

# COMPARING CREDIBILITY ESTIMATES OF HEALTH INSURANCE CLAIMS COSTS

Gilbert W. Fellingham,\* H. Dennis Tolley,<sup>†</sup> and Thomas N. Herzog<sup>‡</sup>

---

## ABSTRACT

We fit a linear mixed model and a Bayesian hierarchical model to data provided by an insurance company located in the Midwest. We used models fit to the 1994 data to predict health insurance claims costs for 1995. We implemented the linear mixed model in SAS and used two different prediction methods to predict 1995 costs. In the linear mixed model we assumed a normal likelihood. In the hierarchical Bayes model, we used Markov chain Monte Carlo methods to obtain posterior distributions of the parameters, as well as predictive distributions of the next year's costs. We assumed the likelihood for this model to be a mixture of a gamma distribution for the nonzero costs, with a point mass for the zero costs. All prediction methods use credibility-type estimators that use relevant information from related experience. The linear mixed model was heavily influenced by the skewed nature of the data. The assumed gamma likelihood of the full Bayesian analysis appeared to underestimate the tails of the distributions. All prediction models underestimated costs for 1995.

---

## 1. BACKGROUND

One of the goals of the Society of Actuaries' Credibility for Health Coverages Task Force organized several years ago was to investigate the utility of credibility theory as a practical method of estimating and updating premium calculations for health insurance products. In connection with the work of this task force, a major health insurance provider provided a database summarizing recent claims experience for select health insurance coverages in Illinois and Wisconsin. These data were made available to the authors to illustrate the calculations and utility of credibility methods.

Credibility methods are actuarial techniques of using data from insurance claims (experience) to estimate claims in the future. Some of these

methods are ad hoc, while others are founded on statistical principles. Both empirical Bayes and full Bayesian methods would be identified by the practicing actuary under the rubric of "credibility" methods. Although credibility methods have been used for years in the casualty insurance area, only recently have these methods been applied to health insurance data. This new application also brings with it a more complex problem in two ways. First, the assumptions associated with the analysis may be more complicated than those of the linear mixed model, which assumes normal data and normal random effects. Second, health measures, risk factors, and type of care show a high level of variety, eventually mandating that credibility be viewed with models containing many covariates and possibly multivariate response profiles. This paper illustrates the use of currently available statistical techniques to initiate solutions to this more difficult application of credibility methods.

## 2. INTRODUCTION

The purpose of this paper is to use both a linear mixed model (LMM) as implemented in SAS (Littell et al. 1996) and a Bayesian hierarchical model (BHM) to estimate health care costs with an eye

---

\* Gilbert W. Fellingham, PhD, is a Professor in the Department of Statistics, Brigham Young University, Provo, UT 86402, e-mail: gwf@byu.edu.

<sup>†</sup> H. Dennis Tolley, ASA, PhD, is a Professor in the Department of Statistics, Brigham Young University, Provo, UT 86402, e-mail: toolley@byu.edu.

<sup>‡</sup> Thomas N. Herzog, ASA, PhD, is Chief Actuary at the Federal Housing Administration for the U.S. Department of Housing and Urban Development, Washington, DC 20410, e-mail: thomas\_n\_herzog@hud.gov.

to predicting expected costs for the coming year. These models could be used in premium calculations for small groups, and in premium calculations for blocks of business in new areas, as well as to calculate experience-based refunds. As methods for health finance improve, the need for a full Bayesian treatment of the problems faced by the industry is expected to grow.

One of the underlying principles of these methods is improvement of the process of estimating, say, the pure premium for a block of business, by “borrowing strength” from related experience. For example, if the exposure for a block is small enough, the experience for the previous years may be limited. In this case estimates of future costs may be based on a combination of previous experience with other, related experience in an effort to mitigate the effects of random variation on the estimation process. Also, the thoughtful use of expert opinion along with current experience in other market areas could be used to improve prediction of expected experience in new markets.

Various paradigms for estimating expected experience have been presented in detail by several authors (see, e.g., Herzog 1999; Klugman 1992). Although these presentations provide the essentials for involved problems, even in these simpler paradigms the examples are restricted to smaller, textbook-type problems. Several authors (Morris and Van Slyke 1978; Venter 1996; Tolley, Nielsen, and Bachelor 1999) have provided computational shortcuts for more involved models.

Two key papers that synthesize several credibility models are Frees et al. (1999) and Frees et al. (2001). In these papers the authors make explicit the relationship between credibility procedures and the parametric statistical methods used for longitudinal and panel data analysis.

An implementation of a full Bayesian approach has been possible only for a limited class of mod-

els until the advent of Markov chain Monte Carlo (MCMC) numerical methods. Scollnik (2001) shows how to implement Bayesian methods in actuarial models using currently available software. The data set we use for illustration is fairly large and requires estimation of nearly 3,600 parameters. The size of the data set mitigates the necessity of eliciting precise prior information. However, the methods would clearly generalize to smaller problems, where more precise prior information would be useful.

### 3. THE DATA

The data set is from a major medical plan, covering a block of medium-sized groups in Illinois and Wisconsin for 1994 and 1995. Each policyholder was part of a group plan. The groups consisted of from 1 to 280 employees with a median size of 24 and an average size of 35.2 in 1994. We have claims information on 40,631 policyholders from 1,177 groups. Policies were of three basic types: (1) employee only (employee), (2) employee plus one individual (spouse), and (3) employee plus multiple individuals (family). Table 1 gives the number of employees and summary information about costs per day for the different policy types. Although the data are dated from a business perspective, they provide the ability to examine a full Bayesian analysis without divulging proprietary information. Only data from 1994 are used for building the model; 46,691 observations from 1995 are used to validate model predictions.

Data consist of claims costs by policyholder. While age and gender of policyholder were known, age and gender of the claimant were known only when the claimant was the policyholder.

Costs were assigned to each policyholder on a yearly basis and not assigned by episode of care or by medical incident. The costs were total costs,

Table 1  
Costs per Day in Dollars for 1994 Summarized by Policy Type

Policy Type	<i>n</i>	Mean	Std. Dev.	Median	Maximum	Proportion Zero Claims
Employee	22,618	2.77	12.74	0	823.45	0.574
Spouse	8,921	6.79	21.01	1.11	643.02	0.315
Family	9,092	7.97	14.81	2.44	277.64	0.211

with deductible and copayments added back in. The total yearly costs were then divided by the number of days of exposure. As per the policy of the company providing the data, all policies with annual claims costs exceeding \$25,000 were excluded from all analyses, including the 1995 prediction set. Large daily costs are still possible if the number of days of exposure were small enough that total costs did not exceed \$25,000.

## 4. THE MODELS

### 4.1 The Linear Mixed Model

The data set did contain some covariate information about the insureds on which to base premium estimates; however, these data were limited. For example, although we knew both the policy type and group for each policy, we did not know the individual identification of any claimant from the policy. Thus, we did not have access to information on gender or age of claimant, or if multiple claims were made on the policy by the same individual or different individuals covered by the policy during the year. Whether or not the policy covered the same people from year to year was not known. Consequently we will use only policy type and group as predictors of cost to illustrate and compare the mixed-model methods with the Bayesian methods of estimating future costs. Knowledge and use of additional policy and claimant-specific data would improve prediction. However, such information would also make the presentation more difficult to follow with the additional detail. We use the data available to present an expository illustration. The methods are readily extended to more involved data sets.

In the classical terminology associated with mixed models (see, e.g., Littell et al. 1996), parameters associated with policy type would be considered “fixed.” This means that our predictions are to be based on the same set of three policy types considered here. On the other hand, groups are “random,” meaning that predictions of costs will not be restricted to only a fixed set of insured groups. In essence, we consider the risk characteristics of a group of policies to be a random draw from a pool of risks. This draw is a one-time event, meaning that whatever risk characteristics were drawn from the pool for a specific group remain the same over time. Thus, experi-

ence gained on a specific group in one year can be used to predict outcomes in future years.

The general form of the linear mixed model (see Frees et al. 1999, 2001) is as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad (4.1)$$

where  $\mathbf{Y}$  is an  $n \times 1$  vector representing the observed data (in this example  $n$  is 40,631),  $\mathbf{X}$  is an  $n \times p$  design matrix of known constants ( $p$  is the number of fixed effects, in this case three policy types),  $\boldsymbol{\beta}$  is a  $p \times 1$  vector representing the fixed effects of the policy types,  $\mathbf{Z}$  is an  $n \times q$  design matrix of known constants ( $q$  is the number of random effects, in this case 1,177 groups),  $\mathbf{u}$  is a  $q \times 1$  vector representing random effects of the groups, and  $\mathbf{e}$  is an  $n \times 1$  vector representing the errors. The distributional assumptions are that  $\mathbf{u}$  and  $\mathbf{e}$  are independent multivariate normal (MVN) random variables with  $\mathbf{u} \sim \text{MVN}(\mathbf{0}, \mathbf{G})$ ,  $\mathbf{e} \sim \text{MVN}(\mathbf{0}, \mathbf{R})$ . The vector  $\boldsymbol{\beta}$  is a vector of fixed regression parameters to be estimated.

Equation (4.1) provides the basis for predicting future costs. Heuristically we can envision the estimation process as follows. We assume values for  $\mathbf{G}$  and  $\mathbf{R}$ , whether known or estimated, and estimates of the vector  $\boldsymbol{\beta}$  obtained from previous data. For any policy type we will know the set of  $X$  values. The group to which the policy belongs will specify the  $Z$  values. We predict the costs of any such policy by plugging into equation (4.1) the estimates of  $\boldsymbol{\beta}$ , the appropriate  $X$  and  $Z$  values, and then drawing a random vector of  $\mathbf{u}$  values and solving. Any temporal effects on costs, such as medical inflation costs or changes in utilization patterns, would be included in the  $X$  values to have an estimate adjustment in the vector  $\boldsymbol{\beta}$ . In this paper we made no such temporal judgment and therefore expect a downward bias in our predictions due to medical inflation.

Concerning the draw of the random vector  $\mathbf{u}$ , one of two conditions obtain. In the first case the group is the same group on which experience already exists (the classical credibility problem). In this case a new draw of risk characteristics is unnecessary since risk characteristics have already been “drawn,” and the effect of these characteristics for the group can be estimated from this experience. In the second case the group is a new group for which no previous experience is available. In this case we simulate the costs by

repeatedly drawing realizations of  $u$  and plugging them into equation (4.1). The resulting set of costs will be an estimate of the distribution of costs one might expect for this group. In actuality, of course, we need not formally make all the draws, but rather substitute for  $u$  its expected value and calculate the predicted costs directly. In either case the estimate of future costs is a linear equation in the  $\beta$  and  $u$  parameters. Equations for making these estimates are given below.

$$\hat{C} = \begin{pmatrix} (X'\hat{V}^{-1}X)^{-1} & -(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Z\hat{G}' \\ -\hat{G}Z'\hat{V}^{-1}X(X'\hat{V}^{-1}X)^{-1} & (Z'\hat{R}^{-1}Z + \hat{G}^{-1})^{-1} + \hat{G}Z'\hat{V}^{-1}X(X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}Z\hat{G}' \end{pmatrix}. \quad (4.5)$$

It follows from the distributional assumptions that  $Y \sim MVN(X\beta, V)$  where  $V = ZGZ' + R$ . If  $V$  were known, the best linear unbiased estimate (the maximum likelihood estimate) of  $\beta$  would be the generalized least squares estimate given as

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y. \quad (4.2)$$

When  $V$  is unknown,  $G$  and  $R$  are usually estimated using either the full likelihood (ML) or a restricted form of the likelihood (REML), and then  $\beta$  is estimated using generalized least squares with  $V$  replaced by  $\hat{V}$ . Since this approximation procedure is well established and programmed in some statistical packages, we will use the notation and representation of estimates common in statistical literature. Note that using the generalized least squares approach assumes that we can get consistent estimates of  $G$  and  $R$  and then use these to form estimates of  $\beta$  and  $u$ . The assumption of normality in  $u$  and  $e$  is used to form an estimation procedure. One might generalize this to an  $m$ -estimate in future work. Here we proceed without concern about this assumption. One might try to appeal to the Central Limit Theorem to reduce the effect of many zero values in the data. However, even with this fairly large data set, the sample sizes within groups are limited. Alternatively, one might try a conditional analysis conditioned on a positive claim amount. In summary, there is need for caution in forming estimates of  $G$  and  $R$  using likelihood methods where many responses are zero.

Using the mixed model shown in equation (4.1), a linear combination  $L$  of the parameters (e.g., for estimates of future costs) is estimated as

$$L \begin{pmatrix} \hat{\beta} \\ \hat{u} \end{pmatrix}, \quad (4.3)$$

where  $\hat{u} = \hat{G}Z'\hat{V}^{-1}(Y - X\hat{\beta})$  and where  $\hat{\beta}$  is as in equation (4.2) with  $V$  replaced by its estimate  $\hat{V}$ . Standard errors of these estimates are computed as

$$\sqrt{L\hat{C}L'}, \quad (4.4)$$

where

See McLean and Sanders (1988) for more details.

To make predictions of future cost, we must specify the  $X$  values signifying the type of policy and an indicator  $Z$  for the group. These make up the  $L$  in equation (4.3). Letting  $X_m$  denote the policy type, and  $Z_m$  the group, then equation (4.3) reduces to

$$\hat{Y} = X_m\hat{\beta} + \hat{C}_m\hat{V}^{-1}(Y - X\hat{\beta}), \quad (4.6)$$

where  $\hat{C}_m = Z_m\hat{G}Z'$ .

Equation (4.6) resembles the classical credibility formula where the overall experience of the block, estimated by the first term on the right-hand side of equation (4.6), is adjusted for the experience of the group, the second term on the right-hand side of equation (4.6). In statistical terminology the estimates produced by equation (4.6) are called empirical best linear unbiased predictions (EBLUPs) for costs. In the case where the policy of interest is from a new group, the indicator vector,  $Z_m$ , will contain only zeros. In this case  $C_m$  will be null, and the second term will be zero. This indicates that there is no credibility adjustment for past experience as there is no past experience.

To quantify the improvement in estimation, we examine the variance of the estimate given by equation (4.6), where the variance of predicted costs is

$$\hat{V}_m - \hat{C}_m\hat{V}^{-1}\hat{C}_m' + [X_m - \hat{C}_m\hat{V}^{-1}X] \times (X'\hat{V}^{-1}X)^{-1}[X_m - \hat{C}_m\hat{V}^{-1}X]', \quad (4.7)$$

where  $\hat{V}_m$  is the model-based variance matrix of  $Y$ .

We note that the value of  $\hat{V}_m$  depends on the general sampling model. In the case we consider,  $\hat{V}_m = \hat{\sigma}_{error}^2 + \hat{\sigma}_{group}^2$ . The two improvements on using a mixed-model credibility estimate are the adjustment

for group-specific experience in equation (4.6) and the reduction of the variance of predicted costs as given by the second term in equation (4.7).

SAS's approach in Proc Mixed is the one mentioned previously, where the variance components  $\mathbf{G}$  and  $\mathbf{R}$  are estimated using either the full likelihood (ML) or a restricted or residual form of the likelihood (REML) (Littell et al. 1996). Using  $\hat{\mathbf{V}}$  instead of  $\mathbf{V}$  in equation (4.2) gives estimated generalized least squares estimates of the  $\boldsymbol{\beta}$ 's. REML is the default method of Proc Mixed and is used in our analysis.

There are at least two different methods of calculating the predictions given by equation (4.6) using SAS. SAS will automatically produce predicted values for any new group that has only the independent (or covariate) terms entered and no dependent variable. The first method is to use the `random` command. In this case SAS will automatically determine the EBLUPs using equation (4.6). However, the variance quoted by SAS is the variance of the estimate (the third term in equation [4.7]). The  $\hat{\mathbf{V}}_m$  term,  $\hat{\sigma}_{error}^2 + \hat{\sigma}_{group}^2$ , must be added to this estimate, and the second term,  $\hat{\sigma}_{group}^2$ , subtracted.

The second SAS method is the `repeated` option. Since the covariance structure in the problem is compound symmetric ( $\sigma_{error}^2 + \sigma_{group}^2$  on the diagonal,  $\sigma_{group}^2$  off the diagonal of the covariance matrix), identical estimates of the variance components (and by extension of  $\hat{\boldsymbol{\beta}}$ ) are available by using the `repeated` statement in SAS, where groups are treated as the factor of repetition. Using this approach, no random effects are estimated, but EBLUPs are still available.

Theoretically the two procedures will yield identical values for the EBLUPs. However, the computational algorithms are not the same, and thus the estimates often do not agree in every decimal place. Also, the degree to which the variance terms are adjusted by SAS for  $\hat{\mathbf{V}}_m$  as noted above is different, although the variance of the estimates will be the same once the adjustment is made. We note here that the `repeated` command in SAS appears to adjust out incorrectly  $\hat{\sigma}_{group}^2$  (the second term in equation [4.7]) when the policy predictions are for policies from a new group. In this case  $\hat{\mathbf{C}}_m \hat{\mathbf{V}}^{-1} \mathbf{C}'_m$  would be 0, not  $\hat{\sigma}_{group}^2$ .

Estimates were computed using both formulations and compared to each observed data point in 1995. We will make reference to this model as LMM when the formulations yield identical re-

sults. However, when referring to predicted values, these two formulations will be called LMM (random) and LMM (repeated).

## 4.2 The Hierarchical Bayes Model

In the Bayesian framework the model consists of the likelihood of the data given the parameters, multiplied by probability densities for each of the parameters. The densities on the parameters are called the "prior" probabilities as they are formulated prior to the collection of the data. Based on Bayes theorem, posterior densities for the parameters given the data then are available from the scaled product of the likelihood and the priors. (For a review of Bayesian methods in general see, e.g., Gelman et al. 1998, Klugman 1992, Scollnik 2001, or Makov 2001.) We call this model hierarchical since subjects are considered to be a random draw of all possible subjects within a group, and each group is considered to be a random draw from all possible groups. It is this sequential nesting that gives the model its hierarchical structure.

Two things must be considered when thinking about the form of the likelihood. Propensity, the probability that a claim is made, differs from group to group and in our data is around 0.60. Thus, about 40% of the data are zeros, representing no claims. We chose to deal with this by having a likelihood with a point mass at zero with probability  $\pi_{g_i}$  for group  $i$ . The parameter  $\pi_{g_i}$  depends on the group membership. It would also be reasonable to include a parameter for propensity based on the policy type, and such a parameterization should be pursued in future analyses. Severity, the cost of a claim given that a claim is paid, is positively skewed. We chose a gamma density for this portion of the likelihood with parameters  $r$  and  $\theta$ . Since we desire to estimate both policy type and group effects for this portion of the likelihood, we write  $r$  as  $r_{p_j} + r_{g_i}$  and  $\theta$  as  $\theta_{p_j} + \theta_{g_i}$ . Thus we estimate a component of the  $r$  parameter for each group ( $r_{g_i}$ ) and each policy type ( $r_{p_j}$ ) and a component of the  $\theta$  parameter for each group ( $\theta_{g_i}$ ) and each policy type ( $\theta_{p_j}$ ). Implicit in this formulation is the assumption that policy type and group effect parameters are additive. This is not the same as assuming additivity in an analysis of variance model, or in assuming that the effect is the sum of random variables. Rather, this additivity assumption is a simple method of generating a prior for  $r$  and  $\theta$  by component realizations of hyperpriors.

Although the components have gamma distributions, their sums, in general, will not. While other formulations are clearly possible, this likelihood seems to us to be reasonable. The likelihood follows using a compound distribution argument:

$$\prod_{i=1}^{1,177} \prod_{j=1}^3 \prod_{k=1}^{j(i)} \left[ \pi_{g_i[y_{ijk}=0]} + (1 - \pi_{g_i}) \times \left( \frac{1}{(\theta_{p_j} + \theta_{g_i})^{(r_{p_j} + r_{g_i})} \Gamma(r_{p_j} + r_{g_i})} \times y_{ijk}^{r_{p_j} + r_{g_i} - 1} e^{[-y_{ijk}/(\theta_{p_j} + \theta_{g_i})]} \right) \right]_{[y_{ijk} > 0]}, \quad (4.8)$$

where

$i$  indexes the group number

$j$  indexes the policy type

$k$  indexes the observation within a specific group and policy type

$j(i)$  is the number of observations of policy type  $j$  within group  $i$

$\pi_{g_i}$  is the propensity parameter for group  $i$

$\theta_{p_j}$  and  $r_{p_j}$  are the severity parameters for policy type  $j$

$\theta_{g_i}$  and  $r_{g_i}$  are the severity parameters for group  $i$  and

$y_{ijk}$  is the cost per day of exposure for each policyholder.

Thus, we have a point mass probability for  $y_{ijk} = 0$ , and a gamma likelihood for  $y_{ijk} > 0$ .

The assignment of prior distributions could be a critical part of an analysis where data are limited. One of the strengths of the full Bayesian approach is the ability that the analyst has to incorporate information from other sources. In this case we had 40,631 data points for 1,177 groups, so the likelihood will dominate the analysis whatever prior information is used. We chose priors that seemed to us to be reasonable, but knowing they have only modest influence on the result.

We assigned each parameter a prior distribution as follows:

$$\theta_{p_j} \sim \text{Gamma}(2, 2)$$

$$r_{p_j} \sim \text{Gamma}(2, 2)$$

$$\pi_{g_i} \sim \text{Beta}(c_g, d_g)$$

$$\theta_{g_i} \sim \text{Gamma}(a_g, b_g)$$

$$r_{g_i} \sim \text{Gamma}(e_g, f_g).$$

The parameters associated with each group, in classical parlance the random factors, have another set of parameters in their prior distributions. In the BHM it is the depth of the hierarchical structure that is used to distinguish between what are known as fixed and random factors in the LMM (see, e.g., Brophy and Joseph 2000; Stangl and Berry 2000). These parameters, known as hyperparameters, also need a prior structure. These priors are called hyperpriors. Since, as we have already noted, the prior distributions will have minimal impact on the posteriors in this case (and by extension of that argument, the hyperpriors will have an even smaller effect), we assigned the same hyperprior to each of the parameters  $a_g$ ,  $b_g$ ,  $c_g$ ,  $d_g$ ,  $e_g$ , and  $f_g$ . We assumed each was distributed as an *Exponential*(1). We note that while the choice of hyperparameters should not materially affect posterior distributions, choices for hyperparameters can affect dramatically convergence in a model of this size if the hyperprior distributions are too steep. We ran the same model with hyperprior distributions of *Exponential*(0.1) with no change in posteriors. However, steeper hyperprior distributions led to MCMC runs that clearly had not converged, even after runs on the order of 80,000 iterations.

This formulation yields a full posterior proportional to

$$e^{-a_g} e^{-b_g} e^{-c_g} e^{-d_g} e^{-e_g} e^{-f_g} \left[ \prod_{j=1}^3 \theta_{p_j} e^{(-\theta_{p_j}/2)} r_{p_j} e^{(-r_{p_j}/2)} \right] \times \left[ \prod_{i=1}^{1,177} \frac{b_g^{-a_g}}{\Gamma(a_g)} \theta_{g_i}^{a_g-1} e^{(-\theta_{g_i}/b_g)} \frac{\Gamma(c_g + d_g)}{\Gamma(c_g) + \Gamma(d_g)} \pi_{g_i}^{c_g-1} \times (1 - \pi_{g_i})^{d_g-1} \frac{f_g^{-e_g}}{\Gamma(e_g)} r_{g_i}^{e_g-1} e^{(-r_{g_i}/f_g)} \right] \times \left[ \prod_{i=1}^{1,177} \prod_{j=1}^3 \prod_{k=1}^{j(i)} \left\{ \pi_{g_i[y_{ijk}=0]} + (1 - \pi_{g_i}) \times \left( \frac{1}{(\theta_{p_j} + \theta_{g_i})^{(r_{p_j} + r_{g_i})} \Gamma(r_{p_j} + r_{g_i})} \times y_{ijk}^{r_{p_j} + r_{g_i} - 1} e^{(-y_{ijk}/(\theta_{p_j} + \theta_{g_i}))} \right) \right\} \right]_{[y_{ijk} > 0]}. \quad (4.9)$$

Posterior distributions for such a complicated model are not available in closed form. Current methods to analyze such a model include MCMC implementation to produce samples from the posterior distributions, which can then be evaluated (Gilks, Richardson, and Spiegelhalter 1995). MCMC is essentially Monte Carlo integration using Markov chains. Monte Carlo integration draws samples from the required distribution and then forms sample averages to approximate expectations. MCMC draws these samples by running a cleverly constructed Markov chain for a long time. There are many ways of constructing these chains, but all of them are special cases of the general framework of Metropolis et al. (1953) and Hastings (1970). Loosely speaking, the MCMC process draws samples from the posterior distributions by sampling throughout the appropriate support in the correct proportions. This is done using a Markov chain with the posterior as its stationary distribution.

More precisely, we first formulated the posterior distribution of each parameter, conditional on the other parameters and assigned an initial value to each parameter. Then a new value is drawn from a “proposal” distribution. The ratio of the values of the complete conditionals computed using the proposed value and the old value of the parameters is computed and compared to a random uniform variate. If the ratio exceeds the random uniform, the proposed value is kept; otherwise the old value is kept. Using this method on each parameter, and cycling through the parameters, yields a distribution that converges to the appropriate posterior for each parameter. For a more complete exposition of this methodology, see Scollnik (2001).

## 5. RESULTS

### 5.1 The Groups

Besides predicting outcomes for all individuals in 1995, we also chose three groups to model in more detail. We selected group 115 primarily because it contained the largest data point, \$823.45 per day, although in other respects it was quite similar to the other groups. Group 980 was included because of high propensity and because it was a small group. Group 1,034 was chosen primarily because the severity data were heavily right skewed, especially in the spouse policy type, and because it was one of the larger groups. In Table 2 we show summary statistics for the raw data in these three groups in 1994.

### 5.2 The Chain

The MCMC procedures of the BHM involve sampling from the posterior distributions of the parameters of interest, as well as from the predictive distributions. Although we know that our methodology guarantees convergence to the appropriate distributions, we do not know with certainty that any particular chain of sampled values has converged to the appropriate place. With a model of this many parameters, convergence may not happen quickly. To maximize the probability of the chains converging, we proceeded as follows.

We ran 70,000 burn-in iterations followed by 10,000 iterations sampled every second one for 5,000 samples from the posterior distributions of the parameters. We saved all 5,000 samples for the policy-type severity parameters but computed only means and standard deviations for the 1,177 group severity and propensity parameters.

Table 2  
**Raw Data Summary Information for Three Selected Groups, 115, 980, and 1,034, for 1994**

Group	Policy Type	N	Mean	Std. Dev.	Minimum	Maximum
115	Individual	26	35.81	161.04	0.00	823.45
115	Spouse	2	29.75	18.20	16.89	42.62
115	Family	4	11.84	13.31	2.43	21.25
980	Individual	4	7.19	12.72	0.08	26.23
980	Spouse	2	33.77	41.82	4.20	63.34
980	Family	6	9.68	8.88	2.31	25.33
1,034	Individual	41	14.56	42.46	0.00	193.58
1,034	Spouse	85	56.92	115.01	0.00	556.41
1,034	Family	13	2.31	3.29	0.00	12.32

We also took 5,000 draws from the predictive distributions for groups 115, 980, and 1,034, as well as 5,000 draws from a predictive distribution for a new group.

We used the *gibbsit* program described by Raftery and Lewis (1995) and the *Bayesian Output Analysis Program* (Smith 2001) to check for sequence convergence. All diagnostic procedures were indicative of convergent chains.

### 5.3 The Parameters

The LMM estimates for the variance components were  $\hat{\mathbf{G}} = 5.44\mathbf{I}_{1177}$  and  $\hat{\mathbf{R}} = 230.0\mathbf{I}_{40631}$ . Group variability is quite small relative to residual variance. Estimates for the fixed effects in this model are expected mean costs for a subject averaged across groups. Estimates  $\pm$  estimated standard errors were \$2.82 per day  $\pm$  \$0.13 for employee only, \$6.76 per day  $\pm$  \$0.18 for spouse, and \$8.04 per day  $\pm$  \$0.18 for family. The LMM (random) model estimates of the random effects for groups 115, 980, and 1,034 are shown in Table 3.

These random effects are reflective of the large claim costs in groups 115 and 1,034. Many of the groups had random effects more like that of group 980, showing little need to take the group effect into account in predicting next year's values. The large random effects of groups 115 and 1,034 are reflected in the predicted values we see for these groups (see Section 5.4).

In Table 4 we display some summary information for the six policy parameters from the BHM. These parameters describe the underlying likelihoods for the nonzero costs for the three policy types. It is interesting to note that the expected values for these distributions (equal to  $r_{p_j}\theta_{p_j}$ , ignoring group effects) increase as policy types change from employee only to employee plus spouse to employee plus family as we would ex-

pect. However, the variances of the distributions (equal to  $r_{p_j}\theta_{p_j}^2$ , ignoring group effects) are largest for employee plus spouse.

In Table 5 we show some information on the group parameters for the three groups (115, 980, and 1,034) that we chose to examine in more detail.

Bayesian estimators borrow strength from all the data (Gelman et al. 1998). This can perhaps most easily be seen in the mean value of the posterior of  $\pi_{g_{980}}$ , which is 0.15 despite the fact that there are no zero values in this group, which would mean the maximum likelihood estimator for these data would be zero. Thus, the Bayesian estimate has borrowed strength from the other data in producing the posterior that is not centered over zero. It certainly appears from Table 5 that the inclusion of group parameters  $r_{g_i}$  and  $\theta_{g_i}$  was an unnecessary addition to the likelihood. These parameters converged to virtually the same (small) values in each group. Estimating the propensity parameters ( $\pi_{g_i}$ ) for each group, however, appears to be of reasonable importance in describing the data.

### 5.4 Predicted Values

Generating predicted values using the LMM (either random or repeated) can be done automatically using SAS. Proc Mixed will produce a predicted value for any data point with covariate information but no value for the dependent variable.

Generating predicted values for future observations is quite straightforward using the BHM paradigm. At each iteration, take the current draws of the parameters for each group and policy type and use these to make a draw from the likelihood (these draws will include both zero and nonzero values). Using this procedure, 5,000 predicted observations were made for all three policy types in groups 115, 980, and 1,034. To get predicted observations for groups not found in 1994, draw group parameters from the distributions defined for the current draws of the hyperparameters, and use these parameters in the likelihood to draw observations. This technique was used to produce 5,000 draws for each policy type in a "new" group. This method essentially integrates over all sources of uncertainty to produce pre-

Table 3

**Estimates of Random Effects ( $\hat{\mathbf{u}}$ ) and Their Standard Errors for Groups 115, 980, and 1,034 for 1994 Using LMM (random)**

Group	Random Effect	Standard Error
115	12.64	1.79
980	1.50	2.06
1,034	25.79	1.13

Table 4  
**Means, Standard Deviations, and Selected Quantiles of Posterior Distributions  
 from the BHM of Policy Parameters  $\theta_{p_j}$  and  $r_{p_j}$**

Parameter	Mean	Std. Dev.	0.025	0.50	0.975
Employee $\theta_{p_1}$	13.3	0.26	12.8	13.3	13.8
Employee $r_{p_1}$	0.49	0.006	0.47	0.49	0.50
Spouse $\theta_{p_2}$	18.4	0.45	17.5	18.4	19.2
Spouse $r_{p_2}$	0.53	0.009	0.52	0.53	0.55
Family $\theta_{p_3}$	15.3	0.34	14.7	15.3	16.0
Family $r_{p_3}$	0.65	0.010	0.63	0.65	0.67

dicted values that display appropriate levels of variability.

In Table 6 we show the actual mean costs and standard deviations using the 1995 data for the three selected groups and for a “new” group. The “new” group consists of all groups that were represented in the data in 1995 that were not represented in 1994. Eighty-five groups are actually represented in the “new” group. In Table 7 we show model predicted mean costs and estimated standard deviations of the predicted values for both the LMM (random) and the LMM (repeated). Also in Table 7 we display the actual predicted density mean and standard deviation for the BHM for the three selected groups as well as the “new” group.

Again, the point estimates for LMM (random) and LMM (repeated) are theoretically equivalent, yet will sometimes differ only because of computational issues. However, the standard deviations are computed differently. As described in Section 4.1, the variance terms for the LMM (random) does not include the first two terms in equation (4.7). The first term in equation (4.7) is  $\hat{\sigma}_{error}^2 + \hat{\sigma}_{group}^2 = 230.0 + 5.44$ . The second term for all groups seen in 1994 is  $\hat{\sigma}_{group}^2 = 5.44$ . For a group not seen in 1994, the second term is 0. Hence, the standard deviation for LMM (repeated) for groups seen in 1994 is obtained by squaring the entry in

column 4 of Table 7 (the standard deviation for LMM (random)), and adding  $230.0 + 5.44 - 5.44$  and then taking the square root. The answer corresponds to that given in column 6. So for group 115, the LMM (repeated) standard deviation is  $\sqrt{1.79^2 + 230.0 + 5.44 - 5.44} = 15.27$ . For a new group, SAS computes the standard deviation of LMM (repeated) as  $\sqrt{2.34^2 + 230.0 + 5.44 - 5.44} = 15.34$ . This is incorrect. The appropriate standard deviation should be  $\sqrt{2.34^2 + 230.0 + 5.44 - 0.00} = 15.52$ .

It is obvious that both the LMM methods are quite sensitive to the outliers. The predicted values for groups 115 and 1,034 are quite large, and some of the data for these two groups are extreme (from Table 2 the largest value in group 115 is 4.89 standard deviations larger than the mean, and the largest value in group 1,034 is 4.34 standard deviations larger than the mean). These large estimates are reflective of the large random effects for these groups we noted in Table 3. In group 980, the group without many extreme values, the standard deviations of LMM (random) reflect the actual data reasonably well.

The LMM predicted values for the “new” group are functions of the fixed effects only. There is no random effect from the previous years (or shrinkage based on the other estimated groups) to move the predicted values higher. Again, the standard

Table 5  
**Means and Standard Deviations for Posterior Distributions from the BHM of Severity Parameters  
 ( $\theta_{g_i}$  and  $r_{g_i}$ ) and Propensity Parameter ( $\pi_{g_i}$ ) for Three Selected Groups**

Group	Mean( $\theta_{g_i}$ )	SD( $\theta_{g_i}$ )	Mean( $r_{g_i}$ )	SD( $r_{g_i}$ )	Mean( $\pi_{g_i}$ )	SD( $\pi_{g_i}$ )
115	0.102	0.019	0.00009	0.00008	0.49	0.06
980	0.101	0.018	0.00009	0.00008	0.15	0.08
1,034	0.104	0.019	0.00009	0.00008	0.50	0.05

Table 6  
**1995 Raw Data Summary Information for Three Selected Groups, 115, 980, and 1,034, and a “New Group”**

Group	Policy Type	N	Data Mean	Data Std. Dev.
115	Individual	27	6.21	12.54
115	Spouse	3	12.51	10.96
115	Family	2	15.62	18.34
980	Individual	4	0.63	0.74
980	Spouse	2	6.99	0.70
980	Family	4	10.08	6.21
1,034	Individual	57	4.16	12.03
1,034	Spouse	91	7.69	12.98
1,034	Family	10	15.50	21.49
New	Individual	1,318	2.68	12.48
New	Spouse	428	6.15	16.54
New	Family	592	8.65	18.48

Note: Raw data for the “new” group include all groups that existed in 1995 that did not exist in 1994.

deviations of LMM (random) are too small, while those for LMM (repeated) are closer to the actual data.

While the LMM estimates for the groups with highly skewed data are in general too large, the BHM estimates for these same groups tend to be too small. This is reflective of the convergence of the group severity parameters to essentially zero. Thus, the BHM is not effectively using the previous year’s data for the estimation of the predicted values in the groups with highly skewed data. The BHM does estimate different propensity param-

eters for the groups. This is shown in the larger estimates for costs in group 980, a group with high propensity to have claims. The BHM point estimates for the “new” group also seem to be very reasonable. However, the BHM predictive density standard deviations are consistently too small relative to the actual variability in the data.

Since we computed a predicted value for all the 46,691 observations in 1995, we also computed both total error (computed as actual minus predicted), and total error as a percentage of total daily costs for the year. These results are shown in Table 8. Both formulations of the LMM, and the BHM, underestimate actual costs accrued in 1995, although the LMM outperforms the BHM.

## 6. DISCUSSION

We believe these methods offer the actuary effective tools to predict costs when covariate data are limited. All the models have “credibility”-type estimators. While the LMM may be derived in a classical framework (see, e.g., Harville 1977), Laird and Ware (1982) show how the LMM (random) estimators may also be thought of as empirical Bayes. The LMM (repeated) estimators are also shrinkage-type estimators. The BHM is a full Bayesian approach.

It appears that the normal likelihood assumption of the LMM may be too sensitive to the outliers often seen in health care insurance costs.

Table 7  
**1995 Predicted Values and Standard Deviations from LMM (random), LMM (repeated), and Predicted Density Summary Information from the BHM for Three Selected Groups, 115, 980, and 1,034, and a “New Group”**

Group	Policy Type	LMM (random) Predicted Value	LMM (random) Predicted Std. Dev.	LMM (repeated) Predicted Value	LMM (repeated) Predicted Std. Dev.	BHM Predicted Density Mean	BHM Predicted Density Std. Dev.
115	Individual	15.46	1.79	15.44	15.27	3.49	7.61
115	Spouse	19.40	1.79	19.38	15.27	4.92	10.53
115	Family	20.68	1.79	20.66	15.27	5.26	10.41
980	Individual	4.33	2.06	4.33	15.30	5.60	8.95
980	Spouse	8.26	2.07	8.26	15.30	8.26	12.92
980	Family	9.54	2.07	9.54	15.30	8.40	11.88
1,034	Individual	28.61	1.13	28.59	15.21	3.16	7.23
1,034	Spouse	32.55	1.13	32.53	15.21	4.98	10.79
1,034	Family	33.83	1.14	33.81	15.21	5.05	10.20
New	Individual	2.82	2.34	2.82	15.34	4.12	8.02
New	Spouse	6.76	2.34	6.76	15.34	6.98	11.81
New	Family	8.04	2.34	8.04	15.34	7.78	11.77

Table 8  
**Total Error and Percentage Error of Actual Cost Minus Model Predicted Cost for 46,691 Observations in 1995 Using LMM (random), LMM (repeated), and the BHM**

Model	Total Error	Percentage Error
LMM (random)	14,029.48	5.9%
LMM (repeated)	14,023.37	5.9
BHM	26,589.38	11.2

The predictive estimates for groups with outliers tend to be too large. Since group variability is small relative to residual error, the estimates are not “shrunk” back to the mean at the level that appears to be necessary. On the other hand, estimators from groups with small group effects tend to underestimate costs.

The BHM, while computationally more intense, gives the actuary the ability to specify a likelihood structure that may be more realistic for this type of data. However, appropriate determination of the likelihood seems to us to be key. In our example we believe that the gamma likelihood for the severity data is not rich enough to capture the extreme variability present in this type of data. We are currently exploring the generalized beta (McDonald and Xu 1995) density as a possibility. While this five-parameter density may be too rich, leading to difficulties with convergence, it is likely that stepping back from this general form will lead to a more appropriate choice. Also, based on our experience, the inclusion of group parameters in the severity portion of the likelihood seems to complicate the model unduly and is not necessary for the likelihood we used. However, use of group severity parameters might be necessary when the likelihood is more appropriately specified. Inclusion of group parameters for propensity seems to be appropriate.

The Bayesian hierarchical model, used with MCMC technology, allows for the fitting of the kinds of general models necessary to predict costs in the highly volatile health insurance industry. We believe this general framework has great potential for wide applicability in insurance.

**ACKNOWLEDGMENTS**

The authors wish to express their appreciation to the anonymous referees and the editor, whose comments greatly improved the manuscript.

**REFERENCES**

BROPHY, JOHN F., AND LAWRENCE JOSEPH. 2000. A Bayesian Meta-analysis of Randomized Mega-trials for the Choice of Thrombolytic Agents in Acute Myocardial Infarction. In *Meta-analysis in Medicine and Health Policy*, edited by D. Stangl and D. Berry, pp. 83–104. New York: Marcel Dekker.

FREES, EDWARD W., VIRGINIA R. YOUNG, AND YU LUO. 1999. A Longitudinal Data Analysis Interpretation of Credibility Models. *Insurance: Mathematics and Economics* 24: 229–48.

———. 2001. Case Studies Using Panel Data Models. *North American Actuarial Journal* 5(4): 24–43.

GELMAN, ANDREW, JOHN B. CARLIN, HAL S. STERN, AND DONALD B. RUBIN. 1998. *Bayesian Data Analysis*. London: Chapman & Hall.

GILKS, WALTER R. 1995. Full Conditional Distributions. In *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, pp. 75–88. London: Chapman & Hall.

GILKS, W. R., SYLVIA RICHARDSON, AND DAVID J. SPIEGELHALTER, EDS. 1995. *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall.

HARVILLE, DAVID A. 1977. Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *Journal of the American Statistical Association* 72: 320–38.

———. 1990. BLUP (Best Linear Unbiased Prediction), and Beyond. In *Advances in Statistical Methods for Genetic Improvement of Livestock*, pp. 239–76. New York: Springer.

HASTINGS, W. K. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* 57: 97–109.

HENDERSON, CHARLES R. 1984. *Applications of Linear Models in Animal Breeding*. University of Guelph.

HERZOG, THOMAS N. 1999. *Introduction to Credibility Theory*. Winsted, Conn.: ACTEX.

KLUGMAN, STUART A. 1992. *Bayesian Statistics in Actuarial Science with Emphasis on Credibility*. Boston: Kluwer.

LAIRD, NAN M., AND JAMES H. WARE. 1982. Random-Effects Models for Longitudinal Data. *Biometrics* 38: 963–74.

LITTELL, RAMON C., GEORGE A. MILLIKEN, WALTER W. STROUP, AND RUSSELL D. WOLFINGER. 1996. *SAS® System for Mixed Models*. Cary, N.C.: SAS Institute.

MAKOV, UDI E. 2001. Principal Applications of Bayesian Methods in Actuarial Science. *North American Actuarial Journal* 5(4): 53–73.

MCDONALD, JAMES B., AND YEXIAO XU. 1995. A Generalization of the Beta Distribution with Applications. *Journal of Econometrics* 66: 133–52.

MCLEAN, ROBERT A., AND WILLIAM L. SANDERS. 1988. Approximating

- Degrees of Freedom for Standard Errors in Mixed Linear Models. In *Proceedings of the Statistical Computing Section*, American Statistical Association Annual Meetings, New Orleans, pp. 50–59.
- METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER, AND E. TELLER. 1953. Equation of State Calculations by Fast Computing Machine. *Journal of Chemical Physics* 21: 1087–91.
- MORRIS, CARL AND LEE VANSLYKE. 1978. Empirical Bayes Methods for Pricing Insurance Classes. In *Proceedings of the Business and Economics Statistics Section*, American Statistical Association, Annual Meetings, San Diego, pp. 579–82.
- RAFTERY, ADRIAN E., AND STEVEN M. LEWIS. 1995. Implementing MCMC. In *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, pp. 115–30. London: Chapman & Hall.
- SCOLLNIK, DAVID P. 2001. Actuarial Modeling with MCMC and BUGS. *North American Actuarial Journal* 5(2): 96–124.
- SMITH, BRIAN J. 2001. *Bayesian Output Analysis Program (BOA) Version 1.0.0 User's Manual*. Online at [www.public-health.uiowa.edu/boa/](http://www.public-health.uiowa.edu/boa/).
- STANGL, DALENE, AND DON BERRY. 2000. Meta-analysis: Past and Present Challenges. In *Meta-analysis in Medicine and Health Policy*, edited by D. Stangl and D. Berry, pp. 1–28. New York: Marcel Dekker.
- TOLLEY, H. DENNIS, MICHAEL D. NIELSEN, AND ROBERT BACHLER. 1999. Credibility Calculations Using Analysis of Variance Computer Routines. *Journal of Actuarial Practice* 7: 223–28.
- VENTER, GARY G. 1996. Credibility. In *Foundations of Casualty Actuarial Science*, pp. 375–483. Arlington, Va.: Casualty Actuarial Society.

*Discussions on this paper can be submitted until July 1, 2005. The authors reserve the right to reply to any discussion. Please see the Submission Guidelines for Authors on the inside back cover for instructions on the submission of discussions.*