

ALLOWANCE FOR COST OF CLAIMS IN BONUS-MALUS SYSTEMS

JEAN PINQUET¹

THEMA, University Paris X, 92001 Nanterre, France

ABSTRACT

The objective of this paper is to make allowance for cost of claims in experience rating. We design here a bonus-malus system for the pure premium of insurance contracts, from a rating based on their individual characteristics. Empirical results are presented, that are drawn from a French data base of automobile insurance contracts.

KEYWORDS

Bayesian and heterogeneous models. Number and cost residuals. Bonus-malus for frequency of claims, average cost per claim, and pure premium.

INTRODUCTION

Bayesian models lead to a posteriori ratemaking of insurance contracts (Bühlmann (1967)). Suppose that the number of claims follows a Poisson distribution. A bonus-malus system for the frequency of claims is obtained if we consider that the parameter follows a gamma distribution (see Lemaire (1985, 1995)). This model may include a ratemaking of policyholders on an individual basis, the parameter of the Poisson distribution depending then on rating factors (see Dionne et al (1989, 1992)).

The allowance for severity of claims in experience rating can be achieved by considering the dichotomy between claims with material damage only, and claims including bodily injury (see Lemaire (1995)). In this model, the number of claims that caused bodily injury follows a binomial distribution, the parameter of which follows a beta distribution.

In this paper, the severity of claims will be taken into account by using their cost. The analysis of cost of claims makes clearly appear a positive correlation between the average cost per claim and the frequency risk (see Renshaw (1994), Pinquet et al (1992)). An a priori ratemaking will therefore be influenced by the allowance for costs. Concerning the third party liability guaranty, it can be noted that.

- The settlement of claims with material damage is performed partly through fixed amount compensations from an insurance company to the third party

¹Thanks to Georges Dionne for motivating this work, as well as Christian Gouriéroux, Eric Renshaw and two anonymous referees for comments. This research received financial support from the Fédération Française des Sociétés d'Assurance.

- The amount of compensations related to claims including bodily injury depends on the social position of the victim

Hence, it is difficult to explain the cost of these claims by the rating factors, and we shall investigate the damage guaranty in the empirical part of the paper

Allowing for cost of claims in bonus-malus systems can be achieved in the following way. Starting from a rating model based on the analysis of number and cost of claims, two heterogeneity components are added. They represent unobserved factors, that are relevant for the explanation of the severity variables. Later on, we shall refer to any variable explained by a rating model (number, cost of claim, total cost of claims, and so on) as a "severity variable". These unobserved factors are, for instance, annual mileage for number distributions, and speed (and the driver's behaviour in general) for number and cost distributions. A bonus-malus coefficient can be related to the credibility estimation of a heterogeneity component

In this paper, costs of claims are supposed to follow gamma or log-normal distributions. The rating factors, as well as the heterogeneity component, are included in the scale parameter of the distribution. Considering that the heterogeneity component also follows a gamma or log-normal distribution, a credibility expression is obtained, which provides a predictor of the average cost per claim for the following period. For instance, a cost-bonus will appear after the first claim if its cost is inferior to the estimation made by the rating model

Experience rating with a bayesian model is possible only if there is enough heterogeneity in the data. For instance, in the negative binomial model without covariates, the estimated variance of the heterogeneity component is equal to zero if the variance of the number of claims is inferior to their mean (see Pinquet et al (1992)). In that case, a priori and a posteriori tariff structures are the same, and the bayesian model fails.

A sufficient condition for the existence of a bonus-malus system derived from a bayesian model is provided in section 2.3. The existence is equivalent to an overdispersion of residuals related to the severity variable. This approach allows one to test for the presence of a hidden information, that is relevant for the explanation of the severity variables.

The heterogeneity on distributions for severity variables, that is not explained by the rating factors, is revealed through experience on policyholders. The paper investigates the rate of this revelation, which is found to be lower for average cost per claim than for the frequency.

For the sample considered here, the unexplained heterogeneity related to costs is stronger for gamma than for log-normal distributions. Besides, the latter family gives a better fit to the data.

If the heterogeneity components on number and cost distributions are independent, the bonus-malus coefficient for pure premium is the product of the coefficients related to frequency and expected cost per claim. But one may think that the behavior of the policyholder influences the two heterogeneity components in a similar way, and so that they are positively correlated.

Lastly, this paper proposes a bonus-malus system for the pure premium of insurance contracts, that admits a correlation between the two components. Although the

likelihood of a model based on number and costs of claims is not analytically tractable in the presence of such a correlation, consistent estimators for the parameters exist. The correlation between the number and cost heterogeneity components appears to be very low for the sample investigated here

1 A PRIORI RATEMAKING

Let us suppose a sample of policyholders indexed by i , the policyholder i being observed during T_i periods. The analysis of the correlation between the number and cost heterogeneity components shows the necessity of considering a non constant number of periods for each policyholder. The working sample is presented in 1.3

1.1 Frequency of claims

We write

$$N_{it} \sim P(\lambda_{it})_{i=1, \dots, T_i}, \lambda_{it} = \exp(w_{it} \alpha)$$

to represent the Poisson model where n_{it} , the outcome of N_{it} , is the number of claims reported by the policyholder i in period t . The parameter λ_{it} is a multiplicative function of the explanatory variables, the line-vector w_{it} represents their values, and α is the column-vector of the related parameters.

The frequency-premium (estimation of the expectation of N_{it}) is denoted as $\hat{\lambda}_{it} = \exp(w_{it} \hat{\alpha})$, and $nres_{it} = n_{it} - \hat{\lambda}_{it}$ is the number-residual for the policyholder i and period t . The maximum likelihood estimator of α is the solution to the equation:

$$\sum_{i,t} nres_{it} w_{it} = 0,$$

which is an orthogonality relation between the explanatory variables and the residuals. The rating factors have in general a finite number of levels, and the explanatory variables are then indicators of these levels. The preceding equation means that, for every sub-sample associated to a given level, the sum of the frequency premiums is equal to the total number of claims. This property means that the preceding model provides the multiplicative tariff structure that does not mutualize the frequency-risk.

One may think of replacing n_{it} by tc_{it} , the total cost of claims (pure premium rate-making) in the likelihood equation. When applied to the working sample, this non probabilistic model shows that the elasticity of the pure premium risk with respect to the frequency risk is greater than one (see section 1.4.1).

1.2 Models for average cost per claim and pure premium

1.2.1 Gamma distributions

Let c_{itj} be the cost of the j^{th} claim reported by the policyholder i in period t ($1 \leq j \leq n_{it}$, if $n_{it} \geq 1$). We shall suppose in the paper that the costs are strictly positive. This assumption gives another reason to discard the third party liability guaranty: owing to fixed amount compensations, a policyholder involved in a claim caused by the third party can make his insurance company earn money.

Considering gamma distributions, we write

$$C_{it} \sim \gamma(d, b_{it}), b_{it} = \exp(z_{it}\beta),$$

or $b_{it} C_{it} \sim \gamma(d)$. The coefficient b_{it} is a scale parameter, a multiplicative function of the covariates, that are represented by the line-vector z_{it} .

Let $\hat{c}_{it} = \hat{d} / \hat{b}_{it} = \hat{d} / \exp(z_{it}\hat{\beta})$ be the estimation of the average cost for each claim reported by the policyholder i in period t . If we suppose that the costs are independent, the maximum likelihood estimator of β is the solution of the following equation.

$$\sum_{i,t} (n_{it} - (tc_{it} / \hat{c}_{it})) z_{it} = \sum_{i,t} c_{it} \text{res}_{it} z_{it} = 0$$

The term $n_{it} - (tc_{it} / \hat{c}_{it})$ is the sum, for the claims reported by the policyholder i in period t , of their cost residual $1 - (c_{it} / \hat{c}_{it})$. It is written res_{it} . The likelihood equation in β can hence be interpreted as an orthogonality relation between the explanatory variables and cost-residuals.

The average cost per claim increases with the frequency risk (see 1.4.2), which confirms the previous conclusions about the risks related to frequency and pure premium

1.2.2 Log-normal distributions

The other distribution family considered in this paper is the normal distribution family for the logarithms of costs

$$\log C_{it} \sim N(z_{it}\beta, \sigma^2) \Leftrightarrow \log C_{it} = z_{it}\beta + \varepsilon_{it}, \varepsilon_{it} \sim N(0, \sigma^2).$$

The likelihood equation giving $\hat{\beta}$ is

$$\sum_{i,t} \left(\sum_j (\log c_{itj} - z_{it}\hat{\beta}) \right) z_{it} = \sum_{i,t} \text{lcre}_{it} z_{it} = 0.$$

This equation is also an orthogonality relation between explanatory variables and residuals.

1.2.3 Pure premium model

The total cost of claims reported by the policyholder i in period t may be written as:

$$TC_{it} = \sum_{j=1}^{N_{it}} C_{itj}$$

It is a sum of N_{it} i.i.d. outcomes from a variable that we denote as C_{it} . The pure premium is: $E(TC_{it}) = E(N_{it}) E(C_{it})$.

1.3 Presentation of the working sample

The sample investigated in the paper is part of the automobile policyholders portfolio of a French insurance company. It is composed of more than a hundred thousand policyholders. The damage guaranty being considered here, only the contracts with that kind of guaranty were kept. Policyholders can be observed over two years, and each anniversary date, changing of vehicle or coverage level entails a new period. Only claims concerning the damage guaranty and closed at the date of obtention of the data base were kept. Reserved costs were thus avoided. The rating factors retained for the estimation of number and cost distributions are

- The characteristics of the vehicle: group, class, age
- The characteristics of the insurance contract: type of use, level of the deductible, geographic zone

Other rating factors are the policyholder's occupation, as well as the year when the period began (in order to allow for a generation effect). These eight rating factors have a finite number of levels, the total number of which is 44. The explanatory variables are binary, and indicate the levels for the policyholders: in order to avoid collinearity, one level is suppressed for each rating factor, the intercept being kept anyway. Therefore, we shall consider $(44-8)+1=37$ covariates. With the notations of the paper, we obtain: $\alpha, \beta \in \mathbb{R}^{37}; w_{it}, z_{it} \in \{0,1\}^{37}$.

The estimated coefficients derived from the rating model depend on the level suppressed for each rating factor. Results that are independent from the suppressions are obtained by dividing the coefficients by their mean in the multiplicative model. These standardized coefficients can be compared with the relative severity of the levels.

The periods having not the same duration, the parameter of the Poisson distribution must be proportional to the duration. The results given on the frequencies remain unchanged if, d_{it} being the duration of period t for the policyholder i , we write:

$$\lambda_{it} = d_{it} \exp(w_{it} \alpha), \text{ and } \hat{\lambda}_{it} = d_{it} \exp(w_{it} \hat{\alpha})$$

The working sample includes 38772 policyholders and 71126 policyholders-periods. These policyholders reported 3493 claims. The average duration of the periods is nine months, and the annual frequency of the claims is 6.7%.

1.4 Empirical results

1.4.1 A priori rating for frequency and pure premium

When applied to the number of claims or their total cost, the Poisson models provide standardized coefficients, that can be compared with the relative severity of the levels. For almost each rating factor, the variance of the coefficients related to the levels is inferior to the variance of the relative severity. For instance, for the "type of use" rating factor, one gets

frequency	relative severity	standardized coefficient
professional use	1.623	1.278
standard use	0.982	0.992

pure premium	relative severity	standardized coefficient
professional use	1.747	1.177
standard use	0.979	0.995

The distributions of the policyholders among the levels of the different rating factors are not independent from one another. Policyholders with a professional use have, for the other rating factors, more risky levels than the other policyholders. The Poisson model does not mutualize the risk: hence these policyholders have, with respect to other rating factors, a level of relative severity equal to $(1.747/1.177) - 1 = 48.4\%$ more than the average, in term of pure premium.

The elasticity of the pure premium with respect to the frequency risk is equal to 1.52 on the sample, and the difference from 1 is significant (the related Student statistic is equal to 5.93). Hence, if the frequency risk is multiplied by two, the average cost per claim increases by $2^{0.52} - 1 = 43.5\%$, and the pure premium increases by 187%.

This positive correlation between the risks on frequency and average cost per claim is observed on each rating factor, except for the geographical zone.

1.4.2 A priori rating for average cost per claim

On the sample of claims, the gamma model leads to the following results (rating factor: type of use)

average cost	relative severity	standardized coefficient
professional use	1.076	0.933
standard use	0.996	1.003

The estimated elasticity of the average cost per claim with respect to the frequency is equal to 0.51, which confirms the results obtained in the preceding section.

2 EXPERIENCE RATING FOR FREQUENCY AND AVERAGE COST PER CLAIM

2.1 Heterogeneous models

In a bayesian framework, the allowance for a hidden information, relevant for the rating of risks, can be performed in the following way

- the starting point is an a priori rating model. If y represents the severity variable(s), the likelihood of y will be written $f_0(y/\theta_1, x)$, where x is the vector of explanatory variables, and θ_1 the vector of parameters related to them
- A heterogeneity component (scalar, or vector) is added to the model, which measures the influence that unobserved variables have on the severity distribution. If u is this component, a distribution of y conditional on u and the explanatory variables is defined, and we denote its likelihood as $f_\pi(y/\theta_1, x, u)$. In practice, the a priori distribution is equal to the distribution defined conditionally on u , for some value u^0 of u : $f_\pi(y/\theta_1, x, u^0) = f_0(y/\theta_1, x) \forall \theta_1, x, y$. If u is a scalar, $u^0 = 0$ or 1, according to the fact that u is included additively or multiplicatively in the conditional distribution

- The credibility estimation of u_i , the heterogeneity component for the policyholder i , leads to a bonus-malus system. It rests on a heterogeneous model, in which u_i is the outcome of a random variable U_i , the $(U_i)_{i=1, \dots, p}$ being i.i.d. and their distribution being parameterized by θ_2 . The likelihood of y_i in the model with heterogeneity is obtained by integrating the conditional likelihood over U_i , that is to say

$$f(y_i / \theta, x_i) = E_{\theta_2} [f_*(y_i / \theta_1, x_i, U_i)],$$

with $\theta = (\theta_1, \theta_2)$. The heterogeneity component vector on number and cost distributions will be denoted, for the policyholder i

$$U_i = \begin{pmatrix} U_{ni} \\ U_{ci} \end{pmatrix},$$

where n stands for the numbers and c for the costs. The link between heterogeneous and bayesian models is made clear in the example that follows

2.2 Examples of heterogeneous models

2.2.1 Number of claims

With the notations of 1.1, the distributions defined conditionally on u_{ni} are

$$N_{ni} \sim P(\lambda_{ni} u_{ni}), \text{ with } U_{ni} \sim \gamma(a, a)$$

in the heterogeneous model. The expectation of U_{ni} is equal to one, and its variance is $1/a$. On a period, the number of claims distribution is negative binomial in the heterogeneous model.

The negative binomial model can be considered as a Poisson model with a random component, if we write $\lambda_{ni} U_{ni} = \tilde{\lambda}_{ni}$. If the intercept is the first of k explanatory variables, and if e_1 is the first vector of the canonical base of \mathbb{R}^k , we have

$$\tilde{\lambda}_{ni} = \exp(w_{ni} \alpha + \log(U_{ni})) = \exp(w_{ni} (\alpha + \log(U_{ni})e_1)) = \exp(w_{ni} \tilde{\alpha}_i)$$

In the last expression of λ_{ni} , the parameter $\tilde{\alpha}_i = \alpha + \log(U_{ni})e_1$ is random, and the formulation is bayesian. But it is less tractable than that of the heterogeneous model, as well for bonus-malus computations as for statistical inference.

2.2.2 Gamma distributions for costs of claims

The heterogeneous models that follow, which allow us to design bonus-malus systems for average cost per claim, suppose the independence of heterogeneity components on the number and costs distributions. The empirical results presented later will make this assumption plausible.

For the gamma model and with the notations of 1.2.1, the distributions conditional on u_{ci} are

$$C_{ij} \sim \gamma(d, b_{ij} u_{ci}), \text{ with } U_{ci} \sim \gamma(\delta, \delta)$$

in the heterogeneous model. The heterogeneity component is included, as the rating factors, in the scale parameter of the distribution.

In the heterogeneous model, one can write $C_{ij} = D_{ij} / (b_u U_{ci})$, with $D_{ij} \sim \gamma(d)$, $U_{ci} \sim \gamma(\delta, \delta)$, D_{ij} and U_{ci} being independent. The variable C_{ij} follows a GB2 distribution (see Cummins et al (1990)), and D_{ij} represents the relative severity of the claim.

2.2.3 Log-normal distributions for costs of claims

With the notations of 1.2.2, the heterogeneous model is

$$\log C_{ij} = z_{it}\beta + \varepsilon_{ij} + U_{ci}, \quad U_{ci} \sim N(0, \sigma_U^2),$$

where the ε_{ij} and U_{ci} are independent. The variable ε_{ij} represents the relative severity of the claim

The heterogeneous model used to design a bonus-malus system for pure premium will be presented after the empirical results related to the preceding models.

2.3 A sufficient condition for the existence of a bonus-malus system derived from a bayesian model

Experience rating with a bayesian model is possible only if there exists enough heterogeneity on the data. Considering for instance the negative binomial model without covariates, the estimated variance of the heterogeneity component is equal to zero if the variance of the number of claims is lower than their mean (see Pinquet et al. (1992)). In that case, a priori and a posteriori tariff structures do not differ, and the bayesian model fails.

A sufficient condition for the existence of a bonus-malus system derived from a bayesian model is provided here: it will be applied later on to the models for number and cost of claims

Let us start from a heterogeneous model, as defined in 2.1. The heterogeneity component is supposed to be scalar, and its distribution is parameterized by the variance σ^2 . The parameters of the model are $\theta = (\theta_1, \sigma^2)$ and we shall write $\hat{\theta}^0 = (\hat{\theta}_1^0, 0)$, $\hat{\theta}_1^0$ being the maximum likelihood estimator of θ_1 in the a priori rating model.

If the right-derivative, with respect to σ^2 , of the log-likelihood is positive in $\hat{\theta}^0$, $\hat{\sigma}^2$ will be positive in the heterogeneous model. The existence of a bonus-malus system is hence related to the sign of a lagrangian, which is part of the score test for nullity of σ^2 (see Rao (1948), Silvey (1959)). With the notations of 2.1, and denoting the lagrangian as \mathcal{L} , one can prove:

$$\begin{aligned} \sum_i \log f(y_i / \hat{\theta}_1^0, \sigma^2, x_i) - \sum_i \log f_0(y_i / \hat{\theta}_1^0, x_i) &= \mathcal{L}\sigma^2 + o(\sigma^2), \text{ with} \\ \mathcal{L} &= \frac{1}{2} \sum_i (res_i^2 - s_i); \\ res_i &= \left(\frac{\partial}{\partial u} \log f_*(y_i / \hat{\theta}_1^0, x_i, u) \right)_{u=u^0}; \quad s_i = - \left(\frac{\partial^2}{\partial u^2} \log f_*(y_i / \hat{\theta}_1^0, x_i, u) \right)_{u=u^0} \end{aligned}$$

See Pinquet (1996b) for a proof, and references to a recent literature. The term res_i is a residual, which is related to those encountered in the likelihood equations for numbers and costs. The condition for existence of a bonus-malus system is

$$\mathcal{L} > 0 \Leftrightarrow \sum_i res_i^2 > \sum_i s_i$$

It can be interpreted as an overdispersion condition on residuals.

2.4 Prediction with heterogeneous models and bonus-malus systems

Let us suppose a policyholder observed on T periods: $Y_T = (y_1, \dots, y_T)$ is the sequence of severity variables, and $X_T = (x_1, \dots, x_T)$ that of the covariates. The sequences X_T and Y_T take the place of x_i and y_i in the preceding sections. The date of forecast T must be explicitated here, and the individual index can be suppressed, since the policyholder can be considered separately. Besides, belonging to the working sample is not mandatory for this policyholder.

We want to predict a risk for the period $T+1$, by means of a heterogeneous model. For the period t , this risk R_t is the expectation of a function of Y_t (y_t is the outcome of Y_t). For instance, Y_t is the sequence of both number and costs of claims in period t , and R_t , the pure premium, is the expectation of the total cost.

We now include a heterogeneity component u , as defined in 2.1. The distribution of Y_t conditional on u depends on θ_1, x_t and u . This applies to R_t , and we can write $R_t = h_{\theta_1}(x_t) g(u)$, for the three types of risk dealt with later (frequency of claims, average cost per claim, pure premium), g being a real-valued function.

A predictor for the risk in period $T+1$ can be written as $h_{\hat{\theta}_1}(x_{T+1}) \hat{g}^{T+1}(u)$, with $\hat{g}^{T+1}(u)$ a credibility estimator of $g(u)$, defined from:

$$\hat{g}^{T+1}(u) = \arg \min_a E_{\theta_2} [(g(U) - a)^2 f_*(Y_T / \theta_1, X_T, U)],$$

$$f_*(Y_T / \theta_1, X_T, U) = \prod_{t=1}^T f_*(y_t / \theta_1, x_t, U).$$

The expectation is taken with respect to U , and one obtains

$$\hat{g}^{T+1}(u) = E_{\theta} [g(U) / X_T, Y_T] = \frac{E_{\theta_2} [g(U) f_*(Y_T / \theta_1, X_T, U)]}{E_{\theta_2} [f_*(Y_T / \theta_1, X_T, U)]},$$

the expectation of $g(U)$ for the posterior distribution of U . Replacing θ_1 and θ_2 by their estimations in the heterogeneous model, we obtain the a posteriori premium

$$\hat{R}_{T+1}^{T+1} = h_{\hat{\theta}_1}(x_{T+1}) E_{\hat{\theta}} [g(U) / X_T, Y_T],$$

computed for period $T+1$. It can be written as

$$\left(h_{\hat{\theta}_1}(x_{T+1}) E_{\hat{\theta}_2} [g(U)] \right) \times \frac{E_{\hat{\theta}} [g(U) / x_1, \dots, x_T; y_1, \dots, y_T]}{E_{\hat{\theta}_2} [g(U)]}$$

The first term is an a priori premium, based on the rating factors of the current period. The second one is a bonus-malus coefficient it appears as the ratio of two expectations of the same variable, computed for prior and posterior distributions. Owing to the equality $E_{\theta}[E_{\theta}(g(U)/X_T, Y_T)] = E_{\theta}[g(U)] = E_{\theta_i}[g(U)]$, the rating is balanced.

2.5 Bonus-malus for frequency of claims

2.5.1 Theoretical results

With the notations of 2.2.1 and 2.4, we write: $y_t = n_t, x_t = w_t, \theta_1 = \alpha; R_t = E(N_t) = \lambda_t u, h_{\theta_1}(x_t) = \lambda_t, g(u) = u; X_T = (w_1, \dots, w_T), Y_T = (n_1, \dots, n_T)$. The posterior distribution of U is a $\gamma(a + \sum_t n_t, a + \sum_t \lambda_t)$ (see Dionne et al (1989, 1992))

Hence:

$$E_{\theta}[U/w_1, \dots, w_T, n_1, \dots, n_T] = \hat{u}^{T+1} = \frac{a + \sum_{t=1}^T n_t}{a + \sum_{t=1}^T \lambda_t} \tag{1}$$

Replacing λ_t by $\hat{\lambda}_t = \exp(w_t \hat{\alpha})$ and a by \hat{a} in equation (1) leads to the bonus-malus coefficient. There will be a frequency-bonus if the estimator of $\hat{u}^{T+1} - 1$ is negative, or if the number-residual $\sum_t (n_t - \hat{\lambda}_t)$ is negative

Considering in equation (1) that N_t follows a Poisson distribution, with a parameter $\lambda_t u, \hat{u}^{T+1}$ converges towards u when T goes to $+\infty$. The heterogeneity on number distributions, which is not explained by the rating factors, is hence revealed completely with time. It may be interesting to investigate the distribution of bonus-malus coefficients on a portfolio of policyholders, as well as its time evolution (see section 2.5.2 for empirical results)

We explicit now the condition for existence of a bonus-malus system for frequencies. On the working sample, and with the notations in 2.2.1, one can write

$$\log f_{\theta}(y_t / \hat{\theta}_1^0, x_t, u) = \sum_t \left[n_{it} (\log \hat{\lambda}_{it} + \log u) - \hat{\lambda}_{it} u - \log(n_{it}!) \right],$$

with $\hat{\lambda}_{it} = \exp(w_{it} \hat{\alpha}^0), \hat{\alpha}^0$ being the estimator of α in the a priori rating model. With the notations of 2.3, and with $u^0 = 1$, we obtain

$$res_{it} = \sum_t (n_{it} - \hat{\lambda}_{it}), s_{it} = \sum_t n_{it}, L > 0 \Leftrightarrow \sum_t nres_{it}^2 > \sum_t n_{it},$$

where $nres_{it} = \sum_t (n_{it} - \hat{\lambda}_{it})$ is the number-residual for policyholder i , and $n_{it} = \sum_t n_{it}$ is the number of claims reported by this policyholder on all periods. This condition means that, considering the total number of claims, its variance is superior to its mean, the variance being calculated conditionally on the explanatory variables. This empirical overdispersion condition can be related to the theoretical overdispersion of the

negative binomial model: if $N_i \sim P(\lambda_i, U_i)$, $U_i \sim \gamma(a, a)$ (with $a = 1/\sigma^2$), one gets: $V(N_i) = \lambda_i + \lambda_i^2 \sigma^2 > \lambda_i = E(N_i)$

A score test for nullity of σ^2 can be performed from the Lagrange multiplier $\mathcal{L} = (1/2) \sum_i (nres_i^2 - n_i)$. The previous remarks allow us to reject the nullity of σ^2 if \mathcal{L} is large enough. If the number of policyholders goes to infinity, $\xi^{\mathcal{L}} = \mathcal{L} / \sqrt{\hat{V}(\mathcal{L})}$ converges towards a $N(0, 1)$ distribution. One can prove that $\hat{V}(\mathcal{L}) = 1/2 \sum_i \hat{\lambda}_i^2$, with $\hat{\lambda}_i = \sum_{ii} \hat{\lambda}_{ii}$. If $u_{1-\varepsilon}$ is the quantile at the level $1 - \varepsilon$ of a $N(0, 1)$ distribution, the null hypothesis $\sigma^2 = 0$ will be rejected at the level ε if $\xi^{\mathcal{L}} \geq u_{1-\varepsilon}$.

Besides, the lagrangian provides an estimator of the parameters. Starting from $\hat{\alpha}^0$ and $\hat{\sigma}^2 = 0$ in the algorithm of the likelihood maximisation, one gets at the following step

$$\hat{\alpha}^1 = \hat{\alpha}^0; \hat{\sigma}^{2^1} = \frac{\mathcal{L}}{\hat{V}(\mathcal{L})} = \frac{\sum_i nres_i^2 - \sum_i n_i}{\sum_i \hat{\lambda}_i^2} = \frac{\sum_i [(n_i - \hat{\lambda}_i)^2 - n_i]}{\sum_i \hat{\lambda}_i^2} \quad (2)$$

The estimators $\hat{\alpha}^1$ and $\hat{\sigma}^{2^1}$ can be shown to be consistent for the negative binomial model (see Pinquet (1996b) for demonstrations)

2.5.2 Empirical results

From the sample described in 1.3, we obtain

$$\sum_i nres_i^2 = \sum_i (n_i - \hat{\lambda}_i)^2 = 3709.24; \sum_i n_i = n = 3493,$$

and experience rating is possible for frequencies. Without explanatory variables (apart from total duration of observation for each policyholder), one obtains: $\sum_i nres_i^2 = 3746.25$. The sum of square of residuals decreases when explanatory variables are added, and the condition for existence of a bonus-malus system is more restrictive when they are present. This is logical because they are a cause of heterogeneity on a priori distributions.

Besides, $\sum_i \hat{\lambda}_i^2 = 389.48$, and the estimator of σ^2 given in (2) is

$$\hat{\sigma}^2 = \frac{\mathcal{L}}{\hat{V}(\mathcal{L})} = \frac{\sum_i nres_i^2 - \sum_i n_i}{\sum_i \hat{\lambda}_i^2} = \frac{216.24}{389.48} = 0.555.$$

As a comparison, the maximum likelihood estimation for the negative binomial model is $\hat{\sigma}^2 = 0.576$. The score test for nullity of σ^2 is based on the statistic

$$\xi^L = \frac{L}{\sqrt{\hat{V}(L)}} = \frac{\sum_i nres_i^2 - \sum_i n_i}{\sqrt{2 \sum_i \hat{\lambda}_i^2}} = \frac{216.24}{\sqrt{778.96}} = 7.75,$$

and the null hypothesis is rejected. Examples of bonus-malus coefficients derived from the credibility formula are developed in actuarial and econometric literature (see Lemaire (1985), Dionne et al (1989,1992)).

Evolution throughout time of bonus-malus coefficients, as well as a posteriori premiums related to them, will be investigated for the risks related to frequency and average cost per claim. We consider here a simulated portfolio, derived from the working sample. In this portfolio, the characteristics of each policyholder in the sample are those of the first period, and we suppose that they remain unchanged. If this assumption does not hold individually, it is however plausible on the whole population. Investigating the distribution of bonus-malus coefficients in the heterogeneous model, one can measure their dispersion on the portfolio by estimating their coefficient of variation after T years (see Pinquet (1996a)). Considering the frequencies, with the tariff structure obtained in 1.4.1 and $\hat{\sigma}^2 = 0.576$, we obtain:

TABLE I
REVEALATION THROUGHOUT TIME OF HETEROGENEITY RELATED TO NUMBER DISTRIBUTIONS

	Coefficients of variation (frequency of claims) a priori premium 0.372				
	T=1	T=5	T=10	T=20	T=+∞
bonus-malus coefficient	0.144	0.300	0.392	0.494	0.759
a posteriori premium	0.411	0.515	0.590	0.673	0.891

The coefficient of variation is a measure of the relative dispersion of bonus-malus coefficients and premiums. Apart from the a priori premium, the elements of the preceding table are an estimation of the expectation in the heterogeneous model. After nine years, the relative dispersion of the bonus-malus coefficients exceeds that of the a priori premium. This means that, after nine years, the heterogeneity revealed by the observation of policyholders becomes more important than that explained by the rating factors.

2.6 Bonus-malus for average cost per claim (gamma distributions)

2.6.1 Theoretical results

With the notations in 2.2.2 and 2.4, we can write: $y_i = (c_{ij})_{j=1, \dots, n_i}$, $x_i = z_i$; $R_i = E(C_{ij}) = d/(b_i u)$; $\theta_i = (\beta, d)$; $h_{\theta_i}(x_i) = d/b_i$; $g(u) = 1/u$. The bonus-malus coefficient on average cost per claim for period $T+1$ is derived from the credibility estimator

of $1/u$. Since the a priori distribution of U is a $\gamma(\delta, \delta)$, with a density proportional to $f_\delta(u) = \exp(-\delta u)u^{\delta-1}$, one gets:

$$f_\delta(u) \times f_*(Y_T/\theta_1, X_T, u) = \exp((\delta + \sum_{i,j} b_i c_{ij})u)u^{d(\sum n_i) + \delta - 1},$$

times a coefficient independent of u . The posterior distribution of U is therefore a $\gamma(\delta + d(\sum_i n_i), \delta + \sum_{i,j} b_i c_{ij})$, and:

$$\widehat{1/u}^{T+1} = E_\theta \left[\frac{1}{U} / X_T, Y_T \right] = \frac{\delta + \sum_{i,j} b_i c_{ij}}{\delta - 1 + d(\sum_i n_i)}$$

We have $E_{\theta_2}(1/U) = \delta/(\delta - 1)$ (we suppose $\delta > 1$, a necessary condition for $1/U$ to have a finite expectation). Omitting the period index, and writing S_T for the set of claims reported by the policyholder during the first T periods, the bonus-malus coefficient is

$$\frac{E_{\hat{\theta}} \left[\frac{1}{U} / X_1, Y_1 \right]}{E_{\hat{\theta}_2} \left[\frac{1}{U} \right]} = \frac{\hat{\eta} + \sum_{j \in S_T} (c_j / E_{\hat{\theta}}(C_j))}{\hat{\eta} + |S_T|}, \tag{3}$$

where we wrote: $\eta = (\delta - 1)/d$, $E_{\theta_2}(C_j) = E_{\theta_2}(d/(b_j U)) = (d/b_j)(\delta/(\delta - 1))$. The rating structure derived from (3) is obviously balanced. Writing $E_{\hat{\theta}}(C_j) = \hat{c}_j$, and $res_T = \sum_{j \in S_T} (1 - (c_j / \hat{c}_j))$ the cost-residual for the policyholder, there will be a cost-bonus if the cost-residual is positive. The bonus is then equal to

$$1 - \frac{\hat{\eta} + \sum_{j \in S_T} c_j / \hat{c}_j}{\hat{\eta} + |S_T|} = \frac{res_T}{\hat{\eta} + |S_T|}$$

The time evolution of the distribution of bonus-malus coefficients is investigated in 2.6.2. Considering the simulated portfolio defined in 2.5.2, the heterogeneity unexplained by the rating factors is revealed more slowly for cost than for number distributions. This is not surprising, as far as no claim means no information on the cost distribution — if there is no correlation between the two heterogeneity components — whereas no claim generates frequency-bonus.

Let us apply to this model the condition allowing experience rating. For the working sample, we denote S_i as the set of claims reported by the policyholder over the T_i periods. One can write

$$\log f_*(y_i / \hat{\theta}_1^0, x_i, u) = \sum_{j \in S_i} (\hat{d}^0 \log u - \hat{b}_{ij}^0 c_{ij} u) + z_i,$$

where z_i does not depend on u . With the notations of 2.3 and with $u^0 = 1$, we obtain:

$$res_i = \sum_{j \in S_i} (\hat{d}^0 - \hat{b}_{ij}^0 c_{ij}); s_i = n_i \hat{d}^0; L > 0 \Leftrightarrow \frac{1}{n} \sum_i cres_i^2 > \frac{1}{\hat{d}^0}$$

The total number of claims over the sample is n , and $cres_i$ is the cost-residual for the policyholder i . This residual is equal to 0 without claims, and otherwise, $cres_i = \sum_{j \in S_i} (1 - (c_{ij} / \hat{c}_{ij}^0)) = \sum_{j \in S_i} cres_{ij}$, where $\hat{c}_{ij}^0 = \hat{d}^0 / \hat{b}_{ij}^0$ is the estimator for the expectation of C_{ij} . Now, we have: $E(1 - (C_{ij} / E(C_{ij})))^2 = V(C_{ij}) / E^2(C_{ij}) = CV^2(C_{ij}) = 1/d$, if $C_{ij} \sim \gamma(d, b_{ij})$. The condition for existence of a bonus-malus system is hence related to the square of coefficients of variation

2.6.2 Empirical results

Considering the working sample, one obtains:

$$\frac{1}{n} \sum_i cres_i^2 = 1.092; \frac{1}{\hat{d}^0} = 0.821,$$

and experience rating for average cost of claims is possible. For the sample of policyholders that reported claims, the maximum likelihood estimators for the GB2 model are.

$$\hat{\delta} = 3.620, \hat{d} = 1.807, \hat{\eta} = (\hat{\delta} - 1) / \hat{d} = 1.45.$$

The bonus (negative in case of malus) related to average cost per claim is equal to $cres_i / (\hat{\eta} + |S_i|)$. It remains equal to zero as long as there are no claims. After the first claim, if we consider the cases where the ratio actual cost-predicted cost is equal, either to 0.5 or to 2, the related cost-residuals are equal to 0.5 and -1 respectively. The multiplicative coefficient $1/(1 + \hat{\eta})$ being equal to 0.408, we obtain a cost-bonus of 20.4% in the first case, and a cost-malus of 40.8% in the second case. This coefficient is independent of the period during which the claim occurs.

The distributions of bonus-malus coefficients and a posteriori premiums can be investigated on the simulated portfolio defined in 2.5.2. With the tariff structures obtained in 1.4.1 and 1.4.2 and $\hat{\delta} = 3.62$, we obtain (see Pinquet (1996a))

TABLE 2
RELATION THROUGHOUT TIME OF HETEROGENEITY RELATED TO COST DISTRIBUTIONS

	Coefficients of variation (expected cost per claim) a priori premium 0.401				
	T=1	T=5	T=10	T=20	T=+∞
bonus-malus coefficient	0.128	0.268	0.356	0.453	0.786
a posteriori premium	0.427	0.504	0.568	0.648	0.937

The relative dispersion of the bonus-malus coefficients exceeds the dispersion of the a priori premium after fourteen years. Unexplained heterogeneity on cost distributions is revealed more slowly than it was for numbers.

2.7 Bonus-malus for average cost per claim (log-normal distributions)

2.7.1 Theoretical results

With the notations in 2.2.2 and 2.4, we write $y_i = (\log c_{ij})_{j=1, \dots, n_i}$; $x_i = z_i$, $\log C_{ij} \sim N(z_i \beta + u, \sigma^2) \Rightarrow R_i = E(C_{ij}) = \exp(z_i \beta + u + (\sigma^2 / 2))$, $\theta_1 = (\beta, \sigma^2)$, $h_{\theta_1}(x_i) = \exp(z_i \beta + (\sigma^2 / 2))$; $g(u) = \exp(u)$. The bonus-malus coefficient is derived from the credibility estimator of $\exp(u)$. Now

$$f_{\sigma_U^2}(u) \times f_*(Y_T / \theta_1, X_T, u) = \exp \left[-\frac{1}{2} \left(\frac{1}{\sigma_U^2} + \frac{m_T}{\sigma^2} \right) \left(u - \frac{tlc_T - E_{\theta_1}(TLC_T)}{m_T + (\sigma^2 / \sigma_U^2)} \right)^2 \right]$$

times a coefficient independent from u . We wrote $m_T = \sum_{i=1}^I n_i$, $tlc_T = \sum_{j \in S_i} \log c_j$, $E_{\theta_1}(TLC_T) = \sum_{j \in S_i} E_{\theta_1}(\log C_j)$; S_T is the set of claims reported by the policyholder during the T periods ($|S_T| = m_T$), and the period index is omitted. Hence, the posterior distribution of U is

$$U / (X_T, Y_T) \sim N \left(\frac{tlc_T - E_{\theta_1}(TLC_T)}{m_T + (\sigma^2 / \sigma_U^2)}, \frac{1}{(1 / \sigma_U^2) + (m_T / \sigma^2)} \right)$$

The bonus-malus coefficient for period $T+1$ is equal to

$$\frac{E_{\hat{\theta}_2}[\exp(U) / X_T, Y_T]}{E_{\hat{\theta}_2}[\exp(U)]} = \exp \left[\frac{lcr_{es_T} - (m_T \hat{\sigma}_U^2 / 2)}{(\hat{\sigma}^2 / \hat{\sigma}_U^2) + m_T} \right],$$

writing $lcr_{es_T} = \sum_{j \in S_i} lcr_{es_j}$, $lcr_{es_j} = \log c_j - E_{\hat{\theta}_1}(\log C_j)$.

The condition for existence of a bonus-malus system is easily interpretable with the log-normal model. We have

$$\log f_*(y_i / \hat{\theta}_1^0, x_i, u) = - \sum_{j \in S_i} \frac{(lcr_{es_j} - u)^2}{2 \hat{\sigma}^2{}^0}$$

plus terms that do not depend on u , with $lcr_{es_j} = \log(c_{ij}) - z_{ij} \hat{\beta}^0$. With $u^0 = 0$ (see 2.3), the existence condition is:

$$\sum_i \frac{(\sum_{j \in S_i} lcr_{es_j})^2}{(\hat{\sigma}^2{}^0)^2} - \frac{n}{\hat{\sigma}^2{}^0} = \frac{1}{(\hat{\sigma}^2{}^0)^2} \left[\sum_i \left(\sum_{j \in S_i} lcr_{es_j} \right)^2 - n \hat{\sigma}^2{}^0 \right] > 0$$

Now, in the a priori rating model, $n \hat{\sigma}^2{}^0 = \sum_{i,j} lcr_{es_j}^2$, with $\hat{\sigma}^2{}^0$ the maximum likelihood estimator of σ^2 . Experience rating is possible if

$\sum_i \left(\sum_{j \in S_i} lres_{ij} \right)^2 - \sum_{i,j} lres_{ij}^2$ is positive, that is to say if

$$\sum_{i/n_i \geq 2} \sum_{j,k \in S_i, j \neq k} lres_{ij} lres_{ik} > 0$$

This condition means that, for claims related to policyholders having reported several of them, cost-residuals have rather the same sign. If the first claim has a cost greater than its prediction, it will be the same on average for the following ones.

One can prove that, if \mathcal{L} is the lagrangian with respect to σ_U^2 , we have

$$\hat{V}(\mathcal{L}) = \frac{\sum_i n_i(n_i - 1)}{2(\hat{\sigma}^2)^2} \Rightarrow \hat{\sigma}_U^2 = \frac{\mathcal{L}}{\hat{V}(\mathcal{L})} = \frac{\sum_{i/n_i \geq 2} \sum_{j,k \in S_i, j \neq k} lres_{ij} lres_{ik}}{\sum_i n_i(n_i - 1)},$$

and that $\hat{\sigma}_U^2$ is an consistent estimator of σ_U^2 (see Pinquet (1996a)). It appears to be the average, for the policyholders having reported several claims, of the product of residuals associated to couples of different claims

2.7.2 Empirical results

From the working sample, we obtain $\sum_{i/n_i \geq 2} \sum_{j,k \in S_i, j \neq k} lres_{ij} lres_{ik} = 100.80$, and experience rating is possible Hence

$$\hat{\sigma}_U^2 = \frac{\sum_{i/n_i \geq 2} \sum_{j,k \in S_i, j \neq k} lres_{ij} lres_{ik}}{\sum_i n_i(n_i - 1)} = \frac{100.80}{590} = 0.171.$$

The nullity of σ_U^2 is tested for with $\xi^L = \mathcal{L} / \sqrt{\hat{V}(\mathcal{L})} = 2.86$ The critical value for a one-sided test at a level of 5% is 1.645, and the null hypothesis is rejected The maximum likelihood estimators of σ_U^2 and σ^2 in the heterogeneous model are: $\hat{\sigma}_U^2 = 0.172$, $\hat{\sigma}^2 = 0.855$.

Bonus-malus coefficients can be computed from the examples considered with the gamma distributions (one claim, and a ratio actual cost-expected cost equal to 0.5 or 2) The residual associated to a claim is the logarithm of the latter ratio In the first case, the bonus-malus coefficient is equal to

$$\exp \left[\frac{lres_T - (ln_T \hat{\sigma}_U^2 / 2)}{(\hat{\sigma}^2 / \hat{\sigma}_U^2) + ln_T} \right] = \exp \left[\frac{-\log 2 - 0.086}{(0.855 / 0.172) + 1} \right] = 0.878,$$

and is associated to a cost-bonus of 12.2% In the second case, the bonus-malus coefficient is equal to 1.107, and implies a cost-malus of 10.7% These results can be compared with 20.4% and 40.8%, the boni and mali derived from the gamma distributions, although the ratios actual cost-expected cost are different in the two models. They

must be different, since the cost-residuals in the gamma and log-normal models are equal to $1 - (c_{ij} / \hat{c}_{ij}^{\text{gamma}})$ and $\log(c_{ij} / \hat{c}_{ij}^{\text{log-normal}})$ respectively, whereas they fulfill the same orthogonality relations with respect to the covariates.

Considering the simulated portfolio defined in 2.5.2, the heterogeneity on cost distributions that is unexplained by the a priori rating model is more important for gamma than for log-normal distributions. This can be seen by comparing the limits of the coefficients of variation for the bonus-malus coefficients, as we did in sections 2.5.2 and 2.6.2. For the GB2 model, this limit is the coefficient of variation of $1/U, U \sim \gamma(\hat{\delta}, \hat{\delta})$ (see Pinquet (1996a)). With $\hat{\delta} = 3.62$, it is equal to $1/\sqrt{\hat{\delta} - 2} = 0.786$. Considering the log-normal model, the limit is the coefficient of variation of $\exp(U), U \sim N(0, \hat{\sigma}_U^2)$.

With $\hat{\sigma}_U^2 = 0.172$, it is equal to $\sqrt{\exp(\hat{\sigma}_U^2) - 1} = 0.433$.

This result can be related to a comparison between the two a priori rating models. If F_{θ_j, x_j} is the continuous distribution function of Y_j (here equal to the cost of the claim j , or its logarithm) $e_j = F_{\theta_j, x_j}(Y_j)$ is uniformly distributed on $[0, 1]$. Computing the residuals $e_j, e_j = F_{\theta_j, x_j}(Y_j)$, and rearranging e_j in the increasing order, by $e_{(1)} \leq \dots \leq e_{(n)}$, we derive the Komolgorov-Smirnov statistic $KS = \sqrt{n} \max_{1 \leq j \leq n} |(j/n) - e_{(j)}|$. We obtain $KS = 2.83$ (resp. $KS = 1.04$) for the gamma (resp. log-normal) distribution family. The latter family seems to fit the data better than the gamma family, and will be retained for the bonus-malus system on pure premium.

The two last results can be related to each other. There is more unexplained heterogeneity for gamma than for log-normal distributions, and the latter provide a better fit to the data. This fact raises a question: is apparent heterogeneity only explained by hidden information, or can it be also explained by the fact that the model does not make the best use of observable information?

3 BONUS-MALUS FOR PURE PREMIUM

3.1 The heterogeneous model

From the preceding results, we shall retain log-normal rather than gamma distributions for costs. Besides, they are better integrated in a heterogeneous model with a joint distribution for the two heterogeneity components related to the number and cost distributions. We retain here a bivariate normal distribution. The parameters of the related heterogeneous model can be estimated consistently, although the likelihood is not analytically tractable.

A way to derive consistent estimators for heterogeneous models is proposed in Pinquet (1996b). It is based on the properties of extremal estimators, the maximum likelihood estimator being of this type. The estimators of the parameters of the a priori

rating model have a limit if the actual distributions include heterogeneity, and this limit is tractable in the model investigated here. Consistent estimators are then obtained from a method of moments using the scores with respect to the variances and the covariances of the heterogeneity components.

The heterogeneous model is hence composed of Poisson distributions on numbers, log-normal distributions on costs, and of bivariate normal distributions for the two heterogeneity components. The notations are the following.

- The distributions conditional on u_{ni} and u_{ci} , the heterogeneity components for number and cost distributions of the policyholder i , are

$$N_{it} \sim P(\lambda_{it} \exp(u_{ni})), \log C_{ijt} = z_{it}\beta + \varepsilon_{ijt} + u_{ci}, \text{ with}$$

$$\lambda_{it} = \exp(w_{it}\alpha), \varepsilon_{ijt} \sim N(0, \sigma^2), t = 1, \dots, T_i; j = 1, \dots, n_{it}$$

- In the heterogeneous model, U_{ni} and U_{ci} follow a bivariate normal distribution with a null expectation and a variance equal to

$$V = \begin{pmatrix} V_{nn} & V_{nc} \\ V_{cn} & V_{cc} \end{pmatrix}.$$

The parameters of the model are

$$\theta_1 = \begin{pmatrix} \alpha \\ \beta \\ \sigma^2 \end{pmatrix}, \theta_2 = \begin{pmatrix} V_{nn} \\ V_{cn} \\ V_{cc} \end{pmatrix}$$

Bonus-malus coefficients are computed in the heterogeneous model from the expression given in section 2.4

$$\frac{E_{\hat{\theta}}[g(U) | \mathcal{X}_T, \mathcal{Y}_T]}{E_{\hat{\theta}_2}[g(U)]} = \frac{E_{\hat{\theta}_2}[g(U) f(\mathcal{Y}_T / \hat{\theta}_1, \mathcal{X}_T, U)]}{E_{\hat{\theta}_2}[g(U)] E_{\hat{\theta}_2}[g(U) f(\mathcal{Y}_T / \hat{\theta}_1, \mathcal{X}_T, U)]} \quad (4)$$

We can write.

- $g(u_n, u_c) = \exp(u_n)$ for frequency
- $g(u_n, u_c) = \exp(u_c)$ for average cost per claim
- $g(u_n, u_c) = \exp(u_n + u_c)$ for pure premium.

because the expectations of N_i , C_{ij} and TC_i are respectively proportional to $\exp(u_n)$, $\exp(u_c)$ and $\exp(u_n + u_c)$, if computed conditionally on u_n and u_c . The mathematical expectations that lead to the bonus-malus coefficients (see equation (4)) can be estimated if we can write $U = f_{\hat{\theta}_2}(S)$, where the distribution of S is independent from θ_2 it is enough to simulate outcomes of S . Such an expression can be obtained by writing the Choleski decomposition of the variances-covariances matrix, i.e.

$$V = \begin{pmatrix} V_{nn} & V_{nc} \\ V_{cn} & V_{cc} \end{pmatrix} = T_\varphi T_\varphi'; T_\varphi = \begin{pmatrix} \varphi_{nn} & 0 \\ \varphi_{cn} & \varphi_{cc} \end{pmatrix} \Rightarrow V = \begin{pmatrix} \varphi_{nn}^2 & \varphi_{nn}\varphi_{cn} \\ \varphi_{nn}\varphi_{cn} & \varphi_{nn}^2 + \varphi_{cc}^2 \end{pmatrix}$$

One can write for the policyholder i

$$U_i = \begin{pmatrix} U_m \\ U_{ci} \end{pmatrix} = T_\varphi S_i; S_i = \begin{pmatrix} S_m \\ S_{ci} \end{pmatrix}, S_i \sim N(0, I_2),$$

and we have $U_i = f_{\theta_2}(S_i)$, φ being related to V , hence to θ_2 . The likelihood used in the bonus-malus expression (see equation (4)) is obtained as the product of the likelihoods related to numbers and costs. With the notations of 2.4, we have

$$\log f_*(Y_i/\theta_i, X_i, U) =$$

$$-\left(\sum_i \lambda_i\right) \exp(U_n) + \left(\sum_i n_i\right) U_n - \sum_{i,j} \frac{(\log c_{ij} - z_i \beta - U_{ci})^2}{2\sigma^2}, \text{ with}$$

$$X_i = (x_1, \dots, x_T); x_i = (w_i, z_i), Y_i = (y_1, \dots, y_T), y_i = (n_i, (c_{ij})_{j=1, \dots, n_i}),$$

plus terms that do not depend on the heterogeneity components. Replacing θ_i by $\hat{\theta}_i$, we obtain

$$f_*(Y_i/\hat{\theta}_i, X_i, U) = \exp(V_T) \times \text{terms independent from } U, \text{ with}$$

$$V_T = -\left(\sum_i \hat{\lambda}_i\right) \exp(U_n) + m_T U_n - \frac{m_T U_c^2 - 2U_c \text{lcres}_T}{2\hat{\sigma}^2} \tag{5}$$

A bonus-malus coefficient for a policyholder and for the period $T+1$ depends then on:

- $\sum_i \hat{\lambda}_i$, which is proportional to the frequency premium of the policyholder on all periods. This premium is equal to

$$\hat{E}(TN_T) = \sum_i \hat{\lambda}_i \hat{E}[\exp(U_n)] = \left(\sum_i \hat{\lambda}_i\right) \exp\left(\frac{\hat{\varphi}_{nm}^2}{2}\right) = \left(\sum_i \hat{\lambda}_i\right) \exp\left(\frac{\hat{V}_{nm}}{2}\right).$$

- m_T , the number of claims reported by the policyholder during the T periods
- lcres_T , the sum of residuals on the logarithm of costs of claims reported by the policyholder. It represents their relative severity.

From equation (4), bonus-malus coefficients on frequency, expected cost per claim, and pure premium are respectively equal to

$$\frac{\hat{E}[\exp(U_n + V_i)]}{\hat{E}[\exp(U_n)] \hat{E}[\exp(V_i)]}, \frac{\hat{E}[\exp(U_c + V_i)]}{\hat{E}[\exp(U_c)] \hat{E}[\exp(V_i)]}, \frac{\hat{E}[\exp(U_n + U_c + V_T)]}{\hat{E}[\exp(U_n + U_c)] \hat{E}[\exp(V_T)]}.$$

The coefficients are estimated by simulations of outcomes of S_n and S_c . For instance, we infer that the estimated covariance

$$\widehat{Cov}\left(\frac{\exp(U_n)}{E[\exp(U_n)]}, \frac{\exp(V_i)}{E[\exp(V_i)]}\right)$$

is a frequency-malus. The existence of boni and mali for the different risks can be interpreted through the sign of estimated covariances.

The a posteriori premium is obtained by the expression given in section 2.4

$$\hat{R}_{t+1}^{T+1} = \left(h_{\hat{\theta}_1}(x_{T+1}) E_{\hat{\theta}_2} [g(U)] \right) \frac{E_{\hat{\theta}} [g(U) / X_T, Y_T]}{E_{\hat{\theta}_2} [g(U)]}$$

The first term is the a priori premium. It is an estimation of

$$\lambda_{T+1} \exp(z_{T+1}\beta) E[\exp(U_n + U_c)] = \exp \left(w_{T+1}\alpha + z_{T+1}\beta + \frac{(\varphi_{nm} + \varphi_{cn})^2 + \varphi_{cc}^2}{2} \right),$$

because $U_n + U_c = (\varphi_{nm} + \varphi_{cn})S_n + \varphi_{cc}S_c$.

Besides, $(\varphi_{nm} + \varphi_{cn})^2 + \varphi_{cc}^2 = V_{nm} + 2V_{cn} + V_{cc}$.

We should have consistent estimators for the parameters, in order to derive bonus-malus coefficients. A method to obtain such estimators was quoted in the introduction. When applied to the preceding model, it leads to the following results.

We write $\hat{\alpha}^0, \hat{\beta}^0, \hat{\sigma}^2$ the estimators of the parameters in the a priori rating model, and $\hat{\lambda}_i = \sum_t \exp(w_t \hat{\alpha}^0), tlc_i = \sum_j \log(c_{ij}), E_{\hat{\theta}_1}(TLC_i) = \sum_u n_u z_u \beta, \hat{t}c_i = E_{\hat{\theta}_1^0}(TLC_i) = \sum_u n_u z_u \hat{\beta}^0$

The variances and covariances of the two heterogeneity components are consistently estimated by:

$$\hat{V}_{nm} = \log(1 + \hat{V}_{nm}^1), \hat{V}_{nm}^1 = \frac{\sum_i (n_i - \hat{\lambda}_i)^2 - n_i}{\sum_i \hat{\lambda}_i^2}; \hat{V}_{cn} = \frac{\sum_i (n_i - \hat{\lambda}_i)(tlc_i - \hat{t}c_i)}{\left(\sum_i \hat{\lambda}_i^2 \right) (1 + \hat{V}_{nm}^1)},$$

$$\hat{V}_{cc} = \frac{\sum_i \left[(tlc_i - \hat{t}c_i)^2 - n_i \hat{\sigma}^2 \right]}{\left(\sum_i \hat{\lambda}_i^2 \right) (1 + \hat{V}_{nm}^1)} - \hat{V}_{cn}^2 \tag{6}$$

Consistent estimators of $\varphi_{nm}, \varphi_{cn}$ and φ_{cc} are given by the solutions of the equation

$$T_{\hat{\varphi}} T'_{\hat{\varphi}} = \hat{V}$$

The estimators of φ are used in the computation of bonus-malus coefficients. remember that $U_i = T_{\varphi} S_i$ ($S_i \sim N(0, I_2)$), and that the coefficients are estimated through simulations of outcomes of S_i . As for the parameters of the a priori rating model, they are consistently estimated by

$$\hat{\alpha} = \hat{\alpha}^0 - \frac{\hat{V}_{nm}}{2} e_{n,1}, \hat{\beta} = \hat{\beta}^0 - \hat{V}_{cn} e_{c,1}, \hat{\sigma}^2 = \hat{\sigma}^2{}^0 - \hat{V}_{cc} \tag{7}$$

The intercepts are supposed to be the first of the k_n and k_c explanatory variables for the number and cost distributions, and $e_{n,1}$ (resp $e_{c,1}$) are the first vectors of the canonical base of \mathbb{R}^{k_n} (resp \mathbb{R}^{k_c})

3.2 Empirical results

The numerical results $\sum_i (n_i - \hat{\lambda}_i)^2 - n_i = 216.24$; $\sum_i \hat{\lambda}_i^2 = 389.48$, already used for bonus-malus on frequencies, lead to.

$$\hat{V}_{nn}^1 = \frac{\sum_i (n_i - \hat{\lambda}_i)^2 - n_i}{\sum_i \hat{\lambda}_i^2} = 0.555, \hat{V}_{nn} = \log(1 + \hat{V}_{nn}^1) = 0.442 \Rightarrow \hat{\phi}_{nn} = \sqrt{\hat{V}_{nn}} = 0.665$$

In this paper, two distribution families are considered for the heterogeneity component related to numbers. We first took into account the gamma, and now the log-normal family (writing the heterogeneity component in a multiplicative way)

Considering an insurance contract without claims, we can compare the boni derived from the two models. The sum $\sum_i \hat{\lambda}_i$ being the cumulated frequency premium in the negative binomial model, the bonus for the policyholder is equal to

$$1 - \frac{\hat{a}}{\hat{a} + \sum_i \hat{\lambda}_i} = \frac{\sum_i \hat{\lambda}_i}{\hat{a} + \sum_i \hat{\lambda}_i} = \frac{\hat{V}_{nn}^1 \sum_i \hat{\lambda}_i}{1 + (\hat{V}_{nn}^1 \sum_i \hat{\lambda}_i)}, (\hat{a} = 1 / \hat{V}_{nn}^1).$$

For the log-normal family, the bonus can be written as

$$- \widehat{Cov} \left(\frac{\exp(U_n)}{E[\exp(U_n)]}, \frac{\exp(V_T)}{E[\exp(V_T)]} \right), U_n = \phi_{nn} S_n, V_T = - \sum_i \hat{\lambda}_i \exp(U_n),$$

with $S_n \sim N(0,1)$. With the values of \hat{V}_{nn}^1 and $\hat{\phi}_{nn}$ computed precendently, one obtains for example

TABLE 3
COMPARISON OF FREQUENCY-BONUS COEFFICIENTS FOR TWO DISTRIBUTIONS ON THE HETEROGENEITY COMPONENT (CONTRACTS WITHOUT CLAIMS REPORTED)

frequency premium	0.05	0.1	0.2	0.5	1	2
bonus (% , gamma distributions)	2.7	5.3	10	21.7	35.7	52.6
bonus (% , log-normal distributions)	2.6	5.1	9.4	19.3	30.3	43.6

The boni derived from log-normal distributions on the heterogeneity component are lower than those derived from the gamma distributions. The difference is all the more important since the frequency premium is high

Let us estimate the covariance between the two heterogeneity components:

$$\sum_i (n_i - \hat{\lambda}_i)(tlc_i - \hat{tl}c_i) = 7.96 \Rightarrow \hat{V}_{cn} = \frac{\sum_i (n_i - \hat{\lambda}_i)(tlc_i - \hat{tl}c_i)}{\left(\sum_i \hat{\lambda}_i^2 \right) (1 + \hat{V}_{mn}^l)} = 0.013.$$

One can think of relating a positive or negative sign of the covariance to the fact that the average cost per claim increases or decreases with the number of claims reported by the policyholder. To see this, suppose that the duration of observation is the same for all the policyholders, and that the intercept is the only explanatory variable for number and cost distributions. We would then have

$$\hat{\lambda}_i = \bar{n}, \hat{tl}c_i = n_i \overline{\log c} \Rightarrow \sum_i (n_i - \hat{\lambda}_i)(tlc_i - \hat{tl}c_i) = \sum_i (n_i - \bar{n})n_i(\overline{\log c}' - \overline{\log c}) = \sum_{i/n_i \geq 2} (n_i - 1)n_i(\overline{\log c}' - \overline{\log c}), \text{ because } \sum_i n_i(\overline{\log c}' - \overline{\log c}) = 0.$$

We wrote $\overline{\log c}'$ for the logarithms of costs of claims reported by the policyholder i , computed on average. The estimator of the covariance would be positive if the average of the logarithms of costs of claims related to the policyholders that reported several of them was superior to the global mean.

On the working sample, the number of claims reported by the policyholder had little influence on the average cost.

The preceding results justify the allowance for a non constant number of periods related to the observation of policyholders. To see this, we remark that the more severe is a claim, the greater is the probability to change the vehicle afterwards. Hence, there is less severity on average for several claims reported on the same car. If policyholders were not kept in the sample after changing cars, a negative bias would appear in the estimation of the correlation coefficient between the heterogeneity components. Now, keeping the policyholder in the sample as long as possible leads us to consider a non constant number of periods.

When computing bonus-malus coefficients for average cost per claim, we used (see 2.7.2)

$$\sum_i \left[(tlc_i - \hat{tl}c_i)^2 - n_i \hat{\sigma}^2 \right] = \sum_{i/n_i \geq 2} \sum_{k \in S_i, j \neq k} lcr_{es_{ij}} lcr_{es_{ik}} = 100.80$$

A bonus-malus system for average cost per claim can be considered if the observation of the ratio actual cost-expected cost for a claim brings information for the following claims. If the last expression is positive, the cost residuals of claims related to policyholders having reported several of them have rather the same sign. The relative severity of a claim is associated to the sign of the residual, and it may be interesting to compare the sign of residuals for claims related to policyholders having reported two of them.

Considering the working sample, we obtain

number of policyholders having reported two claims	negative residual (second claim)	positive residual (second claim)
negative residual (first claim)	74	46
positive residual (first claim)	36	70

The sign of the residual does not change for 64% of policyholders having reported two claims

From equation (6), we infer

$$\hat{V}_{cc} = \frac{\sum_i (tlc_i - t\hat{c}_i)^2 - n_i \hat{\sigma}^2}{\left(\sum_i \hat{\lambda}_i^2\right) (1 + \hat{V}_{mm})} - \hat{V}_{cn}^2 = 0.166, \text{ and } \hat{r}_{cn} = \frac{\hat{V}_{cn}}{\sqrt{\hat{V}_{cc} \hat{V}_{nn}}} = 0.048$$

The correlation coefficient between the heterogeneity components is positive, but close to zero. Hence

$$\hat{V}_{cn} = \hat{\phi}_{nn} \hat{\phi}_{cn} \Rightarrow \hat{\phi}_{cn} = 0.020, \hat{V}_{cc} = \hat{\phi}_{cn}^2 + \hat{\phi}_{cc}^2 \Rightarrow \hat{\phi}_{cc} = 0.407$$

The boni for average cost per claim and pure premium for the contracts without claims can be computed, and results can be compared to those obtained for frequency. From the expressions

$$- \widehat{Cov} \left(\frac{\exp(U_c)}{E[\exp(U_c)]}, \frac{\exp(V_T)}{E[\exp(V_T)]} \right), - \widehat{Cov} \left(\frac{\exp(U_n + U_c)}{E[\exp(U_n + U_c)]}, \frac{\exp(V_T)}{E[\exp(V_T)]} \right)$$

we obtain

TABLE 4
BONI FOR AVERAGE COST PER CLAIM AND PURE PREMIUM (CONTRACTS WITHOUT CLAIM REPORTED)

frequency premium	0.05	0.1	0.2	0.5	1	2
average cost per claim bonus (%)	0.1	0.1	0.2	0.5	0.9	1.5
pure premium bonus (%)	2.7	5.3	9.7	19.9	31.2	44.7

Because of the positive correlation between the two heterogeneity components, a cost-bonus appears in the absence of claims, but it is very low.

We now compute bonus-malus coefficients for policyholders that reported one claim. They are a function of the cost-residual $lcr_{es7} = \log(c_1) - z_1 \hat{\beta}$ (c_1 is the cost of the claim, and z_1 represents the policyholder's characteristics when the claim occurred), and of the frequency premium. From equations (5) and (7), we have

$$V_T = -\sum_i \hat{\lambda}_i \exp(U_n) + U_n - \frac{U_c^2 - 2U_c \text{lcres}_T}{2\hat{\sigma}^2},$$

$$\hat{\sigma}^2 = \hat{\sigma}^2 - \hat{V}_{cc} = \frac{\sum \text{lcres}_{ij}^2}{n} - \hat{V}_{cc} = \frac{3588}{3493} - 0.166 = 0.861$$

We recall that the bonus-malus coefficients on frequency, expected cost per claim and pure premium are respectively equal to

$$\frac{\hat{E}[\exp(U_n + V_T)]}{\hat{E}[\exp(U_n)] \hat{E}[\exp(V_T)]}, \frac{\hat{E}[\exp(U_c + V_T)]}{\hat{E}[\exp(U_c)] \hat{E}[\exp(V_T)]}, \frac{\hat{E}[\exp(U_n + U_c + V_T)]}{\hat{E}[\exp(U_n + U_c)] \hat{E}[\exp(V_T)]}.$$

We obtain for example (the bonus-malus coefficients are given in percentage)

TABLE 5
BONUS-MALUS COEFFICIENTS (POLICYHOLDERS HAVING REPORTED ONE CLAIM)

frequency coefficient <i>lcres_T</i>	frequency premium					
	0.05	0.1	0.2	0.5	1	2
-1	147.4	142.1	133.1	113.9	94.5	73.4
-0.5	148.4	143	133.8	114.5	95	73.7
0	149.3	143.7	134.6	115	95.3	74
0.5	150.1	144.6	135.3	115.6	95.7	74.3
1	151	145.6	136	116.1	96.2	74.6

average cost per claim coefficient <i>lcres_T</i>	frequency premium					
	0.05	0.1	0.2	0.5	1	2
-1	84.8	84.7	84.6	84.3	84	83.5
-0.5	92	91.9	91.7	91.4	91	90.5
0	99.7	99.6	99.5	99.1	98.7	98.1
0.5	108.1	108	107.8	107.5	107	106.4
1	117.1	117	116.9	116.5	116	115.4

pure premium coefficient <i>lcres_T</i>	frequency premium					
	0.05	0.1	0.2	0.5	1	2
-1	124.6	120	112.2	95.6	78.9	60.9
-0.5	136.1	131	122.3	104.2	86	66.3
0	148.4	142.7	133.3	113.5	93.5	72.2
0.5	161.8	155.7	145.4	123.7	101.9	78.5
1	176.6	170	158.4	134.7	111	85.4

Because of the positive correlation between the two heterogeneity components, the frequency coefficients increase with the cost-residual, which is related to the severity of the claim. In the same way, the coefficients related to average cost per claim decrease with the frequency premium, but these variations are very low. Because of the correlation, the coefficients related to pure premium are not equal to the product of the

coefficients for frequency and expected cost per claim. Here also, differences are very low

4. CONCLUDING REMARKS

We recall the main results obtained in this paper

- The unexplained heterogeneity with respect to the cost distributions depends strongly on the choice of the distribution family.
- Besides, it is revealed more slowly throughout time than for number distributions
- On the working sample, the correlation between the heterogeneity components on the number and cost distributions is very low.

In the long run, it would be desirable to relax the assumption of invariance of the heterogeneity components with respect to time. Because of this invariance, the age of claims has no influence on the bonus-malus coefficients. Now, the fact that an ancient claim has the same influence on the coefficients that a recent one is questionable. The allowance for an innovation at each period for the heterogeneity components would raise new problems, and would make it necessary to observe policyholders on many periods.

REFERENCES

- BUHLMANN, H (1967) Experience Rating and Credibility *ASTIN Bulletin* **4**, 199-207
- CUMMINS, J. D., DIONNE, G., MC DONALD, J. B. and PRITCHETT, B. M. (1990) Application of the GB2 Distribution in Modelling Insurance Loss Processes *Insurance Mathematics and Economics* **9**, 257-272
- DIONNE, G. and VANASSE, C. (1989) A Generalization of Automobile Insurance Rating Models: The Negative Binomial Distribution with a Regression Component *ASTIN Bulletin* **19**, 199-212
- DIONNE, G. and VANASSE, C. (1992) Automobile Insurance Ratemaking in the Presence of Asymmetrical Information *Journal of Applied Econometrics* **7**, 149-165
- LEMAIRE, J. (1985) *Automobile Insurance Actuarial Models*. Huebner International Series on Risk, Insurance and Economic Security
- LEMAIRE, J. (1995) *Bonus-Malus Systems in Automobile Insurance*. Huebner International Series on Risk, Insurance and Economic Security
- PINQUET, J., ROBERT, J. C., PESTRE, G. and MONTOCCHIO, L. (1992) Tarification a Priori et a Posteriori des Risques en Assurance Automobile *Mémoire au Centre d'Etudes Actuarielles*
- PINQUET, J. (1996a) Allowance for Costs of Claims in Bonus-Malus Systems *Proceedings of the ASTIN colloquium, Copenhagen 1996*
- PINQUET, J. (1996b) Hétérogénéité Inexpliquée *Document de travail THEMA n° 11*
- RAO, C. R. (1948) Large Sample Tests of Statistical Hypothesis Concerning Several Parameters with Applications to Problems of Estimation *Proceedings of the Cambridge Philosophical Society* **44**, 50-57
- RENSHAW, E. A. (1994) Modelling the Claims Process in the Presence of Covariates *ASTIN Bulletin* **24**, 265-285
- SILVEY, S. D. (1959) The Lagrange Multiplier Test *Annals of Mathematical Statistics* **30**, 389-407

JEAN PINQUET

Université de Paris X, U F R de Sciences Economiques
 200, Avenue de la République 92001 NANTERRE CEDEX
 Phone 33 1 49 81 72 45, Fax: 33 1 40 97 71 42,
 E-mail. pinquet@u-paris10.fr

