partial \psi_k}\right)\right], \tag{4.4}$$

where $\mathbf{I}(\boldsymbol{\psi})$ is the expected or observed information matrix, for $u, v, k \in \{1, 2, \cdots, d\}$, and the quantity $\nu_{u,v,k} = \partial \mathrm{I}_{uv}/\partial \psi_k$ denotes the third-order cumulant. Finally, the quantities $\mathrm{I}_{uv}$ and $\mathrm{I}^{uv}$ are the entries $(u, v)$ of $\mathbf{I}(\boldsymbol{\psi})$ and $\mathbf{I}^{-1}(\boldsymbol{\psi})$, respectively. The penalized MLE $\boldsymbol{\psi}^*$ is then obtained by solving the modified score equation $\mathbf{U}^*(\boldsymbol{\psi}) = \mathbf{0}$. Note that $\mathbf{U}^*(\boldsymbol{\psi})$ is related to the complete likelihood in (3.8). Then, the version of the expected complete penalized likelihood function is $\tilde{L}_c^*\left(\boldsymbol{\psi}, \mathbf{g}^{(m)}\right) = \tilde{L}_c\left(\boldsymbol{\psi}, \mathbf{g}^{(m)}\right)|\mathbf{I}(\boldsymbol{\psi})|^{1/2}$. The logarithmic version is

$$\tilde{\ell}_c^*\left(\boldsymbol{\psi}, \mathbf{g}^{(m)}\right) = \tilde{\ell}_c\left(\boldsymbol{\psi}, \mathbf{g}^{(m)}\right) + 1/2 \log|\mathbf{I}(\boldsymbol{\psi})|. \tag{4.5}$$

The quantity $|\mathbf{I}(\boldsymbol{\psi})|^{1/2}$ in (4.5) is the penalty term. For exponential family with canonical parameterization, Firth (1993) showed that the imposed penalty term might be interpreted as the non-informative Jeffrey's invariant prior (Jeffreys, 1946) widely used in the Bayesian context. In this case, the maximum of $\tilde{\ell}_c^*\left(\boldsymbol{\psi}, \mathbf{g}^{(m)}\right)$ can be seen as the posterior mode under the Jeffrey's prior. From the complete likelihood function in Equation (3.8), the information matrices $\mathbf{I}_1(\cdot)$ and $\mathbf{I}_2(\cdot)$ depend only on $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$, respectively. These aspects suggest that (3.9) and (3.10) can be penalized separately. In this case, $\tilde{\ell}_1^*\left(\boldsymbol{\gamma}, \mathbf{g}^{(m)}\right) = \tilde{\ell}_1\left(\boldsymbol{\gamma}, \mathbf{g}^{(m)}\right) + 1/2 \log|\mathbf{I}_1(\boldsymbol{\gamma})|$ and $\tilde{\ell}_2^*\left(\boldsymbol{\theta}, \mathbf{g}^{(m)}\right) = \tilde{\ell}_2\left(\boldsymbol{\theta}, \mathbf{g}^{(m)}\right) + 1/2 \log|\mathbf{I}_2(\boldsymbol{\theta})|$. Note that, the term $1/2 \log|\mathbf{I}_1(\boldsymbol{\gamma})|$ depends on components expressed in (3.14); therefore, it can be maximized by making $\pi(\xi) = 1/2$, which implies that $\boldsymbol{\gamma} = \mathbf{0}$.

In term of notation, denote $\kappa_{3i}$ as the third-order cumulant for the partially observed random variable, such that, $\kappa_{3i} = \partial \kappa_{2i}/\partial \gamma_s = z_{si}\left[1 - 2\pi(\xi_i)\right]\kappa_{2i}$, with $\kappa_{2i} = \pi(\xi_i)\left[1 - \pi(\xi_i)\right]$ and $\mathbf{w}(\gamma_s) = \partial \mathbf{w}(\boldsymbol{\gamma})/\partial \gamma_s = \mathrm{diag}\left(\{\kappa_{3i}\}_{i=1,\cdots,n}\right)$. Note that the diagonal matrix $\mathbf{w}(\boldsymbol{\gamma})$ was defined in Section 3.3. From Equation (4.3), the penalty term for the incidence part is $\mathrm{A}(\gamma_s) = \frac{1}{2}\sum_{i=1}^n z_{si} h_i\left(\frac{\kappa_{3i}}{\kappa_{2i}}\right) = \sum_{i=1}^n z_{si} h_i (1/2 - \pi(\xi_i))$. As a result, the modified score function can be expressed in the following way

$$
\begin{aligned}
\mathrm{U}_1^*(\gamma_s) &= \mathrm{U}_1(\gamma_s) + \mathrm{A}(\gamma_s) \\
&= \sum_{i=1}^{n} z_{si}\left[g_i^{(m)} - \pi(\xi_i)\right] + \sum_{i=1}^{n} z_{si} h_i \left(1/2 - \pi(\xi_i)\right) \\
&= \sum_{i=1}^{n} z_{si}\left\{\left[g_i^{(m)} - \pi(\xi_i)\right] + h_i\left[1/2 - \pi(\xi_i)\right]\right\}. \qquad (4.6)
\end{aligned}
$$

In this formulation, $h_i$ denotes the leverage values for the $i$-th observation, i.e., $i$-th diagonal element of the hat matrix $\mathbf{H}(\boldsymbol{\gamma}) = \mathbf{w}(\boldsymbol{\gamma})\mathbf{z}\left[\mathbf{z}^\top \mathbf{w}(\boldsymbol{\gamma})\mathbf{z}\right]^{-1}\mathbf{z}^\top$. A solution to the modified equation in (4.6) can be obtained by using the usual `glm` package in `R`, if the values $h_i$, for $i = 1, 2, \cdots, n$ are known. However, in practice it is more likely that the values $h_i$ are unknown, since they are function of an unknown coefficients vector $\boldsymbol{\gamma}$. Then, the `logistf` package available in `R` (Heinze et al., 2013), or an iterative procedure, can be used to obtain the solution to the modified score equation. Note that the expression (4.6) reduces to (3.13), when $h_i = 0$ or when $\pi(\xi_i) = 1/2$.

The third-order cumulants computation for the latency part depends on the distribution specified to model the survival times (parametric or semiparametric specifications). The general expression is given by:

$$
\begin{aligned}
\mathrm{U}_2^*(\theta_r) &= \mathrm{U}_2(\theta_r) + A\left(\theta_r\right) \\
&= \mathrm{U}_2(\theta_r) + \frac{1}{2}\mathrm{tr}\left\{\mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\left(\frac{\partial \mathbf{J}(\boldsymbol{\theta})}{\partial \theta_r}\right)\right\}, \qquad (4.7)
\end{aligned}
$$

where $r = 1, \cdots, p+2$ (for parametric specification), and $r = 1, \cdots, p$, otherwise. The corresponding expressions of the third-order cumulants $\nu_{u,v,r} = \partial \mathrm{J}_{uv}/\partial\theta_r$ for the parametric mixture cure rate model are given in Appendix A.1, where $\mathrm{J}_{uv}$ is the entry $(u, v)$ of the observed information matrix $\mathbf{J}(\boldsymbol{\theta})$ related to the latency part. Similarly, under the semiparametric specification, the dimension of $\boldsymbol{\theta}$ reduces then, Equation (4.7) can be expressed as

$$\begin{aligned}
\mathrm{U}_2^*(\beta_r) &= \mathrm{U}_2(\beta_r) + \mathrm{A}(\beta_r) \\
&= \sum_{i=1}^{n} \delta_i \left[ x_{(i)r} - \sum_{j \in R_{(t_i)}} x_{jr} w_j \right] + \frac{1}{2} \sum_{u=1}^{p} \sum_{v=1}^{p} \mathrm{J}^{uv} \nu_{u,v,r},
\end{aligned}$$

where $\mathrm{J}^{uv}$ is the entry $(u, v)$ of $\mathbf{J}^{-1}(\boldsymbol{\beta})$ in which its components are given in (3.22). The third-order cumulants $\nu_{u,v,r} = \partial \mathrm{J}_{uv}/\partial \beta_r$ are now expressed as

$$\begin{aligned}
\nu_{u,v,r} = \sum_{i=1}^{n} \delta_i \Bigg\{ & \sum_{j \in R_i} x_{jr} x_{ju} x_{jv} w_j - \left( \sum_{j \in R_i} x_{jr} x_{ju} w_j \right) \left( \sum_{j \in R_i} x_{jv} w_j \right) \\
& - \left( \sum_{j \in R_i} x_{jr} x_{jv} w_j \right) \left( \sum_{j \in R_i} x_{ju} w_j \right) - \left( \sum_{j \in R_i} x_{ju} x_{jv} w_j \right) \left( \sum_{j \in R_i} x_{jr} w_j \right) \\
& + 2 \left( \sum_{j \in R_i} x_{jr} w_j \right) \left( \sum_{j \in R_i} x_{ju} w_j \right) \left( \sum_{j \in R_i} x_{jv} w_j \right) \Bigg\}.
\end{aligned}$$

The penalized maximum likelihood estimates are obtained by replacing $\mathbf{U}(\boldsymbol{\psi})$ by $\mathbf{U}^*(\boldsymbol{\psi})$ in the iterative NR procedure given in (3.12) and adopting the same information matrices of non-penalized cases. The evaluation is now made with respect to the penalized MLE $\hat{\boldsymbol{\psi}}^* = \left(\hat{\boldsymbol{\theta}}^{*\top}, \hat{\boldsymbol{\gamma}}^{*\top}\right)^\top$. The general idea, presented in Firth (1992) and Firth (1993) is that the first-order asymptotic covariance matrix of $\hat{\boldsymbol{\psi}}^*$ is the same as the one for the usual MLE $\hat{\boldsymbol{\psi}}$, namely $\mathbf{I}_2(\hat{\boldsymbol{\theta}}^*)$ and $\mathbf{I}_1(\hat{\boldsymbol{\gamma}}^*)$.

Based on empirical findings of Heinze and Schemper (2001), Zorn (2005) and Lin et al. (2013), and the proofs in Firth (1993), the penalized MLEs always exist and they are unique, even in the presence of the ML issue. In order to illustrate the finiteness of the penalized MLEs, we plot again, the profiled score function, for both usual and modified cases (under the parametric mixture cure rate model). Figure 4.1 clearly shows that, finite values for the MLEs are obtained after modifying the score functions (using the Firth method). The penalty term was able to pull down the profiled score function in order to assume negative values, meaning that finite solutions for the MLEs can be found.

Similarly, the profiled score functions for both usual and modified cases related to

the semiparametric mixture cure rate model are illustrated in Figure 4.2. An additional implicit method was proposed by Kenne Pagui et al. (2017), which is based on the median distribution of the MLEs. Other variants of the Firth correction method are also available in the literature; see Elgmati et al. (2015) and Lima and Cribari-Neto (2016) for more details.

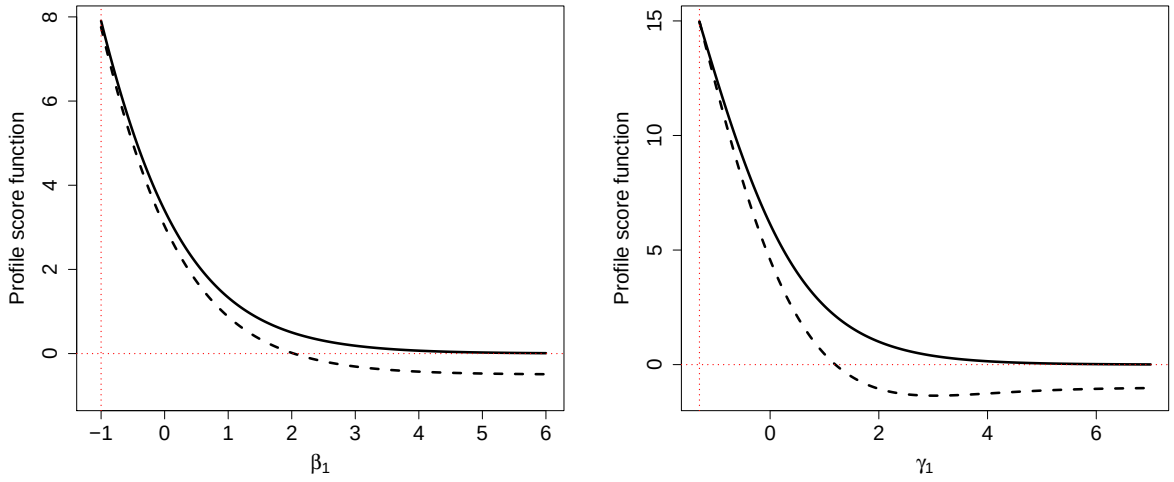


Figure 4.1: Behavior of the modified score functions for the coefficient related to the binary covariate mitosis in the latency part (left panel) and incidence part (right panel); $U(\psi_k)$ (solid line) and $U^*(\psi_k)$ (dashed line).

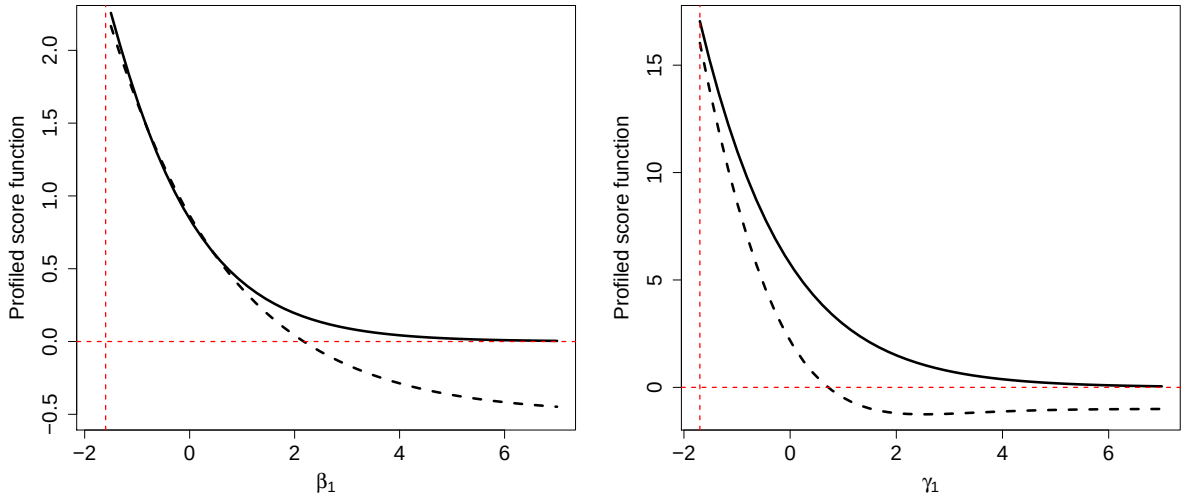


Figure 4.2: Behavior of the modified score functions (under the semiparametric specification) for the coefficient related to the binary covariate mitosis in the latency part (left panel) and incidence part (right panel); $U(\psi_k)$ (solid line) and $U^*(\psi_k)$ (dashed line).

One main property of the penalized estimates, which is also shared by the estimator from asymptotic bias correction, is that it has the same asymptotic distribution as the original MLEs. In other words, they are normally distributed with mean $\boldsymbol{\psi}$ and variance-covariance matrix $\mathbf{J}_{obs}(\boldsymbol{\psi})$. In the next section, we present a general approach to compute the information matrix in the standard mixture cure fraction model.

## 4.2 Standard Error Estimation

A common critique for the EM algorithm is that, due to its primary conceptual power in converting a maximization problem involving a complex likelihood function, into a sequence of separate estimation procedures, the standard errors of the resulting MLEs are not automatically available when using the algorithm.

The standard error is a good measure of the amount of information that the data can provide about the parameters being estimated. The way to obtain the information matrix for $\hat{\boldsymbol{\psi}}$, is based on the inverse of the observed covariance matrix $\mathbf{J}_{obs}\left(\boldsymbol{\psi}\right) = -\partial^2 \ell\left(\boldsymbol{\psi}|\mathcal{D}_{obs}\right)/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top$, which depends on the observed likelihood function given in (3.7). However, this strategy may be difficult, or at least tedious, especially in situations involving a high-dimensional set of parameters and an infinite-dimensional component $H_0(t|\cdot)$.

Due to the difficulty of estimating the standard errors, a variety of methods are available to simplify this procedure, such as the bootstrap method proposed in Efron and Tibshirani (1993), and largely applied in the context involving the mixture cure rate models by Peng (2003), Cai et al. (2012) and Niu and Peng (2013), the so called sandwich variance estimation method, suggested by Rosen et al. (2000) for situations involving a general mixture of marginal models, and the Louis' method (Louis, 1982), largely discussed in the literature. The general idea of the Louis' method, which will be considered in this thesis, is that the observed information $\mathbf{J}_{obs}\left(\boldsymbol{\psi}\right)$ can be obtained as the difference between the information based on the augmented data, that is, $\mathbf{J}_{\mathcal{D}_c}\left(\boldsymbol{\psi}\right) = -\partial^2 \ell_c\left(\boldsymbol{\psi}|\mathcal{D}_c\right)/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top$ and that regarding the distribution of the partially observed random variable, conditioned

to the observed data $\mathbf{J}_{Y|\mathcal{D}_{obs}}(\boldsymbol{\psi}) = -\partial^2 \ell_{Y|\mathcal{D}_{obs}}(\boldsymbol{\psi}, Y|\mathcal{D}_{obs})/\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}^\top$. In other words, the observed information matrix has the form $\mathbf{J}_{obs}(\boldsymbol{\psi}) = \mathbf{J}_{\mathcal{D}_c}(\boldsymbol{\psi}) - \mathbf{J}_{Y|\mathcal{D}_{obs}}(\boldsymbol{\psi})$, see Louis (1982) and Givens and Mallick (2013) for more details.

In this thesis, the approach proposed by Sy and Taylor (2000) to compute the observed information matrix in the standard mixture cure model will be considered. The procedure can be seen as a variant of the Louis method, being computationally more attractive, and similar to the traditional Louis approach, it takes into account the complete and missing data distributions. In short, the observed score function for $\psi_k$ can be expressed as $\mathrm{U}_{obs}(\psi_k) = \partial\ell_{obs}(\boldsymbol{\psi}|D_{obs})/\partial\psi_k = \partial\ell_c(\boldsymbol{\psi}|D_c)/\partial\psi_k$, now evaluated at $Y_i = g_i$. Then, for $s, b \in \{0, 1, \cdots, q\}$, the observed score function related to the incidence part is

$$\mathrm{U}_{obs}(\gamma_s) = \left.\frac{\partial\ell_c(\boldsymbol{\psi}|\mathcal{D}_c)}{\partial\gamma_s}\right|_{Y_i=g_i} = \sum_{i=1}^{n}\left\{z_{ib}(g_i - \pi(\xi_i))\right\}.$$

Similarly, the observed score function for the latency part under the parametric specification are:

$$\begin{aligned}
\mathrm{U}_{obs}(\beta_r) &= \left.\frac{\partial\ell_c(\boldsymbol{\psi}|\mathcal{D}_c)}{\partial\beta_r}\right|_{Y_i=g_i} = \sum_{i=1}^{n} x_{ri} g_i \left[\delta_i - \lambda t_i^{\alpha}\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right], \\
\mathrm{U}_{obs}(\lambda) &= \left.\frac{\partial\ell_c(\boldsymbol{\psi}|\mathcal{D}_c)}{\partial\lambda}\right|_{Y_i=g_i} = \sum_{i=1}^{n} g_i \left[\frac{\delta_i}{\lambda} - t_i^{\alpha}\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\right], \\
\mathrm{U}_{obs}(\alpha) &= \left.\frac{\partial\ell_c(\boldsymbol{\psi}|\mathcal{D}_c)}{\partial\alpha}\right|_{Y_i=g_i} = \sum_{i=1}^{n} g_i \left[\delta_i\left(\frac{1}{\alpha} + \log t_i\right) - t_i^{\alpha}\lambda\exp\left(\mathbf{x}_i^\top\boldsymbol{\beta}\right)\log(t_i)\right],
\end{aligned}$$

where $r, u \in \{1, 2, \cdots, p\}$. The corresponding components of the observed covariance matrix $\mathbf{J}_{obs}(\boldsymbol{\psi}) = -\partial\mathrm{U}_{obs}(\boldsymbol{\psi})/\partial\boldsymbol{\psi}$, in the parametric mixture cure fraction model, are given in Appendix A.2. In addition, assume the following quantities $D_{j,0} = \sum_{j\in R_{(t_i)}} Y_j \exp\left(\mathbf{x}_j^\top\boldsymbol{\beta}\right)$ and $D_{j,r} = \sum_{j\in R_{(t_i)}} x_{jr} Y_j \exp\left(\mathbf{x}_j^\top\boldsymbol{\beta}\right)$. Then, the observed score functions in the semiparametric mixture cure model is

$$\mathrm{U}_{obs}(\beta_r) \;=\; \sum_{i=1}^{n} \delta_i \left\{ x_{ir} - \frac{D_{j,r}}{D_{j,0}} \right\} \Bigg|_{Y_i = g_i}.$$

The components of the $d \times d$ covariance matrix for the observed data, when using the semiparametric specifications in the latency distribution, are given in Appendix A.3. The general discussion concerning the observed covariance matrix in the semiparametric mixture cure rate models are presented in Sy and Taylor (2000) and Sy and Taylor (2001).

## 4.3 Asymptotic Properties

The theoretical properties illustrated in this work, for the parametric mixture cure model, are also discussed in our recently published paper Almeida et al. (2021a). In short, the required regularity conditions and main asymptotic results are established in this section. In addition, we outline the proofs of some related theorems. Denote by $\boldsymbol{\psi}_0 = (\boldsymbol{\theta}_0^\top, \boldsymbol{\gamma}_0^\top)^\top$ the vector of true values and $\hat{\boldsymbol{\psi}} = \left(\hat{\boldsymbol{\theta}}^\top, \hat{\boldsymbol{\gamma}}^\top\right)^\top$ the corresponding MLEs based on the complete data likelihood $L_c(\boldsymbol{\psi}, \mathbf{g}^{(m)})$. The following regularity conditions are necessary to derive the asymptotic properties.

- C1. The vectors of covariates $\mathbf{x}_i \in \mathbb{R}^p$ and $\mathbf{z}_i \in \mathbb{R}^{q+1}$ are uniformly bounded, i.e., there exists a finite and positive constant $\epsilon_0$, such that $|x_i| < \epsilon_0$ and $|z_i| < \epsilon_0$, for $i \in \{1, \cdots, n\}$.

- C2. Conditional on the bounded vectors of covariates, the right censoring $C_i$ and the survival times $T_i$ are independent, and for the unobserved random variable $Y_i$, it is necessary to assume that $P\left(C_i = \infty \text{ and } T_i = \infty | \mathbf{x}_i, \mathbf{z}_i\right) > 0$.

- C3. The true parameter value $\boldsymbol{\psi}_0$ lies within a known compact set $\mathcal{B} \subset \mathbb{R}^d$ such that, for a some constant $B_0$, we have:

$$\mathcal{B} = \left\{ (\boldsymbol{\beta}, \lambda, \alpha, \boldsymbol{\gamma}) \;:\; |\boldsymbol{\beta}| \leq B_0,\; |\boldsymbol{\gamma}| \leq B_0,\; \lambda \text{ and } \alpha \text{ are bounded away for } \mathbb{R}^+ \right\}.$$

- C4. The true baseline distribution function for the susceptible group, say $F_0(t|\cdot)$, is non-decreasing and differentiable with $f_0(t|\cdot) = \partial F_0(t|\cdot)/\partial t > 0$, $\forall\, t \in \mathbb{R}^+$.

The condition C1 means that, if $\mathbf{x}_i$ and $\mathbf{z}_i$ are observed vectors of covariates, there exists $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$ such that, $\mathbf{x}_i^\top \tilde{\boldsymbol{\beta}} = 0$ and $\mathbf{z}_i^\top \tilde{\boldsymbol{\gamma}} = 0$ almost surely, which is equivalent to the assumption of the linear independence of $\mathbf{x}_i$ and of $\mathbf{z}_i$. The following assumption $P\left(C_i = \infty \text{ and } T_i = \infty | \mathbf{x}_i, \mathbf{z}_i\right) > 0$ in C2 ensures that, some subjects in the data set are cured and they are not right-censored. In situations where C2 is not true, Li et al. (2001) and Hanin and Huang (2014) propose to evaluate one of following assumptions:

- A1. The regression parameters, except the intercept in $\boldsymbol{\gamma}$, cannot all be zero;
- A2. The survival function $S(t|\cdot)$ must have a parametric form;
- A3. There exists a constant $\tau_0$ that satisfies $P\left(C_i = \tau_0 \text{ and } T_i = \tau_0 | \mathbf{x}_i, \mathbf{z}_i\right) > 0$.

Note that A1 is related to the model identification and can be proved by induction, which requires that the survival function for the susceptible group in Equation (3.2) is proper. The assumption A2 is clearly valid in the context where the latency part is modeled parametrically. Finally, A3 is equivalent to C2, when $\tau_0 = \infty$. Condition C4 can be easily verified through a parametric specification of the latency distribution. Next, we present some important theorems related to the asymptotic properties of the MLEs in $\hat{\boldsymbol{\psi}}$. We would like to point out that, these theorems have already been proved in the literature; see, for example, Murphy et al. (1994), Scharfstein et al. (1998), Fang et al. (2005) and Lu (2010).

**Theorem 1** (Existence and uniqueness)**.** *Under the conditions C1–C4 and in the absence of the ML issue, the quantities* $\hat{\boldsymbol{\psi}} = \left(\hat{\boldsymbol{\beta}}^\top, \hat{\lambda}, \hat{\alpha}, \hat{\boldsymbol{\gamma}}^\top\right)^\top$ *maximize the complete likelihood function* $L_c(\hat{\boldsymbol{\psi}}|\mathcal{D}_c)$ *over the parameter set*

$$\left\{(\boldsymbol{\beta}, \lambda, \alpha, \boldsymbol{\gamma}) : \text{ where } \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\gamma} \in \mathbb{R}^{q+1}, \alpha \in \mathbb{R}^+ \text{ and } \lambda \in \mathbb{R}^+\right\}.$$

*Proof of Theorem 1*: Assume the absence of the ML issue in the data set. Under the parametric specification of both parts of the model, the existence and uniqueness of the $\hat{\boldsymbol{\psi}}$ can be obtained following the compactness and bounded assumptions in C1–C3. One can

also observed that $H(t|\cdot)$ is a non-decreasing and right-continuous bounded function with jumps in all uncensored times. As a result, the complete likelihood function $L_c(\boldsymbol{\psi}|\mathcal{D}_c)$ is continuous in the parameter space, and has a maximum in the compact subspace $\mathbb{R}^p \times \mathbb{R}^{q+1} \times \mathbb{R}^+ \times \mathbb{R}^+$. Using the continuity assumption of $L_c(\boldsymbol{\psi}|\mathcal{D}_c)$ in the parameter space, following the Helley's theorem (Ash and Doleans-Dade, 2000) and considering the compactness assumption of $\mathcal{B}$, the MLE $\hat{\boldsymbol{\psi}}$ exists and it is unique; see Murphy et al. (1997) and Fang et al. (2005) for a more details. The uniqueness of $\hat{\boldsymbol{\psi}}$ and other technical aspects can also be found in Scharfstein et al. (1998) and Kosorok et al. (2004).

**Theorem 2** (Consistency)**.** *Suppose that conditions C1–C4 are satisfied, then* $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| \to 0$, $||\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0|| \to 0$, $||\hat{\lambda} - \lambda_0|| \to 0$ *and* $||\hat{\alpha} - \alpha_0|| \to 0$ *almost surely, where* $||\cdot||$ *is the Euclidean norm.*

*Proof of Theorem 2*: Suppose the existence of $\tilde{\boldsymbol{\psi}}$ such that, the subsequence $\left\{\hat{\boldsymbol{\psi}}_{nk}\right\}$ of $\left\{\hat{\boldsymbol{\psi}}_n\right\}$ converges in probability to $\tilde{\boldsymbol{\psi}}$ as $n \to \infty$. Then, under the parametric specification, it can be shown by induction that the difference $\left\{L_{nc}(\tilde{\boldsymbol{\psi}}) - L_{nc}(\boldsymbol{\psi}_0)\right\} \geq 0$ converges to the negative of the Kullback-Leibler divergence between $\tilde{\boldsymbol{\psi}}$ and $\boldsymbol{\psi}_0$ as $n \to \infty$. This implies that $\tilde{\boldsymbol{\psi}} = \boldsymbol{\psi}_0$ with probability one. Consequently, the subsequence $\left\{\hat{\boldsymbol{\psi}}_{nk}\right\}$ must converge to $\boldsymbol{\psi}_0$. Using the Helley's theorem (Ash and Doleans-Dade, 2000) and under the assumption of model identifiability, $\hat{\boldsymbol{\psi}}$ must converges to $\boldsymbol{\psi}_0$ in probability. The term $L_{nc}(\tilde{\boldsymbol{\psi}})$ resembles the complete log-likelihood function in Equation (3.8). Additional details can be found in Murphy et al. (1997) and Lu (2010).

**Theorem 3** (Asymptotic normality)**.** *Assume that the conditions C1–C4 hold. Then, the asymptotic distribution of* $\hat{\boldsymbol{\psi}}$ *is such that* $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0, \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0, \hat{\lambda} - \lambda_0, \hat{\alpha} - \alpha_0\right) \xrightarrow{d} N\left(\boldsymbol{0}, \boldsymbol{J}_{obs}^{-1}(\boldsymbol{\psi}_0)\right)$, *where* $\boldsymbol{J}_{obs}(\cdot)$ *is the observed information matrix mentioned in Section 4.2.*

*Proof of Theorem 3*: The Theorem 2 provides the foundations for inferences about $\boldsymbol{\psi}_0$. Under the conditions C1–C4, and using the results of the Theorem 1, the asymptotic normality of $\hat{\boldsymbol{\psi}}$ can be obtained by setting $\mathbf{U}_n(\boldsymbol{\psi}) = \partial \log L_{nc}(\boldsymbol{\psi})/\partial\boldsymbol{\psi}$ as the operator of the score function and $\mathbf{U}(\boldsymbol{\psi}_0) = \mathbb{E}\left(\mathbf{U}_n(\boldsymbol{\psi}_0)\right)$ the corresponding true value. Denote by $\mathcal{C} = \{\mathbf{c}_\beta, c_\lambda, c_\alpha, \mathbf{c}_\gamma\}$ the $d$-dimensional vector, where $\mathbf{c}_\beta$ and $\mathbf{c}_\gamma$ are the $p$- and $(q+1)$-

dimensional random vectors, $c_\lambda$ and $c_\alpha$ are scalar quantities. For $k = 1, \cdots, d$, we have that $c_{\psi_k} = \max\left(|\hat{\psi}_k|, 1\right) \text{sign}\left(\hat{\psi}_k\right)/\sqrt{n}$ converges in probability to 0, as $n \to \infty$. Let $\sigma = (\sigma_\beta, \sigma_\lambda, \sigma_\alpha, \sigma_\gamma)$ be a consistent linear operator and continuously invertible on the compact space $\mathbb{R}^p \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{R}^{q+1}$, with $\sigma^{-1} = \left(\sigma_\beta^{-1}, \sigma_\lambda^{-1}, \sigma_\alpha^{-1}, \sigma_\gamma^{-1}\right)$. Assume that $\mathbf{f}_r$, $\mathbf{e}_s$ and $(a_1, a_2)$ are $p$-, $(q+1)$- and 2-dimensional binary vectors, respectively. In addition, consider that $\mathbf{J}_\beta = \left(\sigma_\beta^{-1}(\mathbf{f}_1, 0, 0, \mathbf{0}), \cdots, \sigma_\beta^{-1}(\mathbf{f}_p, 0, 0, \mathbf{0})\right)$, $\mathbf{J}_\gamma = \left(\sigma_\gamma^{-1}(\mathbf{0}, 0, 0, \mathbf{e}_0), \ldots, \sigma_\gamma^{-1}(\mathbf{0}, 0, 0, \mathbf{e}_q)\right)$ and $\mathbf{J}_{\lambda\alpha} = \left(\sigma_\lambda^{-1}(\mathbf{0}, a_1, 0, \mathbf{0}), \sigma_\alpha^{-1}(\mathbf{0}, 0, a_2, \mathbf{0})\right)$ are submatrices of $\mathbf{J}_{obs}(\cdot)$. Based on Theorem 2.2 in Murphy et al. (1997), we have that

$$\begin{aligned}
&\sqrt{n}\left(\mathbf{c}_\beta^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \mathbf{c}_\gamma^\top(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) + c_\lambda\left(\hat{\lambda} - \lambda_0\right) + c_\alpha\left(\hat{\alpha} - \alpha_0\right)\right) \\
&= \sqrt{n}\left(\mathbf{U}_n(\boldsymbol{\psi}_0)[\sigma^{-1}(\mathcal{C})] - \mathbf{U}(\boldsymbol{\psi}_0)[\sigma^{-1}(\mathcal{C})]\right) + o_p(1). \qquad (4.8)
\end{aligned}$$

This converges to 0 in probability uniformly in the set $\mathcal{C}$. From Equation (4.8) and the central limit theorem, assuming $\mathbf{c}_\gamma = \mathbf{0}$, $\mathbf{c}_\lambda = 0$ and $\mathbf{c}_\alpha = 0$ then, we can write $\sqrt{n}\mathbf{c}_\beta^\top\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) = \sqrt{n}\left(\mathbf{U}_n(\boldsymbol{\beta}_0)[\sigma^{-1}\left(\mathbf{c}_\beta, 0, 0, \mathbf{0}\right)] - \mathbf{U}(\boldsymbol{\beta}_0)[\sigma^{-1}(\mathbf{c}_\beta, \mathbf{0}, 0, 0)]\right) + o_p(1)$. In addition, one can say that $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{J}_\beta$. Similarly, assume that $\mathbf{c}_\beta = \mathbf{0}$, $\mathbf{c}_\lambda = 0$ and $\mathbf{c}_\alpha = 0$, then $\sqrt{n}\left(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{J}_\gamma\right)$. In order to complete the proof, assume that $\mathbf{c}_\beta = \mathbf{0}$, $\mathbf{c}_\gamma = \mathbf{0}$, then
$\sqrt{n}\left(\left(\hat{\lambda}, \hat{\alpha}\right) - (\lambda_0, \alpha_0)\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{J}_{\lambda\alpha}\right)$. Again, Murphy et al. (1997) and Fang et al. (2005) are key references for the complete proof.

In order to incorporate the semiparametric approach, the previously defined theorems 1-3, the conditions C1–C4, and the corresponding assumptions A1–A4, might be adapted in order to accommodate the unknown cumulative hazard function $\tilde{H}\left(t|\cdot\right) = H\left(t_i|\cdot\right)$, for $i \in \{1, 2, \cdots, n\}$.Again, this function is strictly non-decreasing, continuous and differentiable, with positive jumps at all uncensored times $t_{(1)} < t_{(2)} < \cdots t_{(d^*)}$ and $\tilde{H}\left(t|\cdot\right) < \infty$, such that the set of the MLEs are now defined as $\left(\hat{S}_0, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$. Recall that the MLE for the survival function $\hat{S}_0$ is given in (3.19). The existence and finiteness of the MLEs, in the absence of ML issue, was proved in Fang et al. (2005) and Lu (2007).

The asymptotic normality can be obtained by taking into account the score equa-

tion and the Fisher information analogues of the parametric mixture cure model described in the Theorems 2 and 3. The consistency and asymptotic normality for $\hat{S}_0$ is also discussed in Murphy et al. (1994), Fang et al. (2005), Lu (2007), Lu (2010) and Barui and Grace (2020), only to mention few works. The theoretical properties of the penalized estimates are the same as those previously described for the original MLE $\hat{\boldsymbol{\psi}}$. These asymptotic results are important for the purpose of inference.

## 4.4 Brief Summary of the Chapter

This chapter discussed the general approach of the Firth modified score function, the standard error estimation based on the Louis method, and some asymptotic properties. In short, because the Firth approach does not depend directly on the original MLE in its application, this method is especially recommended in researches involving rare events where it is common to obtain small samples with strongly unbalanced dichotomous covariates. These aspects imply the existence of a non-zero probability of obtaining infinite estimates. In addition, another interesting aspect regarding the implicit methods concerns the fact that the penalized MLEs share the same information matrix and asymptotic properties as the usual MLEs. The variant discussed in this chapter to compute the observed information matrix is a viable alternative, since it does not require obtaining the distribution of the latent variable conditioned to the observed data. In the next chapter, we propose to investigate other penalty functions (prior distributions) in order to deal with the ML issue.

# Chapter 5

# BAYESIAN ANALYSIS

The Bayesian penalization procedure for the target mixture cure model has some advantages over the frequentist (or classical) methods, especially in terms of precision of the estimates. Additional benefits in using this estimation method are: $(i)$ it allows us to make exact inference about the unknown vector of parameters $\boldsymbol{\psi}$, without requiring any asymptotic assumption about estimates; $(ii)$ two sources of information as combined (prior and likelihood) in the inference procedure; $(iii)$ the sampling methods for Bayesian estimation have a more natural interpretation of parameter uncertainty, and $(iv)$ it allows a model comparison across non-nested alternatives, and recent sampling estimation developments have facilitated new methods of model choice in contrast with the frequentist techniques (Bayarri and Berger, 2004; Chib and Jeliazkov, 2005).

Another feature that differentiates these two approaches is that under the frequentist vision the sample data $\mathcal{D}_n$, defined in Section 2.2.1, are taken as random while the $d$-dimensional parameter set $\boldsymbol{\psi}$ are taken as fixed. In contrast, the Bayesian framework assumes that the parameter set is a random variable, following a probability distribution $p(\boldsymbol{\psi})$ which is commonly referred to as the prior. The knowledge about anything unknown can be expressed probabilistically through the prior distribution. The Bayesian methodology is especially attractive for survival analysis due to its natural treatment of censoring and truncation schemes, and the probabilistic quantification of relevant com-

ponents of the model structure without using asymptotic assumptions (Ibrahim et al., 2001; Sinha et al., 2003).

The prior distribution expresses the uncertainty about $\boldsymbol{\psi}$ before observing the data, and brings extra information/additional knowledge about the parameter. A natural question in Bayesian inference is: how to get the prior knowledge about $\boldsymbol{\psi}$ before the data are collected? In order to answer this question, we recommend the following references Raiffa and Schlaifer (1961), Bernardo and Smith (2000), Paulino et al. (2003), Lambert et al. (2005), Gelman (2008) and Migon et al. (2014), which present a general discussion regarding prior elicitation.

A basic division of the possible choices of prior distributions in terms of information level is: non-informative (also known as reference prior), weakly informative (or vague) and informative. The so-called informative prior contains a substantive amount information about the vector of parameters $\boldsymbol{\psi}$. When assuming a non-informative prior, we are intended to let the data dominate the posterior distribution thus, if contains no information about $\boldsymbol{\psi}$. In addition, the likelihood function plays a very important role in the Bayesian inference, because it expresses the information about $\boldsymbol{\psi}$ coming from the data $\mathcal{D}_n$ based on the chosen statistical model. The prior and the likelihood are combined in the Bayesian inference through the Bayes' theorem.

The Bayes' rule provides a method to update the prior beliefs about the unknown quantity $\boldsymbol{\psi}$. This update is done by incorporating the information from the observed data in the likelihood function $L\left(\mathcal{D}_n|\boldsymbol{\psi}\right)$. The result of the Bayes' rule is the so-called posterior distribution $p\left(\boldsymbol{\psi}|\mathcal{D}_n\right)$. More specifically, the conditional probability distribution of $\boldsymbol{\psi}$, given the data, can be expressed in the following manner

$$p\left(\boldsymbol{\psi}|\mathcal{D}_n\right) = p^{-1}(\mathcal{D}_n) \times L\left(\boldsymbol{\psi}|\mathcal{D}_n\right) p(\boldsymbol{\psi}), \tag{5.1}$$

where $p\left(\mathcal{D}_n\right)$ is a normalizing constant (or Bayesian evidence), necessary to ensure that expression (5.1) integrates or sums out to one, according to the nature of the parametric space (continuous or discrete). Note that the expression in (5.1) can be seen as weighted distribution, with $p\left(\boldsymbol{\psi}\right)$ acting as the weighting function. The posterior distribution is

the basis for all inferential statements about any unknown parameter or quantity. Thus, expression (5.1) represents the full answer to the estimation problem, because it gives a complete description of our state of knowledge and our uncertainty about the values of the unknown parameters. The logarithm of the posterior distribution is

$$\log p\left(\boldsymbol{\psi}|\mathcal{D}_n\right) = \log L\left(\boldsymbol{\psi}|\mathcal{D}_n\right) + \log p\left(\boldsymbol{\psi}\right) + \log\left(p^{-1}\left(\mathcal{D}_n\right)\right), \tag{5.2}$$

which has a similar form as the penalized log-likelihood function under the Jeffreys prior (Jeffreys, 1946) presented in Equation (4.5) with $p(\boldsymbol{\psi}) = |\mathbf{I}(\boldsymbol{\psi})|^{1/2}$. As mentioned in Chapter 1, our goal is to investigate other penalty functions (seen as prior distributions) in order the lead with the ML issue in the standard mixture cure rate model. Studies exploring the Bayesian approach to overcome the ML issue under the non-mixture models can be easily found in the literature, even in the logistic (Greenland and Mansournia, 2015; Discacciati et al., 2015) or Cox regression model (Lin et al., 2013; Almeida et al., 2018).

In the study of the ML problem, the advantage of using the Bayesian estimation procedure over the frequentist is that it allow us to choose a different prior distributions to evaluate distinct information levels of the penalty term. From Figure 5.1, the use of the Bayesian approach can be interpreted in the following manner: if $L\left(\boldsymbol{\psi}|\mathcal{D}_n\right)$ is monotone with respect to $\psi_k$ (first panel), this behavior can change as a consequence of the impact of the prior distribution (acting as a penalty). This modification is seen through the posterior distribution $p\left(\psi_k|\mathcal{D}_n\right)$ (second panel). In addition, improper or vague priors tends to produce identical estimates to their obtained when using the non-penalized likelihood function, see Almeida et al. (2018).

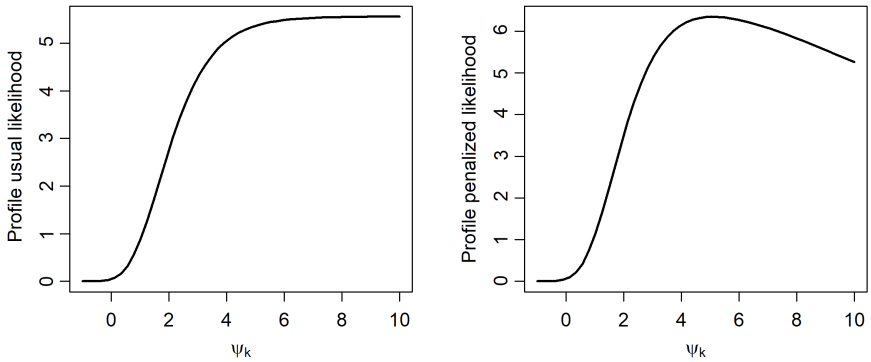


Figure 5.1: Behavior of the log-likelihood function. The usual profile likelihood (first panel) and penalized profile likelihood (second panel) under the Bayesian framework.

## 5.1 Prior Specifications

The prior elicitation is one of the most important step in the Bayesian framework. Two basic family of independent prior densities for the regression coefficients $\beta_u$ and $\gamma_s$ will be discussed here. The first is the normal prior, with log-density function given by $\log p\left(\beta_u|m_u,\nu_u\right) \propto -\left(\beta_u - m_u\right)^2/2\nu_u^2$ (for the latency part) and $\log p(\gamma_s|m_s,\nu_s) = -\left(\gamma_s - m_s\right)^2/2\nu_s^2$ (for the incidence part), with $u \in \{1,\cdots,p\}$, $s \in \{0,\cdots,q\}$, $p,q < d$.

The degree of flatness of the normal curve is controlled by the variance, where small values for $\nu_u^2$ or $\nu_s^2$ means that, the prior distributions concentrates probabilistic mass around the means ($m_u$ or $m_s$), respectively. As $\nu_u^2$ or $\nu_s^2$ increases, the prior becomes increasingly vague (high prior uncertainty). The second family is the log-$F$ proposed by Prentice (1975) and later extended in Greenland (2007) and Greenland and Mansournia (2015) to overcome the weaknesses encountered when using the Firth correction in situations involving the separation problem. In short, denote by $2r_1$ and $2r_2$ the degrees of freedom of the $F$-distribution, then $\log p(\psi_k|r_1,r_2) \propto r_1\psi_k - (r_1+r_2)\log\left(r_2 + r_1 e^{\psi_k}\right)$ denotes the logarithm of the kernel of the log-$F(r_1,r_2)$ distribution with $r_1$ and $r_2$ degrees of freedom (Dupuis, 2001). This distribution encompasses a variety of special cases including the Log-normal, Extreme-values, Weibull, Log-logistic, Generalized gamma, among others.

The unpenalized likelihood function might be obtained when using log-$F(0,0)$ in expression (5.2), meaning that the posterior estimate is equivalent to the usual MLE.

Likewise, Greenland and Mansournia (2015) argues that the log-$F(1,1)$ becomes a Firth's penalty under the logistic regression model. The mean and variance of the log-$F$ prior are given by $\mathbb{E}(\psi_k) = \Psi(r_1) - \Psi(r_2) + \log(r_2/r_1)$ and $\text{Var}(\psi_k) = \Psi^{'}(r_1) + \Psi^{'}(r_2)$, being $\Psi(\cdot)$ and $\Psi^{'}(\cdot)$ the digamma and trigamma functions, respectively. In this case, the symmetry and prior information level are controlled by the degrees of freedom. Thus, the log-$F$ prior becomes symmetric around zero, if $r_1 = r_2 = r^*$, which approaches the normality with large background information as $r^*$ increases, and weakly informative when assuming small values for $r^*$. If $r_1 < r_2$ the log-$F$ distribution becomes left-skewed, whereas the right-skewed prior might be obtained when $r_1 > r_2$.

In order to complete the prior specifications, we set the gamma prior for the scale and shape parameters of the Weibull distribution, with log-density proportional to the following expressions: $\log p\left(\lambda|a_\lambda, b_\lambda\right) = \left(a_\lambda - 1\right)\log\left(\lambda\right) - \lambda b_\lambda$ and $\log p\left(\alpha|a_\alpha, b_\alpha\right) = \left(a_\alpha - 1\right)\log\left(\alpha\right) - \alpha b_\alpha$, for scale and shape parameters, respectively. Note that the prior mean and variance are given by $a/b$ and $a/b^2$. Here, the information level is controlled by the prior hyperparameters $(a_\lambda, b_\lambda)$ and $(a_\alpha, b_\alpha)$.

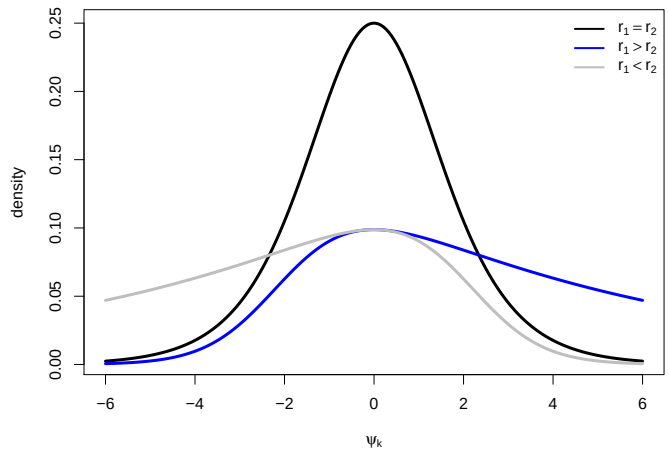


Figure 5.2: Different shapes of the log-$F(r_1, r_2)$ density function.

The main aim when performing the Bayesian inference is to summarize the information from the posterior distribution, see Paulino et al. (2003), Robert et al. (2010) and Givens and Mallick (2013) for more detailed approaches. A major limitation is that the calculation of the Bayes rule to obtain the posterior distribution often requires the

integration of a complex or high-dimensional function, which can be analytically difficult to handle. The Markov Chain Monte Carlo (MCMC) methods (Gamerman and Lopes, 2006), are widely used computational algorithms to allow indirect sampling from a posterior distribution only known up to the normalizing constant. The main feature that differentiates the MCMC methods from indirect sampling approaches is that it uses iterative simulation algorithms building a Markov chain. In other words, the generated values in iteration "$l$" are used to generate the new values in the iteration "$l+1$", which means that some level of dependence may exist between consecutive samples (Ehlers, 2007). In short, the algorithm is built around a stochastic process with stationary distribution being $p(\boldsymbol{\psi}|\mathcal{D}_n)$.

## 5.2 Markov Chain Monte Carlo

The general idea behind the MCMC methods is to transform the static problem (or intractable integrals) under consideration into a dynamic problem built with respect to a stochastic process that is easy to simulate, and that converges to the original distribution. This process is, in general, a homogeneous Markov chain whose equilibrium distribution is the posterior distribution (Paulino et al., 2003). Formally speaking, the MCMC methods provide the sampler that generates the sequence of observations that can be thought of as evolving over time towards the target distribution. This class of methods includes, for example, the Metropolis-Hastings (M-H) (Metropolis et al., 1953; Hastings, 1970) and the Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990).

The M-H and the Gibbs sampling algorithms are the most popular MCMC methods. The first one takes into account a proposal distribution to generate candidates and an accept-reject criteria to evaluate these candidates. The Gibbs sampler explores the full conditional posterior distributions to sample from the target joint posterior. In some studies, some of the full conditionals may not be available through the Bayes rule, then M-H steps can be included to sample within the Gibbs.

### 5.2.1 Metropolis-Hastings Algorithm

The M-H algorithm is a general term for a family of Markov chain simulation techniques that are useful for sampling from Bayesian posterior distributions. The method is an adaptation of a random walk with an acceptance/rejection rule to converge to the specified target distribution, and was originally introduced by Metropolis et al. (1953) and later generalized by Hastings (1970).

The main idea of the M-H algorithm is to generate a Markov chain $\left\{\boldsymbol{\psi}^{(l)}|l=0,1,\cdots,S\right\}$ such that its stationary distribution is the one given in (5.1). The algorithm must specify, for a given current state $\boldsymbol{\psi}^{(l-1)}$, how to generate the next state $\boldsymbol{\psi}^{(l)}$ based on two-stages. In the first stage, a candidate $\boldsymbol{\psi}^{prop}$ is generated from the proposal distribution $q\left(\cdot|\boldsymbol{\psi}^{prop}\right)$, which depends on the current state of the Markov chain $\boldsymbol{\psi}^{(l-1)}$. In the second stage, the proposed value can be accepted or rejected with probability

$$\alpha\left(\boldsymbol{\psi}^{(l-1)},\boldsymbol{\psi}^{prop}\right)=\min\left\{1,\frac{q\left(\boldsymbol{\psi}^{(l-1)}|\boldsymbol{\psi}^{prop}\right)p\left(\boldsymbol{\psi}^{prop}|\mathcal{D}_n\right)}{q\left(\boldsymbol{\psi}^{prop}|\boldsymbol{\psi}^{(l-1)}\right)p\left(\boldsymbol{\psi}^{(l-1)}|\mathcal{D}_n\right)}\right\}.$$

The general illustration of this sampling technique is in Algorithm 2.

**Algorithm 2** Metropolis-Hastings sampler

Start with any value $\boldsymbol{\psi}^{(0)} = \left(\psi_1^{(0)}, \psi_2^{(0)}, \cdots, \psi_d^{(0)}\right)^\top$ and set the chain counter as $l = 0$.

**for** Iteration $l = 1, 2, \cdots, S$ **do**

Sample $\boldsymbol{\psi}^{prop}$ from $q\left(\cdot|\boldsymbol{\psi}^{(l-1)}\right)$.

Acceptance Probability:

$\alpha\left(\boldsymbol{\psi}^{(l-1)}, \boldsymbol{\psi}^{prop}\right)$.

Generate $u \sim \mathcal{U}(0, 1)$.

**if** $u < \alpha\left(\boldsymbol{\psi}^{(l)}, \boldsymbol{\psi}^{(l-1)}\right)$ **then**

Accept the proposal: $\boldsymbol{\psi}^{(l)} \leftarrow \boldsymbol{\psi}^{prop}$.

**else**

Reject the proposal: $\boldsymbol{\psi}^{(l)} \leftarrow \boldsymbol{\psi}^{(l-1)}$.

**end if**

Increment the chain counter: $l = l + 1$.

**end for**

**output** $\left\{\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \cdots, \boldsymbol{\psi}^{(N_s)}\right\}$.

The choice of the proposal distribution is very flexible, but the generated chain must satisfy certain regularity conditions. In short, this proposal distribution must be chosen so that the chain covers the support of the stationary distribution, and guarantees that the produced candidate values are not accepted or rejected too frequently (Givens and Mallick, 2013). A typical choice is a distribution with symmetric arguments, such that $q\left(\boldsymbol{\psi}^{(l-1)}|\boldsymbol{\psi}^{prop}\right) = q\left(\boldsymbol{\psi}^{prop}|\boldsymbol{\psi}^{(l-1)}\right)$, which establishes the Metropolis algorithm (Metropolis et al., 1953). The "symmetric arguments" feature in the proposal distribution simplifies the algorithm, so that the acceptance probability in Algorithm 2 can be reexpressed as

$$\alpha\left(\boldsymbol{\psi}^{(l-1)}, \boldsymbol{\psi}^{prop}\right) = \min\left\{1, p\left(\boldsymbol{\psi}^{prop}|\mathcal{D}_n\right)/p\left(\boldsymbol{\psi}^{(l-1)}|\mathcal{D}_n\right)\right\}.$$

### 5.2.2 Gibbs Simpling Algorithm

The Gibbs sampler was introduced and named by Geman and Geman (1984) as an algorithm for simulating complex and high-dimensional multivariate distributions, which appear in image reconstruction problems. However, it is a general method that can be applied to a much wider class of distributions, such as sampling from the posterior

distribution (Gelfand and Smith, 1990).

The Gibbs sampler which is apparently the most used MCMC scheme to sampling from the posterior distribution, assumes that the full conditional distributions $p\left(\psi_k|\boldsymbol{\psi}_{-k}, \mathcal{D}_n\right)$ are available and completely known to be sampled. Here, $\boldsymbol{\psi}_{-k} = (\psi_1, \cdots, \psi_{k-1}, \psi_{k+1}, \cdots,$ represents all the components of $\boldsymbol{\psi}$, except the $k$-th element. The first step when using this algorithm is to analytically derive the posterior conditional distributions $p\left(\psi_k|\boldsymbol{\psi}_{-k}, \mathcal{D}_n\right)$ for each random quantity. Then, the posterior samples are obtained from the target joint posterior by iteratively sampling a value from the corresponding full conditional distribution, while all other quantities are fixed to their current values, such as presented in Algorithm 3.

**Algorithm 3** Gibbs sampler

Start with any value $\boldsymbol{\psi}^{(0)} = \left(\psi_1^{(0)}, \psi_2^{(0)}, \cdots, \psi_d^{(0)}\right)^\top$ and set the chain counter as $l = 0$.

**for** Iteration $l = 1, 2, \cdots, S$ **do**

Generate

$$\psi_1^{(1)} \sim p\left(\psi_1|\mathcal{D}_n, \psi_2^{(0)}, \cdots, \psi_d^{(0)}\right)$$
$$\psi_2^{(1)} \sim p\left(\psi_2|\mathcal{D}_n, \psi_1^{(1)}, \psi_3^{(0)} \cdots, \psi_d^{(0)}\right)$$
$$\psi_3^{(1)} \sim p\left(\psi_3|\mathcal{D}_n, \psi_1^{(1)}, \psi_2^{(1)}, \psi_4^{(0)} \cdots, \psi_d^{(0)}\right)$$
$$\vdots$$
$$\psi_d^{(1)} \sim p\left(\psi_d|\mathcal{D}_n, \psi_1^{(1)}, \psi_2^{(1)}, \cdots, \psi_{d-1}^{(1)}\right).$$

Increment the chain counter: $l = l + 1$.

**end for**

**output** $\left\{\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \cdots, \boldsymbol{\psi}^{(N_s)}\right\}$.

One curious feature of the Gibbs sampler, not shared by the M-H, is the idempotent property, i.e., the effect of multiple updates has the same effect of just one. This is because the update never changes the posterior distribution, hence the result of many repetitions of the same Gibbs update results in $\psi_k^{(l)}$, having the conditional distribution $p\left(\boldsymbol{\psi}|\mathcal{D}_n\right)$ just like the result of a single update. As the number of iterations increases, the chain approaches to its equilibrium condition, and the convergence is then assumed to hold approximately (Migon et al., 2014).

Despite to its simplicity in the application of the algorithm, as well as its wide applicability, there are several limitations related to this sampling algorithm. First, even if we have the full joint posterior distribution, in some situations, it may not be possible or practical to sampler directly from the full conditional distributions for each parameter in the model. Second, even if we have the full posterior conditionals for each scalar parameter $\psi_k$, it might be that they are not of a known form and, therefore, there is not a straightforward way to draw samples from them. Finally, there are cases in which the Gibbs sampler will not be very efficient, i.e., the mixing of the sampling chain might be very slow, meaning that the algorithm may spend a long time exploring a local region with high density, and thus take very long to explore the whole parameter space.

In some applications, persistent dependence of the chain on its starting values can seriously impact its performance. Therefore it may be sensible to omit some of the initial realizations (burn-in period) of the chain for purpose of the posterior inference. The burn-in period is an essential component of MCMC applications. In addition, many MCMC diagnostics methods were proposed in the literature. A general way to monitor the convergence of a Markov chain to the target distribution can be done through visual inspection of the traceplot graph displaying the trajectory of the chain. Additional theoretical details are also presented in Gelman et al. (2014).

### 5.2.3 Bayesian Interval Estimation

An useful method to summarize the posterior distribution is based on the credible interval (or credible set), which is the Bayesian analogue of the confidence intervals. Formally, the credible set for the posterior distribution is defined in the following manner. Denote $C_\epsilon$ as the subset of the parameter space of $\psi_k$, such that, the $100(1-\epsilon)\%$ credible interval meet the condition

$$P\left(\psi_k \in C_\epsilon | \mathcal{D}_n\right) = \int_{C_\epsilon} p(\psi_k|\mathcal{D}_n) d\psi_k = 1 - \epsilon. \tag{5.3}$$

In other words, the credible set for $\psi_k$ is the region $C_\epsilon = \left(Q_{\epsilon/2}; Q_{1-\epsilon/2}\right)$ with respect to $p(\psi_k|\mathcal{D}_n)$. This interval reflects the variation of the posterior estimate around

the posterior mean. It is important to note that, the credible intervals are not unique, i.e., we can easily define $C_\epsilon$ in different ways to cover varying parts of $\psi_k$ and still meet the probabilistic condition in (5.3). However, a reasonable choice is the one with minimum size (length, area or volume), resulting in the highest posterior density (HPD) region (Bernardo, 2005).

The $100(1-\epsilon)\%$ HPD region is the subset of the support of the posterior distribution for a particular parameter $\psi_k$ that meets the criteria: $C_\epsilon = \{\psi_k : p(\psi_k|\mathcal{D}_n) > Q\}$, being $Q$ the largest number such that:

$$P\left(\psi_k \in C_\epsilon|\mathcal{D}_n\right) = \int_{\psi_k:p(\psi_k|\mathcal{D}_n)>Q} p(\psi_k|\mathcal{D}_n)d\psi_k = 1-\epsilon. \tag{5.4}$$

When the posterior distribution is unimodal and symmetric, the HPD region coincides with the credibility interval, and have the same interpretation as discussed previously. Otherwise, if the posterior distribution is multimodal, the HPD may be a union of distinct intervals (or regions in the higher-dimensional case).

## 5.3 Brief Summary of the Chapter

In this chapter, we described some important elements of the Bayesian estimation, which includes the elicitation of some flexible prior distributions. The methods described in this chapter essentially aim to investigate the impact of different levels of prior information in terms of performance of the posterior estimates for situations involving the ML issue. Algorithms to allow indirect sampling from unknown posterior distributions and some elements of interval estimation in the Bayesian context were also described. The next chapter is presents simulation studies based on the methodology discussed in the thesis until this point.

# Chapter 6

# SIMULATION STUDY

The purpose of this study is to compare the finite sample performance of the proposed estimation methods, namely frequentist approach, under the Firth modified score function and Bayesian inference (based on different prior information levels). In this case, artificial data sets are generated, under the same conditions, from the logistic-Weibull mixture cure model, and then fitted using the proposed frameworks. We consider two independent covariates in the regression structure: $\mathbf{x}_i = (x_{1i}, x_{2i})^\top$ for $i = 1, 2, \cdots, n$. The first covariate is binary and configured to induce the ML issue in the simulated data. We set exactly the first five observed values in $\{x_{11}, x_{12}, \cdots, x_{1n}\}$ to be 1 and the remaining cases are considered to be 0. This same choice was also adopted in Almeida et al. (2018) to control the ML occurrence in the MC replications, for the situation involving the non-mixture model. The second covariate $\{x_{21}, x_{22}, \cdots, x_{2n}\}$ is randomly generated from the standard normal distribution for all $i$.

The survival times $T_i^*$ for the susceptible subjects are generated from the Weibull distribution with scale and shape parameters given by $\lambda_i^* = \lambda \exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)$ and $\alpha$, respectively. In this case, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$ and $\lambda = \exp(\beta_0)$, when using the parametric specification into the latency part, and $\lambda = 1$ for the semiparametric case. The censoring times $C_i$ are generated according to the specification of the conditional baseline hazard function. In addition, the censoring indicator for the susceptible group is $\omega_i^* = 1_{\{T_i^* \leq C_i\}}$,

then, for the overall population, the censoring indicator $\delta_i$ takes the value 0 for cured subjects ($Y_i = 0$) and $\omega_i^*$ otherwise. Note that because the cured individuals are among the censored, the survival times for the overall population is such that $T_i = \min\{T_i^*, C_i\}$ for susceptible individuals ($Y_i = 1$), and $T_i = C_i$ otherwise.

The susceptibility indicator $Y_i$ is generated from the Bernoulli distribution with the probability of success $\pi(\xi_i) = 1/[1 + \exp(-\xi_i)]$, which can also be seen as the proportion of susceptible subjects, with $\xi_i = \mathbf{z}_i^\top \boldsymbol{\gamma}$, and $\mathbf{z}_i = \left(\mathbf{1}, \mathbf{x}_i^\top\right)^\top$ being the covariates vectors affecting the incidence part, which is assumed to be completely equal to $\mathbf{x}_i$, apart from the column of 1's included to accommodate the intercept in the logistic regression model. According to the different specifications for the baseline hazard function described in Chapter 3, we consider two scenarios (SC's) for each case. The first one having a high proportion of samples with the ML issue ($SC1$:$\approx$ 75%), and the second having low proportion of samples with the mentioned problem ($SC2$:$\approx$ 25%). In both situations, the censoring and cured rates are fixed around 70% and 60%, respectively. Recall that in the melanoma data set, the censoring rate and cured proportion related to the mitosis risk factor (see Figure 1.1 given in Chapter 1) are $\approx 85\%$ and $\approx 70\%$, respectively.

These scenarios were obtained by setting a vector of possible values: raging between $(-5, 5)$ for the regression coefficients $\beta_r$ and $\gamma_s$, with $r = 1, 2$ and $s = 0, 1, 2$; and the interval $(0.01, 10)$ for $\lambda$ and $\alpha$ (scale and shape parameters of the Weibull distribution) and $\tau$ (parameter used to control the censoring rate). Using the regression structure with two covariates, as previously indicated, different choices of the mentioned vectors are established by choosing different combinations of values for the involved parameters. Next, 1,000 artificial data sets (sample size $n$ is fixed) are generated, and they are evaluated to see how often the level 1 of the dichotomous covariate $x_{1i}$ is not related to any failure (leading to the ML issue). For each data set, the analysis also accounts for the number of times in which $z_{1i}$ perfectly predicts the binary response (cure indicator); this means "level 1" of the response being exclusively related to the same category of $z_{1i}$ (separation problem). In addition, the censoring rate $\%\{\delta_i = 0\}$ and the proportion of the cured subjects $\%\{Y_i = 0\}$ are determined for each artificial data set. Results are

summarized through the quantities: proportion of samples with the ML issue; proportion of samples with the separation problem; average censoring rate based on 1,000 values; average cured proportion based on 1,000 values. Then, the resulting parameters configurations are presented separately according to the distribution specified for the latency part, for both scenarios.

($i$) **Logistic – Weibull mixture cure model:** when using the parametric specification for the latency part, we control the censoring rate and cured proportion in the MC replications by generating the censoring times $C_i$ from $\mathcal{U}(0, \tau)$. Consider the following configurations:

- SC1: $\boldsymbol{\beta} = (-2.00, 1.56)^\top$, $\boldsymbol{\gamma} = (-0.31, -2.00, 0.31)^\top$ and $\tau = 3.00$.
- SC2: $\boldsymbol{\beta} = (-2.00, 2.20)^\top$, $\boldsymbol{\gamma} = (-0.73, -1.00, 2.20)^\top$ and $\tau = 2.64$.

The scale and shape parameters are the same in both scenarios, taking the values $\lambda = 8.79$ and $\alpha = 7.57$, respectively.

($ii$) **Logistic – Semiparametric mixture cure model:** under this specification for the lifetime distribution, we control the censoring rate and cured proportion in the MC replications by generating the censored times from the exponential distribution with mean $1/\tau$. Once again, the quantity $\tau$ will be chosen in order to obtain the following desirable scenarios:

- SC1: $\boldsymbol{\beta} = (-1.00, 1.56)^\top$, $\boldsymbol{\gamma} = (-0.67, -2.00, -0.67)^\top$ and $\alpha = 6.00$.
- SC2: $\boldsymbol{\beta} = (-1.00, 1.00)^\top$, $\boldsymbol{\gamma} = (-1.00, -1.00, 2.00)^\top$ and $\alpha = 4.88$.

In this case, the ML/SP issue, censoring rate and the proportion of cure individuals are controlled by setting the quantity $\tau = 0.10$, for both scenarios.

These MC simulation studies are based on the analysis of 1,000 samples generated in the MC scheme. For each particular scenario, we consider the following sample sizes $n$: 50, 200, 600 and 1,000. Some performance measurements to be considered are: root of the mean square error (RMSE), percentage of relative bias (%RB) and coverage probability (CP). Additionally, we investigate the MC standard errors $\hat{\sigma}_{mc}$ in a com-

parison with the asymptotic standard errors $\hat{\sigma}_{as}$ estimated through the Louis method (see again Section 4.2) for the frequentist approach. The coverage probability indicates the proportion of the 95% Wald-type confidence intervals $\hat{\psi}_k \pm 1.96 \times SE\left(\hat{\psi}_k\right)$, or the 95% HPD regions obtained for each MC replication, that contain the true value of the parameter being estimated. Consider $\text{RMSE}_{\hat{\boldsymbol{\psi}}^*} = \left[\sum_{b=1}^{R}(\hat{\boldsymbol{\psi}}_b^* - \boldsymbol{\psi})^2/R\right]^{1/2}$ and $\%\text{RB}_{\hat{\boldsymbol{\psi}}^*} = (100/R)\sum_{b=1}^{R}\left(\hat{\boldsymbol{\psi}}_b^* - \boldsymbol{\psi}\right)/|\boldsymbol{\psi}|$, where $\hat{\boldsymbol{\psi}}^*$ is the penalized MLE or the posterior estimate for $\boldsymbol{\psi}$ and $R$ is the number of MC replications. The Monte Carlo standard error (MCSE) is obtained by $\hat{\sigma}_{mc} = \left[\sum_{b=1}^{R}\left(\hat{\boldsymbol{\psi}}_b^* - \hat{\boldsymbol{\psi}}_{mc}^*\right)^2/R\right]^{1/2}$, with $\hat{\boldsymbol{\psi}}_{mc}^* = \sum_{b=1}^{R}\hat{\boldsymbol{\psi}}_b^*/R$. This measure is an indication of how much error is related to the estimate due to the fact that Monte Carlo method is used.

## 6.1 Computational Methods

Under the frequentist approach, the iterative NR procedure was carried out by setting the maximum number of iteration at 50 and a tolerance level equal to $\zeta$=$10^{-5}$. Specific information about the convergence of the estimation algorithm can be found in Tables 6.1 and 6.2. The mean central processing unit (CPU) times are the average of the computational times obtained in the 1,000 MC replications by using the `proc.time()` function in `R`. In order to ensure the model identification and reasonable estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we complete the upper tail of the baseline survival function given in equation (3.19) by assuming a smooth decay towards zero, starting from the largest uncensored time point. This fact will impose a proper survival function for the susceptible individuals.

MCMC methods are required to deal with the Bayesian framework in the context of mixture cure modeling. For computational reasons, we obtain the posterior marginal distributions by using the complete data likelihood function in Equation (3.8). Since the full posterior distributions for $\beta_r$, $\gamma_s$, $\lambda$ and $\alpha$ are only known up to their normalizing constants, the M-H algorithm is applied to allow the indirect sampling of these parameters (see Algorithm 2). Such as recommended in the literature, the acceptance rate were fixed around $40 - 50\%$; see Roberts et al. (1997) for more details. Formally speaking, there are theoretical arguments indicating that the optimal acceptance rate is 44% for one

dimensional parameter set, and has a limit of 23.4% as the dimension goes to infinity. In this case we have an additional step, where $Y_i$ will be treated as another parameter in the model, and its posterior distribution needs to be computed as well. That is, for censored cases, the quantity $Y_i$ follows the Bernoulli distribution with probability of success

$$P\left(Y_i = 1|\boldsymbol{\psi}, T_i > t_i, \delta_i = 0, \mathcal{D}_{obs}\right) = \frac{P\left(Y_i = 1|\mathbf{z}_i, \boldsymbol{\gamma}\right) S\left(t_i|\mathbf{x}_i, Y_i = 1, \boldsymbol{\theta}\right)}{P\left(Y_i = 0|\mathbf{z}_i, \boldsymbol{\gamma}\right) + P\left(Y_i = 1|\mathbf{z}_i, \boldsymbol{\gamma}\right) S\left(t_i|\mathbf{x}_i, Y_i = 1, \boldsymbol{\theta}\right)}.$$

For each MC replication 6,000 iterations were considered, in which the first 1,000 are assumed to be the burn-in period. The algorithm convergence was monitored by using the traceplots of the chains (for each parameter separately). Additionally, we summarize our point estimates as the posterior mean by calculating $\bar{\psi}_k = \sum_{l=1}^{S} \psi_k^{(l)}/S$, with $\psi_k^{(\cdot)} \in \left\{\psi_k^{(1)}, \cdots, \psi_k^{(S)}\right\}$, where $S$ is the number of posterior samples after the burn-in period. As a result, we have the vector of the posterior means $\left\{\bar{\psi}_k^{(1)}, \cdots, \bar{\psi}_k^{(R)}\right\}$. By the Central Limit Theorem, $\bar{\psi}_k|\mathcal{D}_c$ is normally distributed with mean $\psi_k = \mathbb{E}\left(\bar{\psi}_k|\mathcal{D}_c\right)$ and standard error $\sigma_{\bar{\psi}_k} = \left[\sum_{b=1}^{R}\left(\bar{\psi}_k^{(b)} - \bar{\psi}_k^{*}\right)^2/R\right]^{1/2}$, being $\bar{\psi}_k^{*} = \sum_{b=1}^{R}\bar{\psi}_k^{(b)}/R$, equivalent to the MC penalized MLE.

A sensitivity analysis was carried out to choose an optimal configuration for the prior hyperparameters that provided small relative bias. Because little information is known about the true value of the parameter $\psi_k$, the prior distribution for the regression coefficients were centered at 0. Additionally, in order to compare the estimated parameters performance, we also consider the normal prior centered at the true value of the unknown parameter. Thus, the following configurations of prior distributions are obtained:

- `Bayes_1`: $p(\beta_r) \sim N(0, 20)$ and $p(\gamma_s) \sim N(0, 10)$;
- `Bayes_2`: $p(\beta_r) \sim N(\beta_r^{true}, 1)$ and $p(\gamma_s) \sim N(\gamma_s^{true}, 1)$;
- `Bayes_3`: $p(\beta_r) \sim \log\text{-}F(0.3, 0.3)$ and $p(\gamma_s) \sim \log\text{-}F(0.6, 0.6)$.

Figure 6.1 illustrates the mentioned configurations of prior distributions according to their information levels. The prior specifications $\lambda$ and $\alpha$, which are related to the

model in $(i)$, are set to be $p(\alpha) \sim G(0.1, 0.1)$ and $p(\lambda) \sim G(0.1, 0.1)$, respectively. All simulations were developed in the same computer (Linux system) with Intel Core i5 processor and 8.00 GB memory, using the `R` programming language (R Core Team, 2021). The simulation results are presented in two parts, according to the specification related to the latency part.

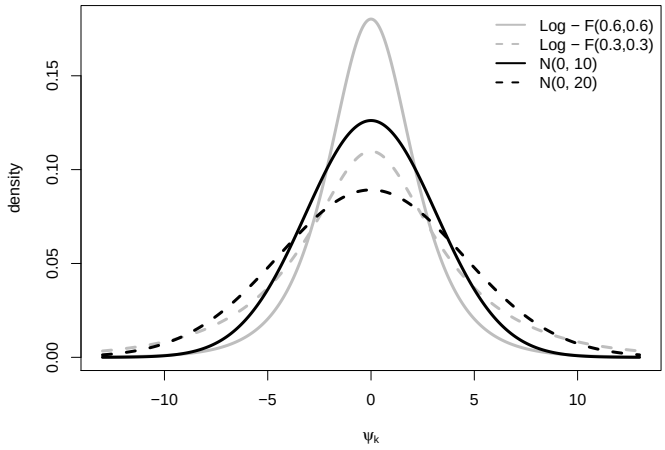


Figure 6.1: Comparing the shapes of the densities (Normal and log-$F$) chosen as prior distributions in the present thesis.

Tables 6.1 and 6.2 illustrate the proportion of convergence samples and the corresponding mean CPU times for the frequentist approach. The results suggest that, for the standard score function (i.e., non-penalized case), the presence of the ML issue in the artificial data sets strongly affects the convergence of the estimation algorithm. The proportions of samples for which the convergence status is achieved is quite low, ranging between $15.10 - 22.80\%$ (in $SC1$) and $58.00 - 66.90\%$ (in $SC2$). In these cases, the lack of convergence occurs for the samples impacted by the ML issue. In the modified score function situation, the convergence of the estimation algorithm was confirmed, on average, within the first 20 iterations. Note also that the percentages $\text{EM}_{cs}(\%)$ are quite high (near or equal to 100%) for all sizes $n$. In terms of computational time, the measurements are similar when comparing the corresponding cases from the standard and the modified score functions. In addition, as expected, the mean CPU times to run the algorithm increases as $n$ increases.

Table 6.1: Comparing the proportion of converged samples when using the standard and modified score functions, and the corresponding mean CPU time (average of the execution time on 1,000 MC replications, given in seconds). The quantity $\text{EM}_{\text{CS}}(\%)$ denotes the proportion of times in which the EM algorithm converged within the 1,000 generated data sets. The parametric mixture cure rate is considered here.

| **Standard score functions** | | | | |
|---|---|---|---|---|
| | Scenario 1 | | Scenario 2 | |
| n | $\text{EM}_{\text{CS}}(\%)$ | Mean CPU times | $\text{EM}_{\text{CS}}(\%)$ | Mean CPU times |
| 50 | 19.800 | 0.043 | 58.600 | 0.032 |
| 200 | 18.100 | 0.043 | 56.500 | 0.058 |
| 600 | 20.000 | 0.227 | 54.300 | 0.209 |
| 1,000 | 21.200 | 0.554 | 50.400 | 0.544 |
| **Modified score functions** | | | | |
| | Scenario 1 | | Scenario 2 | |
| n | $\text{EM}_{\text{CS}}(\%)$ | Mean CPU times | $\text{EM}_{\text{CS}}(\%)$ | Mean CPU times |
| 50 | 98.700 | 0.090 | 95.500 | 0.072 |
| 200 | 99.900 | 0.155 | 99.100 | 0.104 |
| 600 | 100.00 | 0.562 | 99.100 | 0.425 |
| 1,000 | 100.00 | 1.775 | 99.700 | 1.681 |

Table 6.2: Comparing the proportion of converged samples when using the standard and modified score functions, and the corresponding mean CPU times (in seconds). The quantity $\text{EM}_{\text{CS}}(\%)$ denotes the proportion of times in which the EM algorithm converged within the 1,000 generated data sets. The semiparametric mixture cure rate is considered here.

| **Standard score functions** | | | | |
|---|---|---|---|---|
| | Scenario 1 | | Scenario 2 | |
| n | $\text{EM}_{\text{CS}}(\%)$ | Mean CPU times | $\text{EM}_{\text{CS}}(\%)$ | Mean CPU times |
| 50 | 15.100 | 0.925 | 58.000 | 0.875 |
| 200 | 19.600 | 4.505 | 59.700 | 4.147 |
| 600 | 19.600 | 24.325 | 63.700 | 18.749 |
| 1,000 | 22.800 | 54.889 | 66.900 | 42.451 |
| **Modified score functions** | | | | |
| | Scenario 1 | | Scenario 2 | |
| n | $\text{EM}_{\text{CS}}(\%)$ | Mean CPU times | $\text{EM}_{\text{CS}}(\%)$ | Mean CPU times |
| 50 | 99.800 | 0.907 | 99.800 | 0.848 |
| 200 | 99.900 | 4.813 | 100.00 | 4.213 |
| 600 | 100.00 | 24.183 | 100.00 | 19.476 |
| 1,000 | 99.900 | 55.877 | 100.00 | 42.228 |

As mentioned earlier, the convergence of the generated MCMC chains were monitored by a simple inspection of the traceplots and autocorrelation functions (ACF) graphs given in Figures C.1.1 to C.1.4 and Figures C.2.1 to C.2.4 of the Appendix C. Results suggests that the convergence is achieved. The general discussion about the MC simulation results are given in next two sections with respect to the specified distribution of the survival times in the susceptible group.

## 6.2 Discussion of Results (Parametric Model)

Figures 6.2 to 6.7 show the behavior of the performance measurements previously mentioned for the scenarios ($SC1$ and $SC2$). The results obtained when using the frequentist approach (under the Firth correction) and Bayesian methods are confronted. In this simulation study, three flexible prior configurations based on different information levels were considered: two based on the normal distribution named as (`Bayes_1` and `Bayes_2`), and one based on the log-$F$ distribution. The values for the MC penalized MLEs and the corresponding posterior mean are summarized in Tables C.1.1 to C.1.5 and Tables C.2.1 to C.2.4 given in Appendix C.

The analysis indicates that a high relative bias and high root of mean square error are associated with the regression coefficients $\beta_1$ and $\gamma_1$ corresponding to the dichotomous covariate (which is the source of the ML/SP issue), especially when using the Firth method. For example, the RB's for $\hat{\beta}_1^*$ and $\hat{\gamma}_1^*$ varies between 43.53 to 49.93% and 83.88 to 86.12% (for $SC1$), 32.50 to 41.50% and 69.13 to 74.84% (for $SC2$), when using the Firth modified score functions. In terms of the magnitude, the RB's for both coefficients decreases when using the Bayesian framework, under the `Bayes_2`, varying around -0.43 to 0.21% and -4.27 to -3.54% (for $SC1$), -0.59 to 1.90% and -4.11 to 0.40% (for $SC2$), respectively. In addition, the magnitude of the RB's obtained when using the `Bayes_3` setting reduces when compared to that obtained in the frequentist case, in both scenarios. Despite this, the relative bias does not decrease, and remains stable at a certain level, even when the sample size increases. This occurs as a consequence of the high imbalance of the binary covariate.

The results obtained for the `Bayes_1` prior setting are quite similar to those based on the Firth correction, for the regression coefficient $\beta_1$. In contrast, high performance in terms of bias reduction was observed for the estimate related to the coefficient $\gamma_1$, under the `Bayes_1` configuration. We can see the RB varying between -6.64 to -1.37% (for $SC1$) and -7.45 to 3.50% (for $SC2$). In both scenarios, the MCSE for the frequentist approach, are quite similar to those obtained in the Bayesian case, under the `Bayes_1` and `Bayes_3` for $\hat{\beta}_1^*$. The same settings indicates different performance on the coefficient $\hat{\gamma}_1^*$. However, under these conditions, large values of the RMSE were obtained for $\hat{\beta}_1^*$, ranging between 1.63 to 1.76 (for $SC1$) and 1.37 to 1.52 (for $SC2$). Additionally, the RMSE varies between 1.769 to 1.81 (for $SC1$) and 1.41 to 1.59 (for $SC2$) for $\hat{\gamma}_1^*$. Small values of MCSE and RMSE are related to the `Bayes_2` prior setting.

As expected, the presence of the ML/SP issue in the data sets affects directly the coverage probability, especially when using the modified score function, under the Firth method. In short, the CP for the coefficients $\beta_1$ and $\gamma_1$, are all underestimated, ranging from 72.60 to 78.40% and 78.40 to 90.40% (for $SC1$); 74.90 to 81.60% (for $SC2$), respectively. For different sample sizes, the $\text{CP}_{\gamma_1}$ is greater than the nominal level ($\approx 98\%$ on average) in $SC2$. The coverage probabilities obtained when using the settings `Bayes_1` and `Bayes_3`, are on average around the nominal value (for both $SC1$ and $SC2$), and are above the nominal, when using the `Bayes_2`. The poor performances of these coefficients, in terms of relative bias and coverage rate, especially in the frequentist case, are due to the high degree of imbalance in the binary covariate (very few values $x_{1i} = 1$). For example, depending on the sample size, we have the following percentages of $x_{1i} = 1$: 10% ($n = 50$), 2.5% ($n = 200$), 0.83% ($n = 600$) and the strongest unbalanced case with 0.50% ($n$ = 1,000). Again, the advantage of this setting is that it allows us to control the percentage of samples affected by the ML issue in the MC schemes.

The mentioned imbalance is behind the instability of the RB's, even in the Bayesian inference. Its important to emphasize that because the covariate vectors are set to be completely equal in both parts of the model, then the ML issue has a double impact in the estimation procedure. On the other hand, good performances are obtained when

using a balanced configuration for $x_1$ in both scenarios (see Figure 6.8). However, the disadvantage in using the balanced setting for our analyses is that many MC replications will not be affected by the target ML issue; this can be see in Figures 6.9. Under the frequentist approach, the authors in Kenne Pagui and Colosimo (2020) also illustrate how the estimators related to the binary covariate are strongly affected by the degree of imbalance, even in the absence of the ML issue.

Figures 6.10 and 6.11 show box-plots, based on the log-absolute ratio of the relative biases, for the parametric mixture cure rate models. Consider log-rRB=log $(|\text{RB}_u/\text{RB}_m|)$, where $\text{RB}_u$ and $\text{RB}_m$ are the relative biases obtained through the usual and the penalized method under the frequentist and Bayesian methods, respectively. The magnitude of the bias reduction is assessed by evaluating the distance between the median and the reference level (horizontal red line). The results related to the coefficients $\beta_1$ and $\gamma_1$ indicates that their $|\text{RB}_u|$ are greater than $|\text{RB}_m|$ for both estimation methods, since the corresponding medians are above the level 0. This is more evident for the coefficient $\beta_1$. The impact of ML issue in $\hat{\lambda}^*$ is clear, especially in small sample size ($n = 50$). The median of the box-plots are slightly above the threshold, and approaches to 0 when the sample size increases.

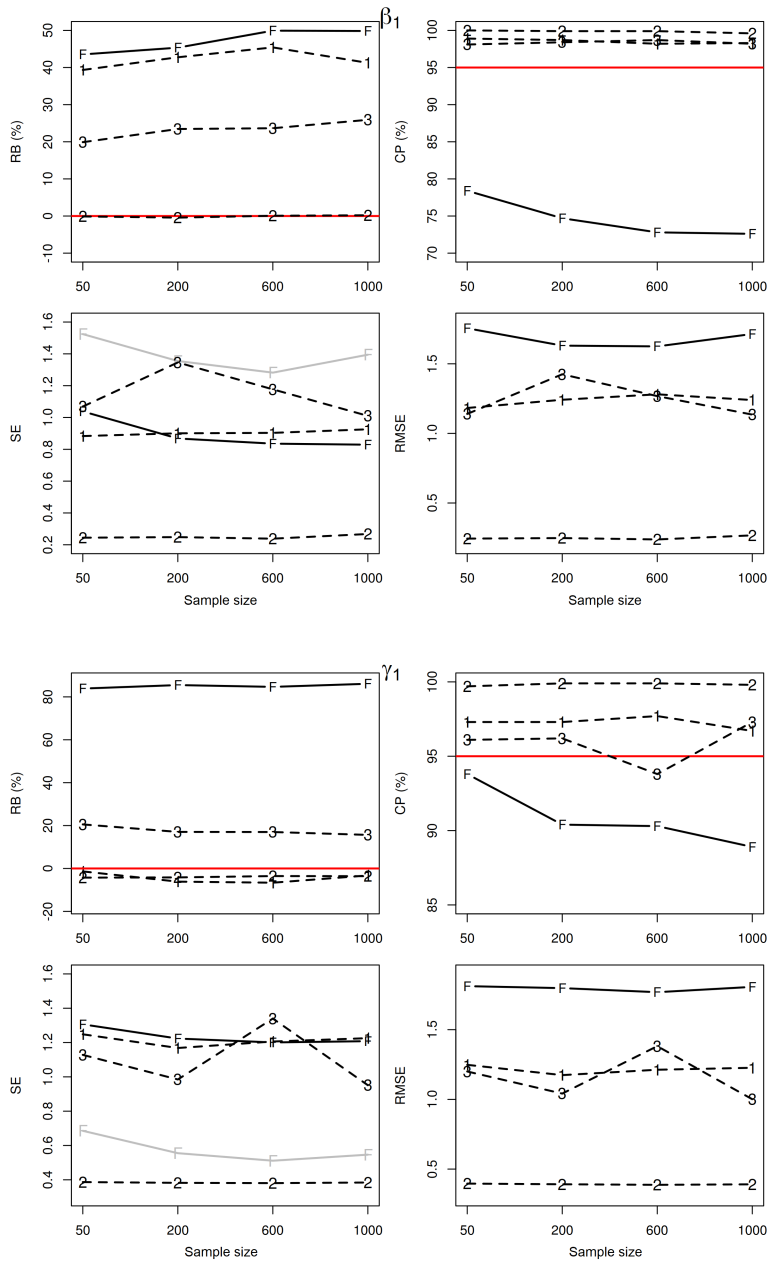


Figure 6.2: Monte Carlo simulation results for the regression coefficients related to the binary covariate $x_{1i}$, based on the $SC1$. The Firth modified score function and the Bayesian approach are applied here. The solid lines "`F`" denote the results obtained when using the Firth method, where the asymptotic and MCSE are denoted by the *black* and *grey solid lines*, respectively. The lines named as "`1`" denote the results obtained when using the `Bayes_1`, "`2`" for the `Bayes_2` and "`3`" for `Bayes_3`.

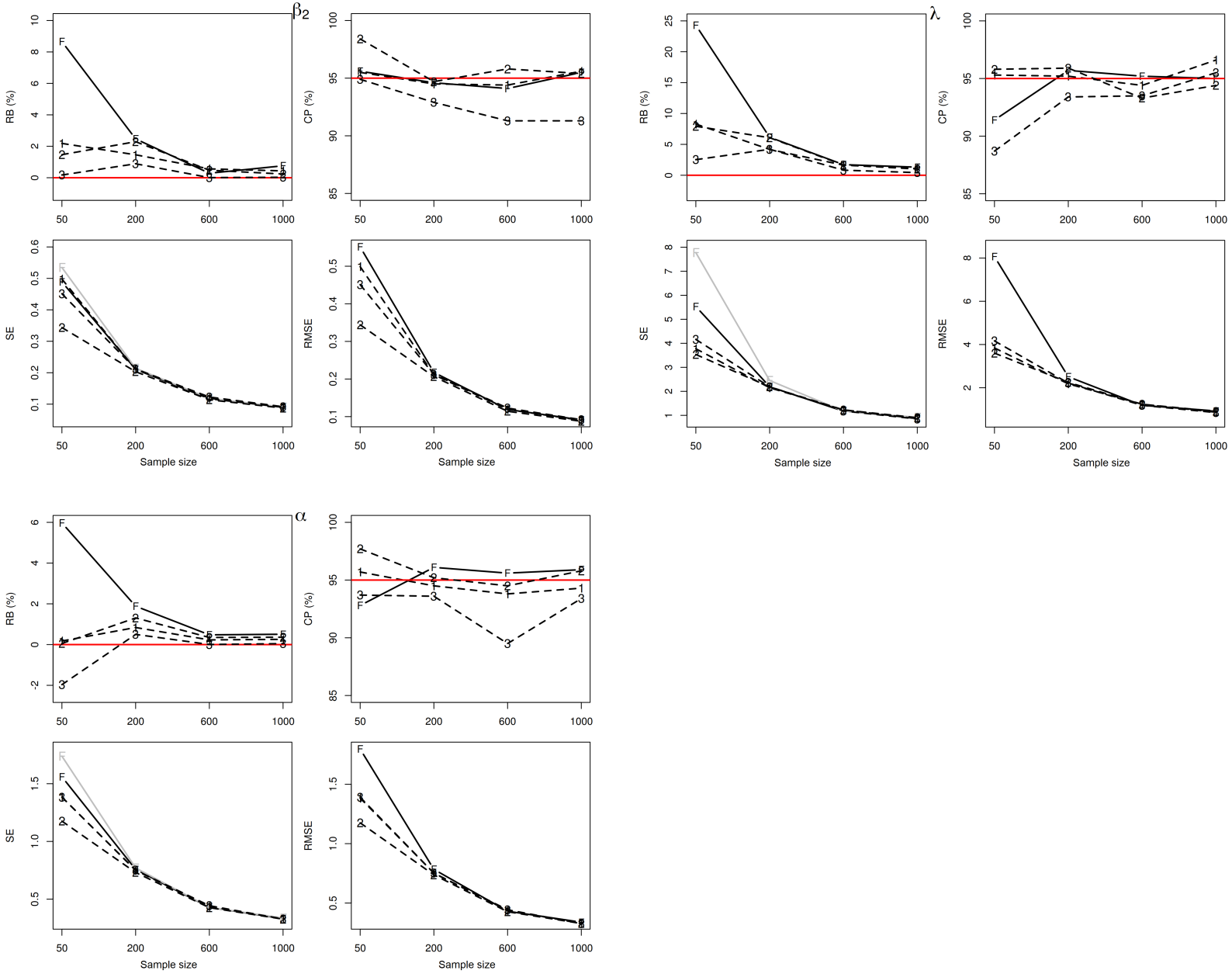


Figure 6.3: Monte Carlo simulation results for the parameters not directly related to the monotone likelihood for the latency distribution, based on the *SC*1. The Firth modified score function and the Bayesian approach are applied here. The solid lines "`F`" denote the results obtained when using the Firth method, where the asymptotic and MCSE are denoted by the *black solid lines* and *grey solid lines*, respectively. The lines named as "`1`" denote the results obtained when using the `Bayes_1`, "`2`" for the `Bayes_2` and "`3`" for `Bayes_3`.

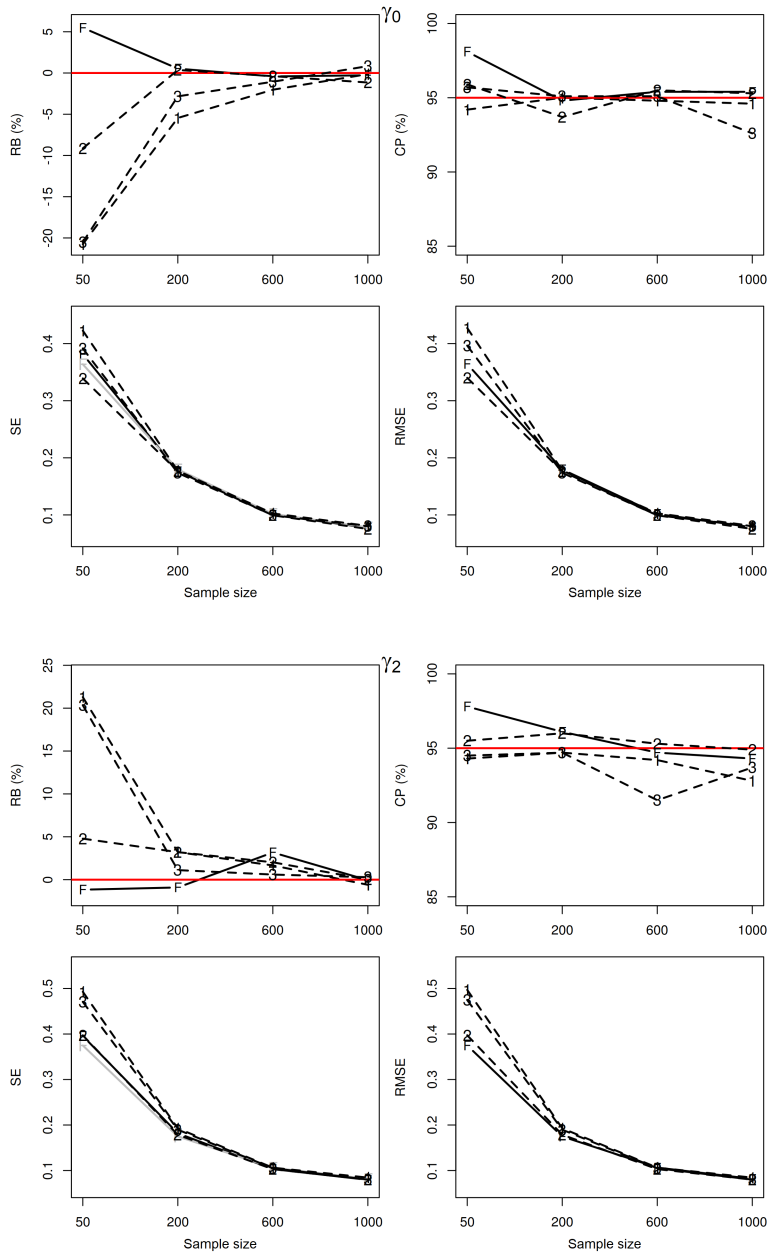


Figure 6.4: Monte Carlo simulation results for the regression coefficient not directly related to the monotone likelihood for the incidence distribution, based on the $SC1$. The Firth modified score function, and the Bayesian approach are applied here. The solid lines "`F`" denote the results obtained when using the Firth method, where the asymptotic and Monte Carlo standard errors are denoted by the *black solid lines* and *grey solid lines*, respectively. The lines named as "`1`" denote the results obtained when using the `Bayes_1`, "`2`" for the `Bayes_2` and "`3`" for `Bayes_3`.

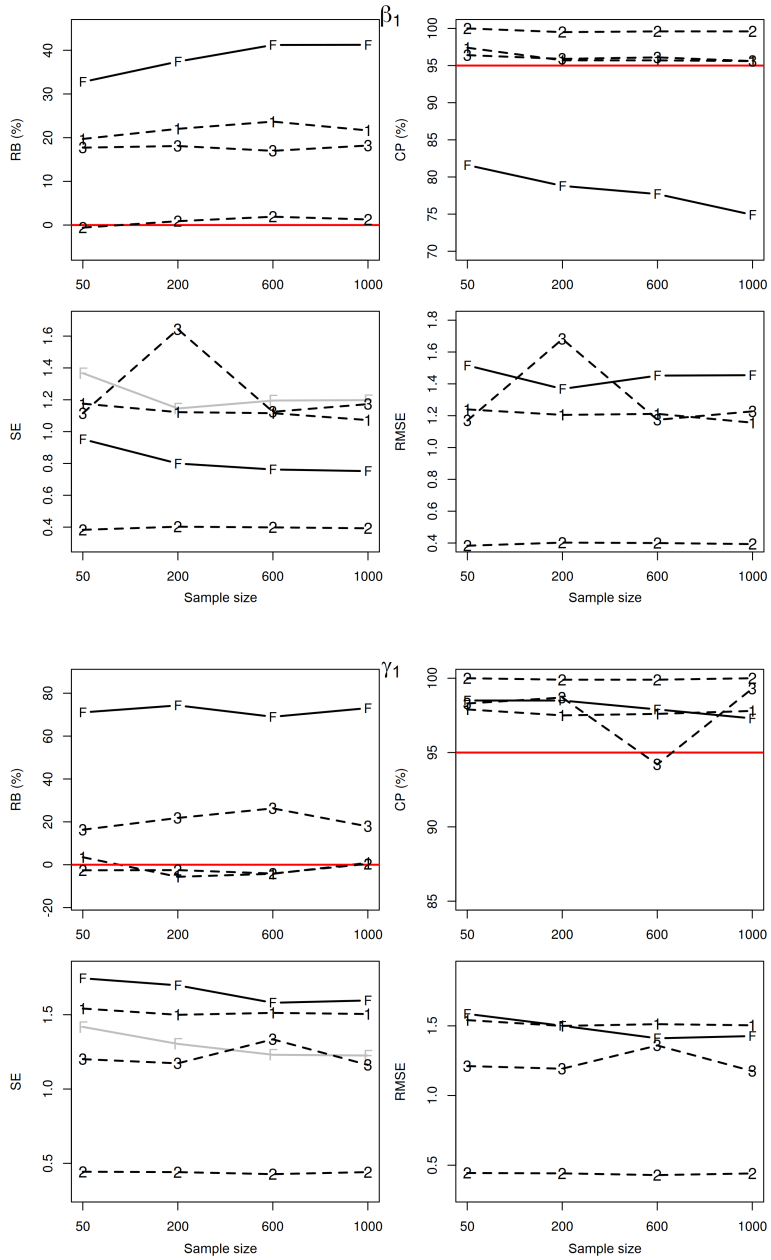


Figure 6.5: Monte Carlo simulation results for the regression coefficients related to the binary covariate $x_{1i}$, based on the $SC2$. The Firth modified score function, and the Bayesian approach are applied here. The solid lines "`F`" denote the results obtained when using the Firth method, where the asymptotic and MCSE are denoted by the *black* and *grey solid lines*, respectively. The lines named as "`1`" denote the results obtained when using the `Bayes_1`, "`2`" for the `Bayes_2` and "`3`" for `Bayes_3`.

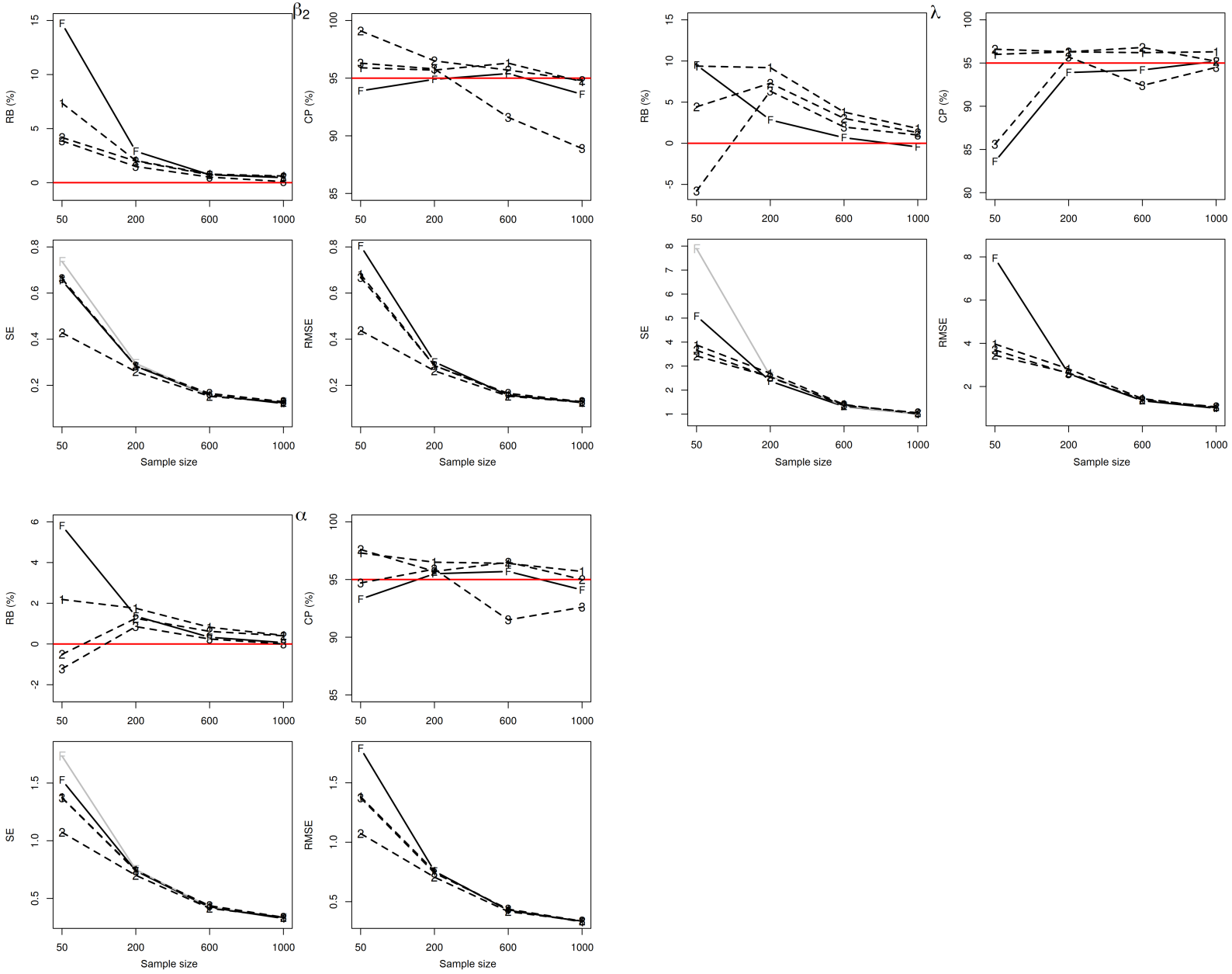


Figure 6.6: Monte Carlo simulation results for the parameters not directly related to the monotone likelihood issue for the latency distribution, based on the *SC*2. The Firth modified score function and the Bayesian approach are applied here. The solid lines "`F`" denote the results obtained when using the Firth method, where the asymptotic and Monte Carlo standard errors are denoted by the *black solid lines* and *grey solid lines*, respectively. The lines named as "`1`" denote the results obtained when using the `Bayes_1`, "`2`" for the `Bayes_2` and "`3`" for `Bayes_3`.

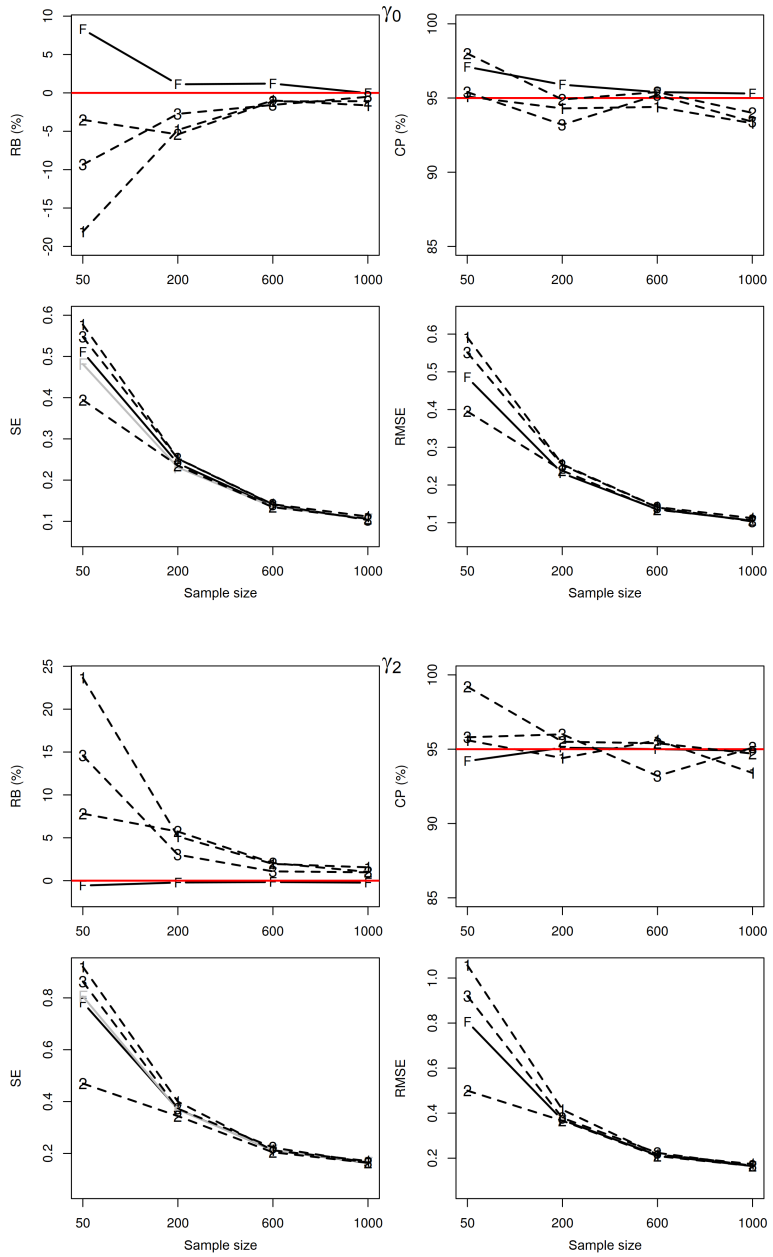


Figure 6.7: Monte Carlo simulation results for the regression coefficient not directly related to the monotone likelihood issue for the incidence distribution, based on the $SC2$. The Firth modified score function and the Bayesian approach are applied here. The solid lines "`F`" denote the results obtained when using the Firth method, where the asymptotic and Monte Carlo standard errors are denoted by the *black solid lines* and *grey solid lines*, respectively. The lines named as "`1`" denote the results obtained when using the `Bayes_1`, "`2`" for the `Bayes_2` and "`3`" for `Bayes_3`.

Note that all parameters not directly connected to the binary covariate( $\beta_2$, $\alpha$, $\gamma_0$ and $\gamma_2$) have a very small RB and RMSE for any sample size, i.e., they show a good

asymptotic properties in both scenarios, and for both estimation methods. The qualities of all performance measurements improves as the sample size increases. Under the Firth method, a small difference is observed between the MCSE and asymptotic standard errors. However, these two measurements are similar when the sample size increases. Furthermore, this similarity is also observed when comparing to the MCSE obtained in the Bayesian case, regardless the considered settings of the prior. In general, the CP are near the nominal level, manly when the sample size is large.

The parameter $\lambda$ have large for small sample size (in $SC1$), when using the modified score functions, where the $\text{RB}_{\hat{\lambda}^*}$ is 24.30% (for $n = 50$) and decreases when the sample size increases. The RMSE is approximately 8.07 (in $SC1$) and 7.94 (in $SC2$). In terms of coverage probability, $\text{CP}_\lambda$ is below the nominal level for $n = 50$ (being 91.40% in $SC1$ and 83.50% in $SC2$). This performance also improves when $n$ is large. The MCSE and asymptotic standard error are also large for a small sample sizes, and they tend to decrease when $n$ increases (see Figures 6.3 and 6.6). A possible reason for these results is the fact that $\hat{\lambda}^* = e^{\hat{\beta}_0^*}$. This close relationship with the intercept $\hat{\beta}_0^*$ is possibly critical to determine a strong impact of the ML/SP issue, especially in small samples. An improvement in terms of bias reduction and RMSE was obtained when using the Bayesian estimation method. For example, when $n = 50$ the RMSE in $SC1$ is 3.56 (for `Bayes_1`), 3.54 (for `Bayes_2`) and 4.16 (for `Bayes_3`). As expected, the box-plots in Figures 6.10 and 6.11, clearly show that the medians are quite close to the reference level 0, i.e., low differences between $|\text{RB}_m|$ and $|\text{RB}_u|$ were detected for the coefficients not directly affected to the ML/SP issue in both scenarios and both estimation procedures. This proximity indicates that small rates of the relative bias reduction were obtained here with respect to the results for the coefficients related to the binary covariate.

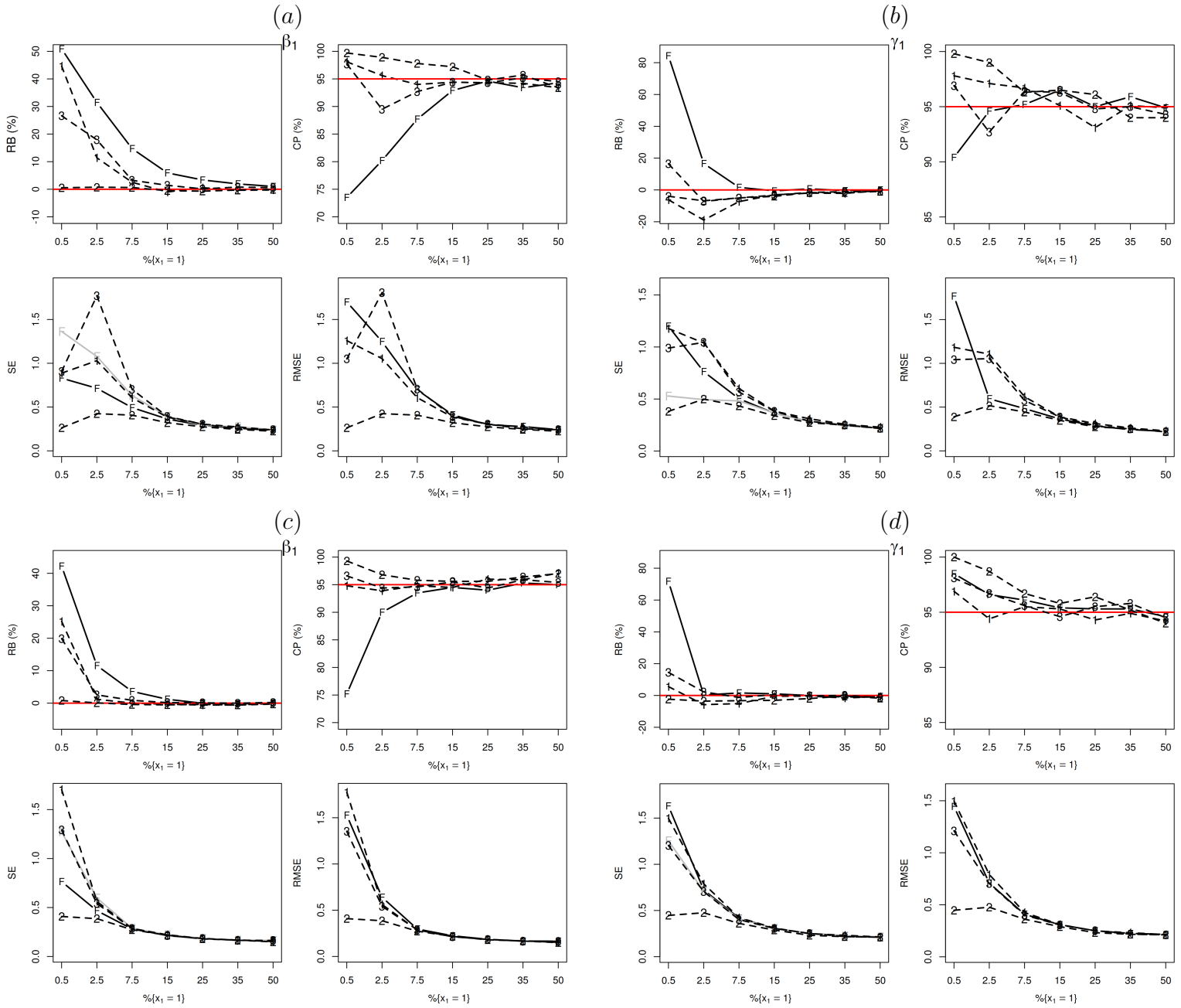


Figure 6.8: Monte Carlo results for the regression coefficients related to the binary covariate ($\beta_1$ and $\gamma_1$). We consider different configurations of the dichotomous covariate, for a fixed sample size ($n$ = 1,000). The most unbalanced case have 0.5% of ones in the binary covariate and the last setting is the balanced configuration with 50% of ones. The results in panels ($a$) and ($b$) are based on the $SC1$ and panels ($c$) and ($d$) are based on the $SC2$. In addition, lines named as "`F`" denote the results obtained when using the Firth method, where the asymptotic and MC standard errors are denoted by the *black solid lines* and *grey solid lines*, respectively. "`1`" for the `Bayes_1`, "`2`" for the `Bayes_2` and "`3`" for the `Bayes_3`.

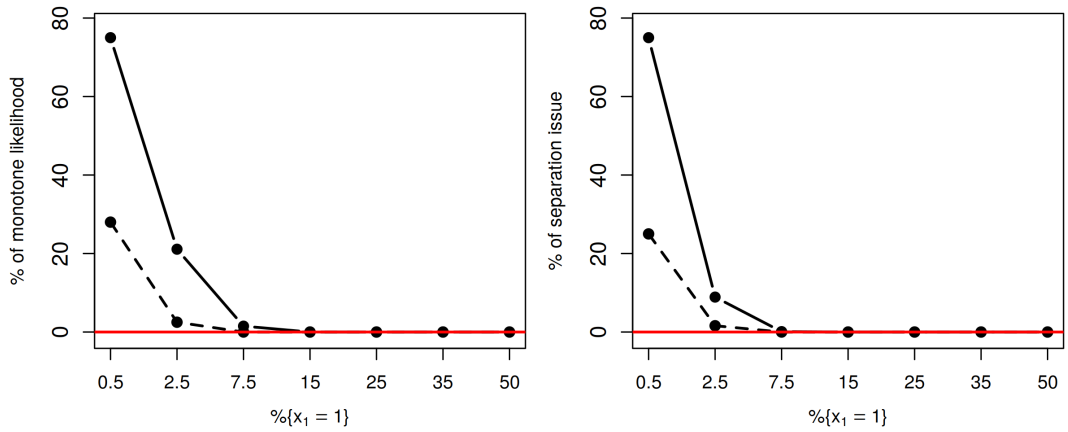


Figure 6.9: Proportions of the monotone likelihood issue in the Monte Carlo replications considering different settings of the binary covariate, for a fixed sample size ($n = 1{,}000$). We denote the scenarios $SC1$ (solid lines) and $SC2$ (dashed lines).

The normality assumption of the penalized MLEs based on the frequentist approach were assessed by using the Quantile-Quantile (Q-Q) plots and the corresponding histogram; see Figures B.1.1 and B.1.2 (Appendix B.1). The results indicate a slight deviation from normality for the coefficients related to the dichotomous covariates, and the estimated value of $\lambda$. This is also a natural consequence of the existing high imbalance degree between the levels of $x_1$, especially for large sample sizes. This influence justifies the gap between the mode and the true value of $\beta_1$ and $\gamma_1$. However, these properties were assessed when using a balanced configuration; see Figures B.1.3 to B.1.6 (Appendix B.1). The asymptotic properties for $\hat{\lambda}^*$ were assessed when the sample increases. As expected, the results show that the coefficients not directly affected by the ML/SP problem are asymptotically normal, with the mode being quite close to the true value in both scenarios.

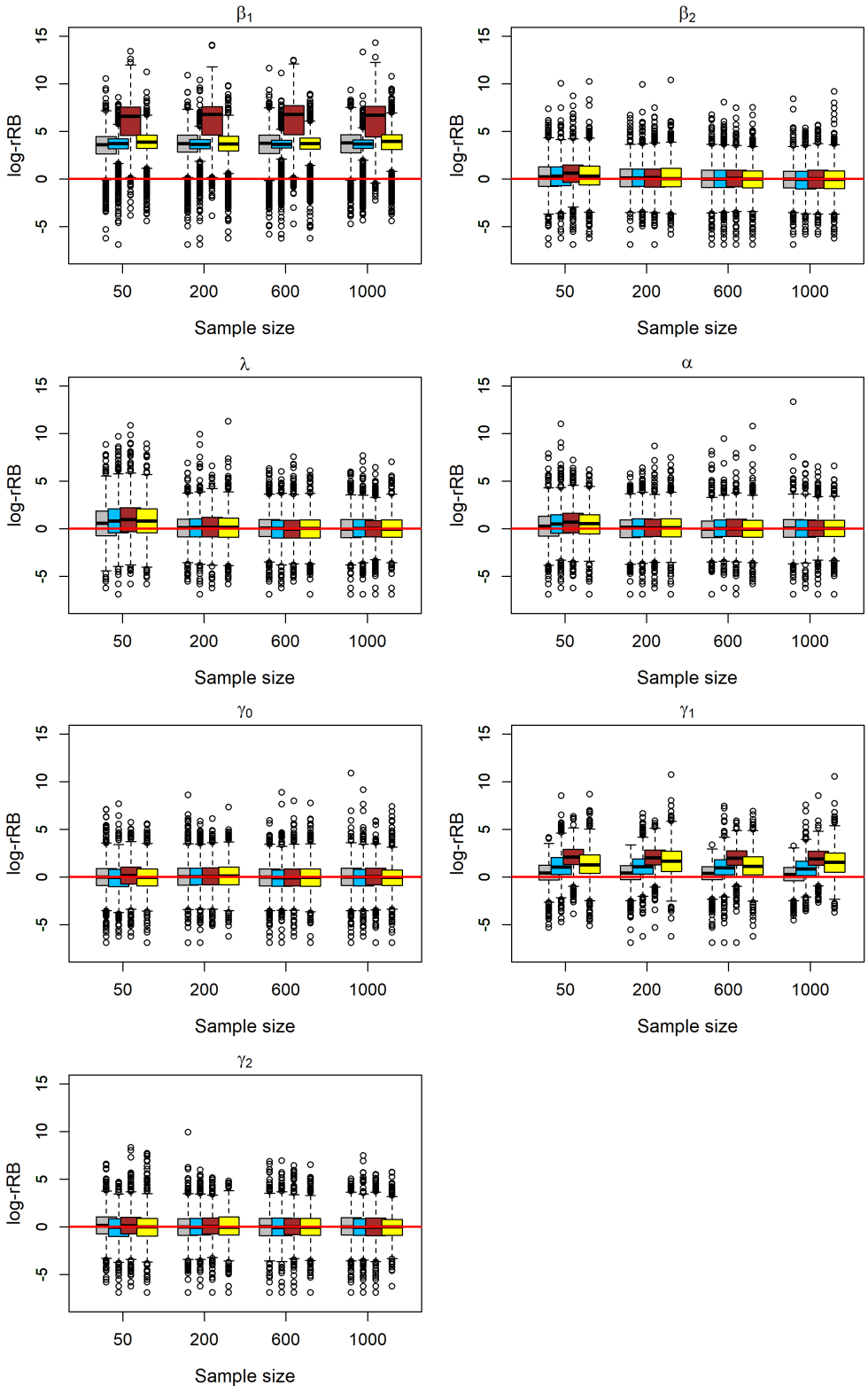


Figure 6.10: Box-plots of the log-absolute ratio of relative biases, log-rRB = $\log\left(|\mathrm{RB}_u/\mathrm{RB}_m|\right)$, with $\mathrm{RB}_u$ and $\mathrm{RB}_m$ being the relative biases obtained through the usual and the penalized cases, under the SC1. The Firth approach is denoted in grey, `Bayes_1` setting in blue, `Bayes_2` in brown and `Bayes_3` in yellow.

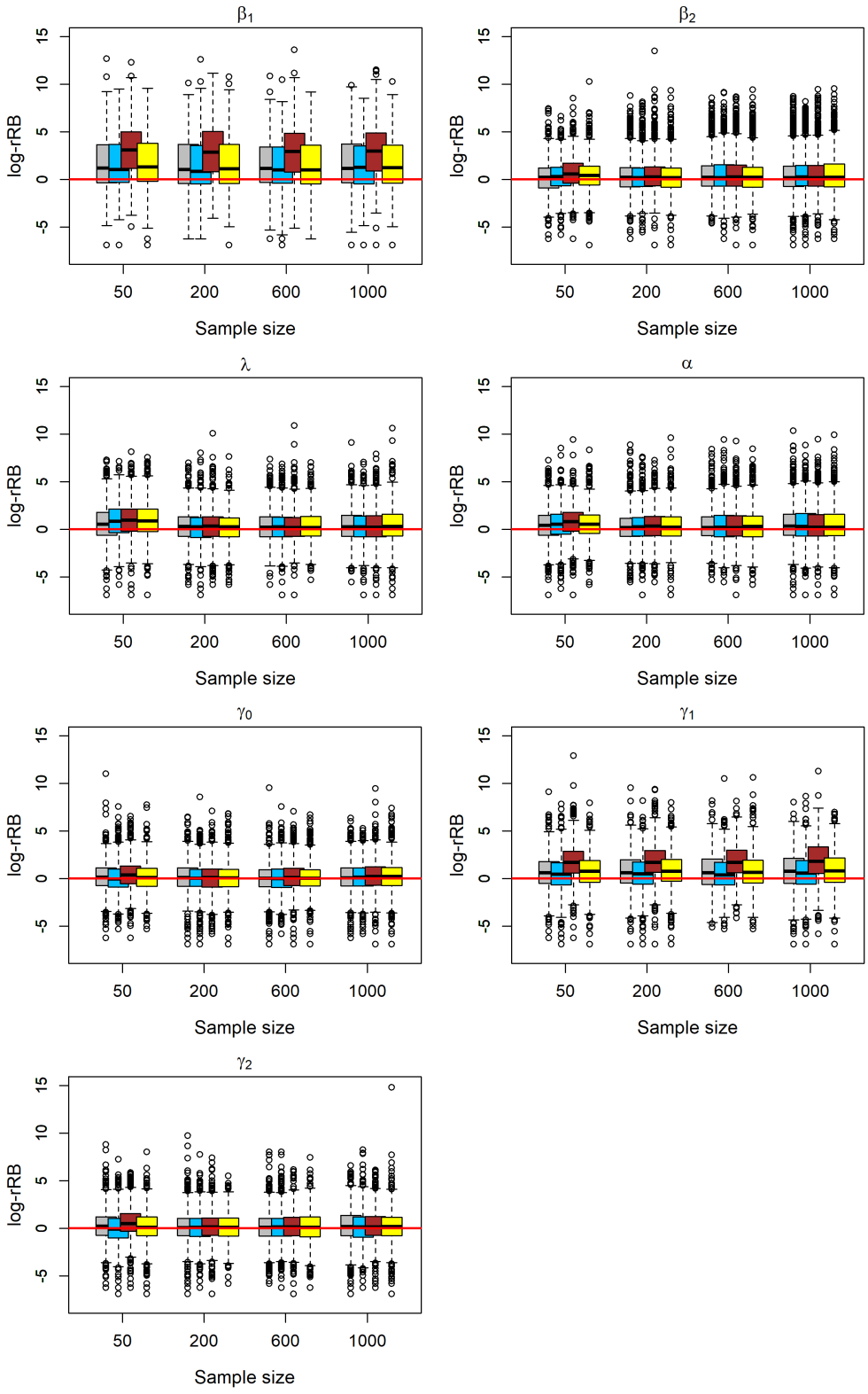


Figure 6.11: Box-plots of the log-absolute ratio of relative biases, log-rRB = $\log(|\mathrm{RB}_u/\mathrm{RB}_m|)$, with $\mathrm{RB}_u$ and $\mathrm{RB}_m$ being the relative biases obtained through the usual and the penalized cases, under the SC2. The Firth approach is denoted in grey, `Bayes_1` setting in blue, `Bayes_2` in brown and `Bayes_3` in yellow.

## 6.3 Discussion of Results (Semiparametric Model)

The main results of the simulation study in the semiparametric mixture cure rate model for both scenarios are summarized in Figures 6.12 to 6.15. The two estimation methods (frequentist and Bayesian) are confronted for both scenarios. The simulation results obtained through the Firth method indicate that the coefficients $\beta_1$ and $\gamma_1$, related to the dichotomous covariate $x_{1i}$ (inducing the ML/SP issue), seem to be underestimated even for large $n$. Their RB's are very high, ranging between 49.38 to 79.29% ($\hat{\beta}_1^*$) and 82.25 to 85.27% ($\hat{\gamma}_1^*$) in $SC1$, and between 35.49 to 65.58% ($\hat{\beta}_1^*$) and 68.66 to 76.25% ($\hat{\gamma}_1^*$) in $SC2$, respectively. Note that MCSE and asymptotic standard errors for these coefficients differ for all $n$. These variabilities tend to be larger than those obtained for the other coefficients. The RMSEs of $\hat{\beta}_1^*$ and $\hat{\gamma}_1^*$ are also high, raging between 1.36 to 1.55 ($\hat{\beta}_1^*$) and 1.71 to 1.84 ($\hat{\gamma}_1^*$) in $SC1$, and between 1.16 to 1.29 ($\hat{\beta}_1^*$) and 1.21 to 1.45 ($\hat{\gamma}_1^*$) in $SC2$. Theses coefficients have coverage rates below the nominal level. The CP in $SC1$ varies from 76.67 to 90.04% ($\beta_1$) and from 84.28 to 88.15% ($\gamma_1$). In contrast, the CP in $SC2$ is around 95.58 to 97.38% ($\gamma_1$), which is closer to the nominal level. From Figures 6.12 and 6.14, it is clear that the CP for $\beta_1$ decreases from 90.63 to 79.54% in $SC2$, and increases with a low rate, from 84.55 to 88.15% in $SC1$ ($\gamma_1$) as $n$ increases.

The influence of the chosen prior distributions, under the flexibility of the proportional hazards model in the standard mixture cure rate model, allowed us to reduce the estimation bias compared to the frequentist case. That is, the estimated RB's for the coefficient $\hat{\beta}_1^*$ and $\hat{\gamma}_1^*$ under the `Bayes_1` setting varies between -14.49 to 0.85% and -0.34 to 5.31% (in $SC1$), -14.49 to 14.47% and 0.59 to 11.15% (in $SC2$). Under this configuration of prior distributions, the magnitude of the RB's reduces significantly, especially when the sample size increases. Additionally, small values of the MCSE and RMSE were observed in $SC1$, where the values range from 0.80 to 1.12 and 0.80 to 1.13 (for $\hat{\beta}_1^*$) and 0.99 to 1.03 and 0.99 to 1.04 (for $\hat{\gamma}_1^*$), respectively. In both scenarios, the CP are relatively above to the nominal level 95%. In terms of the performance, the `Bayes_3` was the one presenting higher values for the coefficients related to ML/SP issue (when using

the $SC1$), among the three configurations of prior distributions. The RB varies from -38.50 to -23.20% (for $\hat{\beta}_1^*$) and 19.30 to 22.99% (for $\hat{\gamma}_1^*$), respectively.

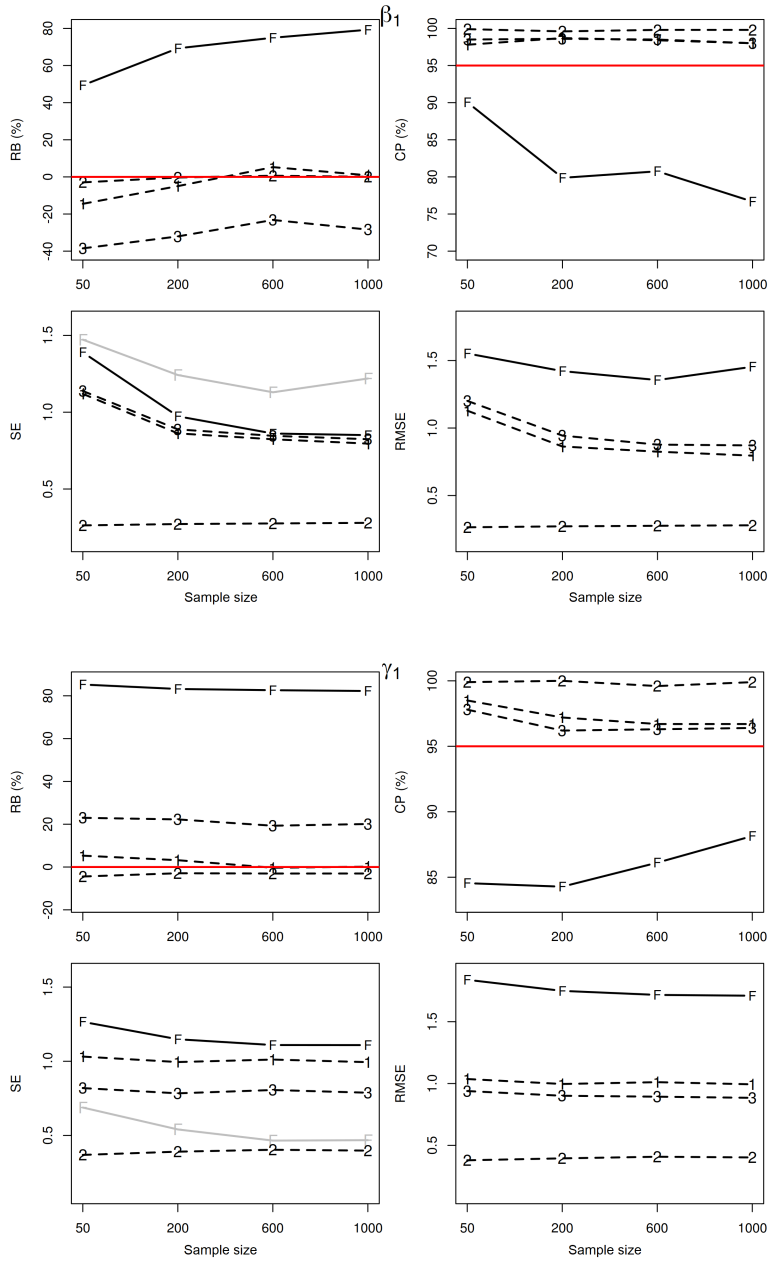


Figure 6.12: Monte Carlo simulation results for the regression coefficients related to the binary covariate $x_{1i}$, based on the $SC1$. The Firth modified score function and the Bayesian approach are applied here. The solid lines "F" denote the results obtained when using the Firth method, where the asymptotic and MCSE are denoted by the *black* and *grey solid lines*, respectively. The lines named as "1" denote the results obtained when using the `Bayes_1`, "2" for the `Bayes_2` and "3" for `Bayes_3`.

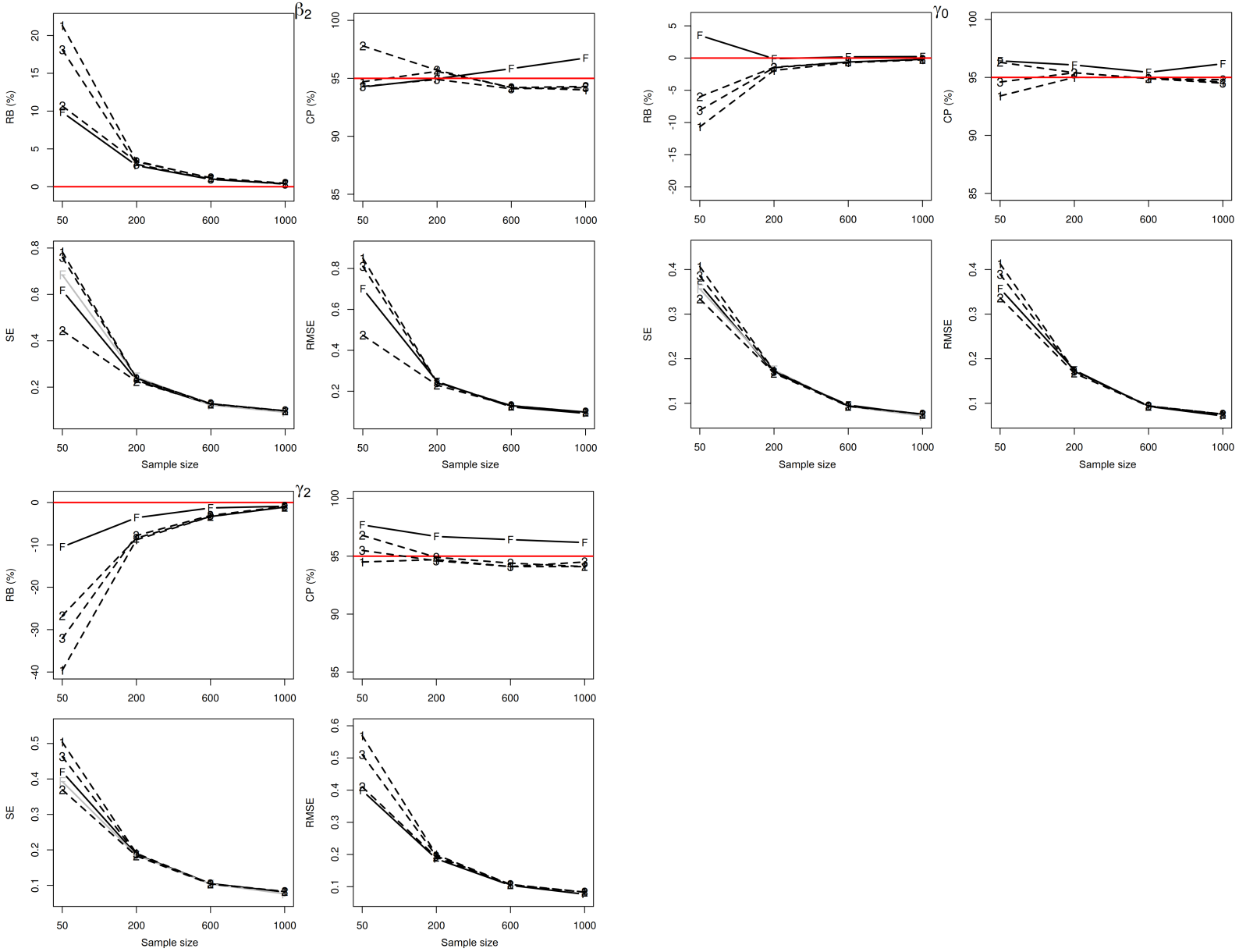


Figure 6.13: Monte Carlo simulation results for the regression coefficients not directly related to the monotone likelihood issue, based on the $SC1$. The Firth modified score function and the Bayesian approach are applied here. The solid lines "`F`" denote the results obtained when using the Firth method, where the asymptotic and MCSE are denoted by the *black* and *grey solid lines*, respectively. The lines named as "`1`" denote the results obtained when using the `Bayes_1`, "`2`" for the `Bayes_2` and "`3`" for `Bayes_3`.

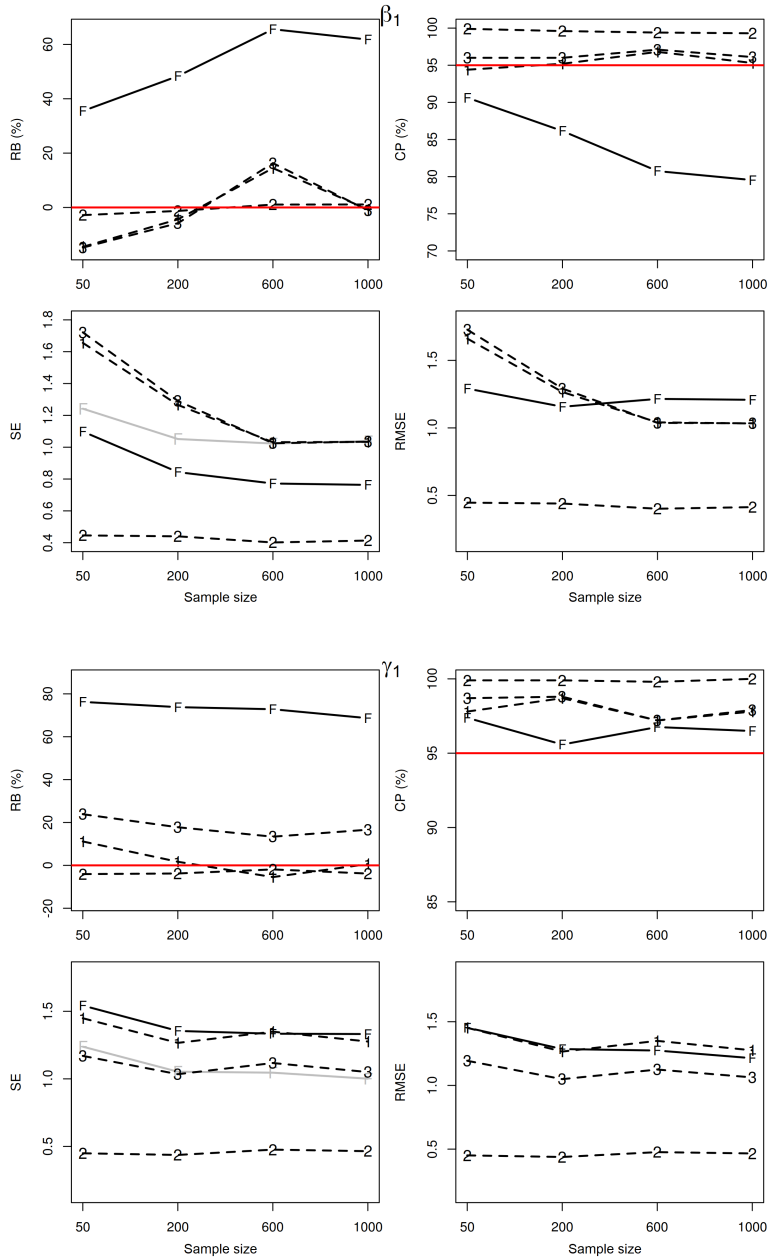


Figure 6.14: Monte Carlo simulation results for the coefficients related to the binary covariate $x_{1i}$, based on the $SC2$. The Firth modified score function and the Bayesian approach are applied here. The solid lines "`F`" denote the results obtained when using the Firth method, where the asymptotic and MCSE are denoted by the *black* and *grey solid lines*, respectively. The lines named as "`1`" denote the results obtained when using the `Bayes_1`, "`2`" for the `Bayes_2` and "`3`" for `Bayes_3`.

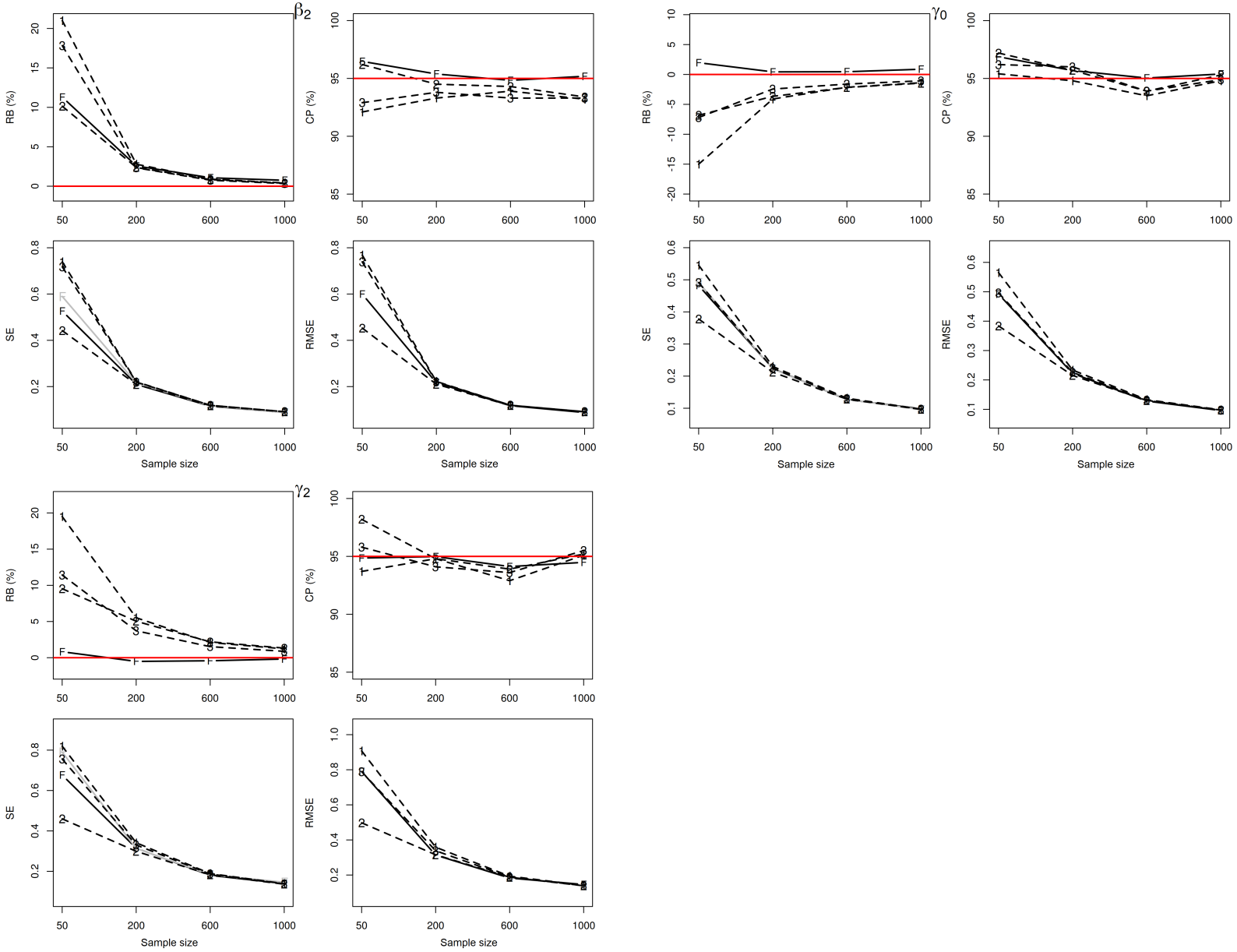


Figure 6.15: Monte Carlo simulation results for the coefficients not directly related to the monotone likelihood issue, based on the *SC*2. The Firth modified score function and the Bayesian approach are applied here. The solid lines "`F`" denote the results obtained when using the Firth method, where the asymptotic and MCSE are denoted by the *black* and *grey solid lines*, respectively. The lines named as "`1`" denote the results obtained when using the `Bayes_1`, "`2`" for the `Bayes_2` and "`3`" for `Bayes_3`.

Note that the configuration that provides poor performance in $SC1$ also indicates results with good asymptotic properties in the $SC2$, where the RB reduces significantly, raging from -14.76 to 16.30%. The CP for both coefficients are relatively around the nominal value. This trend of the bias reduction is shared by the configuration `Bayes_2` in both scenarios, with small MCSE, raging from 0.26 to 0.28 (for $\hat{\beta}_1^*$) and 0.40 to 0.45 (for $\hat{\gamma}_1^*$) in $SC1$, and between 0.44 to 0.48 (for $\hat{\beta}_1^*$) and 0.82 to 0.89 (for $\hat{\gamma}_1^*$) in $SC2$. In both scenarios, all CP are above to the nominal value. These results were expected, because in this setting the prior distributions are centered at the true values with small variability. The instability of the performance measurements for the regression coefficients related to the binary covariate is a consequence of the high imbalance degree, such as discussed in Section 6.2, see also the Figure 6.16. The impact of the imbalance degree appeared to be greater in the frequentist framework. This point confirms the idea that although the Firth correction ensures finite estimates, the method tends to provide biased results and high standard errors when the ML/SP is present, as mentioned in Chapter 1.

As discussed in the parametric case (Section 6.2), the normal Q-Q plots and the corresponding histogram related to the penalized MLEs, given in Figures B.2.1 and B.2.2 (Appendix B.2), indicate a slight deviation from normality for the coefficients related to the dichotomous covariate ($\beta_1$ and $\gamma_1$). This is a natural consequence of the existing high imbalance degree between the levels of $x_1$, especially for large sample sizes. Once again, the asymptotic normality for these coefficients were assessed when using a balanced configuration, see Figures B.2.3 to B.2.6. Note that regression coefficients not directly affected by the ML/SP problem are asymptotically normal, with the mode being quite close to the true value.

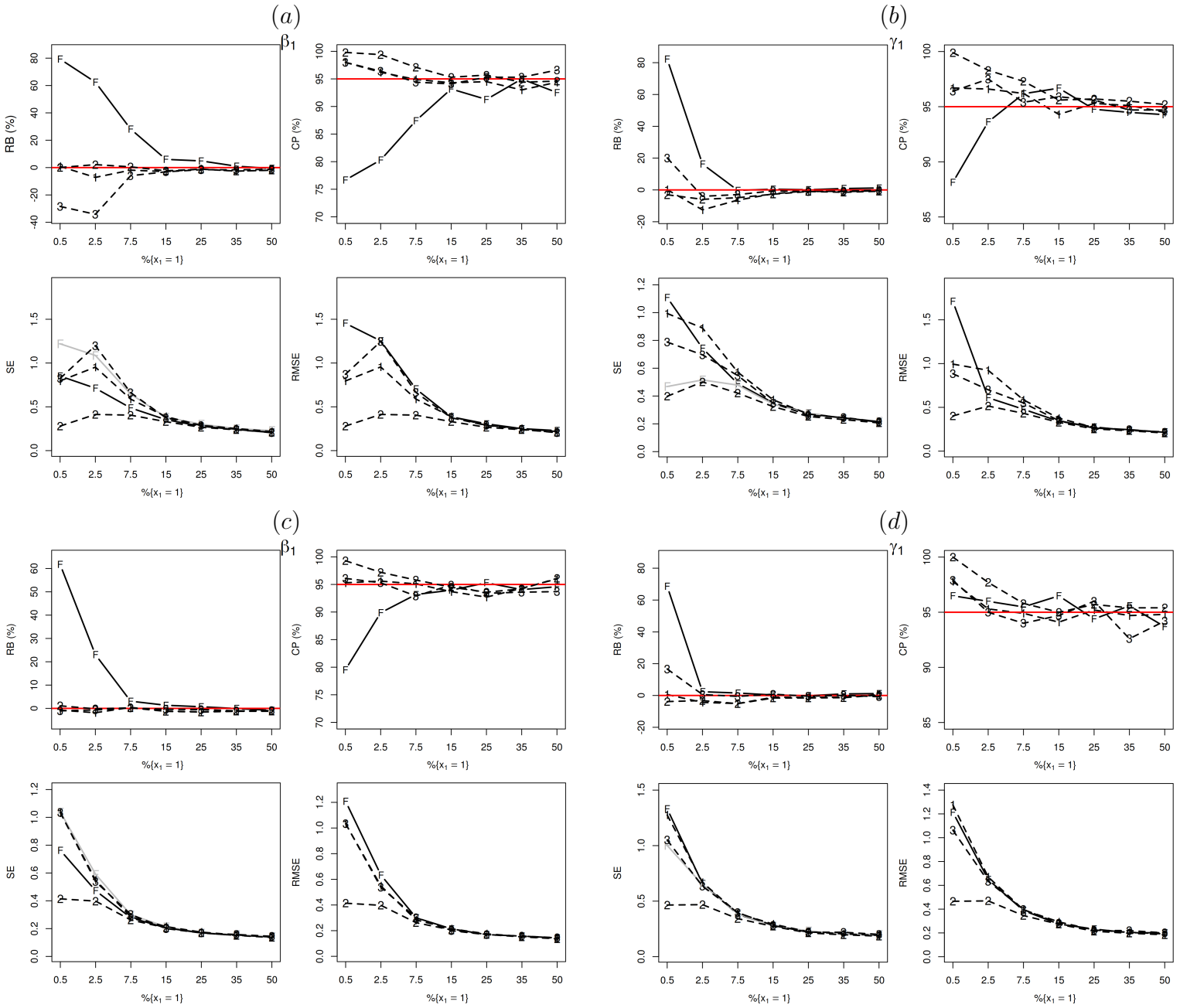


Figure 6.16: Monte Carlo results for the regression coefficients related to the binary covariate ($\beta_1$ and $\gamma_1$). We consider different configurations of the dichotomous covariate, for a fixed sample size ($n$ = 1,000). The most unbalanced case have 0.5% of ones in the binary covariate and the last setting is the balanced configuration with 50% of ones. The results in panels ($a$) and ($b$) are based on the $SC1$ and panels ($c$) and ($d$) are based on the $SC2$. In addition, lines named as "`F`" denote the results obtained when using the Firth method, where the asymptotic and MC standard errors are denoted by the *black solid lines* and *grey solid lines*, respectively. "`1`" for the `Bayes_1`, "`2`" for the `Bayes_2` and "`3`" for the `Bayes_3`.

Results related to the coefficients $\beta_2$, $\gamma_0$ and $\gamma_2$ have better behavior in terms of performance measurements in both scenarios, and both estimation methods. Recall that these coefficients are not directly connected to the dichotomous covariate inducing the ML/SP problem. Their MC estimates are close to the true values, especially for large $n$. In $SC1$, greater values of the RB's are related to the small sample size ($n = 50$), varying on average around 19.67% (for $\hat{\beta}_2^*$), -9.19% (for $\hat{\gamma}_0^*$) and -35.94% (for $\hat{\gamma}_2^*$) under the three considered prior configurations. A similar situation was observed in $SC2$, where the RB's are on average 19.01% (for $\hat{\beta}_2^*$), $-11.91\%$ (for $\hat{\gamma}_0^*$) and 16.31% (for $\hat{\gamma}_2^*$). Additionally, the results based on the frequentist estimation method indicates that small RB's are related to the following estimates $\hat{\gamma}_0^*$ and $\hat{\gamma}_2^*$, for any sample size. A high RB is detected for $\hat{\beta}_2^*$ (11.29%) when $n = 50$. Note that in both situations, the RB's move towards zero, when $n$ increases.

The MCSE's, for both estimation methods, are close to the corresponding asymptotic standard errors, and they tend to decrease as $n$ increases in both scenarios. The RMSE's are small and close to the MCSE's. As expected, the CP's are ranging around the nominal level, which occurs for both estimation methods. The magnitude of the bias reduction, when using the panelized approach (frequentist and Bayesian methods), was also assessed by analyzing the box-plots in Figures 6.17 and 6.18. In short, the results indicate similar behaviors of the parametric case, where the proximity of the median and the line 0 is also observed for the regression coefficients not directly affected to ML/SP issue.

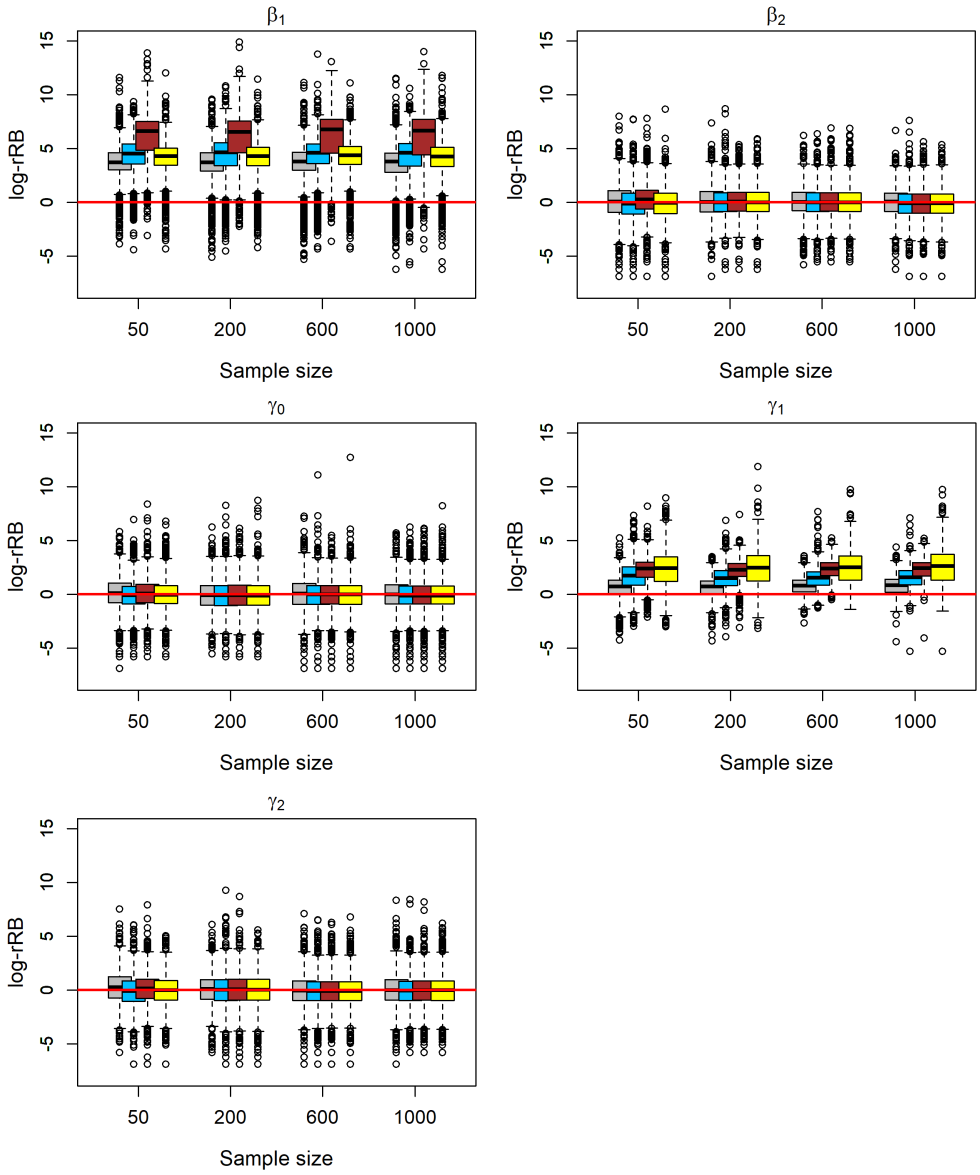


Figure 6.17: Box-plots of the log-absolute ratio of relative biases, log-rRB = $\log\left(|\mathrm{RB}_u/\mathrm{RB}_m|\right)$, with $\mathrm{RB}_u$ and $\mathrm{RB}_m$ being the relative biases obtained through the usual and the penalized cases, under the SC1. The Firth approach is denoted in grey, `Bayes_1` in blue, `Bayes_2` in brown and `Bayes_3` in yellow.

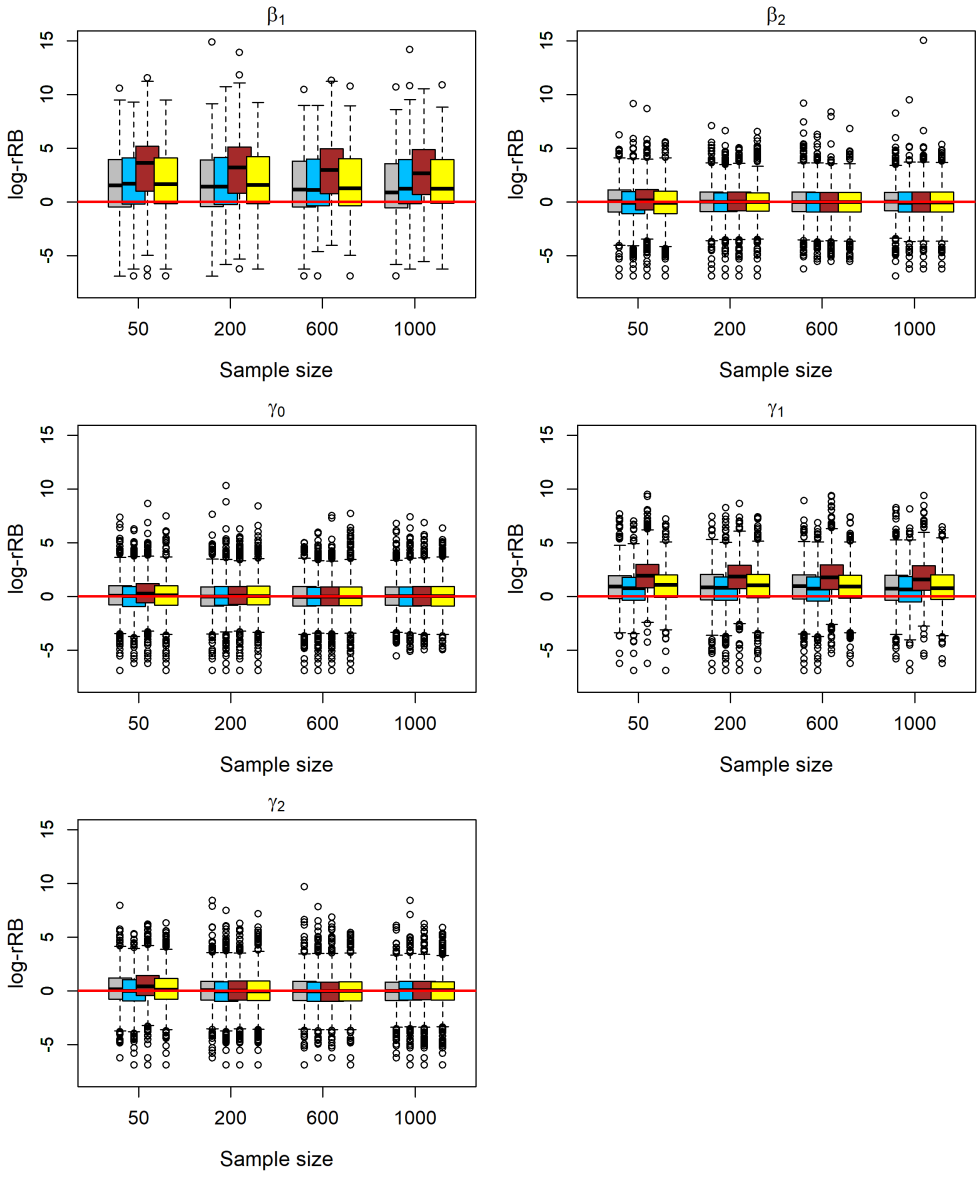


Figure 6.18: Box-plots of the log-absolute ratio of relative biases, log-rRB = $\log(|\mathrm{RB}_u/\mathrm{RB}_m|)$, with $\mathrm{RB}_u$ and $\mathrm{RB}_m$ being the relative biases obtained through the usual and the penalized cases, under the SC2. The Firth approach is denoted in grey, `Bayes_1` in blue, `Bayes_2` in brown and `Bayes_3` in yellow.

## 6.4 Brief Summary of the Chapter

This chapter presented some simulation studies in the context involving cure fraction models. Two scenarios based on different proportion of the ML/SP issue in the MC replications were considered, for both parametric and semiparametric mixture cure frac-

tion models. For each situation, the results obtained under the Firth correction and Bayesian estimation methods were confronted. In short, the results indicate that good performance measurements were obtained in the Bayesian inference under the considered prior settings. This performance improves in the semiparametric mixture cure fraction model due to its flexibility. The next chapter aims to apply the studied methods to investigate the melanoma data set, which have both the ML issue and cured individuals.

# Chapter 7

# REAL DATA APPLICATION

This section is devoted to the analysis of the melanoma data set briefly described in Chapter 1. In this data set, 221 patients have complete information for all risk factors. The maximum and minimum observed failure times are 1.13 and 30.87 weeks, respectively. The non-penalized and penalized approaches (Bayesian and frequentist) are considered to estimate the regression coefficients. The study includes five prognostics factors decomposed into seven variables. The first two variables are $(i)$ $x_{1i}$ mitosis (yes = 1, no = 0) and $(ii)$ $x_{2i}$ gender (female = 1, male = 0). The next two variables are binary representing levels of histological types: $(iii)$ $x_{31i}$ for "nodular" and $(iv)$ $x_{32i}$ for "acral lentiginous"; the reference category is "extensive malign + superficial spreading". Other two variables are binary and represent levels of the Breslow index: $(v)$ $x_{41i}$ for "$1-4$ mm" and $(vi)$ $x_{42i}$ for "$>4$ mm"; the reference level is "$<1$ mm". Finally, the last variable is $(vii)$ $x_{5i}$ ulceration (yes = 1, no = 0). The variable $x_{1i}$ (mitosis) is the one causing the ML problem as displayed in Figure 1.1.

The event under study is the metastasis occurrence and the survival time is measured from the diagnosis of primary melanoma to the date of the last visit (right censoring) or the date of the metastasis detection. The vectors of covariates affecting the latency part is denoted by $\mathbf{x}_i = (x_{1i}, x_{2i}, x_{31i}, x_{32i}, x_{41i}, x_{42i}, x_{5i})^\top$. The corresponding vector of regression coefficients is $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_{31}, \beta_{32}, \beta_{41}, \beta_{42}, \beta_5)^\top$. The same covari-

ates are considered in the incidence part, therefore, $\mathbf{z}_i = (\mathbf{1}, \mathbf{x}_i^\top)^\top$ and the corresponding vector of coefficients is $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \gamma_{31}, \gamma_{32}, \gamma_{41}, \gamma_{42}, \gamma_5)^\top$.

A descriptive analysis shows that the median survival time and the interquartile range for all individuals with metastasis are 4.60 and 11.14 weeks, respectively. The censoring rate is approximately 85%. Figures 1.1 and 7.1 suggests the presence of cured individuals in the melanoma data set, since a long plateaus containing numerous data points are observed in the survival curves related to the seven risk factors. In addition, for each categorical risk factor we have: mitosis (yes = 64.71%, no = 35.29%), gender (female = 62.44%, male = 37.56%), histological type (nodular = 17.20%, acral lentiginous = 10.40%, reference level = 72.40%), Breslow index (1-4mm = 29.41%, >4mm = 9.96%, reference level = 60.63%) and ulceration (yes = 20.36%, no = 79.64%). The main results are summarized in Tables 7.1 and 7.2; Figures 7.2 and 7.3 given bellow, and Figures D.1.4 to D.2.4 in Appendix D. The first step to analyze the melanoma data set under the Bayesian approach was related to the visual inspection of the convergence through the traceplots given in Figures D.1.1 to D.1.3 (for parametric mixture cure model) and Figures D.2.1 and D.2.2 (for the semiparametric mixture cure model).

The results obtained when using the previous specification $\mathcal{G}(0.1, 0.1)$ for the parameters $\lambda$ and $\alpha$, indicate a slight lack of convergence in the parameters related to the ML/SP issue. Furthermore, the posterior mean of the coefficient $\beta_1$ appeared with a negative sign, which does not make sense in clinical terms. This may be occurring due to the small information level in the previously mentioned prior. The same problem was also observed when other prior settings were tested, for example $\mathcal{G}(1, 1)$, $\mathcal{G}(0.01, 0.01)$, $\mathcal{G}(0.001, 0.001)$, $\mathcal{G}(1000, 1000)$. Another possible explanation may be the fact that we are dealing only with dichotomous covariates in both regression structures and, therefore, the impact of the ML issue tends to be very stronger than in cases where there are continuous variables in the regression structures.

In order to overcome this issue, and obtain a clinically interpretable coefficient for the mitosis factor, we consider a highly informative prior distribution for the parameters $\lambda$ and $\alpha$, that is, $\mathcal{G}(1, 1000)$. In addition, we also consider other informative prior distri-

butions for the regression coefficients $\beta_r$ and $\gamma_s$ (for $r = 1, 2, \cdots, p$ and $s = 0, 1, \cdots, q$), namely the $N(0, 1)$ and log-$F(9, 9)$. The last two prior distributions performed very poorly in the sensitivity analysis (simulation study). However, this choice aim to evaluate how the MLEs for the coefficients directly affected by ML issue vary when changing the prior information level. Note that, from the expression presented in Section 5.1, the variance of the log-$F(9, 9)$ prior is approximately 0.24. All MCMC setups are the same considered in the simulation studies discussed in Section 6.1.

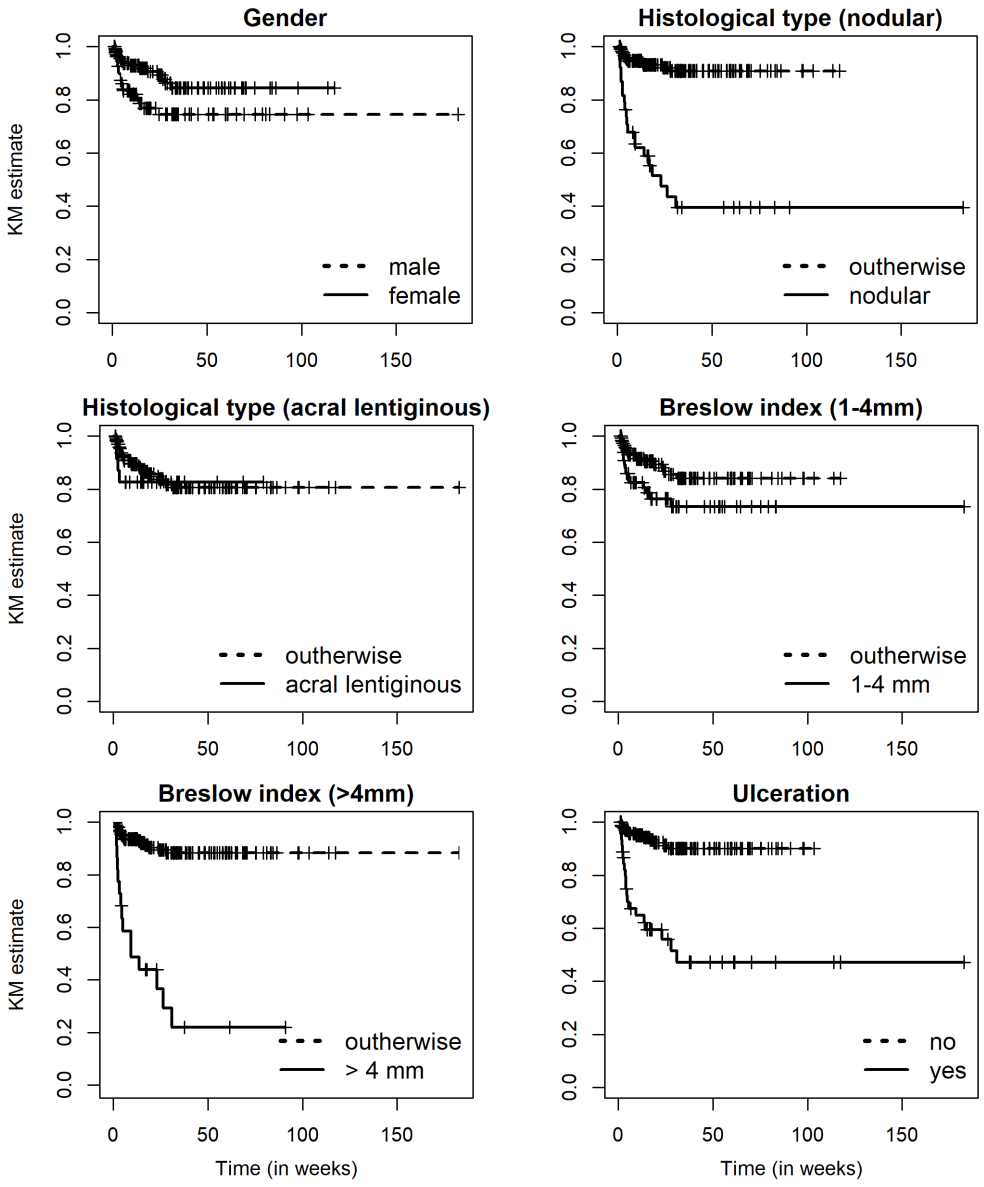


Figure 7.1: Kaplan-Meier estimates for additional risk factors investigated in the melanoma data set.

From the KM estimates in Figure 7.1, we observe a clear plateau in the right tail (above to 0) in all survival curves regarding to the mentioned risk factors related to the melanoma data set. This means the presence of cured individuals, where the height is an estimator of the cured proportion.

Table 7.1 presents the results for the penalized and non-penalized likelihoods functions, based on both estimation methods. The Weibull and logistic regression models are considered for the latency and incidence parts, respectively. Note that the usual score function provides high values for the MLEs and their corresponding standard errors for the quantities $\beta_1$, $\lambda$, $\gamma_0$ and $\gamma_1$. These results are explained by their connection with the dummy variable "mitosis" promoting the ML issue in the melanoma data set. In fact, the MLEs and standard errors of these coefficients diverge to infinity (see Figures 4.1 and 4.2). Consequently, it provides a non-significant coefficient for the previously mentioned quantities, in which their 95% Wald-type confidence intervals (CI) are large, includes 0, and lead to the following percentiles: $-0.04$ (for $\hat{\beta}_1$), 0.01 (for $\hat{\lambda}$), $-0.14$ (for $\hat{\gamma}_0$), and 0.08 (for $\hat{\gamma}_1$). Note that the percentiles are calculated as $z = \hat{\psi}_k / SE(\hat{\psi}_k)$.

The results obtained when using the FC method indicate that the coefficient $\beta_1$ becomes significant and important to explain the time until metastasis occurrence; their 95% Wald-type confidence interval is $(0.375; 3.637)$. Additionally, all HPD intervals indicate that the mitosis risk factor is important to model the time to the metastasis occurrence (see Figure 7.2), except the results obtained when using the `Bayes_1` setting with 95% HPD interval given $(-0.440; 7.675)$. In addition, the results indicate that the mitosis factor is not important to model the incidence distribution. In short, these findings are in agreement with the clinical point of view, where subjects diagnosed with mitosis have a greater chance to develop metastasis than becoming cured (Paek et al., 2007).

Note that the sign of $\hat{\beta}_1^*$ in the penalized case (for both Bayesian and frequentist) is positive, meaning a larger risk to developing metastasis is related to individuals with mitosis. Therefore, they are not likely to be cured, when compared to those without mitosis. Note that this interpretation cannot be obtained through the approach based

on the usual score function. Additionally, $\hat{\beta}_2^*$ is negative, suggesting that female subjects are less likely to develop metastasis than the males. These conclusions for mitosis and gender are in accordance with other references in biomedical research such as Damato et al. (2011) and Arce et al. (2014). The ML also affects the standard error related to $\hat{\lambda} = e^{\hat{\beta}_0}$. From the 95% Wald-type confidence intervals and HPD regions plotted in Figures D.1.4 (given in Appendix D.1), it is clear that the parameters $\lambda$ and $\alpha$ are both significant under the FC correction and Bayesian framework. The results of the confidence intervals and HPD regions also indicate that other risk factor were detected as important through at least one estimation method.

In the logistic regression, similar conclusions can be drawn. Recall that the covariates explaining the cure rate are the same ones used in the latency part. The presence of mitosis risk factor affects the estimates of the coefficient $\gamma_1$, where the resulting MLE and its standard error are determined to be infinity, i.e., the true effect of the risk factor is not well-represented in the logistic regression part. Similar results are also obtained for the intercept $\gamma_0$, showing that this coefficient is highly affected by the SP issue. However, finite estimates are found when applying the FC method and Bayesian framework. In this case, the parameter $\gamma_0$ became significant with 95% interval being $(-6.364; -0.714)$. Note that $e^{\hat{\gamma}_1^*} \approx 3.290$ represents the increment in the odds to develop metastasis for a subject with mitosis compared to the reference level "no mitosis", holding the other covariates fixed. This increment reduces when using the `Bayes_1` setting ($e^{0.243} \approx 1.275$).

Table 7.1: Results for the analysis of the melanoma data set based on the non-penalized (standard approach) and penalized likelihood, under the Firth and Bayesian methods, for both parts of the model (latency and incidence), based on the parametric mixture cure fraction model. The standard errors of the estimated coefficients are given in parentheses.

| | **Weibull regression (latency distribution)** | | | |
|---|---|---|---|---|
| | Non-penalized | Firth method | Bayes_1 | Bayes_3 |
| $\beta_1$ | -7.313 (204.310) | 2.006 (0.832) | 3.059 (2.005) | 3.120 (1.555) |
| $\beta_2$ | -0.630 (0.442) | -0.602 (0.399) | -0.254 (0.660) | -0.415 (0.552) |
| $\beta_{31}$ | 0.434 (0.486) | 0.326 (0.430) | 0.825 (0.875) | 0.981 (0.684) |
| $\beta_{32}$ | 2.443 (0.785) | 2.302 (0.748) | 0.092 (1.067) | 0.110 (0.852) |
| $\beta_{41}$ | 0.608 (0.867) | 0.518 (0.739) | 1.724 (1.246) | 1.367 (1.134) |
| $\beta_{42}$ | -0.092 (0.905) | -0.171 (0.777) | 2.282 (1.223) | 2.056 (1.193) |
| $\beta_5$ | 0.184 (0.430) | 0.169 (0.390) | 0.549 (0.681) | 0.545 (0.577) |
| $\lambda$ | 3.967 (277.086) | 0.006 (0.002) | 0.002 (0.001) | 0.002 (0.001) |
| $\alpha$ | 1.292 (0.183) | 1.300 (0.180) | 0.033 (0.006) | 0.033 (0.006) |
| | **Logistic regression (incidence distribution)** | | | |
| | Non-penalized | Firth method | Bayes_1 | Bayes_3 |
| $\gamma_0$ | -13.614 (100.836) | -3.535 (1.439) | 0.025 (1.146) | 2.071 (2.219) |
| $\gamma_1$ | 12.647 (168.327) | 1.187 (1.509) | 0.243 (1.418) | 1.880 (2.641) |
| $\gamma_2$ | -0.679 (0.564) | -0.632 (0.524) | -0.162 (1.207) | 1.198 (2.226) |
| $\gamma_{31}$ | 1.522 (0.646) | 1.455 (0.608) | 0.600 (2.172) | 1.330 (2.434) |
| $\gamma_{32}$ | -0.461 (0.925) | -0.287 (0.835) | -0.031 (1.867) | 0.291 (2.608) |
| $\gamma_{41}$ | 0.883 (0.793) | 0.828 (0.721) | 0.110 (1.811) | 0.714 (2.467) |
| $\gamma_{42}$ | 2.715 (0.988) | 2.412 (0.903) | 1.531 (2.497) | 1.205 (2.452) |
| $\gamma_5$ | 0.923 (0.603) | 0.868 (0.566) | 0.526 (1.940) | 1.089 (2.595) |

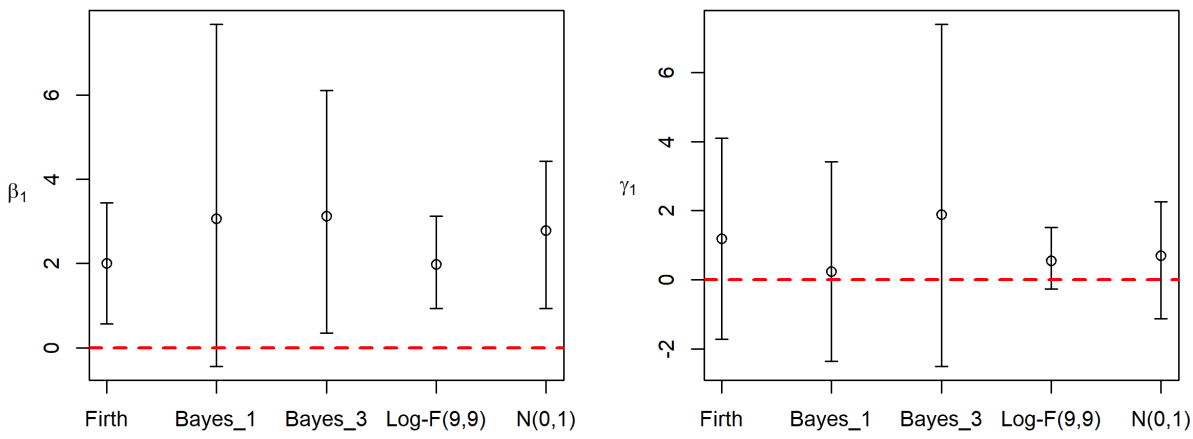


Figure 7.2: Comparing 95% Wald-type confidence intervals (Firth method) and 95% HPD regions (Bayesian framework) for the coefficients related to the mitosis risk factor ($\beta_1$ and $\gamma_1$) related to the parametric mixture cure rate model. The small circle denote the posterior mean (Bayesian) or penalized maximum likelihood estimates (Firth method).

Table 7.2 summarizes the main results of the data analysis under the semiparametric mixture cure model. Once again, it is clear that the presence of the ML problem affects the estimation of the regression coefficient directly related to the mitosis risk factor, since the usual score function provides high values for the regression coefficients related to the mitosis factor ($\beta_1$ and $\gamma_1$) and for their corresponding standard errors. Note that, this is a key variable, since it is the one promoting the ML issue. Note also that, in practice, the MLEs and their standard errors do not have finite values in the case without penalization. Consequently, it provides a non-significant regression coefficients $\beta_1$ and $\gamma_1$, in which the 95% Wald-type confidence interval (CI) is large and includes 0.

The model fit assuming the Firth modified score function determines finite estimates for $\beta_1$ (2.330), $\gamma_0$ (-2.972) and $\gamma_1$ (0.596). Under the Bayesian framework, finite estimates were also obtained: $\beta_1$ (4.356 for `Bayes_1` and 1.846 for `Bayes_3`) and $\gamma_1$ (0.378 for `Bayes_1` and 1.964 for `Bayes_3`). The estimate related to the intercept changes the sign when using the `Bayes_1` setting ($\hat{\gamma}_0^* = 0.550$). In terms of significance, the results indicate that the mitosis risk factor has a positive effect to model the latency distribution, for both estimation method, when considering the settings `Bayes_1` and `N(0,1)`, respectively. In addition, the null effect of the risk factor was observed for the incidence part; see the 95% Wald-type confidence and HPD intervals plotted in Figure 7.3. These results are in line with the conclusions from some works in the biomedical research field, and are also in agreement with those obtained in the parametric cases. According to the results, the intercept term $\gamma_0$ of the logistic model is also significant, when using the frequentist approach and Bayesian method (settings `Bayes_3` and `N(0,1)`). The positive estimate for $\beta_1$ in both estimation methods suggests that, the presence of mitosis in patients diagnosed with cutaneous melanoma increases their hazards for developing metastasis than the hazard to be cured.

The 95% Wald-type confidence intervals and HPD regions given in Figure D.2.3 (Appendix D.2) indicate that the category "acral lentiginous", of the Histological type risk factor, is important to model the latency distribution under the Firth method and the `Bayes_3` setting. Additionally, the categories "nodular", of the Histological type risk

factor, and "1–4mm", of the Breslow index, were detected with significant effect to model the incidence distribution, for both estimation methods and under the settings `Bayes_3` and `log-`$F$`(9,9)` in the Bayesian case. Note that these risk factors show null effect for the incidence part, under the mentioned settings. Furthermore, the ulceration factor was detected with a significant effect in the incidence part under the non-penalized approach. However, this factor becomes non-significant under the penalized approach (Firth method and Bayesian case).

In addition, the category ">4mm" of the Breslow index, was detected with a positive effect in the incidence part. Their 95% Wald-type confidence interval for $\gamma_{42}$, is $(0.935; 4.099)$ in Firth correction; the 95% HPD region are $(0.704; 4.752)$ for `Bayes_3` and $(0.091; 1.741)$ for `Log-`$F$`(9,9)` prior. In contrast, the same category was detected to be significant in the latency distribution. This fact is clinically relevant, since it is in agreement with other previous studies conducted without considering the cure fraction (Cherobin et al., 2018). The reported results also show that the risk factors, that are not directly related to the ML issue, have similar estimates, especially for the usual score and Firth modified approaches.

Finally, note that, in the non-penalized case, the expected log-likelihood functions $\tilde{\ell}_1(\boldsymbol{\gamma}, \mathbf{g}^{(m)})$ and $\tilde{\ell}_2(\boldsymbol{\theta}, \mathbf{g}^{(m)})$ are not maximized at the points $\hat{\gamma}_1$ and $\hat{\beta}_1$. In fact, the non-penalized MLEs for these coefficients do not exist. In practice, the outcomes from the optimization algorithm are obtained when a pre-specified maximum number of iterations is achieved. The remaining regression coefficients, not directly related to the ML/SP issue, have finite estimates and the expected log-likelihood functions are properly maximized (with similar performance) for both non-penalized and penalized versions (under the Firth method).

Table 7.2: Results for the analysis of the melanoma data set based on the non-penalized (standard approach) and penalized likelihood, under the Firth and Bayesian methods, for both parts of the model (latency and incidence) based on the semiparametric mixture cure fraction model. The standard errors of the estimated coefficients are given in parentheses.

| | **Cox regression (latency distribution)** | | | |
|---|---|---|---|---|
| | Non-penalized | Firth method | Bayes_1 | Bayes_3 |
| $\beta_1$ | 50.466 ($7.082 \times 10^{15}$) | 2.330 (0.715) | 4.356 (2.648) | 1.846 (5.018) |
| $\beta_2$ | -0.790 (0.368) | -0.684 (0.379) | -0.586 (0.372) | -0.713 (0.493) |
| $\beta_{31}$ | 0.654 (0.374) | 0.362 (0.502) | 0.832 (0.468) | 0.357 (0.444) |
| $\beta_{32}$ | 0.610 (0.766) | 2.226 (0.502) | 0.213 (0.701) | 1.979 (0.858) |
| $\beta_{41}$ | 0.166 (0.396) | 0.362 (0.457) | 1.350 (0.826) | 0.626 (0.779) |
| $\beta_{42}$ | -0.047 (0.493) | -0.253 (0.615) | 2.078 (0.852) | 0.064 (0.769) |
| $\beta_5$ | 0.151 (0.518) | -0.054 (0.449) | 0.703 (0.401) | -0.056 (0.437) |
| | **Logistic regression (incidence distribution)** | | | |
| | Non-penalized | Firth method | Bayes_1 | Bayes_3 |
| $\gamma_0$ | -13.282 (78.859) | -2.972 (0.615) | 0.550 (0.640) | -4.322 (1.527) |
| $\gamma_1$ | 11.056 (78.860) | 0.596 (0.672) | 0.378 (0.687) | 1.964 (1.618) |
| $\gamma_2$ | -0.485 (0.477) | -0.549 (0.448) | -0.168 (0.688) | -0.731 (0.658) |
| $\gamma_{31}$ | 1.156 (0.542) | 1.400 (0.539) | 0.956 (1.050) | 1.583 (0.762) |
| $\gamma_{32}$ | -0.129 (0.758) | -0.457 (0.783) | 0.097 (0.958) | -0.640 (0.933) |
| $\gamma_{41}$ | 1.153 (0.587) | 0.902 (0.605) | -0.131 (0.834) | 0.686 (0.809) |
| $\gamma_{42}$ | 2.715 (0.832) | 2.517 (0.807) | 0.817 (1.403) | 2.652 (1.049) |
| $\gamma_5$ | 1.092 (0.505) | 0.951 (0.501) | 0.294 (0.975) | 1.098 (0.735) |

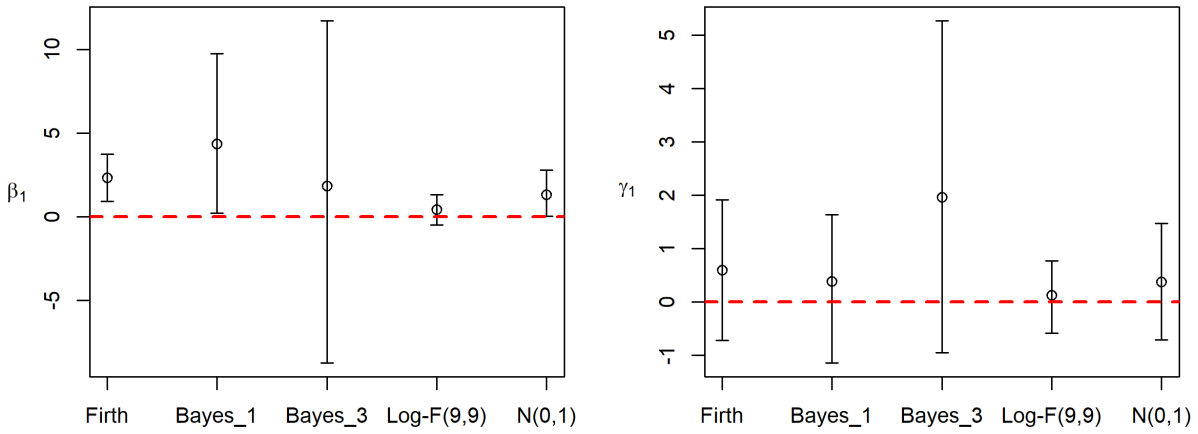


Figure 7.3: Comparing 95% Wald-type confidence intervals (Firth method) and 95% HPD regions (Bayesian framework) for the coefficients related to the mitosis risk factor ($\beta_1$ and $\gamma_1$) related to the semiparametric mixture cure rate model. The small circles denote the posterior mean (Bayesian) or penalized maximum likelihood estimates (Firth method).

## 7.1 Brief Summary of the Chapter

This chapter shows the data analysis based on the melanoma data set, which motivate this work. The presence of ML issue in the data set is evident, since it led to an overestimation of the MLEs and their uncertainty levels for the coefficients directly affected to this phenomenon. As a consequence, it resulted in non-significance of the coefficients related to the "mitosis", one of the most important risk factors. Therefore, these conclusions changes when using the penalized likelihood function for both approaches (frequentist and Bayesian). In the parametric mixture cure fraction model, certain levels of prior information provided a negative sign for the coefficient related to the mitosis factor. This result does not make any sense in clinical terms. However, it was possible to get around this problem by choosing a very informative prior for the shape and scale parameters of the Weibull distribution. The next chapter presents the final remarks.

# Chapter 8

# CONCLUSIONS

Monotone likelihood or separation problem occurs due to special conditions on the data set. Its occurrence and implications is a topic rarely addressed in the context of cure rate models. The main goal of this thesis is to discuss the results obtained through two estimation procedures (Bayesian and frequentist) from a mixture cure model fitted to data sets affected by the ML/SP problem. The study is developed under the parametric and semiparametric specifications for the latency part through the Cox PH model. Some attention is given to a comparative analysis confronting the MLEs obtained via the usual score function, leading to divergent estimates and the penalized case (through the Firth and Bayesian approaches) to reduce bias through a penalized likelihood function. The results indicate that the ML/SP issue should not be ignored and its presence can mislead the conclusions from the model fit without penalization. In particular, the impact of some covariates may be incorrectly detected as significant, leading to wrong interpretations. The estimation procedure based on augmented data, and implemented through an EM algorithm, allowed us to obtain the MLEs based on distinct parts of the likelihood.

The penalization proposed by Firth could then be applied to impose separate modifications in terms of score function. In order to consider the semiparametric framework, a zero-tail constraint for the baseline survival was required to avoid numerical instability. In a MC simulation study based on artificial data sets, the results from the Firth modi-

fied case indicate finite estimates for the coefficient connected to the covariate $x_1$ directly related to the ML problem. However, the performance in terms of RB and CP are poor, even for large $n$. This phenomenon is explained by the high imbalance of $x_1$ in the tested simulations scenarios; see again Sections 6.2 and 6.3, and also consider the discussion in Kenne Pagui and Colosimo (2020). On the other hand, greater improvement in terms of bias reduction and CP around the nominal level was obtained in the Bayesian inference under some prior settings. This improvement in term of the performances is more evident in the semiparametric mixture cure model due to its flexibility. In both estimation methods, asymptotic properties can be better observed, when moving the configuration of the dichotomous covariate from the unbalanced to balanced case. The disadvantage of the balanced setting for our analyses is that it implies in losing the ML structure in the MC replications. As expected, parameters not directly related to the ML/SP issue are well estimated, and this quality improves as the sample size increases.

The discussed real application is related to a melanoma data set configured with a cure rate fraction and containing the ML/SP issue (all individuals with metastasis indicate the same level of mitosis). According to specialists, "mitosis" is indeed an important factor to explain the metastasis occurrence, however, using the non-penalized approach, the mitosis factor is detected as non-significant for both parts of the model (latency and incidence). As expected, the usual score function provides large values for the MLEs and its standard errors for the coefficients related to this particular covariate. Important changes are observed in the model fit when the modified score function is considered via the Firth correction and Bayesian framework. As an example, the mitosis factor (not significant without penalization) is detected with a significant impact on the latency part of the penalized mixture cure model; the effect of mitosis remains not significant for the incidence part. Other alterations of significance in the comparison "non-penalized vs. penalized" can also be observed for other coefficients in the model.

# Appendix

This appendix is dived in four parts. The Appendix A, shows the general expressions of the components related to the observed information matrices, including the second and third-order cumulants for both parametric and semiparametric mixture cure fraction models. In Appendix B, we present the extra results for the frequentist estimation method, based on same scenarios (1 and 2) of the MC simulation study developed in Chapter 6. Appendix C shows the marginal posterior distributions for each parameter, and additional MC simulation results based on the Bayesian and frequentist estimation methods for both scenarios (1 and 2) and both specifications for the latency distribution (parametric and semiparametric). Finally, Appendix D is devoted to present additional results of the melanoma data analysis.

## Appendix A. Additional Theorical Details

In this section, we present the third-order cumulants $\nu_{u,v,r}$ corresponding to the generic entry of the matrix $\partial \mathbf{J}(\boldsymbol{\theta})/\partial\theta_r$, for $r = 1, 2, \cdots, p+2$ (for the parametric specification), see Section 4.1. The general details to obtain the covariance matrix of the obtained estimates (see Section 4.2) are presented in appendices A.1, A.2 and A.3.

### Appendix A.1. Third-order Cumulants

Here, we present the third-order cumulants related to the latency part of the model, when the parametric model is specified. These quantities are the derivative of the generic entry $(r, u)$ of the matrix $\mathbf{J}(\boldsymbol{\theta})$.

$$
\begin{aligned}
\mathrm{J}_{\beta_r\beta_u\beta_v} &= \sum_{i=1}^{n} g_i^{(m)} x_{ri}x_{ui}x_{vi}t_i^{\alpha}\lambda\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta}) & \mathrm{J}_{\alpha\alpha\lambda} &= \sum_{i=1}^{n} g_i^{(m)} t_i^{\alpha}\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})\log^2(t_i),\\
\mathrm{J}_{\lambda\beta_r\beta_v} &= \sum_{i=1}^{n} g_i^{(m)} x_{ri}x_{vi}t_i^{\alpha}\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta}) & \mathrm{J}_{\lambda\lambda\lambda} &= -\sum_{i=1}^{n}\frac{2\delta_i g_i^{(m)}}{\lambda^3},\\
\mathrm{J}_{\alpha\beta_r\beta_v} &= \sum_{i=1}^{n} g_i^{(m)} x_{ri}x_{vi}t_i^{\alpha}\lambda\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})\log(t_i) & \mathrm{J}_{\lambda\beta_r\lambda} &= 0,\\
\mathrm{J}_{\alpha\alpha\beta_v} &= \sum_{i=1}^{n} g_i^{(m)} x_{vi}t_i^{\alpha}\lambda\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})\log^2(t_i) & \mathrm{J}_{\lambda\lambda\alpha} &= 0,\\
\mathrm{J}_{\alpha\beta_r\lambda} &= \sum_{i=1}^{n} g_i^{(m)} x_{ri}t_i^{\alpha}\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})\log(t_i) & \mathrm{J}_{\alpha\alpha\alpha} &= \sum_{i=1}^{n} g_i^{(m)}\left\{t_i^{\alpha}\lambda\exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})\log^3(t_i) - \frac{2\delta_i}{\alpha^3}\right\}.
\end{aligned}
$$

**Appendix A.2. Components of the Observed Information Matrix (Parametric Mixture Cure Fraction Model)**

Elements of the observed information matrix $\mathbf{J}_{obs}(\boldsymbol{\psi})$ obtained through the Louis method, when the latency distribution is modeled parametrically (see reference in the main text of the thesis). Here $\xi_i = \mathbf{z}_i^\top \boldsymbol{\gamma}$ follow the same notation presented in the Section 3.1, for $i = 1, 2, \cdots, n$.

$$
\begin{aligned}
\mathrm{J}^{obs}_{\gamma_s\gamma_b} &= -\sum_{i=1}^{n} z_{si}\left[\left(\frac{\partial g_i}{\partial \gamma_b}\right) - \frac{\partial \pi(\xi_i)}{\partial \gamma_b}\right] = -\sum_{i=1}^{n} z_{si}\left\{\left(\frac{\partial g_i}{\partial \gamma_b}\right) - z_{bi}\pi(\xi_i)\left[1-\pi(\xi_i)\right]\right\},\\
\mathrm{J}^{obs}_{\beta_r\beta_u} &= -\sum_{i=1}^{n} x_{ri}\left[\left(\frac{\partial g_i}{\partial \beta_u}\right)\left(\delta_i - \lambda\exp(\mathbf{x}_i^\top\boldsymbol{\beta})t_i^\alpha\right) - g_i x_{ui} t_i^\alpha \lambda \exp(\mathbf{x}_i^\top\boldsymbol{\beta})\right],\\
\mathrm{J}^{obs}_{\lambda\lambda} &= -\sum_{i=1}^{n}\left[\left(\frac{\partial g_i}{\partial \lambda}\right)\left(\frac{\delta_i}{\lambda} - t_i^\alpha\exp(\mathbf{x}_i^\top\boldsymbol{\beta})\right) - \frac{g_i\delta_i}{\lambda^2}\right],\\
\mathrm{J}^{obs}_{\lambda\beta_r} &= -\sum_{i=1}^{n}\left[\left(\frac{\partial g_i}{\partial \beta_r}\right)\left(\frac{\delta_i}{\lambda} - t_i^\alpha\exp(\mathbf{x}_i^\top\boldsymbol{\beta})\right) - g_i x_{ri} t_i^\alpha\exp(\mathbf{x}_i^\top\boldsymbol{\beta})\right],\\
\mathrm{J}^{obs}_{\alpha\alpha} &= -\sum_{i=1}^{n}\left[g_i\left(\frac{\delta_i}{\alpha^2} + \lambda\exp(\mathbf{x}_i^\top\boldsymbol{\beta})t_i^\alpha\log^2(t_i)\right)\right]\\
&\quad + \left(-\sum_{i=1}^{n}\left[\left(\frac{\partial g_i}{\partial \alpha}\right)\left[\delta_i\left(\frac{1}{\alpha} + \log(t_i)\right) - \lambda t_i^\alpha\exp(\mathbf{x}_i^\top\boldsymbol{\beta})\log(t_i)\right]\right]\right),\\
\mathrm{J}^{obs}_{\lambda\alpha} &= -\sum_{i=1}^{n}\left[\left(\frac{\partial g_i}{\partial \alpha}\right)\left(\frac{\delta_i}{\lambda} - t_i^\alpha\exp(\mathbf{x}_i^\top\boldsymbol{\beta})\right) - g_i t_i^\alpha\exp(\mathbf{x}_i^\top\boldsymbol{\beta})\log(t_i)\right],\\
\mathrm{J}^{obs}_{\alpha\beta_r} &= -\sum_{i=1}^{n}\left(\frac{\partial g_i}{\partial \beta_r}\right)\left[\delta_i\left(\frac{1}{\alpha} + \log(t_i)\right) - t_i^\alpha\lambda\exp(\mathbf{x}_i^\top\boldsymbol{\beta})\log(t_i)\right]\\
&\quad + \left(-\sum_{i=1}^{n} g_i x_{ri} t_i^\alpha\lambda\exp(\mathbf{x}_i^\top\boldsymbol{\beta})\log(t_i)\right),\\
\mathrm{J}^{obs}_{\gamma_s\beta_r} &= -\sum_{i=1}^{n} z_{si}\left(\frac{\partial g_i}{\partial \beta_r}\right), \quad \mathrm{J}^{obs}_{\gamma_s\lambda} = -\sum_{i=1}^{n} z_{si}\left(\frac{\partial g_i}{\partial \lambda}\right), \quad \mathrm{J}^{obs}_{\gamma_s\alpha} = -\sum_{i=1}^{n} z_{si}\left(\frac{\partial g_i}{\partial \alpha}\right),\\
\frac{\partial g_i}{\partial \beta_r} &= -(1-\delta_i)x_{ri}t_i^\alpha\lambda\exp(\mathbf{x}_i^\top\boldsymbol{\beta})g_i(1-g_i),
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial g_i}{\partial \alpha} &= -(1-\delta_i) t_i^{\alpha} \log(t_i) \lambda \exp(\mathbf{x}_i^{\top}\boldsymbol{\beta}) g_i (1-g_i), \\
\frac{\partial g_i}{\partial \lambda} &= -(1-\delta_i) t_i^{\alpha} \exp(\mathbf{x}_i^{\top}\boldsymbol{\beta}) g_i (1-g_i), \quad \frac{\partial g_i}{\partial \gamma_s} = (1-\delta_i) z_{si} g_i (1-g_i), \\
g_i &= \delta_i + (1-\delta_i) \times \exp\left[x_i - t_i^{\alpha}\lambda \exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})\right] / \left(1 + \exp\left[\xi_i - t_i^{\alpha}\lambda \exp(\mathbf{x}_i^{\top}\boldsymbol{\beta})\right]\right).
\end{aligned}
$$

## Appendix A.3. Components of the Observed Information Matrix (Semiparametric Mixture Cure Fraction Model)

Elements of the $d \times d$ observed information matrix $\mathbf{J}_{obs}(\boldsymbol{\psi})$ obtained through the Louis method, under the semiparametric specification for the latency distribution (see reference in the main text of the thesis). Denoting $D_{j,0} = \sum_{j \in R_i} g_j \exp\left(\mathbf{x}_j^{\top}\boldsymbol{\beta}\right)$ and $D_{j,r} = \sum_{j \in R_i} x_{jr} g_j \exp\left(\mathbf{x}_j^{\top}\boldsymbol{\beta}\right)$, the observed information matrix have the following components:

$$
\begin{aligned}
\mathrm{J}^{obs}_{\gamma_s\gamma_b} &= -\sum_{i=1}^{n}\left[z_{ib}\left(\frac{\partial g_i}{\partial \gamma_s} - \frac{\partial \pi(\xi_i)}{\partial \gamma_s}\right)\right], \\
\mathrm{J}^{obs}_{\beta_r\beta_u} &= \sum_{i=1}^{n}\delta_i\left[\frac{D^*_{j,ru}}{D_{j,0}} - \frac{D^*_{j,u}D_{j,r}}{D^2_{j,0}}\right], \\
\mathrm{J}^{obs}_{\gamma_b\beta_u} &= -\sum_{i=1}^{n}\left[z_{ib}\left(\frac{\partial g_i}{\partial \beta_u}\right)\right], \\
D^*_{j,ru} &= \sum_{j \in R_i}\left[\left(\frac{\partial g_j}{\partial \beta_u} + x_{ju} g_j\right) x_{jr} \exp\left(\mathbf{x}_j^{\top}\boldsymbol{\beta}\right)\right], \\
D^*_{j,u} &= \sum_{j \in R_i}\left[\left(\frac{\partial g_j}{\partial \beta_u} + x_{ju} g_j\right) \exp\left(\mathbf{x}_j^{\top}\boldsymbol{\beta}\right)\right], \\
\frac{\partial g_i}{\partial \beta_u} &= -(1-\delta_i)\left[\left(\frac{\partial H_0(t_i|\cdot)}{\partial \beta_u}\right) + x_{iu} H_0(t_i|\cdot)\right] g_i(1-g_i)\exp\left(\mathbf{x}_i^{\top}\boldsymbol{\beta}\right), \\
\frac{\partial H_0(t_i|\cdot)}{\partial \beta_r} &= \sum_{i:t_{(i)} \le t}\left(\frac{\partial h_{0i}}{\partial \beta_r}\right) = -\sum_{i:t_{(i)} \le t}\left[\frac{\delta_i\, D_{j,r}}{D^2_{j,0}}\right],
\end{aligned}
$$

$$
\begin{aligned}
g_i &= \delta_i + (1-\delta_i)\ \exp\left\{\mathbf{z}_i^\top \boldsymbol{\gamma} - H_0(t_i|\cdot)\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\} / \left[1+\exp\left\{\mathbf{z}_i^\top \boldsymbol{\gamma} - H_0(t_i|\cdot)\exp\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right)\right\}\right], \\
\frac{\partial g_i}{\partial \gamma_s} &= (1-\delta_i) z_{is} g_i (1-g_i).
\end{aligned}
$$

**Appendix B: Additional Details of the MC Study: frequentist approach**

This section displays the numerical details associated with the results related to the frequentist approach, for both parametric and semiparametric specifications in the latency distribution. Additionally, Figures B.1.1 and B.1.2, and Figures B.2.1 and B.2.2 contain the normal Q-Q plots and the corresponding histogram (true values are the horizontal red lines) for the main simulation study.

In Figures B.1.3 to B.1.6 and Figures B.2.3 to B.2.6, we illustrate the normal Q-Q plots and the respective histograms obtained by assuming different configurations of the dichotomous covariate $x_{1i}$. That is, the new cases account for the following percentages of $\{x_{1i} = 1\}$, $i = 1, 2, \ldots, n$: 0.5% (highly unbalanced), 2.5%, 7.5%, 15%, 25%, 35% and 50%. All the mentioned results are presented in Appendices B.1 and B.2.

**Appendix B.1. Results Based on the Parametric Approach**

This section displays the additional details of the MC study discussed in Chapter 6.

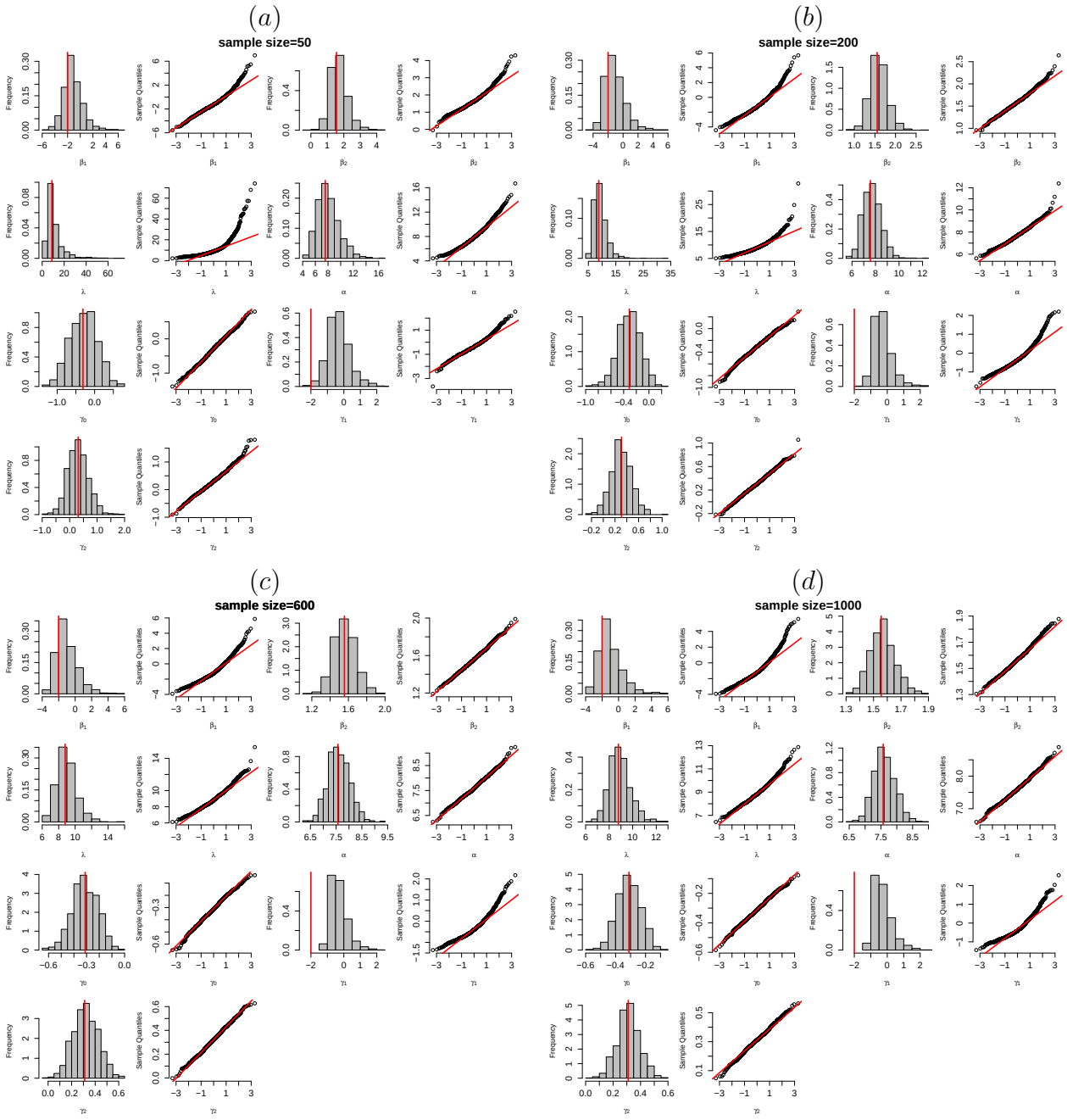


Figure B.1.1: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates (Scenario 1, latency part). Sample sizes: 50 ($a$), 200 ($b$), 600 ($c$) and 1,000 ($d$). The vertical red line in the histogram is the true value of the parameter.

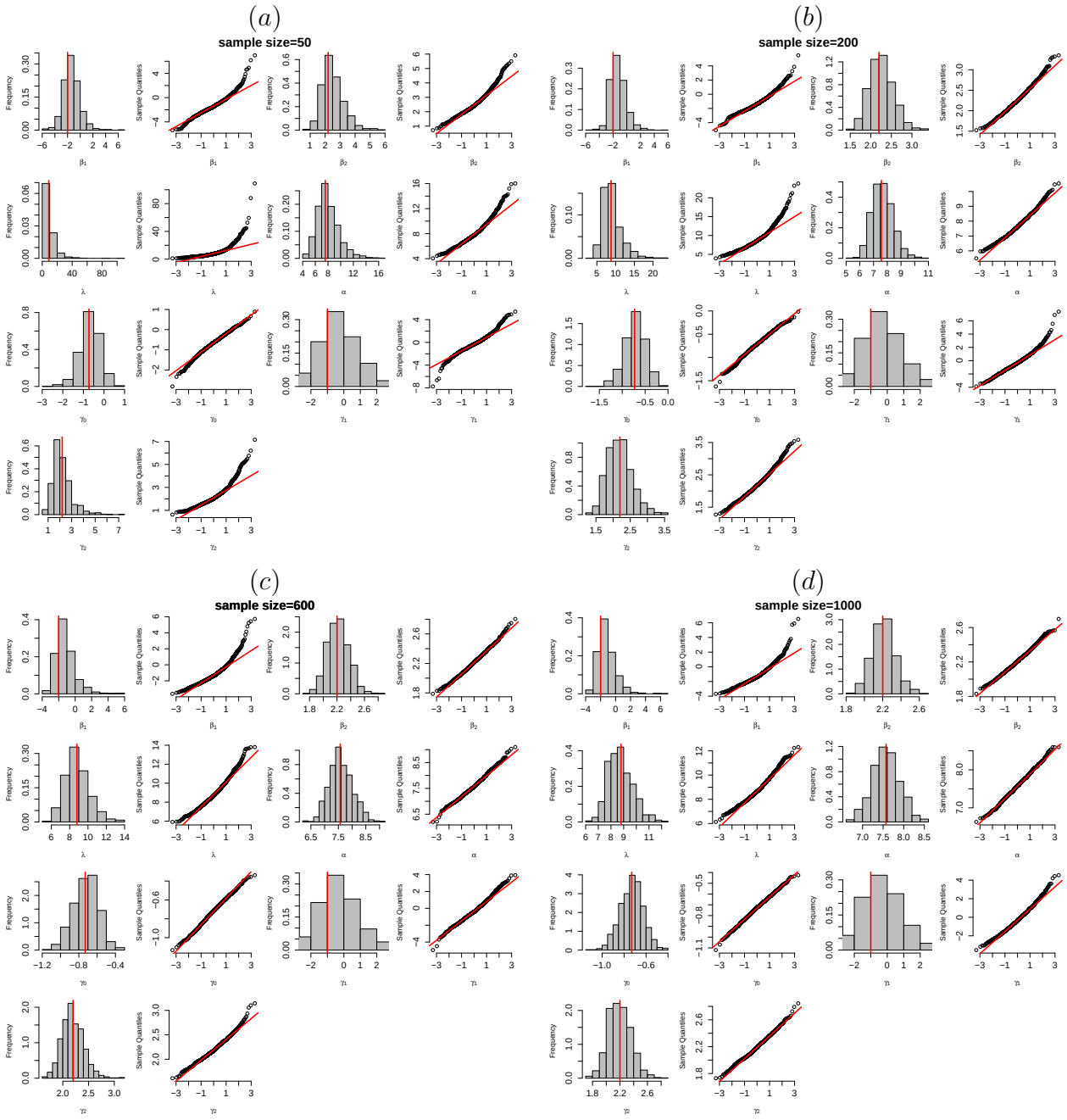


Figure B.1.2: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates (Scenario 2, latency part). Sample sizes: 50 ($a$), 200 ($b$), 600 ($c$) and 1,000 ($d$). The vertical red line in the histogram is the true value of the parameter.

**Histogram and Q-Q Plots for the Considered Cases**

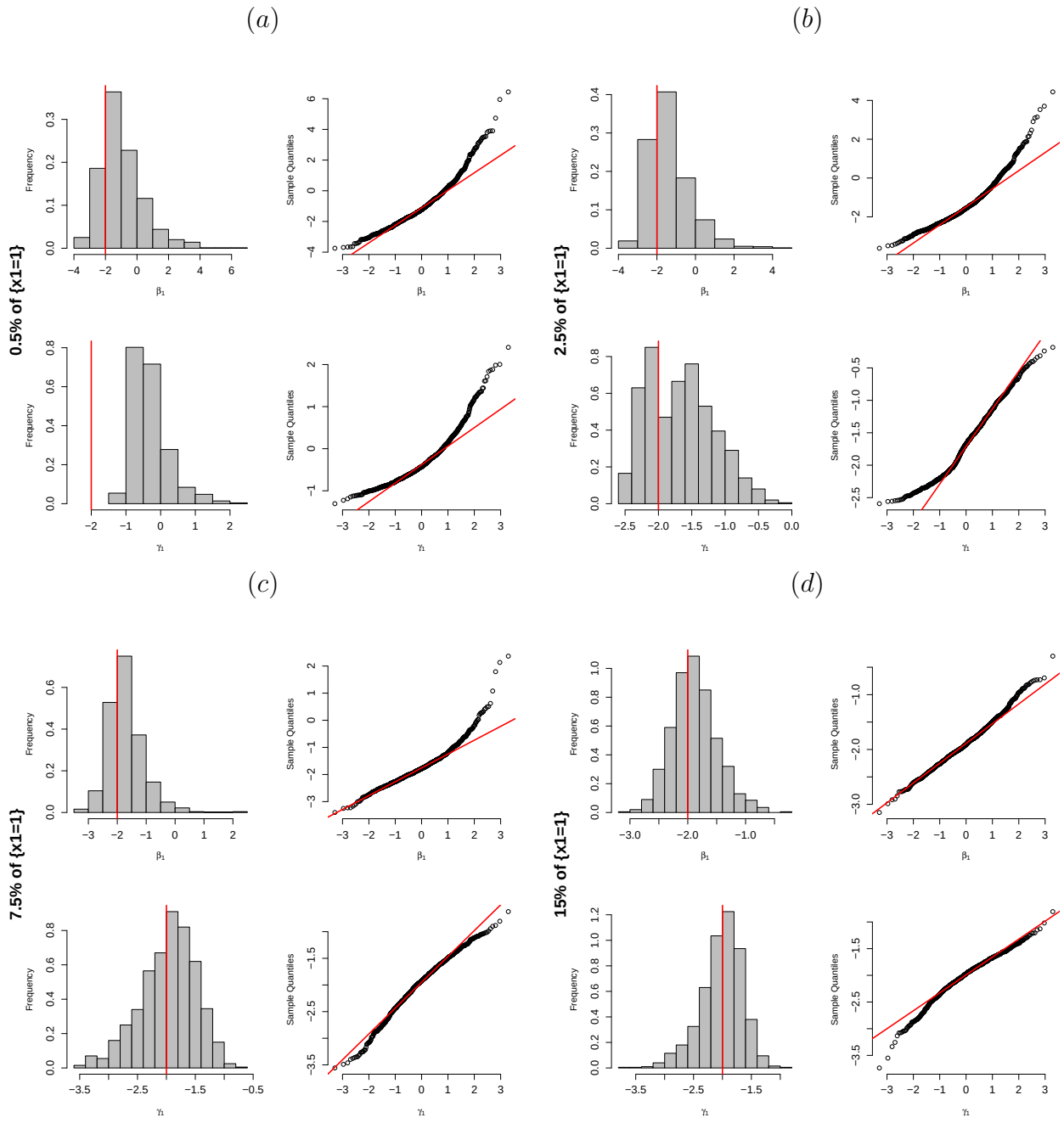


Figure B.1.3: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates for $\beta_1$ and $\gamma_1$ (Scenario 1). Percentages of $x_{1i} = 1$: 0.5% $(a)$, 2.5% $(b)$, 7.5% $(c)$ and 15% $(d)$. The vertical red line in the histogram is the true value of the parameter.

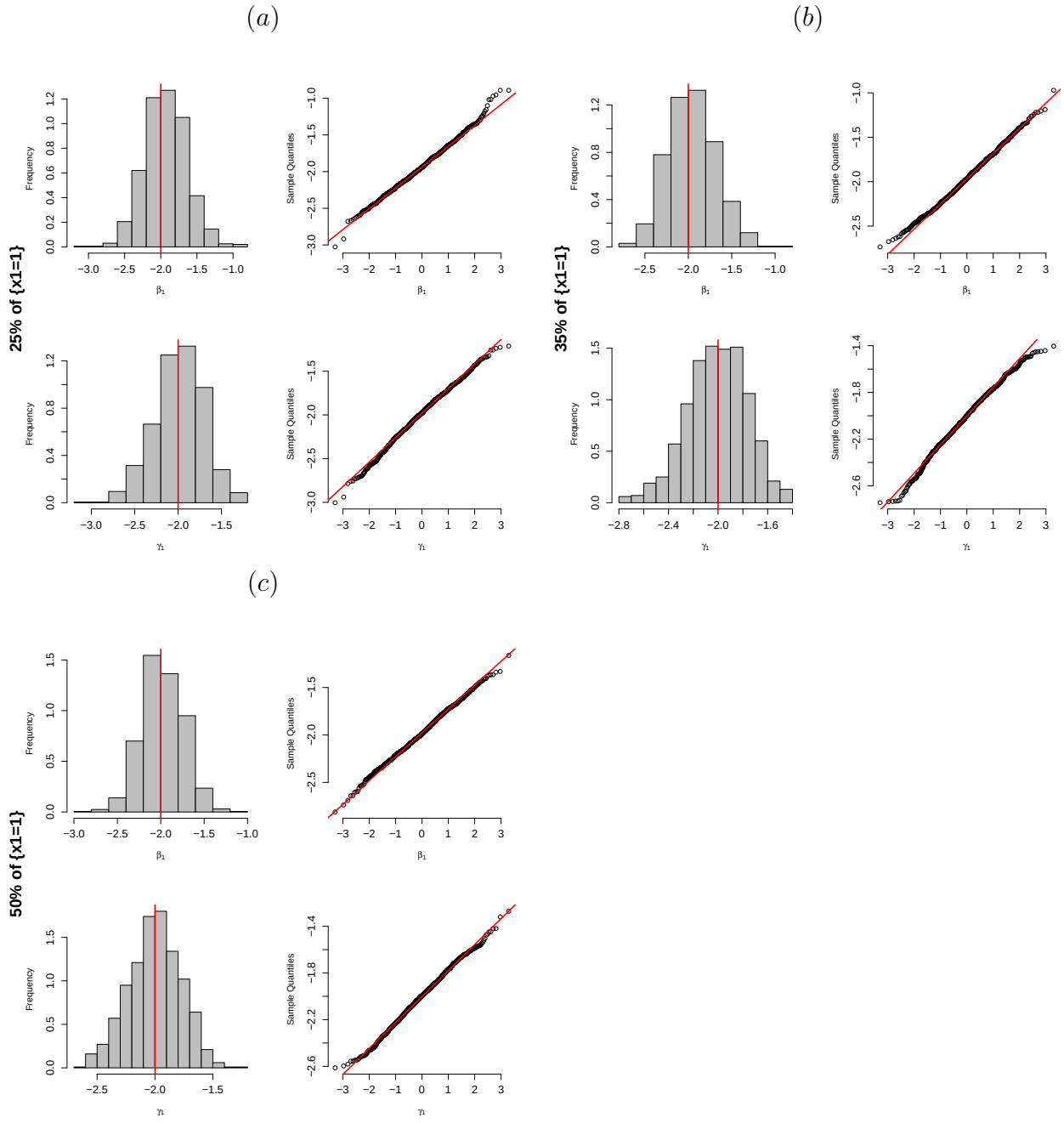


Figure B.1.4: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates for $\beta_1$ and $\gamma_1$ (Scenario 1). Percentages of $x_{1i} = 1$: 25% ($a$), 35% ($b$) and 50% ($c$). The vertical red line in the histogram is the true value of the parameter.

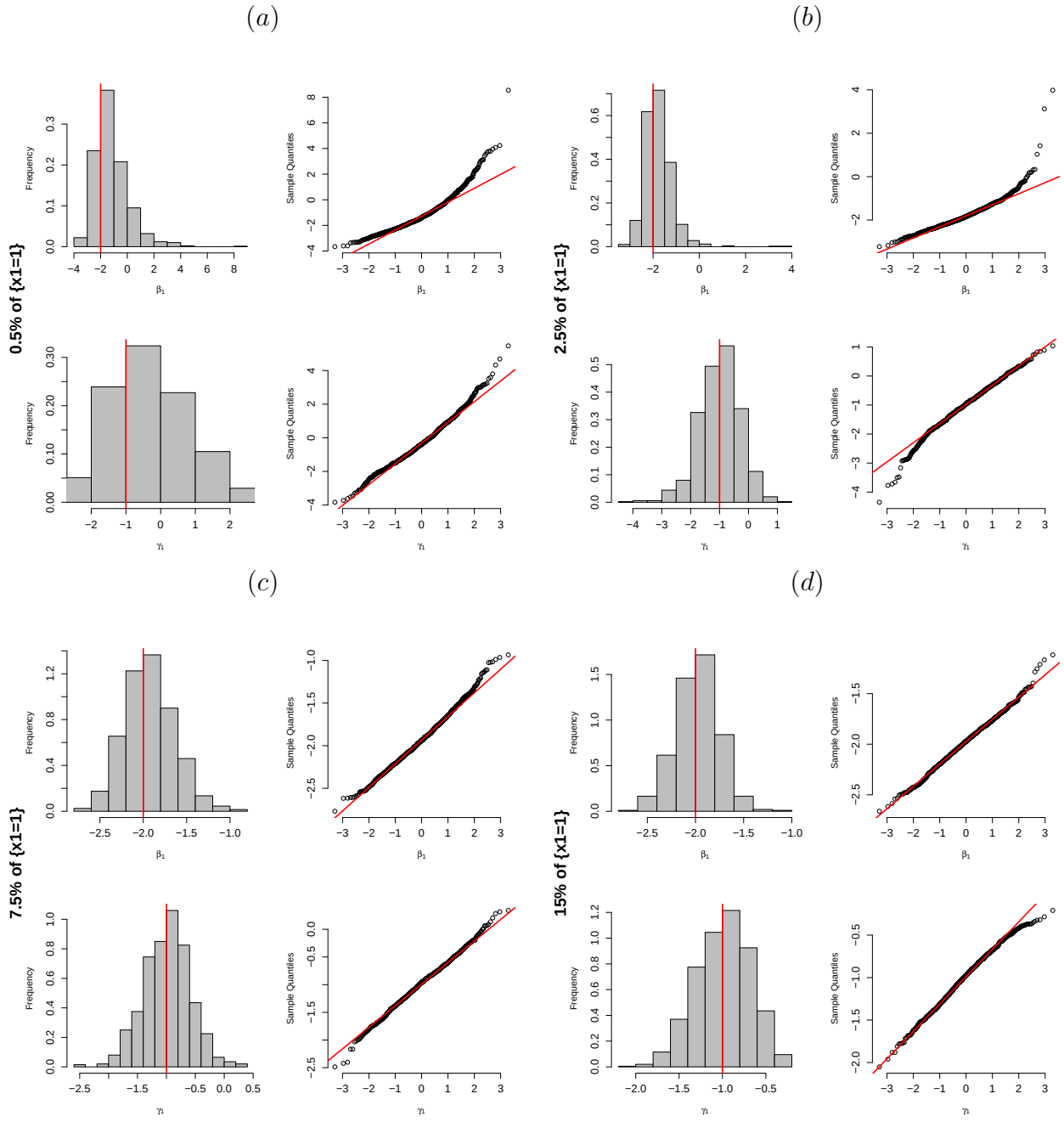


Figure B.1.5: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates for $\beta_1$ and $\gamma_1$ (Scenario 2). Percentages of $x_{1i} = 1$: 0.5% ($a$), 2.5% ($b$), 7.5% ($c$) and 15% ($d$). The vertical red line in the histogram is the true value of the parameter.

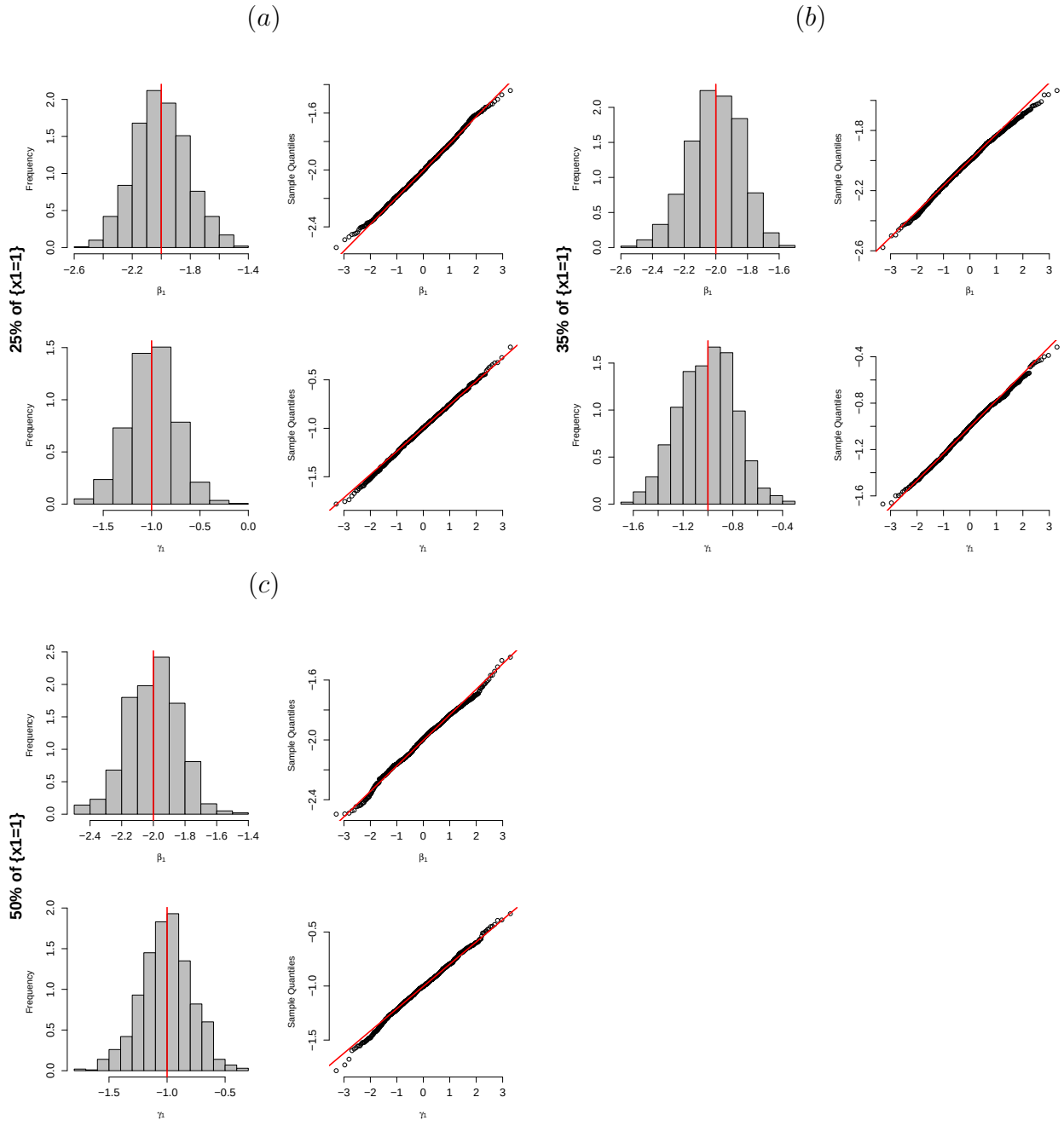


Figure B.1.6: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates for $\beta_1$ and $\gamma_1$ (Scenario 2). Percentages of $x_{1i} = 1$: 25% ($a$), 35% ($b$) and 50% ($c$). The vertical red line in the histogram is the true value of the parameter.

### Appendix B.2. Results for the Semiparametric Approach

This section displays the additional details of the MC simulation results obtained when using the semiparametric mixture cure fraction models.

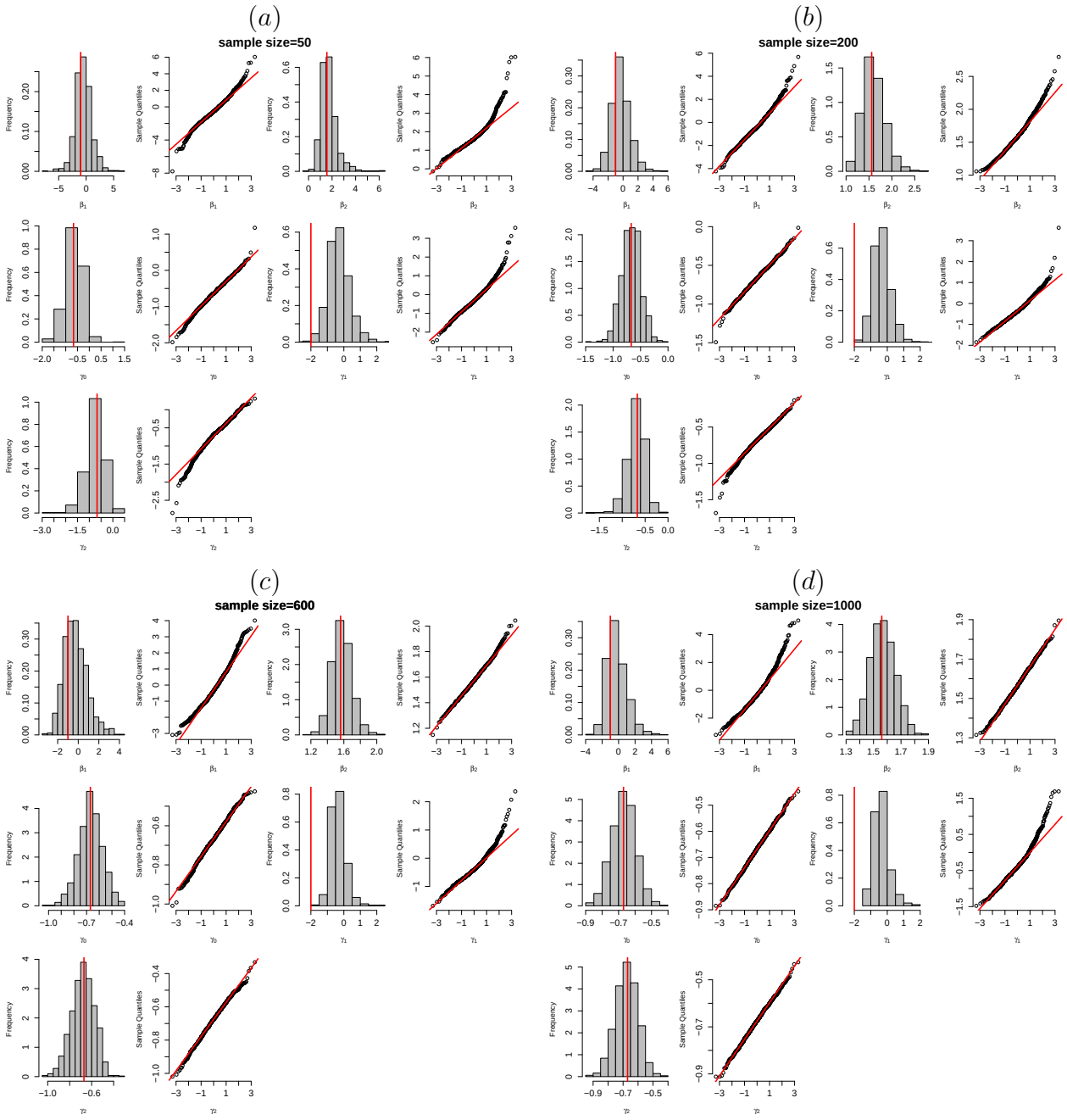


Figure B.2.1: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates (Scenario 1, latency part). Sample sizes: 50 ($a$), 200 ($b$), 600 ($c$) and 1,000 ($d$). The vertical red line in the histogram is the true value of the parameter.

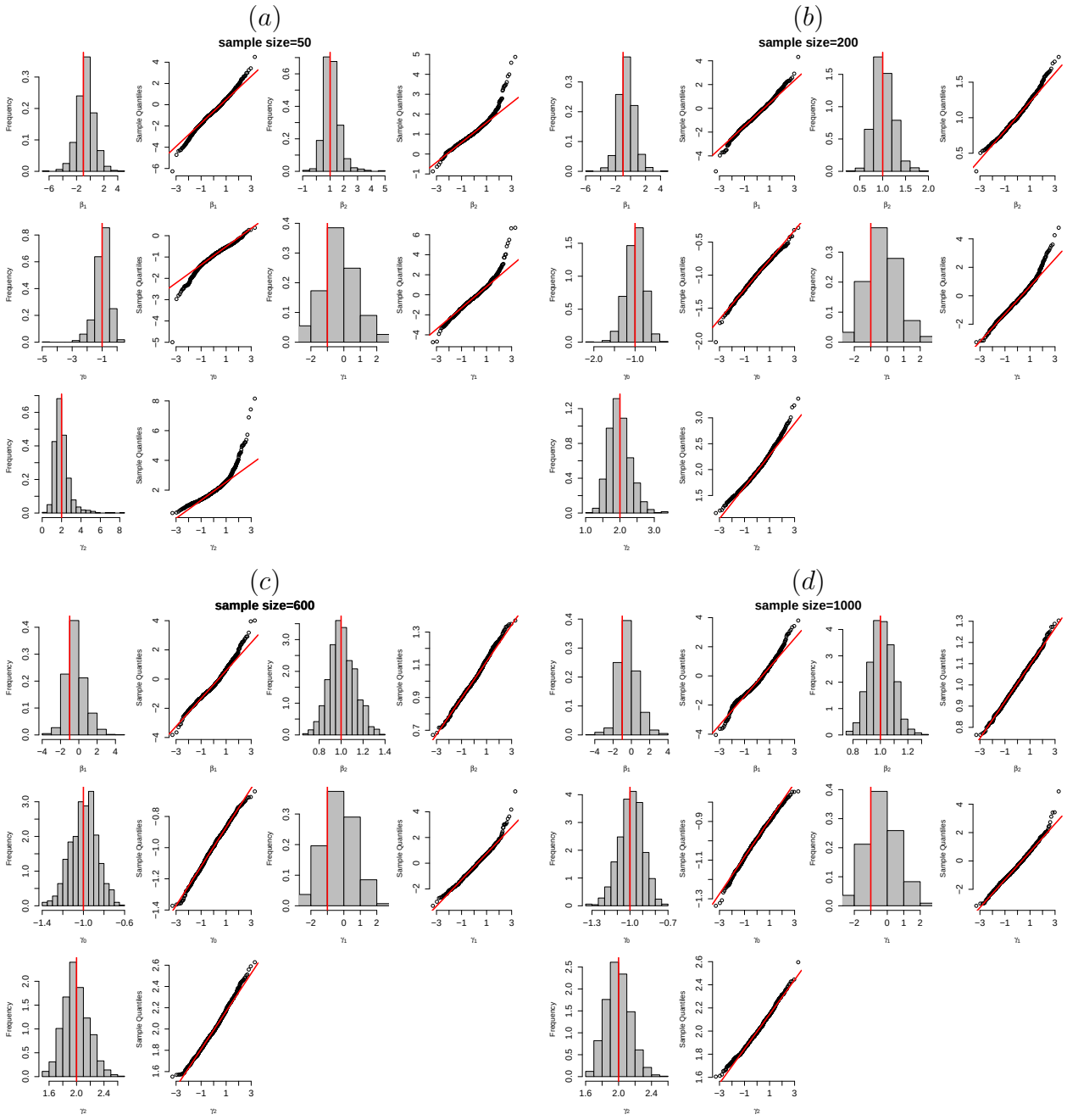


Figure B.2.2: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates (Scenario 2, latency part). Sample sizes: 50 ($a$), 200 ($b$), 600 ($c$) and 1,000 ($d$). The vertical red line in the histogram is the true value of the parameter.

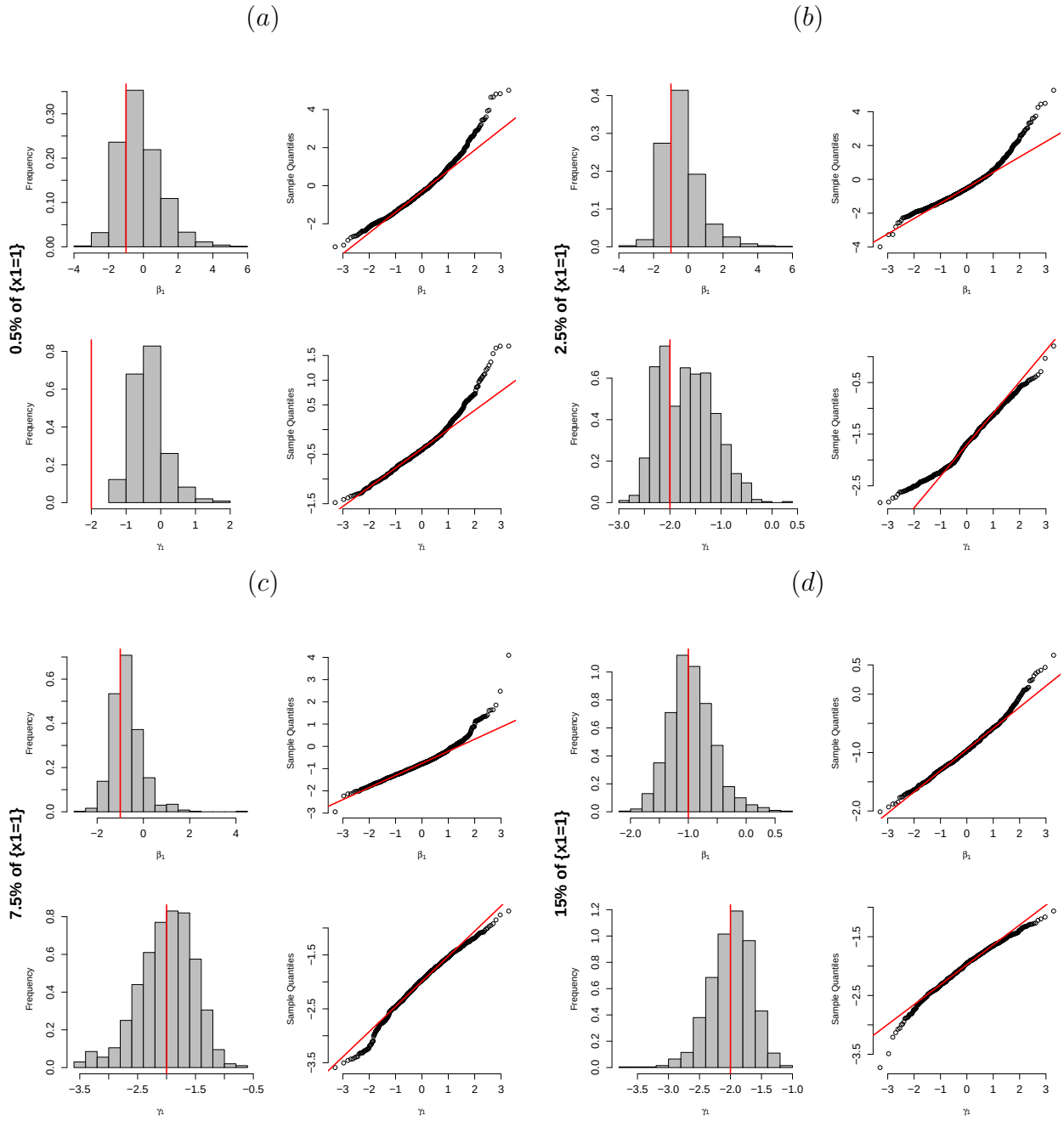


Figure B.2.3: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates for $\beta_1$ and $\gamma_1$ (Scenario 1). Percentages of $x_{1i} = 1$: 0.5% $(a)$, 2.5% $(b)$, 7.5% $(c)$ and 15% $(d)$. The vertical red line in the histogram is the true value of the parameter.

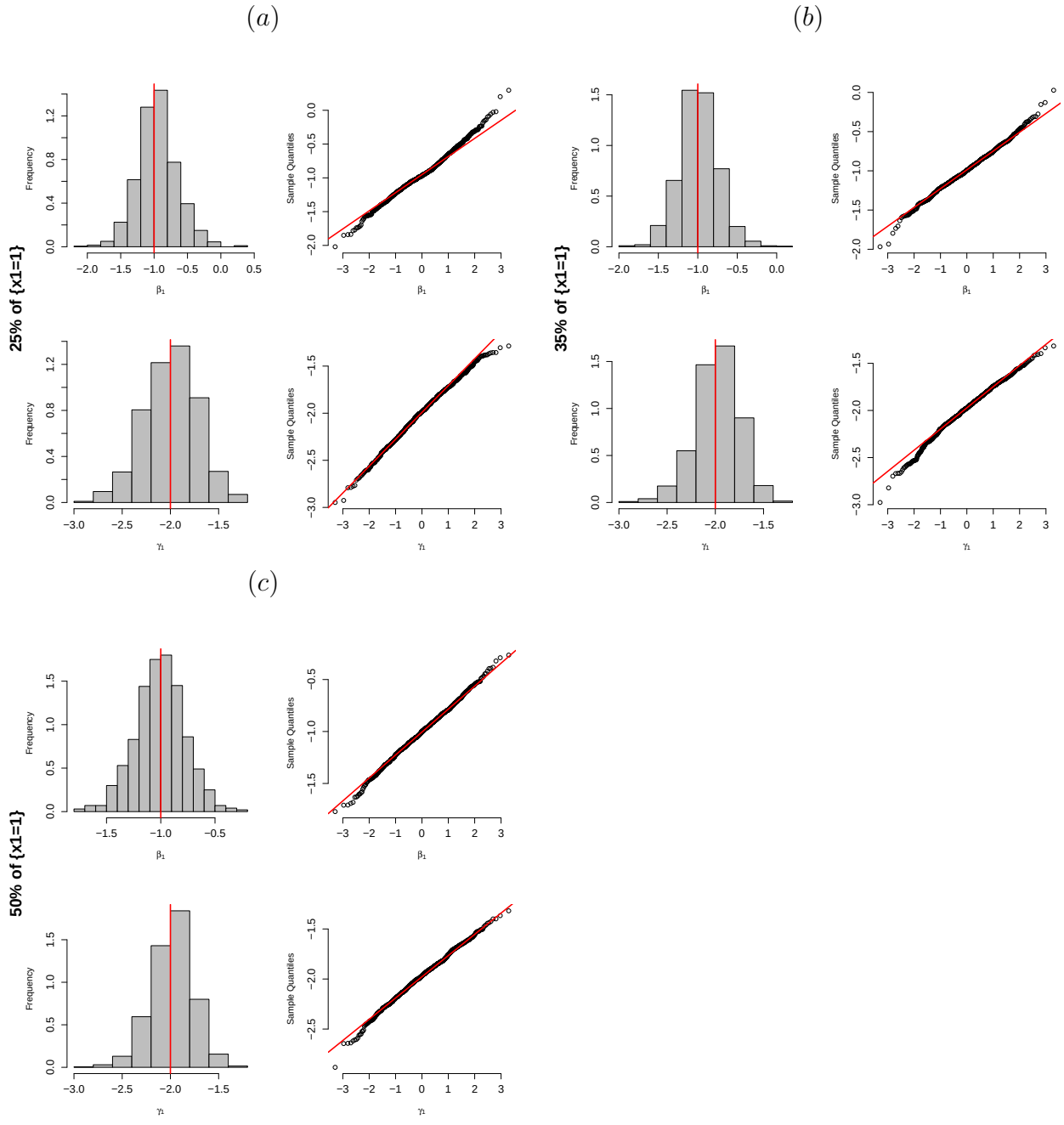


Figure B.2.4: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates for $\beta_1$ and $\gamma_1$ (Scenario 1). Percentages of $x_{1i} = 1$: 25% ($a$), 35% ($b$) and 50% ($c$). The vertical red line in the histogram is the true value of the parameter.

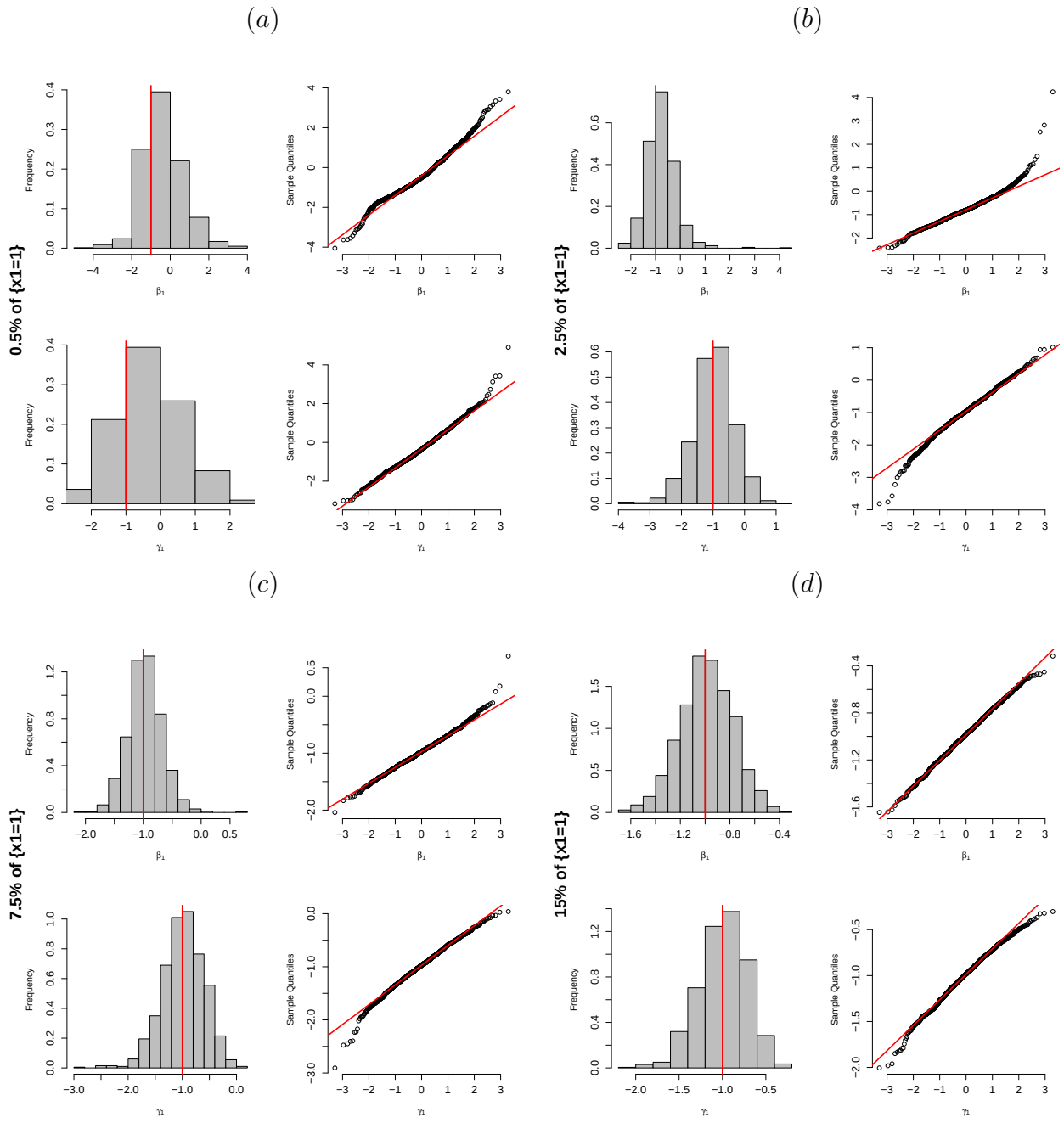


Figure B.2.5: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates for $\beta_1$ and $\gamma_1$ (Scenario 2). Percentages of $x_{1i} = 1$: 0.5% ($a$), 2.5% ($b$), 7.5% ($c$) and 15% ($d$). The vertical red line in the histogram is the true value of the parameter.

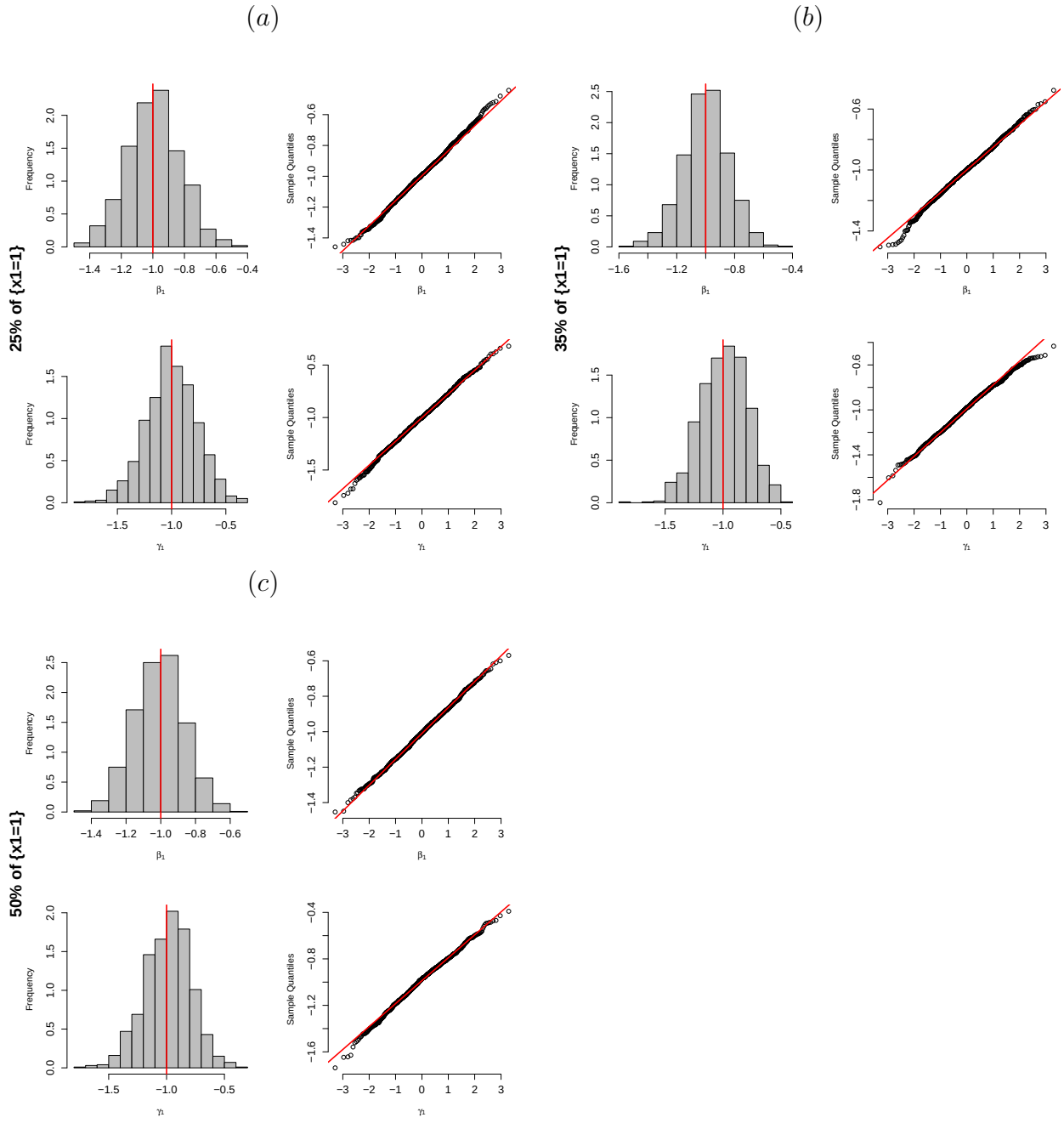


Figure B.2.6: Histogram and normal Q-Q plot of the penalized maximum likelihood estimates for $\beta_1$ and $\gamma_1$ (Scenario 2). Percentages of $x_{1i} = 1$: 25% ($a$), 35% ($b$) and 50% ($c$). The vertical red line in the histogram is the true value of the parameter.

## Appendix C. Additional Details of the Bayesian Approach

From the posterior density in (5.1), and under the augmented data likelihood function presented in the expression (3.8), the joint posterior density for the parametric mixture cure fraction model has the form:

$$p\left(\boldsymbol{\beta},\lambda,\alpha|\mathcal{D}_c\right)=\left\{\prod_{i=1}^{n}\left[\alpha\lambda t_i^{\alpha-1}\exp\left(\mathbf{x}_i^{\top}\boldsymbol{\beta}\right)\right]^{\delta_i}\exp\left(-Y_i\lambda t_i^{\alpha}\exp\left(\mathbf{x}_i^{\top}\boldsymbol{\beta}\right)\right)\right\}p\left(\boldsymbol{\beta}\right)p\left(\boldsymbol{\gamma}\right)p\left(\lambda\right)p\left(\alpha\right),$$

resulting in the following conditional posterior densities:

$$\begin{aligned}
p\left(\boldsymbol{\beta}|\lambda,\alpha,\mathcal{D}_c\right) &\propto \left\{\prod_{i=1}^{n}\exp\left(\delta_i\mathbf{x}_i^{\top}\boldsymbol{\beta}-Y_i\lambda t_i^{\alpha}\exp\left(\mathbf{x}_i^{\top}\boldsymbol{\beta}\right)\right)\right\}p\left(\boldsymbol{\beta}\right),\\
p\left(\lambda|\boldsymbol{\beta},\alpha,\mathcal{D}_c\right) &\propto \left\{\prod_{i=1}^{n}\lambda^{\delta_i}\exp\left(-Y_i\lambda t_i^{\alpha}\exp\left(\mathbf{x}_i^{\top}\boldsymbol{\beta}\right)\right)\right\}p\left(\lambda\right),\\
p\left(\alpha|\boldsymbol{\beta},\lambda,\mathcal{D}_c\right) &\propto \left\{\prod_{i=1}^{n}\left[\alpha t_i^{\alpha-1}\right]^{\delta_i}\exp\left(-Y_i\lambda t_i^{\alpha}\exp\left(\mathbf{x}_i^{\top}\boldsymbol{\beta}\right)\right)\right\}p\left(\alpha\right).
\end{aligned}$$

The full conditional posterior density, when using the semiparametric specification into the latency part is:

$$p\left(\boldsymbol{\beta}|\mathcal{D}_c\right)=\prod_{i=1}^{d^*}\left(\frac{\exp\left(\mathbf{x}_{(i)}^{\top}\boldsymbol{\beta}\right)}{\sum\limits_{j\in R_{(t_i)}}Y_j\exp\left(\mathbf{x}_j^{\top}\boldsymbol{\beta}\right)}\right)p\left(\boldsymbol{\beta}\right),$$

where $d^*$ denote the number of distinct uncensored times, given in the Section 2.2.2. For both cases, the full conditional posterior density for the coefficients vector $\boldsymbol{\gamma}$, related to the incidence part, has the form

$$p\left(\boldsymbol{\gamma}|\mathcal{D}_c\right)\propto\prod_{i=1}^{n}\left(\frac{\exp\left(Y_i\mathbf{z}_i^{\top}\boldsymbol{\gamma}\right)}{1+\exp\left(\mathbf{z}_i^{\top}\boldsymbol{\gamma}\right)}\right)p\left(\boldsymbol{\gamma}\right).$$

The independent prior distributions $p\left(\boldsymbol{\beta}\right)$, $p\left(\boldsymbol{\gamma}\right)$, $p\left(\lambda\right)$ and $p\left(\alpha\right)$ are there described in the Section 5.1. The partially observed random variable $Y_i$ has the posterior

distribution described in the Section 6.1.

This section displays some additional details of the MC simulation results obtained through the parametric mixture cure fraction models. The numerical values of the results in Figures 6.2 to 6.15 (Chapter 6) are summarized in the Tables C.1.1 to C.2.4.In Tables C.1.5–C.1.6 and C.2.5–C.2.6, we illustrate the numerical summary containing the values of the performance measurements obtained by assuming different configurations of the dichotomous covariate $x_{1i}$. That is, the new cases account for the following percentages of $\{x_{1i} = 1\}$, $i = 1, 2, \ldots, n$: 0.5% (highly unbalanced), 2.5%, 7.5%, 15%, 25%, 35% and 50%.

**Appendix C.1. Results for the Parametric Mixture Cure Fraction Model**

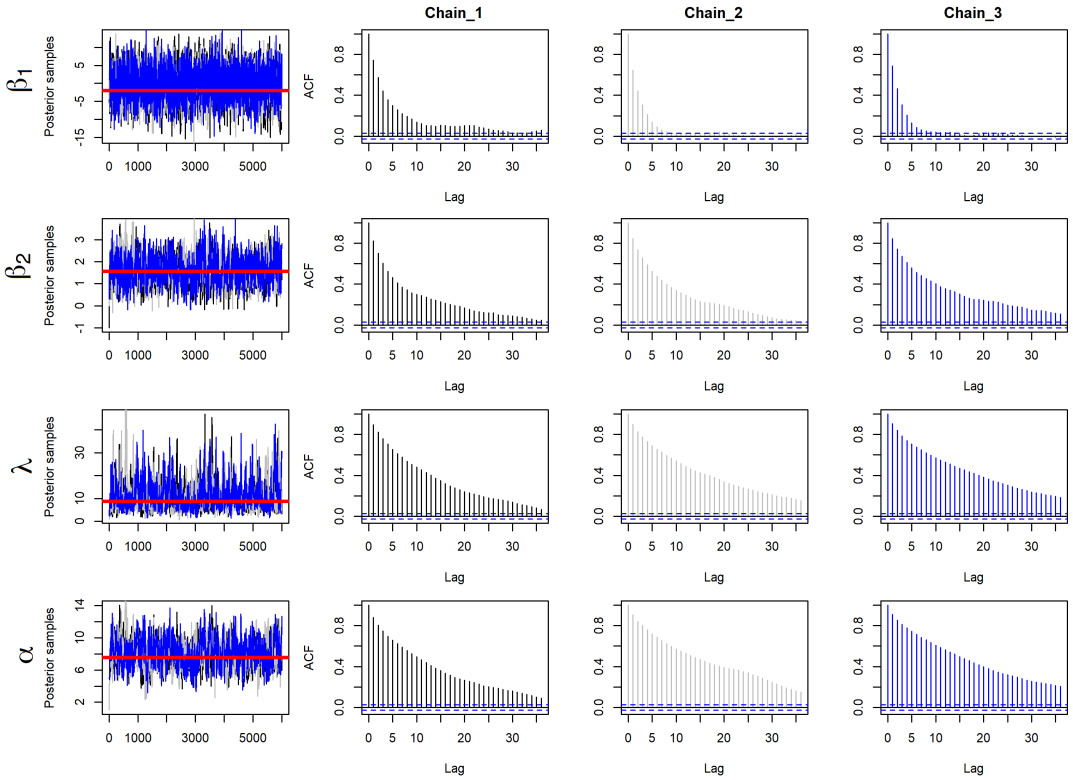


Figure C.1.1: Traceplots and ACF graphs for the Markov chain of the parameters related to the latency part. The SC1 is considered under the `Bayes_1` setting, for a single sample size ($n = 50$). The red line indicates the true value of the corresponding parameter.

The Figures C.1.1 to C.1.4, shows clearly that, the sampling method is mixing well as the whole of the posterior distribution is being visited in a short period, i.e., the considered

posterior distribution converge to the equilibrium distribution. In generally, high values for the ACF measure are observed in the first lags, especially for coefficients related to the latency part, when using the parametric specification.

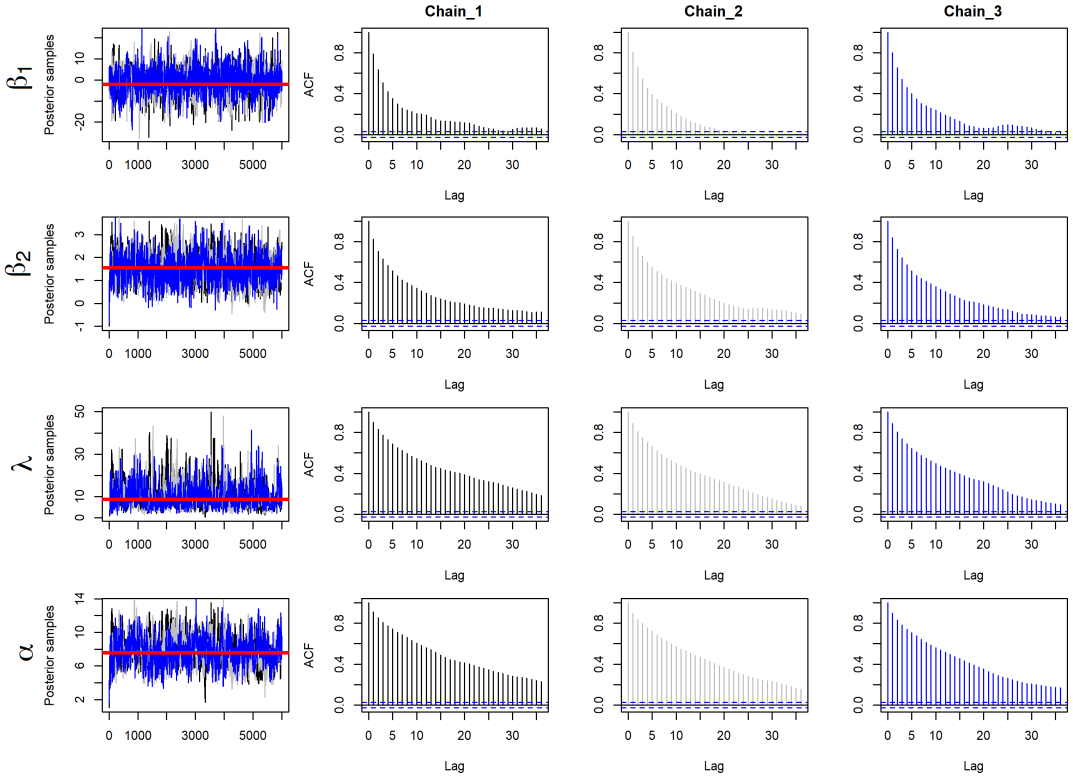


Figure C.1.2: Traceplots and ACF graphs for the Markov chain of the parameters related to the latency part. The SC1 is considered under the `Bayes_3` setting, for a single sample size ($n = 50$). The red line indicates the true value of the corresponding parameter.

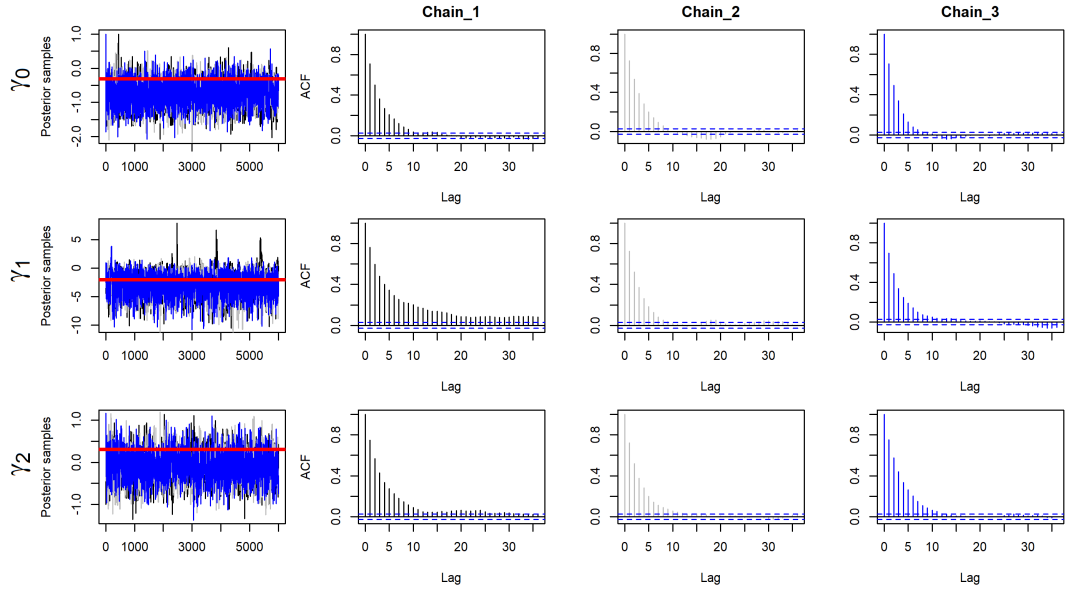


Figure C.1.3: Traceplots and ACF graphs for the Markov chain of the parameters related to the incidence part. The SC1 is considered under the `Bayes_1` setting, for a single sample size ($n = 50$). The red line indicates the true value of the corresponding parameter.

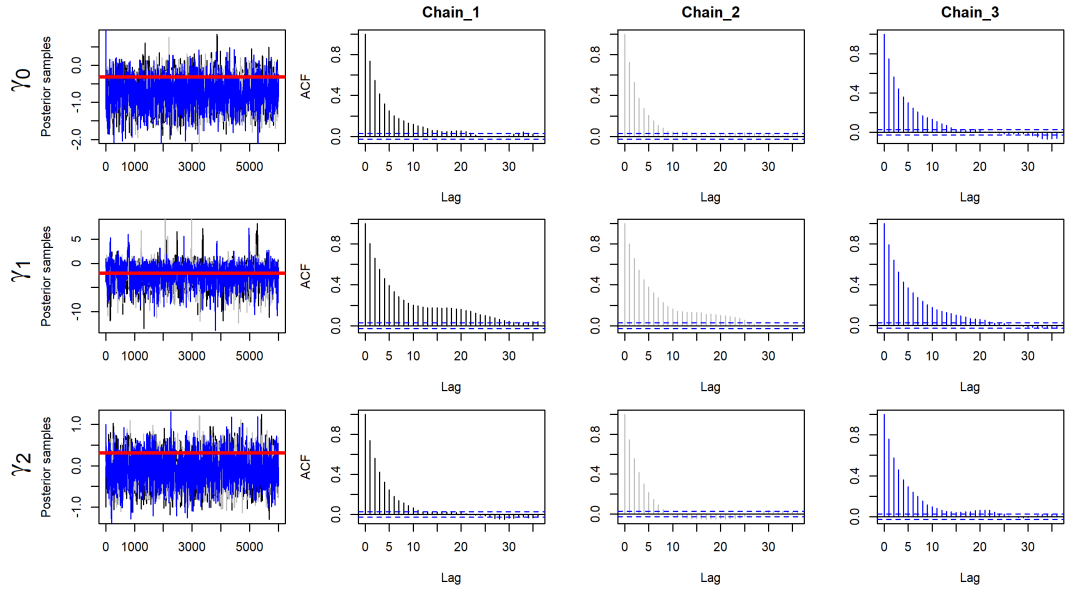


Figure C.1.4: Traceplots and ACF graphs for the Markov chain for the parameters related to the incidence part. The SC1 is considered under the `Bayes_3` setting, for a single sample size ($n = 50$). The red line indicates the true value of the corresponding parameter.

Table C.1.1: *MC estimates (frequentist and Bayesian) and performance measurements for the coefficients related to the ML issue in both part of the model ($\beta_1$ and $\gamma_1$) based on the SC1. The real values are shown near the parameter name.*

| **Scenario 1** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_1 = -2.00$ | | | | | | $\gamma_1 = -2.00$ | | | | | |
| $n$ | | Mean $\hat{\beta}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 50 | Firth | -1.129 | 43.532 | 1.038 | 1.524 | 1.755 | 78.400 | -0.322 | 83.879 | 1.306 | 0.686 | 1.812 | 78.400 |
| | Bayes_1 | -1.213 | 39.348 | – | 0.883 | 1.183 | 98.900 | -2.027 | -1.374 | – | 1.248 | 1.248 | 97.300 |
| | Bayes_2 | -2.002 | -0.114 | – | 0.245 | 0.244 | 99.900 | -2.085 | -4.273 | – | 0.387 | 0.396 | 99.700 |
| | Bayes_3 | -1.602 | 19.888 | – | 1.069 | 1.140 | 98.100 | -1.589 | 20.544 | – | 1.128 | 1.200 | 96.100 |
| 200 | Firth | -1.093 | 45.352 | 0.868 | 1.355 | 1.630 | 74.700 | -0.290 | 85.509 | 1.223 | 0.555 | 1.798 | 90.400 |
| | Bayes_1 | -1.145 | 42.765 | – | 0.901 | 1.242 | 98.700 | -2.123 | -6.139 | – | 1.168 | 1.174 | 97.300 |
| | Bayes_2 | -2.009 | -0.429 | – | 0.248 | 0.248 | 99.900 | -2.083 | -4.167 | – | 0.382 | 0.391 | 99.900 |
| | Bayes_3 | -1.531 | 23.444 | – | 1.346 | 1.425 | 98.400 | -1.659 | 17.052 | – | 0.985 | 1.042 | 96.200 |
| 600 | Firth | -1.001 | 49.931 | 0.836 | 1.282 | 1.625 | 72.800 | -0.307 | 84.670 | 1.201 | 0.511 | 1.769 | 90.300 |
| | Bayes_1 | -1.092 | 45.420 | – | 0.903 | 1.281 | 98.200 | -2.133 | -6.636 | – | 1.206 | 1.213 | 97.700 |
| | Bayes_2 | -1.999 | 0.072 | – | 0.238 | 0.238 | 99.900 | -2.071 | -3.540 | – | 0.381 | 0.387 | 99.900 |
| | Bayes_3 | -1.527 | 23.648 | – | 1.177 | 1.268 | 98.700 | -1.660 | 16.997 | – | 1.340 | 1.382 | 93.800 |
| 1,000 | Firth | -1.003 | 49.838 | 0.829 | 1.395 | 1.714 | 72.600 | -0.278 | 86.119 | 1.208 | 0.547 | 1.807 | 88.900 |
| | Bayes_1 | -1.175 | 41.251 | – | 0.926 | 1.240 | 98.300 | -2.065 | -3.266 | – | 1.226 | 1.227 | 96.700 |
| | Bayes_2 | -1.996 | 0.206 | – | 0.268 | 0.268 | 99.600 | -2.072 | -3.594 | – | 0.384 | 0.391 | 99.900 |
| | Bayes_3 | -1.482 | 25.915 | – | 1.010 | 1.135 | 98.200 | -1.687 | 15.657 | – | 0.951 | 1.001 | 97.300 |

Table C.1.2: *MC estimates (frequentist and Bayesian) and performance measurements for the coefficients not directly related to the ML issue based on the SC1. The real values are shown near the parameter name.*

**Scenario 1**

| | | $\beta_2 = 1.56$ | | | | | | $\lambda = 8.79$ | | | | | | $\alpha = 7.57$ | | | | | | $\gamma_0 = -0.31$ | | | | | | $\gamma_2 = 0.31$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | | Mean $\hat{\beta}_2$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\lambda}$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\alpha}$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_0$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_2$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 50 | Firth | 1.691 | 8.657 | 0.489 | 0.535 | 0.551 | 95.600 | 10.921 | 24.295 | 5.532 | 7.778 | 8.065 | 91.400 | 8.023 | 5.974 | 1.560 | 1.738 | 1.796 | 92.800 | -0.294 | 5.484 | 0.381 | 0.364 | 0.364 | 98.100 | 0.307 | -1.169 | 0.396 | 0.375 | 0.375 | 97.800 |
| | Bayes_1 | 1.594 | 2.183 | – | 0.498 | 0.499 | 95.500 | 9.519 | 8.291 | – | 3.773 | 3.841 | 95.300 | 7.583 | 0.175 | – | 1.384 | 1.384 | 95.700 | -0.374 | -20.750 | – | 0.423 | 0.427 | 94.200 | 0.376 | 21.288 | – | 0.493 | 0.497 | 94.300 |
| | Bayes_2 | 1.583 | 1.466 | – | 0.344 | 0.344 | 98.400 | 9.487 | 7.929 | – | 3.535 | 3.601 | 95.800 | 7.574 | 0.055 | – | 1.176 | 1.175 | 97.700 | -0.338 | -9.136 | – | 0.339 | 0.340 | 95.900 | 0.325 | 4.784 | – | 0.397 | 0.397 | 95.500 |
| | Bayes_3 | 1.563 | 0.172 | | 0.451 | 0.451 | 94.900 | 9.012 | 2.524 | – | 4.161 | 4.165 | 88.700 | 7.421 | -1.967 | – | 1.383 | 1.391 | 93.700 | -0.374 | -20.521 | – | 0.391 | 0.396 | 95.700 | 0.373 | 20.337 | – | 0.471 | 0.475 | 94.500 |
| 200 | Firth | 1.594 | 2.470 | 0.211 | 0.214 | 0.218 | 94.600 | 9.326 | 6.150 | 2.194 | 2.450 | 2.509 | 95.700 | 7.714 | 1.883 | 0.757 | 0.770 | 0.784 | 96.100 | -0.309 | 0.529 | 0.177 | 0.180 | 0.180 | 94.800 | 0.308 | -0.909 | 0.182 | 0.175 | 0.175 | 96.100 |
| | Bayes_1 | 1.583 | 1.459 | – | 0.210 | 0.211 | 94.500 | 9.158 | 4.181 | – | 2.149 | 2.179 | 95.200 | 7.633 | 0.838 | – | 0.749 | 0.751 | 94.500 | -0.327 | -5.451 | – | 0.177 | 0.178 | 95.000 | 0.320 | 3.211 | – | 0.193 | 0.193 | 94.700 |
| | Bayes_2 | 1.596 | 2.292 | – | 0.203 | 0.206 | 94.700 | 9.324 | 6.078 | – | 2.184 | 2.248 | 95.900 | 7.669 | 1.308 | – | 0.731 | 0.737 | 95.200 | -0.309 | 0.369 | – | 0.176 | 0.176 | 93.700 | 0.320 | 3.209 | – | 0.178 | 0.179 | 96.000 |
| | Bayes_3 | 1.574 | 0.888 | – | 0.212 | 0.212 | 92.900 | 9.160 | 4.205 | – | 2.188 | 2.218 | 93.400 | 7.608 | 0.500 | – | 0.751 | 0.752 | 93.600 | -0.319 | -2.831 | – | 0.174 | 0.174 | 95.100 | 0.313 | 1.114 | – | 0.190 | 0.190 | 94.700 |
| 600 | Firth | 1.560 | 0.286 | 0.117 | 0.119 | 0.119 | 94.100 | 8.937 | 1.721 | 1.174 | 1.176 | 1.186 | 95.200 | 7.607 | 0.478 | 0.429 | 0.430 | 0.431 | 95.600 | -0.312 | -0.415 | 0.100 | 0.102 | 0.102 | 95.400 | 0.321 | 3.172 | 0.103 | 0.106 | 0.106 | 94.700 |
| | Bayes_1 | 1.569 | 0.564 | – | 0.119 | 0.119 | 94.400 | 8.934 | 1.633 | – | 1.230 | 1.237 | 94.400 | 7.587 | 0.230 | – | 0.437 | 0.437 | 93.800 | -0.316 | -2.028 | – | 0.103 | 0.103 | 94.800 | 0.315 | 1.659 | – | 0.106 | 0.106 | 94.200 |
| | Bayes_2 | 1.567 | 0.478 | – | 0.114 | 0.114 | 95.800 | 8.935 | 1.655 | – | 1.227 | 1.235 | 93.300 | 7.596 | 0.349 | – | 0.426 | 0.427 | 94.500 | -0.311 | -0.381 | – | 0.099 | 0.099 | 95.500 | 0.316 | 2.026 | – | 0.103 | 0.103 | 95.300 |
| | Bayes_3 | 1.560 | 0.010 | – | 0.123 | 0.123 | 91.300 | 8.862 | 0.820 | – | 1.183 | 1.185 | 93.500 | 7.570 | 0.005 | – | 0.444 | 0.443 | 89.500 | -0.313 | -1.059 | – | 0.099 | 0.099 | 95.100 | 0.312 | 0.596 | – | 0.105 | 0.105 | 91.500 |
| 1,000 | Firth | 1.568 | 0.760 | 0.090 | 0.091 | 0.091 | 95.500 | 8.901 | 1.307 | 0.899 | 0.918 | 0.926 | 95.000 | 7.609 | 0.507 | 0.332 | 0.336 | 0.338 | 95.900 | -0.312 | -0.274 | 0.077 | 0.078 | 0.078 | 95.400 | 0.311 | -0.061 | 0.080 | 0.082 | 0.082 | 94.300 |
| | Bayes_1 | 1.567 | 0.436 | – | 0.090 | 0.090 | 95.600 | 8.885 | 1.083 | – | 0.862 | 0.867 | 96.600 | 7.589 | 0.257 | – | 0.329 | 0.330 | 94.300 | -0.310 | -0.134 | – | 0.081 | 0.081 | 94.600 | 0.308 | -0.598 | – | 0.084 | 0.084 | 92.800 |
| | Bayes_2 | 1.564 | 0.236 | – | 0.088 | 0.088 | 95.400 | 8.879 | 1.009 | – | 0.900 | 0.904 | 94.400 | 7.597 | 0.362 | – | 0.327 | 0.328 | 95.800 | -0.314 | -1.152 | – | 0.075 | 0.075 | 95.300 | 0.310 | 0.009 | – | 0.080 | 0.080 | 94.900 |
| | Bayes_3 | 1.560 | 0.029 | – | 0.092 | 0.092 | 91.300 | 8.828 | 0.430 | – | 0.848 | 0.849 | 95.500 | 7.574 | 0.046 | – | 0.327 | 0.327 | 93.400 | -0.307 | 0.835 | – | 0.080 | 0.080 | 92.600 | 0.311 | 0.286 | – | 0.080 | 0.080 | 93.700 |

Table C.1.3: *MC estimates (frequentist and Bayesian) and performance measurements for the coefficients related to the ML issue in both part of the model ($\beta_1$ and $\gamma_1$) based on the SC2. The real values are shown near the parameter name.*

| **Scenario 2** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_1 = -2.00$ | | | | | | $\gamma_1 = -1.00$ | | | | | |
| $n$ | | Mean $\hat{\beta}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 50 | Firth | -1.345 | 32.750 | 0.952 | 1.367 | 1.516 | 81.600 | -0.289 | 71.052 | 1.745 | 1.418 | 1.586 | 98.500 |
| | Bayes_1 | -1.606 | 19.708 | – | 1.176 | 1.239 | 97.400 | -0.965 | 3.502 | – | 1.542 | 1.541 | 97.900 |
| | Bayes_2 | -2.012 | -0.589 | – | 0.383 | 0.383 | 99.900 | -1.026 | -2.596 | – | 0.444 | 0.444 | 99.900 |
| | Bayes_3 | -1.646 | 17.713 | – | 1.115 | 1.170 | 96.400 | -0.837 | 16.326 | – | 1.201 | 1.211 | 98.300 |
| 200 | Firth | -1.252 | 37.408 | 0.799 | 1.145 | 1.368 | 78.800 | -0.257 | 74.316 | 1.698 | 1.304 | 1.501 | 98.500 |
| | Bayes_1 | -1.560 | 22.020 | – | 1.122 | 1.205 | 95.700 | -1.056 | -5.634 | – | 1.499 | 1.500 | 97.500 |
| | Bayes_2 | -1.982 | 0.886 | – | 0.403 | 0.403 | 99.500 | -1.025 | -2.509 | – | 0.441 | 0.441 | 99.900 |
| | Bayes_3 | -1.638 | 18.088 | – | 1.643 | 1.681 | 95.900 | -0.782 | 21.820 | – | 1.173 | 1.192 | 98.700 |
| 600 | Firth | -1.176 | 41.212 | 0.762 | 1.195 | 1.452 | 77.700 | -0.310 | 69.005 | 1.580 | 1.230 | 1.411 | 97.900 |
| | Bayes_1 | -1.527 | 23.669 | – | 1.116 | 1.211 | 97.600 | -1.075 | -7.452 | – | 1.512 | 1.486 | 98.100 |
| | Bayes_2 | -1.962 | 1.904 | – | 0.398 | 0.400 | 99.600 | -1.041 | -4.105 | – | 0.427 | 0.429 | 99.900 |
| | Bayes_3 | -1.660 | 16.983 | – | 1.124 | 1.174 | 96.100 | -0.737 | 26.280 | – | 1.334 | 1.359 | 94.200 |
| 1,000 | Firth | -1.175 | 41.248 | 0.752 | 1.198 | 1.454 | 74.900 | -0.269 | 73.067 | 1.596 | 1.226 | 1.427 | 97.297 |
| | Bayes_1 | -1.567 | 21.649 | – | 1.072 | 1.156 | 95.600 | -0.991 | 0.886 | – | 1.505 | 1.504 | 97.800 |
| | Bayes_2 | -1.974 | 1.277 | – | 0.393 | 0.393 | 99.600 | -0.996 | 0.404 | – | 0.441 | 0.441 | 99.990 |
| | Bayes_3 | -1.636 | 18.200 | – | 1.173 | 1.227 | 95.600 | -0.821 | 17.867 | – | 1.162 | 1.175 | 99.300 |

Table C.1.4: *MC estimates (frequentist and Bayesian) and performance measurements for the coefficients not directly related to the ML issue based on the SC2. The real values are shown near the parameter name.*

**Scenario 2**

| | | $\beta_2 = 2.20$ | | | | | | $\lambda = 8.79$ | | | | | | $\alpha = 7.57$ | | | | | | $\gamma_0 = -0.73$ | | | | | | $\gamma_2 = 2.20$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | | Mean $\hat{\beta}_2$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\lambda}$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\alpha}$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_0$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_2$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 50 | Firth | 2.524 | 14.723 | 0.654 | 0.737 | 0.805 | 93.900 | 9.626 | 9.510 | 5.088 | 7.895 | 7.939 | 83.600 | 8.010 | 5.813 | 1.526 | 1.734 | 1.789 | 93.300 | -0.669 | 8.292 | 0.511 | 0.482 | 0.485 | 97.100 | 2.187 | -0.572 | 0.782 | 0.805 | 0.806 | 94.200 |
| | Bayes_1 | 2.362 | 7.355 | – | 0.662 | 0.681 | 95.900 | 9.614 | 9.377 | – | 3.882 | 3.966 | 96.000 | 7.736 | 2.187 | – | 1.374 | 1.383 | 97.300 | -0.862 | -18.090 | – | 0.577 | 0.592 | 95.100 | 2.720 | 23.645 | – | 0.920 | 1.056 | 95.600 |
| | Bayes_2 | 2.292 | 4.177 | – | 0.428 | 0.437 | 99.100 | 9.179 | 4.429 | – | 3.426 | 3.446 | 96.600 | 7.532 | -0.508 | – | 1.073 | 1.073 | 97.600 | -0.756 | -3.505 | – | 0.395 | 0.395 | 98.000 | 2.372 | 7.812 | – | 0.470 | 0.500 | 99.200 |
| | Bayes_3 | 2.285 | 3.863 | – | 0.662 | 0.668 | 96.300 | 8.281 | -5.792 | – | 3.660 | 3.693 | 85.600 | 7.478 | -1.218 | – | 1.371 | 1.374 | 94.700 | -0.798 | -9.345 | – | 0.548 | 0.552 | 95.400 | 2.521 | 14.576 | – | 0.863 | 0.920 | 95.800 |
| 200 | Firth | 2.264 | 2.901 | 0.283 | 0.295 | 0.301 | 94.900 | 9.039 | 2.834 | 2.372 | 2.582 | 2.594 | 93.900 | 7.673 | 1.366 | 0.738 | 0.750 | 0.757 | 95.500 | -0.722 | 1.119 | 0.241 | 0.231 | 0.231 | 95.900 | 2.195 | -0.213 | 0.370 | 0.369 | 0.369 | 95.100 |
| | Bayes_1 | 2.245 | 2.038 | – | 0.283 | 0.286 | 95.700 | 9.598 | 9.188 | – | 2.697 | 2.814 | 96.300 | 7.703 | 1.754 | – | 0.745 | 0.756 | 96.500 | -0.765 | -4.832 | – | 0.252 | 0.255 | 94.300 | 2.314 | 5.160 | – | 0.400 | 0.415 | 94.400 |
| | Bayes_2 | 2.244 | 1.990 | – | 0.259 | 0.263 | 96.500 | 9.432 | 7.302 | – | 2.552 | 2.631 | 96.300 | 7.665 | 1.260 | – | 0.700 | 0.706 | 95.700 | -0.769 | -5.407 | – | 0.235 | 0.238 | 94.900 | 2.326 | 5.725 | – | 0.344 | 0.336 | 95.500 |
| | Bayes_3 | 2.233 | 1.490 | – | 0.284 | 0.285 | 95.800 | 9.347 | 6.337 | – | 2.534 | 2.593 | 95.700 | 7.635 | 0.855 | – | 0.739 | 0.742 | 95.900 | -0.750 | -2.735 | – | 0.252 | 0.253 | 93.200 | 2.267 | 3.026 | – | 0.373 | 0.378 | 96.000 |
| 600 | Firth | 2.216 | 0.730 | 0.157 | 0.155 | 0.156 | 95.400 | 8.850 | 0.678 | 1.318 | 1.339 | 1.341 | 94.200 | 7.595 | 0.331 | 0.423 | 0.427 | 0.427 | 95.700 | -0.721 | 1.213 | 0.136 | 0.135 | 0.136 | 95.400 | 2.196 | -0.162 | 0.212 | 0.213 | 0.213 | 95.000 |
| | Bayes_1 | 2.217 | 0.784 | – | 0.158 | 0.158 | 96.300 | 9.123 | 3.789 | – | 1.401 | 1.439 | 96.200 | 7.632 | 0.823 | – | 0.428 | 0.432 | 96.400 | -0.737 | -0.989 | – | 0.141 | 0.141 | 94.400 | 2.243 | 1.949 | – | 0.212 | 0.216 | 95.600 |
| | Bayes_2 | 2.216 | 0.723 | – | 0.152 | 0.153 | 95.700 | 9.056 | 3.027 | – | 1.354 | 1.379 | 96.800 | 7.617 | 0.619 | – | 0.414 | 0.417 | 96.500 | -0.738 | -1.115 | – | 0.134 | 0.134 | 95.400 | 2.245 | 2.054 | – | 0.204 | 0.209 | 95.400 |
| | Bayes_3 | 2.211 | 0.511 | – | 0.165 | 0.165 | 91.600 | 8.964 | 1.975 | – | 1.383 | 1.393 | 92.400 | 7.589 | 0.246 | – | 0.437 | 0.437 | 91.500 | -0.741 | -1.568 | – | 0.140 | 0.141 | 95.200 | 2.224 | 1.083 | – | 0.223 | 0.224 | 93.200 |
| 1,000 | Firth | 2.210 | 0.466 | 0.121 | 0.127 | 0.127 | 93.600 | 8.751 | -0.447 | 1.004 | 0.993 | 0.993 | 95.200 | 7.575 | 0.069 | 0.326 | 0.333 | 0.333 | 94.100 | -0.730 | -0.031 | 0.106 | 0.105 | 0.105 | 95.300 | 2.195 | -0.228 | 0.164 | 0.165 | 0.165 | 94.900 |
| | Bayes_1 | 2.214 | 0.616 | – | 0.124 | 0.125 | 94.700 | 8.948 | 1.793 | – | 1.007 | 1.019 | 96.300 | 7.602 | 0.422 | – | 0.328 | 0.330 | 95.700 | -0.742 | -1.630 | – | 0.112 | 0.112 | 93.300 | 2.234 | 1.557 | – | 0.171 | 0.174 | 93.400 |
| | Bayes_2 | 2.212 | 0.543 | – | 0.125 | 0.126 | 94.800 | 8.903 | 1.290 | – | 1.056 | 1.062 | 95.200 | 7.601 | 0.404 | – | 0.336 | 0.337 | 95.000 | -0.738 | -1.076 | – | 0.106 | 0.106 | 94.000 | 2.223 | 1.024 | – | 0.164 | 0.165 | 94.700 |
| | Bayes_3 | 2.202 | 0.097 | – | 0.129 | 0.129 | 88.900 | 8.878 | 1.003 | – | 1.047 | 1.050 | 94.500 | 7.570 | -0.001 | – | 0.337 | 0.336 | 92.600 | -0.734 | -0.510 | – | 0.103 | 0.103 | 93.400 | 2.221 | 0.974 | – | 0.165 | 0.166 | 95.100 |

Table C.1.5: MC estimates (frequentist and Bayesian) and performance measurements for the coefficients ($\beta_1$ and $\gamma_1$) related to the binary covariate ($x_{1i}$, $i = 1, 2, \ldots, n$) for both estimation methods. The sample size is fixed ($n$ = 1,000). Different percentages of ones in $x_{1i}$ are explored.

| **Scenario 1** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_1 = -2.00$ | | | | | | $\gamma_1 = -2.00$ | | | | | |
| $\%\{x_{1i} = 1\}$ | | Mean $\hat{\beta}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 0.5 | Firth | -1.003 | 49.838 | 0.829 | 1.395 | 1.714 | 72.600 | -0.278 | 86.119 | 1.208 | 0.547 | 1.807 | 88.900 |
| | Bayes_1 | -1.175 | 41.251 | – | 0.926 | 1.240 | 98.300 | -2.065 | -3.266 | – | 1.226 | 1.227 | 96.700 |
| | Bayes_2 | -1.996 | 0.206 | – | 0.268 | 0.268 | 99.600 | -2.072 | -3.594 | – | 0.384 | 0.391 | 99.900 |
| | Bayes_3 | -1.482 | 25.915 | – | 1.010 | 1.135 | 98.200 | -1.687 | 15.657 | – | 0.951 | 1.001 | 97.300 |
| 2.5 | Firth | -1.370 | 31.505 | 0.714 | 1.077 | 1.248 | 80.200 | -1.667 | 16.630 | 0.763 | 0.494 | 0.595 | 94.600 |
| | Bayes_1 | -1.772 | 11.402 | – | 1.029 | 1.053 | 95.600 | -2.379 | -18.932 | – | 1.039 | 1.105 | 97.100 |
| | Bayes_2 | -1.985 | 0.748 | – | 0.425 | 0.425 | 98.900 | -2.136 | -6.817 | – | 0.499 | 0.517 | 99.000 |
| | Bayes_3 | -1.642 | 17.912 | – | 1.771 | 1.806 | 89.500 | -2.144 | -7.219 | – | 1.044 | 1.054 | 92.700 |
| 7.5 | Firth | -1.706 | 14.723 | 0.494 | 0.631 | 0.696 | 87.700 | -1.968 | 1.624 | 0.503 | 0.482 | 0.483 | 95.200 |
| | Bayes_1 | -1.949 | 2.526 | – | 0.605 | 0.607 | 94.000 | -2.147 | -7.338 | – | 0.601 | 0.619 | 96.700 |
| | Bayes_2 | -1.987 | 0.631 | – | 0.408 | 0.408 | 97.800 | -2.100 | -4.989 | – | 0.433 | 0.444 | 96.300 |
| | Bayes_3 | -1.935 | 3.238 | – | 0.702 | 0.705 | 92.700 | -2.101 | -5.039 | – | 0.567 | 0.576 | 96.400 |
| 15 | Firth | -1.881 | 5.932 | 0.364 | 0.394 | 0.411 | 92.900 | -2.012 | -0.593 | 0.363 | 0.365 | 0.365 | 96.500 |
| | Bayes_1 | -2.020 | -0.982 | – | 0.384 | 0.385 | 94.400 | -2.061 | -3.061 | – | 0.383 | 0.388 | 95.100 |
| | Bayes_2 | -2.011 | -0.531 | – | 0.321 | 0.321 | 97.200 | -2.081 | -4.045 | – | 0.339 | 0.348 | 96.500 |
| | Bayes_3 | -1.971 | 1.433 | – | 0.392 | 0.393 | 94.400 | -2.065 | -3.235 | – | 0.382 | 0.387 | 96.300 |
| 25 | Firth | -1.933 | 3.367 | 0.290 | 0.291 | 0.298 | 94.600 | -1.986 | 0.699 | 0.285 | 0.284 | 0.285 | 95.000 |
| | Bayes_1 | -1.989 | 0.566 | – | 0.305 | 0.305 | 94.300 | -2.040 | -1.995 | – | 0.310 | 0.312 | 93.100 |
| | Bayes_2 | -2.014 | -0.691 | – | 0.274 | 0.274 | 94.800 | -2.034 | -1.707 | – | 0.272 | 0.274 | 96.100 |
| | Bayes_3 | -1.997 | 0.127 | – | 0.306 | 0.305 | 94.300 | -2.031 | -1.568 | – | 0.286 | 0.288 | 94.800 |
| 35 | Firth | -1.961 | 1.935 | 0.258 | 0.275 | 0.277 | 93.400 | -2.008 | -0.423 | 0.249 | 0.242 | 0.242 | 95.900 |
| | Bayes_1 | -2.007 | -0.372 | – | 0.273 | 0.273 | 94.200 | -2.043 | -2.154 | – | 0.257 | 0.261 | 95.100 |
| | Bayes_2 | -2.006 | -0.299 | – | 0.245 | 0.245 | 95.700 | -2.033 | -1.662 | – | 0.250 | 0.252 | 94.000 |
| | Bayes_3 | -1.984 | 0.791 | – | 0.251 | 0.251 | 95.200 | -2.022 | -1.112 | – | 0.249 | 0.250 | 95.000 |
| 50 | Firth | -1.979 | 1.027 | 0.234 | 0.241 | 0.242 | 94.300 | -2.001 | -0.054 | 0.221 | 0.224 | 0.224 | 94.900 |
| | Bayes_1 | -2.005 | -0.272 | – | 0.238 | 0.238 | 93.400 | -2.010 | -0.480 | – | 0.229 | 0.229 | 94.900 |
| | Bayes_2 | -2.001 | -0.060 | – | 0.224 | 0.224 | 93.400 | -2.019 | -0.943 | – | 0.218 | 0.219 | 94.000 |
| | Bayes_3 | -1.989 | 0.566 | – | 0.242 | 0.242 | 94.500 | -2.018 | -0.876 | – | 0.217 | 0.217 | 94.300 |

Table C.1.6: MC estimates (frequentist and Bayesian) and performance measurements for the coefficients ($\beta_1$ and $\gamma_1$) related to the binary covariate ($x_{1i}$, $i = 1, 2, \ldots, n$) for both estimation methods. The sample size is fixed ($n$ = 1,000). Different percentages of ones in $x_{1i}$ are explored.

| **Scenario 2** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_1 = -2.00$ | | | | | | $\gamma_1 = -1.00$ | | | | | |
| $\%\{x_{1i} = 1\}$ | | Mean $\hat{\beta}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 0.5 | Firth | -1.175 | 41.248 | 0.752 | 1.198 | 1.454 | 74.900 | -0.269 | 73.067 | 1.596 | 1.226 | 1.427 | 97.297 |
| | Bayes_1 | -1.567 | 21.649 | – | 1.072 | 1.156 | 95.600 | -0.991 | 0.886 | – | 1.505 | 1.504 | 97.800 |
| | Bayes_2 | -1.974 | 1.277 | – | 0.393 | 0.393 | 99.600 | -0.996 | 0.404 | – | 0.441 | 0.441 | 99.990 |
| | Bayes_3 | -1.636 | 18.200 | – | 1.173 | 1.227 | 95.600 | -0.821 | 17.867 | – | 1.162 | 1.175 | 99.300 |
| 2.5 | Firth | -1.768 | 11.585 | 0.468 | 0.597 | 0.640 | 90.000 | -0.997 | 0.309 | 0.723 | 0.709 | 0.709 | 96.600 |
| | Bayes_1 | -1.976 | 1.193 | – | 0.562 | 0.562 | 93.900 | -1.058 | -5.774 | – | 0.789 | 0.791 | 94.400 |
| | Bayes_2 | -1.998 | 0.088 | – | 0.386 | 0.386 | 96.800 | -1.035 | -3.546 | – | 0.476 | 0.477 | 98.700 |
| | Bayes_3 | -1.971 | 1.442 | – | 0.633 | 0.634 | 92.700 | -0.980 | 1.965 | – | 0.732 | 0.732 | 94.200 |
| 7.5 | Firth | -1.928 | 3.609 | 0.281 | 0.288 | 0.297 | 93.500 | -0.984 | 1.598 | 0.417 | 0.408 | 0.408 | 96.100 |
| | Bayes_1 | -2.000 | 0.019 | – | 0.286 | 0.286 | 94.800 | -1.051 | -5.087 | – | 0.423 | 0.426 | 95.500 |
| | Bayes_2 | -2.008 | -0.417 | – | 0.273 | 0.273 | 95.800 | -1.033 | -3.270 | – | 0.360 | 0.361 | 96.700 |
| | Bayes_3 | -1.997 | 0.132 | – | 0.290 | 0.290 | 95.600 | -1.004 | -0.418 | – | 0.401 | 0.401 | 95.300 |
| 15 | Firth | -1.976 | 1.192 | 0.214 | 0.222 | 0.223 | 94.500 | -0.990 | 1.029 | 0.306 | 0.309 | 0.309 | 95.400 |
| | Bayes_1 | -2.007 | -0.343 | – | 0.221 | 0.221 | 94.500 | -1.005 | -0.502 | – | 0.310 | 0.310 | 95.300 |
| | Bayes_2 | -2.011 | -0.531 | – | 0.216 | 0.217 | 95.600 | -1.030 | -3.042 | – | 0.289 | 0.290 | 95.800 |
| | Bayes_3 | -1.988 | 0.619 | – | 0.219 | 0.219 | 94.900 | -1.012 | -1.207 | – | 0.310 | 0.311 | 93.600 |
| 25 | Firth | -2.001 | -0.025 | 0.182 | 0.185 | 0.185 | 94.000 | -1.000 | -0.012 | 0.251 | 0.251 | 0.251 | 95.300 |
| | Bayes_1 | -2.012 | -0.581 | – | 0.183 | 0.183 | 96.000 | -1.004 | -0.388 | – | 0.249 | 0.249 | 94.300 |
| | Bayes_2 | -2.008 | -0.395 | – | 0.180 | 0.180 | 95.600 | -1.019 | -1.926 | – | 0.232 | 0.232 | 96.400 |
| | Bayes_3 | -1.989 | 0.535 | – | 0.184 | 0.184 | 95.500 | -0.990 | 0.952 | – | 0.239 | 0.239 | 95.400 |
| 35 | Firth | -2.005 | -0.229 | 0.169 | 0.167 | 0.167 | 95.300 | -1.010 | -1.020 | 0.227 | 0.225 | 0.225 | 95.300 |
| | Bayes_1 | -2.011 | -0.562 | – | 0.168 | 0.168 | 95.900 | -1.016 | -1.571 | – | 0.231 | 0.232 | 94.900 |
| | Bayes_2 | -2.012 | -0.599 | – | 0.166 | 0.167 | 96.400 | -1.004 | -0.352 | – | 0.217 | 0.216 | 95.200 |
| | Bayes_3 | -2.005 | -0.269 | – | 0.175 | 0.175 | 93.800 | -1.003 | -0.315 | – | 0.231 | 0.231 | 94.900 |
| 50 | Firth | -2.000 | 0.007 | 0.161 | 0.165 | 0.165 | 95.000 | -1.007 | -0.675 | 0.214 | 0.216 | 0.216 | 94.600 |
| | Bayes_1 | -2.002 | -0.094 | – | 0.153 | 0.153 | 97.100 | -1.007 | -0.695 | – | 0.214 | 0.214 | 94.200 |
| | Bayes_2 | -2.006 | -0.276 | – | 0.150 | 0.150 | 97.000 | -1.017 | -1.712 | – | 0.211 | 0.212 | 94.000 |
| | Bayes_3 | -1.998 | 0.101 | – | 0.168 | 0.168 | 93.900 | -1.002 | -0.210 | – | 0.207 | 0.207 | 95.500 |

## Appendix C.2. Results for the Semiparametric Mixture Cure Fraction Model

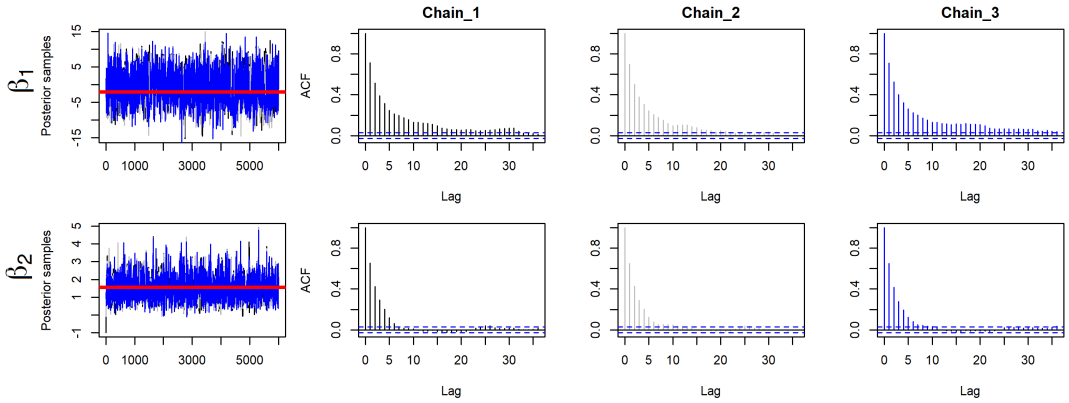


Figure C.2.1: Traceplots and ACF graphs for the Markov chain of the parameters related to the latency part. The SC1 is considered under the `Bayes_1` setting for a single sample size ($n = 50$). The red line indicates the true value of the corresponding parameter.

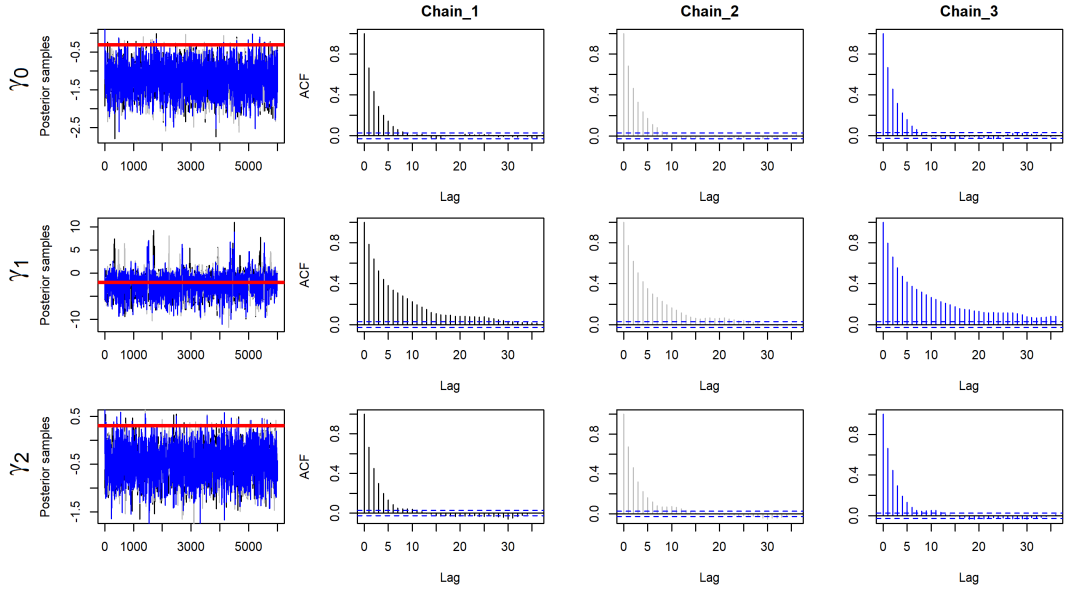


Figure C.2.2: Traceplots and ACF graphs for the Markov chain of the considered parameters based on the parametric mixture cure model. The SC1 is considered under the `Bayes_1` setting for a single sample size ($n = 50$). The red line indicates the true value of the corresponding parameter.

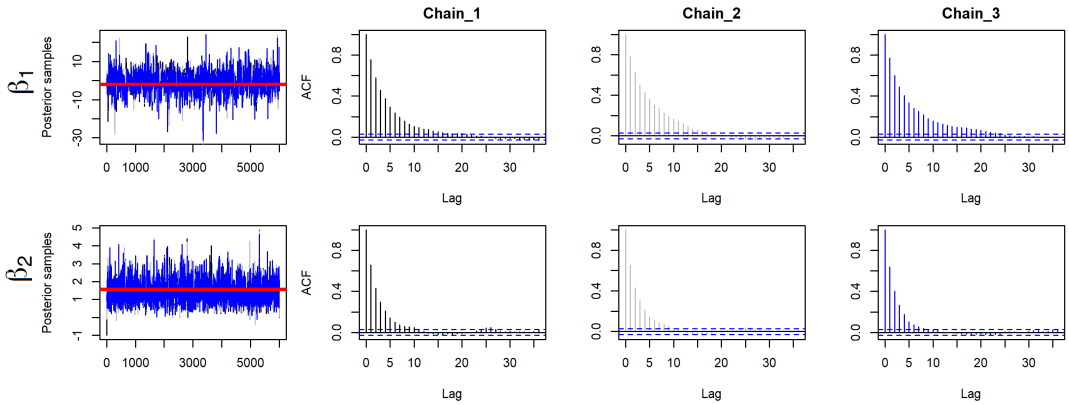


Figure C.2.3: Traceplots and ACF graphs for the Markov chain of the parameters related to the latency part. The SC1 is considered under the `Bayes_3` setting for a single sample size ($n = 50$). The red line indicates the true value of the corresponding parameter.

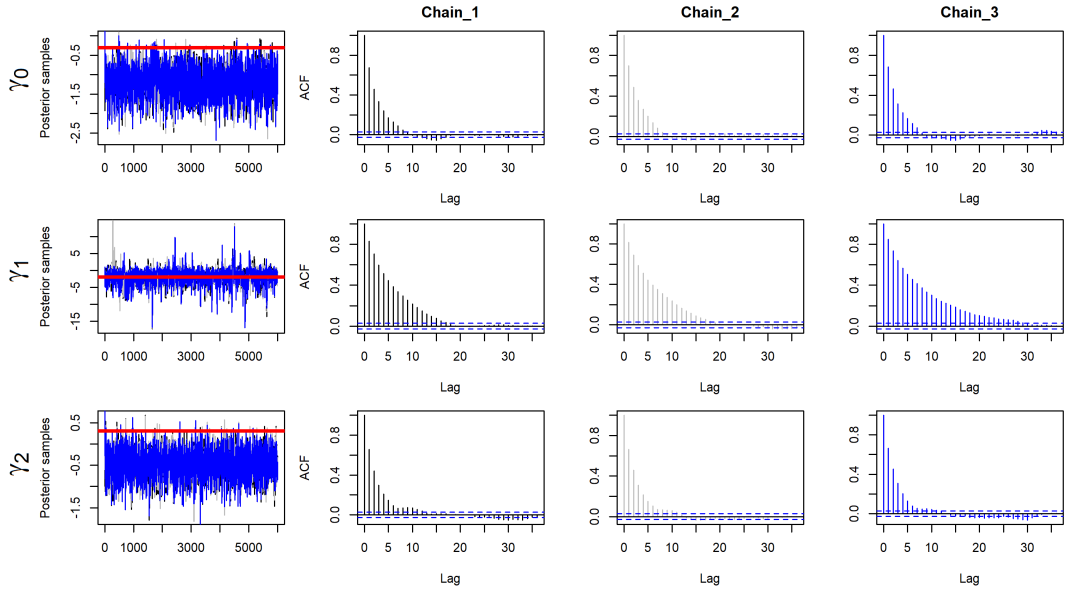


Figure C.2.4: Traceplots and ACF graphs for the Markov chain of the parameters related to the incidence part. The SC1 is considered under the `Bayes_3` setting for a single sample size ($n = 50$). The red line indicates the true value of the corresponding parameter.

Table C.2.1: *MC estimates (frequentist and Bayesian) and performance measurements for the coefficients related to the ML issue in both part of the model ($\beta_1$ and $\gamma_1$) based on the SC1. The real values are shown near the parameter name.*

| **Scenario 1** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_1 = -1.00$ | | | | | | $\gamma_1 = -2.00$ | | | | | |
| $n$ | | Mean $\hat{\beta}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 50 | Firth | -0.506 | 49.378 | 1.390 | 1.472 | 1.552 | 90.038 | -0.295 | 85.269 | 1.266 | 0.688 | 1.839 | 84.547 |
| | Bayes_1 | -1.145 | -14.495 | – | 1.120 | 1.128 | 97.800 | -1.894 | 5.308 | – | 1.032 | 1.037 | 98.500 |
| | Bayes_2 | -1.030 | -2.951 | – | 0.263 | 0.265 | 99.900 | -2.089 | -4.428 | – | 0.369 | 0.379 | 99.900 |
| | Bayes_3 | -1.385 | -38.497 | – | 1.138 | 1.201 | 98.500 | -1.540 | 22.993 | – | 0.819 | 0.939 | 97.800 |
| 200 | Firth | -0.307 | 69.275 | 0.973 | 1.242 | 1.421 | 79.874 | -0.336 | 83.187 | 1.148 | 0.541 | 1.749 | 84.277 |
| | Bayes_1 | -1.050 | -4.961 | – | 0.862 | 0.863 | 98.700 | -1.936 | 3.188 | – | 0.995 | 0.996 | 97.200 |
| | Bayes_2 | -1.003 | -0.313 | – | 0.272 | 0.272 | 99.600 | -2.058 | -2.909 | – | 0.391 | 0.395 | 99.900 |
| | Bayes_3 | -1.321 | -32.058 | – | 0.888 | 0.944 | 98.600 | -1.555 | 22.249 | – | 0.784 | 0.901 | 96.200 |
| 600 | Firth | -0.250 | 74.964 | 0.861 | 1.130 | 1.355 | 80.754 | -0.348 | 82.617 | 1.109 | 0.465 | 1.717 | 86.111 |
| | Bayes_1 | -0.947 | 5.264 | – | 0.824 | 0.825 | 98.400 | -2.007 | -0.340 | – | 1.012 | 1.011 | 96.700 |
| | Bayes_2 | -0.994 | 0.625 | – | 0.276 | 0.276 | 99.800 | -2.061 | -3.038 | – | 0.404 | 0.408 | 99.600 |
| | Bayes_3 | -1.232 | -23.199 | – | 0.846 | 0.877 | 98.500 | -1.614 | 19.299 | – | 0.806 | 0.894 | 96.300 |
| 1,000 | Firth | -0.207 | 79.292 | 0.851 | 1.219 | 1.454 | 76.673 | -0.355 | 82.245 | 1.109 | 0.468 | 1.710 | 88.145 |
| | Bayes_1 | -0.992 | 0.849 | – | 0.796 | 0.795 | 98.000 | -1.997 | 0.169 | – | 0.994 | 0.994 | 96.700 |
| | Bayes_2 | -0.999 | 0.060 | – | 0.280 | 0.280 | 99.800 | -2.061 | -3.029 | – | 0.398 | 0.402 | 99.900 |
| | Bayes_3 | -1.284 | -28.391 | – | 0.824 | 0.871 | 98.000 | -1.598 | 20.098 | – | 0.788 | 0.884 | 96.400 |

Table C.2.2: *MC estimates (frequentist and Bayesian) and performance measurements for the coefficients not directly related to the ML issue based on the SC1. The real values are shown near the parameter name.*

**Scenario 1**

| | | $\beta_2 = 1.56$ | | | | | | $\gamma_0 = -0.67$ | | | | | | $\gamma_2 = -0.67$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | | Mean $\hat{\beta}_2$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_0$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_2$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 50 | Firth | 1.714 | 9.840 | 0.617 | 0.684 | 0.701 | 94.253 | -0.646 | 3.525 | 0.367 | 0.356 | 0.357 | 96.424 | -0.740 | -10.480 | 0.420 | 0.393 | 0.399 | 97.701 |
| | Bayes_1 | 1.892 | 21.299 | – | 0.783 | 0.851 | 94.700 | -0.742 | -10.683 | – | 0.407 | 0.413 | 93.400 | -0.935 | -39.596 | – | 0.504 | 0.569 | 94.500 |
| | Bayes_2 | 1.726 | 10.669 | – | 0.444 | 0.474 | 97.800 | -0.710 | -6.011 | – | 0.334 | 0.336 | 96.300 | -0.849 | -26.706 | – | 0.370 | 0.410 | 96.800 |
| | Bayes_3 | 1.842 | 18.104 | – | 0.760 | 0.810 | 94.300 | -0.724 | -8.107 | – | 0.386 | 0.389 | 94.600 | -0.885 | -32.092 | – | 0.463 | 0.510 | 95.500 |
| 200 | Firth | 1.604 | 2.794 | 0.229 | 0.244 | 0.247 | 94.969 | -0.671 | -0.106 | 0.169 | 0.175 | 0.175 | 96.069 | -0.694 | -3.594 | 0.187 | 0.185 | 0.187 | 96.698 |
| | Bayes_1 | 1.613 | 3.377 | – | 0.238 | 0.244 | 95.600 | -0.683 | -1.912 | – | 0.173 | 0.173 | 95.000 | -0.729 | -8.808 | – | 0.190 | 0.199 | 94.700 |
| | Bayes_2 | 1.611 | 3.294 | – | 0.224 | 0.230 | 95.700 | -0.680 | -1.419 | – | 0.167 | 0.168 | 95.400 | -0.726 | -8.414 | – | 0.183 | 0.191 | 94.900 |
| | Bayes_3 | 1.606 | 2.959 | – | 0.237 | 0.241 | 94.900 | -0.680 | -1.494 | – | 0.172 | 0.172 | 95.400 | -0.722 | -7.808 | – | 0.188 | 0.195 | 94.600 |
| 600 | Firth | 1.575 | 0.954 | 0.124 | 0.122 | 0.123 | 95.833 | -0.669 | 0.187 | 0.096 | 0.093 | 0.093 | 95.437 | -0.679 | -1.285 | 0.105 | 0.104 | 0.104 | 96.429 |
| | Bayes_1 | 1.578 | 1.140 | – | 0.129 | 0.130 | 94.200 | -0.675 | -0.718 | – | 0.094 | 0.094 | 95.000 | -0.692 | -3.335 | – | 0.105 | 0.105 | 94.100 |
| | Bayes_2 | 1.578 | 1.174 | – | 0.126 | 0.128 | 94.200 | -0.674 | -0.602 | – | 0.093 | 0.093 | 94.900 | -0.692 | -3.308 | – | 0.103 | 0.106 | 94.400 |
| | Bayes_3 | 1.576 | 1.017 | – | 0.128 | 0.129 | 94.100 | -0.674 | -0.609 | – | 0.094 | 0.094 | 94.900 | -0.690 | -3.042 | – | 0.104 | 0.106 | 94.100 |
| 1,000 | Firth | 1.566 | 0.397 | 0.095 | 0.092 | 0.092 | 96.750 | -0.668 | 0.230 | 0.074 | 0.072 | 0.072 | 96.176 | -0.676 | -0.887 | 0.081 | 0.076 | 0.076 | 96.176 |
| | Bayes_1 | 1.566 | 0.413 | – | 0.098 | 0.098 | 94.000 | -0.672 | -0.274 | – | 0.076 | 0.076 | 94.600 | -0.677 | -1.095 | – | 0.083 | 0.083 | 94.100 |
| | Bayes_2 | 1.567 | 0.453 | – | 0.097 | 0.097 | 94.300 | -0.671 | -0.219 | – | 0.075 | 0.075 | 94.800 | -0.677 | -1.100 | – | 0.082 | 0.083 | 94.100 |
| | Bayes_3 | 1.565 | 0.338 | – | 0.098 | 0.098 | 94.200 | -0.671 | -0.202 | – | 0.076 | 0.076 | 94.500 | -0.676 | -0.931 | – | 0.083 | 0.083 | 94.500 |

Table C.2.3: *MC estimates (frequentist and Bayesian) and performance measurements for the coefficients related to the ML issue in both part of the model ($\beta_1$ and $\gamma_1$) based on the SC2. The real values are shown near the parameter name.*

| **Scenario 2** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_1 = -1.00$ | | | | | | $\gamma_1 = -1.00$ | | | | | |
| $n$ | | Mean $\hat{\beta}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 50 | Firth | -0.645 | 35.489 | 1.098 | 1.242 | 1.291 | 90.625 | -0.237 | 76.253 | 1.543 | 1.238 | 1.454 | 97.379 |
| | Bayes_1 | -1.145 | -14.492 | – | 1.655 | 1.661 | 94.400 | -0.888 | 11.154 | – | 1.449 | 1.452 | 97.800 |
| | Bayes_2 | -1.028 | -2.815 | – | 0.446 | 0.447 | 99.900 | -1.041 | -4.071 | – | 0.449 | 0.451 | 99.900 |
| | Bayes_3 | -1.148 | -14.760 | – | 1.721 | 1.727 | 96.000 | -0.761 | 23.868 | – | 1.170 | 1.193 | 98.700 |
| 200 | Firth | -0.517 | 48.328 | 0.843 | 1.051 | 1.157 | 86.145 | -0.262 | 73.809 | 1.355 | 1.053 | 1.285 | 95.582 |
| | Bayes_1 | -1.044 | -4.374 | – | 1.266 | 1.266 | 95.200 | -0.983 | 1.684 | – | 1.266 | 1.266 | 98.700 |
| | Bayes_2 | -1.013 | -1.276 | – | 0.440 | 0.440 | 99.600 | -1.038 | -3.783 | – | 0.437 | 0.438 | 99.900 |
| | Bayes_3 | -1.058 | -5.793 | – | 1.291 | 1.291 | 96.000 | -0.822 | 17.787 | – | 1.035 | 1.049 | 98.800 |
| 600 | Firth | -0.344 | 65.583 | 0.772 | 1.023 | 1.215 | 80.750 | -0.272 | 72.831 | 1.334 | 1.046 | 1.274 | 96.758 |
| | Bayes_1 | -0.855 | 14.469 | – | 1.031 | 1.041 | 96.800 | -1.055 | -5.458 | – | 1.349 | 1.350 | 97.200 |
| | Bayes_2 | -0.989 | 1.060 | – | 0.402 | 0.402 | 99.400 | -1.019 | -1.886 | – | 0.476 | 0.476 | 99.800 |
| | Bayes_3 | -0.837 | 16.304 | – | 1.025 | 1.037 | 97.100 | -0.866 | 13.382 | – | 1.117 | 1.125 | 97.200 |
| 1,000 | Firth | -0.382 | 61.775 | 0.763 | 1.039 | 1.208 | 79.539 | -0.313 | 68.662 | 1.332 | 1.001 | 1.214 | 96.489 |
| | Bayes_1 | -1.008 | -0.769 | – | 1.033 | 1.033 | 95.300 | -1.014 | 0.594 | – | 1.276 | 1.276 | 97.800 |
| | Bayes_2 | -0.989 | 1.104 | – | 0.414 | 0.414 | 99.300 | -1.038 | -3.827 | – | 0.465 | 0.466 | 99.900 |
| | Bayes_3 | -1.011 | -1.052 | – | 1.035 | 1.034 | 96.100 | -0.834 | 16.635 | – | 1.050 | 1.063 | 97.900 |

Table C.2.4: *MC estimates (frequentist and Bayesian) and performance measurements for the coefficients not directly related to the ML issue based on the SC2. The real values are shown near the parameter name.*

| **Scenario 2** | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_2 = 1.00$ | | | | | | $\gamma_0 = -1.00$ | | | | | | $\gamma_2 = 2.00$ | | | | | |
| $n$ | | Mean $\hat{\beta}_2$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_0$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | CP (%) | Mean $\hat{\gamma}_2$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 50 | Firth | 1.113 | 11.290 | 0.526 | 0.589 | 0.600 | 96.472 | -0.980 | 1.966 | 0.482 | 0.492 | 0.492 | 96.875 | 2.016 | 0.822 | 0.677 | 0.790 | 0.790 | 94.859 |
| | Bayes_1 | 1.210 | 20.999 | – | 0.740 | 0.768 | 92.100 | -1.150 | -15.004 | – | 0.545 | 0.565 | 95.400 | 2.390 | 19.506 | – | 0.820 | 0.907 | 93.700 |
| | Bayes_2 | 1.101 | 10.137 | – | 0.442 | 0.453 | 96.200 | -1.068 | -6.820 | – | 0.378 | 0.384 | 97.200 | 2.191 | 9.534 | – | 0.460 | 0.498 | 98.200 |
| | Bayes_3 | 1.178 | 17.817 | – | 0.717 | 0.739 | 92.900 | -1.072 | -7.201 | – | 0.491 | 0.496 | 96.200 | 2.229 | 11.426 | – | 0.756 | 0.789 | 95.800 |
| 200 | Firth | 1.026 | 2.567 | 0.209 | 0.215 | 0.216 | 95.382 | -0.996 | 0.432 | 0.223 | 0.222 | 0.222 | 95.683 | 1.990 | -0.511 | 0.315 | 0.314 | 0.314 | 94.980 |
| | Bayes_1 | 1.028 | 2.787 | – | 0.221 | 0.222 | 93.300 | -1.041 | -4.104 | – | 0.230 | 0.234 | 94.800 | 2.110 | 5.524 | – | 0.341 | 0.339 | 94.800 |
| | Bayes_2 | 1.024 | 2.373 | – | 0.209 | 0.210 | 94.500 | -1.037 | -3.66 | – | 0.212 | 0.215 | 95.700 | 2.101 | 5.039 | – | 0.299 | 0.315 | 94.800 |
| | Bayes_3 | 1.023 | 2.332 | – | 0.219 | 0.220 | 93.800 | -1.024 | -2.438 | – | 0.224 | 0.226 | 96.000 | 2.074 | 3.719 | – | 0.330 | 0.338 | 94.100 |
| 600 | Firth | 1.011 | 1.070 | 0.115 | 0.117 | 0.117 | 94.833 | -0.995 | 0.455 | 0.127 | 0.128 | 0.128 | 95.035 | 1.992 | -0.424 | 0.180 | 0.183 | 0.183 | 94.124 |
| | Bayes_1 | 1.009 | 0.896 | – | 0.119 | 0.119 | 93.900 | -1.021 | -2.139 | – | 0.131 | 0.133 | 93.500 | 2.042 | 2.099 | – | 0.189 | 0.193 | 92.900 |
| | Bayes_2 | 1.008 | 0.837 | – | 0.117 | 0.118 | 94.300 | -1.022 | -2.169 | – | 0.128 | 0.130 | 93.900 | 2.044 | 2.194 | – | 0.182 | 0.187 | 93.900 |
| | Bayes_3 | 1.008 | 0.760 | – | 0.119 | 0.119 | 93.300 | -1.016 | -1.610 | – | 0.130 | 0.131 | 93.900 | 2.031 | 1.527 | – | 0.186 | 0.189 | 93.600 |
| 1,000 | Firth | 1.007 | 0.749 | 0.088 | 0.087 | 0.088 | 95.186 | -0.991 | 0.876 | 0.098 | 0.096 | 0.096 | 95.386 | 1.996 | -0.182 | 0.139 | 0.146 | 0.146 | 94.483 |
| | Bayes_1 | 1.004 | 0.415 | – | 0.092 | 0.092 | 93.200 | -1.014 | -1.384 | – | 0.097 | 0.098 | 94.800 | 2.025 | 1.227 | – | 0.139 | 0.141 | 95.100 |
| | Bayes_2 | 1.004 | 0.383 | – | 0.091 | 0.091 | 93.400 | -1.015 | -1.458 | – | 0.096 | 0.097 | 95.300 | 2.027 | 1.327 | – | 0.136 | 0.139 | 95.200 |
| | Bayes_3 | 1.003 | 0.337 | – | 0.091 | 0.091 | 93.300 | -1.011 | -1.064 | – | 0.097 | 0.097 | 94.900 | 2.018 | 0.882 | – | 0.138 | 0.139 | 95.500 |

Table C.2.5: MC estimates (frequentist and Bayesian) and performance measurements for the coefficients ($\beta_1$ and $\gamma_1$) related to the binary covariate ($x_{1i}$, $i = 1, 2, \ldots, n$) for both estimation methods. The sample size is fixed ($n$ = 1,000). Different percentages of ones in $x_{1i}$ are explored.

| **Scenario 1** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_1 = -1.00$ | | | | | | $\gamma_1 = -2.00$ | | | | | |
| $\%\{x_{1i} = 1\}$ | | Mean $\hat{\beta}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 0.5 | Firth | -0.207 | 79.292 | 0.851 | 1.219 | 1.454 | 76.673 | -0.355 | 82.245 | 1.109 | 0.468 | 1.710 | 88.145 |
| | Bayes_1 | -0.992 | 0.849 | – | 0.796 | 0.795 | 98.000 | -1.997 | 0.169 | – | 0.994 | 0.994 | 96.700 |
| | Bayes_2 | -0.999 | 0.060 | – | 0.280 | 0.280 | 99.800 | -2.061 | -3.029 | – | 0.398 | 0.402 | 99.900 |
| | Bayes_3 | -1.284 | -28.391 | – | 0.824 | 0.871 | 98.000 | -1.598 | 20.098 | – | 0.788 | 0.884 | 96.400 |
| 2.5 | Firth | -0.375 | 62.458 | 0.713 | 1.087 | 1.253 | 80.300 | -1.675 | 16.249 | 0.744 | 0.516 | 0.609 | 93.621 |
| | Bayes_1 | -1.072 | -7.174 | – | 0.953 | 0.955 | 96.200 | -2.250 | -12.523 | – | 0.889 | 0.923 | 96.600 |
| | Bayes_2 | -0.978 | 2.173 | – | 0.413 | 0.413 | 99.400 | -2.118 | -5.904 | – | 0.501 | 0.515 | 98.300 |
| | Bayes_3 | -1.342 | -34.152 | – | 1.198 | 1.245 | 96.400 | -2.080 | -4.017 | – | 0.696 | 0.700 | 97.500 |
| 7.5 | Firth | -0.718 | 28.175 | 0.488 | 0.646 | 0.705 | 87.432 | -2.005 | -0.270 | 0.489 | 0.478 | 0.478 | 96.175 |
| | Bayes_1 | -1.019 | -1.875 | – | 0.588 | 0.588 | 94.900 | -2.128 | -6.383 | – | 0.571 | 0.585 | 96.200 |
| | Bayes_2 | -0.994 | 0.555 | – | 0.405 | 0.405 | 97.100 | -2.098 | -4.891 | – | 0.421 | 0.432 | 97.300 |
| | Bayes_3 | -1.059 | -5.950 | – | 0.662 | 0.664 | 94.400 | -2.058 | -2.924 | – | 0.547 | 0.550 | 95.400 |
| 15 | Firth | -0.940 | 6.008 | 0.351 | 0.381 | 0.386 | 93.173 | -1.990 | 0.483 | 0.351 | 0.342 | 0.342 | 96.679 |
| | Bayes_1 | -1.029 | -2.922 | – | 0.388 | 0.389 | 94.300 | -2.049 | -2.446 | – | 0.373 | 0.376 | 94.300 |
| | Bayes_2 | -1.023 | -2.339 | – | 0.330 | 0.331 | 95.300 | -2.053 | -2.670 | – | 0.324 | 0.328 | 95.600 |
| | Bayes_3 | -1.031 | -3.082 | – | 0.374 | 0.375 | 94.100 | -2.012 | -0.583 | – | 0.349 | 0.349 | 95.900 |
| 25 | Firth | -0.951 | 4.884 | 0.276 | 0.302 | 0.306 | 91.274 | -1.997 | 0.147 | 0.276 | 0.276 | 0.276 | 94.764 |
| | Bayes_1 | -1.011 | -1.093 | – | 0.289 | 0.289 | 94.500 | -2.010 | -0.522 | – | 0.270 | 0.270 | 95.300 |
| | Bayes_2 | -1.011 | -1.133 | – | 0.266 | 0.266 | 95.700 | -2.020 | -0.998 | – | 0.253 | 0.253 | 95.700 |
| | Bayes_3 | -1.015 | -1.532 | – | 0.287 | 0.288 | 95.300 | -2.021 | -1.032 | – | 0.262 | 0.263 | 95.600 |
| 35 | Firth | -0.990 | 1.023 | 0.240 | 0.251 | 0.251 | 94.983 | -1.982 | 0.884 | 0.240 | 0.235 | 0.236 | 94.482 |
| | Bayes_1 | -1.026 | -2.588 | – | 0.252 | 0.253 | 93.000 | -2.022 | -1.098 | – | 0.243 | 0.244 | 95.100 |
| | Bayes_2 | -1.026 | -2.564 | – | 0.237 | 0.239 | 94.400 | -2.029 | -1.454 | – | 0.231 | 0.232 | 95.500 |
| | Bayes_3 | -1.016 | -1.639 | – | 0.243 | 0.244 | 95.300 | -2.015 | -0.746 | – | 0.246 | 0.246 | 94.700 |
| 50 | Firth | -1.007 | -0.654 | 0.213 | 0.229 | 0.229 | 92.528 | -1.975 | 1.241 | 0.213 | 0.219 | 0.220 | 94.277 |
| | Bayes_1 | -1.019 | -1.873 | – | 0.217 | 0.218 | 94.400 | -2.010 | -0.507 | – | 0.215 | 0.215 | 94.500 |
| | Bayes_2 | -1.019 | -1.949 | – | 0.208 | 0.209 | 94.600 | -2.017 | -0.839 | – | 0.206 | 0.207 | 95.200 |
| | Bayes_3 | -1.013 | -1.260 | – | 0.204 | 0.204 | 96.600 | -2.003 | -0.171 | – | 0.215 | 0.215 | 94.700 |

Table C.2.6: MC estimates (frequentist and Bayesian) and performance measurements for the coefficients ($\beta_1$ and $\gamma_1$) related to the binary covariate ($x_{1i}$, $i = 1, 2, \ldots, n$) for both estimation methods. The sample size is fixed ($n$ = 1,000). Different percentages of ones in $x_{1i}$ are explored.

| **Scenario 2** | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\beta_1 = -1.00$ | | | | | | $\gamma_1 = -1.00$ | | | | | |
| $\%\{x_{1i} = 1\}$ | | Mean $\hat{\beta}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP | Mean $\hat{\gamma}_1$ | %RB | $\sigma^{(as)}$ | $\sigma^{(mc)}$ | RMSE | %CP |
| 0.5 | Firth | -0.382 | 61.775 | 0.763 | 1.039 | 1.208 | 79.539 | -0.313 | 68.662 | 1.332 | 1.001 | 1.214 | 96.489 |
| | Bayes_1 | -1.008 | -0.769 | – | 1.033 | 1.033 | 95.300 | -1.014 | 0.594 | – | 1.276 | 1.276 | 97.800 |
| | Bayes_2 | -0.989 | 1.104 | – | 0.414 | 0.414 | 99.300 | -1.038 | -3.827 | – | 0.465 | 0.466 | 99.900 |
| | Bayes_3 | -1.011 | -1.052 | – | 1.035 | 1.034 | 96.100 | -0.834 | 16.635 | – | 1.050 | 1.063 | 97.900 |
| 2.5 | Firth | -0.769 | 23.089 | 0.473 | 0.590 | 0.634 | 89.870 | -0.976 | 2.362 | 0.658 | 0.649 | 0.649 | 95.482 |
| | Bayes_1 | -1.019 | -1.858 | – | 0.549 | 0.549 | 95.600 | -1.044 | -4.353 | – | 0.669 | 0.670 | 95.300 |
| | Bayes_2 | -1.002 | -0.187 | – | 0.398 | 0.397 | 97.200 | -1.032 | -3.218 | – | 0.468 | 0.469 | 97.700 |
| | Bayes_3 | -1.007 | -0.671 | – | 0.538 | 0.537 | 95.300 | -0.994 | 0.588 | – | 0.634 | 0.634 | 95.000 |
| 7.5 | Firth | -0.969 | 3.090 | 0.278 | 0.299 | 0.300 | 93.173 | -0.985 | 1.542 | 0.382 | 0.388 | 0.388 | 96.482 |
| | Bayes_1 | -0.996 | 0.417 | – | 0.284 | 0.283 | 95.100 | -1.051 | -5.073 | – | 0.391 | 0.394 | 94.900 |
| | Bayes_2 | -0.998 | 0.235 | – | 0.262 | 0.262 | 95.800 | -1.05 | -5.145 | – | 0.342 | 0.346 | 95.800 |
| | Bayes_3 | -0.998 | 0.192 | – | 0.301 | 0.301 | 92.900 | -1.006 | -0.562 | – | 0.397 | 0.397 | 94.000 |
| 15 | Firth | -0.987 | 1.345 | 0.205 | 0.214 | 0.214 | 93.970 | -0.997 | 0.309 | 0.279 | 0.275 | 0.275 | 95.291 |
| | Bayes_1 | -1.012 | -1.161 | – | 0.214 | 0.215 | 93.700 | -1.011 | -1.099 | – | 0.293 | 0.293 | 94.100 |
| | Bayes_2 | -1.013 | -1.271 | – | 0.206 | 0.206 | 94.500 | -1.016 | -1.593 | – | 0.273 | 0.273 | 95.000 |
| | Bayes_3 | -1.004 | -0.400 | – | 0.200 | 0.200 | 94.900 | -0.992 | 0.796 | – | 0.281 | 0.281 | 94.700 |
| 25 | Firth | -0.993 | 0.705 | 0.168 | 0.169 | 0.169 | 95.295 | -1.001 | -0.144 | 0.229 | 0.233 | 0.232 | 94.394 |
| | Bayes_1 | -1.015 | -1.473 | – | 0.175 | 0.175 | 92.700 | -1.012 | -1.187 | – | 0.224 | 0.224 | 95.200 |
| | Bayes_2 | -1.016 | -1.012 | – | 0.170 | 0.170 | 93.600 | -1.016 | -1.016 | – | 0.213 | 0.214 | 95.700 |
| | Bayes_3 | -1.005 | -0.510 | – | 0.169 | 0.169 | 93.400 | -1.004 | -0.438 | – | 0.220 | 0.220 | 96.000 |
| 35 | Firth | -1.001 | -0.056 | 0.152 | 0.158 | 0.158 | 94.070 | -0.990 | 0.982 | 0.206 | 0.201 | 0.202 | 95.578 |
| | Bayes_1 | -1.012 | -1.165 | – | 0.154 | 0.155 | 94.300 | -1.011 | -1.082 | – | 0.205 | 0.205 | 94.700 |
| | Bayes_2 | -1.012 | -1.224 | – | 0.151 | 0.151 | 94.200 | -1.014 | -1.379 | – | 0.197 | 0.197 | 95.400 |
| | Bayes_3 | -1.011 | -1.069 | – | 0.158 | 0.158 | 93.600 | -0.999 | 0.079 | – | 0.220 | 0.219 | 92.600 |
| 50 | Firth | -1.009 | -0.856 | 0.141 | 0.143 | 0.144 | 94.589 | -0.998 | 1.199 | 0.193 | 0.201 | 0.202 | 93.687 |
| | Bayes_1 | -1.012 | -1.152 | – | 0.139 | 0.140 | 96.000 | -0.994 | 0.581 | – | 0.192 | 0.192 | 94.800 |
| | Bayes_2 | -1.012 | -1.230 | – | 0.136 | 0.137 | 96.100 | -0.997 | 0.282 | – | 0.185 | 0.185 | 95.400 |
| | Bayes_3 | -1.009 | -0.944 | – | 0.145 | 0.145 | 93.700 | -1.008 | -0.753 | – | 0.201 | 0.201 | 94.200 |

## Appendix D. Additional Results for Data Analysis

This section shows the additional details of the melanoma data analysis illustrated in Chapter 7, for both parametric and semiparametric mixture cure fraction models. The general details of the Markov chain convergence and the corresponding confidence intervals or credible regions plots are presented.

### Appendix D.1. Plots for the Parametric Mixture Cure Fraction Model

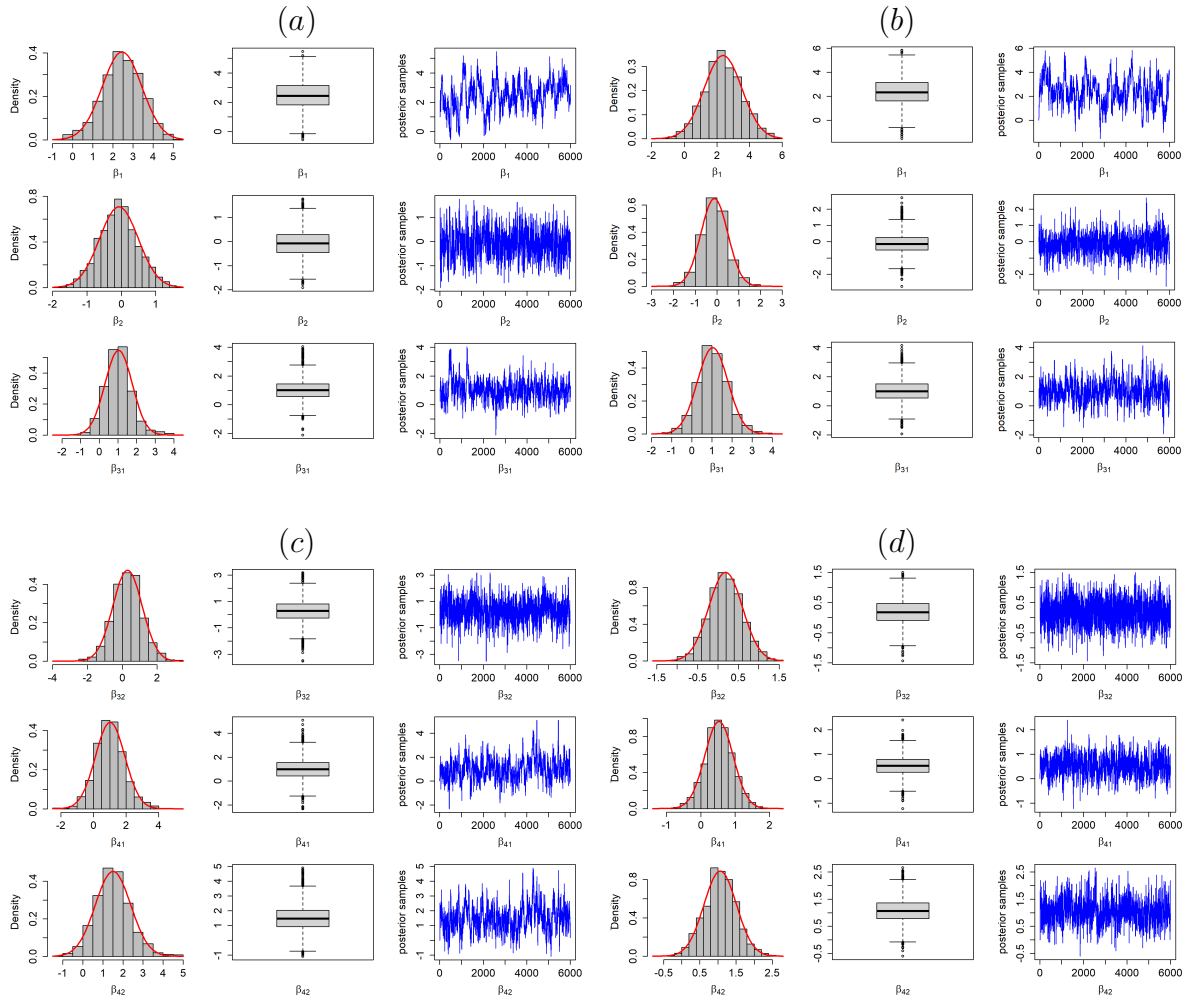


Figure D.1.1: Graphs related to the Markov chain for the regression coefficients $\beta_1$ (Mitosis); $\beta_2$ (Gender); $\beta_{31}$ (Histologial type: "nodular"); $\beta_{32}$ (Histologial type: "acral lentiginous"); $\beta_{41}$ (Breslow index: "1-4mm") and $\beta_{42}$ (Breslow index: ">4mm") related to the latency part. The melanoma data set were considered under the `Bayes_1` (panels $a$ and $c$) and `Bayes_3` (panels $b$ and $d$), under the parametric specification for the latency distribution.

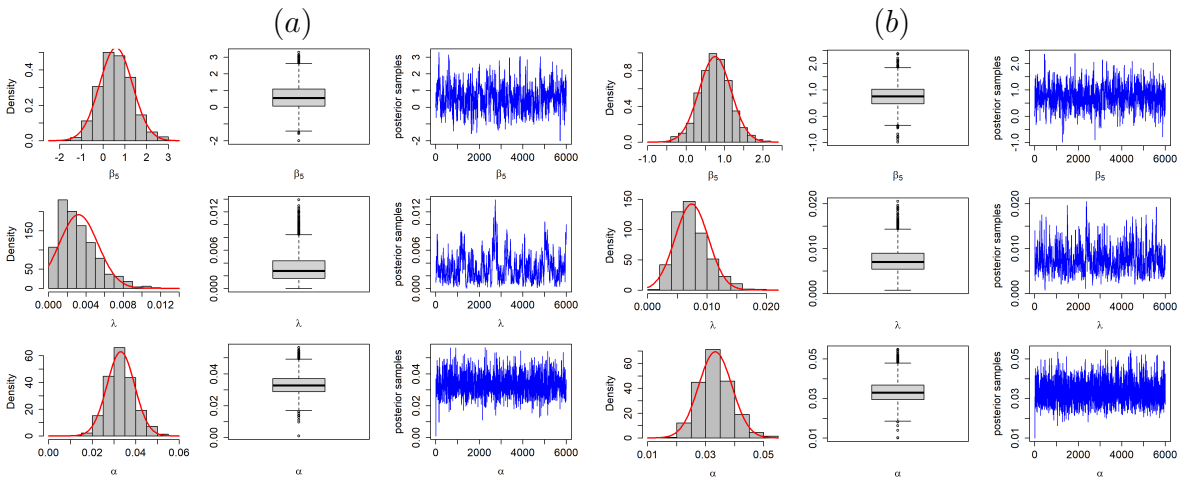


Figure D.1.2: Graphs related to the Markov chain for the quantities related to the latency part, namely: $\beta_5$ (Ulceration), including the shape and scale parameters ($\alpha$ and $\lambda$). The melanoma data set were considered under the `Bayes_1` (panel $a$) and `Bayes_3` (panel $b$), based on the parametric specification for the latency distribution.

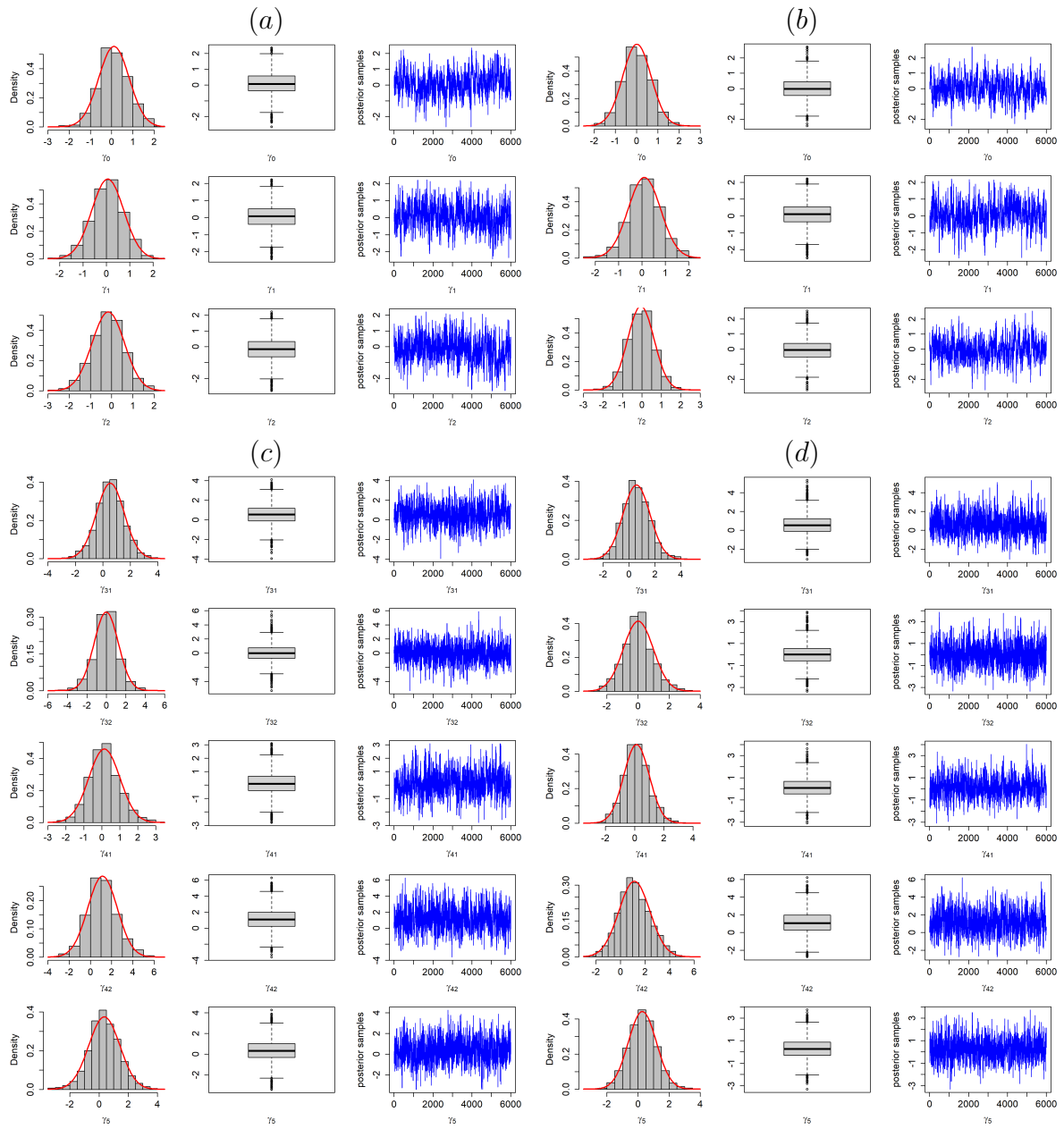


Figure D.1.3: Graphs related to the Markov chain for the coefficients related to the incidence part, namely: $\gamma_0$ (Intercept); $\gamma_1$ (Mitosis); $\gamma_2$ (Gender); $\gamma_{31}$ (Histologial type: "nodular"); $\gamma_{32}$ (Histologial type: "acral lentiginous"); $\gamma_{41}$ (Breslow index: "1-4mm") and $\gamma_{42}$ (Breslow index: ">4mm") and $\gamma_5$ (Ulceration). The melanoma data set were considered under the `Bayes_1` (panels $a$ and $c$) and `Bayes_3` (panels $b$ and $d$), under the parametric specification for the latency distribution.

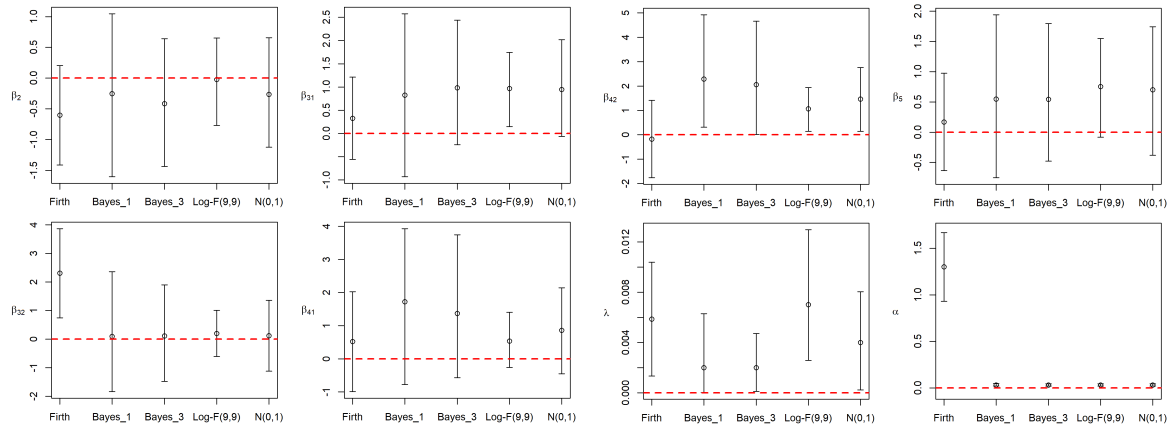


Figure D.1.4: Comparing 95% confidence intervals and 95% credible regions for the regression coefficients: $\beta_2$ (Gender); $\beta_{31}$ (Histologial type: "nodular"); $\beta_{32}$ (Histologial type: "acral lentiginous"); $\beta_{41}$ (Breslow index: "1-4mm") and $\beta_{42}$ (Breslow index: ">4mm"); $\beta_5$ (Ulceration), including the shape and scale parameters ($\alpha$ and $\lambda$) of the parametric mixture cure fraction model. The small circles denote the posterior estimate or penalized maximum likelihood estimates, according to the estimation method.

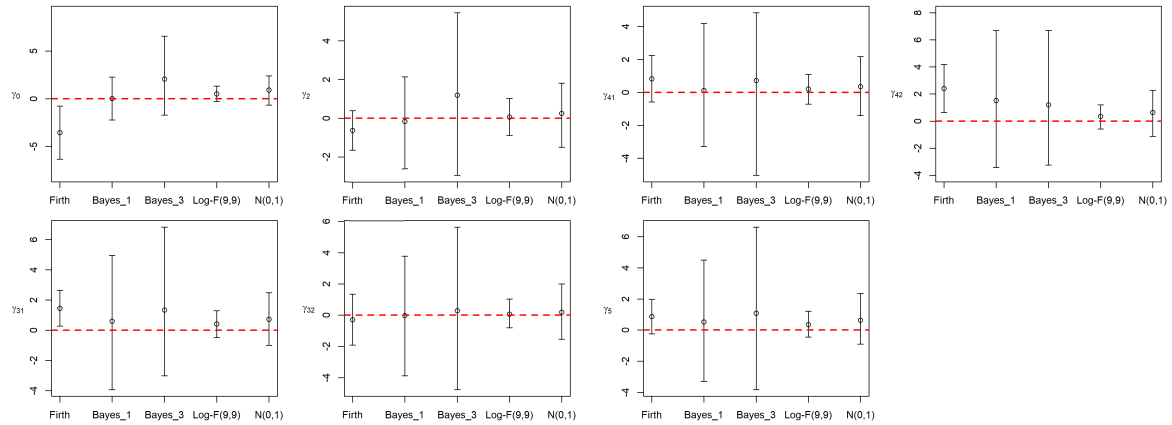


Figure D.1.5: Comparing 95% confidence intervals and 95% credible regions for the regression coefficients: $\gamma_0$ (Intercept); $\gamma_2$ (Gender); $\gamma_{31}$ (Histologial type: "nodular"); $\gamma_{32}$ (Histologial type: "acral lentiginous"); $\gamma_{41}$ (Breslow index: "1-4mm") and $\gamma_{42}$ (Breslow index: ">4mm") and $\gamma_5$ (Ulceration). The small circles denote the posterior estimate or penalized maximum likelihood estimates, according to the estimation method.

## Appendix D.2. Plots for the Semiparametric Mixture Cure Fraction Model

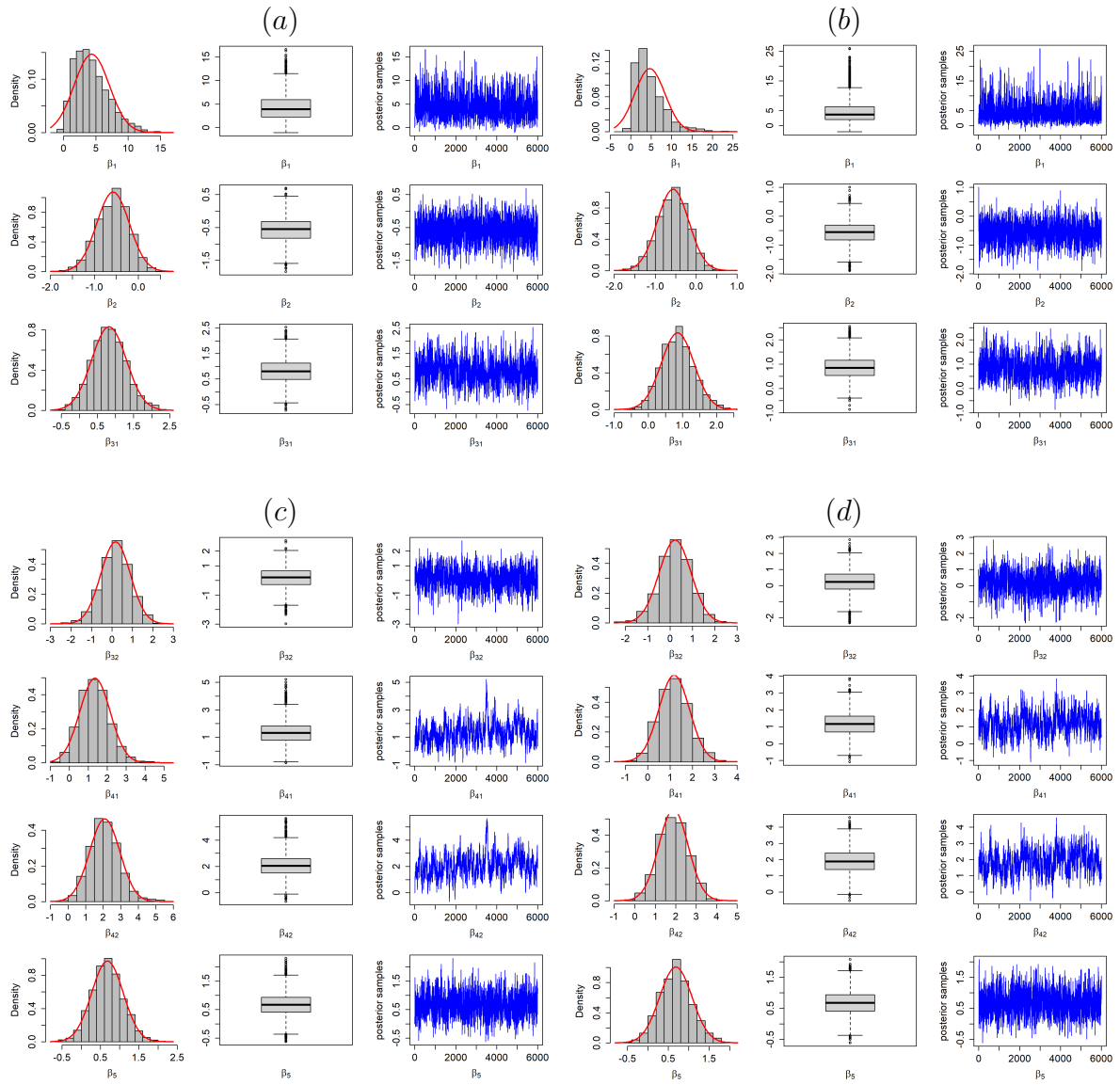


Figure D.2.1: Graphs related to the Markov chain for the coefficients related to the latency part, namely: $\beta_1$ (Mitosis); $\beta_2$ (Gender); $\beta_{31}$ (Histologial type: "nodular"); $\beta_{32}$ (Histologial type: "acral lentiginous"); $\beta_{41}$ (Breslow index: "1-4mm") and $\beta_{42}$ (Breslow index: ">4mm") and $\beta_5$ (Ulceration). The melanoma data set were considered under the `Bayes_1` (panels *a* and *c*) and `Bayes_3` (panels *b* and *d*), under the semiparametric specification for the latency distribution.

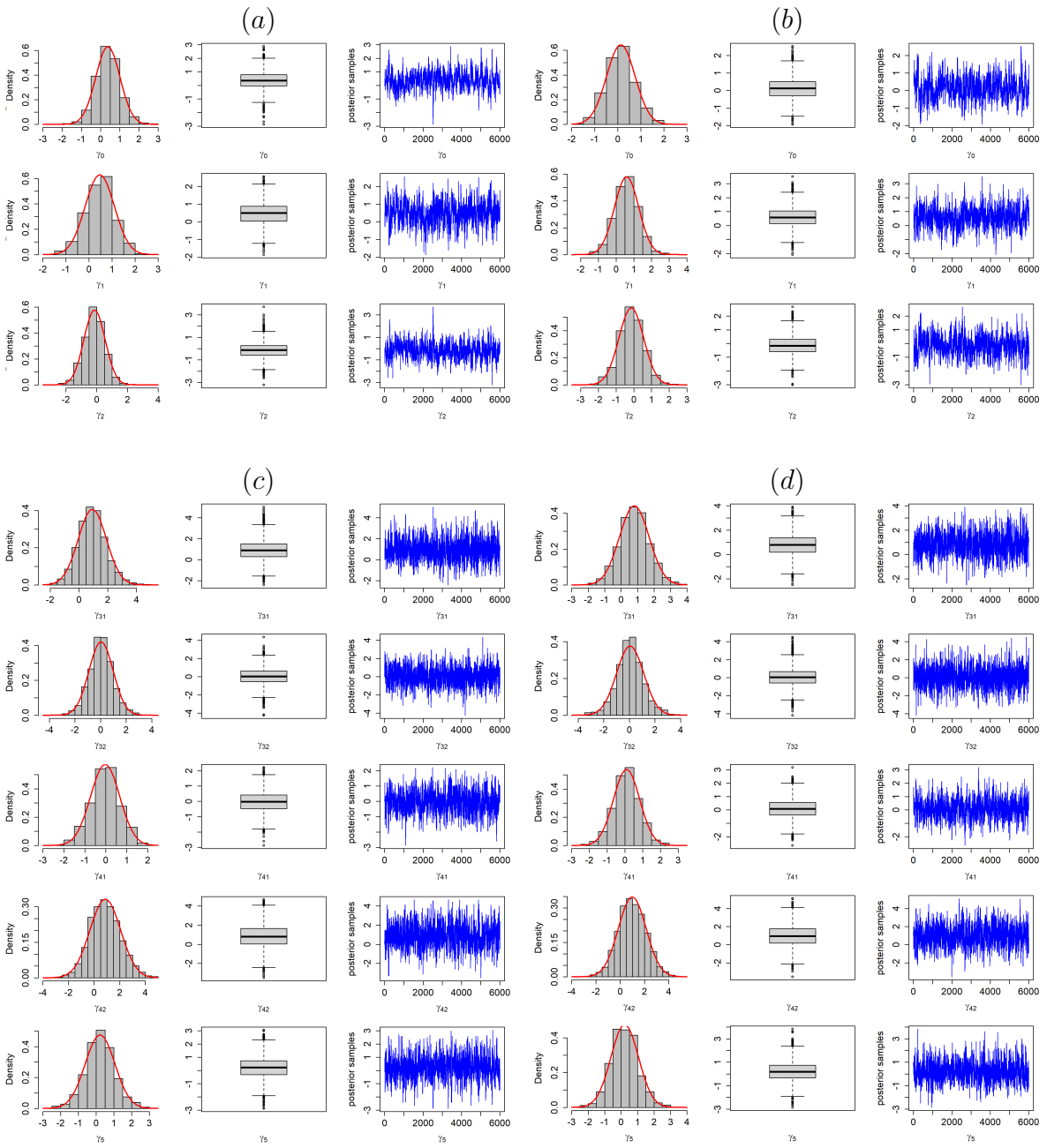


Figure D.2.2: Graphs related to the Markov chain for the coefficients related to the incidence part, namely: $\gamma_0$; $\gamma_1$ (Mitosis); $\gamma_2$ (Gender); $\gamma_{31}$ (Histologial type: "nodular"); $\gamma_{32}$ (Histologial type: "acral lentiginous"); $\gamma_{41}$ (Breslow index: "1-4mm") and $\gamma_{42}$ (Breslow index: ">4mm") and $\gamma_5$ (Ulceration). The melanoma data set were considered under the `Bayes_1` (panels *a* and *c*) and `Bayes_3` (panels *b* and *d*), under the semiparametric specification for the latency distribution.

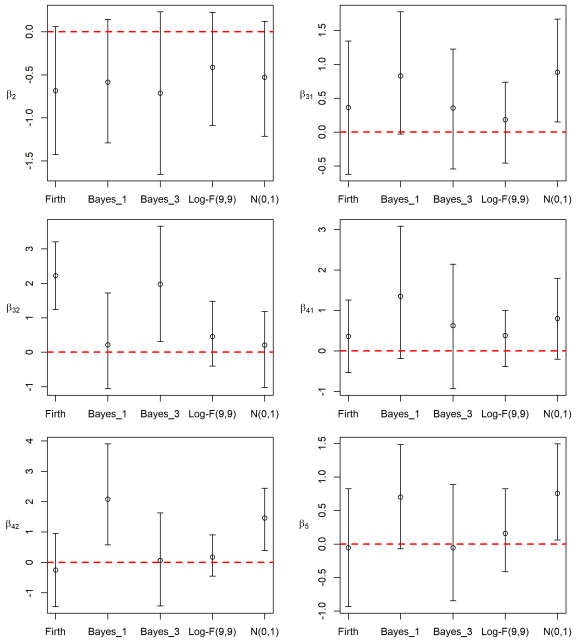


Figure D.2.3: Comparing 95% confidence intervals and 95% credible regions for the regression coefficients: $\beta_2$ (Gender); $\beta_{31}$ (Histologial type: "nodular"); $\beta_{32}$ (Histologial type: "acral lentiginous"); $\beta_{41}$ (Breslow index: "1-4mm") and $\beta_{42}$ (Breslow index: ">4mm"); $\beta_5$ (Ulceration) of the semiparametric mixture cure fraction model. The small circles denote the posterior mean (for Bayesian framework) or penalized maximum likelihood estimates (for the Firth method).

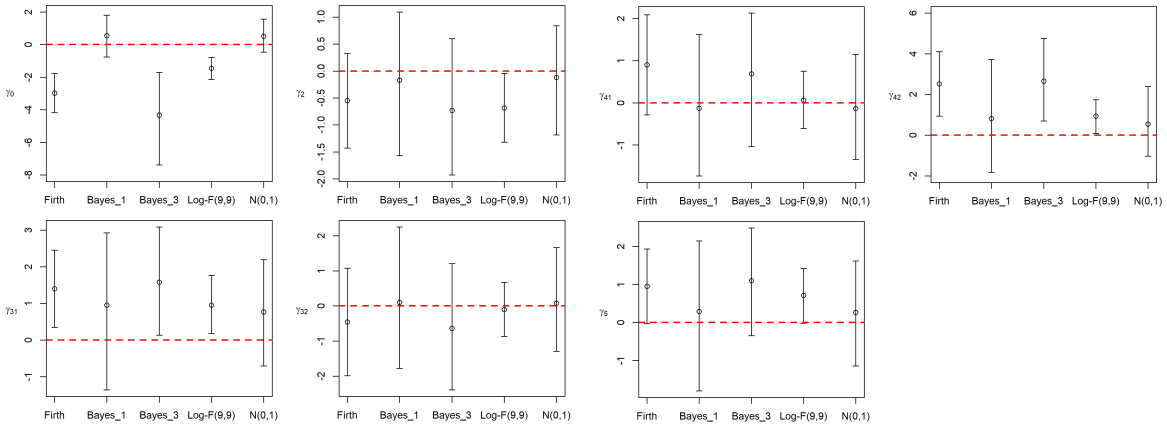


Figure D.2.4: Comparing 95% confidence intervals and 95% credible regions for the regression coefficients: $\gamma_0$; $\gamma_2$ (Gender); $\gamma_{31}$ (Histologial type: "nodular"); $\gamma_{32}$ (Histologial type: "acral lentiginous"); $\gamma_{41}$ (Breslow index: "1-4mm") and $\gamma_{42}$ (Breslow index: ">4mm") and $\gamma_5$ (Ulceration). The small circles denote the posterior mean (for Bayesian framework) or penalized maximum likelihood estimates (for the Firth method).

# Bibliography


Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726.

Achcar, J. A., Coelho-Barros, E. A., and Mazucheli, J. (2012). Cure fraction models using mixture and non-mixture models. *Mathematical Publications*, 51(1):1–9.

Agresti, A. and Yang, M.-C. (1987). An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics & Data Analysis*, 5(1):9–21.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrics*, 71(1):1–10.

Allison, P. D. (2010). *Survival analysis using SAS: a practical guide*. SAS Institute, North Carolina, 2 edition.

Almeida, F. M., Colosimo, E. A., and Mayrink, V. D. (2018). Prior specifications to handle the monotone likelihood problem in the Cox regression model. *Statistics and Its Interface*, 11(4):687–698.

Almeida, F. M., Colosimo, E. A., and Mayrink, V. D. (2021a). Firth adjusted score function for monotone likelihood in the mixture cure fraction model. *Lifetime Data Analysis*, 27(1):131–155.

Almeida, F. M., Colosimo, E. A., and Mayrink, V. D. (2021b). Modified score function for monotone likelihood in the semiparametric mixture cure model. *Biometrical Journal*, page (In press).

Andersen, P. K. (1991). Survival analysis 1982–1991: the second decade of the proportional hazards regression model. *Statistics in Medicine*, 10(12):1931–1941.

Arce, P. M., Camilon, P. R., Stokes, W. A., Nguyen, S. A., and Lentsch, E. J. (2014). Is sex an independent prognostic factor in cutaneous head and neck melanoma? *The Laryngoscope*, 124(6):1363–1367.

Ash, R. B. and Doleans-Dade, C. A. (2000). *Probability and measure theory*. Academic Press, San Diego, 2 edition.

Bartlett, M. (1953). Approximate confidence intervals. *Biometrika*, 40(1/2):12–19. https://doi.org/10.2307/2333091.

Barui, S. and Grace, Y. Y. (2020). Semiparametric methods for survival data with measurement error under additive hazards cure rate models. *Lifetime Data Analysis*, 26(3):421–450.

Bayarri, M. J. and Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, 19(1):58–80.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515.

Bernardo, J. M. (2005). Intrinsic credible regions: An objective Bayesian approach to interval estimation. *Test*, 14(2):317–384.

Bernardo, J. M. and Smith, A. F. (2000). *Bayesian theory*. John Wiley & Sons, New York, 1 edition.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B*, 11(1):15–53.

Bryson, M. C. and Johnson, M. E. (1981). The incidence of monotone likelihood in the Cox model. *Technometrics*, 23(4):381–383.

Bull, S. B., Mak, C., and Greenwood, C. M. (2002). A modified score function estimator for multinomial logistic regression in small samples. *Computational Statistics & Data Analysis*, 39(1):57–74.

Cai, C., Zou, Y., Peng, Y., and Zhang, J. (2012). smcure: An R-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 108(3):1255–1260.

Castro, M. d., Cancho, V. G., and Rodrigues, J. (2009). A bayesian long-term survival model parametrized in the cured fraction. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 51(3):443–455.

Chen, M.-H., Ibrahim, J. G., and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association*, 94.

Chen, T. and Du, P. (2018). Promotion time cure rate model with nonparametric form of covariate effects. *Statistics in Medicine*, 37(10):1625–1635.

Cherobin, A. C. F. P., Wainstein, A. J. A., Colosimo, E. A., Goulart, E. M. A., and Bittencourt, F. V. (2018). Prognostic factors for metastasis in cutaneous melanoma. *Anais Brasileiros de Dermatologia*, 93(1):19–26.

Chib, S. and Jeliazkov, I. (2005). Accept–reject Metropolis–Hastings sampling and marginal likelihood estimation. *Statistica Neerlandica*, 59(1):30–44.

Cho, M., Schenker, N., Taylor, J. M., and Zhuang, D. (2001). Survival analysis with long-term survivors and partially observed covariates. *Canadian Journal of Statistics*, 29(3):421–436.

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991). Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association*, 86(413):68–78.

Collett, D. (2015). *Modelling survival data in medical research*. Chapman and Hall/CRC, United Kingdom, 1 edition.

Colosimo, E. A. and Giolo, S. R. (2006). *Análise de sobrevivência aplicada*. ABE-Projeto Fisher, São Paulo, 1 edition.

Corbière, F. and Joly, P. (2007). A SAS macro for parametric and semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine*, 85(2):173–180.

Cordeiro, G. M. and McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society, Series B*, 53(3):629–643.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34(2):187–202.

Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, Series B*, 30(2):248–265.

Damato, B., Eleuteri, A., Taktak, A. F., and Coupland, S. E. (2011). Estimating prognosis for survival after treatment of choroidal melanoma. *Progress in Retinal and Eye Research*, 30(5):285–295.

Demarqui, F. N., Dey, D. K., Loschi, R. H., and Colosimo, E. A. (2014). Fully semiparametric Bayesian approach for modeling survival data with cure fraction. *Biometrical Journal*, 56(2):198–218.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–22.

Discacciati, A., Orsini, N., and Greenland, S. (2015). Approximate Bayesian logistic regression via penalized likelihood by data augmentation. *The Stata Journal*, 15(3):712–736.

Dupuis, D. (2001). Fitting log-F models robustly, with an application to the analysis of extreme values. *Computational Statistics & Data Analysis*, 35(3):321–333.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. CRC press, New York, 1 edition.

Ehlers, R. S. (2007). *Inferência Bayesiana*. IME-USP. URL http://conteudo.icmc.usp.br/pessoas/ehlers/bayes/bayes.

Elgmati, E., Fiaccone, R. L., Henderson, R., and Matthews, J. N. (2015). Penalised logistic regression and dynamic prediction for discrete-time recurrent event data. *Lifetime Data Analysis*, 21(4):542–560.

Fang, H.-B., Li, G., and Sun, J. (2005). Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics*, 32(1):59–75.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14(3):257–262.

Fijorek, K. and Sokołowski, A. (2012). Separation-resistant and bias-reduced logistic regression: Statistica macro. *Journal of Statistical Software*, 47(1):1–12.

Firth, D. (1992). Bias reduction, the Jeffreys prior and glim. in in l. fahrmeir, b. francis, r. gilchrist, and g. tutz (eds.), advances in glim and statistical modelling: Proceedings of the glim 92 conference, munich. 78:91–100. URL https://link.springer.com/chapter/10.1007/978-1-4612-2952-0_15.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38.

Gamerman, D. and Lopes, H. F. (2006). *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Chapman and Hall/CRC, New York, 2 edition.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27(15):2865–2873.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian data analysis*. Chapman and Hall/CRC, New York, 3 edition.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Givens, G. H. and Mallick, B. K. (2013). *Computational Statistics*. Wiley Online Library, New Jersey, 2 edition.

Greenland, S. (2007). Prior data for non-normal priors. *Statistics in Medicine*, 26(19):3578–3590.

Greenland, S. and Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, 34(23):3133–3143.

Guo, S. (2010). *Survival analysis*. Oxford University Press, New York, 1 edition.

Hanin, L. and Huang, L.-S. (2014). Identifiability of cure models revisited. *Journal of Multivariate Analysis*, 130:261–274.

Hastings, W. K. (1970). Monte Carlo sampling methods using markov chains and their applications. *The Stata Journal*, 1(57):97–109.

He, Z. and Emura, T. (2019). The COM-Poisson cure rate model for survival data-computational aspects. *Journal of the Chinese Statistical Association*, 57(1):1–42.

Heinze, G. and Ploner, M. (2003). Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. *Computer Methods and Programs in Biomedicine*, 71(2):181–187.

Heinze, G., Ploner, M., Dunkler, D., and Southworth, H. (2013). logistf: Firths bias reduced logistic regression. *R Package Version 2.1*, 1:21.

Heinze, G. and Schemper, M. (2001). A solution to the problem of monotone likelihood in Cox regression. *Biometrics*, 57(1):114–119.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer, New York, 1 edition.

Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association*, 86(416):981–986.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*. John Wiley & Sons, New Jersey, 2 edition.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Kenne Pagui, E., Salvan, A., and Sartori, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika*, 104(4):923–938.

Kenne Pagui, E. C. and Colosimo, E. A. (2020). Adjusted score functions for monotone likelihood in the Cox regression model. *Statistics in Medicine*, 39(10):1558–1572.

Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer, New York, 2 edition.

Kleinbaum, D. G. and Klein, M. (2012). *Survival analysis*. Springer, New York, 2 edition.

Kosmidis, I. (2014). Bias in parametric estimation: reduction and useful side-effects. *Wiley Interdisciplinary Reviews:Computational Statistics*, 6(3):185–196.

Kosorok, M. R., Lee, B. L., Fine, J. P., et al. (2004). Robust inference for univariate proportional hazards frailty regression models. *The Annals of Statistics*, 32(4):1448–1491.

Kuk, A. Y. C. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541.

Lambert, P. and Bremhorst, V. (2019). Estimation and identification issues in the promotion time cure model when the same covariates influence long-and short-term survival. *Biometrical Journal*, 61(2):275–289.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and Jones, D. R. (2005). How vague is vague? a simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24(15):2401–2428.

Lawless, J. F. (2003). *Statistical models and methods for lifetime data*. John Wiley & Sons, New Jersey, 2 edition.

Li, C.-S. and Taylor, J. M. (2002). A semi-parametric accelerated failure time cure model. *Statistics in Medicine*, 21(21):3235–3247.

Li, C.-S., Taylor, J. M., and Sy, J. P. (2001). Identifiability of cure models. *Statistics & Probability Letters*, 54(4):389–395.

Lima, V. M. and Cribari-Neto, F. (2016). Penalized maximum likelihood estimation in the modified extended Weibull distribution. *Communications in Statistics - Simulation and Computation*, 48(2):334–349.

Lin, I.-F., Chang, W. P., and Liao, Y.-N. (2013). Shrinkage methods enhanced the accuracy of parameter estimation using Cox models with small number of events. *Journal of Clinical Epidemiology*, 66(7):743–751.

Liu, M., Lu, W., and Shao, Y. (2006). Interval mapping of quantitative trait loci for time-to-event data with the proportional hazards mixture cure model. *Biometrics*, 62(4):1053–1061.

Liu, X. (2012). *Survival analysis: models and applications*. John Wiley & Sons, Hoboken, 1 edition.

Liu, Y., Hu, T., and Sun, J. (2017). Regression analysis of current status data in the presence of a cured subgroup and dependent censoring. *Lifetime Data Analysis*, 23(4):626–650.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44(2):226–233.

Lu, W. (2007). Maximum likelihood estimation in the proportional hazards cure model. *Annals of the Institute of Statistical Mathematics*, 60(3):545–574.

Lu, W. (2010). Efficient estimation for an accelerated failure time model with a cure fraction. *Statistica Sinica*, 20(2):661–674.

Ma, Y. and Yin, G. (2008). Cure rate model with mismeasured covariates under transformation. *Journal of the American Statistical Association*, 103(482):743–756.

Maller, R. A. and Zhou, S. (1994). Testing for sufficient follow-up and outliers in survival data. *Journal of the American Statistical Association*, 89(428):1499–1506.

Masud, A., Tu, W., and Yu, Z. (2018). Variable selection for mixture and promotion time cure rate models. *Statistical Methods in Medical Research*, 27(7):2185–2199.

Meeker, W. Q. (1987). Limited failure population life tests: application to integrated circuit reliability. *Technometrics*, 29(1):51–65.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Migon, H. S., Gamerman, D., and Louzada, F. (2014). *Statistical inference: an integrated approach*. CRC press, London, 2 edition.

Murphy, S., Rossini, A., and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):968–976.

Murphy, S. A. et al. (1994). Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics*, 22(2):712–731.

Naseri, P., Baghestani, A. R., Momenyan, N., and Akbari, M. E. (2018). Application of a mixture cure fraction model based on the generalized modified weibull distribution for analyzing survival of patients with breast cancer. *International Journal of Cancer Management*, 11(5). https://doi.org/10.5812%2Fijcm.62863.

Nash, J. C. (2014). *Nonlinear Parameter Optimization using R Tools*. Wiley, Chichester, 1 edition.

Nelson, W. (1969). Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1):27–52.

Niu, Y. and Peng, Y. (2013). A semiparametric marginal mixture cure model for clustered survival data. *Statistics in Medicine*, 32(14):2364–2373.

Paek, S. C., Griffith, K. A., Johnson, T. M., Sondak, V. K., Wong, S. L., Chang, A. E., Cimmino, V. M., Lowe, L., Bradford, C. R., Rees, R. S., et al. (2007). The impact of factors beyond breslow depth on predicting sentinel lymph node positivity in melanoma. *American Cancer Society*, 109(1):100–108.

Paulino, C. D., Turkman, M. A. A., and Murteira, B. (2003). *Estatística Bayesiana*. Fundação Calouste Gulbenkian, Lisboa, 1 edition.

Peng, Y. (2003). Estimating baseline distribution in proportional hazards cure models. *Computational Statistics and Data Analysis*, 42(1-2):187–201.

Peng, Y. and Dear, K. B. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243.

Peng, Y., Dear, K. B., and Denham, J. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, 17(8):813–830.

Pianto, D. M. and Cribari-Neto, F. (2011). Dealing with monotone likelihood in a model for speckled data. *Computational Statistics and Data Analysis*, 55(3):1394–1409.

Portier, F., El Ghouch, A., Van Keilegom, I., et al. (2017). Efficiency and bootstrap in the promotion time cure model. *Bernoulli*, 23(4B):3437–3468.

Prentice, R. L. (1975). Discrimination among some parametric models. *Biometrika*, 62(3):607–614.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raiffa, H. and Schlaifer, R. (1961). *Applied statistical decision theory*. Wiley, New York, 1 edition.

Rainey, C. (2016). Dealing with separation in logistic regression models. *Political Analysis*, 24(3):339–355.

Rizzo, M. L. (2007). *Statistical computing with R*. Chapman and Hall/CRC, London, 1 edition.

Robert, C. P., Casella, G., and Casella, G. (2010). *Introducing Monte Carlo methods with R*. Springer, New York, 1 edition.

Roberts, G. O., Gelman, A., Gilks, W. R., et al. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120.

Rosen, O., Jiang, W., and Tanner, M. A. (2000). Mixtures of marginal models. *Biometrika*, 87(2):391–404.

Scharfstein, D. O., Tsiatis, A. A., and Gilbert, P. B. (1998). Semiparametric efficient estimation in the generalized odds-rate class of regression models for right-censored time-to-event data. *Lifetime Data Analysis*, 4(4):355–391.

Schmidt, P. and Witte, A. D. (1989). Predicting criminal recidivism using split population survival time models. *Journal of Econometrics*, 40(1):141–159.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society, Series B*, 43(3):310–313.

Sinha, D., Ibrahim, J. G., and Chen, M.-H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika*, 90(3):629–641.

Sy, J. and Taylor, J. (2001). Standard errors for the Cox proportional hazards cure model. *Mathematical and Computer Modelling*, 33(12-13):1237–1251.

Sy, J. P. and Taylor, J. M. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236.

Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, 51(3):899–907.

Tsodikov, A. D., Ibrahim, J. G., and Yakovlev, A. Y. (2003). Estimating cure rates from survival data: an alternative to two-component mixture models. *Journal of the American Statistical Association*, 98(464):1063–1078.

Wang, P., Tong, X., and Sun, J. (2018). A semiparametric regression cure model for doubly censored data. *Lifetime Data Analysis*, 24(3):492–508.

Wu, J., de Castro, M., Schifano, E. D., and Chen, M.-H. (2018). Assessing covariate effects using Jeffreys-type prior in the Cox model in the presence of a monotone partial likelihood. *Journal of Statistical Theory and Practice*, 12(1):23–41.

Yakovlev, A. Y. and Tsodikov, A. D. (1996). Stochastic model of tumor latency and their biostatistical applications. *World Scientific*. Singapore. https://doi.org/10.1142/2420.

Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of permanent employment in Japan. *Journal of the American Statistical Association*, 87(418):284–292.

Yin, G. and Ibrahim, J. G. (2005). Cure rate models: a unified approach. *Canadian Journal of Statistics*, 33(4):559–570.

Zaeran, E., Azizmohammad Looha, M., Amini, P., Azimi, T., and Mahmoudi, M. (2019). Evaluating long-term survival of patients with esophageal cancer using parametric non-mixture cure rate models. *Journal of Advanced Medical and Biomedical Research*, 27(120):43–50.

Zeng, D., Yin, G., and Ibrahim, J. G. (2006). Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association*, 101(474):670–684.

Zorn, C. (2005). A solution to separation in binary response models. *Political Analysis*, 13(2):157–170.