Chapter 1

Bayesian Regression Models

Luis E. Nieto-Barajas and Enrique de Alba

Chapter preview. In this chapter we approach many of the topics of the previous chapters, but from a Bayesian viewpoint. Initially we cover the foundations of Bayesian inference. We then describe the Bayesian linear and generalized regression models. We concentrate on the regression models with zero-one and count response and illustrate the models with real datasets. We also cover hierarchical prior specifications in the context of mixed models. We finish with a description of a semiparametric linear regression model with a nonparametric specification of the error term. We also illustrate its advantage with respect to the fully parametric setting using a real data set.

1.1 Introduction

The use of Bayesian concepts and techniques in actuarial science dates back to Whitney (1918) who laid the foundations for what is now called empirical Bayes credibility. He mentions that the solution of the problem "depends upon the use of inverse probabilities". This is the term used by T. Bayes in his original paper. However, Ove Lundberg was apparently the first one that realized the importance of Bayesian procedures (Lundberg, 1940). In addition, Bailey (1950) put forth a clear and strong argument in favor of using Bayesian methods in actuarial science. To date, the Bayesian methodology is used in various

areas within actuarial science, see for example Klugman (1992), Makov (2001), Makov *et al.* (1996) and Scollnik (2001).

Bayesian methods have several advantages that make them appealing for their use in actuarial science. First, they allow the actuary to formally incorporate expert or existing prior information. This prior information can be in the form of global or industry-wide information (experience) or in the form of tables. In this respect it is indeed surprising that Bayesian methods are not used more extensively, since there is a wealth of "objective" prior information available to the actuary. In fact, the "structure distribution" frequently used in credibility was originally formulated in a Bayesian framework (Bühlmann, 1967).

The second, advantage of Bayesian methods is that the analysis is always done by means of the complete probability distribution for the quantities of interest, either the parameters, or the future values of a random variable. Actuarial science is a field where adequate understanding and knowledge of the complete distribution is essential. In addition to expected values we are usually looking at certain characteristics of probability distributions, e.g. ruin probability, extreme values, value at risk (VaR), and so on.

From a theoretical point of view, Bayesian methods have an axiomatic foundation and are derived from first principles (Bernardo and Smith, 2000). From a practical perspective, Bayesian inference is the process of fitting a probability model to a set of data and summarizing the uncertainty by a probability distribution on the parameters of the model and on unobserved quantities, such as predictions for new observations. A fundamental feature of Bayesian inference is the direct quantification of uncertainty. To carry it out, the actuary must set up a full probability model (a joint probability distribution) for all observable and unobservable quantities in a given problem. This model should be consistent with knowledge about the process being analyzed. Then, Bayesian inference about the parameters in the model or about unobserved data are made in terms of probability statements that are conditional on the observed data (posterior distributions). Hence, these methods provide a full distributional profile for the parameters, or other quantities of interest, so that the features of their distribution are readily apparent, for example, nonnormality, skewness, tail behavior, or others. To obtain these posterior distributions, Bayesian methods combine the prior available information, no matter how limited, with the theoretical models for the variables of interest. Therefore Bayesian models automatically account for all the uncertainty in the

parameters.

1.2 The Bayesian Paradigm

Bayesian theory is developed from the axiomatic system of the foundations of decision theory. In some references the dual concepts of probability and utility are formally defined and analyzed. Probabilities are considered "degrees of belief" of the analyst about the occurrence of a given event and the criterion of maximizing expected utility is seen to be the only criterion compatible with the axiomatic system. Statistical inference is viewed as a particular decision problem, and statistical inference, whether estimation or prediction, must follow the laws of probability. As a result, the uncertainty of all unknown quantities is described in terms of probability distributions which implies that these quantities are treated as random variables. The fact that parameters have a distribution function allow the application of Bayes Theorem to combine information coming from the data with prior information about the parameters. For a comprehensive exposition on the foundations see Bernardo and Smith (2000) and references therein.

The ensuing methodology establishes how to formally combine an initial (prior) degree of belief of a researcher with currently measured, observed data, in such a way that it updates the initial degree of belief. The result is named posterior belief. This process is called Bayesian inference since the updating process is carried out through the application of Bayes Theorem. The posterior belief is proportional to the product of the two types of information, the prior information about the parameters in the model, and the information provided by the data. This second part is usually thought of as the objective portion of the posterior belief. We explain this process as follows:

Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be independent random variables, each of them coming from a probability model with density function $f(y_i|\theta)$, where θ is a parameter vector that characterizes the form of the density. Then $f(\mathbf{y}|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$ is the joint probability density of $\mathbf{y}$ given θ that is usually referred to as the likelihood function. Prior available information on the parameter is described through a prior distribution $f(\theta)$ that must be specified or

modeled by the actuary. Then, from a purely probabilistic point of view, it follows that

$$\boxed{f(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta) f(\theta)}{f(\mathbf{y})}}$$

where $f(\mathbf{y})$ is the marginal joint density of $\mathbf{y}$ defined as $f(\mathbf{y}) = \int f(\mathbf{y}|\theta) f(\theta)\, \mathrm{d}\theta$ if θ is continuous, and $f(\mathbf{y}) = \sum_\theta f(\mathbf{y}|\theta) f(\theta)$ if θ is discrete. This is Bayes' Theorem that rules the updating of the information. Considering that $f(\mathbf{y})$ is just a constant for θ, then the updating mechanism can be simply written as $f(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta) f(\theta)$, where $\propto$ indicates proportionality. In other words, the posterior distribution of the parameters, conditional on the observed data, is proportional to the product of the likelihood function and the prior degree of belief. Any inference on the parameters is now carried out using the posterior distribution $f(\theta|\mathbf{y})$.

As was mentioned above, the only criterion for optimal decision making, consistent with the axiomatic system, is the maximization of the expected utility. Alternatively, this criterion is equivalently replaced by the minimization of a loss function. Therefore, in the Bayesian framework parameter estimation is done by minimizing the expected value of a specified loss function $l(\hat{\theta}, \theta)$ with respect to $\hat{\theta}$, where the expected value is taken with respect to the posterior distribution of the parameter θ given the data $\mathbf{y}$. In particular, a quadratic loss function $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$ leads to the posterior mean $\hat{\theta} = \mathrm{E}(\theta|\mathbf{y})$ as an optimal estimate for the parameter. On the other hand, a linear loss function $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$ yields to the median of the posterior distribution as an optimal estimate $\hat{\theta}$ for θ.

Nuisance parameters are handled in a very straightforward fashion within the Bayesian settting via marginalization. For example, if the parameter has two components, say $\theta = (\phi, \lambda)$ where ϕ is the parameter of interest and λ is the nuisance parameter, inference is done using the marginal posterior distribution $f(\phi|\mathbf{y}) = \int f(\phi, \lambda|\mathbf{y})\, \mathrm{d}\lambda$.

When the main purpose of modeling is prediction, then the observed data $\mathbf{y}$ are used to predict future observations y_F by means of the posterior predictive distribution. Assuming continuous random variables to simplify presentation, the predictive distribution is defined as

$$\boxed{f(y_F|\mathbf{y}) = \int f(y_F|\theta) f(\theta|\mathbf{y})\, \mathrm{d}\theta} \tag{1.1}$$

The parameters in the model have been marginalized (integrated out). Therefore, only information in the observed data is used in prediction. Finally, the optimal point predictor $\hat{y_F}$, assuming a quadratic loss function, is the mean of the predictive distribution $E(y_F|\mathbf{y})$.

To summarize, the Bayesian inference method can be thought of as comprising the following principal steps:

a Specify the prior beliefs in terms of a probability model. This should reflect what is known about the parameter prior to observing the data.

b Compute the likelihood function in terms of the probability model that gave rise to the data. This contains the observed information about the parameters.

c Apply Bayes' Theorem to derive the posterior density. This posterior belief expresses what we know about the parameters after observing the data together with the prior belief.

d Derive appropriate inference statements about the parameter from the posterior distribution, and about future observations from the posterior predictive distribution.

There is a vast literature on how to specify a prior distribution. One of the most common approaches is to use the family of natural conjugate priors. A prior distribution $f(\theta)$ is said to be a natural conjugate for θ if, when combining it with the sample information, $f(\theta)$ and the resulting posterior $f(\theta|\mathbf{y})$ belong to the same family. These priors can be used to produce vague or diffuse priors, which reflect knowing little or having no prior information about the parameter, or to produce informative priors that reflect the prior knowledge of the actuary. In either case this is achieved by setting the parameters of the prior to an appropriate value.

1.3 Generalized linear models

1.3.1 Linear models

The linear regression model is a way of expressing the relationship between a dependent or response variable y and a set of $p-1$ independent or explanatory variables $\mathbf{x}' = (1, x_1, \ldots, x_{p-1})$, via a linear combination with coefficients $\boldsymbol{\beta}' = (\beta_0, \ldots, \beta_{p-1})$ of the form $\boldsymbol{\beta}'\mathbf{x} = \beta_0 + \beta_1 x_1 +$

$\cdots + \beta_{p-1}x_{p-1}$. This relationship can be expressed in terms of a linear equation with an additive random error ϵ such that

$$y = \boldsymbol{\beta}'\mathbf{x} + \epsilon, \tag{1.2}$$

where ϵ is interpreted as a measurement error and is assumed to have zero mean and constant precision (reciprocal of the variance). If we further assume that the error comes from a normal distribution, then $\epsilon \sim \mathrm{N}(0, \tau)$.

The normal assumption in the error term implies that the response variable y, conditional on $\mathbf{x}$, also follows a normal model and thus can take any possible value in the real line. Generalized linear models extend this assumption to response variables with positive, bounded or discrete outcomes. For example, if one is desired to describe the behavior of the amounts in insurance claims or the number of claims in a certain period of time, the normal assumption would not be adequate in either case since the claims cannot be negative and the number of claims are positive and discrete.

The role played by the explanatory variables $\mathbf{x}$ in the linear normal (linear regression) model is to help in understanding the average or mean behavior of the response variable y. This is why the (conditional) expected value $\mathrm{E}(y|\mathbf{x})$ is equal to a linear combination of the explanatory variables $\boldsymbol{\beta}'\mathbf{x}$. This justifies the name *regression to the mean* of the linear regression model (1.2).

The linear regression model is a particular case of the larger class of generalized linear models. We will discuss its properties and Bayesian inference in the following sections.

1.3.2 Generalized linear models

A Bayesian generalized linear model is a generalized linear model together with a specification of the prior believes of the unknown parameters. It can be said that generalized linear models are also regression models to the mean (of y) but in a non linear form since the parameter space of $\mathrm{E}(y|\mathbf{x})$ is not necessarily the whole real line. Let us start by recalling the form of a generalized linear model. In order to account for all possible kinds of response variables (positive, bounded, discrete, etc.) the model describes the probabilistic behavior of the responses with a member of the exponential family. Then, for a sample of independent

random variables $y_1, y_2, \ldots, y_n$, each of them comes from the model

$$f(y_i \mid \theta_i, \phi_i) = b(y_i, \phi_i) \exp[\phi_i\{y_i\theta_i - a(\theta_i)\}], \tag{1.3}$$

where $a(\cdot)$ and $b(\cdot)$ are two monotonic functions. The parameters θ_i and ϕ_i are known as natural and dispersion parameters, respectively. It is not difficult to show that the mean and variance of y_i can be expressed in terms of derivatives of function $a(\cdot)$ as follows:

$$\mu_i = \mathrm{E}(y_i) = a'(\theta_i) \text{ and } \sigma_i^2 = \mathrm{Var}(y_i) = \frac{a''(\theta_i)}{\phi_i}.$$

Here prime and double prime denote first and second derivative, respectively.

Each individual i has its own set of explanatory variables $\mathbf{x}_i$, $i = 1, \ldots, n$. These will be combined in a single value through a linear combination with coefficients $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_{p-1})$ forming what is called the *linear predictor* $\eta_i = \boldsymbol{\beta}'\mathbf{x}_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i,p-1}$. We note that linearity means linear in the coefficients since the linear predictor could well be a polynomial of order $p-1$ of the form $\beta_0 + \beta_1 x_i + \cdots + \beta_{p-1}x_i^{p-1}$ with a single explanatory variable for individual i, x_i.

The idea of the generalized linear models is to model the mean of the response variable, $\mu_i = \mathrm{E}(y_i)$, in terms of the explanatory variables via the linear predictor η_i and an appropriate transformation $g(\cdot)$, that is, $\eta_i = g(\mu_i)$. The function $g(\cdot)$ is called *link function* because it links the explanatory variables with the response. At the same time, the link function adjust the parameter space of μ_i to correspond to the values of the predictor η_i, which is typically the real line. This can be seen as $\mu_i = g^{-1}(\eta_i)$. A particular choice for the link function $g(\cdot)$ is to take $g^{-1}(\cdot) = a'(\cdot)$. In this case the linear predictor η_i becomes equal to the natural parameter θ_i and $g(\cdot)$ is called canonical link function. Other options for the link function are available, as long as the domain of the function $g(\cdot)$ corresponds to the parameter space of μ_i and the image to the real numbers. Let us consider a couple of examples to illustrate these ideas.

It can be shown that the normal linear regression model is also a generalized linear model. To see this we take $y_i \sim \mathrm{N}(\mu_i, \tau_i)$ parameterized in terms of mean μ_i and precision (reciprocal of the variance) τ_i. The density function is

$$f(y_i \mid \mu_i, \tau_i) = (2\pi/\tau_i)^{-1/2} \exp\left\{-\frac{\tau_i}{2}(y_i - \mu_i)^2\right\}$$

for $y_i \in \mathbb{R}$, $\mu_i \in \mathbb{R}$ and $\tau_i > 0$. Writing the normal density as in (1.3) we get

$$\boxed{\begin{array}{ll} \phi_i = \tau_i, & b(y_i, \phi_i) = (2\pi/\phi_i)^{-1/2} \exp\left\{\frac{\phi_i}{2} y_i^2\right\} \\ \theta_i = \mu_i, & a(\theta_i) = \frac{\theta_i^2}{2} \end{array}}$$

In this case $a'(\theta_i) = \theta_i$ and thus the canonical link is $g(\mu_i) = \mu_i$. Therefore the mean μ_i is modeled directly with the linear predictor η_i obtaining the linear model $\mu_i = \boldsymbol{\beta}'\mathbf{x}_i$.

A second example, suitable for response variables in the positive real line (as is the case for the claim amounts in insurance companies), is to consider a response with gamma distribution. Using a mean parameterization of the gamma, that is, $y_i \sim \text{Ga}(\alpha_i, \alpha_i/\mu_i)$ such that $\text{E}(y_i) = \mu_i$, then the density function is of the form

$$f(y_i \mid \alpha_i, \mu_i) = \left(\frac{\alpha_i}{\mu_i}\right)^{\alpha_i} \frac{1}{\Gamma(\alpha_i)} y_i^{\alpha_i - 1} e^{-\frac{\alpha_i}{\mu_i} y_i},$$

for $y_i > 0$ and $\alpha_i, \mu_i > 0$. Writing this gamma density as in (1.3) we get

$$\boxed{\begin{array}{ll} \phi_i = \alpha_i, & b(y_i, \phi_i) = \frac{\phi_i^{\phi_i}}{\Gamma(\phi_i)} y_i^{\phi_i - 1} \\ \theta_i = -\frac{1}{\mu_i}, & a(\theta_i) = \log\left(-\frac{1}{\theta_i}\right) \end{array}}$$

Computing the derivative of the function $a(\cdot)$ we obtain $a'(\theta_i) = 1/\theta_i$, implying a canonical link $g(\mu_i) = -1/\mu_i$. We note that this link has a problem, its domain is fine since it corresponds to the parameter space of μ_i, however, the image of $g(\cdot)$ is the negative numbers and not the real line. An alternative link function that overcomes this flaw is to take $g(\mu_i) = \log(\mu_i)$, where the domain and image are as desired. More examples for response variables in $\{0, 1\}$ and for count variables will be presented in the following sections.

Once we have defined the sampling model for the data, a Bayesian model is completed by assigning prior distributions to the unknown quantities. A typical assumption in the previously defined generalized linear models, is to consider a common dispersion parameter $\phi_i = \phi$ for all individuals $i = 1, \ldots, n$. In this case, the set of unknown parameters in the model is $(\boldsymbol{\beta}, \phi)$. According to West (1985), conjugate priors for these parameters are available only in very special cases. In general, posterior distributions are not available in close forms so the choice of the prior has to do with the simplicity to accommodate prior believes. In this context a normal prior for the vector $\boldsymbol{\beta}$ has been the common choice, together with a gamma

prior for the precision ϕ. Typically assuming independence a-priori among the β_j elements of $\boldsymbol{\beta}$ and between $\boldsymbol{\beta}$ and ϕ. The normal and gamma are well known models that allow the user to specify prior believes in terms of mean and precision (reciprocal of variance). That is, we can take a priori $\beta_j \sim \mathrm{N}(b_0, t_0)$, where b_0 is the prior mean and t_0 the prior precision for β_j. In the case of little or none information about β_j we can set $b_0 = 0$ together with t_0 close to zero, say 0.1, 0.01 or 0.001, for $j = 1, \ldots, p$. For the sampling precision parameter ϕ, if μ_ϕ and σ^2_ϕ represent the prior mean and variance, then we can take $\phi \sim \mathrm{Ga}(a_0, a_1)$ with $a_0 = \mu^2_\phi/\sigma^2_\phi$ and $a_1 = \mu_\phi/\sigma^2_\phi$. Again, in the case of little or none prior information about ϕ, we can set $a_0 = a_1$ equal to a small value, say 0.1, 0.01 or 0.001, in such a way that ϕ has prior mean one and large/small prior variance/precision. Alternatively, more diffuse priors are also considered for the coefficients β_j, for instance a student-t prior or even a cauchy prior (Gelman *et al.*, 2008).

Posterior inference of the parameters $(\boldsymbol{\beta}, \phi)$ requires to combine the information provided by the data, summarized in the likelihood function, and the prior distributions. The likelihood function is constructed by the product of the density (1.3) of the response variables as a function of the explanatory variables and the parameters, that is, $\mathrm{lik}(\boldsymbol{\beta}, \phi) = \prod_{i=1}^{n} f(y_i|\, \theta_i, \phi)$. Remember that the explanatory variables enter the model via the natural parameter θ_i, which in the case of using the canonical link $\theta_i = \boldsymbol{\beta}'\mathbf{x}_i$, otherwise θ_i is replaced with an appropriate function of the linear predictor η_i. Finally, the posterior distribution $f(\boldsymbol{\beta}, \phi|\, \text{data})$ is proportional to the product of this likelihood function $\mathrm{lik}(\boldsymbol{\beta}, \phi)$ and the prior distributions $f(\boldsymbol{\beta}, \phi)$. Point estimates and credible intervals are obtained as summaries from this posterior distribution. Those summaries are obtained numerically via a MCMC sampling algorithm or via EM techniques. The former can be implemented in `OpenBugs` within R through the library `R2OpenBUGS`. The latter is implemented in the R command `bayesglm` from the package `arm` (data analysis using regression and multilevel/hierarchical models). Both are available in `CRAN`.

Another aspect of interest when using generalized regression models is prediction of future outcomes. This inference problem is addressed naturally in the Bayesian approach by computing the predictive distribution for a future observation y_F. If a new individual has explanatory variables $\mathbf{x}_F$, and assuming the canonical link, then $\theta_F = \boldsymbol{\beta}'\mathbf{x}_F$ and the predictive distribution will be the weighted average of the density $f(y_F|\, \theta_F, \phi)$ with respect to

the posterior distribution $f(\boldsymbol{\beta}, \phi|\,\text{data})$ as in (1.1). Point or interval predictions are produced using summaries from this predictive distribution. This is usually done numerically.

Example 14.1. The insurance market in Mexico operates in different classes. Seven of these are: Accident and sickness (ACC), agriculture and livestock (AGR), automobiles (AUT), major medical expenses (MED), fire (FIR), liability and professional risks (LIA) and health (HEA). It is of interest to the insurance companies to predict claim amounts y_i in terms of the premiums written x_i. The insurance industry regulator in Mexico gathers the information from all different insurance companies every year and makes the information available in its web page http://www.cnsf.gob.mx/. The information is available for all 32 Mexican States, and in some cases from abroad. In total, for the year 2010, we have $i = 1, \ldots, n$ with $n = 228$ observations classified by insurance sector. The dataset can be found in the Web Appendix of the book. A dispersion diagram of the 228 observations in logarithmic scale is presented in Figure 1.1. From the graph we can see that all sectors together follow a common pattern and a single line could potentially serve for fitting the data. In fact, the least square estimates are 0.0008 for the intercept and 0.85 for the slope.

Let us assume that the logarithm of the claim amounts $\log(y_i)$ follows a normal distribution with mean μ_i and constant precision τ, that is $\log(y_i) \sim \text{N}(\mu_i, \tau)$. We model the mean level μ_i in terms of a linear combination of the premiums written in log scale $\log(x_i)$ and class indicators z_{ji}, for $j = 2, \ldots, 7$, where for example z_{2i} takes the value of one if observation i belongs to sector 2 (AGR), and so on, following the order of the sectors in the previous paragraph. These sector indicators will serve to determine possible differences in the intercepts and the slopes by including the interactions $\log(x_i) * z_{ji}$. The mean level is thus modeled as $\mu_i = \alpha_1 + \sum_{j=2}^{7} \alpha_j z_{ji} + \beta_1 \log(x_i) + \sum_{j=2}^{7} \beta_j \log(x_j) * z_{ji}$. Note that to avoid indetermination of the model, the indicator for sector one is not present, so sector one has been taken as baseline. An individual i coming from sector one is identified by assigning zeroes to all sector indicators z_{ji}, $j = 2, \ldots, 7$. For the model coefficients we assign vague normal priors centered at zero and with small precision, that is, $\alpha_j \sim \text{N}(0, 0.001)$ and $\beta_j \sim \text{N}(0, 0.001)$ independently for $j = 1, \ldots, 7$. For the common precision of the observations we take $\tau \sim \text{Ga}(0.001, 0.001)$ such that τ has mean one and large variance a-priori. The R (Bugs) code of this model is given in Table 1.1.

Posterior estimates of the model coefficients and their credible intervals are presented

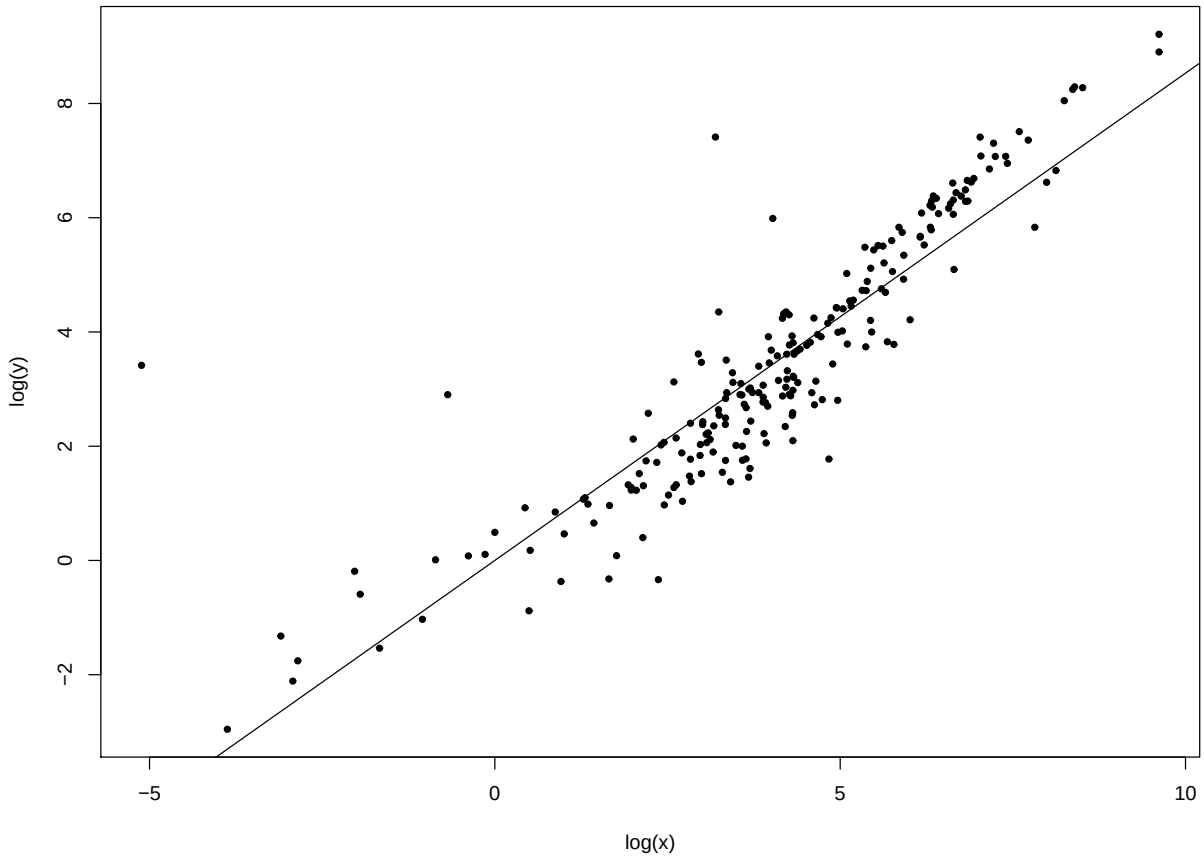

Figure 1.1: Dispersion diagram of severity amounts y_i versus premium written x_i in logarithmic scale for $i = 1, \ldots, 228$ individuals. Straight line corresponds to a least square fit to the data.

Table 1.1: Bugs code for model of Example 14.1.

```
model{
#Likelihood
for (i in 1:n){
y[i]~dnorm(mu[i],tau)
mu[i]<-a[1]+a[2]*z2[i]+a[3]*z3[i]+a[4]*z4[i]+a[5]*z5[i]+a[6]*z6[i]+a[7]*z7[i]
      +b[1]*x[i]+b[2]*x[i]*z2[i]+b[3]*x[i]*z3[i]+b[4]*x[i]*z4[i]+b[5]*x[i]*z5[i]
      +b[6]*x[i]*z6[i]+b[7]*x[i]*z7[i]
}
#Priors
for (j in 1:7){
a[j]~dnorm(0,0.001)
b[j]~dnorm(0,0.001)
}
tau~dgamma(0.001,0.001)
}
```

in the second and third columns in Table 1.2. If the hypothesis of a single regression line for all sectors were true then coefficients α_j and β_j for $j = 2, \ldots, 7$ would all need to be zero. As we can see from the table, except for α_5, the rest of the coefficients are all different from zero implying different intercepts $\alpha_1 + \alpha_j$ and different slopes $\beta_1 + \beta_j$ for each sector j. These differences can be better appreciated graphically in Figure 1.2 where each colored line corresponds to a different sector. From the graph it is noticeable that sector ACC, represented by the black line, is the one that deviates the most from the general pattern of Figure 1.1. This large difference is mostly explained by the extreme observation with coordinates $(-5.11, 3.41)$ in logarithmic scale. An alternative model that is more robust to extreme observations is the semiparametric regresion model with a Polya tree prior for the errors, that will be described in Section 1.5. Finally, posterior mean of the observations precision τ is 1.55 with 95% credible interval (CI) $(1.26, 1.86)$.

1.3.3 Bayesian regression with zero-one dependent variables

In actuarial science and risk management, it is of interest to estimate the probability of default. For instance, when assigning a personal credit (loan) to an individual, the finan-

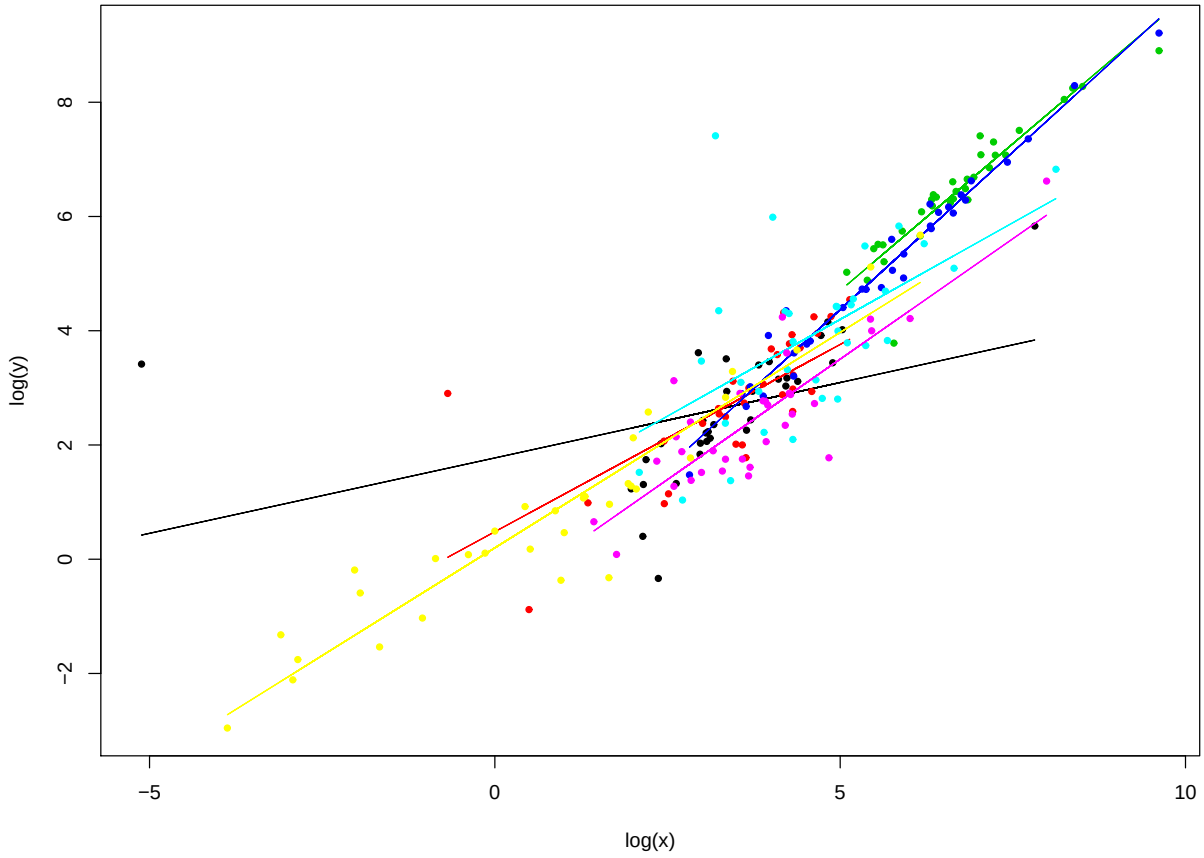

Figure 1.2: Dispersion diagram of claim amounts y_i versus premiums written x_i in logarithmic scale by class. Straight lines corresponds to model fit by sector. Colors indicate different sector: black (ACC), red (AGR), green (AUT), blue (MED), cyan (FIR), magenta (LIA) and yellow (HEA).

Table 1.2: Posterior estimates and credible intervals of multiple regression coefficients in the insurance dataset. Parametric and semiparametric formulations.

	Parametric		Semiparametric	
Coef.	Mean	95% CI	Mean	95% CI
α_1	1.77	(1.34, 2.20)	0.48	(-0.10, 1.63)
α_2	-1.29	(-2.24, -0.32)	-1.00	(-2.20, -0.23)
α_3	-2.21	(-4.22, -0.15)	-0.45	(-2.01, 0.63)
α_4	-2.92	(-4.01, -1.72)	-1.79	(-2.99, -1.01)
α_5	-0.96	(-2.11, 0.15)	-0.66	(-2.30, 0.54)
α_6	-2.49	(-3.48, -1.54)	-1.47	(-2.46, -0.63)
α_7	-1.57	(-2.11, -1.06)	-0.11	(-1.19, 0.57)
β_1	0.26	(0.14, 0.38)	0.64	(0.41, 0.80)
β_2	0.39	(0.13, 0.64)	0.31	(0.12, 0.57)
β_3	0.76	(0.45, 1.06)	0.32	(0.13, 0.62)
β_4	0.84	(0.61, 1.04)	0.48	(0.27, 0.73)
β_5	0.41	(0.17, 0.67)	0.24	(-0.01, 0.57)
β_6	0.58	(0.35, 0.84)	0.25	(0.01, 0.45)
β_7	0.49	(0.33, 0.65)	0.13	(-0.04, 0.31)

cial institution needs to quantify the risk of default, according to the individual personal characteristics and financial history. This problem can be modeled by assuming a zero-one (Bernoulli) response variable y_i, with probability of success (default) μ_i, that is, $y_i \sim \text{Ber}(\mu_i)$ with density function given by

$$f(y_i \mid \mu_i) = \mu^{y_i}(1-\mu_i)^{1-y_i},$$

for $y_i \in \{0, 1\}$ and $\mu_i \in (0, 1)$. Writing this density as in (1.3), to identify the model as a member of the exponential family, we get

$$\begin{array}{|cc|}\hline \phi_i = 1, & b(y_i, \phi_i) = 1 \\ \theta_i = \log\left\{\frac{\mu_i}{1-\mu_i}\right\}, & a(\theta_i) = \log(1+e^{\theta_i}) \\ \hline \end{array}$$

The first derivative of function $a(\cdot)$ is $a'(\theta_i) = e^{\theta_i}/(1+e^{\theta_i})$. Inverting this function to obtain the canonical link we get $g(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$.

A generalized linear model for a Bernoulli response with canonical link is called *logistic regression model*. Recall that other link functions can be used as long as the domain of

function $g(\cdot)$ corresponds to the parameter space and the image to the real line. Since the parameter space of μ_i is the interval $(0, 1)$, any function that transform the $(0, 1)$ into the real numbers is a suitable link function. In the basic probability courses we learnt that cumulative distribution functions (c.d.f.) for continuous random variables are functions with real domain and $(0, 1)$ image. Therefore, the inverse of any continuous c.d.f. $F(\cdot)$ can be a link function, that is, $g(\cdot) = F^{-1}(\cdot)$. In particular if $F = \Phi$, the c.d.f. of a standard normal, then $g(\cdot) = \Phi^{-1}(\cdot)$ produces the *probit regression model*. In fact, the inverse of the canonical link corresponds to the c.d.f. of a logistic random variable, and thus the name of logistic regression. Other two common link functions are the log-log link $g(\mu_i) = \log\{-\log(\mu_i)\}$ and complementary log-log link $g(\mu_i) = \log\{-\log(1 - \mu_i)\}$. This latter link corresponds to the inverse of the c.d.f. of a extreme value distribution. Whatever the link function we choose, for a given vector of explanatory variables of individual i, the probability of success is expressed in terms of the explanatory variables as $\mu_i = g^{-1}(\eta_i) = F(\boldsymbol{\beta}'\mathbf{x}_i)$.

Sometimes several individuals share the same value of the explanatory variables, or it is also possible that available information is grouped and the covariate information is only available at the group level. This is the case, for example, in insurance groups where it is assumed that all individuals in the same group show similar risk characteristics and the number of claims y_i out of n_i members in group i is reported. In such a case it is of interest to estimate the severity (probability of presenting a claim) π_i for group i with characteristics $\mathbf{x}_i$. These kind of data can also be modeled with a generalized linear model by assuming $y_i \sim \text{Bin}(n_i, \pi_i)$ and $\pi_i = F(\boldsymbol{\beta}'\mathbf{x}_i)$ with a specific choice of continuous c.d.f. $F^{-1}(\cdot)$ as link function.

For both models, Bernoulli and Binomial, the precision parameter ϕ_i is equal to one, and for the grouped data, the number of individual in the group n_i is assumed known. This leaves us with one set of unknown parameters $\boldsymbol{\beta}$. For each β_j, $j = 1, \ldots, p$, we assign normal and student-t prior distributions as suggested above.

Example 14.2. The Mexican Central Bank is responsible for issuing the required number of bills for the well functioning of the economy. Table 1.3 contains the information on the number of bills in circulation and the number of fake bills, both in million pieces, for different denominations ($20, $50, $100, $200 and $500 Mexican pesos). This information is available anually from the year 2000 to 2011. Let us disregard temporal dependence and assume that

Table 1.3: Number of bills in circulation (C) and number of fake bills (F), in million pieces, for different bill denominations for years 2000 to 2011.

Year	C20	F20	C50	F50	C100	F100	C200	F200	C500	F500
2000	2182.1	14.8	3141.4	179.7	2779.4	178.5	4163.4	83.9	1100.7	26.6
2001	2092.6	13.1	2900.9	150.5	2795.0	136.8	4745.5	64.4	1335.0	20.9
2002	2182.4	18.1	3026.5	109.7	3155.5	64.2	5192.1	97.3	1802.1	35.7
2003	2449.1	9.4	4245.0	140.9	4455.4	60.1	4870.4	77.6	2352.4	42.9
2004	2545.8	1.5	4031.8	149.2	4951.7	117.8	5087.4	80.5	3028.0	34.0
2005	2707.8	1.0	3420.2	249.3	4411.0	142.9	5422.1	117.8	3522.5	43.6
2006	2877.4	0.7	3615.2	215.1	4625.9	106.5	5935.6	88.8	4190.9	70.9
2007	2959.8	0.6	3847.5	122.5	4768.0	77.1	6358.0	78.6	4889.7	90.5
2008	3360.9	1.0	3892.8	59.4	4830.2	87.6	6850.7	97.7	5682.5	91.7
2009	3578.6	3.2	4129.0	28.3	4872.5	81.0	7314.7	136.3	6934.4	91.2
2010	3707.6	2.7	4197.3	67.9	5210.0	101.2	7505.1	139.7	7799.3	96.4
2011	3858.8	1.3	4375.1	208.9	5416.0	88.7	7528.1	120.1	8907.4	89.7

Source: Banco de México. http://www.banxico.org.mx/estadisticas/index.html.

the number of fake bills y_i follows a binomial distribution with parameters n_i, the number of circulating bills, and π_i, the proportion of fake bills with respect to the real bills circulating, that is $y_i \sim \text{Bin}(n_i, \pi_i)$, for $i = 1, \ldots, n$ with $n = 60$ observations. To identify the bill denomination we construct auxiliary dummy variables, say $x50_i, x100_i, x200_i, x500_i, x1000_i$ such that they take the value of one if the observation i corresponds to the bill denomination they represent and zero otherwise. Note that \$20 pesos bills are identified when all dummy variables take the value of zero. We then define a linear predictor of the form $\eta_i = \beta_1 + \beta_2 x50_i + \beta_3 x100_i + \beta_4 x200_i + \beta_5 x500_i$. We will compare the logistic and standard normal links. These two links imply $\pi_i = e^{\eta_i}/(1 + e^{\eta_i})$ and $\pi_i = \Phi(\eta_i)$ respectively. For the prior distributions we consider two alternatives $\beta_j \sim \text{N}(0, 0.001)$ and $\beta_j \sim \text{St}(0, 0.001, 3)$ for $j = 1, \ldots, 5$. Note that the third parameter in the student-t distribution correspond to the degrees of freedom which has to be greater than 2. We chose 3 to avoid numerical problems. This model is translated into R (Bugs) code as shown in Table 1.4.

To compare the different models, the library `R2OpenBUGS` automatically computes the deviance which is defined as $-2 \log \text{lik}$, so a smaller value indicates a better fit. Table 1.5 shows posterior expected values for the deviance of the competing models. As can be seen from the table, the logit link is preferred to the probit link for this particular dataset, regardless of the prior distribution used. When comparing the two priors, there is little

Table 1.4: Bugs code for model of Example 14.2.

```
model{
#Likelihood
for (i in 1:n){
y[i]~dbin(pi[i],e[i])
logit(pi[i])<-b[1]+b[2]*x50[i]+b[3]*x100[i]+b[4]*x200[i]+b[5]*x500[i]
#probit(pi[i])<-b[1]+b[2]*x50[i]+b[3]*x100[i]+b[4]*x200[i]+b[5]*x500[i]
}
#Priors
for (j in 1:5){
b[j]~dnorm(0,0.001)
#b[j]~dt(0,0.001,3)}
#Useful parameters
for (j in 1:5){p[j]<-pi[12*(j-1)+1]*1000}}
```

difference in the fit, with the probit model slightly more sensitive to the choice of the prior. To better see the impact of the prior in the posterior estimates, we compare the two extreme models, the best fit achieved by the logit-normal model versus the worst fit obtained by the probit-student-t model. We define the rate of fake bills for every thousand of circulating bills as $p_j = \pi_{12(j-1)+1} \times 1000$, for each of the five bill denominations $j = 1, \ldots, 5$. Figure 1.3 shows a graph comparing these estimates. The central dots correspond to posterior means and the vertical lines to 95% credible intervals with limits obtained as the 2.5% and 97.5% quantiles from the posterior distribution, respectively. For each denomination the graph shows two lines. The left (black) line corresponds to the best fit (logit-normal), whereas the right (red) line to the worst fit (probit-student-t). In all cases, the credible intervals are larger when using the student-t prior. This is due to to the heavier tails with respect to the normal. Interpreting the estimated rates, the \$50 pesos bill has the largest fake rate with 37.5 fake bills for every 1000 circulating bills. On the other hand, the \$20 pesos bill shows the smallest rate with almost 2 fake bills for every 1000 circulating.

1.3.4 Bayesian regression with count dependent variables

Another common problem in actuarial science is the study of counting or count data. For example, the number of claims that an insured individual can file during a calendar year. The

Table 1.5: Deviance posterior means for four binomial generalized models fitted to the Mexican Central Bank data.

	Link	
Prior	logit	probit
Normal	1290.71	1358.16
Student-t	1290.75	1366.92

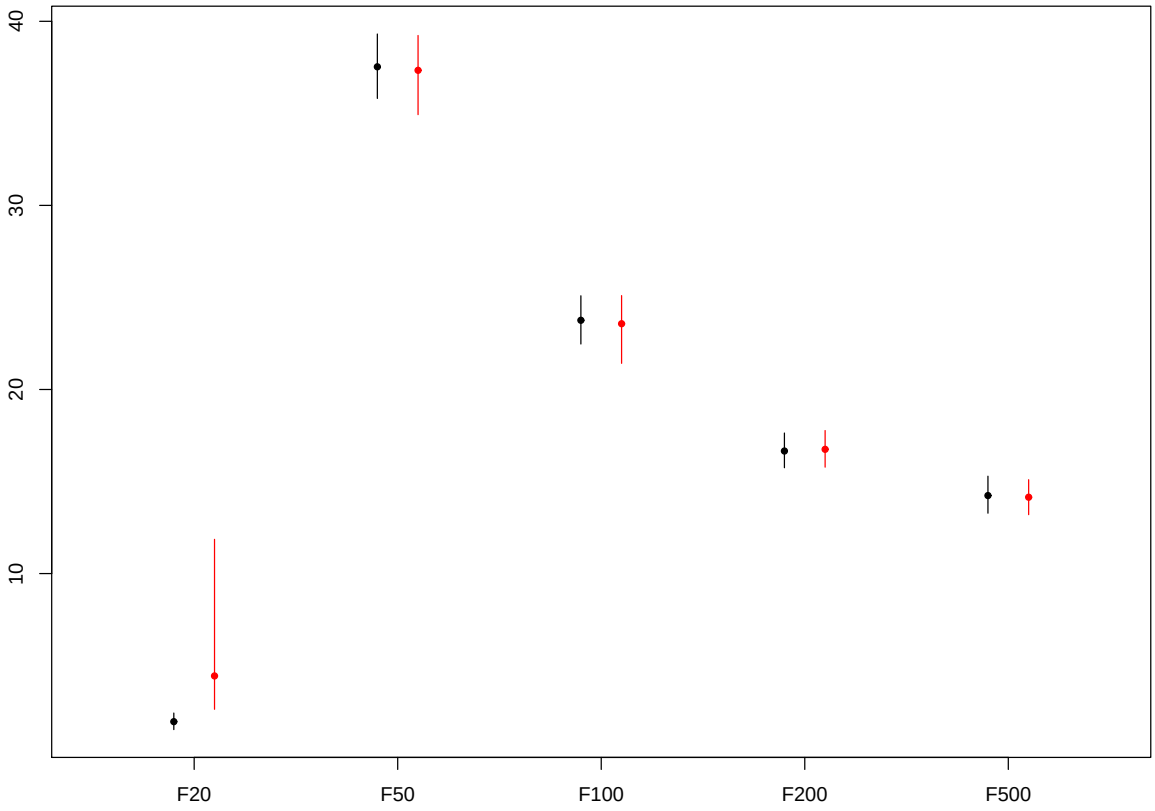

Figure 1.3: Posterior estimates of parameters p_i, $i = 1, \ldots, 5$. Central dots are posterior means and vertical lines denote 95% credible intervals. For each bill denomination, the left (black) line corresponds to the best fit and the right (red) line to the worst fit according to Table 1.5.

natural assumption for a counting response variable y_i is a Poisson model, that is $y_i \sim \text{Po}(\mu_i)$. This model has density function given by

$$f(y_i \mid \mu_i) = e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!},$$

for $y_i = 0, 1, \ldots$ and $\mu_i > 0$. Identifying this density as (1.3) we obtain

$$\boxed{\begin{array}{cc} \phi_i = 1, & b(y_i, \phi_i) = \frac{1}{y_i!} \\ \theta_i = \log(\mu_i), & a(\theta_i) = e^{\theta_i} \end{array}}$$

Thus the canonical link obtained as the inverse of the derivative of function $a(\cdot)$ is $g(\mu_i) = \log(\mu_i)$. Therefore, the mean of the response variable is modeled as $\mu_i = e^{\boldsymbol{\beta}'\mathbf{x}_i}$. This model is also known as *Poisson regression model*.

Sometimes, instead of modeling the mean μ_i of a Poisson response variable, it is of interest to model the rate of occurrence of events λ_i relative to a known number exposed or at risk e_i, such that $\mu_i = e_i\lambda_i$. In this case the rate is modeled through the explanatory variables as $\lambda_i = e^{\boldsymbol{\beta}'\mathbf{x}_i}$. For instance, in mortality studies, the maternity mortality ratio is defined as the rate of maternity deaths for every 100 thousand births. In such studies the response variable y_i is the number of maternity deaths, e_i is the number of births (in 100 thousands) and thus λ_i becomes the maternity mortality ratio.

As mentioned before, in regression models with count dependent variables, the Poisson model is the common assumption, however this model assumes that the mean and variance of the responses are the same, that is $\text{E}(y_i) = \text{Var}(y_i) = \mu_i$. In practice, this assumption is not always satisfied by the data, due to an effect of overdispersion ($\text{Var}(y_i) > \text{E}(y_i)$). To account for overdispersion in a dataset, a different model for the responses has to be used. The negative binomial is the typical alternative for modeling counting data in the presence of overdispersion, i.e., $y_i \sim \text{NB}(r_i, \pi_i)$. To give the parameters of the negative binomial the same interpretation as in the Poisson model, the integer parameter r_i has to coincide with the number of exposed e_i and the probability of success π_i with $1/(1+\lambda_i)$. This implies that $\text{E}(y_i) = e_i\lambda_i$ and $\text{Var}(y_i) = e_i\lambda_i(1+\lambda_i)$. The quantity $1+\lambda_i = \text{Var}(y_i)/\text{E}(y_i)$ is a measure of the amount of overdispersion present in the data. Finally, the rate λ_i is modeled in terms of the explanatory variables $\mathbf{x}_i$ as in the Poisson model.

In the Bayesian literature (e.g. Bernardo and Smith, 2000) it is well known that a negative binomial distribution is a particular case of a Poisson-gamma distribution. The latter

gets its name since it can be obtained as a mixture of a Poisson distribution with respect to a gamma distribution. For the particular parameterization of our negative binomial model $y_i \sim \text{NB}(e_i, 1/(1+\lambda_i))$, we can obtain the same model by considering a conditional Poisson distribution $y_i|t_i \sim \text{Po}(t_i\lambda_i)$ and marginal distribution $t_i \sim \text{Ga}(e_i, 1)$. Writing the negative binomial in this form allows us to consider the overdispersion case within the non-overdispersed Poisson setting by taking $t_i = e_i$ fixed if no overdispersion is present and $t_i \sim \text{Ga}(e_i, 1)$ random in the case of overdispersion. This construction of the negative binomial represents a hierarchical model that will be explained in detail in Section 1.4.

Here, as in the previous model with Bernoulli or binomial response variables, the only set of unknown parameters is $\boldsymbol{\beta}$, so normal and student-t distribution are used to represent prior knowledge.

Example 14.3. Consider the bills dataset presented in Table 1.3. The number of fake \$20 pesos bills is reported in variable F20. These numbers are shown as empty dots in Figure 1.4 and linked with a solid black line. As can be seen, there is a drastic drop of level in the number of fake bills before and after year 2003. This change is explained by the fact that in the early months of 2003 the Bank of Mexico released a new \$20 pesos bill made of polymer instead of regular money paper which is more difficult to counterfeit. To model these data we propose a generalized linear Poisson regression model that accounts for a change in the level. Specifically, we assume that the number of fake \$20 pesos bills y_i follows a Poisson distribution with rate or intensity μ_i, that is, $y_i \sim \text{Po}(\mu_i)$, with $log(\mu_i) = \beta_1 + \beta_2 I(t_i \geq \alpha)$, for observations $i = 1, \ldots, 12$. Note that t_i corresponds to the year reported in Table 1.3. We consider $\text{N}(0, 0.001)$ independent priors for β_j, $j = 1, 2$ and a uniform discrete prior for α on the set $\{2000, 2001, \ldots, 2011\}$. The corresponding R (Bugs) code is presented in Table 1.6.

The red lines, solid and dotted in Figure 1.4, correspond to the rate μ_i point estimates and 95% credible intervals, respectively, for all years. Years 2000 to 2003 (inclusive) have a common rate of 13.78 million fake pieces per year with a 95% credible interval of $(10.54, 17.72)$, whereas from 2004 to 2011 show a rate of 1.28 million pieces with a 95% credible interval of $(0.73, 2.16)$. The estimated change year α was 2004, denoted with a blue (dotted-dashed) vertical line in Figure 1.4. Although the new \$20 pesos bills were introduced at the beginning of 2003, the impact on the counterfeit rate was reflected from 2004 on. The number of

Table 1.6: Bugs code for model of Example 14.3.

```
model{
#Likelihood
for (i in 1:n){
y[i]~dpois(mu[i])
log(mu[i])<-b[1]+b[2]*step(t[i]-a)}
#Priors
a<-c+1999
c~dcat(p[])
for (j in 1:12){p[j]<-1/12}
for (j in 1:2){b[j]~dnorm(0,0.001)}
}
```

fake bills in 2003 was more similar to the previous years than the following years. This was captured by the model.

1.4 Mixed and hierarchical models

1.4.1 Mixed models

In the previous section, general regression models for different forms of the response variable were introduced. Those models are also known as *fixed effects models* and assume that the observed individuals $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ are independent. In some applications, response variables are observed in time (longitudinal models), in space (spatial models) or are clustered in groups (repeated measurements). All these cases assume certain kind of dependence among observations. *Mixed effects models*, or simply *mixed models* account for dependence among observations by introducing random (unobserved) effects in the model. The general specification of a mixed model assumes two sets of explanatory variables $\mathbf{x}_i$ and $\mathbf{z}_i$ such that the former is associated to fixed coefficients $\boldsymbol{\beta}$ and the latter to random coefficients $\boldsymbol{\alpha}_i$. Thus, in the context of a generalized linear model, the mixed model has a linear predictor of the form $\eta_i = \boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\alpha}_i'\mathbf{z}_i$. This is again, linked to the response variable, through an appropriate link function such that $g(\mu_i) = \eta_i$.

To better understand how dependence is introduced in a mixed model, let us consider a nested structure for the observations, say y_{ij}, where $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$. For

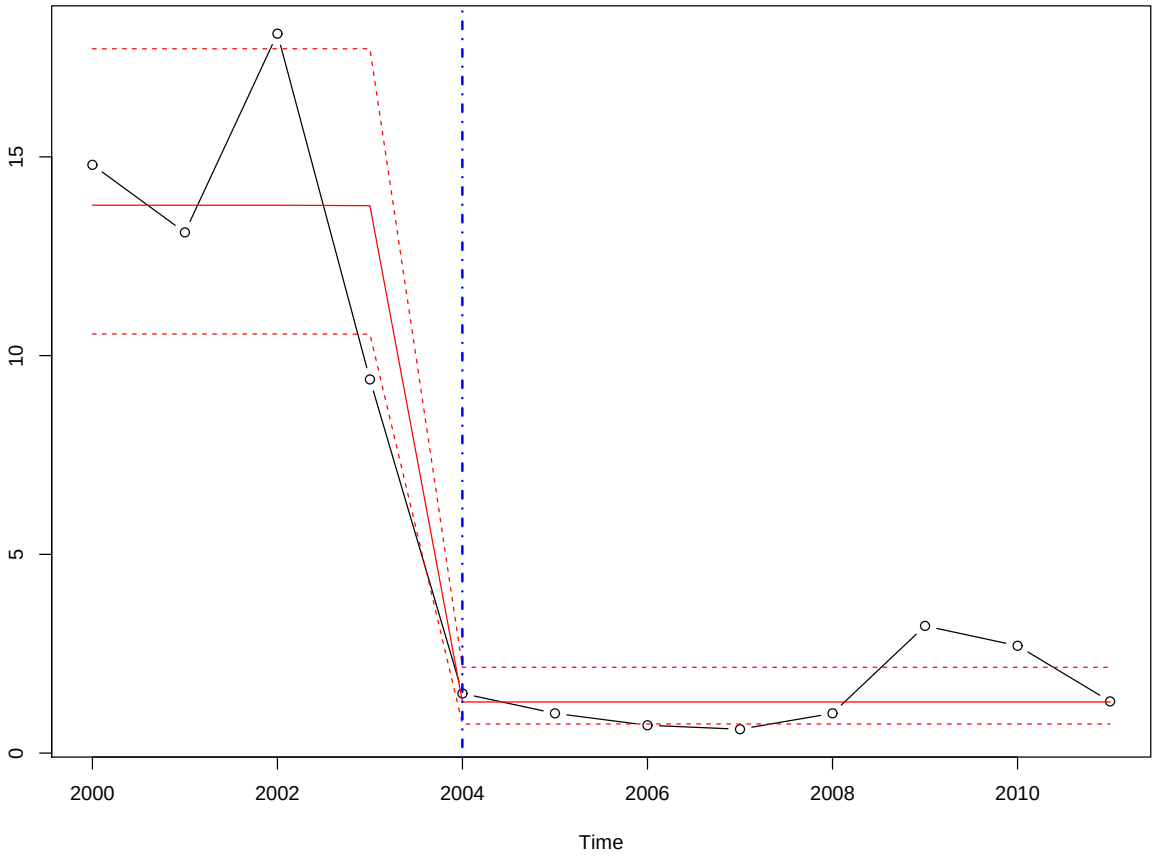

Figure 1.4: Number of fake $20 peso bills from 2000 to 2011. Empty dots (linked with solid black lines) represent observed values (F20). Red straight lines correspond to rate estimates (solid line) and 95% credible intervals (dotted lines). Vertical blue dashed-dotted line corresponds to year change estimate.

example, individual i could file a total of n_i number of claims during a year, with j denoting the specific claim. In this case, a mixed model for the claim amounts y_{ij}, in a generalized linear setting, would have linear predictor $\eta_{ij} = \boldsymbol{\beta}'\mathbf{x}_{ij} + \alpha_i$. Note that the parameter $\boldsymbol{\beta}$ is the fixed effect component common to all individuals, whereas α_i is a random effect common to all claims j made by the same individual i, and thus introducing a dependence in those claims made by the same individual.

Specifications for the random effects α_i's may vary according to the application. They could simply be $\alpha_i \overset{\text{iid}}{\sim} \text{N}(0, \tau)$, which is the typical specification in repeated measurements, clustered observations and longitudinal models. Alternative specifications include spatial effects $(\alpha_1, \ldots, \alpha_n) \sim \mathcal{CAR}(\rho, \tau)$, where $\mathcal{CAR}$ stands for a conditionally autorregressive model with association parameter ρ and precision τ. This model is a multivariate normal whose precision matrix is based on the spatial neighborhood structure. We refer the reader to Banerjee *et al.* (2004) for details. Or temporal effects $\alpha_i = \gamma\alpha_{i-1} + \nu_i$, with $\nu_i \overset{\text{iid}}{\sim} \text{N}(0, \tau)$, following a dynamic equation to account for dependence in time. We refer the reader to West and Harrison (1997) for details. In the following subsection we describe an alternative specification for the random effects that is based on the idea of exchangeability.

1.4.2 Hierarchical models

According to Gelman *et al.* (2004) hierarchical models are the most powerful tool for data analysis. Hierarchical specifications are usually helpful to specify a joint prior distribution for a set of parameters. However, they are also useful for specifying the distribution of random effects in a mixed model.

To describe the construction of a hierarchical model let us consider a simple scenario with response variables y_i, for $i = 1, \ldots, n$, where the distribution of each y_i depends on a parameter θ, that is $f(y_i|\theta)$. This scenario assumes that there is a unique parameter θ common to all individuals. So inference on θ will be based on a prior distribution $f(\theta)$ and all observations y_i's (as in a traditional Bayesian analysis with i.i.d. observations). On the other hand, a completely different specification of the problem would be to assume that the distribution of each individual i has its own parameter θ_i, that is $f(y_i|\theta_i)$. In this case, if we further take independent priors $f(\theta_i)$ for all $i = 1, \ldots, n$, inference on θ_i will only depend on its prior and the single observation y_i, like having n separate analysis. Hierarchical

models present a compromise between these two extreme scenarios by (i) allowing to have heterogeneity in the parameters by keeping a different θ_i for each y_i, and (ii) allowing to pull strength across different observations to increase precision in the estimation of the θ_i's.

We achieve (i) and (ii) by considering an exchangeable prior distribution for the vector $\boldsymbol{\theta}' = (\theta_i, \ldots, \theta_n)$. Exchangeability can be interpreted as a symmetric condition in the prior such that each θ_i has the same marginal distribution and the dependence among any pair (θ_i, θ_j) is the same. We achieve this symmetry in the prior with a two level hierarchical representation of the form:

$$\begin{array}{c} \theta_i|\psi \overset{\text{iid}}{\sim} f(\theta_i \mid \psi), \quad i = 1, \ldots, n \\ \psi \sim f(\psi) \end{array}$$

The parameter ψ is called hyper-parameter and plays the role of an anchor of the θ_i's. Conditional on ψ, the θ_i's are independent and when ψ is marginalized the θ_i's become dependent, i.e., $f(\theta_i, \ldots, \theta_n) = \int \prod_{i=1}^n f(\theta_i|\psi) f(\psi) \mathrm{d}\psi$. The hierarchical model is completed by specifying the distribution of the data, which in general would be $y_i \sim f(y_i|\theta_i)$ independently for $i = 1, \ldots, n$. This specification is analogous to the so called structure distribution that is frequently used in actuarial science, specifically in credibility theory.

Hierarchical models are particularly useful for meta-analysis, where information coming from different studios y_i is linked via a hierarchical prior distribution on the different parameters $(\theta_1, \ldots, \theta_n)$. Global or population inference from all studies is usually summarized in terms of the hyper-parameter ψ.

Example 14.4. Regarding the bills dataset of Table 1.3, consider now that the individuals i are the different bill denominations for $i = 1, \ldots, n$ with $n = 5$. For each bill denomination i we have $n_i = 12$ observations $j = 1, \ldots, n_i$ corresponding to the 12 years. For each observed number of fake bills y_{ij} we assume a Poisson model of the form $y_{ij} \sim \text{Po}(\mu_i)$ with $\log(\mu_i) = \beta_i$. Here we have two options, take independent priors for each β_i, say $\beta_i \sim \text{N}(0, 0.001)$ for $i = 1, \ldots, 5$, or take an exchangeable prior for the vector $(\beta_1, \ldots, \beta_5)$ with hierarchical representation given by $\beta_j|\beta_0, \tau \sim \text{N}(\beta_0, \tau)$ with $\beta_0 \sim \text{N}(0, 0.001)$ and $\tau \sim \text{Ga}(10, 1)$. Here the crucial parameter is τ. Since τ is a precision parameter for the β_i's, a small value would imply a large uncertainty and will allow a broad combination of information across different individuals i, whereas a large value reduces the uncertainty around β_0 and constraints the

Table 1.7: Bugs code for model of Example 14.4.

```
model{
#Likelihood for (i in 1:5){
for (j in
1:n){y[i,j]~dpois(mu[i])}
log(mu[i])<-b[i]
}
#Priors
for (i in 1:5){
b[i]~dnorm(0,0.001)
#b[i]~dnorm(b0,tau)
}
#b0~dnorm(0,0.001)
#tau~dgamma(10,1)
}
```

sharing of information across different i's. The prior we took for τ, Ga$(10, 1)$ is a slightly informative prior that allows a moderate sharing of information. The R (Bugs) code of this model is presented in Table 1.7.

Posterior estimates for the two prior choices, independent and hierarchical, are reported in Table 1.8 and presented in Figure 1.5. Numerical values are very similar when using both priors. If we concentrate in the point estimates of the μ_i's we can see that for those denominations with the smallest rates (F20 and F500) their point estimates increase when using a hierarchical prior with respect to those with independent priors. On the other hand, for those denominations with the largest rates (F50 and F100) their point estimates decrease. These effects are the result of sharing information between models and the estimates tend to compromise among all pieces of information, but at the same time respect the differences. An advantage of using a hierarchical prior is that the mean parameter of the coefficients, β_0, concentrates the population information coming from all parameters β_i. Posterior estimate of $\mu = e^{\beta_0}$, the population counterfeit rate, is reported in the last row of Table 1.8. This estimate is compared with that obtained from considering that all observations come from the same model, that is, $y_{ij} \sim$ Po(μ) with $log(\mu) = \beta$ and prior $\beta \sim$ N$(0, 0.001)$. The estimate of μ from this latter model is also included in the last row in Table 1.8 under the

Table 1.8: Posterior estimates of fake rates for different bill denominations under distinct scenarios.

		Independent		Hierarchical	
Coef.	Variable	Mean	95% CI	Mean	95% CI
μ_1	F20	5.62	(4.39, 7.06)	5.74	(4.47, 7.20)
μ_2	F50	140.16	(133.65, 147.10)	140.15	(133.70, 147.10)
μ_3	F100	103.57	(97.79, 109.40)	103.53	(97.85, 109.30)
μ_4	F200	98.58	(93.03, 104.05)	98.54	(92.96, 104.20)
μ_5	F500	61.17	(56.83, 65.68)	61.22	(56.81, 65.73)
μ	F	81.76	(79.54, 84.12)	58.05	(32.04, 96.06)

column of Independent prior. These two estimate show great differences. The model that assumes that all observations come from the same model with a single rate μ produces an interval estimate which is very narrow showing an enormous precision, whereas the interval estimate obtained with the hierarchical model aknowledges the uncertainty coming from the different denomination rates μ_i producing an overall counterfeit rate estimate for μ more realistic with a lot less precision. This effect can better be appreciated in the last pair of intervals in Figure 1.5.

1.5 Nonparametric regression

1.5.1 Bayesian nonparametric ideas

The concepts of parametric and nonparametric statistics refer to assumptions that are placed on the distribution of the available observations. One might assume that a particular dataset was generated from a normal distribution with unknown mean and precision. This is a parametric asumption since the $\mathrm{N}(\mu, \tau)$ defines a parametric family. In general, a parametric asumption would mean that the dataset is assumed to be generated from a member of a parametric family. Once a parametric asumption has been placed to a dataset, the objective is to estimate the unknown quantities, which is typically a finite number of parameters that define the model. A nonparametric asumption would imply that the dataset is not generated from a member of a particular parametric family, it is assumed to be generated from an unknown density (distribution) f (F). Since the whole f is unknown, we can say

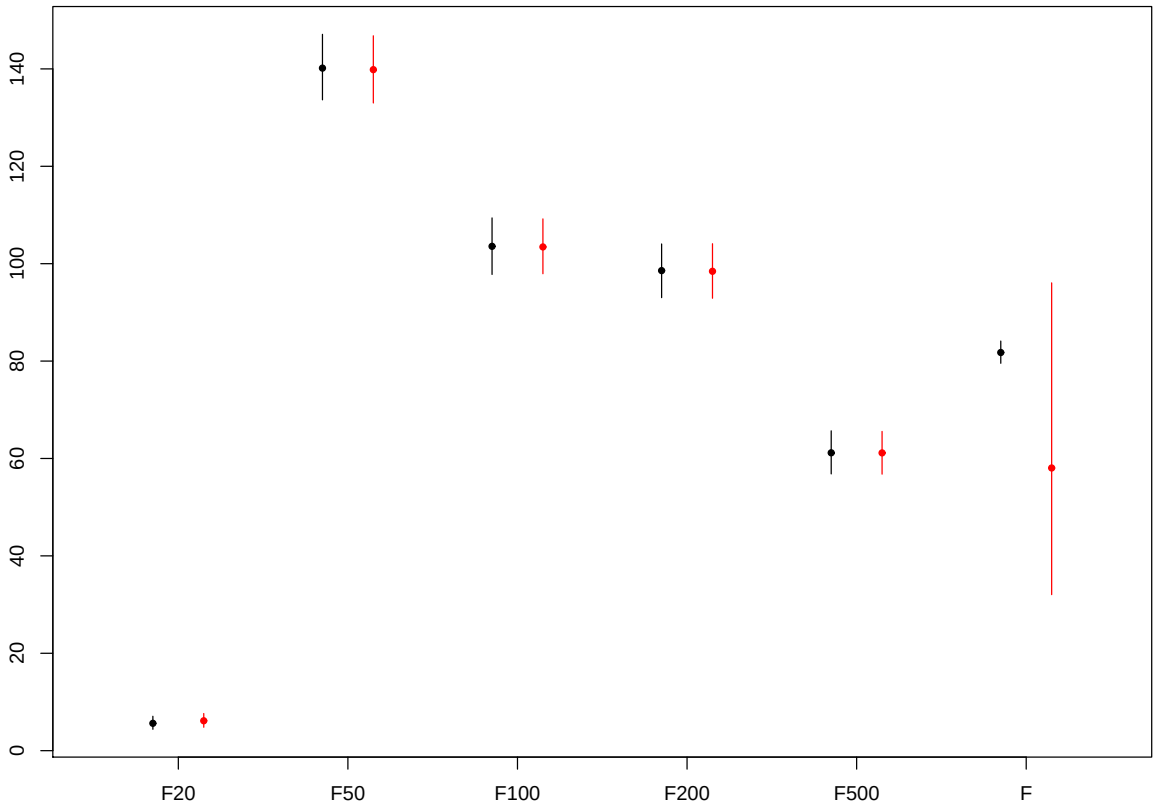

Figure 1.5: Posterior estimates of fake rates for different bill denominations. vertical lines correspond to 95% credible intervals and big dots to posterior means. Black (left) lines are obtained with independent priors and red (right) lines with hierarchical prior.

that the number of parameters to estimate is infinity, all $f(y)$'s values at any specific point y.

The way the Bayesian paradigm treats the unknown quantities is to determine a prior distribution and via Bayes Theorem update the prior knowledge with the information given by the data. In a nonparametric assumption the unknown quantities are the whole f (or F), so one is required to place a prior distribution on f. The way we achieve this is by using stochastic processes whose paths are density (or distribution) functions. This leads to the concept of nonparametric priors or random probability measures, since once a stochastic process has been chosen for f, any probability calculated from it, say $\mathrm{P}(Y \in B) = \int_B f(y)\mathrm{d}y$ is a random variable. According to Ferguson (1973) a nonparametric prior must have large support in the sense that any fixed probability distribution can be arbitrarily approximated by a realization from the prior.

1.5.2 Polya tree prior

One of the simplest and most powerful nonparametric priors is the Polya tree (Lavine, 1992). To start building the picture, let us consider a probability histogram. This is an estimate of a density function where the sampling space is partitioned in intervals and a bar is placed on top of each interval whose area represents the probability of lying in that particular interval. Imagine that the area assigned to each interval is a random variable such that the sum of all areas (random variables) is constrained to be one (almost surely), then this would be a realization of a (finite) Polya tree. This behaviour is illustrated in Figure 1.6. The formal definition of a Polya tree has been postponed to the Appendix 1.6.

1.5.3 Semiparametric linear regression

Let us recall the linear regresion model of Section 1.3, where the conditional distribution of the response variable y_i given a set of explanatory variables $\mathbf{x}_i$ is given by $y_i|\mathbf{x}_i \sim \mathrm{N}(\boldsymbol{\beta}'\mathbf{x}_i, \tau)$. If we consider the linear equation with an additive error $\epsilon_i \sim \mathrm{N}(0, \tau)$, the same model is re-expressed as $y_i = \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i$. This linear parametric model can be converted in a semiparametric model by relaxing the distribution f_ϵ of the errors ϵ_i to be nonparametric, for instance, a Polya tree. The normal model can be our prior mean and we can control how uncertain we are about the normal distribution of the errors by controling the precision

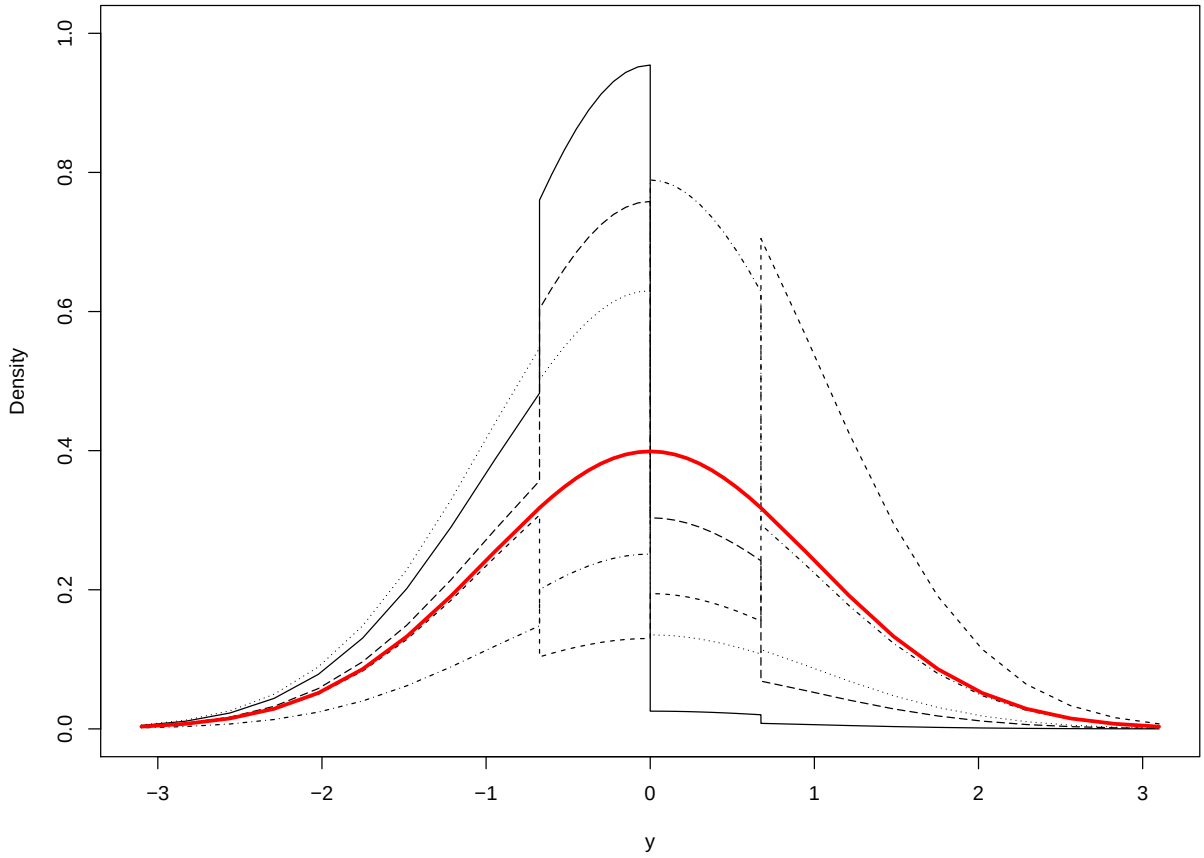

Figure 1.6: Five realizations (black thin lines) of a finute $\mathcal{PT}$ with $M = 2$ levels and centered in a $\mathrm{N}(0, 1)$ density (red thick line), with $a = 1$.

parameter a. For $a \to \infty$ we go back to the parametric normal regression model as a limiting case.

The semiparametric regression model is then defined as

$$y_i = \boldsymbol{\beta}'\mathbf{x}_i + \epsilon_i, \quad \epsilon_i \mid f_\epsilon \sim f_\epsilon, \quad f_\epsilon \sim \mathcal{PT}(\Pi, \mathcal{A}), \tag{1.4}$$

with $f_0 = \mathrm{N}(0, \tau)$ and $\alpha_{mj} = am^2$ and $a > 0$. A further adjustment needs to be done. Since $\mathrm{E}(f_\epsilon) = f_0$ this implies that $\epsilon_i \sim \mathrm{N}(0, \tau)$ marginally (on average). That is, not always $\mathrm{E}(\epsilon_i) = 0$, only on average. We can force the Polya tree to be centered at zero always (with probability 1) by fixing the first partition of the tree to be $B_{11} = (-\infty, 0]$ and $B_{12} = (0, \infty)$, and taking $\theta_{11} = \theta_{12} = 1/2$ with probability 1. This implies that the median of the random density f_ϵ of each ϵ_i is zero. This is verified by noting that the median of ϵ_i is zero iff $\mathrm{P}(\epsilon_i \leq 0) = \mathrm{P}(\epsilon_i \in B_{11}) = \theta_{11} = 1/2$. Therefore, the semiparametric regression model (1.4) is not a mean regression model but a *median regression model* since $\boldsymbol{\beta}'\mathbf{x}_i$ becomes the median of the response variable y_i.

Model (1.4) has two unknown quantities $\boldsymbol{\beta}$ and f_ϵ. Our prior knowldge on f_ϵ has been placed through the Polya tree, so we also require a prior distribution for $\boldsymbol{\beta}$. The common assumption is to take $\beta_j \sim \mathrm{N}(b_0, t_0)$, independently for $j = 1, \ldots, p$, as in most (generalized) regression models. The likelihood function for this semiparametric model is a function of the prior parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ as follows:

$$\mathrm{lik}(\boldsymbol{\beta}, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{x}) = \prod_{i=1}^{n} f(y_i|\mathbf{x}_i) = \prod_{i=1}^{n} f_\epsilon(y_i - \boldsymbol{\beta}'\mathbf{x}_i) = \prod_{i=1}^{n} \prod_{m} \theta_{m,j_{\epsilon_i}},$$

with $\epsilon_i = y_i - \boldsymbol{\beta}'\mathbf{x}_i$. Posterior distributions for $(\boldsymbol{\beta}, f_\epsilon)$ will be characterized conditionally. $f_\epsilon|\boldsymbol{\beta}, \mathbf{y}$ is another Polya tree with the distribution of the parameters $\boldsymbol{\theta}$ updates with ϵ_i's as observations. $\boldsymbol{\beta}|f_\epsilon, \mathbf{y}$ is just proportional to the product of the likelihood and the prior and Metropolis-Hastings steps will be required for sampling from it. Fortunately, posterior inference with this semiparametric model is implemented in the function `PTlm` from the R library `DPpackage`.

Example 14.5. Consider the insurance dataset described in Example 14.1. This dataset consisted of claim amounts y_i, premiums written x_i and class indicators z_{ji}, for $j = 1, \ldots, 7$ sectors and $i = 1, \ldots, 228$ observations. In Example 14.1 a linear model in the logarithmic

Table 1.9: R code for model of Example 14.5.

```
# Initial state
state<-NULL
# MCMC parameters
nburn<-500; nsave<-5000; nskip<-2; ndisplay<-500
mcmc<-list(nburn=nburn,nsave=nsave,nskip=nskip,ndisplay=ndisplay)
# Prior information
prior<-list(alpha=1,beta0=rep(0,14),Sbeta0=diag(1000,14),tau1=0.01,tau2=0.01,M=6)
# Fit the model
fit<-PTlm(formula=log(y)~z2+z3+z4+z5+z6+z7+log(x)+z2*log(x)+z3*log(x)+z4*log(x)
+z5*log(x)+z6*log(x)+z7*log(x),prior=prior,mcmc=mcmc,state=state,status=TRUE)
# Summary with HPD and Credibility intervals
summary(fit)
```

scale was suggested to describe the severity amounts in terms of the premiums written by sector. Let us consider the same linear predictor, but instead of assuming normality of the errors (responses) we will consider nonparametric errors with a Polya tree prior. The new model becomes $y_i = \alpha_1 + \sum_{j=2}^{7} \alpha_j z_{ji} + \beta_1 \log(x_i) + \sum_{j=2}^{7} \beta_j \log(x_j) * z_{ji} + \epsilon_i$, with $\epsilon_i | f_\epsilon \sim f_\epsilon$ and $f_\epsilon \sim \mathcal{PT}(\Pi, \mathcal{A})$.

The Polya tree is centered on $f_0 = \mathrm{N}(0, \tau)$, as in the parametric model, with $\alpha_{mj} = am^2$. We took $a = 1$ and assigned a prior to the error precision $\tau \sim \mathrm{Ga}(0.01, 0.01)$. It is worth mentioning that this latter prior induces a different specification of the Polya tree partitions $\Pi = \{B_{mj}\}$ for every value of τ, producing a mixing over the partitions and thus implying smoothed paths of the tree. We specify a finite tree with a number of partition levels $M = 6$. For the model coefficients we used the same priors as in Example 14.1, that is, $\alpha_j \sim \mathrm{N}(0, 0.001)$ and $\beta_j \sim \mathrm{N}(0, 0.001)$, for $j = 1, \ldots, 7$. The specifications for implementing this model in R with the use of the `DPpackage` library are presented in Table 1.9.

Remember that the semiparametric regression model just defined is a median regression model with an enhanced flexibility in the specification of the errors. Posterior estimates of model coefficients are included in the last two columns of Table 1.2. The point estimates of all coefficients are numerically different from those obtained with the parametric model, but only few of them are statistically different. Estimates of β_1 and β_7 present intervals that do not intercept between the parametric and the nonparametric scenarios, so implying

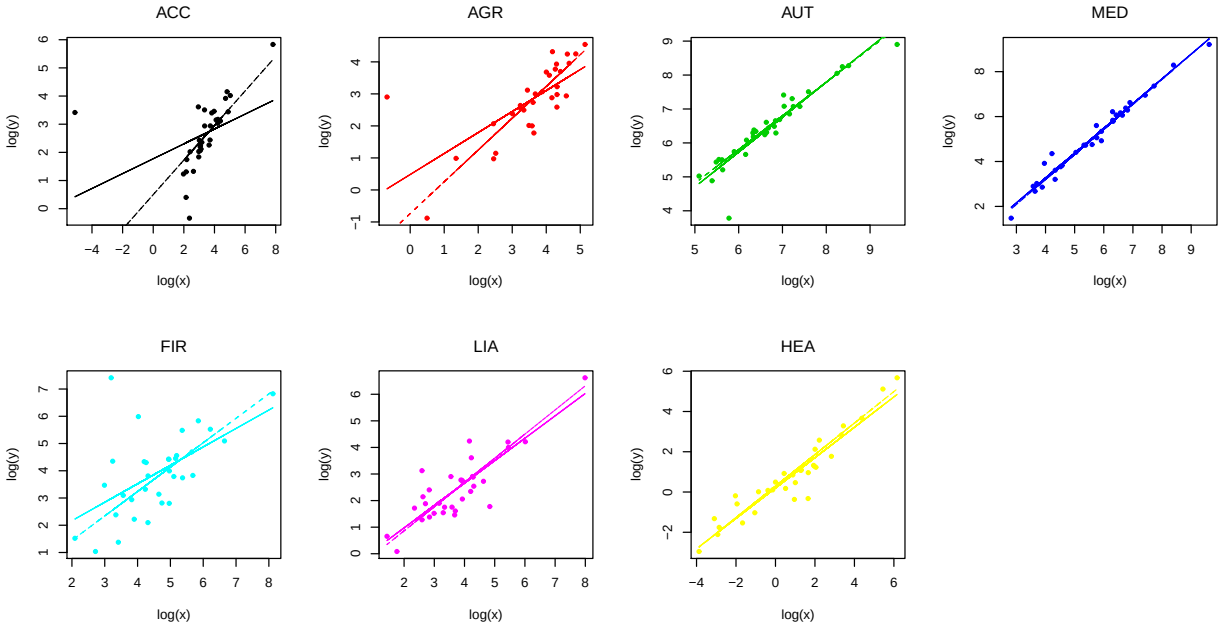

Figure 1.7: Dispersion diagrams of severity amounts y_i versus premiums written x_i by insurance class $j = 1, \ldots, 7$. Solid line (parametric fitting) and dotted line (nonparametric fitting).

a difference in the slope relationships between $\log(y_i)$ and $\log(x_i)$ for insurance sectors ACC and HEA comparing parametric and nonparametric fit. Figure 1.7 compares the paramertric and the nonparametric fittings for the seven sectors in different panels. The major differences among the two fittings are in the first two sectors ACC and AGR. In both cases there is presence of an extreme observation that pulls the parametric fit, whereas the nonparametric model is less sensitive to extreme observations producing a more realistic fit consistent with non-extreme observations.

1.6 Appendix

To formally define a Polya tree prior let $\Pi = \{B_{mj}\}$ be a set of binary nested partitions of $\mathbb{R}$ such that at level $m = 1, 2, \ldots$ we have a partition of $\mathbb{R}$ with 2^m elements and the index j, $j = 1, \ldots, 2^m$, identifies the element of the partition at level m. For example, at level one ($m = 1$), we have a partition of 2^1 elements B_{11} and B_{12}. At level two ($m = 2$) we have a partition of $2^2 = 4$ elements B_{21}, B_{22}, B_{23} and B_{24} such that (B_{21}, B_{22}) are a partition of

B_{11} and (B_{23}, B_{24}) are a partition of B_{12}. In general, at level m, B_{mj} is partitioned into $(B_{m+1,2j-1}, B_{m+1,2j})$ at level $m+1$ with $B_{m+1,2j-1} \cap B_{m+1,2j} = \emptyset$. Figure 1.8 presents a diagram of these nested partitions for levels $m = 1, 2, 3$.

Let $\boldsymbol{\theta} = \{\theta_{mj}\}$ be a set of parameters such that each θ_{mj} is associated to the set B_{mj}. The parameter θ_{mj} determines the conditional probability of a random variable Y being in the set B_{mj} given that it belonges to the father, $B_{m,(j+1)/2}$ if j is odd, or $B_{m,j/2}$ if j is even. For example, $\theta_{21} = \mathrm{P}(Y \in B_{21} | Y \in B_{11})$. Since the two subsets of a father set form a partition of the set, the conditional probabilities must sum to one. In the example, $\theta_{21} + \theta_{22} = 1$, where $\theta_{22} = \mathrm{P}(Y \in B_{22} | Y \in B_{11})$. In general $\theta_{m,2j} = 1 - \theta_{m,2j-1}$ for $j = 1, \ldots, 2^{m-1}$. Therefore, for the sets at level m, the probability of Y belonging to the set B_{mj} is just the product of all conditional probabilities θ_{mj}, one for each level, where the set B_{mj} belong to. In notation,

$$\mathrm{P}(Y \in B_{mj}) = \prod_{k=1}^{m} \theta_{m-k+1, r(m-k+1)},$$

where $r(k-1) = \lceil (r(k)/2) \rceil$is a recursive decreasing formula whose initial value is $r(m) = j$ and locates the set B_{mj} with its ancestors upwards in the tree. $\lceil \cdot \rceil$ denotes the ceiling function. For example, $\mathrm{P}(Y \in B_{21}) = \theta_{21}\theta_{11}$. If we continue the partitions down to infinity, we can define the density $f(y|\theta)$ foe every $y \in \mathbb{R}$ in terms of the parameters $\boldsymbol{\theta}$.

The Polya tree is then defined as the prior distribution for the density $f(y|\theta)$. Since $\boldsymbol{\theta}$ is an infinite set, then $f(y|\boldsymbol{\theta})$ is nonparametric (or infinitely parametric). Because θ_{mj} are (conditional) probabilities, they must be in the interval $(0, 1)$, so a natural prior is a beta distribution. Therefore $\theta_{mj} \sim \mathrm{Be}(\alpha_{m,j}, \alpha_{m,j+1})$. If we denote by $\mathcal{A} = \{\alpha_{mj}\}$ the set of all α parameters, then we can denote by $\mathcal{PT}(\Pi, \mathcal{A})$ a Polya tree prior for the density $f(y)$ or for the probability measure $\mathrm{P}(\cdot)$.

The Polya tree prior is defined in terms of the partitions Π and nonnegative parameters $\mathcal{A}$. These two sets must reflect our prior knowledge about the unknown density $f(\cdot)$. If we know that the true $f(\cdot)$ should be around a $f_0(\cdot)$ density, e.g. a $\mathrm{N}(0, 1)$ density, we can make the prior to satisfy $\mathrm{E}(f) = f_0$ in the following way (e.g. Hanson and Johnson, 2002): Take the partition elements B_{mj} to correspond to the dyadic quantiles of f_0, i.e.,

$$B_{mj} = \left[F_0^{-1}\left(\frac{j-1}{2^m}\right), F_0^{-1}\left(\frac{j}{2^m}\right) \right], \tag{1.5}$$

for $j = 1, \ldots, 2^m$, with $F_0^{-1}(0) = -\infty$ and $F_0^{-1}(1) = \infty$ and F_0 the distribution function

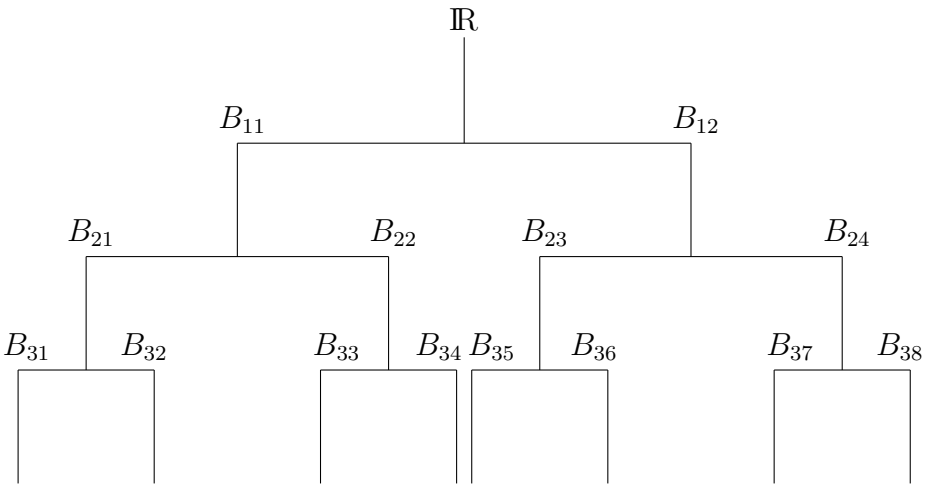

Figure 1.8: Diagram of nested partitions of $\mathbb{R}$ for three levels.

corresponding to density f_0; and take $\alpha_{mj} = a\, m^2$ (constant within each level m) such that $\theta_{m,2j-1} \sim \text{Be}(am^2, am^2)$ independently for $j = 1, \ldots, 2^{m-1}$. This particular choice of the α_{mj} parameters defines an almost surely continuous prior (Ferguson, 1974). The parameter a plays the role of a precision parameter, larger values of a make the prior to concentrate closer to the mean f_0, whereas smaller values make the prior to be "more nonparametric" since the prior will place larger variance around the mean f_0.

To better understand the Polya tree, Figure 1.6 presents five realizations of a finite Polya tree prior with a total of $M = 2$ levels, producing 4 elements partitioning the real line. These subsets were defined using the quartiles of a $\text{N}(0, 1)$ density as in (1.5). Since we stop partitioning at a finite level M, the density of the points inside the sets B_{Mj} needs to be spread, either uniformly (forming a histogram), or according to f_0, as in Figure 1.6 with $f_0 = \text{N}(0, 1)$. This is achieved by defining $f(y \mid \boldsymbol{\theta}) = 2^M f_0(y) \prod_{m=1}^{M} \theta_{m,j_y}$, where the pair (m, j_y) identifies the set B at level m that contains the point y. Each realization of a (finite) Polya tree corresponds to a "histogram" which results from a random perturbation of the centering density f_0 in the sets at level M, B_{Mj}.

Apart form the intuitive definition of a Polya tree, it has the advantage that its posterior representation is conjugate, following another Polya tree with updated parameters $\mathcal{A}$. For a

sample $y_1, \ldots, y_n$ of size n such that $y_i|f \sim f$ and $f \sim \mathcal{PT}(\Pi, \mathcal{A})$ then $f \mid \mathbf{y} \sim \mathcal{PT}(\Pi, \mathcal{A}^*)$ with $\alpha^*_{mj} = \alpha_{mj} + n_{mj}$ where $n_{mj} = \sum_{i=1}^{n} I(y_i \in B_{mj})$ is the number of observations y_i's that fall in the set B_{mj}.

Bibliography

Bailey, A. (1950). Credibility procedures laplace's generalization of bayes' rule and the combination of collateral knowledge with observed data. *Proceedings of the Casualty Actuarial Society*, **37**, 7–23.

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchicla modeling and analysis for spatial data*. Boca Raton: Chapman and Hall.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian theory*. New York: Wiley.

Bühlmann, H. (1967). Experience rating and probability. *ASTIN Bulletin*, **4**, 199–207.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**, 615–629.

Gelman, A., Carlin, J., stern, H., and Rubin, D. (2004). *Bayesian data analysis*. Coca Raton: Chapman and Hall.

Gelman, A., Jakulin, A., Grazia-Pittau, M., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, **2**, 1360–1383.

Hanson, T. and Johnson, W. (2002). Modeling regression error with a mixture of polya trees. *Journal of the American Statistical Association*, **97**, 1020–1033.

Klugman, S. (1992). *Bayesian Statistics in Actuarial Science: With Emphasis on Credibility*. Huebner International Series on Risk, Insurance, and Economic Security. Kluwer Academic Publishers.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling. *Annals of Statistics*, **20**, 1222–1235.

Lundberg, O. (1940). *On random processes and their application to sickness and accident statistics*. Almqvist and Wiksells, Uppsala.

Makov, U. E. (2001). Principal applications of bayesian methods in actuarial science: a perspective. *North American Actuarial Journal*, **5**(4), 53–73.

Makov, U. E., Smith, A. F. M., and Liu, Y.-H. (1996). Bayesian methods in actuarial science. *Journal of the Royal Statistical Society. Series D (The Statistician)*, **45**(4), 503–515.

Scollnik, D. (2001). Actuarial modeling with mcmc and bugs. *North American Actuarial Journal*, **5**(4), 96–125.

West, M. (1985). Generalized linear models: scale parameters, outlier accomodation and prior distributions (with discussion). In D. L. J.M. Bernardo, M.H. DeGroot and A. Smith, editors, *Bayesian Statistics 2*, pages 531–558.

West, M. and Harrison, J. (1997). *Bayesian forecasting and dynamic models*. New York: Springer.

Whitney, A. W. (1918). The theory of experience rating. *Proceedings of the Casualty Actuarial Society*, **4**, 274–292.