Bayesian Analysis of a Health Insurance Model

Helio S. Migon
Institute of Mathematics, Federal University of Rio de janeiro, migon@im.ufrj.br

Edison M.O. Penna
Centro de Eletr6nica da Marinha, emopenna@uol.com.br

Bayesian Analysis of a Health Insurance Model

Helio S. Migon,* and Edison M.O. Penna†

Abstract‡

We consider the problem of determining health insurance premiums based on past information on size of loss, number of losses, and size of population at risk. The size of loss and the number of losses are treated as mutually independent random variables. The number of losses is assumed to follow a Poisson process, and the loss sizes are independent and identically distributed non-negative random variables, and the population at risk is assumed to follow a non-linear growth model. An expression for the premium is obtained through maximization of the insurer's expected utility under a Bayesian model. The parameter estimation process is based on Monte Carlo Markov chain (MCMC). Our methology is applied to two real data sets.

Key words and phrases: *collective risk model, aggregate loss, rate making, predictive distribution, stochastic simulation, Monte Carlo Markov chain (MCMC)*

*Helio S. Migon, Ph.D., is a professor of statistics at the Institute of Mathematics, Federal University of Rio de Janeiro. He received his Ph.D. in statistics from University of Warwick, U.K. Dr. Migon is an associate editor of the *Brazilian Journal of Probability and Statistics* and of *Operation Research* (of the Brazilian Operational Research Society) and a former president of the ABE-Brazilian Statistical Association.

Dr. Migon's address is: Universidade Federal do Rio de Janeiro (UFRJ), Operational Research Section/COPPE, CP.: 68507, CEP 21941-972 - Rio de Janeiro RJ, BRAZIL. E-mail: *migon@im.ufrj.br*

†Edison M. O. Penna, M.Sc., is an engineer and head of the production department at Centro de Eletrônica da Marinha, a facility for logistic support. He was an officer in the Brazilian navy.

Mr. Penna's address is: Centro de Eletrônica da Marinha, Marinha do Brasil, Praça Barão de Ladário s/n Ilha das Cobras - AMRJ - Ed. 26, CEP.: 20.180-001, Rio de Janeiro RJ, BRAZIL.
E-mail: *emopenna@uol.com.br*

‡This research is partially supported by grants from the Brazilian Ministry of Science and Technology (Pronex), FAPERJ, and CNPq - Brazil. We acknowledge useful discussions with Mariana A. B. Santos (UFRJ) and Fernando S. Moura (UFRJ) and thank the referees for their many comments that have improved this paper.

1 Introduction

The main aim of a health insurance company is to generate enough premiums to cover losses due to expenses, medical payments (e.g., visits to physicians, diagnostic tests, physical therapy, and hospitalizations), and to produce profits. The premium charged for an individual health insurance contract is based on, among other factors, the insured person's age, health history, size of deductible, health plan chosen (Pai, 1997). Therefore pricing actuaries must use past information to develop a probabilistic model of the important uncertainties involved in the loss process.

In developing a health insurance model there will be many areas of uncertainty. For example, care must be taken to avoid adverse selection (where mainly unhealthy individuals are the predominant clients) and to resolve the conflicting interests of the doctor, the policyholder, and the insurer (e.g., in a fee-for-service plan, a doctor may seek unnecessary diagnostic tests to boost income while protecting against malpractice claims). Thus the insurer must establish a statistical control model to help to reduce unnecessary expenditures (Rosenberg, 2001). In spite of this, many actuarial models do not fully contemplate the uncertainties involved such as those due to parameter estimation (Migon and Gamerman, 1999).

The above problems associated with pricing health insurance can be dealt with in the Bayesian paradigm.[1] Regardless of the details of a particular model, the Bayesian approach requires that, before data are observed and the posterior distribution is evaluated, a prior distribution for the parameters involved in the premium calculation be specified. In specifying a prior distribution, there is plenty of room to incorporate expert opinions as well as to include industry-wide information. As mentioned in Hogg and Klugman (1984, page 14), "...actuaries are encouraged to introduce any sound a priori beliefs in the inferences whether Bayesian or not."

A Bayesian approach is adopted in this paper. We take fully into consideration all the uncertainty involved in determining the predictive distribution, which is the distribution of future observations conditional on observed values. Specifically we takes into consideration the uncertainty due to estimation of the parameters that form the basis for determining the premium. The premium is obtained via a maximization of the expected utility. The computations are done using

[1] Many authors have used a Bayesian approach to actuarial modeling, e.g., DuMouchell (1983); Herzod (1994); Makov, Smith, and Liu (1996); Haberman and Renshaw (1996); and Pai (1997).

WinBugs, i.e., Bayesian inference using the Gibbs sampler (Spiegelhalter et al., 2000). The major attractiveness of sampling-based methods is their conceptual simplicity and ease of implementation by users with available computing resources, not demanding any numerical analytic expertise. A review of some aspects of Bayesian data analysis in the context of actuarial models implemented and analyzed using Markov chain Monte Carlo techniques using WinBugs can be found in Scollnik (2001 and 2002) and Ntzoufras and Dellaportas (2002).

The remainder of this paper is organized as follows: the risk model and the Bayesian models are presented in Section 2, where the prior distributions are introduced and the estimation paradigm is presented. Alternative models and some numerical applications are presented in Section 3. Section 4 concludes with some remarks.

2 The Bayesian Framework

The basic insurance risk model used is the classic compound Poisson model that is commonly used in actuarial risk theory (e.g., Embrechts, Kluppelberg, and Mikosch, 1997). The model is briefly described as follows: Consider a single person insured for the unit time period $(t-1, t)$. Let N_t denote the number of losses and X_t denote and the aggregate loss produced by this person for $t = 1, \cdots, T$. It follows that

$$X_t = \sum_{j=1}^{N_t} Z_{t,j}$$

where $Z_{t,j}$ is the amount of the j^{th} loss in $(t-1, t)$. The main assumptions of the classic compound Poisson model are (i) the number of losses produced by this person in any interval is a Poisson process with rate λ, (ii) the loss sizes are independent and identically distributed (i.i.d.) non-negative random variables, and (iii) the number and size of losses are mutually independent. Clearly N_t is a Poisson random variable with mean λ and the $Z_{t,j}$s are i.i.d. Specifically we assume the $Z_{t,j}$s are exponentially distributed with finite mean $1/\theta$.

Next we consider a portfolio of such insured persons. Let $\pi_{a,t}$ denote the number of persons age a who are insured in the time interval $(t-1, t)$. We assume that $\pi_{a,t}$ has a normal distribution with mean $\mu_{a,t}$ and variance σ_a^2. This normal assumption is for simplicity and ease of computations (Migon and Gamerman, 1993).

To summarize, our model is mathematically described as follows: Consider a health insurance portfolio consisting of persons of various ages who are placed in one of c different age classes labeled from 1 to c. Let $N_{a,t}$ and $X_{a,t}$ denote the number of losses and the aggregate loss, respectively, produced by the insured persons age a in the period $(t-1, t)$. If $N_{a,t}$ is Poisson with mean $\lambda_a > 0$ and the losses are i.i.d. exponential variables with mean $1/\theta_a$

$$[N_{a,t} | \lambda_a, \pi_{a,t}] \sim \text{Poisson with mean } \pi_{a,t}\lambda_a \tag{1}$$

$$[X_{a,t} | N_{a,t} = n_{a,t}, \theta_a] \sim \text{Gamma}(n_{a,t}, \theta_a) \tag{2}$$

where Gamma(α, β) denotes the pdf of a gamma distribution with mean α/β and variance α/β^2

$$\pi_{a,t} \sim \text{N}(\mu_{a,t}, \sigma_a^2) \tag{3}$$

for $t = 1, 2, \ldots$ and $a = 1, 2, \ldots, c$. The population model model used is

$$\mu_{a,t} = \left(\beta_{a,0} + \beta_{a,1}e^{-\beta_{a,2}t}\right)^{1/\phi}, \tag{4}$$

where ϕ is usually chosen as 1, -1, or 0, corresponding to the modified exponential, logistic, and Gompertz growth models (Migon and Gamerman, 1993). Of course, the case of $\phi = 0$ must be viewed as the limit when ϕ tends to zero and corresponds to the logarithm of $\mu_{a,t}$. In this paper only the logistic time evolution model ($\phi = -1$) will be taken into consideration. It is worth noting that the model in equation (3) can be interpreted as a non-linear regression with time as the explanatory variable, for each age class.

Let us suppose that some past information is available for the time periods $(t-1, t)$ and the information is in the form $(\boldsymbol{n}_t, \boldsymbol{x}_t, \boldsymbol{\pi}_t)$, for $t = 1, \cdots, T$, where $\boldsymbol{n}_t = (n_{1,t}, \ldots, n_{A,t})^\intercal$ with $n_{a,t}$ representing the observed number of losses, $\boldsymbol{x}_t = (x_{1,t}, \ldots, x_{A,t})^\intercal$ with $x_{a,t}$ representing the observed aggregate loss, and $\boldsymbol{\pi}_t = (\pi_{1,t}, \ldots, n_{A,t})^\intercal$ with $\pi_{a,t}$ representing the observed number of insureds in age class a in time $(t-1, t)$. The type of health care service will not be taken into account just to keep the notation simple.

The main concern at this stage is to obtain the full predictive distribution of the total loss for each age class a, X_a at time $T+h, h = 1, \cdots, H$, where H is the given planning horizon. To be more specific we need to obtain the distribution of $X_{a,T+h}$ given all the available information. The total loss amount up to the time horizon $T+H$ is obtained as an aggregation over the age classes and the time horizon, given:

$$X_{T+H} = \sum_{a=1}^{c} \sum_{h=1}^{H} X_{a,T+h}. \tag{5}$$

This will be the key quantity used to define the premium, meaning that the premium will be quoted today to cover all future losses incurred.

Assuming that the total size of losses in age class a and at time t are independent given $n_{a,t}$ and θ_a, and that the number of losses is independent of time and age, given the population $\pi_{a,t}$ and λ_a, the likelihood function follows as:

$$l(\boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2 | \boldsymbol{D}_T) \propto \prod_{a=1}^{c} \prod_{t=1}^{T} \frac{\theta_a^{n_{a,t}} \lambda_a^{n_{a,t}}}{(\sigma_a^2)^{1/2}} \times \exp\left[-\theta_a x_{a,t} - \lambda_a \pi_{a,t} - \frac{(\pi_{a,t} - \mu_{a,t})^2}{2\sigma_a^2} \right] \tag{6}$$

where $\boldsymbol{D}_T = \{(\boldsymbol{x}_t, \boldsymbol{n}_t, \boldsymbol{\pi}_t), t = 1, \cdots, T\}$ represents all data available, $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ are $A \times 1$ vectors, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_a, \cdots, \boldsymbol{\beta}_A)^{\intercal}$, where $\boldsymbol{\beta}_a = (\beta_{a,0}, \beta_{a,1}, \beta_{a,2})'$ is a vector describing the insured population time evolution in the a^{th} age class.

In order to conduct a Bayesian analysis, one needs to define a prior distribution over the parameter space. A proper prior distribution will be adopted with the assumption of independence among the parameters in each age class:

$$p(\theta_a, \lambda_a, \boldsymbol{\beta}_a, \sigma_a^{-2}) = p(\theta_a) p(\lambda_a) p(\boldsymbol{\beta}_a) p(\sigma_a^{-2}), \tag{7}$$

which is a non-structured prior distribution for the parameters of the model of equations (1) to (3) for each age class. Alternatively a hierarchical prior (Moura and Migon, 2002 and Migon and Moura, 2005) could be introduced to borrow strength from the age class.

Although in many applications the prior distribution is carefully elicited by the research (for example Garthwaite, Kadane, and O'Hagan, 2004), in this paper the hyperparameters are chosen in such a way that a relatively non-informative but proper prior is implied. The use of an improper prior in general can cause problems such as inability to evaluate meaningful Bayes factors or even the lack of existence of the posterior distribution itself, as mentioned by Gelfand (1995, Chapter 9, page 148).

It is natural to model θ_a as Gamma(A_a, B_a), where $A_a, B_a > 0$ are chosen in such a way that the prior is sufficiently vague. As $E[\theta_a] = A_a/B_a$ and $Var[\theta_a] = A_a/B_a^2$, the values of (A_a, B_a) can be easily obtained. A conjugate prior distribution proposed for λ_a is also gamma with parameters $C_a, D_a > 0$. Those quantities can be assessed as described before. Finally the prior distribution for the regression coefficients in the time evolution of the population mean is

$$\boldsymbol{\beta}_a \sim \mathrm{N}(\mathbf{0}, \tau_0^2 \boldsymbol{I}) \tag{8}$$

where $\boldsymbol{I}$ is the 3×3 identity matrix and

$$\sigma_a^{-2} \sim \mathrm{Gamma}(n_0/2, n_0 s_0^2/2) \tag{9}$$

$n_0, s_0^2 > 0$ for all age classes.

The next step is to combine the likelihood function of equation (6) and the joint prior distribution in equation (7) to obtain the joint posterior distribution. Unfortunately the joint posterior distribution is not available in a closed form, so a Monte Carlo Markov chain (MCMC) sampler is employed to generate drawings from this distribution. The method used to make inferences about the parameters is the Gibbs sampler, which is a MCMC scheme where the transition kernel is formed by the full conditional distributions (Gamerman and Lopes, 2006). Roughly speaking, it consists of generating sequential drawings from the full conditional posterior distributions. The relevant issue related to MCMC is to ensure the empirical distribution of the parameters has achieved its limit distribution (Gamerman and Lopes, 2006). The posterior distribution of any quantity of interest (i.e., any function of the parameters) can easily be obtained in the MCMC process. As our main interests are in (i) the total cumulative loss for the planning horizon, (ii) the future values of the loss number and size, and (iii) the insured population's evolution, they can be jointly generated via the MCMC algorithm.

As can be observed, the full conditional distributions are available in a closed form for all parameters, except $\boldsymbol{\beta}_a$. For these parameters the Gibbs sampler can easily be implemented. The full conditional posterior distribution of λ_a is given as

$$[\lambda_a | \boldsymbol{\theta}, \boldsymbol{\beta}_a, \sigma_a^2, D_T] \sim \mathrm{Gamma}(C_{a,1}, D_{a,1}), \tag{10}$$

where

$$C_{a,1} = \sum_{t=1}^{T} n_{a,t} + C_a \quad \text{and} \quad D_{a,1} = \sum_{t=1}^{T} \pi_{a,t} + D_a. \tag{11}$$

From equations (6) and (7) the full conditional posterior distribution of θ_a, is

$$[\theta_a | \boldsymbol{\lambda}, \boldsymbol{\beta}_a, \sigma_a^2, D_T] \sim \mathrm{Gamma}(A_{a,1}, B_{a,1}) \tag{12}$$

where

$$A_{a,1} = \sum_{t=1}^{T} n_{a,t} + A_a \quad \text{and} \quad B_{a,1} = \sum_{t=1}^{T} x_{a,t} + B_a. \tag{13}$$

As the full conditional posterior distributions for $\boldsymbol{\beta}_a$s are not available in closed form, a Metropolis within Gibbs algorithm is used to successively sample from the full conditional posterior distribution for the $\boldsymbol{\beta}_a$s as implemented in WinBugs. Finally, to complete the inference steps, the predictive distribution for $(\boldsymbol{n}, \boldsymbol{x}, \boldsymbol{\pi})_{T+h}$ is obtained from equations (1) and (3), conditional on the parameters generated as described before.

3 A Practical Application

The model described in Section 2 will be applied to two data sets consisting of the experience of two relatively new small self administered Brazilian health care plans called the Northeast Health Company (NHC) and Southeast Health Company (SHC). The data sets consist of monthly observations on the number of losses, the aggregate of the observed losses, and the number of insured individuals for each age class, i.e., $(\boldsymbol{n}_t, \boldsymbol{x}_t, \boldsymbol{\pi}_t)$). The data from SHC consists of 15 monthly observations (from March 1997 up to February 1998), while the NHC data consists of 23 monthly observations (from August 1998 to June 2000). The age classes used are: age class 1 is age 0 to 18, age class 2 is age 18 to 35, age class 3 is age 35 to 45, age class 4 is age 45 to 55, age class 5 is age 55 to 65, age class 6 is age 65 to 75, and age class 7 is age 75 and over. Tables A1, A2, and A3 in the Appendix show the monthly aggregate losses, number of losses, and population size for each age class for service 1 for Northeast Health Company and Southeast Health Company.

3.1 Premium Estimation Methods

Three different methods are used to determine the risk-loaded premium: one is based on classical assumptions and the other two are Bayesian in nature. For the Bayesian approaches, the predictive distribution over a planning horizon is considered, which implies that all uncertainties involved in the insurance business are included. The first method relies on standard asymptotic approximations for equation (5), the second method is a special case of the full Bayesian model called the semi-predictive approach, while the third method is the full Bayesian model. In each method, however, the premium is defined as the 97.5% percentile of the predictive distribution of the aggregate loss.

These three methods will now be discussed in detail: The **first** method uses the well known expression for the mean and variance of X (see, for example, Bowers et al., 1997, Chapter 12):

$$\mathbb{E}[X] = \mathbb{E}[N]\mathbb{E}[Z] \quad \text{and} \quad \mathbb{V}\text{ar}[X] = \mathbb{E}[N]\mathbb{V}\text{ar}[Z] + (\mathbb{E}[Z])^2\mathbb{V}\text{ar}[N]$$

where Z represents one of the i.i.d. random variables characterizing the loss value. The parameters λ and θ involved in the mean and variance are estimated via maximum likelihood as:

$$\hat{\lambda}_a = \frac{\sum_{t=1}^{T} n_{a,t}}{\sum_{t=1}^{T} \pi_{a,t}} \quad \text{and} \quad \frac{\hat{\theta}_a = \sum_{t=1}^{T} n_{a,t}}{\sum_{t=1}^{T} x_{a,t}}. \tag{14}$$

The estimates of $\mathbb{E}[X]$ and $\mathbb{V}\text{ar}[X]$ then follow from the invariance properties of the maximum likelihood estimator (Migon and Gamerman, 1999).

The **second** method, which we call the semi-predictive method, is closely related to the non-compound collective model (Gómez-Déniz et al., 1999) that assumes only the number of losses is stochastic. This is a very useful practical simplifying assumption because in practice the prices of most medical services are often negotiated between the insurer and the provider and are set in advance for the period. Under fixed price conditions, the posterior and predictive distributions can be developed in a closed form. Recalling that in the model that $N_{a,t}|\lambda_a$ has a Poisson with mean $\lambda_a\pi_{a,t}$, where $\pi_{a,t}$ is known, and the prior distribution of λ_a is Gamma(C_a, D_a), then the posterior and predictive distributions are

$$\begin{aligned} \lambda_a|\boldsymbol{D}_T &\sim \text{Gamma}(C_{a,1}, D_{a,1}) \quad \text{and} \\ N_{T+h}|\boldsymbol{D}_T &\sim \text{NBin}(C_{a,1}, \frac{D_{a,1}}{D_{a,1} + \pi_{a,T+h}}) \end{aligned} \tag{15}$$

where NBin(n, θ) represents the probability function of a negative binomial distribution with mean $n(1-\theta)/\theta$ and variance $n(1-\theta)/\theta^2$, $C_{a,1} = C_a + \sum n_{a,t}$, and $D_{a,1} = D_a + \sum \pi_{a,t}$ (Migon and Gamerman, 1999, page 249). Using a square error loss function, the best point estimator is then the posterior mean and the Bayes risk corresponds to the variance of λ_a, given by:

$$\tilde{\lambda}_a = E[\lambda_a|\boldsymbol{D}_T] = C_{a,1}/D_{a,1} \quad \text{and} \quad \mathbb{V}\text{ar}[\lambda_a|\boldsymbol{D}_T] = C_{a,1}/D_{a,1}^2. \quad (16)$$

Note that if the insured population is allowed to evolve over time, then computationally intensive procedures must be used to make the inferences.

The **third** method, which is the full Bayesian model, corresponds to the model described by equations (1), (2), and (3). The inference in this case necessarily needs the implementation of an MCMC algorithm because the posterior and predictive distributions for the quantities of interest are not available in closed form.

A useful tool in the early stages of model building is a directed acyclic graphic (DAG), which is also called an influence diagram. DAGs are useful in determining the full conditional distributions involved in MCMC schemes (Gilks et al., 1995). In fact, one can obtain WinBugs code from a DAG. In a DAG quantities of interest are represented as nodes and arrows run into nodes from their direct influences. A double arrow represents stochastic dependence, while a single one denotes a functional relationship. There are two types of nodes: those representing known deterministic quantities (square symbols) and those representing stochastic (circles). Recall that the number of losses in age class a at time t, $N_{a,t}$, is a random quantity with parametric distribution depending on the expected number of losses per policyholder, λ_a. If the insured population is given, the posterior and predictive distributions can be obtained in closed form. For example, Figure 1 is a representation of the collective risk model under the semi-predictive approach, i.e., in Figure 1(a) $\pi_{a,t}$ is known while in Figure 1(b) it is stochastic. From Figure 1(a) we see just how simple the DAG is in this case. Assuming the insured population is also a random quantity characterized by a mean $\mu_{a,t}$ and a precision σ_a^{-2}, as stated in equation (3), the MCMC method is needed to make the inferences feasible.

Figure 2 shows the DAG obtained for the full Bayesian model. The DAG presented in Figure 2(b) is useful in determining the full conditional distributions involved in the Gibbs sampler scheme (Gilks et al., 1995) assuming the population evolution is unknown, while Figure 2(a) represents the full predictive model assuming the population evolution is known. Even in this simple case the posterior and predictive distributions are not obtained in closed form. Note that Figure 2(b) includes the components of the insured population evolution described by a generalized linear regression model.

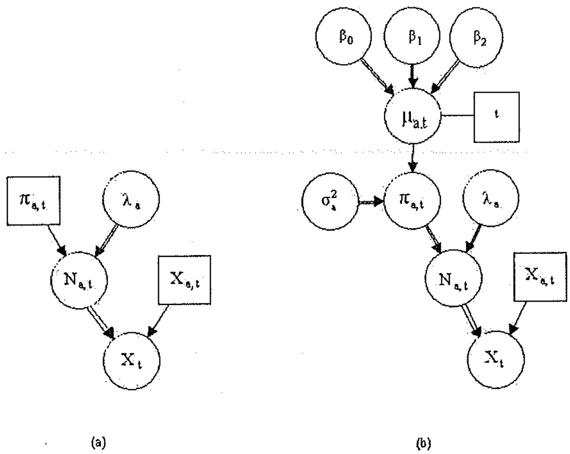

Figure 1: DAG Under the Semi Predictive Approach. *Notes*: This is an influence diagram for the number and size of losses assuming that the evolution of the insured population is: (a) known and (b) unknown.

3.2 The Calculated Premiums

The convergence of the MCMC process was assessed by different criteria proposed in the literature, thus assuring that the results presented are reliable (e.g., Gamerman and Lopes, 2006, Chapter 5). Some statistical tests were done in order to assess the convergence of the Gibbs sampler sequences. It is worth mentioning that based on three chains with 2,500 runs, including a burn-in of 500, the Gelman and Rubin (1992) criterion exhibits convergence before 1000 iterations were drawn. The convergence was also confirmed by many other graphical outputs. Nevertheless, the results presented in this paper are obtained by pooling over the three chains, corresponding to the final 6,000 draws. The predictive density of the total loss value obtained under the assumption of a full predictive Bayesian model shows some evidence that these distributions are asymmetric to the right, at least for the SHC.

The assumptions used in the model's development were consistent with the data sets analyzed. For example the coefficients in the population evolution model, equation (4), are all significantly different from zero. Also the assumption of independent and exponentially distributed loss value are confirmed by the goodness of fit of the assumed composed Poisson model. The predictive distributions obtained via the MCMC method are clearly non-symmetric, confirming that asymptotic normality is not appropriate. Although our models assume that the in-

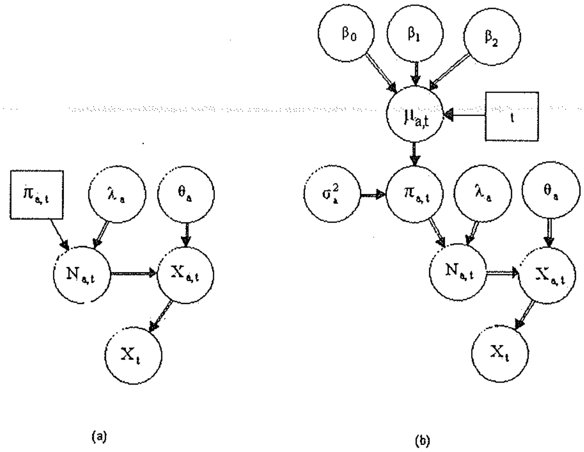

Figure 2: DAG Under the Full Predictive Approach. *Notes*: This is an influence diagram for the number and size of losses assuming that the evolution of the insured population is: (a) known and (b) unknown.

tensity of losses is age dependent, the proposed prior is not structured, i.e., the age classes are considered to be independent. The data show the loss intensities and the expected value of losses are age dependent.

A summary of the predictive distribution, which is useful for setting premiums, is presented in Table 1. The main results obtained are based on $T = 15$ and $T = 23$ monthly data sets for the SHC and NHC, respectively. It is worth pointing out that the global pure premiums are almost the same, although the global premiums (the 97.5% percentile of the predictive distribution) are quite different.

The individual premium for SHC and NHC are compared on a monthly basis in Table 2. All figures were obtained based on, respectively, the 12 and 23 month experience of the SHC and NHC and are quoted in U.S. dollars. The premium presented corresponds to the 97.5% percentile of the predictive distribution, accumulated over a three month horizon ($H = 3$), in the Bayesian model and in the normal approximation. The choice of the 97.5% percentile as the premium corresponds to the maximization of the expected value of a very particular utility function, called the modified absolute deviation (Moura and Migon, 2002). The classical and the semi-predictive models differ only slightly. The reason could be that neither take into account all the variability involved. The full Bayesian model, in turn, presents a bigger premium value than the previous methods for all age classes. This must be a consequence of the asymmetry of the predictive distribution and also of the consider-

Table 1
Summary of Predictive Distribution
Over a Three Month Horizon (H = 3)

Variable	Mean	2.5%	Median	97.5%
Based on SHC Data:				
Premium	45.45	30.93	44.58	64.96
Number of Losses	1845	1663	1843	2033
Size of Losses (in 1,000s)	230.10	156.00	225.80	327.10
Insured Population	5064	4801	5062	5350
Based on NHC Data:				
Premium	45.67	42.51	45.64	48.98
Number of Losses	31490	29840	31460	33180
Size of Losses (in 1,000s) (10^3)	465.10	440.40	464.70	490.50
Insured Population	10190	9646	10180	10780

ation of all the uncertainties involved. The last column corresponds to an equivalent monthly premium that is constant for all age classes: the global premium. Of course these figures correspond to the premium without the administrative cost and the insurer's profits.

The results obtained are not surprising. The risk premium increases almost steadily, from the age class 2 up to 7. The global premium is around \$45 in the NHC and \$60 dollars in the SHC, which seems quite reasonable and a little bit cheaper than the prices they charge in the market. This difference in the global premium is expected, because medical care is in general cheaper in the Northeast.

Figure 3 shows the posterior mean for λ_a, the expected monthly number of physician consultations per insured, for each age class ($a = 1, \cdots, 7$), based on the 15 data points available for the SHC. The expected number of losses increases with age, which is not a surprise. For example, the number of physician consultations is around 0.20 per month in age class 45⊢55 (age class 4), increasing to 0.25 in age class 65⊢ 75 (age class 6), representing an annual expected rate of 3 and 4 visits per year respectively, which seems reasonable.

4 Concluding Remarks

We discussed the implementation of the collective risk model in a Bayesian setting using stochastic simulation techniques. A practical

Table 2
Individual Premium Comparison for the SHC and NHC

	Age Class							$P^{(G)}$
	1	2	3	4	5	7	7	
Based on SHC Data:								
$P^{(N)}$	17.35	12.02	21.28	38.93	98.10	82.90	335.04	60.40
$P^{(S)}$	19.08	10.86	19.61	35.91	106.9	99.37	341.0	50.47
$P^{(F)}$	30.53	14.84	25.7	49.54	158.1	120.3	475.0	64.96
Based on NHC Data:								
$P^{(N)}$	15.40	23.54	28.00	39.69	30.66	58.18	83.73	31.86
$P^{(S)}$	24.35	39.25	45.07	47.71	56.51	77.86	93.67	45.90
$P^{(F)}$	29.51	45.88	52.3	56.69	66.27	92.56	106.1	48.98

Notes: $P^{(N)}$ denotes the normal approximation, $P^{(S)}$ denotes the semi predictive model, $P^{(F)}$ denotes the full predictive model, and $P^{(G)}$ denotes the global premium.

example was provided using two small data sets taken from the claims experience of two small Brazilian health care plans.

The stochastic simulation techniques used make the inferences almost straightforward. The implementation of these models in WinBugs is extremely simple, and the computing time is almost insignificant. Our main recommendation is to use the full Bayesian model. Given the asymmetry of the loss distribution, the assumption of asymptotic normality should be avoided.

The full Bayesian model described in equations (1) to (5) could be extended in many directions. For example, the population evolution could be modeled via generalized growth curves as in Migon and Gamerman (1993). This is a very broad class of growth models including the logistic and Gompertz as special cases. Keep in mind that the main goal is to input a structured prior to contemplate the possibility of an exchangeable structure among age classes. The same could be true for the other parameters in the model, such as the claim intensity and the expected value of each claim.

Other extensions that deserve some comments are to consider different distributions for the claim amounts and to allow the portfolio to be composed of dependent risks (Goovaerts and Dhaene, 1996). Censoring and truncation could play an important role when deductibles and policy limits are included in the model (Pai, 1997). Some of the extensions proposed here are considered in Moura and Migon (2005).

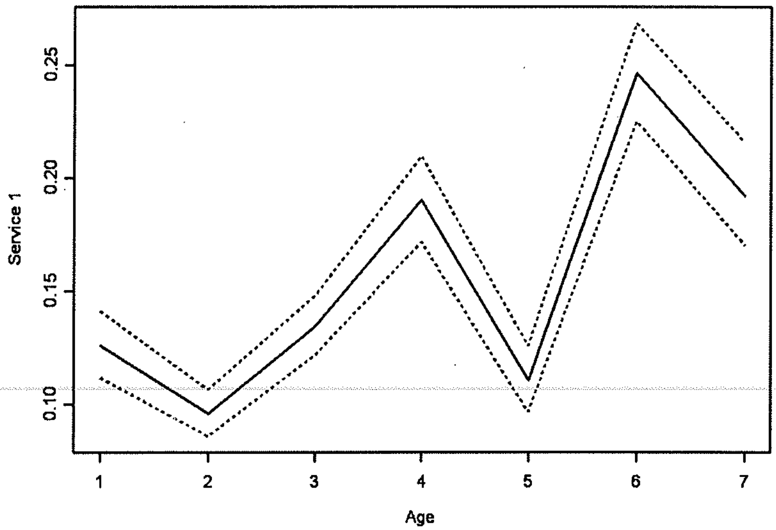

Figure 3: The Posterior Mean of Physician Consultations, λ_a, (Bold Line) and 95% Bayesian Confidence Interval of λ_a (Dotted Lines) for Various Age Classes Using SHC Data

References

Bowers, N.L., Gerber, H.U., Hickman, J.C., Jones, D.A., and Nesbitt, C.J. *Actuarial Mathematics.* Schaumburg, IL: The Society of Actuaries, 1997.

DuMouchel, W.H. "The 1982 Massachusetts Automobile Insurance Classification Scheme." *The Statistician* 32 (1986): 69–81.

Embrechts, P., Kluppelberg, C., and Mikosch, T. *Modelling Extremal Events for Insurance and Finance.* Berlin, Germany: Springer Verlag, 1997.

Gamerman, D and Lopes, H.F. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* London, United Kingdom: Chapman Hall, 2006.

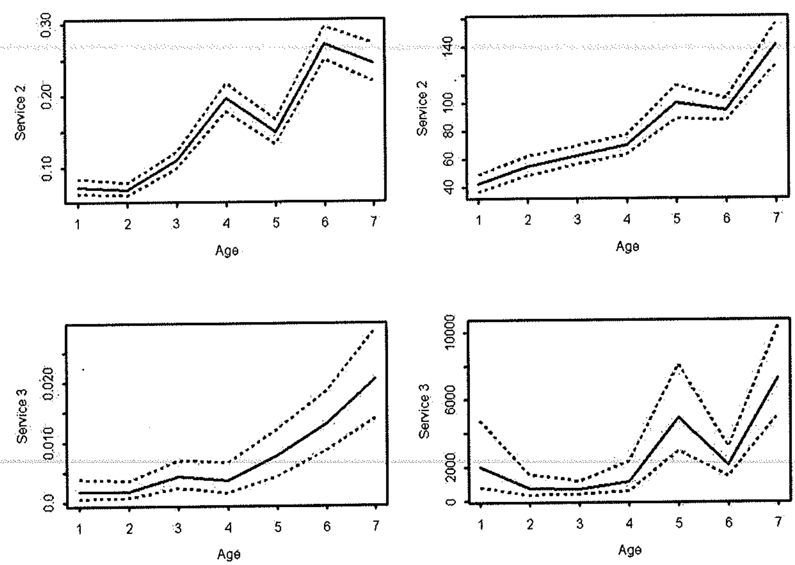

Figure 4: The Posterior Mean and 95% Bayesian Confidence Interval of λ_a and $1/\theta_a$ for Various Age Classes Using SHC Data. *Notes*: Top left figure shows λ_a (bold line) and its 95% Bayesian confidence interval (dotted lines) while top right figure shows $1/\theta_a$ (bold line) and its 95% Bayesian confidence interval (dotted lines)for service 2. Similarly, bottom left figure shows λ_a (bold line) and its 95% Bayesian confidence interval (dotted lines) while bottom right figure shows $1/\theta_a$ (bold line) and its 95% Bayesian confidence interval (dotted lines)for service 3.

Garthwaite, P.H., Kadane, J.B., and O'Hagan, A. "Statistical Methods for Eliciting Probability Distributions." *Journal of the American Statistical Association* 100 (2005): 680–701.

Gelfand, A.E. "Model Determination Using Sampling Methods." In *Markov Chain Monte Carlo in Practice*. (eds. W.R. Gilks, S. Richardson, and D.J. Spiegelhalter.) London, United Kingdom: Chapman-Hall, 1995: 145–162.

Gelman, A. and Rubin, D. B. "Inference from Iterative Simulation Using Multiple Sequences.", *Statistical Science* 7, no. 4 (1992): 457–511.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. *Markov Chain Monte Carlo in Practice*. London, United Kingdom: Chapman-Hall, 1995.

Gómez-Déniz, E., Hernandez-Bastida, A., and Vazquez-Polo, F.J. "The Esscher Premium Principle in Risk Theory: A Bayesian Sensitivity Study." *Insurance: Mathematics and Economics* 25, no. 3 (1999): 387-395.

Goovaerts, M.J. and Dhaene, J. "The Compound Poisson Approximation for a Portfolio of Dependent Risks." *Insurance: Mathematics and Economics* 18, no. 1 (1996): 81-85.

Haberman S and Renshaw A.E. "Generalized Linear Models and Actuarial Science." *The Statistician* 45, no. 4 (1996): 407-436.

Herzog, T.N. *Introduction to Credibility Theory.* Winsted, CT: Actex Publishers, 1999.

Hogg, R.V. and Klugman, S.A. *Loss Distributions.* New York, NY: Wiley, 1984.

Makov, U.E., Smith, A.F.M., and Liu, Y.H. "Bayesian Methods in Actuarial Science." *The Statistician* 45, no. 4 (1996): 503-515.

Migon, H.S. and Gamerman, D. *Statistical Inference: An Integrated Approach.* London, United Kingdom: Arnold, 1999.

Migon, H.S. and Gamerman, D. "Generalized Exponential Growth Model—A Bayesian Approach." *Journal of Forecasting* 12 (1993): 573-584.

Moura, F.A.S. and Migon, H.S. "Bayesian Hierarchical Models Applied to Credibity Theory." Technical Report No. 150. Rio de Janeiro, Brazil: LES/Universidade Federal do Rio de Janeiro, 2004.

Moura, F.A.S. and Migon, H.S. "Hierarchical Bayesian Collective Risk Model: An Application to Health Insurance." *Insurance: Mathematics and Economics* 36, no. 2 (2005): 119-135.

Ntzoufras, I. and Dellaportas, P. "Bayesian Modelling of Outstanding Liabilities Incorporating Claim Count Uncertainty." *North American Actuarial Journal* 6 (2002): 113-128.

Pai, S. "Bayesian Analysis of Compound Loss Distributions." *Journal of Econometrics* 79, no. 1 (1997): 129-146.

Penna, E.M.O. "Decisão Sob Risco: Um Método de Previsão por Simulação de Monte Carlo Via Cadeias de Markov Aplicado a um Modelo Atuarial de Saúde." (Master's Thesis) COPPE, Universidade Federal do Rio de Janeiro, Brazil, 2000.

Rosenberg, M. A. "A Statistical Method for Monitoring a Change in the Rate of Nonacceptable Inpatient Claims." *North American Actuarial Journal* 5, no. 4 (2001): 74-83.

Scollnik, D.P.M. "Actuarial Modeling with MCMC and BUGS." *North American Actuarial Journal* 5, no. 2 (2001): 96–125.

Scollnik, D.P.M. "Modeling Size-of-Loss Distributions for Exact Data in WinBUGS." *Journal of Actuarial Practice* 10 (2002): 202–227.

Spiegelhalter, D.J., Thomas, A., and Best, N.G. *WinBUGS Version 1.3 User Manual.* Cambridge, United Kingdom: MRC Biostatistics Unit, 2000.

Appendix

Table A1
Service 1 Aggregate Monthly Loss Data for NHC and SHC with Losses Rounded to the Nearest Dollar

	Age Class						
Month	1	2	3	4	5	6	7
Northeast Health Company (NHC)							
Aug/98	637	653	522	851	659	772	245
Sep	1402	504	676	702	715	872	327
Oct	1910	1114	940	1637	992	905	729
Nov	1224	789	784	849	889	1039	442
Dec	1716	733	1100	1625	1906	1176	658
Jan/99	2696	1294	1430	2705	1883	1404	858
Feb	3120	1348	1108	2990	1948	1169	755
Marc	4269	2934	2184	4690	2708	2390	1658
Apr	4628	3388	2940	5488	3073	1979	1315
May	5463	3883	3935	4770	3783	2005	1648
Jun	6345	3605	3708	6080	4048	2623	1333
Jul	6061	3670	3550	6098	4105	2850	1490
Aug	5815	4455	3190	3190	5643	5113	2608
Sep	4540	3398	2995	2995	5430	4495	2845
Oct	4070	3518	3415	3415	5100	4158	2833
Nov	3550	3463	3263	3263	4935	3951	2683
Dec	3581	3055	2623	2623	4987	3173	2108
Jan/00	5108	3680	3130	3130	5343	4045	2490
Feb	6367	5715	4302	4302	6665	4738	3080
Marc	3048	3487	3159	3159	5941	4145	2763
Apr	3841	2971	2100	2100	4062	3449	1737
May	5009	3866	2819	2819	6042	4160	3150
Jun	444	333	529	529	858	558	683
Southeast Health Company (SHC)							
Mar/97	70	89	117	40	40	109	24
Apr	0	55	0	66	20	153	44
May	60	125	60	169	84	106	166
Jun	218	354	403	286	273	346	313
Jul	404	280	911	431	313	534	589
Aug	380	492	1.081	699	443	1122	540
Sep	944	828	674	661	430	1845	537
Nov	649	482	792	649	607	1745	676
Dec	546	1016	938	1274	560	1060	752
Jan/98	619	836	1618	976	542	1.634	821
Feb	932	1.037	1049	927	555	1527	802
Mar	880	885	1011	870	505	952	443
Apr	1013	953	1126	1182	516	861	754
Jun	292	378	319	504	356	661	177
Jul	316	329	287	375	196	485	128

Table A2
Loss Frequency Data for Service 1 for NHC and SHC

Month	Age Class 1	2	3	4	5	6	7
Northeast Health Company (NHC)							
Aug/98	35	36	29	47	36	42	13
Sep	77	28	37	39	38	46	17
Oct	92	53	45	81	49	44	36
Nov	53	37	40	43	43	54	22
Dec	67	29	44	65	74	47	26
Jan/99	106	52	56	108	75	57	36
Feb	125	55	44	119	77	46	29
Marc	169	121	87	191	109	96	67
Apr	186	136	118	224	122	80	53
May	223	153	157	193	153	81	66
Jun	264	145	150	248	170	109	54
Jul	241	144	142	243	163	114	59
Aug	249	194	140	236	218	111	78
Sep	208	146	126	229	182	121	83
Oct	214	155	149	231	188	130	77
Nov	183	188	165	245	200	135	88
Dec	205	148	133	242	175	105	78
Jan/00	238	159	146	247	178	109	88
Feb	280	235	188	287	205	129	113
Marc	213	195	160	273	196	137	131
Apr	244	181	150	292	232	135	94
May	334	216	169	345	226	174	108
Jun	286	214	188	346	256	173	95
Southeast Health Company (SHC)							
Mar/97	3	4	5	2	2	5	1
Apr	0	3	0	3	1	8	2
May	3	4	3	7	4	5	8
Jun	11	17	17	13	12	17	15
Jul	20	14	41	21	15	23	21
Aug	19	25	37	34	20	44	26
Sep	32	28	29	31	18	52	23
Oct	23	20	32	26	23	60	27
Nov	22	43	38	50	18	40	28
Dec	25	35	55	36	22	63	31
Jan/98	37	39	41	37	20	58	31
Feb	36	25	41	36	20	38	18
Marc	41	35	46	45	20	34	28
Apr	12	16	13	21	13	26	7
May	13	13	12	14	8	19	5

Table A3
Population Data for Service 1 for NHC and SHC

Month	Age Class 1	2	3	4	5	6	7
Northeast Health Company (NHC)							
Aug/98	273	160	122	209	177	126	63
Sep	296	168	137	224	198	135	76
Oct	306	176	142	231	208	139	80
Nov	315	182	151	242	215	140	84
Dec	339	187	163	263	224	145	91
Jan/99	356	192	169	269	229	147	92
Feb	526	269	250	427	303	149	103
Marc	551	290	263	458	324	151	107
Apr	589	314	275	493	349	158	111
May	674	354	331	583	421	195	127
Jun	687	356	341	596	432	198	127
Jul	739	399	388	647	461	225	157
Aug	749	410	391	660	466	231	162
Sep	758	419	399	672	470	234	162
Oct	770	430	406	686	479	241	167
Nov	777	436	410	701	487	244	170
Dec	783	441	411	711	489	250	175
Jan/00	788	443	418	721	501	253	180
Feb	796	451	419	725	504	256	183
Marc	815	467	422	741	517	257	186
Apr	833	475	431	757	526	260	192
May	862	491	445	774	541	267	198
Jun	878	508	454	785	550	269	201
Southeast Health Company (SHC)							
Mar/97	89	98	98	107	94	108	68
Apr	91	107	104	108	106	106	70
May	96	123	122	108	116	122	74
Jun	100	136	133	109	118	105	75
Jul	104	143	125	111	119	105	79
Aug	108	190	182	112	121	86	80
Sep	131	209	204	116	125	86	90
Oct	142	233	228	116	128	87	89
Nov	147	238	236	117	127	86	88
Dec	171	261	254	118	127	106	97
Jan	196	282	260	121	133	149	111
Feb	209	293	259	142	138	194	113
Marc	233	329	263	162	152	207	119
Apr	262	347	283	183	166	219	127
May	282	362	299	250	185	232	131