View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Concordia University Research Repository

Computational Bayesian Methods for Insurance Premium Estimation

Oscar Alberto Quijano Xacur

A Thesis
for The Department of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy (Mathematics) at
Concordia University
Montreal, Quebec, Canada

July 2019

© Oscar Alberto Quijano Xacur, 2019

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: **Oscar Alberto Quijano Xacur**

Entitled: **Computational Bayesian Methods for Insurance Premium Estimation**

and submitted in partial fulfillment of the requirements for the degree of

Doctor Of Philosophy

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

________________________________ Chair

Mary Appezzato

________________________________ External Examiner

Arthur Charpentier

________________________________ External to Program

Dennis Kira

________________________________ Examiner

Mélina Mailhot

________________________________ Examiner

Yogendra P. Chaubey

________________________________ Thesis Supervisor

José Garrido

Approved by Cody Hyndman

Chair of Department or Graduate Program Director

André G. Roy

Dean of Faculty

Date of Defense August 20, 2019

Abstract

Computational Bayesian Methods for Insurance Premium Estimation

Oscar Alberto Quijano Xacur, Ph.D.
Concordia University, 2019

Bayesian Inference is used to develop a credibility estimator and a method to compute insurance premium risk loadings. Algorithms to apply both methods to Generalized Linear Models (GLMs) are provided. We call our credibility estimator the *entropic premium*. It is a Bayesian point estimator that uses the relative entropy as the loss function. The risk measures Value-at-Risk (VaR) and Tail-Value-at-Risk (TVaR) are used to determine premium risk loadings. Our method considers the number of insureds and their durations as random variables. A distribution to model the duration of risks is introduced. We call it *unifed*, it has support on the interval $(0,1)$, it is an exponential dispersion family and it can be used as the response distribution of a GLM.

Contribution from Authors

Bayesian Credibility for GLMs

Oscar Alberto Quijano Xacur	Research, writing, editing and proof-reading.
José Garrido	Research supervisor, funding and proof-reading.

Contents

List of Figures

List of Tables

List of Abbreviations

EDF Exponential Dispersion Family

GLM Generalized Linear Model

MCMC Markov Chain Monte Carlo

HMC Hamiltonian Monte Carlo

PRNG Pseudo Random Number Generator

PPD Posterior Predictive Distribution

SLLN Strong Law of Large Numbers

VaR Value at Risk

TVaR Tail Value at Risk

Notation

- Vectors, unless otherwise stated, are assumed to be column vectors.

- $\log(x)$ means natural logarithm of x.

- Given a positive measure m, it will be said that something is true modulo m, with notation $[m]$, if it is true except on some set E with $m(E) = 0$.

- If μ and ν are measures, we use $\mu \ll \nu$ to express that μ is absolutely continuous with respect to ν. When $\mu \ll \nu$ and $\nu \ll \mu$ we will write $\mu \equiv \nu$ and we will say that μ and ν are equivalent.

- If $(\Omega, \mathcal{F}, m)$ is a measure space, $L^1(\Omega, \mathcal{F}, m)$ is the set of all the $\mathcal{F}$-measurable functions such that
$$\int |f(x)| dm(x) < \infty.$$
When Ω and $\mathcal{F}$ are clear from the context we use $L^1(m)$ instead of $L^1(\Omega, \mathcal{F}, m)$.

- We use $\mathbb{N}$ to denote the natural numbers. It is assumed that they start at 1, i.e. $\mathbb{N} = \{1, 2, \ldots\}$.

- $\mathbb{E}$ and $\mathbb{V}$ are used for the expectation and variance of a random variable, while $\mathbf{V}$ is used to denote the variance function of an Exponential Dispersion Family.

- $N(\mu, \sigma^2)T(a, b)$ denotes the normal distribution with mean μ and variance σ^2 truncated on the interval (a, b).

Chapter 1

Introduction

Insurance is a collective effort to mitigate the risk of some financial loss. The idea is for a group of individuals, the *insureds*, to contribute some money, the *premium*, into a fund called the reserve. The reserve is then used to pay for the losses incurred by the insureds. Thus, the money contributed by many pays for the losses of a few. This allows the premiums to be much lower than the potential loss.

The reserve is managed by an entity called the *insurer.* We refer to the set of insureds as the *insurance portfolio* or simply the *portfolio.* This thesis focuses on non-life insurance where the financial loss is about casualties that occur to a specific good or property (e.g. a car or a house). This object is referred to as the *risk.* The insurance conditions agreed upon by the insured and the insurer are expressed in a contract called the *insurance policy* or simply the *policy.*

One can never be certain that the reserve will be enough to pay for all future losses. Thus, the fundamental question of insurance is: what should the premium be in order for the reserve to be likely to pay for all losses? We leave this concept vague for now. In Chapter 9 we make the phrase “likely to pay” more precise.

Assumptions about the insurance portfolio have to be made in order to answer the fundamental question. In actuarial mathematics there is a set of assumptions that we consider to be a building block for the rest of the theory. These assumptions define what we call the *homoegenous portfolio.*

Definition *Let n be the number of risks in an insurance portfolio and $S_1, \ldots, S_n$ be* *random variables representing the future losses of each risk. We say that the portfolio is* homogeneous *when the S_i's are independent, identically distributed and with finite mean.*

New methods and ideas are usually first tested in the homogeneous portfolio and then generalized to more complex situations. We adhere to this practice along this work.

The Strong Law of Large Numbers (SLLN) gives us a starting point to look for premiums that result in solvent reserves. It tells us that with probability one

$$\frac{S_1 + \cdots + S_n}{n} \to \mu,$$

where $\mu = \mathbb{E}[S_1]$. Actuaries call μ the *pure premium.*

We now know that the pure premium is too low to obtain a solvent reserve. A classical paper that exemplifies the insufficiency of the pure premium is de Finetti (1939), where a gambler's ruin setting based on de Moivre (1756) is used. In Chapter 9 we prove the insufficiency of the pure premium from a different viewpoint.

The premium should then be the pure premium plus some additional amount. We refer to this additional amount as the *risk loading*. Of course, there has been a lot of effort to determine how much should the risk loading be. Premium principles and risk measures have been used for this purpose (see for instance Young (2006) and Hardy (2006)).

The areas of study in actuarial mathematics presented so far propose models that assume some known underlying probability distribution or at least some known quantities about them like the mean or a quantile. When one wants to apply such models to a real problem it is necessary to estimate these quantities based on past data. Thus we enter in the realm of statistics.

The theory of statistics tells us that estimations based on few observations are not reliable. Credibility theory is the branch of actuarial mathematics that aims to solve the problem of estimating the pure premium when there are few observations. Chapter 1 gives an account of the existing methods and ideas in credibility theory.

An important extension of the homogeneous portfolio has been the relaxation of the identically distributed part. Such portfolios are called heterogeneous. Different approaches have been proposed in the actuarial literature, but all of them (to the best of our knowledge) have a common factor: they start by dividing the portfolio into (approximately) homogeneous groups. This is known as *segmenting* the portfolio or simply *segmentation.* Nowadays Generalized Linear Models (GLMs) are widely used in practice for this purpose. They are introduced in Chapter 3.

The new contributions of this thesis are contained in Chapters 6, 7, 8 and 9.

Chapter 6 talks about linear credibility for GLMs. Linear estimators have played a central role in credibility theory. *Exact* linear credibility means choosing a prior whose posterior mean is a linear function of the sample average. Previously existing linear credibility results for GLM are not exact. The research for this thesis started with the search of a prior that would give linear credibility for GLMs. We ended up proving that such a prior does not exist, which is the main result of Chapter 6.

Chapter 7 introduces entropic credibility. It is a Bayesian point estimator that uses the relative entropy as loss function instead of the usual square error loss. The relative entropy and its properties are introduced in Chapter 4.

In Chapter 8 we present a new probability distribution: the unifed. It has support on $(0, 1)$ and it can be used as the response distribution of a GLM. We propose it for modelling the duration of policies in an insurance portfolio.

In Chapter 9 we propose methods to compute the premium risk loading using GLMs.

Along this work Bayesian statistics are used for parameter estimation. Partly because it is a natural framework for credibility problems but also because we find the interpretation of Bayesian estimates more natural than the frequentist one. In Chapter 5 we give a brief introduction to Bayesian statistics and we discuss the differences of interpretation between the Bayesian and the frequentist approach. Markov Chain Monte Carlo (MCMC) algorithms are necessary for the application of all our results in any practical situation. They are introduced in Chapter 2.

We think of a model as a lens through which one looks at reality. In consequence, we see model assessment as a tool for understanding what aspects of reality are replicated reasonably well by a model and which ones are not. For this reason, when assessing the goodness of fit, we have deliberately avoided the use of asymptotic distributions that arise from the assumption that the chosen parametric family is the *true* one. We prefer the use of replicated samples, which we introduce in Chapter 6.

In the applied examples we focus on illustrating new concepts and algorithms and we have left cross-validation out of the examples. Nevertheless, we do recommend the use of cross-validation in any predictive modeling context where our proposed procedures are used.

The examples given in Chapters 7, 8 and 9 were coded in R (R Core Team (2019)) and stan (Stan Development Team (2018)). We created an R package that can be used to reproduce the results of all our examples. It can be downloaded from `https://gitlab.com/oquijano/mythesis`.

Chapter 2

Credibility Theory

Usually the first step taken by insurers is to segment their portfolio into homogeneous classes. Then it proceeds to estimate the risk premium and loading for each class. The first essential problem in credibility theory can be formulated in this way: how many observations are needed in a class in order for such estimations to be reliable (in some sense)?

Assuming that we can answer this question, let us introduce some terminology. For those classes in which there is a large enough number of observations we say that the estimation is *credible* or that we have *full credibility.* Otherwise we say that the estimation is not credible or that we have *partial credibility.* Additionally, we will call *criterion for full credibility* or *full credibility criterion* any method that aims to answer this question.

The second essential problem in credibility theory is: given a group for which we do not have sufficient policyholders for a fully credible estimation, how can we get a reliable estimation of the pure premium?

Following the development of credibility theory, we make the distinction between three different types of estimation.

We call *credibility estimator* or *credibility premium* the final estimation that we consider more reliable than the sample average $\bar{S}_n$.

We call *empirical estimator* or *empirical premium*, an estimator coming from a sample of the population of interest.

We call *manual estimator* or *manual premium* an out-of-sample estimator. Typically it comes from previous or out-of-sample experience of the same insurer with similar risks or from pooled information from different insurers.

Traditionally, credibility estimators have been obtained by combining in some way empirical and manual estimators. We call *credibility formula* any expression involving the data and the manual estimator with the purpose of obtaining a credibility estimator.

2.1 Overview of Credibility Results

The historical development of credibility theory can be divided into two parts. On the one hand we have results concerning criteria for full credibility and on the other hand we have those that assume partial credibility and develop a credibility formula.

Early articles on full credibility criteria for the pure premium are Mowbray (1914) and Whitney (1918). Both of these papers develop models in the context of workers compensation. A well known paper with a more general scope is Perryman (1932). The criteria introduced in this paper depends on two parameters k and p which are interpreted as "the observed pure premium should be within $100k\%$ of the expected pure premium with probability p". In this sense, it is said that an estimation $\hat{\mu}$ of μ is fully credible (k, p) if

$$\mathrm{P}(|\hat{\mu} - \mu| \leq k\mu) \geq p. \tag{2.1}$$

By using the Central Limit Theorem, Perryman assumed that there were enough observations for $\hat{\mu}$ to be normally distributed and in this way he found for what minimum sample size (2.1) is satisfied. Mayerson et al. (1968) generalized Perryman's work by dropping the assumption of normality for $\hat{\mu}$. They did it in a distribution-free manner. Only some moments of the distribution are required. They achieved this using the Cornish-Fisher expansion. Many decades later Schmitter (2004) was the first to give a full credibility criterion for pure premiums estimated by GLMs.

With regards to partial credibility, for a long time premium formulas were not derived mathematically and heuristic methods were used. These eventually gave birth to the most developed and popular credibility formula: the linear credibility premium.

Let us denote by $\hat{\mu}$ the empirical premium, M the manual premium and P the credibility premium. Linear credibility consists in setting

$$P = z\hat{\mu} + (1 - z)M, \tag{2.2}$$

for some credibility weight $z \in [0, 1]$. This formulation is intuitive and easy to interpret, it is a weighted average between the empirical premium and the manual premium. Having formulated (2.2) the problem is now to find an "optimal" value of z to use. The first mathematically sound method for finding z is given in Bühlmann (1967). In that article, it is assumed that there is some random parameter θ on which the distribution of S_1 depends. Let $\mu(\theta) = \mathbb{E}[S_1|\theta]$ and $\sigma^2(\theta) = \mathbb{V}(S_1|\theta)$. Under this setting and given a sample $S_1, S_2, \ldots, S_n$, Bühlmann found for which values of a and b the objective function

$$\mathbb{E}\left[\left(\mathbb{E}[\mu(\theta)|S_1, \ldots, S_n] - [a + b\bar{S_n}]\right)^2\right]$$

is minimized. He found that the best approximation is given by $b = \frac{n}{n+k}$ and $a = (1 - b)\mathbb{E}[\mu(\theta)]$, where $k = \frac{\mathbb{E}[\sigma^2(\theta)]}{\mathbb{V}[\mu(\theta)]}$. This justifies and gives meaning to (2.2) with $z = \frac{n}{n+k}$ and $M = \mathbb{E}[\mu(\theta)]$. A strong point of this result is that the only things needed about the prior distribution of θ are $\mathbb{E}[\mu(\theta)]$, $\mathbb{E}[\sigma^2(\theta)]$ and $\mathbb{V}[\mu(\theta)]$ and no further assumptions about the shape of this prior distribution are made.

The down side of this result is that it is an approximation, and we do not know its accuracy. An important paper that shed some light on this issue is Jewell (1974). In this paper, the loss distribution is assumed to belong to some natural exponential family, i.e. its density or probability function is given by

$$f(y|\theta) = a(y) \exp(\theta y - \kappa(\theta)), \qquad \theta \in \Theta, y \in \mathcal{Y} \tag{2.3}$$

for some parameter space Θ, set $\mathcal{Y}$ and real-valued functions a and κ. As in Bühlmann's result, let us assume that θ is a random variable over Θ, Jewell considered the following prior

$$\pi_{n_0,x_0}(\theta) \propto \exp(n_0\{x_0\theta - \kappa(\theta)\}), \qquad \theta \in \Theta, \tag{2.4}$$

where n_0 and x_0 are some parameters. The possible values for x_0 and n_0 depend on κ. Note that the parametrization used here for (2.3) and (2.4) is not the same one Jewell used; ours is taken from Diaconis and Ylvisaker (1979). Given a sample $S_1, S_2, \ldots, S_n$ of size n from an exponential family distribution (2.3), Jewell showed that (2.4) is a conjugate prior with posterior parameters $n_0 + n$ and $(n_0 x_0 + \sum_{i=1}^{n} S_i)/(n_0 + n)$, corresponding to n_0 and x_0, respectively. For $\mu(\theta) = \mathbb{E}[S_1|\theta]$, he proved that the prior mean is given by

$$\mathbb{E}[\mu(\theta)] = x_0,$$

and therefore, the corresponding posterior mean is

$$\mathbb{E}[\mu(\theta)|S_1, S_2, \ldots, S_n] = \left(\frac{n_0}{n_0 + n}\right) x_0 + \left(\frac{n}{n_0 + n}\right) \bar{S}_n. \tag{2.5}$$

This shows that for the distributions considered by Jewell, the credibility premium in (2.5) is linear in $S_1, \ldots, S_n$. Thus, for these cases, Bühlmann's formula is not an approximation, it is exact.

The Bühlmann-Straub model (Bühlmann and Straub (1970)), generalized these results by considering that not all observations are equally precise. They did it by introducing some weights for each observation. Their credibility factor depends on the sum of the weights rather than the number of observations.

Jewell's hierarchical model (Jewell (1975)) generalized further by introducing a tree structure for dividing the portfolio into homogeneous groups before computing the credibility estimators.

The Bayesian methods discussed so far use the usual square distance loss function for obtaining point estimators. In Gómez Déniz (2006), linear credibility estimators are found using weight balanced loss functions which have the form

$$L(a, x) = wh(x)(\delta_0 - a)^2 + (1 - w)h(x)(x - a)^2,$$

where $w \in (0, 1)$ must be fixed, $h(x)$ is some positive weight function and δ_0 is some function of the observed data. Najafabadi (2010) considered best linear approximations to Bayesian point estimators coming from some arbitrary loss function ρ. Let δ_π be the Bayesian point estimator coming from the prior π and loss function ρ. They focused on minimizing

$$\mathbb{E}[(\delta_\pi(X) - \delta_\alpha(X))^2],$$

where $\delta_\alpha(X) = \alpha\bar{X} + (1-\alpha)\mu$ for some $\alpha \in (0, 1)$, and $\mu := \mathbb{E}_\pi[\theta]$. They found such estimator under certain conditions for the loss function and the loss distribution.

The main developments in credibility theory have remained centered around linear credibility and exponential families. For instance, there is a generalization of Jewell's result for multivariate exponential families. It can be found in Diaconis and Ylvisaker (1979). There have also been some developments outside of this mainstream. For example De Vylder (1996) developed a theory for non-linear credibility.

2.2 Credibility for Regression Models

There exist some credibility results for regression models in the literature. We discuss here those of Hachemeister (1975) and De Vylder (1985), reviewing only their results with regard to credibility formulae.

Hachemeister considered linear regression for different classes of policyholders. He did this in order to find the different inflation trends on worker's compensation claims among different states in the US.

For all the classes the covariates and also the design matrix which we denote with $\boldsymbol{X}$ are the same. For class j, the mean of the response vector $\boldsymbol{Y_j}$ depends on some random parameter θ_j through the relation

$$\mathbb{E}[\boldsymbol{Y_j}|\theta_j] = \boldsymbol{X}\boldsymbol{\beta}(\theta_j), \tag{2.6}$$

where $\boldsymbol{\beta}(\theta_j)$ is a vector of regression coefficients for the class. The credibility estimator of $\boldsymbol{\beta}(\theta_j)$, say $\boldsymbol{B_j}$, is of the form

$$\boldsymbol{B_j} = \boldsymbol{Z_j}\hat{\boldsymbol{\beta}}_j + (\boldsymbol{I} - \boldsymbol{Z_j})\boldsymbol{b}, \tag{2.7}$$

where $\boldsymbol{Z_j} = \mathrm{diag}(z_{j1}, \ldots, z_{jk})$ is a matrix with credibility weights on the diagonal, $\hat{\boldsymbol{\beta}}_j$ is the estimated vector of regression coefficients for class j, before credibility, and $\boldsymbol{b} = \mathbb{E}[\boldsymbol{\beta}(\theta_j)]$. The mean $\boldsymbol{b}$ does not change among different classes because the θ_j's are assumed to be identically distributed. With this setup, Hachemeister's procedure consists in finding the best $\boldsymbol{Z_j}$. He does this by finding the $\boldsymbol{Z_j}$ that minimizes the distance between $\boldsymbol{B_j}$ and $\boldsymbol{\beta}(\theta_j)$. The distance used is the one induced by an inner product of the form $\langle \boldsymbol{U}, \boldsymbol{V} \rangle = \mathbb{E}[\boldsymbol{U}^T \boldsymbol{\Sigma} \boldsymbol{V}]$, where $\boldsymbol{\Sigma}$ is some positive definite matrix specified in the model.

De Vylder generalized this by allowing non-linear functions of the regression coefficients, i.e. he replaced (2.6) with

$$\mathbb{E}[\boldsymbol{Y}|\theta_j] = \boldsymbol{f}(\boldsymbol{\beta}(\theta_j)), \tag{2.8}$$

for some function $\boldsymbol{f}$. The credibility formula De Vylder used in this model is still of the same form as in (2.7), and the method used to find the optimal weight matrix is also similar.

Pitselis (2004) extends Hachemeister's and De Vylder's regression models by using robust inference methods. It is important to point out that all these models are distribution free and consist in finding the best linear approximation under some criteria.

Credibility for GLMs

Even though we have not yet introduced GLMs, it is possible to review the existing credibility results for GLMs.

For this purpose we only need the following characteristics of GLMs:

- A GLM is a regression model, in which some distribution is assumed for the response vector Y.
- It is a nonlinear regression model in which the mean vector, $\boldsymbol{\mu} = \mathbb{E}[Y|X]$, is related to the linear predictor by the equation

  $$\boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta}, \tag{2.9}$$

  where $\boldsymbol{g}$ is some one-to-one map called the *link function.*

There are just a few credibility results for GLMs, some propose a full credibility criteria and others propose a partial credibility formula.

To the best of our knowledge only two articles consider a full credibility criteria for GLMs: Schmitter (2004) and Garrido and Zhou (2009).

The existing results for finding a partial credibility estimators rely on the introduction of random effects (or factors). This requires a modification of (2.9) as

$$\boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{T}\boldsymbol{u},$$

where $\boldsymbol{T}$ is some given matrix and $\boldsymbol{u}$ is a random vector whose entries are called random effects. There are several articles that treat this subject, for instance Nelder and Verrall (1997) and Antonio and Beirlant (2007). In Ohlsson (2008) random effects are used for obtaining linear credibility estimators, but they only used variables that are a Multi Level Factor (MLF), which they define as a categorical variable that

1. Has many classes with few observations in some or all of them.

2. The classes do not posses any inherent ordering and therefore there is no simple way to join some of them in order to increase the number of observations in each segment.

Chapter 3

Monte Carlo Methods

Monte Carlo (MC) methods are a class of computational algorithms based on random sampling. Their most common application are to numerical integration and optimization problems. Here we focus on the numerical integration part.

MC integration is justified by the Strong Law of Large Numbers (SLLN). The following is a version of the SLLN written in a MC suggestive way.

Theorem 3.1. *Let $\{X_n\}_{n\in\mathbb{N}}$ be an independent and identically distributed (iid) sequence of random variables in some probability space $(\Omega, \mathcal{F}, \mathrm{P})$. Let g be a map such that $g(X_1) \in L^1(\mathrm{P})$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} g(X_i) \to \int g(x)d\mathrm{P}(x) \qquad [\mathrm{P}]. \tag{3.1}$$

Where [P] means that this is true with probability one with respect to P. Thus, one way of approximating an integral that involves a probability measure P is the following:

1. A large sample of size N, $\{x_n\}_{n=1}^{N}$, is simulated from the desired distribution.
2. The function g whose integral we want to approximate is applied to each simulated value. In this way the sequence $\{g(x_n)\}_{n=1}^{N}$ is obtained.
3. The average $\frac{1}{N}\sum_{n=1}^{N} g(x_n)$ is computed to approximate $\int g(x)d\mathrm{P}(x)$.

The phrase "a large sample of size N" in Step 1 is somewhat imprecise. It is in fact not possible to determine with total certainty how big N should be for the error to be smaller than some desired bound. Nevertheless it is possible to compute an asymptotic confidence interval for the value of the integral. Details for the univariate case can be found in Section 3.2 of Robert and Casella (2004).

Example 3.1. *Let us approximate the integral of $\int_0^1 x^2 dx$ with the MC method. Notice that this is the expectation of U^2 where U has a uniform distribution on the interval $(0, 1)$. For this purpose we simulate 10,000 independent uniform random variables and take the average of their squares. This can be done in the statistical software R with two lines of code:*

```
x <- runif(1E4)
mean(x^2)
```

```
0.330421339950455
```

We see that the approximation is close to the real value of the integral 1/3.

The method exposed here assumes that one is able to simulate random variables from the distribution of interest. Often with Bayesian methods it happens that one has a function that is proportional to the density of interest i.e. we have the density up to a normalizing constant. Other times we may have the density but we are unable to use it to get simulated values from it. This can happen for example when it is hard to evaluate the inverse of the cdf. In these cases it is not possible to use the MC method as explained in this section. However, such cases can be approached with a Markov Chain Monte Carlo (MCMC) method.

3.1 Markov Chains

For completeness, in this section we give a brief introduction to Markov chains in a general state space. We restrict ourselves to the concepts needed to introduce MCMC. Our exposition is based on Athreya and Lahiri (2006) and Robert and Casella (2004).

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(\mathcal{X}, \mathcal{B})$ some measurable space. Let $\{X_n\}_{n \geq 0}$ be a sequence of random variables from Ω to S, and for each $n \geq 0$ let $\mathcal{F}_n = \sigma\langle X_0, \ldots, X_n \rangle$, i.e. $\mathcal{F}_n$ is the $\sigma-$algebra generated by the random variables $X_0, \ldots, X_n$.

Definition 3.1. *The sequence of random variables* $\{X_n\}_{n\geq 0}$ *is called a Markov chain if for any* $A \in \mathcal{B}$,

$$\mathrm{P}(X_{n+1} \in A|\mathcal{F}_n) = \mathrm{P}(X_{n+1} \in A|\sigma\langle X_n\rangle) \qquad [\mathrm{P}], \tag{3.2}$$

for all $n \geq 0$ *and for any initial distribution* P_0 *of* X_0.

We focus on Markov chains that have a transition probability function.

Definition 3.2. *A function* $P : \mathcal{X} \times \mathcal{B} \to [0, 1]$ *is called a transition probability function on* S *if*

i) For all $x \in S$, $P(x, \cdot)$ *is a probability measure on* $(\mathcal{X}, \mathcal{B})$.

ii) For all $A \in \mathcal{B}$, $P(\cdot, A)$ *is a* $\mathcal{B}-$*measurable function from* $\mathcal{X}$ *to* $[0, 1]$.

Definition 3.3. *We say that a Markov chain* $\{X_n\}_{n\geq 0}$ *has transition function* $\mathrm{P}(\cdot, \cdot)$ *if*

$$\mathrm{P}(X_{n+1} \in A|\sigma\langle X_n\rangle) = P(X_n, A), \qquad \textit{for all } n \in \mathbb{N}. \tag{3.3}$$

It has been proved that Markov chains that satisfy some general conditions have a transition function. In what follows the existence of the transition function is assumed.

From (3.3), we see that for any n, $P(x, A)$ is the probability of $X_{n+1} \in A$ given $X_n = x$. In other words, given that the chain is in x, $P(x, A)$ is the probability of entering A in the next step. It is also possible to find the probabilities of entering some set after n steps. For this purpose let us define the sequence of functions $\{P^{(n)}(\cdot, \cdot)\}_{n\geq 0}$ as follows

$$P^{(n)}(x, A) = \begin{cases} I_A(x) & \text{if } n = 0, \\ \int_S \mathrm{P}^{(n-1)}(y, A)\mathrm{P}(x, dy) & \text{if } n \geq 1, \end{cases} \tag{3.4}$$

with $P^{(1)}(\cdot, \cdot) = P(\cdot, \cdot)$. It can be proven that given $X_0 = x$,

$$\mathrm{P}(X_n \in A) = P^{(n)}(x, A), \qquad \text{for all } n \geq 0.$$

This motivates the following definition.

Definition 3.4. $P^{(n)}(\cdot, \cdot)$ *as defined in* (3.4) *is called the n-step transition function generated by* $P(\cdot, \cdot)$.

We are interested in two specific aspects about the behaviour of Markov chains, which we formulate as questions: is there a probability measure π for which

1. given a π-measurable function f
$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \to \int f d\pi? \tag{3.5}$$
and

2. are there some conditions under which
$$P^{(n)}(x,\cdot) \to \pi(\cdot), \qquad \text{for all } x \in \mathcal{X}? \tag{3.6}$$

In what follows we introduce sufficient conditions for (3.5) and (3.6), and we also specify the type of convergence for both of them. The measure π for which we can have these properties is the *stationary distribution.*

Definition 3.5. *A probability measure π on $(\mathcal{X}, \mathcal{B})$ is called stationary for a transition function $P(\cdot,\cdot)$ if*
$$\pi(A) = \int P(x, A)\pi(dx), \quad \textit{for all } A \in \mathcal{B}.$$

Let us talk now about *irreducibility*, which is related to the question: given some $A \in \mathcal{B}$, is it possible that the chain enters A at some point? We start by formalizing the phrase "entering A".

Definition 3.6. *Let $\{X_n\}_{n\geq 0}$ be a Markov chain taking values in the measurable space $(\mathcal{X}, \mathcal{B})$. For any $A \in \mathcal{B}$ the first entrance time to A is defined as*
$$T_A = \begin{cases} \infty & \textit{if } X_n \notin A \textit{ for all } n, \\ \min\{n : n \geq 1, X_n \in A\} & \textit{otherwise.} \end{cases}$$

Now, a reference measure ϕ is needed for defining the sets A we are interested in.

Definition 3.7. *Let ϕ be a non-zero $\sigma-$finite measure on $(\mathcal{X}, \mathcal{B})$. A Markov chain $\{X_n\}_{n\geq 0}$ taking values in $(\mathcal{X}, \mathcal{B})$ is ϕ-irreducible, or Harris irreducible with reference measure ϕ, if for any $A \in \mathcal{B}$*
$$\phi(A) > 0 \Rightarrow \mathrm{P}_x(T_A < \infty) > 0, \qquad \textit{for all } x \in \mathcal{X},$$
where $\mathrm{P}_x(T_A < \infty) = \mathrm{P}(T_A < \infty | X_0 = x)$.

Thus, a ϕ-irreducible chain enters the ϕ-non-null sets with positive probability.

Recurrence is a stronger case of irreducibility. Given $A \in \mathcal{B}$, will the chain enter A at some point with probability one? Again here we use a reference measure for defining the sets of interest.

Definition 3.8. *Let ϕ be a non-zero $\sigma-$finite measure on $(\mathcal{X}, \mathcal{B})$. A Markov chain is ϕ-recurrent if for $A \in \mathcal{B}$,*

$$\phi(A) > 0 \Rightarrow \mathrm{P}_x(T_A < \infty) = 1, \qquad \textit{for all } x \in \mathcal{X}.$$

Definition 3.9. *A Markov chain is* Harris recurrent *if it is ϕ-recurrent for some measure ϕ.*

We are now able to state sufficient conditions for (3.5).

Theorem 3.2. *Let $\{X_n\}_{n\geq 0}$ be a Harris recurrent Markov chain on $(\mathcal{X}, \mathcal{B})$ and transition function $P(\cdot, \cdot)$. Assume that $\mathcal{B}$ is countably generated and that π is a stationary probability measure for $P(\cdot, \cdot)$. Then*

1. *π is unique.*

2. *For all $f \in L^1(\mathcal{X}, \mathcal{B}, \pi)$ and $x \in \mathcal{X}$,*

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \rightarrow \int f d\pi \qquad [P_x].$$

One extra condition is needed in order to guarantee (3.6). Suppose there is a partition of the space $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_{d-1}$, with $d \geq 2$ such that if for some n, $X_n \in \mathcal{X}_i$, then with probability one $X_{n+1} \in \mathcal{X}_{[(i+1) \mod d]}$ (here $a \mod b$ means the remainder of a divided by b). For a chain to satisfy (3.6), such a partition must not exist, in other words the chain has to be *aperiodic*.

Definition 3.10. *A Markov chain with transition probability function P is* aperiodic *if there is no partition $\mathcal{X} = \mathcal{X}_0 \cup \mathcal{X}_1 \cup \ldots \cup \mathcal{X}_{d-1}$ with $d \geq 2$, such that*

$$\begin{aligned} & P(x, \mathcal{X}_0) = 1, && \textit{for all } x \in \mathcal{X}_{d-1}. \\ \textit{and} \qquad & P(x, \mathcal{X}_{i+1}) = 1, && \textit{for all } x \in \mathcal{X}_i, \textit{ where } i \in \{0, \ldots, d-2\}. \end{aligned}$$

The type of convergence that we can obtain for (3.6) is for the total variation distance.

Definition 3.11. *Given two probability measures μ and ν on some probability space $(\mathcal{X}, \mathcal{B})$, the total variation distance between μ and ν is given by*

$$\|\mu - \nu\|_{TV} = \sup_{A \in \mathcal{B}} |\mu(A) - \nu(A)|.$$

Theorem 3.3. *Let $\{X_n\}_{n \geq 0}$ be a Markov chain that satisfies the conditions of Theorem 3.2. If $\{X_n\}_{n \geq 0}$ is also aperiodic, then*

$$\|P^{(n)}(x, \cdot) - \pi(\cdot)\|_{TV} \to 0, \qquad \text{as } n \to \infty.$$

3.2 Markov Chain Monte Carlo

In this section we show how Markov chains can be used to obtain simulations from an arbitraty distribution π.

The idea of MCMC is to simulate a Markov chain whose stationary distribution is π and for which (3.5) and (3.6) hold. There is more than one way to achieve this and to our knowledge the most popular MCMC methods are the Metropolis-Hastings algorithm, the Gibss sampler and Hamiltonian Monte Carlo.

With the purpose of showing the reader how one can simulate a Markov chain with some desired stationary distribution we present now the simplest (and yet brilliant) of the three methods mentioned above: the Metropolis-Hastings algorithm. It is first necessary to choose what is called a *proposal distribution*. This is a measurable function $q(\cdot|\cdot) : (\mathcal{X} \times \mathcal{X}, \mathcal{B} \times \mathcal{B}) \to [0, \infty)$ such that for each x, $\int q(y|x) dm(y) = 1$.

In the MCMC jargon, π is called the *target distribution*. Let $(\mathcal{X}, \mathcal{B})$ be the measurable space where π is defined and assume that there is a $\sigma-$finite measure m such that $d\pi(x) = f(x) dm(x)$ for some function f.

Given the target and proposal distribution, the algorithm consists in creating a Markov chain $\{X_n\}$ as follows.

Metropolis-Hastings algorithm.

1. Given $X_n = x$, generate a random variable Y_n from the density $q(\cdot|x)$.

2. Take
$$X_{n+1} = \begin{cases} Y_n & \text{with probability } p(x, Y_n) \\ X_n & \text{with probability } 1 - p(x, Y_n), \end{cases}$$
where
$$p(x, y) = \min\left\{\frac{f(y)q(x|y)}{f(x)q(y|x)}, 1\right\}.$$

Notice that at each step, the algorithm either accepts the proposed value Y_n with probability $p(x, Y_n)$ or it stays at the same value.

Definition 3.12. *When a Metropolis-Hastings algorithm is run, the proportion of times that the algorithm accepts the proposed value Y_n is called the* acceptance rate*.*

In order to have (3.5) and (3.6), we need the chain to be aperiodic and Harris recurrent. In what follows we discuss sufficient conditions for which this happens. Proofs can be found in Section 7.3.2 of Robert and Casella (2004).

A sufficient condition for a Metropolis chain $\{X_n\}_{n\geq 0}$ to be aperiodic, is that for some n, the event $\mathrm{P}(X_{n+1} = X_n) > 0$. This is equivalent to

$$\mathrm{P}(f(X_n)q(X_n|Y_n) < f(Y_n)q(X_n|Y_n)) < 1. \tag{3.7}$$

A practical way to check this condition is to simulate the chain and check that the acceptance rate is less than one. A sufficient condition for the chain to be Harris recurrent is that

$$q(x|y) > 0 \text{ for every } (x, y) \in \mathcal{X} \times \mathcal{X}. \tag{3.8}$$

From the discussion above and Theorems 3.2 and 3.3, the following result follows.

Theorem 3.4. *Let $\{X_n\}$ be a Markov chain generated with the Metropolis algorithm with target distribution π satisfying* (3.7) *and* (3.8)*. Then, for any $f \in L^1(\pi)$ and $x \in S$,*

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \to \int f d\pi \qquad [P_x]$$

$$\text{and} \qquad \|\mathrm{P}^{(n)}(x, \cdot) - \pi(\cdot)\| \to 0, \qquad \text{as } n \to \infty.$$

Notice that $p(x, y)$ depends on f through $\frac{f(y)}{f(x)}$. Therefore the algorithm works even if we know f only up to a multiplicative constant. This is very useful in Bayesian statistics since often one has the posterior distribution specified only up to a normalizing constant.

In this thesis we mainly use Hamiltonian Monte Carlo (HMC). It is harder to implement (if one has to do it on its own), but it converges faster to the stationary distribution than the Metropolis-Hastings algorithm, specially in higher dimensions. Specifically we use the R interface to stan (Stan Development Team (2018)), which allows for Bayesian inference using HMC (although the HMC details are transparent for the user, one only needs to define the model). We do not explain the details of HMC here; we consider Neal (2010) to be a good reference.

Convergence Diagnostics

We discuss now how to assess the convergence of simulated chains. The methods shown here work for any MCMC method. They help checking whether (3.2) and (3.3) have occurred.

When MCMC is performed, there is a first batch of simulations that are not kept. This is called the *burnin* or *warmup* period and its purpose is to bring the n-step transition function close to the target distribution (as in (3.6)). After this is done, one cannot be sure that the chain has converged to the target distribution. Nevertheless there are some tests that provide evidence in favor or against the convergence. Here we discuss three: the traceplot, the running mean and the autocorrelation.

For each test, we give an example suggesting convergence and another one where it suggests the contrary. For the examples we use Jewell's prior for the gamma distribution.

When the gamma distribution is expressed with the exponential dispersion family(EDF) parametrization (4.1) it takes the following form:

$$f(x|\theta, \phi) = \frac{x^{\frac{1}{\phi}-1}}{\Gamma(\frac{1}{\phi})\phi^{\frac{1}{\phi}}} \exp\left(\frac{\theta x + \log(-\theta)}{\phi}\right), \qquad \phi > 0, \theta < 0, x > 0. \tag{3.9}$$

From the properties of EDF's (discussed in Chapter 4), if X has density (3.9)

$$\mathbb{E}[X] = \mu = -\frac{1}{\theta}, \qquad \mathbb{V}(X) = \phi\mu^2 = \frac{\phi}{\theta^2}.$$

Then, according to (2.4), Jewell's prior for the gamma distribution has density

$$\pi_{n_0,x_0}(\theta) \propto (-\theta)^{n_0} e^{n_0 x_0 \theta}, \qquad \theta < 0,\, n_0, x_0 > 0. \tag{3.10}$$

This is one of the common cases in which there is a constant missing for the density to integrate to 1 and where MCMC is very convenient. In the examples that follow we use (3.10) as target density with $x_0 = 200$ and $n_0 = 20$.

Traceplot The traceplot is a graph of the simulated values against time. When convergence has been reached, the plot should look like an i.i.d. plot, i.e. the observations should not seem correlated and they should be taking values in all the regions in the support of the target distribution. When this is the case it is usually said that the chain is *mixing well.* Figure 3.1 was generated using the Metropolis algorithm with the target density in (3.10). The proposal distribution is normal with mean equal to the current state of the chain and variance 0.00001. An example of a chain that does not mix well is given in Figure 3.2. It was generated with a chain similar to the one used for Figure 3.1, with the only difference that a variance of 0.01 (instead of 0.00001) was used for the proposal distribution.

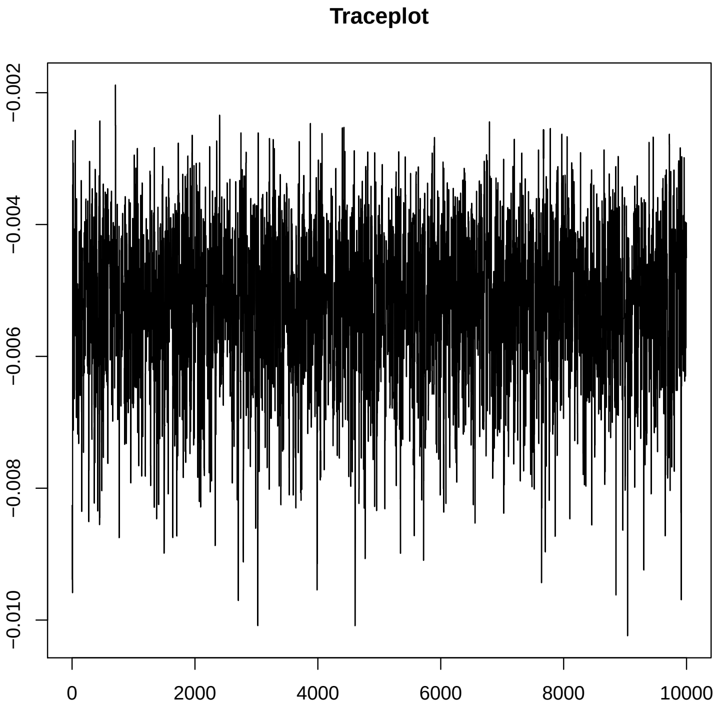


Figure 3.1: Good mixing example

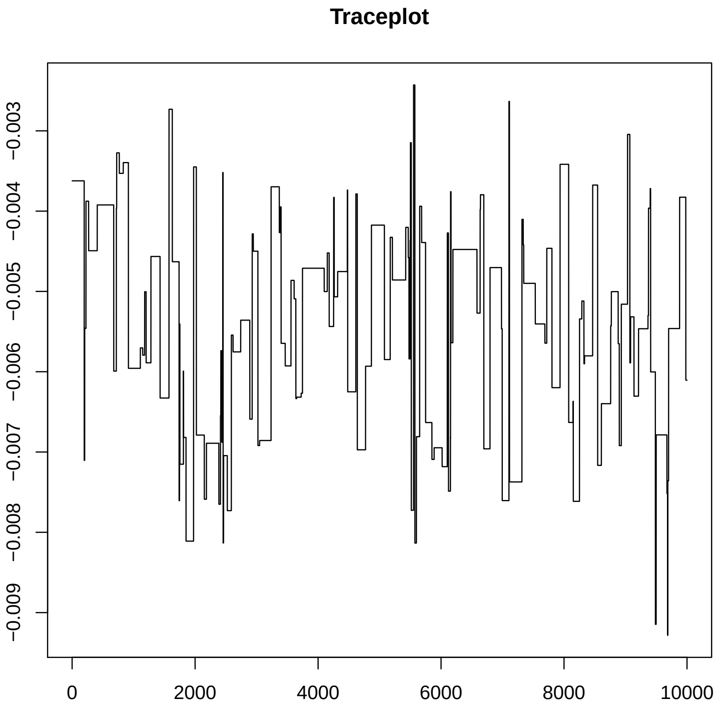


Figure 3.2: Bad mixing example

Running Means The running mean at time n is the mean of the first n values of the simulated chain. A graph of the running mean against time shows whether the mean appears to stabilize as the number of simulations increase. When this is the case, it is evidence that a convergence of type (3.5) is being reached. For this graph it is useful to run parallel chains (this is, to simulate independent chains with the same transition kernel) and plot their running means together to see if they all stabilize at the same value. Figure 3.3 shows the running graph of two parallel chains. They were simulated using stan.

Autocorrelation Plots The k-th lag autocorrelation is defined as

$$\rho_k = \frac{\sum_{i=1}^{n-k}(X_i - \bar{X})(X_{i+k} - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

It gives the correlation between every simulated value X_i and its k-th lag X_{i+k}. For a converging chain, we expect to see ρ_k get closer to zero as k increases. When this is not the case, it is a signal of bad mixing. Figure 3.4 shows an autocorrelation plot example for a converging chain and for a bad mixing one.

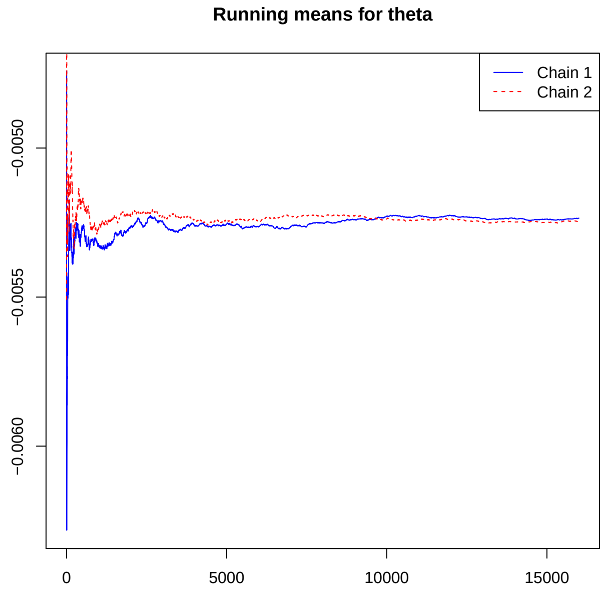


Figure 3.3: Running means of two parallel chains

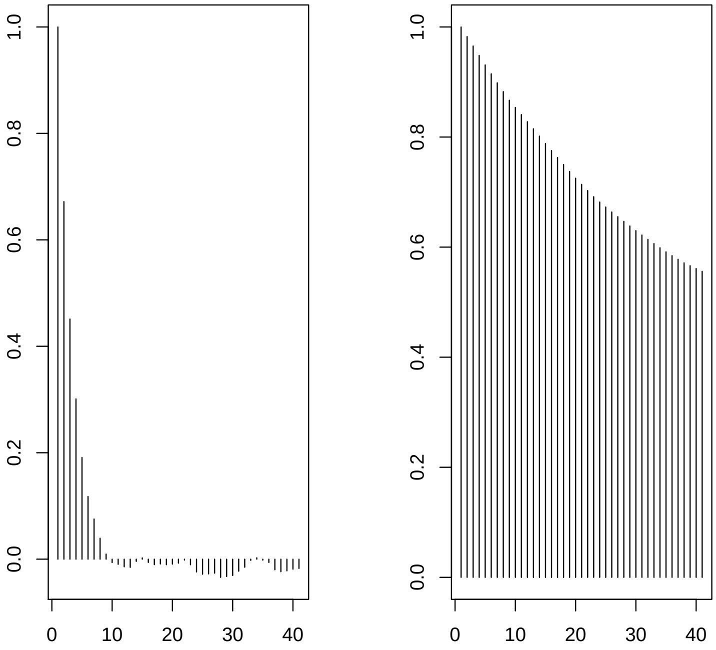
3.3 Monte Carlo Methods and Pseudo Random Numbers

To introduce the Monte Carlo method we used the SLLN (Theorem 3.1) as a justification. Nevertheless the “random” numbers given by a computer come from deterministic algorithms that depend only on a initial parameter called the seed. These algorithms are called *Pseudo Random Number Generators (PRNGs)*. A very common methodology for simulating random numbers from a specific desired distribution is the following:

Simulation Method 1 (SM1)

1. A PRNG is used to simulate a sequence $\{x_i\}_{i=1}^n$ of independent uniform(0,1) numbers.

2. A transformation T is applied to each simulated number in the sequence above. T is chosen in such a way that if $U \sim$ uniform(0,1), then $T(U)$ has the desired distribution. In this way we obtain the sequence $\{T(x_i)\}_{i=1}^n$ and this is our simulated sequence from the desired distribution.



(a) Converging chain

(b) Slow mixing chain

Figure 3.4: Autocorrelation plots

Example 3.2. *Let us assume that we have a PRNG that simulates independent uniform(0,1) variates and we want to simulate 5 numbers coming from an exponential distribution with mean 1. It can be seen that the transformation $T(u) = -\log(1-u)$ has the characteristic explained in Step 2 above. Thus, we first generate 5 numbers with the PRNG. In R this can be achieved with the following code:*

```
(u <- runif(5))
```

```
[1] 0.4299841 0.1144878 0.8944409 0.2644449 0.7343649
```

This is our simulated uniform sequence. Now we apply T to each number above:

```
-log(1-u)
```

```
[1] 0.5620910 0.1215890 2.2484845 0.3071298 1.3256318
```

and this is a simulation from an exponential(1) distribution. Notice that R has its own function **rexp** *for generating an exponential distribution, but it uses a different algorithm.*

Now, why do MC methods work with numbers generated from these algorithms? It is clear that the SLLN does not apply since the numbers are not truly random (whatever this means). The answer to this question comes from convergence results in dynamical systems. In what follows we outline how this works. Our exposition is based on Góra and Boyarsky (1997, Chap. 3).

Definition 3.13. *Let $(\Omega, \mathcal{F}, m)$ be a probability space. A measurable transformation $\tau : \Omega \to \Omega$ is said to preserve m if $m(\tau^{-1}(A)) = m(A)$ for all $A \in \mathcal{F}$.*

Definition 3.14. *Let $(\Omega, \mathcal{F}, m)$ be a probability space and let $\tau : \Omega \to \Omega$ preserve m. The quadruple $(\Omega, \mathcal{F}, m, \tau)$ is called a dynamical system.*

Definition 3.15. *A measure preserving transformation $\tau : (\Omega, \mathcal{F}, m) \to (\Omega, \mathcal{F}, m)$ is called ergodic if for any $A \in \mathcal{F}$ with $\tau^{-1}(A) = A$, either $m(A) = 0$ or $m(A^c) = 0$.*

In dynamical systems the properties of sequences defined with successive applications of some map τ are studied, i.e. sequences of the type

$$x_0, \tau(x_0), (\tau \circ \tau)(x_0), \ldots, \tau^{\circ n}(x_0), \ldots$$

where $\tau^{\circ n}$ means the composition of τ with itself n times and x_0 is some initial value. The following is a corollary of Birkhoff's ergodic theorem. It is the main result of our discussion.

Theorem 3.5. *Let $(\Omega, \mathcal{F}, m, \tau)$ be a dynamical system with τ ergodic and let g be map in $L^1(m)$. Then for $m-$almost every x_0*

$$\frac{1}{n} \sum_{i=1}^{n} g(\tau^{\circ i}(x_0)) \to \int g dm. \tag{3.11}$$

This theorem gives sufficient conditions for Monte Carlo integration to converge with a sequence of pseudo random numbers. Let us see how this is the case for SM1.

Assume that we want to use the Monte Carlo method to approximate $\mathbb{E}[h(Y)]$ for some measurable function h and a random variable Y that follows some distribution D. Assume also that $\mathbb{E}[|h(Y)|] < \infty$, so the L^1 condition is satisfied. According to the MC methodology we generate N simulations $\{y_n\}_{n=0}^{N-1}$ from the distribution D and then we compute $\sum_{n=0}^{N-1} h(y_n)/N$.

Let us now use SM1 to generate $\{y_n\}_{n=0}^{N-1}$. Suppose you have a dynamical system with the Lebesgue measure on $[0,1]$ as the invariant measure. Let x_0 be an initial value for which (3.11) is true, and let T be such that if $U \sim \text{uniform}(0,1)$, then $T(U) \sim D$. Then a simulated sequence of D is given by $T(x_0), T(\tau(x_0)), \ldots, T(\tau^{\circ(N-1)}(x_0))$. By Theorem 3.5, we have that

$$\frac{1}{N}\sum_{i=0}^{N-1}(h \circ T)(\tau^{\circ i}(x_0)) \to \int_0^1 (h \circ T)(x)dx.$$

By the change of variable $y = T(x)$, we have that $\int_0^1 (h \circ T)(x)dx = \int h dm_D = \mathbb{E}[h(Y)]$, where m_D is the probability measure associated with the distribution D. Thus, if we take $y_n = T(\tau^{\circ n}(x_0))$, for $n = 0, \ldots, N-1$, we get that

$$\sum_{n=0}^{N-1} h(y_n)/N \to \mathbb{E}[h(Y)],$$

which shows that the method converges to the right value.

Chapter 4

Exponential Families and GLMs

In practice insurance portfolios are never homogeneous. An important part of modern pricing is to segment portfolios into aproximately homogeneous groups. From a modelling point of view this allows us to use the knowledge of the properties of homogeneous groups. A *fairness* argument is also sensible here: each individual should pay according to the risk they represent. This argument is used for in Bühlmann (1967) where it is formulated as follows: "... each class of risk with *equal observed risk performance* should pay its own way".

Generalized Linear Models (GLMs) are widely used in practice for segmenting heterogeneous portfolios and estimate the pure premiums of the resulting classes. In this chapter we give an introduction to GLMs where we emphasize the properties needed to develop our credibility estimate in Chapter 8.

4.1 Exponential Dispersion Families

The definitions and results from this section are based on Jørgensen (1997). A reproductive Exponential Dispersion Family (EDF), is a collection of probability distributions with densities of the form

$$f(y|\theta, \lambda) = a(y, \lambda) \exp(\lambda(\theta y - \kappa(\theta))), \qquad \theta \in \Theta, \lambda \in \Lambda, y \in \mathcal{Y}, \tag{4.1}$$

where θ and Θ are called the canonical parameter and canonical space, respectively, λ and Λ are the index parameter and set, respectively, and the support of f is $\mathcal{Y} \subset \mathbb{R}$. Also, Θ must be an interval and $\Lambda \subset (0, \infty)$. κ is a function called the cumulant generator of the family; it is assumed twice continuously differentiable and $\kappa'' > 0$.

Throughout this text whenever the terms *exponential families* or *reproductive families* are used, they refer to (4.1).

Many well known continuous and discrete families of distributions, or transformations of these can be written as an exponential family. The following table shows some examples of well known families of distributions and their respective values of θ, λ, Θ, Λ and κ when written in the form (4.1).

Distribution	θ	λ	Θ	Λ	$\kappa(\theta)$
Binomial(n, p)	$\ln\left(\frac{p}{1-p}\right)$	n	$\mathbb{R}$	$\mathbb{N}$	$\ln\left(\frac{1+\exp(\theta)}{2}\right)$
Poisson(λ)	$\ln(\lambda)$	$--$	$\mathbb{R}$	$--$	$\exp(\theta) - 1$
Gamma(α, β)	$1 - \beta$	α	$(-\infty, 1)$	$\mathbb{R}$	$\ln\left(\frac{1}{1-\theta}\right)$

GLMs allow to fit regression models with a response from an exponential family. Some properties of reproductive EDF's are presented in what follows. We focus on those properties that are essential for the development of GLMs.

The first property concerns the likelihood equation. Assume that we have a random sample from (4.1). Writing the likelihood equation, one can see that the mle for θ does not depend on λ. This property is exploited in estimation procedures for GLMs.

Let Y be a random variable whose density can be written as in (4.1). A neat property of the reproductive families is that for $\theta \in \text{int}\Theta$ (here int stands for interior),

$$\mathbb{E}[Y] = \dot{\kappa}(\theta) \qquad \text{and} \qquad \mathbb{V}[Y] = \frac{\ddot{\kappa}(\theta)}{\lambda}, \tag{4.2}$$

where $\dot{\kappa} = \kappa'$ and $\ddot{\kappa} = \dot{\kappa}' = \kappa''$. From these two properties we see that the mean and variance are strongly related through κ. The variance function of the family highlights this dependence.

Definition 4.1. *Given a reproductive exponential dispersion family, the mean domain of the family is defined as*

$$\Omega = \{\mu = \dot{\kappa}(\theta) : \theta \in \text{int}\Theta\},$$

which consists of all the means of the family for which (4.2) *holds.*

Definition 4.2. *The unit variance function of an exponential family is the function* $\mathbf{V} : \Omega \to (0, \infty)$*, with*

$$\mathbf{V}(\mu) = (\ddot{\kappa} \circ \dot{\kappa}^{-1})(\mu).$$

Remark 4.1. *We use the symbol* $\mathbb{V}$ *for the variance of a random variable while* $\mathbf{V}$ *is used for the variance function of an exponential family.*

Two important properties of the unit variance function are:

1. $\mathbb{V}[X] = \dfrac{\mathbf{V}(\mu)}{\lambda}$. The name unit variance comes from the fact that $\mathbb{V}[X] = \mathbf{V}(\mu)$ for $\lambda = 1$.

2. The unit variance function characterizes the family, i.e. two different exponential dispersion families cannot have the same unit variance function.

Now consider $\dot{\kappa}$ and Θ. The function $\dot{\kappa}$ is always continuous and one-to-one. By the continuity of $\dot{\kappa}$, as Θ is an interval, then so is Ω. Since $\mu = \dot{\kappa}(\theta)$ and $\dot{\kappa}$ is invertible, when Θ is an open interval, we can reparametrize the family in terms of $(\mu, \lambda) \in \Omega \times \Lambda$. This is called the mean value parametrization of the family. When Θ is not open, i.e. the interval contains at least one of its endpoints, the mean value parametrization can be extended by continuity to the endpoints. Thus, in this way, it is always possible to reparametrize an exponential family with the mean value parametrization.

The support of an EDF depends only on the value of λ. For a given family, let C_λ be the convex support of any member of the family with index parameter λ. We define the convex support of the family as

$$C = \bigcup_{\lambda \in \Lambda} C_\lambda.$$

Definition 4.3. *The unit deviance function of an exponential dispersion family, is defined as* $d : C \times \Omega \to [0, \infty)$ *with*

$$d(y, \mu) = 2 \left[\sup_{\theta \in \Theta} \{\theta y - \kappa(\theta)\} - y \dot{\kappa}^{-1}(\mu) + \kappa(\dot{\kappa}^{-1}(\mu)) \right]. \tag{4.3}$$

The unit deviance function plays a very important role in the theory of GLMs. In fact, the model assessment of a GLM is through hypothesis tests that are based on the asymptotic behavior of this unit deviance function. Some of its important properties are:

1. The unit deviance and variance functions are related by the equation

$$\frac{\partial^2}{\partial \mu^2} d(\mu, \mu) = \frac{2}{\mathbf{V}(\mu)}.$$

2. The unit deviance function characterizes the family.

3. The mean value parametrization of a reproductive exponential dispersion family can be written as

$$p(y; \mu, \lambda) = c(y, \lambda) \exp\left(-\frac{\lambda}{2} d(y, \mu) \right), \tag{4.4}$$

for some function c.

Regular exponential families are an important particular case of exponential families.

Definition 4.4. *A reproductive exponential dispersion model is called regular if its canonical space Θ is open.*

Regular families have some important properties that will be used later:

1. For any given family, we have $\Omega \subset C_\lambda$, and hence $\Omega \subset C$. For regular families we have $\Omega = C$.

2. When $y, \mu \in \Omega$, the unit deviance function can be written as

$$d(y, \mu) = 2 \left[y\{\dot{\kappa}^{-1}(y) - \dot{\kappa}^{-1}(\mu)\} - \kappa(\dot{\kappa}^{-1}(y)) + \kappa(\dot{\kappa}^{-1}(\mu)) \right], \tag{4.5}$$

which is easier to work with than the original definition since the sup in (4.3) disappears so (4.5) can be used as the definition of the unit deviance for regular families.

3. For $y, \mu \in \Omega$, the deviance can be written as

$$d(y, \mu) = 2 \int_\mu^y \frac{(y - t)}{\mathbf{V}(t)} dt. \tag{4.6}$$

As in the previous point, for regular families (4.6) is equivalent to (4.3).

4.1.1 Weights and Data Aggregation

There is a decomposition of the index parameter in (4.1) that has been appropriate in several contexts; one of them being GLMs. It consists in taking $\lambda = \frac{w}{\phi}$. w is known as the weight and ϕ as the dispersion parameter. The weight is assumed to be known and it is not considered a new parameter of the distribution. Thus, allowing ourselves the little abuse of notation of using the same function name a, (4.1) becomes

$$f(y|\theta, \phi) = a(y, \phi) \exp\left(\frac{w}{\phi}\{y\theta - \kappa(\theta)\}\right). \tag{4.7}$$

There is a useful property of reproductive exponential dispersion families parametrized as above that allows for data aggregation. Jørgensen's notation (from Jørgensen (1997)) is very convenient for expressing this property: given a fixed exponential family, if Y has mean μ and density given by (4.7), we say that it is $ED(\mu, \phi/w)$ distributed. The property is then as follows: if $Y_1, Y_2, \cdots, Y_n$ are independent, and $Y_i \sim ED(\mu, \phi/w_i)$, then

$$\bar{Y} = \frac{w_1 Y_1 + \cdots + w_n Y_n}{w_+} \sim ED(\mu, \phi/w_+), \qquad w_+ = \sum_{i=1}^{n} w_i. \tag{4.8}$$

4.1.2 A Note on Aggregating Discrete Exponential Dispersion Models

There are two usual parametrizations of exponential dispersion families. (4.1) gives the density of *reproductive* EDFs and it is used for continuous distributions. Discrete distributions are usually parametrized as *additive* EDFs, whose densities have the form

$$f(y|\theta, \phi) = a(y, \phi) \exp\left(y\theta - \lambda\kappa(\theta)\right), \qquad \theta \in \Theta, \lambda \in \Lambda. \tag{4.9}$$

Both parametrizations are defined and discussed in Jørgensen (1997) and Jørgensen (1992). An aggregation property different than (4.8) holds for additive EDFs. In Jørgensen's notation, given a fixed additive EDF with density (4.9) and mean μ, we say that it follows a $ED^*(\mu, \lambda)$. If $Y_1, \ldots, Y_n$ are independent and $Y_i \sim ED^*(\mu, \lambda_i)$, then

$$Y_1 + \cdots + Y_n \sim ED^*(\mu, \lambda_+), \qquad \lambda_+ = \lambda_1 + \cdots + \lambda_n.$$

As shown in the next section, GLMs assume the reproductive parametrization (see also Nelder and Wedderburn (1972)). Now, for many discrete EDFs, the dispersion parameter has a known value. Specifically, for the Poisson, Bernoulli and negative binomial distributions $\Lambda = \{1\}$. This makes (4.1) and (4.9) the same parametrization and allows such distributions to enter the GLMs framework. Nevertheless, it is important to be aware that for GLMs with a discrete response distribution, one cannot aggregate data using (4.8). The properties of the Poisson distribution allow to use an offset for this purpose (see for example Kaas et al. (2008)) and quasi-likelihood can be used for other discrete distributions.

4.2 GLMs

In a GLM the response variable is assumed to follow a reproductive exponential dispersion family with density (4.7). Notice that since (4.7) is equivalent to (4.1) with $\lambda = \frac{w}{\phi}$, then the mean and variance of the response variable can be expressed as $\mu = \kappa'(\theta)$ and $\sigma^2 = \phi\kappa''(\theta)/w$, respectively. It is further assumed that there is a vector of explanatory variables, also known as covariates, $\boldsymbol{x} = (x_1, \cdots, x_p)$, a vector of coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ and a function g such that

$$g(\mu) = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p. \tag{4.10}$$

It is useful for further developments to express the canonical parameter θ in terms of the coefficients. Since $\mu = \dot{\kappa}(\theta)$ then:

$$\begin{aligned}(g \circ \dot{\kappa})(\theta) &= \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p \\ \theta &= (g \circ \dot{\kappa})^{-1}(\beta_0 + x_1\beta_1 + \cdots + x_p\beta_p).\end{aligned} \tag{4.11}$$

Notice that the population can be divided into different classes according to the values of the explanatory variables. Thus, given a sample, we can group together all the observations that share the same values of explanatory variables and aggregate them with (4.8). It is important to mention that with this grouping there is no loss of information for estimating the mean since $\bar{Y}$ is a sufficient statistic for θ (but not for ϕ, thus some information is lost for the estimation of ϕ).

After aggregating, let m be the number of classes and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_m)$ a vector whose entries are the different values of θ over all classes. Each class is assumed independent of all others and therefore the density of the sample can be expressed as

$$f(\boldsymbol{y}|\boldsymbol{\theta},\phi) = \boldsymbol{a}(\boldsymbol{y},\phi)\exp\left(\frac{\boldsymbol{y}^T W\boldsymbol{\theta} - \mathbf{1}^T W\boldsymbol{\kappa}(\boldsymbol{\theta})}{\phi}\right), \qquad \boldsymbol{y}\in\mathbb{R}^m, \tag{4.12}$$

where $\boldsymbol{\kappa}(\boldsymbol{\theta}) = (\kappa(\theta_1),\cdots,\kappa(\theta_m))$, $W = \operatorname{diag}(w_1,\cdots,w_m)$, w_i is the sum of all the weights in the i-th class, $\mathbf{1} = (1,\cdots,1)$ and $\boldsymbol{a}(\boldsymbol{y},\phi) = \prod_{i=1}^m (a(y_i,\frac{w_i}{\phi}))$. In order to express $\boldsymbol{\theta}$ in terms of $\boldsymbol{\beta}$, we define the following maps

$$\boldsymbol{\mu} = \dot{\boldsymbol{\kappa}}(\boldsymbol{\theta}) = \begin{pmatrix} \dot{\kappa}(\theta_1) \\ \vdots \\ \dot{\kappa}(\theta_m) \end{pmatrix}, \quad G(\boldsymbol{\mu}) = G\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} = \begin{pmatrix} g(\mu_1) \\ \vdots \\ g(\mu_m) \end{pmatrix},$$

and the design matrix

$$X = \begin{pmatrix} 1 & \boldsymbol{x_1}^T \\ & \vdots \\ 1 & \boldsymbol{x_m}^T \end{pmatrix},$$

where $\boldsymbol{x}_i$ is the vector of explanatory variables for the i-th class. With all this definitions, we have that

$$\begin{aligned} G(\boldsymbol{\mu}) &= X\boldsymbol{\beta}, \\ (G\circ\dot{\boldsymbol{\kappa}})(\boldsymbol{\theta}) &= X\boldsymbol{\beta}, \\ \boldsymbol{\theta} &= (G\circ\dot{\boldsymbol{\kappa}})^{-1}(X\boldsymbol{\beta}). \end{aligned} \tag{4.13}$$

When $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ have the same dimension we say that the model is *saturated.* In this case one can find a value of $\boldsymbol{\beta}$ for which the predicted means are equal to the observed means. In practical applications the dimension of $\boldsymbol{\beta}$ is usually less than the dimension of $\boldsymbol{\mu}$. This is called a *non-saturated* model.

It is useful to reparametrize (4.12) in terms of the mean vector $\boldsymbol{\mu}$ instead of $\boldsymbol{\theta}$. Using the mean value parametrization (see (4.4)), (4.12) can be reparametrized as

$$f(\boldsymbol{y}|\boldsymbol{\mu},\phi) = \boldsymbol{c}(y,\phi)\exp\left(-\frac{1}{2\phi}D(\boldsymbol{y},\boldsymbol{\mu})\right), \tag{4.14}$$

where $\boldsymbol{c}(\boldsymbol{y},\phi) = \prod_{i=1}^m c(y_i,\frac{\phi}{w_i})$, and

$$D(\boldsymbol{y},\boldsymbol{\mu}) = \sum_{i=1}^m w_i d(y_i,\mu_i). \tag{4.15}$$

D is called the deviance of the model. We give here some of its properties:

- Given a sample, finding the mle of $\boldsymbol{\theta}$ is equivalent to finding the value of $\boldsymbol{\beta}$ that minimizes the deviance.

- D can be used to estimate the dispersion parameter (although it is not the only method). The deviance estimator of ϕ is given by
$$\hat{\phi} = \frac{D(\boldsymbol{y}, \boldsymbol{\mu})}{n - p}.$$

- The asymptotic distribution of D plays an important role in model assessment and variable selection.

For further details about the use and properties of the deviance we recommend Jørgensen (1992).

Remark 4.2. *GLMs allow categorical and continuous variables. In an insurance context their different values divide the population into homogeneous groups. Since it is desirable to have numerous observations in each group, it is a common practice to divide continuous variables into intervals so they can be treated as categorical.*

Chapter 5

Relative entropy

In this section, let m_i, $i = 1, 2$ be probability measures with $dm_i(x) = f_i(x)ds(x)$, for some functions f_1, f_2 and some probability measure s. It is also assumed that $m_1 \equiv m_2 \equiv s$, where $m_1 \equiv m_2$ means that m_1 and m_2 are equivalent measures.

Definition 5.1. *The relative entropy of m_2 from m_1 is defined as*

$$D(m_1||m_2) = \mathbb{E}_{m_1}\left[\log\left(\frac{f_1(X)}{f_2(X)}\right)\right] = \int \log\left(\frac{f_1(x)}{f_2(x)}\right) dm_1(x). \tag{5.1}$$

The definition above was introduced by Kullback and Leibler (1951). $D(\cdot||\cdot)$ is often called the Kullback-Leibler divergence, although what they called divergence is the sum $D(m_1||m_2) + D(m_2||m_1)$ ($I(1:2) + I(2:1)$ in their own notation). Therefore we have decided to use another common name used for (5.1): relative entropy.

We give now a statistical interpretation of this definition. It is taken from Kullback (1968), Chapter 1. Assume we have two hypotheses H_1 and H_2 about the distribution of X; H_i is the hypothesis that X is distributed according to m_i, $i = 1, 2$. Assume also prior probabilities $\mathrm{P}(H_i)$, $i = 1, 2$ for these hypotheses. One can show that the posterior probabilities $\mathrm{P}(H_i|X)$ satisfy

$$\mathrm{P}(H_i|X) = \frac{\mathrm{P}(H_i)f_i(X)}{\mathrm{P}(H_1)f_1(X) + \mathrm{P}(H_2)f_2(X)} \qquad [s], \qquad i = 1, 2.$$

From this relation, we can obtain

$$\log\left(\frac{f_1(X)}{f_2(X)}\right) = \log\left(\frac{\mathrm{P}(H_1|X)}{\mathrm{P}(H_2|X)}\right) - \log\left(\frac{\mathrm{P}(H_1)}{\mathrm{P}(H_2)}\right). \tag{5.2}$$

Now, $\frac{\mathrm{P}(H_1)}{\mathrm{P}(H_2)}$ and $\frac{\mathrm{P}(H_1|X)}{\mathrm{P}(H_2|X)}$ are the prior and posterior odds of H_1, respectively. Hence, the right hand side in (5.2), is the difference between the logarithm of the odds of H_1 after and before observing the value of X. Thus, given $X = x$, the likelihood ratio $\log\left(\frac{f_1(x)}{f_2(x)}\right)$ is defined as the information for discriminating in favour of H_1 against H_2. $D(m_1||m_2)$ is the integral with respect to m_1 of the left hand side in (5.2). Therefore, $D(m_1||m_2)$ is the mean information in favor of H_1 against H_2 per observation from m_1.

5.1 Properties

Additivity

The relative entropy of a vector of independent random variables is equal to the sum of the relative entropies of the marginal distributions. In other words, if H_i states that for measurable sets A and B, $m_i(A \times B) = m_{ix}(A)m_{iy}(B)$ for $i = 1, 2$, then

$$D(m_1||m_2) = D(m_{1x}||m_{2x}) + D(m_{1y}||m_{2y}).$$

Convexity

Theorem 5.1. $D(m_1||m_2) \geq 0$ *with equality if and only if* $m_1 = m_2$.

Invariance

Let $(\Omega_1, \mathcal{F}, m_i)$ and $(\Omega_2, \mathcal{G}, \nu_i)$, for $i = 1, 2$, be probability spaces and $T : \Omega_1 \to \Omega_2$ a measurable transformation such that $\nu_i(G) = m_i(T^{-1}(G))$, for $G \in \mathcal{G}$. Define also $\gamma(G) = s(T^{-1}(G))$. Since $m_1 \equiv m_2 \equiv s$, then $\nu_1 \equiv \nu_2 \equiv \gamma$. This implies, by Radon-Nykodim's theorem that there exist g_1 and g_2 such that

$$\nu_i(G) = \int_G g_i(y) d\gamma(y), \qquad G \in \mathcal{G}.$$

With these definitions in mind, the following theorem asserts the invariance property of the relative entropy. Its proof can be found in Chapter 2 of Kullback (1968).

Theorem 5.2. $D(m_1||m_2) = D(\nu_1||\nu_2)$ *if and only if* T *is a bijective transformation.*

Lower Bound: The Total Variation Distance

The relative entropy is not a metric since it is not symmetric and it does not satisfy the triangle inequality. Nevertheless it is usually interpreted as a measure of how different two probability measures are. It also provides a bound on the total variation between two probability measures.

Theorem 5.3. *Let $d_{TV}(P,Q) = \sup_{B\in\mathcal{F}} |\mathrm{P}(B) - Q(B)|$ be the total variation distance in some measurable space $(\Omega, \mathcal{F})$, for two probability measures P and Q. Then*

$$D(P||Q) \geq \frac{d_{TV}(P,Q)^2}{2}.$$

A proof of this result can be found in Kemperman (1969).

5.2 Relative Entropy and Maximum Likelihood Estimation

Consider the case when a parametric model is assumed in some given situation. Let Θ be the parameter space and ν_θ the model's probability measure for a given parameter $\theta \in \Theta$. In this section we give a proof that, asymptotically, a maximum likelihood estimator (mle) minimizes the entropy between the true distribution and the assumed model. This property was introduced in Akaike (1973), but the result is only commented in the paper with no formal proof nor specifying sufficient conditions. We give in this section a set of assumptions under which the result is true and prove it.

Let m be the true probability measure, the first assumption is:

(A1) For every $\theta \in \Theta$, m and ν_θ are absolutely continuous with respect to some common probability measure s.

Due to this assumption there exists a measurable function f such that $dm(x) = f(x)ds(x)$, and for every $\theta \in \Theta$ there exists f_θ such that $d\nu_\theta(x) = f_\theta(x)ds(x)$. This takes us to our second assumption:

(A2) $\log(f_\theta(X)) \in L^1(m)$ for every $\theta \in \Theta$.

Given that $X_1, \cdots, X_n$ are iid and m-distributed, the log-likelihood function for our assumed model is defined as

$$\ell_n(\theta) = \sum_{i=1}^{n} \log(f_\theta(X_i)).$$

Notice that by Assumption 2 and the strong law of large numbers, we have that

$$\frac{\ell_n(\theta)}{n} \longrightarrow \mathbb{E}_m[\log(f_\theta(X_1))] \qquad [m]. \tag{5.3}$$

For every $N \in \mathbb{N}$, define D^N as

$$D^N(m||\nu_\theta) = \mathbb{E}_m[\log(f(X_1))] - \frac{\ell_N(\theta)}{N},$$

and note that by (5.3) we have for every $\theta \in \Theta$, that

$$D^N(m||\nu_\theta) \longrightarrow D(m||\nu_\theta) \qquad [m]. \tag{5.4}$$

Let $\hat{\theta}_n$ be a maximum likelihood estimator of θ. In other words,

$$\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}}\, \ell_n(\theta),$$

which also means then that $\hat{\theta}_N$ minimizes $D^N(m||\nu_\theta)$ over the set $\theta \in \Theta$. We state now the third assumption:

(A3) There exists an mle for θ for every $n \in \mathbb{N}$.

The fourth and final assumption simply states that it is possible to minimize the entropy between the true distribution and the assumed model:

(A4) There exists $\theta_0 \in \Theta$ such that for every $\theta \in \Theta$, $D(m||\nu_{\theta_0}) \leq D(m||\nu_\theta)$.

Remark 5.1. *It is not being assumed that m belongs to the hypothesized model; i.e. we do not suppose that there exists $\theta^* \in \Theta$ such that $f = f_{\theta^*}$.*

Notation 5.1. *To simplify notation in the following proofs, $D^N(\theta)$ and $D(\theta)$ are used for $D^N(m||\nu_\theta)$ and $D(m||\nu_\theta)$, respectively, through the remaining part of this section.*

In the results that follow, we make reference to the sets A, $\{B_n\}_{n\in\mathbb{N}}$ and B defined as follows. A is the set of all the elements in the parameter space that minimize D. i.e.

$$A = \{\theta_0 \in \Theta : D(\theta_0) \leq D(\theta) \text{ for all } \theta \in \Theta\}. \tag{5.5}$$

For each $n \geq 1$,

$$B_n = \{\theta \in \Theta : D(\theta) - D(\theta_0) > 1/n \text{ for some } \theta_0 \in \mathbb{N}\}, \tag{5.6}$$

and then define B as

$$B = \bigcup_{n\in\mathbb{N}} B_n = \{\theta : D(\theta) - D(\theta_0) > 0 \text{ for } \theta_0 \in A\}. \tag{5.7}$$

Notice that for every n, $B_n \subset B_{n+1}$ and that $A = B^c$.

Lemma 5.1. *Let $\theta_1 \in B$ and $\theta_0 \in A$. Then m-a.s., there exists $M \in \mathbb{N}$ such that for every $N \geq M$,*

$$\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{f_{\theta_0}(X_i)}{f_{\theta_1}(X_i)}\right) > 0. \tag{5.8}$$

Proof. (5.8) can be expressed in terms of D^N as follows

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{f_{\theta_0}(X_i)}{f_{\theta_1}(X_i)}\right) &= \frac{1}{N}\sum_{i=1}^{N}\log(f_{\theta_0}(X_i)) - \frac{1}{N}\sum_{i=1}^{N}\log(f_{\theta_1}(X_i)) \\
&= \left(\mathbb{E}_m[\log(f(X_1))] - \frac{1}{N}\sum_{i=1}^{N}\log(f_{\theta_1}(X_i))\right) \\
&\quad - \left(\mathbb{E}_m[\log(f(X_1))] - \frac{1}{N}\sum_{i=1}^{N}\log(f_{\theta_0}(X_i))\right) \\
&= D^N(\theta_1) - D^N(\theta_0).
\end{aligned} \tag{5.9}$$

By (5.4), we have that

$$(D^N(\theta_1) - D^N(\theta_0)) \to (D(\theta_1) - D(\theta_0)), \qquad \text{as } N \to \infty \qquad [m]. \tag{5.10}$$

Since $\theta_1 \in B$, there exists $n \in \mathbb{N}$ such that $\theta_1 \in B_n$. Also, by (5.10), m-a.s. there exists $M \in \mathbb{N}$ such that for every $N \geq M$

$$|(D^N(\theta_1) - D^N(\theta_0)) - (D(\theta_1) - D(\theta_0))| < \frac{1}{2n}.$$

By the definition of B_n, this implies that

$$D^N(\theta_1) - D^N(\theta_0) \geq \frac{1}{2n} > 0.$$

□

Theorem 5.4. *Let $\theta_1 \in B$ and $\theta_0 \in A$. Then, m-a.s., there exists $M \in \mathbb{N}$ such that for $N \geq M$,*

$$D^N(\theta_0) \leq D^N(\theta_1).$$

Proof. By Lemma 5.1, there exists M such that for $N \geq M$,

$$\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{f_{\theta_0}(X_i)}{f_{\theta_1}(X_i)}\right) > 0,$$

then,

$$\begin{aligned}
D^N(\theta_1) &= \mathbb{E}_m[\log(f(X_1))] - \frac{1}{N}\sum_{i=1}^{N}\log(f_{\theta_1}(X_i)) \\
&= \mathbb{E}_m[\log(f(X_1))] - \frac{1}{N}\sum_{i=1}^{N}\log(f_{\theta_0}(X_i)) \\
&+ \frac{1}{N}\sum_{i=1}^{N}\log(f_{\theta_0}(X_i)) - \frac{1}{N}\sum_{i=1}^{N}\log(f_{\theta_1}(X_i)) \\
&= D^N(\theta_0) + \frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{f_{\theta_0}(X_i)}{f_{\theta_1}(X_i)}\right) \\
&\geq D^N(\theta_0).
\end{aligned}$$

□

Chapter 6

Bayesian Models

In this chapter we give a brief introduction to Bayesian statistics. This is a very broad topic, so we focus on the ideas and methods that we consider indispensable for the Bayesian Credibility developed in later chapters. We assume that the reader is familiar with frequentist (non-Bayesian) point estimation, confidence intervals and hypothesis tests, as well as the concept of prior and posterior distributions from Bayesian statistics. Our exposition emphasizes the differences of interpretation between the frequentist and Bayesian approaches. The interpretation parts are based on Chapter 1 of Verbraak (1990), and Sections 6.4 and 6.5 are based on Chapter 6 of Gelman et al. (2014)

Along this chapter the following notation is used: θ is the parameter for which inference is needed, it can take values in the set Θ and can either be a scalar or a vector. In order to avoid introducing excessive notation θ can either be a fixed value or a random variable. We believe that the context in which it is used makes it clear in each case. Y is a random variable (possibly a vector), that can take values in the set $\mathcal{Y}$ and whose density $f(y|\theta)$ depends on the parameter θ. Here $\pi(\theta)$ is the prior density of θ. n is the size of the observed sample which we denote $\boldsymbol{y} = (y_1, \ldots, y_n)$. $\pi(\theta|\boldsymbol{y})$ represents the posterior density (notice that we use π for both the prior and the posterior density but the condition on $\boldsymbol{y}$ makes it clear when we mean the posterior).

6.1 Interpretation of Probabilities

There is a difference in how probabilities are interpreted between the frequentist and Bayesian approach. For a frequentist, a probability is an objective physical measure about the likelihood of an event. For a Bayesian, a probability is a measure of how likely one believes in the occurrence of some outcome, i.e. it is a subjective measure of the likelihood of an event.

To dig into the implications of these interpretations imagine that someone tosses a coin and covers it before you can see what happened. What is the probability of the outcome being heads? A frequentist would say either zero or one, depending on the result that is ignored. For a frequentist the randomness was over when the coin landed and at this point it is either head or not. A Bayesian would say one half since his measure of how likely heads is has not changed even if the experiment has already happened.

Considering the difference between these two interpretations might seem a mere philosophical exercise with no practical use, but we consider that it is essential for the correct interpretation and assessment of models.

6.2 Point Estimation

Bayesian point estimation consists in minimizing a risk function. Let us assume that the parametric family we have chosen for our data is correct, but we do not know what the true value of the parameter is.

Definition 6.1. *A loss function is a map $\mathcal{L} : \Theta \times \Theta \to [0, \infty)$ with $\mathcal{L}(\theta, \theta) = 0$ for every $\theta \in \Theta$.*

The intuition behind a loss function is that if θ_0 is the true parameter, $\mathcal{L}(\theta_0, \theta_1)$ represents the "cost" of taking θ to be θ_1 instead of θ_0.

Definition 6.2. *The risk function associated with the loss function $\mathcal{L}$, is the map $\mathcal{R} : \Theta \to [0, \infty)$ with*

$$\mathcal{R}(\theta) = \int_\Theta \mathcal{L}(\theta, \theta_1)\pi(\theta_1|y)dm(\theta_1).$$

Thus, $\mathcal{R}(\theta)$ represents the expected loss with the respect to the posterior distribution if it is assumed that the true parameter is θ. Once a prior and a loss function have been fixed, the

Bayesian point estimator θ^* of θ_0 is defined as the element of Θ that minimizes the expected loss, i.e.

$$\theta^* = \operatorname*{argmin}_{\theta \in \Theta} \mathcal{R}(\theta).$$

6.3 Confidence Regions

A Bayesian confidence region of level p is any region of Θ that has probability p with respect to the posterior distribution. This concept is similar to frequentist's confidence regions. The main difference is that the latter ones depend on pivotal quantities (a statistic whose distribution does not depend on θ) or asymptotic distributions. In spite of this similarity, there are important differences in interpretation.

In frequentist confidence intervals the confidence level is a probability only before the sample is observed. Once the sample is observed and the region is delimited, it is not a probability anymore. Imagine for example that we want to build a confidence interval for the mean μ of a univariate normal population. A confidence interval of level α is given by $(T_1, T_2) = (\bar{X} - t_{n-1,\frac{\alpha}{2}}, \bar{X} + t_{n-1,\frac{\alpha}{2}})$. Before the sample is observed we state that $\mathrm{P}(T_1 \leq \mu \leq T_2) = 0.95$. Now, assume that after observing the sample T_1 and T_2 take the values 5 and 15 respectively. The usual interpretation here is: *the interval (5,15) was computed with a procedure that captures the real mean 95% of the time.* The word probability is not used in the interpretation of the observed interval. The reader might remember from early statistics classes or books the phrase "A 95% confidence interval does not mean that the true parameter is in there with probability 0.95. The parameter is either there or not". This sounds puzzling at first, but it becomes clear when one takes into consideration the frequentist interpretation of probability. The analogy with the coin example in Section 6.1 is very helpful here.

From a Bayesian perspective, confidence regions are always assigned a probability.

6.4 Test Quantities and Posterior Predictive p-values

In this section we describe a way to assess the fit of a Bayesian model. For this we need the *posterior predictive distribution (PPD)*, which is the distribution of a future observation given the model. Its density is given by

$$f(y|\boldsymbol{y}) = \int_{\Theta} f(y|\theta)\pi(\theta|\boldsymbol{y})d\nu(\theta),$$

where ν is a dominating measure of the distribution of θ. The model assessment consists in comparing the observed sample with a hypothetical sample coming from the PPD. The comparison is done through the use of *test quantities* and *replicated samples.*

Definition 6.3. *A* test quantity *T is a function of the sample and the parameter, i.e. $T : \mathcal{Y}^n \times \Theta \to \mathbb{R}$.*

Definition 6.4. *A* replicated sample *is a sample of the same size as the observed sample whose elements come from the PPD. We will use the notation $\boldsymbol{y}^{rep}$ to denote a replicated sample.*

Given a test quantity T, its Posterior Predictive p-value p_B is defined as the probability that a replicated sample is more extreme than the observed data, i.e.

$$p_B = \mathrm{P}(T(\boldsymbol{y}^{rep}, \theta) \geq T(\boldsymbol{y}, \theta)|\boldsymbol{y}), \tag{6.1}$$

where the probability is taken according to the joint distribution of $(\boldsymbol{y}^{rep}, \theta)$ given the observed sample $\boldsymbol{y}$. Thus, the marginal distribution of θ in (6.1) is the posterior distribution and the one of $\boldsymbol{y}^{rep}$ is the posterior predictive distribution. In most cases p_B is hard to compute; Monte Carlo methods are useful for this since (6.1) can be expressed as an expectation:

$$p_B = \mathbb{E}\big[I\left\{T(\boldsymbol{y}^{rep}, \theta) \geq T(\boldsymbol{y}, \theta)\right\}\big].$$

A value of p_B that is close to 0 or 1 means that $T(\boldsymbol{y}, \theta)$ is an extreme value with respect to the model. Thus, it is desirable for p_B to be away from 0 or 1. Notice that this is different from frequentist's p-values that need to be close to zero. Notice also that we have not talked about type I and type II errors. This is because we are not trying to accept or reject the model but rather trying to understand its limits of applicability.

Usually several test quantities are used for a given model. Each test quantity corresponds to a specific aspect of the data. If $T(\boldsymbol{y}, \theta)$ does not seem consistent with replicated samples, it gives evidence that the model does not fit well that particular aspect. For this purpose we will not only consider the value of p_B but also the distribution of $T(\boldsymbol{y}^{rep}, \theta) - T(\boldsymbol{y}, \theta)$ (for which we can obtain an histogram through simulations).

6.5 Diagnostics for Bayesian GLMs

We discuss here diagnostic methods for Bayesian GLMs. They are all used for model assessment in later chapters.

6.5.1 Residuals

In classical linear models residuals are defined as the difference between observed and predicted values. From now on we refer to these as *response* residuals.

In frequentist GLMs response residuals are not used much since there is no reference distribution to which they can be compared. Thus deviance residuals are often used because there is an asymptotic theoretical behaviour to which they can be compared.

In Bayesian GLMs response residuals are useful for model diagnostics since they can be compared with replicated residuals, i.e. residuals coming from replicated samples of the posterior predictive distribution. Thus, a diagnostic plot for the fit of a Bayesian GLM is a graph with the observed residuals against the predicted value with confidence bands based on quantiles of the replicated residuals.

It is also useful to plot the normalized residuals also known as Pearson residuals. They are defined as

$$\frac{y_i - \mu_i}{\sqrt{V(\mu_i)}}, \qquad i = 1, \ldots, m, \tag{6.2}$$

where y_i is the i-th observed response, μ_i is the i-th predicted value (the predicted mean of the i-th class) and V is the variance function of the response distribution.

It is also useful to compare the observed residuals plot with the plot of some sets of replicated residuals. In this comparison one wants to see if the observed residuals present some pattern that seems very different from the replicated ones.

It is also possible to use the sample mean and variance of the residuals as test quantities of the model.

6.5.2 Variable Selection

There exist Bayesian tools for variable selection and model choice such as the Bayesian factor and model projection (see Robert (2007, Chap. 7) for an exposition of methods).

In frequentist GLMs there is a standard test for variable significance: for each coefficient β_j, one tests the null hypothesis $\beta_j = 0$ using the asymptotic distribution of the mle estimation.

We propose here a test that resembles the frequentist test. For a coefficient β_j we find the largest probability p_{β_j} for which there is an interval I that does not include zero with $\mathrm{P}(\beta_j \in I) = p_{\beta_j}$, where the probability is taken with respect to the posterior probability of β_j. Notice that

$$p_{\beta_j} = \max(\mathrm{P}(\beta_j \geq 0), 1 - \mathrm{P}(\beta_j \geq 0))$$

and that significance of β_j is suggested when p_{β_j} is close to 1. Thus, in order to keep things close to the familiar frequentist test, we report $1 - p_{\beta_j}$ which suggests significance when its value is close to zero.

Definition 6.5. *For a GLM coefficient β_j, we define the coefficient significance as*

$$q_{\beta_j} := 1 - p_{\beta_j} = \min(\mathrm{P}(\beta_j \geq 0), 1 - \mathrm{P}(\beta_j \geq 0)).$$

The intuition behind the definition of q_{β_j} works when the distribution is unimodal. The interpretation is more complicated when the distribution is multimodal and there is a positive and a negative mode. We recommend graphing the density of β_j for the correct interpretation of q_{β_j}.

6.6 On Prior Selection

Ideally the prior distribution should reflect our knowledge about the parameters in the chosen model for the problem at hand. Often this is hard or impossible because the complexity of the chosen model does not allow to translate prior information in terms of the parameters of the model. In these cases it is customary to use the often called "non-informative priors". The idea is to choose a prior that reflects lack of information in order to "let the data speak for itself" (Gelman et al., 2014, Sec. 2.8). The problem is that defining a prior that reflects lack of information is much more complicated than it seems (see Irony and Singpurwalla (1997) for a high level discussion on the topic). In the following paragraphs we talk about why this is the case and then we comment on the priors we use in the examples of the following chapters.

Often when the term non-informative prior is used one refers to *flat priors*. There are two aspects about flat priors we consider relevant to discuss. First that they are actually informative and second that they are often *improper*, which means that the total probability mass is infinite instead of one (later we explain how this makes sense in a Bayesian context).

Suppose you have some parametric family of distributions whose members have density $f(y|\theta)$ where the parameter θ can take values on the interval $(1, 2)$. The flat prior for θ is the uniform distribution on $(1, 2)$. It is sometimes claimed that such prior reflects lack of information since it gives equal weight to all possible values of θ. Let us reparametrize the model with the one-to-one transformation $\phi = \frac{1}{\theta}$ and denote with $f(y|\phi)$ the densities of the new parametrization. By applying the change of variable formula, we see that the prior for ϕ that is equivalent to the flat prior for θ is $\pi(\phi) = \frac{1}{\phi^2}$ for ϕ in $(0.5, 1)$ (this is an absolute abuse of notation of the use of ϕ but we think what we mean is clear). The latter prior has a decreasing shape and it is therefore informative. This shows that flat priors only appear to be non-informative since this "non-informativity" depends on the chosen parametrization.

The use of improper priors is not uncommon. They can be used when, after applying Bayes formula, one gets a proper posterior. A typical example of improper prior is a flat prior on an infinite interval which in some cases yield a proper posterior. This seems unnatural and one might argue that it would make more sense to restrict the domain of the improper prior to a big compact set. In this way we work with a proper prior that is interpreatable. This is a sensible point. Now, it turns out that a proper posterior coming from an improper prior is the limit of posteriors coming from proper priors (see Theorem 1 in Bernardo (2005b)). This implies that it is possible to obtain sensible posteriors (see question 8 in Irony and Singpurwalla (1997)) from improper priors.

There is an area of research in Bayesian statistics that looks into finding priors that minimize as much as possible the influence of the prior on the posterior regardless of the parametrization. The first successful attempt for doing this was the Jefrreys prior (see Jeffreys (1946)). An important problem with Jefrreys prior is that it is only defined for one-parameter distributions with no clear way about how to generalize it. A successful generalization to several parameters is the reference prior introduced in Bernardo (1979). We had originally thought to use reference priors for our examples, but the mathematical complexity required to derive them for the coefficients of a GLM made us take a different approach.

In our applied examples we use *weakly* informative priors (see Section 2.9 in Gelman et al. (2014)). Similarly to flat priors, the idea is to use a diffuse prior but rather than doing it over all possible values of the parameters, we use a proper prior concentrated around *natural boundaries* coming from the problem at hand. This approach can be criticized in a similar way than flat priors: they are only weakly informative in appearance since, if the parametrization changes, they turn out to be quite informative. The answer to this criticism is that some parametrizations are more useful than others. In some of them, the parameters can be interpreted and this allows to mathematically express our knowledge or little information we have about it. This argument can also be used in favor of flat priors. So, why bother thinking about how to formulate the *natural boundaries* for the parameter? When such boundaries exist, a flat prior places an infinite weight outside those boundaries. The corresponding prior may then place a significant probability on extreme parameter values and the posterior might end up being too diffuse. A very instructive case study published in the stan documentation is Betancourt (2017).

In the next chapter we use GLMs for modelling the frequency and severity of a car insurance dataset. We use weakly informative priors fro for the coefficients. When the parameters can take any real value we use a normal distribution with a mean and variance consistent with what we consider common sense barriers for the model. If a parameter takes positive values we use a normal distribution truncated on $(0, \infty)$.

Most of our examples use a log link function and thus yield a multiplicative model. From anecdotal evidence we know that the yearly frequency of car insurance is typically smaller than 0.3. Thus, for the intercept of the model, β_0, we use a mean of -1.5 (since $\exp(-1.5) \approx 0.22$) and a standard deviation of one. For the other parameters we use a standard normal distribution. For the severity model we choose the mean of the intercept based on the observed response values (we discuss the details in Chapter 7) and for the rest of the coefficients we use mean 0 and standard deviation 2. This gives enough variability to the multiplicative factors that result from this values. For the dispersion parameter we use a normal distribution with mean 10 and standard deviation 20. truncated on $(0, \infty)$.

When arguing in favor of using weakly informative priors we said that some parametrizations are more useful than others. Now that we have presented the priors we have chosen for our examples, a natural question is what makes the chosen parametrization of the GLM better than others? We actually think that the best would be to be able to formulate a weakly informative prior for $\boldsymbol{\mu}$, the vector of means. We find it to be the most natural parameter, specially since it is the one we are most interested in estimating. Unfortunately there is a technical difficulty for this. Even though $\boldsymbol{\mu} \in \mathbb{R}^m$, it cannot take any value in $\mathbb{R}^m$ since $\boldsymbol{\mu} = G(^{-1}(X\boldsymbol{\beta}))$ (see (4.13)),where $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$. Thus a prior for $\boldsymbol{\mu}$ should have support on the set $G(^{-1}(X\boldsymbol{\beta}))$ and it is not easy to formulate a prior in such space.

Chapter 7

Exact Linear Credibility for GLMs is Impossible

Most credibility research has been centered around linear estimators and GLMs are not excluded from this. For instance, in Nelder and Verrall (1997), Ohlsson (2008) and Antonio and Beirlant (2007), they obtain linear credibility estimators by adding random effects to the GLM.

Before starting to work on the entropic credibility estimator of next chapter, we were interested in knowing whether random effects were the only way of obtaining linear estimators for the mean. More precisely, we wondered whether Jewell's result could be extended to GLMs. This chapter addresses the question: is there a prior for the regression coefficients $\boldsymbol{\beta}$ for which the posterior mean is a weighted mean between an out–of–sample estimate and the sample mean of a GLM? We found that the answer is no (the results of this chapter have appeared in Quijano Xacur and Garrido (2018)).

There are two ways in which the question from the previous paragraph can be interpreted. One could think of it as all m dimensions having the same credibility factor, that is, the credibility premium $\hat{\boldsymbol{\mu}}_c$ is given by

$$\hat{\boldsymbol{\mu}}_c = z\bar{\boldsymbol{y}} + (1-z)\boldsymbol{M}, \tag{7.1}$$

where $\bar{\boldsymbol{y}}$ is the GLM observed sample mean (i.e. a vector for which (4.8) applies to each coordinate), $\boldsymbol{M}$ is a vector of out–of–sample "manual" premiums as coordinates and $z \in (0,1)$

is the credibility factor. We call this interpretation *Linear Credibility of Type 1.*

The other interpretation is to give a different credibility factor to each coordinate, in a similar way to Hachemesiter's model (see (2.7)). This is

$$\hat{\boldsymbol{\mu}}_c = Z\bar{\boldsymbol{y}} + (I - Z)\boldsymbol{M}, \tag{7.2}$$

where $\hat{\boldsymbol{\mu}}_c$, $\bar{\boldsymbol{y}}$ and $\boldsymbol{M}$ are as in (7.1), but $Z = \operatorname{diag}(z_1, \dots, z_m)$, where z_i is the credibility factor of the i-th class and I is the identity matrix. We call this interpretation *Linear Credibility of Type 2.* Note that linear credibility of Type 1 is a special case of linear credibility of Type 2.

7.1 Linear Credibility of Type 1 is Impossible

Jewell's prior in (2.4) is a conjugate prior to (2.3) that gives linear credibility premiums based on the posterior mean. Diaconis and Ylvisaker (1979) generalized Jewell's result to multivariate exponential dispersion models. Since a GLM assumes that the response vector follows such a distribution, it could be conjectured that this automatically implies linear credibility for GLMs. In what follows we show that this is not the case.

After adapting the conjugate prior discussed in Diaconis and Ylvisaker (1979) to correspond to (4.12) so as to consider weights we get

$$\pi_{n_0,\boldsymbol{x_0}}(\boldsymbol{\theta}) \propto \exp\left(n_0\{\boldsymbol{x_0}^T W\boldsymbol{\theta} - \mathbf{1}^T W\boldsymbol{k}(\boldsymbol{\theta})\}\right)\mathbb{I}_{\Theta^m}(\boldsymbol{\theta}), \tag{7.3}$$

where $\Theta^m = \left\{(\theta_1 \dots \theta_m)^T : \theta_1, \dots, \theta_m \in \Theta\right\}$; $n_0 > 0$ and $\boldsymbol{x_0} \in \Omega^m$ are the parameters of the prior distribution, $\mathbb{I}_{\Theta^m}$ is an indicator function and $\Omega^m = \left\{(\mu_1 \dots \mu_m)^T : \mu_1, \dots, \mu_m \in \Omega\right\}$. Theorem 3 of Diaconis and Ylvisaker (1979) proves that if the support of (4.12) contains an interval then (7.3) is the only prior that gives linear credibility. This implies that for any continuous response (and also for any Tweedie distribution), (7.3) is the only prior that gives linear credibility. In the paper it is also proven that (7.3) is the unique prior that gives linear credibility for the binomial distribution and in Johnson (1957) the same is proven for the Poisson distribution.

As shown in (4.13), $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are related by $\boldsymbol{\theta} = ((G \circ \dot{\boldsymbol{\kappa}})^{-1} \circ X)(\boldsymbol{\beta})$. Thus, when a prior for $\boldsymbol{\beta}$ is chosen, a distribution is induced on $\boldsymbol{\theta}$. In what follows we refer to this distribution as the *induced prior* on $\boldsymbol{\theta}$. We have then that for continuous and Tweedie distributions and for the Poisson and negative binomial, a prior on $\boldsymbol{\beta}$ gives linear credibility if and only if the induced prior on $\boldsymbol{\theta}$ is (7.3).

Our strategy to prove the impossibility of linear credibility of type 1 is to show that no prior of $\boldsymbol{\beta}$ induces a prior on $\boldsymbol{\theta}$ that has density (7.3). We can see that this is the case by focusing on the support of the induced distribution. Since the support of (7.3) is Θ^m, then it is enough to prove that the support of every induced prior of $\boldsymbol{\theta}$ is different than Θ^m.

Proposition 7.1.1. *For any prior of $\boldsymbol{\beta}$ in a non-saturated GLM, the support of the induced prior of $\boldsymbol{\theta}$ is a proper subset of Θ^m.*

Proof. Since there is no restriction for the value of $\boldsymbol{\beta}$, it can take any value on $\mathbb{R}^{p+1}$. This is represented on the left rectangle of Figure 7.1.

$X\boldsymbol{\beta}$ can take values in $R(X)$, where $R(X)$ is the range of X. Since $\dim(R(X)) = p+1 < m$, then $R(X) \subsetneq \mathbb{R}^m$. This is represented in the middle rectangle of Figure 7.1.

Let S be the support of the induced prior on $\boldsymbol{\theta}$. Then $S \subset (G \circ \dot{\boldsymbol{\kappa}})^{-1}(R(X)) := \{(G \circ \dot{\boldsymbol{\kappa}})^{-1}(X\boldsymbol{\beta}) : \boldsymbol{\beta} \in \mathbb{R}^{p+1}\}$ (subset but not equality since some values of $\mathbb{R}^{p+1}$ may not be in the support of $\boldsymbol{\beta}$). Now, $(G \circ \dot{\boldsymbol{\kappa}})^{-1}$ is a bijective function. Let $\boldsymbol{p}$ be a point in $\mathbb{R}^m$ that is not in $R(X)$ and let $\boldsymbol{q} = (G \circ \dot{\boldsymbol{\kappa}})^{-1}(\boldsymbol{p})$. Then $\boldsymbol{q} \in \Theta^m$ but $\boldsymbol{q} \in S$ because otherwise $(G \circ \dot{\boldsymbol{\kappa}})^{-1}$ would not be one-to-one. This proves that $(G \circ \dot{\boldsymbol{\kappa}})^{-1}(R(X)) \subsetneq \Theta^m$ and therefore also that $S \subsetneq \Theta^m$. This is represented in the right rectangle of Figure 7.1.

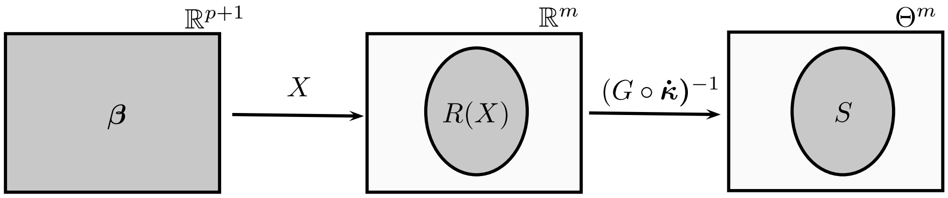


Figure 7.1: From left to right the grey zone represents the values that $\boldsymbol{\beta}$, $R(X)$ and S can take, respectively.

□

Now, on a different but related note, it is possible to generalize (7.3) in a way that allows to obtain conjugate priors that are suitable for GLMs. Define

$$\pi_1(\boldsymbol{\theta}) \propto h(\boldsymbol{\theta}) \exp\left(n_0\{\boldsymbol{x_0}^T W\boldsymbol{\theta} - \mathbf{1}^T W\boldsymbol{k}(\boldsymbol{\theta})\}\right)\mathbb{I}_{\Theta^m}(\boldsymbol{\theta}),$$

where h is some integrable function for which the integral on the right hand–side above is finite and denote this distribution by $D_{conj}(n_0, \boldsymbol{x_0})$.

Proposition 7.1.2. *π_1 is a conjugate prior to* (4.12) *with posterior distribution*

$$D_{conj}\left(n_0 + \frac{1}{\phi}, \frac{1}{n_0\phi + 1}\bar{\boldsymbol{y}} + \frac{\phi n_0}{\phi n_0 + 1}\boldsymbol{x_0}\right).$$

Proof. Let $\pi_1(\cdot|\boldsymbol{y})$ denote the posterior of π_1. Then, by definition of the posterior:

$$\begin{aligned}
\pi_1(\cdot|\boldsymbol{y}) &\propto \pi_1(\boldsymbol{\theta})f(\boldsymbol{y}|\boldsymbol{\theta},\phi) \\
&\propto h(\boldsymbol{\theta})\exp\left(n_0\boldsymbol{x_0}^T W\boldsymbol{\theta} + \frac{\boldsymbol{y}^T W\boldsymbol{\theta}}{\phi} - n_0\mathbf{1}^T W\boldsymbol{k}(\boldsymbol{\theta}) - \frac{\mathbf{1}^T W\boldsymbol{\kappa}(\boldsymbol{\theta})}{\phi}\right) \\
&= h(\boldsymbol{\theta})\exp\left((n_0\boldsymbol{x_0}^T + \frac{\boldsymbol{y}^T}{\phi})W\boldsymbol{\theta} - \left(n_0 + \frac{1}{\phi}\right)W\boldsymbol{\kappa}(\boldsymbol{\theta})\right) \\
&= h(\boldsymbol{\theta})\exp\left(\left(n_0 + \frac{1}{\phi}\right)\left\{\frac{n_0\boldsymbol{x_0}^T + \frac{\boldsymbol{y}^T}{\phi}}{n_0 + \frac{1}{\phi}}W\boldsymbol{\theta} - \mathbf{1}^T W\boldsymbol{\kappa}(\boldsymbol{\theta})\right\}\right) \\
&= h(\boldsymbol{\theta})\exp\left(\left(n_0 + \frac{1}{\phi}\right)\left\{\frac{\phi n_0\boldsymbol{x_0}^T + \boldsymbol{y}^T}{\phi n_0 + 1}W\boldsymbol{\theta} - \mathbf{1}^T W\boldsymbol{\kappa}(\boldsymbol{\theta})\right\}\right),
\end{aligned}$$

which proves the result. □

Now, in order for π_1 to overcome the problems that do not allow π in (7.3) to be used as a prior for GLMs, it is only necessary to chose h such that π_1 is outside of $(G \circ \dot{\boldsymbol{\kappa}})^{-1}(R(X))$ with probability zero. This way π_1 has the "right" support and there is a distribution of $\boldsymbol{\beta}$ that gives this distribution when transformed with $((G \circ \dot{\boldsymbol{\kappa}})^{-1} \circ X)$.

Two important remarks about π_1:

1. It does not give linear credibility (since this is impossible as has been shown above).

2. It is not easy to find an analytic expression for $\boldsymbol{\mu}$ (although this might be possible for some choices of π). Thus most likely one has to use some numerical method or MCMC in order to find the posterior means, but this defeats the purpose of using a conjugate prior.

7.2 Linear Credibility of Type 2 is Sometimes Feasible

Since the model is a GLM, there should be a $\hat{\boldsymbol{\beta}}_c$ such that $\hat{\boldsymbol{\mu}}_c = G^{-1}(X\hat{\boldsymbol{\beta}}_c)$. Thus, (7.2) becomes

$$G^{-1}(X\hat{\boldsymbol{\beta}}_c) = Z\bar{\boldsymbol{y}} + (I - Z)\boldsymbol{M}. \tag{7.4}$$

It turns out that for non saturated models (i.e. $\dim(\boldsymbol{\beta}) < \dim(\boldsymbol{\mu})$), the existence of some $\hat{\boldsymbol{\beta}}_c$ for which (7.4) can be satisfied depends on the observed sample. We demonstrate why this is the case with a simple example in dimension 2.

Consider a situation in which you divide your population in only 2 segments using a binary covariate with no intercept (otherwise we would have a saturated model). The design matrix in this case would be

$$X = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \hat{\beta}_c \in \mathbb{R}.$$

Then, assuming a log–link function, the left hand side of (7.4) can be expressed as

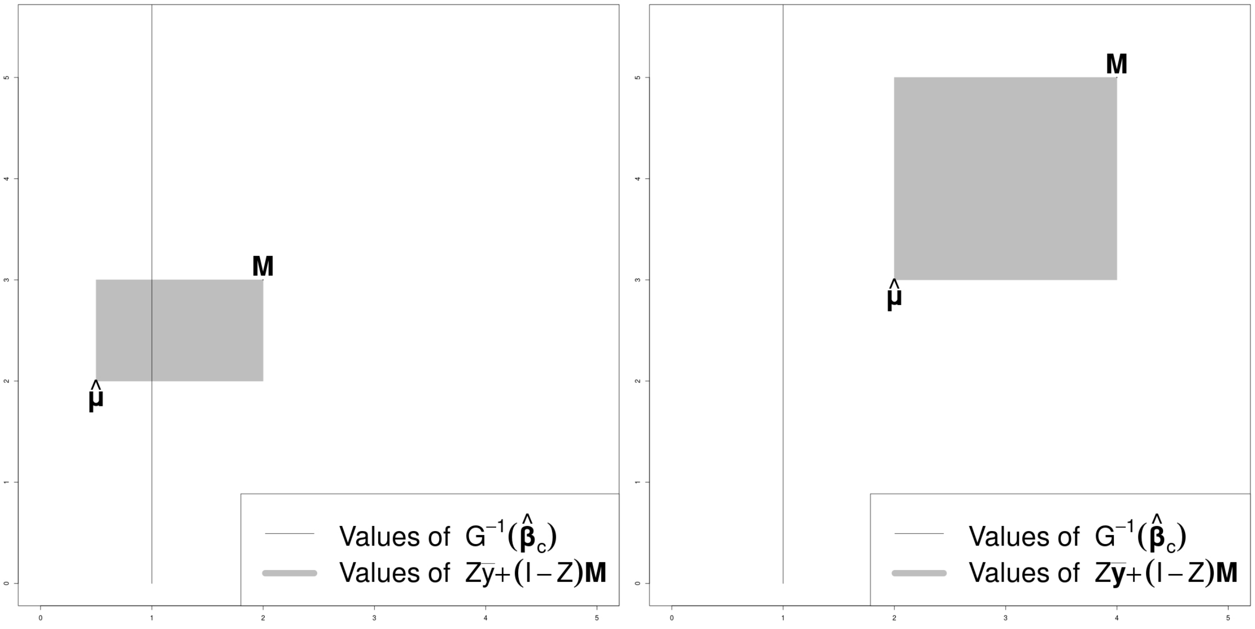
$$\hat{\boldsymbol{\mu}}_c = G^{-1}(X\hat{\beta}_c) = G^{-1}\begin{pmatrix} 0 \\ \hat{\beta}_c \end{pmatrix} = \begin{pmatrix} \exp(0) \\ \exp(\hat{\beta}_c) \end{pmatrix} = \begin{pmatrix} 1 \\ \exp(\hat{\beta}_c) \end{pmatrix}.$$

If we graphed it we would see that the left hand side of (7.4) takes values only on the half upper side of the vertical line $x = 1$.

Imagine now two scenarios. In Scenario 1, $\bar{\boldsymbol{y}} = (0.5, 2)$ and $\boldsymbol{M} = (2, 3)$, while in Scenario 2, $\bar{\boldsymbol{y}} = (2, 3)$ and $\boldsymbol{M} = (4, 5)$.

As the values of the elements of Z vary, the right hand side of (7.2) can take the values of the rectangle defined by $\bar{\boldsymbol{y}}$ and $\boldsymbol{M}$. Figure 7.2 shows graphs with the possible values of the left and right hand side of (7.2) for each scenario.

In both graphs, the vertical line represents the values of $\hat{\boldsymbol{\mu}}_c$. The rectangle represents all the possible values that $Z\bar{\boldsymbol{y}} + (I - Z)\boldsymbol{M}$ can take as the entries in the diagonal of Z vary from 0 to 1. In order to have exact linear credibility of type 2, it is necessary for the line and the rectangle to intersect. This is because the points of intersection, correspond to combinations of values of $\boldsymbol{\beta}_c$ and Z for which (7.4) holds. If there is no intersection it is not possible to have linear credibility of type 2.



(a) Scenario 1

(b) Scenario 2

Figure 7.2: Values of the left and right hand side of (7.4) in both scenarios

The graph of Scenario 1 shows that (7.2) is satisfied for some values of Z, while in the graph for Scenario 2 it is impossible to satisfy (7.2).

The results of this section show that Jewell's result cannot be generalized to GLM's. That is, no prior for the parameters of a GLM guarantee linear credibility for all observed samples.

Chapter 8

Entropic Credibility

As a Bayesian model, linear credibility (described in Section 2.1) is rather artificial. Note, for instance, that the only aspect of the posterior distribution that we use is its mean and we disregard all the other information provided by it. Furthermore the adequacy of Jewell's prior in any given situation is usually not discussed. The main focus has been the fact that Jewell's prior yields a linear credibility premium that we can easily compute. This convenience was crucial when Bühlman and Jewell originally published their work since computing power was scarce and expensive. Nowadays not only computing is cheap, but also sophisticated simulation software is available for anyone on the internet.

We propose in this chapter a Bayesian point estimator for credibility. We advocate the use of a prior that expresses the out-of-sample information rather than one that gives a credibility formula (some of the results of this chapter have appeared in Quijano Xacur and Garrido (2018)).

Note that with this method, there is no need of a full credibility criterion. In fact the terms full credibility and partial credibility (see Chapter 2) loose their relevance.

For these ideas to be applicable in practice one relies on MCMC. In consequence our method has the following two limitations:

Convergence of MCMC: There are cases in which even after a long warmup period, the simulated Markov chains do not show convergence. The simulations obtained in these cases are not reliable and should not be used for estimation.

Computational power and time: The computations required to perform MCMC in higher dimensions are very demanding and require substantial computing resources. Thus powerful computers are necessary and even then sometimes one might need to wait hours or days before the simulations are completed.

For GLMs that use only categorical covariates, aggregating the data (as described in Section 4.2) often reduces significantly the dimension of the simulations. This reduces the time for computations considerably and helps deal with the second limitation above.

8.1 The Entropic Estimator

A posterior distribution is more informative than a point estimation since it reflects our uncertainty about the true parameter. Now, in insurance, it is necessary to charge a premium, which is a point estimate. In this section we define the point estimators that we propose as credibility premiums.

Consider a parametric family of distributions with a parameter $\boldsymbol{\theta}$ to be estimated. Assume that $\boldsymbol{\theta}_0$ is the "true" parameter. As commented in section 6.2 in Bayesian point estimation one first chooses a loss function $\mathcal{L}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ that represents the cost of estimating $\boldsymbol{\theta}$ to be $\boldsymbol{\theta}_1$ instead of $\boldsymbol{\theta}_0$. Now, since $\boldsymbol{\theta}_0$ is not known, we define a risk function as

$$\mathcal{R}(\theta) = \mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_0, \boldsymbol{\theta})\right],$$

where the expectation is taken with respect to the posterior distribution of $\boldsymbol{\theta}$. Then the point estimator $\hat{\boldsymbol{\theta}}$ of θ_0 is the value of θ that minimizes $\mathcal{R}$. That is:

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmin}}\, \mathcal{R}(\boldsymbol{\theta}).$$

The entropic estimator is defined as the Bayesian point estimator when the loss function $\mathcal{L}$ is the relative entropy of the distribution with the real parameter $\boldsymbol{\theta_0}$ over the estimated one.

More precisely, assume that the density of a random vector $\boldsymbol{Y}$ depends on a parameter $\boldsymbol{\theta}$. Denote with $f(\boldsymbol{Y}|\boldsymbol{\theta}_1)$ the density of $\boldsymbol{Y}$ when $\boldsymbol{\theta}$ takes some value $\boldsymbol{\theta}_1$. The loss function is defined as

$$\mathcal{L}(\boldsymbol{\theta}_0, \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}_0}\left[\log\left(\frac{f(\boldsymbol{Y}|\boldsymbol{\theta}_0)}{f(\boldsymbol{Y}|\boldsymbol{\theta})}\right)\right].$$

The corresponding risk function is then defined as

$$\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}\left[\log\left(\frac{f(\boldsymbol{Y}|\boldsymbol{\theta}_0)}{f(\boldsymbol{Y}|\boldsymbol{\theta})}\right)\right] := \mathbb{E}_{\pi}\left[\mathbb{E}_{\boldsymbol{\theta}_0}\left[\log\left(\frac{f(\boldsymbol{Y}|\boldsymbol{\theta}_0)}{f(\boldsymbol{Y}|\boldsymbol{\theta})}\right)\right]\right],$$

where $\mathbb{E}_{\pi}$ is the expectation taken with respect to the posterior distribution of $\boldsymbol{\theta}$.

Definition 8.1. *The entropic estimator is defined as*

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathbb{E}\left[\log\left(\frac{f(\boldsymbol{Y}|\boldsymbol{\theta}_0)}{f(\boldsymbol{Y}|\boldsymbol{\theta})}\right)\right],$$

where $\mathbb{E}[\cdot] = \mathbb{E}_{\pi}[\mathbb{E}_{\boldsymbol{\theta}_0}[\cdot]]$.

8.1.1 Invariant Estimators

In parametric point estimation we start with some observed data and a set of candidate probability distributions characterized by some parameter θ. When we use the word *estimator* we mean the value of θ we choose after applying some optimality criterion to the data and the set of candidate distributions.

For example the optimality criterion in maximum likelihood estimators is maximizing the likelihood function while the one for Bayesian point estimation it is minimizing the risk function.

We are interested in finding optimality criteria that gives estimators with good properties. *Invariance* is one of them (this terminology is consistent with Bernardo (2005a)).

Definition 8.2. *Fix some optimality criterion. Suppose $C(\theta)$ is a candidate set of distributions parametrized by θ and $C(\alpha)$ is the same candidate set but parametrized by $\alpha = g(\theta)$, where g is some bijective map. For some observed data, let θ^* be the estimator for the θ-parametrization and α^* be the estimator for the α-parametrization. The estimator is said to be* invariant *if the distribution corresponding to θ^* in $C(\theta)$ is the same than the distribution corresponding to α^* in $C(\alpha)$ for every map g.*

When we choose a specific value of the parameter θ we are in the end picking an element of the candidate set of distributions to describe our observations. Thus, invariant estimators always choose the same distribution regardless of the chosen parametrization of the candidate set. In contrast non-invariant estimators pick different distributions for different parametrizations.

Even though it is hard to define what it means for a distribution from a candidate set to be the *best* for some observed data, we can safely say that it is not related to a specific parametrization. This is because *best* goes more on the lines of the distribution being able to replicate some aspect or aspects of the observed data and the chosen parametrization is irrelevant for this. Thus, non-invariant measures are by design incompatible with what we would expect from "best fitting the data". This makes invariance if not a minimum requirement at least a very desirable property.

The entropic estimator is invariant. This is a direct consequence of the invariant property of the relative entropy (see Section 5.1). The Maximum Likelihood estimator is also known to be invariant.

The Bayesian point estimator with a square loss, i.e. the posterior mean, is known to not be invariant but it is widely used. This is because it is one of the few loss functions for which, at least theoretically, we know what value minimizes the risk function and we have methods to find it. Moving away from this now familiar ground makes it hard and often impossible for finding a point estimator. For instance, later we will be computing entropic estimators for GLMs, but it took a PhD thesis to find an algorithm to compute it!

8.2 Entropic Estimators for univariate EDFs

In this section, we focus on entropic estimators for univariate EDFs and their relation to linear credibility. The main result of the section is Theorem 8.1, which is preceded by two technical lemmas that show properties of the unit deviance that are fundamental for finding entropic estimators of exponential families and GLMs.

Lemma 8.1. *Let d be the unit deviance of a univariate EDF in* (4.5). *Then, there exist* *functions d_1 and d_2 such that for $(y, \mu) \in \Omega \times \Omega$, we have the following decomposition:*

$$d(y, \mu) = d_1(y) + d_2(y, \mu). \tag{8.1}$$

Moreover, d_2 has the property that if Y is a random variable with support in Ω, and μ is fixed, then

$$\mathbb{E}\big[d_2(Y, \mu)\big] = d_2\big(\mathbb{E}[Y], \mu\big).$$

Proof. Let d be the unit deviance function of the response distribution and $y, \mu \in \Omega$. By (4.5), we have that

$$\begin{aligned} d(y, \mu) &= 2\left[y\left\{\dot{\kappa}^{-1}(y) - \dot{\kappa}^{-1}(\mu)\right\} - \kappa(\dot{\kappa}^{-1}(y)) + \kappa(\dot{\kappa}^{-1}(\mu))\right] \\ &= 2\left[y\dot{\kappa}^{-1}(y) - \kappa(\dot{\kappa}^{-1}(y))\right] + 2\left[\kappa(\dot{\kappa}^{-1}(\mu)) - y\dot{\kappa}^{-1}(\mu)\right] \\ &= d_1(y) + d_2(y, \mu), \end{aligned}$$

where $d_1(y) = 2\left[y\dot{\kappa}^{-1}(y) - \kappa(\dot{\kappa}^{-1}(y))\right]$ and $d_2(y, \mu) = 2\left[\kappa(\dot{\kappa}^{-1}(\mu)) - y\dot{\kappa}^{-1}(\mu)\right]$. Now, let Y be a random variable with support in Ω and $\mu \in \Omega$ be fixed, then

$$\begin{aligned} \mathbb{E}[d_2(Y, \mu)] &= \mathbb{E}\left[2\left[\kappa(\dot{\kappa}^{-1}(\mu)) - Y\dot{\kappa}^{-1}(\mu)\right]\right] \\ &= 2\left[\kappa(\dot{\kappa}^{-1}(\mu)) - \mathbb{E}[Y]\dot{\kappa}^{-1}(\mu)\right] \\ &= d_2(\mathbb{E}[Y], \mu). \end{aligned}$$

□

Lemma 8.2. *Let y be fixed, then*

1. *the value of μ that minimizes $d_2(y, \mu)$ is the same one that minimizes $d(y, \mu)$.*

2. *$d_2(y, \mu)$ is minimized when $\mu = y$.*

Proof. The first part is a direct consequence of (8.1). Since y is fixed, minimizing the right hand side is equivalent to minimizing d_2 and therefore the claim is true.

The unit deviance d is such that (see Chapter 1 of Jørgensen (1997)) $d(y, \mu) > 0$ if $y \neq \mu$ and $d(y, \mu) = 0$ when $y = \mu$. Thus $d(y, \mu)$ is minimized when $y = \mu$ and by Part 1 above the same applies to $d_2(y, \mu)$.

□

Distribution	$d(y,\mu)$	$d_1(y)$	$d_2(y,\mu)$
Normal	$(y-\mu)^2$	y^2	$\mu^2 - 2y\mu$
Poisson	$2\left\{y\log\left(\frac{y}{\mu}\right) - (y-\mu)\right\}$	$2y[\log(y) - 1]$	$2\left[\mu - y\log(\mu)\right]$
Gamma	$2\left\{\log\left(\frac{\mu}{y}\right) + \frac{y}{\mu} - 1\right\}$	$2\left[\frac{y}{\mu} + \log(\mu)\right]$	$2\left[\frac{y}{\mu} + \log\left(\frac{\mu}{y}\right) - 1\right]$

Table 8.1: Deviance decomposition of some common EDF's

Table 8.1 shows d, d_1 and d_2 for the normal, Poisson and gamma distributions.

Theorem 8.1. *Let Y be a random variable whose density is given by* (4.4) *for some unknown values of μ and ϕ, $\pi(\mu,\phi)$ be a prior distribution for (μ,ϕ), the vector $\boldsymbol{y} = (y_1 \dots y_n)^T$ be a conditionally i.i.d. sample given (μ,ϕ), and $\pi(\mu,\phi|\boldsymbol{y})$ the corresponding posterior. The entropic estimator $\hat{\mu}$ of μ is then given by*

$$\hat{\mu} = \mathbb{E}[Y|\boldsymbol{y}] = \mathbb{E}_\pi\left[\mathbb{E}_{\mu,\phi}[Y]\right],$$

where $\mathbb{E}_\pi$ represents the expectation with respect to the posterior distribution and $\mathbb{E}_{\mu,\phi}$ represent the expectations with respect to fixed values of (μ,ϕ).

Proof. Let (μ_0,ϕ_0) be the true parameters. By Lemma 8.1, the entropic risk measure can be expressed as

$$\begin{aligned}
\mathcal{R}(\mu,\phi) &= \mathbb{E}\left[\log\left(\frac{f(Y|\mu_0,\phi_0)}{f(Y|\mu,\phi)}\right)\right] \\
&= \mathbb{E}\left[\log\left(\frac{c(Y,\phi_0)\exp\left(-\frac{1}{2\phi_0}d(Y,\mu_0)\right)}{c(Y,\phi)\exp\left(-\frac{1}{2\phi}d(Y,\mu)\right)}\right)\right] \\
&= \mathbb{E}\left[\log\left(c(Y,\phi_0)\exp\left(-\frac{1}{2\phi_0}d(Y,\mu_0)\right)\right)\right] \\
&\quad - \mathbb{E}[\log(c(Y,\phi))] + \frac{1}{2\phi}\mathbb{E}[d(Y,\mu)] \\
&= \mathbb{E}\left[\log\left(c(Y,\phi_0)\exp\left(-\frac{1}{2\phi_0}d(Y,\mu_0)\right)\right)\right] \\
&\quad - \mathbb{E}[\log(c(Y,\phi))] + \frac{1}{2\phi}\mathbb{E}[d_1(Y)] + \frac{1}{2\phi}d_2(\mathbb{E}[Y],\mu).
\end{aligned}$$

Note that, regardless of the value of ϕ, the value of μ that minimizes the expression above is the same one that minimizes the simpler function

$$\mathcal{R}_1(\mu) = d_2\big(\mathbb{E}[Y],\mu\big).$$

Then, by Part 2 of Lemma 8.2, the entropic estimator of μ is given by $\hat{\mu} = \mathbb{E}[Y]$. □

This result shows that for univariate EDFs the posterior mean not only minimizes the expected square error risk, but also the posterior entropic risk. The results from next section show that this property does not generalize to non-saturated GLMs. Realize that a direct consequence of Theorem 8.1 is that Jewell's estimator in Jewell (1974) is an entropic estimator.

Corollary 8.1. *The linear credibility estimator* (2.5) *is the entropic estimator when* ϕ *is assumed known and* (2.4) *is used as prior for* θ*.*

8.3 Entropic Credibility for GLMs

Let us now focus on entropic credibility for GLMs. We start by enunciating the following technical lemma, which is an extension of Lemma 8.1 to higher dimensions.

Lemma 8.3. *Let* D *be the deviance of a GLM (see* (4.15)*). Then there exist functions* D_1 *and* D_2 *such that for* $(\boldsymbol{y}, \boldsymbol{\mu}) \in \Omega^m \times \Omega^m$

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = D_1(\boldsymbol{y}) + D_2(\boldsymbol{y}, \boldsymbol{\mu}). \tag{8.2}$$

Moreover, D_2 *has the property that if* $\boldsymbol{Y}$ *is a random vector with support in* Ω^m *and* $\boldsymbol{\mu} \in \Omega^m$ *is fixed, then*

$$\mathbb{E}[D_2(\boldsymbol{Y}, \boldsymbol{\mu})] = D_2(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu}). \tag{8.3}$$

Proof. From the definition of D and Lemma 8.1, we have that

$$\begin{aligned} D(\boldsymbol{y}, \boldsymbol{\mu}) &= \sum_{i=1}^{m} w_i d(y_i, \mu_i) \\ &= \sum_{i=1}^{m} w_i (d_1(y_i) + d_2(y_i, \mu_i)) \\ &= \sum_{i=1}^{m} w_i d_1(y_i) + \sum_{i=1}^{m} w_i d_2(y_i, \mu_i) \\ &= D_1(\boldsymbol{y}) + D_2(\boldsymbol{y}, \boldsymbol{\mu}), \end{aligned}$$

where $D_1(\boldsymbol{y}) = \sum_{i=1}^{m} w_i d_1(y_i)$ and $D_2(\boldsymbol{y}, \boldsymbol{\mu}) = \sum_{i=1}^{m} w_i d_2(y_i, \mu_i)$. This proves (8.2). Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)$ be a random vector with support on Ω^m, and $\boldsymbol{\mu} \in \Omega^m$ be fixed. Then

$$\mathbb{E}[D_2(\boldsymbol{Y}, \boldsymbol{\mu})] = \mathbb{E}[\sum_{i=1}^{m} w_i d_2(Y_i, \mu_i)] = \sum_{i=1}^{m} w_i d_2(\mathbb{E}[Y_i], \mu_i) = D_2(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu}),$$

which proves (8.3). □

In what follows, an arbitrary prior π is assumed (not necessarily conjugate) with posterior

$$\pi(\boldsymbol{\beta}, \phi | \boldsymbol{y}) \propto f(\boldsymbol{y} | \boldsymbol{\beta}, \phi) \pi(\boldsymbol{\beta}, \phi), \tag{8.4}$$

where f is as in (4.12) or, equivalently (4.14), depending on the chosen parameterization. As stated, $\mathbb{E}_\pi[\cdot]$ denotes expectation with respect to the posterior measure. Whenever the expectation symbol is used without a subindex, it means expectation with respect to the predictive posterior distribution, i.e. $\mathbb{E}[\cdot] = \mathbb{E}_\pi[\mathbb{E}_{\boldsymbol{\beta},\phi}(\cdot)]$, where $\mathbb{E}_{\boldsymbol{\beta},\phi}(\cdot)$ means expectation with respect to the density in (4.12) with fixed coefficients vector $\boldsymbol{\beta}$ and fixed dispersion parameter ϕ.

Theorem 8.2. *The entropic estimator $\boldsymbol{\beta}^*$ of the coefficients of a Bayesian GLM are equal to the maximum likelihood estimator of a frequentist GLM with the same covariates, response distribution and weights, but with an observed response vector equal to $\mathbb{E}[\boldsymbol{Y}]$.*

Proof. Let $(\boldsymbol{\beta}_0, \phi_0)$ represent the true parameters and $(\boldsymbol{\beta}, \phi)$ some fixed values. We use here for f the mean value parameterization in (4.14). Then, the risk function is given by

$$\mathcal{R}(\boldsymbol{\beta}, \phi) = \mathbb{E}_\pi\{\mathcal{L}[(\boldsymbol{\beta}_0, \phi_0), (\boldsymbol{\beta}, \phi)]\} = \mathbb{E}\left[\log\left(\frac{f(\boldsymbol{Y}|\boldsymbol{\mu}_0, \phi_0)}{f(\boldsymbol{Y}|\boldsymbol{\mu}, \phi)}\right)\right],$$

where $\boldsymbol{\mu} = G^{-1}(X\boldsymbol{\beta})$ and $\boldsymbol{\mu}_0 = G^{-1}(X\boldsymbol{\beta_0})$. Then, by Lemma 8.3 and (4.14) the expression above becomes

$$
\begin{aligned}
\mathcal{R}(\boldsymbol{\beta}, \phi) &= \mathbb{E}\left[\log(C(\boldsymbol{Y}, \phi_0)) - \frac{1}{2\phi_0} D(\boldsymbol{Y}, \boldsymbol{\mu}_0)\right] - \mathbb{E}[\log(C(\boldsymbol{Y}, \phi))] \\
&\quad + \frac{1}{2\phi}\mathbb{E}[D(\boldsymbol{Y}, \boldsymbol{\mu})] \\
&= \mathbb{E}\left[\log(C(\boldsymbol{Y}, \phi_0)) - \frac{1}{2\phi_0} D(\boldsymbol{Y}, \boldsymbol{\mu}_0)\right] \\
&\quad - \mathbb{E}[\log(C(\boldsymbol{Y}, \phi))] + \frac{1}{2\phi}\mathbb{E}[D_1(\boldsymbol{Y}) + D_2(\boldsymbol{Y}, \boldsymbol{\mu})] \\
&= \mathbb{E}\left[\log(C(\boldsymbol{Y}, \phi_0)) - \frac{1}{2\phi_0} D(\boldsymbol{Y}, \boldsymbol{\mu}_0) + \frac{1}{2\phi} D_1(\boldsymbol{Y})\right] \\
&\quad - \mathbb{E}[\log(C(\boldsymbol{Y}, \phi))] + \frac{1}{2\phi}\mathbb{E}[D_2(\boldsymbol{Y}, \boldsymbol{\mu})] \\
&= \mathbb{E}\left[\log(C(\boldsymbol{Y}, \phi_0)) - \frac{1}{2\phi_0} D(\boldsymbol{Y}, \boldsymbol{\mu}_0) + \frac{1}{2\phi} D_1(\boldsymbol{Y})\right] \\
&\quad - \mathbb{E}[\log(C(\boldsymbol{Y}, \phi))] + \frac{1}{2\phi} D_2(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu}). \qquad (8.5)
\end{aligned}
$$

The Bayesian point–estimator of $(\boldsymbol{\beta}_0, \phi_0)$ is given by the vector $(\boldsymbol{\beta}^*, \phi^*)$ that minimizes $\mathcal{R}$. Let us first focus on finding $\boldsymbol{\beta}^*$. Note that this is equivalent to minimizing

$$\mathcal{R}_1(\boldsymbol{\beta}) = D_2(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu}). \qquad (8.6)$$

Compare now the minimization of $\mathcal{R}_1(\boldsymbol{\beta})$ with a different optimization problem for which the solution method is well known. Consider a frequentist (non–Bayesian) GLM with the same response distribution, explanatory variables and weights. Imagine a sample under this model in which the observed response vector is equal to $\mathbb{E}[\boldsymbol{Y}]$. Using the mean value parameterization and Lemma 8.3, the log–likelihood function based on such a sample is given by

$$
\begin{aligned}
\ell(\boldsymbol{\beta}, \phi) &= \log(C(\mathbb{E}[\boldsymbol{Y}], \phi)) - \frac{1}{2\phi} D(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu}) \\
&= \log(C(\mathbb{E}[\boldsymbol{Y}], \phi)) - \frac{1}{2\phi} D_1(\mathbb{E}[\boldsymbol{Y}]) - \frac{1}{2\phi} D_2(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu}),
\end{aligned}
$$

where $\boldsymbol{\mu} = G^{-1}(X\boldsymbol{\beta})$. Since the only term that depends on $\boldsymbol{\beta}$ is the third one, then maximizing $\ell(\boldsymbol{\beta}, \phi)$ is equivalent to minimizing $D_2(\mathbb{E}[\boldsymbol{Y}], \boldsymbol{\mu})$, i.e. the same as minimizing $\mathcal{R}_1(\boldsymbol{\beta})$. Hence, by obtaining the mle of the regression coefficients of this hypothetical frequentist GLM, we obtain $\boldsymbol{\beta}^*$ (or conclude that there is no solution, whenever this is the case). □

Once $\boldsymbol{\beta}^*$ has been found, the invariance property of the relative entropy allows to find the entropic premium straightforwardly.

Corollary 8.2. *If $\boldsymbol{\beta}^*$ is the entropic estimator of the coefficients of a Bayesian GLM, then the entropic premium is given by*

$$\boldsymbol{\mu}^* = G^{-1}(X\boldsymbol{\beta}^*). \tag{8.7}$$

Remark 8.1. *For a saturated model, i.e. when the dimension of $\boldsymbol{\beta}$ is equal to the dimension of $\boldsymbol{Y}$ (in other words $m = p + 1$), the entropic premium is equal to $\mathbb{E}[\boldsymbol{Y}]$. This is because in a saturated model, the predicted mean is equal to the observed response mean.*

8.3.1 Estimation of the Dispersion Parameter

It is important to remark that the credibility estimator from the previous section takes into consideration the uncertainty of the dispersion parameter. This is because the posterior distribution of $\boldsymbol{\beta}$ depends on the posterior of ϕ.

This differs from classical credibility results where the dispersion parameter is considered known (e.g. Jewell (1974) and Diaconis and Ylvisaker (1979)). To the best of our knowledge there is only one article that considers a prior distribution for the dispersion parameter, Landsman and Makov (1998), about which we have the following remarks:

1. The exponential distribution for the index parameter is justified using the principle of maximum entropy. The authors maximize the continuous entropy (that is entropy for continuous random variables), and use it for the index parameter assuming a known mean. Now, the continuous entropy does not have good properties as a measure of information. For instance it is not invariant under bijective transformations, which implies that one can loose or gain information by just transforming a random variable. Thus, the principle of maximum entropy is not a valid justification for the exponential distribution. Nevertheless, it is a valid prior and one can use it in those cases where it reflects properly the out of sample information.

2. A more serious problem exists with their result in Theorem 2; the integrals in (7) are carried out assuming that λ is exponential with mean λ_0. In other words, these

are computed assuming the prior distribution for λ. This is erroneous since it is the posterior distribution that should be used in this integral. This would be justified if the prior for λ were natural conjugate. In this way the posterior of λ would also be exponential, but the parameter of the posterior would be different than the parameter of the prior, in this case.

We have not found a general procedure for obtaining the entropic estimator of the dispersion parameter. We discuss here the cases for which it can be found and present the difficulties in obtaining a general solution. Notice that a point–estimator for ϕ is not necessary to obtain the credibility premium or its uncertainty (which is accounted for in the posterior distribution).

Suppose that the credibility premium $\boldsymbol{\mu}^*$ has been obtained. From (8.5), one can see that finding the entropic estimator ϕ^* of the dispersion parameter is equivalent to minimizing

$$\mathcal{R}_2(\phi) = -\mathbb{E}\left[\log(C(\boldsymbol{Y}, \phi))\right] + \frac{1}{2\phi}\mathbb{E}[D(\boldsymbol{Y}, \boldsymbol{\mu}^*)], \tag{8.8}$$

where $\boldsymbol{\mu}^* = G^{-1}(X\boldsymbol{\beta}^*)$. There are standard methods for minimizing univariate functions, but $\mathcal{R}_2$ is more difficult because the first expectation in (8.8) depends on ϕ. We consider first a special case where this minimization is rather straightforward. This is when there exists a function $H : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ such that

$$-\mathbb{E}[\log(C(\boldsymbol{Y}, \phi))] = H(\mathbb{E}[\boldsymbol{Y}], \phi), \tag{8.9}$$

for every ϕ. In this case the problem simplifies considerably because once $\mathbb{E}[\boldsymbol{Y}]$ and $\mathbb{E}[D(\boldsymbol{Y}, \boldsymbol{\mu}^*)]$ have been found (most likely by simulations), then it is possible to use standard methods to find ϕ^*, since (8.8) becomes

$$\mathcal{R}_2(\phi) = H(\mathbb{E}[\boldsymbol{Y}], \phi) + \frac{1}{2\phi}\mathbb{E}[D(\boldsymbol{Y}, \boldsymbol{\mu}^*)], \tag{8.10}$$

which is simple to evaluate.

A case worth mentioning when (8.9) occurs is when the response distribution is a proper dispersion model (see Jørgensen (1997, Chap. 5)), i.e. when c in (4.4) can be decomposed as

$$c(y, \phi) = d(y)e(\phi), \tag{8.11}$$

for some functions d and e. Then, the first term on the right hand side of (8.8) becomes

$$\begin{aligned}-\mathbb{E}[\log(C(\boldsymbol{Y},\phi))] &= -\mathbb{E}\left[\prod_{i=1}^{m}\log\left(c\left(Y_i,\frac{\phi}{w_i}\right)\right)\right]\\ &= -\mathbb{E}\left[\prod_{i=1}^{m}\log\left(d(Y_i)e\left(\frac{\phi}{w_i}\right)\right)\right]\\ &= -\sum_{i=1}^{m}\mathbb{E}[\log(d(Y_i))] - \sum_{i=1}^{m}\log\left(e\left(\frac{\phi}{w_i}\right)\right).\end{aligned}$$

Since $\sum_{i=1}^{m}\mathbb{E}[\log(d(Y_i))]$ does not depend on ϕ, the problem reduces to minimizing

$$\mathcal{R}_3(\phi) = -\sum_{i=1}^{m}\log\left(e\left(\frac{\phi}{w_i}\right)\right) + \frac{1}{2\phi}\mathbb{E}[D(\boldsymbol{Y},\boldsymbol{\mu}^*)],$$

which can be done using standard optimization methods. Now, it is known that there are only three exponential dispersion models for which the factorization in (8.11) holds: the gamma, inverse Gaussian and normal distributions (this result is commented in Jørgensen (1997, Chap. 5) and proven in Daniels (1980)). Table 8.2 gives e is for these three models.

Distribution	Normal	Gamma	Inverse Gaussian
$e(\phi)$	$\phi^{-1/2}$	$\dfrac{e^{-1/\phi}}{\Gamma(\frac{1}{\phi})\phi^{1/\phi}}$	$\phi^{-1/2}$

Table 8.2: $e(\phi)$ for the three proper exponential dispersion families

Let us now consider the general case where (8.10) does not hold. Again MCMC methods can be helpful. Let $\boldsymbol{Y}^1,\ldots,\boldsymbol{Y}^N$ be N simulations of $\boldsymbol{Y}$ from the posterior predictive distribution (superscripts are used since Y_i was already defined to be the i-th entry of $\boldsymbol{Y}$). Now define

$$\tilde{\mathcal{R}}_N(\phi) = -\frac{1}{N}\sum_{i=1}^{N}\log(C(\boldsymbol{Y}^i,\phi)) + \frac{1}{2\phi N}\sum_{i=1}^{N}D(Y^i,\boldsymbol{\mu}^*),$$

then, for every fixed ϕ

$$\lim_{N\to\infty}\tilde{\mathcal{R}}_N(\phi) = \mathcal{R}_2(\phi) \qquad \text{a.s..} \tag{8.12}$$

Let $\tilde{\phi}_N = \operatorname{argmin}\tilde{\mathcal{R}}_N(\phi)$. Since $\tilde{\mathcal{R}}_N$ is simple to evaluate with a computer, standard univariate optimization methods can be used to find $\tilde{\phi}_N$. The question now is whether $\tilde{\phi}_N$ converges to ϕ^* as $N\to\infty$? We have not found easy–to–check sufficient conditions that guarantee convergence, although the following theorem might be useful in some cases.

Proposition 8.3.1. *If the convergence in* (8.12) *is uniform almost surely w.r.t.* ϕ*, then*

$$\mathcal{R}_2(\phi^*) = \lim_{N\to\infty} \tilde{\mathcal{R}}_N(\tilde{\phi}_N) \qquad a.s.$$

Proof. On the one hand we have that for every $n \in \mathbb{N}$

$$\begin{aligned} & \mathcal{R}_2(\phi^*) \leq \mathcal{R}_2(\tilde{\phi}_n) \\ \therefore \qquad & \mathcal{R}_2(\phi^*) \leq \liminf_n \mathcal{R}_2(\tilde{\phi}_n). \end{aligned}$$

On the other hand, let $\epsilon > 0$, since $\tilde{\mathcal{R}}_N \to \mathcal{R}_2$ uniformly a.s., then with probability one there exists $M > 0$, such that for every $n \geq M$,

$$|\tilde{\mathcal{R}}_n(\tilde{\phi}_n) - \mathcal{R}(\tilde{\phi}_n)| < \epsilon.$$

By the definition of $\tilde{\phi}_n$,

$$\tilde{\mathcal{R}}_n(\tilde{\phi}_n) \leq \mathcal{R}_n(\phi^*), \qquad \text{for all } n \in \mathbb{N}.$$

Then for every $n \geq N$,

$$\mathcal{R}_2(\tilde{\phi}_n) - \epsilon < \mathcal{R}_n(\phi^*),$$

thus

$$\begin{aligned} & \limsup_n \mathcal{R}_2(\tilde{\phi}_n) - \epsilon \leq \limsup \mathcal{R}_n(\phi^*) \\ \therefore \qquad & \limsup_n \mathcal{R}_2(\tilde{\phi}_n) - \epsilon \leq \mathcal{R}_2(\phi^*) \qquad a.s. \end{aligned}$$

Since this is true for $\epsilon > 0$, this implies that

$$\limsup_n \mathcal{R}_2(\tilde{\phi}_n) \leq \mathcal{R}_2(\phi^*),$$

and therefore

$$\mathcal{R}_2(\phi^*) = \lim_{n\to\infty} \mathcal{R}(\tilde{\phi}_n) \qquad a.s.$$

□

8.4 On the Applicability of the Entropic Premium

The previous section showed how one can find the entropic premium of a GLM, theoretically. In this section we address its practicability. In other words, we address the following question: is entropic credibility feasible for real–life datasets?

From Proposition 8.2 and Corollary 8.2, we have that once the response distribution, explanatory variables and prior have been chosen, the following steps give the entropic premium:

1. Find $\mathbb{E}[\boldsymbol{Y}]$ (see the paragraph preceding Proposition 8.2 for the definition of $\mathbb{E}[\boldsymbol{Y}]$).

2. Fit a frequentist GLM with the same covariates, response distribution and weights, but with observed response vector equal to $\mathbb{E}[\boldsymbol{Y}]$. This gives $\boldsymbol{\beta}^*$, the entropic estimator of the coefficients.

3. Find the entropic mean using (8.7).

Steps 2 and 3 are simple: one can perform these computations in R (see R Core Team (2019)) without major problems. The difficult part is Step 1. To the best of our knowledge, the simplest way to solve this problem is using Markov Chain Monte Carlo (MCMC). In the paragraphs that follow we give some recommendations on how to use this method.

It is important to consider that the greater m (the dimension of $\mathbb{E}[\boldsymbol{Y}]$), the more demanding the computations (both in terms of memory and CPU). Thus, it is very useful to first aggregate the data as in (4.8). This can drastically reduce m and turn an infeasible computation into something manageable.

A continuous variable can make data aggregation useless, especially when this happens with several different variables in the dataset. In such cases one should consider converting the support of these variables into intervals, hence transforming them into categorical variables.

Using Bayesian methods for variable selection can be time consuming. This is because one would need to run MCMC simulations for each combination of variables. A pragmatic approach to deal with this is to choose the variables using a frequentist GLM (which is much faster to fit). The resulting combination of variables can be used to build a starting model in the Bayesian case.

8.5 Vehicle Insurance Example

In this section we use the R interface to STAN to find entropic credibility estimators for a frequency and a severity model for a publicly available dataset. The main purpose of this section is to show that it is feasible to obtain credibility premiums.

The dataset appears in de Jong and Heller (2008). It is based on one year policies from 2004 or 2005 and it consists of 67,856 policies. The dataset can be downloaded from the companion site of the book: `http://www.acst.mq.edu.au/GLMsforInsuranceData`, it is the dataset called Car. Table 8.3 shows the description of the dataset variables provided by the website. The data is also provided by the R package *insuranceData* (Wolny-Dominiak and Trzesiok (2014)), which can be installed directly from CRAN, where it corresponds to the dataset called `dataCar`.

Variable name	Description
`veh_value`	vehicle value, in \$10,000s
`exposure`	0-1
`clm`	occurrence of claim (0 = no, 1 = yes)
`numclaims`	number of claims
`claimcst0`	claim amount (0 if no claim)

`veh_body`	vehicle body, coded as
	BUS
	CONVT = convertible
	COUPE
	HBACK = hatchback
	HDTOP = hardtop
	MCARA = motorized caravan
	MIBUS = minibus
	PANVN = panel van
	RDSTR = roadster
	SEDAN
	STNWG = station wagon
	TRUCK
	UTE - utility

`veh_age`	age of vehicle: 1 (youngest), 2, 3, 4
`gender`	gender of driver: M, F
`area`	driver's area of residence: A, B, C, D, E, F
`agecat`	driver's age category: 1 (youngest), 2, 3, 4, 5, 6

Table 8.3: Vehicle insurance variables

There is one continuous variable: `veh_value`. As suggested in Remark 4.2 we divide this variable into three intervals $[0, 1.2)$, $[1.2, 1.86)$ and $[1.86, \infty)$, which we label as `P1`, `P2` and `P3` respectively.

8.5.1 Frequency Model

Table 8.4 contains the information of the Bayesian GLM we used for the frequency. Note that the value between parenthesis at the right of each explanatory variable corresponds to the reference category used in the model. After aggregating the data for this model the observations were reduced from 67,856 to 212.

Model information		MCMC information	
Response distribution	Poisson	**No. of chains**	4
Weight variable	`exposure`	**Warmup period**	$3,000$
Covariates	`veh_value(P1)`	**Simulations kept(per chain)**	$47,000$
	`veh_body(HBACK)`		
	`agecat(1)`		
Link function	log		
Prior	$\beta_0 \sim N(-1.5, 1)$		
$i \geq 1$,	$\beta_i \sim N(0, 1)$.		
	all independent		

Table 8.4: Frequency model.

The diagnostic plots of our MCMC simulations for the betas showed convergence, but due to the number of parameters we omit them here. We found the entropic estimate of the betas. Table 8.5 shows their estimated values an their significance (Definition 6.5).

For assessing the model we computed its residuals. Figure 8.1 shows plots of the response and normalized residuals. A 0.95 confidence region is shown in both plots.

Except for the clear outliers present in the normalized residuals, the graph suggests a good fit. Looking further into the outliers we realized that they correspond to classes with exposure close to zero. Thus, the observed extreme value is explained by the fact that

Variable	Estimated Betas	Significance	Variable	Estimated Betas	Significance
(Intercept)	-1.7079	0	RDSTR	0.3671	0.345
veh_valueP2	0.1116	0.001	SEDAN	0.0176	0.319
veh_valueP3	0.2353	0	STNWG	-0.0307	0.247
BUS	0.8786	0.011	TRUCK	-0.0824	0.179
CONVT	-0.6559	0.073	UTE	-0.2322	0
COUPE	0.4036	0.001	agecat2	-0.1736	0.001
HDTOP	0.0323	0.371	agecat3	-0.2339	0
MCARA	0.4618	0.06	agecat4	-0.2618	0
MIBUS	-0.1316	0.178	agecat5	-0.4822	0
PANVN	0.0486	0.362	agecat6	-0.4655	0

Table 8.5: Summary table of frequency model estimated coefficients

the difference between the observed and predicted values is magnified when divided by the exposure. Hence, these outliers do not suggest a lack of fit of the model.

Our next diagnostic plot is presented in Figure 8.2. It shows the observed residuals and 8 randomly chosen sets of replicated residuals. In this kind of graph, one tries to see whether there is some pattern in the observed data that is not present in the replicated ones. This does not appear to be the case in this example.

We use two test quantities based on the residuals: their mean and variance. Their respective *p*-values are 0.502 and 0.325. Both are far from 0 and 1 which does not suggest a lack of fit.

A common issue with frequency models and specially with the Poisson distribution, is that the model does not describe well the probability of zero accidents. It is not straightforward to come up with a test for the prediction of the number of zeros for aggregated data. We devised a way to do it, but it is computationally intensive.

The test consists in using the PPD to obtain replicates of the non aggregated test set (this is why it is computationally intensive), then for each replication we count the number of zeros (this is, the number of policies with no accidents) for each class. We perform two diagnostics for the number of zeros using these replications:

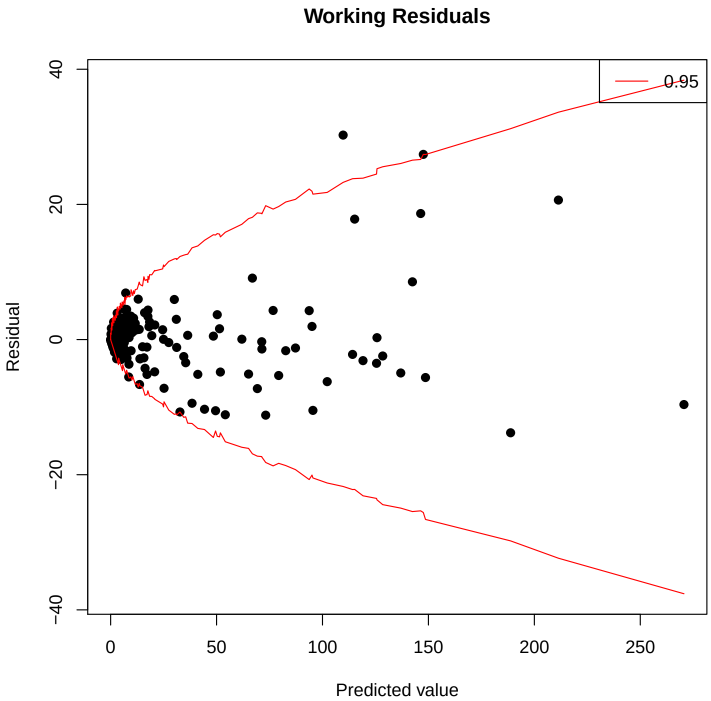


Figure 8.1: Residuals for frequency model

1. **A *zero residuals* plot:** For each class we plot the observed number of zeros minus the expected number of zeros along with a confidence region with some fixed confidence level.

2. **zeros test:** We use the total number of zeros in the portfolio as a test quantity and compute its p-value.

Figure 8.3 shows the plot of zero residuals and the p-value of the zeros test is 63232. They both suggest a good fit of the model for predicting the number of zeros.

Figure 8.4 shows a graph of the entropic frequency estimates in increasing order for all the classes and compares it to premiums obtained using a frequentist GLM. The left y-axis gives the values of these estimates. They gray bars represent the sum of the durations of all the observations in the class. The right y-axis gives the values of the gray bars.

8.5.2 Severity Model

Table 8.6 shows the information of the Bayesian GLM used for the severity. After aggregating the data for this model the number of observations was reduced to 101. The average observed

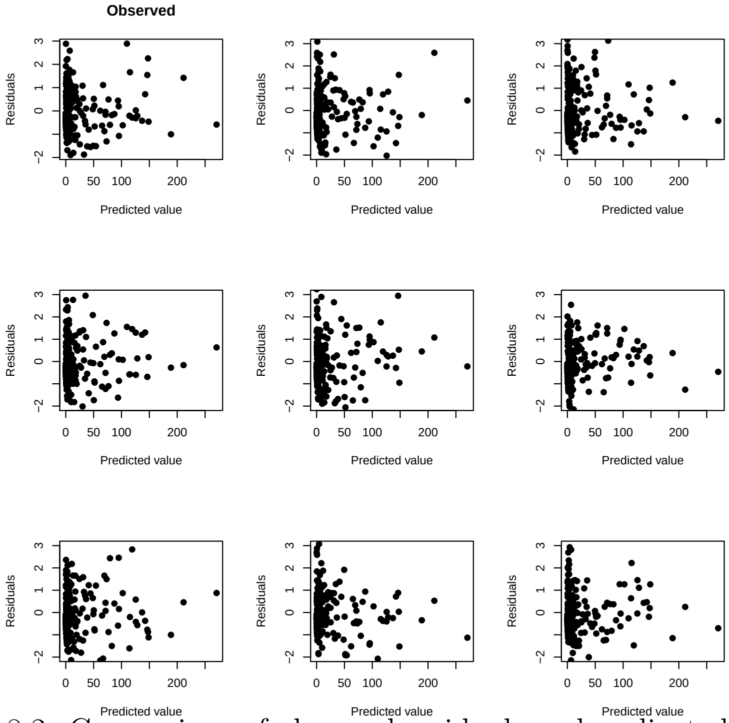


Figure 8.2: Comparison of observed residuals and replicated residuals

response is 2456 whose log is 7.8. Thus, for β_0 we use a normal distribution with mean 8 and variance 9. For the other coefficients we use a standard normal distribution.

The entropic betas and their significance are shown in Table 8.7.

Figure 8.5 shows the plots of response and normalized residuals of the model. The p-values of their mean and variance are 0.39 and 0.564, respectively.

Figure 8.6 shows a graph of the entropic severity estimates in increasing order for all the classes and compares it to premiums obtained using a frequentist GLM. The left y-axis gives the values of these estimates. They gray bars represent the total number of claims in the class. The right y-axis gives the values of the gray bars.

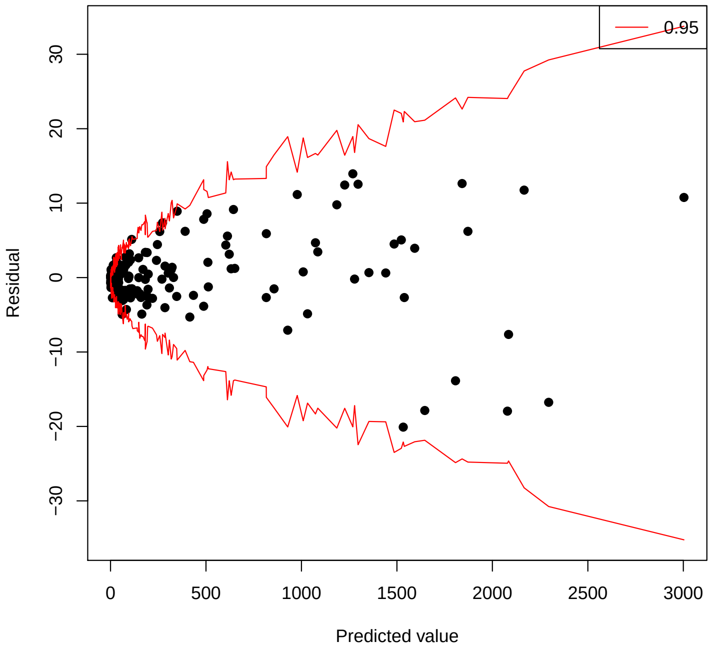


Figure 8.3: Residuals for the number of zeros

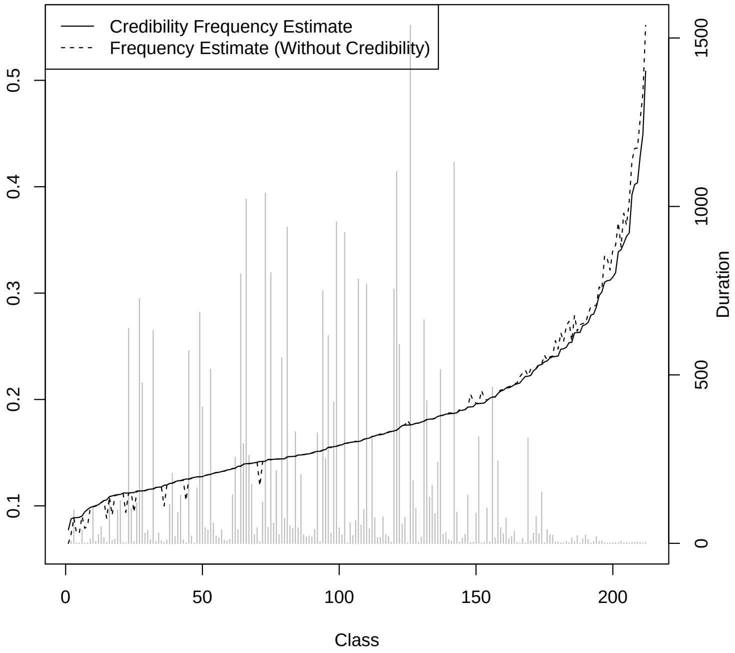


Figure 8.4: Entropic and frequentist frequency estimation comparison

Model information		MCMC information	
Response distribution	gamma	**No. of chains**	4
Weight variable	`numclaims`	**Warmup period**	$3,000$
Covariates	`agecat(1)`		
	`gender(F)`	**Simulations kept**	$47,000$
	`area(ABCD)`	**(per chain)**	
	`veh_value(P1)`		
Link function	log		
Prior	$\beta_0 \sim N(8,9)$,		
$i \geq 1$,	$\beta_i \sim N(0,4)$,		
$\phi \sim$	$N(10,20)T(0,\infty)$,		
	all independent.		

Table 8.6: Severity model.

Variable	Estimated Betas	Significance	Variable	Estimated Betas	Significance
(Intercept)	7.8394	0	genderM	0.1779	0.001
agecat2	-0.192	0.037	areaE	0.1557	0.068
agecat3	-0.2932	0.003	areaF	0.3988	0
agecat4	-0.2843	0.003	veh_valueP2	-0.1208	0.045
agecat5	-0.4041	0	veh_valueP3	-0.1518	0.017
agecat6	-0.3175	0.009			

Table 8.7: Summary table of severity model estimated coefficients

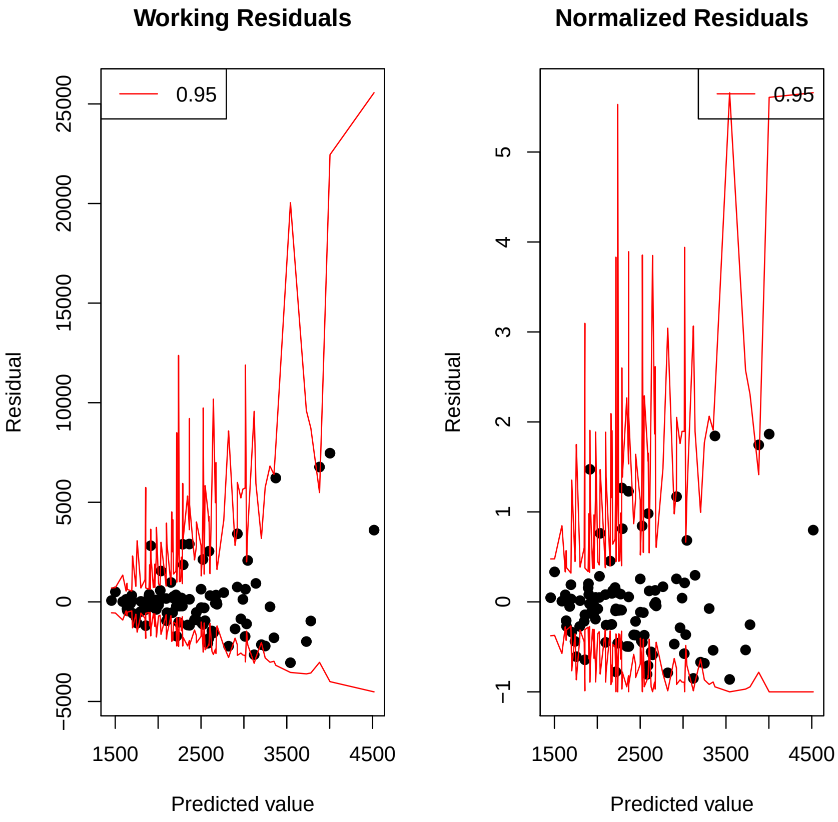


Figure 8.5: Residuals for severity model

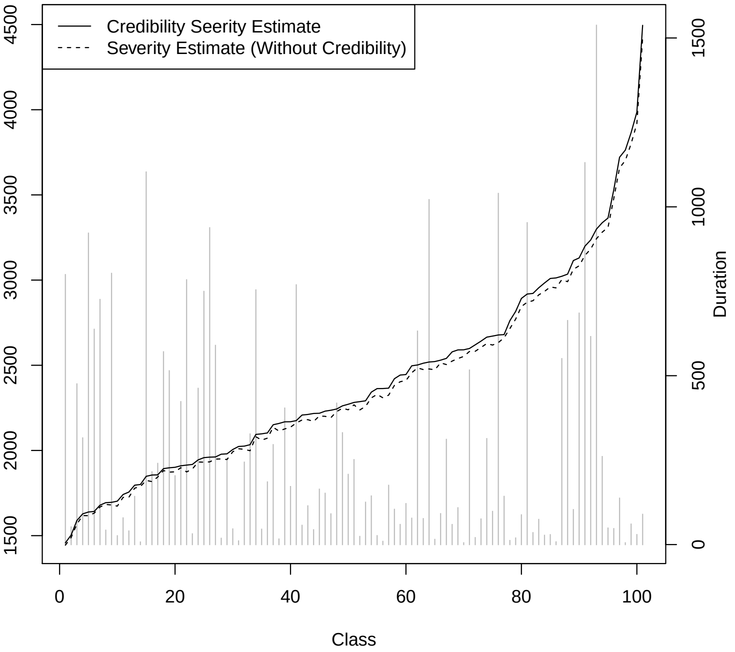


Figure 8.6: Entropic and frequentist severity estimation comparison

Chapter 9

The Unifed Distribution

The next chapter talks about risk loading estimation. One of the important points discussed there is that uncertainty about the duration should be included in the method used for deciding the risk loading. Thus, a model for the duration is required.

In order for a duration model to work in our context we need it to have support on (0,1) and ideally it should be able to deal with different insurance portfolio classes so it can be used with heterogeneous observations. A good candidate that satisfies these requirements is the beta regression (see Section 9.7). Nevertheless it has the drawback that it is not suitable for data aggregation (see Section 9.7.1) and this makes the MCMC algorithm take significantly more time.

We created a new distribution that we have called *unifed*. It is simpler and much less versatile than the beta distribution. Nevertheless it's shape reassembles histograms of observed durations and it is able to fit exposure data reasonably well. Moreover it is a EDF and therefore (4.8) can be used for aggregating data and it can also be used as the response distribution of a GLM.

We have developed an R (R Core Team (2019)) package called `unifed` (Quijano Xacur (2019b)) for working with this distribution. Among other features it allows to use the unifed distribution with the `glm()` function in R. The package is part of the Comprehensive R Archive Network (CRAN). The code and documentation of the version in CRAN can be seen at `https://CRAN.R-project.org/package=unifed`. For the development of the code the following gitlab repository is used: `https://gitlab.com/oquijano/unifed`. A vignette for introducing the distribution and the package has also been prepared. The vignette for the version in CRAN is at `https://cran.r-project.org/web/packages/unifed/` and the one of the development version can be found at `https://oquijano.gitlab.io/unifed/` (they will be the same unless new features are added to the development version that have not been added to CRAN yet).

In this chapter we introduce the unifed distribution, we talk about the numerical challenges that had to be solved for implementing software to work with it. We illustrate also some graphs to show that it's shape is able to adapt to exposure data. We give an example of a unifed GLM applied to heterogeneous data. We also comment on the difficulties of aggregating data for the commonly used alternative to the unifed, the beta distribution.

9.1 Definition

The unifed is the Exponential Dispersion Family (EDF) generated by the uniform distribution (see Chapters 2 and 3 of Jørgensen (1997) to see how an EDF can be generated from a moment generating function). From there it's name, unifed=**unif**orm+**ed**f. What follows contains references to some functions in the in the *unifed* R package. `This font format` is used for those references.

To express the density of the *unifed* we need the density of the distribution of the sum of n independent $uniform(0,1)$ random variables. This corresponds to the Irwin-Hall distribution (see Johnson et al. (1995)) and its density function is

$$h(y;n) = \frac{1}{(n-1)!}\sum_{k=0}^{\lfloor y \rfloor}(-1)^k \binom{n}{k}(y-k)^{n-1}, \qquad y \in [0,n], n \in \mathbb{N}, \tag{9.1}$$

where $\lfloor y \rfloor$ denotes the floor of y. The canonical and index spaces of the unifed family are $\Theta = \mathbb{R}$ and $\Lambda = \{1,2,3,4\ldots\}$. The cumulant generator is

$$\kappa(\theta) = \begin{cases} \log\left(\frac{e^\theta - 1}{\theta}\right) & \text{if } \theta \neq 0, \\ 0 & \text{if } \theta = 0. \end{cases} \tag{9.2}$$

We denote the unifed distribution with canonical parameter θ and dispersion parameter ϕ with $unifed(\theta, \phi)$. It's density is given by

$$f(x;\theta,\phi) = \frac{h(x/\phi, 1/\phi)}{\phi} \exp\left(\frac{x\theta - \kappa(\theta)}{\phi}\right), \tag{9.3}$$

where h and κ are as in (9.1) and (9.2), respectively and $x \in [0,1], \theta \in \mathbb{R}, \phi \in \left\{1, \frac{1}{2}, \frac{1}{3}, \ldots\right\}$.

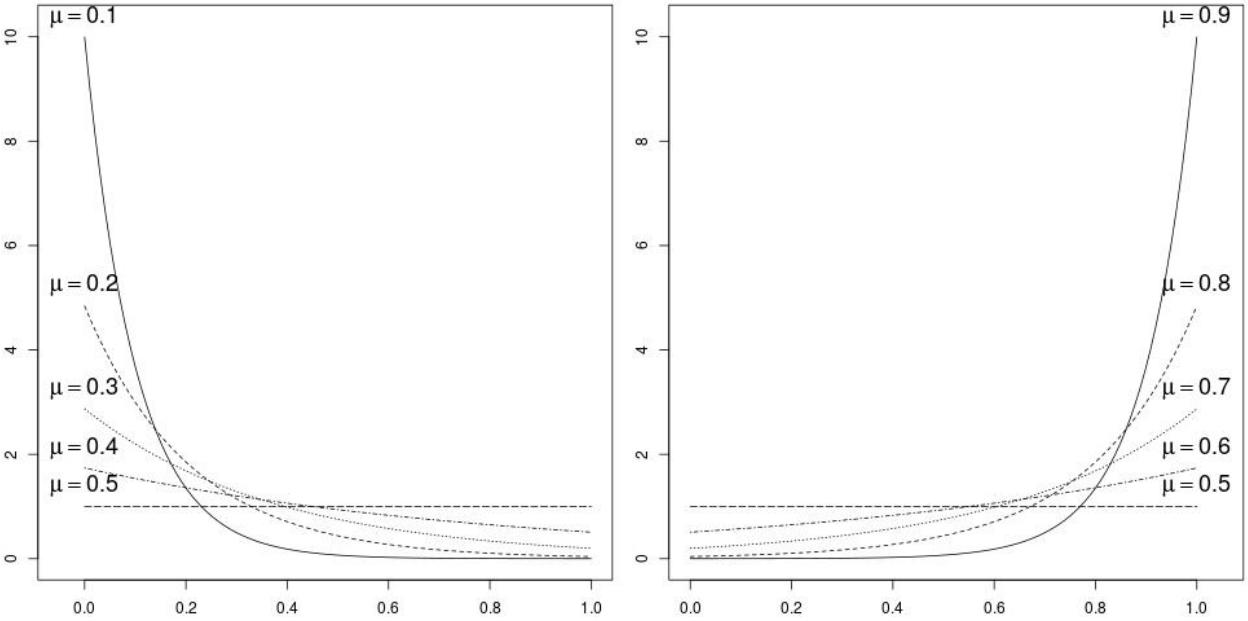


Figure 9.1: Density of the unifed for different values of its mean μ

The unifed package does not contain an implementation of (9.3). This is because we did not find a numerically stable way to compute h. To show this, the package includes the function `dirwin.hall` that computes h. Table 9.1 shows the results we get by calling this function with n set to 50 and varying the values of y. The changes of sign indicate that a float overflow is happening.

Code	Result
`dirwin.hall(35,50)`	0.0674864
`dirwin.hall(36,50)`	-13.12745
`dirwin.hall(37,50)`	45.44388
`dirwin.hall(38,50)`	-37.44488

Table 9.1: Float overflow of the Irwin-Hall implementation.

Thus, the package calls unifed distribution the 1-parameter special case of (9.3) where

$\phi = 1$, which we denote with $unifed(\theta)$. This simplifies the density to

$$f(x;\theta) = \begin{cases} \frac{\theta}{e^\theta - 1} e^{x\theta} & \text{if } \theta \neq 0 \\ 1 & \text{if } \theta = 0 \end{cases} \qquad \text{for } x \in (0,1). \tag{9.4}$$

The functions `dunifed`, `punifed`, `qunifed` and `runifed`, give the density, distribution, quantile and simulation functions, respectively of this simplified version. The mean and variance of each element of the family are given by

$$\mathbb{E}[X] = \dot{\kappa}(\theta) = \begin{cases} \frac{(\theta-1)e^\theta + 1}{\theta(e^\theta - 1)} & \text{if } \theta \neq 0, \\ \frac{1}{2} & \text{if } \theta = 0, \end{cases} \tag{9.5}$$

$$\mathbb{V}[X] = \ddot{\kappa}(\theta) = \begin{cases} \left(\frac{e^{2\theta} - (\theta+2)e^\theta + 1}{\theta^2 (e^\theta - 1)^2}\right) & \text{if } \theta \neq 0, \\ \frac{1}{12} & \text{if } \theta = 0, \end{cases} \tag{9.6}$$

where $\dot{\kappa}$ and $\ddot{\kappa}$ are the first and second derivative of κ, respectively. We have not been able to find an analytical expression for the inverse function $\dot{\kappa}^{-1}$. Thus, it has not been possible either to find analytical expressions for the variance function and unit deviance of the unifed. Nevertheless, the *unifed* package contains the function `unifed.kappa.prime.inverse` that uses the Newthon Raphson method to implement the inverse of $\dot{\kappa}$. This allows us to get a numerical solution for the variance function by using the relation $\mathbf{V}(\mu) = \ddot{\kappa}(\dot{\kappa}^{-1}(\mu))$. This is implemented in the function `unifed.varf`. Figure 9.2 shows a plot of the variance function.

Similarly, since the unifed is a regular EDF, we can compute the unit deviance by using the following relation (commented in (4.5))

$$d(y,\mu) = 2\left[y\{\dot{\kappa}^{-1}(y) - \dot{\kappa}^{-1}(\mu)\} - \kappa(\dot{\kappa}^{-1}(y)) + \kappa(\dot{\kappa}^{-1}(\mu))\right]. \tag{9.7}$$

The function `unifed.unit.deviance` computes the unit deviance using (9.7). As mentioned in Chapter 4 (see (4.4)), the unit deviance can be used to reparametrize the distribution in terms of it's mean and dispersion parameter. We denote with $unifed^*(\mu,\phi)$ the unifed distribution with mean μ and dispersion parameter ϕ and when $\phi = 1$ we write simply $unifed^*(\mu)$.

Figure 9.1 shows plots of the unifed distribution for different values of its mean. We can see that except for $\mu = 0.5$, it is always monotone. For $\mu < 0.5$ it is strictly decreasing and

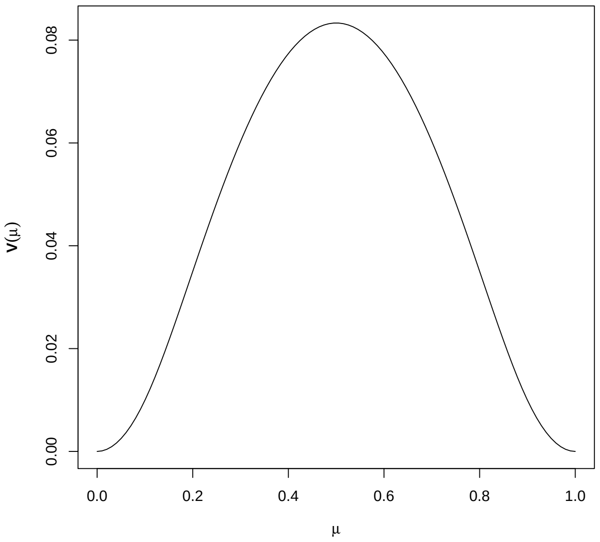


Figure 9.2: Variance function of the Unifed

the mode is at zero. For $\mu > 0.5$ it is strictly increasing and the mode is at one. The R code used for producing this plot can be found at `https://gitlab.com/oquijano/unifed/snippets/1786224`.

It is possible to use the unifed as the response distribution for a GLM. In this case, the dispersion parameter ϕ must be fixed to 1 and the weight of each class is the number of observations in the class. The unifed R package (Quijano Xacur (2019b)) provides the function `unifed` that returns a family object than can be used inside the `glm` function.

9.2 Numerical and Computational Considerations for Software Implementations

Functions have to be numerically stable in the support so simulations are not stopped due to numerical errors. A few numerical and computational issues were found while working on the unifed R package. This section reports these problems and the solutions used to mitigate them. This is done for two purposes: first it is useful for understanding the exact behavior of the R package. Second, it would be useful for someone wanting to work with the unifed distribution in a language different than R.

9.2.1 Newthon-Raphson Overflows

As mentioned in the previous section $\dot{\kappa}^{-1}$, which takes values in [0,1] is implemented using the Newthon-Raphson method. When the argument take values very close to 0.5, the function returns strange values. This is because the operations involved for computing the value of the function are too extreme for the precision of the computer.

In order to solve the problem around 0.5, this function returns 0 (which is the image of 0.5), for every μ with $|\mu - 0.5| < 10^{-5}$.

9.2.2 Cumulant Generator Blowing Up

In the previous section we saw that the cumulant generator of the unifed is

$$\kappa(\theta) = \begin{cases} \log\left(\frac{e^\theta - 1}{\theta}\right) & \text{if } \theta \neq 0, \\ 0 & \text{if } \theta = 0. \end{cases}$$

The difficulty of implement κ in a computer is that e^θ exceeds the largest possible float very fast. For instance, in R `exp(710)` returns `Inf`.

To avoid this problem the function that implements κ in the R package, `unifed.kappa`, uses the following approximation function

$$\kappa_{mod}(\theta) = \begin{cases} \kappa(\theta) & \text{if } \theta \leq 50, \\ \theta - \log(\theta) & \text{if } \theta > 50. \end{cases}$$

The justification for this approximation is that if y is defined as

$$y = \log\left(\frac{e^\theta - 1}{\theta}\right),$$

then also,

$$y = \theta - \log(\theta + e^{-y}).$$

Now, κ is an increasing function and y goes to infinity as θ goes to infinity. Thus for large values of θ the term e^{-y} is close to zero. We use 50 as the threshold for the approximation since evaluating `log( ( exp(50) - 1) / 50 ) - ( 50 - log (50) )` in R returns zero already.

9.2.3 Implementing $\dot{\kappa}$

From (9.5) we see that $\dot{\kappa}$ has e^{θ} in the numerator and denominator. This is a problem when θ takes large values. To solve this, the implementation of the software uses the following equivalent expression for $\dot{\kappa}$

$$\dot{\kappa}(\theta) = \frac{1}{1-e^{-\theta}} - \frac{1}{\theta}.$$

9.2.4 Implementing $\ddot{\kappa}$

$\ddot{\kappa}$ also has the problem of having exponential functions in the numerator and denominator that gives problems for large values of θ. It can be rewritten as

$$\ddot{\kappa}(\theta) = \frac{1}{\theta^2} - \frac{e^{-\theta}}{\left(e^{-\theta}-1\right)^2},$$

which is not a problem anymore when $\theta \to \infty$ although now the numerator of the second term goes to infinity when $\theta \to -\infty$. Now, the limit of the second term when $\theta \to -\infty$ is zero. Thus, the package implements the following approximation

$$\ddot{\kappa}_{mod}(\theta) = \begin{cases} \frac{1}{\theta^2} - \frac{e^{-\theta}}{\left(e^{-\theta}-1\right)^2} & \text{if } \theta > -100, \\ \frac{1}{\theta^2} & \text{otherwise.} \end{cases}$$

9.2.5 Cumulative Distribution Function

By taking the integral of (9.4) we get that for $\theta \neq 0$ the cumulative distribution of the unifed is

$$F(x;\theta) = \begin{cases} 0 & \text{if } x \leq 0, \\ \frac{e^{\theta x}-1}{e^{\theta}-1} & \text{if } x \in (0,1), \\ 1 & \text{if } x \geq 1. \end{cases}$$

There is again the problem of having powers of e in the numerator and denominator. We care particularly about the log of the function above since it is what stan uses for truncating distributions. The function `unifed_lcdf` in the stan code uses the following implementation to prevent problems when $x \in (0,1)$ and large θ

$$\log(F(x;\theta)) = \begin{cases} \log\left(\frac{e^{x\theta}-1}{e^{\theta}-1}\right) & \text{if } \theta \leq 50, \\ (x-1)\theta + \log(1-e^{-\theta x}) - \log(1-e^{-\theta}) & \text{otherwise.} \end{cases}$$

9.3 Modelling Duration with the Unifed

As mentioned at the beginning of this chapter, we propose the unifed as a distribution for modelling duration. In this section we talk about some usual properties of duration observations and show some simple examples.

Definition 9.1. *The* duration *is defined as the amount of time measured in years that a policy is in force.*

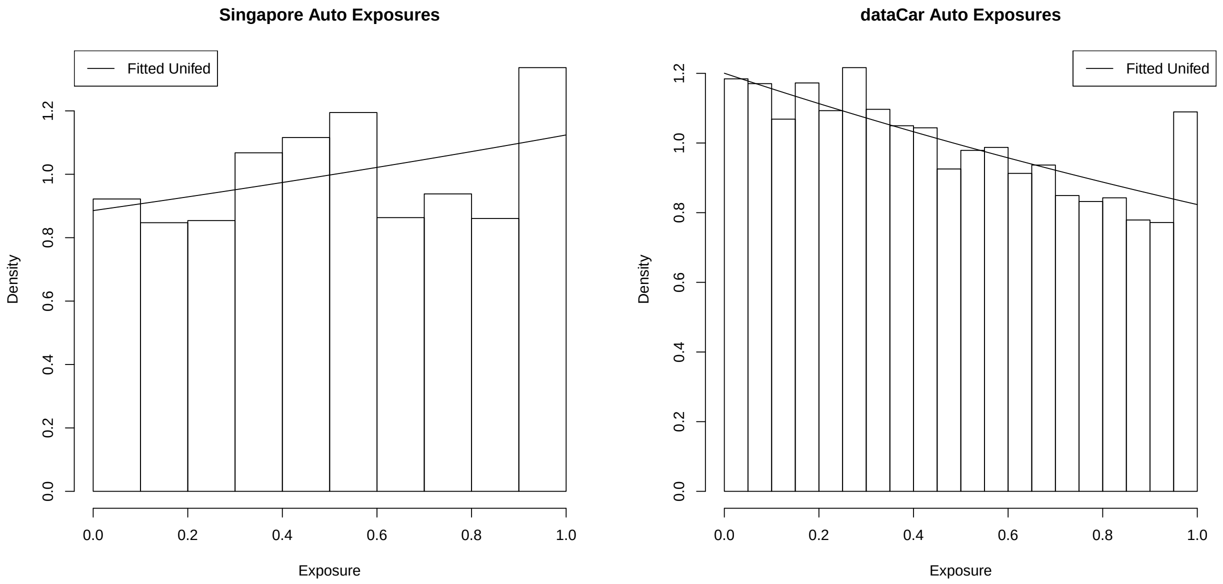
Insurers usually analyze their experience (i.e. their losses, earned premium and exposure) in one year periods. Depending on the analysis one chooses one of two different interpretations of *one year*: *accident year* and *policy year*.

Under *accident year* one reports losses, earned premiums and exposures that occurred from the 1st of January to the 31st of December of the year in question. Under *policy year* one reports losses, earned premiums and exposures with respect to a year after the policy's effective start date.

To make this clear we give an example: imagine a policy that started on July 1st, 2017 and expire on June 30th, 2018 and had two losses, one in October 2017 with a cost of \$300.00 and another one in February 2018 with a cost of \$200.00. For the accident year 2017, this policy is reported to have exposure of 0.5 (since it started in July), and one loss of \$300.00. For the policy year 2017, it is reported to have exposure one and two losses, one for \$300.00 and another one of \$200.00. Suppose the policy is renewed and cancelled on January 1st 2019 with no further claims. For the accident year 2019, this policy has exposure 1. This is because since the policy was renewed, it was in force for the entire year. In contrast, for the policy year 2018 it as exposure 0.5. This is because when the policy year starts to run on the date of renewal.

Since usually most insureds renew their policies, it is common to see an important proportion of exposures with value one for both policy and accident year.

Figure 9.3 shows two histograms of exposure data. The data used for them comes from the R package *insuranceData* (Wolny-Dominiak and Trzesiok (2014)). The exposures used for generating Figure 9.5a come from the dataset called *SingaporeAuto* and the source is the General Insurance Association of Singapore. It includes several characteristics for explaining



(a)

(b)

Figure 9.3: Two histograms of exposure data from the insuranceData R package

automobile accident frequency. The histogram in Figure 9.5b comes from the dataset called *dataCar* in the same R package and it first appeared in de Jong and Heller (2008). Along with both histograms we show the density of a unifed distribution with the maximum likelihood estimate for the canonical parameter θ.

9.4 Data Aggregation

As an EDF it is possible to summarize a unifed sample without loss of information for estimating either its mean or canonical parameter.

Let $Y_1, \cdots, Y_n$ be an i.i.d. sample from a $unifed^*(\mu)$. From the properties of exponential families, we know that $\bar{Y}$, the sample mean, is a sufficient statistic for the mean. By (4.8) we also know that $\bar{Y} \sim unifed^*(\mu, 1/n)$. This implies that for any inference about μ (or θ), we will arrive to the same conclusion whether we use the likelihood function of $\bar{Y}$ or the one corresponding to the entire sample.

Moreover, since maximizing the likelihood function is equivalent to minimizing the deviance, the maximum likelihood extimator (mle) of an i.i.d. unifed sample is given by $\hat{\mu} = \bar{Y}$.

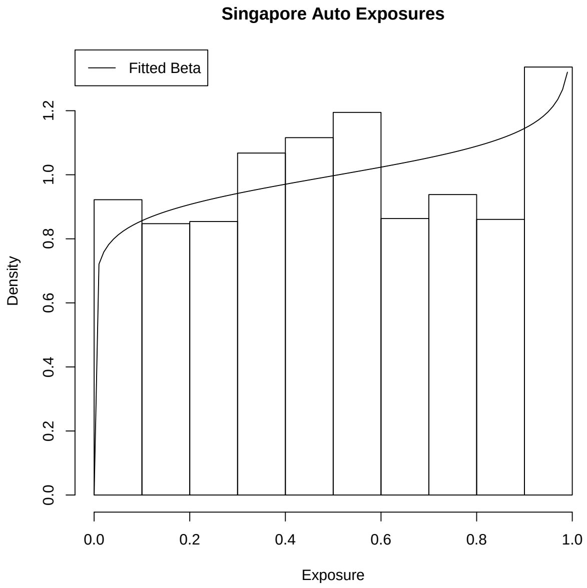


Figure 9.4: Beta fit for Singapore auto histogram

Let us compare this to the mle computation for a beta sample. It does not have a sufficient statistic with known distribution (see more about this in Section 9.7.1). Thus, given an i.i.d. sample coming from the beta distribution one has to use the entire sample for inference. A numerical method is necessary for finding the mle from such a sample, but there are fast software implementations for doing this even for large samples. An example is the function `beta.mle` provided by the R package *Rfast* (Papadakis et al. (2018)), which was used to produce Figure 9.4, which is the same histogram than in Figure 9.5a, but now with a beta fit instead of the unifed.

The situation is not the same for Bayesian estimation where the reduced sample size provided by the sufficient statistic makes the difference much more noticeable. This is because the likelihood function is simplified to one term and therefore much faster to evaluate.

At this point you might be thinking: "wait a second, all of this sounds very nice, but the distribution of $\bar{Y}$ is $unifed^*(\mu, \frac{1}{n})$ whose density (9.1) we cannot evaluate due to the numerical stability problems for evaluating h". True, but there is a fix for this. Suppose you have an observed sample $y_1, \cdots, y_n$ from a unifed distribution and let π be some prior for θ.

By (4.8) and (9.1) the likelihood function of this sample is given by

$$L(\theta|y_1, \cdots, y_n) = nh(n\bar{y}, n)\exp(n[\bar{y}\theta - \kappa(\theta)]).$$

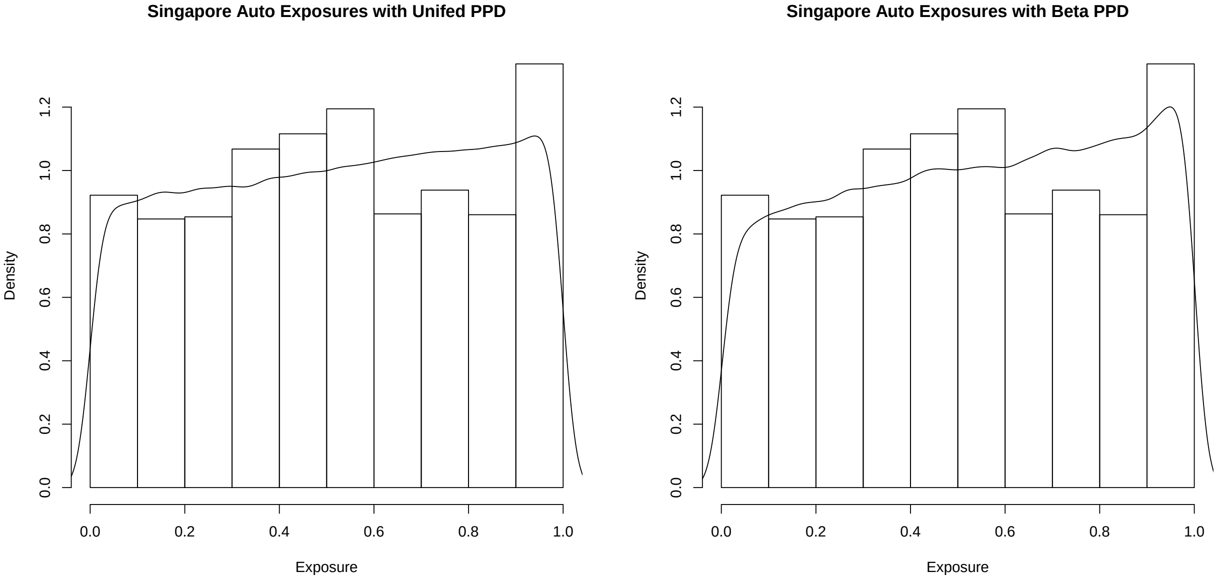
Then the posterior distribution can be expressed as

$$\pi(\theta|y_1, \cdots, y_n) \propto nh(n\bar{y}, n)\exp(n[\bar{y}\theta - \kappa(\theta)])\pi(\theta).$$

Now, $nh(n\bar{y}, n)$ does not depend on θ. Since the Hamiltonian Monte Carlo (HMC) requires the distribution up to a multiplicative constant, we can use instead the following proportional relation to feed into the HMC

$$\pi(\theta|y_1, \cdots, y_n) \propto \exp(n[\bar{y}\theta - \kappa(\theta)])\pi(\theta),$$

which does not contain the problematic h.



(a)

(b)

Figure 9.5: Estimated posterior predictive distribution for Unifed and Beta distributions for Singapore auto exposures

To illustrate the difference in computational time we obtained simulations of the posterior predictive distribution (PPD) for the Singapore auto exposures that contains 7483 observations. For both distributions we ran four parallel chains on a 4-cores computer with a warmup period of 3,000 and we kept the following 47,000 observations. This gives a total of 188,000 simulations for each distribution. We used a normal distribution with mean 0 and standard deviation of 20 as the prior for the canonical parameter of the unifed, while truncated normal distributions on $(0, \infty)$ with mean 4 and standard deviation of 20 were used for both parameters of the beta distribution. The simulations took 1.024 seconds for the unifed chains and 210.014 seconds (3.5 minutes) for the beta distribution. Figure 9.5 shows a gaussian kernel smoothed density along with the histogram of the observations for each case.

9.5 The Unifed GLM

For a unifed distribution the dispersion parameter ϕ can take values in $\{1, \frac{1}{2}, \cdots\}$. As mentiones in Section 9.1 evaluating the density for $\phi \neq 1$ is problematic. Thus, in this thesis we set $\phi = 1$ for every unifed GLM. The unifed glm family provided in the R package does the same.

In the frequentist case, if one is interested in being able to use other values of ϕ in a specific case, we recommend to first estimate the betas with $\phi = 1$ and then to use some criterion for choosing the best value of ϕ. Remember that the β's found this way are optimal regardless of the value of ϕ.

Using the terminology from Section 4.2 with $\phi = 1$, depending on whether we parametrize with respect to $\boldsymbol{\beta}$ or $\boldsymbol{\mu}$, we can express the likelihood function of the unifed GLM in the the following two ways

$$f(\boldsymbol{y}|\boldsymbol{\mu}, \phi) = \boldsymbol{c}(\boldsymbol{y}, 1) \exp\left(-\frac{1}{2} D(\boldsymbol{y}, \boldsymbol{\mu})\right) \tag{9.8}$$

$$f(\boldsymbol{y}|\boldsymbol{\theta}, \phi) = \boldsymbol{a}(\boldsymbol{y}, 1) \exp\left(\boldsymbol{y}^T W \boldsymbol{\theta} - \mathbf{1}^T W \boldsymbol{\kappa}(\boldsymbol{\theta})\right). \tag{9.9}$$

In the discussion that follows we need to make reference to the expresions inside the exp function of both equations above. We introduce then the following notation for the power in

(9.9). Define the function E as

$$\begin{aligned} E(\boldsymbol{y}, \boldsymbol{\theta}) : &= \boldsymbol{y}^T W \boldsymbol{\theta} - \mathbf{1}^T W \boldsymbol{\kappa}(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{m} w_i (y_i \theta_i - \kappa(\theta_i)). \end{aligned}$$

Even though both parametrizations define the same model, there is a big difference in the time required to fit each of them. This is because the implementation of $\dot{\kappa}^{-1}$ uses the Newton-Raphson algorithm and therefore it is slow to evaluate. Thus, we should use the parametrization that minimizes the number of times $\dot{\kappa}^{-1}$ is called.

Let $\boldsymbol{\eta} = X\boldsymbol{\beta}$, with $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_m)$. If we use the parametrization (9.8), we need to evaluate

$$D(\boldsymbol{y}, \boldsymbol{\mu}) = \sum_{i=1}^{m} w_i d(y_i, \mu_i) = \sum_{i=1}^{m} w_i d(y_i, g^{-1}(\eta_i)),$$

where g is the link function. As mentioned in Section 9.1, d is implemented as

$$d(y, \mu) = 2\left[y\{\dot{\kappa}^{-1}(y) - \dot{\kappa}^{-1}(\mu)\} - \kappa(\dot{\kappa}^{-1}(y)) + \kappa(\dot{\kappa}^{-1}(\mu)) \right],$$

which requires two calls to $\dot{\kappa}^{-1}$, one for y and one for μ. Therefore D requires $2m$ calls to $\dot{\kappa}^{-1}$.

The parametrization (9.9), requires to evaluate $E(\boldsymbol{y}, \boldsymbol{\theta}) = \sum_{i=1}^{m} w_i (y_i \theta_i - \kappa(\theta_i))$. From (4.13), we have that $\theta_i = \dot{\kappa}^{-1}(\theta_i)$ for each i. Thus, one call to $\dot{\kappa}^{-1}$ is required for each term in E which gives a total of m calls for E.

We should use then the canonical parametrization (9.9) for unifed GLMs. As an additional advantage, notice that if one chooses the link function to be $g = \dot{\kappa}^{-1}$, then $\theta_i = \eta_i$. In this case no calls to κ^{-1} are needed, and fitting is much faster. This is the *canonical* link for the unifed.

The `glm` function in R, wich gives the mle estimator of the regression coefficients, uses the mean-value parametrization. The current version (1.1.0) of the unifed package depends on this function for fitting a unifed GLM and therefore it does not provide the computational advantages of the canonical parametrization. Nevertheless, all the cases we have tried so far have only taken a few seconds to fit. If considered advantageous enough, future versions will include a fit function relying on the canonical parametrization.

For the Bayesian case, let $\pi(\boldsymbol{\beta})$ be a prior for the regression coefficients. The posterior can then be expressed as

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \propto \exp(E(\boldsymbol{y}, \boldsymbol{\theta}))\pi(\boldsymbol{\beta}).$$

The *unifed* R package comes with stan function called `unifed_glm_lp` which implements the above posterior. This implementation do profit from the computational advantages of the canonical parametrization.

9.6 Applied Example

To illustrate the use of the unifed GLM we use it to model the exposure of the vehicle insurance example introduced in Section 8.5.

We regroup two explanatory variables; the `area` classes `A`, `B`, `C`, `D` and `E` are merged into on class that we call `ABCDE`. Thus `area` can now take the values `ABCDE` or `F`. Similarly, classes `1` and `2` of `agecat` are merged together and are given the name `12`. The other classes are left intact.

We use `gender`, `veh_age` and the modified versions explained above of `agecat` and `area` as explanatory variables. After aggregating the observations using (4.8) we reduce the dataset to 80 observations. We first fit a frequentist GLM to this reduce dataset and comment on the model fit. We then fit a Bayesian GLM and find the entropic estimator of the betas.

9.6.1 Frequentist GLM

Table 9.2 (exported from R using the package `xtable` Dahl et al. (2018)) shows the summary provided by the `glm` function of R. We see that all the variables included are significant.

A deviance χ^2 test for goodness of fit is commonly used for GLMs. The null hypothesis is that the data is distributed according to the fitted GLM. Assuming the null hypothesis for this example implies that the residual deviance reported at the bottom of Table 9.2 follows a χ^2 distribution with 70 degrees of freedom. The p-value for this example is $\mathbb{P}(\chi^2_{70} \geq 92.506) = 0.0371$, which suggests to reject the model. Now, the detail with this test is that the χ^2 distribution for the residual deviance is asymptotic on the smallest weight of all GLM classes going to infinity (Jørgensen, 1992, Section 3.6). The smallest observed weight here

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3241	0.0133	-24.46	0.0000 ***
genderM	0.0288	0.0090	3.20	0.0014 **
agecat3	0.0525	0.0125	4.20	0.0000 ***
agecat4	0.0573	0.0124	4.61	0.0000 ***
agecat5	0.1036	0.0140	7.38	0.0000 ***
agecat6	0.0683	0.0167	4.10	0.0000 ***
areaF	0.0807	0.0200	4.05	0.0001 ***
veh_age2	0.1708	0.0138	12.40	0.0000 ***
veh_age3	0.1613	0.0133	12.17	0.0000 ***
veh_age4	0.1551	0.0134	11.56	0.0000 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for unifed family taken to be 1)

Null deviance: 376.034 on 79 degrees of freedom

Residual deviance: 92.506 on 70 degrees of freedom

Table 9.2: Summary of frequentist GLM

is 4 and it corresponds to the class with `gender`=F, `agecat`=6, `area`=F and `veh_age`=1. Therefore the χ^2 test for this example is not reliable and therefore we do not consider it evidence against the model.

Figure 9.6 shows the deviance residuals of this model. It suggests a good fit since they do not show any apparent pattern.

9.6.2 Bayesian GLM

Table 9.3 shows the information of the model. Table 9.4 shows the entropic coefficients and Figure 9.7 the response residuals. The p-values of the mean and variance of the residuals used as test quantities are 0.174 and 0.171, respectively.

9.7 The Beta Regression

The beta regression (Ferrari and Cribari-Neto (2004)) is a versatile model for applications with a response variable on the unit interval. Moreover, the well documented R package `betareg` (Cribari-Neto and Zeileis (2010)) makes it a practical tool in many applications.

The density of the beta distribution contains a large variety of shapes. In Ferrari and Cribari-Neto (2004) the beta density is reparameterized as

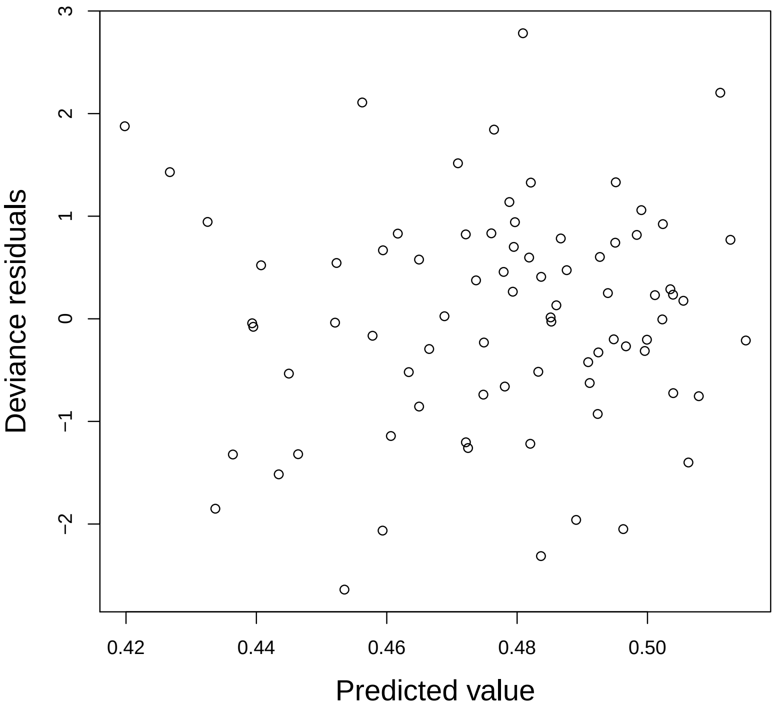


Figure 9.6: Residuals of Unifed GLM

$$f(y) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \qquad 0 < y < 1, \tag{9.10}$$

with $0 < \mu < 1$ and $\phi > 0$, and the distribution is denoted by $\mathcal{B}(\mu, \phi)$. Under this parametrization, if $Y \sim \mathcal{B}(\mu, \phi)$, the mean and variance are

$$\mathbb{E}[Y] = \mu \quad \text{and} \quad \mathbb{V}[Y] = \frac{\mu(1-\mu)}{1+\phi}. \tag{9.11}$$

Here ϕ is called the precision parameter of the distribution. In the beta regression model it is assumed that the response variable is a vector $Y = (Y_1, \ldots, Y_m)$, in which $Y_i \sim \mathcal{B}(\mu_i, \phi)$ for $i = 1, \ldots, m$. The $Y_i's$ are assumed independent to each other. The explanatory variables are incorporated to the model through the relation

$$g(\mu_i) = \boldsymbol{x_i}^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a vector of parameters and $\boldsymbol{x_i}$ is a vector of regresors. Here $g : (0, 1) \rightarrow \mathbb{R}$ is invertible and is called the link function.

Model information		MCMC information	
Response distribution	unifed	**No. of chains**	4
Weight variable	`numclaims`	**Warmup period**	$3,000$
Covariates	`agecat(1)`		
	`gender(F)`	**Simulations kept**	$47,000$
	`area(ABCDE)`	**(per chain)**	
	`veh_age(1)`		
Prior	betas are i.i.d. $N(0,20)$.		

Table 9.3: Exposure Bayesian model

Variable	Estimated Betas	Significance	Variable	Estimated Betas	Significance
(Intercept)	-0.9777	0	agecat6	0.2068	0
genderM	0.0867	0.001	areaF	0.2433	0
agecat3	0.1588	0	veh_age2	0.5164	0
agecat4	0.1736	0	veh_age3	0.488	0
agecat5	0.3125	0	veh_age4	0.4694	0

Table 9.4: Summary table of Bayesian exposure model estimated coefficients

Then Simas et al. (2010) generalized this model to allow the precision parameter to vary among classes in a similar way to the double generalized linear models (see Smyth and Verbyla (1999)). More specifically, in this case the response vector $Y = (Y_1, \dots, Y_m)$ is such that $Y_i \sim \mathcal{B}(\mu_i, \phi_i)$, independently and

$$g_1(\mu_i) = \boldsymbol{x_i}^T \boldsymbol{\beta},$$
$$g_2(\phi_i) = \boldsymbol{z_i}^T \boldsymbol{\gamma},$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are regression coefficients.

These regression models offer great flexibility when the response variable lies in the interval $(0,1)$, and both are implemented in the R package `betareg` (R Core Team (2019), Cribari-Neto and Zeileis (2010)).

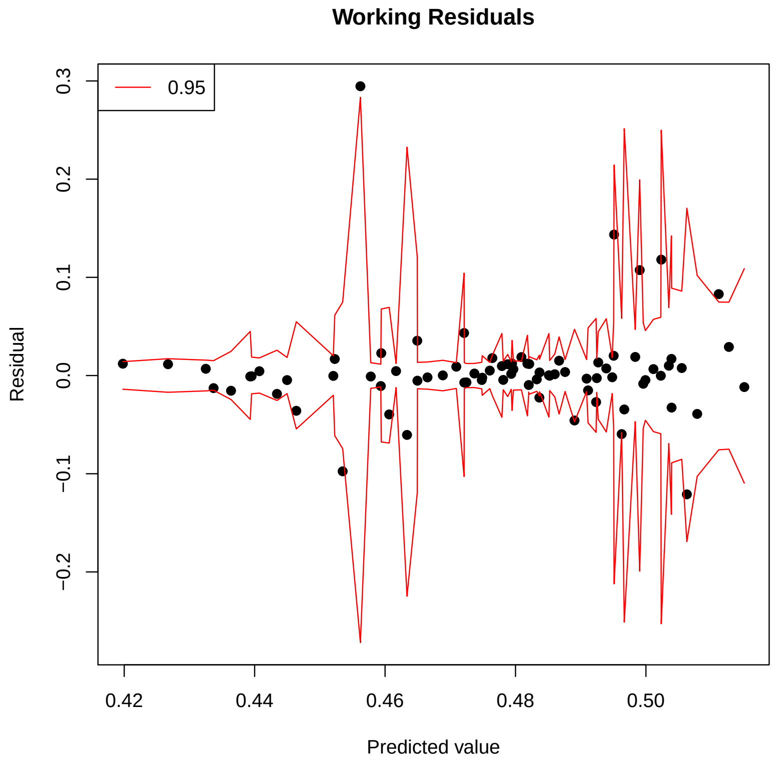


Figure 9.7: Residuals of Bayesian Unifed GLM

The beta distribution is not an exponential family and therefore the beta regression is not a GLM. Nevertheless this parametrization of the model chosen by the authors along with (9.11) give it a similar look and feel.

9.7.1 On the Difficulties of Data Aggregation for the Beta Regression

Data aggregation gives a practical advantage when working with large datasets. For GLMs this is straightforward due to two properties of $\bar{Y}$ in (4.8):

- $\bar{Y}$ is a sufficient statistic for μ
- The distribution of $\bar{Y}$ belongs to the same family than the Y_i's in (4.8).

We do not know any statistic with these two properties for the beta distribution. For instance, let $Y_1, \ldots, Y_n$ be an i.i.d sample from a $\mathcal{B}(\mu, \phi)$ distribution. The joint likelihood function of this sample is then

$$f(\boldsymbol{y}) = \left(\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}\right)^n \left(\prod_{i=1}^{n} y_i\right)^{\mu\phi-1} \left(\prod_{i=1}^{n}(1-y_i)\right)^{(1-\mu)\phi-1},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$. This density can be rearranged as follows

$$f(\boldsymbol{y}) = \left(\frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}\right)^n \left[\prod_{i=1}^{n} \frac{(1-y_i)^{\phi-1}}{y_i}\right] \left(\prod_{i=1}^{n} \frac{y_i}{1-y_i}\right)^{\mu\phi}$$

The factorization theorem (see (Hogg et al., 2005, Chapter 7)), implies that $T = \prod_{i=1}^{n} \frac{y_i}{1-y_i}$ is sufficient for μ. Now, the distribution of T, which is not beta, would be needed to use T for data aggregation. In other words, a regression model whose response distribution is a family that includes the distribution of T for every n would need to be developed.

9.7.2 Differences Between the Unifed GLM and the Beta Regression

The unifed density does not have the variety of shapes that the beta density has. To see this compare the shapes shown in Figure 9.1 with the shapes for the beta distribution shown in Ferrari and Cribari-Neto (2004). Thus, the beta regression adapts to more shapes than a unifed GLM and even more so if regresors are used for the dispersion parameter.

In those cases where a beta regression and a unifed GLM give similar good fit, the parsimony principle suggests to pick the unifed GLM, since it has one less parameter; the dispersion parameter is known for the unifed GLM.

From a numerical point of view, the unifed GLM has the advantage that it is possible to use (4.8) for data reduction. This is a practical advantage when dealing with large datasets specially if simulations of the response vector need to be performed.

Chapter 10

Risk Loadings for GLMs

10.1 Introduction

As mentioned in Chapter 1, insurers charge the pure premium plus a risk loading. Premium principles and risk measures are common tools for computing risk loadings (see Kaas et al. (2008)).

In this chapter we propose methods for computing the risk loading. We combine the Bayesian approach we have followed along this work with two widely used risk measures: the Value at Risk(Var) and the Conditional Tail Value at Risk(TVaR).

All our computations rely heavily on simulations. This allows us to incorporate uncertainty around the future number of clients and their durations. We first develop these ideas for a homogeneous portfolio and then we integrate them with GLMs in order to obtain risk premiums for heterogeneous data.

10.2 The Homogeneous Portfolio

10.2.1 Definitions and Notation

Consider a homogeneous portfolio or risk class of a portfolio over a fixed period (for example one year). We define the following notation:

- n - The number of risks in the portfolio, also called the size of the portfolio.

- S_i - the total loss of the i-th risk over the period, where $i \in \{1, \dots, n\}$.
- μ - The pure premium of each risk in the portfolio, i.e. $\mu = \mathbb{E}[S_1]$.
- ℓ - The individual premium risk loading.
- P - The individual loaded premium charged to each risk, i.e. $P = \mu + \ell$. Expenses and profit are not considered in this paper.
- L_n - Total loss of the portfolio over the period, i.e. $L_n = S_1 + \cdots + S_n$. The sub-index n makes reference to the size of the portfolio explicitly.
- $\bar{L}_n$ - The average loss of the portfolio over the period, i.e. $\bar{L}_n = \frac{L_n}{n}$.
- B_n - The balance of the portfolio, i.e. $B_n = nP - L_n = n(\mu + \ell) - L_n$.

We assume that the random variables $S_1, \dots, S_n$ are independent and identically distributed with $\mathbb{E}[S_1] = \mu < \infty$ and $\mathbb{V}[S_1] = \sigma^2$.

We say that the portfolio is viable (in the sense that it will be sufficient to cover liabilities) if $B_n > 0$ and we call $\mathrm{P}(B_n > 0)$ the probability of viability. B_n is the money that remains (or that is owed) after all premiums have been collected and all claims have been paid. Note that we do not consider the time of arrivals of premiums or of claims and therefore it is a simpler concept than the one studied in ruin theory.

We also assume that the same premium P is paid for all risks in this homogeneous portfolio, or risk class of a portfolio (henceforth called portfolio for simplicity).

10.2.2 Asymptotic Interpretation of Pure Premium and Risk Loading

In this section "asymptotic" is interpreted as the behaviour "when the number of risks grows to infinity".

Substantial effort is spent in estimating the pure premium of a portfolio. This is not the amount charged to a policyholder for a risk transfer, but it is used as a starting point to obtain the charged premium.

What makes $\mu = \mathbb{E}[S_1]$ such an important quantity for actuaries to dedicate so much effort to estimate it? The Strong Law of Large Numbers (SLLN) is often used as a justification for this. Now, what the SLLN tells us is that with probability one

$$\frac{S_1 + \cdots + S_n}{n} \to \mu \qquad \text{as} \qquad n \to \infty. \tag{10.1}$$

One could think that this implies that if the insurer charges the pure premium, then asymptotically it will break even. This would mean that if $\ell = 0$, then with probability one

$$B_n \to 0 \qquad \text{as} \qquad n \to \infty. \tag{10.2}$$

Now, not only (10.1) does not imply (10.2) but it actually implies something very different and much worse, as seen in the following result.

Proposition 10.2.1. *Let $\ell = 0$, $\sigma < \infty$ and $M > 0$ be an arbitrary positive value, then*

$$\lim_{n\to\infty} \mathrm{P}(|B_n| > M) = 1.$$

Proof. Let $\delta > 0$ be arbitrary, define $Z_n = \frac{\sqrt{n}}{\sigma}\left(\mu - \bar{L}_n\right)$ and note that $B_n = \sigma\sqrt{n}Z_n$ when $\ell = 0$. By the Central Limit Theorem,

$$\lim_{n\to\infty} \mathrm{P}(|Z_n| \leq \delta) = 1 - 2\Phi(-\delta), \tag{10.3}$$

where Φ is the cumulative distribution function of the standard normal distribution. There exists $N > 0$ such that for every $n \geq N$ we have $\frac{M}{\sigma\sqrt{n}} < \delta$ and

$$0 \leq \mathrm{P}(|B_n| \leq M) = \mathrm{P}\left(|Z_n| \leq \frac{M}{\sigma\sqrt{n}}\right) \leq \mathrm{P}(|Z_n| \leq \delta).$$

This implies that

$$0 \leq \limsup_n \mathrm{P}(|B_n| \leq M) \leq 1 - 2\Phi(-\delta).$$

Since the expression above is true for every $\delta > 0$, we have then that

$$\limsup_n \mathrm{P}(|B_n| \leq M) \leq \lim_{\delta\to 0} 1 - 2\Phi(-\delta) = 0$$
$$\therefore \limsup_n \mathrm{P}(|B_n| \leq M) = 0.$$

Using an analogous argument we also have that $\liminf_n \mathrm{P}(|B_n| \leq M) = 0$, and therefore $\lim_{n\to\infty} \mathrm{P}(|B_n| \leq M) = 0$. Which is equivalent to the conclusion of the proposition. □

Remark 10.1. *The stronger statement* $\mathrm{P}(\lim_{n\to\infty} |B_n| > M) = 1$ *is not true, since as* n *grows there will be oscillations between positive and negative values of* B_n*. Thus, there is no steady growth towards infinity.*

Since Proposition 10.2.1 holds for every $M > 0$, it can be interpreted as follows: when the individual premium loading $\ell = 0$, the larger the number of risks in the portfolio, n, the more certain the occurrence of a large discrepancy between the claims cost and the collected premiums. Now, what happens with the probability of viability as $n \to \infty$?

Proposition 10.2.2. *Suppose* $\sigma < \infty$*, for* $\ell = 0$*,*

$$\lim_{n\to\infty} \mathrm{P}(B_n > 0) = \frac{1}{2}.$$

Proof. Let Z_n be as in Proposition 10.2.1. As $\mathrm{P}(B_n > 0) = \mathrm{P}(\sigma\sqrt{n}Z_n > 0) = \mathrm{P}(Z_n > 0)$, then for any $n \geq 1$

$$\lim_{n\to\infty} \mathrm{P}(B_n > 0) = \lim_{n\to\infty} \mathrm{P}(Z_n > 0) = \Phi(0) = \frac{1}{2}.$$

□

Thus, if there is a large number of risks and the pure premium is charged, the portfolio will be viable with probability $\frac{1}{2}$. This is unacceptably low. Let us now introduce a risk loading in the analysis and see how this probability changes. Assume that the insureds pay $P = \mu + \ell$, where ℓ can be positive or negative. Then

$$\begin{aligned} \mathrm{P}(B_n > 0) &= \mathrm{P}(nP - L_n > 0) \\ &= \mathrm{P}\left(\frac{L_n}{n} - P < 0\right) \\ &= \mathrm{P}\left(\frac{\sqrt{n}}{\sigma}(\bar{L}_n - \mu) - \frac{\sqrt{n}}{\sigma}\ell < 0\right) \\ &= \mathrm{P}\left(Z_n \leq \frac{\sqrt{n}\ell}{\sigma}\right). \end{aligned}$$

In the expression above Z_n converges in distribution to a standard normal. If $\ell > 0$, then $\frac{\sqrt{n}\ell}{\sigma}$ goes to infinity as $n \to \infty$ and therefore $\mathrm{P}\left(Z_n \leq \frac{\sqrt{n}\ell}{\sigma}\right) \to 1$. On the other hand, when $\ell < 0$, then $\frac{\sqrt{n}\ell}{\sigma}$ goes to minus infinity an therefore $\mathrm{P}\left(Z_n \leq \frac{\sqrt{n}\ell}{\sigma}\right) \to 0$ as $n \to \infty$. In summary, we have derived the asymptotic probability of viability for all possible values of ℓ:

$$\lim_{n\to\infty} \mathrm{P}(B_n > 0) = \begin{cases} 1 & \text{if } \ell > 0 \\ \frac{1}{2} & \text{if } \ell = 0 \\ 0 & \text{if } \ell < 0 \end{cases} . \tag{10.4}$$

We can see that, in the asymptotic case, any positive loading guarantees viability, no matter how small it is.

For all results shown so far we have assumed finite variance. It turns out that for the cases $\ell < 0$ and $\ell > 0$ it is possible to drop the finite variance assumption and also have a stronger type of convergence.

Proposition 10.2.3. *With probability one,*

$$\lim_{n\to\infty} B_n = \begin{cases} \infty & \textit{if } \ell > 0 \\ -\infty & \textit{if } \ell < 0 \end{cases} \tag{10.5}$$

Proof. By the SLLN, with probability one there exists $N > 0$ such that for every $n \geq N$, $|\mu - \bar{L}_n| < \frac{\ell}{2}$. For such n, we have that $\ell + (\mu - \bar{L}_n) \geq \frac{\ell}{2}$, and therefore that

$$B_n = n\left[\ell + (\mu - \bar{L}_n)\right] \geq \frac{n\ell}{2}.$$

Taking limits on both sides when n goes to infinity, we get that

$$\lim_{n\to\infty} B_n = \infty,$$

with probability one. An analogous argument can be used for $\ell < 0$. □

10.2.3 The Finite Case

From the previous section we see that in the asymptotic case, determining the loading is not a problem; we can choose any positive number, no matter how small and that will do.

It is not the same for the finite case. When n is finite, the probability of viability is a function of ℓ and n. This is

$$\mathrm{P}(B_n > 0) = \mathrm{P}(n(\mu + \ell) - L_n > 0) = p(n, \ell),$$

for some function $p : \mathbb{N} \times \mathbb{R} \to [0,1]$. The expression above suggests a natural way for choosing the risk loading: Pick an $\alpha > 0$ and choose the smallest ℓ for which

$$\mathrm{P}(n(\mu + \ell) - L_n \geq 0) \geq \alpha. \tag{10.6}$$

This is equivalent to

$$\mathrm{P}\left(\frac{L_n}{n} \leq \mu + \ell\right) \geq \alpha, \tag{10.7}$$

which implies

$$\ell = (\bar{L}_n)_\alpha - \mu,$$

where $(\bar{L}_n)_\alpha$ is the α-th quantile of $\bar{L}_n$. This motivates the following definition.

Definition 10.1. *The* quantile risk loading *at level α is defined as*

$$\ell_q(\alpha) = (\bar{L}_n)_\alpha - \mu.$$

The quantile premium *(QP) is the loaded premium that corresponds to a quantile risk loading and is denoted with $P_q(\alpha)$, that is*

$$P_q(\alpha) = \mu + \ell_q(\alpha) = (\bar{L}_n)_\alpha. \tag{10.8}$$

Note that charging $P_q(\alpha)$ to each policyholder implies that the premiums collected are equal to the value at risk (VaR) (see Hardy (2006)) at confidence level α of the portfolio since

$$\mathrm{VaR}[L_n;\alpha] = n\mathrm{VaR}[\bar{L}_n;\alpha] = n(\bar{L}_n)_\alpha = nP(\alpha).$$

Remark 10.2. *Note that this idea can be transformed in a straightforward way to obtain a risk loading based on the loss ratio rather than on the aggregate loading B_n. Fix a target loss ratio δ and a probability α, then choose the smallest ℓ for which*

$$\mathrm{P}\left(\frac{L_n}{n(\mu + \ell)} \leq \delta\right) \geq \alpha,$$

which is given by

$$\ell = \frac{(\bar{L}_n)_\alpha}{\delta} - \mu.$$

A criticism of VaR is that it does not take into consideration how bad the $(1-\alpha)\%$ worst case scenarios can be. In our context this means that the quantile premium does not consider what the loss would be if the portfolio is not viable. The tail value at risk (TVaR) is a risk measure that was designed to address this issue. We call here the *tail quantile premium* the loaded premium each insured should pay so that the total collected premiums are equal to the TVaR.

Definition 10.2. *The* tail quantile premium *(TQP) is defined as*

$$P_t(\alpha) = \frac{1}{1-\alpha}\int_{\alpha}^{1} P_q(x)dx, \tag{10.9}$$

Similarly, the tail quantile risk loading *is defined as*

$$\ell_t(\alpha) = P_t(\alpha) - \mu. \tag{10.10}$$

An equivalent and easier to compute expression than (10.9) is

$$P_t(\alpha) = P_q(\alpha) + \frac{1}{1-\alpha}\mathbb{E}\left[\max\left(0, \bar{L}_n - P_q(\alpha)\right)\right].$$

A direct consequence of the expression above is that $P_t(\alpha) \geq P_q(\alpha)$ for every $\alpha > 0$.

An argument often used to choose TVaR over VaR is that TVaR is sub-additive (see Hardy (2006)). We think that when deciding on the loading for a specific portfolio, one should see the results provided by each method and interpret them. One should use this understanding along with other relevant factors (e.g. competitiveness of prices, risk aversion/appetite) for deciding on the final loading. Thus, in the examples given in this chapter we simply report the values of both methods.

10.2.4 Adding Uncertainty on the Portfolio Size

In the discussion from the previous section it was natural to assume that n, the size of the portfolio, is known. In practice this is never the case since risks enter and leave the portfolio as time goes on.

Suppose that n is a random variable. Notice that the discussion from Section 10.2.3 is still valid and we can use Definitions 10.1 and 10.2 to find the quantile and tail quantile premiums respectively. One needs only to draw samples from the assumed distribution of n, when generating simulations for $\bar{L}_n$.

We do not propose here a specific model for the number of clients. Insurers usually have have retention models and projections of sales that can be used to find sensible distributions for n.

Another element that was not incorporated in the discussion in the previous section is how long each risk stays in the portfolio. Two concepts related to the premium derive from the duration.

Definition 10.3. *The* earned premium *of a policyholder is defined as its duration multiplied by its annual premium, i.e. wP.*

Definition 10.4. *The* unearned premium *is defined as the annual premium minus the earned premium, i.e. $(1 - w)P$.*

It is a common practice to return the unearned premium to the insured after a cancellation. We propose two different ways to compute the loading depending on whether the unearned premium is returned or not.

Loading when the Unearned Premium is Returned

In this case only the earned premium is considered as collected, when analyzing the balance of a portfolio for a specific period of time.

For a specific year, let w_i be the duration of the i-th policy in the portfolio and let $w_+ = w_1 + \cdots + w_n$, where n is the size of the portfolio. In this context P denotes the annual premium. The balance at the end of the year is then defined as

$$B_n^* = w_+ P - L_n, \tag{10.11}$$

where L_n is the sum of the losses during the year. (10.11) implies that when considering durations the average loss of the portfolio should be defined as

$$\bar{L}_n^* = \frac{L_n}{w_+}.$$

Note that we are using the superscript * for the balance and the average losses when w_+ is used, instead of the number of premiums for the total collected premium and the denominator of the average losses.

With this modification the quantile and tail quantile definitions (Definitions 10.1 and 10.2) still work if $\bar{L}_n^*$ replaces $\bar{L}_n$ in (10.8) and (10.10). This is, the quantile and tail quantile premiums when the unearned premium is returned is defined as

$$P_q^*(\alpha) = (\bar{L}_n^*)_\alpha, \tag{10.12}$$

$$P_t^*(\alpha) = P_q^*(\alpha) + \frac{1}{1-\alpha}\mathbb{E}\left[\max\left(0, \bar{L}_n^* - P_q^*(\alpha)\right)\right]. \tag{10.13}$$

In the coming sections we show examples where we find the quantile and tail quantile premiums in which we consider the number of risks and their exposures as random.

Loading when the Unearned Premium is not Returned

If the unearned premium is not returned, we propose a method that lowers the premium charged to each insured. If one has a model for duration, one can use it in the analysis to price knowing that some insureds will not be covered for the entire time estipulated in order to charge less to everybody.

It can be argued that by doing this one would modify the behavior of the portfolio since insureds that cancel in order to get back the unearned premium would not cancel anymore. This is true, but it is also true that many cancellations are out of the control of the insured. For instance a home insurance policy is cancelled if the owner moves to a different place, or a car insurance policy is cancelled after the sale of the vehicle. Thus, there is a proportion of cancellations that would still occur even if the unearned premium is not returned.

Under this scheme, Definitions 10.1 and 10.2 of the quantile and tail quantile premiums are used without any modification. There is one detail that requires care; for the method to work a model is needed for the time at which policyholders cancel. Therefore, policy year has to be used to model the duration under this scheme.

Remark 10.3. *Either policy year or accident year can be used in a frequency or severity model. When the unearned premium is returned, either one can be used. By contrast, if the* *unearned premium is not returned, then to use the method described here, policy year must be used for the duration.*

10.3 Homogeneous Example

In this section we find the quantile and tail quantile premiums based on a simulated homogeneous dataset with 30 observations. It might seem artificial to do this with simulated data. The point is mainly to show how to perform the required simulations and to provide R and stan code to do it.

Duration	Claims	Loss 1	Loss 2	Duration	Claims	Loss 1	Loss 2
0.85	0	0.00	0.00	0.93	0	0.00	0.00
0.63	1	942.43	0.00	0.91	0	0.00	0.00
0.84	0	0.00	0.00	1.00	0	0.00	0.00
0.85	0	0.00	0.00	0.90	0	0.00	0.00
0.92	0	0.00	0.00	0.93	1	993.02	0.00
0.80	0	0.00	0.00	0.99	2	956.97	1086.41
0.91	0	0.00	0.00	0.61	0	0.00	0.00
0.90	1	1002.87	0.00	0.89	0	0.00	0.00
0.96	0	0.00	0.00	0.78	0	0.00	0.00
1.00	0	0.00	0.00	0.94	0	0.00	0.00
0.79	1	992.25	0.00	0.99	1	953.23	0.00
0.52	0	0.00	0.00	0.91	0	0.00	0.00
0.99	0	0.00	0.00	0.90	0	0.00	0.00
0.88	0	0.00	0.00	1.00	0	0.00	0.00
0.93	0	0.00	0.00	0.98	0	0.00	0.00

Table 10.1: Simulated Homogeneous Policy Year Losses

Table 10.1 shows the simulated losses in a policy year format. For each hypothetical risk, the data was generated as follows:

1. The duration was simulated using a unifed distribution with mean 0.9.

2. The number of accidents were simulated using a Poisson distribution with mean equal to 0.1 times the duration.

3. The cost of accidents was generated using a gamma distribution with mean $1,000$ and variance $1,500$.

Three models were fit to this dataset, one for the frequency, one for the severity and one for the exposure. In order to obtain simulations from the posterior distributions of the parameters of this model, stan was used with 4 parallel chains, a warmup period of 30,000 and the next 70,000 simulations were kept. The Hamiltonian Monte Carlo showed convergence for all the chains. We omit here the diagnostic plots.

10.3.1 Frequency Model

We assume that the annual number of accidents of any insured in the portfolio follows a Poisson distribution with mean λ. So, for a client with duration w, we assume that the claim counts follow a Poisson distribution with mean λw. We use a $(0, \infty)$ truncated normal distribution with mean 0 and standard deviation 10.

Mean: We use here the weighted mean defined as

$$\mu^* = \frac{\sum_{i=1}^N n_i}{w_+}, \qquad \text{where } w_+ = \sum_{i=1}^N w_i.$$

Variance: We used here a weighted version of the sample variance:

$$S_w^2 = \frac{1}{N}\sum_{i=1}^N (n_i - w_i\mu^*)^2.$$

Note that S_w^2 is not an unbiased estimator of the annual variance (i.e. the variance of the number of accidents in a period of one year). Nevertheless it is proportional to an unbiased estimator, and this is good enough for our testing purposes since the histogram and p-value for S_w^2 have the same shape and value, than those for the unbiased estimator.

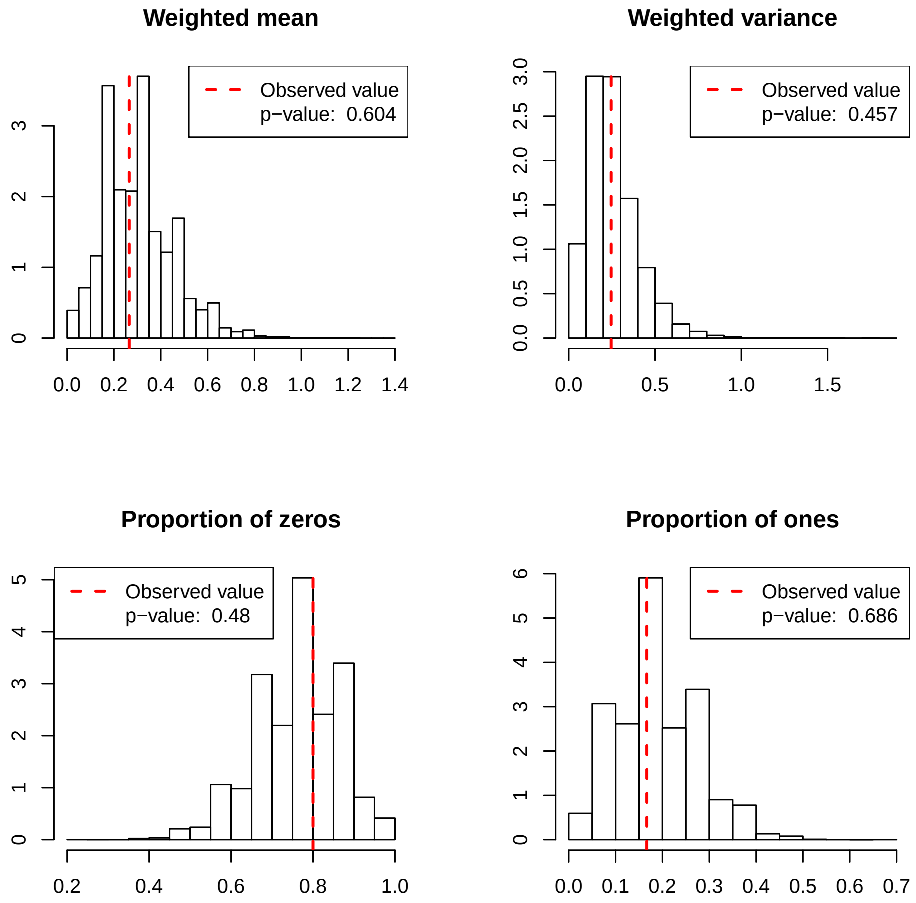


Figure 10.1: Test quantities for frequency model

Proportion of zeros: The observed proportion of zero accidents is usually high in car insurance. It is important to check that the model fits well the number of no accidents. We also use the proportion of ones (i.e. proportion of observations with one accident) as a test quantity in this example.

Figure 10.1 shows the histograms of the four test quantities mentioned above. They suggest a good fit of the model to the observed data.

10.3.2 Severity Model

Assume that the loss severity of an accident follows a gamma distribution. In order to be able to define sensible priors of the gamma mean and variance we scaled the observed losses,

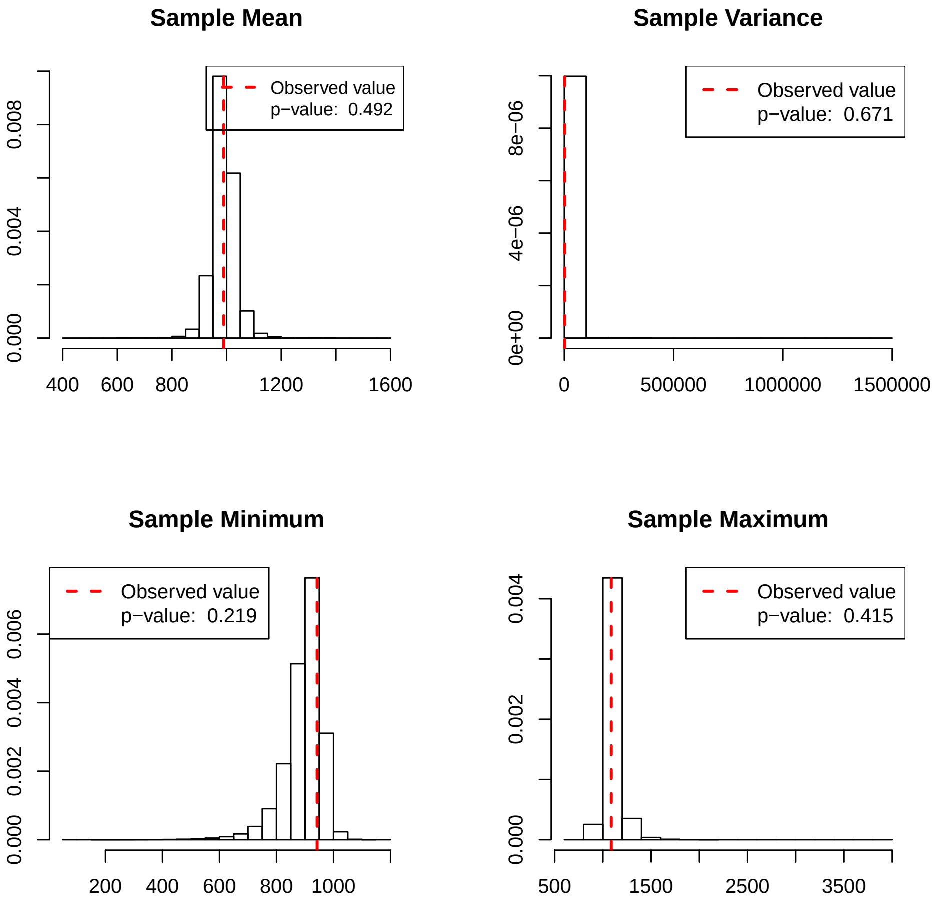


Figure 10.2: Test Quantities for the Severity Model

dividing them by 500. Note that the scaled losses also follow a gamma distribution as it is a scale family.

For the mean and variance of the scaled losses we use a normal prior with mean 0 and standard deviation 1, truncated to $(0, \infty)$.

We use here four test quantities: the mean, standard deviation, sample minimum and sample maximum. By contrast to the frequency model, our observations are considered i.i.d.. The histograms and p-values of the four quantities are shown in Figure 10.2. These graphs suggest a good fit, so we use this model for estimation.

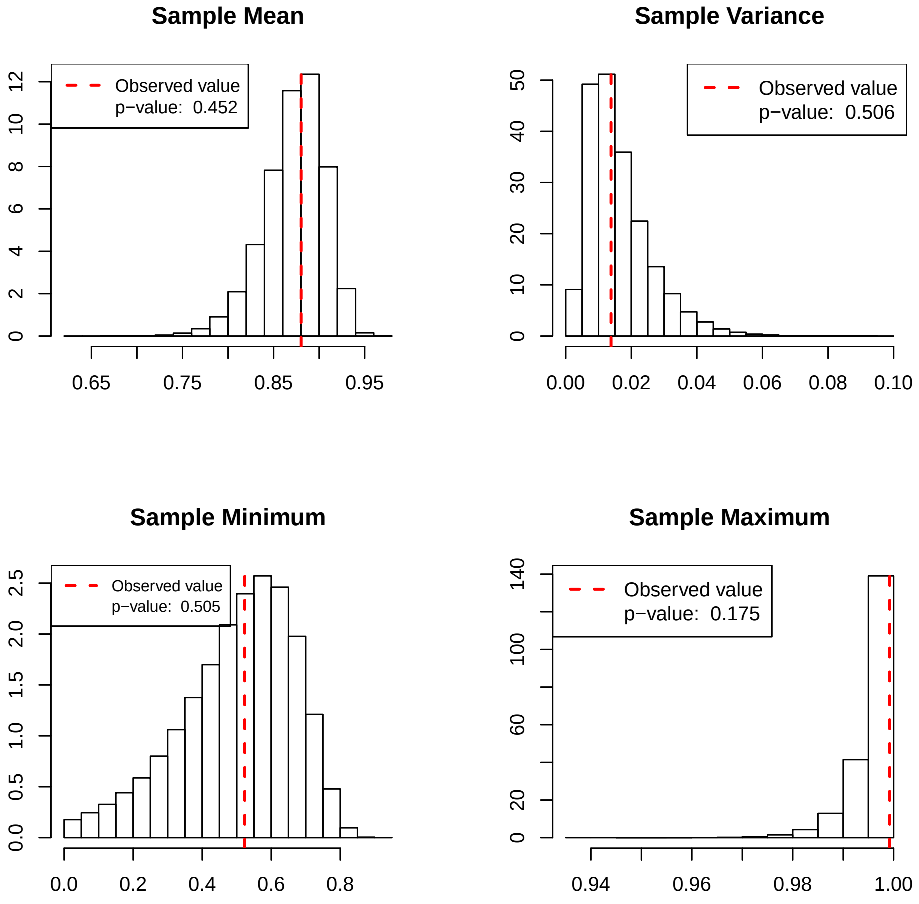


Figure 10.3: Test Quantities for the Exposure Model

10.3.3 Duration Model

Assuming durations that follow a unifed distribution, we use a truncated normal$(0, 1)$ prior with mean 0.5 and variance 10 on μ. Four test quantities are used: the sample mean, sample variance, sample minimum and sample maximum. The histograms of these test quantities and their respective p-values are given in Figure 10.3.

10.3.4 Premium Loading

In this example we assume that the unearned premium is returned to the clients after a cancellation. In order to show how the premium changes when the uncertainty about the number of risks and exposure is considered, the quantile and tail quantile loadings are computed here in three different ways:

Method 1 The number of risks for next year is assumed to be 30 and all the exposures are assumed to be 1.

Method 2 The number of risks is considered unknown and all exposures are assumed to be 1.

Method 3 The number of risks is considered unknown and the exposure assumed to follow the model from the previous section.

For simplicity, in Methods 2 and 3, it is assumed that the number of risks for the following year follows a negative binomial distribution with mean 30 and variance 39. In a real life situation one could use a renewal model along with sales projection in order to have a more accurate model.

Table 10.2 shows the quantile and tail quantile premiums for the three methods at $\alpha = 0.99$.

Method	Quantile Premium	Tail Quantile Premium
1	716.54	799.82
2	723.37	811.53
3	735.37	824.74

Table 10.2: Loaded Premiums

10.4 Heterogeneous case

This section combines the ideas from the previous section with GLMs to obtain loadings for an heterogeneous portfolio. We start by defining some notation to describe the behavior of the heterogeneous data. Note that some symbols introduced in Section 10.2.1 are repeated here, but have a different meaning in the heterogeneous case; context should make the meaning clear.

So far two different premiums, the quantile and tail quantile premium, were defined and two different schemes to use them: returning the unearned premium or not. This gives four different ways for computing the premium. Since we often need to refer to which way is being considered, we introduce the following notation:

$C(q,r)$ quantile premium with return of the unearned premium.

$C(t,r)$ tail quantile premium with return of the unearned premium.

$C(q,nr)$ quantile premium with no return of the unearned premium.

$C(t,nr)$ tail quantile premium with no return of the unearned premium.

The C stands for "case", the argument, q or t, stands for quantile or tail quantile, respectively and the second argument, r or nr, stands for "return the unearned premium" or "no return of the earned premium", respectively. Finally, to refer to more general cases we drop one argument; $C(q)$ to either $C(q,r)$ or $C(q,nr)$, and $C(nr)$ refers to either $C(q,nr)$ of $C(t,nr)$, while $C(*)$ is for any of the four cases.

Then the following notation describes the outcome of the whole portfolio. Let m be the number of classes. For $i = 1, \dots, m$ define

- m denotes the number of different classes defined by the explanatory variables of the GLM.
- L^i the total loss of the i-th class.
- $$w_i = \begin{cases} \text{Number of risks in the i-th class,} & \text{if } C(nr), \\ \text{Sum of all durations in the i-th class,} & \text{if } C(r). \end{cases}$$
- $\bar{L}^i = \frac{L^i}{w_i}$ is the average loss of the portfolio and $\bar{L}^i_\alpha$ the α-th quantile of $\bar{L}^i$. Note that the denominator is the number of risks for $C(nr)$ and the duration for $C(r)$.
- L the total portfolio loss, i.e. $L = \sum_{i=1}^m L^i$.
- $P_i(\alpha)$ is the premium charged for each risk in the i-th class, i.e.

$$P_i(\alpha) = \begin{cases} \bar{L}^i_\alpha, & \text{for } C(q), \\ \bar{L}^i_\alpha + \frac{1}{1-\alpha}\mathbb{E}\left[\max\left(0, \bar{L}^i - \bar{L}^i_\alpha\right)\right] & \text{for } C(t). \end{cases} \tag{10.14}$$

- For $C(*)$, $P(\alpha)$ is the total amount of premiums received for all classes. In other words

$$P(\alpha) = \sum_{i=1}^{m} w_i P_i(\alpha).$$

- Let L_α α-th quantile of L. We define the global portfolio premium at level α as

$$GP(\alpha) = \begin{cases} L_\alpha, & \text{for } C(q), \\ L_\alpha + \frac{1}{1-\alpha}\mathbb{E}\left[\max\left(0, L - L_\alpha\right)\right] & \text{for } C(t). \end{cases} \tag{10.15}$$

  In other words, $GP(\alpha)$ corresponds to VaR for $C(q)$, but for $C(t)$ it corresponds to TVaR at level α.

Section 10.2.3 presents two different methods to compute premiums. From the definitions above, note that $P(\alpha)$ is the premium obtained by first applying one of these methods to each class and then adding them all. By contrast, $GP(\alpha)$ corresponds to these methods applied directly to the whole portfolio.

10.4.1 Risk Loading with known Number of Risks and Exposures

When the number of risks and exposures are known, $P(\alpha)$ is a fixed number and since neither the quantile nor the tail quantile premiums are additive, we have that in general $P(\alpha) \neq GP(\alpha)$.

$GP(\alpha)$ is the total premium amount from the whole portfolio. The challenge here is to decide how much to charge to each risk on each class to get a total of $GP(\alpha)$ (premium allocation).

We propose to do this by finding first a number β that solves the following equation:

$$P(\beta) = GP(\alpha). \tag{10.16}$$

Note that a solution β always exists when the loss distribution is continuous. Once a β has been obtained, charge $P_i(\beta)$ to each risk on the i-th class.

To guarantee the numerical stability of the procedure to find a value of β that satisfies (10.16) we propose an algorithm for which the input is a target α and a tolerance level tol, while the output is a value β^* such that

$$GP(\alpha) \leq P(\beta^*) < GP(\alpha)(1 + tol). \tag{10.17}$$

Algorithm 1 Find β^* for known number of risks and exposures

```
Require: Initial value for $target.alpha$
Require: Initial value for $tol$
$min.bound \leftarrow 0$
$max.bound \leftarrow 1$
$p.beta \leftarrow 0$
while $p.beta < GP(\alpha)$ OR $p.beta \geq GP(\alpha)(1 + tol)$ do
    $\beta \leftarrow (min.bound + max.bound)/2$
    for $i$ in $1 : m$ do
        $P_i(\beta) \leftarrow$ premium at level $\beta$ for the $i$-th class.
    end for
    $p.beta \leftarrow \sum_{i=1}^{m} w_i P_i(\beta)$
    if $p.beta < global.premium$ then
        $min.bound \leftarrow \beta$
    else
        $max.bound \leftarrow \beta$
    end if
end while
return $\beta$, $P_i(\beta)$ for each $i \in \{1, \cdots, m\}$, $p.beta$
```

Algorithm 1 relies on the fact that $P(\beta)$ is monotonically increasing to find β^*.

10.4.2 Risk Loading with Random Number of Risks and Exposures

This case is more complicated than the previous section since there does not exist a β that satisfies (10.16). This is because $P(\beta)$ is a random variable that depends on the number of risks in each class and their respective durations.

For $C(q)$, we propose to find the smallest value of β such that

$$\mathrm{P}(L \leq P(\beta)) \geq \alpha. \tag{10.18}$$

For $C(t)$ it is more complicated. In order to define an intuitive generalization of (10.16) for $C(t)$, we develop first an equivalent formulation of (10.15). The Conditional Tail Expectation (CTE) of the loss L at level α is defined as

$$\mathrm{CTE}(L,\alpha) = \mathbb{E}[L|L > L_\alpha]. \tag{10.19}$$

Denote with F_L the cumulative distribution function of L. For the same value of α (see Chapter 5 of Kaas et al. (2008)) the CTE is greater or equal than the TVaR, and they coincide when $\alpha = F_L(L_\alpha)$. L is a continuous random variable except at 0. Thus, if $\alpha_0 = P(L = 0)$, the CTE and TVaR of L coincide for every $\alpha > \alpha_0$. In other words, for every $\alpha > \alpha_0$,

$$L_\alpha + \frac{1}{1-\alpha}\mathbb{E}\left[\max\left(0, L - L_\alpha\right)\right] = \mathbb{E}[L|L > L_\alpha].$$

Since it is unlikely for the portfolio to not have any losses, α_0 is close to zero. In contrast, we want a global level α close to 1 for the portfolio. It is then safe to give the following equivalent formulation of (10.15)

$$GP(\alpha) = \begin{cases} L_\alpha, & \text{for } C(q), \\ \mathbb{E}[L|L > L_\alpha] & \text{for } C(t). \end{cases} \tag{10.20}$$

With this reformulation of $GP(\alpha)$, we can now give a sensible generalization of (10.16) for the tail quantile premium. For $C(t)$, we propose to find the smallest value of β for which

$$\mathbb{E}[P(\beta) - L \mid L > L_\alpha] \geq 0. \tag{10.21}$$

Note that (10.18) and (10.21) are equivalent to (10.16) when the number of risks and the exposures are fixed.

For the numerical stability of the algorithm to find such a β, it is necessary to consider a level of tolerance to deviations from the target value. Since $GP(\alpha)$ is a fixed number and $P(\beta)$ is a random variable, the tolerance around $GP(\alpha)$ cannot be defined as in the previous section. Here the tolerance tol is defined around level α. For $C(q)$ we give an algorithm that finds a value β^* such that

$$\alpha \leq \mathrm{P}(L \leq P(\beta^*)) < \alpha + tol. \tag{10.22}$$

For $C(t)$, note that (10.21) is equivalent to

$$\mathbb{E}\left[L \mid L > L_\alpha\right] \leq \mathbb{E}\left[P(\beta) \mid L > L_\alpha\right].$$

Thus, we propose an algorithm that finds a β^* such that

$$\mathbb{E}\left[L \mid L > L_\alpha\right] \leq \mathbb{E}\left[P(\beta^*) \mid L > L_\alpha\right] \leq \mathbb{E}\left[L \mid L > L_{\alpha+tol}\right]. \tag{10.23}$$

In order to write a unique algorithm that include $C(q)$ and $C(t)$, we define the following three functions given a target α and a tolerance level tol:

$$stop(\beta) = \begin{cases} \alpha \leq \mathrm{P}(L \leq P(\beta)) < \alpha + tol, & \text{for } C(q), \\ \\ \mathbb{E}\big[L \mid L > L_\alpha\big] \leq \mathbb{E}\big[P(\beta) \mid L > L_\alpha\big] < \mathbb{E}\big[L \mid L > L_{\alpha+tol}\big], & \text{for } C(t), \end{cases}$$

$$smaller(\beta) = \begin{cases} \mathrm{P}(L \leq P(\beta)) \geq \alpha + tol, & \text{for } C(q), \\ \\ \mathbb{E}\big[P(\beta) \mid L > L_\alpha\big] \geq \mathbb{E}\big[L \mid L > L_{\alpha+tol}\big], & \text{for } C(t), \end{cases}$$

$$bigger(\beta) = \begin{cases} \mathrm{P}(L \leq P(\beta)) < \alpha, & \text{for } C(q), \\ \\ \mathbb{E}\big[P(\beta) \mid L > L_\alpha\big] < \mathbb{E}\big[L \mid L > L_\alpha\big], & \text{for } C(t), \end{cases}$$

There three functions return either `TRUE` or `FALSE`. They are used in Algorithm 2 for changing the behavior at each iteration. The function *stop* is used to decide whether the algorithm can stop or not. When *smaller* returns `TRUE` it indicates that the value of β should be smaller at the next iteration. Similarly, when *bigger* returns `TRUE`, β should be bigger at the next iteration.

Note the ambiguity here on what is meant with class. This is because we have several models, i.e. frequency, severity duration and number of risks, and each has its own classes. Here with class we mean a specific combination of values of the explanatory variables for all the models. In order to obtain the necessary simulations in this setting, what is first needed is what is called the *prediction table*.

Algorithm 2 Find β^* for random number of risks and exposures

```
Require: Initial value for $target.alpha$
Require: Initial value for $tol$
$min.bound \leftarrow 0$
$max.bound \leftarrow 1$
while TRUE do
  $\beta \leftarrow (min.bound + max.bound)/2$
  for $i$ in $1 : m$ do
    $P_i(\beta) \leftarrow$ premium at level $\beta$ for the $i$-th class.
  end for
  if $stop(beta)$ then
    BREAK WHILE
  else
    if $smaller(beta)$ then
      $max.bound \leftarrow \beta$
    else
      $min.bound \leftarrow \beta$
    end if
  end if
end while
return $\beta$, $P_i(\beta)$ for each $i \in \{1, \cdots, m\}$
```

Algorithm 1 relies on the fact that $P(\beta)$ is monotonic to find such a β^*. It has two initial variables: $target.alpha$, which corresponds to α in (10.22) and tol.

Definition 10.5. *A* prediction table *contains the modelling data aggregated with respect to all the explanatory variables of all the models. If the same variable is used in more than one model with different re-groupings, there must be a column for each different regrouping.*

Note that each line of the prediction table corresponds to a different class. For each model we simulate the GLM coefficients. They can then be used to obtain simulations of the mean or canonical parameter and the dispersion parameter of each line of the prediction table.

These in turn can be used to simulate the response of each model.

10.5 Car Example

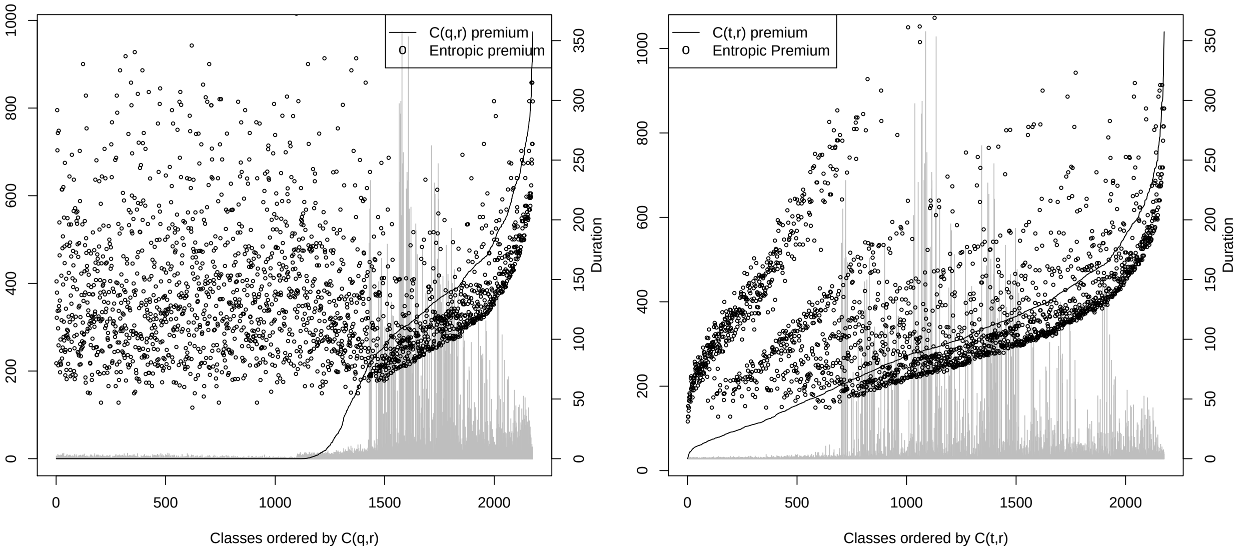
The loaded premiums can now be obtained for the observations contained in a car insurance dataset. We use the same frequency and severity models we got in Section 8.5 and the same duration model than in Section 9.6.2. So the same explanatory variables and priors were used, but the simulations were rerun using a stan program with all the models programmed. For each class, it is assumed that the number of insureds for the following year follows a negative binomial distribution with mean equal to the observed number of insureds and a standard deviation equal to the mean divided by two. The stan code file for the illustration used here is `src/stan_files/loading_heterogeneous_example.stan` in the package `mythesis` (Quijano Xacur (2019a)).

An important detail about this example is that the data seems to use the accident year format. This is not mentioned explicitly on the website where data is published, but the shape of the exposures histogram seems to indicate it. Policy year histograms usually have a much taller bar at 1, as most insureds renew their policy. We still compute the premiums using the provided durations, but in practice one should make sure to use policy year durations for $C(nr)$.

After creating the prediction table (see Definition 10.5) $2,176$ classes were obtained. We computed different premiums to all classe with a target α of 0.99 was used for all.

Let us focus on $C(q,r)$ and $C(t,r)$. Their obtained values of β^* were 0.735 and 0.189, respectively.

Figure 10.4 shows graphs comparing them to the entropic premium for each class. In both plots the right y-axis corresponds to the total duration of a class and the left y-axis to premium values. The total duration of each class is shown with the gray vertical bars. The x-axis correspond to the different classes. In Figure 10.5a the classes are sorted in increasing order of the $C(q,r)$ premiums. Similarly, in Figure 10.5b, the classes are ordered in increasing order of the $C(t,r)$ premiums. The entropic premium of each class is shown with a small circle.



(a) $C(q,r)$ and entropic premius

(b) $C(t,r)$ and entropic premius

Figure 10.4: Premiums that return the unearned premium

The first thing that jumps to the eye when seeing Figure 10.4 is that there are more than $1,000$ classes for which the quantile premium is zero. Another important thing to notice, and that is a common denominator in both Figures 10.5a and 10.5b is that for the classes with the lowest premiums, the entropic premium is larger than the loaded premium. Note that the premiums we found satisfy our optimality criteria (see (10.18) and (10.21)) and therefore we expect these premiums to make the portfolio viable. Then, it seems that the classes with higher premiums are subsidizing the classes with lower premiums at very high levels. This is undesirable since we want the premium to reflect the risk properly. The next section explains why this happens and proposes a solution, which is implemented to this example in Section 10.7.

10.6 Managing the Subsidy from High Premium Classes Towards Low Premium Classes

We have created an interactive app for illustrating the arguments of this section. The app can be started locally by calling the function `run.premium.comparissons.app` in the R package `mythesis` accompanying this thesis. It is also available on the website `https://oquijano.shinyapps.io/premiums_visualization/` where it can be used without installing anything in the local computer.

The app assumes that there is a homogeneous portfolio with uncertainty about the number of risks and durations. It is assumed that the number of risks follows a negative binomial distribution. For each risk, the duration follows a unifed distribution. The number of accidents follows a Poisson distribution whose parameter is a constant value λ times the duration. Finally, the cost of each claim is supposed to follow a gamma distribution. The left panel of the app has controls to set and change the parameters of all these distributions. It also has a field that allows to change the number of simulations used to compute the results. Finally, it has two controls for setting the maximum y value shown in the two graphs present in the app.

The one on the top is about premiums for which the unearned premium is returned and the one below when it is not. The x-axis of both graphs represents different premium levels. Each graph has three lines. In the upper graph the lines correspond to the $C(q,r)$ and $C(t,r)$ premiums and the annual pure premium. Here annual pure premium means the expected value of the loss of a single risk assuming it's duration will be one. The three lines in the lower graph correspond to the $C(q,nr)$ and $C(t,nr)$ premiums and the pure premium. Here with pure premium we mean the expectation of the loss of a single risk considering the randomness of the duration.

Set the mean of the number of insureds to a low value, say between 1 and 5. Note that the quantile premium is zero for most quantiles and it gets more extreme if the mean of the duration is decreased. This happens because when the total duration of the portfolio is small, the probability of no accidents is very high. Thus, the quantile of the loss will be zero for all values less or equal to that probability. This explains why so many classes have a $C(q,r)$

premium of zeros. In Figure 10.5a the gray lines show that the duration is very low for the classes with zero premium. In fact, 30% of the classes have an expected total duration of less than one and 64% less than five.

Set now the mean of the number of risks to a high value. Note that all tail quantile premiums are higher than the entropic premium for all quantiles. Now start lowering that mean to small values and note that at some point you start getting $C(t,r)$ premiums smaller than the quantile premium for the lower quantiles. Since the β^* is 0.189 and there are many classes with very low expected duration this explains why there are many classes for which the $C(t,r)$ premium is lower than the entropic premium.

Thus, the classes that are paying less than the entropic premium are the ones with small duration. This implies that in the end the classes with large duration are subsidizing the classes with small duration.

A solution to this problem is to divide the risk classes into different categories according to their duration and compute the β^* for each category separately. The function `compute.loadings.heterogeneous.example` of the R package `mythesis`, can receive a vector representing interval boundaries for the expected class durations which defines categories for the different risk classes. The β^* for each category is computed along with the entropic premiums for each category.

10.6.1 On Why not Using the CTE for Each Risk Class

In Section 10.4.2 we used the CTE to define the level of a global optimality criterion in $C(t)$. For the total loss of the portfolio this turned out to be equivalent to the TVaR. In a similar way, one might consider to use the CTE instead of the TVaR for the premium of each risk class. This is, imagine we change $P_i(\alpha)$ in (10.14) for

$$P_i(\alpha) = \begin{cases} \bar{L}^i_\alpha, & \text{for } C(q), \\ \mathbb{E}[\bar{L}^i | \bar{L}^i > \bar{L}^i_\alpha] & \text{for } C(t). \end{cases} \tag{10.24}$$

It turns out this change can make Algorithms 1 and 2 not converge in certain situations. In the following paragraphs we explain how this happens.

We showed already how grouping the risk classes according to their total expected duration can solve the problem of classes with higher exposure subsidizing classes with small exposure. Consider a category of classes with small exposure. Let m_0 be the number of classes in that category and let $i_1, \dots, i_{m_0}$ be the indices of those categories among the original m classes. For each k in $1, \dots, m_0$, define $\alpha_0^{i_k}$

$$\alpha_0^{i_k} = \mathrm{P}(\bar{L}^{i_k} = 0).$$

Then, for any such k and $\beta < \alpha_0^{i_k}$,

$$\mathbb{E}[\bar{L}^i | \bar{L}^i > \bar{L}_\beta] = \mathbb{E}[\bar{L}^i | \bar{L}^i > \bar{L}_{\alpha_0^{i_k}}].$$

In other words, for $C(t)$ we have that for every $\beta < \alpha_0^{i_k}$ $P_{i_k}(\beta) = P_{i_k}(\alpha_0^{i_k})$. Denote with $P^{cat}(\beta)$ and L^{cat} the total collected premium at level β and the total losses in this category, respectively. This is $P^{cat}(\beta) = \sum_{k=1}^{m_0} P_{i_k}(\beta)$ and $L^{cat} = \sum_{k=1}^{m_0} L^{i_k}$. Define $\alpha_0 := \min(\alpha_0^{i_1}, \dots, \alpha_0^{i_{m_0}})$, and notice that for every $\beta < \alpha_0$,

$$P^{cat}(\beta) = P^{cat}(\alpha_0).$$

Suppose that for some target α you have that

$$\mathbb{E}\left[P^{cat}(\alpha_0) \mid L^{cat} > L^{cat}_\alpha\right] > \mathbb{E}\left[L^{cat} \mid L > L^{cat}_{\alpha+tol}\right], \tag{10.25}$$

and notice that this implies that for every $\beta < \alpha_0$,

$$\mathbb{E}\left[P^{cat}(\beta) \mid L^{cat} > L^{cat}_\alpha\right] > \mathbb{E}\left[L^{cat} \mid L > L^{cat}_{\alpha+tol}\right].$$

In this case, no matter how small β gets, it will be impossible to satisfy (10.23) and the algorithm would not converge. Thus, if one used the CTE for the premium of each risk class, the minimum possible premium is $P(\alpha_0)$. If this value is greater than the upper bound of our stop criterion (10.23), then algorithm for finding the level of each individual class would not converge. For this reason we recommend to use TVaR for the premiums of each homogeneous risk class in the GLM.

10.7 Car Example: Second Attempt

We defined categories according to the expected duration for the classes of the example in Section 10.5. The intervals used for defining the classes are $(0, 1]$, $(1, 2]$, $(2, 5]$, $(5, \infty)$. Table 10.3 shows the β^* obtained for each category and each type of premium.

	cqr	ctr	cqnr	ctnr
(0,1]	0.97	0.82	0.97	0.82
(1,2]	0.94	0.65	0.93	0.65
(2,5]	0.87	0.40	0.87	0.40
(5,Inf]	0.68	0.14	0.68	0.15

Table 10.3: β^* by premium type and category

Figure 10.5 shows the premiums for all classes for each type of premium. $C(r)$ premiums are compared to the entropic estimator of the annual pure premium. $C(nr)$ premiums are compared to the entropic estimator of the pure premium including uncertainty about the duration.

The first thing to notice is that we do not have zero as the quantile premiums for many classes as before. In the new premiums there is one class with zero $C(q, r)$ premium and also one class with zero $C(q, nr)$ premium. One can also notice that the $C(q)$ premiums are still subsidizing many classes. We mean this in the sense that the loaded premium is less than the estimated pure premium. For the $C(q, r)$ premiums there are 824 subsidized classes containing $6,306$ risks which represent 9.29% of all risks in the porfolio. For $C(q, nr)$ premiums there are 831 subsidized classes containing $6,249$ risks which represent 9.21% of all risks in the porfolio.

For the $C(t)$ premiums we see that there are very few subsidised classes and for those that are, there loaded premium is close to the entropic premium. For $C(t, r)$ premiums there are 3 subsidized classes containing 7 risks which represent 0.0103% percent of he portfolio. For $C(t, nr)$ premiums there are 0 subsidized classes containing 0 risks which represent 0% percent of the portfolio.

Let us now make a comparison between the $C(r, t)$ premium the expected value principle (see Chapter 5 of Kaas et al. (2008)) for risk loading. We take the premium to be 20% more than the annual entropic premium. Figure 10.6 compares $C(t, r)$ premiums with the expected mean premium principle.

Finally, suppose that each insured in the dataset had paid the premium that has been estimated here, for each of the four cases. Table 10.4 reports the total loss incurred, along

	Total Portfolio Amount
Loss	9,314,604.443
$C(q, r)$ premium	$11,287,972.96$
$C(q, nr)$ premium	$11,306,463.36$
$C(t, r)$ premium	$11,328,213.78$
$C(t, nr)$ premium	$11,349,109.44$
Expected mean principle	$12,044,134.71$

Table 10.4: Total Collected Premiums

with the total premium that would have been collected in each case.

10.8 Recommendations for Further Improvements

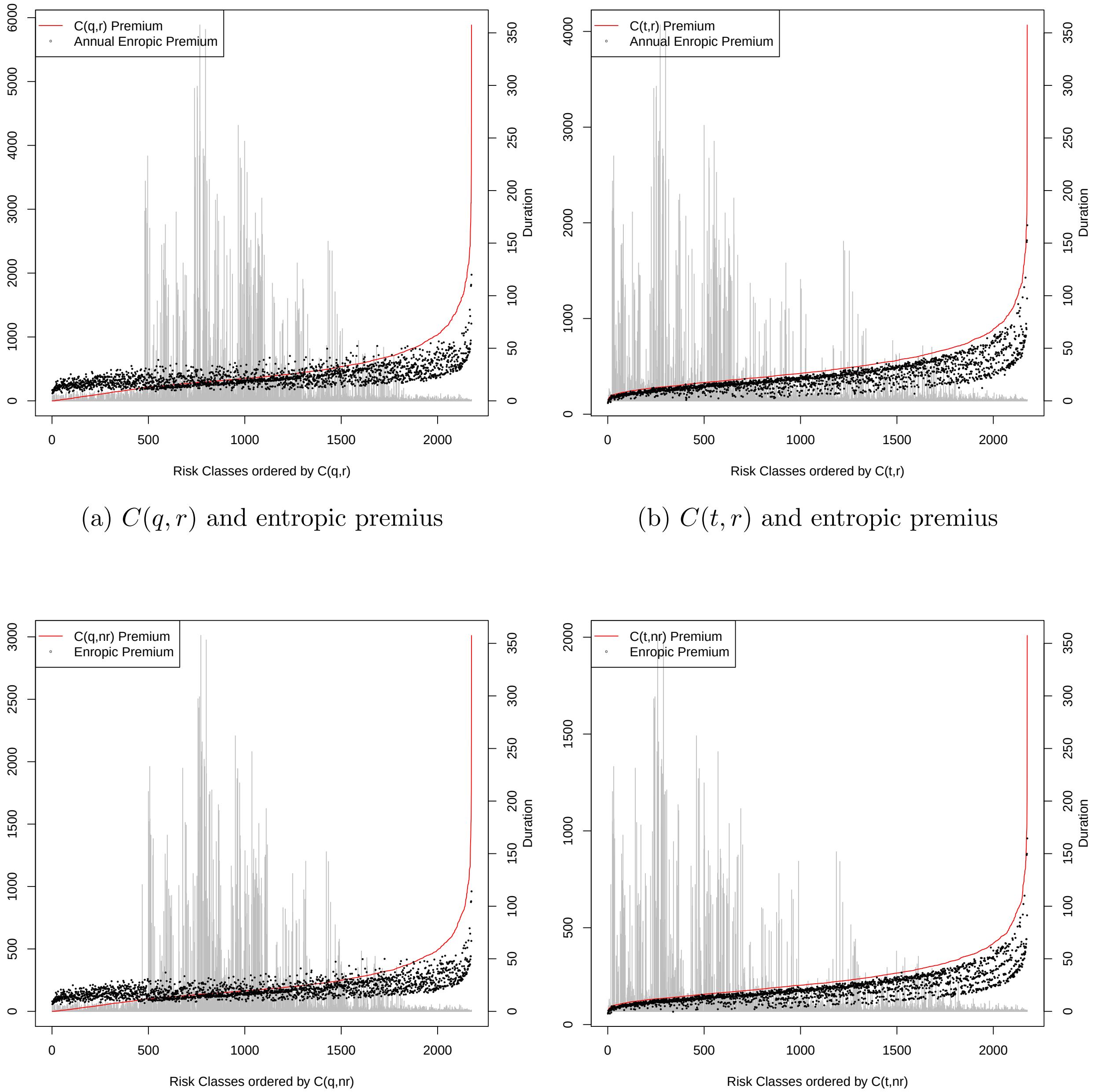
We would like to comment on possible ways to improve the results from the previous section. We dot now show their effects on our current example, but we think they could be useful in other examples in practice.

One might be interested in getting rid of all the subsidized classes or at least to have less of them. For this purpose one could play with the intervals used to categorize the risk classes according to their expected duration. If this does not give the desired, one can use the premium at the desired level (in the example 0.99) and apply the procedures from Sections 10.4 and 10.6 to the rest of the classes.

There is something to notice with the duration model. The average durations of the risk classes in the prediction table are very different from the durations in the data aggregated for estimating the parameters of the duration model. This makes that for many classes, the estimated expected duration is very far from the observed value in the prediction table. Table 10.5 shows summary statistics of the values of average durations in the prediction table and the expected class durations. In similar situations one might try to look for additional explanatory variables for improving the duration model. In this particular example adding all the available covariates does not solve the problem.

	Minimum	25th quantile	Median	Mean	75th quantile	Maximum
Observed	0.00274	0.382	0.466	0.467	0.549	0.999
Predicted	0.42	0.463	0.476	0.473	0.485	0.515

Table 10.5: Comparison between observed and predicted class durations



(a) $C(q, r)$ and entropic premius

(b) $C(t, r)$ and entropic premius

(c) $C(q, nr)$ and entropic premius

(d) $C(t, nr)$ and entropic premius

Figure 10.5: Premiums for classes categorized by expected duration.

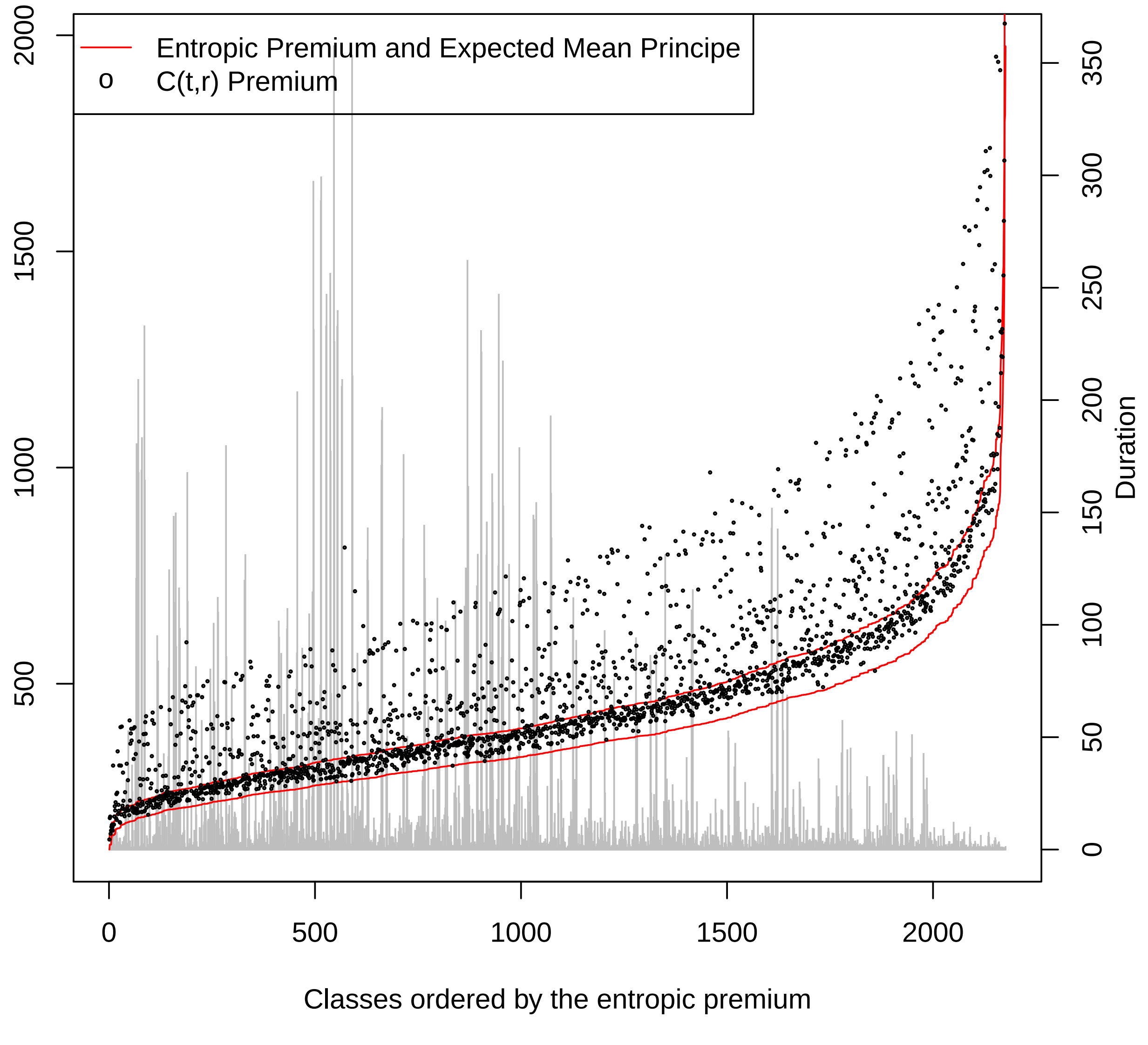


Figure 10.6: Comparison between the $C(t, r)$ premium, the entropic premium and the expected mean premium principle.

Summary and Ideas for Future Development

This thesis uses a Bayesian approach to develop a credibility estimator and a method to compute premium risk loadings. Both methods can be used with GLMs.

We called our credibility estimator the *entropic premium*. It is a Bayesian point estimator that uses the relative entropy as loss function. Unlike the posterior mean, the entropic premium is invariant. This is, it is consistent among different parametrizations of the model. For univariate EDF's, if a conjugate prior is used and the dispersion parameter is considered known, the entropic premium coincides with Jewell's linear credibility estimator. An algorithm for finding the entropic estimator for GLMs is provided and we give an applied example to show it's feasibility in practical situations. The development of the entropic estimator is preceded by a proposition that shows that exact linear credibility for GLMs is impossible.

We use the VaR and TVaR on the entire insurance portfolio to compute risk loadings. Our method considers the number of insureds and their durations as random variables. We introduced a new distribution called *unifed* for modeling the duration of each risk. The advantage of the unifed over other existing distributions with support on $(0, 1)$, like the beta distribution, is that it allows to aggregate data without loss of information. This can significantly reduce the dimension of the problem at hand and thus simplify the number of required computations. We consider two types of premium schemes for our method. One in which the unearned premium is returned to the policyholder and another one in which it is not returned.An algorithm for applying this method for obtaining the premium risk loadings for the classes of a GLM is provided.

In the Chapter 8 we mention that our approach has the advantage that one can use a prior that reflects the out-of-sample information rather than one that gives a linear formula. Nevertheless, we use weakly informative priors for all our examples. A way to find sensible priors based on out-of-sample information would be a good complement to this work.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory*, Budapest, pp. 267–281. Akadémiai Kiado.

Antonio, K. and J. Beirlant (2007). Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics 40*(1), 58 – 76.

Athreya, K. and S. Lahiri (2006). *Measure Theory and Probability Theory.* Springer-Verlag New York.

Bernardo, J. M. (1979, 01). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B 41.*

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

Bernardo, J. M. (2005b). Reference analysis. In *In Handbook of Statistics 25.* Elsevier.

Betancourt, M. (2017, January). How the shape of a weakly informative prior affects inferences. `https://mc-stan.org/users/documentation/case-studies/weakly_informative_shapes.html`.

Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin 4*(3), 99–207.

Bühlmann, H. and E. Straub (1970). Glaubwürdigkeit für schadensätze. *Bulletin of the Swiss Association of Actuaries*, 111–133.

Cribari-Neto, F. and A. Zeileis (2010). Beta regression in R. *Journal of Statistical Software 34*(2), 1–24.

Dahl, D. B., D. Scott, C. Roosen, A. Magnusson, and J. Swinton (2018). *xtable: Export Tables to LaTeX or HTML.* R package version 1.8-3.

Daniels, H. E. (1980). Exact saddlepoint approximations. *Biometrika 67*(1), 59–63.

de Finetti, B. (1939). La teoria del rischio e il problema della rovina dei giocatori. *Giornale dell'istituto Italiano degli Attuari 10*, 41–51.

de Jong, P. and G. Z. Heller (2008). *Generalized Linear Models for Insurance Data.* Cambridge University Press. Cambridge Books Online.

de Moivre, A. (1756). *The doctrine of chances: or, A method of calculating the probabilities of events in play.* London: printed for A. Millar.

De Vylder, F. (1985). Non-linear regression in credibility theory. *Insurance: Mathematics and Economics 4*(3), 163–172.

De Vylder, F. (1996). *Advanced Risk Theory — a Self Contained Introduction.* Editions de L'Université de Bruxelles.

Diaconis, P. and D. Ylvisaker (1979). Conjugate priors for exponential families. *The Annals of Statistics 7*(2), 269–281.

Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics 31*(7), 799–815.

Garrido, J. and J. Zhou (2009). Full credibility with generalized linear and mixed models. *ASTIN Bulletin 39*, 61–80.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis.* Chapman and Hall/CRC.

Gómez Déniz, E. (2006). On the use of the weighted balanced loss function to obtain credibility premiums. In *International Conference on Mathematical and Statistical Modeling in Honor of Enrique Castillo*, pp. 1–12.

Góra, P. and A. Boyarsky (1997). *Laws of Chaos.* Birkhäuser Boston.

Hachemeister, C. A. (1975). Credibility for regression models with application to trend. *Proc. of the Berkeley Actuarial Research Conference on Credibility*, 129–163.

Hardy, M. (2006). *An Introduction to Risk Measures for Actuarial Applications.* Society of Actuaries. Construction and Evaluation of Actuarial Models Study Note.

Hogg, R. V., J. W. McKean, and A. T. Craig (2005). *Introduction to Mathematical Statistics* (6th ed.). Pearson.

Irony, T. Z. and N. D. Singpurwalla (1997). Non-informative priors do not exist a dialogue with José M. Bernardo. *Journal of Statistical Planning and Inference 65*(1), 159 – 177.

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.

Jewell, W. S. (1974). Credible means are exact Bayesian for exponential families. *ASTIN Bulletin 8*(1), 77–90.

Jewell, W. S. (1975). he use of collateral data in credibility theory: a hierarchical model. *Giornale dell'Istituto Italiano degli Attuari*.

Johnson, N. (1957). Uniqueness of a result in the theory of accident proneness. *Biometrika 44*, 530–531.

Johnson, N., S. Kotz, and N. Balakrishnan (1995). *Continuous univariate distributions*, Volume 2 of *Wiley series in probability and mathematical statistics: Applied probability and statistics.* Wiley & Sons.

Jørgensen, B. (1992). *The Theory of Exponential Dispersion Models and Analysis of Deviance.* Instituto de Matemática Pura e Aplicada, (IMPA), Brazil.

Jørgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall, London.

Kaas, R., M. Goovaerts, J. Dhaene, and M. Denuit (2008). *Modern Actuarial Risk Theory: Using R*. Springer.

Kemperman, J. H. B. (1969). On the optimum rate of transmitting information. *The Annals of Mathematical Statistics 40*(6), 2156–2177.

Kullback, S. (1968). *Information Theory and Statistics*. Dover Publications, New York.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*(1), 79–86.

Landsman, Z. M. and U. E. Makov (1998). Exponential dispersion models and credibility. *Scandinavian Actuarial Journal 1998*(1), 89–96.

Mayerson, A., D. A. Jones, and J. N. L. Bowers (1968). On the credibility of the pure premium. *Proceedings of the Casualty Actuarial Society LV*(103 & 104), 175–185.

Mowbray, A. (1914). How extensive a payroll exposure is necessary to give a dependable pure premium? *Proceedings of the Casualty Actuarial Society I*(1), 24–30.

Najafabadi, A. T. P. (2010). A new approach to the credibility formula. *Insurance: Mathematics and Economics 46*(2), 334 – 338.

Neal, R. M. (2010). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo 54*, 113–162.

Nelder, J. and R. Verrall (1997). Credibility theory and generalized linear models. *ASTIN Bulletin 27*(1), 71–82.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A, General 135*, 370–384.

Ohlsson, E. (2008). Combining generalized linear models and credibility models in practice. *Scandinavian Actuarial Journal 2008*(4), 301–314.

Papadakis, M., M. Tsagris, M. Dimitriadis, S. Fafalios, I. Tsamardinos, M. Fasiolo, G. Borboudakis, J. Burkardt, C. Zou, K. Lakiotaki, and C. Chatzipantsiou. (2018). *Rfast: A Collection of Efficient and Extremely Fast R Functions.* R package version 1.9.2.

Perryman, F. (1932). Some notes on credibility. *Proceedings of the Casualty Actuarial Society XIX*(39 & 40), 65–84.

Pitselis, G. (2004). De Vylder's robust nonlinear regression credibility. *Belgian Actuarial Bulletin 4*(1), 44–49.

Quijano Xacur, O. A. (2019a). mythesis. R package.

Quijano Xacur, O. A. (2019b). *unifed: The Unifed Distribution.* R package version 1.1.

Quijano Xacur, O. A. and J. Garrido (2018). Bayesian credibility for glms. *Insurance: Mathematics and Economics 83*, 180 – 189.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods.* Springer-Verlag New York.

Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation* (2nd ed.). Springer.

Schmitter, H. (2004). The sample size needed for the calculation of a GLM tariff. *ASTIN Bulletin 34*(1).

Simas, A. B., W. Barreto-Souza, and A. V. Rocha (2010). Improved estimators for a general class of beta regression models. *Computational Statistics & Data Analysis 54*(2), 348 – 366.

Smyth, G. K. and A. P. Verbyla (1999). Double generalized linear models: approximate reml and diagnostics. *Proceedings of the 14th International Workshop on Statistical Modelling*, 66–80.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.2.

Verbraak, H. (1990). *The Logic of Objective Bayesianism.* Amsterdam, The Netherlands: Insist Publishing Consultancy.

Whitney, A. (1918). The theory of experience rating. *Proceedings of the Casualty Actuarial Society IV*(9 & 10), 274–292.

Wolny-Dominiak, A. and M. Trzesiok (2014). *insuranceData: A Collection of Insurance Datasets Useful in Risk Classification in Non-life Insurance.* R package version 1.0.

Young, V. R. (2006). *Premium Principles.* John Wiley & Sons, Ltd.